Manuscript Details

Manuscript number	SETA_2020_427
Title	A new software tool to facilitate environmental modelling procedures based on meteorological data management
Article type	Full Length Article

Abstract

Meteorological data are extensively used to perform environmental learning. Machine Learning (ML) techniques, a valuable support in many research areas, require datasets containing information related to the topic under study, which are not always available in an appropriate format, and its preparation and pre-processing implies a lot of time and effort by the 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 by customisable procedures, in terms of predictive and objective variables and temporal resolution. These datasets can be used by ML methodologies for prediction tasks (classification or regression) that support improvement of sustainability energy production, design of production systems as WECs or environmental modelling, among others.

Keywords	Meteorological data;Reanalysis data;Creating datasets;Environmental prediction;Marine energy.
Taxonomy	Wave Power, Renewable Energy Forecasting

Corresponding Author	Antonio Manuel Gómez-Orellana
Order of Authors	Antonio Manuel Gómez-Orellana, Juan Carlos Fernández, Manuel Dorado-Moreno, P.A. Gutierrez, Cesar Hervás-Martínez

Submission Files Included in this PDF

File Name [File Type]

CoverLetter.pdf [Cover Letter]

letter-response-editor.pdf [Response to Reviewers]

Highlights.pdf [Highlights]

GraphicalAbstract.pdf [Graphical Abstract]

Manuscript.pdf [Manuscript File]

DeclarationOfInterestStatement.pdf [Conflict of Interest]

DownloadSPAMDA_for_editor_and_reviewers.pdf [Supplementary Material]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

University of Cordoba.
Department of Computer Science and Numerical Analysis.
Rabanales Campus, "Albert Einstein" building, 3rd floor.
14071 – Córdoba (Spain).
Ph. +34 957218349 Fax +34 957218630

Sustainable Energy Technologies and Assessments

Cordoba, Spain, April 21, 2020.

Dear Editor,

Please find attached a manuscript, entitled "*A new software tool to facilitate environmental modelling procedures based on meteorological data management*" submitted for possible publication in Sustainable Energy Technologies and Assessments.

We state that it is our original work, and it has not been submitted, nor it is under consideration in any other forum.

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.

Declaration for any conflicts of interest

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Novelty of this work

This paper presents a new open source tool named SPAMDA for the creation, in an easy way, without the need of knowledge in programming languages and using a graphical interface, of datasets integrated by meteorological variables from two sources of information. These datasets can later be used with Machine Learning methodologies in classification and prediction studies, such as prediction of significant wave height and energy flux in coastal and ocean areas. This tool, among other functionalities (Section 3 of the paper describes concrete and comprehensive information of the developed software), allows managing and storing datasets with different configurations, in terms of predictive and objective variables and different temporal resolution, integrating two different sources of information available for the scientific community (NDBC and NNRP). Note that this data integration process has an extensive casuistry giving rise to incomplete datasets and it requires laborious pre-processing procedures, that implies a lot of time and effort by the researchers, that usually lead to errors (Appendix A of this work shows the casuistry and some problems of the mentioned data integration process).

To our knowledge, there are no software tools that can be used for integrating these two sources of information with the functionalities and options offered by SPAMDA, which, by means of an user-friendly GUI, facilitates the management and storage of buoys and reanalysis data, time and geographical coordinates conversion of both kind of data, missing values handling (dates or measurements not recorded), pre-processing tasks, the generation of datasets to be used by ML techniques in prediction tasks (classification or regression) including time horizon selection and output discretisation, among others. Therefore, the developed software tool will allow researchers focusing on the study of the meteorological aspects of the observations, providing a series of novelties and functionalities briefly described at the end of the Introduction section.

On the other hand, increasingly, Machine Learning techniques are been applied to tackle wave characterisation, feasibility assessment of wave energy extraction and designing offshore structures, and this software allows the creation of datasets ready to use with the Weka Data Mining tool (integrated in the software itself) or in .csv format for being used with any other prediction tool. So, we hope that SPAMDA can be used by researchers when carrying out environmental modelling related to energy, atmospheric or oceanic studies or design of production systems as WECs (Wave Energy Converters, that need a proper prediction of waves in order to maximise the wave energy extraction), among others, using the two sources of information described. Moreover, it could manage other sources of reanalysis data with different spatial and temporal resolution with a reasonable development effort.

Relationship with previous publications in Sustainable Energy Technologies and Assessments

Regarding the aims and scope of this journal, this paper is related to the use of new technologies and their utility and support on the conversion and production of energy, efficiency and the sustainability of the environment. The characterisation and prediction of flux of energy generated by the movement of the waves from meteorological data falls within the aforementioned concepts, since it allows knowing, before the installation of WECs and energy extraction systems, whether a certain coastal or oceanic area could be exploited to produce renewable and sustainable energy, as well as knowing the design and implementation needs of these systems.

SETA journal has published papers related to wave energy prediction, significant wave height, software tools on sustainable energy, WECs and applied Machine Learning methodologies, so we think that this work could be considered for evaluation in this journal.

Should you have any doubts or question about this submission, please do not hesitate in contacting us,

Best regards,

Antonio M. Gómez-Orellana, Juan C. Fernández, Manuel Dorado-Moreno, Pedro A. Gutiérrez, César Hervás-Martínez
Department of Computer Science and Numerical Analysis, University of Cordoba, 14071, Córdoba, Spain.
email: {am.gomez@uco.es, jfcaballero@uco.es, manuel.dorado@uco.es, pagutierrez@uco.es, chervas@uco.es}

Paper title:

A new software tool to facilitate environmental modelling procedures based on meteorological data management

Paper ref: SETA_2020_427

Authors:

Antonio M. Gómez-Orellana, Juan C. Fernández, Manuel Dorado-Moreno, Pedro A. Gutiérrez, César Hervás-Martínez.

Response to the Editor:

We would like to thank the Editor for giving us the opportunity of resubmit our paper. Following the indications, we have addressed the points required.

- Editor comments are written in ***black italic.***
- Author's comments are written in blue.

Editor Comments:

1. I have performed an initial review of your manuscript and would like to see more explicitly what the novelty is about. The cover letter is missing this important information, and also about how it relates to previous publications in our Journal; the cover letter should also include a declaration for any conflicts of interest. You are requested to more explicitly describe what the novelty of your work is about throughout the main manuscript and a statement of this to be included in the cover letter.

First of all, we would like to thank the editor for the suggestions given to improve the paper. According to them, we have included these issues in the cover letter, and we have also taken the opportunity to write some of these clarifications in the paper. Concretely, the Introduction (last paragraphs before the beginning of Section 2), the Conclusions and the Abstract have been modified.

University of Cordoba.
Department of Computer Science and Numerical Analysis.
Rabanales Campus, "Albert Einstein" building, 3rd floor.
14071 – Córdoba (Spain).
Ph. +34 957218349 Fax +34 957218630

Sustainable Energy Technologies and Assessments

A new software tool to facilitate environmental modelling procedures based on meteorological data management

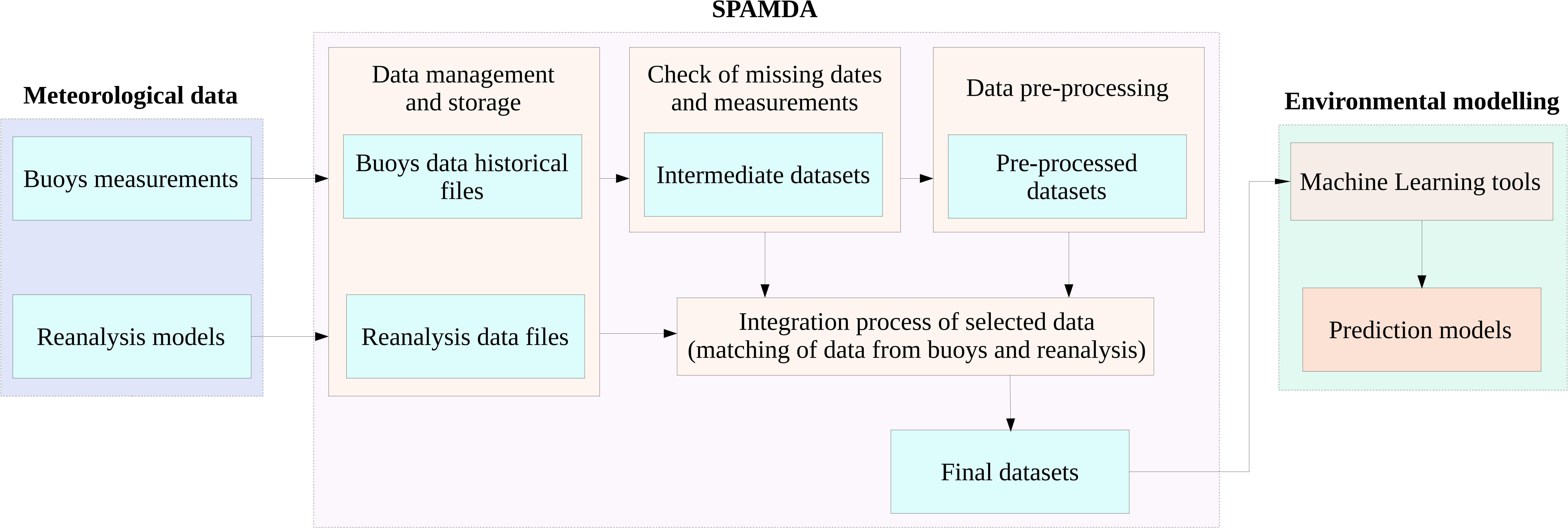
Highlights

- A software tool to assist researchers in environmental modelling studies.
- Data processing and management: researchers can focus on their studies.
- Data integration from two sources: buoys measurements and reanalysis models.
- The content of datasets can be customised depending on researchers needs.
- Datasets are ready to be used in prediction tasks by Machine Learning tools.

Antonio M. Gómez-Orellana

Department of Computer Science and Numerical Analysis, University of Cordoba, 14071, Córdoba, Spain.

email: am.gomez@uco.es


SPAMDA
Meteorological data
Buoys measurements
Reanalysis models
Data management and storage
Buoys data historical files
Reanalysis data files
Check of missing dates and measurements
Intermediate datasets
Data pre-processing
Pre-processed datasets
Integration process of selected data (matching of data from buoys and reanalysis)
Final datasets
Environmental modelling
Machine Learning tools
Prediction models

A new software tool to facilitate environmental modelling procedures based on meteorological data management

Antonio M. Gómez-Orellana[a,*], Juan C. Fernández[a,**], Manuel Dorado-Moreno[a], Pedro A. Gutiérrez[a], César Hervás-Martínez[a]

[a] *Department of Computer Science and Numerical Analysis, University of Cordoba, 14071, Córdoba, Spain.*

Abstract

Meteorological data are extensively used to perform environmental learning. Machine Learning (ML) techniques, a valuable support in many research areas, require datasets containing information related to the topic under study, which are not always available in an appropriate format, and its preparation and pre-processing implies a lot of time and effort by the 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 by customisable procedures, in terms of predictive and objective variables and temporal resolution. These datasets can be used by ML methodologies for prediction tasks (classification or regression) that support improvement of sustainability energy production, design of production systems as WECs or environmental modelling, among others.

Keywords: Meteorological data, Reanalysis data, Creating datasets, Environmental prediction, Marine energy

*Corresponding author.
**Corresponding author.
Email addresses: `am.gomez@uco.es` (Antonio M. Gómez-Orellana), `jfcaballero@uco.es` (Juan C. Fernández), `manuel.dorado@uco.es` (Manuel Dorado-Moreno), `pagutierrez@uco.es` (Pedro A. Gutiérrez), `chervas@uco.es` (César Hervás-Martínez)

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], assessment and control of wind turbine [3], wind power ramp events prediction [4], evaluation of salinity-gradient energy harvesting [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.

On the other hand, 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 [9], a software tool for designing solar water heating systems is developed, which simulates different situations and finds the best technical and economical solution. In [10], 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 [11] in order to facilitate data exchange and improve operation and maintenance practices. In [12], a load management application to maximize the used energy of thermal systems and reduce the costs and emissions is developed. A risk assessment tool to improve safety standards and emergency management in onshore wind farms, is presented in [13]. Raabe et al. [14] 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. [15] 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, and it includes ocean thermal energy, marine tidal current energy and wave energy, among others. Its benefits and great potential [16] 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 [17, 18]. Wave energy exhibits a more stable power supply than wind energy and even solar energy. In recent years *Wave Energy Converters* (WECs) [19] 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 [20] or seawater desalination plants [21], 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 [22]. As a consequence of this complexity, many aspects of WEC design, deployment and operation [23, 24, 25] need a proper prediction of waves, in order to maximise the wave energy extraction. 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). Additionally, wave predictions are also helpful for designing offshore structures [26], operational works in the sea [27], providing technical information to the coastal and coral reef planners in developing countries [28], etc.

Currently, and as a support to traditional study procedures, Machine Learning (ML) techniques [29, 30] are being widely used in numerous research fields related to classification, regression and optimisation tasks, obtaining significant improvements in the performance of the results [31, 32, 33]. 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) [34] software tool provides researchers with a wide collection of ML algorithms. In this way, ML techniques have been already applied to tackle wave characterisation, accurately estimating H_s and T_e parameters [35, 36], given that robustness of ML methods can tackle the previously explained difficulties in wave energy prediction. The problem is that, in order to apply these methods, 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 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 [37, 38, 39], then a data integration process, denominated as the matching process in this document, has to be carried out by the researchers to manually create the datasets with the needed information. Moreover, depending on the subject and the 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 for ML techniques, 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, and 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. 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) [40] and the *National Centers for Environmental Prediction* (NCEP)/*National Center for Atmospheric Research* (NCAR) *Reanalysis Project* (NNRP or R1) [41, 42]. The 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 the researchers focus on the study of the meteorological aspects of the observations. The datasets obtained are ready to be used as input for ML techniques in prediction tasks (classification or regression), although the 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) [43], which is the one used by WEKA. Besides, the datasets can also be generated in *Comma-Separated Values* (CSV) format, enabling the researchers to use others tools.

Up to now, and to the best of our knowledge, this is the first software tool addressing the problem previously discussed, integrating meteorological data from NDBC and NNRP. The novelties and functionalities that SPAMDA offers to the researchers 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 ML techniques, output discretisation or prediction horizon, among others.

- It makes the researcher focus on environmental modelling, without having to worry about mechanical tasks, avoiding laborious pre-processing procedures that imply a lot of time and effort.
- It manages the extensive casuistry of data integration which can lead to incomplete datasets.
- It avoids possible researcher errors in the intermediate steps of the process, such as time and geographical coordinates conversion or missing values handling (dates or measurements not recorded), among others.
- It provides information about the quality and quantity of the data.
- It includes different pre-processing tasks, such as normalisation and missing data recovery.
- It facilitates data management and well-organised storage of the datasets.
- The created datasets can be easily used by ML tools.
- 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 novel support tool to researchers when carrying out environmental modelling related to energy, design of production systems as WECs (that need a proper prediction of waves in order to maximise the wave energy extraction) and atmospheric or oceanic studies, among others.

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 [44], 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 [45]. 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 5.

- 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, 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.

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.

The reanalysis data is available in the NNRP website [46], 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 [47], 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.

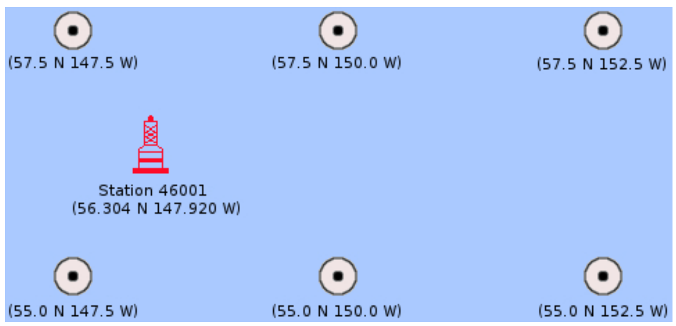


Figure 2: Example of a six sub-grid reanalysis nodes around the *Station 46001*.

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 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).

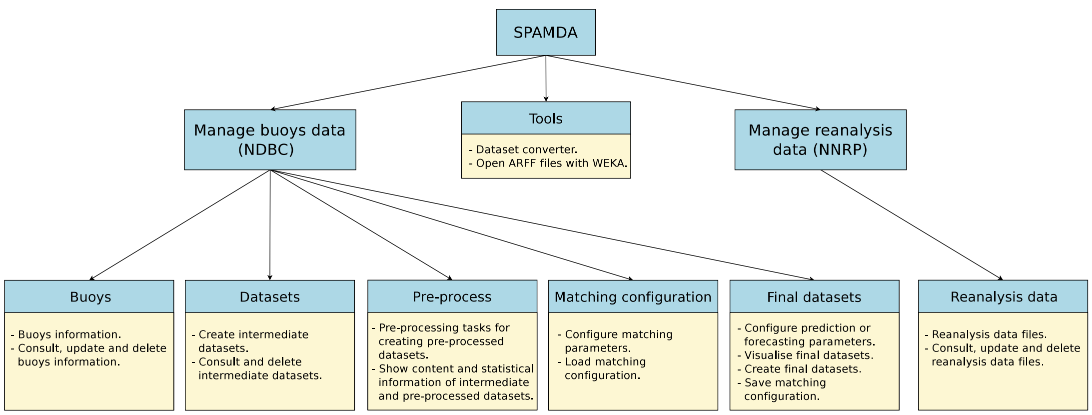


Figure 3: Brief outline of the functionality provided by SPAMDA.

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, the researchers can keep the reanalysis data files updated for their studies. Such files will be used, depending on the 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.

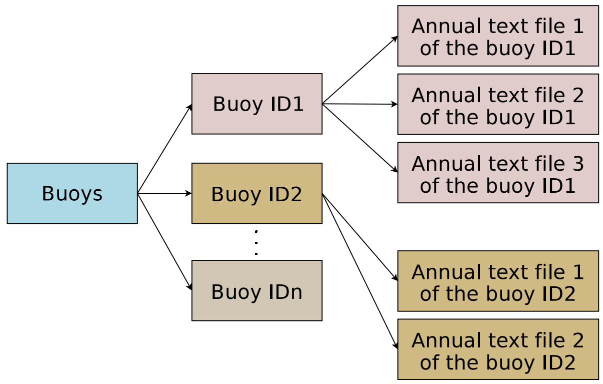


Figure 4: Example of entering two buoys with its annual text files.

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, the 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.

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 dat can be

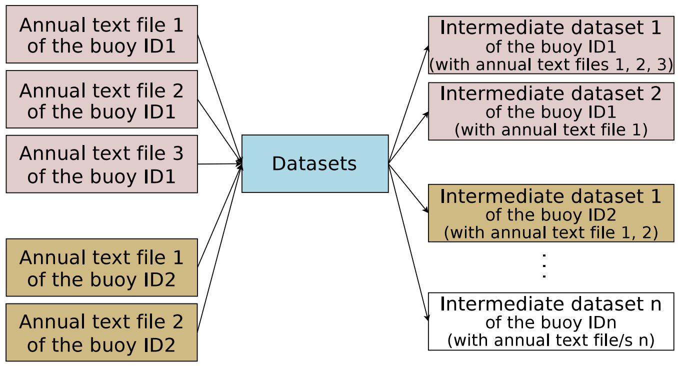


Figure 5: Example of the creation of the intermediate datasets.

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 the 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 the researchers can create as many pre-processed datasets as they consider opportune.

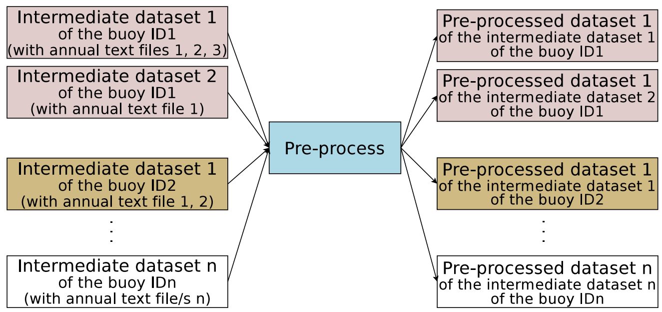


Figure 6: Example of the creation of pre-processed datasets.

3.4. Matching configuration

In order to use the data provided by the two sources of information described in Section 2, it is necessary to carry out a data integration process, denominated as the matching process in this document, to merge and format such data. 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 × 147.5 W and 55.0 N × 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 5 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

Date	57.5 N 147.5 W Pres	55.0 N 147.5 W 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 the 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* [38]: 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 [46], 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, the 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, the researcher 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 [48] 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:

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

  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:

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

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 $\times$ 147.5 W and 55.0 N $\times$ 147.5 W.

– *'N' (one per each reanalysis node)*: As many final datasets as the number of nearest N reanalysis nodes configured by the 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 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 $\times$ 147.5 W, 55.0 N $\times$ 147.5 W, 57.5 N $\times$ 150.0 W and 55.0 N $\times$ 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 matched information considering the data shown in Fig. 7 and using the following matching configuration[1]:

- Variable WVHT as attribute to predict.
- Variable Pres as reanalysis input attribute.
- Variable WSDP as buoy input attribute.

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

- 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

Figure 8: An example of the matched information 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, 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 the 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 is lost 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. 8, 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 better understanding, considering the matched information shown in Fig. 8, an example of the creation of one *Regression* dataset is shown in Fig. 9. 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.

```
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
```

Figure 9: An example of the creation of a *Regression* dataset, with a prediction horizon of 6h and without synchronisation.

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. 8, an example of the creation of the same dataset but applying synchronisation (see Eq. 5) is shown in Fig. 10.

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 values) for the attribute

```
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
```

Figure 10: An example of the creation of a *Regression* dataset, with a prediction horizon of 6h and with synchronisation.

selected as output (attribute to be predicted). SPAMDA allows the 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. 8, an example of the creation of a *Classification* dataset is shown in Fig. 11. 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. 11

Class	**Description**	**Inferior [**	**Superior)**
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

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

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

discretised according to the thresholds. 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* [43], which is used by WEKA. SPAMDA allows the 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 the 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, the 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. 12.

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.

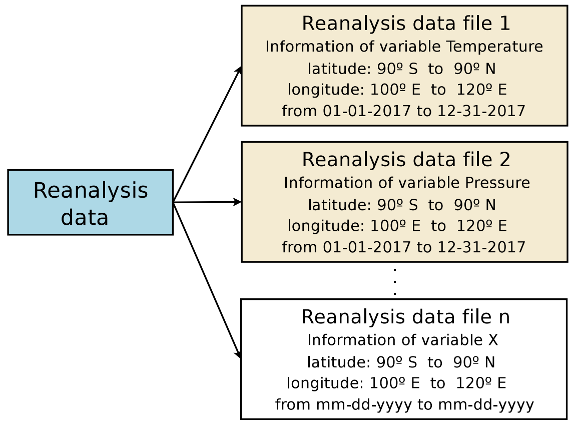


Figure 12: Example of entering two reanalysis data files.

4. Case study

This section will describe how the application works in a practical approach. To do so, an example showing how to create a fully processed dataset (final dataset), starting from the raw data, will be described. The objective of this final dataset is to be used with ML algorithms for environmental modelling, concretely, to classify waves in the Gulf of Alaska depending on their height. In this case, the problem will be 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, energy flux range forecasting, missing buoy data reconstruction, or extreme significant wave heights detection, among others.

4.1. Obtaining the final dataset

The data collected to perform this example 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 mea-

surements: 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 $\times$ 147.5 W, 57.5 N $\times$ 150 W, 55 N $\times$ 147.5 W and 55 N $\times$ 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[2], the researcher can open SPAMDA. In Figure 13, the main window is shown. In order to input the reanalysis data which will be used in further steps for creating the final dataset, the researcher has to select the option *Manage reanalysis data*.

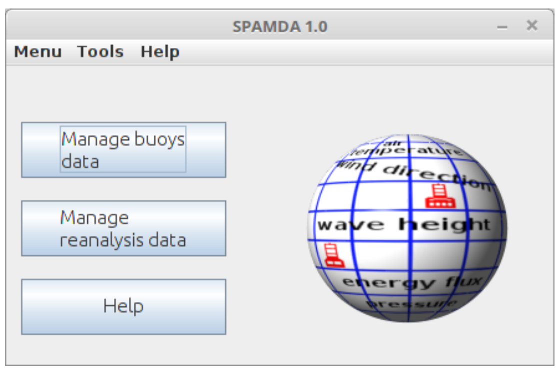


Figure 13: SPAMDA main window.

Then, the window of Figure 14 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 window can be closed and the user can go back to the main window to continue entering the information related to the buoy under study.

After that, the researcher has to select *Manage buoys data* to open the

[2]Further instructions for downloading this data can be found in the user manual of the application.

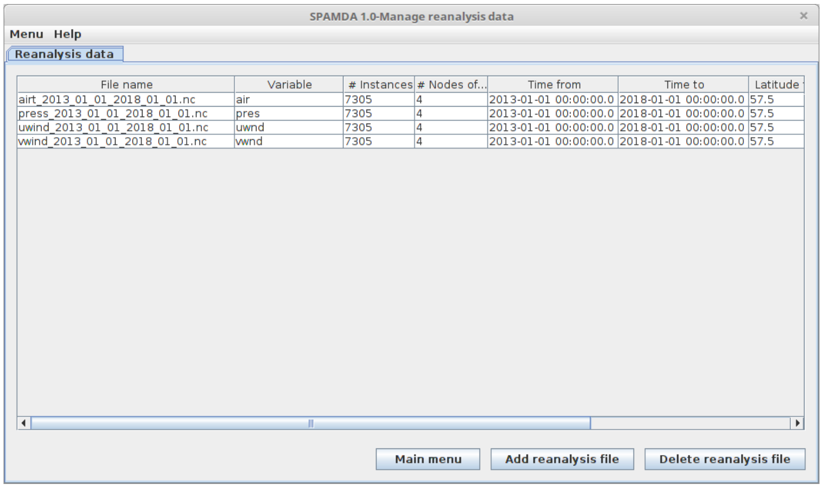


Figure 14: *Manage reanalysis data* window: downloaded files containing the four closest reanalysis nodes.

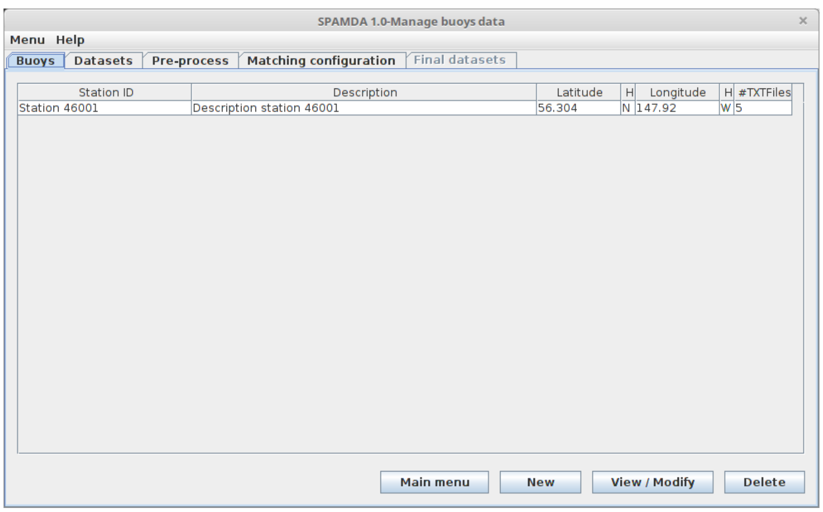


Figure 15: *Buoys* tab: buoy ID 46001.

window shown in Figure 15, 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 window shown in Figure 16 pops up.

Here the information about the buoy has to be included: the *Station ID*, its

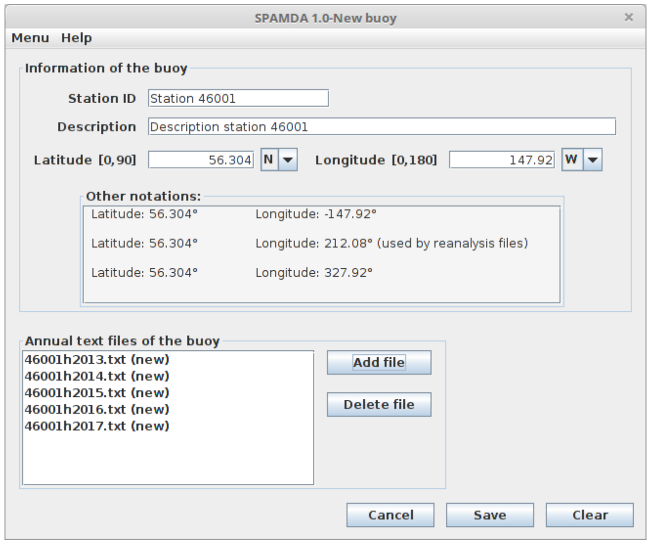


Figure 16: *New buoy* window: information of the buoy ID 46001.

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 introduced, it is necessary to click on the *Save* button to insert the buoy in SPAMDA database. After that, the window 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 Figure 15) to switch to the corresponding view (see Figure 17). 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 Figure 18 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,

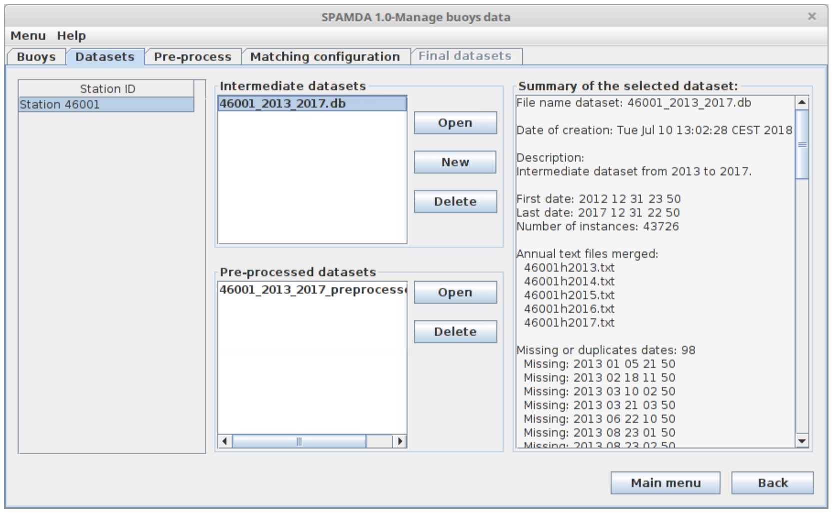


Figure 17: *Datasets* tab: intermediate datasets of the buoy ID 46001.

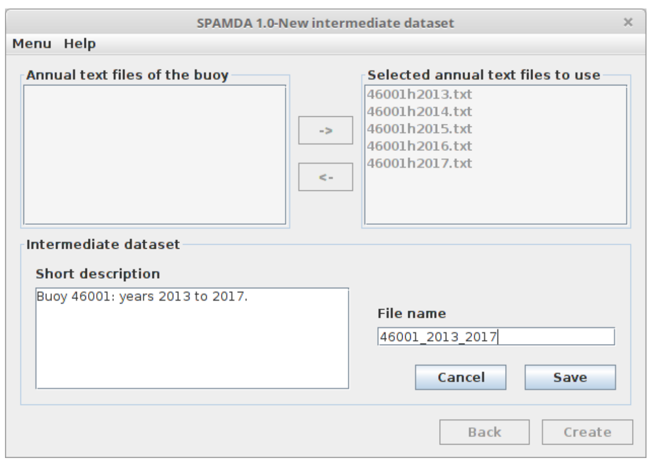


Figure 18: *New intermediate dataset* view: creating the intermediate dataset with 5 annual text files.

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 Figure 17), and then the button *Open* is clicked to jump to the tab *Pre-process* (shown in Figure 19).

In *Pre-process* tab, relevant statistical information about the selected dataset

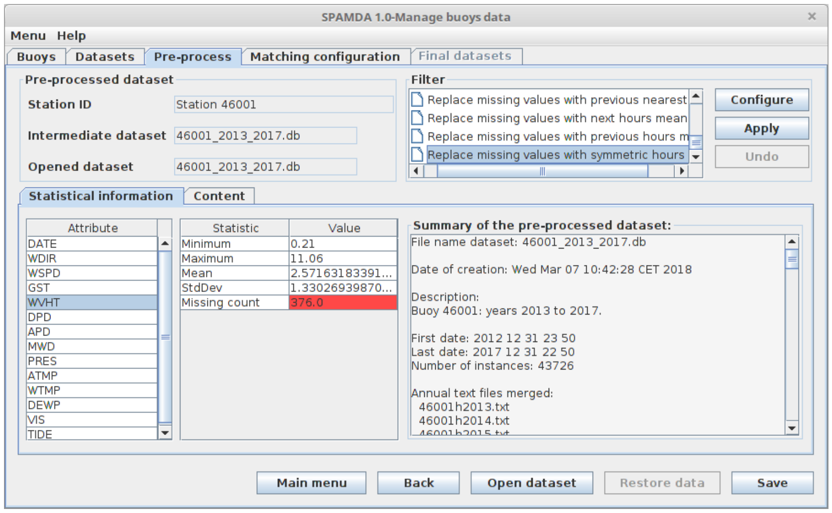


Figure 19: *Pre-process* tab: preprocessing the created intermediate dataset.

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 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 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.

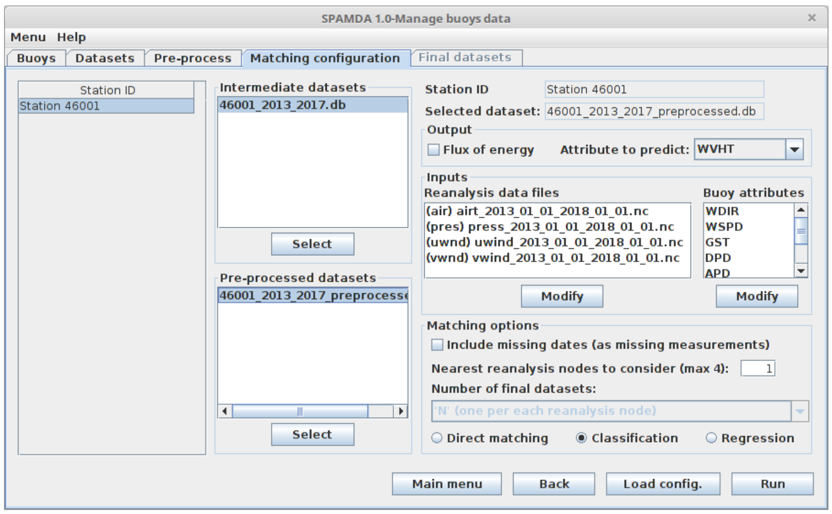


Figure 20: *Matching configuration* tab: parameters for the data integration of the intermediate dataset and the reanalysis files.

The next step is to click on the *Matching configuration* tab, to open the view shown in Figure 20. In this view, the researcher can customise (or load) the parameters of the matching process according to their needs and select the prediction task (described in Section 3.4) that the final dataset will be used for. In this example, 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.

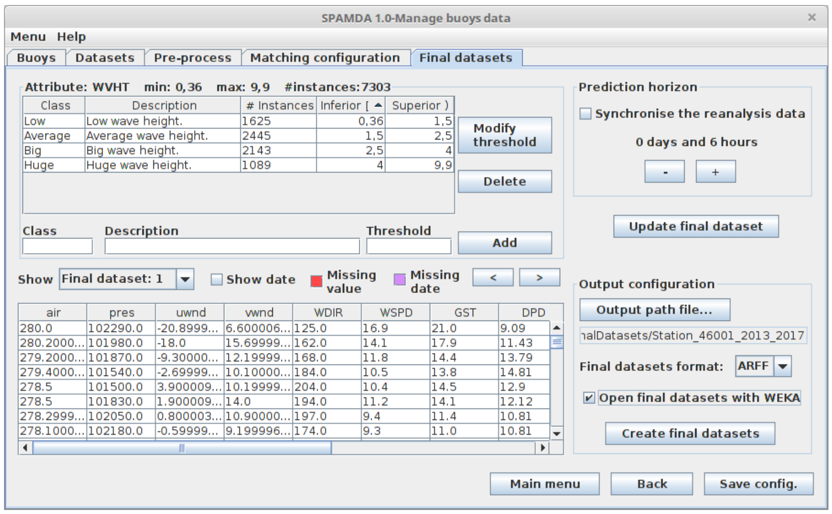


Figure 21: ***Final datasets*** tab: content of the final dataset created after data integration and discretisation of the output variable in 4 classes.

After configuring the matching process, the researcher can click on the *Run* button to jump to the view shown in Figure 21 and proceed to define the final dataset structure according to the selected prediction task. Given that, in the previous window, *Classification* was selected, the researcher can now add, modify or delete the thresholds (usually defined by an expert) for discretising the output variable. 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, 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. 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. For this case study, the following configuration was applied:

- Thresholds: see Table 2.
- Prediction horizon: 6 hours
- Synchronisation: Disabled

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. Obtaining models with ML algorithms

Now, it will be described how to obtain wave classification models using the final dataset previously created with SPAMDA. The modelling will be performed using WEKA as ML tool, which can be opened through SPAMDA, as shown in Fig. 22. Nevertheless, as mentioned above, the researcher can create the final dataset in CSV format in order to use other ML tools, such as KEEL, Python or R, among others.

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

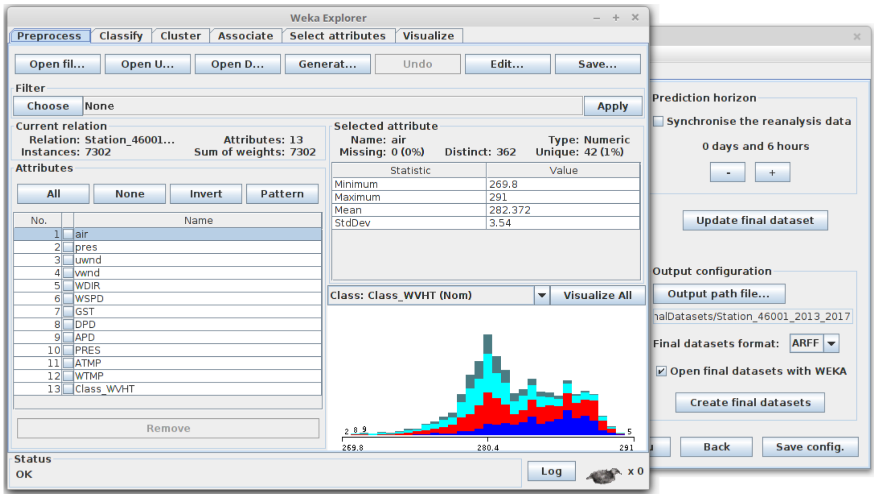


Figure 22: Final dataset opened with the environment Explorer of WEKA.

instances). Previously to the learning phase, the attributes are normalised to avoid that some attributes could be more important than others because their values could be on a larger scale.

The classification algorithms that will be considered for the wave modelling are Logistic Regression [49], C4.5 [50], Random Forest [51], Support Vector Machine [52] and Multilayer Perceptron [53], which will be applied with the default values of the parameters configured in WEKA. Given that Logistic Regression and C4.5 algorithms are deterministic, it is only necessary to launch one execution 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.

As can be seen, Random Forest and Multilayer Perceptron algorithms have achieved similar accuracy, but the performance of the latter is slightly better.

Although it has been performed a first classification approach, to illustrate the use of datasets created with SPAMDA for environmental modelling employing ML algorithms, both models have obtained a good performance despite the problem tackled is to predict wave height six hours in advance.

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**

4.3. Important remarks

In this section, it has been described how to use SPAMDA to create a final dataset with the aim of classifying waves. However, using the same data described in Section 4.1 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 prediction can be addressed by changing the time horizon, wave 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 the researchers from performing repetitive tasks but also prevents them from making possible errors. In this way, the 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. These datasets, that will be ready to be used as input for ML techniques in prediction tasks (classification or regression), are easily obtained by means of the selection of different input parameters, such as predictive and objective variables, output discretisation or prediction horizon. As a result, the 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 ML technique to use), it avoids errors and reduces the time needed. In this way, the researchers will be able to have more in-depth analysis, which could result in more complete conclusions about the issue under study.

In order to improve SPAMDA, some future work could be focused on new functional modules for managing meteorological data of different formats [54], 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 [55]. 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*) [56]. 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.

Acknowledgments

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. We also thank to NVIDIA Corporation for the transfer of computational resources for research works.

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.

Additional material

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

References

[1] M. S. Anis, B. Jamil, M. A. Ansari, E. Bellos, 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 31 (2019) 179–198. doi:10.1016/j.seta.2018.12.009.

[2] V. Laface, F. Arena, C. G. Soares, Directional analysis of sea storms, Ocean Engineering 107 (Supplement C) (2015) 45–53. doi:10.1016/j.oceaneng.2015.07.027.

[3] O. Dahhani, A. El-Jouni, I. Boumhidi, Assessment and control of wind turbine by support vector machines, Sustainable Energy Technologies and Assessments 27 (2018) 167–179. doi:10.1016/j.seta.2018.04.006.

[4] M. Dorado-Moreno, L. Cornejo-Bueno, P. Gutiérrez, L. Prieto, C. Hervás-Martínez, S. Salcedo-Sanz, Robust estimation of wind power ramp events with reservoir computing, Renewable Energy 111 (2017) 428–437. doi:10.1016/j.renene.2017.04.016.

[5] O. Reyes-Mendoza, O. Alvarez-Silva, X. Chiappa-Carrara, C. Enriquez, Variability of the thermohaline structure of a coastal hypersaline lagoon and the implications for salinity gradient energy harvesting, Sustainable Energy Technologies and Assessments 38 (2020) 100645. doi:10.1016/j.seta.2020.100645.

[6] H. L. Fuchs, G. P. Gerbi, Seascape-level variation in turbulence- and wave-generated hydrodynamic signals experienced by plankton, Progress in Oceanography 141 (Supplement C) (2016) 109–129. doi:10.1016/j.pocean.2015.12.010.

[7] V. d. P. R. da Silva, R. A. e Silva, E. P. Cavalcanti, C. C. Braga, P. V. de Azevedo, V. P. Singh, E. R. R. Pereira, Trends in solar radiation in NCEP/NCAR database and measurements in northeastern Brazil, Solar Energy 84 (10) (2010) 1852–1862. doi:10.1016/j.solener.2010.07.011.

[8] B. Gouldby, F. J. Méndez, Y. Guanche, A. Rueda, R. Mínguez, A methodology for deriving extreme nearshore sea conditions for structural design and flood risk analysis, Coastal Engineering 88 (Supplement C) (2014) 15–26. doi:10.1016/j.coastaleng.2014.01.012.

[9] C. E. C. Nogueira, M. L. Vidotto, F. Toniazzo, G. Debastiani, Software for designing solar water heating systems, Renewable and Sustainable Energy Reviews 58 (Supplement C) (2016) 361–375. doi:10.1016/j.rser.2015.12.346.

[10] C. K. Lo, Y. S. Lim, F. A. Rahman, New integrated simulation tool for the optimum design of bifacial solar panel with reflectors on a specific site, Renewable Energy 81 (Supplement C) (2015) 293–307. doi:10.1016/j.renene.2015.03.047.

[11] T. H. Nguyen, A. Prinz, T. Friisø, R. Nossum, I. Tyapin, A framework for data integration of offshore wind farms, Renewable Energy 60 (Supplement C) (2013) 150–161. doi:10.1016/j.renene.2013.05.002.

[12] S. H. Shehadeh, A. Moh'd, H. H. Aly, M. El-Hawary, An intelligent load management application for solar boiler system, Sustainable Energy Technologies and Assessments 38 (2020) 100644. doi:10.1016/j.seta.2020.100644.

[13] D. A. Garcia, D. Bruschi, A risk assessment tool for improving safety standards and emergency management in Italian onshore wind farms, Sustainable Energy Technologies and Assessments 18 (2016) 48–58. doi:10.1016/j.seta.2016.09.009.

[14] A. L. A. Raabe, A. H. d. F. Klein, M. González, R. Medina, MEPBAY and SMC: Software tools to support different operational levels of headland-bay beach in coastal engineering projects, Coastal Engineering 57 (2) (2010) 213–226, hydrodynamics and Applications of Headland-Bay Beaches. doi:10.1016/j.coastaleng.2009.10.008.

[15] S. Motahhir, A. E. Hammoumi, A. E. Ghzizal, A. Derouich, Open hardware/software test bench for solar tracker with virtual instrumentation, Sustainable Energy Technologies and Assessments 31 (2019) 9–16. doi:10.1016/j.seta.2018.11.003.

[16] M. Zeyringer, B. Fais, I. Keppo, J. Price, The potential of marine energy technologies in the UK – Evaluation from a systems perspective, Renewable Energy 115 (Supplement C) (2018) 1281–1293. doi:10.1016/j.renene.2017.07.092.

[17] M. Bhattacharya, S. A. Churchill, S. R. Paramati, The dynamic impact of renewable energy and institutions on economic output and CO2 emissions across regions, Renewable Energy 111 (Supplement C) (2017) 157–167. doi:10.1016/j.renene.2017.03.102.

[18] M. Brede, B. J. de Vries, The energy transition in a climate-constrained world: Regional vs. global optimization, Environmental Modelling & Software 44 (2013) 44–61, thematic Issue on Innovative Approaches to Global Change Modelling. doi:10.1016/j.envsoft.2012.07.011.

[19] A. F. d. O. Falcão, Wave energy utilization: A review of the technologies, Renewable and Sustainable Energy Reviews 14 (3) (2010) 899–918. doi:10.1016/j.rser.2009.11.003.

[20] S. Oliveira-Pinto, P. Rosa-Santos, F. Taveira-Pinto, Electricity supply to offshore oil and gas platforms from renewable ocean wave energy: Overview and case study analysis, Energy Conversion and Management 186 (2019) 556–569. doi:10.1016/j.enconman.2019.02.050.

[21] L. F. Prieto, G. R. Rodríguez, J. S. Rodríguez, Wave energy to power a desalination plant in the north of Gran Canaria Island: Wave resource, socioeconomic and environmental assessment, Journal of Environmental Management 231 (2019) 546–551. doi:10.1016/j.jenvman.2018.10.071.

[22] M. K. Ochi, Ocean Waves: The Stochastic Approach, Cambridge Ocean Technology Series, Cambridge University Press, 1998. doi:10.1017/CBO9780511529559.

[23] S. Crowley, R. Porter, D. J. Taunton, P. A. Wilson, Modelling of the WITT wave energy converter, Renewable Energy 115 (Supplement C) (2018) 159–174. doi:10.1016/j.renene.2017.08.004.

[24] O. Abdelkhalik, R. Robinett, S. Zou, G. Bacelli, R. Coe, D. Bull, D. Wilson, U. Korde, On the control design of wave energy converters with wave prediction, Journal of Ocean Engineering and Marine Energy 2 (4) (2016) 473–483. doi:10.1007/s40722-016-0048-4.

[25] J. V. Ringwood, G. Bacelli, F. Fusco, Energy-Maximizing Control of Wave-Energy Converters: The Development of Control System Technology to

Optimize Their Operation, IEEE Control Systems 34 (5) (2014) 30–55. doi:10.1109/MCS.2014.2333253.

[26] K. Chatziioannou, V. Katsardi, A. Koukouselis, E. Mistakidis, The effect of nonlinear wave-structure and soil-structure interactions in the design of an offshore structure, Marine Structures 52 (Supplement C) (2017) 126–152. doi:10.1016/j.marstruc.2016.11.003.

[27] Y. Dalgic, I. Lazakis, I. Dinwoodie, D. McMillan, M. Revie, Advanced logistics planning for offshore wind farm operation and maintenance activities, Ocean Engineering 101 (Supplement C) (2015) 211–226. doi:10.1016/j.oceaneng.2015.04.040.

[28] D. P. Callaghan, T. E. Baldock, B. Shabani, P. J. Mumby, Communicating physics-based wave model predictions of coral reefs using Bayesian belief networks, Environmental Modelling & Software 108 (2018) 123–132. doi:10.1016/j.envsoft.2018.07.021.

[29] E. Alpaydin, Introduction to Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2004.

[30] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[31] S. Delfani, M. Esmaeili, M. Karami, Application of artificial neural network for performance prediction of a nanofluid-based direct absorption solar collector, Sustainable Energy Technologies and Assessments 36 (2019) 100559. doi:10.1016/j.seta.2019.100559.

[32] Nasruddin, Sholahudin, P. Satrio, T. M. I. Mahlia, N. Giannetti, K. Saito, Optimization of HVAC system energy consumption in a building using artificial neural network and multi-objective genetic algorithm, Sustainable Energy Technologies and Assessments 35 (2019) 48–57. doi:10.1016/j.seta.2019.06.002.

[33] Y. Guo, J. Wang, H. Chen, G. Li, J. Liu, C. Xu, R. Huang, Y. Huang, Machine learning-based thermal response time ahead energy demand prediction for building heating systems, Applied Energy 221 (2018) 16–27. doi:10.1016/j.apenergy.2018.03.125.

[34] M. A. H. Eibe Frank, I. H. Witten, The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques (2016).

[35] A. M. Durán-Rosal, J. C. Fernández, P. A. Gutiérrez, C. Hervás-Martínez, Detection and prediction of segments containing extreme significant wave heights, Ocean Engineering 142 (Supplement C) (2017) 268–279. doi:10.1016/j.oceaneng.2017.07.009.

[36] N. K. kumar, R. Savitha, A. A. Mamun, Regional ocean wave height prediction using sequential learning neural networks, Ocean Engineering 129 (2017) 605–612. doi:10.1016/j.oceaneng.2016.10.033.

[37] L. Johansson, V. Epitropou, K. Karatzas, A. Karppinen, L. Wanner, S. Vrochidis, A. Bassoukos, J. Kukkonen, I. Kompatsiaris, Fusion of meteorological and air quality data extracted from the web for personalized environmental information services, Environmental Modelling & Software 64 (2015) 143–155. doi:10.1016/j.envsoft.2014.11.021.

[38] J. C. Fernández, S. Salcedo-Sanz, P. A. Gutiérrez, E. Alexandre, C. Hervás-Martínez, Significant wave height and energy flux range forecast with machine learning classifiers, Engineering Applications of Artificial Intelligence 43 (Supplement C) (2015) 44–53. doi:10.1016/j.engappai.2015.03.012.

[39] J. Adams, S. Flora, Correlating seabird movements with ocean winds: linking satellite telemetry with ocean scatterometry, Marine Biology 157 (4) (2010) 915–929. doi:10.1007/s00227-009-1367-y.

[40] National Data Buoy Center, National Oceanic and Atmospheric Administration of the USA (NOAA), http://www.ndbc.noaa.gov/, (Accessed 19 November 2018).

[41] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, D. Joseph, The NCEP/NCAR 40-Year Reanalysis Project, Bulletin of the American Meteorological Society 77 (3) (1996) 437–471. doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

[42] R. Kistler, W. Collins, S. Saha, G. White, J. Woollen, E. Kalnay, M. Chelliah, W. Ebisuzaki, M. Kanamitsu, V. Kousky, H. van den Dool, R. Jenne, M. Fiorino, The NCEP–NCAR 50–Year Reanalysis: Monthly Means CD–ROM and Documentation, Bulletin of the American Meteorological Society 82 (2) (2001) 247–267. doi:10.1175/1520-0477(2001)082<0247:TNNYRM>2.3.CO;2.

[43] The WEKA Data Mining Software: Attribute-Relation File Format (ARFF), https://www.cs.waikato.ac.nz/ml/weka/arff.html, (Accessed 18 December 2018).

[44] National Data Buoy Center, NDBC - Historical NDBC Data, http://www.ndbc.noaa.gov/historical_data.shtml, (Accessed 15 January 2019).

[45] National Data Buoy Center, NDBC - Important NDBC Web Site Changes, http://www.ndbc.noaa.gov/mods.shtml, (Accessed 15 January 2019).

[46] 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).

[47] Unidata, Network Common Data Form (NetCDF) version 4.6.10 [software].

Boulder, CO: UCAR/Unidata., https://doi.org/10.5065/D6H70CW6 (2017).

[48] M. J. de Smith, M. F. Goodchild, P. A. Longley, Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools, 3rd Edition, Matador, 2009.

[49] D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant, Applied logistic regression, Vol. 398, John Wiley & Sons, 2013.

[50] J. R. Quinlan, C4. 5: programs for machine learning, Elsevier, 2014.

[51] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[52] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[53] S. Haykin, Neural networks: a comprehensive foundation, Prentice Hall PTR, 1994.

[54] National Data Buoy Center, NDBC - Measurement Descriptions and Units, https://www.ndbc.noaa.gov/measdes.shtml, (Accessed 15 January 2019).

[55] A. M. Durán-Rosal, C. Hervás-Martínez, A. J. Tallón-Ballesteros, A. C. Martínez-Estudillo, S. Salcedo-Sanz, Massive missing data reconstruction in ocean buoys with evolutionary product unit neural networks, Ocean Engineering 117 (2016) 292–301, jCR(2016): 1.894 Position: 2/14 (Q1) Category: ENGINEERING, MARINE. doi:10.1016/j.oceaneng.2016.03.053.

[56] J. Alcalá-Fdez, L. Sánchez, S. García, M. J. del Jesús, S. Ventura, J. M. G. i Guiu, J. M. Otero, C. Romero, J. Bacardit, V. M. R. Santos, J. C. Fernández, F. Herrera, KEEL: a software tool to assess evolutionary algorithms for data mining problems, Soft Comput. 13 (2009) 307–318. doi:10.1007/s00500-008-0323-y.

Appendix A. Managing 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. 23:

- 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 the 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. 24 shows the status of the creation of an intermediate dataset with the information of Fig 23. 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

```
       #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 23: A fragment of an annual text file with different missing value examples.

[54] 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, the 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.

SPAMDA takes into account this casuistry when carrying out the matching process. An example is given in Fig. 25. 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

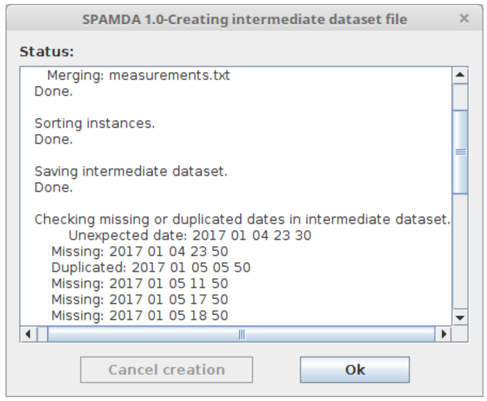


Figure 24: Status of the creation of the intermediate dataset for the example of Fig 23.

the selection made by the researchers in the parameter *Include missing dates*, this instance will be included in the final dataset (with missing values for buoy variables) or not.

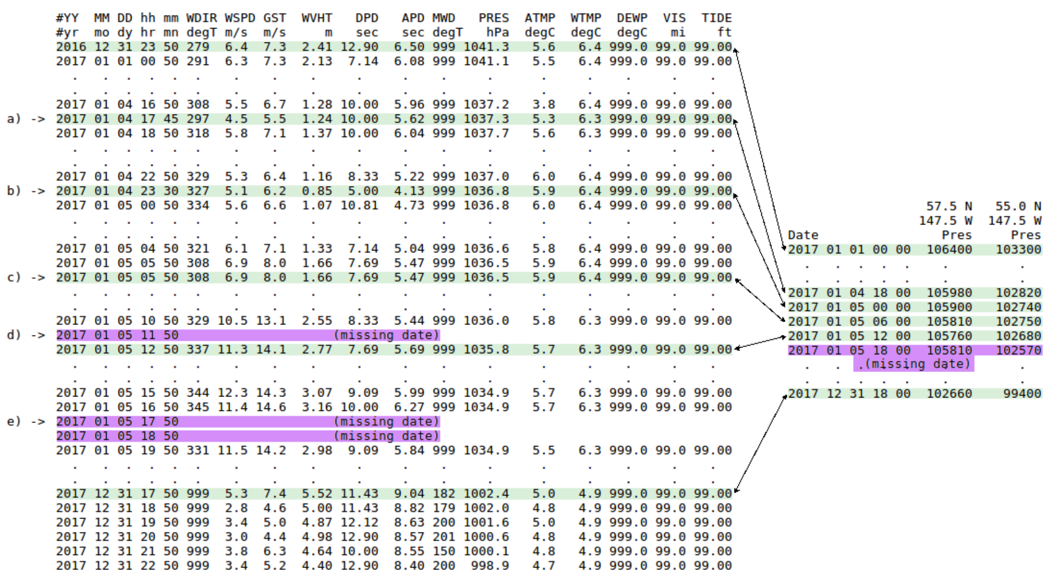


Figure 25: Matching the measurements (left) and the reanalysis data (right).

University of Cordoba.
Department of Computer Science and Numerical Analysis.
Rabanales Campus, "Albert Einstein" building, 3rd floor.
14071 – Córdoba (Spain).
Ph. +34 957218349 Fax +34 957218630

Sustainable Energy Technologies and Assessments

Cordoba, Spain, April 21, 2020.

Dear Editor,

Please find attached a manuscript, entitled "*A new software tool to facilitate environmental modelling procedures based on meteorological data management*" submitted for possible publication in Sustainable Energy Technologies and Assessments.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Best regards,

Antonio M. Gómez-Orellana, Juan C. Fernández, Manuel Dorado-Moreno, Pedro A. Gutiérrez, César Hervás-Martínez

Department of Computer Science and Numerical Analysis, University of Cordoba, 14071, Córdoba, Spain.
email: {am.gomez@uco.es, jfcaballero@uco.es, manuel.dorado@uco.es, pagutierrez@uco.es, chervas@uco.es}

University of Cordoba.
Department of Computer Science and Numerical Analysis.
Rabanales Campus, "Albert Einstein" building, 3rd floor.
14071 – Córdoba (Spain).
Ph. +34 957218349 Fax +34 957218630

Sustainable Energy Technologies and Assessments

A new software tool to facilitate environmental modelling procedures based on meteorological data management

Open source tool:

SPAMDA is a software tool that has been developed in the AYRNA research group (https://www.uco.es/grupos/ayrna/index.php/en) of the University of Cordoba, Spain.

In the distribution submitted to the reviewers the source code is not available, if the paper is considered suitable for publication in this journal, it will be made available to the scientific community as free software under the terms of the **GNU General Public License** as published by the Free Software Foundation, **either version 3** of the License (**GPL3**), or any later version.

The source code and the software tool will be available at: https://github.com/ayrna (as it is mentioned in Section "Additional material" in the manuscript.)

Instructions for downloading and installing SPAMDA Software Tool.

Step 1: Download the file SPAMDA.zip from the following url:

https://drive.google.com/file/d/1p1j7lokYYwWlSSdlNV8csSLU52LqTtyL/view?usp=sharing

or

https://drive.google.com/file/d/10QyTYT7GwDalIFGT0d8XNOBOIKdLc9Zo/view?usp=sharing

Step 2: Create a folder and copy the downloaded file inside it.

Step 3: Unzip the file SPAMDA.zip

Step 4: Go to SPAMDA folder.

Step 5: Read the file README:

"Section 1. System requirements."

"Section 3. Running SPAMDA."