A Comparison of Global and Local Statistical and Machine Learning Techniques in Estimating Flash Flood Susceptibility

Jing Yao
Urban Big Data Centre, School of Social and Political Sciences, University of Glasgow, UK

Ziqi Li [1]
Department of Geography, Florida State University, Tallahassee, FL, USA

Xiaoxiang Zhang
Department of Geographic Information Science, College of Hydrology and Water Resources, Hohai University, Nanjing, China

Changjun Liu
Department of Flood and Drought Disaster Reduction, China Institute of Water Resources and Hydropower Research, Beijing, China

Liliang Ren
State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, College of Hydrology and Water Resources, Hohai University, Nanjing, China

Abstract

Flash floods, as a type of devastating natural disasters, can cause significant damage to infrastructure, agriculture, and people's livelihoods. Mapping flash flood susceptibility has long been an effective measure to help with the development of flash flood risk reduction and management strategies. Recent studies have shown that machine learning (ML) techniques perform better than traditional statistical and process-based models in estimating flash flood susceptibility. However, a major limitation of standard ML models is that they ignore the local geographic context where flash floods occur. To address this limitation, we developed a local Geographically Weighted Random Forest (GWRF) model and compared its performance against other global and local statistical and ML alternatives using an empirical flash floods model of Jiangxi Province, China.

2012 ACM Subject Classification Computing methodologies → Machine learning

Keywords and phrases Machine Learning, Spatial Statistics, Flash floods, Susceptibility

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.86

Category Short Paper

Funding This work was supported by the National Key Research and Development Program of China [grant numbers No. 2019YFC1510601]; the Economic and Social Research Council, UK [grant number ES/P011020/1, ES/S007105/1].

1 Introduction

Flash floods are one of the most devastating natural disasters, which often occurs within a short period of time and can be caused by a variety of factors such as intense rainfall, rapid snow melt, landslides, and dam failure. Given their rapid speed and strong force, flash floods can cause significant damages to properties, infrastructures, and even loss of life. As a result, flash flood risk mitigation and management are of fundamental importance if sustainable

[1] Corresponding author

12th International Conference on Geographic Information Science (GIScience 2023).
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 86; pp. 86:1–86:6
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

development is to be achieved. Flash flood susceptibility estimation has long been an effective means adopted by practitioners and policymakers to assist with development of flood risk reduction strategies, land use planning and emergency resource deployment [6] [8].

Common approaches that have been widely adopted in the estimation of flash flood susceptibility include statistical, hydrodynamic models and geographical information system (GIS) based spatial analyses. The examples of statistical models include regression analysis, frequency ratio, weights-of-evidence, and analytical hierarchy process, among others [6]. Hydrodynamic models usually predict the propensity of an area to flash floods by simulating the water flow during a rainfall event [13]. GIS-based approaches often combine potential factors that contribute to flash floods (e.g., rainfall topography and land use) to identify areas at risk, mainly utilizing remote sensing images [12].

In recent years, with the emergency of big data (e.g., weather and water levels) collected by various sensors as well as the advances in high-performance computing techniques, artificial intelligence (AI) particularly machine learning (ML) has been increasingly applied in evaluating and predicting flash flood susceptibility [9]. Common ML approaches such as support vector machine (SVM), random forest (RF), neural network (NN) have demonstrated better performance than traditional methods like statistical and hydrodynamic models [8] [2] [1]. However, a major limitation of existing ML approaches is that they ignore the geographic nature of flash floods. Often, the same set of hyperparameters are employed for all observations without considering the geographic context of each flash flood event. It is worth mentioning that there have been several recent developments in GeoAI that incorporate spatiality into modeling.

[3] developed Geographical Random Forests (GRF), in which a separate RF model is fitted for each location. One limitation of GRF is that, although it considers the local nature of the phenomenon, it does not allow geographical weighting in the training, which ignores the distance-decay effect for most geographical processes. [4] improved GRF to incorporate geographical weighting, but the prediction process for unseen data is less explicit and does not allow the weighting kernel to vary spatially. [5] developed a Geographically Weighted Neural Network (GWNN) model, in which geographical weighting is imposed on the loss function during model training. However, GWNN does not allow hyperparameters to vary spatially, thus failing to account for local variations in the underlying processes.

To this end, in this paper, we address limitations in recent GeoAI developments by allowing geographical weighting in model training and prediction as well as allowing hyperparameters, which include both the model hyperparameters and the bandwidth parameter that controls the geographical weighting, to vary spatially. In this regard, both complex spatial and non-spatial processes can be fully considered. We use a random forest model as an example of this generic local modelling framework, which can be naturally extended to other popular models such as neural networks and gradient boosting, for both regression and classification tasks. We benchmark its performance against other global and local statistical and ML alternatives with an empirical flash flood model of Jiangxi Province, China.

2 Methods

Four models are included in comparison to predict a binary flash flood occurrence: 1) logistic regression (LR); 2) geographically weighted logistic regression (GWLR); 3) random forest (RF) and 4) geographically weighted RF (GWRF), They represent the four quadrants of model (as shown in Table 1)) types consisting of global/local and statistical/ML, respectively.

Table 1 Four model types.

Model Type	Global	Local
Statistical	Logistic Regression (LR)	GW Logistic Regression (GWLR)
Machine learning	Random Forest (RF)	GW Random Forest (GWRF)

Listing 1 GWRF Algorithm

```
For each location in all locations:
    1. Find a set of hyperparameters and local bandwidth that
       minimises geographically weighted loss with a 5-fold cross
       validation;
    2. Train the local RF model using the best set of
       hypterparameters and local bandwidth;
    3. Use the local RF to predict at any unknown locations weighted
       by its distance away from  unknown locations;

Sum of all the distance weighted predictions to be the final
    predictions.
```

LR is a global statistical model used to predict binary outcomes. It's a linear model with a logit link function that transforms continuous outcomes into probabilities bounded between 0 and 1. GWLR is a local statistical approach that accounts for location-specific effects when generating the outcome of interest. It fits a geographically weighted logistic regression model at each location using a distance decay kernel governed by a kernel function and kernel bandwidth. This approach allows for parameters in the model to vary spatially. RF is a machine learning algorithm that utilizes ensemble learning methods to make predictions by combining multiple decision trees. While RF is widely used in various applications due to its flexible and accurate predictions, it's considered a global model since the same hyperparameters that govern the tree structure remain constant regardless of geographic location. The last model GWRF is the proposed approach. It trains a separate local RF model at each location allowing different hyperparameters for the RF model and bandwidth for geographical weighting. Each local RF is optimised using a geographically weighted loss function. Then the prediction at an unseen location can be computed as the distance weighted predictions from all RFs. The specific training and prediction process are described as follows:

LR and RF are implemented using the sklearn python package [11], GWLR is fitted using the mgwr python package [10], and GWLR is implemented using both sklearn and mgwr. Code and data that produce the results can be found at this repository: `https://anonymous.4open.science/r/global_local_ML_GIScience-48F9`.

3 A Case Study of Jiangxi Province, China

3.1 Study area

The case study area is Jiangxi, a province in south-eastern China. Jiangxi has long been one of the places suffering flash floods every year in China, which is primarily due to its unique geography and climate. It is located in a mountainous region with over 3,000 rivers and lakes, which accounts for 78% of the total area. The largest freshwater lake in China, Poyang Lake, is located in the north of the province. Further, Jiangxi is in a subtropical climate zone and experiences a high amount of rainfall during the monsoon season from May to September. Flash flood risk reduction and management is a major challenge to local government with respect to sustainable development. In addition to dams and other flood control infrastructure, mapping flash flood susceptibility has become an effective measure to assist with land use planning as well as to improve public knowledge of flash floods.

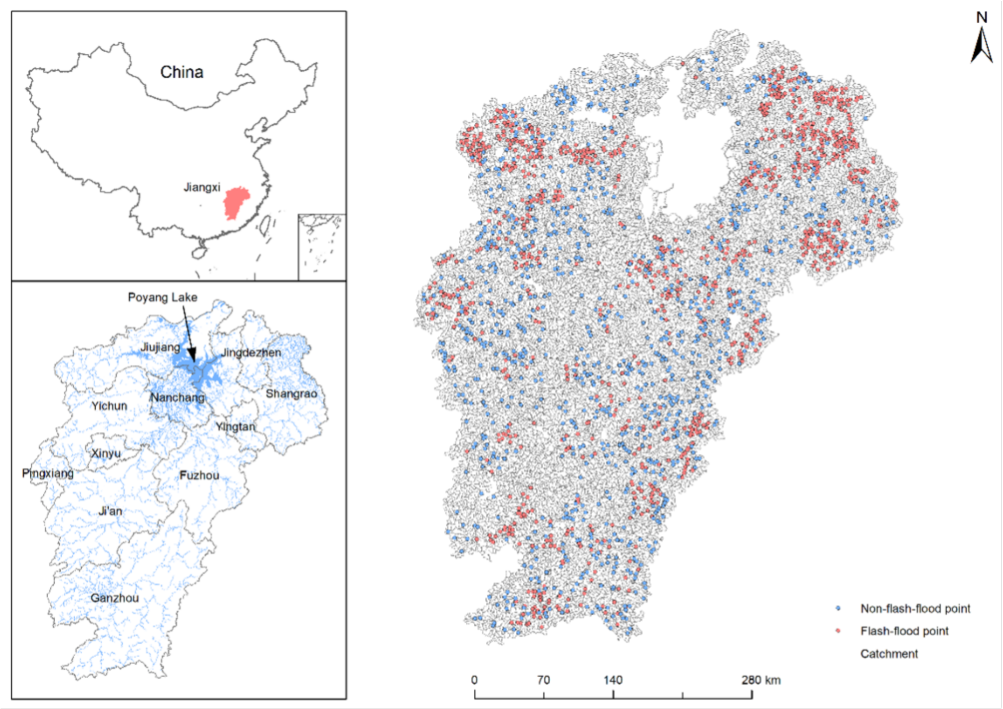

Figure 1 Historical flash flood events in Jiangxi Province, China.

3.2 Data

The main dataset used in this research is the flash flood inventory map provided by the Flood Control and Drought Relief Division, Emergency Management Department of Jiangxi, which contains historical flash floods in Jiangxi during 1950-2015. Among 12,388 catchments within the province, 940 contain historical flash flood events. Accordingly, 971 catchments without historical flash floods are randomly selected across space. The final dataset contains 1,911 observations labelled either 1 (flash floods) or 0 (non-flash floods). The resulting flash floods distribution map can be seen in figure 1.

In addition, four ancillary datasets are used to derive potential factors that contribute to flash floods, including the DEM dataset of China (2014), Statistical Parameter Atlas of Rainstorms in China (2010), River System in China (2012) and the Landsat 7 Collection 1 Tier 1 Annual NDVI Composite. Based on those datasets, 10 influencing factors are calculated or extracted: slope, elevation, shape factor, concentration gradient, topographic wetness index, rainfall, peak discharges per unit area, time of concentration, normalized difference vegetation index (NDVI) and distance to the nearest river, which are selected based on previous studies and data availability.

3.3 Results

The dataset was split 80/20, with 20% of the unseen data being used for out-of-sample accuracy assessment, the results of which are shown in Table 2. Three accuracy measures are included:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Precision} = (\text{TP}/\text{TP} + \text{FP})$$

where TP is True Positive; TN is True Negative; FP is False Positive; and FN is False Negative.

Table 2 Accuracy, recall and precision for four models.

Model	Accuracy	Recall	Precision
LR	0.70	0.80	0.66
GWLR	0.75	0.60	0.86
RF	0.81	0.65	0.96
GWRF	0.85	0.76	0.91

Regarding the overall accuracy of models, local models have been observed to have approximately a 5% advantage over their global counterparts. This suggests that allowing parameters to vary spatially can lead to an increase in model accuracy. Furthermore, machine learning (ML) approaches have been found to be approximately 10% more accurate than statistical approaches, indicating that complex non-linear and interaction effects are present and can be captured by ML but not by statistical approaches. The proposed GWRF, which allows for non-linearity, interaction, and spatial heterogeneity, has emerged as the best-performing model, achieving a promising overall accuracy of 85%. Additionally, the GWRF model demonstrates the second-highest precision and recall, resulting in a more well-rounded and balanced performance in estimating flash flood occurrences.

4 Summary

Flash floods can pose significant threats to the environment, properties, and life. Recent advances in AI particularly ML techniques provides new opportunities for assessing and estimating the susceptibility of flash floods – an effective measure that can help with designing flash flood risk reduction strategies. This research develops a novel Geographically Weighted Random Forest (GWRF) within a generalisable local ML framework and compares against other local and global statistical and machine learning approaches in estimating flash flood susceptibility. The preliminary results show that GWRF has the best performance among others with higher accuracy and more balanced precision and recall. The initial findings suggest the importance of incorporating geographic space into ML approaches to improve model performance. However, one drawback of ML is its black-box nature, which limits interpretability. The recent development of eXplainable AI methods (XAI) offers opportunities to estimate the effects of ML models and has been demonstrated to be effective when modeling spatial data [7]. The next step of this research is to investigate the explainability of the ML model to explore spatial and non-spatial relationships, enhancing better understanding of flash flood processes.

References

1 Jialei Chen, Guoru Huang, and Wenjie Chen. Towards better flood risk management: Assessing flood risk and investigating the potential mechanism based on machine learning models. *Journal of environmental management*, 293:112810, 2021.

2 Romulus Costache, Haoyuan Hong, and Quoc Bao Pham. Comparative assessment of the flash-flood potential within small mountain catchments using bivariate statistics and their novel hybrid integration with machine learning models. *Science of The Total Environment*, 711:134514, 2020.

3 Stefanos Georganos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholus Mboga, Eléonore Wolff, and Stamatis Kalogirou. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, 36(2):121–136, 2021.

4 Stefanos Georganos and Stamatis Kalogirou. A forest of forests: A spatially weighted and computationally efficient formulation of geographical random forests. *ISPRS International Journal of Geo-Information*, 11(9):471, 2022.

5 Julian Hagenauer and Marco Helbich. A geographically weighted artificial neural network. *International Journal of Geographical Information Science*, 36(2):215–235, 2022.

6 Khabat Khosravi, Hamid Reza Pourghasemi, Kamran Chapi, and Masoumeh Bahri. Flash flood susceptibility analysis and its mapping using different bivariate models in iran: a comparison between shannon's entropy, statistical index, and weighting factor models. *Environmental monitoring and assessment*, 188:1–21, 2016.

7 Ziqi Li. Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost. *Computers, Environment and Urban Systems*, 96:101845, 2022.

8 Meihong Ma, Changjun Liu, Gang Zhao, Hongjie Xie, Pengfei Jia, Dacheng Wang, Huixiao Wang, and Yang Hong. Flash flood risk analysis based on machine learning techniques in the yunnan province, china. *Remote Sensing*, 11(2):170, 2019.

9 Amir Mosavi, Pinar Ozturk, and Kwok-wing Chau. Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536, 2018.

10 Taylor M Oshan, Ziqi Li, Wei Kang, Levi J Wolf, and A Stewart Fotheringham. mgwr: A python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information*, 8(6):269, 2019.

11 Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

12 Binh Thai Pham, Mohammadtaghi Avand, Saeid Janizadeh, Tran Van Phong, Nadhir Al-Ansari, Lanh Si Ho, Sumit Das, Hiep Van Le, Ata Amini, Saeid Khosrobeigi Bozchaloei, et al. Gis based hybrid computational approaches for flash flood susceptibility assessment. *Water*, 12(3):683, 2020.

13 Seann Reed, John Schaake, and Ziya Zhang. A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *Journal of hydrology*, 337(3-4):402–420, 2007.