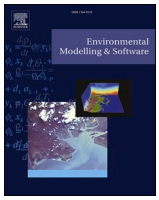
Environmental Modelling and Software 167 (2023) 105781

Contents lists available at ScienceDirect

Environmental Modelling and Software

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



Towards a monitoring system of the sea state based on microseism and machine learning

Vittorio Minio [a,*], Alfio Marco Borzì [a], Susanna Saitta [b], Salvatore Alparone [c], Andrea Cannata [a,c], Giuseppe Ciraolo [d], Danilo Contrafatto [c], Sebastiano D'Amico [e], Giuseppe Di Grazia [c], Graziano Larocca [c], Flavio Cannavò [c]

[a] Dipartimento di Scienze Biologiche, Geologiche ed Ambientali - Sezione di Scienze della Terra, Università degli Studi di Catania, Corso Italia, 57, 95129, Catania, Italy
[b] Dipartimento di Matematica e Informatica, Università di Catania, Viale Andrea Doria, 6, 95100, Catania, Italy
[c] Istituto Nazionale di Geofisica e Vulcanologia - Sezione di Catania, Osservatorio Etneo, Piazza Roma, 2, 95125, Catania, Italy
[d] Dipartimento di Ingegneria Civile, Ambientale, Aerospaziale, dei Materiali, Università di Palermo, Viale delle Scienze, Building 8, 90128, Palermo, Italy
[e] Department of Geosciences, University of Malta, Msida, 2080, Malta

ARTICLE INFO

Handling Editor: Daniel P Ames

Keywords:
Microseism
Significant wave height
Correlation coefficient
Array analysis
Machine learning

ABSTRACT

In this work, we exploited the ubiquitous seismic noise generated by energy transfer from the sea to the solid Earth (called microseism) to infer the significant wave height data, with the aim of developing a microseism-based monitoring system of the Sicily Channel. We used a combined approach based on statistical analysis and machine learning by using seismic and sea state data (provided by the hindcast maps), recorded between 2018 and 2021.Through spectral and amplitude analysis, we observed that microseism was influenced by the conditions of the seas surrounding Sicily. Correlation analysis demonstrates that microseism mostly originates from sources located up to 400 km from the coastlines. Moreover, employing machine learning algorithms, we successfully reconstruct spatial and temporal sea wave distributions using microseism data. Among the tested methods, the Random Forest algorithm yields the best results, with an R^2 value of 0.89 and a mean prediction error of about 0.21 m.

Software and data availability

The codes that were used for the analysis of the seismic and sea state data using python language (version 3.10) based on ObsPy and scikit-learn libraries can be found in Github: https://github.com/VittorioMinio93/shwpredict. This repository was created by Vittorio Minio (Email: vittorio.minio@phd.unict.it) in 2023 and contains program codes (40 KB). Development environment and code testing were as follows:

- OS: Windows 11 Pro 64-bit
- CPU: 2 CPU AMD EPYC 7713 64core 225 W 2.0 GHz
- RAM: 640 GB RAM TrueDDR4 3200MHz (optional memory expansion of up to 8 TB)
- GPU: 2 x GPU A100 40 GB PCIe Gen 40

The seismic data can be downloaded free of charge from European Integrated Data Archive (EIDA; http://www.orfeus-eu.org/data/eida/, last access May 2023). Sea state data can be downloaded by using E.U. Copernicus Marine Service Information (https://doi.org/10.25423/cmcc/medsea_multiyear_wav_006_012, last access May 2023). The earthquake catalogue that was used for the analysis can be downloaded from the United States Geological Survey (USGS; https://earthquake.usgs.gov/fdsnws/event, last access May 2023).

1. Introduction

Microseism is the continuous and omnipresent seismic signal on Earth, and it is generated by the ocean-solid earth interaction (Tanimoto et al., 2015). Considering its spectral content and source mechanism (Haubrich and McCamy, 1969), it can be classified as (*i*) primary microseism (PM), (*ii*) secondary microseism (SM), and (*iii*) short-period secondary microseism (SPSM).

The PM (period 13–20 s) is generated when ocean gravity waves

* Corresponding author. Corso Italia 57, 95129, Catania, Italy.
E-mail address: vittorio.minio@phd.unict.it (V. Minio).

https://doi.org/10.1016/j.envsoft.2023.105781
Received 25 May 2023; Received in revised form 16 July 2023; Accepted 27 July 2023
Available online 29 July 2023
1364-8152/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

reach shallow water in coastal regions and interact with the sloping seafloor (Hasselmann, 1963). In this case, the wave energy can be converted into seismic energy through vertical pressure variations, which have periods similar to the incident ocean gravity waves. The SM (period 5–10 s) could be explained by the superposition of ocean waves of equal periods travelling in opposite directions, generating standing gravity waves of half the period (Longuet-Higgins, 1950). These standing waves cause non-linear perturbations that propagate without attenuation to the ocean bottom where they are converted to seismic energy (Hasselmann, 1963). The SPSM (period 2–5 s) is characterized by sources generally linked to local sea state and wave activity (Bromirski et al., 2005).

At temperate latitudes, microseism amplitudes are characterized by significant annual periodicity with maxima during the autumn-winter seasons and minima spring-summers (Aster et al., 2008). This pattern is different in the Glacial Arctic Sea and the Southern Ocean where, during the winters because of the sea ice, the oceanic waves cannot efficiently excite seismic energy (Aster et al., 2008; Cannata et al., 2019).

Several authors have investigated the empirical relationship between microseism amplitudes and ocean wave height (e.g., Ardhuin et al., 2012) to predict the significant wave height along the coastline. Other authors have developed physical models of the generation of the different types of microseism from the sea state (e.g., Ardhuin et al., 2015).

Considering the availability of seismic and sea wave data in the Sicilian areas and seas, the relationship between microseism and sea waves has been investigated in such areas by some authors. For instance, De Caro et al. (2014) studied the spectral content of microseism recorded in the Ionian and Tyrrhenian Seas. Cannata et al. (2020) and Moschella et al. (2020) explored the microseism recorded along the coastline of Eastern Sicily in terms of spectral features, amplitude seasonal pattern, source location, and relationship with the sea wave activity by machine learning-based algorithms able to provide significant wave height information from microseism recordings.

The monitoring of the sea state is a fundamental task for economic activities in coastal areas, such as transportation, tourism, and the design of infrastructures (e.g., Von Storch et al., 2015; Ferretti et al., 2018). In recent years, the routine monitoring of wave height for marine risk assessment and mitigation has become compelling due to global warming that is making sea waves stronger and, hence, the extreme wave events more intense and frequent (Reguero et al., 2019). In particular, the Mediterranean Sea has been considered one of the most responsive regions to global warming, which could favor the intensification of the Mediterranean Tropical-Like Cyclones, also called Medicanes, characterized by intense winds, heavy precipitation, and high sea waves (e.g., Ivan et al., 2018; Portmann et al., 2020; Lagouvardos et al., 2022). The last feature, combined with the fact that Medicanes occur in a closed basin with a high-density population along the coastlines, can potentially lead to severe socio-economic consequences. To reduce the impacts of these events, the development of more advanced monitoring systems of the sea state becomes necessary. For the first time, the relationship between microseism and Medicanes was taken into account by Borzì et al. (2022) who investigated the relationship between SM and SPSM and the Medicane Apollo, that occurred in the late-October 2021 in the Ionian Sea, in a way to locate the microseism source, by using array techniques and a grid search method based on the seismic amplitude decay. In this context, machine learning (ML) techniques can play an important role in improving the existing monitoring systems. Indeed, ML techniques are designed to extract information directly from data using well defined optimization rules and help unravel hidden relationships between distinct parameters, as well as to build predictive models (e.g., Kuhn and Johnson, 2013; Kong et al., 2018; Mayfield et al., 2020; Chen et al., 2021).

In this work, we will show the results of analyses performed to acquire information necessary to develop a monitoring system of the sea state, in terms of significant wave height, in the Sicily Channel based on microseism. In particular, we will present a comprehensive analysis of this seismic signal, including spectral, amplitude, correlation with significant wave height, and array analysis, aimed to unravel the unique features of the microseism and explore its relationship with the state of the surrounding seas. By establishing a solid understanding of these fundamental aspects, we lay the groundwork for developing a robust microseism-based predictive model of the sea state by using ML techniques and interpreting the final results.

2. Data and methods

2.1. Seismic data

To study microseism, we utilized seismic signals recorded during 2018–2021 by the three components of 14 stations belonging to permanent seismic networks managed by the Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Etneo (INGV-OE) and the Department of Geosciences at the University of Malta (see Fig. 1a and b). These stations are equipped with broadband three-component seismometers that record at a sampling rate of 100 Hz (refer to Table A1). The stations were selected based on three criteria: (*i*) short distance from the coastlines of the Tyrrhenian Sea, Ionian Sea, and Sicilian Channel Sea; (*ii*) availability of continuous recordings during the 2018–2021 period with minimal data gaps; (*iii*) broadband sensors capable of recording the entire microseism band.

Furthermore, for microseism array analysis, we used seismic signals recorded by the Mt. Etna monitoring permanent network run by INGV-OE, specifically those recorded by 15 stations equipped with broadband three-component Trillium 40-s seismometers (Nanometrics™), which record at a sampling rate of 100 Hz (see Fig. 1a,c and Table A2).

2.2. Sea state data

To quantify the relationship between microseisms and wave height time series in the Sicilian Seas, we obtained sea wave data from the Copernicus Marine Environment Monitoring Service (CMEMS) for the period spanning 2018–2021. Specifically, we used the "MEDSEA_HINDCAST_WAV_006_012" product, which is the hindcast output of the Mediterranean Sea Waves forecasting system. This product has an hourly temporal resolution and a spatial resolution of $1/24^{\circ}$. We focused on the significant wave height data, which roughly corresponds to the average of the highest one-third of the waves (Steele and Mettlach, 1993).

2.3. Spectral and amplitude analysis

To characterize the spectral features of seismic signals recorded by the three-component of 14 selected stations, the short-time Fourier transform was performed. This involved moving an 81.92-s-long time window along the entire length of the traces and calculating a spectrum for each non-overlapping position of the window. To obtain daily spectra, all spectra from the same day were averaged using Bartlett's method (Bartlett, 1948). The daily spectra were then collected and displayed as spectrograms, which are 3D plots with time on the x-axis, frequency or period on the y-axis, and power spectral density (PSD) indicated by a color scale. Single spectra were also computed for each three-component station as the median of all daily spectra, to obtain information on the spectral content of the seismic signals recorded by different stations throughout the period of interest.

Temporal variability of the seismic data was investigated by calculating the daily root-mean-square (RMS; Kenney and Keeping, 1962) amplitude of the seismic signal band-pass filtered in 14 different frequency ranges: (*i*) 0.05–0.1 Hz; (*ii*) 0.1–0.2 Hz; (*iii*) 0.2–0.35 Hz; (*iv*) 0.35–0.5 Hz, (*v*) 0.5–0.65 Hz; (*vi*) 0.65–0.8 Hz; (*vii*) 0.8–0.95 Hz; (*viii*) 0.95–1.10 Hz; (*ix*) 1.10–1.25 Hz; (*x*) 1.25–1.40 Hz; (*xi*) 1.40–1.55 Hz;

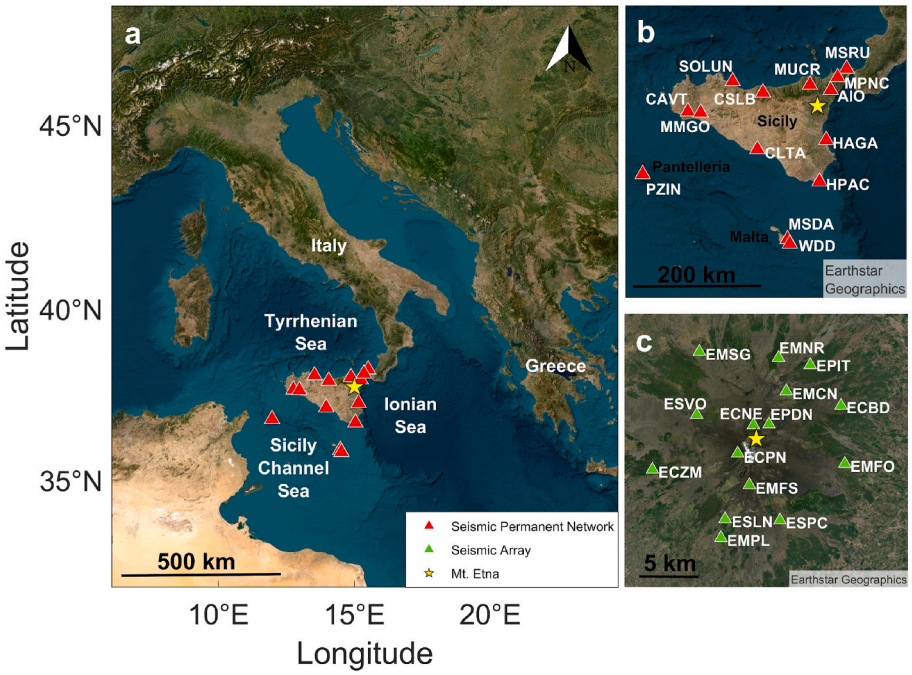


Fig. 1. Location of the seismic station used in this work. a) Map of part of the Mediterranean sea showing the location of seismic stations used for the analysis. b) Map of the Sicilian area with a selection of the broadband seismic stations available in the INGV-OE and University of Malta databases and used in the spectral, amplitude, correlation and machine learning analysis. c) Map of the summit area of Mt. Etna with a selection of the broadband seismic stations managed by INGV-OE and used in the array analysis. In (a), (b) and (c), the base maps were retrieved from Earthstar Geographics®. In (a), (b), and (c), the yellow star represents the roughly position of the summit area of the Mt. Etna.

(*xii*) 1.55–1.70 Hz; (*xiii*) 1.70–1.85 Hz; (*xiv*) 1.85–2.00 Hz. In particular, the first four frequency bands cover the typical band of microseism. The daily sampling of the RMS amplitude time series was chosen to investigate amplitude variations of microseism in the long term.

2.4. Correlation analysis between microseism amplitude and significant wave height

Following previous studies (e.g., Bromirski, 2001), the correlation coefficients between the time series of seismic RMS amplitudes and those of the significant wave height were calculated. These coefficients were computed for each grid cell of the hindcast maps for the 2018–2021 period to obtain information about their spatial variability. This kind of analysis provides some information about the location of the main sources of the microseism recorded by the selected stations. Following the idea of Craig et al. (2016), we used the Spearman correlation coefficient to explore the non-linear dependence between seismic RMS amplitudes and significant wave heights. The Spearman correlation coefficient is defined as a nonparametric measure of rank correlation (Craig et al., 2016).

2.5. Analysis of spatial distribution of microseism amplitude

As one of the major issues in the development of the microseism-based monitoring system of the sea state in the Sicily Channel is the "noise" generated by sea wave activity in the Ionian and Tyrrhenian Seas, we calculated the spatial distribution of the microseism amplitude during the time intervals characterized by different sea wave features. These time intervals were selected based on one of the following conditions: (*i*) intense sea wave activity in the Sicily Channel but not in the Ionian and Tyrrhenian Seas; (*ii*) intense sea wave activity in the Tyrrhenian Sea but not in the Sicily Channel and Ionian Sea; (*iii*) intense sea wave activity in the Ionian Sea but not in the Sicily Channel and Tyrrhenian Sea. To do that, we made use of the significant wave height information for the Mediterranean Sea and the 2018–2021 period. In this case, we specifically selected days characterized by mean height in the first area higher than 90th percentile (2.1 m), and mean height in the other two areas lower than 76th percentile (1.4 m). Overall, three days turned out to be clearly characterized by these peculiar features: 15 June 2018, 29 January 2020, and 30 October 2021.

2.6. Array analysis

To locate the source of the microseisms in the Mediterranean Sea, we utilized fifteen stations from the Mt. Etna seismic permanent network to define a roughly circular array (see Fig. 1a,c). We performed array analysis (e.g., Rost and Thomas, 2002) to measure the apparent velocity and back azimuth of the arriving wavefront of the microseism signals for the PM, SM, and SPSM frequency bands.

Assuming a planar propagation of the wavefront across the array, the resolution of the array analysis depends on the geometry and size of the array, as well as the relationship between the sensor-source distances and the wavelength of the signal of interest (Havskov and Alguacil, 2016). To ensure the correct processing of the data, three conditions had to be satisfied in the spatial configuration of the array. Firstly, the aperture of the array should be greater than a quarter of the signal wavelength that we want to analyze (Aster and Scott, 1993). Secondly, to avoid spatial aliasing, the wavelength of the signal should be at least comparable with the array interspacing (Asten and Henstridge, 1984). Lastly, distances between the array receivers and the source of the signal must be greater than one wavelength (Havskov and Alguacil, 2016).

To plan the geometry of the array needed to investigate microseism signals, the Array Response Function (ARF) was calculated using the Beam Pattern Function (Capon, 1969) for the frequency range of the microseisms (Supplementary materials, Figure A1). In the slowness domain, ARF describes the resolution and sensitivity of a plane wave vertically impinging at the array with a slowness of 0 s/km and with a specific frequency/wavelength. ARF depends only on the relative position of the array elements for a given slowness-frequency (Capon, 1969). The position and the height of the peaks provide information about the capability of the array to acquire a coherent wavefield for a specific frequency range.

In this study, we used the f-k (frequency-wavenumber) analysis to locate the source of microseism signals (e.g., Borzì et al., 2022). Array analysis on microseism was performed during the time intervals characterized by different sea wave features, as shown in Section 2.5. Overall, we followed a series of processing steps on the seismic signals. First, we applied demeaning and detrending. Then, we filtered the

signals to isolate the frequency band of interest for microseisms. We subdivided the data into 120-s tapered windows, and excluded windows with amplitude transients, such as volcano-tectonic earthquakes, long-period events, very long-period events, and regional earthquakes, which were detected using the STA/LTA technique (Trnkoczy, 2012). Finally, we applied the f-k analysis to each window using a slowness grid search, ranging from −1 to 1 s/km in the east and north components of the slowness vector, with a spacing of 0.05 s/km.

2.7. Regression analysis by machine learning

To build models able to predict the sea state using microseism, we performed a regression analysis that in the Machine Learning (ML) field falls under the so-called supervised learning wherein the algorithm is trained with both input features (seismic RMS amplitudes) and output/target values (significant wave height data). It aims to establish a relationship among variables by estimating how input variables affect the others in output.

To provide reliable predictive models, the method we applied is composed of three macro-steps (Fig. 2): (*i*) pre-processing, (*ii*) training and cross-validation, and (*iii*) testing and evaluation.

Phase (*i*) is crucial for optimizing the performance of any ML process (Kuhn and Johnson, 2013). This step may involve several actions, depending on the type of dataset being used. As input features we considered the hourly RMS of the seismic signals filtered in the above-mentioned frequency bands for each station and component, thus in total 588 features (14 bands for 14 stations).

First, the input features affected by a great amount of data gaps (greater than 14%) were deleted from the dataset. On the contrary, when the number of missing data is low, a linear interpolation (e.g., McKinney, 2010) was applied for data imputation. Concurrently, an earthquake catalogue from the United States Geological Survey (USGS; https://earthquake.usgs.gov/fdsnws/event, last access May 2023) data archive is used to delete data including teleseisms (with a magnitude greater than 7.0) and regional earthquakes (with a magnitude greater than 5.5). Indeed, such earthquakes produce strong seismic signals with a broad frequency range which include some frequencies similar to those of the microseism signals (e.g., Tanimoto et al., 2015; Anthony et al., 2020). In this way, models can be built without introducing ineffective artifacts or reducing the statistical power of the analysis (e.g., Ferretti et al., 2018). Then, Box-Cox power transformation (Box and Cox, 1964) is applied to features showing high values of skewness. This method can optimize the data's resemblance to a normal distribution, improving the accuracy of predictions made using linear regression and facilitating the training of ML nonparametric models. Successively, features and target data were normalized to avoid different orders of magnitude having an impact on the model. In this case, the minimum and maximum values of the un-normalized data are used for rescaling (Han et al., 2022). Finally, seismic (features) and sea wave state (target) data were randomly and temporally subdivided into N non-consecutive chunks composing training and testing sets. Indeed, another problem concerns the correct sampling of the datasets, especially during the splitting into training and testing ones. Considering the input features and targets are time series with low autocorrelation function decay, training data sufficiently close in time to testing one could affect the model's capability to generalize to an independent/unseen dataset, leading to the development of overfitting phenomena (e.g., Cook and Ranstam, 2016). To avoid this problem, we split the 4 years of data into 40 segments/chunks that were combined to form the training (70%) and testing (30%) sets. In addition, we prepared the validation set by using about 10% of the available test.

Concerning step (*ii*), we used the following three ML techniques to build predictive models: random forest (RF) regression, Light Gradient Boosting (LGB), and K-nearest neighbors (KNN) regression.

RF is based on decision trees often used for classification and regression (Ho, 1995). It operates by using a bagging method (Breiman, 1996, 2001) for the aggregation of the results. Indeed, it builds in

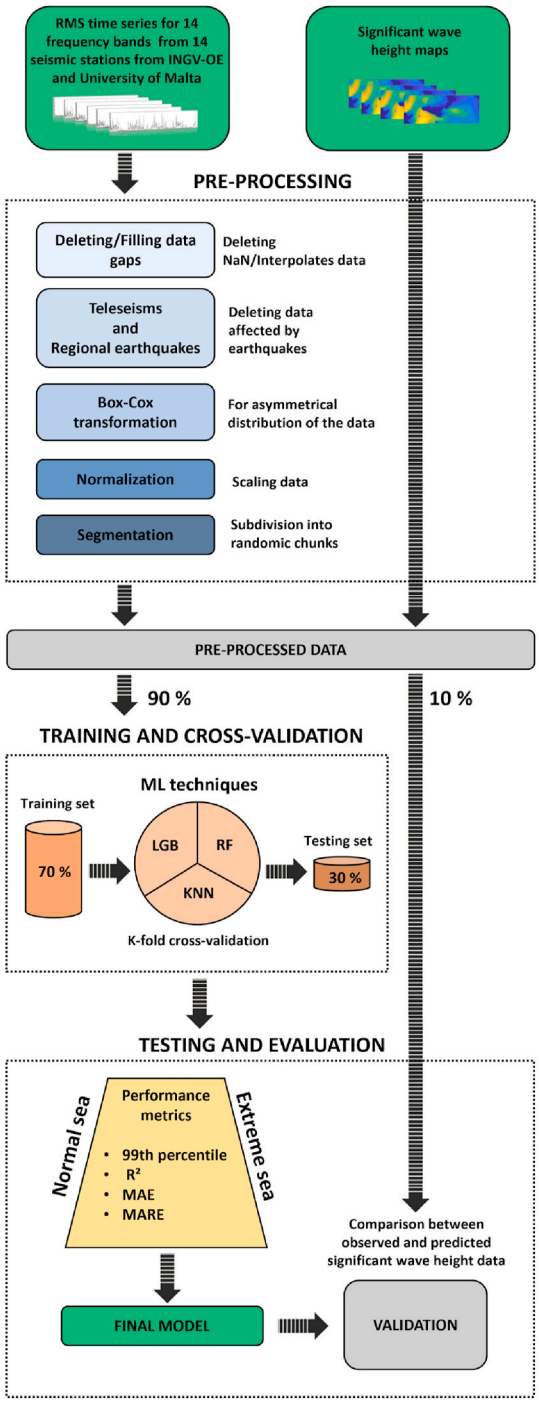


Fig. 2. Process workflow of the ML analysis. The process is described from data collection to the evaluation of the predictive model.

parallel decision trees during the training. To build each decision tree, it randomly selects *j* data points from the training set. The prediction for a new data point, in the case of regression, will be the average of the N trees predictions for that point.

LGB is very similar to RF except for the method used to aggregate the data, i.e. the boosting (Bauer and Kohavi, 1999; Natekin and Knoll, 2013). When LGB builds each sequence tree, it uses the histogram-based algorithm to create the optimal split point. To reduce the complexity of this method, it downsamples the data and features using Gradient-based One Side Sampling (GOSS; Ke et al., 2017) and Exclusive Feature Bundling (EFB; Ke et al., 2017). GOSS is a novel sampling method that downsamples the instances on the basis of gradients. It retains instances

with large gradients while performing random sampling on instances with small gradients. The EFB serves to speed up tree learning. LGB achieves this goal by identifying in the data the features that are mutually exclusive, that is, that never take zero values simultaneously, and bundling them into a single feature.

K-nearest neighbors (KNN) is a non-parametric method that can be used for both classification and regression tasks (Altman, 1992). In KNN regression, the prediction of a new sample is made by considering the K-closest samples in the training set (Altman, 1992; Kuhn and Johnson, 2013). Essentially, the output of a new input is determined by taking the average of the values of its K-nearest neighbors in the feature space of the training set.

Before training the model, we performed a grid search to identify the optimal hyperparameters for our data (e.g., Larochelle et al., 2007; Hinton, 2012). For a specific model, we calculated the best combination of all hyperparameters by defining a grid with the values to be tested and evaluating the model performance in terms of mean absolute error (MAE, Willmott and Matsuura, 2005) on the test sets. To avoid overfitting, we applied cross-validation (Kuhn and Johnson, 2013), which we will discuss later. A resume of the best hyperparameters resulting from the application of this technique is reported in the Supplementary Materials (Tables A3).

Step (*ii*) included also the evaluation of the best ML model by performing the k-fold cross-validation (Kuhn and Johnson, 2013, García et al., 2019). It implies partitioning the original input and output datasets into k subsets and constraining a model for each partition. Indeed, the hold-out method (i.e. splitting of the dataset into training and testing set) is repeated k times by using a part of the data for training the model and a part for testing it. For each time, one of the k subsets is used as the testing set and the other k-1 subsets are put together to form a training set. The error estimation, which in our case will be given by the absolute difference between the predicted wave height and the measured one, is averaged over all k trials to get the total effectiveness of our model. Thus, every data point gets to be in a testing set exactly once and gets to be in a training set k-1 time. In particular, we used this technique both in the hyperparameters searching phase, and in the training phase of each model. In both cases, RMS amplitude and significant height wave time series were partitioned into 3 subsets ($k = 3$).

For the phase (*iii*), each model is tested on the testing dataset, which is composed of input data unused in the training process. The difference between the true output and that predicted by the models allows the evaluation of the quality of this according to different metrics. To evaluate the goodness of the trained models we used the three metrics:

- The 99th percentile of the absolute and relative errors.
- The coefficient of determination (e.g., Draper and Smith, 1998), more commonly R^2, which is an index that measures the relationship between the variability of the data and the correctness of the model used. If the R^2 score is close to 1, then the model is able to return a reliable prediction.
- The mean absolute error (MAE, Willmott and Matsuura, 2005), which is the average of the absolute values of the individual prediction errors across all instances in the test set. Each prediction error is the difference between the true value and the predicted value for the instance.
- The mean absolute relative errors (MARE).

Through the metrics described above, we defined the behavior of the model for data recorded during sea normal conditions and extreme events (e.g., Miglietta, 2019; Faranda et al., 2022). For this purpose, during the evaluation phase, we selected the samples of the target variable (significant wave height) lower than their 99th percentile to extract significant height values in normal conditions. On the contrary, samples of the target variable greater than their 99th percentile are selected to define sea conditions that are typical of an extreme event.

Finally, the final model was trained with the whole dataset from January 2018–December 2021, except for September–December 2018. Indeed, this period composed the validation set and it was used to verify the model's capabilities to predict new entries.

3. Results

The spectral analysis has shown that most of the highest peaks were focused below 5 s over the whole period (Fig. 3), although smaller spectral peaks were recognized above 10 s during brief time intervals. Significant changes of the spectral features ranged between 2 and 10 s, with higher amplitudes during the winter periods. Indeed, a seasonal modulation was observed on the long term, with maxima during winter and minima during summer (Fig. 3). It is worth noting that most seismic energy was observed in the SPSM frequency band, while the PM or SM exhibited smaller spectral amplitudes (Fig. 4). Averagely, the peaks between 2 and 5 s (0.2–0.5 Hz) were the strongest among the typical frequency bands of microseism activity, as shown in Fig. 4. Indeed, the absolute maximum value was observed around the 2 s, while the spectral amplitudes decreased for increasing periods (Fig. 4) up to 10 s. A smaller peak was focused between 10 and 20 s (Fig. 3). Most of these observations can be also recognized in Fig. 5, which shows the temporal variability of the seismic amplitude in the typical frequency bands of the microseism. In particular, the highest RMS amplitude values were observed in the SPSM frequency range (Fig. 5c), for which the seasonal modulation was more pronounced than the PM (Fig. 5a) and SM (Fig. 5b). Indeed, in the PM frequency range, microseism amplitudes reached the smallest values. However, high and sudden variations of RMS amplitude values were observed in the PM frequency bands (Fig. 5a), as also shown in the spectrograms (Fig. 3).

Regarding the correlation analysis, maps, gathering together the correlation values obtained in the nodes of the whole Mediterranean Sea, were obtained for the three components of each station and the 14 different frequency bands (Fig. 6). In particular, the marks shown in Fig. 6 represent the barycentres of the areas with correlation coefficients greater than 95th percentile. Furthermore, considering the distance from the station recording the microseism to the sea grid cell providing the significant wave height data, cross-plots showing the frequency in the x-axis, the correlation coefficient in the y-axis, and the distance in the color of the marks were obtained (Fig. 7). The highest values of the correlation factor were observed between 0.2 and 0.5 Hz that represented the frequency band of SPSM (Fig. 6). The portions of the Mediterranean Sea, showing the maximum value of correlation, were the ones closest to the stations (Fig. 7). In addition, Fig. 7 clearly shows a decrease in the correlation values with increasing distance from the station recording the microseism to the sea grid cell providing the significant wave height data and with increasing frequency of analysis used for the RMS amplitude calculation.

Considering the criteria described in Section 2.5, an average value of RMS amplitude was obtained per each station in the frequency bands of the microseism and in each time interval selected. A map showing the spatial distribution of RMS amplitude and average significant wave height values for each frequency band and per each time interval is shown in Fig. 8. These maps highlighted a good match between the spatial distributions of significant wave heights and RMS amplitudes, especially in the 0.2–0.5 Hz frequency band (see c,f,i panels). In particular, when the Sicily Channel was characterized by higher sea waves than the Ionian and Tyrrhenian Seas, maximum RMS amplitudes were observed along the Sicily southern coastlines (Fig. 8a,b,c). When the highest wave heights were observed in the Tyrrhenian Sea, as expected, the highest seismic amplitudes were mostly recorded along the Sicily Northern coastlines (Fig. 8d,e,f). On the other hand, when the Ionian Sea recorded the highestwave heights, the distribution of RMS amplitudes exhibited fewer clear results because of high amplitudes observed also along the southern coastlines (Fig. 8g,h,i).

Concerning the array analysis, the Mt. Etna seismic permanent network turned out to be a reliable array to locate the microseism

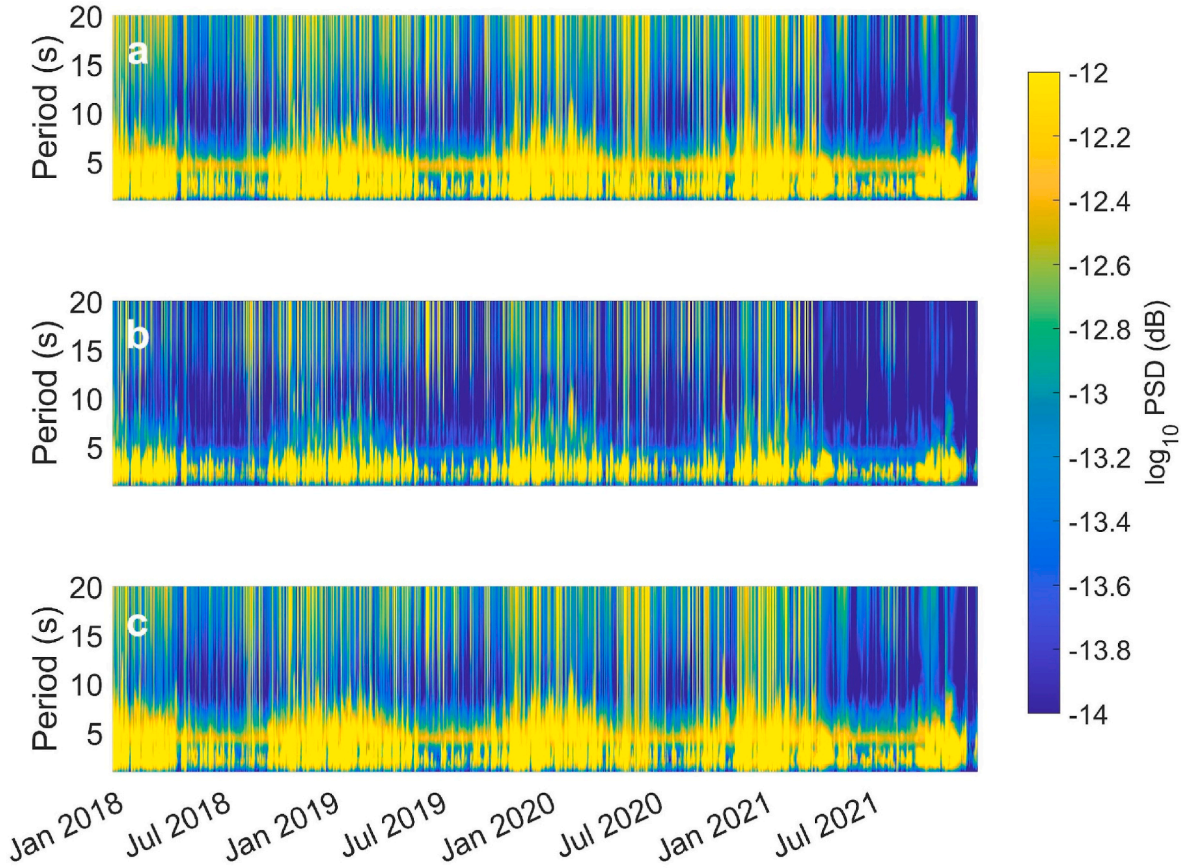


Fig. 3. Temporal variability of the spectral features of microseism during 2018–2021 period. Examples of spectrograms of the seismic signal recorded by the East (a), Vertical (b) and North (c) components of CAVT station located along the Southern Sicily coastlines.

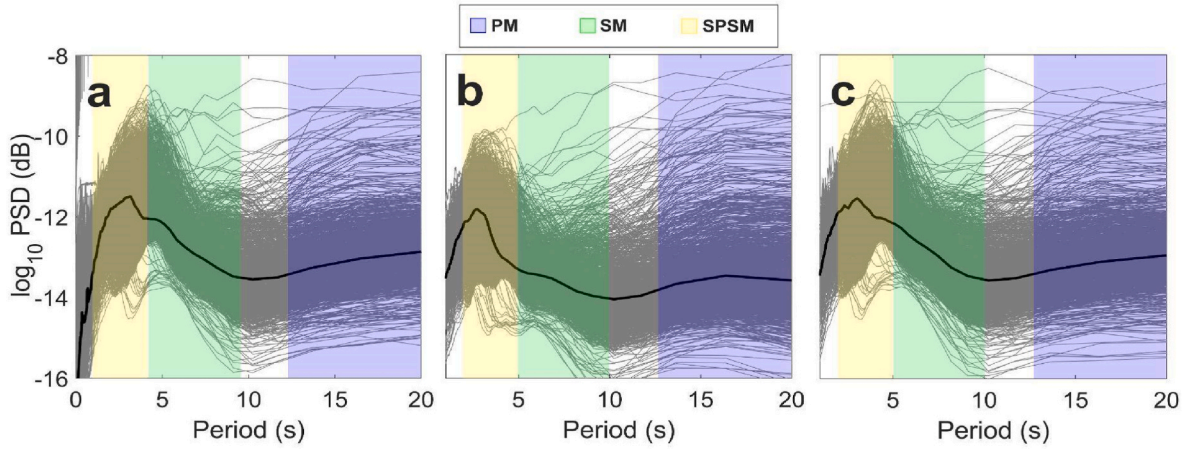


Fig. 4. Spectral features of microseism. Examples of daily spectra of the seismic signals recorded by the East (a), Vertical (b) and North (c) components of CAVT station (grey lines) during 2018–2021 period. For each diagram, the thick black line represents the median spectrum, while the colored boxes refer to the period band for the PM (blue), SM (Green), and SPSM (yellow), based on literature concerning microseism (e.g., Hasselmann, 1963; Bromirski et al., 2005).

sources in the SM band (Fig. 9b,e,h). Instead, back azimuth and apparent velocity values turned out to be sometimes ambiguous (e.g., Fig. 8c) and more scattered (e.g., Fig. 9a,d,g) in the SPSM and PM bands, respectively. During stormy days in the Sicily Channel Sea, back azimuth values, calculated in the SM band, pointed toward the southern coastline (Fig. 9b). Alternatively, when the highest significant wave heights are observed in the Tyrrhenian Sea, the back azimuth values rotate pointing north-westward (Fig. 9e). In addition, when the Medicane Apollo hit the Sicilian Ionian coastlines (Fig. 9h), back azimuth values indicate the Catania Gulf, pointing south-eastward. As for the apparent seismic velocity values, the histograms in Figure A2 show values of ~1.5–2.0 km/s.

As for the regression analysis, we found the RF algorithm to be the most effective approach for generating reliable predictions of the sea state in the Sicily Channel (Fig. 10), especially in terms of R^2 and errors of analysis (Figs. 11 and 12). Although the model's performance is still not optimal for both normal (in terms of relative error, Fig. 10c) and extreme (in terms of absolute error, Fig. 10b' and 10d') conditions, the

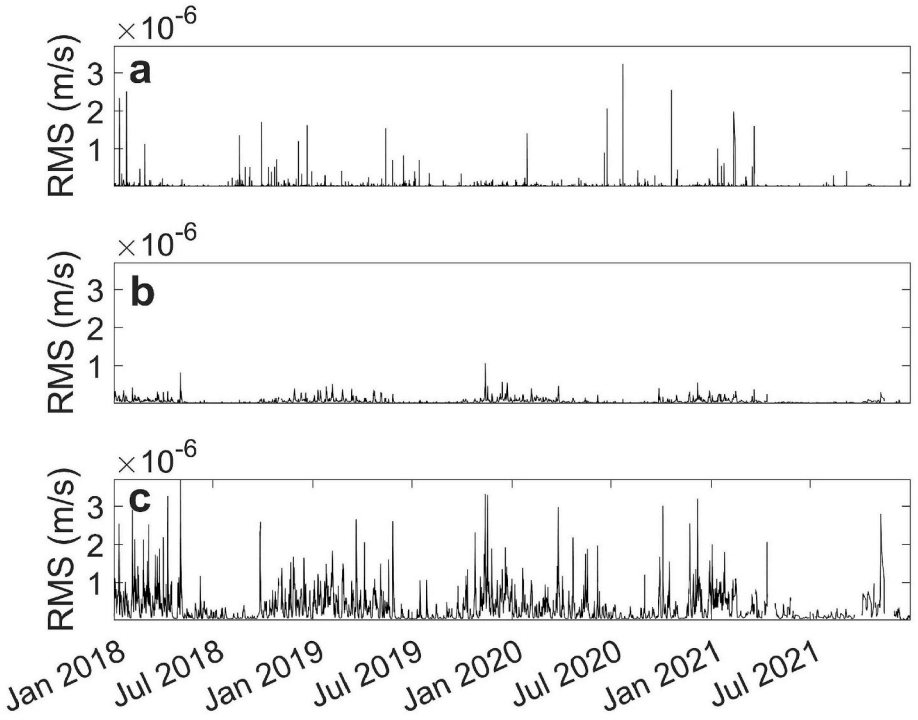


Fig. 5. Temporal variability of the microseism amplitudes during 2018–2021 period. Examples of RMS amplitude time series of the seismic signal recorded by the vertical component of CAVT station for PM (a), SM (b) and SPSM (c) frequency bands, based on literature concerning microseism (e.g., Hasselmann, 1963; Bromirski et al., 2005).

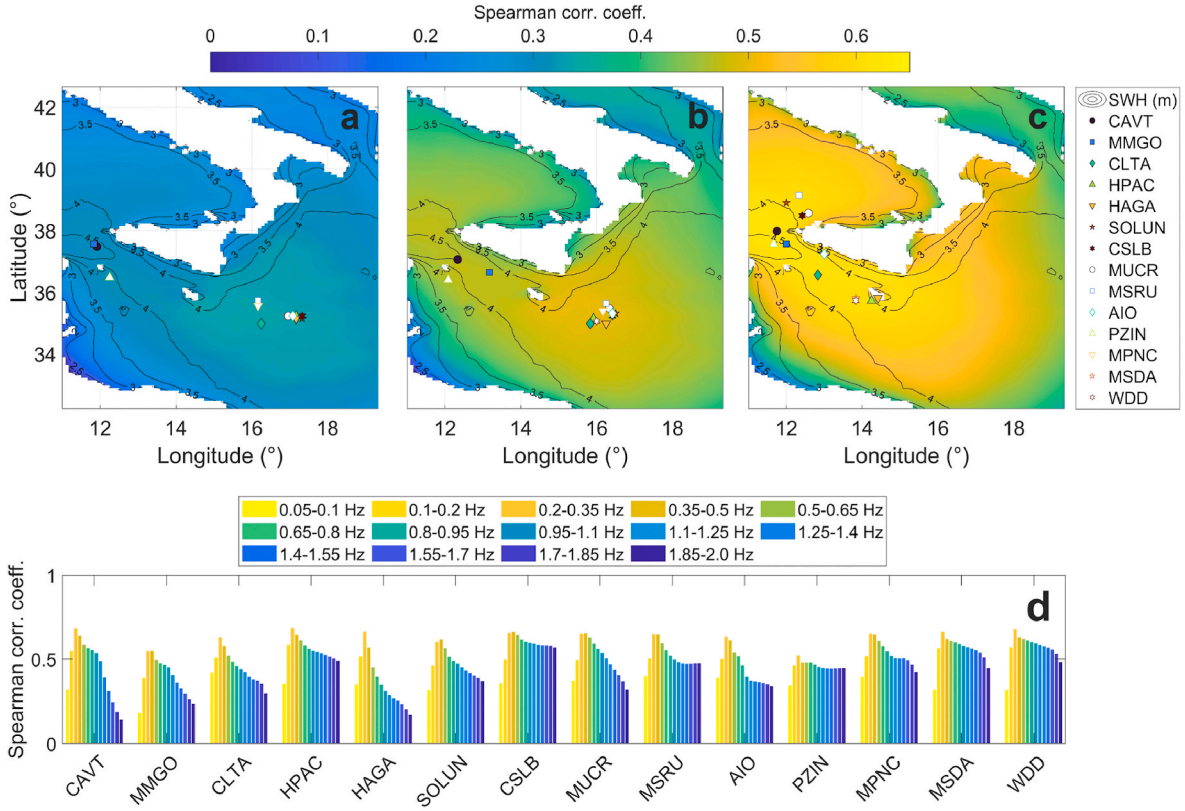


Fig. 6. Correlation between seismic and sea state data. Example of correlation maps obtained for the vertical component of all of the stations for the (a) PM, (b) SM and (c) SPSM frequency bands and for 2018–2021 period. In (a), (b) and (c) the colored marks represent the barycentres of the areas with correlation coefficients greater than 95th percentile and calculated for each station. The colored surface refers to the average of the correlation maps obtained for each station. The black contouring refers to the 99th percentile of the significant wave height in meters for the period 2018–2021. d) Maximum Spearman correlation coefficients computed between significant wave height and seismic RMS amplitudes for each station (only vertical component) and frequency band for the period 2018–2021.

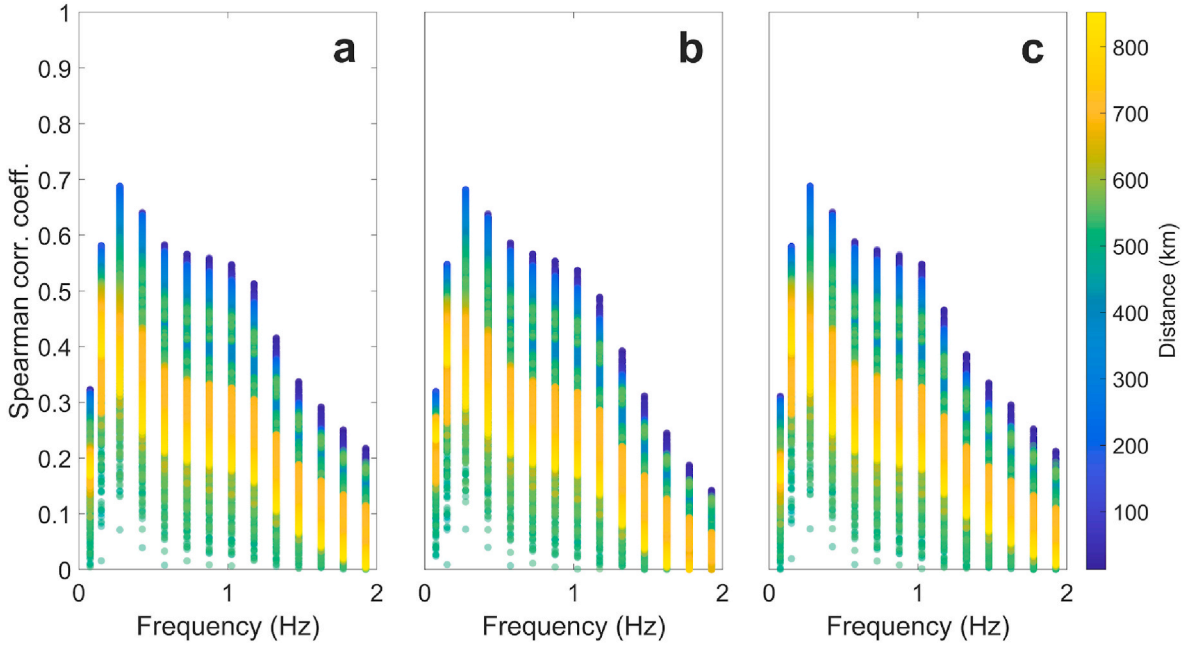


Fig. 7. Results of the correlation analysis. Cross-plots showing the relationship between Spearman correlation coefficient, computed between significant wave height and seismic RMS amplitudes for the East (a), Vertical (b) and North (c) components of CAVT station, the distance between the seismic station and the sea grid cell and the frequencies of analysis used for the RMS amplitude calculation.

R^2 (Fig. 11) indicated that the model can accurately predict sea state under any conditions. In particular, the results in Fig. 10 showed that for both normal and extreme sea cases the average absolute relative error (Fig. 10e,e') was distributed almost uniformly below unity indicating a good average predictive ability of the model in relation to any sea condition. Only the Ionian coast had higher average relative error values, and this also reflected a distribution of relative error values that extended to larger values in the same area, as shown by the 99th percentile relative error map (Fig. 10c,c'). In absolute terms, the mean errors in the normal sea state cases were at significantly low values (below 1 m, see Fig. 10d) with the distribution of absolute errors reaching values just above the meter in the western parts of the Sicilian Channel (see Fig. 10b). In cases of extreme seas, the mean absolute error was mostly below 2 m in the whole area, but the Ionian coast suffered in this case from a distribution of absolute errors that reached at the 99th percentile even high values of discrepancy (even on the order of 4–6 m, see Fig. 10b'). However, it should be noted that not all the regions of the Sicily Channel are predicted with the same level of accuracy. Areas with the highest prediction errors or lowest R^2 values were generally found in regions with poor station azimuthal coverage and at greater distances (>200 km) from the seismic stations or the coastlines (Fig. 11 and Fig. A4). Finally, in the case of RF algorithm, the comparison between the observed and predicted significant wave height data during the testing period, September–December 2018 (Fig. 13), showed very similar values in terms of spatial distribution (Fig. 13a and b) and time series (Fig. 13d). It was also confirmed by the low error values (0.21 ± 0.23 m) (Fig. 13c) and the high values of R^2 equal to 0.89 (Fig. 13e).

4. Discussion

Spectra of the microseism recorded by stations deployed along the Sicilian coastline showed very high amplitudes of the SPSM (Figs. 3 and 4). Considering seismic stations were very close to the sea coastline, SPSM was recorded very well due to the presence of nearby seismic sources related to local sea state and wave activity (e.g., Bromirski et al., 2005; Chen et al., 2011; Moschella et al., 2020; Cannata et al., 2020).

Results obtained from spectral and amplitude analysis indicate a clear relationship between microseism and sea state data. The seasonal modulation shown in the spectrograms and RMS amplitude time series (Figs. 3 and 5) has been observed at temperate latitudes in all areas around the world (e.g., Aster et al., 2008; Stutzmann et al., 2009). As expected for the Northern Hemisphere, considering the sea was stormier in winter, seismic stations showed a clear contamination by microseism due to the more efficient energy transfer from the sea to the solid Earth, although some sudden changes of PM (Figs. 3 and 5a) on the short term may be related to the occurrence of teleseisms or regional earthquakes (e.g., Tanimoto et al., 2015; Anthony et al., 2020).

From the results obtained by performing correlation analysis (Figs. 6 and 7), a good match between the spatial distributions of significant wave heights and RMS amplitudes was shown. It was possible to infer as the seismic sources generating microseism were located close to the recording stations (at distances up to 400 km), especially for the SPSM band. Indeed, being characterized by higher frequency content, the SPSM showed a quick attenuation with distance (e.g., Bromirski et al., 2005), although distant areas may contribute to its generation (e.g., Beucler et al., 2015; Becker et al., 2020). As shown in Fig. 8, microseism space distribution amplitude was affected by the conditions of the seas surrounding Sicily in terms of significant wave height.

However, the correlation analysis shows only the dominant source areas over the whole analyzed time period (2018–2021), and it may be affected by the limited temporal and spatial resolution (Essen et al., 1999). Despite the low correlation values observed in the SM and PM bands, we do not exclude the possibility that part of the recorded microseism was generated in open sea or nearby the Mediterranean coasts. Indeed, the source of the SM can partially be associated with wave–wave interaction mechanisms in the deep ocean (e.g., Cessaro, 1994; Chevrot et al., 2007); while PM can be also generated by remote source regions from the seismic stations. For example, Gualtieri et al. (2019) performed a global-scale simulation to understand the location of the dominant source areas of the PM. For some stations, they observed how most of the PM at low frequency (~0.05 Hz) was generated by source area located at coastlines thousands of kilometres away from the sensors. In particular, they observed that the slope of the bathymetry in shallow water plays a key role in the generation and propagation of PM.

Based on the ARF, the approximately circular array shows a strong response for both the PM (Figure A1a) and SM (Figure A1b) cases. This is

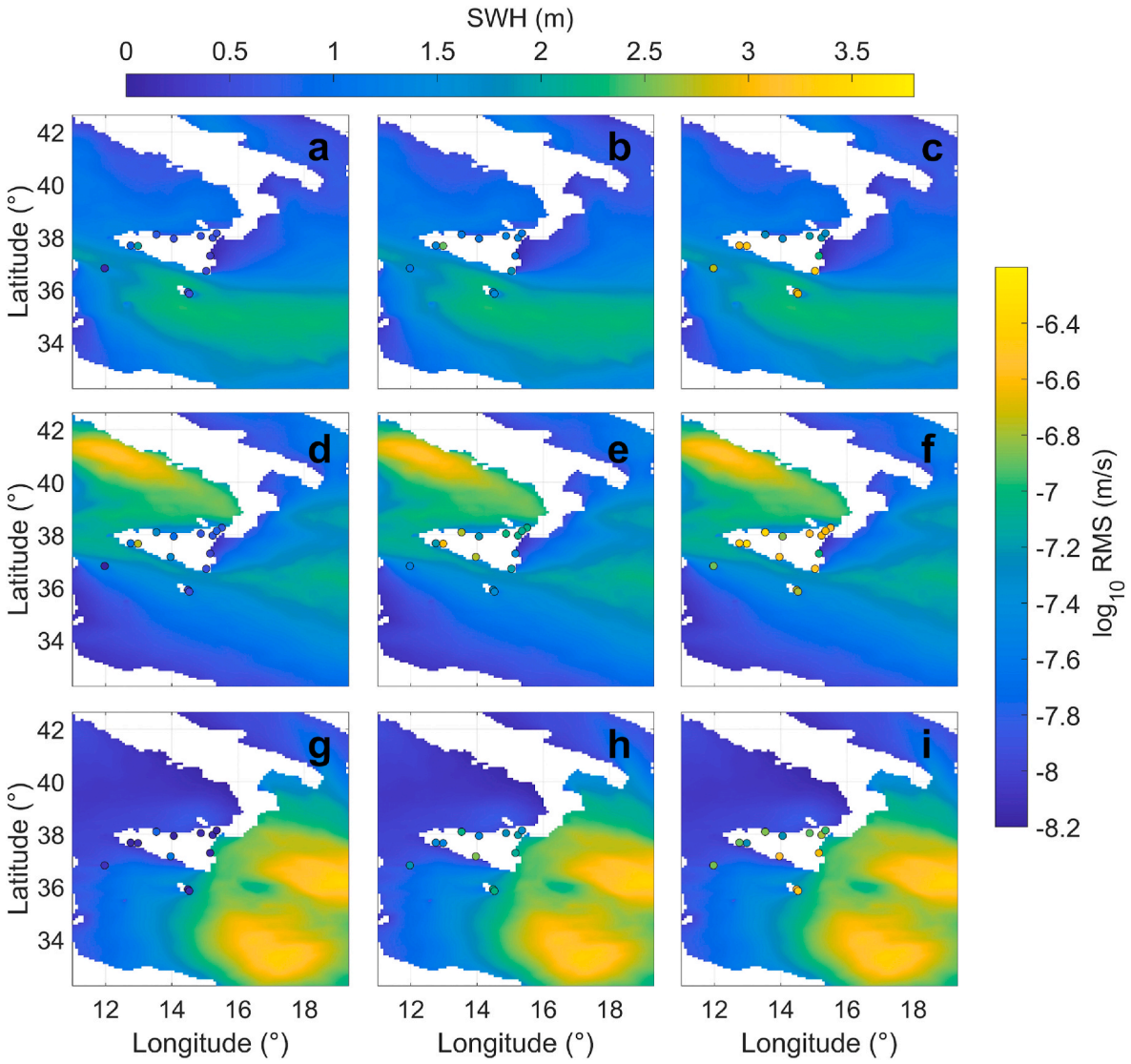


Fig. 8. Spatial distribution of seismic and sea state data. Maps of the spatial distribution of the daily average significant wave height (SWH) and of the daily average RMS amplitude values of the seismic signals (colored dots; only vertical component) recorded during (a, b, c) 15 June 2018 (the Sicily Channel has the higher wave height), (d, e, f) 29 January 2020 (the Tyrrhenian Sea has the higher wave height) and (g, h, i) 30 October 2021 (the Ionian Sea has the higher wave height). The RMS amplitude values refer to the (a, d, g) PM, (b, e, h) SM and (c, f, i) SPSM frequency bands.

because the wavelengths of PM (~26 km) and SM (~10 km) are comparable to the aperture of the array (~16 km) and the interspacing between sensors (~6 km), particularly if the velocity of the S-waves (Vs) is around 2 km/s in the top few kilometers of the crust of Mount Etna (e.g., Patanè et al., 1994). Furthermore, if the microseism sources are located at minimum distances of approximately 20 km, 45 km, and 100 km from the center of the array (distances from the Ionian Sea, Tyrrhenian Sea, and Sicily Channel Sea, respectively), the Etna circular array should be capable of locating the microseism sources assuming a planar wavefront.

The Mt. Etna seismic permanent network turned out to be a reliable array to locate the microseism sources in the SM band (Fig. 9). It was observed that the SM sources appeared to be located in extended areas in the Tyrrhenian, the Ionian and the Sicily Channel Seas. Although some results may be considered as ambiguous (e.g., Fig. 9c), we also were able to locate microseism sources in the SPSM band, especially in areas nearby the northern and eastern coastlines (e.g., Fig. 9f,i). The lack of information about the SPSM source in the Sicily Channel (Fig. 9c) may depend on the great distance between the southern coastline and the centre of the Etnean array (~100 km) or the lower depth of the basin, causing energy losses by both scattering and transfer to the solid Earth (e.g., Bromirski et al., 2013). These factors may reduce the signal-to-noise ratio of the seismic signals recorded by the seismic stations, and the array analysis may give ambiguous back azimuths pointing toward other areas. For instance, the array may be more influenced by nearby seismic sources located in the Tyrrhenian and Ionian Seas. At these shorter distances (~20 and ~45 km from the Ionian Sea and Tyrrhenian Sea, respectively), the SPSM wavefield may be affected by less distortions and energy losses during the path (e.g., Bromirski et al., 2005), favouring the tracking of the seismic sources in these areas rather than in the Sicily Channel Sea. Instead, the results obtained for the PM band (Fig. 9a,d,g) may be related only to the low seismic energy observed in this frequency band (e.g., Gualtieri et al., 2019; Moschella et al., 2020; Cannata et al., 2020), as also shown in Figs. 4 and 5. Indeed, seismic signals impinge at the array with a low signal-to-noise ratio, affecting the capability of the array to identify a coherent wavefield and return accurate back azimuths (e.g., Rost and Thomas, 2002).

Overall, the results of array analysis agree with some previous studies demonstrating that the back azimuth of microseism is well-correlated to ocean-wave heights (e.g., Chevrot et al., 2007; Chen

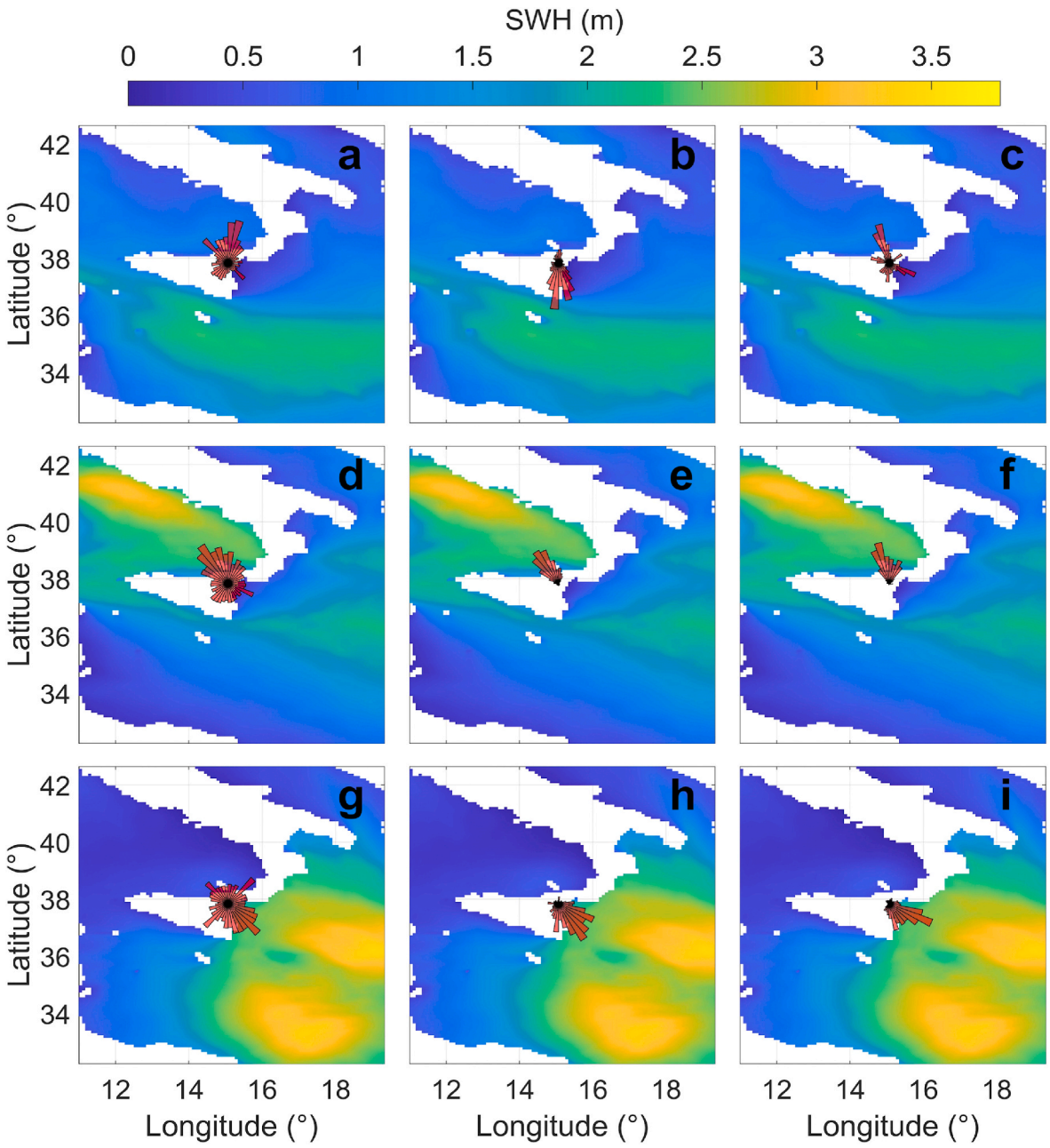


Fig. 9. Results of the array analysis. Average hindcast maps showing the significant wave heights (in m) during the period of the most intense sea wave activity in (a, b, c) the Sicily Channel (15 June 2018), (d, e, f) the Tyrrhenian (29 January 2020) and (g, h, i) Ionian Seas (30 October 2021). The rose diagram, located at the center of the Etna seismic permanent network (see Fig. 1c), shows the distribution of the back azimuth values computed by f–k analysis for the (a, d, g) PM, (b, e, h) SM and (c, f, i) SPSM frequency bands.

et al., 2011). Results obtained in the SM band are compatible with Borzì et al. (2022), who investigated the microseism recorded in the eastern part of Sicily during the Medicane Apollo (25 October–5 November 2021). In this case, the authors showed a good match between the positions of the Medicane/minor storms and the seismic source of the microseism highlighted by array analysis. Concerning the SPSM band, the results are also in agreement with Moschella et al. (2020), who explored the microseism acquired along the coastlines of Eastern Sicily. In their work, the SPSM sources were located in the shallow waters of the Catania Gulf and the Northern Sicily coastlines, especially when the highest significant wave heights were observed in the Ionian Sea and the Tyrrhenian Sea, respectively. Concerning the seismic apparent velocity estimated in the SM and SPSM band (Figure A2), the values of ~ 1.0–3.0 km/s agree with the surface wave velocity, as retrieved by investigating the ambient seismic noise in the coastlines of the Sicily (Moschella et al., 2020; Borzì et al., 2022), in the northeast of the Netherlands (Kimman et al., 2012), in the New Zealand (Brooks et al., 2009) and in the Valley of Mexico (Rivet et al., 2015).

Finally, ML techniques were reliable to reconstruct maps of significant wave height by using microseism recorded by distinct seismic stations and in different frequency bands. Specifically, such methods allowed us to predict the significant wave height in the Sicily Channel with fairly low error (Fig. 10) by using microseism recorded at distinct seismic stations in different frequency bands. Choosing the most appropriate model has played a crucial role in our study. Through the use of cross-validation, we were able to identify the model that would best generalize to an independent dataset. In particular, the RF algorithm represents the ML technique showing the best performance,

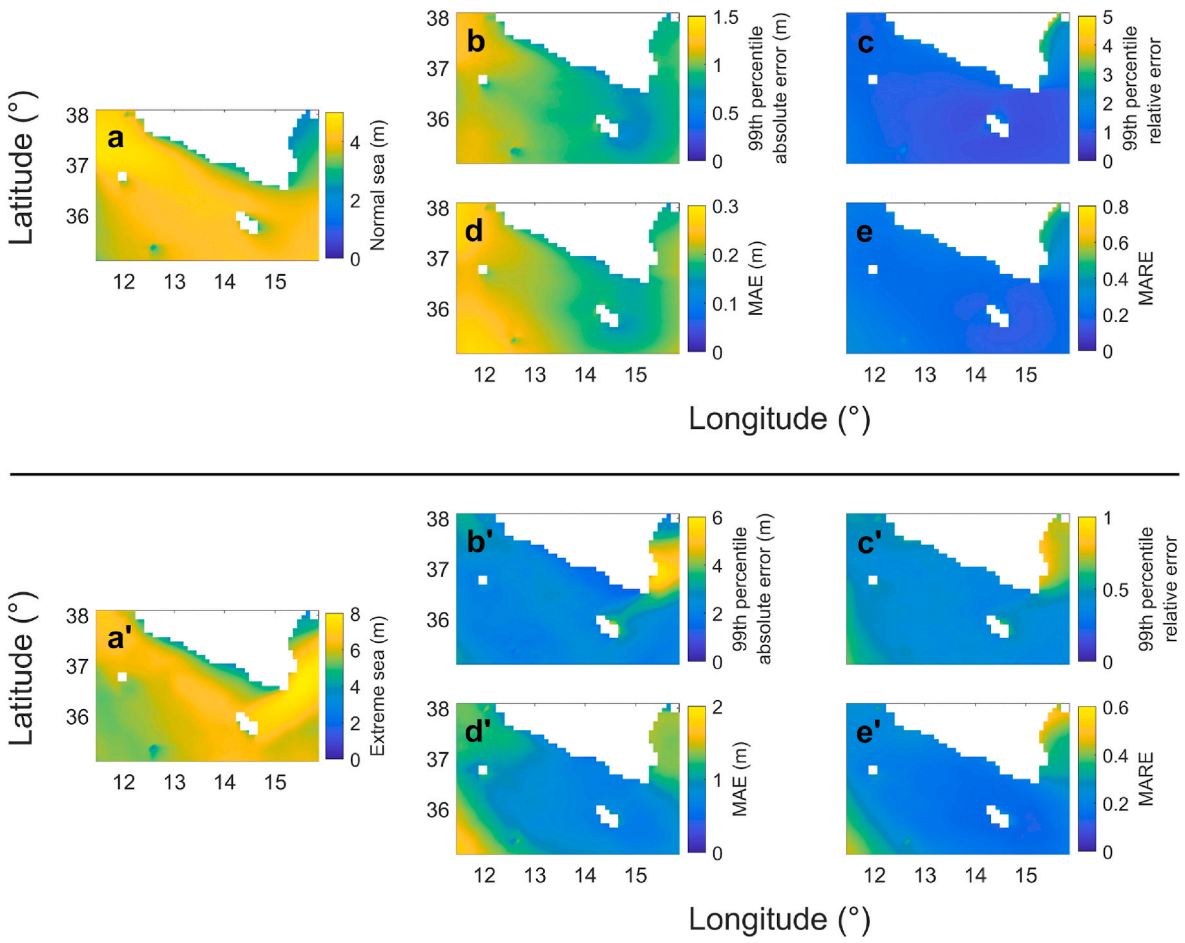


Fig. 10. Maps of errors for sea normal and extreme conditions. They were obtained through the difference between the observed and the predicted values. In this case, the RF algorithm is used to build the final model. a) Space distribution of the 99th percentile of the SWH calculated in the testing period. b) 99th percentile of the absolute errors for sea normal condition. c) 99th percentile of the relative errors for sea normal condition. d) MAE for sea normal conditions. e) MARE for sea normal conditions. a') Space distribution of the 99.99th percentile of the SWH calculated in the testing period. b') 99th percentile of the absolute errors for extreme events. c') 99th percentile of the relative errors for extreme events. d') MAE for extreme events. e') MARE for extreme events.

especially in terms of R^2 and relative errors (Figs. 11 and 12). This can be related to some factors, such as: (*i*) RF's performance is less affected by parameter selection (e.g., Li et al., 2011; Kuhn and Johnson, 2013); (*ii*) the RF algorithm is less sensitive to outliers and noise (e.g., Breiman, 2001); (*iii*) RF learner is able to deal with input and output data characterized by a non-linear relationship, such as in the case of the microseism amplitude and the significant wave height (e.g., Essen et al., 2003; Craig et al., 2016); (*iv*), RF algorithm uses an ensemble of decision trees avoiding as better as possible the overfitting phenomena (e.g., Li et al., 2011).

Not all the regions of the Sicily Channel show the same level of prediction accuracy (Fig. 11 and Fig. A4). The results may be affected by the different distances between the seismic stations and the points in which the significant wave height is measured. Indeed, the ML model can provide accurate predictions up to 200 km from the coastlines, as shown in Figure A4. As SPSM showed the highest amplitude values among the microseism frequency bands (Figs. 3, 4 and 5) and its sources are located in the waters nearby the Sicilian coastlines (Figs. 6 and 7), microseism mostly contains information for the sea state reconstruction about areas located close to the seismic stations,as described in the literature concerning SPSM (e.g., Bromirski et al., 2005; Chen et al., 2011; Gualtieri et al., 2015; Moschella et al., 2020; Cannata et al., 2020). However, further regions from coastline may contribute to the prediction of sea state, where SPSM may be generated in deeper waters (e.g., Beucler et al., 2015; Becker et al., 2020). Results may be affected by the spatial distribution of the SWH values especially in areas very close to the coastline, such as the Ionian nearshore. Indeed, in this area (Fig. 10a), the average value of SWH is about 0.49 m in spite of the higher values observed in the same region but at greater distances from the coastline (average SWH of about 0.79). This difference, in terms of average SWH, may reduce the level of prediction accuracy in the Ionian Sea, as shown in the spatial distribution of the relative errors (Fig. 10c,e, c',e') and R^2 (Fig. 11 and Fig. A4). Indeed, considering the minimum distance between the coastlines (Figure A4b) or the seismic stations (Figure A4c) and the sea grid cell, small values of R2 can be recognized up to ~40 km from the Ionian coastline. Coastal areas are often characterised by complex wave interactions due to bathymetry, topography, and nearshore processes (e.g., Holman, 1995; Heege et al., 2016; Davidson-Arnott et al., 2019; Gaeta et al., 2020). Therefore, these dynamics may introduce challenges in accurately predicting SWH values, particularly near the coastline, and may require the use of local meteorological conditions and measured wave parameters from a local buoy to improve the wave prediction (e.g., Callens et al., 2020; Cutroneo et al., 2021).

Considering the comparison between the observed and the predicted significant wave height data (Fig. 13), RF algorithm provides reliable prediction of the output data for new data entries. Indeed, it may slightly overestimate the predicted values under normal conditions of the sea state (Fig. 13d,e). However, it may underestimate the predicted values for extreme conditions of the sea (Fig. 13d,e), as observed during the evaluation phase (Fig. 10). This reduced ability to track extreme wave cases well may be due to the imbalanced training data set. Indeed, the

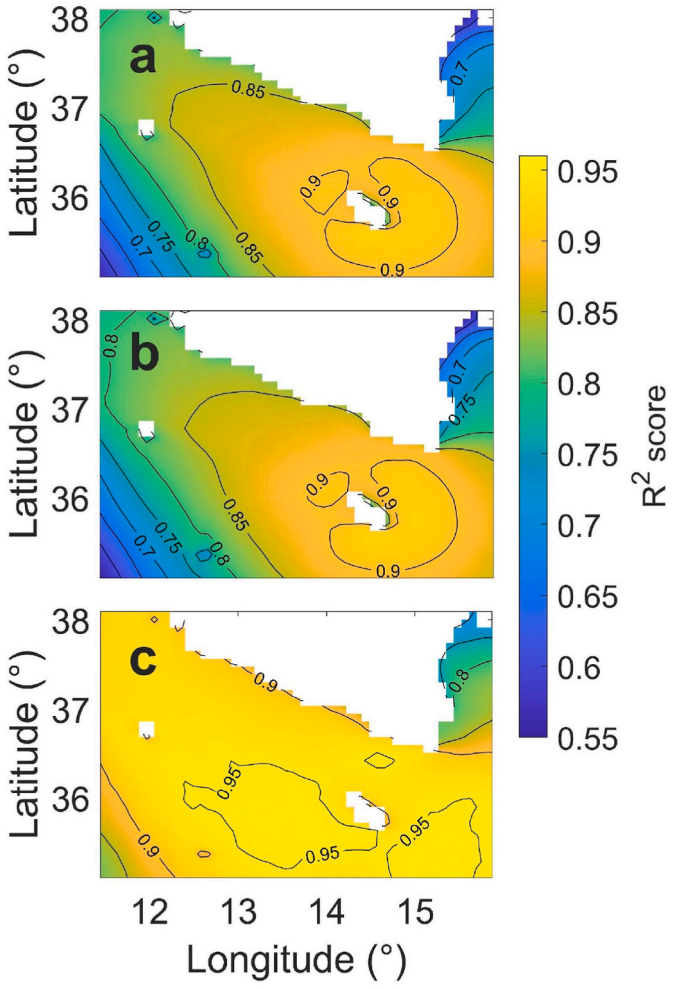


Fig. 11. Map of the coefficient of determination (R^2). It was produced using k-fold cross-validation and RF algorithm. a) Overall R^2 computed between the predicted and the observed data. b) R^2 computed for sea normal condition data. c) R^2 computed for sea extreme condition data.

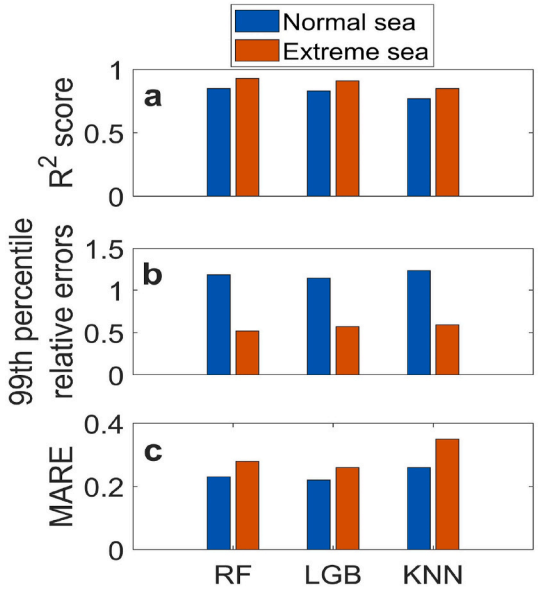


Fig. 12. Comparison of the model performances. They were estimated by using k-fold cross-validation for sea normal and extreme conditions. a) Average coefficient of determination (R^2) computed between predicted and observed values. b) Average of the 99th percentile of the relative errors. c) Average MARE.

extreme cases, as defined here, represent only 1% of the data set (Supplementary materials, Figure A3). In addition, the underestimation of the predicted values may be related to the lack of meteorological parameters in the predictive model, as shown in Cutroneo et al., (2021). Obtaining an accurate prediction of such phenomena represents a big challenge and has become an active contemporary topic in applied mathematics (e.g., Viotti and Dias, 2014; Zhang et al., 2022) and in deep neural networks (e.g., Krawczyk, 2016; Liu et al., 2016; Lagerquist et al., 2019; Xiao et al., 2019). Therefore, for future developments, we do not exclude the use of further machine learning strategies to improve the ability to track extreme cases.

The results obtained through ML techniques may be potential to improve existing marine monitoring systems, which employ various instruments and technologies to capture and analyze key oceanographic parameters. Our proposed ML model based on seismic signals may offer several advantages. Indeed, seismic networks provide continuous recording of seismic signals, and have a much higher temporal resolution than radar altimeters and synthetic aperture radar (e.g., Moreira et al., 2013; Orasi et al., 2018; Quartly et al., 2021). Seismic stations have also lower costs of installation and maintenance and a better spatial coverage with respect to most of the instruments routinely used to monitor the sea waves through situ measurements, such as wave buoys (e.g., Fu and Cazenave, 2000; Holthuijsen, 2010). In contrast, the development of a reliable ML model based on seismic signals may present some challenges. As shown in this work, a considerable amount of data may be required for training an appropriate ML technique over a wide range of sea conditions.

The integration of seismic data with existing marine monitoring systems may present a valuable opportunity to enhance sea state prediction and its implications across multiple domains. In particular, the incorporation of seismic data can contribute to more accurate wave resource assessments (e.g., Hashemi and Neill, 2014) and provide valuable insights for coastal zone planning, protection, and management, including the evaluation of energy resources (Wandres et al., 2017). Furthermore, it may play a crucial role in ensuring maritime safety, especially in areas where the interplay between strong currents and wind-generated waves is prominent (e.g., Ardhuin et al., 2017). Large-scale oceanic currents have the potential to generate hazardous sea states that pose substantial risks to navigation, people, infrastructures, and buildings (e.g., Diakakis et al., 2023). Therefore, future research should focus on refining the model's accuracy, addressing technical integration challenges, and conducting comprehensive validation studies to establish the reliability and practical utility of the microseism in marine monitoring applications.

5. Conclusions

The results of analyses described in this work provide information necessary to develop a monitoring system of the sea state, in terms of significant wave height, based on microseism. We retrieved the spatial and temporal features of the microseism recorded along the Sicilian coastlines. By using array analysis, it was possible to locate microseism sources in the Mediterranean Sea. We demonstrated how it is possible to predict sea state data for the Sicily Channel by using ML methods and microseism. For these purposes, we used the microseism recorded between 2018 and 2021 by 14 seismic stations, located along the Sicilian coastlines, and the sea state data provided by CMEMS.

These are the main results: (*i*) by correlation analysis, we showed how the SPSM recorded by the considered stations was well correlated with the sea state in the Sicilian seas, reaching the highest correlation values at frequency 0.2–0.5 Hz; (*ii*) in the SPSM frequency band, the correlation coefficient reached the highest values for distances up to 400 km, showing how the SPSM was mostly generated by sources

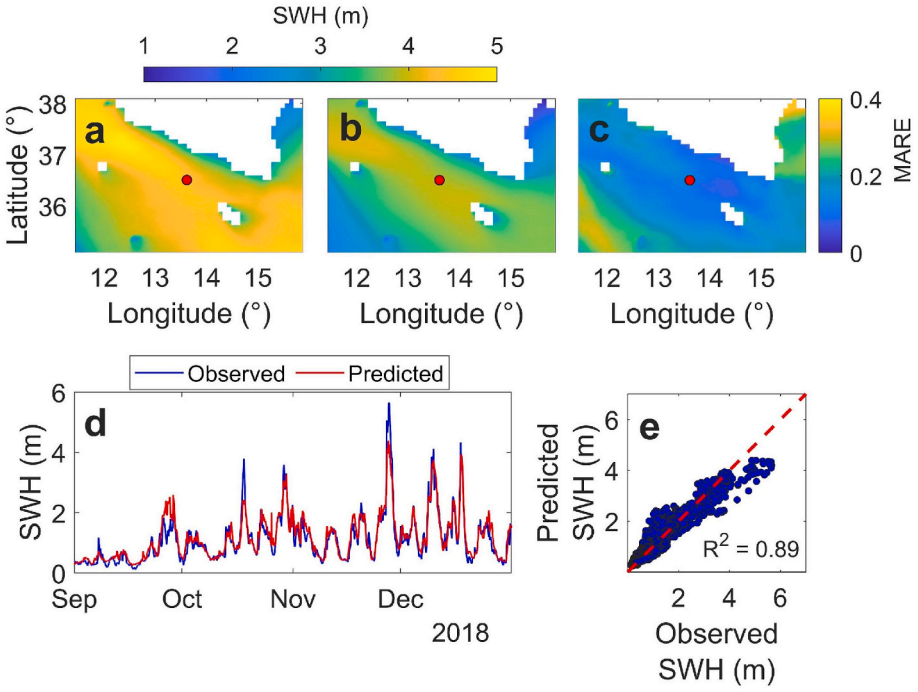


Fig. 13. Comparison between the observed and the predicted significant wave height values for the period September-December 2018. a) Space distribution of the 99th percentile of the observed SWH calculated in the validation period. b) Space distribution of the 99th percentile of the predicted SWH calculated in the validation period. c) MAE related to the relative errors for the values shown in (a) and (b). d) Observed (blue line) and predicted (red line) significant wave height time series referring to the red marks (Longitude 13.61°E, Latitude 35.50°N) shown in (a), (b), and (c). e) Cross plot showing the observed versus the predicted significant wave heights shown in (d). The red dashed line in (e) is the $y = x$ line. The value of the determination coefficient (R^2) is also reported in the bottom right corner of the diagram.

located close to the stations; (*iii*) space distribution amplitude and source position of microseism were affected by the conditions of the seas surrounding Sicily in terms of significant wave height; (*iv*) Random Forest is the best method used to build the predictive model, with a value of R^2 equal to 0.89 and mean prediction error of about 0.21 ± 0.23 m; (*v*) Random Forest is able to provide accurate results for regions located at maximum distance of about 200 km from the coastlines and from the seismic stations; (*vi*) ML methods are useful to develop a system providing a near real-time prediction of the sea state; (*vii*) our model may have the potential to augment existing marine monitoring systems.

Author contributions

V.M., F.C and A.C. designed the research; V.M., S.S. and F.C. performed the machine learning analysis; V.M., S.A., A.M.B., A.C., G.D. and S.D. performed the seismic analysis; V.M., A.M.B. and S.S. wrote the paper; V.M., S.A., G.L. and D.C. dealt with the new seismic installation; G.C. lead the main project funding this research and helped to interpret the sea data; all the authors discussed the results and edited the paper under the supervision of F.C.

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 are open-access and available on E.U. Copernicus Marine Service Information, United States Geological Survey, and European Data Archive databases.

Acknowledgments

The authors would like to thank Prof. Daniel Ames for his comments on preliminary versions of this manuscript, as well as the two anonymous reviewers of this paper, whose comments have helped to improve its quality significantly. The authors thank the i-waveNET "Implementation of an innovative system for monitoring the state of the sea in climate change scenarios" project, funded by the Interreg Italia-Malta Programme (https://iwavenet.eu/; notice February 2019 Axis 3; project code C2-3.2-106). A.M.B. thanks the PON "Ricerca e Innovazione 2014–2020 Azione IV.5 - Dottorati su tematiche green". This study has been conducted using E.U. Copernicus Marine Service Information (https://doi.org/10.25423/cmcc/medsea_multiyear_wav_006_012, last access May 2023) and an earthquake catalogue from the United States Geological Survey (USGS; https://earthquake.usgs.gov/fdsnws/event, last access May 2023).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2023.105781.

References

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Statistician 46, 175–185. https://doi.org/10.1080/00031305.1992.10475879.

Anthony, R.E., Ringler, A.T., Wilson, D.C., Bahavar, M., Koper, K.D., 2020. How processing methodologies can distort and bias power spectral density estimates of seismic background noise. Seismol Res. Lett. 91 (3), 1694–1706. https://doi.org/10.1785/0220190212.

Ardhuin, F., Balanche, A., Stutzmann, E., Obrebski, M., 2012. From seismic noise to ocean wave parameters: general methods and validation. J. Geophys. Res. 117, C05002 https://doi.org/10.1029/2011JC007449.

Ardhuin, F., Gualtieri, L., Stutzmann, E., 2015. How ocean waves rock the Earth: two mechanisms explain microseisms with periods 3 to 300 s. Geophys. Res. Lett. 42 (3), 765–772. https://doi.org/10.1002/2014GL062782.

Ardhuin, F., Gille, S.T., Menemenlis, D., Rocha, C.B., Rascle, N., Chapron, B., Gula, J., Molemaker, J., 2017. Small-scale open ocean currents have large effects on wind wave heights. J. Geophys. Res. Oceans. 122, 4500–4517. https://doi.org/10.1002/2016JC012413.

Asten, M.W., Henstridge, J.D., 1984. Array estimators and the use of microseism for reconnaissance of sedimentary basins. Geophysics 49, 1828–1837. https://doi.org/10.1190/1.1441596.

Aster, R.C., Scott, J., 1993. Comprehensive characterization of waveform similarity in microearthquake data sets. Bull. Seismol. Soc. Am. 83, 1307–1314. https://doi.org/10.1785/BSSA0830041307.

Aster, R.C., McNamara, D.E., Bromirski, P.D., 2008. Multidecadal climate-induced variability in microseisms. Seismol Res. Lett. 79, 194–202. https://doi.org/10.1785/gssrl.79.2.194.

Bartlett, M., 1948. Smoothing periodograms from time-series with continuous spectra. Nature 161, 686–687. https://doi.org/10.1038/161686a0.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. 36, 105–139. https://doi.org/10.1023/A:1007515423169.

Beucler, É., Mocquet, A., Schimmel, M., Chevrot, S., Quillard, O., Vergne, J., Sylvander, M., 2015. Observation of deep water microseisms in the North Atlantic Ocean using tide modulations. Geophys. Res. Lett. 42, 316–322. https://doi.org/10.1002/2014GL062347.

Becker, D., Cristiano, L., Peikert, J., Kruse, T., Dethof, F., Hadziioannou, C., Meier, T., 2020. Temporal modulation of the local microseism in the North Sea. J. Geophys. Res. Solid Earth 125 (10), e2020JB019770. https://doi.org/10.1029/2020JB019770.

Borzì, A.M., Minio, V., Cannavò, F., Cavallaro, A., D'Amico, S., Gauci, A., De Plaen, R., Lecocq, T., Nardone, G., Orasi, A., Picone, M., Cannata, A., 2022. Monitoring extreme meteo-marine events in the Mediterranean area using the microseism (Medicane Apollo case study). Sci. Rep. 12 (1), 21363 https://doi.org/10.1038/s41598-022-25395-9.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. Ser. A Stat. Soc. 26 (2), 211–252. http://www.jstor.org/stable/2984418.

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

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

Bromirski, P.D., 2001. Vibrations from the "perfect storm". G-cubed 2 (7), 1030. https://doi.org/10.1029/2000GC000119.

Bromirski, P.D., Duennebier, F.K., Stephen, R.A., 2005. Mid-ocean microseisms. G-cubed 6 (4), Q04009. https://doi.org/10.1029/2004GC000768.

Bromirski, P.D., Stephen, R.A., Gerstoft, P., 2013. Are deep-ocean-generated surface-wave microseisms observed on land? J. Geophys. Res. Solid Earth 118, 3610–3629. https://doi.org/10.1002/jgrb.50268.

Brooks, L.A., Townend, J., Gerstoft, P., Bannister, S., Carter, L., 2009. Fundamental and higher-mode Rayleigh wave characteristics of ambient seismic noise in New Zealand. Geophys. Res. Lett. 36 (23), L23303 https://doi.org/10.1029/2009GL040434.

Callens, A., Morichon, D., Abadie, S., Delpey, M., Liquet, B., 2020. Using Random forest and Gradient boosting trees to improve wave forecast at a specific location. Appl. Ocean Res. 104, 102339 https://doi.org/10.1016/j.apor.2020.102339.

Cannata, A., Cannavò, F., Moschella, S., Gresta, S., Spina, L., 2019. Exploring the link between microseism and sea ice in Antarctica by using machine learning. Sci. Rep. 9 (1), 13050 https://doi.org/10.1038/s41598-019-49586-z.

Cannata, A., Cannavò, F., Moschella, S., Di Grazia, G., Nardone, G., Orasi, A., Picone, M., Ferla, M., Gresta, S., 2020. Unravelling the relationship between microseisms and spatial distribution of sea wave height by statistical and machine learning approaches. Rem. Sens. 12 (5), 761. https://doi.org/10.3390/rs12050761.

Capon, J., 1969. High resolution frequency-wavenumber spectrum analysis. Proc. IEEE 57, 1408–1418. https://doi.org/10.1109/PROC.1969.7278.

Cessaro, R.K., 1994. Sources of primary and secondary microseisms. Bull. Seismol. Soc. Am. 84 (1), 142–148. https://doi.org/10.1785/BSSA0840010142.

Chen, Y.-N., Gung, Y., You, S.-H., Hung, S.-H., Chiao, L.-Y., Huang, T.-Y., Chen, Y.-L., Liang, W.-T., Jan, S., 2011. Characteristics of short period secondary microseisms (SPSM) in Taiwan: the influence of shallow ocean strait on SPSM. Geophys. Res. Lett. 38 (4), L04305 https://doi.org/10.1029/2010GL046290.

Chen, J., Pillai, A.C., Johanning, L., Ashton, I., 2021. Using machine learning to derive spatial wave data: a case study for a marine energy site. Environ. Model. Software 142, 105066. https://doi.org/10.1016/j.envsoft.2021.105066.

Chevrot, S., Sylvander, M., Benahmed, S., Ponsolles, C., Lefevre, J.M., Paradis, D., 2007. Source locations of secondary microseisms in western Europe: evidence for both coastal and pelagic sources. J. Geophys. Res. 112, B11301 https://doi.org/10.1029/2007JB005059.

Cook, J.A., Ranstam, J., 2016. Overfitting. Br. J. Surg. 103 (13), 1814. https://doi.org/10.1002/bjs.10244.

Craig, D., Bean, C., Lokmer, I., Möllhoff, M., 2016. Correlation of wavefield separated ocean-generated microseisms with North Atlantic Source regions. Bull. Seismol. Soc. Am. 106, 1002–1010. https://doi.org/10.1785/0120150181.

Cutroneo, L., Ferretti, G., Barani, S., Scafidi, D., De Leo, F., Besio, G., Capello, M., 2021. Near real-time monitoring of significant sea wave height through microseism recordings: analysis of an exceptional sea storm event. J. Mar. Sci. Eng. 9 (3), 319. https://doi.org/10.3390/jmse9030319.

Davidson-Arnott, R., Bauer, B., Houser, C., 2019. Introduction to Coastal Processes and Geomorphology, second ed. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/9781108546126.

De Caro, M., Monna, S., Frugoni, F., Beranzoli, L., Favali, P., 2014. Seafloor seismic noise at central eastern mediterranean sites. Seismol Res. Lett. 85, 1019–1033. https://doi.org/10.1785/0220130203.

Diakakis, M., Mavroulis, S., Filis, C., Lozios, S., Vassilakis, E., Naoum, G., Soukis, K., Konsolaki, A., Kotsi, E., Theodorakatou, D., Skourtsos, E., Kranis, H., Gogou, M., Spyrou, N.I., Katsetsiadou, K.-N., Lekkas, E., 2023. Impacts of Medicanes on geomorphology and infrastructure in the eastern mediterranean, the case of Medicane ianos and the ionian islands in western Greece. Water 15 (6), 1026. https://doi.org/10.3390/w15061026.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, third ed. John Wiley and Sons. ISBN 9780471170822.

Essen, H.H., Klussmann, J., Herber, R., Grevemeyer, I., 1999. Does microseisms in Hamburg (Germany) reflect the wave climate in the North Atlantic? Dtsch. Hydrogr. Zeitschrift. 51 (1), 33–45. https://doi.org/10.1007/BF02763955.

Essen, H.-H., Krüger, F., Dahm, T., Grevemeyer, I., 2003. On the generation of secondary microseisms observed in northern and central Europe. J. Geophys. Res. Space Phys. 108, 2506. https://doi.org/10.1029/2002JB002338.

Faranda, D., Bourdin, S., Ginesta, M., Krouma, M., Messori, G., Noyelle, R., Pons, F., Yiou, P., Messori, G., 2022. A climate-change attribution retrospective of some impactful weather extremes of 2021. Weather Clim. Dynam. 3 (4), 1311–1340. https://doi.org/10.5194/wcd-3-1311-2022.

Ferretti, G., Barani, S., Scafidi, D., Capello, M., Cutroneo, L., Vagge, G., Besio, G., 2018. Near real-time monitoring of significant sea wave height through microseism recordings: an application in the Ligurian Sea (Italy). Ocean Coast Manag. 165, 185–194. https://doi.org/10.1016/j.ocecoaman.2018.08.023.

Fu, L.L., Cazenave, A., 2000. Satellite Altimetry and Earth Sciences: A Handbook of Techniques and Applications, vol. 69. Academic Press, New York.

García, V., Sánchez, J., Marqués, A., 2019. Synergetic application of multi-criteria decision-making models to credit granting decision problems. Appl. Sci. 9, 5052. https://doi.org/10.3390/app9235052.

Gaeta, M.G., Samaras, A.G., Archetti, R., 2020. Numerical investigation of thermal discharge to coastal areas: a case study in South Italy. Environ. Model. Software 124, 104596. https://doi.org/10.1016/j.envsoft.2019.104596.

Gualtieri, L., Stutzmann, E., Capdeville, Y., Farra, V., Mangeney, A., Morelli, A., 2015. On the shaping factors of the secondary microseismic wavefield. J. Geophys. Res. 120, 6241–6262. https://doi.org/10.1002/2015jb012157.

Gualtieri, L., Stutzmann, E., Juretzek, C., Hadziioannou, C., Ardhuin, F., 2019. Global scale analysis and modelling of primary microseisms. Geophys. J. Int. 218, 560–572. https://doi.org/10.1093/gji/ggz161.

Han, J., Pei, J., Tong, H., 2022. Data Mining: Concepts and Techniques. Morgan Kaufmann.

Hashemi, M.R., Neill, S.P., 2014. The role of tides in shelf-scale simulations of the wave energy resource. Renew. Energy 69, 300–310. https://doi.org/10.1016/j.renene.2014.03.052.

Hasselmann, K., 1963. A statistical analysis of the generation of microseisms. Rev. Geophys. 1 (2), 177–210. https://doi.org/10.1029/RG001i002p00177.

Haubrich, R.A., McCamy, K., 1969. Microseisms: coastal and pelagic sources. Rev. Geophys. 7 (3), 539–571. https://doi.org/10.2183/pjab.93.026.

Havskov, J., Alguacil, G., 2016. Seismic arrays. In: Instrumentation in Earthquake Seismology. Springer, Berlin, Germany, pp. 11–70. https://doi.org/10.1007/978-3-319-21314-9.

Heege, T., Bergin, M., Hartmann, K., Schenk, K., 2016. Satellite services for coastal applications. In: Wright, D.J. (Ed.), Ocean Solutions, Earth Solutions, second ed. Esri, Redlands, CA, USA, pp. 357–368. https://doi.org/10.17128/9781589484603_18.

Hinton, G.E., 2012. A practical guide to training restricted Boltzmann machines. In: Montavon, G., Orr, G.B., Müller, K.R. (Eds.), Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science. Springer, Berlin, Germany, pp. 599–619. https://doi.org/10.1007/978-3-642-35289-8_32.

Ho, T.K., 1995. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE, Montreal, QC, Canada, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994.

Holman, R., 1995. Nearshore processes. Rev. Geophys. 33 (S2), 1237–1247. https://doi.org/10.1029/95RG00297.

Holthuijsen, L.H., 2010. Waves in Oceanic and Coastal Waters. Cambridge university press., Cambridge.

Ivan, R., Oliver, J., Mia, F., Renato, F., 2018. A study of GPS positioning error associated with tropospheric delay during Numa Mediterranean cyclone. Int. J. Traffic aTransp. Eng. 8, 282–293. https://doi.org/10.7708/ijtte.2018.8(3).03.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. In: Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates Inc., San Mateo, CA (USA).

Kenney, J.F., Keeping, E.S., 1962. Root mean square. In: Mathematics of Statistics, third ed. Van Nostrand Company, Princeton, New Jersey, pp. 59–60.

Kimman, W.P., Campman, X., Trampert, J., 2012. Characteristics of seismic noise: fundamental and higher mode energy observed in the Northeast of The Netherlands. Bull. Seismol. Soc. Am. 102, 1388–1399. https://doi.org/10.1785/0120110069.

Kong, Q., Trugman, D.T., Ross, Z.E., Bianco, M.J., Meade, B.J., Gerstoft, P., 2018. Machine learning in seismology: turning data into insights. Seismol Res. Lett. 90, 3–14. https://doi.org/10.1785/0220180259.

Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Prog. Artif. Intell. 5, 221–232. https://doi.org/10.1007/s13748-016-0094-0.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling, first ed. Springer, New York, NY, USA. https://doi.org/10.1007/978-1-4614-6849-3.

Lagerquist, R., McGovern, A., Gagne, D.J., 2019. Deep learning for spatially explicit prediction of synoptic-scale fronts. Weather Forecast. 34 (4), 1137–1160. https://doi.org/10.1175/WAF-D-18-0183.1.

Lagouvardos, K., Karagiannidis, A., Dafis, S., Kalimeris, A., Kotroni, V., 2022. Ianos—a hurricane in the mediterranean. Bull. Am. Meteorol. Soc. 103, 1621–1636. https://doi.org/10.1175/BAMS-D-20-0274.1.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., Bengio, Y., 2007. An empirical evaluation of deep architectures on problems with many factors of variation. In: Proceedings of the 24th International Conference on Machine Learning, pp. 473–480. https://doi.org/10.1145/1273496.1273556.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Software 26, 1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004.

Liu, Y., Racah, E., Prabhat, Correa, J., Khosrowshahi, A., Lavers, D.A., Kunkel, K.E., Wehner, M.F., Collins, W.D., 2016. Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets, vol. 1605, 01156. https://doi.org/10.48550/arXiv.1605.01156. ArXiv. abs.

Longuet-Higgins, M.S., 1950. A theory of the origin of microseisms. Philos. Trans. Royal Soc. A. 243 (857), 1–35. https://doi.org/10.1098/rsta.1950.0012.

Mayfield, H.J., Smith, C., Gallagher, M., Hockings, M., 2020. Considerations for selecting a machine learning technique for predicting deforestation. Environ. Model. Software 131, 104741. https://doi.org/10.1016/j.envsoft.2020.104741.

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., Papathanassiou, K.P., 2013. A tutorial on synthetic aperture radar. IEEE Trans. Geosci. Rem. Sens. 1, 6–43. https://doi.org/10.1109/MGRS.2013.2248301.

McKinney, W., 2010. Data structures for statistical computing in Python. In: van der Walt, S., Millman, J. (Eds.), Proceedings of the 9th Python in Science Conference, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Miglietta, M.M., 2019. Mediterranean tropical-like Cyclones (Medicanes). Atmosphere 10, 206. https://doi.org/10.3390/atmos10040206.

Moschella, S., Cannata, A., Cannavò, F., Di Grazia, G., Nardone, G., Orasi, A., Picone, M., Ferla, M., Gresta, S., 2020. Insights into microseism sources by array and machine learning techniques: ionian and Tyrrhenian Sea case of study. Front. Earth Sci. 8, 114. https://doi.org/10.3389/feart.2020.00114.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorob. 7, 21. https://doi.org/10.3389/fnbot.2013.00021.

Orasi, A., Picone, M., Drago, A., Capodici, F., Gauci, A., Nardone, G., Inghilesi, R., Azzopardi, J., Galea, A., Ciraolo, G., Musulin, J.S., Alonso-Martirena, A., 2018. HF radar for wind waves measurements in the Malta-Sicily Channel. Measurement 128, 446–454. https://doi.org/10.1016/j.measurement.2018.06.060.

Patanè, D., Privitera, E., Ferrucci, F., Gresta, S., 1994. Seismic activity leading to the 1991-1993 eruption of Mt. Etna and its tectonic implications. Acta Vulcanol. 4, 47–55.

Portmann, R., González-Alemán, J., Sprenger, M., Wernli, H., 2020. How an uncertain short-wave perturbation on the North Atlantic wave guide affects the forecast of an intense Mediterranean cyclone (Medicane Zorbas). Weather Clim. Dyn. 1, 597615 https://doi.org/10.5194/wcd-1-597-2020.

Quartly, G.D., Chen, G., Nencioli, F., Morrow, R., Picot, N., 2021. An overview of requirements, procedures and current advances in the calibration/validation of radar altimeters. Rem. Sens. 13 (1), 125. https://doi.org/10.3390/rs13010125.

Reguero, B.G., Losada, I.J., Méndez, F.J., 2019. A recent increase in global wave power as a consequence of oceanic warming. Nat. Commun. 10 (1), 205. https://doi.org/10.1038/s41467-018-08066-0.

Rivet, D., Campillo, M., Sanchez-Sesma, F., Shapiro, N.M., Singh, S.K., 2015. Identification of surface wave higher modes using a methodology based on seismic noise and coda waves. Geophys. J. Int. 203, 856–868. https://doi.org/10.1093/gji/ggv339.

Rost, S., Thomas, C., 2002. Array seismology: method and application. Rev. Geophys. 40 (3), 1008. https://doi.org/10.1029/2000RG000100.

Steele, K.E., Mettlach, T., 1993. NDBC wave data – current and planned. In: Ocean Wave Measurement and Analysis, pp. 198–207.

Stutzmann, E., Schimmel, M., Patau, G., Maggi, A., 2009. Global climate imprint on seismic noise. G-cubed 10, Q11004. https://doi.org/10.1029/2009GC002619.

Tanimoto, T., Hadziioannou, C., Igel, H., Wasserman, J., Schreiber, U., Gebauer, A., 2015. Estimate of Rayleigh-to-Love wave ratio in the secondary microseism by colocated ring laser and seismograph. Geophys. Res. Lett. 42 (8), 2650–2655. https://doi.org/10.1016/b978-044452748-6.00075-4.

Trnkoczy, A., 2012. Understanding and parameter setting of STA/LTA trigger algorithm. In: Bormann, P. (Ed.), New Manual of Seismological Observatory Practice 2 (NMSOP-2). Potsdam, pp. 1–20. https://doi.org/10.2312/GFZ.NMSOP-2_IS_8.1.

Viotti, C., Dias, F., 2014. Extreme waves induced by strong depth transitions: fully nonlinear results. Phys. Fluids 26 (5), 051705. https://doi.org/10.1063/1.4880659.

Von Storch, H., Emeis, K., Meinke, I., Kannen, A., Matthias, V., Ratter, B., Stanev, E., Weisse, R., Wirtz, K., 2015. Making coastal research useful – cases from practice. Oceanologia 57 (1), 3–16. https://doi.org/10.1016/j.oceano.2014.09.001.

Wandres, M., Wijeratne, E.M.S., Cosoli, S., Pattiaratchi, C., 2017. The effect of the Leeuwin Current on offshore surface gravity waves in southwest western Australia. J. Geophys. Res. Oceans. 122, 9047–9067. https://doi.org/10.1002/2017JC013006.

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82. https://doi.org/10.3354/cr030079.

Xiao, C., Chen, N., Hu, C., Wang, K., Xu, Z., Cai, Y., Xu, L., Chen, Z., Gong, J., 2019. A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data. Environ. Model. Software 120, 104502. https://doi.org/10.1016/j.envsoft.2019.104502.

Zhang, M., Yang, X., Cleverly, J., Huete, A., Zhang, H., Yu, Q., 2022. Heat wave tracker: a multi-method, multi-source heat wave measurement toolkit based on Google Earth Engine. Environ. Model. Software 147, 105255. https://doi.org/10.1016/j.envsoft.2021.105255.