Science of the Total Environment 851 (2022) 158002

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

# A novel flood risk mapping approach with machine learning considering geomorphic and socio-economic vulnerability dimensions

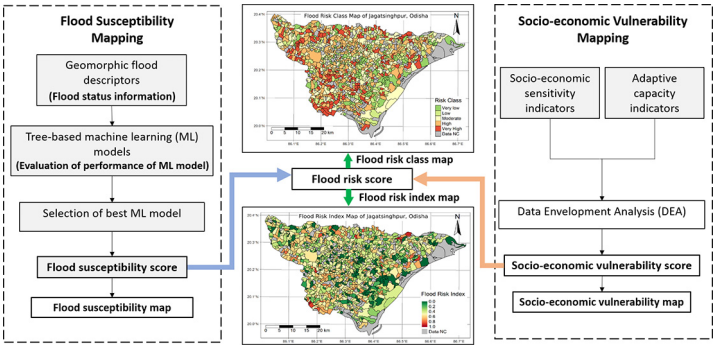
Prakhar Deroliya [a], Mousumi Ghosh [b], Mohit P. Mohanty [a,c], Subimal Ghosh [b,d], K.H.V. Durga Rao [f], Subhankar Karmakar [a,b,e,*]

[a] Environmental Science and Engineering Department, Indian Institute of Technology Bombay, Mumbai 400076, India
[b] Interdisciplinary Program in Climate Studies, Indian Institute of Technology Bombay, Mumbai 400076, India
[c] Department of Water Resources Development and Management, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand 247667, India
[d] Department of Civil engineering, Indian Institute of Technology Bombay, Mumbai 400076, India
[e] Centre for Urban Science and Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India
[f] Disaster Management Support Group, National Remote Sensing Centre, Indian Space Research Organization, Hyderabad, India

## HIGHLIGHTS

- A novel approach for *flood risk* mapping under resource-constrained scenarios is proposed.
- Tree-based ML models are evaluated to estimate *flood susceptibility* using DTM-derived GFDs.
- An efficient DEA-based approach is employed to map *socio-economic vulnerability*.
- *Flood risk* is derived by combining *flood susceptibility* and *socio-economic vulnerability*.
- A GIS-based *flood risk* map is developed at the finest administrative-level.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

Editor: Fernando A.L. Pacheco

*Keywords:*
Data envelopment analysis
Flood risk assessment
Flood susceptibility mapping
Geomorphic approach
Supervised learning
Vulnerability mapping

## ABSTRACT

Quantifying flood hazards by employing hydraulic/hydrodynamic models for flood risk mapping is a widely implemented non-structural flood management strategy. However, the unavailability of multi-domain and multi-dimensional input data and expensive computational resources limit its application in resource-constrained regions. The fifth and sixth IPCC assessment reports recommend including *vulnerability* and *exposure* components along with *hazards* for capturing *risk* on human-environment systems from natural and anthropogenic sources. In this context, the present study showcases a novel *flood risk* mapping approach that considers a combination of geomorphic flood descriptor (GFD)-based *flood susceptibility* and often neglected *socio-economic vulnerability* components. Three popular Machine Learning (ML) models, namely Decision Tree (DT), Random Forest (RF), and Gradient-boosted Decision Trees (GBDT), are evaluated for their abilities to combine digital terrain model-derived GFDs for quantifying flood susceptibility in a flood-prone district, Jagatsinghpur, located in the lower Mahanadi River basin, India. The area under

*Abbreviations:* AUC, Area under ROC curve; CCR, Charnes-Cooper-Rhodes; CI, Convergence index; DEA, Data envelopment analysis; DEM, Digital elevation model; DI, Downslope index; DT, Decision tree; DTM, Digital terrain model; FI, Feature importance; FP, False positives; FN, False negatives; FR, Flood risk; FRI, Flood risk index; FS, Flood susceptibility; GBDT, Gradient-boosted decision trees; GFD, Geomorphic flood descriptor; GFI, Geomorphic flood index (stream); HHM, Hydraulic-hydrodynamic modeling; LGFI, Geomorphic Flood Index (local); MFGFD, Multicollinearity-free geomorphic flood descriptors; MRVBF, Multi-resolution valley bottom flatness index; RF, Random forest; ROC, Receiver operating characteristic; SEV, Socio-economic vulnerability; SU, Spatial unit; TI, Topographic index; TN, True negatives; TP, True positives; VDCN, Vertical distance to channel network; VIF, Variance inflation factor; wd, Floodwater depth; wv, Floodwater velocity.

* Corresponding author at: Environmental Science and Engineering Department, Indian Institute of Technology Bombay, Mumbai 400076, India.
*E-mail address:* skarmakar@iitb.ac.in (S. Karmakar).

http://dx.doi.org/10.1016/j.scitotenv.2022.158002
Received 28 May 2022; Received in revised form 31 July 2022; Accepted 9 August 2022
Available online 17 August 2022

receiver operating characteristics curve (*AUC*) along with Cohen's kappa are used to identify the best ML model. It is observed that the RF model performs better compared to the other two models on both training and testing datasets, with *AUC* score of 0.88 on each. The socio-economic vulnerability assessment follows an indicator-based approach by employing the Charnes-Cooper-Rhodes (CCR) model of Data Envelopment Analysis (DEA), an efficient non-parametric ranking method. It combines the district's relevant *socio-economic sensitivity* and *adaptive capacity* indicators. The *flood risk* classes at the most refined administrative scale, i.e., village level, are determined with the Jenks natural breaks algorithm using *flood susceptibility* and *socio-economic vulnerability* scores estimated by the RF and CCR-DEA models, respectively. It was observed that >40 % of the villages spread over Jagatsinghpur face *high* and *very high* flood risk. The proposed novel framework is generic and can be used to derive a wide variety of *flood susceptibility*, *vulnerability*, and subsequently *risk* maps under a data-constrained scenario. Furthermore, since this approach is relatively data and computationally parsimonious, it can be easily implemented over large regions. The exhaustive flood maps will facilitate effective flood control and floodplain planning.

## 1. Introduction

Floods are the most widespread and the third most calamitous natural hazard globally. There were nearly 1298 major flood events between 2010 and 2019 in the world, causing US$33.7 billion in economic loss (IFRC, 2020). In India alone, >100,000 people lost their lives between 1953 and 2018 due to major flood events, which incurred ₹3000 billion worth of damage to public utilities (CWC, 2019). Alarmingly, the country also has the most socio-economically deprived population exposed to flood risk (Rentschler and Salhab, 2020). Unfortunately, with the continual rise of global temperature (Nilawar and Waikar, 2019; Schiermeier, 2011), urbanization (Shastri et al., 2015), deforestation (Bradshaw et al., 2007), and greenhouse emissions (IPCC, 2013), both the intensity and the frequency of flood events are expected to rise, which is a severe threat especially for developing nations such as India. Thus, flood risk assessment is critical strategy for developing appropriate resilience pathways (Mehryar and Surminski, 2022) and for stakeholders' allocation and management of water resources (Kapetas et al., 2019).

"Risk" is defined as a measure of the probability and severity of adverse effects (Lowrance, 1976). Objectively, "flood risk" arises due to the interaction of hazard, exposure, and vulnerability dimensions (Kron, 2005; Barredo et al., 2007; IPCC, 2021). Mapping flood risk by aggregating hazard, exposure, and vulnerability helps to derive crucial information for informed decision-making (Prinos et al., 2008). In the context of flood risk, socio-economic vulnerability assessment is often neglected, but is a vital component for understanding 'to whom' and 'to what degree' the adverse impacts prevail in a society (Mishra and Sinha, 2020; Mohanty et al., 2020a,b). Especially developing nations face high flood risk due to their populace's high socio-economic sensitivity and low adaptive capacity (Ward et al., 2017). It highlights the need for administrative authorities, researchers, and water resource planners to come together to strategize and implement various flood management options to lessen the risk and damages associated with these hydro-climatic extremes (Dilley et al., 2005). Structural and non-structural measures are the two primary options for flood risk management and mitigation (Mohapatra and Singh, 2003; Yamini et al., 2020; Kuriqi and Hysa, 2021; Ardiclioglu et al., 2022). Of these options, the latter is more sustainable due to its reversibility, acceptability, and environment-friendliness (Kundzewicz, 2002). Assessment and mapping of flood risk elements is one such well-known non-structural measure and is used to identify flood hotspots (Díez-Herrero et al., 2009). Flood hazard is a key element of flood risk and has been characterized by flood extent (Manfreda et al., 2014), floodwater depth ($w_d$) and floodwater velocity ($w_v$), the combination of $w_d$ and $w_v$ (Mohanty et al., 2020a,b), insurance rate at a region (Burby, 2001), flooding area (Shareef and Abdulrazzaq, 2021), flooding susceptibility (Rahman et al., 2019), pedological distribution (Sangwan and Merwade, 2015), bankfull discharge (Chau and Thanh, 2021), among others. Hydraulic-hydrodynamic modeling (HHM) approaches have been employed for estimating flood extent, $w_d$, and $w_v$ at fine resolutions, i.e., 1 m resolution. However, these complex models not only require multi-domain and multi-dimensional input data but also require intensive computational facility and financial resources. Therefore, their application is limited in data- and resource-constrained scenario, mostly an issue of developing nations (Samela et al., 2018). Various alternative approaches have been researched and adopted to map flood hazard under such circumstances. Many of these alternative approaches use digital terrain models (DTM) and other remotely sensed products to estimate flood hazard (Noman et al., 2001; Pradhan, 2010).

These approaches suggest that a region's underlying geomorphological signature influences its flooding behavior (Hack and Goodlett, 1960). Various geomorphic flood descriptors (GFDs) such as topographic index (Beven and Kirkby, 1979), multi-resolution valley bottom flatness index (Gallant and Dowling, 2003), downslope index (Hjerdt et al., 2004), modified topographic index (Manfreda et al., 2011), height above the nearest drainage (Nobre et al., 2011), geomorphic flood index (Manfreda et al., 2014), DTM-derived flood depth (Nardi et al., 2006), and others represent a variety of hydro-geomorphological relationships. These descriptors have been extensively used for estimating natural hazards such as floods (Ghosh et al., 2019; Papaioannou et al., 2015), landslides (Ayalew and Yamagishi, 2005), and even forest fires (Pourghasemi et al., 2020). The extensive adoption of GFDs for quantifying the severity of these natural hazards is primarily because of the unrestricted and near-global availability of remotely sensed products such as DTMs (Gorokhovich and Voustianiouk, 2006). In the context of flood risk assessment, GFDs have been used for flood extent mapping (Degiorgis et al., 2012), estimation of flood depth (Manfreda and Samela, 2019), and flood hazard and susceptibility mapping (Pradhan, 2010; Mishra and Sinha, 2020).

Flood susceptibility has often been used to characterize flood hazard. It is the propensity of a region to flooding due to its physical attributes (Vojtek and Vojteková, 2019). A recently popular way to estimate flood susceptibility is by using machine learning (ML) models, which can combine geographic information system (GIS) layers of physical attributes into a flood susceptibility layer. This approach is more accessible and economical compared to HHM approaches because it requires less data and resources. Generally, a group of preselected ML models is trained on a small sample of training data consisting of the values of physical attributes at a location with its corresponding flooding status (flooded or non-flooded). Post-training, the probabilistic output of the best ML model is generalized over the study area. The need to test a group of models arises because no single model works on all possible data distributions (Kotsiantis et al., 2007). Numerous ML models such as logistic regression (Pradhan, 2010), support vector machines (Tehrany et al., 2015), tree-based classifiers (Lee et al., 2017), multilayer perceptron (Janizadeh et al., 2019), convolutional neural networks (Wang et al., 2020), and others have been used for delineating flood hotspots. ML models have also been used for modeling stage-discharge-sediment (Kumar et al., 2022) and mapping flood probability (Avand et al., 2022). Since ML-based flood susceptibility assessment approach is affordable considering the size and variety of data required, it can be employed to estimate the hazard component of risk. However, majority of these studies have not considered vulnerability, which plays a governing role in sensitivity, adaptability and resilience of a society, during flood events.

Vulnerability resides at the core of risk, and it impacts various aspects of ecosystem such as hydroclimatic extremes (Vittal et al., 2020), agriculture (Sharma et al., 2020), groundwater (Kazakis et al., 2015), and others. It refers to the susceptibility of individuals, communities, or systems to risk (United Nations International Strategy for Disaster Reduction, 2009). Socio-economic vulnerability is expressed as a combination of the socio-economic sensitivity and the lack of adaptive capacity of the populace in a region. Including socio-economic vulnerability in flood risk assessment is especially timely in the era of socio-hydrology but is still largely ignored (Mohanty et al., 2020a,b). Conventionally used vulnerability assessment methods, such as the analytical hierarchical process (AHP), the technique for order performance by similarity to ideal solution (TOPSIS), among others, introduce subjectivity because these methods require expert consultation to agree upon the weights of vulnerability indicators (Sherly et al., 2015). It is understood that choices such as selecting vulnerability indicators and aggregation method are subjective and largely depend on data and resource availability, causing the elimination of subjectivity in vulnerability assessment impossible. Nevertheless, the subjectivity due to the weighting scheme of indicators derived from expert consultation, can be removed by using data envelopment analysis (DEA) models, which have been widely adopted for socio-economic vulnerability assessment (Sherly et al., 2015; Vittal et al., 2020).

In light of the existing literature, the current study proposes a novel framework based on ML and DEA for flood risk quantification to address the above-mentioned gaps. The framework considers hazard as well as socio-economic vulnerability under data-scarce situations. ML-based flood susceptibility mapping approach is utilized as an affordable alternative to conventional HHM approaches, as it overcomes the computational cost and time associated with HHM approaches significantly without compromising accuracy. In a noteworthy step, the study considers socio-economic vulnerability dimensions as an integral component of flood risk that is usually neglected in the literature on flood risk management. In doing this, our study aligns its approach for flood risk mapping according to the definition recommended in IPCC's Fifth and Sixth Assessment Reports. Hence, the present research considering both these dimensions within a sophisticated framework is a crucial step toward flood risk mapping, advantageous for flood-prone regions in low, and middle-income nations. This study is demonstrated over a severely flood-prone region in the lower Mahanadi River basin in India to derive flood risk map at the finest administrative level. The primary goal of this work is to develop a framework that: (a) requires less data considering both the availability of reliable public data at fine resolution and (b) maps flood risk accurately with low subjectivity. In previous studies, large-scale floodplain delineation tools that require only DTM-derived flood descriptors have been developed (Samela et al., 2018). With this context, the proposed framework requires only the DTM of the area under study for flood susceptibility mapping, which makes the framework data parsimonious without sacrificing accuracy. Additionally, these descriptors intrinsically capture hydro-geomorphological, lithological, and even geo-environmental attributes (MacMillan et al., 2004; Odeha et al., 1994; Yang et al., 2005). Simultaneously, socio-economic vulnerability of the entire study area based on the available demographic information from recent census data for an efficient non-parametric DEA is quantified. Finally, the socio-economic vulnerability and flood susceptibility are amalgamated and mapped in GIS platform to derive flood risk maps at the village level. The proposed framework is data and computationally parsimonious, which promises its selection as an effective tool for evaluating flood risk over hydrologically and geographically complex areas. This research article is organized in five sections. Section 1 provides an insight into the relevant literature followed by the objectives of the current study. The description of the study area and data used are described in Section 2. The proposed methodology which illustrates the estimation of flood susceptibility and socio-economic vulnerability is presented in Section 3. Section 4 describes results and discussion pertaining to the study. The major outcomes of the study are enumerated in Section 5.

## 2. Study area and data description

Jagatsinghpur district (19°58′ N to 20°23′ N and 86°3′ E to 86°45′ E), located in Odisha state (Fig. 1), is one of the most flood-affected districts of India (Ghosh et al., 2019). It spreads for an area of 1668 km$^2$ and accommodates a population of around 11,36,971 (Census of India, 2011). The district has 1320 villages. The Mahanadi river and its distributaries delineate the northern side of the district, whereas the southern side is surrounded by the Devi, the Kathajodi, and the Biluakhai rivers. The district is part of both the Mahanadi and Devi deltas. Its eastern boundary aligns with the Bay of Bengal. The district is situated in a coastal plain zone as per agro-climatic classification and in deltaic alluvial plains of the river system with several estuaries, creeks on the coastal belt. Three types of

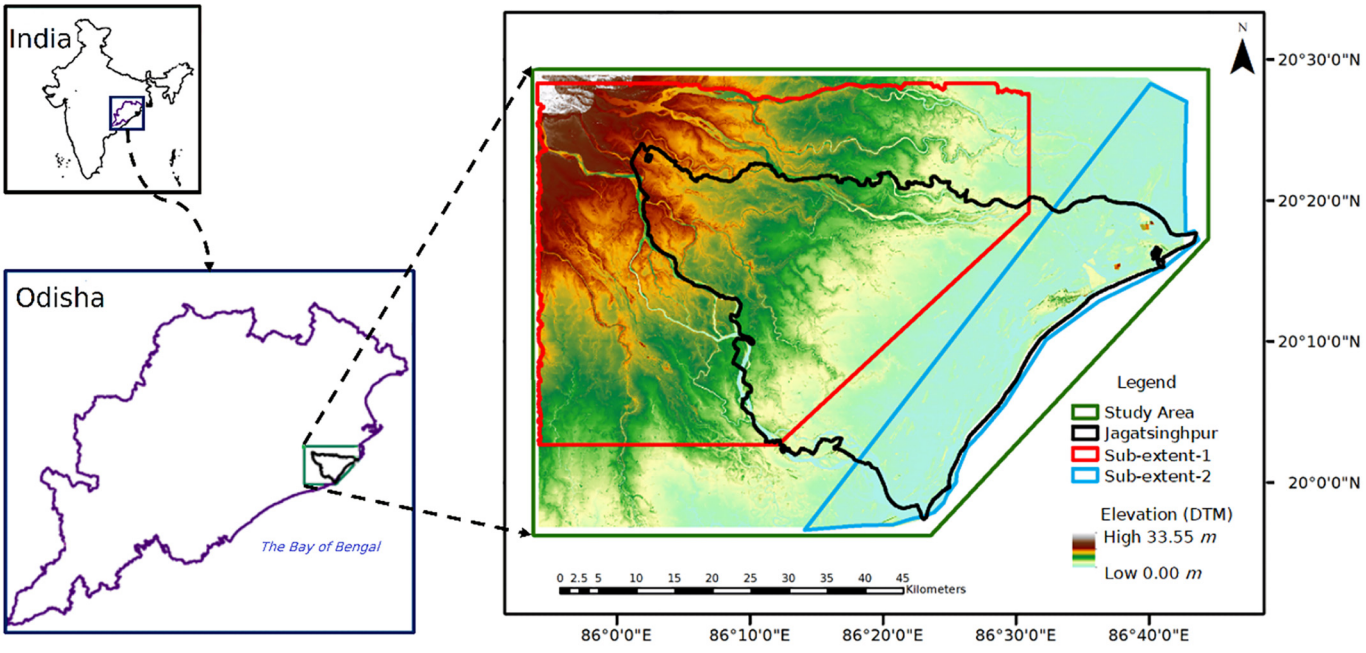



**Fig. 1.** Location map of Jagatsinghpur district. *Flooded* and *non-flooded* points were collected from Sub-extent-1.

soils of the district are laterite (89,700 thousand ha), coastal saline (53,700 thousand ha), and deep alluvial (20,500 thousand ha). The average annual rainfall is around 1500 mm (Baliarsingh et al., 2018). The district is one of the highly populated coastal districts of Odisha with temperate climate (maximum temperature: 38 °C and minimum temperature: 12 °C). Roughly 80 % of the annual rainfall occurs in the monsoon period. The district is prone to cyclonic rainfalls during the monsoons and has been experiencing natural calamities like floods and cyclones which significantly hinder its economic development.

The economy of the district is predominantly agrarian as nearly 80 % of its population depends on agriculture- and cultivation-related activities for their survival. The land use land cover of the district primarily comprises nearly 63 % agricultural land, 3.5 % forest cover, and 32 % under waste land (DPMU, 2017). Paddy is the primary food crop grown, while cotton, sugarcane, and turmeric are the major commercial crops. Owing to the flat terrain (elevation: 0–25 m above mean-sea level) of the district and heavy sediment load from upstream, the surrounding and inhabiting rivers and rivulets quickly overflow during the monsoon season (Mishra and Mishra, 2010). Agriculture and cultivation-dependent populations are among the worst affected communities by natural hazards, especially by droughts and floods (Mishra and Mishra, 2010). Hence, the agriculture and cultivation-driven economy and the geographical position of the district render it socio-economically vulnerable and physically susceptible to floods. Such geomorphological and spatial configuration of the district, warm and humid climate and high rainfall renders it vulnerable to a trio of flooding (coastal, fluvial and pluvial). The studies conducted over this region on flood risk management such as Mohanty et al. (2020a,b) and Gusain et al. (2020) have considered flood hazard and vulnerability components to quantify risk. However, the computational and data requirements in these frameworks are computationally demanding. The proposed novel approach attempts to overcome these limitations by adopting an ML-DEA-based approach to quantify flood risk at the finest administrative scale (village level) over the study area.

In the present study, the DTM of the district was generated from the indigenously developed national digital elevation model, CartoDEM (generated using CartoSat-1 stereo pairs, obtained from National Remote Sensing Centre (NRSC), Hyderabad, India) (Muralikrishnan et al., 2013), as shown in Fig. 1, which is used to generate the GFDs of the framework (Refer Table S-1). The socio-economic data for Jagatsinghpur town and 1072 villages was collected from the Census of India. Table S-2 enumerates the SEV indicators used for the present study. Recently, Mohanty et al. (2020a,b), Mohanty and Karmakar (2021) carried out a comprehensive fine-scale $w_d$ and $w_v$ mapping of the district using coupled 1D-2D hydrodynamic modeling. The $w_d$ grid at 50 years return period from Mohanty et al. (2020a,b) was considered as the ground truth for identifying flooded and non-flooded locations and for validating the performance of the flood susceptibility mapping module of the framework.

## 3. Methodology

The proposed framework is illustrated in Fig. 2. It consists of three components: (i) estimating flood susceptibility using ML, (ii) calculating

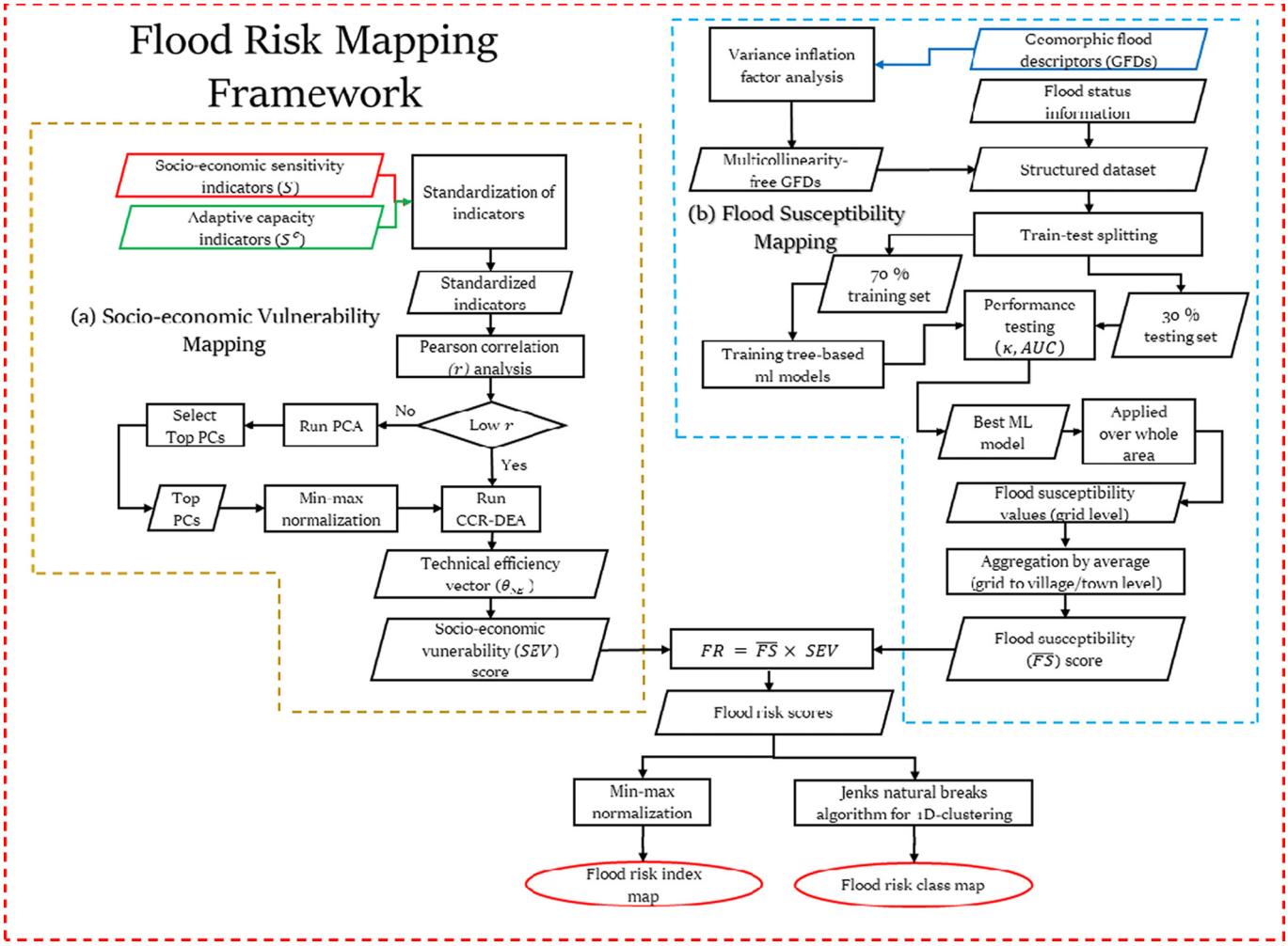



**Fig. 2.** Proposed flood risk mapping framework used in the study.

socio-economic vulnerability using DEA, and (iii) quantification and mapping of flood risk. Based on an exhaustive literature survey (Manfreda et al., 2014; Papaioannou et al., 2015; Pradhan, 2010), twenty GFDs, which require only DTM or its derivatives as input(s) and can be computed with open-source tools, are selected to predict flood susceptibility using ML. Among these descriptors *DTM* and *MRVBF* capture flatness and lowness at a location. *VDCN* captures flatness relative to stream network. $S_l$, $\kappa_h$, $\kappa_p$, $\kappa_t$, and *CI* capture localized flow behavior, whereas global flow behavior is captured by flow accumulation (local) ($A_l$), flow accumulation (stream) ($A_r$), flow depth (stream) ($h_r$), flow depth (local) ($h_l$), distance to the nearest stream (*D*), elevation difference to the nearest stream (*H*), vertical distance to the channel network (*VDCN*), topographic index (*TI*), and downslope index (*DI*). All GFDs of the framework are explained in detail in Table S-1.

As mentioned earlier, socio-economic vulnerability can be represented by an integrated resultant of socio-economic sensitivity and adaptive capacity. A relatively higher value of a socio-economic sensitivity indicator, for example population density of a sub-region, in a region indicates a comparatively higher socio-economic vulnerability in the sub-region along the dimension expressed by that indicator. Likewise, a relatively higher value of an adaptive capacity indicator, for example, literacy rate, results in a comparatively lower socio-economic vulnerability in the sub-region along the dimension it represents. All SEV indicators must be judiciously chosen per the socio-economic scenario of area under study. In the present study, relevant socio-economic sensitivity and adaptive capacity indicators were picked to estimate the socio-economic vulnerability of Jagatsinghpur district at village level using the CCR-DEA model, an efficient non-parametric ranking technique. The following subsections describe the three components of the proposed framework.

### 3.1. Estimation of flood susceptibility using ML models

*DTM* is the primary input GFD for flood susceptibility assessment in the present geomorphic framework. The remaining nineteen GFDs are computed in GIS environment, using an open-source software package, QGIS (QGIS.org, 2021). The framework employs tree-based ML models to combine the grids of GFDs into flood susceptibility grids. These models can capture non-linear and complex relationships between input features and target labels, making them ideal candidates for mapping the relationship between GFDs and flooding behavior at a location. They have also been popularly adopted in flood susceptibility studies (Lee et al., 2017; Tehrany et al., 2013) because of their non-parametric nature (Song and Ying, 2015) and inherent ability to capture feature interaction (Oh, 2019). The framework poses flood susceptibility estimation as a binary classification problem: classification of a location as either flooded or non-flooded based on its GFD values. However, multicollinearity among GFDs can impact model performance and interpretability; therefore, multicollinear GFDs are discarded and the remaining GFDs, referred to as multicollinearity-free GFDs (MFGFDs), are used as input features. Flooding information available for a small set of locations of the area under study, along with the corresponding values of GFDs at these locations, is used to create a tabular dataset. The dataset is further split into training (70 %) and testing (30 %) datasets, ensuring class balance. Class balance means there is an equal number of flooded and non-flooded data points in each sub-dataset. Maintaining class balance is recommended, especially if area under the receiver operating characteristic curve (*AUC*) is used as a metric to evaluate model performance (Saito and Rehmsmeier, 2015). The ML models are trained on the training dataset, and their classification performance is evaluated on both datasets. Using tree-based ML models ensures complex, non-linear relationships of these features with flood susceptibility and interaction among these descriptors are learned without increasing data redundancy and computational burden of modeling.

Flooding information required to prepare datasets can be obtained from previous local studies, historical flood data and maps, and other reliable sources. In the present study, flooded and non-flooded locations were identified from the $w_d$ grid of the district prepared by Mohanty et al. (2020a,b) using an HHM approach. The details of the approach are discussed in Mohanty et al. (2020a,b) and Ghosh et al. (2021). It should be noted that, ideally, observed flood maps should be utilized for training ML models. However, due to unavailability of observed flood data for the concerned study area, flood information derived from a comprehensive HHM framework (Mohanty et al., 2020a) is utilized here. Furthermore, the proposed ML framework is designed specifically for preliminary flood risk assessment in data- and resource-scarce regions and not to replace the traditional HHM approaches that can compute dynamic flood attributes such as flood velocity and momentum even in near-real time. ML models are not expected to learn these dynamic attributes with GFDs which are static. Yet, they likely learn the relative static representation because GFDs are selected explicitly for expressing hydro-geomorphological properties that influence these dynamic attributes. The following sub-sections explain the flood susceptibility assessment part of the framework in detail.

#### 3.1.1. Removing multicollinearity

Multicollinearity occurs when high intercorrelations exist among more than two GFDs. Multicollinearity is expected among GFDs because they are generated from one input, DTM. It is required to be removed to improve model performance and interpretability (Mansfield and Helms, 1982). The framework uses the variance inflation factor (VIF) analysis to identify multicollinear GFDs owing to its prominent use in flood susceptibility studies (Khosravi et al., 2018). After performing VIF analysis, multicollinearity-free geomorphic flood descriptors (MFGFDs) are used as input features in the ML models. The procedure for performing VIF analysis is enumerated below.

1. An ordinary least square regression model is fit considering the *i*th GFD ($g_i$) as a dependent variable and the rest of GFDs as explanatory variables.
2. The procedure is repeated for all the descriptors. Each regression model gives the coefficient of determination ($R_i^2$). The VIF score of the $i^{th}$ flood descriptor ($VIF_i$) is then calculated using the equation below:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} \tag{1}$$

3. Finally, only the descriptors having lower than a preset cut-off of VIF score (Kutner et al., 2005) are considered free from multicollinearity. The cut-off value is set to five in the framework.

#### 3.1.2. Tree-based models

In the present study, three tree-based ML models: (a) decision tree (DT), (b) random forest (RF), and (c) gradient-boosted decision trees (GBDT) were assessed for their ability to combine the grids of MFGFDs into the grid of flood susceptibility values. These models can learn non-linear and complex relationships between MFGFDs and flooding status. If flooding information at *M* different locations of the area under study is available, a tabular dataset is prepared using the values of MFGFDs ($G_{mf}$ in number) ($\boldsymbol{X} = \boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_i}, \ldots \boldsymbol{x_M}$ and $\boldsymbol{x_i} \epsilon R^{G_{mf}}$ with their corresponding flooding status values $\boldsymbol{Y} = y_1, y_2, \ldots y_i, \ldots, y_M$ and $y_i = \{0, 1\}$ where 0 denotes non-flooded location and 1 denotes flooded location. The dataset is further split into training and testing datasets. The objective of an ML model is to estimate a classification function $F(\boldsymbol{x})$ capable of producing *y* as accurately as possible with the training dataset. The three models are briefly explained in the subsequent paragraphs. It should be stressed that a location here refers to a square region represented by the resolution at which GFDs are used.

A tree is an undirected graph with no cycles, and a node in a tree is a decision point. A binary tree is a special tree having at most two decision outcomes at any given decision node. Decision-tree learning is a set of non-parametric supervised ML algorithms that are used for both classification and regression. It is based on the concept of entropy in information

theory (Shannon, 1948). The entropy of a dataset ($B$) that has $c$ classes/categories/types of instances is given by Eq. (2).

$$E(B) = \sum_{i=1}^{c} - p_i \, log_2 p_i \tag{2}$$

where $p_i$ is the proportion of instances of the $i$th class in $B$.

The classification and regression tree (CART) (Breiman et al., 1984) is a widely used decision-tree algorithm that has been previously adopted for flood susceptibility mapping (Rahmati and Pourghasemi, 2017). It can model rule-based relationships between physical attributes and flooding status at a location in the form of a binary tree without requiring any strict assumptions about the data distribution. Advantageously, it also inherently considers feature interaction. If a GFD ($g$) is split at a threshold ($g_\tau$) on a node of the tree, it divides the dataset ($B$) having $m$ data points at that node into $B_{left}$ and $B_{right}$ sub-datasets each having $m_{left}$ and $m_{right}$ data points, respectively. The information gain ($I$), reduction in the overall entropy, from the split is given by the following equation.

$$I(B, g_\tau) = E(B) - \frac{m_{left}}{m} . E\left(B_{left}\right) - \frac{m_{right}}{m} . E\left(B_{right}\right) \tag{3}$$

The trained binary tree comprises a root node, internal nodes, and terminal nodes. Each non-terminal node of the tree makes an optimal binary decision ($g_\tau^*$) that separates the dataset available at that node to maximize information gain (see Eq. (3)) rendering the subsequent sub-datasets at the daughter nodes as homogenous as possible. Each internal node follows the same rule. However, no further splitting is done at the terminal nodes. These nodes either have a predefined entropy level or can be finalized from pruning to reduce overfitting, which occurs when the model performance is worse on the testing dataset than on the training dataset. The CART algorithm was used for decision tree learning for the present study, and the trained decision tree (DT) was pruned with the minimum cost-complexity pruning algorithm to reduce overfitting (Breiman et al., 1984).

An ensemble can improve the classification performance of a single decision tree without losing its benefits. Bootstrapped aggregation (bagging) and boosted tree ensembles have been used in flood susceptibility mapping (Chen et al., 2020; Lee et al., 2017). Random forest (RF), introduced by Breiman (2001), is an ensemble of decision trees trained on bootstrapped training data samples. In the proposed framework, each tree is trained on a bootstrapped sample of data (by replacement) with a random subset of MFGFDs. The flooding probability given by each trained tree is considered to estimate the flooding probability, which is interpreted as flood susceptibility value ($FS_l$), at a location $l$ using its values of MFGFDs ($\boldsymbol{x_l}$). If $T$ such trees are trained on the dataset, the flooding probability ($FS_l$) at $l$ is given by the following expression:

$$FS_l = \frac{\sum_{t=1}^{T} FS_{lt}}{T} \tag{4}$$

where $FS_l$ is the flood susceptibility value at location $l$, and $FS_{lt}$ is the probability given by $t \in T$ tree that $l$ belongs to the flooded class.

Unlike bagging in RF, boosting attempts to improve the accuracy of a weak learner by repeatedly training the weak learner on various distributions of the training data and combining the trained weak learners into a robust composite learner (Schapire, 1999). A shallow decision tree, a decision tree with less depth (usually 10–12), is a weak leaner. It performs poorly on both training and testing datasets. Shallow decision trees can be boosted into gradient-boosted decision trees (GBDT), which is a highly robust and interpretable model for regression and classification (Friedman, 2001). It is an additive boosted model that optimizes an arbitrary differentiable loss function and uses shallow decision trees. It builds one decision tree at a time to fit the residual errors of the trees before it and has been extensively used for its high prediction power and computational performance. A comprehensive mathematical description of GBDT can be found in Si et al. (2017). In the present study, all the three tree-based models were based on CART, and their classification performance was evaluated using the metrics described in the following sub-section.

### 3.1.3. Evaluation of classification performance and feature importance

As per the framework, a location can belong to either flooded or non-flooded class which are denoted as 1 and 0 respectively. The ML models learn to predict the probability that a location belongs to flooded class using the values of MFGFDs ($\boldsymbol{x_l}$) at the location. This probability value is interpreted as the flood susceptibility value at the location. In the context of the proposed framework, a threshold ($\tau$) converts the probabilistic output (flood susceptibility value) of an ML model to its deterministic output (a given location belongs to flooded class if flood susceptibility value is greater than or equal to $\tau$). A confusion matrix (Kohavi and Provost, 1998) is derived after applying the threshold (assumed 0.5) on flood susceptibility values. The matrix consists of four values: true positives ($TP$), false positives ($FP$), true negatives ($TN$), and false negatives ($FN$).

True positives ($TP$): It is the number of known flooded locations that are predicted to belong to flooded class.

False positives ($FP$): It is the number of known non-flooded locations that are predicted to belong to flooded class.

True negatives ($TN$): It is the number of known non-flooded locations that are predicted to belong non-flooded class.

False negatives ($FN$): It is the number of known flooded locations that are predicted to belong to non-flooded class.

The proposed framework uses a commonly adopted threshold-dependent metric, Cohen's kappa ($\kappa$) (Cohen, 1960) and a popular threshold-independent metric, area under receiver operating characteristic curve ($AUC$) (Fawcett, 2006) to evaluate the performance of ML models. The two metrics have been widely used in flood susceptibility studies (Chapi et al., 2017; Tehrany et al., 2015). $\kappa$ represents the agreement between the prediction made by a trained ML model and the ground reality and can be computed using the equations provided below:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

where $p_o$ is the relative observed agreement between the ML model and the ground reality, and $p_e$ is the hypothetical probability of such agreement by chance. They can be calculated using the equations below.

$$p_o = \frac{TP + TN}{TP + TN + FN + FP} \tag{6}$$

$$p_e = \frac{(TP + FN)(TP + FP) + (FP + TN)(FN + TN)}{(TP + TN + FN + FP)^2} \tag{7}$$

Another approach to evaluate binary classification performance is by using receiver operating characteristic (ROC) curves. Varying $\tau$ generates multiple confusion matrices, which can used to calculate true positive ($\boldsymbol{r_{tp}}$) and false positive ($\boldsymbol{r_{fp}}$) rates. These rates are required to plot ROC curves and to calculate $AUC$ scores. $r_{tp}$ and $r_{fp}$ for a particular threshold are calculated as follows:

$$r_{tp} = TP/\mathrm{P} \tag{8}$$

$$P = TP + FN \tag{9}$$

$$r_{fp} = FP/N \tag{10}$$

$$N = TN + FP \tag{11}$$

$AUC$: Area under ROC curve is the probability that a trained ML can correctly classify a location into flooded or non-flooded class with the values of MFGFDs at the location.

The ML model with the highest $\kappa$ and $AUC$ values on both training and testing datasets is used to generate a grid of flood susceptibility values of the study area by applying the trained model on the grids of MFGFDs.

The global contribution of each MFGFD in the performance of a trained ML model indicates its importance in the model (Williamson et al., 2021). The present framework defines the feature importance score (*FI*) for a given MFGFD used in a trained ML model as the percentage contribution of that feature to the performance of the ML model.

### 3.2. Estimation of socio-economic vulnerability using data envelopment analysis

Socio-economic vulnerability has two components: sensitivity and adaptive capacity. The former component is determined by the indicators that adversely affect the results of a catastrophic event, for example, the population density of a community. On the other hand, adaptive capacity indicators describe the ability of a community to prepare for and respond to the impacts of present and foreseeable hazards (Smit and Wandel, 2006). As mentioned earlier, the inclusion of socio-economic vulnerability into flood risk estimation is equally essential. Therefore, the proposed framework aggregates relevant socio-economic sensitivity and adaptive capacity indicators of the study area to quantify its socioeconomic vulnerability, as per IPCC recommendation, using the Charles-Cooper-Rhodes (CCR) DEA model. Nevertheless, if a specific type of vulnerability such as infrastructural vulnerability is to be incorporated into risk, the framework is flexible enough to accommodate it. The preprocessing steps and the description of the model are provided in subsequent sub-sections.

#### 3.2.1. Standardization of SEV indicators

Socio-economic data is usually not available at the grid level where GFDs are computed. It is often available at the administrative level such as district, village, or block level. In the present framework, spatial unit (SU) term is used to refer to the level of administrative units. As per the framework, two sets of SEV indicators: $S$, the set of socio-economic sensitivity indicators and $S^c$, the set of adaptive capacity indicators, respectively, are standardized using the equation shown below:

$$a_{ju}^{std} = \begin{cases} \dfrac{a_{ju} - a_j^{min}}{a_j^{max} - a_j^{min}}, j \subseteq S \\ \dfrac{a_j^{max} - a_{ju}}{a_j^{max} - a_j^{min}}, j \subseteq S^c \end{cases} \tag{12a}$$

where $a_{ju}$ is the value of the $j$th indicator for the $u$th spatial unit (SU) and is converted into $a_{ju}^{std}$ which is the standardized value of the $j$th indicator for the $u$th SU. $a_j^{min}$ and $a_j^{max}$ are the minimum and the maximum values of the $j$th indicator across all SUs, respectively.

The standardization ensures that all indictors have a common scale and unit, are non-negative (Karmakar et al., 2010), and can be fed into an input-oriented DEA model (Sherly et al., 2015). The standardized values of SEV indicators for each SU are coalesced into its socio-economic vulnerability score using the CCR-DEA model, owing to its widespread use in vulnerability assessment, especially from natural hazards (Liu et al., 2018; Wei et al., 2004). The model is described in the subsequent sub-section.

Similarly, the standardization of GFDs at the grid level can be performed for their visual comparison using the equation below.

$$g_{fl}^{std} = \begin{cases} \dfrac{g_{fl} - g_f^{min}}{g_f^{max} - g_f^{min}}, f \subseteq G \\ \dfrac{g_f^{max} - g_{fl}}{g_f^{max} - g_f^{min}}, f \subseteq G^c \end{cases} \tag{12b}$$

where $g_{fl}$ is the value of the $f$th GFD at the $l$th location and is converted into $g_{fl}^{std}$ which is the standardized value of the $f$th GFD at the $l$th location. $g_f^{min}$ and $g_f^{max}$ are the minimum and the maximum values of the $f$th GFD across all locations, respectively. $G$ and $G^c$ are the sets of GFDs (see Table-S1) that are positively or negatively related to the flooding propensity at a location, respectively. It should be noted tree-based models, especially when used for classification, do not require standardization as a preprocessing step.

#### 3.2.2. The Charles-Cooper-Rhodes (CCR) data envelopment analysis (DEA) model

A unit, from a set of homogenous units, that produces $y$ outputs with $x$ inputs can be evaluated for its relative efficiency using data envelopment analysis (DEA), which is a non-parametric technique for evaluating the relative efficiency of homogenous units. DEA is often implemented when it is difficult to estimate absolute measures of efficiency. The proposed framework uses the slack-based input-oriented CCR-DEA model (Charnes et al., 1978). The CCR-DEA model considers constant returns to scale (CRS) meaning outputs linearly scale with inputs.

In the context of the framework, the standardized SEV indicators ($a^{std}$) are inputs to the DEA model. The $j$th SEV indicator value for the $u$th SU is denoted by $a_{ju}^{std}$ ($j = 1, 2, \ldots, x$). $Z$ is the total number of SUs. Because vulnerability loss data is seldom available, the output of the model is assumed to be unity for all SUs. The mathematical representation of the CCR-DEA model to evaluate the relative efficiency of the $u$th SU in the context of the proposed framework is described below.

$$min\left[\theta_u - \epsilon\left(\Sigma_{j=1}^{x} s_j^- + s^+\right)\right] \tag{13}$$

subject to

$$\sum_{z=1}^{Z} \lambda_z a_{jz}^{std} + s_j^- = \theta_u a_{ju}^{std} \; (j = 1, 2, \ldots, x) \forall u$$

$$\sum_{z=1}^{Z} \lambda_z - s^+ = 1 \forall u$$

$$s_j^- \geq 0 \; (j = 1, 2, \ldots, x)$$

$$s^+ \geq 0$$

$$\lambda_z \geq 0 \;\; (z = 1, 2, \ldots, Z)$$

where $\theta_u$ is the relative efficiency of the $u$th SU; $a_{jz}^{std}$ is the value of the $j$th standardized SEV indicator for the $z$th SU. $\epsilon > 0$ is a non-Archimedean value, and $s_j^-$ and $s^+$ are the slack and surplus variables, respectively. Solving the optimization problem, the optimum value of $\theta_u$ ($\theta_u^*$) for the $u$th SU is obtained and is called its CCR efficiency.

Prior to using the model, it should be ensured that the total number of SUs must be more than twice the product of the total number of inputs and outputs to prevent the curse of dimensionality (Dyson et al., 2001). Therefore, principal component analysis (PCA), a popular dimension reduction technique, may be required to decorrelate the standardized SEV indicators and to reduce the number of the indicators before feeding them into the DEA model if a significant linear correlation (Pearson's correlation coefficient $\geq 0.8$) exists among the standardized SEV indicators (Nataraja and Johnson, 2011; Sherly et al., 2015). Weakly correlated indicators may otherwise alter the dominant attributes of the principal components, corrupting the efficiency estimation (Yap et al., 2013). The CCR-DEA model provides the CCR efficiency ($\theta^*$) of each SU, which can be interpreted as the efficiency of an SU to contribute to SEV. $\theta_u^*$ of $u$th SU is converted into its socio-economic vulnerability score ($SEV_u$) using the following equation.

$$SEV_u = \left(1 - \theta_u^*\right) \tag{14}$$

It should be noted that the output to the CCR-DEA model is kept uniform (set to one) across all spatial units (SUs) because vulnerability loss data is seldom available. Still, vulnerability loss metrics such as annual fatalities attributed to floods if available can be added as outputs to the DEA model, which will improve vulnerability ranking of SUs.

### 3.3. Estimation of flood risk using flood susceptibility and SEV scores

In the proposed framework, flood risk is defined as the product of flood susceptibility and socio-economic vulnerability. A similar approach to combine susceptibility and vulnerability has recently been adopted by

Ghosh et al. (2019) and Mishra and Sinha (2020) and is consistent with the IPCC (2021) definition of risk. Since socio-economic vulnerability scores and flood susceptibility values are calculated at two different spatial levels (the former at the SU level and the latter at the grid level), flood susceptibility values are resampled from the grid level to the SU level to keep spatial consistency, using the equation below.

$$\overline{FS_u} = \frac{\Sigma_{l=1}^{L} FS_l}{L} \tag{15a}$$

where $\overline{FS_u}$ is the flood susceptibility score of the $u$th SU, $L$ is the total number of locations in the $u$th spatial unit, $FS_l$ is the flood susceptibility value predicted by the best ML model at the $l$th location in the $u$th SU. If waterbody information of the study area is available, it can be included to enrich flood susceptibility scores using Eq. (15b).

$$\overline{FS_u} = \frac{\Sigma_{k=1}^{K} FS_k + W}{K + W} \tag{15b}$$

where $FS_k$ is the flood susceptibility value predicted by the best ML model at the $k$th non-waterbody location in the $u$th SU. $K$ and $W$ are the total number of non-waterbody and waterbody locations in the $u$th SU, respectively.

Thereafter, Eq. (16) is used to calculate the flood risk score of each SU by multiplying its flood susceptibility and socio-economic vulnerability scores. Furthermore, the flood risk score of each SU is converted into its corresponding flood risk index (FRI) value using min-max normalization (Eq. (17)).

$$FR_u = \overline{FS_u} \times SEV_u \tag{16}$$

$$FRI_u = \frac{FR_u - FR_{min}}{FR_{max} - FR_{min}} \tag{17}$$

where $FR_u$ and $FRI_u$ are the flood risk score and the flood risk index value at the $u$th location, respectively. $FR_{min}$ and $FR_{max}$ are the minimum and maximum flood risk scores out of all SUs, respectively. Additionally, five flood risk classes: "*very low*," "*low*," "*medium*," "*high*," and "*very high*" flood risk, can be delineated with flood risk scores, for linguistic representation and ease in communication with stakeholders. The framework employs the Jenks natural breaks algorithm (Jenks, 1967) for delineating these five classes. The algorithm is an iterative algorithm that minimizes the intra-cluster variance of each class and maximizes the overall inter-cluster variance for a predefined number of classes and has been extensively adopted in flood hazard, risk, and vulnerability mapping (Mishra and Sinha, 2020; Toosi et al., 2019).

## 4. Results and discussion

The present study, for the first time, employs a combination of ML and DEA (ML-DEA) to quantify flood risk over a geomorphologically complex region at the finest administrative level. The $w_d$ grid of the study area was available at 20 m × 20 m resolution; therefore, the DTM grid of the study area was resampled and the other GFD grids were computed at the same resolution using QGIS. Relevant socio-economic sensitivity and adaptive capacity indicators (see Table S-3) were judiciously chosen to capture the socio-economic characteristics of the district. The flood susceptibility and socio-economic vulnerability scores of the district were estimated as per the methodologies depicted in Fig. 2b and a, respectively. These scores were subsequently utilized to quantify *FRI* values and to delineate flood risk classes at the spatial unit level. In this study, spatial units were villages and Jagatsinghpur town of the district.

### 4.1. Data exploration, pre-processing, and model training

The GFDs discussed in Table S-1 were standardized as per Eq. (12b) as shown in Fig. S-1a and b to visually understand the influence of each GFD. It may be noted that a relatively higher standardized value indicates a relatively higher flood propensity. It is observed that $h_l$, $\kappa_h$, $\kappa_p$, $\kappa_t$, $A_l$, and $C_A$ vary little spatially. *DTM* and *MRVBF* have high values in the coastal region showing their ability to capture the propensity of coastal floods. Multicollinearity among GFDs is expected as mentioned in Section 3.1.1. Thus, VIF analysis was performed at the grid level to assess multicollinearity among GFDs. The VIF scores obtained are provided in Table 1 in decreasing order. Five GFDs: *LGFI*, *GFI*, $h_r$, $h_l$, and $A_r$, were found to be multicollinear, hence, discarded. The remaining 15 GFDs (MFGFDs) were used as input features in the ML models. Table 1 and Fig. S1-b show that *LGFI* and *GFI* are the top two multicollinear GFDs and highly correlated, respectively. The high correlation between these two GFDs in the present study is because the study area is a small convex-divergent catchment. Since both GFDs have the same denominator, $H$, in their formulations (See Table-S1), while their corresponding numerators vary a little in the study area; therefore, the denominator dominates their values. However, it should be noted that the influence of GFDs is expected to vary depending on the study area. To derive the supervisory data to train and test ML models, two thousand locations (1000 flooded) were randomly sampled from the portion of the district not dominated by coastal floods (Sub-extent-1) (See Fig. 1) to fix the flow direction (from upstream to coastline). Sub-extent-1 was demarcated using the $w_d$ grid (return period: 50 years) obtained from Mohanty et al. (2020a,b). The $w_d$ grid was considered as the ground truth for identifying flooded and non-flooded locations. A tabular dataset was prepared from the values of MFGFDs and the flooding status values (flooded or non-flooded) at these sampled locations. It was further divided into the training dataset (70 %: 1400 data points) and the testing dataset (30 %: 600 data points). The three tree-based models (DT, RF, and GBDT), briefly described in Section 3.1.2, were trained on the training dataset. The performance of these models was compared using $\kappa$ and *AUC* scores as explained in Section 3.1.3.

**Table 1**
Geomorphic flood descriptors with their corresponding variance inflation factor (VIF) scores in deceasing order.

| GFD | VIF |
|---|---|
| *LGFI* | 16.72 |
| *GFI* | 16.23 |
| $h_r$ | 13.58 |
| $h_l$ | 11.31 |
| $A_r$ | 9.48 |
| $C_A$ | 4.62 |
| *DTM* | 3.30 |
| *H* | 3.30 |
| $A_l$ | 3.01 |
| *MRVBF* | 2.95 |
| *TI* | 2.67 |
| $C_B$ | 2.26 |
| *VDNS* | 2.17 |
| $S_l$ | 2.09 |
| $\kappa_h$ | 1.99 |
| $\kappa_p$ | 1.79 |
| *CI* | 1.77 |
| *DI* | 1.68 |
| $\kappa_t$ | 1.49 |
| *D* | 1.37 |

#### 4.1.1. Selection of the best ML model

Table 2 enlists the values of performance metrics $\kappa$ and *AUC* for the three ML models on the training and testing datasets. Fig. 3 shows the

**Table 2**
Classification performance of DT, RF, and GBDT models on training and testing.

| Performance metric | $\kappa$ | | | *AUC* | | |
|---|---|---|---|---|---|---|
| ML models | DT | RF | GBDT | DT | RF | GBDT |
| Training | 0.59 | 0.69 | 0.62 | 0.75 | 0.88 | 0.83 |
| Testing | 0.60 | 0.68 | 0.67 | 0.75 | 0.88 | 0.85 |

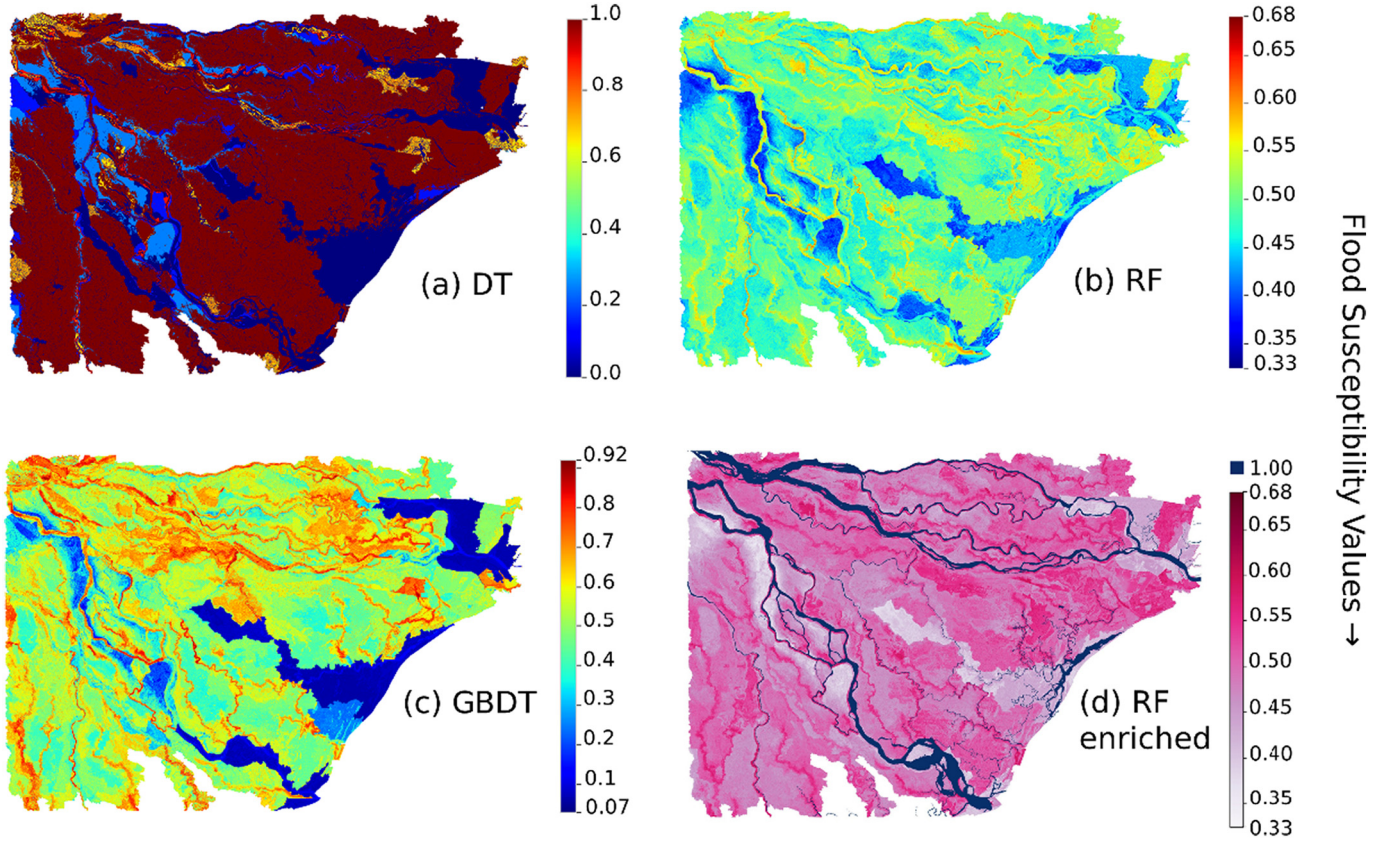



**Fig. 3.** Flood susceptibility grids produced using a) DT, b) RF, c) GBDT model, d) RF enriched with already available waterbodies information.

flood susceptibility grids developed using the three ML models. It may be observed that the metrics corresponding to the DT model are the lowest indicating its poor performance. The GBDT model provides a significant performance improvement. Nonetheless, the RF model is observed to perform better than the other two models on both datasets. Additionally, ROC curves were generated using the testing dataset with the trained models as shown in Fig. 4. It is evident that the DT model suffered from early retrieval issue and had the poorest *AUC* among the models, whereas the RF model provided the best performance. The flood susceptibility grid generated by the RF model was enriched with the information of the waterbodies of the district considering the presence of water in these waterbodies for most of the year as shown in Fig. 3(d). Thereafter, the enriched grid was resampled to the SU level using Eq. (15b). The resampled flood susceptibility values are referred to as flood susceptibility scores ($\overline{FS}$) and were used to generate the flood susceptibility map of the district shown in Fig. 5. Eragari, Nuapada, Itatikiri, Bachhalo, and Saharadia areas were among the most flood susceptible villages, whereas Biraballavpur, Kalioda, Nepur, Gopa, and Talia were the least flood susceptible villages in the district.

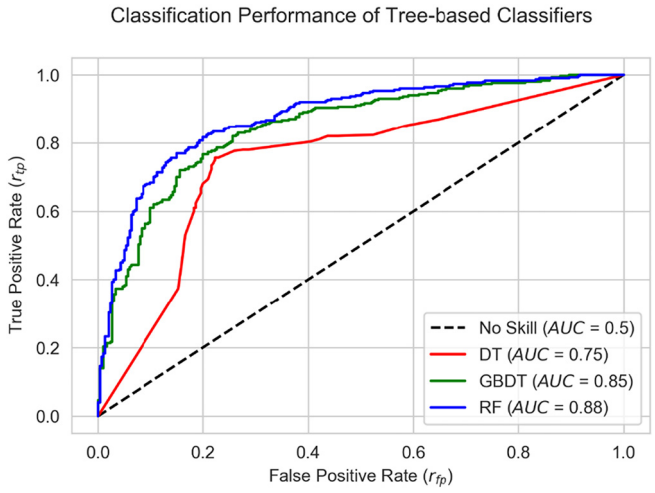



**Fig. 4.** The receiver operating characteristic (ROC) curves on the testing dataset for the no-skill classifier, DT, GBDT, and RF models.

Table 3 represents the feature importance scores of geomorphic flood descriptors used in the RF model, which shows that *DTM*, *MRVBF*, and *VDCN* have higher feature importance than the other GFDs. In fact, these three GFDs contribute >40 % to the performance of the RF model.

The framework uses tree-based models because they provide a good balance of their ability to learn complex relationships between input features and target labels and data requirement. However, interpreting them is not straightforward. Although the importance of each MFGFD can be computed at global level (see Table 3), the contribution of MFGFDs to flood susceptibility values may vary from location to location. Therefore, in the future, instance-based interpretability techniques can be added to the framework. These techniques allow for understanding how MFGFDs influence flood susceptibility values locally, hence, strengthening confidence in modeling. Additionally, they can be used by experts to evaluate ML models and to develop personalized flood measures for each SU. Using tree-based ML models ensures complex, non-linear relationships of these features with flood susceptibility and interaction among these descriptors are learned without increasing data redundancy and computational burden of modeling.

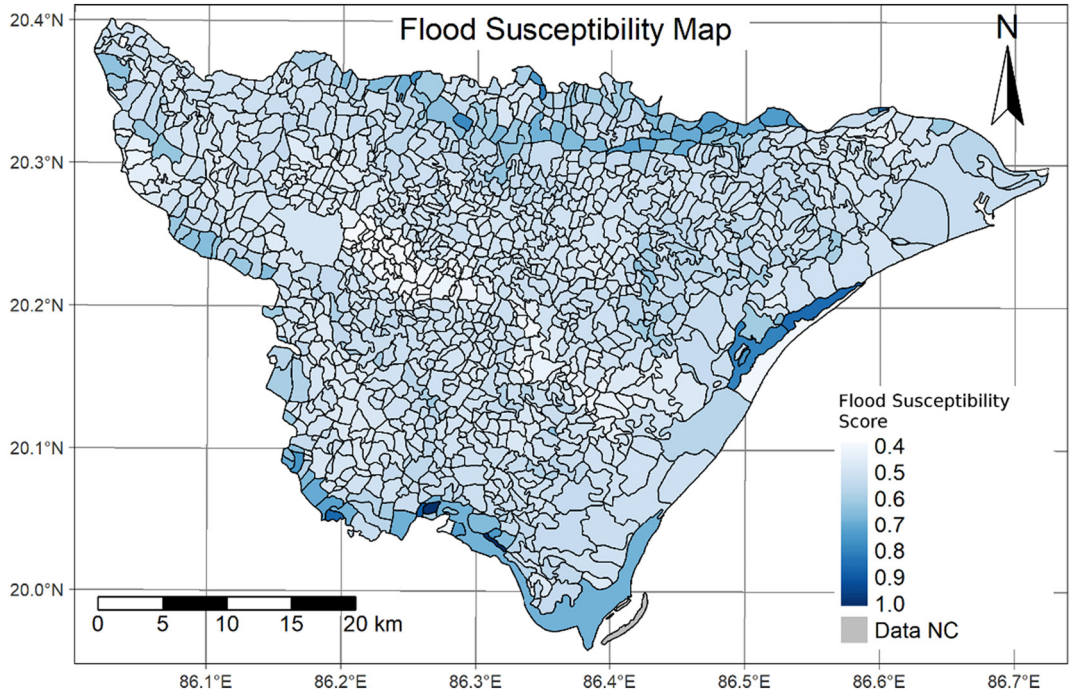



**Fig. 5.** The flood susceptibility map of Jagatsinghpur, Odisha, based on the flood susceptibility scores ($\overline{FS}$).

### 4.2. Socio-economic vulnerability and flood risk mapping

Pearson correlation coefficients between all pairs of SEV indicators were calculated as shown in Fig. S-2. High intercorrelations among these indicators were not found. Therefore, the CCR-DEA model was used to aggregate all available indicators after standardization without performing PCA. The CCR efficiency ($\theta^*$) of each SU was converted into its socio-economic vulnerability score using Eq. (14). Following this, the Jenks natural breaks algorithm was applied to demarcate five vulnerability classes. According to it, 29.13 %, 17.7 %, 19.5 %, 20.6 %, and 13 % of villages belong to "*very low*," "*low*," "*moderate*," "*high*," and "*very high*" vulnerability classes, respectively. Fig. 6 shows the socio-economic map of the district. Sanamandasahi and Gopalpur are found to be the most socio-economically vulnerable villages.

Next, the socio-economic vulnerability score (*SEV*) and the flood susceptibility score ($\overline{FS}$) of each SU were multiplied to obtain its flood risk score (*FR*) as per Eq. (16). Thereafter, the flood risk scores were converted into the flood risk index values using Eq. (17). The flood risk scores were also used to identify flood risk classes using the Jenks natural breaks algorithm. The maps of flood risk index and flood risk class of the district are shown in Figs. 7 and 8, respectively. Bilipada, Jayasankhapur, Saharadia, Jotta, and Paruna are the five most flood risk prone villages, whereas Keruapada, Chhelikulia, Chariakana, Nalanga, and Salio were the least prone to flood risk. As per the results, 16.68 % of villages are subject to "very low" flood risk, 15.47 % to "low," 24.5 % to "moderate," 27.6 % to "high," and 15.75 % are subject to "very high" flood risk. The extensive and detailed flood related information determined at the finest administrative scale over Jagatsinghpur district may be referred to by the Regional District Emergency Centre, Mahanadi River Basin Organization, policy makers, and town-planners to adopt appropriate land-use options and flood control measures for minimizing risk and increasing resilience of the affected communities.

**Table 3**
Feature importance scores (*FI*) of geomorphic flood descriptors (GFDs) used in the RF model.

| GFD | *FI* (%) | GFD | *FI* (%) | GFD | *FI* (%) |
|---|---|---|---|---|---|
| *DTM* | 15.9 | *D* | 6.5 | $\kappa_p$ | 4.3 |
| *MRVBF* | 15.1 | $C_B$ | 4.8 | *CI* | 4.2 |
| *VDNS* | 14.3 | *TI* | 4.6 | $\kappa_t$ | 3.9 |
| *H* | 6.7 | $S_l$ | 4.5 | $A_l$ | 2.2 |
| $C_A$ | 6.7 | $\kappa_h$ | 4.4 | *DI* | 2.1 |

The proposed framework uses DTM-derived flood descriptors at the same spatial level because it improves the reliability of the resultant flood susceptibility grid. All tree-based ML models and the CCR-DEA are non-parametric, which widens the framework's applicability. The free and near-global availability of good quality DTMs such as SRTM and ASTERDEM ensure the easy availability of reliable DTM. All flood descriptors have been preselected for the framework based on an in-depth literature review. Nevertheless, socio-economic sensitivity and adaptive capacity indicators must be judiciously chosen per the study area's socio-economic context and data availability. In the present study, the SEV indicators were selected to represent the agrarian economy of the district, the literacy and employment characteristics of its populace, and the presence of marginal population in the district. The probability density functions of each GFD conditioned on the flooding status were plotted as shown in Fig. S-3, to understand its influence on floods in the study area. In the figure, it may be observed that *DTM*, *MRVBF*, and *VDCN* separate flooded locations from non-flooded locations better than the other GFDs. Yet, none of the GFDs can be used as a sole descriptor for classification, which validates our choice of choosing many descriptors over one. Nevertheless, it should be noted that the influence of GFDs is expected to change per study area. It is easier to use all the GFDs in the ML models after accounting for multicollinearity than to identify a universal flood descriptor, as has been extensively researched in the past (Degiorgis et al., 2012; Manfreda et al., 2014). It is well-known that neighborhood characteristics at a location also affect flood susceptibility, which tree-based models cannot inherently capture, so in the future, computer vision-based techniques can be used to incorporate these characteristics into the framework. It is emphasized that the proposed framework is primarily designed to preliminarily assess flood risk in data- and resource-scarce regions. It cannot replace the traditional HHM approaches because GFDs cannot capture the

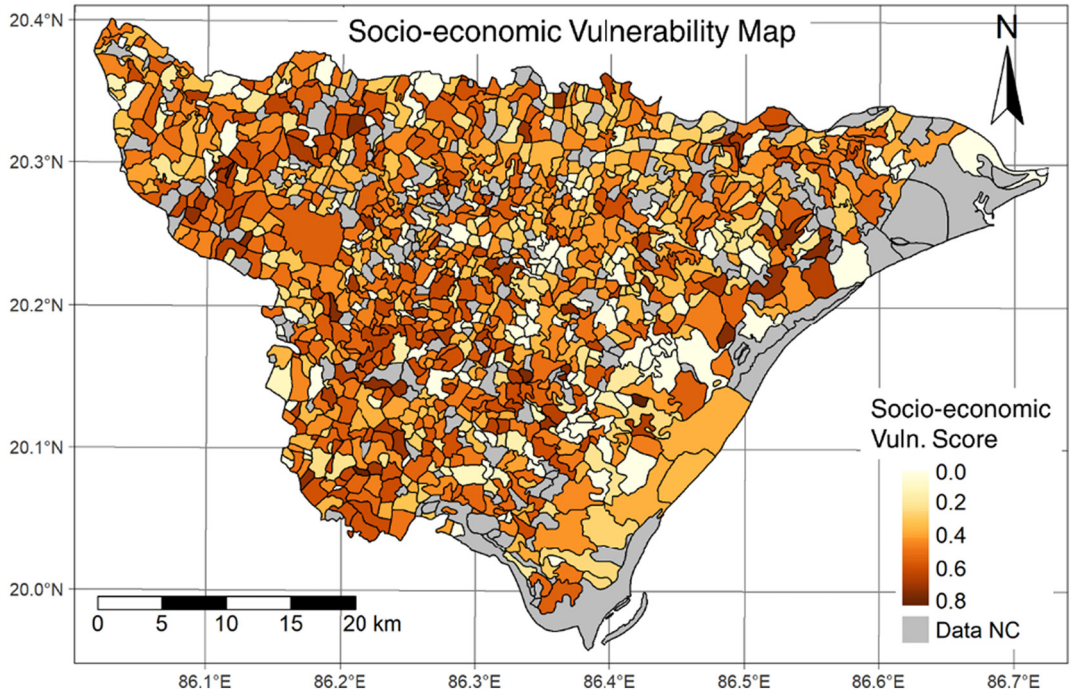



**Fig. 6.** Socio-economic vulnerability map produced by converting the technical efficiency scores into the socio-economic vulnerability scores.

dynamics of floodwater; however, the framework can be used to identify flood risk hotspots for further hydrodynamic modeling at a local scale, hence, significantly reducing computational cost.

## 5. Conclusions

The present study proposes a novel framework of flood risk mapping through a machine learning-based approach using DTM-based GFDs as well as socio-economic sensitivity and adaptive capacity indicators. The exhaustive framework is demonstrated over Jagatsinghpur district, part of the highly flood prone lower Mahanadi basin in India. Three extensively-used tree-based ML models are assessed for their ability to combine 15 multicollinearity-free DTM-derived GFDs into flood susceptibility value, which is a proxy of geomorphic flood hazard. It is observed that the RF model performs better compared to the other two models by providing the highest $AUC$ and $\kappa$ values ($AUC_{RF} = 0.88$; $\kappa_{RF} = 0.68$), and therefore selected for quantifying the flood susceptibility for the study area at the finest administrative scale, i.e., village level. On the other hand, socio-economic vulnerability is derived by considering relevant socio-economic sensitivity and adaptive capacity indicators using the CCR-DEA model

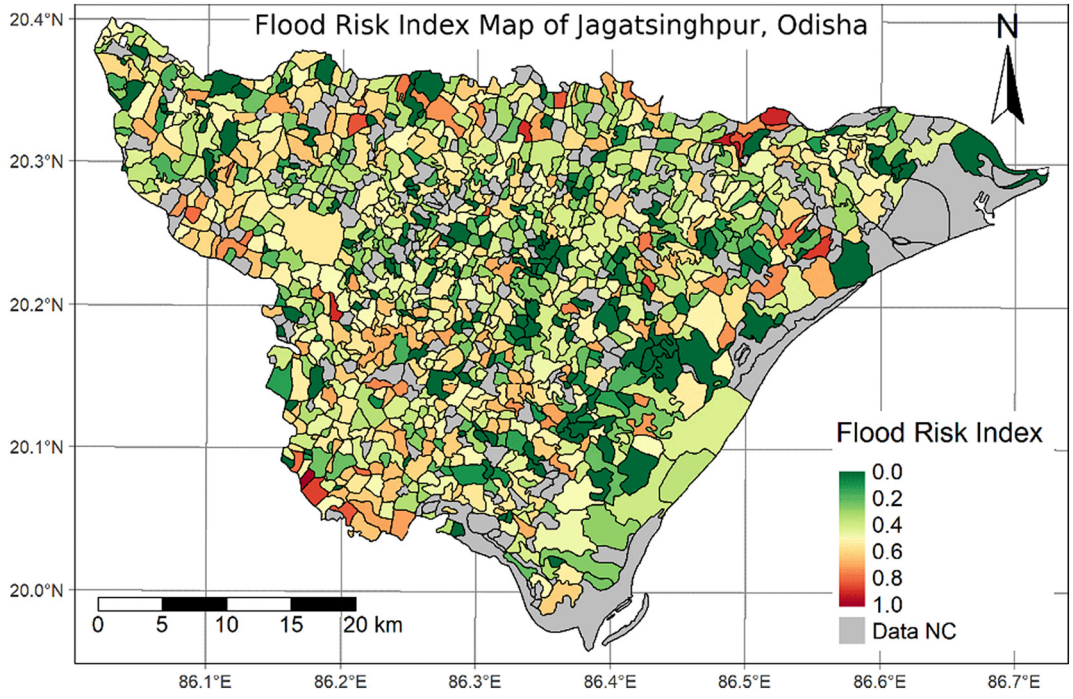



**Fig. 7.** Flood risk map based on flood risk index values calculated after normalizing flood risk scores.

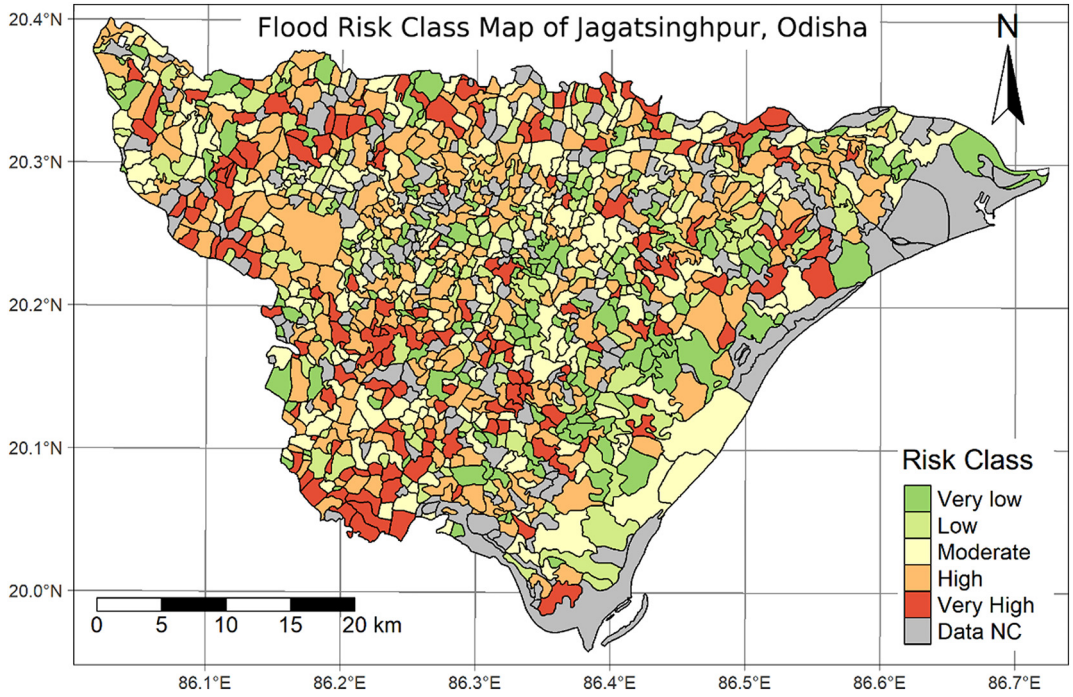



**Fig. 8.** Flood risk map based on flood risk classes demarcated by one-dimensional clustering performed with the Jenks natural breaks algorithm.

which eliminates subjectivity arising from weighting indicators by experts. Further, five flood risk classes are demarcated with the Jenks natural breaks algorithm on the flood risk scores obtained by multiplying the flood susceptibility and socio-economic vulnerability scores. It is observed that 16.68 % of villages are subject to "*very low*" flood risk, 15.47 % to "*low*," 24.5 % to "*moderate*," 27.6 % to "*high*," and 15.75 % are subject to "*very high*" flood risk. The proposed framework is computationally efficient as it primarily requires only DTM and a set of known flooded locations over a study area for estimating flood susceptibility. Therefore, it is an affordable alternative to computationally expensive HHM approaches and can be used when intricacies related to flood dynamics are not required. The proposed framework, owing to the consideration of a wide array of geomorphological features and socio-economic characteristics of disaster-facing population, promises its wide applicability, especially over data- and resource-constrained large flood-prone regions.

## CRediT authorship contribution statement

**Prakhar Deroliya**: Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization.

**Mousumi Ghosh**: Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization.

**M.P. Mohanty**: Formal analysis, Investigation, Writing - review & editing.

**S. Ghosh**: Resources, Writing - review & editing, Funding acquisition, Supervision.

**K.H.V. Durga Rao**: Resources, Funding acquisition, Writing - review & editing.

**S. Karmakar**: Conceptualization, Methodology, Resources, Data curation, Visualization, Investigation, Writing - review & editing, Funding acquisition, Project administration, Supervision.

## Data availability

Data will be made available on request.

## Declaration of competing interest

None.

## Acknowledgements

The work presented here is supported by ISRO-IIT(B)-Space Technology Cell (STC) through sponsored project (RD/0119-ISROC00-001). The authors also thank National Remote Sensing Centre (NRSC), Hyderabad; Department of Water Resources (DoWR), Govt. of Odisha; and Odisha Space Applications Centre (ORSAC), Odisha, for providing relevant data for carrying out the research. The authors are grateful to NRSC for allowing the access to CartoDEM. The support toward computational resources has been provided by IIT Bombay. The authors also express their gratitude to Professor Fernando A.L. Pacheco (Associate Editor) and the three anonymous Reviewers for their constructive suggestions, which helped to improve the overall quality of the manuscript.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.158002.

## References

Ardiclioglu, M., Hadi, A.M., Periku, E., Kuriqi, A., 2022. Experimental and numerical investigation of bridge configuration effect on hydraulic regime. Int. J. Civ. Eng. 1–11. https://doi.org/10.1007/s40999-022-00715-2.

Avand, M., Kuriqi, A., Khazaei, M., Ghorbanzadeh, O., 2022. DEM resolution effects on machine learning performance for flood probability mapping. J. Hydro Environ. Res. 40, 1–16. https://doi.org/10.1016/j.jher.2021.10.002.

Ayalew, L., Yamagishi, H., 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology 65 (1–2), 15–31. https://doi.org/10.1016/j.geomorph.2004.06.010.

Baliarsingh, A., Panigrahi, G., Kamar, B., Pattanaik, P., Mohapatra, A.K.B., 2018. Risk proof crop planning based on rainfall probability in Jagatsinghpur district of Odisha. J. Pharm. Innov. 7 (8), 182–186.

Barredo, J.I., De Roo, A., Lavalle, C., 2007. Flood risk mapping at European scale. Water Sci. Technol. 56 (4), 11–17. http://natural-hazards.jrc.it/downloads/pdf/ec_jrc_riskmapping.pdf.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. J. 24 (1), 43–69. https://doi.org/10.1080/02626667909491834.
Bradshaw, C.J., Sodhi, N.S., PEH, K.S.H., Brook, B.W., 2007. Global evidence that deforestation amplifies flood risk and severity in the developing world. Global Change Biology 13 (11), 2379–2395. https://doi.org/10.1111/j.1365-2486.2007.01446.x.
Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.I., 1984. Classification and Regression Trees. Wadsworth, Belmont, Calif https://doi.org/10.1201/9781315139470.
Burby, R.J., 2001. Flood insurance and floodplain management: the US experience. Global Environ. Change B. Environ. Hazard 3 (3), 111–122. https://doi.org/10.3763/ehaz.2001.0310.
Census of India, 2011. Provisional Population Totals. Census of India. Office of the Registrar General and Census Commissioner, India.
Central Water Commission, 2019. Water and Related Statistics. Water Resources Information Systems Directorate, New Delhi.
Chapi, K., Singh, V.P., Shirzadi, A., Shahabi, H., Bui, D.T., Pham, B.T, Khosravi, K., 2017. A novel hybrid artificial intelligence approach for flood susceptibility assessment. Environ. Model. Softw. 95, 229–245. https://doi.org/10.1016/j.envsoft.2017.06.012.
Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring the efficiency of decision-making units. Eur. J. Oper. Res. 2 (6), 429–444. https://doi.org/10.1016/0377-2217(78)90138-8.
Chau, T.K., Thanh, N.T., 2021. Primarily results of a real-time flash flood warning system in Vietnam. CivilEngineering Journal 7 (4), 747–762. https://doi.org/10.28991/cej-2021-03091687.
Chen, W., Li, Y., Xue, W., Shahabi, H., Li, S., Hong, H., Ahmad, B.B., 2020. Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. Sci. Total Environ. 701, 134979. https://doi.org/10.1016/j.scitotenv.2019.134979.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46. https://doi.org/10.1177/2F001316446002000104.
Degiorgis, M., Gnecco, G., Gorni, S., Roth, G., Sanguineti, M., Taramasso, A.C., 2012. Classifiers for the detection of flood-prone areas using remote sensed elevation data. J. Hydrol. 470, 302–315. https://doi.org/10.1016/j.jhydrol.2012.09.006.
Díez-Herrero et al., n.d.A. Díez-Herrero L. Lain-Huerta M. Llorente-Isidro n.d.A Handbook on Flood Hazard Mapping Methodologies. Publications of the Geological Survey of Spain, Series Geological Hazards/Geotechnics No. 2, Madrid, Spain, 190 pp.
Dilley, M., Chen, R.S., Deichmann, U., Lerner-Lam, A.L., Arnold, M., 2005. Natural Disaster Hotspots: A Global Risk Analysis. The World Bank, Washington DC, US.
DPMU, 2017. Comprehensive District Plan for 2017-18. District Planning and Monitoring Unit, Jagatsinghpur, Odisha, India.
Dyson, R.G., Allen, R., Camanho, A.S., Podinovski, V.V., Sarrico, C.S., Shale, E.A., 2001. Pitfalls and protocols in DEA. Eur. J. Oper. Res. 132 (2), 245–259. https://doi.org/10.1016/S0377-2217(00)00149-1.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232. https://www.jstor.org/stable/2699986.
Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39 (12). https://doi.org/10.1029/2002WR001426.
Ghosh, A., Das, S., Ghosh, T., Hazra, S., 2019. Risk of extreme events in delta environment: a case study of the Mahanadi delta. Sci. Total Environ. 664, 713–723. https://doi.org/10.1016/j.scitotenv.2019.01.390.
Ghosh, M., Mohanty, M.P., Kishore, P., Karmakar, S., 2021. Performance evaluation of potential inland flood management options through a three-way linked hydrodynamic modelling framework for a coastal urban watershed. Hydrol. Res. 52 (1), 61–77. https://doi.org/10.2166/nh.2020.123.
Gorokhovich, Y., Voustianiouk, A., 2006. Accuracy assessment of the processed SRTM-based elevation data by CGIAR using field data from USA and Thailand and its relation to the terrain characteristics. Remote Sens. Environ. 104 (4), 409–415. https://doi.org/10.1016/j.rse.2006.05.012.
Gusain, A., Mohanty, M.P., Ghosh, S., Chatterjee, C., Karmakar, S., 2020. Capturing transformation of flood hazard over a large River Basin under changing climate using a top-down approach. Sci. Total Environ. 726, 138600. https://doi.org/10.1016/j.scitotenv.2020.138600.
Hack, J.T., Goodlett, J.C., 1960. Geomorphology and forest ecology of a mountain region in the central Appalachians. United States Geological Survey Professional Paper No. 347. United States Government Printing Office, Washington DC, US https://doi.org/10.3133/pp347.
Hjerdt, K.N., McDonnell, J.J., Seibert, J., Rodhe, A., 2004. A new topographic index to quantify downslope controls on local drainage. Water Resour. Res. 40 (5), 1–6. https://doi.org/10.1029/2004WR003130.
IFRC, 2020. World Disasters Report 2020: Come Heat or High Water. International Federation of Red Cross and Red Crescent Societies, Geneva, Switzerland.
IPCC, 2013. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. https://doi.org/10.1017/CBO9781107415324 1535 pp.
IPCC, 2021. In: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J.B.R., Maycock, T.K., Waterfield, T., Yelekçi, O., Yu, R., Zhou, B. (Eds.), Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press In Press.
Janizadeh, S., Avand, M., Jaafari, A., Phong, T.V., Bayat, M., Ahmadisharaf, E., Lee, S., 2019. Prediction success of machine learning methods for flash flood susceptibility mapping in the Tafresh watershed, Iran. Sustainability 11 (19), 5426. https://doi.org/10.3390/su11195426.
Jenks, G.F., 1967. The data model concept in statistical mapping. International Yearbook of Cartography 7, 186–190. https://doi.org/10.3130/aije.68.71_3.
Kapetas, L., Kazakis, N., Voudouris, K., McNicholl, D., 2019. Water allocation and governance in multi-stakeholder environments: insight from Axios DeltaGreece. Science of The Total Environment 695, 133831. https://doi.org/10.1016/j.scitotenv.2019.133831.
Karmakar, S., Simonovic, S.P., Peck, A., Black, J., 2010. An information system for risk-vulnerability assessment to flood. J. Geogr. Inf. Syst. 2 (3), 129–146. https://doi.org/10.4236/jgis.2010.23020.
Kazakis, N., Oikonomidis, D., Voudouris, K.S., 2015. Groundwater vulnerability and pollution risk assessment with disparate models in karstic, porous, and fissured rock aquifers using remote sensing techniques and GIS in anthemountas basinGreece. Environmental earth sciences 74 (7), 6199–6209. https://doi.org/10.1007/s12665-015-4641-y.
Khosravi, K., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Bui, D.T., 2018. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci. Total Environ. 627, 744–755. https://doi.org/10.1016/j.scitotenv.2018.01.266.
Kohavi, R., Provost, F., 1998. Confusion matrix. Machine learning 30 (2–3), 271–274. https://doi.org/10.1023/A:1017181826899.
Kotsiantis, S.B., Zaharakis, I., Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. EmergingArtificial Intelligence Applications in Computer Engineering. 160, pp. 3–24. https://doi.org/10.5555/1566770.1566773 1.
Kron, W., 2005. Flood risk= hazard• values• vulnerability. Water Int. 30 (1), 58–68.
Kumar, M., Kumar, P., Kumar, A., Elbeltagi, A., Kuriqi, A., 2022. Modeling stage–discharge–sediment using support vector machine and artificial neural network coupled with wavelet transform. Appl. Water Sci. 12 (5), 1–21. https://doi.org/10.1007/s13201-022-01621-7.
Kundzewicz, Z.W., 2002. Non-structural flood protection and sustainability. Water Int. 27 (1), 3–13. https://doi.org/10.1080/02508060208686972.
Kuriqi, A., Hysa, A., 2021. Multidimensional aspects of floods: nature-based mitigation measures from basin to river reach scale. In: Ferreira, C.S.S., Kalantari, Z., Hartmann, T., Pereira, P. (Eds.), Nature-based Solutions for Flood Mitigation. The Handbook of Environmental Chemistry. 107. Springer, Cham. https://doi.org/10.1007/698_2021_773.
Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W., 2005. Applied Linear Statistical ModelsVol. 5. McGraw-Hill Irwin, New York, US.
Lee, S., Kim, J.C., Jung, H.S., Lee, M.J., Lee, S., 2017. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. Geomat. Nat. Haz. Risk 8 (2), 1185–1203. https://doi.org/10.1080/19475705.2017.1308971.
Liu, Y., Wei, J., Xu, J., Ouyang, Z., 2018. Evaluation of the moderate earthquake resilience of counties in China based on a three-stage DEA model. Nat. Hazards 91 (2), 587–609. https://doi.org/10.1007/s11069-017-3142-6.
Lowrance, W.W., 1976. Of AcceptableRisk: Science and the Determination of Safety. William Kaufmann, Inc, Menlo Park, California, US. http://hdl.handle.net/10822/772475.
MacMillan, R.A., Jones, R.K., McNabb, D.H., 2004. Defining a hierarchy of spatial entities for environmental analysis and modeling using digital elevation models (DEMs). Comput. Environ. Urban. Syst. 28 (3), 175–200. https://doi.org/10.1016/S0198-9715(03)00019-X.
Manfreda, S., Samela, C., 2019. A digital elevation model-based method for a rapid estimation of flood inundation depth. J. Flood Risk Manage. 12, e12541. https://doi.org/10.1111/jfr3.12541.
Manfreda, S., Di Leo, M., Sole, A., 2011. Detection of flood-prone areas using digital elevation models. J. Hydrol. Eng. 16 (10), 781–790. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000367.
Manfreda, S., Nardi, F., Samela, C., Grimaldi, S., Taramasso, A.C., Roth, G., Sole, A., 2014. Investigation on the use of geomorphic approaches for the delineation of flood prone areas. J. Hydrol. 517, 863–876. https://doi.org/10.1016/j.jhydrol.2014.06.009.
Mansfield, Edward R., Helms, Billy P., 1982. Detecting multicollinearity. Am. Stat. 36 (3a), 158–160. https://doi.org/10.1080/00031305.1982.10482818.
Mehryar, S., Surminski, S., 2022. Investigating flood resilience perceptions and supporting collective decision-making through fuzzy cognitive mapping. Science of The Total Environment. https://doi.org/10.1016/j.scitotenv.2022.155854 in press.
Mishra, S.K., Mishra, N., 2010. Vulnerability and adaptation analysis in flood affected areas of Orissa. Social Change 40 (2), 175–193. https://doi.org/10.1177/2F004908571004000205.
Mishra, K., Sinha, R., 2020. Flood risk assessment in the Kosi megafan using multi-criteria decision analysis: a hydro-geomorphic approach. Geomorphology 350, 106861. https://doi.org/10.1016/j.geomorph.2019.106861.
Mohanty, M.P., Vittal, H., Yadav, V., Ghosh, S., Rao, G.S., Karmakar, S., 2020a. A new bivariate risk classifier for flood management considering hazard and socio-economic dimensions. J. Environ. Manag. 255, 109733. https://doi.org/10.1016/j.jenvman.2019.109733.
Mohanty, M.P., Indu, J., Ghosh, S., Rao, G., Nithya, S., Karmakar, S., 2020b. Sensitivity of various topographic data in flood management: Implications on inundation mapping over large data-scarce regions. J. Hydrology 590, 125523. https://doi.org/10.1016/j.jhydrol.2020.125523.
Mohanty, M.P., Karmakar, S., 2021. WebFRIS: an efficient web-based decision support tool to disseminate end-to-end risk information for flood management. J. Environ. Manag. 288, 112456. https://doi.org/10.1016/j.jenvman.2021.112456.
Mohapatra, P.K., Singh, R.D., 2003. Flood management in India. Flood Problem and Management in South Asia. Springer, Dordrecht, pp. 131–143 https://doi.org/10.1007/978-94-017-0137-2_6.
Muralikrishnan, S., Pillai, A., Narender, B., Reddy, S., Venkataraman, V.R., Dadhwal, V.K., 2013. Validation of Indian national DEM from Cartosat-1 data. J. Indian Soc. Remote Sens. 41 (1), 1–13. https://doi.org/10.1007/s12524-012-0212-9.
Nardi, F., Vivoni, E.R., Grimaldi, S., 2006. Investigating a floodplain scaling relation using a hydrogeomorphic delineation method. Water Resour. Res. 42 (9), W09409. https://doi.org/10.1029/2005WR004155.
Nataraja, N.R., Johnson, A.L., 2011. Guidelines for using variable selection techniques in data envelopment analysis. Eur. J. Oper. Res. 215 (3), 662–669. https://doi.org/10.1016/j.ejor.2011.06.045.

Nilawar, A.P., Waikar, M.L., 2019. Impacts of climate change on streamflow and sediment concentration under RCP 4.5 and 8.5: a case study in Purna river basin India. Science of the total environment 650, 2685–2696. https://doi.org/10.1016/j.scitotenv.2018.09.334.

Nobre, A.D., Cuartas, L.A., Hodnett, M., Rennó, C.D., Rodrigues, G., Silveira, A., Saleska, S., 2011. Height above the nearest drainage–a hydrologically relevant new terrain model. J. Hydrol. 404 (1–2), 13–29. https://doi.org/10.1016/j.jhydrol.2011.03.051.

Noman, N.S., Nelson, E.J., Zundel, A.K., 2001. Review of automated floodplain delineation from digital terrain models. J. Water Resour. Plan. Manag. 127 (6), 394–402. https://doi.org/10.1061/(ASCE)0733-9496(2001)127:6(394).

Odeha, I.O.A., McBratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63 (3–4), 197–214. https://doi.org/10.1016/0016-7061(94)90063-9.

Oh, S., 2019. Feature interaction in terms of prediction performance. Appl. Sci. 9 (23), 5191. https://doi.org/10.3390/app9235191.

Papaioannou, G., Vasiliades, L., Loukas, A., 2015. Multi-criteria analysis framework for potential flood prone areas mapping. Water Resour. Manag. 29 (2), 399–418. https://doi.org/10.1007/s11269-014-0817-6.

Pourghasemi, H.R., Kariminejad, N., Amiri, M., Edalat, M., Zarafshar, M., Blaschke, T., Cerda, A., 2020. Assessing and mapping multi-hazard risk susceptibility using a machine learning technique. Sci. Rep. 10 (1), 1–11. https://doi.org/10.1038/s41598-020-60191-3.

Pradhan, B., 2010. Flood susceptible mapping and risk area delineation using logistic regression, GIS and remote sensing. J. Spat. Hydrol. 9 (2). https://doi.org/10.1007/s12665-009-0245-8.

Prinos, P., Kortenhaus, A., Swerpel, B., Jiménez, J.A., Samuels, P., 2008. Review of flood hazard mapping. Integrated Flood Risk Analysis and Management Methodologies. FLOODsite, HR Wallingford, The United Kingdom in press.

QGIS.org, 2021. QGIS Geographic Information System. QGIS Association. http://www.qgis.org.

Rahman, M., Ningsheng, C., Islam, M.M., Dewan, A., Iqbal, J., Washakh, R.M.A., Shufeng, T., 2019. Flood susceptibility assessment in Bangladesh using machine learning and multi-criteria decision analysis. Earth Syst. Environ. 3 (3), 585–601. https://doi.org/10.1007/s41748-019-00123-y.

Rahmati, O., Pourghasemi, H.R., 2017. Identification of critical flood prone areas in data-scarce and ungauged regions: a comparison of three data mining models. Water Resour. Manag. 31 (5), 1473–1487. https://doi.org/10.1007/s11269-017-1589-6.

Rentschler, J., Salhab, M., 2020, Octoberr. Poverty and Shared Prosperity 2020 (No. 9447). World Bank Group https://doi.org/10.1596/1813-9450-9447. https://documents1.worldbank.org/curated/en/669141603288540994/pdf/People-in-Harms-Way-Flood-Exposure-and-Poverty-in-189-Countries.pdf.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one 10 (3), e0118432. https://doi.org/10.1371/journal.pone.0118432.

Samela, C., Albano, R., Sole, A., Manfreda, S., 2018. A GIS tool for cost-effective delineation of flood-prone areas. Comput. Environ. Urban. Syst. 70, 43–52. https://doi.org/10.1016/j.compenvurbsys.2018.01.013.

Sangwan, N., Merwade, V., 2015. A faster and economical approach to floodplain mapping using soil information. J. Am. Water Resour. Assoc. 51 (5), 1286–1304. https://doi.org/10.1111/1752-1688.12306.

Schapire, R.E., 1999, Julyy. A brief introduction to boosting. Ijcai 99, 1401–1406 doi: 10.1.1.640.9942.

Schiermeier, Q., 2011. Increased flood risk linked to global warming. Nature 470 (7334), 316. https://doi.org/10.1038/470316a.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423. https://doi.org/10.1145/584091.584093.

Shareef, M.E., Abdulrazzaq, D.G., 2021. River flood modelling for flooding risk mitigation in Iraq. Civil Eng. J. 7 (10), 1702–1715. 10.28991/cej-2021-03091754.

Sharma, T., Vittal, H., Karmakar, S., Ghosh, S., 2020. Increasing agricultural risk to hydro-climatic extremes in India. Environ. Res. Lett. 15 (3), 034010. https://doi.org/10.1088/1748-9326/ab63e1.

Shastri, H., Paul, S., Ghosh, S., Karmakar, S., 2015. Impacts of urbanization on indian summer monsoon rainfall extremes. J. Geophys. Res.-Atmos. 120 (2), 496–516. https://doi.org/10.1002/2014JD022061.

Sherly, M.A., Karmakar, S., Parthasarathy, D., Chan, T., Rau, C., 2015. Disaster vulnerability mapping for a densely populated coastal urban area: an application to Mumbai, India. Ann. Assoc. Am. Geogr. 105 (6), 1198–1220. https://doi.org/10.1080/00045608.2015.1072792.

Si, S., Zhang, H., Keerthi, S.S., Mahajan, D., Dhillon, I.S., Hsieh, C.J., 2017. Gradient boosted decision trees for high dimensional sparse output. Proceedings of the 34th International Conference on Machine Learning (ICML). 70, pp. 3182–3190 190 pp https://proceedings.mlr.press/v70/si17a.html.

Smit, B., Wandel, J., 2006. Adaptation, adaptive capacity and vulnerability. Glob. Environ. Chang. 16 (3), 282–292. https://doi.org/10.1016/j.gloenvcha.2006.03.008.

Song, Y.Y., Ying, L.U., 2015. Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry 27 (2), 130. https://doi.org/10.11919/2Fj.issn.1002-0829.215044.

Tehrany, M.S., Pradhan, B., Jebur, M.N., 2013. Spatial prediction of flood susceptible areas using rule-based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J. Hydrol. 504, 69–79. https://doi.org/10.1016/j.jhydrol.2013.09.034.

Tehrany, M.S., Pradhan, B., Mansor, S., Ahmad, N., 2015. Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. Catena 125, 91–101. https://doi.org/10.1016/j.catena.2014.10.017.

Toosi, A.S., Calbimonte, G.H., Nouri, H., Alaghmand, S., 2019. River basin-scale flood hazard assessment using a modified multi-criteria decision analysis approach: a case study. J. Hydrol. 574, 660–671. https://doi.org/10.1016/j.jhydrol.2019.04.072.

United Nations International Strategy for Disaster Reduction, 2009. Terminology on Disaster Risk Reduction.

Vittal, H., Karmakar, S., Ghosh, S., Murtugudde, R., 2020. A comprehensive India-wide social vulnerability analysis: highlighting its influence on hydro-climatic risk. Environ. Res. Lett. 15 (1), 014005. https://doi.org/10.1088/1748-9326/ab6499.

Vojtek, M., Vojteková, J., 2019. Flood susceptibility mapping on a national scale in Slovakia using the analytical hierarchy process. Water 11 (2), 364. https://doi.org/10.3390/w11020364.

Wang, Y., Fang, Z., Hong, H., Peng, L., 2020. Flood susceptibility mapping using convolutional neural network frameworks. J. Hydrol. 582, 124482. https://doi.org/10.1016/j.jhydrol.2019.124482.

Ward, P., Aerts, J.C.J.H., Botzen, W.J., Bates, P., Kwadijk, J., Hallegatte, S., Winsemius, H., 2017. Future costs and benefits of river flood protection in urban areas: a global framework. Nat. Clim. Chang. 6, 381–385. https://doi.org/10.1038/nclimate3350.

Wei, Y.M., Fan, Y., Lu, C., Tsai, H.T., 2004. The assessment of vulnerability to natural disasters in China by using the DEA method. Environ. Impact Assess. Rev. 24 (4), 427–439. https://doi.org/10.1016/j.eiar.2003.12.003.

Williamson, B.D., Gilbert, P.B., Carone, M., Simon, N., 2021. Nonparametric variable importance assessment using machine learning techniques. Biometrics 77 (1), 9–22. https://doi.org/10.1111/biom.13392.

Yamini, O.A., Kavianpour, M.R., Movahedi, A., 2020. Performance of hydrodynamics flow on flip buckets spillway for flood control in large dam reservoirs. J. Human Earth Future 1 (1), 39–47. https://doi.org/10.28991/HEF-2020-01-01-05.

Yang, X., Chapman, G.A., Young, M.A., Gray, J.M., 2005, Decemberr. Using compound topographic index to delineate soil landscape facets from digital elevation models for comprehensive coastal assessment. MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, pp. 1511–1517 DOI:10.1.1.640.6132.

Yap, G.L.C., Ismail, W.R., Isa, Z., 2013. An alternative approach to reduce dimensionality in data envelopment analysis. J. Mod. Appl. Stat. Methods 12 (1), 17. https://doi.org/10.22237/jmasm/1367381760.