POLITECNICO DI TORINO

Dipartimento di Ingegneria Energetica
Corso di Laurea Magistrale in Ingegneria Energetica e Nucleare

Tesi di Laurea Magistrale

From Wind to Wave energy resource: forecasting methods analysis

Relatore:

Prof. Giovanni Bracco

Correlatore:

Ing. Fenu Beatrice

Ing. Cervelli Giulia

Candidato:

Palmieri Marco

Matr. S288432

Anno accademico 2021-2022

Acknowledgments

I would like to thank Professor Giovanni Bracco for the opportunity he gave me to work on such an important topic for the future. A big thanks to Beatrice Fenu and Giulia Cervelli for excellently guiding me in this work.

A special thanks to my parents who supported and encouraged me from the first to the last day.

Summary

The need of a more sustainable future led humanity to investigate new renewable energy sources, such as waves. The greatest advantage of the waves over other alternative energy sources is that it is easily predictable through the study of winds. Adequate wave predictions are crucial for the proper mapping and design of the wave farms that will contribute to renewable energy generation in the future.
The thesis aims to provide a general overview of wave prediction methods and to analyze some of them. Three different methods are built and compared: an empirical method, that is the Sverdrup-Munk-Bretschneider (SMB) method, a multiple regression method and an artificial intelligence method, specifically a machine learning one called Artificial Neural Network (ANN).
The three methods are applied in a case study in which the selected site is the Island of Pantelleria. In the first phase, the models will be built using data obtained from ERA5, that is the fifth generation ECMWF atmospheric reanalysis of the global climate. Subsequently, data acquired from a buoy located on the site of interest will be implemented and the results will be compared to the previous ones. The goal is to find the most suitable prediction method for the selected site.

Table of contents

List of figures

List of tables

List of Abbreviations

LCOE	Levelized Cost Of Electricity
RES	Renewable Energy Sources
EU	European Union
H_s	Significant wave height
T_p	Mean wave period
WEC	Wave Energy Converter
PTO	Power Take-Off
SMB	Sverdrup-Munk-Bretschneider method
U	Horizontal wind speed
F	Fetch length
FDS	Fully Developed Sea
RMSE	Root Mean Square Error
AI	Artificial Intelligence
ANN	Artificial Neural Network

Chapter 1

Introduction

1.1 Context analysis

One of the greatest that the humanity have ever faced is undoubtedly the climate changes. The global warming, that we are experiencing more every day, is the result of decades of excessive exploitation of the resources. A global economy is based and developed on a vicious cycle: low costs due to not considering the negative externalities (indirect cost of an activity to an uninvolved third party like the society or environment, the first example are the emissions) [1], the consequently growing demand and market, and thus an ever greater exploitation to follow the demand and the opportunities. This scenario, repeated for many years without concrete control measures, led to the tragic current situation.
There is no time to avoid the climate change, but we can only mitigate that with immediate actions, aiming to revolutionize an entire world based on the over-exploitation of limited resources.
One of the main global target is the use of renewable resources for the energy production, especially electrical energy, substituting the fossil ones that generate too high emissions. It is a very hard challenge considering the brief available time.
The major renewable power plant in the world are hydroelectric, wind and photovoltaic ones [2]. The continuous development of these plants aiming to increase the electrical energy production led to the grid parity for them, it means that these energy source can generate power at a Levelized Cost of Electricity (LCOE) that is

less than or equal to the price of the electricity. For that reason, all the countries are pushing toward the construction of these plants in order to increase the percentage of renewable energy production.
Despite that, the RES have several disadvantages that have to be solved in order to reach renewable energy target in the short term and the total abolition of the fossil sources in the long term [3]. For hydroelectric power plant, the main issue is the environmental impact and there are already plants built in major part of profitable sites, so the growth of this renewable sources will not be so great, especially if we are talking about the Italy. For wind and photovoltaic plants, the main issues are the power generation depending on natural resources that are uncontrollable and the large land use.
According to the report of REN21 [2], the world percentage of renewable energy up to 2020 is around the 22% and it is more or less the same percentage referring to the European Union. The European Commission stated that EU wants to accelerate the take-up of renewables to contribute and reach the goal of reducing net greenhouse gas emissions by at least 55% by 2030 [4]. It presented new 2030 climate targets, seeking to increase the current target (32%) to at least 40% renewable energy sources (RES) in the EU's overall energy mix. A huge challenge that can be achieved through policy and incentives from the governments.
In order to achieve the environmental goals, the research is oriented toward the construction of increasingly efficient and higher power plants. This implies engineering problems solving, like the construction of bigger infrastructure, the need of new materials with better performance, or to find a solution for the intermittency of natural element, for example through the installation of storage systems. In addition to that, there are also social problems that must be taken into account, first of all the environmental impact. We are performing this energy transition and world revolution precisely to limit the climate changes, and if we do not consider the environment safeguard during this process, the achievement of predefined goals will be impossible.
One of the biggest problems is the land use [3]. Indeed, the construction of new farms, especially wind and photovoltaic ones, requires a lot of land. Considering that most of the space is already taken up by farmland, buildings and infrastructure, the situation become quite tricky.

A recent study, carried out by Saunders [5], presents findings from academic research related to the land use impacts of wind and photovoltaic plants. The author concludes that developing a 100% renewable energy system would be challenging from a land use policy perspective. A power system that relies exclusively on wind, water, and solar generation will necessarily have much greater land use impacts than one that uses fuel with higher power density like fossil ones. Moreover, electrifying other sectors, the demand would increase too, requiring a larger system. There may be enough land to physically accommodate an all-renewable energy system, but the scale of the physical footprint could be daunting. The main problem is the low power density of these energy sources. Indeed, the median power density of natural gas generation is roughly 80 times than the solar generation and well over 200 times than the wind electricity.

A study conducted in the 2017 assessed the viability of a 100% solar energy system for 27 European Union (EU) member states and 13 non-EU countries (including the United States) [6]. All 40 countries together would account for 65% of global population and 90% of global gross domestic product. The results of the research are that a 100% solar energy system in the EU would require about 45% of the EU's total land area. Moreover, six countries, including Belgium, Denmark, and the United Kingdom, would not have sufficient land area to supply their final energy demand with solar power alone. While in America a fully solar energy system would require more than five times the total area occupied by all land already built. The authors conclude that the land use challenges associated with photovoltaic development are greatest in "northern latitudes with high population densities and high electricity consumption per capita". Thankfully at those latitudes the wind farms are very efficient, but the land use remains a big problem.

At the light of above, it is important to not overlook this aspect and a great solution is to move toward the offshore.

For what concern the wind farms, the offshore is already economically affordable, especially in shallow water, but research and investments are needed to reach higher efficient and energy production for the farms. Moreover, the marine energy cloud play a key role in the energy transition and the exploitation of these resources like tides and waves will be fundamental. The Europe is moving in that direction and future development plans are being considered.

1.2 Wave energy potential

Marine energy is the youngest renewable energy and it is an enormous untapped reserve with inexhaustible potential. Fully exploiting the ocean and sea power, all of the energy consumption forecast by the International Energy Agency (IEA) by 2035 would already be covered [7].
The Europe is incentivizing the development of required technologies [8]. A first general plan includes the exploitation of wave energy in the Mediterranean ares, while offshore wind and tidal farms will be placed mainly in the North Europe.
An Italian study, conducted by ISPRA (Istituto Superiore per la Protezione e la Ricerca Ambientale), estimated wave power flux per unit crest averaged over the entire 10 years simulation period in the Mediterranean [9]. The most productive area, with an average ave power above 12 kW/m, is located in the western Mediterranean, between casts of Sardinia, Balearic Islands and North Africa. Also, the region located in the Sicily channel, off the north-western and southern Sicilian coasts, is very productive with an average wave energy flux per unit crest that reaches 9 kW/m.
Italy can play an important role for the development of technologies and the growth of this new renewable source. Currently, the drawback is the highest LCOE of all the renewable energies because the wave energy technologies are in the pre-commercial prototype and demonstration stage [10]. The research is proceeding in order to make the wave farms economically available and to reach the grid parity as done for the photovoltaic and wind.
The advantages of this energy are considerable and in future it will be a fundamental part of the energy mix of the coastal areas. The wave energy is quite consistent and is shown to be much better than other sources. Waves are hardly interrupted and almost always in motion, so the electricity generation from wave energy is a more reliable source compared to wind power, since wind is not constantly blowing. Probably the biggest advantage of wave power, compared to the other alternative energy sources, is that it is easily predictable and it is quite easy to calculate the amount that it can produce. Certainly, also this source is characterized by season

variation, but the trend is more or less similar from year to year.
The energy flux that could be potentially extracted from the waves (expressed in W/m) is proportional to the square of the significant wave height H_s and to the mean wave period T_p [11]. The former is defined as the average wave height, from trough to crest, of the highest one-third of the waves, while the latter is defined as the mean of all wave periods in a time-series representing a certain sea state [12]. Consequently, predicting wave height and wave period allows to estimate indirectly the energy flux too.

1.3 Wave formation and energy converters

Waves are most commonly caused by wind [13]. Wind driven waves, or surface waves, are created by the friction between wind and surface water. The capillary waves are the first waves to form when the wind blows and they are the tiniest waves on the surface of ocean. As wind wins keeps blowing, capillary waves develop and they offer more surface area to the wind becoming longer and travel faster. Beyond a length of 1.73 cm, the waves transition from capillary waves to surface gravity waves. First we have chop waves, that are short-crested waves, and they break readily at the crest. Then, there are wind waves with increasing period until the fully developed state is reached.
Two fundamental factors to reach the fully developed sea are the constant wind blow and the fetch length long enough. The latter is defined as the length of water over which a given wind has blown without obstruction, (it will be deepened during the empirical model discussion, section 2.1).
Other types of waves are the swell waves that are not generated by distant weather systems, seiches that are standing waves in an enclosed or partially enclosed body of water, tsunamis and tidal waves.
The waves have a big amount of energy that can be converted into electrical or mechanical energy using the Wave Energy Converter (WEC). They can be classified into different categories [14]: point absorbers, terminators, attenuators, oscillating wave surge, submerged pressure differential, oscillating water column, bulge and rotating mass devices.

The power take-off (PTO) of a converter is defined as the mechanism with which the absorbed energy by the primary converter is transformed into usable electricity [15]. They can be hydraulics, that are cheaper but with less efficiency, or direct driven, that are more expensive but more efficient.
Several technologies are already developed and prototypes are being tested, while other typologies and solution are being studied. As said, this renewable energy is still under developmental stage and more years of research are needed, but it is growing fast.
Research conducted by the University of Exeter showed that accurate wave prediction combined with constrained optimal control can greatly improve the efficiency of a wave energy converter [16]. In particular, studying different control strategies for a point absorber, it was found that the energy generated could be doubled using new methods for predicting wave power. In addition, prediction is also useful to prevent WECs from being damaged by large waves, especially during winter storms where they have to be shut down.
An accurate wave forecasting results extremely important both long and short term [17]. During the planning of a farm, the estimation of the potential production is crucial in order to determine the size of the farm and consequently to manage in the best way the construction. Moreover, knowing how much energy can be exploited from these sources, it is possible to evaluate also the contributions of other renewable energy sources to reach the 100% of green energy for that area or, in general, for the country.
On the other hand, a short term forecasting, in the range of hours or days, is important not so much for the planning and construction, as for the management and energy mix optimization. Indeed, as well known, many renewables, especially wind and photovoltaic, can exhibit large fluctuations in the production and a forecast of the available resources in the range of hours will allow a better management in terms of the resource quantity and from an economic point of view too.
The goals of this work are to provide a general overview on the more diffused forecasting models and, subsequently, to analyze and compare some of them. The studied are is off Pantelleria, an island in the Strait of Sicily. The aim is to find the most accurate model for that area and to identify pros and cons of each one.

Chapter 2

State of the art of forecasting methods

The wave forecasting has been developed since 1970s. Many countries have been interested in order to improve their maritime activities, like shipping, fisheries, offshore mining, commerce, coastal engineering, construction and others. However it was not so easy to access to the available material on wave forecasting methodology. For this reason the World Meteorological Organization (WMO) published in 1988 the *Guide to Wave Analysis and Forecasting* [18]. It was an update and replacement of the *Handbook on Wave Analysis and Forecasting* published in 1976. Subsequently, the Guide was updated in 1998 (second edition) and 2018 (third edition) that is the current one. The last version takes into account the huge improvements in wave modelling over the last two decades.
All the models developed in these years have been divided into two macro-categories: empirical model and numerical model:

- The empirical formulae are not derived from an analysis of all physical processes, but rather from analysis of large datasets. They represent wave growth from known the wind field properties (as wind speed and direction, fetch and duration). Often the variables are all made dimensionless to simplify the comparison and diagram for manual wave forecasting are constructed in order to estimate the wave height knowing the input parameters.

- In the physical models, the processes affecting the energy of the waves need to

be identified. Wave energy at a given location is changed through advection (rate of energy propagated into and away from the location), the wave-energy gains from the external environment (mainly the wind) and wave-energy losses due to dissipation. In wave modelling, the usual approach is to represent these influences as a wave-energy conservation equation and then to solve it. The sources of wave energy (gains and losses) are identified as three major processes: the external gains (S_{in}), the dissipative loss (S_{ds}) and the shifting of energy within the spectrum due to weakly non-linearwave–wave interactions (S_{nl}) [18].

2.1 Empirical models

In the past, many empirical formulae for wave forecasting have been derived from big visually observed datasets. Nowadays the formulae derive from wave measurements and they are more accurate. The main variables are wind velocity, winds duration, fetch length and width, and initial wave characteristics. The objective is to forecast wave heights, periods and directions.
First of all, the wind variability is analyzed since it is the main driven factor. Generally, if the wind direction changes by 30° or less, wave heights and periods are computed as if no change has occurred and the wave direction is assumed to be aligned with the mean direction. With higher direction changes, the already formed waves are treated as swell and the new ones are estimated in relation to the new wind direction. The limit angle of 30° is a free assumption and even higher value can be examined in some cases (however it is suggested to not overcome 60°) [19]. Also when the wind speed drops below the value needed to maintain the height of existing waves, they become swell waves and the new ones are calculated for lower wind speed and combined with the existing swell waves.
Then the fetch is estimated. Two different types can be defined: geographical fetch and effective fetch [20]. The former is defined as the geographical distance between the point of interest and the nearest land along a certain direction. By doing so for all the direction included in the concerned sector, an irregularly shaped polar diagram can be constructed. But for the wave calculation, also the fetch along the

directions close to mean wind direction are important. Indeed, the wind transmits energy to the sea surface along a bundle of directions centered around a prevailing one. Therefore, the weighted average of all the geographical fetches is considered. This weighted average is called effective fetch and it is defined as:

$$F_eff,w = \frac{\sum_{\phi_i=\phi_w-\theta}^{\phi_w+\theta} F_i \cdot cos^{n+1} \cdot (\phi_i - \phi_w)}{\sum_{\phi_i=\phi_w-\theta}^{\phi_w+\theta} F_i \cdot cos^{n}\dot{(}\phi_i - \phi_w)} \tag{2.1}$$

where the symbols represent the following parameters:

$F_{eff,w}$	length of the effective fetch relative to the direction ϕ_w;
F_i	length of the geographic fetch relative to the i-th direction ϕ_i;
ϕ_w	mean direction (referring to geographic north) of possible origin of the wind responsible for the wave generation;
$\phi_w - \theta \leq \phi_i \leq \phi_w + \theta$	i-th direction (referring to geographic North) relative to a sector of 2θ considered around the direction ϕ_w;
θ	Amplitude of the sector of possible wave origin (the Seymour method is the most accredited and it refers to a value of $\theta = \pm 90°$);
n	exponential term (for the Mediterranean seas it is usual to assume a value equal to 2).

The equation is derived from the indirect wave reconstruction theory known as the S.M.B. method (Sverdrup, Munk and Bretsheneider, 1947) and its subsequent updates (Saville 1954, Seymour 1977, Smith 1991) [18].
After the wind and fetch calculation, the contribute of the swell waves is added. A swell height decays about 25% in 12 hours and 40% in 24 hours is usually considered. Additionally, swell heights are almost unaffected by opposing winds and for distant swell angular spreading usually dominates over dispersion effects. Generally, if the study is conducted in the Mediterranean seas, the swell waves contribution can be neglect because it is an enclosed basin, and the distances are not large enough to

allow swell waves to develop.
Depending on the chosen site, there may be other factors that influence the waves. For example, waves moving in the same currents direction increase in wavelength and reduce steepness, while waves moving into opposing currents reduce wavelength and steepness [18]. Also high/low stability in the lower boundary layer will reduce/enhance surface wind stress, resulting in lower/higher waves. Another important factor to consider is the temperature: lower temperatures result in higher density air at constant pressure and this effect is linear with absolute temperature. It can be significant at middle to high latitudes, where winter temperatures can often be 30°C or more colder than those in summer, and can give surface wind-stress levels higher by 10% or more for the same wind speeds in winter.
The wave forecast can be typically divided in two steps [18]: analysis, diagnosis and prognosis.

1. Analysis step: check current observed conditions and recent observed trends and compare to the model initialization.
   First, it is important to check the wind state. For example if the wind does not blow, the wave field almost certainly will be off too. Then, look for error magnitudes because it will give a guide on how to modify the model forecast. Secondly, check all wave observations. Look for other wave components from manual observations or wave spectra. With this, it might be possible to identify wave fields not resolved by other sources, and compare with the way the model partitioned the wave fields.

2. Diagnosis and prognosis step: check discrepancies in the forecast domain, finding the cause in order to give a strategy for modifying the forecast.
   For wind discrepancies, for instance, an approaching hurricane that was underforecast should imply a low bias to wave height or peak period in the model.
   For wave discrepancies not attributable to wind discrepancies, it is important to observe if waves are too weak or too strong for the winds. For weaker waves, the causes might include oceanic or tidal currents moving with the wind, reducing the wind stress on the water. For stronger waves, a possible cause might be due to non-locally generated swell moving into the forecast domain, or already there and leaving/dissipating.

Once the model has been modified to eliminate the discrepancies observations and model forecasts, it is important to monitor and repeat the cycle as necessary.

2.1.1 Sverdrup-Munk-Bretschneider (SMB) method

The best known and most proven empirical method for wave forecasting is the Sverdrup-Munk-Bretschneider (SMB) method. It is based on an energy balance and the three fundamental factors involved that you need to know are [20]:

- the wind speed U (at the conventional altitude of 10 meters at sea level);
- the fetch length F;
- the duration of the wind t.

The basic simplifying assumption is that throughout the duration t, the wind has uniform speed and direction over the entire fetch F.
For an given wind speed the wave motion is limited either by the length of the fetch (steady state) or by the duration of the wind (transient state). If F, t, and d can be ideally treated as infinite, the FDS (Fully Developed Sea) conditions are achieved: in this case, a balance is reached between the increase in energy provided by the wind and the dissipations due to friction, turbulence, and breaking. The average wave characteristics are held constant for constant wind speed. In practice, FDS conditions are rare and are achieved only in the ocean and for not very high wind speeds [20].
Given the wind speed, it is necessary to check which of the other two factors (duration, fetch) is more limiting and consider the one that causes the lowest wave height. Then the main wave characteristics (H_s and T_p) can be derived through the use of the practical graph shown in Figure 2.1.

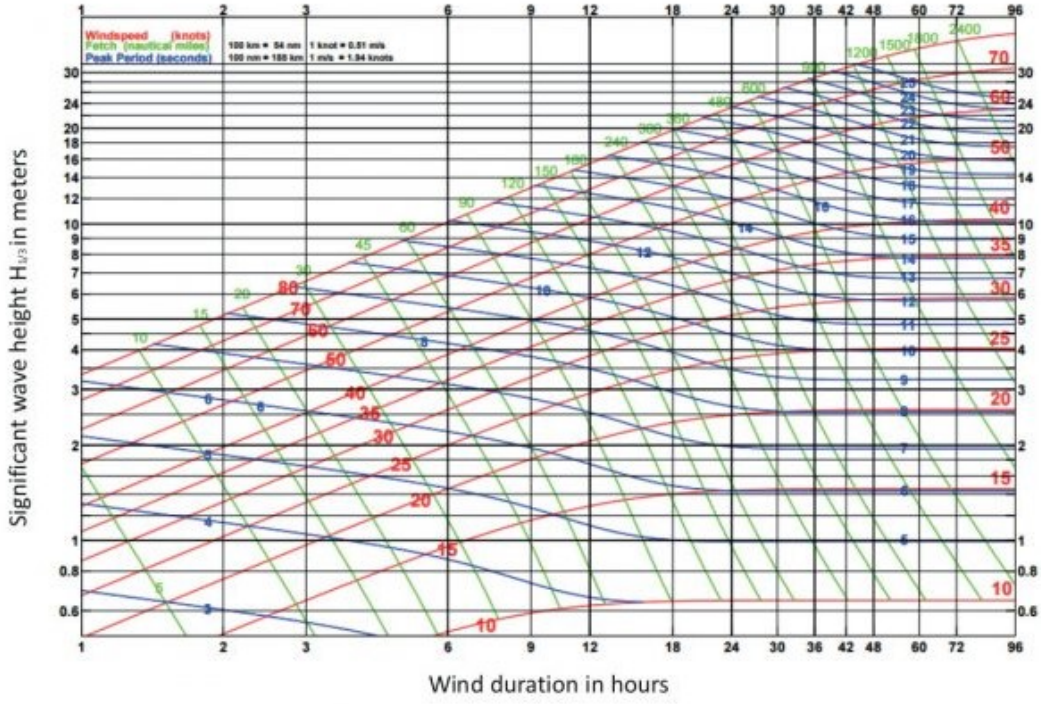


Figure 2.1: Manual wave-growth nomogram.

Source: Breugem and Holthuijsen (2007)

The SMB procedure can also be used analytically and can thus be automated to continuously process wind data collected over the years from weather stations, buoys, satellites or other sources [18].
This model is most often applied to predict wave characteristics in closed or semi-closed basins. The accuracy of this model is highly variable because the equations applied are general and may be better suited for some sites and less so for others.
In a study conducted in a Bay of Bengal in 2015 [21], the SMB method has been applied to predict significant wave height and peak wave period during cyclonic storms. In addition to this, another empirical method, called Young, and a numerical one have been also applied. The results show that the estimated H_s by empirical methods fits well with the model simulated value, while concerning the wave period, the SMB overestimates a bit the prediction. The author concludes that the empirical equations provide good predictions especially for the wave height.

In another study conducted in the Mediterranean, specifically in an area off the northern coast of Sicily, the SMB method is applied to predict the wave height and the results are compared with a buoy measurements [19]. The author's conclusion is that, even with some limitations observed, the SMB model can still be considered as a reliable tool in providing usable values when wave measurements are not available. In conclusion, empirical models can be a viable alternative for wave prediction. The implementation of the method will be analytically described in the section 3.3.1 and the results will be subsequently discussed and compared to other methods.

2.2 Numerical models

Using atmospheric and oceanic information, numerical models are more and more improving to provide accurate sea state predictions allowing a faster growth of most maritime sectors. Companies operating in this sector have an increasing need for wave information. For example, the offshore oil industry needs wave data for fatigue analysis, operational planning and marine operations. In recent years, also the marine energy exploitation is becoming more and more important for the energy transition, (as previously discussed in par.1.2), and to accelerate its growth, an accurate wave forecasting is critical.
A key input parameter to the wave models is the wind energy transferred to waves through surface stresses, that is roughly proportional the square of the wind speed. Therefore, an error in wind specification can lead to a large error in the wave energy and, subsequently, in parameters such as significant wave height. In order to achieve accurate wave model output, the quality of the wind fields is extremely important. Generally, two different approaches can be followed to model changing waves effectively [18]:

1. A mathematical representation. To characterize sea conditions, the surface waves are discretized in large random element that requires a statistical description. The scales (time step or grid length) must be small enough to resolve the wave evolution.
   The most common representation of the wave field is the sea surface variance

density spectrum E(f,θ), where f is the frequency and θ is the direction. The wave field is discretized in f and θ under the assumption of linear wave theory. Then, each component can be regarded as a sinusoidal wave for which there is a well known theory.
Following this approach, sea state parameters can be derived from this spectrum. Some of the most important are the significant wave height, the peak frequency, the primary wave direction, the zero crossing period, the directional spreading.

2. Simpler models may be built around direct estimation of the significant wave height with directional characteristics often derived directly from the wind. These processes are described by the response of useful statistical quantities, such as the wave spectrum, but not all the processes are fully understood. For this reason, empirical results are used within wave model too. Although the current research trend is to develop good physics-based representation of these processes. In some models, the performance can be adjusted by altering empirical constants. A general representation is given by Figure 2.6.

The most general formulation for computer models based on the elements in Figure 2.6 involves the spectral energy balance equation, which describes the development of the surface gravity wave field in time and space:

$$\frac{\partial E}{\partial t} + \nabla \cdot (c_g E) = S = S_{in} + S_{nl} + S_{ds} \tag{2.2}$$

where E = E(f,θ,x,t) is the five dimensional wave spectrum. It is function of the frequency f, direction θ, time t and space x. The variable $c_g = c_g(f)$ is the deep water group velocity and it is function of its discrete intrinsic frequency f. S is the net source function and, under the assumption of deep water, it can be divided in three different contributions: S_{in} the energy input with the wind; S_{nl} the non linear energy transfer by wave–wave interactions; S_{ds} the dissipation that can be split into a set of terms based on the different dissipation processes considered in the model. The main goal of wave modelling is to solve the energy balance equation 2.2. Based on the management of the non linear source term S_{nl}, the wave model can be classified in a first, second or third generation:

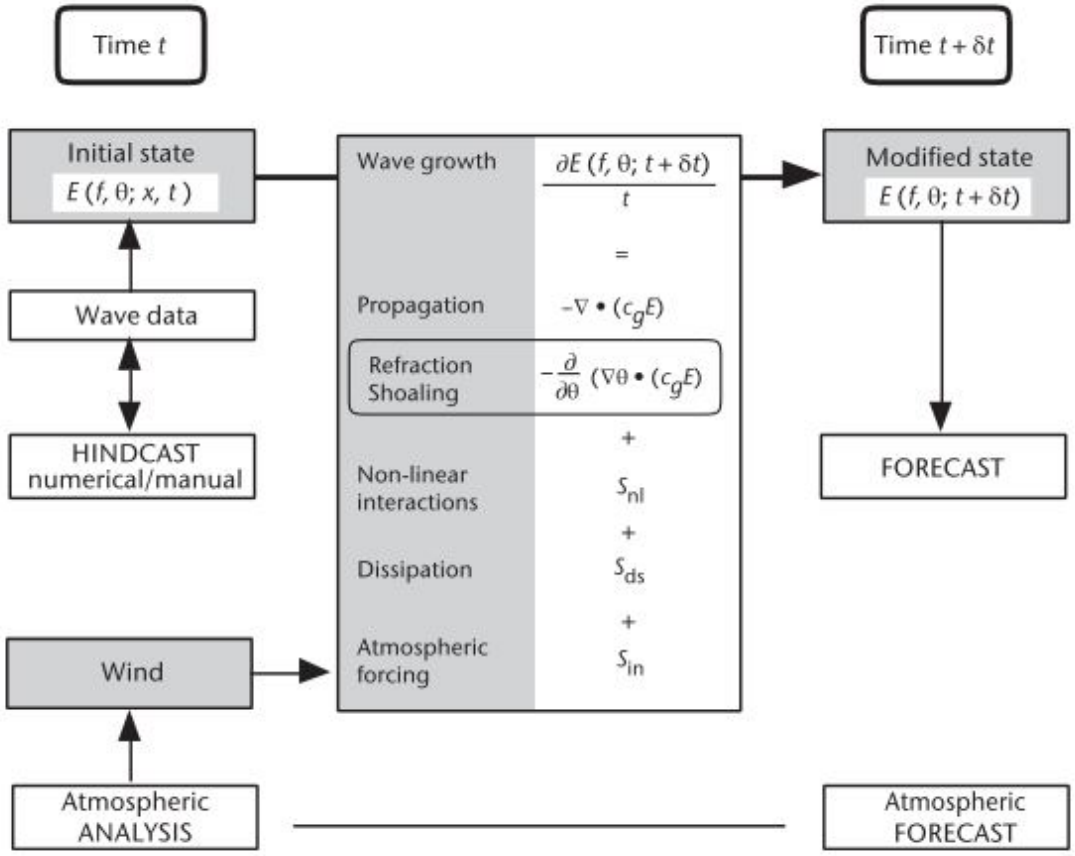


Figure 2.2: Wave modelling.

Source: Guide to Wave Analysis and Forecasting (2018)

- First generation models, also called spectral models. They do not have an explicit S_{nl} term. It is implicitly expressed through the S_{in} and S_{ds} terms.

- Second generation model, also called parametric model. They handle the S_{nl} term by parametric methods, for example, by applying a reference spectrum (as the JONSWAP or the Pierson–Moskowitz spectrum [22]) to reorganize the energy over the frequencies after wave growth and dissipation.

- Third generation models. They calculate the non linear energy transfers explicitly, although it is usually necessary to make analytic and numerical approximations to not have too high computational time.

The first and second generation wave models can be calibrated to give reasonable results in most wind situation. However, some shortcomings are present in these models, particularly in extreme wind and wave situation or when an accurate and reliable wave forecasts are most important.
The main difference between the second and third generation wave models is that, in the latter, the wave energy balance equation is solved without constraints on the shape of the wave spectrum, while in the former there is a reference spectrum. This is achieved trying to make an accurate calculation of the S_{nl} term.
Thanks to increasingly powerful computers, scientists began to develop a new, third generation of wave models that explicitly calculated each of the identified mechanisms in wave evolution. This led to the development of WAM (1994) [23] that has shown good results in extreme wind and wave conditions. Than other models have been developed like SWAN (2007) [24] and WAVEWATCH III (2016) [25].
In the recent year, these third generation wave models have been coupled with atmosphere models to achieve more and more accurate predictions.

2.2.1 Coupled atmosphere-ocean model

The term couple model refers to the simultaneous use of two or more models in order to take into account all equations and variables considered by the various ones, trying to build a model as close as possible to the reality.
For what concern the wave forecasting, the most used are the two-coupled models and they are composed of the union of atmospheric and ocean models. There are also three-coupled models that add to the previous two a wave model, like the WAM or the WAVEWATCH III.
A software interface is used to match the different models and it is called "coupler". It provides a flexible way of linking component models and controlling the exchange and interpolation of coupling fields.
The main advantage of coupled models is that changes in one model can directly influence the other models. For example, considering only the ocean model, the absence of any feedback on the atmospheric forcing variables, like winds, temperature and humidity, can cause many inaccuracies [18].

On the other hand, probably the biggest challenge of coupled modelling is the associated complexity: different component models need to be linked together in a flexible way, but they must be scientifically sensible and computationally affordable too. Other complications are that these models often use different horizontal grids, moreover they are developed by different communities and have different priorities, timescales and management for their development. So the decisions on the coupling approach will be driven by a whole variety of factors.
The big debate of these models is the compromise between resolution and complexity. Nowadays, there are no models with good performance on present-day climate estimation, that simulate aspects like storm tracks, hydrological cycle, clouds, and that simultaneously include aspects of the earth system, most notably the carbon cycle which may become increasingly important in future.
In a report published in 2018, Chris Harris perform an analysis considering different timescale [26]: climate (from 10 to 100 years); seasonal (3-12 months) and decadal (1-20 years); short to medium range (1-2 weeks).

- Climate (from 10 to 100 years). The new frontier of climate models is the move towards earth system models of increasing complexity that takes into account the whole variety of different physical and bio-geochemical model components. Such models have to include a representation of the carbon cycle too. The individual terms in the global carbon cycle are very large and small changes could have a significant impact on the amount of carbon dioxide which remains in the atmosphere, and consequently on the amount of future warming.
  These models are very expensive to run and additional components, that have poor historical observations or with intrinsic longer timescales like ice shelves, increase the complexity and the computational time.

- Seasonal (3-12 months) and decadal (1-20 years). Seasonal (and decadal) forecasting systems make use of combination of model simulations to provide probabilistic forecast information. On these timescales, the initial conditions atmospheric layer become less important. On the other hand, the aspects of the land surface and sea ice and the low frequency forcing from the ocean, provide the dominant contributions to predictability.

- Short to medium range (1-2 weeks). This timescale is the one on which coupled

atmosphere-ocean modeling has been the most developed in the operational oceanography and numerical weather forecasting over the past decade. For regional modelling, coupled systems can not be considered as "new frontier" and there are a lot of models already developed. For what concern global configuration there are few one. The main are CCMEP (Canadian Centre for Meteorological and Environmental Prediction) that provide both ocean forecasting and numerical weather prediction, ECMWF (European Centre for Medium Range Weather Forecasting) that at the moment provide only numerical weather prediction and UK Met Office that provides ocean forecasts. However, other models are under development.

The major applications of short timescales models are similar to those using traditional forecasting systems but some forecasting improvements has been observed and that may open up for new products. In particular, more accurate predictions of tropical storms and their tracks, is potentially of great significance.

A coupled model application will be briefly described below.
A study conducted in the 2020 provided a statistical assessment of advanced two-way coupled atmosphere-ocean wave modelling system, namely CHAOS [27]. It consists of the Weather Research Forecasting (WRF) [28] model as atmospheric component and the Wave model (WAM) as ocean wave component, coupled through the OASIS3-MCT coupler [29]. The analysis was performed in the Mediterranean and Black Sea and it was set up to perform a continuous simulation during one year. Two configuration have been compared:

1. the 1-way coupling mode, the ocean wave component uses the wind data produced by the atmospheric component, so the data are only transmitted from the atmospheric to the ocean level;

2. the 2-way coupling mode, the ocean wave model uses the u and v components of wind at 10 meters produced by the atmospheric model. Meanwhile, the atmospheric one uses sea state information to estimate roughness length, so there is a bilateral exchange of information.

The results show that the simulations in 2-way coupling mode produce more realistic results and the predicted values follows slightly better than the observed

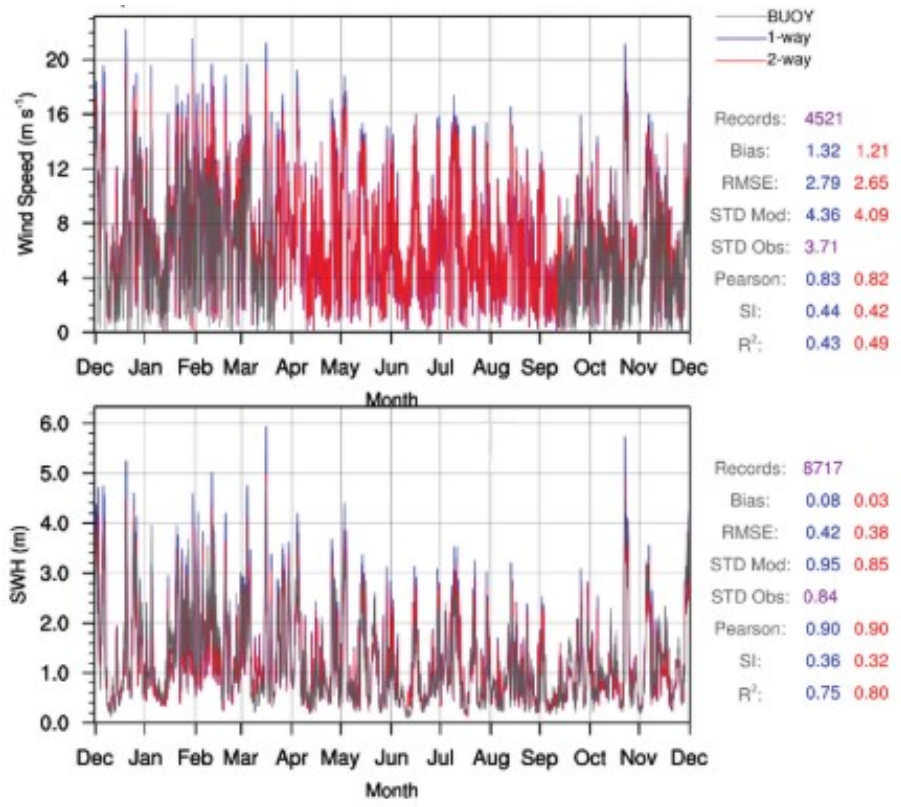


Figure 2.3: Results coupled atmosphere-ocean wave modelling system, namely CHAOS [27].

ones (Figure 2.3). Referring to the buoy observations, the 2-way mode reduces the Root Mean Square Error (RMSE) 1.2% for wind speed reaching achieving a value of 2.65, while for significant wave height the performances are even better and the RMSE is reduced by 6.3% compared to 1-way configuration with a total value of 0.38.

2.2.2 Statistical approach, multiple regression

To solve the main problem of the coupled model, i.e., the computational time and the need of supercomputers to run, the use of statistical models is adopted in many studies present in the literature. One of the most used is the multiple regression. Regression analysis is a technique used to analyze a dataset consisting of a dependent variable and one or more independent variables [30]. The purpose is to estimate any

relationship, i.e., the regression equation, that exists between the dependent variable and the independent ones. The former results a function of the latter plus an error term. This error is a random variable and represents uncontrollable and unpredictable variation in the dependent variable .
Statistical methods include also the autoregressive (AR), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA).
To build this kind of model, existing data are needed to find the relation between the different variables, that generally are wind peed, wind direction and temperature (in some cases also the pressure but it is strictly correlated to the temperature and so may be redundant).
On the other hand, they are extremely less complexity and, if the study is conducted in a specific spot, where a lot of variables can be assumed constant (morphology, water deep, coordinates and so the seasons and many others), the results of these model are often comparable to the others types. Obviously statistical models are hard to apply if the goal is to build a general global model for forecasting.
As said, maybe the main disadvantage compared the coupled models is the need of a big dataset to build it and to find the correlation coefficients between the variables. Obtaining these data can be costly both from an economic and time point of view. If the data are not available, a measuring program must be implemented, like the installation of some buoys, but, especially on regional basis, the costs involved and the period spent waiting for a reasonable amount of data to be collected, are unacceptable.
For this reason, hindcasts are playing an increasing role in marine climatology. Wave hindcasting refers to predicting surface waves for a past wind event [31] and, if the period of the hindcast is sufficiently long, the database can also be used for analysis to long-return periods. Many times, the relations or models used for predictions for the past are the same used for future event prediction (forecasting). Therefore in the absence of large dataset, the adopted strategy is the following one:

- develop a model to hindcast the data of the studied spot;
- use those data to forecast wave data based on the time you want to consider (longer forecasts require larger dataset);
- Check if the prediction are congruent with the observations. If so, future

data should also be considered reliable, otherwise a new model should be constructed.

A study conducted by K. Gunaydın compares two different methods to estimate the wave height at an Atlantic Ocean site: an Artificial Neural Network (ANN) method (it belongs to the machine learning models described in the section 2.2.3.1) and a regression method [32]. Waves and meteorological data were collected by different buoys and different ANN and regression method configurations has been analyzed.
The best configuration results are reported in figure 2.4.

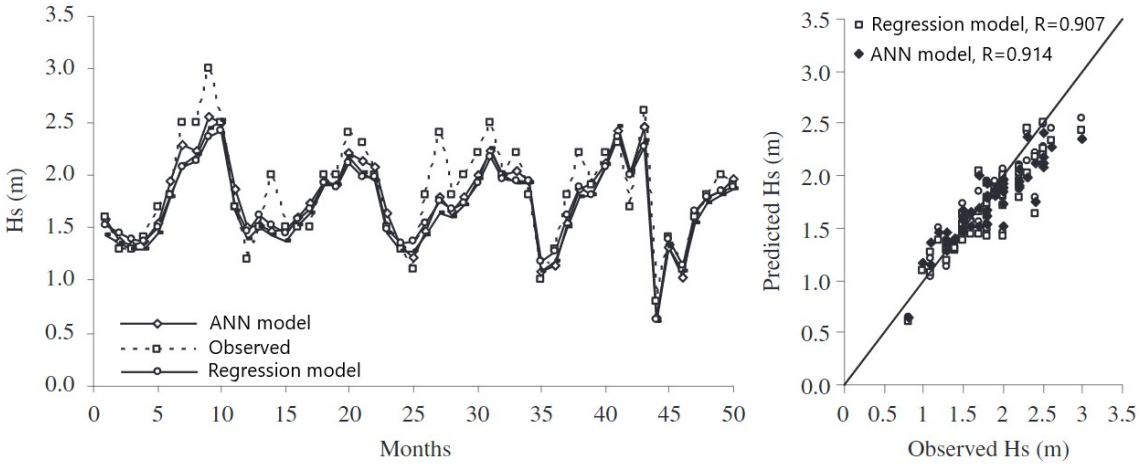


Figure 2.4: Results ANN method and regression method [32]

At the end of his work the author states that the differences are small between the regression model and ANN model performances. Therefore, these two approaches can be both used under the conditions of used wave and meteorological data.
This is further evidence that regression methods can achieve very good results if the study is focused on a specific spot. Moreover, they are quite easy to implement and in most cases the problem is only the data retrieval.

2.2.3 Artificial Intelligence methods for forecasting

The first example of Artificial Intelligence dates back to 1951 with a successful AI program written by Christopher Strachey [33], later director of the Programming

Research Group at the University of Oxford. The term "Artificial Intelligence" had been used to describe machines that mimic the "human" cognitive skills that are associated with the human mind, mainly "learning" and "problem-solving".
Machine learning, as defined by Tom M. Mitchell [34] is the study of computer algorithms that allow computer programs to automatically improve through experience. An algorithm can be imagined as a set of rules or instructions that a computer can process. Machine learning algorithms learn by experience, analogous to how humans do. For example, after having seen multiple examples of an object, a machine learning algorithm can become able to recognize that object in new, previously unseen scenarios.
Probably the main advantage of machine learning over the regression model is that it allows to be merged more data in order to perform a more accurate prediction [35]. Machine learning methods can adapt to changes in the dataset and are less easily manipulated than traditional ones that become less accurate over time. Also for future prospective, machine learning is becoming more accessible as the technology advances. There are many software platforms that allow anyone to a model without any prior knowledge. On the other hand, traditional methods often require specialized knowledge and training.
In the last decade, with the development of Artificial Intelligence, various new AI methods have been developed. Some of the most used are reported in the following sections.

2.2.3.1 Artificial Neural network model

Probably the most widely used in terms of forecasting are the ANN models, especially when there are many experimental data collected. They could deal with non-linear and complex problems in terms of classification or forecasting.
ANNs are composed of artificial neurons which are conceptually deduced from natural neurons. Each artificial neuron has inputs and produces a single output which can be sent to multiple other neurons. The inputs can be the feature values of a sample of external data, while the outputs of the final output neurons of the neural network accomplish the task [36].

The neurons are generally organized into multiple layers, particularly in deep learning. Neurons of one layer connect only to neurons of the directly preceding and directly following layers [37]. The connections are called edges and weights, that change as learning proceeds, are defined for each neuron and edge some. The first layer is the input one that receives external data. The last is the output layer and it produces the ultimate results. In between them are zero or more hidden layers and multiple connection patterns are possible between two different layers. They can be 'fully connected', with every neuron in one layer connecting to every neuron in the next layer, or they can be pooling, where a group of neurons in one layer connect to a single neuron in the next layer (Figure 2.5).

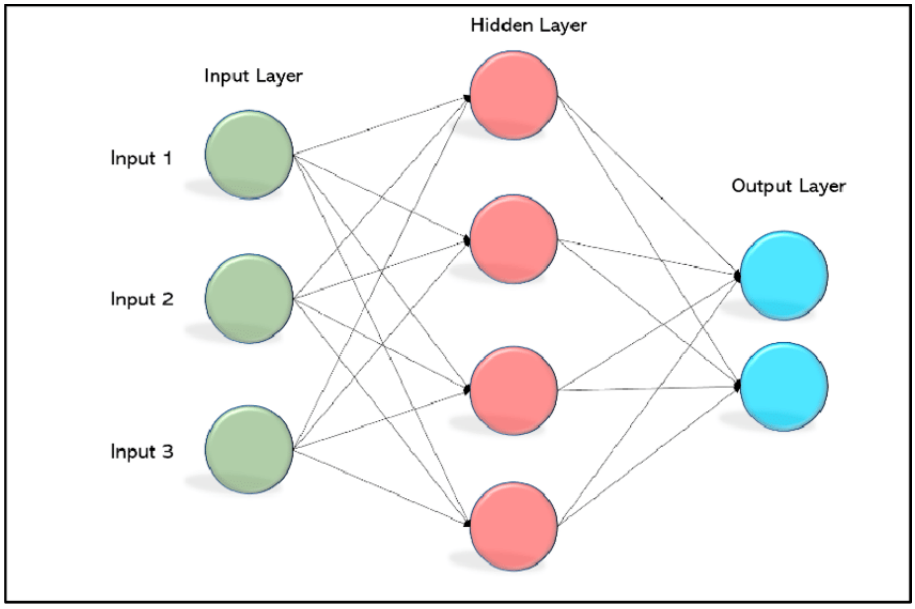


Figure 2.5: Artificial neural network representation.

In literature, many example of ANN model applications are available, especially for wind and wave forecasting. One example has been described in the previous paragraph 2.2.2, where an ANN method has been compared to a regression method [32]. The performances are very good, particularly if the study refers to a specific region.

The big advantage of using Artificial Neural Networks (ANN) is that they have a great ability to detect complex non-linear relationships between dependent and independent variables and to detect all possible interactions between predictor variables

too. Moreover, they can handle large amount of dataset.
An issue of neural networks is that they are black-box systems generating results based on experience and not on specified programs, so modifications are quite difficult. Moreover, as for the regression model, dependency on data is one of the leading disadvantages of Neural Networks. If there are errors in the data, the result will be faulty, which poses serious threats.

2.2.3.2 Decision Trees

In the field of Artificial Intelligence, the Decision Tree model is used to arrive at a conclusion based on the data from past decisions [34]. It is a simple and efficient model that can be applied for both regression and classification problems. Decision Tree is so called because the way the data are broken down into smaller portions is similar to the structure of a tree.

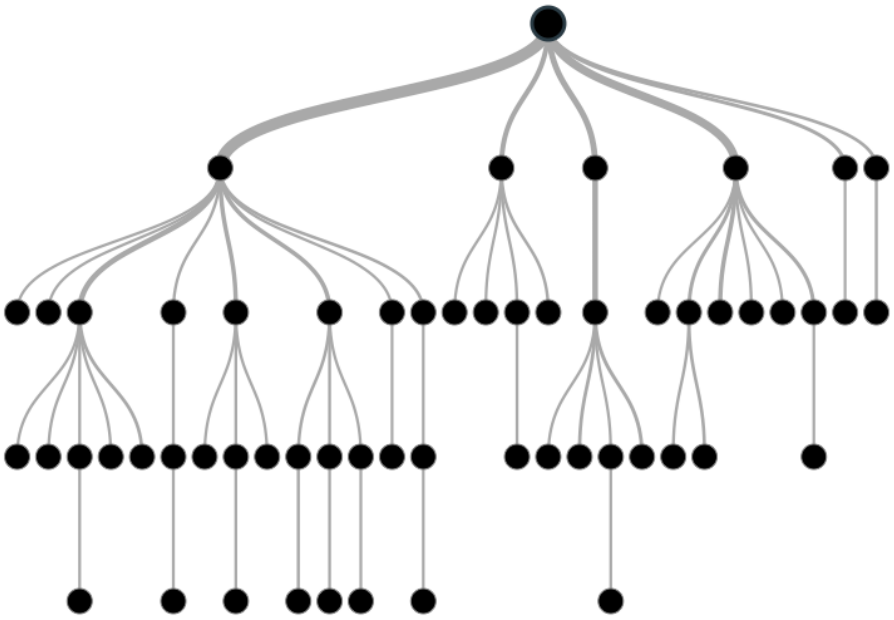


Figure 2.6: Decision tree structure.

Source: https://wiki.pathmind.com/decision-tree

Decision trees can enable to identify relations between two or more variables. Moreover, they can make classifications based on both numerical and categorical variables. A disadvantage of the model is that decision trees can accept continuous numerical input, they are not a practical way to predict such values. Another common problem is the overfitting. It refers to a model that models the training data too well. A model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.

2.2.3.3 Support Vector Machine

SVM, or Support Vector Machine, is a quick and efficient model that excels in analyzing limited amounts of data [38]. It is applicable for binary classification problems. Compared to newer technologies such as Artificial Neural Networks, SVM is faster and performs better with a data-set of limited samples, such as in text classification problems.
SVM works by mapping data to a high-dimensional feature space, thereby data points can be categorized even when the data are not linearly separable. First, a separator between the categories is found and successively the data are transformed in order to draw the separator as a hyperplane. In this way, characteristics of new data can be used to predict the group to which a new record should belong.

Chapter 3

Site and data description

The energy potential of waves is extremely high and, as previously described in section 1.2. The waves will be an increasingly important resource for renewable energy production, but, at the moment, there is still a significant need for the development of technologies. In this regard, search sites are increasing all over the world.
In Italy, one of the sites with the highest interest in development of marine energy is the Island of Pantelleria. This has been chosen as the site of this work, where the different wave prediction methods will be applied and studied.
The aim of this chapter is to provide the characteristic of the selected location. A brief description of the geography and mainly of the energy system and energy potential of the island is presented.
Subsequently, the resource of wind and wave data is described. The required data are obtained from the open access dataset of the ECMWF (European Centre for Medium-Range Weather Forecasts) called ERA5 [39]. The data acquisition is a fundamental step for the future implementation of the forecasting models.
In the end, the models analyzed in this work are presented. The models are of different types: empirical, statistical or artificial intelligence models (as described in the previous chapter).

3.1 Site description

3.1.1 Geography and demography

Pantelleria is an island in the Sicily Region. It is located in the Strait of Sicily, Mediterranian Sea, at 10 km South of Sicily and 65 km North-East of Tunisia and its area measures 83 square kilometers (Figure 3.1).

Figure 3.1: Pantelleria Island.

The Island, of volcanic origin, has a mostly flat area in the northern part, where the main town (Pantelleria) and port are located [40]. The rest of the territory is characterized by a variable topography, with the presence of numerous terraces for agricultural practice.

Pantelleria benefits from a Mediterranean climate that implies hot summers and mild winters. Rainfall is mainly concentrated in the winter season.

In particular, the exposure to the wind coming from the NW direction, called Maestrale, strongly characterizes the winter period, with recorded wind peaks of over 35 m/s and high intensity swells.

The island, due to its inclination toward tourism, has great seasonal variability in attendance. It counts about 7800 inhabitants in winter and about 30,000 in the

summertime distributed mainly in three population centers: Pantelleria, Khamma-Tracino and Scauri.

3.1.2 Energetic system

Because of its great distance from the coast, the Island of Pantelleria has a logistically complex and expensive energy supply, both in terms of the electric carrier and fossil fuels in general. The island is not interconnected to the mainland and the electricity is currently produced by diesel generators. Consequently, the means of transport use gasoline and diesel engines.
Pantelleria has a very high availability of RES, among the largest in Italy, as well as a great variety of them. The resources of the greatest interest are solar, wind and wave.
The solar source is mainly related to latitude and it presents an annual global solar radiation of more than 2100 kWh/m^2.
Its position in the center of the Sicilian Channel ensures a high windiness for the Island of Pantelleria, among the highest in Italy. Especially the winter months are characterized by very high average wind speeds, but summer also has important peaks. The average wind speed is more than 7 m/s at 25 m above the sea level [41].
Because of the strong correlation with wind, the wave resource is also important. The average annual incident energy flux in the area in the northwest of the island is about 7 kW/m, referred to the unit length of the wave front.
Thanks to this great availability of renewable resources, Pantelleria has become one of the best site to study and improve the energy transition in the Mediterranean sea. The island has been selected as one of 26 European islands to receive European Commission support for energy transition through the Clean Energy for EU Islands Secretariat [42]. In the 2020 the *Agenda per la Transizione Energetica* has been stipulated where all the key points are described to achieve the total decarbonization by 2050 [41].

Nowadays, about 7% of electricity is produced from about 1 MW of small photovoltaic systems [41]. In 2022, all school buildings, town halls, libraries and other

Table 3.1: Pantelleria energy targets [41].

Year	**RES penetration [%]**	**Self-sufficiency of residential sector [%]**	**CO2 emissions reduction since 2018 [%]**	**Energy self-sufficiency [%]**
2018	1%	1%	0%	1%
2025	20%	10%	-15%	11%
2030	35%	25%	-30%	20%
2035	50%	45%	-45%	30%
2040	65%	60%	-60%	45%
2045	80%	70%	-80%	70%
2050	100%	80%	-100%	100%

buildings will be equipped with photovoltaic systems. Moreover, the Regional Government, the Municipality and the National Park are purchasing electric buses for the island's public transport.
The optimal energy mix identified for Pantelleria is reported in figure 3.2.

- 40% of the electricity produced from photovoltaic plants, both distributed and centralized. The plan provides for the installation of 15 MW of photovoltaic systems, including 7 MW distributed (on rooftops and pergolas) and 8 MW in medium-sized centralized systems, including offshore.
- 41% of the electricity produced from wind power plants. 6 MW consisting of offshore wind power plants on floating platform (2 MW size turbines), while 300 kW of onshore medium size wind power plants (60 kW turbines).
- 4% of the energy produced from wave power plants through 14 wave energy

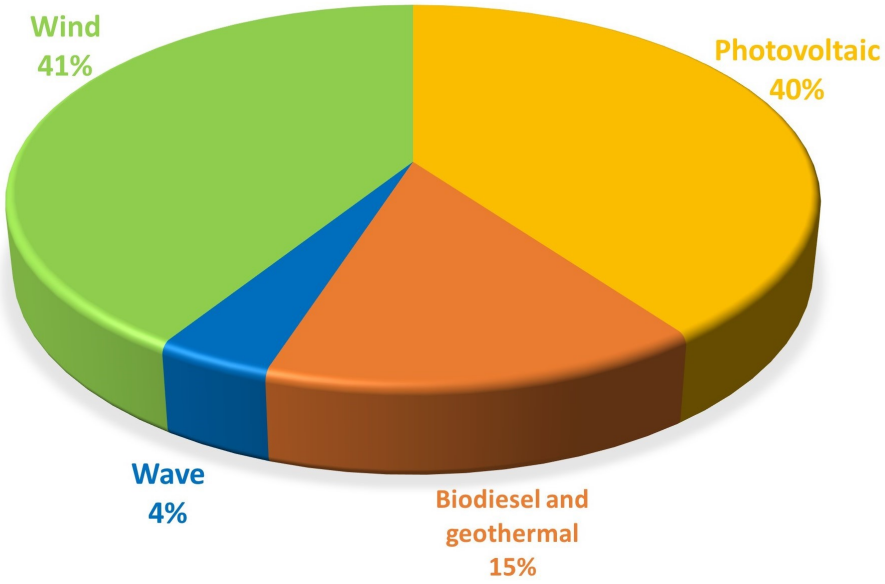


Figure 3.2: Planned electricity mix for Pantelleria.

converters (WEC). Each one has a size of 1.7 MW.

- 15% of electricity from programmable RES (imported biodiesel and possible biomass harvested on site) and geothermal source (which can provide for part of the base load).

One of the main problems of an energy system with very high penetration of non-programmable RES is undoubtedly the time matching between electricity supply and demand. Given the inconvenience of installing a disproportionate amount of storage systems, it becomes necessary to apply so-called curtailment, that is the forced cut-off of available power from RES in case energy supply exceeds demand and there are no storage systems ready to receive energy. Since as the curtailment rate decreases the need for storage systems increases exponentially, it will therefore be necessary to accept that a certain amount of energy from non-programmable renewable sources will not be fed into the grid. The rate of curtailment may be around 6 % on an annual basis.

3.2 Wind and wave data

The reliable data acquisition from the available resources is fundamental for planning future scenarios, as previously described for the island of Pantelleria. The predefined targets for the following years are based on a careful analysis of those data, considering many aspects: economic and financial, administrative, legislative, social. Therefore the resource of the data has to be as accurate as possible to avoid future problems scenarios that were not taken into account.
Also in this work the acquisition of precise data is crucial. The aim is to develop wave prediction model and, without accurate data, the models will never be able to perform good prediction and the errors would be too high to really consider those as feasible.
Fortunately there are several methods to collect meteorological data and these sources can also be compared in order to check the goodness of the data.

3.2.1 Wind data sources

The best way to get wind information is to install an anemometer in the selected site and to collect data for at least one year. An anemometer is a device that measures wind speed and direction and there are many different types [43]. The main ones are:

- Cup anemometers. They use a vertical axis shaft with cups and rotational speed varies with wind speed. So the number of pulses in a certain time gives mean wind speed. Generally they are coupled with a vane for wind direction measurement. They measure data for a specific point and at the height where they are installed.

- Propeller anemometers. Very similar to the cup anemometer, but they use a a horizontal axis turbine. So they register data of both speed and direction.

- Sonic anemometers. They measure wind speed based on the time of sonic pulses between pairs of transducers. Also for these devices the measurement is in one point.

- Acoustic Doppler sensors (SODAR). They measure the data for an entire surface and not only a point as the previous ones. They are based on the Doppler effect of an acoustic signal to detect the air flow in the atmospheric boundary layer and deduce the wind speed and direction. These are classified as remote sensing devices (RSDs) are easily relocated in several points.

- Light Doppler sensors (LIDAR). Very similar to the SODAR but they use a laser signal instead of an acoustic one. They are also more accurate and therefore more used than the previous ones.

However, installing a tower for the anemometers or the use of SODAR or LIDAR devices can be expensive and the amount of data collected by on-site anemometer can be not worth the cost. Therefore, free data is often used to estimate the wind field. The free mapping tools based on existing local data are often the best information available, but it is important to emphasize that these sources do not give as accurate data as on-site device. The greatest advantage lies in quickly obtaining a large amount of data even in large time intervals (months or years). This makes it possible to reduce, if not eliminate, the cost of data acquisition and to conduct accurate analyses immediately.
There are several ways available to collect wind data, from software to web application. Maybe the most used source, especially to perform preliminary analysis on wind characteristics of the site, is the *Global Wind Atlas* web portal [44]. It is owned by the Wind Energy Department of the Technical University of Denmark (DTU Wind Energy) and it provides free access to data on wind power density and wind speed at multiple heights using the latest historical weather data and modeling, at an output resolution of 250 meters. The Global Wind Atlas is based on a coupling of mesoscale to microscale modeling using meteorological re-analysis data provided by the European Centre for Medium-Range Weather Forecasts. Users may download poster maps, GIS data and Generalized Wind Climate.

As show in figure 3.3, large scale atmospheric data from re-analysis datasets are used as an input into medium scale mesoscale atmospheric models. Then, the output from the mesoscale modeling is generalized and used as microscale modeling input. In the end, taking into account a high resolution topography, the output of the microscale modeling is the wind prediction.

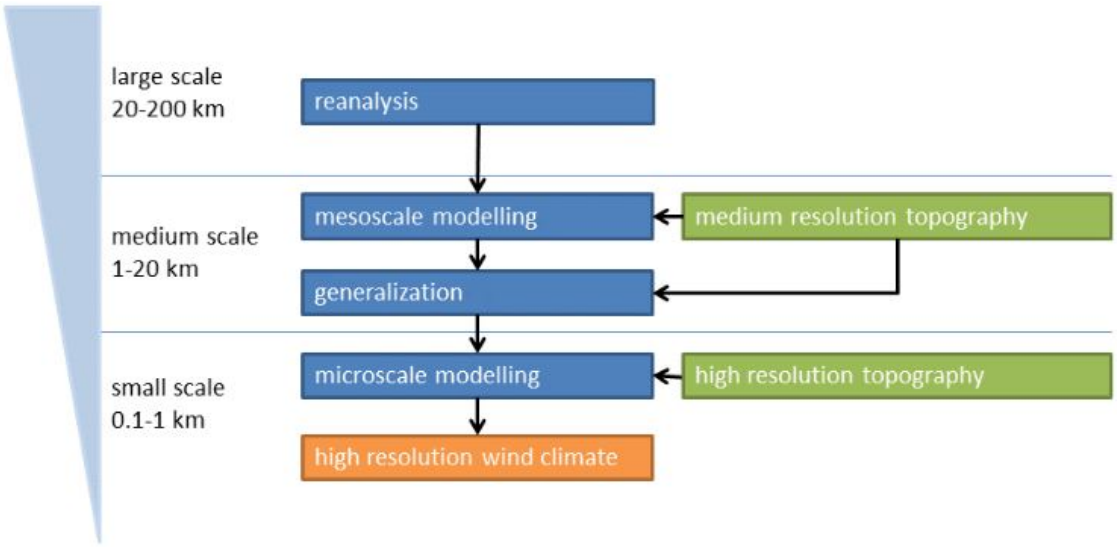


Figure 3.3: Schematic showing the methodology of the GWA is downscaling [44].

A very similar wind atlas is the *New European Wind Atlas* [45], that is an updated version of the old *European Wind Atlas* developed in the last years. The mesoscale modelling of the NEWA covers the entire EU plus Turkey and 100 km offshore as well as the complete North and Baltic Seas. The microscale modelling is similar to that used in the Global Wind Atlas, except a finer 50 m calculation grid spacing is used. The NAWE web page provides ways for user to display map layers, perform analysis and plot charts, and download data.
Another alternative to get wind data is the software *The Wind Data Generator* (WDG) [46]. It is a wind energy software tool to obtain a wind data at any location, any height of interest with a resolution between 3 km and 10 km. The numerical weather prediction model implemented in the software is the Weather Research and Forecasting (WRF) model. It is one of the most used model, as discussed also during the coupled models description (section 2.2.1) because it is an open-sources code.
The ERA5 dataset is a further source widely used for data acquisition. It is the fifth generation ECMWF atmospheric reanalysis of the global climate covering the period from January 1959 to present [39]. ERA5 is developed by the Copernicus Climate Change Service and it provides hourly estimates of a large number of atmospheric, land and oceanic climate variables. This sources will be explained in more detail in

a dedicated section (3.2.3).
The last alternative is to consider satellite measurements. This way is generally less used than the others because very often the data are not in open source and it is more complicated to obtain them. Moreover, the resolution is not so high, especially for wind data, if compared to the other sources.

3.2.2 Wave data sources

Also for the wave data acquisition there are several methods. The most accurate one is with no doubt the measurements on the site the most used devices are [47]:

- Accelerometer buoy. A buoy floating in the water moves up and down with the waves and the device measures wave accelerations. Integrating twice the acceleration, the displacement versus time can be obtained. However this device is useful to measure large waves and it cannot respond to capillary or small gravity waves due to its mass.

- Pressure gage. An underwater device that is sensitive to the amount of water between it and the surface. Because of wave movement, the pressure changes in the point directly above the gage. A measure of pressure versus time is equivalent to wave height versus time measurement. This device is coupled with a recording meter generally located on-shore.

- Step-resistance gage. A thin vertical support is partially immersed in the water and on that there is a set of exposed electrodes, each one in series with a fixed resistor. As the water moves up on the gage staff, more electrodes are shorted out and the total gage resistance decreases. In this way, wave height is simply correlated to a measurement of resistance versus time.

- Acoustic Doppler current profiler (ADCP). It is a device able to measure wave height, wave direction and also the current profile using the Doppler effect of sound waves scattered back from particles within the water column. It is positioned on the seafloor in relatively shallow water (depth up to 100 m). To measure the water surface movements are used sound waves (SONAR), infrared beams or radio waves (RADAR).

As for the wind in situ measurements, the issues are the installation and maintenance costs, especially for some devices like buoys and ADCP, and the time needed to record all the data. Moreover, long term buoy wave measurement networks are still relatively few and far between. For example, along the Italian coast there are only 15 buoys installed by the Italian wave measurement system called RON (*Rete Ondametrica Nazionale*). To conduct an analysis on a specific site, these data are very often insufficient and, for these reason, also other sources are considered to obtain wave data.

An alternative are the satellite data. They are acquired through the back-scattered signal from satellite altimeters [48]. They can provide measurements with accuracy close to buoy from an orbit of typically 1,000 km, but they need to be calibrated and quality controlled: each satellite altimeter has to be validated in order to remove the altimeter-dependent biases on significant wave height. This is generally done by comparing with long term offshore buoy data. The measurements are usually made each second and the satellite velocity is about 6 km/s. This results in enormous amounts of wave data worldwide, and millions of new observations available each month. This makes satellite data extremely suitable if the goal is to perform a global analysis, instead, these data are maybe not the best choice for studies in specific regions because they are not so frequent since they depend on the satellite path.

Another option for the wave data acquisition is through numerical models. Third-generational spectral wave models are based on solving the spectral action balance equation in space and time.

The data are easily accessed through web portals that represent a good alternative for the fast data procurement. The most used are two: *ERA5* and *MetOcean Data Portal*. The former will be detailed described in a dedicated section because it is the source of the data used for this work (section 3.2.3). The latter is a paywall dataset owned by the Danish research organization DHI. It is based on MIKE 21 Spectral Waves model also developed by DHI [49]. The model simulates the growth, decay and transformation of wind-waves and swells in offshore and coastal areas. This portal offers instant access to metocean data and analytics of nearly 40 years all around the world with a hourly time resolution. In addition to wave information, there are available also information on offshore wind and currents. MIKE 21 SW is particularly applicable for wave prediction and analysis on regional scale and

especially local scale due to its irregular resolution: coarse spatial and temporal resolution is used for the regional part of the mesh and a high resolution and depth adaptive mesh describe the shallow water environment near the coastline (see figure 3.4).

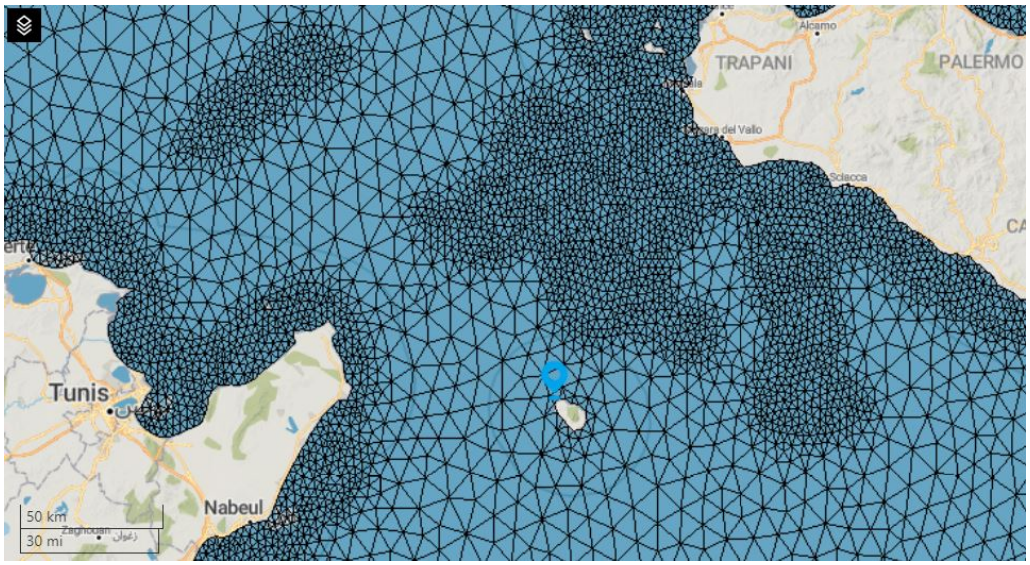


Figure 3.4: Example of MIKE 21 SW computational mesh.

3.2.3 ERA5 sources

ERA5 is source of the wind and wave data used in this work to build the wave forecasting models presented in the following section (3.3). It contains estimates of atmospheric parameters such as air temperature, pressure and wind velocity and direction at different altitudes, and surface parameters such as rainfall, sea-surface temperature, wave height and others.
ERA5 is the fifth generation ECMWF atmospheric reanalysis for the global climate and weather. The term "reanalysis" refers to a continually updating grid dataset that represents the state of the atmosphere, taking into account observations and outputs of numerical weather prediction models from past to present-day [39].
The model data are combined with observations from all around the world to obtain a complete and consistent dataset using the laws of physics. This principle, called data assimilation, is a methodology used by numerical weather prediction centers, where

every so many hours (12 hours at ECMWF) a previous forecast is combined with newly available observations in order to produce a new best state of the atmosphere. This is called analysis and from that an updated, improved forecast is issued.
Reanalysis works in the same way, but the resolution is reduced to allow for the provision of a dataset covering back several decades. A great advantage is that reanalysis does not aim to provide timely forecasts, so there is more time to collect observations. This allows to go further back in time and include more accurate data of the original observations, improving the quality of the reanalysis.
The data has been regridded to a regular latitude-longitude grid of 0.25° for the reanalysis covering the Earth on a 30 km grid and resolve the atmosphere using 137 levels from the surface up to a height of 80 km, as shown in Figure 3.5. (For ocean waves a regular grid of 0.5° is employed).

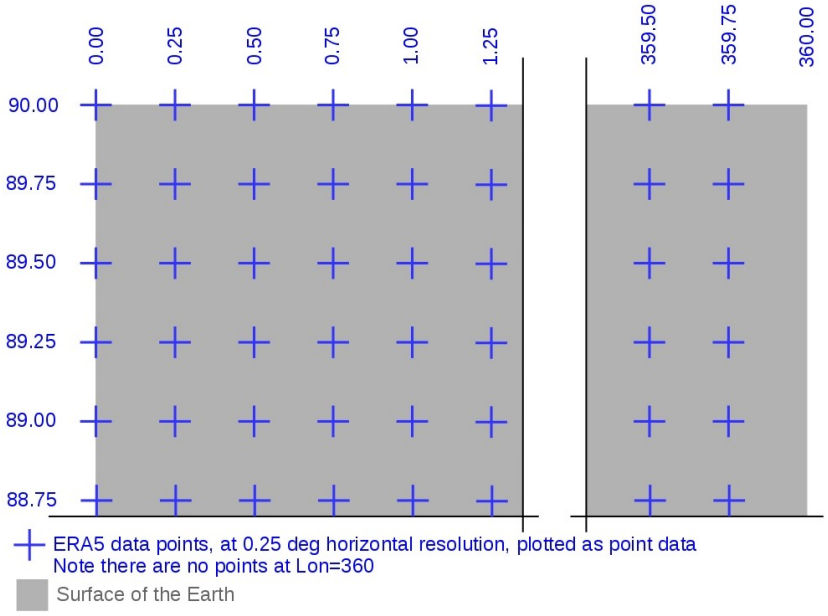


Figure 3.5: Visualization of regular lat/long grid ERA5.

Source: ECMWF portal

ERA5 also includes information about uncertainties for all variables at reduced spatial and temporal resolutions. The new ERA5 data are updated for quality assurance within 3 months of real time.
Presented below are the main parameters extracted from ERA5 and used to build

the forecasting models:

- 10 m U component of wind. This parameter has units of meters per second (m/s) and it is the horizontal speed of air moving towards the east, at a height of ten meters above the surface of the Earth. Care should be taken when comparing this parameter with measured values, because the latter vary on small space and time scales and are affected by the local terrain, vegetation and other factors that are represented only on average in the ECMWF Integrated Forecasting System (IFS). The speed and direction of the horizontal 10 m wind can be estimated combining this parameter with the vertical component of the wind at a height of ten meters.

- 10 m V component of wind. This parameter has units of meters per second (m/s). It is the horizontal speed of air moving towards the north, at a height of ten meters above the surface of the Earth. As for the U component, care should be taken when comparing this parameter with observations and the two components can be combined to give the speed and direction of the horizontal 10 m wind.

- 2 m temperature. This parameter is the temperature of air at two meters above the surface of land, sea or inland waters, and it has units of Kelvin (K). It is calculated by interpolating between the lowest model level and the Earth's surface, taking account of the atmospheric conditions.

- Mean direction of wind waves. The wave spectrum can be decomposed into wind-waves, which are directly affected by local winds, and swell, the waves that were generated by the wind at a different location and time. This parameter takes account of wind-waves only. It is the mean over all frequencies and directions of the total wind-wave spectrum. The units are degrees true, which means the direction relative to the geographic location of the North Pole, so zero degree means "coming from the North" and 90 degrees means "coming from the East".

- Significant height of wind waves. This parameter represents the average height of the highest third of surface ocean/sea waves generated by the local wind. It

represents the vertical distance between the wave crest and the wave trough. Also this parameter takes account of wind-waves only and the wind-wave spectrum is obtained by only considering the components of the two-dimensional wave spectrum that are still under the influence of the local wind.

- Mean period of wind waves. This parameter is the average time it takes for two consecutive wave crests, on the surface of the ocean/sea generated by local winds, to go through a fixed point. As the others wave parameters, it considers only wind waves and it is the mean over all frequencies and directions of the total wind-sea spectrum.

These data, obtained over a period of ten years (from 2012 to 2021), will be used to build the different forecast models presented in the following section.

3.3 Description of implemented models

In this work, basically three different models will be analyzed and compared. Precisely the following will be studied:

- an empirical model that is the Sverdrup-Munk-Bretschneider (SMB);
- a multiple regression, classifiable as a statistical model;
- a deep learning technology that is the Artificial Neural Network (ANN).

3.3.1 Sverdrup-Munk-Bretschneider (SMB) model

As reported earlier in the section 2.1.1, the SMB is based on the resolution of the energy balance. In the literature, several equations can be found to derive the wave height and period, but they all depend on three variables: the wind speed U, the fetch length F and the duration of the wind event t.
The first step is to determine the length of fetch in all the directions starting from the studied point. Generally, for studies concerning the Mediterranean Sea, an

important assumption is done: a maximum fetch length is considered [50]. This is done because in the Mediterranean wind usually does not have constant direction and intensity over the entire length of the fetch. Indeed the major of the wind events, and consequently the highest wave height, are due to barometric differences and in this condition fetch loses its meaning. So, some wind variations could occur even if no obstacles are present in a given direction. For this reason, excessive distances are avoided. In some studies this value is assumed from 450 to 500 km. In this work, because at North-West of Pantelleria there is a wide section in with the wind blow basically without obstacle, the maximum value of the fetch is fixed to 500 km.
Using the tools of Google Earth software [51], the fetch is estimated all around the Pantelleria (from one direction to the next one there is an angle of 5°).

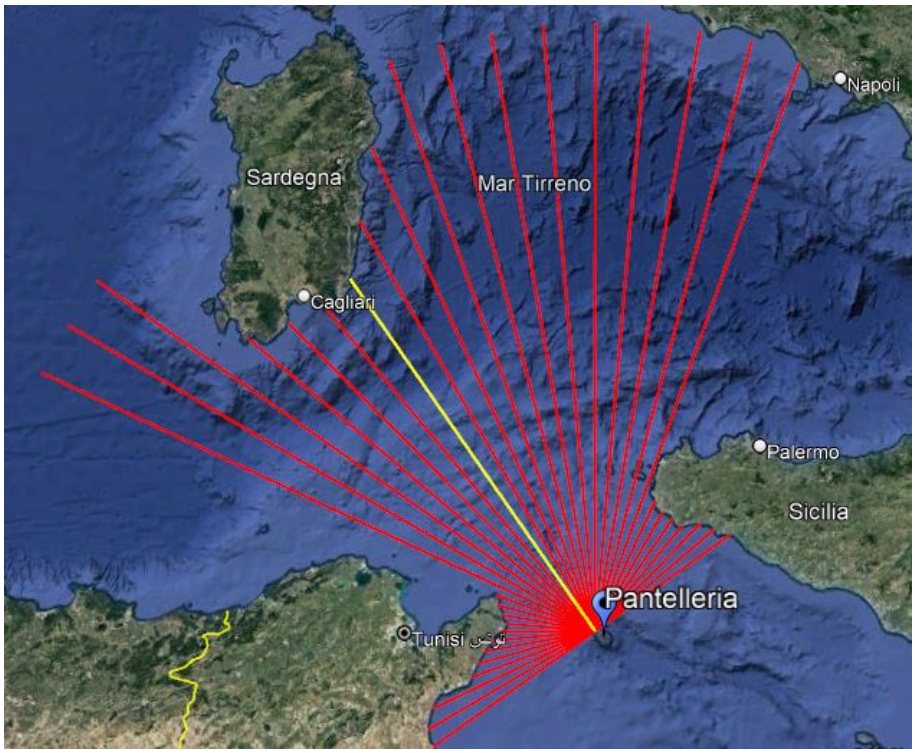


Figure 3.6: Fetch length from the the Island of Pantelleria.

In the Figure 3.6, the fetch from Pantellria is represented only for the North-West sector to provide a clearer view, but, obviously, in the model have been considered all the directions, so the South-East sector too.
As discussed in the section 2.1.1, these fetch values are used to calculate the effective

fetch in all directions using the equation 2.1.
The next step is the evaluation of the limited time t_{lim}. It depends on the fetch and especially on the wind speed (less time is required to reach the FDS condition at higher speeds) and it is needed to estimate if the limiting variable is the fetch or the duration of the event. An event ends when the wind stops blowing or when it changes the direction of an angle greater than 60°. This angle is arbitrary, generally varies from 30° to 60°, and it must be chosen according to the conditions of the case study [19].
There can be basically two scenarios:

- if the duration of the event is grater than t_{lim}, the limited variable is the fetch;
- if the duration of the event is less than t_{lim}, the limited variable is the time.

Another scenario could be when neither time nor fetch are limiting value and the FDS condition is reached. But, as mentioned above, in the Mediterranean there are continuous variation in the wind speed and direction and this scenario is very rare. For that, in this work the state of the sea will always be treated as limited by event duration or fetch.
The used equation are the ones reported in a USA study conducted in 2009 [52]. These equations have been chosen from the many in the literature because they present the best performances in the considered site.
The limited time t_{lim} is derived from the following equation:

$$\frac{g\ t_{lim}}{U} = 6.5882\ \exp\left\{\left[0.0161\left(\ln\frac{gF}{U^2}\right)^2 - 0.3692\left(\ln\frac{gF}{U^2}\right) + 2.2024\right]^{0.5} + 0.8798\left(\ln\frac{gF}{U^2}\right)\right\} \quad (3.1)$$

Where:

g is gravitational acceleration constant equal to 9.81 m/s^2;

U is the horizontal wind speed in meters per second;

F is the effective fetch length in kilometers.

For each time sample, so every three hours, the wind speed and fetch are already defined and the limited time can be evaluated.
If the duration of the event results greater than t_{lim}, there is a fetch limited condition and the significant wave height H_s and wave period T_p can be estimated using the following equations [52] [21]:

$$\frac{g\ H_s}{U^2} = 0.283\ \tanh\left[0.0125\ \left(\frac{gF}{U^2}\right)^{0.42}\right] \tag{3.2}$$

$$\frac{g\ T_p}{U} = 0.24\pi\ \tanh\left[0.077\ \left(\frac{gF}{U^2}\right)^{0.25}\right] \tag{3.3}$$

Also in this case U represents the horizontal wind speed in meters per second, F the effective fetch in meters and g the gravitational acceleration constant. The resulting significant wave height is expressed in meters.
The second scenario is then the time event results less than t_{lim}. In this case the equation 3.1 must be used to evaluate the limited fetch, i.e. a fictitious fetch value that is less than the real value in order to consider the lack of time to have a fully developed sea. To estimate this value, the t_{lim} must be replaced by the wind time, the wind speed is given and the only unknown in the equation results in the fetch F. Evaluated the limited fetch, this value is substituted in the equations 3.2 and 3.3 to calculate the significant wave height.
The ERA5 data, used in this work, are collected every three hours and also the limited time, evaluated with the equation 3.1, turns out to be a few hours in most conditions. Due to the very similar order of magnitude, it is reasonable to assume that the fetch is the dominant limited variable. This aspect will be better explained in the section 4.1.2 where the results of the SMB model are presented.

3.3.2 Multiple regression model

The second typology analyzed is a statistical model, precisely a multiple regression one. The aim of the regression analysis is to study a dataset in order to estimate

any functional relationship between the dependent variable and the independent variables [30]. The dependent variable in the regression equation is a function of the independent variables plus an error term. The latter is a random variable and represents uncontrollable and unpredictable variation in the dependent variable. The regression model is widely used for prediction and forecasting. It can be applied without any knowledge of the physical processes that generated the data and in this case the model is an empirical one.
The simplest regression method is the linear one in with the dependent variable y is correlated to one independent variable x.

$$y = \beta_0 + \beta_1 x \tag{3.4}$$

The coefficient β_1 represents the relationship between independent and the dependent variable and β_0 is the intercept value of the function.
To build a regression model, first of all the model estimation must be defined. It refers to how the correlation between variables is investigated [53]. The most used method is the Ordinary Least Squares (OLS). The working principle is to fit each data i to a regression line (Figure 3.7) in order to minimize the sum of the squared distances to it. This distance represents the error or residual e, and it is the difference between the predicted value $\hat{y}$ and the real value y, that is the one observed (see equation 3.5).

$$e_i = y_i - \hat{y}_i \tag{3.5}$$

These distances are squared to avoid that negative distances (i.e., below the regression line) cancel positive distances (i.e., above the regression line). Moreover, by using the square, the observations that are far from the regression line are emphasized much more than the observations close to the regression line[53].
The aim of the OLS method is to minimize the sum of the squared residual SSR. This results a set of simultaneous linear equations in the parameters, which are solved to obtain the estimates of parameters $\hat{\beta}_0$ and $\hat{\beta}_1$, i.e., the coefficients of the regression equation.

Figure 3.7: A visual explanation of regression analysis.

Source: Towards Data Science

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{3.6}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3.7}$$

In the equations 3.6 and 3.7, the term $\bar{x}$ and $\bar{y}$ represent the mean value of all dependent and independent variables, respectively.
A fundamental hypothesis for the application of the OLS method is that the error term has a constant variance all along the line [53]. Under this assumption, the variance $\hat{\sigma}^2$ can be estimated (equation 3.8). It is also known as Mean Square Error (MSE).

$$\hat{\sigma}^2 = \frac{SSR}{n-2} = MSE \tag{3.8}$$

with n equal to the number of data considered.
The square root of this value represents the Root Mean Square Error (RMSE). It is widely used in forecasting and regression analysis to show how concentrated the

data are around the line of best fit. A lower RMSE is indicative of a more accurate model. When more than one variable must be correlated to the dependent one, a more general method of regression is used that is the multiple regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_v x_v \tag{3.9}$$

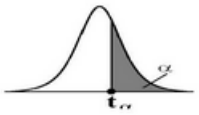
The equation 3.9 represents the most general expression of the multiple regression model in which the dependent variable y is correlated to v independent variables x. The coefficients β represent the relationship between each independent variable and the dependent one and β_0 is the x value when all other variables are zero and represents the intercept value of the function. Also for the multiple regression, the most used method to estimate the coefficients is the OLS. The equation for the variance estimation is similar to the previous one (3.8), but it takes into account more than one β coefficient:

$$\hat{\sigma}^2 = \frac{SSR}{n-(k+1)} \tag{3.10}$$

where k is the number of the β coefficients and n is always equal to the number of data considered.
For any regression model application is very important to verify the acceptability of the coefficients obtained. Indeed a statistical method, called hypothesis test, must be applied to verify if these values can be assumed significantly different from zero. A very common test is the T-test [54]. It checks the relationship between the target variable and every predictor variable independently, one by one, and it associates each correlation with a value t. The latter is equal to the correlation coefficient β_i of the independent variable i divided by the standard error of the estimated coefficient. Referring to the probability density Student's t-distribution, each t-statistic is associated with a p-value by using of proper distribution tables (Figure 3.8).

The p-value shows the probability, for a hypothesis assumed to be true (in the case of regression it is the goodness of β coefficients), that the observed result is due to the randomness introduced by sampling, or if this result is statistically significant. The more this value is low the more the variable is significant. Generally the

Percentage Points of the *t* Distribution; $t_{\nu,\alpha}$
$P(T > t_{\nu,\alpha}) = \alpha$



ν	α													
	0.40	0.30	0.20	0.15	0.10	0.05	0.025	0.02	0.015	0.01	0.0075	0.005	0.0025	0.0005
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706	15.895	21.205	31.821	42.434	63.657	127.322	636.590
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303	4.849	5.643	6.965	8.073	9.925	14.089	31.598
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145	2.264	2.415	2.624	2.771	2.977	3.326	4.140

Figure 3.8: Student's t-distribution table.

Source: https://www.scribbr.com/statistics/students-t-table/

limiting value of the p-value is 0,05 and if an higher value is observed, the variable is rejected. In the section 4.1.3, where the results of the multiple regression model are discussed, the p-values associated in all the independent variables are shown too.
The last important step is to verify the absence of collinearity among the independent variables [53]. The collinearity is a data issue due to an high correlation between two independent variables. This scenario should absolutely be avoided and in case of collinearity, one of the two variables should be eliminated. Also three or more variables could be correlated and in that case we talk about multicollinearity. This aspect will be taken up and shown in the section 4.1.3 where the model results are discussed.

3.3.3 Artificial Neural Network model

The Artificial Neural Network is a machine learning model that, differently from normal models, "learn" and acquire information directly from data without relying on predefined equations.
As described in the section 2.2.3.1, the model is composed by a network of several nodes, called artificial neurons, organized in different layers and connected each other. As the process goes on and "learns", different weights are defined for each

node and connection trying to achieve the lowest possible error deriving the outputs from the inputs data.
The aim of this work is to determine the wave characteristic using atmospheric input parameters, such as wind velocity and air temperature. To do that, the use of a discrete dynamic model was chosen. This is a model in which the input data can be current or past ones, acquired with a certain sample time. The MATLAB neural network toolbox will be used to build the model [55].
The first step is the data ordering. All the inputs are organized in a matrix in which each row contains all the data of one input variable. The resulting matrix has a number of rows equal to the number of variables in input and as many columns as there are the observed data.
As said, also past data can be model inputs to forecast the wave characteristic for a given instant. Therefore several configurations are possible depending on the number past event considered. Obviously the results are different from one configuration to another, but not necessarily the best will be the one with the most inputs. The ANN model results are presented in the section 4.1.4, and this aspect will be explained in more detail.
Defined the input matrix, the dataset is split in three different blocks used for the three phases of the model, i.e., the training, validation and test phase. The definition of training, validation and test is a general technique use to reduce model overfitting [56].
The overfitting is the big issue of the neural networks model. It occurs when the model outputs match the observations almost perfectly, but when the same model is used to predict new data, the accuracy is very poor. Moreover, in that condition, the model takes into account even the noise of measurements and this need to be avoided if the aim is to obtain a model as general as possible. The key is to find the best compromise between complexity and error.
To do that, training, validation and test phase are used [56]:

- In the training, the weights, that represent the relationship between ANN inputs and outputs, are investigated. In this phase the model "learn" from the input data and the weights are modified at each iteration in order to reduce the errors.

- The data used in the validation are the base to calculate the error, i.e., the difference between the observed value and the one estimated by the model. The results of each iteration is compared to the previous one to point the model in the right direction, that is to reduce the error.

- Also in the test phase the errors are estimated and they are compared to the value obtained in the validation. If the latter carries forward the model, the goal of test phase is to stop the model and to avoid the overfitting. Indeed, if the error in the validation is much less than error in the test, it means that the model has good performance for the data used in the learning phase but not for the new data, so it is overfitting.

Typically in the literature the recommended percentages are around 70% of dataset for the training, 20% for the validation and 10% for the test. The training phase has the biggest percentage because a large amount of data is required to optimize the model. However, in this study tens of thousands of data are collected over ten years (more than twenty-nine thousand data). For this reason a lower percentage of training is adopted in order to increase the validation phase trying to make the model as accurate as possible without overfitting. In particular, 30% of data is used for the training, 50% for the validation and 20% for the test.
The training algorithm is called "Adam" and it is provided by the neural networks MATLAB toolbox. This algorithm is coupled with a technique called minibatch. It consists in the subdivision of the dataset into equally sized subsets, i.e., the batches, and for each one the training is performed and the model weights are estimated. The training of the subsequent batch will be initialized using data taken from the previous one. Therefore, the whole dataset is analyzed but piece by piece and, in this way, the error is minimized for one batch of data at a time and not for all of them, avoiding the overfitting.
For the analyzed ANN model, the size of the batch is set to 128. Also higher values could be used but a lower one allows for a more robust model preventing the main problem of machine learning model, i.e., the overfitting. (It is important to emphasize that the dimension must always be a power of two).
Once the data have been acquired, the last step before performing the training is to randomize the data, so to shuffle the order of the data and do not take them from

first to last chronologically. In this way, during the training of each batch, the data considered are taken from different time periods avoiding considering trends that may occur in specific times. For example, in our case small changes in the ERA5 data acquisition over ten years may generate little different trends. Therefore, to prevent the model from fitting them and to make it as general as possible, data shuffling is performed.
Then, the structure of the network is defined. In addition to the input and output layers, the size and the number of hidden layers must be chosen. A rule of thumb is to set the size at 2^n where n is the number of input variables. In this analysis, since past events were also considered as input, the size has been reduced to 2^4 in order to avoid too high computational time. Instead, the number of hidden layers is generally set to the number of inputs, but also this number has been reduced for the same reason and it is set to four. Despite that indexes reduction, the performance of the model remains more or less the same.
The following step is the activation function definition. The aim of this function is to transform the summed weighted input from the node into an output value to be fed to the next hidden layer [57]. Moreover, it decides whether a neuron should be activated or not, in other words, it determines if the neuron's input is important or not in the output prediction using simpler mathematical operations. In this network is used the rectifier or ReLU (rectified linear unit) as activated function. This function does not activate all the neurons at the same time. Indeed, if the output of the linear transformation is less than zero, the neurons will be deactivated [57].

The main ReLU advantage is the higher computational efficiency than other functions, like sigmoid and tanh, since not all the neurons are activated. Moreover it allows for a generic structure suitable for most of the problems.
In the end, a hyperbolic tangent function (tanh) is applied before the output layer. The goal is to give a range to the output variable filtering outlayer and extreme events.
The structure of the neural network is simplified in the figure 3.10.

The model inputs can be different variables x_i (the index i represents the number of these) and several past events p, compared to the analyzed time t, can be considered. Moreover, the previous values of the dependent variable y could be inputs

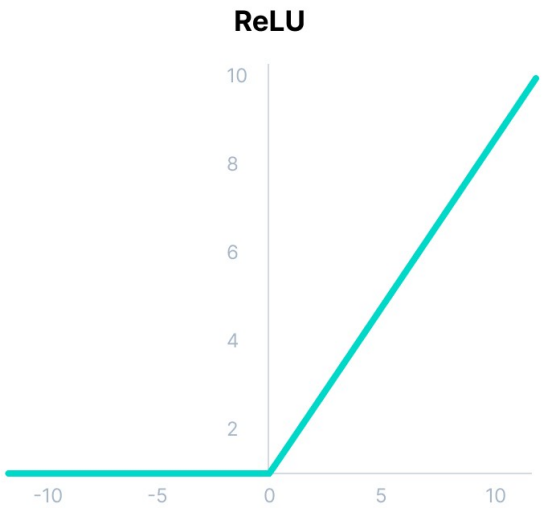


Figure 3.9: ReLU activated function.

Source: V7 Labs

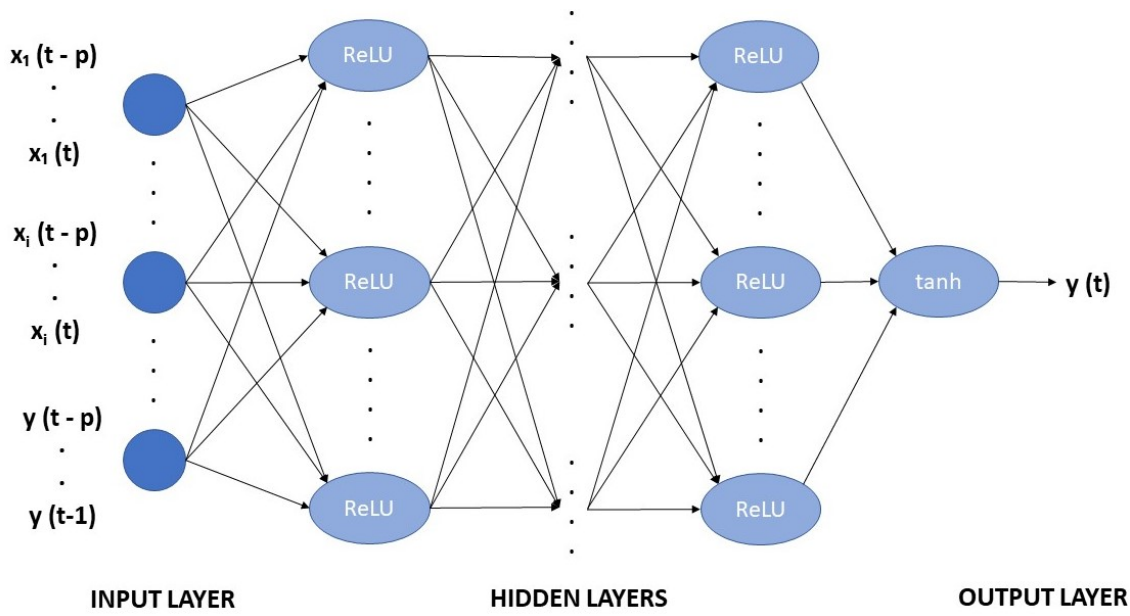


Figure 3.10: Artificial Neural Network structure.

to predict the value of $y(t)$. The hidden layers are composed by ReLU activated function as previously explained and the last layer before the output consists of a hyperbolic tangent.

To build this model in the MATLAB environment, the specific neural network toolbox is used. Firstly the data are normalized. Then, the connection and ReLU layers are added. After these, the tanh function coupled with a scaling layer are inserted. The scaling layer is needed otherwise the obtained value will be between -1 and 1 as the hyperbolic tangent. In the end there is the output, that is the model prediction. Different model configurations, especially on the input layer, will be analyzed and compared in order to find the best set up. In the section 4.1.4 all the ANN model results will be presented.

Chapter 4

Results and comparison of the models

Once the models have been described, the next step is to present and analyze the results. Different model configuration, especially for the regression and ANN models will be studied to achieve the best performance. Then, the outcomes of each model will be compared in order to find the best one for the Pantelleria Island.

4.1 Presentation of the results

4.1.1 Wind and wave data in Pantelleria

First of all, the wind and wave characteristic are described. They are derived extrapolating the data of the Pantelleria area from the ERA5 database. As previously mentioned, the data are collected over ten years, from the 2012 to 2021, and using MATLAB environment, these data are employed to plot the wind rose and wave rose of the site. They are represented respectively in Figures 4.1 and 4.2.

The wind rose points out how the wind blows mainly from the North-West (NW) and South-East (SE) directions. It fits perfectly with the morphology of the surrounding area: Pantelleria is located in the center of the Strait of Sicily, which extends in the same directions. This confirms the correct extrapolation and reliability of the data from the ERA5 database.

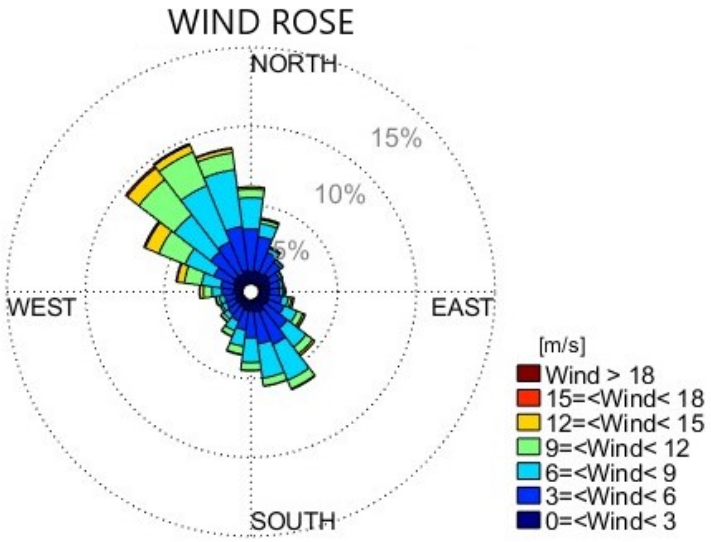


Figure 4.1: Wind rose Pantelleria Island.

The wind rose has been divided in 24 sectors having each an angle 15 degrees. The concentric lines show the percentage of time the wind blew in a given direction compared to the total data (the sum of all percentages must equal to one hundred). Analyzing the figure, it appears that more than 45% of the wind comes from the sectors to the North-West, and these are where the highest speeds are recorded. Indeed, velocity peaks above 12 m/s are present with non-negligible frequencies, while for the other directions, these velocities are either completely absent or hardly distinguishable on the graph (South-East direction) which means uncommon. Also the wind speeds between 9 and 12 m/s (green areas in the rose) are also far more frequent when the wind blows from the North-West than in other directions. Instead, speeds between 6 m/s and 9 m/s (light blue areas in the rose) are recorded in both the North-West and South-East sectors, while lighter winds are present all around the island with a slightly higher frequency for the areas mentioned previously.
In light of above, the best areas for future offshore wind farm projects and installations seem to be those in the NW of the island. However, it must be considered that these areas have the highest population density, so possible impacts such as visual impacts, must be carefully evaluated and resolved. The results found for wind

characteristics are also fully reflected on the wave ones, shown in figure 4.2, due to the strong dependence between wave generation and wind.

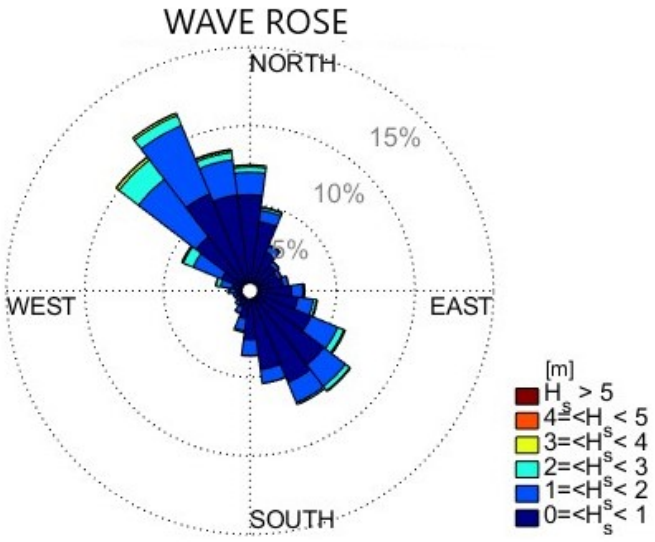


Figure 4.2: Wave rose Pantelleria Island.

As the wind rose, also the wave rose has been divided in 24 sectors having each an angle 15 degrees and the significant wave height H_s and the frequency of each direction are represented. Also for the wave, the predominant directions are North-West and South-Est ones, while the waves coming from North-Est and South-West are almost negligible. The greatest significant wave heights are recorded at NW, indeed only in these sectors a value of H_s between two and three meters (light blue areas in the rose) is recorded with an acceptable frequency. Moreover, also values between one and two meters (blue areas in the graph) are more frequent from NW than from SE. Lower waves, with H_s between zero and one meter, are the majority and they have more or less the same frequency from both NW and SE directions. Although the selected site is in the Mediterranean, considerable wave heights are recorded. Certainly waves are not comparable to ocean ones, but they can still provide a great energy potential that can be exploited. Therefore, it is crucial to design wave energy converter able to harness the stored energy even in the smallest waves since, despite they have lower energy potential, they are the most frequent.

The described dataset was exploited to build the three prediction models presented in section 3.3: Sverdrup-Munk-Bretschneider (SMB) empirical model, multiple regression model and Artificial Neural Network model. Their performance will be shown and discussed in the following sections and the two main analyzed parameters will be the significant wave height H_s and the mean wave period T_p.

4.1.2 SMB model results

As described in the section 3.3.1, the implementation of the model requires the resolution of two empirical equations, 3.2 and 3.3, in order to predict the significant wave height and the mean wave period knowing the wind velocity and the fetch.
In this model, it is very important to analyze the limiting variable, i.e., the one that does not allow to achieve the fully developed sea state. It can be the fetch length or the duration of the wind. To chose between these values, the evaluation of the limited time t_{lim} is crucial (equation 3.1). If the event duration is higher than t_{lim}, the limiting variable results the fetch, otherwise is the time event.
The outcome of this analysis, reported in the section 3.3.1, show that it is reasonable to consider fetch as a limiting variable in any case. Indeed, the limited time has an order of magnitude of a few hours and, as the data are collected every three hours and assuming constant wind during in that time frame, it can be concluded that the time elapsed between measurements is sufficient to reach a fully developed state from the time perspective. Moreover, without this assumption, the performance of the model would be very poor in cases where time is taken as the limiting value. This is due to the large time span between observations, which does not allow for an accurate estimate of when the limiting time t_{lim} is reached. It is important to stress that this assumption is not as acceptable for shorter time intervals, such as hourly acquired data. In that case the duration of the event is important since at least three or four observations of roughly constant wind are needed to have a time event higher than t_{lim} and to consider the fetch as limiting.
The SMB model is then applied to the ERA5 dataset in order to forecast the significant wave height and the mean wave period. The model predictions are compared to the observed values in order to evaluate the performances.

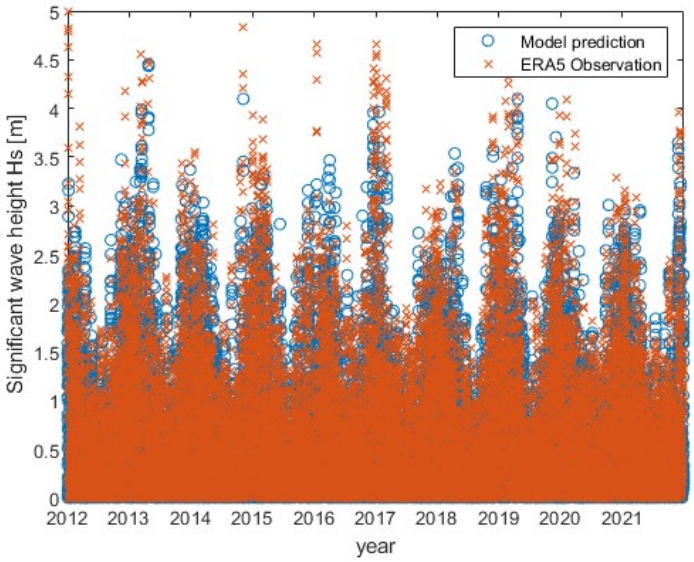


Figure 4.3: SMB model results, significant wave height prediction.

The figure 4.3 shows the ERA5 observations in orange and the predictions of the SMB model in blue for all the analyzed period, from 2012 to 2021. The first important result is that the model predictions seem to match quite well with the acquired data. Moreover, the graph shows a certain periodicity of wave height: there are periods in which the maximum height is around one meter and others with values above 3.5 meters. This is essentially due to the seasonal climate. Indeed, in the Mediterranean, the height varies significantly between seasons as reported in a study published in 2021 in the journal *Frontiers in Marine Science* [58]. The results are obtained from the analysis of ERA5 data over the past 40 years and they are shown in figure 4.4.

The season in which there are the highest wave height values is winter. Indeed, in this period the winds are typically stronger [59]. The highest values are observed in the West, between Sardinia and the Balearic Islands, while for the Strait of Sicily, i.e., the region studied in this thesis, maximum values are around 3.5 and 4 meters. The authors of the article report that extreme conditions are generally up to four times greater than the typical ones during this season. Regarding the summer period, the highest H_s are much lower, around 1.5 and 2 meters in the Strait of Sicily, and the regions west of Sardinia remain those with the greatest values. Spring and autumn,

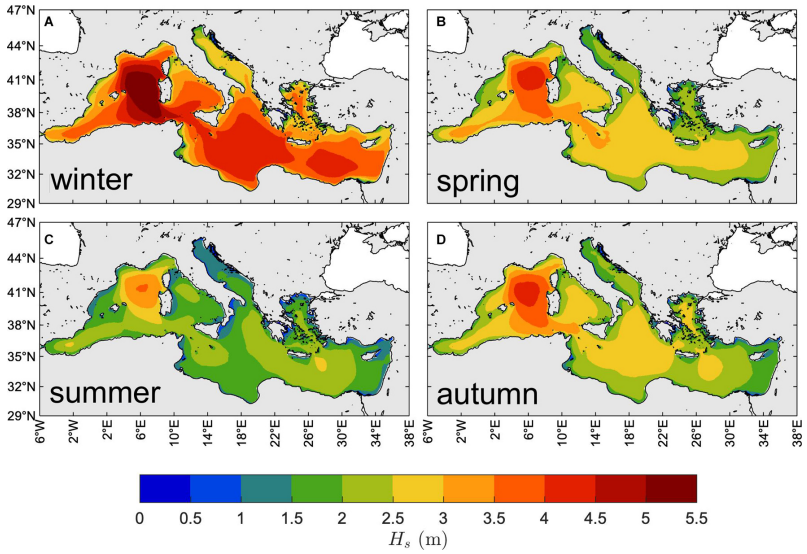


Figure 4.4: Extreme events significant wave height for each season, Mediterranean Sea [58].

on the other hand, are characterized by very similar conditions with intermediate values between summer and winter.
These data fit fairly well with the results of the SMB model, confirming that wave height predictions can be considered quite good.
To conduct a more detailed analysis on the results and to provide a model error estimation, the predictions are directly compared with the ERA5 values for each time instant and the following graph is built.

In the figure 4.5, all data are represented on a diagram called scatter plot, where on the x-axis are the values given by ERA5 and on the y-axis is the value of the SMB model for the same date. For better reading, the bisector $x = y$ is also shown of the graph. Indeed, the closer the points are to the bisector, the more accurate the model is, providing predictions similar to the real values.
Looking at the graph, we can see that most of the points follow the bisector and they can be regarded as good values. On the other hand there are several points under the bisector. They indicate that for those data the model forecasts are underestimated and in the practice there are higher values. For some points, the SMB model reports wave heights of a few centimeters or almost zero when the actual values are also a few meters (the values closest to the x-axis). These incorrect predictions greatly

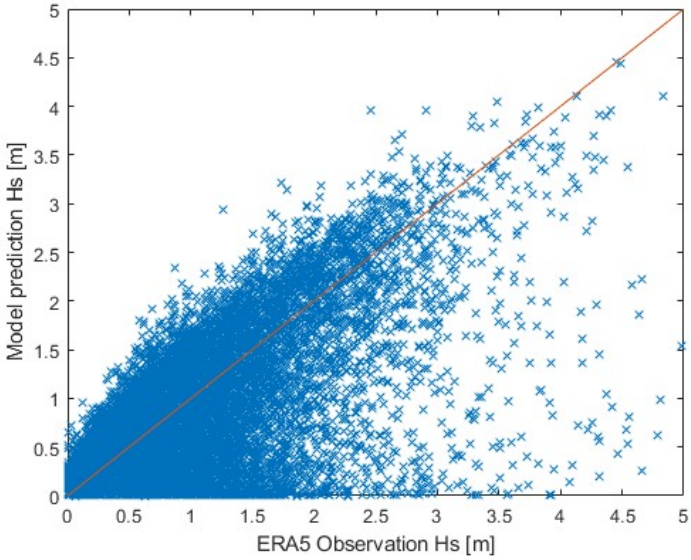


Figure 4.5: Comparison of SMB H_s prediction and ERA5 observation.

reduce model performance.
The explanation for these inaccuracies can be found in the structure of the model. As described in the section 3.3.1, the significant wave height H_s and the mean wave period T_p are estimated in each time instant knowing the wind velocity and the fetch. Therefore, the great shortcoming of this model is the consideration of previous data as inputs needed to predict future values. Indeed, whether the wind decreases or stops, the predicted H_s will be very low or zero. But, if high H_s had been recorded in the previous period, it is reasonable to assume that the wave height does not immediately decrease but there is a transient. This aspect is not considered in the empirical model SMB and it greatly reduces the accuracy.
To evaluate the performance of the model, the RMSE is calculated. It is an index with the same size of the considered value and it is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}} \tag{4.1}$$

Where:

x_i is the real value of the i-th data (obtained by ERA5 database)

$\hat{x_i}$ is the model prediction of the i-th data

N is the number of analyzed data.

The minimum value of the RMSE is zero. In that case the model predictions are perfectly coincident with the actual values. So, the lower the RMSE is, the more accurate the model is. For the significant wave height, the RMSE results 0.48 meters. This is quite a high result considering that the wave height is around a few meters. For this reason, the SMB does not seem to be the most suitable method for making accurate predictions, but it is still applicable to estimate trends over long periods of analysis. Probably by reducing the period studied the uncertainty of the results would be further emphasized.
For the mean wave period T_p, the predictions turn out to be even less accurate.

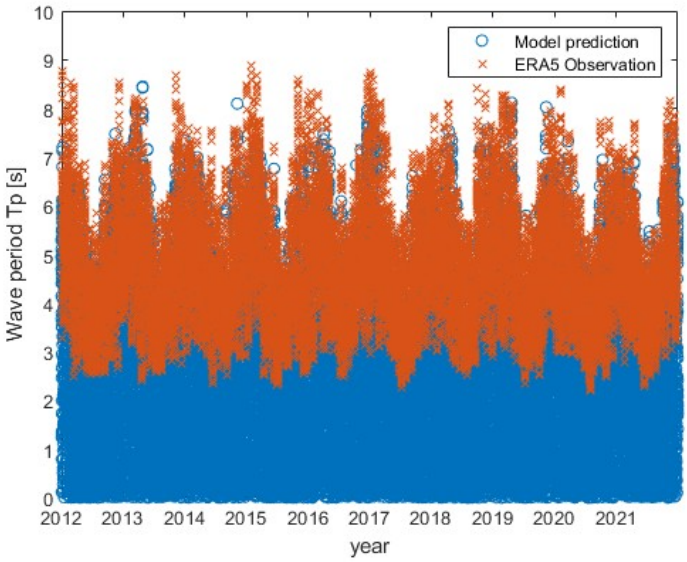


Figure 4.6: SMB model results, mean wave period prediction.

The Figure 4.6 clearly shows that the ERA5 T_p values never drop below 2 seconds, while the model predictions also reach zero values. As explained for the wave height, this is essentially due to the lack of the previous data as inputs to forecast the output. For the wave period this aspect is even more important than the wave height. Moreover, an underestimation of the data can also be observed for the peaks in the graph. The Figure 4.7 highlights these aspects even better.

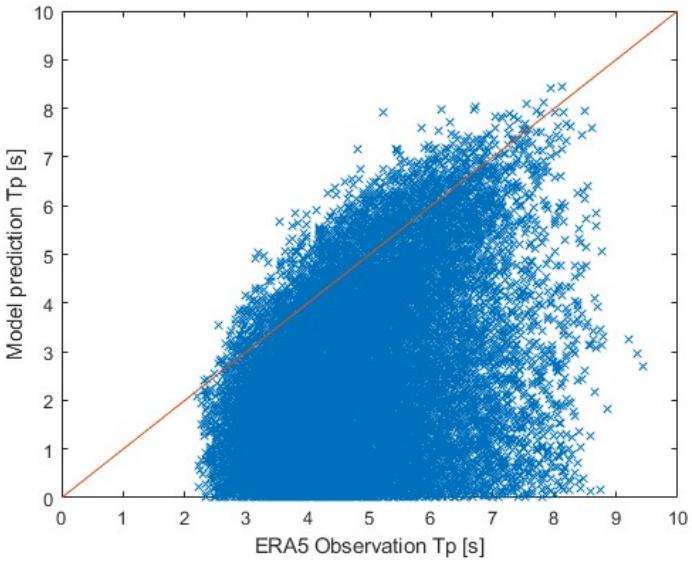


Figure 4.7: Comparison of SMB T_p prediction and ERA5 observation.

Most points fall below the bisector showing an underestimation. Moreover, the x component (the ERA5 observation) never goes under two seconds, while the forecasts, shown on y-axis, are spread from zero to eight seconds. In weak or no wind conditions, the empirical model predicts very low or null values of T_p that are very far from the real ones. This greatly reduces the performance, and the model is practically unsuitable for providing reliable forecasts over the period. The RMSE results 2.82 seconds, definitely too high considering a range of T_p about two and eight seconds.

In the end, the SMB empirical model turned out to be acceptable to predict the trend and values of the significant wave height in the studied site, i.e., the Pantelleria Island, while the performance of wave period estimation are very low.

In the next sections the results of the other two models will be presented trying to get more accurate predictions.

4.1.3 Multiple regression model results

The purpose of the regression model is to estimate the dependence, if it exists, of a variable with respect to one or more quantities, called independent or explanatory variables. In the study, the multiple regression is implemented since the independent variables are more than one. The model is explained in the section 3.3.2. The outputs are the coefficients of the equation that allows the computation of the dependent variable as a function of the others. The simplest is the equation 3.9 in which a linear dependence is expressed, but even more complex equations can also be applied to improve the model. An example is the quadratic one. In this section, the linear and the quadratic multiple regression model will be applied and compared in order to find the best predictions for the significant wave height and mean wave period.
Because the model considers more than one independent variable, the first step is to verify the absence of collinearity among them, i.e. no dependency. To do that, a MATLAB function, called "collintest" will be employed.
In addition to wind characteristics, i.e., wind speed and direction, atmospheric data are also chosen as useful variables for predicting wave characteristics, particularly temperature and pressure. Given the great correlation between wind, air temperature and air pressure, collinearity analysis is even more meaningful and critical to eliminate possible unnecessary variables.
To perform this analysis, two indexes have to be defined [60]:

- Condition index tolerance (CondIdx). It is used to decide which indices are large enough to infer a near dependency in the data. Large indices identify near dependencies among the specified variables. The value is set to ten. When a higher number is observed, there is collinearity and to o more variables are correlated.

- Variance-decomposition proportion tolerance (tolProp). It is used to decide which variables are involved in any near dependency. The index is set to 0.5 and when two or more variables present a higher value, they must be considered and correlated and the redundant ones must be eliminated.

The results of analysis, obtained through the MATLAB tool, are shown in the

table 4.1.

Table 4.1: Collinearity analysis regression model.

CondIdx	Wind speed	Temperature	Wind direction	Pressure
1	0.012	0	0.012	0
5.018	0.633	0.005	0.079	0.004
5.047	0.313	0	0.906	0
154.578	0.041	0.998	0.002	0.998

The last row of the table presents a CondIdx higher than the limit value of ten. The two variables that are strongly correlated turn out to be the temperature and the pressure, since the index tolProp results higher than 0.5 (see figure 4.8).

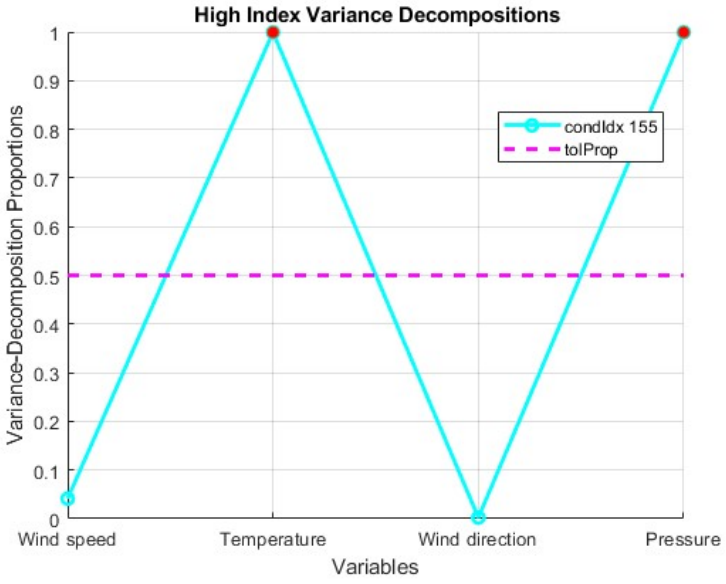


Figure 4.8: Collinearity analysis regression model.

The results are expected since the dependence between temperature and pressure is well known, however they confirm that the collinearity analysis was carried

out correctly.
It can be concluded that one variable between temperature and pressure should be eliminated and, no less important, that wind speed and direction turn out to be independent from the other variables. Therefore, the multiple regression model will be applied considering wind speed, wind direction and temperature as inputs trying to predict the wave characteristics.

4.1.3.1 Linear multiple regression model

Firstly is studied the linear model, therefore the model output equation will similar to the equation 3.9. The variables x_1, x_2 and x_3 are respectively refereed to wind speed [m/s], temperature [K] and wind direction [°]. The model has been applied to predict the significant wave height H_s. As described in section 3.3.2, for each variable coefficient β_i is been calculated the corresponding p-value in order to determine if that coefficient is acceptable or not. For all of them, the p-value results lower than 0.05 and almost near to zero, so they are valid. The equation for the H_s forecast, found by the linear regression model, results:

$$H_s = 1.204 + 0.193\, x_1 - 0.006\, x_2 - 8.553 \cdot 10^{-5}\, x_3 \qquad (4.2)$$

The coefficient of the variable x_3, that is the wind direction, is much lower than the other two. The direction has been added as input to consider the different fetch lengths. Indeed, specific wind directions where the fetch was higher could generate larger waves and, from the model results, this does not seem so relevant for the predictions. It should be noted, however, that the range of that variable is between 0 and 360 degrees, but even with high values, the third term of the equation remains very low compared to the other two. The main variable affecting wave height is wind speed, as expected. Also, the air temperature plays a key role in the wave prediction. Indeed, it is expressed in Kelvin and the second term of the equation has orders of magnitude comparable with the first one.
All the data are implemented to predict the wave height over the analyzed period and the model results are plotted 4.9.

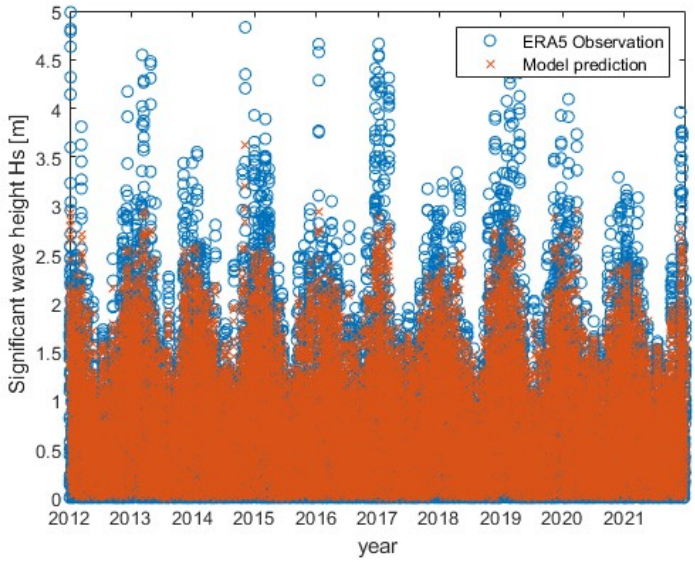


Figure 4.9: Linear multiple regression model results, significant wave height prediction.

The model predictions match well the seasonal trend, but they seem a bit underestimated compared to ERA5 observation. As done in section 4.1.2, the outputs are plotted on a scatter plot where on the x-axis are the values given by ERA5 and on the y-axis is the value of the model 4.10.

For lower wave heights, between zero and two meters, the prediction follow quite well the bisector and maybe they are a bit overestimated (most of the points are above the bisector). Going toward higher values of Hs, the values are increasingly underestimated, which greatly reduces the performance of the model. The RMSE results 0.234 meters, quite good compared to the SMB model.
The described trend suggests that a linear correlation is not sufficient to obtain very accurate predictions, especially for high wave heights. Thus models with higher orders of magnitude must be implemented to improve the regression model.
Regarding the mean wave period, the third coefficient of the regression equation presents a p-value of 0.258, higher than the threshold of 0.05. So it must be considered as null (p-value analysis, section 3.3.2). The model equation results:

$$T_p \;=\; 32.407 \;+\; 0.100\; x_1 \;-\; 0.097\; x_2 \tag{4.3}$$

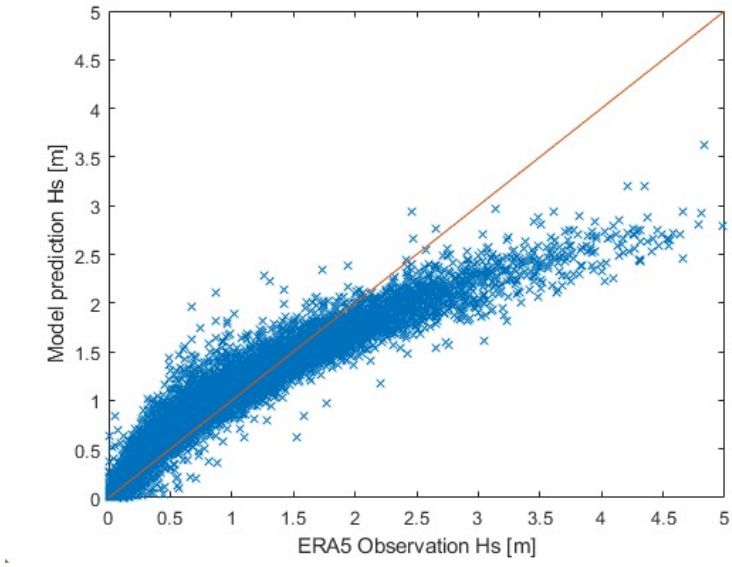


Figure 4.10: Comparison of linear multiple regression H_s prediction and ERA5 observation.

All the other coefficients of the T_p are verified and the model predictions are reported in the Figures 4.11 and 4.12.

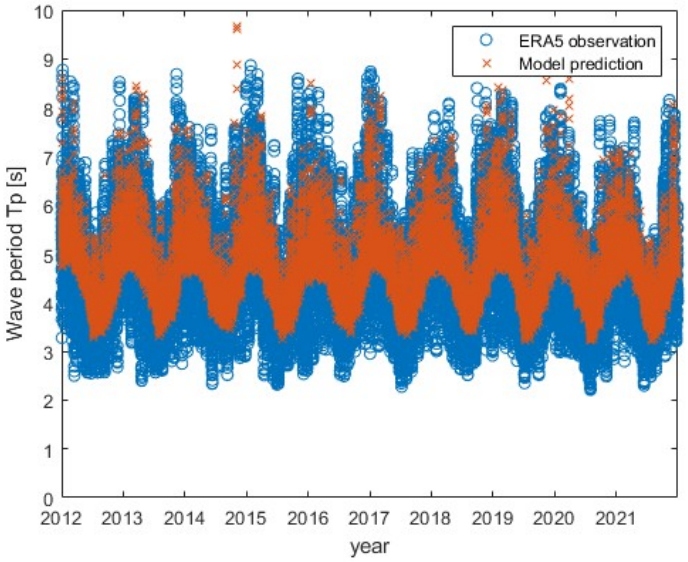


Figure 4.11: Linear multiple regression model results, mean wave period prediction.

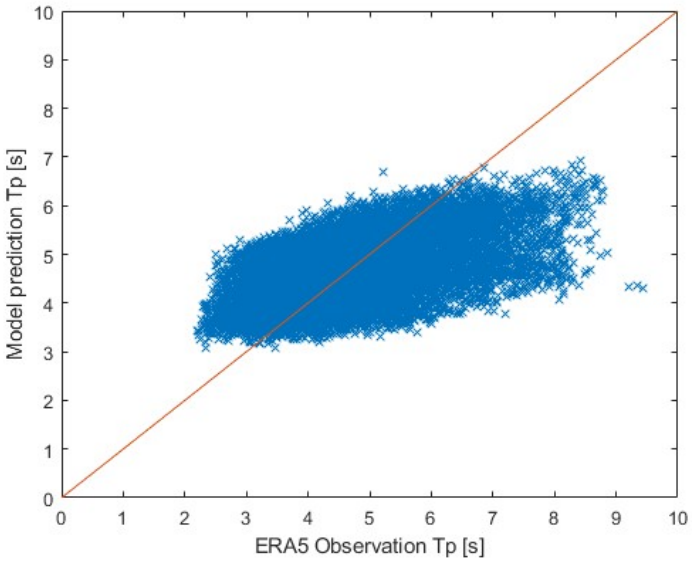


Figure 4.12: Comparison of linear multiple regression T_p prediction and ERA5 observation.

The model fits too much seasonal trend, as shown in Figure 4.11, but the predictions are poor especially for values near the minimum or maximum. It presents overestimates for low values of the period and underestimates for high values. Consequently, the accuracy decreases and the RMSE results 0.982 seconds, a more acceptable value than the SMB model but still very high. Also in this case the main reason is the non consideration of previous data as inputs.

4.1.3.2 Quadratic multiple regression model

To improve the performance of the multiple regression model, a higher order of magnitude relationship is sought, i.e., a quadratic one. In this model on the variable combinations are investigated and the coefficients are evaluated for each combination and the equation will be of the type:

$$y = \beta_0 + \beta_1 x_1 x_2 + \beta_2 x_1 x_3 + \beta_3 x_2 x_3 + \beta_4 {x_1}^2 + \beta_5 {x_2}^2 + \beta_6 {x_3}^2 \tag{4.4}$$

Regarding the wave height predictions, the third coefficient, related to the variables x_2 and x_3, presents a p-value higher than 0.05 and must be regarded as null. The resulting H_s prediction equation is:

$$H_s = 26.901 - 4.285 \cdot 10^{-4} x_1 x_2 - 4.351 \cdot 10^{-5} x_1 x_3 + 1.486 \cdot 10^{-2} {x_1}^2 + \\ + 3.135 \cdot 10^{-4} {x_2}^2 + 9.299 \cdot 10^{-7} {x_3}^2 \tag{4.5}$$

The performance improvement is huge. The RMSE results 0.134 meters, extremely low and the model can be considered very accurate. The improvement is even more clear by analyzing the figure 4.13.

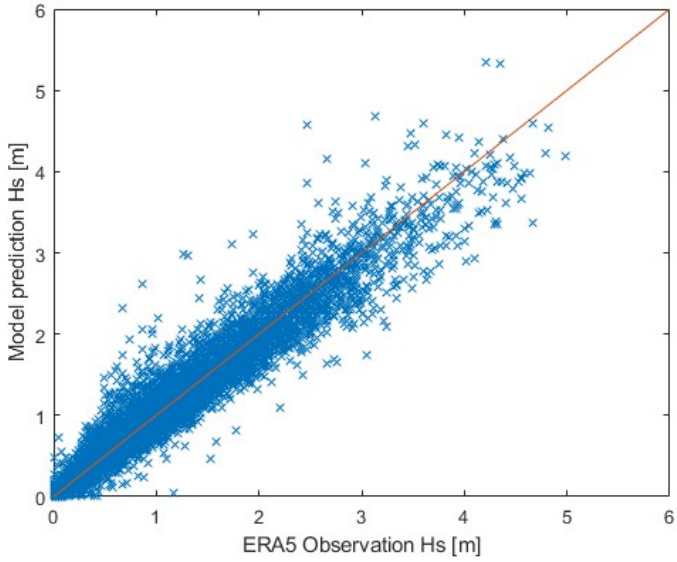


Figure 4.13: Comparison of quadratic multiple regression H_s prediction and ERA5 observation.

The data perfectly follow the bisector for both low values of Hs and high values. The spread of the points around the bisector is quite constant. That means that

the error is roughly similar whatever the value of H_s. The results show that the quadratic model is definitely better than the linear model for significant wave height forecasting. Indeed, the average error has more than halved and a prediction with a mean error of around 13 centimeters is to be considered absolutely very accurate. However, it is important to emphasize that this type of model can suffer from overfitting, thus fitting the dataset very well but having poorer performance for new data. The correlation found should therefore also be checked for future observations. If the performance is equally good, it can be concluded that the model is not overfitting and the correlation 5.2 can be used to predict wave height.
On the other hand, for the mean wave period prediction the model performances improve but not by much. The correlation found is:

$$\begin{aligned} T_p \;=\; 129.07 \;+\; 1.083 \cdot 10^{-2}\; x_1 x_2 \;-\; 8.218 \cdot 10^{-5}\; x_1 x_3 \;-\; 4.294 \cdot 10^{-5}\; x_2 x_3 \;+ \\ +\; 2.483 \cdot 10^{-2} {x_1}^2 \;+\; 9.452 \cdot 10^{-4}\; {x_2}^2 \;+\; 5.983 \cdot 10^{-6}\; {x_3}^2 \end{aligned} \tag{4.6}$$

All the coefficients present an acceptable p-value. The RMSE of the model decreases a bit compared to the linear one but remains very high, 0.875 seconds. Also in the figure 4.14 it is evident that the performance of quadratic regression in predicting T_p does not improve so much.

The forecast gets a little better for high T_p values compared to the linear multiple regression model. However, an overestimation of low T_p values and an underestimation for higher one are still present. As the linear regression, the model fits the seasonal trends very well, but the errors still remain high.
In the end, the regression model analyzed does not seem suitable for accurately predicting the period of the wave, while the accuracy on wave height forecast is very high. To improve the model, even the previous data with respect to the analyzed instant should be considered. This aspect will be implemented in the ANN model explained in the following section.

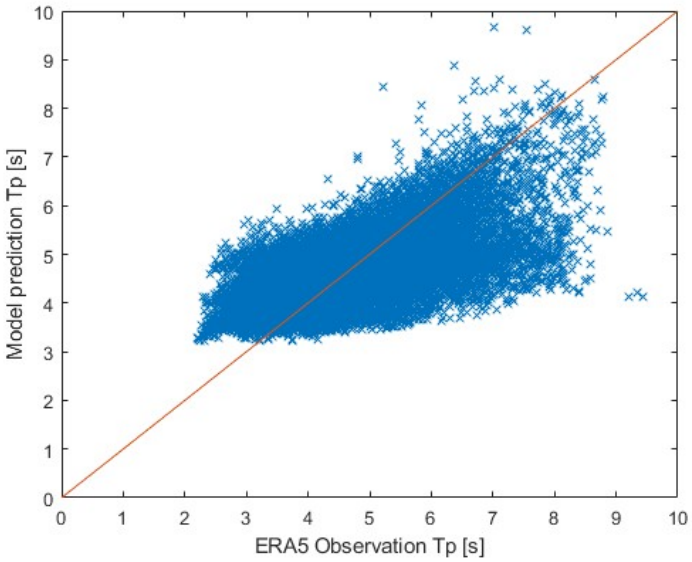


Figure 4.14: Comparison of quadratic multiple regression T_p prediction and ERA5 observation.

4.1.4 Artificial Neural Network model results

The last analyzed model is the Artificial Neural Network. The structure has been discussed in the section 3.3.3 and the model results are shown in this section. In light of the collinearity analysis conducted in the section 4.1.3, the parameters chosen as input are wind speed, wind direction and temperature.
Implementing the model, it is possible to see the error trend throughout the training and validation phase, as shown in Figure 4.15.

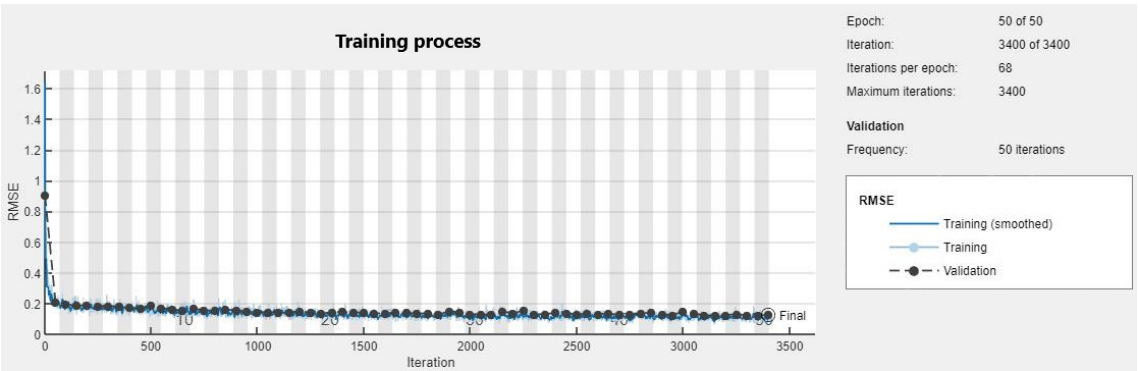


Figure 4.15: Training process neural network MATLAB tool.

On the x-axis there are the iterations. One iteration corresponds to the analysis of a batch, i.e. 128 data, as explained in the section 3.3.3. The training dataset consists of 30% of all data, corresponding to 68 batches. When all the batches are analyzed, an epoch of the training phase comes to an end. The training process stops when 50 epochs are reached, which is when the whole training dataset has been analyzed 50 times.
At the same time as the training process, validation is also performed and the error is tracked by comparing it to the training one. The Figure 4.15 shows that the former follows well the latter. This means that the model has been applied correctly.
Since the model initialization is random, the last RMSE value after 50 epochs varies slightly from process to process. For this reason, multiple tests are run and an average of the final vales is calculated to derive the accuracy of the model (see table 4.2).
A huge advantage of this model is that it can take into account previous data as inputs to calculate the future instant. So an analysis was carried out. The purpose was to see whether that led an improvement and if so, how many past events should be considered to achieve the best performance, i.e., the lowest error.
The first variable analyzed is the significant wave height and the RMSE values of training process for different model configurations are summarized in Table 4.2.

The best model configurations is considering the observations immediately preceding to perform the prediction. Considering that time between two consecutive observations are three hours, it can be concluded that even with great changes in wind and weather conditions, the previous observed values influence the subsequent sea state. This thesis finds further confirmation by analyzing even earlier data, thus six and nine hours earlier. Such a long time is sufficient to have a complete development of the new conditions, and thus observations from many hours back are redundant and worsen the performance of the model.
The improvement compared to not considering any past event is really huge. Indeed, the RMSE is reduced to about one third (from 0.40 meters to 0.13 meters). With such a low error, the ANN model analyzed is to be considered extremely accurate. Moreover, the model also has no overfitting. The RMSE of the testing process, which does not analyze the model construction data but new ones, is very similar to the RMSE of the training. Therefore, the ANN does not adapt to the construction

Table 4.2: RMSE of training process for different ANN model H_s forecasting configurations.

Attempt	**RMSE with no previous event considered [m]**	**RMSE with 1 previous event considered [m]**	**RMSE with 2 previous event considered [m]**	**RMSE with 3 previous event considered [m]**
1	0.42925	0.11673	0.1802	0.1837
2	0.39781	0.1384	0.18058	0.18034
3	0.36441	0.12608	0.19803	0.17112
4	0.41939	0.13318	0.17849	0.18853
5	0.38061	0.13405	0.19117	0.17384
mean	0.39935	**0.12969**	0.18569	0.17951

data, but maintains very good performance even in predicting new data.
Figure 4.16 shows the results of test process for the configuration with one previous observation data as inputs, that is the best one. It highlights the accuracy of the model very well: the predictions turn out to be extremely close to the ERA5 data, whatever the value of wave height (all the points are very close to the bisector).

The ANN model results very good even for the mean wave period T_p predictions. As for the H_s analysis, several model configurations depending on the previous observation considered as inputs has been compared. More than one run was performed given the random initialization, and the results are shown in Table 4.3.

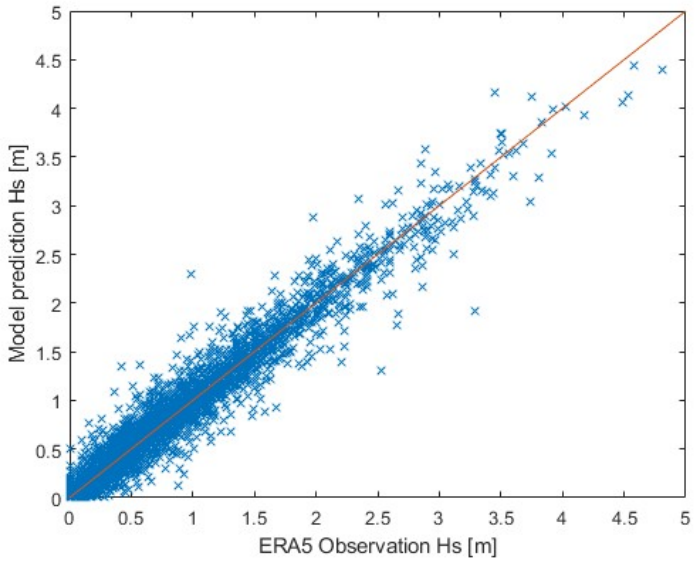


Figure 4.16: Comparison of ANN H_s prediction and ERA5 observation.

Table 4.3: RMSE of training process for different ANN model T_p forecasting configurations.

Attempt	**RMSE with no previous event considered** [s]	**RMSE with 1 previous event considered** [s]	**RMSE with 2 previous event considered** [s]	**RMSE with 3 previous event considered** [s]
1	0.9221	0.2227	0.3727	0.4623
2	0.9092	0.2269	0.3760	0.4542
3	0.8964	0.2178	0.3685	0.4462
4	0.8975	0.2213	0.3719	0.4581
5	0.9028	0.2096	0.3742	0.4608
mean	0.9056	**0.2197**	0.3727	0.4563

Considering the previous observed data as input turns out to be even more important in T_p forecast than the H_s prediction. Indeed, the RMSE is reduced to a quarter of the case without previous data, from 0.9 seconds to 0.2 seconds. This is the best results with regard to period forecast and a prediction with that accuracy is to be considered extremely precise. Also in this case, considering data from six or nine hours earlier the model performance worsens.
The test process RMSE results is very similar to the training RMSE and also in this case the model does not suffer from overfitting.

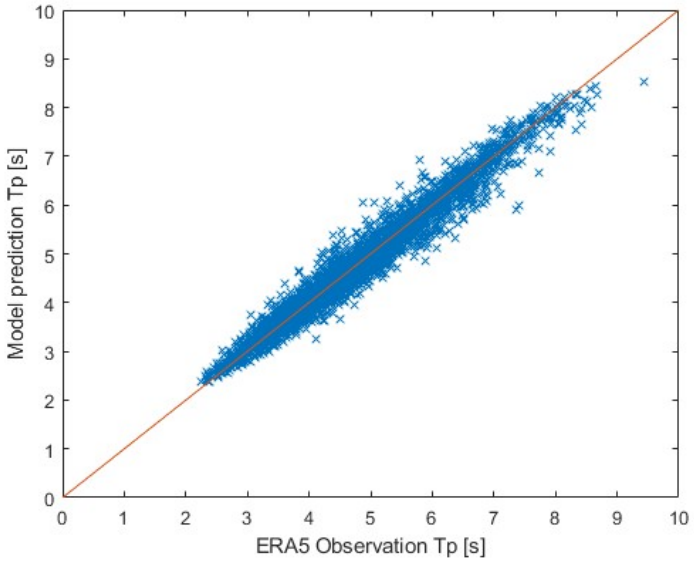


Figure 4.17: Comparison of ANN T_p prediction and ERA5 observation.

4.2 Comparison of the different model results

The three wave prediction models described in this analysis present results sometimes very different and sometimes comparable. In this section they are compared to make further observations and to verify the goodness of the models.
The Figure 4.18 shows the model results previously reported.

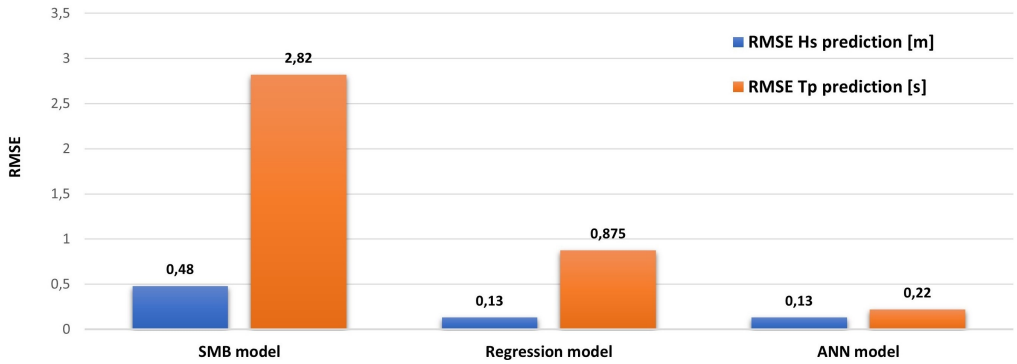
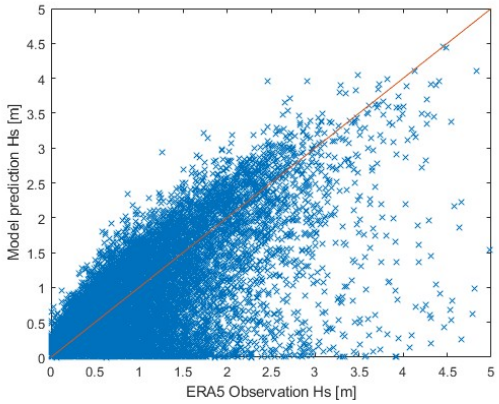


Figure 4.18: Model performances with ERA5 data implementation.

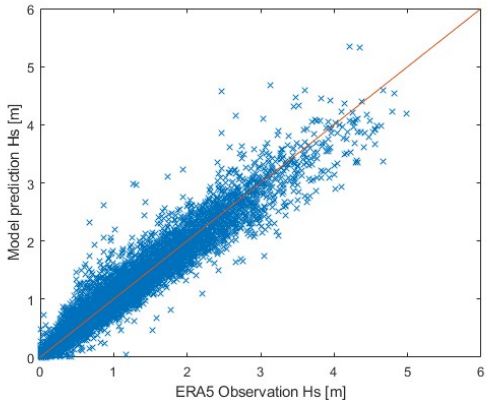
4.2.1 Comparison of H_s forecasts

In Figure 4.19 the model outputs related to wave height predictions are shown again.

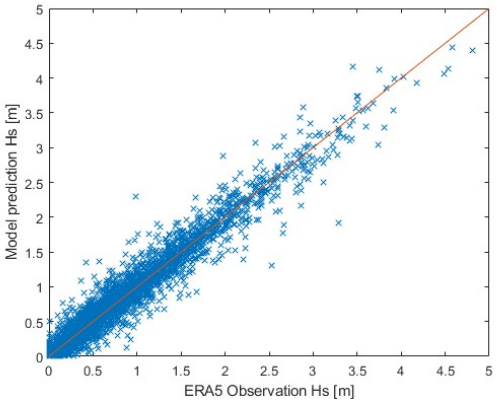
The models with the best accuracy are the ANN and the quadratic multiple regression. They present a RMSE respectively of 0.130 meters and 0.134 meters. Considering that the results of the ANN model have small fluctuations due to the randomization of the initial data, it can be concluded that the performance is practically identical. While the SMB model has a RMSE much higher, around 0.48 meters.
It should be emphasized, however, that the extreme accuracy of the ANN model is to be considered more acceptable than the regression model. Indeed, the former was constructed in such a way to avoid overfitting as much as possible by randomizing the time sequence and analyzing the dataset part by part, making the results as general as possible. The same cannot be said for the regression results. The correlation found should be verified for new data. If the predictions of the new data have more or less the same error, the outputs of the regression model are completely reliable, otherwise the regression model would suffer from overfitting and a new relationship should be calculated.
Analyzing the figures 4.19b and 4.19c, it can be observed that the spread of points with respect to the bisector is larger for the regression model than for the ANN, especially at higher wave heights. This means that in the former there are predictions with higher error and other predictions that are more accurate. The mean error



(a) SMB model



(b) Quadratic regression model



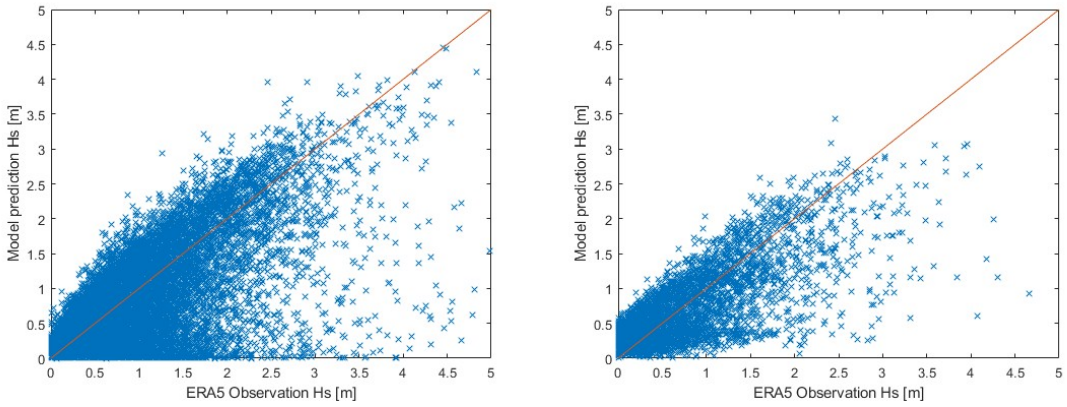
(c) ANN model

Figure 4.19: H_s prediction of different models.

does not vary but, in general, a model with a more consistent error and thus a lower spread of outputs is preferred, in this case it is the ANN model.

The SMB presents the worst results and Figure 4.19a shows that a lot of points are far from bisector, so they are very different from the observed data. In particular, most of the predictions with high error are underestimates (points below the bisector). In some cases, wave height forecasts are almost zero when the actual observations are also very high (points near the x-axis). As described in section 4.1.2, the main reason is that in this model previous data are not considered as input to calculate H_s. However, the goodness of the model is confirmed by comparing it with the ANN model. Indeed, if previous data are not considered even in the latter, the

obtained results are very similar, as shown in Figure 4.20



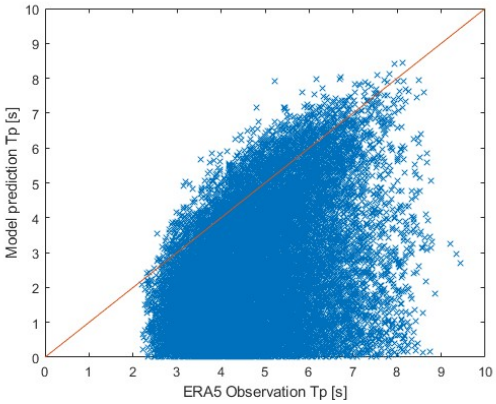
(a) SMB model (b) ANN with no previous data

Figure 4.20: H_s prediction of SMB and ANN model with no previous data considered.

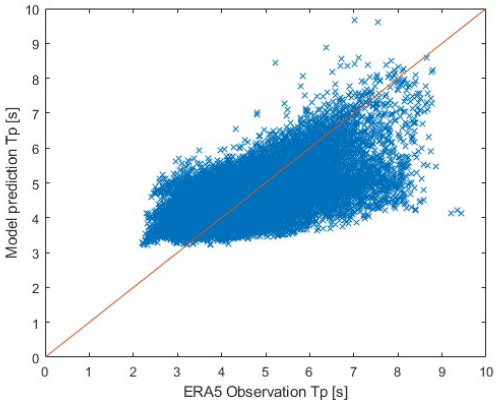
Even in this configuration of the ANN model there are a lot of underestimated points far from bisector. They are less spread out, but this is due to the greater complexity of the model, compared to an empirical one. The ANN adjusts itself according to the results and, consequently, the performance is slightly better. However, the overall RMSE are comparable, around 0.48 meters for the SMB and around 0.40 meters for the ANN. Even the regression model does not consider previous data, but the results are far more accurate, as said its RMSE is around 0.134 meters. This could indicate the presence of overfitting in that model.
For what has been described, it can be concluded that the best model to perform wave height prediction is ANN. It is extremely accurate and the reliability of the results is very high given the structure of the model, which aims to eliminate overfitting as much as possible.

4.2.2 Comparison of T_p forecasts

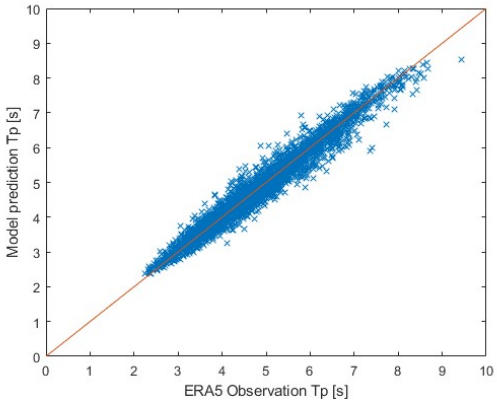
Regarding the mean wave period, the results of the three models are reported in figure 4.21.



(a) SMB model



(b) Quadratic regression model



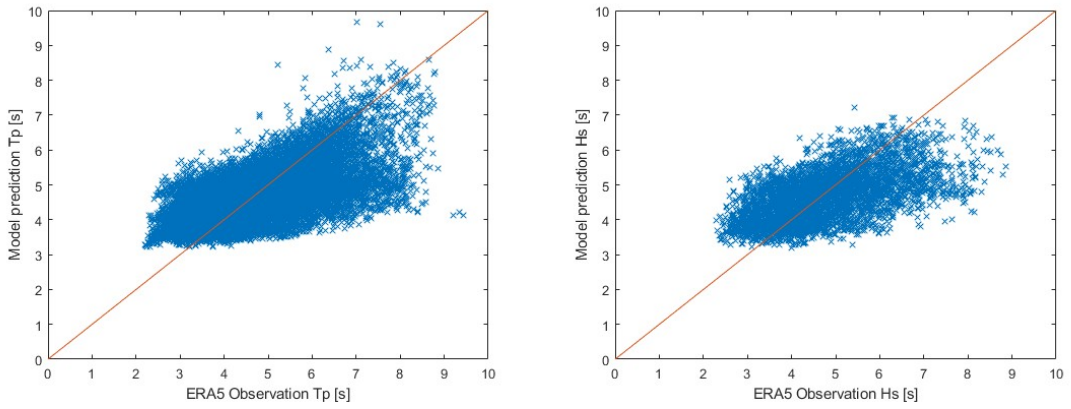
(c) ANN model

Figure 4.21: H_s prediction of different models.

It is evident that the only model that can provide accurate estimates of T_p is the ANN. The SMB model suffers from a strong underestimation of the data, even reporting in many cases values close to zero, when the observed values never have a period below two seconds. As repeatedly pointed out, the main reason is the disregard of previous instants, which result even more important for the T_p prediction than the H_s. The RMSE of this model is around 2.82 seconds, an extremely high value considering that T_p varies roughly between two and eight seconds. For this reason, the results are to be considered unreliable, and the empirical model does not seem suitable for predicting the wave period at this site.

Also the quadratic multiple regression model, which does not present the previous

data as inputs, has quite bad performance. There are no more predictions below two seconds but there is a strong overestimation for low T_p (points above bisector, Figure 4.21b) and a strong underestimation for high T_p (points below figure bisector, Figure 4.21b). This results in improved performance compared with the SMB model, and the RMSE is lowered to 0.88 seconds. Despite that, the predictions provided by the regression are not very accurate. Consequently, the model results far better in forecasting wave height than period.
Even for the ANN model, if the previous data are not considered, the performance is lower and comparable with the regression model (see Figure 4.22).



(a) Regression with no previous data

(b) ANN with no previous data

Figure 4.22: T_p prediction of regression and ANN model with no previous data considered.

The RMSE is more or less the same, around 0.9 seconds, considering that for ANN there is a randomization of the initial data and thus the results vary slightly for each run of the model. The points are less spread in the ANN, but the trend is very similar and there is always an overestimate at low T_p and an underestimate at high T_p. This is further evidence that considering the previous instant is critical for correct prediction of the data.
Indeed, the accuracy increases dramatically when previous data are implemented in the ANN model as inputs (Figure 4.21c). This turns out to be the configuration with the most acceptable results and the RMSE is reduced to a bit more than 0.2 seconds.

Chapter 5

Buoy data implementation and model performance analysis

The results presented in the previous chapter were obtained through the analysis of ERA5 data. These data are supplemented with some actual measurements from different stations, but for the most part they are derived by applying a physical model called ECMWF (see section 3.2.3). The goals of this chapter are to verify the accuracy of the ERA5 data for the selected site by comparing them with those acquired from a buoy off the Pantelleria coast, and then to implement the buoy data into the previously constructed models to analyze the performance.
Buoy data are not available for the entire duration of the period analyzed (2012 to 2021), but only for a few months. These data will then be compared with the corresponding ones given by ERA5 in order to compare the two acquisition methods.

5.1 Buoy data description

The new data are direct measurements from a subsurface buoy with an Acoustic Wave and Current Profiler (AWAC), developed by Nortek [61].

The main advantage of a subsurface buoy is to have the instrument close enough to the surface to make high-quality wave measurements e to be safe from the dangers of surface exposure.

(a) (b)

Figure 5.1: AWAC subsurface buoy.

Source: https://www.nortekgroup.com/products/awac-600-khz

To estimate the non-directional wave characteristics (e.g. wave height and wave period), the device uses the Acoustic Surface Tracking (AST). This technology provides accurate measurements of the distance between the AWAC and the water surface. The method used to detect the surface can be divided in simple sequential steps [62]:

1. transmit a pulse of a given length;
2. specify a reception window that covers the range of all possible wave heights;
3. discretize the reception window into multiple cells (e.g. of 2.5 cm);
4. apply a matching filter over a series of cells to detect the surface;
5. use quadratic interpolation to accurately estimate the location of the surface.

Once the time series of the surface has been obtained, the traditional zero-upcrossing method is applied to estimate the extreme waves, while spectral methods are used for all other wave.
The tests conducted to compare AST with other technologies have shown that the

former results a robust technique and works very well with the AWACs deployed on a subsurface buoy [61]. Moreover, AST does not suffer from the attenuation effects associated with increasing depth.
The AWAC-AST buoy is located in the North-West of the Island at about 700 meters from the coastline. The exact coordinates are 36°49' N and 11°55' E.
However, the data available from the buoy are limited and cover only a few months: from 20th April to 24th August 2015; from 26th September to 4th December 2015; from 18th July to 16th September 2019. These data will be compared with the corresponding data given by ERA5, and then the models will be tested by implementing the new data and comparing the results with those previously obtained.

5.2 Buoy data implementation and model results

The first step is to compare the buoy data with ERA5 ones to find any differences. As mentioned, the data recorded by the buoy are for only a few months for a total of just over two thousand measurements. However, this number is sufficient to compare the two different methods of data acquisition and implement them in the three prediction models previously constructed.
The Figure 5.2 shows the results of ERA5 and buoy data comparison.

The Figures 5.2a and 5.2b are related to the significant height of the wave. The first shows the data series from April to August 2015, while the second compares all available data with a scatter plot. Instead, the Figures 5.2c and 5.2d are related to the mean wave period.
It is evident that the ERA5 data for both H_s and T_p are severely underestimated when compared with the buoy readings (in the scatter plots the most of the points are above the bisector). The underestimation increases as H_s and T_p values increase. This shows that it would be very important to have more than one source of a site's data, so they can be compared and a more accurate analysis can be conducted. Obviously buoy records are the best option, but very often there are no buoys at or near the site under consideration, and installation costs can be very high. In addition, it would be necessary to wait as long as it takes for enough data to be recorded, in contrast to a dataset such as ERA5 that makes data available even from many years.

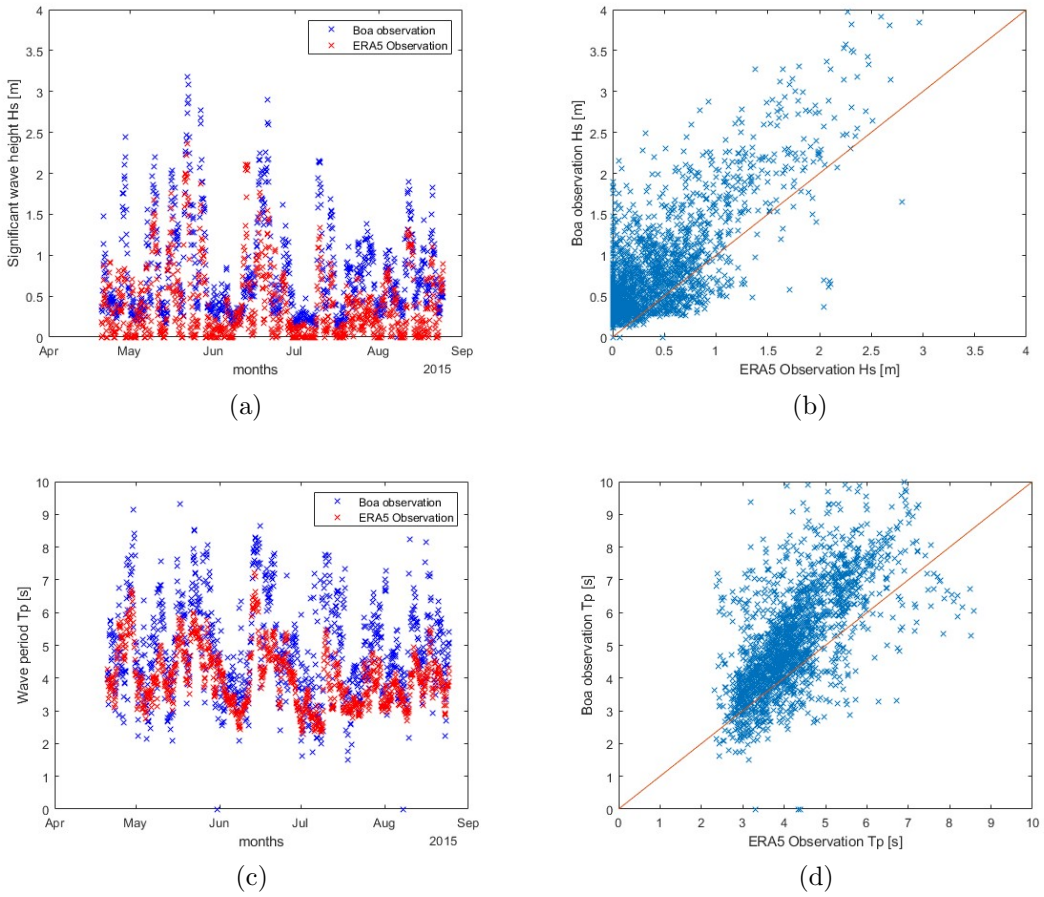


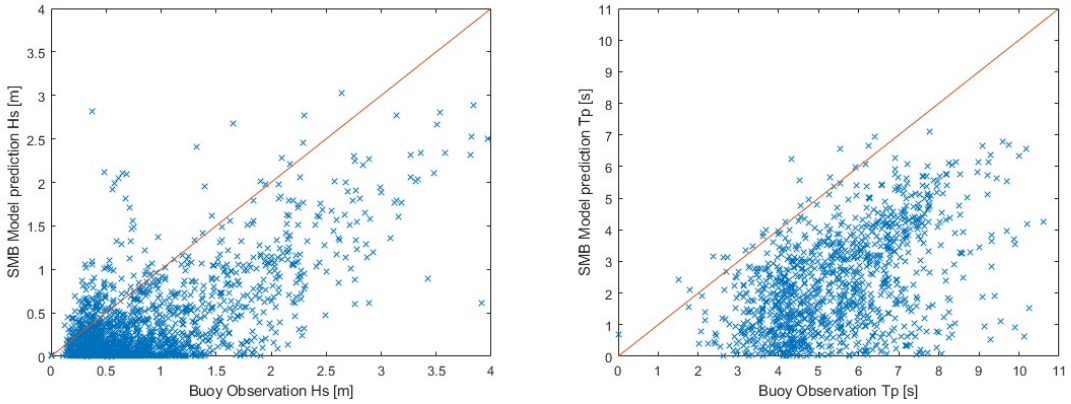
Figure 5.2: Comparison between buoy and ERA5 data

It can be concluded, therefore, that data from a buoy are the best and certainly the most accurate, but in the absence of these, where one must rely on other sources, it is important to consider any underestimation in order to conduct analysis as accurately as possible. However, it is important to note that in other spots ERA5 data may be more accurate and not present such a marked difference from on-site surveys.

Then, AWAC-AST buoy data has been implemented in the three constructed models (empirical, regression, and ANN) to evaluate the new performances.

5.2.1 SMB model results with buoy data implementation

The results of the SMB model are shown in figure 5.3.



(a) H_s predictions (b) T_p predictions

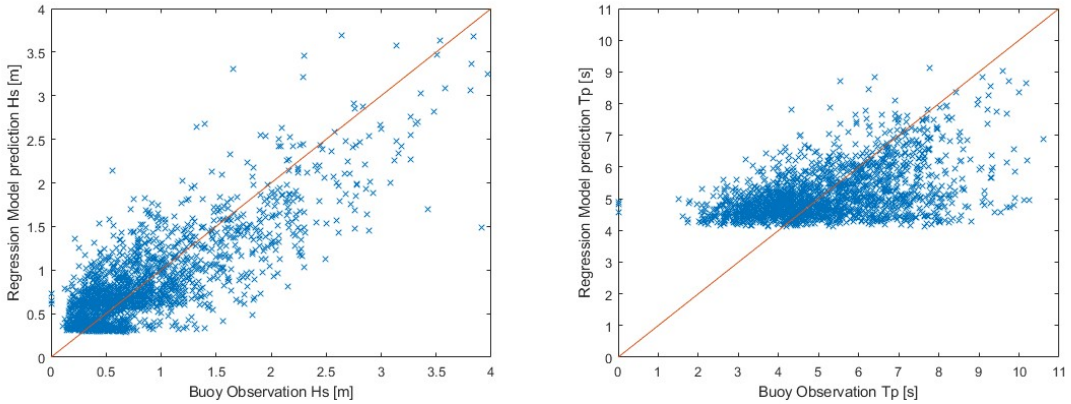
Figure 5.3: SMB model performances with buoy data implementation.

The equations for predicting wave height and period are unchanged. It is clear that in this case the empirical model is not adequate and the predictions provided are severely underestimated. As for the wave height, in many places and predictions provided are about zero, while the buoy readings turn out to be a few tens of centimeters up to a meter and a half (data close to the x-axis in figure 5.3a). Also for the period forecasting, there is a total underestimation. Almost all points result below the bisector, Figure 5.3b.

5.2.2 Regression model results with buoy data implementation

The quadratic multiple regression model is applied and the results are presented in Figure 5.4.

Regarding the wave height prediction, the Figure 5.4a shows an improvement compared with the empirical model. The points follow the bisector fairly well, even though their spread is high and the error with respect to the buoy detections is



(a) H_s predictions (b) T_p predictions

Figure 5.4: Quadratic multiple regression model performances with buoy data implementation.

still significant. In fact, the RMSE is around 0.36 meters. The resulting regression equations for the H_s forecasts is:

$$H_s = 116.26 - 4.378 \cdot 10^{-3}\, x_1 x_2 + 5.792 \cdot 10^{-4}\, x_1 x_3 + 5.671 \cdot 10^{-3}\, {x_1}^2 + \\ + 1.282 \cdot 10^{-4}\, {x_3}^2 + 8.019 \cdot 10^{-6}\, {x_3}^2 \quad (5.1)$$

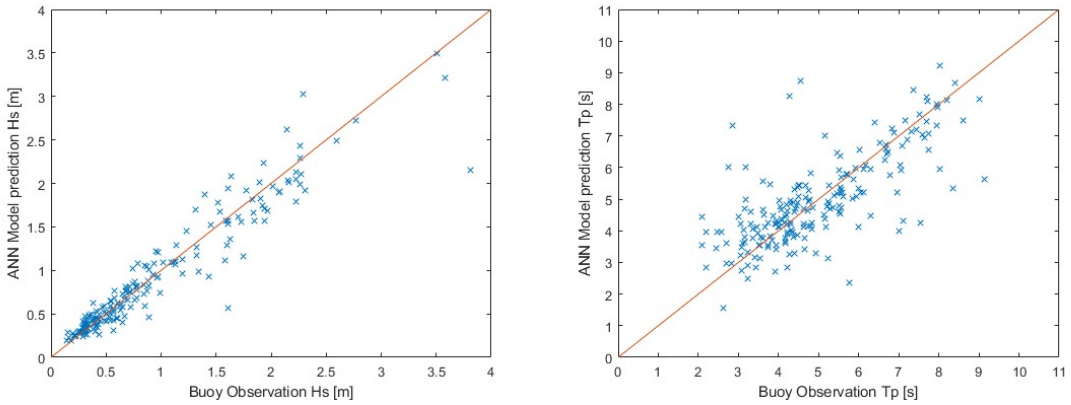
where x_1 is the wind speed, x_2 is the air temperature and x_3 is the wind direction. All the coefficients in the equation are acceptable and have a p-value less than 0.05 except for the one for the variable "$x_1 x_3$" which, indeed, is not present in the equation 5.1. Analyzing the T_p predictions of the quadratic regression model (Figure 5.4b), the model does not seem very suitable, and the RMSE is about 1.5 seconds. As discussed in the section 4.1.3 this is mainly due to the lack of previous data as model inputs.
The resulting T_p regression equation is:

$$T_p = -7.193 - 1.826 \cdot 10^{-3}\, x_1 x_2 + 1.636 \cdot 10^{-3}\, x_1 x_3 + 6.314 \cdot 10^{-3} {x_1}^2 + \\ + 6.074 \cdot 10^{-6}\, {x_3}^2 \quad (5.2)$$

All omitted coefficients have an unacceptable p-value. However, the error of this model is too high to consider its results as significant.

5.2.3 ANN model results with buoy data implementation

The last studied model is the artificial neural network. The best configuration is used, so the one in which data from the previous observation are considered as model inputs. Since the dataset available from the buoy is around two thousand observations, far less than the previous dataset, the training validation and test phase percentages have been reset to obtain a more robust training phase. Thus 70% of the data was used for training, 20% for validation, and 10% for testing. The total structure of the model is the same as described 3.3.3. The buoy data are implemented and the result are shown in Figure 5.5.



(a) H_s predictions (b) T_p predictions

Figure 5.5: Quadratic multiple regression model performances with buoy data implementation.

Also in this case, the ANN model provides the best predictions for both wave height and period. For H_s prediction (Figure 5.5a), the model is very accurate. The points are all very close to the bisector, especially for low values of H_s, which means that the error between prediction and buoy detection is very small. While for T_p forecast, the ANN model is a bit less accurate and to points are a little more spread

out and further away from the bisector (Figure 5.5b), but still maintaining the right trend. Therefore the predictions for the period are to be considered significant and can provide a general idea of real values. Certainly the ANN turns out to be the best model in terms of T_p forecasting.
As discussed in section 4.1.4, the accuracy slightly depends on the initialization values that are randomized. So several models run have been done and an average RMSE has been calculated. For the H_s the RMSE results around 0.18 meters and for the T_p around one second.
All the model performances are briefly summarized in the Figure 5.6.

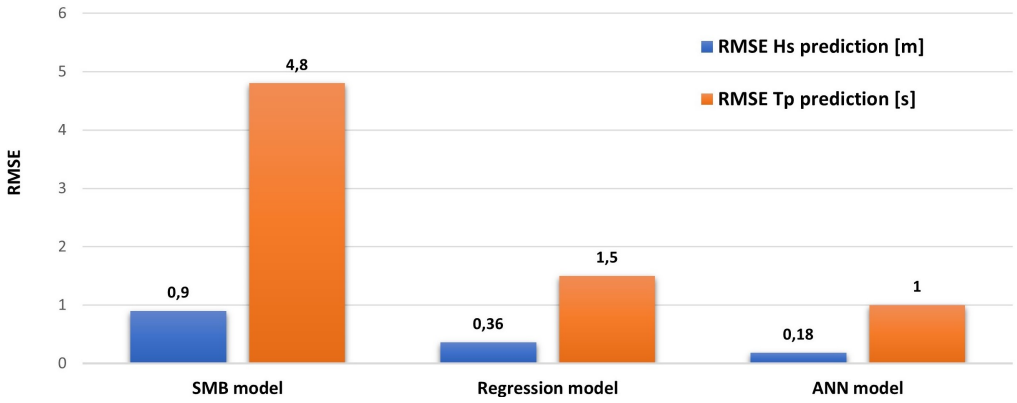


Figure 5.6: Model performances with buoy data implementation.

5.3 Model performance comparison between ERA5 data and buoy data

In this last section, the results obtained from the buoy data are compared with those obtained using ERA5 data. The goal is to test whether and how much the performance of a model varies when data from different sources are implemented.
Analyzing the empirical SMB model, the performances drop dramatically by implementing buoy data. Indeed the RMSE for the H_s predictions increases from an initial 0.48 meters with ERA5 data to about 0.9 meters with buoy observations, almost a doubling of the average error, mainly due to the large underestimation of outputs.

The empirical equations that constitute this model are general in nature and do not vary from one context to another. Obviously, each site may have profoundly different characteristics, and variables that influence wave characteristics may not be considered in this model. For this reason, the SMB may have good results in some studies and less significant results in others. Similarly, ERA5 dataset is based on a physical model and on solving of specific equations.
Probably, given the purely theoretical nature, the equations of the SMB empirical model and the equations for calculating the ERA5 data are quite related. This makes the SMB fairly accurate when using such data as inputs. It is important to note that a model for calculating wave characteristics can be extremely complex, while the SMB greatly simplifies that complexity. For this reason, the results will never be extremely accurate, but when it works it can provide a general estimate and range of the actual data.
Regarding the prediction of the period, in no case can the SMB be considered reliable. However, the results get much worse considering the buoy observations and the RMSE increases from about 2.8 seconds to 4.8 seconds.
Talking about the regression model, again the results are less accurate by implementing the buoy data. The RMSE related to H_s increases from an extremely low value of 0.13 meters to 0.36 meters. The error practically triples, and the main reason is that the predictions are greatly underestimated compared to the first model. There may be several reasons for this.
First, the dataset provided by the buoy is much smaller than that analyzed previously (more than 10 times smaller). This makes the training process less effective since there is fewer data to analyze and undoubtedly the error may increase.
Also, as repeatedly stated in paragraph 4.1.3, the regression model built on the ERA5 data may suffer from overfitting, thus having extreme accuracy for the analyzed data but being less accurate with predictions of new data. In fact, no real countermeasure was taken in this model, and such a large amount of data could lead to this problem.
A further factor might be the lack of previous data as inputs, which in the case of buoy measurements might acquire even more importance. Probably the loss of accuracy of the model is not to be attributed to one cause but to a mix of them.
For the period forecast, the regression model already presented insignificant results

with a RMSE of 0.87 seconds, and by implementing the buoy data, this comes to 1.5 seconds. So it can be concluded that this model is totally unsuitable for T_p forecasting.
The last model is the ANN that presents the best results by implementing buoy measurements. The Figure 5.7 compares model results with ERA5 data to those with buoy data.

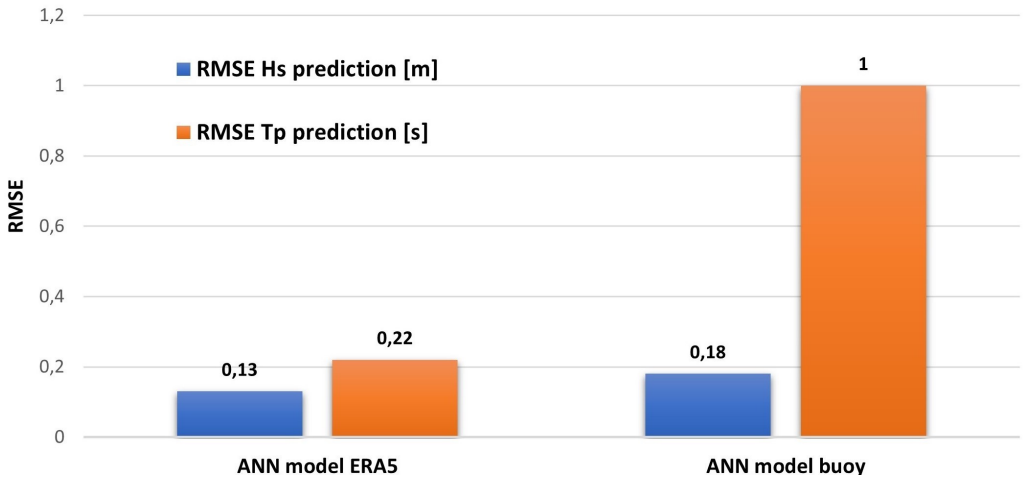
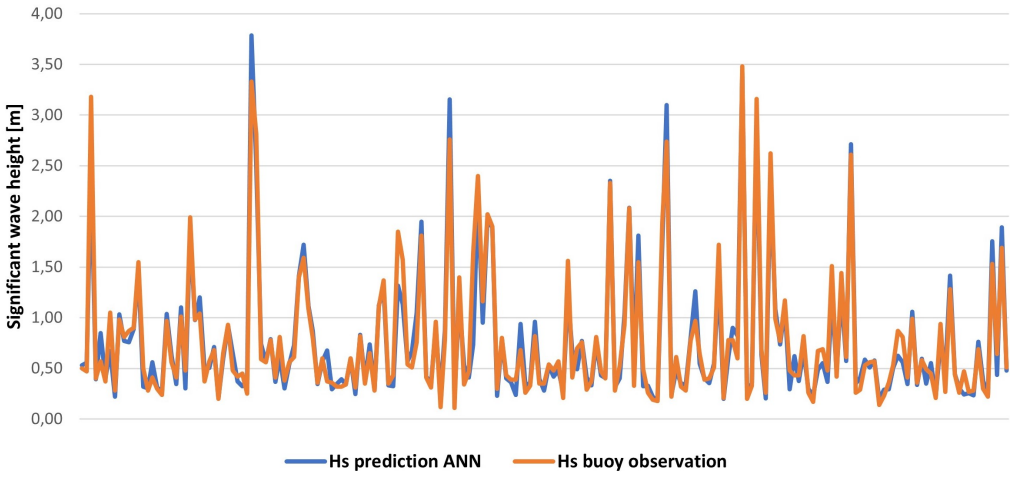


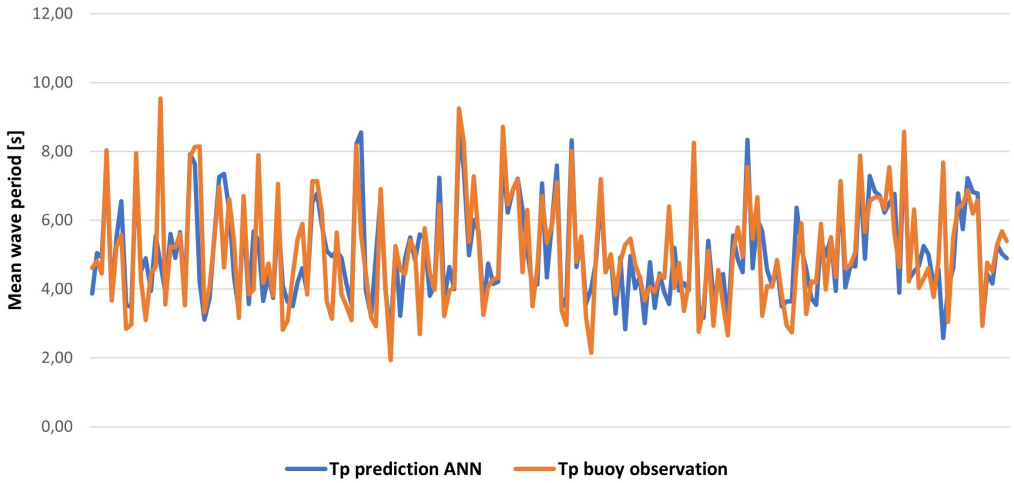
Figure 5.7: Comparison ANN model results between with ERA5 and buoy data implementation.

The wave height predictions remain extremely accurate although the error increases slightly from about 0.13 meters to 0.18 meters. Probably the main reason is the much smaller number of data implemented in the model, which makes it slightly less precise.
Regarding the T_p, the forecast error increases dramatically from 0.18 with ERA5 data to about one second with the new buoy data. Despite the very high error, the forecasts remain fairly consistent with the buoy observations (see Figure 5.5b). The points are more or less all spread around the bisector, and there is no area with high underestimation or overestimation of the predictions.
The goodness of the predictions provided by this model can be clearly seen by analyzing from the graphs in Figure 5.8.

Wave height predictions follow the observations provided by the buoy extremely well. The largest errors occur at some peaks where measurements differ by a few



(a) H_s



(b) T_p

Figure 5.8: ANN model predictions with respect to buoy observations.

tens of centimeters.

Regarding the mean wave period prediction, more pronounced differences can be observed from the actual value. However, overall the model follows the trend of the buoy data quite well.

In the end, the ANN model turns out to be the best considering the buoy data: for the height prediction it remains extremely accurate, while for the period prediction, despite the higher RMSE, it can provide a significant indication of the period. Certainly by increasing the number of observations the performances of the model

would be further increased.

Chapter 6

Conclusions

The wave energy resource represents a key asset for renewable energy production in the future. Research and development of increasingly efficient technologies to maximize the exploitation of this resource is one of the goals of the energy transition. Improved wave forecasting can also be crucial in this process. Indeed, it results extremely important during the planning of a farm to determine the potential production and consequently to plan the size and construction. Moreover, the wave forecasting is also important during the farm operation to greatly improve the efficiency of a wave energy converter and to predict the energy generated for managing and optimizing the energy mix.
The aim of the study is to test three different methods for the prediction of wave characteristics, knowing the atmospheric parameters, i.e., wind speed, wind direction and temperature. The considered models are an empirical one called Sverdrup-Munk-Bretschneider (SMB), a multiple regression model and a machine learning Artificial Neural Network (ANN). The site selected for analysis is one of the most pioneering sites for marine energy research and exploitation in Italy, i. e., the Island of Pantelleria. The implemented data have been obtained from ERA5, the fifth generation ECMWF atmospheric reanalysis of the global climate.
The results show that the best model is the ANN for both the significant wave height prediction and the mean wave period. The model RMSEs are respectively 0.13 meters and 0.22 seconds. In addition, due to the specific way the ANN has been built, it does not suffer from overfitting, a typical problem of machine learning

models.
The multiple regression shows the best results with a quadratic relationship between the dependent variable and the independent variables, especially regarding the wave height. The corresponding RMSE of the predictions is around 0.13 meters, like the ANN model. It is important to note that such a model may suffer from overfitting. For the mean wave period, the forecasts present high errors (RMSE around 0.9 seconds) and they are to be considered not very significant.
The SMB is the simplest method and, despite that, has acceptable results for the H_s predictions. The RMSE is around 0.48 meters, a fairly large value, but the empirical model can be a good tool for studying wave height trends. Regarding the period prediction the SMB proved to be totally unsuitable, managing to roughly predict only annual trends but presenting a very high RMSE of 2.8 seconds.
Then, the three models are compared implementing the data an AWAC-AST buoy installed on the site. For the H_s prediction the RMSEs are 0.9 meters for the SMB, 0.36 meters for the regression and 0.18 meters, for the ANN. Regarding the T_p forecasts, RMSE values are of 4.8 seconds for the SMB, 1.5 seconds for the regression and around one 1 second for the ANN.
The performance of all three models worsens. The SMB has too high errors, and the results totally lose significance. The regression model has acceptable predictions only for the wave height, but the error is three times larger than in the previous case. Only the ANN presents significant results. H_s increases by only a few centimeters compared with the ERA5 data implementation, while for the period, even if the results are no so accurate, they are still quite significant since they follow the buoy observations fairly well. Probably results so good are because it is the only model that takes into account data from previous forecasts as input to predict the future instant.
The loss of performance can also be attributed to fewer data available from the buoy compared with ERA5. Indeed, the measurements available from the buoy are for only a few months, with a total of just over 2,000 data, while the data previously obtained amounted to more than 29,000.
Certainly, more observations allow the regression model and ANN to reduce errors. Therefore, it would be appropriate to recalculate the performances when more measurements are available from the buoy.

Moreover, it is important to emphasize that the results and relationships found are correlated with the selected site. The ANN and regression models intrinsically fit the data that will certainly depend on site characteristics. The results obtained are therefore specific to Pantelleria. If other sites have to be analyzed, the found regression relationships should be recalculated or at least the accuracy should be checked for the new spot, while the ANN model should be rerun with the new data.

Bibliography

[1] Andrzej Bielecki et al. "The externalities of energy production in the context of development of clean energy generation". In: *Environ. Sci. Pollut. Res.* 27.11 (2020), pp. 11506–11530. ISSN: 16147499. DOI: 10.1007/S11356-020-07625-7/TABLES/2.

[2] REN21. *Renewables 2022 Global status report*. 2022. ISBN: 9783948393045.

[3] Frank R. Spellman. *Environmental impacts of renewable energy*. CRC Press, 2014, pp. 1–458. ISBN: 9781482249477. DOI: 10.1201/B17744.

[4] European Commission. *Renewable energy targets*. URL: https://energy.ec.europa.eu/topics/renewable-energy/renewable-energy-directive-targets-and-rules/renewable-energy-targets_en (visited on 08/15/2022).

[5] Paul Saunders. *Land Use Requirements of Solar and Wind Power Generation: Understanding a Decade of Academic Research.-Japan-South Korea cooperation on energy technology View project Land Use-Solar and Wind Power View project*. 2020. ISBN: 9781735933504. URL: https://www.researchgate.net/publication/345638945.

[6] Iñigo Capellán-Pérez, Carlos de Castro, and Iñaki Arto. *Assessing vulnerabilities and limits in the transition to renewable energies: Land requirements under 100% solar energy scenarios*. 2017. DOI: 10.1016/j.rser.2017.03.137.

[7] Enel Green Power. *Marine energy*. URL: https://www.enelgreenpower.com/learning-hub/renewable-energies/marine-energy (visited on 09/21/2022).

[8] European Commission. *EU strategy on offshore renewable energy*. URL: https://energy.ec.europa.eu/topics/renewable-energy/offshore-wind-and-ocean-energy_en (visited on 08/15/2022).

[9] Luca Liberti, Adriana Carillo, and Gianmaria Sannino. "Wave energy resource assessment in the Mediterranean, the Italian perspective". In: *Renew. Energy* 50 (2013), pp. 938–949. ISSN: 09601481. DOI: 10.1016/j.renene.2012.08.023.

[10] Muhammed Zafar, Haider Ali Khan, and Muhammad Aziz. "Harvesting Energy from Ocean: Technologies and Perspectives". In: *Energies 2022, Vol. 15, Page 3456* 15.9 (2022), p. 3456. ISSN: 1996-1073. DOI: 10.3390/EN15093456. URL: https://www.mdpi.com/1996-1073/15/9/3456/htmhttps://www.mdpi.com/1996-1073/15/9/3456.

[11] P. Pinson, G. Reikard, and J. R. Bidlot. "Probabilistic forecasting of the wave energy flux". In: *Appl. Energy* 93 (2012), pp. 364–370. ISSN: 03062619. DOI: 10.1016/J.APENERGY.2011.12.040.

[12] E. B.L. Mackay. "Resource assessment for wave energy". In: *Compr. Renew. Energy* 8 (2012), pp. 11–77. DOI: 10.1016/B978-0-08-087872-0.00803-9.

[13] Alessandro Toffoli and Elzbieta M. Bitner-Gregersen. "Types of Ocean Surface Waves, Wave Classification". In: *Encycl. Marit. Offshore Eng.* (2017), pp. 1–8. DOI: 10.1002/9781118476406.EMOE077.

[14] Dolores Esteban, José-Santos López-Gutiérrez, and Vicente Negro. "Classification of Wave Energy Converters". In: *Recent Adv. Petrochem Sci.* 2.4 (2017). DOI: 10.19080/RAPSCI.2017.02.555593.

[15] Amélie Têtu and A Têtu. "Power Take-Off Systems for WECs". In: (2017), pp. 203–220. DOI: 10.1007/978-3-319-39889-1_8. URL: https://link.springer.com/chapter/10.1007/978-3-319-39889-1_8.

[16] Guang Li et al. "Wave energy converter control by wave prediction and dynamic programming". In: *Renew. Energy* 48 (2012), pp. 392–403. ISSN: 09601481. DOI: 10.1016/j.renene.2012.05.003. URL: http://dx.doi.org/10.1016/j.renene.2012.05.003.

[17] ECMWF. *Ocean wave forecasting*. 2020. URL: https://www.ecmwf.int/en/about/media-centre/focus/2020/fact-sheet-ocean-wave-forecasting (visited on 10/01/2022).

[18] World Meteorological Organization. *Guide to Wave Analysis and Forecasting*. Tech. rep. 2018.

[19] Carlo Lo Re, Marcella Cannarozzo, and Giovanni Battista Ferreri. "Present-day use of an empirical wave prediction method". In: *Proc. Inst. Civ. Eng. Marit. Eng.* 169.1 (2016), pp. 3–14. ISSN: 17517737. DOI: 10.1680/jmaen.15.00006.

[20] Reale F. Dentale F., Pugliese Carratelli E. *Modellistica del moto ondoso e formazione delle onde*. Tech. rep. 2018.

[21] Sisir Kumar Patra et al. "Estimation and Validation of Offshore Wave Characteristics of Bay of Bengal Cyclones (2008-2009)". In: *Aquat. Procedia* 4.Icwrcoe (2015), pp. 1522–1528. ISSN: 2214241X. DOI: 10.1016/j.aqpro.2015.02.197. URL: http://dx.doi.org/10.1016/j.aqpro.2015.02.197.

[22] Robert H Stewart. *Introduction to physical oceanography*. Robert H. Stewart, 2008.

[23] Heinz Günther, Susanne Hasselmann, and Paem Janssen. *The WAM Model cycle 4*. Tech. rep. 1992.

[24] N. Booij, L. H. Holthuijsen, and R. C. Ris. "'SWAN' wave model for shallow water". In: *Proc. Coast. Eng. Conf.* 1 (1997), pp. 668–676. ISSN: 08938717. DOI: 10.1061/9780784402429.053.

[25] The WAVEWATCH III Development Group WW3DG. "User manual and system documentation of WAVEWATCH III version 6.07". In: *NOAA / NWS / NCEP / MMAB Tech. Note* 333 (2019), p. 311.

[26] Chris Harris. "Coupled Atmosphere-Ocean Modelling". In: *New Front. Oper. Oceanogr.* GODAE OceanView, 2018. DOI: 10.17125/gov2018.ch16.

[27] George Varlas and Petros Spyrou, Christoszadopoulos, Anastasios Korres, Gerasimos Katsafados. "One-year assessment of the CHAOS two-way coupled atmosphere-ocean wave modelling system over the Mediterranean and Black Seas". In: *Mediterr. Mar. Sci.* 21.2 (2020), pp. 372–385. ISSN: 17916763. DOI: 10.12681/mms.21344.

[28] William C Skamarock et al. *A Description of the Advanced Research WRF Model Version 4*. Tech. rep. URL: http://library.ucar.edu/research/publish-technote.

[29] Sophie Valcke Cerfacs. *OASIS3-MCT User Guide OASIS3-MCT 3.0 How to get documentation ? How to get assistance? Phone Numbers and Electronic Mail Adresses*. Tech. rep. 2015. URL: http://oasis.enes.org.

[30] *Regression analysis - Wikipedia*. URL: https://en.wikipedia.org/wiki/Regression_analysis (visited on 09/13/2022).

[31] Dilip K Barua. "Wave Hindcasting". In: *Encycl. Coast. Sci.* Ed. by Charles W Finkl and Christopher Makowski. Cham: Springer International Publishing, 2019, pp. 1859–1864. ISBN: 978-3-319-93806-6. DOI: 10.1007/978-3-319-93806-6_347.

[32] Kemal Günaydin. "The estimation of monthly mean significant wave heights by using artificial neural network and regression methods". In: *Ocean Eng.* 35.14-15 (2008), pp. 1406–1415. ISSN: 00298018. DOI: 10.1016/j.oceaneng.2008.07.008.

[33] Britannica. *artificial intelligence - Alan Turing and the beginning of AI*. URL: https://www.britannica.com/technology/artificial-intelligence/Alan-Turing-and-the-beginning-of-AI (visited on 09/15/2022).

[34] Tom M. Mitchell. *Machine learning*. Vol. 45. 13. 2017, pp. 40–48. ISBN: 026201243X. URL: https://books.google.ca/books?id=EoYBngEACAAJ&dq=mitchell+machine+learning+1997&hl=en&sa=X&ved=0ahUKEwiomdqfj8TkAhWGslkKHRC\\bAtoQ6AEIKjAA.

[35] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. "Statistical and Machine Learning forecasting methods: Concerns and ways forward". In: *PLoS One* 13.3 (2018), e0194889. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0194889.

[36] *Artificial neural network - Wikipedia*. URL: https://en.wikipedia.org/wiki/Artificial_neural_network (visited on 08/24/2022).

[37] Zhihua Zhang. *Artificial Neural Network Multivariate*. 2019, pp. 1–31. ISBN: 9783319673400. URL: https://neoteric.eu/blog/10-use-cases-of-ai-in-manufacturing/.

[38] Wen-Yeau Chang. "A Literature Review of Wind Forecasting Methods". In: *J. Power Energy Eng.* 02.04 (2014), pp. 161–168. ISSN: 2327-588X. DOI: 10.4236/jpee.2014.24023.

[39] *ERA5 hourly data on single levels from 1959 to present*. URL: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview (visited on 09/06/2022).

[40] Wikipedia. *Pantelleria*. URL: https://it.wikipedia.org/wiki/Pantelleria (visited on 09/20/2022).

[41] Clean energy for EU islands. "Agenda per la transizione energetica. Isola di Pantelleria". In: (2020).

[42] European Union website. *Pantelleria — Clean energy for EU islands*. URL: https://clean-energy-islands.ec.europa.eu/countries/italy/pantelleria (visited on 09/02/2022).

[43] *Different Types of Anemometers*. URL: https://sciencing.com/different-types-anemometers-8526081.html (visited on 09/03/2022).

[44] *Global Wind Atlas*. URL: https://globalwindatlas.info/about/method.

[45] *New European Wind Atlas*. URL: https://map.neweuropeanwindatlas.eu/about (visited on 09/04/2022).

[46] *Wind Data Generator - Wikipedia*. URL: https://en.wikipedia.org/wiki/Wind_Data_Generator (visited on 09/04/2022).

[47] I. Katz. "Ocean wave measurements". In: *Syria Stud.* 7.1 (2015), pp. 37–72. ISSN: 17549469. arXiv: arXiv:1011.1669v3. URL: https://www.researchgate.net/publication/269107473_What_is_governance/link/548173090cf22525dcb61443/download\%0Ahttp://www.econ.upf.edu/~reynal/Civilwars_12December2010.pdf\%0Ahttps://think-asia.org/handle/11540/8282\%0Ahttps://www.jstor.org/stable/41857625.

[48] Steve Barstow et al. "USE OF SATELLITE WAVE DATA IN THE WORLD-WAVES PROJECT". In: *Gayana (Concepción)* 68.2 (2004), pp. 40–47. ISSN: 0717-6538. DOI: 10.4067/S0717-65382004000200007. URL: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-65382004000200007&lng=en&nrm=iso&tlng=en.

[49] Danish Hydraulic Institute (DHI). "Mike 21 Wave Modelling". In: *Dhi* (2015), pp. 1–18.

[50] Alberto Fortelli et al. "Analysis of sea storm events in the mediterranean sea: The case study of 28 december 2020 sea storm in the gulf of Naples, Italy". In: *Appl. Sci.* 11.23 (2021). ISSN: 20763417. DOI: 10.3390/app112311460.

[51] Google LLC. *Google Earth Pro*. 2022.

[52] A Etemad-Shahidi, M H Kazeminezhad, and S J Mousavi. *On The Prediction of Wave Parameters Using Simplified Methods*. Tech. rep. 56. 2009, pp. 505–509.

[53] E. Mooi M. Sarstedt. *Regresion Analysis*. 2018, pp. 209–256. ISBN: 9783662567074. DOI: 10.1007/978-3-662-56707-4.

[54] Pennsylvania State University. *The Multiple Linear Regression Model*. URL: https://online.stat.psu.edu/stat462/node/131/ (visited on 09/15/2022).

[55] The Math Works. *MATLAB*. 2022.

[56] Tarang Shah. *About Train, Validation and Test Sets in Machine Learning*. URL: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7 (visited on 09/16/2022).

[57] *Activation Functions in Neural Networks*. URL: https://www.v7labs.com/blog/neural-networks-activation-functions (visited on 09/17/2022).

[58] Francesco Barbariol et al. "Wind Waves in the Mediterranean Sea: An ERA5 Reanalysis Wind-Based Climatology". In: *Front. Mar. Sci.* 8 (2021), pp. 1–42. ISSN: 22967745. DOI: 10.3389/fmars.2021.760614.

[59] S. Zecchetto and F. De Biasio. "Sea surface winds over the Mediterranean basin from satellite data (2000-04): Meso- and local-scale features on annual and seasonal time scales". In: *J. Appl. Meteorol. Climatol.* 46.6 (2007), pp. 814–827. ISSN: 15588424. DOI: 10.1175/JAM2498.1.

[60] MathWorks. *Collinetest*. URL: https://it.mathworks.com/help/econ/collintest.html (visited on 09/22/2022).

[61] Torstein Pedersen, Eric Siegel, and Jon Wood. "Directional wave measurements from a subsurface buoy with an acoustic wave and current profiler (AWAC)". In: *Ocean. Conf. Rec.* (2007). ISSN: 01977385. DOI: 10.1109/OCEANS.2007.4449153.

[62] Torstein Pedersen and Atle Lohrmann. "Possibilities and limitations of acoustic surface tracking". In: *Ocean '04 - MTS/IEEE Techno-Ocean '04 Bridg. across Ocean. - Conf. Proc.* 3 (2004), pp. 1428–1434. DOI: 10.1109/oceans.2004.1406331.