Simulating flood risk in Tampa Bay using a machine learning driven approach

Article · December 2024

DOI: 10.1038/s44304-024-00045-4

5 authors, including:

Hemal Dey
University of Alabama
8 PUBLICATIONS 89 CITATIONS
SEE PROFILE

Md Munjurul Haque
University of Alabama
10 PUBLICATIONS 87 CITATIONS
SEE PROFILE

Wanyun Shao
University of Alabama
63 PUBLICATIONS 1,366 CITATIONS
SEE PROFILE

Matthew S. VanDyke
University of Alabama
33 PUBLICATIONS 422 CITATIONS
SEE PROFILE

https://doi.org/10.1038/s44304-024-00045-4

Simulating flood risk in Tampa Bay using a machine learning driven approach

Check for updates

Hemal Dey[1,2], Md Munjurul Haque[1,2], Wanyun Shao[1,2] ✉, Matthew VanDyke[2,3] & Feng Hao[4]

Machine learning (ML) models can simulate flood risk by identifying critical non-linear relationships between flood damage locations and flood risk factors (FRFs). To explore it, Tampa Bay, Florida, is selected as a test site. The study's goal is to simulate flood risk and identify dominant FRFs using historical flood damage data as target variable, with 16 FRFs as predictor variables. Five different ML models such as decision tree (DT), support vector machine (SVM), adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost), and random forest (RF) were adopted. RF classifies 2.42% of Tampa Bay as very high risk and 2.54% as high risk, while XGBoost classifies 3.85% as very high risk and 1.11% as high risk. Moreover, the communities reside at low altitudes and near the waterbodies, with dense man-made infrastructure, are at high flood risk. This study introduces a comprehensive framework for flood risk assessment and helps policymakers mitigate flood risk.

Flooding is a common disaster around the world[1–3], having enormous detrimental impacts on society[4,5]. The frequency and intensity of floods are rising due to climate change and the consequential damage is dramatically increasing as a result of elevated exposure[6]. A great amount of research is focused on risk assessment and mitigation strategies[7,8]. Flooding, as a natural disaster influenced by numerous factors, is challenging to prevent entirely. However, the effects can be significantly reduced through the implementation of risk mitigation strategies[9] including by providing risk decision-makers with relevant, accurate risk assessment information that can aid in establishing effective emergency management protocols and communicating with the public to save lives and property. With the increasing frequency and intensity of flooding, it has become imperative to mitigate flood risks effectively. It is thus vital to first accurately assess flood risks. Flood risk assessment contributes significantly to managing floods effectively. These assessments enable us to identify possible threats at both the global and local levels, providing vital information for mapping high-risk areas. This vital information serves as input to targeted interventions, enabling more efficient and effective mitigation and management of flood impacts[10]. The interaction between naturally occurring events and vulnerable populations leads to complex challenges, specifically, as does the need for different types of authorities and decision-makers to have information relevant to the flood risk decisions they make[11] (e.g., when to deploy specific resources during a flood event versus comprehensive planning in advance of a forecasted event). Flood risk assessment is accordingly a multifaceted process encompassing the hazard engendered by naturally occurring events and the social determinants that affect how these events impact human communities[12]. Researchers have long called for an integrated approach that considers both physical dimensions (hazard and exposure) and social dimensions (vulnerability)[13–15]. An integrated approach to flood risk assessment that includes both physical and social aspects serves as a foundation for policy recommendations and loss minimization[14].

The core task in flood risk assessment is to identify the vulnerable locations to flooding to assist sustainable flood planning and prevent losses[16]. High accuracy in the identification of flood-risk zones has a positive correlation with effective flood risk mitigation[17]. Flood risk assessment often integrates topological factors, hydrological factors, socioeconomic factors, and climate change impacts into one analytical framework. Integrating these factors into flood risk assessments can lead to the development of better-informed floodplain management strategies, disaster preparedness plans, and infrastructure investments that prioritize the needs of vulnerable communities. Vulnerability can be significantly increased if a system or community is more exposed to a particular hazard. Additionally, their degree of susceptibility and capacity for adaptation or recovery are also deeply connected to their vulnerability[12]. Understanding the interaction among flood hazards, exposure and vulnerabilities is thus essential for building resilience and mitigating the potential consequences of flooding in the dynamic and at-risk region. Flood risk assessment for mitigation purposes existed long before the data science boom of the past decade. However, recent technological advancements have made this research increasingly popular and practical for effective risk mitigation[16,18]. Geographic Information System (GIS) and Remote Sensing techniques serve as essential tools to analyze and integrate flood-related information[19]. Integrating machine learning (ML) classification approaches into the flood risk assessments will offer better prospects of improved accuracy, shorter computation times, and

[1]Department of Geography & the Environment, University of Alabama, Tuscaloosa, AL, USA. [2]Alabama Water Institute, University of Alabama, Tuscaloosa, AL, USA. [3]Department of Advertising and Public Relations, University of Alabama, Tuscaloosa, AL, USA. [4]Department of Sociology & Interdisciplinary Social Sciences, University of South Florida, Tampa, FL, USA. ✉e-mail: wshao1@ua.edu

reductions in the overall expenses associated with model development[20]. Numerous studies have used ML algorithms to assess flood risk[8,16,18,21–23]. Several researchers have explored the application of ML models and decision-making algorithms for flood risk assessment, hazard mapping, and vulnerability assessment, by employing various techniques such as random forest (RF), support vector machine (SVM), neural networks, and logistic regression to analyze flood factors[24–29].

Despite these advancements, no studies have yet considered past flood damage data and a wide variety of flood risk factors (FRFs) from both physical and social dimensions using advanced ML models to simulate flood risks. Hence, the present study offers three novelties. First, this study introduced a unique approach of assessing flood risk by utilizing past flood damage data as a target variable and a diverse range of FRFs from hazards, exposure, and vulnerability components of flood risk as predictors. This approach can yield robust and more accurate results as demonstrated by Yarveysi et al.[30] and Dey et al.[29]. Based on an extensive literature review, nine hazard factors including elevation, slope, aspect, curvature, precipitation, normalized difference vegetation index (NDVI), distance from waterbodies, topographic wetness index (TWI), and drainage density, and three exposure factors, namely, building footprint density, road network density, and normalized difference built-up index (NDBI) and, four vulnerability factors such as median income, percentage of Hispanic population, percentage of Black (African American) population, and percentage of people with no school completion were used as FRFs. Second, this study examines how accurate the ML models are in predicting flood risk and further compares the simulated flood risk maps (FRMs) with the Federal Emergency Management Agency's (FEMA's) 100-year floodplain map. Third, this study highlights some additional uses of ML models in flood risk assessment, aiding policymakers in flood mitigation by identifying key FRFs and vulnerable populations and their locations based on historical flood damage in Tampa Bay.

To implement our analytical strategies, Tampa Bay, Florida in the USA was selected as a test bed. Tampa Bay is located on the Gulf Coast of Florida (Fig. 1), and the area has a population of over three million, according to the 2020 Census. The coastline's location and proximity to the Gulf make the region vulnerable to extreme weather events. According to the urban adaptation assessment data from the University of Notre Dame, Tampa Bay is exposed to the risk of flooding and sea level rise. Studies have revealed that different communities in Tampa Bay are highly vulnerable due to sea level rise[31,32]. In the past decade, several hurricanes have passed Tampa and devastated places close to Tampa Bay. Hurricane Idalia (Category 3) landed on the north coast of Tampa, and Hurricanes Ian and Irma, both Category 4, landed on the south coast of Tampa. These events brought heavy rainfall, human casualties, and substantial damage to the economy. The historical average cost of flood events between 2011 and 2015 was over $500 thousand, and the projected cost due to sea level rise is over $700 million by 2040. Considering the significant recent history with flooding and its sizable population, it is an ideal study location for us to test our methods of assessing flood risk using advanced ML algorithms.

Results

Models' accuracy assessment

The process of evaluating the ML models has two major parts. First, all these ML models were evaluated by the AUC–ROC curve (Fig. 2). Later, several evaluation metrics including overall accuracy, precision score, recall score,

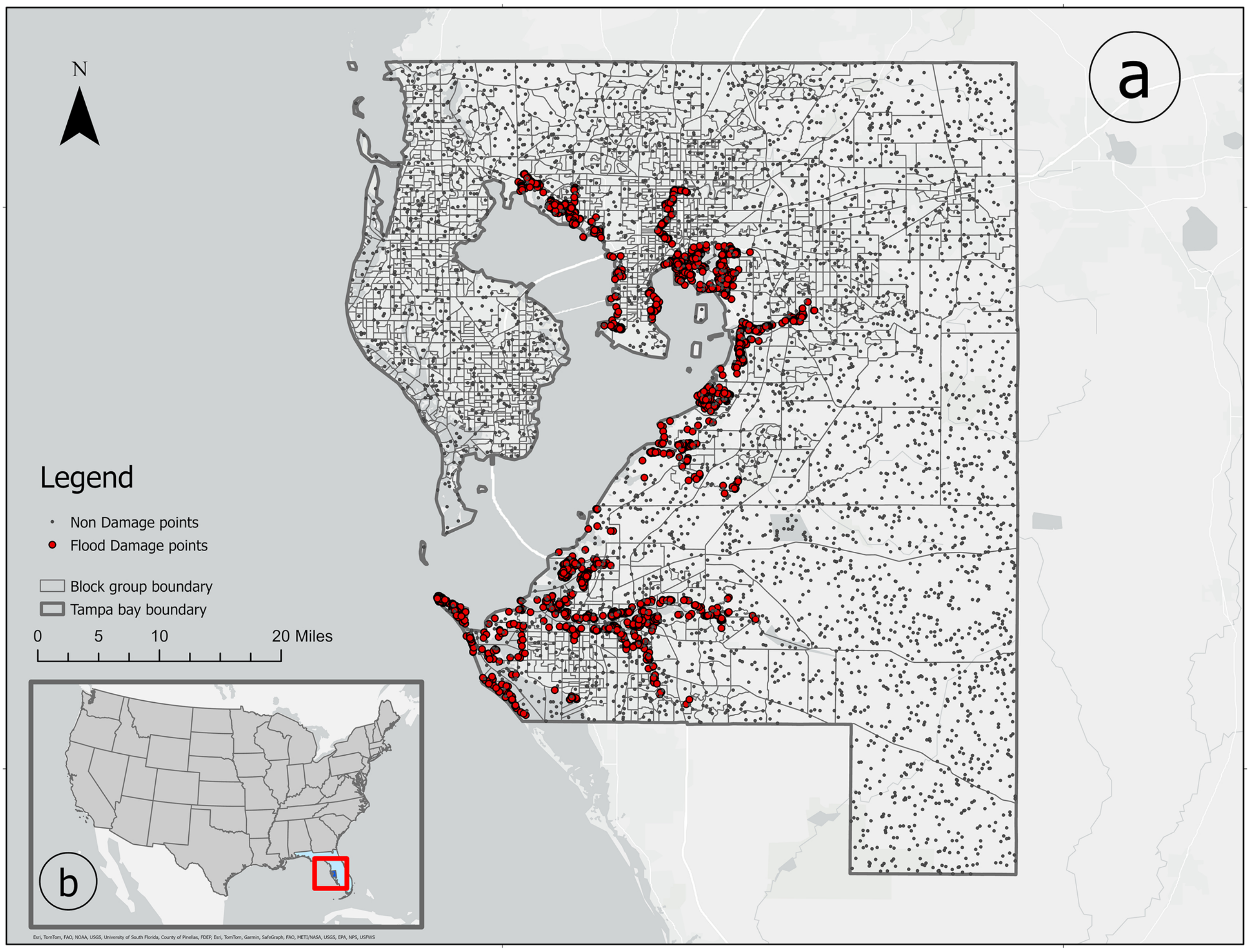
N
a
Legend
• Non Damage points
● Flood Damage points
Block group boundary
Tampa bay boundary
0
5
10
20 Miles
b
Esri, TomTom, FAO, NOAA, USGS, University of South Florida, County of Pinellas, FDEP, Esri, TomTom, Garmin, SafeGraph, FAO, METI/NASA, USGS, EPA, NPS, USFWS

Fig. 1 | The location of the study area. a Boundary of Tampa Bay with flood damage and non-damage points. Red dots indicate the flood damage points while gray dots indicate non-damage points. **b** The location of Tampa Bay and Florida State in the context of CONUS. This map was generated in ArcGIS Pro 2.4.0.

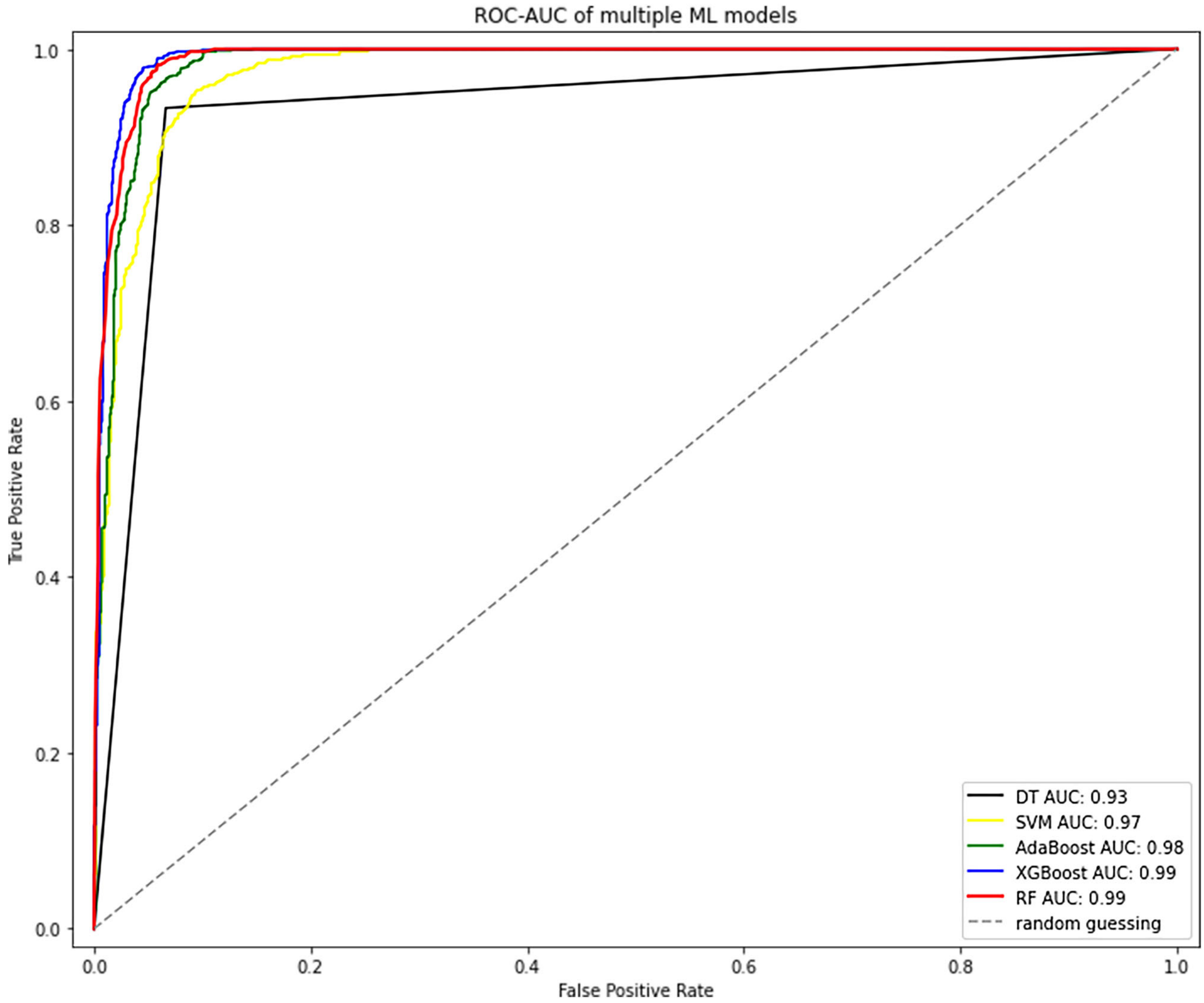

Fig. 2 | The ROC–AUC curve of ML models. The *X*-axis represents the false positive rate (1 – specificity) and *Y*-axis represents the true positive rate (sensitivity). The red, blue, green, yellow, and black lines represent the AUC curves for the RF, XGBoost, AdaBoost, SVM, and DT models, respectively. The gray dotted line indicates the AUC curve for random guessing. This analysis was conducted and plotted using *roc_curve* function in Python 3.11.7.

F-1 score, Kappa score, and Jaccard score were examined to further evaluate the accuracy of these ML models (Table 1).

According to the ROC–AUC curve analysis, the XGBoost and RF model achieved the highest AUC score of 0.99. On the other hand, the AdaBoost model achieved the second-best AUC score of 0.98. Meanwhile, SVM and DT yielded the lowest score among the models, standing at 0.97 and 0.93, respectively.

The overall accuracy test reveals that XGBoost and RF secured the highest score of 0.96 while Adaboost, DT, and SVM achieved 0.95, 0.93, and 0.92, respectively (Table 1). For precision score, XGBoost secured the best score of 0.94, followed by RF and DT at 0.93, and Adaboost at 0.92, and SVM at 0.87. For recall score, RF achieved the highest accuracy (0.99) compared to SVM (0.98), XGBoost (0.98), Adaboost (0.98), and DT (0.93). The F-1 score for both RF and XGboost were the highest at 0.96 with Adaboost, DT, and SVM at 0.95, 0.93, and 0.92. XGBoost had a Kappa score of 0.93, RF scored 0.92, Adaboost achieved 0.90, while DT and SVM scored 0.87 and 0.83, respectively. For the Jaccard score, XGBoost received a score of 0.93 and RF scored 0.92 while AdaBoost, DT, and SVM achieved 0.90, 0.87, and 0.85, respectively.

Based on the AUC–ROC curve and the results from multiple evaluation metrics, all ML models demonstrated strong performance in flood risk assessment. However, given the slightly better accuracy of the RF and XGBoost models, both were selected to simulate FRMs for Tampa Bay.

Table 1 | Accuracy assessment of ML models using multiple evaluation metrics

Metrics	DT	SVM	AdaBoost	XGBoost	RF
Accuracy	0.93	0.92	0.95	0.96	0.96
Precision	0.93	0.87	0.92	0.94	0.93
Recall	0.93	0.98	0.98	0.98	0.99
F-1 score	0.93	0.92	0.95	0.96	0.96
Kappa score	0.87	0.83	0.90	0.93	0.92
Jaccard score	0.87	0.85	0.90	0.93	0.92

Major contributing factors of flood risk

The findings of this study indicate that elevation is the dominant contributor (39%) to flood risk in Tampa Bay followed by distance to the river and waterbodies (25%) (Fig. 3). In addition, building footprint density contributes 7%, road network density contributes 5%, TWI contributes 4%, slope and drainage density each contribute 3%, respectively. However, the rest of the factors have very small contributions.

This study further used SHapley Additive exPlanations (SHAP) technique to interpret the importance and contribution of each FRF. These SHAP

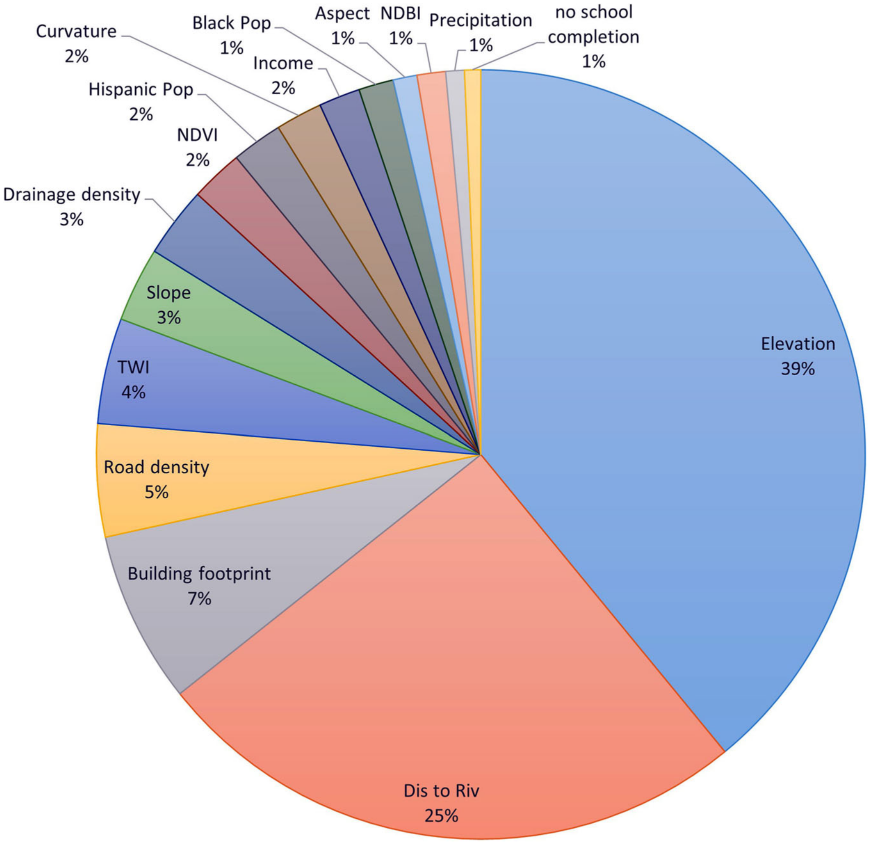

Fig. 3 | Contributions of different flood risk factors on flood risk. This figure depicts the percentages of contribution of different FRFs in flood risk predictions. This analysis was conducted using *feature_importances_* function in Python 3.11.7.

values not only demonstrate the importance of each feature but also indicate whether their contributions are positive or negative. According to the graph of SHAP method, the regions located in low-lying altitudes are at high flood risks, as flood water naturally flows downward due to gravitational forces and accumulate rapidly from runoff (Fig. 4). Moreover, areas located near waterbodies are inherently more prone to flooding due to the proximity to potential sources of water overflow. Both low elevation and proximity to waterbodies render those coastal regions highly prone to flooding (Fig. 4), especially when heavy rainfall and tropical storms surge cause water to overflow.

High density of building and road network increases the risk of flooding, whereas low density reduces this risk for two reasons. Dense man-made infrastructure is often associated with urbanization, which indicates intense human activities such as construction of settlement and thus significant alteration of the land surface. Urban land mostly made up of concrete absorbs water less efficiently for the reason. Meanwhile, a congested community with high buildings and road density suggests increased exposure to hazards. Whenever a disaster strikes, more of the population assets and infrastructure would be susceptible to the devastating impacts. All these conditioning factors have turned out to be prime determinants of flood risk in Tampa Bay.

Considering the contributions of all these FRFs, it can be summarized that geographical location, topographic and demographic conditions of a region highly contribute to amplify flood risk in a region. In Tampa Bay region, densely populated communities located in low altitude near the coast face a particularly high risk of flooding.

Flood risk simulation using RF and XGBoost model

Table 2 displays the distribution of flood risk area in Tampa Bay classified by the RF and XGBoost models. The findings from the RF model indicate that 2.42% of the area of Tampa Bay is at a very high risk for flooding while the XGBoost models classified 3.85% regions at very high risk. These areas are mainly adjacent to Tampa Bay areas which marked as red in Fig. 5.

In addition, the RF model and XGBoost model classified 2.54% and 1.11%, respectively, of the total Tampa Bay area as high risk, marked in orange color. Furthermore, 2.78% of the area is at moderate flood risk according to the RF model while XGBoost model classified 1.04% areas in the category of moderate flood risk. However, RF model indicates 4.32% regions are at low risk of flood, on the other hand, XGBoost indicates 1.34% areas at low flood risk. The remaining portion of the Tampa Bay is classified as very low risk for flooding by the RF and XGBoost model. These results indicate that both the RF and XGBoost models identify almost similar proportions of areas classified as high and very high risk for flooding. Nevertheless, the RF model exhibits a more detailed spectrum of risk levels compared to the XGBoost model.

Moreover, based on the analysis, it is found that 10.29% of Tampa Bay's total population are exposed to a significant level of flood risk, falling either into very high or high flood risk categories. Within this group, 13.98% are Hispanic origin and 6.38% are identified as Black population.

Comparing the flood risk map with FEMA's 100-year floodplain map

Hurricane Irma (2017) was a record-breaking Category 5 hurricane in the Gulf of Mexico and Caribbean Sea. Its maximum storm surge in Florida, USA had a return period ranging from 110 to 283 years[33]. It generated peak storm surge with a 92–109 year return period in the Florida Keys and 41–254 years along Florida's Gulf Coast. However, most of the gauging stations in Gulf Coast experienced storm surges with a return period of nearly 100 years during Hurricane Irma[33]. In addition, the FEMA 100-year floodplain is a key policy tool that directly guides federal flood insurance purchase and mitigation in the USA[34,35]. So, this study aims to compare the FRM, generated from Hurricane Irma-induced past flood damage data, with FEMA's 100-year floodplain map for Tampa Bay.

To compare them we overlaid the FRMs from RF and XGBoost models with FEMA's 100-year floodplain map. It is observed that the high-risk areas identified by RF and XGBoost model coincide with the FEMA's 100-year flood zones (Fig. 6). All the identified high and very high flood risk areas lay within the boundary of the FEMA's 100-year floodplain. Meanwhile, our

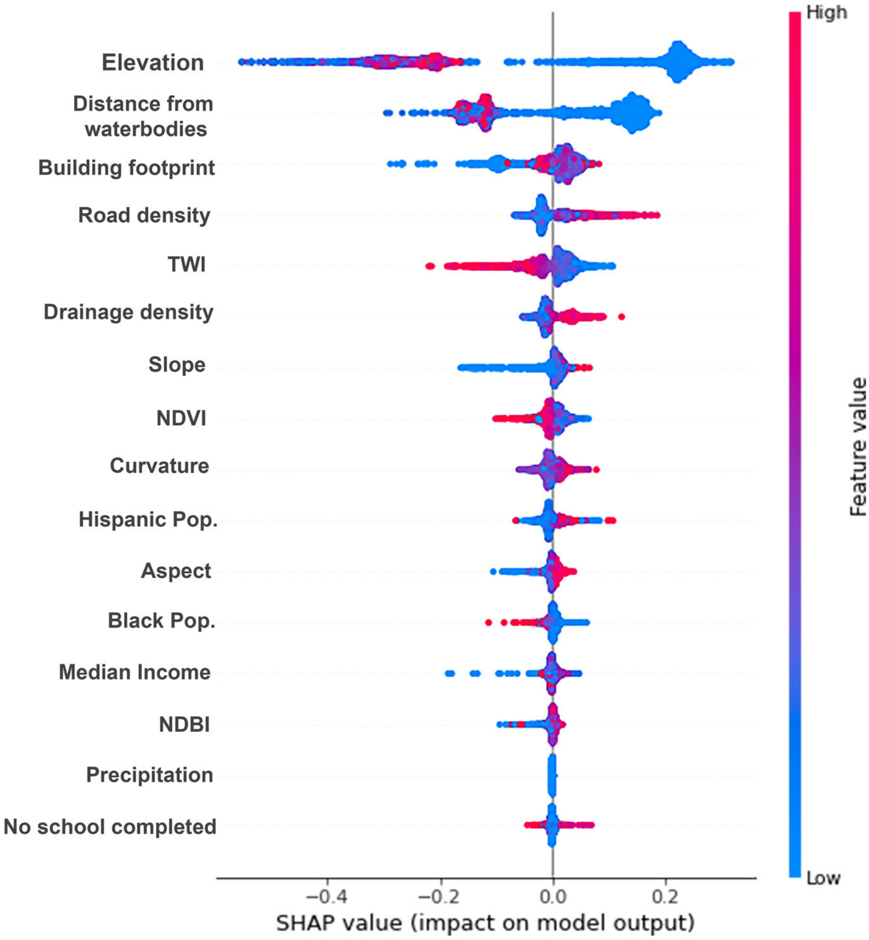

Fig. 4 | The SHAP interpretation of FRFs contributions. The *X*-axis displays the SHAP values, where a positive value indicates a higher contribution to flood risk and a negative value indicates a lower contribution. The primary *Y*-axis lists the names of the FRFs, while the secondary *Y*-axis shows the corresponding value ranges for these factors. Red dots indicate higher values of FRFs, while blue dots indicate lower values. The plot was generated using *shap* function in Python 3.11.7.

Table 2 | The classification results of flood risk from RF and XGBoost model

Risk category	RF		XGBoost	
	Area (sqkm)	%	Area (sqkm)	%
Very low	5693.49	87.92	5999.82	92.66
Low	280.04	4.32	86.74	1.34
Moderate	180.55	2.78	67.64	1.04
High	164.73	2.54	71.81	1.11
Very high	156.62	2.42	249.44	3.85

FRM distinguishes from FEMA's 100-year floodplain map in a critical manner. FEMA's flood zone designation is binary, determining one area either inside or outside a special flood hazard area (SFHA). This binary designation lacks nuances, leading to inaccurate estimation of flood risks, which further misleads risk mitigation behaviors[34,35]. Conversely, our flood risk assessment provides a spectrum which can more effectively guide planning and policy-making efforts to prioritize high and very high-risk zones, while the surrounding areas that are categorized as moderate or low risks can be considered with less urgency accordingly. This nuanced understanding can help efficient allocation of resources, especially with strained resources during an emergency.

Discussion

In this study, we considered Tampa Bay as a pilot study area, considering various risk factors such as its low-lying topography, coastal location, and history of severe weather events that contribute to frequent flooding. A total of five different ML models, including DT, SVM, AdaBoost, XGBoost, and RF, were trained and tested using flood damage data and FRFs to evaluate the effectiveness of each model in flood risk assessment. This study validated the results with a widely accepted method the AUC–ROC curve and several evaluation metrics including overall accuracy, precision score, recall score, F-1 score, kappa score and Jaccard score. Both the AUC–ROC curve and the other evaluation metrics supported that all ML models performed very well while RF and XGBoost model performed slightly better than others. However, it cannot be conclusively stated that the RF and XGBoost model are the universally superior ML model for other regions around the world. Any variations in the geographical location would result in a different dataset, and the performance of these models would intrinsically depend on the characteristics of that corresponding dataset. In the literature, a few papers established RF model as superior model in flood susceptibility assessment[29]. While a few other studies found that XGBoost is highly accurate in performing flood hazard assessment[23,36–38].

Both RF and XGBoost models suggest that nearly 5% of Tampa Bay, especially the areas adjacent to the coast, is at a very high or high risk (Fig. 5). This very high- and high-risk region from both models coincides with FEMA-designed 100-year flood plain (Fig. 6). Different from FEMA's 100-year flood map, our FRM presents a broader spectrum of risks, including very low, low, moderate, high, and very high-risk zones, providing a nuanced understanding of flood risk to decision-makers. A significant portion of Tampa Bay's population (10.29%) inhabits this area, among them 13.98% is Hispanic and 6.38% is Black, posing a potential threat to a large community.

Among the 16 FRFs, elevation (39%) was reported as the most dominant factor. This finding is consistent with that of many studies. Specifically, Desalegn and Mulu[39] did flood risk assessment in Ethiopia,

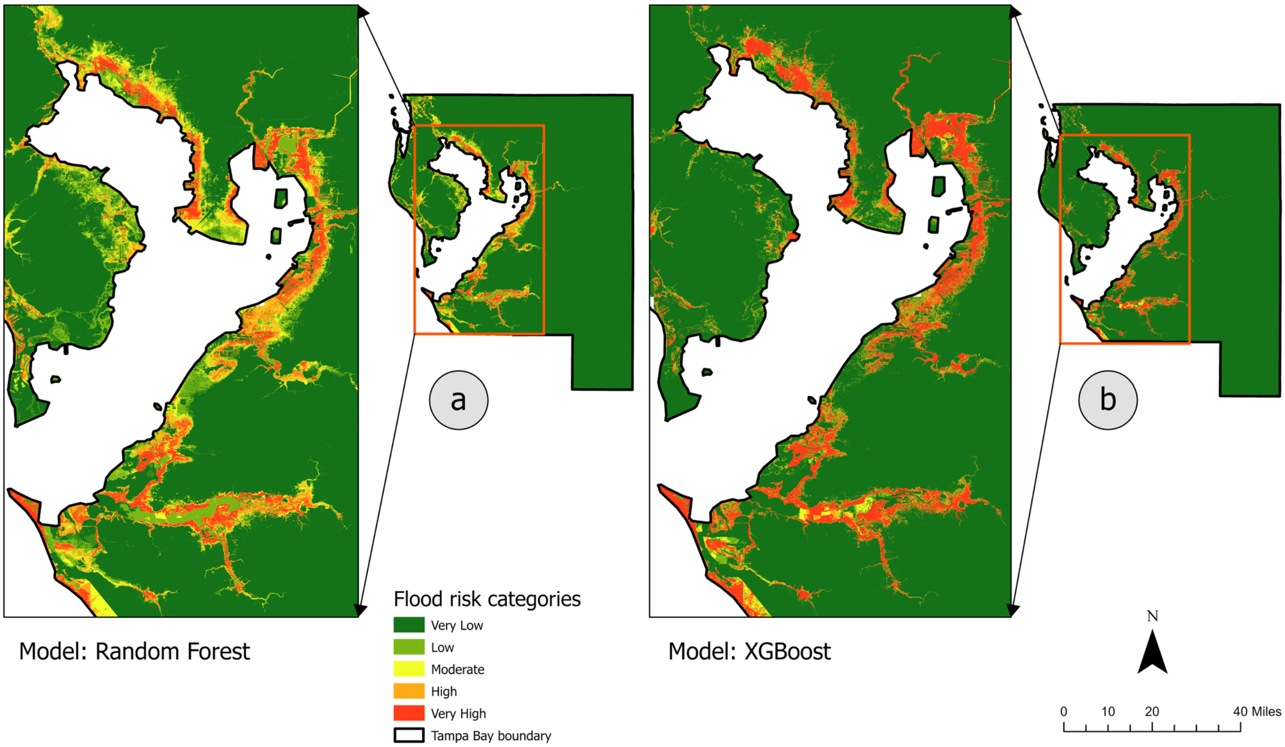

Fig. 5 | Simulated flood risk maps of Tampa Bay. a Flood risk map simulated by RF model. **b** Flood risk map simulated by XGBoost model. These figures represent the spatial distribution of flood risk zones in Tampa Bay at different levels of risk. Where, green color indicates very low risk, light green represents low risk, yellow color represents moderate risk, orange color represents high risk, and red color represents very high risk of flooding. The simulation was conducted in Python 3.11.7 and the maps were generated in ArcGIS Pro 2.4.0.

Ziarh et al.[40] in Malaysia, Hoque et al.[41] in Bangladesh, Vojtek and Vojteková[42] in Slovakia, Dey et al.[29] in New Orleans, Louisiana and Mukherjee and Singh[43] in Harris County, TX. They all reported that elevation is the most important factor determining flooding. This study also revealed that distance from the river or coast (25%) is the second most important factor for flood susceptibility in Tampa Bay. Many studies reported identical findings[44–47]. Communities residing near rivers or coastlines face a substantially higher flood risk compared to those farther inland. This increased risk is primarily due to the combined impact of coastal storm surges and heavy rainfall, which can exponentially escalate the flooding threat during high precipitation or extreme weather events[48]. Additionally, highly dense built-up areas with dense building and road network was reported as a substantial contributor to flooding in Tampa Bay. Higher infrastructure density increases the exposure of a community and ultimately contributes to escalating flood risks by preventing the natural drainage system of water and reducing the percolation process.

However, the findings indicate that the probability of flood risk is primarily driven by hazard factors (i.e., low-lying areas and proximity to waterbodies) and exposure factors (building footprint density and road network density) while vulnerability-related factors, such as socioeconomic conditions, appear to have less influence. The reason could be that this study trained ML models using flood property damage data. As a result, it makes sense that hazard and exposure factors had seemed the most contributing factors, since property damage is more closely related to hazard and exposure components than to vulnerability. If the ML models were trained using indicators like the flood fatality or death toll data, socioeconomic factors would likely have shown a more significant impact. Future studies could consider using flood fatality data to train the ML models. Despite the limitation, this study adopted state of art methods and incorporated all critical components of flood risk when simulating it. The final visualization of flood risk distribution in Tampa Bay presents a comprehensive FRM.

This identification of areas with flood risk will inform policymakers about the existing condition, which is invaluable for sustainable flood risk management. Without identifying a potential threat, the goal of effective flood risk management is unachievable. Moreover, ML models have identified several very high-risk zones, where the presence of substantial manmade infrastructure significantly elevates the overall flood risk to these communities. If no adequate mitigation measures are taken, these communities will be confronted with increased exposure to elevated flood risks. While the scientific community widely recognizes that elevation, distance from rivers or coasts, and dense urbanization are key factors influencing flood risk, we still advocate for their consideration in specific local flood risk management. These factors are crucial for both risk assessment and mitigation efforts. These hazards, exposure, and vulnerability indicators combined are meaningful variables for flood risk planning because these data are readily available (or likely accessible) to risk decision-makers in their own jurisdictions. This information can be used to develop proactive flood planning measures and mitigation strategies, including public communication and information strategies in heavily populated areas, that are sensitive and specific to the physical and social characteristics present in authorities' jurisdictions; moreover, flood risk planning and response strategies can be dynamic depending on local, regional, and broader fluctuations in these flood risk predictor variables. Policymakers and planners in the respective areas should consider these factors when planning for and managing flood risks, especially due to likely exacerbations promoted by climate change. Low-lying areas are more prone to flooding[9,49]. This understanding is important for policymakers for pinpointing areas at high risks of flooding, which will assist them to plan exactly where they should offer intervention for more stringent flood management measures. Informed by factors such as proximity to rivers or coasts, policymakers can enhance land use planning and development by creating flood-resistant infrastructure and enforcing zoning and robust building codes designed for resilience[50]. They can also strategically allocate resources to establish monitoring and early warning systems and ensure that emergency services are available and accessible in these high-risk areas, increasing the community's safety and reducing the potential impact of flooding. By considering

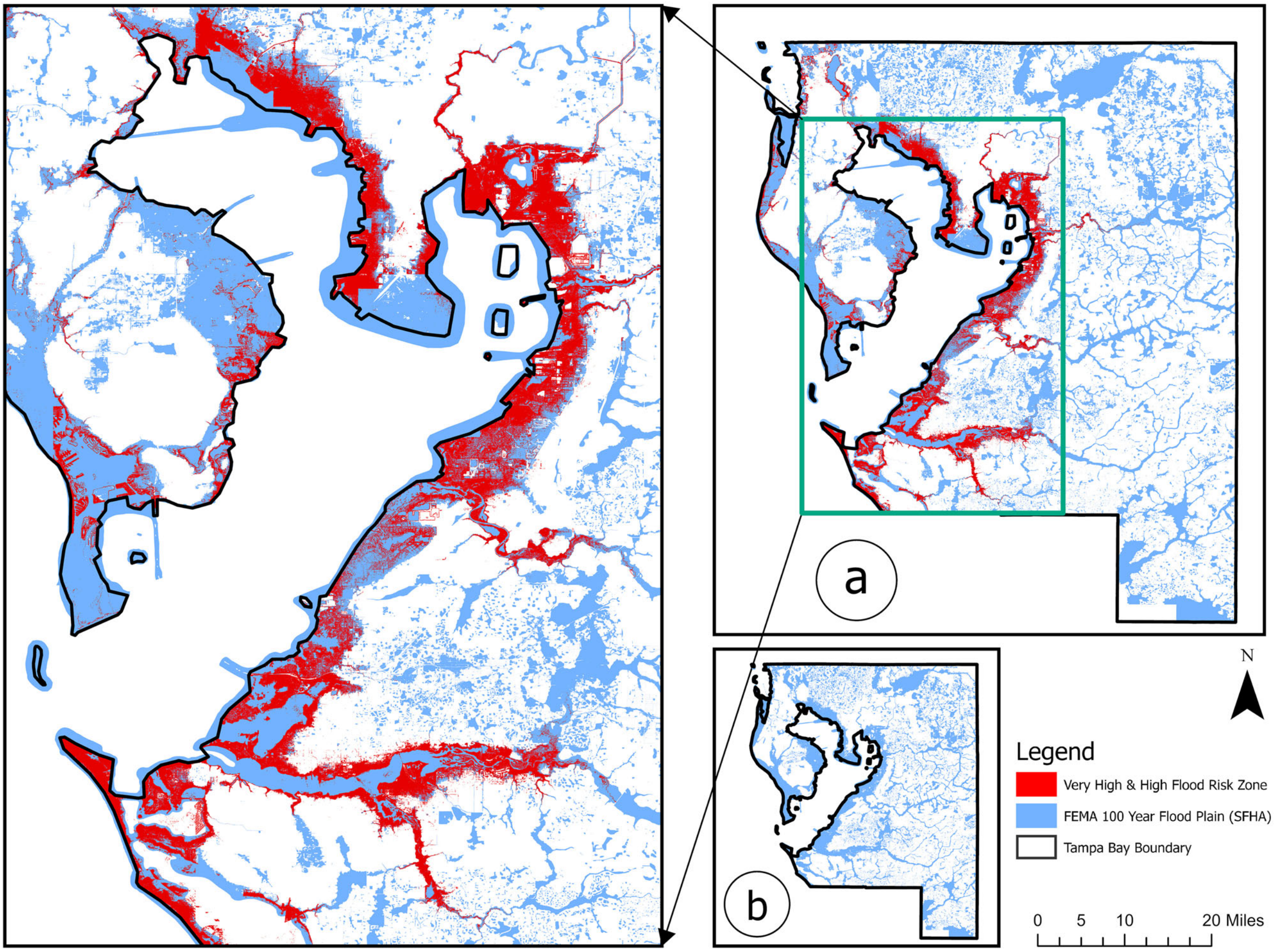
a
b
N
Legend
🟥 Very High & High Flood Risk Zone
🟦 FEMA 100 Year Flood Plain (SFHA)
□ Tampa Bay Boundary
0 5 10 20 Miles

Fig. 6 | The comparison between simulated very high and high flood risk areas (marked as red) by the RF and XGBoost model and FEMA's 100-year floodplain (marked as blue) of Tampa Bay. a An overlaid FRMs and FEMA's 100-year flood plain map. **b** A FEMA's 100-year floodplain map of Tampa Bay highlighting special flood hazard area (marked as blue). This map was generated in ArcGIS Pro 2.4.0.

building footprint and road network density, policymakers can effectively plan timely evacuation protocols and allocate resources in advance, ensuring efficient and successful flood management[51].

In the context of global climate change, sea level rise, and increased social vulnerability, disasters are more frequent, and the risk of severe impacts is high. To address this, enhancing community resilience through infrastructural, social, economic, and institutional measures can be an effective flood risk management strategy. This study demonstrates an effective approach to assessing flood risk by considering a wide spectrum of FRFs using advanced supervised ML algorithms. Through this approach, we were able to identify the zones under very high and high flood risks in Tampa Bay, along with the identification of responsible FRFs, in addition to potential exposed populations. Such risk identification procedures will enable us to take precautions and to better prepare for floods at both household and institutional levels. The risk identification is believed to ultimately advance flood resilience. All the findings will contribute to flood risk management both scientifically and practically. The findings can be meaningful in the vast scientific field for flood management studies around the world and the approach will directly assist the policymakers during their decision-making regarding flood risk management.

Methods

Flood risk modeling

To perform flood risk modeling, it is crucial to first define flood risk. Flood risk is defined as the function of the likelihood of a flood event and the potential loss of its negative impact on a community. The Intergovernmental Panel on Climate Change (IPCC) has characterized flood risk as the function of the interactions of three components: hazard, vulnerability, and exposure[25,52,53]. Hazard refers to the probability of flood inundation and the magnitude of flooding events that are influenced by the natural characteristics of a region. Exposure refers to the presence of population, assets, and infrastructures that are likely impacted by flooding events. Vulnerability refers to the susceptibility of a community or incapacity of a society to deal with the adverse impacts of flooding events. The combination of these three components can provide comprehensive information about flood risk[52]. Therefore, it is essential to consider causative factors from hazard, exposure, and vulnerability for flood risk modeling of the area. This study aimed to simulate flood risk by considering nine causative factors related to hazard, three causative factors related to exposure, and five causative factors related to vulnerability. These causative factors from three aspects also known as FRF will be the predictor variables. The past flood damage induced by Hurricane Irma as a Category 5 hurricane in September 2017 will be the target variable in the simulations.

Data sources

This study acquired several datasets from different dimensions of flood risk such as hazard, exposure, and vulnerability (Table 3). Majority of the hazards and exposure-related data are secondary or remotely sensed data which were collected from various sources at different spatial resolutions. Census data were collected from the US Census

Bureau (https://data.census.gov/cedsci) at block group level. The datasets from both physical and social dimensions were utilized to simulate the FRM. Since they had different spatial resolutions, they were all resampled into a 10 m spatial resolution and converted into GCS NAD 1983 before flood risk simulation.

Methodology

To simulate flood risk using ML models, two fundamental components were needed: flood damage points (the target variable) and FRFs (the predictors).

Flood damage points. Flood damage points are the specific locations where any kind of damage was recorded due to flooding. Flood damage arises because of the interaction between flood hazard, the physical aspects of flooding (flood water depth and runoff velocity), and social vulnerability, the socioeconomic conditions of a community (vulnerable individuals and infrastructure). Thus, any ML model trained on historical flood damage data has the potential of predicting future flood risk by detecting the non-linear relationship between flood damage points and various FRFs. This study collected flood damage data from the following data source: https://disasters-geoplatform.hub.arcgis.com/pages/historical-damage-assessment-database. These flood damage points were recorded in 2017 during Hurricane Irma by the FEMA. The flood damage points were divided into four categories: Affected, Minor, Major, and Destroyed. The level of damage was assessed based on flood depths at each structure determined by the most accurate flood depth grid during the damage assessment. A total of 3892 flood damage points (marked as red points) have been used as flood inventory points. Later, 3892 non-flood damage points (marked as gray points) were randomly generated and merged with flood damage points. Non-damage points were utilized alongside flood damage points as input in a more balanced dataset to train ML models. This improves model generalization by reducing biases. Overall, 7784 points were utilized in this study (Fig. 1). Flood damage points were labeled as 1 and non-flood damage points were labeled as 0. Initially, these points were divided into a 70:30 ratio where 70% of the data were

Table 3 | Description of datasets used in this study

Dataset	Data sources	Temporal resolution	Spatial resolution/ Data types	Data output
Hurricane Irma-induced flood damage data	FEMA	2017	Points	Flood damage points
The National Flood Hazard Layer (NFHL)	FEMA	2024	Polygon	FEMA 100-year flood plain map (SFHA)
JRC Global Surface Water	Pekel et al.[72]	2020	30 m	Distance from waterbodies
3DEP DEM	USGS	–	10 m	Elevation, slope, aspect, curvature, TWI, drainage density
UCSB-CHG/CHIRPS/DAILY	CHIRPS	2017	5566 m	Precipitation
Sentinel 2	ESA	2017	10 m	NDVI, NDBI
US Building Footprints	Microsoft	2020	Polygon	Building density
Primary and Secondary Road	US Census Bureau	2023	Polyline	Road density
Census data	US Census Bureau	2020	Polygon (Block groups)	Vulnerability factors

Table 4 | Rationality of selecting flood risk factors

Relation between flood risk and factors			
	Hazards		**Exposure**
Elevation	Elevation and flood occurrence have an inverse correlation. It plays an important role in directing flood waters flow with low-lying regions that tend to experience a higher flood risk as gravitational force drives floodwater to flow from high to low elevated regions [55,56,62,73,74]	Building density	Higher building density can amplify both the likelihood and severity of urban flooding. As higher building density reduced the amount of permeable surface and natural drainage system that hinder surface runoff
Slope	The flood water accumulation process, infiltration process, and sedimentation process depend on the topographic slope as it has a direct effect on the speed and direction of flood water[75,76]	Distance to roads	Concrete roads are also impermeable that prevents water percolation process and increased surface runoff
Aspect	The amount of rainfall and surface runoff depends on the topographic aspect because it refers to the specific direction of a topographic slope or land surface[77]	NDBI	NDBI is positively correlated to flooding, as excessive amount inefficiently planned urbanization makes people and property more susceptible to flooding[77,78]
Curvature	The curvature of a land surface influences the water budget of floodplains. It also helps to differentiate the regions where surface runoff diverges and converges [79]	Vulnerability	
Precipitation	Precipitation led rivers to overflow and inundate surrounding regions. Higher precipitation is often responsible for infrastructural damage and life loss[41,75,80]	Median income	Demographic and socioeconomic factors are highly responsible for total flood risk[81–83]. For example, income is highly correlated with individuals housing and surrounding conditions which indirectly increase the high risk of flood damage. Higher income peoples have more resources to prevent flood damage
NDVI	NDVI is negatively correlated with flood risk because dense vegetation cover obstructs surface water flow and promotes infiltration and percolation processes. While sparse vegetation and bare land facilitate rapid and unrestricted surface water flow into human settlements[62]	Percentage of Hispanic population	Through their socioeconomic status, language barriers, limited access to resources, and occupational exposure Hispanic population potentially increasing their flood risks

Table 4 (continued) | Rationality of selecting flood risk factors

Relation between flood risk and factors			
	Hazards		**Exposure**
Distance from waterbodies	Flood risk of a human settlement is strongly negatively linked with its proximity to waterbodies as the regions close to river and waterbodies tend to experience more inundation by flood water[74,77]	Percentage of Black (African American) population	Through their socioeconomic status, language barriers, limited access to resources, and occupational exposure Black population potentially increasing their flood risks
TWI	TWI is a conceptual hydrological model that provides an indication of topographic wetness characteristics. It helps to understand surface flow as wetter regions tend to have rapid water flow compared to dryer regions[80]	Percentage of people with no school completion	People with no schooling experience have very low awareness, fewer chances of employment, lower income, and less access to resources, all of which lead to a higher risk of flooding
Drainage density	Drainage density, a ratio between total channel length and basin area, influences surface runoff direction and water discharge. Higher drainage density refers to the high probability of flood risk[78]		

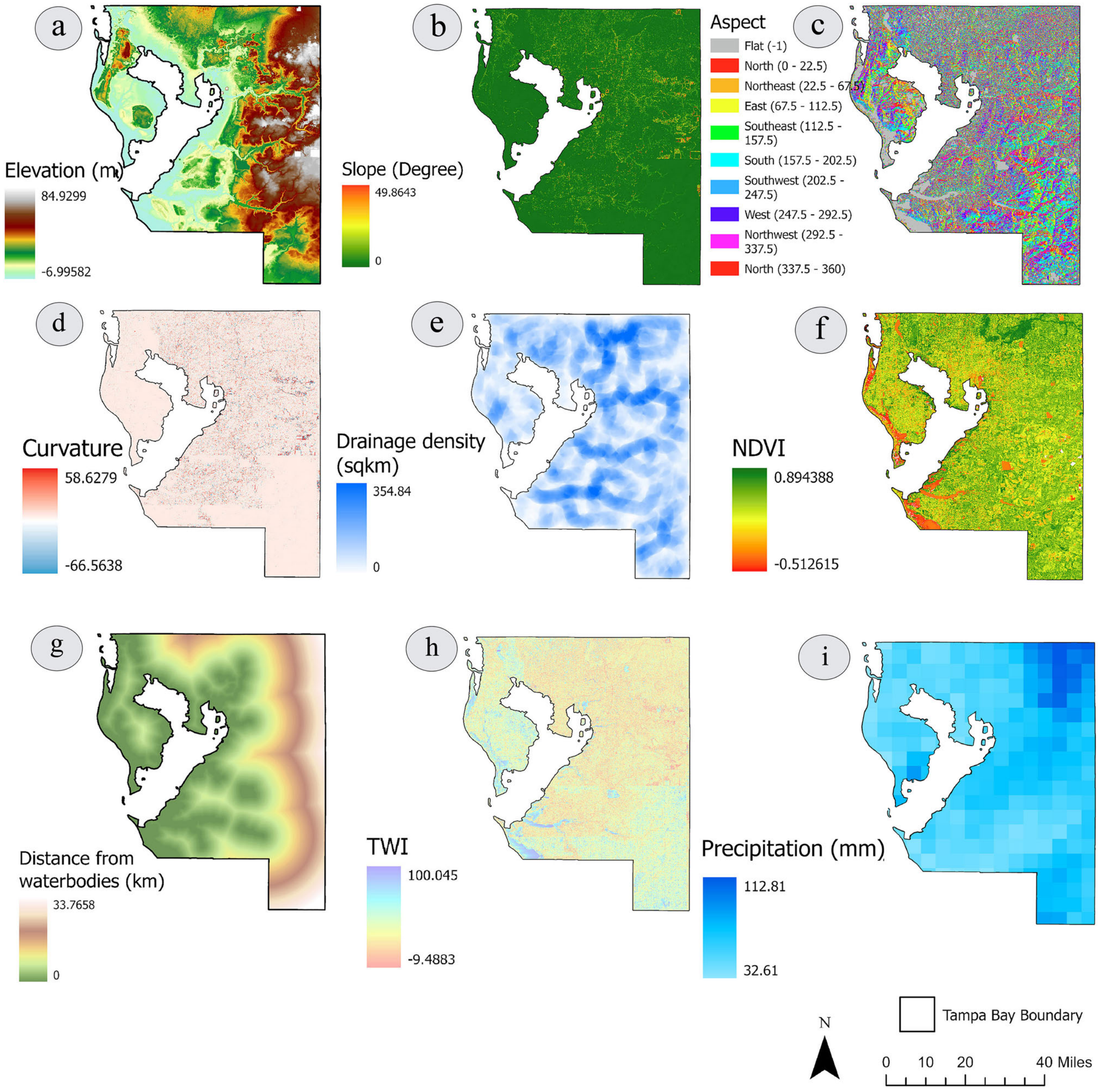

Fig. 7 | Maps of hazard-related factors of flood risk. a Elevation; (**b**) slope; (**c**) aspect; (**d**) curvature; (**e**) drainage density; (**f**) NDVI; (**g**) distance from waterbodies; (**h**) TWI; (**i**) precipitation. These maps were generated in ArcGIS Pro 2.4.0.

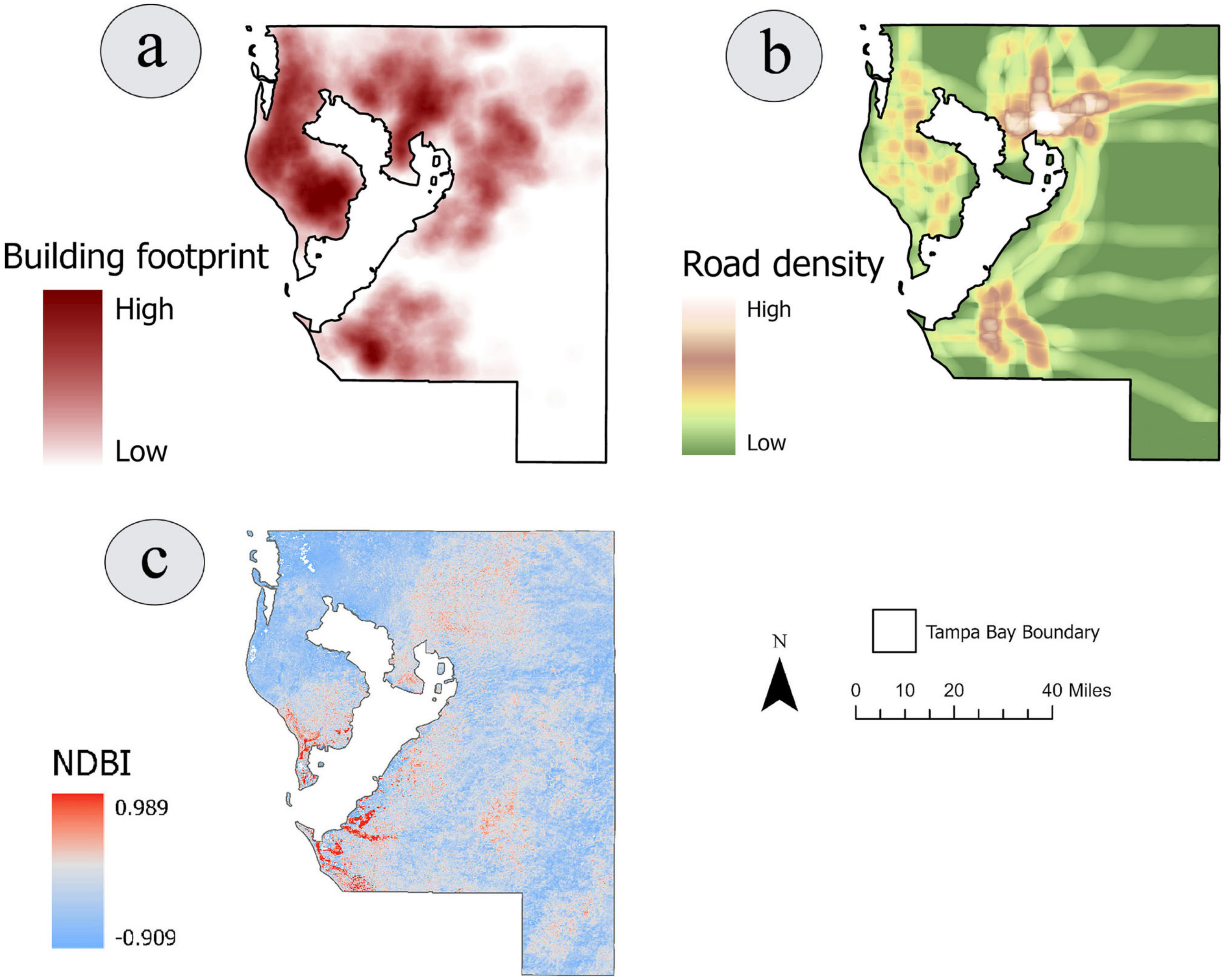

Fig. 8 | Maps of exposure-related factors of flood risk. **a** Building footprint density; (**b**) road network density; (**c**) NDBI. These maps were generated in ArcGIS Pro 2.4.0.

used as a training dataset and 30% were used as a testing dataset during flood risk simulation.

Flood risk factors. FRF are the factors that have direct or indirect influences on flood risk. A total of 16 FRFs have been selected and used in this study. The rationality of choosing these FRFs is briefly described in Table 4. Among them nine factors were taken from the hazards component (Fig. 7), three factors were taken from the exposure component (Fig. 8), and four factors were chosen from the vulnerability component (Fig. 9).

Before selecting these FRFs, this study underwent a multicollinearity analysis to check if any severe correlation exists among them. Conducting a multicollinearity analysis is a crucial step prior to selecting FRFs for flood risk simulation. The value of multicollinearity analysis ranges from −1 to 1. The high correlation between two factors indicates high similarities in data and it has a negative effect on flood risk predictions. So, the factors that have high correlation values above 0.8 and below −0.8 will make predictions inaccurate[54]. This study thus adopted a threshold value ± 0.8 in selecting FRFs. According to multicollinearity analysis, there is no severe correlation exists among these factors (Fig. 10).

This study also checked multicollinearity using variance inflation factors (VIF). When VIF is larger than 10, it indicates multicollinearity problem among FRFs[55]. However, there are no severe correlations among them (Fig. 11). Following that, all these selected FRFs were standardized using the *Z*-scores equation (Eq. 1) before training and testing the ML models.

$$z = \frac{(x - \mu)}{\sigma} \tag{1}$$

where z is the normalized value, x is the value of each data FRF variable, μ is the mean value of the FRF variable, and σ is the standard deviation of the FRF variable.

Machine learning models. This study adopted a total of five ML models including DT, SVM, AdaBoost, XGBoost, and RF. This study employed the default classifier for each ML model to provide an equal basis for evaluation. These models are briefly described below.

Decision tree (DT). Decision trees algorithm can effectively solve classification, regression, and multi-output problems by fitting complex datasets[56]. The DT model utilizes hierarchical structure to identify the pattern in data and establish a decision-making rule for estimating the relationship between the independent and dependent variables[57]. Each DT comprises root nodes, child nodes, and leaf nodes while leaf nodes provide the final prediction. DT models can identify complex relationships between variables and handling both categorical and continuous data without any strict data distribution assumption[58]. This study used the DT classifier from the SciKit Learn package[59].

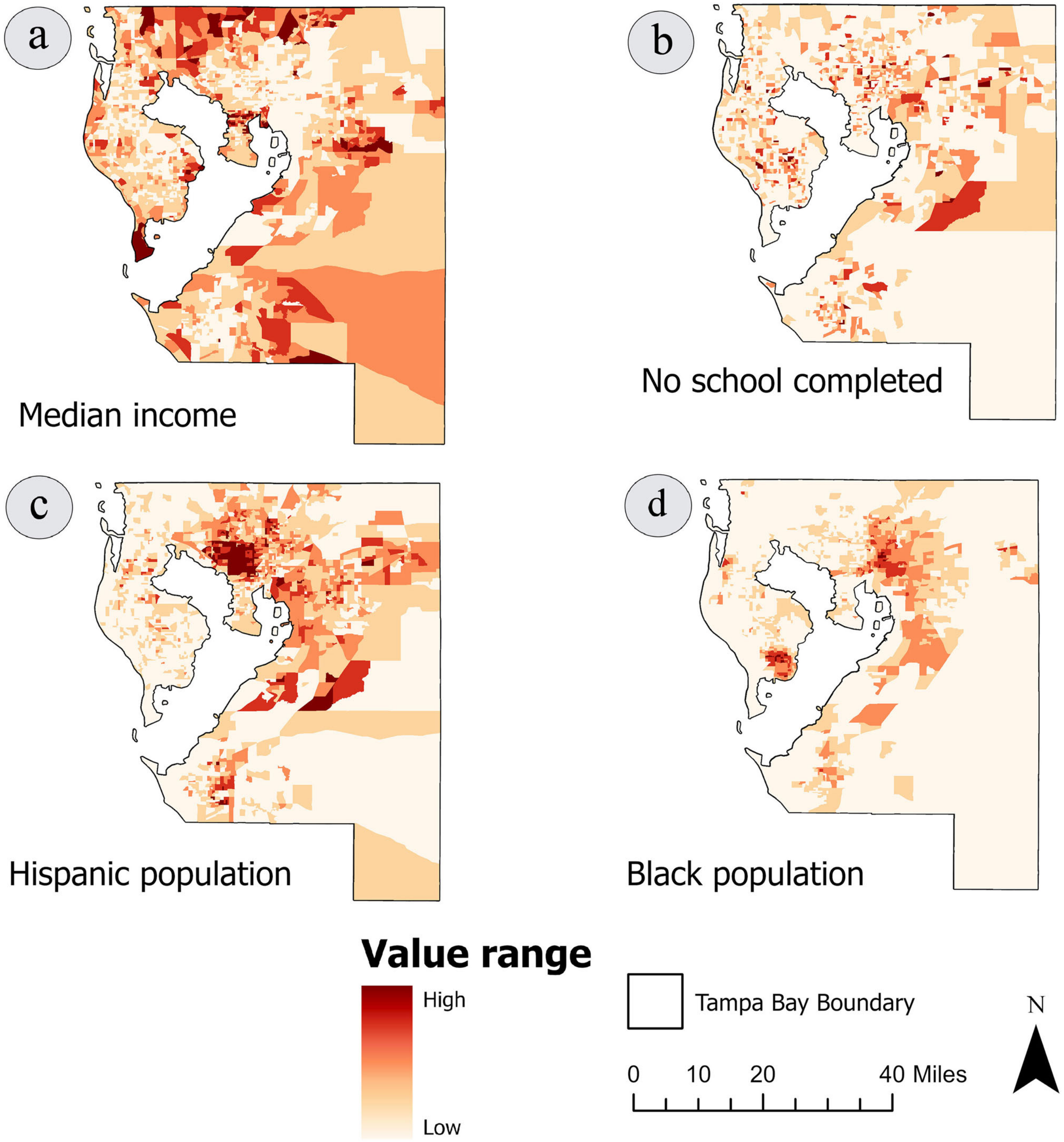

Fig. 9 | Maps of vulnerability-related factors of flood risk. **a** Median income; (**b**) percentage of population with no school completion; (**c**) percentage of Hispanic population; (**d**) percentage of Black (African American) population. These maps were generated in ArcGIS Pro 2.4.0.

Support vector machine (SVM). SVM is a powerful and versatile supervised ML algorithm, able to perform both linear/non-linear classification, and regression. It is also capable of detecting outliers[56]. It is based on statistical learning theory and the structural risk minimization principle that uses training dataset to map original input space into high dimensional feature space[58]. Later, an optimal hyperplane is generated by maximizing the margins of class boundaries to separate the points of different classes. The point above the optimal hyperplane is labeled as +1 and the points below the hyperplane are labeled as −1[60]. The closest training points to the optimal hyperplane are known as support vectors. Once the decision boundary is generated, it can be used to classify new data[61]. This study utilized the SVM classifier from the SciKit Learn package[59].

Adaptive boosting (AdaBoost). Adaptive boosting, commonly known as AdaBoost, is an ensemble ML algorithm that turns a weak learner into a strong learner. A new predictor pays more attention to predecessor training instances that are under-fitted[56]. This algorithm initially trains a base classifier such as DT by assigning equal weight to all training instances and tries to predict them. Later, this algorithm tries to increase the relative weight of misclassified instances and decrease the weights of correctly classified instances[62]. After that, this algorithm trains another classifier with updated weights and makes predictions on the training set and updates the instance weights again. This process continues until a perfect predictor is found[56]. This algorithm provides significant benefits, including solving binary class problems, multi-class single or multi-label problems, and regression

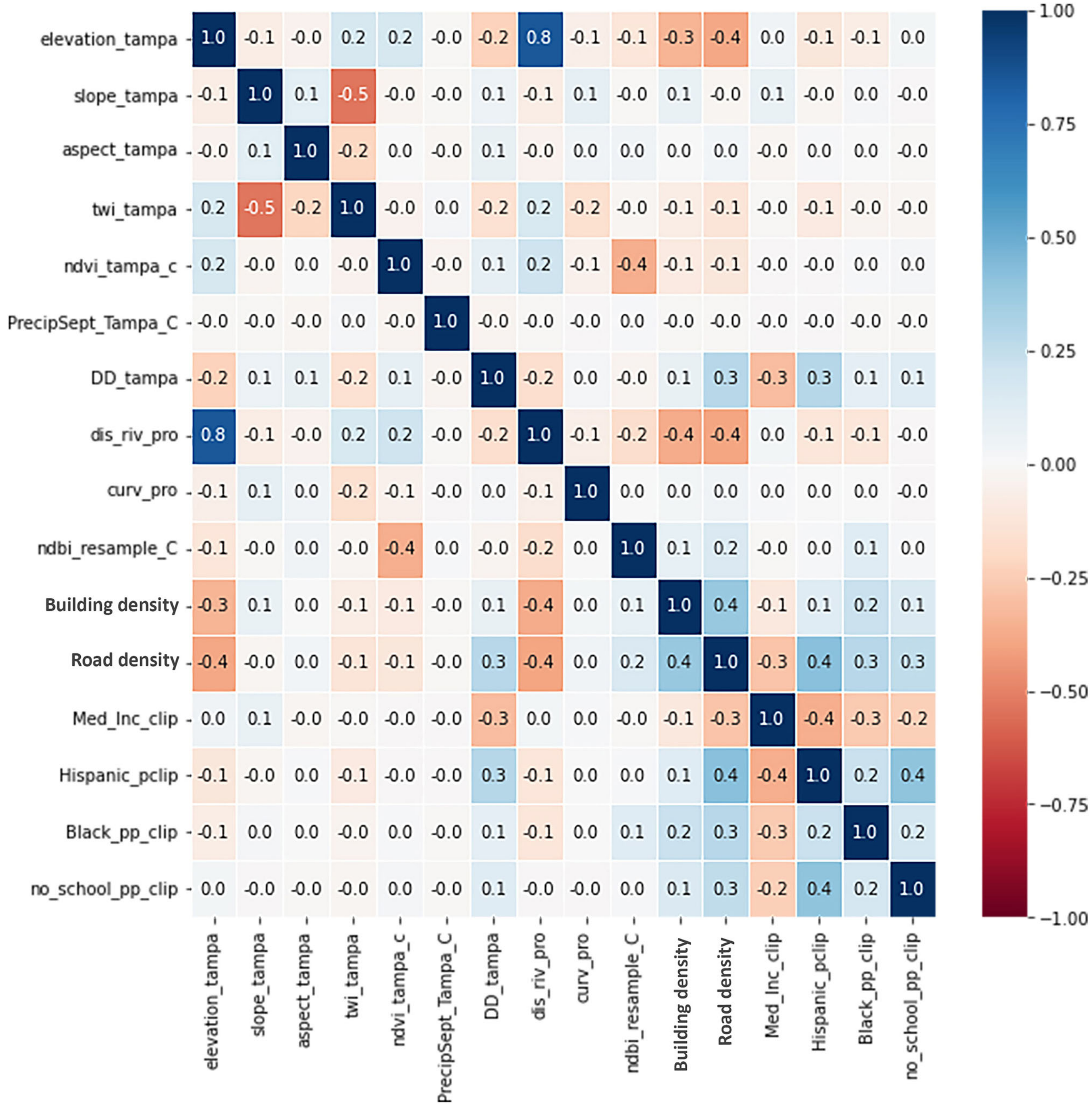

Fig. 10 | Multicollinearity analysis among flood risk factors. This figure displays the correlation among each FRFs. The *X*-axis and primary *Y*-axis represent the name of FRFs. The secondary *Y*-axis represents the value range of correlation analysis (−1 to +1). Blue color indicates positive correlation and red color indicates negative correlation. The data were processed and plotted in Python 3.11.7.

problems[63]. In this study, *AdaBoost classifier* has been used from the SciKit Learn package[59].

Extreme gradient boosting (XGBoost). Extreme gradient boosting is a scalable ML tree-boosting system proposed by Chen and Guestrin[64]. It is mainly designed for superior performance and speed. Instead of averaging independent DTs, XGBoost builds a sequence of DTs by targeting prediction errors or residuals from highly uncertain samples generated by previous tree models[65]. Its numerous tunable parameters and hyperparameter help to increase the predictive accuracy by reducing the overfitting issue and predictive variability[66]. The XGBoost model was conducted by using *XGBclassifier* from the xgboost package.

Random forest (RF). RF, introduced by Breiman[67], is a robust ensemble supervised ML algorithm which can solve both classification and regression problems[56]. This model utilizes bootstrap sampling method to select a subset randomly from training dataset and constructs a DT by recurrently splitting the data based on the best split for each bootstrap sample[68]. Each DT consists of root, child, and leaf nodes. Leaf nodes provide final prediction through majority voting for classification or mean assessment for regression[69]. RF models offer several benefits, such as the abilities to identify significant features, to achieve high predictive accuracy, to reduce overfitting, to handle high dimensional data, insensitivity to noise, and to manage missing value[70,71]. This study utilized the RF classifier from the SciKit Learn package[59]. The default parameters of these ML models that used in this study are listed in Table 5.

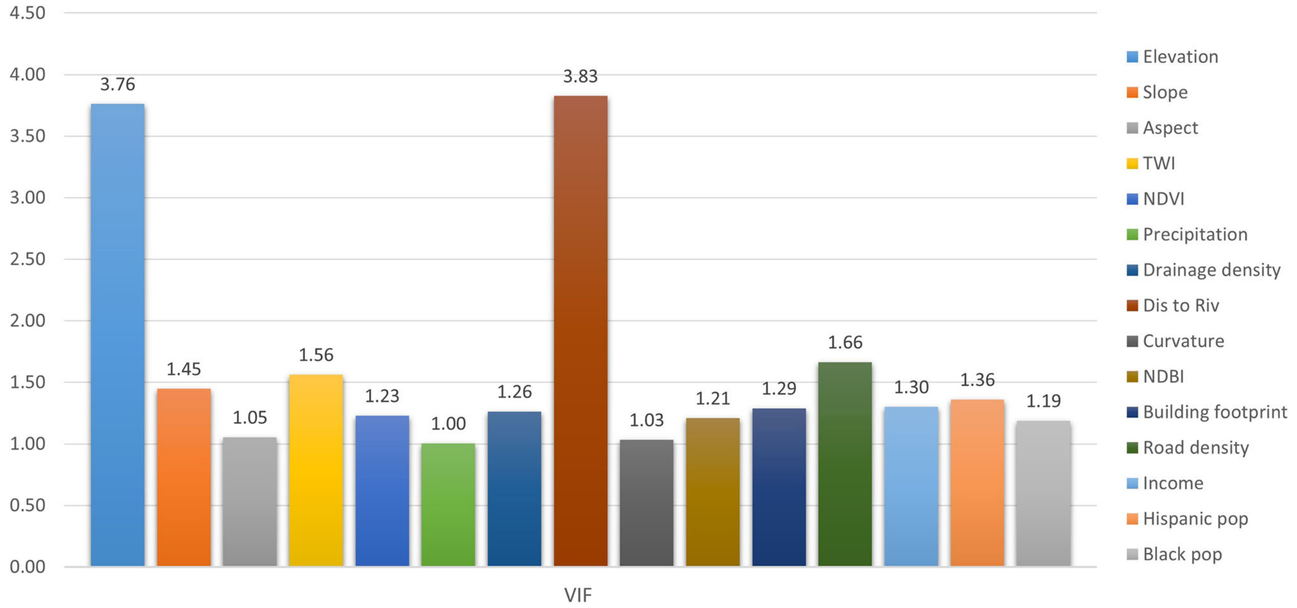

Fig. 11 | VIF score for flood risk factors. *Y*-axis represents the range of VIF score. The VIF analysis was conducted in Python 3.11.7.

Table 5 | The default parameters of the machine learning models used for flood risk simulations

Model name	Description of parameters
DT	criterion = gini, splitter = best, max_depth = None, max_features = None, random_state = 42
SVM	C = 1.0, kernel = rbf; degree = 3, gamma = scale, probability = True, tolerance = 0.001, random_state = 42
AdaBoost	n_estimators = 50, learning_rate = 1.0, algorithm = SAMME.R, random_state = 42
XGBoost	n_estimators = 100, learning_rate = 0.3, max_depth = 6, gamma = 0, booster: gbtree, random_state = 42
RF	n_estimators = 100, criterion = gini, max_depth = None, max_features = sqrt random_state = 42

All these five ML models were trained and tested using flood damage points (target variable) and FRFs (predictor variable). The RF and XGBoost were selected based on their slight better performance on ROC–AUC curve and evaluation metrics for further flood risk simulation. To simulate flood risk index (FRI), *predict_proba* function was used on the RF and XGBoost model. The FRI for each pixel generated from the simulation ranges from 0 to 1, representing the gradual range of flood risk probability. The value close to 0 indicates low flood risk and value close to 1 indicates high flood risk. This study applied equal interval classification method to classify the FRI into five different risk categories including very low (0–0.2), low (0.2–0.4), moderate (0.4–0.6), high (0.6–0.8), and very high (0.8–1.0). Finally, the spatial distribution of flood risk areas across different risk categories was mapped and documented in tables.

Data availability

Data are available and can be obtained by emailing a request to the corresponding author.

Received: 15 February 2024; Accepted: 22 November 2024;
Published online: 06 December 2024

References

1. Rentschler, J., Salhab, M. & Jafino, B. A. Flood exposure and poverty in 188 countries. *Nat. Commun.* **13**, 3527 (2022).
2. Salman, A. M. & Li, Y. Flood risk assessment, future trend modeling, and risk communication: a review of ongoing research. *Nat. Hazards Rev.* **19**, 04018011 (2018).
3. Shah, A. A., Ye, J., Abid, M., Khan, J. & Amir, S. M. Flood hazards: household vulnerability and resilience in disaster-prone districts of Khyber Pakhtunkhwa province, Pakistan. *Nat. Hazards* **93**, 147–165 (2018).
4. Talbot, C. J. et al. The impact of flooding on aquatic ecosystem services. *Biogeochemistry* **141**, 439–461 (2018).
5. Dottori, F. et al. Increased human and economic losses from river flooding with anthropogenic warming. *Nat. Clim. Change* **8**, 781–786 (2018).
6. Bhattarai, S., Parajuli, P. B. & To, F. Comparison of flood frequency at different climatic scenarios in forested coastal watersheds. *Climate* **11**, 41 (2023).
7. Zaharia, L., Costache, R., Prăvălie, R. & Ioana-Toroimac, G. Mapping flood and flooding potential indices: a methodological approach to identifying areas susceptible to flood and flooding risk. Case study: the Prahova catchment (Romania). *Front. Earth Sci.* **11**, 229–247 (2017).
8. Dey, H., Shao, W., Haque, M. M. & VanDyke, M. Enhancing flood risk analysis in Harris County: integrating flood susceptibility and social vulnerability mapping. *J. Geovisualization Spat. Anal.* **8**, 19 (2024).
9. Haque, M. M., Islam, S., Sikder, M. B. & Islam, M. S. Community flood resilience assessment in Jamuna floodplain: a case study in Jamalpur District Bangladesh. *Int. J. Disaster Risk Reduct.* **72**, 102861 (2022).
10. De Moel, H. et al. Flood risk assessments at different spatial scales. *Mitig. Adapt. Strateg. Glob. Change* **20**, 865–890 (2015).
11. VanDyke, M. S., Armstrong, C. L. & Bareford, K. How risk decision-makers interpret and use flood forecast information: assessing the

Mississippi River Outlook email product. *J. Risk Res.* **24**, 1239–1250 (2021).

12. Turner, B. L. et al. A framework for vulnerability analysis in sustainability science. *Proc. Natl. Acad. Sci. USA* **100**, 8074–8079 (2003).
13. Brown, J. & Damery, S. Managing flood risk in the UK: towards an integration of social and technical perspectives. *Trans. Inst. Br. Geogr.* **27**, 412–426 (2002).
14. Gain, A., Mojtahed, V., Biscaro, C., Balbi, S. & Giupponi, C. An integrated approach of flood risk assessment in the eastern part of Dhaka City. *Nat. Hazards* **79**, 1499–1530 (2015).
15. Shao, W., Jackson, N., Ha, H., and, N. & Winemiller, T. Community vulnerability to floods and hurricanes in the U.S. Gulf Coast. *Disasters* **44**, 518–547 (2020).
16. Chen, J., Huang, G. & Chen, W. Towards better flood risk management: assessing flood risk and investigating the potential mechanism based on machine learning models. *J. Environ. Manag.* **293**, 112810 (2021).
17. Firoozishahmirzadi, P., Rahimi, S. & Esmaeili Seraji, Z. Application of machine learning models for flood risk assessment and producing map to identify flood prone areas: literature review. *Int. J. Data Envel. Anal.* **9**, 43–88 (2021).
18. Bui, Q. D. et al. Flood risk mapping and analysis using an integrated framework of machine learning models and analytic hierarchy process. *Risk Anal.* **43**, 1478–1495 (2023).
19. Kabenge, M., Elaru, J., Wang, H. & Li, F. Characterizing flood hazard risk in data-scarce areas, using a remote sensing and GIS-based flood hazard index. *Nat. Hazards* **89**, 1369–1387 (2017).
20. Wagenaar, D. et al. Invited perspectives: how machine learning will change flood risk and impact assessment. *Nat. Hazards Earth Syst. Sci.* **20**, 1149–1161 (2020).
21. Pham, B. T. et al. Flood risk assessment using deep learning integrated with multi-criteria decision analysis. *Knowl.-Based Syst.* **219**, 106899 (2021).
22. Rafiei-Sardooi, E., Azareh, A., Choubin, B., Mosavi, A. H. & Clague, J. J. Evaluating urban flood risk using hybrid method of TOPSIS and machine learning. *Int. J. Disaster Risk Reduct.* **66**, 102614 (2021).
23. Madhuri, R., Sistla, S. & Srinivasa Raju, K. Application of machine learning algorithms for flood susceptibility assessment and risk management. *J. Water Clim. Change* **12**, 2608–2623 (2021).
24. Darabi, H. et al. Urban flood risk mapping using the GARP and QUEST models: a comparative study of machine learning techniques. *J. Hydrol.* **569**, 142–154 (2019).
25. Eini, M., Kaboli, H. S., Rashidian, M. & Hedayat, H. Hazard and vulnerability in urban flood risk mapping: machine learning techniques and considering the role of urban districts. *Int. J. Disaster Risk Reduct.* **50**, 101687 (2020).
26. Liu, J. et al. Assessment of flood susceptibility mapping using support vector machine, logistic regression and their ensemble techniques in the Belt and Road region. *Geocarto Int.* **37**, 9817–9846 (2022).
27. Prakash, A. J., Begam, S., Vilímek, V., Mudi, S. & Das, P. Development of an automated method for flood inundation monitoring, flood hazard, and soil erosion susceptibility assessment using machine learning and AHP–MCE techniques. *Geoenviron. Disasters* **11**, 14 (2024).
28. Taromideh, F., Fazloula, R., Choubin, B., Emadi, A. & Berndtsson, R. Urban flood-risk assessment: integration of decision-making and machine learning. *Sustainability* **14**, 4483 (2022).
29. Dey, H., Shao, W., Moradkhani, H., Keim, B. D. & Peter, B. G. Urban flood susceptibility mapping using frequency ratio and multiple decision tree-based machine learning models. *Nat. Hazards* **120**, 10365–10393 (2024).
30. Yarveysi, F., Alipour, A., Moftakhari, H., Jafarzadegan, K. & Moradkhani, H. Block-level vulnerability assessment reveals disproportionate impacts of natural hazards across the conterminous United States. *Nat. Commun.* **14**, 4222 (2023).
31. Xie, W. & Meng, Q. An integrated PCA–AHP method to assess urban social vulnerability to sea level rise risks in Tampa, Florida. *Sustainability* **15**, 2400 (2023).
32. Fu, X. & Peng, Z. R. Assessing the sea-level rise vulnerability in coastal communities: a case study in the Tampa Bay Region, US. *Cities* **88**, 144–154 (2019).
33. Bacopoulos, P. Extreme low and high waters due to a large and powerful tropical cyclone: Hurricane Irma (2017). *Nat. Hazards* **98**, 939–968 (2019).
34. Shao, W., Feng, K. & Lin, N. Predicting support for flood mitigation based on flood insurance purchase behavior. *Environ. Res. Lett.* **14**, 054014 (2019).
35. Shao, W. et al. Understanding the effects of past flood events, perceived and estimated flood risks on individuals' voluntary flood insurance purchase behaviors. *Water Res.* **108**, 391–400 (2017).
36. El-Magd, S. A. A., Pradhan, B. & Alamri, A. Machine learning algorithm for flash flood prediction mapping in Wadi El-Laqeita and surroundings, Central Eastern Desert, Egypt. *Arab. J. Geosci.* **14**, 1–14 (2021).
37. Ma, M. et al. XGBoost-based method for flash flood risk assessment. *J. Hydrol.* **598**, 126382 (2021).
38. Sanders, W., Li, D., Li, W. & Fang, Z. N. Data-driven flood alert system (FAS) using extreme gradient boosting (XGBoost) to forecast flood stages. *Water* **14**, 747 (2022).
39. Desalegn, H. & Mulu, A. Flood vulnerability assessment using GIS at Fetam watershed, upper Abbay basin, Ethiopia. *Heliyon* **7**, e05865 (2021).
40. Ziarh, G. F., Asaduzzaman, M., Dewan, A., Nashwan, M. S. & Shahid, S. Integration of catastrophe and entropy theories for flood risk mapping in peninsular Malaysia. *J. Flood Risk Manag.* **14**, e12686 (2021).
41. Hoque, M. A. A., Ahmed, N., Pradhan, B. & Roy, S. Assessment of coastal vulnerability to multi-hazardous events using geospatial techniques along the eastern coast of Bangladesh. *Ocean Coast. Manag.* **181**, 104898 (2019).
42. Vojtek, M. & Vojteková, J. Flood susceptibility mapping on a national scale in Slovakia using the analytical hierarchy process. *Water* **11**, 364 (2019).
43. Mukherjee, F. & Singh, D. Detecting flood prone areas in Harris County: a GIS based analysis. *GeoJournal* **85**, 647–663 (2020).
44. Vafakhah, M., Mohammad Hasani Loor, S., Pourghasemi, H. & Katebikord, A. Comparing performance of random forest and adaptive neuro-fuzzy inference system data mining models for flood susceptibility mapping. *Arab. J. Geosci.* **13**, 1–16 (2020).
45. Haque, M. M., Islam, S., Sikder, M. B., Islam, M. S. & Tabassum, A. Assessment of flood vulnerability in Jamuna floodplain: a case study in Jamalpur district, Bangladesh. *Nat. Hazards* **116**, 341–363 (2023).
46. Choubin, B. et al. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Sci. Total Environ.* **651**, 2087–2096 (2019).
47. Mojaddadi, H., Pradhan, B., Nampak, H., Ahmad, N. & Ghazali, A. H. B. Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS. *Geomat. Nat. Hazards Risk* **8**, 1080–1102 (2017).
48. Torresan, S., Critto, A., Rizzi, J. & Marcomini, A. Assessment of coastal vulnerability to climate change hazards at the regional scale: the case study of the North Adriatic Sea. *Nat. Hazards Earth Syst. Sci.* **12**, 2347–2368 (2012).
49. Rhubart, D. & Sun, Y. The social correlates of flood risk: variation along the US rural–urban continuum. *Popul. Environ.* **43**, 232–256 (2021).

50. Song, J. et al. Resilience-vulnerability balance to urban flooding: a case study in a densely populated coastal city in China. *Cities* **95**, 102381 (2019).
51. Wang, Z., Huang, J., Wang, H., Kang, J. & Cao, W. Analysis of flood evacuation process in vulnerable community with mutual aid mechanism: an agent-based simulation framework. *Int. J. Environ. Res. Public Health* **17**, 560 (2020).
52. Koks, E. E., Jongman, B., Husby, T. G. & Botzen, W. J. Combining hazard, exposure and social vulnerability to provide lessons for flood risk management. *Environ. Sci. Policy* **47**, 42–52 (2015).
53. Bin, L., Xu, K., Pan, H., Zhuang, Y. & Shen, R. Urban flood risk assessment characterizing the relationship among hazard, exposure, and vulnerability. *Environ. Sci. Pollut. Res.* **30**, 86463–86477 (2023).
54. Özay, B. & Orhan, O. Flood susceptibility mapping by best–worst and logistic regression methods in Mersin, Turkey. *Environ. Sci. Pollut. Res.* **30**, 45151–45170 (2023).
55. Khosravi, K. et al. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Sci. Total Environ.* **627**, 744–755 (2018).
56. Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* ("O'Reilly Media, Inc.", 2022).
57. Han, J., Kim, J., Park, S., Son, S. & Ryu, M. Seismic vulnerability assessment and mapping of Gyeongju, South Korea using frequency ratio, decision tree, and random forest. *Sustainability* **12**, 7787 (2020).
58. Tien Bui, D., Pradhan, B., Lofman, O. & Revhaug, I. Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naive Bayes Models. *Math. Probl. Eng.* **2012**, 974638 (2012).
59. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
60. Tehrany, M. S., Lee, M. J., Pradhan, B., Jebur, M. N. & Lee, S. Flood susceptibility mapping using integrated bivariate and multivariate statistical models. *Environ. Earth Sci.* **72**, 4001–4015 (2014).
61. Tehrany, M. S., Pradhan, B. & Jebur, M. N. Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. *Stoch. Environ. Res. Risk Assess.* **29**, 1149–1165 (2015).
62. Al-Abadi, A. M. Mapping flood susceptibility in an arid region of southern Iraq using ensemble machine learning classifiers: a comparative study. *Arab. J. Geosci.* **11**, 1–19 (2018).
63. Hong, H. et al. Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *Catena* **163**, 399–413 (2018).
64. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery*. pp. 785–794 (2016).
65. Abedi, R., Costache, R., Shafizadeh-Moghadam, H. & Pham, Q. B. Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. *Geocarto Int.* **37**, 5479–5496 (2022).
66. Hasanuzzaman, M., Islam, A., Bera, B. & Shit, P. K. A comparison of performance measures of three machine learning algorithms for flood susceptibility mapping of river Silabati (tropical river, India). *Phys. Chem. Earth A/B/C* **127**, 103198 (2022).
67. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
68. Elmahdy, S. I., Mohamed, M. M., Ali, T. A., Abdalla, J. E. D. & Abouleish, M. Land subsidence and sinkholes susceptibility mapping and analysis using random forest and frequency ratio models in Al Ain, UAE. *Geocarto Int.* **37**, 315–331 (2022).
69. Lee, S., Kim, J. C., Jung, H. S., Lee, M. J. & Lee, S. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomat. Nat. Hazards Risk* **8**, 1185–1203 (2017).
70. Youssef, A. M., Pourghasemi, H. R. & El-Haddad, B. A. Advanced machine learning algorithms for flood susceptibility modeling—performance comparison: Red Sea, Egypt. *Environ. Sci. Pollut. Res.* **29**, 66768–66792 (2022).
71. Amare, S. et al. Susceptibility to gully erosion: applying random forest (RF) and frequency ratio (FR) approaches to a small catchment in Ethiopia. *Water* **13**, 216 (2021).
72. Pekel, J. F., Cottam, A., Gorelick, N. & Belward, A. S. High-resolution mapping of global surface water and its long-term changes. *Nature* **540**, 418–422 (2016).
73. Samanta, S., Pal, D. K. & Palsamanta, B. Flood susceptibility analysis through remote sensing, GIS and frequency ratio model. *Appl. Water Sci.* **8**, 66 (2018).
74. Rahmati, O. et al. Development of novel hybridized models for urban flood susceptibility mapping. *Sci. Rep.* **10**, 12937 (2020).
75. Rahmati, O., Pourghasemi, H. R. & Zeinivand, H. Flood susceptibility mapping using frequency ratio and weights-of-evidence models in the Golastan Province, Iran. *Geocarto Int.* **31**, 42–70 (2016).
76. Farhadi, H. & Najafzadeh, M. Flood risk mapping by remote sensing data and random forest technique. *Water* **13**, 3115 (2021).
77. Sarkar, D. & Mondal, P. Flood vulnerability mapping using frequency ratio (FR) model: a case study on Kulik river basin, Indo-Bangladesh Barind region. *Appl. Water Sci.* **10**, 1–13 (2020).
78. Islam, A. R. M. T. et al. Flood susceptibility modelling using advanced ensemble machine learning models. *Geosci. Front.* **12**, 101075 (2021).
79. Ali, S. A., Khatun, R., Ahmad, A. & Ahmad, S. N. Application of GIS-based analytic hierarchy process and frequency ratio model to flood vulnerable mapping and risk area estimation at Sundarban region, India. *Model. Earth Syst. Environ.* **5**, 1083–1102 (2019).
80. Darabi, H. et al. Urban flood risk mapping using data-driven geospatial techniques for a flood-prone case area in Iran. *Hydrol. Res.* **51**, 127–142 (2020).
81. Dey, H., Shao, W., Pan, S. & Tian, H. The spatiotemporal patterns of community vulnerability in the US Mobile Bay from 2000–2020. *Appl. Spat. Anal. Policy* **17**, 371–392 (2023).
82. Kusmiyarti, T. B., Wiguna, P. P. K. & Dewi, N. R. Flood risk analysis in Denpasar City, Bali, Indonesia. In *IOP Conference Series: Earth and Environmental Science* Vol. 123, p. 012012 (IOP Publishing, 2018).
83. Rufat, S., Tate, E., Burton, C. G. & Maroof, A. S. Social vulnerability to floods: review of case studies and implications for measurement. *Int. J. Disaster Risk Reduct.* **14**, 470–486 (2015).

Acknowledgements

Funding for this project was provided by the National Oceanic and Atmospheric Administration (NOAA), awarded to the Cooperative Institute for Research on Hydrology (CIROH) through the NOAA Cooperative Agreement with The University of Alabama, NA22NWS4320003.

Author contributions

H.D. and W.S. conceived the study with inspiration from their previous studies. H.D. collected and analyzed the data, prepared the figures and tables. H.D., M.H., W.S., M.V., and F.H. contributed to the writing of the manuscript text. H.D., M.H., and W.S. contributed to the editing of the manuscript text. W.S. and M.V. acquired the funding.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Wanyun Shao.

Reprints and permissions information is available at http://www.nature.com/reprints