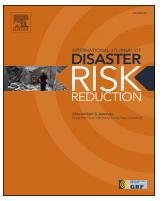
International Journal of Disaster Risk Reduction 98 (2023) 104110

Contents lists available at ScienceDirect

International Journal of Disaster Risk Reduction

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



An empirical flood fatality model for Italy using random forest algorithm

Mina Yazdani [a, c, *], Christian N. Gencarelli [b], Paola Salvati [c], Daniela Molinari [a]

[a] Department of Civil and Environmental Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milan, Italy
[b] Italian National Research Council (CNR), Institute of Environmental Geology and Geoengineering (IGAG), 20131, Milan,Italy
[c] Italian National Research Council (CNR), Research Institute for Geo-Hydrological Protection (IRPI), 06128, Perugia, Italy

ARTICLE INFO

Keywords:
Flood fatalities
Flood damage model
Random forest
Po river

ABSTRACT

Due to the increasing occurrence of natural hazards, such as floods, a significant number of lives are lost each year worldwide. The risk of experiencing catastrophic losses from flooding is exacerbated due to the changing climate and increasing anthropogenic activities. Consequently, predicting the conditions leading to fatalities is crucial in the assessment of flood risk. However, the existing modeling capabilities in this field are limited, emphasizing the critical need for the development of such tools. Here, we show that the occurrence of flood fatalities can be estimated using a random forest (RF) algorithm applied to nine explanatory variables characterizing each fatality. Furthermore, by converting the RF model outcomes into a user-friendly tool, it is possible to predict the probability of the occurrence of flood-related fatalities, based on variables describing hazard intensity and the environmental and sociodemographic conditions contributing to such events. Our results represent an initial attempt towards a predictive model of flood fatalities in the Italian context. They reveal the key factors that together influence flood fatalities, enabling the prediction of such occurrences. These findings can serve as a foundational framework for quantitatively assessing the risk to the population from such events and as a valuable resource for identifying strategies to mitigate flood risk.

1. Introduction

Flood fatalities have become a growing global concern, due to the increasing number of people being affected by these occurrences [1]. This issue is further exacerbated by the increasing frequency of intense precipitation events attributed to climate change [2], leading to a larger portion of the world population being exposed to flood hazards [3]. Furthermore, continuous population and urban growth in flood prone areas [4], with limited sustainable flood control measures in place, aggravate the consequences of climate change, especially the loss of life [5]. In this regard, the reduction of losses caused by natural disasters, including the number of fatalities, is an expected outcome of the Sendai Framework for Disaster Risk Reduction 2015–2030 [6]. The possibility of predicting and estimating the expected loss of life in flood prone regions is highly desired in both developed and developing countries. Identifying vulnerable communities that are potentially affected by these events, as well as the areas with the highest risk of flooding, are among the top priorities of decision-makers in flood risk management.

* Corresponding author. Italian National Research Council (CNR), Research Institute for Geo-Hydrological Protection (IRPI), 06128, Perugia, Italy.
E-mail addresses: mina.yazdani@irpi.cnr.it (M. Yazdani), christian.gencarelli@igag.cnr.it (C.N. Gencarelli), paola.salvati@irpi.cnr.it (P. Salvati), daniela.molinari@polimi.it (D. Molinari).

https://doi.org/10.1016/j.ijdrr.2023.104110
Received 13 July 2023; Received in revised form 27 October 2023; Accepted 30 October 2023
Available online 31 October 2023

In this research, we focus on analyzing the human consequences of flood events with the goal of developing a flood fatality model, in the context of Italy, in particular focusing on the Po River District in Northern Italy which covers the largest Italian hydrographic basin with a total area of approximately 82,700 km^2.

Different authors have previously worked on the topic of flood fatalities by considering various aspects of these events. Based on these studies, Petrucci [1], identifies two primary categories of factors contributing to flood fatalities: environmental factors and victim characteristics. Environmental factors consider elements that are related to flood intensity and the geographical features of the affected location, while victim characteristics describe the vulnerability of the individuals affected by the fatal flood events. It is worth noting that other studies have developed flood fatality prediction models, using various univariable or multivariable approaches. These models offer valuable insights into the potential loss of life during floods. However, among the available models, no specific one has been calibrated or validated in the Italian context; nonetheless, existing models are hardly transferable to Italy due to the strong dependency of contributing factors of flood fatalities on local conditions.

For the objective of this study, random forest (RF) [7], a supervised machine learning algorithm based on an ensemble of multiple decision trees, is utilized to analyze victim records over a time span of 50 years (Fig. 1). This model was chosen to consider multiple significant variables for flood fatalities. This algorithm is often used for complex classification and regression problems and has previously been implemented in multi-variable flood damage modelling and risk assessments. For example, Wang et al. [8]; used RF to develop a flood hazard risk assessment model and to identify the most significant indicators of flood risk, while Wagenaar et al. [9] adopted this algorithm to model flood damage to residential buildings. Terti et al. [10] applied RF to study vehicle-related flash flood fatalities in the United States.

In our study, the RF algorithm was used to create a tool for the estimation of flood fatalities by solving the classification problem of fatality/non-fatality. The remainder of this paper is organized as follows: First, in Section 1.1, we examine the existing knowledge on flood fatality mechanisms and drivers, which serves as the starting point for the selection of model variables. Then a brief description of the investigated area is supplied (Section 2). In Section 3, we present the initial dataset at the base of the model along with the analyses implemented to prepare the data for the application of the RF algorithm. Subsequently, we explain the model training and validation (Section 4). The analysis results are presented in Section 5, including the best runs of the model, the measures used to compare the performance of the model under these runs, and the final model setup. The next part of the paper (Section 6) presents a critical analysis of the results and their translation in a user-friendly tool to estimate the probability of a fatality for a given scenario. Finally, we summarize the main results of the study (Section 7).

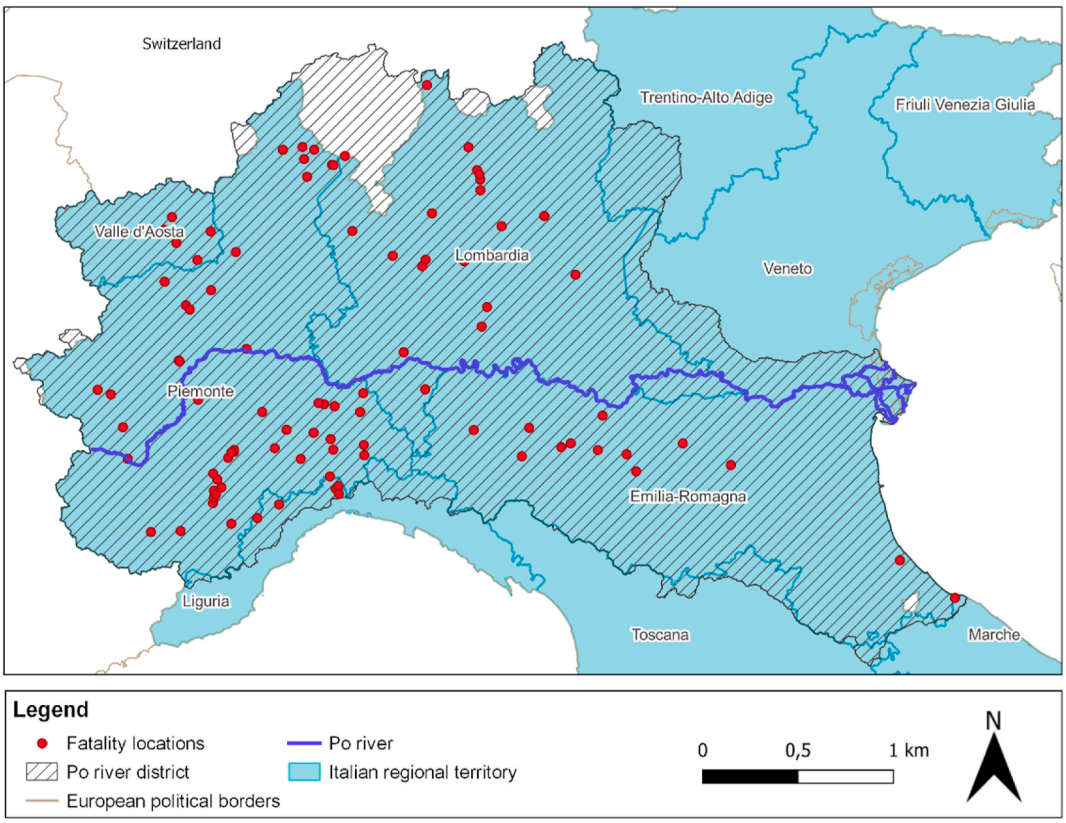


Fig. 1. Study area and geographical distribution of flood fatality in the 50-year period from 1970 to 2019.

1.1. Background on the drivers of flood fatality

Several studies have been conducted to determine the number of flood fatalities caused by flood events and/or to characterize the fatalities [1,11,12]; and references therein. As briefly mentioned before, in the review article by Petrucci [1]; two main groups of contributing factors were identified as drivers of flood fatalities: environmental factors, and victim characteristics. The first group relates to both the hazard conditions and the geomorphological characteristics of the location where fatalities occur. These include the characteristics related to the topography and the morphology of the flood plain, the catchment size [13]; [14]; [15], the presence of obstructions [13]; [1,5], the state of land cover [15,13,14]; which could influence the runoff features, and the flood depth and velocity, namely, two of the most important hazard characteristics influencing the degree of damage to humans. Spitalar et al. [15], Diakakis and Deligiannakis [16], and Terti et al. [17] add that several elements such as the place where the accident occurs (i.e., rural or urban areas) contributes to the impact of the flood. In rural areas, the ability of the individuals to get to safety can be negatively affected by a lack of fast response units for rescue and evacuation, or the absence of mitigating structures [15]. Alternatively, an increase in the concentration of human activities in urban areas can amplify risk factors [18]. The accident location can be interpreted as indoor/outdoor areas, which affect the probability of loss of life [12,19,20]. Finally, the distance of individuals from the watercourse [13], and the time of the accident [18,20–22] are considered important factors; the latter, determines the visibility conditions [15] during a flood event, which affects the ability of individuals to judge the depth and the speed of the flowing water [1].

The second group of contributing factors are related to the characteristics of fatalities. Significant factors that describe the conditions of victims are mentioned in the literature as the age and gender [1,11,12]; and references therein), their economic situations and level of education, the state of residency of the victims [11,23] and whether or not they adopt hazardous behaviors [1,12]; [24]; [25]. However, the relative importance of each factor may differ depending on the context and the country where the fatalities are being investigated.

In addition to studies on contributing factors, several scholars have focused on the development of prediction models for flood fatalities. The first and simplest ones consider the water depth as the only explanatory variable [26–28]. Subsequently, more complex models were developed. For example Jonkman [29], suggested a methodology to determine loss of life due to sea and riverine floods in the Netherlands which considers the effects of water depth, flow velocity and the possibilities for evacuation [30]. Penning et al. [13] introduces a model to estimate flood-related injury and loss of life by developing a 'hazard rating' for the different floodplain zones (based on depth, velocity, and a debris factor associated with each zone), a score for the 'area vulnerability' (for example, the flooding lead time), the population at risk, and the population's vulnerability (in terms of elderly and sick or disabled). Jonkman et al. [30] also proposed a model by analyzing flood characteristics, such as water depth, rise rate and flow velocity, followed by the estimation of the number of exposed people. In another study Brazdova and Riha [18], proposed a model for the estimation of the number of human losses during river floods in central Europe, based on the data from 19 past floods in the region. In this approach, they estimated the number of fatalities, combining three groups of contributing factors representing the material loss due to a flood event (including factors related to the flood hazard as well as the number of inhabitants in the exposed area), the preparedness of the society to face floods, and the methods to warn the population. Terti et al. [10] suggests a random forest classifier to determine the likelihood of loss of life for vehicle-related flood deaths in the United States, as a function of representative variables, such as the age of the driver, the density of the road system, the maximum duration of precipitation, etc. Recently Alfieri et al. [31], presented a physically-based modeling framework to estimate the seasonal pattern of population exposure, and mortality rates from river floods in all the countries worldwide at a spatial resolution of up to 1 km. This model was derived using state-of-the-art global stream flow reanalysis data. The resulting inundation maps were then combined with the information on the population exposed, to estimate the mortality.

2. Study area

The Po river district shown in Fig. 1 includes three of the largest Italian regions, namely Piedmont, Lombardy, and Emilia Romagna, covering 86.5 % of the District. In addition, the district includes the Valle d'Aosta region and parts of Trentino Alto Adige, Veneto, Liguria, Tuscany and Marche. Approximately 20 million people (1/3 of the Italian population) reside in the Po district where 37 % of the national Gross Domestic Product (GDP) is produced. Based on the Pinna Modified Köppen climatic classification of Italy [32] which is assumed to be still valid, the study area is characterized by a high climate variability, spanning from the coldest classes (DW and H) to the temperate sub-continental climate class (Cfsa). The Po plain experiences a temperate climate with warm summers and the mean temperature of the coldest month ranges from −1.5 °C to −3 °C. The average annual precipitation is about 1080 mm [33], with high seasonal variability among the different physiographic areas. The hydrographic network that characterizes the Po basin has a considerable extension, comprising 35 main sub-basins, characterized by a complex water system formed by the natural and artificial surface water networks [34]. The plains cover 42 % of the territory, with the remaining 58 % consisting of mountains and hills.

Consequently, suitable models are required to estimate the risk of flood fatalities in the area. Given that it covers a large part of the Italian territory and the diverse climatic, topographic, and geomorphological conditions it encompasses [35], this region provides an ideal context for the development of our fatality model.

3. Method

The flowchart of the work done in this study is presented in Fig. 2. It includes different steps that are described in the following sections. In brief, the method entails 1) identifying the key contributing factors, i.e., explanatory variables for estimating flood-related fatalities based on extensive review of the literature. We then compared these findings with the information available in the

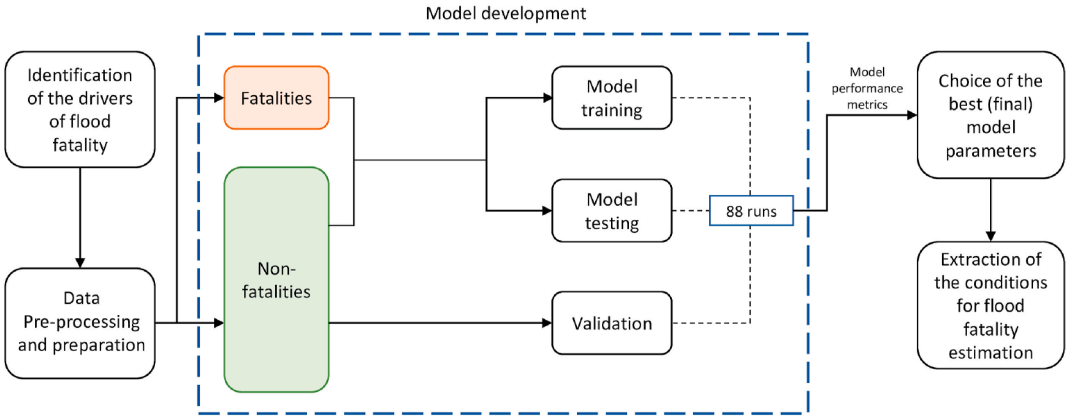


Fig. 2. Flowchart of the study. The final dataset consists of 127 fatalities, and 1270 non-fatalities. The model training and testing is performed using the fatalities dataset, together with a part of the non-fatalities, selected randomly. The model validation is carried out utilizing the remaining part of the Non-fatality records.

victim records (hereinafter referred to as the "Fatality" dataset). 2) Following this comparison, two additional processes were necessary to enhance the characterization of the fatalities: (i) the pre-processing of available data to ensure their suitability for use by the RF algorithm and (ii) the definition of new variables related to factors for which no information was available in the original dataset, along with the corresponding assignment of values. Next, to enable the RF algorithm to properly assess the role and the significance of the different explanatory variables, we created a synthetic dataset consisting of records for individuals who were involved in the event but managed to survive (these operations are indicated in Fig. 2 as "data pre-processing and preparation"). The dataset of fatalities combined with this synthetic dataset (hereinafter referred to as the "Non-fatality" dataset) was then used as the input for the creation and testing of the RF model.

3.1. Data

The data used in this study is partially derived from the dataset presented in the work of Salvati et al. [12]. Their research investigated fatal flood events in Italy during the 50-year period of 1965–2014, which is freely accessible (https://osf.io/sm8tv/files/osfstorage). Additionally, records related to the events occurring between 2014 and 2019 were acquired from the POLARIS website (https://polaris.irpi.cnr.it/), which publishes accurate information on geo-hydrological risk to the Italian population, such as periodic reports on fatal landslide and flood events, and the number of fatalities [36]. For this current study, the data coming from the sources mentioned above, along with the geographical locations of the fatalities, was extracted for the Po river District. This dataset comprises 138 records of flood fatalities due to 45 flood events from 1970 to 2019. The geographical distribution of the fatalities is presented in Fig. 1.

The information available in the dataset is listed below. For some of the fatalities, not all of the listed information was available.

- Event identification: information such as the date of the event, the region, the province, the municipality associated with the fatality, and an identification code that relates each record to its associated location point.
- Situational parameters: "Time" and "Darkness" describing the situation in which the fatality occurred; in particular, information regarding time of the accident and the visibility conditions. Time is described both in a 24-h clock system, or in descriptive terms such as "early morning", while the Darkness condition is reported as "Yes" or "No".
- Victim identification: information on the age and gender of the victims.
- Place of the accident: information regarding the location where the victim was found, such as on the street, inside a flooded room, or on a bridge.
- Circumstances of the fatality: "Dynamic" and "Manner" of the fatality, explaining how the accident happened, e.g., if the person was stuck in a flooded room, or had fallen in the watercourse, etc.
- Inappropriate behavior: whether a victim was engaged in inappropriate behaviors, such as retrieving possessions, going to the lower levels of the building, walking in a flooded pathway, etc.
- Cause of death: specifying how the victim died; for example, through drowning, electrocution, or a heart attack.

Of the 138 fatalities, the majority, accounting for 60.9 %, were male individuals. Most victims were between 15-29 and 65–85 years old. Moreover, 65 % of people who died belong to the Piedmont Italian region, followed by the Lombardy region (18 %). The statistics regarding the places where the victims were found showed that most victims were discovered in outdoor locations, especially streets (30.7 %).

3.2. Data pre-processing

To obtain a more refined set of data records with well-defined explanatory variables for flood fatalities, suitable for RF analysis, a number of modifications were conducted on the initial dataset. They include (i) eliminating data related to time and darkness conditions as these fields contained a large number of missing values; (ii) removing records where the cause of death was unrelated to the

current analysis, such as cases attributed to factors like heart attacks; (iii) excluding the information on manner, dynamic, and inappropriate behavior for the current analysis, as these aspects could potentially overshadow the impact of other factors (e.g. the inappropriate behavior of an individual in case of a flood fatality might disproportionately influence the significance of the morphology of the area); (iv) reinterpreting and reclassifying available information to obtain well-defined variables, which can be categorical or continuous (see Table 1); and (v) defining and calculating additional significant explanatory variables, according to suggestions from the literature on contributing factors of flood fatalities, as presented in Section 1.1. The values of these variables were evaluated based on the location of each victim record, and the logic and criteria for choosing them are explained as follows:

- Morphological Zone: the characteristics of the flood plain, such as the morphology of the area, influence critical factors such as the depth and velocity of the flooding water [13], as well as the time of concentration of the basin. Moreover, it can be considered as a proxy to characterize the slope, which affects the flow of water and the stability of individuals during a flood event. Therefore, based on the spatial data produced by the Po River District Authority [37], each fatality record is associated with a morphological zone class described as "Mountain" or "Plain".
- Corine Land Cover (CLC): the land cover of an area influences the hydrological processes, such as runoff and infiltration. In urban areas, extensive built-up and compacted landscapes can affect the volume and the velocity of the runoff, aggravating the hazard conditions in a flood event. Conversely, urban areas provide more opportunities for shelter and rescue [1]. In this study, the land use data associated with each fatality record was derived from CLC (2018) [38], at the second level classes. The majority of the fatalities in our dataset were linked to the urban fabric.
- Distance from the river: the depth and velocity of the floodwater can be influenced by the distance from the river source of the flooding which was determined here utilizing the measurement tool in Google Earth Pro (2020).
- Density of the buildings: based on the building footprints in areas where the fatalities occurred, we calculated the density of buildings in each affected municipality, to obtain a proxy for representing population density. This was necessary since obtaining precise information about the population density at the desired scale of analysis due to the time span of the data records was challenging. Furthermore, building density can serve as a proxy for assessing the level of urbanization in the flooded area, complementing the data on land use.
- Flood hazard scenario: the flood hazard maps developed as part of the Po District's Flood Risk Management Plan (FRMP) [37] were used to link each record with a level of hazard probability at the fatality location. These maps categorize the flood hazard in the Po river district based on three probability scenarios (low, medium, and high probabilities of flood occurrence). It is important to note that the quality of these flood hazard zonings varies unevenly within each river basin authority (RBA) and across all RBAs. As mentioned by Marchesini et al. [39]; in many of the basins in Northern Italy, most of the maps do not cover the whole extent of an RBA with the same accuracy. They are limited to the main river channels, and the areas where vulnerable elements are most abundant. Therefore, in the cases where the fatality points are outside the available hazard scenario maps, a value "Outside" is assigned to this variable.

Table 1
List of variables used as indicators of flood fatality.

Parameter	Categories	Description
Gender	F	Gender of the victims (based on ISTAT (2011) dataset); feminine and masculine
	M	
Age	≤ 14	
	15–29	Age of the victims (based on ISTAT (2011) dataset); categorical variable
	30–49	
	50–64	
	65–74	
	≥ 75	
Place	Bridge	Bridge: The victims were found on a bridge that collapsed
	Building	Building: The victims were found inside a generic building (house, cellar, shop, etc.)
	Campsite	Campsite: The victims were found in a camping area.
	Outdoor	Outdoor: The victims were found outdoors, whether in a vehicle, or on foot.
	River	River: The victims were found near the embankments (in a 1 m distance from the river) or inside the river.
Morphological zone	Mountain	Indicates the morphological zone corresponding to the location of the fatality event; categorical variable
	Plain	
Hazard Scenario	H	Describing the probability of flood scenario based on the maps produced within FRMP.
	M	H: frequent floods, return period 30–50 years
	L	M: floods with medium frequency, return period 100–200 years
	Outside	L: rare floods, return period up to 500 years
Corine Land Cover	15 possible categories	Categorical variable
Distance from the river	All	Distance of the fatality location from the river causing the flood event, continuous variable
Density of the buildings	All	Density of the buildings with respect to the municipal area; continuous variable
Return period	15 possible categories	Return period of the max 24-hr precipitation; categorical variable
Solid transport	Yes	Indicating the presence of solid material; categorical variable
	No	
	NA	

– The return period of the maximum daily precipitation: as described in the introduction, hazard intensity is a critical factor influencing flood fatalities. Each fatality record in the dataset is associated with a flood event that occurred in the Po river district; however, information on the intensity of such floods is not provided. To address this, we analyzed the grey literature (e.g., technical reports, agency data, and newspapers) produced from 1970 to 2019 to gather data on hazard intensity. However, due to the 50-year time span of the flood events in the dataset, the level of information collected was not uniform, especially due to the scarcity of information on the events dating back to the 1970s. Thus, we made a compromise and selected the parameters that could uniformly represent flood hazard across all events in the dataset. After thorough investigation, we found that the maximum daily precipitation value could be retrieved for almost all flood events, making it a suitable indicator of the severity of the flood events. Similar indices have also been identified by Alfieri and Thielen [40] as effective proxy predictors of flash floods. Thus, using this data, the return period of the maximum daily precipitation causing the flood event was obtained to serve as a proxy for the return period of the flood.
– Solid transport: following the suggestion of previous studies [13], [1]; which highlight the relevance of bed load and solid material carried during floods in relation to fatalities, we conducted an investigation using various sources to retrieve this variable for our dataset. The availability and reporting methods varied across different flood events. As a result, we categorized the variable representing the presence of solid material transported by the flood into three qualitative classes ("Yes": indicating the occurrence of solid material transport, "No": indicating no solid material transport, and "NA": indicating no evidence of such information).

A summary of the explanatory variables, their implications and their possible values is presented in Table 1.

3.2.1. A synthetic dataset of non-fatalities

To assess the significance of each variable in causing flood fatalities, it is necessary to integrate the fatalities dataset with data related to individuals who were affected by the same flood event, but survived. Since such a dataset was not available for the flood events considered in this study, a synthetic dataset of non-fatalities was created to fulfill this objective. The synthetic dataset was created by randomly selecting values from the frequency distribution of explanatory variables associated with the flooded area. However, since the exact extent of the flooded area was not available for all the events, hazard scenario maps corresponding to the return period of the flood (as described in Section 3.2) were used as proxies. These maps were used to estimate the perimeter of the flooded area, constrained to the boundaries of the municipality where the fatality occurred, as well as the upstream and downstream municipalities. Using this approach, for each record in the fatalities dataset, 10 non-fatality records were created, characterizing individuals (and environmental circumstances) who were affected by the same flood but survived. This process resulted in a complete dataset consisting of 127 fatalities and 1270 non-fatalities, used for the modelling process. A detailed description of the creation of this dataset is available in the Supplementary material.

4. Model development

The model development process is a classification task, where each data record in the total dataset is characterized by ten input variables (explanatory variables), and a target variable indicating whether or not a fatality occurred. The process leading to the creation of the RF model was implemented in RStudio (2020) - version 4.1.2, using the function "randomForest" of the randomForest CRAN package [7,41].

Multiple runs of the model were conducted by varying the significant model parameters. The model's performance in each run was evaluated using two distinct datasets and various performance metrics to identify the most suitable model proposed in this study.

4.1. Model training and validation

As discussed previously, the total dataset consists of 1397 data records, with 9 % representing fatalities and 91 % representing non-fatalities. These statistics show that the class distribution of the target variable is unbalanced. This could cause the model to be biased towards the majority class resulting in overfitting issues. To avoid this problem, we implemented random undersampling (RUS), as proposed by Yap et al. [42]; in which the records in the majority class were eliminated randomly to achieve an equal distribution with the minority class. To control the balance in the data records used for model training and validation, we defined the parameter K (K = 1, 2, …,10) that specifies the ratio between the number of non-fatalities and that of fatalities, as expressed in Equation (1). The model was run multiple times with different K values, randomly sub-setting from the non-fatality records, to assess its performance.

$$K = \frac{(Number\ of\ Non-fatalities)}{(Number\ of\ Fatalities)} \quad \text{Eq. (1)}$$

Multiple combinations of training and test sets were considered, each time applying a different ratio between the two sets. Specifically, from the 1270 records in the synthetic dataset of non-fatalities, (127*K) random samples were used in combination with the 127 fatality records to train the model (the portion of data used for this purpose is called "Training set"). The model's performance was then tested on the "Test set". The remaining part of the non-fatalities dataset (1270–127*K) that was not used for model training, was employed to conduct the model validation. The algorithm allows the user to define the number of the random decision trees that are created, wherein 500 trees were chosen as the optimal number for the development of the algorithm for this study.

The model's outcome is a classification of the target variable for each record in the test and validation data sets, as either Fatality or Non-fatality. By comparing what the model predicts on each record and the real value of the target variable in the test and validation data sets, we evaluated the prediction ability of the model. For this purpose, we investigated six different performance measures.

These include the classification accuracy (Acc), the true positive rate (TPR), the true negative rate (TNR), and the false negative ratio (FN-ratio, which defines the number of false negatives-missed alarms with respect to all the fatalities), on the test dataset. Additionally, the validation dataset includes two measures: true negative ratio ($TNR_{validation}$) and false positive ratio ($FPR_{validation}$). The equations for these measures are provided in the appendix B.

5. Results

A total of 88 sensitivity runs were conducted to calibrate the RF model parameters for optimal results. In each run, three model parameters were modified: (i) the parameter K (controlling the number of data records used for model training, testing and validation), (ii) the ratio of training and test dataset, and (iii) the number of considered explanatory variables. The model was run both with all input variables and with each variable removed individually. The 88 runs comprised 11 main cases: one with all variables included and 10 cases involving the removal of one variable. For each main case, eight sensitivity analyses were performed considering (i) and (ii). The results obtained for all the runs in terms of performance measures are available in Table S1 in the Supplementary Material. The choice of the best-fit model was based on finding a balance between Acc, TPR, TNR, and FN-ratio. Subsequently, we considered the results of the validation as the next performance measure. The performance of two of these runs, which were chosen as the best cases, are reported in Table 2 as runs 1.2.2, and 2.2.2. These runs achieved acceptable classification accuracy on the test set while maintaining a low FN-ratio, indicating effective performance in minimizing missed alarms, which, in our case, pertain to fatalities. Furthermore, they show high performance on the validation data ($TNR_{validation}$ values).

The results for run 1.2.2 show an FN-ratio of 0.222 (6 out of 27 fatalities in the test set), and a classification accuracy of 88.3 %, indicating that the model could correctly classify the fatalities by a ratio of 77.8 % (indicated by the TPR); additionally, for the non-fatalities in the test set, the TNR was 94 %. Considering the performance measures of the validation dataset, the $TNR_{validation}$ was 93.7 %, implying that the model could correctly predict 952 of the non-fatality cases among a total of 1016 records, and misclassified 64 cases. Thus, we conclude that the initial choice of the 10 explanatory variables for flood fatality was reasonable because an acceptable level of performance was achieved in this run, where all predictors were used in model training.

In run 2.2.2 the model was created based on 9 predictor variables, excluding the "Gender" parameter. As presented in Table 2, the FN-ratio is 0.185 (containing five out of 27 fatalities in the test set), and the classification accuracy of the model is 89.6 %, with TNR and TPR values of 81.5 and 94 %, respectively; this is considered one of the best calibrations for the model. Considering the results of the validation, the model correctly classifies 947 non-fatality cases out of 1016 records ($TNR_{validation}$:93.2 %), and misclassifies 69 cases ($FPR_{validation}$:6.8 %).

5.1. Interpreting the results into rules

Tree ensembles such as RF are accurate machine learning algorithms, but they are usually complex and difficult to interpret. We created a tool to enable an easy interpretation of the results, which can be used by non-experts/decision-makers and organizations interested in estimating and mitigating flood fatality, through the inTrees package in R (CRAN) [43]. This package introduces a framework that extracts and summarizes rules that govern the splits in each tree in the tree ensemble (in a descriptive way and with the support of logical operators) and calculates frequent variable interactions [44]. He states that "a rule can be expressed as C ⇒ T, where C, referred to as the condition of the rule, is a conjunction of variable-value pairs and T is the outcome of the rule", and that "the rules extracted from a tree ensemble are a combination of rules extracted from each decision tree in the tree ensemble." In fact, the rules identify the combinations of model variables that are likely to result in an estimation of fatality or non-fatality; extracting these rules can ensure a more straightforward interpretation of the results by the model. For the adaptation of the inTrees algorithm on the final model characterized in Table 2 (run 2.2.2), first, the rules were extracted from the 500 trees generated by the model; subsequently, based on the metrics of those rules, they were classified according to their qualities. Next, each rule was pruned (i.e., purified from irrelevant or redundant variable-value pairs). Finally, the rules extracted from the tree ensemble were summarized into a rule-based learner. Therefore, after executing the inTrees package, we extracted 34 rules for our flood fatality model from an initial number of 12,917 rules (i.e., before pruning). The rules are listed in Table S2 in the supplementary material. Each rule is characterized by its metrics (frequency or accuracy of the rule, error, and length that identifies the number of variable-value pairs) and the conditions (explanatory variable-value pairs) of that rule leading to a prediction of fatality ('1') or non-fatality ('0').

Table 2
Summary of the performance measures for runs 1.2.2, 2.2.2. and 9.1.2. Run 1.2.2, described as the base case, refers to the RF model with all the 10 explanatory variables, whereas in the other cases, one variable is removed.

Runs				Testing				Validation	
Label	Description	K	training/test	Acc	FN-ratio	TPR (%)	TNR (%)	$TNR_{validation}$	$FPR_{validation}$
run 1.2.2	Base case	2	0.80	88.3	0.222	77.8	94.0	93.7	6.3
run 2.2.2	Removing Gender	2	0.80	89.6	0.185	81.5	94.0	93.2	6.8
run 9.1.2	Removing Distance	1	0.80	68.6	0.250	75.0	60.9	68.4	31.6

6. Discussion

6.1. Sensitivity of the predictors

To better interpret the results, and identify the most important features in describing the model response, we used the function "varImpPlot" from the randomForest CRAN package [41], which enables the measurement and plotting of the importance of the variables used in the model. This information is crucial for a better assessment of the logic of the model and evaluating the accuracy of this logic. One of the measures for representing the variable importance in the varImpPlot function is the mean decrease accuracy (MDA). This measure is computed by permuting out-of-bag (OOB) data [7]: where a sub-sample with replacements is used to create training samples to train the model. The remaining part of the observations (not used to fit a given tree) is referred to as the OOB observations. For each tree, the error rate on the OOB portion of the data was recorded, after permuting each explicative variable. To obtain the MDA for the explicative variables in each model run, the difference between the case where the variable is included in the model and that where it is permuted is then averaged over all the trees and normalized by the standard deviation of the differences. This measure shows the extent to which the accuracy of the model decreases when each explicative variable is excluded from the model [41]. The greater this decrease, the higher is the importance of the variable for the model. The plots of the variable importance, obtained for the two runs in Table 2 are shown in Fig. 3.

Based on the plots related to the runs 1.2.2 and 2.2.2, the "Distance" variable has the highest MDA, implying that it has the highest importance in the classification problem. However, we should note that the Distance variable holds some degree of uncertainty. The accuracy of geographical coordinates for the fatality records in the initial database might differ based on the precision of the information about the victim's location. In fact, as mentioned by Guzzetti [45]; the fatality dataset used here is a non-instrumental record, and the uncertainties and the incompleteness in the formation are known biases of this type of catalogue. This issue was observed in a few of the sites used for this study, where inconsistencies were observed between the geo-spatial location of the fatalities and the "Place" variable. For this reason, we decided to investigate the results of the model for the case where the variable Distance is excluded from the input variables. As listed in Table 2, run 9.1.2 has the best performance measures for this case. In this run, the accuracy of the model in classifying the records on the test set was 68.6 %, with the TPR and the TNR were 75 and 60.9 %, respectively. This exhibits a low performance compared to other runs (runs 2.2.2 and 1.2.2.), highlighting the importance of the distance variable, despite some sites showing low accuracy. It's crucial to emphasize that the dataset's accuracy level, in terms of the measure we utilized, exceeds that of many other datasets reporting data at e.g., the municipal scale. However, a quantitative measure of geographical uncertainty related to the distance variable becomes fundamental when extending and implementing this model in different geographical areas. This is particularly important for potential applications of the model.

As mentioned before, this is a first attempt to model the occurrence of flood fatalities based on detailed variables characterizing the fatality conditions. For this purpose, we used the available data on flood fatalities in the Po District that is a sub-set of the national catalogue that was initially developed for other purposes (e.g., regional risk assessment, national flood mortality evaluation) [46,47], and for which the information on the exact site of the accident is not always recorded, especially in the early part of the catalogue. Despite these limitations, the results we obtained from applying the RF algorithm to this dataset demonstrated a high performance. This indicates the possibility to transfer this model to other contexts for which datasets with accurate information might not be available.

Additionally, the variable Gender displayed the lowest importance in run 1.2.2; hence, removing it from the input variables of the model, as in run 2.2.2 was considered acceptable. This is in line with the study of Terti et al. [10]; in which they excluded this parameter based on their risk hypothesis of flash floods. However, it is noteworthy that the variable of gender exhibits a low level of importance in our developed model, contrasting with the findings by Salvati et al. [12]; elaborated for Italy. Their study revealed an overrepresentation of male fatalities when comparing the fatality distribution by gender and age to the corresponding population groups

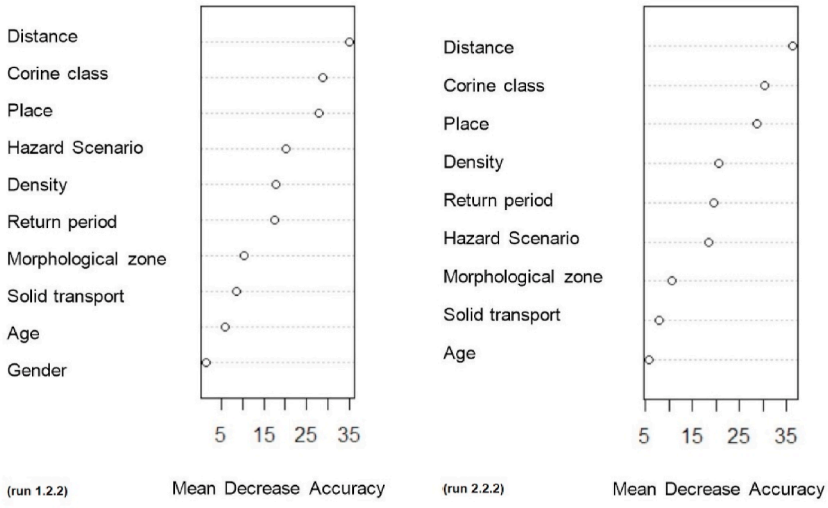


Fig. 3. Variable importance for runs 1.2.2 and 1.2.2 of the RF model. Table 1 presents the description of the variables.

based on national census data, using a multinomial probability mass function. The authors argued that the overrepresentation of males with respect to females could be attributed to a different propensity to engage in risky behavior. Other studies have revealed larger mortality rates of male individuals as well in Greece [48]. Furthermore, in non-developed countries many authors have found that gender largely influences mortality [49–51]. In these contexts, they found that females were more vulnerable when involved in flood disasters resulting in a large number of female fatalities. These findings originally influenced the choice of the explanatory variables for the development of our method. The low significance of gender in this model can be explained when analyzed collectively with other explanatory variables of our model that influence the occurrence of a fatality, describing the environmental condition and the intensity of the flood event. The goal of the model is to estimate flood fatalities, independently of the gender. The environmental variables (distance, land cover, place, hazard scenario, building density, morphological zone) in combination with those representing the intensity of the hazard (return period of the maximum 24-hr precipitation and solid transport) have demonstrated a high level of significance in the estimation of fatalities suggesting that for future implementation of the model, these variables should be represented with a finer level of detail.

One of the most important results of the model was the extraction of the rules allowing the prediction of the fatalities based on the combination of variable-value pairs. For this purpose, we have transformed these rules into a tool that is available in the supplementary material as an Excel spreadsheet. This tool evaluates the probability of a fatality based on the number of rules satisfied for a possible fatality condition, represented by the values of the selected variables. This first attempt to assess the risk of fatalities is of great interest, especially where the previously conducted hazard and risk zoning, constantly updated by the FRMP, is missing; this greatly supports and validates the existing hazard and risk zoning. Moreover, this tool can be used to offer suggestions and guidance to decision-makers to identify the most critical combination of variables, leading to fatal occurrences. These outcomes can be utilized to plan the effective mitigation measures to reduce flood mortality rate, which we consider the most relevant impact on society in this context.

6.2. Limitations of the model

As an initial effort to predict flood fatalities in the Italian context, our model still has some limitations that affect its reliability, which should be overcome in future studies. First, during the preparation of the dataset, replicating the exact conditions of fatality or non-fatality occurrence was not always possible, mainly due to the challenges in retrieving information related to the oldest events. Accordingly: (i) the two sets of information regarding the age and the gender of the records in the non-fatalities dataset were derived from the census data of the year 2011 [52]; whereas (ii) the variables of CLC and density of the buildings, used in both the fatalities and the non-fatalities data sets, were derived, from the CLC database of 2018 and cadastral maps of the present day. In fact, the non-fatalities dataset used in this study is a synthetic one created to compensate for the lack of an actual database on the individuals that managed to survive previous floods. Here, we emphasize the importance of appropriate data collection for such events. Lack of appropriate data was also an issue for other explicative variables, which were either ignored (e.g., Time and Darkness) or approximated; one example is the information on the intensity of the hazard that is often missing in the initial dataset. Our choice of using the return period of the maximum daily precipitation as a proxy for flood intensity might misrepresent the real conditions, especially for the flood events that occurred in smaller basins, as the different sub-basins within the Po river district have different sizes and thus, different values of time of concentration. However, as explained in Section 3.2, only the 24-h precipitation records were accessible for all flood events linked to fatalities. Notably, the catalogue of historical, sparse, point data on flood fatalities in Italy is believed to be the most complete during the period from 1965 to the present day [12]. However, in this study, since the model was applied to the dataset limited to the Po district, to ensure a dataset of reasonable size, it was not possible to use a more recent portion of the catalogue where the precipitation records were available in shorter (sub-daily) durations. To solve this issue in future elaborations of this work, extending the territory of the study to the entirety of Italy can shorten the temporal period of the dataset to a more recent one, without reducing the size of the fatality dataset.

Another issue regarding the explanatory variables is related to the derivation of the hazard scenario. For the fatality points that were outside the hazard scenario areas, we assigned a value defined as “Outside”; this was because either the fatality occurred outside of the available flood scenario maps, or the maps had not been elaborated for the area by FRMP (see Section 3.2). The hazard maps, initially developed for major river branches and areas with preliminary flood risk assessments, were subsequently mandated by the flood directive to undergo continuous updates for a more comprehensive flood risk assessment. This implies that the variable we utilized may undergo changes in the future. Therefore, we highlight the importance of having homogeneous products of hazard estimation for the investigated territory in future elaborations of this work. Additionally, the use of supplementary maps when available can help to overcome such limitations. Overall, as already emphasized, the quality of the input variables has a significant impact on the quality of the model performance.

We acknowledge the inherent limitations and uncertainties associated with this study, given that this is a preliminary effort to model the occurrence of a flood fatality in the Italian context. When estimating the probability of loss of life using this tool, users should consider these limitations. Therefore, we recommend using the tool (available in the Supplementary materials) primarily for comparative purposes, such as assessing the relative levels of risk among different areas, rather than relying on it for precise quantitative evaluations. By adopting this approach, users can ensure the model's appropriate application.

7. Conclusions

In this study, we have developed a model to predict the occurrence of flood fatalities in the Italian context, given a set of variables including environmental conditions, the flood intensity and socio-demographic variables. For this purpose, we utilized a dataset com-

prising 127 flood fatalities coupled with a synthetic dataset of 1270 non-fatalities as inputs for a Random Forest (RF) model. To enhance the model's interpretability, we extracted rules delineating the conditions under which a fatality is more likely to occur. These rules were then transformed into a tool, to estimate the probability of fatality during events characterized by specific variable values. The proposed model is a first attempt to estimate the probability of fatality in the Italian context, and it provides a proxy for the quantitative estimation of the risk to the population. We acknowledge the significance of this estimation, particularly in the context of emergency management planning and in the development of flood risk mitigation strategies to choose the most effective intervention to reduce the risk to people. The case study used to develop the model includes the largest Italian river basin; however, it covers a portion of the Italian territory. For possible application of the model to the entire Italian territory or to other countries, one should consider the possible uncertainties it might introduce. The reliability and the use of the model is contingent upon the quality and accuracy of the input data used, necessitating further refinement and validation in future research endeavors.

Author contributions

Conceptualization and supervision of the work was done by Daniela Molinari and Christian N. Gencarelli. The initial dataset was created by Paola Salvati. Further data collection and elaboration, as well as model development, were performed by Mina Yazdani. All authors contributed to the investigation of the results. The original draft was written by Mina Yazdani. Contributions to the final draft were made by Daniela Molinari, Christian N. Gencarelli and Paola Salvati.

Financial support

The work has been carried out within the context of the MOVIDA project (https://sites.google.com/view/movida-project). Mina Yazdani was supported by a grant awarded by the Italian National Research Council (CNR), project DTA.AD003.474 "Cambiamento climatico: mitigazione del rischio per uno sviluppo sostenibile (quota FOE 2019)".

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Daniela Molinari reports financial support was provided by MOVIDA project (lead by Po River District Authority).

Data availability

Data will be made available on request.

Acknowledgements

Authors acknowledge with gratitude Francesco Ballio, from Politecnico di Milano- Department of Civil and Environmental Engineering, and Simone Sterlacchini, Marco Zazzeri, and Mohammed Hammouti, from Italian National Research Council- Institute of Environmental Geology and Geoengineering, for their useful suggestions and critical reviews during the development of the project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijdrr.2023.104110.

Appendix B. Equations used in the work

The model's performance measures include: the classification Accuracy (Acc), the True Positive Rate (TPR), the True Negative Rate (TNR), the False Negative ratio (FN-ratio), the True Negative ratio (TNR validation) and the False Positive ratio ($FPR_{validation}$), shown in equations A.1 to A.6.

- True positive (TP): The outcome of the test, correctly indicating the presence of a positive condition
- True negative (TN): The outcome of the test, correctly indicating the presence of a negative condition
- False positive (FP), false alarm: The outcome of the test, wrongly indicating the presence of a positive condition
- False negative (FN), missed alarm: The outcome of the test, wrongly indicating the presence of a negative condition

Metrics on the test dataset:

- Classification Accuracy (Acc): It is a measure that describes the number of the correct predictions the model makes with respect to all of the predictions.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Eq. (A.1)}$$

- False Negative Ratio (FN-tatio): A ratio that defines the number of false negatives (missed alarms) with respect to all of the deaths in the test data. It is a measure that is very important in the context of this study, since a high number of missed fatalities in the results does not represent a good performance by the model.

$$FN - ratio = \frac{FN}{N^{\circ}\ of\ "fatalities"\ in\ test\ set} \quad \text{Eq. (A.2)}$$

- True Positive Rate (TPR): This measure considers the ability of the model to correctly classify the “Fatalities”.

$$TPR = \frac{TP}{TP + FN} \quad \text{Eq. (A.3)}$$

- True Negative Rate (TNR): It describes the ability of the model to correctly classify the “Non-fatality” cases.

$$TNR = \frac{TN}{TN + FP} \quad \text{Eq. (A.4)}$$

Metrics on the Validation dataset:

- The two measures of True Negative ratio ($TNR_{validation}$) and the False Positive ratio ($FPR_{validation}$), are described in equations A.5 and A.6.

$$FPR_{validation} = \frac{FP}{TN + FP} \quad \text{Eq. (A.5)}$$

$$TNR_{validation} = \frac{TN}{TN + FP} \quad \text{Eq. (A.6)}$$

References

[1] O. Petrucci, Review article: factors leading to the occurrence of flood fatalities: a systematic review of research papers published between 2010 and 2020, Nat. Hazards Earth Syst. Sci. 22 (2022) 71–83, https://doi.org/10.5194/nhess-22-71-2022.

[2] G. Myhre, K. Alterskjar, C.W. Stjern, et al., Frequency of extreme precipitation increases extensively with event rareness under global warming, Sci. Rep. 9 (2019) 16063, https://doi.org/10.1038/s41598-019-52277-4.

[3] IPCC, Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press. Cambridge University Press, Cambridge, UK and New York, NY, USA, 2022, p. 3056, https://doi.org/10.1017/9781009325844.

[4] J. Rentschler, P. Avner, M. Marconcini, et al., Global evidence of rapid urban growth in flood zones since 1985, Nature 622 (2023) 87–92 https://doi.org/10.1038/s41586-023-06468-9, 2023.

[5] S.N. Jonkman, Global perspectives on loss of human life caused by floods, Nat. Hazards 34 (2005) 151–175, https://doi.org/10.1007/425s11069-004-8891-3.

[6] Sendai Framework for Disaster Risk Reduction 2015-2030. In: UN world conference on disaster risk reduction, 2015 March 14–18, Sendai, Japan. Geneva: United Nations Office for Disaster Risk Reduction, http://www.wcdrr.org/uploads/Sendai_Framework_for_Disaster_Risk_Reduction_2015-2030.pdf,2015.

[7] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[8] Z. Wang, C. Lai, X. Chen, B. Yang, S. Zhao, X. Bai, Flood hazard risk assessment model based on random forest, J. Hydrol. 527 (2015) 1130–1141, https://doi.org/10.1016/j.jhydrol.2015.06.008.

[9] D. Wagenaar, J. Jong, L. Bouwer, Multi-variable flood damage modelling with limited data using supervised learning approaches, Nat. Hazards Earth Syst. Sci. 17 (2017) 1683–1696, https://doi.org/10.5194/nhess-17-1683-2017.

[10] G. Terti, I. Ruin, J. Gourley, P. Kirstetter, Z. Flamig, J. Blanchet, A. Arthur, S. Anquetin, Toward probabilistic prediction of flash flood human impacts, Risk Anal. 39 (2019) 140–161, https://doi.org/10.1111/risa.12921.

[11] K. Papagiannaki, O. Petrucci, M. Diakakis, et al., Developing a large-scale dataset of flood fatalities for territories in the Euro-Mediterranean region, FFEM-DB. Scientific Data 9 (1) (2022 Apr) 166, https://doi.org/10.1038/s41597-022-01273-x,2022.

[12] P. Salvati, O. Petrucci, M. Rossi, C. Bianchi, A. Pasqua, F. Guzzetti, Gender, age and circumstances analysis of flood and landslide fatalities in Italy, Sci. Total Environ. 610 (2018) 867–879, https://doi.org/10.1016/j.scitotenv.2017.08.064.

[13] E. Penning-Rowsell, P. Floyd, D. Ramsbottom, S. Surendran, Estimating injury and loss of life in floods: a deterministic framework, Nat. Hazards 36 (2005) 43–64, https://doi.org/10.1007/s11069-004-4538-7.

[14] J.D. Creutin, M. Borga, C. Lutoff, A. Scolobig, I. Ruin, L. Creton-Cazanave, Catchment dynamics and social response during flash floods: the potential of radar rainfall monitoring for warning procedures, Meteorol. Appl. 16 (2009) 115–125 https://doi.org/10.1002/met.128, 2009.

[15] M. Spitalar, J. Gourley, C. Lutoff, P. Kirstetter, M. Brilly, N. Carr, Analysis of flash flood parameters and human impacts in the US from 2006 to 2012, J. Hydrol. 519 (A) (2014) 863–870, https://doi.org/10.1016/j.jhydrol.2014.07.004.

[16] M. Diakakis, G. Deligiannakis, Vehicle-related flood fatalities in Greece, Environ. Hazards 12 (3–4) (2013) 278–290, https://doi.org/10.1080/17477891.2013.832651.

[17] G. Terti, I. Ruin, S. Anquetin, J. Gourley, A situation-based analysis of flash flood fatalities in the United States, Bull. Am. Meteorol. Soc. 98 (2017) 333–345, https://doi.org/10.1175/BAMS-D-15-00276.1.

[18] M. Brazdova, J. Riha, A simple model for the estimation of the number of fatalities due to floods in central Europe, Nat. Hazards Earth Syst. Sci. 14 (2014) 1663–1676, https://doi.org/10.5194/nhess-14-1663-2014.

[19] M. Diakakis, K. Papagiannaki, Characteristics of indoor flood fatalities: evidence from Greece, Sustainability 13 (15) (2021) 8612, https://doi.org/10.3390/su13158612.

[20] O. Petrucci, P. Salvati, L. Aceto, C. Bianchi, A. Pasqua, M. Rossi, F. Guzzetti, The vulnerability of people to damaging hydrogeological events in the calabria region (southern Italy), Int. J. Environ. Res. Publ. Health 15 (1) (2018) 48 https://doi.org/10.3390/ijerph15010048, 2018.

[21] M. Spitalar, M. Brilly, D. Kos, A. Ziberna, Analysis of flood fatalities–slovenian illustration, Water 12 (1) (2020) 64 https://doi.org/10.4853390/w12010064, 2020.

[22] M. DeKay, G. McClelland, Predicting loss of life in cases of dam failure and flash flood, Risk Anal. 13 (2) (1993) 193–205, https://doi.org/10.1111/j.1539-6924.1993.tb01069.x.

[23] O. Petrucci, L. Aceto, C. Bianchi, et al., Flood fatalities in Europe, 1980–2018: variability, features, and lessons to learn, Water (8) (2019) 1682 https://doi.org/10.3390/w11081682, 2019.

[24] G. FitzGerald, W. Du, A. Jamal, M. Clark, X.-Y. Hou, Flood fatalities in contemporary Australia (1997–2008), Emergency Medicine Australasia 22 (2010) 180–186. https://doi.org/10.1111/j.1742-6723.2010.01284.x.

[25] S.N. Jonkman, I. Kelman, An analysis of the causes and circumstances of flood disaster deaths, Disasters 29 (2005) 75–97, https://doi.org/10.1111/j.0361-3666.2005.00275.x.

[26] E. Boyd, M. Levitan, I. Heerden, Further Specification of the Dose-Response Relationship for Flood Fatality Estimation. US-Bangladesh Workshop on Innovation

in Windstorm and Storm Surge Mitigation Construction, National Science Foundation and Ministry of Disaster & Relief, Government of Bangladesh, Dhaka, 2005, pp. 19–21.
[27] J.A. Duiser, Een Verkennend Onderzoek Naar Methoden Ter Bepaling Van Inundatieschade Bij Doorbraak, TNO Report Ref, vols. 82–644, 1989.
[28] P. Waarts, Methode voor de bepaling van het aantal doden als gevolg van inundatie, Report TNO B-91-1099, 1992, pp. 5–15.
[29] S.N. Jonkman, Overstromingsrisico's: een onderzoek naar de toepasbaarheid van risicomaten, MSc. Thesis, TU Delft, 2001.
[30] S.N. Jonkman, J. Vrijling, A. Vrouwenvelder, Methods for the estimation of loss of life due to floods: a literature review and a proposal for a new method, Nat. Hazards 46 (2008) 353–389, https://doi.org/10.1007/s11069-008-9227-5.
[31] L. Alfieri, F. Dottori, P. Salamon, H. Wu, L. Feyen, Global modeling of seasonal mortality rates from river floods, Earth's Future 8 (2020) e2020EF001541, https://doi.org/10.1029/2020EF001541.
[32] Mario Pinna, Contributo alla classificazione del clima in Italia, Riv. Geogr. Ital. 77 (2) (1970) 129–152.
[33] ADBPO, Authority of Po river basin district, Piano di Bilancio Idrico del distretto idrografico del fiume Po, 2021. https://www.adbpo.it/.
[34] ADBPO, Authority of Po river basin district, Piano di Gestione del Distretto idrografico del fiume Po riesame a aggiornamento al, 2015. https://www.adbpo.it/.
[35] F. Guzzetti, P. Reichenbach, Towards a definition of topographic divisions for Italy, Geomorphology 11 (1994) 57–74, https://doi.org/10.1016/0169-555X(94)90042-6.
[36] P. Salvati, U. Pernice, C. Bianchi, I. Marchesini, F. Fiorucci, F. Guzzetti, Communication strategies to address geohydrological risks: the POLARIS web initiative in Italy, Nat. Hazards Earth Syst. Sci. 16 (2016) 1487–1497, https://doi.org/10.5194/nhess-16-1487-2016.
[37] ADBPO, Authority of Po river basin district, Piano di Gestione Rischio Alluvioni, https://www.adbpo.it/, last access: July 2021.
[38] CORINE Land Cover 2018 (vector/raster 100 m), Europe, 6-yearly, European Union, Copernicus Land Monitoring Service 2018, European Environment Agency (EEA). https://doi.org/10.2909/71c95a07-e296-44fc-b22b-415f42acfdf0.
[39] I. Marchesini, P. Salvati, M. Rossi, M. Donnini, S. Sterlacchini, F. Guzzetti, Data-driven flood hazard zonation of Italy, J. Environ. Manag. 294 (2021) 112986, https://doi.org/10.1016/j.jenvman.2021.112986.
[40] L. Alfieri, J. Thielen, A European precipitation index for extreme rain-storm and flash flood early warning, Met. Apps 22 (2015) 3–13, https://doi.org/10.1002/met.1328.
[41] A. Liaw, M. Wiener, Classification and regression by randomForest, R. News 2 (3) (2002) 18–22. https://CRAN.R-project.org/doc/Rnews/.
[42] B. Yap, K. Rani, H. Rahman, S. Fong, Z. Khairudin, N. Abdullah, An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets, in: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), 2014, pp. 13–22, https://doi.org/10.1007/978-981-4585-18-7_2.
[43] H. Deng, X. Guan & V. Khotilovich : inTrees: Interpret Tree Ensembles. R package version 1.2., https://CRAN.R-project.org/package: inTrees.
[44] H. Deng, Interpreting tree ensembles with inTrees, Int. J. Data Sci. Anal. 7 (2019) 277–287, https://doi.org/10.1007/s41060-018-0144-8.
[45] F. Guzzetti, Landslide fatalities and the evaluation of landslide risk in, Italy. Eng. Geol. 58 (2000) 89–107, https://doi.org/10.1016/S0013-7952(00)00047-8.
[46] P. Salvati, I. Marchesini, V. Balducci, C. Bianchi, F. Guzzetti, A new digital catalogue of harmful landslides and floods in Italy, 19-25 September 2011, in: C. Margottini, P. Canuti, K. Sassa (Eds.), Landslide Science and Practice. Proceedings of the Second World Landslide Forum, Spatial Analysis and Modelling, vol. 3, 2013, pp. 409–414 Rome.
[47] P. Salvati, C. Bianchi, M. Rossi, F. Guzzetti, Societal landslide and flood risk in Italy, Nat. Hazards Earth Syst. Sci. 10 (2010) 465–483, https://doi.org/10.5194/nhess-10-465-2010.
[48] M. Diakakis, G. Deligiannakis, Flood fatalities in Greece: 1970–2010, J. Flood Risk Manag. 10 (1) (2015) 115–123, https://doi.org/10.1111/jfr3.12166.2015.
[49] D.G. Sapir, Natural and man-made disasters: the vulnerability of women-headed households and children without families, World Health Stat. Q. 46 (4) (1993) 227–233 PMID: 8017082.
[50] E. Pradhan, K. West, J. Katz, S. LeClerq, S. Khatry, S. Shrestha, Risk of flood-related mortality in Nepal, Disasters 31 (2007) 57–70, https://doi.org/10.1111/j.1467-7717.2007.00340.x.
[51] K. Alderman, L.R. Turner, S. Tong, Floods and human health: a systematic review, Environ. Int. 47 (2012) 37–47, https://doi.org/10.1016/j.envint.2012.06.003.2012.
[52] ISTAT (Istituto Nazionale di Statistica), https://www.istat.it/it/archivio/104317, last access: March 2021.