ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
INGEGNERIA CIVILE, CHIMICA, AMBIENTALE E DEI MATERIALI

Ciclo XXXVI

08/A1 - Idraulica, Idrologia, Costruzioni Idrauliche e Marittime
ICAR/02 - Costruzioni Idrauliche e Marittime e Idrologia

DATA-DRIVEN MULTIVARIATE FLOOD HAZARD MODELING

Presentata da:
Andrea Magnini

Supervisore:
Prof. Attilio Castellarin

Cosupervisore:
Prof. Alessio Domeneghetti

Coordinatore del Dottorato:
Prof. Alessandro Tugnoli

Esame Finale - Anno 2024

Abstract

In recent decades, advancements in technology have led to a wealth of computational tools for statistical analysis and an abundance of large open datasets. Can we combine multiple open source input information types with the help of statistical analysis techniques to enhance traditional models for flood hazard assessment and modeling? The present Dissertation tackles this issue by showing how these elements can be exploited to improve modelling accuracy in two distinct yet interconnected domains of flood hazard assessment.

In the first Part of the Dissertation, unsupervised artificial neural networks are employed as regional models for sub-daily rainfall extremes. The aim of the models is to learn a robust relation to estimate locally the parameters of a Gumbel distribution for representing the frequency of extreme rainfall depths for any duration in the 1-24h range. The prediction depends on a set of twenty morphoclimatic descriptors. The application of the models focuses on a large study area in north-central Italy, for which an extensive dataset of annual maximum series is available. Validation is performed over an independent set of 100 gauges, where locally fitted Gumbel distributions are used as reference. A conventional approach from the literature, where Gumbel parameters are functions of mean annual precipitation (MAP), is used as benchmark.

Are ANNs effective in RFA for sub-daily rainfall extremes? Is the combination of a set of morpho-climatic indexes helpful for describing the local frequency regime? Our results show that multivariate ANNs may remarkably improve the estimation of percentiles relative to the benchmark approach. Finally, we show that the very nature of the proposed ANN models makes them suitable for interpolating predicted sub-daily rainfall quantiles across space and time-aggregation intervals.

In the second Part of the Dissertation, decision trees are used to combine a selected blend of input geomorphic descriptors for predicting flood hazard (FH). This kind of approaches is commonly referred to as DEM-based in the literature, as it depends on geomophic information retrieved from DEMs. Relative to existing DEM-based approaches, the method proposed here is innovative, as it relies on the combination of three characteristics: (1) simple multivariate models, (2) a set of exclusively DEM-based descriptors as

input, (3) an existing flood hazard map as reference information. What is the potential and accuracy of multivariate approaches relative to univariate ones? Can we effectively use these methods for extrapolation purposes, i.e., FH assessment outside the region used for setting up the model? Is it possible to exploit natural characteristics of these methods for enhancing and completing FH information from the target map?

First, the methods are applied to a wide study area in northern Italy, represented with the MERIT DEM, at ∼90m horizontal resolution. Here, the potential of multivariate approaches relative to the performance of a selected univariate model is assessed and discussed, also on the basis of multiple extrapolation experiments, where models are tested outside their training region. The results show that multivariate approaches may (a) significantly enhance floodprone areas delineation relative to univariate ones, (b) provide accurate predictions of expected inundation depths, and (c) produce encouraging results in extrapolation.

Second, the whole of Italy is studied, and represented with the EU DEM with 25m horizontal resolution. The validation of the proposed model against a mix of multiple sources of independent information confirms the benefits of considering multiple geomorphic descriptors, and shows the potential of DEM-based models for completing the information of imperfect reference flood hazard maps, and the advantages of continuous representation of hazard over binary flood maps.

Contents

List of Figures

List of Tables

Chapter 1

The rationale of the Dissertation

The European Water Charter of the Council of Europe, proclaimed in 1968, stated unequivocally, "*There is no life without water. It is a treasure indispensable to all human activity*". However, when extreme events occur in conjunction with inefficient risk management strategies, water can transform from an essential resource into a significant threat to human life.

The Center for Research on Epidemiology of Disasters (CRED) conducts annual analyses of natural disasters worldwide, providing estimates of their impacts. CRED's observations underscore the alarming frequency of floods as a natural disaster. According to the Emergency Events database, known as EM-DAT and maintained by CRED, nearly half of all recorded events from 2000 to 2019 were attributed to floods (i.e., 3'254 out of 7'348 recorded events, CRED and UNDRR, 2020).

Given the historical role of floods as a threat to human life and an impediment to economic growth, it is essential to recognize that their impact is expected to be even more pronounced in future climate scenarios. In fact, climate change is inequivocably leading to an increase of global mean temperature. As mean temperatures continue to rise, the atmosphere experiences a heightened capacity to hold water, resulting in increased solubility of water vapor. In situations favoring droplet formation, this amplified moisture content leads to larger and more concentrated rainfall within shorter timeframes. Consequently, the incidence of extreme rainfall events is increasing in response to the effects of climate change (e.g., Seneviratne et al., 2012).

This Dissertation is dedicated to two closely interconnected research areas of flood hazard assessment. The first topic is regional frequency analysis of extreme rainfall events. This entails the development of a model to transfer information from gauged sites to ungauged ones. The second topic revolves around flood hazard assessment and

mapping, with an emphasis on geomorphic information and non-physical models.

Despite the apparent differences between these two topics, the same underlying philosophy guides the research conducted in both cases: the objective is to leverage machine learning techniques to construct multivariate models that make use of large and open datasets as input data. These models are designed to address the challenges presented by large and complex study areas. Since the innovative models proposed in the next chapters are designed to combine the aforementioned characteristics, we call them "data driven" and "multivariate".

The primary objective of this research is to explore the potential and limitations of machine learning approaches in two distinct cases within the realm of flood hazard assessment and mapping. Can we exploit the advantages of machine learning techniques for (1) extending the area of applicability of our models, (2) increasing the amount of effective input information, (3) reducing our dependence on non-freely accessible datasets, and (4) decreasing the effort needed for setting up the models? Through the analyses described in the next chapters, a significant contribution to address these research questions is given.

Why should we consider simplified data-driven models in contrast to physically-based models? Indeed, robust knowledge has been achieved about the physical mechanisms that lead to the formation of clouds and peak flows in river discharge. Sophisticated climate models can simulate the formation of clouds and precipitation (e.g., Bonan et al., 2002), hydrological models can reproduce the hydrological cycle from rainfall to river discharge (Hartmann et al., 2014; Singh, 2018), and hydraulic models can propagate peak flows along the river network and reconstruct inundations of the floodplains (e.g., Bates et al., 2010). However, employing most of these models requires numerous assumptions about the underlying physics of the system, and their application demands a profound level of detail regarding boundary conditions, primarily obtained from infrequently available field measurements. Moreover, even if we manage to address the major issues related to model assumptions and input data, setting up these models remains a resource-intensive task. Additionally, computational requirements can be substantial, while the inherent uncertainty in model assumptions and boundary conditions inevitably leads to uncertainties in the results. In summary, when developing models for hydrological applications, large efforts are needed, while output accuracy is not guaranteed (e.g., Mendoza et al., 2015). For this reason, simplified and conceptual models are largely abundant in the literature (e.g., Salvadore et al., 2015; Hartmann et al., 2014; Annis et al., 2020b; Petroselli and Grimaldi, 2018; Manfreda et al., 2015).

Notably, techniques for remote sensing are significantly advancing, providing information and products that are then converted into freely accessible and open datasets (e.g., Yamazaki et al., 2017; Dottori et al., 2021; Gallaun et al., 2019). These datasets are typically in raster format, meaning that the Earth's surface is discretized into pixels. As a result, the horizontal resolution often falls short compared to measurements from field surveys. Nonetheless, remote sensing products and datasets are easily accessible, regularly updated, and validated according to rigorous public standards (e.g., Takaku and Tadono, 2017; Mukherjee et al., 2013; Garcia G., 2015).

Additionally, the availability and flexibility of computational tools and models for artificial intelligence applications are rapidly increasing. Nowadays, a wide variety of methods fall under the popular term 'machine learning.' In essence, it involves training a machine to develop an algorithm to solve a specific problem. By observing input data and following specific directives provided by developers, the machine learns a model to generate an output based on a given input. While physical processes governing the phenomena of interest are implicitly modeled, yet control over the equations used in the model is limited. Nevertheless, machine learning empowers users to construct models based on available data and avoids imposing preconceived assumptions about system behavior (Hastie et al., 2009; Zounemat-Kermani et al., 2021).

This Dissertation presents and discusses innovative multivariate data-driven models for flood hazard assessment and mapping. They can be called "data-driven" and "multivariate", as they largely benefit from the wealth of multiple publicly available sources of information. They leverage the flexibility of machine learning techniques for representing complex phenomena in largely extended and morphoclimatically complex study areas. We do not aim to discuss the potential of machine learning in general, or to extensively compare different machine learning methods. Instead, we are interested in analysing the benefits from using specific techniques in the two fields considered here, that are extreme rainfal regionalization and geomorphic flood hazard modelling.

Our models are compared with benchmarking approaches that are well described in the scientific literature, including both data-driven and physical models. The analysis of our results demonstrates how the proposed techniques, which permit the effective utilization of a broader range of input variables and the coverage of larger study areas, can bring substantial advantages to flood hazard assessment and mapping (Magnini et al., 2022, 2023, 2024).

This Thesis is structured into three main parts. The introductory part provides an overview of key machine learning concepts. Two numbered parts follow. The first Part concerns the regional frequency anaylsis of sub-daily rainfall extremes, and contains an introduction to the problem, a chapter for describing the state of the art, and a chapter showing the application of a machine learning approach to this field. The second Part is about geomorphic flood hazard mapping, and contains an introduction to the topic, two chapters for the state of the art, and a chapter describing the application of ML approaches in this area. Finally, a conclusive chapter summarises the main findings delineated in the first and second Parts.

Chapter 2

An introduction to machine learning

The term "machine learning" refers to a branch of artificial intelligence and computer science which focuses on data-driven algorithms to imitate the way humans learn, with a gradual improvement of accuracy. The term was coined and made popular during the late 1950s by Arthur Lee Samuel, who was an employee at the International Business Machine, a multinational corporation in the field of technology, computer software and hardware. Different articles in the scientific community may not always agree on precise definitions, but in general *artificial intelligence* (AI), whose *machine learning* (ML) is a branch, is considered as a subset of the main field of *statistical learning*. Tipically, in statistical learning one is more interested in the relationships between the input and the output of a specific problem, whereas in machine learning the focus is on the accuracy of the results. The term "deep learning" refers to specific cases of ML, including complex models as artificial neural networks.

Today, ML approaches have been applied to a large number of problems, including text analysis, computer vision, business planning and email filtering, leading to generative models, capable to create new texts and images (e.g, OpenAI, 2023).

In general, ML implies learning from data (Hastie et al., 2009). In a typical scenario, there is an outcome variable, which can be quantitative (such as the inundation water depth) or categorical (such as susceptibility/non-susceptibility to floods), that we wish to predict based on a set of *features*, or *descriptors* (such as elevation of terrain and distance from the river network). The *training set* of data is used to build a prediction model, or learner. In this set, the feature measurements are observed for a set of objects (such as pixels of a raster spatial domain), while the outcome may be either known or unknown, depending on the nature the specific application. After the training phase, the model

can be used to predict the outcome for new unseen objects. A good learner is one that accurately predicts such an outcome. A large number of different models are available, each one with its own advantages. Each model has some internal parameters (or a specific configuration) that is learned during the training phase. In the next Sections, an overview of the procedures for statistical learning is given, and some of the most common machine learning models are described.

2.1 The statistical learning procedure

In general, suppose that Y is the expected quantitative outcome, and $X = (X_1, X_2, .., X_m)$ is a set of m features or descriptors. It can be assumed that the relationship between Y and X has the form

$$Y = f(X) + \epsilon \tag{2.1}$$

where f is some fixed but unknown function of the input features and ϵ is a random error term, which is independent of X and has mean zero. Through machine learning techniques, an estimate $\hat{f}$ of f is found, which yields to the prediction $\hat{Y} = \hat{f}(X)$. Then, it can be demostrated that

$$E[(Y - \hat{Y})^2] = E[(f(X) - \hat{f}(X))^2] + Var(\epsilon) \tag{2.2}$$

where $E[(Y - \hat{Y})^2]$ is the average of the squared difference between the predicted and actual value of Y. Equation 2.2 shows that the error consists of two parts: the first one, $E[(f(X) - \hat{f}(X))^2]$, depends on the estimate of f, and is reducible throught the selection of the most appropriate model. The second, $Var(\epsilon)$, is irreducible, as it depends on a series of external factors, as measurement errors, lack of data, or inappropriate descriptors.

Once the model has been trained, and an estimate $\hat{f}$ of f is obtained, it needs to be tested on another set of data, which is usually called *test set*. Then, it can be applied to different sets of data. Typically, in ML applications one is not interested in the form of $\hat{f}$, which is treated as a "black box". Most of times, the actual aim is to obtain a model capable to generalize, which means predicting accurately with new sets of data.

In general, the best solution is always a balance in the *Bias-variance tradeoff*. *Variance* refers to the amount by which $\hat{f}$ would change if it was estimated with a different training set. *Bias* refers to the error that is introduced by approximating a real-life problem. Thus, using a more flexible model, which means with a higher number of parameters (thus, more complex), usually significantly reduces the bias, but increases the

variance. Initially, increasing flexibility results in faster decrease of bias than increase in variance. Hence, the accuracy for the test set improves. However, at some point the more the model is adapted to the training set (i.e., lower bias), the less its prediction is expected to be accurate with a test set (i.e., higher variance).

In the most simple and common case, the outcome of the target variable (Y, see equation 2.1) is observed and known for the training set, but in several ML applications it is not available. Nevertheless, the problem to be solved by the model can be described as in equation 2.1, yet with unknown Y.

Depending on the nature of the problem, and consequently on the available data, there are three different learning paradigms: supervised learning, unsupervised learning, and reinforcement learning.

2.1.1 Supervised learning

Supervised learning algorithms build a model for a set of data where not only the input features (or descriptors) are available, but also the outcome variable (or variables), which is also called as *target variable*, is present. Thus, having a matrix $A_{n\times m}$ of the training set, whose n rows are the individual samples and m columns are the observed features, the supervised learning algorithms searches for a function $f(A, parameters)$ whose result is as close as possible to the vector Y of the target outcome, with length n. This is achieved by iteratively trying values for the parameters of f until the optimization of an objective function is reached.

The two most common applications of supervised learning are regression and classification, which occur when the target variable is continuous or categorical, respectively. The most common objective function for regression is mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{2.3}$$

where n is the number of elements where the target and input variables have been observed, and x_i the set of features for each element i-th element.
With reference to classification, the most common objective function is accuracy (ACC):

$$ACC = \frac{1}{n}\sum_{i=1}^{n} f_{acc}(x_i) \tag{2.4}$$

where f_{acc} is 1 when the i-th element is correctly labelled, and 0 otherwise.

2.1.2 Unsupervised learning

Unsupervised learning is a more complex problem than the supervised one. It occurs when the available data do not include an observed target variable.

One of its most important applications is *clustering*, which consists of dividing a dataset in a number of subsets characterized by high similarity in the m-dimensional space of the m input features. Several algorithms for clustering are available (e.g., DBSCAN, hierarchical clustering, k-means nearest neighbour), but in general it can be defined as a multi-objective optimization problem. In fact, not only the distance between clusters, but also the number of clusters and their internal homogeneity need to be optimized.

The second most common application of unsupervised learning is density estimation, which consists of generating a synthetic a probability density function to describe a set of data. In this case, the most common method is "maximum likelihood estimation", which means minimizing the logaritmic likelihood (logLH)

$$LogLH = log(\prod_{i=1}^{n}(P(x_i|\theta)) = \sum_{i=1}^{n}(log(P(x_i|\theta)) \tag{2.5}$$

where $P(x_i|\theta)$ is the probability to observe x_i given a set of parameters θ for the probability density function.

2.1.3 Reinforcement learning

The third statistical learning method is reinforcement (Sutton and Barto, 2018). Unlike supervised learning, it does not require the presentation of input/output pairs. Also, in contrast to unsupervised learning, it does not explicitly correct sub-optimal actions. This methodology is primarily applied to problems where an exact solution does not exist. Consequently, the computation and optimization of an objective function with precise values for exact solutions are unattainable.

Typically, this kind of algorithms use dynamic programming techniques, as the learning process is a sequence of trial-and-error steps. Within these algorithms, the learner, or decision maker (called *agent*) takes some *actions* to respond to stimulation from an

interactive *environment*. Based on the *reward* for his actions, the agent learns to improve the rules leading to its actions (*policy*). During this trial-and-error process, reinforcement learning involves a trade-off between exploration and exploitation. In fact, the agent must explore different actions to discover the best strategy (exploration) for maximizing the rewards. In the meanwhile, it must also exploit the current knowledge to make optimal decisions (exploitation).

Some of the most common applications of reinforcement learning are game theory, vehicles with automatic driving systems and robotics.

2.2 Decision trees (DTs)

Decision trees (DTs) are popular supervised ML techniques (Breiman et al., 1984; Hastie et al., 2009), as they are very effective in solving many kinds of classification or regression problems based on an easily-interpretable logic.

DTs search for a relation between input and target output by means of a recursive splitting, which is done through a set of nodes organized in a tree structure. Being the input of a DT a data matrix $A_{n\times m}$, whose n rows are the individual samples and m columns the observed features (or attributes), each node corresponds to a test to be performed on a single attribute in the m columns. Depending on the outcome of the tests on the nodes, the data (originally with n rows) is splitted into two subsets (i.e., two matrices with n_1 and $n - n_1$ rows): each subset is forwarded to one of a set of "child" nodes (see Figure 2.1). Leaves are the last nodes; each one is labeled with an output value, such as a class or a number, representing the tree's output for an input vector that reaches that leaf via the tree's structure.

Training a decision tree consists in determining its structure, the splitting rule on each node, and the labels on the leaves. Most training algorithms operate by recursively splitting the training set, measuring the quality of each partition with object functions that reflect the degree of uniformity of the output values (see Sect. 9.4.2). Repeatedly, tests leading to the best partition are chosen, and child nodes are created accordingly. When some termination criterion is reached, e.g. a set in the partition is perfectly uniform or a maximum depth has been reached, the last nodes become leaves and they are labeled either with the most frequent class value (discrete case, or classification) or with the average of the output values (numeric case, or regression).

An additional useful feature of DTs is that they can predict class probabilities of the input samples. This will be hereinafter referred to as *p-value*. Indeed, the *p-value* can

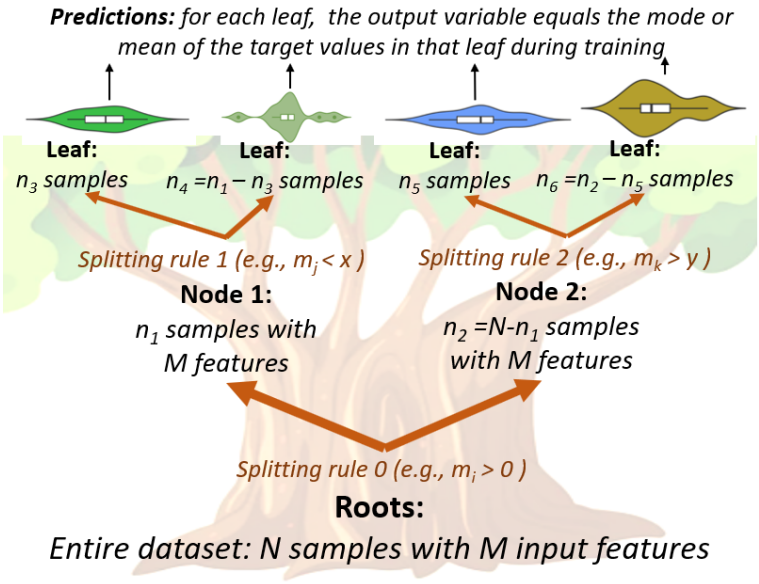

Figure 2.1: Example structure of a decision tree for a given dataset with N samples and M features, having seven nodes in total: one root node, two decision nodes, and four leaves, resulting in an overall depth of three (i.e., longest path from roots to leaves). Figure adatped from Magnini et al. (2022)

be easily obtained for any class in a classification problem, and consists of the fraction of samples of that class in each leaf (computed during the training).

To the aims of the present Dissertation, it is worth underlining also that the number of possible outcomes that the prediction can assume corresponds to the number of leaves. Therefore, the range of target values of a DT is finite, and any regression problem is inevitably approximated by a classification problem, since any continuous numeric interval is "discretized" into a series of fixed values, or *class*, whose number corresponds to the number of the leaves. For each class, based on the frequency of that class over the

The most important parameters to adjust for decision trees are their depth and the minimum number of samples per leaf. Deeper trees contain more splits, leading to less biased models, but usually with a high variance (i.e., bad performances with a validation set of data). Reducing the complexity of a DT, aiming to improve its generalization accuracy, is called *pruning*, and is usually achieved by increasing the minimum number of samples per leaf.

2.3 Artificial neural networks (ANNs)

ANNs are among the most common machine learning models (Hastie et al., 2009). They imitate the structure of human brains, and consist of successive layers, each one

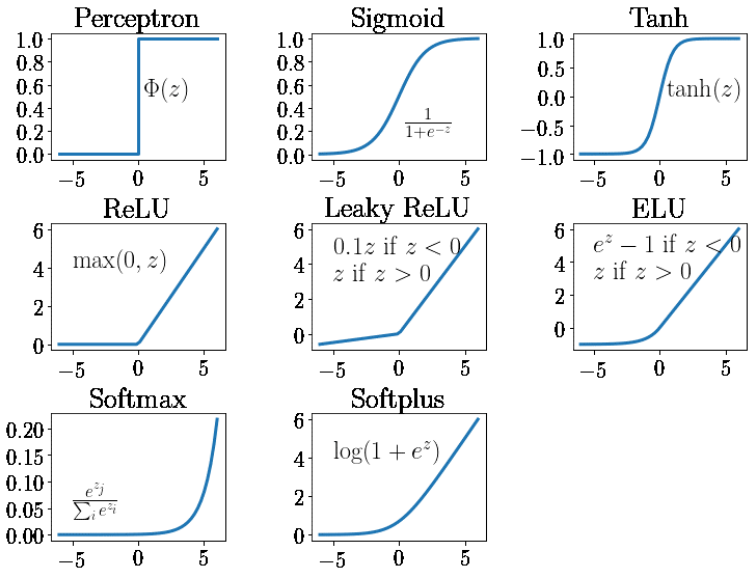

Figure 2.2: Most commonly adopted activation functions for artificial neural networks

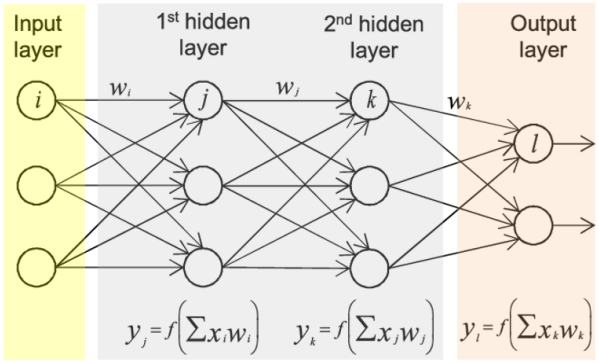

Figure 2.3: Example structure of an artificial neural network, with three input features, two hidden layers, and two target variables

containing a given number of inter-connected units, called neurons (Figure 2.3).

Referring to each single j-th neuron, its output y_j is a function f_j of the linear combination of the n_{input} input factors $x_{i,j}$, as follows:

$$y_j = f_j\left(\sum_{i=1}^{n_{input}} (w_{i,j} \cdot x_{i,j})\right) \tag{2.6}$$

Where w_i are the weights, or coefficients of the input factors, and f_i (usually called the "activation function") can assume a variety of shapes (see Figure 2.2). Usually, the relu or the sigmoid function are adopted (e.g., Han and Moraga, 1995).

The n_{input} input factors for each neuron are the outputs from the previous layer, while for the first layer the input are the features (or descriptors) themselves.

The most common approach for training neural networks is *back-propagation*, also called *gradient descent*. Given a certain objective function, or measure of fit, $R(\theta)$ (usually to be minimized), depending on a set of parameters θ, the gradient descent consists of a series of steps: (1) the loss $R(\theta_0)$ is computed with a trial set of parameters θ_0; (2) the derivatives of $R(\theta)$ with respect to the parameters are computed; (3) the parameters are updated according to the gradient descent (i.e., back-propagation).

The training phase consists of an iterative process, composed of *epochs*. First, the training set is divided into a number of subsamples, called *batches*. At each epoch, each batch in turn is used for observing the data and guessing values for the model's parameters θ (i.e., the weights and the activation functions). Before moving to the next batch, the target variables are computed using data from the previous batch, and the parameter values are updated using the gradient of the fit measure, $R(\theta)$. The increment/decrement value adopted for computing the derivatives of $R(\theta)$ with respect to the parameters is called *learning rate*.

Thus, at any epoch, the parameters are updated as many times as the number of batches. When the number of observations (size) of the batch is between 1 and the total number of the sample, which is the most frequent situation, the optimization algorithm is called *mini-batch gradiend descent*.

With respect to decision trees, ANNs are definitely more complex, as a larger number of parameters need to be defined by the user, which govern both the structure and the learning procedure. First, the size of batches, number of epochs, learning rate and measure of fit need to be carefully set, as they have a significant influence on the accuracy of the final model. Second, the structure of ANNs can be adapted to specific cases by modifiying the number of hidden layers (i.e., the layers between the input and output). This allows to increase complexity and flexibility of the model, reducing the bias (see above the bias-variance tradeoff).

The nature of ANNs make them suitable to solve a large variety of non-linear complex problems. However, they are not easily interpretable, which means that the presence of non-linearity and interconnections among neurons make impossible to follow the path of single features through the ANN layers. Hence, there is not a direct way to estimate the influence of each input feature.

2.4 Bagging

One of the most common approaches to reduce the variance of a ML model is bootstrap aggregation, which is also called "bagging" (Breiman, 1996). In fact, given a set of n independent observations $X_1, X_2, .., X_n$, each one with variance σ^2, the variance of the mean value $\bar{X}$ is σ^2/n: thus, averaging a set of observations reduces variance.
Bagging consists of a series of steps:

1. Generating series of B training sets through bootstrap. Given a sample of n elements, bootstrap consists of randomly selecting n_b elements from the original sample, with the possibility to draw multiple times the same observations. Thus, it is possible to obtain B n_b-dimensional samples from the same population

2. For each sample generated from bootstrap, training a ML model (e.g., a decision tree)

3. Averaging the predictions of the B models

The third step is usually called "stacking", and can be also performed by assuming the mode of the predictions.
Two of the most common applications of bagging concern decision trees, leading to random forests, and ANNs, leading to ensembles of ANNs.

Part 1

Data-driven multivariate regional frequency modeling of sub-daily rainfall extremes

Chapter 3

Introduction to the first Part of the Dissertation

Several hydrological applications, such as the design and management of stormwater drainage systems, combined sewer overflows and flood control systems require an accurate estimation of the design storm (e.g., Claps et al., 2022; Camorani et al., 2005). The latter can be defined as the rainfall depth associated with a given duration and non-exceedance probability (commonly expressed in terms of return period). To produce an accurate estimation of the design storm, modelling the frequency regime of rainfall extremes in the location of interest is needed (e.g., Koutsoyiannis, 2007; Persiano et al., 2020). Timeseries of observed rainfall extremes (e.g., annual sequences of maximum rainfall depths for given durations), when locally available, are in many cases too short to perform robust at-site frequency analysis. This limitation is often addressed by means of regional frequency analysis (RFA), that consists in transferring observed data from other gauged locations to the target site (see e.g., Di Baldassarre et al., 2006; Castellarin et al., 2009; Blöschl, 2011).

In general, RFA consists of two main phases: (i) the delineation of a homogenous pooling-group of sites (i.e., region), containing gauged sites that are similar to the target one, and (ii) the definition of a regional model to transfer the information from the homogeneous region to the target site (Grimaldi et al., 2011). The scientific literature reports on a large number of different methods for conducting RFA of rainfall extremes (see e.g., Svensson and Jones, 2010). The homogeneous region is defined based on some hydrological similarity criteria (see e.g. Castellarin(2001), and can be considered as fixed (i.e., the gauged sites are divided into fixed clusters) or specific for any target site, as in the region-of-influence approach (Burn, 1990). The regional transfer function defined in the second phase of RFA is highly variable depending on the specific approach:

on the one hand, the target variable could be a specific percentile (e.g., Ouali et al., 2016), a parameter of a probability distribution (e.g., Soltani et al., 2017), a statistical moment or L-moment (e.g., Modarres and Sarhadi, 2011; Ngongondo et al., 2011), or the complete time series itself (e.g., Requena et al., 2017, 2018); on the other hand, the observations of the gauged sites could be used alone or with some covariates of the target variable. Literature also reports on several methods that do not require the definition of a homogeneous region (methods based on regression, e.g., Brath et al. 2003, or interpolation, (e.g., Claps et al., 2022).

Regarding the application of RFA to the estimation of flood quantiles, the scientific community has proposed several approaches that make use of advanced artificial intelligence (AI) techniques. Linear and non-linear techniques have been discussed for the definition of a homogeneous region (e.g., Ouarda et al., 2001; Ouali et al., 2016). Models for regional flood frequency analysis can consider many morphological and climatic covariates (Msilini et al., 2022), consider non-linearity of the input-output relations (Ouarda and Shu, 2009), and represent the interaction between the input variables (e.g., Msilini et al., 2020). On the contrary, the literature does not report on many AI-aided RFA methods for modelling the frequency regime of extreme precipitation.

The preference for one model or another strongly depends on the specific case. Moreover, since certain knowledge of the frequency regime of a highly stochastic event is not possible, incontrovertible evaluation of regional models cannot be obtained (Di Baldassarre et al., 2009; Velázquez et al., 2011). However, it is clear which are the characteristics that a good regionalization method should have. First, the ability to profitably use as much recorded information (i.e., observed data) as possible. This is due to the nature of available official gauging networks, that consist of unevenly distributed rain gauges, and timeseries that are often short or fragmented (Libertino et al., 2018; Kidd et al., 2017). As a result, very short sequences are often discarded in regional analysis (e.g., Di Baldassarre et al., 2006). Second, the model should be as flexible as possible. Due to the fast development of technology and science, the amount, location and type of data available is highly dynamic. Thus, a model that can be easily adapted to these changes has a great advantage.

This first part of the present Thesis is divided into three chapters. The first is the present Introduction, the second is dedicated to the state of the art of regional frequency analysis for rainfall extremes, and the third to the application of an innovative machine learning RFA model for rainfall extremes.

Chapter 4

Regional frequency analysis (RFA) of rainfall extremes - State of the art

As said in the Introduction to this Part, several water management infrastructures require the knowledge of the design rainfall, which is the rainfall associated with a given duration of the event and a given non-exceedance probability. In the best case, the design rainfall is obtained through local frequency analysis (LFA). Supposing a long and complete rainfall depth time series is available at the site of interest, LFA consists of a series of steps. First, extreme values are extracted. These can be the single maximum value for each year (annual maxima approach), or all the values that exceed a selected threshold (i.e., peak-over-threshold approach). Second, a probability distribution is fitted on the maximum data. Finally, intensity-duration-frequency (IDF) relationships are derived by extracting quantiles from the probability distributions (see Koutsoyiannis et al., 1998; Brath et al., 2003).

This type of application requires time series referring to different time intervals (generally, between 1 and 24 hours), and its reliability depends on the length of the time series. However, the availability of long and complete time series is limited to a small portion of Earth's surface (Kidd et al., 2017; Libertino et al., 2018). In fact, it is very common to have short time series due to newly installed measure stations, or fragmented records due to interruptions of the measurement.

Thus, regional frequency analysis (RFA) is adopted. It consists of transferring to ungauged sites information from other gauged locations based on some hydrological similarity (Hosking and Wallis, 1997), performing a "space-for-time substitution" (e.g., see Claps et al., 2022). Indeed, this transfer process is very delicate, and requires some hydrological considerations and assumptions. In the following sections, the most widely adopted methods for RFA of rainfall extremes will be described.

4.1 Compilation of extreme values samples

Before performing both local and regional frequency analysis, some pre-processing steps for the rainfall datasets are generally necessary. At a given location, rainfall measurements consist of the depth of precipitated water for a unit surface during time. While for analogic historical devices this results in a continuous line, the modern electronic ones provide observations every 1-2 minutes.

Since for hydrological applications the duration associated with a given precipitation event is a crucial aspect, the measures need to be summed over a moving window of width d over the time axis.

The result is the time series of the rainfall depth associated with a d-wide time interval, which is also called "complete duration series". This is actually different from the *duration* of the precipitation event itself, yet still the most common way to obtain this kind of data. Therefore, the two terms *time interval* and *duration* will be used in the next chapters of this Thesis as synonims, even if this is not properly true (Koutsoyiannis et al., 1998).

The second processing step is to extract the extreme values from the time series for a given duration. As mentioned in the Introduction, the first method is the peak-over-threshold (POT), which consists of selecting all the measurements exceeding a fixed threshold. The major advantage of this method is that various values can be selected for every year, depending on the threshold, leading to longer time series of the maxima, which in turn can provide valuable results when dealing with heavy-tail distributions (Madsen et al., 1997; Marani, 2003). However, the approach is very sensitive to the identification of the theshold (Claps and Laio, 2003).

The second method is the block maxima, which consists of selecting the maximum event over a time period (or block). Commonly, a block of one year is used, leading to the extraction of the annual maximum series (AMS). Following the suggestions of Gumbel (1954), this approach is the most commonly adopted, also due to the wider availability of AMS for older time series.

4.2 Commonly adopted probability distributions and statistics

Once the timeseries of the extreme values are obtained, statistical analysis is performed. Both for LFA and RFA, this usually consists of selecting and fitting the most appropriate probability distribution.

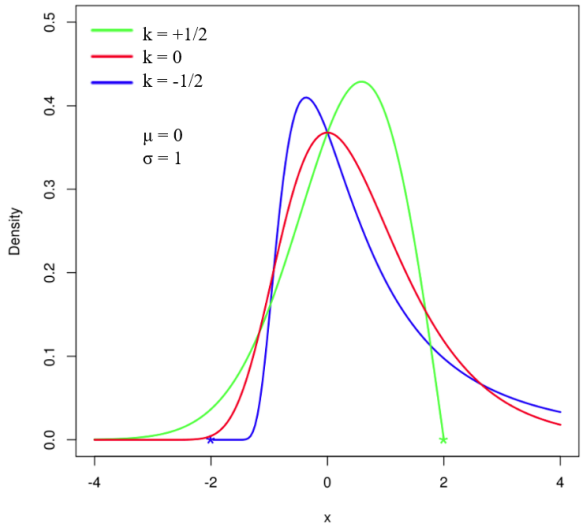

Figure 4.1: Generalized extreme value distributions with different values for the shape parameter k. For all distributions, the mean μ is 0, and the standard deviation σ is 1. Asterisks mark support endpoints

The scientific field that studies the statistical behaviour of the extreme values of a stochastic variable (as rainfall can be considered) is named *extreme value theory*, or *extreme value analysis* (EVA). With specific reference to hydrology, the main founders of EVA were Jenkinson, Gumbel, and later Chow. They showed, in the 50s and 60s, that when the number of observations of annual maxima n (see Section 4.1) tends to infinity, their probability function converges to one of three possible asymptotes. All three asymptotes can be described by a single mathematical expression introduced by Jenkinson (1955) and become known as the *generalized extreme value* (GEV) distribution.

The GEV distribution is characterized by three parameters, that are the location, ξ_{GEV}, the scale, α_{GEV}, and the shape k (see eq. 4.1 and 4.2). Its cumulate probability distribution, $F(x)$, is reported in equation 4.1 following the formulation in Hosking and Wallis (1997) for k:

$$F(x) = -e^{-e^{-y}} \tag{4.1}$$

where

$$y = \begin{cases} -k^{-1}ln[1 - k(x - \xi)/\alpha_{GEV}], & k \neq 0 \\ (x - \xi_{GEV})/\alpha_{GEV}, & k = 0 \end{cases}$$

The shape parameter k, also called the "third parameter" controls the upper tail and the support of the distribution (see Figure 4.2). When $k = 0$, the type I distribution of maxima (EV1), or *Gumbel distribution* is obtained (equation 4.2).

$$F(x) = exp(-exp(-\frac{x - \xi_{EV1}}{\alpha_{EV1}})) \tag{4.2}$$

The Gumbel distribution is widely used in the literature for modeling the AMS statistics (e.g., Piper et al., 2016; Maity, 2018; Caldas-Alvarez et al., 2022), as (1) it has only two parameters, which leads to lower complexity compared to the GEV, and (2) its support is $(-\infty, +\infty)$.

When $k < 0$ the support is $(\xi_{GEV} + \alpha_{GEV}/k) < x < \infty$ and the distribution is called type II or Frèchet extreme value distribution. Finally, when $k > 0$ the distribution is called type III, or reversed Weibull extreme value distribution, and its support is $-\infty < x \leq (\xi_{GEV} + \alpha_{GEV}/k)$. This third case, when the distribution has an upper limit, is the rarest one for AMS analysis.

The range of distributions adopted for EVA in hydrology includes also other distributions, such as the generalized logistic, generalized Pareto, or the lognormal. More details on these distributions can be found in Hosking and Wallis (1997) and Coles (2001), whereas the present Dissertation focuses exclusively on the GEV and its simplified case, the Gumbel, which are the most common (e.g., Svensson and Jones, 2010) and will be be used for the analyses described later.

4.2.1 L-moments

L-moments are statistical properties of measured timeseries and theoretical distributions that became very popular in hydrology, as they can effectively be used for (1) testing the global heretogeneity of a sample (i.e., a group of timeseries, or a supposed homogeneous region), (2) measuring the discordancy of each single timeseries with respect to the sample, (3) measuring the goodness-of-fit of a theoretical probability distribution on a sample or a single timeseries, and (4) fitting a given theoretical probability distribution on a sample or a single timeseries. These concepts are clearly described in Hosking and Wallis (1997), and are widely known in hydrology. Thus, in the present Dissertation only a quick resume is given.

The characteristics of a given timeseries and the shape of a given probability distibution have traditionally been described by their *moments*. These include the mean (μ), which describes the center of location of the distribution, the standard deviation (σ), which describes the dispersion of the distribution about its center, the coefficient of variation ($CV = \sigma/\mu$), the skewness, which describes the asymmetry, and the kurtosis, which describe the propensity to produce outliers (or "tailedness"). In general, moments m_r of order $r > 1$ can be obtained from a timeseries of n measurements (e.g., annual

maxima) by equation 4.3:

$$m_r = n^{-1} \sum_{i=1}^{n} (x_i - m_1)^r \tag{4.3}$$

where m_1 is the mean.

L-moments are an alternative system that arose from the work of Greenwood et al. (1979) and was extensively investigated and described by Hosking and Wallis during the 70s, 80s and 90s. Let $x_{1:n} \leq x_{2:n} \leq ... \leq x_{n:n}$ be an ordered sample, the unbiased sample estimations l_r for the L-moments with order r, λ_r, are defined as:

$$l_{r+1} = \sum_{k=0}^{r} p_{r,k}^{sh} b_k \tag{4.4}$$

where b_k is the unbiased estimator for the probability weighted moment, defined as follows:

$$b_r = n^{-1} \binom{n-1}{r}^{-1} \sum_{j=r+1}^{n} \binom{j-1}{r} x_{j:n} = n^{-1} \sum_{j=r+1}^{n} \frac{(j-1)(j-2)...(j-r)}{(n-1)(n-2)...(n-r)} x_{j:n}$$

and the coefficients $p_{r,k}^{sh}$ are defined as:

$$p_{r,k}^{sh} = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} = \frac{(-1)^{r-k}(r+k)!}{(k!)^2(r-k)!}$$

The most important L-moments are λ_1, equal to the mean, λ_2, or L-scale, $L-CV = \lambda_2/\lambda_1$, or L-coefficient of variation, $\tau_3 = \lambda_3/\lambda_2$, or L-CS or L-skewness, and $\tau_4 = \lambda_4/\lambda_2$, or L-kurtosis. Their informative content on the underlying distribution is similar to the counterparts of the moments (i.e., scale, coefficient of variation, skewness and kurtosis). Nevertheless, the L-moments have the advantage of being fixed, or in a fixed relationship, for the most important 2- or 3-paremeter theoretical probability distributions (see Figure 4.2). This characteristic is extremely useful for selecting the most appropriate probabiliy distribution for a given timeseries (LFA) or a given sample of timeseries (RFA).

More details on the definition and statistical properties of L-moments can be found in Hosking and Wallis (1997). However, comparing equations 4.3 and 4.4 it is evident that L-moments basically consists of linear combinations of the observations, while traditional moments involve non-linear combinations (i.e., exponents > 1). Thus, L-moments are less biased, more robust to outliers and can characterize a wider range of distributions, which makes them particularly attractive for statistical inference (e.g., Di Baldassarre et al., 2006; Schaefer, 1990).

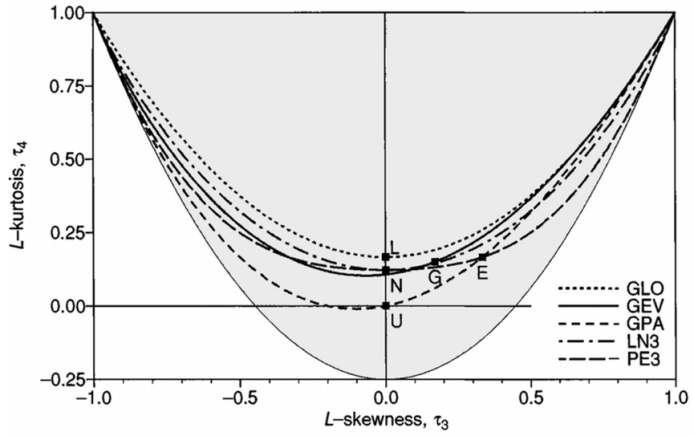

Figure 4.2: L-moments ratio diagram for the most widely used theoretical probability distributions: uniform (U), logistic (L), normal (N), exponential (E), Gumbel (G), generalized logistic (GLO), generalized extreme value (GEV), generalizeed Pareto (GPA), lognormal (LN3), and Pearson type III (PE3). Adpted from Hosking and Wallis (1997)

4.3 Approaches to statistical regionalization

A large variety of regionalization methods exist, but most of them rely on the "index storm approach", which was orginally developed for floods, and named "index flood" (Dalrymple, 1960).

Accordingly, the rainfall depth $h(d,T)$ associated with a given duration d and return period T, is the product of a scale factor m_d, that is called the storm index, and the dimensionless rainfall depth $h'(d,T)$, that is called growth factor (see Equation 4.5). The scale factor is site dependent and is usually estimated by averaging the available measurements at the target station or exploiting regional information through interpolation techniques. The growth factor is derived from a regional relation that is assumed to be valid for the entire homogeneous group of sites, which requires to be defined through RFA.

$$h(d,T) = m_d \cdot h'(d,T) \tag{4.5}$$

A crucial aspect of regional models is the way different durations are handled. Indeed, physical phenomena that origin rainfall with different durations are very different, as convective cloud formation leads to short precipitation, while stratiform clouds cause long events. However, when performing frequency analysis, it is very common to assume that the same probability distribution can be adopted for all the durations, which is called the "simple-scaling assumption" (Burlando and Rosso, 1996). In the case of the storm index method, this lead to a simplified model, where $h'(T)$ does not depend on the duration.

$$h(d,T) = m_d \cdot h'(T) \tag{4.6}$$

A second, fundamental aspect for the classification of RFA models is the variable that is regionalized. A very common approach is to regionalize specific quantiles for a given duration with simple or multi-scaling assumption (Svensson and Jones, 2010; Soltani et al., 2017), while other authors suggest to regionalize the parameters of a target probability distribution, or the sample ordinary moments or L-moments (e.g., Modarres and Sarhadi, 2011; Ngongondo et al., 2011), or the complete time series itself (e.g., Requena et al., 2017, 2018).

The third major element for the classification of RFA models is the way the pooling group of gauged stations is (or not) selected. In fact, since a regional model aims to transfer hydrological information from a pooling group of gauged stations to ungauged ones, some criteria for assessing the reliability of this transfer are necessary. This is usually accomplished by selecting the gauged stations as a homogeneous region, based on some homogeneity and similarity tests. This will be addressed in the next sub-section.

4.3.1 Definition of a homogeneous region

Generally, a homogeneous region in the context of the RFA is defined as the domain where the statistical properties of the investigated variable can be considered constant. This can be declined in two different ways: (1) the same regional function with fixed parameters to estimate the investigated variable; (2) the investigated variable has a fixed valued over all the homogeneous region.

Thus, with specific reference to the storm index method, a given homogeneous region could have either the property of having fixed parameters for a regional function for the growth factor (e.g., $h'(d,T) = f(parameters)$) or a the one of fixed growth factor over all the area.

The procedures for the delineation of a homogeneous region evolved significantly over time. Originally, fixed and non-overlapping geographical regions were adopted. These were defined by splitting the study area into smaller regions based on some morphological or climatological criteria (Cole, 1966). In the following years, it was proposed to pool stations based on some geomorphological descriptors, such as latitude, longitude, elevation or distance from the sea (e.g., Acreman and Wiltshire, 1989). In this way, it was possible to minimize subjectivity in the delineation criteria and take into account a

larger number of driving factors.

This second method became indeed more suitable with the invention and optimization of cluster analysis tools (e.g., Modarres and Sarhadi, 2011; Ngongondo et al., 2011; Le Gall et al., 2022).

A third method, which is called region-of-influence (ROI), was proposed by Burn (1990), and still widely adopted (Svensson and Jones, 2010). It relies on non-fixed regions, as single specific regions are defined for each ungauged site of the study area.

Regardless to the method adopted for the delineation, the heterogeneity of the region (i.e., presence of outliers and discordant data) needs to be tested, as well as the homogeneity with the ungauged sites. To do this, specific test defined by Hosking and Wallis (1997), considering the L-moments ratios (i.e., the dimensionless L-variance, or L-CV, L-skewness, or L-CS, and L-kurtosis) are iteratively used (e.g., Castellarin et al., 2005; Di Baldassarre et al., 2006).

4.3.2 Regionless methods

In some cases, the variable to regionalize has a high variability, so that is not convenient to make any assumption of the homogenenity of a region. This leads to a different class of RFA approaches that can be named "regionless", as they do not require the delineation of a homogeneous region, and that is based on interpolation techniques. These methods are generally used for the estimation of the index rainfall, and can be divided into interpolation and geostatistical methods.

The simplest interpolation techniques is based on Thiessen polygons. Given a certain variable to regionalize, for each ungauged site, the record of the closest guaged site is assigned (Goovaerts, 2000). Other methods include splines or thin plate splines (Carey-Smith et al., 2018), or bilinear surface smoothings (Malamos and Koutsoyiannis, 2016).

Another important method is the inverse distance wieghting method, that assign to ungauged locations a weighted average of the nearby measurements, whose weights decrease with the distance from the target location. Thus, the distance is the only factor considered when regionalizing the information, leading to excessively weighting of the measurements around the target location in some cases (i.e., the "bull's eye effect"). To overcome this major drawback, geostatistical methods, also referred as kriging, have been proposed.

The first kriging techinque was developed by Krige and Matheron (see Matheron, 1963). It takes advantage of the spatial correlation between neighbouring stations to

perform an estimate in an ungauged site. First, the spatial variability of the estimates is described with a *variogram*. This is defined as the half of the variance along the space of the target variable (Z) in two points spaced h, described as in the following equation

$$\gamma(h) = \frac{1}{2}E\{[Z(x+h) - Z(x)]^2\} \tag{4.7}$$

where x is the location and E is the expected value. The kriging technique is a best linear unbiased estimator (BLUE): "best" as it leads to the minimum variance, and "unbiased" as the expected value of the estimation bias error is zero. The linear condition allows the evaluation of the unknown value $\hat{z_0}$ in position x_0 as a weighted average of the nearby measurements

$$\hat{z_0} = \sum_{i=1}^{n} \lambda_i z(x_i) \tag{4.8}$$

where the weights of the interpolation λ_i are defined by fitting a model to the sample variogram (eq. 4.7).

In the simplest case, called ordinary kriging (OK, see Prudhomme and Reed, 1999), the mean value of Z is assumed to be constant but unknown in the neighbour of x_0.

Since OK does not allow a good representation of heterogeneity and anisotropy of rainfall field, other techniques have been proposed. For instance, the presence of significant trends of Z, that break the assumption of constant mean, can be removed, in order to apply the OK on the detrended values (i.e., detrended kriging, or regression kriging Prudhomme and Reed, 1999). In this case, the prediction is obtained with the equation

$$\hat{z_0} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j(x_0) + \sum_{i=1}^{n} \lambda_i \epsilon(x_i) \tag{4.9}$$

where β_0 and β_j are the intercept and coefficients of the regression, X_j are the p independent variables.

The kriging with external drift (or KED, Wackernagel, 1998) and the universal kriging (or UK, first introduced by Matheron, 1969) are special cases of kriging where the mean is not constant over the spatial domain, and a trend is recognized wtih an external variable (in the most common case, the elevation). Usually, the term UK is reserved for the case where the trend (or drift) is modelled exclusively as a function of the coordinates, while in KED it depends also on some auxiliary variables (Hengl et al., 2003). Both KED and UK take advantage of the correlation with an external variable, but the deterministic portion of the approach (the regression) and the stochastic one (the kriging) are modelled simoultaneously (Hengl, 2007).

Cokriging was introduced to improve the estimation of a variable using another spatially correlated variable with better spatio-temporal availability (Ahmed and De Marsily, 1987). It can be considered as an extension of the OK that allows modelling multivariate data with a multivariate variogram and a covariance model. With this approach, measurements may not cover all the sample locations: the data can be available either at the same or at different points for each variable, leading to isotopic (in the first case) or heterotopic (in the second case) datasets. As recommended by Hengl et al. (2003), cokriging should be used to improve the prediction if the number of secondary variables is low and if they are not available at all sample locations

Geostatistical methods perform significantly better than traditional interpolation approach, and are therefore used for several applications of RFA for rainfall (Deidda et al., 2021; Libertino et al., 2018; Svensson and Jones, 2010; Bostan et al., 2012).

Chapter 5

AI-based morphoclimatic regional frequency modelling of sub-daily rainfall extremes

5.1 Introduction

In the present Chapter, the potential of a new AI-based approach to RFA of rainfall extremes is investigated and discussed, mainly following Magnini et al. (2024). It is based on ensembles of unsupervised artificial neural networks (ANNs), that are able to predict the parameters of a selected extreme value probability distribution of the dimensionless extreme rainfall for any duration between 1 and 24 hours.

Following the general framework of the widely adopted storm index method (e.g., Di Baldassarre et al., 2006), the frequency regime of the dimensionless extreme rainfall is regionalized. This is the extreme rainfall depth timeseries divided by its mean value at each site for a given duration. In the present study we focused on the Gumbel distribution, but the approach could consider different models. The proposed method is simple, flexible, and innovative thanks to some characteristics. First, no clustering or target-pooling of available rain gauges is performed: all available annual sequences of maximum rainfall depth are used jointly. Second, no filter on a minimum length of annual sequences is needed and very short sequences (even with two observations) can be used. Third, training is performed simultaneously on all available durations, which leads to advanced interpolation of time-aggregation intervals capability, a very useful feature for practical applications such as the construction of intensity-duration-frequency curves (see Koutsoyiannis et al., 1998; Brath et al., 2003). Fourth, the modelled extreme value distribution can be predicted at ungauged locations within the study region, based on

available morphoclimatic information.

The proposed approach is tested in the present study by implementing it through four kinds of ANNs with increasing complexity. The first makes use of the mean annual precipitation (MAP) alone (MAP-ANN), which is a classical proxy for frequency regime of rainfall annual maxima (see e.g., Schaefer, 1990; Alila, 1999; Castellarin et al., 2009). The second relies on an extended set of twenty morphoclimatic characteristics of the site of interest, including MAP (EXT-ANN). The third (EXT-PCA-ANN) and the fourth (EXT-CCA-ANN) models are fed on pre-preprocessed versions of the same twenty input descriptors, that are obtained through principal component analysis (PCA), and canonical correlation analysis (CCA, see e.g., Di Prinzio et al., 2011).
We make use of a large dataset of gauged stations located in northern and central Italy. For each station, rainfall annual maximum series (AMS) for five different time-aggregation intervals, or durations for the sake of brevity, are available (i.e., 1h, 3h, 6h, 12h and 24h). The maximum length of the AMS series is 90 years. In particular, 2238 stations representing a wide range of morphological and climatic conditions are used to train the four models. Validation of the four regional models is performed using data collected at 100 independent raingauges. The validation considers a traditional RFA method based on L-moments and MAP (e.g., Di Baldassarre et al., 2006) as the baseline regional approach (hereafter also referred to as MAP-Lm), as well as the newly proposed models (i.e., MAP-ANN, EXT-ANN, EXT-PCA-ANN, EXT-CCA-ANN), and compare their predictions with those resulting from at-site frequency analyses (i.e. locally estimated Gumbel distributions).

Finally, the study shows a preliminary application of the proposed EXT-ANN model that adopts the 3-parameter Generalized Extreme Value (GEV) distribution (Jenkinson, 1955). In this preliminary application, the parameter that controls the skewness of the GEV distribution (i.e., the shape parameter) is regionalized through geostatistical interpolation procedure (Hengl, 2007), while the remaining two parameters are derived from the prediction of the EXT-ANN model. Testing the proposed approach for a 3-parameter distribution is important. The scientific literature clearly indicates that in some geographical and climatic contexts the flexibility of a 2-parameter Gumbel distribution, even though widely adopted in previous works (see e.g., Grieser et al., 2007; Svensson and Jones, 2010; Van den Besselaar et al., 2013; Piper et al., 2016; Maity, 2018; Caldas-Alvarez et al., 2022), is not enough for producing an accurate representation of the frequency of rainfall extremes (e.g., Koutsoyiannis and Baloutsos, 2000; Koutsoyiannis, 2004; Papalexiou et al., 2018).

Two main research questions are addressed within this research: (1) are ANNs useful and effective in RFA of rainfall extremes? (2) Are morphological indices helpful in describing the local frequency regime of sub-daily rainfall extremes?

5.2 Methods

The present Section describes the methodologies adopted to set up all the regional models considered in the study. It is divided into two parts: the first briefly summarizes the storm index model with the L-moments approach, that is considered as baseline; the second illustrates the theoretic principles of the ANN approach that originates the four AI-based models. More detail on the models set-up is given in Section 4.

5.2.1 Storm index method with L-moments approach

As said in Chapter 4, the storm index method is one of the most commonly adopted models for RFA (e.g., Brath et al., 2003). Following the original version by Dalrymple (1960), several different applications of the storm index method have been proposed. The MAP-Lm model set up in the present study strictly follows the methods in Di Baldassarre et al. (2006), which in turn strongly relies on the findings of Schaefer (1990) and Alila (1999). These authors studied the extreme dimensionless rainfall depth, which is obtained by dividing the dimensional data by the mean depth associated with the same duration at the same station. They found that statistical moments and L-moments of extreme dimensionless rainfall are in relation to MAP. Thus, homogenous groups of stations can be identified according to their values of MAP, which can be used as a substitute of the geographical location and as proxy of extreme precipitation. Based on the findings by Hosking and Wallis (1997, 1993), L-moments should be preferred for RFA to traditional moments as they are more robust to outliers, can characterize a wider range of distributions and are less subject to estimation bias.

In particular, the MAP-Lm model aims to estimate the regional growth factor of the dimensionless rainfall by means of a Gumbel distribution (i.e., generalized extreme value distribution with zero-value shape parameter; see cumulative density function $F(x)$, eq. 4.2) for each one of the considered durations (i.e., 1, 3, 6, 12, 24h). It is assumed that the whole study area can be described by the same regional laws between local MAP value and statistical moments of rainfall extremes; accordingly, the procedure can be summarized in three major steps: (1) group gauged stations by their MAP values through a moving window of 100 stations; (2) plot the average MAP and L-CV (i.e., L-coefficient of variation); (3) fit the regional function L-CV(MAP) as solution of a least

squares problem. This procedure is applied for each duration to all stations with more than five years of measurement.

In detail, the relationship between L-CV and MAP is modelled by adopting a Horton-type curve (eq. 2).

$$L - CV = a + (b - a) \cdot exp(-c \cdot MAP) \tag{5.1}$$

Once the a, b and c regional parameters are found, L-CV can be obtained in the target location as a function of MAP with eq. 5.1. Then, using the relations described in Hosking and Wallis (1997), i.e., eq. 5.2 and 5.3, the location and scale parameters (ξ and α) of the desired local frequency distribution, which is a Gumbel distribution in this case, are estimated.

$$scale = \alpha = \lambda_2 / ln(2) = (L - CV) \cdot \lambda_2 / ln(2) \tag{5.2}$$

$$location = \xi = \lambda_1 - \gamma \cdot \alpha \tag{5.3}$$

Where λ_1, λ_2 and γ are in this order the L-moment of the first and second order (i.e., mean and the standard deviation of the dimensionless AMS observed) and the Euler's constant (i.e., 1.504).

Despite relying on a single proxy for extreme rainfall, this model has been shown to be rather accurate (Di Baldassarre(2006) over the region for which it was proposed and tested. Moreover, the general framework is rather flexible since local MAP values can be easily retrieved for various regions of the world. For these reasons, it is selected as the baseline for evaluating the performance of the AI-based approaches.

5.2.2 ANN approach

Artificial intelligence has been profitably used by several authors for enhancing RFA for floods (e.g., Msilini et al., 2020; Ouali et al., 2016; Ouarda and Shu, 2009; Shu and Ouarda, 2007), and the accurate choice of models and manipulation of data shows encouraging results. However, the investigation of artificial intelligence techniques for rainfall RFA is not yet well documented.

In the present study, artificial intelligence is used to enhance efficiency in data exploitation and combination. We consider ensembles of unsupervised ANNs; four different models are set up: while fed with different input data, they are based on the same macro-structure. All the ANN models follow four general guidelines: (1) exploiting simultaneously the AMS with all the available durations; (2) using short timeseries as well as long ones; (3) minimizing negative logarithmic likelihood as objective function

(see below); (4) predicting Gumbel distributions as target. Thus, differently from the MAP-Lm model, the timeseries from all the stations are pooled together for training and validating the ANN models. The aim of these ANNs, as for the MAP-Lm model, is to estimate the growth factor of the storm index framework. This is done by finding the best parameters for Gumbel probability distributions of the dimensionless extreme rainfall for any duration at any location, that correspond to the minimum negative logarithmic likelihood.

ANNs are among the most common machine learning models (Hastie et al., 2009), capable of high accuracy in a wide range of problems, inluding hydrological applications (e.g., Mosavi et al., 2018).

In this application case, the activation function f_i (see eq. 2.6) is assumed in the present study as a sigmoid function, as it is commonly done (e.g., Han and Moraga, 1995). The input factors for neurons of the first layer are the descriptors themselves, and each year of measurement for any station and any duration is considered as a single observed element.

Before the training, the available dataset is divided into a "training/testing set" (the larger one), and a "validation set" (the smaller one, used only for the validation). During the training phase, the training/testing set is randomly divided into two subsets: a larger one that will be referred to as "training set" and a smaller one, that is referred to as "testing set". Common proportions of the original training/testing dataset are 80% for the training and 20% for the testing set (e.g., see Xu and Goodacre, 2018). The best hyperparameter set (i.e., the weights $w_{i,j}$ for each i-th neuron) is searched while observing the training set and minimizing the negative logarithmic likelihood function computed on the remaining data included in the testing set (i.e., backpropagation). In this case, equation 2.5 becomes equation 5.4.

$$LogLH = log(\prod_{k=1}^{m}(p(x_k)) = \sum_{k=1}^{m}(log(p(x_k)) \tag{5.4}$$

Where m is the testing set.

Since ANNs are complex and accurate models, they are likely to learn how to perfectly reproduce the training set while being inaccurate with other datasets. This is referred to as overfitting. To improve the generalization ability and stability of a single ANN, an ANN ensemble can be used (see Chapter 2). In the present study, ANNs are generated through bagging, and averaging is used for merging the results. Thus, for each single ANN, the same initial training/testing set is randomly split, as discussed above, so that the optimal hyperparameter set is searched by training the model on the training set

while optimizing the objective function for the testing set. This method is a simple and effective way to obtain ANN ensemble models (see e.g., Shu and Burn, 2004; Shu and Ouarda, 2007).

5.3 Study region and morphoclimatic descriptors

5.3.1 Study region

The methods explained above are applied on a dataset of 2338 gauged locations. These are located in a wide geographical area in northern and central Italy (Figure 5.1), where a variety of climatic and morphological systems can be found. The north is dominated by the Alps, the highest Italian mountain chain, with a mean elevation of 2500 m.a.s.l., and highest peaks until 4800 m.a.s.l. The largest Italian plain, the Po plain, stretches in the southern border of the Alps, following the course of the Po river from the northwest to the northeast, where low coasts are located. The southern border of the Po plain is marked by the Northern Apennines, whose maximum peak is 2165 m.a.s.l. Within the study area, one of the factors that seem to have the largest effect on the precipitation regime is altitude (Allamano et al., 2009; Marra et al., 2021; Mazzoglio et al., 2022).

Three different datasets (Table 5.1) are used in the present study to derive the input information and set up the models described in Section 5.2. First, the Annual Maxima Series (AMS), the variable whose probability distribution needs to be estimated, are retrieved from the dataset I2-RED (Mazzoglio et al., 2020). It includes annual maximum rainfall depths for 1, 3, 6, 12 and 24 consecutive hours from 2338 weather stations across the study area, recorded between 1916 and 2019. While 2238 stations have been selected and used as training/testing set for the five models (i.e. 80% for training and 20% for testing, see Section 5.2.2), the remaining 100 are used as a validation set. The selection of the validation set is based on three main guidelines: (1) to identify a significant number of timeseries, so to have an informative validation set for evaluating the RFA models' performance; (2) to have gauges that are representative of the entire dataset in terms of location, local climate and morphological conditions (see Figure 5.1.c and 5.1.d); (3) to have long timeseries for using as reference validation values the at-site predictions of rainfall quantiles (Figure 5.1.b). Both groups have stations with at least 50 years of measurement (i.e., 248 for the training/testing set and 29 for the validation set, see Figure 5.1.b).

The second dataset used is the multi-error removed improved terrain model (MERIT, see Figure 5.1 Yamazaki et al., 2017), that was used to derive the morphological descrip-

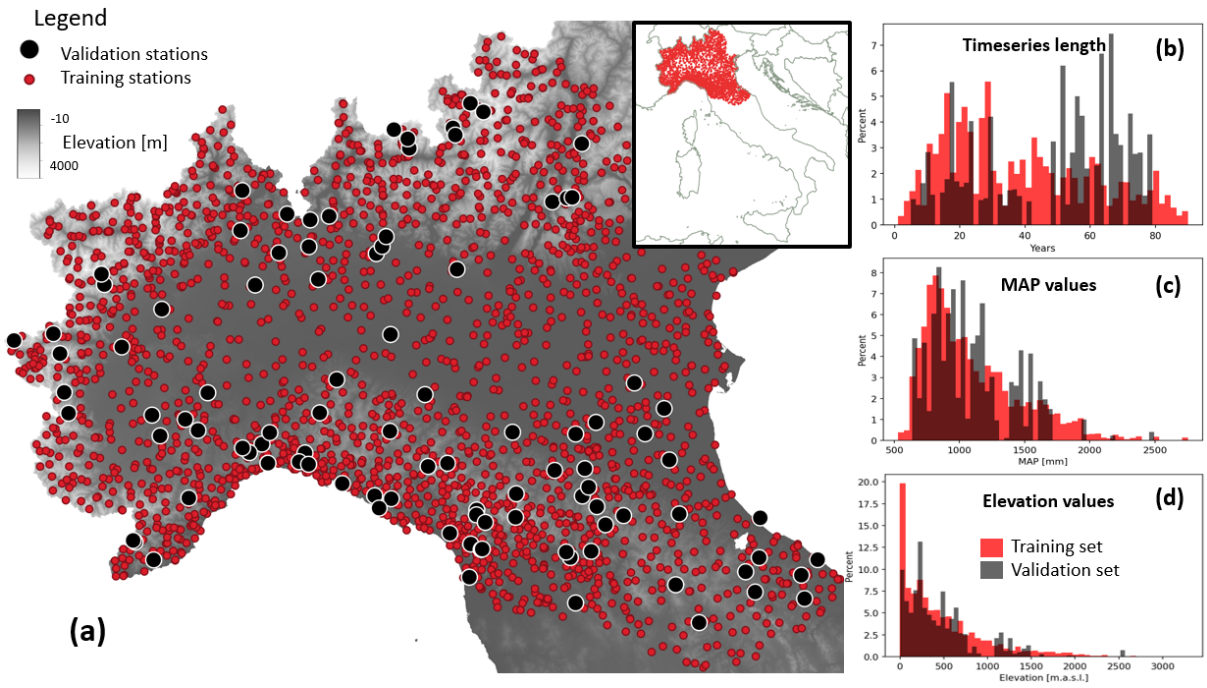

Figure 5.1: Study area, training/testing (red dots) and validation (black dots) raingauges (a); sample frequency distribution (%) of several characteristics for the training/testing (red bars) and validation (grey bars) raingauges: timeseries length (b) for training and validation set, mean annual precipitation (or MAP, (b)), and elevation (d). Adapted from Magnini et al. (2024)

tors (lines 1 to 14 of Table 5.2). Third, the climatic information comes from the ISPRA BIGBANG dataset (Braca et al., 2019), that contains, among other variables, a 1km raster representation of the annual totals of cumulative liquid and snow precipitation over the 1951-2019 time interval.

5.3.2 Morphoclimatic descriptors

The descriptors adopted in the present study were selected based on Mazzoglio et al. (2022), which explored the influence of climatic and morphological descriptors on the statistics of rainfall extremes. They can be divided into three groups: morphological, climatic and geographical. The first group includes descriptors of the elevation, slope and aspect for a buffer area with 1km radius around the station of interest (i.e., from 1 to 6 of Table 5.1), and of the orography and distance between the station and the sea coastlines (i.e., from 7 to 14 of Table 5.1). The climatic descriptors include mean annual rainfall and snow, and their multi-year standard deviation (i.e., from 15 to 18 of Table 5.1). The geographic descriptors consist of longitude and latitude (i.e., 19 and 20 of Table 5.1).

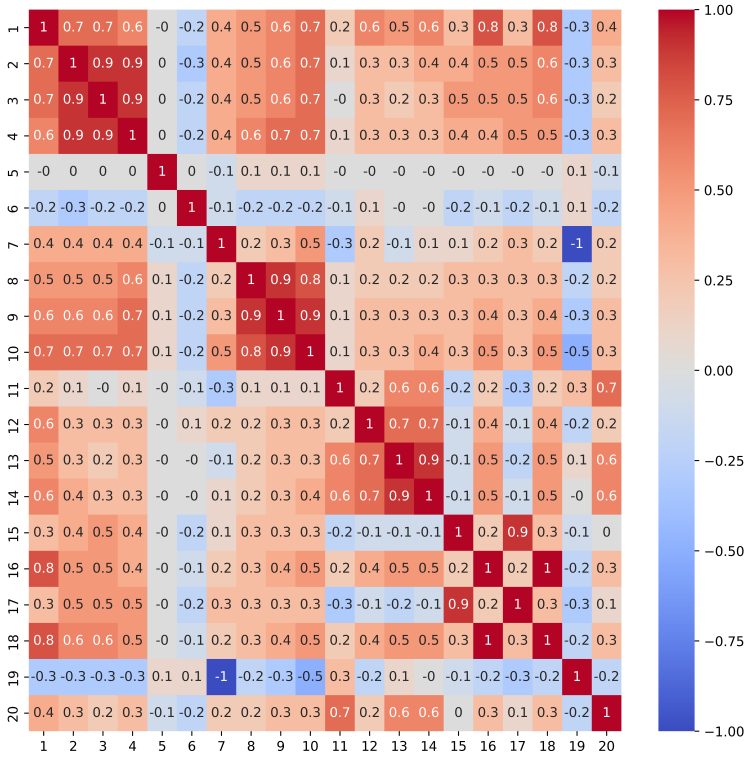

Figure 5.2: Correlation matrix (i.e., matrix whose elements are empirical Pearson correlation coefficients) of input descriptors, reported in the same order as in table 5.1. Adapted from Magnini et al. (2024)

The study descriptors can be obtained through GIS processing procedures of freely available datasets: a digital elevation model (DEM) is needed for retrieving the morphological descriptors, a precipitation dataset for the climatic group, and the coordinate of the gauged stations themselves for the geographical descriptors.

Some descriptors show significant inter-correlation, as illustrated by the correlation matrix depicted in Figure 5.2. Each element X_{ij} of this square and symmetric matrix is the Pearson's correlation coefficient (PCC) of descriptors X_i and X_j. The PCC varies between –1 and 1, and the higher its absolute value is, the higher is the correlation between the two variables to whom it is referred.

$$PCC(X_i, X_j) = \frac{cov(X_i, X_j)}{var(X_i) \cdot var(X_j)} \tag{5.5}$$

Several groups of highly inter-correlated variables are evident (e.g., 1-4, see Figure 5.2), and the mean altitude (descriptor 1) is strongly correlated with most of the other descriptors. This characteristic of the dataset is common, and several authors, as Di Prinzio et al. (2011), showed that pre-processing input datasets by means of Principal Component Analysis (PCA, Jolliffe, 2002) or Canonical Correlation Analysis (CCA,

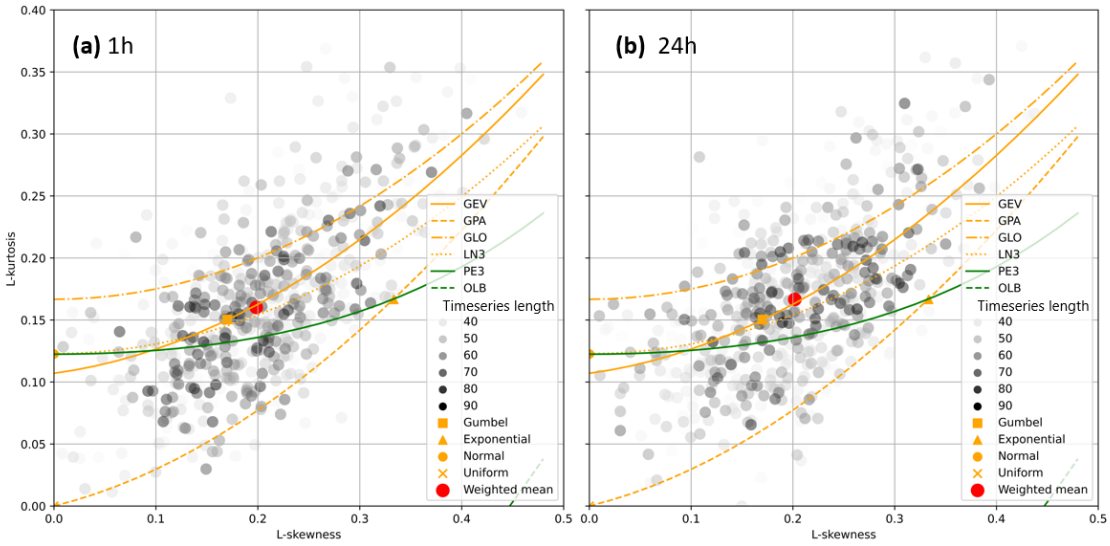

Figure 5.3: L-moments ratio diagram of the 2338 gauged stations (i.e., training/testing and validation set) for annual maximum series with 1h (a) and 24h (b) duration. Adapted from Magnini et al. (2024)

Hotelling 1935), and removing redundant information may improve the training efficiency of data-driven methods and may result in better predictions. To test the convenience of preprocessing techniques for the present study case, PCA and CCA are adopted for two out of the four ANN models trained (see Section 5.4). Finally, climate indexes used in our study (i.e., descriptors 15, 16, 17 and 18) are long-term averages referring to a given climate time-window (i.e. 1951-2019). Hence, we do not consider possible non-stationarities in our study, which is an interesting subject for future developments (see e.g. Persiano et al., 2020, for indications of signals on non-stationarity in sub-daily rainfall extremes for the study region).

5.3.3 Gumbel target frequency distribution

Figure 5.3 illustrates the L-moments ratio diagram (see e.g. Hosking and Wallis, 1997) of the study the rainfall annual maximum series at 1h and 24h durations. Regardless of the duration, a high variability of sample L-moments is evident, as it was expected; the location of the weighted average is a point on the GEV line, very near to the point indicating the theoretical L-moments of the Gumbel distribution.

Thus, the choice of a regional distribution type is restricted to be either a Gumbel or a GEV. Several studies (e.g., Koutsoyiannis and Baloutsos, 2000; Koutsoyiannis, 2004; Papalexiou and Koutsoyiannis, 2013) showed how the latter should be preferred for a better estimation of the upper tail in some geographical and climatic contexts. However, these studies also made it clear that the estimation of the shape parameter of a GEV dis-

tribution is affected by high uncertainty, especially when short time series are used. Due to this reason, several recent studies proposing RFA approaches still resort to the more robust Gumbel distribution (e.g., Svensson and Jones, 2010; Maity, 2018; Ouarda et al., 2019; Caldas-Alvarez et al., 2022). Thus, our study thoroughly assesses the viability of the Gumbel target distribution for the proposed ANN RFA models. Nevertheless, for the sake of generalization, we also briefly present a possible adaptation of the proposed approach to the GEV distribution.

5.4 Regional ANN models

5.4.1 ANN models with Gumbel target frequency distribution

We set up and analyze four different regional ANN models, all consisting of ensembles of ANNs. As the MAP-Lm model (see Sect. 5.2.1), they all aim to produce a regional estimate of the local growth factor for the dimensionless extreme rainfall depth associated with in a given duration (or time aggregation interval). Therefore, they are trained on dimensionless AMS of rainfall depths for a set duration, that are obtained by dividing the original annual sequence by its sample mean (see details later on).

The first model is referred to as MAP-ANN and is fed exclusively on the MAP descriptor. Its application and validation allow it to investigate the effect of exploiting the same input information as in the MAP–Lm method (i.e., MAP) with a different model (i.e., ANN) that exploits all available timeseries of the training set and is simultaneously trained with all available durations (i.e., 1h, 3h, 6h, 12h and 24h). The second model, EXT-ANN, is fed on the extended dataset composed of all the descriptors considered (see Table 5.1). It is an example of both a multivariate approach to RFA of rainfall extremes and a machine learning-based way to exploit multiple input information. The third and fourth models (EXT-PCA-ANN and EXT-CCA-ANN, respectively) make use of preprocessed versions of the same input descriptors of EXT-ANN through PCA and CCA, respectively.

PCA is a statistical technique for reducing the dimensionality of a dataset (Jolliffe, 2002, Di Prinzio(2011). This is accomplished by linearly transforming the data into a new coordinate system where most of the variation in the data can be described with fewer dimensions than the initial dataset. This consists of a change of basis of the original data matrix, and the new dimensions are the principal components (or PCs). Given a set of r variables ($X = (X_1, X_2, ..., X_r)$), the covariance matrix is computed, where each element aij is the covariance between the i-th and j-th variables. The PCs are eigenvectors of the covariance matrix of the original data. The higher the number

of selected PCs, the higher amount of variance of the initial matrix is caught. In this study, the variance chosen for variable shrinking is 80.5%, that corresponds to five PCs: these are the input covariates for the EXT-PCA-ANN model.

CCA (Hotelling 1935, Di Prinzio(2011) is a multivariate analysis technique used to identify the possible correlations between two groups of variables. It consists of a linear transformation of two groups of random variables into pairs of canonical variables, which are established in such a way that the correlations between each pair are maximized. With specific reference to our case, the set of right-hand random variables X consists of the r input descriptors ($X = (X_1, X_2, ..., X_r)$), where r equals twenty). As the left-hand set of variables, the s L-coefficients of variation of the AMS for each station for the analyzed durations are selected ($Y = (L{-}CV_{1h}, L{-}CV_{3h}, L{-}CV_{6h}, L{-}CV_{12h}, L{-}CV_{24h})$, where s is five). The objective of CCA is to construct linear combinations V_i and W_i (called canonical variables) of the variables X and Y, as follows:

$$V_i = A_{i1} \cdot X_1 + A_{i2} \cdot X_2 + ... + A_{i20} \cdot X_{20} \tag{5.6}$$

$$W_i = B_{i1} \cdot L - CV_{1h} + B_{i2} \cdot L - CV_{3h} + ... + B_{i5} \cdot L - CV_{24h} \tag{5.7}$$

where $i = 1, ..., p$, with $p = min(r, s)$. The first weights vectors A1 and B1 maximize the correlation coefficients between resulting canonical variables, under constraints of unit variance. Once the first pair of canonical variables is identified, other pairs can be obtained under the constraint that the correlation between V_i and W_i is 0 (where $i \neq j$). The five canonical variables derived from the canonical transformation of the twenty input descriptors are used as input covariates of the EXT-CCA-ANN model.

The EXT-PCA-ANN and EXT-CCA-ANN models provide the opportunity to assess the effect of PCA and CCA pre-processing techniques. It will be discussed whether the preprocessed input descriptors are able to effectively reproduce the variability of the problem, while the noise in the real signal is absent.

The workflow for setting-up and validating the ensemble ANN models is summarized in Figure 5.4. For a stable training of the ANNs, the input to the four regional ANN models, which include the dimensionless AMS and the considered morphoclimatic descriptors, is standardized (see e.g., Milligan and Cooper, 1988; Jain et al., 2005). Each standardized input ($X_{i,st}$) is obtained from the original input (X_i) by subtracting the regional mean (μ_i) and by dividing this difference by the regional standard deviation (σ_i).

$$X_{i,st} = \frac{X_i - \mu_i}{\sigma_i} \tag{5.8}$$

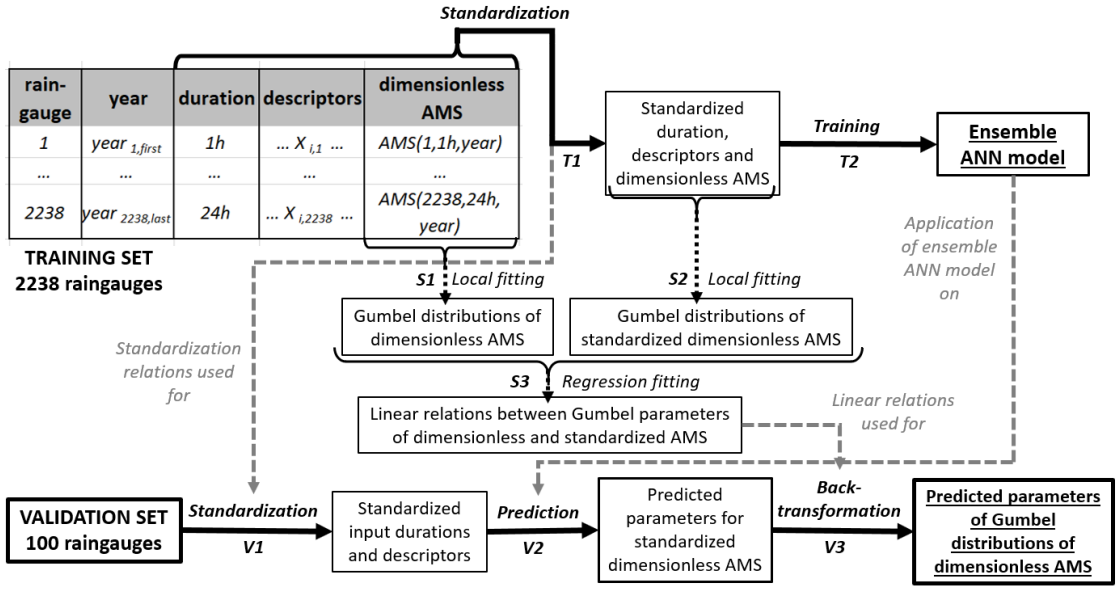

Figure 5.4: Workflow for setting-up and validating the ANN models. Main processes for training (T1 and T2) and validation (V1, V2 and V3) are marked with solid black arrows; side processes (i.e., S1, S2 and S3) marked with dotted black arrows. Models and relations defined in the training phase and used for validation are marked with dashed grey arrows. Adapted from Magnini et al. (2024)

The same relations used for standardizing the training set (with the same μ_i and σ_i, see T1 in Fig. 3), are used also for the validation set (process V1 in Fig. 3). Accordingly, the ensemble ANN models predict Gumbel parameters for standardized and dimensionless distributions at the 100 validation sites, and these parameters need to be back-transformed to the dimensionless space (see process V3, Fig. 3). This is done through two empirical linear relations (i.e., one for the location parameter and one for the scale parameter) between the parameters of the locally fitted Gumbel distribution that model the frequency of the dimensionless AMS and the standardized AMS (see S1, S2 and S3 of Fig. 3).

After some preliminary experiments with different structures of the models, ensembles of 15 ANNs, each one with four layers, were found to be a good balance between prediction accuracy and computational resources required for training.

We tested different proportions for splitting the training/testing set into the training and testing set (e.g., 70%-30% and 80%-20%), but we did not observe significant variations in the results. Thus, we opted for the 80%-20% configuration, since it used a larger amount of data for the training set.

5.4.2 ANN models with GEV target frequency distribution: a preliminary assessment

The approach we propose is general, and can be adapted to any distribution. In particular, the usage of more flexible, 3-parameter probability distributions may allow a better representation of the highest rainfall percentiles. Nevertheless, the estimation of three parameters might be highly uncertain, even when state-of-the-art fitting methods are adopted for at-site frequency analysis (see Section 5.3.3). Hence, in case the proposed ANN approach makes use of a 3-parameter distribution (or even a 4-parameter one), the usage of very short timeseries should be carefully considered both for training the models and for validating them. The findings of these test experiments may be different from case to case. Thus, in this Section we show a preliminary adaptation of the ANN models with the GEV distribution, that has to be intended as a demonstration of the flexibility of the proposed approach.

The GEV distribution is characterized by three parameters, that are the location, ξ_{GEV}, the scale, α_{GEV}, and the shape k (see eq. 4.1 and 4.2 in Chapter 4). The third parameter (i.e., the shape) controls the upper tail of the distribution and its support, and is directly linked to the skewness of the distribution (e.g. expressed in terms of L-CS, as seen in Chapter 4). The remaining two parameters, location and scale, depend on the third one, but are also linked to the first and second L-moments, similarly to the Gumbel case. The GEV distribution and the mathematical relationships between its parameters and the L-moments can be found in Hosking and Wallis (1997):

$$k \approx 7.8590\left(\frac{2}{3+L-CS}-\frac{ln(2)}{ln(3)}\right)+2.9554\left(\frac{2}{3+L-CS}-\frac{ln(2)}{ln(3)}\right)^2 \tag{5.9}$$

$$\alpha_{GEV}=\frac{\lambda_2 k}{(1-2^{-k})\Gamma(1+k)} \tag{5.10}$$

$$\xi_{GEV}=\lambda_1-\alpha_{GEV}\frac{(1-\Gamma(1+k))}{k} \tag{5.11}$$

where Γ denotes the gamma function

$$\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt \tag{5.12}$$

We adapted our EXT-ANN model to the GEV distribution and validated it for the same 100 validation sites by following four steps:

1. The sample L-coefficient of skewness (L-CS) is computed for the timeseries within

the training/test set having at least 30 years of data

2. The sample L-CS values from step 1 are regionalized across the study area by a geostatistical interpolation technique (Hengl, 2007)

3. For any given raingauge belonging to the 100-sites validation set,

   (a) The shape parameter of the GEV is estimated based on the regionalized L-CS value

   (b) The local L-CV value is obtained from the Gumbel scale parameter predicted by EXT-ANN with eq. 5.2

   (c) The L-CV value from step 4 and shape parameter from step 3 are used for deriving the remaining parameters of the GEV distribution

4. The resulting ANN model is compared with a GEV distribution whose shape parameter results from previous step 3.a, while location and scale parameters are fitted using an at-site maximum likelihood procedure

Step 2 adopts the ordinary kriging (OK) method as preliminary analyses using more complex approaches (i.e., kriging with external drift, universal kriging, see Hengl, 2007) did not improve our results. Details on the OK method can be found in Chapter 4 and in several studies (e.g., Shehu et al., 2023; Hengl, 2007). This application of the EXT-ANN model will be hereinafter referred to as EXT-ANN-GEV.

5.5 Performance metrics used in validation

The evaluation of the five models (i.e., baseline, MAP-ANN, EXT-ANN, EXT-PCA-ANN, EXT-CCA-ANN) is conducted by considering three aspects of the models' output Gumbel distributions: the scale parameters, and the 80th and 99th percentiles. The true values are the ones related to the Gumbel probability distributions fitted on the validation dataset with the maximum likelihood method.

Three metrics are computed to evaluate the models' performance from a global point of view: relative BIAS (BIASr), root mean squared error (RMSE), and Pearson's correlation coefficient (PCC, see Section 3). The first two are commonly used in literature (e.g., Msilini et al., 2020; Ghamariadyan and Imteaz, 2021; Shu and Ouarda, 2007), and quantify the systematic error of the models (BIASr) and the gap between the predicted and expected values of the considered variables (RMSE). Differently, the PCC does not take into account the actual values of the variables, since it simply measures the degree of

linearity of the relationship between the empirical reference values and the corresponding model's predictions.

$$BIASr = \sum_{i=1}^{n} (\frac{y_i - y_{i,pred}}{y_i}) \tag{5.13}$$

Where n is the total number of validation observations (i.e., for a given duration, the sum of the total years of annual maxima records n_j over the n_t validation stations):

$$n = \sum_{j=1}^{n_t} n_j \tag{5.14}$$

Each annual maximum is considered as a single i-th observation. Thus, any station is counted as many times as its timeseries length (see also Figure 5.4), and final metrics mainly depend on the longest timeseries. While y_i is the true value of the output variable (i.e., the one related to the fitted Gumbel distributions) for the i-th observation, $y_{i,pred}$ is the value of the output variable obtained with the regional model.

One more metric is computed to evaluate models' accuracy for each single station and each single duration. This is herein referred to as percent relative error (PRE) and defined as in eq. 14.

$$PRE[\%] = \frac{y_{pred} - y}{y} \cdot 100 \tag{5.15}$$

Positive values of PRE represent overestimation with respect to the true values (y), while negative ones account for underestimation.

5.6 Validation of the Regional Models

After training, the five models described in Sections 2 and 4 (i.e., MAP-Lm, MAP-ANN, EXT-ANN, EXT-PCA-ANN and EXT-CCA-ANN) are used to predict Gumbel distributions for the dimensionless annual maximum rainfall at the locations of the validation stations. We considered the performance metrics associated with the estimation of the Gumbel scale parameters and two dimensionless rainfall quantiles, namely the quantiles associated with the 0.8 and 0.99 non exceedance probabilities.

Concerning the metrics for the scale parameters (see Table 5.2), it is possible to observe that RMSE and PCC present a similar behavior across durations and models, which differs from the outcomes in terms of BIASr. Indeed, RMSE and PCC tend to show optimal values for the same regional model, which is often different from the regional model characterized by the smallest BIASr. The value of BIASr is generally between a

few % and 9% for the scale parameters. Overall, the estimation of a regional model for the 1h duration seems to be more complex than for higher durations, as it is pointed out by higher values of the PCC for 12 and 24h. The MAP-Lm model is the least accurate according to the RMSE and PCC metrics, while the MAP-ANN model is the second-to-least accurate, but it is still slightly better than the MAP-Lm. Even though the EXT-ANN, EXT-PCA-ANN and EXT-CCA-ANN models have similar performances, the preprocessing of data leads to a better performance of EXT-PCA-ANN and EXT-CCA-ANN compared to the EXT-ANN model for the 1h duration, while the EXT-ANN is the best one for durations of 6, 12 and 24h.

The metrics computed for the 80th and 99th percentiles are reported in Table 5.2.

It is generally possible to observe a good agreement between the metrics listed in Table 5.2 (scale parameter) and in Table 5.3 (rainfall percentiles). The EXT-ANN, EXT-PCA-ANN and EXT-CCA-ANN models show a similar performance according to all metrics. As for the scale parameter, BIASr values are discordant with the other two metrics, but always very small in all cases (a few % at most). For longer durations (i.e., 6, 12 and 24h) EXT-ANN is the best performing model according to RMSE and PCC; differently, for short durations (i.e., 1 or 3h), the best performing model depends on the metric being considered, but the models with preprocessing usually outperform the EXT-ANN.

Since the overall best performing model is in general achieved by the EXT-ANN model, we carried out a detailed analysis of its behavior. Figure 5.5 reports the geographical distribution of the PRE of the 99th percentile predicted for 1 and 24h durations (panels a and b, respectively). First, no clear geographical pattern of the prediction error is visible: the goodness of the prediction for both 1 and 24h does not seem to be linked to elevation, nor geographical location, and shows similar geographical variability for both durations. Absolute values of PRE (—PRE—) are higher than 50% in one case only for the set of 100 validation locations. Most of stations have low —PRE— values (i.e., —PRE—< 20% for 44 and 43 validation locations for 1h and 24h duration, respectively). The number of validation locations showing 20%<—PRE—<50% for a duration of 1h (i.e., 9) is larger than for a 24h duration (i.e., 6).

The same analysis of PRE for the 99th percentile obtained from the MAP-Lm model is presented in Figure 5.6. As for the EXT-ANN model, no clear geographical pattern of the PRE is observed and most of the stations have PRE between –20% and 20% (i.e., 17+37+35 for 1h, 19+33+37 for 24h). In particular, the stations with PRE between –5% and 5% are 37 for 1h and 33 for 24h (lower numbers when compared to Figure 5.5), while the stations with PRE>20% or <20% are 11 for both time-intervals (i.e., 3+7+1

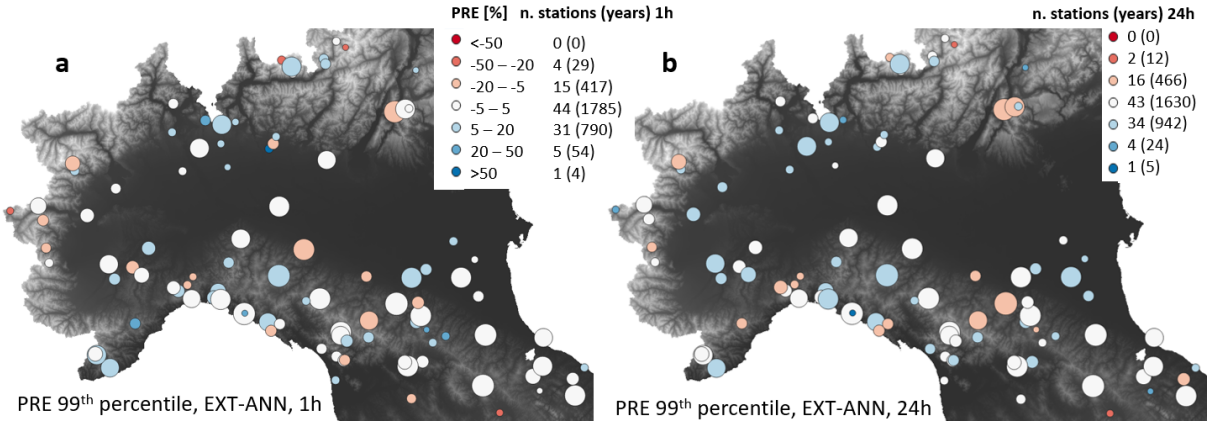

Figure 5.5: Percent relative error (PRE) of EXT-ANN dimensionless 99th percentiles at 100 validation raingauges for 1h (a) and 24h (b) durations; larger circles represent longer annual maximum series. The number of raingauges (overall station-years of data) is reported for each PRE category. Adapted from Magnini et al. (2024)

for 1h and 6+4+1 for 24h, higher values when compared to Figure 5.5).

Finally, the results of model EXT-ANN are used to obtain the location and scale parameters of a GEV distribution (see Section 5.4.2). Table 5.3 reports the global metrics obtained for the 80th and 99th percentiles of dimensionless rainfall depth (columns "EXT-ANN-GEV"). These are relative to empirical predictions of the same percentiles adopting a GEV distribution and the hybrid local/regional estimator described in Section 5.4. While BIASr and RMSE are very similar to the previous case of application with the Gumbel ANN (columns "EXT-ANN" of Table 5.2), PCC values are significantly lower, with the exception of 1h, which has the highest PCC.

5.7 Interpolation across space and time-aggregation interval

One of the most innovative and useful aspects of our AI-based approach is its capability to provide predictions of the dimensionless rainfall distribution in any location (spatial interpolation) and for any time-aggregation interval (i.e., duration) between 1 and 24 hours (time-aggregation interpolation).

As an example, four stations in different geographic and climatic contexts (Table 5.2) are selected for time-aggregation interpolation. Figure 5.7 shows the Depth Duration Frequency (DDF) curves obtained with the EXT-ANN model in the four raingauges. Dimensionality was reintroduced, and consistency among percentiles ensured, by multiplying the predicted dimensionless percentiles by the mean extreme precipitation for

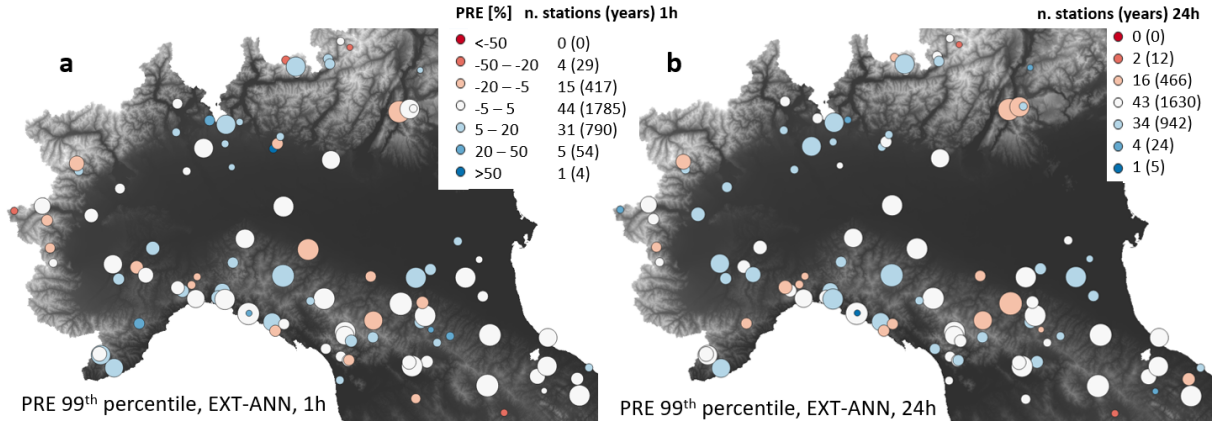

Figure 5.6: Percent relative error (PRE) of MAP-Lm dimensionless 99th percentiles at 100 validation raingauges for 1h (a) and 24h (b) durations; larger circles represent longer annual maximum series. The number of raingauges (overall station-years of data) is reported for each PRE category. Adapted from Magnini et al. (2024)

each duration (eq. 4.5). The latter was obtained by applying the scale-invariance hypothesis to the mean extreme precipitation (see Burlando and Rosso, 1996) and using a power scale law between time-aggregation and mean precipitation. Since the focus of the present study is regional modelling of the growth factor of the storm index method, estimation of the index rainfall (i.e., the mean extreme rainfall depth, see Section 5.2) with multiple scaling is not discussed.

Some observations can be highlighted. First, in stations 9086, 5143 and 16126, the EXT-ANN model has greater accuracy at 1h and 24h, while it is not fully capable of reproducing the fitted model at 6h and 12h; this confirms the metrics in Tables 5.2 and 5.3. Station 17020 is a case of underestimation of the EXT-ANN (see also same station in Figure 5.4, with PRE in -20% - -5% range), where the traditional MAP-Lm model performs better than the ANN-based one. The MAP-Lm approach has similar performances in all four stations: greater errors for longer return periods and longer durations.

Regional ensemble ANN models presented here can produce a spatial interpolation of estimated frequency distributions (i.e., Gumbel distribution in this study) based on the gridded discretization of the study area used for retrieving the local values of the morphoclimatic indices. Panels (a) and (b) of Figure 5.8 show the scale parameters predicted by the EXT-ANN model for the dimensionless distributions of 1h and 24h annual maximum rainfall depths over the drainage area of an Apennine catchment in north-central Italy (i.e., Panaro river basin, drainage area 2300 km2). It can be observed that the average value is higher for 1h duration than 24h (i.e., 0.28 against 0.25). The relation between the predicted scale and the elevation is directly proportional to the

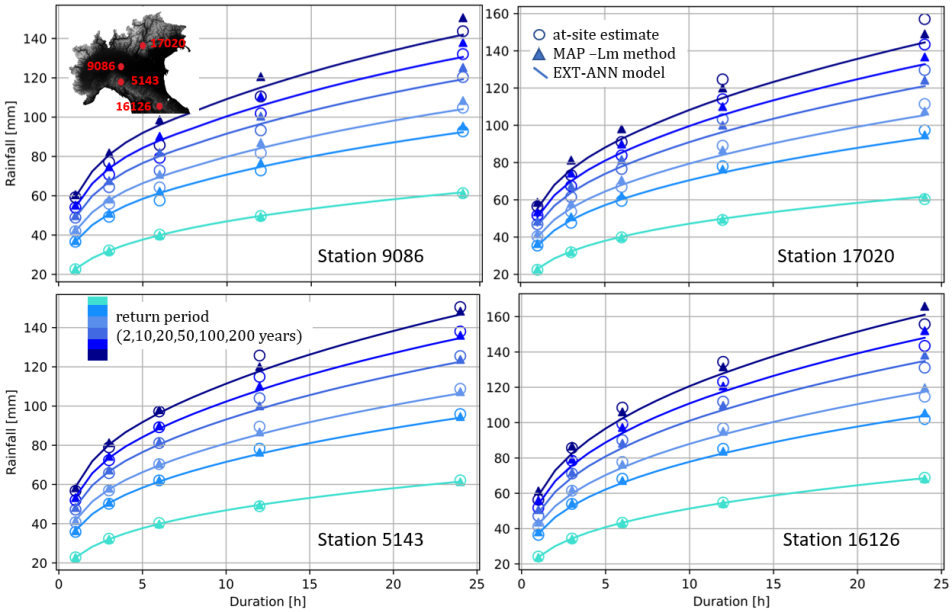

Figure 5.7: DDF obtained with EXT-ANN and MAP-Lm models for stations 9086, 17020, 5143, and 16126 (see also Table 5.2). Adapted from Magnini et al. (2024)

MAP and is further explored in panels (c) and (d). Here, a rather evident decrease of the scale parameter for 1h duration is observed when the elevation is growing, while no clear regression is obtained for 24h. This seems to be in agreement with recent findings of Marra et al. (2021), who observed a significant decreasing trend of the Weibull scale parameters with elevation for sub-hourly durations, while no significant trend was detected for longer time-intervals.

5.8 Discussion

The proposed approach aims at improving the predicting ability of the traditional L-moments storm index model with the action of three combined strategies: (1) exploiting complex non-linear regional functions (i.e., through ANN ensembles), (2) increasing the amount of data used for training the regional models, and (3) increasing the number of proxies for extreme precipitation. The first two points are discussed through the comparison between MAP-Lm and MAP-ANN (subsection 8.1), while the third one regards the EXT-ANN, EXT-PCA-ANN and EXT-CCA-ANN models (subsection 8.2).

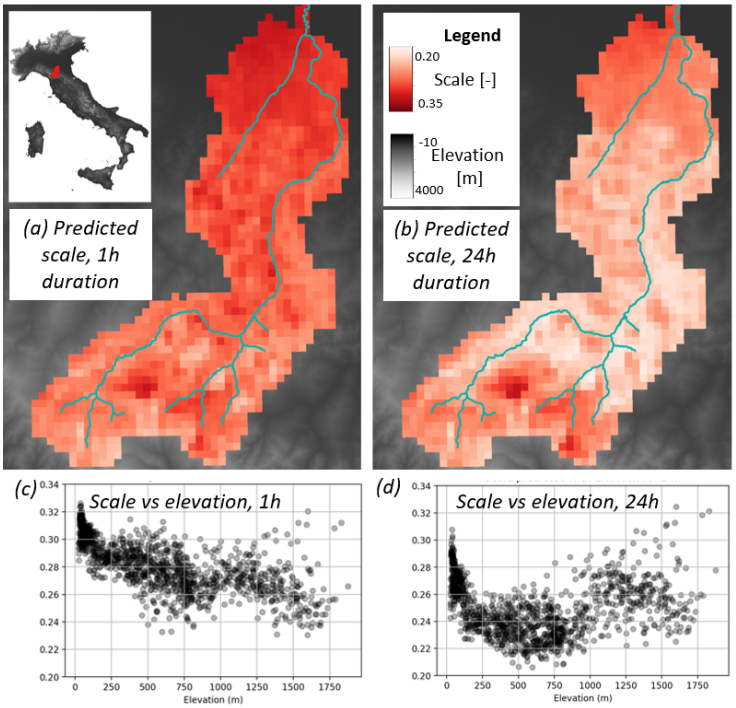

Figure 5.8: Raster-based EXT-ANN prediction of Gumbel scale parameters for 1h (a) and 24h (b), obtained for an example river catchment in the study area (i.e., Panaro river catchment); the main river network is reported in light blue. Scatterplot of scale parameters against elevation values for 1h (c), and 24h (d), from raster-based prediction.

5.8.1 Comparing the univariate benchmark and AI-based models

Metrics in Tables 2 and 3 show that BIASr and RMSE are very similar for MAP-Lm and EXT-ANN, while PCC has significant variations and, more in general, rather low values (i.e., lower than 0.60). This is due to the different nature of the three metrics (see Section 5.4): a small difference in BIASr and RMSE confirms that prediction errors from the two approaches are very similar. Anyway, the higher PCC values (with the exception of 1h time-interval) suggest that changing model type and data management strategy introduces some benefits (i.e., positive correlation between expected and predicted variables).

Overall, it is evident that the MAP-ANN model has unsatisfactory accuracy (i.e., maximum PCC 0.204 for scale with 12h time-interval). However, this could be considered a good result when carefully looking at the regional relation developed within the MAP-Lm framework (see Figure 5.9 for 1h and 24h cases). In particular, the L-CV(MAP) relations found in the present study (see red lines in Figure 5.9) do not show clear and strong relationships between L-CV and MAP, especially for the 24h case. The comparison with an empirical relation reported in the literature (i.e., Di Baldassarre(2006,

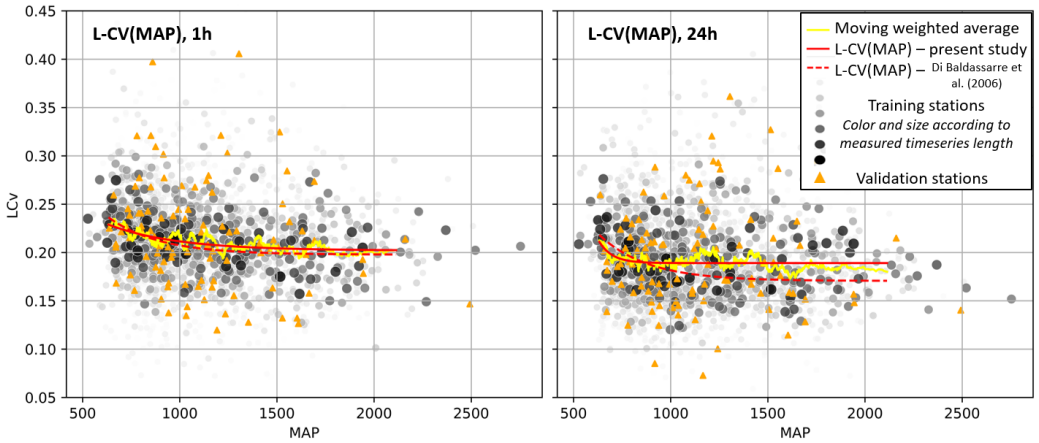

Figure 5.9: Local MAP values and empirical L-CV of 1-hour and 24-hour annual maxima across the study area (dots); moving weighted average (yellow line); Horton-type regional relationship $L-CV(MAP)$ in eq. 2 fitted to the moving weighted average (red solid line) and found by Di Baldassarre et al. (2006) over north-central Italy (red dashed line). Adapted from Magnini et al. (2024)

see dotted red lines in Figure 5.9) suggests that when a large and morphologically and climatically complex region is considered, the classical MAP-Lm approach may not be a viable regionalization strategy. The discrepancies between the relationship identified in this study and that of Di Baldassarre et al. (2006) is more evident for long time aggregation intervals and less pronounced for short durations. In fact, 24h annual maxima are generally associated with frontal disturbances in the study region, and show a complex geographical variability of annual maxima statistics (see e.g. Mazzoglio(2022). Differently, 1h annual maxima are mostly the result of convective storms, which have more spatially homogeneous statistics (see e.g. Schaefer, 1990, Alila 1999, Di Baldassarre et al. 2006).

5.8.2 The multivariate AI-based models

For any time-aggregation interval (i.e., duration), the EXT-ANN, EXT-PCA-ANN and EXT-CCA-ANN models outperform the MAP-Lm and MAP-ANN when referring to the global metrics (Tables 2 and 3). The difference is particularly significant for the PCC, which shows that the EXT- models (that make use of an EXTended set of descriptors) are much more effective in capturing the overall trend of the variables (i.e., location, scale, 80th and 99th percentiles) within the study area. However, the maximum value of PCC for the scale parameter is still very low for 1h (i.e., 0.275), and lower than 0.6 for 24h. This clearly shows that the EXT- models are more accurate when modelling rainfall phenomena with longer durations, as it is also evident from the higher number

of high absolute PRE values (i.e., Percent Relative Error) for 1h (i.e., 10 cases in total out of 100) than for 24h (i.e., 7 cases) for the EXT-ANN (Figure 5.5). A possible reason for these results is that convection phenomena are less influenced by the morpholology (see, e.g., Schaefer, 1990, and Alila, 1999), leading to lower predictive power of most of the input descriptors. Anyway, it is worth highlighting that PRE values <-20% or >20% are mainly observed in Figure 5.5.a in stations where long data series are not available and the expected values (i.e., locally fitted Gumbel distributions) are less reliable, while numerous stations with longer AMS show good results (i.e., PRE values between –5% and 5%).

Also, it is interesting to note that percentiles obtained from the MAP-Lm approach are rather similar to the ones from local data frequency analysis in some cases (see Figure 5.7), even if the model is very simple. However, when looking at the PRE values for the dimensionless 99th percentile (see Figure 5.6), and comparing it with the one predicted with EXT-ANN, the lower accuracy of the former MAP-Lm is evident. First, gauged locations associated with high PRE values for the 99th percentiles are more for MAP-Lm (i.e., 11 for 1h, and 11 for 24h, see Figure 5.6) than for the EXT-ANN (see above). Accordingly, locations associated with lower PRE values (i.e., 37 for 1h and 33 for 24h) are less (i.e., for EXT-ANN, 44 for 1h and 43 for 24h).

Regarding the gridded EXT-ANN predictions (Figure 5.8), it is difficult to objectively compare our results with previous knowledge, as studies on the link between parameters of Gumbel distributions of dimensionless annual maxima and orography are lacking. Marra et al. (2021) observed the relations between scale parameter of Weibull distributions of ordinary events and elevation in Israel. This seems to be aligned with the trends of the linear and non-linear relations found in the present study between the scale parameter and elevation for 1h and 24h (panels (c) and (d) of Figure 5.8).

Since we use a large number of gauged sites that are described by several correlated indices and annual maximum series of very different length, the real signal of the regional function is expected to be disturbed by a certain noise. Thus, the actual impact of preprocessing morphological and climate descriptors is worth analyzing. Multivariate preprocessing techniques generally allow to reduce the noise, so that the models train to reproduce the signal. However, it is remarkable here that data preprocessing (i.e., principal component analysis, or PCA, and canonical correlation analysis, or CCA) seems to have a positive impact just for the 1h duration, while for longer durations the EXT-ANN model performs the best (see Tables 2 and 3). This is a positive result, as it clearly shows how the ensemble ANNs can successfully handle large datasets with several couples of variables that are strongly inter-correlated (see dark colored cells in Figure 2).

In general, it is not possible to understand whether or not preprocessing the input data should be preferred to using the raw dataset, but some conclusions can be drawn. First, preprocessing can impact the performance of the model, in a positive but also negative way. Second, PCA and CCA have similar performances. Third, in absolute terms, all EXT- regional models have very similar performances and prediction accuracy. Fourth, given the similar performance, it has to be mentioned that reducing the dimensionality of the problem and making the models simpler through preprocessing techniques has significant computational advantages in the training phase.

Concerning the preliminary adaptation of the proposed ANN approach to the GEV (i.e., Generalized Extreme Value) distribution, the results are encouraging, as the BIASr and RMSE are very similar to the ones computed for the Gumbel (Table 5.2). However, a correct estimation of the higher order statistical moments, and therefore of shape parameters of 3- and 4-parameter distributions, remains a critical aspect for the use of more flexible probability distributions. The sampling variability of the predictions of GEV parameters used in this study is probably the main reason for the lower PCC values obtained for the EXT-ANN-GEV relative to the EXT-ANN, which indicates a weaker linearity between ANN predictions of dimensionless GEV rainfall percentiles and their empirical validation counterparts. To further investigate this aspect, we computed the same performance metrics by considering only the validation stations with timeseries longer than 40 years. For the sake of brevity, these results are not reported in Table 5.3, but an improvement of the performance was observed, leading to PCC values aligned to those associated with the best performing Gumbel ANN models. This confirms that the record length strongly affects the training and validation of the EXT-ANN-GEV approach. Although we showcase the adaptability of the proposed ANN approach to probability distributions with more than two parameters, dedicated studies are needed for assessing the impact and benefit of short timeseries for the models' training and validation.

In conclusion, the proposed AI-based approach shows satisfactory accuracy relative to classical regionalization methods, and significantly superior performances for time-aggregation intervals equal to or longer than 12h. It also has the advantages of being applicable over a very large study area, and allowing to model any time-aggregation interval between 1 and 24 hours, which automatizes the construction of duration-depth-frequency curves. However, some drawbacks and margins of improvements are still present. First, the accuracy with 1h is still low, which could be improved with a more complex model architecture. Second, some of the input variables (i.e., morphological and climatic descriptors) are not easy to retrieve and compute and not always available, which affects

model applicability. In particular, descriptors related to the distance from the two coasts (i.e., variables 7-14, Table 5.1) can require significant GIS-computational resources to be retrieved for a large number of points. This can be a limit specially in a raster-based or gridded application aimed to produce spatially interpolated maps as in Figure 5.8.

Such a problem could be solved with a reliable ranking of the input descriptors' influence on the final results. However, no direct method exists for input features' ranking in ANNs, and the weights computed for the PCA and CCA are not informative, as in disagreement. Given also the case-specific meaning of this analysis, which is in contrast with the general aims of this study, this point needs to be addressed by future works.

Finally, the present work considered the 2-parameter Gumbel distribution. More flexible statistical models, as the 3-parameter Generalized Extreme Value distribution (Jenkinson, 1955), or the 4-parameter Kappa distribution (Blum et al., 2017) could eventually be used. Preliminary experiments performed with a GEV distribution highlight strong potential of the approach. However, further testing of the robustness of our approach is needed for models with more than 2 parameters in which for instance the skewness (3-parameter) and the skewness and kurtosis (4-parameter) need to be modelled.

5.9 Conclusions

Regional frequency analysis (RFA) is commonly adopted for estimating extreme hydrological variables such as floods or extreme rainfall where local measurements are unavailable or insufficient for at-site frequency analysis. Different approaches have been proposed for the RFA of rainfall extremes, each one characterized by specific advantages and disadvantages (see e.g. Claps et al., 2022). One of the most common drawbacks is that regional models specifically refer to a single duration or a single exceedance probability. Several approaches require the definition of a homogeneous region where the model is trained; this leads to higher accuracy. However, the applicability of the model is then limited to locations that are hydrologically similar to the homogeneous group used in the training. Moreover, most models require filtering the available gauged stations based on the length of the measured timeseries to perform reliable frequency analysis. These aspects lead to discarding a significant amount of data, which could turn out to be detrimental to the accuracy of the regional predictions in some cases.

This study proposes a new approach for estimating the growth factor within the storm index framework for extreme rainfall RFA. This is the dimensionless percentile associated with a given duration and exceedance probability. By multiplying the growth

factor by the mean extreme rainfall with the same duration, the dimensional percentile can be obtained (Dalrymple, 1960). Our approach is based on ensemble unsupervised artificial neural networks (ANNs), that are capable of predicting the location and scale parameters of a Gumbel distribution of the dimensionless rainfall for any sub-daily time aggregation interval (duration) in the 1-24h range.

The study area consists of a large region in north-central Italy, where a wide variety of morpho-climatic contexts are present. From the I2-RED freely available dataset (Mazzoglio et al., 2020), 2338 gauged stations are selected, where measurements of annual maximum rainfall depths are available in the 1931-2019 record period for the 1, 3, 6, 12, 24h time-intervals (i.e., durations). To train the models 2238 gauged stations are used, while the remaining 100 serve as validation set.

Following one approach proposed in the literature (Di Baldassarre et al., 2006) that showed good results over a significant portion of the selected study area, a baseline regional model is developed (MAP-Lm). This consists of a relationship between the L-coefficient of variation and the MAP. Then, four applications of the ANN-based approach are set up: the first (MAP-ANN) makes use of the mean annual precipitation as unique input covariate; the second (EXT-ANN) makes use of an extended number of twenty variables, including morphology (e.g., elevation, slope, aspect, distance from the coast), climate (i.e., mean and standard deviation of snow and liquid precipitation) and geographical coordinates of the stations. The fourth and fifth models (EXT-PCA-ANN and EXT-CCA-ANN, respectively) make use of the same extended dataset, but apply two different preprocessing strategies of the input morphoclimatic covariates (or descriptors), namely principal component analysis (PCA) and canonical correlation analysis (CCA).

This method is innovative for several reasons. First, it does not require the identification of a homogeneous group of sites for model training and application. Second, it uses all available annual maximum data, regardless of the length of the annual sequence (which can be very short, and even two observations). Third, training is simultaneously performed for all durations. These characteristics lead to high interpolation ability, meaning that a single model can predict Gumbel distributions for the extreme rainfall in every point in the spatial domain, and for any duration in the 1-24h range.

The performances of the regional models are analyzed through global metrics (e.g., Pearson correlation coefficient, or PCC), that sum up prediction accuracy over all the validation set, and through the percent relative error (PRE) at each single validation station. Results indicate that the classical approach MAP-Lm appears to have low accuracy when applied to a very large and morpho-climatically heterogeneous region (i.e., PCC 0.1 for 99th dimensionless percentiles of annual maximum depth for a 1h

duration, and PCC -0.1 for a 24h duration). The MAP-ANN method, which uses the same input information (i.e., MAP) but a different approach (i.e., simultaneous use of durations, and ANNs ensembles) shows a slightly better performance, but still low accuracy. When twenty descriptors of the local morphoclimatic conditions are used, ANN-based models show a significant improvement over the MAP-Lm and MAP-ANN (i.e., PCC 0.3 for the 99th dimensionless percentile for 1h, and 0.5 for 24h). Even if the maximum PCC are still low, when considering the local PRE over the 100 validation stations, the improvement of the new approach is evident, as the number of stations with low relative error (i.e., PRE between –5% and 5%) increases (i.e., 44 and 43 for EXT-ANN at 1 and 24h durations, respectively, versus 37 and 33 for the MAP-Lm).

Also, ensemble ANNs show good ability to handle complex and heterogeneous datasets, even without data preprocessing. PCA and CCA seem to have a slight positive effect in modelling short duration extremes, while their impact is limited for longer durations.

In conclusion, based on the outcomes of our study we can affirm that using ensemble ANN models with a few traditional descriptors (i.e., local MAP value as in Schaefer 1990 and similar and more recent regional studies) does not lead to significant advantages over a traditional method (i.e., statistics of extremes rainfall event as empirical functions of local MAP value). However, when combined with multiple morphological and climatic descriptors, the improvement can be remarkable, particularly for annual maximum rainfall depths associated with longer time-aggregation intervals (between 12 and 24 hours in this study). Time and space interpolation ability of the ANNs over the 1-24h range and across the entire study area enable practitioners to directly obtain depth-duration-frequency curves or raster maps of rainfall extremes associated with a given duration and exceedance probability.

Future analyses should build on this preliminary study and address some of the current limitations of the approach. First, methods should be further developed in order to improve the accuracy for extremes originated by convective events. Second more flexible distributions should be considered (e.g., Generalized Extreme Value). Preliminary experiments in this direction produced encouraging results. Third, some additional research should aim at identifying the most effective and descriptive morphoclimatic indices, including alternative or complementary information to the descriptors considered in this study.

Table 5.1: Input descriptors and source datasets used for deriving them (references for MERIT DEM, BIGBANG and I2-RED datasets are Yamazaki et al., 2017; Braca et al., 2019; Mazzoglio et al., 2020, in this order)

Descriptor	Description	Information origin
1	mean altitude within a 1km distance from the gauged location	MERIT DEM
2	standard deviation of the altitude within a 1km distance from the gauged location	MERIT DEM
3	mean slope (I.e., ration between vertical and horizontal distance) within a 1km distance from the gauged location	MERIT DEM
4	standard deviation of the slope within a 1km distance from the gauged location	MERIT DEM
5	mean aspect (I.e., direction of maximum slope) within a 1km distance from the gauged location	MERIT DEM
6	standard deviation of the aspect within a 1km distance from the gauged location	MERIT DEM
7	minimum distance from the Adriatic coast	MERIT DEM
8	mean elevation within the distance between the gauged location and the Adriatic coast	MERIT DEM
9	standard deviation of elevation within the distance line between the gauged location and the Adriatic coast	MERIT DEM
10	Maximum elevation within the distance line between the gauged location and the Adriatic coast	MERIT DEM
11	minimum distance from the Tyrrhenian coast	MERIT DEM
12	mean elevation within the distance between the gauged location and the Tyrrhenian coast	MERIT DEM
13	standard deviation of elevation within the distance line between the gauged location and the Tyrrhenian coast	MERIT DEM
14	Maximum elevation within the distance line between the gauged location and the Tyrrhenian coast	MERIT DEM
15	Mean annual precipitation	BIGBANG dataset
16	Mean annual snow precipitation	BIGBANG dataset
17	Standard deviation of annual precipitation within the 1919-2019 record period	BIGBANG dataset
18	Standard deviation of annual snow precipitation within the 1919-2019 record period	BIGBANG dataset
19	Longitude	I2-RED
20	Latitude	I2-RED

.

Table 5.2: Performance metrics for estimated scale parameter for Gumbel distributions of dimensionless annual maxima at 100 validation points. The best values for each metric are marked with bold characters

	Scale				
Metrics	**MAP-Lm**	**MAP-ANN**	**EXT-ANN**	**EXT-ANN-PCA**	**EXT-ANN-CCA**
1h					
BIASr	-0.075	-0.038	-0.051	**-0.025**	-0.046
RMSE	0.050	0.049	0.050	**0.048**	**0.048**
PCC	0.120	0.105	0.242	0.252	**0.275**
3h					
BIASr	-0.083	-0.046	-0.053	**-0.044**	-0.047
RMSE	0.050	0.047	**0.044**	**0.044**	**0.044**
PCC	-0.031	0.015	**0.422**	0.387	0.394
6h					
BIASr	-0.081	-0.048	-0.047	-0.060	**-0.038**
RMSE	0.050	0.048	**0.043**	0.045	0.044
PCC	-0.037	0.176	**0.453**	0.387	0.384
12h					
BIASr	-0.068	-0.046	**-0.029**	-0.051	-0.030
RMSE	0.049	0.048	**0.042**	0.045	0.045
PCC	-0.063	0.204	**0.463**	0.337	0.350
24h					
BIASr	-0.079	-0.061	**-0.045**	-0.050	**-0.045**
RMSE	0.048	0.047	**0.039**	0.042	0.042
PCC	0.030	0.139	**0.548**	0.421	0.420

Table 5.3: Performance metrics for estimated 80th and 99th percentiles of dimensionless annual maxima at 100 validation points. For each duration, the best case among the models MAP-Lm, EXT-ANN, EXT-PCA-ANN and EXT-CCA-ANN is marked in bold for each metric, while the worst is in italic. The column EXT-ANN-GEV reports the metrics for a demonstration of the adaptability of EXT-ANN to the GEV distribution.

	80th percentile						**99th percentile**					
	MAP-Lm	**MAP-ANN**	**EXT-ANN**	**EXT-ANN-PCA**	**EXT-ANN-CCA**	**EXT-ANN-GEV**	**MAP-Lm**	**MAP-ANN**	**EXT-ANN**	**EXT-ANN-PCA**	**EXT-ANN-CCA**	**EXT-ANN-GEV**
			1h						1h			
BIASr	-0.009	-0.003	-0.004	**0.001**	-0.004	0.008	-0.020	-0.007	-0.013	**0.000**	-0.011	0.022
RMSE	0.044	0.043	0.043	**0.042**	**0.042**	0.049	0.198	0.194	0.196	0.191	**0.189**	0.197
PCC	0.103	0.077	0.240	**0.268**	0.251	0.237	0.105	0.101	0.242	0.257	**0.272**	0.378
			3h						3h			
BIASr	-0.012	**-0.009**	-0.011	**-0.009**	-0.010	0.031	-0.021	**-0.019**	-0.024 -	0.019	-0.020	0.060
RMSE	0.045	0.040	0.038	**0.037**	0.038	0.088	0.201	0.186	**0.172**	**0.172**	**0.172**	0.276
PCC	-0.096	0.034	**0.405**	0.375	0.361	0.202	-0.082	0.020	**0.420**	0.387	0.389	-0.015
			6h						6h			
BIASr	-0.007	-0.009	-0.009	-0.011	**-0.006**	0.030	**-0.013**	-0.019	-0.020	-0.025	-0.014	0.057
RMSE	0.045	0.041	**0.038**	0.040	0.039	0.096	0.206	0.187	**0.171**	0.178	0.175	0.279
PCC	-0.152	0.175	**0.395**	0.319	0.307	0.284	-0.129	0.175	**0.442**	0.375	0.369	0.082
			12h						12h			
BIASr	-0.006	-0.008	-0.005	-0.008	**-0.004**	0.028	-0.011	-0.017	**-0.010**	-0.020	**-0.010**	0.053
RMSE	0.045	0.042	**0.038**	0.041	0.040	0.098	0.205	0.188	**0.169**	0.181	0.178	0.275
PCC	-0.087	0.146	**0.394**	0.242	0.255	0.235	-0.092	0.194	**0.450**	0.318	0.331	0.142
			24h						24h			
BIASr	**-0.006**	-0.010	-0.007	-0.008	-0.007	0.011	**-0.014**	-0.025	-0.018	-0.020	-0.017	0.020
RMSE	0.044	0.041	**0.036**	0.039	0.039	0.062	0.198	0.185	**0.156**	0.170	0.169	0.205
PCC	-0.070	0.103	**0.474**	0.319	0.316	0.444	-0.065	0.134	**0.534**	0.401	0.400	0.201

Table 5.4: Main characteristics of the four stations adopted for the time-aggregation interpolation application through EXT-ANN model.

Code	**Location**	**Record length**	**Mean elevation [m a.s.l.]**	**Minimum distance from Adriatic coast [km]**	**Minimum distance from Thyrrenic coast [km]**	**MAP [mm]**	**MA-Snow [mm]**
9086	Codogno	64	60.28	202.49	98.18	802.58	7.28
17020	Folgaria	58	1114.7	109.60	225.73	1219.16	193.83
5143	Isola di Palanzano Centrale	70	723.13	167.62	43.18	1416.89	39.28
16126	La Verna	76	1080.65	62.29	124.64	1174.25	163.38

Part 2

Data-driven multivariate flood hazard modeling and mapping

Chapter 6

Introduction to the second Part of the Dissertation

Every year flood events worldwide cause vast economic losses, as well as heavy social and environmental impacts, which have been steadily increasing over the last five decades (Jongman et al., 2014; Guha-Sapir et al., 2016), mainly because of the complex interaction between the intensification of extreme hydrological events due to climate change (e.g., Brunetti et al., 2002; Uboldi and Lussana, 2018) and anthropogenic pressure (i.e., land-use and land-cover modifications, see Di Baldassarre et al., 2013; Domeneghetti et al., 2015; Requena et al., 2017). Thus, nowadays, successful flood hazard mapping for flood hazard management is a major task for the whole scientific community (Alfieri et al., 2014; Dottori et al., 2016). Traditional methods to assess fluvial flood hazard rely on hydrological and hydraulic numerical models, whose improvement allows to simulate any scenario for different geometrical or hydrological conditions, obtaining very accurate results (Horritt and Bates, 2002; Costabile et al., 2012; Bellos and Tsakiris, 2016). However, a high amount of hydrologic and hydraulic input information is required to adequately describe the geometry and hydraulic behaviour of the system, thus considerable effort and computation capacity are needed. Consequently, numerical models are unsuitable for high resolution and large-scale applications and in data-scarce regions. To overcome this issue, other mapping techniques have been proposed that take advantage of the wealth of topographic information contained in digital elevation models (DEMs): flood-related geomorphic descriptors (or features, or indices) can be derived from DEMs and used to obtain a measure of flood hazard.

A large variety of descriptors has been tested singularly (see e.g., Manfreda et al., 2015; Samela et al., 2017) or in selected blends (Degiorgis et al., 2012; Gnecco et al., 2017), while some others mixed these indices with information from other sources (e.g.,

Wang et al., 2015; Lee et al., 2017; Khosravi et al., 2018; Arabameri et al., 2019; Janizadeh et al., 2019; Costache et al., 2020). These studies suggest that data-driven flood hazard mapping has a remarkable potential. However, in most of the studies, the reference flood hazard information used to set up the models consists of a dataset of isolated historical events observed in the study area (Lee et al., 2017; Khosravi et al., 2018; Janizadeh et al., 2019; Arabameri et al., 2019; Costache et al., 2020), leading to case-specific prediction skills.

This kind of techniques are highly efficient in terms of computation time, and can be used on very large study areas (even at continental scale, see Nardi et al., 2019). However, they do not capture flow dynamics, and hence their application is limited to large scale and low detail cases. Overall, DEM-based models are very useful as preliminary flood hazard mapping tools in data-scarce contexts and in application to large areas, but cannot yet effectively substitute the traditional models, especially when detailed results are required. Nevertheless, if a strong and reliable relation to derive flood hazard from GDs is obtained, the model could be easily applied in extrapolation to any region where the same relation is supposed to be valid (Tavares da Costa et al., 2020).

This Part of the present Thesis investigates the potential and limitations of DEM-based flood hazard mapping models, with particular attention to the comparison between a single-descriptor model and a multivariate one. First, an innovative DEM-based approach is set-up over Northern Italy, and tested in geographical extrapolation (Magnini et al., 2022), which consists of applying the models on totally new areas with respect to the calibration phase. This represents a strong innovation with reference to the existing literature, where detailed tests on the trasportability of these methods outside the calibration region are not present. Second, the capability of the approach is further investigated nation-wide for Italy, and its natural capabilities are used to solve heterogeneities and inconsistencies of the national official flood hazard map (Magnini et al., 2023).

Chapter 7

Hydrodynamic models for flood hazard mapping

Hydrodynamic or hydraulic models are the most common methods for estimating flood hazard (e.g., Bates and De Roo, 2000; Neal et al., 2012; Horritt and Bates, 2002; Neelz and Pender, 2013), as they enable the user to accurately reproduce the river hydraulics and floodplain inundation dynamics. This is is achieved by the resolution of physical equations through numeric methods. Usually, a significant amount of input data is necessary, that can require long time and big effort for being retrieved. According to their dimensionality, hydraulic models can be divided into 1D, 2D and 3D. While 1D, 2D and 1D-2D mixed models are widely applied for flood hazard estimation, 3D models are adopted for studying vertically stratified fluid properties (e.g., temperature, salinity or sediment transport). Thus, they are commonly used for ecological studies or hydrodynamic simulation of complex hydraulic structures as dams.

7.1 1D hydrodynamic models

One-dimensional models rely on the evaluation of the water volume flowing through the channel cross-sections. Most models solve full 1D Saint Venant equations for conservation of mass and momentum (Neelz and Pender, 2013):

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} = 0 \tag{7.1}$$

$$\frac{1}{A}\frac{\partial Q}{\partial t} + \frac{1}{A}\frac{\partial}{\partial x}(\frac{Q^2}{A}) + g\frac{\partial h}{\partial x} - g(S_0 - S_f) = 0 \tag{7.2}$$

Where Q, t, h, g, S_f, S_0, A, and x are discharge, time, water depth, the gravitational

acceleration, the friction slope, the channel bed slope, the flow cross-section area and the distance between cross-sections, in this order.
The equations are solved by assuming that the flow is parallel to the centerline of the channel and that varies slowly in the cross-sections. In 1D models, the entire riverine system is represented as a succession of cross sections: in each cross-section, the information about the river morphology is used to calculate flow parameters, while the space between two consecutive cross-sections is used as the spatial step for the numerical solution of the equations 7.1 and 7.2.

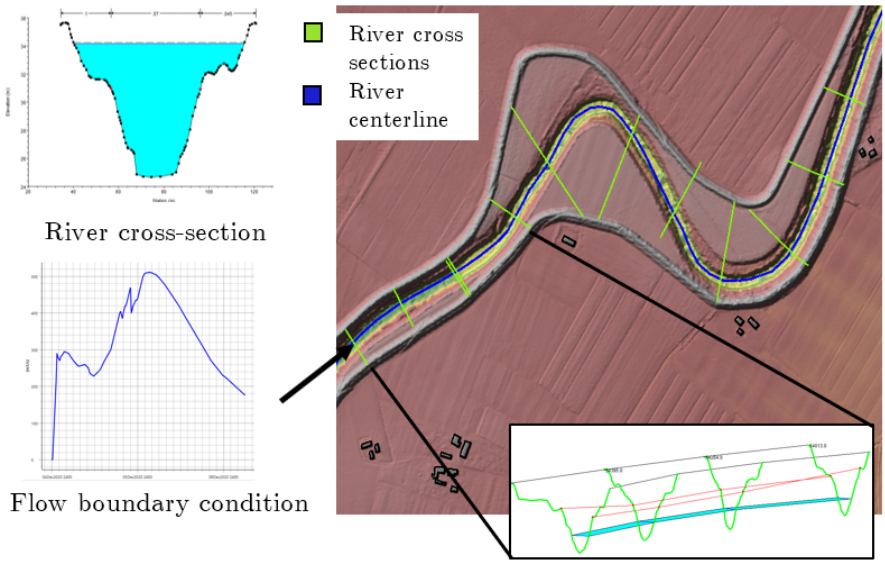

Figure 7.1: Example of 1D geometry and boundary conditions. Upper right panel is a top view on the channel and cross-sections. Upstream boundary conditions are represented in the lower left panel. The flow characteristics computed are shown in the bottom right and top left panels.

Input to 1D models include riverbed characteristics at each cross section (i.e., elevation and roughness of riverbed, banks and adiacent relevant areas) and flow boundary condition (i.e., the hydrograph of the water flow entering the system).

The numerical modelling scheme depends on the specific model. The majority adopts finite-difference method, as HEC-RAS 1D (U.S. Army corps of Engineers), SOBEK 1D Flow, InfoWorks RS 1D, Mike 11 (Danish Hydraulic Institute).
Overall, the 1D numerical scheme leads to significantly lower computation times than 2D models. However, since flow characteristics are computed only within cross-sections, the interaction with hydraulic structures as weir and bridges and the inundation of floodplains can be poorly described.

7.2 2D hydrodynamic models

Two-dimensional models rely on a complete 2D representation of the terrain though a grid composed of a large number of cells. Thus, river dynamics are exclusively driven by the topography by depth averaging Navier-Stokes shallow water equations. These are, again, conservation of mass and momentum:

$$\frac{\partial h}{\partial t} + \frac{\partial (hu)}{\partial x} + \frac{\partial (hv)}{\partial y} = 0 \tag{7.3}$$

$$\frac{\partial hu}{\partial t} + \frac{\partial}{\partial x}(hu^2 + \frac{1}{2}gh^2) + \frac{\partial (huv)}{\partial y} = 0 \tag{7.4}$$

$$\frac{\partial hv}{\partial t} + \frac{\partial (huv)}{\partial x} + \frac{\partial}{\partial y}(hv^2 + \frac{1}{2}gh^2) = 0 \tag{7.5}$$

Where x and y are spatial dimensions, estimates of u, v and h over space and time. Different numerical schemes are commonly implemented for solving these equations over all the cells of the grid, that can assume various shapes according to the model.

Very frequently, a mixed model scheme 1D-2D is adopted (e.g., TELEMAC 2D, Infoworks RS, and HEC-RAS 2D), due to a lower computational demand with respect to only 2D models. This usually consists of using a 1D scheme to obtain flow characteristics within the riverbed, and implementing a 2D model for floodplain inundation wherever it is necessary (see, e.g., Monteleone et al., 2023).

7.3 Flood hazard mapping through hydrodynamic models

The accurate set-up of hydrodynamic models allows to reproduce real evets (e.g., Monteleone et al., 2023) or simulate hypothetical scenarios (e.g., Shustikova et al., 2020). Flood hazard modelling is obtained from flood inundation mapping when a specific non-exceedance probability (or return period) is associated with the flood event. In case a single river is considered, this is typically achieved by adopting a flow hydrograph with the desired return period as boundary condition for the system, or simulating the failure of hydraulic structures. However, when a large spatial domain is considered, where a large number of rivers and minor streams are present (e.g., regional, national or even continental scale), assessing flood hazard is a more complex task. Typically, multiple flood simulations are run in this case, and an ensemble of the results from the single

models (i.e., flow depth and extension of inundated areas) is created.

A relevant example of hydrodynamic flood hazard mapping over a large area is the one adopted by Alfieri et al. (2014), Dottori et al. (2016), and Dottori et al. (2022) for a coverage including the whole European continent. This impressive work was performed through a cascading simulation approach, composed of the following consecutive steps:

1. Distributed hydrological model setup and calibration
2. Simulation of a long-term discharge time series and derivation of peakflows with selected return period
3. Downscaling to 100 m spatial resolution and derivation of design flood hydrographs
4. Floodplain hydraulic simulations
5. Merging of output flood depth single maps

Before starting the hydrodynamic modeling phase, the research for obtaining the input design flood hydrographs for the hydrodynamic models is conducted (steps 1, 2 and 3). First, the LISFLOOD software (Van Der Knijff et al., 2010) with European coverage at a grid resolution of 5km is calibrated (step 1). Second, a 21-year continuous discharge time series is generated using the calibrated model (step 2). The meteorological input data for LISFLOOD is obtained by combining point measurements from the Monitoring Agricultural Resources agro-meteorological database (Rijks et al., 1998), the World Meteorological Organization (WMO) synoptic observations (`http://www.wmo.int/pages/prog/www/`), and the German Weather Service (`http://www.dwd.de/`) network.
In step 3, the output of the hydrological model is downscaled at the resolution of 100m, and the generated time series of flow discharge evaluated statistically by fitting extreme value distributions on their annual maxima (i.e., Gumbel distributions) for each pixel of the raster domain falling within the river network.
In step 4, design flood hydrographs from step 3 are then used to perform small-scale floodplain hydrodynamic simulations every 5 km along the river network using the 2D modeling LISFLOOD-FP software (Bates et al., 2010). For this step, only river sections with a contributing area larger than 500 km^2 are considered. Despite this semplification, a large number of single models are set-up and run (i.e., more that 37'000 in Alfieri et al., 2014).
Finally, the single inundation maps from step 4 are merged for obtaining a single hazard map for the whole Europe, whose return period corresponds to the one selected in step 3.

As evident from the described cascade process, the effort required for hydrodynamic flood hazard assessment is significant, and its amount exponentially increases for very extended study areas. Nevertheless, simplifying assumptions are always necessary (e.g., the resolution of 5km for the generation of flow discharge timeseries, or the minimum contributing area of 500 km^2), and can introduce significant inaccuracy in the results, which can be either justified or not based on the specific application case. An evident example is Figure 7.2, where the flood map obtained by Alfieri et al. (2014) with a return period of 500 years is showed for Italy, and the lack of flood hazard information for a large portion of the river network is striking.

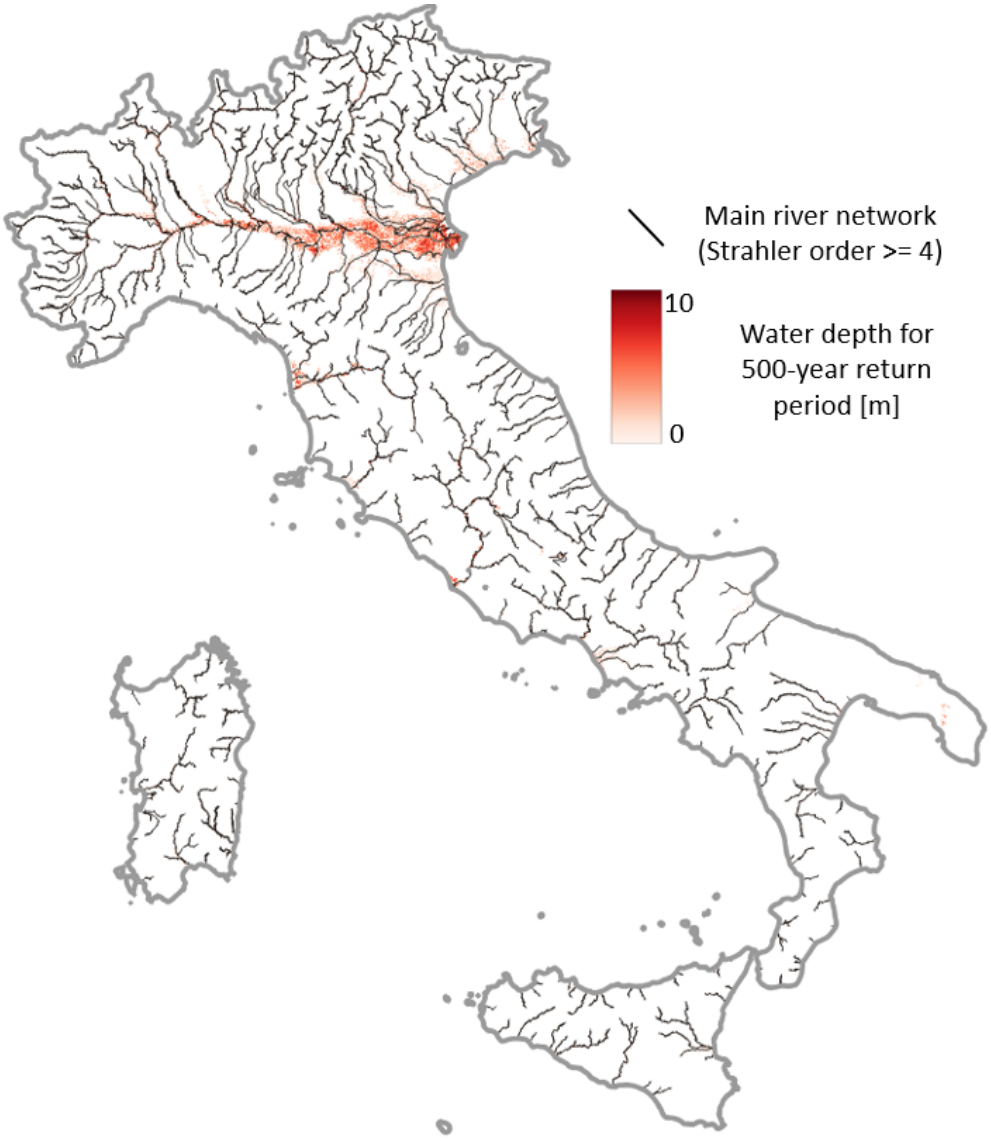

Figure 7.2: Water depth (red scale colors) from flood hazard map with 500-year return period from Alfieri et al. (2014) over Italy. In black, main river network

Chapter 8

DEM-based flood hazard modeling and mapping

Since a large amount of hydrologic and hydraulic input information is required for their set up, numerical models are often unsuitable for large-scale applications and data-scarce regions. Thus, alternative mapping techniques have been proposed which mainly rely on topographic information contained in digital elevation models (DEMs). Important advantages of DEM-based flood hazard mapping methods are their flexibility and, in principle, their general applicability to any flood-prone area where a reliable DEM is available, as well as their low computational costs relative to numerical models. However, two main drawbacks must be highlighted: first, DEM-based methods do not consider the water dynamics, and second, they need a pre-existing reliable reference flood hazard map, which may or may not be available for the area of interest.

During calibration (or training) of DEM-based models, a function f is learned between a certain number of descriptors of terrain morphology, used as indepentend variables (or covariates GDs) and flood susceptibility f_s.

$$f_s = f(GDs, parameters) \tag{8.1}$$

The type of function, the number and nature of the GDs, and the way the optimal parameters of the functions are found depend on the specific DEM-based model. All DEM-based models need some flood hazard information as reference (or target) for the calibration. During this phase, the best parameters are found based on the optimization of an objective function that represents the accuracy in reproducing the reference flood hazard information.

One of the most common ways to classify the DEM-based models is the number of GDs used as covariates, which leads to the definition of univariate and multivariate models.

8.1 Univariate DEM-based models

The first DEM-based approaches proposed in the literature consider a single geomorphic descriptor, or index (see e.g., Williams et al., 2000; Noman et al., 2001; Dodov and Foufoula-Georgiou, 2006; Nardi et al., 2006; Manfreda et al., 2011, 2014, 2015; Samela et al., 2017; De Risi et al., 2018), that is used as a binary classifier to distinguish between flood-prone and flood-free areas through the definition of a threshold value. In this case, the function f of the GD is very simple, and the only parameter is the threshold value (th):

$$f(x) = \begin{cases} \text{flood-prone}, & \text{if } x \geq \text{th} \\ \text{non-flood-prone}, & \text{otherwise} \end{cases}$$

Where x is the value of the selected geomorphic descriptor at a specific point of the study area (typically, a pixel within a raster domain), and the $\geq$ character could be either $>$,$\leq$, or $<$, depending on the case.

The optimal threshold value, th, is identified by means of an iterative calibration procedure which optimizes the agreement of the binary map with reference pre-existing flood hazard information.
Several authors (see e.g., Manfreda et al., 2015; Samela et al., 2017) highlight that the performance of the considered geomorphic descriptor (or index) can change according to the geographical context of the application.

8.1.1 Geomorphic descriptors (GDs)

Topographical rasterized information contained in DEMs can be used to extract GDs adopting several algorithms available in the literature (e.g., Tarboton et al., 1991). These descriptors vary spatially, assuming different values for different pixels within the domain, while being constant in time.
They can be divided into two broad categories: (1) single features, if they represent simple terrain characteristics, and (2) composite indices, if they are derived based on a combination of other features. Several descriptors were proposed and compared by the authors as single covariates for DEM-based models (see e.g., Manfreda et al., 2015; Samela et al., 2017). Some of the most popular and effective can be found in the list below, which consist first of three single indices, and then three composite:

1. **Local slope (sd8)**, estimated for each cell as the maximum slope among the eight

possible flow directions and computed as the ratio between the vertical and the horizontal differences. It can be defined either in radiants or in degrees

2. **Upstream contributing area ($\mathbf{A}_d$)**, also called "accumulation area". It can be either in m^2 or in number of drained pixels

3. **Horizontal distance from the nearest stream (D)**, defined as the length of the path that hydrologically connects each cell to the nearest cell of the river network (Figure 8.1). It can be either in m or in number of pixels.

4. **Height above the nearest drainage (HAND)**, defined as the vertical difference between a given cell and the hydrologically nearest cell belonging to the river network (Rennó et al., 2008, , see Figure 8.1). As the D, either in m or in number of pixels.

5. **Modified topographic index ($\mathbf{TI}_m$)**, derived from the modification proposed by Manfreda et al. (2008) to the index originally introduced by Kirkby (1975), and defined as follows:

$$TI_m = \ln\left(\frac{a_d^n}{\tan(\beta)}\right) \tag{8.2}$$

where a_d is the drained area per unit contour length, $\tan(\beta)$ is the local gradient, n is an exponent <1

6. **Geomorphic flood index (GFI)**, defined as the ratio between the term h_r and HAND. The numerator represents the water depth, computed in the hydrologically nearest stream chapter with a hydraulic scale relation ($h_r \approx bA_r^n$, where A_r is the contributing area in the considered stream chapter, see Figure 8.1). Coefficient b and exponent n can be taken from the literature (Nardi et al., 2006) or appropriately estimated with calibration (e.g., see the recent work from Annis et al., 2022, , where the dependence of b and n on catchment morphology and climate characteristics).

$$GFI = \ln\left(\frac{h_r}{HAND}\right) \tag{8.3}$$

7. Alternative version of the GFI, hereinafter referred to as **local geomorphic flood index (LGFI)**, defined as:

$$LGFI = \ln\left(\frac{h_l}{HAND}\right) \tag{8.4}$$

where the water depth h_l is computed with reference to the contributing area of the considered pixel

In addition, the elevation itself [m a.s.l.], is frequently used as a GD.

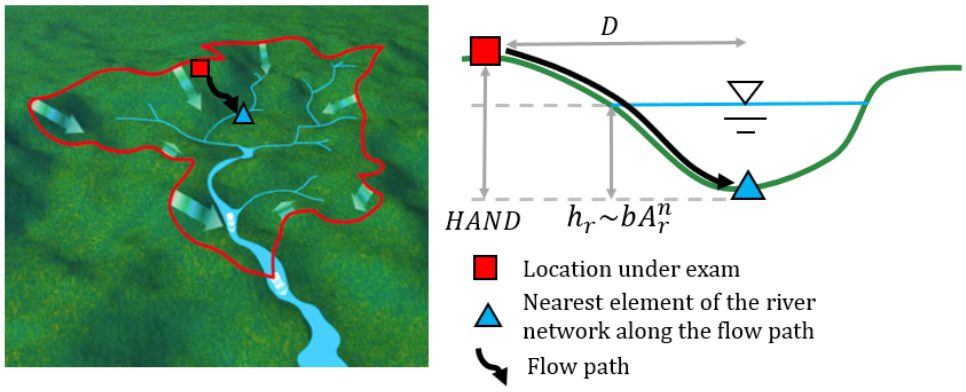

Figure 8.1: Schematic representation of some of the most effective and popular geomorphic descriptors: distance (D) and elevation difference ($HAND$) with respect to the nearest section on the river network along the flow path (or drainage direction); water depth computed with a scale relation (h_r)

8.1.2 Calibration

The calibration phase is extremely important, and relies on two aspects: the objective function and the reference information. With specific reference to univariate models, several metrics have been used and proposed for finding the optimal threshold, which consists in a constrained binary classification problem. Some of the most common are the accuracy (ACC), precision (or positive predictive value, PPV), recall (or true positive ratio, TPR) and true skill score (TSS; Youden, 1950; Everitt, 2002):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{8.5}$$

$$PPV = \frac{TP}{TP + FP} \tag{8.6}$$

$$TPR = \frac{TP}{TP + FN} \tag{8.7}$$

$$TSS = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \tag{8.8}$$

where TP, TN, FP, FN are respectively true positive, true negative, false positive and false negative predictions of the model. All of these objective functions vary between 0 (no skill) and 1 (optimal value).

With reference to the flood hazard reference information, two cases are possible. The first is when a coherent flood hazard map is available over the study area. In this case, all the segments of the river network are supposed to be associated with flood

hazard for a specific return period (or non-exceedance probability). Alternatively, only isolated events are available, that can be considered simoultaneously. In this second case, the return period and characteristics of the events can be highly heterogeneous, as the dataset could consist of observed events reconstructed from satellite images or from more hydraulic models, or of a mix of the two.

Aditionally, the study region could be used just partially for the calibration. In fact, previous studies (e.g., Tavares da Costa et al., 2019; Gnecco et al., 2017; Marchesini et al., 2021) have highlighted that the DEM-based classification of regions into flood-prone and flood-free zones is more effective if the calibration is done on meaningful areas. This is due to the different importance of far-from-river and close-to-river pixels in the computation of the objective function. Thus, some authors calibrated their models referring to the portion of the study area that is closest to the river network, and that we will term *calibration area* (Gnecco et al., 2017; Tavares da Costa et al., 2019; Marchesini et al., 2021).

Two methods are prevalent for the definition of the calibration area: a buffer with a fixed radius along the flood susceptible areas of the reference flood hazard maps, or a buffer with variable radius along the river network. In the second case, the variability can be determined by the height above the nearest stream section (HAND), the contributing area or the stream order.

8.2 Multivariate DEM-based models

A second class of DEM-based approaches to be investigated can be named as multivariate, as they rely on the combination of different GDs. The relation between the combination of GDs and flood hazard can be searched through numerous statistical methods. Commonly, machine learning (ML; Breiman et al., 1984) models are used, often ensembled with multi-criteria decision-making techniques (Triantaphyllou, 2000; Ho et al., 2010).

Regarding the covariates, some authors (Degiorgis et al., 2012; Gnecco et al., 2017) have tested a blend of GDs, while some others mixed these indices with information on land use, soil geology and climate, and compared different combination strategies (e.g., Wang et al., 2015; Lee et al., 2017; Khosravi et al., 2018; Arabameri et al., 2019; Janizadeh et al., 2019; Costache et al., 2020).

The objective function is variable, depending on the specific approach adopted. Usu-

ally, the MSE or ACC are used, in case the target is a countinuous characterization of flood susceptibility or a binary classification into flood-prone/non-flood-prone.

In most of the recent studies about multivariate DEM-based flood hazard modelling, the reference information used to set up the models consists of a dataset of isolated historical events observed in the study area (Lee et al., 2017; Khosravi et al., 2018; Janizadeh et al., 2019; Arabameri et al., 2019; Costache et al., 2020). Thus, DEM-based techinques are used to combine punctual informations from single flood events to derive continuous raster flood susceptibility maps. In other words, information from hydraulic simulations and hazard is completely absent from the reference dataset used for the calibration.

As evident, multivariate models present a significant advantage over the univariate ones: since the dynamics of flood events is not modelled, the usage of multiple information types is expected to improve the accuracy of such simplified models. However, considering multiple layers of raster information for large study areas, as typically done with DEM-based applications, can increase exponentially the computational costs.

Thus, additional studies about univariate models have been recently published (e.g., Nardi et al., 2019; Lindersson et al., 2021; Annis and Nardi, 2021; Annis et al., 2022), suggesting that data-driven flood hazard mapping has a remarkable potential not just for multivariate models, but also for the univariate ones, which still remain a valid and low-effort option.

Chapter 9

Machine-Learning blends of geomorphic descriptors: value and limitations for flood hazard assessment across large floodplains

9.1 Introduction

In this Chapter, mainly relying on Magnini et al. (2022) and Magnini et al. (2023), an innovative approach for multivariate DEM-based flood hazard mapping is proposed and discussed through two consecutive applications with the same methodology yet different focus. In both cases, we consider large study areas, characterized by markedly varied morphological, hydrological and climatic conditions.

First, Northern Italy is considered (i.e., $10^5 km^2$), and the $\sim 90m$ resolution, hydrologically-corrected, MERIT DEM (Yamazaki et al., 2017) is used for deriving a set of GDs. We then use decision trees (Hastie et al., 2009) for assessing flood hazard associated with a given probability of occurrence (i.e., return period) in terms of (a) delineation of flood-prone and flood-free areas, and (b) prediction of expected inundation water depth (as a measure for flood intensity).

The simultaneous combination of the five following meaningful elements makes our study different from all previous works in literature. First, only strictly easy-to-retrieve, DEM-based GDs are used to assess flood hazard, in contrast with several studies in which also other information is considered. Second, both generation of binary flood susceptibility maps and prediction of expected maximum inundation water depth are analyzed, setting up parallel models. Third, decision trees are trained using pre-existing flood

hazard maps as target information, in contrast with the discontinuous datasets of historical events mostly used to train machine learning models for flood hazard estimation (Lee et al., 2017; Khosravi et al., 2018; Janizadeh et al., 2019; Arabameri et al., 2019; Costache et al., 2020). Fourth, a univariate geomorphological approach for identification of flood-prone and flood-free areas (i.e., GFI) is compared with the proposed multivariate approach: this allows us to analyse the actual enhancement resulting from the use of multiple GDs. Fifth, predictive skill of the multivariate DEM-based flood hazard approach is assessed in extrapolation by applying models trained on specific geographical areas to different regions with dissimilar morphological and/or hydrological features. This last aspect is highly important for possible future applications to data-scarce environments in extrapolation mode.

In the second application, the whole of Italy is adopted as study area. The initial phase of this study consists of the analysis of the available datasets for selecting the most appropriate input DEM and reference flood-hazard map. For both, an effective framework for the selection is proposed, exploiting EU-Hydro (Gallaun et al., 2019) as reference river network. Second, the same methods described above are used to derive input GDs and set up a univariate and a multivariate DEM-based model. After the calibration over specific fractions of the study area, the two models are applied to the whole region. Finally, we validate both models by referring to independent information (i.e., datasets that were not used for training or calibration). Namely, this consists of three inundation extents produced by the same number of recent flood events and delineated on the basis of remote sensing data (local validation), as well as a synthetic inundation scenario obtained through 2D hydrodynamic numerical modelling. The latter was generated as the envelope of several levee-breaching simulations along a 300-km branch of the major Italian river.

Differently from what previously done, the aim is not the assessment of model's performance in test areas that differ from the training ones, but instead in areas where the calibration flood map may be inhomogeneous or inaccurate -which is a common situation in practice. This is done by referring to validation sources (e.g. detailed output of hydrodynamic model runs) that are alternative to the available target flood maps.

Usually, DEM-based methods are used at large scales to obtain binary maps, which delineate maximum flood extent associated with a given return period, and they are considered auxiliary tools of the more accurate hydraulic models. However, by nature DEM-based flood hazard models can efficiently handle secondary river networks (see e.g., Nardi et al., 2019) and can produce spatially continuous and highly homogeneous characterization of flood hazard. Some authors (e.g., Avand et al., 2022; Deroliya et al.,

2022; Costache et al., 2020) exploited multivariate DEM-based approaches for a spatially continuous estimation of flood susceptibility. Nevertheless, their models were calibrated on a number of independent inundation events, instead of a coherent flood hazard map with a given return period.

Through the application to the entire Italy, we investigate the unexplored potential of these features of DEM-based approaches for enhancing the flood hazard information with respect to heterogeneities and inconsistencies that are present in existing calibration maps covering all the study area. Moreover, the presentation of a baseline reference framework to address the selection of the most appropriate DEM and reference hazard map is an innovative and useful element for future studies on the topic.

By assuming the abovementioned characteristics, these two studies aim to advance previous knowledge on the potential of ML techniques for combining GDs to derive accurate flood susceptibility maps across large geographical regions. More precisely, we want to investigate four main research questions: (1) can we profit from a blend of various GDs for flood hazard assessment and mapping relative to a univariate approach? (2) Can we use simple ML techniques for effectively blending multiple GDs? (3) Are these techniques capable of providing a reliable assessment of flood hazard over large geographical areas when used in geographical extrapolation? What are the desired characteristics of the training region/watershed to make the trained model as general as possible? (4) Can we use DEM-based models to enhance existing flood hazard maps?

The present Chapter is organized as follows: Section 9.2 describes the general methods adopted for both the applications (i.e., GDs and decision trees); the study area, the detailed framework of the analyses and the results obtained for Northern Italy are illustrated in Sections 9.3, 9.4 and 9.5, in this order. Then, the details for the application to all Italy, including the methods for the selection of the input DEM and target flood hazard map, are described in Section 9.6. Section 9.7 shows the results for the selection of DEM and reference hazard map, the application of the models to the study area and their validation. Additional investigations about the application at a national scale relative to a recent catastrophic event are presented in Section 9.8. Finally, all the results are discussed in Section 9.9 and summarised in Section 9.10.

9.2 Methods

The analyses conducted in the study are based on two main elements: geomorphic descriptors (GDs) and decision trees (DTs); simplicity and replicability of these elements represent a fundamental aspect and an important advantage of this contribute.

Aiming to estimate flood hazard output variables (i.e., flood-susceptibility and maximum expected water depth), DTs models combine several selected DEM-derived input features (GDs), based on the availability of target information (i.e., flood hazard reference maps). Consistent with the aims of our study, we set up two different types of DTs: classifier DTs to solve the classification problem relative to flood-extent delineation, and regressor DTs to solve the regression problem of water depth estimation. Classifier and regressor models use the same input GDs, but require different target flood hazard maps. The software we use for the training is Scikit-learn (Pedregosa et al., 2011), open source library for Python 3.6 or later (Van Rossum and Drake Jr, 1995).

9.2.1 Geomorphic descriptors

As input variables for the above-mentioned models, in our study we use the ground elevation in meters a.s.l. itself (as retrieved from the DEM) together with six GDs, the first three of which are single indices, while the remaining three are composite (see Chapter 8):

1. **Local slope (sd8)**, pure number
2. **Horizontal distance from the nearest stream (D)**, in number of pixels
3. **Height above the nearest drainage (HAND)**, in meters
4. **Modified topographic index ($\mathbf{TI}_m$)**, pure number
5. **Geomorphic flood index (GFI)**, pure number. Coefficient b and exponent n can be appropriately estimated with calibration or taken from the literature (Nardi et al., 2006).
6. **local geomorphic flood index (LGFI)**, pure number

The choice of the above mentioned GDs is due to different reasons. First, previous studies (e.g., Manfreda et al., 2015; Samela et al., 2017) clearly showed that D and HAND are the most descriptive single-feature indices for flood hazard mapping, sufficiently accurate in mountainous regions, but still inadequate over predominantly flat areas, whereas, among composite feature indices, GFI and LGFI show good performance in both the geographical contexts. Also, in several studies (e.g., Wang et al., 2015; Lee et al., 2017; Khosravi et al., 2018; Janizadeh et al., 2019; Costache et al., 2020), elevation retrieved from DEM shows to have a strong influence on flood occurrence. Slope appears to be the most important index in Khosravi et al. (2018) and Costache et al. (2020), and

among the most influent ones in Arabameri et al. (2019). The adoption of TI_m is based on Manfreda et al. (2008), who highligthed a strong correlation between the index and the occurrence of inundation events.

Indeed, we believe that the selected set of GDs provides DT models with a rather exhaustive description of the study area morphology. In fact, slope and TI_m may influence the infiltration time, and consequently the runoff; elevation is not only strongly linked to the runoff, but also to climatic conditions; D and HAND consider the horizontal and vertical proximity to the river network, and GFI and LGFI combine this information with an estimation of the water depth in the nearest stream.
Overall, for the aim of a multivariate analysis, this combination should enable one to consider two comprehensive pieces of information by looking into the morphology (i.e. elevation, sd8, TI_m) and hydrology (i.e., by accounting for the river network; i.e. D, HAND, GFI, LGFI) of the study region.

9.2.2 Decision trees

Decision trees (DTs, see Chapter 2) are one of the machine learning models most frequently used in hydrology (e.g., Mosavi et al., 2018). There are multiple reasons why this technique was selected for the present research. First, their simplicity makes DTs low-computational-effort models, meaning that they can be suitable for applications over large study areas. Second, DTs are very convenient for modelling non-linear relations betwenn input and output, as it is in the case between GDs and flood-susceptibility. Third, they are easily interpretable, meaning that it is possible to exactly understand the path that leads a given input x_i to a certain output y_i. This last characteristic is not common to all ML models, and allows the user to also obtain a ranking of the relevance of the input features (i.e., the GDs) in the optimal configuration obtained during the training.

The combination of strictly DEM-based GDs through simple machine learning techniques aims to investigate an approach that can be implemented in low times, and just using common programming and GIS tools.

9.3 Testing the approach: application to Northern Italy

The study area includes most of Northern Italy and a little part of Switzerland, having a total extent of about $10^5 km^2$. Many different geographical subsystems can be found within this surface: the Alps, located in the North, lie in about $5 \cdot 10^4 km^2$, with average elevation of 2500 m a.s.l. and a mainly rocky soil. This mountain range also hosts several big lakes, as Garda, Maggiore and Iseo Lakes. The Apennines, in the southern portion, have lower altitudes than the Alps, and more permeable soils. The Po Valley, the largest floodplain in Italy, stretches from West to East, covering an area of about $4.6 \cdot 10^4 km^2$, going from the Alps and the Apennines to the Adriatic Sea (see Figure 9.1).
The study area is mostly occupied by the Po river basin, that is the largest in Italy. Moreover, other important rivers are the Adige, Brenta, Reno and Bacchiglione.
For this large and predominantly flat region, floods represent a major issue, also considering its high population density and presence of strategic industrial and agricultural assets (ISPRA, 2018; Persiano et al., 2020).
The DEM used to represent the study area is the freely-available Multi-Error-Remover Improved-Terrain model (MERIT; see Yamazaki et al., 2017). This choice was made for two reasons. First, MERIT should be quite reliable for hydrological applications, as it is the product of several processing operations and corrections on previously available DEMs (i.e., NASA SRTM3 and JAXA AW3D), some of which specifically addressing hydrological consistency (e.g. agreement between modelled and real stream-network). The second reason is that its resolution is 3 arcseconds, which corresponds to $\sim 90m$ at the equator. These characteristics enabled us to perform an accurate computation of geomorphic indices, while reducing the computational costs.

Two different freely-available reference flood hazard maps have been used to train the ML models. The first, used for the classification problem (i.e., delineation of flood-extent), has been produced by the Italian Institute for Environmental Protection and Research (ISPRA) in 2017 to fulfill the Floods Directive of the European Parliament (2007/60/EC). An updated version of the map was finalized in 2020 and released by ISPRA in 2021. However, the analyses described in the Dissertation rely on the version realized in 2017, and released in 2018, which was the only available at that time. This map (hereinafter referred to as PGRA P1) refers to a return period of about 500 years and comes from the merge of different hazard maps produced by local authorities, which explains its heterogeneity. Detailed flood hazard mapping characterizes some areas (e.g.,

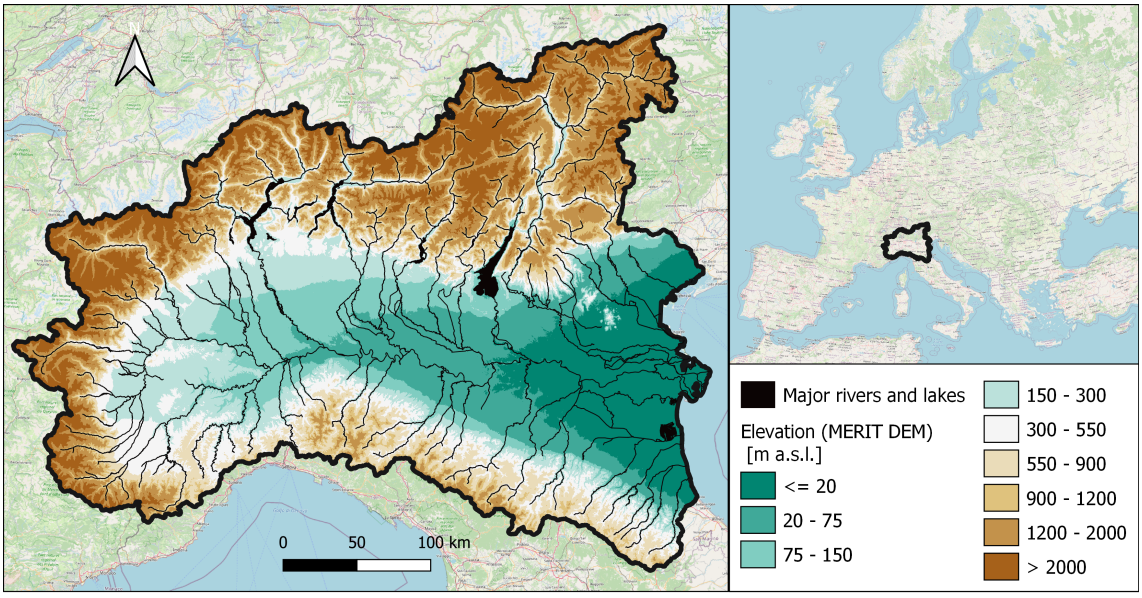

Figure 9.1: MERIT DEM for the study area, with major rivers and lakes marked in black (left); study area in the European context (right; map from ©OpenStreetMap contributors (2017), distributed under the Open Data Commons Open Database License (ODbL) v1.0). Adapted from Magnini et al. (2022)

see the northwestern portion of the study area in Figure 9.2), while lacking information affects other zones (e.g., see the northeastern portion of the study area in Figure 9.2). In the reminder of this study we term exhaustiveness the degree of detail by which flood hazard is defined and captured for minor streams.

The second map (see Figure 9.3), used for the regression problem (i.e., estimation of water depth), is made available by the study from the Joint Research Centre (JRC) of the European Commission according to the methods described by Dottori et al. (2016) and summarised in Chapter 7. It refers to a return period of 100 years, and thus, it will be named JRC 100 in the remainder of the study. Differently from PGRA P1, JRC 100 provides information in terms of water depth and is uniform throughout the study area, yet evenly incomplete and less accurate for minor streams, as it comes from the merger of several numerical simulations, which considered only river catchments with drainage area higher than $500 km^2$ (see Dottori et al., 2016, and Chapter 7).

9.4 Framework of the analysis

This chapter provides an overview of the four macro-phases of the present study, namely:

1. **Data selection and preparation**

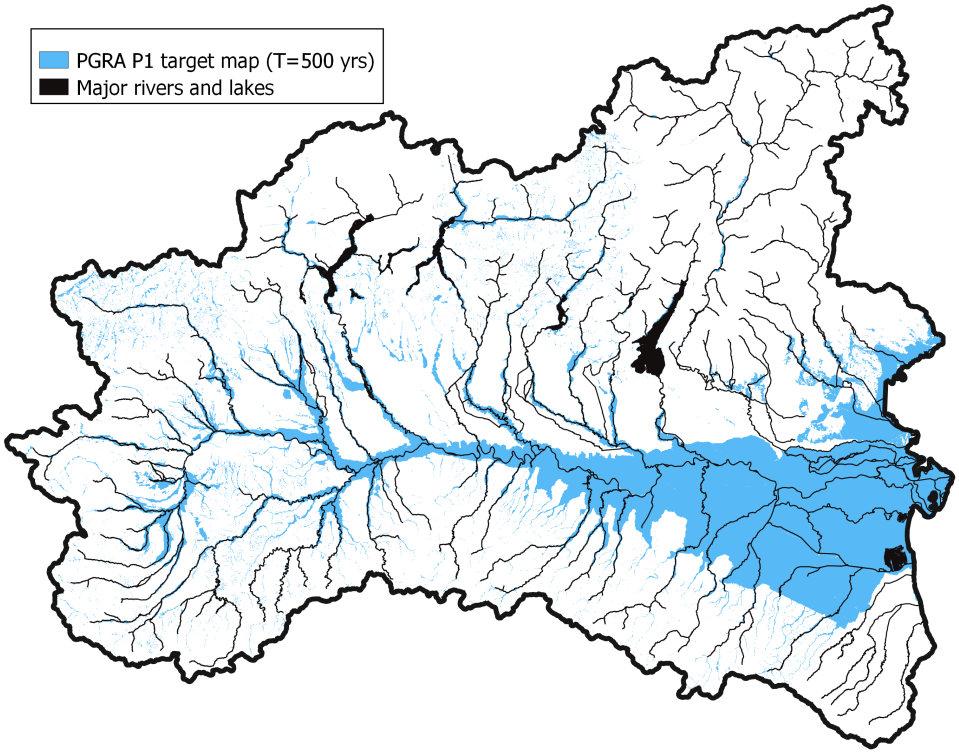

Figure 9.2: Binary flood hazard target map with return period ∼500 years, made available by ISPRA in 2018 (ISPRA, 2018) and termed PGRA P1 in this study. Adapted from Magnini et al. (2022)

   (a) Selection of the DEM and computation of geomorphic indices with terrain analysis

   (b) Selection of the flood hazard target map

2. **Preliminary analyses**

   (a) Definition and preparation of the calibration area

   (b) Selection of performance metrics and objective functions

3. **Implementation of the univariate approach (benchmark approach)**: set-up of GFI optimal threshold in randomly-selected 85% of calibration area

4. **Testing multivariate approach with two different modes:**

Testing mode	***Training set***	***Testing set***
Geographical interpolation	Randomly selected 85% of calibration area	Randomly selected 15% of calibration area
Geographical extrapolation	Geographical sub-region of the calibration area	Geographical remainder of the calibration area

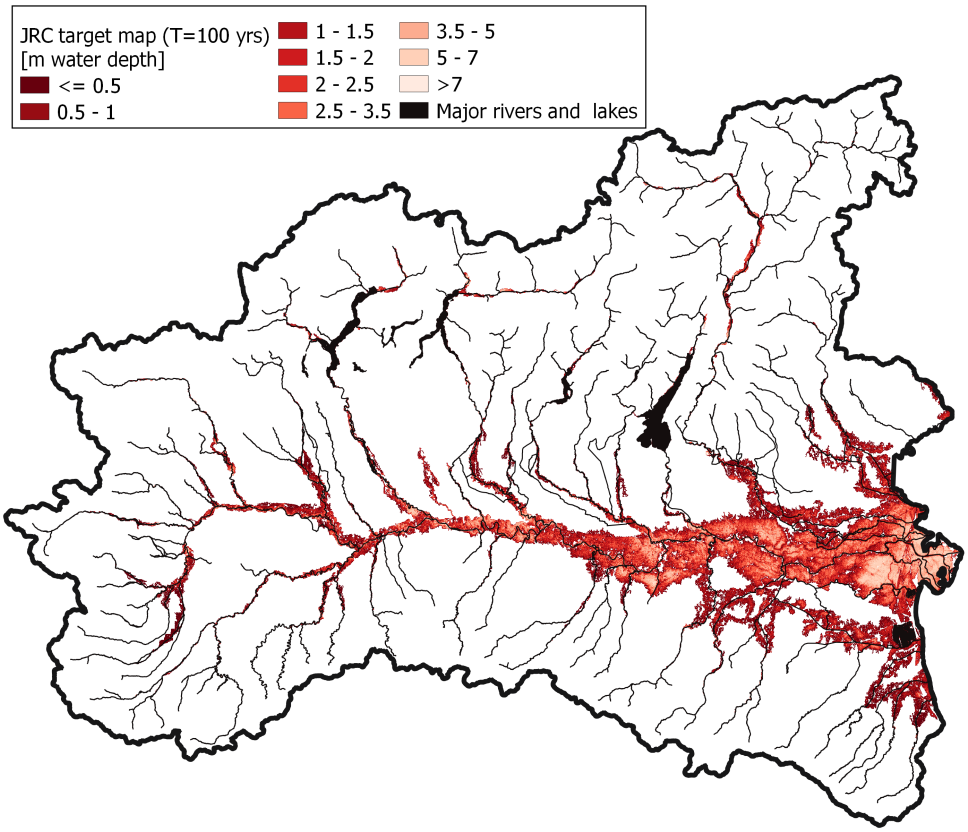

Figure 9.3: Water depth for the target 100-year flood hazard map obtained by Dottori et al. (2016), termed JRC 100 in this study (colour classes in the legend are used for data visualization only). Adapted from (Magnini et al., 2022)

Macro-phase (1) of the study consists of the preparation of input data, which is a fundamental step for the success of machine learning algorithms; specific criteria are used to select the GDs (Sect. 9.2.1), the accuracy and horizontal resolution of the DEM, and the target flood-hazard datasets (Sect. 9.3). Phase (2) (i.e., preliminary analyses) is necessary for defining some important aspects for the succesful set-up of DEM-based models: the calibration area (Sect. 9.4.1), the objective functions and the performance metrics for evaluating the results (Sect. 9.4.2). Phase (3) identifies the benchmarking approach, i.e. a univariate DEM-based model for classification of flood susceptible areas, to be used as comparison for the successive analysis. This model is built up according to the indications reported in the literature, and considers the GFI descriptor alone, as it is found to be the most versatile and accurate by many authors (e.g., Samela et al., 2017).

The main results of the study are obtained in phase (4), as the DEM-based multivariate approach is tested in two different ways. First, two DTs are set-up (i.e., one classifier DT and one regressor DT) using training and test sets with the same statistical distribution of input features. This represents an ideal case (termed here as *geographical*

interpolation mode), in which the training and test sets have very similar morphoclimatic characteristics.

Second, four sub-portions of the study area are selected based on specific morphoclimatic conditions, and then, eight more DTs are trained on these areas (i.e., one classifier DT and one regressor DT for each training area) and tested on the complement to the study region (see Sect. 9.4.3). This represents a data-scarce case (termed here as *geographical extrapolation mode*), in which morphoclimatic characteristics of training and test sets may be rather different.

9.4.1 Calibration area

Since previous studies (e.g., Tavares da Costa et al., 2019) have highlighted the benefits from defining a calibration area, in the present study, training and testing of the models have been performed referring to a portion of the entire study area.

Different methods to define this area have been tested during the preliminay analyses of phase (2), finding that the most effective way, representing a good trade-off between the calibration efficiency and the ease of identification, is to refer to a constant-radius buffer around the target flood hazard map. In particular, based on sensitivity analyses that clearly showed that the radius value has a non-negligible impact on the accuracy of the trained model, a 2 km radius has been selected for the PGRA P1 target map, and a 5 km radius for the JRC 100 map (see Figure 9.4). Thus, during our analyses, all the pixels falling outside the 2 km and 5 km calibration buffer areas are neglected when fitting the models and evaluating the results for all classification and regression problems, respectively.

9.4.2 Objective functions and performance metrics

Specific objective functions are used to train the DTs for classification and regression, while other performance metrics are computed to evaluate their predictions during the validation. With regards to the classification problem, the objective function, used during the training of the DTs to assess the quality of each split, is the Gini impurity ($I_G(p)$), that varies between 0 (the optimal value) and 1 (Hastie et al., 2009). At each step, the Gini impurity measures how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution in the subset. Given the number of target classes J, and the fraction of items labeled with class i in

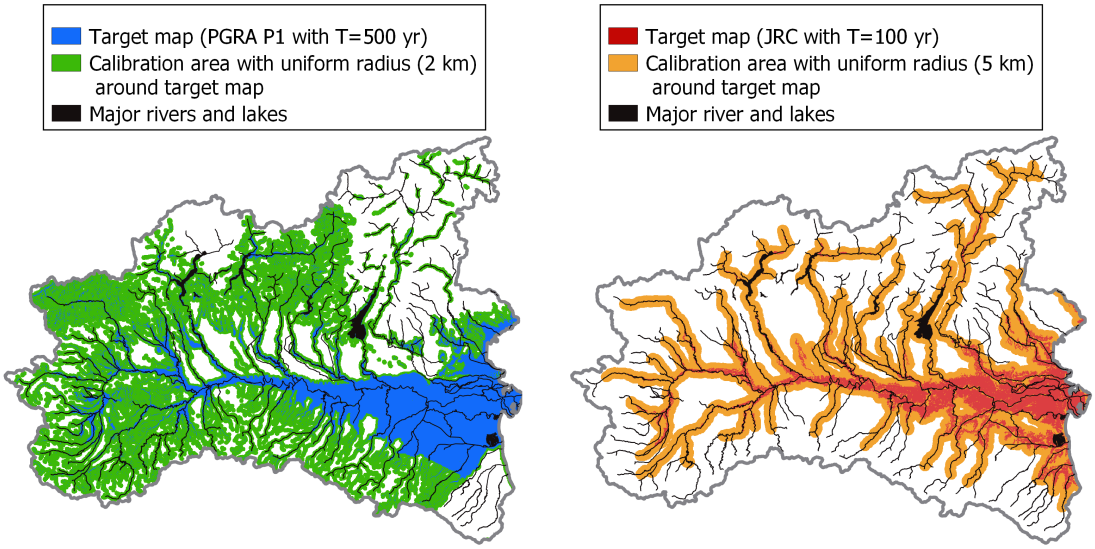

Figure 9.4: Calibration areas: 2km buffer (green) and PGRA P1 flood-prone areas (blue) used for the classification problem (left); 5km buffer (orange) and JRC 100 flood-prone areas (red) used for the regression problem (right). Adapted from Magnini et al. (2022)

the set p_i, the Gini impurity is defined as follows:

$$I_G(p) = \sum_{i=1}^{J} p_i \cdot (1 - p_i) \tag{9.1}$$

To perform implementation of the univariate approach and parameterize the multivariate classifier DTs, and evaluate the results, we use the true skill statistic (or TSS, see 8). TSS has been successfully used by several authors in different applications (Bartholmes et al., 2009; Alfieri et al., 2012; Tavares da Costa et al., 2019). During preliminary analyses of phase (2), some experiments suggested to prefer this metric to accuracy (ACC, see below), which showed to be less sensitive to model modifications (i.e., different calibration areas, input information, tree depth) and goodness (lower extension of FP and FN areas).
Other metrics used for analysing the results are accuracy (ACC), precision (or positive predictive value, PPV), and recall (or true positive ratio, TPR). All the three are very common in evaluating the performance of a classifier (e.g., Manfreda et al., 2015; Samela et al., 2017). They all vary between 0 and 1 (see 8).

With regards to the regression problem, the objective function to minimize during the training is the well-established mean squared error (MSE, see eq. 2.3 in Chapter 2). The metric mainly used to evaluate the results and parameterize the multivariate regressor DTs is the determination coefficient R^2, that varies between $-\infty$ and 1 (the

optimal value). It measures the improvement of the predicted values relative to the mean of the input samples ($\overline{y}$), defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{9.2}$$

The last considered metric is the mean absolute error (MAE), defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9.3}$$

Lastly, we use the Gini importance (GI) to measure the importance of each factor (i.e., each GD) in the trained models (both classifier and regressor DTs), which is defined for the *j-th* factor as the total decrease in node impurity (I_{G_i}), weighted by the fraction of samples reaching that node (n_i). Although this measure is largely used for its speed of computation, it has the drawback of neglecting the weakest factor when two related factors are used, which has to be taken into account when discussing the results.

$$GI_j = \sum_{i=1}^{N_j}\frac{(I_{G_i} - I_{G_{i-1}})}{n_i} \tag{9.4}$$

where N_j is the number of nodes where a condition on the *j-th* factor is used as splitting rule.

9.4.3 Training and testing strategy

All models considered in this study are trained and tested in different sub-domains of their calibration area based on two different strategies.

For the univariate model and the interpolation DTs, the pixels of the calibration area have been randomly split in 85% for the training and 15% for the test set, based on established proportion adopted for machine learning algorithms (Mosavi et al., 2018). This produces two datasets with millions of pixels, both with very diverse ranges of input and target information.

During the extrapolation analyses, training is performed in turn on four different portions of the overall calibration area. To avoid dividing any catchement into a part for training and one for testing, the delineation of these areas follows catchment boundaries, as well as precise geographical and hydrological criteria (see Figure 9.5):

- **Area A** includes the Alpine catchments and the northern portion of the Po river floodplain. The complementary test area includes all the Apennines, a lower moun-

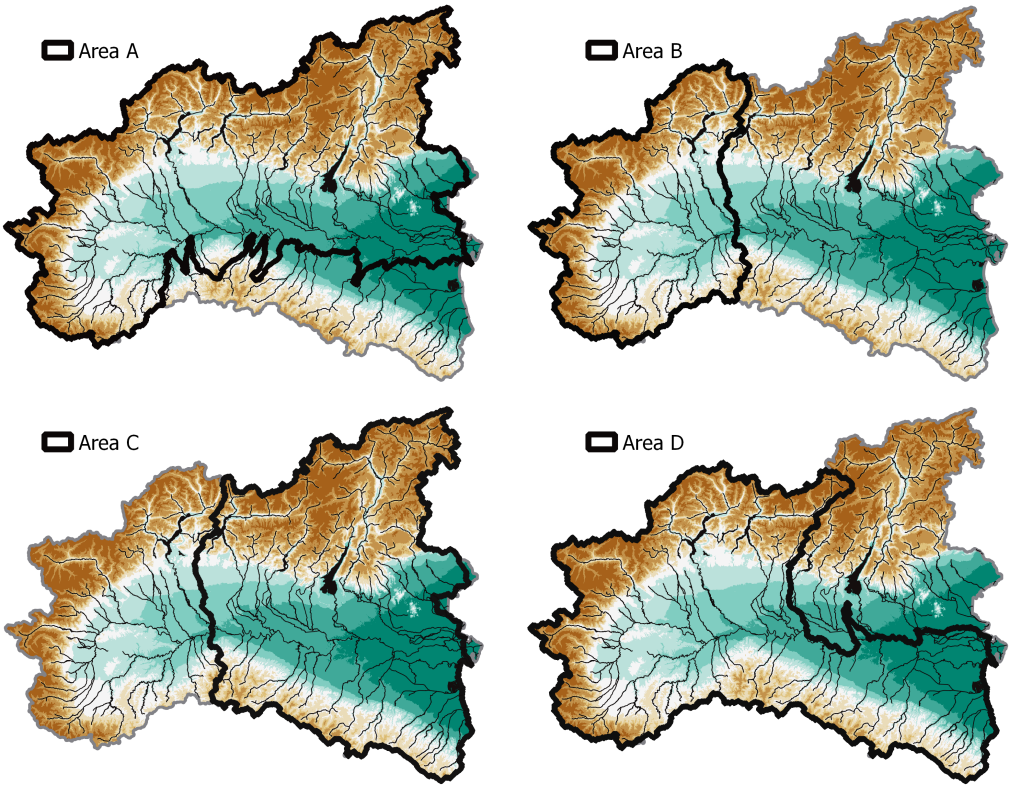

Figure 9.5: Training areas (bold contour) used for the geographical extrapolation experiments performed in phase (4), with major rivers and lakes higlighted in black. Adapted from (Magnini et al., 2022)

tain range, and the southern part of Po plain, where smaller river catchments are located

- **Area B** includes catchments in the upstream sector of the Po river basin, representing part of the Alps and of the Apennines. The complementary test area includes most of the Po plain, and part of the Alps and Apennines

- **Area C** is the complementary of area B and consists of the downstream portion of the Po river basin

- **Area D** includes the Apennines, Western and Central Alps and the entire Po streamline. Its complementary test area contains a rather small part of the Po plain, the Western Alps and the flood plain of the Adige, Brenta and Bacchiglione rivers

Before training DTs, k-fold cross-validation (CV) is performed to optimize models hyper-parameters, namely: the maximum tree depth and the minimum number of records in any leaf node. K-fold CV is a widely used method for model parameterization and

selection (Hastie et al., 2009), and consists in dividing the training set into k folds and then performing two consecutive operations: (1) training of the model using $k-1$ folds, and (2) validation of the model using the remaining fold. These two steps are repeated for k times, for all the combinations of the k folds of the training data.

9.5 Results of the application to Northern Italy

The reliability of the predictions of the models is assessed by performance metrics that refer to (a) the training set and (b) the test set. While the metrics computed for the training set assess the reliability in reproducing the observed target map, the ones regarding the test set measure the ability of the model when applied to a different sample than the one used in training (i.e. validation of the model). In order to find out the relevance of each input GD in the DTs' structure, the Gini importance (see Sect. 9.4.2) for each model is reported in Table 9.3, and will be better discussed in Sect. 9.9.

9.5.1 Delineation of flood-prone areas in interpolation mode

Figure 9.6 represents the flood susceptibility map obtained with the classifier DT model trained within the random 85% of the 2 km buffer calibration area (i.e. multivariate flood susceptibility map). To understand the quality of the proposed approach and profitably discuss the results, Figure 9.7 illustrates the map produced with the univariate benchmark approach set up in the same area. Relevant performance metrics for multivariate and univariate models are reported in rows 1 and 2 of Table 9.1, respectively.
Figure 9.6 and Table 1 highlight that the DT flood susceptibility map is strongly consistent with the target map PGRA P1. Also, the model produces a rather detailed mapping across floodplains of minor streams (i.e. exhaustiveness, as defined in Sect. 9.3). In particular, it can be observed in Figure 9.6 that the zones where the target map has high exhaustiveness (e.g., northwestern portion of the study area) are mapped with slightly lower exhaustiveness by the DT model, while the DT output is more detailed in floodplain of minor streams than the target map, where the latter is lacking exhaustiveness (e.g., northeastern part).

Figure 9.6 shows that GFI uniformly and exhaustively estimates flood susceptibility along all minor streams in mountain areas, but tends to severely overestimate the size of flood-prone areas in predominantly flat regions.
The first line of Table 9.3 reports the Gini importance for the classifier DT: HAND scores about 65%, followed by elevation (16.5%) and GFI (10.5%).

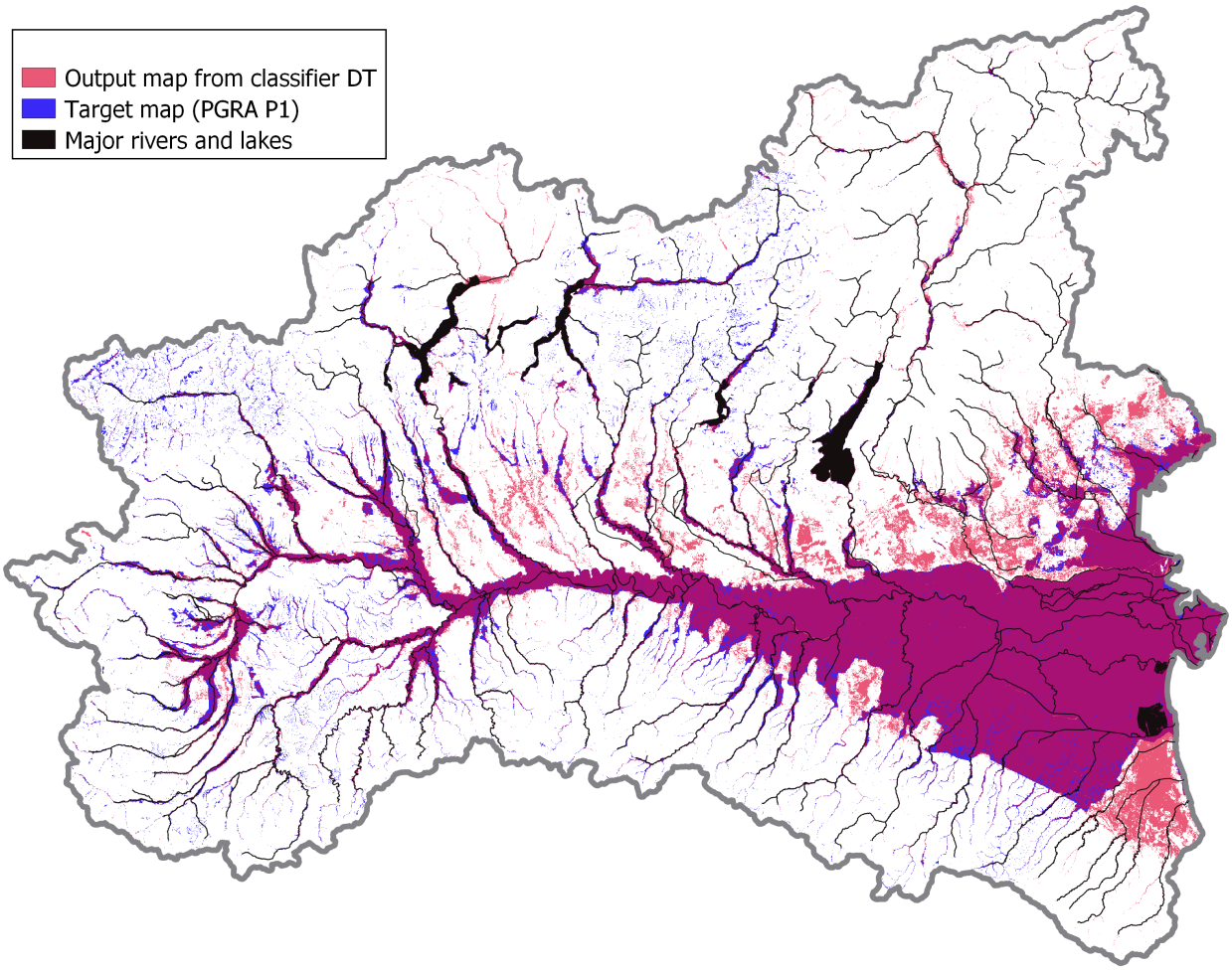

Figure 9.6: Multivariate 500-year flood susceptibility map for the study area (red); target flood hazard map (PGRA P1, blue); purple indicates overlaying areas. Adapted from Magnini et al. (2022)

9.5.2 Prediction of flood hazard intensity in interpolation mode

Figure 9.8 illustrates expected maximum inundation water depths as predicted through the regressor DT trained within the random 85% of the 5 km buffer calibration area; relevant performance metrics can be found in the first row of Table 9.2. Figure 9.8 and Table 9.2 show good performance of the DT model for the regression problem. It is worth noting here that the exhaustiveness of the DT water-depth map is considerably higher than that of the reference map (i.e. JRC 100). This result was expected due to the focus of JRC 100 on larger catchments.

The data density plot in Figure 9.9 depicts the relationship between target and predicted water depths for the test set focusing on true positives (i.e. both target and predicted water depths are higher than 0.0 m) and neglecting water depths higher than $3.5m$ (neglected pairs, beyond axes' limits, are 4.2% of the total).

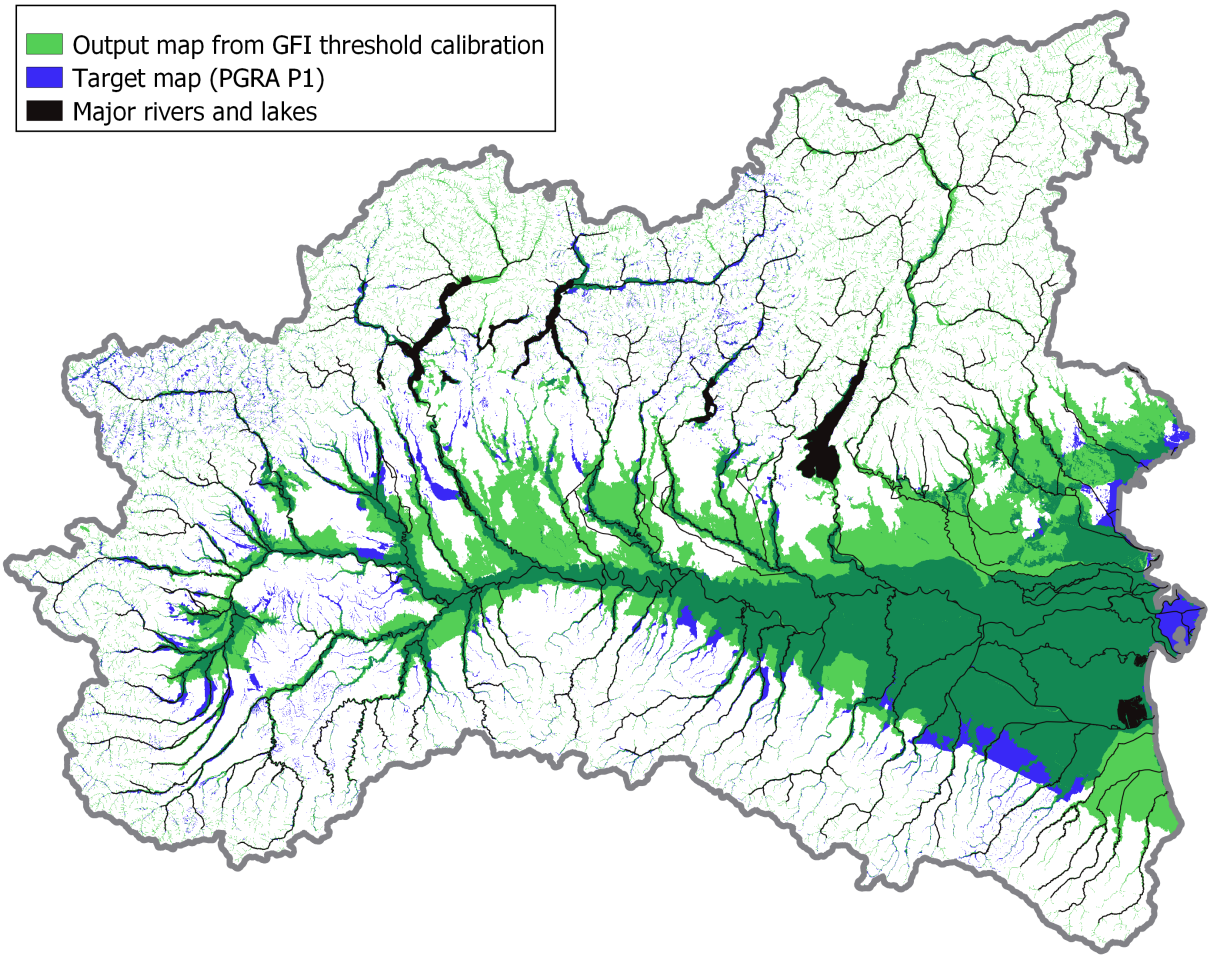

Figure 9.7: Binary flood susceptibility map resulting from a univariate analysis (morphometric index: GFI, light green); target flood hazard map (PGRA P1, blue); dark green indicates overlaying areas. Adapted from Magnini et al. (2022)

The second row of Table 9.3 shows that the most informative GD is GFI (63.7%), followed by elevation (20.7%) and slope (5.4%).

9.5.3 Multivariate flood hazard modelling in extrapolation mode

Tables 9.1 (rows 3-6) and 9.2 (rows 2-5) report performance metrics for the geographical extrapolation experiments for the classification and regression problems, respectively, while Figures 9.10 and 9.11 depict the corresponding DT output maps.
With regards to the classification problem (Table 9.1), the performance metrics highlight a generalized good agreement with the target map. Figure 9.10 and chapter "Training performance" of Table 9.1 show that all models can accurately reproduce the target map in the training area, but they are quite inaccurate in the test area, as it is evident the difference between the two. In fact, concerning the test area, Table 9.1 shows that

Table 9.1: Classification problem: performance metrics for the multivariate (classifier DTs) and univariate (classifier GFI) flood susceptibility maps; target flood hazard map for both approaches: PGRA P1. The reported values have been converted from the interval 0-1 to the percentage notation. The best testing metrics values are reported in bold, the worst ones in italic (the first line should be compared with the second one; the last four lines should be compared to each other)

Model	**Training performance**				**Test performance**			
	TSS	ACC	PPV	TPR	TSS	ACC	PPV	TPR
Classifier DT - interpolation	80%	93%	89%	84%	**78%**	**92%**	**88%**	*83%*
Classifier GFI - benchmark	69%	84%	66%	87%	*69%*	*84%*	*66%*	**87%**
Classifier DT trained in A	75%	92%	86%	78%	56%	83%	**88%**	61%
Classifier DT trained in B	61%	93%	82%	64%	**65%**	85%	80%	75%
Classifier DT trained in C	82%	92%	89%	88%	*33%*	**88%**	71%	*35%*
Classifier DT trained in D	80%	94%	91%	93%	63%	*79%*	*53%*	**87%**

Table 9.2: Regression problem: performance metrics for the multivariate water-depth output maps obtained with the regressor DTs (target flood hazard map: JRC 100); the best testing metrics values are reported in bold, the worst ones in italics

Model	**Training performance**			**Test performance**		
	R^2	MSE	MAE	R^2	MSE	MAE
Regressor DT - interpolation	0.726	0.227	0.393	**0.692**	**0.242**	**0.439**
Regressor DT trained in A	0.709	0.240	0.443	-0.029	1.100	0.547
Regressor DT trained in B	0.606	0.145	0.284	*-2.110*	*5.208*	*1.283*
Regressor DT trained in C	0.711	0.281	0.467	**0.333**	**0.623**	**0.264**
Regressor DT trained in D	0.741	0.251	0.380	0.175	1.109	0.417

according to the true skill score (TSS), the best prediction in the test area is obtained using B as training area (TSS=65%), followed by D (TSS=63%) and A (TSS=56%), respectively.

The same table shows that the best results are obtained when training on area C if one focuses on accuracy (ACC=88%), followed by B (ACC=85%) and A (ACC=83%). According to precision (PPV), the best result is obtained training the model on area A (PPV=88%), while it is D according to recall (TPR=87%).

Concerning the regression problem, worse predictive skill in geographical extrapolation is observed in Table 9.2. Differently from the classification, performance metrics for the regression problem are in good agreement among each other, showing that area C has the better results, while area B is the worst. On the other hand, Figure 9.11 suggests that water depth estimation in the test area is quite reliable in all the cases, with the exception of the DT trained in area B.

Focusing on Gini importance, Table 9.3 clearly shows that regressor DTs (rows 7-10) are characterized by similar structures regardless of the training areas: GFI is always ranked

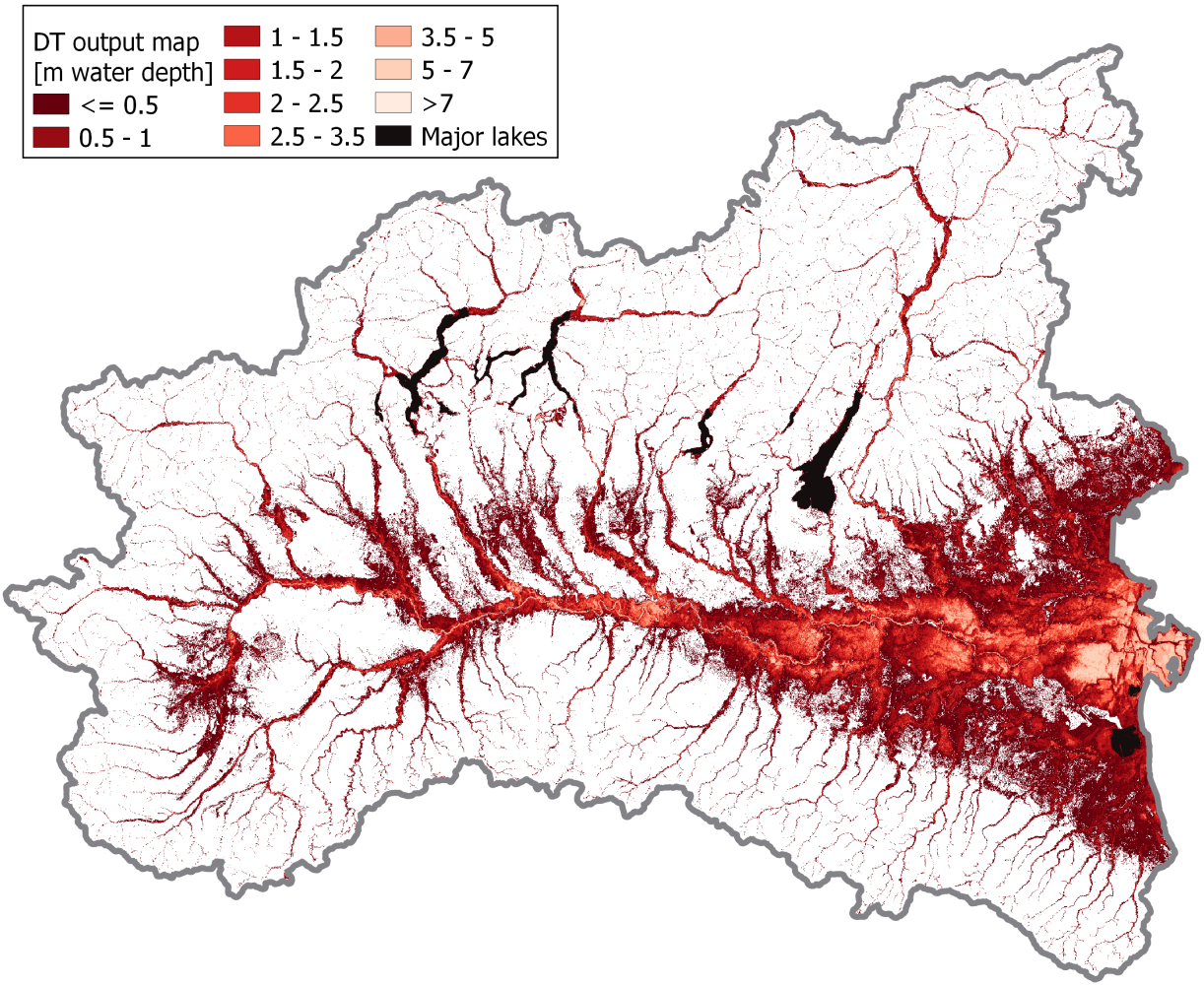

Figure 9.8: Multivariate water-depth hazard map obtained with regressor DT in interpolation mode (target flood hazard map: JRC 100). Adapted from Magnini et al. (2022)

first in terms of relevance, followed by elevation and slope. This is not true for the classification problem (rows 3-6): in this case, classifiers DTs identified four different training areas have different structures, in which the most informative geomorphic descriptor can be alternatively GFI, or HAND, or the elevation; this latter is always ranked second.

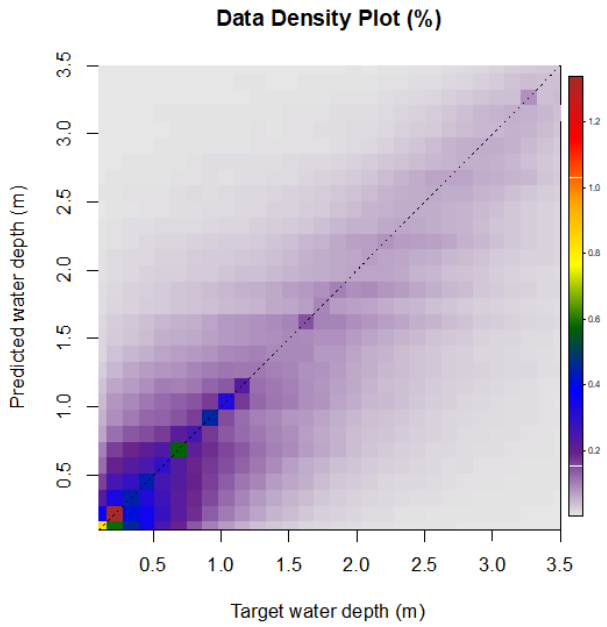

Figure 9.9: Data density plot (%) for target vs. predicted expected maximum water depth (target values: empirical JRC 100; predicted values: regressor DT applied to the test set). Adapted from Magnini et al. (2022)

Table 9.3: Gini importance of the selected input features computed for the DTs trained in phase (4); the highest value for each DT is highlighted in bold, the lowest in italic

Model	elevation	sd8	D	HAND	GFI	LGFI	TI_m
Classifier DT - interpolation	16.5%	3.5%	2.8%	**65.6%**	10.5%	0.6%	*0.4%*
Regressor DT - interpolation	20.7%	5.4%	2.0%	4.8%	**63.7%**	1.8%	*1.6%*
Classifier DT trained in A	10.2%	6.8%	2.2%	8.0%	**71.6%**	*0.3%*	0.8%
Classifier DT trained in B	9.8%	9.8%	3.8%	**60.0%**	11.8%	4.2%	*0.4%*
Classifier DT trained in C	**74.3%**	2.3%	1.7%	9.7%	11.1%	0.6%	*0.1%*
Classifier DT trained in D	18.5%	2.8%	1.4%	**69.5%**	7.1%	0.4%	*0.3%*
Regressor DT trained in A	14.3%	3.6%	1.8%	3.5%	**73.2%**	2.3%	*1.3%*
Regressor DT trained in B	18.9%	3.8%	2.6%	4.2%	**66.7%**	2.0%	*1.9%*
Regressor DT trained in C	17.8%	3.1%	1.9%	4.3%	**69.2%**	2.5%	*1.2%*
Regressor DT trained in D	14.3%	3.9%	1.3%	4.0%	**74.7%**	*0.9%*	*0.9%*

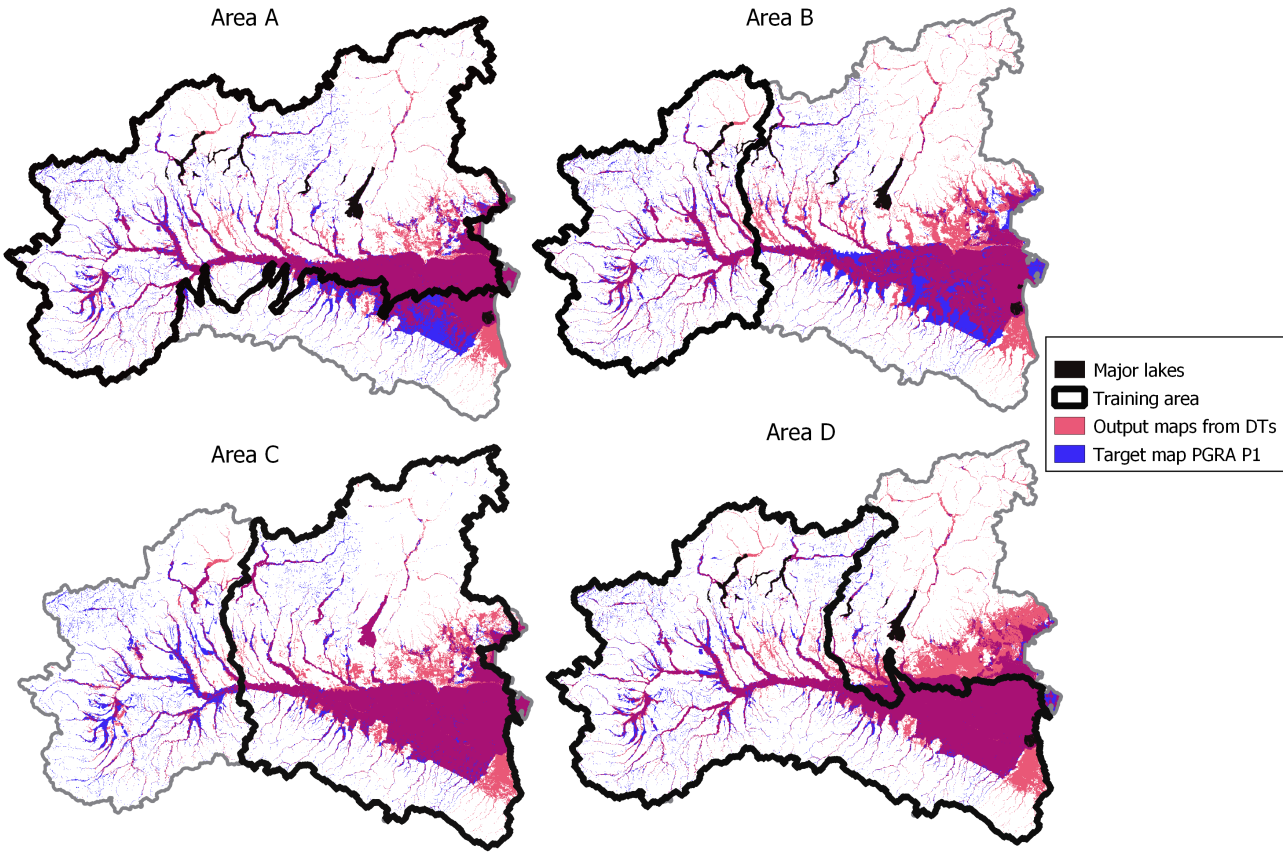

Figure 9.10: Geographical extrapolation for the classification problem: multivariate flood susceptibility maps obtained from classifier DTs (red); target flood hazard map (PGRA P1, blue); purple indicates overlaying areas. Adapted from Magnini et al. (2022)

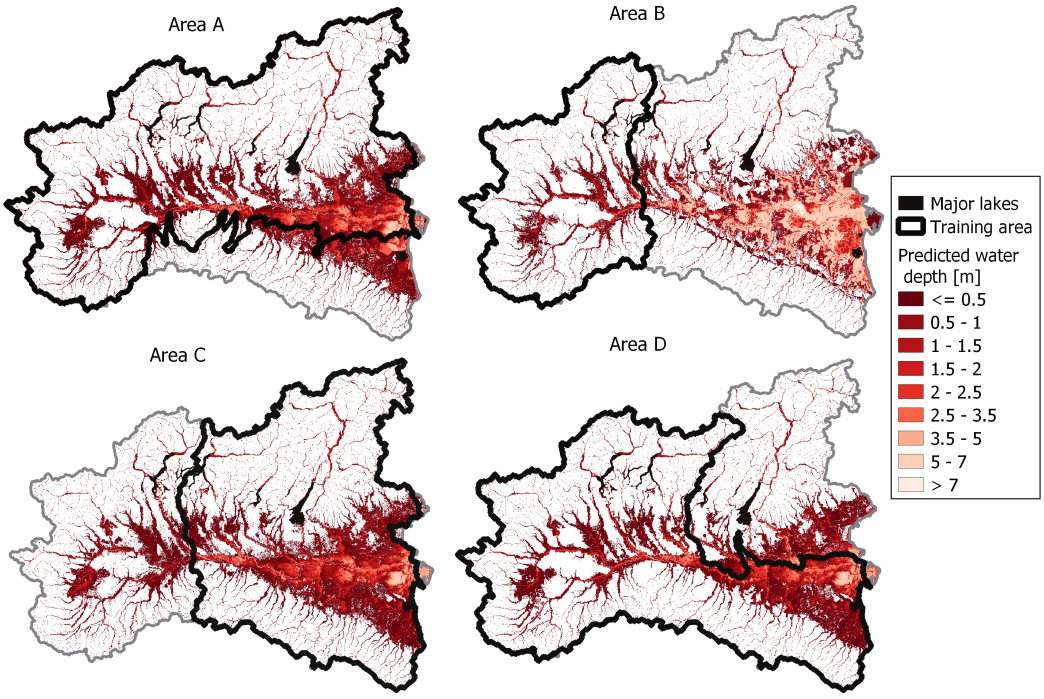

Figure 9.11: Geographical extrapolation for the regression problem: multivariate flood susceptibility maps obtained from regressor DTs (see also Figure 9.3, target flood hazard map: JRC 100). Adapted from Magnini et al. (2022)

9.6 Deploying the technology: application to Italy

The methods adopted for the set-up of the univariate and the multivariate DEM-based models for this second study case, focusing on the whole of Italy, are mainly the same as for the previous application to Northern Italy. The main differencies are the following points:

1. A more objective and extended approach for the input DEM and target flood hazard map is presented (i.e., phases (1.a) and (1.b) in the framework presented at Section 9.4)

2. Preliminary analyses for the selection of the appropriate buffering distance for the univariate model led to the value of 200 meters (i.e., phase (2.a) in the framework at Section 9.4)

3. The implementation of the univariate approach is more sophisticated, and consists of two phases: first, the geographical domain is divided into hydro-climatic districts; second, for each in turn, the best threshold for the GFI is searched by means of optimization of the TSS (compare with phase (3) in the framework at Section 9.4)

4. The multivariate model is set-up by considering only six geomorphic descriptors out of the seven adopted for Northern Italy (Section 9.2.1). The discarded descriptor is the modified topographic index (Manfreda et al., 2008), as it is the one with the lowest influence in the models (see Table 9.3)

5. The univariate and multivariate models are validated only in interpolation, but against a more extended set of independent flood hazard datasets (compare with phase (4) in the framework at Section 9.4)

These changes to the framework of the analyses are due to the fact that the study area is notably more extended than for Northern Italy. Modification (4) aims to reduce computational demand for data processing and models training, while (1), (2) and (3) Specific subsections follow, that are dedicated to the study area, the validation datasets, and the methods for the selection of the input DEM and target flood hazard map.

9.6.1 Second study area: Italy

The selected study area consists of the whole Italian peninsula and islands, that approximately amount at 301'103km^2 (see Figure 9.12). The overall extent and the large

variety of geographical and climatic conditions make Italy an interesting and complex study area for large scale flood-hazard modelling. In fact, the length of the peninsula and mostly mountainous hinterland make the climate highly diverse, ranging from humid subtropical to humid continental and oceanic (e.g., see climate classification in Cui et al., 2021). The Alps extend from the north-west to the north-east, and are the highest mountain range (i.e., twelve peaks higher than 3500m a.s.l., and highest peak at 4809m a.s.l.). The Apennines stretch from North to South for almost the whole peninsula (i.e., about 1300km) and have lower peaks compared to the Alps (i.e., maximum peak is 2900m a.s.l.). The largest plain, located in the north, is the floodplain of the Po river, that is also the main river in Italy (i.e., length of ∼650km, and mean discharge of ∼1500 m^3/s).

Italy is an interesting case study for flood modelling, as a large portion of its territory is subject to floods: 5.4% with high probability hazard, which corresponds to 2.4 million of people exposed, while 14% to low probability hazard, corresponding to 12.2 million of people (Trigila et al., 2021).

9.6.2 Validation datasets

After being trained (i.e., calibrated) in the calibration areas (i.e., 200m and 2km buffer areas, as described in Section 2), the DEM-based models are applied to the whole study region (i.e., even outside buffer areas). The validation is performed by comparing the model outputs with new information, consisting of two types of datasets: (1) three inundation maps delineated from satellite data, and (2) one envelope flood hazard map obtained from the merger of several 2D hydrodynamic simulations. These all are associated with a return period that is approximately the same as, or lower than the reference flood hazard map. Based on the return period and the location, which corresponds to the most flood susceptible area in Italy, the four datasets can be adopted for an effective validation.

Concerning the first validation dataset, two recent flood events are selected: the inundation event that occurred between 19th and 24th October 2019 in Alessandria Province (AL, Piedmont region), and the one that occurred between 15th and 19th November 2019 in Bologna province (BO, Emilia-Romagna region). These events are described by three Sentinel-1 SAR (Synthetic Aperture Radar) images; the corresponding inundation extents are delineated, within the present study, through a change detection method expressly developed, partially derived from Canty (2019), and successfully validated with ground truth data. Finally, three inundation maps, named as AL 21/10/19, BO 20/11/19 and BO 21/11/19, are obtained and used for the validation of the target and modelled flood maps (see panels 1(a), 2(b), 3(c) of Figure 9.12).

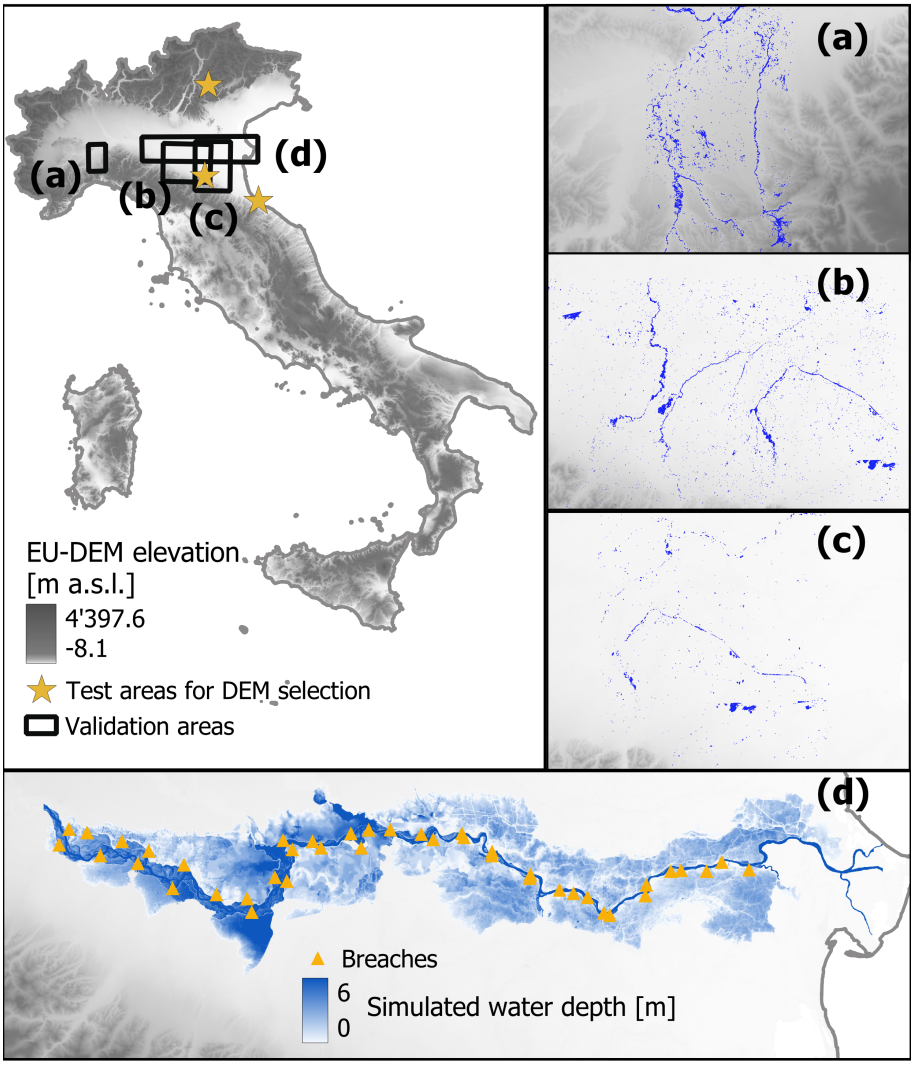

Figure 9.12: EU-DEM in Italy (top-left panel); validation datasets: inundation maps associated with AL21/10/19 (panel 1(a)), BO20/11/19 (panel 2(b)), BO21/11/19 (panel 3(c)) events; catastrophic inundation scenario along the middle lower portion of the Po River in terms of maximum simulated water depths (panel 4(d)). Adapted from Magnini et al. (2023)

The second validation dataset consists of a single flood hazard map obtained from a series of simulations run with the two-dimensional hydrodynamic model LISFLOOD-FP (Bates and De Roo, 2000; Neal et al., 2012; Shustikova et al., 2020). The flood hazard map of 50m horizontal resolution was produced by merging 42 model simulations into one map (i.e., each simulation is associated with a return period higher than 200 years, but lower than 500 years, see e.g. Domeneghetti et al., 2015). Each simulation represents hypothetical levee breaches on both banks of the river (21 on the right and 21 on the left bank; see panel 4 of Figure 9.12).

9.6.3 Analysis of available DEMs

Nowadays, DEMs can be developed from a variety of different surveying techniques, characterized by specific advantages and disadvantages. Above all, the onset of modern satellite remote sensing techniques allowed the creation of freely distributed DEMs with global or semi-global coverage. However, the accuracy of the DEMs is often affected by inevitable errors, that are associated with the techniques and algorithms used for creating the DEMs, and the characteristics of the terrain, such as morphology and land cover (Mukherjee et al., 2013; Thomas et al., 2014). Producers of global and national DEMs carry out quality assessments on their products, yet usually, only the global RMSE (i.e., root mean square error, see formula below eq. 9) is provided as a measure of accuracy, which gives no information about the accuracy over specific areas of interest or geomorphological contexts. Thus, choosing the right terrain model for a study can be difficult without performing a specific data quality analysis (Florinsky et al., 2018; Patel et al., 2016; Tavares da Costa et al., 2019; Thomas et al., 2014).

For this reason, seven different DEMs with a spatial resolution finer than 100m, obtained with various techniques and covering the whole Italy, are considered and assessed. Namely, they are: SRTMGL1 (Farr et al., 2007), ASTER GDEM (Tachikawa et al., 2011; Abrams, 2016; Gesch et al., 2016), ALOS AW3D30 (Tadono et al., 2016; Takaku and Tadono, 2017), TINITALY (Favalli and Pareschi, 2004; Tarquini et al., 2012), EU-DEM (Bashfield and Keim, 2011; Garcia G., 2015), HydroSHEDS DEM (Lehner and Grill, 2013) and MERIT DEM (Yamazaki et al., 2017). The spatial resolution of these DEMs is: 1 arc second (∼30 m) for STRMGL1, ASTER GDEM and ALOS AW3D30; 10 m for TINITALY; 25 m for EU-DEM and 3 arc second (∼90 m) for HydroSHEDS DEM and MERIT DEM.

The seven DEMs are tested over three areas of interest (see upper left panel of Figure 9.12), each one with different morphological and land-cover characteristics. These are, from north to south, Valsugana (a valley in a predominantly mountainous region in

northern Italy), the territory of Bologna Municipality and surroundings (an area characterized by flatland, hills and urban zones), and the territory of Rimini and surroundings (a coastal urban area on the Adriatic Sea). On these areas, the national DEMs are compared with the high-resolution DEMs obtained by the airborne LiDAR surveys performed for the Special Remote Sensing Plan (see Costabile, 2010), carried out by the Italian Ministry of the Environment and for Protection of the Land and Sea, which we assume here as ground truth. These reference DEMs have slightly different characteristics for each area of interest, with the resolution varying in a 1-2 m range, and vertical accuracy being between 15 and 30 cm. A series of tests are carried out to evaluate the vertical accuracy of the DEMs in respect to the reference LiDAR for the three areas, considering different terrain slopes (i.e., $\leq$5°, 5°-10°, 10°-30° and $\geq$30°), land cover type (i.e., urban, forest, low vegetation and crop fields, bare land) and HAND values (i.e., HAND $\leq$ 3 m and HAND $\geq$ 5 m). When performing these tests, the LiDAR datasets are resampled to the resolution of the tested DEM with a bilinear method.

In addition to the vertical accuracy, hydraulic consistency is also evaluated by comparing the Topographic Wetness Index (or TWI, see Beven and Kirkby, 1979) with the EU-Hydro photo-interpreted river network dataset (Gallaun et al., 2019), made available by the Copernicus Programme from the European Union (https://www.copernicus.eu/en). This choice is derived from preliminary analyses, which pointed out the strong agreement between the EU-Hydro dataset and the real Italian stream network. The TWI (see eq. 8) is directly computed from the DEMs via the SCA (i.e., specific catchment area, defined as the contributing area per unit width of contour, in meters), which represents the tendency of a pixel to receive water, and the slope ϕ, which represents the tendency to drain. Thus, it is strongly related to the water flow direction (Mattivi et al., 2019) and can be used as a qualitative evaluation of hydraulic consistency.

$$TWI = ln(\frac{SCA}{tan(\phi)}) \tag{9.5}$$

The vertical accuracy of the models is obtained by computing the residuals, $e(x, y)$, between the two DEMs (i.e., $e(x, y) = f'(x, y)$–$f(x, y)$, where $f'(x, y)$ is the surface of the DEM under analysis and $f(x, y)$ is the surface of the reference DEM). We refer to widely used performance metrics to quantify the accuracy of each DEM; in particular, we consider the linear error with 90% confidence (LE90), and the RMSE (eq. 9).

9.6.4 Analysis of available reference flood-hazard maps

In recent years, the large demand of reliable flood hazard maps of different scale (e.g. national, continental, global) exhilarated the development of various flood modelling methods and frameworks (Pappenberger et al., 2012; Winsemius et al., 2013; Yamazaki et al., 2011; Sampson et al., 2015; Dottori et al., 2016, e.g.,). As a result, plenty of flood hazard maps are available, but the comparison of different large scale models is not straightforward (Ward et al., 2015; Trigg et al., 2016; Lindersson et al., 2021), making critical the selection of a target map.

Here, we consider the two hazard maps used by Magnini et al., 2022 and we compare them in Italy based on their capillarity in representing flood hazard along the national river network. Indeed, having target flood hazard information for a greater number of minor streams, which have lower accumulation area, can lead to more effective training of DEM-based models (Magnini et al., 2022, e.g., see). The first map is the European-scale map made available by the European Joint Research Centre (JRC; see Alfieri et al., 2014) with a 500-year return period and 100m resolution, that has been developed as a component of the Copernicus European Flood Awareness System (EFAS, www.efas.eu).

The second map has been provided in 2018 by the Italian Institute for Environmental Protection and Research (ISPRA; see ISPRA, 2018) to fulfil the EU Floods Directive of the European Commission (2007/60/EC). It refers to a return period of about 500 years and is the merger of different hazard maps produced by local authorities. Resulting from local maps obtained with different methodologies, the ISPRA flood hazard map has notable heterogeneity: detailed flood hazard mapping characterizes some regions (e.g., see the north-western region of Figure 9.13), while mapping is sparser in other ones (e.g., see the north-eastern portion of the study area in Figure 9.13).

The evaluation of these two maps is carried out by considering the overlay with the EU-Hydro river network at a national scale(as for the DEM selection phase), that is selected as reference. Two steps are needed: first, the EU-Hydro shapefile is converted to a raster file with the same resolution and dimension of the considered hazard map. Second, computing the ratio between the number of flood-prone pixels falling on the river network and the total number of river-network pixels. This is done for separately for the different Strahler orders, and is used as a measure of capillarity and completeness of the target datasets for training DEM-based flood hazard models.

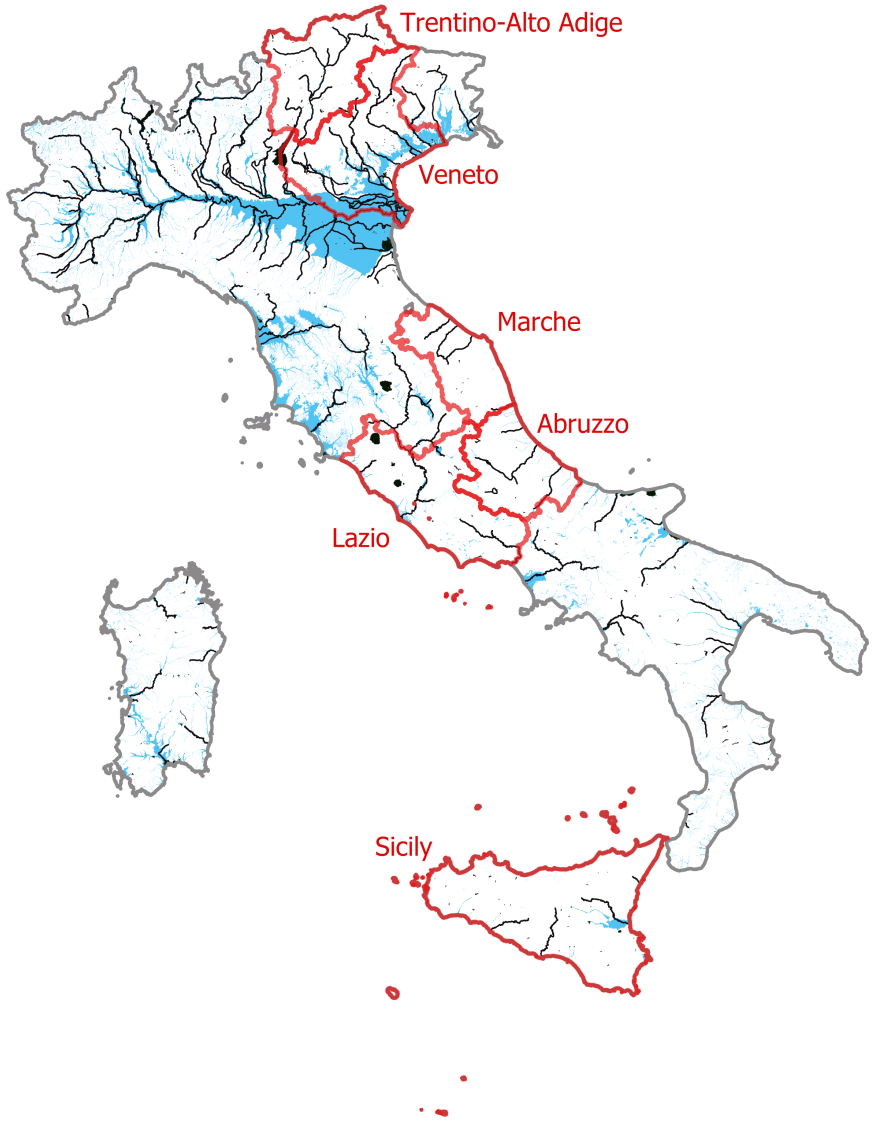

Figure 9.13: Flood hazard map with 500-years return period (light blue) released by ISPRA in 2018; in black: lakes and major rivers (Strahler order $\geqslant$5), from EU-Hydro dataset (©European Union, Copernicus Land Monitoring Service 2021, European Environment Agency (EEA)). In red: six Italian regions (see Section 7.1). Adapted from Magnini et al. (2023)

9.7 Results of the application to Italy

9.7.1 Selection of input DEM

Metrics for the evaluation of the vertical accuracy of the considered DEMs (i.e., RMSE and LE90, see Section 9.6.3) in the test areas (i.e., Valsugana, Bologna and Rimini, from North to South in Figure 9.12) are computed with reference to the LiDAR measurements (see Section 9.6.3) and are reported in Table 1. The three DEMs with the lowest errors are TINITALY, SRTM and MERIT for Valsugana, SRTM, HydroSHEDS, and MERIT for Bologna, and EU-DEM, HydroSHEDS and MERIT for Rimini. It can be also noticed that vertical accuracy for EU-DEM and SRTM is very similar in Bologna. Metrics computed for the different terrain slopes, land cover types and HAND values are not reported for the sake of brevity, as they confirm what seen in Table 9.4.

Table 9.4: metrics for vertical accuracy of the considered DEMs. Higher values (corresponding to worst accuracy) are marked with darker coloured cells.

		Valsugana		**Bologna**		**Rimini**	
DEM	**Resolution**	**RMSE**	**LE90**	**RMSE**	**LE90**	**RMSE**	**LE90**
TINITALY	10	8.236	12.141	4.511	6.475	2.832	4.412
EU-DEM	25	18.722	26.865	3.968	6.091	2.041	3.185
SRTM	30	15.764	21.083	3.203	4.665	2.239	3.537
ASTGTM	30	16.621	24.577	6.769	9.885	5.683	8.917
AW3D30	30	12.93	20.597	5.264	8.329	2.967	4.431
HydroSHEDS	90	30.166	44.597	3.355	4.549	2.108	3.362
MERIT	90	15.043	21.955	3.466	4.598	1.98	3.074

The analysis of TWI clearly shows that the hydraulic consistency of TINITALY is very low, while the other DEMs have similar performance. Thus, TWI is useful to exclude inappropriate DEMs, but not for numerically ranking the best ones. For the sake of brevity, only a detailed example of the comparison between TINITALY and EU-DEM within the Bologna test area is shown (Figure 9.14).

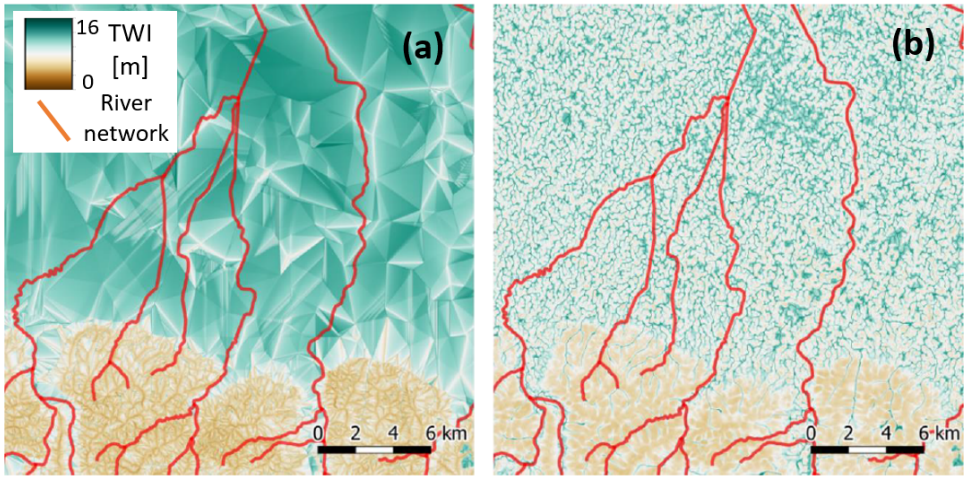

Figure 9.14: Focus on part of the Bologna test area. TWI computed from TINITALY (left panel) and EU-DEM (right panel); in red, river network from EU-Hydro dataset (©European Union, Copernicus Land Monitoring Service 2021, European Environment Agency (EEA)). Adapted from Magnini et al. (2023)

Given the combination of good performances in flat areas (i.e., Bologna and Rimini), the good hydraulic consistency, and the good resolution, the EU-DEM is selected as the most appropriate DEM for the present study. This decision is also enforced by the nature of EU-DEM, which was produced by hydraulic conditioning on the EU-Hydro river network (Bashfield and Keim, 2011), the same dataset largely used for analysing the results of the present study.

9.7.2 Selection of reference flood hazard map

Quantification of the agreement between the considered flood hazard maps (i.e., the ISPRA and JRC maps) and EU-Hydro river network over the study area is reported in Table 9.5. It is evident that flood prone areas in the ISPRA map comprehend a much higher portion of the river network than in the JRC map. This advantage is more significant for stream segments with lower Strahler orders, which is taken as evidence that, due to the threshold source area of 500km^2, the JRC map neglects a significative number of minor streams. As a result, the ISPRA map is selected as the target flood hazard map.

9.7.3 Reproduction of target hazard maps

Panels (a) and (c) of Figure 9.15 represent the standardized GFI (i.e., rescaled GFI values so that the maximum is equal to 1, highest susceptibility of being flooded, and the minimum is equal to 0, lowest susceptibility), and the p-value computed by the decision

Table 9.5: Percentage of overlay between EU-Hydro river network and considered reference flood hazard maps. Darker colour means higher percentage, which in turn means better agreement between the two datasets.

Map	**Strahler**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
ISPRA	10.30%	23.10%	37.60%	51.70%	64.80%	58.10%	99.60%	97.50%
JRC	0.02%	2.90%	3.20%	14.70%	58.80%	69.40%	90.60%	91.40%

tree (varying from 1, highest probability for a pixel of being classified as "floodable", and 0, lowest probability; see Chapter 2). These are the starting points from which the univariate and multivariate models derive their binary output maps, respectively. (see Panels (b) and (d) of Figure 9.15) represent respectively the flood-susceptibility maps obtained with the univariate and multivariate DEM-based flood hazard models. Both figures show strong similarity between the target and the output maps (i.e., dark redblue areas in panels (b) and (d) of Figure 9.15), which is particularly evident for the multivariate model.

Table 9.6 reports the performance metrics of the two DEM-based binary maps. It is evident that the multivariate method (second line) leads to better metrics (except for the TPR), but the difference with the univariate one (first line) is low, suggesting that the two binary outputs are very similar.

Table 9.6: Performance metrics for the DEM-based hazard maps computed for testing pixels located inside a 200m buffer area around the target flood hazard map. Highest and lowest values for each column are marked in bold and italic, respectively.

Model	**TSS**	**ACC**	**PPV**	**TPR**	**F1**
Univariate	*0.528*	*0.762*	*0.830*	**0.754**	*0.790*
Multivariate	**0.596**	**0.781**	**0.905**	*0.705*	**0.792**

An additional evaluation is performed by computing the overlay between the flood-prone areas of the DEM-based binary maps and the EU-Hydro river network (Table 9.7), as done for the selection of the reference ISPRA map (Section 9.5). The univariate model detects flood-susceptibility on a significantly higher portion of the river network than in the ISPRA map. The exception is Strahler order 8, where the overlay drops to 70.72%, suggesting inaccuracy of the model. The increase in the overlay for the multivariate model is lower, but still significant for Strahler orders from 3 to 6. For the others, the percentage of flood prone areas over the river network is substantially the same as in the reference ISPRA map.

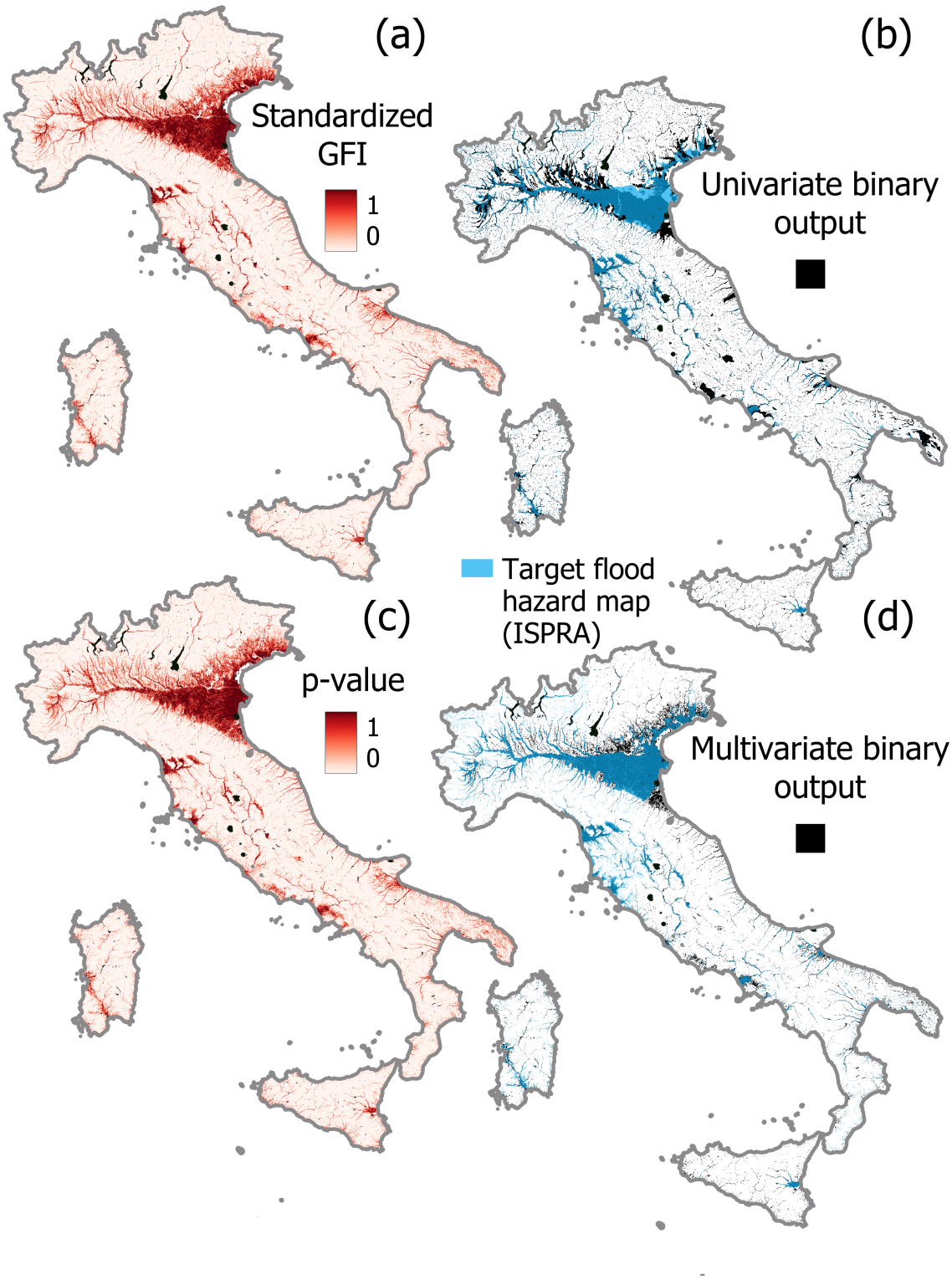

Figure 9.15: Output of DEM-based models - univariate model: standardized GFI values (red scale, panel (a)), binary flood hazard map (panel (b), black); multivariate model: p-value (red scale, panel (c)), binary flood hazard map (panel (d), black). In panels (b) and (c), blue represents the target ISPRA hazard map; dark blue identifies overlapping areas between the target and model maps. Adapted from Magnini et al. (2023)

Table 9.7: Percentage of overlay between EU-Hydro river network, reference ISPRA flood hazard and binary outputs from DEM-based models. Darker colour means higher percentage (see Table 2).

Map	**Strahler**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
ISPRA	10.30%	23.10%	37.60%	51.70%	64.80%	58.10%	99.60%	97.50%
Univariate	40.60%	69.70%	82.81%	91.05%	92.14%	84.64%	98.61%	70.72%
Multivariate	9.38%	22.03%	45.43%	71.41%	78.84%	73.64%	95.77%	95.95%

9.7.4 Validation against observed inundation extents and envelope flood hazard map

In Table 9.8 the agreement between the DEM-based models and the validation maps is showed, by means of the overlaying flood-susceptible areas. Considering the inundation maps (see the first three sections of Table 9.8 from the left), it is evident that the flood susceptibility map from the multivariate model has the lowest agreement with the validation dataset. However, with the exception of the AL 21/10/19 event, that is more poorly represented in all the maps, the overlaying percentage is always more than 85%. Differently, the envelope validation map has higher agreement with the multivariate model (i.e., 94.45%) than with the univariate (77.19%).

Table 9.8: Overlap between binary flood hazard maps (Target: reference flood hazard map, PGRA; Univariate: GFI DEM-based model; Multivariate: Decision Tree DEM-based model) and validation maps (i.e., observed inundation extents retrieved from satellite data, inundation scenario from 2D hydrodynamic modelling). Highest and lowest values for each column are marked in bold and italic, respectively.

	Observed inundation events						**Synthetic inundation scenario**	
	AL 21/10/19		**BO 20/11/19**		**BO 21/11/19**			
Model	**km^2**	**%**	**km^2**	**%**	**km^2**	**%**	**km^2**	**%**
Target	16.16	76.79	37.43	98.59	14.93	99.60	3347.62	99.72
Univariate	14.63	69.51	33.52	88.30	14.46	96.43	2591.19	77.19
Multivariate	12.06	57.30	33.09	87.17	13.25	88.36	3170.73	94.45

Regarding the values assumed by the standardized GFI and the multivariate p-value over the areas detected by the validation maps (9.16), higher median values of the latter are observed (i.e., p-value higher than 0.6, while standardized GFI lower than 0.4). It is again evident here that the event occurred in Alessandria is more critical to be modelled than the one in Bologna, as both the p-value and the standardized GFI are partially under the classification threshold over the area.

The continuous indexes and binary output maps of the DEM-based models can be

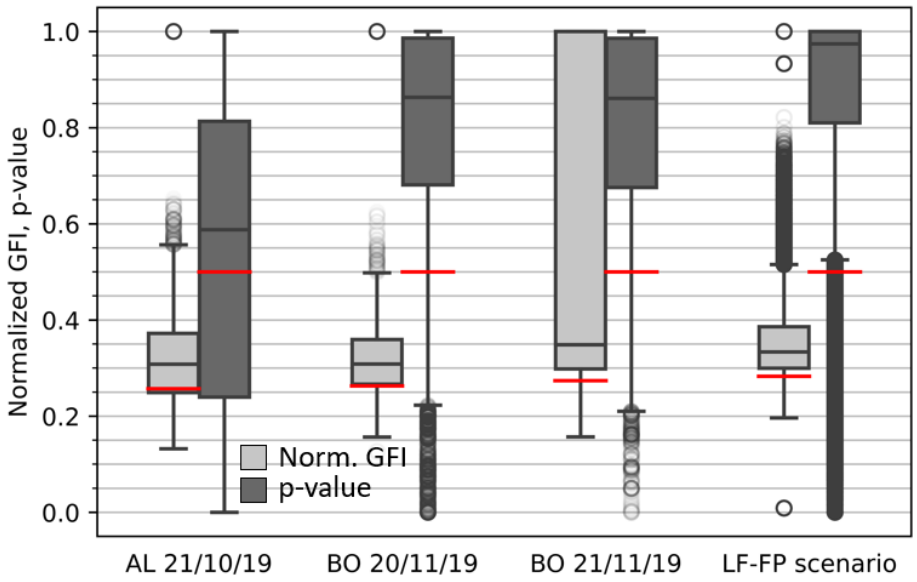

Figure 9.16: Boxplot of standardized GFI (univariate model, light grey) and p-value (multivariate model, dark grey) within the four inundated areas used in validation. Red lines indicate the thresholds for the DT classification (i.e., 0.5 at national level) and for the GFI classification (i.e., 0.265, 0.260, 0.249, 0.283 for AL 21/10/19, BO 20/11/19, BO 21/11/19 and the 2D envelope area, in this order). Adapted from Magnini et al. (2023)

better examined with specific focus on the validation areas of the AL 21/20/19 event and the synthetic scenario (illustrative examples in 9.17 and 9.18, respectively). From this analysis, two points emerge: (1) the multivariate model leads to discontinuous floodplain delineation, while sharp floodplain boundaries are produced by the univariate one (upper panels of Figures 6 and 7), and (2) the multivariate model is more efficient in characterizing flood hazard outside the main river network (panels (d) and (e) of 9.17 and 9.18).

9.8 Informativeness relative to the catastrophic event in Emilia-Romagna in May 2023

In this Section, the four outputs obtained from the analyses described in Sections 9.6 and 9.7 (i.e., the standardized GFI and univariate binary flood hazard map, and the p-value and multivariate binary flood hazard map) are further examined. In particular, their informativeness is investigated relative to a recent record-breaking event, occurred over Eastern Emilia-Romagna region (Northern Italy) in May 2023.

This event is different from the other ones considered in Sections 9.6 and 9.7, as it (1) is a real event, as the AL 21/10/19, BO 20/11/19 and BO 21/11/19, and (2) has

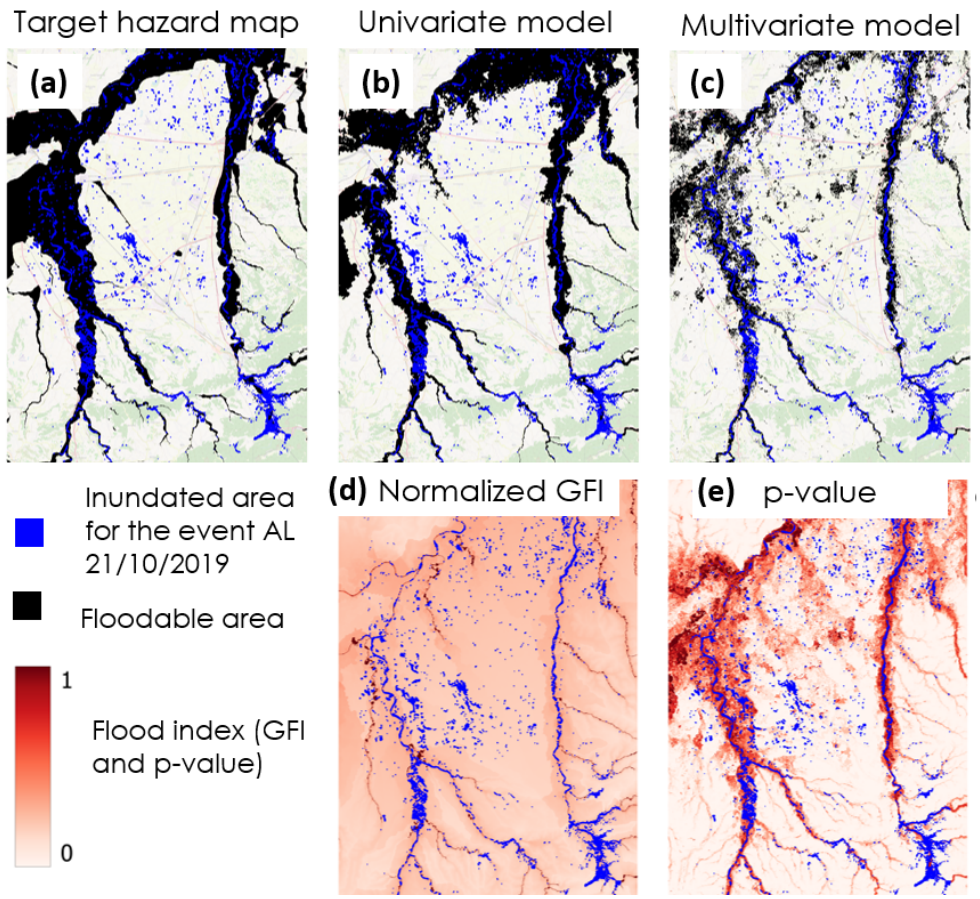

Figure 9.17: Upper panels: comparison of the floodable area (black) according to the target flood hazard, and DEM-based binary outputs with observed inundated areas (blue) (upper panels); lower panels: standardized GFI values and p-value (colour scale) compared with inundated areas (blue) for the AL 21/10/19 event. Adapted from Magnini et al. (2023)

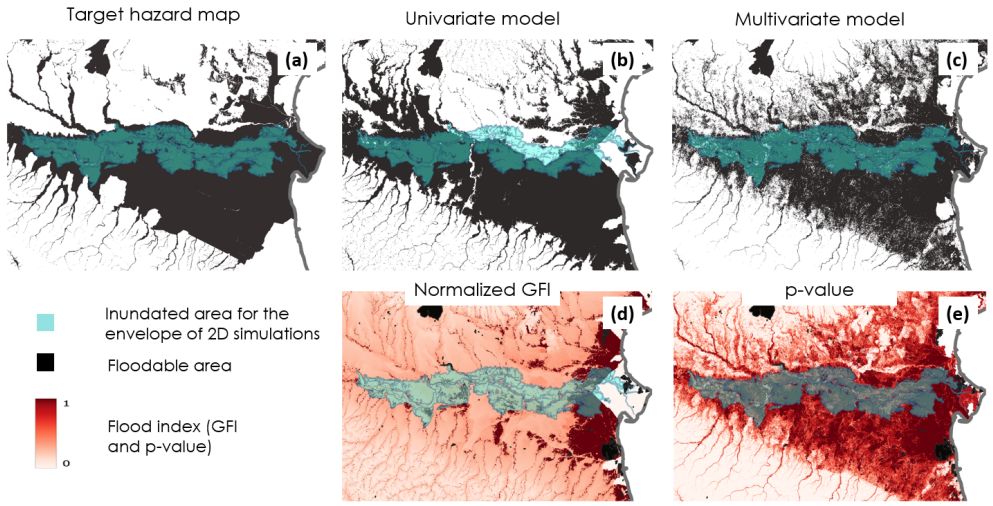

Figure 9.18: Comparison between envelope the synthetic inundation scenario (transparent light blue) and binary flood hazard maps (upper panels, from left to right: target map, univariate model, and multivariate model), and continuous flood-susceptibility indices (bottom panels, from left to right: standardized GFI of the univariate model, and p-value of the multivariate model). Adapted from Magnini et al. (2023)

a medium scale, as the synthetic scenario generated with LF-FP simulations, but (3) it involves several river catchments, which were almost simoultaneously subject to flooding.

During May 2023, two severe precipitation events occurred since May 2nd to 3rd, and since 16th to 17th. Both the rainfall events were caused by a cyclonic disturbancy, that brought large warm air masses to clash into the Appennines. Precipitation with unprecedented spatial extension and time persistency occurred, with return periods in some cases longer than 500 years (see Brath et al., 2023). The rapid succession of the two events led simoultaneously 23 rivers to overtop, since the soil was oversaturated by the first event. In total, about $540 km^2$ were flooded.

The investigation presented in this Section is based on the extension of the flooded areas after the event occurred on May 18th 2023. This dataset results from the merge of the products from the Copernicus Emergency Management Service's Mapping component. These consist of the outputs for post-processing operations from the acquisitions from different sensors at different times since 16th of May to 1st of June 2023. More detail is given in the official technical report (`https://emergency.copernicus.eu/mapping/ems/information-bulletin-167-copernicus-emergency-management-service-activities-following-latest`). The maximum extension of the flooded area is reported in Figure 9.19, where the ISPRA map released in 2017 is also represented. It is evident that a large percentage of the flooded areas are labelled as non-flood-prone according to the ISPRA map. This inconsistency is not present in the new version of the map, released in 2021 (Trigila et al., 2021), but the present analyses exclusively relate

on the 2017 dataset, since it served as target for training the models.

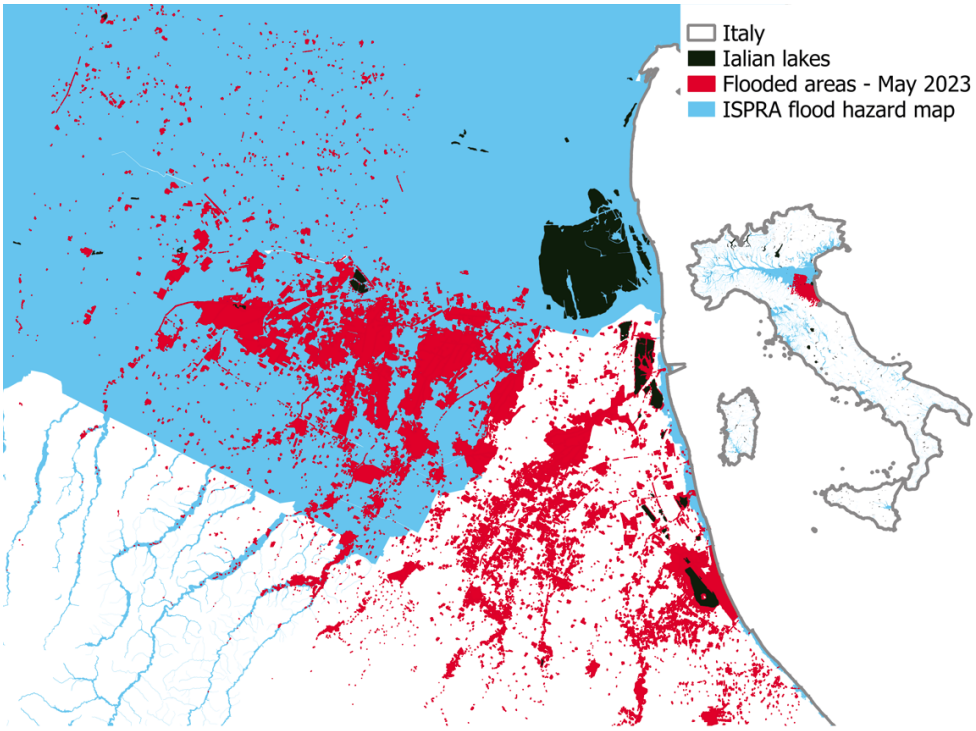

Figure 9.19: Boxplot of standardized GFI (univariate model, (a)) and p-value (multivariate model, (b)) within the inundated areas in Romagna region during May 2023. Red lines indicate the thresholds for the GFI classification (i.e., 0.27 as mean value) and for the DT classification (i.e., 0.5 at national level)

The analyses follow the same scheme adopted for the validation against the three inundation maps AL 21/10/19, BO 20/11/19, BO 21/11/19 and the LF-FP scenario (see Section 9.7.4). First, the percentage of overlap is reported in Table 9.9. Here, what observed in Section 9.7.4 is exacerbated, as the ISPRA map has a very low overlap (see also Figure 9.19). The univariate binary map is the one with the best overlap, while the multivariate one has characteristics in between the other two maps.

Table 9.9: Overlap between binary flood hazard maps (Target: reference flood hazard map by ISPRA, 2018; Univariate: GFI DEM-based model; Multivariate: Decision Tree DEM-based model) and validation map about Romagna region in May 2023. Highest and lowest values for each column are marked in bold and italic, respectively.

	Observed inundation event Romagna, May 2023	
Model	**km^2**	**%**
Target (ISPRA, 2018)	*316.67*	*58.47*
Univariate	**489.83**	**90.47**
Multivariate	426.89	78.81

Second, the values of the standardized GFI and the p-value are examined in the boxplots of Figure 9.20. Here, it is showed that the variability of the GFI values within the flooded areas is much wider than the one of the p-value; thus, the major extension of the flood-prone areas in the univariate binary map results from a lower threshold.

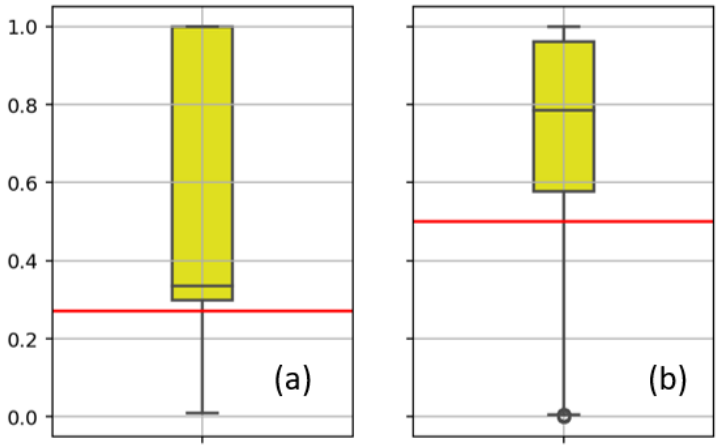

Figure 9.20: Boxplot of standardized GFI (univariate model, (a)) and p-value (multivariate model, (b)) within the inundated areas in Eastern Emilia-Romagna during May 2023. Red lines indicate the thresholds for the GFI classification (i.e., 0.27 as mean value) and for the DT classification (i.e., 0.5 at national level)

Accordingly, looking at the values of the two indexes of flood-proness in detail in the study areas, as it is done in Figure 9.21, one can promptly notice the improvement of the informative content due to the multivariate model with respect to the univariate one.

9.9 Discussion

The Discussion of the results is divided into six subsections. First, the methodologies adopted for the selection of the input DEM and flood hazard map are discussed. Then, three subsections are dedicated to the research questions that mainly are addressed within the application to Northern Italy. Namely, these are: (1) Can we profit from a blend of various geomorphic descriptors for flood hazard assessment and mapping? (2) Can we use simple ML techniques for effectively blending multiple GDs? (3) Are these techniques capable of providing a reliable assessment of flood hazard over large areas in extrapolation? Finally, the main research question for the application over all Italy is discussed: Can we use DEM-based models to enhance existing flood hazard maps?

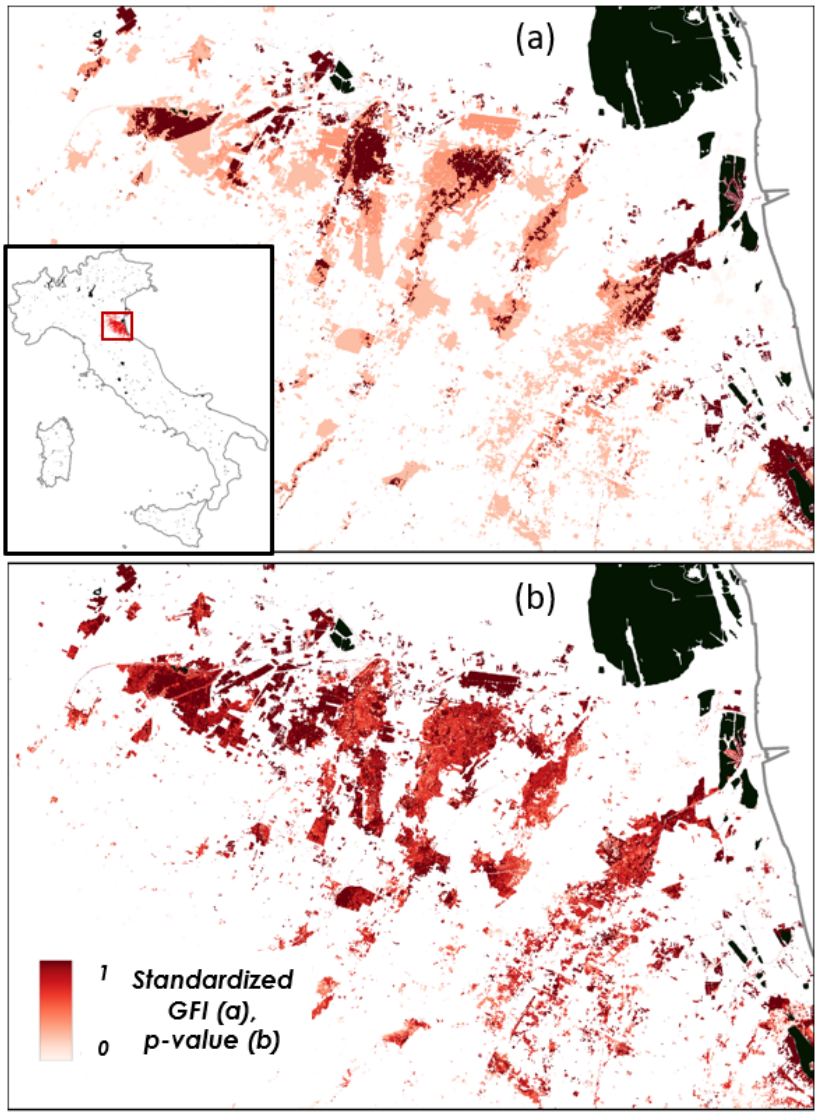

Figure 9.21: Standardized GFI values (a) and p-value (b) within part of the inundated areas (i.e., the province of Ravenna) in Eastern Emilia-Romagna during May 2023

9.9.1 Selection of the input DEM

On the importance of the selection of the appropriate DEM for calibrating a DEM-based flood hazard model, several studies are already present (e.g., Tavares da Costa et al., 2019). For the study on Northern Italy, the selection of the MERIT DEM is performed with a simple framework, based on two points: its declared hydrological consistency and and its horizontal resolution. This method is adapt in its specific application case, as the focus is to assess the goodness of the proposed models (i.e., the usage of DTs, the set of input GDs, the predictive power in extrapolation). Indeed, a DEM with finer resolution would have made increase the time and effort for processing the input data and training the input models (see Section 9.3).

In the study over the whole of Italy, a more sophisticated framework for performing this operation is given, trying to decrease subjectivity (see Section 9.6.3). The method presented is valuable as it considers several characteristics of the DEMs: vertical accuracy, hydraulic consistency and resolution. In fact, even if the vertical accuracy is the simplest way to evaluate a DEM, the calibration of a DEM-based model requires either good reliability of the river network extracted. Also, a DEM with higher vertical accuracy in flat areas, where the flood hazard is generally more difficult to estimate, has to be preferred over one with higher accuracy in mountainous contexts.

Indeed, one of the critical aspects of the proposed method is that some ground-truth data for evaluating vertical accuracy of DEMs could be not available. In this case, the selection should be based not only on the horizontal resolution, but also on the hydraulic consistency (e.g., Magnini et al., 2022). As an example, the river network or some geomorphic descriptors can be derived from the considered DEMs and compared with a reference river network (Tavares da Costa et al., 2019, see). Alternatively, studies showing that a specific DEM has good performance over the target area can be followed.

9.9.2 Selection of the reference flood hazard map

An objective evaluation of the most appropriate reference flood hazard map is very difficult, as each map has its own advantages and weaknesses (see Section 9.6.4). In these studies, a framework for a quantitative selection is given. The Authors are aware the coverage of river network extent and the accuracy of flood hazard estimation are two extremely different concepts. However, it is evident that in the case of the present study, there are several areas of Italy where the minor streams are susceptible to floods with 500years according to the ISPRA map (see, e.g., the extreme north-western spot of Italy), while in JRC modelling they are not. Thus, it is reasonable to assume that these

minor streams are accurately modelled in the ISPRA map, and that they do not appear as flood-susceptible in the JRC map due to the source area threshold (see Chapter 7). Thus, evaluating the portion of modelled river network that is covered by flood prone areas can be an effective way to quantify the comprehensiveness of the hydraulic modelling, and is used here as a selection method for the reference hazard map.
This approach can be suitable for large study areas, as in the present study, but for medium to small scale applications other aspects of the reference maps should also be considered (e.g., the way hydraulic structures were modelled). Finally, the proposed method requires to have a river network dataset to assume as reference, which should be rather feasible, thanks to open source online resources as EU-Hydro. Nevertheless, in case no dataset is available, a user-defined river network can be obtained by manual extraction from satellite data.

9.9.3 Can we profit from a blend of various geomorphic descriptors for flood hazard assessment and mapping?

The first goal of the present research is the evaluation of the improvement which can be obtained by applying a machine-learning aided multivariate DEM-based flood hazard assessment relative to a univariate DEM-based approach.
This research question is addressed by comparing the results obtained with the univariate and multivariate DEM-based models, and is conveniently divided into two sections, each one dedicated to a study area.

Application to Northern Italy

First, regarding the classification problem (i.e., differentiation between flood-prone and flood-free areas), the outcomes reported in Figures 9.6-9.7 and Table 9.1 (rows 1-2) suggest that the combination of multiple geomorphic descriptors (GDs) increases the comprehensiveness of the morphological description of the study area. The resulting multivariate data-driven model can reproduce the reference flood hazard map in a significantly enhanced way relative to a univariate approach adopting a single GD. This is particularly visible from the lower extension of wrongly-predicted areas (i.e., false positive, or FP, and false negative, or FN) in the classifier DT output map (ligth red and blue areas in Figure 9.6) relative to the GFI output map (light green and blue areas in Figure 9.7).

Second, concerning the regression problem (i.e., prediction of the flood intensity, such as the expected maximum water-depth associated with a given probability of occurrence)

the regressor DT considered for interpolation shows high accuracy in reproducing the target map.

Also, it is worth highlighting that regressor DTs provide a direct estimate of this variable, relative to the traditional univariate DEM-based approaches, which usually requires the prior delineation of flood extent to compute water depth, as the elevation difference between the flood-extent border and each pixel (see Manfreda and Samela, 2019).

Figure 9.9 highlights that the correlation between the predicted and target water depths can be improved, yet it also clearly shows that predictions for the test set are unbiased. It is worth mentioning here that the diagram neglects the true negatives (i.e. target and predicted water depths are equal to $0.0m$; 49.78% of the cases), false positives (i.e. only predicted water depths are equal to $0.0m$; 22.37% of the cases) and false negatives (i.e. only target water depth are equal to $0.0m$; 0.08%). While the occurrence of the most concerning cases (false negatives) is very limited, predictions show significant margins for improvement as far as the false positives are concerned. Nevertheless, it should also be recalled here that the target map by its own very nature neglects smaller streams (contributing area has to be higher than $500km^2$), whereas the decision tree regressor looks at morphology only and provides water depth predictions also for smaller streams (i.e. higher exhaustiveness, see Figure 9.8).

One of the most interesting aspects is the relevance that each GD assumes in the regressor DTs (see Table 9.3). It can be observed that all models rely mainly on one single GD, with Gini importance always in excess of 60%, but still, the multivariate analysis leads to significantly better results relative to the univariate one. Also, it is important to highlight that:

- While regressor DTs tend to depend mainly on the GFI, classifier DTs depend on HAND
- While the input GDs have quite a similar Gini importance hierarchy in regressor DTs, classifiers DTs assume different hierarchical structures depending on the considered training area
- All models agree in giving low Gini importance to LGFI and TI_m, probably due to redundant information relative to GFI
- Elevation is very often ranked second, always associated with significant importance

Overall, this suggests that regressor DTs tend to operate by correcting a baseline estimate that mostly relies on the GFI value. On the other hand, classifier DTs obtain their

results by following different rules depending on the training data, and often prefer using lower-levels features relative to more complex indicators such as the GFI. This sensitivity to the training area makes difficult to set a priori weights to the GDs when building up the models. It should be kept in mind, however, that different Gini importances do not necessarily imply radically different classification rules, due to the existing correlations between the input features. Ideally, dedicated feature selection and importance analysis algorithms should be used to obtain deeper insight on how the different models come to their conclusions; we plan to investigate this line as part of future work.

Application to all Italy

When comparing the binary maps obtained across the entire study area (Table 9.6), the multivariate model achieves slightly better metrics than the univariate one, confirming what observed for Northern Italy. When focusing on smaller scales, the univariate model performs slightly better for single inundation events, yet significantly worse at a river branch scale (Table 9.8, sections "Observed inundation events" and "Synthetic inundation scenario", respectively).

On one hand, better overlap between inundation events and the univariate binary map confirms the validity of GFI. In fact, while its computation is straightforward, its accuracy for inundation susceptibility can be locally very significant. In particular, this is true when the thresholding is performed through a watershed-wise strategy, as in this case. On the other hand, the multivariate model trained nationwide shows higher accuracy in reproducing large inundation scenarios obtained through hydrodynamic modelling (Figure 9.18). This confirms what observed for Northern Italy, and is a clear indication that considering a variety of morphological descriptors as opposed to a single index leads to a better delineation of the envelope of all possible inundation events, which is the true objective of geomorphic floodplain delineation (see also dark blue areas in panels (b) and (d) of Figure 9.15 on this point).

Differently than for Northern Italy, probably due to the usage of two distinct DEMs, numerous discontinuities (i.e., isolated non-floodable pixels or small pixel clusters in floodable areas and vice versa) can be identified in the multivariate binary map (right-upper corners of Figures 9.17 and 9.18) by looking at specific areas in more detail. These discontinuities result from the combination of the p-value thresholding (p-value=0.5), that is required by the production of a binary map, with the nature of an approach based on decision trees. In fact, as this multivariate approach is pixel-based, it does not explicitly enforce spatial coherence of the output. These isolated pixels are behind the lower metrics obtained while validating the multivariate flood-hazard map against the

inundated areas for specific flood events (columns "Observed inundation events" in Table 9.8). On the contrary, discontinuities are not present in the univariate flood-hazard map, due to the hydrological topologic consistence that characterises GFI. In fact, the contributing area increases and the elevation above the nearest river-pixel decreases moving downstream, implying a monotonic increase of GFI in the same direction. Nevertheless, the GFI local descriptiveness of the actual susceptibility of a pixel of being inundated may vary from region to region. Thus, relying only on GFI can lead to improved accuracy in mountainous areas and upper river segments (central upper panel in Figure 9.17), and simultaneously to very significant and spatially broad inconsistencies in predominantly flat large areas (see e.g. the area North to the Po river in the central upper panel of Figure 9.18). These findings are in line with the literature, where consistence of univariate models is found to be higher in floodplain areas unaltered by humans (Nardi et al., 2018, e.g.,) and influenced by river stream order (Annis et al., 2019).

9.9.4 Can we use simple ML techniques for effectively blending multiple GDs?

The second research question of the study over Northern Italy is wheter it is possible to obtain a good estimation of flood hazard by combining multiple GDs with low-complexity machine learning models. Differently from several other contributes in the literature, we do not focus on model complexity nor on the comparison of different models (Wang et al., 2015; Khosravi et al., 2018; Mosavi et al., 2018; Arabameri et al., 2019; Costache et al., 2020). Instead, we prefer to select one simple model type (i.e., decision trees, DTs) and focus on the combination of the five innovative elements listed in the Introduction; in this way, we can analyse the influence on the multivariate DEM-based approach of the preliminary steps, consisting in data pre-processing (i.e., selection and manipulation of input features, target maps, training set and test set). This is highly important, because machine learning models do not reproduce the dynamics of the water, as such, their performance is strictly linked to the data used for the training, that need to be handled very carefully.

As it is highlighted in Section 9.9.3, the outcomes of the study over Northern Italy (Figures 9.7-9.8, Tables 9.1-9.2) clearly show that DTs can effectively reproduce the target information (Figures 9.2-9.3) with high accuracy for both classification and regression problem, even if the resolution of the MERIT DEM (Yamazaki et al., 2017), from which the input GDs have been retrieved, is not very high.

Indeed, even if regressor DTs necessarily implicate discretization of the output vari-

able, in the present study for Northern Italy large datasets and appropriate tree depth allow us to obtain wide ranges of different water depth values.
Moreover, the trained DTs estimate flood hazard associated with different minor streams that are neglected in the target maps (see red areas in Figure 9.6; compare Figure 9.8 with Figure 9.3). Due to the absence of information in these areas, it is not possible to assess the goodness of the models output, but this tendency of completing target information could be a key aspect for future applications to data-scarce or inaccurately mapped regions, and thus, it could be considered as a promising characteristic of the models. Additional considerations on this aspect follow in the next sections of the Discussion.

Overall, it is possible to observe that DTs are effective tools to combine GDs and estimate flood hazard. This indicates that proper data handling has a strong influence on the accuracy of the final estimation, which is comparable to the choice of a given machine learning technique. In particular, we want to underline two elements of the presented approach that have great importance on the predictive skill.

First, the utilization of flood hazard maps as target results in a large number of pixels for the training and test set, and therefore a very broad spectrum of hydrological/morphological characteristics, which represent a much more informative dataset relative to isolated points used by other authors for training more complex models (Lee et al., 2017; Khosravi et al., 2018; Arabameri et al., 2019; Janizadeh et al., 2019).

Second, a sensible identification of a calibration area is very important for a successful training, as it allows to neglect irrelevant pixels. To this aim, a preliminary sensitivity analysis might be very useful for identifying the optimal buffering radius around the target map (see Section 9.4.1), even if different approaches are proposed in the literature (e.g., Degiorgis et al., 2012). Indeed, in the case of application of DEM-based methods in data-scarce areas, where local flood-hazard modelling datasets may not be available, global or continental flood hazard maps produced by the European Joint Research Centre (Dottori et al., 2016, 2021) can be used as a target.

These observations lead to the choice of using the same methodology for the application over all Italy. In this case, the larger extension of the study area and the finer resolution of the EU-DEM entail processing a huge amount of pixels from extremely different morpho-climatical systems.

The discontinuities in predicted flood susceptibility described in Section 9.9.3 reveal that DTs can produce spatial inconsistencies in some cases. This is probably due to the size of the study area and the nature of the input DEM, and partially solved by considering a continuous characterization of flood hazard instead of a binary mapping

(Figures 9.15, 9.16). However, despite this drawback, the multivariate output has good evaluation metrics (Tables 9.6 and 9.8), and DTs provide a better estimation of flood hazard over large flat areas (Section 9.9.3). This is remarkable, as the calibration of the univariate model is regionally-wise, while the one of DTs is unique for the whole of Italy.

Overall, the results obtained from this second study confirm the effectiveness of multivariate DTs for (a) reproducing accurately and (b) completing the target information.

9.9.5 Are these techniques capable of providing a reliable assessment of flood hazard over large areas in extrapolation?

The evaluation of prediction accuracy for geographical extrapolation (i.e., applying the models in geographical areas, or watersheds, that have not been considered for parameterization and training) is a key and characteristic aspect of the application over Northern Italy.

On the one hand, performing predictions with new input data is a major problem for machine learning models; on the other hand, reaching good predictive skills in extrapolation is needed for future practical applications to data-scarce environments. What is more interesting about this part is to understand the link between training and test performances: if the relationship between input and target values, learnt by the model during the training, is also valid for the extrapolation region, accurate test predictions are obtained, but this depends strongly on the choice of input and target datasets for the training, which can be very difficult. Before addressing this very issue, a careful discussion of the resulting metrics and maps is required, as their interpretation is not straightforward.

With reference to the classification problem, each metric suggests a different training area as the best case, and this highlights how difficult it is to choose a single metric for describing the goodness of a model for a binary classification. Figure 9.10 and TSS values in rows 4-5 of Table 9.2 could suggest that Area B (test TSS=65%) has better extrapolation performance than Area C (test TSS=33%). On the contrary, ACC is similar for the two cases, and higher for Area C (ACC=88%) than for Area B (ACC=85%), suggesting that TSS is a more informative metric than ACC in representing the model performance. On the other hand, precision and recall appear to be quite unbalanced metrics, as areas A and D lead to test prediction with considerable overextension of FN and FP values, respectively (see Figure 9.10). Differently, regression metrics agree in pointing at the DT trained in B as the best case (Table 9.3). However, the absolute values of R^2, that

depicts low-accuracy test predictions, do not reflect other metrics (MSE and MAE) and the output maps (Figure 9.11).

As expected, the choice of the training area has great influence on prediction accuracy. This is particularly visible for the classification problem: in Figure 9.10, the difference between metrics for training and test is striking. Nevertheless, this difference becomes less clear for the regression problem (Figure 9.11). The same observations are confirmed by Table 9.3, where evidence is given of different structures for the classifiers DTs, while the regressor DTs are all very similar. More in detail, the obtained results show that the extent of the training area has less importance than the quality of the input data that it contains. Perfect examples of this remark are classifiers DTs trained in A and D: even if both A and D are very wide, prediction over the test area is affected by considerable errors. This happens because A does not include any part of the Apennines, while D ignores a large flat area in the eastearn coast, meaning that any geographical system corresponds to a specific relationship between input GDs and flood susceptibility, and thus it cannot be fully represented by a model trained with very different datasets. The comparison between area B and C is also meaningful: while the training in B leads to good test predictions for the classification, it is the worst case for the regression (the opposite is valid for C). This is probably due to the fact that area B contains useful information to delineate flood-prone areas, as it represents the upstream chapter of Po river, but cannot adequately train a regressor DT, as it lacks high target values (i.e., high inundation water depths). To sum up, GDs combination with DTs is capable to provide quite a reliable estimation of flood hazard (i.e., flood-prone areas and maximum water depth) in extrapolation mode, but a careful choice of the training area is needed, where target and input dataset is complete and representative for the test area.

9.9.6 Can we use DEM-based models to enhance existing flood hazard maps?

Resolution of weaknesses of the reference flood hazard map

Finally, the increase in the extension of flood-prone areas in the DEM-based outputs with respect to the reference ISPRA map should be discussed. This is a major task, as no method exists to evaluate geomorphic-based flood hazard information where no reference is available from hydraulic models. However, it is quite reasonable to expect areas close to rivers to be flood-susceptible when considering a 500-years return period (see also Section 9.9.2); thus, the overlap between the flood-prone areas and the river network is considered in the present study.

At a national scale, only the overlap for the DEM-based outputs can be compared just with the one for the reference map (Table 9.7), while at a regional scale, the JRC map can also be used (Table 9.10). In fact, in some regions the JRC has a major overlap with the EU-Hydro river network than the ISPRA map, showing inconsistencies of the latter. In these regions (Figure 9.13), the overlap values for the DEM-based models and JRC map are very similar, and both higher that the ISPRA map (Table 9.10). This shows that geomorphic approaches can effectively use the reference information collected elsewhere to accurately predict flood susceptibility where exact reference is not available. Globally (Table 9.7), it is possible to consider as an advantage the increase in overlap for high Strahler orders (i.e., 3, 4, 5 and 6) observed for the multivariate model, as it represents an advance with respect to the probable inaccuracy of the reference ISPRA map. Differently, the huge overlap for minor streams (i.e., Strahler orders 1, 2 and 3) for the univariate model should be considered suspiciously.

The results from the further investigation against the event occurred in May 2023 in the Romagna region are in line with the previous observations. Both the DEM-based models have greater overlap with the flooded area: while the univariate has about 100% overlap, the multivariate has characteristics more similar to the target ISPRA map, even if with a significant improvement of the agreement with the validation dataset (see Table 9.9).

It is important to mention that the updated of the ISPRA flood hazard map, released in November 2021 (Trigila et al., 2021), does not have the same inconsistencies as the version of 2017. However, the analyses described are still useful and valid, as the presence of flood hazard maps with low accuracy is very frequent worldwide.

These findings seem to confirm that DEM-based models, in particular the multivariate ones, can be useful to complete flood hazard information where it is already available but with some inconsistencies (see also Lindersson et al., 2021).

Floodplain delineation vs flood hazard characterization

To conclude, concerning the informativeness of a binary geomorphological flood hazard modelling, as opposed to a continuous representation of flood hazard, it is worth comparing the standardized GFI values and the corresponding p-values from the multivariate approach. By looking at the left panels of Figure 9.15, and the lower panels of Figures 9.17 and 9.18, all adopting the same colour scale, it is possible to observe that both the indices assume higher values in floodplains and river proximity, correctly estimating flood susceptibility in the entire spatial domain. However, the differences in the strength of the regional patterns produced by the two modelling approaches are rather

Table 9.10: Percentage of overlap between EU-Hydro river network, reference ISPRA flood hazard, JRC map, and binary outputs from DEM-based models. Darker colour means higher percentage (see Table 9.5). Only regions and Strahler orders where overlap for JRC is significantly higher than for ISPRA are reported. When regions do not have rivers with 6 Strahler order, values are missing.

Region	Map	Strahler 3	Strahler 4	Strahler 5	Strahler 6
Trentino Alto-Adige	ISPRA	3.34%	10.21%	16.38%	5.05%
	JRC	5.92%	4.85%	46.32%	71.34%
	Univariate	57.61%	73.46%	87.89%	92.46%
	Multivariate	16.08%	39.28%	52.47%	53.73%
Veneto	ISPRA	11.03%	4.37%	16.41%	20.50%
	JRC	10.62%	3.37%	49.66%	55.44%
	Univariate	54.46%	65.95%	72.71%	65.02%
	Multivariate	29.70%	48.05%	63.41%	59.13%
Marche	ISPRA	2.52%	2.21%	0.19%	-
	JRC	3.16%	19.92%	47.55%	-
	Univariate	87.47%	92.65%	92.73%	-
	Multivariate	60.39%	80.16%	74.62%	-
Lazio	ISPRA	14.78%	35.56%	76.87%	48.63%
	JRC	4.90%	22.54%	79.57%	83.57%
	Univariate	72.47%	86.91%	98.72%	91.30%
	Multivariate	48.41%	72.10%	90.69%	84.72%
Abruzzo	ISPRA	23.36%	34.12%	49.23%	66.34%
	JRC	2.01%	5.65%	61.13%	87.04%
	Univariate	74.03%	85.97%	92.59%	95.90%
	Multivariate	35.59%	64.65%	72.39%	89.31%
Sicily	ISPRA	10.17%	35.59%	19.18%	-
	JRC	1.01%	16.90%	78.50%	-
	Univariate	84.28%	93.31%	97.90%	-
	Multivariate	59.39%	82.19%	84.16%	-

striking. On one hand, the p-values show a smooth and gradual decrease moving further away from the river network, and higher values in floodplains relative to standardized GFI values.

On the other hand, very high standardized GFI values can be found in well-defined areas and very close to the river network. Then, the univariate flood susceptibility decreases rather abruptly when moving from the river network to its immediate proximity, having an overall variability between 0.2 and 0.0 from a river floodplain to a mountain peak. This is confirmed by the boxplots of Figure 9.16, showing that the inundated areas of the studied flood events are characterized by lower standardized GFI values and narrower ranges relative to the p-values associated with the multivariate model. An additional confirmation is the low GFI thresholds (red lines in Figure 9.16) obtained in calibration for the four considered areas, and the corresponding wider floodplains de-

lineated through the univariate flood hazard model.The investigation against the catastrophic flood in Emilia-Romagna in May 2023 is an additional confirmation. Here, the GFI values are still lower than the p-value in average, but their range is wider (Figures 9.20 and 9.21). This is due to the extension of the inundations, that include even areas far from the main river bed. Hence, a better agreement between the p-value and the inundation extent is confirmed.

The above considerations can support a more efficient and effective use of DEM-based flood hazard modelling products. In fact, most of the scientific literature focuses on the capability of these models to reproduce binary target flood-hazard maps. However, these models produce a great deal of information on susceptibility to inundation that is mostly lost if binary mapping, which we could also refer to as "floodplain delineation", is preferred to a continuous representation of flood susceptibility, which instead is closer to a "flood hazard mapping" in the strict sense. Just a few studies (e.g., Avand et al., 2022; Deroliya et al., 2022; Costache et al., 2020) show the application of DEM-based approaches for obtaining spatially continuous flood-susceptibility maps, but their models are trained on a pool of single inundation events. Thus, their analyses do not extensively focus on the information gain with respect to floodplain delineation, and rather show how their approaches can combine sparse information into a coherent output. Differently, we train our models on binary flood hazard maps, and examine how these methods can improve information on flood hazard maps that is already available.

In this context, continuous representation of flood susceptibility can be used for the entire study domain, or limited to a buffer of delineated floodplains (i.e., binary maps), proving a graphical representation of the uncertainty of the binary map itself. Based on the outcomes of the present study, when a continuous spatial representation of flood hazard is considered, a simple multivariate approach to geomorphic flood hazard modelling seems to be associated with much higher potential and informativeness of a univariate modelling adopting a single, and yet very effective, morphological index.

Overall, the analyses above show the effectiveness of a new application of DEM-based models. So far, most of the literature describes how they can mimic pre-existing flood hazard maps (e.g., Nardi et al., 2018; Manfreda et al., 2014), derive spatially continuous estimation of flood susceptibility from multiple single measurements of inundation events (e.g., Costache et al., 2020; Avand et al., 2022), or predict flood hazard over data-scarce regions (e.g., Magnini et al., 2022). Differently, we gave evidence that DEM-based modelling can be used for enhancing incomplete or inexact information contained in national target flood hazard maps. This is possible thanks to their natural capabilities to produce spatially homogeneous and continuous flood-susceptibility maps, which seem

to gain robustness when using a variety of morphometric indices instead of a single one. So far, the main limitations of geomorphic approaches, in particular in coastal and flat areas as observed by other authors (e.g., Lindersson et al., 2021), remain unsolved. Also, evaluating the goodness of the additional information produced by DEM-based models with respect to the reference maps (e.g., minor streams) is very challenging. In these cases, appropriate use of global flood hazard maps and ancillary datasets (e.g., JRC flood maps and EU-Hydro river network) can be resolutive.

9.10 Conclusions

This Chapter analyzes and compares data-driven and resource-efficient methods for assessing and mapping riverine flood hazard across large geographical areas. It illustrates the potential and limitations of combining different geomorphic descriptors by means of decision trees for delineating flood prone areas and for predicting the expected maximum water depths for a given return period.

Two consecutive applications of a common methodology are outlined. First, we focus on a large study area in Northern Italy (size $\sim 10^5 km^2$) containing Western, Central and part of the Eastern Italian Alps, part of the Northern Apennines and the floodplains of a complex river-system including the main rivers Po, Adige, Brenta, Bacchiglione and Reno. The morphology of the study area is described by the Multi-Error-Remover Improved-Terrain model (MERIT DEM; Yamazaki et al., 2017), with a 90-meter resolution, approximately. Decision trees are trained using as input features the geomorphic descriptors retrieved from the MERIT DEM, and as target maps two different datasets: one representing flood extent with a reference return period of 500 years, and one representing expected maximum water depth for a 100-year return period scenario.

In the second application, the whole of Italy is considered; the EU-DEM, with 25-meter horizontal resolution, is used to describe the study area, while the official flood hazard map for Italy is used as reference for the training. In both the studies, univariate and multivariate DEM-based models are built up. The univariate models rely only on the geomorphic flood index (or GFI; Manfreda et al., 2014), which is one of the most accurate and versatile (e.g., see Samela et al., 2017). The multivariate models use a data-driven blend of various DEM-based indices, including GFI.

Relative to previous studies focusing on morphometric floodplain delineation and flood-hazard mapping (see e.g., Dodov and Foufoula-Georgiou, 2006; Nardi et al., 2006; Manfreda et al., 2011, 2014, 2015; Samela et al., 2017; De Risi et al., 2018) and

machine-learning aided multivariate flood hazard mapping (see e.g., Gnecco et al., 2017; Arabameri et al., 2019; Janizadeh et al., 2019; Costache et al., 2020), our analyses represent a relevant advacement of knowledge. In fact, several elements are simultaneously combined: (a) an effective framework for selecting the appropriate input DEM and reference flood hazard map is presented, relying on the open-source EU-Hydro river network dataset (Gallaun et al., 2019); (b) only strictly DEM-based morphometric data and indices are used for predicting flood hazard; (c) morphological characterization of flood hazard associated with a given probability of occurrence is studied separately as a classification problem (i.e., generation of binary flood hazard maps) and as a regression problem (i.e., prediction of expected maximum inundation water depth); (d) machine learning models (i.e., decision trees) are trained using pre-exhisting flood hazard maps as target information; (e) univariate geomorphological assessment of flood hazard (i.e., one geomorphic descriptor used as predictor) is thoroughly compared with a multivariate assessment, in which several DEM-based geomorphic descriptors are blended together by means of decision trees; (f) potential and accuracy of DEM-based flood hazard prediction is assessed in geographical extrapolation by applying models trained on specific geographical areas to different areas having diverse morphologic and/or hydrological features; (g) abilities of the models is investigated to resolve heterogeneities and inconsistencies of the same reference map used for training.

In particular, we address four main science questions: (1) can we profit from a blend of geomorphic descriptors to perform flood hazard mapping with respect to a univariate DEM-based approach? (2) Are decision trees a valid tool for combining multiple geomorphic descriptors? (3) Is this approach capable to predict flood hazard over large areas in geographical extrapolation? (4) Can we use DEM-based models to enhance existing flood hazard maps?

With reference to the first and second questions, the models are trained and tested with different sets, consisting respectively in randomly-selected 85% and 15% of the pixels contained in a buffer area around the flood-prone areas of the target maps. The results obtained for the classification problem (i.e., binary flood susceptibility mapping, or floodplain delineation) show high performance metrics in validation relative to the univariate approach. In particular, the combination of DEM-based descriptors leads to much more accurate results in flood-prone areas delineation over predominantly flat regions. Concerning the regression problem (i.e., estimating maximum inundation water depth associated with a 100-year return period), good performances are confirmed in validation as well.

Also, with reference to the third question, we test the proposed approach in a second mode, which we termed geographical extrapolation. We delineate four different subregions of Northern Italy to train classifier and regressor decision trees by selecting four areas belonging to four different hydrologically-coherent geographical systems. When tested on the remainders of the first study area, the four different models show different extrapolation performances depending on the morphological features (e.g. Apennines vs. Alps) and the broadness of the hydrological conditions included in the training subregions. In particular, concerning the classification problem, models trained in areas containing headwater catchments of the main rivers can extrapolate better over the downsteam portions of the basins than vice versa. Concerning the regression problem, the selection of the training area must rely not only on these morphological and hydrological features, but also on the availability of a sufficently wide range of values for the target variable (i.e., maximum water depth in our case) within this area, in order to adequately train the model. This means that training in headwater catchment areas performs very poorly for extrapolating maximum water depth across downstream floodplains.

Concerning the fourth research question, we focus on a national scale, and analyse the ability of DEM-based models of handling the secondary river network, and producing spatially-continuous and homogeneous characterization of flood hazard. These advantages are naturally offered by geomorphic methods, but yet just partially exploited (e.g., see Deroliya et al., 2022), and not fully discussed. Accordingly, we validate the univariate and multivariate models against independent information, that is remotely-sensed inundated areas during three different flood events, and a synthetic catastrophic inundation scenario obtained as the envelope of several 2D hydrodynamic simulations. In brief, it is observed that (a) multivariate DEM-based models can be used to complete flood hazard information (with reasonable uncertainty) where the reference map is incomplete or inaccurate; (b) the spatially continuous representation of flood susceptibility (flood hazard mapping), should always be preferred to a binary representation (floodplain delineation) as it provides a wealth of information, e.g. on uncertainty and descriptiveness of the simplified DEM-based model; (c) in case spatially continuous flood susceptibility maps are to be prepared, multivariate approaches (e.g., p-value from a decision-tree classifier) seem to be preferable to univariate ones (e.g., GFI alone) due to their higher descriptiveness and information content.

In conclusion, multivariate DEM-based analysis by means of decision trees is very effective in estimating flood hazard relative to univariate approach, and that these tech-

niques have good potential in extrapolation mode as well. Also, we suggest a more effective use of national flood hazard maps obtained from DEM-based models: instead of mimicking the national target flood hazard maps, which are often heterogeneous and inexact, we should rather exploit the training information to produce spatially homogeneous and continuous flood-susceptibility maps.

Different elements of this work can be further examined in future studies, in order to deepen the collective knowledge and understanding of the DEM-based multivariate techniques. First, classifier and regressor decision trees could be compared with other multivariate approaches, whose training is based on different target maps (e.g., inundation maps derived from satellite products). Second, finer resolution DEMs could be used, in order to increase the accuracy of the morphological description of the study area (see, e.g., Annis et al., 2020a). Third, to further enhance the input information, soil and climate data (e.g., permeability and precipitation) could be added beside geomorphic descriptors. Finally, more complex machine learning models should be tested, for better characterizing the impact of selecting a given technique on the accuracy of flood hazard assessment.

Chapter 10

Final remarks

The present Thesis describes the application of innovative and efficient data-driven approaches to two distinct, yet closely related, research areas of flood hazard assessment: regional frequency analysis of rainfall extremes, and flood hazard mapping.

In both of these research domains, the existing scientific literature is rich in terms of approaches aiming to solve complex problems through models of reasonable complexity. These approaches benefit from low computational efforts and ease of access to the necessary input information, but are susceptible to significant inconsistencies in certain scenarios. In the research detailed in the preceding Chapters, innovative data-driven approaches are introduced, aiming to relax assumptions and limitations of the previously adopted methods. The new models take advantage of machine learning techniques, which allow to increase the number of input variables relative to the benchmark approaches, and capture non-linear relationships. Importantly, all the models employed rely exclusively on open datasets and are viable for applications to large study areas.

In detail, this Dissertation addresses the following reseach questions: Can we adopt machine learning (ML) techniques to enhance well-established flood hazard models? Can we exploit intrinsic capabilities of ML for extending our study areas while maintaining general validity? Can we rely solely on open-access input information? Can we benefit from the change by transitioning from univariate to multivariate models? The research described in the present Dissertation aims at addressing these very complex and not yet exhaustively answered questions.

In Part 1, the focus is regional frequency analysis (RFA) of sub-hourly rainfall extremes over a large study area in Northern Italy. In the framework of the widely adopted storm index method (Dalrymple, 1960), the benchmark approach considered for the regionalization of the L-moments is the one from Castellarin et al. (2005); Di Baldassarre

et al. (2006). Based on the findings of Schaefer (1990) and Alila (1999), this approach consists of modelling a simple exponentially decaying relation where the L-CS and L-CV (i.e., L-coefficient of skewness and L-coefficient of variation, see Hosking and Wallis, 1997) depend only on the mean annual precipitation (MAP). In other words, the mean cumulative yearly precipitation, a readily available and widely observed variable, is used as a proxy for extreme precipitation. This, in turn, eliminates the need for extensive time series data for precise local estimations of extreme rainfall statistics, which are estimated on a regional basis instead.

Chapter 5 describes an innovative approach for RFA that makes use of ensembles of unsupervised artificial neural networks (ANN). Four models were set up with a different set of input descriptors: the first (MAP-ANN) makes use of the MAP solely, the second (EXT-ANN) relies on an extended set of twenty variables, inclunding distance from the coast, orographic barriers, slope, and elevation (Magnini et al., 2024). The third and fourth build on a pre-processed version of the extended set through principal component analysis (PCA) and canonical correlation analysis (CCA), respectively. The aim of the ANN models is to find a regional relation that links the parameters of a Gumbel distribution for the annual maximum rainfall depths to the input descriptors of climate and morphology for any duration within 1-24 hours.

The validation over a set of 100 gauging stations reveals that the improvement with respect to the benchmark approach is significant, specially when considering longer durations. Moreover, the very nature of the proposed ANN models makes them suitable for interpolating predicted subdaily rainfall quantiles across time-aggregation intervals and space. Indeed, the proposed framework has some limitations. First, a notable challenge lies in the relatively low correlation between predicted and locally fitted quantiles, which is particularly evident for short durations. Second, the number of input descriptors, while beneficial for model accuracy, can complicate the practical application of the models. Third, the utilization of the 2-parameter Gumbel distribution may locally result in underestimation of the highest quantiles. However, despite these inherent limitations, which make necessary further investigation in subsequent studies, the potential presented by machine learning in this context remains significant. Notably, this approach dispenses with the need to assume specific shapes for the relationships between morpho-climatic descriptors and statistical outputs, and it allows for the inclusion of a wide range of input features that may capture any pertinent factors.

Part 2 focuses on simplified methods for flood hazard mapping. These are usually referred to as DEM-based or geomorphic approaches, since they mainly rely on geomorphic

descriptors of the study area retrieved from digital elevation models (DEMs).

In this context, it is common practice in the literature to employ machine learning (ML) algorithms to integrate various data types, which frequently include non-DEM-based features such as geology, precipitation and land use. However, most of the multivariate DEM-based models in the literature are calibrated (or trained) using a reference dataset derived from isolated flood events. On the one hand, this leads to flood susceptibility prediction that is not directly associated with a specific return period, and, on the other hand, to locally satisfactory predictive skills yet with uncertain generalization abilities.

In Chapter 9, an innovative DEM-based approach is proposed and extensively discussed (Magnini et al., 2022). It is based on a blend of seven exclusively DEM-based geomorphic descriptors, that are combined with decision trees to reproduce and extrapolate flood hazard information contained in pre-existing target maps. This approach draws inspiration from various characteristics found in existing literature, carefully selected to create a model whose set-up process is as straightforward as possible.
First, the methods are applied to a large and morpho-climatically complex study are in Northern Italy. During the evaluation of model accuracy, the potential of the models when applied to new areas is tested, showing for the first time in the literature the potential and the limitations of such an approach for a real application to data-scarce areas.

Through a second, nation-wide application to the whole of Italy, we take an additional step by exploring a novel utilization of DEM-based methods (Magnini et al., 2023). Differently from the previous literature, in this case the models are applied to the same study area where training is performed. Then, through a detailed comparison with respect to a set of validation datasets, the ability of DEM-based models to exploit their natural features to enhance flood hazard mapping over the study region is investigated. Final results show the potential of these methods for completing the information of imperfect reference flood hazard maps, and the advantages of continuous representation of hazard over binary flood maps. Thus, new encouraging pathways for data-driven methods for flood hazard mappings are delineated.

Indeed, the aim of the data-driven techniques proposed, described and discussed within this Dissertation is not meant to be an alternative to accurate and well established physically-based models. Instead, the results obtained clearly exemplify how existing ML techniques can be effectively leveraged to process multiple open datasets and improve accuracy for flood hazard assessment and mapping.

It is showed how either complex (i.e., ANNs) and simple (i.e., decision trees) models can improve the performance of benchmark approaches. Multivariate functions of the input descriptors are built for representing flood hazard and rainfall over large and morphoclimatically complex study areas. In general, the results obtained are encouraging, as seen for the longer durations in sub-daily precipitation events and the capability of DEM-based models to extend flood hazard information where target data are not available. However, much research is needed in those areas in which the proposed models show significan margins of improvement, as it is observed for shorter durations in rainfall RFA, and for discontinuous flood hazard characterization from decision trees.

Bibliography

Abrams, M. (2016). Aster global dem version 3, and new aster water body dataset. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLI-B4:107–110.

Acreman, M. and Wiltshire, S. (1989). *The regions are dead. Long live the regions. Methods of identifying and dispensing with regions for flood frequency analysis.* IAHS-AISH Publication 187.

Ahmed, S. and De Marsily, G. (1987). Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. *Water Resources Research*, 23:1717–1737.

Alfieri, L., Salamon, P., Bianchi, A., Neal, J., Bates, P., and Feyen, L. (2014). Advances in pan-european flood hazard mapping. *Hydrological Processes*, 28(13):4067–4077.

Alfieri, L., Salamon, P., Pappenberger, F., Wetterhall, F., and Thielen, J. (2012). Operational early warning systems for water-related hazards in europe. *Environmental Science & Policy*, 21:35–49.

Alila, Y. (1999). A hierarchical approach for the regionalization of precipitation annual maxima in canada. *Journal of Geophysical Research: Atmospheres*, 104(D24):31645–31655.

Allamano, P., Claps, P., Laio, F., and Thea, C. (2009). A data-based assessment of the dependence of short-duration precipitation on elevation. *Physics and Chemistry of the Earth, Parts A/B/C*, 34(10):635–641.

Annis, A., Karpack, M., Morrison, R., and Nardi, F. (2022). On the influence of river basin morphology and climate on hydrogeomorphic floodplain delineations. *Advances in Water Resources*, 159:104078.

Annis, A. and Nardi, F. (2021). Gfplain and multi-source data assimilation modeling: Conceptualization of a flood forecasting framework supported by hydrogeomorphic floodplain rapid mapping. *Hydrology*, 8(4).

Annis, A., Nardi, F., Morrison, R. R., and Castelli, F. (2019). Investigating hydrogeomorphic floodplain mapping performance with varying dtm resolution and stream order. *Hydrological Sciences Journal*, 64(5):525–538.

Annis, A., Nardi, F., Petroselli, A., Apollonio, C., Arcangeletti, E., Tauro, F., Belli, C., Bianconi, R., and Grimaldi, S. (2020a). Uav-dems for small-scale flood hazard mapping. *Water (Switzerland)*, 12(6).

Annis, A., Nardi, F., Volpi, E., and Fiori, A. (2020b). Quantifying the relative impact of hydrological and hydraulic modelling parameterizations on uncertainty of inundation maps. *Hydrological Sciences Journal*, 65(4):507 – 523.

Arabameri, A., Rezaei, K., Cerdà, A., Conoscenti, C., and Kalantari, Z. (2019). A comparison of statistical methods and multi-criteria decision making to map flood hazard susceptibility in northern iran. *Science of The Total Environment*, 660:443–458.

Avand, M., Kuriqi, A., Khazaei, M., and Ghorbanzadeh, O. (2022). Dem resolution effects on machine learning performance for flood probability mapping. *Journal of Hydro-environment Research*, 40:1–16.

Bartholmes, J. C., Thielen, J., Ramos, M. H., and Gentilini, S. (2009). The european flood alert system efas – part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrology and Earth System Sciences*, 13(2):141–153.

Bashfield, A. and Keim, A. (2011). Continent-wide dem creation for the european union. Cited by: 39.

Bates, P. D. and De Roo, A. (2000). A simple raster-based model for flood inundation simulation. *Journal of Hydrology*, 236(1):54–77.

Bates, P. D., Horritt, M. S., and Fewtrell, T. J. (2010). A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. *Journal of Hydrology*, 387(1):33–45.

Bellos, V. and Tsakiris, G. (2016). A hybrid method for flood simulation in small catchments combining hydrodynamic and hydrological techniques. *Journal of Hydrology*, 540:331–339.

Blöschl, G. (2011). 2.19 - scaling and regionalization in hydrology. In Wilderer, P., editor, *Treatise on Water Science*, pages 519–535. Elsevier, Oxford.

Blum, A. G., Archfield, S. A., and Vogel, R. M. (2017). On the probability distribution of daily streamflow in the united states. *Hydrology and Earth System Sciences*, 21(6):3093–3103.

Bonan, G. B., Oleson, K. W., Vertenstein, M., Levis, S., Yongjiu Dai, X. Z., Dickinson, R. E., and Yang, Z.-L. (2002). The land surface climatology of the community land model coupled to the ncar community climate model. *Journal of Climate*, 15(22):3123 – 3149.

Bostan, P., Heuvelink, G., and Akyurek, S. (2012). Comparison of regression and kriging techniques for mapping the average annual precipitation of turkey. *International Journal of Applied Earth Observation and Geoinformation*, 19:115–126.

Braca, G., Bussettini, M., Ducci, D., Lastoria, B., and Mariani, S. (2019). Evaluation of national and regional groundwater resources under climate change scenarios using a gis-based water budget procedure. *Rendiconti Lincei Sci. Fis. E Nat.*, 30:109–123.

Brath, A., Casagli, N., Marani, M., Mercogliano, P., and Motta, R. (2023). Rapporto della commissione tecnico-scientifica istituita con deliberazione della giunta regionale n. 984/2023 e determinazione dirigenziale 14641/2023, al fine di analizzare gli eventi meteorologici estremi del mese di maggio 2023. Technical report, Emilia-Romagna region.

Brath, A., Castellarin, A., and Montanari, A. (2003). Assessing the reliability of regional depth-duration-frequency equations for gaged and ungaged sites. *Water Resources Research*, 39(12).

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification And Regression Trees*. Repr. ed. Chapman & Hall [u.a.], Boca Raton.

Brunetti, M., Maugeri, M., Nanni, T., and Navarra, A. (2002). Droughts and extreme events in regional daily italian precipitation series. *International Journal of Climatology*, 22(5):543–558.

Burlando, P. and Rosso, R. (1996). Scaling and muitiscaling models of depth-duration-frequency curves for storm precipitation. *Journal of Hydrology*, 187(1):45–64. Fractals, scaling and nonlinear variability in hydrology.

Burn, D. H. (1990). An appraisal of the "region of influence" approach to flood frequency analysis. *Hydrological Sciences Journal*, 35(2):149–165.

Caldas-Alvarez, A., Augenstein, M., Ayzel, G., Barfus, K., Cherian, R., Dillenardt, L., Fauer, F., Feldmann, H., Heistermann, M., Karwat, A., Kaspar, F., Kreibich, H., Lucio-Eceiza, E. E., Meredith, E. P., Mohr, S., Niermann, D., Pfahl, S., Ruff, F., Rust, H. W., Schoppa, L., Schwitalla, T., Steidl, S., Thieken, A. H., Tradowsky, J. S., Wulfmeyer, V., and Quaas, J. (2022). Meteorological, impact and climate perspectives of the 29 june 2017 heavy precipitation event in the berlin metropolitan area. *Natural Hazards and Earth System Sciences*, 22(11):3701–3724.

Camorani, G., Castellarin, A., and Brath, A. (2005). Effects of land-use changes on the hydrologic response of reclamation systems. *Physics and Chemistry of the Earth, Parts A/B/C*, 30(8):561–574. Assessment of Anthropogenic Impacts on Water Quality.

Canty, M. (2019). *Image Analysis, Classification and Change Detection in Remote Sensing: With Algorithms for Python*. CRC Press, fourth edition (4th ed.) edition.

Carey-Smith, T., Henderson, R., and Singh, S. (2018). High intensity rainfall design system version 4. Technical report, National Institute of Water and Atmospheric Research Ltd (NIWA).

Castellarin, A., Di Baldassarre, G., Brath, A., and Galeati, G. (2005). Legame sperimentale tra piovosità annuale e regime di frequenza degli estremi di precipitazione. *L'Acqua*, pages 9–22.

Castellarin, A., Merz, R., and Blöschl, G. (2009). Probabilistic envelope curves for extreme rainfall events. *Journal of Hydrology*, 378:263–271.

Claps, P., Ganora, D., and Mazzoglio, P. (2022). Chapter 11 - rainfall regionalization techniques. In Morbidelli, R., editor, *Rainfall*, pages 327–350. Elsevier.

Claps, P. and Laio, F. (2003). Can continuous streamflow data support flood frequency analysis? an alternative to the partial duration series approach. *Water Resources Research*, 39(8).

Cole, G. (1966). An application of the regional analysis of flood flows. *River Flood Hydrology*, page 39–57.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, UK.

Costabile, P., Costanzo, C., and Macchione, F. (2012). Comparative analysis of overland flow models using finite volume schemes. *Journal of Hydroinformatics*, 14:122–135.

Costabile, S. (2010). Geoportale nazionale: il piano straordinario di telerilevamento per l'ambiente. *GEOmedia*, 14(3).

Costache, R., Pham, Q., Avand, M., Thuy Linh, N., Vojtek, M., Vojteková, J., Lee, S., Khoi, D., Thao Nhi, P., and Dung, T. (2020). Novel hybrid models between bivariate statistics, artificial neural networks and boosting algorithms for flood susceptibility assessment. *Journal of Environmental Management*, 265:110485.

CRED and UNDRR (2020). Human cost of disasters. in an overview of the last 20 years: 2000–2019. Technical report, Centre for Research on the Epidemiology of Disaster (CRED), Brussels, Belgium.

Cui, D., Liang, S., Wang, D., and Liu, Z. (2021). A 1 km global dataset of historical (1979–2013) and future (2020–2100) köppen–geiger climate classification and bioclimatic variables. *Earth System Science Data*, 13(11):5087–5114.

Dalrymple, T. (1960). *Flood frequency analysis*. U.S. Geol. Surv. Water Supply Pap.

De Risi, R., Jalayer, F., De Paola, F., and Lindley, S. (2018). Delineation of flooding risk hotspots based on digital elevation model, calculated and historical flooding extents: the case of ouagadougou. *Stochastic Environmental Risk Assessment*, 32:1545–1559.

Degiorgis, M., Gnecco, G., Gorni, S., Roth, G., Sanguineti, M., and Taramasso, A. (2012). Classifiers for the detection of flood-prone areas using remote sensed elevation data. *Journal of Hydrology*, 470-471:302–315.

Deidda, R., Hellies, M., and Langousis, A. (2021). A critical analysis of the shortcomings in spatial frequency analysis of rainfall extremes based on homogeneous regions and a comparison with a hierarchical boundaryless approach. *Stoch Environ Res Risk Assess*, 35:2605–2628.

Deroliya, P., Ghosh, M., Mohanty, M., Ghosh, S., Durga Rao, K., and Karmakar, S. (2022). A novel flood risk mapping approach with machine learning considering geomorphic and socio-economic vulnerability dimensions. *Science of The Total Environment*, 851:158002.

Di Baldassarre, G., Castellarin, A., and Brath, A. (2006). Relationships between statistics of rainfall extremes and mean annual precipitation: an application for design-storm estimation in northern central italy. *Hydrology and Earth System Sciences*, 10(4):589–601.

Di Baldassarre, G., Kooy, M., Kemerink, J. S., and Brandimarte, L. (2013). Towards understanding the dynamic behaviour of floodplains as human-water systems. *Hydrology and Earth System Sciences*, 17(8):3235–3244.

Di Baldassarre, G., Laio, F., and Montanari, A. (2009). Design flood estimation using model selection criteria. *Physics and Chemistry of the Earth, Parts A/B/C*, 34(10):606–611. Recent developments of statistical tools for hydrological application.

Di Prinzio, M., Castellarin, A., and Toth, E. (2011). Data-driven catchment classification: application to the pub problem. *Hydrology and Earth System Sciences*, 15(6):1921–1935.

Dodov, B. and Foufoula-Georgiou, E. (2006). Floodplain morphometry extraction from a high-resolution digital elevation model: a simple algorithm for regional analysis studies. *IEEE Geoscience and Remote Sensing Letters*, 3(3):410–413.

Domeneghetti, A., Carisi, F., Castellarin, A., and Brath, A. (2015). Evolution of flood risk over large areas: Quantitative assessment for the po river. *Journal of Hydrology*, 527:809–823.

Dottori, F., Alfieri, L., Bianchi, A., Skoien, J., and Salamon, P. (2021). A new dataset of river flood hazard maps for europe and the mediterranean basin region. *Earth System Science Data Discussions*, 2021:1–35.

Dottori, F., Alfieri, L., Bianchi, A., Skoien, J., and Salamon, P. (2022). A new dataset of river flood hazard maps for europe and the mediterranean basin. *Earth System Science Data*, 14(4):1549 – 1569.

Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., and Feyen, L. (2016). Development and evaluation of a framework for global flood hazard mapping. *Advances in Water Resources*, 94:87–102.

Everitt, B. S. (2002). The cambridge dictionary of statistics. Cambridge, United Kingdom.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, 45(2).

Favalli, M. and Pareschi, M. T. (2004). Digital elevation model construction from structured topographic data: The dest algorithm. *Journal of Geophysical Research: Earth Surface*, 109(F4).

Florinsky, I., Skrypitsyna, T., and Luschikova, O. (2018). Comparative accuracy of the aw3d30 dsm, aster gdem, and srtm1 dem: A case study on the zaoksky testing ground, central european russia. *Remote Sensing Letters*, 9(7):706–714.

Gallaun, H., Dohr, K., Puhm, M., Stumpf, A., and Hugè, J. (2019). *EU-Hydro - River Net User Guide 1.3*. European Environment Agency.

Garcia G., J. C. (2015). *EU-DEM Upgrade - Documentation EEA User Manual*. Indra Sistemas S.A.

Gesch, D., Oimoen, M., Danielson, J., and Meyer, D. (2016). Validation of the aster global digital elevation model version 3 over the conterminous united states. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B4:143–148.

Ghamariadyan, M. and Imteaz, M. (2021). Prediction of seasonal rainfall with one-year lead time using climate indices: A wavelet neural network scheme. *Water Resour. Manag.*, page 5347–5365.

Gnecco, G., Morisi, R., Roth, G., Sanguineti, M., and Taramasso, A. (2017). Supervised and semi-supervised classifiers for the detection of flood-prone areas. *Soft Computing*, 21:3673–3685.

Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228(1):113–129.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15(5):1049–1054.

Grieser, J., Staeger, T., and Schönwiese, C.-D. (2007). Estimates and uncertainties of return periods of extreme daily precipitation in germany. *Meteorologische Zeitschrift*, 16(5):553–564.

Grimaldi, S., Kao, S.-C., Castellarin, A., Papalexiou, S.-M., Viglione, A., Laio, F., Aksoy, H., and Gedikli, A. (2011). 2.18 - statistical hydrology. In Wilderer, P., editor, *Treatise on Water Science*, pages 479–517. Elsevier, Oxford.

Guha-Sapir, D., Hoyois, P., Wallemacq, P., and Below, R. (2016). Annual disaster statistical review 2016: The numbers and trends. Technical report, CRED, Brussels, Belgium.

Gumbel, E. J. (1954). *Statistical theory of extreme values and some practical applications*. United States Government Printing Office.

Han, J. and Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. In Mira, J. and Sandoval, F., editors, *From Natural to Artificial Neural Computation*, pages 195–201, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hartmann, A., Goldscheider, N., Wagener, T., Lange, J., and Weiler, M. (2014). Karst water resources in a changing world: Review of hydrological modeling approaches. *Reviews of Geophysics*, 52(3):218 – 242.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York, New York, NY.

Hengl, T. (2007). *A Practical Guide to Geostatistical Mapping of Environmental Variables*. Office for Official Publications of the European Communities, Luxembourg (Luxembourg).

Hengl, T., Heuvelink, G., and Stein, A. (2003). Comparison of kriging with external drift and regression-kriging. *Technical Note, ITC*.

Ho, W., Xu, X., and Dey, P. K. (2010). Multi-criteria decision making approaches for supplier evaluation and selection: A literature review. *European Journal of Operational Research*, 202(1):16–24.

Horritt, M. and Bates, P. (2002). Evaluation of 1d and 2d numerical models for predicting river flood inundation. *Journal of Hydrology*, 268(1):87–99.

Hosking, J. and Wallis, J. (1993). Some statistics useful in regional frequency analysis. *Water Resour. Res.*, 29:271–281.

Hosking, J. and Wallis, J. (1997). *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press, Cambridge, UK, 1st ed. edition.

ISPRA (2018). Landslides and floods in italy: Hazard and risk indicators – summary report 2018. Technical report, ISPRA Reports 287/bis/2018.

Jain, A., Nandakumar, K., and Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285.

Janizadeh, S., Avand, M., Jaafari, A., Phong, T. V., Bayat, M., Ahmadisharaf, E., Prakash, I., Pham, B. T., and Lee, S. (2019). Prediction success of machine learning methods for flash flood susceptibility mapping in the tafresh watershed, iran. *Sustainability*, 11(19).

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348):158–171.

Jolliffe, I. (2002). *Principal component analysis*. Springer, New York, 2nd ed. edition.

Jongman, B., Koks, E. E., Husby, T. G., and Ward, P. J. (2014). Increasing flood exposure in the netherlands: implications for risk financing. *Natural Hazards and Earth System Sciences*, 14(5):1245–1255.

Khosravi, K., Pham, B., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., and Tien Bui, D. (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at haraz watershed, northern iran. *Science of The Total Environment*, 627:744–755.

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G., and Kirschbaum, D. B. (2017). So, how much of the earth's surface is covered by rain gauges? *Bulletin of the American Meteorological Society*, 98(1):69 – 78.

Kirkby, M. (1975). *Processes in physical and human geography*, chapter Hydrograph modelling strategies, page 69–90. Heinemann, Oxford.

Koutsoyiannis, D. (2004). Statistics of extremes and estimation of extreme rainfall: I. theoretical investigation / statistiques de valeurs extrêmes et estimation de précipitations extrêmes: I. recherche théorique. *Hydrological Sciences Journal*, 49(4):–590.

Koutsoyiannis, D. (2007). *A Critical Review of Probability of Extreme Rainfall: Principles and Models*, pages 139–166. Taylor and Fraincis, London.

Koutsoyiannis, D. and Baloutsos, G. (2000). Analysis of a long record of annual maximum rainfall in athens, greece, and design rainfall inferences. *Natural Hazards*, 22:29–48.

Koutsoyiannis, D., Kozonis, D., and Manetas, A. (1998). A mathematical framework for studying rainfall intensity-duration-frequency relationships. *Journal of Hydrology*, 206(1):118–135.

Le Gall, P., Favre, A.-C., Naveau, P., and Prieur, C. (2022). Improved regional frequency analysis of rainfall data. *Weather and Climate Extremes*, 36:100456.

Lee, S., Kim, J.-C., Jung, H.-S., Lee, M. J., and Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in seoul metropolitan city, korea. *Geomatics, Natural Hazards and Risk*, 8(2):1185–1203.

Lehner, B. and Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, 27(15):2171–2186.

Libertino, A., Allamano, P., Laio, F., and Claps, P. (2018). Regional-scale analysis of extreme precipitation from short and fragmented records. *Advances in Water Resources*, 112:147–159.

Lindersson, S., Brandimarte, L., Mård, J., and Di Baldassarre, G. (2021). Global riverine flood risk – how do hydrogeomorphic floodplain maps compare to flood hazard maps? *Natural Hazards and Earth System Sciences*, 21(10):2921–2948.

Madsen, H., Pearson, C. P., and Rosbjerg, D. (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 2. regional modeling. *Water Resources Research*, 33(4):759–769.

Magnini, A., Lombardi, M., Bujari, A., Mattivi, P., Shustikova, J., Patella, M., Bitelli, G., Bellavista, P., Lo Conti, F., Tirri, A., Bagli, S., Mazzotti, P., and Castellarin, A. (2023). Geomorphic flood hazard mapping: from floodplain delineation to flood-hazard characterization. *Hydrological Sciences Journal*, Accepted for publication.

Magnini, A., Lombardi, M., Ouarda, T. B. M. J., and Castellarin, A. (2024). Ai-driven morphoclimatic regional frequency modelling of sub-daily rainfall extremes. *Journal of Hydrology*, 631:130808.

Magnini, A., Lombardi, M., Persiano, S., Tirri, A., Lo Conti, F., and Castellarin, A. (2022). Machine-learning blends of geomorphic descriptors: value and limitations for flood hazard assessment across large floodplains. *Natural Hazards and Earth System Sciences*, 22(4):1469–1486.

Maity, R. (2018). *Statistical methods in hydrology and hydroclimatology*. Springer Nature Singapore Pte Ltd.

Malamos, N. and Koutsoyiannis, D. (2016). Bilinear surface smoothing for spatial interpolation with optional incorporation of an explanatory variable. part 2: Application to synthesized and rainfall data. *Hydrological Sciences Journal*, 61:527–540.

Manfreda, S., Di Leo, M., and Sole, A. (2011). Detection of flood-prone areas using digital elevation models. *Journal of Hydrologic Engineering*, 16(10):781–790.

Manfreda, S., Nardi, F., Samela, C., Grimaldi, S., Taramasso, A. C., Roth, G., and Sole, A. (2014). Investigation on the use of geomorphic approaches for the delineation of flood prone areas. *Journal of Hydrology*, 517:863–876.

Manfreda, S. and Samela, C. (2019). A digital elevation model based method for a rapid estimation of flood inundation depth. *Journal of Flood Risk Management*, 12.

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

Manfreda, S., Sole, A., and Fiorentino, M. (2008). Can the basin morphology alone provide an insight into floodplain delineation? In Proverbs, D., Brebbia, C. A., and Penning-Roswell, E., editors, *Flood Recovery, Innovation and Response I*, page 47–56. WITpress, London, England.

Marani, M. (2003). On the correlation structure of continuous and discrete point rainfall. *Water Resources Research*, 39(5).

Marchesini, I., Salvati, P., Rossi, M., Donnini, M., Sterlacchini, S., and Guzzetti, F. (2021). Data-driven flood hazard zonation of italy. *Journal of Environmental Management*, 294:112986.

Marra, F., Armon, M., Borga, M., and Morin, E. (2021). Orographic effect on extreme precipitation statistics peaks at hourly time scales. *Geophysical Research Letters*, 48(5).

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58:1246–1266.

Matheron, G. (1969). *Le krigeage universel. Vol. 1.* Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris, Fontainebleau.

Mattivi, P., Franci, F., Lambertini, A., and Bitelli, G. (2019). Twi computation: a comparison of different open source giss. *Open geospatial data, softw. stand.*, 4(6).

Mazzoglio, P., Butera, I., Alvioli, M., and Claps, P. (2022). The role of morphology in the spatial distribution of short-duration rainfall extremes in italy. *Hydrology and Earth System Sciences*, 26(6):1659–1672.

Mazzoglio, P., Butera, I., and Claps, P. (2020). I2-red: A massive update and quality control of the italian annual extreme rainfall dataset. *Water*, 12(12).

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., and Gupta, H. (2015). Are we unnecessarily constraining the agility of complex process-based models? *Water Resources Research*, 51(1):716 – 728.

Milligan, G. and Cooper, M. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, 5:181–204.

Modarres, R. and Sarhadi, A. (2011). Statistically-based regionalization of rainfall climates of iran. *Global and Planetary Change*, 75(1):67–75.

Monteleone, B., Giusti, R., Magnini, A., Arosio, M., Domeneghetti, A., Borzì, I., Petruccelli, N., Castellarin, A., and Martina, M. (2023). Estimations of crop losses due to flood using multiple sources of information and models: the case study of the panaro river. *Water*, 15.

Mosavi, A., Ozturk, P., and Chau, K.-w. (2018). Flood prediction using machine learning models: Literature review. *Water*, 10(11).

Msilini, A., Masselot, P., and Ouarda, T. B. M. J. (2020). Regional frequency analysis at ungauged sites with multivariate adaptive regression splines. *Journal of Hydrometeorology*, 21(12):2777 – 2792.

Msilini, A., Ouarda, T., and Masselot, P. (2022). Evaluation of additional physiographical variables characterising drainage network systems in regional frequency analysis, a quebec watersheds case-study. *Stoch Environ Res Risk Assess*, 36:331–351.

Mukherjee, S., Joshi, P., Mukherjee, S., Ghosh, A., Garg, R., and Mukhopadhyay, A. (2013). Evaluation of vertical accuracy of open source digital elevation model (dem). *International Journal of Applied Earth Observation and Geoinformation*, 21:205–217.

Nardi, F., Annis, A., Di Baldassarre, G., Vivoni, E. R., and Grimaldi, S. (2019). Gfplain250m, a global high-resolution dataset of earth's floodplains. *Sci Data*, 6.

Nardi, F., Morrison, R. R., Annis, A., and Grantham, T. E. (2018). Hydrologic scaling for hydrogeomorphic floodplain mapping: Insights into human-induced floodplain disconnectivity. *River Research and Applications*, 34(7):675–685.

Nardi, F., Vivoni, E. R., and Grimaldi, S. (2006). Investigating a floodplain scaling relation using a hydrogeomorphic delineation method. *Water Resources Research*, 42(9).

Neal, J., Schumann, G., and Bates, P. (2012). A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas. *Water Resources Research*, 48(11).

Neelz, S. and Pender, G. (2013). *Benchmarking of 2D Hydraulic Modelling Packages*. Environment Agency, Horison House, Deanery Road, Bristol.

Ngongondo, C. S., Xu, C.-Y., Tallaksen, L. M., Alemaw, B., and Chirwa, T. (2011). Regional frequency analysis of rainfall extremes in southern malawi using the index rainfall and l-moments approaches. *Stochastic Environmental Research and Risk Assessment*, 25(7):939 – 955.

Noman, N. S., Nelson, E. J., and Zundel, A. K. (2001). Review of automated floodplain delineation from digital terrain models. *Journal of Water Resources Planning and Management*, 127(6):394–402.

OpenAI (2023). Chatgpt. https://chat.openai.com.

OpenStreetMap contributors (2017). Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

Ouali, D., Chebana, F., and Ouarda, T. (2016). Non-linear canonical correlation analysis in regional frequency analysis. *Stochastic Environmental Research and Risk Assessment*, 30(2):449 – 462.

Ouarda, T. B., Girard, C., Cavadias, G. S., and Bobée, B. (2001). Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology*, 254(1):157–173.

Ouarda, T. B. M. J. and Shu, C. (2009). Regional low-flow frequency analysis using single and ensemble artificial neural networks. *Water Resources Research*, 45(11).

Ouarda, T. B. M. J., Yousef, L. A., and Charron, C. (2019). Non-stationary intensity-duration-frequency curves integrating information concerning teleconnections and climate change. *International Journal of Climatology*, 39(4):2306–2323.

Papalexiou, S. M., AghaKouchak, A., and Foufoula-Georgiou, E. (2018). A diagnostic framework for understanding climatology of tails of hourly precipitation extremes in the united states. *Water Resources Research*, 54(9):6725–6738.

Papalexiou, S. M. and Koutsoyiannis, D. (2013). Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, 49(1):187–201.

Pappenberger, F., Dutra, E., Wetterhall, F., and Cloke, H. L. (2012). Deriving global flood hazard maps of fluvial floods through a physical model cascade. *Hydrology and Earth System Sciences*, 16(11):4143–4156.

Patel, A., Katiyar, S., and Prasad, V. (2016). Performances evaluation of different open source dem using differential global positioning system (dgps). *The Egyptian Journal of Remote Sensing and Space Science*, 19(1):7–16.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Persiano, S., Ferri, E., Antolini, G., Domeneghetti, A., Pavan, V., and Castellarin, A. (2020). Changes in seasonality and magnitude of sub-daily rainfall extremes in emilia-romagna (italy) and potential influence on regional rainfall frequency estimation. *Journal of Hydrology: Regional Studies*, 32:100751.

Petroselli, A. and Grimaldi, S. (2018). Design hydrograph estimation in small and fully ungauged basins: a preliminary assessment of the eba4sub framework. *Journal of Flood Risk Management*, 11:S197 – S210.

Piper, D., Kunz, M., Ehmele, F., Mohr, S., Mühr, B., Kron, A., and Daniell, J. (2016). Exceptional sequence of severe thunderstorms and related flash floods in may and june 2016 in germany – part 1: Meteorological background. *Natural Hazards and Earth System Sciences*, 16(12):2835–2850.

Prudhomme, C. and Reed, D. W. (1999). Mapping extreme rainfall in a mountainous region using geostatistical techniques: a case study in scotland. *International Journal of Climatology*, 19:1337–1356.

Rennó, C. D., Nobre, A. D., Cuartas, L. A., Soares, J. V., Hodnett, M. G., Tomasella, J., and Waterloo, M. J. (2008). Hand, a new terrain descriptor using srtm-dem: Mapping

terra-firme rainforest environments in amazonia. *Remote Sensing of Environment*, 112(9):3469–3481.

Requena, A., Prosdocimi, I., Kjeldsen, T., and Mediero, L. (2017). A bivariate trend analysis to investigate the effect of increasing urbanisation on flood characteristics. *Hydrology Research*, 48(1):802–821.

Requena, A. I., Chebana, F., and Ouarda, T. B. (2018). A functional framework for flow-duration-curve and daily streamflow estimation at ungauged sites. *Advances in Water Resources*, 113:328–340.

Rijks, D., Terres, J., and Vossen, P. (1998). Agrometeorological applications forregional crop monitoring and production assessment.

Salvadore, E., Bronders, J., and Batelaan, O. (2015). Hydrological modelling of urbanized catchments: A review and future directions. *Journal of Hydrology*, 529(P1):62 – 81.

Samela, C., Troy, T. J., and Manfreda, S. (2017). Geomorphic classifiers for flood-prone areas delineation for data-scarce environments. *Advances in Water Resources*, 102:13–28.

Sampson, C. C., Smith, A. M., Bates, P. D., Neal, J. C., Alfieri, L., and Freer, J. E. (2015). A high-resolution global flood hazard model. *Water Resources Research*, 51(9):7358–7381.

Schaefer, M. (1990). Regional analyses of precipitation annual maxima in washington state. *Water Resour. Res.*, 26:119–131.

Seneviratne, S., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., Mcinnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., and Zhang, X. (2012). *Changes in climate extremes and their impacts on the natural physical environment: An overview of the IPCC SREX report*, pages 12566–. Cambridge University Press, Cambridge, UK; New York, NY, USA.

Shehu, B., Willems, W., Stockel, H., Thiele, L.-B., and Haberlandt, U. (2023). Regionalisation of rainfall depth–duration–frequency curves with different data types in germany. *Hydrology and Earth System Sciences*, 27(5):1109–1132.

Shu, C. and Burn, D. H. (2004). Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resources Research*, 40(9).

Shu, C. and Ouarda, T. B. M. J. (2007). Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. *Water Resources Research*, 43(7).

Shustikova, I., Neal, J. C., Domeneghetti, A., Bates, P. D., Vorogushyn, S., and Castellarin, A. (2020). Levee breaching: A new extension to the lisflood-fp model. *Water*, 12(4).

Singh, V. P. (2018). Hydrologic modeling: progress and future directions. *Geoscience Letters*, 5(1).

Soltani, S., Helfi, R., Almasi, P., and Modarres, R. (2017). Regionalization of rainfall intensity-duration-frequency using a simple scaling model. *Water Resour. Manag.*, 31:4253–4273.

Sutton, R. and Barto, A. G. (2018). *Reinforcement Learning, an Introduction*. The MIT Press, second edition edition.

Svensson, C. and Jones, D. (2010). Review of rainfall frequency estimation methods. *Journal of Flood Risk Management*, 3(4):296–313.

Tachikawa, T., Hato, M., Kaku, M., and Iwasaki, A. (2011). Characteristics of aster gdem version 2. In *2011 IEEE International Geoscience and Remote Sensing Symposium*, pages 3657–3660.

Tadono, T., Nagai, H., Ishida, H., Oda, F., Naito, S., Minakawa, K., and Iwamoto, H. (2016). Generation of the 30 m-mesh global digital surface model by alos prism. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B4:157–162.

Takaku, J. and Tadono, T. (2017). Quality updates of 'aw3d' global dsm generated from alos prism. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5666–5669.

Tarboton, D. G., Bras, R. L., and Rodriguez-Iturbe, I. (1991). On the extraction of channel networks from digital elevation data. *Hydrological Processes*, 5(1):81–100.

Tarquini, S., Vinci, S., Favalli, M., Doumaz, F., Fornaciai, A., and Nannipieri, L. (2012). Release of a 10-m-resolution dem for the italian territory: Comparison with global-coverage dems and anaglyph-mode exploration via the web. *Computers and Geosciences*, 38(1):168–170.

Tavares da Costa, R., Manfreda, S., Luzzi, V., Samela, C., Mazzoli, P., Castellarin, A., and Bagli, S. (2019). A web application for hydrogeomorphic flood hazard mapping. *Environmental Modelling and Software*, 118:172–186.

Tavares da Costa, R., Mazzoli, P., and Bagli, S. (2019). Limitations posed by free dems in watershed studies: The case of river tanaro in italy. *Frontiers in Earth Science*, 7.

Tavares da Costa, R., Zanardo, S., Bagli, S., Hilberts, A. G. J., Manfreda, S., Samela, C., and Castellarin, A. (2020). Predictive modeling of envelope flood extents using geomorphic and climatic-hydrologic catchment characteristics. *Water Resources Research*, 56(9):e2019WR026453. e2019WR026453 2019WR026453.

Thomas, J., Joseph, S., Thrivikramji, K., and Arunkumar, K. (2014). Sensitivity of digital elevation models: The scenario from two tropical mountain river basins of the western ghats, india. *Geoscience Frontiers*, 5(6):893–909.

Triantaphyllou, E. (2000). *Multi-Criteria Decision Making Methods: A Comparative Study*. Applied Optimization. Springer, Boston, MA, 1 edition.

Trigg, M. A., Birch, C. E., Neal, J. C., Bates, P. D., Smith, A., Sampson, C. C., Yamazaki, D., Hirabayashi, Y., Pappenberger, F., Dutra, E., Ward, P. J., Winsemius, H. C., Salamon, P., Dottori, F., Rudari, R., Kappes, M. S., Simpson, A. L., Hadzilacos, G., and Fewtrell, T. J. (2016). The credibility challenge for global fluvial flood risk analysis. *Environmental Research Letters*, 11(9):094014.

Trigila, A., Iadanza, C., Lastoria, B., Bussettini, M., and Barbano, A. (2021). Landslides and floods in italy: Hazard and risk indicators – summary report 2021. Technical report, ISPRA Reports 356/2021.

Uboldi, F. and Lussana, C. (2018). Evidence of non-stationarity in a local climatology of rainfall extremes in northern italy. *International Journal of Climatology*, 38(1):506–516.

Van den Besselaar, E. J. M., Klein Tank, A. M. G., and Buishand, T. A. (2013). Trends in european precipitation extremes over 1951–2010. *International Journal of Climatology*, 33(12):2682–2689.

Van Der Knijff, J. M., Younis, J., and De Roo, A. P. J. (2010). Lisflood: a gis-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, 24(2):189–212.

Van Rossum, G. and Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Velázquez, J. A., Anctil, F., Ramos, M. H., and Perrin, C. (2011). Can a multi-model approach improve hydrological ensemble forecasting? a study on 29 french catchments using 16 hydrological model structures. *Advances in Geosciences*, 29:33–42.

Wackernagel, H. (1998). *Multivariate geostatistics: an introduction with applications*. Springer-Verlag, 2nd edition edition.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., and Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527:1130–1141.

Ward, P., Jongman, B., Salamon, P., Simpson, A., Bates, P., De Groeve, T., Muis, S., de Perez, E., Rudari, R., Trigg, M., and Winsemius, H. (2015). Usefulness and limitations of global flood risk models. *Nat. Clim. Change*, 5:712–715.

Williams, W., Jensen, M., Winne, J., and Redmond, R. (2000). An automated technique for delineating and characterizing valley-bottom settings. *Environmental Monitoring and Assessment*, 64:105–114.

Winsemius, H. C., Van Beek, L. P. H., Jongman, B., Ward, P. J., and Bouwman, A. (2013). A framework for global river flood risk assessments. *Hydrology and Earth System Sciences*, 17(5):1871–1892.

Xu, Y. and Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *J. Anal. Test.*, 2:249–262.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., and Bates, P. D. (2017). A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, 44(11):5844–5853.

Yamazaki, D., Kanae, S., Kim, H., and Oki, T. (2011). A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resources Research*, 47(4).

Youden, W. (1950). Index for rating diagnostic tests. *Cancer*, 3:32–35.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology*, 598.

Acknowledgements

This work would not have been possible without the help of my supervisor Prof. Attilio Castellarin. His eternally positive spirit provided me invaluable guidance, no less than his scientific rigor and knowledge. A further thanks goes to Prof. Michele Lombardi for his friendly, yet irreplaceable assistance with programming, and to Prof. Taha B. M. J. Ouarda for allowing the wonderful experience of landing to Quebec and joining INRS.

Thanks to all the members of Costruzioni Idrauliche at DICAM, for being not just colleagues but friends, and for sharing laughter and support, instead of just work and office space.

It is crucial to recognize and appreciate the hard work of researchers, developers, and scientists who strive every day to provide open-source datasets and analysis tools to people worldwide. This is true for various software like Python, QGIS, GRASS, and RStudio and their packages and plugins, and for datasets such as the MERIT DEM, EU-DEM, BIGBANG and EU-HYDRO. The primary aim of science should be the progress of humanity, not profit. I express my sincere gratitude and admiration to all the individuals and institutions that still uphold a non-profit mindset.

I would like to give a special recognition to Bologna, which has been the city of my life so far, even though it isn't where I was born and raised. Throughout its cozy streets and magnificent squares, I have encountered countless people and ideas. I hope that some of them will stay with me forever.

Finally, my sincere gratitude goes to my family and my friends. There's no better luck than having them with me during this dear, crazy life.

May man be eternally a seeker. May he not be content with things, with comforts and with chatter. May his gaze rise high, always asking new questions and seeking new meanings.