water

MDPI

Article

Flood Risk Evaluation in the Middle Reaches of the Yangtze River Based on Eigenvector Spatial Filtering Poisson Regression

Tao Fang, Yumin Chen *, Huangyuan Tan, Jiping Cao, Jiaxin Liao and Liheng Huang

School of Resource and Environment Science, Wuhan University, Wuhan 430079, China; fountaintop@whu.edu.cn (T.F.); tanhuangyuan@whu.edu.cn (H.T.); caojiping@whu.edu.cn (J.C.); 2018282050169@whu.edu.cn (J.L.); 2014301110077@whu.edu.cn (L.H.)

* Correspondence: ymchen@whu.edu.cn; Tel.: +86-27-687-783-86

Received: 20 August 2019; Accepted: 17 September 2019; Published: 21 September 2019

Abstract: A Poisson regression based on eigenvector spatial filtering (ESF) is proposed to evaluate the flood risk in the middle reaches of the Yangtze River in China. Regression analysis is employed to model the relationship between the frequency of flood alarming events observed by hydrological stations and hazard-causing factors from 2005 to 2012. Eight factors, including elevation (ELE), slope (SLO), elevation standard deviation (ESD), river density (DEN), distance to mainstream (DIST), NDVI, annual mean rainfall (RAIN), mean annual maximum of three-day accumulated precipitation (ACC) and frequency of extreme rainfall (EXE) are selected and integrated into a GIS environment for the identification of flood-prone basins. ESF-based Poisson regression (ESFPS) can filter out the spatial autocorrelation. The methodology includes construction of a spatial weight matrix, testing of spatial autocorrelation, decomposition of eigenvectors, stepwise selection of eigenvectors and calculation of regression coefficients. Compared with the pseudo R squared obtained by PS (0.56), ESFPS exhibits better fitness with a value of 0.78, which increases by approximately 39.3%. ESFPS identifies six significant factors including ELE, DEN, EXE, DIST, ACC and NDVI, in which ACC and NDVI are the first two main factors. The method can provide decision support for flood risk relief and hydrologic station planning.

Keywords: spatial autocorrelation; Poisson regression; eigenvector spatial filtering method; flood risk evaluation

1. Introduction

Floods are the most common and destructive natural disaster around the world. The dam-breaking and waterlogging events caused by floods have caused great losses of human lives and property throughout history [1]. When the water level measured in large parts of a river becomes too high and exceeds its maximum capacity, flooding could occur. A recent study ranked floods as the greatest threat to 616 cities around the world over earthquakes and storms [2]. There are many types of floods such as river floods, flash floods, urban floods, sewer flooding and coastal flooding in China. Affected by the monsoon climate and geographical conditions, China is seriously impacted by flood disasters, especially flash floods and river floods [3]. Floods in the region has a wide range of impacts and can be sudden and strong, occur for long periods, frequent or seasonal. These obvious characteristics make floods one of the most important factors restricting China's economic and social development. Hence, management and mitigation before and after a flood disaster are necessary and significant. The evaluation of risks in flood-prone areas and the construction of hydrological stations are effective means of prevention and monitoring before floods. Areas susceptible to flood must be detected in advance so that urbanization and industrialization can be prevented in these areas. In addition, some

watersheds prone to flooding should be monitored with hydrological stations in the case of dangerous water levels.

Floods in different regions at different times are often caused directly by different factors [4] while climate and human activities are common factors [5]. Against the background of climate change in recent years, extreme rainy weather brings concentrated, intense and consecutive rainfall that quickly transforms into sediment and runoff [6,7]. Many cities were recently reported to encounter the highest recorded rainfall in history because massive urbanization and unwise city planning have replaced many forests and rivers with industrialized areas [8]. Vegetation is able to contain soil and some rainfall [9]. Conversely, large scale urban expansion reduces city permeability, However, the drainage functions developed in cities cannot reach the level required to swiftly discharge runoff in time [10].

Two categories of flood models have been researched and developed which is either physically based or statistically based [11]. The Storm Water Management Model, a good representative of many hydrological models, has been widely employed to simulate the flow and flood processes [12]. Apart from detailed data and complicated calculations [13], these models can only simulate the extent and velocity of flood inundation over small-scale areas through one dimensional or two-dimensional hydraulic models and require much time and hardware to compute [14]. Both of two categories have been conducted in the many previous study based on the Yangtze River [15].

Regarding statistical models that only consider correlation and cause-and-effect relationships, qualitative or semiquantitative methods such as multicriteria analytic hierarchy process (AHP) are frequently used to model the risks of disasters such as floods or landslides [16,17]. A previous study combined the AHP with a Bayesian Network to research the flooding risk in Guangzhou [18]. These models are used widely because they have good calculation performance with simpler statistical theory. Other studies also use machine learning methods such as decision tree, SVM (support vector machine) and artificial neural network (ANN) to predict the spatial distribution of flood risk over large-scale regions with accuracies higher than 80% [19,20]. Expert weighting of the AHP is relatively subjective and the different factors in different regions at different times that result in disasters tend to show distinct relative importance, hence, experiences about the weights of factors in certain areas cannot be transfer directly to others [21]. Machine learning algorithms such as ensembled decision tree model and ANN is more complicated and just like a black box so that its network structure and weight parameters are not interpretable [22]. Moreover, these models ignore the spatial autocorrelation that exists flood disasters. Considering spatial effects in these models will probably further increase their performance.

Multivariate regression usually involves several spatial factors and non-spatial attribute data. Geographic Information System (GIS) is a powerful tool to collect, process, manage and analyze spatially referenced layers in table, vector or raster format [23], and quantization calculation can mostly be easily integrated into it. Moreover, the visualization of GIS can make risk assessment accessible and easy to understand. Risk analysis of natural disasters integrated with GIS has been applied widely in the recent decades [24].

Poisson regression has been commonly applied to estimate disasters or disease risks and perform factor analysis [25,26]. The method is very suitable for modeling the data on the accumulated count of events that take place with low probability, such as diseases rate and death rates, especially when the observations approximately follow the Poisson distribution. Moreover, simple models such as generalized linear regression including Poisson regression are more suitable for small dataset and have more interpretability than complex machine learning algorithms. Aforementioned statistical models consider a weighted score or a probability as the metric of flood risk while the study considers the count of flood alarming events in history as the metric of flood risk because the frequency is also representative of occurrence rate when different hydrological stations have the same number of observations. The new data from hydrological stations currently available, allow a much better spatial pattern analysis. Moreover, some count observations often exhibit spatial autocorrelation which exists in most geospatial processes [27], according to the First Law of Geography [28]. Failure to consider

spatial effects of flood alarming events based on the nature of the basin's connectivity in the regression model will lead to model uncertainty, thus, spatial autocorrelation must be included in the regression model [29,30]. Whether spatial autocorrelation exists in the observations of geographical units can be tested using Moran's I.

Eigenvector spatial filtering method is proposed to account for spatial effects [31]. This method selects a subset of eigenvectors from the spatial weight matrix that represent the spatial distribution pattern and then adds them to the ordinary Poisson regression model as independent proxy variables [32]. The linear combination of these eigenvectors filters the spatial autocorrelation out of the observations, which enables the observations in different geographical units to be independent [33,34].

In this paper, we propose the eigenvector spatial filtering Poisson regression (ESFPS) model for the estimation of flood risk, using the frequency of flood alarming events observed by hydrological stations in the middle reaches of the Yangtze River in China. Independent hazard-causing variables include elevation (ELE), slope (SLO), elevation standard deviation (ESD), river density (DEN), distance to mainstream (DIST), normalized difference vegetation index (NDVI), annual mean rainfall, mean annual maximum of three-day accumulated precipitation (ACC) and frequency of extreme rainfall (EXE), because these factors are considered to aggravate or trigger flooding or are closely associated with flood risk in the study area. The model results will be compared with the results of Poisson regression and negative binomial regression, and the best model will be used to predict the flood risk throughout the basin and analyze the spatial distribution pattern of flood risk.

2. Study Area and Data

2.1. Study Area

The focus of this study is the middle reaches of the Yangtze River spanning 108°24′–117°26′ E and 24°33′–33°14′ N and present a typical subtropical monsoon climate. The central drainage basin of the Yangtze River mainly covers three provinces—Hubei, Hunan and Jiangxi—in the central part of China, as shown in Figure 1. The weather is relatively humid with four clear seasons, and rainfall is especially abundant and is concentrated from April to August. As shown in Figure 2, according to the rainfall records from six meteorological stations in the three provinces, most of the monthly mean precipitation in the different provinces exceeds 150 mm during this period. Moreover, there are large amounts of variegated rivers and lakes, of which Dongting Lake in Hunan and Hubei Provinces and Poyang Lake in Jiangxi Province are China's largest and second-largest freshwater lake, respectively. In addition, many hydrologic stations and dams have been constructed along the Yangtze River and its main tributaries including the Han River, the Xiang River, the Yuan River and the Gan River in the study region. As a strategically important economic area, the region plays an extremely significant role in flood monitoring and forecasting. The varied topography, subtropical monsoon climate and a complex water system make this area among the highest flood prone areas especially where several destructive dam-break floods and many waterlogging events have occurred in history. The main flooding types in the area are river flooding, flash flooding and urban waterlogging caused by the former two types.

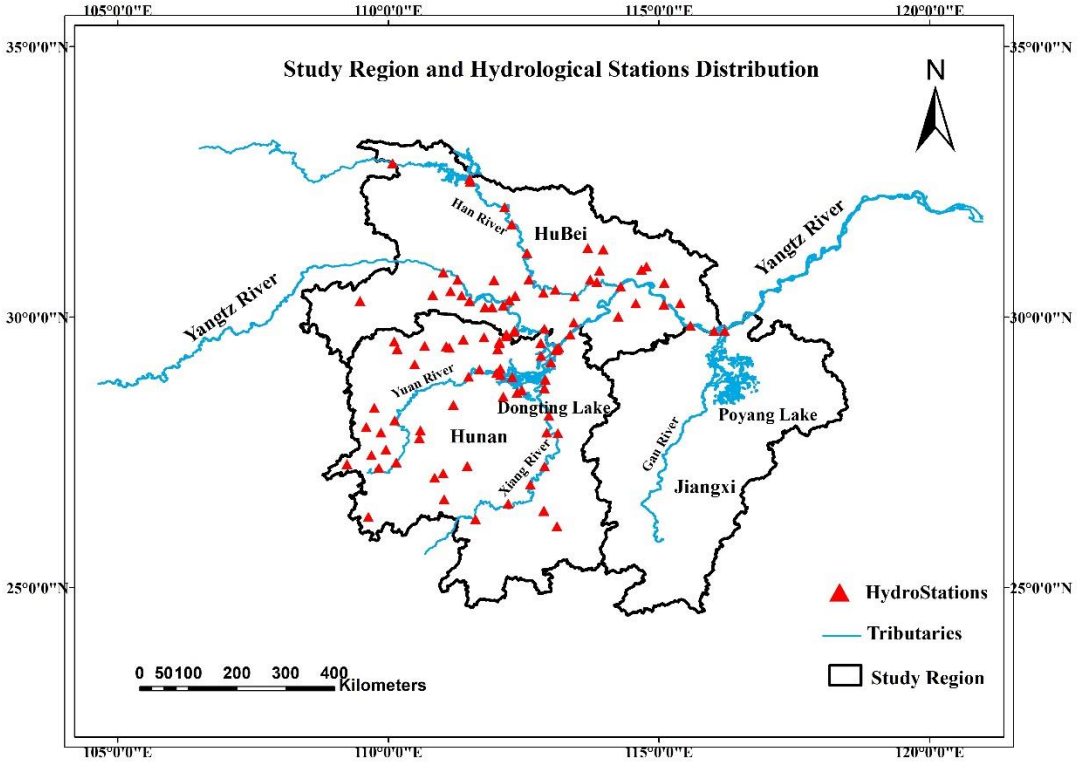

Figure 1. Study region and hydrological stations.

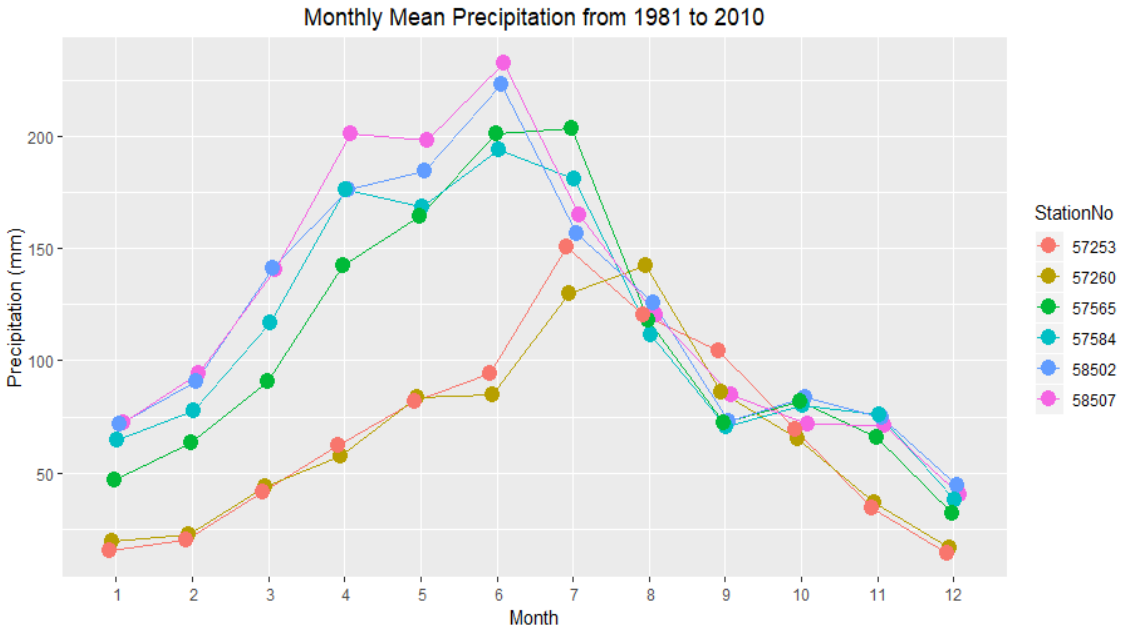

Figure 2. Monthly mean precipitation at six meteorological stations.

2.2. Data Source

In China, the major hindrance to the study of flood risk is the unavailability and scarcity of hydrological data and potential flood-related factors in digital GIS format. Geographically referenced data about hydrological stations were obtained from a list from Baidu Wenku, which contains the names, coordinates and other detailed information of all hydrologic stations around China. As Table 1 shows, daily hydrological observation data at 8:00 a.m. were acquired by means of web spider from

the Hydrological Information Inquiry website built by the Hydrology Bureau from Hunan and Hubei Province. Monthly mean rainfall from 1981 to 2010 and different types of meteorological statistical data can be accessed in the National Meteorological Information Center. Topological data using SRTM DEM with 30 m resolution and MODIS NDVI raster data with 500 m resolution can be acquired from Geographical Space Cloud (http://www.gscloud.cn/). Raster data of precipitation interpolated by meteorological observations in vector format seems invalid and inaccurate due to the sparsity of meteorological stations in the study region. Hence, a nearly real-time dataset named TRMM_3B42RT short for 3-h $0.25^{\circ} \times 0.25^{\circ}$ merged TRMM and other satellites estimates in raster format is applied and processed to calculate statistical factors concerning precipitation during the period from 2005 to 2012 (https://precip.gsfc.nasa.gov/index.html). The vector data of the river network included in the Chinese National Basic Geographic Information System data are open to the public on the website of the National Geomatics Centre of China.

Table 1. Data sources of the study.

Data Product Name	Resolution	Data Source
GDEMDEM 30 M	30 m	http://www.gscloud.cn/
MODIS NDVI	500 m	http://www.gscloud.cn/
Hydrological Observation Data	daily	http://61.187.56.156/wap/index_sq.asp
TRMM_3B42RT	0.25 degree	https://precip.gsfc.nasa.gov/index.html
River Network	vector	http://ngcc.sbsm.gov.cn/

2.3. Derivation of Hazard Factors

The flood risk at a hydrological station is represented by the historical frequency of flood alarm events which occur when dangerous water levels higher than the historical warning level are observed. Hazard factors that affect the risk of flooding in the middle reaches of the Yangtze River mainly include topological factors, hydrological factors and trigger factors. Topological factors such as elevation, slope and elevation standard deviation come from the DEM. Hydrological factors include river density and distance to mainstream. Vegetation is also an important factor. Trigger factors are mainly derived from rainfall data because flash floods and river floods caused by concentrated heavy rainfall are dominant in the study area.

2.3.1. Frequency of Flood Alarming Events

In hydrological observations, daily reports usually contain river names, hydrological observatory names, observation time, water level, flow, rise and fall of floodwater, warning level, highest water level in history and location information. Obviously, water level that is equal or even higher than warning level at a certain time indicates a flood alarming event, which indicates an excessively high level of water and requires the attention of people to prevent flooding from occurring. In the study area, approximately more than one hundred hydrological stations have been built and functioned normally to monitor water regimens. Every province in the middle reaches of the Yangtze River has set up an open website to display and query hydrological observation data in real time. The study employs the web crawler technology to gather hydrological observation report data at 8:00 a.m. every day from 2005 to 2012. After data processing and cleaning in consideration of the validity of every observation record and functionality of every hydrological station, 104 hydrological stations and their daily report data are included for research and analysis, totaling 303,80 records during the study period. Each observation record at a certain station is considered a flood alarming event if the water level is equal to or greater than the warning level. Then, we count the frequency of all flood alarming events totaling 1647 observed by these hydrological stations. From the histogram of flood alarming events in Figure 3, most of hydrological stations have low occurrence rate of flood alarming events close to 0 and large parts of hydrological stations observed 0 times of flood alarming events in the study period. Few

stations have high frequencies. In most cases, the larger the frequency is and the lower the number of hydrological stations is.

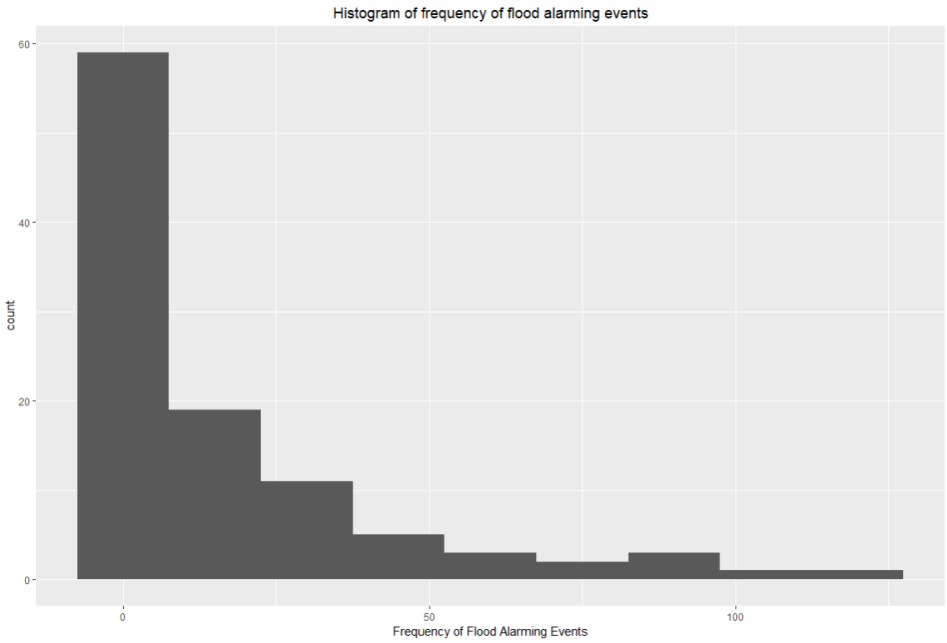

Figure 3. Histogram of the frequency of flood alarming events.

2.3.2. Topological Factors

Elevation is the most fundamental feature indicating the characteristics of the underlying surface. It is commonly acknowledged that water tends to accumulate in lower areas that are difficult to flow out from and that a hydrological station built in a relatively low area is prone to frequent flood alarming events.

Moreover, almost all of the previous studies that focused on the evaluation of flood risk considered elevation to be an important factor [17,21,35]. As is shown in Figure 4, the altitude of the study area ranges from −194 m to 5495 m and fluctuates largely. Slope can be defined as the steepness or gradient within a series of adjacent units of terrain, slope is usually measured as an angle in degrees or as a percentage and indicates the variation in elevation. In general, river currents flow faster over steep areas, reducing the hazard of flooding, while water easily accumulates on flat terrains, leading to a high and dangerous water level. Therefore, the slope of the area where hydrological stations are built must be taken into account [36]. The distribution of slope in the study area can be seen in Figure 5. Elevation standard deviation is calculated by some adjacent cells such as the neighboring cells in a 3×3 window, which represents a locally limited area. In reality, a relatively spacious areas are still susceptible to flooding because it has some local subareas with greater slope. Instead, another measurement of the elevation variation within larger ranges that needs to be included is the elevation standard deviation [18,21]. Corresponding to the resolution of the DEM and our study scale, this study uses a 5×5 neighboring window that represents an area of 150 m^2.

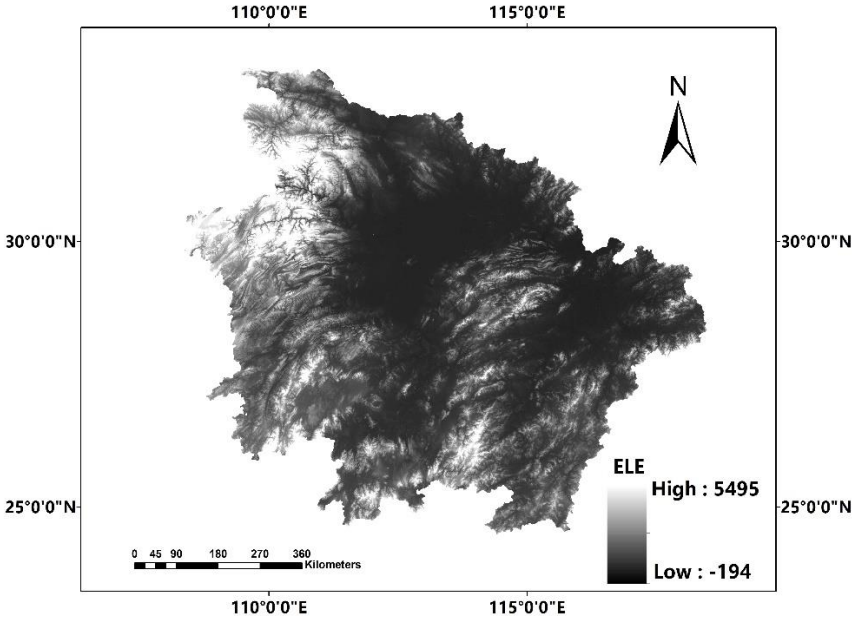

Figure 4. Elevation.

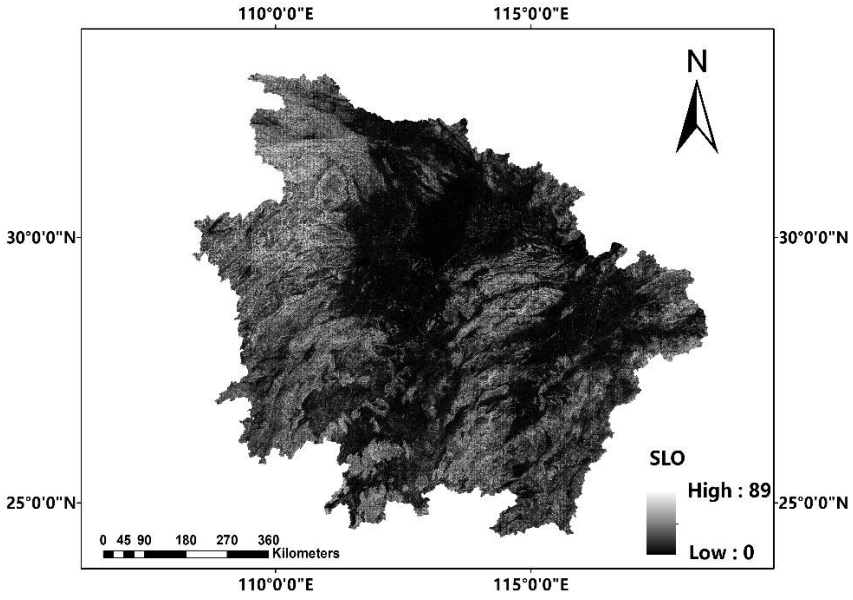

Figure 5. Slope.

The calculation method for elevation standard deviation (ESD) is shown in Equation (1):

$$\mathrm{ESD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\left(E_i - \overline{E}\right)^2\right.} \quad (1)$$

In the formula, E_i and $\overline{E}$ represent the cell value and mean within the window respectively and n is the number of cells in the window. Elevation standard deviation can be generated by focal statistics on DEM as seen in Figure 6.

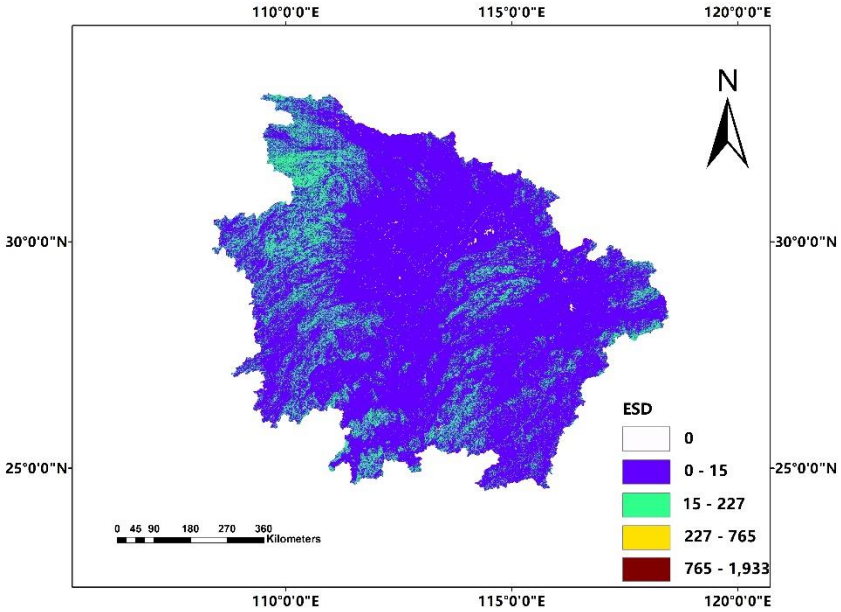

Figure 6. Elevation standard deviation.

2.3.3. Hydrological Factors

The density of a river system refers to the ratio between the summed length of the mainstream and tributaries in the basin and the area of the basin which represents the river's length within a unit of area. Climatology also reveals that the shapes of rivers have their own characteristics that allow for the discharge of surface runoff in time. River density is an important factor in controlling the occurrences of floods and river flow because these areas with high densities of rivers are prone to flooding [18]. Raster data of the river density are obtained by using the line density analysis to convert the vector data of the river network in the study region to raster data in Figure 7.

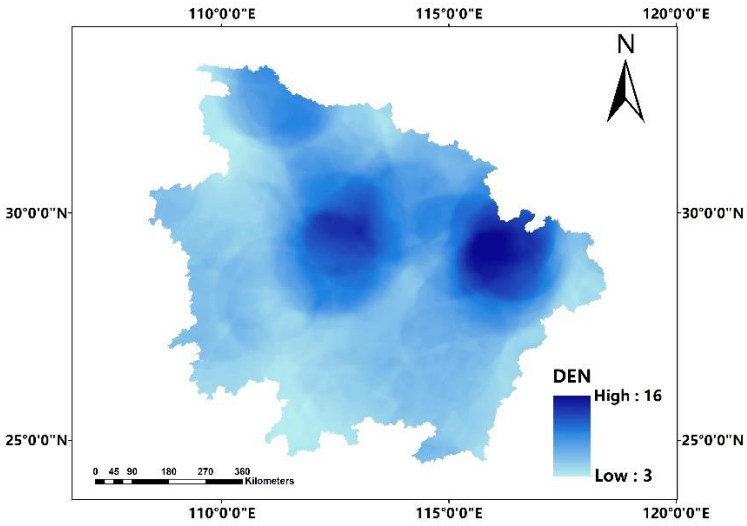

Figure 7. River Density.

Drainage proximity is measured by the distance to the mainstream or tributaries. The occurrences of flood alarming events and even flooding are related to the distribution of the drainage system [18]. A raster layer of the distance to mainstream is generated by performing Euclidean distance analysis on the mainstream network using vector data in ArcGIS10.2.1 as shown in Figure 8.

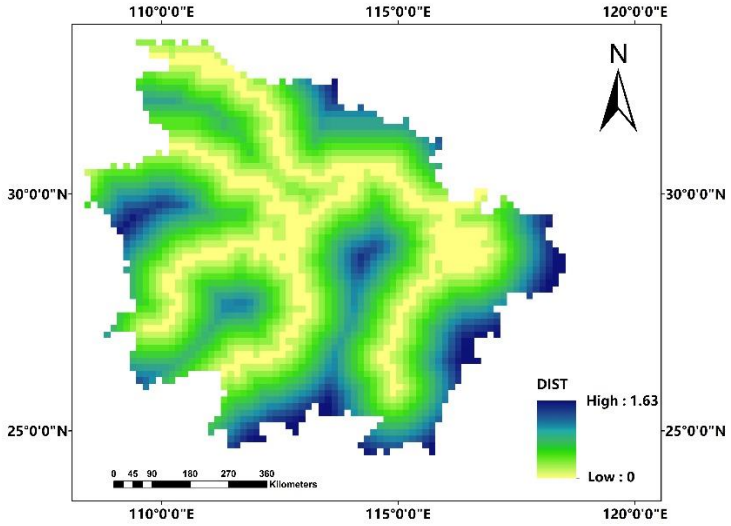

Figure 8. Extreme rainfall frequency.

2.3.4. NDVI

The effects of forest vegetation change on runoff and flooding are very significant: As forest coverage increases, the ability of river basins to intercept rainfall runoff is significantly enhanced, and the amount of rainfall converted to runoff is reduced. Meanwhile, flood peak flow decreases. Changes to vegetation cover can change soil properties and reduce soil erosion. The vegetation and the corresponding soil water absorption reduce the amount of flooding [37]. Vegetation blocks and intercepts floods and protects slopes. Therefore, NDVI, which is widely used in vegetation research of remote sensing is the best indicator factor of plant growth status and spatial distribution of vegetation density as shown in Figure 9.

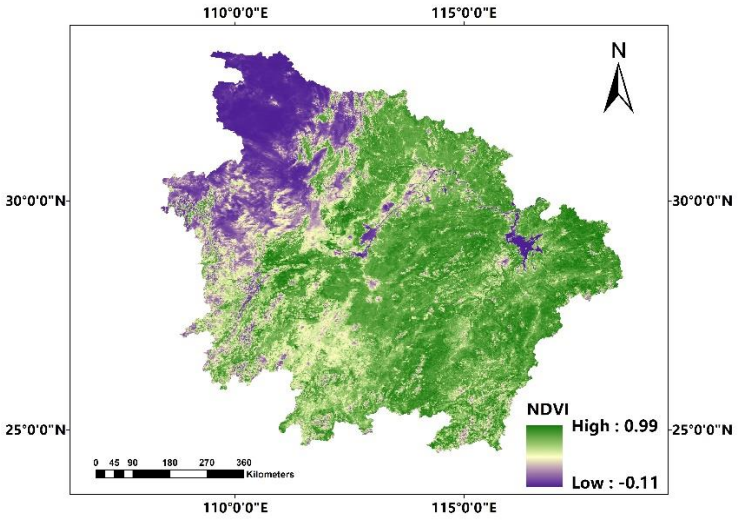

Figure 9. NDVI.

2.3.5. Trigger Factors

Statistical factors related to precipitation are derived from the processing and calculation of raster data, in which the value of every cell signifies a 3-h rainfall rate(mm/h) based on map algebra. This type of rainfall dataset has 0.25° × 0.25° resolution, which geographically corresponds to 25 km × 25 km area that is similar to a county. Many state-of-the-art studies apply real-time data for more accurate

simulation and analysis [38]. Considering its fine temporal resolution and better accuracy than rainfall raster data interpolated by rainfall observations of weather stations, the dataset is an ideal source for researchers to calculate trigger factors of flood. First, we accumulated six raster layers of the rainfall rate to represent daily precipitation. Second, every three consecutive raster layers were summed by using map algebra, and eventually, the raster of annual mean rainfall is produced. The detailed steps for processing the raster data are shown in Figure 10:

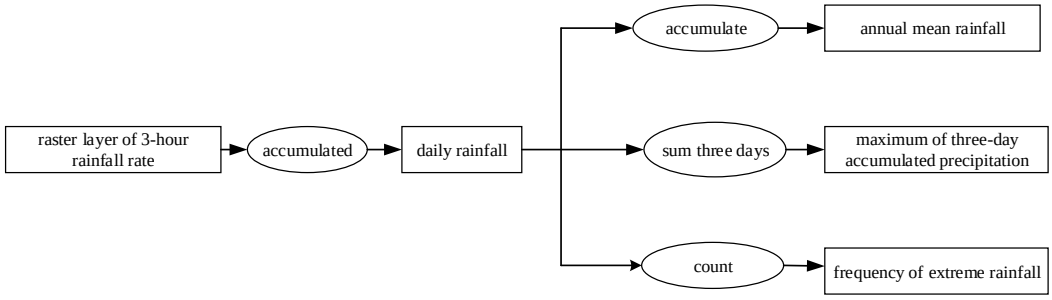

Figure 10. Flow chart of rainfall raster processing.

The raster data of the 3-h rainfall rate is accumulated ArcGIS10.2, and then, the derived raster data of daily rainfall are accumulated or counted further, and ArcPy is fit for efficient and fast batch processing.

Although annual mean rainfall is not directly related to flood alarming events and floods, it represents a meteorological pattern of different regions in the long run. Some previous studies have included similar factors that represent rainfall intensity such as annual mean rainfall or the Modified Fournier Index (MFI) [22]. The spatial distribution of annual mean rainfall is as shown in Figure 11.

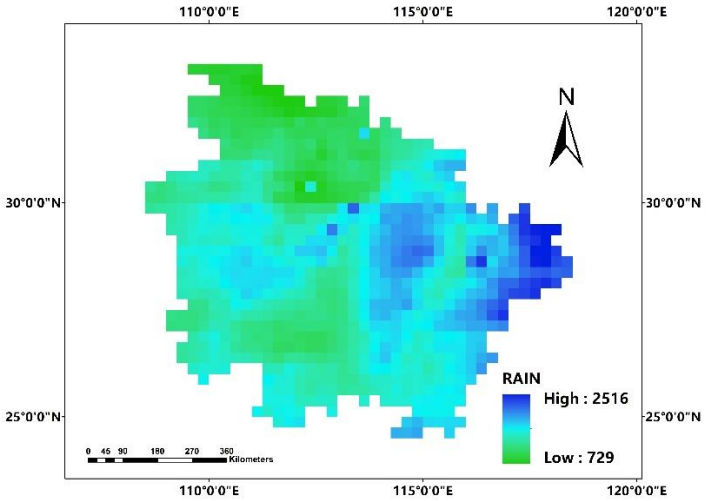

Figure 11. Annual mean rainfall.

It is well known for everyone that flood alarming events and flooding are often caused by continuous and intense rainfall in short periods. According to the resources of recorded historic flood, the maximum of three-day consecutive precipitation plays the most essential role in occurrences of dangerous water levels and floods. Therefore, the mean of the annual maximum of three-day accumulated precipitation is necessary to consider the occurrence of flood alarming events from a short-term perspective. As is seen in Figure 12, the factor ranges from 84 mm to 249 mm.

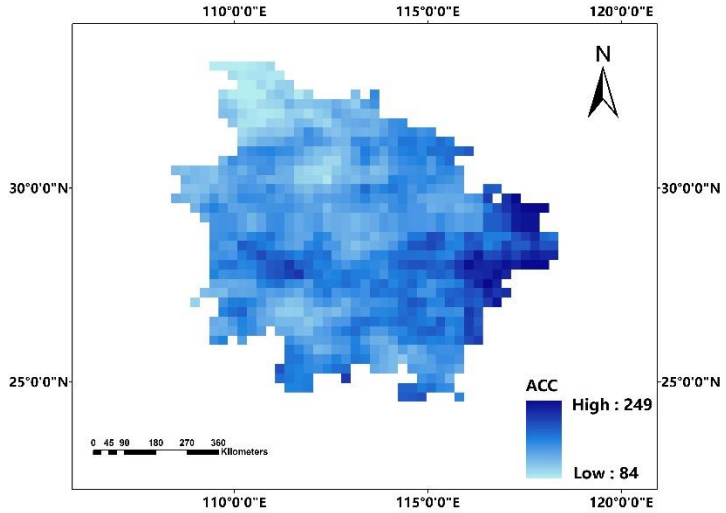

Figure 12. Mean annual maximum of three-day accumulated precipitation.

In meteorology, precipitation that is equal to or greater than 50 mm in 24 h is defined as torrential rain according to the official national standard released by China Meteorological Administration. The rise of river levels is more likely after extreme rainfall. Theoretically, more frequent extreme rainfall means a higher likelihood of flood alarming events and even flooding at worst. The count of heavy precipitation was considered by some previous studies that tried to evaluate flood risk [18]. The frequency of extreme rainfall in raster format is shown in Figure 13.

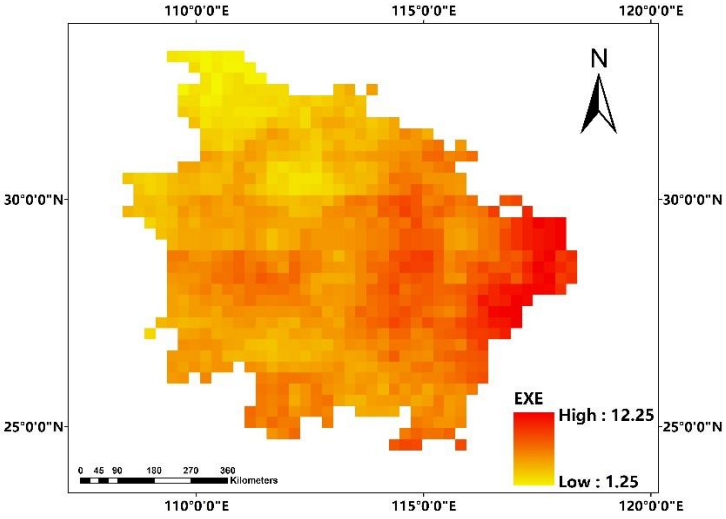

Figure 13. Frequency of extreme rainfall.

3. Methodology

The methodology of this study includes four steps: factor selection, eigenvector spatial filtering Poisson regression (ESFPS), model assessment and flood risk mapping as is shown in Figure 14. The second step is also divided into four parts: construction of a spatial weight matrix, testing of spatial autocorrelation, calculation of eigenvectors, normalization of independent variables, eigenvector selection and calculation of coefficients. The performance of ESFPS will be compared with that of the Poisson Regression (PS) and the negative binomial regression (NB). The model metrices for assessment includes fitness accuracy and generalization error. Afterwards, the best regression model is applied to

compare their importance by weights of different factors that pass the significance test in a regression. Finally, this model will be used to predict the flood risk in the middle reaches of the Yangtze River.

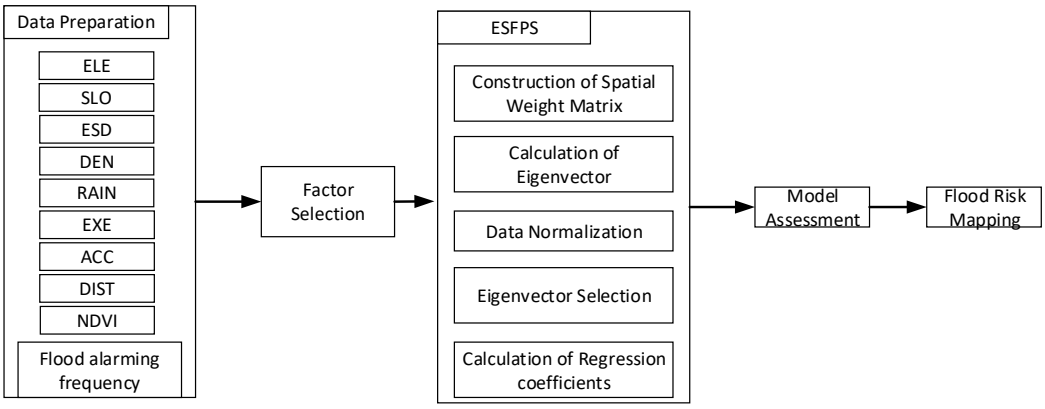

Figure 14. Flowchart of the study.

3.1. Factors Selection

Multicollinearity problems of factors are often encountered in regression problems, hence, a multicollinearity test must be performed before the regression [39]. Variance inflation factor (VIF) and tolerance are suggested as indicators of multicollinearity. To calculate the collinearity of these independent variables, every independent variable is used as the dependent variable and the remaining independent variables are used as the independent variable in the same regression equation. Tolerance and variance inflation factors can be calculated by Equations (2) and (3):

$$\text{tolerance} = 1 - R_j^2 \tag{2}$$

$$\text{VIF} = \frac{1}{tolerance} \tag{3}$$

R_j is the coefficient of determination of a regression of the j-th independent variable on other independent variables. If the tolerance is close to 0, the collinearity problem is quite serious. The larger the VIF is, the more severe the collinearity is. Generally, if the VIF is less than 5, the collinearity is not serious, and the standard for a broader point is less than 10 [40].

3.2. Eigenvector Spatial Filtering Poisson Regression

3.2.1. Construction of Spatial Weight Matrix

In this study, Thiessen polygons are applied to define the neighboring relationship between two different geographical units. In climatology, Thiessen polygons, also called Voronoi diagrams are applied to estimate the precipitation in the neighborhood area of a point-shape unit [41]. This study considers that the area covered by a river within a Thiessen polygon has similar hydrological conditions based on the basic fact that close hydrological stations on the same drainage area usually encounter high water levels and flood alarming events at the same time when flooding occurs. When a spatial weighted matrix W is constructed, the queen neighborhood is often applied to determine whether two units is neighbor each other [42]. Two polygons that share common points or edges are considered neighbors. the value in the matrix is set to 1 if the i-th unit is linked to the j-th unit, otherwise, the value is set to 0.

As is seen in Figure 15, the larger the red bubble is and the darker the color of the polygon, the more frequently flood alarming events occurred. Geographical units with high frequency tend to accumulate with each other and Thiessen polygons with low frequency are close to each other.

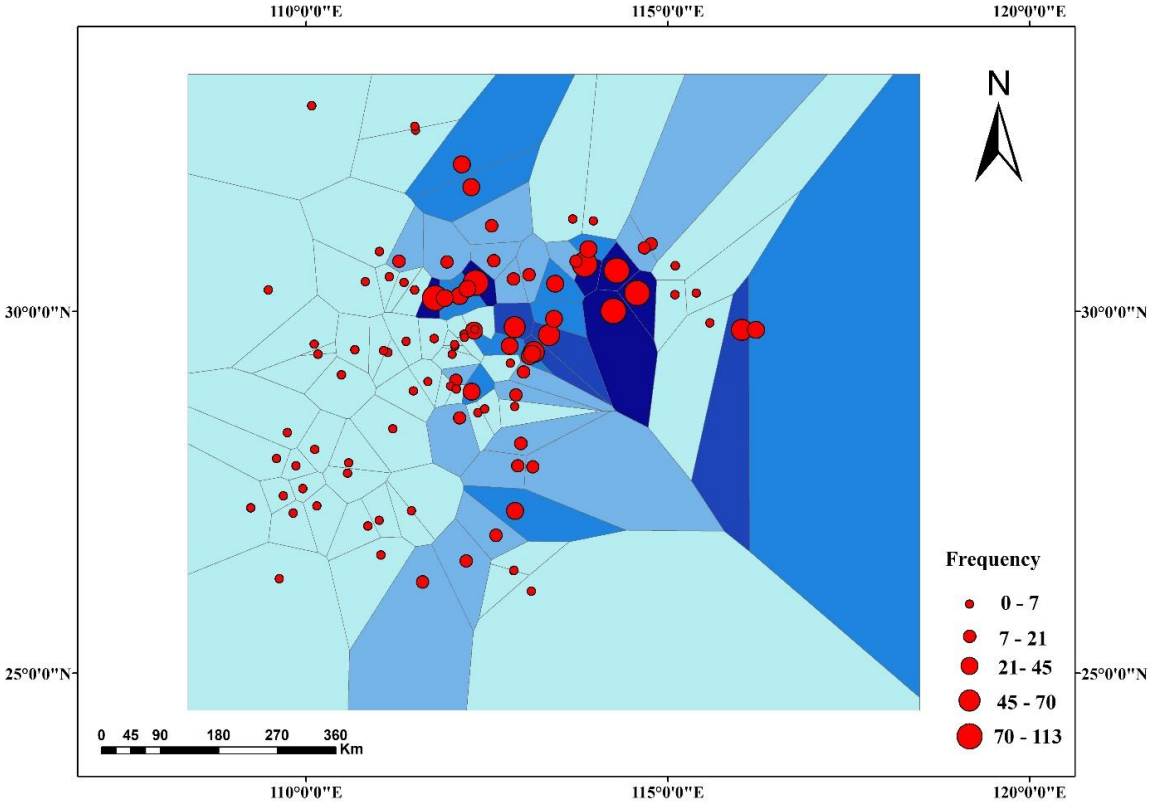

Figure 15. Graduated frequency and distribution of flood alarming events.

After performing a spatial autocorrelation test on the frequency of flood alarming events and the residuals of frequency calculated by ordinary least squares (OLS), relatively significant autocorrelation is found in Table 2: the Moran's coefficients are 0.32 and 0.23, respectively, which are higher than the expected values, −0.01 and −0.03. The R square of 0.2623 means that ordinary linear regression is not fit for modeling. A *p*-value that is close to zero means that significant spatial autocorrelation exists and eigenvector spatial filtering algorithms can be applied.

Table 2. Results of the spatial autocorrelation test.

	Moran's I	Expectation	Variance	*p*-Value	R Squared
Frequency	0.3246	−0.0097	0.0040	1.547×10^{-7}	
OLS residual	0.2255	−0.039	0.0043	2.743×10^{-5}	0.2623

3.2.2. Calculation of Eigenvectors

Before the calculation of eigenvectors, the raw spatial neighborhood matrix must be row-standardized by Equation (4):

$$C = (I - \frac{11^T}{n})W(I - \frac{11^T}{n}) \quad (4)$$

where n is the number of studied units, W is the spatial neighborhood matrix of $n \times n$, I is an identity matrix of $n \times n$, 1 is a vector of $n \times 1$ whose values are all 1, and C is called the centralized weighted matrix. The decomposition of the centralized weighted matrix can be expressed as Equation (5):

$$C = E\Lambda E^T \quad (5)$$

The decomposition of C generates eigenfunctions that contain n eigenvectors and n corresponding eigenvalues [33]. The n eigenvectors can be denoted as $E = (E_1, E_2, \cdots, E_n)$ and each of eigenvector is capable of capturing latent spatial autocorrelation at different scales [34]. Λ is an n-by-n diagonal matrix, and its diagonal elements are n eigenvalues of C, which can be denoted in descending order as $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n)$. It should be noted that all eigenvectors are orthogonal and uncorrelated [34]. Moreover, Moran's I of every eigenvector is calculated by Equation (6):

$$\text{Moran's I}_i = \frac{n}{1^T C 1}\lambda_i \tag{6}$$

where λ_i is the i-th eigenvalue.

3.2.3. Z-Score Normalization

In a multicriterion evaluation system, the evaluation criteria usually have different dimensions and orders of magnitude due to the different natures. When the levels between the indicators differ greatly, if the analysis is performed directly with the original index values, the role of the higher-value indicators in the comprehensive analysis will be highlighted, and the effect of the lower-level indicators will be relatively weakened. Therefore, to ensure the reliability of the results and compare the weights of different factors, the original indicator data must be standardized [43]. In addition, normalization can improve the convergence speed of the model. The normalization of data involved scaling the data down to a small specific interval.

The methods of normalization include min-max normalization and z-score normalization. Z-score normalization is applied to transform nine factors before regression because the minimum and maximum of each factor are unknow in the current datasets. The Z-score transformation of the j-th factor of the i-th sample X_{ij} is as Equation (7):

$$Z_{ij} = \frac{X_{ij} - \overline{X_j}}{S_j} \tag{7}$$

where $\overline{X_j}$ and S_j are the mean and standard deviation of the j-th factor.

3.2.4. Eigenvector Selection

We choose eigenvectors whose Moran's I values are greater than or equal to a threshold as eigenvectors with larger Moran's I values indicates more spatial autocorrelation, and the threshold has been researched in some previous studies on spatial autocorrelation [32]. The parameter can also be lifted to limit the number of candidate eigenvectors and complexity of the final model and 0.25 is chosen for the threshold. In the statistical model, stepwise regression is an automatic method of selecting among independent variables under certain criterion [44]. The AIC is a comprehensive metric of goodness of fit and complexity of model and a model with lower value of AIC mean a better one. Based on the Poisson regression model, forward stepwise selection of independent variables including factors and spatial synthetic variables is chosen to conduct and if the new model has a lower AIC than the previous model and passes the significance test of residuals of spatial autocorrelation after an eigenvector is added to the model, the eigenvector will be included in the final model. If the eigenvector cannot lower the AIC of the current regression model or significant spatial autocorrelation of the residuals still exists between the predicted values output by the current model and ground truth values, the eigenvector will be omitted. When all eigenvectors have been checked, the final Poisson model is output. The detailed process is shown on the flowchart in Figure 16.

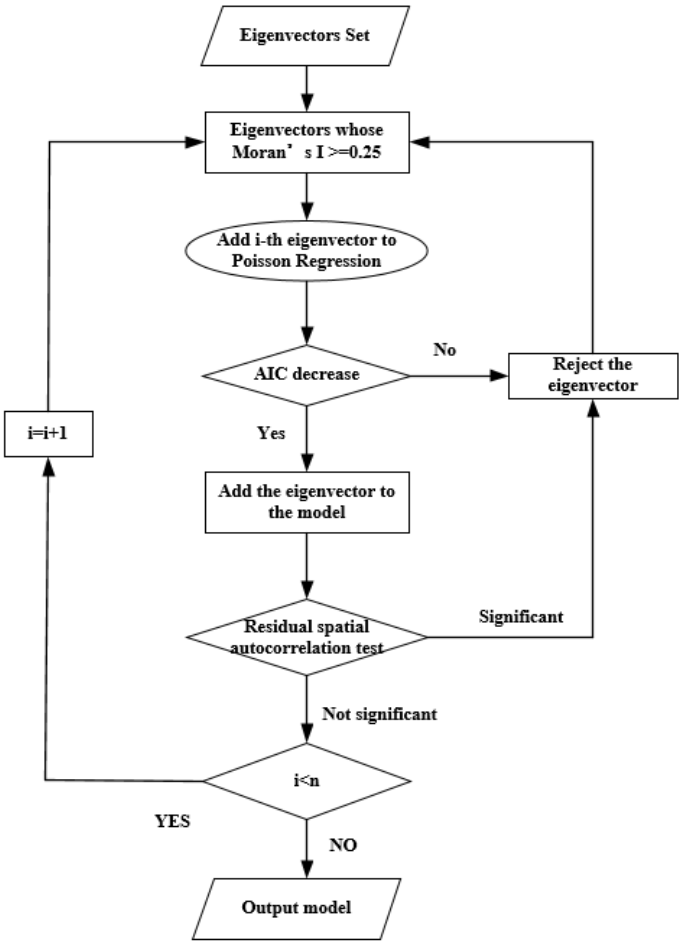

Figure 16. Flowchart of eigenvector selection.

3.2.5. Coefficient Calculation

Different from multivariate linear regression whose response variable is a continuous variable, Poisson regression assumes that the response variable, which is the frequency of flood alarming events in the study, is an observed count that follows the Poisson distribution, whose possible values of Y are non-negative integers. More importantly, this model is based on the hypothesis that a large count or frequency is rare. In the model, the Poisson incidence rate μ is regressed by its k predictors X_i (ELE, SLO, ESD, DEN, DIST, NDVI, RAIN, ACC, EXE) and coefficients $\beta_1, \beta_2, \ldots, \beta_9$ are estimated weight coefficients from flood alarming datasets. Poisson regression for sample i is written as Equations (8) and (9):

$$\Pr(Y_i = y_i | y_i, t_i) = \frac{e^{-\mu_i t_i}(\mu_i t_i)^{y_i}}{y_i!} \tag{8}$$

$$\mu_i = t_i \times \mu\left(X_i{}^T \beta\right) = t_i \times \exp(\beta_1 ELE_i + \beta_2 SLO_i + \ldots + \beta_9 EXE_i) \tag{9}$$

where t_i can be seen as an intercept of linear combinations of factors. Therefore, we can reasonably assume that frequent flood alarming events occurred with low possibility at a certain hydrological station, while most hydrological stations encountered few records of flood alarming events as illustrated in the frequency distribution of flood alarming events in Figure 3.

Similarly, the coefficient of ESFPS can be formulated as Equation (10):

$$\mu_i = t_i \times exp(\beta_1 ELE_i + \beta_2 ESD_i + \ldots + \beta_9 EXE_i + \sum_{k=1}^{l} \alpha_k E_k) \tag{10}$$

where E_k is the selected eigenvectors by the forward stepwise algorithm, and l is the number of the selected eigenvectors.

3.3. Model Assessment

The Poisson model (PS) and negative binomial regression (NB) are used to compare with ESFPS, and these models are provided with the same hazard-causing factors and data processing. Comparison of model assessment is conducted mainly by two aspects, performance of fitness and generalization error. Performance of fitness is assessed by pseudo R-squared and AIC while the metric of generalization error is the mean square error of leave-one-out cross validation.

3.3.1. Accuracy of Model Fitting

The R-squared value which is widely used in linear regression cannot be extended to Poisson regression models, and some pseudo R-squared models have been proposed to measure the interpretation of the Poisson model. In Poisson regression, the most popular pseudo R-squared measure is the function of the log-likelihoods of three models as shown in Equation (11):

$$Pseudo\ R^2 = \frac{LL_{fit} - LL_0}{LL_{max} - LL_0} \tag{11}$$

where LL_0 is the log-likelihood of the intercept-only model, LL_{fit} is the log-likelihood of the current model, and LL_{max} is the maximum log-likelihood possible.

3.3.2. AIC

The Akaike information criterion is a measure of the relative quality of statistical models for a given set of data. The AIC of one model is calculated as Equation (12):

$$\mathrm{AIC} = 2k - 2\ln\left(\hat{L}\right) \tag{12}$$

where k is the number of parameters and $\hat{L}$ represents the maximum likelihood value for the model. Therefore, this metric can serve as the means for model comparison and selection [45]. Among several candidate models for the samples, the preferred model is the one with the minimum AIC value. In this study, AIC is not only used in the comparison of the final candidate models, but also taken as a prespecified criterion in the stepwise algorithms.

3.3.3. Leave-One-Out Cross Validation (LOOCV)

In supervised learning applications of statistical learning theory, the out-of-sample error or generalized error is a measurement of how accurately a model is capable of predicting a response value for previously unseen data [46]. It is noteworthy that stepwise regression encounters the issue of overfitting because the algorithm will search a very large number of candidate models using brute force. As for LOOCV conducted every time, only one sample is used as the test dataset and the other samples are used as training dataset. If we have k samples, we will train k times and test k times. Although this method demands frequent and complex computation, it has a high utilization rate of samples and is especially suitable for small datasets.

3.4. Flood Risk Mapping

Given hazard-causing factors in other parts of the studied region, a regression model can be used to predict the expected frequency of flood alarming events. The expected frequency obtained by Poisson regression also represents the risk of flood alarming events in the format of frequency while the probability output by the logistic regression indicates the flood risk in the format of probability [47]. After comparing the Poisson model, negative binomial model and ESFPS model, the best one will

be selected to predict the expected frequency of flood alarming events throughout the central basin of the Yangtze river. In the GIS system, the linear vector data of river network is firstly transferred to point-shape vector that represents the range of the middle reaches of the Yangtze river. Secondly, these points extract the corresponding value from the raster of factors that pass the significance test. Thirdly, ESFPS transform these factors into a weighted score and output the expected frequency. Finally, point-shape vector among the whole study area is converted back into the raster format for visualization.

4. Results

4.1. Factor Selection

In multicollinearity diagnosis, the variance inflation factor and correlation coefficient matrix are calculated, and the results are plotted respectively in Table 3 and Figure 17. The correlation coefficient between slope (SLO) and elevation standard deviation (ESD) is 0. 97, close to 1 and VIF of slope and elevation standard deviance is 10.994 and 10.460 therefore one of these two factors must be excluded. The results of multicollinearity diagnosis of the remaining eight factors can be seen in Table 4 after slope is excluded and the result shows that there is hardly strong multicollinearity in these remaining eight factors.

Table 3. Variance inflation factor.

Factor	ELE	SLO	ESD	DEN	RAIN	EXE	DIST	ACC	NDVI
VIF	1.299	10.944	10.460	1.783	2.678	6.095	1.156	5.339	1.090

Figure 17. correlation coefficient matrix.

Table 4. Variance inflation factor after factor selection.

Factor	ELE	ESD	DEN	RAIN	EXE	DIST	ACC	NDVI
VIF	1.331	1.236	1.837	2.804	5.980	1.155	5.265	1.099

4.2. Results of Model Coefficients

The maximum threshold of the *p*-value that passes significant tests is set to 0.05. As shown in Table 5, the Poisson model considers seven factors including ELE, ESD, DIST, RAIN, EXE, ACC and NDVI significant while the negative binomial model only accepts ELE, DIST, ACC as significant factors. ESF-based Poisson regression model considers six factors including ELE, DEN, DIST, EXE, ACC and NDVI significant. Among the eight factors, according to their estimated coefficients and significance in Table 5, three models all shows that elevation, elevation standard deviation, distance to mainstream and NDVI are negatively correlated with the frequency of flood alarming events while the frequency of extreme rainfall is positively related to it. Moreover, annual mean rainfall cannot pass the significance test of NB and ESFPS while river density cannot pass the test of PS and NB, and ACC, DIST and ELE are all significant in three models.

Table 5. Results of the three regressions.

Model	**PS**		**NB**		**ESFPS**	
Factors	**Coefficient**	***p*-Value**	**Coefficient**	***p*-Value**	**Coefficient**	***p*-Value**
ELE	−1.58272	0 ***	−0.9996	0.001683 **	−0.0066	0 ***
ESD	−0.25210	0 ***	−0.19299	0.368332	−0.0769	0.28752
DEN	−0.06189	0.11095	0.07266	0.749230	0.0630	0.000915 **
DIST	−0.56152	0 ***	−0.79702	0.000618 ***	−0.0187	0 ***
RAIN	−0.11253	0.00136 **	−0.03248	0.897488	0.0006	0.5612
EXE	0.59830	0 ***	0.58400	0.053373	1.0334	0 ***
ACC	2.876	0 ***	−0.70197	0.014845 *	5.6894	0 ***
NDVI	−0.2313	0 ***	−0.18582	0.252906	−2.3345	0 ***
Intercept	1.7188	0 ***	1.8193	0 ***	1.4003	0 ***

In Table 5, the signs of significance ***, **, * and · represent that the *p*-value is less than the threshold of 0, 0.001 and 0.05, respectively.

Regression equation of the PS model is written as Equation (13):

$$\begin{aligned}\mu_i = \exp(&-1.58272 \times ELE_i + (-0.2521) \times ESD_i + (-0.56152) \times DIST_i + (-0.11253) \times RAIN_i \\ &+0.5983 \times EXE_i + 2.876 \times ACC_i + (-0.0312) \times NDVI_i + 1.7188)\end{aligned} \tag{13}$$

Regression equation of the NB model is written as Equation (14):

$$\mu_i = \exp((-0.996) \times ELE_i + (-0.79702) \times DIST_i + (-0.70197) \times ACC_i + 1.8193) \tag{14}$$

Regression equation of the ESFPS model is written as Equation (15):

$$\begin{aligned}\mu_i = \exp(&(-0.0066) \times ELE_i + 0.063 \times DEN_i + (-0.0187) \times DIST_i \\ &+1.033 \times EXE_i + 5.6894 \times ACC_i + (-2.335) \times NDVI_i + 4.193 \times EV1 \\ &+8.377 \times EV2 + (-4.048) \times EV7 + 1.4003)\end{aligned} \tag{15}$$

In the equations of the ESFPS model, $EV1$, $EV2$ and $EV7$ are the selected eigenvectors as seen in Table 6.

Table 6. Eigenvectors selected by ESFPS.

Eigenvectors	**Coefficient**	***p*-Value**
EV1	4.1934	0.039604 *
EV2	8.3770	0.000133 **
EV7	−4.0476	0.002899 ***

The signs of significance **, * and * represent that the *p*-value is less than the threshold of 0, 0.001 and 0.05, respectively.

4.3. Model Assessment

The Poisson regression model, the negative binomial regression model and the eigenvector spatial filtering Poisson regression fit the samples with the performance as shown in Table 7:

Table 7. Performances of three models.

Model	AIC	Pseudo R Squared	LOOCV	Moran's I (Expectation)
PS	1979.4	0.56	817.9884	−0.01929 (−0.0097)
NB	944.86	0.47	608.1647	0.05737 (−0.0097)
ESFPS	1040.3	0.78	430.4137	0.01958 (−0.0097)

Moran's I for the residuals of the predicted frequency of flood alarming events are all reduced to be closer to the expected value (−0.0097), and the results shows that no spatial autocorrelation exists in the residuals. While the pseudo R squared of PS (0.56) is greater than that of NB (0.47), the AIC of PS, 1979.4 is greater than that of NB, 944.86. ESFPS has the best fitness with the Pseudo R Squared, 0.78, which obviously outperforms PS and NB. Generalized error of ESFPS, 430.4137, calculated by LOOCV is less than that of NB, 608.1647 and both of them are less than generalized error of PS, 817.9884. Although the AIC of ESFPS 1040.3 is 10.1% greater than that of NB, 944.86, ESFPS is much better than PS and NB because ESFPS has better performance of fitness accuracy and generalized accuracy. ESFPS improves pseudo R squared from 0.56 to 0.78, by 39.3% and reduces generalized error from 817.9884 to 430.4137.

4.4. Spatial Pattern of Flood Risk

After the ESFPS model with better accuracy is obtained, the expected frequency throughout the basin can be calculated by the ESFPS model because the frequency of flood alarming events represents the risk of flood occurrences at a hydrological station. More hydrological stations should be planned and built in the place under high risk while obsolete stations have to be removed or repaired. As the river network in the format of raster shows in Figure 18, the darker the color is, the higher the flood risk of the river is. In the upper branch of the Hanjiang River, Poyang Lake District, Zuzhang River and Dongting Lake District, the red pixels indicate high flood risk and yellow pixels indicate an ordinary hazard level. Hydrology departments should pay attention to these sub-basins and strengthen their hydrological infrastructure. In general, sub-basins with high river density and frequent extreme consecutive rainfalls are susceptible to flood alarming events.

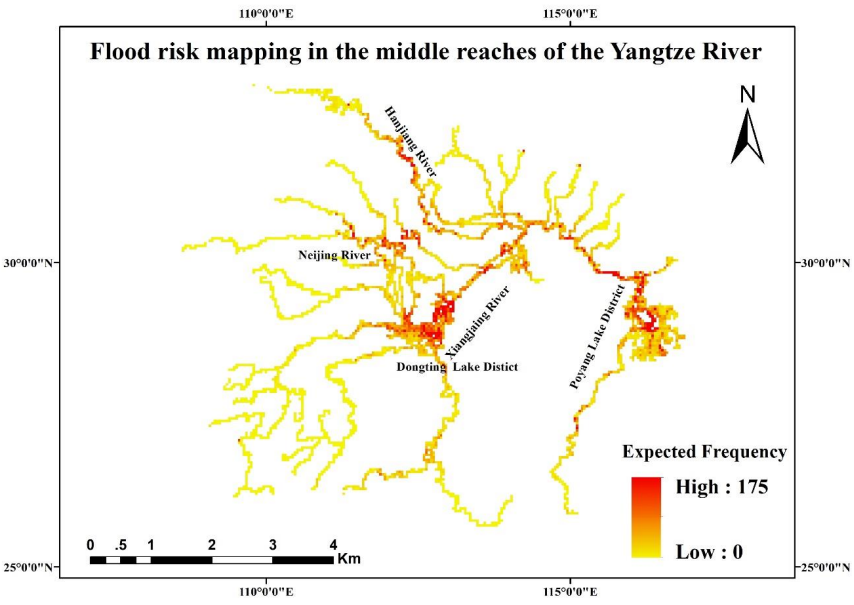

Figure 18. Predicted frequency of flood alarming risk.

5. Discussion

5.1. Improvement in the Accuracy of the Flood Risk EVALUATION Model

This study models the relationship between the frequency of flood alarming events observed on hydrological stations and nine factors using the Poisson regression model. Inclusion of spatial autocorrelation in the regression model increases the accuracy of fitting and prediction. The pseudo R squared is improved by 39.3% from 0.56 to 0.78. Regarding the generalization performance of these models, LOOCV of the ESFPS model is better than that of the Poisson model and negative binomial model. Moran's I of residuals for the ESFPS model is near to its expected value. In general, ESF-based Poisson regression is better than other counterparts.

5.2. Determination of Significant Factors

Both the PS model and the ESFPS model consider elevation, frequency of extreme rainfall, distance to mainstreams, NDVI and mean annual maximum of three-day accumulated precipitation as significant factors. Among the eight factors, according to their estimated coefficients and significance, elevation, elevation standard deviation, distance to the mainstream and NDVI are negatively correlated with the frequency of flood alarming events while frequency extreme rainfall, elevation and river density is positively related with it. Extreme rainfall events represented by frequency of extreme rainfall and mean annual maximum of three-day accumulated precipitation together play a significant role in flood events. These results are consistent with many previous researches and human experiences. Six factors that pass significant test in the ESFPS model are ranked according to the absolute value of weight coefficients in Table 8. Comparing their weights of these factors, accumulation and NDVI are the two main impact factors, with the weights of (+5.6894) and (−2.3345). Hence, short-term heavy rainfall is a leading trigger for the flood in the study area and vegetation has an important resistance to flood.

Table 8. Rank of factors.

Factor	Weight	Rank
ACC	5.6894	1
NDVI	−2.3345	2
EXE	1.0334	3
DIST	−0.0187	4
DEN	0.063	5
ELE	−0.0066	6

5.3. Limitations and Future Enhancements

The complexity of the ESFPS model is greater than that of the PS model because it requires more time to select eigenvectors. In addition, The study only focuses on the factors on the earth surface and trigger factors, and the type of soil and lithology is often considered as important factors in the previous study [48]. Flood is also influenced by how much water can be stored in the soil from previous flooding and local rainfall. Different types of soil and lithology have different water retention capacities. The study initially oversees these two factors because it tries to map flood risk and the flood-alarming events of the hydrological stations on the water instead of the whole regions spanned by the middle Yangtze River. Social and economic factors such as land use and constructions of dikes and drainage facilities should have been included in the model because the capacity of discharging runoff swiftly are represented by these factors. Subsequent research should also consider social factors such as the type of land use [49] and investment in hydrological infrastructures such as dikes and pumping stations [21]. From the perspective of data, hydrological data that ranges from 2005 to 2012 lack timeliness and the flood risk map output by the model can function as an early warning system if new recent data in larger size are used. Moreover, it is more reasonable for the conceptualization

of neighborhood relationship for the spatial weight matrix to consider the topology of river network because different segments belonging to the same river possibly share similar hydrological observations and flood alarming events could be observed by hydrological stations in downstream segments and upstream segments at the same time. Hence, soil type, density of dikes and pumping stations and new neighborhood relationship will be considered in the further research.

6. Conclusions

The study proposes a Poisson regression based on eigenvector spatial filtering to model the flood risk at hydrological stations using the frequency of flood alarming events. The frequency of flood alarming events and eight factors including elevation, elevation standard deviation, river density, annual mean rainfall, mean annual maximum of three-day accumulated precipitation, frequency of extreme rainfall and NDVI are used to train the regression model. The PS, NB and ESFPS models reach the conclusion that elevation, frequency of extreme rainfall, distance to rivers, NDVI and mean annual accumulation of three-day accumulated precipitation are important factors leading to high flood risk. In the study area, heavy consecutive rainfall is the chief culprit of the flood and vegetation can reduce flood risk efficiently. Moreover, Inclusion of autocorrelation of the frequency of flood alarming events into the model robustly improved the accuracy of fitness and prediction. Stepwise regression using the AIC can efficiently select the eigenvectors that represent spatial distribution pattern of flood alarming frequency. Eigenvectors with high eigenvalues and Moran's I are accepted in the final ESFPS model. The ESFPS model can be used to map flood risks throughout basins according to the expected value of frequency that the model output after the data related to the eight factors are input into the model. In the study, flood risk mapping output by ESFPS and new recent data can function as an early warning system and help identify where hydrological stations should be built and strengthened for monitoring of water level.

Author Contributions: conceptualization, Y.C.; methodology, Y.C. and T.F.; software, J.L. and L.H.; validation, H.T.; formal analysis, T.F.; investigation, J.C.; resources, T.F.; data curation, H.T.; writing—original draft preparation, T.F.; writing—review and editing, Y.C.

Funding: This research was funded by Ministry of Science and Technology of the People's Republic of China, grant number [2017YFB0503704] and the National Nature Science Foundation of China, grant number [41671380]. And the APC was funded by [2017YFB0503704].

Acknowledgments: This work was supported by National Key R&D Program of China: [grant number 2017YFB0503704] and the National Nature Science Foundation of China [grant number 41671380].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Petrucci, O.; Aceto, L.; Bianchi, C.; Bigot, V.; Brázdil, R.; Pereira, S.; Kahraman, A.; Kılıç, Ö.; Kotroni, V.; Llasat, M.C.; et al. Flood fatalities in Europe, 1980–2018: Variability, features, and lessons to learn. *Water* **2019**, *11*, 1682. [CrossRef]
2. Schelske, O.; Sundermann, L.; Hausmann, P. *Mind the Risk—A global Ranking of Cities Under Threat from Natural Disasters*; Swiss Reinsurance Company Ltd.: Zurich, Switzerland, 2013.
3. Zhang, L.; Geng, J.; Fan, C. The comprehensive analysis of flood disasters losses in china from 2000 to 2010. *IOP Conf. Ser. Mater. Sci. Eng.* **2018**, *466*, 012023. [CrossRef]
4. Snedaker, S.; Rima, C. Chapter 4—Risk assessment. In *Business Continuity and Disaster Recovery Planning for It Professionals*, 2nd ed.; Snedaker, S., Rima, C., Eds.; Syngress: Boston, MA, USA, 2014; pp. 151–224.
5. He, Y.; Pappenberger, F.; Manful, D.; Cloke, H.; Bates, P.; Wetterhall, F.; Parkes, B. 5.16—Flood inundation dynamics and socioeconomic vulnerability under environmental change. *Clim. Vulnerability* **2013**, 241–255. [CrossRef]
6. Pei, F.; Wu, C.; Qu, A.; Xia, Y.; Wang, K.; Zhou, Y. Changes in extreme precipitation: A case study in the middle and lower reaches of the yangtze river in China. *Water* **2017**, *9*, 943. [CrossRef]

7. Hsieh, S.-H.; Liu, L.-W.; Chung, W.-G.; Wang, Y.-M. Sensitivity analysis on the rising relation between short-term rainfall and groundwater table adjacent to an artificial recharge lake. *Water* **2019**, *11*, 1704. [CrossRef]
8. Hashizume, M. 1.10—precipitation and flood hazards: Health effects, risks, and impacts. *Clim. Vulnerability* **2013**, 115–124. [CrossRef]
9. Legesse, D.; Vallet-Coulomb, C.; Gasse, F. Hydrological response of a catchment to climate and land use changes in tropical africa: Case study south central ethiopia. *J. Hydrol.* **2003**, *275*, 67–85. [CrossRef]
10. Chen, Y.R.; Yeh, C.-H.; Yu, B. Integrated application of the analytic hierarchy process and the geographic information system for flood risk assessment and flood plain management in Taiwan. *Nat. Hazards* **2011**, *59*, 1261–1276. [CrossRef]
11. Chau, K.W.; Wu, C.L.; Li, Y.S. Comparison of several flood forecasting models in yangtze river. *J. Hydrol. Eng.* **2005**, *10*, 485–491. [CrossRef]
12. Bisht, D.S.; Chatterjee, C.; Kalakoti, S.; Upadhyay, P.; Sahoo, M.; Panda, A. Modeling urban floods and drainage using swmm and mike urban: A case study. *Nat. Hazards* **2016**, *84*, 749–776. [CrossRef]
13. Sharma, S.K.; Kwak, Y.J.; Kumar, R.; Sarma, B. Analysis of hydrological sensitivity for flood risk assessment. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 51. [CrossRef]
14. Neelz, S.N.; Pender, G. *Benchmarking of 2d Hydraulic Modelling Packages*; Environment Agency: Bristol, UK, 2010.
15. Lu, C.; Zhou, J.; He, Z.; Yuan, S. Evaluating typical flood risks in yangtze river economic belt: Application of a flood risk mapping framework. *Nat. Hazards* **2018**, *94*, 1187–1210. [CrossRef]
16. Malczewski, J. A gis-based approach to multiple criteria group decision-making. *Int. J. Geogr. Inf. Syst.* **1996**, *10*, 955–971. [CrossRef]
17. Wang, Y.; Li, Z.; Tang, Z.; Zeng, G. A gis-based spatial multi-criteria approach for flood risk assessment in the dongting lake region, Hunan, central China. *Water Resour. Manag.* **2011**, *25*, 3465–3484. [CrossRef]
18. Chen, Y.; Liu, R.; Barrett, D.; Gao, L.; Zhou, M.; Renzullo, L.; Emelyanova, I. A spatial assessment framework for evaluating flood risk under extreme climates. *Sci. Total Environ.* **2015**, *538*, 512–523. [CrossRef] [PubMed]
19. Tehrany, M.S.; Pradhan, B.; Mansor, S.; Ahmad, N. Flood susceptibility assessment using gis-based support vector machine model with different kernel types. *Catena* **2015**, *125*, 91–101. [CrossRef]
20. Xiong, J.; Li, J.; Cheng, W.; Wang, N.; Guo, L. A gis-based support vector machine model for flash flood vulnerability assessment and mapping in China. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 297. [CrossRef]
21. Xiao, Y.; Yi, S.; Tang, Z. Integrated flood hazard assessment based on spatial ordered weighted averaging method considering spatial heterogeneity of risk preference. *Sci. Total Environ.* **2017**, *599–600*, 1034–1046. [CrossRef]
22. Kourgialas, N.N.; Karatzas, G.P. A national scale flood hazard mapping methodology: The case of Greece—Protection and adaptation policy approaches. *Sci. Total Environ.* **2017**, *601–602*, 441–452. [CrossRef]
23. Leggett, D.J.; Jones, A. The application of gis for flood defence in the anglian region: Developing for the future. *Int. J. Geogr. Inf. Syst.* **1996**, *10*, 103–116. [CrossRef]
24. Dawod, G.M.; Mirza, M.N.; Al-Ghamdi, K.A. Gis-based estimation of flood hazard impacts on road network in Makkah city, Saudi Arabia. *Environ. Earth Sci.* **2012**, *67*, 2205–2215. [CrossRef]
25. Mandallaz, D.; Ye, R. Prediction of forest fires with poisson models. *Can. J. For. Res.* **1997**, *27*, 1685–1694. [CrossRef]
26. Wahiduzzaman, M.; Yeasmin, A. Statistical forecasting of tropical cyclone landfall activities over the north Indian ocean rim countries. *Atmos. Res.* **2019**, *227*, 89–100. [CrossRef]
27. Betts, M.G.; Diamond, A.W.; Forbes, G.J.; Villard, M.A.; Gunn, J.S. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. *Ecol. Model.* **2006**, *191*, 197–224. [CrossRef]
28. Tobler, W. On the first law of geography: A reply. *Ann. Assoc. Am. Geogr.* **2004**, *94*, 304–310. [CrossRef]
29. Getis, A. *Spatial Filtering in a Regression Framework: Examples Using Data on Urban Crime, Regional Inequality, and Government Expenditures*; Springer-Verlag: Berlin, German, 2010.
30. Getis, A.; Griffith, D.A. Comparative spatial filtering in regression analysis. *Geogr. Anal.* **2002**, *34*, 130–140. [CrossRef]
31. Murakami, D.; Griffith, D.A. Random effects specifications in eigenvector spatial filtering: A simulation study. *J. Geogr. Syst.* **2015**, *17*, 1–21. [CrossRef]
32. Chun, Y.; Griffith, D.A.; Lee, M.; Sinha, P. Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. *J. Geogr. Syst.* **2016**, *18*, 67–85. [CrossRef]

33. Griffith, D.A. A linear regression solution to the spatial autocorrelation problem. *J. Geogr. Syst.* **2000**, *2*, 141–156. [CrossRef]
34. Griffith, D.; Chun, Y. *Spatial Autocorrelation and Spatial Filtering*; Springer-Verlag: Berlin, German, 2003.
35. Kourgialas, N.N. A flood risk decision making approach for mediterranean tree crops using gis; climate change effects and flood-tolerant species. *Environ. Sci. Policy* **2016**, *63*, 132–142. [CrossRef]
36. Wu, Y.; Zhong, P.A.; Zhang, Y.; Xu, B.; Ma, B.; Yan, K. Integrated flood risk assessment and zonation method: A case study in huaihe river basin, China. *Nat. Hazards* **2015**, *78*, 635–651. [CrossRef]
37. Rawat, P.K. Impacts of climate change and hydrological hazards on monsoon crop patterns in the lesser himalaya: A watershed based study. *Int. J. Disaster Risk Sci.* **2012**, *3*, 98–112. [CrossRef]
38. Jiang, S.; Ren, L.; Hong, Y.; Yang, X.; Ma, M.; Zhang, Y.; Yuan, F. Improvement of multi-satellite real-time precipitation products for ensemble streamflow simulation in a middle latitude basin in south China. *Water Resour. Manag.* **2014**, *28*, 2259–2278. [CrossRef]
39. Farrar, D.E.; Glauber, R.R. Multicollinearity in regression analysis: The problem revisited. *Rev. Econ. Stat.* **1967**, *49*, 92–107. [CrossRef]
40. Kutner, M.H.; Nachtsheim, C.J.; Neter, J. *Applied Linear Regression Models*; McGraw-Hill Irwin: New York, NY, USA, 2004.
41. Zhu, Q.A.; Zhang, W.C.; Zhao, D.Z. Topography-based spatial daily precipitation interpolation by means of prism and thiessen polygon analysis. *Sci. Geogr. Sin.* **2005**, *25*, 233–238.
42. Zhang, J.; Li, B.; Chen, Y.; Chen, M.; Fang, T.; Liu, Y. Eigenvector spatial filtering regression modeling of ground $PM_{2.5}$ concentrations using remotely sensed data. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1228. [CrossRef]
43. Grus, J. *Data Science from Scratch: First Principles with Python*; O'Reilly Media: New York, NY, USA, 2015.
44. Hocking, R.R. A biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics* **1976**, *32*, 1–49. [CrossRef]
45. Akaike, H. IEEE xplore abstract—A new look at the statistical model identification. *IEEE Autom. Control Trans.* **1974**, *19*, 716–723. [CrossRef]
46. Kohavi, R. A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995.
47. Youssef, A.M.; Pradhan, B.; Sefry, S.A. Flash flood susceptibility assessment in Jeddah city (Kingdom of Saudi Arabia) using bivariate and multivariate statistical models. *Environ. Earth Sci.* **2016**, *75*, 12. [CrossRef]
48. Tehrany, M.S.; Pradhan, B.; Jebur, M.N. Spatial prediction of flood susceptible areas using rule based decision tree (dt) and a novel ensemble bivariate and multivariate statistical models in gis. *J. Hydrol.* **2013**, *504*, 69–79. [CrossRef]
49. Kazakis, N.; Kougias, I.; Patsialis, T. Assessment of flood hazard areas at a regional scale using an index-based approach and analytical hierarchy process: Application in rhodope–evros region, Greece. *Sci. Total Environ.* **2015**, *538*, 555–563. [CrossRef] [PubMed]