



# The influences of the spatial extent selection for non-landslide samples on statistical-based landslide susceptibility modelling: a case study of Anhui Province in China

Chuhan Wang<sup>1,2</sup> · Qigen Lin<sup>1</sup> · Leibin Wang<sup>3</sup> · Tong Jiang<sup>1</sup> · Buda Su<sup>1</sup> · Yanjun Wang<sup>1</sup> · Sanjit Kumar Mondal<sup>1</sup> · Jinlong Huang<sup>1</sup> · Ying Wang<sup>4</sup>

Received: 24 August 2021 / Accepted: 26 January 2022 / Published online: 18 February 2022  
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

## Abstract

Landslide susceptibility assessment is crucial to the development of appropriate strategies to mitigate the risk of landslide fatalities and economic losses. The selection of spatial extent for non-landslide samples has an important role in the statistical-based landslide susceptibility modelling (LSM). In this study, two different non-landslides sampling areas (the entire area and the mountainous area of Anhui Province, China) were designed to explore the influences of the different spatial extent for non-landslides sampling on LSM. Six categories of influencing factors including climatic, morphological, geological, hydrological, vegetation, and human activities were considered. The dominant influencing factors that are more closely related to the distribution of historical landslides were selected based on the GeoDetector. The landslide inventory samples and the non-landslide samples generated on two selected areas were divided into a training set (70%) and a validation set (30%) for establishing the entire area LR model (EareaLR) and the mountainous area LR model (MareaLR) based on the logistic regression (LR) model. The performance of the models was evaluated by the confusion matrix and the area under the receiver operating characteristic curve (AUROC). The results showed that the EareaLR model outperformed the MareaLR model by various evaluation metrics and the appearance of the final landslide susceptibility map. Hence, we conclude that the potential influence of the spatial extent of the non-landslide sample selection needs to be taken into account while comparing the reliability of different data-driven landslide susceptibility models.

**Keywords** Non-landslide · Sampling strategy · Landslide susceptibility · GeoDetector · Logistic regression (LR) model · China

---

✉ Qigen Lin  
linqigen@nuist.edu.cn

Extended author information available on the last page of the article

## 1 Introduction

Landslide disasters are widely distributed around the world and resulting in serious casualties (Petley 2012; Froude and Petley 2018). In recent years, recorded non-seismic induced landslides cause approximately 4300 deaths worldwide each year (Petley 2012; Froude and Petley 2018). China is one of the countries where landslide disasters are most frequent. There were 88,084 reported landslides, resulting in 3226 injuries and 4856 deaths, and severe direct economic loss (approximately 4.5 billion Chinese Yuan) during the period from 2010 to 2019 (National Bureau of Statistics 2020). Statistics of the Fatal Landslide Event Inventory of China (FLEIC) show that there were about 650 fatalities each year resulting from landslides during the period from 2000 to 2018, which accounts for about a quarter of all fatalities caused by natural disasters in the same period (Lin and Wang 2018; National Bureau of Statistics 2020). For Anhui Province, there were 23 reported rainfall triggered fatal landslides resulting in 117 fatalities according to the FLEIC. The recorded fatal landslides in the past ten years have become more frequent (Lin and Wang 2018). Moreover, the landslide occurrence is projected to increase due to more extreme precipitation events under climate change scenarios and more exposed populations under the rapid urbanization processes in mountainous areas (Gariano and Guzzetti 2016; Lin et al. 2017b; Lin et al. 2020; Zhang et al. 2021). Therefore, it is crucial to better identify the high-risk areas of landslide in Anhui Province for formulating effective landslide risk mitigation measures.

Landslide susceptibility is defined as the propensity for landslides to happen in an area, i.e., the spatial probability of landslide occurrence (Guzzetti et al. 1999). Landslide susceptibility modeling is an assessment of the spatial probability of landslide occurrence on a single landslide or the areas of different spatial scales using qualitative or quantitative methods, by considering many landslides environmental factors (Guzzetti et al. 1999; Guzzetti et al. 2005; Guzzetti et al. 2006). At present, statistical-based (Logistic regression, LR; Generalized additive model, GAM, etc.) and machine learning-based (Random Forest, RF; Support Vector Machine, SVM, etc.) landslide susceptibility mapping is the most widely used methods (Brenning et al. 2015; Steger et al. 2016, 2017; Lin et al. 2017a; Xiao et al. 2020; Zhou et al. 2021). In the process of a statistical landslide susceptibility modeling, many aspects such as the sampling strategies for landslides and non-landslides (Regmi et al. 2014; Nefeslioglu et al. 2008; Hussin et al. 2016; Pourghasemi et al. 2020; Shao et al. 2020; Sameen et al. 2020), the selection of the landslide influencing factors (Iovine et al. 2014), the quality of the landslide inventory database (Steger and Glade, 2017; Lin et al. 2021), the selection of statistical classification method (Devkota et al. 2013; Goetz et al. 2015) and the evaluation indicators (Chung and Fabbri 2003; Guzzetti et al. 2006; Frattini et al. 2010) would influence the final results. Among these studies, it has been demonstrated that the non-landslide sample selection would have a significant effect on landslide susceptibility models (Shao et al. 2020; Xie et al. 2021; Zhou et al. 2021). However, the potential effects of spatial extent selection for the non-landslide samples on the statistical-based landslide susceptibility modeling have seldom been investigated (Steger and Glade 2017). Thus, the hypothesis that this study attempts to investigate is whether the selection of the spatial extent for non-landslide samples will have an influence on the validation results of the statistical landslide susceptibility model, the modelled relationship of the explanatory factors, and the appearance of the derived susceptibility map. For this purpose, the Anhui Province, located in East China is selected as the study area, which has diverse geomorphological types and apparent spatial heterogeneity of landslide distribution. Then,

the dominant influencing factors of the landslide were selected based on the GeoDetector (Wang and Xu 2017). The respectively generated non-landslide samples from the entire study area (i.e., region-wide including flatter terrain areas and mountainous area) and the mountainous study area combined with landslide samples to form the training and validation datasets of the entire area and the training and validation datasets of mountainous area, respectively, and introduced into the Logistic Regression (LR) model to build the entire area LR model (EareaLR) and the mountainous area LR model (MareaLR). Finally, the performances of the two landslide susceptibility models based on different non-landslide sampling spatial extents are explored using multiple evaluation indicators, and the influence of the selection of the spatial extent for non-landslide samples on statistical-based landslide susceptibility models is discussed.

## 2 Study area and dataset

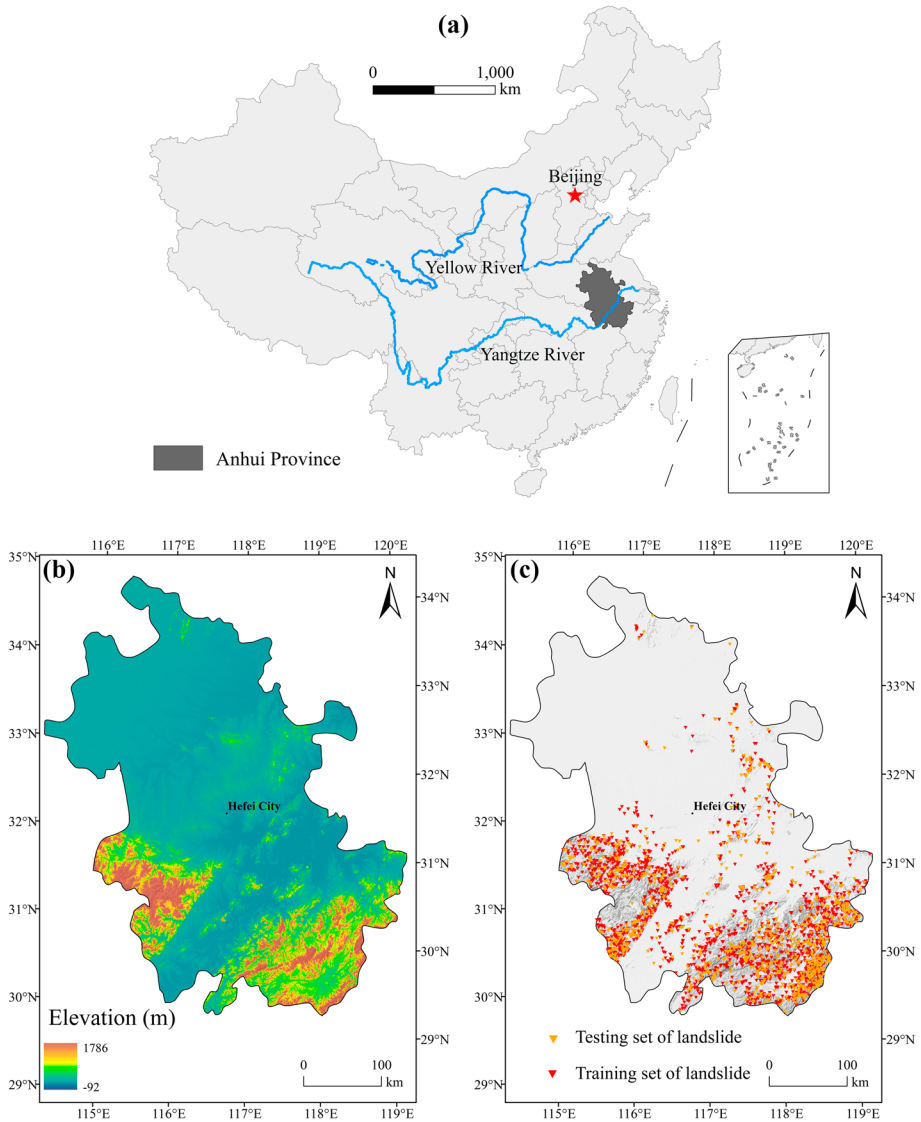
### 2.1 Study area

Anhui Province is in the East China, surrounded by the Huang-Huai-Hai Plain in the southern part, the eastern side includes the extension of the Qinling Mountains, the plain in the middle and lower reaches of the Yangtze River basin and the Jiangnan hills are in the northern part, covering an area of about 140,000 km<sup>2</sup>. It has various geomorphic types, such as high terrain in the south, low terrain in the north, and steep mountains in mountainous areas. The elevation of the study area varies from – 92 to 1786 m above sea level (Fig. 1b). The Yangtze and Huaihe rivers are the main drainage systems in the study area. The climate regime is divided into two categories with the Huai River as the dividing line. The north belongs to the warm, temperate, semi-humid monsoon climate, while the south belongs to the subtropical humid monsoon climate. The mean annual precipitation is approximately 800–1600mm, which is significantly affected by the monsoon climate, with continuous rainfall and heavy rainfall during the rainy season.

Landslides have occurred frequently in Anhui Province in recent years. For instance, multiple landslides were triggered by heavy rainfall in Yuexi and Jinzhai counties in July 2016 (He et al. 2019). In 2017, landslide was triggered by continuous rainfall in Jixi county in June. Landslide in Taihu Province Road in March and landslides caused by strong rainfall in Jinzhai and Huoshan counties in June in 2020. In 2021, a landslide caused by heavy rainfall under the influence of a typhoon in Ningguo city in July. The recorded landslides are mainly concentrated in the mountainous areas of southern and southwestern Anhui Province (Fig. 1c).

### 2.2 Landslide inventory and influencing factors

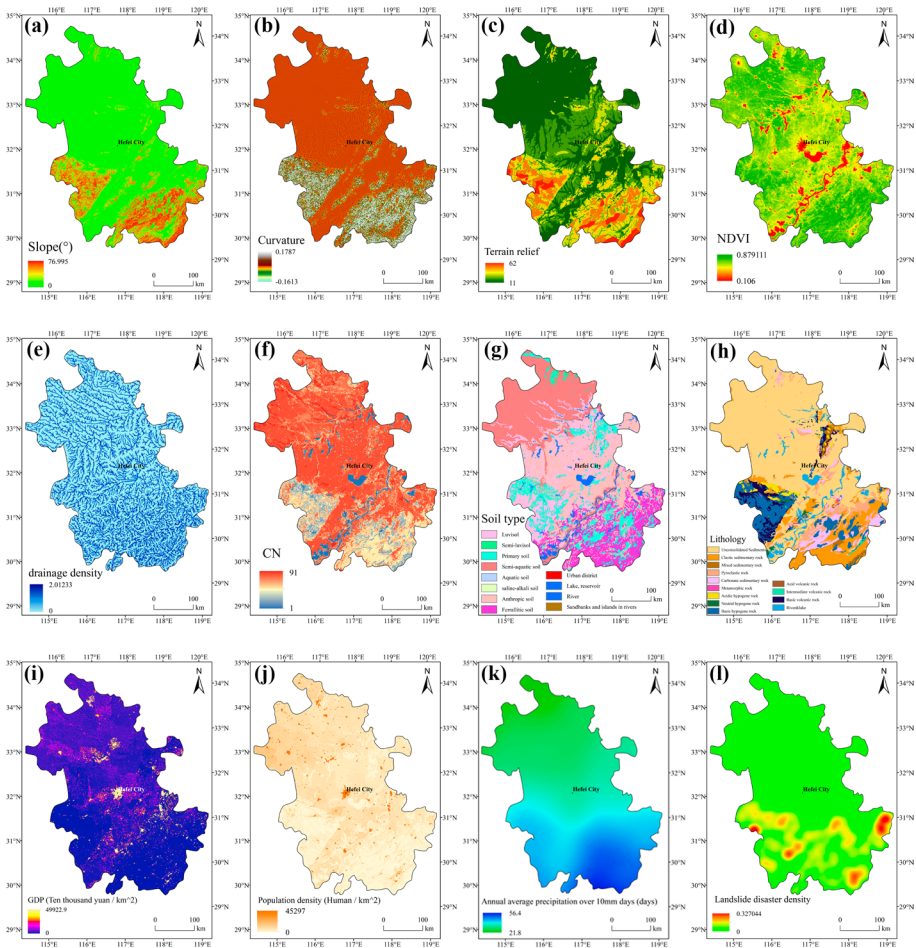
The historical landslide inventory in this study was obtained from the National Geological Hazard Detailed Survey Database conducted by the China Geological Survey. The China Geological Survey investigated six categories of geological hazards in detail, including landslides, debris flow, rockfall, ground fissure, ground subsidence, and ground collapse. A total of 3286 rainfall-triggered landslides of the slide-type movement obtained from the National Geological Hazard Detailed Survey Database are used for statistical landslide susceptibility mapping in this study. Each landslide is mapped as a point and contains information including the latitude, longitude, geographical location, landslide type, geological



**Fig. 1** Study area of Anhui Province, China and historical landslide inventory **a** Location of the study area **b** topography of Anhui Province **c** spatial distribution of historical landslide inventory in Anhui Province

environmental conditions, natural environment conditions, and causes of the landslide (Lin et al. 2021). The spatial distribution of the landslide inventory is shown in Fig. 1c. The landslide density derived from the landslide inventory (Fig. 2l) is used to analyze the relationship with the distribution of each influencing factor in the GeoDetector model, and is not used as the landslide predictor in the landslide susceptibility model.

The landslide environmental factors can be divided into climatic factors, geological and soil factors, topographic factors, vegetation coverage factors, hydrological factors, and human activities factors (Reichenbach et al. 2018). Climatic factors are mainly used



**Fig. 2** Spatial distribution of landslide density and influencing factors **a** Slope **b** curvature **c** terrain relief **d** NDVI **e** drainage density **f** CN **g** soil type **h** lithology **i** GDP **j** population density **k** average annual precipitation over 10 mm days **l** landslide density

to represent the spatial variation in precipitation patterns that trigger landslides. The precipitation indicators including average annual precipitation (1981–2010), maximum annual daily precipitation, annual average daily maximum precipitation, annual average precipitation over 10mm (days), annual average precipitation over 25mm (days), and annual average precipitation over 50mm (days), annual average precipitation above 95% quantiles (days) and annual 95% quantile daily precipitation were calculated based on CN05.1 grid observation daily precipitation data (Wu and Gao 2013). Since the spatial patterns of these factors are relatively similar, this study here only presents the annual average precipitation over 10 mm (days) (Fig. 2k), which is more closely related to the distribution of landslides.

Geological and soil factors include soil types and lithology. Soil type affects the degree of drainage, erosion, and damage which influence the occurrence of landslides (Nhu et al. 2020). Lithology is often regarded as one of the significant factors in landslide susceptibility modeling due to it directly affects the strength and permeability of rock and soil

(Kavzoglu et al. 2014). The soil type data was obtained from the Resource and Environmental Science and Data Center (RESDC) and the lithology data was obtained from the Global Lithological Map Database (Hartmann and Moosdorf 2012). Morphological factors are used to characterize the geomorphic characteristics of landslide occurrence. The digital elevation model data in this study was the SRTM 30 m DEM (USGS 2015). Three terrain indicators namely slope, terrain relief, and curvature were derived based on the DEM (Fig. 2a and b, and c). The Normalized Difference Vegetation Index (NDVI) was selected to represent the vegetation coverage that influences the occurrence of landslides (Choi et al. 2012). The annual average normalized vegetation index (NDVI) (1981–2010) product with a spatial resolution of 1 km was obtained from the RESDC. Curve number (CN) and drainage density were selected as the hydrological factors. CN is a comprehensive dimensionless parameter that can reflect the characteristics of the underlying surface of the basin before rainfall, it is related to soil type, previous soil moisture, land use, slope, and other influencing factors (Wang and Xu 2017). And the greater the drainage density, the greater the impact of river slope cutting and the higher the possibility of landslide occurrences. The drainage density data was obtained from the HydroSHEDS data set (Linke et al. 2019) and the CN data was obtained from Zhao et al. (2018). Human activities also have significant influences on landslide occurrence (Wang et al. 2018). The high population density and the production of GDP are often accompanied by the exploitation and destruction of natural resources (Wang et al. 2018). Two indicators including gross domestic product (GDP) and population density were selected as a proxy to represent the potential influences of human activities on the landslide occurrence. The GDP and population density data were obtained from the RESDC.

### 3 Methods

The methodological framework is shown in Fig. 3, it mainly includes four steps. Firstly, the dominant influencing factors with high correlation on landslide distribution are analyzed and selected based on the GeoDetector (Wang and Xu 2017). Then, the selected influencing factors are introduced into the logistic regression (LR) model to establish the landslide models based on different non-landslide samples generated from the entire area and the mountainous area of Anhui Province, separately. Thirdly, the model performance is evaluated by using confusion matrix, ROC curve and k-fold cross-validation. Finally, the established model is applied to landslide susceptibility mapping, and the influence of the spatial extent selection for non-landslide samples on statistical-based landslide susceptibility mapping is compared and discussed. Each method is described in detail in the following sections.

#### 3.1 Selection of landslide influencing factors based on GeoDetector

GeoDetector is developed by Wang and Xu (2017) to detect spatial differentiation based on statistical principles. It is a tool for detecting the spatial differentiation which can quantitatively reveal the driving forces behind spatial differentiation (Wang et al. 2010). GeoDetector detects the interpretation degree of the spatial distribution of the dependent variable  $Y$

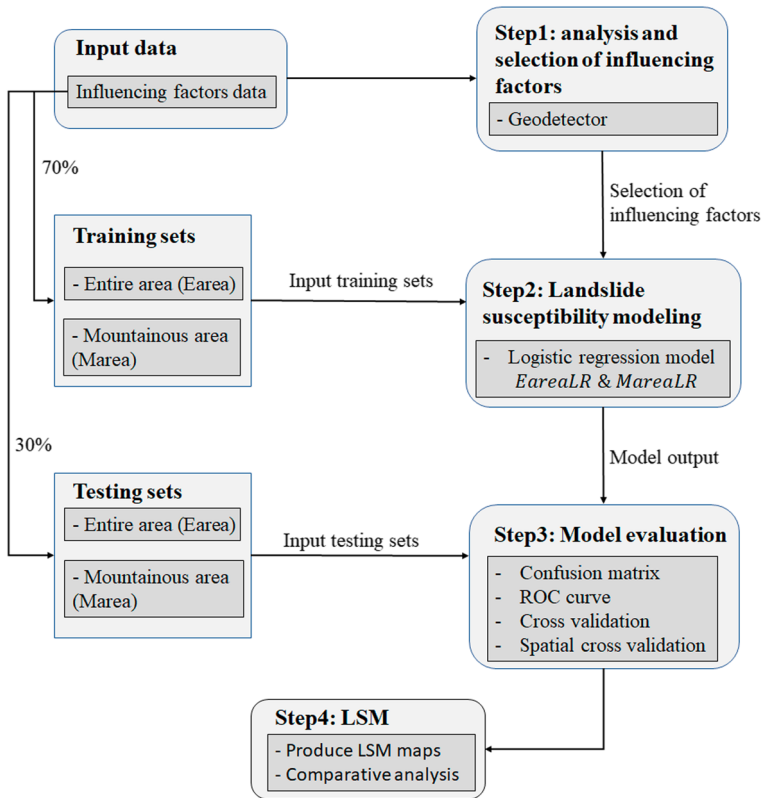


Fig. 3 The methodological framework of this study

by the respective variables  $X_i$ , and tests the confidence degree of the respective variables. The degree of interpretation is represented by  $q$ . Equation (1) is as follows:

$$q = 1 - \frac{1}{N\sigma^2} \sum_{h=1}^L N_h \sigma_h^2 = 1 - \frac{SSW}{SST}, q \in [0, 1] \tag{1}$$

where  $SSW = \sum_{h=1}^L N_h \sigma_h^2$ ,  $SST = N\sigma^2$ ;  $h = 1, \dots, L$  refer to the stratification of variable  $Y$  or factor  $X$ , that is, classification or partition,  $N_h$  refers to the number of units in layer  $h$  and  $N$  refers to the number of units in the entire area,  $\sigma_h^2$  and  $\sigma^2$ , respectively refer to the variance of the  $Y$  value of the layer  $h$  and the variance of the  $Y$  value of the entire region.  $SSW$  represents the within the sum of squares (SSW) and  $SST$  represents the total sum of squares (SST).

The closer the  $q$ -value is to 1, that is, the larger the  $q$ -value, the more significant the spatial distribution of  $Y$  are explained by the self-variable  $X_i$ .

A simple transformation of  $q$  satisfies the non-central F distribution with the following Eqs. (2) and (3):

$$F = \frac{q(N - L)}{(1 - q)(L - 1)} \sim F(L - 1, N - L; \lambda) \tag{2}$$

$$\lambda = \frac{1}{\sigma^2} \left[ \sum_{h=1}^L \bar{Y}_h^2 - \frac{1}{N} \left( \sum_{h=1}^L \sqrt{N_h} \bar{Y}_h \right)^2 \right] \quad (3)$$

where the  $\lambda$  refers to a non-central parameter,  $\bar{Y}_h$  refers to the mean value of  $Y$  value of the  $h$  layer. According to Eq. (3), the GeoDetector software can be used to evaluate whether the  $q$ -value is significant.

The relationship between the spatial distribution of historical landslides and the candidate landslide environmental factors in Anhui Province is quantitatively analyzed by GeoDetector. The spatial distribution of historical landslides (Fig. 21) as the dependent variable in GeoDetector is produced with a spatial resolution of 1 km based on the ArcGIS spatial analysis tool kernel density. At the same time, all candidate landslide environmental factors were resampled to the same spatial resolution of 1 km. Regarding the categorical independent variables, the relationship between them and landslide density was analyzed directly by the GeoDetector. For continuous independent variables, the classification number was set to 3–10 categories, and three classification methods of quantile classification, natural breaks classification, and isometric classification were used in turn. The optimal discrete classification method and the number of classification categories for continuous variables are determined by the highest  $q$  value. Then the relationship between the historical landslide density and the landslide environmental factors under each optimal discrete classification is analyzed.

### 3.2 Logistic regression model

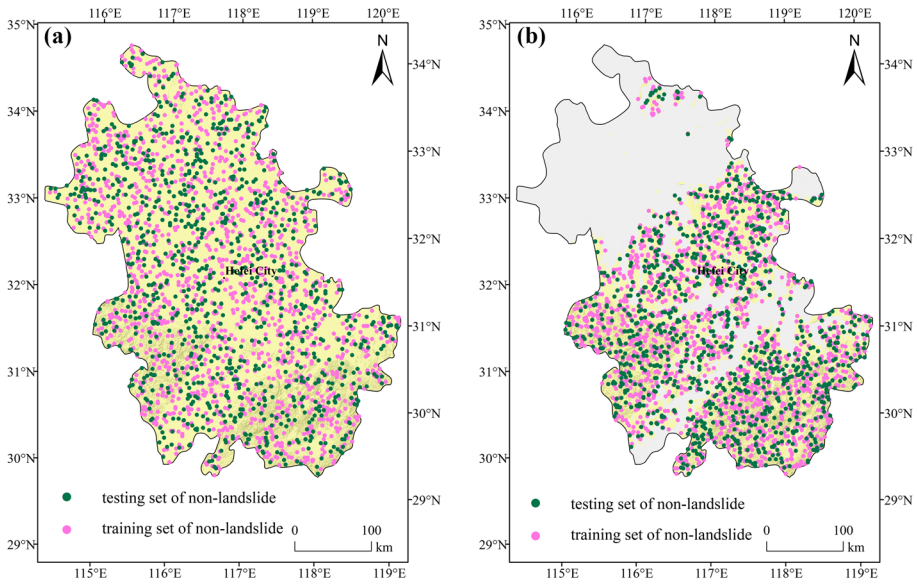
Logistic regression (LR) model is a kind of regression analysis with a binary dependent variable. In the landslide susceptibility assessment, various landslide environmental factors can be used as categorical or continuous independent variables (Wang et al. 2016), and the occurrence of landslide can be represented by 0 (absence of landslide) and 1 (presence of landslide) as dependent variable (Kalantar et al. 2018; Ma et al. 2019). The advantage of the LR model is that the dependent variable can be binary, and the result ranges from 0 to 1. Independent variables can be continuous or categorical variables. The continuous variables can be directly added to the model, and categorical variables can be transformed into dummy variables, and then, introduced into the model. The formula of the LR model is as follows:

$$P(y = 1|x_1, \dots, x_i) = \frac{e^{\epsilon_i}}{1 + e^{\epsilon_i}} \quad (4)$$

$$\epsilon_i = \alpha + \sum_{k=1}^k \beta_k x_k \quad (5)$$

where  $P(y = 1|x_1, \dots, x_i)$  refers to the occurrence probability of the event,  $x_i$  is the independent variable,  $\epsilon_i$  refers to the linear function of the factors influencing the occurrence of the event,  $\alpha$  refers to the logarithm of the ratio of event occurrence to non-occurrence without the influence of any factor,  $\beta_k$  refers to the logistic regression coefficient representing the change value of the occurrence/non-occurrence probability ratio when the independent variable changes. According to this model, the occurrence probability  $P$  of landslide is





**Fig. 4** Non-landslide sampling spatial extent on **a** the entire area and **b** the mountainous area

assumed to be [0, 1]. The occurrence probability of landslide is the dependent variable, the landslide environmental factors  $x_1, \dots, x_i$  are the independent variables. The logistic regression model of the occurrence probability of landslide is as follows:

$$P = \frac{e^{\alpha + \sum_{k=1}^k \beta_k x_k}}{1 + e^{\alpha + \sum_{k=1}^k \beta_k x_k}} \tag{6}$$

Training set and validation set are needed for statistical-based landslide susceptibility modeling. For this study, the landslide inventory was randomly divided into 70% (2300 landslides) of the training dataset and the remaining 30% (986 landslides) of the validation dataset (Steger et al. 2021). Regarding the non-landslide samples, the same number of non-landslides were generated for training and validation datasets, respectively from different spatial extents (i.e., the selected entire area and the mountainous area of Anhui Province) as shown in Fig. 4a and b. The extent of the mountainous area for this study was derived from a geomorphological classification map obtained from the RESDC. The GeoDetector is implemented based on the “GD” package in R programming language with the version of 3.6.3 and the LR model is implemented based on the “stats” package in R (Song et al. 2020; R Core Team 2020).

### 3.3 Evaluation methods

The confusion matrix, receiver operating characteristic (ROC) curve, k-fold non-spatial cross-validation (CV), and k-fold spatial cross-validation (SCV) were used to evaluate the predicted performance of the established landslide susceptibility models (Frattini et al. 2010). The match ratio between the prediction value of the models and the corresponding actual value is presented in a table called the confusion matrix (Table 1). The ROC curve takes each prediction result as a possible judgment threshold and calculates

**Table 1** Confusion matrix

Actual value	Predicted value		Percentage correct
	0	1	
0	$f_{11}$	$f_{12}$	$\frac{f_{11}}{f_{11}+f_{12}}$
1	$f_{21}$	$f_{22}$	$\frac{f_{22}}{f_{21}+f_{22}}$
Overall percentage	$\frac{f_{11}+f_{22}}{f_{11}+f_{12}+f_{21}+f_{22}}$		

the corresponding sensitivity and specificity according to the threshold, and it draws the abscissa with the false positive rate (1-specificity) and the ordinate with the true positive rate (sensitivity) (Frattini et al. 2010). The area under the ROC curve is the AUROC value. The AUROC value range is [0.5, 1]. The larger the value is, the higher the predicted ability of the model is (Frattini et al. 2010). The  $K$ -fold non-spatial cross-validation refers to randomly divide the original data set into  $K$  subsets with the same number of samples, and combine  $K-1$  of them as the training data set to train the model in turn. The remaining one subset is the test dataset. The  $MSE_i$  of the model on the test dataset is calculated in each test, and the  $MSE_i$  obtained after  $k$  times of test are averaged to obtain the  $MSE$  to evaluate the robustness of prediction performance (Brenning 2012). The  $K$ -fold spatial cross-validation refers to randomly divide the original data set into  $K$  subsets with the same number of samples by considering the spatial location of each sample. Then the same as the non-spatial CV, the average of the MSE was calculated to evaluate the robustness of prediction performance by considering the spatial distribution of the samples

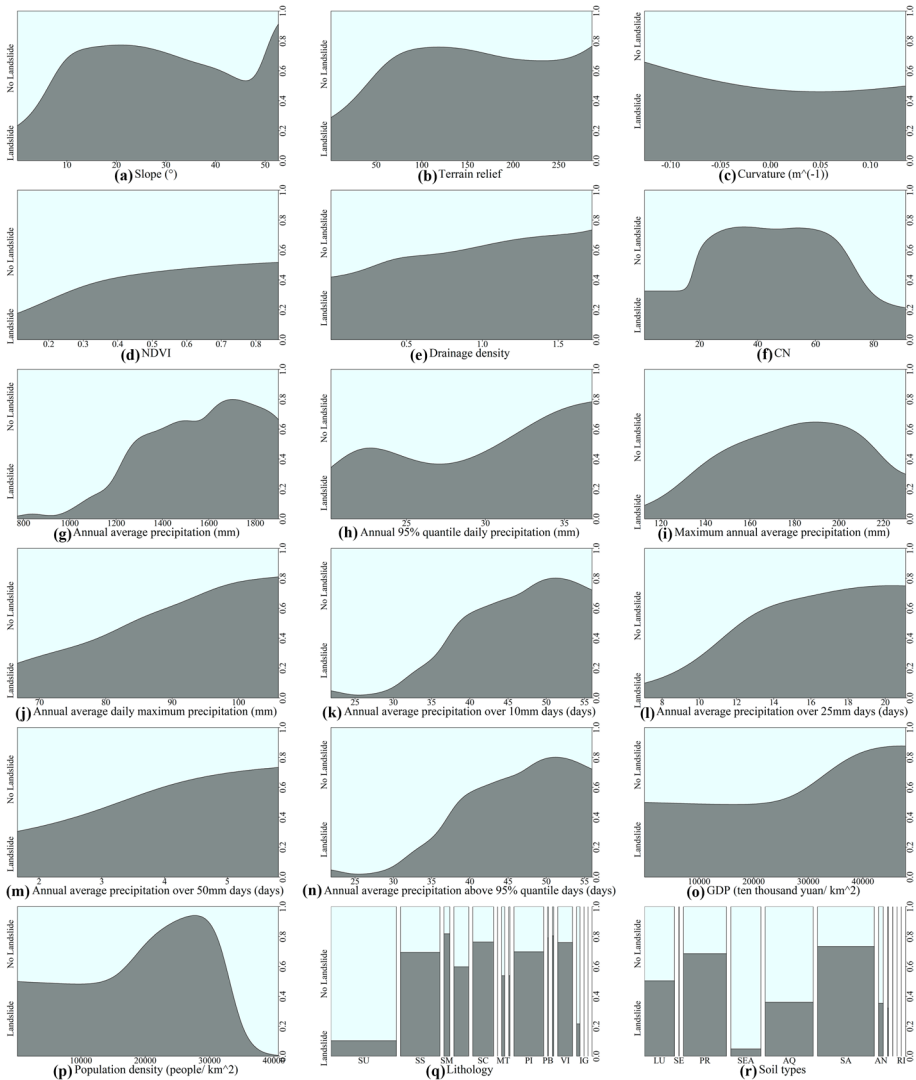
$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k MSE_i \quad (7)$$

## 4 Results and discussion

### 4.1 Analysis of influencing factors of landslide susceptibility

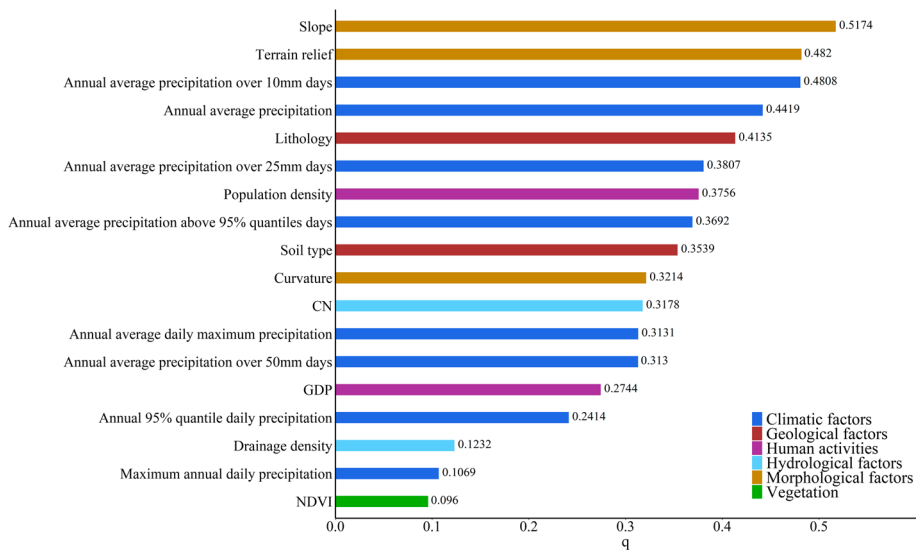
#### 4.1.1 Relationship between landslide density and various influencing factors

In terms of morphological factors, Fig. 5 shows that the frequency of landslides in Anhui Province increases with the increase in slope and terrain relief. For the curvature, the landslide appears more in the areas where the curvature is concave or convex. For hydrological factors, most of the landslides occur in areas with a CN value of 20–70, and the density of landslides rises as the drainage density increases. In terms of vegetation coverage, the frequency of landslides increases with the increase of NDVI. In terms of the climatic factors, the greater the annual average precipitation and the higher the number of extreme precipitation days, the higher the occurrence of landslides. Among the soil types, landslides mainly occur in luvisol, primary soil, and ferrallitic soil areas. For lithology, landslides



**Fig. 5** Conditional frequency plots displaying empirical relations between landslide and various environmental factors **a** Slope **b** terrain relief **c** curvature **d** NDVI **e** drainage density **f** CN **g** annual average precipitation **h** annual 95% quantile daily precipitation **i** maximum annual average precipitation **j** annual average daily maximum precipitation **k** annual average precipitation over 10mm days **l** annual average precipitation over 25mm days **m** annual average precipitation over 50mm days **n** annual average precipitation above 95% quantile days **o** GDP **p** population density **r** lithology **s** soil type

mainly occur in sedimentary rocks, hypogene rocks, and volcanic rock areas. For population density, the conditional probability of a landslide shows an increase followed by a decrease with increasing population density. While for GDP, the conditional probability of landslide increases with the increase of GDP.



**Fig. 6** The  $q$ -value of each landslide environmental factors

#### 4.1.2 The $q$ -value of each landslide environmental factor

The effects of landslide influencing factors with the spatial distribution of landslide density in Anhui Province were detected using the GeoDetector (Fig. 6). The  $q$ -value of each environmental factor is greater than zero and significant with a 95% confidence level, indicating that there is a significant relationship between the zoning of each landslide environmental factor and the spatial distribution of the landslides in Anhui Province. The  $q$ -value of each influencing factor is sorted from large to small in Fig. 6. The first two variables with the strongest interpretation of the spatial distribution of landslides in Anhui Province are morphological factors, which reflects that the morphology has the most obvious influence on landslides in Anhui Province. The following is the annual average precipitation days greater than 10mm and the average annual precipitation, these two environmental factors reflect the spatial pattern of extreme precipitation and long-term precipitation in Anhui Province, indicating that rainfall triggering conditions have a significant influence on the occurrence of landslide in Anhui Province.

For the selection of factors environmental landslides, if the  $q$ -value of the GeoDetector was used as the only criteria, more than half of the selected environmental factors would be climatic factors. This would result in multicollinearity during the modeling process, which may make the model prediction distorted. To avoid such problems, the dominant landslide environmental factors were selected according to their categories as follows: Climatic (precipitation), geological and soil (soil and lithology), topography (slope, terrain relief and curvature), vegetation coverage (NDVI), hydrological (drainage density and CN), human activities (population density and GDP). In each category, the  $q$  value was compared, and the dominant landslide environmental factor with the highest  $q$  value in each category was selected to enter the model training. Finally, the dominant landslide environmental factors were selected as follows: annual average precipitation days greater than 10mm, lithology, slope, NDVI, and population density.

The dominant landslide influencing factors were selected in this work based on the Geo-Detector which has been applied for identifying the conditional factors of landslide susceptibility modeling (Yang et al. 2019; Rong et al. 2020; Xie et al. 2021). The results in this study show that the slope and extreme precipitation are most closely related to the spatial distribution of historical landslides in Anhui Province. While Xie et al. (2021) show that the elevation, slope and land use are the most relevant factors for landslide susceptibility mapping (LSM) in Xiaojin County, Sichuan Province, China. Yang et al. (2019) indicate that the most closely related landslide environmental factors in the surrounding areas of Duwen highway in Sichuan Province are lithology, slope, elevation, seismic intensity, and distance from settlements. Rong et al. (2020) show that the dominant influencing factors of the occurrence of landslides in Shuicheng County, Guizhou Province, China are elevation, slope, lithology, and distance to faults. The comparison with the existing studies shows that the dominant influencing factors of the spatial distribution of historical landslides have significant spatial heterogeneity in different regions. The GeoDetetor is applicable to identify the dominant influencing factors in different study areas. However, selecting the dominant factors into the model purely by the order of importance obtained by GeoDetetor may lead to some co-linearity troubles. It would be more appropriate to incorporate the heuristics to select the dominant factors that are eventually employed for modeling.

## 4.2 Landslide susceptibility modeling results

### 4.2.1 Landslide susceptibility models

The five dominant influencing factors were finally selected and introduced into the logistic regression model. For lithology, the Unconsolidated Sediments are regarded as the reference category. The coefficients of the influencing factors of the EareaLR model and the MareaLR model are shown in Table 2. The results of the variables of the two models show that the slope and the annual average precipitation days greater than 10mm will increase the occurrence possibility of landslide, and the NDVI will reduce the occurrence possibility of landslide. Regarding the lithological types, most lithology types such as acidic volcanic rock, basic volcanic rock, and mixed volcanic rock will increase the probability of landslide occurrence compared with unconsolidated sedimentary rock. The two models are in line in the direction of the influence of each factor on landslide occurrence. However, comparing the coefficient sizes of the two models reveals that the coefficients of MareaLR model are significantly lower than those of EareaLR model for each lithological category and slope. This is mainly because the EareaLR model selects the non-landslide samples in the entire region (i.e., region-wide including flatter terrain areas) so that the non-landslide samples are more commonly located in the lithological category of Unconsolidated Sediments (Fig. 2h) where landslides are less likely to occur and also in areas with lower slopes (Fig. 2a). It results in the EareaLR model having a better explanatory capability in terms of the predictors lithology and slope.

### 4.2.2 Evaluation results of the landslide susceptibility models

The confusion matrix reflects the prediction accuracy of the landslide susceptibility model. By setting the threshold value to 0.5, the model prediction result  $P$  is divided into the occurrence of landslide ( $P \geq 0.5$ ) and non-occurrence of landslide ( $P < 0.5$ ) (Xiong et al. 2019). The accuracy of the confusion matrix of the training and validation set for the

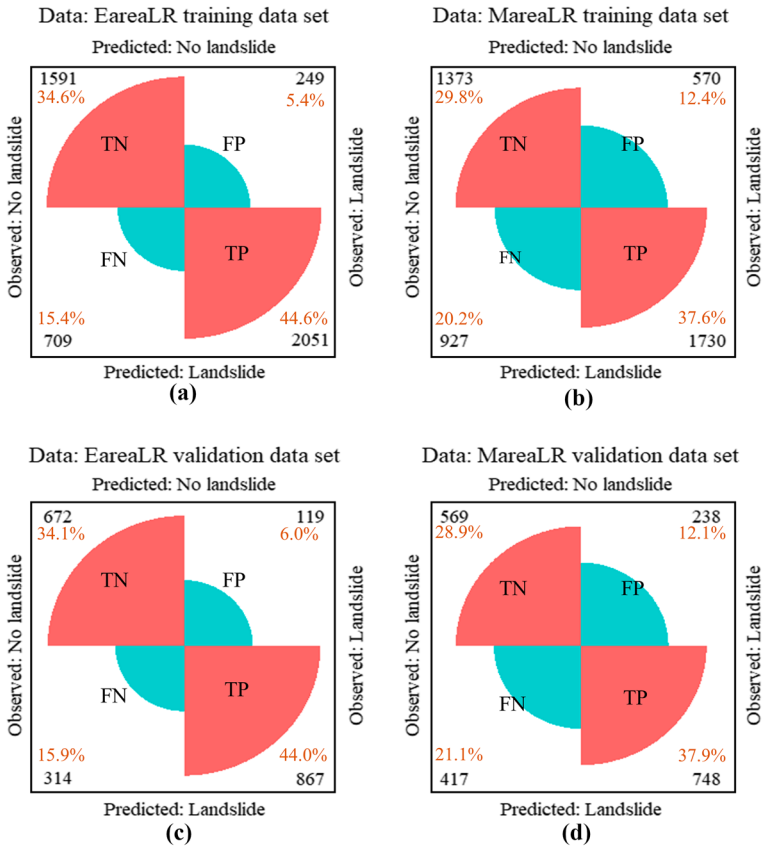
**Table 2** Coefficients of the influencing factors

Independent variable	EareaLR			MareaLR		
	$\beta$	Sig.	Exp( $\beta$ )	$\beta$	Sig.	Exp( $\beta$ )
Intercept	-3.755	0.0000	0.023	-2.732	0.0000	0.065
Slope (°)	0.026	0.0000	1.027	0.005	0.3399	1.005
Annual average precipitation days greater than 10 mm (days)	0.114	0.0000	1.121	0.104	0.0000	1.110
Lithology		0.0000			0.0000	
Clastic sedimentary rock	1.179	0.0000	3.251	0.822	0.0000	2.276
Mixed sedimentary rock	2.182	0.0000	8.864	1.249	0.0000	3.487
Pyroclastic rock	1.451	0.0000	4.267	1.006	0.0000	2.735
Carbonate sedimentary rock	1.862	0.0000	6.437	1.566	0.0000	4.787
Acidic hypogene rock	1.805	0.0000	6.080	0.974	0.0000	2.650
Neutral hypogene rock	1.757	0.0000	5.795	0.719	0.0477	2.053
Basic hypogene rock	1.789	0.0000	5.983	1.251	0.0000	3.494
Acid volcanic rock	2.444	0.0000	11.519	2.405	0.0003	11.078
Intermediate volcanic rock	1.938	0.0009	6.945	1.512	0.0004	4.536
Basic volcanic rock	2.354	0.0000	10.528	1.961	0.0000	7.106
Population density (human/km <sup>2</sup> )	0.000	0.2950	1.000	0.000	0.3190	1.000
NDVI	-3.111	0.0000	0.045	-3.647	0.0000	0.026

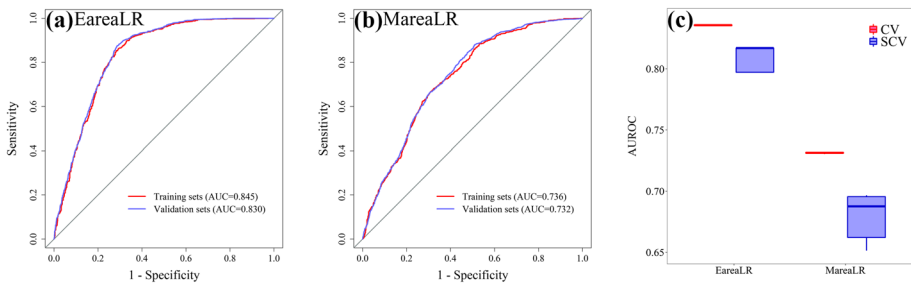
EareaLR model and the MareaLR model are shown in Fig. 7. It is found that the accuracy of the EareaLR model is much higher than that of the MareaLR model. The overall accuracy is 79.2% of the training dataset and is 78.1% of the validation dataset for the EareaLR model. While the overall accuracy of the training set of the MareaLR model is only 67.4%, and the corresponding accuracy of the validation set is only 66.8%. Compared with the MareaLR model, the classification accuracy of the EareaLR model is improved by 11.8% in the training dataset and 11.3% in the validation dataset. This shows that the classification accuracy of the training dataset and validation dataset of the EareaLR model is better than that of the MareaLR model.

The ROC curves of the EareaLR model and the MareaLR model are shown in Fig. 8. The area under the ROC curve (AUROC) of the training dataset and validation dataset of the MareaLR model are 0.736 and 0.732, respectively. Compared with the MareaLR model, the EareaLR model has higher accuracy in landslide susceptibility modeling. The AUROC of the training dataset and the validation dataset of the EareaLR model are 0.845 and 0.830, respectively, which are 10.9 and 9.8% higher than that of the MareaLR model. This shows that compared with the spatial selection of non-landslide samples in the mountainous area of Anhui Province, the landslide susceptibility model trained by the spatial selection of non-landslide samples in the entire area of Anhui Province has a higher prediction accuracy.

To further evaluate the robustness of the models, the CV and SCV (5 folds 30 times, total 150 times) were used to evaluate the performance of the EareaLR model and the MareaLR model. The results are shown in Fig. 8. The CV average of AUROC for the EareaLR model is 0.834, and the SCV average of AUROC is 0.797. The CV average of AUROC for the MareaLR model is 0.732, and the SCV average of AUROC is 0.678. Compared with the MareaLR model, the CV and SCV average of AUROC for the EareaLR model

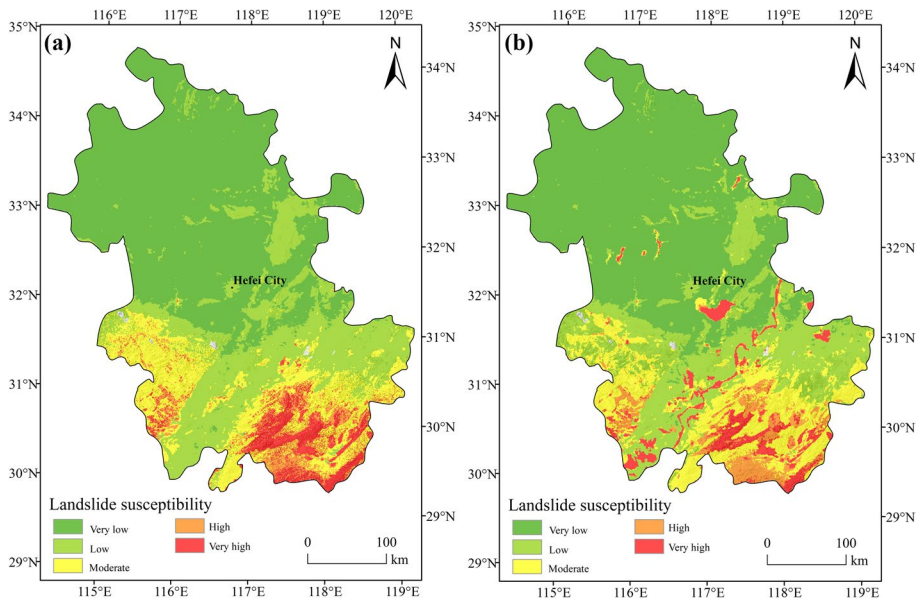


**Fig. 7** The confusion matrix of the training data set of **a** EareaLR and **b** MareaLR and the validation data set of **c** EareaLR and **d** MareaLR



**Fig. 8** The ROC curves of **a** the EareaLR model and **b** the MareaLR model and **c** the CV and SCV of the EareaLR and the MareaLR model

are increased by 10.21 and 11.85%, respectively, which suggested that the selection area of non-landslide samples has more robustness modeling performance and higher prediction ability in the entire area of Anhui Province (i.e., region-wide including flatter terrain areas).



**Fig. 9** Spatial distribution of landslide susceptibility in Anhui Province obtained by **a** the EareaLR model and **b** the MareaLR model

The EareaLR model and the MareaLR model were applied to the landslide susceptibility mapping in Anhui Province, and the landslide susceptibility values were divided by using the quantile method corresponding to the 50%, 75%, 90%, and 95% quantiles. The landslide susceptibility levels were reclassified into five levels such as very low, low, moderate, high, and very high. The results are shown in Fig. 9.

In comparison between Fig. 9a and b, most of the areas with very high and high landslide susceptibility areas produced by the MareaLR model are distributed in high-altitude mountainous areas and scattered in flatter terrain areas in northern Anhui Province (Fig. 9b). In the landslide susceptibility map produced by the EareaLR model, landslide susceptibility areas with very high and high susceptibility levels were mainly distributed in the southern mountainous area and part of the southwestern mountainous area (Fig. 9a). Compared with the recorded distribution of historical landslides in Anhui Province (Fig. 1c) and the geomorphology characteristics of Anhui Province, the spatial distribution of landslide susceptibility produced by the EareaLR model was geomorphologically more plausible.

The statistical results of the distribution of landslide susceptibility levels for the EareaLR model and MareaLR model are shown in Table 3. About 80% of the historical landslides are distributed in very high, high, and moderate landslide susceptibility areas produced by the EareaLR model, which only accounts for 25% of the total area of Anhui Province. And the area with low and very low landslide susceptibility levels accounted for about 75% of Anhui Province, which only accounted for about 20% of historical landslides. The area with the moderate, high and very high landslide susceptibility produced by the MareaLR model also accounted for 25% of the total area of Anhui Province, but the historical landslides in this area only account for 74.66%, and the area with low and very



**Table 3** Distribution of historical landslides in various landslide susceptibility levels in Anhui Province

Landslide susceptibility	Area proportion (%)	Number of landslides (proportion %)		Landslide density (landslides/10000 m <sup>2</sup> )	
		EareaLR	MareaLR	EareaLR	MareaLR
Very low	50.0	140 (4.29%)	127 (3.89%)	0.201	0.182
Low	25.0	517 (15.82%)	701 (21.46%)	1.482	2.010
Moderate	15.0	1295 (39.64%)	1374 (42.06%)	6.189	6.567
High	5.0	603 (18.46%)	548 (16.77%)	8.645	7.857
Very high	5.0	712 (21.79%)	517 (15.82%)	10.208	7.412

low susceptibility levels also accounts for 75% of the area of Anhui Province, and there are about 25% of historical landslides.

Comparing the results of the recorded landslides distribution of each susceptibility level of the EareaLR model and the MareaLR model, it is found that the moderate, high and very high susceptibility areas of the landslide susceptibility model calibrated by the selection area of non-landslide samples in the entire area of Anhui Province can explain more historical landslides and there are fewer landslides in low and very low susceptibility areas when the area of each susceptibility zone is the same. This suggests that the landslide susceptibility model established by taking the entire area as the selected spatial extent of non-landslide samples has better prediction performance.

Various evaluation metrics reveal that the EareaLR model performs significantly better than the MareaLR model in terms of predictive performance. This is attributed to the fact that the non-landslide samples in the EareaLR model are selected on a region-wide extent, and these non-landslide samples have a large number of lithological categories (Unconsolidated Sediments) that differ from the landslide samples. At the same time, these non-landslide samples are mostly located in low-slope areas. These two together enable the lithology and slope predictors to better discriminate between landslide and non-landslide samples in the EareaLR model compared to the MareaLR model, resulting in higher prediction performance. This high predictive performance and the capability to discriminate between landslide and non-landslide samples in the EareaLR model is further presented in the appearance of the final landslide susceptibility map. It results in landslide susceptibility maps obtained from the EareaLR model that better coincide with historical landslide distributions and propensity to predict lower landslide susceptibility levels in Unconsolidated Sediments lithological categories and low-slope areas. These results suggest that when comparing the performance of different data-driven landslide susceptibility models, the extent of non-landslide sample selection also has a significant effect, and the performance of the models cannot be judged by only quantitative evaluation metrics.

There are no universal criteria for non-landslide sampling schemes in the statistical-based landslide susceptibility model (Dou et al. 2020). It has been demonstrated that different non-landslide sampling strategies have significant effects on landslide susceptibility modeling results, such as different non-landslide/landslide sample ratios (Shao et al. 2020), different sampling intensity (Shao et al. 2020), different sampling scales (Yi et al. 2020) and different sampling method (Erener et al. 2017). This study attempts to explore the influence of different spatial extent selection for non-landslide samples on the statistical-based landslide susceptibility model. The results showed that different spatial extent of the non-landslide samples had significant effects on the explanatory power of the predictors,

the predictive performance of the model, and the appearance of the final landslide susceptibility map. The entire area (i.e., region-wide including flatter terrain areas) for non-landslide sampling could yield better evaluation performance and more geomorphologically plausible distribution of landslide susceptibility. The results of this study adequately validate, at a larger spatial scale and based on more abundant evaluation metrics, the argument proposed in the literature that the introduction of “trivial terrain” into data-driven landslide susceptibility models can enhance the predictive performance of geomorphologically meaningless models (Steger and Glade 2017).

## 5 Conclusions

This study investigates the potential influence of the selection of the spatial extent for non-landslide samples on the landslide susceptibility models. The entire area (i.e., region-wide including flatter terrain areas) and the mountainous area of Anhui Province were designed for different non-landslides sampling areas. Then, the EareaLR model and MareaLR model representing different non-landslides sampling spatial extent were built based on the logistic regression model and the selected dominant landslide environmental factors. The performances of the two models for landslide susceptibility modeling and mapping were compared and evaluated. The main conclusions are as follows:

The method of GeoDetector can effectively identify the dominant landslide environmental factors and quantify the relationship between each environmental factor and the spatial distribution of historical landslides in Anhui Province. However, determining the dominant factor to be introduced into the model solely by quantitative metrics such as importance obtained by GeoDetector may increase the likelihood of introducing biased variables or the trouble of introducing multiple redundant variables. It would be more appropriate to incorporate a heuristic approach to select the ultimately dominant factors for modeling.

Different spatial extent of the non-landslide samples had significant effects on the explanatory power of the predictors, the predictive performance of the model, and the appearance of the final landslide susceptibility map. The non-landslide samples in the EareaLR model are selected on a region-wide extent to enable the lithology and slope predictors to better discriminate between landslide and non-landslide samples in the EareaLR model, thus resulting in higher prediction performance and more geomorphologically plausible appearance of the final landslide susceptibility map. This suggests that when compared to the performance of different data-driven landslide susceptibility models, the performance of the models cannot be judged by quantitative evaluation metrics alone but also the spatial extent of non-landslide sample selection needs to be considered.

**Funding** This research was supported by the Second Tibetan Plateau Scientific Expedition and Research Program (Grant Nos. 2019QZKK0906, 2019QZKK0606), the Project from Anhui Climate Center “Risk assessment technique of meteorological disaster in Anhui Province”, the Startup Foundation for Introducing Talent of NUIST.

## Declarations

**Conflict of interest** Information provided in this manuscript does not have any conflict or competing interests.

**Consent for publication:** All authors consent to participate in this publication.

**Consent to participate** All authors consent to participate in this research.

## References

- Brenning A (2012) Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the R package *sperrorst*. In: &nbsp;2012 IEEE international geoscience and remote sensing symposium, pp 5372–5375. <https://doi.org/10.1109/IGARSS.2012.6352393>
- Brenning A, Schwinn M, Ruiz-Páez AP, Muenchow J (2015) Landslide susceptibility near highways is increased by 1 order of magnitude in the andes of southern Ecuador, Loja province. *Nat Hazards Earth Syst Sci* 15(1):45–57. <https://doi.org/10.5194/nhess-15-45-2015>
- Choi J, Oh HJ, Lee HJ, Lee C, Lee S (2012) Combining landslide susceptibility maps obtained from frequency ratio, logistic regression, and artificial neural network models using ASTER images and GIS. *Eng Geol* 124:12–23. <https://doi.org/10.1016/j.enggeo.2011.09.011>
- Chung CJF, Fabbri AG (2003) Validation of spatial prediction models for landslide hazard mapping. *Nat Hazards* 30(3):451–472. <https://doi.org/10.1023/B:NHAZ.0000007172.62651.2b>
- Devkota KC, Regmi AD, Pourghasemi HR, Yoshida K, Pradhan B, Ryu IC, Dhital MR, Althuwaynee OF (2013) Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling–Narayanghat road section in Nepal Himalaya. *Nat Hazards* 65(1):135–165. <https://doi.org/10.1007/s11069-012-0347-6>
- Dou J, Yunus AP, Merghadi A, Shirzadi A, Nguyen H, Hussain Y, Avtar R, Chen Y, Pham BT, Yamagishi H (2020) Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning. *Sci Total Environ* 720:137320. <https://doi.org/10.1016/j.scitotenv.2020.137320>
- Erener A, Sivas AA, Selcuk-Kestel AS, Düzgün HS (2017) Analysis of training sample selection strategies for regression-based quantitative landslide susceptibility mapping methods. *Comput Geosci* 104:62–74. <https://doi.org/10.1016/j.cageo.2017.03.022>
- Fratini P, Crosta G, Carrara A (2010) Techniques for evaluating the performance of landslide susceptibility models. *Eng Geol* 111(1–4):62–72. <https://doi.org/10.1016/j.enggeo.2009.12.004>
- Froude MJ, Petley DN (2018) Global fatal landslide occurrence from 2004 to 2016. *Nat Hazards Earth Syst Sci* 18(8):2161–2181. <https://doi.org/10.5194/nhess-18-2161-2018>
- Gariano SL, Guzzetti F (2016) Landslides in a changing climate. *Earth Sci Rev* 162:227–252. <https://doi.org/10.1016/j.earscirev.2016.08.011>
- Goetz JN, Brenning A, Petschko H, Leopold P (2015) Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput Geosci* 81:1–11. <https://doi.org/10.1016/j.cageo.2015.04.007>
- Guzzetti F, Carrara A, Cardinali M, Reichenbach P (1999) Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study. *Central Italy Geomorphology* 31(1–4):181–216
- Guzzetti F, Reichenbach P, Cardinali M, Galli M, Ardizzone F (2005) Probabilistic landslide hazard assessment at the basin scale. *Geomorphology* 72(1–4):272–299. <https://doi.org/10.1016/j.geomorph.2005.06.002>
- Guzzetti F, Reichenbach P, Ardizzone F, Cardinali M, Galli M (2006) Estimating the quality of landslide susceptibility models. *Geomorphology* 81(1–2):166–184. <https://doi.org/10.1016/j.geomorph.2006.04.007>
- Hartmann J, Moosdorf N (2012) Global lithological map database v1.0
- He H, Hu D, Sun Q, Zhu L, Liu Y (2019) A landslide susceptibility assessment method based on GIS technology and an AHP-weighted information content method: a case study of southern Anhui, China. *ISPRS Int J Geo Inf* 8(6):266. <https://doi.org/10.3390/ijgi8060266>
- Hussin HY, Zumpano V, Reichenbach P, Sterlacchini S, Micu M, van Westen C, Bălteanu D (2016) Different landslide sampling strategies in a grid-based bi-variate statistical susceptibility model. *Geomorphology* 253:508–523. <https://doi.org/10.1016/j.geomorph.2015.10.030>
- Iovine GG, Greco R, Gariano SL, Pellegrino AD, Terranova OG (2014) Shallow-landslide susceptibility in the Costa Viola mountain ridge (southern Calabria, Italy) with considerations on the role of causal factors. *Nat Hazards* 73(1):111–136. <https://doi.org/10.1007/s11069-014-1129-0>
- Jebur MN, Pradhan B, Tehrany MS, He H, Hu D, Sun Q, Zhu L, Liu Y (2014) 2019. A landslide susceptibility assessment method based on GIS technology and an AHP-weighted information content


- method: a case study of southern Anhui, China. *ISPRS Int J Geo Inf* 8(6): 266. <https://doi.org/10.3390/ijgi8060266>
- Kalantar B, Pradhan B, Naghibi SA, Motevalli A, Mansor S (2018) Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). *Geomat Nat Hazards Risk* 9(1):49–69. <https://doi.org/10.1080/19475705.2017.1407368>
- Kavzoglu T, Sahin EK, Colkesen I (2014) Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides* 11(3):425–439. <https://doi.org/10.1007/s10346-013-0391-7>
- Lin Q, Wang Y (2018) Spatial and temporal analysis of a fatal landslide inventory in China from 1950 to 2016. *Landslides* 15:2357–2372. <https://doi.org/10.1007/s10346-018-1037-6>
- Lin L, Lin Q, Wang Y (2017a) Landslide susceptibility mapping on a global scale using the method of logistic regression. *Natural Hazards and Earth System Sciences* 17(8):1411–1424. <https://doi.org/10.5194/nhess-17-1411-2017>
- Lin Q, Wang Y, Liu T, Zhu Y, Sui Q (2017b) The vulnerability of people to landslides: a case study on the relationship between the casualties and volume of landslides in China. *Int J Environ Res Public Health* 14(2):212. <https://doi.org/10.3390/ijerph14020212>
- Lin Q, Wang Y, Glade T, Zhang J, Zhang Y (2020) Assessing the spatiotemporal impact of climate change on event rainfall characteristics influencing landslide occurrences based on multiple GCM projections in China. *Clim Change* 162(2):761–779. <https://doi.org/10.1007/s10584-020-02750-1>
- Lin Q, Lima P, Steger S, Glade T, Jiang T, Zhang J, Liu T, Wang Y (2021) National-scale data-driven rainfall induced landslide susceptibility mapping for China by accounting for incomplete landslide data. *Geosci Front* 12(6):101248. <https://doi.org/10.1016/j.gsf.2021.101248>
- Linke S, Lehner B, Dallaire CO, Ariwi J, Grill G, Anand M, Beames P, Burchard-levine V, Maxwell S, Moidu H, Tan F, Thieme M (2019) Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Sci Data* 6(1):1–15. <https://doi.org/10.6084/m9.figshare.10059551>
- Ma M, Liu C, Zhao G, Xie H, Jia P, Wang D, Wang H, Hong Y (2019) Flash flood risk analysis based on machine learning techniques in the Yunnan Province, China. *Remote Sens* 11(2):170. <https://doi.org/10.3390/rs11020170>
- National Bureau of Statistics (2020) China statistical yearbook. China Statistics Press, Beijing
- Nefeslioglu HA, Gokceoglu C, Sonmez H (2008) An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps. *Eng Geol* 97(3–4):171–191. <https://doi.org/10.1016/j.enggeo.2008.01.004>
- Nhu VH, Hoang ND, Nguyen H, Ngo PTT, Bui TT, Hoa PV, Samui P, Bui DT (2020) Effectiveness assessment of Keras based deep learning with different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area. *Catena* 188:104458. <https://doi.org/10.1016/j.catena.2020.104458>
- Peng W, Hongwei D (2020) Study on flood hazard risk zoning based on GIS and logistic regression model. *Adv Earth Sci* 35(10):1064–1072. <https://doi.org/10.11867/j.issn.1001-8166.2020.082>
- Petley D (2012) Global patterns of loss of life from landslides. *Geology* 40(10):927–930. <https://doi.org/10.1130/G33217.1>
- Pourghasemi HR, Kornejady A, Kerle N, Shabani F (2020) Investigating the effects of different landslide positioning techniques, landslide partitioning approaches, and presence-absence balances on landslide susceptibility mapping. *Catena* 187:104364. <https://doi.org/10.1016/j.catena.2019.104364>
- R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Regmi NR, Giardino JR, McDonald EV, Vitek JD (2014) A comparison of logistic regression-based models of susceptibility to landslides in western Colorado. *USA Landslides* 11(2):247–262. <https://doi.org/10.1007/s10346-012-0380-2>
- Reichenbach P, Rossi M, Malamud DB, Mihir M, Guzzetti F (2018) A review of statistically-based landslide susceptibility models. *Earth Sci Rev* 180:60–91. <https://doi.org/10.1016/j.earscirev.2018.03.001>
- Rong G, Li K, Han L, Alu S, Zhang J, Zhang Y (2020) Hazard mapping of the rainfall–landslides disaster Chain based on GeoDetector and Bayesian network models in Shuicheng County, China. *Water* 12(9):2572. <https://doi.org/10.3390/w12092572>
- Sameen MI, Pradhan B, Bui DT, Alamri AM (2020) Systematic sample subdividing strategy for training landslide susceptibility models. *Catena* 187:104358. <https://doi.org/10.1016/j.catena.2019.104358>

- Shao X, Ma S, Xu C, Zhou Q (2020) Effects of sampling intensity and non-slide/slide sample ratio on the occurrence probability of coseismic landslides. *Geomorphology* 363:107222. <https://doi.org/10.1016/j.geomorph.2020.107222>
- Song Y, Wang J, Ge Y, Xu C (2020) An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: cases with different types of spatial data. *GIScience Remote Sens* 57(5):593–610. <https://doi.org/10.1080/15481603.2020.1760434>
- Steger S, Brenning A, Bell R, Petschko H, Glade T (2016) Exploring discrepancies between quantitative validation results and the geomorphic plausibility of statistical landslide susceptibility maps. *Geomorphology* 262:8–23. <https://doi.org/10.1016/j.geomorph.2016.03.015>
- Steger S, Glade T (2017) The challenge of “trivial areas” in statistical landslide susceptibility modelling. Workshop on World Landslide Forum. pp 803–808. [https://doi.org/10.1007/978-3-319-53498-5\\_92](https://doi.org/10.1007/978-3-319-53498-5_92)
- Steger S, Mair V, Kofler C, Pittore M, Zebisch M, Schneiderbauer S (2021) Correlation does not imply geomorphic causation in data-driven landslide susceptibility modelling—Benefits of exploring landslide data collection effects. *Sci Total Environ* 776:145935. <https://doi.org/10.1016/j.scitotenv.2021.145935>
- Sun D, Shi S, Wen H, Xu J, Zhou X, Wu J (2021) A hybrid optimization method of factor screening predicated on GeoDetector and Random Forest for Landslide Susceptibility Mapping. *Geomorphology* 379:107623. <https://doi.org/10.1016/j.geomorph.2021.107623>
- USGS (2015) Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global. US Geological Survey. <https://doi.org/10.5066/F7P7R7TFT>
- Varnes DJ (1984) Landslide hazard zonation: a review of principles and practice. United Nations Educational, Scientific and Cultural Organization, Paris
- Wang JF, Xu CD (2017) Geodetector: principle and prospective. *Acta Geogr Sin* 72(1):116–134. <https://doi.org/10.11821/dlxb201701010>
- Wang JF, Li XH, Christakos G, Liao YL, Zhang T, Gu X, Zheng XY (2010) Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *Int J Geogr Inf Sci* 24(1):107–127. <https://doi.org/10.1080/13658810802443457>
- Wang Y, Song C, Lin Q, Li J (2016) Occurrence probability assessment of earthquake-triggered landslides with Newmark displacement values and logistic regression: the Wenchuan earthquake, China. *Geomorphology* 258:108–119. <https://doi.org/10.1016/j.geomorph.2016.01.004>
- Wang Y, Lin Q, Shi P (2018) Spatial pattern and influencing factors of landslide casualty events. *J Geog Sci* 28(3):259–274. <https://doi.org/10.6038/cjg20130406>
- Wu J, Gao XJ (2013) A gridded daily observation dataset over China region and comparison with the other datasets. *Chin J Geophys* 56(4):1102–1111. <https://doi.org/10.6038/cjg20130406>
- Xiao T, Segoni S, Chen L, Yin K, Casagli N (2020) A step beyond landslide susceptibility maps: a simple method to investigate and explain the different outcomes obtained by different approaches. *Landslides* 17(3):627–640. <https://doi.org/10.1007/s10346-019-01299-0>
- Xie W, Nie W, Saffari P, Robledo LF, Descote PY, Jian W (2021) Landslide hazard assessment based on Bayesian optimization—support vector machine in Nanping City. *China Nat Hazards*. <https://doi.org/10.1007/s11069-021-04862-y>
- Xie W, Li X, Jian W, Yang Y, Liu H, Robledo LF, Nie W (2021) A novel hybrid method for landslide susceptibility mapping-based geodetector and machine learning cluster: a case of Xiaojin County, China. *ISPRS Int J Geo Inf* 10(2):93. <https://doi.org/10.3390/ijgi10020093>
- Xiong J, Li J, Cheng W, Zhou C, Guo L, Zhang X, Wang N, Li W (2019) Spatial-temporal distribution and the influencing factors of mountain flood disaster in southwest China. *Acta Geogr Sin* 74:1374–1391. <https://doi.org/10.11821/dlxb201907008>
- Yang J, Song C, Yang Y, Xu C, Guo F, Xie L (2019) New method for landslide susceptibility mapping supported by spatial logistic regression and GeoDetector: a case study of Duwen Highway Basin, Sichuan Province, China. *Geomorphology* 324:62–71. <https://doi.org/10.1016/j.geomorph.2018.09.019>
- Yi Y, Zhang Z, Zhang W, Jia H, Zhang J (2020) Landslide susceptibility mapping using multiscale sampling strategy and convolutional neural network: a case study in Jiuzhaigou region. *Catena* 195:104851. <https://doi.org/10.1016/j.catena.2020.104851>
- Zhang Y, Wang Y, Chen Y, Xu Y, Zhang G, Lin Q, Luo R (2021) Projection of changes in flash flood occurrence under climate change at tourist attractions. *J Hydrol* 595:126039. <https://doi.org/10.1016/j.jhydrol.2021.126039>
- Zhao G, Pang B, Xu Z, Yue J, Tu T (2018) Mapping flood susceptibility in mountainous areas on a national scale in China. *Sci Total Environ* 615:1133–1142. <https://doi.org/10.1016/j.scitotenv.2017.10.037>
- Zhou X, Wen H, Zhang Y, Xu J, Zhang W (2021) Landslide susceptibility mapping using hybrid random forest with GeoDetector and RFE for factor optimization. *Geosci Front* 12(5):101211. <https://doi.org/10.1016/j.gsf.2021.101211>

Zhu AX, Miao Y, Liu J, Bai S, Zeng C, Ma T, Hong H (2019) A similarity-based approach to sampling absence data for landslide susceptibility mapping using data-driven methods. *Catena* 183:104188. <https://doi.org/10.1016/j.catena.2019.104188>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Chuhan Wang<sup>1,2</sup> · Qigen Lin<sup>1</sup>  · Leibin Wang<sup>3</sup> · Tong Jiang<sup>1</sup> · Buda Su<sup>1</sup> · Yanjun Wang<sup>1</sup> · Sanjit Kumar Mondal<sup>1</sup> · Jinlong Huang<sup>1</sup> · Ying Wang<sup>4</sup>

<sup>1</sup> Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters / Institute for Disaster Risk Management / School of Geographical Sciences, Nanjing University of Information Science & Technology, Nanjing 210044, China

<sup>2</sup> School of Geography, South China Normal University, Guangzhou 510631, China

<sup>3</sup> School of Geographical Sciences, Hebei Normal University, Shijiazhuang 050024, China

<sup>4</sup> Key Laboratory of Environmental Change and Natural Disaster of Ministry of Education / Academy of Disaster Reduction and Emergency Management, Ministry of Emergency Management and Ministry of Education, Beijing Normal University, Beijing 100875, China