UNIVERSITÀ DEGLI STUDI GENOVA

DIPARTIMENTO DI SCIENZE DELLA TERRA,
DELL'AMBIENTE E DELLA VITA
CORSO DI LAUREA MAGISTRALE IN SCIENZE GEOLOGICHE

EXPLORING THE USE OF MACHINE LEARNING TECHNIQUES FOR WAVE NEARSHORE DOWNSCALING

Relatori:
Marco Ferrari
Melisa Menendez
Hector Lobeto

Correlatori:
Pierluigi Brandolini

Candidato:
Edoardo Susini

Anno accademico 2022/23

Summary

1-Introduction

Coastlines are highly dynamic and changing environments that show large temporal and spatial variability in response to the action of different and complex coastal processes essentially linked to waves, currents, sea level and winds.

Buitrago et al.[1] define Coastal Hazard as "any process (natural or anthropogenic) that can cause loss or harm to life, property, services or heritage placed in its path." According to Bevacqua et al.[2] coastal vulnerability is "a spatial concept that identifies people and places that are susceptible to disturbance resulting from coastal hazard". Natural hazards in coastal areas have been rising dramatically in recent years due to increasing human presence and sudden climate change[3], such as flooding, rising sea levels, and increased frequency and intensity of storms. Knowledge of long-term wave and wind characteristics is a key aspect to guarantee coastal environment safety and protection and facilitate all marine activities[3].

The characterization of wave climate requires long-term time series of wave parameters at a particular location. These time series are usually obtained through wave propagation models capable of simulating the transformations undergone by wave motion in transfer from deep water to shallow water[4]. Wave predictions in deep waters have experienced significant developments during the last few decades and the skill of the state-of-the-art models has been shown to be generally good. However, the predictions are very sensitive to the wind fields used as has been demonstrated by previous studies[5]. Nowadays, the quality of the wind fields over the oceans is generally good, but for the enclosed basins, where the surface winds are affected by the presence of land, the skill of the wind models decreases[6]. Wave hindcasts in climate research and coastal engineering have multiplied in recent years[7], becoming one of the most powerful means of reconstructing the time series of the main parameters describing wave characteristics in coastal areas.

These hindcast databases have the advantage of providing good spatial resolution[7] and allowing the reconstruction of continuous time series with homogenous forcing[8] that cover a very long period (even more than 40 years)[7].

However, the process is fraught with certain problems that can be summarized in the three points below:

- They can show quality problems.
- The accuracy of the description of wave properties is low in shallow water as the spatial resolution is too limited.
- Interaction with bathymetry generates wave transformations that are generally not considered by models.

The main technique used to estimate the value of wave parameters near the coast is "wave downscaling". Wave downscaling can be defined as the process of refining and detailing wave data from a coarse resolution scale (e.g., global data) to a higher, localized resolution scale (e.g., regional, or coastal data). This process aims to retrieve accurate wave conditions in specific areas, such as coastal zones.

There are three main approaches for processing downscaling:

1. Statistical downscaling: it uses statistical and mathematical techniques (e.g.,) to adjust low resolution data to a better resolution[9].
2. Dynamical downscaling: it requires the implementation of a wave propagation model to simulate the transformation processes that waves undergo while approaching the coast (e.g. reflection, bottom friction, shoaling diffraction, breaking) from deep water to shallow water[7].
3. Hybrid downscaling: it involves integrating the previous two approaches.

When performing a wave downscaling, it is necessary to find a compromise between adequate spatial resolution and sustainable computational impact[10]; choosing one type of downscaling over another significantly affects the latter.

Final outcome of a wave downscaling is the time series reconstruction of the wave parameters. Long (i.e. for several decades) continuous hourly time series without gaps are needed to characterize the mean and extreme wave conditions and its climate variations. The resulting wave time series from simulations require however comparison against recorded values from in situ measurements (e.g. bouys). This procedure is named data validation.

When using hybrid or statistical wave downscaling approaches, the time series reconstruction can be achieved through different mathematical techniques. Two main types of techniques used for this reconstruction are:

- The application of mathematical equations that allow to reconstruct the complete dataset through regression methods, such as interpolation functions. One multivariate technique belonging to this group is the Radial Basis Function (RBF).
- The implementation of machine learning techniques involving the use of Artificial Neural Networks (ANN). The introduction of a limited number of inputs within the neural model enables the reconstruction of the time series through activation functions solved into the structure.

1.1-Work aims

The main objective of this work is to investigate the use of the artificial intelligence (neuronal networks) in hybrid wave downscaling techniques. To this end, the author of this study has the following secondary objectives:

- To learn and understand the physical processes associated with wave generation and propagation.
- To apply a hybrid downscaling technique to a coastal study area, which combines numerical modeling with mathematical techniques.
- To investigate the use of machine learning techniques to optimize hybrid wave downscaling to obtain continuous time series at target coastal locations.

1.2-Software used

The present work was performed entirely using the support of MATLAB software. MATLAB is a programming platform design specifically for engineers and scientists. It uses a matrix-based language allowing the most natural expression of computational mathematics. This software allows to analyze data, develop algorithms, and create models. For this work, the software was used to implement statistical analysis, to manage and generate data, and to obtain graphs

and plots. In addition, the capabilities of some characteristic toolboxes, such as the Machine Learning Toolbox, were exploited for the creation and the implementation of neural networks.

1.3-Framework of the study area

The study area is the Andalusian coastal stretch, between the Almeria harbor and the promontory of Cabo de Gata, in Spain (Figure 1).

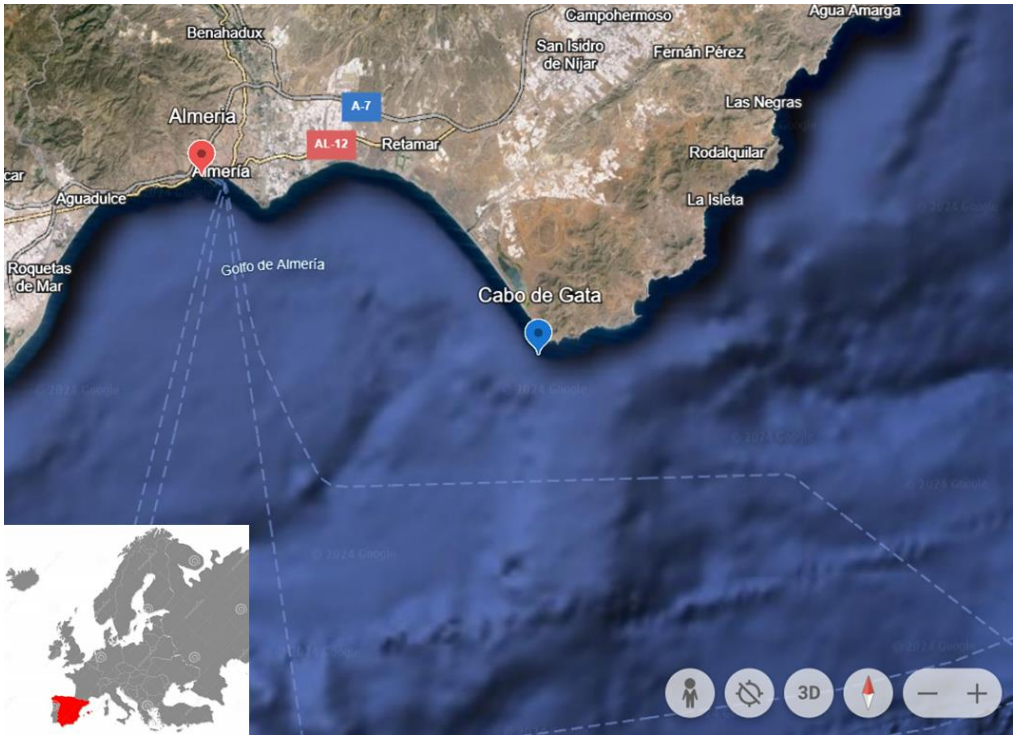

Figure 1- Study area. The location of the two buoys used for validation, Almeria (red placeholder) and Cabo de Gata (blue placeholder), is also displayed. Figure obtained from Google Earth.

1.3.1-The Mediterranean Sea

The Mediterranean Sea is characterized by a steady increase in coastal vulnerability, due to a steep rise in extreme weather events, that result in a surge of humans' loss, and human-induced settlement pressure. It is therefore very important to know how the evolution of these phenomena modifies the vulnerability and coastal risk[11] and the impact it has on marine activities[12].

The western part of the Mediterranean is typified by intense socio-economic activity and concentration of marine traffic, partly because it is a connecting zone between the Atlantic Ocean, the Red Sea and the Black Sea[12].

All the aspects highlighted so far testify to the fundamental importance of a correct and precise analysis of the marine dynamics in this basin.

Currently, several studies have assessed the spatial and temporal trends of key marine parameters in the Mediterranean Sea. The impact of waves on coastal areas or infrastructure is not only connected to the energetic properties of the wave motion, but also to the persistence of certain conditions[12].

1.3.2-The Andalusian Coast

In the Mediterranean basin, the development of economic activities close to the coast has particularly characterized the Spanish coastal area near the Costa del Sol, generating a significant increase in the number of people and property subjected to wave action and more generally to all those natural phenomena related to ongoing climate change[13].

In order to protect human and commercial activities near the coast, it is therefore essential to carry out studies that include an accurate characterization of the coastal environment[14].

The Andalusian area is tectonically controlled by the Betic Range, a mountain system located in southern Spain; the mountains in this chain also reach significant altitudes near the coast, and this aspect markedly influences the climate and coastal orography.

Rivers and streams flow through the study area, which, thanks to the input of sediment, promote the generation of small floodplains near the coast[15].

However, the regulatory plans of these basins have led to the construction of structures such as dams and reservoirs resulting in the limitation of fluvial sediment supply to the coasts and aggravation of erosion problems[16].

Regarding the grain-size characteristics of the beaches, the Andalusian coast is quite heterogeneous: the provinces of Cadiz, Malaga, and Granada are characterized by medium/coarse-grained sands and intermediate to reflexive

morpho-dynamic regimes[15]. These regimes are characteristic of beaches that have coarse-sized pebbles and reflect much of the energy of incoming waves[14]. Meanwhile, the Almeria area has beaches composed of quartz-rich fine-grained sands that exhibit dissipative morpho-dynamic regimes, characterized by the presence of bars and good stability of the beach morphology[17].

Beaches in this area are frequently interrupted by the presence of pocket beaches, small beaches located within two headlands (natural or artificial) that have an important influence on littoral dynamics.[17]

In terms of tidal characteristics, the Andalusian coast has a purely microtidal environment, with tidal ranges generally less than 20 cm.[15]

Storm events are frequent, especially in the season from November to March. The Gibraltar area is mainly affected by storms coming from the east, while the Malaga and Almeria area is affected by both easterly and western storms.[14]

The main coastal drift, related to the characteristics of the wind and the dominant sea in terms of direction and intensity, has an east-west trend.[14]

The Andalusian coast has been affected in recent decades by an important industrial and tourist development, with recreational tourism activities (about 41% of the coastal land concerned), commercial/port (2.5%) and industrial (2.3%). Areas characterized by significant natural development are also common (4.7%), while among the activities of the marine environment, the importance of fishing (0.9%) should be emphasized.[15] This intense development of nearshore activities and services results in a steady development of studies related to the prediction of marine parameters in the coastal environment.

1.4-Thesis outline

In the second chapter, an overview of the main wave parameters analyzed in this study is provided and the wave propagation characteristics offshore the coastal study area will be described.

The third chapter describes the data used in this study.

In the fourth chapter, the methodology used to develop a high-resolution wave downscaling of the main wave parameters is described, focusing on the

techniques used to reduce data dimensionality and the techniques used for the time-series reconstruction at the target points.

In the fifth chapter, the results obtained using the previously described approach will be presented. Both reconstruction techniques will be compared with the data measured by buoys to assess which methodology is more appropriate.

Finally, in the last chapter, conclusions are presented, highlighting the strengths of the methodology and possible future developments of the work.

2-Wave climate

Wave climate is the long-term statistical description of waves in a particular region or location. It encompasses the frequency, intensity, and characteristics of waves observed over an extended period, typically spanning several years or decades. Wave climate provides valuable information about the typical wave conditions experienced in a given area, including seasonal variations, extreme events, and trends over time.

According to Holthuijsen (2008)[18], a wave is the «profile of the surface elevation between two successive downward zero-crossings of the elevation (zero = mean of surface elevations)» while the surface elevation is «the instantaneous elevation of the sea surface (i.e., at any one moment in time) relative to some reference level ». A scheme of these definitions is shown in Figure 2.

The description of the main parameters related to wind-waves requires an assumption: for relatively limited time intervals (e.g., 1 hour), or sea states, the wave motion can be studied considering it statistically stationary[18]. By adopting this assumption, the wave field can be described by integrated wave parameters, which represent an average value of the main wave characteristics (e.g., significant wave height, mean period, mean direction for the wave height, period, and direction, respectively).

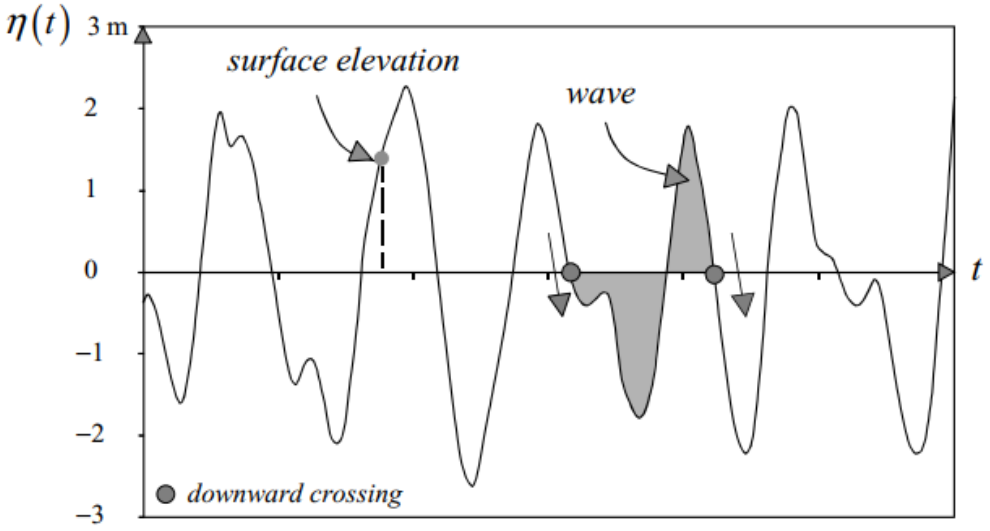

Figure 2-Definition of wave and elevation surface in a time record (source: Holthuijsen, 2008)

To accurately describe ocean waves, it is necessary to introduce the concept of wave spectrum.

The wave spectrum is a mathematical representation of the energy distribution of ocean waves as a function of their frequency (or period) and direction. The most accurate representation of the energy associated with the wave spectrum is that provided by the variance density equation, which describes the wave spectrum by considering the variation of its amplitude spectrum:

$$E(f) = \lim_{\Delta f \to 0} \frac{1}{\Delta f} E\{\frac{1}{2}\underline{\alpha}^2\} \quad (1)$$

The wave spectrum provides a theoretical basis for understanding and predicting the sea state. A sea state refers to the prevailing conditions of the ocean surface at a specific location and time, encompassing various parameters that describe the behavior of waves. It is determined by factors such as wind speed, duration, and direction, as well as the interaction between waves, currents, and the ocean floor. Sea state parameters include significant wave height and wave period,

2.1-Wave spectral parameters: wave height and wave period

Spectral parameters are derived from the spectral moments of the wave spectrum.

The parameter that most immediately describes wave motion characteristics is the wave height (H), «the vertical distance between the highest and the lowest surface elevation in a wave»[18]. In a wave record with N waves, the mean wave height $\overline{H}$ is then readily defined as:

$$\overline{H} = \frac{1}{N}\sum_{i=1}^{N} H_i \quad (2)$$

where *i* is a single wave within the wave record.

The significant wave height (H_S) is frequently used instead of the individual wave height as it gives integrated information on the waves reaching the coast over a sea state. The H_S, from a spectral point of view, is a statistical measure of wave height within the spectrum. H_S is calculated from the first spectral moment, Hm0, defined as the square root of the second moment of the wave spectrum.:

$$H_s = 4\sqrt{\mathrm{Hm0}} \tag{3}$$

where H_S is the significant wave height and Hm0 is the zeroth spectral moment. Hm0 is calculated from the energy distribution within the wave spectrum by the following formula:

$$Hm0 = \sum_{i=1}^{N} E_i \tag{4}$$

where E_i is the energy associated with each frequency band in the wave spectrum, N is the total number of frequency bands in the spectrum.

The wave period is defined as «the interval between one zero-down crossing and the next».[18]

The mean period of a sea state can be obtained through the following equation:

$$T_0 = \frac{1}{N}\sum_{j=1}^{N} T_{0,j} \tag{5}$$

where *i* is a single wave within the wave record.

From a spectral point of view the wave period can be described as follows.

The peak period (Tp) is typically obtained from the spectral peak frequency, which corresponds to the frequency with the maximum energy in the spectrum. The mean wave period (Tm02) can be derived from the second spectral moment, representing a measure of the central tendency of wave periods:

$$Tm02 = \frac{\sum_{i=0}^{N} T_i E_i^2}{\sum_{i=1}^{N} E_i^2} \tag{6}$$

Where T_i represents the period of each wave component in the spectrum, E_i the energy associated with each wave component and N is the total number of wave components.

2.2-Wave propagation

In the context of coastal engineering studies, it is essential to understand the mechanisms that govern wave propagation in deep and surface waters. Particularly, the changes the wave undergoes in its motion toward the coast as it interacts with the bathymetry are of great importance.

As the movement toward the coast progresses, the wave profile tends to become steeper with an increase in wave amplitude and a decrease in length. Then, wave breaking occurs, with dissipation of the energy transported by the wave[19].

The basis in the field of wave propagation is the linear theory[18], which can be applied to describe wave propagation in both oceanic and coastal environments. Linearity implies that there is no interaction between the waves during their motion, and the main assumption for applying the theory is that the waves exhibit a small amplitude relative to the wavelength and depth of the water[18]. Water is considered as an ideal fluid, whose movement is controlled only by the earth's gravitational force; this assumption allows water to be identified as an incompressible fluid with constant density and no viscosity.[18]

There are two equations that govern the linear theory: a mass balance equation and a momentum balance equation.

The continuity equation is derived from the mass balance equation:

$$\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} + \frac{\partial u_z}{\partial z} = 0 \qquad (7)$$

This is a linear equation expressed in terms of the velocities of water particles u_x, u_y and u_z.

The moment balance equation for the three main directions is represented as follows:

$$\frac{\partial u_x}{\partial t} = -\frac{1}{\rho}\frac{\partial p}{\partial x} \tag{8}$$

$$\frac{\partial u_y}{\partial t} = -\frac{1}{\rho}\frac{\partial p}{\partial y} \tag{9}$$

$$\frac{\partial u_z}{\partial t} = -\frac{1}{\rho}\frac{\partial p}{\partial z} - g \tag{10}$$

In the z-direction the term g appears as the volume weight of water acts in that direction.

Wave propagation is closely related to the wind action in open sea. This action is a very complex phenomenon that can be represented by a three-dimensional vector that varies randomly in time and space.

To understand the influence of wind in wave formation and propagation, it is necessary to introduce the concept of fetch. Fetch is defined as the distance across which the wind can blow in a constant and uniform direction across the sea surface without encountering obstacles. The greater the fetch, the higher and more energetic the waves generated will be.[18]

When waves approach the shoreline their main characteristics, such as amplitude and direction, undergo changes related to the fact that the depth is limited. Various phenomena develop in shallow waters:

- Shoaling: a phenomenon that causes a change in the direction of wave propagation due to a change in group velocity, i.e., the propagation speed of a wave train[18].
- Refraction: modification that generates a change in the direction of wave propagation, mainly due to depth-induced changes in the phase velocity along the wave crest[18].
- Diffraction: a phenomenon similar to refraction, which is, however, generated by the presence of obstacles such as islands and headlands[18].

- Reflection: modification that causes a partial return of wave energy toward the point where it originated. The triggering of this phenomenon is because the wave may encounter obstacles such as rocky headlands, defense works, and changes in the topographic features of the seabed during its path[18].

The most complex phenomenon to describe, however, is wave breaking. There are several types of breaking, which can be estimated through the Iribarren number (or surf similarity parameter, ε)[18]:

$$\varepsilon = tan\alpha/\sqrt{H/L_{\infty}} \tag{11}$$

By calculating this parameter in deep water:

$$\varepsilon_0 = tan\alpha/\sqrt{H_{\infty}/L_{\infty}} \tag{12}$$

with H_{∞} representing the wave height in deep water and $L_{\infty} = gT/^2\pi$, the following types of breaks are identified:

- Spilling: if $\varepsilon_0 < 0.5$
- Plunging: if $0.5 < \varepsilon_0 > 3.3$
- Collapsing or surging: $\varepsilon_0 > 3.3$

The Iribarren number influences not only the breaking mechanism but also the way wave reflection occurs and the run-up of the wave on the beach (i.e., the maximum vertical distance the wave reaches on the beach).

The breaking process causes a limitation of wave height in surface waters; this aspect has heavy implications in the assessment of H_S with consequent impacts in the design of coastal defense works.

3-Data

In this chapter, the databases employed for the study are described. Section 3.1 provides information about the wind data, while section 3.2 describes wave data from observations (3.2.1) and hindcast (3.2.2). Finally, section 3.3 details the bathymetric information used.

3.1 Wind Data

The main forcing force driving the formation of wave motion is wind. Wind-generated waves are formed when the wind blows over the surface of the water and transfers kinetic energy to the water.

In recent years, the number of measured observations of wind and atmospheric pressure at sea level has increased significantly over the North Atlantic European coast and shelf. However, measurements from in-situ instruments, although the most reliable, lack the spatial resolution necessary to undertake global and regional scale studies. Since the 1980s, satellite data provide good spatial coverage, but with discontinuous temporal measurements. Given these problems, the use of data from calibrated global atmospheric numerical models has become very popular, as it provides a spatially and temporally consistent set of atmospheric variables over a long period.

One of the most recent global reanalysis products is ERA-5, developed by the European Centre for Medium-term Climate Prediction Facility (ECMWF), with hourly temporal resolution and spatial resolution of 0.25^{o}[20]. The wind fields from ERA5 are used in this study as input forcings or the numerical wave modelling. The ERA5 reanalysis product has been validated against in-situ records. It has been used in multiple studies, showing, for example, its skill to reproduce weather types[21] and its skill to model wind power more accurately than previous models, such as MERRA-2[22].

Like any reanalysis database, ERA5 combines observational data with coupled models of the climate subsystems through the process of data assimilation. It uses the IFS Cycle 41r2 4D-Var data assimilation system and currently covers the period from 1979 to the present.

3.2-Wave Data

To develop this study, it has been used two sources of wave data: those measured in situ from bouy gauges, and those obtained through wave models from numerical simulations (hindcast).

3.2.1) In situ measurements

Instrumental wave measurements provide valuable information on the behavior of the waves. The in-situ records provide local information that require specific quality control, often providing data for periods of time that are too short to carry out a rigorous climate study (for which at least three decades of information are recommended to characterize natural climate variability).

In this work, the observational data have been used to validate the wave dynamics databases used as boundary conditions in the downscaling process, and to validate the databases generated. Regarding the latter, the comparison with in-situ records allows for validation of the generated databases, both for the mean values and different magnitudes of each variable, including a validation of extreme events.

The buoys of Almeria and Cabo de Gata have been used in this work to validate the data. The former is located within the harbor of Almeria, while the latter is close to a promontory. More details about the characteristics of the buoys are given in the table below (Table 1). Additionally, the map in Figure 3 shows their location along the Andalusian coast.

Buoy	Coordinates (°) [Lon-Lat]	Code	BuoyNet	Observed period	Depth (m)
Almeria	2.48W-36.83N	1537	REDCOS	27/7/2000-6/9/2006	15
Cabo de Gata	2.20W-36.71N	1518	REDCOS	10/4/91-6/5/2012	35

Table 1-Buoy characteristics

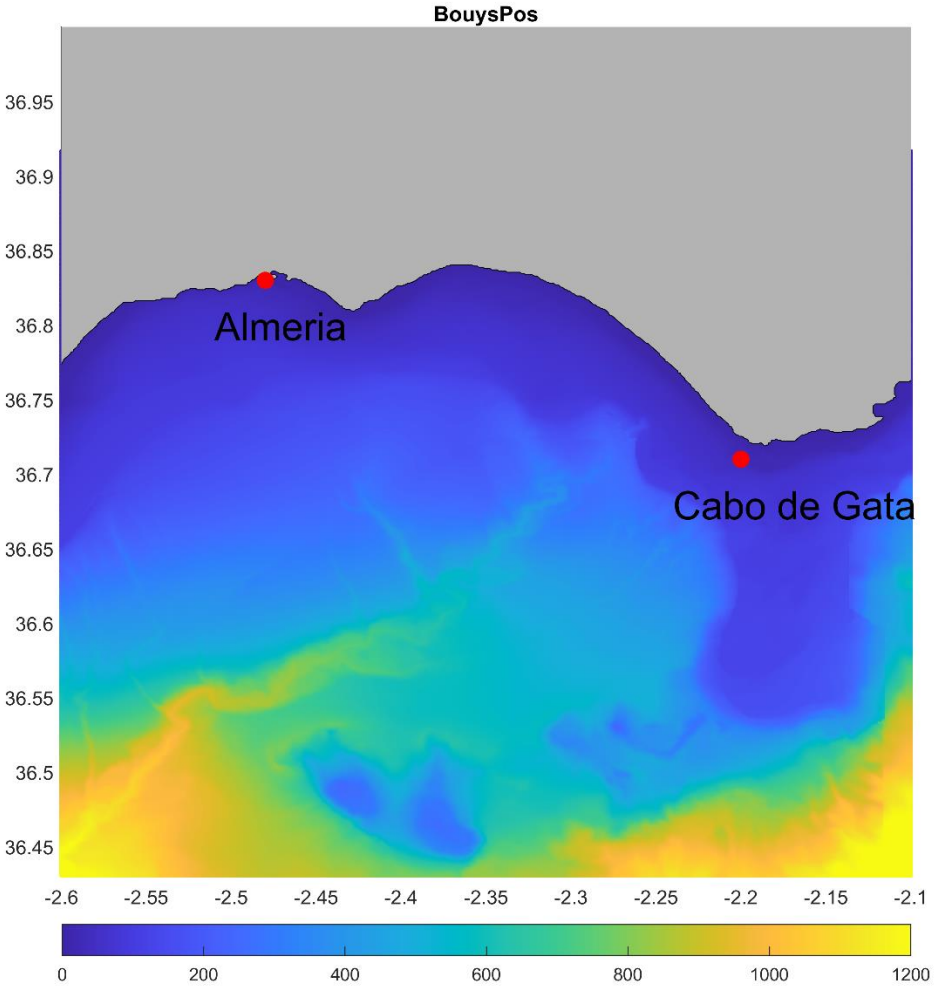

Figure 3-Buoy position along the Andalusian coastline. The bathymetry of the area is also shown.

3.2.2) Hindcast wave data

In recent years, wind wave hindcasts have become the most widely used tool for obtaining marine weather/climate information to be exploited in coastal areas[23]. Hindcasting is a technique that uses models to retrospectively simulate past marine conditions.

To develop the nearshore downscaling, a wave dataset to be used as boundary conditions is required. Since the size of nearshore downscaling grid is frequently small for computational purposes, the waves reaching the coast are likely not

generated within the grid. Thus, waves are introduced in the contour of the grid from a regional wave hindcast database (ROW database, Figure 4).

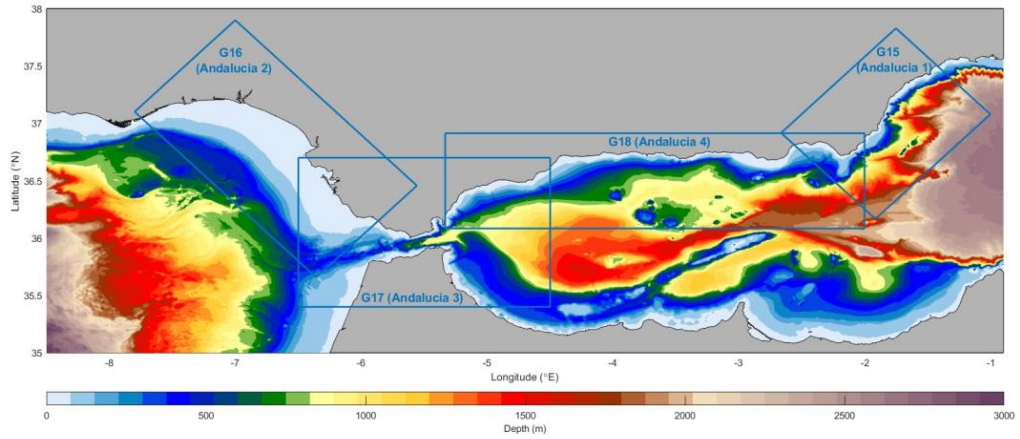

Figure 4-Domains of the wave hindcast ROW, with 1 km horizontal resolution.

ROW is a set of historical wave reconstructions with 1 km of resolution developed at IHCantabria, which is produced through a dynamic wave downscaling of the global Global Ocean Waves (GOW) product also developed at the same institution. The particular ROW product covering the Andalusian coast spans for the historical period 01/1985-12/2019 with hourly resolution. This database was generated using the SWAN (Simulating WAves Nearshore) numerical model on a grid with a spatial resolution of 1 km covering the entire Andalusian coastal area.

3.3-Bathymetry

To generate the domains, bathymetric and shoreline information from different sources has been integrated. It is summarized below:

- DTM 1st coverage (2008-2015) with 5 m grid pitch from the IGN.
- Bathymetric data from the eco-cartographic studies of the Spanish Ministry for Ecological Transition and Demographic Challenge (MITERD) for the coastline province of Almería and Granada (2008-2009).
- EMODnet (2020) in those areas where no ecocartes were available, such as the province of Huelva and offshore areas. The link is made through a

coastline generated from the IGN coastline in natural areas and the IGN municipal boundary in anthropic areas, always anthropic environments, always supported by the National Plan of Aerial Orthophotography (PNOA) of the most up to date.

The work was carried out considering the bathymetry information referenced to the Alicante Mean Sea Level (AMSL), which is the zero of the eco-cards and coastline.

4-Methods

This chapter presents the downscaling methodology used in this work, with particular focus on the reconstruction approach used to obtain the time series at the target points. Section 4.1 describes the numerical wave propagation setup (subsection 4.1.1) and the case selection process (subsection 4.1.2.). Section 4.2 describes the approaches to reconstruct of time series using RBF (4.2.1) and ANN (4.2.2). Section 4.3 presents a description of the data validation process, analyzing some representative error metrics.

4.1-Hybrid downscaling Approach

The generation of a high-resolution coastal wave database (hereinafter DOW, Downscaling Ocean Waves) using a hybrid downscaling methodology allows to significantly reduce the computational effort with respect to a dynamic downscaling (for example, a wave hindcast using SWAN propagations for a long historical period). The approach defined as DOW has been developed by IHCantabria[7] and applied in multiple projects for the last decade.

The DOW database analyzed here covers the period 1985-2019 on the coast reaching 100 meters of spatial resolution.

Figure 5 shows the DOW development methodology. The process will be fully explained in the next sections.

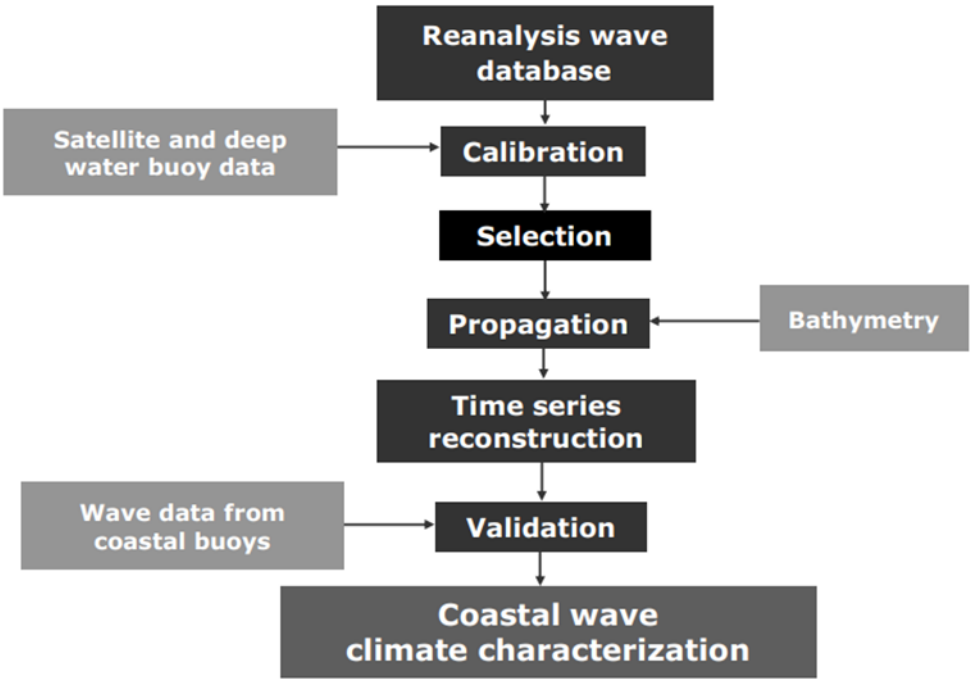

Figure 5-DOW development methodology.

4.1.1-Selection of Cases: data standardization, dimensionality reduction (PCA) and clustering (MDA)

The hybrid approach is based on simulating a limited number of sea states independently, so that the full time series can be reconstructed through interpolation afterwards. Simulation cases have been chosen using the maximum dissimilarity algorithm (MDA) clustering algorithm.

Wave and wind data are considered as input variables to perform the clustering. The points in which wind and wave data are obtained from ERA5 wind fields and ROW database, respectively, are shown in Figure 7. In particular, the wave variables considered to conduct the classification are H_S, T_p, T_{m02}, Dir.; the wind variables considered are E(v) and N(v), representing the wind speed along the east-west and north-south directions, respectively.

As mentioned in Chapter 4.1 the selection based on these variables leads to different problems in data processing, which can be summarized as follows:

- The variables involved may have very different ranges of variation, which may consequently result in different weights of the variables during clustering.
- The number of variables involved in the process is too large to be managed in the clustering process.
- Hindcast data simulations returns very good spatial resolution but at the same time requires unsustainable computational effort.

The first problem is solved applying a data standardization technique that allows to homogenize all variables involved, rescaling to a null mean and a standard deviation of 1. This is achieved following the formula below:

$$Z = \frac{x - X}{\sigma} \tag{13}$$

where Z is the rescaled value, x is the value of the variable in a specified point, X is the mean of the values of the variable, and σ is the standard deviation.

The dimensionality reduction of variables is performed using the PCA. This technique allows to eliminate redundant information with a minimum loss of data, compress and transform it to a new space. Each PC represent a certain percentage of variability and the first PC represent the most variance.

$$\sigma_t = \sigma_{PC1} + \sigma_{PC2} + \sigma_{PC3} + \cdots + \sigma_{PCN}$$

$$\sigma_{PC1} > \sigma_{PC2} > \sigma_{PC3} > \cdots > \sigma_{PCN} \tag{14}$$

The goal of PCA is to find the minimum number of components that will guarantee a given variance of the total data.
The original data can be expressed as a linear combination of PCs and EOFs (Empirical Orthogonal Function)[7], a useful function to identify dominant modes of temporal and spatial variation in the data:

$$X(x,t_i) = EOF_1(x) \times PC_1(x) + EOF_2(x) \times PC_2(x) + \cdots + EOF_d(x) \times PC_d(x) \quad (15)$$

Obviously, the variance is more accurately described as the number of PCs considered increases, but the aim of this application is to select a limited number of components that ensure a certain percentage of variability. To estimate the appropriate number of PCs, the root mean square error (RMSE) of the offshore wave and wind condition is performed, progressively increasing the number of PCs and explained variance[7].

At the end of this process, it became clear that 99% of the variability of the data in the analysis performed is accurately described by 16 components. The components thus identified are representative of the 16 points used as forcings for the SWAN model during the DOW (Figure 6).

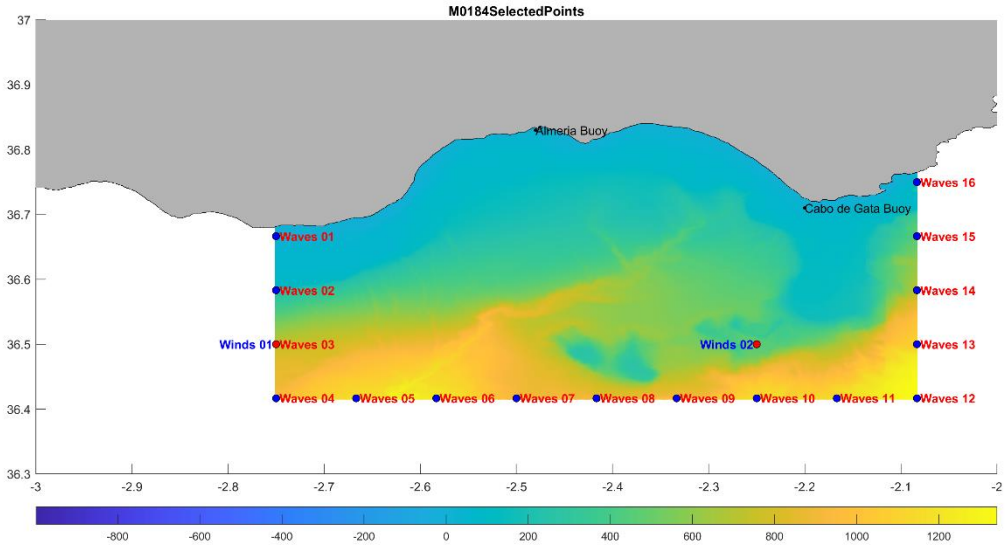

Figure 6-SWAN model input point distribution; the 16 points representing the wave forcing (obtained via PCA) and the 2 representing the wind forcing (obtained from the ERA-5 model) for DOW.

The last problem is solved by applying a data clustering algorithm (Maximum Dissimilarity Algorithm, MDA) at the points identified by PCA (Figure 7) to

identify a subset of temporal data that is representative of the maximum dissimilarity within a given database [19].

MDA is a hierarchical-agglomerative clustering method. Hierarchical-agglomerative methods are clustering algorithms that initially consider each point in the data set as an individual cluster. They then progressively join the closest clusters together to create a hierarchical cluster structure. This technique allows us to identify from a sample of M data, the subset of N data that explain the highest variability of the total set.

The goal of the MDA is to minimize the redundancy between the selected data. In this work it was chosen to select 500 representative cases. The case from among the 306792 with the highest H_S is chosen as the starting data for clustering, to keep within the data processed by SWAN the value that is thought to produce the greatest damage in the studied stretch of coastline.

The algorithm next identifies within the cases the one that is most dissimilar to the previously identified; in this case dissimilarity is assessed not only based on the characteristics of H_S but also in relation to the other wave variables (e.g., wave period, wave direction, ecc.) and wind.

Mathematically, the most dissimilar case compared to the previous one is calculated through an equation that allows us to obtain the Euclidean distance between the selected case and all the others:

$$D_i = \sum_{j=1}^{N-1} \left\| x_i - x_j \right\|; j = 1, \dots., N \tag{16}$$

Where N is the number of cases chosen as representative (500 in this work).

The most dissimilar case will be the one that presents the greatest distance compared to the first case.

The newly selected case becomes the current, and the process continues iteratively until the required 500 cases are identified. Then the set will be representative of a wide range of possible situations.

The 500 cases identified through the application of the MDA to the wave and wind fields represented in the PCA are then numerically simulated with SWAN

(Section 4.1.1) to obtain information on the main marine variables of interest at a very high resolution (about 100 m) within the reference mesh. The target domain for this work is M0184 (Figure 7).

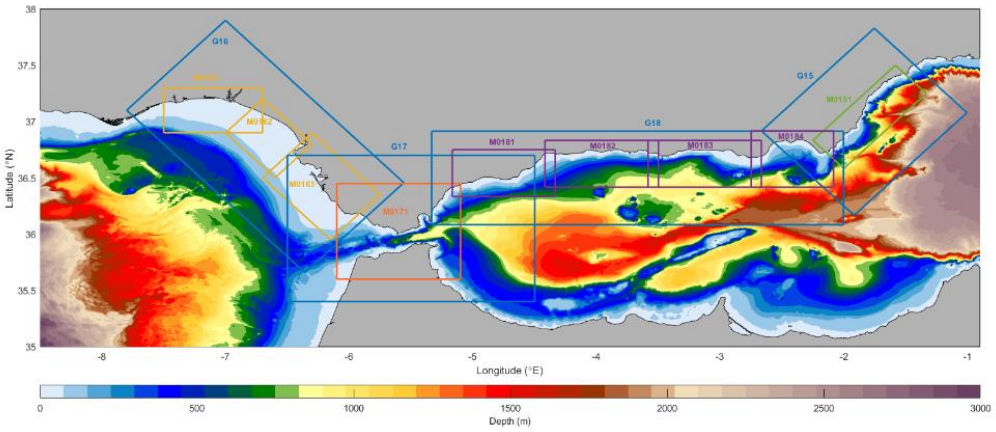

Figure 7-Position of external domain (ROW) and internal domain (DOW, 100 m of resolution)

4.1.2-Numerical Wave Propagation Setup (SWAN)

SWAN is a third-generation spectral model developed by NOAA/NCEP. Third-generation wave models can simulate a wide range of wave conditions and are used in various applications, including weather forecasting, coastal engineering, maritime operations, and oceanography. They are an essential tool for understanding and predicting ocean wave behavior, ensuring safety at sea, and making informed decisions related to coastal and offshore activities.

SWAN is based on the solution of the phase-averaged wave action equation. This equation allows to correctly simulate the processes of refraction, scattering, dissipation with the bottom, dissipation by white capping, break-up, non-linear interactions, wind wave generation at regional scales and diffraction. The use of implicit schemes for solving the equations and the inclusion of specific terms for shallow water make it the most suitable tool for high-resolution coastal wave propagation.

The model is based on the following equation[24]:

$$\frac{\partial}{\partial t}N \times + \frac{\partial}{\partial x}c_x N + \frac{\partial}{\partial y}c_y N + \frac{\partial}{\partial \sigma}c_\sigma N + \frac{\partial}{\partial \theta}c_\theta N = \frac{S}{\sigma} \qquad (17)$$

Where N (σ, θ; x, y, t) is the wave action density that is a function of frequency σ, direction θ, horizontal coordinate x, y, and time t[24]. The first three terms on the left describe the propagation and evolution of N over time and in the two horizontal directions[9]. The fourth term is representative of the shift in frequency induced by changes in current and depth. The fifth quantifies the refraction phenomenon induced[24]. S, on the other hand, is representative of the effects of generation, dissipation, and non-linear wave-wave interaction[24]:

$$S = S_{in} + S_{nl3} + S_{nl4} + S_{ds,w} + S_{ds,b} + S_{ds,br} \quad (18)$$

Si_n represents the wind input, which is considered the main factor in the equation[9]. S_{nl3} and S_{nl4} are parameters that consider the non-linear interactions that characterize wave motion in deep and surface water, respectively. The last three parameters are instead representative of the phenomena of white capping, bottom roughness and depth-induced wave breaking.

The SWAN version 41.20 was implemented in its stationary mode and sensitivity analyses were performed to find the best configuration of the numerical model. The data used by the model in the runs have been full-spectrum data, discretized in 48 directions and at 40 frequencies ranging from 1.5 Hz to 0.0324Hz, thus ensuring that all variability was captured. Parameters associated with breakdown dissipation (BREAK=0.73), dissipation by whitecapping (WCAP=2.36E-05) and bottom friction (FRIC=0.038) are configured for the specific analysis area. The rest of the parameters were configured following the recommended default values.

4.2-Reconstruction of time series

The following subsections present in detail the methods of time series reconstruction by interpolation functions such as the RBF (4.2.1) and ANN (4.2.2).

4.2.1-Radial Basis Function

An RBF is a function whose value depends on the radial distance from a central point. This function is often used in various contexts, including interpolation, function approximation, and machine learning.

To reconstruct the time series of waves during the historical period (1985 - 2019) in the proximity of the Andalusian coast, non-linear interpolation techniques based on RBFs are used, applying the methodology described in Camus et al., 2011. This scheme is very advantageous when dealing with sparse and multivariate data[19].

The methodology is based on the calculation of an RBF approximation function, formed by a linear combination of symmetric radial functions, centered at the points given by the propagated cases (where the function at these points results in the exact value of the propagated parameter). In total, as many RBF approximation functions will be calculated as many wave parameters have been used.

More in detail, through the application of an approximation function obtained by a weighted sum of radially symmetric basic functions at the 500 points derived from the MDA algorithm, the implementation of the RBF function is carried out:

$$RBF(x) = p(x) + \sum_{j=1}^{M} a_j \phi(\left|\left|x - x_j\right|\right|) \qquad (19)$$

Where ϕ is the RBF, p(x) is a monomial basis formed by a number of monomials of degree one equal to the size of the data and a monomial of degree zero, being $b = \{b_0,\ b_1, \dots ,\ b_n\}$ the coefficient of these monomials. M is the number of cases selected with MDA and x_j is the associated real-valued function for every case $f_i = f(x_j),\ j = 1, \dots, M$. Finally, a_j and b are the RBF and monomial coefficient respectively.[19]

The goal of the RBF application is therefore the reconstruction of the time series for all parameters of interest at any point within the mesh using an interpolation function.

The calculation process of the function begins with the normalization of all variables considered. At this point, each deep-water sea state is defined through the expression $D_j = \{H_i, T_i, \theta_i, W_i, \beta_i\}; 1, \dots, N$, and each selected case is expressed using the function $D_j = \{H_j^D, T_j^D, \theta_j^D, W_j^D, \beta_j^D\}; j = 1, \dots, M$ The interpolation function will be calculated as follows:

$$RBF(X_i) = p(X_i) + \sum_{j=}^{M} a_j \phi(\left|\left|X_i - D_j\right|\right|) \tag{20}$$

With $pX_i = b_0 + b_1 H_1 + b_2 T_1 + b_3 \theta_1 + b_4 W_i + b_5 \beta_i$ and ϕ is a Gaussian function defining the shape parameter c:

$$\phi\left(\left|\left|X_i - D_j\right|\right|\right) = \exp\left(-\frac{\left|\left|X_i - D_j\right|\right|^2}{2c^2}\right) \tag{21}$$

Finally, the time series are transferred from deep to surface water at the point of interest by using the following functions:

$$H_{sp,i} = RBF_H\left(\{D_j H_{sp,j}(j = 1, \dots, M)\}, X_i\right); i = 1, \dots, N \tag{22}$$

$$T_{mp,i} = RBF_T\left(\{D_j T_{mp,i}(j = 1, \dots, M)\}, X_i\right); i = 1, \dots, N \tag{23}$$

$$\theta_{xp,i} = RBF_{\theta_x}\left(\{D_j \theta_{xp,i}(j = 1, \dots, M)\}, X_i\right); i = 1, \dots, N \tag{24}$$

$$\theta_{yp,i} = RBF_{\theta_y}\left(\{D_j \theta_{yp,i}(j = 1, \dots, M)\}, X_i\right); i = 1, \dots, N \tag{25}$$

The result is the reconstructed time series at a specific location in shallow water:

$$X_{p,i}^* = \{H_{sp,i} T_{mp,i}, \theta_{mp,i}\}; i = 1, \dots, N \tag{26}$$

The time series reconstruction using RBF is carried out at the points represented by the Almeria and Cabo de Gata buoys (Figure 3). Unfortunately, the two buoys under consideration do not have records of wave direction, so that only the significant wave height (H_S), peak period (Tp), and average period (Tm02) will be considered.

After reconstructing the time series for the parameters of interest, it was decided to derive scatter plots representative of the comparison between the data obtained using the RBF and the data measured from the buoys.

4.2.2-Artificial Neural Networks

In recent years, the development of sophisticated machine learning systems has made it possible to test an alternative method for reconstructing time series.

In particular, ANNs are widely used, as they are capable of studying the non-linear behavior that characterizes wave motion when the transformation phenomena involving it occur upon reaching shallow water[25].

Browne et. all, have defined the ANNs as "a flexible learning architecture which rely on the presentation of input and target data, rather than a theoretical model, for the estimation of an underlying physical relationship".

The development of these systems is modeled based on the characteristics of the human brain[26]. Precisely because of their similarity to biological neural networks, they do not need to be pre-programmed to perform a given task as their functioning is based on adaptive learning[27].

ANN are networks formed by groups of individual elements, called neurons, which operate in parallel and are organized in level, called layers. Any neuron receives an input, processes it and produces an output[28].

Neural networks typically have three layers:

- Input layer: the input layer receives as input raw data representing the force variables of the neural model. These data undergo a transformation, which will be analyzed in more detail later, and are then transported into the hidden layer through the connection between these two layers[29].

- Hidden layer: the main calculations are implemented and the different dependencies between the variables are extracted in this layer[29]. Essentially, hidden layers play a crucial role in enabling neural networks to model complex relationships in data, making it possible to process information and generate output.
- Output layer: this is the layer that yields the value of the interest variable[29]. In summary, the output layer is crucial for generating the results that are usable by the network based on its learning capabilities and the requirements of the specific problem. Thoughtful design of this layer is essential for achieving optimal performance in addressing the given problem.

The neural network used in this work belongs to the category of feedforward networks. These neural networks have a simple structure and are defined as 'unidirectional' because the data flow cross through the network without loops or feedback (Figure 8).

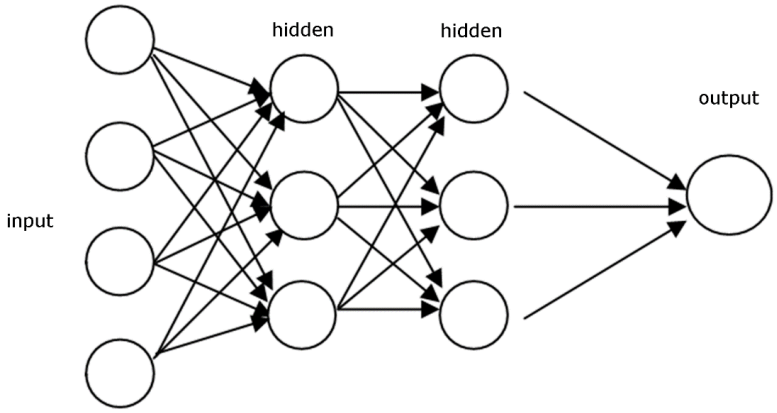

Figure 8-Feedforward neural network structure.

However, there are several critical aspects in managing and training a neural network with these characteristics:

- **Data set partitioning:** in the context of neural networks, it is common practice to divide the input data into three different subsets, represented by the training set, the test set and the validation set.

The training set is the portion of the dataset used to train the neural network; the network tries to capture relationships in the training data to minimize errors so that it generalizes well to new similar data.

The validation set is used during training to evaluate the network performance on data that were not used for training. This set is important to monitor the network's behavior on previously unanalyzed data and to prevent overfitting. Overfitting is a phenomenon in which a machine learning model fits the training data too well but does not generalize correctly to new inputs. In other words, the model learns the details and noise present in the training data to the point that its ability to make accurate predictions on unobserved data diminishes.

The test set is used at the end of the training process, after the network has been fully trained. This set represents completely new data, which the network has never seen before. The goal is to evaluate the network's performance on new data to make sure that the model generalizes well and is not too specific to the training data. The results of the test set provide an estimate of the model's effectiveness.

In this work, the outputs of the 500 cases simulated using SWAN model were divided as follows: 70% of the data was used for training while the remaining 30% was split equally between validation and test sets.

The division of the 500-input data into the three sets just described is done randomly, so that each time the network is processed, the data processing and the results obtained are different. This has implications for the stability of the model, but it ensures that the network is better able to adapt to different inputs.

- **Weights initialization**: Once the partitioning of the dataset is done, the initialization of connection weights between the various neurons takes place. These are parameters that are learned by the network during training and influence the performance of the model. In the present work, the method of initializing the weights used is the Ngueyen-Widrow method. This is a process that involves initialization to ensure that the active region of each neuron is causally but uniformly distributed within

the input space of the layer; this aspect is crucial to ensure that each neuron is sensitive to different inputs in a uniform manner, preventing some neurons from over-specializing on specific patterns during the initial training phase.

- **Activation/transfer function**: Activation functions are used to introduce nonlinearities into neural networks. This aspect allows the network to learn complex relationships in the data, enabling it to fit more intricate patterns that are thus adaptable to more data. In a nutshell, the activation function defines the output of a neuron relative to the weighted sum of its inputs.

  In this thesis, the activation function used is the hyperbolic tangent function. In addition to introducing as mentioned nonlinearity, the function returns output data that are in a range between -1 and 1; this limited range allows for greater stability of the neural model.

- **Optimization algorithm:** Optimization algorithms are methods used to update the weights of the neural network during the training period.

  The main objective of an optimization algorithm is to limit the cost function. This function is representative of the discrepancy between the values obtained by the neural model and the values to which one would like the model to approach, in this case the measured parameters. Limiting the cost function through the optimization algorithm should therefore return results progressively closer to those measured as the model's training progresses, thus minimizing error.

  The optimization algorithm used in this work is the Levenberg-Marquardt algorithm. It acts progressively on the weights connecting the various neurons allow progressive minimization of the error.

A relevant aspect to consider when training a neural network is the model stability. To pursue this, in addition to working on the parameters described above, the structural characteristics of the network in terms of the number of neurons, layers and iterations must also be considered. In fact, these characteristics have an important influence on the results obtained from the model; a very simple model, characterized by a few neurons, layers and

iterations, may lead to inaccurate prediction of the parameters of interest while a more complex model could lead to overfitting issues, also increasing the computational time required to process the network.

The stability of the model is then related to how the input data are treated. Previously it was seen that the input data were divided into the training, test, and validation sets randomly. This provides the network with good adaptability to different input data from each other by preventing it from being too specific and unable to adapt to the input of new data. At the same time, however, instability tends to be generated because the partitioning of the input data into the network is different each time the network is processed.

All scatter plots and time series presented in Chapter 5.2 results from the iteration giving the minimum test error. The test error is a measure of the model adaptability to new training data so minimizing it is indicative of a model that provides better stability and increased performance.

4.3-Validation against in-situ data

Several error metrics have been calculated for evaluating the performance of the reconstruction against buoy data. The error metrics considered are:

- BIAS, that represents the discrepancy between the average value predicted by the model and the actual value of the parameter:

$$BIAS = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n} \quad (27)$$

- RMSE (Root Mean Square Error), a measure of the dispersion of model predictions from actual values. It is calculated by taking the square root of the mean of the squares of the differences between predicted and actual values. In essence, it provides an estimate of the standard deviation of the model errors. A lower RMSE indicates greater accuracy of the model in its predictions.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \quad (28)$$

- SI (Scatter Index), which is indicative of the dispersion of the data concerning a given regression line. A lower value indicates less dispersion of the data around the regression line while a higher value indicates greater dispersion resulting in a less accurate fit of the model used to the data.

$$SI = \frac{RMSE}{\bar{x}_i} \quad (29)$$

- CORR (Correlation Coefficient), which measures the strength and direction of the linear relationship between two variables. It takes values between -1 and 1, where 1 indicates perfect positive correlation, -1 indicates perfect negative correlation and 0 indicates no linear correlation.

$$CORR = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (30)$$

- R2 (Determination Coefficient), a statistical measure that represents the proportion of variance in the data dependent on the independent variable in the regression model. In other words, R2 indicates how well variations in the dependent variable can be explained by variations in the independent variable. An R2 closer to 1 indicates a good fit of the model to the data, while an R2 closer to 0 indicates a worse fit.

$$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (31)$$

5) Results

This chapter reports the results obtained from the previously described analyses. Section 5.1 presents the results of RBF reconstruction, while section 5.2 reports those obtained through ANN reconstruction. Section 5.3 compares the results obtained from the two different reconstruction methods.

5.1) RBF reconstruction

Below are reported the scatter plots obtained by performing a reconstruction of the H_S (Figure 9), Tm02 (Figure 10), and T_p (Figure 11) parameters through RBFs for the two buoys under analysis. The scatter plots show the pairs of hourly data values (colored according to their probability of occurrence) and the percentiles of the distribution (black and red diamonds). In the graphs are also shown information on key error metrics (see Chapter 4.3)

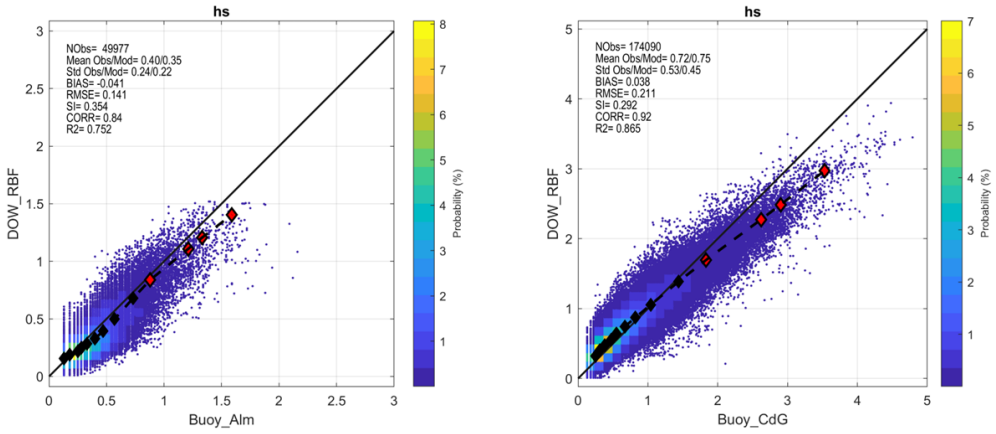

Figure 9-Comparison of H_S estimated parameter using the DOW-RBF approach for the buoys of Almeria (left) and Gabo de Gata (right)

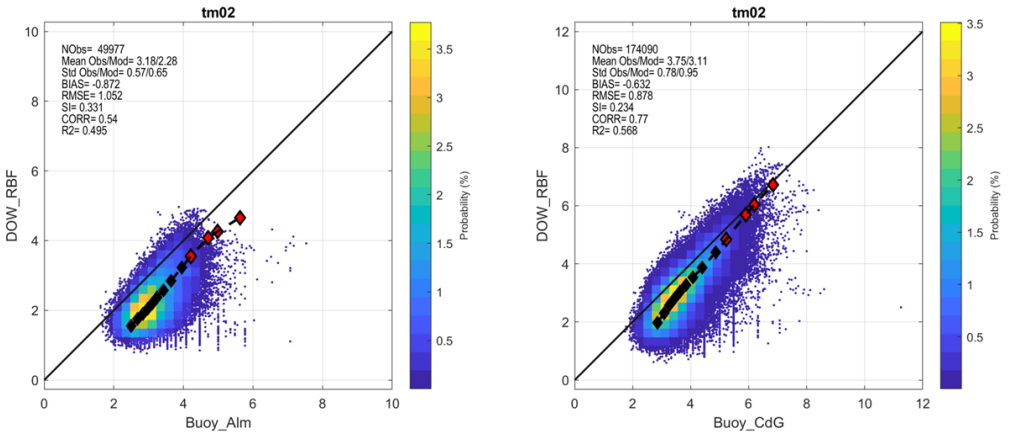

Figure 10- Comparison of Tm_{02} estimated parameter using the DOW-RBF approach for the buoys of Almeria (left) and Gabo de Gata (right).

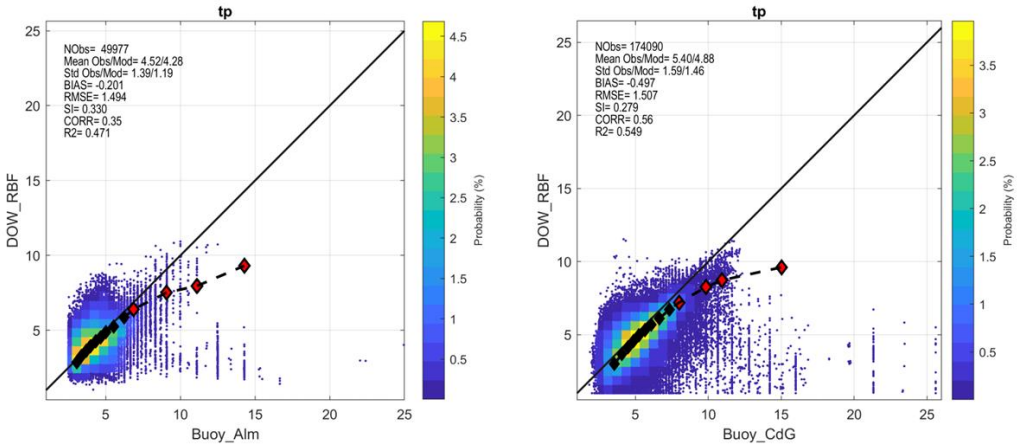

Figure 11- Comparison of T_p estimated parameter using the DOW-RBF approach for the buoys of Almeria (left) and Gabo de Gata (right).

It can be seen from the plots that the H_S for the Cabo de Gata buoy is the most accurately reconstructed parameter, while the model struggles to return an adequate representation of Tm02 for the Cabo de Gata buoy and both parameters for the Almeria buoy.

According to these results, the subsequent investigations will focus on the H_S parameter.

Scatter plot in Figure 12 shows the comparison between the H_S data obtained from the reconstruction through the RBF and the data measured from the Cabo de Gata buoy.

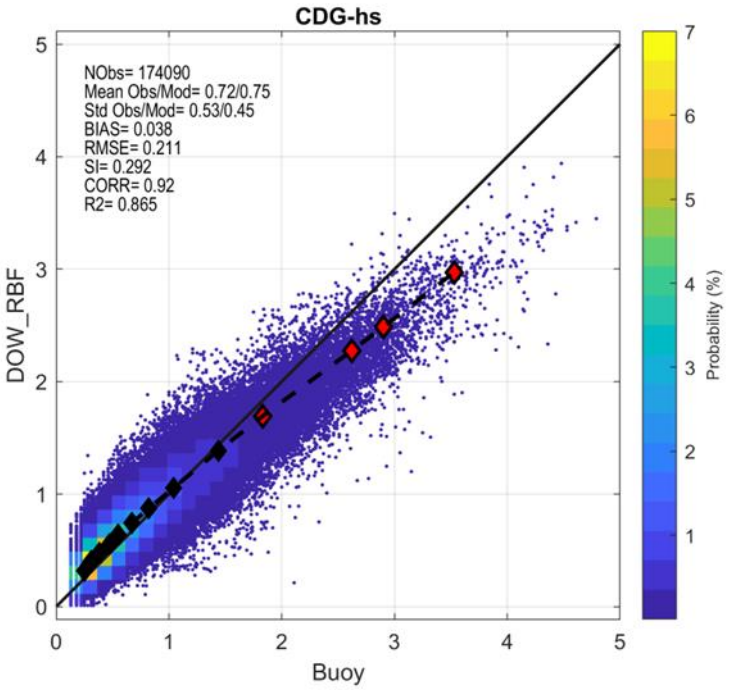

Figure 12- H_S scatter plot obtained using DOW-RBF approach for the buoy of Cabo de Gata buoy.

The reconstruction through the RBF tends to underestimate the values measured from the buoy. Many of the data, especially for the higher wave height values, are on the right-hand side of the graph. This indicates that the reconstruction returns lower wave height values than those measured by the instrument.

Figure 13 shows the entire time series for the period 1991-2012, the period for which the Cabo de Gata buoy data are available.

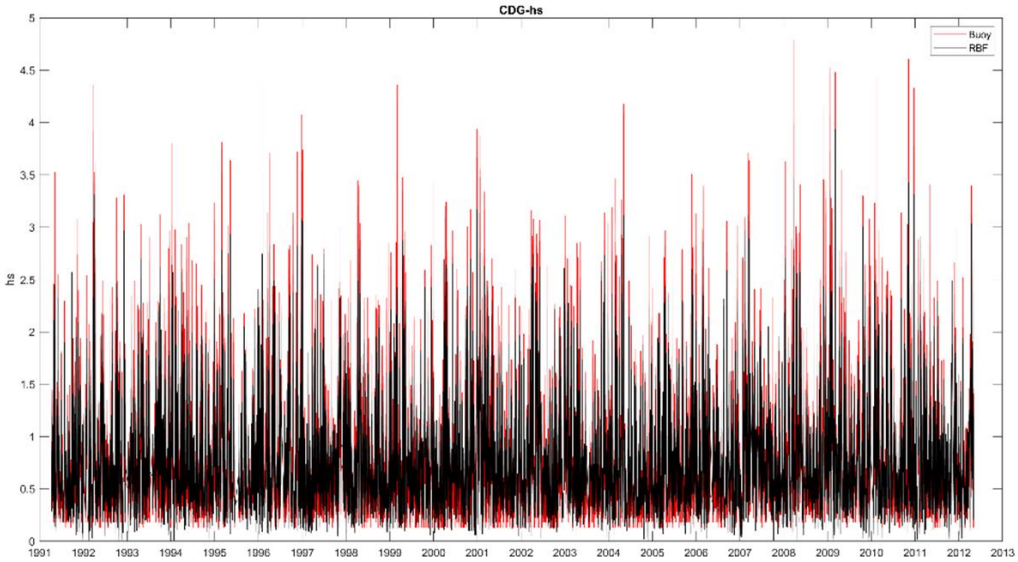

Figure 13-Comparisons between the time series measured from the Cabo de Gata buoy (red line) and that reconstructed at the same point with the DOW-RBF (black line) during the period 1991-2012 for the H_S parameter.

In conformity with the previous analysis, it is evident from the graphs that the wave height data reconstructed through the RBF method are generally lower than those measured from the buoy. This is particularly evident for the higher H_S values.

For example, the maximum wave height value measured by the instrument during the above period in the year 2008 is about 4.8 meters. The same value obtained through RBF reconstruction is about 3.3 meters.

5.2) ANN reconstruction

The reconstruction through ANN aims to reduce the discrepancies with respect to the data measured by the buoy encountered with the RBF method.

Initially, we chose to process the network with multiple combinations of neurons, layers, and number of iterations to see how the model fits to the input data (Table 2).

	Neurons	Layers	Iterations
ANN	2	1	5
	4	2	
	6	3	
	8	4	

Table 2- Combinations of neurons, layers, and iterations used.

The number of iterations used in this first phase of analysis is quite low and has been set at 5, so as not to burden the analyses too much from a computational time point of view, particularly with regard to more complex neural networks. From the network processing, it can be seen that the reconstruction algorithm has difficulties to fit the Almeria buoy, as already seen in the analysis performed with the RBF (Figures 9-10-11), especially regarding the mean period, but also with regard to the wave height. Two examples are shown for both variables in Figure 14, using a simple neural network consisting of 2 neurons and 2 layers and a more complex one consisting of 8 neurons and 4 layers (Figure 15).

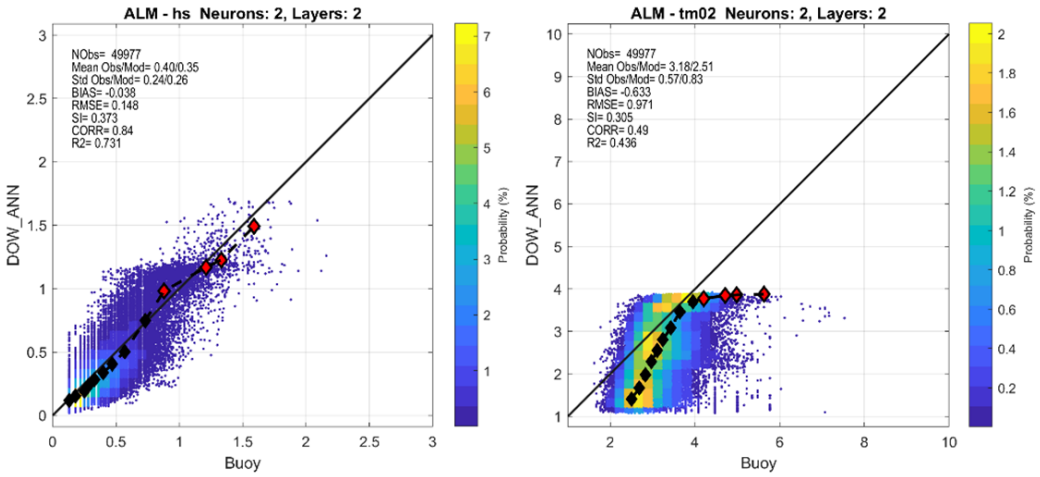

Figure 14- Scatter plot of the H_S and Tm02 parameters estimated using the DOW-ANN approach (2 neurons, 2 layers) for the Almeria buoy.

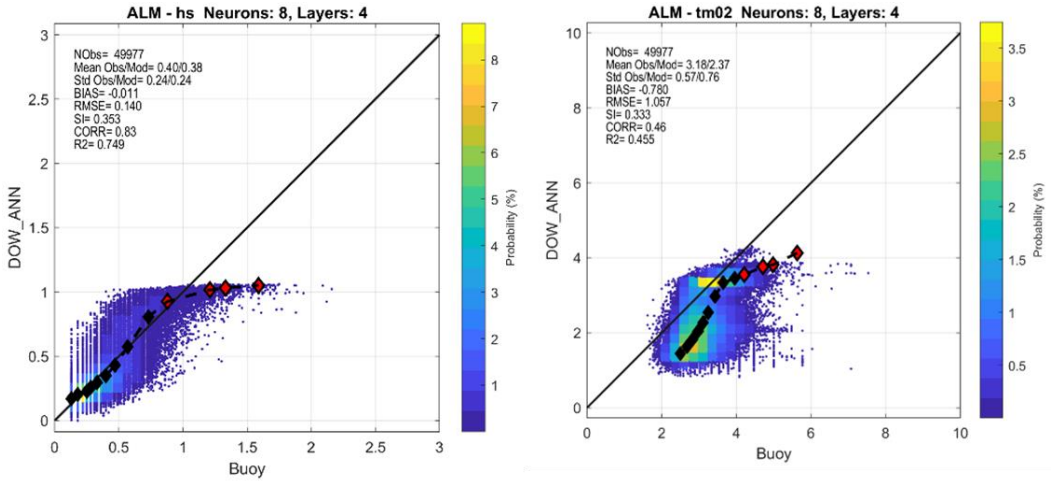

Figure 15- Scatter plot of the H_S and the Tm02 parameters estimated using the DOW-ANN approach (8 neurons, 4 layers) for the Almeria buoy.

Reconstructing the time series at the Almeria buoy is complex due to the characteristics of the area in which the buoy is located. The instrument is located within the port of Almeria. Marine dynamics near areas such as harbors are complex due to phenomena that modify wave properties, especially for the wave period. Indeed, reflection phenomena are triggered within ports that models fail to interpret correctly. As proof of this, the data obtained from the reconstruction performed in the vicinity of the Cabo de Gata buoy, which geographically is not very far from that of Almeria, agree much better to those measured by the instrumentation. To demonstrate this, scatter plots obtained near the Cabo de Gata buoy are shown below. The same combinations of neurons and layers seen previously are shown, i.e., 2 neurons and 2 layers (Figure 16) and 8 neurons and 4 layers (Figure 17).

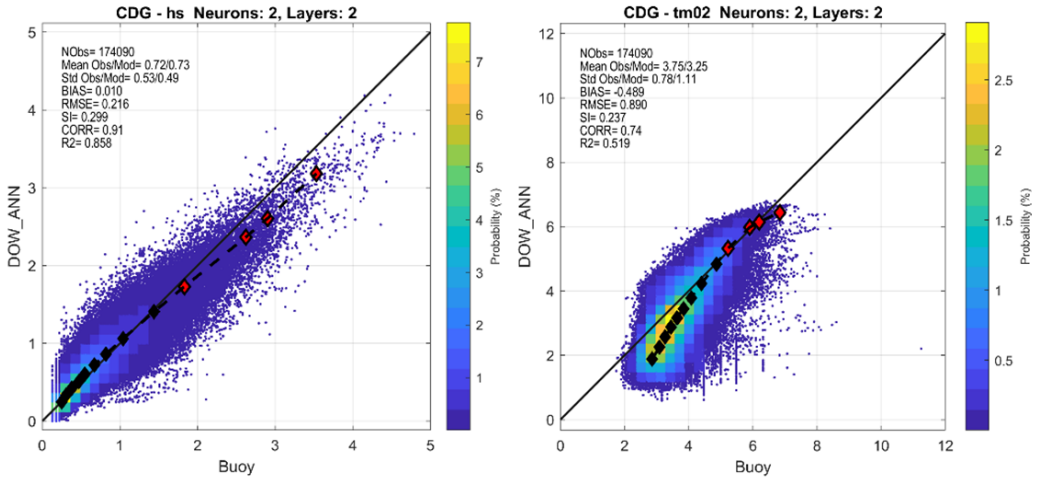

Figure 16- Scatter plot of the H_S and the Tm02 parameters estimated using the DOW-ANN approach (2 neurons, 2 layers) for the Cabo de Gata buoy.

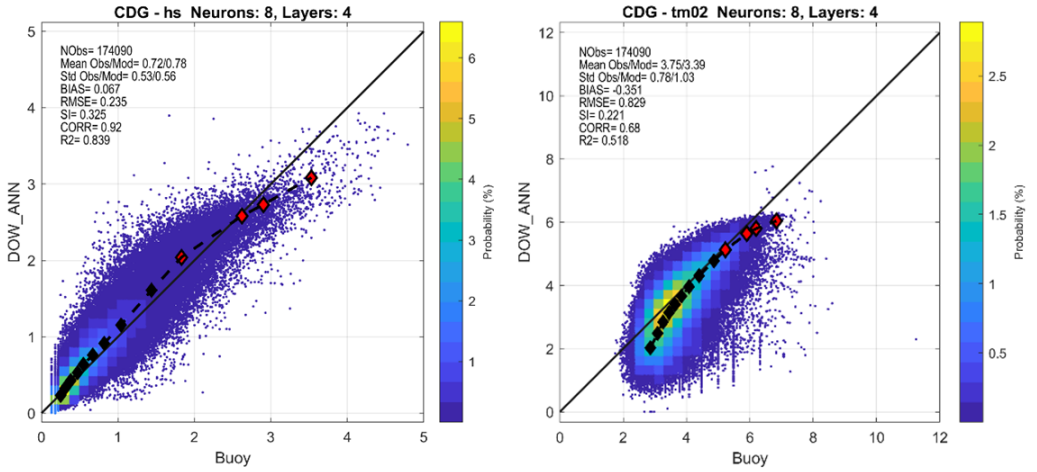

Figure 17- Scatter plot of the H_S and the Tm02 parameters estimated using the DOW-ANN approach (8 neurons, 4 layers) for the Cabo de Gata buoy.

The reconstruction through ANN confirmed the results obtained with RBFs, showing that the most accurately reconstructed parameter is the H_S for the Cabo de Gata buoy. Thus, as it was the case for the RBF (Chapter 5.1), the ANN analysis will focus on the H_S.

Once defined the target location and variable (H_S in Cabo de Gata) multiple ANN were tested with the combinations of neurons and layers seen before, but with a variable number of iterations in order to evaluate the impact of these three parameters on the results. After processing the network several times, it was seen that an increase in the number of neurons and layers does not correspond to more

reliable estimates for the parameter of interest. Some scatter plots related to the H_S parameter for the Cabo de Gata buoy are given as examples (Figures 18-21).

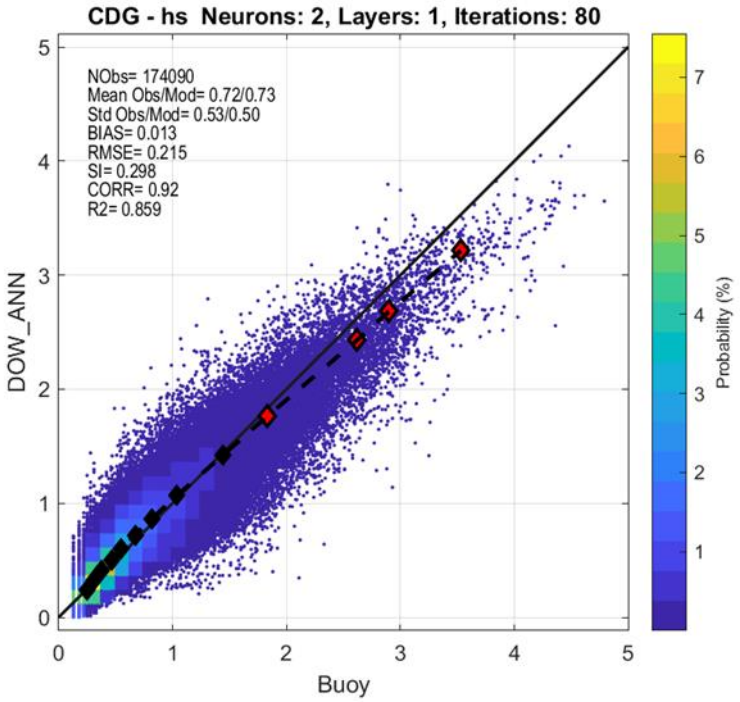

Figure 18-H_S scatter plot obtained using the DOW-ANN approach (2 neurons, 1 layer with 80 iterations) for the Cabo de Gata buoy.

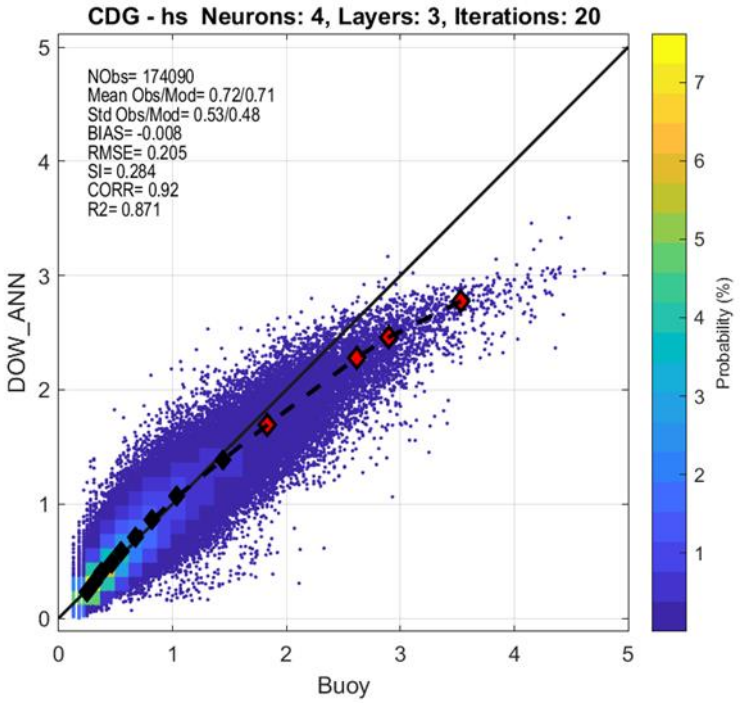

Figure 19- H_S scatter plot obtained using the DOW-ANN approach (4 neurons, 3 layer with 20 iterations) for the Cabo de Gata buoy.

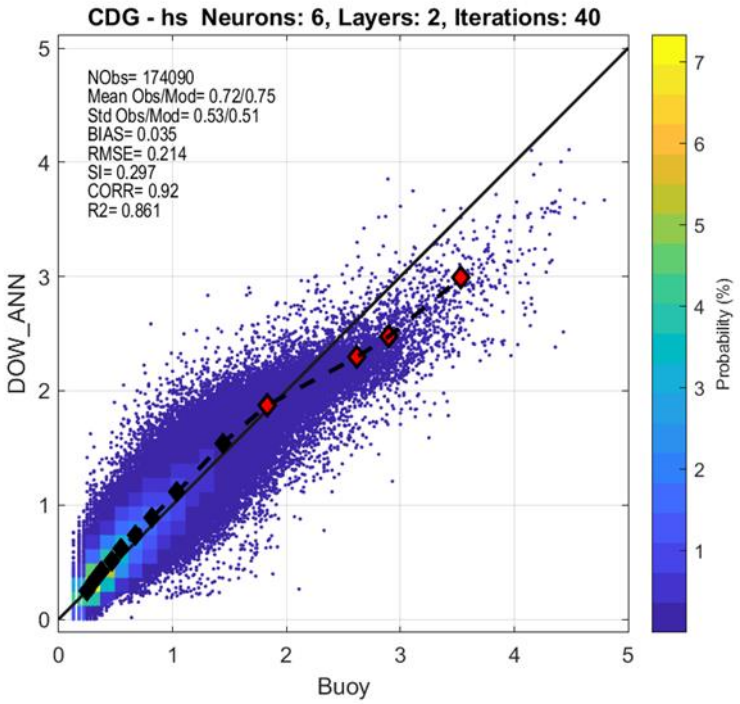

Figure 20- H_S scatter plot obtained using the DOW-ANN approach (6 neurons, 2 layer with 40 iterations) for the Cabo de Gata buoy.

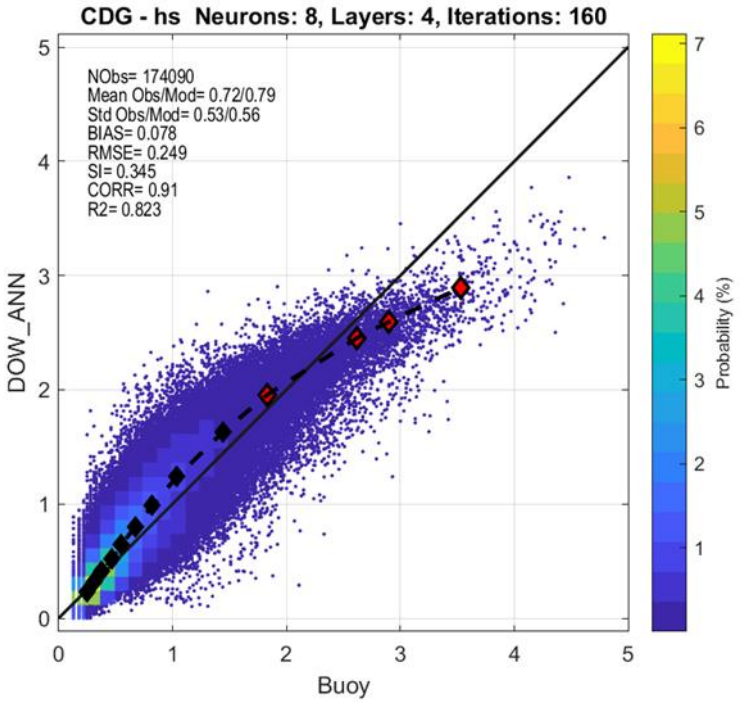

Figure 21- H_S scatter plot obtained using the DOW-ANN approach (8 neurons, 4 layers with 160 iterations) for the Cabo de Gata buoy.

As the number of neurons and layers increases, the network tends to return less satisfactory results, suggesting overfitting issues.

The number of iterations, which is representative of the number of times the training algorithm crosses the entire dataset during the learning process, markedly influences the stability of the model. During each iteration, the neural network updates its parameters (weights and bias) to progressively limit the difference between the returned predictions and the measured values. The appropriate number of iterations is closely related to factors such as the complexity of the problem, the dimensionality of the data that are part of the training set, the structural characteristics of the neural network, and the optimization algorithm used.

To evaluate the most appropriate number of iterations for the case study, it was decided to analyze the minimum test error in relation to the number of iterations for each combination of neurons and layers. (Figure 22).

As seen earlier, test error is a parameter that returns important information about the stability of the neural model.

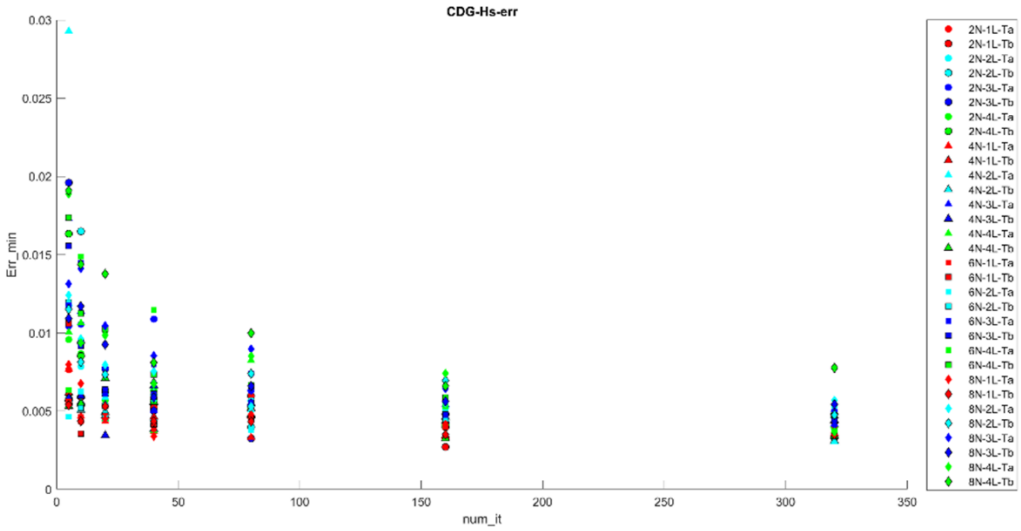

Figure 22- Minimum test error related to the number of iterations for each combination of neurons and layers.

From Figure 22, it can be seen that the error tends to stabilize sharply around 120 iterations. Moreover, this confirms that for the same number of iterations the networks minimizing the test error are the simplest ones (low number of neurons and layers).

Considering this, we chose to study the simplest networks in more detail, and more specifically those that had a number of neurons equal to two. To prove their stability, the network was processed several times. Figure 23 shows the results from one of the network processing as an example.

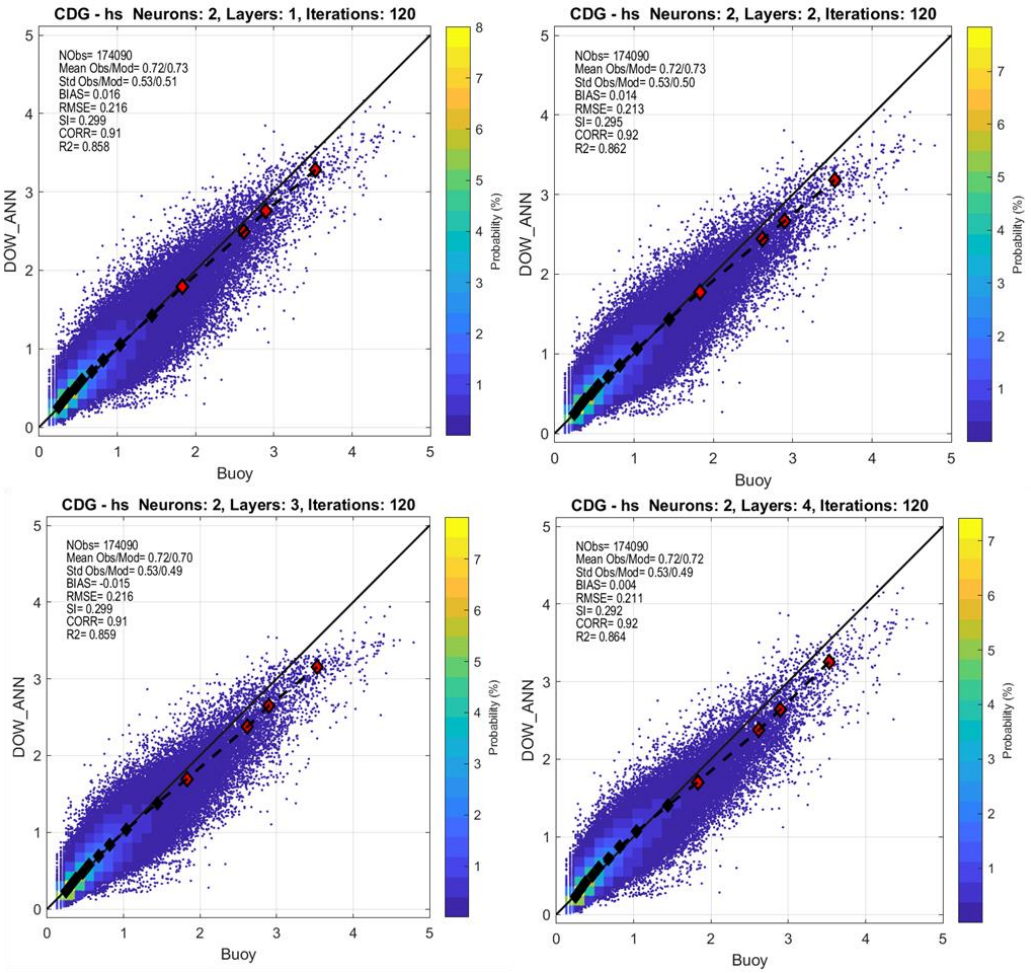

Figure 23- H_S scatter plots obtained using the DOW-ANN approach (2 neurons and a number of layers ranging from 1 to 4, with 120 iterations) for the Cabo de Gata buoy.

It was therefore decided to use a neural network consisting of 2 neurons and 1 layer with 120 iterations. Again, the network was tested several times to evaluate its stability. Figure 24 shows the scatter plots obtained from this analysis, performed 8 times.

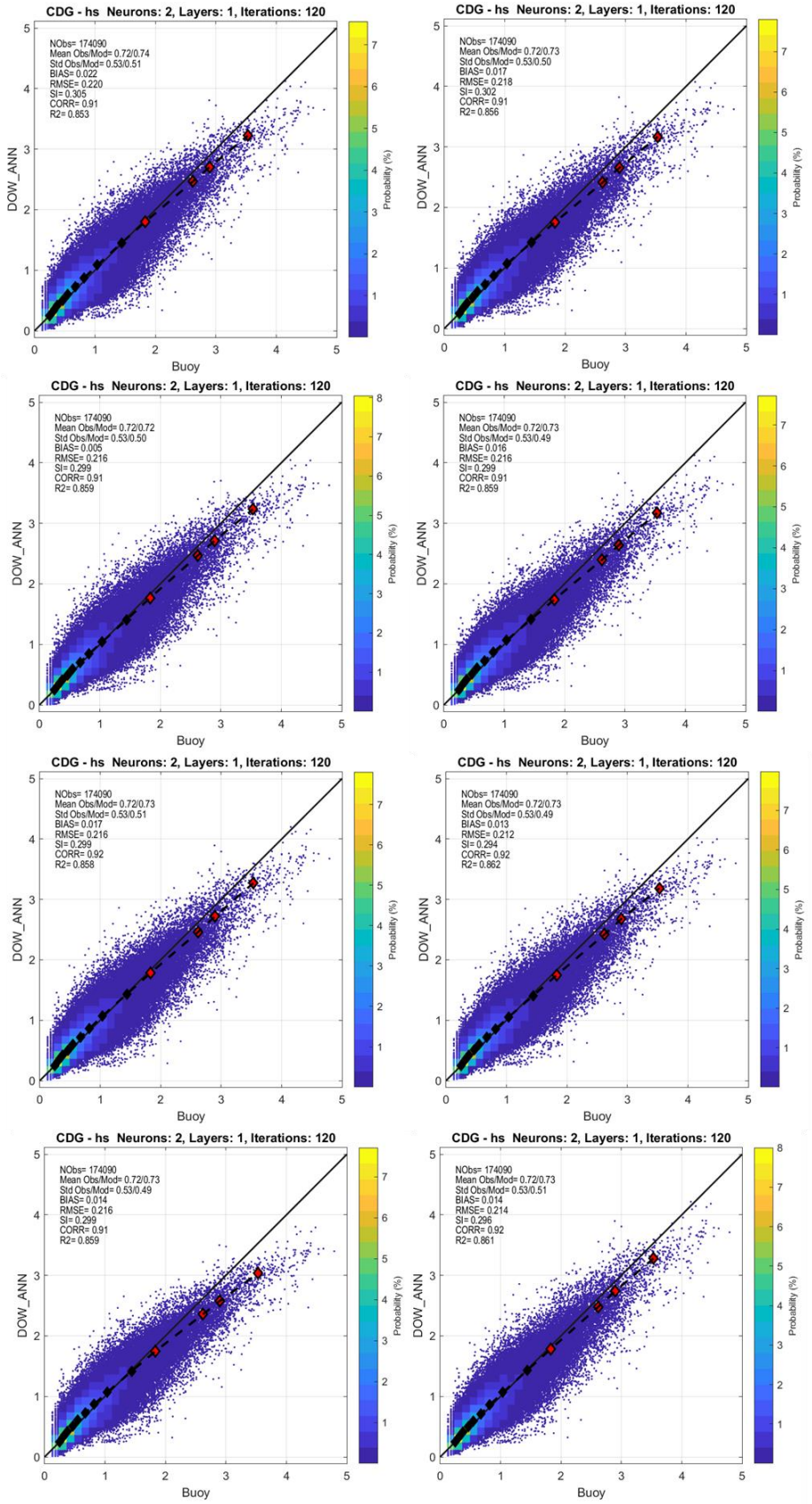

Figure 24- H_S scatter plots obtained using the DOW-ANN approach (2 neurons, 1 layer with 120 iterations) for the Cabo de Gata buoy (8 tests).

The complete time series of the H_S parameter related to the last scatter plot in the previous figure is shown in Figure 25.

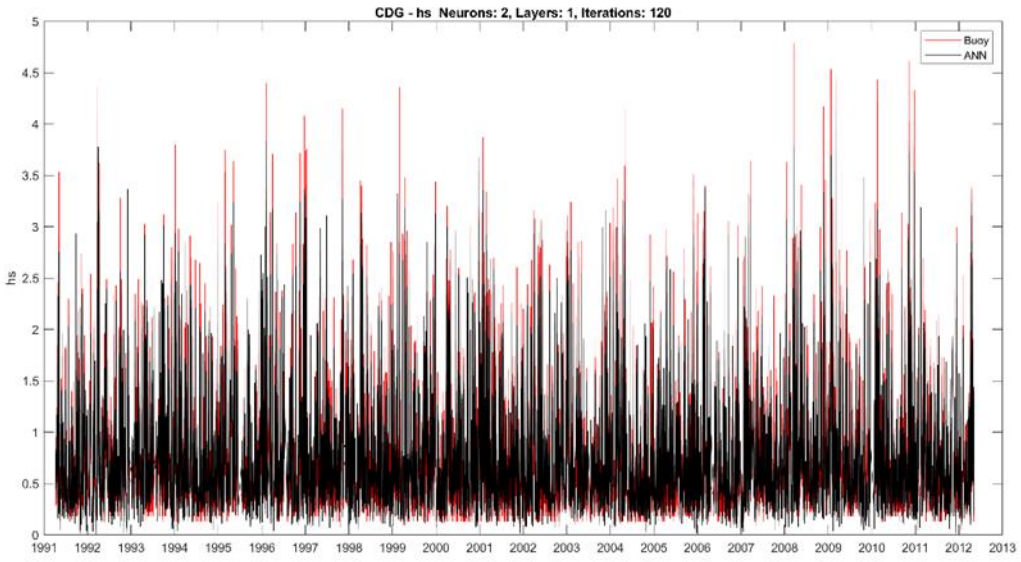

Figure 25- Comparison between the time series measured from the Cabo de Gato buoy (red line) and that reconstructed at the same point with DOW-ANN (black line) during the period 1991-2012 for the H_S parameter (8° test).

By analyzing the plots obtained and the error metrics introduced in the Chapter 4.2, it can be seen that the neural model has acquired good stability through the use of the structure formed by 2 neurons, 1 layer, and a number of iterations equal to 120 (Table 3).

ANN	BIAS	RMSE	SI	CORR	R2
Test 1	0.022	0.220	0.305	0.91	0.853
Test 2	0.017	0.218	0.302	0.91	0.856
Test 3	0.005	0.216	0.299	0.91	0.859
Test 4	0.016	0.216	0.299	0.91	0.859
Test 5	0.017	0.216	0.299	0.92	0.858
Test 6	0.013	0.212	0.294	0.92	0.862
Test 7	0.014	0.216	0.299	0.91	0.859
Test 8	0.014	0.214	0.296	0.92	0.861

Table 3- Error metrics of each DOW-ANN test (2 neurons, 1 layer, 120 iterations).

With regard to H_S, the reconstruction with ANN, as it was seen for RBFs, tends to underestimate the values measured by the buoy (Figure 24). Looking at the maximum H_S value measured from the Cabo de Gata buoy, the reconstruction using ANN returns a value approximately equal to 3.7 m.

5.3) Reconstruction skill comparison

A very important aspect of the results obtained with reconstruction using ANN is the better accuracy regarding the most extreme values with respect to the reconstruction using RBFs. Indeed, all the tests performed with ANN return better results in comparison with the reconstruction pursued with the RBF. It can be seen from the scatter plots in Figure 26 how the highest values of H_S are closer to the datum recorded by the buoy in the reconstruction performed with ANN than that obtained by employing RBFs. The scatter plot for the ANN reconstruction shown in the Figure 26 is referring to last test shown in Figure 24.

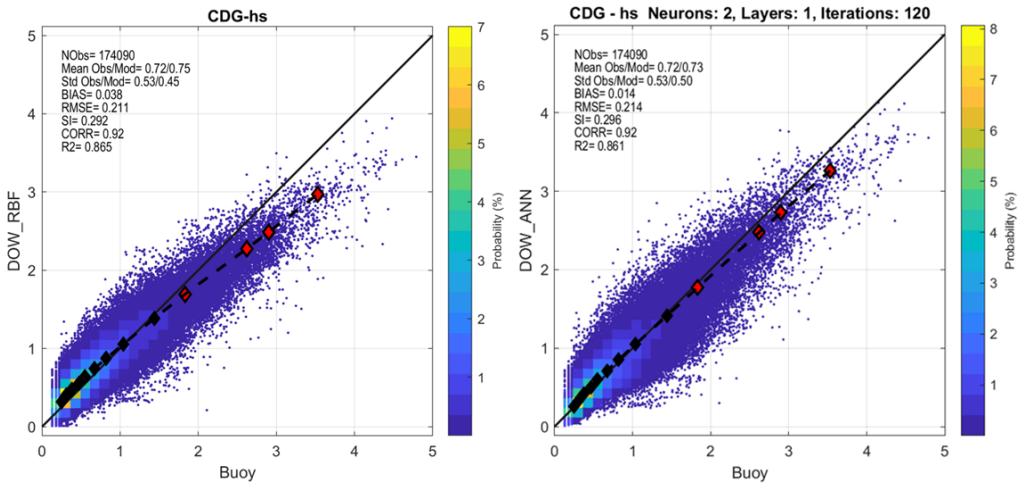

Figure 26- H_S scatter plot obtained using the DOW-RBF approach (left) and the DOW-ANN approach (left) for the Cabo de Gata buoy.

The comparison of the two scatter plots obtained with the RBF and ANN reconstruction techniques allows for a more immediate visualization of the difference between the maximum estimated H_S value and the maximum value measured by the buoy. In fact, the first method returns a value approximately equal to 3.3 m, the second approximately equal to 3.7 m. Although this may seem a small difference, it may be critical for the design of coastal defense works.

As seen already in the comparison of the two-time series reconstruction methods performed earlier, the ANN seems to return more accurate estimates of extreme wave height conditions, i.e. potentially more destructive events.

Therefore, it was decided to perform an analysis that considers only the higher quantiles of the H_S parameter. This test was performed using the results obtained from the last experiment conducted with the neural network chosen for this study (last scatter plot in Figure 24).

Three tests were performed, considering Hs values measured from the Cabo de Gata buoy above quantile 80, 90 and 90, respectively. The 99 was not selected because the number of data used would have been very limited, not ensuring adequate statistical variability.

The analysis on the highest quantiles pursues the goal of confirming the improvement obtained from reconstructions carried out with ANN compared with those obtained derived from the use of RBF, making the difference in reconstructing extreme values more easily describable and representable.

Below are the scatter plots for both methods used in this study, in the framework of the analysis at quantile 80 (Figure 27), 90 (Figure 28), and 95 (Figure 29).

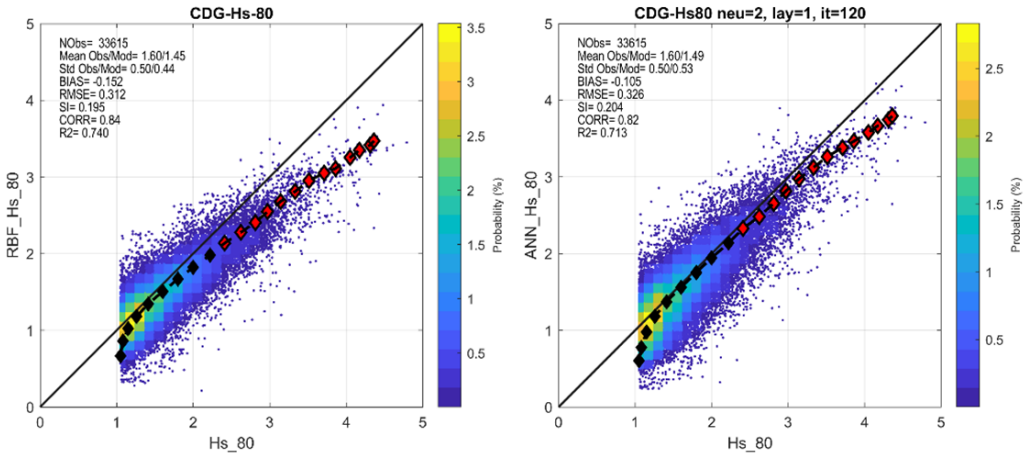

Figure 27- H_S scatter plot at quantile 80 obtained using the DOW-RBF approach (left) and the DOW-ANN approach (right) for the Cabo de Gata buoy.

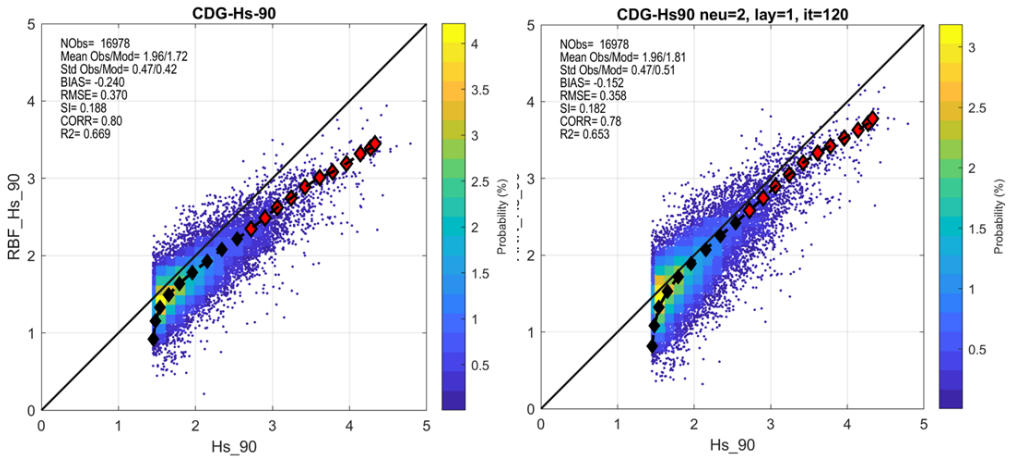

Figure 28- H_S scatter plot at quantile 90 obtained using the DOW-RBF approach (left) and the DOW-ANN approach (right) for the Cabo de Gata buoy.

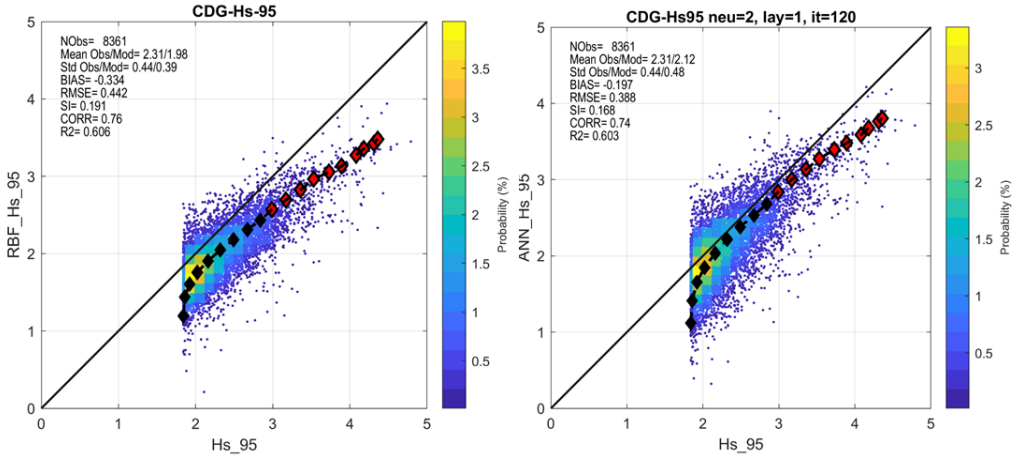

Figure 29- H_S scatter plot at quantile 95 obtained using the DOW-RBF approach (left) and the DOW-ANN approach (right) for the Cabo de Gata buoy.

Results show clear differences between the outcomes from reconstruction through ANN and RBF. This can be appreciated both visually, but also through the analysis of the main error metrics calculated.

In particular, the discrepancy between the BIAS values obtained with the two reconstruction techniques is quite stark and allows us to understand how, especially in the analysis of extreme values, the reconstruction by means of ANN gives values significantly closer to those measured by the buoy (Table 4).

BIAS	RBF	ANN
Quantile 80	-0,152	-0,105
Quantile 90	-0,240	-0,152
Quantile 95	-0,334	-0,197

Table 4- BIAS values at quantiles 80, 90 and 95 obtained using the DOW-RBF (left column) and the DOW-ANN approach (right column).

As a further confirmation, we also note that by focusing the analysis on progressively higher quantiles, the difference in the BIAS value between the results obtained applying the two reconstruction methods is higher.

The RMSE also shows an interesting trend in values. As can be seen from Table 5, for higher quantiles the ANN tends to give results that are more comparable to those measured by the buoy.

RMSE	RBF	ANN
Quantile 80	0.312	0.326
Quantile 90	0.370	0.358
Quantile 95	0.442	0.388

Table 5- RMSE values at quantiles 80, 90 and 95 obtained using the DOW-RBF (left column) and the DOW-ANN approach (right column).

Regarding SI, it shows a downward trend for higher quantiles in the reconstruction with ANN, while it is relatively stable in the reconstruction with RBF (Table 6). This testifies that for higher quantiles the data returned by ANN shows less dispersion than that of RBF.

SI	RBF	ANN
Quantile 80	0.195	0.204
Quantile 90	0.188	0.182
Quantile 95	0.191	0.168

Table 6- SI values at quantiles 80, 90 and 95 obtained using the DOW-RBF (left column) and the DOW-ANN approach (right column).

With regard to the other error metrics (CORR and R2), there are no particular differences between one reconstruction method and another.

Another interesting aspect when comparing the two methodologies presented is the computational time required to perform the reconstruction.
The RBF method takes about 40 seconds to process. The implementation of the neural network with the characteristics selected for this specific work (2 neurons, 1 layer, 120 iterations) enabled the reconstruction of the time series in approximately 37 seconds (always referring to the test chosen for the previous analysis). The computational times used to perform each of the two reconstructions are shown in Table 7.

	RBF	ANN
Computational time (s)	39,96	37,16

Table 7-Time comparison between RBF and ANN approaches.

6) Conclusions

This work demonstrates the applicability of two techniques to reconstruct wave hourly time series, RBF and ANN, to obtain high-resolution wave parameters at the coast.

The application of the RBF has shown how difficult it is to get an appropriate description of wave parameters in areas characterized by complex morphologies (such as, for example, the port area of Almeria), as wave modification phenomena are not properly simulated by wave propagation models.

The implementation of machine learning techniques for the reconstruction of wave time series resulted in more accurate estimates of the H_S parameter for the Cabo de Gata buoy, with respect to RBF technique. Although the data obtained through both techniques show a general underestimation compared to the measurements taken from the buoy, the use of the ANN resulted in more similar values to those measured by the instrument. This aspect has important implications in the design of coastal defense works and the protection of structures located near the seashore.

The ANN selected in this work, consisting of 2 neurons, 1 layer and 120 iterations proved to be stable each time it was processed. Further testing and more specific analysis could lead to an even more accurate estimation of the wave parameters of interest and a progressive limitation of the computational time spent processing the network.

It was chosen to intervene only in the structural features of the network (number of neurons, layers, and iterations), while no changes were made in other aspects that may still have some bearing on the stability and accuracy of the network. These include, for example, the breakdown of the input data, the initialization of the weights, the activation function, and the optimization algorithms used.

The processing times of the two reconstruction techniques are comparable, although that employed by RBF is stable in that the reconstruction is always done in the same way, while that of ANN varies between about 35 and 50 seconds in the 8 times the network chosen for the case study was processed. This variability is due to the way the partitioning of the input data takes place, which is different each time the reconstruction is performed.

The analysis performed on the most extreme values, of maximum interest in the engineering design context, suggests that ANN are more adequate than the RBF for reconstructing high values of the H_S parameter. This finding confirms again what was seen previously; once network stability issues are resolved, the application of machine learning techniques to reconstruct time series of marine parameters can provide a more accurate coastal wave time series.

7) Bibliography

1. Rangel-Buitrago N, Neal WJ, Bonetti J, Anfuso G, de Jonge VN. Vulnerability assessments as a tool for the coastal and marine hazards management: An overview. *Ocean Coast Manag*. 2020;189:105134. doi:10.1016/j.ocecoaman.2020.105134
2. Bevacqua A, Yu D, Zhang Y. Coastal vulnerability: Evolving concepts in understanding vulnerable people and places. *Environ Sci Policy*. 2018;82(November 2017):19-29. doi:10.1016/j.envsci.2018.01.006
3. Amarouche K, Akpınar A, Bachari NEI, Çakmak RE, Houma F. Evaluation of a high-resolution wave hindcast model SWAN for the West Mediterranean basin. *Appl Ocean Res*. 2019;84:225-241. doi:10.1016/j.apor.2019.01.014
4. Perez J, Menendez M, Losada IJ. GOW2: A global wave hindcast for coastal applications. *Coast Eng*. 2017;124(March):1-11. doi:10.1016/j.coastaleng.2017.03.005
5. Teixeira JC, Abreu MP, Stares CG. Uncertainty of ocean wave hindcasts due to wind modeling. *J Offshore Mech Arct Eng*. 1995;117(4):294-297. doi:10.1115/1.2827237
6. Ponce de León S, Guedes Soares C. Sensitivity of wave model predictions to wind fields in the Western Mediterranean sea. *Coast Eng*. 2008;55(11):920-929. doi:10.1016/j.coastaleng.2008.02.023
7. Camus P, Mendez FJ, Medina R, Tomas A, Izaguirre C. High resolution downscaled ocean waves (DOW) reanalysis in coastal areas. *Coast Eng*. 2013;72:56-68. doi:10.1016/j.coastaleng.2012.09.002
8. Mínguez R, Espejo A, Tomás A, Méndez FJ, Losada IJ. Directional calibration of wave reanalysis databases using instrumental data. *J Atmos Ocean Technol*. 2011;28(11):1466-1485. doi:10.1175/JTECH-D-11-00008.1
9. Adytia D, Saepudin D, Tarwidi D, et al. Modelling of Deep Learning-Based Downscaling for Wave Forecasting in Coastal Area. *Water (Switzerland)*. 2023;15(1). doi:10.3390/w15010204
10. Vannucchi V, Taddei S, Capecchi V, Bendoni M, Brandini C. Dynamical downscaling of era5 data on the north-western mediterranean sea: From atmosphere to high-resolution coastal wave climate. *J Mar Sci Eng*. 2021;9(2):1-29. doi:10.3390/jmse9020208
11. Satta A, Puddu M, Venturini S, Giupponi C. Assessment of coastal risks to climate change related impacts at the regional scale: The case of the Mediterranean region. *Int J Disaster Risk Reduct*. 2017;24:284-296. doi:10.1016/j.ijdrr.2017.06.018
12. Amarouche K, Akpınar A, Semedo A. Wave storm events in the Western Mediterranean Sea over four decades. *Ocean Model*. 2022;170. doi:10.1016/j.ocemod.2021.101933
13. Bacon S, Carter DJT. Wave climate changes in the North Atlantic and North Sea. *Int J Climatol*. 1991;11(5):545-558. doi:10.1002/joc.3370110507
14. Molina R, Manno G, Re C Lo, Anfuso G, Ciraolo G. Storm energy flux characterization along the mediterranean coast of Andalusia (Spain). *Water (Switzerland)*. 2019;11(3). doi:10.3390/w11030509
15. Molina R, Manno G, Re C Lo, Anfuso G, Ciraolo G. A methodological approach to determine sound response modalities to coastal erosion processes in mediterranean Andalusia (Spain). *J Mar Sci Eng*. 2020;8(3). doi:10.3390/jmse8030154

16. Molina R, Anfuso G, Manno G, Prieto FJG. The Mediterranean coast of Andalusia (Spain): Medium-term evolution and impacts of coastal structures. *Sustain*. 2019;11(13). doi:10.3390/su11133539
17. Williams AT, Micallef A, Anfuso G, Gallego-Fernandez JB. Andalusia, Spain: An Assessment of Coastal Scenery. *Landsc Res*. 2012;37(3):327-349. doi:10.1080/01426397.2011.590586
18. Holthuijsen LH. *Waves in Oceanic and Coastal Waters*.; 2008.
19. Camus P, Mendez FJ, Medina R. A hybrid efficient method to downscale wave climate to coastal areas. *Coast Eng*. 2011;58(9):851-862. doi:10.1016/j.coastaleng.2011.05.007
20. Hersbach H, Bell B, Berrisford P, et al. The ERA5 global reanalysis. *Q J R Meteorol Soc*. 2020;146(730):1999-2049. doi:10.1002/qj.3803
21. Susini S, Menendez M, Eguia P, Blanco JM. Climate Change Impact on the Offshore Wind Energy Over the North Sea and the Irish Sea. *Front Energy Res*. 2022;10(May):1-17. doi:10.3389/fenrg.2022.881146
22. Olauson J. ERA5: The new champion of wind power modelling? *Renew Energy*. 2018;126:322-331. doi:10.1016/j.renene.2018.03.056
23. Reguero BG, Menéndez M, Méndez FJ, Mínguez R, Losada IJ. A Global Ocean Wave (GOW) calibrated reanalysis from 1948 onwards. *Coast Eng*. 2012;65:38-55. doi:10.1016/j.coastaleng.2012.03.003
24. Ris RC, Holthuijsen LH, Booij N. A third-generation wave model for coastal regions 2. Verification. *J Geophys Res Ocean*. 1999;104(C4):7667-7681. doi:10.1029/1998jc900123
25. Kalra R, Deo MC, Kumar R, Agarwal VK. Artificial neural network to translate offshore satellite wave data to coastal locations. *Ocean Eng*. 2005;32(16):1917-1932. doi:10.1016/j.oceaneng.2005.01.007
26. Afzal MS, Kumar L, Chugh V, Kumar Y, Zuhair M. Prediction of significant wave height using machine learning and its application to extreme wave analysis. *J Earth Syst Sci*. 2023;132(2). doi:10.1007/s12040-023-02058-5
27. Rajindas KP, Shashikala AP. Development of hybrid wave transformation methodology and its application on Kerala Coast, India. *J Earth Syst Sci*. 2021;130(2). doi:10.1007/s12040-021-01612-3
28. Rodriguez-Delgado C, Bergillos RJ, Iglesias G. An artificial neural network model of coastal erosion mitigation through wave farms. *Environ Model Softw*. 2019;119:390-399. doi:10.1016/j.envsoft.2019.07.010
29. Shamshirband S, Mosavi A, Rabczuk T, Nabipour N, Chau K wing. Prediction of significant wave height; comparison between nested grid numerical model, and machine learning models of artificial neural networks, extreme learning and support vector machines. *Eng Appl Comput Fluid Mech*. 2020;14(1):805-817. doi:10.1080/19942060.2020.1773932