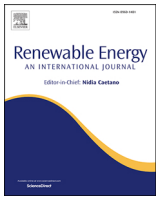
Renewable Energy 261 (2026) 125280

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene



Transformation of reanalysis data for improved long-term estimation of wind speed and direction at a target site

José A. Carta, Pedro Cabrera *

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

ARTICLE INFO

Keywords:
Measure-correlate-predict method
Machine learning techniques
Reanalysis data
Wind speed
Wind direction
Wind power density

ABSTRACT

This paper proposes the use of measure-correlate-predict (MCP) methods based on supervised machine learning (ML) techniques to transform reanalysis data from ERA5 and MERRA2, aiming to improve the long-term estimation of wind speed and direction at locations with limited on-site measurements. The study analyzes models that directly estimate the target variables—wind speed and direction—as well as two-stage models that first estimate the Cartesian components of wind velocity and subsequently transform them into polar coordinates.

As a case study, hourly mean wind data recorded between 2001 and 2023 at 10 m above ground level are used. The data were collected from an anemometric station located on the island of Gran Canaria (Canary Archipelago, Spain).

Key findings include the following: (a) Reanalysis data underestimate actual wind speeds and fail to adequately represent the mean wind direction; (b) although reanalysis data poorly represent the daily wind speed profile, the MCP model significantly corrects this, achieving a Pearson correlation of 0.994; (c) the MCP method minimizes the differences between observed and estimated values (7.2 m/s vs. 7.13 m/s, and 4.49° vs. 4.50°, respectively); (d) the combination of ERA5 and MERRA2 consistently yields the lowest estimation errors, regardless of model type; (e) artificial neural networks outperform other ML techniques in all scenarios; and (f) the proposed method reduces the mean relative error in wind power density estimation to 13.89 %, compared to 43 % and 63.1 % using MERRA2 and ERA5 alone, respectively.

1. Introduction

The accurate estimation of wind resource characteristics at a target site (TS) is essential for energy system planning and the development of economically viable wind farms [1]. When conducting feasibility studies for the installation of a wind farm at a target site, it is necessary to estimate the energy that will be generated by the wind farm over the course of its lifetime [2]. According to Landberg et al. [3], a minimum of 5–10 years of data is required to assess the long-term wind resource. Hiester and Pennell [4] emphasize that at least 10 years of measurements are essential to accurately estimate mean wind power at a target site.

Several authors have highlighted that long observational records are required to characterize the full temporal variability of the wind resource. Burton et al. [5] stress that multi-year and even multi-decadal datasets are desirable for defining a reliable wind climate. Baker et al. [6] quantified annual and seasonal variations in mean wind speed and wind turbine energy production using long-term records. Klink [7] analyzed 22–35-year series at several stations in Minnesota and showed that trends and interannual variability can markedly affect the distribution of wind speeds. Yet, in practice, such long-term datasets are rarely available at target sites, where typically only short-term measurement campaigns—often limited to one year or even a few months—are feasible due to time and cost constraints.

In general, at least one year of data is required to capture the seasonal behaviour of wind. In Brazil, as noted by Miguel et al. [8], the extension of measurement campaigns from 24 to 36 months in 2017 reflects a stronger commitment to capturing wind variability and improving resource assessment accuracy.

The availability of long-term data series is often limited by the high cost of resource measurement campaigns and the urgency to obtain this information within a short timeframe to initiate the necessary procedures for the installation and commissioning of the wind farm. To address this challenge, both the scientific literature and the wind industry have adopted measure-correlate-predict (MCP) methods [9]. To estimate the long-term wind resource at a target site, MCP methods use the

* Corresponding author.
E-mail address: pedro.cabrerasantana@ulpgc.es (P. Cabrera).

https://doi.org/10.1016/j.renene.2026.125280
Received 8 April 2025; Received in revised form 7 December 2025; Accepted 11 January 2026
Available online 20 January 2026

short-term data that is available for that site in combination with long-term data obtained from nearby meteorological stations.

In situations where suitable reference meteorological stations are unavailable, several studies have examined the direct use of reanalysis products for wind-resource assessment. Olauson [10] showed that ERA5 provides substantially improved temporal consistency and reduced bias compared to earlier reanalyses, making it a strong candidate for long-term wind modelling. Rabbani and Zeeshan [11]evaluated MERRA-2 for wind-energy applications in Pakistan and found that, although regional patterns are well reproduced, local discrepancies persist. Pryor and Barthelmie [12] analyzed extreme wind speeds globally and highlighted marked regional differences in reanalysis performance, which are particularly relevant for wind-energy design. Cai and Bréon [13] a [13] ssessed wind-power potential in climate-change scenarios and noted that reanalysis datasets capture large-scale signals but may miss finer-scale features. Görmüş et al. [14] analyzed multi-decadal offshore wind characteristics in the Mediterranean using reanalysis data, demonstrating their usefulness for long-term resource estimation while acknowledging site-specific limitations. Climate reanalyses combine historical measurements from various observation systems with numerical prediction models to produce time series records of multiple climate variables on global or regional three-dimensional grids.

According to Watson [15] and Sheridan et al. [16], two of the most well-known global reanalyses used in renewable energy resource analysis are the Modern-Era Reanalysis for Research and Applications (MERRA2) [17] and ERA5 [18]. MERRA2, produced by NASA's Global Modeling and Assimilation Office, provides hourly wind speeds at three heights above ground level: 2 m, 10 m, and 50 m. ERA5, a global climate reanalysis product developed by the European Centre for Medium-Range Weather Forecasts, offers hourly mean wind speeds at two heights above ground level: 10 m and 100 m.

Gualtieri [19] carried out a critical review of the state of the art on the uncertainties associated with the direct use of reanalysis data for wind resource assessment. The author highlights that reanalysis data (particularly from ERA5) are sufficiently reliable for offshore and flat onshore sites. However, he also indicates that at certain sites the reanalysis data may differ significantly from the actual measurements at the site of interest. Samal [20] compared wind speed data measured on a 50 m mast to data from MERRA2 in the state of Odisha (India), observing significant discrepancies in hourly, monthly, and seasonal variations. However, no study was conducted on wind directions, nor were any suggestions made as to how to improve the usability of the data available in MERRA2 [20].

In this context, MCP methods present an opportunity to establish, over a short training period, the relationship between reanalysis data used as a reference and wind data recorded at a selected ground site. The goal is to transform the reanalysis data to more accurately reflect the real long-term behavior of the wind speed and direction at the selected ground site.

1.1. Aims and originality of this paper

This study proposes a novel approach to reconstructing the historical (long-term) behaviour of wind speed and direction at a target site when only limited short-term measurements are available. The reconstructed series provide a consistent representation of the past wind regime—including its daily, seasonal and interannual variability—which is essential for estimating the long-term energy production of a wind turbine installation. The methodology relies on the use of reanalysis datasets (ERA5 and/or MERRA2) as the reference source from which the long-term series are reconstructed. This is particularly relevant because, although widely used, reanalysis products do not always reproduce local observations accurately and therefore require bias correction before they can be employed with confidence. Within this framework, the proposed approach also enables an assessment, for a given site, of which reanalysis dataset (ERA5, MERRA2 or their combination) is most suitable once corrected through ML-based MCP models. The main contributions and original aspects of the work are summarized below:

i. **Exploration of direct and two-stage MCP models for handling linear and circular variables.** Beyond the conventional two-stage approach—where wind's Cartesian components (x and y) are estimated first and subsequently converted into polar coordinates (magnitude and direction) [21] —this study also employs MCP models that directly predict wind speed (a linear variable) and wind direction (a circular variable [22]). The objective is to determine which strategy (direct or two-stage) is more suitable for estimating wind speed and direction. Within the two-stage framework, two variants are analyzed for the first time: **single-output models**, where each wind component (V_x and V_y) is predicted by an independent model, and **dual-output models**, where both components are estimated simultaneously by a single model. To the best of our knowledge, this comparative analysis has not been previously conducted.
ii. **Evaluation of different ML techniques within MCP methods.** Several of the most commonly applied supervised ML techniques in MCP contexts—RF, SVR, XGB, and ANN—are systematically tested in order to identify which method provides the most effective bias correction and the most robust long-term predictions. This comparative perspective is essential because the choice of ML technique can significantly affect model performance, yet it has received little attention in previous MCP studies using reanalysis data.
iii. **Comprehensive assessment of training period length and selection.** All available years in series (2001–2023) were rotated as training periods, with the remaining years used for testing. An analogous procedure was applied for 2, 3, and 4 years of training (e.g., 2001–2002 for training and 2003–2023 for testing, and so on), enabling an extensive evaluation of how both the duration and the specific choice of the training period affect model performance. The results obtained for 1, 2, 3, and 4 years are presented and discussed, with emphasis on the practical implications for planning measurement campaigns and for the applicability of the models in sites with limited on-site data.

2. Method and meteorological data

A block diagram illustrating the proposed method, covering the process from data collection to result analysis, is shown in Fig. 1.

2.1. Overview of the method

The first task in the process is the collection of data from the selected sources, which in this case include reanalysis data from MERRA2 and ERA5 as well as data recorded by a ground-based anemometric tower.

The second task focuses on comparing the data from the three sources to identify potential discrepancies between them. In this task, wind speed, wind direction, and mean wind power density data from the TS are compared with the corresponding data from the reanalysis sources.

The third task involves the selection of four ML techniques. Using each technique and with data from the reanalysis sources (individually and combined), two types of MCP models are constructed. The first type, referred to as direct prediction models, directly estimate wind speed (a linear variable) or wind direction (a circular variable). The second type, known as two-stage models, firstly predict the Cartesian x and y components of the wind and then determine its polar coordinates (i.e., the modulus of the wind speed and its direction). Within the two-stage approach, we considered two variants: (i) **single-output models**, where independent models are trained for V_x and V_y, and (ii) **dual-output models**, where a single model simultaneously estimates both V_x

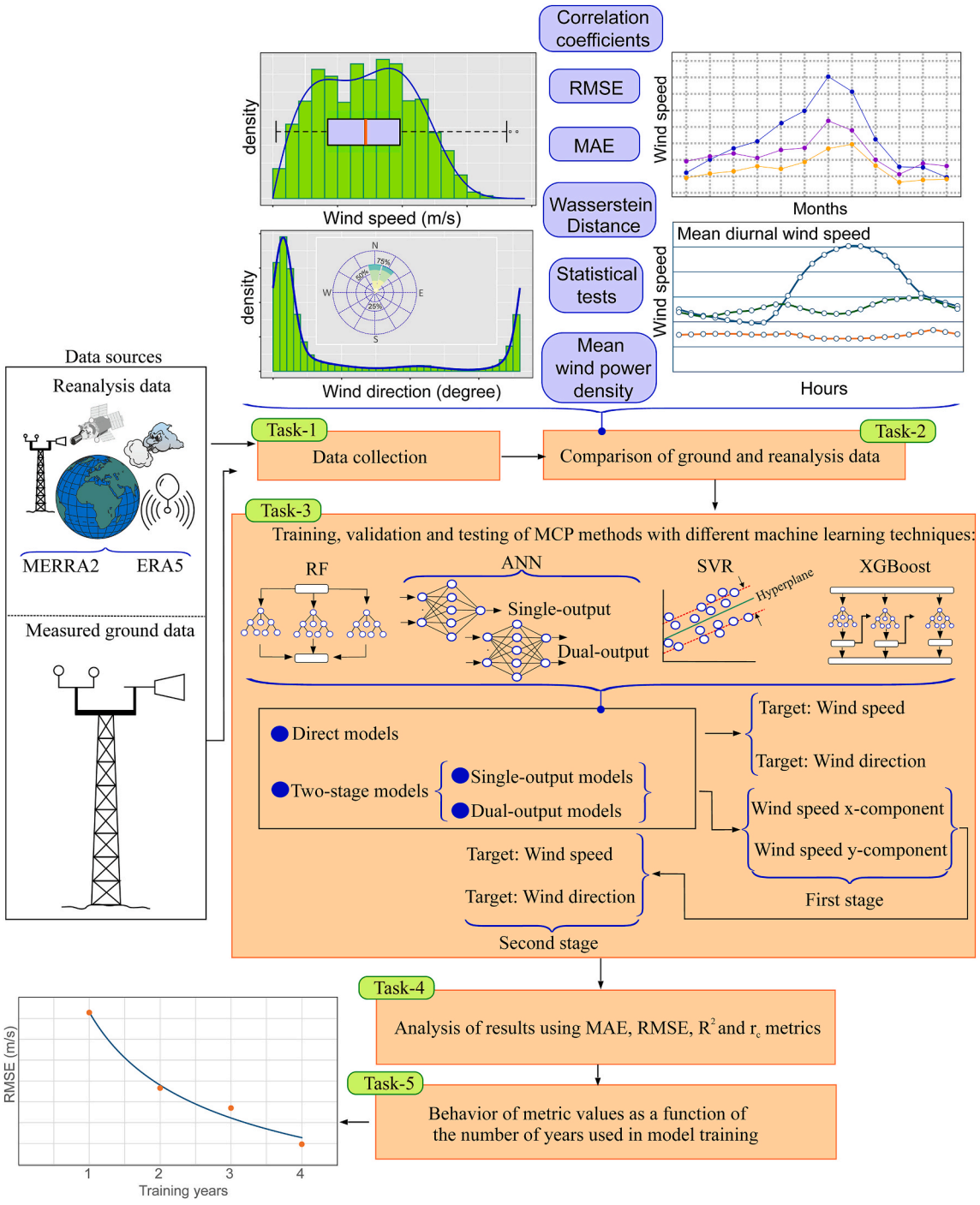


Fig. 1. Schematic representation of the process used to estimate long-term wind speed and direction at a target site using MCP methods with reanalysis data as input.

and V_y.

The fourth task analyzes the error and association metrics obtained from the different models. The fifth task examines the trend in error and association metrics as the number of years used for training and data validation increases.

2.2. Task-1: data collection

The data collected from the reanalysis sources are maintained in their original format. Specifically, for the period from January 1, 2001, to December 31, 2023, the recorded variables include the date (year, month, day, and time) and the Cartesian components of wind speed (V_x and V_y, in m/s) at a height of 10 m above ground level. For the ground-based source, the meteorological variables recorded are wind speeds (V, in m/s) and directions (θ, in degrees), measured using a cup anemometer and a wind vane installed on an anemometer tower at a height of 10 m above ground level. This tower is located on the island of Gran Canaria (Canary Archipelago, Spain) at UTM coordinates 27° 55′ 04" N latitude, 15° 23′ 43" W longitude. If necessary, the Cartesian components of wind speed (V_x and V_y) are converted to polar coordinates (V and θ). North is defined as θ = 0°, and clockwise rotation is considered positive. To calculate V and θ, Eq. (1) and Eq. (2) are used.

$$V = \sqrt{V_x^2 + V_y^2} \tag{1}$$

$$\theta = \begin{cases} \tan^{-1}\left(\frac{V_x}{V_y}\right), if \quad V_y > 0, V_x \geq 0 \\ \frac{\pi}{2}, if \quad V_y = 0, V_x > 0 \\ \tan^{-1}\left(\frac{V_x}{V_y}\right) + \pi, if \quad V_y < 0 \\ \tan^{-1}\left(\frac{V_x}{V_y}\right), if \quad V_y \geq 0, V_x < 0 \\ undefined, if \quad V_y = 0, V_x = 0 \end{cases} \tag{2}$$

Similarly, when using two-stage models, the wind speed modulus (V) recorded at the ground-based anemometer station is decomposed into its Cartesian components: $V_x = V\sin(\theta)$ and $V_y = V\cos(\theta)$.

2.3. Task-2: Comparison of ground and reanalysis data

The comparison of the data obtained from the three sources is performed from four perspectives: a) From the perspective of the frequency distributions of wind speeds and directions; b) From the perspective of the seasonal and daily evolution of wind speed; c) From the perspective of the differences in mean wind power densities; and d) From the perspective of local accuracy.

2.3.1. Comparisons between probability density functions

To facilitate the evaluation of the differences between the probability histograms, the corresponding probability density functions are determined. In this context, the univariate continuous parametric probability density functions used are specific to the two types of variables: wind speed and wind direction:

a) For the wind speed (V), the two-component mixture Weibull distribution, whose probability density function is given by Eq. (3) [23].

$$PDF_v(v, \alpha_1, \beta_1, \alpha_2, \beta_2, \omega) = \omega\left[\frac{\alpha_1}{\beta_1}\left(\frac{v}{\beta_1}\right)^{\alpha_1 - 1} e^{-\left(\frac{v}{\beta_1}\right)^{\alpha_1}}\right] + (1-\omega)\left[\frac{\alpha_2}{\beta_2}\left(\frac{v}{\beta_2}\right)^{\alpha_2 - 1} e^{-\left(\frac{v}{\beta_2}\right)^{\alpha_2}}\right] \tag{3}$$

where $0 \leq \omega \leq 1$ is a mixture parameter, $\alpha_1 > 0$ and $\alpha_2 > 0$ are shape parameters and $\beta_1 > 0$ and $\beta_2 > 0$ are scale parameters.

b) For the circular variables (θ), a finite mixture of $M = 5$ von Mises distributions [24] whose probability density function is given by Eq. (4):

$$PDF_\theta\left(\theta, \kappa_j, \mu_j, \omega_j\right) = \sum_{j=1}^{M} \frac{\omega_j}{2\pi I_o(\kappa_j)} \exp\left[\kappa_j \cos\left(\theta - \mu_j\right)\right]; \; 0 \leq \theta \leq 2\pi \tag{4}$$

where the ω_j are nonnegative quantities that sum to one, Eq. (5), $I_o(\kappa_j)$ is the modified Bessel function of the first kind and order zero [25], and $\kappa_j \geq 0$ and $0 \leq \mu_j \leq 2\pi$ are real parameters.

$$0 \leq \omega_j \leq 1; \;\; (j = 1, \ldots, M) \text{ and } \sum_{j=1}^{M} \omega_j = 1 \tag{5}$$

The maximum likelihood method [23] is used to estimate the parameters of the different distributions. The Cramér-von Mises (CvM) test is used to test the goodness-of-fit of the distributions of circular variables to the experimental histograms, and the Anderson-Darling (A-D) test in the case of the distributions of linear variables [26].

Three tests are employed to compare the global reanalysis wind speed distributions with the global TS wind speed distribution: the A-D test [26], the Kolmogorov-Smirnov (K-S) test [26], and the energy distance (E-D) test for equality of distributions [27]. The E-D metric is particularly useful when the distributions are not normally distributed or have unequal variances. This test calculates a statistic based on the mean distance between points within and between samples [27].

To compare the reanalysis wind direction distributions with the TS wind direction distribution, the same tests are applied, except that the A-D test is replaced with the CvM test [26].

In addition, the Wasserstein distance (WD) metric, also known as the optimal transport distance or earth mover's distance [28], is used to analyze the geometric similarity between the global distributions of wind speeds and directions in the TS data and those provided by ERA5 and MERRA2. If F and G are the cumulative distribution functions (CDFs) of two distributions, the metric is defined as shown in Eq. (6). We set $p = 1$, as our goal is to analyze general differences.

$$W_p(F, G) = \left(\int |F(x) - G(x)|^p dx\right)^{1/p} \tag{6}$$

2.3.2. Comparisons of wind speed seasonal variations and daily means

The degree of correlation between wind speed seasonal variations and daily means is determined using Pearson's correlation coefficient, as defined in Eq. (7). This metric is widely used in MCP methods [9].

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \frac{\sum_{i=1}^{n} x_i}{n}\right)\left(y_i - \frac{\sum_{i=1}^{n} y_i}{n}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \frac{\sum_{i=1}^{n} x_i}{n}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \frac{\sum_{i=1}^{n} y_i}{n}\right)^2}} \; r \begin{cases} = 1 & \text{Perfect positive linear relationship} \\ = -1 & \text{Perfect negative linear relationship} \\ = 0 & \text{No linear relation (though non} - \text{linear relations may exist).} \end{cases} \tag{7}$$

In Eq. (7), x_i and y_i are the individual values of two variables X and Y, respectively.

2.3.3. Comparison of mean wind power densities

The mean wind power densities ($\overline{WPD}$) are estimated using Eq. (8) [29].

$$\overline{WPD} = \frac{1}{2n}\sum_{i=1}^{n} \rho_i v_i^3 \tag{8}$$

In Eq. (8), ρ_i are the air densities. Most authors [30,31]use Eq. (8), assuming that air density is constant over time, and employ the standard value of 1.225 kg m^{-3}, corresponding to standard atmospheric conditions (completely dry air, and mean pressure and air temperature at sea level of 1013.25 hPa and 15 °C, respectively).

2.3.4. Analysis of the accuracy of wind speed and direction data

The errors generated when representing the TS variable data using reanalysis variable data are analyzed. For linear target variables (wind

speed), the error and association metrics used are the mean absolute error (MAE), Eq. (9), the root mean squared error (RMSE), Eq. (10), and the coefficient of determination (R^2), Eq. (11), as these metrics are widely used in wind resource estimation [21]. The RMSE is used to evaluate large deviations, while the MAE is more robust to outliers. In this task, R^2 serves as a measure of how much of the variability in the TS data is explained by the reanalysis data.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \frac{\sum_{i=1}^{n} y_i}{n}\right)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

In Eq. (9), Eq. (10), and Eq. (11), y_i is the observed (true) value, $\hat{y}_i$ is the predicted value, and n is the total number of observations.

For circular target variables (wind direction) defined within the range [0°,360°], it is crucial to use specific error metrics that account for the circular nature of the data. To address this, the Circular_Difference function is defined in Eq. (12). Based on Eq. (12), custom metrics, referred to as RMSE_Circular (Eq. (13)) and MAE_Circular (Eq. (14)), are introduced.

$$\text{CircularDifference}(\theta_i, \hat{\theta}_i) = min(|\theta_i - \hat{\theta}_i|, 360^\circ - |\theta_i - \hat{\theta}_i|) \tag{12}$$

$$\text{RMSE}_{\text{Circular}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\text{Circular}_{\text{Difference}(\theta_i,\hat{\theta}_i)}\right]^2} \tag{13}$$

$$\text{MAE}_{\text{Circular}} = \frac{1}{n}\sum_{i=1}^{n}\left|\text{Circular}_{\text{Difference}(\theta_i,\hat{\theta}_i)}\right| \tag{14}$$

In Eq. (12), Eq. (13) and Eq. (14), θ_i is the observed (actual) value, $\hat{\theta}_i$ i is the predicted value, and n the total number of observations. The degree of association between wind directions is estimated with the circular-circular correlation coefficient of Jammalamadaka and Sarma [32], Eq. (15). The original formulation was proposed by Jammalamadaka & Sarma [32], and is described in detail by Jammalamadaka & SenGupta **(2001)** [33].

$$r_c = \frac{\sum_{i=1}^{n}\sin(\phi_i - \overline{\phi})\cdot\sin(\varphi_i - \overline{\varphi})}{\sqrt{\left(\sum_{i=1}^{n}\sin^2(\phi_i - \overline{\phi})\right)\cdot\left(\sum_{i=1}^{n} sin^2((\varphi_i - \overline{\varphi}))\right)}} \tag{15}$$

In Eq. (15), $(\phi_1, \varphi_1), \ldots, (\phi_n, \varphi_n)$ represent the samples of n data from the directions (in radians) of two sources and $\overline{\phi}$ and $\overline{\varphi}$ are the sample mean directions, Eq. (16), [33].

$$\overline{\phi} = \text{atan}\,2\left(\sum_{i=1}^{n} sin(\phi_i), \sum_{i=1}^{n} cos(\phi_i)\right); \overline{\varphi} = \text{atan}\,2\left(\sum_{i=1}^{n} sin(\varphi_i), \sum_{i=1}^{n} cos(\varphi_i)\right) \tag{16}$$

These global error and association metrics (MAE, RMSE, MAE_Circular, RMSE_Circular and r_c) are standard in MCP applications. They quantify how effectively the MCP models reduce the discrepancies between reanalysis datasets and local observations and are widely used in the literature to evaluate the performance of MCP-based reconstructions of long-term wind conditions

2.4. Task-3: Training, test and validation of the different MCP methods based on ML techniques

The proposed models for estimating the target variables were developed using multiple regression, Eq. (17).

$$Y_t = f(X_t) = f\left(\overbrace{V_{x,t}, V_{y,t}, V_t}^{MERRA2}, \overbrace{V_{x,t}, V_{y,t}, V_t}^{ERA5}, H_c, H_s, S_c, S_s\right) \tag{17}$$

In the functional forms of the model, $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ are the input variables, the subscript t indicates the instant evaluated, and $Y_t = (V, \theta, V_x \text{ or } V_y)$ represents the estimated response variable V, θ, V_x or V_y of the TS. H_c and H_s, are harmonic transformations of the hour of the day (H), while S_c and S_s denote harmonic transformations of the seasonal cycle. These transformations allow the models to capture cyclic patterns and avoid artificial discontinuities that could be misinterpreted by ML techniques.

Two alternative formulations were considered for the seasonal cycle: Eq. (18), based on the month M, i.e. a monthly harmonic, and Eq. (19), based on the day of year d, i.e. a daily harmonic.

$$H_c = \cos\left(\frac{2\pi H}{24}\right);\ H_s = \sin\left(\frac{2\pi H}{24}\right);\ S_c = \cos\left(\frac{2\pi M}{12}\right);\ S_s = \sin\left(\frac{2\pi M}{12}\right) \tag{18}$$

$$H_c = \cos\left(\frac{2\pi H}{24}\right);\ H_s = \sin\left(\frac{2\pi H}{24}\right);\ S_c = \cos\left(\frac{2\pi d}{365}\right);\ S_s = \sin\left(\frac{2\pi d}{365}\right) \tag{19}$$

The function $Y_t = (V, \theta, V_x \text{ or } V_y)$ is valid for estimating each of the target variables. Input variables can be drawn from a single reanalysis data source (MERRA2 or ERA5) or simultaneously from both data sources (MERRA2 and ERA5).

A broader discussion of alternative methodological options, including additional variables, spatial extensions, and other complementary approaches, is provided in Section 4 (Limitations).

Fig. 2 presents a block diagram that schematizes the training, validation, and testing processes for the various MCP models using the different ML techniques considered.

The process is summarized in three steps, each represented by a number enclosed in a circle. The first step involves determining the optimal hyperparameters for the model under consideration. The data are divided into K = 5 folds to train and evaluate the model using cross-validation, ensuring robustness and minimizing the risk of overfitting. In each iteration, one-fold is used as the validation set, while the remaining folds are used for training. Next, the model is defined using the selected ML technique, and the error metric to be employed in the training and validation process of the MCP model is specified.

The proposed ML techniques are as follows [34]: random forest (RF), selected for its robustness against overfitting [35] and its strong performance in previous MCP studies [36]; extreme gradient boosting (XGBoost), chosen for its high accuracy and computational efficiency [34]; support vector regression (SVR), included for its effectiveness in solving regression problems with complex and nonlinear relationships, as well as its strong results in previous MCP research [29]; and artificial neural network (ANN), utilized for its ability to learn complex patterns and its extensive application in MCP problems [37].

In the case of circular target variables, the custom metric RMSE_Circular, Eq. (12), is used to account for the circular nature of the target variables. This metric is directly employed as an evaluation criterion during the model training and validation process.

For each ML technique used, the hyperparameter search space is defined, meaning that key values are explored to optimize model

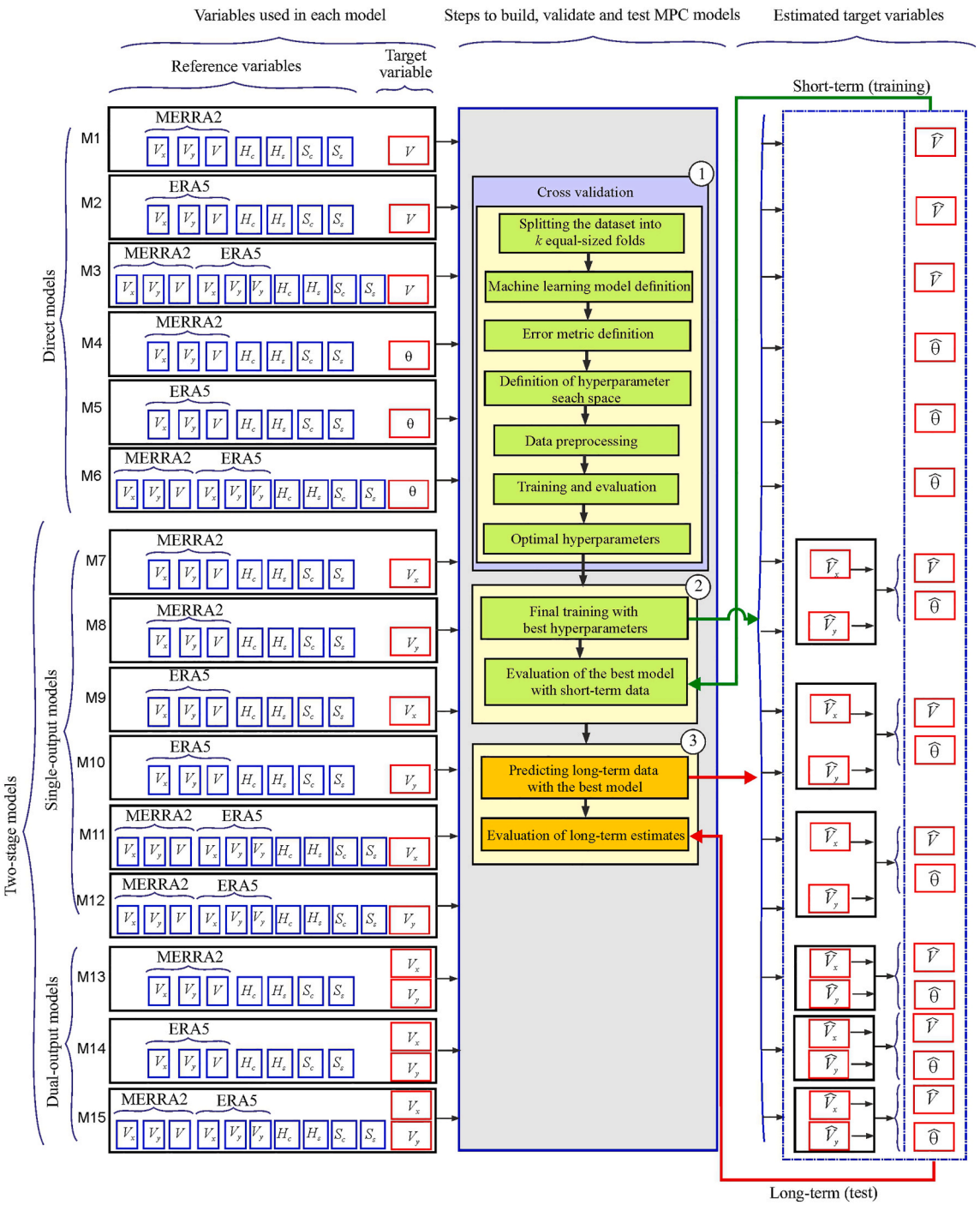


Fig. 2. General configuration of the methodological procedure followed.

performance. In the case of RFs, the hyperparameters explored include the number of trees (num.trees), the number of randomly selected variables in each split (mtry), and the maximum depth of the trees (max. depth) [38].

In the case of XGBoost, the hyperparameters defined include the number of trees (nrounds), learning rate (eta), maximum tree depth (max_depth), instance and feature subsampling (subsample, colsample_bytree), and regularization parameters (gamma, lambda, and alpha) [38]. For SVR, different values of the key parameters (C and σ) are tested using a predefined grid [38]. For ANN, combinations of hyperparameters are explored, including the number of hidden layers (between 1 and 3), the number of neurons per layer (between 10 and 200), the dropout rate in each layer (between 0 and 0.5), the learning rate (0.001, 0.0001, or 0.00001), and the number of epochs (with a maximum value of 1000). The random selection of hidden neurons is programmed to follow a pyramidal shape, where the number of neurons decreases from the first to the last layer. This structure adheres to the so-called geometric pyramid rule [39]. The early stopping parameters are also defined as follows: patience, the number of epochs without improvement before stopping training, is set to 10; and min_delta, the minimum change in RMSE improvement, or RMSE_Circular, considered significant, is set to 10^{-3}.

The artificial neural networks used in this study are fully connected feed-forward multilayer perceptrons. Hidden layers employ the ReLU activation function, while the output layer is linear for V, V_x, and V_y, and adapted to circular metrics for θ. Training was performed using the

Adam optimizer together with early stopping, as described above.

To ensure comparability across ML techniques, hyperparameter combinations were generated through a randomized but controlled procedure and stored in catalogs for reproducibility. This guarantees that each technique was explored under equivalent conditions, avoiding biases due to unequal search spaces. In the case of ANN, catalogs were created separately for the direction task and for the other targets, with adjusted ranges, while maintaining the pyramidal constraint on the number of neurons.

In XGBoost and RF, data standardization is not required due to the nature of the underlying algorithms. However, for SVR and ANN data preprocessing is necessary to ensure that the predictor variables are on the same scale and to facilitate optimal model performance. For these models, the training data are standardized (centered and scaled), and the same transformation parameters are applied to the validation and test data. In addition, for ANN and SVR the target variables were also standardized during training and subsequently rescaled to their original units for the computation of evaluation metrics.

Initially, each target variable (V, θ, V_x, V_y) was estimated with an independent model, in order to maintain methodological consistency across all ML techniques. This choice is particularly relevant because V is a linear variable, whereas θ is circular and requires dedicated loss functions and validation metrics, which are not natively implemented in most ML frameworks.

Each ML technique optimizes a specific loss function during training. For RF, the splitting criterion is based on minimizing the variance within nodes, which is equivalent to squared error loss in regression tasks. For XGB, the default squared error objective was used. For SVR, training is based on the ε–insensitive loss, which penalizes deviations larger than ε while ignoring smaller residuals. For ANN, the loss function corresponds to the mean squared error (MSE) for linear variables, while for circular variables (θ) a circular loss based on the minimum angular difference between observed and predicted values was implemented. In the case of dual-output ANN models (V_x and V_y), the losses of both outputs were combined either as an unweighted mean or as a variance-weighted mean, as detailed above.

In the case of ANN, however, we also implemented **dual-output models** for V_x and V_y in order to directly compare their performance against the single-output approach. For these dual-output ANN models, two alternative strategies were tested for combining the losses of both outputs: Unweighted mean, Eq. (20) and Variance-weighted mean, Eq. (21).

$$\mathscr{L} = \frac{1}{2}\left(\mathscr{L}_{V_x} + \mathscr{L}_{V_y}\right) \tag{20}$$

$$\mathscr{L} = \frac{\sigma^2_{V_x}}{\sigma^2_{V_x} + \sigma^2_{V_y}}\mathscr{L}_{V_x} + \frac{\sigma^2_{V_y}}{\sigma^2_{V_x} + \sigma^2_{V_y}}\mathscr{L}_{V_y} \tag{21}$$

where $\mathscr{L}_{V_x}$ and $\mathscr{L}_{V_y}$ denote the mean squared error for each component, and $\sigma^2_{V_x}$ and $\sigma^2_{V_y}$ are their sample variances in the training set. The second formulation ensures that the more variable component exerts proportionally greater influence during optimization.

While ANNs allow for true multi-output designs with shared parameters, the situation differs for the other ML techniques considered. **RF** in *scikit-learn* natively support multi-output regression, since each tree leaf can store a vector of values. Nevertheless, this option was not used here in order to maintain methodological comparability across techniques. For **SVR** and **XGBoost**, no native multi-output regression implementation is available. Their extension to multi-output relies on wrappers such as MultiOutputRegressor, which simply train an independent model for each target variable without parameter sharing or reduction in overall complexity. For this reason, in this study the multi-output formulation was restricted to ANNs only.

Although this study tested dual-output ANN models for the linear components V_x and V_y, no joint multi-output design was implemented for V and θ. This decision is motivated by their heterogeneous statistical nature: V is a linear variable, usually modeled with squared-error loss, whereas θ is circular and requires specialized losses based on angular differences. Combining both targets in a single training process would demand a careful weighting or normalization of such heterogeneous losses, and may even require transforming θ into its Cartesian components (cosθ, sinθ) to avoid discontinuities. Given these technical challenges, and to ensure methodological comparability across ML techniques, V and θ were modeled independently in this study.

In step 2, the model is trained with the best parameters and evaluated using all available short-term data. For the evaluation, we use the metrics MAE, (Eq. (8)), RMSE (Eq. (9)), and R^2 (Eq. (10)) for the linear target variables, and MAE_Circular (Eq. (13)), RMSE_Circular (Eq. (12)), and r_c (Eq. (14)) for the circular target variables. Table A.1 in Appendix A lists the main R libraries used to define and train each ML technique. In step 3, the best model (with the best parameters) is applied to estimate the values of the long-term target variable.

2.5. Task-4: Analysis of results using statistics metrics

The differences in the metric values obtained by the different MCP models analyzed are evaluated.

In addition to the comparative analysis from the perspective of local accuracy, this task includes, as in Task-2, a comparison between the data

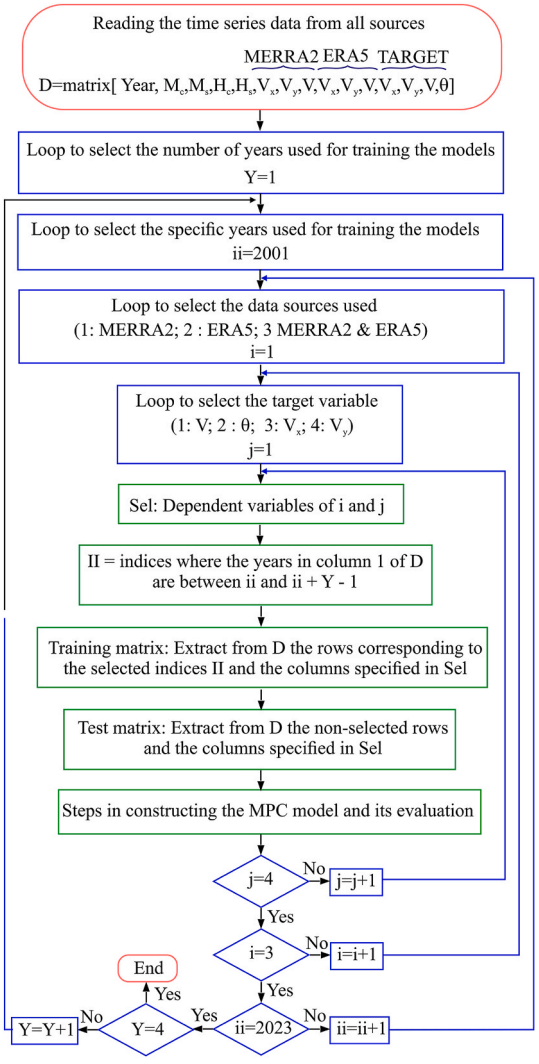


Fig. 3. Algorithm used to select training and test years, reference data sources and models trained, validated and tested with each ML technique.

measured at the TS and the data estimated by the best MCP models. The comparison is made from the following perspectives: (a) frequency distributions of wind speeds and directions; (b) seasonal and daily evolution of wind speed; and (c) differences in mean wind power densities.

2.6. Task-5: Analysis of performance

The steps outlined in Fig. 2 are performed using one year of data for training (short term) and the remaining years for testing (long term).

Fig. 3 shows the algorithm employed for this process, which achieves its objective when Y = 1. In this task, the goal is to analyze the effect on the error and association metrics of the number of years used for training and validation. The algorithm depicted in Fig. 3 represents the procedure followed to achieve this, with the objective reached when Y = 4.

3. Results and discussion

This section presents the results of the analyses conducted based on the tasks described in the methodology, as outlined in Fig. 1.

3.1. Task-2: Comparison of the data collected in the three data sources (ground and two different reanalysis data)

The following subsections present and analyze the results obtained in Task-2 from the perspectives outlined in Fig. 1.

3.1.1. Probability density functions of wind speed and wind direction

The comparison between the probability distributions of wind speeds at the TS and those derived from ERA5 and MERRA2 (Fig. 4) reveals statistically significant differences (Table A.2). The observed series exhibits a bimodal regime, whereas both reanalyses show unimodal distributions with lower means (7.2 m/s at the TS, 6.6 m/s in MERRA2, and 5.5 m/s in ERA5). These results demonstrate that both reanalyses systematically underestimate the mean wind speed at the site.

Linear correlations between reanalysis data and the TS are moderate ($r \approx 0.72$–0.74), while ERA5 and MERRA2 are highly correlated with each other ($r = 0.92$). It should be noted that these correlations were calculated from the hourly wind speed series used to construct the histograms, rather than directly from the histograms.

Regarding wind direction, Fig. 5 shows marked discrepancies between reanalyses and on-site measurements, with predominant patterns poorly represented. Table A.3 reports the parameters of the fitted wind direction distributions and the *p*-values of the goodness-of-fit tests. The circular–circular correlation coefficients are very low, and in some cases even negative, between MERRA2 and ERA5 and between MERRA2 and the TS. The coefficient proposed by Jammalamadaka and SenGupta [33] is a robust measure: small or negative values indicate the absence of a simple circular–linear relationship, although they do not rule out more complex dependencies influenced by other factors [22].

Detailed statistical tests (A–D, K–S, E–D, and CvM) and distance metrics (WD) quantifying differences between distributions are provided in Appendix A (Tables A.4 and A.5). These results confirm that the ERA5 and MERRA2 distributions differ significantly from those observed at the TS. Altogether, these discrepancies highlight the limitations of using raw reanalysis data to represent local wind conditions and provide the rationale for applying ML-based MCP models in the following tasks.

3.1.2. Wind power densities

The absolute percentage error (APE) obtained between the ($\overline{WPD}$) calculated using TS wind speeds and those calculated with MERRA2 and

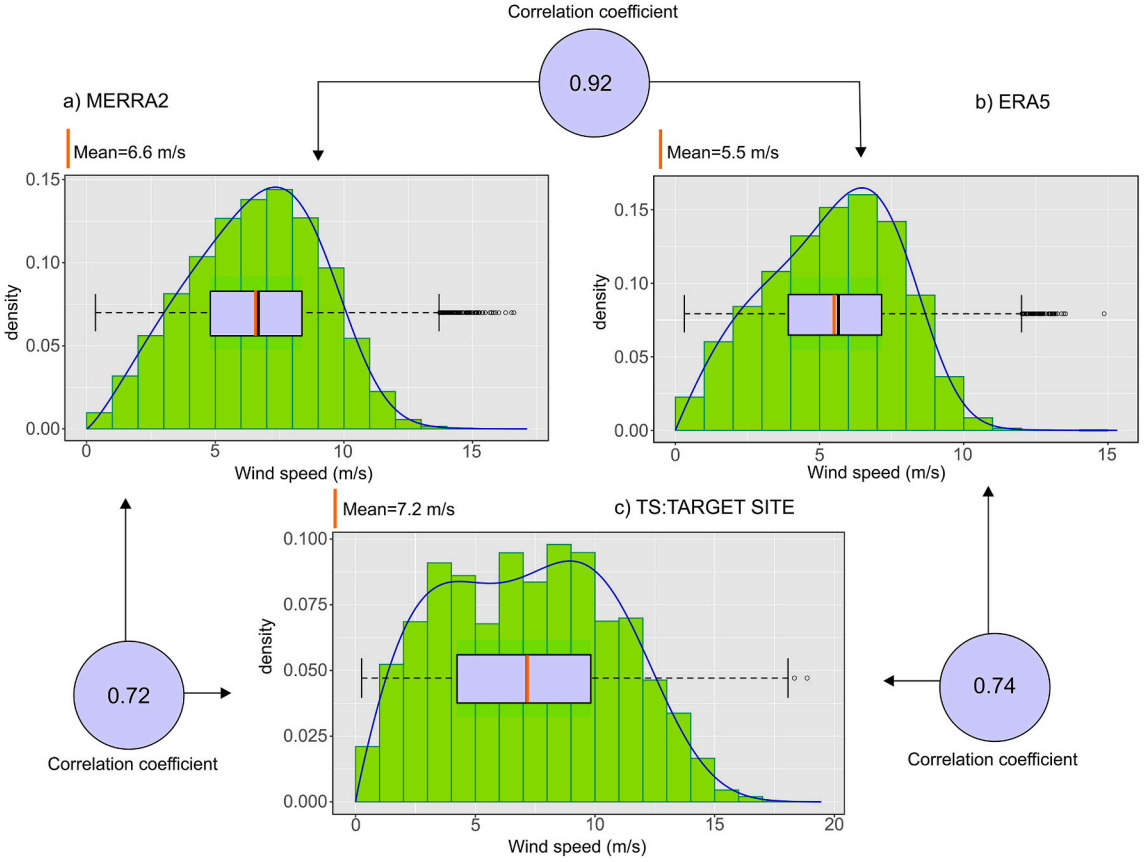


Fig. 4. Histograms and probability density functions of wind speeds recorded in: (a) MERRA2, (b) ERA5, and (c) the target site. The correlations were calculated from the hourly wind speed series used to construct the histograms.

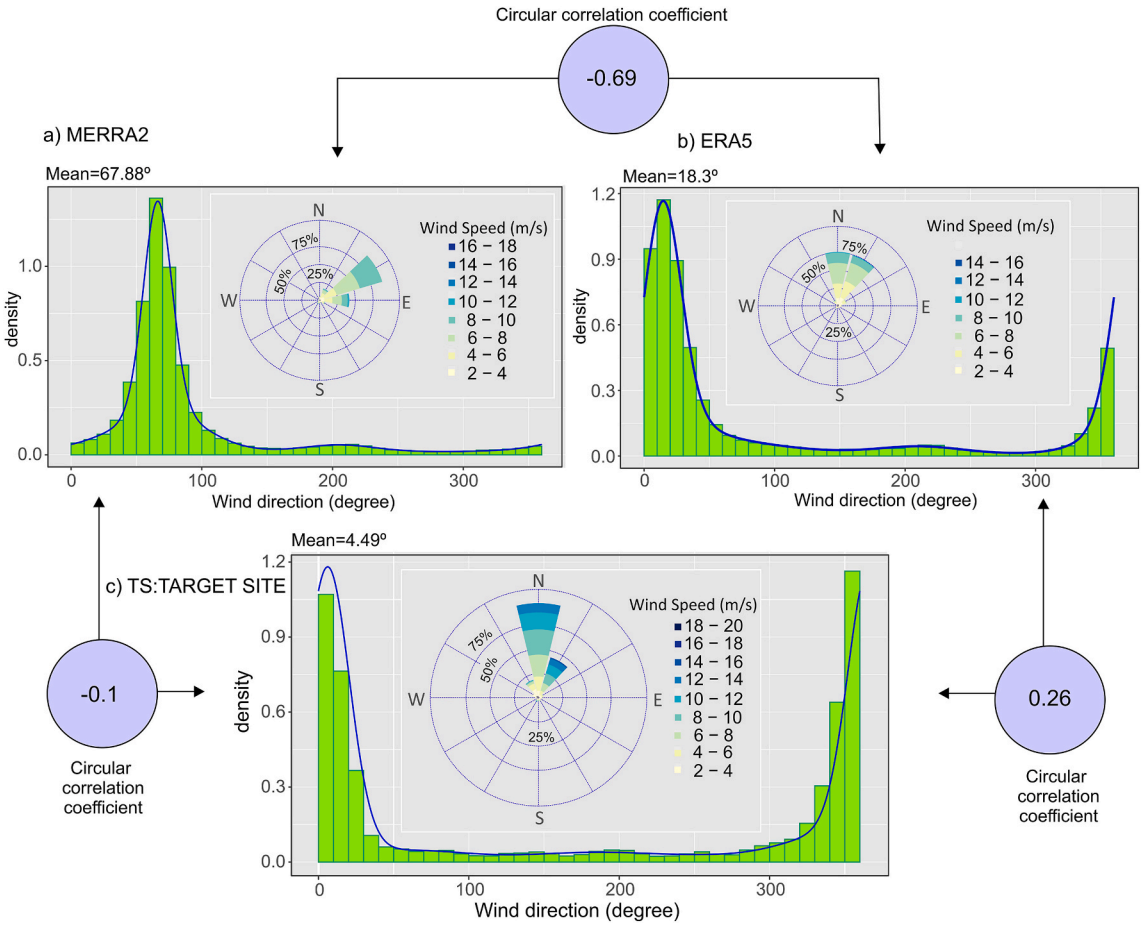


Fig. 5. Histograms and probability density functions of wind directions recorded in: (a) MERRA2, (b) ERA5, and (c) the target site. The correlations were calculated from the hourly wind direction series used to construct the histograms.

ERA5 wind speeds (Eq.(21)) were 39 % and 63.1 %, respectively. The mean $\overline{WPD}$ is an important indicator of wind energy potential and is commonly included in regional wind resource maps as valuable preliminary information to identify potentially attractive sites for wind project installations [29]. In this context, the APE values indicate significant percentage differences between the reanalysis data and the TS data. These findings align with the results reported by Samal [20] in his study using MERRA2 data and measurements recorded in the Indian state of Odisha.

$$APE = \left| \overline{WPD}_{TS} - \overline{WPD}_{Reanalysis(MERRA2\ or\ ERA5} \frac{}{\overline{WPD}_{TS}} \right| \times 100 \quad (21)$$

3.1.3. Seasonal and daily wind speed variation

Fig. 6 shows the mean daily wind speeds derived from the three data sources. The Pearson correlation coefficients reveal a very weak relationship between the reanalysis datasets and the TS, indicating that MERRA2 and ERA5 do not reproduce the actual daily wind speed profile. This emphasizes the need for model-estimated data to approximate the real daily cycle and to capture seasonal patterns more accurately.

Such improvements are essential in applications where the hourly wind-power profile must be realistically represented. One study [40] demonstrated that accurate hourly wind data are crucial for the optimal sizing of stand-alone wind-powered desalination systems. Another work [41] showed that realistic temporal wind profiles are also required when assessing the carbon footprint of desalination processes in island grids with limited flexibility.

The seasonal evolution of monthly mean wind speeds is presented in Fig. 7. All three data sources show the highest values in June, July, and August, but during these months the differences between reanalysis and TS are most pronounced. In general, reanalysis data consistently underestimate the observed wind speeds. Nevertheless, the Pearson correlation coefficients between the monthly mean values at the TS and those from MERRA2 and ERA5 are 0.929 and 0.949, respectively, confirming a strong relationship in the representation of seasonal variability.

These results confirm that raw reanalysis data fail to reproduce the local daily wind cycle and systematically underestimate seasonal wind speeds, even though the overall seasonal trends are well captured. Such discrepancies highlight the need for bias-correction methods, which will be addressed through the ML-based MCP models evaluated in the following sections.

3.1.4. Local accuracy

Table 1 shows the error and correlation metrics obtained by comparing reanalysis wind speed and direction data with TS measurements. For wind speeds, ERA5 shows higher MAE and RMSE values than MERRA2, but also a slightly higher R^2 (54.3 % vs. 52.0 %). For wind directions, ERA5 achieves lower MAE and RMSE errors and a higher correlation coefficient rcr_crc than MERRA2.

These results indicate that, although both reanalysis datasets contain useful information, their direct use is affected by substantial errors in both speed and direction. This reinforces the need for bias-correction methods to fully exploit the potential of reanalysis data. In the next sections, ML-based MCP models are applied precisely to address these discrepancies.

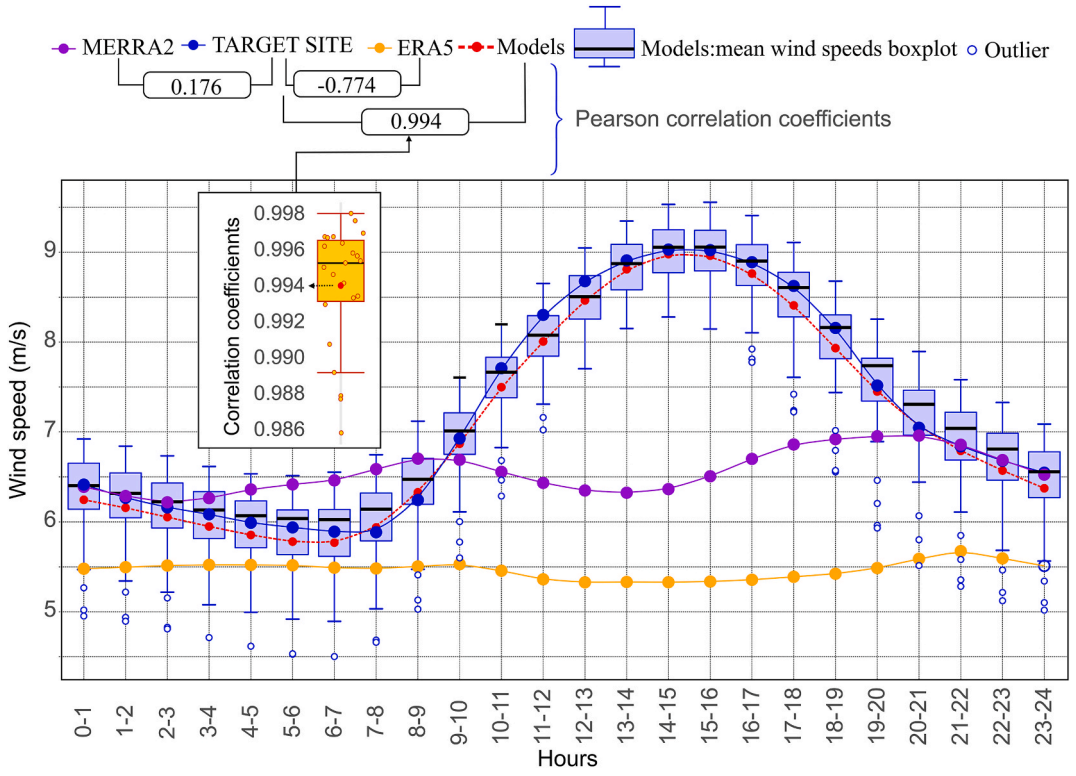


Fig. 6. Mean daily wind speeds calculated using the three data sources and the best MCP model.

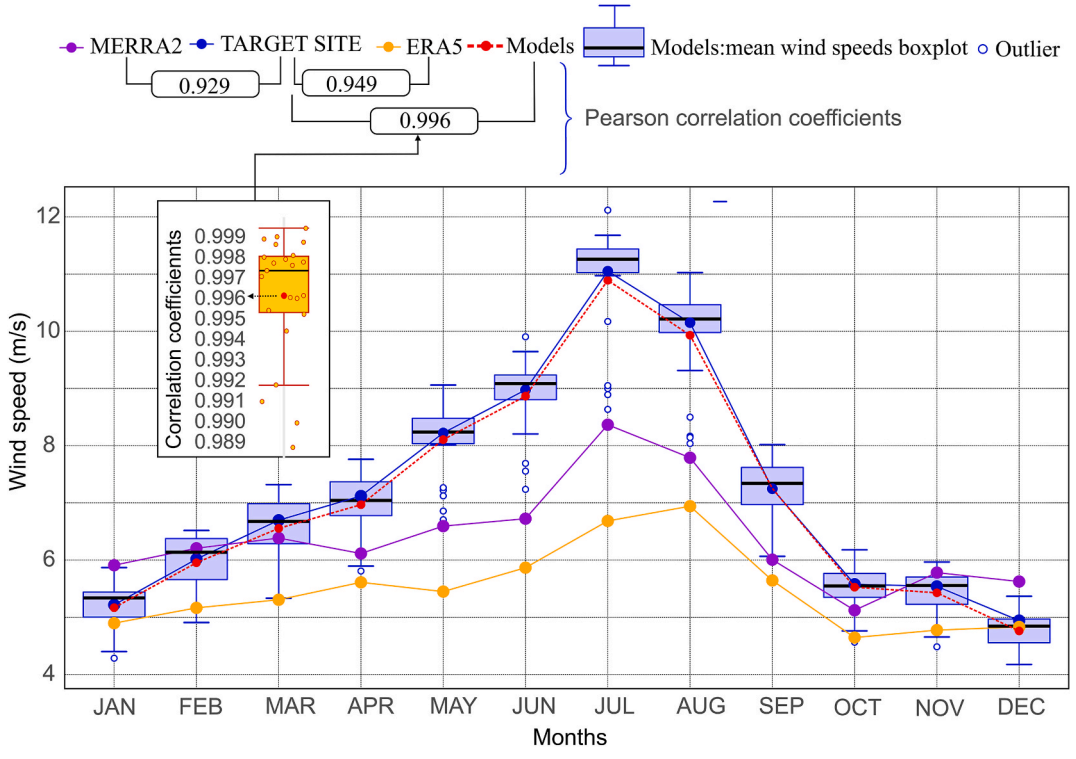


Fig. 7. Seasonal variation of mean wind speeds calculated using the three data sources and the best MCP model.

3.2. Training, test and validation of the different MCP methods using ML (Task −3)

Table 2 presents the mean values of the metrics obtained during the testing process (long term) for the two types of MCP models (direct and single-output two-stage models) and the three data sources (MERRA2, ERA5, and MERRA2 & ERA5) using each of the ML techniques considered (RF, XGBoost, SVR, and ANN).

Artificial neural networks (ANN) consistently achieve the lowest errors (MAE, RMSE, MAE_Circular, RMSE_Circular) and the highest

Table 1
Error and association metrics between TS data and reanalysis data.

Sources	Wind speeds			Wind directions		
	MAE (m/s)	RMSE (m/s)	R^2 (%)	MAE (degree)	RMSE (degree)	r_c
MERRA2-TS	2.05	2.61	52.00	67.3	75.1	−0.0967
ERA5-TS	2.38	3.05	54.30	27.62	43.11	0.2575

coefficients of determination (R^2 and r_c) in almost all configurations, confirming their superior ability to capture the nonlinear relationships between reanalysis inputs and local wind conditions. While RF and XGBoost occasionally perform comparably for wind speed estimation in direct models (as confirmed by Wilcoxon tests), ANN remains the only technique that systematically excels for both linear and circular targets.

For wind speeds, direct models tend to provide slightly lower global errors, while two-stage models capture directional behavior more accurately, yielding higher circular correlations. This complementarity highlights the importance of evaluating both approaches when correcting reanalysis biases.

An additional point of interest concerns the relative behavior of ERA5 and MERRA2. Before post-processing, MERRA2 wind speeds are closer to the TS values than those of ERA5. However, once corrected with ANN-based MCP models, ERA5 systematically outperforms MERRA2, both in direct and two-stage formulations. This inversion highlights the fact that raw accuracy of reanalysis data does not necessarily translate into better performance after bias correction. It also explains why ERA5-based methods generally surpass MERRA2-based ones in the corrected results. Whether this behavior is specific to the conditions of the present site or represents a broader pattern is an open question. Comparative studies at other locations with different climatic and topographic settings would be needed to determine its generality, which constitutes an important direction for future research.

The analysis of the hyperparameter catalogs (Appendix A, Table A.6) reveals systematic patterns:

a) Models estimating wind direction required, on average, larger hidden layers than those estimating wind speed, reflecting the added complexity of circular targets.
b) Two-stage models typically needed more neurons than direct models for wind speed estimation, likely because they must reconstruct intermediate Cartesian components before converting them to polar form.

Table 2
Mean values of error and correlation metrics obtained during the testing process (long-term) for different ML techniques (RF, XGBoost, SVR, ANN), model types (direct and two-stage), and reanalysis sources (MERRA2, ERA5, and their combination). The upper half of the table reports results for **wind speed** (MAE, RMSE, R^2), while the lower half corresponds to **wind direction** (MAE, RMSE, circular correlation coefficient r_c). Boldface indicates the best performance within each metric set. The last three columns show the adjusted p-values from Wilcoxon tests comparing ANN with the other ML techniques.

Model type	Source data	Target	Metric	Machine learning technique				Test. *p*-values		
				RF	XGBoost	SVR	ANN	RF/ANN	XGBoost/ANN	SVR/ANN
Direct models	MERRA2	V	MAE (m/s)	1.49	1.49	1.62	**1.48**	0.324	0.453	<0.001
	ERA5			1.42	**1.40**	1.54	**1.40**	0.048	0.800	<0.001
	MERRA2&ERA5			**1.38**	1.40	1.64	1.39	0.700	0.209	<0.001
	MERRA2		RMSE (m/s)	1.91	1.91	2.08	**1.89**	0.016	0.068	<0.001
	ERA5			1.83	1.81	1.99	**1.79**	0.001	0.133	<0.001
	MERRA2&ERA5			1.78	1.80	2.10	**1.77**	0.481	0.003	<0.001
	MERRA2		R^2 (%)	75.35	75.25	70.10	**75.79**	0.012	0.064	<0.001
	ERA5			77.50	78.01	72.65	**78.43**	<0.001	0.076	<0.001
	MERRA2&ERA5			78.75	78.16	69.47	**79.00**	0.133	0.002	<0.001
Two-stage models	MERRA2	V	MAE (m/s)	1.55	1.57	1.76	**1.51**	<0.001	<0.001	<0.001
	ERA5			1.47	1.48	1.64	**1.43**	<0.001	<0.001	<0.001
	MERRA2&ERA5			1.44	1.45	1.75	**1.41**	0.007	0.001	<0.001
	MERRA2		RMSE (m/s)	2.01	2.03	2.32	**1.94**	<0.001	<0.001	<0.001
	ERA5			1.92	1.93	2.18	**1.84**	<0.001	<0.001	<0.001
	MERRA2&ERA5			1.87	1.88	2.32	**1.81**	<0.001	<0.001	<0.001
	MERRA2		R^2 (%)	73.99	73.43	65.41	**76.12**	<0.001	<0.001	<0.001
	ERA5			76.25	75.95	69.33	**78.48**	<0.001	<0.001	<0.001
	MERRA2&ERA5			77.53	77.26	65.92	**79.14**	<0.001	<0.001	<0.001
Direct models	MERRA2	θ	MAE (degree)	71.56	69.20	52.76	**20.28**	<0.001	<0.001	<0.001
	ERA5			69.64	67.54	51.93	**19.71**	<0.001	<0.001	<0.001
	MERRA2&ERA5			71.63	70.76	59.60	**19.72**	<0.001	<0.001	<0.001
	MERRA2		RMSE (degree)	87.73	86.24	71.17	**34.74**	<0.001	<0.001	<0.001
	ERA5			86.40	84.93	69.78	**33.84**	<0.001	<0.001	<0.001
	MERRA2&ERA5			88.03	87.64	74.79	**33.94**	<0.001	<0.001	<0.001
	MERRA2		r_c	0.001	0.224	0.210	**0.610**	<0.001	<0.001	<0.001
	ERA5			0.083	0.248	0.179	**0.627**	<0.001	<0.001	<0.001
	MERRA2&ERA5			−0.04	0.105	0.046	**0.619**	<0.001	<0.001	<0.001
Two-stage models	MERRA2	θ	MAE (degree)	19.93	20.33	23.78	**19.24**	<0.001	<0.001	<0.001
	ERA5			19.14	19.38	22.46	**18.49**	<0.001	<0.001	<0.001
	MERRA2&ERA5			18.64	18.99	24.37	**18.30**	<0.001	<0.001	<0.001
	MERRA2		RMSE (degree)	34.42	35.18	41.72	**33.75**	<0.001	<0.001	<0.001
	ERA5			33.20	33.71	39.68	**32.56**	<0.001	<0.001	<0.001
	MERRA2&ERA5			32.56	33.14	43.28	**32.34**	<0.001	<0.001	<0.001
	MERRA2		r_c	0.608	0.605	0.562	**0.631**	<0.001	<0.001	<0.001
	ERA5			0.626	0.626	0.586	**0.647**	<0.001	<0.001	<0.001
	MERRA2&ERA5			0.633	0.627	0.562	**0.649**	<0.001	<0.001	<0.001

c) Dropout regularization and learning rates varied moderately across tasks, but no single configuration dominated, indicating that ANN performance is robust across a range of parameterizations.
d) The number of epochs required for convergence (Appendix A, Fig. A1) was generally higher for direct models estimating wind speed, suggesting a more gradual learning process compared to other tasks.

To further investigate methodological alternatives, dual-output ANN models (predicting V_x and V_y simultaneously) were compared with single-output models (separate training for each component). Interestingly, all variants—single-output, dual-output with unweighted mean loss, and dual-output with variance-weighted loss—converged to the same optimal hyperparameter configurations and produced identical results in this case study. This indicates that, under the parameter ranges explored, learning both components jointly did not provide measurable improvements in generalization.

This outcome can be explained by several technical factors:

a) Separate models allow each network to dedicate its full capacity to a single variable, while dual-output models must share hidden representations, which may not be optimal for both outputs.
b) V_x and V_y exhibit different statistical distributions, so a joint loss may force compromises, "sacrificing" accuracy in one output to improve the other.
c) Balancing losses is non-trivial: if one component has larger variance, it may dominate training unless explicitly weighted.
d) Limited model capacity can restrict the ability of a dual-output network to simultaneously capture both patterns, while two single-output models double the effective capacity.

Despite the identical numerical performance, dual-output models remain attractive because they reduce computational cost: only one network must be trained, leading to shorter runtimes and simpler deployment. Thus, in practice, dual-output architectures can be recommended when efficiency is a priority.

By contrast, no joint multi-output models were tested for wind speed (V) and wind direction (θ). Although such an approach could in principle improve the physical coherence of the predictions by capturing their interdependence, it presents important challenges. V is a linear target, typically optimized with squared-error loss, whereas θ is circular and requires dedicated angular loss functions. Combining both in a single training process would thus require careful normalization or weighting of heterogeneous errors, and may even involve transforming θ into its Cartesian components (cosθ, sinθ) to avoid discontinuities. These difficulties, together with the need to ensure methodological comparability across techniques, motivated the decision to train separate models for V and θ in the present study.

Regarding the seasonal encoding discussed in Section 2.4, two alternative formulations of the annual cycle were tested: a monthly harmonic pair, Eq. (18), and a daily harmonic pair, Eq. (19). Both approaches provide a correct cyclic encoding, avoiding artificial discontinuities at the end of the year. While the daily formulation offers finer resolution of the annual cycle, the results obtained in this study did not show significant differences in predictive accuracy compared to the monthly representation.

This suggests that, for large-scale seasonal modulation, a monthly harmonic is parsimonious and sufficiently robust against interannual variability, whereas the daily formulation may become relevant in contexts where higher temporal resolution is critical. Extending this comparison to multiple sites with different climatic regimes would be necessary to determine whether daily harmonics consistently improve performance, representing a promising line of future work.

Overall, the results from Task 3 demonstrate that ANN-based MCP models provide the most reliable corrections of reanalysis biases. The comparative analysis of modeling alternatives (direct vs. two-stage, single-vs. dual-output, monthly vs. daily harmonics) shows that, although no significant accuracy gains were obtained from the more

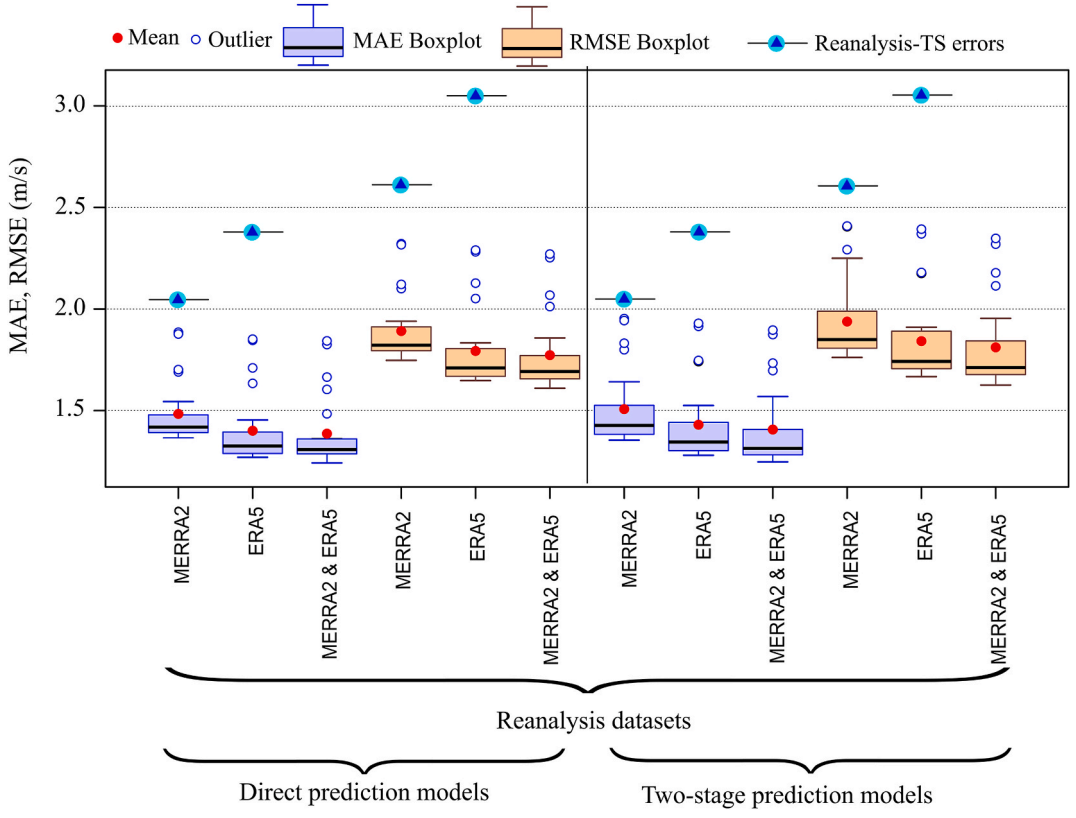


Fig. 8. Values of long-term wind speed error metrics.

complex formulations in this case, they offer practical advantages (e.g., computational efficiency) and remain promising directions for future studies.

3.3. Task-4: Analysis of the methods based on fundamental statistical metrics

The following subsections show the results obtained in Task-4 from the points of view indicated in Fig. 1.

3.3.1. Local accuracy

Fig. 8 shows the error metrics (MAE and RMSE) obtained when comparing wind speed estimates from the MCP models with TS data. Errors are presented as a function of the ANN-based MCP method (direct or two-stage) and the input data sources (MERRA2, ERA5, or their combination). The lowest average errors are consistently achieved when both reanalysis datasets are combined. Moreover, all MCP-based estimates reduce the errors reported in Table 1, confirming the added value of bias correction compared with the raw reanalysis data.

Direct models generally yield lower mean errors than two-stage models for wind speed. Significant differences between the two approaches are observed, except in the case of MAE when using MERRA2 or MERRA2 & ERA5 as inputs. This suggests that, while both approaches are effective, direct models are slightly better suited for speed estimation.

Fig. 9 presents the error metrics (MAE and RMSE) for wind direction. In contrast to wind speed, two-stage models clearly outperform direct models, producing lower mean errors across all reanalysis inputs. The smallest errors are again obtained when combining MERRA2 and ERA5. For direct models, no significant differences are observed between ERA5 alone and the combined dataset. As with wind speed, the MCP-corrected results are substantially better than those from raw reanalysis, regardless of the training year.

Fig. 10 shows the correlation metrics: R^2 for wind speed (left plots) and the circular correlation coefficient (r_c) for wind direction (right plots). For wind speed, the highest R^2 values are obtained when both reanalysis sources are combined, but no significant difference exists between direct (79 %) and two-stage (79.14 %) models. These results indicate that approximately 79 % of the observed wind speed variability can be explained by the predictors. Models trained with ERA5 consistently yield higher R^2 values than those using MERRA2, a reversal of the pattern observed in the raw data (Table 1), which highlights the stronger corrective potential of ERA5 after ANN-based MCP processing.

For wind direction, two-stage models achieve the highest *rc* values, confirming their superiority in capturing directional behavior. No significant differences are found between ERA5 and the combined dataset in this case. To facilitate interpretation, the comparison of circular correlations in Fig. 10 was performed using a broken y-axis, which enhances the readability of differences in the central range while still displaying outliers to provide a complete picture of the results.

Overall, the results from Task-4 confirm that ANN-based MCP models substantially reduce the errors of raw reanalysis data and capture both wind speed and direction with good accuracy. Nevertheless, model performance is not only influenced by the choice of approach (direct vs. two-stage) or data source (MERRA2, ERA5, or their combination), but also by the amount of on-site data available for training. The following section therefore examines the impact of training period length, a critical factor for practical applications where measurement campaigns are typically short.

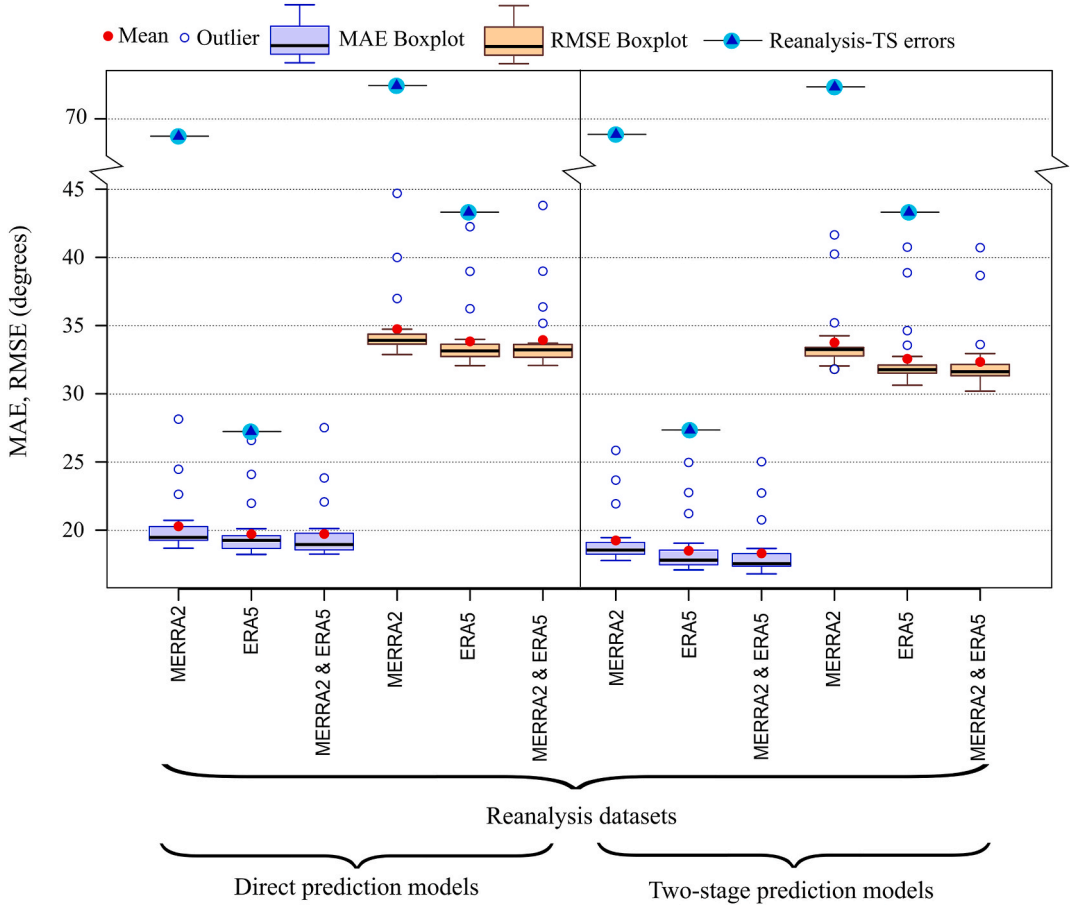


Fig. 9. Values of long-term wind direction error metrics.

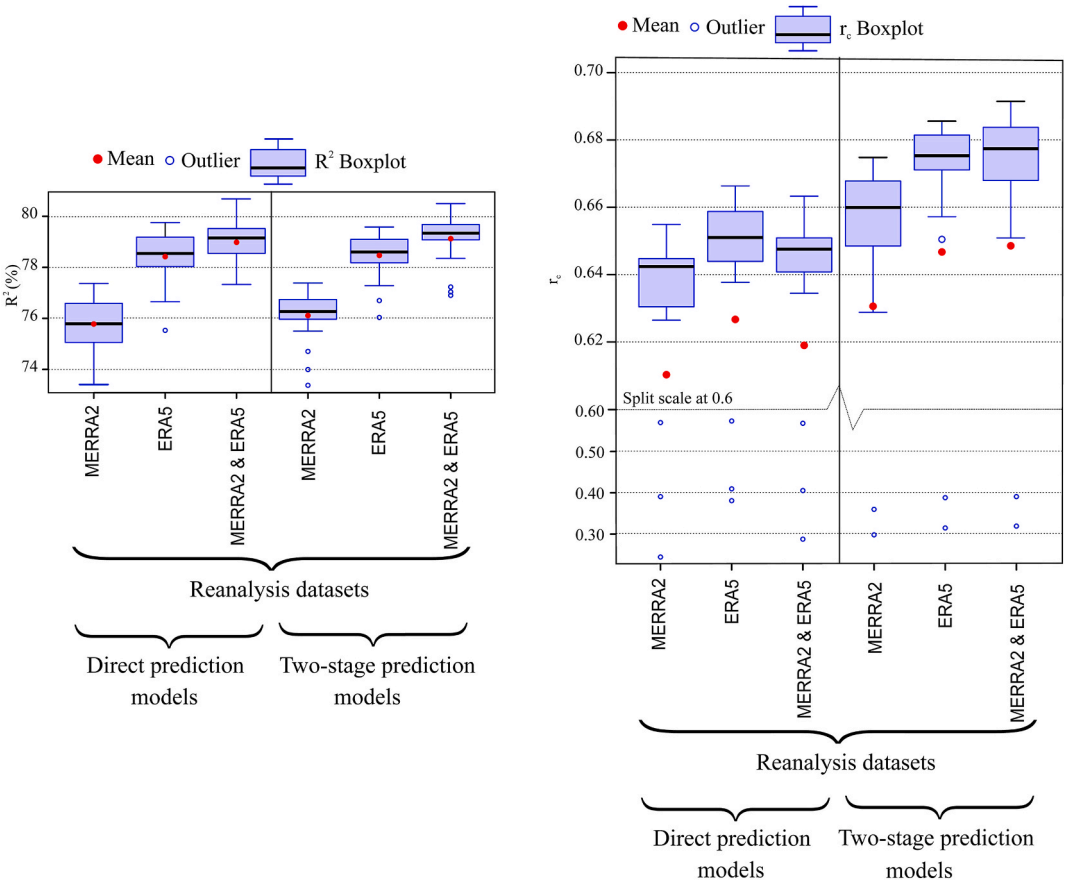


Fig. 10. Correlation metrics of ANN-based MCP models compared with TS data. The left plots show the determination coefficient (R^2) for wind speed, while the right plots display the circular correlation coefficient (r_c) for wind direction. Results are presented for direct and two-stage models using MERRA2, ERA5, and their combination (MERRA2 & ERA5) as input sources. To enhance readability, the right-hand plots use a split Y-axis scale at $r_c = 0.6$, which emphasizes the central distribution while still displaying the outliers for completeness.

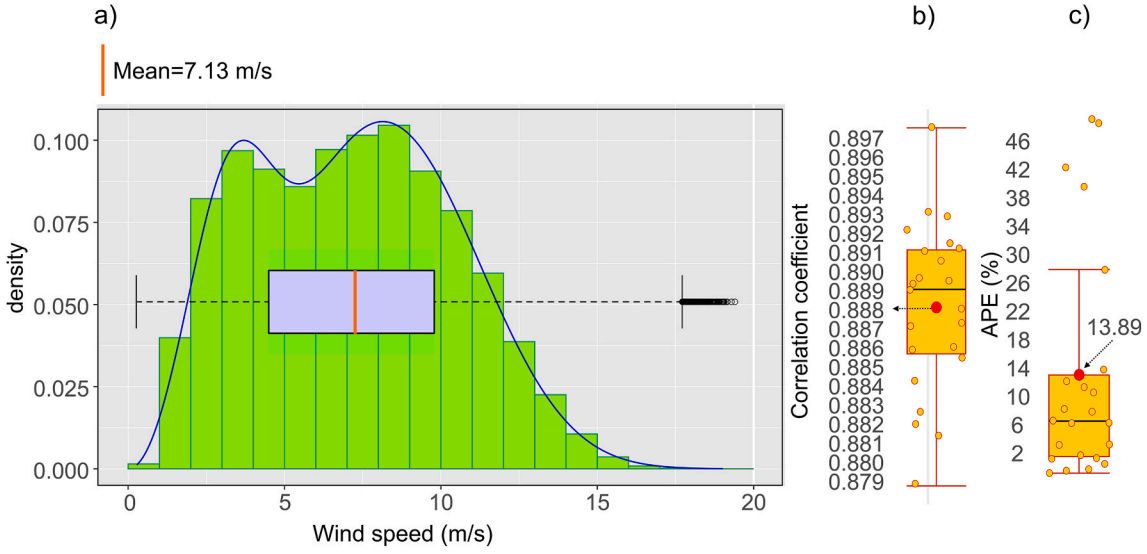


Fig. 11. a) Probability density function fitted to the wind data generated by the MCP model; b) Boxplot of correlation coefficients between the target site wind speeds and those estimated with the MCP model; c) Boxplot of the absolute percentage error (APE) values estimated using the target site wind speeds and the wind speeds generated by the MCP models.

3.3.2. Wind speed and direction distribution functions

The ability of MCP-based models to reproduce the statistical distribution of wind variables was assessed by comparing the estimated and observed series at the TS.

For **wind speeds,** the direct model with combined reanalysis inputs (MERRA2 & ERA5) provided the best performance, as also reflected in

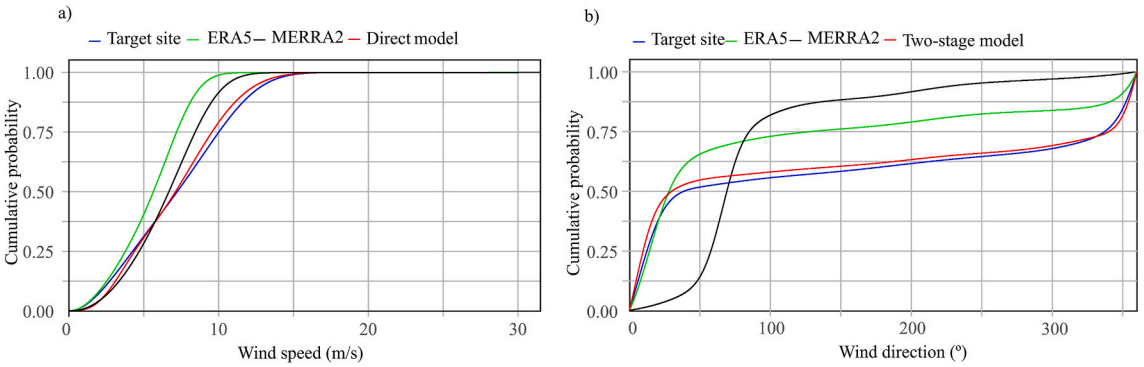


Fig. 12. Cumulative probabilities: a) Wind speed, b) Wind direction.

the lowest MAE and RMSE values (Fig. 8). Its fitted probability density function closely reproduced the bimodal pattern observed in the TS data (Fig. 11a).

The parameters of the fitted distributions are summarized in Table A.7, while the results of the formal goodness-of-fit tests are given in Table A.8. In all cases, the A–D, K–S, and E–D tests did not reject the null hypothesis of equality between the estimated and observed distributions at the 5 % level. The Wasserstein distance was also significantly lower than for raw reanalysis data, confirming the improved alignment. This is consistent with the cumulative probability curves (Fig. 12a), which show that MCP-corrected wind speeds follow the TS distribution much more closely than the raw reanalyses. The hourly correlations reached a mean value of 0.888 (Fig. 11b), a level classified as "good" in the MCP literature [37]. The mean wind speed obtained with the direct model (7.13 m/s) was practically identical to the observed value (7.2 m/s).

For wind directions, the two-stage model with combined inputs (MERRA2 & ERA5) yielded the closest agreement with the TS distribution. Its fitted probability density function (Fig. 13a) and the cumulative distribution (Fig. 12b) illustrate the strong alignment achieved after bias correction. The parameters of the distributions are provided in Table A.9, while the outcomes of the CvM, K–S, and E–D tests are shown in Table A.10. Again, the null hypothesis of equality between distributions was not rejected, and the Wasserstein distance (0.13) was much smaller than for raw reanalysis data. The mean circular correlation coefficient between estimated and observed hourly directions was 0.648 (Fig. 13b), a substantial improvement over the raw correlations (Table 1). The estimated mean wind direction (4.58°) was also nearly identical to the observed mean (4.49° at the TS), which has practical implications for turbine layout design, as proper alignment with prevailing winds helps reduce wake effects and increase energy efficiency.

In summary, both wind speed and wind direction distributions generated by ML-based MCP models can be considered statistically indistinguishable from TS observations, providing a robust basis for long-term wind resource assessment.

3.3.3. *Wind power densities*

As shown in Fig. 11c, the mean absolute percentage error (APE) between the wind power density ($\overline{WPD}$) estimated by the direct MCP model and that calculated from TS wind speeds is 13.89 %. This represents a substantial improvement compared to the APE values obtained when using raw reanalysis data (39 % for MERRA2 and 63.1 % for ERA5; see subsection 3.1.2). These findings confirm that MCP-based models are able to substantially reduce the bias of reanalysis data, providing wind power density values that are much closer to local observations. Since ($\overline{WPD}$) is a key indicator for assessing the economic viability of wind projects, this improvement highlights the practical relevance of applying ML-based MCP models in wind resource assessment.

3.3.4. *Seasonal evolution and daily mean behavior of wind speed*

The MCP models also succeeded in reproducing the temporal structure of wind speed variability. As shown in Fig. 6, the Pearson correlation between monthly mean wind speeds estimated with the direct model and those recorded at the TS was **0.996**, a much higher value than those obtained when comparing the TS with MERRA2 or ERA5 alone. Similarly, the correlation between daily mean wind speeds estimated by the direct model and those observed at the TS reached **0.994** (Fig. 7). These results indicate that the MCP methodology not only corrects mean

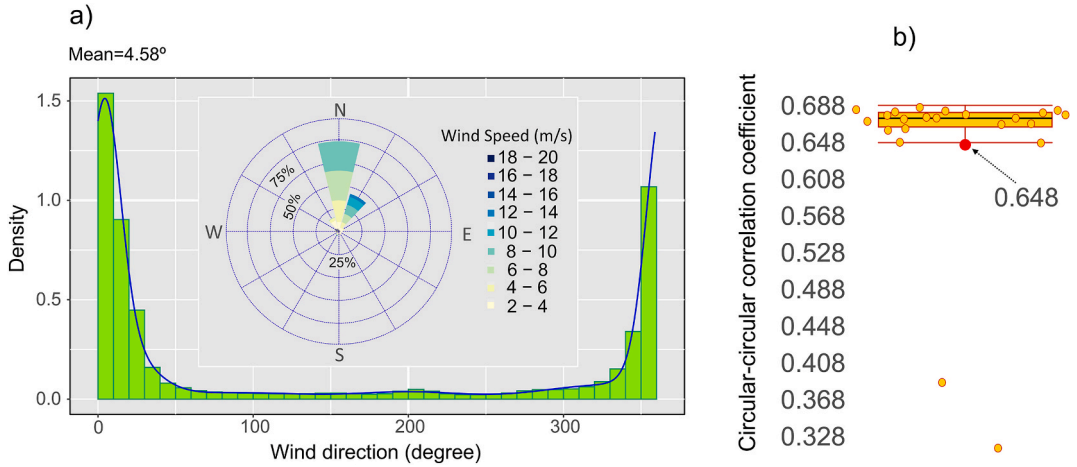


Fig. 13. a) Probability density function fitted to the wind direction data generated by the MCP model, b) Boxplot of correlation coefficients between the target site wind directions and those estimated by the MCP model.

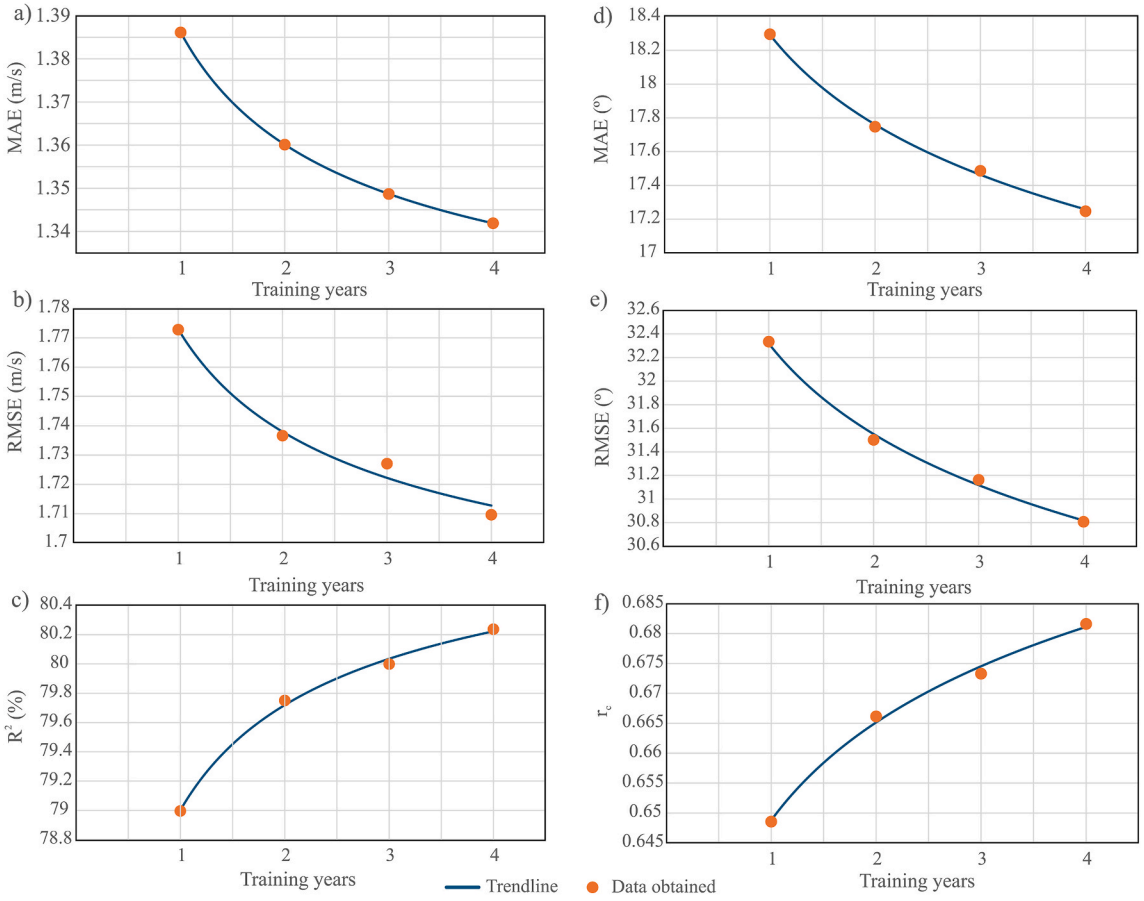


Fig. 14. Trend of a) MAE, b) RMS, c) R^2, d) MAE_Circular, e) RMSE_Circular, and f) circular correlation coefficient (r_c) metrics as the number of years used in the training of the MCP models increases.

values but also preserves the seasonal cycle and the intraday profile of wind speeds. This capacity is particularly important for energy system applications, where matching wind generation with demand and planning storage depends on capturing realistic hourly and seasonal dynamics.

3.4. *Results of the analysis of performance (Task-5)*

Fig. 14 shows the evolution of long-term error (MAE, RMSE, MAE_Circular, RMSE_Circular) and association metrics (R^2, r_c) as the training window increases from one to four years (Y = 1–4). A consistent trend can be observed: error metrics decrease, while association metrics increase with longer training periods. This reflects the fact that larger training datasets allow the models to better capture the statistical relationships between predictors and target variables, leading to improved generalization and stability.

These results have important implications for the design of measurement campaigns. In the case studied, using only one year of training data already provides reasonably accurate estimates, but incorporating two or more years yields additional gains in robustness. This observation is consistent with the recommendation of Fadigas et al. [8], who emphasized the importance of extending measurement campaigns from 24 to 36 months in Brazil to ensure reliable wind resource assessments. Our findings confirm that longer training periods enhance the generalization ability of ML-based MCP models, reducing sensitivity to the specific choice of training year.

However, this technical improvement must be balanced against the practical costs of measurement campaigns. Extending the data collection period at the target site implies higher economic costs and, more critically, longer delays in decision-making for wind project development.

In practice, long on-site measurement campaigns can delay investment decisions, especially where wind farm development is regulated through competitive tenders with strict deadlines between call publication and project submission. For this reason, developers often have no more than one year to collect wind data at a candidate site. While such a period is generally sufficient to characterize the seasonal cycle, it is insufficient to capture interannual variability, which typically requires substantially longer observational records.

In this context, our results highlight an important trade-off: while two or more years of training data improve the performance of MCP models, in practice many projects must rely on shorter campaigns (often $\leq$1 year). The fact that our models already achieve high correlations ($R^2 \approx 0.78$, $r_c \approx 0.63$) and low errors with Y = 1 is therefore a key finding, as it shows that reliable bias correction of reanalysis data can be achieved even under the typical constraints faced by developers. Longer training windows, when available, remain desirable for research or strategic planning, but the practical feasibility of one-year campaigns makes them the default option in most real-world wind energy tenders.

4. Limitations

The present study was designed to evaluate the ability of ML-based MCP methods to correct the bias of a single reanalysis grid point at a target site. This choice is consistent with common practice in wind resource assessment, where reanalysis data are often used as a proxy for local conditions. However, several methodological alternatives remain unexplored.

First, the use of neighboring reanalysis grid cells could provide additional spatial context, although selecting the most representative nodes (closest vs. most correlated with the target site) would require a

dedicated study. As indicated in Ref. [37], the use of a high number of reference stations (neighboring reanalysis grid cells) may result in overspecification with its associated negative effects. These include, among others, an increase in the estimation error and/or overfitting, which could be detrimental to the generalization capacity of the model when handling new data (prediction). Therefore, it would be necessary to analyze the benefits of feature selection [37]. It would also be appropriate to carry out a global sensitivity analysis method applied to wind speed and direction prediction models [42]. The relevance of a global sensitivity analysis is that it allows quantification of the contribution of the uncertainty of each input variable of the estimation model to the uncertainty of the model response [42]. Given the potential correlation between data from neighboring reanalysis grid cells, the global sensitivity method should consider the dependency among the input variables [42].

Second, it should be noted that the native spatial resolution is approximately 30 km for ERA5 and 50 km for MERRA2. Consequently, the target site may be located at a considerable distance from the corresponding reanalysis node, which can affect the accuracy with which it represents local conditions. Depending on the distance to the grid node, as well as hourly variations in wind speed and direction, discrepancies may arise between the reanalysis series and the target site measurements, potentially leading to a lack of synchronization between them. In this context, the option of incorporating lagged predictors to capture delayed or advanced dependencies between the reference series and the target series deserves to be explored. Another possibility is to apply spatial interpolation methods between neighboring reanalysis nodes. Deterministic approaches, which estimate values based on nearby samples (e.g., Inverse Distance Weighting [43]), or geostatistical approaches, which rely on statistical models and spatial autocorrelation to provide accuracy estimates (e.g., kriging [43]), could be used to estimate wind characteristics at the exact coordinates of the target site. These interpolated values could then serve as input to the ML models, potentially improving their representativeness.

Third, the incorporation of additional meteorological variables (e.g., pressure, temperature, solar radiation, relative humidity, etc.) into the input space of the ML models could be analyzed in order to improve the physical representation of the system and potentially reduce error metrics.

Fourth, ERA5 provides wind data at 10 m and 100 m above ground level, while MERRA2 provides data at three heights: 2 m, 10 m, and 50 m. It is of interest to analyze the error metrics when including wind speeds at multiple altitudes as predictors. In this context, it should be noted that feature selection methods may discard some heights due to the high correlation among them [37]. It is worth noting that some authors [44] have suggested that the heights above ground level (agl) at which data are collected at the reference and target sites should be similar.

Another important limitation is the lack of results from ML techniques using nearby ground-based reference stations to compare against those obtained with reanalysis data. In this regard, it is worth noting that the main conclusion of the studies conducted by Brower [45], where reanalysis data and direct observations were compared, was that reanalysis data were not reliable enough for use in MCP methods. However, Schwartz et al. [46] carried out an initial assessment of the usefulness of including reanalysis data in wind resource evaluation methodologies and obtained encouraging results, while emphasizing that unresolved issues remain that justify further research. We believe that such comparisons would be valuable, although to reach broader conclusions a more systematic evaluation across multiple sites with diverse terrains would be required. This is because the representativeness of results from nearby stations depends largely on local orography, which limits their generalizability.

In addition, the present study did not explore the multi-output capabilities of ML techniques other than ANN. Although **RFs** natively allow multi-output prediction, and **SVR/XGBoost** can be extended via wrappers, these approaches do not share parameters across outputs and are effectively equivalent to training independent models. Assessing the potential of such variants, as well as future multitask implementations in algorithms beyond ANN, constitutes a promising line of research.

Furthermore, no joint multi-output models were developed for wind speed (V) and wind direction (θ). While such an approach could enhance physical consistency by modeling both variables simultaneously, it raises important technical challenges. V is a linear target typically optimized with squared-error loss, whereas θ is circular and requires specialized angular losses. Combining them in a single architecture would therefore demand careful balancing of heterogeneous objectives or a transformation of θ into its Cartesian components (cosθ, sinθ) to avoid discontinuities. For these reasons, V and θ were modeled independently in this study, but extending multi-output formulations to heterogeneous targets remains an important direction for future research.

Finally, it should be noted that the comparative behavior of reanalysis datasets may vary across sites. In the present study, raw MERRA2 wind speeds were initially closer to the observed values than ERA5, but after ML-based bias correction, ERA5 consistently outperformed MERRA2. Whether this inversion is a site-specific phenomenon or reflects a more general pattern remains an open question. Systematic comparative studies at other locations, encompassing diverse climatic and topographic conditions, would be needed to establish the generality of this finding.

Addressing these aspects lies beyond the scope of the present work, but they represent promising avenues for future research. Their inclusion could complement and expand the approach presented here, further strengthening the methodological robustness and applicability of ML-based MCP methods.

5. Conclusions

This study shows that machine-learning-based MCP methods can substantially improve the reconstruction of long-term wind speed and direction at sites with limited on-site measurements. By correcting the systematic discrepancies of ERA5 and MERRA-2, the proposed approach produces site-adapted wind series that more accurately reflect the observed mean values (7.2 m/s vs. 7.13 m/s for wind speed and 4.49° vs. 4.50° for direction) and their daily and seasonal variability, achieving a correlation of 0.994 for the daily mean cycle.

Among the tested techniques, artificial neural networks consistently achieved the best performance, and the combination of ERA5 and MERRA-2 yielded the lowest errors across all model variants. The bias-corrected MCP models reduced the mean relative error in wind power density to 13.9 %, compared with 39 % (MERRA-2) and 63.1 % (ERA5) from raw reanalysis data, demonstrating a substantial gain in long-term resource representativeness.

The comparison between direct and two-stage MCP formulations shows similar overall performance, while the choice of ML technique and training-period length has a clear influence on accuracy. These findings indicate that the proposed methodology provides a practical and robust framework for improving the usability of reanalysis data in wind-resource feasibility studies, especially at locations where long-term ground measurements are unavailable.

CRediT authorship contribution statement

José A. Carta: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pedro Cabrera:** Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research has been co-funded with ERDF funds through the INTERREG MAC 2021–2027 programme in the RESMAC project (1/MAC/2/2.2/0011). This work made use of wind speed and direction data provided by the Spanish State Meteorological Agency (AEMET), Ministry of Agriculture, Food and Environment. No funding sources had any influence on study design, collection, analysis, or interpretation of data, manuscript preparation, or the decision to submit for publication.

Appendix A

A.1. R libraries used

Table A.1 lists the R libraries employed to implement and train the machine learning (ML) models (ANN, SVR, XGBoost, RF).

Table A.1
R code libraries used for the implementation and training of the different machine learning (ML) techniques

ML	Library	Reference
ANN	torch	https://cran.r-project.org/web/packages/torch/index.html
SVR	caret	https://cran.r-project.org/web/packages/caret/index.html
	Kernlab	https://cran.r-project.org/web/packages/kernlab/index.html
XGBoost	mlr3	https://cran.r-project.org/web/packages/mlr3/index.html
RF	mlr3	https://cran.r-project.org/web/packages/mlr3/index.html

A.2. Probability distributions of wind data

This section provides the parameters and goodness-of-fit test results for the wind speed and direction distributions from the target site and reanalysis datasets (ERA5 and MERRA2).

Table A.2 reports the parameters and Anderson–Darling (A–D) test results for the **wind speed pdfs** obtained from the target site and the reanalysis datasets (ERA5 and MERRA2). In all cases, the A–D test returned p-values above 0.05, indicating no statistical evidence to reject the null hypothesis of similarity between the empirical and fitted distributions. Table A.3 presents the corresponding results for the **wind direction pdfs**, together with the Cramér–von Mises (CvM) test. Although these tests also support similarity, visual inspection reveals noticeable discrepancies, particularly between ERA5 and the target site, which highlights the limitations of reanalysis data in capturing local directional patterns.

Table A.2
Parameters of the wind speed distributions of the data sources and the *p*-values obtained from the goodness-of-fit test used.

Data sources	Parameters					A-D statistic	p-value
	α_1	β_1	α_2	β_2	ω		
	–	m s^{-1}	–	m s^{-1}	–		–
MERRA2	2.23	4.45	5.94	8.47	0.478	0.8055	0.931
ERA5	1.97	4.53	4.48	7.25	0.435	0.921	0.933
Target site	1.92	4.73	3.89	10.37	0.416	2.300	0.647

Table A.3
Parameters of the wind direction distributions generated by the MCP models, and the *p*-values obtained from the goodness-of-fit test used.

		MERRA2	ERA5	TARGET
κ_1	–	27.5192899	16.8648556	16.179670
μ_1	(rad)	1.1565837	0.2536739	0.1117812
ω_1	–	0.53855894	0.60876372	0.69480834
κ_2	–	4.0460901	2.8216698	1.643063
μ_2	(rad)	−2.7217488	−2.5940214	−3.1043191
ω_2	–	0.05431074	0.06069787	0.07180155
κ_3	–	0.3030963	1.4530359	1.137917
μ_3	(rad)	−0.6507579	1.1163983	−1.4023937
ω_3	–	0.07373193	0.11668085	0.04990065

(*continued on next page*)

Table A.3 (*continued*)

		MERRA2	ERA5	TARGET
κ_4	–	2.6728175	0.8041157	2.418890
μ_4	(rad)	0.8700034	1.0141130	1.1245219
ω_4	–	0.13399201	0.03814863	0.06825952
κ_5	–	3.7265521	4.1479544	3.899613
μ_5	(rad)	1.3348655	0.2662494	−0.3872888
ω_5	–	0.19940638	0.17570893	0.11522994
CvM statistic		0.2535	0.10399	8.12
p-value		0.65	0.814	0.492

A.3. Distributions generated by MCP models

Tables A.4 and A.5 report the parameters of the wind speed and wind direction pdfs estimated with direct and two-stage MCP models, together with the corresponding goodness-of-fit test results. The analysis indicates that direct models tend to reproduce more accurately the global shape of the wind speed pdf, whereas two-stage models provide a better adjustment of the wind direction pdf. These results are consistent with the discussion in Section 3, where the complementary strengths of both approaches are emphasized.

Table A.4
Tests and metrics used to compare two wind speed distributions.

Anderson-Darling test. *p*-values		
	MERRA2	ERA5
Target site	A-D = 24; $p < 0.001$	A-D = 68.8; $p < 0.001$
Kolmogorov-Smirnov test		
Target site	K-S = 0.17044; $p < 0.001$	K-S = 0.30226; $p < 0.001$
Energy distance test for equality of distributions		
Target site	E-D = 83.07; $p < 0.001$	E-D = 275.20; $p < 0.001$
Wasserstein distance		
Target site	1.0131	1.7581

Table A.5
Statistical tests and metrics used to compare two wind direction distributions.

Cramér-von Mises test. *p*-values		
	MERRA2	ERA5
Target site	CvM = 11026.918; $p < 0.001$	CvM = 3215.78; $p < 0.001$
Kolmogorov-Smirnov test		
Target site	K-S = 0.42876; $p < 0.001$	K-S = 0.17843; $p < 0.001$
Energy distance test for equality of distributions		
Target site	E-D = 22053.836; $p < 0.001$	E-D = 6431.56; $p < 0.001$
Wasserstein distance		
Target site	1.725	0.9216

A.4. Statistical comparison of distributions

Table A.6 summarizes the outcomes of several tests—including Anderson–Darling, Kolmogorov–Smirnov, Energy distance, and Wasserstein distance—used to compare the pdfs generated by MCP models with those observed at the target site. The reported values confirm that the application of MCP substantially improves the agreement between reanalysis-based and observed distributions, both for wind speed and direction, thereby supporting the effectiveness of the proposed methodology.

Table A.6
Best hyperparameters (frequencies of occurrence in parentheses).

Model type	Source data	Feature target	Number of neurons per hidden layer			Dropout rate			Learning rate
			Layer_1	Layer_2	Layer_3	Dropout_1	Dropout_2	Dropout_3	
Direct models	MERRA2	V	63 (17)	54 (17)	0(17)	0.1120 (17)	0.1791 (17)	- (17)	0.001 (23)
			159 (6)	104 (6)	53(6)	0.2738 (6)	0.0376 (6)	0.1876 (6)	
	ERA5	V	63 (6)	54 (6)	0(6)	0.1120 (6)	0.1791 (6)	- (6)	0.001 (23)

(*continued on next page*)

Table A.6 (*continued*)

Model type	Source data	Feature target	Number of neurons per hidden layer			Dropout rate			Learning rate
			Layer_1	Layer_2	Layer_3	Dropout_1	Dropout_2	Dropout_3	
			159 (17)	104 (17)	53(17)	0.2738 (17)	0.0376 (17)	0.1876 (17)	
	MERRA2&ERA5	V	63 (14)	54 (14)	0(14)	0.1120 (14)	0.1791 (14)	- (14)	0.001 (23)
			159 (9)	104 (9)	53(9)	0.2738 (9)	0.0376 (9)	0.1876 (9)	
	MERRA2	θ	63 (1)	54 (1)	0(1)	0.1120 (1)	0.1791 (1)	- (1)	0.001 (23)
			159 (22)	104 (22)	53(22)	0.2738 (22)	0.0376 (22)	0.1876 (22)	
	ERA5	θ	63 (1)	54 (1)	0(1)	0.1120 (1)	0.1791 (1)	- (1)	0.001 (23)
			159 (22)	104 (22)	53(22)	0.2738 (22)	0.0376 (22)	0.1876 (22)	
	MERRA2&ERA5	θ	159 (23)	104 (23)	53(23)	0.2738 (23)	0.0376 (23)	0.1876 (23)	0.001 (23)
Two-stage models	MERRA2	V_x	159 (23)	104 (23)	53(23)	0.2738 (23)	0.0376 (23)	0.1876 (23)	0.001 (23)
		V_y	159 (22)	104 (22)	53(22)	0.2738 (22)	0.0376 (22)	0.1876 (22)	0.001 (23)
			63(1)	54(1)	0(1)	0.1120 (1)	0.1791 (1)	-(1)	
	ERA5	V_x	159 (23)	104 (23)	53(23)	0.2738 (23)	0.0376 (23)	0.1876 (23)	0.001 (23)
		V_y	159 (22)	104 (22)	53(22)	0.2738 (22)	0.0376 (22)	0.1876 (22)	0.001 (23)
			63(1)	54(1)	0(1)	0.1120 (1)	0.1791 (1)	-(1)	
	MERRA2&ERA5	V_x	159 (23)	104 (23)	53(23)	0.2738 (23)	0.0376 (23)	0.1876 (23)	0.001 (23)
		V_y	159 (22)	104 (22)	53(22)	0.2738 (22)	0.0376 (22)	0.1876 (22)	0.001 (23)
			63(1)	54(1)	0(1)	0.1120 (1)	0.1791 (1)	-(1)	

A.5. ANN architectures and hyperparameters

The ANN models employed in this study correspond to fully connected feed-forward networks with ReLU activation functions in the hidden layers. Table A.7 presents the most frequent optimal hyperparameter configurations identified during training, including the number of neurons per layer, dropout rates, and learning rate. Fig. A1 illustrates the evolution of the training process and highlights the best epochs selected by early stopping, providing additional insight into the stability and convergence of the ANN models.

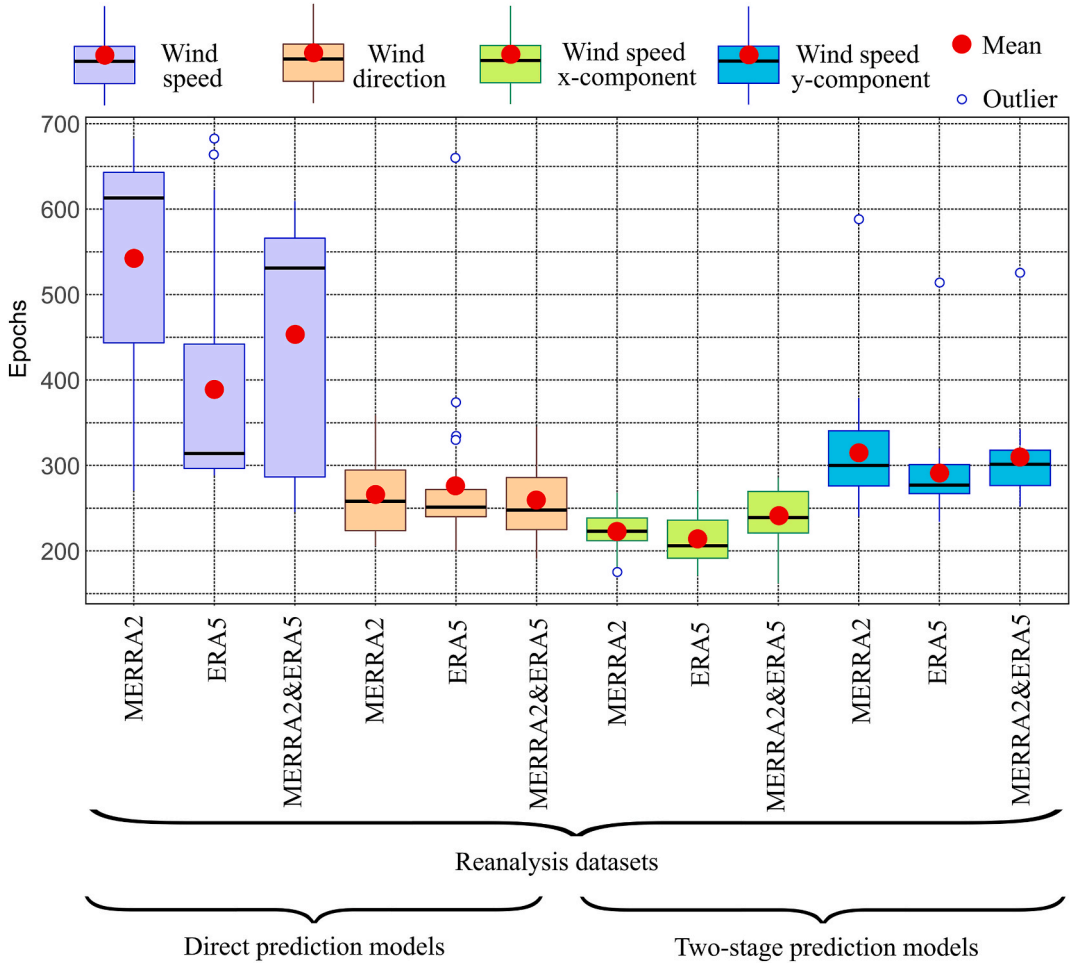


Fig. A.1. Best epochs detected in the training stages.

Table A.7
Parameters of the wind speed distributions generated by the MCP models and the *p*-values obtained from the goodness-of-fit test used.

Model	Data sources	Parameters					A-D statistic	*p*-value
		α_1	β_1	α_2	β_2	ω		
		–	m s^{-1}	–	m s^{-1}	–		–
Direct	MERRA2	3.097218	3.898027	3.207062	9.158304	0.224331	5.63184197812734	0.678
	ERA5	3.128615	3.609815	3.182110	9.119546	0.2189469	5.81914340389994	0.663
	MERRA2&ERA5	2.975513	3.655812	3.187744	9.166803	0.2185944	5.7613296060299	0.659
Two-stage	MERRA2	2.007278	3.573671	3.705527	9.483398	0.3323665	5.5397650443119	0.737
	ERA5	2.007188	3.361209	3.524680	9.393152	0.3036387	4.88330523152035	0.792
	MERRA2&ERA5	1.964464	3.431325	3.598121	9.465688	0.3129155	4.49977791657147	0.82

A.6. Additional tests for wind direction

Complementary results for the wind direction pdfs are provided in Tables A.8 and A.9, which include the Cramér–von Mises, Kolmogorov–Smirnov, Energy distance, and Wasserstein distance statistics comparing the direct and two-stage MCP models against the target site. These results confirm the conclusions presented in the main text, namely that two-stage models offer a superior representation of directional patterns, while direct models are more effective for wind speed estimation.

Table A.8
Tests and metrics used to compare two wind speed distributions.

Anderson-Darling test. *p*-values		
	Direct model	Two-stage model
Target site	A-D = 2.21; $p = 0.0682$	A-D = 2.7; $p = 0.0422$
Kolmogorov-Smirnov Test		
Target site	K-S = 0.04261; $p = 0.5061$	K-S = 0.0479; $p = 0.3381$
Energy distance test for equality of distributions		
Target site	E = 6.6743; $p = 0.14537$	E = 10.21173; $p = 0.0664$
Wasserstein distance		
Target site	0.3049	0.4125

Table A. 9
Parameters of the wind direction distributions constructed from the data estimated using the two MCP models (direct and two-stage) and the *p*-values obtained from the goodness-of-fit test employed.

Parameters		Direct models			Two-stage models		
		MERRA2	ERA5	MERRA2&ERA5	MERRA2	ERA5	MERRA2&ERA5
κ_1	–	143.556742	144.679467	117.275420	35.978766	32.023274	30.218896
μ_1	(rad)	0.02991201	0.02538258	0.03869844	0.06288515	0.06738307	0.07044619
ω_1	–	0.31974868	0.33136176	0.38477353	0.53626463	0.57703988	0.59218436
κ_2	–	2.344836	2.361173	2.342035	2.159977	2.293838	2.873558
μ_2	(rad)	1.46360509	1.35368931	1.41956171	−2.71862142	−2.65891637	−2.76626113
ω_2	–	0.06645413	0.08224934	0.08782369	0.06344519	0.05862097	0.05110372
κ_3	–	1.241879	1.901517	11.836226	1.349540	1.467506	1.624546
μ_3	(rad)	−0.95676818	−0.22650858	0.18919369	1.40986160	1.62558722	1.59177336
ω_3	–	0.10631832	0.16973824	0.17321678	0.06699457	0.06547607	0.06717850
κ_4	–	2.567656	17.030297	1.513234	2.606945	2.519059	2.617477
μ_4	(rad)	0.06741906	0.16551743	−0.90421319	−0.45238071	−0.53139667	−0.68706373
ω_4	–	0.19992365	0.33729302	0.14426601	0.11093712	0.10444880	0.10495889
κ_5	–	17.673995	1.080062	6.951453	10.360101	9.351302	8.686815
μ_5	(rad)	0.14773743	−0.91309823	0.11365550	0.24275401	0.26740822	0.26607832
ω_5	–	0.30755522	0.07935764	0.20991999	0.22235850	0.19441426	0.18457453
CvM statistic		0.3930	0.792	0.4974	0.5703	1.359	0.5299
p-value		0.9	0.556	0.858	0.766	0.562	0.704

Table A.10
Tests and metrics used to compare two wind direction distributions.

Cramér von Mises test. *p*-values		
	Direct model	Two-stage model
Target site	CvM = 118.59; $p = 0.21096$	CvM = 74.57; $p = 0.3363$
Kolmogorov-Smirnov test		
Target site	K-S = 0.10919; $p < 0.001$	K-S = 0.06258; $p = 0.1046$
Energy distance test for equality of distributions		
Target site	E = 237.18; $p = 0.2106$	E = 149.14; $p = 0.3433$
Wasserstein distance		
Target site	0.1372	0.1317

References

[1] B. Elshafei, A. Popov, D. Giddings, Enhanced offshore wind resource assessment using hybrid data fusion and numerical models, Energy 310 (2024) 133208, https://doi.org/10.1016/j.energy.2024.133208.

[2] S. Velázquez, J.A. Carta, J.M. Matías, Comparison between ANNs and linear MCP algorithms in the long-term estimation of the cost per kW h produced by a wind turbine at a candidate site: a case study in the Canary Islands, Appl. Energy 88 (2011) 3869–3881, https://doi.org/10.1016/J.APENERGY.2011.05.007.

[3] L. Landberg, L. Myllerup, O. Rathmann, E.L. Petersen, B.H. Jørgensen, J. Badger, N. G. Mortensen, Wind resource Estimation-An overview, Wind Energy 6 (2003) 261–271, https://doi.org/10.1002/we.94.

[4] T. Hiester, E. Pennell, The Siting Handbook for Large Wind Energy Systems, First, WindBooks, New York, 1981.

[5] T. Burton, D. Sharpe, N. Jenkins, E. Bossanyi, WIND ENERGY HANBOOK, Wiley, 2011.

[6] R.W. Baker, S.N. Walker, J.E. Wade, Annual and seasonal variations in mean wind speed and wind turbine energy production, Sol. Energy 45 (1990) 285–289, https://doi.org/10.1016/0038-092X(90)90013-3.

[7] K. Klink, NOTES AND CORRESPONDENCE Trends and Interannual Variability of Wind Speed Distributions in Minnesota, 2001.

[8] J.V.P. Miguel, E.A. Fadigas, I.L. Sauer, The influence of the wind measurement campaign duration on a measure-correlate-predict (MCP)-based wind resource assessment, Energies 12 (2019) 3606, https://doi.org/10.3390/EN12193606. Page 3606 12 (2019).

[9] J.A. Carta, S. Velázquez, P. Cabrera, A review of measure-correlate-predict (MCP) methods used to estimate long-term wind characteristics at a target site, Renew. Sustain. Energy Rev. 27 (2013) 362–400, https://doi.org/10.1016/J.RSER.2013.07.004.

[10] J. Olauson, ERA5: the new champion of wind power modelling? Renew. Energy 126 (2018) 322–331, https://doi.org/10.1016/J.RENENE.2018.03.056.

[11] R. Rabbani, M. Zeeshan, Exploring the suitability of MERRA-2 reanalysis data for wind energy estimation, analysis of wind characteristics and energy potential assessment for selected sites in Pakistan, Renew. Energy 154 (2020) 1240–1251, https://doi.org/10.1016/J.RENENE.2020.03.100.

[12] S.C. Pryor, R.J. Barthelmie, A global assessment of extreme wind speeds for wind energy applications, Nat. Energy 6 (2021) 268–276, https://doi.org/10.1038/s41560-020-00773-7.

[13] Y. Cai, F.M. Bréon, Wind power potential and intermittency issues in the context of climate change, Energy Convers. Manag. 240 (2021) 114276, https://doi.org/10.1016/J.ENCONMAN.2021.114276.

[14] T. Görmüş, B. Aydoğan, B. Ayat, Offshore wind power potential analysis for different wind turbines in the Mediterranean region, 1959–2020, Energy Convers. Manag. 274 (2022) 116470, https://doi.org/10.1016/J.ENCONMAN.2022.116470.

[15] Simon Watson, Handbook of Wind Resource Assessment, First, John Wiley & Sons Inc., 2023. https://www.wiley.com/en-us/Handbook+of+Wind+Resource+Assessment-p-9781119055396. (Accessed 14 November 2024).

[16] L.M. Sheridan, C. Phillips, A.C. Orrell, L.K. Berg, H. Tinnesand, R.K. Rai, S. Zisman, D. Duplyakin, J.E. Flaherty, Validation of wind resource and energy production simulations for small wind turbines in the United States, Wind Energy Sci. 7 (2022) 659–676, https://doi.org/10.5194/WES-7-659-2022.

[17] R. Gelaro, W. McCarty, M.J. Suárez, R. Todling, A. Molod, L. Takacs, C.A. Randles, A. Darmenov, M.G. Bosilovich, R. Reichle, K. Wargan, L. Coy, R. Cullather, C. Draper, S. Akella, V. Buchard, A. Conaty, A.M. da Silva, W. Gu, G.K. Kim, R. Koster, R. Lucchesi, D. Merkova, J.E. Nielsen, G. Partyka, S. Pawson, W. Putman, M. Rienecker, S.D. Schubert, M. Sienkiewicz, B. Zhao, The modern-era retrospective analysis for research and applications, version 2 (MERRA-2), J. Clim. 30 (2017) 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1.

[18] Climate reanalysis | Copernicus, (n.d.). https://climate.copernicus.eu/climate-reanalysis (accessed November 14, 2024).

[19] G. Gualtieri, Analysing the uncertainties of reanalysis data used for wind resource assessment: a critical review, Renew. Sustain. Energy Rev. 167 (2022) 112741, https://doi.org/10.1016/J.RSER.2022.112741.

[20] R.K. Samal, Assessment of wind energy potential using reanalysis data: a comparison with mast measurements, J. Clean. Prod. 313 (2021), https://doi.org/10.1016/j.jclepro.2021.127933.

[21] J.W. Ding, M.J. Chuang, J.S. Tseng, I.Y.L. Hsieh, Reanalysis and ground station data: advanced data preprocessing in deep learning for wind power prediction, Appl. Energy 375 (2024) 124129, https://doi.org/10.1016/J.APENERGY.2024.124129.

[22] K.V. Mardia, P.E. Jupp, Directional statistics, directional statistics, 1–432, https://doi.org/10.1002/9780470316979, 2008.

[23] J.A. Carta, P. Ramírez, S. Velázquez, A review of wind speed probability distributions used in wind energy analysis: case studies in the Canary Islands, Renew. Sustain. Energy Rev. 13 (2009) 933–955, https://doi.org/10.1016/J.RSER.2008.05.005.

[24] J.A. Carta, S. Díaz, A. Castañeda, A global sensitivity analysis method applied to wind farm power output estimation models, Appl. Energy 280 (2020) 115968, https://doi.org/10.1016/j.apenergy.2020.115968.

[25] J.A. Carta, C. Bueno, P. Ramírez, Statistical modelling of directional wind speeds using mixtures of von Mises distributions: case study, Energy Convers. Manag. 49 (2008), https://doi.org/10.1016/j.enconman.2007.10.017.

[26] RalphB. D'Agostino, Goodness-of-Fit techniques, goodness-of-fit techniques. https://doi.org/10.1201/9780203753064/GOODNESS-FIT-TECHNIQUES-RALPHB-AGOSTINO, 2017.

[27] M.L. Rizzo, G.J. Székely, DISCO analysis: a nonparametric extension of analysis of variance, 4 1034–1055, https://doi.org/10.1214/09-AOAS245, 2010.

[28] Y. Cai, L.H. Lim, Distances between probability distributions of different dimensions, IEEE Trans. Inf. Theor. 68 (2022) 4020–4031, https://doi.org/10.1109/TIT.2022.3148923.

[29] S. Díaz, J.A. Carta, J.M. Matías, Comparison of several measure-correlate-predict models using support vector regression techniques to estimate wind power densities. A case study, Energy Convers. Manag. 140 (2017) 334–354, https://doi.org/10.1016/J.ENCONMAN.2017.02.064.

[30] S.M. Weekes, A.S. Tomlin, Data efficient measure-correlate-predict approaches to wind resource assessment for small-scale wind energy, Renew. Energy 63 (2014) 162–171, https://doi.org/10.1016/J.RENENE.2013.08.033.

[31] A. Dinler, A new low-correlation MCP (measure-correlate-predict) method for wind energy forecasting, Energy 63 (2013) 152–160, https://doi.org/10.1016/J.ENERGY.2013.10.007.

[32] S.R. Jammalamadaka, Y.R. Sarma, A correlation coefficient for angular variables, Statist. Theor. Data Anal. II (1988) 349–364.

[33] S.R. Jammalamadaka, A. SenGupta, Topics in circular statistics, 5, https://doi.org/10.1142/4031, 2001.

[34] Trevor Hastie, Robert Tibshirani, J.H. Friedman, The elements of statistical learning : data mining, inference, and prediction, 745, https://books.google.com/books/about/The_Elements_of_Statistical_Learning.html?hl=es&id=eBSgoAEACAAJ, 2009. (Accessed 15 November 2024).

[35] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[36] S. Díaz, J.A. Carta, J.M. Matías, Performance assessment of five MCP models proposed for the estimation of long-term wind turbine power outputs at a target site using three machine learning techniques, Appl. Energy 209 (2018) 455–477, https://doi.org/10.1016/J.APENERGY.2017.11.007.

[37] J.A. Carta, P. Cabrera, J.M. Matías, F. Castellano, Comparison of feature selection methods using ANNs in MCP-wind speed methods. A case study, Appl. Energy 158 (2015) 490–507, https://doi.org/10.1016/J.APENERGY.2015.08.102.

[38] B. Boehmke, B. Greenwell, Hands-On machine learning with R, Hands-On machine learning with R. https://doi.org/10.1201/9780367816377, 2019.

[39] Practical Neural Network Recipies in C++ Masters, Morgan Kaufmann, 2014. http://www.sciencedirect.com:5070/book/9780080514338/practical-neural-network-recipies-in-c-and-and. (Accessed 19 November 2024).

[40] J.A. Carta, P. Cabrera, Optimal sizing of stand-alone wind-powered seawater reverse osmosis plants without use of massive energy storage, Appl. Energy 304 (2021) 117888, https://doi.org/10.1016/J.APENERGY.2021.117888.
[41] P. Cabrera, J.A. Carta, C. Matos, E. Rosales-Asensio, H. Lund, Reduced desalination carbon footprint on islands with weak electricity grids. The case of Gran Canaria, Appl. Energy 358 (2024) 122564, https://doi.org/10.1016/J.APENERGY.2023.122564.
[42] J.A. Carta, S. Díaz, A. Castañeda, A global sensitivity analysis method applied to wind farm power output estimation models, Appl. Energy 280 (2020) 115968, https://doi.org/10.1016/J.APENERGY.2020.115968.
[43] C. Munyati, N.I. Sinthumule, Comparative suitability of ordinary kriging and Inverse distance weighted interpolation for indicating intactness gradients on threatened savannah woodland and forest stands, Environ. Sustain. Indic. 12 (2021) 100151, https://doi.org/10.1016/J.INDIC.2021.100151.
[44] O. Probst, D. Cárdenas, State of the art and trends in wind resource assessment, Energies 3 (2010) 1087–1141, https://doi.org/10.3390/EN3061087, 3 (2010) 1087–1141.
[45] M. Brower, The use of NCEP/NCAR reanalysis data in MCP, in: European Wind Energy Conference & Exhibition, 2006 p. Greece; 27 February–2 March.
[46] M. Schwartz, R. George, D. Elliott, The use of reanalysis data for wind resource assessment at the national renewable energy laboratory. http://www.doe.gov/bridge/home.html, 1999. (Accessed 19 September 2025).