Machine Learning Framework for Flood Susceptibility Modeling in a Fast-Growing Urban City of Southern India

A. L. Achu, Girish Gopinath, and U. Surendran

Abstract Flooding in urban areas often results severe loss of life and property and has many negative socio-economic impacts. Therefore, identifying the flood prone areas is necessary for future flood hazard mitigation, early warning, and land use planning for infrastructure developments in urban areas. In this study, flood susceptibility modeling is carried out for Kozhikode urban and per-urban area, which is severely affected by 2018 Kerala flood. To begin with, a flood inventory map is prepared with 307 flood location points marked immediately after 2018 flood. Thereafter, the inventory is randomly classified into 70% for model training and remaining 30% for model testing. In addition, twelve independent variables such as land use/land cover, soil texture, lithology, elevation, slope angle, slope aspect, valley depth, topographical wetness index, profile curvature, plan curvature, convergence index, and channel network base level were prepared and used. Subsequently, final modeling is carried out using these flood conditioning factors and flood inventory locations using machine learning random forest method. The result shows that ~ 13.78% of the study area is very highly susceptible to the occurrence of flood. The predicted model shows 85.2% accuracy (ROC-AUC) in training phase and 78.5% in testing phase. Therefore, the model is trustworthy and can be used for future hazard mitigation and land use planning in Kozhikode urban and per-urban area.

Keywords Flood · Machine learning · Kozhikode · Kerala

A. L. Achu (✉) · G. Gopinath
Department of Climate Variability and Aquatic Ecosystems, Kerala University of Fisheries and Ocean Studies (KUFOS), Kochi, Kerala 682508, India
e-mail: achu.geomatics@gmail.com

G. Gopinath
e-mail: ggkufos@ac.in

U. Surendran
Land and Water Management Research Group, Centre for Water Resources Development and Management (CWRDM), Kozhikode, Kerala 673571, India
e-mail: suren@cwrdm.org

P. V. Timbadiya et al. (eds.), *Geospatial and Soft Computing Techniques*, Lecture Notes in Civil Engineering 339, https://doi.org/10.1007/978-981-99-1901-7_43

1 Introduction

Among the different natural calamities, floods are most frequent and affecting millions of peoples across the globe. Urban flooding is a global concern, and it does not just mean "the flooding that happens in an urbanized area." The Federal Emergency Management Agency (FEMA) report 2016 defines urban flooding as: the inundation of property in a built environment, particularly in more densely populated areas, caused by rain falling on increased amounts of impervious surfaces and overwhelming the capacity of drainage systems. Flooding causes huge loss to life and property across the world. Between 2011 and 2012 alone, floods affected around 200 million people and caused economic losses of about $95 billion. Hence, it is of paramount importance to manage floods and reduce their risk, which requires flood prediction and computation of inundation areas [6]. Flood is a complex phenomenon, and hence, predicting the same is difficult [13]. For predicting the probability of flood and for mitigating and managing future flood hazard, modeling flood susceptibility is an essential procedure [10]. To model the flood susceptibility, multi-sourced dataset is required.

With the development of remote sensing techniques, multi-temporal and multi-sourced data have been widely used to predict flood susceptibility with GIS techniques [6]. However, in recent times, with the introduction of the concept of big data analytics and machine learning, accuracy and reliability of flood susceptibility mapping are improved significantly. Many researchers used different machine learning techniques to assess the flood [7, 11, 18]. Methods including random forest [1], support vector machine [17], artificial neural network [3], logistic regression [14, 15] have been widely used for flood prediction. In 2018, Kerala witnessed extreme rainfall event caused huge flooding and numeroous landslides across the state. Kozhikode was one of the coastal cities which was severely affected by flooding during 2018. Absence of flood susceptibility map was one the reason for extended causality, and hence in this study, random forest method is used to model flood susceptibility in Kozhikode urban cluster area. The proposed study will be useful for future hazard mitigation.

2 Materials and Methods

2.1 Study Area

Urban clusters in the Kozhikode District on the southwest coast of India were chosen for the study. Kozhikode urban cluster (KUC) is the largest urban agglomeration in Malabar region (northern Kerala) with an area of 197 km^2. As per the census report 2011 [5], KUC is the second-order urban zone with a population density of 3746 persons/km^2 (State Urbanization Report—Kerala 2012), and it is projected to increase population density and infrastructure development in near future and [9].

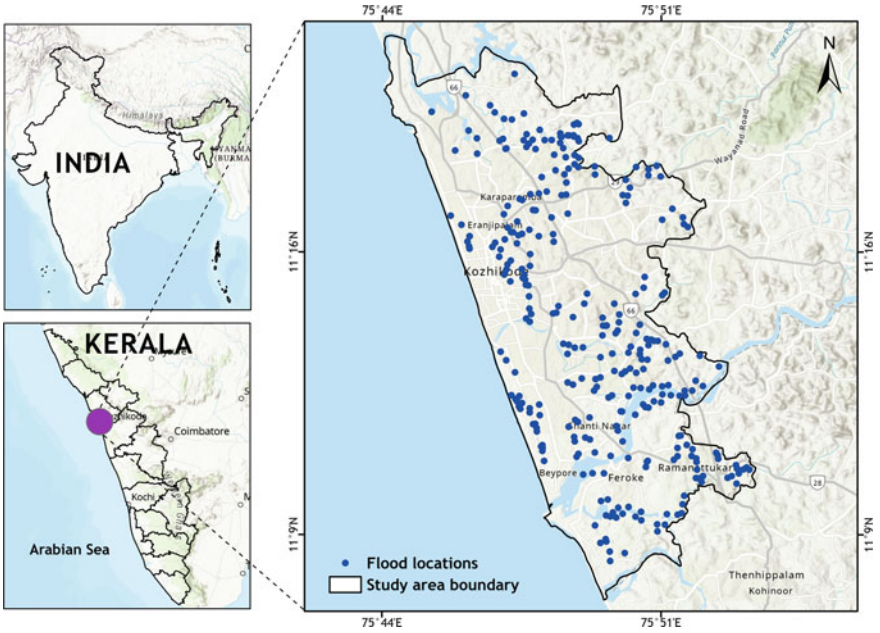

Fig. 1 Location map of the KUC with flood locations

Physiographically KUC is a part of Western coastal plains of Kerala with undulating topography lies between 11° 7′ 27.46″ N to 11° 21′ 17.91″ N latitudes and 75° 44′ 13.09″ E to 75° 52′ 9.11″ E longitudes (Fig. 1). During 2018 Kerala flood, KUC was affected severely, displacing millions. During June–August 2018, KUC received 2898 mm rainfall against its normal average of 2250 mm which caused intense flooding in river valleys and low-lying areas of KUC [16].

2.2 Spatial Database

Accurate and reliable flood inventory datasets are essential for flood susceptibility modeling [8]. In this study, flood inventory marking was carried out with intense field visits immediately after 2018 flood. Flooded areas were marked using handheld GPS and flooding height was also measured and attributed to the locations. A total of 307 flood locations were marked which range 0.11 m–2.48 m and used in this study. The locations were randomly divided into 70–30% for model building and model testing. Besides, a ten-fold cross-validation is implemented to avoid over fitting.

The construction of flood susceptibility modeling is a complex decision-making process which involves many geo-environmental variables [8]. In this study, twelve geo-environmental variables including elevation, slope angle, lithology, soil texture, land use/land cover, slope aspect, Topographic Wetness Index (TWI), Valley depth,

Channel network base level (CNBL), Convergence Index, Plan curvature, and Profile curvature were selected on the basis of expert opinion and literature [12, 18, 19].

SRTM Digital Elevation Model (DEM, 30 M) is used to represent elevation of the study area, and other DEM derivatives such as slope angle, slope aspect, TWI, valley depth, CNBL, CI, plan curvature, and profile curvature were derived. The KUC is a low-lying area where elevation ranges from ~ 0 to 90 m above mean sea level (Fig. 2a). Slope angle is an important parameter which determines flow velocity and concentration. KUC is a gently sloping terrain where slope angle ranges from 0 to 20.75° (Fig. 2b). Conventional parameters such as lithology, soil texture, and land use/land cover data are gathered from Geological Survey of India, Kerala State Soil Survey Organization, and Kerala State land Use Board, respectively. Charnokite group of rocks and tertiary deposits of Sand and Silt is the major lithology found in the study area followed minor patches of migmatite complex (Fig. 2c). Gravelly clay is the dominating soil texture found in the study area followed clay and sandy soils (Fig. 2d). KUC has different land use classes including agricultural area, built-up land, waste lands, wetlands, and water bodies (Fig. 2e).

Slope aspect of the study area is shown in Fig. 2f which shows nine slope directions; however, flat and northern slopes are the major slope directions present in KUC. TWI is another important terrain parameter which represents soil moisture concentration at a given point. In the study area, TWI values range from 4.25 to 20.88 (unitless) (Fig. 2g). The KUC has a valley depth which ranges from 0.05 to 56.48 m (Fig. 2h), and in general, valleys having higher depth are considered as higher flood susceptible area. Channel network base level (CNBL) is another important parameter used for flood susceptibility modeling. In the study area, CNBL values range from 0 to 29.55 (Fig. 1f). Convergence index (CI) is a terrain parameter which shows the structure of the relief as a set of convergent areas (channels) and divergent areas (ridges). It represents the convergence or divergence of overland flow. In the study area, CI values range from −93.58 to 96.99 (Fig. 2j). The present study used plan and profile curvatures for modeling (Fig. 3k and l). In general, profile curvature is defined as curvature parallel to the direction of the maximum slope, whereas plan curvature is perpendicular to the direction of the maximum slope.

2.3 Random Forest Method (RF)

RF is a powerful machine learning method proposed by Breiman [4]. RF is a decision tree-based (DT) model that can be used for both classification and regression problems. RF is an ensemble DT model which operates by constructing a multitude of decision trees at the training time and outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Random-decision forests correct for the decision tree habit of overfitting to a training set [2, 15]. In this study, flood susceptibility is treated as a binary classification, i.e., flood occurrences (1) and non-occurrences (0). Consider training set $D = ((A1, B1),\ldots..,(An, Bn))$ that consists of n vectors, $A = \in X$ where X is a set

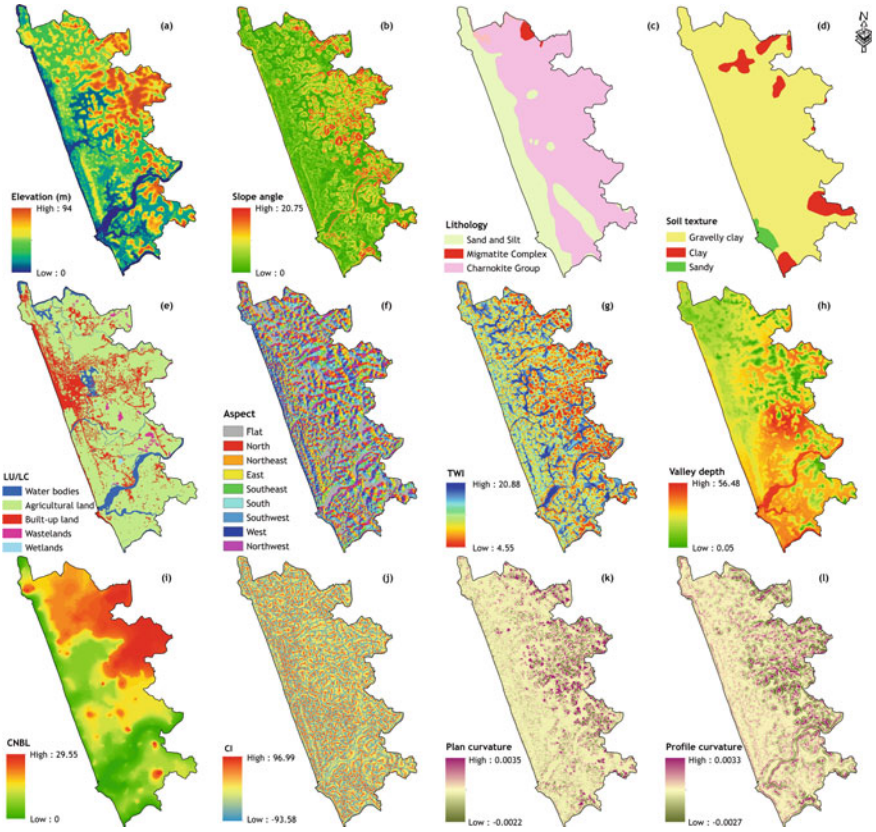

Fig. 2 Flood conditioning factors selected for the flood susceptibility modeling

of numerical or symbolic observations, and $B = \in Y$ where Y a set of class labels (here flood and non-flood). For classification problems, a classifier is a mapping $X \rightarrow Y$ [6]. The RF is working with two processes; the first is Breiman's "bagging" idea and the second is Ho's "random selection features. Bagging is an ensemble machine learning procedure to improve the prediction accuracy of a weak classifier by creating a set of classifiers.

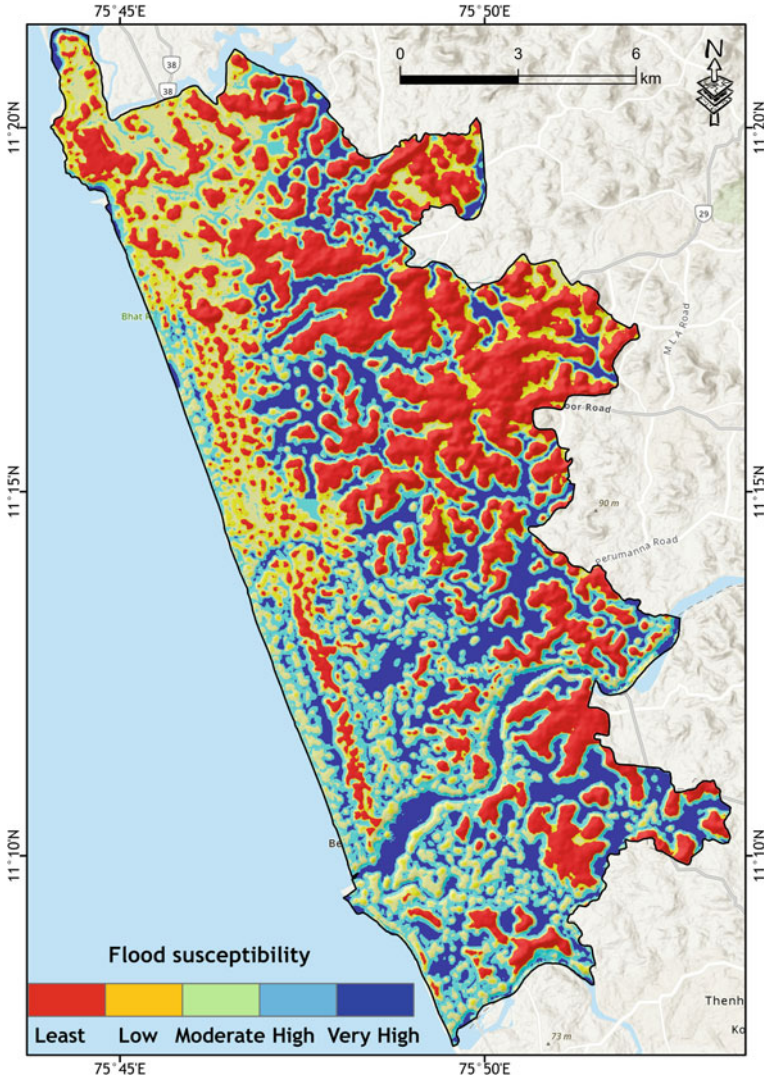

Fig. 3 Flood susceptibility map of the study area

3 Results and Discussions

3.1 Model Training and Validation

In the present study, the trained RF model is evaluated using different statistical methods before projecting to the geographical extend. The model performance evaluation in training and testing sections is summarized in Table 1. During training

Table 1 Model performance in training and testing sections

Models	TP	TN	FP	FN	N	Sensitivity	Specificity	Accuracy	K	AUC	RMSE
Training	186	159	29	56	430	0.769	0.846	0.802	0.605	0.852	0.380
Testing	88	59	4	33	184	0.727	0.937	0.799	0.598	0.785	0.407

Table 2 Model performance and error estimation in 10 fold cross-validations

Folds	Accuracy	AUC	RMSE
CV_1	0.916	0.937	0.372
CV_2	0.947	0.988	0.348
CV_3	0.682	0.943	0.352
CV_4	0.864	0.894	0.366
CV_5	0.894	0.818	0.398
CV_6	0.706	0.751	0.421
CV_7	1.000	1.000	0.283
CV_8	0.857	0.711	0.411
CV_9	0.733	0.768	0.413
CV_10	0.863	0.933	0.349

section, the ability of classifying flood locations or the sensitivity value is 0.769 whereas specificity higher value of 0.845. In the testing mode also, specificity value is higher than the sensitivity value (0.937 and 0.727, respectively) which shows that the model has better ability to classify the non-flood occurrences. It should be noted that overall accuracy in both training and testing sections is nearly same (i.e., 0.802 and 0.799, respectively) (Table 1). Kappa index also shows negligible differences in both training and testing phases. In the case of AUC values, which is generally considered as a robust measure of classification accuracy, model obtained a decent AUC value of 0.852 in training section and 0.785 in testing section.

In the case of RMSE, training phase shows lowest value (0.380) than validation phase (0.407). Besides to cross-check the efficiency of trained model, a ten-fold CV was implemented and summarized in Table 2, which shows overall good performance. Therefore, the model is finally projected for the entire study area.

3.2 Flood Susceptibility Modelling

The flood susceptibility map is prepared by transferring the probability of flood occurrences in the study area, which is further classified into five zones such as least susceptible area, low, moderate, high, and very high susceptibility areas (Fig. 3). About 34.31% of the study area is categorized under least flood susceptibility zone followed by 13.20% in low susceptibility, 19.93% in moderate susceptibility, 18.78% in high susceptibility, and 13.78% in very high susceptibility area.

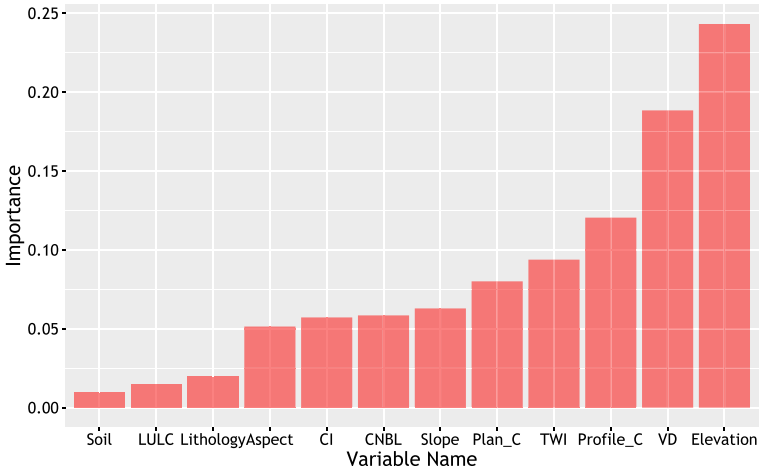

Fig. 4 Variable importance for the occurrence of flood in the study area

Variable importance analysis using random forest model is also carried out to identify the significant variables which influence the occurrence of flood in KUC. As shown in Fig. 4, elevation is the most important parameter which influences the flood occurrence followed by valley depth profile curvature and TWI. Other parameters such as plan curvature, slope, CNBL, CI, and slope aspect have moderate influence. Lithology, soil, and LULC are the least important parameters which affect the flood occurrences. In general, terrain parameters are the major flood influencing factors of KUC.

4 Conclusions

Urban flood susceptibility is estimated using RF method in a fast-growing urban agglomeration in southern India. Flood inundation locations are collected using field work, and thereafter, twelve independent variables such as land use/land cover, soil texture, lithology, elevation, slope angle, slope aspect, valley depth, topographical wetness index, profile curvature, plan curvature, convergence index, and channel network base level were analyzed and used for flood susceptibility mapping. The model obtained a decent AUC value of 0.852 in training section and 0.785 in testing section. Thereafter, the model is projected to geographical extend and classified into five zones such as least susceptible area, low, moderate, high, and very high. About 34.31% of the study area is categorized under least flood susceptibility zone followed by 13.20% in low susceptibility, 19.93% in moderate susceptibility, 18.78% in high susceptibility, and 13.78% in very high susceptibility area. The proposed flood susceptibility map is trust worthy for future infrastructure building and hazard mitigation.

References

1. Abedi R, Costache R, Shafizadeh-Moghadam H, Pham QB (2021) Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. Geocarto Int:1–18
2. Achu AL, Thomas J, Aju CD, Gopinath G, Kumar S, Reghunath R (2021) Machine-learning modelling of fire susceptibility in a forest-agriculture mosaic landscape of southern India. Ecol Inform 64:101348. https://doi.org/10.1016/j.ecoinf.2021.101348
3. Ahmed N, Hoque MAA, Arabameri A, Pal SC, Chakrabortty R, Jui J (2021) Flood susceptibility mapping in Brahmaputra floodplain of Bangladesh using deep boost, deep learning neural network, and artificial neural network. Geocarto Int:1–22
4. Breiman L (2001) Random forests. Machine Learning 45(1):5–32
5. Census of India (2011) District Census Handbook, Kozhikode. Series-33, Part-XII-B
6. Chapi K, Singh VP, Shirzadi A, Shahabi H, Bui DT, Pham BT, Khosravi K (2017) A novel hybrid artificial intelligence approach for flood susceptibility assessment. Environ Model Softw 95:229–245
7. Fang Z, Wang Y, Peng L, Hong H (2021) Predicting flood susceptibility using LSTM neural networks. J Hydrol 594:125734
8. Islam ARMT, Talukdar S, Mahato S, Kundu S, Eibek KU, Pham QB, Linh NTT (2021) Flood susceptibility modelling using advanced ensemble machine learning models. Geosci Front 12(3):101075
9. Jesiya NP, Gopinath G (2019) A customized fuzzy AHP-GIS based DRASTIC-L model for intrinsic groundwater vulnerability assessment of urban and peri urban phreatic aquifer clusters. Groundw Sustain Dev 8:654–666
10. Kourgialas NN, Karatzas GP (2011) Flood management and a GIS modelling method to assess flood-hazard areas—a case study. Hydrol Sci J J des Sci Hydrol 56(2):212–225
11. Luu C, Pham BT, Van Phong T, Costache R, Nguyen HD, Amiri M, Trinh PT (2021) GIS-based ensemble computational models for flood susceptibility prediction in the Quang Binh Province, Vietnam. J Hydrol 599:126500
12. Panahi M, Jaafari A, Shirzadi A, Shahabi H, Rahmati O, Omidvar E, Bui DT (2021) Deep learning neural networks for spatially explicit prediction of flash flood probability. Geosci Front 12(3):101076
13. Pappenberger F, Matgen P, Beven KJ, Henry JB, Pfister L (2006) Influence of uncertain boundary conditions and model structure on flood inundation predictions. Adv Water Resour 29(10):1430–1449
14. Pham BT, Phong TV, Nguyen HD, Qi C, Al-Ansari N, Amini A, Tien Bui D (2020) A comparative study of kernel logistic regression, radial basis function classifier, multinomial naïve Bayes, and logistic model tree for flash flood susceptibility mapping. Water 12(1):239
15. Pham QB, Achour Y, Ali SA, Parvin F, Vojtek M, Vojteková J, Anh DT (2021) A comparison among fuzzy multi-criteria decision making, bivariate, multivariate and machine learning models in landslide susceptibility mapping. Geomatics, Nat Hazards Risk 12(1):1741–1777
16. Shankar MA, Bindu CA (2021) Appraising the need for disaster mitigation in existing planning documents of Municipal Corporations of Kerala in the event of past disasters. In IOP Conf Ser Mater Sci Eng 1114(1):012039). IOP Publishing
17. Tehrany MS, Pradhan B, Jebur MN (2014) Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. J Hydrol 512:332–343
18. Wang Y, Fang Z, Hong H, Costache R, Tang X (2021) Flood susceptibility mapping by integrating frequency ratio and index of entropy with multilayer perceptron and classification and regression tree. J Environ Manage 289:112449
19. Zzaman RU, Nowreen S, Billah M, Islam AS (2021) Flood hazard mapping of Sangu River basin in Bangladesh using multi-criteria analysis of hydro-geomorphological factors. J Flood Risk Manage 14(3):e12715