Mapping the risk zoning of storm flood disaster based on heterogeneous data and a machine learning algorithm in Xinjiang, China

Yan Liu1 | Xinyu Lu1 | Yuanzhi Yao2 | Ni Wang1 | Yanyun Guo3 | Chunrong Ji1 | Jianhui Xu4
1Institute of Desert Meteorology, China Meteorological Administration, Urumqi, China
2School of Forestry and Wildlife Science, Auburn University, Auburn, Alabama
3Agricultural Meteorological Station, Xinjiang Uygur Autonomous Region Meteorological Bureau, Xinjiang Agricultural Meteorological Station, Urumqi, China
4Key Laboratory of Guangdong for Utilization of Remote Sensing and Geographical Information System, Guangdong Open Laboratory of Geospatial Information Technology and Application, Guangzhou Institute of Geography, Guangzhou, China

Abstract
Mapping flood risk zone is an essential task in the arid region for sustainable water resources management. Due to the lack of hydrological and meteorological information and disaster event inventory in Xinjiang, China, storm flood disaster (SFD) risk zoning is an effective technique in investigating the potential impact of SFD. In this study, the statistics about natural, social, and risk related to SFD are collated. With the help of the compiled inventory data, a disaster risk assessment model for storm flood is proposed for the Xinjiang region based on the random forest (RF) algorithm. Randomly selected negative and positive samples from the historical SFD locations are composed of five different total samples. The overall prediction accuracy of the five sample groups attained 83.48%, indicating that the proposed RF model can well capture the spatial distribution of SFD in Xinjiang. It should also be noted that the spatial heterogeneity and complexity of SFD had a significant effect on its spatial distribution in Xinjiang. There are spatial distribution characteristics of lowland plains and high plateaus; the main mountainous regions, plains in the middle-lower reaches of major rivers, and areas surrounding major lakes are prone to flooding. The variable importance RF indicates that the disaster risk is mainly affected by the following factors, including hazard factors, catastrophic intensity, population density, as well as economic development in the affected area. Besides, latitude, longitude, agricultural acreage, road density, distance from rivers, and the maximum monthly precipitation account for most of the increase in storm flooding disasters, and they are the main
1 | INTRODUCTION

Storm flood disaster (SFD) is one of the most serious natural disasters in China. They brought various negative impacts and significantly jeopardised social-economic growth, as well as livelihoods and human health (Ding & Zhang, 2009). The damage of natural flooding disasters has been found directly correlated to precipitation intensity, frequency, and accumulation of precipitation over a given period. Due to inadequate drainage and the accumulation of runoff, an excessive amount of accumulated water can bring large areas into high-risk zones (Ma, 1994). Moreover, under the context of global warming, the frequency and intensity of extreme rainfall have been aggravating (Guan, Wang, Zheng, et al., 2014; Qin & Stocker, 2014; Su, Song, & Zhang, 2005; Xu & Lu, 2007; Xu, Mao, & Lu, 2008), which leads to a certain increase of the high-risk zones of flooding disaster (Xu, Zhang, Zhou, et al., 2014). Flood disaster risk zoning, which ensembles the fundamental natural, social, and risk information related to flood disasters for a particular area, is a crucial technique for effective flood risk management. By strengthening climate monitoring and expanding investigations of SFD, the effects of these flooding disasters can be reduced, and beneficial to the national macro-control of flood prevention and disaster reduction.

The pioneering studies have made significant progress in mapping risk zoning of rainstorms and floods during the past 20 years. Studies on zonation and evaluation of rainstorm flood disaster risk have been conducted in different spatial scales, including watershed scale (Chen, 2014; Liu, Li, Liu, et al., 2008; Sanyal & Lu, 2005; Xie, Tian, & Lu, 2015), provincial regional scale (Duan, He, Nover, et al., 2016; Gan, Fan, Xiao, et al., 2017; Guan & Chen, 2007; Su, Xiao, Wei, et al., 2012; Wang Mao, Li, et al., 2011) and national scale (Islam & Sado, 2000; Washington, respectively. However, limited by the complex topography and landforms, the in-situ observation networks or facilities in mountainous areas of Xinjiang are very sparse, and the inventory-based flood event database is the most reliable source in flood risk assessment. Currently, the flood risk assessment in Xinjiang mainly focuses on the classification of disaster-prone rainstorm grade based on single precipitation index (Wang et al., 2011), interannual variation trend and influencing climate factors (Jiang, Hu, & Yang, 2004; Li, 2003; Wang, Cui, & Yao, 2008; Wu & Zhang, 2003). However, due to geographical factors and the difficulty in obtaining those datasets, few studies have been focused on the risk assessment and regionalization of rainstorm and flood disasters in Xinjiang.

Storm and flood risk zoning technologies include cloud model (Wan, Yin, Sun, et al., 2017), weighted comprehensive evaluation method, and GIS (Qi, Liu, Yang, et al., 2012; Tang & Zhu, 2005). Statistical models are also a useful tool for drawing and determining flood disaster risk zoning, based on flood investigation events and flood interpretation factors (Haruyama, Shigeko, Ohokura, et al., 1996; Sanyal & Lu, 2005; Zhao, Gao, Huang, & He, 2017). Because the nonlinear relationships exist between evaluation indices and the level of disaster risk, machine learning algorithms are needed when assessing SFD. Random forest (RF), a classification and regression tree-based machine learning algorithm (Breiman, 2001), is suitable for multivariate prediction. This method effectively resolves the nonlinear problems, and has been widely utilised in ecological environment monitoring (Chen & Ishwaran, 2012; Dong, Xi-Bing, & Peng, 2013; Mihăilescu, Gui, Toma, Popescu, & Sporea, 2013; Tesfamariam & Liu, 2010; Zhao, Tie, Yang, Cai, & Cao, 2012). Lai, Chen, Zhao, Wang, and Wu (2015) and Wang, Lai, Chen, et al. (2015) used the RF models in flood risk assessments on the Dongjiang River Basin in Jiangxi Province, China and, found that RF classification results were more reliable and had more robust data mining capabilities compared to support vector machines (SVMs). Feng, Liu, and Gong (2015) also found that RF outperformed the maximum likelihood (ML) method and the artificial neural network (ANN) in urban flood risk mapping. Additionally, the RF method had an excellent tolerance to noise and outliers to avoid overfitting. RF, therefore, had the advantages in computational efficiency, scalability, robustness, and accuracy among the most widely used machine learning algorithms. However, due to the spatial heterogeneity and complexity of flood disasters in the Xinjiang Region, the RF-based flood

triggering point for SFD in Xinjiang. The proposed model provides some insight into the disaster in the mountainous region, and gives useful guidance for the national macro-control of flood prevention and disaster reduction.

**KEYWORDS**

random forests, storm flood disaster, Xinjiang, zoning maps
forecasting method cannot be directly applied. The spatial variability of floods, the intrinsic uncertainty of the flood event database, and potential deviations challenge the quality of studying SFD in Xinjiang region.

In this study, we compiled a survey database of SFD in the township level from 1989 to 2017 across the Xinjiang region. Based on the feature of SFD in Xinjiang extracted from the database, 17 evaluation indices were selected, including extreme values of daily precipitation, monthly average precipitation extremes, days with daily precipitation $\geq 25$ mm extremes, elevations and coefficients of variation, relative elevation, slopes, distances from rivers and roads, runoff curves (curve number), gross domestic product (GDP), cultivated land density, population density, and vegetation coverage, etc. A risk assessment model based on RF was explicitly built for SFD risk in Xinjiang. We mapped a flooding disaster risk zoning map with a spatial resolution of 1 km and analysed the feature of risk for each zone.

2 | STUDY AREA AND DATA

2.1 | Study area

As an arid and semi-arid region in the northwest inland of China, Xinjiang has its unique landscapes such as mountainous fields, oases and the Gobi desert, where oases on the edge of basins and along river systems, and account for about 8% of the total area in Xinjiang and desert of vast territory accounts for about 26.9% (Figure 1). Since this hinterland surrounded by towering mountains is far from the ocean, marine current rarely reaches its central area. Consequently, it brings Xinjiang a temperate continental climate, arid weather and sparse precipitation that leads 155 mm natural precipitation and $2,429 \times 10^8$ m$^3$ total precipitation annually on average (Wang, Huang, Liang, et al., 2018). The precipitation in Xinjiang is relatively low but concentrates in a certain area. For example, precipitation is scarce in plain areas, but about 84% of precipitation occurs in the mountainous fields. Precipitation and melting snow in the mountains provide plenty of water for the plain oases, but the unique climatic and topographical features in this area have resulted in frequent storm flood disasters.

2.2 | Data description and processing

Based on the principle that natural and social attributes should be general and macro, and the evaluation units should be complete and systematic, this study proposed a comprehensive evaluation index system for SFD by integrating four categories of indices, which contain natural, social, disaster and disaster-causing factors. The average state parameters of 17 evaluation indices under this system were adopted to reflect the natural, social environment, and disaster mitigation capacity of the rainstorm and flood disaster in Xinjiang (Figure 2).

- Natural indices: These indices are comprised of hydrology, terrain, and landform factors (including

![FIGURE 1 Geospatial location of the study area, flood locations and occurrence frequency from 1984 to 2017 from climate center of Xinjiang meteorological bureau](image)
FIGURE 2  Evaluation indices of the random forest assessment model for SFD in Xinjiang
vegetation). The hydrology factor uses the dimensionless runoff curve number (CN) to describe interception, infiltration, and surface storage processes of the water. As one of the comprehensive index indicating surface characteristics before rainfall, CN is related to soil moisture, terrain slope, vegetation type, soil type, and land use status. It is utilised to predict direct runoff or infiltration caused by excessive precipitation. Its value, ranging from 0 to 100, depends on the characteristics of soil type, land use, and land cover in the drainage basin (Zeng, Tang, Hong, et al., 2017). The terrain factors include digital elevation model (DEM) data, elevation variation coefficient, relative elevation, and slope, where DEM is a 90 m terrain product derived from the shuttle radar topography mission (SRTM). By using ArcMap, DEM data was resampled, and the relative elevation, slope, and elevation coefficient of variation were also calculated to comprehensively reflect the local terrain features in a 1 km × 1 km area. Landform factors include vegetation coverage and multi-layer particle-size distribution (PSD; 0–30 cm, i.e., sand, silt, and clay content) in China. PSD is an essential physical property that affects many vital attributes of soil, and it has a specific relationship with flood propagation (Wei, Dai, Liu, Ye, & Yuan, 2012). Monthly MODIS vegetation index products (MOD13A3) from 2003 to 2017 were used in this study, and the annual maximum normalised difference vegetation index (NDVI), deriving by the maximum synthesis method (MVC), was used to estimate the vegetation coverage. This value ranges from −1 to 1.

- Social indices include agriculture, residential, industrial and commercial, transportation systems, shipping, public facilities, etc. This study mainly considered cultivated land area, GDP and population per kilometre, distance from rivers, and road network density. GDP and population data with a spatial resolution of 1 km in 2014 were sourced from the National Science and Technology Infrastructure: National Earth System Science Data Center. The cultivated land data were
mainly derived from the Xinjiang Economic Statistical Yearbook, 2014. The inverse distance weighted (IDW) method was used to interpolate a 1 km × 1 km raster data in ArcMap. Distance from a river and road density were calculated by using the ArcMap-Density module.

- Disaster indices: These indices quantitatively describe the occurrence of SFD. Based on the original disaster data from 1984 to 2017 provided by the Xinjiang Climate Center, a data set of historical occurrences of the disaster at the township level was compiled, and spatial distribution of SFD was also mapped in ArcMap (Figure 1).
- Disaster-causing indices: Precipitation intensity and frequency, are chosen as the primary triggers of SFD. More significant precipitation and precipitation variability may result in a higher possibility of causing a disaster. Rainstorm in Xinjiang was defined as a daily precipitation volume ≥24 mm. This study adopted the daily precipitation data from 105 meteorological stations in Xinjiang from 1961 to 2017. Based on the above data, the annual maximum precipitation, maximum monthly precipitation, and days with precipitation ≥25 mm were calculated. Spatial interpolation of the three was undertaken by using ArcMap to obtain 1 km × 1 km of grid data.

Finally, the 17 evaluation indices were projected to a unified coordinate system of WGS84 ellipsoidal projection, and their values were normalised. These data were resampled to a grid size of 1 km × 1 km in ArcGIS. The entire study area was divided into 4,901,312 grids, as shown in Figure 2.

3 | METHODS

3.1 | The flood disaster assessment model based on RF

Random forest (RF) is an integrated machine learning algorithm (Ensemble Learning), which was proposed by Leo Breiman (2001). By combining multiple weak classifiers, the final results were voted or taken as a mean, which made the results of the RF model have high accuracy and generalisation performance. RF collected a fixed number of samples from the training set by self-help (bootstrap) resampling technique, but after each sample was collected, the samples were put back. That is to say, the samples collected before are likely to continue to be collected after being put back. As a result, the content of each sample set's content is different, and the k classification tree is generated by this self-help method to form a random forest to achieve a random sample. In each random sampling of the Bagging, approximately 36.8% of the data in the training set were not sampled. This part of the data is called out-of-bag data (out of bag, OOB). These data are not involved in the fitting of the training set model, so they can be used to detect the model's generalisation ability.

Taking the SFD in Xinjiang as an example, the basic principle of RF is to assume that there are N levels of rainstorm and flood risk, and M number of characteristic factors. Firstly, this study used a bootstrap method to randomly select samples from the sample set to construct a new training sample set. Each time it resampled and picked the samples that were not chosen before creating out-of-bag (OOB) data as the test sample set. Secondly, the Mtry (Mtry ≤ M) number of feature factors was randomly selected from the M number of feature factors. Each classification tree was the most representative feature factor with the best classification ability, and each tree would grow to the maximum without any pruning. This process was repeated many times to generate Ntree classification trees. Finally, the generated Ntree was aggregated to obtain a random forest. Meanwhile, the mode of classification result was used as the classification result of the random forest, that is, the risk level of SFD.

In this study, the randomForest package in R language platform was implemented to construct an RF assessment model for zoning SFD in Xinjiang. The objective was to determine the optimal value of Ntree and Mtry.

3.2 | Sample selection and creation

Sample selection was an important step in developing an RF assessment model for SFD. Firstly, based on the natural breakpoint method, this study classified the number of historical disaster occurrences into four levels: high-risk areas, general-risk areas, low-risk areas, and non-risk areas. They represented the risk levels of the disaster. Positive samples were made based on 947 historical rainstorms and flood disasters in Xinjiang at the township level from 1984 to 2017. Secondly, after considering the geographical similarity and the effects of lakes and rivers, buffer zones of 50, 20, 5, and 2 km were created around the locations of historical disasters, rivers, lakes, reservoirs, and canals, respectively. The area outside buffer zones was the sampleable area for negative samples. A 1 km × 1 km grid was created with the fishnet tool in ArcMap. The non-disaster and no-risk zones would put in areas where negative samples were randomly collected under the same number of positive samples outside the buffer zone, that is, where disasters would not occur. To avoid accidents, these steps were repeated five times; thus random sampling was also repeated five times in the sampleable area of negative samples. Meanwhile, five different total samples (including training and test samples) were obtained after combining them with positive
samples. Finally, 1894 total samples were extracted under 17 evaluation indices by using ArcMap.

The natural breakpoint method is a statistical method of grading and classification based on the statistical distribution of numerical statistics, which could group similar values appropriately and maximise the differences among various classes. The calculation formula for this method is as follows:

\[ SSD_{i-j} = \frac{1}{N} \sum_{k=1}^{j} (A[k] - \text{mean}_{i-j})^2 (1 \leq i < j \leq N) \quad (1) \]

where \( A \) is an array, and its length is \( N \); and \( \text{mean}_{i-j} \) is the average value of each level.

3.3 | Model implementation and validation

The training sample is the core part of constructing the RF assessment model for SFD. Five sample groups were randomly divided under the R language platform. Among them, 70% of the sample was used as the training set for model construction, and the remaining 30% was used as the test set for testing the model (Rodrigues & Riva, 2014). Through establishing the relationship between storm flood risk and the indices, the RF model created the corresponding classification rules and classified the data to be measured. Eventually, this study got the risk zoning of rainstorm and flooding disaster of different regions in Xinjiang.

4 | RESULTS AND DISCUSSION

4.1 | Parameter sensitivity analysis

The application of the RF model is based on its training and testing models. As described in Section 3.1, the core problem of RF modelling is to determine the number of decision trees (Ntree), which defaults to 500, and the number of random variables (Mtry) for splitting the nodes, which defaults to \( \log N \) (\( N \) is the feature number). Low Ntree value may result in a high error rate will result; High Ntree value may result in a more complex model and decrease its efficiency.

4.1.1 | Determination of the Mtry value

In this study, five groups of samples were separately tested under the premise of a larger number of decision trees; it evaluated RF simulation accuracy when Mtry had different values. The specific steps were: (a) Set Ntree = 500. Since 17 evaluation indices were selected, set Mtry value from 1 to 17, respectively, and model training is performed to obtain 17 RF simulation accuracy; (b) Set Ntree = 1,000 and repeat step one. (c) Compare the RF simulation accuracy of five groups of samples when Ntree = 500 and Ntree = 1,000, and determine the optimal Mtry value according to its accuracy. The average and maximum accuracy of RF training for the five sample sets under the different Mtry values are shown in Figure 3. When Ntree was 500 or 1,000, with the increase of Mtry value, the accuracy of the model obtained by OOB unbiased estimation of five sample sets significantly increased and then slowly decreased. When Mtry was set to 6, simulation accuracy was the highest. Simulation accuracy slowly decreased after peaking but maintained an accuracy above 84%. Based on this result, the Mtry value was set to 6.

4.1.2 | Determination of the Ntree value

After the Mtry value was determined, the next step was to determine the optimal Ntree value. Here, the accuracy
of the RF model using various Ntree values with an interval of 100 was compared. Figure 4 shows the relationship between RF simulation accuracy of five sample sets and the number of decision trees for the different Ntree values. It is indicated that when Mtry = 6, the increase of the Ntree value resulted in continuously improving model accuracy. When Ntree = 1,000, the accuracy of the RF model under the five random sample sets was more than 84.5%. When the Ntree values >1,000, accuracy fluctuated around 84%. Therefore, the Ntree is set to 1,000 in this study.

4.2 | Feature ranking based on learning models (model-based ranking)

The frequency and intensity of flood disasters can be affected by many factors. Feature selection can not only reduce the number of features and achieve dimensionality reduction to empower the model and reduce the risk of overfitting, as well as enhance the relationship between a feature and feature value. After having processed the data described in Section 2.2, for example, outliers, missing values, and conversion data, it is necessary to determine whether all of the data could be input into the RF model as feature variables for training.

4.2.1 | Feature importance analysis

This study used the mean accuracy reduction method to evaluate the importance of the 17 evaluation indices. This method estimated the importance of a certain feature variable by analysing the increase of the OOB error after its noise is added. For variable X, if the OOB accuracy is greatly reduced after random noise is added, it means that the X has a significant impact on the sample classification result, that is, X has a relatively high degree of importance; if the accuracy increases after the noise are added, it means that X has a relatively low degree of importance. Figure 5 shows the importance of ranking of five random sample sets after RF calculations. According to the determination process of the Mtry and Ntree in Sections 4.1.1 and 4.1.2, when Mtry = 6 and Ntree = 1,000, random sample set four had the highest RF accuracy, where its training sample simulation accuracy was 84.98%, and the test sample prediction accuracy was 82.43%. As shown in Figure 5, the importance ranking of 17 variables under five sample sets was consistent. To quantitatively evaluate the importance of each index, this study used a sample set four as an example (Figure 6).

The disaster-prone environment jointly influences the rainstorm and flood disasters in Xinjiang (altitude, slope, geomorphologic features), disaster-causing factors (extreme precipitation), and vulnerable disaster-bearing bodies (population and GDP). Their formation depends not only on the intensity and frequency of disaster-causing factors, but also on the natural environment and socio-economic background. The influence degree of different disaster-bearing bodies is different under the action of the same meteorological disaster-causing factors. Figure 6 represents the ranking of the evaluation factors of flood risk grade in Xinjiang based on the mean accuracy reduction method.

Results in Figure 6 show that:

1. Latitude, longitude, agricultural acreage, road density, maximum monthly precipitation, and distance from the river, these six are the most important explanatory factors, which records >40% of the contribution. In particular, latitude and longitude are as high as 68.21% and 56.42%, respectively, indicating that
FIGURE 5  The importance ranking of the feature variables of random forest model for zoning SFD risk in Xinjiang, based on the mean accuracy reduction method (sample sets 1–5)

FIGURE 6  Importance ranking of feature variables in the random forest model for zoning SFD risk in Xinjiang based on the mean accuracy reduction method (sample set 4)

Geographical location has the most significant impact on and contribution to the flood occurrence.

2. Maximum monthly precipitation (41.37%), annual maximum precipitation (35.87%) and days with precipitation ≥25 mm (33.84%) have a relatively significant impact on the occurrence of SFD in Xinjiang. Three types of precipitation indices reveal the precipitation intensity and frequency distribution of the
region. The more days with daily precipitation \( \geq 25 \text{ mm} \), the more frequent of heavy rain events; the higher average precipitation, the stronger intensity of local heavy rain events. They are related to higher risks of rainstorms.

3. The contribution of population and GDP per kilometre is 36.64% and 38.18%, respectively. In general, the magnitude of damage caused by floods depends on the economic value. Population and GDP indicate the level of urbanisation of this area: the larger the GDP and denser the population, the more significant damages flooding may cause. Floods of the same magnitude occur in economically developed and densely populated areas, leading to far more significant losses than those occurring in less-developed and sparsely populated areas.

4. Among topographic and hydrological factors, the contribution of elevation to the occurrence of flood disasters is the greatest, which is 32.83%. The coefficients of elevation variation, slope, and relative elevation that characterise topographic changes are 23.81%, 22.48%, and 22.90%, respectively. Results also show that multi-layered soil particle-size distribution (23.47%) also has a specific impact on rainstorm and flooding events. CN has the smallest impact on the risk of these disasters (13.3%). Elevation and slope determine runoff formation time and its volume. What is more, the topography is closely connected to the degree of flood risk, which reveals that the lower the terrain elevation, the smaller the change in terrain; the higher the absolute elevation, the smaller the standard deviation of relative elevation, and more possibilities of triggering flood disaster. In mountainous regions, surface runoffs are merged and route into the valley, which is not easy to form a flood of the vast area; in a flat region, runoffs are prone to exceed the local discharge capacity, which in turn results in large-area water accumulation and flooding.

5. The contribution of vegetation coverage records 37.67%. Because of its good soil and water conservation, the higher the regional vegetation coverage, the lower risk of flood disaster.

4.2.2 Feature rationality analysis

To further test the rationality of these evaluation indices, 17 indices were randomly combined and were input into the RF model separately. According to the parameter sensitivity analysis (Section 4.1), this study set \( \text{Mtry} = 6 \) and \( \text{Ntree} = 1,000 \), and five sample sets were analysed respectively. As more evaluation indices were introduced, out-of-bag error rates (OOB error) for the five sample sets under different evaluation indices generally decreased (Figure 7). After considering more explanatory indices, the efficiency of the RF model continued to increase; when all evaluation indices were input into the RF model, OOB error gained the smallest value, indicating that flood risk is an overall result of flood threats and socioeconomic vulnerability. The magnitude and extent of a disaster’s losses depend on such points as disaster-causing factors, disaster intensity, economy, population density in affected areas, and so on.
4.3 Model accuracy analysis

In the study, the percentage of training sample set classified by the RF model was used to measure the fitting degree of the model and the set (Table 1), and the accuracy of the RF model was then evaluated. As can be seen from Table 1, the average accuracy of the training samples for the RF model constructed by five different sample sets was 84.75%, and the average accuracy of the test samples was 83.48%. Five RF models all met the accuracy requirements for the risk zoning of rainstorm and flooding disaster in Xinjiang.

4.4 Zoning characteristics of SFD in Xinjiang

Here, 83,322,304 samples based on 17 evaluation indices around the study area were put into the RF model, and it got the map of risk zoning of the rainstorm and flooding disaster in Xinjiang (Figure 8). As shown in the figure, the risk of SFD is classified into four parts:

1. Key-risk area: It covers about 16% of the study area and is mainly distributed in mountainous watersheds, around major lakes and plains in the middle-lower reaches of the major river. They were an extensive flood risk zone with widespread influence, the Yili River Valley, the Wulun River Basin, and the Kaidu River Basin show different characteristics. In the Yili River Valley, fast-flowing runoffs formed in high and steep areas converge into the river valley. Subsequently, when water comes to a relatively flat area at a step junction, flooding would quickly occur. The risk zones with a relatively small impact are concentrated in such areas as the Kuitun River Basin, the Manas River Basin, the Hutubi River Basin, the Urumqi River Basin, and the foothill belt of the Xiaohe Area on the northern slope of the Tianshan Mountains. These areas with vast river network are densely-populated economic zone in Xinjiang, where its industries are relatively concentrated, and agriculture is well developed. Once flood occurs, the high urbanisation areas would suffer huge losses. Therefore, they have the highest risk level. Based on flood statistical data from 1984 to 2018, all these areas mentioned above have occurred severe or relatively severe floods.

2. General-risk area: It covers about 11% of the study area and mainly distributes in the Irtysh River Basin in northern Altay. Snowmelt floods in this area may happen in spring. Due to poor flood drainage infrastructure, flooding events are easily formed in the mountainous fields of the Emin River Basin (Tacheng Area), Zhuxiaoho River Basin, and the Kashgar River Basin (Hami Area). Currently, because of the stepping up of urbanisation and increase of population density, flooding disaster leads to relatively high losses in these areas.

3. Low-risk area: It covers about 6% of the study area and is mainly distributed in the Bortala River Basin,
Emin and Baiyang River Basin, Manas, Hutubi and Aksu River Basin, Kuitun River Basin on the northern slope of the Tianshan Mountains, and the plain area of the lower reaches of the major rivers in the Gar river basin. The population in these areas is relatively sparse, and the economic development is on the average level. Therefore flood leads to relatively low losses.

4. No-risk area: It covers about 67% of the study area and mainly distributes in the desert of the Junggar Basin, the desert of the Tarim Basin, the Balikutun-Yiwu Basin, and the Qiangtang and Cheercheng River Basin in the northern foot of the Kunlun Mountains. Surrounded by mountains and plateaus which blocked humid air from the ocean, The Junggar Basin and the Tarim Basin are the driest hydrological regions in China. Scarce precipitation and strong evaporation in these areas result in fewer river networks. Thus, there is no flooding in these areas. Because of the rare visits and underdeveloped economy, these areas show the lowest level of flood risk.

5 | CONCLUSIONS

Based on rainstorm and flood disaster data of Xinjiang from 1984 to 2017 at the township level and on basic natural, social and risk-related information, this study developed an RF-based risk assessment model for assessing the risk of SFD in Xinjiang. This study constructed a 1 km × 1 km rainstorm and flood disaster zoning map for Xinjiang, and the following conclusions are drawn:

1. Combined effects of multiple indices lead to SFD in Xinjiang. The RF explanatory variables indicate that geographical location has the greatest impact on the occurrence of SFD, and floods have significant spatial variability. Latitude, longitude, agricultural acreage, roading density, distance from the river, and the maximum monthly precipitation have great effects on triggering disasters. In addition, precipitation is the disaster-causing factor. The heavier precipitation and greater its variability, the more likely to result in a disaster. In areas mainly dominated by agriculture, damages of farmland are a major loss. In areas with relatively developed industries, damages of the property based on their dense population, road network, and asset distribution are severe. Curve number has the smallest impact on the risk of SFD, but its contribution is also more than 10%, which reflects different conditions of runoff on the underlying surface.

2. RF model describes the relationship between the number of flood events and evaluation factors, and it has a relatively high classification accuracy. Two random sampling methods, random sampling based on samples and on features, are introduced in the RF model. This improves the accuracy and stability of the model, reduces the sensitivity of noise and abnormal value, and effectively avoids overfitting. With more evaluation indices involved in, OOB error generally shows a downward trend. With more explanatory factors considered, the efficiency of the RF model steps up, eventually reaching an average simulation accuracy of 83.48%. All 17 indices analysis indicates that flood risk is a comprehensive result of its threats and socio-economic vulnerability. Besides, the magnitude and degree of disaster losses depend on many aspects, such as hazard factors, catastrophic intensity, economic development, and population density of the affected area.

3. The results of the RF model indicate that SFD in Xinjiang has an uneven and complex spatial distribution. High-risk areas are concentrated in mountainous basins, plains of the middle-lower reaches of the major rivers, and areas surrounding major lakes. Different watersheds have diverse characteristics and mainly show a spatial feature that there are more lowlands in plains and less mountainous fields in plateaus.

ACKNOWLEDGEMENTS

Our recent work has been funded by the science & Technology Basic Resources Investigation Program of China 2017FY100503, the National Natural Science Foundation of China 42071075, the National Natural Science Foundation of China - Xinjiang Joint Fund U1703121, the Natural Science Foundation of of Qinghai Province under grant 2020-ZJ-731, the Strategic Priority Research Program of Chinese Academy of Sciences XDA20100306, and the National Natural Science Foundation of China 41975146.

Before that, we are thankful to the science team of NASA for the DEM data (http://reverb.echo.nasa.gov/reverb/), the Xinjiang Meteorology Center for the observational meteorological data and storm flood disasters data and acknowledgement for the data support from “National Earth System Science Data Center, National Science & Technology Infrastructure of China (http://www.geodata.cn).”

CONFLICT OF INTEREST

The authors declare no conflict of interest.
DATA AVAILABILITY STATEMENT
The data used to support the findings of this study are available from the corresponding author upon request.

ORCID
Yan Liu https://orcid.org/0000-0002-9775-1913

REFERENCES
Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:101093304324

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6), 323–329. https://doi.org/10.1016/j.ygeno.2012.04.003

Chen, Z. (2014). Study on cause and prevention technology of mountain torrent disaster in small watershed (in Chinese). Guangzhou: South China University of Technology.

Ding, Y., & Zhang, J. (2009). Rainstorm and flood (in Chinese). Beijing: China Meteorological Press.

Dong, L. J., Xi-Bing, L. I., & Peng, K. (2013). Prediction of rockburst classification using random Forest. *Transactions of Nonferrous Metals Society of China*, 23(2), 472–477. https://doi.org/10.1016/S1003-6326(13)62487-5

Duan, W., He, B., Nover, D., Fan, J., Yang, G., Chen, W., & Liu, C. (2016). Floods and associated socioeconomic damages in China over the last century. *Natural Hazards*, 82(1), 401–413.

Feng, Q., Liu, J., & Gong, J. (2015). Urban flood mapping based on unmanned aerial vehicle remote sensing and random Forest classifier - A case of Yuyao, China. *Water*, 7(4), 1437–1455. https://doi.org/10.3390/w7041437

Gan, W., Fan, J., Xiao, T., Jing, F., zhan, Z., & Wang, L. (2017). Research and establishment of a comprehensive evaluation model of rainstorm process in Sichuan Province (in Chinese). *Journal of Glaciology and Geocryology*, 39(6), 1336–1344.

Guan, M., & Chen, X. (2007). Research of regional torrent risk zonation in Jiangxi Province. *Torrential rain and Disaster*, 26(4), 339–343.

Guan, Y., Wang, B., Zheng, F., & Zhang, P. (2014). Characteristics of meteorological disasters in China during 2012. *Journal of Natural Disasters*, 23(1), 24–31.

Haruyama S., Ohokura H., Simking T., & Ramphín. (1996). Geo-morphological zoning for flood inundation using satellite data. *Geo Journal*, 38(3), 273–278.

Islam, M. M., & Sado, K. (2000). Flood hazard assessment in Bangladesh using NOAA AVHRR data with geographical information system. *Hydrological Processes*, 14(3), 605–620.

Jiang, F., Hu, R., & Yang, Y. (2004). Abrupt change in the time sequences of flood disaster in Xinjiang and its possible climatic reasons. *Journal of Glaciology and Geocryology*, 26(6), 674–681.

Lai, C., Chen, X., Zhao, S., Wang, Z., & Wu, X. (2015). A flood risk assessment model based on random forest and its application. *Journal of Hydraulic Engineering*, 46(01), 58–66.

Li, Y. (2003). Change of river flood and disaster in Xinjiang during past 10 years[J]. *Journal of Glaciology and Geocryology*, 25(3), 312–316.

Liu, J., Li, J., Liu, J., & Cao, Y. (2008). Integrated GIS/AHP-based flood risk assessment: A case study of Huaihe River basin in China[J]. *Journal of Natural Disaster*, 17(6), 110–114.

Ma, Z. (1994). *Major natural disasters and the countermeasures of disasters mitigation in China* (pp. 11–21). Beijing: Science Press.

Mihăilescu, D. M., Gui, V., Toma, C. I., Popescu, A., & Sporea, I. (2013). Computer aided diagnosis method for steatosis rating in ultrasound images using random forests. *Medical Ultrasonography*, 15(15), 184–190. https://doi.org/10.11152/mu.2013.2066.153.dmm1vg2

Qi, Y., Liu, Y., Yang, Z., Xu, Z., & Fang, M. (2012). GIS-based analysis of landslide and debris flow hazard in Lanzhou. *Journal of Glaciology and Geocryology*, 34(1), 96–114.

Qin, D., & Stocker, T. (2014). Highlights of the first working group report of the fifth assessment report of the IPCC. *Progress in Climate Change Research*, 10(01), 1–6.

Rodrigues, M., & Riva, J. D. L. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software*, 57, 192–201. https://doi.org/10.1016/j.envsoft.2014.03.003

Sanyal, J., & Lu, X. X. (2005). Remote sensing and GIS-based flood vulnerability assessment of human settlements: A case study of Gangetic West Bengal, India. *Hydrological Processes*, 19(18), 3699–3716. https://doi.org/10.1002/hyp.5852

Shi, Y., Shen, Y., Kang, E., Li, D., Ding, Y., Zhang, G., & Hu, R. (2007). Recent and future climate change in Northwest China. *Climatic Change*, 80(3/4), 379–393.

Su, J., Xiao, Z., Wei, B., Zhao, Y., Zhang, F., & Liu, L. (2012). The risk zoning of rainstorm based on GIS in Lanzhou of Gansu Province. *Journal of Arid Meteorology*, 30(4), 650–655.

Su, L., Song, Y., & Zhang, Z. (2005). Response of climate and ecological environment to global warming in the northern slope of Tianshan Mountains in recent 40 years. *Arid Land Geography*, 03, 342–346.

Tang, C., & Zhu, J. (2005). A GIS based regional torrent risk zonation. *Acta Geographica Sinica*, 60(1), 87–94.

Tesfamariam, S., & Liu, Z. (2010). Earthquake induced damage classification for reinforced concrete buildings. *Structural Safety*, 32(2), 154–164.10.1016/j.strusafe.2009.10.002.

Wan, X., Yin, W., Sun, P., Tan, B. I. (2017). Risk division assessment of rainfall-flood disasters based on cloud model. *Journal of Natural Disasters*, 26(4), 77–83.

Wang, H., Mao, W.-Y., Li, Y.-P., Yu, X.-J., & Bai, S.-q. (2011). Risk assessment of torrential rain which easily causes damage in Xinjiang region. *Journal of Glaciology and Geocryology*, 33(6), 1407–1413.

Wang, Q., Cui, c., & Yao, Y. (2008). Variation trends and multi-scale analysis of flood affected area in various regions of Xinjiang. *Acta Geographica Sinica*, 63(7), 769–779.

Wang, Y., Huang, X., Liang, H., Sun, Y., Feng, Q., & Liang, T. (2018). Tracking snow variations in the northern hemisphere using multi-source remote sensing data (2000-2015). *Remote Sensor*, 10, 136.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 27, 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008

Wei, S., Dai, Y., Liu, B., Ye, A., & Yuan, H. (2012). A soil particle-size distribution dataset for regional land and climate modeling in China. *Geoderma*, 171-172, 85–91. https://doi.org/10.1016/j.geoderma.2011.01.013

Wu, S., & Zhang, G. (2003). Preliminary approach on the floods and their calamity changing tendency in Xinjiang region. *Journal of Glaciology and Geocryology*, 25(2), 199–203.
Xie, W., Tian, H., & Lu, Y. (2015). Risk evaluation of rainstorm and flood disasters in Datong river basin based on the flood area model. *Torrential Rain and Disaster, 34*(4), 384–387.

Xu, Y., Zhang, B., Zhou, B. T., Dong, S. Y., Yu, L., & Li, R. K. (2014). Projected risk of flooding disaster in China based on CMIP5 models. *Progresses Inquisitions de Mutatione Climatis, 10*(4), 268–275 (in Chinese).

Xu, Y.-h., & Lu, Y.-y. (2007). Study on countermeasures for disaster prevention and mitigation of floods and droughts in Xinjiang in the early 21st century. *Desert and Oasis Meteorology, 05*, 54–58.

Xu, Y., Mao, W., & Lu, Y. (2008). A comprehensive review of recent changes in meteorological disasters in Xinjiang and disaster prevention and mitigation. *Desert and Oasis Meteorology, 01*, 50–54.

Zeng, Z., Tang, G., Hong, Y., Zeng C., & Yang Y. (2017). Development of an NRCS curve number global dataset using the latest geospatial remote sensing data for worldwide hydrologic applications. *Remote Sensing Letters, 8*(6), 528–536. https://doi.org/10.1080/2150704X.2017.1297544

Zhao, S., Gao, G., Huang, D., & He, W. (2017). Analysis of the characteristics of meteorological disasters in China from 2004 to 2013. *Journal of Meteorology and Environment, 33*(01), 101–107.

Zhao, T., Tie, G., Yang, D., Cai, X., & Cao, Y. (2012). Study on runoff forecast of the upper reaches of the Yangtze River in dry season based on random Forest model. *Journal of Hydroelectric Engineering, 31*(03), 18–24.

**How to cite this article:** Liu Y, Lu X, Yao Y, et al. Mapping the risk zoning of storm flood disaster based on heterogeneous data and a machine learning algorithm in Xinjiang, China. *J Flood Risk Management*. 2021;14:e12671. https://doi.org/10.1111/jfr3.12671