Science of the Total Environment 1002 (2025) 180592

Contents lists available at ScienceDirect

Science of the Total Environment

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

Mapping flood susceptibility using Random Forest exploiting satellite observations and geomorphic features

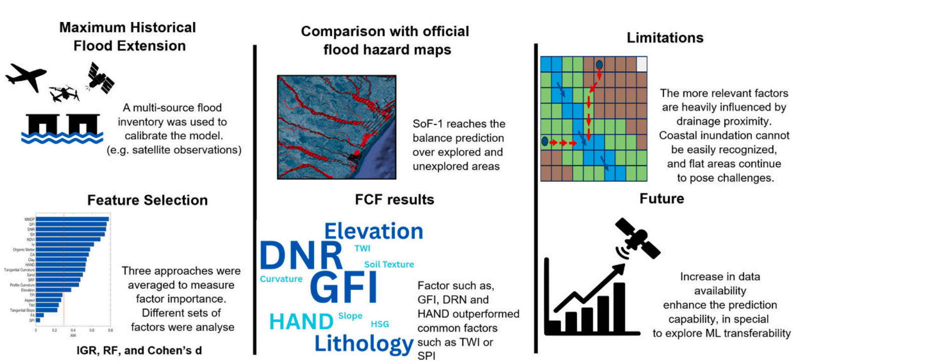
Jorge Saavedra Navarro [a,*], Ruodan Zhuang [a], Cinzia Albertini [a], Salvatore Manfreda [a]

[a] Department of Civil, Construction and Environmental Engineering (DICEA), University of Naples Federico II, Naples, Italy

HIGHLIGHTS

- A multi-source flood inventory was used for susceptibility modelling.
- AMI was introduced to evaluate the relative information of flood conditioning factors.
- A maximum of 10 predictors was enough for a good flood representation.
- GFI reduces the overestimation in flooded rivers.
- The official flood hazard maps were utilized to assess the generalisation abilities.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Geomorphic Flood Index
Flood conditioning factors
Flood susceptibility
Random Forest
Satellite observations

ABSTRACT

Flood events are among the most destructive natural hazards, requiring comprehensive risk management strategies to mitigate their impact on society and the environment. This study uses the potential of the Random Forest (RF) model to assess the flood susceptibility in Italy, evaluating 26 potential flood conditioning factors (FCF). A holistic strategy called Average Merit of Information (AMI) was employed to maximize the information contained within FCFs. At the same time, correlation issues were addressed using the Pearson correlation index and the Variance Inflation Factor (VIF). Satellite observations and regional records of historical flood events were adopted to calibrate the model and represent the maximum flood extension. Eleven sets of factors (SoF) were evaluated using a validation set and compared with official flood hazard maps. The RF model trained with SoF-1 (mean maximum daily precipitation (MMDP), the Geomorphic Flood Index (GFI), distance from the nearest river (DNR), elevation, lithology, soil properties, Normalized Difference Vegetation Index (NDVI), and land cover) demonstrated superior generalisation capacity compared to other SoFs. The inclusion of GFI significantly improved prediction accuracy in most unexplored areas, though challenges persist in flat regions and some areas without information. Ultimately, integrating updated satellite-derived information, complementary datasets, and adequate predictors facilitates the accurate identification of flood-prone areas, streamlining computational processes and providing decision-makers with preliminary analysis.

* Corresponding author.
E-mail address: jorgeandres.saavedranavarro@unina.it (J.S. Navarro).

https://doi.org/10.1016/j.scitotenv.2025.180592
Received 19 May 2025; Received in revised form 21 September 2025; Accepted 21 September 2025
Available online 24 September 2025

1. Introduction

Floods are among the most devastating natural disasters, triggering human fatalities, population displacement, and economic losses (Smith, 2015). Between 2000 and 2019 an increment of 23 % in flood occurrences was observed above the previous annual average of 163 flood events, which has led to economic losses of more than 600 billion dollars, accounting for 22 % of the total losses from disasters on a global scale (WMO, 2021).

In this context, flood hazard and susceptibility mapping is a crucial tool for reducing risk and the implementing effective risk management strategies. Hydrodynamic and hydraulic models have long been used to generate flood hazard maps, supporting informed land use planning. These models incorporate physically based equations and detailed terrain definitions to delineate flood-prone areas and estimate flood depths and flow velocities for different return periods (Karim et al., 2023).

However, such models pose significant computational and financial challenges, particularly when applied at large spatial scales. Their applicability is further constrained in ungauged catchments, where hydrological data is scarce (Grimaldi et al., 2013). Moreover, given the rapid evolution of both morphology and climatic conditions, continuous data updates are essential to maintain model accuracy and reliability.

To overcome these limitations, emerging approaches exploiting geomorphic methods, remote sensing, and machine learning (ML) techniques are becoming increasingly popular in flood studies (Kaya and Derin, 2023; Munawar et al., 2022).

Among the available approaches, geomorphic methods use Digital Elevation Models (DEMs) to identify flood-prone areas analysing the landscape's topographic and morphological characteristics. The effectiveness of various morphological variables and composite indices for flood mapping has been assessed in several studies (Magnini et al., 2023; Degiorgis et al., 2012; Manfreda et al., 2015; Nobre et al., 2016). Among them, the Geomorphic Flood Index (GFI, Manfreda et al., 2015; Samela et al., 2017) and Height Above the Nearest Drainage (HAND, Nobre et al., 2016; Aristizabal et al., 2023) have proven to be particularly reliable in this respect and have been successfully applied to large-scale applications in several countries.

In addition, remote sensing observations have made it possible to map natural disasters, including flood events, across multiple spatial and temporal scales (Schumann et al., 2023). Such data offer a significant amount of information that could be used for training existing ML learning algorithms. This provides an alternative approach to accurately representing flood hazards and susceptibility (Balestra et al., 2022; El-Haddad et al., 2021).

However, the effectiveness of machine learning (ML) in flood susceptibility mapping strongly depends on the quality and quantity of training data. Many existing studies have focused on relatively small areas and have often relied on hydraulic simulations or limited flood observations (Chen et al., 2023; Khosravi et al., 2018, 2019; Shahabi et al., 2020). Instead, the true potential of ML emerges when it is applied to large-scale, multi-source datasets, which allow for more robust and transferable models.

For example, Zhao et al. (2018) demonstrated this by training Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) models using a comprehensive database of historical floods recorded in China over 50 years. Their results confirmed the high capability of ML to model susceptibility across diverse terrains. In another large-scale application, Marchesini et al. (2021) used a different data source, inundation maps derived from modelling, to successfully train a logistic regression model across 23 Italian catchments. Magnini et al. (2023) used three satellite observations and an official flood hazard map as validation sets to assess ML (Decision Tree) and geomorphic (GFI) models. Therefore, integrating diverse data sources (such as historical records) across the broadest possible scope could be a key strategy for mitigating the issue of poor data availability and for building more robust, widely applicable ML flood susceptibility models.

Another crucial step in flood modelling is the selection of the appropriate number and type of predictors. However, there is no clear consensus on which FCF to use or how many (Ado et al., 2022). Kaya and Derin (2023) identified a frequent use of ten factors for flood susceptibility mapping, with a minimum of 5 and a maximum of 21, among 170 research articles analysed without finding a relationship between method and type or amount of FCF. The most used factors include altitude, rainfall, slope, aspect, land use/cover, distance to the river, and NDVI. For example, Marchesini et al. (2021) demonstrated that a reduced number of hydro-morphometric factors is sufficient to construct reliable models, most of which outperformed models built with many variables. Similar patterns have been reported in other studies where some predictors were removed without compromising the accuracy of the models (Esfandiari et al., 2020; Khosravi et al., 2018; Wang et al., 2015).

Tehrany et al. (2019), for example, compared two sets of FCFs, respectively including eight factors (altitude, slope, aspect, curvature, Stream Power Index (SPI), Topographic Wetness Index (TWI), Topographic Roughness Index (TRI), and Sediment Transport Index (STI)) and 12 factors (adding geology, land use/cover, soil, distance from river and road), in two ML models (i.e., Decision Tree and SVM) for flood-prone areas identification. Both models performed the best with eight conditioning factors. Similarly, Esfandiari et al. (2020) tested 30 combinations of four to 12 factors using a Random Forest model and found that five factors (altitude, slope, aspect, distance to river and land use/cover) achieved optimal performance. These findings reinforce the conclusion of Marchesini et al. (2021) that a carefully selected subset of FCFs can achieve robust flood susceptibility predictions.

Last but not least, the model selection varies across different flood modelling studies (Karim et al., 2023; Kaya and Derin, 2023; Munawar et al., 2022). For instance, these include ML, statistical and Multi Criteria Decision Making (MCDM) models. While the choice of the model depends on the study, RF has proven to be a robust option for flood modelling (Chang et al., 2019; Vafakhah et al., 2020; Zhao et al., 2018). RF models were introduced by Breiman (2001), who highlighted their advantages, including the high efficiency when handling large databases, low computational cost, and the ability to process a large number of variables without losing efficiency. RF can also estimate the importance of each variable and generate an internal estimate of the generalisation error (out-of-bag error, Breiman, 1996), which is useful for analysing the relationship between floods and conditioning factors. Furthermore, RF offers high performance with only a few parameters to optimise, enabling variable importance estimation and algorithms for handling missing values. Finally, RF is relatively robust to outliers and noise (Kulkarni and Lowe, 2016), which makes it particularly suitable for handling gaps or inaccuracies in historical flood records.

Based on those points (flood source and flood conditioning factors), in the present research, a flood susceptibility map of the Italian territory is produced by implementing an RF model and exploiting a source merge of satellite data (the Copernicus Emergency Management Service — CEMS) and historical flood records (event reconstruction, aerophotography, others; from Italy's administrative entities, etc.) from 1920 to 2023. In addition, a preliminary and thorough screening of 26 variables spanning hydrologic, topographic, and categorical information was performed to select the most relevant FCF for flood modelling. The impact of the selected factors on the performance of the RF model was then evaluated, and their prediction capabilities were validated using three official flood hazard maps for three different period return. The objectives of this study involve assessing a multi-source flood inventory – that can be updated as the number of observations increases over time – as well as reducing the number of factors used to predict flood-prone areas, while retaining the most significant FCFs.

The results provide a comprehensive overview of flood susceptibility modelling across Italy, representing essential information for municipal planners, environmental agencies, insurance companies, and the general public.

2. Materials and methods

2.1. Study area

The area selected for the current investigation (Fig. 1) encompasses the entire Italian territory, covering a total area of approximately 302,073 km^2 with a coastline of 7456 km. This diverse geography and climate make it an ideal region for studying flood susceptibility modelling. The territory comprises 35.2 % mountains (above 600 m), 23.2 % hills, and 41.6 % flat areas, resulting in varied hydrological conditions. Climatic influences include differences in latitude, Atlantic winds in the north, and hot African winds in the south. The Alpine Chain intensifies precipitation in northern and central regions by condensing Mediterranean vapour.

2.2. Methodology

The methodology was carried out in three main phases, as illustrated in the flowchart (Fig. 2). The first phase consists of database construction, which involves collecting all necessary information, including the calculation and extraction of 26 potential conditioning factors. Additionally, flood inventory maps are compiled from multiple sources.

The second phase focuses on factor analysis and model optimisation. Initially, calibration and validation pools are extracted from the flood inventory. Marginal hazard areas were estimated to delimit non-flooded points. The analysis of factors was carried out in five steps, importance quantification based on AMI, correlation and multicollinearity analysis, creation of Sets of Factors (SoFs), influence factor analysis using a Jackknife-based test and optimal number of predictors. Finally, a straightforward optimisation of the Random Forest (RF) model parameters is conducted.

The third phase comprises a thorough analysis of the model predictions. This is achieved first by using the independent validation dataset (from the flood inventory map) and second, through a comparative analysis of the resulting susceptibility map against official flood hazard maps to assess its accuracy and reliability.

2.2.1. Data collection

2.2.1.1. Flood conditioning factors. The primary task in flood susceptibility modelling is the selection of the most relevant predictors from available FCF. The selection aims to retain the maximum relevance while avoiding redundancy and noise (Khosravi et al., 2019; Ado et al., 2022). However, appropriate predictor selection is inherently linked to the geo-environmental characteristics of the study area (Kaya and Derin, 2023). Based on this, 26 potential flood conditioning factors (21 continuous and 5 categorical factors) were selected from the literature, considering their possible relevance to flood generation. Among them, 16 FCF were derived from the HydroSHEDS v1 data, obtained from the Shuttle Radar Topography Mission (STRM) DEM at a resolution of 3 arc-seconds (~90 m at the equator) (Lehner et al., 2008). These include elevation, tangential slope, aspect, tangential and profile curvature, TWI, SPI, Topographic Position Index (TPI), Surface Roughness Factor (SRF), geomorphology (terrain forms), and geomorphic features, HAND, GFI, potential floodwater depth (hr), Distance to the nearest rivers (DNR), Contributing Area (CA), and Flow Accumulation (FA). In addition, 8 variables related to soil properties and land surface condition were obtained from various databases: the European Soil Database & Soil Properties, which provides the percentage of organic matter, sand, silt, and clay particles of the topsoil at 500 m resolution (Ballabio et al., 2016); Soil texture information from Google Earth Engine

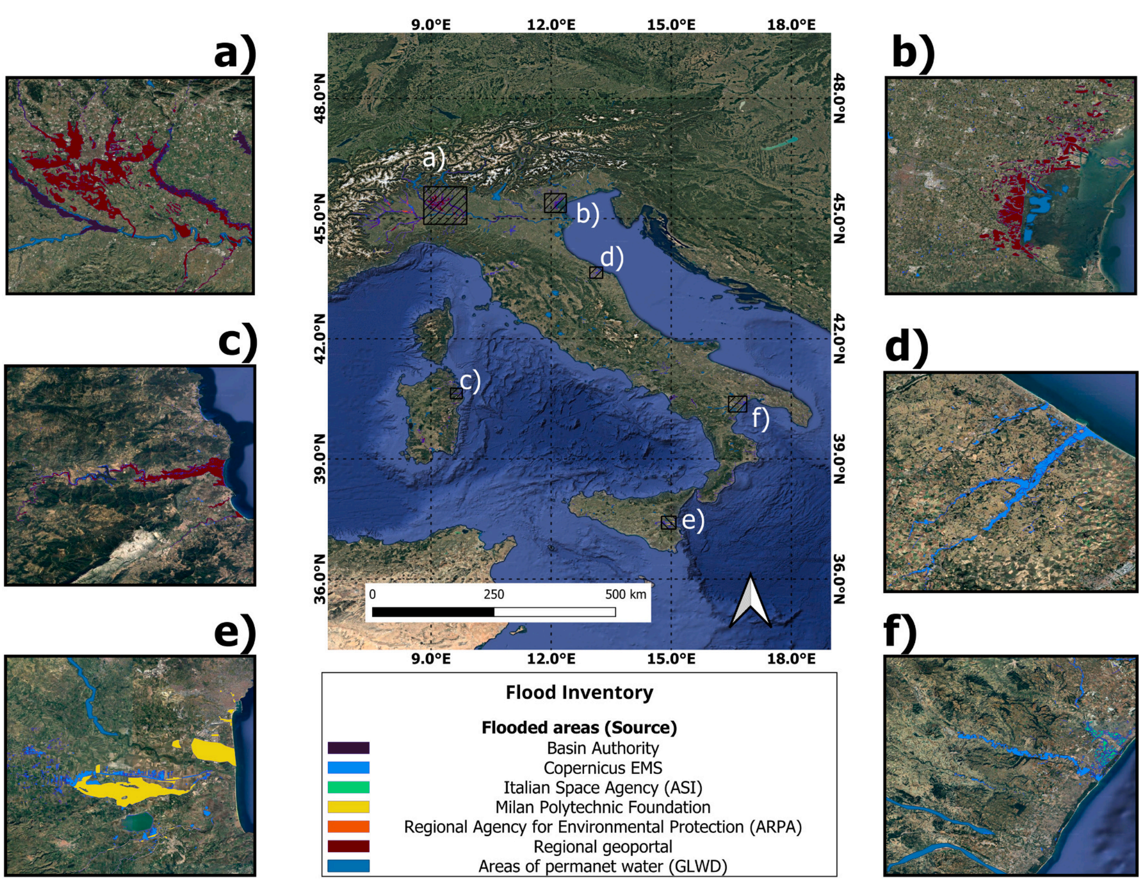


Fig. 1. Overview of the study area and the flood Inventory for the model training. The six insets highlight specific regions of the Italian territory.

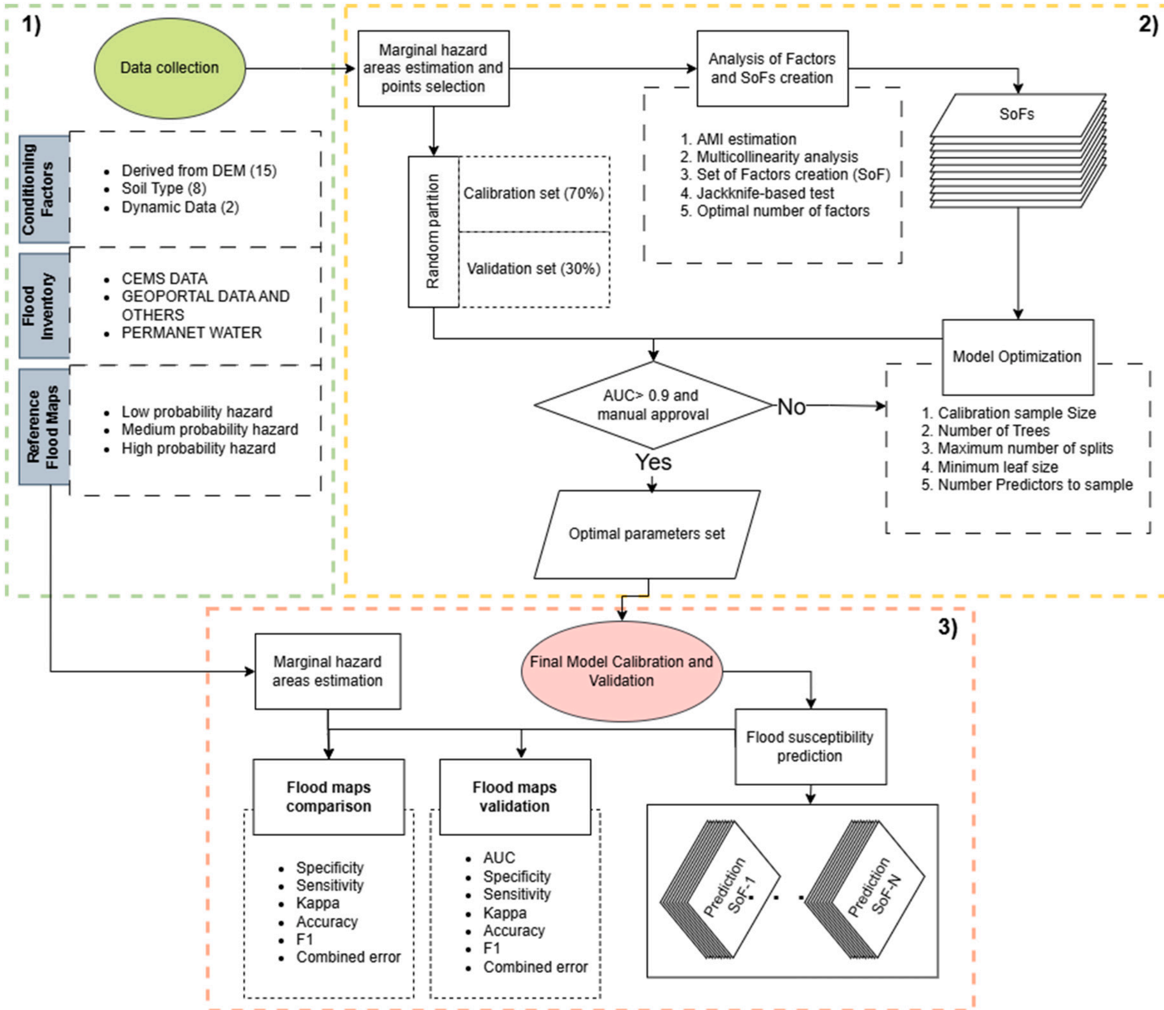


Fig. 2. Flowchart of the methodology, which is subdivided into three phases: 1) database construction – collection of all the necessary information – Inputs (Flood Conditioning Factors) and Outputs (flood maps) (green dashed block); 2) selection of non-flooded points (calibration and validation sets), factors selection (optimum number of predictors), definition of the Set of Factors (SoF), model optimisation, and assessment of model parameters (yellow dashed block). And 3) flood susceptibility mapping, assessment of the SoF predictions (orange dashed block).

"OpenLandMap Soil Texture Class"; Hydrologic Soil Groups (HSG) from The National Aeronautics and Space Administration (NASA) EARTHDATA service at 250 m resolution (Ross et al., 2018); a geolithological map from the WFS service of the "Geoportale Nazionale" by the Italian Ministry of the Environment and Energy Security; and Land cover (LC) from the Environmental Systems Research Institute (ESRI) with a 10 m resolution concerning 2023 (Karra et al., 2021). Finally, two dynamic datasets were used: hourly precipitation data (Muñoz Sabater, 2019) from 1979 to 2022 extracted from the Copernicus Climate Change Service (covering more than 40 years), and the NDVI from 2001 to 2022 obtained from Google Earth Engine MOD13Q1.061 Terra Vegetation Indices 16-Day Global 250 m (covering from the available date 01/01/2001). The difference in time coverage should not influence performance, as shown in Zhao et al. (2018). All factors were projected to the WG84-UTM 32 N coordinate system and resampled to a common resolution of approximately 90 m. A detailed description of each factor is presented in the supplementary material (Table S1).

It is worthwhile to emphasise certain details regarding the processing and selection of FCF. First, the factors from DEM, i.e., GFI, HAND, CA, hr and DNR, were calculated in MATLAB. GFI, HAND and D have demonstrated a high capability in flooded area recognition (Degiorgis et al., 2012; Manfreda et al., 2015; Magnini et al., 2023). Tangential slope, aspect, curvature (both profile and tangential), TPI, and SRF were calculated using Topotoolbox in MATLAB (Schwanghart and Kuhn, 2010). TWI, SPI and geomorphology were obtained from QGIS with GRASS extension (r.watershed and r.geomorphon). TWI is widely used to determine the location and extent of saturated regions due to topography, SPI measures power erosion and intensity of surface runoff, and geomorphology characterises the terrain in ten shapes (flat, peak, ridge, shoulder, spur, slope, hollow, footslope, valley, and pit). These factors play an important role in the hydrological process, controlling the evaporation process, flow velocity and terrain characterisation (Hoang and Liou, 2024). Two key dynamic factors were processed to identify conditions favourable for flood generation. First, we calculated the Mean Maximum Daily Precipitation (MMDP) from hourly precipitation data for each year. The MMDP indicates areas with higher water availability, a primary driver of flood events (Chen et al., 2021). Second, we determined the mean minimum Normalised Difference Vegetation Index (NDVI) for the period 2001–2022. Low NDVI values, representing poor vegetation density (Zhao et al., 2018; Khosravi et al., 2018, 2019).

2.2.1.2. Flood inventory. The production of an updated flood inventory is crucial in the assessment of natural hazards and in the prevention of flood events (Pradhan et al., 2016; Tien Bui and Hoang, 2017). In this study, a multi-source flood inventory was used, considering the criterion that "*flooded areas must correspond to a specific event*" (day, month, and year, at least month and year), assuming the same importance for each flood event. The flood inventory was used to determine the maximum flood extents over time and to calibrate and validate the RF model.

The flood inventory used in this study comprised a total of 280 flood records from various sources, spanning from 1926 to 2023 (Fig. 1). In particular, 33 events were obtained from the Copernicus EMS (https://emergency.copernicus.eu/), 24 events from Basin Authorities (Campania, Calabria, and Lombardia), three events from the Italian Space Agency, 106 from the Milan Polytechnic Foundation, 107 from regional geoportals (Liguria, Lombardia, Piemonte, Sardegna, Trentino-Alto Adige, and Veneto Regions), and 7 from the Regional Agency for Environmental Protection (Piemonte). Finally, permanent flooded areas extracted from the Global Lakes and Wetlands Database (Lehner and Döll, 2004) were included. Concerning CEMS data, each event was analysed to ensure the extraction of flooded areas, removing information such as landslide data and inundation from dike breaks.

2.2.1.3. Reference flood hazard maps. The reference flood maps comprised a merger of different hydrodynamic simulations, developed by the Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA), Italy, as reported in the 2020 document. The maps were created in compliance with Article 6 of the Flood Directive 2007/60/EC, which mandates the preparation of flood hazard and risk maps at the basin or management unit scale (Smith, 2015).

The maps are provided in shapefile format at the national scale, detailing flood extents under three probability scenarios (https://idrogeo.isprambiente.it/app/page/open-data): High Probability Hazard (HPH, return period between 20 and 50 years); Medium Probability Hazard (MPH, return period between 100 and 200 years); and Low Probability Hazard (LPH, return period higher than 200 years). These maps contain more information than the validation pool, which only covers the observation areas, and can help enrich the flood conditioning factor analysis, providing a better understanding of their implications over unexplored areas.

To carry out the comparison (see Section 2.2.3), the ISPRA maps were preliminarily rasterised to ~90 m and reprojected to WG84-UTM 32 N.

2.2.2. Random Forest model implementation

RF is a supervised ML model for solving classification and regression problems based on a non-parametric statistical method that uses multiple decision trees combined with the bootstrap aggregating (bagging) approach (Breiman, 1996). A set of predictors and labelled data describing the class to be predicted (e.g., flooded and non-flooded classes) are required for model calibration, together with the optimisation of some model parameters.

The bagging technique consists of randomly selecting different subsets of the predictors to train the model and aggregating results to make the final prediction. This method allows for reducing overfitting and improving generalisation. Additionally, during the training of each decision tree, a random sample of the training data, called out-of-bag (OOB), is excluded and used exclusively to evaluate the performance of that specific tree.

In this study, two types of output were defined from the RF model. The first is a binary output, with values of 1 and 0, describing flooded and non-flooded areas, respectively. The binary map represents flood-prone areas and is obtained by considering all the instances with a probability greater than 0.5. The second one is the probability of an instance being flooded, and it is used to represent the level of flood susceptibility over the study area.

2.2.2.1. Marginal hazard areas delimitation and training points selection. Selecting of flooded and non-flooded points is a crucial step in training the RF model. While the identification of flooded areas is relatively straightforward, there is no clear definition in the literature of a robust and standard method for selecting non-flooded training points. Indeed, non-flooded areas are often identified using a random procedure. For instance, Avand et al. (2022) used a random selection method by extracting points from high-altitude zones representing areas with a low probability of flooding. Conversely, Chen et al. (2023), Choubin et al. (2019), and Tehrany et al. (2019) randomly extracted non-flooded points from areas with no recorded floods, applying no additional criteria. This procedure is important because the absence of flood records in an area does not imply that it will never be affected by flooding in the future.

Therefore, the selection of non-flooded points in the present study has been carefully defined, in the attempt to isolate the portion of the river system where the flooding has been observed, while removing all the remaining portions where observations are unavailable. For this reason, the concept of “marginal hazard area”, first introduced by Degiorgis et al. (2012), has been utilized. These areas (in orange in Fig. 3) represent the portion of the catchment directly connected to the river and where the flood has been observed, but which has not been flooded. This allows the analysis to focus only on the portion of the river basin where a clear distinction can be made between flooded and non-flooded zones (e.g., Manfreda et al., 2015).

In the present work, marginal hazard areas are define as the non-flooded points and have been recognised as such in three ways: 1) areas that are directly drained by the main river network; 2) areas that are outside of the flooded areas; and 3) areas that do not flow through the secondary or third river network (this condition only applies to areas not subject to flooding). Due to the specificity of the adopted dataset, two additional conditions were applied: 4) The pixels nearest to the flooded areas were removed due to the uncertainty associated with satellite-based flood detection. Specifically, this is due to the observed noise layers describing the flood extent, which are affected by waterbody filtering and vegetation masking. In addition, the resampling to a 90-m resolution may result in the loss of relevant flood information, especially along the boundaries of flooded areas. 5) A buffer within the marginal hazard areas around the flooded areas was created to reduce the number of non-flooded points further and avoid the selection of pixels that could be flooded but not recognised. The final non-flooded

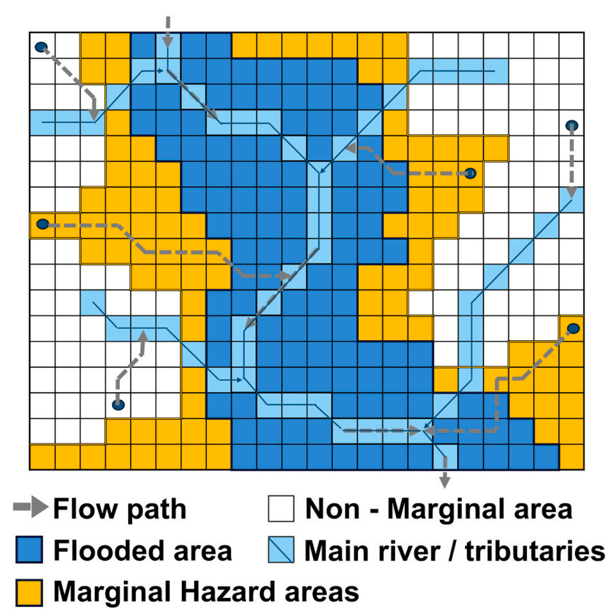


Fig. 3. Definition of marginal hazard areas. The orange colour corresponds with all the cells drained in a flooded river cell. Dark blue colour corresponds to flooded cells, and light blue colour to river cells. Finally, white cells correspond to all the cells drained in a non-flooded river cell (non-marginal hazard area) and grey rows show the drainage direction.

and flooded areas were divided into two pools, for calibration and validation, corresponding to 70 % and 30 % of the total, respectively.

2.2.2.2. Analysis of factors and Set of Factors (SoFs) creation. The selection of conditioning factors is essential in the creation of susceptibility maps using ML models (Wang et al., 2015; Khosravi et al., 2019). Indeed, in the current work, particular attention was given to the maximisation of information and the minimisation of redundancy and high correlations among predictors to implement the RF model using the most influential and descriptive ones.

This study applied three different methods to measure the relative importance of the 21 continuous factors: the Information Gain Ratio (IGR), Cohen's d index, and the Importance Quantification (IQ) based on the ML method (Chen et al., 2023; Zhao et al., 2018; Wang et al., 2015; El-Haddad et al., 2021). In addition, to obtain a more representative measure of the information provided by each factor, the Average Merit of Information (AMI) metric was applied (the name represents a modification of the perspective of Average Merit, where frequently IGR is used). AMI was introduced to address the uncertainty using only a single metric for information quantification. After identifying the most informative factors, Pearson's correlation coefficient and VIF metric were used to remove highly correlated variables (Khosravi et al., 2019; Avand et al., 2022). Separately, Categorical factors were excluded from this preliminary analysis to avoid noise in the results. The statistical methods used in this section, such as Cohen's d index and the Pearson correlation coefficient, are designed for continuous data and may not adequately represent categorical variables. Its importance was not ignored. Instead, Categorical factors were assessed later, directly within the final model, alongside the most relevant continuous factors identified in this initial stage.

As the initial step in our analysis, we used the AMI metric (Eq. 1) to filter and retain only the most significant continuous conditioning factors. The AMI is an intuitive composite index designed to provide a holistic measure of factor importance by integrating three distinct sources of information:

– The Information Gain Ratio (IGR), representing the factor overall contribution to the system (Yu et al., 2024). Higher IGR values are associated with a significant influence of the factor (Khosravi et al., 2018).
– Cohen's d index, quantifying the factor ability to distinguish between flooded and non-flooded area distributions by considering the mean and standard deviation (see Cohen, 1988)
– The Importance Quantification (IQ-RF), measuring the factor predictive power within the RF model. The “AllSplits” option was used, corresponding to the “PredictorSelection” parameter of the TreeBagger function, which selects the predictor that maximises the split criterion gain across all possible splits of all predictors. Default internal parameters were used (100 trees and the whole calibration pool).

To calculate the AMI for each factor, the values from these three-component metrics were first normalised to a standard scale of 0 to 1 using the max-min approach. The three normalised scores were then averaged (see Eq. 1), ensuring each source of information contributed equally to the final assessment. Establishing a clear selection criterion: only factors with an AMI value of 0.3 or greater were retained for further analysis. This threshold was determined after a thorough evaluation, removing factors that provide little to no relevant information. In practice, variables falling below this threshold typically had at least two of the three component metrics with values near zero. For studies considering a smaller initial set of factors, a lower threshold value may be more appropriate. After retaining the most relevant factors, the Pearson's correlation coefficient and VIF were estimated, excluding those with a VIF value greater than 10 (Khosravi et al., 2019) and a Pearson correlation coefficient value greater than 0.7 (Chen et al., 2023).

Different Sets of Factors (SoFs) were created using the most important continuous FCF to evaluate their capability to predict not only in areas with flood records, but also in areas without them, meaning outside of the model information (generalisation capability). This was done while minimising redundancy caused by multicollinearity or irrelevant information. Several of the considered FCF, such as Elevation and HAND, CA and hr, or GFI, hr and HAND, are known to be interdependent (see Table S1 for reference equations), and have been identified in previous studies as independent relevant factor for flood-prone area mapping (e.g Degiorgis et al., 2012).

Once the SoFs were defined, the influence of each categorical factor was evaluated. Using a Jackknife-based test (Choubin et al., 2019). The influence of factor i was measured by the Reduction in Performance based on the Kappa index (RPi, Eq. 2), which involves comparing an RF model calibrated with all factors to an RF model calibrated excluding factor i. At the same time, the optimal number of conditioning factors within each SoF was defined. The conditioning factors were ranked by their influence; the first four most relevant factors (those resulting in the highest reduction in performance) were used as a reference. Then, one by one, the remaining factors were added to determine the highest performance based on the AUC index.

$$\mathrm{AMI_i} = \frac{\left(\mathrm{IQ_{RF_i}}^{\mathrm{normalized}} + \mathrm{Cohen's\ d_i^{normalized}} + \mathrm{IGR_i^{normalized}}\right)}{3} \tag{1}$$

$$RP_\mathrm{i} = (\mathrm{Kappa}_{all} - \mathrm{Kappa}_{i})/(\mathrm{Kappa}_{all})^{*}100\ (\%) \tag{2}$$

2.2.2.3. Calibration, validation, and optimisation of Random Forest algorithm. The RF model was implemented using the TreeBagger function in MATLAB (R2023b Update 7), which does not consider an automatic optimisation function, a process that is essential for reducing generalisation errors (Rahmati et al., 2016; Wang et al., 2015). Six key parameters were selected for optimisation and are presented in Table 1.

We developed a sequential, stepwise procedure to tune the model. This approach combines iterative resampling with manual inspection to ensure robust and reliable results. The entire process is based on a dataset of approximately 1,480,000 points (balanced between flooded and non-flooded instances), which was partitioned into a calibration pool (70 %) and a validation pool (30 %).

The optimisation process started with the identification of an appropriate calibration sample size ranging from 500 to 50,000 points. To obtain a stable and representative performance estimate for each sample size, an iterative resampling method was used. For each sample size, the process of randomly sampling points from the calibration and

Table 1
Internal parameters description of the Random Forest model for the optimisation process.

Parameter	Range/step	Description
Training sample size	500–50,000/500	Number of flooded and non-flooded pixels
NumTrees	10–500/10	Total number of decision trees
MaxNumSplits	1–1000/10	The maximum number of splits (or nodes) that each tree can have
MinLeafSize	1–50/1	The minimum number of observations that a terminal leaf (node) in a tree must have
Num Predictors To Sample	1 – optimum number of factors / 1	The number of variables to consider for splitting at each tree node. By default, this property is the square root of the total number of variables for classification trees, and one-third of the total number of variables for regression trees. A range from 1 to the optimal number of variables is defined.

validation pools was repeated 100 times. For Example, for a sample size of 1000 calibration points, 428 validation points were used, corresponding to a hypothetical split of 70 % calibration and 30 % validation. The model performance was then averaged across these 100 runs. These iterations were then manually inspected until no relevant changes in flood extension were observed and a minimum AUC value of 0.9 was reached, achieving a balance between high performance and computational efficiency. Once the optimal sample size had been established, each of the other model parameters was evaluated individually using the same validation procedure. This stepwise approach guarantees a consistent representation of flooded areas and minimises the risk of misinterpreting the model performance indices.

2.2.3. *Prediction and comparison assessment*

To evaluate the Random Forest model performance in mapping flood susceptibility, a comprehensive set of metrics was used. The Area Under the Curve (AUC), which measures the model overall ability to discriminate between flooded and non-flooded zones. To assess specific aspects of this, Sensitivity was used to quantify the correct identification of truly flood-prone areas, while Specificity measured the correct classification of non-flooded areas, helping to avoid false alarms. The Kappa index provided a robust assessment of agreement between predictions and reality, accounting for agreement by chance, which is vital in imbalanced datasets. While overall Accuracy was considered, the F1, the harmonic mean of precision and sensitivity, offered a more balanced view where both false positives and negatives are critical. Finally, a custom metric called Combined Error (rfp + [1 − rtp]), defined as the sum of the false positive rate and one minus the true positive rate, was introduced to synthesise both overprediction and omission errors into a single, intuitive measure of total misclassification.

Following optimisation, the final model for each SoF was used to generate 10 separate prediction maps. Each map was created using a different, randomly selected calibration set of the optimal size. To enhance stability and reduce prediction variance, these ten maps were averaged to produce a single final flood susceptibility map for each SoF. For validation, a highly robust procedure was implemented. Each of the ten model runs was independently evaluated against 50 different validation sets based on a suite of metrics, including AUC, Specificity, Sensitivity, Kappa, Accuracy, F1-score, and Combined Error. The final, averaged susceptibility maps were classified into five hazard levels (very low to very high) using the quantile method.

The flood-prone area maps (1 and 0 values) – each SoF – were compared with each return period of the official flood hazard maps from ISPRA (see Section 2.2.1.3). Performance metrics were then calculated to evaluate how well the model generalises across different geographic contexts. A spatial comparison of its performance was conducted across various administrative districts in Italy, allowing us to assess its regional variability and reliability.

3. Results and discussion

3.1. *Features selection*

3.1.1. *Importance of factors based on AMI*

The importance of factors was assessed with AMI (introduced in Section 2.2.2.2) over 21 continuous factors. Fig. 4 shows the rank of AMI for the conditioning factors, with MMDP as the most important factor, followed by GFI, DNR, and Silt. Considering the threshold of 0.3, TPI, Aspect, TWI, Tangential Slope, FA, and SPI were removed in this step.

According to the importance quantification by RF, the most relevant variables are MMDP, DNR, hr, and Silt. The IGR emphasises the significance of Silt, NDVI, MMDP, Sand, Clay, and GFI, while Cohen's d identifies HAND and GFI as top-ranking factors. All these variables exhibit normalised values above 0.9, showcasing variations in the factors highlighted by each approach. However, certain factors, such as GFI, NDVI, and DNR, consistently stand out across methods.

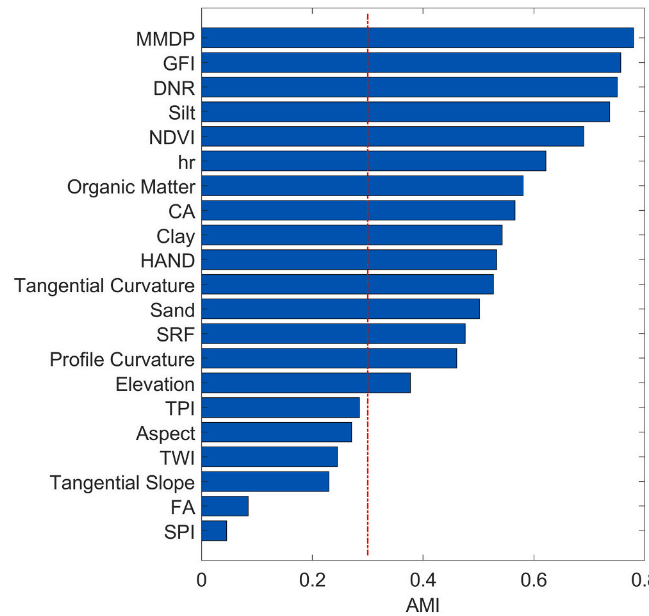


Fig. 4. Relative importance of continuous conditioning factors based on the Average Merit of Information (AMI). The AMI metric was applied exclusively to continuous factors (ranging from −inf to +inf) to determine their predictive power. The red vertical line represents the established threshold for excluding factors with minimal contribution. To ensure a clear assessment, categorical factors were not included in this ranking and are evaluated in the following sections.

MMDP, the most important factor for AMI, varies in its formulation and significance depending on the region and analytical method. For instance, Wang et al. (2015) found that Maximum Three-Day Precipitation (M3DP) ranks as 1st. According to the studies of Avand et al. (2022) and Chen et al. (2023), the Annual Average Rainfall ranks as 7th and 1st, respectively; Hoang and Liou (2024) used Average Total Precipitation during the rainy season, which ranked as 9th. Metrics with shorter temporal windows, such as MMDP, often receive greater importance in RF due to their association with flash flood events. However, longer-term precipitation metrics, such as those used by Chen et al. (2023), can also hold high relevance depending on the modelling context. DNR ranked third overall in this study and consistently appears among the most relevant factors regardless of the evaluation method. This aligns with findings from studies such as Shahabi et al. (2020) and Vafakhah et al. (2020), which indicate that DNR frequently plays a key role in flood susceptibility analysis. Conversely, factors like SPI and FA consistently exhibited low importance across all approaches (AMI $<$ 0.3). Other factors, such as Aspect, TWI, TPI and tangential slope, showed mixed results (Tehrany et al., 2019). Aspect, for instance, is commonly included in flood susceptibility models but does not consistently achieve high importance (Avand et al., 2022), as observed in this study. This variability in ranking factor importance highlights the relevance of information quantification; a multi-perspective approach reduces uncertainties regarding approach selection.

Cohen's d provides an effective and interesting method for assessing factor importance. These results align with the highest performances in the linear binary classification models (Degiorgis et al., 2012), where HAND and DNR achieved high AUC scores and low error rates, while Manfreda et al. (2015) demonstrated that GFI often outperformed HAND and DNR in flood dynamics representation, a result corroborated by other studies (Samela et al., 2017; Manfreda and Samela, 2019). In these steps, 15 continuous factors remain for the next steps. The results of normalising values regarding RF, IGR, and Cohen's d can be found in the supplemental material (Table S2).

3.1.2. SoFs creation

VIF and Pearson's correlation analyses were carried out on the remaining 15 continuous factors to remove redundancy and noise.

Regarding VIF, Silt, Clay, and Sand reached the highest value, i.e., greater than 10 (Fig. S1). Meanwhile, according to Pearson's correlation index, the factors with the highest correlation (greater than 0.7) were Sand-Clay (−0.73), Sand-Silt (0.73), and CA-hr (0.88) (Fig. S2). Based on this evidence, Sand was removed from the general set and with the new configuration, Clay and Silt reached a VIF value lower than 10. Regarding CA-hr correlation, both were evaluated separately in the next step, avoiding the correlation noise.

Eventually, SoFs were created as explained in Section 2.2.2.2, considering the continuous and categorical factors (14 continuous and 5 categorical). Finally, eleven SoFs were proposed and presented in Table 2.

3.1.3. Optimal number of predictors for each Set of Factors (SoF)

Table 3 presents the results of the Jackknife test for determining the optimal number of predictors. The RP, expressed as a percentage for each factor in each SoF, is shown. Positive values indicate a decrease in performance when the factor is removed, while negative values signify an improvement (see Eq. 2).

Among all factors, DNR was identified as one of the most influential in each SoF, with a performance decrease ranging from 4.58 % to 9.36 %, followed by MMDP, which ranges from 2.42 % to 5.38 %. Conversely, removing SRF, tangential and profile curvature, and geomorphology resulted in a performance increase of no more than 2 % in each of the SoFs. The absence of lithology and organic matter results in a slight reduction in performance, while the impact of other factors varied across SoFs with no significant effects.

The Jackknife-based test further validated the critical importance of MMDP and DNR, regardless of the model structure. Factors such as Curvature, Geomorphology, Soil Texture, and HSG showed minimal influence or even improved model performance when excluded, suggesting they can be omitted to optimise the model. Similar trends were noted in Chaubin et al. (2019), where Aspect, HSG, and Curvature had insignificant impacts across models, while DNR, River Density, and Slope were the most impactful factors. Instead, Rahmati et al. (2016) found less impact from slope (percentage and aspect), plan curvature, and TWI. A higher impact was observed on elevation and lithology. Based on the RP, a ranking was created for each SoF, from the highest to the lowest percentage. According to AUC, 6 to 10 factors are necessary to achieve the best average performance in each SoF, with 7 factors being the most frequent. Table 4 summarises the optimum number of conditioning factors and the performance of each SoF. The SoF-3, SoF-6, SoF-10, and SoF-11 outperformed the others.

3.2. Optimisation of the RF algorithm

With each SoF defined, the model optimisation was carried out. The results showed a similar behaviour when changing the set of RF parameters among all the SoFs. Fig. 5 shows an example for SoF-1. The most influential parameters were the calibration sample size and the maximum number of splits. The model achieved a performance greater than or equal to 0.9 with 1000 calibration points and a maximum number of splits of 200. Regarding the number of trees, an increment in performance was observed up to 80 trees, and then slight variations were observed. The number of predictors to sample showed an increment with a maximum between 3 and 5. Finally, increasing the minimum leaf size only decreased the model performance.

Through manual checking, no significant changes were observed in flood-prone areas, from 10,000 points used for model calibration. The following parameters were defined: 100 trees, 200 as the maximum number of splits, 1 as the minimum leaf size, and 4 as the number of predictors to sample. The increment calibration sample size was allowed to capture the variability of the terrain, ensuring a better understanding of the flood events.

3.3. Flood maps performance and comparison

The final flood map for each SoF was estimated by averaging ten flood maps carried out from ten different calibration subsets (Table 5). The validation was performed over 50 different validation subsets (see Section 2.2.3).

The evaluation metrics (AUC, Specificity, Sensitivity, Kappa, Accuracy, F1, and Combined Error) reveal how each SoF contributes to the model performance. SoF-3, SoF-5, SoF-10, and SoF-11 consistently reached the highest values across all metrics, indicating strong and balanced performance in identifying flood-prone and non-flood-prone areas.

The best-performing SoFs exhibit a shared trait: they incorporate combinations such as hr or CA and HAND or GFI, indicating that these sets of factors are notably effective. However, some SoFs with simpler combinations such us SoF-6, SoF-7, and SoF-8 showed slightly lower but competitive results. This suggests that even with reduced complexity (e.g., using only GFI, Elevation, or HAND), the models can still perform reasonably well.

Thus, although more intricate SoFs might enhance performance to

Table 2
Eleven sets of conditioning factors (SoF).

	SoF										
	1	2	3	4	5	6	7	8	9	10	11
CA				X	X					X	
hr		X	X								X
HAND			X		X		X		X		
GFI	X							X	X	X	X
Elevation	X	X		X		X					
DNR	X	X	X	X	X	X	X	X	X	X	X
MMDP	X	X	X	X	X	X	X	X	X	X	X
Lithology	X	X	X	X	X	X	X	X	X	X	X
Organic matter	X	X	X	X	X	X	X	X	X	X	X
LC	X	X	X	X	X	X	X	X	X	X	X
NDVI	X	X	X	X	X	X	X	X	X	X	X
Clay	X	X	X	X	X	X	X	X	X	X	X
Silt	X	X	X	X	X	X	X	X	X	X	X
HSG	X	X	X	X	X	X	X	X	X	X	X
Soil texture	X	X	X	X	X	X	X	X	X	X	X
Geomorphology	X	X	X	X	X	X	X	X	X	X	X
Profile curvature	X	X	X	X	X	X	X	X	X	X	X
Tangential curvature	X	X	X	X	X	X	X	X	X	X	X
SRF	X	X	X	X	X	X	X	X	X	X	X

Table 3
Results of the Jackknife-based test, expressed as Reduction Performance (percentage) for each SoF.

	SoF										
	1	2	3	4	5	6	7	8	9	10	11
CA	–	–	–	5.17	6.06	–	–	–	–	6.61	–
hr	–	5.56	6.16	–	–	–	–	–	–	–	6.67
HAND	–	–	6.12	–	7.19	–	6.94	–	3.81	–	–
GFI	3.59	–	–	–	–	–	–	5.93	2.71	6.09	6.14
Elevation	2.42	2.96	–	2.62	–	4.42	–	–	–	–	–
DNR	5.16	4.58	6.04	5.02	6.53	7.25	8.58	9.36	7.53	6.46	6.56
MMDP	3.76	3.06	2.67	3.10	2.56	5.38	4.30	4.20	3.46	2.81	2.42
Lithology	1.21	1.78	1.03	1.37	1.14	1.59	1.51	1.65	1.09	1.14	1.01
Organic matter	1.16	0.76	0.70	0.57	0.86	1.79	1.37	1.41	1.05	0.31	0.71
LC	0.33	0.47	−0.15	0.41	0.28	0.86	0.66	0.67	0.13	0.13	0.23
NDVI	0.33	0.30	−0.14	0.01	−0.02	1.11	0.77	0.80	0.45	0.01	0.00
Clay	−0.06	0.05	−0.21	−0.25	0.07	−0.24	0.50	0.33	0.66	0.15	0.05
Silt	0.11	0.29	0.27	0.20	0.06	0.17	0.52	0.47	0.37	0.13	0.00
HSG	−0.41	0.08	−0.41	0.07	−0.31	−0.27	−0.17	−0.09	−0.14	−0.53	−0.43
Soil texture	−0.32	−0.26	−0.09	−0.24	−0.19	0.12	−0.08	−0.09	0.04	−0.41	−0.09
Geomorphology	−0.74	−0.22	−0.52	−0.32	−0.37	−0.49	−0.45	−0.42	−0.36	−0.60	−0.43
Profile curvature	−1.02	−0.88	−0.79	−1.34	−0.87	−1.48	−0.74	−0.98	−1.08	−0.92	−0.85
Tangential curvature	−0.95	−0.86	−0.73	−0.97	−0.76	−1.19	−0.73	−0.94	−0.80	−0.79	−1.26
SRF	−1.13	−1.52	−1.28	−1.99	−1.21	−1.72	−0.95	−1.16	−0.95	−1.08	−1.34

Table 4
Optimum number of conditioning factors and mean performance values for the eleven Set of Factors (SoFs). In bold, the best values in the corresponding SoF are highlighted.

	SoF										
	1	2	3	4	5	6	7	8	9	10	11
Optimum number of factors	9	7	8	7	7	7	9	9	10	7	6
AUCROC	0.95	0.95	**0.96**	0.95	**0.96**	0.94	0.94	0.94	0.95	**0.96**	**0.96**
Specificity	0.87	0.89	0.90	0.89	**0.91**	0.86	0.88	0.86	0.88	0.90	0.90
Sensitivity	0.88	0.89	**0.89**	0.88	**0.89**	0.88	0.86	0.87	0.87	**0.89**	**0.89**
Kappa	0.76	0.78	**0.80**	0.77	**0.80**	0.74	0.74	0.73	0.76	**0.79**	**0.79**

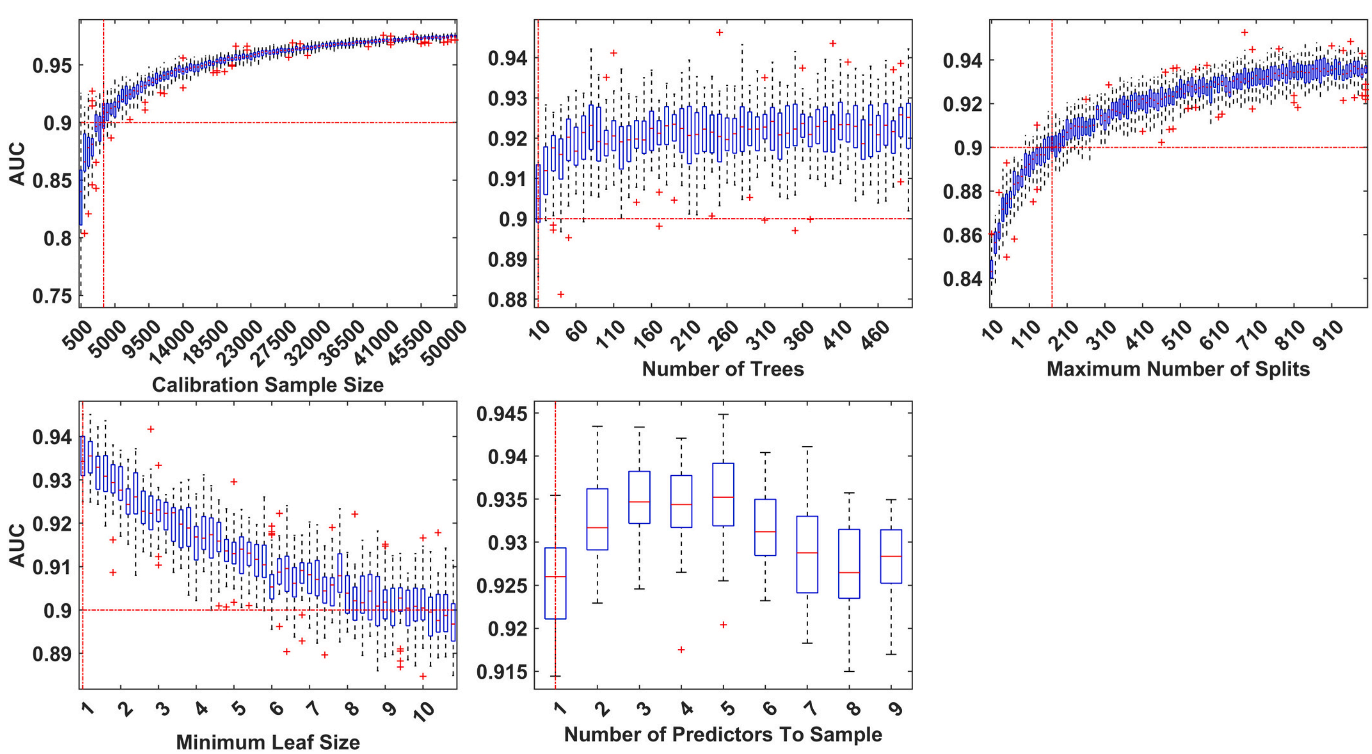


Fig. 5. Results of the Random Forest model parameters optimisation.

Table 5
Mean performance of flood maps over validation subsets for each Set of Factors (SoFs). In bold, the best values in the corresponding SoFs are reported.

SoF	AUC	Specificity	Sensitivity	Kappa	Accuracy	F1	Combined error
1	0.94	0.86	0.88	0.74	0.87	0.87	0.26
2	0.95	0.88	0.88	0.76	0.88	0.88	0.24
3	**0.96**	**0.90**	**0.89**	**0.79**	**0.89**	**0.89**	**0.21**
4	0.95	0.88	0.88	0.76	0.88	0.88	0.24
5	**0.96**	**0.90**	**0.89**	**0.79**	**0.89**	**0.89**	**0.21**
6	0.93	0.85	0.87	0.71	0.86	0.86	0.29
7	0.94	0.87	0.86	0.73	0.87	0.86	0.27
8	0.93	0.85	0.87	0.72	0.86	0.86	0.28
9	0.95	0.88	0.87	0.75	0.88	0.88	0.25
10	0.96	0.90	0.88	0.78	0.89	0.89	0.22
11	0.96	0.90	0.88	0.78	0.89	0.89	0.22

some degree, simpler combinations could present a beneficial compromise by minimising redundancy (for example, GFI and hr can deliver overlapping information) and enhancing interpretability without significantly affecting accuracy.

To provide a robust assessment of each SoF's spatial accuracy and generalisation capabilities, the model predictions were compared against the official ISPRA Flood Hazard Maps. This evaluation was conducted across three distinct hazard levels (LPH, MPH, HPH) using a comprehensive suite of metrics (Specificity, Sensitivity, Kappa, F1-score, etc.), with detailed results shown in Table 6.

The comparison represents a particularly rigorous test of the model performance. The official ISPRA maps are geographically more extensive than the multi-source flood inventory used to train the model. Therefore, this validation step critically evaluates the model ability to accurately extrapolate predictions to areas where no initial flood data was available, offering key insights into its real-world utility and revealing a complex but balanced performance profile for each SoF.

It is possible to observe that individual metrics, such as Kappa and Specificity, reached a low value for each hazard level; this should not be interpreted as an overall model weakness. Other metrics, such as Sensitivity, Accuracy, F1 and Combined Error, provide complementary insights that reveal a more balanced scenario. For instance, Sensitivity increases with hazard level, exceeding 0.90 for HPH in most SoFs, indicating a strong capability to identify high-risk zones correctly. In contrast, Specificity is higher at LPH, confirming reliable detection of safe areas. This inverse trend reflects the natural trade-off in flood modelling, where enhancing true positive detection at high hazard levels may lead to more false positives in lower-risk zones. Meanwhile, F1 (which integrates both false positive and false negatives) maintains consistently high values across all hazard levels, peaking at 0.89 in HPH for SoF-1 and SoF-8, and rarely dropping below 0.82, reinforcing the notion of a balance trade-off between sensitivity and precision across the study area. Accuracy also follows this trend, improving with the hazard level and reaching a maximum of 0.82 in HPH. Finally, the Combined Error is lowest in LPH scenario, suggesting the models are particularly conservative when predicting low-hazard areas.

In general, SoF-1 and SoF-8 reached the highest comparison performance with the more robust behaviour. Both used the GFI in their configuration, but the absence of Elevation in SoF-8 led to a slightly lower performance.

These findings demonstrate that relying on a single validation method, such as testing against an independent dataset, is insufficient for assessing model robustness in complex, large-scale applications. Incorporating a second validation step (specifically, benchmarking predictions against authoritative hazard maps) provides crucial, complementary evidence of a model performance, particularly in areas not covered by the training data. This dual-validation strategy allows for a deeper understanding of the model behaviour across different hazard levels.

3.3.1. Capability in unexplored areas

When a model is trained using data from a calibration pool (in this case, flood event records), the performance on this set can be excellent, but this does not imply good generalisation to unexplored areas (i.e., areas outside the training zone or in regions with different conditions). This may be due to factors such as overfitting, training data bias, or non-representative samples. For example, SoF-3 and SoF-5 outperformed the other SoFs during calibration. However, they failed to demonstrate optimal recognition of flood-prone areas outside the training domain, revealing a limited generalisation capability in the comparison phase.

By contrast, SoF-1, which slightly underperformed SoF-3 and SoF-5 during calibration, outperformed all other SoFs when compared to reference maps. This suggests that SoF-1 is a more balanced and robust model for capturing variability across the study area. It likely incorporates features that better reflect the underlying flood dynamics across diverse regions, whereas other SoFs may be overfit to known conditions, reducing their adaptability to geographic and climatic variations not present in the training set. SoF-1 differs from the others only

Table 6
Quantitative comparison between flood-prone areas (prediction) and hazard maps (observed) over the entire study area for each Set of Factors (SoFs) and hazard level (i.e., low probability hazard — LPH, medium probability hazard — MPH, high probability hazard — HPH). In bold, the best values in the corresponding SoFs are reported.

SoF	Specificity			Sensitivity			Kappa			Accuracy			F1			Combined Error		
	LPH	MPH	HPH	LPH	MPH	HPH	LPH	MPH	HPH	LPH	MPH	HPH	LPH	MPH	HPH	LPH	MPH	HPH
1	**0.63**	**0.52**	**0.36**	0.80	0.86	0.92	**0.34**	**0.34**	**0.30**	**0.77**	**0.80**	**0.82**	**0.86**	**0.88**	**0.89**	**0.57**	**0.62**	**0.73**
2	0.51	0.40	0.26	0.79	0.84	0.90	0.26	0.24	0.19	0.73	0.75	0.77	0.82	0.85	0.86	0.70	0.76	0.83
3	0.52	0.43	0.29	**0.82**	**0.87**	**0.93**	0.33	0.32	0.26	0.74	0.76	0.77	0.82	0.84	0.86	0.66	0.70	0.78
4	0.50	0.40	0.26	0.79	0.84	0.90	0.26	0.24	0.19	0.73	0.75	0.77	0.82	0.84	0.86	0.71	0.76	0.84
5	0.52	0.43	0.29	**0.82**	**0.87**	**0.93**	0.34	0.32	0.26	0.74	0.76	0.76	0.82	0.84	0.85	0.66	0.70	0.79
6	0.58	0.47	0.31	0.78	0.84	0.90	0.27	0.26	0.23	0.75	0.79	0.81	0.84	0.87	0.89	0.63	0.69	0.78
7	0.58	0.47	0.31	0.81	0.87	0.92	0.36	0.34	0.27	0.76	0.78	0.79	0.84	0.86	0.87	0.61	0.66	0.77
8	0.62	0.51	0.35	0.80	0.86	0.92	**0.34**	**0.34**	**0.30**	**0.77**	**0.80**	**0.82**	0.85	0.88	0.89	0.58	0.63	0.73
9	0.57	0.47	0.31	0.82	0.87	0.92	0.36	0.35	0.28	0.76	0.78	0.79	0.84	0.86	0.87	0.61	0.66	0.76
10	0.54	0.43	0.30	0.82	0.87	0.92	0.34	0.32	0.26	0.75	0.76	0.77	0.83	0.85	0.86	0.65	0.70	0.78
11	0.53	0.42	0.29	0.82	0.87	0.92	0.34	0.31	0.25	0.74	0.75	0.76	0.83	0.84	0.85	0.66	0.71	0.79

in the inclusion of GFI or Elevation, indicating that these factors may better represent flood behaviour in both explored and unexplored areas.

The comparison maps assess the model performance across different geographic zones and configurations (SoF-1, SoF-5, SoF-7) at various probability thresholds (LPH, MPH, HPH), revealing several key insights (Table 7). Overall, the model demonstrates high sensitivity and accuracy, indicating that it is effective in correctly identifying positive cases. However, this comes at the cost of low specificity, indicating a high rate of false positives (specifically, Fig. 6a and b). Performance varies significantly across different geographic zones, with the Northern Apennine District generally showing the most balanced results. The different SoF configurations exhibit subtle yet discernible differences in performance, with SoF-1 often achieving the highest Kappa and F1 scores, indicating a better balance between sensitivity and specificity compared to SoF-5 and SoF-7.

Across all zones and configurations, the model consistently exhibits high Sensitivity values. Generally, above 0.80, and often exceeding 0.90, especially for MPH and HPH. This indicates the model is highly successful at identifying true positive cases. Moderate to High Accuracy and F1 Scores, suggesting a good balance between precision and recall.

Low specificity is one of the model primary weaknesses. Specificity values are frequently below 0.60 and in some cases as low as 0.23. This means the model struggles to accurately identify true negatives, leading to a high number of false alarms. Also, Kappa values are generally low (often below 0.40). This is a direct consequence of the high number of false positives indicated by the low specificity.

There is considerable variation in model performance across the different districts. Sicily (1 % of data in the flood inventory) and Sardinia (7 %) Districts consistently show some of the highest sensitivity and F1 values, particularly for SoF-1 and SoF-7. However, their specificity is moderate. The Northern Apennine (3 %) District often presents the most balanced performance, with some of the highest specificity values across all SoF levels, while still maintaining respectable sensitivity and F1 values. This suggests the model is most reliable in this region. Oriental Alps (31 %) and Po River (21 %) Districts tend to show the lowest performance, particularly in terms of Kappa and F1 scores, indicating a less effective model in these areas.

In summary, SoF-1 generally appears to be the best-performing configuration, often reaching the highest Kappa and F1 values, especially in the Sicily and Northern Apennine districts. It strikes a better balance between the competing demands of sensitivity and specificity. SoF-5 tends to have the lowest performance metrics overall, particularly in terms of Kappa and specificity. SoF-7 shows performance that is generally between that of SoF-1 and SoF-5. For a visual overview, refer to Fig. 6 and the supplemental material (Figs. S3–S12) for the comparison of all SoFs.

3.4. Flood susceptibility maps

Based on the comparison, a susceptibility map is provided exclusively for SoF-1, which represents the most similar flood-prone area map. Fig. 7 displays the flood susceptibility map based on the quantile method, and Fig. 8 further analyses these susceptibility levels across each district as percentages.

A higher flood risk is observed in the Po River District, where over 50 % of the area falls into the high and very high susceptibility categories. This is consistent with the significant number of flood events historically recorded in this district. Similarly, the Oriental Alps District also shows a high concentration of areas with elevated flood risk, with over 50 % of its territory categorised in high and very high susceptibility areas. These results align with the topographic and hydrological complexities of these regions, where steep slopes and flat floodplains coexist, amplifying flood risk.

In contrast, districts like the Northern Apennine, Central Apennine, and Sicily exhibit more than 50 % of their areas categorised in the very low and low susceptibility classes. This lower risk level suggests a more

Table 7
Quantitative comparison between flood-prone areas (prediction of SoF-1, SoF-5, SoF-7) and hazard maps (observed) over different districts in Italy for each hazard level (i.e., low probability hazard — LPH, medium probability hazard — MPH, high probability hazard — HPH).

	Zone	Specificity			Sensitivity			Kappa			Accuracy			F1			Combined Error		
		LPH	MPH	HPH	LPH	MPH	HPH	LPH	MPH	HPH	LPH	MPH	HPH	LPH	MPH	HPH	LPH	MPH	HPH
SoF-1	Sicily District	0.57	0.53	0.43	0.93	0.94	0.95	0.51	0.50	0.44	0.87	0.87	0.87	0.92	0.93	0.92	0.50	0.53	0.62
	Sardinia District	0.57	0.40	0.35	0.89	0.94	0.95	0.37	0.33	0.32	0.86	0.88	0.89	0.92	0.93	0.94	0.53	0.67	0.70
	Meridionale Apennine District	0.48	0.46	0.38	0.89	0.89	0.91	0.30	0.29	0.28	0.85	0.85	0.86	0.91	0.92	0.92	0.63	0.65	0.70
	Central Apennine District	0.49	0.44	0.35	0.88	0.93	0.93	0.32	0.37	0.29	0.83	0.87	0.87	0.90	0.93	0.93	0.63	0.64	0.72
	Oriental Alps District	0.70	0.52	0.47	0.69	0.81	0.83	0.18	0.20	0.21	0.69	0.78	0.79	0.80	0.87	0.87	0.60	0.67	0.71
	Northern Apennine District	0.87	0.69	0.47	0.73	0.84	0.91	0.26	0.33	0.34	0.74	0.82	0.87	0.84	0.90	0.93	0.40	0.47	0.61
	Po River District	0.67	0.54	0.33	0.70	0.78	0.94	0.33	0.32	0.31	0.69	0.71	0.75	0.76	0.79	0.84	0.64	0.68	0.74
SoF-5	Sicily District	0.48	0.45	0.38	0.93	0.95	0.96	0.45	0.45	0.42	0.84	0.84	0.84	0.90	0.91	0.91	0.59	0.60	0.66
	Sardinia District	0.40	0.26	0.23	0.91	0.94	0.95	0.33	0.25	0.23	0.81	0.81	0.82	0.89	0.89	0.89	0.70	0.79	0.81
	Meridionale Apennine District	0.34	0.32	0.29	0.90	0.91	0.93	0.26	0.26	0.26	0.79	0.79	0.80	0.87	0.87	0.88	0.77	0.77	0.78
	Central Apennine District	0.38	0.31	0.23	0.89	0.93	0.93	0.30	0.29	0.20	0.79	0.81	0.79	0.87	0.89	0.88	0.72	0.75	0.84
	Oriental Alps District	0.49	0.37	0.35	0.69	0.82	0.83	0.15	0.18	0.17	0.65	0.73	0.74	0.76	0.83	0.84	0.81	0.81	0.82
	Northern Apennine District	0.64	0.44	0.28	0.78	0.86	0.93	0.38	0.32	0.25	0.75	0.76	0.77	0.82	0.85	0.86	0.58	0.70	0.79
	Po River District	0.61	0.50	0.29	0.72	0.80	0.94	0.33	0.32	0.26	0.68	0.69	0.70	0.73	0.76	0.80	0.67	0.69	0.77
SoF-7	Sicily District	0.53	0.48	0.39	0.94	0.95	0.96	0.52	0.49	0.43	0.86	0.86	0.85	0.91	0.91	0.91	0.53	0.57	0.64
	Sardinia District	0.55	0.37	0.32	0.90	0.94	0.95	0.41	0.34	0.32	0.86	0.87	0.87	0.92	0.93	0.93	0.54	0.69	0.73
	Meridionale Apennine District	0.45	0.43	0.36	0.89	0.90	0.92	0.31	0.31	0.29	0.84	0.84	0.85	0.91	0.91	0.91	0.66	0.67	0.72
	Central Apennine District	0.41	0.35	0.26	0.89	0.94	0.94	0.31	0.34	0.24	0.81	0.83	0.82	0.88	0.90	0.90	0.70	0.71	0.80
	Oriental Alps District	0.62	0.49	0.45	0.69	0.81	0.83	0.17	0.21	0.21	0.68	0.78	0.78	0.80	0.87	0.87	0.69	0.70	0.72
	Northern Apennine District	0.81	0.60	0.39	0.76	0.86	0.92	0.37	0.38	0.34	0.76	0.82	0.84	0.85	0.89	0.91	0.43	0.54	0.69
	Po River District	0.60	0.48	0.28	0.71	0.80	0.94	0.31	0.29	0.25	0.67	0.67	0.68	0.72	0.75	0.78	0.69	0.72	0.78

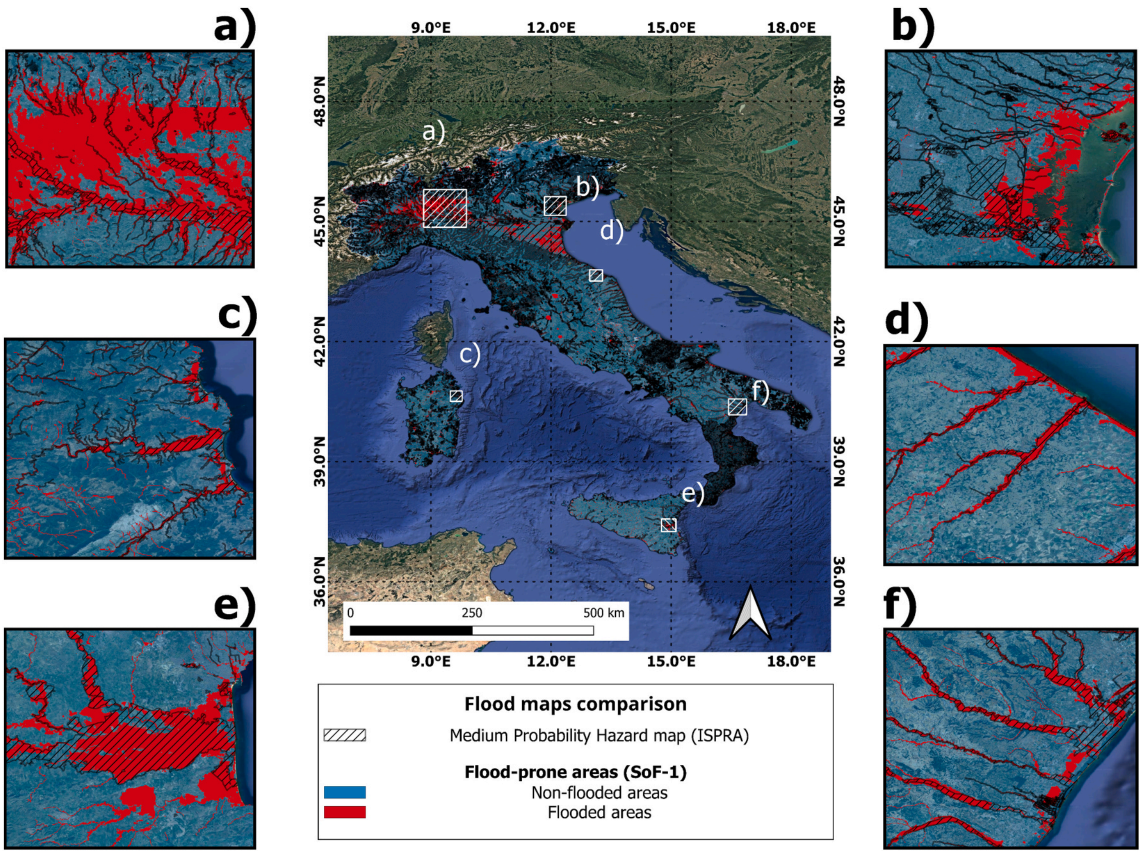


Fig. 6. Flood maps comparison. Flood prediction from SoF-1 (probability equal or greater than 0.5) and Medium Probability Hazard map (MPH) from ISPRA. Featuring six insets that provide detailed views and zoom-in perspectives of example areas as identified in Fig. 1.

stable hydrological environment or more effective flood management systems. In the case of the Northern Apennine District, its homogeneous terrain and fewer recorded flood events may contribute to this outcome. Similarly, Sicily, despite its exposure to coastal hazards, benefits from localised geographic features that limit extensive flood risk. These findings demonstrate an evident regional variability in flood susceptibility across Italy, highlighting the need for district-specific flood risk management strategies.

3.5. Limitations and future directions

While this study presents an in-depth, large-scale analysis of flood conditioning factors with robust results, it is essential to acknowledge certain limitations that, in turn, open new avenues for future research.

A fundamental limitation of ML models is that they process data as vectors without inherently understanding the spatial relationships between them (e.g., the effect of rainfall on lowland areas). The results of this study highlight this issue: indices such as GFI, HAND and DNR are superior to traditional factors (TWI, SPI and slope) due to their formulation (see Table S1). GFI and HAND do not represent a local process, as TWI and SPI do by using tangential slope as a local factor. Instead, they encode the direct relationship of each pixel to the drainage network, which is the primary conduit for fluvial floods. In particular, GFI integrates the influence of the upstream catchment via the river's flow accumulation (HR) with the pixel's vertical position relative to the river (HAND). This provides a simpler yet more powerful representation of the flood process for the model to learn from. GFI even outperforms HAND in this study, as well as in those of Manfreda et al. (2015) and Yilmaz et al. (2025).

This inspires the main future direction — the design or modification of geomorphic and hydrological indices (for instance, GFI in coastal areas by Albertini et al., 2021 or modified HAND by Aristizabal et al., 2023). The goal is to create simple predictors that translate complex physical processes (such as runoff potential or the influence of distant precipitation) into a format that machine learning models can effectively use, thereby enhancing their predictive power without the need for computationally expensive hydraulic simulations.

Other important limitations include:

- Source Data Uncertainty: Uncertainties are present in both the flood inventory, where all events are treated equally, and in dynamic predictors (e.g., NDVI, precipitation), as these reflect general behaviour (Khosravi et al., 2018; Zhao et al., 2018) rather than the specific conditions of each flood event. What may not be a real representation for model training. The same phenomenon occurred with categorical factors, such as HSG, LU, or soil properties, which also change over time but are usually used without considering these changes.
- Drainage Network Dependency and DEM Quality: The high performance of GFI and HAND implies a strong dependency on the quality of the drainage network extracted from the DEM. This presents two challenges:
  - The model is unable to accurately predict non-fluvial flooding, and the backwater effects were not considered (e.g., coastal floods, see Fig. 6f).
  - Prediction accuracy decreases in flat areas (such as the Po River valley), where drainage density is high. Using FABDEM or other conditioned DEM (Neal et al., 2023) and the appropriate identification of the drainage network may improve the predictions.
- Terrain Representation: The 90 m DEM used, while hydrologically conditioned, does not include flood defence structures, which can

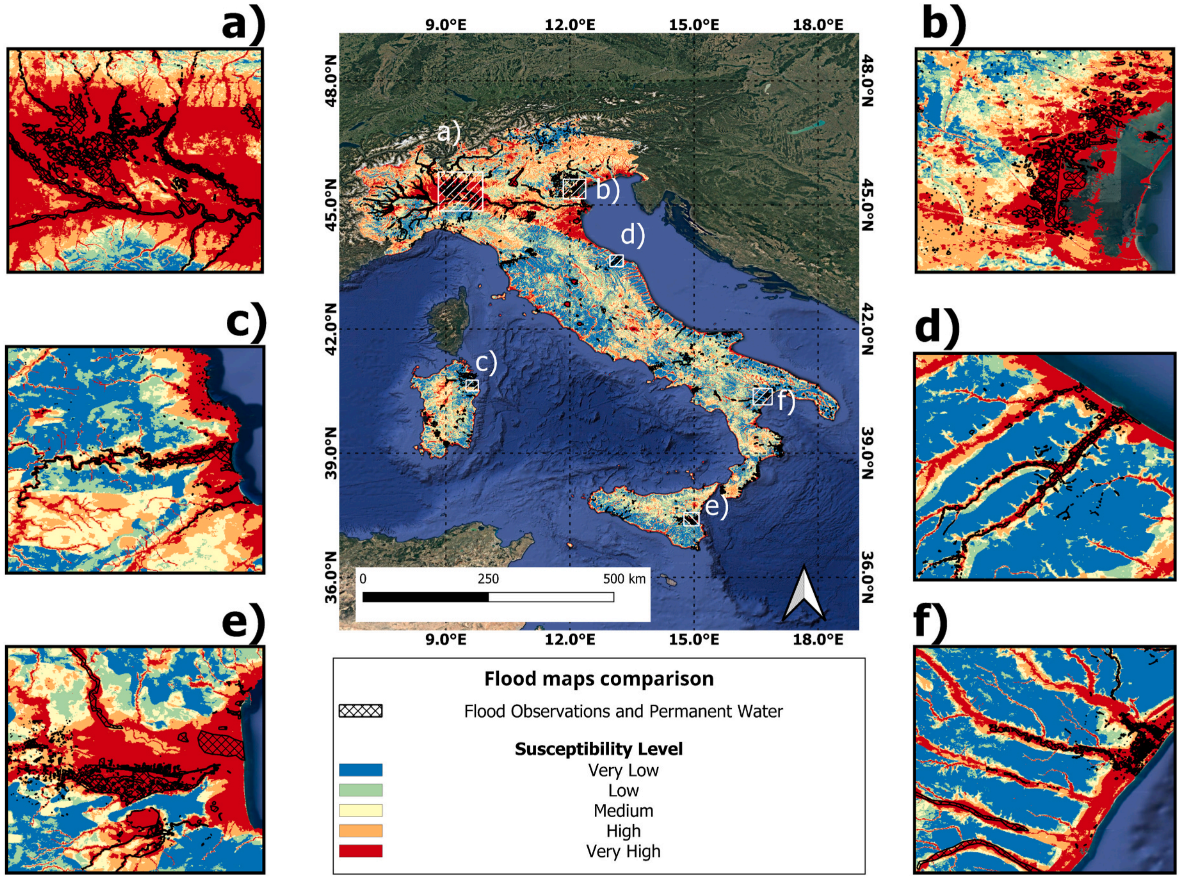


Fig. 7. Flood susceptibility map from SoF-1 of the entire Italian territory, featuring six insets that provide detailed views and zoom-in perspectives of example areas as identified in Fig. 1.

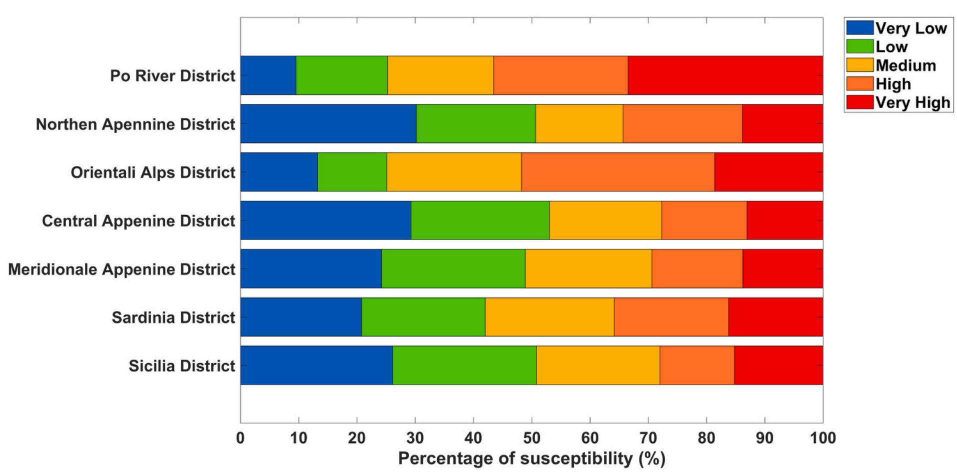


Fig. 8. Percentage of susceptibility levels over the main River Basin Districts in Italy.

lead to overly conservative and unrepresentative predictions in protected areas (Avand et al., 2022).

4. Conclusions

This study demonstrates the effectiveness of an optimised Random Forest (RF) model for mapping flood susceptibility on large-scale. A key finding is that a multi-source flood inventory (1926–2023) is a viable calibration source and that its quality is expected to improve as more data becomes available, including satellite, drone, and camera observations. The use of an external data source, such as ISPRA's hazard maps, was fundamental for conducting an objective performance analysis and evaluating of the model's behaviour in districts with limited historical data, such as Sicily and Sardinia. However, there remains considerable scope for improvement.

Factor relevance analysis reveals that geomorphic factors, such as DNR, GFI, and HAND, are more relevant than traditional factors such as TWI, SPI, and Tangential Slope. This highlights the importance of using factors specifically designed for flood analysis. Additionally, the AMI metric indicates that more than one method should be used to select the final predictors effectively. In this context, we identified an optimal range of six to ten factors that achieves maximum performance across the different SoFs. However, it was observed that all SoFs tend to overestimate flood-prone areas, with SoF-1 exhibiting the least overestimation.

The selection of FCF for model training requires thorough analysis and evolution in its application, as discussed in the limitations section. The results suggest focusing on the creation of stronger indices that can extract more information about the hydrological process, such as the GFI, which is still the subject of further investigations and improvements.

Future studies should focus on explainable AI and emerging remote sensing observations. Data-driven models such as this one may prove particularly valuable in light of the rapid changes observed in the local dynamics of hydrological extremes in recent years. Further efforts should also be directed towards assessing the transferability of machine learning (ML) models to improve predictions in ungauged areas. Continuously refining susceptibility mapping is essential for informed decision-making, supporting proactive measures to mitigate flood risks, strengthen environmental resilience and safeguard community well-being.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2025.180592.

CRediT authorship contribution statement

Jorge Saavedra Navarro: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ruodan Zhuang:** Writing – review & editing, Visualization, Software, Methodology, Investigation. **Cinzia Albertini:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology. **Salvatore Manfreda:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing 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.

Acknowledgments

This study was carried out within the RETURN Extended Partnership funded by the European Union Next-Generation EU (National Recovery and Resilience Plan—NRRP, Mission 4, Component 2, Investment 1.3-D.D. 1243 2/8/2022, PE0000005), the ESA MOST Dragon VI programme (DTE-CLIMATE: Digital Twin Earth Approach for Monitoring and Modelling Climate Change in Water, Energy and Carbon Cycles in Eurasia), and the CHANCES project (An Integrated Modelling Approach for Mitigating Climate Change Effects through Enhanced Weathering in Southern Italy — CUP E53D23021850001).

Data availability

Data will be made available on request.

References

Ado, M., Amitab, K., Maji, A.K., Jasińska, E., Gono, R., Leonowicz, Z., Jasiński, M., 2022. Landslide susceptibility mapping using machine learning: a literature survey. Remote Sens. 14 (13), 3029. https://doi.org/10.3390/rs14133029.

Albertini, C., Miglino, D., Iacobellis, V., De Paola, F., Manfreda, S., 2021. Delineation of flood-prone areas in cliffed coastal regions through a procedure based on the geomorphic flood index. J. Flood Risk Manag. 15 (1), e12766. https://doi.org/10.1111/jfr3.12766.

Aristizabal, F., Salas, F., Petrochenkov, G., Grout, T., Avant, B., Bates, B., Spies, R., Chadwick, N., Wills, Z., Judge, J., 2023. Extending height above nearest drainage to model multiple fluvial sources in flood inundation mapping applications for the U.S. National Water Model. Water Resour. Res. 59 (5). https://doi.org/10.1029/2022WR032039 e2022WR032039.

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.

Balestra, F., Del Vecchio, M., Pirone, D., Pedone, M.A., Spina, D., Manfreda, S., Menduni, G., Bignami, D.F., 2022. Flood Susceptibility Mapping Using a Deep Neural Network Model: The Case Study of Southern Italy, 5, p. 36. https://doi.org/10.3390/environsciproc202202103.

Ballabio, C., Panagos, P., Montanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110–123.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. https://doi.org/10.1007/BF00058655.

Breiman, L., 2001. Random Forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Chang, K.-T., Merghadi, A., Yunus, A.P., Pham, B.T., Dou, J., 2019. Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques. Sci. Rep. 9 (1), 12296. https://doi.org/10.1038/s41598-019-48773-2.

Chen, J., Huang, G., Chen, W., 2021. Towards better flood risk management: assessing flood risk and investigating the potential mechanism based on machine learning models. J. Environ. Manag. 293, 112810. https://doi.org/10.1016/j.jenvman.2021.112810.

Chen, Y., Zhang, X., Yang, K., Zeng, S., Hong, A., 2023. Modeling rules of regional flash flood susceptibility prediction using different machine learning models. Front. Earth Sci. 11, 1117004. https://doi.org/10.3389/feart.2023.1117004.

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., Mosavi, A., 2019. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Sci. Total Environ. 651, 2087–2096. https://doi.org/10.1016/j.scitotenv.2018.10.064.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. L. Erlbaum Associates.

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. 470–471, 302–315. https://doi.org/10.1016/j.jhydrol.2012.09.006.

El-Haddad, B.A., Youssef, A.M., Pourghasemi, H.R., Pradhan, B., El-Shater, A.-H., El-Khashab, M.H., 2021. Flood susceptibility prediction using four machine learning techniques and comparison of their performance at Wadi Qena Basin, Egypt. Nat. Hazards 105 (1), 83–114. https://doi.org/10.1007/s11069-020-04296-y.

Esfandiari, M., Jabari, S., McGrath, H., Coleman, D., 2020. Flood mapping using random forest and identifying the essential conditioning factors; a case study in Fredericton, New Brunswick, Canada. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-3-2020, 2020 editionXXIV ISPRS Congress, Commission III (Volume V-3-2020), pp. 609–615. https://doi.org/10.5194/isprs-annals-V-3-2020-609-2020.

Grimaldi, S., Petroselli, A., Arcangeletti, E., Nardi, F., 2013. Flood mapping in ungauged basins using fully continuous hydrologic–hydraulic modeling. J. Hydrol. 487, 39–47. https://doi.org/10.1016/j.jhydrol.2013.02.023.

Hoang, D.-V., Liou, Y.-A., 2024. Assessing the influence of human activities on flash flood susceptibility in mountainous regions of Vietnam. Ecol. Indic. 158, 111417. https://doi.org/10.1016/j.ecolind.2023.111417.

Karim, F., Armin, M.A., Ahmedt-Aristizabal, D., Tychsen-Smith, L., Petersson, L., 2023. A review of hydrodynamic and machine learning approaches for flood inundation modeling. Water 15 (3), 566. https://doi.org/10.3390/w15030566.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J.C., Mathis, M., Brumby, S.P., 2021. Global land use/land cover with Sentinel 2 and deep learning. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, 11–16 July 2021, pp. 4704–4707. https://doi.org/10.1109/IGARSS47720.2021.9553499.

Kaya, C.M., Derin, L., 2023. Parameters and methods used in flood susceptibility mapping: a review. J. Water Clim. Change 14 (6), 1935–1960. https://doi.org/10.2166/wcc.2023.035.

Khosravi, K., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., Tien Bui, D., 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.

Khosravi, K., Shahabi, H., Pham, B.T., Adamowski, J., Shirzadi, A., Pradhan, B., Dou, J., Ly, H.-B., Gróf, G., Ho, H.L., Hong, H., Chapi, K., Prakash, I., 2019. A comparative assessment of flood susceptibility modeling using Multi-Criteria Decision-making Analysis and Machine Learning Methods. J. Hydrol. 573, 311–323. https://doi.org/10.1016/j.jhydrol.2019.03.073.

Kulkarni, A.D., Lowe, B., 2016. Random forest algorithm for land cover classification. Int. J. Recent Innov. Trends Comput. Commun. 4 (3).

Lehner, B., Döll, P., 2004. Development and validation of a global database of lakes, reservoirs and wetlands. J. Hydrol. 296 (1), 1–22. https://doi.org/10.1016/j.jhydrol.2004.03.028.

Lehner, B., Verdin, K., Jarvis, A., 2008. New global hydrography derived from spaceborne elevation data. EOS Trans. Am. Geophys. Union 89 (10), 93–94. https://doi.org/10.1029/2008eo100001.

Magnini, A., Lombardi, M., Bujari, A., Mattivi, P., Shustikova, I., Persiano, S., Patella, M., Bitelli, G., Bellavista, P., Lo Conti, F., Tirri, A., Bagli, S., Mazzoli, P., Castellarin, A., 2023. Geomorphic flood hazard mapping: from floodplain delineation to flood hazard characterization. Hydrol. Sci. J. 68 (16), 2388–2403. https://doi.org/10.1080/02626667.2023.2269909.

Manfreda, S., Samela, C., 2019. A digital elevation model based method for a rapid estimation of flood inundation depth. J. Flood Risk Manag. 12, e12541. https://doi.org/10.1111/jfr3.12541.

Manfreda, S., Samela, C., Gioia, A., Consoli, G., Iacobellis, V., Giuzio, L., Cantisani, A., Sole, A., 2015. Flood-prone areas assessment using linear binary classifiers based on flood maps obtained from 1D and 2D hydraulic models. Nat. Hazards 79 (2), 735–754.

Marchesini, I., Salvati, P., Rossi, M., Donnini, M., Sterlacchini, S., Guzzetti, F., 2021. Data-driven flood hazard zonation of Italy. J. Environ. Manag. 294, 112986. https://doi.org/10.1016/j.jenvman.2021.112986.

Munawar, H.S., Hammad, A.W.A., Waller, S.T., 2022. Remote sensing methods for flood prediction: a review. Sensors 22 (3), 960. https://doi.org/10.3390/s22030960.

Muñoz Sabater, J., 2019. ERA5-Land hourly data from 1950 to present. In: Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.e2161bac (Accessed on 29-Feb-2024).

Neal, J., Hawker, L., Uhe, P., Paulo, L., Sosa, J., Savage, J., Sampson, C., 2023. FABDEM V1-2. University of Bristol. https://doi.org/10.5523/bris.s5hqmjcdj8yo2ibzi9b4ew3sn.

Nobre, A.D., Cuartas, L.A., Momo, M.R., Severo, D.L., Pinheiro, A., Nobre, C.A., 2016. HAND contour: a new proxy predictor of inundation extent. Hydrol. Process. 30, 320–333. https://doi.org/10.1002/hyp.10581.

Pradhan, B., Tehrany, M.S., Jebur, M.N., 2016. A new semiautomated detection mapping of flood extent from TerraSAR-X satellite image using rule-based classification and Taguchi optimization techniques. IEEE Trans. Geosci. Remote Sens. 54 (7), 4331–4342. https://doi.org/10.1109/TGRS.2016.2539957.

Rahmati, O., Pourghasemi, H.R., Melesse, A.M., 2016. Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran Region, Iran. CATENA 137, 360–372. https://doi.org/10.1016/j.catena.2015.10.010.

Ross, C.W., Prihodko, L., Anchang, J., Kumar, S., Ji, W., Hanan, N.P., 2018. HYSOGs250m, global gridded hydrologic soil groups for curve-number-based runoff modeling. Sci. Data 5, 180091. https://doi.org/10.1038/sdata.2018.91.

Samela, C., Troy, T.J., Manfreda, S., 2017. Geomorphic classifiers for flood-prone areas delineation for data-scarce environments. Adv. Water Resour. 102, 13–28. https://doi.org/10.1016/j.advwatres.2017.01.007.

Schumann, G., Giustarini, L., Tarpanelli, A., Jarihani, B., Martinis, S., 2023. Flood modeling and prediction using earth observation data. Surv. Geophys. 44 (5), 1553–1578. https://doi.org/10.1007/s10712-022-09751-y.

Schwanghart, W., Kuhn, N.J., 2010. Topotoolbox: a set of Matlab functions for topographic analysis. Environ. Model. Softw. 25 (6), 770–781. https://doi.org/10.1016/j.envsoft.2009.12.002.

Shahabi, H., Shirzadi, A., Ghaderi, K., Omidvar, E., Al-Ansari, N., Clague, J.J., Geertsema, M., Khosravi, K., Amini, A., Bahrami, S., Rahmati, O., Habibi, K., Mohammadi, A., Nguyen, H., Melesse, A.M., Ahmad, B.B., Ahmad, A., 2020. Flood detection and susceptibility mapping using sentinel-1 remote sensing data and a machine learning approach: hybrid intelligence of bagging ensemble based on k-nearest neighbor classifier. Remote Sens. 12 (2), 266. https://doi.org/10.3390/rs12020266.

Smith, R., 2015. Directive 2008/94/EC of the European Parliament and of the Council of 22 October 2008. In: Smith, R. (Ed.), Core EU Legislation. Macmillan Education UK, pp. 423–426. https://doi.org/10.1007/978-1-137-54482-7_44.

Tehrany, M.S., Jones, S., Shabani, F., 2019. Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. CATENA 175, 174–192. https://doi.org/10.1016/j.catena.2018.12.011.

Tien Bui, D., Hoang, N.-D., 2017. A Bayesian framework based on a Gaussian mixture model and radial-basis-function Fisher discriminant analysis (BayGmmKda V1.1) for spatial prediction of floods. Geosci. Model Dev. 10 (9), 3391–3409. https://doi.org/10.5194/gmd-10-3391-2017.

Vafakhah, M., Mohammad Hasani Loor, S., Pourghasemi, H., Katebikord, A., 2020. Comparing performance of random forest and adaptive neuro-fuzzy inference system data mining models for flood susceptibility mapping. Arab. J. Geosci. 13 (11), 417. https://doi.org/10.1007/s12517-020-05363-1.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., Bai, X., 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.

WMO, 2021. 2021 State of Climate Services: Water. World Meteorological Organization. https://doi.org/10.4060/cb6241en.

Yilmaz, K.K., Sassi, M., Zanardo, S., Tillmanns, S., Hilberts, A., 2025. Evaluation of multiple geomorphic flood inundation mapping techniques over the Eastern United States. In: EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-19781. https://doi.org/10.5194/egusphere-egu25-19781.

Yu, Y., Lv, M., Qian, J., Lv, J., Miao, D., 2024. Fuzzy information gain ratio-based multi-label feature selection with label correlation. Int. J. Mach. Learn. Cybern. https://doi.org/10.1007/s13042-023-02060-9.

Zhao, G., Pang, B., Xu, Z., Yue, J., Tu, T., 2018. Mapping flood susceptibility in mountainous areas on a national scale in China. Sci. Total Environ. 615, 1133–1142. https://doi.org/10.1016/j.scitotenv.2017.10.037.