Article

Building Suitable Datasets for Soft Computing and Machine Learning Techniques from Meteorological Data Integration: A Case Study for Predicting Significant Wave Height and Energy Flux

Antonio Manuel Gómez-Orellana [1*], Juan Carlos Fernández [1*], Manuel Dorado-Moreno [1*], Pedro Antonio Gutiérrez [1*] and César Hervás-Martínez [1*]

[1] Department of Computer Science and Numerical Analysis, University of Cordoba, 14071, Córdoba, Spain.
* Correspondence: am.gomez@uco.es (A.M.G.-O); jfcaballero@uco.es (J.C.F.); manuel.dorado@uco.es (M.D.-M.); pagutierrez@uco.es (P.A.G.); chervas@uco.es (C.H.-M.)

Version December 23, 2020 submitted to Energies

Abstract: Meteorological data are extensively used to perform environmental learning. Soft Computing (SC) and Machine Learning (ML) techniques represent a valuable support in many research areas, but require datasets containing information related to the topic under study. Such datasets are not always available in an appropriate format and its preparation and pre-processing implies a lot of time and effort by researchers. This paper presents a novel software tool with an user-friendly GUI to create datasets by means of management and data integration of meteorological observations from two well-known sources of information: the *National Data Buoy Center* and the *National Centers for Environmental Prediction and for Atmospheric Research Reanalysis Project*. Such datasets can be created using buoys and reanalysis data through customisable procedures, in terms of temporal resolution, predictive and objective variables, and can be used by SC and ML methodologies for prediction tasks (classification or regression). The objective is providing the research community with an automated and versatile system for the casuistry that entails well-formed and quality data integration, potentially leading to better prediction models. The software tool can be used as a supporting tool for coastal and ocean engineering applications, sustainable energy production or environmental modelling; as well as for decision making in the design and construction of coastal protection structures, marine transportation, offshore industry, ocean energy converters and efficient operation of offshore and coastal engineering activities. Finally, to illustrate the applicability of the proposed tool, a case study to classify waves depending on their significant height and to predict energy flux in the Gulf of Alaska is presented.

Keywords: Environmental Prediction; Renewable Energy Resource Evaluation; Meteorological Data; Reanalysis Data; Marine Energy; Soft Computing

1. Introduction

A better understanding of the environment is of vital importance for science, contributing not only to more efficient exploitation of natural resources but also to the development of new strategies aimed at its protection. In that sense, meteorological observations provide an essential and valuable source of information which is widely used by researchers to address environmental learning, comprehension, prediction and conservation in numerous oceanic and atmospheric studies of a wide variety of areas (e.g. energy, climate change, agriculture, etc.). Some specific examples of the diversity of fields in which meteorological data can be used in are, among others: estimation of global solar radiation based on sunshine duration [1], directional analysis of sea storms [2], estimation of hybrid energy systems taking into account economic and environmental objectives[3], wind power ramp events prediction [4], sea surface temperature prediction [5], study of the responses exhibited by plankton to fluid motions [6], trends in solar radiation [7] or simulation of extreme near shore sea conditions [8]. All these studies require a prior data collection and its adaptation to a specific format that allows the interpretation of them.

Once quality and well-formed data are obtained, these can be used to extract information and build prediction models that explain the behavior of a certain problem. The choice of the appropriate model, either in engineering problems or in any other problem, is also an important factor in addition to data [9,10]. Although Soft Computing (SC) and Machine Learning (ML) techniques have the ability to handle uncertainty in data and are extensively used for modelling purposes, the real challenge in modelling studies is due to the inadequacy of data, since the adequacy of the models depend mainly on the the quality of the information used, so that if a researcher does not have quality data there will be no quality models.

Continuing with this line, special purpose software is usually developed to help researchers to advance in their studies related to energy and environmental modelling, becoming a great support for decision-making in the exploitation and protection of the environment. In [11], a software package in R called "ForecastTB" is developed to compare the accuracy of different prediction methods as related to the characteristics of a time series dataset, presenting the software as a stepping stone in ML automation modelling. In [12], an integrated simulation tool for the optimum design of bifacial solar panel with reflectors is presented. This tool can also be used to analyse the performance of the solar cells. A framework for data integration of offshore wind farms is implemented in [13] in order to facilitate data exchange and improve operation and maintenance practices. In [14], a new software tool named "Storage LCA Tool" is presented for comparing PCM - phase change materials- storage systems with traditional systems that do not entail energy storage, being beneficial in order to support decision-making on energy concepts for buildings. A risk assessment tool to improve safety standards and emergency management in onshore wind farms, is presented in [15]. Raabe et al. [16] developed two software tools, *Model of Equilibrium of Bay Beaches* (MEPBAY) and *Coastal Modelling System* (SMC), to support different operational levels of headland-bay beach in coastal engineering projects, and Motahhir et al. [17] developed an open hardware/software test bench for solar tracker.

Marine energy prediction is currently a hot topic where meteorological data is used in. Marine Renewable Energy (MRE) is one of the most important renewable and sustainable energy sources available in our environment [18], and it includes ocean thermal energy, marine tidal current energy and wave energy, among others. Its benefits and great potential [19] make it one of the most relevant natural resources, playing a crucial role not only in the reduction of the emission of greenhouse gases but also in all other aspects involved in the difficult challenge of the transition to a low carbon footprint society [20–22]. Wave energy exhibits a more stable power supply than wind energy and even solar energy. In recent years *Wave Energy Converters* (WECs) [23] have been developed and widely installed to transform this wave energy into electricity, which can be injected into the electric network or supplied to existing offshore oil and gas platforms [24] or seawater desalination plants [25], among others. WECs are mechanical devices that convert kinetic energy into electrical energy by means of either the vertical oscillation of waves or the linear motion of them. Nevertheless, waves are difficult to be characterised due to their stochastic nature, because of the influence of a large number of environmental factors that exert on them [26]. As a consequence of this complexity, many aspects of WEC design, deployment and operation [27–29] need a proper prediction of waves [30,31], in order to maximise the wave energy extraction [32]. For this purpose WECs use wave *flux of energy* (F_e) which can be calculated from the two most important wave parameters in this regard: *significant wave height* (H_s) and *wave energy period* (T_e).

Currently, and as a support to traditional study procedures, SC and ML techniques [33,34] are being widely used in numerous research fields related to classification, regression and optimisation tasks, obtaining significant improvements in the performance of the results, either in engineering

[10], energy or environmental problems [35–37]. SC and ML methodologies can be used not only by experienced computer scientists but also by other researchers. For example, the well-known *Waikato Environment for Knowledge Analysis* (WEKA) [38] software tool provides researchers with a wide collection of ML algorithms. ML techniques have been already applied to tackle wave characterisation, accurately estimating H_s and T_e parameters [39,40], given that robustness of ML methods can tackle the previously explained difficulties in wave energy prediction. In [41], a reliable ML model based on multiple linear regression and covariant-weighted least square estimation for H_s modelling is presented in order to forecast half-hourly nearly real-time significant wave height values. In [42], an approach for feature selection problems is developed and applied to a specific problem of H_s and F_e prediction in oceanic buoys, obtaining excellent results in terms of prediction quality. In [43], a Bayesian Network system provides an useful tool for the decision making process of installation and maintenance operations in offshore wind farms using predictions of H_s, among others. In [44], several ML methods are implemented and compared for the prediction of H_s in the Persian Gulf, the extreme learning machine (ELM) providing the best results. The problem is that, in order to apply ML and SC techniques, it is essential to obtain datasets with relevant information about the issue under study, used to infer knowledge. Usually, these datasets are not publicly available in a friendly format, and their generation is the first step needed.

The information to create these datasets related with MRE can be obtained from meteorological observations, but such information may be available in a inappropriate format and even contain missing values or measurements. Consequently, it is usually required to perform pre-processing tasks for improving the quality of the data, such as the replacement of missing values, outlier detection or data normalisation, among others. Furthermore, if more than one source of information is used to achieve a better characterisation of the problem under study [45–47], then a data integration process, denominated as the matching process in this document, has to be carried out by researchers to manually create the datasets with the needed information. Given that such process is of great relevance and has an extensive casuistry, the present work has been specially focused on it. Moreover, depending on the subject and the SC and ML technique to be applied, or even if the researcher considers other factors in order to improve the results or have more in-depth conclusions, the datasets would have to be updated afterwards. In summary, many important details and different intermediate steps have to be considered when creating suitable datasets, specially when data integration is required, resulting in an extremely tedious task.

The main purpose of this paper is to present a new open source tool for the creation of datasets integrated by meteorological variables from two sources of information. Given that the tool provides an user-friendly graphical interface, no knowledge in programming languages is needed. It also prevents researchers from performing the mentioned tedious work and greatly simplify all the steps involved in it, avoiding possible errors in the intermediate steps, at least as a preliminary study in certain areas where some kind of environmental prediction is needed. The meteorological data used by the tool come from two well-known sources of information: the *National Oceanic and Atmospheric Administration* (NOAA) *National Data Buoy Center* (NDBC) [48] and the *National Centers for Environmental Prediction* (NCEP)/*National Center for Atmospheric Research* (NCAR) *Reanalysis Project* (NNRP or R1) [49,50]. The open source software tool presented in this work is named SPAMDA (Software for Pre-processing and Analysis of Meteorological DAta to build datasets). As SPAMDA performs all this data processing, it reduces the time involving these tasks and allows researchers focus on the study of the meteorological aspects of the observations. The datasets obtained are ready to be used as input for SC and ML techniques in prediction tasks (classification or regression), although researchers can use them for other purposes. These datasets contain one or more meteorological variables as inputs and one variable as target (variable to be predicted). The format of the generated datasets will be *Attribute-Relation File Format* (ARFF) [51], which is the one used by WEKA. Besides, the datasets can also be generated in *Comma-Separated Values* (CSV) format, enabling researchers to use others tools.

In order to address the problem previously discussed, meteorological data integration from NDBC and NNRP and the casuistry that it entails, SPAMDA offers to researchers novelties and functionalities that will be detailed in Section 3, although some of them are briefly summarised below:

- The generation of datasets becomes a very easy and customisable task by means of the selection of different input parameters, such as predictive and objective variables, classification and regression, output discretisation (useful for ordinal regression) or prediction horizon, among others.
- The created datasets can be easily used by SC and ML tools.
- It makes the researcher focus on environmental modelling, without having to worry about the development of scripts or mechanical tasks, avoiding laborious pre-processing procedures, that imply a lot of time and effort in early stages of the research.
- It avoids possible researcher errors in the intermediate steps of the process, such as geographical coordinates conversion, missing values handling (dates or measurements not recorded) or different temporal resolution of the data collected, among others.
- It provides information about the quality and quantity of the data. SPAMDA allows preliminary studies of missing values (dates or measurements not recorded) in buoys managed by NDBC, so that the researcher can have an idea of the quality of the data recorded by the buoys and about their suitability for the intended purpose. In any case, SPAMDA allows data integration taking into account such missing values when needed by the user.
- Estimation of the amount of energy flux that can be produced at different prediction horizons: short-term, mid-term or long-term. Although this work does not focus on models performance, it should be taken into account that models tend to generalise worse with greater prediction horizons.
- It manages the extensive casuistry of data integration which can lead to incomplete datasets, described in Appendix A.
- Possibility of selecting one or more reanalysis nodes near the localisation under study, which could provide a better description of the problem to achieve more accurate models.
- Although pre-processing is not the main objective of SPAMDA, the tool also provides some basic pre-processing filters on buoys measurements, such as normalisation and missing data recovery.
- It facilitates data management and well-organised storage of the datasets. Environmental studies in different geographical locations can be carried out by merely introducing and using other collected data.
- SPAMDA is distributed as open source tool, its modular design allows the implementation of new modules for managing meteorological data from others sources, benefiting future renewable energy and environmental research.
- It includes a user-friendly GUI, facilitating and greatly simplifying data management, and it is integrated with the Explorer environment of WEKA.
- It is multi-platform, and it can be used on any computer with Java regardless of the operating system.

Therefore, the functionalities and characteristics that SPAMDA offers make it a supporting tool for researchers, which could be used in applications related to coastal and ocean engineering, and also in marine energy prediction. In [3], the estimation of energy supply sources in hybrid energy systems is based on the amount of energy that can be obtained by a marine energy system within a prediction horizon. Regulation of WECs to avoid malfunction or breakage, depending on the significant wave height and/or energy flux expected, as well as the possibility of reconfiguring them in order to maximise the wave energy extraction, is studied in [27,28]. The prediction of the energy that could be obtained from a certain maritime location is considered in [24,25] in order to know whether it is convenient to install WECs as power supply in marine structures, such as offshore oil and gas platforms or seawater desalination plants. In [52], significant wave height forecasting is applied for decision-making in exploitation and environmental protection for the construction of marine energy storage plants, future strategies on renewable energy and coastal planning. Other examples of application are: design of offshore structures and ports [53], decision-making and risk assessment about operational works in the sea [54], security systems for structures or naval security [55].

This paper is organised as follows: Section 2 describes the sources of information used by SPAMDA for creating datasets. Section 3 describes in detail the features of the software tool. Section 4 shows a case study describing the use of SPAMDA in a practical approach. Section 5 provides the final conclusions and future work.

2. Meteorological data sources

The data provided by the above-mentioned sources of information of SPAMDA is described below:

- NDBC is a part of the *National Weather Service* (NWS). NDBC designs, develops, operates, and maintains a network of data collecting buoys (stations). The mission of the network is to collect real-time marine meteorological and oceanographic observations, such as H_s, dominant wave period, or wind speed and direction, among others.

  The buoys maintained by NDBC are deployed in the coastal and offshore waters around oceans and seas, and they are equipped with assorted sensors which allow them to perform different measurements. The information collected by the buoys is available on the NDBC website [56], and it is divided into different groups. One of them corresponds to standard meteorological information of the historical data collected by each buoy, which can be downloaded as annual text files and whose format was adopted by NDBC since January 2007 [57]. These files contain hourly measurements per day from 00:50 to 23:50 UTC (Universal Time Coordinated) and from 23:50 31th December of the previous desired year to 22:50 31th December of the desired year. In Table 1, a comprehensive measurement description and the corresponding units are provided as a summary for the reader. A fragment of one of these files, which contains the measurements collected during year 2017 by the buoy identified as *Station 46001* in NDBC, is shown in Fig. 1. Each column corresponds to a meteorological variable or attribute, and each row or instance corresponds to the values of the measurements collected by the buoy for each attribute at a specific date and time.

```
#YY  MM DD hh mm WDIR WSPD GST  WVHT   DPD   APD MWD   PRES  ATMP  WTMP  DEWP  VIS  TIDE
#yr  mo dy hr mn degT m/s  m/s     m   sec   sec degT   hPa  degC  degC  degC   mi    ft
2016 12 31 23 50 279  6.4  7.3  2.41 12.90  6.50 999 1041.3   5.6   6.4 999.0 99.0 99.00
2017 01 01 00 50 291  6.3  7.3  2.13  7.14  6.08 999 1041.1   5.5   6.4 999.0 99.0 99.00
2017 01 01 01 50 293  5.3  6.6  2.39  7.69  6.64 999 1041.2   5.5   6.4 999.0 99.0 99.00
  .   .  .  .  .   .    .    .     .     .     .   .      .     .     .     .    .     .
  .   .  .  .  .   .    .    .     .     .     .   .      .     .     .     .    .     .
  .   .  .  .  .   .    .    .     .     .     .   .      .     .     .     .    .     .
  .   .  .  .  .   .    .    .     .     .     .   .      .     .     .     .    .     .
2017 12 31 20 50 999  3.0  4.4  4.98 12.90  8.57 201 1000.6   4.8   4.9 999.0 99.0 99.00
2017 12 31 21 50 999  3.8  6.3  4.64 10.00  8.55 150 1000.1   4.8   4.9 999.0 99.0 99.00
2017 12 31 22 50 999  3.4  5.2  4.40 12.90  8.40 200  998.9   4.7   4.9 999.0 99.0 99.00
```

Figure 1. A fragment of an annual text file of the *Station 46001*.

  Note that the data collected by the network of buoys may be incomplete due to diverse circumstances such as the weather conditions in which the buoys have to operate, failures or malfunctioning elements of the buoys, among others. Accordingly, it may be the situation that some of the measurements are completely missing (missing date or instance) or partially missing (some measurements not recorded), by a buoy or by a set of buoys, once in a while or over a period of time. It may be also possible that the measurements have been recorded at a time different from the expected one. These aspects have to be taken into account when creating the datasets. This casuistry is explained in detail in Appendix A.
- NNRP provides three-dimensional global reanalysis of numerous meteorological variables (e.g. air temperature, components South-North and West-East of wind speed, relative humidity,

Table 1. Measurements descriptions and units of each meteorological variable or attribute collected by the buoys.

Attribute	Units	Description
WDIR	degT	Wind direction (the direction the wind is coming from in degrees clockwise from true North) during the same period used for WSPD.
WSPD	m/s	Wind speed (m/s) averaged over an eight-minute period for buoys and a two-minute period for land stations. Reported Hourly.
GST	m/s	Peak 5 or 8 second gust speed (m/s) measured during the eight-minute or two-minute period.
WVHT	m	Significant wave height (meters) is calculated as the average of the highest one-third of all of the wave heights during the 20-minute sampling period.
DPD	sec	Dominant wave period (seconds) is the period with the maximum wave energy.
APD	sec	Average wave period (seconds) of all waves during the 20-minute period.
MWD	degT	The direction from which the waves at the dominant period (DPD) are coming. The units are degrees from true North, increasing clockwise, with North as 0 (zero) degrees and East as 90 degrees.
PRES	hPa	Sea level pressure (hPa). For C-MAN sites and Great Lakes buoys, the recorded pressure is reduced to sea level using the method described in NWS Technical Procedures Bulletin 291 (11/14/80).
ATMP	degC	Air temperature (Celsius degrees).
WTMP	degC	Sea surface temperature (Celsius degrees). For buoys the depth is referenced to the hull's waterline. For fixed platforms it varies with tide, but is referenced to, or near Mean Lower Low Water (MLLW).
DEWP	degC	Dewpoint temperature taken at the same height as the air temperature measurement.
VIS	nmi	Station visibility (nautical miles). Note that buoy stations are limited to reports from 0 to 1.6 nmi.
TIDE	ft	The water level in feet above or below MLLW.

pressure, etc.), which is available monthly, daily and every 6 hours at 00 Z (Zulu time), 06 Z, 12 Z and 18 Z from 1948 on a global 2.5° x 2.5° grid. Weather observations are from different sources, such as ships, satellites and radar, among others. Reanalysis data is created assimilating such observations using the same climate model throughout the entire reanalysis period in order to reduce the effects of modelling changes on climate statistics. Such information has become a substantial support of the needs of the research community, even more in locations where instrumental (real time) data is not available.

The reanalysis data is available in the NNRP website [58], which it is accessible through different sections. Such data can be fully (a global 2.5° x 2.5° grid) or partially (only the desired reanalysis nodes or sub-grid) downloaded as *Network Common Data Form* (NetCDF) files [59], a special binary format for representing scientific data, which provides a description of the file contents and also includes the spatial and temporal properties of the data. Each reanalysis file contains the values of a meteorological variable estimated by a mathematical model for each reanalysis node. For a better understanding, in Fig. 2 an approximate representation of a sub-grid containing six reanalysis nodes around the geographical location of a buoy (obtained from NDBC) is shown.

Therefore, with both sources of information, which complement each other, and carrying out a matching process, SPAMDA will create datasets for prediction tasks. In this way, the dataset input

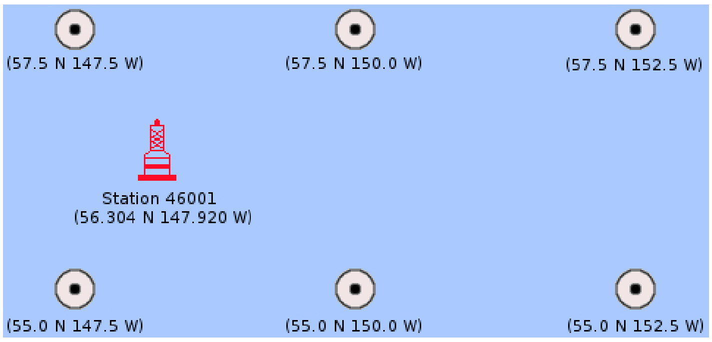


Figure 2. Example of a six sub-grid reanalysis nodes around the *Station 46001*.

variables will be one or more reanalysis variables from NNRP and one or more measurements from NDBC. The dataset output variable will always be one measurement from NDBC.

3. SPAMDA

SPAMDA combines meteorological information from NDBC and NNRP to obtain new datasets for oceanic and atmospheric studies. In order to do so, SPAMDA manages three different types of datasets, which will be described in detail in the following sections, but are briefly introduced bellow for giving the reader a better general understanding:

- *Intermediate datasets*: They contain the meteorological observations from NDBC.
- *Pre-processed datasets*: They are obtained as a result of pre-processing tasks performed on the intermediate datasets.
- *Final datasets*: Created by merging an intermediate or pre-processed dataset (which contain the information from NDBC) with the reanalysis data from NNRP. This procedure is referenced in SPAMDA as matching process and will be carried out according to the study to be performed (classification or regression).

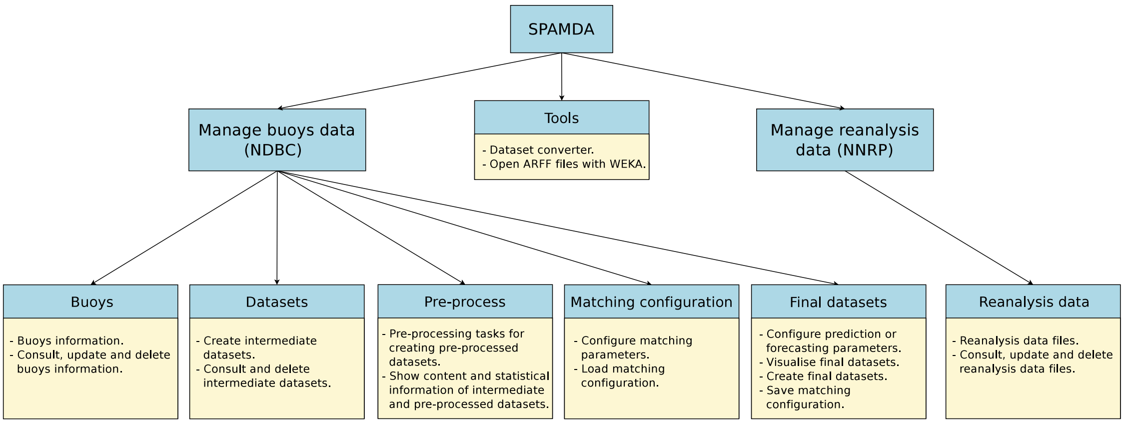


Figure 3. Brief outline of the functionality provided by SPAMDA.

SPAMDA consists of three main functional modules, whose main features, represented in Fig. 3, are the following:

- *Manage buoys data*: The aim of this module is to provide features for the management and analysis of the information related to the buoys from NDBC. This includes:
  1. Entering and updating the information of each buoy.
  2. Creation of intermediate datasets with the collected measurements.
  3. Pre-processing tasks for obtaining the pre-processed datasets.
  4. Matching process to merge the information from NDBC and NNRP.

    5. Creation of the final datasets accordingly to the ML technique to use (classification or regression).

- *Manage reanalysis data*: This module is used for the management of the reanalysis data provided by the NNRP. In this way, researchers can keep the reanalysis data files updated for their studies. Such files will be used, depending on researchers needs, in the matching process when obtaining the final datasets.
- *Tools*: This module includes features for converting intermediate or pre-processed datasets to ARFF or CSV format and for opening ARFF files with WEKA software.

In the following subsections each integrated functional module is described in detail.

3.1. Buoys

When a new buoy is included in SPAMDA the following information, which can be obtained from NDBC, is requested:

- *Station ID*: An alphanumeric identifier that allows easy identification of the buoy.
- *Description*: A short description of the buoy.
- *Latitude*: North or South geographical localisation (degrees) of the buoy.
- *Longitude*: West or East geographical localisation (degrees) of the buoy.
- *Measurements files*: The above-mentioned annual text files of the standard meteorological information collected by the buoy and downloaded from the NDBC website. This will be used for the creation of the intermediate datasets. One file per year is expected.

For clarification, an example is presented in Fig. 4, where the buoy ID1 has three annual text files and the buoy ID2 has two annual text files.

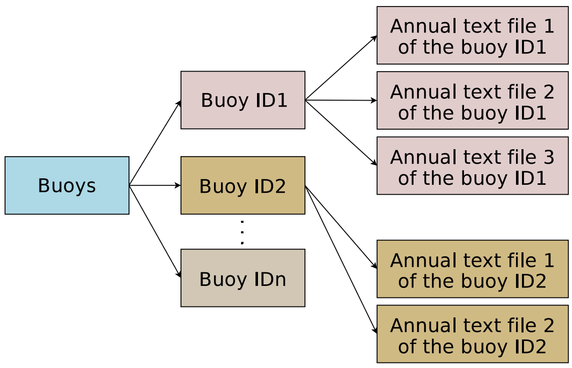


Figure 4. Example of entering two buoys with its annual text files.

3.2. Datasets

Once a buoy has been included as described in Section 3.1, it is possible to create datasets with one or more annual text files, which are referenced in SPAMDA as intermediate datasets. In this module, researchers can manage intermediate datasets of each buoy, which are the baseline for their studies, by creating new ones or deleting the unnecessary ones.

When an intermediate dataset is created, it is associated with its corresponding buoy. Besides, a summary of its content is also created, providing relevant information such as the number of instances, the dates of the first and last measurements, the annual text files included, and the missing and duplicated dates.

An example where three intermediate datasets have been created is presented in Fig. 5. The two intermediate datasets of the buoy ID1 contain meteorological data of different years, and the intermediate dataset of the buoy ID2 contains meteorological data of two years. For each buoy, as many intermediate datasets as needed can be created.

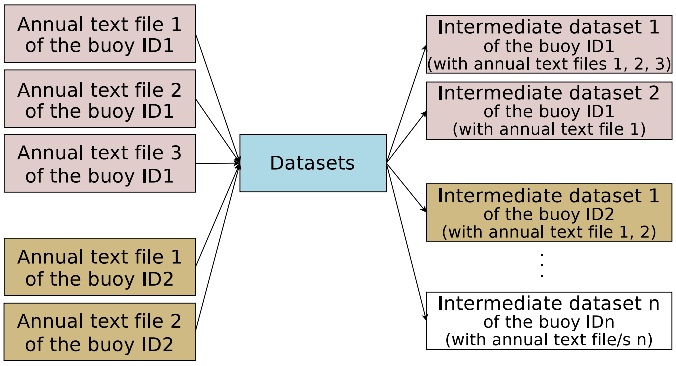


Figure 5. Example of the creation of the intermediate datasets.

3.3. Pre-process

Data pre-processing prepares the raw data (intermediate datasets) to be able to be treated correctly by ML algorithms. In this way, the quality of data can be improved prior to the learning phase, by applying pre-processing tasks (filters). The result will be referenced as pre-processed datasets.

SPAMDA provides several filters grouped in three categories, *Attribute*, *Instance* and *Recover missing data*, including the configuration of their parameters and a short description of them:

- *Attribute*: All theses filters can be applied to the attributes (variables of the buoy from NDBC) of the intermediate dataset.
  - *Normalize*: This filter normalises all numeric values of each attribute. The resulting values are by default in the interval [0,1].
  - *Remove*: It removes an attribute or a range of them.
  - *RemoveByName*: It removes attributes based on a regular expression matched against their names.
  - *ReplaceMissingValues*: For each attribute, all the missing values will be replaced by the average value of the attribute.
  - *ReplaceMissingWithUserConstant*: This filter replaces all the missing values of the attributes with an user-supplied constant value.
- *Instance*: All theses filters can be applied to the instances (hourly measurements of the buoy from NDBC) of the intermediate dataset.
  - *RemoveDuplicates*: With this filter, all duplicated instances are removed.
  - *RemoveWithValues*: This filter removes all the instances that match the attribute and the value supplied by the user.
  - *SubsetByExpression*: It removes all the instances which do not match a user-specified expression.
- *Recover missing data*: All these filters can be applied to the instances of the intermediate dataset.
  - *Replace missing values with next nearest hour*: The missing values of each attribute are replaced with the next nearest non missing value.
  - *Replace missing values with previous nearest hour*: This filter replaces the missing values of each attribute with the previous nearest non missing value.
  - *Replace missing values with next n hours mean*: The missing values of each attribute are replaced with the next n nearest non missing values mean, where n can be configured by the user.
  - *Replace missing values with previous n hours mean*: This filter replaces the missing values of each attribute in the intermediate dataset with the previous n nearest non missing values mean.
  - *Replace missing values with symmetric n hours mean*: The missing values of each attribute in the intermediate dataset are replaced with the n previous and n next non missing values mean.

SPAMDA allows researchers to undo the last filter applied or to restore the initial content of the intermediate dataset. Besides, the content and relevant statistical information (number of instances with missing values, minimum and maximum values, mean and standard deviation) of the intermediate and the pre-processed datasets can be visualised in this module.

Fig. 6 shows an example where the intermediate datasets 1 and 2 of the buoy ID1 have been pre-processed, obtaining as a result the pre-processed dataset 1 of each one. The intermediate dataset 1 of the buoy ID2 has been also pre-processed. *Pre-processed dataset n* represents that researchers can create as many pre-processed datasets as they consider opportune.

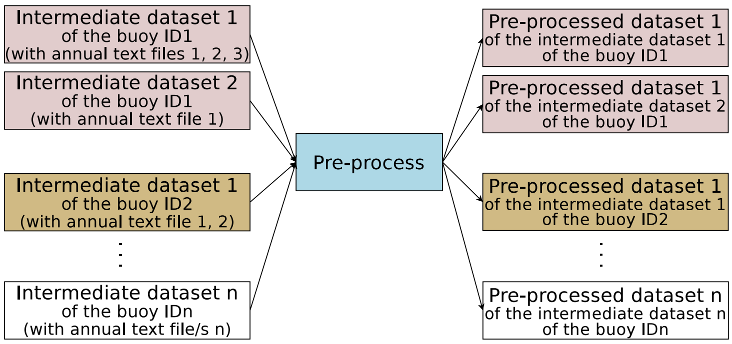


Figure 6. Example of the creation of pre-processed datasets.

Nevertheless, further pre-processing tasks can be performed after obtaining the final datasets by means of the Explorer environment of WEKA or other tools.

3.4. Matching configuration

The automatic integration of the data provided by the two sources of information described in Section 2, to merge and format such data, is denominated as the matching process in this document. Such process is one of the most powerful and remarkable features of this software tool due to its great relevance and extensive casuistry. In this sense, SPAMDA has been developed to provide great flexibility to researchers.

The matching procedure is performed using an intermediate or pre-processed dataset, which includes the measurements collected by a buoy from NDBC, and the needed reanalysis data files from NNRP. Note that SPAMDA is able to manage the NetCDF binary format for handling the information stored in the reanalysis files.

Such process merges the information of both sources that match on time, but, given that the measurements of the buoys are hourly collected from 00:50 to 23:50 UTC, and the reanalysis data is available every 6 hours at 00 Z, 06 Z, 12 Z and 18 Z, the matching can only be carried every 6 hours (discarding the rest of measurements from the buoy data). Besides, and since there is still a difference of 10 minutes, the matching with the reanalysis data will be performed with the nearest buoy measurement (before or after) within a maximum of 60 minutes of difference. Finally, the matched instances of both sources will form the final datasets.

Fig. 7 presents an example of matching with the measurements collected during 2017 by *Station 46001* (NDBC) and the reanalysis data (NNRP) of the variable *pressure* for reanalysis nodes 57.5 N $\times$ 147.5 W and 55.0 N $\times$ 147.5 W in the same year. In this way, only the instances from both sources that are linked with arrows (highlighted in green colour) will be used in the creation of the final datasets. Although the reanalysis dates have been presented in a human readable format, note that reanalysis dates are stored in hours from 01-01-1800, and they have to be transformed for comparison taking into account the time zone. Such transformation is automatically done by SPAMDA when matching the instances.

The reader can check in Appendix A an example with a more complex case of the procedure.

#YY	MM	DD	hh	mm	WDIR	WSPD	GST	WVHT	DPD	APD	MWD	PRES	ATMP	WTMP	DEWP	VIS	TIDE
#yr	mo	dy	hr	mn	degT	m/s	m/s	m	sec	sec	degT	hPa	degC	degC	degC	mi	ft
2016	12	31	23	50	279	6.4	7.3	2.41	12.90	6.50	999	1041.3	5.6	6.4	999.0	99.0	99.00
2017	01	01	00	50	291	6.3	7.3	2.13	7.14	6.08	999	1041.1	5.5	6.4	999.0	99.0	99.00
2017	01	01	01	50	293	5.3	6.6	2.39	7.69	6.64	999	1041.2	5.5	6.4	999.0	99.0	99.00
2017	01	01	02	50	297	3.9	4.9	2.57	14.81	7.39	999	1041.1	5.5	6.4	999.0	99.0	99.00
2017	01	01	03	50	298	5.7	6.7	1.83	7.69	5.80	999	1041.4	5.6	6.4	999.0	99.0	99.00
2017	01	01	04	50	281	5.0	6.1	2.29	13.79	7.14	999	1041.5	5.6	6.4	999.0	99.0	99.00
2017	01	01	05	50	293	6.1	7.6	2.17	12.90	6.88	999	1041.7	5.6	6.4	999.0	99.0	99.00
2017	01	01	06	50	314	4.2	5.4	2.24	12.12	7.09	999	1042.1	5.5	6.4	999.0	99.0	99.00
2017	01	01	07	50	297	4.4	5.2	1.97	13.79	6.64	999	1041.9	5.5	6.4	999.0	99.0	99.00
2017	01	01	08	50	287	4.6	5.6	2.06	12.90	7.26	999	1041.8	5.5	6.4	999.0	99.0	99.00
2017	01	01	09	50	304	3.2	4.1	1.82	12.12	6.65	999	1041.5	5.6	6.4	999.0	99.0	99.00
2017	01	01	10	50	274	2.9	3.9	1.92	12.90	6.88	999	1041.4	5.5	6.4	999.0	99.0	99.00
2017	01	01	11	50	287	1.3	1.7	1.60	12.90	6.66	999	1041.1	5.4	6.4	999.0	99.0	99.00
2017	01	01	12	50	260	2.0	3.0	1.67	12.90	6.95	999	1040.7	5.4	6.4	999.0	99.0	99.00
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2017	12	31	11	50	999	14.5	17.7	5.70	9.09	7.67	134	999.3	5.4	4.9	999.0	99.0	99.00
2017	12	31	12	50	999	11.8	14.7	6.21	10.00	8.52	125	1001.0	5.5	4.9	999.0	99.0	99.00
2017	12	31	13	50	999	10.1	12.6	5.52	10.81	8.49	165	1002.0	5.3	4.9	999.0	99.0	99.00
2017	12	31	14	50	999	8.6	10.5	5.47	10.00	8.61	150	1002.5	5.3	4.9	999.0	99.0	99.00
2017	12	31	15	50	999	7.1	9.1	5.83	12.90	9.27	216	1002.7	5.2	4.9	999.0	99.0	99.00
2017	12	31	16	50	999	6.4	8.4	5.82	11.43	9.35	177	1002.7	5.1	4.9	999.0	99.0	99.00
2017	12	31	17	50	999	5.3	7.4	5.52	11.43	9.04	182	1002.4	5.0	4.9	999.0	99.0	99.00
2017	12	31	18	50	999	2.8	4.6	5.00	11.43	8.82	179	1002.0	4.8	4.9	999.0	99.0	99.00
2017	12	31	19	50	999	3.4	5.0	4.87	12.12	8.63	200	1001.6	5.0	4.9	999.0	99.0	99.00
2017	12	31	20	50	999	3.0	4.4	4.98	12.90	8.57	201	1000.6	4.8	4.9	999.0	99.0	99.00
2017	12	31	21	50	999	3.8	6.3	4.64	10.00	8.55	150	1000.1	4.8	4.9	999.0	99.0	99.00
2017	12	31	22	50	999	3.4	5.2	4.40	12.90	8.40	200	998.9	4.7	4.9	999.0	99.0	99.00

	57.5 N 147.5 W	55.0 N 147.5 W
Date	Pres	Pres
2017 01 01 00 00	106400	103300
2017 01 01 06 00	106400	103280
2017 01 01 12 00	106380	103220
2017 01 01 18 00	106270	103030
2017 01 02 00 00	106150	103030
2017 01 02 06 00	106030	102920
2017 01 02 12 00	105960	102820
2017 01 02 18 00	105980	102850
2017 01 03 00 00	106010	102840
2017 01 03 06 00	106080	102880
2017 01 03 12 00	106100	102960
2017 01 03 18 00	106090	102920
2017 01 04 00 00	106070	102920
2017 01 04 06 00	106040	102810
. . . . .	.	.
. . . . .	.	.
. . . . .	.	.
2017 12 31 12 00	102520	99040
2017 12 31 18 00	102660	99400

Figure 7. An example of matching the data from NDBC (left) and NNRP (right).

SPAMDA allows researchers to perform a customisable matching process, for obtaining as many different versions of the same meteorological data as needed. Prediction tasks are based on the estimation of the output attribute using the information provided by the input attributes. Depending on the task, the datasets must be prepared and configured differently:

- *Classification*: The final datasets will be ready to use as input for ML classifiers, requiring a nominal output attribute, whose specific preparation is detailed in Section 3.5.
- *Regression*: The final datasets will be ready to use as input for regression methods, requiring a real output attribute, whose preparation is also explained in Section 3.5.
- *Direct matching*: In this case the inputs attributes have a direct correspondence with the output attribute, and it is not necessary to perform any additional preparation. Both input and target attributes are synchronised in time, in such a way that the final dataset is not intended for prediction purposes. For example, the final datasets may be used in lost data recovering tasks, in correlation studies, in descriptive analyses, etc.

The following parameters can be specified for the matching process:

- *Flux of energy* [46]: When the F_e is selected, it will be used as output. This attribute is not collected by the buoys, but it can be calculated from two wave parameters: H_s and T_e, which are collected as WVHT and APD attributes, respectively, and were described in Table 1. In this way, SPAMDA obtains the F_e of each instance using the following equation:

$$F_e = 0.49 \cdot H_s^2 \cdot T_e, \tag{1}$$

  where F_e is measured in kilowatts per meter, H_s is measured in meters and T_e is measured in seconds. Note also that F_e is defined in Eq. 1 as an average energy flux (H_s is a kind of average wave height), though for simplicity it will be referred just as flux of energy.
- *Attribute to predict*: Instead of using F_e, researchers can select any of the attributes collected by the buoys as output (e.g. significant wave height, WVHT, wind direction, WDIR, sea level pressure, PRES, etc.). Therefore, they can conduct different studies by selecting one attribute or other.
- *Reanalysis data files*: In order to have a possible better description of the problem under study, more than one reanalysis variable can be considered as input. Remember that these files have to be previously downloaded from the NNRP website [58], which should set the range of dates (temporal properties) and the desired sub-grid (spatial properties, see Fig. 2) for each variable of reanalysis.

In that sense, the reanalysis data files must have the same spatial and temporal properties but related to different variables. SPAMDA simplifies this task by showing the reanalysis data files that are compatibles each other, and checking that the selection made by the researches meets that condition.

- *Buoys attributes*: In addition to the reanalysis variables, the final datasets will also include the selected attributes as inputs (of the intermediate or pre-processed dataset used), providing a possible better characterisation of the problem under study, although it will depend on how correlated the attributes are.
- *Include missing dates*: As above-mentioned, the information collected by a buoy may be incomplete due to measurements not recorded by it. As a consequence, the matching of instances between both sources of information may not be possible (missing dates). In that situation, researchers can consider two options: 1) discard the instances affected or 2) include them. In the latter case, the final datasets will contain the affected instances, but the measurements of the buoy will be stored as missing values in WEKA format, denoted as «?».
- *Nearest reanalysis nodes to consider*: As already shown in Fig. 2 (which represents six reanalysis nodes), the reanalysis data files may contain information of several reanalysis nodes. In this way, researchers can:

  – Consider all the reanalysis nodes contained in each file: in this case, the information provided by each reanalysis node contained in each selected reanalysis data file will be used.
  – Consider only some of the reanalysis nodes contained in each file: in this case, only the information of the N closest reanalysis nodes to the buoy will be used (N given by the user). To do that, SPAMDA uses the *Haversine* equation [60] to calculate the distance from each reanalysis node to the location of the buoy and obtain the closest ones. Haversine equation is also known as the great circle distance and performs calculation from main point to destination point with a trigonometric function using latitude and longitude:

$$\begin{aligned} d(p_0, p_j) \;=\; & \arccos(\sin(lat_0) \cdot \sin(lat_j) \\ & \cdot \cos(lon_0 - lon_j) + \cos(lat_0) \\ & \cdot \cos(lat_j)), \end{aligned} \tag{2}$$

  where p_0 is the buoy geographical location, p_j stands for the location of each reanalysis node, and lat and lon are the latitude and longitude of the points, respectively.

- *Number of final datasets*: Depending on the number of nearest reanalysis nodes to consider, the number of final datasets to create and the content of them can be configured according to the following options:

  – *One (using weighted mean of the N nearest reanalysis nodes)*: Only one final dataset will be created, which will contain the attributes (the selected one as output and the selected ones as inputs) of the intermediate or pre-processed dataset used, along with a weighted mean of each variable of the reanalysis data used (one per selected reanalysis data file). This weighted mean is obtained by SPAMDA and uses Eq. 2 to obtain the distance from each reanalysis node to the location of the buoy. Once the distances have been calculated they are inverted and normalised as follows:

$$w_i = \frac{d(p_0, p_i)}{\sum_{j=1}^{N} d(p_0, p_j)}, \quad i = 1, \ldots, N. \tag{3}$$

  With these weights, a weighted mean of each variable of reanalysis is obtained for each of the N nodes. Therefore, the closest reanalysis nodes to the localisation of the buoy will provide more information.

Considering as example the two nearest reanalysis nodes represented in Fig. 2 and the reanalysis variables air temperature and pressure, the weighted mean of each reanalysis variable will be calculated using the reanalysis nodes 57.5 N × 147.5 W and 55.0 N × 147.5 W.

– *'N' (one per each reanalysis node)*: As many final datasets as the number of nearest *N* reanalysis nodes configured by researcher will be created. Therefore, each final dataset will contain the value of each reanalysis variable used of the nearest corresponding reanalysis node, along with the selected attributes of the intermediate or pre-processed dataset used. In this way, researchers can perform comparison studies depending on the reanalysis node considered, to achieve better performance for the problem under study.

In this case, and considering as example the four closest reanalysis nodes (see Fig. 2) and the reanalysis variables air temperature and pressure, four final datasets will be created, containing each one the information of both reanalysis variables of the corresponding reanalysis node: 57.5 N × 147.5 W, 55.0 N × 147.5 W, 57.5 N × 150.0 W and 55.0 N × 150.0 W, along with the selected attributes of the intermediate or pre-processed dataset used.

Once the matching parameters have been described, for a better understanding of them, Fig. 8 presents an example of the data integration considering the data shown in Fig. 7 and using the following configuration[1]:

- Attribute to predict: variable WVHT (Fig. 8a) / flux of energy (Fig. 8b).
- Variable Pres as reanalysis input attribute.
- Variable WSDP as buoy input attribute.
- Not including missing dates.
- Considering the closest reanalysis node.
- Task to be used: *Direct matching*.

Date	Pres	WSPD	WVHT
2017 01 01 00 00	106400	6.4	2.41
2017 01 01 06 00	106400	6.1	2.17
2017 01 01 12 00	106380	1.3	1.60
2017 01 01 18 00	106270	2.4	1.33
2017 01 02 00 00	106150	1.2	0.94
2017 01 02 06 00	106030	3.0	1.42
2017 01 02 12 00	105960	3.1	1.99
2017 01 02 18 00	105980	2.5	2.52
2017 01 03 00 00	106010	3.5	2.03
2017 01 03 06 00	106080	5.1	1.84
2017 01 03 12 00	106100	5.4	1.81
2017 01 03 18 00	106090	5.6	1.60
2017 01 04 00 00	106070	7.0	1.54
2017 01 04 06 00	106040	7.4	1.73
. . . . .	.	.	.
. . . . .	.	.	.
. . . . .	.	.	.
2017 12 31 12 00	102520	14.5	5.70
2017 12 31 18 00	102660	5.3	5.52

(a) attribute to predict: WVHT

Date	Pres	WSPD	FLUXOFENERGY
2017 01 01 00 00	106400	6.4	15.874644
2017 01 01 06 00	106400	6.1	8.354304
2017 01 01 12 00	106380	1.3	5.954648
2017 01 01 18 00	106270	2.4	2.76664
2017 01 02 00 00	106150	1.2	8.50699
2017 01 02 06 00	106030	3.0	17.561063
2017 01 02 12 00	105960	3.1	30.619089
2017 01 02 18 00	105980	2.5	16.699123
2017 01 03 00 00	106010	3.5	13.852182
2017 01 03 06 00	106080	5.1	11.927297
2017 01 03 12 00	106100	5.4	8.856064
2017 01 03 18 00	106090	5.6	6.77495
2017 01 04 00 00	106070	7.0	8.271178
2017 01 04 06 00	106040	7.4	7.96014
. . . . .	.	.	.
. . . . .	.	.	.
. . . . .	.	.	.
2017 12 31 12 00	102520	14.5	122.107167
2017 12 31 18 00	102660	5.3	134.971684

(b) attribute to predict: flux of energy

Figure 8. Example of data integration for *Direct matching*.

3.5. Final datasets

Once the matching process has been performed with the desired configuration, it is necessary to prepare the matched information for the desired prediction task (*Regression* or *Classification*), obtaining as a result the final datasets. Remember that *Direct matching*, as it was described in Section 3.4,

[1] Note that the date is shown just for a better understanding, but it will not be included in the final dataset.

performs a direct correspondence between the attributes used as inputs and the output one, and it is not necessary to carry out any preparation.

SPAMDA allows researchers to make such preparation by means of the following options:

- *Prediction horizon* (Classification and Regression): This option indicates the time gap for moving backward the attribute to predict (output attribute). In this way, the input attributes (variables of the buoy and reanalysis data) will be used to predict the output attribute in a specific future time (e.g. +6h, +12h, +18h, +1 day, etc.).

  The minimum interval for increasing and decreasing the prediction horizon is 6h (due to reanalysis data temporal resolution) [4], the same interval used when the matching process is carried out. Therefore, for each increment of the prediction horizon, an instance of the dataset is lost (as this future information is not available). As the minimum prediction horizon is 6h, at least one instance will be lost. The relation between the inputs and the attribute to predict will be defined as follows:

  $$o_{t+\Delta t} = \phi(\mathbf{b}_t, \mathbf{r}_t), \tag{4}$$

  where t represents the time instant to study and Δt the prediction horizon; o is the attribute to be predicted, $\mathbf{b}_t$ is the vector containing the selected NDBC variables and $\mathbf{r}_t$ is the vector containing the selected reanalysis variables. In this way and considering the matched information shown in Fig. 8a, WVHT is o, the vector $\mathbf{b}$ contains the variable WSPD and the vector $\mathbf{r}$ contains Pres.

  Optionally, the reanalysis variables can be synchronised with the attribute to predict. Given that these variables are estimated by a mathematical model, we can obtain very good future estimations, which can improve the performance of the results. In this case, the relation between the inputs and the attribute to predict would be:

  $$o_{t+\Delta t} = \phi(\mathbf{b}_t, \mathbf{r}_{t+\Delta t}). \tag{5}$$

  Note that the selected NDBC variables as input cannot be synchronised with the attribute to predict.

  For the sake of clarity, considering the matched information shown in Fig. 8a, an example of building a dataset for a *Regression* task is shown in Fig. 9a. As mentioned earlier, this prediction task requires a real output variable (in this case, WVHT, the last one). The options considered for the preparation of each final dataset are the following:

  – Do not synchronise the reanalysis data (see Eq. 4 for the relation between the inputs and the output).
  – A prediction horizon of 6h.

  Note that, due to prediction horizon is 6h, the values of WVHT attribute are moved backward one instance (up). As a consequence, the last instance (2017/12/31 18:00) is lost and is not included in the final dataset. Besides, and because the reanalysis data has not been synchronised, the values of the Pres and WSPD variables are at the same time instant (t in Eq. 4).

  Moreover, considering again the matched information shown in Fig. 8a, an example of the creation of the same dataset but applying synchronisation (see Eq. 5) is shown in Fig. 9b.

  Again, and due to the prediction horizon selected (6h), the values of the WVHT attribute are moved backward one instance (up) and the last instance (2017/12/31 18:00) is not included in the final dataset. But now, the values of the Pres variable are also moved backward one instance (due to the synchronisation). Therefore, in this case, Pres is at the same time instant as the attribute to predict ($t + \Delta t$ in Eq. 5).
- *Thresholds of the output attribute* (Classification): Since the values of the variables collected by the buoys are real numbers, it is necessary to discretise them (convert them from real to nominal

```
Date              Pres    WSPD   WVHT
2017 01 01 00 00  106400   6.4   2.17
2017 01 01 06 00  106400   6.1   1.60
2017 01 01 12 00  106380   1.3   1.33
2017 01 01 18 00  106270   2.4   0.94
2017 01 02 00 00  106150   1.2   1.42
2017 01 02 06 00  106030   3.0   1.99
2017 01 02 12 00  105960   3.1   2.52
2017 01 02 18 00  105980   2.5   2.03
2017 01 03 00 00  106010   3.5   1.84
2017 01 03 06 00  106080   5.1   1.81
2017 01 03 12 00  106100   5.4   1.60
2017 01 03 18 00  106090   5.6   1.54
2017 01 04 00 00  106070   7.0   1.73
2017 01 04 06 00  106040   7.4   1.64
  .    .  .  .  .    .      .      .
  .    .  .  .  .    .      .      .
  .    .  .  .  .    .      .      .
2017 12 31 12 00  102520  14.5   5.52
```

(a) without synchronisation

```
Date              Pres    WSPD   WVHT
2017 01 01 00 00  106400   6.4   2.17
2017 01 01 06 00  106380   6.1   1.60
2017 01 01 12 00  106270   1.3   1.33
2017 01 01 18 00  106150   2.4   0.94
2017 01 02 00 00  106030   1.2   1.42
2017 01 02 06 00  105960   3.0   1.99
2017 01 02 12 00  105980   3.1   2.52
2017 01 02 18 00  106010   2.5   2.03
2017 01 03 00 00  106080   3.5   1.84
2017 01 03 06 00  106100   5.1   1.81
2017 01 03 12 00  106090   5.4   1.60
2017 01 03 18 00  106070   5.6   1.54
2017 01 04 00 00  106040   7.0   1.73
2017 01 04 06 00  105950   7.4   1.64
  .    .  .  .  .    .      .      .
  .    .  .  .  .    .      .      .
  .    .  .  .  .    .      .      .
2017 12 31 12 00  102660  14.5   5.52
```

(b) with synchronisation

Figure 9. Example of the creation of a *Regression* dataset with a prediction horizon of 6h.

values) for the attribute selected as output (attribute to be predicted). SPAMDA allows researchers to perform this process by defining the necessary classes with their thresholds, which will be used to carry out such discretisation.

Considering again the matched information shown in Fig. 8a, an example of the creation of a *Classification* dataset is shown in Fig. 10. The options considered for the preparation of the final dataset are the following:

– Do not synchronise the reanalysis data.
– A prediction horizon of 6h.
– The thresholds shown in Table 2.

Table 2. Thresholds for the classification example represented in Fig. 10

Class	Description	Lower end [	Upper end)
Low	Low wave height	0.36	1.5
Average	Average wave height	1.5	2.5
Big	Big wave height	2.5	4.0
Huge	Huge wave height	4.0	9.9

```
Date              Pres    WSPD   Class_WVHT
2017 01 01 00 00  106400   6.4   Average
2017 01 01 06 00  106400   6.1   Average
2017 01 01 12 00  106380   1.3   Low
2017 01 01 18 00  106270   2.4   Low
2017 01 02 00 00  106150   1.2   Low
2017 01 02 06 00  106030   3.0   Average
2017 01 02 12 00  105960   3.1   Big
2017 01 02 18 00  105980   2.5   Average
2017 01 03 00 00  106010   3.5   Average
2017 01 03 06 00  106080   5.1   Average
2017 01 03 12 00  106100   5.4   Average
2017 01 03 18 00  106090   5.6   Average
2017 01 04 00 00  106070   7.0   Average
2017 01 04 06 00  106040   7.4   Average
  .    .  .  .  .    .      .      .
  .    .  .  .  .    .      .      .
  .    .  .  .  .    .      .      .
2017 12 31 12 00  102520  14.5   Huge
```

Figure 10. An example of the creation a *Classification* dataset, with a prediction horizon of 6h and without synchronisation.

Note that the attribute to be predicted has been renamed to *Class_WVHT* to show that it is now a nominal variable, because its values have been discretised according to the thresholds (usually defined by an expert). Besides, and due to the 6h prediction horizon, the last instance is lost (2017/12/31 18:00) and the values of the attribute *Class_WVHT* are moved backward one instance (up). As the reanalysis data have not been synchronised, the values of the Pres and WSPD variables are at the same time instant (t in Eq. 4).

The content of the final datasets, obtained as the result of the preparation of the matched data, can be visualised to check everything before saving them on disk. Such preparation can be performed as many times as required and considering the different options in each moment. Although the date will not be included in the final datasets, it can be shown to properly check the matching.

Finally, it is necessary to define the output configuration to create the final datasets:

- *Output path file*: Name of the final datasets and folder to save them on disk.
- *Final datasets format*:
  - *ARFF*: *Attribute-Relation File Format* [51], which is used by WEKA. SPAMDA allows researchers to directly open the final datasets in the Explorer environment of WEKA (in the same context of work), enabling them to choose the most appropriate ML method to tackle the problem under study.
  - *CSV*: *Comma-Separated Values*. This format is included in order to consider other different tasks of software tools.

A text file that summarises the configuration used in matching process and in the preparation of the matched data is also generated. It can be saved and loaded, enabling researchers to resume their studies at any other time.

3.6. Manage reanalysis data

As mentioned in Section 2, the reanalysis data files provided by NNRP contain the estimated values by a mathematical model of one meteorological variable.

In this module (see Fig. 3), SPAMDA includes features for entering new files and deleting the unnecessary ones. Besides, useful information about the content of each reanalysis file can be consulted such as name of the file and the reanalysis variable, number of instances and reanalysis nodes, initial and final time, latitude and longitude. All these fields summarise the temporal and spatial properties of the data. Thus, researcher can quickly and easily identify each reanalysis file entered in SPAMDA.

An example where two reanalysis data files have been entered in SPAMDA is shown in Fig. 11.

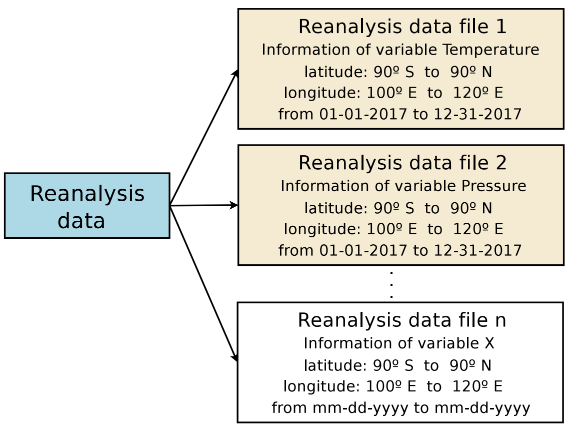


Figure 11. Example of entering two reanalysis data files.

3.7. Tools

SPAMDA also contains another module that provides two utilities: one of them is *Dataset converter* used for converting the desired intermediate or pre-processed datasets to ARFF or CSV formats; the other utility can be used for opening ARFF files with WEKA Explorer environment, which is useful for easily checking the results of different configurations of the pre-processing.

4. A case study applied to Gulf of Alaska

This section describes how SPAMDA works in a practical approach showing two examples to create fully processed datasets (final datasets) starting from the raw data. The objective of these final datasets is to be used with SC and ML algorithms for environmental modelling, in this case, to classify waves depending on their height and to predict energy flux in the Gulf of Alaska.

On the one hand, wave classification is tackled as a multi-class classification problem, given that any continuous variable can be discretised in different classes. Such waves modelling can be applied with different purposes, such as missing buoy data reconstruction, extreme significant wave heights detection or decision-making and risk assessment about operational works in the sea.

On the other hand, energy flux prediction is addressed as a regression problem. Energy flux prediction is related to marine energy and it is useful to characterise the wave energy production from WECs facilities, which could be injected into the electric network or supplied to existing marine platforms.

4.1. Gathering the information and introducing it in SPAMDA

The data collected to perform this case study is:

1. The measurements obtained from 2013 to 2017 by the buoy with ID 46001, placed in the Gulf of Alaska, which are provided by NDBC as annual text files. This data is publicly available at the NDBC website.
2. Complementary information collected from reanalysis data containing air temperature (air), pressure (pres) and two components of wind speed measurements: South-North (vwind) and West-East (uwind). This information will be collected from the four closest reanalysis nodes surrounding the geographical location of the buoy. This data is publicly available at the NNRP website and can be downloaded in NetCDF format. Concretely, the closest reanalysis nodes downloaded are 57.5 N × 147.5 W, 57.5 N × 150 W, 55 N × 147.5 W and 55 N × 150 W. However, as will be seen later, only the information from the nearest node will be used in the data integration process.

After gathering the information described above, researchers can open SPAMDA. In Fig. 12, the main view is shown. In order to input the reanalysis data which will be used in further steps for creating the final dataset, researchers has to select the option ***Manage reanalysis data***.

Then, the view of Fig. 13 is shown. Here, using the buttons located at the bottom, it is possible to add, delete or consult any data from the different reanalysis files. Once the information has been introduced in the application, this view can be closed and the user can go back to the main view to continue entering the information related to the buoy under study.

After that, the researcher has to select ***Manage buoys data*** to open the view shown in Fig. 14, where several tabs are available. In ***Buoys*** tab, the researcher can consult, modify, add or delete different data related to the buoy.

In order to enter such data, click on the ***New*** button, and then the view shown in Fig. 15 pops up.

Here the information about the buoy has to be included: the *Station ID*, its description, geographical localisation and the corresponding annual text files. In this case, the files containing the data from year 2013 to 2017 are inserted by clicking on the ***Add file*** button. Once the data has been

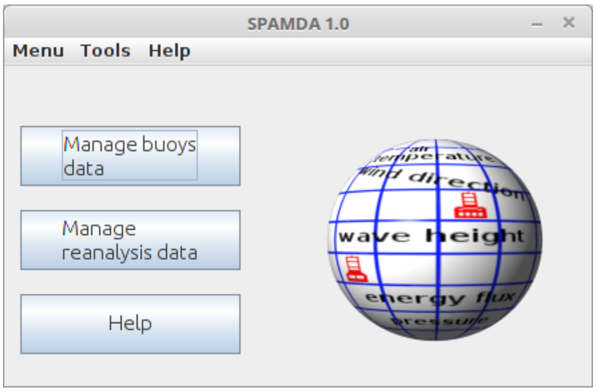


Figure 12. SPAMDA main view.

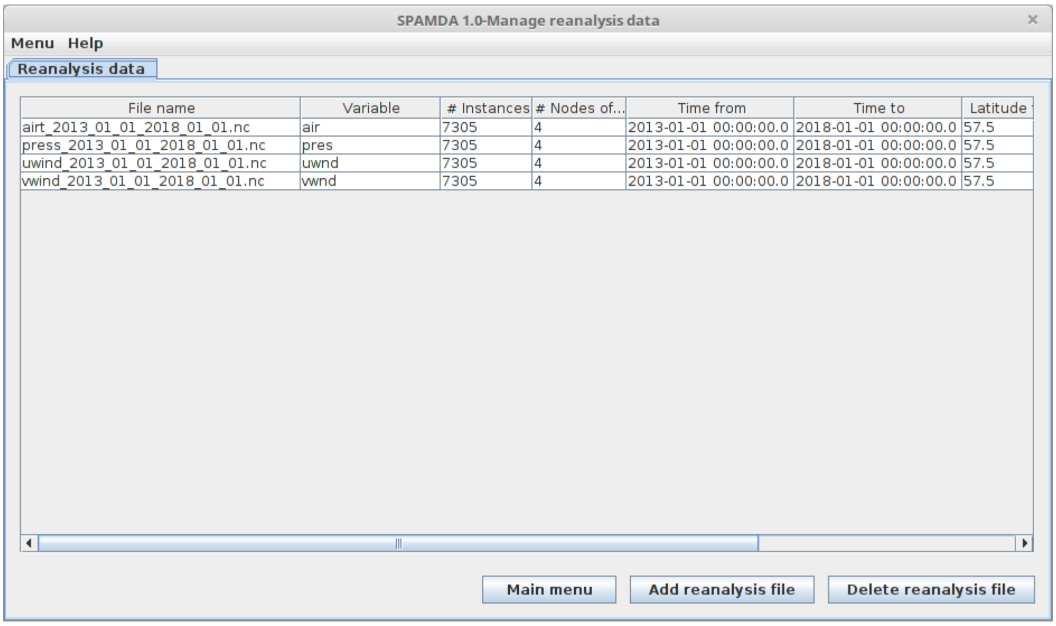


Figure 13. *Manage reanalysis data* view: downloaded files containing the four closest reanalysis nodes.

introduced, it is necessary to click on the ***Save*** button to insert the buoy in SPAMDA database. After that, the view can be closed.

To create the intermediate dataset, the researcher has to double-click on the buoy under study or click on the ***Datasets*** tab (see Fig. 14) to switch to the corresponding view (see Fig. 16). In this view, the researcher can delete or consult a summary of each intermediate or pre-processed dataset by selecting it from the corresponding list. It can also create new ones. To proceed with the creation of the intermediate dataset, the user clicks on the ***New*** button, and the view shown in Fig. 17 appears.

Here the researcher can select the annual text files to be included in the intermediate dataset, by clicking on the **->** and **<-** buttons. In this case, all the files introduced before, which correspond to the buoy under study, are selected. When the file selection is finished, ***Create*** button has to be clicked in order to introduce the description and the file name of the current intermediate dataset, and then, with the ***Save*** button, the creation process starts, showing the status of the process during it. After that, in order to prepare the intermediate dataset, the dataset is selected (see Fig. 16), and then the button ***Open*** is clicked to jump to the tab ***Pre-process*** (shown in Fig. 18).

In ***Pre-process*** tab, relevant statistical information about the selected dataset is shown, and also the content of the dataset can be consulted, providing the researcher the capacity to evaluate the pre-processing being performed. Here the researcher can apply (and configure) the necessary filters (explained in Section 3.3) to the selected dataset, and, in the bottom part, the main statistics of the

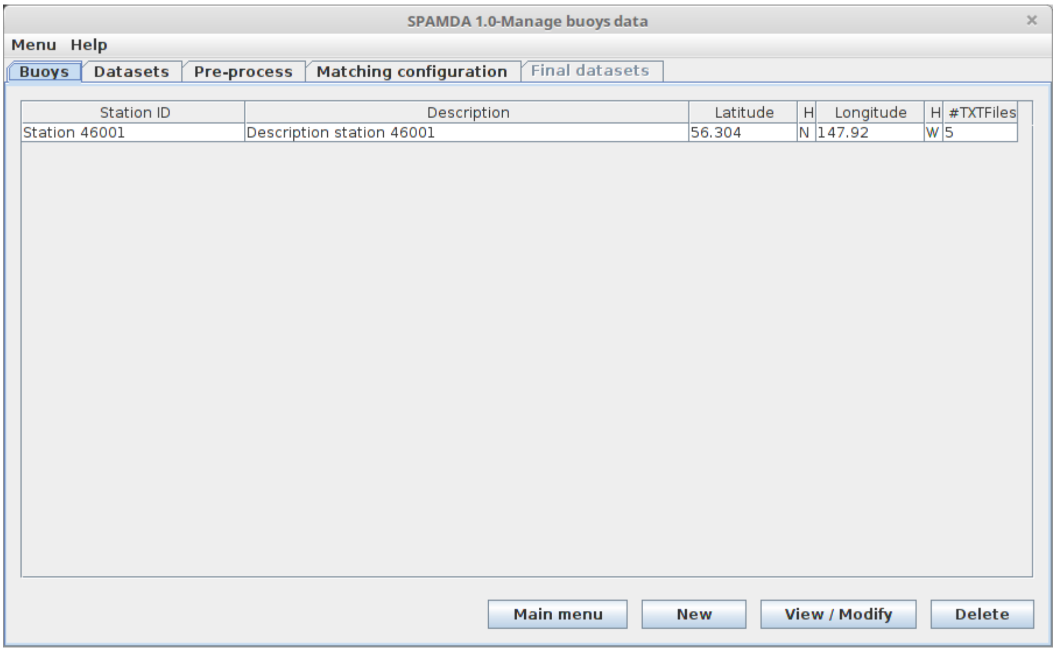


Figure 14. *Buoys* tab: buoy ID 46001.

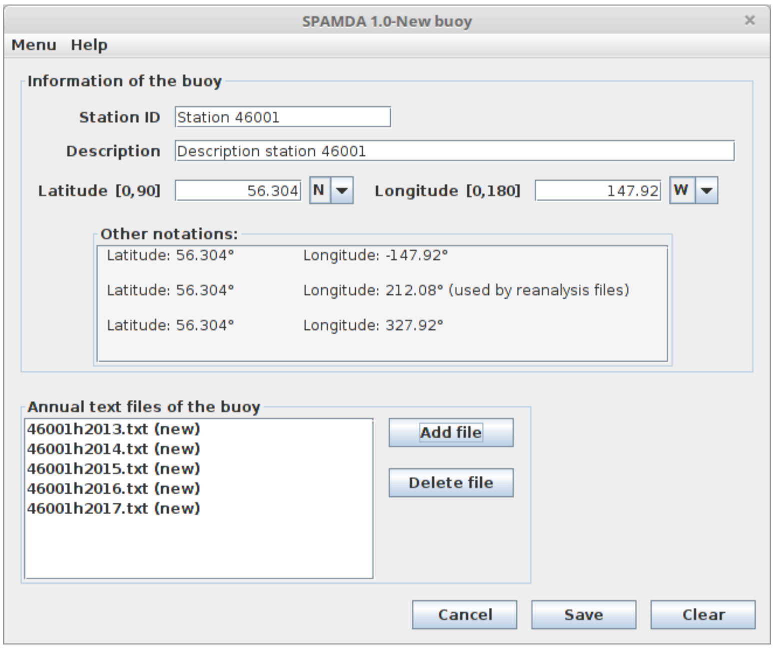


Figure 15. *New buoy* view: information of the buoy ID 46001.

dataset are displayed, which can be used to observe the changes produced when applying a filter. As mentioned earlier, this case study is focused on classifying waves considering their height, so any missing data from wave height (376 values) and the remaining attributes are recovered, using the filter *Replace missing values with symmetric* 3 *hours mean*. Furthermore, the attributes MWD, DEWP, VIS and TIDE are removed from the dataset by applying the filter *RemoveByName*, since the first two had more than 92% of missing data and the last two 100%. After finishing the pre-processing of the dataset, the researcher can click on the ***Save*** button, to introduce the description and file name for the current pre-processed dataset.

At this point, the researcher has registered the buoy in SPAMDA, then entered its raw data and selected the required data for the problem (intermediate dataset). Finally, the data has been

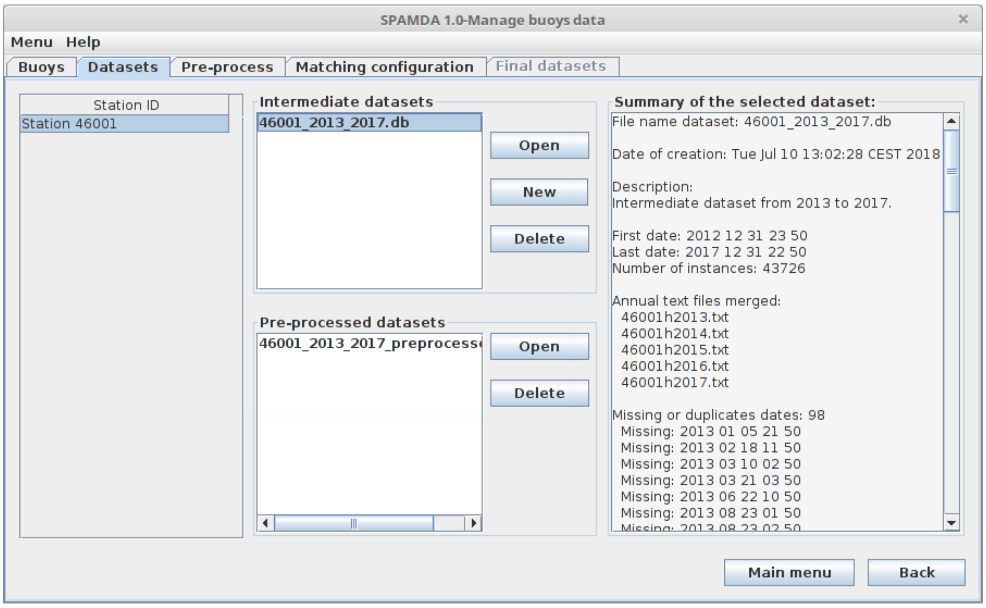


Figure 16. *Datasets* tab: intermediate datasets of the buoy ID 46001.

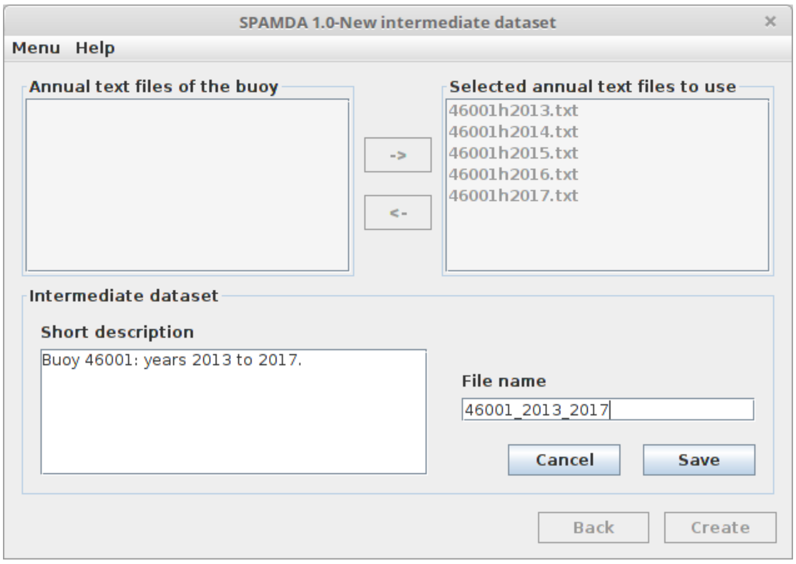


Figure 17. *New intermediate dataset* view: creating the intermediate dataset with 5 annual text files.

pre-processed in order to be ready for its future use in ML algorithms. In order to achieve a more accurate description of the problem under study, a matching process can be carried out to merge the processed data from NDBC with the reanalysis data (also entered previously) from NNRP.

The next step is to customise (or load) the parameters of the matching process according to the problem being studied and to select the prediction task (described in Section 3.4) that the final dataset will be used for, in this case, waves classification or energy flux prediction.

4.2. *Waves classification*

As mentioned above, the objective of the final dataset is to be used with SC and ML algorithms to classify waves depending on their significant height. Following sections describe the procedure of performing the data integration provided by SPAMDA, modelling wave height by using classification algorithms available in WEKA.

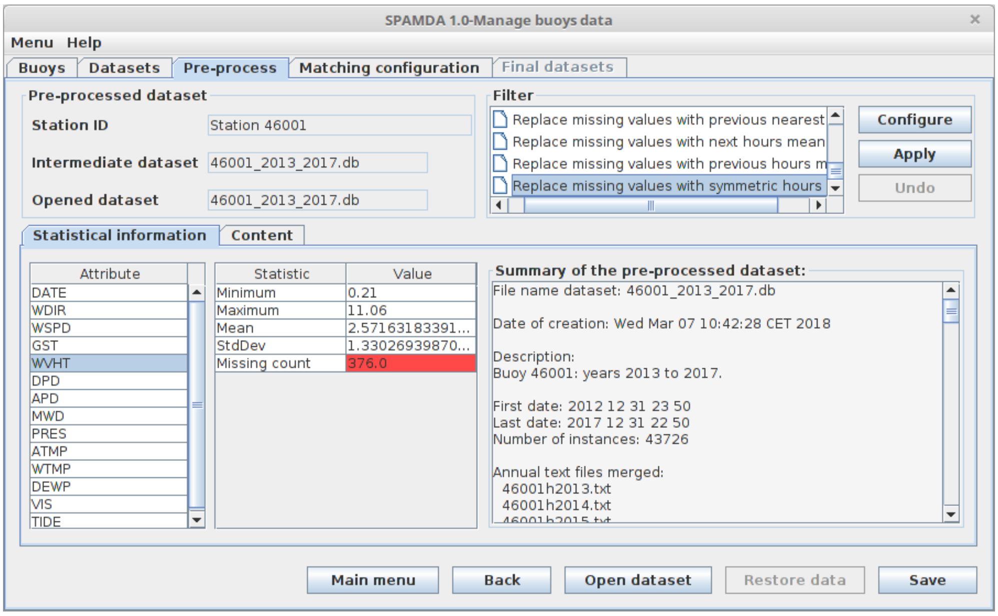


Figure 18. *Pre-process* tab: preprocessing the created intermediate dataset.

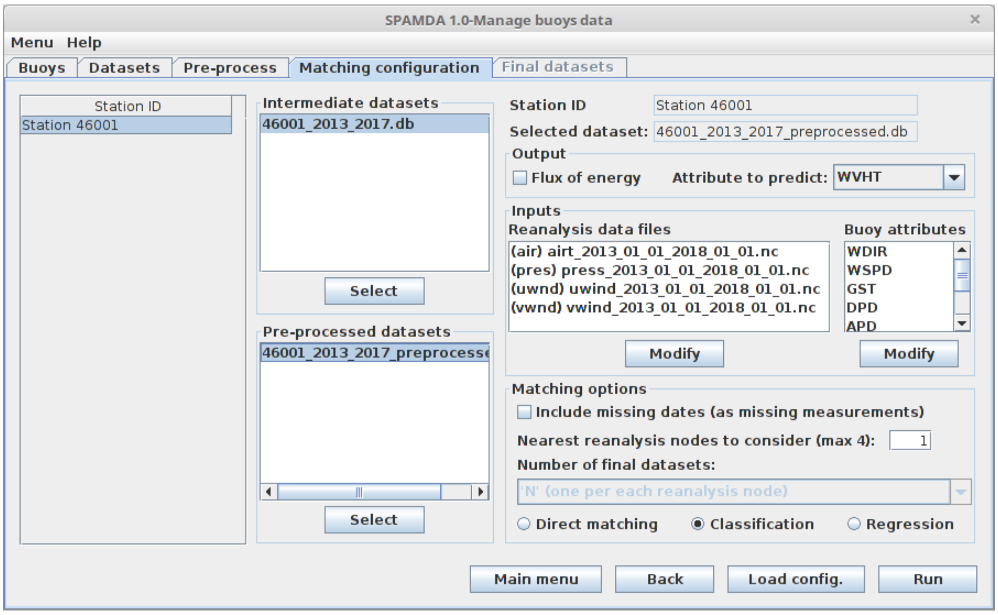


Figure 19. *Matching configuration* tab: parameters for the data integration of the intermediate dataset and the reanalysis files (waves classification).

4.2.1. Obtaining the final dataset

By clicking on the ***Matching configuration*** tab, the view shown in Fig. 19 will be opened. In this view, the researcher can configure the parameters of the data integration process. For this problem, the following parameters were selected:

- Attribute to predict: WVHT.
- Reanalysis data: Air, pressure, u-wind and v-wind.
- Buoy attributes to be used as inputs: WDIR, WSPD, GST, DPD, APD, PRES, ATMP and WTMP (see Table 1).
- Reanalysis nodes to consider: 1 (only the closest reanalysis node will be used).
- Number of final datasets: In this example that option is disabled, because only one reanalysis node is considered.
- Prediction task: Classification.

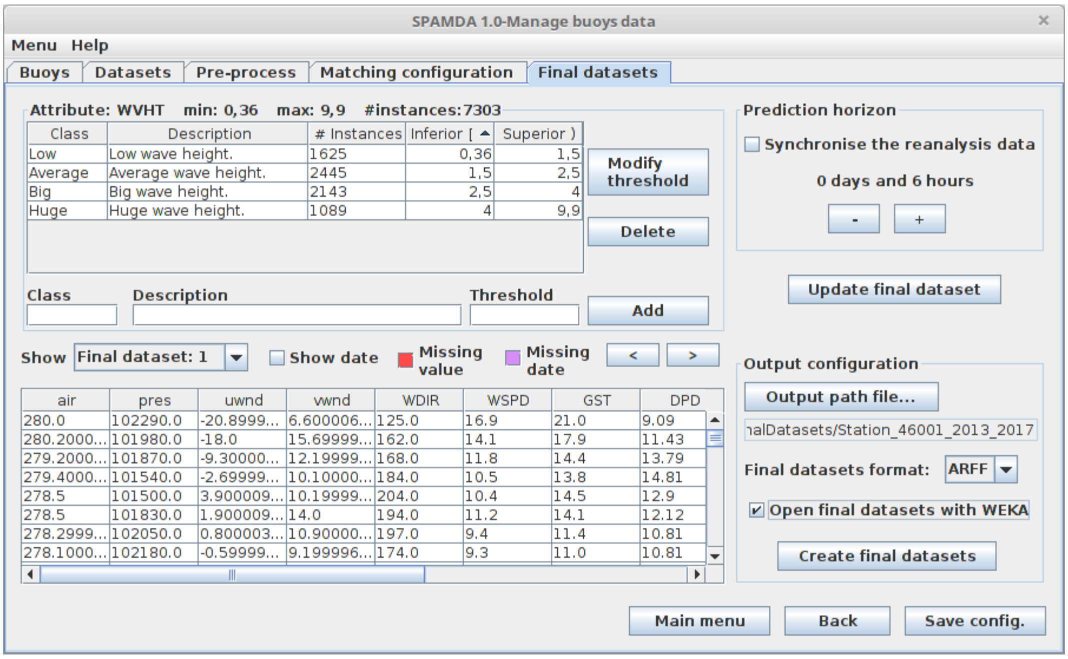


Figure 20. *Final datasets* tab: content of the final dataset created after data integration and discretisation of the output variable in 4 classes.

After configuring the matching process, the researcher can click on the ***Run*** button to jump to the view shown in Fig. 20 and proceed to define the final dataset structure according to the selected prediction task. Given that, in the previous view (Fig. 19), *Classification* was selected, the researcher can now add, modify or delete the thresholds (usually defined by an expert) for discretising the output variable (top left of Fig. 20). After this, the next step is to set the time horizon desired (6 hours by default) and also to activate (if desired) the synchronisation (in time) of reanalysis variables with the output (top right of Fig. 20), as explained in Section 3.5. Then the researcher can click on the ***Update final dataset*** button to see the content shown in the bottom left corner (NDBC observations, NNRP variables, missing values, dates). Finally, after checking that everything is correct, the last step would be to select the name and path of the dataset file, and its output format (CSV or ARFF) and click on the ***Create final datasets*** button (bottom right of Fig. 20). For this example, the following configuration was applied:

- Thresholds: see Table 2.
- Prediction horizon: 6 hours.
- Synchronisation: Disabled.
- Final dataset format: ARFF.

At this point, the final dataset would be created according to the tailored configuration and stored in the computer of the researcher, which already can apply the ML techniques to address the problem of wave classification. Concretely, the final dataset consists of 7302 instances and whose distribution is represented in Table 3.

Table 3. Distribution of instances of the final dataset

Year	Number of instances
2013	1460
2014	1460
2015	1460
2016	1464
2017	1458
	7302

4.2.2. Obtaining classification models with ML algorithms

Now, the process to obtain wave classification models is described using the final dataset previously created with SPAMDA. The modelling will be performed using WEKA as SC and ML tool, which can be opened through SPAMDA, as shown in Fig. 21. Nevertheless, as mentioned above, the researcher can create the final dataset in CSV format in order to use any other ML tools, such as KEEL, Python or R, among others.

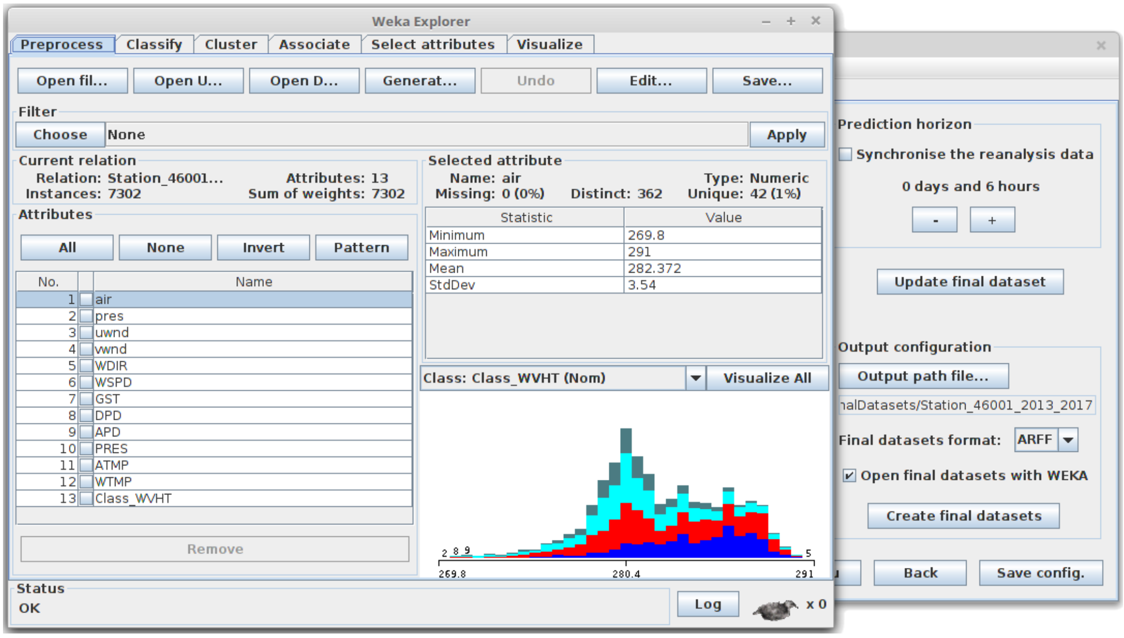


Figure 21. Final dataset opened with the environment Explorer of WEKA (waves classification).

Since the final dataset is a time series of meteorological data (collected from 2013 to 2017), a hold-out scheme (60% train / 40% test) will be used. In this way, years from 2013 to 2015 will be used for the training phase (4380 instances) whereas 2016 and 2017 years will be used for the test phase (2922 instances). Previous to the learning phase, the attributes are normalised to avoid that some attributes dominate others because of a larger scale.

The classification algorithms that will be considered for wave modelling are Logistic Regression [61], C4.5 [62], Random Forest [63], Support Vector Machine [64] and Multilayer Perceptron [65], which will be applied with the default values of the parameters provided by WEKA. Given that Logistic Regression and C4.5 algorithms are deterministic, only one run will be considered for each one. However, Random Forest, Support Vector Machine and Multilayer Perceptron algorithms have a stochastic component, so, in this case, 30 executions for each one will be carried out. The results obtained are shown in Table 4.

Table 4. Results (mean±SD) obtained by the algorithms

Algorithm	**Accuracy (CCR)**	**Kappa**
Logistic Regression	59.0691	0.44447
C4.5	61.7385	0.47852
Random Forest	*68.6516 ± 0.3083*	*0.57040 ± 0.0042*
Support Vector Machine	61.0016 ± 0.0522	0.46770 ± 0.0007
Multilayer Perceptron	**69.7045 ± 1.3033**	**0.58576 ± 0.0178**

As can be seen, Random Forest and Multilayer Perceptron algorithms have achieved similar accuracy, but the performance of the latter is slightly better. Although this is an illustrative classification example using datasets built with SPAMDA, both models have obtained good performance, despite the problem tackled is difficult (prediction is approached six hours in advance).

4.3. Energy flux prediction

As mentioned above, the final dataset of this example is also used with SC and ML algorithms to predict flux of energy. Following sections explain the process of performing the data integration provided by SPAMDA to build the final dataset, modelling the flux of energy by using regression algorithms available in WEKA.

4.3.1. Obtaining the final dataset

The researcher can configure the parameters of the data integration by clicking on the ***Matching configuration*** tab. For this problem, as shown in Fig. 22, the following parameters were selected:

- Attribute to predict: Flux of energy.
- Reanalysis data: Air, pressure, u-wind and v-wind.
- Buoy attributes to be used as inputs: WDIR, WSPD, GST, DPD, APD, PRES, ATMP and WTMP (see Table 1).
- Reanalysis nodes to consider: 1 (only the closest reanalysis node will be used).
- Number of final datasets: In this example, this option is disabled, because only one reanalysis node is considered.
- Prediction task: Regression.

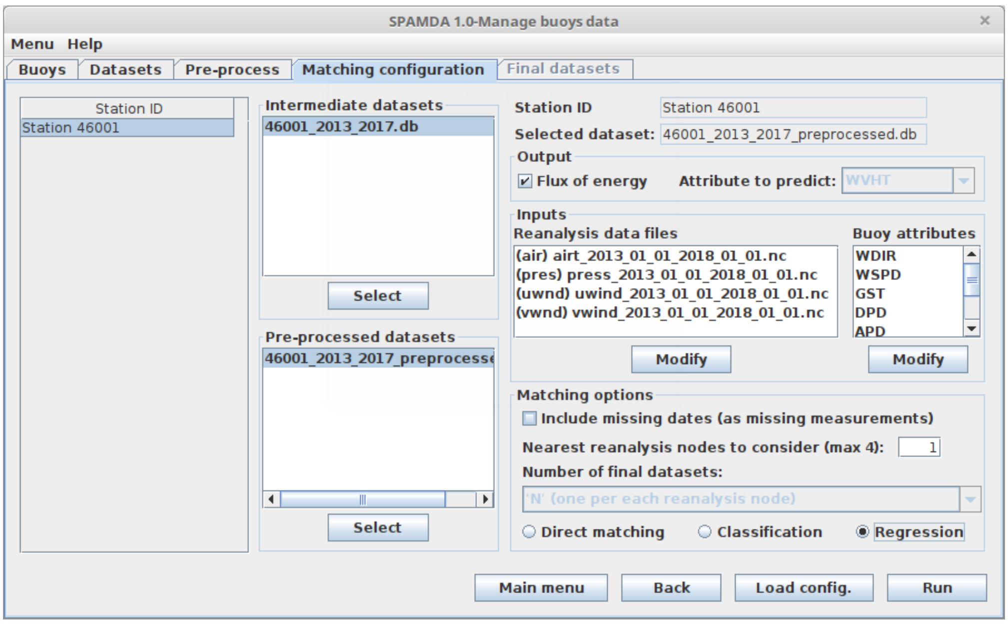


Figure 22. *Matching configuration* tab: parameters for the data integration of the intermediate dataset and the reanalysis files (energy flux prediction).

After configuring the parameters of the matching process, the next step is to define the final dataset structure according to the selected prediction task. Researcher can click on the ***Run*** button to jump to the view shown in Fig. 23. Note that, the thresholds for discretising the output variable (top left of Fig. 23) are disabled due to, in this case, energy flux prediction is a regression problem.

By default, the time horizon is set to 6 hours, that is, the energy flux prediction will be performed 6 hours in advance (top right of Fig. 23), but researchers can increase such time horizon depending on their needs. The synchronisation (in time) of reanalysis variables with the output (explained in Section

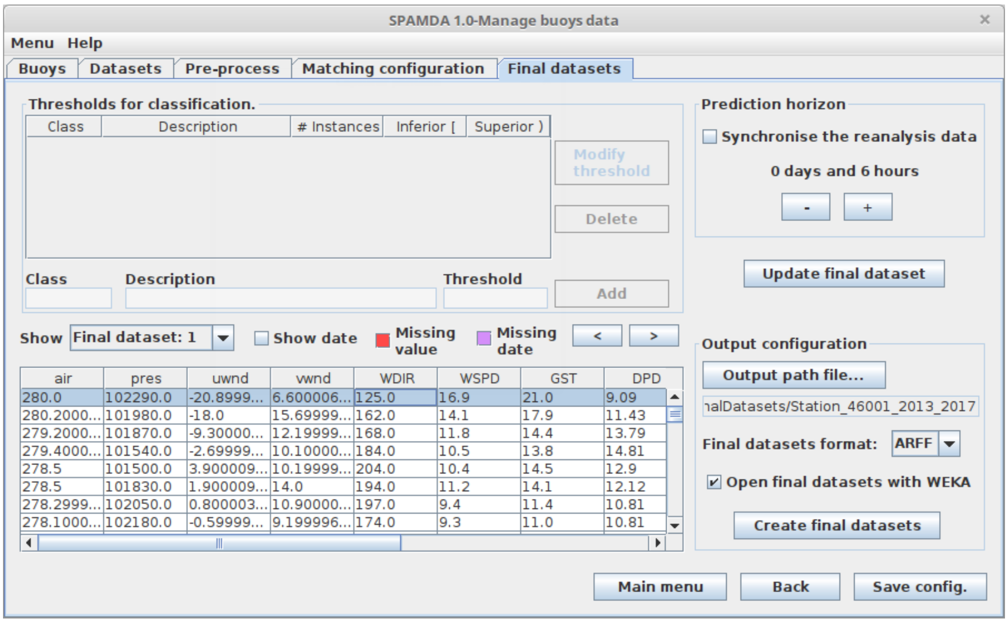


Figure 23. *Final datasets* tab: content of the final dataset created after data integration.

3.5) can be set in this view. By clicking on the ***Update final dataset*** button, researchers can preview the content of the final dataset (bottom left corner of Fig. 23). Finally, the last step would be to set the name, path and output format (CSV or ARFF) of the dataset file, and then the user should click on the ***Create final datasets*** button (bottom right of Fig. 23). For this example, the following configuration was applied:

- Prediction horizon: 6 hours.
- Synchronisation: Disabled.
- Final dataset format: ARFF.

After that, the final dataset would be created and stored in the computer of the researcher according to the introduced configuration, ready to be used as input for SC and ML techniques to tackle the problem of energy flux prediction. The number of instances (7302) and the distribution of the final dataset (Table 3) are the same as in the previous example (waves classification) since the data used to create the final dataset and the time horizon selected (6h) are the same.

4.3.2. Obtaining prediction models with ML algorithms

In this example, WEKA is used as SC and ML tool to obtain energy flux prediction models, as shown in Fig. 24. Nonetheless, the final dataset can be created in CSV format so that the researcher can use any other SC and ML tool.

For this problem, the same partitioning scheme used in the wave classification problem is considered (60% train / 40% test), that is, years from 2013 to 2015 for the training phase (4380 instances) and years 2016 and 2017 for the test phase (2922 instances). Again, the attributes are normalised prior to the learning phase.

To perform the energy flux modelling, one execution will be run for the deterministic algorithm Linear Regression [34], whereas 30 executions will be considered for the stochastic ones: Random Forest [63], Support Vector Machine [64] and Multilayer Perceptron [65]. Table 5 shows the results obtained by each algorithm using the default values of the algorithm parameters provided by WEKA.

As can be checked, Random Forest has achieved the best performance in terms of Root mean squared error. The standard deviation of the results obtained by Multilayer Perceptron indicates that

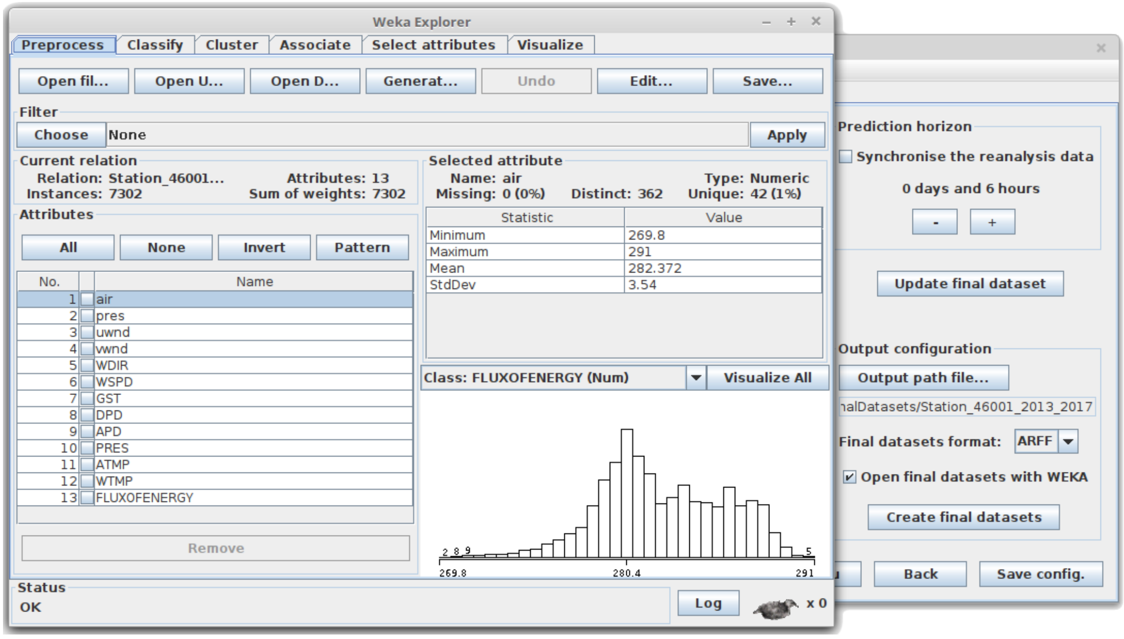


Figure 24. Final dataset opened with the environment Explorer of WEKA (energy flux prediction).

Table 5. Results (mean±SD) obtained by the algorithms

Algorithm	Root mean squared error	Correlation coefficient
Linear Regression	29.6368	0.7296
Random Forest	**23.4353 ± 0.1313**	*0.8408 ± 0.0021*
Support Vector Machine	31.3008 ± 0.1197	0.7275 ± 0.0015
Multilayer Perceptron	*27.1151 ± 7.5536*	**0.8444 ± 0.0193**

this algorithm may has been slightly affected by its stochastic component. However, both Multilayer Perceptron and Random Forest have obtained an excellent Correlation coefficient.

The objective of this case study has been to illustrate the use of datasets created with SPAMDA to address an energy flux prediction problem. An exhaustive comparison of state-of-the-art regression algorithms is not the purpose of this paper. However, note that Multilayer Perceptron and Random Forest algorithms have achieved very good results despite the fact that the energy flux prediction has been performed with a time horizon of 6h.

4.4. *Important remarks*

In this section, it has been described how to use SPAMDA to create final datasets with the aim of classifying waves and predicting flux of energy. However, using the same data described in Section 4 the researcher can quickly address other objectives or different studies by merely tailoring the matching configuration of the data integration process. For example, longer-term wave or energy flux prediction can be addressed by changing the time horizon, waves modelling can be approached from another perspective by creating the final dataset for regression, or environmental modelling can be focused in diverse fields by changing the output meteorological variable.

Furthermore, environmental modelling in other geographical location can be carried out by merely using other collected data.

As SPAMDA performs all data processing and management to create the datasets, it not only prevents researchers from performing repetitive tasks but also prevents them from making possible errors. In this way, researchers can focus on the studies they are carrying out.

5. Conclusions

A new open source tool named SPAMDA with an user-friendly GUI for creating datasets using meteorological data from NDBC and NNRP has been presented in this work. The aim of the tool presented in this work is to provide the research community with an automated, customisable and robust integration for NDBC and NNRP data, serving as a tool for analysis and decision support in marine energy and engineering applications, among others.

Studies on marine energy using ML and SC methodologies apply specific algorithms (extreme learning machine, metaheuristics, Bayesian networks, neural networks, etc) on data using custom-made implementations or scripts developed in some programming language; but they do not allow to build datasets in an automated way ready to be used as input for prediction tasks (classification or regression). These datasets can be easily obtained with SPAMDA by means of the selection of different input parameters, such as predictive and objective variables, output discretisation or prediction horizon. As a result, researchers will benefit from significant support when carrying out environmental modelling related to energy, atmospheric or oceanic studies, among others. Moreover, given that SPAMDA simplifies all the intermediate steps involved in the creation of datasets and manages the extensive casuistry of the data integration (such as entering the meteorological information, managing with the incomplete data, pre-processing tasks, the customisable matching process to merge the data and the preparation of the datasets according to the SC or ML technique to use), it avoids errors and reduces the time needed. In this way, researchers will be able to have more in-depth analysis, which could result in more complete conclusions about the issue under study.

The case study described in Section 4 illustrates how SPAMDA can be used by researchers in a practical approach for environmental modelling, concretely, to classify waves in the Gulf of Alaska depending on their height. The case study also covers an example of energy flux prediction, to predict the wave energy that could be exploited by WECs facilities six hours in advance, although such time horizon is customisable. Given that this work does not focus on models performance, a more extensive validation or comparison study of the results obtained in both examples has not been carried out.

In order to improve SPAMDA, some future work could be focused on new functional modules for managing meteorological data of different formats [66], so that the developed tool can be extended to any other research, new pre-processing functionalities such as filters to analyse the correlation between attributes or new functional modules for recovering missing values using nearby buoys data [67]. Furthermore, the developed software could manage other sources of reanalysis data (with different spatial and temporal resolution), and new output formats for the datasets which could be used as input by other tools for ML such as KEEL (*Knowledge Extraction based on Evolutionary Learning*) [68]. However, such new functionalities can be developed with a reasonable effort to be able to manage each particular casuistry. For example, when dealing with incomplete data, interpreting different data and files structures or carrying out the matching process of two environmental data sources.

Additional material

The source code and the software tool are available at https://github.com/ayrna.

Author Contributions: Conceptualization, Formal analysis and Investigation, Antonio Manuel Gómez-Orellana, Juan Carlos Fernández, Manuel Dorado-Moreno, Pedro Antonio Gutiérrez and César Hervás-Martínez; Funding acquisition, Project administration, Resources and Supervision, Pedro Antonio Gutiérrez and César Hervás-Martínez; Methodology, Antonio Manuel Gómez-Orellana, Juan Carlos Fernández, Manuel Dorado-Moreno, Pedro Antonio Gutiérrez and César Hervás-Martínez; Software, Antonio Manuel Gómez-Orellana, Juan Carlos Fernández and Manuel Dorado-Moreno; Validation, Antonio Manuel Gómez-Orellana, Juan Carlos Fernández, Manuel Dorado-Moreno, Pedro Antonio Gutiérrez and César Hervás-Martínez; Writing – original draft, Antonio Manuel Gómez-Orellana, Juan Carlos Fernández and Manuel Dorado-Moreno.

Funding: This work has been partially subsidised by the projects with references TIN2017-85887-C2-1-P of the Spanish Ministry of Economy and Competitiveness (MINECO), UCO-1261651 of the “Consejería de Economía, Conocimiento, Empresas y Universidad” of the “Junta de Andalucía” (Spain) and FEDER funds of the European Union.

Acknowledgments: The authors also thank to NOAA/OAR/ESRL PSD, Boulder, Colorado, USA for the NCEP Reanalysis data provided from their Web site at https://www.esrl.noaa.gov/psd/, to NOAA/NDBC by its data that were collected and made freely available, to University of Waikato for the WEKA (Waikato Environment for Knowledge Analysis) software tool, to University Corporation for Atmospheric Research/Unidata for the NetCDF (network Common Data Form) Java library and to QOS.ch for the SLF4J (Simple Logging Facade for Java) library.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

F_e	Flux of energy
H_s	Significant wave height
T_e	Wave energy period
p_0	Geographical location of the buoy
p_j	Geographical location of each reanalysis node
lat	Latitude of the point
lon	Longitude of the point
o_t	The attribute to be predicted at the time instant to study
Δt	The prediction horizon
$\mathbf{b}_t$	The vector containing the selected NDBC variables
$\mathbf{r}_t$	The vector containing the selected reanalysis variables

Appendix A. Managing the casuistry of incomplete data

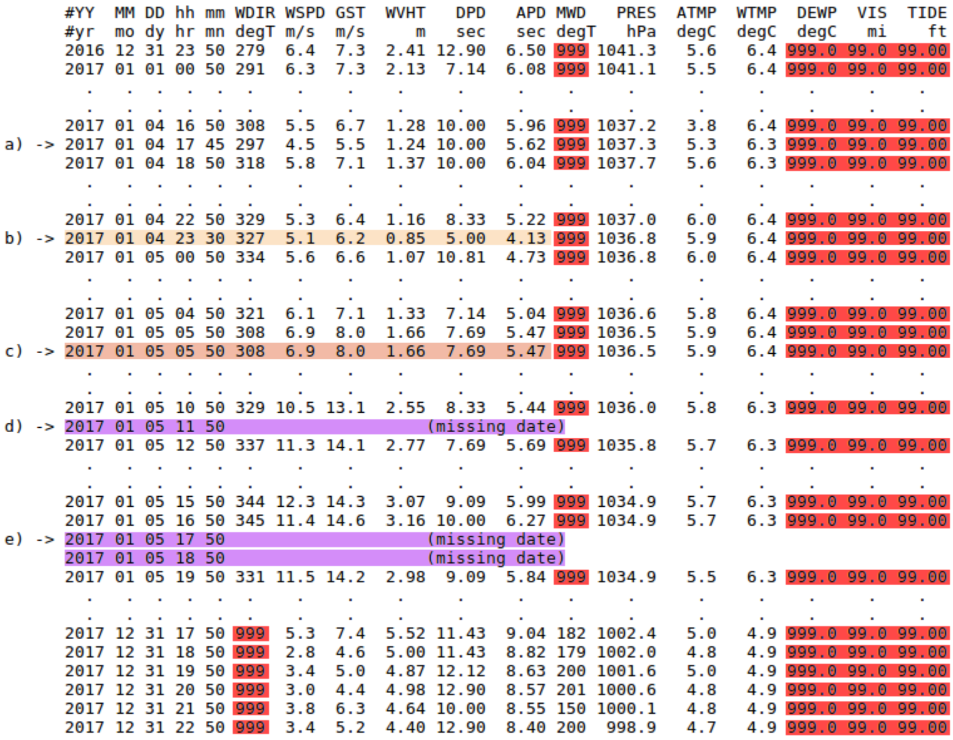
In this appendix, we describe how SPAMDA deals with incomplete data when creating intermediate datasets and performing the matching process.

The measurements collected by the buoys may be incomplete or recorded at a different time than the expected one, due to the weather conditions in which the buoys have to operate. To illustrate this casuistry, the following examples are shown in Fig. A1:

- In the instance marked with a), the measurement of 17:50 was collected at 17:45, 5 minutes earlier.
- In the instance marked with b), the measurement of 23:50 was collected at 23:30, 20 minutes earlier.
- In the instance marked with c), the measurement of 05:50 is duplicated.
- In the instance marked with d), the measurement of 11:50 is missing (missing date or instance).
- In the instance marked with e), the measurement of 17:50 and 18:50 are missing (missing dates or instances).
- Missing values highlighted in red colour.

SPAMDA has been designed to tackle these situations, and it informs researchers of any incidence found while reading the annual text files for creating the intermediate datasets. For the case of measurements that were recorded at a different time than expected, it has been established a time gap of 6 minutes (10% of an hour). Therefore, if the time difference exceeds such value the date will be considered as an unexpected.

Fig. A2 shows the status of the creation of an intermediate dataset with the information of Fig A1. Note that the instance marked with a) has not been informed by SPAMDA as an unexpected date because its time difference is less than 6 minutes. Depending on the affected attribute, NDBC uses a specific value [66] to indicate the presence of lost data (e.g. 99 for VIS and TIDE attributes, 999 for DEWP, MWD and WDIR, etc.). SPAMDA interprets these specific values and, after creating the intermediate dataset, researchers can check if it contains missing values by visualising its statistical information or content. Remember that SPAMDA provides several filters for recovering missing data, which were described in subsection 3.2.



```
         #YY  MM DD hh mm WDIR WSPD GST  WVHT   DPD   APD MWD   PRES  ATMP  WTMP  DEWP  VIS  TIDE
         #yr  mo dy hr mn degT m/s  m/s     m   sec   sec degT   hPa  degC  degC  degC   mi    ft
         2016 12 31 23 50 279  6.4  7.3  2.41 12.90  6.50 999 1041.3   5.6   6.4 999.0 99.0 99.00
         2017 01 01 00 50 291  6.3  7.3  2.13  7.14  6.08 999 1041.1   5.5   6.4 999.0 99.0 99.00
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
         2017 01 04 16 50 308  5.5  6.7  1.28 10.00  5.96 999 1037.2   3.8   6.4 999.0 99.0 99.00
a) ->    2017 01 04 17 45 297  4.5  5.5  1.24 10.00  5.62 999 1037.3   5.3   6.3 999.0 99.0 99.00
         2017 01 04 18 50 318  5.8  7.1  1.37 10.00  6.04 999 1037.7   5.6   6.3 999.0 99.0 99.00
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
         2017 01 04 22 50 329  5.3  6.4  1.16  8.33  5.22 999 1037.0   6.0   6.4 999.0 99.0 99.00
b) ->    2017 01 04 23 30 327  5.1  6.2  0.85  5.00  4.13 999 1036.8   5.9   6.4 999.0 99.0 99.00
         2017 01 05 00 50 334  5.6  6.6  1.07 10.81  4.73 999 1036.8   6.0   6.4 999.0 99.0 99.00
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
         2017 01 05 04 50 321  6.1  7.1  1.33  7.14  5.04 999 1036.6   5.8   6.4 999.0 99.0 99.00
         2017 01 05 05 50 308  6.9  8.0  1.66  7.69  5.47 999 1036.5   5.9   6.4 999.0 99.0 99.00
c) ->    2017 01 05 05 50 308  6.9  8.0  1.66  7.69  5.47 999 1036.5   5.9   6.4 999.0 99.0 99.00
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
         2017 01 05 10 50 329 10.5 13.1  2.55  8.33  5.44 999 1036.0   5.8   6.3 999.0 99.0 99.00
d) ->    2017 01 05 11 50                    (missing date)
         2017 01 05 12 50 337 11.3 14.1  2.77  7.69  5.69 999 1035.8   5.7   6.3 999.0 99.0 99.00
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
         2017 01 05 15 50 344 12.3 14.3  3.07  9.09  5.99 999 1034.9   5.7   6.3 999.0 99.0 99.00
         2017 01 05 16 50 345 11.4 14.6  3.16 10.00  6.27 999 1034.9   5.7   6.3 999.0 99.0 99.00
e) ->    2017 01 05 17 50                    (missing date)
         2017 01 05 18 50                    (missing date)
         2017 01 05 19 50 331 11.5 14.2  2.98  9.09  5.84 999 1034.9   5.5   6.3 999.0 99.0 99.00
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
           .    .  .  .  .  .    .    .     .     .     .   .      .     .     .     .    .     .
         2017 12 31 17 50 999  5.3  7.4  5.52 11.43  9.04 182 1002.4   5.0   4.9 999.0 99.0 99.00
         2017 12 31 18 50 999  2.8  4.6  5.00 11.43  8.82 179 1002.0   4.8   4.9 999.0 99.0 99.00
         2017 12 31 19 50 999  3.4  5.0  4.87 12.12  8.63 200 1001.6   5.0   4.9 999.0 99.0 99.00
         2017 12 31 20 50 999  3.0  4.4  4.98 12.90  8.57 201 1000.6   4.8   4.9 999.0 99.0 99.00
         2017 12 31 21 50 999  3.8  6.3  4.64 10.00  8.55 150 1000.1   4.8   4.9 999.0 99.0 99.00
         2017 12 31 22 50 999  3.4  5.2  4.40 12.90  8.40 200  998.9   4.7   4.9 999.0 99.0 99.00
```

Figure A1. A fragment of an annual text file with different missing value examples.

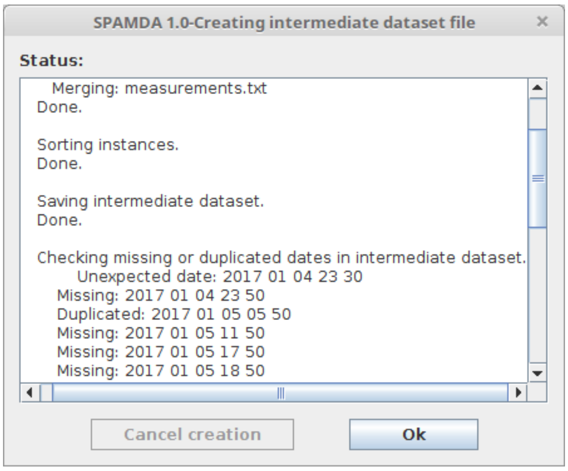


Figure A2. Status of the creation of the intermediate dataset for the example of Fig A1.

SPAMDA takes into account this casuistry when carrying out the matching process. An example is given in Fig. A3. As above-mentioned, the matching process is performed with the nearest measurement (previous or next) within a maximum of 60 minutes of difference. However, in the instance marked with *e*), given that the measurements dates 01/05/2017 17:50 and 01/05/2017 18:50 are missing, the reanalysis date 01/05/2017 18:00 cannot be matched with buoy data (this date is highlighted in mauve colour in the Figure). Depending on the selection made by researchers in the parameter *Include missing dates*, this instance will be included in the final dataset (with missing values for buoy variables) or not.

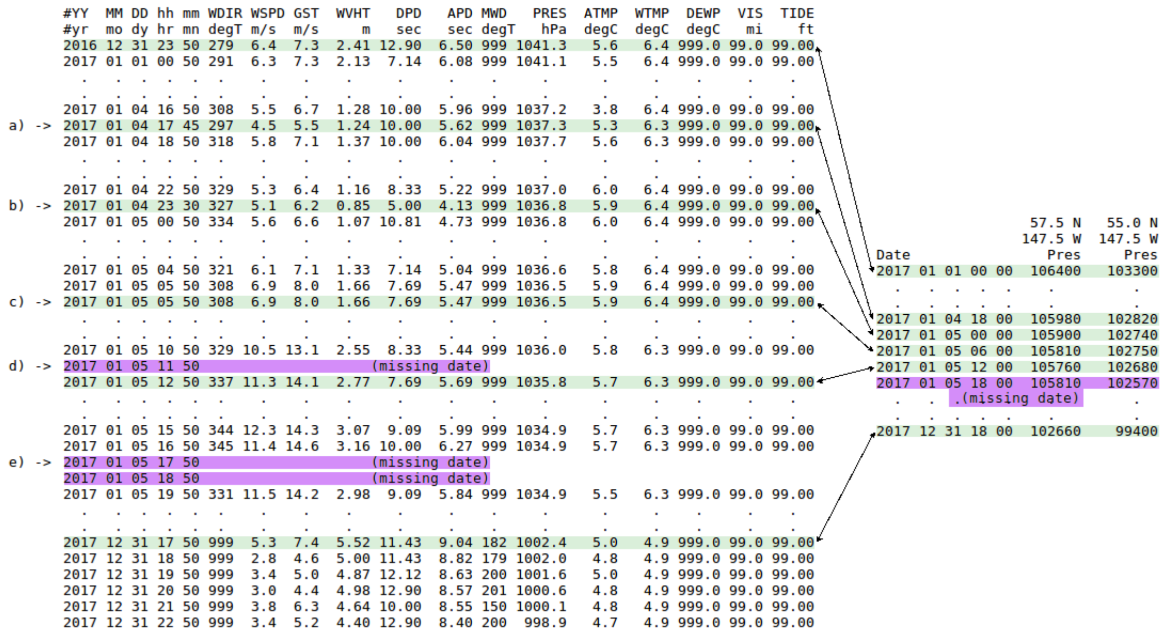


Figure A3. Matching the measurements (left) and the reanalysis data (right).

References

1. Anis, M.S.; Jamil, B.; Ansari, M.A.; Bellos, E. Generalized models for estimation of global solar radiation based on sunshine duration and detailed comparison with the existing: A case study for India. *Sustainable Energy Technologies and Assessments* **2019**, *31*, 179–198. doi:10.1016/j.seta.2018.12.009.
2. Laface, V.; Arena, F.; Soares, C.G. Directional analysis of sea storms. *Ocean Engineering* **2015**, *107*, 45–53. doi:10.1016/j.oceaneng.2015.07.027.
3. Shivam, K.; Tzou, J.C.; Wu, S.C. Multi-Objective Sizing Optimization of a Grid-Connected Solar–Wind Hybrid System Using Climate Classification: A Case Study of Four Locations in Southern Taiwan. *Energies* **2020**, *13*, 2505. doi:10.3390/en13102505.
4. Dorado-Moreno, M.; Cornejo-Bueno, L.; Gutiérrez, P.A.; Prieto, L.; Hervás-Martínez, C.; Salcedo-Sanz, S. Robust estimation of wind power ramp events with reservoir computing. *Renewable Energy* **2017**, *111*, 428–437. doi:10.1016/j.renene.2017.04.016.
5. He, Q.; Zha, C.; Song, W.; Hao, Z.; Du, Y.; Liotta, A.; Perra, C. Improved Particle Swarm Optimization for Sea Surface Temperature Prediction. *Energies* **2020**, *13*, 1369. doi:10.3390/en13061369.
6. Fuchs, H.L.; Gerbi, G.P. Seascape-level variation in turbulence- and wave-generated hydrodynamic signals experienced by plankton. *Progress in Oceanography* **2016**, *141*, 109–129. doi:10.1016/j.pocean.2015.12.010.
7. da Silva, V.d.P.R.; Araújo e Silva, R.; Cavalcanti, E.P.; Braga Campos, C.; Vieira de Azevedo, P.; Singh, V.P.; Rodrigues Pereira, E.R. Trends in solar radiation in NCEP/NCAR database and measurements in northeastern Brazil. *Solar Energy* **2010**, *84*, 1852–1862. doi:10.1016/j.solener.2010.07.011.
8. Gouldby, B.; Méndez, F.J.; Guanche, Y.; Rueda, A.; Mínguez, R. A methodology for deriving extreme nearshore sea conditions for structural design and flood risk analysis. *Coastal Engineering* **2014**, *88*, 15–26. doi:10.1016/j.coastaleng.2014.01.012.
9. Alizadeh, R.; Jia, L.; Nellippallil, A.B.; Wang, G.; Hao, J.; Allen, J.K.; Mistree, F. Ensemble of surrogates and cross-validation for rapid and accurate predictions using small data sets. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **2019**, *33*, 484–501. doi:10.1017/S089006041900026X.
10. Alizadeh, R.; Allen, J.K.; Mistree, F. Managing computational complexity using surrogate models: a critical review. *Research in Engineering Design* **2020**, *31*, 275–298. doi:10.1007/s00163-020-00336-7.
11. Dhanraj Bokde, N.; Mundher Yaseen, Z.; Bruun Andersen, G. ForecastTB—An R Package as a Test-Bench for Time Series Forecasting—Application of Wind Speed and Solar Radiation Modeling. *Energies* **2020**, *13*, 2578. doi:10.3390/en13102578.

12. Lo, C.K.; Lim, Y.S.; Rahman, F.A. New integrated simulation tool for the optimum design of bifacial solar panel with reflectors on a specific site. *Renewable Energy* **2015**, *81*, 293–307. doi:10.1016/j.renene.2015.03.047.
13. Nguyen, T.H.; Prinz, A.; Friisø, T.; Nossum, R.; Tyapin, I. A framework for data integration of offshore wind farms. *Renewable Energy* **2013**, *60*, 150–161. doi:10.1016/j.renene.2013.05.002.
14. Di Bari, R.; Horn, R.; Nienborg, B.; Klinker, F.; Kieseritzky, E.; Pawelz, F. The Environmental Potential of Phase Change Materials in Building Applications. A Multiple Case Investigation Based on Life Cycle Assessment and Building Simulation. *Energies* **2020**, *13*, 3045. doi:10.3390/en13123045.
15. Astiaso Garcia, D.; Bruschi, D. A risk assessment tool for improving safety standards and emergency management in Italian onshore wind farms. *Sustainable Energy Technologies and Assessments* **2016**, *18*, 48–58. doi:10.1016/j.seta.2016.09.009.
16. Raabe, A.L.A.; Klein, A.H.d.F.; González, M.; Medina, R. MEPBAY and SMC: Software tools to support different operational levels of headland-bay beach in coastal engineering projects. *Coastal Engineering* **2010**, *57*, 213–226. doi:10.1016/j.coastaleng.2009.10.008.
17. Motahhir, S.; EL Hammoumi, A.; EL Ghzizal, A.; Derouich, A. Open hardware/software test bench for solar tracker with virtual instrumentation. *Sustainable Energy Technologies and Assessments* **2019**, *31*, 9–16. doi:10.1016/j.seta.2018.11.003.
18. Cascajo, R.; García, E.; Quiles, E.; Correcher, A.; Morant, F. Integration of Marine Wave Energy Converters into Seaports: A Case Study in the Port of Valencia. *Energies* **2019**, *12*, 787. doi:10.3390/en12050787.
19. Zeyringer, M.; Fais, B.; Keppo, I.; Price, J. The potential of marine energy technologies in the UK – Evaluation from a systems perspective. *Renewable Energy* **2018**, *115*, 1281–1293. doi:10.1016/j.renene.2017.07.092.
20. De Jong, M.; Hoppe, T.; Noori, N. City Branding, Sustainable Urban Development and the Rentier State. How do Qatar, Abu Dhabi and Dubai present Themselves in the Age of Post Oil and Global Warming? *Energies* **2019**, *12*, 1657. doi:10.3390/en12091657.
21. Brede, M.; de Vries, B.J.M. The energy transition in a climate-constrained world: Regional vs. global optimization. *Environmental Modelling & Software* **2013**, *44*, 44–61. doi:10.1016/j.envsoft.2012.07.011.
22. Alizadeh, R.; Lund, P.D.; Soltanisehat, L. Outlook on biofuels in future studies: A systematic literature review. *Renewable and Sustainable Energy Reviews* **2020**, *134*, 110326. doi:doi.org/10.1016/j.rser.2020.110326.
23. Falcão, A.F.d.O. Wave energy utilization: A review of the technologies. *Renewable and Sustainable Energy Reviews* **2010**, *14*, 899–918. doi:10.1016/j.rser.2009.11.003.
24. Oliveira-Pinto, S.; Rosa-Santos, P.; Taveira-Pinto, F. Electricity supply to offshore oil and gas platforms from renewable ocean wave energy: Overview and case study analysis. *Energy Conversion and Management* **2019**, *186*, 556–569. doi:10.1016/j.enconman.2019.02.050.
25. Fernández Prieto, L.; Rodríguez Rodríguez, G.; Schallenberg Rodríguez, J. Wave energy to power a desalination plant in the north of Gran Canaria Island: Wave resource, socioeconomic and environmental assessment. *Journal of Environmental Management* **2019**, *231*, 546–551. doi:10.1016/j.jenvman.2018.10.071.
26. Ochi, M.K. *Ocean Waves: The Stochastic Approach*; Cambridge Ocean Technology Series, Cambridge University Press, 1998.
27. Crowley, S.; Porter, R.; Taunton, D.J.; Wilson, P.A. Modelling of the WITT wave energy converter. *Renewable Energy* **2018**, *115*, 159–174. doi:10.1016/j.renene.2017.08.004.
28. Abdelkhalik, O.; Robinett, R.; Zou, S.; Bacelli, G.; Coe, R.; Bull, D.; Wilson, D.; Korde, U. On the control design of wave energy converters with wave prediction. *Journal of Ocean Engineering and Marine Energy* **2016**, *2*, 473–483. doi:10.1007/s40722-016-0048-4.
29. Ringwood, J.V.; Bacelli, G.; Fusco, F. Energy-Maximizing Control of Wave-Energy Converters: The Development of Control System Technology to Optimize Their Operation. *IEEE Control Systems* **2014**, *34*, 30–55. doi:10.1109/MCS.2014.2333253.
30. Wei, C.C. Nearshore Wave Predictions Using Data Mining Techniques during Typhoons: A Case Study near Taiwan's Northeastern Coast. *Energies* **2018**, *11*, 11. doi:10.3390/en11010011.
31. Kaloop, M.R.; Kumar, D.; Zarzoura, F.; Roy, B.; Hu, J.W. A wavelet - Particle swarm optimization - Extreme learning machine hybrid modeling for significant wave height prediction. *Ocean Engineering* **2020**, *213*, 107777. doi:doi.org/10.1016/j.oceaneng.2020.107777.
32. Rusu, L. Assessment of the Wave Energy in the Black Sea Based on a 15-Year Hindcast with Data Assimilation. *Energies* **2015**, *8*, 10370–10388. doi:10.3390/en80910370.

33. Rhee, S.Y.; Park, J.; Inoue, A., Eds. *Soft Computing in Machine Learning*; Springer, 2014.
34. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer-Verlag New York, Inc.: Secaucus, NJ, USA, 2006.
35. Chang, F.J.; Hsu, K.; Chang, L.C., Eds. *Flood Forecasting Using Machine Learning Methods*; MPDI, 2019.
36. Dineva, A.; Mosavi, A.; Faizollahzadeh Ardabili, S.; Vajda, I.; Shamshirband, S.; Rabczuk, T.; Chau, K.W. Review of Soft Computing Models in Design and Control of Rotating Electrical Machines. *Energies* **2019**, *12*, 1049. doi:10.3390/en12061049.
37. Guo, Y.; Wang, J.; Chen, H.; Li, G.; Liu, J.; Xu, C.; Huang, R.; Huang, Y. Machine learning-based thermal response time ahead energy demand prediction for building heating systems. *Applied Energy* **2018**, *221*, 16–27. doi:10.1016/j.apenergy.2018.03.125.
38. Frank, E.; Hall, M.A.; Witten, I.H. The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, 2016.
39. Durán-Rosal, A.M.; Fernández, J.C.; Gutiérrez, P.A.; Hervás-Martínez, C. Detection and prediction of segments containing extreme significant wave heights. *Ocean Engineering* **2017**, *142*, 268–279. doi:10.1016/j.oceaneng.2017.07.009.
40. Kumar, N.K.; Savitha, R.; Al Mamun, A. Regional ocean wave height prediction using sequential learning neural networks. *Ocean Engineering* **2017**, *129*, 605–612. doi:10.1016/j.oceaneng.2016.10.033.
41. Ali, M.; Prasad, R.; Xiang, Y.; Deo, R.C. Near real-time significant wave height forecasting with hybridized multiple linear regression algorithms. *Renewable and Sustainable Energy Reviews* **2020**, *132*, 110003. doi:doi.org/10.1016/j.rser.2020.110003.
42. Cornejo-Bueno, L.; Nieto-Borge, J.; García-Díaz, P.; Rodríguez, G.; Salcedo-Sanz, S. Significant wave height and energy flux prediction for marine energy applications: A grouping genetic algorithm – Extreme Learning Machine approach. *Renewable Energy* **2016**, *97*, 380 – 389. doi:doi.org/10.1016/j.renene.2016.05.094.
43. Emmanouil, S.; Aguilar, S.G.; Nane, G.F.; Schouten, J.J. Statistical models for improving significant wave height predictions in offshore operations. *Ocean Engineering* **2020**, *206*, 107249. doi:doi.org/10.1016/j.oceaneng.2020.107249.
44. Shamshirband, S.; Mosavi, A.; Rabczuk, T.; Nabipour, N.; wing Chau, K. 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. *Engineering Applications of Computational Fluid Mechanics* **2020**, *14*, 805–817. doi:10.1080/19942060.2020.1773932.
45. Johansson, L.; Epitropou, V.; Karatzas, K.; Karppinen, A.; Wanner, L.; Vrochidis, S.; Bassoukos, A.; Kukkonen, J.; Kompatsiaris, I. Fusion of meteorological and air quality data extracted from the web for personalized environmental information services. *Environmental Modelling & Software* **2015**, *64*, 143–155. doi:10.1016/j.envsoft.2014.11.021.
46. Fernández, J.C.; Salcedo-Sanz, S.; Gutiérrez, P.A.; Alexandre, E.; Hervás-Martínez, C. Significant wave height and energy flux range forecast with machine learning classifiers. *Engineering Applications of Artificial Intelligence* **2015**, *43*, 44–53. doi:10.1016/j.engappai.2015.03.012.
47. Adams, J.; Flora, S. Correlating seabird movements with ocean winds: linking satellite telemetry with ocean scatterometry. *Marine Biology* **2010**, *157*, 915–929. doi:10.1007/s00227-009-1367-y.
48. National Data Buoy Center. National Oceanic and Atmospheric Administration of the USA (NOAA). http://www.ndbc.noaa.gov/. (Accessed 10 December 2020).
49. Kalnay, E.; Kanamitsu, M.; Kistler, R.; Collins, W.; Deaven, D.; Gandin, L.; Iredell, M.; Saha, S.; White, G.; Woollen, J.; Zhu, Y.; Leetmaa, A.; Reynolds, R.; Chelliah, M.; Ebisuzaki, W.; Higgins, W.; Janowiak, J.; Mo, K.C.; Ropelewski, C.; Wang, J.; Jenne, R.; Joseph, D. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society* **1996**, *77*, 437–471. doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.
50. Kistler, R.; Collins, W.; Saha, S.; White, G.; Woollen, J.; Kalnay, E.; Chelliah, M.; Ebisuzaki, W.; Kanamitsu, M.; Kousky, V.; van den Dool, H.; Jenne, R.; Fiorino, M. The NCEP–NCAR 50–Year Reanalysis: Monthly Means CD–ROM and Documentation. *Bulletin of the American Meteorological Society* **2001**, *82*, 247–267. doi:10.1175/1520-0477(2001)082<0247:TNNYRM>2.3.CO;2.
51. The WEKA Data Mining Software: Attribute-Relation File Format (ARFF). https://www.cs.waikato.ac.nz/ml/weka/arff.html. (Accessed 10 December 2020).

52. Ali, M.; Prasad, R. Significant wave height forecasting via an extreme learning machine model integrated with improved complete ensemble empirical mode decomposition. *Renewable and Sustainable Energy Reviews* **2019**, *104*, 281 – 295. doi:10.1016/j.rser.2019.01.014.
53. Chatziioannou, K.; Katsardi, V.; Koukouselis, A.; Mistakidis, E. The effect of nonlinear wave-structure and soil-structure interactions in the design of an offshore structure. *Marine Structures* **2017**, *52*, 126–152. doi:10.1016/j.marstruc.2016.11.003.
54. Dalgic, Y.; Lazakis, I.; Dinwoodie, I.; McMillan, D.; Revie, M. Advanced logistics planning for offshore wind farm operation and maintenance activities. *Ocean Engineering* **2015**, *101*, 211–226. doi:10.1016/j.oceaneng.2015.04.040.
55. Spaulding, M.L.; Grilli, A.; Damon, C.; Crean, T.; Fugate, G.; Oakley, B.A.; Stempel, P. STORMTOOLS: Coastal Environmental Risk Index (CERI). *Journal of Marine Science and Engineering* **2016**, *4*, 54. doi:10.3390/jmse4030054.
56. National Data Buoy Center. NDBC - Historical NDBC Data. http://www.ndbc.noaa.gov/historical_data.shtml. (Accessed 10 December 2020).
57. National Data Buoy Center. NDBC - Important NDBC Web Site Changes. http://www.ndbc.noaa.gov/mods.shtml. (Accessed 10 December 2020).
58. NOAA/OAR/ESRL PSD. ESRL : PSD : NCEP/NCAR Reanalysis 1. https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html. (Accessed 15 January 2019).
59. Unidata. Network Common Data Form (NetCDF) version 4.6.10 [software]. Boulder, CO: UCAR/Unidata., 2017. doi:10.5065/D6H70CW6.
60. de Smith, M.J.; Goodchild, M.F.; Longley, P.A. *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*, 3rd revised ed.; Matador, 2009.
61. Hosmer Jr, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied logistic regression, 3rd Edition*; John Wiley & Sons, 2013.
62. Quinlan, J.R. *C4. 5: Programs for machine learning*; Morgan Kaufmann, 1992.
63. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32. doi:10.1023/A:1010933404324.
64. Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20*, 273–297. doi:10.1007/BF00994018.
65. Haykin, S. *Neural networks: a comprehensive foundation*; Prentice Hall PTR, 1994.
66. National Data Buoy Center. NDBC - Measurement Descriptions and Units. https://www.ndbc.noaa.gov/measdes.shtml. (Accessed 10 December 2020).
67. Durán-Rosal, A.M.; Hervás-Martínez, C.; Tallón-Ballesteros, A.J.; Martínez-Estudillo, A.C.; Salcedo-Sanz, S. Massive missing data reconstruction in ocean buoys with evolutionary product unit neural networks. *Ocean Engineering* **2016**, *117*, 292–301. doi:10.1016/j.oceaneng.2016.03.053.
68. Alcalá-Fdez, J.; Sánchez, L.; García, S.; del Jesús, M.J.; Ventura, S.; Garrell, J.M.; Otero, J.; Romero, C.; Bacardit, J.; Rivas, V.M.; Fernández, J.C.; Herrera, F. KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Comput.* **2009**, *13*, 307–318. doi:10.1007/s00500-008-0323-y.