water

MDPI

Article

Flood Risk Assessment of Global Watersheds Based on Multiple Machine Learning Models

Xiangnan Li [1,2], Denghua Yan [1,2], Kun Wang [1,2], Baisha Weng [1,2,*], Tianling Qin [1,2] and Siyu Liu [3]

1 State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100038, China
2 Water Resources Department, China Institute of Water Resources and Hydropower Research, Beijing 100038, China
3 China Institute of Water Resources and Hydropower Research, Beijing 100038, China
* Correspondence: baishaweng@163.com

Received: 26 June 2019; Accepted: 7 August 2019; Published: 10 August 2019

Abstract: Machine learning algorithms are becoming more and more popular in natural disaster assessment. Although the technology has been tested in flood susceptibility analysis of several watersheds, research on global flood disaster risk assessment based on machine learning methods is still rare. Considering that the watershed is the basic unit of water management, the purpose of this study was to conduct a risk assessment of floods in the global fourth-level watersheds. Thirteen conditioning factors were selected, including: maximum daily precipitation, precipitation concentration degree, altitude, slope, relief degree of land surface, soil type, Manning coefficient, proportion of forest and shrubland, proportion of artificial surface, proportion of cropland, drainage density, population, and gross domestic product. Four machine learning algorithms were selected in this study: logistic regression, naive Bayes, AdaBoost, and random forest. The global susceptibility assessment model was constructed based on four machine learning algorithms, thirteen conditioning factors, and global flood inventories. The evaluation results of the model show that the random forest performed better in the test, and is an efficient and reliable tool in flood susceptibility assessment. Sensitivity analysis of the conditioning factors showed that precipitation concentration degree and Manning coefficient were the main factors affecting flood risk in the watersheds. The susceptibility map showed that fourth-level watersheds in the global high-risk area accounted for a large proportion of the total watersheds. With the increase of extreme hydrological events caused by climate change, global flood disasters are still one of the most threatening natural disasters. The global flood susceptibility map from this study can provide a reference for global flood management.

Keywords: machine learning; global fourth-level watersheds; flood susceptibility

1. Introduction

Global climate change is contributing to an increase in extreme weather events, resulting in numerous natural disasters, of which flood is the most devastating disaster [1,2]. Floods threaten residential lives and property, change the natural environment, pollute water resources, and have a profound impact on human society and ecosystems [3,4]. Flood disaster refers to the damage caused by floods to human life and production activities, including casualties, disappearances, or economic losses (the main object of this study was the flood disasters caused by precipitation) [1,2]. According to statistics, the annual economic losses caused by flood disasters in the world amount to 50 billion dollars, and the number of people affected by the disasters is nearly 100 million [5]. Flood disasters occur more frequently than other natural disasters such as landslides, volcanoes, or earthquakes,

and the scope of their impact is broader [6]. The flood risk in a basin is often related to the local precipitation characteristics, the underlying surface conditions, and the adaptability of the basin to disasters [7,8]. Widespread increases in heavy precipitation events have been observed, even in places where total amounts have decreased [9,10]. For example, statistically significant increases in the occurrence of heavy precipitation have been observed across Europe and North America [11,12]. An increase in population and the rapid development of urban construction also increase the risk of flooding in a basin [13,14]. Areas with more population density, more agricultural land, or more concentrated river networks are often more prone to flood disasters. Therefore, carrying out river basin flood risk assessment on a global scale is important for reducing flood disasters and watershed management [8,15].

Various methods have been used to identify and evaluate flood-susceptible areas. For example, some studies have used multi-criteria decision analysis methods, including the analytic hierarchy process and the expert scoring method [16–18]. These methods are based on expert knowledge and are susceptible to uncertainty [19]. Physically based models such as VIC and MIKE models at a regional scale, and other hydrological models at the continental and global scale have also been used to study floods, and have shown great advantages in regional or global flood process research [8,20–23]. Recently, machine learning methods such as artificial neural networks (ANN), support vector machines (SVM), and decision trees (DT) have been applied to flood hazard assessments, which can identify and evaluate flood-prone areas based on the training and testing of large amounts of data [24–27]. By learning the relationship between flooding occurrence and the explanatory factors from the historical flooding records, the machine learning models avoid the subjective determination of weights [28]. The physical model uses a simplified parameter to characterize the physical law, which can simulate a natural or time-continuous phenomenon. This is its advantage over machine learning models. However, in the face of the complexity of global floods, flood models at a global scale often require a large number of model parameters, repeated model debugging, and long computing time [19,21–23,26–28]. Machine learning methods may be a good choice for faster access to global flood hazard assessments.

However, current flood risk assessment based on machine learning methods is always concentrated in a single watershed. For example, Tehrany et al. used SVM with different kernel types in flood susceptibility mapping in Kuala Terengganu, Malaysia [29]. Zhao et al. applied a semi-supervised machine learning model in urban flood susceptibility assessment in Beijing, China [28]. In these machine learning models, the input training samples are only attributes of a flood occurrence point or non-occurrence point in the basin. In fact, the occurrence of floods is due to the comprehensive properties of the basin, rather than the attributes of a single sample point. In addition, machine learning methods have rarely been used for global flood risk assessment, so it is necessary to conduct flood risk assessments for global watersheds.

In this study, a machine learning model for global flood risk assessment was built based on 60,863 fourth-level watersheds, four machine learning methods including logistic regression, naive Bayes, AdaBoost, and random forest, and 13 conditioning factors. Based on this, a flood disaster susceptibility map of the global watersheds was constructed to provide reference for global watershed management and flood disaster identification.

2. Data and Methods

2.1. Data

2.1.1. Global Fourth-Level Watersheds

The basin is the basic unit of hydrological management. Many institutions around the world have obtained global multi-level watershed mapping through different technical means [30,31]. Due to the remote sensing topographic data error, low spatial resolution, and small ground fluctuation, it is difficult to obtain real digital river networks in plains areas based on digital elevation data and GIS technology, which has a series of adverse effects on the subsequent calculation and evaluation [32].

For plains areas, we adopted the "stream burning" method. Specific steps included: (1) When the Google Earth image was enlarged to the finest resolution, the center line of the river was drawn manually according to the real river image using the line drawing function in the Google Earth. (2) The original DEM (Digital Elevation Model) was modified based on the correct digital rivers obtained from Google Earth using the stream burning method. (3) After revising the DEM, the correct digital river network was rebuilt by the standard hydrological processes of ArcGIS (D8 method).

Based on the topological relationship of the first-level to fourth-level rivers, we coded the river networks from outlet to source, from large to small, and from coarse to fine. We used the end points of the rivers to obtain the watershed boundaries. The watershed inherited the code of the corresponding river. The global first-level to fourth-level rivers and corresponding watersheds dataset has been published on the Figshare data platform (https://doi.org/10.6084/m9.figshare.8044184.v3) [33]. The global watershed classification included 60,863 fourth-level watersheds, as shown in Figure 1.

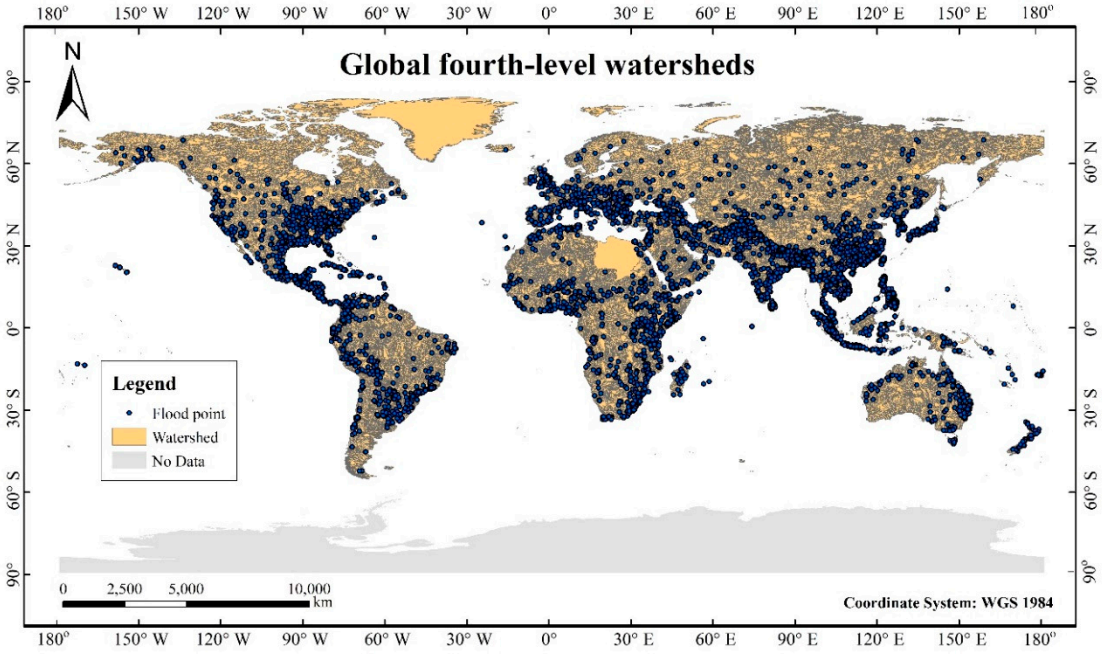

Figure 1. Global fourth-level watersheds and the location of flood inventories.

2.1.2. Flood Disaster Inventory

Accurate analysis of flood susceptibility requires a precise flood inventory map that shows the locations of flood occurrences [34]. There are several existing flood databases, such as the International Disaster Database (EM-DAT) and the Global Active Archive of Large Flood Events, and in other studies at the global scale, flood extent observations or detailed reconstructions of 2D hydraulic models have been used [22,35]. Various non-conventional sources of information (such as amateur videos, photographs, news reports, etc.) also provide data for the reconstruction of flood events [36]. The flood disaster inventory data in this study were derived from the Global Active Archive of Large Flood Events, Dartmouth Flood Observatory, University of Colorado [37]. The database is supported by NASA, the Japanese Space Agency, and the European Space Agency, and is widely used worldwide. The archive has recorded a large number of flood disaster data since 1985, mostly from news, governmental, instrumental, and remote sensing sources, and provides accurate geographical locations of flood disasters. Therefore, we selected this database as the flood disasters sample data.

For this study, 4730 flood disaster data from January 1985 to March 2019 were selected. Based on ArcGIS 10.5 software, the flood sample dataset and the global fourth-level watershed dataset were superimposed and analyzed, and 3335 watersheds with flood disasters were obtained. The distribution of the flood sample points is shown in Figure 1. Sample points where no flood has occurred also have

a great influence on the model results. However, we were unable to obtain sample points that had not experienced flood disasters from the existing database. The general method was based on existing data; non-flooding sample points were randomly selected in the remaining unrecorded flood areas, but this method often leads to false identification [24–27]. After all, the existing database cannot accurately record all flood disaster samples. We referred to previous studies and added conditions to samples that had not experienced flooding [28]. Through literature and previous studies, it was found that deserts or ice fields were less likely to have flood or flood damage [1,3,37]. Therefore, 1500 watersheds were randomly selected in the desert and ice fields. Although these sample data made the conditions of non-flooding more severe, they helped to improve the flood control standards of the basin accordingly. Values of 1 and 0 were assigned to indicate the existence and absence of flood disaster, respectively. These data samples were randomly divided into a training dataset (70%) and a testing dataset (30%) for the machine learning model.

2.1.3. Flood Conditioning Factors

Identifying the conditioning factors is a key step for flood susceptibility assessment. Thirteen conditioning factors were selected in this study by reviewing previous studies and investigating the mechanisms of flood, including maximum daily precipitation (MDP), precipitation concentration degree (PCD), altitude, slope, relief degree of land surface (RDLS), soil type (ST), Manning coefficient (MC), proportion of forest and shrubland (PFS), proportion of artificial surface (PAS), proportion of cropland (PC), drainage density (DD), population, and gross domestic product (GDP) [19,27–29]. These data are calculated based on data such as digital elevation and land use. It is worth noting that, unlike previous studies, this study used the average (or major) value of the conditioning factor for each watershed as input data. That is, based on the spatial statistics module of ArcGIS, we calculated the average (or major) value of the raster data corresponding to the watershed. When training and testing, these data were normalized according to the following formula:

$$F = \frac{f}{f_{max}} \tag{1}$$

where f is the original value of a certain conditioning factor and f_{max} is the maximum value of the factor.

1. Maximum daily precipitation

Precipitation is the direct factor affecting the occurrence of floods. The NCEP reanalysis data is a complete, comprehensive dataset produced by the National Centers for Environmental Prediction, which contains global precipitation data with a spatial resolution of 2.5° and a time resolution of 6 h [38]. NCEP reanalysis data are widely used worldwide. Based on the dataset, the MDP from 1985 to 2017 was calculated, and the data were resampled to the resolution of 0.01° by the resample module of ArcGIS. The resampling algorithm was "NEAREST," which minimized changes to pixel values since no new values were created. The global distribution of average MDP for each fourth-level watershed is shown in Figure 2a.

2. Precipitation concentration degree

PCD is an indicator that reflects the distribution of precipitation over time. The more concentrated precipitation is, the higher the frequency of occurrence of heavy precipitation [39]. Taking the year as the calculation period, according to the principle of vector analysis, the precipitation was decomposed into vectors in the x and y directions. The precipitation in a month was the length of the vector, and the azimuth of the corresponding month was the direction of the vector. The azimuth angle was 360° throughout the study period, and the azimuth distribution of each month was evenly distributed (Table 1). PCD can be calculated according to the following formula:

$$PCD = \frac{\sqrt{R_x^2 + R_y^2}}{R} \tag{2}$$

$$R_x = \sum_{i=1}^{n} r_i \sin \theta_i \quad (3)$$

$$R_y = \sum_{i=1}^{n} r_i \cos \theta_i \quad (4)$$

where R is the total precipitation, r_i is the precipitation of the i-th month, and θ_i is the azimuth corresponding to the month.

Table 1. Azimuth angle of each month.

Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Azimuth angle	15	45	75	105	135	165	195	225	255	285	315	345

It was found that PCD reflected the concentration of total precipitation in a certain research period, and the value ranged from 0 to 1. If the precipitation was concentrated in a certain month, the ratio of the length of its composite vector to the total amount of precipitation was 1, and the PCD was the maximum. If the monthly precipitation was equal, the modulus of the composite vector was 0, and the PCD was the minimum.

Based on the NCEP reanalysis data, the global average PCD data from 1985 to 2017 were obtained. The average PCD data were then resampled to the resolution of 0.01° by ArcGIS with the NEAREST algorithm. The global distribution of average PCD for each fourth-level watershed is shown in Figure 2b.

3. Altitude

Altitude is also one of the most important factors affecting flood disasters [40]. In general, altitude and flood are inversely related; that is, floods are more likely to occur in areas with lower altitudes. A global digital elevation map (DEM) with 90 × 90 m pixel size was created in ArcGIS. Based on the spatial statistics module of ArcGIS, the average altitude of each watershed was obtained. The global distribution of average altitude for each fourth-level watershed is shown in Figure 2c.

4. Slope

Slope is an important geomorphological feature that triggers flood disasters [19,41]. Slope directly affects the generation of surface runoff and the infiltration of precipitation, and river basins with large slopes in mountainous areas are often more prone to flood disaster. A global slope map with 90 × 90 m pixel size was created in ArcGIS. Based on the spatial statistics module of ArcGIS, the average slope of each watershed was obtained. The global distribution of average slope for each fourth-level watershed is shown in Figure 2d.

5. Relief degree of land surface

The RDLS is the difference between the highest altitude point and the lowest altitude point in the watershed. It represents an important indicator of regional topographical features and landform types. A global RDLS map with 90 × 90 m pixel size was created in ArcGIS. Based on the spatial statistics module of ArcGIS, the average slope of each watershed was obtained. The global distribution of RDLS for each fourth-level watershed is shown in Figure 2e.

6. Soil type

Soil depth, soil texture, and soil porosity are the main factors affecting surface runoff. They mainly affect runoff generation by changing the infiltration characteristics and water holding characteristics of the soil. Different soil types indicate different soil properties, so soil type was selected as a conditioning factor. Soil data were derived from the Harmonized World Soil Database v1.0 of FAO (Food and Agriculture Organization of the United Nations) [42]. The Harmonized World Soil Database is a 30 arc–second raster database with over 15,000 different soil mapping units that combines existing regional and national updates of soil information worldwide. A global ST map was built in ArcGIS

with a resolution of 0.01°. Based on the spatial statistics module of ArcGIS, the major soil type of each watershed was obtained. The global distribution of the main ST for each fourth-level watershed is shown in Figure 2f.

7. Manning coefficient

For a long time, the Manning equation has been applied in the analysis of river flow resistance. In practice, the Manning coefficient is usually used to reflect the resistance characteristics of different underlying surface conditions, which are the key parameters affecting flow concentration and flood evolution. According to the reference, different Manning coefficient values are assigned to different land use characteristics, and the average Manning coefficient of each fourth-level watershed was calculated by the following formula [43,44]:

$$M = \frac{\sum_{i=1}^{n} m_i p_i}{n} \quad (5)$$

where m_i denotes the Manning coefficient of the land use type i, p_i denotes the area ratio of the land use type i to the watershed, and n is the number of land use types in the watershed.

The land use data in this paper were derived from the land use raster data of globe30, which is produced by the Chinese government using remote sensing data (www.globeland30.com) [45]. Ten types of land use were included in the data: land surface waters, wetlands, woodlands, grasslands, shrubs, artificial surface, arable land, glaciers and permanent snow, tundra, bare land.

According to the calculation results and ArcGIS, the Manning coefficient distribution map of the global fourth-level watershed was obtained, as shown in Figure 2g.

8. Proportion of forest and shrubland

Vegetation is one of the key factors affecting runoff. Good vegetation can play a role in preventing soil erosion, conserving water sources, and alleviating floods. Based on the land use data of globe30, the proportion of forest area and shrub area in each fourth-level watershed was calculated. The global distribution of the PFS for each fourth-level watershed is shown in Figure 2h.

9. Proportion of artificial surface

Artificial surface is where the population is mainly concentrated. These areas generally have a small infiltration coefficient and are more likely to form surface runoff and floods. Based on the land use data of globe30, the proportion of artificial surface area in each fourth-level watershed was calculated. The global distribution of the PAS for each fourth-level watershed is shown in Figure 2i.

10. Proportion of cropland

Most of the cropland in the world is distributed in flat areas which are more conducive to flood concentration. In addition, agricultural loss is also an important indicator for evaluating flood disasters. Based on the land use data of globe30, the proportion of cropland area in each fourth-level watershed was calculated. The global distribution of the PC for each fourth-level watershed is shown in Figure 2j.

11. Drainage density

Drainage density is a basic feature of a river system, which has effects on peak flows when rainfall occurs in a watershed [46]. Drainage density refers to the ratio of total river length to watershed area in the basin. In this study, river lengths were not calculated for rivers below level four. Based on the river network data in Section 2.1.1, the drainage density of each fourth-level basin was calculated. The global distribution of the DD for each fourth-level watershed is shown in Figure 2k.

12. Population

The distribution of population is directly related to flood disasters, and the consequences of disasters caused by floods in densely populated areas are also greater. Considering that the most

accurate demographic data is released by national governments, based on governmental data, we revised the population data released by the World Bank and FAO to obtain more accurate global population density distribution data [47,48]. The population of each fourth-level watershed was obtained in ArcGIS, as shown in Figure 2l.

13. Gross Domestic Product

GDP is also closely related to flood disasters. Areas with higher GDP tend to have a stronger ability to adapt to disasters, but the consequences of disasters may be greater. Based on GDP data from the World Bank and ArcGIS, the GDP of each of fourth-level watershed in the world was obtained, as shown in Figure 2m [49].

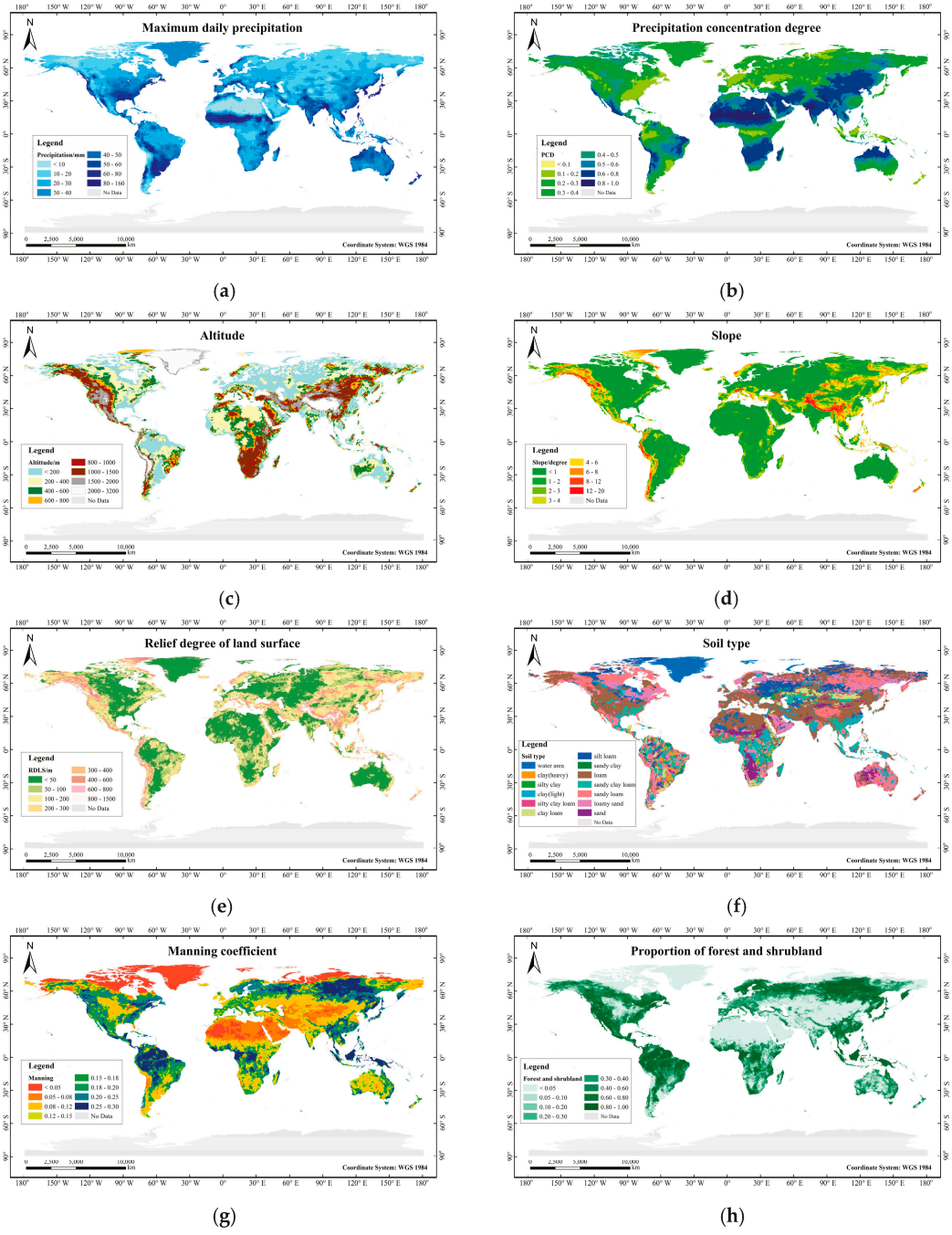

Figure 2. *Cont.*

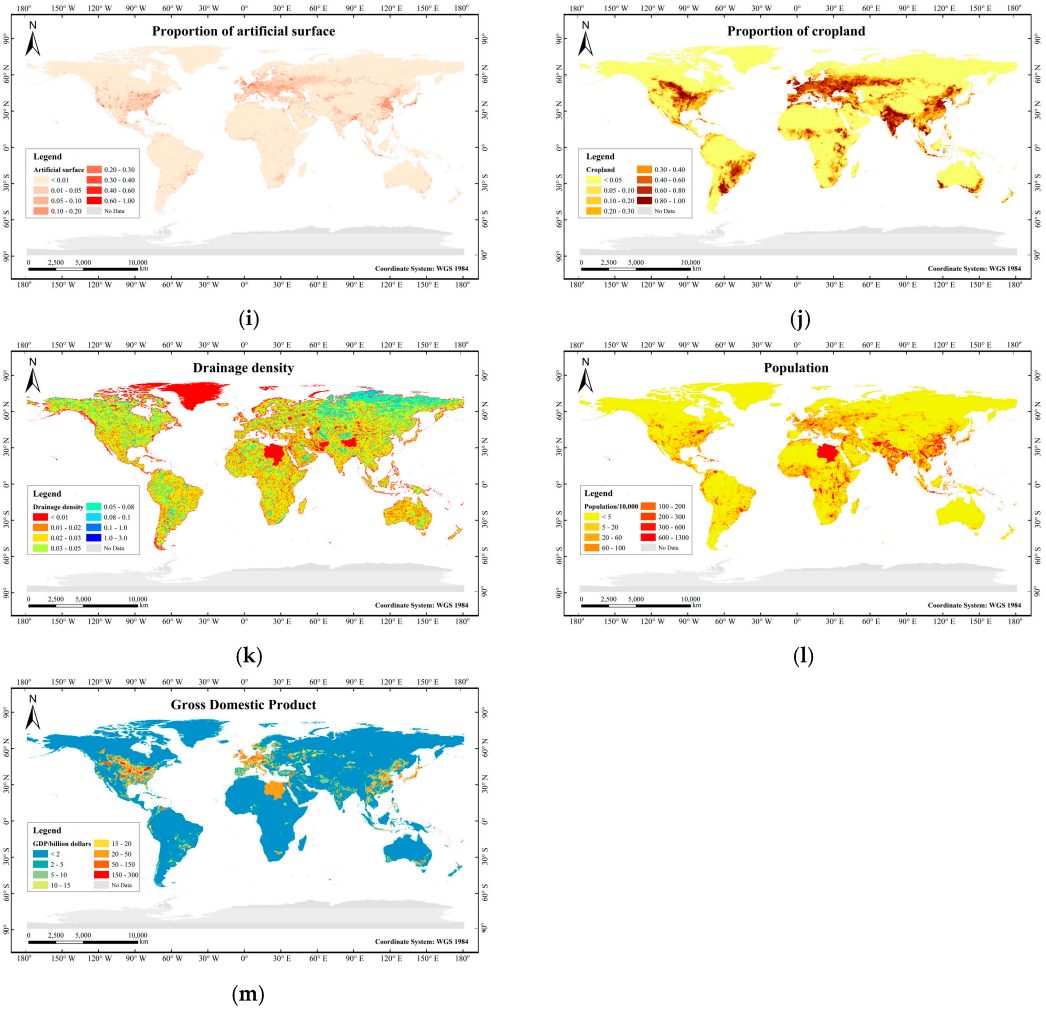

Figure 2. Conditioning factors of this study: (**a**) maximum daily precipitation (MDP), (**b**) precipitation concentration degree (PCD), (**c**) altitude, (**d**) slope, (**e**) relief degree of land surface (RDLS), (**f**) soil type (ST), (**g**) Manning coefficient (MC), (**h**) proportion of forest and shrubland (PFS), (**i**) proportion of artificial surface (PAS), (**j**) proportion of cropland (PC), (**k**) drainage density (DD), (**l**) population, and (**m**) gross domestic product (GDP).

2.2. Methods

2.2.1. Logistic Regression

Logistic regression (LR) was developed by Cox and is a non-linear regression model for solving binary problems [50,51]. Its dependent variable has a value between 0 and 1, so it can explain the probability of certain phenomena occurring. The logistic regression model is widely used in flood risk assessment. When the result of the logistic regression is closer to 1, the probability of a flood disaster is greater. The basic form of the logistic regression model is:

$$p = \frac{1}{1 + e^{-z}} \tag{6}$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{7}$$

where p is the probability of flooding disaster, β_i (i = 0, 1, 2, . . . , n) represents the regression coefficient of the model, and x_i (i = 0, 1, 2, . . . , n) represents different conditioning factors.

2.2.2. Naive Bayes

The naive Bayes (NB) model is a classification algorithm based on the Bayesian theorem and feature condition independent hypothesis [52,53]. The Bayesian conditional probability formula is:

$$p(Y_K|X) = \frac{p(X|Y_K)p(Y_K)}{p(X)} \tag{8}$$

Suppose the samples of the model are: $(x_1^1, x_2^1, \ldots, x_n^1, y_1), (x_1^2, x_2^2, \ldots, x_n^2, y_2), \ldots, (x_1^m, x_2^m, \ldots, x_n^m, y_m)$; that is, there are m samples, each sample has n features, and the model output has k categories, defined as $C1, C2, \ldots, Ck$. The prior probability and conditional probability of the naive Bayes can be obtained from the samples as follows:

$$p(Y = C_k)(k = 1, 2, \ldots, K) \tag{9}$$

$$p(X = x|Y = C_k) = p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|Y = C_k) \tag{10}$$

When making predictions, it is necessary to calculate k conditional probabilities according to Equation (11), and then find the category corresponding to the largest conditional probability as the judgment result. Equation (11) can be obtained according to Equations (9) and (10):

$$p(Y = C_k|X = X^{test}) = \frac{p(X = X^{test}|Y = C_k)p(Y = C_k)}{p(X = X^{test})} \tag{11}$$

In flood risk assessment, when the conditional probability of a flood occurrence is greater than the flood non-occurrence, the test sample is judged to be flooding, otherwise, no flooding.

2.2.3. AdaBoost

The adaptive boosting algorithm (AB) is an iterative algorithm [54,55]. The core idea is to train different weak classifiers for the same training set, and then combine these weak classifiers to form a stronger final classifier. The model mainly includes the following steps.

First, the weight distribution of the training data is initialized, and each training sample is given the same weight at the beginning:

$$w = \frac{1}{N} \tag{12}$$

The initial distribution of the training sample set is:

$$D(i) = (w_1, w_2, \ldots, w_n) = (\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}) \tag{13}$$

The algorithm iteratively calculates $t = 1, \ldots, T$, and selects the weak classifier h with the smallest error rate as the t-th basic classifier Ht. The error of the weak classifier on the distribution D_t is:

$$e_t = p(H_t(x_i) \neq y_i) = \sum_{i=1}^{N} w_{ti} I(H_t(x_i) \neq y_i) \tag{14}$$

The weight of the weak classifier in the final classifier is:

$$\alpha_t = \frac{1}{2} \ln(\frac{1 - e_t}{e_t}) \tag{15}$$

The weight distribution D_{t+1} of the training sample is updated:

$$D_{t+1} = \frac{D_t(i) e^{-\alpha_t y_i H_t(x_i)}}{2\sqrt{e_t(1 - e_t)}} \tag{16}$$

Finally, the weak classifiers are combined according to the weight α_t of the weak classifiers to obtain the final strong classifier:

$$H = sign(\sum_{t=1}^{T} \alpha_t H_t(x)) \tag{17}$$

2.2.4. Random Forest

Random forest (RF), developed by Breiman, is an ensemble machine learning algorithm which uses a large number of classification or regression trees to make a prediction [56]. In this study, the response variable was modeled using a regression tree. It generates different sets of samples by sampling with replacement, and generates multiple corresponding regression tree training models, and then determines the data classification according to the voting results of multiple classifiers.

In the training of the regression tree, rules based on the response variables are established to classify the observations until the prediction has the smallest possible node deviation. The rule of regression trees is a collection of linear partitions of observed data that together create a nonlinear decision surface. One of the main problems with regression trees is that they tend to overfit the training data, and therefore perform poorly when given unknown data [57]. Random forest is a way to address this weakness. When an individual regression tree is trained in the random forest algorithm, a portion of the input records and predictor variables are randomly selected as input to the training. A set of regression trees is created after multiple sampling exercises, and each set of regression trees is only a training result for a randomly selected subset. It is obviously not advisable to use a full sample to train decision trees, because full sample training ignores the laws of local samples.

2.2.5. Evaluation Methods

In order to evaluate the effects of the four models, we selected the four indicators: precision (P), recall (R), F-score (F), and the area under the ROC (Receiver Operating Characteristic) curve (AUC) to evaluate the model [28]. These indicators can represent the ability of the model to identify flood hazard risks.

$$P = \frac{TP}{TP + FP} \tag{18}$$

$$R = \frac{TP}{TP + FN} \tag{19}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{20}$$

where TP (true positive) is the number of samples correctly classified as flooding watersheds; TN (true negative) is the number of samples correctly classified as non-flooding watersheds; FP (false positive) is the number of samples incorrectly classified to flooding watersheds; and FN (false negative) is the number of samples incorrectly classified as non-flooding watersheds.

The ROC curve is a curve with the FP rate as the X-axis and TP as the Y-axis. AUC represents the area under the curve, and the higher the AUC value, the better the model performance.

In addition, based on the standard machine learning integration model Weka 3.8 software, we also compared the calculation time of the four models.

2.2.6. Sensitivity Analysis of Conditioning Factors

Assessing the contribution of different conditioning factors to flood disaster risk is important for the management of floods. This study used the sensitivity of AUC to conditioning factors to analyze its contribution to flood disaster. The sensitivity analysis was based on the Jackknife test, which is accepted to have a high capability for a broad range of practical problems [58]. The percentage of relative decrease (PRD) of the AUC was used to evaluate the contribution of each conditioning factor, as follows [19]:

$$PRD_i = 100 \times \frac{|AUC_{all} - AUC_i|}{AUC_{all}} \quad (21)$$

where AUC_{all} represents the AUC value when predicted with all conditioning factors. AUC_i and PRD_i are, respectively, the AUC value and the percentage of relative decrease of AUC when the i-th factor has been removed from the prediction process. The larger the value of PRD, the greater the effect of a factor on the result.

In order to evaluate the consistency of the PRD rankings in different models, we calculated the SD (standard deviation) of the PRD rankings of each conditioning factor in the four models. The smaller the standard deviation, the higher the consistency of the ranking.

$$SD = \sqrt{\frac{\sum_{i=1}^{4} (R_i - \overline{R})^2}{4}} \quad (22)$$

where R_i is the ranking of a conditioning factor in the model and?R is the average ranking among the four models.

3. Results

3.1. Model Analysis

The evaluation indicators of the four machine learning models for the prediction results of the testing dataset (30%) are shown in Table 2. It was found that for precision, whether in a flood zone or a non-flood zone, the RF was the best, and the values of P were 0.979 and 0.927, respectively. AB had the lowest P value in the flood zone evaluation, which was 0.929, and NB had the lowest P value in the non-flood zone evaluation, which was 0.728. For recall, RF's performance was also the best, with values of 0.966 and 0.954 in the flood zones and non-flood zones, respectively. NB had the lowest R value in the flood zone evaluation, which was 0.838, and AB had the lowest R value in the non-flood zone evaluation, which was 0.844. For F-score, RF performed best, while NB had the lowest F for both flood and non-flood evaluation. Figure 3 shows the ROC curves of the four models in the evaluation flood zone. It was found that the AUC value of RF was the largest. For the simulation time, the calculation time of the four models was relatively short, and they took only a few seconds to calculate. In summary, the four models performed well, while the RF model performed best in assessing global flood risk.

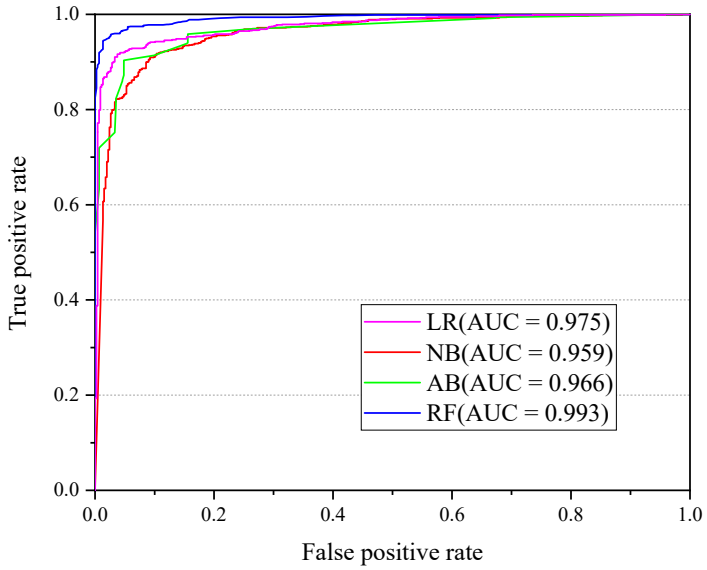

Figure 3. The ROC (Receiver Operating Characteristic) curves of the four models.

Table 2. Evaluation index of the four models.

Index	Logistic Regression (LR)		Naive Bayes (NB)		AdaBoost (AB)		Random Forest (RF)	
	Flood	**Non-Flood**	**Flood**	**Non-Flood**	**Flood**	**Non-Flood**	**Flood**	**Non-Flood**
Precision (P)	0.959	0.868	0.972	0.728	0.929	0.867	0.979	0.927
Recall (R)	0.939	0.912	0.838	0.947	0.941	0.844	0.966	0.954
F-score (F)	0.948	0.890	0.900	0.823	0.935	0.855	0.972	0.940
Time (s)	0.6		1.13		0.4		1.66	

3.2. Global Flood Susceptibility Map

A flood susceptibility map is important for spatial flood prediction and watershed management. Through four machine learning models and conditioning factor datasets of the global fourth-level watersheds, the characteristics of flood disasters in the global fourth-level watersheds were calculated, and, finally, the flood susceptibility map was obtained. The flood vulnerability map shows the flood susceptibility level of the global fourth-level watersheds based on the flood disaster data from January 1985 to March 2019. Based on the natural segmentation method of ArcGIS, the risk level of the flood susceptibility map was divided into five categories, including: lowest, low, medium, high, and highest, as shown in Figure 4. This segmentation method grouped data according to the inherent characteristics of the data, based on the principle of minimizing intra-group differences and maximizing inter-group differences for data sets.

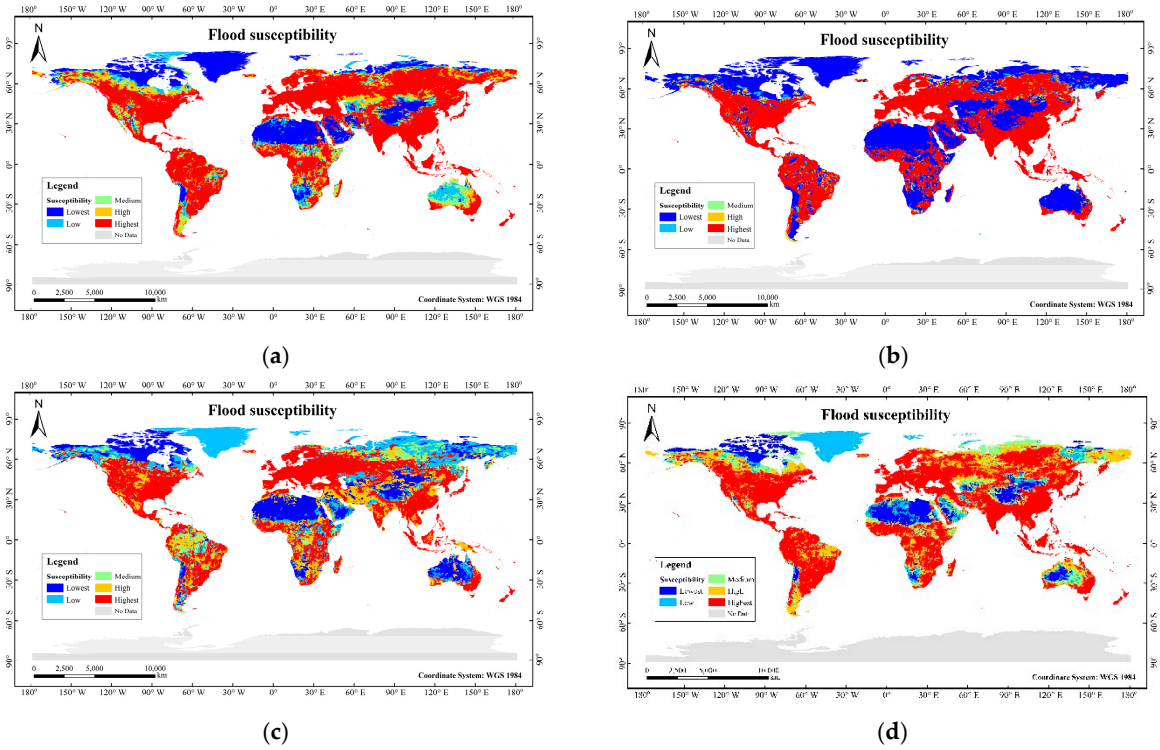

Figure 4. The global flood susceptibility map: (**a**) LR, (**b**) NB, (**c**) AB, and (**d**) RF.

It was found that although the global flood susceptibility maps obtained by different machine learning models were slightly different, the identification of high-risk areas was basically the same. This is similar to previous related research, that is, different machine learning methods have little effect on flood susceptibility maps [19,27,28].

According to the flood susceptibility map, we obtained the flood risk situation of each fourth-level watershed. Table 3 shows the number of fourth-level watersheds and percentage of area for different susceptibility levels. Among the four models, the area with high flood risk accounted for nearly 50% of

the global fourth-level watersheds. Floods around the world are still very severe, with high-risk areas in southern North America, northern and eastern South America, most of Europe, Southeast Asia, Central Africa and northeastern Australia. Northern North America, southwestern South America, the deserts of northern Africa, northwestern China and central Australia are low-risk areas for floods.

Table 3. Susceptibility level for each model of fourth-level watersheds.

Susceptibility Level	Number of Fourth-Level Watersheds				Percentage of Area (%)			
	LR	NB	AB	RF	LR	NB	AB	RF
Lowest	10,670	26,186	12,893	5952	17.10%	38.80%	18.60%	10.30%
Low	5725	1892	11,082	5026	7.40%	2.40%	14.50%	8.20%
Medium	5172	1158	5756	8101	7.40%	1.70%	7.00%	9.40%
High	6811	1598	9091	10,476	9.80%	2.40%	16.20%	15.70%
Highest	32,485	30,029	22,041	31,308	58.30%	54.70%	43.70%	56.40%

3.3. Assessment of Sensitivity of Conditioning Factors

In flood risk analysis, it is very important to choose the appropriate conditioning factors. This study analyzed the contribution of the 13 selected conditioning factors to the flood susceptibility model. According to Equation (21), machine learning models with different conditioning factors were constructed, and the PRDs of AUCs were calculated. The PRD of each conditioning factor is shown in Figure 5. It can be seen that the PRD results of different machine learning models were slightly different. In the LR and NB models, the MC and PCD were assessed as the most important factor, compared to GDP and PCD in the AB model, and altitude and PCD in the RF model. The standard deviation of the PRD rankings of each conditioning factor is shown in Figure 6. It was found that the standard deviations of PCD, ST, population, and PC in the four models were small, indicating that the PRD rankings of these conditioning factors were basically the same for these four models. The PRD rankings of GDP, PFS, and MDP showed large differences.

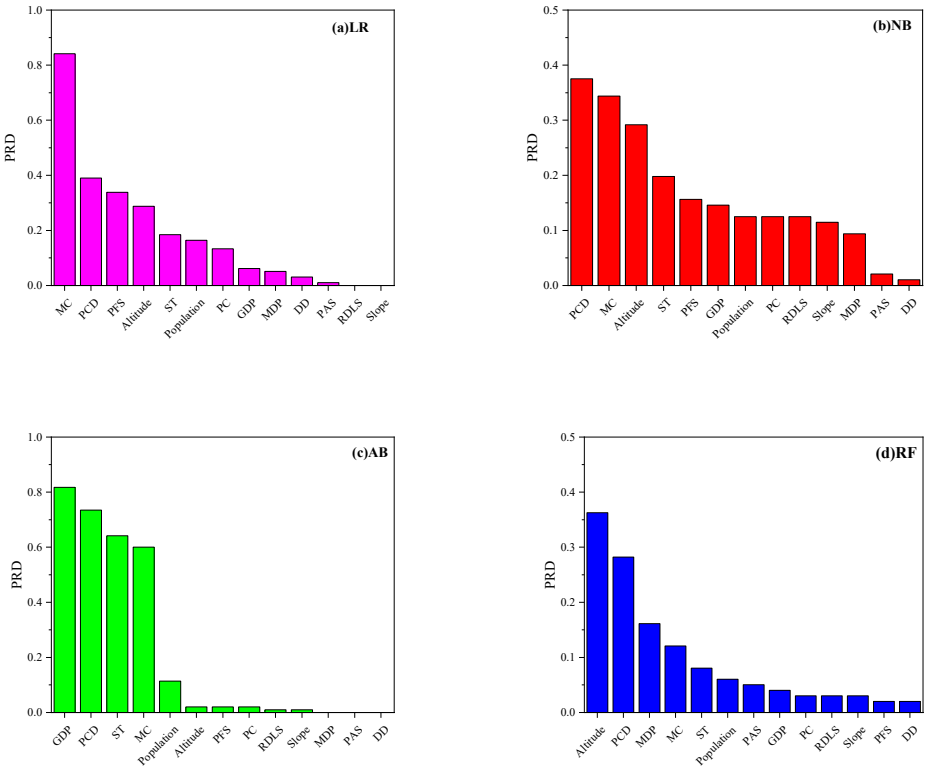

Figure 5. The percentage of relative decrease (PRD) of conditioning factors: (**a**) LR, (**b**) NB, (**c**) AB, and (**d**) RF.

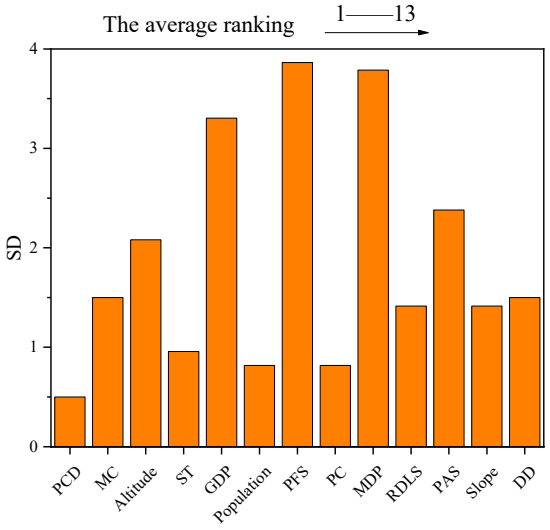

Figure 6. The standard deviations (SD) of PRD rankings of each conditioning factor.

In general, the MC and PCD were relatively important factors in the four models, while the DD and slope had little effect on the results. This is different from previous research results for a certain watershed, mainly because the average slope of the watershed was used as a topographic parameter in the global flood risk assessment, which may lead to the extinction of the extreme values.

4. Discussions

With the continuous development of machine learning algorithms, their applications in the field of hydrology are becoming more and more extensive [27–29,59–61]. Flood susceptibility maps, as an important basis for watershed planning and management, have also evolved from traditional human judgment to statistical analysis methods based on big data. However, research on global flood susceptibility maps is still relatively rare. Most flood susceptibility maps at this stage are for a specific basin. There are two problems with this method:

1. Selection of flood sample points. In a certain basin, accurately defining the criteria and location of flood disasters is very difficult, because general flood disasters are often large-scale [29].

2. Data set of conditioning factors. When conducting a risk assessment for a single watershed, the input conditioning factors are often the values of the flood sample points. In fact, the occurrence of a flood disaster at any location within the watershed is not only affected by the conditioning factors of the flood point, but by the influence of factors in the confluence area above the point.

These problems are likely to produce erroneous results in the judgment of flood susceptibility in a watershed. Although machine learning provides us with an analytical method, this method cannot violate the principles of hydrology.

This study obtained results for flood disaster risk assessment in the global fourth-level watersheds. However, global flood risk assessment obtained by machine learning methods is only a static result rather than a dynamic one, and the physically based model has an advantage. A physically based model can often give more detailed information on flood hazards, such as flow and submerged range, while current machine learning studies focus on qualitative assessment of flood hazards [21–23]. These qualitative evaluations can only provide limited reference for watershed management. In the future, the combination of machine learning and physical models will be better able to provide flood disaster assessment information.

The performance of machine learning models is highly dependent on the accuracy of labeled flood inventories. Apart from labeled flood inventories, the distribution and size of non-flood sample data affect model accuracy [27,28]. Future studies should explore the effective utilization of both flood and non-flood inventories and other massive information to improve the accuracy of results.

Due to the huge amount of global data, the conditioning factors selected in this paper may not have fully contained all the key factors. In addition, when making predictions, the average value of the conditioning factors of the watershed was selected as the independent variable, which may have caused some error in the result. The selection of non-flood sample points may also have caused an overestimation of flood susceptibility level. In the study of flood susceptibility maps, how to obtain accurate flood and non-flood sample points, and how to obtain the flood disaster conditioning factors corresponding to the points is still a problem worth exploring in the future.

5. Conclusions

Based on the four machine learning models of logistic regression, naive Bayes, AdaBoost, and random forest, this paper conducted flood risk assessment on global fourth-level watersheds and obtained global flood susceptibility maps. The results show that the random forest model performed best for prediction. According to the susceptibility map, fourth-level watersheds in the global high-risk area account for a large proportion. As the extreme hydrological events caused by climate change increase, this threat may not be relieved in the near future. Global flood disaster is still one of the most threatening natural disasters. Sensitivity analysis of conditioning factors showed that precipitation concentration degree and Manning coefficient were the main factors affecting flood risk in watersheds. The methods and ideas of this study can provide reference for flood management worldwide.

Author Contributions: Conceptualization, X.L. and D.Y.; Data curation, K.W.; Methodology, X.L., K.W., B.W. and T.Q.; Writing—original draft, X.L.; Writing—review & editing, S.L.

Funding: The study was supported by the National Key Research and Development Program of China (No. 2016YFA0601503), the National Natural Science Foundation of China (No. 51725905 and No. 91547209), and the National Natural Science Foundation of China (No. 41571037 and No. 51879276).

Acknowledgments: We are very grateful for the basic data provided by relevant research institutions. We also thank the editors and anonymous reviewers.

Conflicts of Interest: The authors declare no conflict of interest in any aspect of the data collection, analysis or the preparation of this paper.

References

1. UN International Strategy for Disaster Reduction (UNISDR). *Reducing Disaster Risks through Science: Issues and Actions*; UN International Strategy for Disaster Reduction: Geneva, Switzerland, 2009.
2. Centre for Research on the Epidemiology of Disasters (CRED). *The Human Cost of Weather-Related Disasters 1995–2015 Report*; Centre for Research on the Epidemiology of Disasters: Brussels, Belgium, 2015.
3. Hirabayashi, Y.; Mahendran, R.; Koirala, S.; Konoshima, L.; Yamazaki, D.; Watanabe, S.; Kim, H.; Kanae, S. Global flood risk under climate change. *Nat. Clim. Chang.* **2013**, *3*, 816. [CrossRef]
4. Dottori, F.; Salamon, P.; Bianchi, A.; Alfieri, L.; Hirpa, F.A. Development and evaluation of a framework for global flood hazard mapping. *Adv. Water Resour.* **2016**, *94*, 87–102. [CrossRef]
5. Re, M. *NatCat SERVICE Database*; Munich RE: Munich, Germany, 2014.
6. Youssef, A.M.; Pradhan, B.; Hassan, A.M. Flash flood risk estimation along the St. Katherine road, southern Sinai, Egypt using GIS based morphometry and satellite imagery. *Environ. Earth Sci.* **2011**, *62*, 611–623. [CrossRef]
7. Ahmadisharaf, E.; Tajrishy, M.; Alamdari, N. Integrating flood hazard into site selection of detention basins using spatial multi-criteria decision-making. *J. Environ. Plann. Manag.* **2016**, *59*, 1397–1417. [CrossRef]
8. Schumann, G.J.P.; Bates, P.D.; Apel, H.; Aronica, G.T. *Global Flood Hazard: Applications in Modeling, Mapping, and Forecasting*; American Geophysical Union: Washington, DC, USA, 2018.
9. Petersen, T.C.; Taylor, M.A.; Demeritte, R.; Duncombe, D.L.; Burton, S.; Thompson, F.; Porter, A.; Mercedes, M.; Villegas, E.; Martis, A.; et al. Recent changes in climate extremes in the Caribbean region. *J. Geophys. Res.* **2002**, *107*, 4601. [CrossRef]
10. Griffiths, G.M.; Salinger, M.J.; Leleu, I. Trends in extreme daily rainfall across the South Pacific and relationship to the South Pacific Convergence Zone. *Int. J. Climatol.* **2003**, *23*, 847–869. [CrossRef]

11. Kunkel, K.E.; Easterling, D.R.; Redmond, K.; Hubbard, K. Temporal variations of extreme precipitation events in the United States: 1895–2000. *Geophys. Res. Lett.* **2003**, *30*, 1900. [CrossRef]
12. Haylock, M.R.; Goodess, C.M. Interannual variability of extreme European winter rainfall and links with mean large-scale circulation. *Int. J. Climatol.* **2004**, *24*, 759–776. [CrossRef]
13. Barbosa, A.E.; Fernandes, J.N.; David, L.M. Key issues for sustainable urban stormwater management. *Water Res.* **2012**, *46*, 6787–6798. [CrossRef]
14. Xu, Z.; Zhao, G. Impact of urbanization on rainfall-runoff processes: Case study in the Liangshui River Basin in Beijing, China. Proc. *Int. Assoc. Hydrol. Sci.* **2016**, *373*, 7–12. [CrossRef]
15. European Commission. Directive 2007/60/EC of the European Parliament and of the Council of 23 October 2007 on the assessment and management of flood risks. *Off. J. Eur. Union L* **2007**, *288*, 27–34.
16. 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]
17. Tang, Z.; Zhang, H.; Yi, S.; Xiao, Y. Assessment of flood susceptible areas using spatially explicit, probabilistic multi-criteria decision analysis. *J. Hydrol.* **2018**, *558*, 144–158. [CrossRef]
18. Vojtek, M.; Vojteková, J. Flood Susceptibility Mapping on a National Scale in Slovakia Using the Analytical Hierarchy Process. *Water* **2019**, *11*, 364. [CrossRef]
19. Choubin, B.; Moradi, E.; Golshan, M.; Adamowski, J.; Hosseini, F.S.; Mosavi, A. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Sci. Total Environ.* **2019**, *651*, 2087–2096. [CrossRef]
20. Graham, D.N.; Butts, M.B. Flexible, integrated watershed modelling with MIKE SHE. In *Watershed Models*; CRC Press: Boca Raton, FL, USA, 2005.
21. Yamazaki, D.; Kanae, S.; Kim, H.; Oki, T. A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resour. Res.* **2011**, *47*, W04501. [CrossRef]
22. Sampson, C.C.; Smith, A.M.; Bates, P.D.; Neal, J.C.; Alfieri, L.; Freer, J.E. A high-resolution global flood hazard model. *Water Resour. Res.* **2015**, *51*, 7358–7381. [CrossRef]
23. Hoch, J.M.; Trigg, M.A. Advancing global flood hazard simulations by improving comparability, benchmarking, and integration of global flood models. *Environ. Res. Lett.* **2019**, *14*, 034001. [CrossRef]
24. Lee, S.; Kim, J.C.; Jung, H.S.; Lee, M.J.; Lee, S. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul Metropolitan City, Korea. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1185–1203. [CrossRef]
25. Dano, U.L.; Balogun, A.L.; Matori, A.N.; Yusouf, K.W.; Abubakar, I.R.; Mohamed, M.A.S.; Aina, Y.A.; Pradhan, B. Flood Susceptibility Mapping Using GIS-Based Analytic Network Process: A Case Study of Perlis, Malaysia. *Water* **2019**, *11*, 615. [CrossRef]
26. Lee, S.; Lee, S.; Lee, M.J.; Jung, H.S. Spatial Assessment of Urban Flood Susceptibility Using Data Mining and Geographic Information System (GIS) Tools. *Sustainability* **2018**, *10*, 648. [CrossRef]
27. Tehrany, M.S.; Pradhan, B.; Jebur, M.N. Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *J. Hydrol.* **2014**, *512*, 332–343. [CrossRef]
28. Zhao, G.; Pang, B.; Xu, Z.X.; Peng, D.Z.; Xu, L.Y. Assessment of urban flood susceptibility using semi-supervised machine learning model. *Sci. Total Environ.* **2019**, *659*, 940–949. [CrossRef]
29. Tehrany, M.S.; Pradhan, B.; Mansor, S.; Ahman, N. Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena* **2015**, *125*, 91–101. [CrossRef]
30. European Environment Agency (EEA). *European Catchments and Rivers Network System (Ecrins)*; European Environment Agency: Kobenhavn, Denmark, 2012.
31. World Wildlife Fund. *HydroSHEDS (Hydrological Data and Maps Based on SHuttle Elevation Derivatives at Multiple Scales)*; World Wildlife Fund: Gland, Switzerland, 2006.
32. Martz, L.W.; Garbrecht, J. The treatment of flat areas and depressions in automated drainage analysis of raster digital elevation models. *Hydrol. Process.* **1998**, *12*, 843–855. [CrossRef]
33. Yan, D.H.; Wang, K.; Qin, T.L.; Weng, B.S.; Wang, H.; Bi, W.X.; Li, X.N. A Data Set of Global River Networks and Corresponding Water Resources Zones Divisions. 2019. Available online: https://doi.org/10.6084/m9.figshare.8044184.v3 (accessed on 10 May 2019).
34. Rahmati, O.; Pourghasemi, H.R.; Zeinivand, H. Flood susceptibility mapping using frequency ratio and weights-of-evidence models in the Golastan Province, Iran. *Geocarto Int.* **2016**, *31*, 42–70. [CrossRef]

35. Wan, Z.M.; Hong, Y.; Khan, S.; Gourley, J.; Flamig, Z.; Kirschbaum, D.; Tang, G.Q. A cloud-based global flood disaster community cyber-infrastructure: Development and demonstration. *Environ. Model. Softw.* **2014**, *58*, 86–94. [CrossRef]
36. Macchione, F.; Costabile, P.; Costanzo, C.; Lorenzo, G.D. Extracting quantitative data from non-conventional information for the hydraulic reconstruction of past urban flood events. A case study. *J. Hydrol.* **2019**, *576*, 443–465. [CrossRef]
37. Brakenridge, G.R. Global Active Archive of Large Flood Events. Dartmouth Flood Observatory, University of Colorado. Available online: http://floodobservatory.colorado.edu/Archives/index.html (accessed on 7 May 2019).
38. NOAA Earth System Research Laboratory's Physical Sciences Division. The NCEP/NCAR Reanalysis Project. Available online: https://www.esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml (accessed on 10 June 2019).
39. Zhang, L.J.; Qian, Y.F. Annual distribution features of the yearly precipitation in China and their interannual variations. *Acta Metall. Sin.* **2003**, *17*, 146–163.
40. Bui, D.T.; Pradhan, B.; Nampak, H.; Bui, Q.T.; Tran, Q.A.; Nguyen, Q.P. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibility modeling in a high-frequency tropical cyclone area using GIS. *J. Hydrol.* **2016**, *540*, 317–330.
41. Meraj, G.; Romshoo, S.A.; Yousuf, A.R.; Altaf, S.; Altaf, F. Assessing the influence of watershed characteristics on the flood vulnerability of Jhelum basin in Kashmir Himalaya. *Nat. Hazards* **2015**, *77*, 153–175. [CrossRef]
42. Fischer, G.F.; Nachtergaele, S.; Prieler, H.T.; van Velthuizen, L.; Verelst, D. *Wiberg. Global Agro-Ecological Zones Assessment for Agriculture (GAEZ 2008)*; IIASA: Laxenburg, Austria; FAO: Rome, Italy, 2008.
43. Li, Z.; Zhang, J.T. Calculation of Field Manning's Roughness Coefficient. *Agric. Water Manag.* **2001**, *49*, 153–161. [CrossRef]
44. Mohamoud, Y.M. Evaluating Manning's roughness coefficients for tilled soils. *J. Hydrol.* **1992**, *135*, 143–156. [CrossRef]
45. Chen, J.; Ban, Y.; Li, S. China: Open access to Earth land-cover map. *Nature* **2014**, *514*, 434.
46. Zhao, G.; Pang, B.; Xu, Z.X.; Yue, J.J.; Tu, T.B. Mapping flood susceptibility in mountainous areas on a national scale in China. *Sci. Total Environ.* **2018**, *615*, 1133–1142. [CrossRef]
47. The World Bank. Population, Total Database. Available online: https://data.worldbank.org/indicator/SP.POP.TOTL (accessed on 10 June 2019).
48. Yan, D.H.; Weng, B.S.; Qin, T.L.; Wang, H.; Li, X.N.; Yang, Y.H.; Wang, K. A Data Set of Distributed Global Population and Water Withdrawal from 1960 to 2017. Available online: https://figshare.com/s/fc2ca2beccf475e963cf (accessed on 10 June 2019).
49. The World Bank. GDP (Current US $) Database. Available online: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD (accessed on 10 June 2019).
50. Cox, D.R. The regression-analysis of binary sequences. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1958**, *20*, 215–242. [CrossRef]
51. Pradhan, B. Flood susceptible mapping and risk area delineation using logistic regression, GIS and remote sensing. *J. Spat. Hydrol.* **2010**, *9*, 1–18.
52. George, H.J.; Pat, L. Estimating Continuous Distributions in Bayesian Classifiers. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montréal, QC, Canada, 18–20 August 1995; pp. 338–345.
53. Das, I.; Stein, A.; Kerle, N.; Dadhwal, V.K. Landslide susceptibility mapping along road corridors in the Indian Himalayas using Bayesian logistic regression models. *Geomorphology* **2012**, *179*, 116–125. [CrossRef]
54. Yoav, F.; Robert, E.S. Experiments with a new boosting algorithm. In Proceedings of the Thirteen International Conference on Machine Learning, Bari, Italy, 3–6 July 1996; pp. 148–156.
55. Cao, Y.; Miao, Q.G.; Liu, J.C.; Gao, L. Advance and Prospects of AdaBoost Algorithm. *Acta Autom. Sin.* **2013**, *39*, 745–758. [CrossRef]
56. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
57. Sadler, J.M.; Goodall, J.L.; Morsy, M.M.; Spencer, K. Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest. *J. Hydrol.* **2018**, *559*, 43–55. [CrossRef]
58. Bandos, A.I.; Guo, B.; Gur, D. Jackknife variance of the partial area under the empirical receiver operating characteristic curve. Stat. *Methods Med. Res.* **2017**, *26*, 528–541. [CrossRef]

59. Chang, L.C.; Chang, F.J.; Yang, S.N.; Kao, I.F.; Ku, Y.Y.; Kuo, C.L.; Amin, I. Building an Intelligent Hydroinformatics Integration Platform for Regional Flood Inundation Warning Systems. *Water* **2019**, *11*, 9. [CrossRef]
60. Tien Bui, D.; Khosravi, K.; Li, S.; Shahabi, H.; Panahi, M.; Singh, V.; Chapi, K.; Shirzadi, A.; Panahi, S.; Chen, W.; et al. New hybrids of ANFIS with Several Optimization Algorithms for flood Susceptibility Modeling. *Water* **2018**, *10*, 1210. [CrossRef]
61. Chang, L.C.; Amin, M.; Yang, S.N.; Chang, F.J. Building ANN-Based Regional Multi-Step-Ahead Flood Inundation Forecast Models. *Water* **2018**, *10*, 1283. [CrossRef]