



# Flood risk mapping for the lower Narmada basin in India: a machine learning and IoT-based framework

Nikunj K. Mangukiya<sup>1</sup> · Ashutosh Sharma<sup>1</sup>

Received: 7 December 2021 / Accepted: 29 March 2022 / Published online: 18 April 2022  
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

## Abstract

Floods have a significant economic, social, and environmental impact in developing countries like India. Settlements in flood hazard zones increase flood risk due to a lack of information and awareness. The present study proposed a machine learning-based framework to identify such flood risk zones for the lower Narmada basin in India. Flood hazard factors like elevation and slope of the terrain, distance from main river network, drainage density, annual average rainfall of the area, and land-use land-cover (LULC) characteristics, as well as flood vulnerability factors like population density, agricultural production, and road–river intersections, were used as predictors in the random forest algorithm to predict the flood depth in the region. Initially, the flood depth obtained from the hydrodynamic model was used as a predict and to train the model and determine the weightage of each predictor. The RandomizedSeachCV technique was used to optimize hyperparameters of the random forest algorithm. The obtained results from variable importance of random forest show that the elevation of the terrain, LULC characteristics, distance from the main river network, and rainfall are the major contributors to cause flood risk in the area. Furthermore, the possibility of using the IoT-based sensor to develop the real-time flood risk mapping framework is described. The developed flood risk map can assist policymakers, stakeholders, and citizens in developing guidelines, taking preventive measures, and avoid unnecessary settlements in flood risk zones.

**Keywords** Flood risk · Machine learning · Random forest · Hazard · Vulnerability

## Abbreviations

CV	Cross-validation
DEM	Digital elevation model
DT	Decision tree
GIS	Geographic information system
HD	Hydrodynamic
IoT	Internet of things

---

✉ Ashutosh Sharma  
ashutosh.sharma@hy.iitr.ac.in

Nikunj K. Mangukiya  
nikunj\_m@hy.iitr.ac.in

<sup>1</sup> Department of Hydrology, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India

LULC	Land-use land-cover
MAE	Mean absolute error
ML	Machine learning
MSE	Mean squared error
$R^2$	Coefficient of determination
RF	Random forest
RMSE	Root mean squared error
SRTM	Shuttle Radar Topography Mission

## 1 Introduction

Floods are one of the most destructive natural disasters worldwide, causing death, damaging infrastructure, disrupting transportation and communication, and affecting people's livelihood (Mohapatra and Singh 2003; Teng et al. 2017; Khosravi et al. 2018; Yadav and Mangukiya 2021). In developing countries like India, floods cause an enormous scale impact on the nation's socioeconomic condition (Mohapatra and Singh 2003; National Disaster Management Authority 2008). The consequences, on the other hand, are not always negative. Floods can provide long-term environmental benefits to areas that have been ravaged by drought (Balica et al. 2013; Teng et al. 2017). Floods are caused by continuous high-intensity rainfall in a specific area for an extended period (Dhar et al. 1981; Kumar et al. 2013; Ramkar and Yadav 2021). Global climate change and anthropogenic activities such as land-use patterns and deforestation have recently increased the frequency and duration of high-intensity rainfall and caused floods in the arid and semiarid river basin (Li et al. 2015; Papaioannou et al. 2016; Mangukiya and Yadav 2021; Yadav and Mangukiya 2021).

Flood risk assessment is becoming increasingly important in today's world for proper flood management (Merz et al. 2014; Manfreda and Samela 2019). Many researchers have made systematic efforts to identify flood risk zones and map inundation areas (Chen et al. 2011; Masood and Takeuchi 2012; Balica et al. 2013; Zehra and Afsar 2016; Kabenge et al. 2017; Ntajal et al. 2017; Darabi et al. 2019; Abdrabo et al. 2020; Ramkar and Yadav 2021). Flood risk indicates the probability of devastating impact and is determined as a product of hazard and vulnerability (Sayers et al. 2002; Winsemius et al. 2013). Geo-morphological and geo-environmental factors determine the likelihood of a flood hazard at a specific location (Khosravi et al. 2016). The potentially vulnerable factors of the flood are humans and exposed economic elements (Jato-Espino et al. 2019). Different researchers have used various factors to determine flood risk based on area-specific characteristics (Darabi et al. 2019; Ramkar and Yadav 2021). The elevation and slope of the terrain, drainage density, distance from the river, rainfall distribution, and soil and land-use land-cover (LULC) characteristics are most widely used as flood hazard factors, while population density and characteristics, industrial and agricultural zone, and exposed transportation network are most commonly used as vulnerable flood factors (Balica et al. 2013; Khosravi et al. 2018; Darabi et al. 2019; Ramkar and Yadav 2021).

Flood risk mapping frameworks include various cloud models, analytical hierarchy process (AHP) combined with multi-criteria decision making (MCDM) approach, weighted comprehensive evaluation techniques, artificial intelligence and machine learning-based approach, and GIS-based methods (Wan et al. 2017; Pham et al. 2021; Liu et al. 2021; Ramkar and Yadav 2021). Depending on the interpretation factors of the flood event,

statistical models are also used for flood risk mapping (Sanyal and Lu 2005). The machine learning (ML) approach has recently increased among all methods due to its ability to assess the existing nonlinear relationship between evaluation factors and flood risk level (Ma et al. 2019; Darabi et al. 2019; Costache 2019; Li et al. 2019; Eini et al. 2020; Pham et al. 2021; Liu et al. 2021). Random forest (RF), the classification and regression trees-based ML algorithm, is well suited for multivariate prediction (Breiman 2001). The RF algorithm is widely used in ecological environment monitoring because it efficiently solves nonlinear problems (Carvalho et al. 2018). The study of flood risk mapping based on RF, on the other hand, is limited and still has a scope to explore for large river basins. Lai et al. (2015) and Wang et al. (2015) have developed a flood risk mapping and assessment framework based on the RF algorithm for Jiangxi Province's river basin, China, and concluded that the RF-based outcomes have more reliability than support vector machines (SVMs) algorithm. Similarly, Feng et al. (2015) have also concluded that the RF algorithm outperformed the artificial neural network (ANN) and the maximum likelihood method for flood risk assessment. The efficiency of the various ML algorithms is also affected by the region-specific data characteristics. However, the RF-based methods have a high tolerance for outliers and noise, avoiding overfitting. The RF, as a result, outperformed the most commonly used ML algorithms in terms of computational efficiency, accuracy, scalability, and robustness (Liu et al. 2021).

Flood risk mapping and assessment on the Indian river basin are limited due to data scarcity in observed stage-discharge data, satellite imagery of the flooded area, and observed flood depth of historical flood events (Mangukiya and Yadav 2021; Yadav and Mangukiya 2021; Ramkar and Yadav 2021). Thus, the objective of the present study is to propose a possible framework for flood risk mapping in lower Narmada basin in India based on available observed and remote sensing datasets. The main challenge in using remote sensing datasets is the heterogeneity of the data source. The available data from different sources have various scales, including a mix of unstructured and structured data (Pollard et al. 2018; Towe et al. 2020). To extract useful information and make a decision based on this heterogeneous data, a definite approach is required for integration and subsequent analyses. In the present study, a machine learning-based approach is adopted to analyze flood risk in the region. The RF algorithm was used to train the model. Additionally, the hyperparameters of the RF algorithm were optimized using the 'Randomized-SearchCV' method. The flood risk map was created using the weightage obtained from the RF feature importance parameter. Furthermore, the possibility of using the IoT-based sensor to develop the real-time flood risk mapping framework is described. The proposed methodology and created flood risk map can help policymakers, stakeholders, and citizens in developing guidelines and frameworks, take preventive measures, and avoid unnecessary settlements in flood risk zones.

## 2 Study area and data

### 2.1 Study area

The Narmada River is the fifth largest and sixth longest river of India. The Narmada Basin covers an area of 98,796 km<sup>2</sup> and is located between longitudes 72°38'–81°43' E and latitudes 21°27'–23°37' N. The lower plain of the basin is bounded by Sardar Sarovar Dam and the coastal region covering 8730 km<sup>2</sup> area mainly in the districts of Narmada, Bharuch,

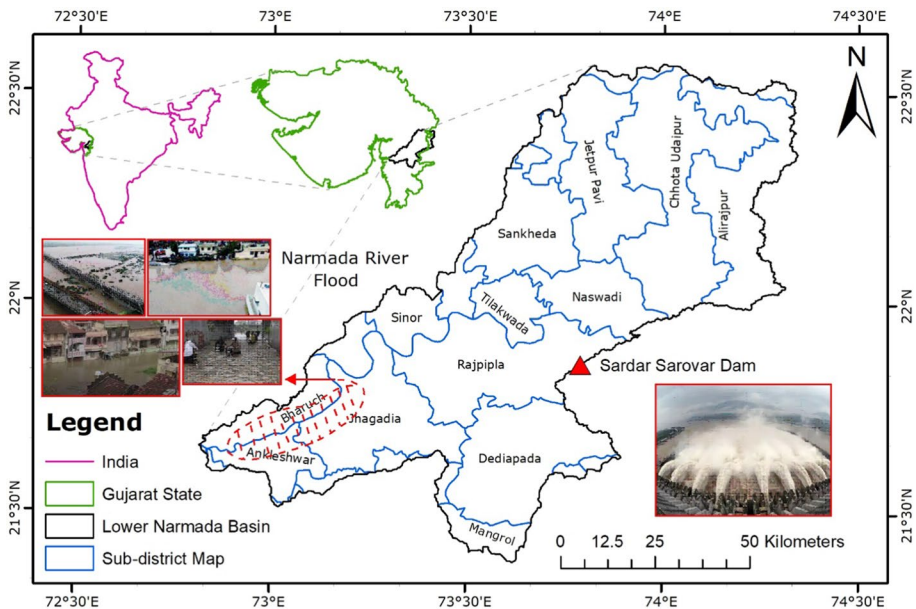
and parts of Vadodara which is more vulnerable to flood. The lower Narmada basin has experienced numerous floods, including severe flooding in 1970, 1973, 1984, 1990, 1994, and 2013. Because of that reason, the lower plain of the coastal region is selected as a study area for the present study (Fig. 1). The upper hilly regions of the basin receive higher annual rainfall (1400–1650 mm), which causes floods in the downstream area even though it is a semiarid zone. The temperature in the lower part is influenced by the sea and varies from 10 to 40 °C in different seasons. The primary LULC class of the lower Narmada basin is agriculture cropland (61%), dense vegetation and forest (21%), and urban land (8%).

## 2.2 Data description and processing

The required data for the present study were collected from an open-source database and government reports as described in Sects. 2.2.1 and 2.2.2. Nine factors were considered for preparing the flood risk map, including flood hazard causing factors and flood vulnerability causing factors. The spatial database of these factors was created and integrated with GIS software for use as input to the ML algorithm.

### 2.2.1 Flood hazard factors

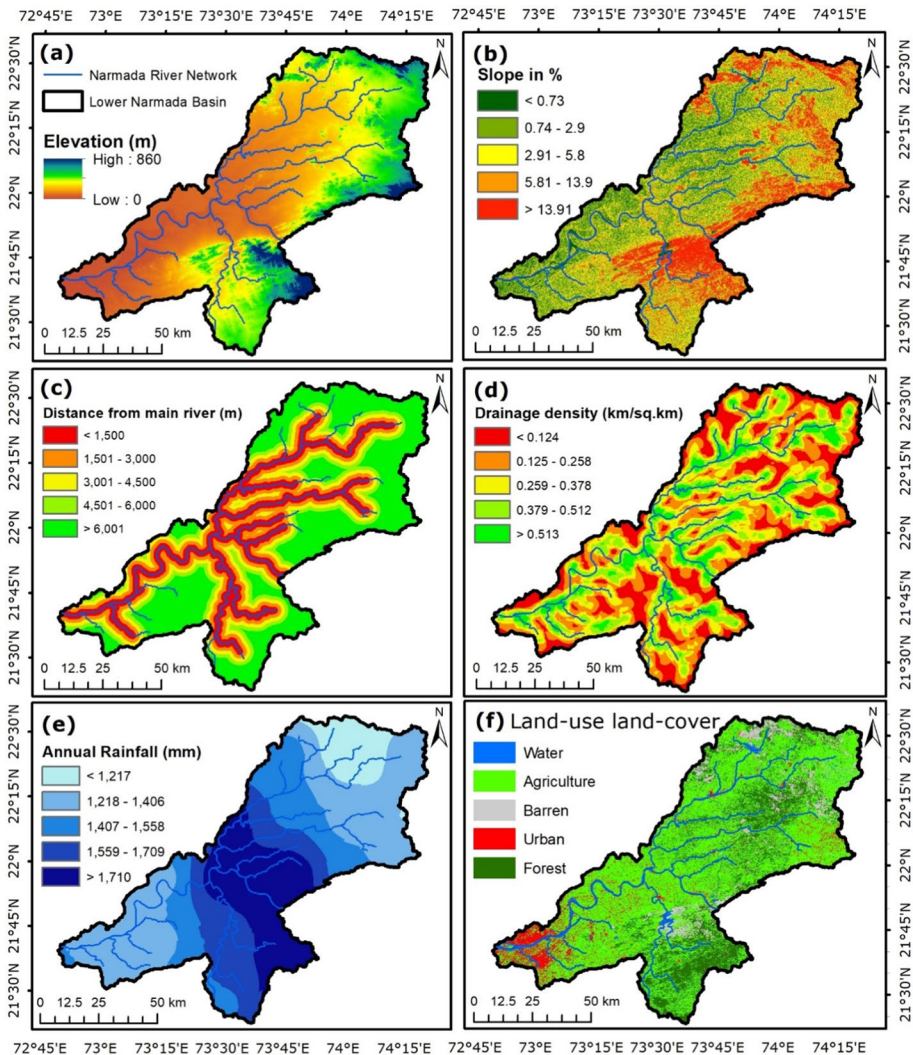
Six different factors, namely elevation of the terrain, percentage change in slope of the terrain, distance from the main river network, drainage density of the area, annual average rainfall distribution of the area, and land-use land-cover characteristics of the area, were considered as flood hazard factors for the study.



**Fig. 1** Location map of the study area. The highlighted lower part of the basin is frequently affected urban flood areas

**2.2.1.1 Elevation of the terrain** The Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) of 1-arc second resolution for the study area was downloaded from the United States Geological Survey (USGS) Earth Explorer portal (<https://earthexplorer.usgs.gov/>). The downloaded DEM was mosaicked together for watershed delineation in ArcGIS v10.8 software. Finally, the DEM was clipped to the extent of the delineated watershed for lower Narmada basin and used for the analysis (Fig. 2a). The terrain elevation ranges from 0 to 860 m, with a mean and standard deviation of 159.35 and 122.28 m.

**2.2.1.2 Percentage change in slope of the terrain** The slope of the terrain affects the speed of draining water and inundation time of the area. The waterlogging of an area can be caused



**Fig. 2** Flood hazard factors: **a** elevation, **b** percentage change in slope, **c** distance from the main river network, **d** drainage density, **e** average annual rainfall distribution, and **f** LULC characteristics of the study area

by a flat or low slope, whereas a steep or high slope can result in quick water drainage (Franci et al. 2016). The DEM was used to calculate the percentage change in slope of the terrain using the ‘Slope’ tool of ‘Spatial analyst toolbox’ in ArcGIS v10.8. The percentage slope of the area ranges from 0 to 185.3%, with a mean and standard deviation of 7.43 and 9.98%, respectively (Fig. 2b). The Sardar Sarovar dam is located in the southeast part of the basin, which has a higher slope, whereas the southwest part of the basin has a lower slope, making this area more flood-prone.

**2.2.1.3 Distance from the main river network** The area closer to the river channel has a high hazard when the water overtops the bank of the river. The hazard of the flood decreases with the increase in distance from the main river network (Khosravi et al. 2016). The DEM was used to delineate the river network by keeping the satellite image as a reference. The ‘Euclidean distance’ tool of ‘Spatial analyst toolbox’ in ArcGIS v10.8 was used to calculate and prepare the distance map from the delineated river network. The Bharuch and Ankleshwar (two major cities of the study area) are located at the downstream end, with high flood hazards due to less distance from the river channel (Fig. 2c).

**2.2.1.4 Drainage density of the area** The ratio of the river network’s total length to the river’s total drainage area is defined as the drainage density of the area. The higher drainage density value represents the exemplary river network which can take less time to drain the rainfall water. The less water travel time can result in draining a larger area simultaneously, increasing the peak of the flood and ultimately increasing the area’s flood hazard. The delineated river network was used to calculate the drainage density of the area using the ‘Line density’ tool of the ‘Spatial analyst toolbox’ in ArcGIS v10.8. The drainage density of the study area ranges from 0 to 0.88 km per square kilometers of the area, with a mean and standard deviation of 0.28 and 0.17 km/km<sup>2</sup>, respectively (Fig. 2d).

**2.2.1.5 Annual average rainfall distribution of the area** The amount of rainfall directly indicates the intensity of flood hazards. The gridded rainfall dataset of 0.25° resolution was downloaded from the India Meteorological Department (IMD, Pune) website. The annual average rainfall distribution of the area was calculated by inverse distance weighting method using ‘IDW’ tool of ‘Spatial analyst toolbox’ in ArcGIS v10.8. The rainfall in the lower Narmada basin ranges from 977 to 1944 mm, with a mean and standard deviation of 1489 and 215 mm, respectively (Fig. 2e). The middle part of the basin receives a higher amount of rainfall which drains into the downstream area and causes the flood.

**2.2.1.6 Land-use land-cover characteristics** The distribution of rainfall water into various hydrological processes such as runoff, infiltration, baseflow, and interflow is governed by the LULC characteristics of the area (Dwarakish and Ganasri 2015). The urban and built-up, agricultural, and barren land results in a higher runoff, increasing flood hazards, while the dense vegetation and forest land result in lesser runoff. The LULC map of the study area was prepared from the Landsat8 satellite imagery of 30 m resolution downloaded from the USGS Earth Explorer Portal. The ‘Maximum likelihood classification’ tool of the ‘Image Classification toolbox’ in ArcGIS v10.8 was used to train and classify the LULC features of the study region. A large number of training samples of various classes (spread throughout the study area) were provided using Google Earth photographs as a background to categorize LULC characteristics in five classes: dense vegetation, urban and built-up area, barren land, agricultural land, and water bodies. Because the LULC characteristics of the study

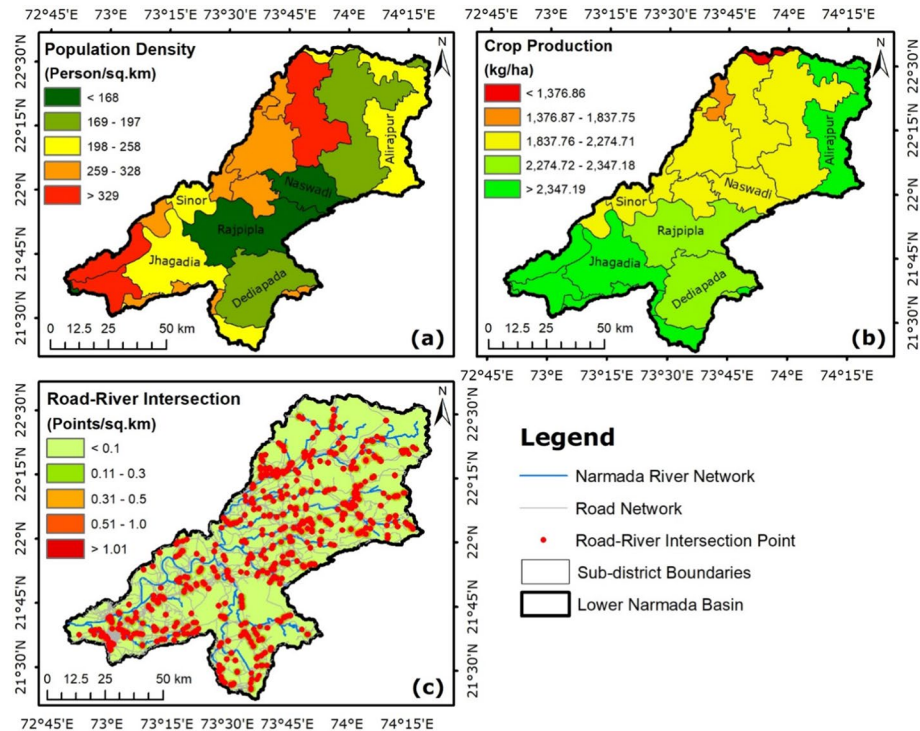


area do not change significantly over time, the LULC map from 2021 was used as a static flood hazard factor in this study. The major LULC characteristic of the study area is agricultural land (61%), followed by dense vegetation (21%), urban and built-up area (8%), barren land (7%), and water bodies (3%) (Fig. 2f).

### 2.2.2 Flood vulnerability factors

Three factors, namely population density of the area as social factor, agricultural production per unit area as economic factor, and road–river intersection points per unit area as physical transportation factor, were considered vulnerable flood factors for the study (Balica et al. 2013; Shivaprasad Sharma et al. 2017).

**2.2.2.1 Population density** Flood causes significant social loss and damage and impacts the economy and infrastructure. For considering social loss, population density is one of the critical factors (Nasiri et al. 2016; Yang et al. 2018). The sub-district-wise population data were collected from the Census of India 2011. This population data was given as input in ArcGIS v10.8 to prepare the population density map (Fig. 3a). The density of population in the study area ranges from 31 to 775 persons per square kilometer. The downstream end



**Fig. 3** Flood vulnerable factors: **a** sub-district-wise population density, **b** average annual crop production, and **c** road–river intersection point density map of the study area

of the basin, where two major cities (Bharuch and Ankleshwar) are located, has a higher population density.

**2.2.2.2 Agricultural production** India is an agricultural country. More than 80% of the rural Indian population depends on agriculture and its allied activities as the primary source of livelihood. Thus, the area's agricultural production is an important economic factor to be considered vulnerable (Chen et al. 2018). The district-wise average annual crop production data were collected from an official government report published by the Directorate of Agriculture, Gujarat State (Directorate of Agriculture 2011), and the website of the Farmer Welfare and Agriculture Development Department, Madhya Pradesh State, India (<http://mpkrishi.mp.gov.in/>). This data was digitized in ArcGIS v10.8 to prepare the agricultural production map of the study area (Fig. 3b). The annual crop production of the study area ranges from 1159 to 7345 kg per unit hectare area.

**2.2.2.3 Road–river intersection density** The transportation links are necessary for rescue work and providing essentials to flood-affected people. Any damage to transportation facilities can isolate the area and increase the flood impact. For the present study, the road and river network intersections are considered vulnerable factors as most of the transportation in the study area takes place by road network. The road network of the study area was downloaded from the DIVA-GIS portal and corrected and validated using Google's hybrid satellite image. The intersection points of the road network and delineated river network were identified using the 'Intersect' tool of 'Analysis toolbox' in ArcGIS v10.8. The density of the intersection point was calculated using the 'Point density' tool of the 'Spatial analyst toolbox' in ArcGIS v10.8 (Fig. 3c).

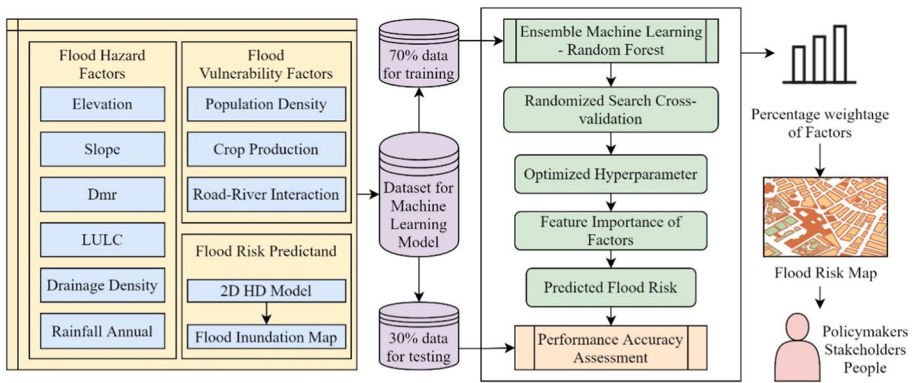
### 3 Methodology

The objective of the present study is to prepare a flood risk map for the lower Narmada basin, which can be helpful for deciding the flood prevention and mitigation strategies and social benefits. The lower Narmada basin has received a historic flood in the year 1994. As the surveyed data of the flood depth and inundation map are not available, the first step of the methodology involves the development of the hydrodynamic model to prepare the flood inundation map of the region. The developed flood inundation map, along with nine flood hazards and vulnerable factors, was then used to train the model and predict the weightage of the predictors using the RF algorithm (Table S1, Supplementary Material). The hyper-parameters of the RF were optimized with the RandomizedSearchCV method to obtain the best results. Finally, the flood risk map was prepared by using the weightage of each predictor. Figure 4 shows the schematic representation of the methodology adopted in the present study.

#### 3.1 Hydrodynamic model

The two-dimensional (2D) hydrodynamic (HD) model was developed for the lower Narmada basin using the Hydrologic Engineering Center's River Analysis System (HEC-RAS v6.1) software. In the absence of the surveyed bathymetry data, the SRTM DEM was given as input for bathymetry data. The discharge data from the Garudeshwar weir (peak





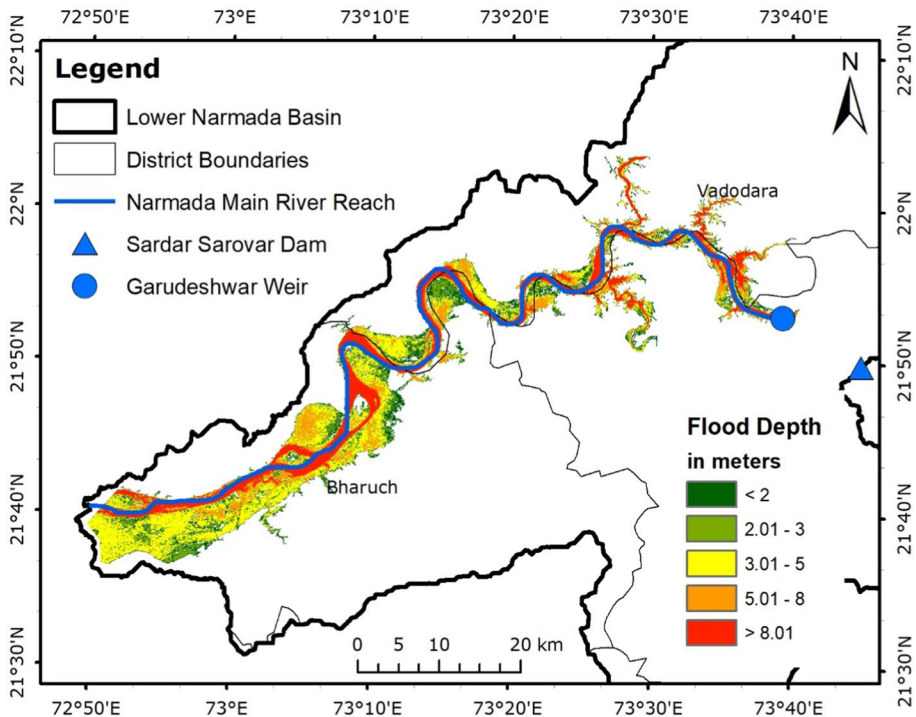
**Fig. 4** Schematic representation of the proposed framework for flood risk mapping. The dataset consists of flood hazard and vulnerability factors as predictors and flood inundation map as predictand was given as input to the RF model for obtaining weightage of each factor to prepare flood risk maps

discharge 32,056 cumec) for 2013 were provided as the upstream boundary condition. The river’s normal slope was given as a downstream boundary condition to simulate the HD model. The roughness coefficient for the bathymetry was subsequently changed to calibrate the HD model by comparing simulated flood depth with the observed water level at the Bharuch gauging station. The calibrated model was then used to develop the historic flood map of the year 1994. The discharge data from the Garudeshwar weir (with peak discharge 60,642 cumec) for 1994 was given as a boundary condition to obtain the flood depth in the region. The developed flood inundation map (Fig. 5) was given as input to the RF algorithm.

### 3.2 Random forest algorithm

RF is the most widely used ensemble machine learning algorithm because of its ability to solve a broad range of regression and classification problems. It was proposed by Breiman (2001). It also includes critical hyperparameters and heuristics for optimizing them. RF is an extension of bagging (bootstrap aggregation) of decision trees. To develop a decision tree (DT) model, RF collects random samples of fixed numbers from the dataset by bootstrap sampling technique. After generating the DT model from one random sample, the sample dataset was again put back. So, the sample collected to develop one DT is likely to be collected again for developing another DT. These ensure that each random sample’s content is different, which results in the variation in each of the developed DT models in the forest. The final results were obtained by taking the mean of the results from all individual DT models of the forest.

Taking the example of the present flood risk study, the RF assumes that  $M$  numbers of flood risk factors cause the flood depth in the particular location. Firstly, the samples were split randomly into two datasets: training (70%) and testing (30%). The RF will randomly choose the sample from the training dataset to build a DT and put the sample back in the training dataset. Secondly, the  $M_{test}$  ( $M_{test} \leq M$ ) numbers of flood risk factors will be randomly selected from the  $M$  numbers. Each DT will be developed by a random sample from the training dataset with random  $M_{test}$  numbers of flood causing factors, and the tree would grow to the maximum depth. The mean squared error (MSE) (Eq. 1) will be



**Fig. 5** Simulated flood depth corresponding to peak discharge of the year 1994 from the Garudeshwar weir. 65% of the inundation area has a flood depth of more than 3 m

calculated at each tree node for splitting the sample into two branches of the tree. In the end, the importance of each flood risk factor for building the DT will be stored. This process will be repeated to develop  $N$  numbers of decision trees ( $N_{tree}$ ). Finally, the developed  $N_{tree}$  will be aggregated to build a random forest, and the mean of the results from  $N_{tree}$  will be obtained as a final result of this random forest.

### 3.3 Optimization of hyperparameter

The RF algorithm can be set up with different hyperparameters; among them, the number of trees in the forest ( $N_{tree}$  or  $n_{estimators}$ ) and the number of features selected for splitting at each node ( $M_{test}$  or  $max\_features$ ) are most important. The maximum depth of the DT ( $max\_depth$ ), the number of data samples required at a node before the split ( $min\_samples\_split$ ), and the number of data samples needed in a leaf to allow the split ( $min\_samples\_leaf$ ) are also optimized in the literature for better results. These hyperparameters can be optimized using two different methods: grid search and randomized search cross-validation. The grid search method iteratively runs all possible combinations of the hyperparameters defined by users, which is time-consuming and necessitates more computation resources. The randomized search method, on the other hand, chooses hyperparameter combinations at random from a predefined subset. As a result, the randomized search method does not necessitate a large amount of computation resources and is also less time-consuming (Paper 2020). Also, the randomized search method greatly simplifies

analysis without significantly sacrificing optimization. Due to this reason, in the present study, different combination value of the hyperparameters was evaluated using random search cross-validation (RandomizedSeachCV) method to find out the optimized value. The RandomizedSearchCV method will randomly take the value of hyperparameter from the defined grid and evaluate each combination of values by many iterations with  $K$ -fold cross-validation (CV). In  $K$ -Fold CV, the training dataset will be split into  $K$  number of groups or folds of equal size. Then the model will be iteratively fitted on ‘ $K-1$ ’ folds and tested on the  $K$ th fold (James et al. 2021). The results of each of the folds will be averaged to obtain final validation metrics. The performance of the different combinations of the hyperparameters was evaluated using statistical indices, i.e., mean squared error (MSE) (Eq. 1), mean absolute error (MAE) (Eq. 2), root mean squared error (RMSE) (Eq. 3), and coefficient of determination ( $R$ -square) (Eq. 4).

$$MSE = \frac{\sum_{i=1}^N (y_{act,i} - y_{pred,i})^2}{N} \tag{1}$$

$$MAE = \frac{\sum_{i=1}^N (y_{pred,i} - y_{act,i})}{N} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_{act,i} - y_{pred,i})^2}{N}} \tag{3}$$

$$R^2 = \left\{ \frac{\sum_{i=1}^N (\bar{y}_{act})(y_{pred,i} - \bar{y}_{pred})}{\sqrt{\sum_{i=1}^N (y_{act,i} - \bar{y}_{act})^2} \cdot \sqrt{\sum_{i=1}^N (y_{pred,i} - \bar{y}_{pred})^2}} \right\}^2 \tag{4}$$

where  $N$  is the number of data points,  $y_{act}$  is the actual flood depth,  $y_{pred}$  is the predicted flood depth by the RF,  $\bar{y}_{act}$  and  $\bar{y}_{pred}$  is the average of actual and predicted flood depth, respectively.

### 3.4 Preparation of flood risk map

The flood risk is the degree of exposure to an undesirable event caused by the combination of natural flood hazard factors with those who are particularly vulnerable to this disaster. In the present study, six flood hazard factors and three vulnerable flood factors are considered. The factors were divided into five classes (Table S2, Supplementary material) for the development of flood hazard and flood vulnerable maps using Jenks natural breaks classification method (Jenks 1967). The weightage of each flood hazard and vulnerable factor obtained by RF was used to prepare the flood hazard and vulnerable raster using the ‘weighted overlay’ tool of ‘Spatial analyst toolbox’ in ArcGIS v10.8. The flood risk raster was prepared by multiplying flood hazard and vulnerable raster using the ‘raster calculator’ tool of ‘Spatial analyst toolbox’ in ArcGIS 10.8. Finally, the flood risk raster was reclassified into four classes to prepare the flood risk map indicating the level of risk at a particular location. Because the values of classes 1 and 2 in flood risk raster do not vary significantly,

both classes were combined together and considered as high-risk zone, whereas the study region in class 5 is primarily mountain ranges with a very low likelihood of flooding and was considered as a no-risk zone.

## 4 Results and discussion

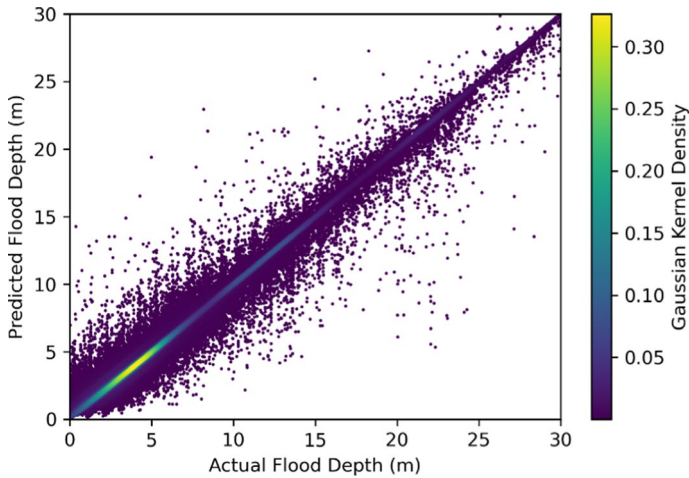
The present study involves nine predictor factors to prepare a risk map. The calibrated and validated flood inundation map simulated using HEC-RAS v6.1 was used as a predictand to train the RF model. Together, all ten datasets were projected to a unified coordinate system of ‘WGS 1984 Universal Transverse Mercator (UTM)’. The dataset was then resampled to a grid resolution of  $30 \times 30$  m using ArcGIS v10.8. These resampled and projected data were then given as input to train the RF model.

### 4.1 Results of random forest model

The RF package developed by the ‘scikit-learn’ in python language was implemented to build the RF model for developing a flood risk map in the present study. The hyperparameter of the RF model was optimized by evaluating combination values from the developed grid (Table 1) using the RandomizedSearchCV method. Together, there were 640 different settings which were evaluated by 300 iterations using the threefold CV method. More number of the iterations can evaluate more settings from the grid, and more CV folds can reduce the probability of overfitting, but it also increases the run time of the model. As a flood can be an instantaneous event, a larger run time should be avoided so that the model can predict the result before the actual event. Keeping this in view, the best hyperparameter combination setting was obtained as:  $n\_estimators = 1000$ ,  $max\_depth = 100$ ,  $max\_features = 'sqrt'$ ,  $min\_samples\_leaf = 2$ , and  $min\_samples\_split = 10$ . The performance of the model with the best hyperparameter setting was evaluated by statistical indices, and found as:  $MAE = 0.34$  m,  $MSE = 0.54$  m<sup>2</sup>,  $RMSE = 0.74$  m, and  $R\text{-square} = 0.98$ . The predicted flood depth by the RF model using the test dataset was compared with the actual flood depth in Fig. 6. As the model is trained and tested on a high-resolution grid dataset (over 600,000 grid points), the predicted flood depth shows the false prediction on some grid points due to homogeneity of the predictor factors in those grid points. But overall, the higher kernel density (Rosenblatt 1956; Parzen 1962) was achieved on the equipotential line ( $y_{act} \cong y_{pred}$ ) in the graph.

**Table 1** Grid of hyperparameter settings for optimization

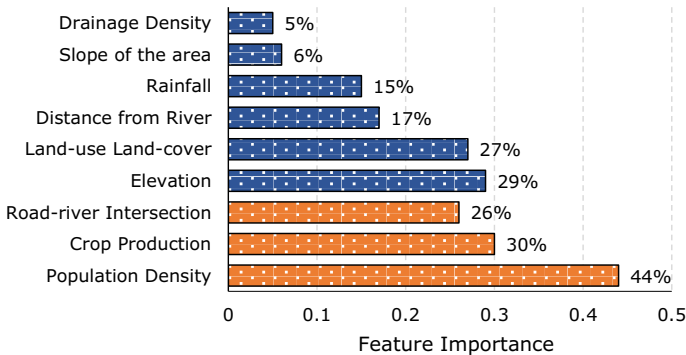
Hyperparameter	Settings
Number of DT in RF (Ntrees or $n\_estimators$ )	[100, 500, 800, 1000]
Maximum depth of DT ( $max\_depth$ )	['none', 20, 50, 100, 200]
Number of features to selected for split ( $max\_features$ )	['auto', 'sqrt']
Number of data samples needed in a leaf to allow the split ( $min\_samples\_leaf$ )	[1, 2, 5, 10]
Number of data samples required at a node before the split ( $min\_samples\_split$ )	[2, 5, 10, 20]



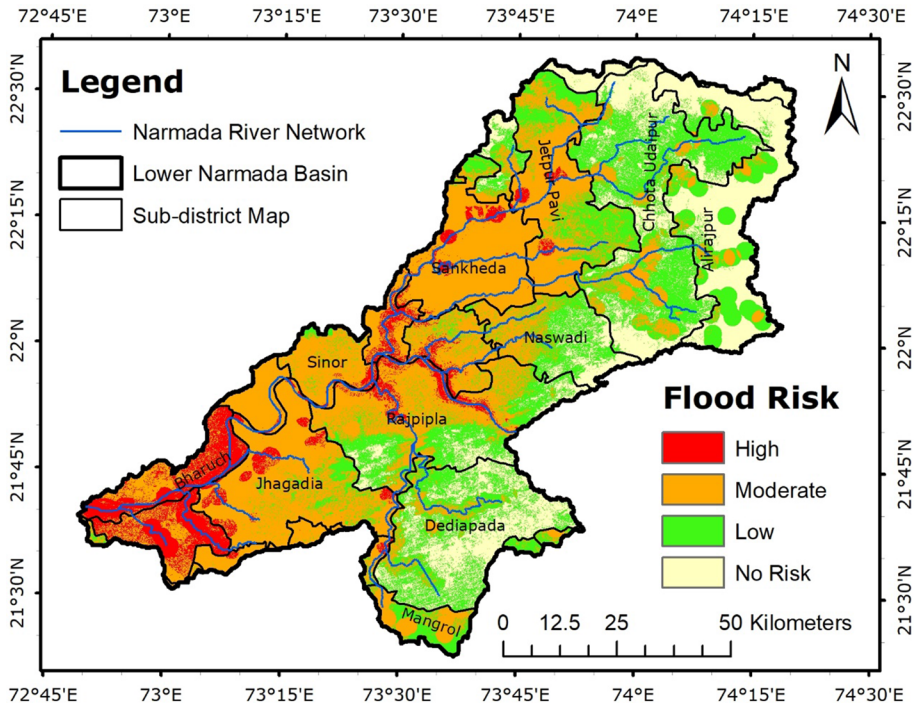
**Fig. 6** The Gaussian Kernel density plot for comparing actual flood depth with predicted flood depth by the RF model. The higher point density was achieved at the equipotential line (1:1) of the plot

### 4.2 Flood risk map

The flood risk map was prepared by using the weightage obtained by the feature importance value of the RF model. The flood hazard raster was prepared by the ‘weighted overlay’ tool of ArcGIS v10.8 by multiplying the percentage weightage of each flood hazard causing factor. Similarly, the vulnerable flood raster was prepared by multiplying the percentage weightage of each vulnerable flood factor. Finally, the flood risk map was prepared by multiplying flood hazard and vulnerable raster using the ‘raster calculator’ tool of ArcGIS v10.8. The obtained percentage weightage from the RF model for all factors is shown in Fig. 7, and the developed flood risk map is shown in Fig. 8. The elevation and LULC characteristics of the region were given the highest importance (29 and 27%) in flood hazard factors by the RF model. The distance from the main river channel and rainfall also has significant importance (17 and 15%) in causing the flood hazard. These



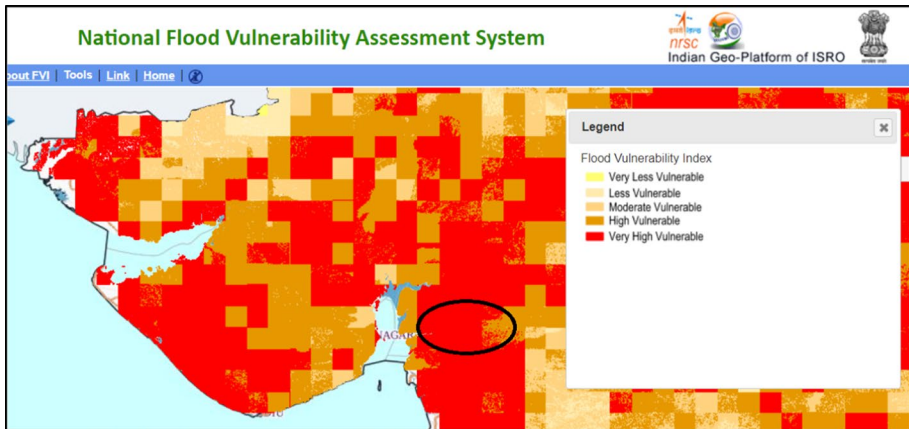
**Fig. 7** The feature importance (weightage) of flood hazard and vulnerability factors obtained from the RF model. The blue bar indicates the flood hazard factors, while the orange bar indicates the flood vulnerability factors



**Fig. 8** The developed flood risk map of the study area by multiplying the flood hazard and vulnerability map. The lower urban area of the basin is prone to high flood risk, while the upper mountainous area is prone to low or no flood risk

results exceptionally would be helpful to decide the development policy in the region. The restrictions and guidelines could be applied for the development near the river banks to reduce the flood hazard. The RF model gave the region's population density and crop production the highest importance (44 and 30%) in flood vulnerable factors. The major cities in the study area are located at the bank of the river, which is more vulnerable to a flood event. The majority of the study area is used for agricultural activities. Many agricultural farms are located even on the delta part of the river due to the higher productive capacity of the land, which makes it more vulnerable to flood. Due to the region's lower elevation and relatively higher rainfall statistics, the study area was considered very high vulnerable by the National Flood Vulnerability Assessment System developed by the Indian geo-platform of the Indian Space Research Organization (ISRO) (Fig. 9). The ISRO's flood vulnerability map is developed using a multi-criteria evaluation technique considering the factors such as probable maximum precipitation obtained by IMD gridded rainfall data, runoff potential obtained by LULC and soil texture, drainage density, and slope of the terrain. The developed flood risk map indicates that Bharuch and Ankleshwar, located in the downstream area, are at high risk due to higher population density and lesser distance from the main river. The majority of the study area, including the sub-district of the Jhagadia, Sinor, Rajpipla, and Sankheda, are at moderate risk due to high rainfall distribution and lower elevation of the region. This result indicates the requirement of a proper storm-water drainage system in a particular area to reduce the flood risk. The region's northern and southern





**Fig. 9** Flood vulnerability map developed by National Flood Vulnerability Assessment System, Indian geo-platform of ISRO using multi-criteria evaluation technique. Source <https://bhuvan-app1.nrsc.gov.in/nfvas/#>

east parts are at low or no risk zone due to their higher elevation and slope, which quickly drains the water from these regions.

In the present study, the developed flood risk map may have various sources of uncertainty, including uncertainty in the given input data and weighing scheme of the proposed ML algorithm. The SRTM DEM is used as the elevation of the terrain, which may contain measurement errors, and this can propagate to other factors such as the slope of the terrain, which is derived from the elevation map. The ground-based survey data would be useful in overcoming this limitation. Natural variability, such as changes in river pathways and LULC, may need to be updated on a regular basis to deal with the uncertainty of these factors. The dynamic factor such as annual rainfall may include measurement errors that must be addressed. However, while developing the flood risk map, the ranks are assigned based on the different class values of these factors, which may account for minor uncertainties in the input data. The ML model and its parameter uncertainty play a critical role in the output from the model. However, the RF model is an ensemble ML algorithm, which may reduce the uncertainty in the output.

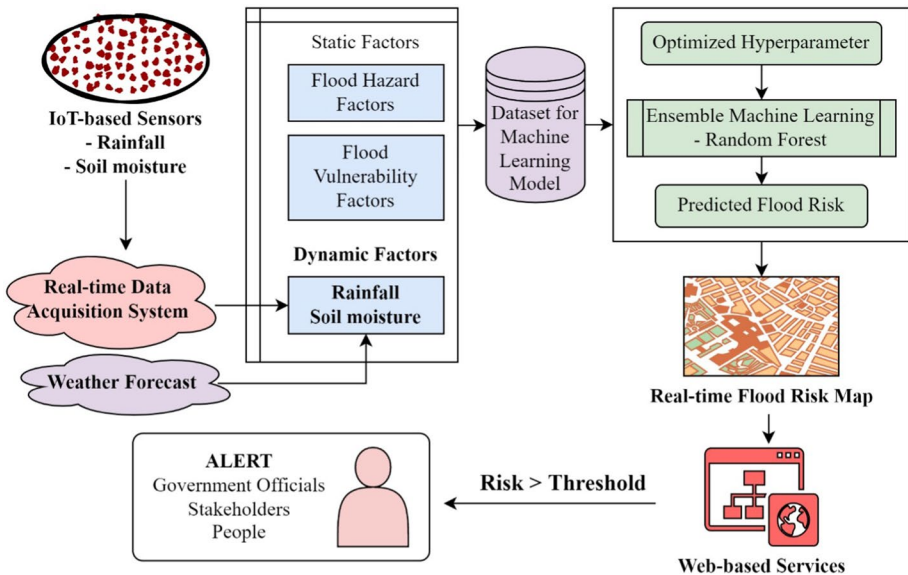
### 4.3 Integrated ML and IoT-based framework for real-time flood risk mapping

The present study demonstrates that the remote sensing dataset can be utilized for flood risk assessment and mapping studies in the data scarce region. The critical challenge in using remote sensing-based big data is to develop an integrated analysis approach that can extract useful information from heterogeneous data. The traditional method of flood risk prediction based on the hydrodynamic model or empirical relationships has its limitation in terms of data integration and utilization. Also, these methods require high computation time for simulating large high-resolution datasets. To overcome this limitation, the ML approach is demonstrated in the present study to develop the integrated framework for flood risk mapping. The developed flood risk map for the lower Narmada basin using the RF model shows that the rainfall distribution of the region is one of the crucial factors which is dynamic for the prediction of flood risk maps. These also indicate the possibility of exploring more such dynamic hydrological process variable factors (such as soil

moisture condition, evaporation, and atmospheric variables), which can influence the flood risk in the region.

The IoT-based sensor network for measuring rainfall and soil moisture datasets can be constructed in the region to collect real-time data. Also, the monitoring stations can be developed to gather information related to flood depth or hazard conditions from the recently developed new paradigm of social IoT (combining the social network with IoT). The dataset collected from the IoT network can be integrated with the remote sensing dataset using a GIS platform or cloud-based services to prepare the unified dataset. The traditional method may not extract helpful information from such dynamic and diversified data. The ML approach, such as the RF model, can be trained and calibrated based on historical data and can be used to extract useful information from the real-time heterogeneous data. However, because the majority of IoT-based devices are low-end, high-end computational resources are required to accomplish such processing. Furthermore, because the ML algorithms detect changes in the given input data and produce the desired output, IoT would be useful to provide real-time input and generate real-time output using ML. Based on the ML model results, the real-time flood risk map can be developed and published with the help of web-based services (Fig. 10).

The IoT network, along with weather forecast data, can also help prepare the early warning system. The IoT network will collect ground-based datasets such as soil moisture and streamflow, which are critical factors in generating floods. The weather forecast, such as precipitation, can be given as input to the ML model for generating future flood risk scenarios. The dataset collected from the IoT network and weather forecast can be clustered based on the spatial location. The region-wise thresholds for risk level can be defined based on historical experiences. The warning system can be set up to send an alert message to authorities and stakeholders when the IoT-based data and weather forecast surpasses the



**Fig. 10** Proposed framework for developing real-time flood risk map using integrated IoT-based sensors and ML algorithm. The real-time data from the IoT-based network and static factors will be given as input to the ML algorithm to prepare a flood risk map

threshold. However, the accuracy of such flood risk forecast will directly depend on the accuracy of the input data. As weather forecasts can be uncertain sometimes, the decision should be made based on historical experiences. In the absence of flood risk, the constructed IoT sensor network can be helpful to monitor the region's irrigation requirement, as the majority of the study area has an agricultural land-use pattern. Such a multi-use of the dataset can result in long-term economic and social benefits for the initially invested resources.

## 5 Conclusions

The present study prepared the flood risk map for the lower Narmada basin using a machine learning algorithm based on available observed and remote sensing-based heterogeneous data. The methodology described in the present study could be helpful in the data-scarce region to predict the flood risk. The traditional methods for flood risk prediction, such as hydrodynamic modeling and empirical relationships, has limitations in utilizing the heterogeneous and diversified dataset. The ensemble machine learning algorithm such as RF can be helpful to extract the information from remote sensing-based heterogeneous data in flood risk mapping. The cross-validation method can be helpful to improve the performance of the model and find the balance between run time and accuracy of the model. The results from the present study indicate that the elevation, distance from the main river, land-use land-cover characteristics, and rainfall are the more significant flood hazard factors. The population density and agricultural production are more exposed and vulnerable to the flood hazard. The developed flood risk map based on the feature importance of the flood hazard and vulnerable factors shows that the two major cities (Bharuch and Ankleshwar) are at high flood risk, and four sub-districts (Jhagadia, Sinor, Rajpipla, and Sankheda) are at medium flood risk in the lower Narmada basin. As rainfall is a dynamic and significant flood risk predictor, the IoT-based sensor network for measuring rainfall and other hydrological process variables can help prepare a real-time flood risk map of the region. The proposed IoT-based framework and machine learning-based methodology to extract helpful information from heterogeneous data would be beneficial for government officials and policymakers to take necessary actions to reduce flood risk in the lower Narmada basin.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11069-022-05347-2>.

**Acknowledgements** The authors are thankful to the India Meteorological Department (IMD), Pune; Sardar Sarovar Narmada Nigam Limited (SSNNL), Gandhinagar; Directorate of Agriculture, Gujarat; Farmer Welfare and Agriculture Development Department, Madhya Pradesh; Census office, Government of India; DIVA-GIS Team for providing necessary data for the study reported in the paper. The authors are grateful to Bhuvan and India's national remote sensing center for providing the flood vulnerability index map. The authors are thankful to anonymous reviewers for their constructive suggestions in improving the quality of the manuscript. The authors appreciate the computational resources provided by Microsoft AI for Earth grant.

**Funding** Microsoft Azure computational resources were provided by the Microsoft AI for Earth grant awarded to Ashutosh Sharma for this work.

**Data availability** The data used to support the findings of this study are available from the corresponding author upon request.

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- Abdrabo KI, Kantoush SA, Saber M et al (2020) Integrated methodology for urban flood risk mapping at the microscale in ungauged regions: a case study of Hurghada. *Egypt Remote Sens* 12:3548. <https://doi.org/10.3390/rs12213548>
- Balica SF, Popescu I, Beevers L, Wright NG (2013) Parametric and physically based modelling techniques for flood risk and vulnerability assessment: a comparison. *Environ Model Softw* 41:84–92. <https://doi.org/10.1016/j.envsoft.2012.11.002>
- Breiman L (2001) Random forests. *Mach Learn*. <https://doi.org/10.1023/A:1010933404324>
- Carvalho J, Santos JPV, Torres RT et al (2018) Tree-based methods: concepts, uses and limitations under the framework of resource selection models. *J Environ Inform*. <https://doi.org/10.3808/jei.201600352>
- Chen YR, Yeh CH, Yu B (2011) Integrated application of the analytic hierarchy process and the geographic information system for flood risk assessment and flood plain management in Taiwan. *Nat Hazards* 59:1261–1276. <https://doi.org/10.1007/s11069-011-9831-7>
- Chen H, Liang Z, Liu Y et al (2018) Effects of drought and flood on crop production in China across 1949–2015: spatial heterogeneity analysis with Bayesian hierarchical modeling. *Nat Hazards* 92:525–541. <https://doi.org/10.1007/s11069-018-3216-0>
- Costache R (2019) flood susceptibility assessment by using bivariate statistics and machine learning models: a useful tool for flood risk management. *Water Resour Manag* 33:3239–3256. <https://doi.org/10.1007/s11269-019-02301-z>
- Darabi H, Choubin B, Rahmati O et al (2019) Urban flood risk mapping using the GARP and QUEST models: a comparative study of machine learning techniques. *J Hydrol* 569:142–154. <https://doi.org/10.1016/j.jhydrol.2018.12.002>
- Dhar ON, Rakhecha PR, Mandal BN, Sangam RB (1981) The rainstorm which caused the morvi dam disaster in august 1979. *Hydrol Sci Bull* 26:71–81. <https://doi.org/10.1080/02626668109490862>
- Directorate of Agriculture (2011) District-wise area, production and yield of important food & non-food crops in Gujarat State. Directorate of Agriculture, Gandhinagar, Gujarat, India
- Dwarakish GS, Ganasri BP (2015) Impact of land use change on hydrological systems: a review of current modeling approaches. *Cogent Geosci* 1:1–18. <https://doi.org/10.1080/23312041.2015.1115691>
- Eini M, Kaboli HS, Rashidian M, Hedayat H (2020) 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. <https://doi.org/10.1016/j.ijdrr.2020.101687>
- Feng Q, Liu J, Gong J (2015) Urban flood mapping based on unmanned aerial vehicle remote sensing and random forest classifier: a case of Yuyao, China. *Water* 7:1437–1455. <https://doi.org/10.3390/w7041437>
- Franci F, Bitelli G, Mandanici E et al (2016) Satellite remote sensing and GIS-based multi-criteria analysis for flood hazard mapping. *Nat Hazards* 83:31–51. <https://doi.org/10.1007/s11069-016-2504-9>
- James G, Witten D, Hastie T, Tibshirani R (2021) An introduction to statistical learning. Springer, New York
- Jato-Espino D, Lobo A, Ascorbe-Salcedo A (2019) Urban flood risk mapping using an optimised additive weighting methodology based on open data. *J Flood Risk Manag*. <https://doi.org/10.1111/jfr3.12533>
- Jenks GF (1967) The data model concept in statistical mapping. *Int Yearb Cartogr* 7:186
- Kabenge M, Elaru J, Wang H, Li F (2017) Characterizing flood hazard risk in data-scarce areas, using a remote sensing and GIS-based flood hazard index. *Nat Hazards* 89:1369–1387. <https://doi.org/10.1007/s11069-017-3024-y>
- Khosravi K, Nohani E, Maroufina E, Pourghasemi HR (2016) A GIS-based flood susceptibility assessment and its mapping in Iran: a comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decision-making technique. *Nat Hazards* 83:947–987. <https://doi.org/10.1007/s11069-016-2357-2>
- Khosravi K, Pham BT, Chapi K et al (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>

- Kumar KN, Rajeevan M, Pai DS et al (2013) On the observed variability of monsoon droughts over India. *Weather Clim Extrem* 1:42–50
- Lai C, Chen X, Zhao S et al (2015) A flood risk assessment model based on Random Forest and its application. *J Hydraul Eng* 46:58–66. <https://doi.org/10.13243/j.cnki.slxh.2015.01.008>
- Li J, Liu X, Chen F (2015) Evaluation of nonstationarity in annual maximum flood series and the associations with large-scale climate patterns and human activities. *Water Resour Manag* 29:1653–1668. <https://doi.org/10.1007/s11269-014-0900-z>
- Li X, Yan D, Wang K et al (2019) Flood risk assessment of global watersheds based on multiple machine learning models. *Water* 11:1654. <https://doi.org/10.3390/w11081654>
- Liu Y, Lu X, Yao Y et al (2021) Mapping the risk zoning of storm flood disaster based on heterogeneous data and a machine learning algorithm in Xinjiang, China. *J Flood Risk Manag*. <https://doi.org/10.1111/jfr3.12671>
- Ma M, Liu C, Zhao G et al (2019) Flash flood risk analysis based on machine learning techniques in the Yunnan Province, China. *Remote Sens* 11:170. <https://doi.org/10.3390/rs11020170>
- Manfreda S, Samela C (2019) A digital elevation model based method for a rapid estimation of flood inundation depth. *J Flood Risk Manag*. <https://doi.org/10.1111/jfr3.12541>
- Mangukiya NK, Yadav SM (2021) Integrating 1D and 2D hydrodynamic models for semi-arid river basin flood simulation. *Int J Hydrol Sci Technol* 1:1. <https://doi.org/10.1504/IJHST.2021.10035928>
- Masood M, Takeuchi K (2012) Assessment of flood hazard, vulnerability and risk of mid-eastern Dhaka using DEM and 1D hydrodynamic model. *Nat Hazards* 61:757–770. <https://doi.org/10.1007/s11069-011-0060-x>
- Merz B, Aerts J, Arnbjerg-Nielsen K et al (2014) Floods and climate: emerging perspectives for flood risk assessment and management. *Nat Hazards Earth Syst Sci* 14:1921–1942. <https://doi.org/10.5194/nhess-14-1921-2014>
- Mohapatra PK, Singh RD (2003) Flood management in India. *Nat Hazards* 28:131–143. <https://doi.org/10.1023/A:1021178000374>
- Nasiri H, Mohd Yusof MJ, Mohammad Ali TA (2016) An overview to flood vulnerability assessment methods. *Sustain Water Resour Manag* 2:331–336. <https://doi.org/10.1007/s40899-016-0051-x>
- National Disaster Management Authority (2008) National disaster management guidelines: management of floods. Government of India, New Delhi, India
- Ntajal J, Lamptey BL, Mahamadou IB, Nyarko BK (2017) Flood disaster risk mapping in the Lower Mono River Basin in Togo, West Africa. *Int J Disaster Risk Reduct* 23:93–103. <https://doi.org/10.1016/j.ijdr.2017.03.015>
- Papaioannou G, Loukas A, Vasilades L, Aronica GT (2016) Flood inundation mapping sensitivity to riverine spatial resolution and modelling approach. *Nat Hazards* 83:117–132. <https://doi.org/10.1007/s11069-016-2382-1>
- Paper D (2020) Hands-on scikit-learn for machine learning applications. Apress, Berkeley
- Parzen E (1962) On estimation of a probability density function and mode. *Ann Math Stat* 33:1065–1076. <https://doi.org/10.1214/aoms/1177704472>
- Pham BT, Luu C, Van PT et al (2021) Flood risk assessment using hybrid artificial intelligence models integrated with multi-criteria decision analysis in Quang Nam Province. *Vietnam J Hydrol* 592:125815. <https://doi.org/10.1016/j.jhydrol.2020.125815>
- Pollard JA, Spencer T, Jude S (2018) Big Data Approaches for coastal flood risk assessment and emergency response. *Wires Clim Chang*. <https://doi.org/10.1002/wcc.543>
- Ramkar P, Yadav SM (2021) Flood risk index in data-scarce river basins using the AHP and GIS approach. *Nat Hazards* 109:1119–1140. <https://doi.org/10.1007/s11069-021-04871-x>
- Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. *Ann Math Stat* 27:832–837. <https://doi.org/10.1214/aoms/1177728190>
- Sanyal J, Lu XX (2005) Remote sensing and GIS-based flood vulnerability assessment of human settlements: a case study of Gangetic West Bengal, India. *Hydrol Process* 19:3699–3716. <https://doi.org/10.1002/hyp.5852>
- Sayers PB, Hall JW, Meadowcroft IC (2002) Towards risk-based flood hazard management in the UK. *Proc Inst Civ Eng Civ Eng* 150:36–42. <https://doi.org/10.1680/cien.2002.150.5.36>
- Shivaprasad Sharma SV, Roy PS, Chakravarthi V, Srinivasa Rao G (2017) Flood risk assessment using multi-criteria analysis: a case study from Kopili River Basin, Assam, India. *Geom Nat Hazards Risk* 5705:1–15. <https://doi.org/10.1080/19475705.2017.1408705>
- Teng J, Jakeman AJ, Vaze J et al (2017) Flood inundation modelling: a review of methods, recent advances and uncertainty analysis. *Environ Model Softw* 90:201–216. <https://doi.org/10.1016/j.envsoft.2017.01.006>

- Towe R, Dean G, Edwards L et al (2020) Rethinking data-driven decision support in flood risk management for a big data age. *J Flood Risk Manag.* <https://doi.org/10.1111/jfr3.12652>
- Wan X, Yin W, Sun P, Tan B (2017) Risk division assessment of rainstorm-flood disasters based on cloud model. *J Nat Disast* 26:77–83. <https://doi.org/10.13577/j.jnd.2017.0409>
- Wang Z, Lai C, Chen X et al (2015) Flood hazard risk assessment model based on random forest. *J Hydrol* 527:1130–1141. <https://doi.org/10.1016/j.jhydrol.2015.06.008>
- Winsemius HC, Van Beek LPH, Jongman B et al (2013) A framework for global river flood risk assessments. *Hydrol Earth Syst Sci* 17:1871–1892. <https://doi.org/10.5194/hess-17-1871-2013>
- Yadav SM, Mangukiya NK (2021) Semi-arid River Basin flood: causes, damages, and measures. In: Proceedings of the fifth international conference in ocean engineering (ICOE 2019). Lecture Notes in Civil Engineering, pp 201–212
- Yang W, Xu K, Lian J et al (2018) Integrated flood vulnerability assessment approach based on TOPSIS and Shannon entropy methods. *Ecol Indic* 89:269–280. <https://doi.org/10.1016/j.ecolind.2018.02.015>
- Zehra S, Afsar S (2016) Flood hazard mapping of lower indus basin using multi-criteria analysis. *J Geosci Environ Prot* 04:54–62. <https://doi.org/10.4236/gep.2016.44008>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.