

Article

Development of a Wind Speed Forecasting Model Using Observed Data and Machine Learning Approaches

Paula Rose de Araújo Santos ¹, Louise Pereira da Silva ¹, Susane Eterna Leite Medeiros ²  and Raphael Abrahão ^{3,*} 

¹ Graduate Program in Mechanical Engineering, Universidade Federal da Paraíba, João Pessoa 58051-900, Brazil; paularose@alumni.cear.ufpb.br (P.R.d.A.S.); louise.silva@alumni.cear.ufpb.br (L.P.d.S.)

² Graduate Program in Civil and Environmental Engineering, Department of Technology, Universidade Estadual de Feira de Santana, Feira de Santana 44036-900, Brazil; selmedeiros@uefs.br

³ Center for Alternative and Renewable Energy, Department of Renewable Energy Engineering, Universidade Federal da Paraíba, João Pessoa 58051-900, Brazil

* Correspondence: raphael@cear.ufpb.br

Abstract

Considering the growing potential of artificial intelligence (AI), its application has become increasingly relevant in climate-related studies and energy assessments. In this study, the Random Forest algorithm was applied to impute missing values in time series of air temperature, wind speed, atmospheric pressure, and wind direction. The performance of the data imputation was evaluated using RMSE, MSE, and MAE metrics, as well as the Kolmogorov–Smirnov (KS) test, which supported the selection of the most appropriate exogenous variable. Subsequently, short-term wind speed forecasting was performed using the SARIMAX model, and monthly energy generation was estimated for the V80/2000, SWT-2.3-101, and S95/2100 wind turbine models. The proposed methodology was applied to data from 50 conventional meteorological stations of the National Institute of Meteorology (INMET) located in Northeast Brazil. The results indicate that the gap-filling procedure was effective, particularly for wind speed and mean air temperature. Moreover, the SARIMAX model demonstrated good forecasting performance at most of the analyzed stations. Overall, the findings suggest that the majority of the locations analyzed present favorable conditions for wind-based electricity generation.

Keywords: SARIMAX; weather forecast; observed data; wind turbine

1. Introduction

In recent years, growing concern over climate change has intensified the search for sustainable energy sources. According to reports by the Intergovernmental Panel on Climate Change (IPCC), anthropogenic activities associated with greenhouse gas emissions are the primary drivers of global warming [1]. The IPCC further highlights that the continued reliance on fossil fuels is the main cause of the current climate crisis, with climate-related impacts proving to be more severe and widespread than previously anticipated. In addition, several studies have emphasized the urgency of transitioning to low-carbon energy systems [2].

Rising global temperatures, the increasing frequency and intensity of extreme events, and heightened variability in weather patterns pose significant challenges, directly affecting human health, climate stability, and energy production [2]. Within this context, the transition from fossil fuel-based energy systems to renewable energy sources has



Academic Editor: Adrian Ilinca

Received: 4 January 2026

Revised: 27 January 2026

Accepted: 2 February 2026

Published: 24 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

emerged as a critical strategy for mitigating greenhouse gas emissions and fostering environmental sustainability.

Wind energy is a promising, sustainable and abundant source, which has the advantage of not using fossil fuels during its operation. The use of wind resources in the global energy matrix is increasing. According to the Global Wind Energy Council (GWEC), there was a new record in the wind sector in 2024, with 117 GW of new installations worldwide [3].

Although the study by Ghamati et al. [4] focuses on tidal turbines operating in an aquatic environment, it demonstrates the strong dependence of turbine performance on external flow conditions imposed by the surrounding medium. This fundamental concept is directly transferable to wind energy systems, where atmospheric variables (such as air temperature) modulate air density and flow structure, thereby influencing wind speed variability. As a result, accurately characterizing and modeling these atmospheric influences is essential for reliable wind resource assessment and for improving the performance of wind speed forecasting models.

In Brazil, the outlook is also positive. Newly installed wind capacity reached 4.8 GW for the third consecutive year, with more than 1000 wind farms in operation and total installed capacity exceeding 30 GW. The country ranked among the top five global markets for new wind power installations in 2023 [3] and remained within the world's top five wind energy markets in 2024 [5].

The Brazilian Northeast subsystem stands out as the country's leading region for wind energy generation, accounting for 90.3% of total wind power generation in 2022 [6]. In subsequent years, this share increased to 92.0% in 2023 and 92.2% in 2024 [7]. The region is predominantly influenced by the South Atlantic trade winds, which are strong, persistent, and highly favorable for wind energy production [8].

Wind energy is one of the main alternatives for renewable energy production, especially in areas with good wind potential, such as the Northeast of Brazil. However, wind seasonality causes uncertainties that can affect the predictability and stability of wind energy supply [9–11].

In recent years, artificial intelligence (AI) techniques have gained prominence and present themselves as a promising tool in different areas. AI consists of a set of advanced computational techniques capable of learning, recognizing patterns, and making decisions, that is, performing tasks that are usually done by human intelligence. In addition, they are able to work with large volumes of data and capture complex relationships [12,13].

Supervised learning algorithms, such as Random Forest (RF) and statistical methods, such as the Seasonal Autoregressive Integrated Moving Average with exogenous variables (SARIMAX), continue to be applied due to their efficiency.

Random Forest is a learning algorithm in which decision trees are built, trained, and combined to make predictions [14]. The technique consists of calculating the average of each individual tree that is generated to obtain the prediction [15]. SARIMAX is a statistical tool that makes future predictions in time series, incorporating exogenous factors that can influence the time series [16].

Permanent monitoring and accurate forecasts are essential and are only possible when reliable and complete time series are available. Considering that meteorological data often contain gaps, it is necessary to fill these gaps before applying methods to perform the forecasts.

Therefore, RF was used to complement data series, preserving their characteristics, and subsequently, SARIMAX was used to project wind speed data for a 24-month horizon at 50 conventional stations located in the Northeast region of Brazil. Finally, the monthly energy generated by the V80/2000, SWT-2.3-101 and S95/2100 wind turbines was calculated.

Accordingly, the main contribution of this study lies in the integrated application of a machine learning algorithm and a statistical modeling approach to perform data imputation and subsequently project wind speeds over a 24-month horizon. The proposed framework evaluates the model's ability to produce reliable forecasts based on observed data and enables the estimation of monthly energy generation for wind turbines using wind speed data in conjunction with their respective power curves.

2. Methodology

2.1. Characterization of the Study Area

Located in the easternmost portion of South America, the Brazilian Northeast region borders the Atlantic Ocean. The region covers an area of 1,552,167.01 km² and had an estimated population of 57,112,096 in 2024 (Figure 1) [17].

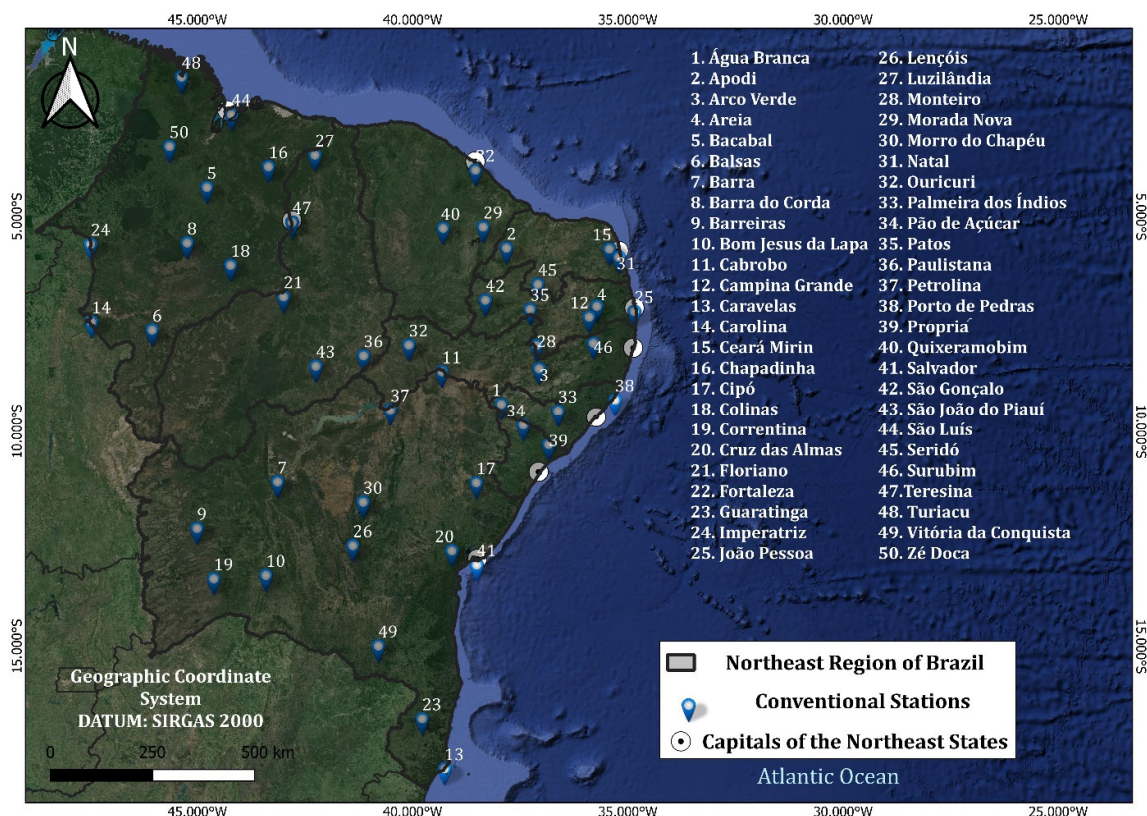


Figure 1. Location map of the Northeast Region of Brazil and the meteorological stations [18,19].

The regional climate is predominantly tropical along the coastline, while semi-arid conditions prevail in inland areas, characterized by low rainfall and prolonged drought periods. Another notable feature is the mean air temperature, which typically ranges between 20 and 28 °C [20–22]. The Brazilian Northeast is strongly influenced by the South Atlantic trade winds, which are persistent, stable, and highly favorable for wind energy production [8].

2.2. Data Preparation

The set of observed data of wind speed (m/s), average air temperature (°C), atmospheric pressure (mb) and wind direction (°) used in the present study was made available by the National Institute of Meteorology (INMET). Data were collected from conventional stations and complemented with data from automatic stations located throughout the Northeast region (Figure 1). It should be noted that the beginning of the analyzed period is

determined by the initial data made available by the station and the final period considered was December 2024 or the last year of available data, with the last 24 months being used to validate the forecasts.

Data from conventional stations often contain gaps; that is, there is a lack of information in certain periods. Before applying the method for forecasting, it is essential to fill these gaps. Initially, the data were tabulated, and gaps were identified. Subsequently, the data from automatic stations were used to complement data from some conventional stations and finally, the Random Forest Regressor technique was applied to fill in the missing data.

2.3. Data Analysis

2.3.1. Random Forest Regressor

The Random Forest Regressor is a machine learning technique widely adopted for missing data imputation due to its ability to combine multiple decision trees, reduce overfitting, and improve predictive accuracy compared to single-tree models [15]. The method relies on random sampling of the training data and the construction of an ensemble of decision trees, which has demonstrated robust performance across a wide range of applications, including datasets with missing values.

Random Forest Regressor was trained using the Scikit-Learn library. Subsequently, the Kolmogorov–Smirnov (KS) test was performed to compare the original and imputed data, and the following evaluation metrics were applied: Mean Absolute Error (MAE) (Equation (1)), Root Mean Squared Error (RMSE) (Equation (2)) and Mean Squared Error (MSE) (Equation (3)):

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - x'_i|, \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - x'_i)^2}, \quad (2)$$

$$MSE = \sum_{i=1}^n (x_i - x'_i)^2, \quad (3)$$

where x_i and x'_i represent the actual results and the expected results, respectively.

Kolmogorov–Smirnov is a statistical and non-parametric test that compares the distribution of two samples and evaluates the similarity between them [23]. The lower the KS, the better the distribution fit. Evaluation metrics are understood as quantitative measures aimed at evaluating the effectiveness and performance of learning models and algorithms [24–26].

In addition to wind speed, the variables average air temperature, atmospheric pressure and wind direction were considered as potential exogenous variables. However, their use depended on the quality of the statistical performance in complementing the data.

2.3.2. Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) Model

The SARIMAX model is employed when it is necessary to incorporate external factors that influence the behavior of a time series [16]. Given the importance of including exogenous variables in the wind speed forecasting process, SARIMAX was selected due to its ability to explicitly integrate external predictors while accounting for seasonal patterns, all with relatively low computational cost.

The SARIMAX model parameters were selected through systematic testing of different parameter configurations and seasonal period combinations. In addition, multiple training–testing splits were evaluated, and the final model was chosen based on a com-

bined criterion incorporating RMSE together with the Akaike (AIC) and Bayesian (BIC) information criteria.

The general form of the SARIMAX model is written as follows [16,27] (Equation (4)):

$$\phi_p(B)\phi_P(B^S)(1-B)^d(1-B^S)^DY_t = c + X_t\beta + \theta_q(B)\theta_Q(B^S)\varepsilon_t, \quad (4)$$

where Y_t is the value of the dependent time series at time t ; ε_t is the residual at time t ; B is the backshift operator ($BY_t = Y_{t-1}$); p , d and q are non-negative integers that denote the order of the autoregressive model, the degree of differentiation, and the order of the moving average model, respectively. In turn, S is the number of periods in each season; P , D , and Q are the autoregressive, differential, and moving average terms for the seasonal part, respectively. In addition, ϕ and θ are autoregressive and moving average coefficients, and β represents how much the exogenous variable (X_t) influences the time series.

In each of the 50 stations located in the Northeast region of Brazil, the wind speed forecasts for 24 months were divided into training and test sets, in three distinct proportions: 80/20, 70/30 and 60/40. Air temperature was considered to be an exogenous variable; in this case, the model performs the tests with and without temperature, providing both results. Then, the results are compared to check which is the best.

Some metrics were used in the evaluation of the model. The Akaike information criterion (AIC) (Equation (5)) and the Bayesian information criterion (BIC) (Equation (6)) were used in conjunction with RMSE to select the best SARIMAX model. Thus, the model sought to minimize error, in addition to avoiding great complexity.

$$AIC = 2(k) - 2\ln(L), \quad (5)$$

$$BIC = (K)\ln(N) - 2\ln(L), \quad (6)$$

where k is the number of model parameters, $\ln(L)$ is the log-likelihood of the model in the data, and N is the sample size [28].

AIC and BIC are used to analyze the fit of the model. AIC is applied with the purpose of identifying the best model regardless of the size of the sample set. In turn, BIC aims to determine the most accurate model considering the sample size. Both penalize the number of parameters in the model in order to avoid overfitting, but BIC does that more rigorously [28].

In addition, it was decided to use the last 24 months of observed data to evaluate and validate the forecasts, as well as the following evaluation metrics: Mean Absolute Error (MAE) (Equation (1)), Root Mean Squared Error (RMSE) (Equation (2)), Mean Squared Error (MSE) (Equation (3)), Mean Absolute Percentage Error (MAPE) (Equation (7)), Mean Absolute Scaled Error (MASE) (Equation (8)) and Nash–Sutcliffe Efficiency Coefficient (NS) (Equation (9)).

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{x_i - \hat{x}_i}{x_i} \right| \cdot 100, \quad (7)$$

$$MASE = \frac{1}{n} \sum_{i=1}^n \left| \frac{x_i - \hat{x}_i}{\frac{1}{n-1} \sum_{i=2}^n |x_i - x_{i-1}|} \right| \cdot 100, \quad (8)$$

$$NS = 1 - \frac{\sum_{i=1}^n (v_{obs} - v_{adj})^2}{\sum_{i=1}^n (v_{obs} - \bar{v}_{obs})^2}, \quad (9)$$

where x_i and $\hat{x}_i$ represent the actual results and the predicted results, respectively, n is the number of observations, and $\bar{v}_{obs}$ is the average of the observed time series.

MAE averages the observed series and the predicted series in order to show the mean absolute errors of the set. In turn, RMSE represents the squared difference between the time series and the predicted series. This metric provides a measure of error dispersion by emphasizing the magnitude of errors [24,25]. MSE represents the average of the squared differences between observed and predicted values. The lowest values point to a high performance of the predicted values [29,30].

MAPE expresses the mean percentage error between the predicted values and the observed values, and is applied to evaluate the performance of the models. If the observed value is very low, any discrepancy can cause MAPE to yield very high results [24], while MASE adjusts for scale differences between datasets by normalizing errors against a reference parameter, such as the mean absolute error of a naïve forecast. It provides a consistent metric for comparing forecasting performance across different time series [31].

NS is usually applied to quantify the quality of the time series. This coefficient can vary from $-\infty$ to 1, and the closer to 1, the better [24]. According to the NS interpretation scale, values greater than 0.75 indicate good model performance, values ranging from 0.36 to 0.75 indicate acceptable performance, and values below 0.36 are regarded as unacceptable [24].

2.4. Logarithmic Extrapolation

As wind speed data from INMET's conventional stations are measured at a height of 10 m, they must be extrapolated to the approximate hub height of wind turbines. In this study, a hub height of 100 m was adopted. According to Monin–Obukhov similarity theory [32], the most accurate wind speed extrapolation methods are those that account for atmospheric stability. Furthermore, Drechsel et al. [33] demonstrated that logarithmic extrapolation to a 100 m hub height provides a reliable and accurate approach.

Wind speed was extrapolated according to Equation (10):

$$W_{100} = W_{10} \frac{\log \frac{z}{z_0}}{\log \frac{z_{ref}}{z_0}}, \quad (10)$$

where W_{100} is the wind speed extrapolated to 100 m height, W_{10} is the observed wind speed at 10 m, z is the height to which the wind speed is being extrapolated (in this case, 100 m), z_0 is the terrain roughness coefficient, and z_{ref} is the reference height (in this case, 10 m).

2.5. Calculation for Generated Electricity

After projecting the data for 24 months, the average monthly wind speed was calculated. Then, together with the power curve of the wind turbine, it was possible to calculate the electricity generated for each location. Table 1 presents the technical specifications of the wind turbine models studied: V80/2000, SWT-2.3-101 and S95/2100 [34].

Table 1. Technical specifications of the wind turbines used in this study.

	Wind Turbines		
	V80/2000	SWT-2.3-101	S95/2100
Manufacturer	Vestas	Siemens	Suzlon
Rated Power	2000 kW	2300 kW	2100 kW
Rotor Diameter	80 m	101 m	95 m
Hub Height	100 m	100 m	100 m

Source: The Wind Power [34].

Figure 2 illustrates the power curves of each wind turbine. The minimum speed at which the wind turbine generates electricity is 3.5 m/s, 3 m/s, and 2.5 m/s in the V80/2000, SWT-2.3-101, and S95/2100 models, respectively. The rated power is 2000 kW (V80/2000), 2300 kW (SWT-2.3-101) and 2100 kW (S95/2100).

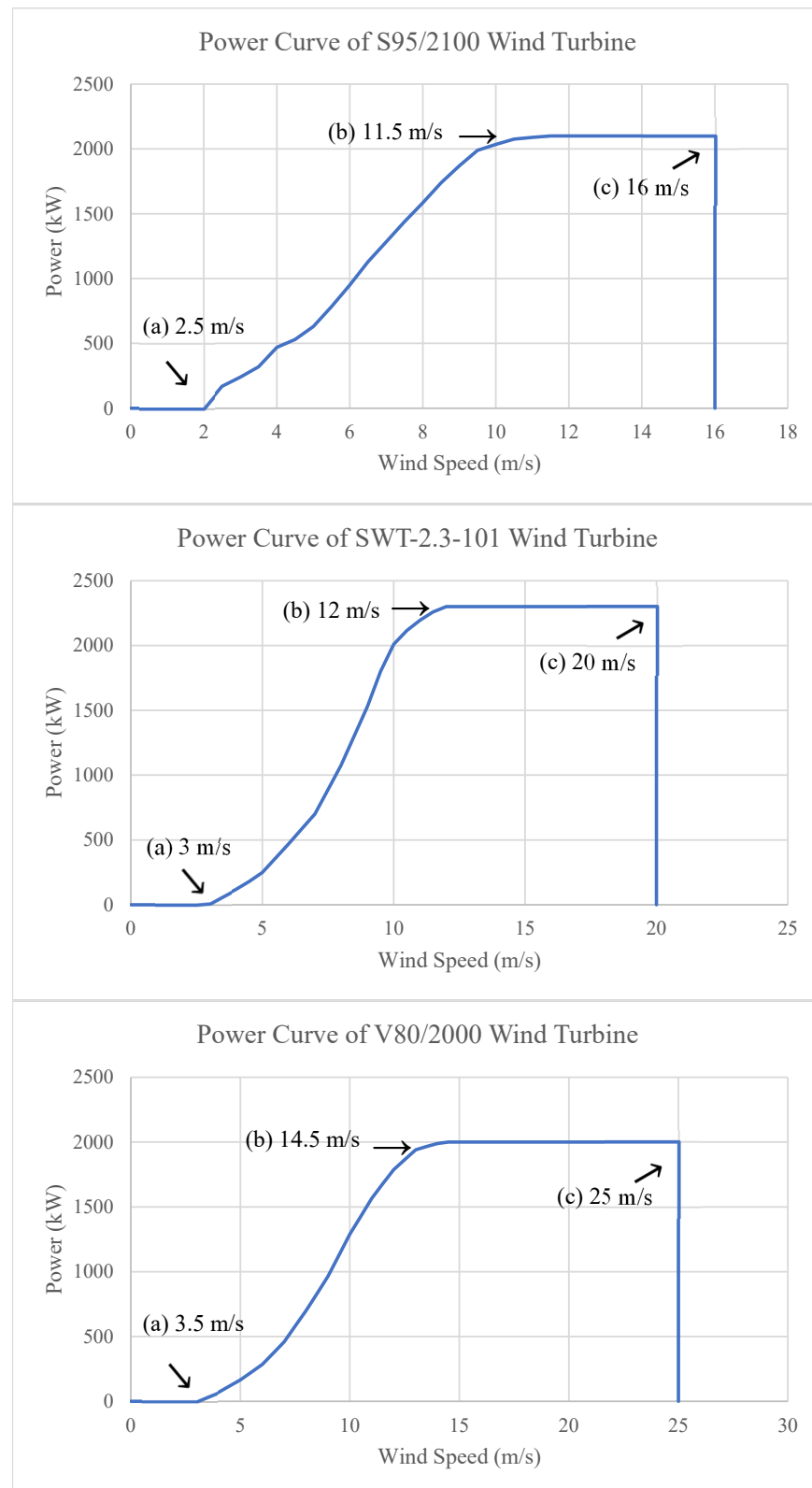


Figure 2. Power curves of wind turbines used in the present study. (a) cut-in speed; (b) rated speed; (c) cut-out speed [34] (The Wind Power, 2025).

The wind turbines of the V80/2000, SWT-2.3-101 and S95/2100 models had different cut-out speed limits. V80/2000 has a cut-out speed of 25 m/s, while SWT-2.3-101 operates with a cut-out speed of 20 m/s, and finally, S95/2100 has a cut-out speed of 16 m/s [34].

The electricity generated was calculated using the turbine power curve and the monthly wind speed in each station, following Equation (11) [35]:

$$E_w = \sum_{i=1}^N P_w U_i (\Delta t) \quad (11)$$

where E_w is the amount of electricity generated by the turbine (kWh); P_w is the turbine power (kW) corresponding to each wind speed; U_i is the wind speed (m/s) in the time interval Δt .

3. Results and Discussion

3.1. Data Filling with Random Forest

The performance of the estimated wind energy generation potential, derived from data imputation using the Random Forest algorithm, was assessed using the RMSE, MSE, and MAE metrics (Table 2).

Table 2. Results of the evaluation metrics.

	RMSE	MSE	MAE
Average air temperature	0.20 to 0.29	0.04 to 0.08	0.16 to 0.23
Average wind speed	0.01 to 0.05	0.0001 to 0.002	0.007 to 0.04
Wind direction	0.06 to 0.19	0.003 to 0.04	0.04 to 0.12
Atmospheric pressure	8.85 to 10.57	78.28 to 111.76	7.05 to 8.45

The results obtained for average air temperature, average wind speed and wind direction indicated that the errors between the observed values and the values imputed by the model were low, showing that the Random Forest method was efficient for these variables (Table 2).

The strong performance of the Random Forest algorithm in imputing average air temperature, wind direction, and mean wind speed is consistent with the findings of Tang and Ishwaran [36], who reported that this approach outperforms traditional imputation techniques such as k-nearest neighbors (KNN).

On the other hand, for atmospheric pressure, RMSE ranged between 8.85 and 10.57, MSE between 78.28 and 111.76, and MAE between 7.05 and 8.45. The amplitude of errors was high, showing that there may be a discrepancy between imputed data and observed data.

The KS test compares the imputed data with the distribution of the observed data. The average temperature and average wind speed parameters showed satisfactory results, demonstrating that the imputation was efficient.

The KS test was applied to check whether the observed values and the imputed values showed a similar distribution. For the average air temperature, the results of the statistical KS test showed values from 0.03 to 0.09, with a p -value between 0.98 and 0.03. For the average wind speed, the statistical KS test ranged between 0.02 and 0.07 and the p -value between 0.07 and 1.00.

The p -values closest to 1 indicated that there is no difference between the distributions; i.e., the model was able to capture the statistical behavior of the data, and the imputation was similar to the observed data. In turn, some values were slightly below 0.05, but remained close to this value. These results below the threshold of 0.05 may indicate that, in some cases, there may be a certain difference between the sets. According to Andrade [37],

the p -value provides a measure of data compatibility with the null hypothesis, but it should not be used alone.

On the other hand, the results of the KS test for the wind direction and atmospheric pressure parameters indicated that the imputation was not done properly. The results of the statistical KS test ranged from 0.10 to 0.51 for wind direction and from 0.31 to 0.42 for atmospheric pressure, while in the p -value, the results were extremely low, ranging from 1.9×10^{-79} to 0.001 for both parameters. These results showed that there are predominantly statistically significant differences between the distributions of the observed and imputed data.

The results of the KS test showed that the data complementation adequately reproduces the wind speed and average air temperature distribution. However, the performance for atmospheric pressure and wind direction is inferior. The poor imputation performance for atmospheric pressure and wind direction can directly affect the quality of the forecasts, as it can generate inaccuracies in the results when applying the SARIMAX model. In turn, the positive performance for the average wind speed and average air temperature variables can contribute to preserving the characteristics of the time series.

After imputation of the missing data and analysis of the statistical results, only variables with satisfactory performance were used as exogenous variables for the wind speed forecasts in the model. The variables wind direction and atmospheric pressure were discarded due to unsatisfactory imputation performance, evidenced by significant differences between the observed and imputed distributions.

Accordingly, only the mean air temperature was selected as an exogenous variable in the model. This choice was not based solely on the satisfactory performance of the data imputation process, but also on the well-established physical relationship between air temperature and wind formation. Thermal gradients induce variations in air density and pressure, which act as primary drivers of atmospheric circulation and, consequently, modulate wind variability [38].

3.2. Results of Forecasts for 24 Months

The results were analyzed by interpreting the values of the metrics used to evaluate the performance of the model in the 50 stations of Northeastern Brazil.

In most seasons, the model showed reliable and consistent performance. The MAE, MSE, RMSE, and MASE values were predominantly less than 1, indicating low error and low deviation from the projected data. These results demonstrated that the model is capable of making good predictions of the data.

Regarding MAPE, the value for most stations was lower than 22%, indicating a relatively low mean error, demonstrating the model's ability to reproduce future data from the observed data. In the specific case of the Seridó station, this result was 34%.

In turn, NS values showed variability between stations, although the values were greater than 0.36 in most of them.

It is worth pointing out that in the São Gonçalo, Barra and São Luís stations, NS values were high, equal to 0.76, 0.75 and 0.72, respectively, while in Paulistana, its value was 0.67 (Figure 3).

In the Floriano and Turiacu stations, NS showed high values of 0.75, while in the Fortaleza, Propriá and Luzilândia stations, NS reached values above 0.80, being 0.81 in the Fortaleza and Luzilândia stations and 0.84 in the Propriá station.

These results demonstrated that the model is capable of making reliable predictions close to the observed values. It is noticeable that there is a good correspondence between the series of observed data and projected data, demonstrating the model's ability to reproduce

the temporal dynamics of these stations. This indicates that the model is capable of properly reproducing the series.

Although the model showed satisfactory results, there was a slight underestimation of the forecasted data in the Areia, Floriano, Seridó and Turiacu stations (Figure 4).



Figure 3. Performance of the average wind speed forecasts for the stations of São Gonçalo, Barra, Paulistana and São Luís for two consecutive years.

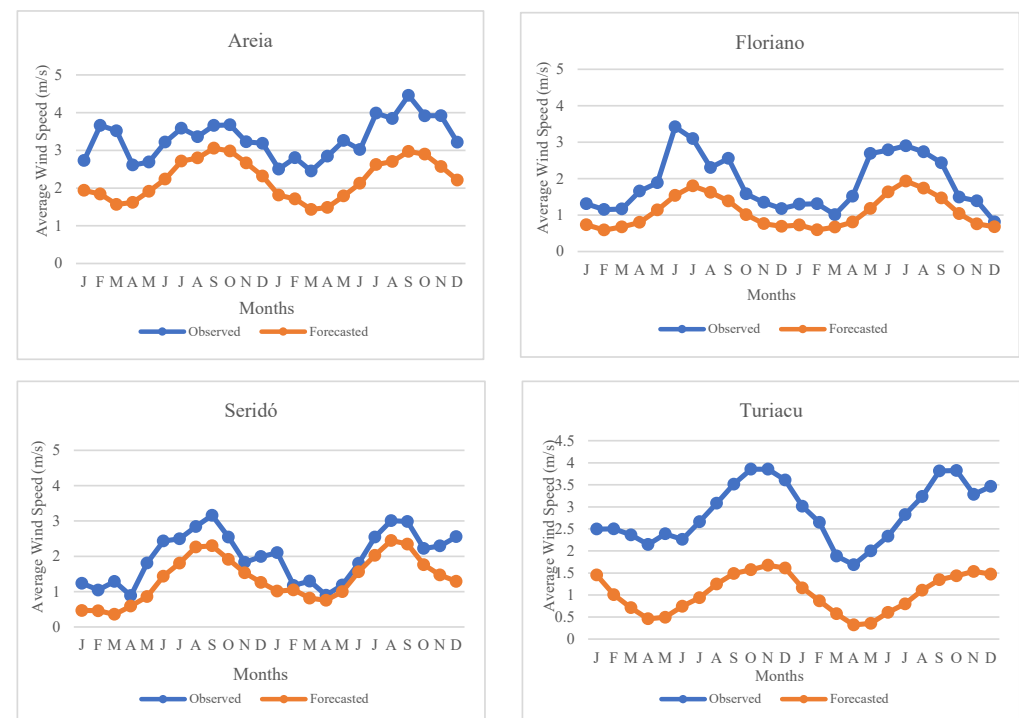


Figure 4. Performance of the average wind speed forecasts for the stations of Areia, Floriano, Seridó and Turiacu for two consecutive years.

For several stations, visual inspection indicates close agreement between the observed and forecasted time series over the analyzed months, although in some cases the model exhibits difficulty in fully capturing short-term oscillations. Periods characterized by higher irregularity reveal limitations in reproducing sharp peaks or abrupt drops. Nevertheless, the model demonstrates overall good forecasting performance, as illustrated in Figure 5.

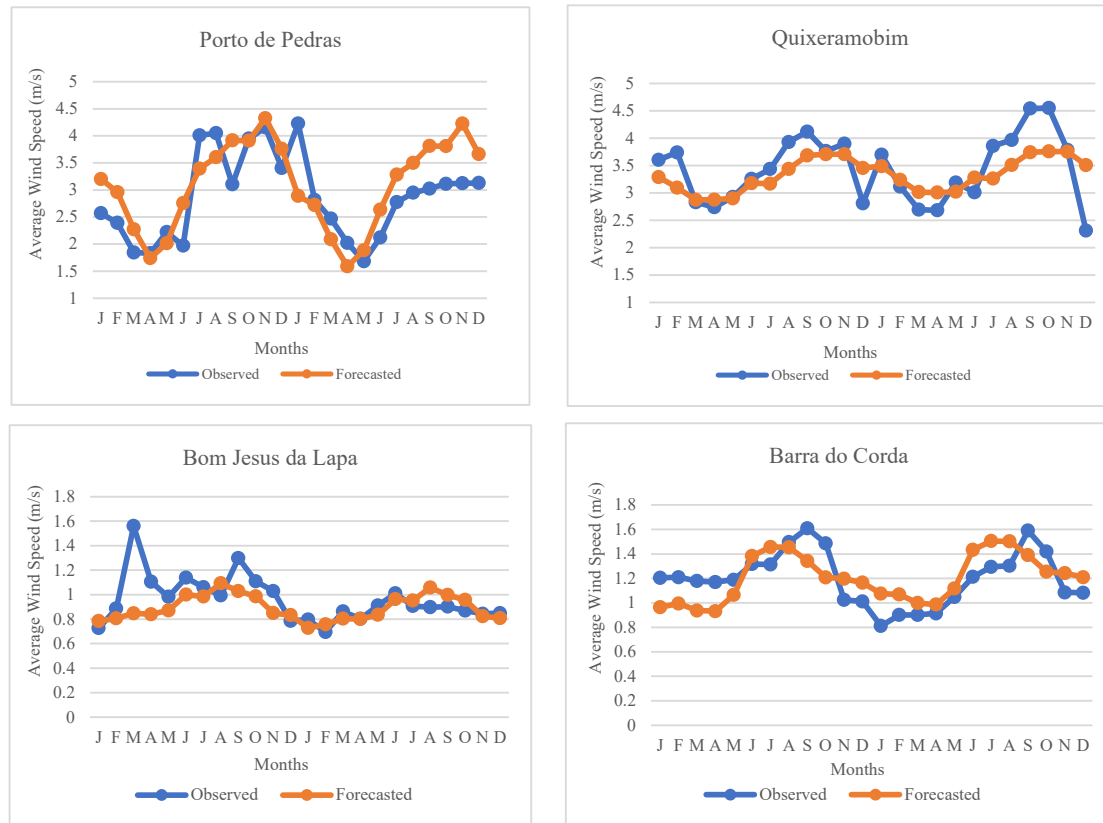


Figure 5. Performance of average wind speed forecasts for the stations of Porto de Pedras, Quixeramobim, Bom Jesus da Lapa and Barra do Corda for two consecutive years.

Although some stations presented Nash–Sutcliffe (NS) values lower than 0.36, the evaluation of the model performance was not restricted to the NS index alone. The assessment also considered the results of the other evaluation metrics employed, in addition to a graphical comparison between observed and predicted time series.

For the stations of Ceará Mirim, Chapadinha, Patos, Vitória da Conquista, Zé Doca, Caravelas, Morro do Chapéu, and Pão de Açúcar, the model adequately captures the seasonal behavior, showing close agreement between observed and forecasted data and a general ability to reproduce the temporal oscillations. Nevertheless, the model exhibits limitations in accurately representing peak values, as increased variability poses challenges to peak prediction (Figure 6).

In the Arco Verde, Balsas, Cipó and Imperatriz stations, the model's performance was unsatisfactory, with NS below 0.36, although the percentage error was between 7% and 19%. In these stations, the model had difficulty reproducing the seasonality of the data. In the Balsas and Cipó stations, the forecast line is virtually linear.

In Arco Verde, the model had difficulty reaching the abrupt peaks, and in Imperatriz the model managed to remain coherent in the first twelve months, but the abrupt change in the data limited its performance (Figure 7).

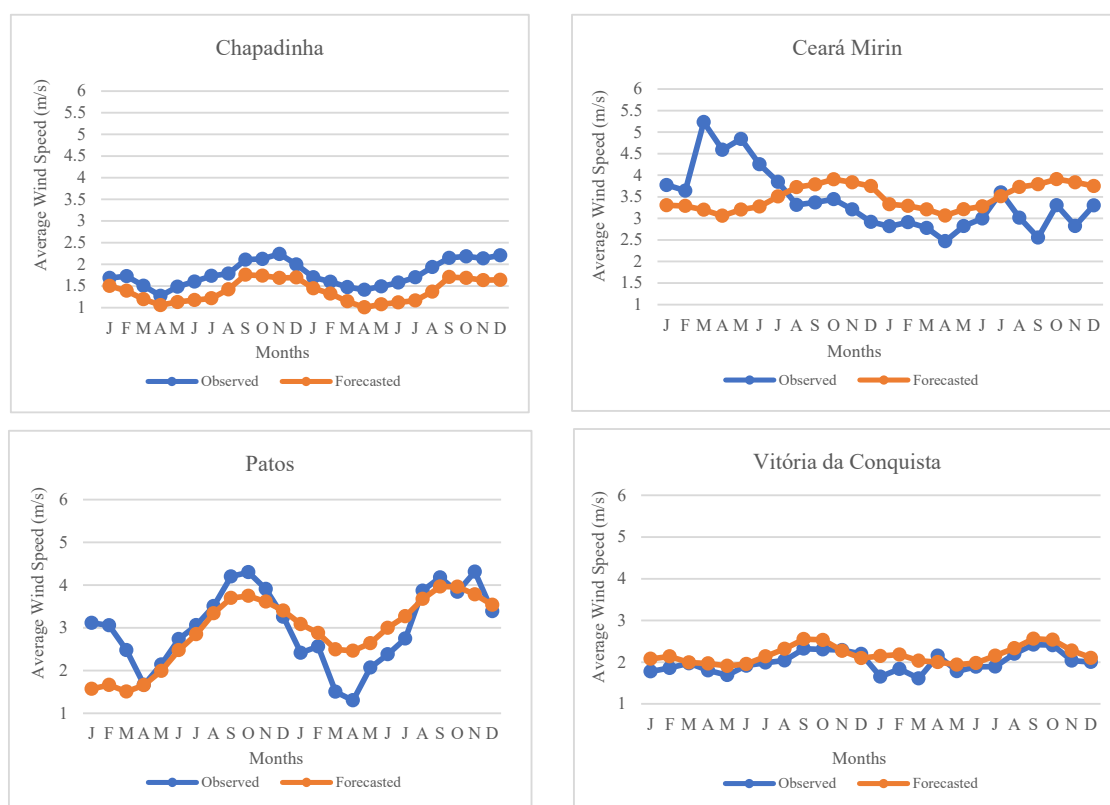


Figure 6. Performance of average wind speed forecasts in two consecutive years for stations that can reproduce the characteristics of the data despite NS less than 0.36.



Figure 7. Performance of average wind speed forecasts for the stations of Arco Verde, Balsas, Cipó and Imperatriz.

The results obtained in this study were compared with those reported by Camelo et al. [24,25,39] who applied hybrid modeling approaches to wind speed forecasting at selected locations in Northeast Brazil. Those studies reported relatively low forecasting errors, with MAPE values of 8.03% for Fortaleza, 7.15% for Natal, and 10.10% for Parnaíba [24]. In contrast, the present study yielded MAPE values ranging from 11% to 21% at the same stations. Similarly, Nascimento et al. [25] reported low MAPE values for Fortaleza, Parnaíba, and Natal.

Regarding the NS efficiency, the values obtained in this study were 0.81 for Fortaleza and 0.41 for Natal. By comparison, Camelo et al. [24] and Camelo et al. [25] reported NS values exceeding 0.80, with values of up to 0.86 for Fortaleza and 0.79 for Natal. These differences in performance may be attributed to variations in modeling strategies, the use of hybrid approaches, local atmospheric characteristics, and differences in data availability and preprocessing procedures, all of which can significantly influence wind forecasting accuracy across regions.

3.3. Potential of Electricity Generated

Monthly electricity generation was estimated using power-curve-based calculations for three wind turbine models (V80/2000, SWT-2.3-101, and S95/2100) combined with 24-month wind speed time series forecasts produced by the SARIMAX model. This approach enabled the estimation of future electricity generation at multiple stations distributed across Northeast Brazil.

At the Água Branca and Natal stations, monthly mean wind speed values exceeded 8 m/s in certain months. These values are comparatively high when contrasted with those reported for several regions in Europe and the United States [40].

At most of the analyzed stations, the mean monthly wind speed exceeds 3.5 m/s. Considering the power curve characteristics of each turbine, these conditions indicate satisfactory potential for electricity generation using the V80/2000 turbine as well as the SWT-2.3-101 and S95/2100 wind turbine models.

When analyzing the power of the wind turbines (Figure 2), it was observed that, although SWT-2.3-101 had the highest rated power, the average winds forecasted for the 24-month period did not exceed 7 m/s. In this case, the S95/2100 wind turbine will generate power sooner and more efficiently than the others.

At some stations, wind speed exhibits pronounced seasonal variability, with periods in which values fall below the turbines' cut-in speed, preventing electricity generation, and other periods during which wind speeds reach levels suitable for effective energy production.

Feng [41] used climate models and analyzed global wind generation capacity considering climate change scenarios. According to the author, South America is expected to show a scenario of low wind yield over time in scenarios of low, medium and high emissions.

In the present study, the stations of Arco Verde, Barra, Barra do Corda, Chapadinha, Correntina, and Floriano exhibited consistently low wind speed values, limiting electricity generation to only a few months of the year. At these locations, mean monthly wind speeds were generally below 3 m/s. Consequently, among the wind turbine models analyzed, only the S95/2100 turbine would be capable of producing electricity, given its lower cut-in speed of 2.5 m/s, at which power generation begins.

At the Bacabal, Balsas, Bom Jesus da Lapa, Colinas, João Pessoa, Palmeira dos Índios, and Salvador stations, mean monthly wind speeds remain below 2.5 m/s, rendering them insufficient for electricity generation using the wind turbines evaluated in this study.

Andrade et al. [35] calculated the electricity generated with data from the Patos and São Gonçalo stations, located in the state of Paraíba, using the wind turbine model SG

3.4–132 for a height of 84 m. The authors observed significant positive trends for wind speed and electricity generation in São Gonçalo, but pointed out that the SG 3.4–132 model is not the most appropriate, since the annual averages of wind speed at the height of the wind turbine do not reach its rated speed. In the present study, at 100 m and applying three models of wind turbines (V80/2000, SWT-2.3-101 and S95/2100) the São Gonçalo station shows relatively low values of average monthly wind speed, but it is still possible to generate electricity, especially for the S95/2100 model, whose cut-in speed is 2.5 m/s and rated speed is 11.5 m/s.

Regarding the Patos station, Andrade et al. [35] also highlighted the increase in wind power generation and an average wind speed of 7 m/s. In the present study, at 100 m height, the wind speed values in the Patos station oscillate over the months; despite that, the conditions of the wind resource are favorable for electricity generation using the V80/2000, SWT-2.3-101 and S95/2100 turbines.

It is important to emphasize that the estimation of the energy generation potential does not account for wake effects between wind turbines, assumes a fixed hub height of 100 m, and does not consider turbine downtime due to maintenance or potential grid constraints. Consequently, the values reported in this study represent an upper-bound, theoretical estimate of wind energy potential, and the actual realizable potential is likely to be lower.

4. Conclusions

This study integrated observed data from 50 conventional meteorological stations using the Random Forest algorithm for data imputation and subsequently produced 24-month wind speed forecasts employing the SARIMAX model. A comprehensive set of evaluation metrics was used to assess both the imputation quality and forecasting performance. The results indicate that the data completion approach was effective for air temperature and wind speed, as evidenced by MAE, MSE, RMSE, and the Kolmogorov–Smirnov test, whereas limited performance was observed for atmospheric pressure and wind direction. Forecasting performance was further evaluated using MAE, MSE, RMSE, MASE, MAPE, and Nash–Sutcliffe efficiency.

For air temperature and wind speed, low MAE, MSE, and RMSE values indicate high agreement and precision relative to the observed data. The KS test also showed that the values did not represent significant differences between the distributions. These results showed that data complementation using Random Forest was successful. On the other hand, for atmospheric pressure and wind direction, the results of MAE, MSE and RMSE were satisfactory, but the KS test showed significant differences between the observed and imputed data. Although the absolute error was low, the distribution of imputed data differs significantly from that of the original data, demonstrating a limitation. Thus, it was decided to use only the average air temperature as an exogenous variable in the forecast of average wind speed data.

The forecast evaluation metrics indicated satisfactory MAE, MSE, and RMSE values. In general, the MASE metric also demonstrated good performance across all stations. MAPE values ranged from 6% to 40%, while the Nash–Sutcliffe efficiency exhibited greater variability, though remaining within acceptable levels for most stations.

In general, the results achieved with the application of the SARIMAX model showed satisfactory performance, as the patterns in most stations are reproduced. In addition, the metrics are predominantly within the expected limits, although reproducing seasonality peaks is difficult in some stations and the results deviate from the data observed over the 24 months in some periods.

It should be noted that the metrics should not be analyzed separately, as each one represents a performance aspect in the model. Therefore, it is essential that the evaluation be carried out jointly for greater reliability of the results; in addition, joint analysis of the metrics is essential to observe the strengths and limitations of the model in each station.

By integrating the power curves of the V80/2000, SWT-2.3-101, and S95/2100 wind turbines with SARIMAX-based forecasts of mean monthly wind speed, this study demonstrates that wind conditions at most of the analyzed stations are suitable for electricity generation using any of the three turbine models. Among them, the S95/2100 turbine shows the broadest applicability due to its lower cut-in wind speed, making it particularly advantageous for sites characterized by moderate or highly variable wind regimes. These findings provide useful insights for wind energy planning and turbine selection, especially in regions with heterogeneous wind resources such as Northeast Brazil.

Author Contributions: Conceptualization, P.R.d.A.S. and R.A.; Methodology, P.R.d.A.S. and R.A.; Software, P.R.d.A.S.; Validation, R.A. and S.E.L.M.; Data Curation, P.R.d.A.S. and L.P.d.S.; Writing—Original Draft Preparation, P.R.d.A.S.; Writing—Review and Editing, R.A., S.E.L.M. and L.P.d.S.; Supervision, R.A. and S.E.L.M.; Funding Acquisition, R.A. and S.E.L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Council for Scientific and Technological Development (CNPq), grant 167824/2022-8 and project 301463/2025-5, and the Paraíba State Research Support Foundation, project 16/2022.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the support of the Laboratory of Environmental and Energy Assessments (LAvAE) of the Federal University of Paraíba (UFPB), the National Council for Scientific and Technological Development (CNPq) and the Paraíba State Research Support Foundation for all the support.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABEEólica	Brazilian Wind Energy Association
AI	Artificial Intelligence
AIC	Akaike information criterion
BIC	Bayesian information criterion
GWEC	Global Wind Energy Council
INMET	National Institute of Meteorology
IPCC	Intergovernmental Panel on Climate Change
KS	Kolmogorov–Smirnov
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MSE	Mean Squared Error
NS	Nash–Sutcliffe Efficiency Coefficient
RF	Random Forest
RMSE	Root Mean Squared Error
SARIMAX	Seasonal AutoRegressive Integrated Moving Average with eXogenous variables
Symbol	Description

B	Backshift Operator
$^{\circ}\text{C}$	Average Air Temperature
E_t	Residual at Time t
E_w	Amount of Electricity Generated by the Turbine
GW	Gigawatts
kWh	Kilowatt-hour
$\ln(L)$	log-likelihood of the model in the Data
mb	Atmospheric Pressure
m/s	Meters per second
N	Sample Size
n	Number of Observations
p, d, q	Non-Negative Integers that Denote the Order of the Autoregressive Model, the Degree Of Differentiation, and the Order of the Moving Average Model, Respectively
P, D, Q	The Autoregressive, Differential, and Moving Average Terms for Tte Seasonal Part, Respectively
P_w	Turbine Power
U_i	Wind Speed (m/s) in the Time Interval Δt
v_{obs}	Average of the Observed Time Series
W_{10}	Observed Wind Speed at 10 m
W_{100}	Wind Speed Extrapolated to 100 m Height
x_i	Actual Results
x^t_i	Expected Results
y_t	Value of the Dependent Time Series at Time t
z	The Height to Which the Wind Speed is Being Extrapolated
z_o	The Terrain Roughness Coefficient
z_{ref}	Reference Height
ϕ and θ	Autoregressive and Moving Average Coefficients

References

1. IPCC. *Climate Change 2021: The Physical Science Basis: Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., et al., Eds.; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2021; p. 2391. Available online: <https://www.ipcc.ch/report/ar6/wg1/> (accessed on 1 February 2026).
2. IPCC. *Climate Change 2023: Synthesis Report: Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Lee, H., Romero, J., Eds.; IPCC: Geneva, Switzerland, 2023; pp. 35–115. [CrossRef]
3. Global Wind Energy Council (GWEC). Global Wind Report 2025. Available online: <https://26973329.fs1.hubspotusercontent-eu1.net/hubfs/26973329/2.%20Reports/Global%20Wind%20Report/GWEC%20Global%20Wind%20Report%202025.pdf> (accessed on 21 August 2025).
4. Ghamati, E.; Kariman, H.; Hoseinzadeh, S. Experimental and computational fluid dynamic study of water flow and submerged depth effects on a tidal turbine performance. *Water* **2023**, *15*, 2312. [CrossRef]
5. Global Wind Energy Council (GWEC). Global Wind Report 2024. Available online: <https://26973329.fs1.hubspotusercontent-eu1.net/hubfs/26973329/2.%20Reports/Global%20Wind%20Report/GWR24.pdf> (accessed on 21 August 2025).
6. Associação Brasileira de Energia Eólica (ABEEólica). *Boletim de Geração Eólica 2022*; ABEEólica: São Paulo, Brazil, 2022.
7. Associação Brasileira de Energia Eólica (ABEEólica). *Boletim Anual Digital 2025*; ABEEólica: São Paulo, Brazil, 2025; Available online: https://abeeolica.org.br/wp-content/uploads/2025/05/424_ABEEOLICA_BOLETIM-ANUAL-DIGITAL-2025_PT_FINAL.pdf (accessed on 11 June 2025).
8. Lucena, J.A.Y.; Lucena, K.A.A. Wind energy in Brazil: An overview and perspectives under the triple bottom line. *Clean Energy* **2019**, *3*, 69–84. [CrossRef]
9. Liu, Z.; Zhang, L. A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings. *Measurement* **2020**, *149*, 107002. [CrossRef]
10. Bilendo, F.; Meyer, A.; Badihi, H.; Lu, N.; Cambron, P.; Jiang, B. Applications and modeling techniques of wind turbine power curve for wind farms—A review. *Energies* **2023**, *16*, 180. [CrossRef]
11. Qin, S.; Liu, D. Distribution characteristics of wind speed relative volatility and its influence on output power. *J. Mar. Sci. Eng.* **2023**, *11*, 967. [CrossRef]

12. Olabi, A.G.; Abdelghafar, A.A.; Maghrabie, H.M.; Sayed, E.T.; Rezk, H.; Al Radi, M.; Obaideen, K.; Abdelkareem, M.A. Application of artificial intelligence for prediction, optimization, and control of thermal energy storage systems. *Therm. Sci. Eng. Prog.* **2023**, *39*, 101730. [CrossRef]
13. Ramos, A.; Carrasco, A.; Fontanet, J.; Herranz, L.E.; Ramos, D.; Díaz, M.; Zazo, J.; Cabellos, O.; Moraleda, J. Artificial intelligence and machine learning applications in the Spanish nuclear field. *Nucl. Eng. Des.* **2024**, *147*, 112842. [CrossRef]
14. Ribeiro, R.; Fanzeres, B. Identifying representative days of solar irradiance and wind speed in Brazil using machine learning techniques. *Energy AI* **2023**, *9*, 100320. [CrossRef]
15. Adewale, M.D.; Ebem, D.U.; Awodele, O.; Sambo-Magaji, A.; Aggrey, E.M.; Okechalu, E.A.; Donatus, R.E.; Olayanju, K.A.; Owolabi, A.F.; Oju, J.U.; et al. Predicting gross domestic product using the ensemble machine learning method. *Syst. Soft Comput.* **2024**, *6*, 200132. [CrossRef]
16. Mulla, S.; Pande, S.; Singh, R. Times Series Forecasting of Monthly Rainfall using Seasonal Auto Regressive Integrated Moving Average with EXogenous Variables (SARIMAX) Model. *Water Resour. Manag.* **2024**, *38*, 1825–1846. [CrossRef]
17. Instituto Brasileiro de Geografia e Estatística (IBGE). Statistical Data. 2025. Available online: <https://www.ibge.gov.br/cidades-e-estados.html?view=municipio> (accessed on 17 November 2025).
18. Instituto Brasileiro de Geografia e Estatística (IBGE). Geosciences Database. 2017. Available online: <https://www.ibge.gov.br/geociencias/downloads-geociencias.html> (accessed on 17 November 2025).
19. Google. *Google Maps*; Google LLC: Mountain View, CA, USA, 2023; Available online: <https://maps.google.com> (accessed on 17 November 2025).
20. Alvares, C.A.; Stape, J.L.; Sentelhas, P.C.; Gonçalves, J.D.M.; Sparovek, G. Köppen's climate classification map for Brazil. *Meteorol. Z.* **2013**, *22*, 711–728. [CrossRef] [PubMed]
21. Lima, F.J.L.; Martins, F.R.; Costa, R.S.; Gonçalves, A.R.; dos Santos, A.P.P.; Pereira, E.B. Seasonal variability and trends of surface solar irradiation in Northeast Brazil. *Sustain. Energy Technol. Assess.* **2019**, *35*, 335–346. [CrossRef]
22. Morales, F.E.C.; Rodrigues, D.T.; Marques, T.V.; Amorim, A.C.B.; Oliveira, P.T.; Silva, C.M.S.; Gonçalves, W.A.; Lucio, P.S. Spatiotemporal Analysis of Extreme Rainfall Frequency in the Northeast Region of Brazil. *Atmosphere* **2023**, *14*, 531. [CrossRef]
23. Wann, J.; Bronkhorst, C.A.; Thoma, D.J. Machine learning enhanced analysis of EBSD data. *Comput. Mater.* **2024**, *10*, 133. [CrossRef]
24. Camelo, H.N.; Lucio, P.S.; Leal, J.B.V., Jr.; de Carvalho, P.C.M. A hybrid model for wind speed forecasting in Northeast Brazil. *Sustain. Energy Technol. Assess.* **2018**, *28*, 65–72. [CrossRef]
25. Camelo, H.N.; Lucio, P.S.; Leal, J.B.V., Jr.; de Carvalho, P.C.M. Proposal for Prediction of Wind Speed through Hybrid Modeling Elaborated from ARIMAX and ANN Models. *Rev. Bras. Meteorol.* **2018**, *33*, 115–129. [CrossRef]
26. De Paula, M.; Casaca, W.; Colnago, M.; da Silva, J.R.; Oliveira, K.; Dias, M.A.; Negri, R. Predicting energy generation in large wind farms using machine learning. *Inventions* **2023**, *8*, 126. [CrossRef]
27. Guo, Y.; Lai, X.; Gan, M. Cyanobacterial biomass prediction using SARIMAX models. *Ecol. Inform.* **2023**, *78*, 102292. [CrossRef]
28. Costa, G.E.M.; Menezes Filho, F.C.M.d.; Canales, F.A.; Fava, M.C.; Brandão, A.R.A.; de Paes, R.P. Assessment of Time Series Models for Mean Discharge Modeling and Forecasting in a Sub-Basin of the Paranaíba River, Brazil. *Hydrology* **2023**, *10*, 208. [CrossRef]
29. Khosravi, A.; Machado, L.; Nunes, R.O. Time-series prediction of wind speed using machine learning. *Appl. Energy* **2018**, *224*, 550–566. [CrossRef]
30. Mattos Netto, P.S.G.; de Oliveira, J.F.L.; de Oliveira Santos, D.S., Jr.; Siqueira, H.V.; Da Nóbrega Marinho, M.H.; Madeiro, F. Hybrid nonlinear system for monthly wind speed forecasting. *IEEE Access* **2020**, *8*, 3032070. [CrossRef]
31. Jainonthee, C.; Sivapirunthep, P.; Pirompu, P.; Punyapornwithaya, V.; Srisawang, S.; Chaosap, C. Modeling and Forecasting Dead-on-Arrival in Broilers Using Time Series Methods: A Case Study from Thailand. *Animals* **2025**, *15*, 1179. [CrossRef] [PubMed]
32. Monin, A.S.; Obukhov, A.M. Basic laws of turbulent mixing in the surface layer of the atmosphere. *Geophys. Obs.* **1954**, *24*, 163–187.
33. Drechsel, S.; Mayr, G.J.; Messner, J.W.; Stauffer, R. Wind speeds at heights crucial for wind energy. *J. Appl. Meteorol. Climatol.* **2012**, *51*, 1602–1617. [CrossRef]
34. The Wind Power. Available online: <https://www.thewindpower.net/> (accessed on 15 September 2025).
35. Andrade, A.R.; Melo, V.F.M.B.; Lucena, D.B.; Abrahão, R. Wind speed trends and electricity generation potential in Brazil. *J. Braz. Soc. Mech. Sci. Eng.* **2021**, *43*, 182. [CrossRef]
36. Tang, F.; Ishwaran, H. Random forest missing data algorithms. *Stat. Anal. Data Min.* **2017**, *10*, 363–377. [CrossRef]
37. Andrade, C. The *p*-value and statistical significance. *Indian J. Psychol. Med.* **2019**, *41*, 210–215. [CrossRef]
38. Wallace, J.M.; Hobbs, P.V. *Atmospheric Science: An Introductory Survey*, 2nd ed.; Elsevier Academic Press: Amsterdam, The Netherlands, 2006.

39. Camelo, H.N.; Sérgio Lucio, P.; Vercosa Leal, J.B., Jr.; Von Glehn dos Santos, D.; Cesar Marques de Carvalho, P. Innovative Hybrid Modeling of Wind Speed Prediction Involving Time-Series Models and Artificial Neural Networks. *Atmosphere* **2018**, *9*, 77. [[CrossRef](#)]
40. Filgueiras, A.; Silva, T.M.V. Wind energy in Brazil—Present and future. *Renew. Sustain. Energy Rev.* **2003**, *7*, 439–451. [[CrossRef](#)]
41. Feng, S. Global wind-power generation capacity under climate change. *Renew. Sustain. Energy Rev.* **2024**, *169*, 113013. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.