International Journal of Disaster Risk Science (2023) 14:636–648
https://doi.org/10.1007/s13753-023-00489-8

www.ijdrs.com
www.springer.com/13753

ARTICLE

A Heterogeneous Sampling Strategy to Model Earthquake-Triggered Landslides

Hui Yang[1] · Peijun Shi[2,3,4] · Duncan Quincey[5] · Wenwen Qi[6] · Wentao Yang[1,5,7]

Accepted: 2 April 2023 / Published online: 22 May 2023

Abstract

Regional modeling of landslide hazards is an essential tool for the assessment and management of risk in mountain environments. Previous studies that have focused on modeling earthquake-triggered landslides report high prediction accuracies. However, it is common to use a validation strategy with an equal number of landslide and non-landslide samples, scattered homogeneously across the study area. Consequently, there are overestimations in the epicenter area, and the spatial pattern of modeled locations does not agree well with real events. In order to improve landslide hazard mapping, we proposed a spatially heterogeneous non-landslide sampling strategy by considering local ratios of landslide to non-landslide area. Coseismic landslides triggered by the 2008 Wenchuan Earthquake on the eastern Tibetan Plateau were used as an example. To assess the performance of the new strategy, we trained two random forest models that shared the same hyperparameters. The first was trained using samples from the new heterogeneous strategy, and the second used the traditional approach. In each case the spatial match between modeled and measured (interpreted) landslides was examined by scatterplot, with a 2 km-by-2 km fishnet. Although the traditional approach achieved higher AUC_{ROC} (0.95) accuracy than the proposed one (0.85), the coefficient of determination (R^2) for the new strategy (0.88) was much higher than for the traditional strategy (0.55). Our results indicate that the proposed strategy outperforms the traditional one when comparing against landslide inventory data. Our work demonstrates that higher prediction accuracies in landslide hazard modeling may be deceptive, and validation of the modeled spatial pattern should be prioritized. The proposed method may also be used to improve the mapping of precipitation-induced landslides. Application of the proposed strategy could benefit precise assessment of landslide risks in mountain environments.

Keywords Earthquake-triggered landslides · Landslide hazard modeling · Machine learning · Model validation · Sampling strategy · Tibetan Plateau

✉ Wentao Yang
yang_wentao@bjfu.edu.cn

[1] School of Soil and Water Conservation, Beijing Forestry University, Beijing 100083, China

[2] State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China

[3] Academy of Disaster Reduction and Emergency Management, Ministry of Emergency Management and Ministry of Education, Beijing Normal University, Beijing 100875, China

[4] Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

[5] School of Geography, University of Leeds, Leeds LS2 9JT, UK

[6] National Institute of Natural Hazards, Beijing 100085, China

[7] Academy of Plateau Science and Sustainability, People's Government of Qinghai Province and Beijing Normal University, Xining 810016, China

1 Introduction

Landslides are major mountain hazards that have been threatening mountain communities around the globe (Petley 2012; Kirschbaum et al. 2015). Landslides induced by an earthquake event could severely increase total casualties compared to earthquakes alone (Yin et al. 2009). Mapping regional landslide hazards is indispensable and paramount for managing landslide risks. Predicting where landslides will occur has been very challenging because landslides have been recognized as stochastic processes (Larsen and Montgomery 2012; Emberson et al. 2016). Although countless efforts have been made on selections of optimal features and models, few practical results have been produced in earthquake-triggered landslide modeling. Regional landslide hazard results are much too general to be used for regional landslide risk analysis.

Landslide hazard mapping should show the temporal and spatial probability of a landslide of a given magnitude. Landslide hazard characteristics include landslide velocity, volume, depth, and other characteristics that may determine the potential of its hazardous influence on the exposed elements, such as people, roads, buildings, and so on (Fell et al. 2008). By considering triggering factor intensity (intensity of seismicity or precipitation), landslide hazard mapping is different from susceptibility mapping by bearing temporal probabilities (van Westen et al. 2008). For earthquake-triggered landslides, temporal probability is determined by the recurrence time of a given seismic intensity. As there is high uncertainty in estimating landslide magnitudes, major efforts have been made to examine the spatial probability of landslides.

Machine learning models are a type of mainstream methods to assess earthquake-triggered landslide hazards (Chen et al. 2018; Qi et al. 2021). Most landslide hazard assessment models follow the logic of using feature layers that include landslide influencing and triggering factors as independent variables to predict landslide hazards (as the dependent variable). Influencing factors are variables that affect the susceptibility to landslides. Land cover type, roads, elevation, slope, aspect, lithology, fault density, and annual precipitation are some of the most frequently used environmental factors (Shu et al. 2019; Tanyas et al. 2022). Urban areas, for example, are more prone to precipitation-induced landslides (Johnston et al. 2021). In landslide hazard models, the intensity of triggers with probability of exceedance of a given time period is also considered as an input layer. Different feature layers have been designed and tested in landslide hazard and susceptibility modeling. Elevation, slope, aspect, land cover type, and lithology are the most frequently used layers in many landslide hazard and susceptibility models.

Efforts on regional landslide hazard mapping have also been focused on the selection of different machine learning models. Most existing machine learning algorithms have been tried in regional landslide hazard modeling. Logistic regression models, support vector machine, artificial neural networks, and random forest models are among the most frequently used machine learning models in regional landslide hazard mapping (Xu et al. 2012). In recent years, deep learning models have become more popular and have also been used in landslide mapping (Wang et al. 2020). Because inputs for these models are tabular data, the advantage of deep learning in automatic feature extraction is limited in landslide hazard modeling. Thus, improvement of landslide hazard mapping by using the best machine learning models is limited.

Positive (landslide) and negative (non-landslide) samples are indispensable data inputs for machine learning models. Dividing landslide/non-landslide samples into training and validating subsets is a commonly used way to evaluate the performance of landslide hazard models. Most hazard modeling results are tested by using the accuracy of the area under the receiver operating characteristic (ROC) curve (Wang et al. 2020; He et al. 2021), balanced accuracy (Nowicki Jessee et al. 2018), or other statistical goodness-of-fit metrics (Nowicki et al. 2014; He et al. 2021). These validation criteria are good indicators to show the robustness of machine learning models, yet they have revealed little on the validity of the mapping results. Direct comparisons between modeled landslide hazards and truly distributed landslides have seldom been mentioned. Almost all maps from existing earthquake-triggered landslide hazard studies show exaggeratedly high hazard values in epicentral areas and unanimously low hazard values in areas far from the epicenters, which is of limited use for decision makers (Allstadt et al. 2018).

Sampling of landslide/non-landslide points in space could affect the mapping results of the spatial pattern of landslide hazards (Tanyas et al. 2019; Pokharel et al. 2021). In general, sampling of the factors that influence landslides occurs within landslide-polygons, with subtle difference between selecting either their geometry feature points (Lombardo et al. 2019; He et al. 2021) or points that have the highest elevation (Jones et al. 2021). A commonly used non-landslide sampling strategy is to select the same number of randomly distributed non-landslide points. There are some studies that have tried to randomly select distributed non-landslide points with numbers proportional to the areal ratio of non-landslide areas in their study area. Because earthquake-triggered landslides are spatially heterogeneous and most are located in the epicenter area, randomly distributed non-landslide sampling points would underrepresent stable slopes in the epicenter area and overrepresent stable slopes further away. These spatially biased sampling strategies are a major reason for the inaccuracy in traditional landslide hazard mapping. Some existing studies have addressed the issue that the quantity of non-landslide samples could affect the accuracy of the landslide modeling results (Shao et al. 2020; Liu et al. 2021; Yang et al. 2022). However, it remains a challenge to select landslide/non-landslide samples to achieve a reliable hazard identification result.

This study aimed to propose a new non-landslide sampling strategy by considering the spatial heterogeneous distribution of earthquake-triggered landslides. Samples produced by the new strategy were used in a random forest model to map landslide hazards for the 2008 Wenchuan Earthquake. The landslide hazard map was further compared with results produced by using the traditional landslide/non-landslide balanced sampling strategy.

2 Data

In this study, we selected the mountainous part of the 2008 Wenchuan Earthquake-affected areas as the study area (Fig. 1), located in the northeastern part of the Tibetan Plateau, bordering the Sichuan Basin to the east. We selected this region for four reasons: (1) The Wenchuan Earthquake triggered the highest known number of earthquake-triggered landslides in the world. Xu et al. (2014) interpreted approximately 200,000 coseismic landslides spread over an area of about 110,000 km^2, and the total area of these coseismic landslides is greater than 1,000 km^2 (Cui et al. 2012); (2) The Mw 7.8 earthquake caused massive casualties (about 100,000), a great number of which (about 20,000) were attributed to earthquake-triggered landslides (Yin et al. 2009); (3) Various sizes of landslides from a few square meters to mega-landslides as large as 10 km^2 (Huang and Fan 2013); and (4) Our selected study area is very large, which ensures that it has various climates, lithology, and landforms to test the models' performances. These characteristics make this study area an ideal place to model earthquake-triggered landslides.

We selected eight layers to perform the landslide hazard analysis (Table 1). We used the 30-m resolution AW3D30 digital elevation model (DEM). Slope and aspect data are derivatives of the DEM, calculated with the ArcGIS software. The 1:2,500,000 geological data were produced by Ye et al. (2017) and released by the China Geological Survey. Lithology and faults in the geological data were used in this study. Land cover type was produced with the Copernicus Global Land Service, which was derived from the European Space Agency's (ESA) Sentinel-2 satellite images. The spatial resolution of the land cover data is 100 m. Peak ground acceleration (PGA) of the 2008 Wenchuan Earthquake was downloaded from the United States Geological Survey (USGS) earthquake catalogue. Annual precipitation data were obtained from the U.S. National Aeronautics and Space

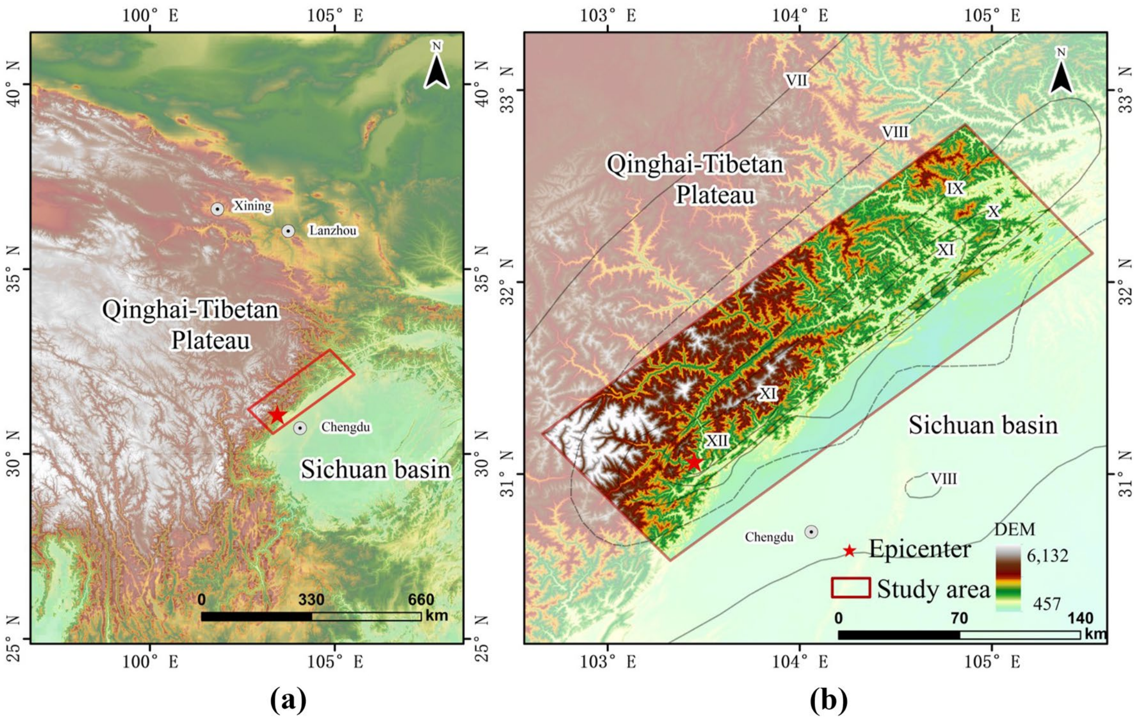

Fig. 1 The study area on the eastern Tibetan Plateau (**a**) covering a major part of the earthquake-affected area with MMI VIII or greater intensity (**b**) *DEM*: Digital elevation model

Table 1 Layers used for the landslide hazard modeling in the 2008 Wenchuan Earthquake study area on the eastern Tibetan Plateau

	Data	Provider	Spatial resolution	Source
1	DEM	AW3D30, JAXA	30 m	https://www.eorc.jaxa.jp/ALOS/en/dataset/aw3d30/aw3d30_e.htm
2	Slope	AW3D30 DEM	30 m	Derived from the AW3D30 DEM
3	Aspect	AW3D30 DEM	30 m	Derived from the AW3D30 DEM
4	Land cover	Copernicus	100 m	https://land.copernicus.eu/global/products/lc
5	Peak ground acceleration	USGS	30 m	https://earthquake.usgs.gov/earthquakes/search/
6	Annual precipitation	NASA GPM	0.1°	https://gpm.nasa.gov/
7	Lithology	Ye et al. (2017)	1:2,500,000	https://doi.org/10.12029/gc2017Z103
8	Faults	Ye et al. (2017)	1:2,500,000	https://doi.org/10.12029/gc2017Z103

Administration (NASA) Global Precipitation Measurement (GPM). The spatial resolution of the GPM data is 0.1°. The annual precipitation was calculated using the monthly data. We selected these parameters because they are some of the most frequently used factors in previous landslide hazard modeling (Shu et al. 2019; Tanyas et al. 2022). These factors include both environmental factors (DEM, slope, aspect, land cover type, lithology, fault density, annual precipitation) and the triggering factor (PGA). These layers cover major factors that influence the occurrence of landslides from different aspects. In addition, the main objective was to compare different sampling strategies, as long as they use the same layers.

3 Method

In the first part of this section, we outline our two sampling strategies for generating landslide and non-landslide samples. Particular attention is given to detailing the production of non-landslide samples, as it shares the same methodology with landslide samples. In the second part, we elucidate the underlying principles and critical parameters utilized in our machine learning model. Finally, in the third part, we expound upon the methods employed to validate the performance of the two sampling strategies.

3.1 Two Strategies for Selecting Non-landslide Samples

To train a machine learning model for landslide hazard analysis, we had to select landslide and non-landslide sampling points, which are used to extract positive and negative samples. For the 2008 Wenchuan Earthquake, Xu et al. (2014) interpreted about 200,000 coseismic landslides. Among these landslides, 23,561 have an area of more than 10,000 m^2. The geometry centers of these landslide polygons were selected as landslide sampling points. Among them, we deleted landslide points that fall outside the landslide polygons, which usually occur in long-shaped landslides (He et al. 2021). We used two datasets to train and validate random forest models. Both datasets have the same landslide points but different non-landslide points.

Non-landslide samples in both strategies were selected as points and had to meet two criteria: (1) they had to be more than 30 m from the boundaries of any landslide polygons; and (2) they had to be more than 100 m from any non-landslide points. The minimum distance from non-landslide points to landslide boundaries was set at 30 m because it avoids possible errors caused by misinterpretation of the landslide boundaries from remote sensing images. For landslide interpretation, the spatial resolution of the remote sensing images used by Xu et al. (2014) ranged from 0.5 to 15 m. The 30 m distance ensures that there are at least two pixels from landslide polygons even in the 15 m images. The distance between non-landslide points was set at 100 m to ensure feature differences in neighboring non-landslide samples. Because the spatial resolution of the feature layers is 30 m, a distance of 100 m ensures that the nearest non-landslide samples are in two different pixels.

The first non-landslide sampling strategy is a traditional way frequently used by previous studies. In this strategy, we randomly generated the same number of non-landslide points as landslide samples. These non-landslide samples are randomly scattered within the study area.

To carry out the second non-landslide sampling strategy, we first used a 2 km-by-2 km fishnet to segment the study area into grids of the same area. Second, we singled out grids that have landslide points. Third, we calculated the areal percentage of all landslides within each grid. Fourth, we divided all the grids that have landslide points into five classes according to the areal percentages of landslides. For each class, $N_{nls,i}$ non-landslide points were randomly produced within their respective spatial extent.

$$N_{nls,i} = \frac{A_{nls,i}}{A_{ls,i}} \times N_{ls,i} \tag{1}$$

where, $N_{ls,i}$ is the number of landslide points in the ith grid; and $A_{ls,i}$ and $A_{nls,i}$ are areas of all landslides and non-landslides in the ith grid, respectively. To make it easy to process, we divided all 2 km squares that have landslide points into five classes. Squares of the same class are merged to a large polygon. We calculated non-landslide points within each polygon and scattered these points evenly within them.

After landslide and non-landslide points were produced, these two datasets were used to extract samples used to train random forest models. There is an equal number of 23,688 landslide and non-landslide samples in the first strategy. The number of landslide samples is the same for both strategies, whereas there are many more non-landslide samples (202,230) produced in the second strategy.

3.2 Random Forest Models

Random forest model is an ensemble model that integrates multiple uncorrelated decision trees and is very popular in landslide mapping (Chen et al. 2018; He et al. 2021). Bias and overfitting can be overcome by many uncorrelated individual trees. These decision trees are grown by randomly selecting different samples and features. A random subset of features is used to grow a decision tree. To train the model, a testing sample is set with replacement, which is also known as the out-of-bag (OOB) sample. This strategy is employed to increase the difference among decision trees and reduce their correlations. Random forest model can be used in both classification and regression problems. In this study, the classification model was used, in which multiple decision trees predict either landslide or non-landslide targets. The final results are determined by majority voting. Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. We used the same hyperparameters for both models with 100 trees, and other parameters as default.

The main steps of the random forest algorithm are as follows: (1) Make a sample pool with N samples. Select N samples from the pool to form a dataset of training samples. During this process, any single sample may be selected multiple times. Those samples that were not selected are called OOB data. These OOB data will be used to evaluate the performance of the model. (2) For each training sample, a decision tree is generated. Assume there are M features in the sample. Then randomly choose the number of F features (F = sqrt(M)) from the M features to form F nodes in the tree. (3) Decision trees are generated using the Classification and Regression Tree (CART) algorithm, each of which grows freely without pruning. (4) Repeat the above steps k times to obtain a total of k training sets and k decision trees. Accordingly, there are k OOB data. (5) These k decision trees are formed into a random forest, which can be used for classification in new data. The final result is decided by voting of k trees in the random forest. In this study, there are 100 trees and a total of 8 features. All other parameters, such as the maximum depth of the decision tree (which is the number of split nodes of the tree), the minimum number of samples, are set by default. To use the random forest model for landslide hazard mapping, we first produced landslide/non-landslide samples using our proposed strategy. Then, we split these samples into training (70% of all samples) and validation (30%) subsamples.

3.3 Validation

We used two validation methods to test the performance of both models. The first method was to use the receiver operating characteristic (ROC) curve, which is a plot between true positive and false positive rates. The area under the receiver operating characteristic curve ($\mathrm{AUC}_{\mathrm{ROC}}$) can be used to quantitatively measure a model's performance, which ranges from 0.5 to 1. An AUC near 0.5 indicates that the model predicts with random values, whereas 1 indicates perfect prediction. The $\mathrm{AUC}_{\mathrm{ROC}}$ has been very popular and used in most landslide hazard/susceptibility models.

We also used a second validation method to measure the spatial match between landslide hazards and interpreted landslides based on the direct comparison of interpreted landslides and predicted landslide hazards in a 2 km-by-2 km fishnet. For each grid of the fishnet, we calculated the mean landslide hazards predicted by both models and compared them with the landslide areal percentage.

4 Results

The results are presented in three subsections. First, we display the landslide and non-landslide points within 2 km-by-2 km grids, which are generated using the two distinct sampling strategies. Second, we present the outcomes of our earthquake-triggered landslide hazard modeling. Last, we showcase the validation results for both sampling strategies.

4.1 Landslide and Non-landslide Points in 2 km-by-2 km Grids

We produced a 2 km-by-2 km fishnet for the Wenchuan Earthquake-affected area. By using the landslide inventory of the Wenchuan Earthquake (Xu et al. 2014), we calculated landslide areal percentages for each grid of the fishnet (Fig. 2). The landslide percentage of each grid shows the probability of that grid being affected by a landslide, and it shows which grid is more hazardous than others. Figure 2 shows the extremely uneven landslide distribution in the study area. Note that, even grids with the highest landslide

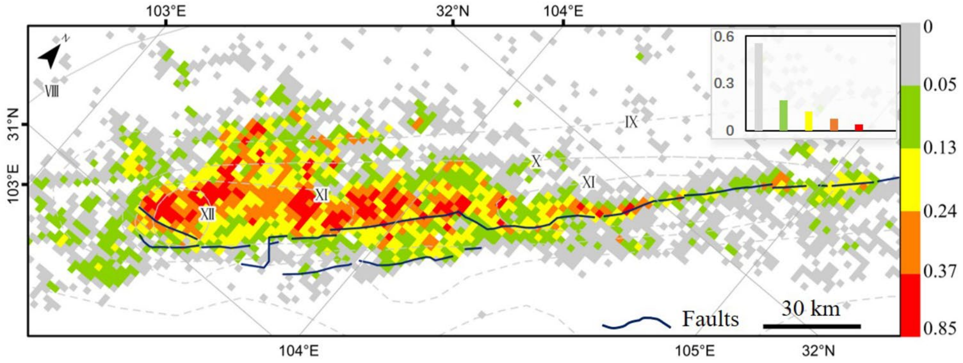

Fig. 2 Landslide areal density (the value within each grid) of the 2008 Wenchuan Earthquake on a fishnet of 2 km-by-2 km in the study area on the eastern Tibetan Plateau. Landslide polygons were manually interpreted by Xu et al. (2014) from high spatial resolution remote sensing images

percentage have non-landslide slopes, whereas some grids that are far from the epicenter have landslides. Most landslides concentrate along active faults. The southern section of the study area has more landslides than the northern section. Figure 2 shows five classes of landslide percentage. Although landslides triggered by the 2008 earthquake are distributed in a large region of about 110,000 km^2, most landslides cluster along the active faults of the earthquake. High landslide percentage grids are located close to the seismogenic faults and the epicenter area, whereas low landslide percentage grids (less than 0.13) are located in peripheral regions that are scattered around the periphery of the study area.

Figure 2 also shows that there are more stable slopes (non-landslide) than landslides for most grids. The highest landslide percentage is less than 0.85, indicating that there are more than 15% non-landslide landscapes even in these most extensive landsliding grids; 99.5% of the grids have landslide percentages of less than 50% and more than half of the grids (55.6%) have the lowest density of less than 0.05; more than 96% of the grids have less than 37% landslide percentages in area.

We used the proposed spatially heterogeneous non-landslide sampling strategy to produce non-landslide points. Figure 3 shows landslide and non-landslide samples in five selected 2 km-by-2 km grids to represent five classes of landslide areal percentages. For each class, the ratio of landslide and non-landslide samples is proportional to their areal ratio. From grid a to grid e, the number of landslide samples (also the areal percentage) decreases, whereas the number of non-landslide samples increases. In comparison to grids b−e, grid a exhibits a higher percentage of area covered by landslides, resulting in a larger number of landslide samples. Conversely, grids b−e have lower landslide

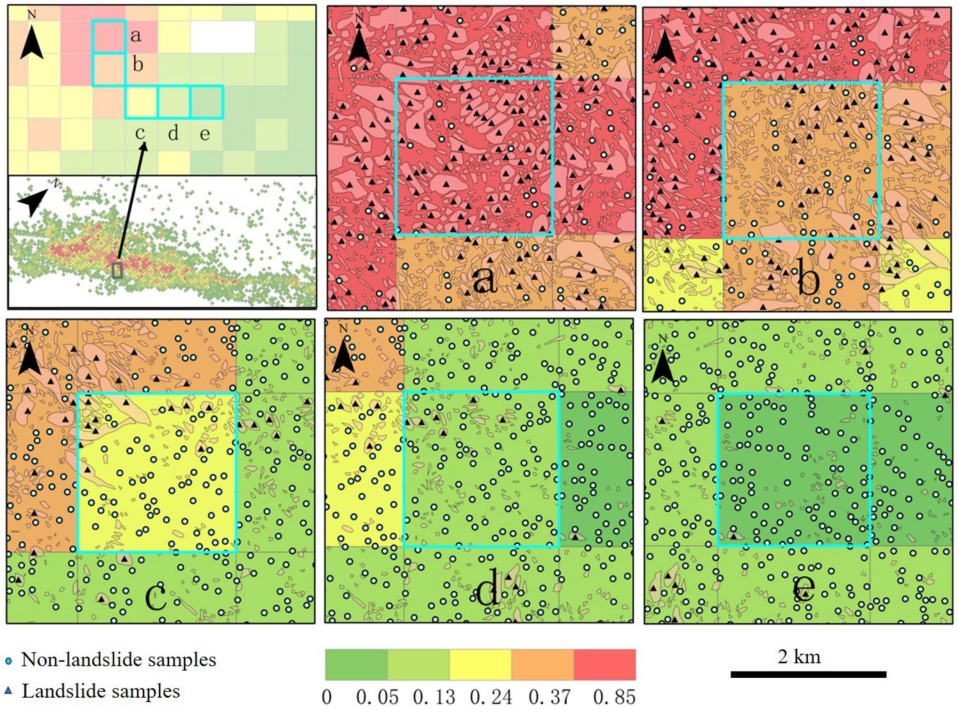

Fig. 3 Landslide and non-landslide samples in five classes of landslide areal density

areal percentages and consequently fewer landslide samples. As shown in Fig. 2, more than 99.5% of the grids have less than 50% landslide areal percentage, almost all grids have more non-landslide samples. The number of non-landslide samples is much larger than the number of landslide samples in this study area.

4.2 Earthquake-Triggered Landslide Hazard Modeling

Figure 4 shows landslide hazard mapping results from both random forest models trained with the two datasets produced by different sampling strategies. The spatial patterns of higher hazard values from the two models are similar: the southern section has higher hazard values than the northern section, and high values concentrate along active faults and in the epicenter. Landslide hazards produced by both models range from 0 to 1. Figure 4a has more high landslide hazard values than Fig. 4b. There is a much larger area of hazard of more than 0.9 in the first map, whereas the area with hazard of more than 0.7 is much smaller in the second map. A large area near the epicenter has hazard values of more than 0.9. In addition, near the epicenter of the earthquake, high values in Fig. 4a are homogeneous, whereas there are some low hazard values among high hazard values in Fig. 4b.

4.3 Hazard Modeling Validation

We selected three subregions to examine the difference between the modeling results from the two different non-landslide sampling strategies. Figure 5-(1, 3, 5) were produced by the strategy to randomly select the same number of landslide/non-landslide samples for the entire study area. Non-landslide samples in other subpanels (Figs. 5-(2, 4, 6) were produced by an unbalanced number of landslide/non-landslide samples. Except for the non-landslide samples, the hyperparameters of the random forest models and landslide samples of both models are the same. There are many more non-landslide samples in our proposed sampling strategy (Figs. 5-(2, 4, 6).

The colored maps of Fig. 5 are landslide hazards produced by the random forest models fed by the same landslide points but two different non-landslide samples. Hazard results produced by the two models in these three selected subregions are very different. In all subpanels of Fig. 5, most landslide polygons overlap with high landslide hazard values. The area of high hazard values in Figs. 5-(1, 3, 5) is much larger than that of Figs. 5-(2, 4, 6). The distribution of high landslide hazards (more than 0.7) produced by the second sampling strategy matches the interpreted landslides by Xu et al. (2014) better than the results from the traditional sampling strategy.

Figure 6 shows the modeled results for the two largest landslides, the Daguangbao landslide (7.3 km^2) and the Wenjiagou landslide (2.9 km^2). The first model consistently overpredicted landslide hazard for these two subareas near the landslides. Not all parts of the Daguangbao landslide have high hazard values in the map produced by our proposed method. This may be because the mechanism of the mega-landslide is distinctly different from other earthquake-triggered landslides (Hu et al. 2019). Higher landslide hazard values can be found in the middle and upper parts of the

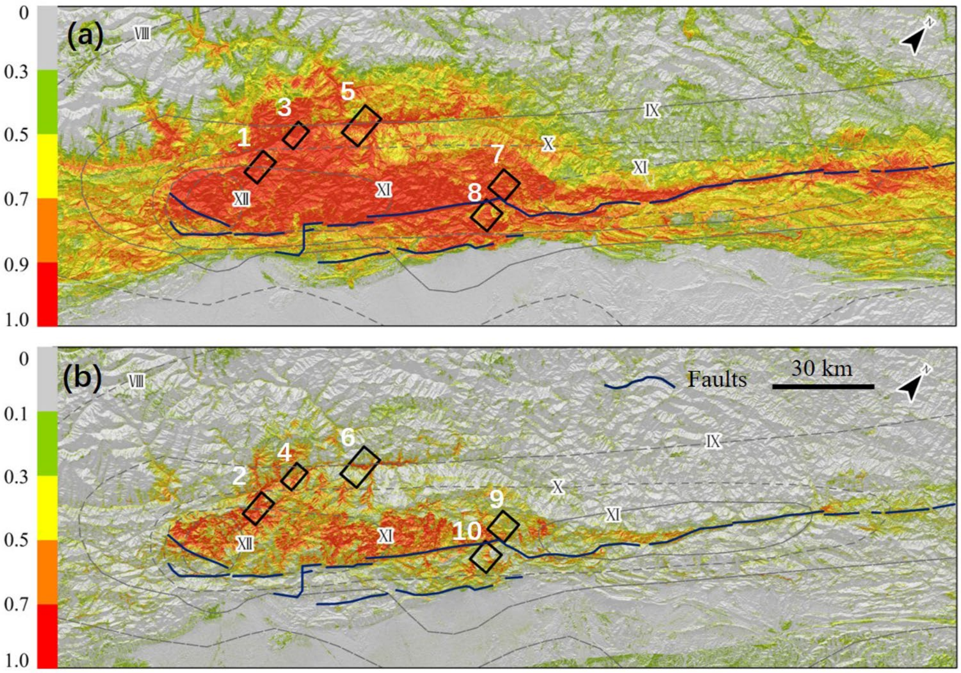

Fig. 4 Landslide hazard maps (1–10) produced for the study area on the eastern Tibetan Plateau by two different non-landslide sampling strategies with the same landslide inventory and machine learning models. **a** By the strategy to randomly select the same number of landslide/non-landslide samples; **b** By an unbalanced number of landslide/non-landslide samples. Pixel values are landslide probabilities predicted by landslide hazard models. Selection of non-landslide samples is the only difference between the two landslide hazard models

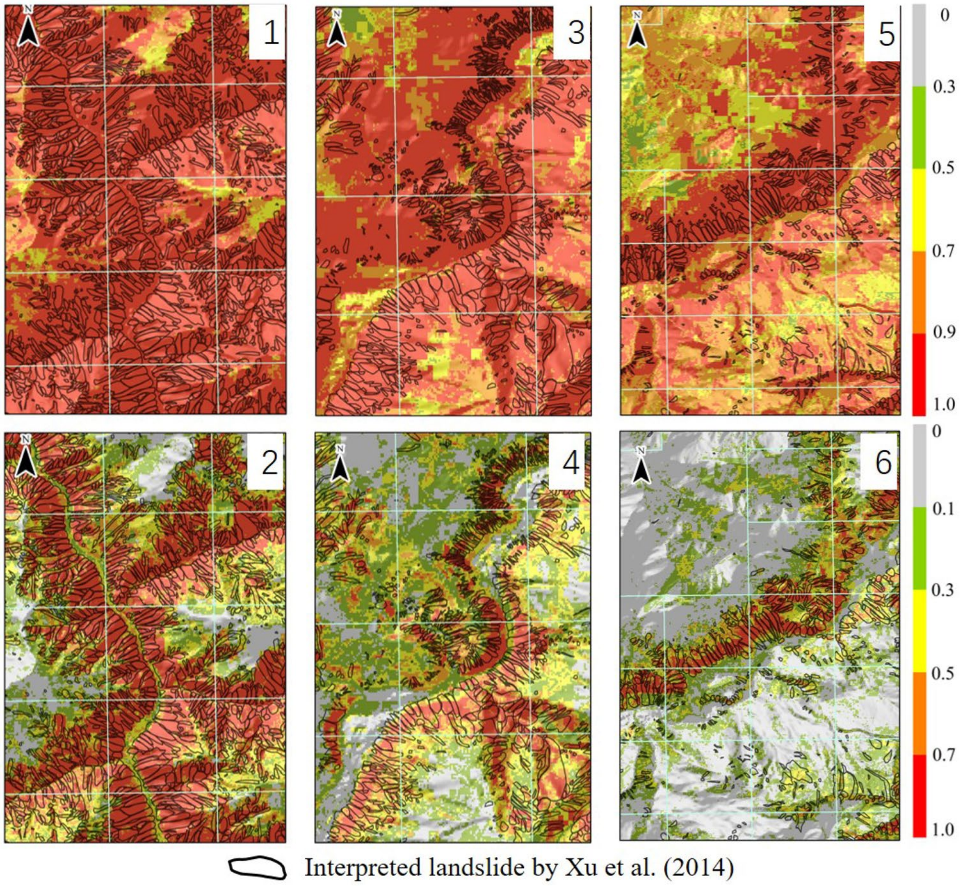

Fig. 5 Distribution of landslide and non-landslide samples produced by two different sampling strategies in three selected subregions of the study area on the eastern Tibetan Plateau. The colors show landslide hazards predicted by the random forest models of the same hyperparameters. Locations of subpanels 1–6 are shown in Fig. 4

Wenjiagou landslide, where the landslide geometric center is located. The lowest landslide hazards are found at the toe of the Wenjiagou landslide, which is impacted by the flow of the landslide material.

The samples were randomly divided into two subdatasets with 70% as training data and the remaining 30% as validation data. The two random forest models were applied in their respective validating datasets to estimate landslide probabilities. These predicted landslide probabilities were then compared with their known labels to determine the predictive skills of these two models. Figure 7 shows the ability of these two models to predict landslide hazards. We used the ROC curves to measure the performance of the two models. We calculated the area under the curve (AUC). The AUC of the first model (0.95) is much higher than that of the second model (0.853), indicating that the performance of the second model is lower than the first. Both AUCs are larger than 0.85, indicating that both models performed well in the validating datasets. In addition, the AUCs for the training datasets of the two models are 0.97 and 0.87, respectively, which are very close to their performances in the validating datasets. These high and stable AUCs of the two models indicate the robustness of both models in landslide hazard modeling.

We further calculated the mean landslide hazard values of both models within each grid of the 2 km-by-2 km fishnet. The mean hazard values of both models in the fishnet are compared with the landslide areal percentage to assess the accuracy of both models in locating high landslide hazards. Figure 8 shows that the coefficient of determinant of the first model (R^2 = 0.55) is much lower than that of the second model (R^2 = 0.88). Although both models overestimated landslide hazards, the relation between the landslide areal percentage and the modeled landslide hazards of the second model is much closer to the 1:1 line than the first model. In addition, there is an upturn tail for high landslide hazard values in the first model, indicating that the distribution of the predicted values of the first model is nonlinear. Compared to high landslide hazards, overestimations of landslide hazards in lower values are much worse.

5 Discussion

This section primarily focuses on analysis and discussion, divided into three parts. First, we examine the enhancement in accuracy of landslide hazard mapping resulting from the implementation of the new strategy. Second, we explore the

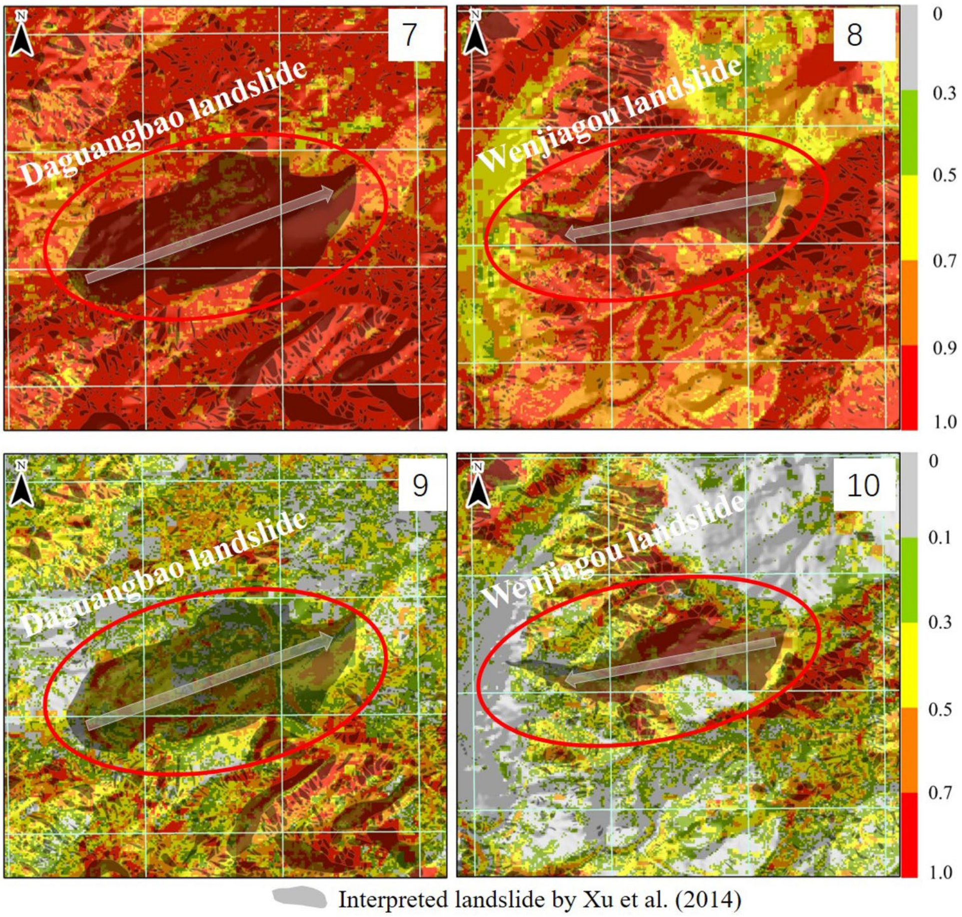

Fig. 6 Modeled landslide hazard for the two largest landslides (the areas of the Daguangbao and Wenjiagou landslides are 7.3 and 2.9 km^2, respectively) in the study area on the eastern Tibetan Plateau. Locations of subpanels 7–10 are shown in Fig. 4

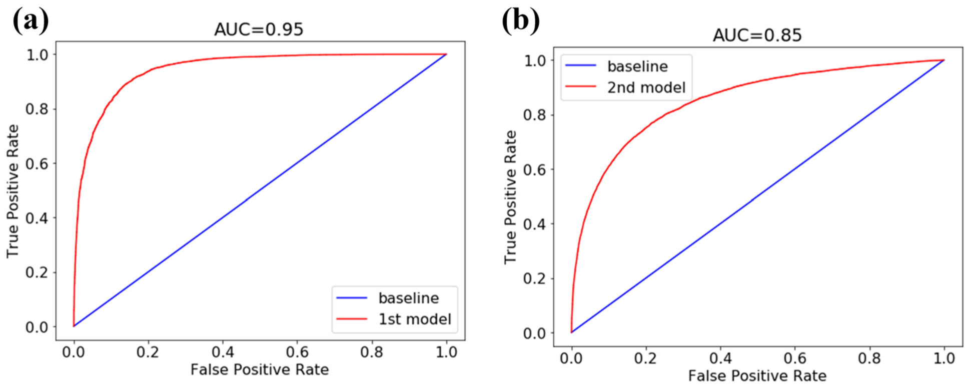

Fig. 7 Receiver operating characteristic (ROC) curves of the two models with two different sampling strategies

potential future applications of the new strategy in rainfall-induced and global earthquake-triggered landslide hazard mapping. Last, we identify and discuss the limitations of the new strategy, as well as propose possible avenues for improvement in the future.

5.1 Improvement of Landslide Hazard Predictions

Machine learning models have been extensively used in landslide hazard analysis (Xu et al. 2012; Chen et al. 2018; Qi et al. 2021). Efforts on earthquake-triggered landslide hazard analysis have been made on designing optimal feature layers and selecting the best machine learning models (Tanyas et al. 2019), yet few have been devoted to optimizing landslide/non-landslide sampling strategies (Shao et al. 2020; Liu et al. 2021; Yang et al. 2022). In this study, we proposed a strategy to produce spatially heterogeneous non-landslide points. Compared to the traditional landslide hazard map with all epicentral areas having high hazard values (Allstadt et al. 2018), our landslide hazard map shows more

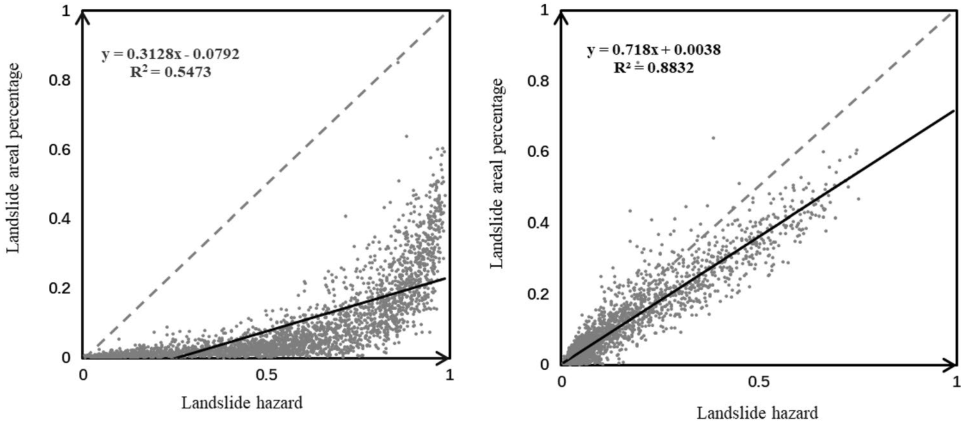

Fig. 8 Scatterplots of landslide areal density and landslide hazards, predicted by random forest models with two different sampling strategies. Each dot represents the landslide areal density and mean landslide hazard value within a grid of the 2 km-by-2 km fishnet

details in the epicentral area. The fact that there is a large portion of the Earth's surface unaffected by earthquake-triggered landslides even in the most severely landslide-affected region (Fig. 2) indicates that traditional landslide hazard models have overestimated the hazard. In addition, there is a significant improvement (R^2 from 0.55 to 0.88) in the match between the spatial pattern of predicted landslide hazard values and interpreted landslides. These results indicate that our strategy to select non-landslide points is very effective in predicting the location of earthquake-triggered landslides at the local scale.

Our model's capability to depict more details in the epicentral area is attributed to our strategy to select non-landslide points. The epicentral area has the largest landslide areal percentage and many landslide points. In the traditional sampling strategy, non-landslide points are randomly distributed. A large number of landslide points and very few non-landslide points in the epicenter would overrepresent landslide hazards. Our strategy fully considered the spatial pattern of landslides and solved the problem of underpresentation of non-landslides in the epicentral area, which is evidenced by the contributions of the used layers (Fig. 9). Figure 9 shows feature importance in the two models. Lithology is the most important feature in both models. In the first model, distance to faults has high ranks, whereas slope and land cover type ranked low. Fault data were derived from a 1:2,500,000 geological map and contain no detailed information. At this scale, distance to fault is correlated to the spatial pattern of earthquake-triggered landslides, that is, more landslides are located near faults than far from the faults. In the second model, slope and land cover type ranked second and fourth, respectively. Both slope and land cover type data are more site-specific and their high ranks in the second model explain well their detailed depiction of high landslide hazard values in the epicentral area.

The difference in feature importance of the two models indicates that our proposed non-landslide sampling strategy captured subtle differences on slope scale. This is different from the traditional sampling strategy, which only ensures reasonable spatial patterns of earthquake-triggered landslides.

Existing studies show that there are no intrinsic differences between different machine learning models. Compared to other machine learning models, random forest works better in dealing with categorical variables (He 2008; Cao 2014) and for nonlinear relationships (Youssef et al. 2015). Variables do not need to be rescaled with the random forest model (Ham et al. 2005). By building multiple unrelated decision tree models, the performance of random forest could be more robust, which leads to high accuracy in landslide hazard modeling (Hong et al. 2017; Liao et al. 2022). Despite the fact that random forest can be computationally intensive for large datasets, it handles nonlinear relationships well (Breiman 2001; Micheletti et al. 2013; Cheng et al. 2021). In addition, it needs fewer samples and is less prone to over-fitting than deep learning models (Catani et al.

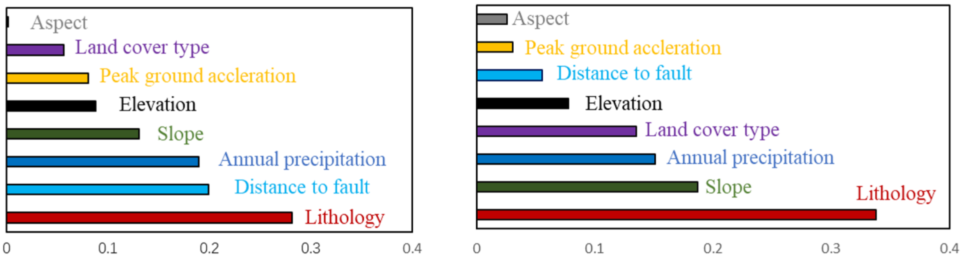

Fig. 9 Feature importance ranked by random forest models with two different sampling strategies

2013; Bui et al. 2020). Similar to other machine learning models, random forest is a black-box algorithm (Breiman 2001; Marteau 2021).

5.2 Implications for Regional Landslide Hazard and Risk Analysis

Our proposed non-landslide sampling strategy results in more reasonable landslide hazard maps, characterized by the good match between high hazard slopes and interpreted landslides. Especially the mean landslide hazard values within the fishnet grids are linearly correlated with landslide areal percentages. The well-matched spatial pattern between the modeled landslide hazard and interpreted landslides means that the result may be directly used by relevant stakeholders.

Using interpreted landslide inventory, Jibson et al. (2000) counted the percentage of failed pixels within a calculated Newmark displacement. They used that percentage as the probability of each pixel impacted by earthquake-triggered landslides. Their model is a hybrid one and requires detailed engineering geology parameters and landslide inventory. In addition, their percentage of failed pixels for a given Newmark displacement may not work for another earthquake. To improve its performance, many other earthquakes that triggered landslides should be used, which would require more detailed engineering geology parameters and known landslide inventories. In comparison, we used a much simpler way in landslide hazard modeling, which only requires easy-to-access data, such as the DEM and lithology maps. This will save lots of time and resources for on-site data collection but achieve high spatial accuracy. The sampling strategy proposed by this study may also be applied to precipitation-induced landslides. Precise landslide hazard mapping could also benefit landslide risk analysis.

5.3 Limitations

Although we used a fishnet to calculate landslide areal percentage within each 2 km-by-2 km grid, we did not produce non-landslide points at the grid level. Instead, we divided all grids into five categories according to their landslide areal percentages. This simplification reduces the effectiveness of the strategy. Despite this, our results are still significantly better than the hazard map produced with the traditional sampling strategy (Fig. 8). In the future, we call for a customized design of non-landslide points by considering the landslide areal percentage of each grid.

Another caveat is that the size of the grid used in this study may not be optimal. The 2 km-by-2 km size of the fishnet was selected subjectively by referring to the size of the landslides induced by the Wenchuan Earthquake (Xu et al. 2014). According to the coseismic landslide inventory interpreted by Xu et al. (2014), the largest landslide in our study area is 7.8 km^2 with a maximum width of 2.2 km. Of the 200,000 coseismic landslides, less than 10 landslides have areas larger than 1 km^2. Therefore, our fishnet of 2 km-by-2 km could ensure that almost no grid is fully occupied by landslides, which means all grids could have proportions for both landslides and non-landslides. Smaller grids may be suitable for precipitation-triggered landslides, which are usually much smaller in size. In addition, use of other irregular divisions (such as the widely used slope unit) (Tanyas et al. 2019) other than the squared grids may also be tried. The partition of squared grids probably cannot represent the natural difference of neighboring grids.

Our study did not differentiate landslide types and different parts of landslides (van Westen et al. 2006). It is possible that the proposed method may be more suitable for translational landslides and avalanches but not perfect for debris flows. In addition, we may expect a better performance by solely modeling landslide sources rather than runout paths. These ameliorations would require better landslide inventory data to differentiate landslide source and runout areas. Other predisposing factors should also meet commensurable quality. For example, higher spatial resolution and accuracy of topographic data are recommended.

This study only used one event and its interpolations to other earthquakes await further testing. Based on our improvement in the modeling performance, we argue that previous overprediction of landslide hazards for the global earthquake-triggered landslide model is not caused by the large number of landslides of the Wenchuan Earthquake (Allstadt et al. 2018). In the future, a global earthquake-triggered landslide hazard map may be produced and validated by incorporating global earthquake-triggered landslide inventories.

The random forest model excelled at modeling nonlinear relationships. The AUC$_{ROC}$ for the second model is much lower than for the first one, which indicates that current globally available layers are good at capturing the general spatial pattern of earthquake-triggered landslides but have difficulty depicting subtle differences between landslides and non-landslides at the slope scale. Further improvements of earthquake-triggered landslide hazard models should include causative layers.

In this study, the spatial resolution of the product is 30 m, which is consistent with topographic layers (DEM and its derivatives). All other data have much coarser resolutions and were resampled at 30 m. In particular the scale of the fault and lithology data is 1:2,500,000, which is approximately equivalent to a spatial resolution of 660 m. Higher spatial resolution of lithology and faults could probably help in producing more refined hazard maps, but we are mainly concerned with how to obtain optimal modeling results with the same accessible data. Despite the huge differences in the spatial resolution of these datasets, we could still carry out a fair comparison between two sampling strategies.

6 Conclusion

Landslide and non-landslide sampling strategies could significantly influence landslide hazard mapping. Our proposed spatially heterogeneous non-landslide sampling strategy significantly improved the spatial pattern prediction of landslide hazards. The ratio of landslide to non-landslide sampling points should be proportional to their areal ratio at the local scale, which will ensure a balanced representation of landslides and non-landslides. The traditional strategy of randomly selecting non-landslides in an earthquake-affected area should be abandoned. Without a reasonable sampling strategy, further design of feature layers or improving machine learning models will offer little help for regional landslide modeling.

Previously used validation methods such as $\mathrm{AUC}_{\mathrm{ROC}}$ or model prediction accuracy are ways to assess the performance of landslide hazard models, but not the spatial validity of landslide hazard predictions. Higher values in these metrics do not guarantee good matches between high landslide hazard values and interpreted landslides. Therefore, direct comparison of predicted landslide hazard values with earthquake-triggered landslide inventory is highly recommended.

Acknowledgements This study was jointly supported by the Fundamental Research Funds for the Central Universities (Grant No.2021ZY46) and the Second Tibetan Plateau Scientific Expedition and Research Program (STEP, Grant No. 2019QZKK0906). Wentao Yang is grateful for the scholarship from the China Scholarships Council (No. 202006515016)

References

Allstadt, K.E., R.W. Jibson, E.M. Thompson, C.I. Massey, D.J. Wald, J.W. Godt, and F.K. Rengers. 2018. Improving near-real-time coseismic landslide models: Lessons learned from the 2016 Kaikōura, New Zealand, Earthquake. *Bulletin of the Seismological Society of America* 108(3B): 1649–1664.

Breiman, L. 2001. Random forests. *Machine Learning* 45: 5–32.

Bui, D.T., P. Tsangaratos, V.T. Nguyen, N.V. Liem, and P.T. Trinh. 2020. Comparing the prediction performance of a deep learning neural network model with conventional machine learning models in landslide susceptibility assessment. *CATENA* 188: 104426.

Cao, Z. 2014. Study on optimization of random forests algorithm. Ph.D. dissertation. Capital University of Economics and Business, Beijing, China (in Chinese).

Catani, F., D. Lagomarsino, S. Segoni, and V. Tofani. 2013. Landslide susceptibility estimation by random forests technique: Sensitivity and scaling issues. *Natural Hazards and Earth System Sciences* 13(11): 2815–2831.

Chen, W., S. Zhang, R. Li, and H. Shahabi. 2018. Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Science of the Total Environment* 644: 1006–1018.

Cheng, Y., T.T. Yu, and N.T. Son. 2021. Random forests for landslide prediction in Tsengwen river watershed central Taiwan. *Remote Sensing* 13(2): 199.

Cui, P., Y. Lin, and C. Chen. 2012. Destruction of vegetation due to geo-hazards and its environmental impacts in the Wenchuan earthquake areas. *Ecological Engineering* 44: 61–69.

Emberson, R., N. Hovius, A. Galy, and O. Marc. 2016. Chemical weathering in active mountain belts controlled by stochastic bedrock landsliding. *Nature Geoscience* 9(1): 42–45.

Fell, R., J. Corominas, C. Bonnard, L. Cascini, E. Leroi, and W.Z. Savage. 2008. Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Engineering Geology* 102(3–4): 85–98.

Ham, J., Y. Chen, M.M. Crawford, and J. Ghosh. 2005. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing* 43(3): 492–501.

He, X. 2008. *Multivariate statistical analysis*. Beijing: People's University of China Press.

He, Q., M. Wang, and K. Liu. 2021. Rapidly assessing earthquake-induced landslide susceptibility on a global scale using random forest. *Geomorphology* 391: 107889.

Hong, H., P. Tsangaratos, I. Ilia, W. Chen, and C. Xu. 2017. Comparing the performance of a Logistic Regression and a Random Forest Model in landslide susceptibility assessments The case of Wuyaun Area, China. In *Advancing culture of living with landslides*, ed. M. Mikos, B. Tiwari, Y. Yin, and K. Sassa, 1043–1050. Cham, Switzerland: Springer.

Hu, W., R. Huang, M. McSaveney, L. Yao, Q. Xu, M. Feng, and X. Zhang. 2019. Superheated steam, hot CO_2 and dynamic recrystallization from frictional heat jointly lubricated a giant landslide: Field and experimental evidence. *Earth and Planetary Science Letters* 510: 85–93.

Huang, R., and X. Fan. 2013. The landslide story. *Nature Geoscience* 6(5): 325–326.

Jibson, R.W., E.L. Harp, and J.A. Michael. 2000. A method for producing digital probabilistic seismic landslide hazard maps. *Engineering Geology* 58(3–4): 271–289.

Johnston, E.C., F.V. Davenport, L. Wang, J.K. Caers, S. Muthukrishnan, M. Burke, and N.S. Diffenbaugh. 2021. Quantifying the effect of precipitation on landslide hazard in urbanized and non-urbanized areas. *Geophysical Research Letters*. https://doi.org/10.1029/2021GL094038.

Jones, J.N., S.J. Boulton, G.L. Bennett, M. Stokes, and M.R. Whitworth. 2021. Temporal variations in landslide distributions following extreme events: Implications for landslide susceptibility modeling. *Journal of Geophysical Research Earth Surface*. https://doi.org/10.1029/2021JF006067.

Kirschbaum, D., T. Stanley, and Y. Zhou. 2015. Spatial and temporal analysis of a global landslide catalog. *Geomorphology* 249: 4–15.

Larsen, I.J., and D.R. Montgomery. 2012. Landslide erosion coupled to tectonics and river incision. *Nature Geoscience* 5(7): 468–473.

Liao, M., H. Wen, and L. Yang. 2022. Identifying the essential conditioning factors of landslide susceptibility models under different grid resolutions using hybrid machine learning: A case of Wushan and Wuxi counties China. *Catena*. https://doi.org/10.1016/j.catena.2022.106428.

Liu, M., J. Liu, S. Xu, T. Zhou, Y. Ma, F. Zhang, and M. Konečný. 2021. Landslide susceptibility mapping with the fusion of multi-feature SVM model based FCM sampling strategy: A case study from Shaanxi Province. *International Journal of Image and Data Fusion* 12(4): 349–366.

Lombardo, L., H. Bakka, H. Tanyas, C. van Westen, P.M. Mai, and R. Huser. 2019. Geostatistical modeling to capture seismic-shaking patterns from earthquake-induced landslides. *Journal of Geophysical Research: Earth Surface* 124: 1958–1980.

Marteau, P.F. 2021. Random partitioning forest for point-wise and collective anomaly detection—Application to network intrusion detection. *IEEE Transactions on Information Forensics and Security* 16: 2157–2172.

Micheletti, N., L. Foresti, S. Robert, M. Leuenberger, A. Pedrazzini, M. Jaboyedoff, and M. Kanevski. 2013. Machine learning feature selection methods for landslide susceptibility mapping. *Mathematical Geosciences* 46(1): 33–57.

Nowicki Jessee, M.A., M.W. Hamburger, K. Allstadt, D.J. Wald, S.M. Robeson, H. Tanyas, M. Hearne, and E.M. Thompson. 2018. A global empirical model for near-real-time assessment of seismically induced landslides. *Journal of Geophysical Research: Earth Surface* 123: 1835–1859.

Nowicki, M.A., D.J. Wald, M.W. Hamburger, M. Hearne, and E.M. Thompson. 2014. Development of a globally applicable model for near real-time prediction of seismically induced landslides. *Engineering Geology* 173: 54–65.

Petley, D. 2012. Global patterns of loss of life from landslides. *Geology* 40(10): 927–930.

Pokharel, B., O.F. Althuwaynee, A. Aydda, S.W. Kim, S. Lim, and H.J. Park. 2021. Spatial clustering and modelling for landslide susceptibility mapping in the north of the Kathmandu Valley, Nepal. *Landslides* 18(4): 1403–1419.

Qi, W., C. Xu, and X. Xu. 2021. AutoGluon: A revolutionary framework for landslide hazard analysis. *Natural Hazards Research* 1(3): 103–108.

Shao, X., S. Ma, C. Xu, and Q. Zhou. 2020. Effects of sampling intensity and non-slide/slide sample ratio on the occurrence probability of coseismic landslides. *Geomorphology*. https://doi.org/10.1016/j.geomorph.2020.107222.

Shu, H., M. Hürlimann, R. Molowny-Horas, M. González, J. Pinyol, C. Abancó, and J. Ma. 2019. Relation between land cover and landslide susceptibility in Val d'Aran, Pyrenees (Spain): Historical aspects, present situation and forward prediction. *Science of the Total Environment* 693: Article 133557

Tanyas, H., T. Görüm, D. Kirschbaum, and L. Lombardo. 2022. Could road constructions be more hazardous than an earthquake in terms of mass movement?. *Natural Hazards* 112: 639–663.

Tanyas, H., M. Rossi, M. Alvioli, C.J. van Westen, and I. Marchesini. 2019. A global slope unit-based method for the near real-time prediction of earthquake-induced landslides. *Geomorphology* 327: 126–146.

van Westen, C.J., E. Castellanos, and S.L. Kuriakose. 2008. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. *Engineering Geology* 102: 112–131.

van Westen, C.J., T.W.J. van Asch, and R. Soeters. 2006. Landslide hazard and risk zonation—Why is it still so difficult?. *Bulletin of Engineering Geology and the Environment* 65: 167–184.

Wang, W., Z. He, Z. Han, Y. Li, J. Dou, and J. Huang. 2020. Mapping the susceptibility to landslides based on the deep belief network: A case study in Sichuan Province, China. *Natural Hazards* 103: 3239–3261.

Xu, C., X. Xu, F. Dai, and A.K. Saraf. 2012. Comparison of different models for susceptibility mapping of earthquake triggered landslides related with the 2008 Wenchuan earthquake in China. *Computers & Geosciences* 46: 317–329.

Xu, C., X. Xu, X. Yao, and F. Dai. 2014. Three (nearly) complete inventories of landslides triggered by the May 12, 2008 Wenchuan Mw 7.9 earthquake of China and their spatial distribution statistical analysis. *Landslides* 11: 441–461.

Yang, C., L.-L. Liu, F. Huang, L. Huang, and X.-M. Wang. 2022. Machine learning-based landslide susceptibility assessment with optimized ratio of landslide to non-landslide samples. *Gondwana Research*. https://doi.org/10.1016/j.gr.2022.05.012.

Ye, T., C. Huang, and Z. Deng. 2017. Spatial database of the 1: 2,500,000 digital geologic map of the People's Republic of China. *Geology in China* 44(S1): 19–24 (in Chinese).

Yin, Y., F. Wang, and P. Sun. 2009. Landslide hazards triggered by the 2008 Wenchuan earthquake, Sichuan. *China. Landslides* 6(2): 139–152.

Youssef, A.M., H.R. Pourghasemi, Z.S. Pourtaghi, and M.M. Al-Katheeri. 2015. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region. *Saudi Arabia. Landslides* 13(5): 839–856.