Big Data Research 36 (2024) 100452

Contents lists available at ScienceDirect

Big Data Research

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

Machine Learning for Tsunami Waves Forecasting Using Regression Trees

Eugenio Cesario [a,b,*], Salvatore Giampá [b], Enrico Baglione [c,e], Louise Cordrie [c], Jacopo Selva [d,c], Domenico Talia [a,b]

[a] University of Calabria, Italy
[b] DtoK Lab, Italy
[c] Istituto Nazionale di Geofisica e Vulcanologia (INGV), Sezione di Bologna, Bologna, Italy
[d] Dipartimento di Scienze della Terra, dell'Ambiente e delle Risorse, Università degli Studi di Napoli 'Federico II', Naples, Italy
[e] Department of Physics and Astronomy, University of Bologna, via Irnerio 46, 40126 Bologna, Italy

ARTICLE INFO

Keywords:
Tsunami forecasting
Machine learning
Regression tree

ABSTRACT

After a seismic event, tsunami early warning systems (TEWSs) try to accurately forecast the maximum height of incident waves at specific target points in front of the coast, so that early warnings can be launched on locations where the impact of tsunami waves can be destructive to deliver aids in these locations in the immediate post-event management. The uncertainty on the forecast can be quantified with ensembles of alternative scenarios. Similarly, in probabilistic tsunami hazard analysis (PTHA) a large number of simulations is required to cover the natural variability of the source process in each location. To improve the accuracy and computational efficiency of tsunami forecasting methods, scientists have recently started to exploit machine learning techniques to process pre-computed simulation data. However, the approaches proposed in literature, mainly based on neural networks, suffer of high training time and limited model explainability. To overtake these issues, this paper describes a machine learning approach based on regression trees to model and forecast tsunami evolutions. The algorithm takes as input a set of simulations forming an ensemble that describes potential benefit regional impact of tsunami source scenarios in a given source area, and it provides predictive models to forecast the tsunami waves for other potential tsunami sources in the same area. The experimental evaluation, performed on the 2003 M6.8 Zemmouri-Boumerdes earthquake and tsunami simulation data, shows that regression trees achieve high forecasting accuracy. Moreover, they provide domain experts with fully-explainable and interpretable models, which are a valuable support for environmental scientists because they describe underlying rules and patterns behind the models and allow for an explicit inspection of their functioning. This can enable a full and trustable exploration of source uncertainty in tsunami early-warning and urgent computing scenarios, with large ensembles of computationally light tsunami simulations.

1. Introduction

Tsunamis can be devastating events, potentially causing huge environmental destruction, losses of human lives and economic collapses. The vast majority of tsunamis are generated by submarine earthquakes, even though many other potential sources for large tsunamis are possible [1]. Tsunami early warning systems (TEWSs) play a fundamental role in managing the risk connected to tsunamis. In particular, after a seismic event that occurs near or under the sea, TEWSs try to accurately forecast the maximum height of incident waves at specific target points in front of the coast. This information, in fact, is crucial to launch early warnings on locations where the impact of tsunami waves can be dangerous (or even destructive), so it is really important that such systems provide their forecasts with short computation time while maintaining a high prediction accuracy [2–4].

Tsunami propagation is controlled by both source geometry and characteristics and bathymetric heterogeneity, making difficult, if not impossible, the development of simplified empirical equations to avoid explicit modeling. On the other hand, large uncertainty typically exists about the specific source model [5,1,6]. For this reason, both computationally-based long-term probabilistic tsunami hazard analysis (PTHA) and short-term probabilistic tsunami forecasting (PTF) are fundamentally based on the explicit simulation of tsunami generation and propagation for a large ensemble of potential sources, covering the

* Corresponding author at: University of Calabria, Italy.
E-mail address: eugenio.cesario@unical.it (E. Cesario).

https://doi.org/10.1016/j.bdr.2024.100452
Received 27 February 2023; Received in revised form 24 November 2023; Accepted 3 April 2024
Available online 16 April 2024

entire space of natural variability [7]. Therefore, the development of tsunami forecast methods necessarily deals with the problem of finding a balance between computational feasibility of accurate tsunami simulations and the need of exploring uncertainties by means of the simulation of large sets of individual tsunami scenarios. The computational effort is usually managed either by simplifying tsunami simulation or by reducing the exploration of uncertainty, or both [6]. The management of this trade-off is at the core of all tsunami forecast methods [8–11].

The management of uncertainty is particularly critical whenever timely, quick and publicly sensitive decisions have to be taken, as in the case of tsunami warning or tsunami urgent computing [2,12,13,4]. In these cases, the time-to-solution is critical, but also an accurate quantification of uncertainty is fundamental to constrain the decision-making phase, as the definition of precautionary measures is strongly related to the capability of exploring the tails of uncertainty distributions of forecasts [4]. In this context, techniques aimed at improving the efficiency of tsunami simulations may play a fundamental role, as they allow for accurate and quick solutions that may help the exploration of uncertainty [2,14–17].

To develop new accurate tsunami forecasting methods, scientists have recently started to exploit machine learning and big data analysis techniques that process precomputed simulation data, in order to extract tsunami predictive models [18–23]. Such models can provide additional knowledge in the form of insights, patterns, rules, which can be used as surrogate models to support experts' decisions and complete uncertainty explorations [17]. The development of such data-driven models is quickly growing in the last times. Specifically, several artificial intelligence approaches based on neural networks have been recently proposed to this end [19,18,24,20].

However, there are some limitations in the use of neural network approaches to perform inundation forecasts. First, the possibility to train in advance such models is limited, and it is typically based on datasets of precomputed simulations with a rather limited source variability. Especially in complex tectonic contexts like the Mediterranean, the Caribbean or the Indonesian archipelago: such areas are characterized by a significant tsunami hazard mainly generated by highly spatially and geometrically variable seismicity with relatively small magnitudes with respect to large subduction zones. These characteristics and the resulting large source variability [7] represent a specific challenge for tsunami warning or hazard quantification [11,7] but also for developing and testing the neural network models. In this context, the possibility to concretely and timely train and check the accuracy of such models in real time is still to be fully understood, and the development of alternative approach is critical for improving the applicability of such approaches in real operational contexts.

A second issue concerns model explainability, that is, the capability of the model to explain the intuition and reasoning behind its decision, and not only providing the user with the forecasting results. In fact, neural network-based predictive models, as well as other unsupervised artificial intelligence statistical methods like Gaussian process-based emulators, are usually used as black-boxes, and this strongly reduces the capability of scientists to understand how the model produces at its prediction. This is a crucial issue for the domain experts, because self-explainable predictive models can represent an important support for scientists. The potential benefit is twofold. First, and most importantly, it allows confirming that the regressor found patterns that are consistent with the (known) physics of the problem, allowing for an immediate confirmation of the goodness of the regression model, even in the quick context of early-warning applications. Indeed, potentially physically non-explainable rules may be symptoms of statistical overfitting or other implementation problems that may bias the estimations. Second, it may provide an alternative way to highlight unexpected patterns at specific source-target couples. This may contribute to improve their theories and refine their mathematical models.

To this end, machine learning approaches not based on neural network, like for example regression trees, may provide important contributions, since the knowledge models extracted by the machine learning algorithm may be inspected by tsunami experts, who may both check the consistency of the found patterns and rules to understand the features that most control tsunami propagation in the area. At the best of our knowledge, there are no examples of this kind of approaches designed for tsunami forecasting.

This paper describes a machine learning approach based on regression trees to model and forecast tsunami simulations. The input data of the analysis is a set of simulations forming an ensemble that describes potential benefit regional impact of tsunami source scenarios in a given source area, where each instance is described by input parameters describing the geometry/kinematic of faults triggering the earthquake (tectonic region, magnitude, longitude, latitude, etc.), and output values (simulation results) corresponding to the estimated heights of tsunami waves at several target points in front of the coast close. Given such simulation data, the approach aims at training a predictive model that can be applied to any potential tsunami source in the area to forecast the height of waves at all the target locations. This model, being based on a regression tree, is computationally light and fully-readable by domain experts, allowing both to inspect the underlying rules for checking its internal consistency and learn from the selected features the leading source characteristics in the area. The experimental evaluation is performed and tested in Western Mediterranean, simulating potential benefit early-warning and urgent-computing computations that would have been required just after the 2003 M6.8 Zemmouri-Boumerdes earthquake and tsunami.

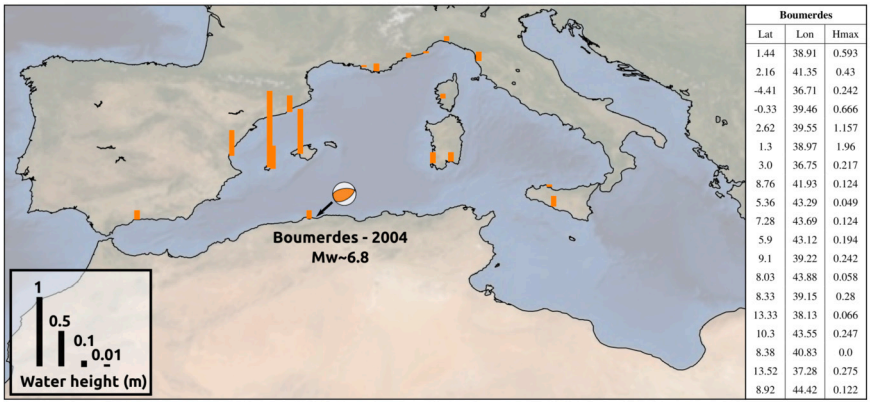
The rest of the paper is organized as follows. Section 2 outlines the problem statement and the goal of this research study. Section 3 presents the proposed approach by describing its main steps in detail. Section 4 provides the experimental evaluation of the approach on simulation data, and a comparative analysis with respect to a baseline. Section 5 discusses our approach in comparison with the most important approaches for tsunami forecasting recently proposed in literature, and the most representative projects in that field of research. Finally, Section 6 concludes the paper and plans future research work.

2. Problem definition and goal

The designed approach aims at defining a fully interpretable and computationally light tsunami simulation tool in a given source area, by training regression trees on a limited dataset of tsunami simulations in the same area. This kind of model may enable a complete and fast exploration of source uncertainty, with potential application both in long-term probabilistic tsunami hazard analysis (PTHA) and in probabilistic tsunami forecasting (PTF) for early warning or urgent computing [11,17,12,4,7,15,5]. The two main advantages of adopting this approach are (*i*) to reduce the number of computationally heavy simulations to be run without limiting the uncertainty exploration, and (*ii*) to provide tsunami predictive models which are fully-explainable and interpretable by domain experts.

To this end, we evaluated the potential of defining this model for a real case study in western Mediterranean, simulating an urgent computing PTF [12,4,17] for the 2003 Mw 6.8 Zemmouri-Boumerdes thrust earthquake and tsunami.

Mw 6.8 earthquake occurred on May 21st, 2003 on the thrust and fold systems that form the Tell Atlas of northern Algeria. This event was the most relevant tsunami event in western Mediterranean in recent times, causing damage at several harbors both in the northern African coast and the Spanish coasts, mainly in the Balearic islands, reaching with significant waves also the French riviera [25–34]. It caused 2,278 casualties on the north African coast [35] and triggered a tsunami causing damage at several harbors in the western side of the Mediterranean basin [25]. It is one of the strongest recent known Mediterranean tsunamis, with observed wave heights of few centimeters to 3 meters. Fig. 1 shows the geographic area involved in this case study, the epicen-



Boumerdes		
Lat	Lon	Hmax
1.44	38.91	0.593
2.16	41.35	0.43
-4.41	36.71	0.242
-0.33	39.46	0.666
2.62	39.55	1.157
1.3	38.97	1.96
3.0	36.75	0.217
8.76	41.93	0.124
5.36	43.29	0.049
7.28	43.69	0.124
5.9	43.12	0.194
9.1	39.22	0.242
8.03	43.88	0.058
8.33	39.15	0.28
13.33	38.13	0.066
10.3	43.55	0.247
8.38	40.83	0.0
13.52	37.28	0.275
8.92	44.42	0.122

Fig. 1. Geographical area of the 2003 Mw 6.8 Zemmouri-Boumerdes case study and wave heights measured in several target locations.

ter of the earthquake and the wave heights measured in several target locations.

For this reason, it is one of the reference events for the tsunami warning centers operating in western Mediterranean [36]. This event has been also taken as reference for the development of the PTF method in [4], in which an ensemble of approximately 15,000 sources is used to describe the uncertainty on the source process in the immediate aftermath of the events (few minutes after origin time). More precisely, the repository collecting the pre-computed scenarios is composed by 15,408 instances [4]. Each scenario represents a different realization of the source parameters (magnitude, hypocenter, fault geometry and kinematics) that correspond to the attributes input to the regressor, as better explained in next section. The Probabilistic Tsunami Forecast associated to an earthquake is based on the results of tsunami simulation of an ensemble of possible earthquake scenarios. This ensemble is built using the short-term information on the earthquake, and has been extracted from the database of simulations of the regional hazard model NEAMTHM18 [11]. These data can be downloaded from TSUMAPS-NEAM documentation website (http://www.tsumaps-neam.eu/documentation/).

Since the tsunami generated by all these sources has been explicitly simulated, we have the opportunity to test the potentiality of our predictive surrogate models. In particular, we randomly selected ensembles of sources with different sizes to train the model, and we quantitatively tested the performance of the predictive model with all the other simulations that represent earthquakes originated in the same area but not included in the training set. The performance has been then evaluated as a function of the size of the training set.

Each scenario simulation is extracted from the NEAMTHM18 propagation database. The tsunami scenarios have been evaluated by combining Gaussian-shaped elementary sea-level elevations that reproduce the scenarios' seafloor deformations [37]. Each elementary source has been modeled with the benchmarked GPU-based nonlinear shallow water Tsunami-HySEA code. The tsunami is propagated over on a regular grid including the whole Mediterranean Sea, using the 30 arc-sec bathymetric model SRTM30+. The modeling of tsunami inundation and run-up is not considered here due to the resolution of the grid and the results are obtained at the 50 meters isobath almost evenly spaced at about 20 kilometers from each other along the coasts of the Mediterranean Sea. At each coastal target point, the maximum tsunami height is estimated from the offshore simulation results with the Green's law. More details can be found in [11,4] and references therein.

3. Input data, output models and the proposed approach

This section provides a proper notation to be used throughout the paper, as well as a definition of the main concepts underlying the proposed approach and the objectives of the analysis.

Input Data. Let D be a dataset collecting simulation data instances, $D = \{d_1, d_2, \ldots, d_N\}$, where each d_i is a tuple representing a tsunami simulation. Specifically, each data tuple is modeled by a $(I + H)$-dimensional attribute vector $A = \langle a_1, \ldots, a_I, a_{I+1}, \ldots, a_{I+H} \rangle$, where the first I attributes describe the input parameters and the remaining H attributes refer to the output results of the simulation.

In the specific case related to tsunami simulation, the dataset of precomputed scenarios adopted in [4] is derived from [11], in which individual scenarios are associated to the following attributes $a_1, \ldots, a_I$:

- *region*: the seismotectonic regions as defined within the NEAMTHM-18 project; the regionalization was built following basic plate tectonics principles and by refining or adapting the regionalization of the European seismic hazard model ([11]).
- *magnitude*: the moment magnitude of the earthquake.
- *longitude*: longitude of the place the earthquake occurs.
- *latitude*: latitude of the place the earthquake occurs.
- *depth of the top*: depth of the upper edge of the fault plane.
- *strike*: angle indicating the orientation in space of the fault defined as the clockwise angle (turning around the normal outgoing from the earth's surface) between the North direction and the positive strike direction.
- *dip*: the dip angle is the angle less than or equal to 90° between the horizontal plane and the fault plane: it gives the direction of the movement on the fault.
- *rake*: the angle that the direction of relative movement (of the hanging-wall with respect to the foot-wall of the fault rupture) forms with the direction of strike, measured counterclockwise from the strike direction.

Many of these attributes are graphically reported in Fig. 2. They describe the geometry and kinematic of faults generating the earthquake (simulation inputs). This input corresponds to all independent input parameters for the precomputed scenarios located of NEAMTHM18 used in this analysis in Western Mediterranean. The other parameters for initializing the tsunami simulations (like fault area, fault length and average slip) were defined deterministically from scaling laws [11] (for the scenarios used here, from [39]). In this area, large magnitudes (e.g. > 8.1) are not included in NEAMTHM18, and the precomputed scenarios do not include non-planar or slip-variable scenarios, which may be instead important for larger magnitude subduction earthquakes in Eastern Mediterranean [11,40].

Further attributes correspond to the expected maximum heights of tsunami waves at the target points in front of the coast (simulation outputs). In particular, the attributes $a_{I+1}, \ldots, a_{I+H}$ correspond to the output fields $hmax_1, \ldots, hmax_H$ of the simulations, where each $hmax_h$ is the maximum height of tsunami waves estimated by the simulator at the h^{th} target point ($h = 1, \ldots, H$).

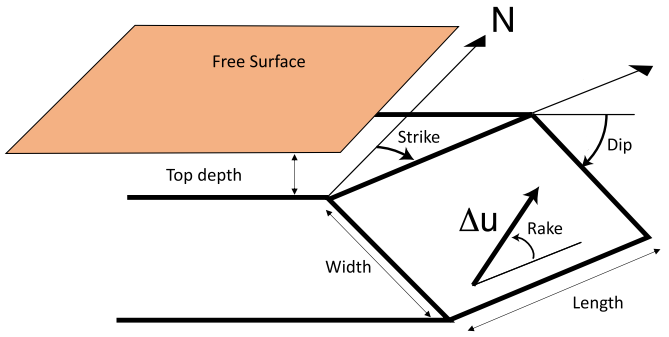


Fig. 2. Simple representation of the fault parametrization, modified from [38]. Under the free surface, coinciding with the sea-bottom for earthquake occurring offshore, the fault orientation is represented: strike, dip and rake angles are reported. The dimensions of the fault plane are reported too. The ΔU vector indicates the slip, which quantifies the amount of displacement between the two faces of the seismic plane (the hanging-wall over the foot-wall).

Output Models. Our goal is to find a regression model for reliably predicting the results of a new simulation (i.e., $hmax_1, \ldots, hmax_H$ values), given an input parameter value setting (i.e., *region, magnitude, ..., average slip*). Formally, we want to extract a set $\mathcal{F}_{hmax}$ of *tsunami predictors*, $\mathcal{F}_{hmax} = \{\mathcal{F}^1_{hmax}, \ldots, \mathcal{F}^H_{hmax}\}$, where each function $\mathcal{F}^h_{hmax} : \mathcal{R}^I \rightarrow \mathcal{R}$, given a specific input parameter setting, forecasts the maximum height of tsunami waves at the h^{th}-target location.

The approach. Now, in order to have a clear view of the whole process, Fig. 3 sketches the general idea of the prediction process through a graphic representation of the designed workflow. The dataset to be analyzed is the set D of collected Tsunami simulation data (represented in the previous described format). The approach computes and returns the set $\mathcal{F}_{hmax}$ of Tsunami predictors. The workflow is composed of three main steps (see Fig. 3), as described in the following. The first step is aimed at splitting the original simulation data in a vertical way, with respect to each specific target location output. In other words, simulation inputs and the h^{th}-output data are gathered in the h^{th}-dataset, for $h = 1, \ldots, H$. At the end of this step, H different datasets are produced, each one containing a vertical projection of D on the h^{th}-target output. The second step consists in the induction of the predictive models, for each target location. In the workflow, this is done by running H regression tree algorithm instances, each one taking in input a dataset built at the previous step. The result consists of H predictive models $\mathcal{F}^1_{hmax}, \ldots, \mathcal{F}^H_{hmax}$, whereas the h^{th}-model represents the Tsunami predictor for the h^{th}-target location. Finally, the third step is aimed at collecting the predictive models extracted at step two. The final result is the whole set $\mathcal{F}_{hmax}$ of Tsunami predictors, which can be used to forecast Tsunami waves at run-time.

Regression Tree learning algorithm. In this work, the Tsunami prediction problem has been modeled as a data-driven task and a *regression tree*-based learning approach [41,42] is exploited to extract the $\mathcal{F}_{hmax}$ predictors. Specifically, we train H regression trees, where the h^{th}-tree is a predictive model to forecast the maximum wave height at the h^{th}-coast location. Now, let us describe more in detail how the regression tree learning algorithm works. A regression tree adopts an axis-parallel hyperplane to iteratively split the attribute data space X into intervals that are as homogeneous as possible with respect to the regressand variable Y. The tree is built according to a top-down induction process, starting from the root, and exploiting a splitting criterion to iteratively determine the splitting attribute $\overline{X}$ and split value $\overline{x}$ that best separate or partition the tuples in D, aimed at reducing the impurity in the data. The sample is split into two subsets D_1 and D_2 (i.e., binary split) by considering all possible partitions $X < \overline{x}$ and $X \geq \overline{x}$, where the split-point $\overline{x}$ is often taken as the midpoint between two adjacent observed values of X [42]. The value $\overline{x}$ is chosen to minimize the collective weighted average of MSE (i.e., mean squared error) of both subsets $MSE(D_1, D_2) = \frac{n1}{n} * MSE(D_1) + \frac{n2}{n} * MSE(D_2)$, which measures how well the points are separated by the split of the dataset D into partitions D_1 and D_2 with $n1$ and $n2$ points. By minimizing $MSE(D_1, D_2)$ the algorithm favors splits into subsets that are homogeneous with respect to the regressed values and heterogeneous with respect to each other subset. Once the first split is chosen, each of the two subsets is split again using the same approach, and the process continues iteratively. This procedure would continue splitting as long as the weighted average of MSE is improved. However, to avoid overfitting several stopping criteria (i.e., minimum number of points in a node, maximum depth of tree, etc.) can be adopted; for example, a common stopping criterion requires that each split must improve the relative MSE error (rMSE) by at least δ_{min}, a predetermined value acting to regularize the cost function of growing the tree by balancing the cost with a penalty for adding additional partitions. Such a criterion is very common and it usually avoids to build overfitting regression trees [42].

An example of regression tree, built by fixing $\delta_{min} = 0.01$, is reported in Fig. 4 (source: [41]); for each split, the absolute MSE and relative (to the first node) rMSE is shown along with δ, i.e. the difference in successive rMSE values. The initial MSE, before to split the root node, is equal to 0.11 (rMSE = 1.0). The first split in the regression tree is done on the boundary value $\overline{x} = 40$, which induces a '$X \geq 40$ and $X > 40$' branching in the data. After the first split, the MSE has been lowered to 0.046 (rMSE = 0.43), resulting in a rMSE reduction equal to $\delta = 0.57$. Then, the second split is done, inducing a '$X < 7$ and $X \geq 7$' branching in the data, and resulting in MSE = 0.025 (further reducing the rMSE of $\delta = 0.19$). Then, the third and fourth splits are done, by reducing the MSE from 0.025 (to 0.016, and then) to 0.010, resulting in rMSE reduction from 0.19 (to 0.09, and then) to 0.05, respectively. Finally, the algorithm evaluates a potential fifth split, which would induce a rMSE improvement equal to 0.007; however, since the tree was built with $\delta_{min} = 0.01$, the split is not done and the growth of the tree terminates.

The prediction of each branch is the average value of the regressand attribute Y within each subset. The regression of a new test instance d_i is done by recursively browsing the extracted regression tree, that is, by evaluating which half-space d_i belongs to, until a leaf node in the regression tree is reached. The regressed value is computed as the average of Y values grouped in the leaf [41].

4. Analysis and experimental results

To evaluate the performance and the effectiveness of the predictive approach described above, we carried out an extensive experimental analysis by executing different tests on the 2003 M 6.8 Zemmouri-Boumerdes earthquake and tsunami, whose details have been described in Section 2. In particular, the experimental evaluation refers to a concrete scenario on which our approach can be applied and the practical usefulness of the system in a real case of urgent computing dealing with a natural disaster.

Specifically, this real-world test benchmark refers to simulations of tsunami waves triggered by the earthquake in specific 1,107 target locations covering the entire Mediterranean area, derived from the dataset of target points of the NEAMTHM18 tsunami hazard model. Fig. 5 shows a zoom to the geographic area involved in this case study: the epicenter of the earthquake is highlighted in blue (on the North Algeria coast, in the Zemmouri-Boumerdes area), and the target locations for the tsunami forecasting are represented as black points (around North Africa and South Europe coasts, which extend toward East also to the Eastern Mediterranean basin). For each location, the goal of the proposed approach is to discover effective predictive models, to be used in case of a seismic event happening in the area, to reliably estimating the maximum height of waves arriving at the coast. Fig. 5 also shows six locations, colored in red, which have been chosen to highlight some specific results to be further discussed in the section. The rest of this section is organized as follows. Section 4.1 describes input data and its gathering, Section 4.2 reports some details about the learning algo-

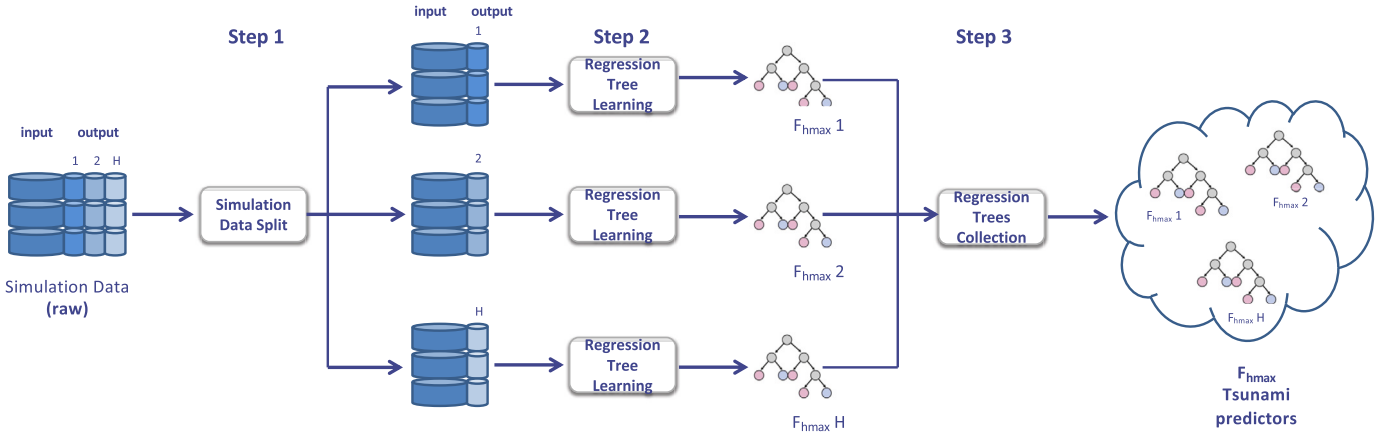
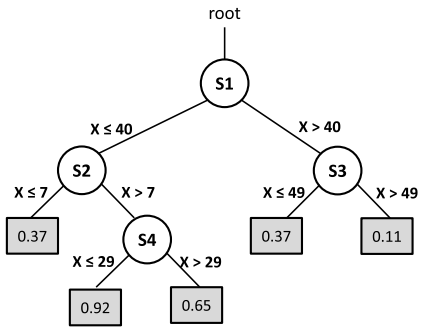


Fig. 3. The workflow including the steps of the analysis process producing the Tsunami predictors.



	MSE	rMSE	δ
Initial state	0.11	1.0	-
split n.1	0.046	0.43	0.57
split n.2	0.025	0.24	0.19
split n.3	0.016	0.15	0.09
split n.4	0.010	0.096	0.05
split n.5 (rejected)	0.009	0.089	0.007

Fig. 4. Regression Tree example [41]. For each split, the absolute MSE and relative MSE are shown, along with the difference in successive rMSE values, δ. The tree was built with cutoff $\delta_{min} = 0.01$, which terminates the growth of the tree.

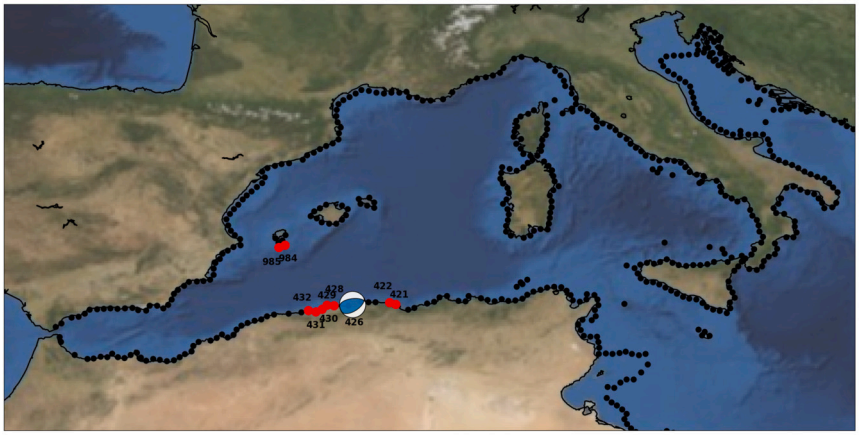


Fig. 5. Geographical representation of the epicenter location (in blue) and the target points (in black), for the 2003 Zemmouri-Boumerdes earthquake case study. Six locations (in red) have been chosen to highlight some specific results to be further discussed in the section.

rithm implementation and models' training, and Section 4.3 describes the evaluation results and a comparison with baselines used for regression analysis.

4.1. Data description

The dataset of precomputed scenarios is composed by 15,408 instances, selected through the PTF workflow as described in Section 2. Such scenarios are subdivided in several homogeneous tectonic regions with adapted discretizations of the events' parameters (*magnitude, position, depth, fault angles, length, area, average slip*). In each tectonic region, all the scenarios corresponding to all the combinations of the parameters values defined within these discretized ranges have been explicitly evaluated. In particular, a probability is attributed to each scenario, which quantifies its consistency with the available information on the studied earthquake and on the region characteristics; it represents the similarity of the scenario to the target event. The ensemble of 15,408 scenarios is produced selecting only the cases associated to a probability larger than a chosen value (2σ). This probability cut-off is defined so that the ensemble is just large enough to produce stable PTF results.

4.2. Training the regressive models

To perform the regression task and its validation, we split the original dataset in two partitions: the training set and the test set. The training set is populated by 10,786 instances (70% of the dataset),

while the test set is composed by 4,622 instances (30% of the dataset). We trained the regression trees from the training set, and we used the trained model to forecast the simulation results on the test set, to assess the quality of the predictions. Furthermore, to have a statistically robust estimation of regression models' performance, we run our tests by implementing the *k-fold cross-validation* methodology, which is a resampling method executing k train-test iterations on different portions of the data. The following section reports more details.

The whole experimental evaluation setting has been developed in Python. In particular, model training has been implemented by the `DecisionTreeRegressor` class of the `scikit-learn` library, which implements a regressor based on decision trees. Specifically, the training algorithm requires some input parameters, whose the most important ones are as follows:

- `max_depth`: the maximum depth of the decision tree. This parameter determines the size of the prediction model. The larger the tree size, the better the approximation of the regressed values. On the other hand, a too large tree results in high computational complexity.
- `criterion`: the function used for measuring the quality of the splits created during the training process; the criterion affects the regression tree model training and has a direct influence on the quality of the results.
- `splitter`: the strategy used to decide which node of the tree must be split; allowed strategies are:
  - `best`: selects the best attribute to split, according to the selected criterion.
  - `random`: selects the best attribute to split from a random subset with the size specified by the max_features parameters. This strategy requires lower computational resources during training.
- `min_samples_split`: the minimum number of samples required to split an existing node. This parameter determines the minimum number of training samples that are aggregated under each tree node. Higher values tend to create short trees that are not very good for regression tasks. Lower values tend to create deeper trees that are better for regression tasks but they require more computational resources. In most cases, by specifying low values, the algorithm tends to split nodes until the maximum depth is reached.

The values of the aforementioned parameters have a direct influence on the quality of the results, as highlighted above. Thus, in order to analyze how they affect predictive model performances, we adopted a parameter sweeping methodology, that is, we run several instances of the learning algorithm by varying their input parameters. Then, we have selected the best result achieved by the model. In particular, we present here the results achieved by fixing `max_depth` = 10, `criterion` = 'squared error', `splitter`='best', and `min_samples_split`=2, which have been assessed through several experimental tests and best suit our application scenario and the considered dataset. Our tests have been carried on a machine hosting four AMD Opteron(TM) 6376 processors (16 cores, 2.3 GHz, 16GB RAM). The average training time of each regression tree amounts to 0.8 secs. The cumulative training time to build the whole set of 1,107 regressors has been measured in about 900 seconds.

4.3. Evaluating the regressive models on the test set

To assess the effectiveness and accuracy of the regressive functions modeled by the regression tree models, we performed an evaluation analysis on the test set by exploiting the models to predict unseen values of the maximum heights of waves for each target location.

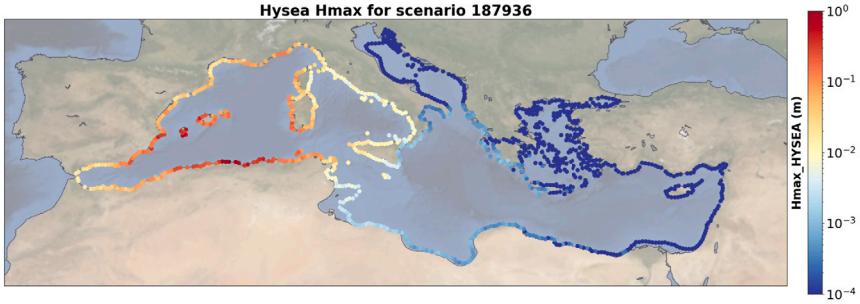
A graphical visualization of the results achieved in our tests is depicted in Fig. 6, which compares the tsunami predictions with the tsunami modeling results for one specific seismic scenario (that is, the best-matching scenario, the one in the ensemble that results closer to the best-guessed source parameters for the seismic event) in the entire geographical area under investigation. Specifically, observed and estimated height of waves is plotted for all target locations, as well as their residual values. We can notice that the observed waves triggered by the earthquake are well estimated by the regression models, and the forecasted values satisfactory reproduce the blue print of the tsunami propagation. In particular, the residual plot shows that the absolute difference between the models is relatively small, with a slight overestimation of the regression model (~ 0.20 cm) only very locally around the earthquake source.

Adopting a different point of view, Fig. 7 shows the estimated vs observed tsunami events (i.e., height of waves), for all scenarios in two specific locations (n. 985 and 922, shown in Fig. 5), where observed and forecasted data are traced in blue and green, respectively. For sake of visualization, the figure shows a partition of events over the whole test set. Considering the faced scenario, the blue line represents the reference (observed) height of waves, while the green line corresponds to the predicted ones. It is evident that the trend forecasted by the regressive model is very similar to that occurring in the observed data. However, the chart shows that low waves are well modeled by the regressive model, while high wave forecasts are affected by an under forecasting bias. For example, by observing Fig. 7(a) we can notice that both real and predicted waves are in general very low; however, the big wave (whose height is 1.49 meters) occurring at the 301^{th} event corresponds to the peak predicted by the regressive model (0.94 meters). Similarly, Fig. 7(b) shows that 1.64 meter high wave really occurred in the 241^{th} event was predicted, by the regressive model, to be 1.21 meter high. From the charts we must notice that, in particular for the big wave cases, the predicted heights may be smaller than the observed data, showing an underforecasting with respect to the real height of waves. However, first the model correctly predicts where big waves are expected (when they occur) by an earthquake, alerting attention on the most destructive cases. On the other hand, the average performance for large tsunamis is rather good, with Mean Average Errors (MAEs) of centimeters for tsunamis in the range 1-2 m, and of tens of centimeters for tsunamis in the range 2-3 m, and negligible Mean Errors (MEs), demonstrating that both over and underestimation occur, and in average the model results not biased.

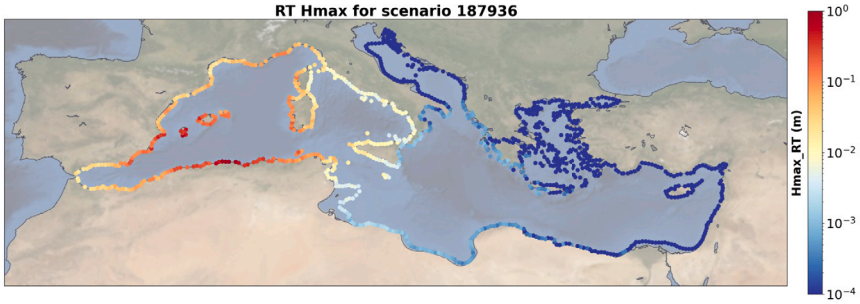
Fig. 8 illustrates the decision tree automatically found by the regressor for target points 432 and 421 (shown in Fig. 5). At higher level, we find the first discriminants for tsunami propagation found by the model. The first parameters found are *latitude, magnitude* and *depth*. While magnitude and depth are known to strongly control tsunami generation, latitude is more peculiar. However, it is a reasonable parameter when we specifically look at the epicentral area, which is characterized by a coastline that develops in the east-west direction at a latitude, in the epicentral area, of approximately 36.8. In this area, latitude is a good discriminant for earthquakes occurring under the sea or under the continent, being epicenters toward the north more tsunamigenic being located under the Mediterranean sea. On the contrary, longitude is found at a much lower priority, demonstrating that it is less relevant for discriminating for the size of the tsunami.

Notably, the most important patterns (the first) correspond to the physics of the problem, demonstrating that the automatic procedure found reasonable patterns to set the models, confirming that there are no mistakes in the implementation nor unwanted overfits, at least at the first order. On the other hand, this establishes a reference trend that can be compared to the regressors in other target points, in order to find potential anomalies. This goes beyond the goals of this paper, but in the future this characteristic can be used to find patterns that may suggest specific corrections to the tsunami modeling scheme and related simplifications.

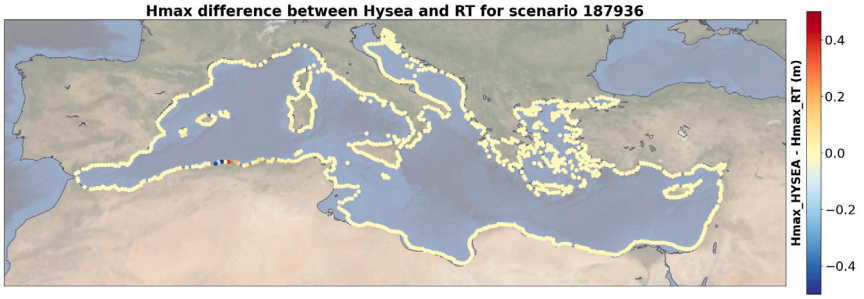
Now, let us give a quantitative evaluation about the accuracy of the regressive model considering the statistics over the entire test set. To do that, we computed several indices, commonly used in the literature to evaluate forecasting accuracy (MSE, MAE, MaxE, EVS, MedAE, ME,



(a) Observe wave heights (Hysea simulations)

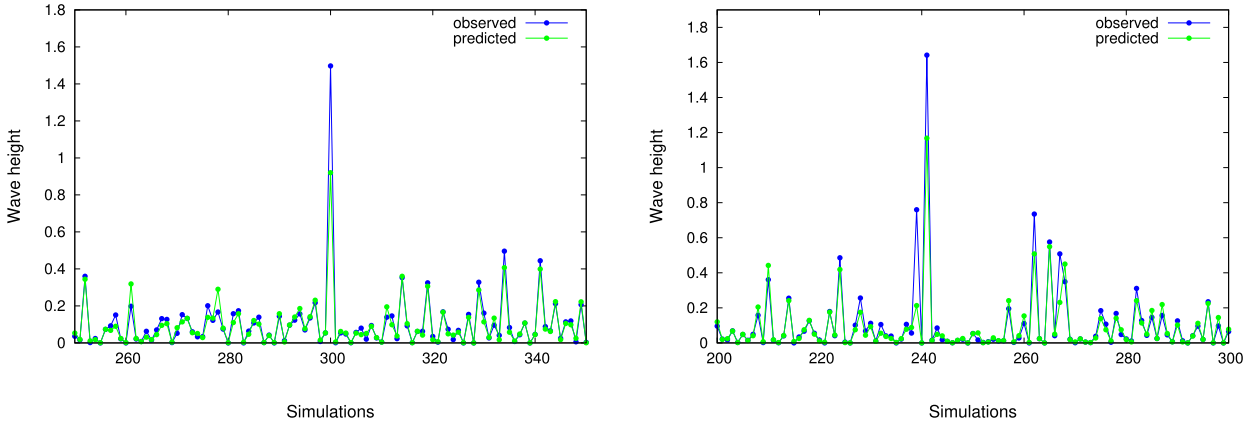


(b) Predicted wave heights (regression tree models)



(c) Residuals between observed and predicted values

Fig. 6. Maximum wave heights at all target locations. For illustration, we report the results for the best-matching scenario (log scale), that is the scenario in the ensemble associated to the largest probability. The chart shows observed, predicted and residual values.



(a) target location n. 985

(b) target location n. 422

Fig. 7. Maximum heights of waves observed and forecasted (blue and green lines) on the test set, for different target locations.

```
|--- lat <= 36.81500053405761718750
|   |--- lat <= 36.59000015258789062500
|   |   |--- magnitude <= 6.93500018119812011719
|   |   |   |--- depth of the top <= 10.83999991416931152344
|   |   |   |   |--- depth of the top <= 4.27999997138977050781
|   |   |   |   |   |--- lon <= 3.56000006198883056641
|   |   |   |   |   |   |--- strike <= 180.00000000000000000000
|   |   |   |   |   |   |   |--- dip <= 20.00000000000000000000
|   |   |   |   |   |   |   |   |--- value: [0.00000000000000000000]
|   |   |   |   |   |   |   |--- dip >  20.00000000000000000000
|   |   |   |   |   |   |   |   |--- strike <= 135.00000000000000000000
|   |   |   |   |   |   |   |   |   |--- strike <= 45.00000000000000000000
|   |   |   |   |   |   |   |   |   |   |--- value: [0.00106631999999999994]
|   |   |   |   |   |   |   |   |   |--- strike >  45.00000000000000000000
|   |   |   |   |   |   |   |   |   |   |--- value: [0.00061411032653061208]
|   |   |   |   |   |   |   |   |--- strike >  135.00000000000000000000
|   |   |   |   |   |   |   |   |   |--- dip <= 80.00000000000000000000
|   |   |   |   |   |   |   |   |   |   |--- value: [0.00165677894736842159]
|   |   |   |   |   |   |   |   |   |--- dip >  80.00000000000000000000
|   |   |   |   |   |   |   |   |   |   |--- value: [0.00071306937500000001]
|   |   |   |   |   |   |--- strike >  180.00000000000000000000
|   |   |   |   |   |   |   |--- strike <= 315.00000000000000000000
|   |   |   |   |   |   |   |   |--- strike <= 225.00000000000000000000
|   |   |   |   |   |   |   |   |   |--- dip <= 40.00000000000000000000
|   |   |   |   |   |   |   |   |   |   |--- value: [0.00044150363636363631]
```

(a) regressor, target location n. 432

```
|--- lat <= 36.81500053405761718750
|   |--- lat <= 36.59000015258789062500
|   |   |--- magnitude <= 6.93500018119812011719
|   |   |   |--- depth of the top <= 10.83999991416931152344
|   |   |   |   |--- lon <= 3.56000006198883056641
|   |   |   |   |   |--- depth of the top <= 4.27999997138977050781
|   |   |   |   |   |   |--- strike <= 180.00000000000000000000
|   |   |   |   |   |   |   |--- dip <= 20.00000000000000000000
|   |   |   |   |   |   |   |   |--- value: [0.00000000000000000000]
|   |   |   |   |   |   |   |--- dip >  20.00000000000000000000
|   |   |   |   |   |   |   |   |--- lon <= 3.28000009059906005859
|   |   |   |   |   |   |   |   |   |--- strike <= 45.00000000000000000000
|   |   |   |   |   |   |   |   |   |   |--- value: [0.00075875283333333334]
|   |   |   |   |   |   |   |   |   |--- strike >  45.00000000000000000000
|   |   |   |   |   |   |   |   |   |   |--- value: [0.00047637723333333339]
|   |   |   |   |   |   |   |   |--- lon >  3.28000009059906005859
|   |   |   |   |   |   |   |   |   |--- rake <= 45.00000000000000000000
|   |   |   |   |   |   |   |   |   |   |--- value: [0.00075801330769230770]
|   |   |   |   |   |   |   |   |   |--- rake >  45.00000000000000000000
|   |   |   |   |   |   |   |   |   |   |--- value: [0.00203325103225806452]
|   |   |   |   |   |   |--- strike >  180.00000000000000000000
|   |   |   |   |   |   |   |--- strike <= 315.00000000000000000000
|   |   |   |   |   |   |   |   |--- strike <= 225.00000000000000000000
|   |   |   |   |   |   |   |   |   |--- dip <= 40.00000000000000000000
|   |   |   |   |   |   |   |   |   |   |--- value: [0.00039618571428571142]
```

(b) regressor, target location n. 421

Fig. 8. Regressor tree examples in text format.

Table 1
Model evaluation metrics.

range (m)	# locs	MSE (m^2)	MaxE (m)	MAE (m)	ME (m)	MeAE (m)	MAPE	AIDA-K
0-0.1	927	8.73E-07	3.56E-03	2.33E-04	-6.17E-07	1.22E-04	7.80E-02	8.30E-01
0.1-0.	67	3.50E-05	5.30E-02	3.42E-03	-1.40E-05	1.85E-03	1.93E-01	9.89E-01
0.2-0.3	38	7.37E-05	8.74E-02	4.86E-03	-1.29E-05	2.56E-03	1.90E-01	9.90E-01
0.3-0.5	38	2.09E-04	1.47E-01	8.04E-03	-1.31E-05	4.17E-03	1.92E-01	9.86E-01
0.5-0.75	19	5.00E-04	2.25E-01	1.26E-02	-7.67E-05	6.69E-03	1.95E-01	9.82E-01
0.75-1	5	8.41E-04	3.50E-01	1.52E-02	-5.66E-05	7.19E-03	2.05E-01	9.75E-01
1-2	9	4.04E-03	6.65E-01	3.09E-02	-1.28E-04	1.36E-02	2.16E-01	9.70E-01
2-3	4	1.36E-02	1.29E+00	5.96E-02	-3.74E-04	2.46E-02	2.34E-01	9.52E-01
0-3	1107	1.07E-04	2.97E-02	1.60E-03	-6.22E-06	7.92E-04	9.70E-02	8.55E-01

Table 2
Baseline evaluation metrics.

range (m)	# locs	MSE (m^2)	MaxE (m)	MAE (m)	ME (m)	MeAE (m)	MAPE	AIDA-K
0-0.1	927	9.79E-06	7.96E-03	9.57E-04	2.66E-07	8.09E-04	5.47E-01	1.46E+02
0.1-0.2	67	3.77E-04	1.20E-01	1.37E-02	-3.23E-05	1.16E-02	4.60E-01	6.29E-01
0.2-0.3	38	8.67E-04	2.09E-01	2.04E-02	2.13E-06	1.72E-02	4.64E-01	6.03E-01
0.3-0.5	38	2.18E-03	3.27E-01	3.20E-02	9.86E-05	2.69E-02	4.66E-01	5.92E-01
0.5-0.75	19	5.19E-03	5.31E-01	4.94E-02	1.53E-04	4.14E-02	4.63E-01	5.95E-01
0.75-1	5	7.21E-03	7.47E-01	5.61E-02	1.37E-04	4.70E-02	4.80E-01	5.57E-01
1-2	9	2.40E-02	1.29E+00	9.90E-02	1.66E-05	8.21E-02	4.85E-01	5.44E-01
2-3	4	6.91E-02	2.31E+00	1.73E-01	-9.02E-04	1.44E-01	5.03E-01	5.01E-01
0-3	1107	7.02E-04	6.36E-02	5.96E-03	1.84E-06	5.01E-03	5.34E-01	1.22E+02

MAPE, AIDA [19]). Also, to make our evaluation more accurate and complete, we performed a comparative analysis of the proposed approach with a baseline approach on the test set. In particular, to have a statistically robust estimation of regression models' performance, we run our tests by implementing the *k-fold cross-validation* methodology ($k = 10$, in our tests), which is a resampling method using different portions of the data to test and train a model on different iterations [42]. Cross-validation is recommended in literature for estimating the performance of a machine learning algorithm due to its relatively low bias and variance, in order to avoid problems like overfitting or selection bias, and to give an insight on how the model will generalize to an independent dataset [42]. Briefly, in the k-fold cross-validation the initial data are randomly partitioned into k mutually exclusive subsets (or folds), $D_1, D_2, \ldots, D_k$, each of approximately equal size. Training and testing is performed k times. At iteration i, for $i = 1, \ldots, k$, partition D_i is reserved as the test set, and the remaining partitions are collectively used as training set to train the model. Finally, all results are combined (i.e., averaged) over the rounds to give an estimate of the overall model's predictive performance.

The values of the error measures are reported in Tables 1 and 2, for the proposed approach and the baseline. In our tests we adopted as baseline a regressor predicting the mean of the outcome variable, which is a ZeroR-based approach [43,44] exploited as basic baseline in regression analysis applications [45]. In particular, we computed a set $\mathcal{B}_{hmax}$ of *baseline predictors*, where each function $\mathcal{B}^h_{hmax}$ is the baseline predictor at the h^{th}-target location, defined as the mean of all observed values at that location. Formally, each $\mathcal{B}^h_{hmax}$ is defined as $\mathcal{B}^h_{hmax} = \frac{1}{N}\sum_{i=1}^{N} y^h_i$, where y^h_i is the i^{th} observation ($i = 1, \ldots, N$) at the h^{th}-target location.

The evaluations have been conducted separately for different flow depths h (i.e. $0\ m < h \leq 0.1\ m$, $0.1\ m < h \leq 0.2\ m$, etc). This is to avoid the introduction of biases due to the fact that many targets (especially in the far field) have small to insignificant tsunami. The purpose of this analysis is to assess the model performance separately for different ranges of maximum observed height, categorized in relation to a practical tsunami early warning system. This allows comparing the performance at similar order of magnitude for the tsunami, without mixing scenarios with practically no tsunami and the ones with significant wave heights. The groups are defined considering the ranges in wave height typically considered in tsunami early warning systems [19]. The groups for higher flow depths ($h > 1$ m) have larger ranges, in order to assure a sufficient number of scenarios for the evaluation of the performance statistics. For completeness, the last row of the table reports also the error metrics computed on the entire test set.

By looking at the values in Table 1 we can make some considerations. First, overall the proposed approach achieves good performance in the tsunami prediction domain. For example, considering the $2\ m < h \leq 3\ m$ range (the most dangerous tsunami scenarios here), test

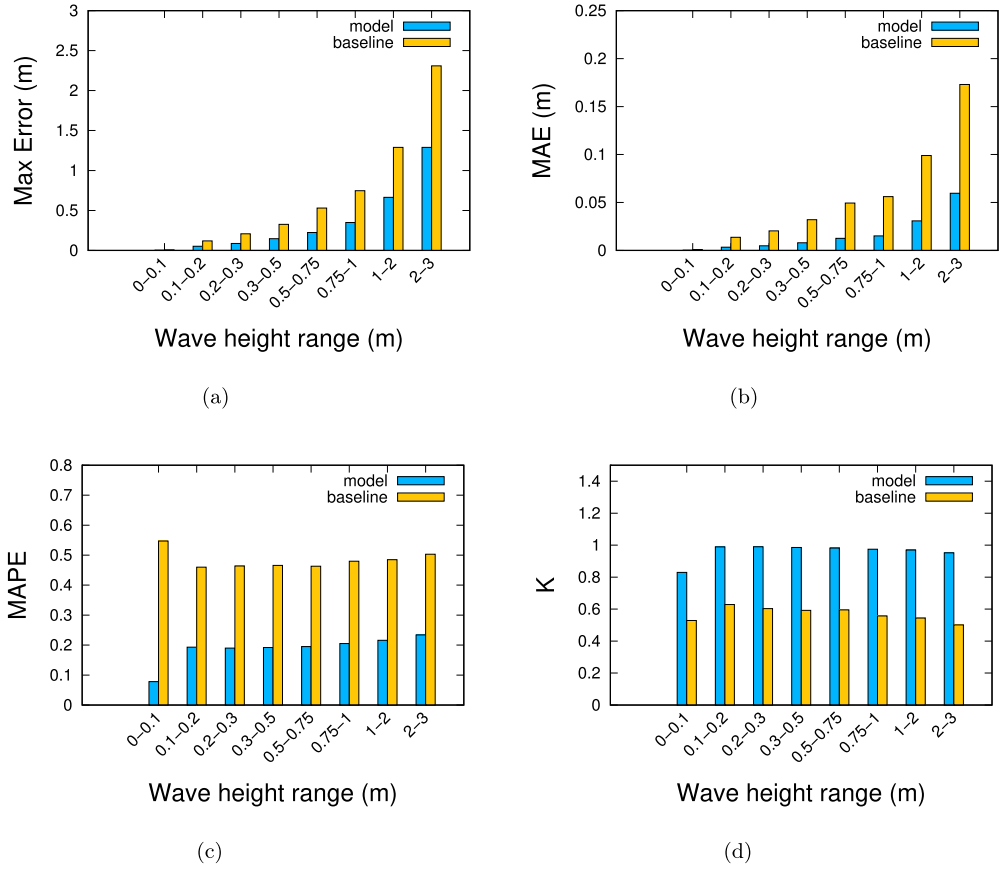


Fig. 9. Comparative performance analysis vs wave height ranges.

results show $MAE = 6.02 * 10^{-2}$ meters and $MAPE = 23.4\%$, which are good values in the considered case. Second, smaller the group's wave magnitude, lower the error measure of the predictive models. For example, considering the $2\ m < h \leq 3\ m$ and $1\ m < h \leq 2\ m$ ranges, the MAE decreases from $6.02 * 10^{-2}$ to $3.08 * 10^{-2}$ meters. A similar trend has been observed also for all other measures and ranges. Finally, also Aida's metric values (specifically used in the tsunami forecasting domain to evaluate the overall coherence of a tsunami model, e.g. see [19]) confirm good performance. In fact, prediction results are generally acceptable when the values of K are within or close to the suggested criteria for satisfactory model performance, which are $0.8 < K < 1.2$ [19]. An underestimation and overestimation of the observations are indicated by the K value larger and smaller than 1, respectively. In our case, regression tree predictions achieve K values very close to 1.0, which is a very remarkable result.

By comparing the values of Tables 1 and 2, we can see that the proposed approach achieves always better predictive performance than the baseline. In fact, considering all error metrics, regression tree models perform better for all of the wave height ranges. This is clearly observed in Fig. 9, which shows some evaluation metrics (i.e., MaxError, MAE, MAPE, AIDA) versus the wave height ranges considered in our tests. As expected, MaxError and MAE (Figs. 9(a) and 9(b)) increase with the wave height magnitude, while MAPE and AIDA's values keep stable values. In all cases, the proposed approach largely achieves better performance than the baseline. Also, as a particular case, Fig. 10 shows the achieved indices, for all locations whose wave depths are within the $1\ m < h \leq 2\ m$ range. We can see that our approach largely achieves a better performance than the baseline, for all the different target locations. These results confirm the appropriateness of the proposed approach based on regression models and its good performance in the tsunami prediction domain.

Fig. 11 shows the distribution of the forecast errors for locations whose wave height is in the $1\ m < h \leq 3\ m$ range (Fig. 11(a)) and all target locations in the test set (Fig. 11(b)), with an overlaid normal curve with mean 0 and the same standard deviation as the distribution of errors. The charts show also the 15^{th}, 50^{th} and 85^{th} percentiles. The plots indicate that, for both cases, forecast errors are normally distributed with the mean centered around zero, suggesting unbiased predictions. As reported in Table 1, the mean error is $-2.04 * 10^{-4}$ m for locations whose wave height is in the $1\ m < h \leq 3\ m$ range, while it is $-2.37 * 10^{-5}$ m for all locations in the test set, which is absolutely consistent with the residual histogram in Fig. 11. Nonetheless, the present results show very good predictive performance considering that the 85% of regression errors are lower than $3.7 * 10^{-2}$ m for locations whose wave height is in the $1\ m < h \leq 3\ m$ range, and $1.4 * 10^{-4}$ m for all locations in the test set, which are highly better results than the existing uncertainty in simulations [37,1].

Now, in addition to the regression analysis and evaluation carried out for different wave height ranges, we analyze here the performance of the regressive models in terms of misclassification rate, that is, the percentage of times observed and forecasted values do not fall in the same bin. This can be especially critical for large wave events, where a classification on a smaller wave height bin (than the real one) can cause an underestimation of the Tsunami event. To do this, we have computed a multi-class confusion matrix, which is a useful tool for analyzing how well a predictive model can recognize instances of different classes. More specifically, given m classes, a confusion matrix CM is a table of size m by m, where an entry $CM[i, j]$ indicates the number of tuples of class i that were labeled by the classifier as class j. For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix, with the rest of the entries being close to zero. Fig. 12 shows the confusion matrix computed through our tests. The total number of predictions is 5,116,554

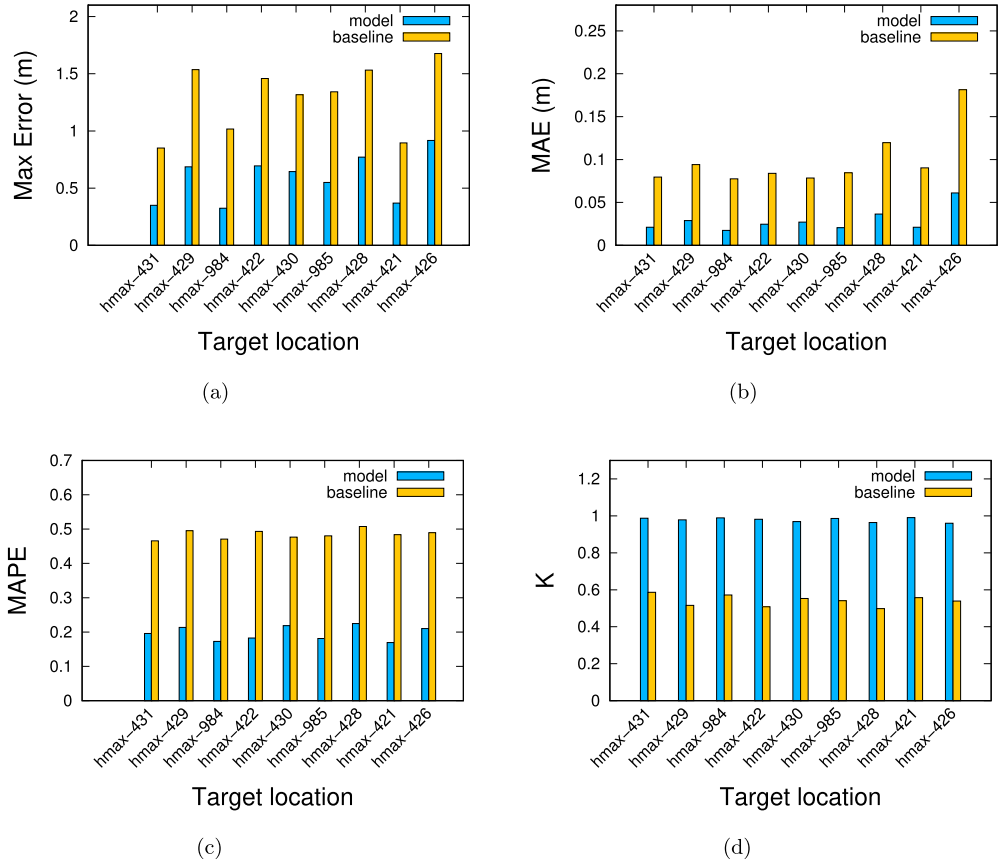
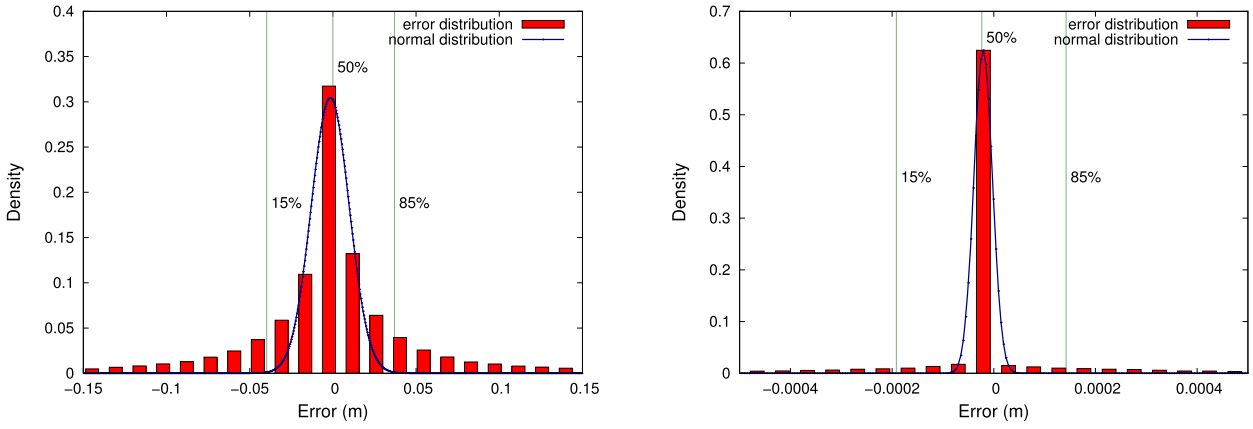


Fig. 10. Performance analysis vs several target locations, whose wave heights are within the 1 $m < h \leq 2$ m range.



(a) Target locations whose wave heights are within the 1 $m < h \leq 3$ m range.

(b) All target locations.

Fig. 11. Distribution of the mean error on the test set, with 15^{th}, 50^{th} and 85^{th} percentiles.

(4,622 test set instances * 1,107 target locations). By observing the matrix, we can notice that the majority of values are on the main diagonal, showing that the most of instances are correctly classified by the predictive models. It is worth nothing that there are also non-zero values occurring in several cells below and above the main diagonal, showing under-forecasting and over-forecasting range classification cases; however, the majority of misclassifications are concentrated on the subdiagonal and superdiagonal, that is lower and upper value ranges. To better clarify this issue, the matrix shows also the percentage distribution of predicted ranges vs actual ranges (values in parentheses). For example, considering the waves in the 0.5 $m < h \leq 0.75$ m height values range, we observe that the 53.1% of cases are predicted to fall in the correct bin, while the 33% and 11% are predicted to fall in the 0.3 $m < h \leq 0.5$ m and 1.0 $m < h \leq 2.0$ m ranges, respectively. Finally, considering the largest (and most dangerous) waves, i.e. 1.0 $m < h \leq 2.0$ m and 2.0 $m < h \leq 3.0$ m ranges, the 57% and 73% of wave height forecasts belong to the correct range, showing good predictive performance in cases of really dangerous tsunami events.

Finally, we discuss here the performance improvement of the regressors with respect to the training-set size used to train them. Fig. 13 shows the variations of MAPE and AIDA as the training-set increases. We can observe that the MAPE strongly decreases for training size up to about 25% of the whole dataset, (i.e., less than 4,000 scenarios out of the 15,408 in the dataset), converging to a stable trend (MAPE <

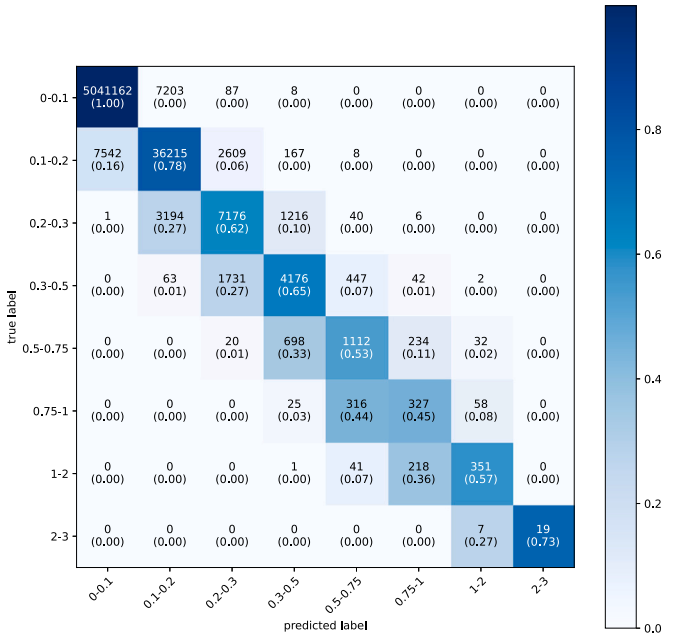


Fig. 12. Confusion matrix for the Tsunami wave predictive model.

10%) for training sizes larger than 25%. Similarly, above 25% also acceptable AIDA's K (> 0.8) are obtained. These metrics variations can be proficiently used to decide when a regressor could be exploited in place of the simulations in a real-time case study like for early warning systems. For example, a first set of regressors can be trained even when a 25% of simulation data are available, rather than waiting for the completion of the whole bunch of simulation data. This could allow to exploit predictive models to launch early warning as quick as possible, which can be an important added value in a tsunami forecasting scenario to reduce damages and save human lives.

5. Related work

In this section we briefly review the most representative research work in the area of tsunami propagation modeling through machine learning and artificial intelligence techniques, and report a critical comparison (on the basis of some specific features) among the method we developed and state-of-art solutions.

A machine learning approach to extract inundation prediction models, based on neural networks, is presented in [19]. In particular, neural network models are trained on a large number of pre-calculated physics-based model results, and they are used immediately after an earthquake to directly estimate the characteristic of (possible) generated tsunamis (i.e., maximum tsunami inundation heights) without running simulation models. The method has been tested on 150 offshore stations encompassing the Japan Trench to simultaneously predict tsunami inundation at seven coastal cities stretching about 100 km along the southern Sanriku coast. The model has been trained using 3,093 hypothetical tsunami scenarios from the megathrust (Mw8.0–9.1) and nearby outer-rise (Mw7.0–8.7) earthquakes, and its predictive accuracy has been tested against 480 unseen scenarios and three near-field historical tsunami events.

Recently, a tsunami forecasting method using a convolutional neural network (CNN) has been proposed in [20]. The method trains a CNN on up-to-date observation data and is exploited to directly forecasting tsunami inundation time series at off-shore locations. Authors highlight that a notable advantage of such a method is that the computational cost of CNN inference is much lower than that of nonlinear tsunami propagation simulations, so it is very feasible for the establishment of early warnings. The experimental evaluation has been carried out on Tohoku data, showing an average maximum tsunami amplitude and tsunami arrival time forecasting errors of 0.4 m and 48 s, respectively, for 1,000 unknown synthetic tsunami scenarios.

An approach based on multi-layer perceptron (MLP) neural networks, exploited to forecast tsunami maximum height and arrival time at given target points, is described in [18]. The experimental evaluation has been done by considering the Horseshoe as seismic source, and Chipiona-Cádiz coast (southwestern Spain) locations as target points, showing a comparative analysis between single and ensemble models.

Liu et al. [24] discuss how three different machine learning approaches can be exploited for forecasting tsunami amplitudes at a set of forecast points, based on hypothetical short-time observations at one or more observation points. Specifically, authors train a support vector machine to predict the maximum amplitude at the target locations, and they also explore the use of two deep convolutional neural networks (i.e., a denoising autoencoder and a variational autoencoder) to predict the full time series at the forecast points. As a case study, models have been trained on synthetic earthquake data, and the experimental evaluation has been performed for an observation point near the entrance of the Strait of Juan de Fuca (Pacific Ocean), and the forecast points in the Salish Sea and in Discovery Bay.

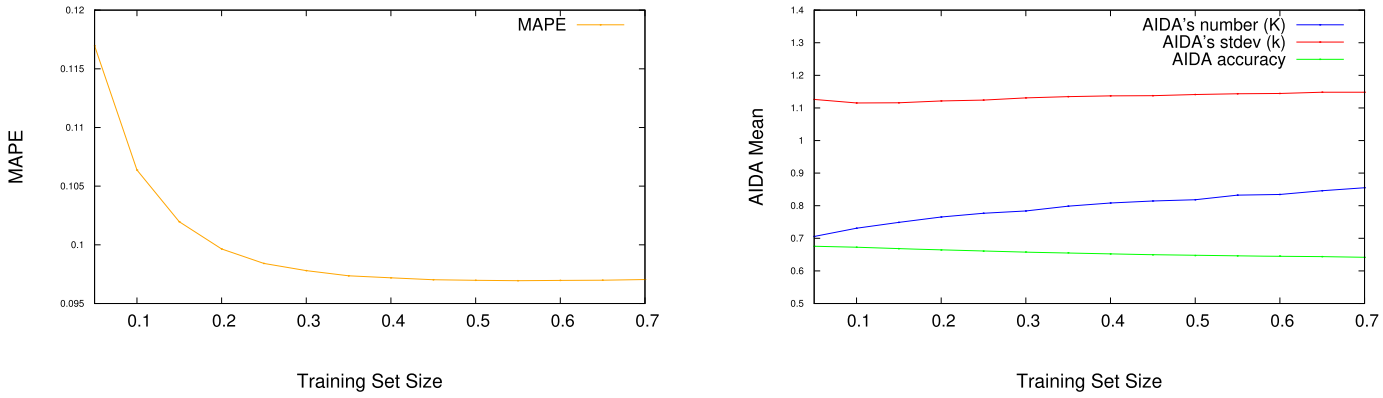
A method to approximate computationally expensive high resolution tsunami simulations with a statistical emulator is discussed in [46]. The paper presents a proof-of-concept case study statistical emulator, based on Gaussian Processes, to estimate the expected tsunami wave elevations and the associated uncertainty in prediction. To estimate Gaussian Processes parameters, authors exploit the Expectation-Maximization algorithm, which is an approach for performing maximum likelihood estimation in the presence of latent variables. It first estimates the values for the latent variables, then optimizes the model and repeats these two steps until convergence. The statistical emulator has been exploited in several research studies, including that one presented in [17], which presents an experimental evaluation carried on the 1945 Makran earthquake considering three areas of interest (Karachi, Chabahar and Muscat). This case study shows that the approach balances the trade-off between computationally expensive simulations and desired accuracy of uncertainties, within given time constraints.

The work presented in [47] aims to assess the capabilities of 1D CNN networks to be used in Tsunami Early Warning. Rather than attempting to estimate a unique network that could map inundation at a large number of points, the focus was set on neural network models (NNM) designed to reproduce time series of tsunami inundation at specific locations. The method has been tested at four specific locations on two bays that differ in their hydrodynamic response. The results showed a good accuracy in predicting on synthetic data not seen by the network before; however, when tested against actual tsunami data, the approach showed lower performance.

Table 3 reports a more detailed and critical comparison among the approach we developed and the above described solutions proposed in the literature. The comparison takes into account several features, as detailed in the following.

Data and application use case. These features differentiate the approaches on the basis of the data and use cases the approaches have been tested on. The proposed approach has been tested on the 2003 Zemmouri-Boumerdes (Algeria) tsunami. The solution proposed in [19] has been tested on the Japan Trench, while that one described in [20] on Tohoku data. The approaches [18] and [17] have been applied on the Horseshoe and Chipiona-Cadiz, and 1945 Makran earthquake, respectively. Finally, the experimental evaluation presented in [24] concerns some target location in the Pacific Ocean.

Forecasting approach. This feature differentiates the algorithms on the basis of the methodology used to perform tsunami forecasting. In particular, our approach exploits regression trees, while those ones presented



(a) MAPE (b) AIDA's number

Fig. 13. MAPE and AIDA's number mean variation as training set size increases.

Table 3
Comparison of several approaches proposed in literature.

Approaches	Data and use case	Forecasting approach	Model Explainability
The proposed approach	2003 Zemmouri-Boumerdes	Regression Tree	Yes
Mulia et al. [19]	Japan Trench	Neural Networks	No or very limited
Makinoshima et al. [20]	Tohoku	Convolutional Neural Networks	No or very limited
Rodriguez et al. [18]	Horseshoe, Chipiona-Cadiz	Multi-layer Perceptron Neural Networks	No or very limited
Sarri et al. [46], Giles et al. [17]	Makran	Gaussian process-based emulator	No or only locally approximated
Liu et al. [24]	Strait of Juan de Fuca, Salish Sea	Support Vector Machines, convolutional neural networks	No or very limited

in [19], [20] and [18] use neural networks, convolutional neural networks and multi-layer perceptron neural networks, respectively. The approach proposed in [24] exploits Support Vector Machine and Convolutional Neural Network approaches. On the other side, the approaches presented in [46] and [17] rely on statistical emulators.

Model explainability. Another important feature we took into account is the model explainability, that is, the capability of the model to explain the intuition and reasoning behind its decision, and not only provide the user with the forecasting result. Considering this aspect, the proposed approach generates regression trees, which are fully-explainable and interpretable models. Differently, the approaches proposed in [19], [20], [18] and [24] exploit neural network models, whose interpretability is very limited because they do not explain how individual neurons work together to arrive at the final output. Finally, the approach proposed in [46] and [17] predicts the highest height of waves at each location through a Gaussian process-based emulator; as stated in [48], such models are necessarily used as black boxes and only local explanations can be done by approximated local linear models, de facto largely limiting the interpretability of the global model.

From the above comparative evaluation, we can summarize the main differences our approach exhibits with respect to the other ones proposed in the literature. The main advantage of the method we designed is its *model explainability*. In fact, the regression tree models allow a user to browse the tree from the root to the leaf (i.e., the regressed value), having also the capability to explain the intuition (splitting attributes, values, etc.) behind its decision. In this way, the tsunami domain expert can visualize the decision path through the tree suffices to understand how and why the model arrives at its prediction. This is undoubtedly an advantage of regression tree models with respect to related models proposed till today. This has two main important implications. On the one side, it allows to make evident potential inconvenient rules based on spurious statistics, like for example overfitting. These features may produce significant biases in the estimations, but they are difficult to find in black-box models. On the contrary, regression trees allow to explicitly inspect the "regularities" that the model exploit, whose physical explainability may be judged by a tsunami expert. On the other side, the detected rules may be used to inform or update simple decision tools like, for example, decision matrices, which are often used to inform tsunami warning for near-source target areas, as it is done by all tsunami service providers in the North East Atlantic, the Mediterranean, and connected seas Tsunami Warning Systems (NEAMTWSs) [49,7]. At the best of our knowledge, this is the first case regression trees are used to perform tsunami forecasting.

6. Conclusion

The increasing availability of tsunami simulation algorithms favors the availability of large amounts of data, whose analysis can produce predictive models to support probabilistic tsunami forecasting (PTF). Such models can be proficiently exploited by tsunami early warning systems (TEWSs) to quickly and accurately forecast dangerous inundation events at the coast, estimate their significant uncertainty, as well as to reduce the computational effort required by probabilistic tsunami hazard analyses (PTHA) to account for the entire natural variability of the potential tsunami sources.

This paper presented a machine learning approach, based on regression trees, to model and forecast tsunami events. The experimental evaluation has been performed on the data relative to the most important recent tsunami occurred in Western Mediterranean, reproducing the potential early-warning and urgent-computing computations that would have been required just after the 2003 M6.8 Zemmouri-Boumerdes earthquake and tsunami. The achieved prediction accuracy ranges from 92% (for wave heights ≤ 0.1 m) to 77% (for wave heights ≥ 1.0 m), showing the appropriateness of the proposed approach based on regression trees and its good performance in the tsunami prediction domain. This approach has also the clear advantage of model explainability, that enables both the potential verification of the physical consistency of the rules adopted by the machine-learning model, and the production of rules that may enhance the comprehension of the tsunamis for the selected target area.

In future work, other research issues will be investigated. For example, we will further explore the extendability of the proposed method to

other areas characterized by large magnitude subduction earthquakes, or the application of spatial analysis approaches for the detection of clusters of target locations having similar tsunami trends. In addition, a parallel implementation of the approach will be integrated in a cloud platform, to take advantage from Cloud computing to reduce execution time and improve speed-up and scale-up.

CRediT authorship contribution statement

Eugenio Cesario: Conceptualization, Supervision, Writing – original draft, Writing – review & editing, Methodology. **Salvatore Giampá:** Software, Validation, Visualization. **Enrico Baglione:** Software, Validation, Visualization. **Louise Cordrie:** Software, Validation, Visualization. **Jacopo Selva:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Domenico Talia:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under Grant Agreement No. 955558. The JU receives support from the European Union's Horizon 2020 research and innovation program, and Spain, Germany, France, Italy, Poland, Switzerland, Norway. We also acknowledge financial support from: PNRR MUR project PE0000013-FAIR. This study was partially supported also by the research project "Assessment of Cascading Events triggered by the Interaction of Natural Hazards and Technological Scenarios involving the release of Hazardous Substances" funded by MIUR – Italian Ministry for Scientific Research under the PRIN 2017 program (grant 2017CEYPS8).

References

[1] A. Grezio, A. Babeyko, M.A. Baptista, J. Behrens, A. Costa, G. Davies, E.L. Geist, S. Glimsdal, F.I. González, J. Griffin, et al., Probabilistic tsunami hazard analysis: multiple sources and global applications, Reviews of Geophysics 55 (4) (2017) 1158–1198.

[2] J. Behrens, A. Androsov, A.Y. Babeyko, S. Harig, F. Klaschka, L. Mentrup, A new multi-sensor approach to simulation assisted tsunami early warning, Natural Hazards and Earth System Sciences 10 (6) (2010) 1085–1100, https://doi.org/10.5194/nhess-10-1085-2010.

[3] E. Bernard, V. Titov, Evolution of tsunami warning systems and products, Philosophical Transactions - Royal Society. Mathematical, Physical and Engineering Sciences 373 (2053) (2015) 20140371.

[4] J. Selva, S. Lorito, M. Volpe, F. Romano, R. Tonini, P. Perfetti, F. Bernardi, M. Taroni, A. Scala, A. Babeyko, F. Løvholt, S.J. Gibbons, J. Macías, M.J. Castro, J.M. González-Vida, C. Sánchez-Linares, H.B. Bayraktar, R. Basili, F.E. Maesano, M.M. Tiberti, F. Mele, A. Piatanesi, A. Amato, Probabilistic tsunami forecasting for early warning, Nature Communications 12 (1) (2021) 56–77.

[5] E.L. Geist, T. Parsons, Probabilistic analysis of tsunami hazards, Natural Hazards 37 (2006) 277–314.

[6] J. Behrens, F. Løvholt, F. Jalayer, S. Lorito, M.A. Salgado-Gálvez, M. Sørensen, S. Abadie, I. Aguirre-Ayerbe, I. Aniel-Quiroga, A. Babeyko, et al., Probabilistic tsunami hazard and risk analysis: a review of research gaps, Frontiers in Earth Science 9 (2021) 628772.

[7] J. Selva, A. Amato, A. Armigliato, R. Basili, F. Bernardi, B. Brizuela, M. Cerminara, M. de'Micheli Vitturi, D. Di Bucci, P. Di Manna, et al., Tsunami risk management for crustal earthquakes and non-seismic sources in Italy, Rivista Del Nuovo Cimento 44 (2) (2021) 69–144.

[8] S. Lorito, J. Selva, R. Basili, F. Romano, M. Tiberti, A. Piatanesi, Probabilistic hazard for seismically induced tsunamis: accuracy and feasibility of inundation maps, Geophysical Journal International 200 (1) (2015) 574–588.

[9] J. Selva, R. Tonini, I. Molinari, M.M. Tiberti, F. Romano, A. Grezio, D. Melini, A. Piatanesi, R. Basili, S. Lorito, Quantification of source uncertainties in seismic probabilistic tsunami hazard analysis (SPTHA), Geophysical Journal International 205 (3) (2016) 1780–1803.

[10] G. Davies, R. Weber, K. Wilson, P. Cummins, From offshore to onshore probabilistic tsunami hazard assessment via efficient Monte Carlo sampling, Geophysical Journal International 230 (3) (2022) 1630–1651.

[11] R. Basili, B. Brizuela, A. Herrero, S. Iqbal, S. Lorito, F.E. Maesano, S. Murphy, P. Perfetti, F. Romano, A. Scala, J. Selva, M. Taroni, M.M. Tiberti, H.K. Thio, R. Tonini, M. Volpe, S. Glimsdal, C.B. Harbitz, F. Løvholt, M.A. Baptista, F. Carrilho, L.M. Matias, R. Omira, A. Babeyko, A. Hoechner, M. Gürbüz, O. Pekcan, A. Yalçıner, M. Canals, G. Lastras, A. Agalos, G. Papadopoulos, I. Triantafyllou, S. Benchekroun, H. Agrebi Jaouadi, S. Ben Abdallah, A. Bouallegue, H. Hamdi, F. Oueslati, A. Amato, A. Armigliato, J. Behrens, G. Davies, D. Di Bucci, M. Dolce, E. Geist, J.M. Gonzalez Vida, M. González, J. Macías Sánchez, C. Meletti, C. Ozer Sozdinler, M. Pagani, T. Parsons, J. Polet, W. Power, M. Sørensen, A. Zaytsev, The making of the NEAM tsunami hazard model 2018 (NEAMTHM18), Frontiers in Earth Science 8 (1) (2021) 56–77.

[12] F. Løvholt, S. Lorito, J. Macias, M. Volpe, J. Selva, S. Gibbons, Urgent tsunami computing, in: 2019 IEEE/ACM HPC for Urgent Decision Making (UrgentHPC), 2019, pp. 45–50.

[13] P.A. Catalan, A. Gubler, J. Cañas, C. Zuñiga, C. Zelaya, L. Pizarro, C. Valdes, R. Mena, E. Toledo, R. Cienfuegos, Design and operational implementation of the integrated tsunami forecast and warning system in Chile (SIPAT), Coastal Engineering Journal 62 (3) (2020) 373–388, https://doi.org/10.1080/21664250.2020.1727402.

[14] L. Blaser, M. Ohrnberger, C. Riggelsen, A. Babeyko, F. Scherbaum, Bayesian networks for tsunami early warning, Geophysical Journal International 185 (3) (2011) 1431–1443.

[15] L. Blaser, M. Ohrnberger, F. Krüger, F. Scherbaum, Probabilistic tsunami threat assessment of 10 recent earthquakes offshore Sumatra, Geophysical Journal International 188 (3) (2012) 1273–1284.

[16] D. Tatsumi, C.A. Calder, T. Tomita, Bayesian near-field tsunami forecasting with uncertainty estimates, Journal of Geophysical Research: Oceans 119 (4) (2014) 2201–2211.

[17] D. Giles, D. Gopinathan, S. Guillas, F. Dias, Faster than real time tsunami warning with associated hazard uncertainties, Frontiers in Earth Science 8 (2021) 597865.

[18] J.F. Rodríguez, J. Macías, M.J. Castro, M. de la Asunción, C. Sánchez-Linares, Use of neural networks for tsunami maximum height and arrival time predictions, GeoHazards 3 (2) (2022) 323–344.

[19] Iyan E. Mulia, Naonori Ueda, Takemasa Miyoshi, Aditya Riadi Gusman, Kenji Satake, Machine learning-based tsunami inundation prediction derived from offshore observations, Nature Communications 13, 5489.

[20] F. Makinoshima, Y. Oishi, T. Yamazaki, Early forecasting of tsunami inundation from tsunami and geodetic observation data with convolutional neural networks, Nature Communications 12, 2253.

[21] D. Salmanidou, S. Guillas, A. Georgiopoulou, F. Dias, Statistical emulation of landslide-induced tsunamis at the Rockall Bank, NE Atlantic, Proceedings of the Royal Society A 473 (2017) 20170026, https://doi.org/10.1098/rspa.2017.0026.

[22] A. Sarri, S. Guillas, F. Dias, Statistical emulation of a tsunami model for sensitivity analysis and uncertainty quantification, Natural Hazards and Earth System Sciences 12 (6) (2012) 2003–2018, https://doi.org/10.5194/nhess-12-2003-2012.

[23] E. Cesario, D. Talia, Distributed data mining patterns and services: an architecture and experiments, Concurrency and Computation 24 (15) (2012) 1751–1774.

[24] C.M. Liu, D. Rim, R. Baraldi, J. Randall, Comparison of machine learning approaches for tsunami forecasting from sparse observations, Pure and Applied Geophysics 178 (2021) 5129–5153.

[25] M. Meghraoui, S. Maouche, B. Chemaa, Z. Cakir, A. Aoudia, A. Harbi, P.-J. Alasset, A. Ayadi, Y. Bouhadad, F. Benhamouda, Coastal uplift and thrust faulting associated with the Mw = 6.8 Zemmouri (Algeria) earthquake of 21 May, 2003, Geophysical Research Letters 31 (19) (2004), https://doi.org/10.1029/2004GL020466.

[26] B. Delouis, M. Vallée, M. Meghraoui, E. Calais, S. Maouche, K. Lammali, A. Mahsas, P. Briole, F. Benhamouda, K. Yelles, Slip distribution of the 2003 Boumerdes-Zemmouri earthquake, Algeria, from teleseismic, GPS, and coastal uplift data, Geophysical Research Letters 31 (18) (2004).

[27] J. Braunmiller, F. Bernardi, The 2003 Boumerdes, Algeria earthquake: Regional moment tensor analysis, Geophysical Research Letters 32 (6) (2005).

[28] F. Semmane, M. Campillo, F. Cotton, Fault location and source process of the Boumerdes, Algeria, earthquake inferred from geodetic and strong motion data, Geophysical Research Letters 32 (1) (2005).

[29] P.-J. Alasset, H. Hébert, S. Maouche, V. Calbini, M. Meghraoui, The tsunami induced by the 2003 Zemmouri earthquake (Mw = 6.9, Algeria): modelling and results, Geophysical Journal International 166 (1) (2006) 213–226.

[30] S. Belabbès, C. Wicks, Z. Çakir, M. Meghraoui, Rupture parameters of the 2003 Zemmouri (Mw 6.8), Algeria, earthquake from joint inversion of interferometric synthetic aperture radar, coastal uplift, and GPS, Journal of Geophysical Research: Solid Earth 114 (B3) (2009).

[31] A. Sahal, J. Roger, S. Allgeyer, B. Lemaire, H. Hébert, F. Schindelé, F. Lavigne, The tsunami triggered by the 21 May 2003 Boumerdes-Zemmouri (Algeria) earthquake:

field investigations on the French Mediterranean coast and tsunami modelling, Natural Hazards and Earth System Sciences 9 (6) (2009) 1823–1834.
[32] M. Heidarzadeh, K. Satake, The 21 May 2003 tsunami in the Western Mediterranean Sea: statistical and wavelet analyses, Pure and Applied Geophysics 170 (2013) 1449–1462.
[33] R. Santos, B. Caldeira, M. Bezzeghoud, J.F. Borges, The rupture process and location of the 2003 Zemmouri–Boumerdes earthquake (Mw 6.8) inferred from seismic and geodetic data, Pure and Applied Geophysics 172 (2015) 2421–2434.
[34] M. Heidarzadeh, Y. Wang, K. Satake, I.E. Mulia, Potential deployment of offshore bottom pressure gauges and adoption of data assimilation for tsunami warning system in the Western Mediterranean Sea, Geoscience Letters 6 (1) (2019) 1–12.
[35] M. Masina, R. Archetti, A. Lamberti, Boumerdès earthquake: numerical investigations of the rupture mechanism effects on the induced tsunami and its impact in harbors, Journal of Marine Science and Engineering 8 (11) (21 may 2003), https://doi.org/10.3390/jmse8110933.
[36] F. Schindelé, A. Gailler, H. Hébert, A. Loevenbruck, E. Gutierrez, A. Monnier, P. Roudil, D. Reymond, L. Rivera, Implementation and challenges of the tsunami warning system in the western Mediterranean, Pure and Applied Geophysics 172 (2015) 821–833.
[37] I. Molinari, R. Tonini, S. Lorito, A. Piatanesi, F. Romano, D. Melini, A. Hoechner, J.M. Gonzàlez Vida, J. Maciás, M.J. Castro, et al., Fast evaluation of tsunami scenarios: uncertainty assessment for a Mediterranean Sea database, Natural Hazards and Earth System Sciences 16 (12) (2016) 2593–2602.
[38] M. Bonafede, Lecture notes, https://www.unibo.it/en/teaching/course-unit-catalogue/course-unit/2017/412275.
[39] M. Leonard, Self-consistent earthquake fault-scaling relations: update and extension to stable continental strike-slip faults, Bulletin of the Seismological Society of America 104 (6) (2014) 2953–2965.
[40] G. Davies, Tsunami variability from uncalibrated stochastic earthquake models: tests against deep ocean observations 2006–2016, Geophysical Journal International 218 (3) (2019) 1939–1960, https://doi.org/10.1093/gji/ggz260.
[41] M.K.N. Altman, Classification and regression trees, Nature Methods 14 (2017) 757–758.
[42] M.K.J. Han, J. Pei, Data Mining Concepts and Techniques, Morgan Kaufmann, 2012.
[43] https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/ZeroR.html.
[44] C. Nasa, S. Suman, Evaluation of different classification techniques for web data, International Journal of Computer Applications 52 (9) (2012) 34–40.
[45] J. Brownlee, Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-to-End, Machine Learning Mastery, 2021.
[46] A. Sarri, S. Guillas, F. Dias, Statistical emulation of a tsunami model for sensitivity analysis and uncertainty quantification, Natural Hazards and Earth System Sciences 12 (6) (2012) 2003–2018.
[47] J. Nunez, et al., Discriminating the occurrence of inundation in tsunami early warning with one-dimensional convolutional neural networks, Tech. rep., 2022.
[48] Y. Yoshikawa, T. Iwata, Gaussian process regression with interpretable sample-wise feature weights, IEEE Transactions on Neural Networks and Learning Systems (2021) 1–15.
[49] A. Amato, A. Avallone, R. Basili, F. Bernardi, B. Brizuela, L. Graziani, A. Herrero, M.C. Lorenzino, S. Lorito, F.M. Mele, A. Michelini, A. Piatanesi, S. Pintore, F. Romano, J. Selva, S. Stramondo, R. Tonini, M. Volpe, From seismic monitoring to tsunami warning in the Mediterranean Sea, Seismological Research Letters 92 (3) (2021) 1796–1816, https://doi.org/10.1785/0220200437.