Comparison of feature selection methods using ANNs in MCP-wind speed methods. A case study

José A. Carta [a,*], Pedro Cabrera [a], José M. Matías [b], Fernando Castellano [c]

[a] Department of Mechanical Engineering, University of Las Palmas de Gran Canaria, Campus de Tafira s/n, 35017 Las Palmas de Gran Canaria, Canary Islands, Spain
[b] Department of Statistics, University of Vigo, Lagoas Marcosende, 36200 Vigo, Spain
[c] Renewable Energies Department, Canary Islands Institute of Technology (ITC), Playa de Pozo Izquierdo s/n, 35119 Santa Lucía – Las Palmas, Spain

HIGHLIGHTS

- An analysis is carried out of the benefits of feature selection in MCP methods which use ANNs.
- The wrapper approach (WA) generated lower mean errors than the filter approach (FA).
- No significant statistical difference was observed between the WA and the FA in certain cases.
- The FA generated models somewhat simpler and more interpretable than the WA.
- The WA displayed better predictive capacity than the FA, but is more computationally intensive.

ARTICLE INFO

ABSTRACT

Recent studies in the field of renewable energies, and specifically in wind resource prediction, have shown growing interest in proposals for Measure–Correlate–Predict (MCP) methods which simultaneously use data recorded at various reference weather stations. In this context, the use of a high number of reference stations may result in overspecification with its associated negative effects. These include, amongst others, an increase in the estimation error and/or overfitting which could be detrimental to the generalisation capacity of the model when handling new data (prediction).

This paper analyses the benefits of feature selection for use with Artificial Neural Network (ANN) techniques with a multilayer perceptron (MLP) structure when the ANNs are used as MCP methods to predict mean hourly wind speeds at a target site. The features considered in this study were the mean hourly wind speeds and directions recorded in 2003 and 2004 at five weather stations in the Canary Archipelago (Spain).

The two feature selection techniques considered in the analysis were the Correlation Feature Selection (CFS), which is a correlation-based filter approach (FA), and an MLP-based wrapper approach (WA). The metrics used to compare the results were the mean absolute error (MAE), the mean absolute percentage error (MAPE) and the index of agreement (IoA).

Evaluation of the mean errors obtained in the 10-fold cross-validation tests for the year used to represent the short-term wind data period resulted in several conclusions. These included, notably, that the WA gave lower mean errors than the FA in 100% of the cases analysed independently of the metric employed. However, the FA resulted in a significant reduction in computational load and considerable enhancement of model interpretability. When very good correlation coefficients were obtained between the target and reference stations, no significant statistical difference was observed at 5% level between the three models (FA, WA and the models constructed with all the variables) in most of the cases analysed.

© 2015 Elsevier Ltd. All rights reserved.

* Corresponding author. Tel.: +34 928 45 14 83; fax: +34 928 45 14 84.
E-mail address: jcarta@dim.ulpgc.es (J.A. Carta).

Nomenclature

A	parameter defined by Eq. (2)
AEMET	State Meteorological Agency of the Ministry of the Environment and Rural and Marine Environs of the Spanish Government (Spanish initials)
ANOVA	analysis of variance
ANNs	artificial neural networks
agl	above ground level
B	parameter defined by Eq. (2)
BH	Benjamini and Hochberg step-up procedure [49]
C	parameter defined by Eq. (2)
CCC	Circular Correlation Coefficient, Eq. (1)
CFS	Correlation Feature Selection
CL	circular-linear correlation coefficient, Eq. (3)
CPU	Central Processing Unit
D, D'	variables that represent wind direction (Degree)
$D1,\ldots,D5$	variables that represent wind directions of weather stations no. 1, ..., 5, respectively
E	parameter defined by Eq. (2)
e	vector which contains the estimated values of the wind speed (m/s) at a target site, Eqs. (5)–(7)
F	parameter defined by Eq. (2)
FA	filter approach
FA-1, FA-2	used to mark the itinerary followed in the use of the filter approach (FA), Fig. 7
FS	full feature set (no feature selection).
G	parameter defined by Eq. (2)
GNU	recursive acronym which stands for *GNU is Not Unix*
H	parameter defined by Eq. (2)
H_0	null hypothesis, Eq. (8)
H_1	alternative hypothesis, Eq. (8)
IoA	index of agreement, Eq. (7)
ITC	Technological Institute of the Canary Islands (Spanish initials)
MAE	mean absolute error, Eq. (5) (m/s)
MAPE	mean absolute percentage error, Eq. (6) (%)
MCP	Measure–Correlate–Predict
MLP	multilayer perceptron
n	number of data, Eqs. (1), (2), (5), (6) and (7).
n_{cv}	number of errors obtained in 10-fold cross-validation.
N	number of reference stations, Fig. 5
o	vector which contains the observed wind speed values (m/s) at a target site, Eqs. (5)–(7)
$\bar{o}$	Arithmetic mean in m/s of observed wind speed values, Eq. (7)
p-value	Is the estimated probability of rejecting the null hypothesis (H_0) of a study question when that null hypothesis is true [50]
Q	parameter defined by Eq. (2)
q	number of neurons in the hidden layer of the neural network, Fig. 5
r_{cs}	correlation coefficient, Eq. (4)
r_{vc}	correlation coefficient, Eq. (4)
r_{vs}	correlation coefficient, Eq. (4)
S	variable representing wind speed (m/s)
$S1,\ldots,S5$	variables which represent the wind speeds of weather stations number 1, ..., 5, respectively
WA	wrapper approach
WA-1, WA-2	Used to mark the itinerary followed in the use of the wrapper approach (WA), Fig. 7
Weka	Waikato Environment for Knowledge Analysis
WS	weather station
WS-k	weather station identified with the number k

Greek letters

α	level of statistical significance
μ_i, μ_j	population means of models i and j, Eq. (8)

1. Introduction

1.1. Background

The feasibility of implementing a wind energy conversion system at a target site depends fundamentally on the wind regime at the site in question [1–3]. In view of this, and as a consequence of the interannual variability of wind speed, long series of wind data are required [1,4–9] to enable estimation of the mean characteristics of the wind resource over the useful life of a wind project with the least degree of uncertainty possible. If only short series of wind data are available for a particular target site, a common way of at least partially overcoming this drawback is the use of methods which attempt to estimate the mean long-term characteristics of the wind resource at the target site by using long historical wind data series recorded at nearby weather stations (WSs). Among such methods are those widely known as MCPs (Measure–Correlate–Predict) [10].

As reported by Carta et al. [10], a considerable number of MCPs have been proposed in the scientific literature and new techniques continue to be formulated [11,12]. Most of these MCP methods use a single reference WS and employ linear equations [13] or probabilistic equations [14]. The parameters of these models are determined with wind data series recorded in a concurrent short-term time period at the target and reference sites and the equations or functions are then used to estimate the long-term wind data at the target site [10].

However, a considerable number of authors [2,12,15–23] argue that it may be beneficial to use several as opposed to a single reference station, as the use of various stations may be able to capture details of the wind resource at a target site that would otherwise have been missed if only one reference station had been used. A variety of data mining techniques have been proposed in this respect which allow simultaneous use of the information provided by a number of reference stations. Perhaps the most notable of the proposed techniques are those based on biologically-inspired algorithms, the so-called artificial neural networks or ANNs [24,25], and more specifically multilayer perceptron (MLP) neural networks [3,11,12,18–23,26,27]. Lopez et al. [21] and Velázquez et al. [22] highlight the importance of using the wind speeds and directions of the reference stations as features, especially in sites of complex terrain. Though Velázquez et al. [22] reported in their studies that predictive capacity tended to increase with the number of reference stations, they also noted that not all the reference stations that were used improved the results. However, these authors did not contemplate the possibility of reducing the number of features, including only the variables (speed or direction) of each reference station that might really be relevant. Aside from this, Brower [2], in reference to linear regression models, warns that the use of a large number of reference stations, especially if they are strongly correlated, may give rise to model overspecification, which could be detrimental to the generalisation capacity of the model when handling new data (prediction).

1.2. Aim of this paper

As previously stated, a growing trend has been noted in proposals that employ MCP methods, which simultaneously use as reference variables wind speed and direction data series registered at various WSs, for long-term wind speed estimation of a target site. However, despite an exhaustive search of the literature, it appears that no proposals have been made to use, within the general framework of strategies of MCP methods, feature selection techniques. According to the bibliography consulted, MCP methods which propose the use of multiple reference WSs employ all the available features without taking into consideration the pros and cons that may be associated with such an action strategy. Accordingly, one of the aims of the present study is to determine the benefits, if any, of feature selection [28–31] when using neural networks as MCP methods. The idea behind this is to offer users and designers of MCP methods a strategy that provides them with greater information and the possibility of improving the use of such methods. Users of traditional MCP methods, which employ a single reference station, tend to turn to rules of thumb when selecting this reference station [10]. Such rules are based on the degree of linear correlation that exists between the wind speeds of the target station and those of the reference station. However, no rules of thumb have been proposed for the selection of various reference stations. In view of this, the present paper will consider a specific MCP problem which onsists of predicting the mean hourly wind speeds at different target sites using the wind speeds and directions from four reference WSs in the Canary Archipelago (Spain) as features. Given the different levels of correlation observed between the features of the WSs used (see Section 2.1), this study also aims to outline rules of thumb that can serve as guidelines when choosing the most appropriate feature selection technique for use in MCP methods using multiple WSs.

In view of their popularity in this type of problem, the ANNs used in this study are MLP neural networks with a hidden layer and a linear-type output layer in accordance with the continuous nature of the wind speed variable.

ANNs are a group of *machine learning* techniques which could fit into the category of non-parametric statistical techniques given their capacity to approximate any continuous function [32]. In consequence, their flexibility is greater than that of parametric techniques, but interpretability is lower and there is also a higher risk of overfitting.

The essential question that lies behind the objectives of this work is whether MLP algorithms are sufficiently capable by themselves of carrying out the necessary selection of features in an MCP context, or whether it is necessary or at the very least beneficial to use selection techniques to complement the MLP algorithms. Three model types are considered in this study. One involves no feature selection (FS – Full Set) and the other two are commonly used feature selection strategies: a filter approach (FA) based on correlation, the so-called Correlation Feature Selection (CFS) [33–35], and a wrapper approach (WA) based on MLPs [29,33]. This study aims to compare and determine which of the above model types gives the best results of the following metrics: the mean absolute error (MAE), the mean absolute percentage error (MAPE) and a refined version of Willmott's dimensionless index of agreement (IoA) [36]. Importantly, this study aims to offer the strongest possible grounds for its conclusions in order to avoid having to treat purely circumstantial differences between the results of the different methods as being due to important structural mechanisms when, in fact, they were simply caused by the randomness of the samples. Therefore, the comparison in this study will not only involve determination, for the particular sample used, of the differences between the results obtained with the different models according to the various metrics in a 10-fold cross-validation

process [33]. An additional comparison will involve statistical hypothesis testing to determine whether there is a significant statistical difference (at 5% level) between the results obtained with the three strategies considered (FA, WA and FS).

The article is structured as follows: the following section describes the materials, including the data sample, the MLP models used and the feature selection techniques that are compared. The methodology employed for the comparison study is then described. Next, the results that were obtained are presented and analysed and, finally, a description is given of the conclusions drawn from the study.

2. Materials

2.1. Meteorological data used

The meteorological data used in this paper (mean hourly wind speeds and directions) were recorded at five WSs installed on three of the seven major islands that make up the Canary Archipelago (Spain) (Fig. 1).

The wind speeds were captured using rotating cup type anemometers situated on masts at 10 m above ground level (agl) located on the coasts of the islands (Fig. 1). The data series used were recorded during the years 2003 and 2004 and were provided by the Technological Institute of the Canary Islands (Spanish initials: ITC)[1] and by the State Meteorological Agency (Spanish initials: AEMET) of the Ministry of Environment and Rural and Marine Affairs of the Government of Spain.

The codes assigned to each WS are shown in the first column of Table 1. The ITC provided the data series from WS-1 and WS-2 and AEMET the data series from WS-3, WS-4 and WS-5. The geographical coordinates of each station (latitude, longitude and altitude) are also shown in Table 1 along with the annual mean and standard deviation of the wind speed.

The predominant wind regime in the Canary Islands is that of the so-called trade winds, which blow mostly in a NE direction. The wind roses of 2003 for each WS are shown in Fig. 2a and c and of 2004 in Fig. 2b and d.

For ease of interpretation, the scatterplot matrix of the 10 variables of the 2003 data sample has been divided into three figures. Fig. 3 represents the scatterplot matrix of the different wind direction variables, while Fig. 4 shows the scatterplot matrix that corresponds to the different wind speed and wind direction variables. The 2004 scatterplots have very similar structures. The sample therefore comprises eight features or input variables (wind speeds and directions of the four reference stations) for each of the five target stations, with a sampling size of 6738 data for 2003 and 8114 for 2004 for each of the variables.

Table 2 shows the correlation between the mean hourly wind speeds of the different WSs in 2003 and 2004. The correlation was quantified by calculating the Pearson product-moment coefficient of linear correlation (commonly called Pearson's correlation coefficient). Pearson's correlation coefficient, in this case, measures the magnitude or strength of the linear association, as well as the direction (rising or falling, depending respectively on whether the sign is positive or negative), between the recorded wind speeds at the target station and reference station [10]. The correlation coefficients recorded in the two years considered in this study ranged between 0.634 and 0.948. The most notable correlations are between wind speed variables S1 and S4, corresponding to WS-1 and WS-4.
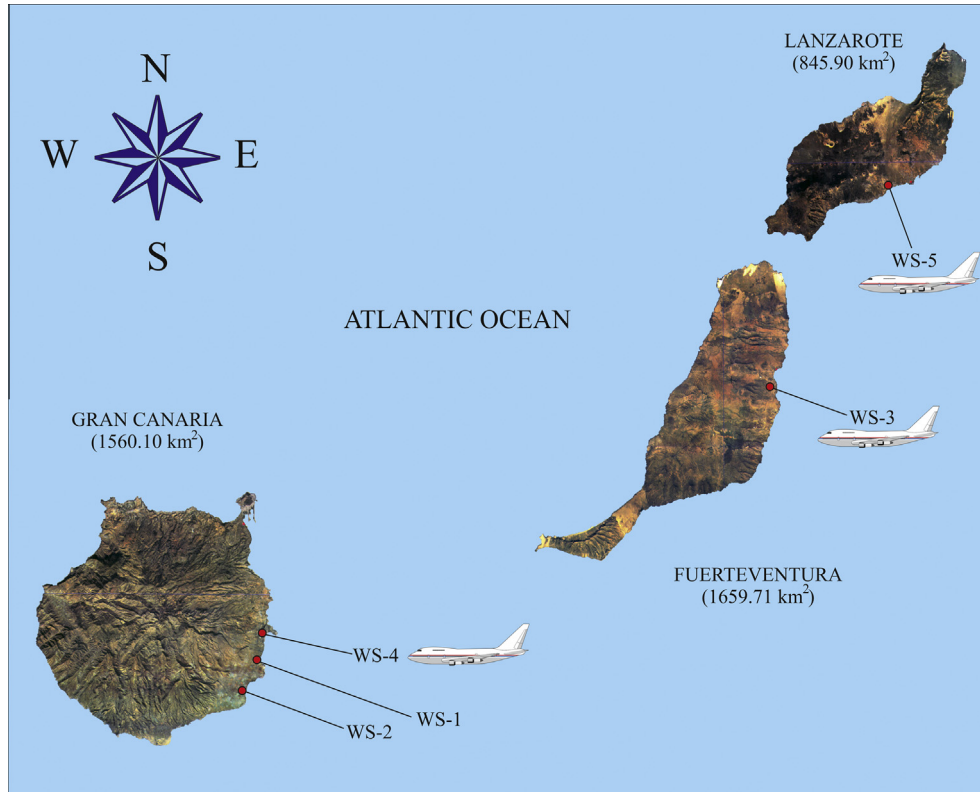

Fig. 1. Location of the weather station used.

Table 1
Weather stations used in the study.

Weather stations	Geographical coordinates			Year 2003		Year 2004	
	Latitude north	Longitude west	Altitude (m)	Annual mean (m/s)	Standard deviation (m/s)	Annual mean (m/s)	Standard deviation (m/s)
WS-1	27°53′24″	15°23′42″	3	6.30	3.30	6.30	3.22
WS-2	27°49′4″	15°25′27″	6	7.37	4.14	7.02	3.86
WS-3	28°27′10″	13°51′54″	29	5.69	2.56	5.65	2.47
WS-4	27°55′44″	15°23′20″	24	7.34	3.67	6.99	3.62
WS-5	28°57′08″	13°36′00″	9	5.83	3.05	5.73	2.83

Table 3 shows the circular-correlation coefficient (CCC) between the mean hourly wind directions of the different stations in 2003 and 2004. The CCC proposed by Fisher and Lee [37,38] to analyse the relationship between two angular variables was used, and is defined in Eq. (1).

$$\text{CCC} = \frac{4(AB - CQ)}{\left[(n^2 - E^2 - F^2)(n^2 - G^2 - H^2)\right]^{1/2}} \quad (1)$$

where n is the number of data and A, B, C, Q, E, F, G and H are given by Eq. (2)

$$\left.\begin{array}{ll} A = \sum_{i=1}^{n} \cos(D_i)\cos(D_i'); & B = \sum_{i=1}^{n} \sin(D_i)\sin(D_i') \\ C = \sum_{i=1}^{n} \cos(D_i)\sin(D_i'); & Q = \sum_{i=1}^{n} \sin(D_i)\cos(D_i') \\ E = \sum_{i=1}^{n} \cos(2D_i) & F = \sum_{i=1}^{n} \sin(2D_i) \\ G = \sum_{i=1}^{n} \cos(2D_i') & H = \sum_{i=1}^{n} \sin(2D_i') \end{array}\right\} \quad (2)$$

D_i and D' represent the n values of the wind direction variables D and D', recorded at two sites.

The CCC takes values ranging between −1 and 1, with zero indicating that no relationship exists between the directions and with 1 and −1 representing the strongest association possible. The hypothesis of non-existence of linear association is rejected if CCC differs greatly from zero.

The range of CCC determined in the two years varies between 0.2 and 0.794. The linear character of the relationships that this coefficient detects can be seen in Fig. 3. The most notable relationships are between the wind direction variables D2 and D4, corresponding to WS-2 and WS-4.

Table 4 shows the circular-linear (CL) correlation coefficient between the mean hourly wind directions (D) and speeds (S) of the different stations in 2003 and 2004. The CL proposed by Mardia [39] and Johnson and Wehrley [40] was used and is formulated in Eq. (3).

$$\text{CL} = \left[\frac{r_{vc}^2 + r_{vs}^2 - 2r_{vc}r_{vs}r_{cs}}{1 - r_{cs}^2}\right]^{1/2} \quad (3)$$

where

$$r_{vc} = \text{corr}(S, \cos D); \quad r_{vs} = \text{corr}(S, \sin D); \quad r_{cs} = \text{corr}(\cos D, \sin D) \quad (4)$$

This coefficient does not reflect the degree of linear relationship between S and D, but rather the degree of a sinusoidal-type non-linear and periodic relationship between D and S (the linear correlation that it reflects is the one between S and the joint variable
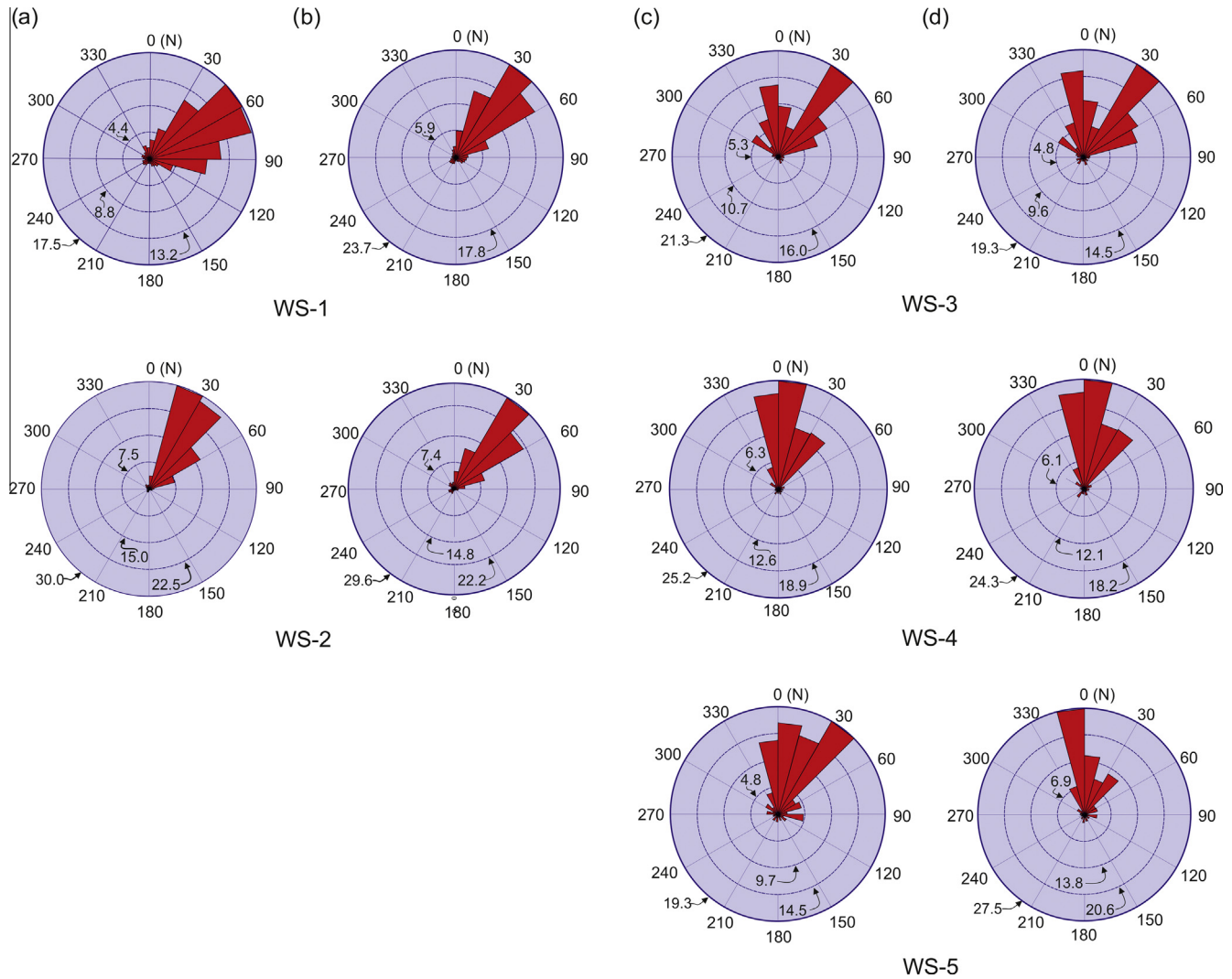
Fig. 2. Wind roses for each of the stations. The (a) and (c) columns represent the wind roses for 2003, and the (b) and (d) columns for 2004.

(sinD, cosD)). This coefficient takes values between 0 and 1, which is to say it does not indicate negative correlation.

The CL correlation coefficients obtained in the two years considered in this study are in the range 0.252–0.521. The non-linear periodic structure, typical of the sinusoidal relationships that this coefficient detects, of the relationships between the wind speed and wind direction variables can be seen in Fig. 4.

2.2. Techniques used in the comparison

2.2.1. Architecture of the ANNs used

Given the dominance in the scientific literature concerned with renewable energies of proposals for biologically-inspired algorithms [3,11,18–23,26,27] to estimate long-term wind speed characteristics using multiple reference stations, the models employed in this paper for such estimation were configured using ANNs.

More specifically, MLP neural networks were used. These possess a multilayer feedforward structure with a single hidden layer comprised of hidden units (neurons) with sigmoid activation functions, and a single neuron output layer with linear activation, as corresponds to the continuous nature of the wind speed at the target site. Moreover, this architecture has shown its capacity to satisfactorily approximate any continuous transformation [32–34] and its use has been proposed by several authors [18,20–22]. This

property, known as the universal approximation property, ensures that an MLP network with a hidden layer can reproduce the structure of relationships that exist between the input variables and the target variable, supposing that the relationships are continuous. In consequence, it is not in theory necessary to use MLP networks with more than one hidden layer. In practice, moreover, networks with more than one hidden layer often make model training and selection processes much more difficult without offering any additional benefits. This is due to the higher degree of non-linearity which is introduced into the target function of the optimisation algorithm and to the addition of an extra parameter for each new hidden layer (the number of neurons in that layer) in what is already a complex model selection process. In this context, only the presence of correlations with a high degree of non-linearity between the input variables and the target variable, together with a large sample size that can reveal these complex relationships, would make such multiple architecture advisable. However, these are not the circumstances of our particular case in view of the moderate complexity of the relationships shown in Fig. 3. A more in-depth discussion of this question which is not limited to the set of techniques used in the present study can be found in [41].

In this paper, the input neurons of the models are fed with series of wind speeds (S) and directions (D) recorded at various reference stations (Fig. 5). If N reference stations are used, then the
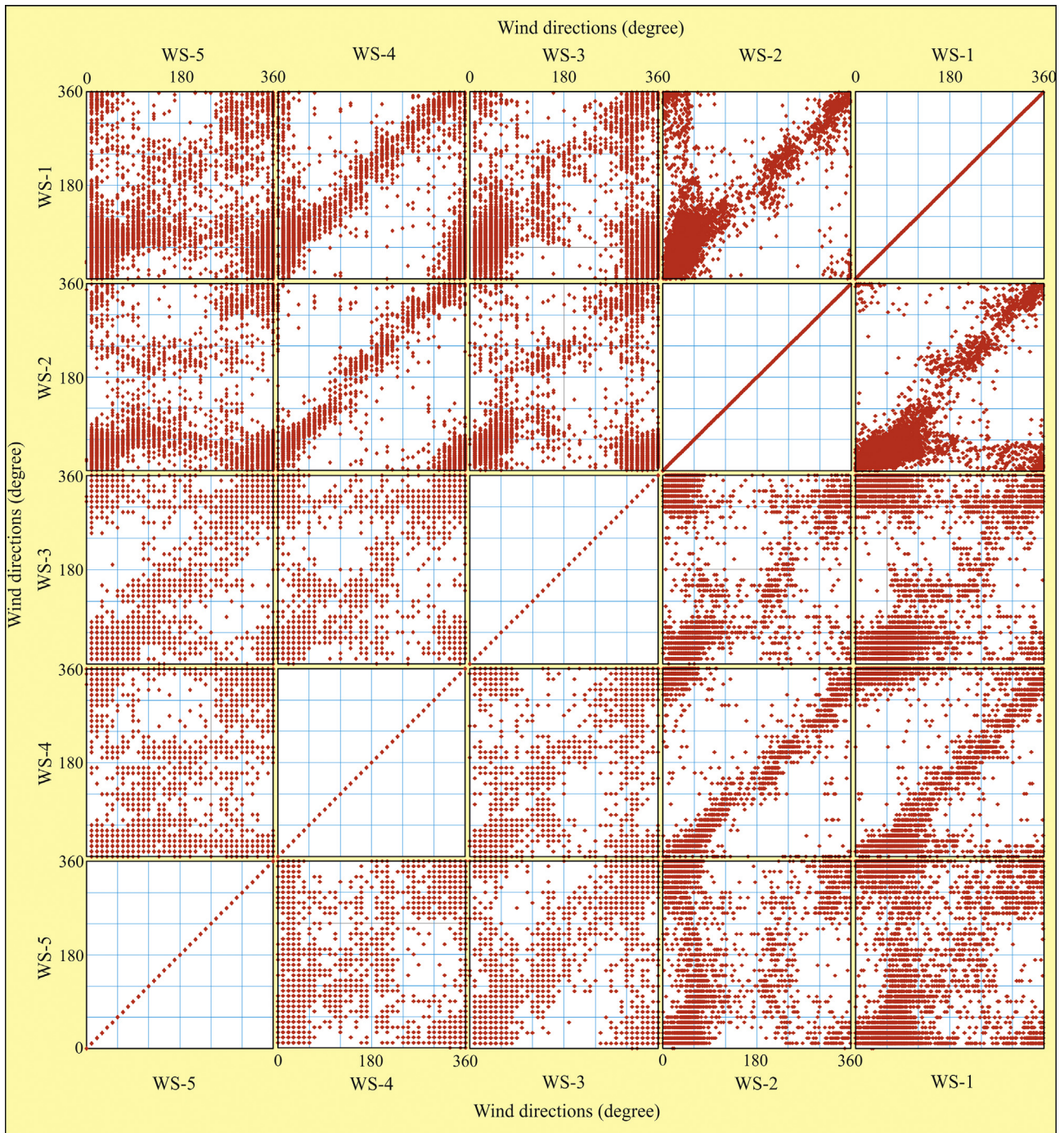
Fig. 3. Scatterplot matrix of the wind directions recorded during 2003 at the five weather stations used in the study.

maximum number of features (input variables) is 2N and the number of neurons of the output layer is one (the target site wind speed).

2.2.2. Selection and training algorithm of the ANN model

The algorithm used in the training of the model of all the MLP neural networks used in this work is the backpropagation algorithm, which aims to optimise estimation of the parameters of the networks. More specifically, we used the *MultilayerPerceptron* algorithm implemented in Weka (Waikato Environment for Knowledge Analysis), free software available under the GNU General Public License [33] and developed by Waikato University (New Zealand).

The 10-fold cross-validation technique was used to estimate the error of the estimation models. This technique is widely used and accepted in the data mining community [33].

The cross-validation mechanism is schematically represented in Fig. 6. This method consists of dividing the data, once randomly ordered, into $n_{cv} = 10$ discrete subsets of similar size. Model learning then takes place using 9 subsets, with the mean prediction error being determined with the different metrics in the remaining subset. The procedure is repeated 10 times, omitting from the training a different subset each time. The final mean prediction error with each metric is obtained by calculating the arithmetic mean of the 10 mean prediction errors obtained in each of the 10 subsets that were successively excluded from the training. The
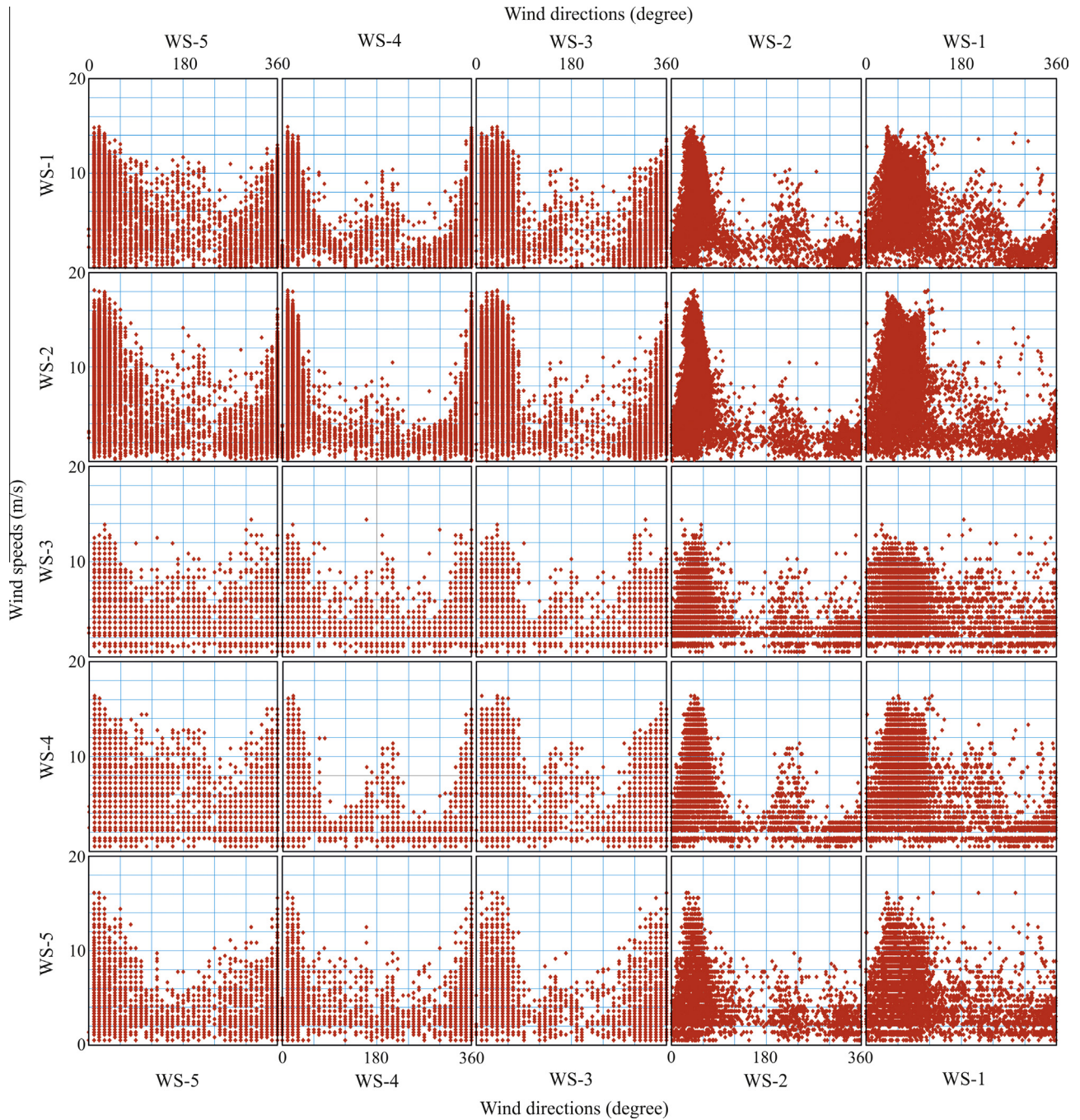
Fig. 4. Scatterplot matrix of the wind speeds and wind directions recorded during 2003 at the five weather stations used in the study.

Table 2
Linear correlation coefficients between the wind speeds of the five anemometer weather stations.

	Year 2003					Year 2004				
	WS-1	WS-2	WS-3	WS-4	WS-5	WS-1	WS-2	WS-3	WS-4	WS-5
WS-1	1.000	0.878	0.698	0.948	0.756	1.000	0.876	0.669	0.930	0.700
WS-2	0.878	1.000	0.644	0.832	0.725	0.876	1.000	0.634	0.828	0.647
WS-3	0.698	0.644	1.000	0.709	0.697	0.669	0.634	1.000	0.667	0.686
WS-4	0.948	0.832	0.709	1.000	0.741	0.930	0.828	0.667	1.000	0.701
WS-5	0.756	0.725	0.697	0.741	1.000	0.700	0.647	0.686	0.701	1.000

variance of the 10 mean prediction errors in the 10 cross-validation groups gives an idea of the variability of these partial means and, therefore, of model performance stability when handling new data.

The number of neurons in the hidden layer (q in Fig. 5) must be specified by the designer when configuring the structure of the network in Weka. Various heuristic rules have been proposed in

Table 3
Circular-correlation coefficients between the wind directions of the five anemometer weather stations.

	Year 2003					Year 2004				
	WS-1	WS-2	WS-3	WS-4	WS-5	WS-1	WS-2	WS-3	WS-4	WS-5
WS-1	1.000	0.575	0.344	0.560	0.200	1.000	0.677	0.405	0.633	0.240
WS-2	0.575	1.000	0.484	0.794	0.335	0.677	1.000	0.461	0.734	0.272
WS-3	0.344	0.484	1.000	0.455	0.469	0.405	0.461	1.000	0.435	0.502
WS-4	0.560	0.794	0.455	1.000	0.304	0.633	0.734	0.435	1.000	0.246
WS-5	0.200	0.335	0.469	0.304	1.000	0.240	0.272	0.502	0.246	1.000

Table 4
Circular-linear correlation coefficients between the wind directions (horizontal) and wind speeds (vertical) of the five anemometer weather stations.

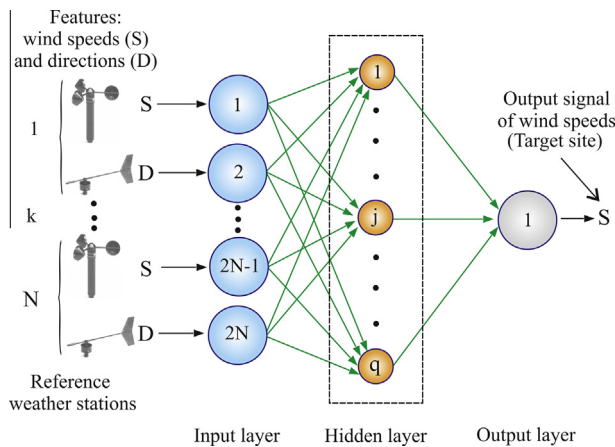| W/D	Year 2003					Year 2004				
	WS-1	WS-2	WS-3	WS-4	WS-5	WS-1	WS-2	WS-3	WS-4	WS-5
WS-1	0.521	0.492	0.520	0.500	0.415	0.411	0.424	0.418	0.399	0.305
WS-2	0.513	0.489	0.576	0.492	0.485	0.435	0.445	0.490	0.437	0.381
WS-3	0.353	0.430	0.422	0.405	0.358	0.270	0.473	0.317	0.291	0.249
WS-4	0.513	0.508	0.505	0.524	0.400	0.412	0.444	0.426	0.433	0.340
WS-5	0.351	0.383	0.412	0.377	0.403	0.252	0.288	0.267	0.265	0.302

Fig. 5. Schematic diagram of an ANN with $2N$ wind speed (S) and angular wind direction (D) input signals of N reference stations and one wind speed (S) output signal of the target station.

the literature [42] that express the number of neurons in the hidden layer as a function of the number of neurons in the input and output layer. The *MultilayerPerceptron* procedure of Weka suggests predefined values for this number of hidden units based on certain general criteria that have shown good behaviour in different data sets. However, given the relevance that this parameter has in the degree of fit and the predictive capacity of the neural networks, it was decided in this work to estimate this parameter by adapting it to the particularities of the sample of wind data used in the study.

Given a set of input variables (features), the number of hidden layer neurons is estimated through the following procedure: the process is initiated by using a minimum number of hidden neurons and this number is increased provided there is a statistically significant improvement (significance level of 5%) in the errors obtained. A nonparametric permutation test for paired data is used for the comparison of the errors of two models [43–46]. Finally, with a view to obtaining the definitive MLP neural network, this optimum number of hidden units is used in the training of an MLP neural network using all the data of the training-validation sample.

2.2.3. Feature selection techniques

The feature selection techniques employed in this work form part of the so-called supervised methods [30], and more specifically to those known as filter methods and wrapper methods. In

general, the basic unit to be evaluated in the feature selection process is not each variable separately but rather a subset of variables, since a variable may not be beneficial in isolation but may be so when accompanied by others. Even redundant variables may be beneficial if participating in a same subset [47].

Filter methods are pre-process techniques and the criterion used to establish the appropriate subset of characteristics is based on measures of quality that are determined, without the intervention of the prediction model, only from an analysis of the data (features and target variable) and the relationships that exist between those data. Different measures of correlation have been proposed depending on the discrete and/or continuous nature of the variables, but generally the linear correlation coefficient is used between continuous variables [35], as is the case in the present study.

The feature evaluation method used in this work for feature selection was the *CfsSubsetEval* [33] and was chosen from among the different methods incorporated in the Weka tool. This type of evaluator requires a search method for feature selection and the one chosen for this study was *ExhaustiveSearch* [33]. *CfsSubsetEval* evaluates the predictive capacity of each individual feature and the degree of redundancy between them, preferring sets of features that are highly correlated with the target variable but with low correlation between themselves. *ExhaustiveSearch* searches through the feature subset space, starting with the empty set, and offers the best subset found.

Wrapper methods [29] select the subsets of variables that produce the highest predictive capacity using a previously chosen learning algorithm. This algorithm may be the same as or different to the algorithm chosen for the problem modelling, though most authors defend the first option so the algorithm can take maximum advantage of the chosen variables. In this work, we use the MLP neural networks as integrated algorithm in the wrapper and as MCP algorithm in the hope that the selection will be more effective if it adapts to the algorithm that the best subset will finally use. The feature evaluation method used for feature selection was the *WrapperSubsetEval*, available in Weka [33] and the search method for feature selection was the previously mentioned *ExhaustiveSearch*.

2.2.4. Metrics used to evaluate numerical prediction of the models

The metrics used in this paper to evaluate the numerical prediction of the proposed models were the MAE, the MAPE and the IoA [36].

MAE is defined by Eq. (5), where the n estimated values are represented by the letter "e" and the n observed values by the letter
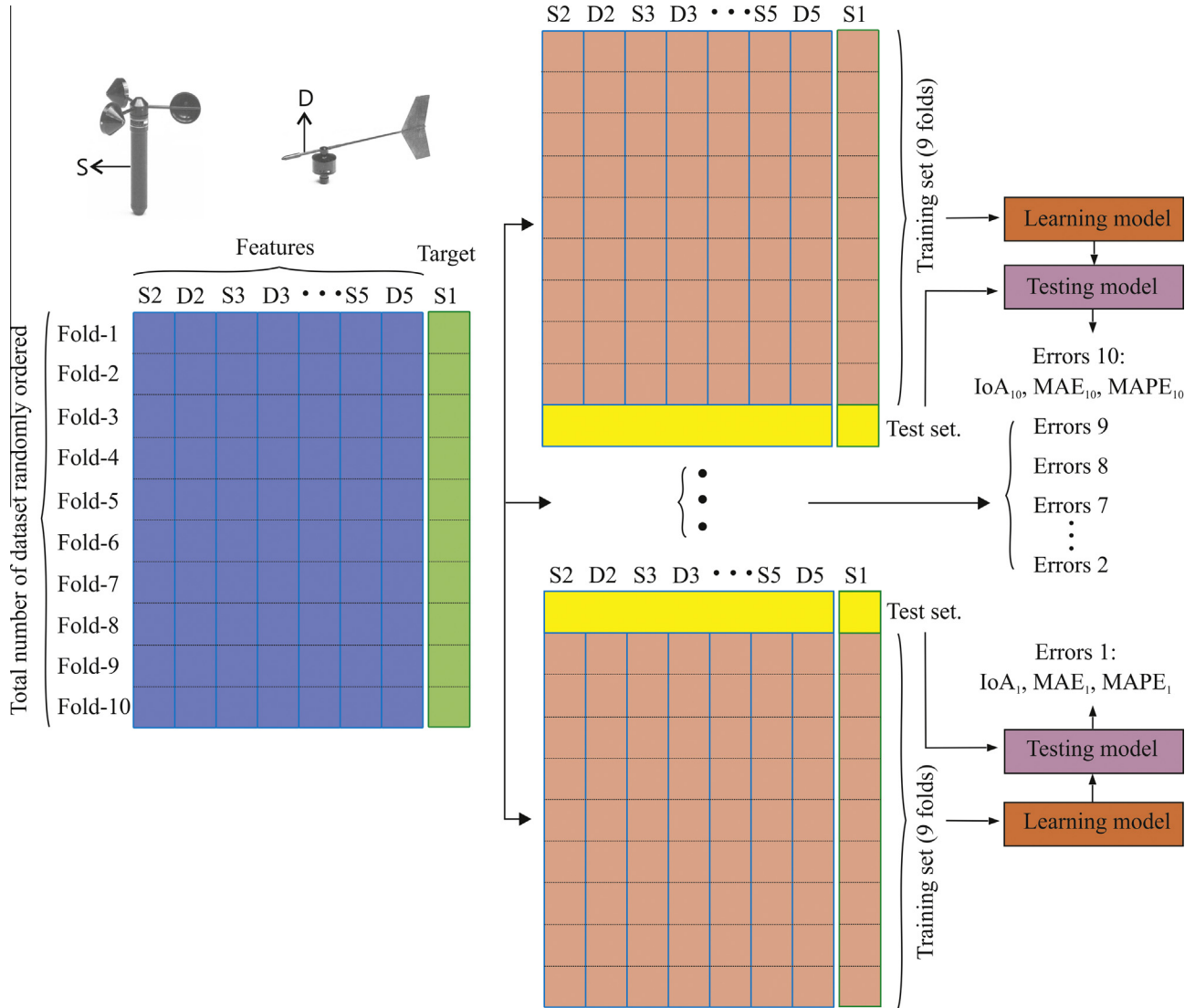
Fig. 6. Schematic representation of the 10-fold cross-validation mechanism used in the study.

"o". MAE is expressed in the same units as the parameters it compares.

$$MAE = \frac{\sum_{i=1}^{n}|e_i - o_i|}{n} \qquad (5)$$

With the MAE, all sizes of error are treated evenly according to their magnitude. MAPE is defined by Eq. (6).

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{e_i - o_i}{o_i}\right| \qquad (6)$$

The MAPE is a relative measurement that expresses the error as a percentage of the observed data.

Willmott et al. [36] propose the IoA that is given by Eq. (7), where $\bar{o}$ is the arithmetic mean of the values observed. The IoA is dimensionless and its value is found in the range from -1 to 1.

$$IoA = \begin{cases} 1 - \dfrac{\sum_{i=1}^{n}|e_i - o_i|}{2\sum_{i=1}^{n}|o_i - \bar{o}|}, & \text{When} \\[2mm] \sum_{i=1}^{n}|e_i - o_i| \leqslant 2\sum_{i=1}^{n}|o_i - \bar{o}| \\[2mm] \dfrac{2\sum_{i=1}^{n}|o_i - \bar{o}|}{\sum_{i=1}^{n}|e_i - o_i|} - 1, & \text{When} \\[2mm] \sum_{i=1}^{n}|e_i - o_i| > 2\sum_{i=1}^{n}|o_i - \bar{o}| \end{cases} \qquad (7)$$

According to Willmott et al. [36], in general, the IoA is more rationally related to model accuracy than other indices in use. They also point out that this index is quite flexible and so is applicable to a wide range of model-performance problems. When IoA values are close to 1 this indicates strong agreement between the results of the model and the observations. If IoA is equal to 0, this means according to Willmott et al. [36] that "the sum of the magnitudes of the errors and the sum of the perfect-model-deviation and observed-deviation magnitudes are equivalent". However, the same authors also point out that "values of IoA near -1 can mean that the model-estimated deviations about o are poor estimates of the observed deviations; but, they also can mean that there simply is little observed variability. As the lower limit of IoA is approached, interpretations should be made cautiously".

2.3. Hardware used for the calculations

Given the methodological procedure used (described in Section 3) to attain the objectives outlined and given the volume of data involved in the study, it is clear that the computational time required would be considerable if using a PC with a low number of microprocessors. It was therefore decided to use a supercomputer, in this case Atlante which forms part of the Spanish

Supercomputing Network and has a distributed-memory cluster [48]. Atlante consists of 84 IBM JS21 compute nodes (blades). Each blade has two dual-core processors at 2.5 GHz running the Linux operating system with 8 GB of memory RAM and 73 GB local disk storage. In total, peak performance of Atlante is 3.09 TFlops. One of the networks that interconnects Atlante (Myrinet Network) has a high bandwidth which it uses for parallel communication applications.

3. Methodology

3.1. Preamble

A general outline of the procedure used is shown in Fig. 7. For easier interpretation the procedure has been particularised for the case in which the short-term series of wind speeds and directions at the reference stations are the four shown in Table 1 and Fig. 1 (with codes WS-2 to WS-5) and the target site is WS-1. The year 2003 is taken as the time period representative of the short-term common to all stations. As indicated by various authors [1,2,4,5,9,10], a long series of wind data at a target site is required to estimate the corresponding long-term wind resource (some authors speak of the need for 20 or 30 years worth of data [1,5,7–10]). Given that no such long historical series of data (which would meet the typical constraints of MCP methods [10]) were available for a significant number of stations which would allow the analysis of different degrees of correlation, the stages indicated in Fig. 7 by a circled '3' and circled '4' will not be taken into consideration in the procedure employed in this paper. In other words, the model construction and testing stages are covered, as the length of the short-term representative series extends to one year and, therefore, the seasonal variation influence of the wind characteristics is considered to be picked up, as is generally recommended [10]. However, long-term hindsighting of the wind conditions of the target station is not carried out.

Though Fig. 7 represents a specific case, the work actually performed was much more extensive in that the cases analysed included each of the five WSs of Table 1 as target site and both 2003 and 2004 were used as the short-term representative year.

3.2. Procedure followed

As can be seen in Fig. 7, the procedure consists of three independent action blocks to produce the different MCP models that result from using MLP neural networks.

One block uses the complete set of variables, which is to say that no type of feature selection takes place. This is indicated by the letters 'FS' (Full Set) enclosed in a polygon in Fig. 7. This block produces MLP neural networks trained with all the variables according to the procedure and model selection described in Sections 2.2.1 and 2.2.2. The other two blocks employ the feature selection methods explained in Section 2.2.3. The FA block is represented in Fig. 7 by the itinerary represented by the initials "FA-1" and "FA-2" enclosed in polygons. MLP neural networks are generated in this process according to the model training and selection algorithm described in Section 2.2.2, using as input the subset of variables with highest merit after an exhaustive search through the space of feature subsets. The WA block is represented in Fig. 7 by the itinerary represented by the initials "WA-1" and "WA-2" enclosed in polygons. This block produces MLP neural networks trained with the model selection and training algorithm described in Section 2.2.2, using as input the subset of variables which produces the best predictive results in a 10-fold cross-validation process using MLP neural networks (also trained as described in Section 2.2.2).

The itinerary of each of the three blocks (FS, FA, WA) leads to the selection of the best model for each strategy. Once the best models for the FS, FA and WA strategies have been selected, a comparison between them is made based on the magnitudes of the mean errors obtained in the cross-validations. This comparison focuses on two aspects. One is based on a comparison of the magnitudes of the metrics obtained in the cross-validations (indicated in Fig. 7 by a circled '1') and the other on the classic statistical analysis of null hypothesis (see Section 3.2.1). The aim is to know whether, from a statistical point of view, there exists a significant difference between the results obtained with the three strategies (this action is indicated by a circled '2' in Fig. 7).

3.2.1. Testing for statistical significance

The purpose of the statistical analysis presented in this subsection is to determine whether the mean error of the ten mean errors obtained with the 10-fold cross-validation of a model is significantly larger or smaller than the mean error of another model. Note that, as three metrics have been proposed (IoA, MAE and MAPE) to evaluate wind speed prediction (Section 2.2.4), there are three samples for each metric (one sample for each model: FA, WA and FS) and each sample has ten data values.

Note also that, in order to rationalise the computational load of the procedure, the same random cross-validation partitions were used in the process to obtain each of the models (FA, WA and FS). The reason behind this was also to try to minimize the variance of the difference between the mean metrics obtained by the three models, as these share the same experimental unit (the same partition of data of each cross-validation iteration: the same nine groups used for the training and the same group excluded for validation in which the metric is evaluated). So, for each metric we have a scenario of analysis of variance with three levels of treatment and three potentially dependent samples (ANOVA within subjects), in which the two-by-two comparisons will be made. As multiple comparisons will be made, the adjusted p-values will be calculated using the procedure proposed by Benjamini and Hochberg [49], normally called the BH step-up procedure.

The decision problem consists of choosing between the null hypothesis H_0 and the alternative hypothesis H_1, Eq. (8), with a significance level of 5%. In the case of the IoA metric, if the null hypothesis is rejected then model i is considered better than model j. In the case of the MAE and MAPE metrics, if the null hypothesis is rejected then model j is considered better than model i. In the case that it is not possible to reject the null hypothesis then it cannot be said that either of the two models that are being compared is better than the other.

$$H_0 : \mu_i = \mu_j; \quad H_1 : \mu_i > \mu_j \tag{8}$$

In Eq. (8), μ_i and μ_j represent the population means of the models i and j, respectively.

To resolve the problem, Eq. (8), both the paired and unpaired parametric Student's t-test [50,51] will be used in this paper and a judgement value will made in view of the correlations that exist between the samples. When the hypothesis of normality is rejected (the modified statistic of Anderson–Darling [52] was applied for this purpose), given that the number of data is small (n_{cv} = 10), non-parametric permutation tests were used for both dependent (paired) samples and independent (unpaired) samples. The choice between them will depend on the correlations that exist between the samples.

The non-parametric permutation tests used in this study were first proposed by Fisher [43] and Pitman [44] and have continued to evolve to the present day on the basis of subsequent work carried out by the same authors [45,46].
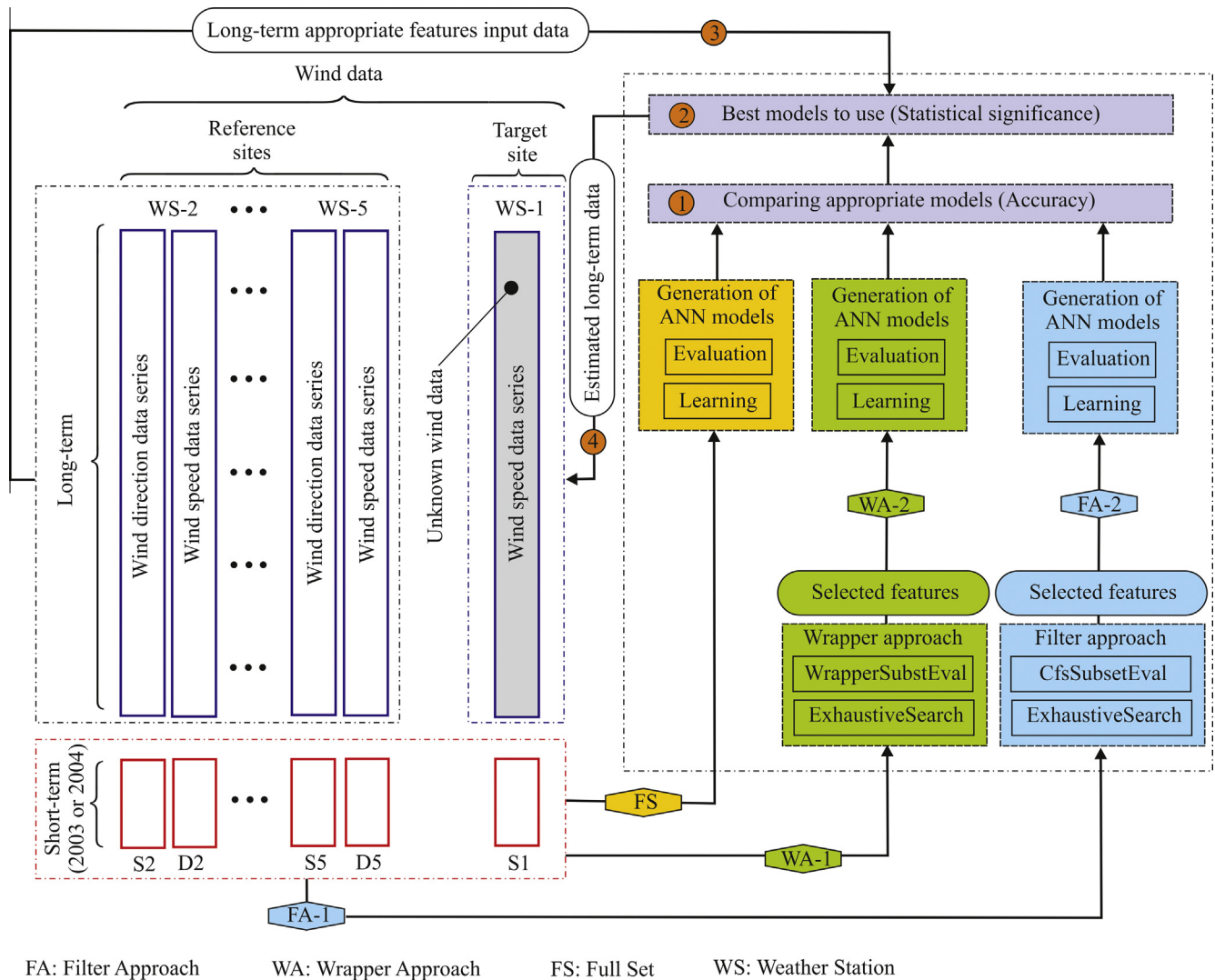
Fig. 7. General configuration of the methodological procedure followed.

The null hypothesis H_0 will be discarded if p-value (adjusted) $\leqslant$ α in favour of the alternative hypothesis H_1.

4. Analysis of results

The most important results obtained by the best models generated following the procedure described in Section 3.2 for the filter, wrapper and full feature set methods are shown in Figs. 8–10. Each of these three figures corresponds to one of the metrics used in the comparison (Section 2.2.4) and is comprised of two graphical representations which each show the mean loss obtained when predicting the mean wind speeds using one of the years as a short-term representative training-validation sample (a on the left for 2003; b on the right for 2004).

Each of these two graphical representations shows, from top to bottom, the optimum number of neurons for the model selected in each strategy, the variables finally selected in each model, and a bar diagram indicating the mean loss in the 10 cross-validation groups used and an interval centred on this mean of the standard deviation of the loss in the cross-validation process.

The computational load required for selection of the MCP models (including feature selection) was counted by measuring the CPU times (using 32 of the 84 blades of the hardware indicated in Section 2.3) needed to run the models. By way of example, the

CPU time required for selection of the MCP model of the target station WS-1 (year 2003) using the IoA metric and FA method was 1019s, while the time required for the same WS and metric in the case of the WA method was 66803s. Expressed another way, the CPU time required for the WA-based model was 6456% higher than for the FA model.

Tables 5–7, each of which corresponds to one of the three metrics (IoA, MAE and MAPE) used in the comparison, show the results of the application of the different statistical hypothesis tests explained in Section 3.2.1 to the losses obtained by the different models (FA, WA and FS) in the cross-validation process used in their training-validation.

The content of each of the three tables is the same, but particularised for each metric. The first column indicates the target station in each MCP problem, the second shows the alternative hypothesis (H_1), the third the data year (short-term representative) used for training-validation of the models whose 10 cross-validation results are being compared, and the fourth the linear correlation coefficient (ρ) obtained between the samples being compared. The fifth, sixth and seventh columns show the p-values of the following statistical tests: the Anderson–Darling Test for normality of the difference between two samples of mean losses produced by the two models being compared, and the Student Test and Permutation Test (both for paired samples). The last column shows the adjusted p-value, following the BH procedure [49] of the paired Permutation
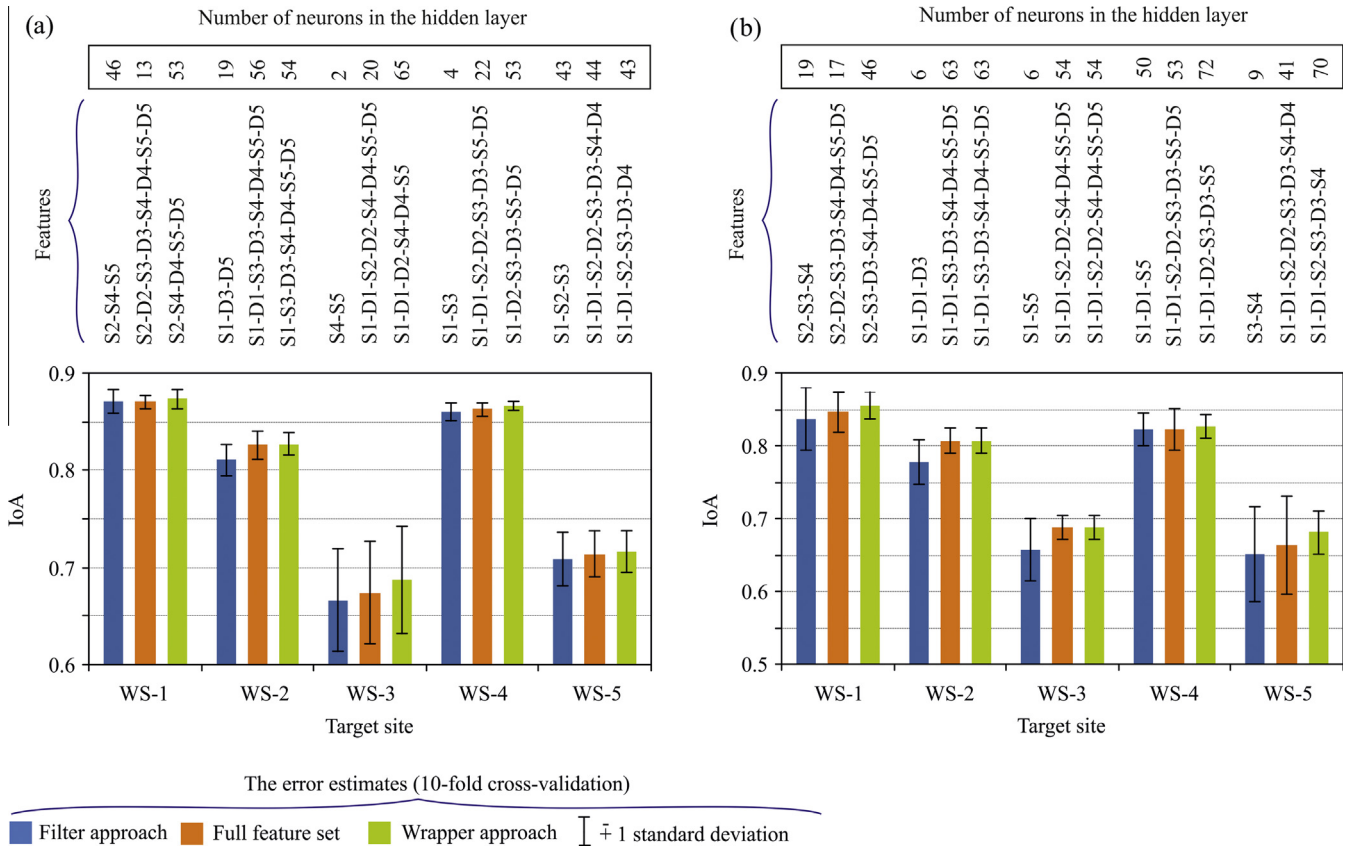
Fig. 8. Results obtained for the IoA metric: (a) Estimation of mean hourly wind speeds (training year: 2003), (b) estimation of mean hourly wind speeds (training year: 2004).

Test. Note that the *p*-values obtained from a total of 90 comparison tests (5 WS × 3 models (FA, WA and FS) × 2 years (2003 and 2004) × 3 metrics (IoA, MAE and MAPE)) were used to calculate the *p*-values shown in the final column of Tables 5–7 with the BH method. Note too that the adjusted *p*-values of Tables 5–7 are written in bold when the null hypothesis is rejected and the alternative hypothesis shown in the second column of these tables is accepted. The adjusted *p*-values which correspond to target station WS-4, year 2003 and alternative hypotheses WA > FA and WA > FS are written in parenthesis in Table 5 in order to show negative and low correlation coefficients between samples. The same thing occurs in Table 6 for the same station and year, in the case of the alternative hypotheses FA > WA and FS > WA. In these cases, the unpaired tests described in Section 3.2.1 (Table 8) were also used in view of these correlations to make a judgement value. The adjusted *p*-values shown in the final column of Table 8 and those shown in parenthesis in Tables 5 and 6 lead to the conclusion that, in the cases analysed, both the paired and unpaired tests indicate that there is no evidence to reject the null hypothesis that there are no significant differences with a significance level of 5%.

An analysis of the results is performed below with a distinction between the results that could be considered of a methodological type, as they are associated with the comparison made between the filter, wrapper and full variable set strategies using MLP neural networks, and those which could be considered specific results of the different MCP problems considered for this comparison in the wind scenario of the Canary Archipelago.

4.1. Methodological type results

The results shown in Figs. 8–10 are synthesised in Fig. 11, where it can be seen that, independently of the metric used (IoA, MAE or MAPE), the WA gave lower mean errors than the FA in

100% of the cases analysed (10 cases: 5 stations × 2 years). It can also be seen that in at least 80% of the cases analysed the WA gave lower mean errors than the FS and that the FS never gave lower errors than the WA. In addition, in at least 80% of the cases analysed the FS gave lower mean errors than the FA. However, the FS generated higher mean errors than the FA in 20% of the cases when the IoA and MAE metrics were used.

Despite the results shown in Fig. 11, it should be noted that, at first sight and considering the magnitudes of the metrics generated by the three models (WA, FA and FS) (Figs. 8–10), no notable difference can be observed between them and, therefore, none of the models can be totally discarded *a priori* without considering other possible benefits. In particular, the FA appears to be competitive at predictive level in the cases in which the errors were lower (WS-1 and WS-4), despite using a smaller number of variables than the other two models. This means that fewer computational resources are required and, no less important, allows for much greater interpretability.

As can be observed in Figs. 8–10, the WA produces models with a higher number of variables than the FA which contribute some information in the prediction and which are useful to the MLP networks. This happens despite the redundancy between variables observed in Section 2.1. So, the MLP networks can manage this redundancy without missing out on the specific information that each variable may provide about the target variable. However, the FA generated simpler models than the WA, as it used between 40% and 75% fewer variables than the WA, and with a smaller number of neurons in the hidden layers. The higher number of neurons of the hidden layer of the MLP produced by the WA has an impact on the number of iterations and the computational load of the training algorithms. So, the MLP training algorithms, in the case of using FA, are faster and potentially more effective at estimating the parameters.
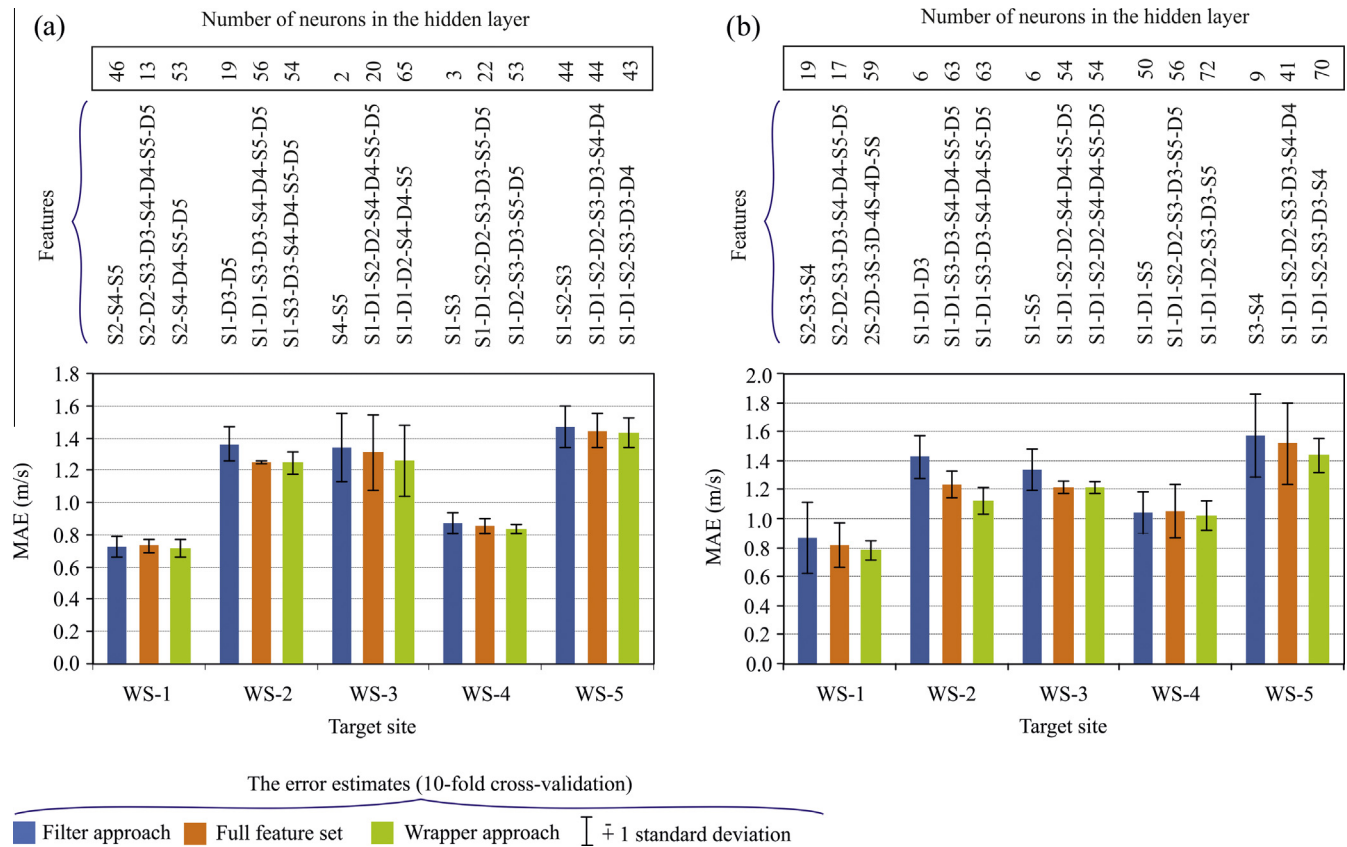
Fig. 9. Results obtained for the MAE metric: (a) Estimation of mean hourly wind speeds (training year: 2003), (b) estimation of mean hourly wind speeds (training year: 2004).

It is also clear from Figs. 8–10 that the differences in the sets of variables selected by both methods (WA and FA) are fundamentally to be found in the direction variables. This is coherent with the larger non-linear component of the relationships between the direction variables and the speed variables which act as target (Fig. 4), resulting in low Pearson's correlation coefficients which are insufficient for the CFS filter method, which uses these coefficients, to select the direction variables in the subset. In all cases, the absolute value of Pearson's correlation coefficient was below 0.5 and in some cases below 0.04 (Table 9). The authors of the present paper consider that it would be convenient for the filter method to use with these pairs of variables the circular-linear correlation coefficient, Eq. (3), which detects these non-linear relationships, with the aim of working under the same conditions as the WA. With this in mind, the authors propose to undertake this methodological improvement in a future study with a view to confirming or qualifying the parsimony differences and the predictive differences between the two strategies that were obtained in the present study.

The results of Tables 5–7 are summarised in Fig. 12 and show that, independently of the metric used, the WA performed significantly better than the FA in 60% of cases and in no case did it perform worse. It can also be seen in Fig. 12 that the WA did not give significantly different results from the FS for the IoA and MAPE metrics in 90% of the cases analysed (in the remaining 10% of cases the WA was significantly better). There were no significant differences between the WA and the FS in 100% of cases with the MAE metric. Meanwhile, the FS was significantly better than the FA in 40% of cases with the IoA and MAE metrics, and never worse. With the MAPE metric, the FS was better than the FA in 20% of cases with no significant differences observed between them in the remaining 80%.

4.2. Results of the different MCP problems posed in the Canary Archipelago

For the analysis of the particular results of the different stations considered as target, a distinction will be made between two groups of stations: one group comprised of WS-1 and WS-4 and a second of WS-2, WS-3 and WS-5.

- Stations WS-1 and WS-4: it can be seen in Figs. 8–10 that the smallest errors in mean hourly wind speed prediction were made for these target stations, independently of the models (WA, FA and FS) used and of the year (2003 or 2004) used to represent the short-term. It can even be observed in the case of these two stations that in most cases analysed (metrics and year) there was no significant statistical difference between the three models (FA, WA and FS) at a 5% level (Tables 5–8). Specifically, in the case of the MAE metric (Tables 6 and 8) in no case is there evidence to reject the null hypothesis at that significance level. The same can be said for the MAPE metric (Table 7) in the case of 2004 (in the case of 2003, the *p*-values are not much lower than 0.05 and there would be no evidence to reject the null hypothesis in the case of a significance level below 4%). It can be deduced from the analyses that an absence of evidence to reject the null hypothesis is fundamentally a consequence of the high degree of linear correlation between the wind speeds recorded at these target stations and the data available at the stations which served as reference (Table 2). The direction variables are less relevant. It can be seen that the direction variables did not intervene in the FA models used in wind speed estimation of the target stations WS-1 and WS-4. The degree of linear correlation influences especially the magnitude of the IoA metric. The highest correlation coefficients
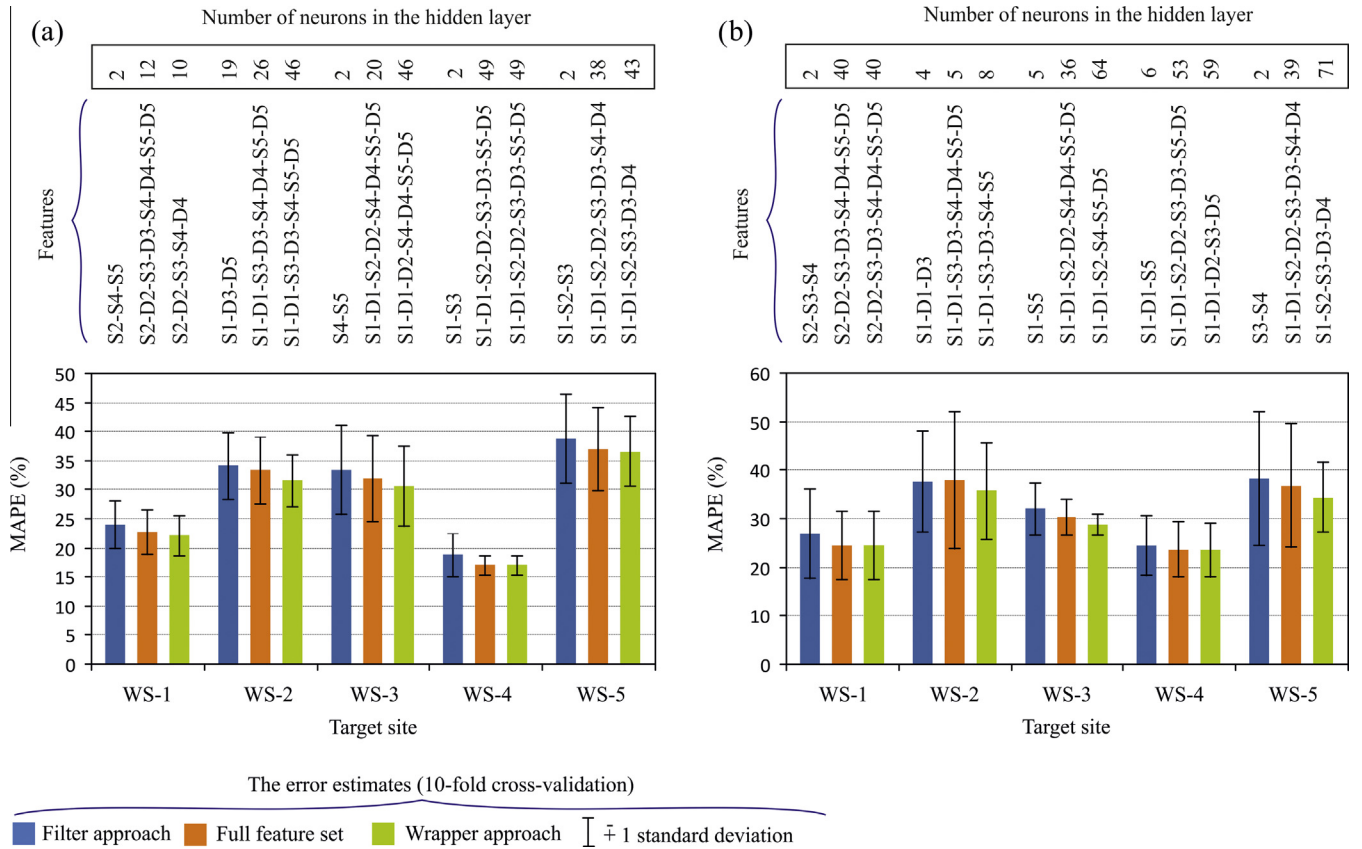
Fig. 10. Results obtained for the MAPE metric: (a) Estimation of mean hourly wind speeds (training year: 2003), (b) estimation of mean hourly wind speeds (training year: 2004).

Table 5
Analysis of statistically significant differences for the IoA metric. Tests for paired samples.

Target site	Alternative hypothesis	Training year	ρ	p-value			p-value (adjusted)
				Anderson Darling	Student test	Permutation test	Permutation Test
WS-1	FA > FS	2003	0.598	0.1044	0.3833	0.3809	0.4223
	WA > FA		0.535	0.2588	0.265	0.2676	0.3211
	WA > FS		0.814	0.5340	0.0616	0.0693	0.1199
	FS > FA	2004	0.976	0.0040	–	0.0313	0.0704
	WA > FA		0.861	0.0001	–	0.0010	**0.0129**
	WA > FS		0.918	0.0035	–	0.0068	**0.0306**
WS-2	FS > FA	2003	0.692	0.8929	0.0014	0.0029	**0.0218**
	WA > FA		0.849	0.5386	0.0001	0.0010	**0.0129**
	WA > FS		0.936	0.0415	–	0.3848	0.4223
	FS > FA	2004	0.362	0.2732	0.0049	0.0039	**0.0219**
	WA > FA		0.362	0.2732	0.0049	0.0039	**0.0219**
	WA > FS		1.000	–	–	1.0000	1.0000
WS-3	FS > FA	2003	0.982	0.1462	0.015	0.0107	**0.0401**
	WA > FA		0.942	0.4396	0.003	0.0029	**0.0218**
	WA > FS		0.946	0.5491	0.0249	0.0273	0.0683
	FS > FA	2004	0.759	0.0013	–	0.0010	**0.0129**
	WA > FA		0.759	0.0013	–	0.0010	**0.0129**
	WA > FS		1.000	–	–	1.0000	1.0000
WS-4	FS > FA	2003	0.621	0.3640	0.1246	0.1260	0.1772
	WA > FA		(−0.259)	0.0028	–	(0.0198)	(0.0540)
	WA > FS		(−0.200)	0.0214	–	(0.1172)	(0.1716)
	FA > FS	2004	0.951	0.0092	–	0.3096	0.3572
	WA > FA		0.930	0.0182	–	0.1338	0.1853
	WA > FS		0.876	0.1256	0.1586	0.1660	0.2134
WS-5	FS > FA	2003	0.939	0.0188	–	0.0488	0.0994
	WA > FA		0.942	0.4312	0.0194	0.0156	**0.0453**
	WA > FS		0.971	0.6267	0.0785	0.0742	0.1237
	FS > FA	2004	0.945	0.1945	0.0674	0.0684	0.1199
	WA > FA		0.915	0.0042	–	0.0049	**0.0245**
	WA > FS		0.915	0.0017	–	0.1494	0.2004

Table 6
Analysis of statistically significant differences for the MAE metric. Tests for paired samples.

Target site	Alternative hypothesis	Training year	ρ	p-value			p-value (Adjusted)
				Anderson Darling	Student Test	Permutation Test	Permutation Test
WS-1	FS > FA	2003	0.588	0.1056	0.3808	0.3799	0.4223
	FA > WA		0.529	0.2595	0.2638	0.2656	0.3211
	FS > WA		0.807	0.5023	0.0614	0.0693	0.1199
	FA > FS	2004	0.978	0.0032	–	0.0313	0.0704
	FA > WA		0.820	0.0000	–	0.0381	0.0836
	FS > WA		0.855	0.0084	–	0.1914	0.2393
WS-2	FA > FS	2003	0.643	0.8646	0.0014	0.0029	**0.0218**
	FA > WA		0.825	0.5538	0.0001	0.001	**0.0129**
	FS > WA		0.928	0.0372	–	0.3926	0.4257
	FA > FS	2004	0.310	0.2997	0.005	0.0039	**0.0219**
	FA > WA		0.310	0.2997	0.005	0.0039	**0.0219**
	FS>WA		1.000	–	–	1.0000	1.0000
WS-3	FA>FS	2003	0.982	0.1670	0.0148	0.0107	**0.0401**
	FA>WA		0.939	0.3731	0.0032	0.0029	**0.0218**
	FS>WA		0.945	0.6174	0.0242	0.0273	0.0683
	FA>FS	2004	0.593	0.0024	–	0.0010	**0.0129**
	FA>WA		0.593	0.0024	–	0.0010	**0.0129**
	FS>WA		1.000	–	–	1.0000	1.0000
WS-4	FA>FS	2003	0.730	0.3036	0.1178	0.1250	0.1772
	FA>WA		(−0.026)	0.0020	–	(0.0196)	(0.0540)
	FS>WA		(−0.141)	0.0227	–	(0.1172)	(0.1716)
	FS>FA	2004	0.959	0.0080	–	0.3027	0.3538
	FA>WA		0.936	0.0170	–	0.1387	0.1891
	FS>WA		0.901	0.1153	0.1576	0.1641	0.2134
WS-5	FS>FA	2003	0.936	0.0201	–	0.041	0.0879
	FA>WA		0.937	0.4396	0.0164	0.0156	**0.0453**
	FS>WA		0.966	0.5866	0.0744	0.0742	0.1237
	FA>FS	2004	0.934	0.1906	0.0691	0.0693	0.1199
	FA>WA		0.886	0.0055	–	0.0049	**0.0245**
	FS>WA		0.892	0.0017	–	0.1514	0.2004

Table 7
Analysis of statistically significant differences for the MAPE metric. Tests for paired samples.

Target site	Alternative hypothesis	Training year	ρ	p-value			p-value (Adjusted)
				Anderson Darling	Student Test	Permutation Test	Permutation Test
WS-1	FA>FS	2003	0.828	0.0354	–	0.0498	0.0994
	FA>WA		0.844	0.5628	0.0135	0.0107	**0.0401**
	FS>WA		0.984	0.8257	0.0158	0.0137	**0.0453**
	FA>FS	2004	0.908	0.0188	–	0.0596	0.1118
	FA>WA		0.908	0.0188	–	0.0596	0.1118
	FS>WA		1.000	–	–	1.0000	1.0000
WS-2	FA>FS	2003	0.871	0.584	0.1962	0.2061	0.2541
	FA>WA		0.876	0.2688	0.0085	0.0059	**0.0279**
	FS>WA		0.765	0.8988	0.083	0.0859	0.1397
	FS>FA	2004	0.987	0.0003	–	0.3721	0.4223
	FA>WA		0.981	0.0787	0.0100	0.002	**0.0218**
	FS>WA		0.980	0.0185	–	0.1123	0.1713
WS-3	FA>FS	2003	0.964	0.8460	0.0282	0.0313	0.0704
	FA>WA		0.880	0.1565	0.0206	0.0225	0.0596
	FS>WA		0.906	0.5339	0.1080	0.1182	0.1716
	FA>FS	2004	0.938	0.4974	0.0177	0.0127	**0.0440**
	FA>WA		0.445	0.0082	–	0.0078	**0.0334**
	FS>WA		0.645	0.0003	–	0.0303	0.0704
WS-4	FA>FS	2003	0.5733	0.0017	–	0.0146	**0.0453**
	FA>WA		0.5733	0.0017	–	0.0146	**0.0453**
	FS>WA		1.000	–	–	1.0000	1.0000
	FA>FS	2004	0.954	0.3466	0.0865	0.0928	0.1465
	FA>WA		0.941	0.0120	–	0.0967	0.1501
	FS>WA		0.965	0.1940	0.4790	0.4795	0.5138
WS-5	FA>FS	2003	0.912	0.1175	0.0479	0.0508	0.0994
	FA>WA		0.959	0.6020	0.0133	0.0117	**0.0421**
	FS>WA		0.973	0.3399	0.2874	0.2881	0.3412
	FA>FS	2004	0.976	0.9226	0.0875	0.0869	0.1397
	FA>WA		0.944	0.0100	–	0.0488	0.0994
	FS>WA		0.974	0.0010	–	0.1738	0.2203

Table 8
Analysis of statistically significant differences. Test for unpaired samples.

Metric	Alternative hypothesis	Training year	p-value				p-value (Adjusted)
			Anderson Darling	Anderson Darling	Student test	Permutation test	Permutation test
IoA	WA>FA	2003	0.0925	0.0001	–	0.0187	0.0526
	WA>FS		0.0925	0.0680	0.0902	0.0888	0.1402
MAE	FA>WA	2003	0.0002	0.9497	–	0.0196	0.0535
	FS>WA		0.0914	0.9497	0.0943	0.0944	0.1440
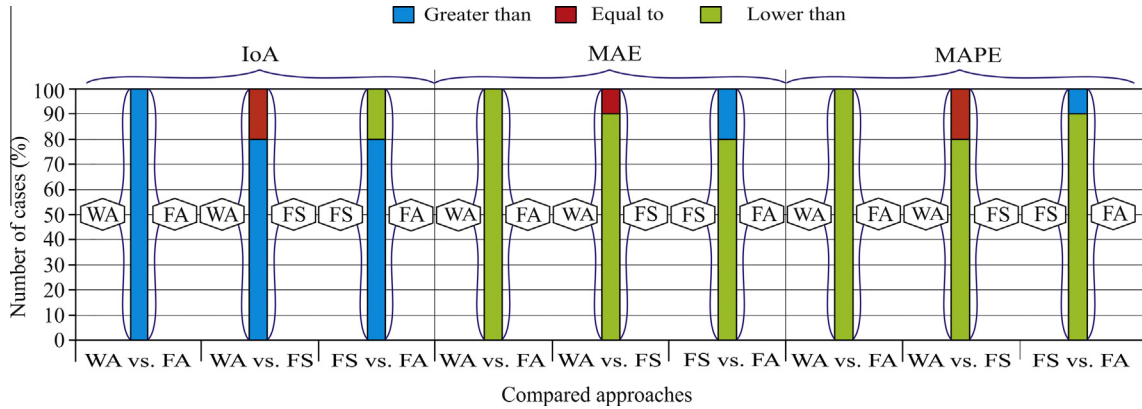
Fig. 11. Comparison between the magnitudes of the metrics of the three models (WA, FA and FS) obtained in the 10-fold cross-validation process, for the ten cases analysed (5 weather stations × 2 years).

Table 9
Pearson's correlation coefficients between the wind speeds (vertical) and wind directions (horizontal) of the five anemometer weather stations.

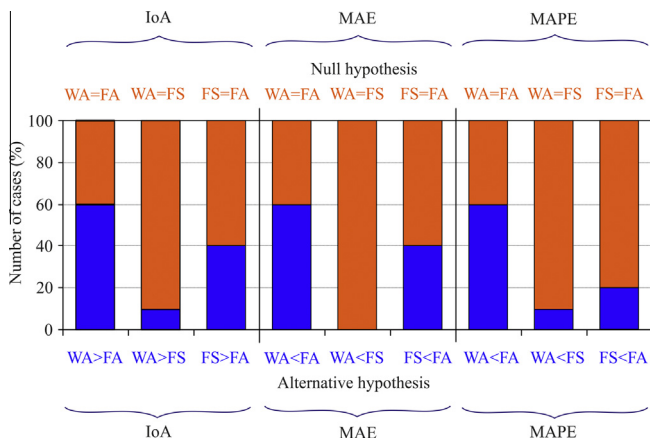| W/D	Year 2003					Year 2004				
	WS-1	WS-2	WS-3	WS-4	WS-5	WS-1	WS-2	WS-3	WS-4	WS-5
WS-1	−0.470	−0.431	−0.410	−0.404	−0.0369	−0.425	−0.404	−0.363	−0.357	0.038
WS-2	−0.431	−0.424	−0.467	−0.342	−0.452	−0.404	−0.423	−0.408	−0.317	−0.061
WS-3	−0.410	−0.467	−0.366	−0.283	−0.281	−0.363	−0.357	−0.297	−0.208	−0.061
WS-4	−0.404	−0.342	−0.283	−0.444	−0.334	−0.357	−0.317	−0.208	−0.384	0.073
WS-5	−0.369	−0.452	−0.281	−0.334	−0.296	0.038	−0.061	−0.061	0.073	0.141

Fig. 12. Comparison between the percentage of cases of the 10 considered (5 weather stations × 2 years) in which the alternative hypotheses are accepted over the null hypotheses for each of the three metrics (IoA, MAE and MAPE) analysed.

observed between the wind speeds of the target stations WS-1 and WS-4 were 0.948 (year 2003) and 0.93 (year 2004). According to the rules of thumb recommended for MCP methods in the literature [10], these correlation coefficients can be classified 'very good'.

- Stations WS-2, WS-3 and WS-5: The highest linear correlation coefficients detected between the target stations WS-2, WS-5 and WS-3 and the stations which were used as reference were 0.878, 0.756 and 0.709, respectively, in year 2003, and 0.876, 0.701 and 0.686, respectively, in year 2004. According to the aforementioned rules of thumb, these correlation coefficients would be classified as 'poor', 'moderate' or 'good' depending on the range (0.6–0.7, 0.7–0.8 and 0.8–0.9, respectively) they are found in.

It can be seen in Figs. 8–10, in the case of the target stations WS-2, WS-3 and WS-5, where the coefficients of linear correlation with the reference stations would be catalogued according to the aforementioned rules of thumb as good (station WS-2) and moderate or poor (stations WS-5 and WS-3), that the direction variables and the redundancy between variables in the MLPs produced by the wrapper method enhance, in comparison with the filter methods, the predictive capacity of the wind speeds. The filter methods used (CFS) renounced redundancy between variables and discarded the use of the direction variables, with the exception of target station WS-2, where the FA models used the direction variables D3 and D5 (year 2003) and D3 and D1 (year 2004). It can be seen in Tables 5–8 that, in the case of these three target stations and for all cases analysed (metrics and years), there are statistically

significant differences (5% significance level) between the errors generated by the filter models and the wrapper models, with the WA errors being lower. Similar conclusions can be drawn when comparing the FS and FA models, except in the case of target station WS-5, where the differences between the errors generated by these models were not statistically significant at a 5% level. It can also be seen in Tables 5–8 that, for these three target stations in no case was there evidence to reject the null hypothesis chosen to compare the WA and the FS.

Although the study undertaken focussed on the short-term period, it should be remembered that the purpose of MCP models is to estimate long-term wind speed series at a target site. In this context, and as mentioned in Section 3.1, MCP methods require a series of conditions to be met for them to be useful [10]. These conditions are as follows: (a) The wind data series must have been recorded by the WSs in compliance with certain standards. In particular, there must have been no changes to the area surrounding any of the WSs (new buildings, installation of wind farms, major changes to the vegetations, etc.) that could have modified the relationships between the wind data of the WSs. (b) The data sets must be statistically stationary (wind behaviour in the future during the working life of the energy project must be analogous to past behaviour). (c) The short-term data series recorded at the target WS must allow conclusions to be drawn about seasonal variations (wind speed and direction), and (d) The wind climate of the different WSs must be similar.

In the framework of the present study, it should be noted that the non-availability of long wind data series meant it was not possible to analyse some of the previously described determinants of MCP methods. Nonetheless, the conclusions drawn from the comparative analysis of WA, FA and FS methods in the short-term periods for each of the two years considered are very similar in terms of predictive ranking, the selection of variables made and even the number of hidden units obtained in the models of the two years. Moreover, this was despite the instability that MLP algorithms usually present due to the multiple local optima in which non-linear optimisation algorithms can terminate. This suggests that the conclusions obtained in this study are applicable to MCP methods if it can be assumed that the wind climate of the WSs fits a sufficiently stable pattern (as is the case of the two years studied).

5. Conclusions

A specific analysis was undertaken in this paper of the usefulness of feature selection methods using ANNs in MCP methods. To date, no similar analysis has been published in the literature and no feature selection methods have been implemented in MCP modules of the software programmes employed in the wind industry. This study compared two general strategies for feature selection in real conditions, the CFS filter method and the wrapper method, and the two were also compared to a model (FS) which made no feature selection. MLP neural networks trained with quadratic loss through the backpropagation algorithm were used as an inductive algorithm.

The comparison was made in the context of an MCP strategy for the prediction of the wind resource at five stations chosen successively as the target site, using the four remaining stations as reference sites. This entailed posing five MCP problems in which the dependent variable was the wind speed at the target site and the independent variables were the wind speeds and directions at the four reference stations. Below, we present the main conclusions obtained from the study:

When a statistical comparison of the wind speeds of the target WS and reference WSs gives Pearson's correlation coefficients which are ranked in the literature as very good (between 0.9 and

1), the FA method can be considered competitive in terms of predictive ability and is more interpretable by the analyst than the WA method. The FA method uses a lower number of features than the WA method and also requires less computational resources. However, the best results of the WA show that when MLP neural networks are being used as a predictive model the redundancy between variables can be beneficial for the prediction as well as lending robustness to the model. As the degree of non-linearity in the relationships between features increases (for example, when the WSs are located in complex terrain where the direction variable may acquire more significance), use of the WA method for feature selection becomes more recommendable, although the FA can still be used as an interpretive tool. However, it must be remembered that using WA methods has a high computational cost. Moreover, these methods require enough data for complex non-linear relationships to show up.

In the case analysed, the differences between the sets of features selected by the WA and FA methods were fundamentally found in the wind direction variables. As the CFS filter method uses Pearson's linear correlation coefficients for feature selection, when these are low (the case of wind direction variables) the CFS method tends to discard them in the subset. For this reason, incorporation in the filter methods is recommended of correlation coefficients of a non-linear nature which are able to detect the non-linear relationships that may exist between the wind direction features and between these and the linear features. In this way, the lower computational load of these methods can be taken advantage of without losing predictive capacity.

References

[1] Jain P. Wind energy engineering. 1st ed. New York: McGraw-Hill; 2011.
[2] Brower MC. Wind resource assessment. 1st ed. New Jersey: Wiley; 2012.
[3] Velázquez S, Carta JA, Matías JM. Comparison between ANNs and linear MCP algorithms in the long-term estimation of the cost per kWh produced by a wind turbine at a candidate site. A case study in the Canary Islands. Appl Energy 2011;88:3869–81.
[4] Hiester TR, Pennell WT. The siting handbook for large wind energy systems. 1st ed. New York: WindBook; 1981.
[5] Justus CG, Mani K, Mikhail AS. Interannual and month-to-month variations of wind speed. J Appl Meteorol 1979;18:913–20.
[6] Landberg L, Myllerup L, Rathmann O, Petersen EL, Jørgensen BH, Niels BJ, et al. Wind resource estimation – an overview. Wind Energy 2003;6:261–71.
[7] Baker R, Walker SN, Wade JE. Annual and seasonal variations in mean wind speed and wind turbine energy production. Sol Energy 1990;45:285–9.
[8] Klink K. Trends and interannual variability of wind speed distributions in Minnesota. J Clim 2002;15:3311–7.
[9] Gasch R, Twele J. Wind power plants. 2nd ed. Berlin: Springer; 2012.
[10] Carta JA, Velázquez S, Cabrera P. A review of measure–correlate–predict (MCP) methods used to estimate long-term wind characteristics at a target site. Renew Sustain Energy Rev 2013;27:362–400.
[11] Jung S, Kwon SD. Weighted error functions in artificial neural networks for improved wind energy potential estimation. Appl Energy 2013;111:778–90.
[12] Zhang J, Chowdhury S, Messac A, Hodge BM. A hybrid measure–correlate–predict method for long-term wind condition assessment. Energy Convers Manage 2014;87:697–710.
[13] Rogers AL, Rogers JW, Manwell JF. Comparison of the performance of four measure–correlate–predict algorithms. J Wind Eng Ind Aerodynam 2005;93:243–64.
[14] Carta JA, Velázquez S. A new probabilistic method to estimate the long-term wind speed characteristics at a potential wind energy conversion site. Energy 2011;36:2671–85.
[15] Casella L. Improving long-term wind speed assessment using joint probability functions applied to three wind data sets. Wind Eng 2012;36:473–84.
[16] Probst O, Cárdenas D. State of the art and trends in wind resource assessment. Energies 2010;3:1087–141.

[17] Carta JA, Velázquez S, Matías JM. Use of Bayesian networks classifiers for long-term mean wind turbine energy output estimation at a potential wind energy conversion site. Energy Convers Manage 2011;52:1137–49.

[18] Öztopal A. Artificial neural network approach to spatial estimation of wind velocity. Energy Convers Manage 2006;47:395–406.

[19] Bilgili M, Sahin B, Yasar A. Application of artificial neural networks for the wind speed prediction of target station using reference stations data. Renew Energy 2007;32:2350–60.

[20] Monfared M, Rastegar H, Kojabadi HM. A new strategy for wind speed forecasting using artificial intelligent methods. Renew Energy 2009;34:845–8.

[21] Lopez P, Velo R, Maseda F. Effect of direction on wind speed estimation in complex terrain using neural networks. Renew Energy 2008;33:2266–72.

[22] Velázquez S, Carta JA, Matías JM. Influence of the input layer signals of ANNs on wind power estimation for a target site: a case study. Renew Sustain Energy Rev 2011;15:1556–66.

[23] Bechrakis DA, Deane JP, McKeogh EJ. Wind resource assessment of an area using short term data correlated to a long term data set. Sol Energy 2004;76:725–32.

[24] Haykin S. Neural networks. A comprehensive foundation. New Jersey (USA): Prentice Hall; 1999.

[25] Bishop CM. Neural networks for pattern recognition. New York: Oxford University Press; 1995.

[26] Kalogirou SA. Artificial neural networks in renewable energy systems applications: a review. Renew Sustain Energy Rev 2001;5:373–401.

[27] Fadare DA. The application of artificial neural networks to mapping of wind speed profile for energy application in Nigeria. Appl Energy 2010;87:934–42.

[28] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artif Intell 1997;97:245–71.

[29] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.

[30] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[31] Liu H, Motoda H. Computational methods of feature selection. Chapman & Hall/CRC; 2008.

[32] Leshno M, Lin VY, Pinkus A, Schocken S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Netw 1993;6:861–7.

[33] Witten IH, Frank E. Data mining. 3rd ed. San Francisco: Elsevier; 2011.

[34] Hall MA. Correlation-based feature selection for machine learning. PhD thesis. Hamilton (New Zealand): Department of Computer Science, University of Waikato; 1999. <www.cs.waikato.ac.nz/~mhall/thesis.pdf> [accessed 16.01.15].

[35] Hall MA. Feature selection for discrete and numeric class machine learning. Working paper 99/4; 1999. <www.cs.waikato.ac.nz/ml/publications/1999/99MH-Feature-Select.pdf> [accessed 16.01.15].

[36] Willmott CJ, Robeson SM, Matsuura K. A refined index of model performance. Int J Climatol 2012;32:2088–94.

[37] Fisher NI, Lee AJ. A correlation coefficient for circular data. Biometrika 1983;70:327–32.

[38] Fisher NI. Statistical analysis of circular data. 1st ed. Cambridge: Cambridge University Press; 1995.

[39] Mardia KV. Linear-circular correlation coefficients and rhythmometry. Biometrika 1976;63:403–5.

[40] Johnson RA, Wehrly T. Measures and models for angular correlation and angular-linear correlation. J Roy Stat Soc Ser B 1977;39:222–9.

[41] Bengio Y, LeCun Y. Scaling learning algorithms toward AI. In: Bottou L, Chapelle O, DeCoste D, Weston J, editors. Large-scale learning machines. Cambridge (MA): The MIT Press; 2007. p. 321–59.

[42] Sheela K, Deepa SN. Review on methods to fix number of hidden neurons in neural Networks. Math Probl Eng 2013;2013:11. http://dx.doi.org/10.1155/2013/425740.

[43] Fisher RA. The design of experiments. 2nd ed. London: Oliver & Boyd; 1937.

[44] Pitman EJG. Significance tests which may be applied to samples from any populations. Suppl J Roy Stat Soc 1937;4:119–30.

[45] Good P. Permutation, parametric and bootstrap tests of hypotheses. 3rd ed. New York: Springer; 2005.

[46] Berry KJ, Johnston JE, Mielke PW. A chronicle of permutation statistical methods. 1st ed. New York: Springer; 2014.

[47] Chakraborty R, Pal NR. Feature selection using a neural network framework with controlled redundancy. IEEE Trans Neural Netw Learn Syst 2015;26:35–50.

[48] http://atlante.itccanarias.org/ [accessed 22.03.14].

[49] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc 1995;57:289–300.

[50] Devore JL. Probability and statistics for engineering and sciences. 8th ed. New York: Brooks/Cole; 2012.

[51] Montgomery DC, Runger GC. Applied statistics and probability for engineers. 6th ed. New York: Wiley; 2014.

[52] D'Agostino RB, Stephens MA. Goodness of fit techniques. 1st ed. New York: Dekker; 1986.