


Article

Mapping Landslide Hazard Risk Using Random Forest Algorithm in Guixi, Jiangxi, China

Yang Zhang ¹ , Weicheng Wu ^{1,*} , Yaozu Qin ¹, Ziyu Lin ¹, Guiliang Zhang ², Renxiang Chen ², Yong Song ², Tao Lang ², Xiaoting Zhou ¹, Wenchao Huangfu ¹, Penghui Ou ¹, Lifeng Xie ¹, Xiaolan Huang ¹, Shanling Peng ¹ and Chongjian Shao ¹

¹ Key Laboratory of Digital Lands and Resources and Faculty of Earth Sciences, East China University of Technology, Nanchang 330013, China; 201810818002@ecut.edu.cn (Y.Z.); qyz60010@ecut.edu.cn (Y.Q.); zylin@ecut.edu.cn (Z.L.); 201900818004@ecut.edu.cn (X.Z.); 201810818004@ecut.edu.cn (W.H.); 201810818013@ecut.edu.cn (P.O.); 201810705007@ecut.edu.cn (L.X.); 201810705009@ecut.edu.cn (X.H.); pshanling@ecut.edu.cn (S.P.); scj350936@ecut.edu.cn (C.S.)

² 264 Geological Team of Jiangxi Nuclear Industry, Ganzhou 341000, China; zgl-63@163.com (G.Z.); crxkcy@163.com (R.C.); songy_6611@163.com (Y.S.); ecitlangtao@163.com (T.L.)

* Correspondence: wuwch@ecut.edu.cn or wuwc030903@sina.com

Received: 10 October 2020; Accepted: 13 November 2020; Published: 23 November 2020



Abstract: Landslide hazards affect the security of human life and property. Mapping the spatial distribution of landslide hazard risk is critical for decision-makers to implement disaster prevention measures. This study aimed to predict and zone landslide hazard risk, using Guixi County in eastern Jiangxi, China, as an example. An integrated dataset composed of 21 geo-information layers, including lithology, rainfall, altitude, slope, distances to faults, roads and rivers, and thickness of the weathering crust, was used to achieve the aim. Non-digital layers were digitized and assigned weights based on their landslide propensity. Landslide locations and non-risk zones (flat areas) were both vectorized as polygons and randomly divided into two groups to create a training set (70%) and a validation set (30%). Using this training set, the Random Forests (RF) algorithm, which is known for its accurate prediction, was applied to the integrated dataset for risk modeling. The results were assessed against the validation set. Overall accuracy of 91.23% and Kappa Coefficient of 0.82 were obtained. The calculated probability for each pixel was consequently graded into different zones for risk mapping. Hence, we conclude that landslide risk zoning using the RF algorithm can serve as a pertinent reference for local government in their disaster prevention and early warning measures.

Keywords: Random Forest; landslide hazard risk; integrated multisource dataset; field sample rasterization; weight assignment

1. Introduction

Landslides are a major natural hazard and can be defined as phenomena in which a rock and soil body on a slope slides down a certain interface under the action of gravity, rainfall, and groundwater. Landslides are one of the most frequent geological disasters in China. In 2019, 6181 geohazards were recorded in China, including 4220 landslides, accounting for 68.27% of the total hazards. These geohazards resulted in 211 deaths, 13 missing persons, and 75 injuries, and a direct economic loss of 2.77 billion yuan (China Geological Survey, 2020) [1]. According to the nationwide distribution of landslides in 2019, Jiangxi province ranks number two. The Ministry of Natural Resources of the People's Republic of China announced that 1747 geological hazards occurred in the first half of 2020, with a direct economic loss of 1.01 billion yuan. It was predicted that the situation will remain severe in the second half of the year (<http://www.mnr.gov.cn/>).

Due to complex natural conditions, spatiotemporal differences, and uncertainties of the landslide mechanism, it is difficult to accurately predict the occurrence time, scale, and impact range of landslides. However, based on the existing geohazard research and available technologies, it is possible to conduct effective landslide risk prediction and mapping. This will assist local authorities to take preventative and early warning measures to reduce damage and loss of life and property.

Recently, machine learning approaches have been applied in risk prediction and mapping. The advantage of the machine learning methods lies in its capacity to deal with a large amount of geospatial data within multi-dimensional and even hyper-dimensional space, and in its ability to achieve accurate prediction and classification (Wu et al. 2016, and 2018) [2,3]. These learning algorithms may provide the probability of the spatial occurrence of a landslide and identify the importance of different geo-environmental causal factors that play a potential role in these landslide events [4]. Several machine learning approaches have been utilized for landslide assessment in the past decade, such as Support Vector Machines (SVMs) [5,6], Artificial Neural Networks (ANNs) [7,8], Deep Learning Neural Networks (DLNNs) [9,10], Convolutional Neural Networks (CNNs) [11], Boosted Regression Trees (BRTs) [12,13], and Random Forests (RFs) [13]. A number of case studies show that these algorithms have a good prediction performance [14,15]; in particular, the RF has gained a high reputation for its outperformance in both classification and prediction compared to other approaches. Therefore, we adopted this algorithm for landslide risk mapping in our study.

Preprocessing of different geo-environmental causal factors is essential to geohazard risk prediction and assessment. At present, no standard exists to address this issue. Some methods use each causal factor as a categorical variable, e.g., a range of values from the same variable. However, prediction is not based on the accurate value of each pixel, which may affect the prediction efficiency of the model. When dealing with linear factors such as faults, roads, and rivers, these factors are not processed in a hierarchical manner, but in buffer zones with a different order of propensity weight in terms of proximity. Furthermore, the range of the buffer zones can be too large to be efficient for accurate prediction in space. Rainfall is a fundamental factor triggering landslides [16,17]; nevertheless, few studies included seasonal rainfall in landslide risk assessment.

For the reasons outlined above, the objectives of this research were to identify a relevant digitization approach to quantify the causal factors for landslide risk prediction and zoning using the RF algorithm for Guixi, Jiangxi Province, and to produce a risk map of landslides, and thus provide support and advice for local governments and decision-makers to implement landslide hazard prevention and early warning measures.

2. Materials and Methods

2.1. Study Area

Guixi is located in the northeast of Jiangxi Province, China, in the middle reaches of the Xinjiang River, and is bordered by the Wuyishan Mountains (Mts) on the south. The study area lies between 27°50′53″ N and 28°37′33″ N in latitude and between 116°57′43″ E and 117°28′06″ E in longitude, covering an area of about 2292 km². Topographically, Guixi is generally characterized by high mountains in the south and low hills in the north, cut by the Xinjiang River running west in the central region. The elevation of the study area varies from 20 to 1504 m above sea level (Figure 1). About 65.3% of the study area has a slope gradient <15°, whereas areas with gradients of 15–25°, 25–35°, 35–45° and >45° account for 19.8%, 10.6%, 3.7%, and 0.7%, respectively.

As a part of the subtropical monsoon climatic zone, Guixi receives an annual rainfall of 1789.3 mm with 163 annual rainfall days on average during the period 1958–2017. The rainy season occurs in March to July, with a mean accumulated rainfall of 1227.1 mm, or 68.6% of the annual rainfall (Figure 2). The annual mean temperature is 18.2 °C. Guixi is one of the major forest resource counties in Jiangxi, with a forest cover rate of 56%. The Guixi National Forest Park, situated in the south of the county, is covered with 2929.93 ha of forests, occupying 98.2% of its total area.

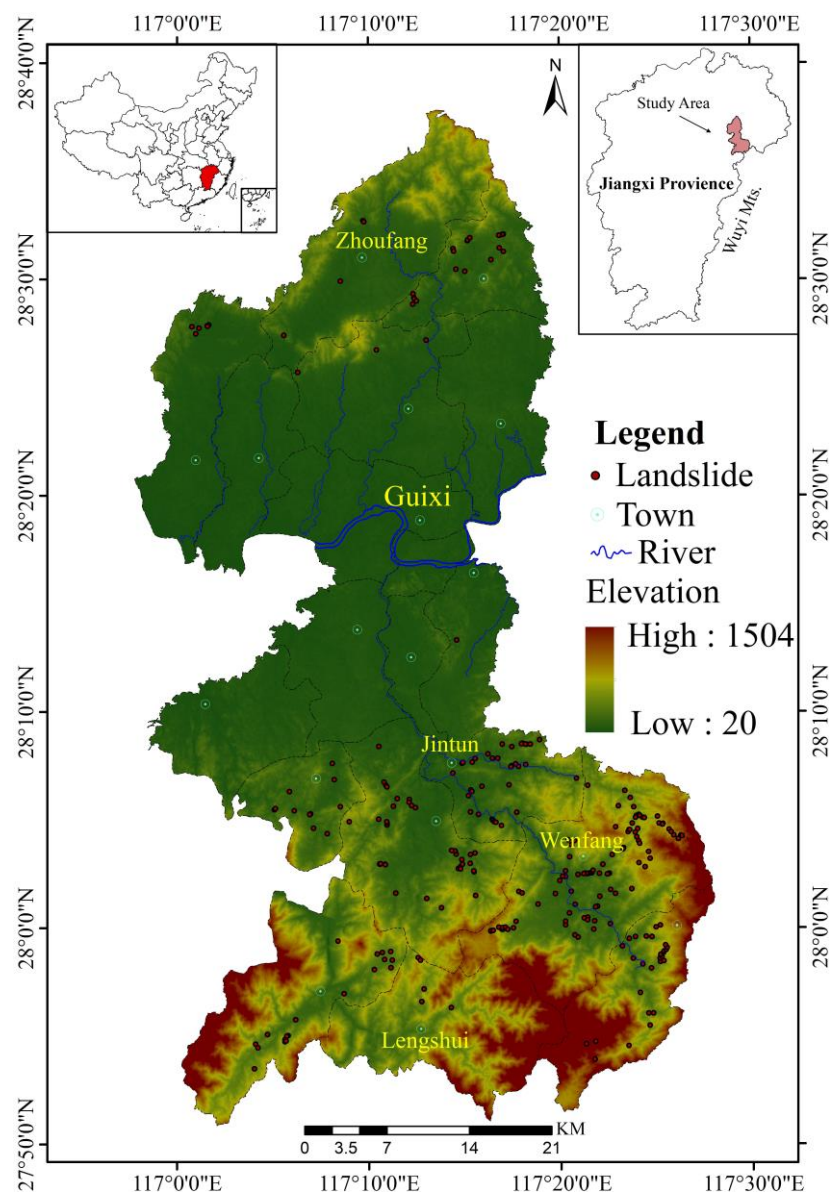


Figure 1. Location of the study area and distribution of the landslides.

Geologically, the formations in the study area include the strata from the Mesoproterozoic to the Cenozoic, with a stratigraphic sequence of Qingbaikou-Cambrian, Carboniferous-Permian, and Triassic-Quaternary. Magmatic rocks are well exposed and mainly distributed in the south, comprising the northern region of the Wuyishan Mts with multiple lithologies, such as basic, acidic, and neutral rocks. In terms of formation time, magmatic activities occurred partly in the Caledonian (around 508–408 ma) and predominantly in the Yanshanian periods (208–65 ma) [18].

According to the historical records, eight earthquakes in total had occurred in Guixi since AD 445, all with a magnitude below 5, and hence, this factor was not considered in our study.

Field investigation revealed that a total of 428 houses, 568 m of highway, 50 m of water channel and 5.1 ha of farmland have been destroyed by landslides in the last ten years in the study area. The damages to social properties were estimated about 3.54 million yuan. However, few efforts have been taken to predict the occurrence of these landslides for disaster reduction purpose.

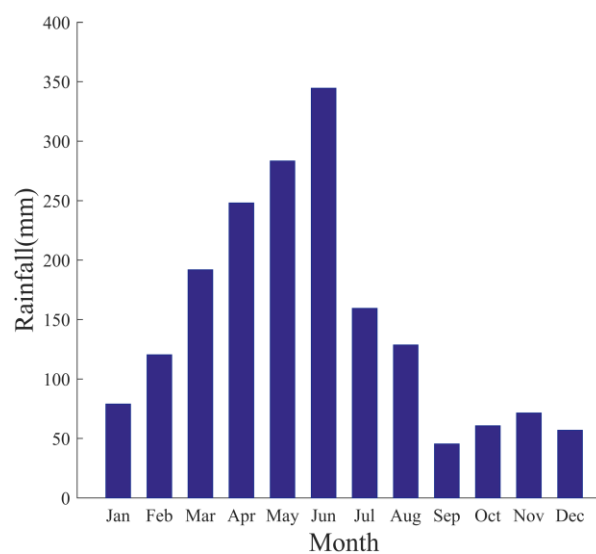


Figure 2. Diagram showing the averaged monthly rainfall in the study area.

2.2. Data and Preparation

2.2.1. Field Data, Training and Validation Sets

The first landslide inventory was conducted in the period September 2014–December 2015 and compiled into the Geological Hazard Survey Report (1:50,000) of Guixi by the 264 Geological Team of Jiangxi Nuclear Industry. The second survey was undertaken by ourselves in July and October 2019, and August 2020.

In this study, a total of 273 landslides that had taken places in the past ten years were identified. They are all small in volume, in which the smallest one is about 5 m³, and the largest one around 20,000 m³ with an average of about 533 m³. As more than 88% of the landslides are less than 900 m³, the landslide sites (points) have been repalced with polygons of 30 m × 30 m in size to facilitate the successive analysis. This field landslide dataset was divided randomly into two groups: training set and validation set, which took up respectively 70% and 30% of the total samples. Against the landslide events, 380 no-risk stable points (defined in the same size of polygons) in lowlands, croplands and urban where slope is < 3° were selected. These no-risk polygons were also separated stochastically into two groups, 70% and 30%, and then incorporated respectively into the training set (191 landslides and 266 non-landslides) and validation set (82 landslides and 114 non-landslides).

As the successive landslide risk mapping was based on a binary classification using RF algorithm, these two classes of samples in both training and validation sets were assigned with a probability value of 1.0 for the occurred landslides, and 0.0 for no-risk samples. And then, these two sets were converted into raster of 30 m size according to the approach proposed by Wu et al. (2018) [3].

2.2.2. Landslide Causal Factors and Integrated Hyper-Dimensional Dataset

Identification, selection and preprocessing of landslide causal factors is a key procedure for risk modeling and zoning. Previous studies have utilized various factors and attempted to reveal their potential roles in landslide events [13]. Based on this and our field knowledge, 21 landslide-related factors such as geological formations, elevation, slope, aspect, plan curvature, profile curvature, thickness of the weathering crust, soil type and texture (clay, sand and silt contents), land use, the normalized difference vegetation index (NDVI), average annual rainfall, March–July rainfall, May–July rainfall, distance to the geological boundaries, distance to faults, distance to roads, and distance to rivers were identified (Table 1). These factors were processed in GIS and all converted to raster with a cell size of 30 m after weight assignment for the non-digital ones.

Table 1. Geo-information layers used as landslide causal factors.

No	Causal Factors	Resolution	Sources
1	Elevation	30 m	GDEM V3 NASA (https://earthdata.nasa.gov/)
2	Slope		DEM-derived
3	Aspect		
4	Plan curvature		
5	Profile curvature		
6	Depth of the weathering crust (soil thickness)	1 km	Kriging interpolation
7	Soil type		Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn/)
8	Soil texture (sand content)		
9	Land use	30 m	Landsat 5 TM
10	NDVI		Landsat 5 TM
11	Average annual rainfall		264 Geological Team of Jiangxi Nuclear Industry
12	March-June rainfall		
13	March-July rainfall		
14	May-July rainfall		
15	June-July rainfall		
16	June-August rainfall	1:50,000	264 Geological Team of Jiangxi Nuclear Industry
17	Lithology		
18	Distance to geological boundaries		
19	Distance to faults	1:5000	Google Earth
20	Distance to roads		
21	Distance to rivers		

Quantification and Weight Assignment

1. Topographic features are critical for landslide hazard risk assessments [19]. A digital elevation model (DEM), ASTER GDEM V003 product of 30 m in resolution, was obtained from the NASA (<https://earthdata.nasa.gov/>) for the study area. This DEM was further used to derive elevation (Figure 1), slope (Figure 3a), aspect, plan curvature and profile curvature.

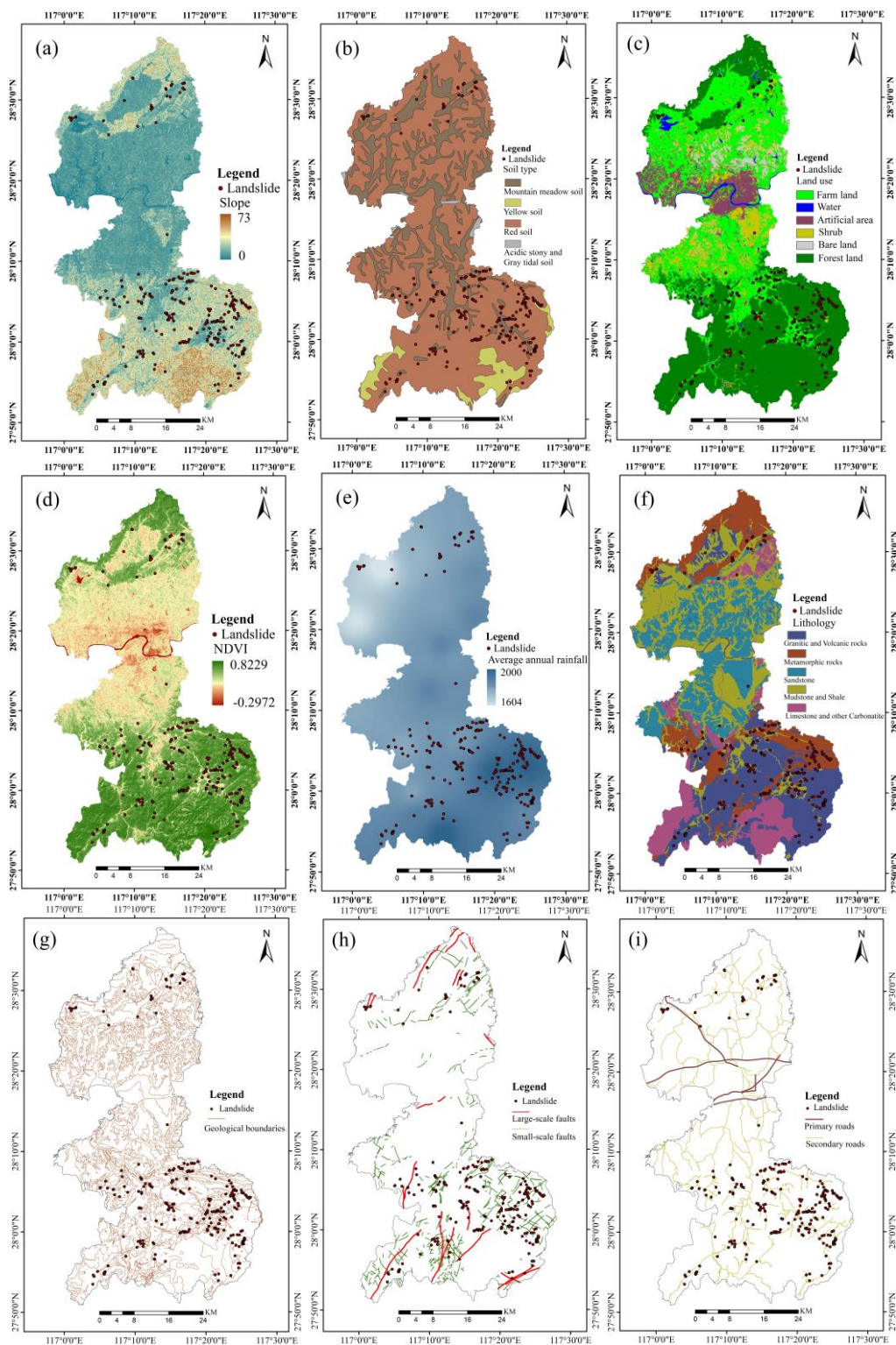


Figure 3. Landslide causal factors used in the study taking the following factors as an example: (a) slope, (b) soil type, (c) land use, (d) NDVI, (e) average annual rainfall, (f) lithology, (g) geological boundaries, (h) faults, (i) roads.

- The weathering crust provides materials and sites for landslides and is the host of the latter, and can be considered as an important controlling factor of landslide event [20–22]. The interaction between this crust and rainfall causes the occurrence of landslides. The survey on the thickness of

the weathering crust in Guixi lacks detailed data except for some landslide profiles. We extracted the ridge and valley lines based on the DEM data and assumed that the lowland plain and valleys had a thick crust of about 10 m, and it decreased as the slope and altitude increased, and at the ridge, it was about 0.5 m. A Kriging interpolation approach was employed to produce the thickness map of the weathering crust.

3. Edaphic factor is also necessary for risk modeling and prediction as it influences the occurrence of landslides [23,24]. Soil type and texture data of the study area were obtained from the Data Center for Resources and Environmental Sciences, CAS (RESDC: <http://www.resdc.cn/>) (Figure 3b). No matter which type of soil, the important feature is the soil texture, i.e., percentage of sands, which influences greatly the soil property. For example, the higher percentage of sands, the higher porosity for rainwater permeation, leading to a higher risk of landslide where clay on the interface may play a role as lubricant. Thus, soil with sand percentage of > 40%, 20–40%, 10–20% and 0–10% were respectively assigned a propensity value of 10, 7, 4, and 1. Then soil map was resampled to pixels with size of 30 m to match the other data.
4. Land use (Figure 3c) is an indicator of human activity that illustrates the relationship between man and environment. The exploitation of land resources has been regarded as an unignorable factor that may affect negatively our environment and the occurrence of landslides [25,26]. The land use map of the study area was produced by using Landsat 5 TM images acquired on May 31 and November 07, 2010, obtained from the Geospatial Data Cloud (<http://www.gscloud.cn/>), using the approaches proposed by Wu et al. (2016) [2,27]. With a mapping accuracy of 91.44%, six land use classes were identified, namely artificial area (urban, rural village and infrastructure), farm land, forests, shrubs, bare land, and water bodies, and were assigned respectively a proneness weight of 0, 0, 1, 4, 10 and 0. Here low slope urban and farmland have lowest proneness, and forest cover has also low propensity while bare land without vegetation protection is the most vulnerable category given the same natural conditions.
5. Vegetation condition and abundance, which can be represented by vegetation index, e.g., the normalized difference vegetation index (NDVI), have been reported of a high correlation with the occurrence of landslides [12]. As a complement to land cover, NDVI (Figure 3d) was selected and included in the analysis of this study. We obtained the late autumn (October 24–November 07, when herbaceous vegetation became withered and most crops were harvested) Landsat 5 TM images of the period 2005–2010, from the same data server as mentioned above. The TM images were atmospherically corrected using the COST model (Chavez 1996; Wu 2003; Wu et al. 2013) [28–30] in which both additive scattering and multiplicative path transmission effects were minimized. NDVI was calculated using the formula $(NIR-R)/(NIR+R)$ [31] from each scene and then averaged to get the multiyear mean NDVI.
6. Rainfall is often considered as a triggering factor of landslide events [32,33]. In this study, the average annual rainfall (Figure 3e), March–June, March–July, May–July, June–July and June–August rainfall were taken into account as hazard-causative factors. Our purpose was to investigate which months' rainfall or combined accumulation is the most important for assessing the landslide risk. The daily rainfall data of the period 2008–2017 from 104 ground stations were acquired and used to create different accumulative monthly rainfall combinations, which were further gridded into raster layers using Inverse Distance Weighted (IDW) interpolation approach.
7. Geological strata, especially, their lithologies (Figure 3f) and bedding, can play different roles in the occurrence of landslides because of their different resistance to weathering and bedding structure, in particular, together with joints and fractures, which may serve as rainfall permeation pathways and slippery interface. The lithological data of the study area were digitized from the Geological Map on a scale of 1/50,000 [18]. The hazard-causative propensity weight of each formation lithology was assigned in terms of its resistance to landslide, e.g., higher resistance formation was assigned with lower weight value or vice versa. More concretely, granitic and volcanic rocks were assigned a weight value of 1, metamorphic rocks 5, sandstone 7, limestone

- and other carbonatite 8, and mudstone and shale 10. The higher value assigned, the higher proneness of the factor tends to contribute to landslides [34].
8. Geological boundaries (Figure 3g) are the connection belts of inhomogeneous geological formations and usually fragile zones that are susceptible to weathering and fracturing. It is thus considered a potential factor influencing the slope stability. Actually, the closer to the boundary the higher risk may exist [9]. The geological boundaries were extracted from the above-mentioned geological map and buffered into different zones as 0–30 m, 30–60 m, 60–90 m and 90–120 m, which were assigned a weight value of 5, 3, 2 and 1, respectively.
 9. Linear features: Faults (Figure 3h) often play an active role in landslide events as they are fractures and subject to water permeation and extensive weathering. This tends to increase the vulnerability of geological bodies and slope instability. Road construction (Figure 3i) is a direct human action on the slope resulting in an instability of the latter. The change in landform and the loss of support from the underlying massif lead to the increase of tension on the upper slope that promotes the development of cracks [4]. Rivers are usually an active factor in modification of landscape by cutting the different geological formations and making their adjacent massif fragile through liquidization. A number of studies revealed that not only rivers but also reservoirs influence the stability of slope [35,36].

In this study, faults, roads and rivers were buffered in line with their scales. For example, small faults, roads and rivers were buffered with distance intervals of 0–30, 30–60, 60–90 and 90–120 m; large-scale faults, main river and reservoirs were buffered with distances of 0–60, 60–120, 120–180 and 180–240 m from the borders. Each buffer zone was assigned a weight in terms of its potential proneness to landslide, e.g., zones 0–30, 30–60, 60–90 and 90–120 m were assigned respectively 10, 7, 4, and 1, and zones 0–60, 60–120, 120–180 and 180–240 m respectively 20, 15, 10, and 5. This assignment was based on the rule that the closer to the linear features, the higher propensity of landslide. These buffers of different linear factors were converted into raster layers of 30 m in cell size.

Integrated Hyper-dimensional Geo-information Set

The above rasterized 21 hazard-causative factors including elevation, slope, aspect, plan curvature, profile curvature, thickness of the weathering crust, soil texture (especially, sand %), land use, NDVI, average annual rainfall, March–July rainfall, May–July rainfall, lithology, distance to faults, distance to roads, and distance to rivers, etc., were stacked together to compose a 21-layer geo-information dataset. Specifically, this is an integrated dataset with 21 dimensions, a realistic hyper-dimensional data space.

2.3. Risk Prediction and Modeling

2.3.1. RF Algorithm

As one of the machine learning approaches, the RF algorithm achieves learning and prediction using an ensemble of growing decision-trees, or rather, of classification and regression trees (CARTs) and their majority voting (Breiman 2001) [37]. One critical technique of this algorithm lies in its bootstrap sampling from the training set to build trees followed with a randomized selection of the input variables to determine the best split for each node. In the meantime, the out-of-bag (OOB) estimates are applied within the RF algorithm to determine the generalization error and the importance of each predictive variable (Breiman, 2001) [37]. Moreover, there shall not be the overfit problem with RF if the number of decision trees (NT) is large enough. In other words, the RF algorithm makes use of the strong law of large numbers, i.e., the more features employed, the less error generated (Breiman 2001; Wu et al. 2018) [3,37]. Thus, NT should be large enough so as to minimize the OOB error of classification or regression to a stable level during the training procedure. Another advantage of the RF algorithm is its capacity to deal with hyper-dimensional data using limited training samples but achieving results of high accuracy. Instead of classification of land cover types, we employed this RF algorithm here to classify the probability of risk and no-risk for each pixel in the whole study area.

2.3.2. Risk Prediction and Modeling

RF modeling was conducted within EnMap-Box, an image processing package developed by DLR (German Aerospace Center) [38]. Using the RF Classification (RFC) function, the integrated 21-layer geo-information dataset was input as predictive variables with the training set for training.

Before risk modeling, a set of parameters have to be set up, for example, NT, the number of randomly selected features at each node, and the stop criteria (for node splitting). How to set up these parameters can be referred to Wu et al. (2018) [3].

Risk modeling was actually a parameterization procedure versus the training set with an internal validation. The generated RF risk model was applied back to the integrated dataset to perform a binary classification of risk probability to derive the probability map.

2.3.3. Verification and Reliability Analysis

To assess the performance of landslide risk modeling, the predicted results were verified against the independent validation set [2,3,39] rather than the training set. Two metrics were used, i.e., overall accuracy (OA) and Kappa Coefficient (KC), which were calculated based on the confusion matrix of the trained landslide risk models versus the validation set. Here KC is a direct indicator of reliability of the risk modeling and prediction [2,3,39,40]. For a value of 0, it indicates a poor consistency between the prediction and the observation, whereas, a value of 1 implies a perfect agreement between the two. The KC-based agreement levels proposed by Landis and Koch [40] were followed: poor (0–0.2), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80) and almost perfect (0.81–1.0).

2.3.4. Assessment of the Importance of Landslide Causal Factors

It is necessary to understand that the role of each hazard-causative factor may differ from one place to another, depending on the assessment model and landslide mechanism in different geo-environments. This implies that a geo-environmental factor may take an active part in landslide prediction in one model in one place but play a tiny role in another elsewhere. Therefore, the contribution of a causative factor is conditional and various. The importance of each factor for the landslide events in this study was evaluated using the OOB ranking procedure of the RF classifier.

3. Results

3.1. NT within the RF Algorithm

The NT affected the predication results when the RF risk modeling was conducted (Table 2). In spite of its capacity to deliver rather accurate prediction when NT was set to 100, the prediction was more robust with higher OA and KC when it was set to 300 and 500. As a confirmation to other authors [2], OA and KC declined slightly when NT was 1000 (Table 2). Hence, 300–500, especially, 300 would be advised to use for NT when tackling landslide risk prediction and zoning.

Table 2. Performance of the RF algorithm with different Number of Trees (NT).

Number of Trees	Overall Accuracy	Kappa Coefficient
100	90.75	81.08
300	91.23	82.02
500	91.07	81.70
1000	90.75	81.04

3.2. Landslide Hazard Risk Map

As shown above, the RF algorithm performed best when NT was set to 300, and thus the modeling results of this case were selected for landslide hazard risk mapping. The computed risk probability, ranging from 0 to 1.0 in each pixel, was classified into five levels, i.e., No risk (0–0.2), Low risk (0.2–0.4),

Median risk (0.4–0.6), High risk (0.6–0.8), and Extremely high risk (0.8–1.0). Thence, the landslide risk zonation map was produced and presented in Figure 4.

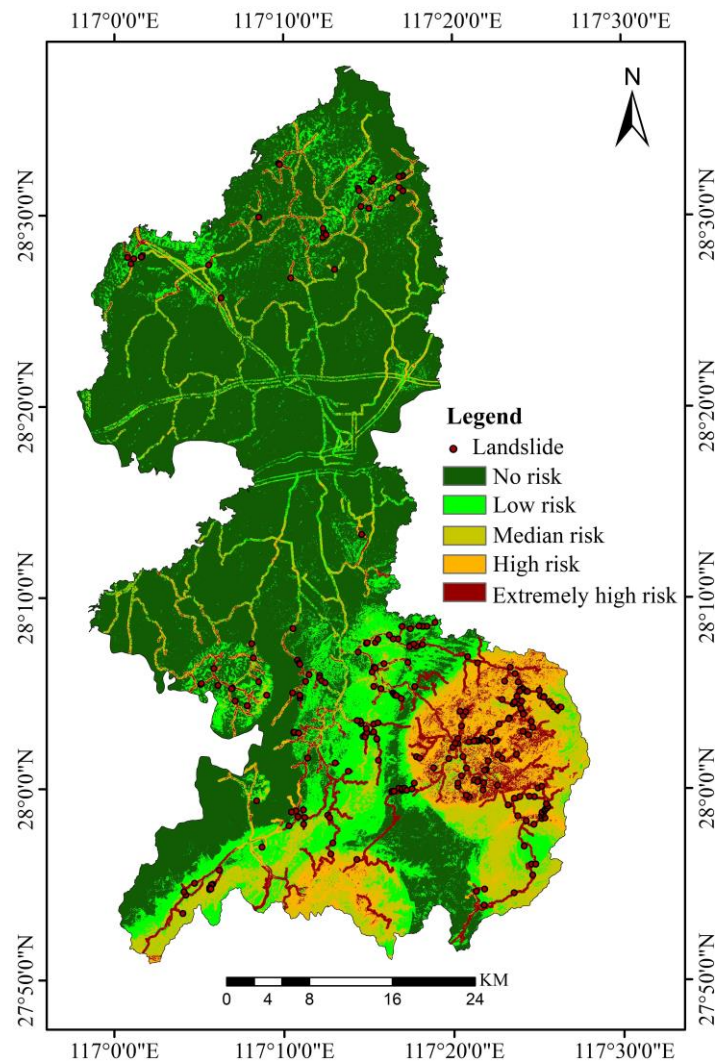


Figure 4. Landslide risk map.

This map shows that the landslide-prone areas are mainly distributed along the roads and in the high slope hilly and mountainous areas in the north and south of Guixi where the predicted High risk and Extremely high risk zones are distributed. There are also Median and High risk zones in the southwest where there is abundant rainfall. Nevertheless, the risk is relatively low in the central part, a plain with gentle topographic relief in Guixi.

It is seen in Figure 4 that two types of landslide risks were predicted, i.e., one is man-made landslides distributed along the roads or cut slopes as a consequence of road construction and housing development, and the other is natural ones distributed in the mountainous slopes in the south and southeast of the study area (Figure 4).

For the modeling result obtained when NT was 300, the OA of this risk map is 91.23%, and KC 0.82 versus the validation set, reaching the “almost perfect (0.81–1.0)” level. As statistics revealed, the number of the observed landslides falling in the zones No risk, Low, Medium, High and Extremely high risk, accounted for, respectively, 0.73%, 1.10%, 2.93%, 2.57% and 92.67% of the total.

3.3. Importance of the Hazard-Causative Factors

As seen in Figure 5, all 21 hazard-causative factors have contributed to the landslide events but the first five, i.e., distance to road, slope, May–July rainfall, average annual rainfall and elevation, comprise 65.45% of the total contribution to the occurrence of landslide disasters. That is to say, they have played a more important role in landslide events than other factors. The contribution of soil type and faults is relatively low.

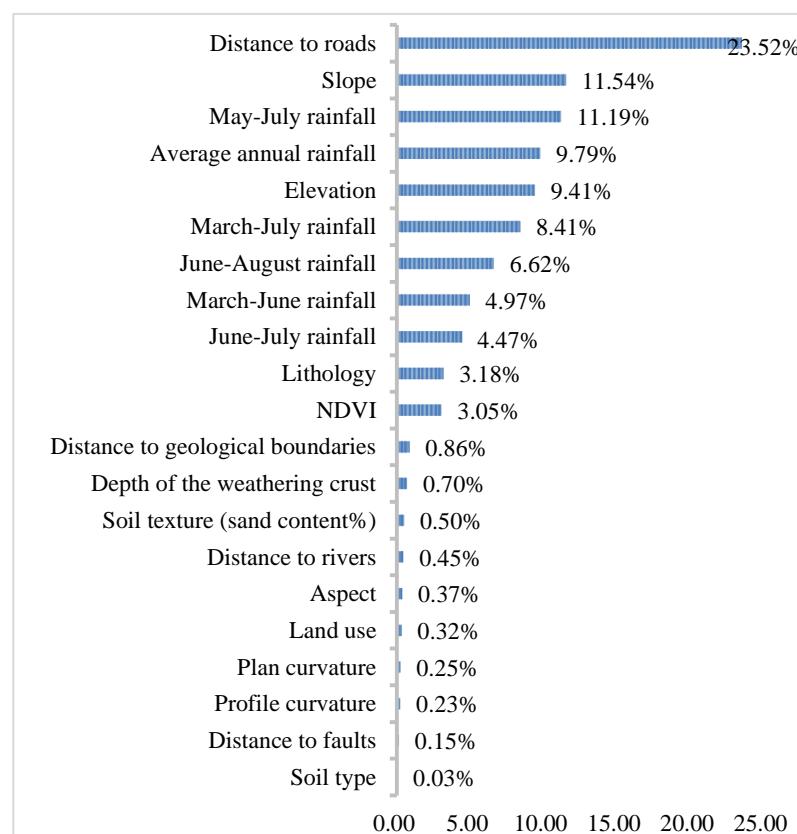


Figure 5. Importance of hazard-causative factors.

Field survey revealed that these landslides were neither triggered by earthquake nor by active tectonic movements but by human activity, e.g., road construction and housing development, and rainfall. Actually, anthropogenic landslides accounted for 98.9%, where the slope ranges from 8° to 25°, much lower than the threshold, 28–35°, of the natural landslides as proposed by Fan et al. (2016) [23]. In addition, 209 landslides, 76.3% of the total, occurred in the talus accumulation with a thickness of about 0.5–10 m. The occurrence time of landslides is mainly in March–July, in particular, in June–July. The number of landslides of these two months accounts for more than 50% of the total.

4. Discussion

4.1. Algorithm for Landslide Risk Assessment

As previously mentioned, a number of data-driven approaches have been applied for landslide risk prediction. Xiong et al. (2020) [13] noted that among the machine learning algorithms, BRTs performed best in debris flow susceptibility assessment in Sichuan Province whereas Chen et al. [15] concluded that RF achieved the best prediction in Chongren, Jiangxi. Actually, one of our parallel research (Zhou et al. under review [34]) conducted in Ruijin, Jiangxi and that of Sun et al. (2020) [19] in Fengjie, Chongqing, both pointed out that RF is capable for providing accurate landslide risk prediction. This study, using

the RF algorithm to fulfill the task with a high satisfactory level, “almost perfect”, confirmed their conclusion. However, care has to be taken while employing different geo-environmental data for RF-based modeling as the landslides used as training samples were mostly small in scale, i.e., less than one Landsat pixel in surface area. It is hence necessary to use high resolution data to highlight such disaster risk while modeling and mapping are conducted, and data with resolution of coarser than 30 m will not be recommended.

4.2. The Different Importance of the Causal Factors

As revealed in Figure 5, distance to roads, slope, rainfall and elevation are the most important factors in landslide events in Guixi. The order of importance of the geo-environmental factors may be different from one site to another, e.g., slope, rock type, distance to river and NDVI [5], slope and distance to roads [6], lithological formation, distance from roads, and NDVI [12], and elevation and annual rainfall [19]. But all these studies point at a fact, that is, slope, distance to road, elevation and rainfall are the commonly important factors causing landslide hazards, and that is what we have uncovered in this study.

Since the road construction constitutes the most active human factor leading to such geohazard, it shall be necessary to design the road system by avoiding the most risky area and taking the geological strata bedding and slope stability into account.

It is worth mentioning that the importance ranking of all factors related to rainfall accounts for 45.45%, which shows a clear relationship of rainfall with the landslide events, especially, the accumulated rainfall of May–July (Figure 5). However, this importance weight seems still underestimated. Theoretically, rainfall is the triggering factor of the most landslides and should have more importance. The exploration on this topic seems not possible until we have grasped the exact occurrence time of these landslides. Only with such information, can we decide how to combine more reasonably the daily or monthly rainfall for risk modeling.

Some factors, such as faults, edaphic features and geological boundaries, used to be considered as necessary. Nevertheless, the factor importance analysis revealed that they were not as significant as expected in this study. Hence, it is possible to optimize the selection of the causal factors in landslide risk modeling and mapping in the similar geo-environment, in particular, when computing capacity is low.

4.3. Landslide Types

In terms of our field survey, the majority of landslides observed is small in volume, provoked by concentrated rainfall superimposed on the road construction and slope cutting for housing development. Rainfall is able to infiltrate into subsurface along the fractures to reach and liquidize the sliding interface, causing landslides.

There exist relatively big and deep landslides with a volume of about 20,000 m³ but driven by different occurrence mechanisms: (1) landslide occurs after the action of the accumulated rainfall, especially, when Quaternary sediment (talus) has a clear interface with the underlying rocks in which the unconformity serves as slide surface after infiltration of rainfall; (2) multicycle landslides at the same place, they begin with small volumes of slide after rainfall but little by little extending out and deepening after the repeated rainfall events, and finally these slides become a big one; (3) big landslide within downhill strata bedding, which does not take place in the talus or weathered crust but inside the geological strata after the bedding surface has been lubricated by the penetrated rainfall when there are faults and joints. The rapidity of the big landslide relies on the dip of the strata bedding. For high dip bedding, rocky landslide may happen quickly as long as rock mass gravity exceeds the resistant friction of the underlying formation. For low dip bedding, the overlying strata and weathered crust do not constitute a rapid slide but a creep moving downward gradually. When there is a slope cutting, the downward movement becomes faster. This was also clearly observed in Ruijin, another city in southern Jiangxi, constituting a threat to the newly established Longzhu Temple and the No 6

Middle School of Ruijin [37]. It is hence essential to take measure to prevent the huge loss and damage before such big hazard happens.

5. Conclusions

This study made use of the RF algorithm for landslide risk modeling, prediction and mapping in Guixi with an integrated 21 geo-environmental factors and field data. During the modeling, we employed machine learning technique to determine which hazard-prone factors are the most important in provoking landslides. We also demonstrated the procedure on how to digitize non-digital geo-environmental factors and assign a weight value in terms of their proximity or propensity so that quantitative analysis and modeling were made possible. This study provides not only key information on how to research landslide mechanism but also operational approach on how to investigate hazard risk to prevent our society from further damage. In particular, our risk zone map with high reliability may serve as a reference for the local governments and decision-makers of Guixi to implement landslide prevention and early warning measures in the landslide-prone areas.

One key finding of this research is the surprising importance of road construction and housing development in landslide events. This reveals the role of human activity in provoking such geo-disaster, and suggests that when designing road systems, more comprehensive slope protection measures and more profound geological investigation on downhill strata bedding should be necessary so that man-made landslides can be minimized.

Author Contributions: Conceptualization, Yang Zhang, Weicheng Wu, Yaozu Qin and Ziyu Lin; methodology, Weicheng Wu and Yang Zhang; software, Weicheng Wu; validation, Yang Zhang, Yaozu Qin, Wenchao Huangfu, and Penghui Ou; formal analysis, Yang Zhang, Xiaoting Zhou, Wenchao Huangfu, and Penghui Ou; investigation, Yang Zhang, Weicheng Wu, Shanling Peng, Chongjian Shao, Penghui Ou, Wenchao Huangfu, Xiaoting Zhou, Lifeng Xie and Xiaolan Huang; resources, Weicheng Wu, Guiliang Zhang, Guiliang Zhang, Yong Song, Zhiling Wang and Tao Lang; data curation, Yang Zhang, Xiaolan Huang, Xiaoting Zhou Renxiang Chen and Penghui Ou; writing—original draft preparation, Yang Zhang and Weicheng Wu; writing—review and editing, Weicheng Wu; visualization, Wenchao Huangfu, Lifeng Xie and Xiaolan Huang; project administration, Weicheng Wu and Guiliang Zhang; funding acquisition, Weicheng Wu. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the Jiangxi Talent Program (Grant No: 900/2120800004) and the Start-up Fund for Scientific Research of the East China University of Technology (Grant No. DHTP2018001) which were both granted to Weicheng Wu.

Acknowledgments: A part of the first-hand landslide observation data and the Geological Map of Guixi on the Scale of 1/50,000 were provided by the 264 Geological Team of Jiangxi Nuclear Industry.

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

References

1. China Geological Survey. Notification on National Geological Hazard in 2019. Available online: http://www.cgs.gov.cn/gzdt/zsdw/202003/t20200331_504559.html (accessed on 15 August 2020).
2. Wu, W.; Zucca, C.; Karam, F.; Liu, G. Enhancing the performance of regional land cover mapping. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 422–432. [CrossRef]
3. Wu, W.; Zucca, C.; Muhaimeed, A.S.; Al-Shafie, W.M.; Al-Quraishi, A.M.F.; Vinay, N.; Zhu, M.; Liu, G. Soil salinity prediction and mapping by machine learning regression in Central Mesopotamia, Iraq. *Land Degrad. Dev.* **2018**, *29*, 4005–4014. [CrossRef]
4. Zhao, Y.; Wang, R.; Jiang, Y.; Liu, H.; Wei, Z. GIS-based logistic regression for rainfall-induced landslide susceptibility mapping under different grid sizes in Yueqing, Southeastern China. *Eng. Geol.* **2019**, *259*, 105147. [CrossRef]
5. Huang, F.; Cao, Z.; Guo, J.; Jiang, S.-H.; Li, S.; Guo, Z. Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping. *Catena* **2020**, *191*, 104580. [CrossRef]

6. Hong, H.; Pradhan, B.; Xu, C.; Bui, D. Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. *Catena* **2015**, *133*, 266–281. [CrossRef]
7. Tan, Q.; Huang, Y.; Hu, J.; Zhou, P.; Hu, J. Application of artificial neural network model based on GIS in geological hazard zoning. *Neural Comput. Appl.* **2020**, *1*, 1–12. [CrossRef]
8. Wang, Y.; Feng, L.; Li, S.; Ren, F.; Du, Q. A hybrid model considering spatial heterogeneity for landslide susceptibility mapping in Zhejiang Province, China. *Catena* **2020**, *188*, 104425. [CrossRef]
9. Bui, D.T.; Tsangaratos, P.; Nguyen, V.T.; Liem, N.V.; Trinh, P.T. Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. *Catena* **2020**, *188*, 104426. [CrossRef]
10. Hua, Y.; Wang, X.; Li, Y.; Xu, P.; Xia, W. Dynamic development of landslide susceptibility based on slope unit and deep neural networks. *Landslides* **2020**, *1*, 1–22. [CrossRef]
11. Fang, Z.; Wang, Y.; Ling, P.; Hong, H. Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping. *Comput. Geosciences* **2020**, *139*, 104470. [CrossRef]
12. Reza, P.H.; Aiding, K.; Norman, K.; Farzin, S. Investigating the effects of different landslide positioning techniques, landslide partitioning approaches, and presence-absence balances on landslide susceptibility mapping. *Catena* **2020**, *187*, 104364. [CrossRef]
13. Xiong, K.; Adhikari, B.R.; Stamatopoulos, C.A.; Zhan, Y.; Wu, S.; Dong, Z.; Di, B. Comparison of Different Machine Learning Methods for Debris Flow Susceptibility Mapping: A Case Study in the Sichuan Province, China. *Remote Sens.* **2020**, *12*, 295. [CrossRef]
14. Wu, X.; Ren, F.; Niu, R. Landslide susceptibility assessment using object mapping units, decision tree, and support vector machine models in the Three Gorges of China. *Environ. Earth Sci.* **2014**, *71*, 4725–4738. [CrossRef]
15. Chen, W.; Xie, X.; Peng, J.; Wang, J.; Duan, Z.; Hong, H. GIS-based landslide susceptibility modelling: A comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models. *Geomat. Nat. Hazards Risk* **2017**, *8*, 950–973. [CrossRef]
16. Segoni, S.; Piciullo, L.; Gariano, S.L. A review of the recent literature on rainfall thresholds for landslide occurrence. *Landslides* **2018**, *15*, 1483–1501. [CrossRef]
17. Huang, R.; Li, W. Formation, distribution and risk control of landslides in China. *J. Rock Mech. Geotech. Eng.* **2011**, *3*, 97–116. [CrossRef]
18. 264 Brigade of the Jiangxi Nuclear Industry Geological Bureau. The Guixi Geological Hazard Survey Project Implemented by Our Team Successfully Passed the Field Acceptance by the Expert Group. Available online: <http://www.hgy264.com/show-27-6127-1.html> (accessed on 27 October 2020).
19. Sun, D.; Wen, H.; Wang, D.; Xu, J. A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. *Geomorphology* **2020**, *362*, 107201. [CrossRef]
20. Hung, L.Q.; Van, N.T.H.; Duc, D.M.; Ha, L.T.C.; Son, P.V.; Khanh, N.H.; Binh, L.T. Landslide susceptibility mapping by combining the analytical hierarchy process and weighted linear combination methods: A case study in the upper Lo River catchment (Vietnam). *Landslides* **2016**, *13*, 1285–1301. [CrossRef]
21. Xi, C. On the red weathering crusts of southern China. *Quaternary Sciences. Quat. Sci. (Chin. Engl. Abstr.)* **1991**, *1*, 1–8.
22. Zhu, X. Red clay and red weathered crust in southern China. *Res. Soil Water Conserv. (Chin. Engl. Abstr.)* **1995**, *4*, 94–101.
23. Fan, L.; Lehmann, P.; Or, D. Effects of soil spatial variability at the hillslope and catchment scales on characteristics of rainfall-induced landslides. *Water Resour. Res.* **2016**, *52*, 1781–1799. [CrossRef]
24. Kitutu, M.G.; Muwanga, A.; Poesen, J.; Deckers, J.A. Influence of soil properties on landslide occurrences in Bududa district, Eastern Uganda. *Afr. J. Agric. Res.* **2009**, *4*, 611–620. [CrossRef]
25. Hong, H.; Liu, J.; Zhu, A.-X. Modeling landslide susceptibility using LogitBoost alternating decision trees and forest by penalizing attributes with the bagging ensemble. *Sci. Total Environ.* **2020**, *718*, 137231. [CrossRef] [PubMed]
26. Napoli, M.D.; Carotenuto, F.; Cevasco, A.; Confuorto, P.; Martire, D.D.; Firpo, M.; Pepe, G.; Raso, E.; Calcaterra, D. Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability. *Landslides* **2020**, *17*, 1897–1914. [CrossRef]

27. Wu, W.; Mhaimeed, A.S.; Al-Shafie, W.M.; Al-Quraishi, A.M.F. Using Radar and Optical Data for Soil Salinity Modeling and Mapping in Central Iraq. In *Environmental Remote Sensing in Iraq*; Fadhil, A.M., Negm, A., Eds.; Springer: Cham, Switzerland, 2019; Chapter 2; pp. 19–40. [\[CrossRef\]](#)
28. Chavez, P.S. Image-Based Atmospheric Corrections-Revisited and Improved. *Photogramm. Eng. Remote Sens.* **1996**, *62*, 1025–1036.
29. Wu, W. Application de la Geomatique au Suivi de la Dynamique Environnementale en Zones Arides. Ph.D. Thesis, Université Paris 1, Paris, France, 2003.
30. Wu, W.; De Pauw, E.; Zucca, C. Use remote sensing to assess impacts of land management policies in the Ordos rangelands in China. *Int. J. Digit. Earth* **2013**, *6* (Suppl. 2), 81–102. [\[CrossRef\]](#)
31. Wu, W. The generalized difference vegetation index (GDVI) for dryland characterization. *Remote Sens.* **2014**, *6*, 1211–1233. [\[CrossRef\]](#)
32. Bordon, M.; Galanti, Y.; Bartelletti, C.; Persichillo, M.G.; Barsanti, M.; Giannecchini, R.; Avanzi, G.D.A.; Cevasco, A.; Brandolini, P.; Galve, J.P. The influence of the inventory on the determination of the rainfall-induced shallow landslides susceptibility using generalized additive models. *Catena* **2020**, *193*, 104630. [\[CrossRef\]](#)
33. Chikalamo, E.E.; Mavrouli, O.C.; Ettema, J.; Westen, C.J.V.; Mustofa, A. Satellite-derived rainfall thresholds for landslide early warning in Bogowonto Catchment, Central Java, Indonesia. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *89*, 102093. [\[CrossRef\]](#)
34. Zhou, X.; Wu, W.; Lin, Z.; Zhang, G.; Chen, R.; Song, Y.; Wang, Z.; Lang, T.; Qin, Y.; Ou, P.; et al. Landslide risk zoning in Ruijin, Jiangxi, China. *Nat. Hazards Earth Syst. Sci. Discuss* **2020**, in press. [\[CrossRef\]](#)
35. Luo, X.; Li, J. Analysis on the Influence of Reservoir Impoundment on the Bank Landslide. *Des. Hydroelectr. Power Stn. (Chin. Engl. Abstr.)* **2003**, *3*, 61–64.
36. Wang, M.; Yan, E. Study on influence of reservoir water impounding on reservoir landslide. *Rock Soil Mech. (Chin. Engl. Abstr.)* **2007**, *12*, 2722–2725. [\[CrossRef\]](#)
37. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
38. Waske, B.; van der Linden, S.; Oldenburg, C.; Jakimow, B.; Rabe, A.; Hostert, P. imageRF—A user-oriented implementation for remote sensing image analysis with Random Forests. *Environ. Model. Softw.* **2012**, *35*, 192–193. [\[CrossRef\]](#)
39. Jakimow, B.; Oldenburg, C.; Rabe, A. *Manual for Application: ImageRF (1.1)*; Universität Bonn and Humboldt Universität zu Berlin: Berlin, Germany, 2012.
40. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [\[CrossRef\]](#) [\[PubMed\]](#)

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).