

Saulo Custodio de Aquino Ferreira

Proposals for the use of reanalysis bases for wind energy modeling in Brazil

Tese de Doutorado

Thesis presented to the Programa de Pós-Graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção.

> Advisor: Prof. Fernando Luiz Cyrino Oliveira Co-advisor: Prof. Paula Medina Maçaira Louro

> > Rio de Janeiro May 2024



Saulo Custodio de Aquino Ferreira

Proposals for the use of reanalysis bases

for wind energy modeling in Brazil

Thesis presented to the Programa de Pós-Graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção. Approved by the Examination Committee:

> Prof. Fernando Luiz Cyrino Oliveira Advisor Departamento de Engenharia Industrial - PUC-Rio

> Prof. Paula Medina Maçaira Louro Co-advisor Departamento de Engenharia Industrial - PUC-Rio

> > Prof. Reinaldo Castro Souza

Departamento de Engenharia Industrial - PUC-Rio

Prof. José Francisco Moreira Pessanha UERJ

> Prof. André Luís Marques Marcato UFJF

Prof. Leonardo Willer de Oliveira UFJF

Prof. Victor Eduardo Leite de Almeida Duca UFF

Rio de Janeiro, May 9th, 2024

All rights reserved.

Saulo Custodio de Aquino Ferreira

Bachelor's in Production Engineering (2014) at Universidade Federal de Juiz de Fora (UFJF). Master's in Electrical Engineering with an emphasis on Power Systems (2017) at UFJF. During his PhD he worked on several consultancy and R&D projects for the electricity market.

Bibliographic Data

Ferreira, Saulo Custodio de Aquino

Proposals for the use of reanalysis bases for wind energy modeling in Brazil / Saulo Custodio de Aquino Ferreira ; advisor: Fernando Luiz Cyrino Oliveira ; coadvisor: Paula Medina Maçaira Louro. – 2024.

88 f. : il. color. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2024.

Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Base de dados de reanálise. 3. Velocidade do vento. 4. Geração eólica. 5. Correção de viés. 6. Simulação. I. Oliveira, Fernando Luiz Cyrino. II. Louro, Paula Medina Maçaira. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

Acknowledgments

First, I would like to thank God for this achievement. Thank you for leading me on the path of good, for giving me strength and wisdom, and for helping me overcome all the obstacles I encountered during my PhD.

I thank my entire family, especially my parents, Sandra Maria de Aquino Ferreira and Sebastião Custodio Ferreira Filho, and my sister, Sandra Carolina de Aquino Ferreira, who were always by my side, supporting me, encouraging me, and not letting me give up on times more difficult.

I sincerely thank my supervisors, Prof. Fernando Luiz Cyrino Oliveira and Prof. Paula Medina Maçaira Louro, for their teachings and encouragement throughout my work.

To anyone who in any way contributed to my journey.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Ferreira, Saulo Custodio de Aquino; Oliveira, Fernando Luiz Cyrino (Advisor). Louro, Paula Medina Maçaira (Co-advisor). **Proposals for the use of reanalysis bases for wind energy modeling in Brazil.** Rio de Janeiro, 2024. 88p. Tese de Doutorado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Brazil's energy landscape has historically relied heavily on renewable sources, notably hydropower, with wind energy emerging as a significant contributor in recent years. Understanding and harnessing the potential of wind energy necessitates robust modeling of its behavior. However, obtaining comprehensive wind speed and generation data, particularly in specific locations of interest, remains a challenge. In the absence of wind speed data, an alternative is to use data from a reanalysis database. They provide long histories of data on climatic and atmospheric variables for different parts of the world, free of charge. Therefore, the first contribution of this work focused on verifying the representativeness of wind speed data made available by MERRA-2 in Brazilian territory. Following literature recommendations, interpolation, extrapolation, and bias correction techniques were used to improve the adequacy of the speeds provided by the reanalysis based on those that occur at the height of the wind farm turbine rotors. In a second contribution, MERRA-2 data was combined with power measured in Brazilian wind farms to model in a stochastic and non-parametric way the relationship between speed and power in wind turbines. For this purpose, clustering, density curve estimation, and simulation techniques were used. Finally, the research culminates in the development of an application within the Shiny environment, offering a user-friendly platform to access and apply the methodologies devised in the preceding analyses. By making these methodologies readily accessible, the application facilitates broader engagement and utilization within the research community and industry practitioners alike.

Keywords

Reanalysis dataset; Wind speed; Wind power; Bias correction; Non-parametric estimation; Simulation.

Resumo

Ferreira, Saulo Custodio de Aquino; Oliveira, Fernando Luiz Cyrino (Orientador); Louro, Paula Medina Maçaira (Co-orientadora). **Propostas do uso de bases de reanálise para modelagem de energia eólica no Brasil**. Rio de Janeiro, 2024. 88p. Tese de Doutorado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

O Brasil sempre foi um país que teve sua matriz elétrica pautada majoritariamente em fontes renováveis, mais especificamente na hídrica. Com passar dos anos, esta tem se diversificado e demonstrado uma maior participação da fonte eólica. Para melhor explorála, pesquisas visando modelar seu comportamento são essenciais. Entretanto, não é sempre que se tem dados de velocidade do vento e de geração eólica disponíveis em quantidade e nas localidades de interesse. Esses dados são primordiais para identificar potenciais locais de instalação de parques eólicos, melhorar o desempenho dos existentes e estimular pesquisas de previsão e simulação da geração eólica que são entradas para auxiliar na melhor performance do planejamento e da operação do setor elétrico brasileiro. Na carência de dados de velocidade do vento, uma alternativa é o uso de dados vindos de base de reanálises. Elas disponibilizam longos históricos de dados de variáveis climáticas e atmosféricas para diversos pontos do globo terrestre e de forma gratuita. Desta forma, a primeira contribuição deste trabalho teve como foco a verificação da representatividade dos dados de velocidade do vento, disponibilizados pelo MERRA-2, no território brasileiro. Seguindo as recomendações da literatura, utilizou-se técnicas de interpolação, extrapolação e correção de viés para melhorar a adequação as velocidades fornecidas pela base de reanalise as que acontecem na altura dos rotores das turbinas dos parques eólicos. Em uma segunda contribuição combinou-se os dados do MERRA-2 com os de potência medidas em parques eólicos brasileiros para modelar de modo estocástico e não paramétrico a relação existente entre a velocidade e potência nas turbinas eólicas. Para isto utilizou-se as técnicas de clusterização, estimação das curvas de densidade e simulação. Por fim, em uma terceira contribuição, desenvolveu-se um aplicativo, no ambiente shiny, para disponibilizar as metodologias desenvolvidas nas duas primeiras contribuições.

Palavras-chave

Dados de base de reanálise; Velocidade do vento; Geração eólica; Correção de viés; Estimação não paramétrica; Simulação.

Contents

1 Introduction	14
2 First Contribution: Validation of the representativeness of wind	
speed time series obtained from reanalysis data for Brazilian territory	18
2.1 Abstract	18
2.2 Introduction	18
2.3 Data and methodology	22
2.3.1 Obtaining reanalysis data	23
2.3.2 Identifying wind turbines	24
2.3.3 Data treatment	25
2.3.4 Interpolation	26
2.3.5 Extrapolation	27
2.3.6 Bias correction	28
2.3.6.1 The Global Wind Atlas	28
2.3.6.2 INMET	29
2.3.7 Evaluation	30
2.4 Results	31
2.4.1 Hourly results	34
2.4.2 Effects of the use of different time scales	37
2.4.3 Impact of different spatial scales	38
2.5 Discussion	40
2.6 Conclusions	42
2.7 Acknowledgments	43
2.8 Appendix A. Supplementary material	43
3 Second Contribution: Joint modeling of wind speed and power via a	
nonparametric approach	44
3.1 Abstract	44
3.2 Introduction	44
3.3 Proposed Method	47
3.3.1 Wind farm data	47
3.3.2 Wind speed data	48

3.3.3 Modeling the speed versus power relationship49

3.3.3.1 Clustering of wind speeds	49
3.3.3.2 Density curve estimation	49
3.3.3.3 Simulation	50
3.3.4 Power curve	51
3.3.5 Evaluation	52
3.4 Data	53
3.4.1 Wind power data	53
3.4.2 Wind turbines	55
3.5 Results and Discussion	56
3.6 Final Considerations	68
3.7 Appendix A. Supplementary material	69
4 Third Contribution: Application to obtain wind speed and wind	

generation data for Brazil.	70
4.1 Introduction	70
4.2 Application	70
4.2.1 Wind Speed	71
4.2.2 Wind Power	74
4.3 Conclusion	80
5 Summary of contributions and avenues for future research	81
References	82

List of Figures

Figure 1.1 Brazilian Electrical Matrix	15
Figure 1.2 Largest wind generations from 2019 to 2021	15
Figure 1.3 Largest installed wind capacity from 2019 to 2022	16
Figure 2.1 Overview of the methods used	23
Figure 2.2 Brazilian territory and installed wind power capacity	24
Figure 2.3 Density curves of wind speed data measured at the turbines	25
Figure 2.4 Monthly boxplot of wind speed data measured at the turbines	26
Figure 2.5 Hourly boxplot of wind speed data measured at the turbines	26
Figure 2.6 Nearest neighbor interpolation	27
Figure 2.7 Selection of data for FCGWA calculation	28
Figure 2.8 Selection of data for FCINMET calculation	29
Figure 2.9 Correlation boxplots of the measured hourly history of the turbines with the treated reanalysis time series	32
Figure 2.10 RMSE boxplots of the measured hourly history of the turbines with the treated reanalysis time series	32
Figure 2.11 MBE boxplots of the measured hourly history of the turbines with the treated reanalysis time series	33
Figure 2.12 MAE boxplots of the measured hourly history of the turbines with the treated reanalysis time series	33
Figure 2.13 Boxplots of Variances diff. of the measured hourly history of the turbines with the treated reanalysis time series	33
Figure 2.14 MERRA-2 data versus measured data scatter plots and histograms	36

Figure 3.1 Steps of the method	47
Figure 3.2 Theoretical wind power curve	51
Figure 3.3 Density curve and boxplot of the wind farms' power data	54
Figure 3.4 Boxplot of the wind power data per hour and month	55
Figure 3.5 Result of clusterization (Wind Farm 5)	58
Figure 3.6 Wind power density curves (Wind Farm 5)	59
Figure 3.7 Dispersion graphs of the historic versus estimated data (Wind Farm 5)	66
Figure 3.8 Polygon graphs of the estimated and historic frequency data (Wind Farm 5)	67
Figure 4.1 Application home screen	71
Figure 4.2 The wind speed tab beginning	72
Figure 4.3 Interpolation result	72
Figure 4.4 Extrapolation step	73
Figure 4.5 Bias correction step	73
Figure 4.6 Type of correction	73
Figure 4.7 Result from the wind speed tab	74
Figure 4.8 The wind power tab beginning	75
Figure 4.9 Input wind speed data	76
Figure 4.10 Building the wind speed data	76
Figure 4.11 Application of the methodology	77
Figure 4.12 Result of modeling the relationship between wind speed	
and wind generation	78
Figure 4.13 Final screen with scenarios and average estimated generation time series	79
Figure 4.14 Final screen with scenarios and average estimated generation time series	79

List of Tables

Table 1.1 Renewable energy scenario in recent years	14
Table 2.1 Treatment of the MERRA-2 time series	32
Table 2.2 Medians of treatment strategies per metric	34
Table 2.3 The range between the first and third quartiles of results of	
treatment strategies per metric	35
Table 2.4 Range of results of treatment strategies by metric	35
Table 2.5 Medians of treatment strategy results	38
Table 2.6 Metric results for different spatial aggregations (Analysis 1)	39
Table 2.6 Metric results for different spatial aggregations (Analysis 2)	39
Table 3.1 Information about the wind farms	53
Table 3.2 Turbine models at the wind farms	56
Table 3.3 Data from the process of adjusting the wind speed data from	
MERRA-2	56
Table 3.4 Turbine models adopted in this study	57
Table 3.5 Number of clusters obtained by each approach	58
Table 3.6 Results reached by the methods according to the evaluation	
metrics	61
Table 3.7 RMSE attained by the methods per month	63
Table 3.8 RMSE attained by the methods per hour	65

List of abbreviations

ANEEL	Agência Nacional de Energia Elétrica				
ANN	Artificial Neural Networks				
BA	Bahia				
BANN	Bayesian Artificial Neural Networks				
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior				
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico				
DTU Wind Energy	The Department of Wind Energy of the Technical University of Denmark				
DISC	Data and Information Services Center				
DISPH	Displacement Height				
DT	Decision Trees				
ECMWF	European Centre for Medium-Range Weather Forecasts				
FAPERJ	Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro				
FDP	Função de Densidade de Probabilidade				
GEOS	Goddard Earth Observing System Model				
GES	Goddard Earth Sciences				
GMAO	Global Modeling and Assimilation Office				
GMCM	Gaussian Mixture Copula Model				

GWA	Global Wind Atlas			
INMET	Instituto Nacional de Meteorologia			
KDE	Kernel Density Estimation			
LOESS	Locally Estimated Scatterplot Smoothing			
MAE	Mean Absolute Error			
MAPE	Mean Absolute Percentage Error			
MBE	Mean Bias Error			
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications Dataset Version 2			
NASA	National Aeronautics and Space Administration			
NetCDF	Network Common Data Form			
ONS	National Electrical System Operator			
PDF	Probability Density Function			
R ²	Coefficient of Determination			
RF	Random Forests			
RMSE	Root Mean Squared Error			
RN	Rio Grande do Norte			
SIGEL/ANEEL	Electric Sector Georeferenced Information System of Brazil's National Electric Energy Agency			
USA	United States of America			

1 Introduction

In recent decades, the impact of climate change caused by fossil fuels and growing concern about the limitations of these resources have motivated research into alternative energy sources. Energy from renewable sources has become an option that is not only ecologically correct but also economically viable to meet global energy demand (DE ASSIS TAVARES et al., 2020; DUCA; FONSECA; CYRINO OLIVEIRA, 2023; FREITAS et al., 2020).

According to data from the International Renewable Energy Agency (IRENA, 2023), in 2021, 7,858.2 TWh were generated worldwide from renewable sources, an increase of 5.4% compared to the previous year, which generated 7,456. 1 TWh. As shown in Table 1.1, this increase has occurred over time, mainly due to the expansion of the installed capacity of these sources. In the years evaluated, it was also noted that China, the United States, and Brazil are the countries with the largest installed capacities and renewable energy production, with Brazil being third on this scale.

Capacity installed (MW)					
	2019	2020	2021	2022	
World	2.543.377,9	2.813.159,2	3.077.238,3	3.371.792,6	
China	789.134,0	929.944,7	1.056.624,2	1.206.588,9	
United States	282.844,8	312.655,1	345.401,3	370.963,7	
Brazil	144.574,6	150.493,1	161.135,9	175.261,9	
Power (GWh)					
	2019	2020	2021	2022	
World	6.994.749,0	7.456.090,0	7.858.208,0	-	
China	2.017.941,2	2.183.034,4	2.444.538,0	-	
United States	787.805,9	848.459,9	886.891,6	-	
Brazil	515.479,5	522.981,6	507.666,6	-	

Table 1.1 Renewable energy scenario in recent years.

Brazil presents a growing investment and expansion of renewable sources in order to follow the global trend. The share of renewable energies (biomass, wind, hydro, solar, and undielectric) in the Brazilian electrical matrix is significantly more significant than the global matrix. While in Brazil, renewable sources represent 83.68%, in the world, it is only 28.60% (EPE, 2023; SIGA ANEEL, 2023). After the 2001 energy crisis, Brazil had a great incentive to diversify its electrical matrix. The current composition of the Brazilian matrix is presented in Figure 1.1, where the growing contribution of wind, solar, and biomass generations is noted (EPE, 2023; SIGA ANEEL, 2023).



Figure 1.1 Brazilian Electrical Matrix.

Wind generation is produced through the passage of wind through turbines, and the amount of energy generated is proportional to wind speed, weather conditions, and the time of day and year. It has two production modes: onshore, which uses turbines installed on land, and offshore, where the turbines are installed on the high seas. Brazil has great potential for exploiting wind energy in both forms, mainly in the Northeast region. Another relevant point of this source about Brazil is that in periods of low reservoir levels, it has its maximum potential, thus enabling seasonal complementarity when combined with the water regime (DUCA; FONSECA; CYRINO OLIVEIRA, 2023; LINS et al., 2023).

Figures 1.2 and 1.3 show the evolution of wind generation and its installed capacity, respectively, over the years and the leading nations exploiting this source. Brazil is the fourth largest producer of wind energy, rising two positions compared to the last two years. Regarding its installed capacity, we are the seventh largest nation, higher than in 2019 and 2020. This demonstrates Brazil's great potential for wind generation, as even with less installed capacity than other nations, we generate more wind energy (IRENA, 2023).



Figure 1.2 Largest wind generations from 2019 to 2021.



Figure 1.3 Largest installed wind capacity from 2019 to 2022.

The diversification of the electrical matrix combined with the incredible insertion of renewable sources imposes more significant challenges for the planning and operation of the electrical sector due to the natural variability of these sources. In an attempt to mitigate these challenges, studies on the stochastic characteristics of sources and their inputs in the territory are essential to allow an understanding of their variability and natural peculiarities, thus enabling reliable and realistic projections of renewable energy generation, which are extremely important for planning to meet demands and costs (MAÇAIRA et al., 2019; PERINI DE SOUZA et al., 2022a).

Based on this scenario, this research attempts to minimize the challenges of wind generation by developing a methodology that improves wind power estimation from turbines. The main strategies adopted are deterministic and parametric; they do not consider that a particular wind speed value can generate different powers (CARRILLO et al., 2013). Whether at different times for the same turbine or the same instant in different turbines but of the same model and spatial region (AL-QURAAN et al., 2022; DE AQUINO FERREIRA; CYRINO OLIVEIRA; MAÇAIRA, 2022).

Extensive historical measurements of variables linked to intermittent sources are necessary to carry out these studies. However, the availability of these data is often insufficient, resulting in the need to search for alternatives, such as reanalysis data (ESTEVES et al., 2019).

The availability and suitability of these variables at different temporal and spatial scales are essential for developing products that feed the various Brazilian energy optimization models, such as NEWAVE, DECOMP, and DESSEM (MAÇAIRA et al., 2019).

Once these limitations are overcome, research involving the prediction and simulation of wind generation becomes viable. These products serve as the basis for correct management of the Brazilian electricity sector, guiding the best choices for expanding and operating the existing structure (AL-DUAIS; AL-SHARPI, 2023; AL-QURAAN et al., 2022; WORTON, 1995).

This work is divided into five chapters. This introduction describes our motivations for developing this study and explains the theme's relevance. The second section is the first product of this PhD, the paper article published in the Energy Journal that deals with validating the MERRA-2 wind speed time series for the territory of Brazil. The third section is the second essay research on developing a non-parametric methodology to estimate wind power from wind speed. It is also intended to be published in Energy Journal (currently submitted). The fourth chapter culminates in the development of an application within the Shiny environment, offering a user-friendly platform to access and apply the methodologies devised in the preceding analyses. Finally, the fifth section presents our final considerations.

2

First Contribution: Validation of the representativeness of wind speed time series obtained from reanalysis data for Brazilian territory

This chapter is based on the first contribution of this thesis, already published in the Energy journal. See https://doi.org/10.1016/j.energy.2022.124746

2.1

Abstract

In recent years, consideration of reanalysis data has gained space and importance globally as a promising alternative for climate studies that suffer from an absence or scarcity of data. Wind speed time series can be obtained from these bases for various purposes, such as inferring the potential of sites for wind power generation. These projections can be useful to analyze the feasibility of building new wind farms and the formation of historical series of wind power generation to enable better planning for existing facilities. Therefore, reliable wind speed time series is essential to obtain accurate projections. The reanalysis databases are characterized for having extended historical series. On the other hand, one of their drawbacks is the arrangement of data in a grid with low spatial resolution, so not cover all points on the Earth's surface. This study aims to verify whether the wind speed time series of the MERRA-2 dataset can represent the values at points in Brazilian territory. For this purpose, we examine the use of strategies for interpolation, extrapolation, and bias correction to overcome these limits and obtain time series that better approximate the most probable values, as suggested in the specialized literature. The results are compared with historic series recorded in Brazil to evaluate the method's applicability and indicate whether the data extracted from MERRA-2, after treatment, provide a relevant representation. This study contributes to the literature by (i) measuring the quality of MERRA-2 data to represent high spatial resolution locations in Brazil, (ii) evaluating the impacts of the natural variability of these wind speed series on the results, (iii) describing new bias correction approaches, (iv) verifying the impact of the temporal and spatial scales utilized on the results, and (v) assessing the results by comparing wind speeds.

2.2

Introduction

In recent years, the Brazilian electricity mix has diversified, with increased participation of renewable sources. In particular, wind generation has grown significantly, especially in the Northeast region. Currently, it is responsible for 10.7% of the installed capacity in the country, which represents 17.82 MW per

month. In 2020, wind power amounted to 56.99 GWh, corresponding to 9.80% of Brazil's generation (ONS, 2021).

This greater inclusion of renewable sources brings significant challenges for the planning and operation of the Brazilian electricity system due to the stochastic nature of the sources. In this sense, reliable time series of renewable energy generation is of great importance to determine the needs and costs of operation, since they allow understanding the variability and peculiarities of each source in each production unit (GONZÁLEZ-APARICIO et al., 2017; HAYES; STOCKS; BLAKERS, 2021).

The generation time series comes from the data measured at power plants. However, these direct measurements are not always available for all locations. Furthermore, in many cases the length of the time series is limited, and the data are incomplete or contain measurement errors. Feasibility studies of new wind farms also require generation projections at candidate's sites based on reliable data. In their absence, they can be constructed synthetically from reanalysis data. Reanalysis databases have become a promising alternative for estimating generation due to data availability for almost the entire Earth surface, with a long data history. Among several available options, these data also include time series of wind speed, the primary input to estimate wind generation (GONZÁLEZ-APARICIO et al., 2017; GRUBER et al., 2019).

Currently, the main datasets are the Modern-Era Retrospective analysis for Research and Applications dataset version 2 (MERRA-2) (GELARO et al., 2017), an evolution of the Modern-Era Retrospective analysis for Research and Applications (MERRA) dataset provided by National Aeronautics and Space Administration (NASA), and ERA5, the fifth version of the European Centre for Medium-Range Weather Forecasts (ECMWF) dataset (GRUBER et al., 2019). MERRA-2 is one of the most widely used reanalysis datasets in the literature to obtain wind speed time series (GRUBER et al., 2019; OLAUSON; BERGKVIST, 2015). However, it has some disadvantages that we intend to overcome in this study. Wind speed data are available at three different heights (2, 10, and 50 m), but the speed contributing to wind generation is at the turbine height, frequently at 100 m (GELARO et al., 2017). In these reanalysis databases, the data are only available in discrete intervals of space and time, with a spatial resolution of approximately 50 km between the points, which can cause regional bias (GRUBER; SCHMIDT, 2019). In general, the reanalysis data have two sources of errors: a systematic one, due to the nature of the physical assimilation model used; and a random one, arising from the local characteristics of the terrain (NEFABAS et al., 2021). To overcome these limitations, previous studies have combined strategies such as interpolation (BOSCH; STAFFELL; HAWKES, 2018; CRADDEN et al., 2017; GRUBER; SCHMIDT, 2019; NEFABAS et al., 2021; RYBERG et al., 2019; STAFFELL; GREEN, 2014; STAFFELL; PFENNINGER, 2016), extrapolation (GRUBER et al., 2019; GUALTIERI, 2021; NEFABAS et al., 2021; RYBERG et al., 2019) and bias correction (BOSCH; STAFFELL; HAWKES, 2018; GRUBER et al., 2019, 2021; GRUBER; SCHMIDT, 2019; NEFABAS et al., 2021; RYBERG et al., 2019; STAFFELL; PFENNINGER, 2016).

There is usually no coincidence between the point of the reanalysis grid and the location under study. Therefore, the interpolation strategy is a way to overcome this problem. Among the main interpolation techniques applied are bilinear interpolation (BOSCH; STAFFELL; HAWKES, 2018; CRADDEN et al., 2017; NEFABAS et al., 2021; RYBERG et al., 2019), nearest-neighbor (BRUNE; KELLER; WAHL, 2021; GRUBER et al., 2021; GRUBER; SCHMIDT, 2019; SHERIDAN et al., 2022; STAFFELL; PFENNINGER, 2016), cubic-spline interpolation (MURCIA et al., 2022), distance weighting (SHERIDAN et al., 2022) and locally estimated scatterplot smoothing (LOESS) (STAFFELL; GREEN, 2014; STAFFELL; PFENNINGER, 2016). Gruber et al. (2019) investigated the possible impacts of different interpolation methods. They concluded that the results were similar and recommended using the nearest neighbor technique because it requires less computational effort.

Extrapolation consists of adapting the value of wind speed to the value corresponding to the height of the wind turbine rotor. The Hellman power law is the wind speed extrapolation approach that has recently achieved the best results and greatest use (GRUBER et al., 2021; GUALTIERI, 2021; MURCIA et al., 2022; NEFABAS et al., 2021; SHERIDAN et al., 2022).

Bias correction is an alternative to minimize the remaining deviation between real and estimated wind speeds. Previous studies have used other data sources that have higher spatial resolutions for correction, such as the Global Wind Atlas (GWA) (BOSCH; STAFFELL; HAWKES, 2018; GRUBER et al., 2019; RYBERG et al., 2019), local country databases (GRUBER et al., 2019), and historical data from power plants (STAFFELL; PFENNINGER, 2016). This bias correction can be achieved by applying a factor that corrects the mean of the time series (GRUBER; SCHMIDT, 2019; MURCIA et al., 2022) or by a statistical scale reduction (GONZÁLEZ-APARICIO et al., 2017; NEFABAS et al., 2021).

Studies typically convert wind speed time series into wind generation series to measure capacity factors of regions (GUALTIERI, 2021; PRYOR; LETSON; BARTHELMIE, 2020; SHERIDAN et al., 2022), or to compare them with historical generation data, when available (GRUBER et al., 2019; NEFABAS et al., 2021), to evaluate the quality of the wind speed time series of the reanalysis bases and the strategies applied to them. The wind speed to wind power conversion occurs through the application of generic curves (GRUBER et al., 2021; RYBERG et al., 2019), or specific power curves of turbines in the wind farms under study (NEFABAS et al., 2021). In recent years, several studies have evaluated the wind speed time series of the reanalysis bases through comparison with measured data from meteorological stations (BRUNE; KELLER; WAHL, 2021; GUALTIERI, 2021; MOLINA; GUTIÉRREZ; SÁNCHEZ, 2021; PRYOR; LETSON; BARTHELMIE, 2020; RABBANI; ZEESHAN, 2020; SHERIDAN et al., 2022) or for wind turbines (BRUNE; KELLER; WAHL, 2021; MURCIA et al., 2022). However, these works focus more on European countries (BRUNE; KELLER; WAHL, 2021; GUALTIERI, 2021; JOURDIER, 2020; MOLINA; GUTIÉRREZ; SÁNCHEZ, 2021; MURCIA et al., 2022), the United States of America (USA) (GUALTIERI, 2021; SHERIDAN et al., 2022), and other countries (JIANG et al., 2021; RABBANI; ZEESHAN, 2020; REN et al., 2019) that have policies to encourage the measurement and availability of wind speed data, unlike Brazil.

The main evaluation metrics used in these works comparing data are Pearson correlation (BRUNE; KELLER; WAHL, 2021; CRADDEN et al., 2017; GRUBER et al., 2019; GUALTIERI, 2021; KHATIBI; KRAUTER, 2021), root mean squared error (RMSE) (BRUNE; KELLER; WAHL, 2021; CRADDEN et al., 2017; KHATIBI; KRAUTER, 2021; NEFABAS et al., 2021; OLAUSON, 2018; RABBANI; ZEESHAN, 2020), mean bias error (MBE) (GRUBER et al., 2019; JIANG et al., 2021; MURCIA et al., 2022; RABBANI; ZEESHAN, 2020), mean absolute error (MAE) (KHATIBI; KRAUTER, 2021; NEFABAS et al., 2021; REN et al., 2019; SHERIDAN et al., 2022), descriptive statistics (GUALTIERI, 2021; MURCIA et al., 2022), curves (BRUNE; KELLER; WAHL, 2021; MURCIA et al., 2022) and parameters of the Weibull distribution (RABBANI; ZEESHAN, 2020).

There is a prevalence of studies that contemplate larger scales, both spatial (continents, countries, and regions) and temporal (year, quarter, month, and day) (OLAUSON; BERGKVIST, 2015; STAFFELL; PFENNINGER, 2016). Most studies have been carried out for the European continent or its countries (BRUNE; KELLER; WAHL, 2021; CRADDEN et al., 2017; GUALTIERI, 2021; JOURDIER, 2020; KHATIBI; KRAUTER, 2021; MOLINA; GUTIÉRREZ; SÁNCHEZ, 2021; MURCIA et al., 2022; OLAUSON, 2018; OLAUSON; BERGKVIST, 2015; RYBERG et al., 2019). There are also studies related to the USA (GUALTIERI, 2021; KHATIBI; KRAUTER, 2021; OLAUSON, 2018; PRYOR; LETSON; BARTHELMIE, 2020; SHERIDAN et al., 2022) and China (JIANG et al., 2021; REN et al., 2019), in addition to preliminary studies in Brazil (GRUBER et al., 2019), South Africa (GRUBER et al., 2021; GUALTIERI, 2021), New Zealand (GRUBER et al., 2021), Pakistan (RABBANI; ZEESHAN, 2020), Iran (GUALTIERI, 2021; KHATIBI; KRAUTER, 2021), Australia (GUALTIERI, 2021; KHATIBI; KRAUTER, 2021) and Ethiopia (NEFABAS et al., 2021). Thus, there is an opportunity to expand these studies to other locations and explore smaller scales, both temporal and spatial.

The present study is focused on Brazil, since we only identified preliminary works (GRUBER et al., 2019, 2021; GRUBER; SCHMIDT, 2019) evaluating the application of reanalysis data. The focus and conclusions of these articles mainly involve validation of the applicability of reanalysis data in places with a low spatial resolution (countries, regions, states). In this respect, Gruber & Schmidt (2019) evaluated the use of ERA5 data for Austria and Brazil, while Gruber et al. (2021) covered Brazil, USA, New Zealand, and South Africa. Gruber et al. (2019) was the only one using MERRA-2 data, exclusively evaluating Brazil, with the analyses and conclusions pertaining to the level of regions and states. In particular Gruber et al. (2021), began by evaluating data from MERRA-2 and ERA5, but when finding better performance of ERA5, the rest of the article relied only on data from that source. However, a considerable portion of the results indicated better performance of MERRA-2 data for Brazil (between 25% and 50%) as well as better performance of MERRA-2 in New Zealand in most cases. Brazil has a large landmass and great regional climate differences, which hampers reaching consensus conclusions about the country. Against this backdrop, the general aim of this article is to provide continuity to these previous studies and further exploring the MERRA-2 data.

The main objective of this article is to verify if the MERRA-2 wind speed time series can satisfactorily represent the historical wind speed time series measured at points in Brazilian territory. This better adjustment of the time series is investigated using interpolation, extrapolation, and bias correction alternatives. In addition, we analyze the consequences of these results in different temporal and spatial scales.

The present study contributes to the literature by giving continuity to the process of validating the MERRA-2 data at greater spatial resolution (turbines and wind farms) in Brazil. In the Brazilian context, the article innovates by determining the quality of the reanalysis time series in Brazilian territory, by comparison with the wind speed time series. Most studies have measured the quality of the MERRA-2 time series by comparing simulated generation and observed generation. The problem is that during this process of transforming wind speed into wind power, many sources of uncertainty are included due to the conversion method and its parameters. These facts impair the quality of the generation time series built and the results of evaluating the MERRA-2 data. Another innovation of this study is a more detailed analysis of the MERRA-2 wind speed time series' variability. The historical time series of different turbines in the same wind farm can present different behaviors due to the stochastic nature of the random variable, wind speed. In this study, these time series are represented by the same wind speed time series obtained from the reanalysis data, thus making it possible to measure the deviation between the historical and MERRA-2 datasets and the natural variation of the process. This enables ascertaining how much the metrics for evaluation of MERRA-2 data can vary due to the natural randomness of wind speed data.

We also make several other contributions. First, we present a new bias correction approach for time frames and analyze its performance. Second, we examine the impacts of the representativeness of the MERRA-2 reanalysis time series when they are temporally and spatially aggregated, comparing the findings with those using other data, a novel analysis. Finally, we propose an innovative methodological approach for aggregation.

This first contribution is divided into five sections. The first part describes our motivations, explains the relevance, and summarizes the relevant literature. The second section details and supports the methodology for adapting the reanalysis data and the respective treatment possibilities. The third section reports the results obtained by evaluating the time series created through the metrics proposed in the methodology, and the fourth section discusses them. Finally, the fifth section presents our final considerations and proposals for future work.

2.3

Data and methodology

As explained in the Introduction, the main objective of this study is to verify whether it is appropriate to use the wind speed time series from the MERRA-2 database for Brazilian territory. The method applied to achieve this goal is shown in Figure 2.1 and has seven main steps. These steps are detailed in subsections 2.3.1 to 2.3.7. The idea underpinning this method is for it to be sufficiently general to generate wind speed time series applicable anywhere in Brazilian territory, with the use of only data in the public domain, while guaranteeing its replicability. Our only use of private data was for validation of the method.



Figure 2.1 Overview of the methods used.

2.3.1

Obtaining reanalysis data

This study uses MERRA-2, which is a global atmospheric reanalysis dataset produced by NASA, more specifically by the Global Modeling and Assimilation Office (GMAO), with historical data series produced with the Goddard Earth Observing System Model (GEOS), which is fed with information collected by NASA satellites. The GEOS model is composed of physical models that are flexibly related to represent several aspects related to earth sciences (GMAO, 2021).

MERRA-2 provides free historical series of several climatic and atmospheric conditions, which can be downloaded directly from the NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC) (GES DISC, 2021). However, the data are available in the NetCDF (Network Common Data Form) format, which requires specific software for manipulation. Alternative scripting programming languages, such as R and Python, available from several works on reanalysis data, allowed us to download and manipulate the data from MERRA-2. The R scripts used by us are the same ones used in Gruber et al. (2019) and are available at GitHub (Supplementary Material).

The MERRA-2 dataset used was "MERRA-2 tavg1_2d_slv_Nx", which consists of single-level diagnostics, assimilation, two-dimensional data, and hourly time-averaged data (GRUBER et al., 2019, 2021; OLAUSON, 2018). These data have been available since 1980 and are updated monthly with hourly resolution. The spatial resolution is approximately 50 km between the points since the data are available in a grid varying by 0.625° in longitude and 0.5° in latitude. The variables selected for this study were the latitudes and longitudes of the MERRA-2 grid points over Brazilian territory with their respective wind speed time series in u and v directions at two different heights (10 m and 50 m), as well as the displacement height (DISPH) (GMAO, 2021).

All elements present on the earth's surface, whether natural or not, can cause some perturbation to the wind profile of these locations, and these terrain peculiarities are addressed in the physical model when the wind speed series at different heights are projected. To represent the impact of terrain conditions on the log wind profile, the MERRA-2 base provides for each point of the grid the DISPH variable, which is an additive correction factor at heights of 2 and 10 m to be applied in the extrapolation process of the MERRA-2 wind speed series (GMAO, 2021).

2.3.2

Identifying wind turbines

Brazil is a mainly tropical country with large territorial extension and a long coastline. It has a favorable climate, including steady trade winds (constant speed regimes, ideal for power generation), mainly in the Northeast region. All these characteristics serve as an incentive for exploring onshore and offshore wind generation (PERINI DE SOUZA et al., 2022b, 2022a). Currently, Brazil has 812 wind farms in operation, of which 87.3% are located in the Northeast region. Bahia state (BA), with 227 wind farms, has the largest number, followed by Rio Grande do Norte (RN) with 216 wind farms. The installed wind generation capacity is 21.8 GW, of which 90.2% is located in the Northeast region, 9.7% in the South region and the rest in the Southeast region. Figure 2.2 presents Brazil's geographic dispersion of installed capacity by state through the blue color scale. There is a greater concentration in RN (30.5%) followed closely by the BA (27.6%) (SIGA ANEEL, 2023).



Figure 2.2 Brazilian territory and installed wind power capacity.

To validate the wind speed time series based on reanalysis data, we compared them with wind speed data measured at turbines. Therefore, we needed to find wind farm operators willing to share such data, namely (i) the wind speed time series measured at turbines and their respective time horizon, (ii) the wind farms' geographic locations, and (iii) height of the wind turbine rotor from the ground.

Through a confidentiality agreement, an operator of wind farms provided 24 wind speed time series obtained from two wind farms in Bahia, 12 from each one. These referred to 2017, with measurements recorded every 10 min at a height of 99.5 m, corresponding to the height of the turbines (2017/01/01 00:00:00 to 2017/12/31 23:50:00).

2.3.3

Data treatment

We used wind speed time series from different bases, with the possibility of measurement errors and missing and/or inconsistent data. Therefore, it was necessary to carry out a step involving preprocessing and validation of these time series.

All the datasets were evaluated for possible missing data and speeds below zero or above 25 m/s, which were disregarded. The MERRA-2 dataset has already been validated, so no treatment was necessary. The database used in the validation did not present any velocity values below zero but did contain 0.53% of values greater than 25 m/s and missing values of 1.98%. This base is formed by 24 time series with daily wind speed at 10-min intervals for a period of one year, in which the number of velocity readings faster than 25 m/s varied between 0.41% and 0.54% and the percentages of missing data ranged from 0.31% to 4.45%, except for one turbine, for which the figure was 12.53%. To build the validation time series with lower resolutions we calculated the average wind speeds among the readings available over the horizon examined, disregarding the invalid data. For example, each hourly time series was formed by taking the average of the six readings in the hour in question, or if there were invalid data, the average was calculated with the number of valid readings, and if all were absent, the situation was classified as missing data.

Figure 2.3, Figure 2.4, Figure 2.5 present some of these wind speed time series measured at the turbines, where some variations can be noted even in the case of series belonging to the same region. Figure 2.3 presents three density curves plotted with data from different turbines: the first curve has three modes, the second has one mode, and the third curve has two modes. Besides this, the principal mode occurs with distinct values among the curves.





26



Figure 2.4, Figure 2.5 present the boxplots of the wind speed data measured at the turbines on a monthly and hourly basis, respectively. These graphs allow visualizing the dispersion of the data in different time frames and the seasonal movement according to the medians and quantiles. Comparison of the data from different turbines does not show large differences in the dispersion of the data for a single time frame, while it does show the repeated seasonal movement of the medians among the months. However, regarding the medians among the hours, some differences can be noted in the behavior, mainly among the readings at the hours 8, 9, 10, 21, 22, and 23. Based on this statistical description of the data from the turbines, we were able to map the different profiles present in this database and verify how this could impact the results.

For bias correction, we used the GWA and the base of Brazil's National Institute of Meteorology (INMET). The GWA supplies mean wind speeds. All the readings were available and had values between 0 and 25 m/s. INMET makes available hourly wind speed time series for its anemometer stations, and no readings were below zero or above 25 m/s. However, there were some cases of missing data, which we disregarded in calculating the correction factors of the time series.

2.3.4

Interpolation

In the interpolation step, the MERRA-2 grid points were associated with the turbines' geographic locations to build a history of the variables relevant to the study. The interpolation technique used was the nearest neighbor, as recommended by Gruber et al. (2019). It consists of associating the location of the wind turbine with the nearest point on the MERRA-2 grid, as demonstrated in Figure 2.6.



2.3.5

Extrapolation

Wind speed is an atmospheric characteristic that varies according to altitude. The wind speed tends to increase with altitude for the same territory and under the same conditions (NEFABAS et al., 2021). Consequently, depending on the height of a wind turbines rotor, it will be subject to a different speed, which causes different energy generation levels. Therefore, it is crucial to identify the hub height of the studied turbines for the correct estimation of the wind speed (STAFFELL; PFENNINGER, 2016).

The extrapolation adjusts the wind speeds provided in the MERRA-2 dataset at heights of 10 and 50 m to the height of the turbine rotor. The increase in speed according to height is exponential rather than linear. Thus, this adjustment can be performed by the Hellman power law, which is expressed in Equation 1. The exponent α is dimensionless and represents the characteristics relevant to the environment (GRUBER et al., 2021).

$$V(h_1) = V(h_2) \cdot \left(\frac{h_1}{h_2}\right)^{\alpha}$$
 (1)

Where:

 $V(h_i)$ is the MERRA-2 wind speed at height h_i , h_i is the height of MERRA-2 wind speed, and. α is the local power exponente.

Thus, the steps of the extrapolation stage are:

1 – Calculate effective wind speeds at 10 and 50 m from their respective components in u and v direction (REN et al., 2019).

2 – Calculate the exponents α through Equation 1 using the height of 50 m and its corresponding effective speed and the height of 10 m plus the DISPH and the associated effective speed. In the MERRA-2 reanalysis data, the wind speeds that correspond to the heights of 2 and 10 m refer to these heights plus the DISPH. This is necessary to correct the log wind profile, which changes at lower heights due to the peculiarities of the relief of each region (GMAO, 2021).

3-Calculate extrapolated wind speeds for the turbine rotor height through Equation 1, using the calculated α exponents and the height of 50 m and their respective effective speeds.

2.3.6

Bias correction

This study uses two datasets as an alternative to correct bias. The approach adopted is correction based on the time series average (GRUBER et al., 2019), described below.

2.3.6.1

The Global Wind Atlas

The GWA is a free web-based application. It is designed to help policymakers, planners, and investors to identify areas suitable for wind power generation virtually anywhere globally. It is the result of a partnership between the Department of Wind Energy of the Technical University of Denmark (DTU Wind Energy) and the World Bank Group (the World Bank and the International Finance Corporation - IFC) (GWA, 2021).

The GWA provides the average wind speeds at five heights (10, 50, 100, 150, and 200 m) for each point, located 250 m apart. Our choice of this information for bias correction was due to its wide use in the literature (BOSCH; STAFFELL; HAWKES, 2018; GRUBER et al., 2019, 2021; NEFABAS et al., 2021; OLAUSON, 2018; RYBERG et al., 2019) and its high spatial resolution.

The GWA data are available in a grid (by latitude and longitude), as are the data in MERRA-2. Therefore, the nearest neighbor technique was used again to identify the GWA grid point closest to the wind turbine. Then, the MERRA-2 grid point with the closest distance to the GWA grid's selected point was chosen to calculate the bias correction factor. Figure 2.7 outlines this step in the process.



The correction factor is the ratio between the averages. The numerator is the average of the wind speed of the GWA that would represent the turbine, and the denominator is the calculated average of the effective wind speeds from MERRA-2 to the MERRA-2 point closest to the GWA point. Since the average supplied by the GWA utilizes wind speed data between 2008 and 2017, we used this time horizon to select the data from MERRA-2. GWA and MERRA-2 data must be at

the same height, in this case, both at 50 m. Equation 2 represents the calculation of the bias correction factor by the GWA, and Equation 3 denotes the correction of the extrapolated MERRA-2 wind speed time series representing the wind turbine.

$$FC_{GWA} = \frac{V_{mean(50m)_{GWA \to Farm}}}{\sum \frac{V(50m)_{MERRA2 \to GWA}}{n}}$$
(2)

$$V_{ext_GWA} = V_{ext} \cdot FC_{GWA} \tag{3}$$

Where:

 FC_{GWA} is the bias correction factor according to the GWA,

 $V_{mean(50m)_{GWA \to Farm}}$ is the GWA mean wind speed at 50 m height at one location, $V(50m)_{MERRA2 \to GWA}$ is the MERRA-2 wind speed at 50 m height at one location, *n* is the total number of times periods,

 V_{ext_GWA} is the MERRA-2 wind speed extrapolated and corrected by the GWA bias correction fator, and.

 V_{ext} is the MERRA-2 wind speed extrapolated to hub height at one location.

2.3.6.2

INMET

The INMET is a Brazilian federal agency whose purpose is to provide meteorological information through monitoring, analysis, and forecasting of the weather and climate. It makes meteorological information available free of charge at its website. INMET has 478 meteorological stations spread across Brazil, which provide hourly wind speed data recorded at 10 m from the ground (INMET, 2021). The INMET wind speed data were selected for bias correction because they are measured data, which can better portray the peculiarities of Brazilian territory, in addition to being freely available, as also described by Gruber et al. (2019).

As indicated in Figure 2.8, once again we used the nearest neighbor method to choose the INMET station and the MERRA-2 data to calculate the bias correction factors according to the INMET dataset (FCINMET). The time horizon adopted for both time series was 2008–2017, as was the case of FCGWA.



For the FCINMET calculation, we analyzed whether there were any missing data in any of the time series. If there were, the respective time frame was

(2)

eliminated from both time series to preserve the equal sizes of the wind speed time series. Next, FCINMET was calculated by using the means of the treated time series, as indicated in Equation 4.

$$FC_{INMET} = \frac{\sum \frac{V(10m)_{INMET \to Farm}}{n}}{\sum \frac{V(10m)_{MERRA2 \to INMET}}{n}}$$
(4)

Where:

FC_{INMET} is the bias correction factor determined by INMET,

 $V(10m)_{INMET \rightarrow Farm}$ is the INMET wind speed at 10 m height at one measurement station,

 $V(10m)_{MERRA2 \rightarrow GWA}$ is the MERRA-2 wind speed at 10 m height at one location, and.

n is the total number of times periods.

As reported by Gruber et al. (2019), if an INMET station is very far from the wind turbine evaluated, the bias correction for its wind speed data may not represent the real situation of the site in question. Therefore, we adopted the same constraint applied by Gruber et al. (2019), only correcting the bias of the INMET data when the distance between the station and the site evaluated was less than 40 km.

To explore all the advantages of the INMET wind speed historical time series, we used four types of bias correction factors for INMET:

1 – General (FCINMET-G): A single factor is calculated, and it is used to multiply all the extrapolated wind speed time series values of MERRA-2 representing the turbine in question.

2 - Monthly (FCINMET-M):12 factors are calculated (one for each month). In the calculation of each factor, only data from its respective month is used. The bias correction of the time series extrapolated from the MERRA-2 values is carried out by ranges, multiplying the factor by the time series values belonging to the same month.

3 – Hourly (FCINMET-H): 24 factors, one for each hour, are applied in the part of the time series extrapolated from its respective hour.

4 – Monthly and Hourly (FCINMET-M/H): 288 factors, one for each hour of each month, with each applied in the part of the extrapolated time series of its respective month and hour.

2.3.7

Evaluation

Several accuracy measures were used to compare the corrected wind speed time series obtained from MERRA-2 with the observed time series from each wind turbine in this study: Pearson correlation coefficients, RMSE, MBE, MAE, and the difference between time series variances (Variances diff.). They were selected based on the main literature sources (GRUBER et al., 2019; NEFABAS et al., 2021; OLAUSON, 2018; RYBERG et al., 2019; STAFFELL; PFENNINGER, 2016).

2.4

Results

This section presents the main results reached, divided into three parts. The first subsection introduces results of the hourly wind speed time series of wind turbine analysis generated by the different methods described in the Methodology section. The second and third subsections complement the study under different temporal and spatial approaches, respectively, but with the same steps used in the method of the first section. We used the R software, version 3.5.1, to tally the data from the various bases, treat them and generate the results (R CORE TEAM, 2018).

We first identified the geographic coordinates of Brazilian territory for which MERRA-2 provides data and extracted the wind speed time series for these sites. In possession of the coordinates of the turbines of the wind farms, we began the interpolation step using the nearest neighbor method, which indicated the same point of MERRA-2 to represent all wind turbines. This fact was expected and is explained by the large spacing between the points of the MERRA-2 grid. This indicates that a single MERRA-2 time series represents innumerable real historical time series. The locations of the turbines are not identified here because of the confidentiality agreement with the wind farms' operator.

In the next step, extrapolation by the Hellman power law, the MERRA-2 time series representing the turbines was adjusted to 99.5 m, which corresponds to the height of the rotor of each turbine.

The GWA grid has high spatial resolution, so when associating the location of the turbines with the grid points, different points were chosen between them, resulting in different time series in the GWA bias correction step. Meanwhile, when correcting the bias using INMET data, all turbines were associated with the same meteorological station, which had an average distance of 35.54 km from them. Therefore, the wind speed time series of the selected INMET station was treated and used jointly with the MERRA-2 time series to construct the general, monthly, hourly, and monthly/hourly correction factors, which were later applied to the bias treatment.

The time series of the MERRA-2 data and the INMET station employed for bias correction covered the period from 2008 to 2017. However, the time series from MERRA-2 after extrapolation and bias correction, which are proposed as historic for the turbines, are only projected for 2017 since the historic dataset measured at the turbines for validation only presents information for that year.

For validation of the method, we treated all the time series measured at the turbines of the wind farms. For each time series measured, six treated time series were generated, as presented in Table 2.1. The comparisons between the time series were made using the five metrics (Pearson correlation, RMSE, MBE, MAE, and Variances diff.) and the results obtained are shown in Figure 2.9, Figure 2.10, Figure 2.11, Figure 2.12, Figure 2.13.

Table 2.1 Treatment of the MERRA-2 time series.

Identification	Treatments				
EXT	Interpolation and Extrapolation				
GWA	Interpolation, Extrapolation and Bias Correction by GWA				
INMET-G	Interpolation, Extrapolation and Bias Correction by INMET – General				
INMET-M	Interpolation, Extrapolation and Bias Correction by INMET – Monthly				
INMET-H	Interpolation, Extrapolation and Bias Correction by INMET – Hourly				
INMET-M/H	Interpolation, Extrapolation and Bias Correction by INMET – Monthly/Hourly				



Figure 2.9 Correlation boxplots of the measured hourly history of the turbines with the treated reanalysis time series.



Figure 2.10 RMSE boxplots of the measured hourly history of the turbines with the treated reanalysis time series.



Figure 2.11 MBE boxplots of the measured hourly history of the turbines with the treated reanalysis time series.



Figure 2.12 MAE boxplots of the measured hourly history of the turbines with the treated reanalysis time series.



Figure 2.13 Boxplots of Variances diff. of the measured hourly history of the turbines with the treated reanalysis time series.

The Pearson correlation coefficient was the only metric used here that is limited between -1 and 1, with 1 indicating strong direct correlation between the data, 0 no correlation, and -1 high inverse correlation, so the closer to 1, the better

was the treatment performance from the point of view of the correlation metric. The other indices measure the difference between the time series, so the closer to zero the metric was, the better the performance of the treated time series. As the performance of each treatment strategy was measured 24 times per metric, a boxplot of these results was generated to demonstrate their dispersion for each treatment method. The median was used to compare the performance between the strategies of each metric.

2.4.1

Hourly results

Figure 2.9 compares the treatments under the correlation metrics and shows better performance of the INMET-H approach, followed by EXT, GWA, and INMET-G, all tied. The dispersion of the results of the correlations for each approach remained similar, with amplitude varying from 0.074 (EXT, GWA, and INMET-G) to 0.091 (INMET-M/H).

Figure 2.10, Figure 2.12 show the RMSE and MAE, respectively, with the best performance generated by INMET-H, followed by EXT. Figure 2.11 shows the MBE with a reversal of positions so that the best performer was EXT followed by INMET-H. Figures. Figure 2.10, Figure 2.11, Figure 2.12, Figure 2.13 demonstrate the worst performance of the GWA. In Figure 2.10, Figure 2.11, Figure 2.12, the GWA has considerably more dispersion of results than in the other methods, mainly when comparing its first and third quartiles with others. Finally, Figure 2.13 demonstrates Variances diff., where INMET-H achieved the best performance, and the second best was INMET-M/H. Another relevant point is that when analyzing the MBE results, the GWA treatment was the only one that presented negative indices, demonstrating it overestimated the wind speeds.

Table 2.2 summarizes the results found for the treatment strategies according to each metric. The best approaches are in blue and the worst in red. The INMET-H treatment was best ranked more often, and the GWA had the worst overall ranking.

Strategies			Metrics		
Strategies	Correlation	RMSE	MBE	MAE	Var diff.
EXT	0.6718	2.3660	0.5584	1.8626	1.8679
GWA	0.6718	3.3024	-1.9596	2.7514	-2.7232
INMET-G	0.6718	2.3907	0.6965	1.8691	2.1217
INMET-M	0.6565	2.4316	0.7127	1.9140	2.3353
INMET-H	0.6903	2.3399	0.6458	1.8316	1.5068
INMET-M/H	0.6671	2.4223	0.6752	1.9007	1.6540

Table 2.2 Medians of treatment strategies per metric.

Table 2.3 shows the range of variation of the results between the first and third quartiles per metric. The dispersion of results was maintained for all treatments except for GWA. For it, the dispersion increased with all metrics except for Variances diff., in which it decreased. For most treatments and all metrics, there was small dispersion between the first and third quartiles, showing that the

comparison between the medians is an adequate approach to decide which treatment to choose for the MERRA-2 data.

Strataging			Metrics		
Strategies	Correlation	RMSE	MBE	MAE	Var diff.
EXT	0.0144	0.1999	0.2819	0.1849	1.6710
GWA	0.0144	0.7274	0.9229	0.6458	1.4909
INMET-G	0.0144	0.1782	0.2820	0.1658	1.6710
INMET-M	0.0148	0.1888	0.2822	0.1673	1.6710
INMET-H	0.0146	0.2064	0.2815	0.1825	1.6710
INMET-M/H	0.0155	0.1806	0.2818	0.1643	1.6710

Table 2.3 The range between the first and third quartiles of results of treatment strategies per metric.

Table 2.4 presents the amplitude of the treatment results according to each metric, showing that depending on the metric and the historical time series of measured wind speed used, the performance of the treatment strategy changed considerably. The impacts were lower for the correlation metrics RMSE and the MAE, where the variation reached maximums of 14% (INMET-M/H), 30% (GWA) and 36% (GWA), respectively, concerning their medians. The MBE and Variances diff. metrics were most impacted by choice of the measured wind speed time series: the first varied up to 397% (EXT) and the second 310% (INMET-H). These results demonstrate the importance of choosing the appropriate evaluation metric and the extent of variability within the historical time series measured for speed of the same wind farm or between neighboring wind farms.

Strategies	Metrics				
	Correlation	RMSE	MBE	MAE	Var diff.
EXT	0.0739	0.4747	2.2142	0.4221	4.6670
GWA	0.0739	0.9960	2.2107	0.9971	4.7347
INMET-G	0.0739	0.5231	2.2145	0.4755	4.6670
INMET-M	0.0772	0.5242	2.2136	0.4731	4.6670
INMET-H	0.0888	0.4714	2.2155	0.4144	4.6670
INMET-M/H	0.0911	0.4706	2.2148	0.4353	4.6670

Table 2.4 Range of results of treatment strategies by metric.

Figure 2.14 presents the relationship between the measured data of the wind farm turbines and their respective treated MERRA-2 data, through a combination of scatter plots and histograms. These visualizations allow better understanding of the scale of the deviations between the measurements and their estimates, and where the greatest concentration of data is. Each graph that composes Figure 2.14 refers to one of the treatment strategies adopted in this study. In order to analyze the quality of the estimates made by the MERRA-2 treatment strategies through these graphs, it is necessary to verify which strategies present the best adjustment of the points to the diagonal of the graph, mainly in the regions that have greater concentrations of data, for which the data are complemented by the histograms.


The worst performance was from the GWA strategy, since it overestimated wind speeds. This can be noticed from the high concentration of data above the diagonal and comparison of the histograms, especially in the final region, where there is a higher density of data in the histogram of the estimates than in the histogram of the measured data. The data coming from MERRA-2 for the studied region are underestimated, as can be seen in the EXT strategy graph, where there is a high concentration of data below the diagonal. Bias correction strategies seek to correct this and the INMET-H strategy presented the best performance. This shows the importance of finding an effective bias correction method, to help control the dispersion and order of magnitude of the data.

The adequacy of the estimates to the measured data can also be analyzed by comparing the shape of the data distribution provided by the histogram. However, the distributions of wind speed data between the turbines already have different shapes, as highlighted in subsection 2.3.3, and in Figure 2.4 the distribution of measured data presented is the joint speed data recorded in all turbines. Comparison of the histograms indicates better adequacy of the histogram pertaining to the INMET-H strategies, followed by the EXT strategy, while the worst adequacy is generated by the GWA strategy. In the Supplementary Material, we provide more graphs to facilitate the comparison of these histograms and the dispersion of these distributions (boxplots).

2.4.2

Effects of the use of different time scales

The specialized literature on validation of reanalysis data contains studies with various time scales. This occurs because depending on the reanalysis dataset or validation method utilized, the data employed are available for different time scales. Therefore, we investigated whether these different time scales could influence the results of the evaluation metrics, and thus the judgment of the quality of the representativeness of the historic series measured by MERRA-2 data.

For this purpose, we temporally aggregated all the wind speed time series used as inputs in our method (the MERRA-2 series, the measured series at the turbines and the INMET series). Therefore, we carried out all the steps of the method in different time scales, including the evaluation through comparison of the wind speeds.

The temporal aggregation of the time series was carried out by calculating the average over each time frame. For example, to convert hourly time series to monthly, we calculated the average wind speed within the hours contained in the month.

All the methods developed and presented above for hourly time series were replicated using daily and monthly data. Table 2.5 shows the medians obtained from the boxplot graphs of each treatment per metric in their respective temporal study. Graphs like those shown in Figure 2.9, Figure 2.10, Figure 2.11, Figure 2.12, Figure 2.13, Figure 2.14 were generated for the daily and monthly studies and their medians are shown in Table 2.5. These graphs can be accessed at Supplementary Material.

Metrics Correlation RMSE MBE MAE Variances diff.	Time	Treatment Strategies										
Metrics	Scale	EXT	GWA	INMET-G	INMET-M	INMET-H	INMET-M/H					
	Hourly	0.6718	0.6718	0.6718	0.6565	0.6903	0.6671					
Correlation	Daily	0.9108	0.9108	0.9108	0.8947	-	-					
Metrics Ti So Correlation Di Mo RMSE Di Mo MBE Di Mo MAE Di Mo Variances diff.	Monthly	0.9765	0.9765	0.9765	0.9185	-	-					
	Hourly	2.3660	3.3024	2.3907	2.4316	2.3399	2.4223					
RMSE	Daily	1.1846	2.1474	1.2257	1.3078	-	-					
	Monthly	0.7418	1.8497	0.8734	0.9959	-	-					
	Hourly	0.5584	-1.9596	0.6965	0.7127	0.6458	0.6752					
MBE	Daily	0.6471	-1.8327	0.7621	0.7861	-	-					
CorrelationHourly Daily0. DailyCorrelationDaily0. MonthlyMonthly0. Hourly2. PailyRMSEDaily1. MonthlyMBEDaily0. MonthlyMAEDaily0. MonthlyMAEDaily0. MonthlyVariances diff.Hourly1. MonthlyMonthly0. Monthly0. Monthly	0.6162	-1.8060	0.7706	0.7855	-	-						
RMSE MBE MAE	Hourly	1.8626	2.7514	1.8691	1.9140	1.8316	1.9007					
MAE	Daily	0.9195	1.9248	0.9827	1.0403	-	-					
	Monthly	0.6665	1.8060	0.7992	0.8219	-	-					
X 7 •	Hourly	1.8679	-2.7232	2.1217	2.3353	1.5068	1.6540					
v ariances diff.	Daily	1.6478	-0.3537	1.7621	2.0314	-	-					
	Monthly	0.3675	-0.5095	0.4294	0.7196	-	-					

Table 2.5 Medians of treatment strategy results.

It can be seen from Table 2.5 that the temporal aggregation of the time series improved the metrics' indices. The MERRA-2 time series best represented the measured time series. All correlation coefficients increased with temporal aggregation for all treatments, and all RMSE and MAE indices decreased. In Variances diff., this trend of decreasing indices with increasing temporal scales was repeated with only one exception in the GWA monthly treatment. The MBE also fell, although to a lesser extent.

2.4.3

Impact of different spatial scales

Another point that can impact the results is the spatial scale used in the analyses. Therefore, we examined whether a MERRA-2 wind speed time series representing larger areas could perform better in the metrics. In Gruber et al. (2019), where performance was measured by comparing the generation time series, the spatial aggregation occurred through the sum of the wind generation of all turbines and wind farms in the evaluated region, before comparing the measured and simulated data (GRUBER et al., 2019). Since we used wind speed data, the spatial aggregation took place by calculating the average wind speed time series using all the time series contained in the analyzed region.

For this investigation, we compared the MERRA-2 time series corrected with (i) the 12 time series measured in the turbines of one of the wind farms (Wind Farm 1 - WF 1), with (ii) the average wind speed time series calculated from the 12 time series of Wind Farm 1's turbines, and with (iii) the average wind speed time series calculated with the time series of all 24 turbines. In (i), the performance of the 12 turbines is given by the median of their results in the metrics. Table 2.6 (first

analysis) compares the WF1 data with the Merra-2 data corrected under different metrics. This same analysis was performed with data from the other wind farm (Wind Farm 2 - WF 2) and is shown in Table 2.7.

Metrics Correlation RMSE MBE MAE Variances	Spatial	Treatment Strategies										
Metrics	Scale	EXT	GWA	INMET-G	INMET-M	INMET-H	INMET-M/H					
MetricsSpatial ScaleEXTCorrelationTurbines F10.6779WF 10.68470.6847Aggregate0.6847Aggregate2.4198WF 12.3590Aggregate2.2421MBETurbines F10.5728MBETurbines F10.5985Aggregate0.3950MAETurbines F11.9296MAEWF 11.8613Aggregate1.7569VariancesWF 12.6278WF 12.2473	0.6779	0.6779	0.6604	0.6903	0.6671							
Correlation	WF 1	0.6870	0.6870	0.6870	0.6700	0.7010	0.6764					
	Aggregate	0.6847	0.6847	0.6847	ET-G INMET-M INMET-H INMET-M/H 779 0.6604 0.6903 0.6671 370 0.6700 0.7010 0.6764 847 0.6678 0.7008 0.6772 418 2.4963 2.4150 2.4966 872 2.4386 2.3507 2.4371 567 2.3040 2.2245 2.3036 113 0.7333 0.6582 0.6909 366 0.7519 0.6859 0.7146 330 0.5483 0.4828 0.5113 447 1.9947 1.9029 1.9756 835 1.9317 1.8366 1.9178 670 1.8164 1.7310 1.8028 816 3,0952 2.2667 2.4139 011 2.7147 1.8862 2.0333 334 1.7470 0.9185 1.0656							
	Turbines F1	2.4198	3.4250	2.4418	2.4963	2.4150	2.4966					
RMSE	WF 1	2.3590	3.3669	2.3872 2.4386		2.3507	2.4371					
	Aggregate	2.2421	2.9101	2.2567	2.3040	2.2245	2.3036					
	Turbines F1	0.5728	-2.0879	0.7113	0.7333	0.6582	0.6909					
MBE	WF 1	0.5985	-2.0613	0.7366	0.7519	0.6859	0.7146					
	Aggregate	0.3950	-1.5056	AINMET-GINMET-MINMET-HINMET-M/H790.67790.66040.69030.6671700.68700.67000.70100.6764470.68470.66780.70080.6772502.44182.49632.41502.4966592.38722.43862.35072.4371012.25672.30402.22452.3036790.71130.73330.65820.6909130.73660.75190.68590.7146560.53300.54830.48280.5113561.94471.99471.90291.9756021.88351.93171.83661.9178341.76701.81641.73101.8028972.88163.09522.26672.4139032.50112.71471.88622.0333891.53341.74700.91851.0656								
	Turbines F1 0.6779 0.6779 0.6779 ion WF 1 0.6870 0.6870 0.6870 Aggregate 0.6847 0.6847 0.6847 0.6847 Turbines F1 2.4198 3.4250 2.441 WF 1 2.3590 3.3669 2.387 Aggregate 2.2421 2.9101 2.256 Turbines F1 0.5728 -2.0879 0.711 WF 1 0.5985 -2.0613 0.736 Aggregate 0.3950 -1.5056 0.533 Turbines F1 1.9296 2.8456 1.944 WF 1 1.8613 2.8102 1.883 Aggregate 1.7569 2.4034 1.76' WF 1 2.6278 -3.1697 2.883 WF 1 2.2473 -3.5503 2.507 Aggregate 1.2796 -2.6889 1.533	1.9447	1.9947	1.9029	1.9756							
MAE	WF 1	1.8613	2.8102	1.8835	1.9317	1.8366	1.9178					
	Aggregate	1.7569	2.4034	1.7670	1.8164	1.7310	1.8028					
T 7 •	Turbines F1	2.6278	-3.1697	2.8816	3,0952	2.2667	2.4139					
Variances diff.	WF 1	2.2473	-3.5503	2.5011	2.7147	1.8862	2.0333					
·····	Aggregate	1.2796	-2.6889	1.5334	1.7470	0.9185	1.0656					

Table 2.6 Metric results for different spatial aggregations (Analysis 1).

Table 2.7 Metric results for different spatial aggregations (Analysis 2)

M - 4	Spatial	•		Treatmen	t Strategies		
Metrics	Scale	EXT	GWA	INMET-G	INMET-M	INMET-H	INMET-M/H
	Turbines F2	0.6694	0.6694	0.6694	0.6542	0.6907	0.6682
Correlation	WF 2	0.6739	0.6739	0.6739	0.6576	0.6921	0.6700
	Aggregate	0.6847	0.6847	0.6847	0.6678	0.7008	0.6772
	Turbines F2	2.2998	2.7017	2.2904	2.3330	2.2771	2.3403
RMSE	WF 2	2.1975	2.7552	2.1970	2.2382	2.1697	2.2386
	Aggregate	2.2421	2.9101	2.2567	2.3040	2.2245	2.3036
	Turbines F2	0.3231	-1,1745	0.4615	0.4774	0.4108	0.4399
MBE	WF 2	0.1862	-1.3079	0.3242	0.3395	0.2739	0.3025
ScaleScaleScaleTurbines F2WF 2AggregateTurbines F2RMSEWF 2AggregateTurbines F2MBEWF 2AggregateTurbines F2MAEWF 2AggregateTurbines F2MAEVariancesdiff.Aggregate	Aggregate	0.3950	-1.5056	0.5330	0.5483	0.4828	0.5113
	Turbines F2	1.8077	2.2083	1.8100	1.8470	1.7820	1.8449
MAE	WF 2	1.7159	2.2674	1.7117	1.7567	1.6855	1.7433
	Aggregate	1.7569	2.4034	1.7670	1.8164	1.7310	1.8028
X 7 •	Turbines F2	1.1485	-1.8971	1.4023	1.6159	0.7874	0.9345
Variances diff.	WF 2	0.6277	-2.4178	0.8815	1.0951	0.2666	0.4138
	Aggregate	1.2796	-2.6889	1.5334	1.7470	0.9185	1.0656

From the correlation metric in Table 2.7 and the RMSE, MAE, and Variances diff. metrics in Table 2.6, it can be stated that spatial aggregation also contributed to improving the representativeness of the MERRA-2 time series concerning the measured time series, since there was improvement of the indices with increasing spatial scale in all of them. In the case of the Pearson correlation coefficient (Table 2.7), it increased with the increase of the represented area, and in the RMSE, MAE, and Variances diff. it decreased as the area covered increased. However, in the other cases, as shown in both Table 2.6, Table 2.7, the inference of improvement with spatial aggregation was not substantial.

2.5

Discussion

The studies found in the literature have used time series in different scales, such as hourly (CANNON et al., 2015; GONZÁLEZ-APARICIO et al., 2017; NEFABAS et al., 2021), daily (GRUBER et al., 2019, 2021; GRUBER; SCHMIDT, 2019), and monthly (CRADDEN et al., 2017; STAFFELL; PFENNINGER, 2016). These variations are due to the measured data used for comparison in the metrics, which are only available for a specific temporal scale. This study generated results in hourly, daily, and monthly scales and indicated that the choice of the temporal scale impacts the results found in the metrics. By increasing the temporal aggregation of the compared time series, the indices improved for most of the metrics.

However, regardless of the time scale used, the results demonstrated that the corrected MERRA-2 time series can satisfactorily represent the wind speed time series measured at the turbines of wind farms. This was mainly the case when comparing the results obtained here with those of Gruber et al. (2019, 2021), Gruber & Schmidt (2019), Nefabas et al. (2021) and Staffell & Pfenninger (2016), although those results applied to generation comparison while we used comparison of wind speeds. Some studies have evaluated the representativeness of the MERRA-2 data by comparison of the wind speeds (BRUNE; KELLER; WAHL, 2021; JIANG et al.. 2021; JOURDIER, 2020; KHATIBI; KRAUTER, 2021; MOLINA; GUTIÉRREZ; SÁNCHEZ, 2021; RABBANI; ZEESHAN, 2020), but none of these works are applied in Brazilian territory. Gruber et al. (2019), who analyzed Brazil using time series on a daily scale, an average correlation of 0.6 was obtained between data from MERRA-2 and data measured at the spatial level of wind farms. In contrast, we achieved a correlation of 0.9 for wind farms and turbines. This shows how promising the results of this work are, and that corrected MERRA-2 data are suitably representative.

Among the different treatment options for the MERRA-2 data, focusing primarily on the hourly time series, the INMET-H strategy was best overall. It had the best performance in four of them and was second ranked in the other. After INMET-H came EXT, which was the MERRA-2 time series only interpolated and extrapolated. It performed best in one metric and was second in three. All strategies performed well in treating the MERRA-2 time series, but GWA was the least recommended because it had the worst performance in four metrics. Due to the negative result for MBE, there was overestimation of wind speeds by the GWA treatment.

Analysis of the treatments of the daily and monthly time series of MERRA-2 showed that the EXT strategy was best, followed by the INMET-G in four metrics for the daily time series and in all the monthly series. For the RMSE, MBE and MAE metrics, the GWA always performed the worst.

Many studies have used GWA data for bias correction (BOSCH; STAFFELL; HAWKES, 2018; GRUBER et al., 2019, 2021; GRUBER; SCHMIDT, 2019; NEFABAS et al., 2021; RYBERG et al., 2019), but this GWA result is contrary to the recommendation of Gruber et al. (2019), who suggested the use of GWA instead of INMET data for correction of wind farm data. Some hypotheses can be suggested to explain these divergences. In Gruber et al. (2019), there are some divergent results within the metrics, which can be explained by several factors, such as: the heterogeneity of the topography of Brazil and its great extension; use of inconsistent INMET time series or stations located very far from the assessed wind farm; use of inconsistent generation time series estimated by ONS (National Electrical System Operator) for comparison with simulated wind farm generation; and the use of some inconsistent geographic locations of the wind farms due to not using the best source to obtain this information (GRUBER et al., 2021). Also reported that the data of the GWA were updated in October 2019, and the quality of the corresponding treatments applied on the ERA5 data for the set of countries considered (South Africa, Brazil, United States and New Zealand) decreased in comparison with the same treatment with the old GWA data. However, the analyses of the use of the MERRA-2 data with bias correction for these two versions of the GWA for Brazil are only presented in the appendix of Gruber et al. (2021), and the best performances were achieved with the new GWA version, unlike our results.

The measured historical time series can have inconsistencies, which can compromise the quality of the results. For example, the measured wind speed time series used had high values, above 25 m/s, which were disregarded in the analysis to overcome inconsistencies. Furthermore, significant variability was found among the measured historical time series, even when pertaining to the same wind farm or neighboring farms. Thus, the result changed significantly depending on the metric used and the selected measurement time series, as shown in Table 2.4. One of the possible reasons for this variability between wind speed time series is the wake effect, which corresponds to the decrease in wind speed after passing through previous turbine blades. We did not consider this wake loss here, but it is described in Murcia et al. (2022) and Nefabas et al. (2021). Thus, one cannot blame only the models of the reanalysis data for the differences found between the MERRA-2 time series and the measured time series. There are many more factors that can cause these differences. Despite all these limitations, we believe the treated MERRA-2 time series represent the measured wind speed time series adequately, so we recommend applying the INMET-H treatment strategy to the MERRA-2 hourly time series and the EXT for the other temporal scales.

Gruber et al. (2021) employed ERA5 data corrected by the old GWA data to evaluate the impacts of spatial and temporal aggregation through the Pearson correlation and RMSE metrics. Although their approach was to evaluate a set of countries (low spatial resolution) and we examined sites in Brazilian territory (high spatial resolution), both studies concluded that temporal and spatial aggregation improved the representativeness of the historic series measured by the series generated by the reanalysis data. However, we also reached this conclusion for the MERRA-2 data, using five bias correction approaches and five different metrics to compare velocities (Gruber et al. (2021) compared energy generation) and applied a different aggregation approach. In contrast Gruber et al. (2021), used hourly time series aggregation resulting from their method while we aggregated the input data.

MERRA-2 data underestimated the variability of wind speed when presenting data with low spatial resolution, especially in areas with complex terrain, as is the case of most Brazilian wind farm sites (CRADDEN et al., 2017; OLAUSON; BERGKVIST, 2015; STAFFELL; PFENNINGER, 2016). Hence there is a need to treat time series to overcome this problem. Studies such as have suggested using spatial aggregation to circumvent the problem and reduce bias. The treatments under the MERRA-2 time series reconciled with spatial aggregation led to even better results. Thus, we used the treatments of the MERRA-2 time series and created measured time series that would represent larger areas to seek better results. As shown in 2.4.3.1, 2.4.3.2, the results were satisfactory for the correlation in analysis 2 (Wind Farm 2) and the RMSE, MAE, and Variances diff. were all satisfactory in analysis 1 (Wind Farm 1).

2.6

Conclusions

Based on the results of this study, the treatment strategies were satisfactory for the MERRA-2 wind speed time series to represent the actual wind speed time series measured at the turbines. The INMET-H treatment is recommended for hourly MERRA-2 time series and the EXT when using other temporal scales. This study's interpolation and extrapolation methods are efficient due to the good results achieved by the time series with the EXT treatment in all metrics. This did not present any type of bias correction based on another data source.

Among the treated time series that received bias correction, we highlight the contribution of databases with measured data to improve the quality of the reanalysis data time series. Bias correction by INMET data filtered on an hourly basis achieved the best values of most metrics. It best represented the historical data of the turbines according to the correlation metrics RMSE, MAE, and Variances diff. Most other articles have suggested using the GWA database to correct the bias of the reanalysis time series. However, its correction of most metrics (RMSE, MBE, MAE, and Variance diff.) presented the worst performance in this study.

The temporal aggregation to apply the method had positive effects, based on the correlation metrics RMSE, MAE and Variances diff., because the results improved as the time series were more temporally aggregated. Spatial aggregation also had satisfactory results but needs to be better evaluated in larger areas.

The results of this study indicate the good potential of the MERRA-2 reanalysis dataset, and show that it is possible to obtain accurate wind speed data for Brazilian territory after applying the strategies presented. These data can represent what happens in turbines, wind farms, regions, states and countries, in order to fill gaps in measured data and facilitate studies. Good primary data can be

used in forecasts and simulations, both of wind speeds and generation, which serve as inputs for feasibility studies of wind farms and energy dispatch models. In the final analysis, this enables improvement of the country's energy planning and stimulates the development and growth of wind generation.

Regarding future work, we suggest applying more bias correction methods, seeking to obtain better performance, applying this method in other locations in Brazil, and using other wind speed data from reanalysis, such as the ERA5 data. The GWA also provides Weibull wind speed distributions, which could also be investigated as an alternative to validate the results. We also recommend using the method presented here in feasibility studies of new wind farms. Finally, we suggest complementing the analyses through the conversion of corrected MERRA-2 wind speed time series into simulated wind generation for comparison with the observed wind generation of turbines. Thus, it will be feasible to measure the dimension of the variability inserted by the generation simulation model.

2.7

Acknowledgments

This work was supported by the Brazilian Coordination for the Improvement of Higher Level Personnel (CAPES) under Grant [number 001]; the Brazilian National Council for Scientific and Technological Development (CNPq) under Grants [numbers 307403/2019-0 and 422470/2021-0], the Carlos Chagas Filho Research Support Foundation of the State of Rio de Janeiro (FAPERJ) under Grants [numbers 202.673/2018, 211.086/2019, 211.645/2021, 201.243/2022 and 201.348/2022] and ANEEL R&D Program and Grupo Energisa under Grant number 06585-1802/2018 (Análise Estocástica de Perdas Técnicas em Sistemas de Energia).

2.8

Appendix A. Supplementary material

Supplementary material (tables, graphs and codes) to this article can be found online at

https://github.com/paulamacaira/Custodio_Cyrino_Macaira_Energy_2022_Representativeness-MERRA2-Brazil.

3 Second Contribution: Joint modeling of wind speed and power via a nonparametric approach

3.1

Abstract

The energy generated by wind farms depends on the wind speed, air density, turbine rotor height and blade area, among many other technical and climatic characteristics. To simplify the calculations, power curves are used to estimate how the power output varies with the wind speed. However, this approximation is one of the sources of uncertainty in wind energy forecasting models because it assumes that the wind speed is distributed in fixed intervals and that a given wind speed will always generate the same power. Therefore, this contribution proposes a new nonparametric method to model better the relationship between wind speed and the corresponding power generated. The steps of this proposal consist of using the Kmeans clustering technique to estimate the wind speed intervals, the kernel density estimation (KDE) method to define the probability density function (PDF) for each interval, and Monte Carlo simulation to infer the power output based on the PDF. We conducted tests of the proposed method through four approaches (single period, monthly, hourly, monthly-hourly) to ascertain its performance. The data came from the MERRA-2 database and five wind farms in northeast Brazil. The proposed method had superior performance than the conventional estimation technique. This study contributes to the literature by (i) proposing a new non-parametric method for modeling the relationship between wind speed and power, (ii) highlighting how probabilistic modeling better represents the natural variability of wind generation when compared to deterministic modeling, (iii) demonstrate how the temporal segregation of data for applying methodologies, to respect the annual and daily seasonality of wind generation, can lead to better method performance, (iv) demonstrate that even within the same region, wind farms can have different generation profiles due to environmental and technical conditions, and (v) evidence of the importance and quality of wind speed data made available by the MERRA-2 reanalysis database.

3.2

Introduction

The growing demand for energy from renewable sources has led to the construction of many wind farms in Brazil and other countries. However, two main factors can hamper the development of the wind power industry: the stochastic nature of wind speed and the uncertainty involving its production (AYODELE; OGUNJUYIGBE, 2015; IPAKCHI; ALBUYEH, 2009). These factors require precise and reliable models that replicate the behavior of local wind generation to support the planning, monitoring and operational management of the power grid and increase the reliability of the electrical network (FANG; WANG, 2017; HAN; YANG; LIU, 2007; REZVANI et al., 2019; THAPAR; AGNIHOTRI; SETHI, 2011).

For the correct dimensioning of generation, it is necessary to understand the functioning of wind turbine generators (AYODELE; OGUNJUYIGBE, 2015). It is well-established that the power from a wind turbine varies in function with the wind speed, air density and turbine blade parameters (size, design, tip speed and pitch angle ratio) (CARRILLO et al., 2013). The manufacturers usually supply a power curve for their models, which is typically generic and related only to the wind speed at the power generated (AL-QURAAN et al., 2022). Nevertheless, this information serves as a parameter for initial monitoring of the performance of a wind farm's turbines (SHETTY; SATHYABHAMA; PAI, 2020).

For managers of wind farms and power systems, it is not advisable to apply the manufacturer's power curve directly since the real working conditions of a turbine can be very different. The power curve made available by the manufacturer is projected for the functioning of a single wind turbine operating under ideal conditions. It does not consider the various types of interference they generate with each other within a wind farm. Hence, a need exists to develop models to estimate wind power in specific generation locations (CARTA; RAMÍREZ; VELÁZQUEZ, 2009; SHETTY; SATHYABHAMA; PAI, 2020).

Researchers have described various deterministic or probabilistic strategies to create reliable estimation models. These power curve models are also subdivided into parametric and nonparametric techniques (CARRILLO et al., 2013; SHOKRZADEH; JAFARI JOZANI; BIBEAU, 2014; SOHONI; GUPTA; NEMA, 2016). The parametric methods are based on mathematical formulations, such as the linearized segmented model, polynomial power curves (GIORSETTO; UTSUROGI, 1983; WEN; ZHENG; DONGHAN, 2009), exponential (MATHEW, 2006), cubic (CARRILLO et al., 2013), maximum principle method, least squares method (SHOKRZADEH; JAFARI JOZANI; BIBEAU, 2014), cubic spline interpolation, dynamic power curve, and models based on probabilistic distributions (SHOKRZADEH; JAFARI JOZANI; BIBEAU, 2014; THAPAR; AGNIHOTRI; SETHI, 2011), among others. In general, they are methods that are easy to apply and calculate the parameters, need little historical data, and have precise adjustments. Their drawbacks are poor accuracy, the need for comprehensive data from the manufacturer and the local conditions, and failure to consider the variability of this data (ESTEVES et al., 2019; KHODABUX et al., 2022).

The nonparametric models include copulas (GILL; STEPHEN; GALLOWAY, 2012; WANG et al., 2014), kernel density estimation, artificial neural networks (MURALIDHARAN et al., 2023), fuzzy logic systems, support vector machines, response surface methodology and data mining algorithms (such as random forest and clusterization) (MURALIDHARAN et al., 2023), among many others (KUSIAK; ZHENG; SONG, 2009; LYDIA et al., 2014; MARVUGLIA; MESSINEO, 2012; SHETTY; SATHYABHAMA; PAI, 2020; ÜSTÜNTAŞ; ŞAHIN, 2008). These do not impose any model specified in advance, are more precise, contemplate the variance of the data and estimate the power curve as closely as possible with the available data subject to smoothing of the fit. However, these methods require long historical data series and are complex to implement (DUCA; FONSECA; CYRINO OLIVEIRA, 2022; SOHONI; GUPTA; NEMA, 2016).

In Muralidharan et al. (2023), the authors used machine learning, artificial neural networks (ANN), decision trees (DT) and random forests (RF) to infer the best alternative to represent the outputs of a wind turbine. The same evaluation is described in Duca et al. (2022) via three dynamic Bayesian models considering

wind speed and power. In turn, Gill et al. (2012) and Wang et al. (2014) proposed the use of copula functions to model this relationship through the measured data, while Gill et al. (2012) also used a probabilistic method to exclude outliers. The article by Lázaro et al. (2022) evaluated the performance of multivariate models based on the Gaussian mixture copula model (GMCM), artificial neural networks and Bayesian artificial neural networks (BANN) to characterize the power curve and estimated power.

Many other studies have also used wind speed distribution as a basis to model the relationship between wind speed and power output. Most of these have recommended using a two-parameter Weibull distribution (WACKER; SEEBASS; SCHLÜTER, 2020). More recent studies have suggested modifications of the Weibull distribution (MILAD et al., 2023), and the use of a four-parameter Kappa probability distribution or five-parameter Wakeby probability distribution to obtain better results (JUNG; SCHINDLER, 2019).

In Carrillo et al. (2013), it described the fitting of polynomial, exponential and cubic equations to represent the power curves of turbines. The authors found the polynomial and cubic approaches to be most precise. Still, the polynomial approach required complex equations, making it hard to find a general expression, leading to the recommendation of the cubic, which only depends on the parameters provided by the manufacturer. Esteves et al. (2019) warned that these mathematical models should only be used for initial evaluation of the power output, because they do not precisely consider the inflection point of the power curve and can result in large forecasting errors. They serve to evaluate new undertakings, dimensioning and optimization of costs. Khodabux et al. (2022) suggested using sigmoid, logistic and Hill functions, while shokrzadeh et al. (2014) proposed the spline regression method to obtain better performance and overcome the problems of determining the curve's inflection point. However, although these mathematical expressions are widely used, there is little evidence that these curves fit the data pertaining to real wind power turbines (GIORSETTO; UTSUROGI, 1983; MATHEW, 2006; WEN; ZHENG; DONGHAN, 2009).

For this reason, this study aims to describe a model that more accurately represents this relationship, requiring only historical data on wind speed and generation. Although the method uses only these two data, we believe it can incorporate the complexity inherent to the random nature of the various data sources that also influence the resulting wind energy.

In this context, we present a method to segment the wind speed ranges through clusterization of data by K-means, mapping of possible wind power outputs associated with each range and constructing the probability density function of the power data by the kernel density estimation method. Finally, we apply Monte Carlo simulation to make the model flexible regarding wind power generation.

Through these methods, we aim to offer a robust and effective way to model the relationship between wind speed and wind power generation, permitting a deeper and more precise understanding of the performance of wind turbines in the field. Once the model is implemented, one only needs wind speed projections to estimate wind power.

This article is divided into five sections including this introduction. In the second section, we present the method used to model the relationship between wind speed and power; in the third section we present the data used in the tests; in the fourth section, we describe and discuss the results; and in the fifth section presents our conclusions and some recommendations for future research.

3.3

Proposed Method

The purpose of this article is to describe a method that can model the nonlinear relationship between the speed of the wind that passes through wind turbines and the power generated by them, thus enabling us to explore the peculiarities of the source and place and depict the probability distribution of generation for a given wind speed value. The steps used in this study to attain this goal are presented in Figure 3.1.



Figure 3.1 Steps of the method.

Section 3.3.1 presents the locations used to test the proposed method. Section 3.3.2 shows the steps employed in our previous contribution (DE AQUINO FERREIRA; CYRINO OLIVEIRA; MAÇAIRA, 2022), covering how to obtain and treat the wind speed reanalysis data from MERRA-2 for points in Brazilian territory, to have data that represent local reality and overcome the shortage of measured data points. We describe the main contributions of this article in sections 3.3.3.1 to 3.3.3.3, regarding the wind power modelling and forecasting method. Section 3.3.4 presents a simple and widely adopted method to compare our proposal here against other methods used to relate wind speed and wind power generation. Finally, section 3.3.5 offers the metrics adopted to evaluate the method.

3.3.1

Wind farm data

To test the method presented in this article, we used information from five wind farms located in the same state in Brazil's Northeast region, three along the coast and the other two in the interior of the state. The information the wind farm managers provided included the geographic locations, installed capacities and generation time series, which cannot be disclosed due to a secrecy agreement. The other information necessary to apply the method, such as the technical characteristics of the turbines used by the wind farms, was obtained from the Electric Sector Georeferenced Information System of Brazil's National Electric Energy Agency (SIGEL/ANEEL) (SIGEL ANEEL, 2023) and The Wind Power database (PIERROT EI, 2024).

These wind farms were chosen due to the convenience of measured data availability and their location in the Northeast region, where installed wind capacity is expanding strongly due to favorable conditions (DE AQUINO FERREIRA; CYRINO OLIVEIRA; MAÇAIRA, 2022; DOS SANTOS et al., 2024; DUCA; FONSECA; CYRINO OLIVEIRA, 2022; GRUBER et al., 2019).

3.3.2

Wind speed data

The wind farms chosen for testing do not have any wind speed history, only historical generation data. However, the power generated by these turbines is directly related to the wind speed at the height of the corresponding rotor, and modeling this relationship is the purpose of this article.

To overcome this limitation, we used reanalysis data from MERRA-2 (GMAO, 2021). It provides a free historical series on climate variables and atmospheric data since 1980, available at grid points covering the Earth's entire land surface. The MERRA-2 databases used most often are "MERRA-2 tavg1_2d_slv_Nx" and "MERRA-2 inst1_2d_asm_Nx" to obtain historic wind speed data. The main difference between them is that the wind speed data are hourly averages in the former, and in the second, the data are collected instantaneously. We adopted the first database for this study because it presented the best results in preliminary tests and the wind power data were also hourly averages (GMAO, 2021).

To adjust the wind speed time series of MERRA-2 to the natural conditions of the turbines in question, we used interpolation, extrapolation and bias correction (DE AQUINO FERREIRA; CYRINO OLIVEIRA; MAÇAIRA, 2022). The interpolation adopted was based on the nearest neighbor technique, which consists of obtaining the data from the point on the MERRA-2 grid closest to the geographic location of the wind farm to represent the speeds that occurred at that point. In turn, we used the power law as the extrapolation technique, to adjust the historic wind speed data of MERRA-2 to the values that occurred at the height of the rotors (GRUBER et al., 2019). Finally, the bias correction involved applying an average hourly correction factor using data from the nearest National Meteorology Institute (INMET) station. This technique consists of calculating the hourly wind speed averages of the MERRA-2 and INMET datasets at the same height (in this case, 10 meters), and calculating the ratio between them each hour, to generate an hourly correction factor to be applied to the extrapolated wind speed time series to adjust it to the real winds measured in Brazilian territory and contemplate the local particularities (GIORSETTO; UTSUROGI, 1983; ÜSTÜNTAŞ; ŞAHIN, 2008). This method is suggested by De Aquino Ferreira et al (2022) to obtain data from MERRA-2 for Brazil.

3.3.3

Modeling the speed versus power relationship

A particular wind speed value can generate different power values from the same turbine at distinct moments. A combination of factors, such as variations in wind direction, relative air humidity, atmospheric pressure and other climatic and atmospheric variables can explain this. However, mapping all these wind generation variations in the function of such variables would make the model highly complex and still would not assure identifying all the factors that can impact the generation (CARRILLO et al., 2013).

Thus, our objective here is to simplify the modeling of wind generation, to enable projecting it in function of a single variable, the wind speed, striving to create a model that explains all the randomness arising from the various data sources solely according to wind speed. To materialize the idea, we segmented the wind speeds that can occur in a wind farm, mapped each range's generation possibilities, and dimensioned the generation based on the speeds we believe will happen. The techniques used to materialize this method are presented in sections 3.3.3.1 to 3.3.3.3.

3.3.3.1

Clustering of wind speeds

We used the K-means clustering technique to identify the wind speed ranges (CELEBI; KINGRAVI; VELA, 2013; MARVUGLIA; MESSINEO, 2012). It groups the wind speed data into *K* clusters (speed range groups), where each data point is assigned to the cluster with more similar data.

The technique's algorithm starts by randomly defining K centroids (centers of clusters). Then the distance is calculated between each data point and the initial K centroids, and the value is attributed to the nearest cluster. Next, new centroids are defined, each being the average of all the data composing the respective cluster. This process is repeated until the centroids remain fixed after multiple iterations (LYDIA et al., 2014).

The K-means method has a premise consisting of the number of clusters, K, to be adopted. Hence, to find the most suitable number of clusters, the elbow method is used, which is applied by gradually increasing the number of clusters, and with each addition verifying whether on average, the standard deviation of the clusters is reduced, such that the accuracy of the centroid reflecting the historic data is increased. Thus, the number of clusters is defined when the variations of the standard deviation become negligible when including another cluster (ESTEVES et al., 2019).

3.3.3.2

Density curve estimation

Since power generation is a continuous variable and we want to understand the frequency of its values for each cluster of velocities, we estimated the probability density function (PDF) for the generation of each cluster. This association is possible because each wind speed value is associated in time with a historic occurrence of wind farm generation. The technique used to estimate the probability density functions is kernel density estimation (KDE), a nonparametric estimation approach. Its main advantage is that it is not necessary to assume the distribution of the sample data in advance, thus avoiding the introduction of subjective previous information (LI et al., 2023; WAHBAH et al., 2019).

Since the KDE involves fitting the density function concerning the data, this process is more precise and robust compared to parametric approaches (LI et al., 2023). If x denotes the sample dataset, f(x) can be expressed as in Equation 1,

$$f(x) = F'(x) = \frac{F(x+h) - F(x-h)}{2h}$$
(1)

where *h* is a non-negative constant, called the bandwidth, and F(x) represents the empirical distribution of the power data that compose the sample.

To assure a better fit of the probability distribution, the sample size is allowed to tend to infinity $(n \to \infty)$ and the bandwidth to zero $(h \to 0)$ (LI et al., 2023; WAHBAH et al., 2019). Hence, the expression for the estimator of f(x) can be defined as in Equation 2,

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$
(2)

where K(.) is the kernel function. Based on Equation 2, note that the determination of the estimates, $\hat{f}(x)$, mainly depends on h and K(.) (CELEBI; KINGRAVI; VELA, 2013).

3.3.3.3

Simulation

We used Monte Carlo simulation to guarantee the variability of the wind power generation in function of the wind speed (WORTON, 1995).

According to the previous steps, each speed range from a cluster has an associated probability density function, for which the cumulative density function is calculated. This randomly generates a value from a uniform [0,1] distribution, which is used in the cumulative function to find the power value that produces the corresponding cumulative probability (SHOKRZADEH; JAFARI JOZANI; BIBEAU, 2014; THAPAR; AGNIHOTRI; SETHI, 2011; VAN RAVENZWAAIJ; CASSEY; BROWN, 2018; WORTON, 1995).

To obtain the following power values, it is necessary to know to what cluster the next wind speed value belongs to find the corresponding cumulative function. Next, a new random probability value is generated for the cumulative function. This procedure is repeated until the finalization of the sequence of wind power generation estimates (WAHBAH et al., 2019).

3.3.4

Power curve

The conventional deterministic approach to obtain the available power of the wind that passes through a turbine's rotors is expressed in Equation 3 (CARRILLO et al., 2013; MACQUEEN, 1967).

$$P_t(v) = \frac{1}{2} A. \rho. v^3$$
(3)

where $P_t(v)$ is the theoretical power generated in watts when the wind speed v in m/s passes through a turbine with rotor area A, in m², and ρ is the air density.

However, the real power generated, $P_r(v)$, is lower than $P_t(v)$ due to mechanical and electrical losses and aerodynamic factors of the blades. The ratio between the powers generates the power coefficient, C_p , which is typically available from the turbine manufacturer. The theoretical maximum value of the power coefficient is 0.593, the Betz limit, but in practice this value is not attained in the turbines, with the maximum value being 0.5 (CARRILLO et al., 2013; THAPAR; AGNIHOTRI; SETHI, 2011).

Figure 3.2 represents the theoretical format of the power curve, the real relationship between wind speed and power (ESTEVES et al., 2019). Equation 4 represents this mathematically.



$$P(v) = \begin{cases} 0, & v < V_{Ci} \\ P^*(v), & V_{Ci} \le v < V_r \\ P_r, & V_r \le v \le V_{Co} \\ 0, & v > V_{Co} \end{cases}$$
(4)

where P(v) is the electric power generated, V_{Ci} is the initial cutoff wind speed, V_r is the nominal wind speed, V_{co} is the final cutoff wind speed, P_r is the nominal power and $P^*(v)$ is the power related nonlinearly with the wind speed.

The format of the nonlinear part of Equation 4 is related to the strategy of controlling the maximum power extraction from the wind, which is normally moderated by a cubic power curve, as expressed in Equation 5, where r is the rotor radius (CARRILLO et al., 2013).

$$P^*(v) = \frac{1}{2}\pi . r^2 . C_p . \rho . v^3$$
(5)

The regions identified in Figure 3.2 are not clearly demarcated in the operation of the wind turbines. The corresponding wind speeds that delimit the regions are found from averages of repeated measurements. This is another limitation of this approach (CARRILLO et al., 2013; DUCA; FONSECA; CYRINO OLIVEIRA, 2022).

3.3.5

Evaluation

The evaluation metrics adopted in this study were root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and coefficient of determination (R²). These are the most commonly used and recommended in the literature (AL-DUAIS; AL-SHARPI, 2023; CARRILLO et al., 2013; DUCA; FONSECA; CYRINO OLIVEIRA, 2023; ESTEVES et al., 2019; HAN; YANG; LIU, 2007; JUNG; SCHINDLER, 2019; LÁZARO; YÜRÜŞEN; MELERO, 2022; LI et al., 2023; MURALIDHARAN et al., 2023; WANG et al., 2019; WORTON, 1995). They measure to what extent the real (observed) power values differ from those predicted by the techniques used, reproducing the degree of dispersion between the two metrics. The RMSE and MAE are expressed per unit (p.u.), while MAPE and R² are expressed in percentage. Equations 6 to 9 are the mathematical formulas of the evaluation indices, where *x* denotes the historic time series of generation in the wind farm and *y* is the time series produced by the technique chosen.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{n}}$$
(6)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|$$
(7)

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{x_i - y_i}{x_i} \right|$$
(8)

$$R^{2} = 1 - \frac{\sum_{i=1}^{n} (x_{i} - y_{i})^{2}}{\sum_{i=1}^{n} (x_{i} - \overline{x_{i}})^{2}}$$
(9)

In the first three metrics, lower values are associated with better performance, while for R^2 , higher values mean a better result.

3.4

Data

In this section, we present the data used in the study and their descriptive analyses.

3.4.1

Wind power data

The historical time series from the wind farms are the hours of active power generation and did not need treatment. There were no missing data, power levels higher than installed capacity, or negative power values. There were differences in the historical dimensions of the wind farms due to their different startup dates. The farms located along the coast are the oldest and have been in operation for 9 years and 2 months (January 1, 2013 to February 28, 2022), while those located inland have only been operating for 4 years and 4 months) (November 1, 2017 to December 28, 2022). The other data necessary to apply the method, such as the technical characteristics of the turbines, were obtained from the SIGEL/ANEEL platform (SIGEL ANEEL, 2023) and are presented in Table 3.1.

Table 5.1	Information		a laillis.			
Wind	Location	Installed	Number	Total	Rotor	Operating
Form	in the	Capacity	of	Height	Diameter	Deriod
1 ai iii	State	(KW)	Turbines	(m)	(m)	I enou
1	Interior	31.500	15	194,00	114,00	November 1,
2	Interior	63.000	30	194,00	114,00	2017 to February 28
						2022 202
3	Coast	10.200	13	98,85	48,00	January 1, 2013
4	Coast	48.000	60	99,60	48,00	to February 28,
5	Coast	4.500	3	123,50	77,00	2022

Table 3.1 Information about the wind farms.

Figures 3.3 and 3.4 depict the characteristics of the historical data on the active power of the wind farms. From Figure 3.3, it can be noted that the density curves of each region are similar. The density curves of the wind farms on the coat are unimodal, with a top value near 10% of installed capacity. In contrast, the curves of the interior facilities are bimodal, with tops at the extremes. The data on the interior wind farms are more variable than those on the coast, as illustrated in the boxplots of Figure 3.3.



Figure 3.3 Density curve and boxplot of the wind farms' power data.

Figure 3.4 presents the boxplots that indicate the generation variability during the different hours and months of operation. The wind farms in the interior have the lowest generation during the afternoon, while those located along the coast have highest generation during that period. Nevertheless, from a monthly standpoint, all the wind farms had higher generation during the year's second half, with peaks in August and September.



Figure 3.4. Boxplot of the wind power data per hour and month.

3.4.2

Wind turbines

The theoretical power curve, although having well-defined regions, has variations at the limits of those regions according to the manufacturer's

specifications. The turbine models used by the wind farms examined in this study are identified in Table 3.2. This information was obtained from a previous study (CARRILLO et al., 2013) and The Wind Power database (PIERROT EI, 2024).

Wind	Manufacturar	Model	Diameter	v _{ci}	v_r	v_{co}	C .
Farm	Ivianulactulei	WIOdel	(m)	(m/s)	(m/s)	(m/s)	$c_{p,máx}$.
1	Gamesa	G114/2100	114	3,5	10	25	0,45
2	Gamesa	G114/2100	114	3,5	10	25	0,45
3	Enercon	E48/800	48	3	14	25	0,50
4	Enercon	E48/800	48	3	14	25	0,50
5	IMPSAr	IV-77-1500	77	3	13	22	0,42

Table 3.2 Turbine models at the wind farms.

3.5

Results and Discussion

In this section, we present the results reached in estimating the power based on the (i) cubic power curve, which is a traditional, deterministic and parametric method; and (ii) our proposed method, which is probabilistic and nonparametric. We present it in four versions: single period, monthly, hourly and monthly-hourly. The difference among them consists of segregating the data according to different temporal approaches to apply the three steps of the method separately. To compare the performance of the modeling and forecasting options, we used the metrics RMSE, MAE, MAPE and R², all with the R software, version 4.3.0 (R CORE TEAM, 2018).

To test the proposed method, we used data from five wind farms in Brazil's Northeast region. The wind farm administrators supplied the time series of the power outputs, while the historic wind speed time series were obtained from the MERRA-2 reanalysis database. To adjust the wind speeds from MERRA-2 to those at the height of the wind farms' turbines, we used the method described in (DE AQUINO FERREIRA; CYRINO OLIVEIRA; MAÇAIRA, 2022), consisting of interpolation, extrapolation and bias correction.

Table 3.3 presents some of the data used and calculated in adjusting the wind speed time series from MERRA-2 to each wind farm's conditions. Gruber et al. (2019) recommend applying the bias correction step to MERRA-2 data only when the INMET station is 40 km from the wind farm. Otherwise, the bias correction is not advisable due to the considerable distance between the station and wind farm, such that the INMET data may not adequately represent the real local conditions of the wind farm.

	Doton Unight	Distance						
Wind Farms	(m)	MERRA-2 – Wind	INMET – Wind					
	(111)	Farm (km)	Farm (km)					
1	137,00	18,64	52,97					
2	137,00	14,95	54,65					
3	74,85	3,41	19,86					
4	75,60	9,01	17,92					
5	85,00	7,61	13,49					

Table 3.3 Data from adjusting the wind speed data from MERRA-2.

In this study, some wind farms are located further than 40 km from an INMET station. Aiming to analyze the impact of using these data in the bias correction, we performed tests with the wind speed time series with and without bias correction.

The other data, such as geographic coordinates of the nearest MERRA-2 grid points to the wind farms and INMET stations are not identified based on a confidentially agreement with the wind farm operator.

With the wind speed data, we carried out the wind power estimations using different techniques and approaches. The theoretical power curve method used the information from the manufacturer presented in Table 3.2 to delineate the regions of zero, constant, and variable generation in terms of the wind speed, as depicted in Figure 3.2.

The modeling of the nonlinear region was performed with the cubic power curve presented in Equation 5, with the parameter ρ set to 1160 kg/m³, as recommended by Silva (2003), and the C_p values were adjusted to enable the formation of a theoretical power curve based on the other parameters of the curve assumed to be known and true (SILVA, 2003). The power coefficient values used for each wind farm are reported in Table 3.4. We compared the power outputs generated by the curves based on the wind speed time series with the historical power time series through the evaluation metrics, and the results are listed in Tables 3.6 to 3.8.

Table 3.4 - Turbine models	adopted in this study
----------------------------	-----------------------

Wind Farm	1	2	3	4	5
C_p	0,3547	0,3547	0,2724	0,2778	0,2528

Our proposed methodological innovation seeks to better estimate the power generated by wind turbines compared to the traditional cubic power curve approach. Its steps involve clusterization of the wind speeds, estimation of the density curves of the power generated broken down into speed range, and simulation of the power estimation. These steps were tested in four approaches: single period, monthly, hourly and monthly-hourly.

In the first step, we grouped the wind speed data of each farm using the Kmeans technique, with clusterization being performed only once with all the data in the single-period approach. In the monthly approach, the clusterization was carried out 12 times, each with the wind speed data for the respective month, while this was 24 times in the hourly approach and 288 times in the monthly-hourly case. In all cases, defining the number of clusters by the elbow method was first necessary. In the single period approach, the number of clusters varied from 20 to 22, in the monthly case it ranged from 17 to 22, in the hourly case from 12 to 20, and in the monthly-hourly case from 3 to 14. Table 3.5 details the number of clusters used in each test. Irrespective of the wind farm and the steps adopted to construct the wind speed time series, the number of clusters defined by the elbow method for the application of the clusterization was practically the same, only diminishing when increasing the segregation of the data to apply our proposed method.

Wind	Speed	A	pproaches of th	e Proposed Me	thod
Farm	Treatment	Singular	Monthly	Hourly	Monthly- hourly
2	EXT	23	18-22	12-20	3-12
1	INMET	20	17-22	17-20	3-11
-	EXT	20	17-21	15-20	5-12
2	INMET	20	18-22	16-20	5-12
-	EXT	21	18-22	14-20	6-12
3	INMET	22	18-21	15-20	5-10
-	EXT	22	18-22	13-20	6-12
4	INMET	21	17-22	13-19	5-12
-	EXT	22	18-22	12-19	5-12
5	INMET	22	18-22	12-20	6-12

Table 3.5 - Number of clusters obtained by each approach.

Figure 3.5 presents the result of applying the elbow method to Wind Farm 5 using a single period with extrapolated speed data (single-EXT) and a dispersion graph of the speed and power data. In the second graph, the different blue shades indicate the different wind speed clusters, permitting visualization of the variability of power levels generated in each speed range. Corresponding graphs for the other farms and approaches are contained in the appendix.



Figure 3.5 Result of clusterization (Wind Farm 5).

We then applied KDE to estimate the probability density function of the power values for each cluster resulting from the previous step. Figure 3.6 contains the PDF resulting from this step for each wind speed interval of Wind Farm 5 by the single-EXT approach. Note that as the wind speed values of the intervals increase, the peak of the density curve shifts toward the end of the interval. This behavior was seen for all the wind farms and approaches, and was expected since faster wind speeds tend to generate more energy. The figures with the probability distributions of all the speed ranges for all the wind farms and approaches are shown in the appendix.



Figure 3.6 Wind power density curves (Wind Farm 5).

It is important to mention that the KDE technique adopted in this work, based on Li et al. (2023) and Wahbah et al. (2019), generates probability distributions that can infer negative power values. Therefore, we tested other methods, such as KDE Beta, but the results were worse, observed by visualizing the graphs and the evaluation metrics.

To emulate the variability of the behavior of a turbine in operation, we performed Monte Carlo simulation, which requires the wind speed time series and the density functions of the clusters. In applying the simulation for a single period, only the PDF utilized varied based on the generating wind speed cluster. At the same time, in the approaches with temporal segmentation, the choice of the PDF changed due to wind speed and instant of time for estimating the wind power.

To measure the quality of the results, we used the metrics RMSE, MAE, MAPE and R² comparing to the active, historical and estimated power time series. To construct the estimates by the method proposed here, we generated 100 power time series simulations and calculated each time point's average power. The number of scenarios adopted was found experimentally by ceasing to increase the number of scenarios when no further gains were obtained based on the evaluation metrics (only increase in computational cost). Table 3.6 depicts the results obtained by the different estimation strategies, with the best result of each test highlighted in boldface. Note that irrespective of the evaluation metric, wind farm analyzed or steps adopted to construct the wind speed time series employed in the test, our proposed method always performed better than the cubic power curve, and the monthly-hourly strategy was the best in all cases.

		_			Method	S	
Metrics	Wind	Speed	Cubic		Pı	oposals	
Wettes	Farm	Treatment	Power Curve	Singular	Monthly	Hourly	Monthly- Hourly
	1	EXT	0,6012	0,2315	0,2210	0,2029	0,1772
	1	INMET	0,4489	0,2428	0,2310	0,2032	0,1771
	2	EXT	0,5971	0,2009	0,1892	0,1775	0,1538
	2	INMET	0,3224	0,2103	0,1942	0,1778	0,1538
DIGE	2	EXT	0,1624	0,1597	0,1546	0,1559	0,1454
RMSE	3	INMET	0,3439	0,1713	0,1640	0,1559	0,1455
		EXT	0,1590	0,1444	0,1385	0,1372	0,1249
	4	INMET	0,2804	0,1482	0,1416	0,1373	0,1248
	_	EXT	0,2590	0,1580	0,1537	0,1330	0,1245
	5	INMET	0,2767	0,1536	0,1494	0,1330	0,1247
	1	EXT	0,4529	0,1883	0,1784	0,1578	0,1349
	1	INMET	0,3697	0,1998	0,1884	0,1581	0,1347
	2	EXT	0,4449	0,1615	0,1509	0,1382	0,1179
	2	INMET	0,2554	0,1707	0,1557	0,1385	0,1180
	2	EXT	0,1245	0,1245	0,1201	0,1206	0,1112
MAE	3	INMET	0,2826	0,1343	0,1276	0,1206	0,1113
		EXT	0,1209	0,1118	0,1060	0,1050	0,0930
	4	INMET	0.2226	0.1142	0.1074	0.1051	0.0930
		EXT	0.2064	0.1230	0.1186	0.1018	0.0949
	5	INMET	0.2209	0.1177	0.1140	0.1019	0.0950
		EXT	0,4648	0,1932	0,1831	0,1620	0,1384
	1	INMET	0,3794	0,2050	0,1934	0,1623	0,1383
	2	EXT	0,4669	0,1695	0,1584	0,1451	0,1238
	2	INMET	0,2681	0,1792	0,1634	0,1453	0,1239
	-	EXT	0,1249	0,1249	0,1205	0,1210	0,1115
MAPE	3	INMET	0,2834	0,1347	0,1280	0,1210	0,1117
		EXT	0,1279	0,1183	0,1122	0,1112	0,0984
	4	INMET	0,2356	0,1209	0,1137	0,1113	0,0984
		EXT	0,2077	0,1238	0,1194	0,1025	0,0955
	5	INMET	0,2223	0,1184	0,1148	0,1025	0,0956
		EXT	-2,8184	0,4338	0,4840	0,5650	0,6682
	1	INMET	-1,1294	0,3769	0,4364	0,5637	0,6687
	2	EXT	-3,5560	0,4842	0,5426	0,5972	0,6979
	2	INMET	-0,3280	0,4348	0,5183	0,5961	0,6978
	2	EXT	0,4201	0,4387	0,4744	0,4657	0,5351
R ²	3	INMET	-1,6019	0,3542	0,4083	0,4651	0,5343
		EXT	0,2889	0,4139	0,4603	0,4704	0,5616
	4	INMET	-1,2102	0,3825	0,4365	0,4696	0,5619
		EXT	-0,8726	0,3029	0,3405	0,5063	0,5673
	5	INMET	-1,1378	0,3416	0,3772	0,5061	0,5660

Table 3.6 – Results reached by the methods according to the evaluation metrics.

With regard to our proposed method, as the segmentation of the data for application of the three techniques increased, the simulations based on the metrics indicated in Table 3.6 improved, with only one exception. In the test of Wind Farm 3 using speed EXT data, the monthly approach had better performance than the hourly, as identified by the four evaluation metrics.

Tables 3.7 and 3.8 detail the results of Table 3.6, where it is possible to see the behavior of the RMSE metric by month and hour in each test, respectively, and verify whether a trend exists other than that observed in Table 3.6. Again, the best performance was achieved by the monthly-hourly approach in all the wind farms in all months and hours. Nevertheless, in a few isolated months and hours, the theoretical power curve method performed better than the proposed technique, with single period and monthly approaches being better than the hourly approach. We performed all these analyses with the metrics MAE, MAPE and R²; the respective tables are in the appendix. The conclusions were the same as those with the RMSE.

Concerning the impact of the evaluation metric in function of the type of treatment applied to the wind speed, this factor did not influence the performance of the new strategies presented here. It mainly did not impair the identification of the best approach to estimate the wind power. Thus, as pointed out in our previous study (DE AQUINO FERREIRA; CYRINO OLIVEIRA; MAÇAIRA, 2022), the wind speed time series data that only underwent interpolation and extrapolation (EXT) also achieved good representation of the series measured at the farms, in some cases better than the series that underwent the third step, bias correction (INMET-H). This might have happened because the quality of the correction is directly related to the quality and quantity of measurements supplied by the INMET station and its distance from the wind farm. It is not always possible to satisfy these factors. For example, the EXT series produced better performance at wind farms 1 and 2, where the maximum distance requirement between the farm and INMET station was not satisfied. At the other wind farms, the best performance between the two treatment options in the monthly case was alternated.

Figures 3.7 and 3.8 compare the behavior of the historic data of the wind farms with the estimates resulting from the techniques. The blue dots and lines in the graphs refer to historical data, while the black ones denote the estimated values. Figure 3.7 presents the dispersion graphs, evidencing the behavior of the power variable in function of the wind speed variable. From these graphs, it is possible to infer that the strategies described in this study managed to better replicate the inherent generation variability for the same speed values, with the highlight on the monthly-hourly proposal, which best captured the existing stochasticity. Figure 3.8 contains the frequency polygon graphs that estimate the density curve based on the frequencies of the power value intervals. Once again, it can be noted that the suggested strategies (single period, monthly, hourly and monthly-hourly) better replicate the power frequencies obtained from the historical data than the cubic power curve.

Wind	Mathad	Speed						Mo	onth					
Farm	Wiethou	Treatment	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Set	Oct	Nov	Dec
	Theoretical	EXT	0,5298	0,4716	0,4596	0,4721	0,5240	0,5509	0,5497	0,6939	0,7611	0,7447	0,7119	0,6194
	Curve	INMET	0,3777	0,3204	0,3133	0,3549	0,4359	0,5188	0,5472	0,5663	0,5303	0,4645	0,4522	0,4353
	Cincular	EXT	0,2376	0,2327	0,2439	0,2345	0,2206	0,2288	0,2212	0,2098	0,2087	0,2328	0,2518	0,2426
	Singular	INMET	0,2477	0,2405	0,2540	0,2391	0,2280	0,2369	0,2326	0,2261	0,2231	0,2476	0,2655	0,2586
1	Monthly	EXT	0,2305	0,2145	0,2238	0,2238	0,2177	0,2165	0,1975	0,1842	0,1999	0,2320	0,2498	0,2398
1	Monuny	INMET	0,2399	0,2211	0,2283	0,2272	0,2246	0,2259	0,2091	0,1974	0,2102	0,2467	0,2623	0,2548
	Hough	EXT	0,2152	0,2208	0,2360	0,2266	0,2104	0,2084	0,2026	0,1745	0,1582	0,1775	0,1917	0,1993
	Hourry	INMET	0,2146	0,2215	0,2369	0,2270	0,2104	0,2081	0,2029	0,1755	0,1590	0,1775	0,1923	0,1998
	Monthly-	EXT	0,2027	0,1996	0,2028	0,2020	0,1938	0,1807	0,1608	0,1361	0,1382	0,1478	0,1602	0,1779
	Hourly	INMET	0,2034	0,1985	0,2025	0,2027	0,1931	0,1809	0,1600	0,1361	0,1384	0,1463	0,1598	0,1785
	Theoretical	EXT	0,5111	0,4643	0,4494	0,4743	0,5428	0,5540	0,5620	0,7154	0,7735	0,7322	0,6850	0,5895
	Curve	INMET	0,2857	0,2441	0,2382	0,2716	0,3149	0,3645	0,3792	0,3812	0,3561	0,3374	0,3368	0,3241
	Singular	EXT	0,2023	0,2007	0,2150	0,2075	0,1971	0,1972	0,1950	0,2018	0,1882	0,1931	0,2084	0,2019
		INMET	0,2092	0,2182	0,2367	0,2283	0,2203	0,2133	0,2070	0,2076	0,1913	0,1924	0,2008	0,1998
2	Monthly	EXT	0,1972	0,1810	0,1877	0,1869	0,1900	0,1891	0,1764	0,1763	0,1775	0,1921	0,2059	0,1998
Z	wonthly	INMET	0,2007	0,1881	0,1946	0,2002	0,2107	0,2064	0,1899	0,1731	0,1744	0,1886	0,1978	0,1992
	Houdy	EXT	0,1817	0,1867	0,2045	0,1966	0,1897	0,1831	0,1818	0,1731	0,1488	0,1493	0,1606	0,1705
	Hourry	INMET	0,1815	0,1880	0,2055	0,1974	0,1908	0,1824	0,1821	0,1726	0,1488	0,1501	0,1606	0,1699
	Monthly-	EXT	0,1742	0,1644	0,1725	0,1650	0,1676	0,1606	0,1486	0,1376	0,1290	0,1227	0,1366	0,1536
	Hourly	INMET	0,1739	0,1648	0,1722	0,1654	0,1685	0,1608	0,1488	0,1375	0,1299	0,1233	0,1349	0,1533
-	Theoretical	EXT	0,1368	0,1344	0,1285	0,1438	0,1588	0,1933	0,2139	0,1981	0,1851	0,1568	0,1390	0,1327
	Curve	INMET	0,2974	0,2875	0,2444	0,2534	0,2588	0,3278	0,3736	0,4292	0,4522	0,4296	0,3635	0,3277
	Singular	EXT	0,1352	0,1344	0,1287	0,1433	0,1595	0,1895	0,1987	0,1948	0,1834	0,1572	0,1385	0,1320
	Singular	INMET	0,1434	0,1430	0,1350	0,1464	0,1584	0,1884	0,1985	0,1975	0,2081	0,1912	0,1718	0,1560
2	Monthly	EXT	0,1335	0,1321	0,1284	0,1420	0,1504	0,1796	0,1901	0,1926	0,1803	0,1472	0,1299	0,1284
3	Monuny	INMET	0,1417	0,1411	0,1298	0,1406	0,1441	0,1816	0,1957	0,1900	0,1951	0,1777	0,1633	0,1506
	Handar	EXT	0,1296	0,1282	0,1198	0,1365	0,1527	0,1856	0,1950	0,1870	0,1808	0,1583	0,1421	0,1329
	Houriy	INMET	0,1302	0,1278	0,1195	0,1361	0,1526	0,1857	0,1953	0,1868	0,1812	0,1588	0,1418	0,1334
	Monthly-	EXT	0,1254	0,1246	0,1179	0,1324	0,1369	0,1701	0,1819	0,1783	0,1754	0,1410	0,1186	0,1199
	Hourly	INMET	0,1250	0,1247	0,1181	0,1336	0,1373	0,1695	0,1828	0,1787	0,1747	0,1413	0,1186	0,1199

Table 3.7 - RMSE attained by the methods per month

	Theoretical	EXT	0,1206	0,1165	0,1170	0,1387	0,1640	0,2044	0,2350	0,2087	0,1767	0,1393	0,1139	0,1115
	Curve	INMET	0,2468	0,2352	0,1985	0,2008	0,2040	0,2573	0,2966	0,3551	0,3755	0,3557	0,2941	0,2683
	Singular	EXT	0,1224	0,1183	0,1123	0,1269	0,1411	0,1686	0,1852	0,1834	0,1692	0,1439	0,1186	0,1155
	Singular	INMET	0,1247	0,1219	0,1126	0,1218	0,1321	0,1583	0,1724	0,1750	0,1838	0,1711	0,1477	0,1360
4	Monthly	EXT	0,1192	0,1149	0,1119	0,1255	0,1327	0,1573	0,1751	0,1809	0,1653	0,1337	0,1099	0,1101
4	Monuny	INMET	0,1233	0,1209	0,1101	0,1169	0,1198	0,1498	0,1673	0,1696	0,1744	0,1586	0,1397	0,1290
	Hourly	EXT	0,1178	0,1121	0,1024	0,1157	0,1280	0,1561	0,1677	0,1625	0,1646	0,1492	0,1299	0,1213
		INMET	0,1177	0,1120	0,1023	0,1159	0,1284	0,1562	0,1687	0,1622	0,1645	0,1494	0,1297	0,1212
	Monthly-	EXT	0,1114	0,1081	0,1013	0,1116	0,1114	0,1362	0,1520	0,1525	0,1582	0,1298	0,1023	0,1045
	Hourly	INMET	0,1116	0,1084	0,1012	0,1116	0,1113	0,1363	0,1513	0,1528	0,1577	0,1297	0,1027	0,1044
	Theoretical	EXT	0,1956	0,1843	0,1849	0,2150	0,2515	0,3200	0,3568	0,3349	0,3237	0,2609	0,1957	0,1915
	Curve	INMET	0,2531	0,2460	0,2060	0,1990	0,2106	0,2541	0,2920	0,3539	0,3400	0,3365	0,2965	0,2707
	Singular	EXT	0,1470	0,1425	0,1320	0,1358	0,1459	0,1650	0,1838	0,1959	0,1867	0,1663	0,1408	0,1359
	Singular	INMET	0,1361	0,1326	0,1239	0,1277	0,1384	0,1562	0,1724	0,1878	0,1813	0,1726	0,1536	0,1420
5	Monthly	EXT	0,1454	0,1397	0,1299	0,1324	0,1384	0,1576	0,1782	0,1928	0,1861	0,1609	0,1310	0,1316
5	wontiny	INMET	0,1352	0,1321	0,1210	0,1214	0,1286	0,1509	0,1702	0,1818	0,1791	0,1664	0,1474	0,1390
		EXT	0,1267	0,1190	0,1111	0,1136	0,1243	0,1451	0,1521	0,1477	0,1474	0,1410	0,1325	0,1261
	Houriy	INMET	0,1272	0,1193	0,1113	0,1138	0,1241	0,1446	0,1521	0,1482	0,1473	0,1406	0,1322	0,1261
	Monthly-	EXT	0,1224	0,1159	0,1088	0,1077	0,1122	0,1316	0,1448	0,1388	0,1428	0,1300	0,1139	0,1160
	Hourly	INMET	0,1229	0,1158	0,1087	0,1089	0,1125	0,1312	0,1445	0,1391	0,1433	0,1297	0,1142	0,1164

Table 3.8 - RMSE attained by the methods per hour

Wind		Speed												Но	urs						-					
Farm	Methods	Treatment	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
	Theoretica	I EXT	0.8585	0.7680	0.6671	0.5637	0.4918	0.4500	0.4154	0.3856	0.3657	0.5635	0.7824	0.7139	0.5953	0.4851	0.3998	0.3363	0.3079	0.3244	0.3817	0.4410	0.7188	0.8698	0.8972	0.8661
	Curve	INMET	0.4374	0.4393	0.4426	0.4511	0.4559	0.4554	0.4679	0.5026	0.4989	0.4709	0.3822	0.2862	0.2212	0.2001	0.2161	0.2578	0.3207	0.4025	0.5509	0.6563	0.6324	0.5978	0.5377	0.4630
		EXT	0.2526	0.2448	0.2342	0.2249	0.2216	0.2178	0.2123	0.2201	0.2151	0.2073	0.2269	0.2340	0.2130	0.1861	0.1640	0.1522	0.1699	0.2133	0.3110	0.3477	0.2546	0.2450	0.2467	0.2501
	Singular	INMET	0.2622	0.2565	0.2440	0.2296	0.2242	0.2181	0.2098	0.2197	0.2167	0.2307	0.1703	0.1951	0.2354	0.2357	0.2132	0.1825	0.1708	0.2090	0.3194	0.3965	0.3024	0.2669	0.2483	0.2516
		EXT	0.2290	0.2216	0.2138	0.2089	0.2098	0.2059	0.2022	0.2065	0.2016	0.1933	0.2181	0.2347	0.2196	0.1981	0.1769	0.1623	0.1684	0.2055	0.2966	0.3355	0.2504	0.2254	0.2224	0.2246
1	Monthly	INMET	0.2300	0.2287	0.2212	0.2128	0.2127	0.2069	0.1998	0.2043	0.2012	0.2106	0.1682	0.2017	0.2406	0.2431	0.2260	0.1937	0.1782	0.2010	0.2981	0.3729	0.2909	0.2510	0.2291	0.2240
		EXT	0.2501	0.2435	0.2297	0.2218	0.2158	0.2128	0.2032	0.2021	0.1957	0.2044	0.1549	0.1319	0.1203	0.1167	0.1259	0.1401	0.1669	0.1958	0.2237	0.2371	0.2292	0.2398	0.2463	0.2471
	Hourly	INMET	0.2495	0.2415	0.2302	0.2213	0.2170	0.2124	0.2034	0.2031	0.1968	0.2062	0.1553	0.1332	0.1188	0.1184	0.1265	0.1409	0.1676	0.1957	0.2252	0.2367	0.2312	0.2389	0.2461	0.2477
	Monthly-	FXT	0 2075	0 1999	0 1979	0.1948	0.1980	0.1960	0 1905	0.1836	0 1806	0.1635	0 1429	0 1270	0 1168	0 1135	0 1244	0.1369	0.1571	0.1783	0.1886	0.1808	0.1811	0.2020	0 2127	0 2130
	Hourly	INMET	0.2064	0.1985	0.1975	0.1948	0.1984	0.1965	0.1904	0.1842	0.1787	0.1650	0.1411	0.1262	0.1166	0.1126	0.1227	0.1370	0.1593	0.1788	0.1896	0.1817	0.1826	0.2037	0.2114	0.2099
	Theoretica	EXT	0.8431	0.7459	0.6414	0.5348	0.4559	0.4094	0.3861	0.3728	0.3606	0.5546	0.7666	0.6952	0.5766	0.4657	0.3856	0.3273	0.2976	0.3145	0.3642	0.4365	0.7875	0.9266	0.9239	0.8679
	Curve	INMET	0.2562	0.2419	0.2759	0.3181	0.3559	0.4008	0.4365	0.4770	0.4874	0.4540	0.3340	0.2397	0.1788	0.1509	0.1545	0.1973	0.2459	0.3041	0.3744	0.3911	0.2730	0.2798	0.2423	0.3135
	<u>.</u>	EXT	0.1993	0.1923	0.1829	0.1781	0.1795	0.1839	0.1901	0.2091	0.2130	0.1986	0.2091	0.2208	0.2011	0.1742	0.1536	0.1499	0.1643	0.2034	0.2590	0.2690	0.2227	0.2143	0.2066	0.2006
	Singular	INMET	0.2187	0.2111	0.2001	0.1907	0.1875	0.1967	0.2152	0.2491	0.2651	0.2639	0.1755	0.1754	0.1906	0,1986	0.1906	0.1754	0.1734	0.1876	0.2233	0.2367	0.2234	0.2261	0.2131	0.2184
_		EXT	0.1813	0.1765	0.1715	0.1670	0.1690	0.1724	0.1769	0.1956	0.2004	0.1973	0.2091	0.2224	0.2078	0.1843	0.1642	0.1524	0.1562	0.1848	0.2398	0.2521	0.1963	0.1779	0.1736	0.1738
2	Monthly	INMET	0.1899	0.1893	0.1824	0.1764	0.1716	0.1795	0.1969	0.2288	0.2452	0.2391	0.1720	0.1805	0.1989	0.2063	0.2002	0.1808	0.1699	0.1792	0.2115	0.2190	0.1816	0.1779	0.1746	0.1804
		EXT	0.1954	0.1869	0.1768	0.1711	0.1675	0.1715	0.1747	0.1863	0.1869	0.1967	0.1608	0.1378	0.1223	0.1175	0.1214	0.1434	0.1634	0.1852	0.2062	0.2143	0.2144	0.2087	0.2001	0.1939
	Hourly	INMET	0.1942	0.1878	0.1776	0.1716	0.1677	0.1710	0.1754	0.1869	0.1882	0.1972	0.1605	0.1386	0.1220	0.1169	0.1211	0.1425	0.1628	0.1851	0.2074	0.2159	0.2141	0.2100	0.1996	0.1938
	Monthly-	EXT	0.1643	0.1578	0.1548	0.1541	0.1534	0.1587	0.1597	0.1710	0.1744	0.1684	0.1503	0.1334	0.1211	0.1148	0.1160	0.1336	0.1453	0.1587	0.1653	0.1573	0.1587	0.1666	0.1646	0.1650
	Hourly	INMET	0.1627	0.1567	0.1557	0.1545	0.1519	0.1576	0.1610	0.1709	0.1739	0.1678	0.1500	0.1321	0.1218	0.1144	0.1172	0.1326	0.1467	0.1585	0.1663	0.1588	0.1586	0.1670	0.1664	0.1650
	Theoretica	EXT	0.1558	0.1580	0.1623	0.1659	0.1677	0.1715	0.1757	0.1831	0.1702	0.1411	0.1300	0.1377	0.1474	0.1539	0.1624	0.1662	0.1687	0.1701	0.1686	0.1698	0.1679	0.1676	0.1638	0.1606
	Curve	INMET	0.3368	0.3288	0.3262	0.3269	0.3258	0.3268	0.3156	0.2998	0.3060	0.3193	0.3316	0.3375	0.3397	0.3434	0.3512	0.3568	0.3620	0.3678	0.3688	0.3788	0.3847	0.3820	0.3672	0.3502
	Singular	EXT	0.1576	0.1597	0.1642	0.1666	0.1687	0.1723	0.1778	0.1854	0.1729	0.1438	0.1309	0.1341	0.1397	0.1458	0.1537	0.1580	0.1592	0.1605	0.1570	0.1622	0.1636	0.1654	0.1620	0.1605
		INMET	0.1657	0.1653	0.1697	0.1746	0.1755	0.1780	0.1756	0.1776	0.1716	0.1625	0.1553	0.1622	0.1652	0.1725	0.1790	0.1807	0.1790	0.1740	0.1658	0.1651	0.1731	0.1770	0.1744	0.1701
		EXT	0.1457	0.1508	0.1556	0.1596	0.1605	0.1649	0.1673	0.1719	0.1602	0.1403	0.1353	0.1381	0.1453	0.1497	0.1573	0.1606	0.1601	0.1604	0.1553	0.1567	0.1556	0.1543	0.1503	0.1482
3	Monthly	INMET	0,1539	0,1551	0,1602	0,1654	0,1658	0,1676	0,1651	0,1631	0,1570	0,1547	0,1532	0,1645	0,1676	0,1706	0,1727	0,1739	0,1712	0,1665	0,1592	0,1640	0,1736	0,1683	0,1621	0,1580
		EXT	0,1558	0,1582	0,1616	0,1653	0,1667	0,1704	0,1714	0,1712	0,1627	0,1390	0,1290	0,1332	0,1391	0,1455	0,1531	0,1557	0,1557	0,1548	0,1508	0,1535	0,1570	0,1596	0,1612	0,1605
	Houriy	INMET	0,1562	0,1581	0,1622	0,1647	0,1666	0,1701	0,1714	0,1723	0,1638	0,1391	0,1289	0,1326	0,1401	0,1453	0,1525	0,1550	0,1549	0,1545	0,1508	0,1543	0,1575	0,1605	0,1614	0,1600
	Monthly- Hourly	EXT	0,1422	0,1447	0,1498	0,1556	0,1545	0,1596	0,1566	0,1513	0,1423	0,1329	0,1260	0,1302	0,1355	0,1413	0,1473	0,1508	0,1510	0,1494	0,1437	0,1441	0,1449	0,1439	0,1425	0,1436
		INMET	0,1436	0,1448	0,1503	0,1561	0,1552	0,1582	0,1576	0,1499	0,1429	0,1329	0,1262	0,1303	0,1370	0,1410	0,1476	0,1501	0,1503	0,1495	0,1441	0,1453	0,1444	0,1439	0,1417	0,1438
	Theoretica	EXT	0,1582	0,1613	0,1655	0,1686	0,1727	0,1771	0,1879	0,1928	0,1638	0,1270	0,1235	0,1324	0,1432	0,1470	0,1535	0,1529	0,1571	0,1612	0,1645	0,1643	0,1570	0,1559	0,1547	0,1557
	Curve	INMET	0,2518	0,2406	0,2374	0,2348	0,2316	0,2310	0,2298	0,2342	0,2546	0,2856	0,3045	0,3102	0,3116	0,3111	0,3127	0,3139	0,3121	0,3081	0,2988	0,3017	0,3058	0,3035	0,2871	0,2678
	Cincular	EXT	0,1405	0,1410	0,1437	0,1453	0,1472	0,1503	0,1588	0,1632	0,1426	0,1274	0,1254	0,1270	0,1365	0,1403	0,1474	0,1495	0,1504	0,1494	0,1445	0,1477	0,1457	0,1483	0,1450	0,1419
	Siriyular	INMET	0,1380	0,1360	0,1384	0,1389	0,1383	0,1400	0,1456	0,1518	0,1446	0,1583	0,1559	0,1496	0,1478	0,1500	0,1552	0,1581	0,1587	0,1568	0,1499	0,1477	0,1476	0,1527	0,1495	0,1428
4	Manthly	EXT	0,1251	0,1271	0,1319	0,1349	0,1357	0,1381	0,1462	0,1490	0,1287	0,1275	0,1312	0,1339	0,1414	0,1458	0,1527	0,1538	0,1527	0,1494	0,1414	0,1422	0,1367	0,1360	0,1301	0,1268
4	Hourly	INMET	0,1257	0,1239	0,1258	0,1272	0,1267	0,1276	0,1318	0,1373	0,1301	0,1544	0,1557	0,1519	0,1516	0,1500	0,1529	0,1525	0,1543	0,1511	0,1470	0,1485	0,1476	0,1454	0,1371	0,1300
		EXT	0,1360	0,1347	0,1369	0,1372	0,1376	0,1381	0,1441	0,1482	0,1367	0,1216	0,1195	0,1229	0,1311	0,1355	0,1407	0,1417	0,1422	0,1420	0,1378	0,1396	0,1400	0,1440	0,1422	0,1393
	Tioutiy	INMET	0,1359	0,1345	0,1369	0,1376	0,1374	0,1383	0,1433	0,1478	0,1375	0,1223	0,1194	0,1234	0,1308	0,1349	0,1412	0,1406	0,1423	0,1422	0,1379	0,1401	0,1401	0,1444	0,1432	0,1397
	Monthly-	EXT	0,1169	0,1165	0,1194	0,1227	0,1205	0,1213	0,1255	0,1255	0,1187	0,1178	0,1162	0,1203	0,1273	0,1302	0,1351	0,1363	0,1371	0,1357	0,1284	0,1289	0,1257	0,1256	0,1216	0,1193
	Hourly	INMET	0,1170	0,1159	0,1205	0,1224	0,1200	0,1216	0,1252	0,1265	0,1183	0,1168	0,1154	0,1207	0,1271	0,1299	0,1357	0,1359	0,1375	0,1362	0,1288	0,1290	0,1253	0,1244	0,1218	0,1193
5	Theoretica	I EXT	0,2785	0,2799	0,2823	0,2850	0,2934	0,3019	0,3155	0,3085	0,2480	0,1740	0,1718	0,1869	0,1966	0,2023	0,2093	0,2177	0,2391	0,2772	0,2939	0,2866	0,2813	0,2689	0,2677	0,2706
	Curve	INMET	0,2036	0,1963	0,1951	0,1934	0,1902	0,1887	0,1941	0,2277	0,2922	0,3560	0,3822	0,3876	0,3869	0,3790	0,3668	0,3440	0,2932	0,2360	0,2177	0,2279	0,2408	0,2440	0,2301	0,2151
	Singular	EXT	0,1297	0,1320	0,1357	0,1360	0,1398	0,1425	0,1541	0,1636	0,1560	0,1761	0,1864	0,1956	0,2055	0,2115	0,2098	0,1959	0,1622	0,1356	0,1296	0,1271	0,1253	0,1234	0,1220	0,1245
		INMET	0,1112	0,1123	0,1113	0,1094	0,1090	0,1114	0,1244	0,1455	0,1667	0,2182	0,2197	0,2017	0,1907	0,1837	0,1789	0,1712	0,1615	0,1682	0,1646	0,1469	0,1248	0,1173	0,1132	0,1122
		EXT	0,1229	0,1256	0,1310	0,1336	0,1356	0,1388	0,1475	0,1505	0,1434	0,1703	0,1844	0,1931	0,2024	0,2092	0,2065	0,1951	0,1587	0,1330	0,1280	0,1263	0,1231	0,1190	0,1157	0,1168
	worniny	INMET	0,1068	0,1087	0,1086	0,1071	0,1073	0,1093	0,1183	0,1329	0,1527	0,2105	0,2134	0,1972	0,1886	0,1823	0,1760	0,1673	0,1559	0,1619	0,1608	0,1483	0,1287	0,1161	0,1091	0,1074
	Hourly	EXT	0,1077	0,1065	0,1084	0,1073	0,1066	0,1071	0,1190	0,1435	0,1551	0,1506	0,1467	0,1529	0,1588	0,1648	0,1674	0,1657	0,1508	0,1317	0,1201	0,1190	0,1170	0,1160	0,1118	0,1086
		INMET	0,1074	0,1067	0,1083	0,1070	0,1066	0,1071	0,1188	0,1438	0,1553	0,1499	0,1468	0,1535	0,1584	0,1649	0,1686	0,1652	0,1508	0,1321	0,1197	0,1187	0,1170	0,1164	0,1121	0,1084
	Monthly-	EXT	0,0999	0,1000	0,1028	0,1018	0,1007	0,1017	0,1065	0,1189	0,1296	0,1408	0,1419	0,1470	0,1523	0,1586	0,1601	0,1583	0,1427	0,1249	0,1139	0,1134	0,1100	0,1090	0,1040	0,1001
	Hourly	INMET	0,1005	0,1001	0,1026	0,1024	0,1002	0,1014	0,1066	0,1192	0,1298	0,1410	0,1417	0,1475	0,1542	0,1587	0,1590	0,1590	0,1436	0,1248	0,1138	0,1127	0,1104	0,1087	0,1042	0,1011



Figure 3.7 Dispersion graphs of the historic versus estimated data (Wind Farm 5).



Figure 3.8 Polygon graphs of the estimated and historic frequency data (Wind Farm 5).

3.6 Final Considerations

Due to the country's favourable conditions, the expanded use of renewable energy sources has led to the more significant insertion of wind power in Brazil. The greater exploitation of this energy source has posed challenges to the modeling and dimensioning of wind farms, because understanding the generation process is essential to the correct planning and operation of these facilities.

Wind speed is the main, but not the only, input variable for wind generation, so modeling the relationship between wind speed and power output is essential for adequate provisioning. The method adopted most often is the power curve, especially the cubic approach. It pertains to the class of deterministic and parametric methods, and its widespread use is due to its simple application combined with good results. However, this method does not replicate the variability of the relationship between wind speed and power output, since in practice, a single wind speed value can generate more than one power value. Other drawbacks of this use are the need to know the air density at the wind farm and turbine height and the latter's efficiency factor. These factors vary in time, and data are often absent. The theoretical power curve also is defined in regions other than where generation occurs, with the generation values being fixed and nonlinear, while in practice the regions where the wind passes are not well defined, and much less static.

All these limitations motivated this work and the development of a method that can better replicate the variability of wind generation in function only of the wind speed, without the need for other data to modeling, such as climate conditions and technical specifications of the turbines. Our proposed method uses clusterization, estimation of the PDF and simulation so that it can be classified as a probabilistic and nonparametric technique. Its main advantage is the need only for historical data on wind power and speed. And even with lack of speed data, this problem can be overcome using MERRA-2 data. The data themselves express the variations due to technical factors of the wind farm and local climate conditions. The disadvantage is the need for measured wind generation data. In the case of new or recently concluded wind farms, these data are unavailable or sparse. This may not reflect what will happen throughout an entire year or in subsequent periods.

To perform the tests in this study, we used information on five wind farms in Brazil's Northeast region (three along the coast and two inland). The data used were the historical active wind power values, and since there were no specific data on wind speed, these values were obtained from the MERRA-2 reanalysis database. The results achieved are promising. According to the four evaluation metrics (RMSE, MAE, MAPE and R²) applied, the performance was best with the monthlyhourly strategy, since it consistently presented the lowest values of RMSE, MAE and MAPE, and highest of R². In all the tests, the worst results were obtained with the traditional cubic power curve approach.

Observation of the dispersion and frequency polygon graphs revealed that the new power estimation technique, by single period, monthly, hourly and monthly-hourly,

managed to replicate the generation variability pattern in relation to the wind speed, to enable users to properly dimension the real future conditions of their undertakings and the nuances of this alternative power source. In particular, the dispersion graphs indicated that the greater the segregation of the data used for application of the technique was, the better the technique managed to replicate the variability of the wind speed data, thus improving the results, according to all evaluation metrics. The polygon frequency graphs also demonstrated that the proposed technique replicated the probability distribution of the historic power series. Hence, this modeling managed to capture the profile of the wind power data and the stochastic behavior of the source.

The need for measured wind speed data was one of the limitations of this study, but in their absence, we successfully used the MERRA-2 dataset as an excellent alternative. We employed two types of wind speed series from MERRA-2. The first was obtained from the steps of interpolation and extrapolation (EXT), while the second involved those two steps along with bias correction (INMET). Both types had good applicability in the four approaches considered. Whenever their performances were compared in the tests, they alternated in having the best result, and in some instances obtained the same result according to each evaluation metric. This was observed in all the tests, even when varying the evaluation metric (RMSE, MAE, MAPE and R²), the wind farms and the analysis details (general, per month and hour).

The wind speed time series of the INMET type had low speed values so that it could have been underestimated after the bias correction step with data from the INMET station. This low estimate could be seen in all wind farms, including those where bias correction was recommended (farms 3, 4 and 5). This aspect could be noted when estimating the power by the power curve method, where in some cases we did not see the complete formation of the power curve and there was no maximum generation specified for the turbine. However, this fact did not affect the new method proposed in this article, since all the power possibilities are inferred from the wind speeds that occur. So, the method overcomes this speed underestimation by working with historic occurrence data.

For future studies, we recommend evaluating this method with other renewable energy sources to test its efficacy, verify further clustering and PDF estimation approaches, and assess if there are better ways to segment the data for application of the new method. Considering a multivariate modeling approach, evaluate using other variables from MERRA-2 and wind speed to estimate wind generation.

3.7

Appendix A. Supplementary material

Supplementary material (tables, graphs and codes) to this contribution can be found online at

https://github.com/saulocustodio/Custodio_Cyrino_Macaira_Energy_2024_New Method_WindSpeed_Vs_WindPower.

4

Third Contribution: Application to obtain wind speed and wind generation data for Brazil.

4.1

Introduction

Brazil has a sizeable territorial extension that presents a diverse climate and different wind regimes. These are the primary renewable resources for wind generation, and their exploitation should be encouraged. To motivate financial incentives for the development of wind farms in Brazil, a correct measurement of wind generation over time is necessary for different points in the territory. For this, time series of wind speed and wind generation are needed.

To encourage and assist in obtaining this data more realistically and appropriately for the Brazilian territory, this thesis worked on obtaining and validating wind speed reanalysis data and creating time series and wind generation scenarios representing the local conditions and peculiarities. The idea was to develop an application to expand and facilitate the use of the methodologies proposed in the first two contributions and enhance the initial objective of the thesis.

The application was developed in Shiny, an R software package that enables the development of web applications.

4.2

Application

The developed application has three tabs, as shown in Figure 4.1. The first tab presents the application and the material used as a base, and its content is shown in Figure 4.1. The second tab aims to create a time series of wind speeds based on the methods used in this thesis's first contribution. The third tab is intended to create wind generation scenarios based on the methodology developed in the second contribution of the thesis.

oposals for the use of reanalysis bases for wind energy modeling in Braz	il
uthors	
ulo Custodio, Fernando Cyrino and Paula Maçaira	
pstract azil has always had its electrical matrix based mainly on renewable sources, specifically hydro. Over the years, this has diversified and demonstrated greater participation of wind sources. To better explore it, research aimed at modeling its behavior is essential. However, it is only sometimes that data wind speed and wind generation is available in quantity and the locations of interest. This data is necessary for identifying potential locations for stalling wind farms, improving the performance of existing ones, and stimulating research into forecasting and simulating wind generation, which are vus to help improve the planning and operation of the Brazilian electricity sector.	
the absence of wind speed data, an alternative is to use data from a reanalysis database. They provide long histories of data on climatic and mospheric variables for different parts of the world, free of charge. Therefore, the first contribution of this work focused on verrifying the presentativeness of wind speed data made available by MERRA-2 in Brazilina territory. Following literature recommendations, interpolation, trapolation, and bias correction techniques were used to improve the adequacy of the speeds provided by the reanalysis based on those that occur at energies that the territory of the world of the WIND SPEED tab, the time series of wind speed is available at any point in Brazilian territory after creasing the data coming from MERRA-2 using the techniques (interpolation, extrapolation, and bias correction) suggested in the first contribution of s work.	
e second contribution proposes modeling the relationship between wind speed and wind generation in a stochastic and nonparametric way based on storical data for both variables. For this purpose, clustering techniques using K-Means, estimation of density curves using KDE, and Monte Canto nutation were used. In the VIND POWER tab, it is possible to develop the relationship between speed and power for any location in Brazil by simply oviding a history of both variables or just wind generation; in the latter case, data from MERRA-2 is used to build the history of wind speed. It is suble to generate future generation scenarios by projecting wind speeds.	
apers	
January Control State	
per submitted in 10.02.24 to Energy: Ferreira, S. C. A; Cyrino Oliveira, F. L.; Maçaira, P. M. Joint modeling of wind speed and power via a nparametric approach. Energy. 2024.	
itHub	
odes can be accessed at https://github.com/paulamacaira/Custodio_Cyrino_Macaira_Energy_2022_Representativeness-MERRA2-Brazil.git ps://github.com/saulocustodio/Custodio_Cyrino_Macaira_Energy_2024_NewMethod_WindSpeed_Vs_WindPower.git	

Figure 4.1 Application home screen.

4.2.1

Wind Speed

The second tab of the "Wind Speed" application aims to provide a MERRA-2 wind speed time series suitable for the conditions of wind farms in Brazil. The methodology steps for obtaining the time series consisted of interpolation and extrapolation, as well as the possibility or not of using bias correction. Figure 4.2 shows the initial layout of the application in the "Wind Speed" tab.

To perform the interpolation, the wind farm's latitude and longitude information is required, which must be inserted in the fields shown in Figure 4.2, followed by pressing the Interpolate button that identifies the MERRA-2 grid point closest to the geographic location of the wind farm.

Home Wind Speed Win	d Power
Vind Speed Tim	e Series based on MERRA-2
This tab allows you to build histo conditions of the location of inter-	rical wind speed time series based on the MERRA-2 reanalysis. This is done through interpolation, extrapolation, and bias correction, which adapt MERRA-2 speeds to the est in Brazil.
The methods presented in this ta reanalysis data for Brazilian te	b are based on the paper. Ferreira, S. C. A.; Cyrino Oliveira, F. L.; Maçaira, P. M. Validation of the representativeness of wind speed time series obtained from rritory. v. 258. p. 124746. Energy. 2022. DOI: https://doi.org/10.1016/j.energy.2022.124746
Interpolation	
It uses the Nearest Neighbor tec in the Brazil quadrant.	inique, which consists of selecting the point on the MERRA-2 grid closest to the wind farm/turbine's geographic coordinates. This study used only the MERRA-2 coordinates
Geographic location of your wind	farm/turbine.
Latitude:*	
-15,25	
*Fram 7.65 (North) to -34.25 (South)	
Longitude:**	
-55,25	
**From -75.95 (West) to -33.40 (East)	
Interpolate	

Figure 4.2 The wind speed tab beginning.

Figure 4.3 shows what the application screen looks like after interpolation. Note that the map indicates the location of the wind farm, MERRA-2 grid point, and the nearest INMET station, which may have its data used in the bias correction step.

After identifying the historical basis for which the MERRA-2 grid point will be used, the period, time scale, and desired height of wind speed data measurements must be indicated. Figure 4.4 shows the fields where this information is included in the application.



Figure 4.3 Interpolation result.
Extrapolatio	on		
The MERRA-2 the Hellman pov	Database ver law e	provides wind spee quation to project th	d history at heights of 2, 10, and 50 meters in u and v components. This work builds the wind speed time series at heights of 10 and 50 meters and uses them i e new time series at the height of the turbine rotor responsible for wind generation.
- The MERRA-2	history a	vailable in this appl	cation is from 2000-01-01 to 2023-11-30.
- The MERRA-2	base use	ed provides informa	tion on its variables per hour. The average information within the time frame is calculated to provide time series at other scales.
Period - Start/End			
2021-01-01	to	2023-11-30	
emporal Scale:			
Hour		•	
Wind Turbine Rote	or Height		
100			

Figure 4.4 Extrapolation step.

The MERRA-2 base only provides wind speed data at 2, 10, and 50 meters, so the extrapolation stage occurs through Hellman's power law that adjusts wind speeds for the reported height. The bias correction procedure using data from the INMET meteorological station closest to the wind farm is only carried out if the user chooses the option "Yes" for the question "Do you want to perform bias correction with INMET data?" (Figure 4.5). If you choose to perform it, you must also choose the type of correction by average, the default being hourly, as it is the one that had the best performance in the tests carried out in the first contribution (Figure 4.6).

Blas Correction	
This step involves inserting local Brazilian characteristics into the stations, and the factor can be SINGLE (the same applied to the e The factors are calculated by the ratio of the average data from th	time series from MERRA-2. This is done by creating correction factors based on data measured at INMET (National Institute of Meteorology) intire series), MONTHLY (one for each month), HOURLY (one for each hour) and MONTHLY-HOURLY (one for each hour of each month), le INMET station to the average data from MERRA-2.
The INMET stations used are the automatic ones that provide he	surly information.
- The nearest neighbor technique is also used to select the INME	F station.
 To construct the factors, data from the INMET and MERRA-2 sta deactivated by INMET. 	tions from 2008-01-01 to 2023-11-30 were used, and in some cases, the period may be shorter due to the date the station was created or
- The literature recommends using bias correction if the INMET st	ation is at most 40 km from the wind farm or turbine.
Distance from the INMET Station to the Wind Farm/Turbine: 1	9.23 Km
Do you want to perform bias correction with INMET data?	
No No	
U TES	

Figure 4.5 Bias correction step.

No		
Yes		
e of Average Correction with INMET data:		
Single Period		
Monthly		
Hourly		
Monthly and Hourly		

Figure 4.6 Type of correction.

To apply the remaining steps of the methodology and visualize the wind speed time series that occurs in your wind farm, click the "Create Output" button, and to obtain the series in an Excel format file, simply click the "Download Output" button. Figure 4.7 shows the final layout that is exposed in the application. In addition to the graph of the evolution of the time series over time, the MERRA-2 base used to create the output is also presented and made available for download via the "Download Full Dataset" button.



Figure 4.7 Result from the wind speed tab.

4.2.2

Wind Power

The second functionality of the application, located in the 'Wind Power' tab, is modeling the relationship between wind speed and the wind power of your turbine/wind farm, which makes it possible to project wind generation and possible scenarios. For its modeling, historical wind speed and power data are necessary; if there is no speed data, the speeds provided by MERRA-2 are used. The methodology adopted is the same as that presented in the second contribution: clustering of speed bands, estimation of power density functions, and simulation of wind power.

Figure 4.8 shows the initial screen of the third application tab, where you can initially load a CSV file with historical wind power data, as shown in the example image, and inform the capacity of the turbine/wind farm. After inserting the file with the historical information, it will appear next to the example image for the user to check the data format. The application also allows you to adjust the CSV file's data separator and decimal places.

ind Power Time Series						
And Fower Time Series						
in this tab, the relationship between wind speed variables speed clustering, estimation of generation density curves,	and wind gener and generation	ation can be n simulation.	nodeled	based on the	history of the	ese data. This modeling is done probabilistically and nonparametrically through
The methods presented in this tab are based on the paper 2024.	Ferreira, S. C	. A.; Cyrino O	lliveira,	F. L.; Maçaira	, P. M. Join	t modeling of wind speed and power via a nonparametric approach. Energy.
Input Data						
To model the relationship between wind speed and wind g data provided by MERRA-2.	eneration, the f	ourly history o	f both va	ariables over t	ne same per	nod must be provided. If you still need wind speed history, we will use the historical
Historical Wind Power Data:"		The file to be	uploade	d must be in	CSV (comn	na-separated values) format and created as per the example below.
Upload No file selected						
	Exa	mple:				Your Data
Power can be supplied in any unit		A	В	C	D	
r anno san ao ampiros n'any area	1	Date	Hour	Power		
CSV file separator:	2	2013-01-01	1	2134.57		
Comma	4	2013-01-01	2	1991.97		
C Containa	5	2013-01-01	3	1630.45		
Semicolon	6	2013-01-01	4	1377.16		
Semicolon		2013-01-01	5	2134.58		
Semicolon Tab	7					
 Semicolon Tab Space 	7	2013-01-01	6	1901.39		
 Semicolon Tab Space 	7 8 9	2013-01-01 2013-01-01	6	1901.39 1991.98		
Semicolon Tab Space The character used to separate decimals:	7 8 9 10	2013-01-01 2013-01-01 2013-01-01 2013-01-01	6 7 8	1901.39 1991.98 1630.46 1377.17		
Semicolon Tab Space The character used to separate decimals: Comma	7 8 9 10 11	2013-01-01 2013-01-01 2013-01-01 2013-01-01	6 7 8 9	1901.39 1991.98 1630.46 1377.17		
Semicolon Tab Space The character used to separate decimals: Comma Point	7 8 9 10 11	2013-01-01 2013-01-01 2013-01-01 2013-01-01	6 7 8 9	1901.39 1991.98 1630.46 1377.17		
Semicolon Tab Space The character used to separate decimals: Comma Point	7 8 9 10 11	2013-01-01 2013-01-01 2013-01-01 2013-01-01 2013-01-01	6789111111	1901.39 1991.98 1630.46 1377.17		
Semicolon Tab Space The character used to separate decimals: Comma Point	7 8 9 10	2013-01-01 2013-01-01 2013-01-01 2013-01-01 2013-01-01 2013-01-01 2013-01-01 2013-01-01 2013-01-01	6 7 8 9 11 12 19	1901.39 1991.98 1630.46 1377.17		
Semicolon Tab Space The character used to separate decimals: Comma Point Wind farm capacity:**	7 8 9 100 111 21 21	2013-01-01 2013-01-01 2013-01-01 2013-01-01 2013-01-01 2013-01-01 2013-01-01	6 7 8 9 11 11 20	1901.39 1991.98 1630.46 1377.17 1551.55 1551.55 1551.55 1377.19 2134.61		
Semicolon Tab Space The character used to separate decimals: Comma Point Wind farm capacity:** 0	7 8 9 10 11 11 15 12 21 22 23	2013 01 01 2013 01 01	6 7 8 9 11 19 20 21	1901.39 1991.98 1630.46 1377.17 1971.91 1377.19 2134.61 1630.49		
Semicolon Tab Space The character used to separate decimals: Comma Point Wind farm capacity:** 0	7 8 9 10 11 11 21 22 23 24	2013 01 01 2013 01 01	6 7 8 9 11 10 19 20 21 22	1901.39 1991.98 1630.46 1377.17 1377.17 1377.19 2133.61 1630.49 1377.20		
Semicolon Tab Space The character used to separate decimals: Comma Point Wind farm capacity:** 0 **Capacity must be provided in the same unit as power.	7 8 9 100 11 11 22 23 24 25	2013 01 01 2013 01 01	6 7 8 9 11 19 20 21 22 23	1901.39 1991.98 1630.46 1377.17 1377.19 1377.19 2134.61 1630.49 1377.20 934.35		

Figure 4.8 The wind power tab beginning.

The user's next step is to inform whether there is a history of wind speeds that occurred at the turbine/wind farm, and the history must be from the same period as the wind energy supplied. Suppose the answer is positive in the application. In that case, a screen will appear, as shown in Figure 4.9, where the user must repeat the same procedure to include generation data, except for the wind farm's capacity. You will see a screen like Figure 4.10 if the answer is negative. It repeats the same steps in the "wind speed" tab to build the wind speed history. After the user provides the geographic coordinates of the wind farm, informs the height of the turbine rotor and whether bias correction will be used with data from the INMET station with its respective modality, he must click the "Create Output" button to create the wind speed time series of the same dimension and period as the generation data.

) No						
Yes						
listorical Wind Speed Data at Turbine Rotor Height:*		The file to be u	uploaded	i must be in (CSV (comm	na-separated values) format and created as per the example below.
Upload No file selected						
	Exa	imple:				Your Data:
Speed can be supplied in any unit.		A	Б	C	D	
	1	Date	Hour	Speed		
SV file separator:	2	2013-01-01	0	6.45		
Comma	3	2013-01-01	1	10.59		
Comma	5	2013-01-01	3	11.31		
semicolon	6	2013-01-01	4	9.22		
) Tab	7	2013-01-01	5	12.03		
Space	8	2013-01-01	6	9.68		
	9	2013-01-01	7	12.04		
De character used to separate decimale:	12					
nie onaroden used to separate decilitais.	13	1010 01 01				
) comma	12	2010 01 01	10	44.70		
Point		1010 01 01	10	5.00		
	21	2013-01-01	19	10.19		
	22	2013-01-01	20	9.02		
	23	2013-01-01	21	12.04		
	2.4	2013-01-01	22	9.22		
	- 64					

Figure 4.9 Input wind speed data.



Figure 4.10 Building the wind speed data.

With the speed and generation of data, the methodology is feasible. The user must select the segmentation to be adopted on the data to apply the modeling and press the "Apply Method" button. The data segmentation pattern is by month and hour, which was the best result in the work presented in the second contribution. This step in the application is represented in Figure 4.11, and in Figure 4.12 are the results achieved by applying clustering methods, density curve estimation, and simulation.

Application of Methods	
Aodeling is carried out through three step	S.
Clustering of wind speeds by K-means;	
Estimation of the probability density func	tion of wind generations in each speed range by KDE (Kernel density estimation),
Simulation of wind generation using Mon	te Carlo
INS MODELING CAN be done SINGLE PERI NONTHLY-HOURLY (segregation by mon	OD (without any data segregation), MONTHLY (data segregation by monin to apply the methods), HOURLY (data separated by hour to carry out the modeling), or th and time to model).
Ins modeling can be done sink-LE PERI IONTHLY-HOURLY (segregation by mon	CD (wimout any data segregation), MONTHLY (data segregation by monin to apply the methods), HOURLY (data separated by nour to carry out the modeling), or th and time to model).
Inite modeling can be done sinvate Pera MONTHLY-HOURLY (segregation by mon Acthodology to use:) Single Period	CD (wimout any daria segregation), MONTHLY (data segregation by monin to apply the methods), HOURLY (data separated by nour to carry out the modeling), or th and time to model).
Inite modeling can be dorte SINGLE PERI XONTHLY-HOURLY (segregation by mon Methodology to use:) Single Period) Monthy Hourthy	CD (winnout any daria segregation), MONTHLY (data segregation by monin to apply the methods), HOURLY (data separated by nour to carry out the modeling), or th and time to model).

Figure 4.11 Application of the methodology.

In Figure 4.12, you can see the number of clusters adopted in each data segmentation and how the data was dispersed by cluster, how the wind power density curve was estimated for each cluster and the comparison of the historical generation with generation estimates simulated through scatter plots and frequency polygons.



Figure 4.12 Result of modeling the relationship between wind speed and wind generation.

To project wind generation and its possible scenarios, it is necessary to provide a wind speed projection that will be applied in the modeling developed. Figure 4.13 shows the location in the application where the wind speed projection must be inserted. As in Figure 4.9, the application shows an example of how the CSV file should be created to be inserted and allows it to be viewed after insertion, in addition to adjusting the formatting of data separation and decimal places. Once the data has been entered correctly, define the generation scenario number you want to create and click the "Create!" button. After the end of the process, a line graph will be presented with the scenarios and the future estimate of wind generation, which is the average of the scenarios (Figure 4.14). The app also allows you to

download this information; select the estimated items you want and click "Download!" an Excel file with these results will be downloaded.

upload. Note selected upload. not muse the provoked in this exame unit as the speed atom to the provoked in th	jection o	of wind speed at the height of the turbine	1	The file to be	uploaded	l must be in (CSV (comr	ma-separated values) format and created as per the example below.
A B C D bit spearator: 2013-01-01 0 6.45 2013-01-01 2 10.58 5 2013-01-01 3 11.01 6 2013-01-01 5 12.02 8 2013-01-01 5 12.02 9 2013-01-01 5 12.02 8 2013-01-01 5 12.02 8 2013-01-01 5 12.02 9 2013-01-01 5 12.02 8 2013-01-01 5 12.02 8 2013-01-01 5 12.02 9 2013-01-01 7 12.04 10 2013-01-01 7 12.04 11 2013-01-01 10 19 21 2013-01-01 2 3.02 22 2013-01-01 2 3.02 22 2013-01-01 2 3.02 22 2013-01-01 2 3.02 22 2013-01-02 0 11.45 00 11.45	pload	No file selected	Exa	mpie:				Your Data
independent muit the provided in the same unit as the speed used in being: i Date Hour Speed i i Date Hour Speed Date Date Hour Speed Date Date Hour Speed Date				A	В	с	D	
seeing 2 2013-01-01 0 6.45 3 2013-01-01 0 6.45 3 2013-01-01 1 8.79 4 2013-01-01 1 8.79 4 2013-01-01 1 8.79 4 2013-01-01 3 11.31 6 2013-01-01 5 12.00 8 2013-01-01 5 12.00 8 2013-01-01 6 9.66 9 2013-01-01 5 12.00 8 2013-01-01 5 12.00 8 2013-01-01 6 9.66 9 2013-01-01 7 2014-01 2013-01-01 7 2013-01-01 9.02 21 2013-01-01 12 12.04 22 2013-01-01 12 12.04 24 2013-01-02 0 11.45 10	nd annoad -	must be remained in the same and as the second and in	1	Date	Hour	Speed		
a 2013-01-01 1 8.79 mma 4 2013-01-01 2 10.58 score 5 2013-01-01 4 9.22 race 7 2013-01-01 6 9.66 9 2013-01-01 6 9.66 9 9 2013-01-01 7 12.04 1 1 1 1 1 1 1 acce 1 2013-01-01 6 9.66 9 2013-01-01 6 9.66 9 2013-01-01 7 12.04 1	ng.	most are provided in the barrier and an the opend diffid in	2	2013-01-0	L 0	6.45		
4 2013-01-01 2 10.58 mma 5 2013-01-01 3 11.31 6 2013-01-01 4 9.22 r 2013-01-01 5 12.00 acce 9 2013-01-01 6 9.60 9 2013-01-01 6 9.60 9 9 2013-01-01 7 12.04 12.04 10 11 19 10.19 12.12 11 22 2013-01-01 19 10.19 12 2013-01-01 19 10.19 12.04 12 2013-01-01 12 12.04 12.04 12 2013-01-01 12 12.04 12.04 12 2013-01-01 22 9.22 22.2 23 2013-01-01 21 12.04 14 2013-01-01 22 9.20 24 2013-01-01 22 9.20 25 2013-01-02 0 11.45			з	2013-01-0	1 1	8.79		
mma 5 2013-01-01 3 11.31 nicolon 6 2013-01-01 4 9.22 nicolon 7 2013-01-01 5 12.00 8 2013-01-01 6 9.66 9 nicolon 9 2013-01-01 6 9.66 stracter used to separate decimals: 1 1 1 1.01 nma 21 2013-01-01 19 10.19 nma 21 2013-01-01 19 10.19 nma 21 2013-01-01 29 9.02 22 2013-01-01 29 9.02 22 23 2013-01-01 20 9.02 24 2013-01-01 21 9.02 25 2013-01-01 21 9.02 26 2013-01-02 0 11.45	e se	parator	-4	2013-01-0	1 2	10.58		
acco 6 2012-01-01 4 9.22 acco 8 2013-01-01 6 9.68 9 2013-01-01 7 12.04 acco 7 12.04 7 12.04 acco 7 12.04 7 12.04 acco 7 12.04 7 12.04 acco 7 2013-01-01 7 12.04 acco 7 2013-01-01 19 10.19 acco 7 2013-01-01 19 10.19 acco 22 2013-01-01 21 12.04 23 2013-01-01 21 12.04 24 24 2013-01-01 21 12.03 26 25 2013-01-02 0 11.45 12.03 acco acco acco acco acco action of Wind Power Time Series: acco acco acco action of Wind Power Time Series: acco acco acco action of Wind Power Time Series: acco acco acco	mana		5	2013-01-0	1 3	11.31		
micolon 7 2013-01-01 5 12.03 b 2013-01-01 6 9.68 9 2013-01-01 7 12.04 haracter used to separate decimals: 7 2013-01-01 19 10.1	ninja		6	2013-01-0	4	9.22		
b 8 2013-01-01 6 9.68 brace 9 2013-01-01 7 12.04 brance 1 1 10.19 10.19 comma 2 2013-01-01 20 9.02 comma 2 2013-01-01 19 10.19 comma 2 2013-01-01 20 9.02 comma 2 2013-01-01 20 12.04 comma 2 2013-01-01 20 11.45 comma 2 2013-01-02 0 11.45 comma 2 2013-01-02 0 11.45 comma 2 2013-01-02 0 11.45 comma 2 2 2 2 </td <td>micolo</td> <td>n</td> <td>7</td> <td>2013-01-0.</td> <td>5</td> <td>12.03</td> <td></td> <th></th>	micolo	n	7	2013-01-0.	5	12.03		
9 2013/01-01 7 12.04 character used to separate decimals: 7 2013/01-01 7 12.04 comma 1 2013/01-01 1 9 0.02 21 2013/01-01 20 9.02 23 2013/01-01 20 22 2013/01-01 20 9.02 23 2013/01-01 20 9.02 23 2013/01-01 20 9.02 23 2013/01-01 20 9.02 24 2013/01-01 21 12.04 20 9.02 23 2013/01-01 21 9.02 25 2013/01-01 21 12.03 26 2013/01-02 0 11.45 section of Wind Power Time Series: 2013/01-02 0 11.45 11.45 11.45 section of Wind Power Time Series: 2013/01-02 0 11.45 11.45 11.45 section of Wind Power Time Series: 2013/01-02 0 11.45 11.45 11.45 section of wind power Time Series: 2013/01-02 11.45 11.45 11.45 11.45 11.45<	ab		8	2013-01-0	6	9.68		
aracter used to separate decimals: 1	ce		9	2013-01-0	7	12.04		
aracter used to separate decimals: 1			12					
staracter used to separate decimals: 10 2010-01-01 19 10.19 21 2010-01-01 20 9.02 23 2010-01-01 24 2010-01-01 24 2010-01-01 22 9.22 2010-01-01 24 2010-01-01 24 2010-01-01 24 2010-01-01 24 2010-01-01 24 2010-01-01 24 2010-01-01 26 2010-01-01 26 2010-01-01 26 2010-01-01 26 2010-01-01 26 2010-01-01 27 9.02 26 2010-01-01 20 9.02 26 2010-01-01 20 9.02 26 2010-01-01 20 9.02 26 2010-01-01 20 9.02 26 2010-01-01 20 9.02 26 2010-01-02 0 11.45 11			12			VILL		
and columna int. columna int. 2013-01-01 22 2013-01-01 22 2013-01-01 22 2013-01-01 22 2013-01-01 24 2013-01-01 25 2013-01-01 26 2013-01-02 21 2013-01-02 2013-01-02 0 11.45	harac	ter used to separate decimals:	15	1010 01 01	17	10.01		
21 2013-01-01 1 10.15 22 2013-01-01 20 9.02 23 2013-01-01 21 12.04 24 2013-01-01 22 9.22 25 2013-01-02 0 11.45 etata you want to download: ata you want to download: ata set set set peed time seties eneration scenarios ata you want to download:	nma				40	40.40		
12 2013-01-01 12 13.04 23 2013-01-01 22 9.22 24 2013-01-01 22 9.22 25 2013-01-01 23 12.03 26 2013-01-02 0 11.45	nt		21	2013-01-0.	1 19	10.19		
add a base of the series generation scenarios : add a base of the series generation scenarios :			22	2013-01-0	20	12.04		
ated ated ated generation scenarios:			28	2013-01-0	22	9.22		
26 2013-01-02 0 11.45 ber of Scenarios:			25	2013-01-0	23	12.03		
e data you want to download: speed time series generation scenarios:			26	2013-01-0	0	11.45		
the data you want to download: d speed time series d generation scenarios	ection	nt Wind Power Time Series:						
Inter of Scenarios: Internet Scenarios: It the data you want to download: It data you want to download: It speed time series It generation scenarios:	Jection (string tonel lille delles.						
calet	nber of	Scenarios:						
teated he data you want to download: speed time series generation scenarios	0							
I the data you want to download: Id speed time series Id generation scenarios	reatel							
nd speed time series nd generation scenarios	the data	uau unet ta daum land:						
nd spectrum mino denies 1d generation scenarios	d enaced	tima cariae						
o generation scenarios	a speed	ume senes						
	d genera	ation scenarios						
	nadl							

Figure 4.13 Final screen with scenarios and average estimated generation time series.



Figure 4.14 Final screen with scenarios and average estimated generation time series.

4.3

Conclusion

The application was developed without any commercial purpose, solely to facilitate and encourage studies on wind generation in Brazil by companies, entities, and researchers. It also facilitates and expands access to MERRA-2 wind speed reanalysis data, as users often present some barriers to extracting and manipulating this data directly from the MERRA-2 base. In addition, it presents the advantage of providing processed data with a focus on the reality of the Brazilian territory.

It provides easy access to the methodologies created in this thesis, enabling researchers to replicate these studies and apply them in their studies, thus enabling comparative work on methods and the continuation of studies.

5 Summary of contributions and avenues for future research

This thesis comprised three contributions involving the promotion of the exploration of wind generation in Brazil, which are found in Chapters 2, 3, and 4. The first study aimed to overcome the need for measured wind speed data for the Brazilian territory because it is the primary resource for estimating wind generation potential. As an alternative, reanalysis data was suggested; thus, the first work validated the MERRA-2 wind speed time series representativeness for points and areas in Brazil. Verifying that this adequacy increases as the MERRA-2 time series is used on larger temporal scales and to represent occurrences in larger areas.

The second work sought to develop a model of the relationship between wind speed and output power in wind turbines. The premises of this modeling were that it should be adaptive to different wind regimes and present only the wind speed variable as input to have a robust but simplified model. The solution created met expectations with a non-parametric and non-deterministic model, which only requires historical wind speed and wind generation.

The third contribution consisted of developing an application to make the methodologies developed in previous works available. The web application was created to facilitate and democratize access to tools that help entrepreneurs, institutions, and researchers. It helps obtain wind speed data and provides wind generation scenarios for locations of interest in the Brazilian territory. In addition to enabling researchers to develop their work and compare their results with those of the methodologies of this thesis.

With all these contributions related to wind generation, it is expected that the products of this thesis can serve as an alternative to feed the Brazilian energy optimization models, NEWAVE, DECOMP, and DESSEM, that encourage more studies related to wind energy and imply greater efficient resource exploitation.

In terms of future work, it is hoped that the thesis will encourage other researchers to develop similar studies for other sources of renewable resources. Develop a comparative study between the ERA5 and MERRA-2 reanalysis bases about the quality of representativeness of wind speed data and others in the Brazilian territory at different temporal and spatial scales to verify whether the same base is better in all conditions and locations and, if not, whether the shortest distance from the geographic coordinates of the measured data and the grid point of the reanalysis base leads to the best result. Investigar o uso de métodos não lineares e/ou não gaussianos para melhorar esta proposta e prever e simular dados de geração a partir da velocidade do vento ou de outras fontes.

References

AL-DUAIS, F. S.; AL-SHARPI, R. S. A unique Markov chain Monte Carlo method for forecasting wind power utilizing time series model. **Alexandria Engineering Journal**, v. 74, p. 51–63, 1 jul. 2023.

AL-QURAAN, A. et al. Power curve modelling of wind turbines- A comparison study. **IET Renewable Power Generation**, v. 16, n. 2, p. 362–374, 22 feb. 2022.

AYODELE, T. R.; OGUNJUYIGBE, A. S. O. Wind energy resource, wind energy conversion system modelling and integration: a survey. **International Journal of Sustainable Energy**, v. 34, n. 10, p. 657–671, 26 nov. 2015.

BOSCH, J.; STAFFELL, I.; HAWKES, A. D. Temporally explicit and spatially resolved global offshore wind energy potentials. **Energy**, v. 163, p. 766–781, 2018.

BRUNE, S.; KELLER, J. D.; WAHL, S. Evaluation of wind speed estimates in reanalyses for wind energy applications. **Advances in Science and Research**, v. 18, p. 115–126, 2021.

CANNON, D. J. et al. Using reanalysis data to quantify extreme wind power generation statistics: A 33 year case study in Great Britain. **Renewable Energy**, v. 75, p. 767–778, 2015.

CARRILLO, C. et al. Review of power curve modelling for wind turbines. **Renewable and Sustainable Energy Reviews**, v. 21, p. 572–581, may 2013.

CARTA, J. A.; RAMÍREZ, P.; VELÁZQUEZ, S. A review of wind speed probability distributions used in wind energy analysis. **Renewable and Sustainable Energy Reviews**, v. 13, n. 5, p. 933–955, jun. 2009.

CELEBI, M. E.; KINGRAVI, H. A.; VELA, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. **Expert Systems with Applications**, v. 40, n. 1, p. 200–210, jan. 2013.

CRADDEN, L. C. et al. A 34-year simulation of wind generation potential for Ireland and the impact of large-scale atmospheric pressure patterns. **Renewable Energy**, v. 106, p. 165–176, 2017.

DE AQUINO FERREIRA, S. C.; CYRINO OLIVEIRA, F. L.; MAÇAIRA, P. M. Validation of the representativeness of wind speed time series obtained from reanalysis data for Brazilian territory. **Energy**, v. 258, 1 nov. 2022.

DE ASSIS TAVARES, L. F. et al. Assessment of the offshore wind technical potential for the Brazilian Southeast and South regions. **Energy**, v. 196, p. 117097, apr. 2020.

DOS SANTOS, F. S. et al. Brazilian wind energy generation potential using mixtures of Weibull distributions. **Renewable and Sustainable Energy Reviews**, v. 189, p. 113990, jan. 2024.

DUCA, V. E. L. A.; FONSECA, T. C. O.; CYRINO OLIVEIRA, F. L. Joint modelling wind speed and power via Bayesian Dynamical models. **Energy**, v. 247, p. 123431, may 2022.

DUCA, V. E. L. A.; FONSECA, T. C. O.; CYRINO OLIVEIRA, F. L. An overview of non-Gaussian state-space models for wind speed data. **Energy**, v. 266, 1 mar. 2023.

EPE. **Empresa de Pesquisa Energética**. Available at: ">https://www.epe.gov.br/pt/abcdenergia/matriz-energetica-e-eletrica#ELETRICA/>. (accessed 9 aug 2023).

ESTEVES, G. et al. Improvements in the Current Brazil's Energy Dispatch Optimization: Load Forecast and Wind Power. Proceedings of the 8th International Conference on Operations Research and Enterprise Systems. Anais...SCITEPRESS - Science and Technology Publications, 2019.

FANG, D.; WANG, J. A novel application of artificial neural network for wind speed estimation. **International Journal of Sustainable Energy**, v. 36, n. 5, p. 415–429, 28 may 2017.

FREITAS, B. B. et al. Cenários para a matriz de geração de eletricidade do Ceará em 2050. **Ciência e Natura**, v. 42, p. e63, 26 may 2020.

GELARO, R. et al. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). **Journal of Climate**, v. 30, n. 14, p. 5419–5454, 2017.

GES DISC. Goddard earth sciences data and information services center. Available at: https://disc.gsfc.nasa.gov/>. (accessed 20 may 2021).

GILL, S.; STEPHEN, B.; GALLOWAY, S. Wind Turbine Condition Assessment Through Power Curve Copula Modeling. **IEEE Transactions on Sustainable Energy**, v. 3, n. 1, p. 94–101, jan. 2012.

GIORSETTO, P.; UTSUROGI, K. Development of a New Procedure for Reliability Modeling of Wind Turbine Generators. **IEEE Transactions on Power Apparatus and Systems**, v. PAS-102, n. 1, p. 134–143, jan. 1983.

GMAO. Global Modeling and Assimilation Office. Modern-Era Retrospective analysis for research and applications. National Aeronatuics and Space Administration. Available at: <https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/>. (accessed 21 may 2021). GONZÁLEZ-APARICIO, I. et al. Simulating European wind power generation applying statistical downscaling to reanalysis data. **Applied Energy**, v. 199, p. 155–168, 2017.

GRUBER, K. et al. Assessing the Global Wind Atlas and local measurements for bias correction of wind power generation simulated from MERRA-2 in Brazil. **Energy**, v. 189, p. 1–13, 2019.

GRUBER, K. et al. Towards global validation of wind power simulations: A multi-country assessment of wind power simulation from MERRA-2 and ERA-5 reanalyses bias-corrected with the Global Wind Atlas. **Energy**, p. 121520, 2021.

GRUBER, K.; SCHMIDT, J. Bias-correcting simulated wind power in Austria and in Brazil from the ERA-5 reanalysis data set with the DTU Wind Atlas. p. 1–15, 2019.

GUALTIERI, G. Reliability of era5 reanalysis data for wind resource assessment: A comparison against tall towers. **Energies**, v. 14, n. 14, 2021.

GWA. **Global Wind Atlas**. Available at: https://globalwindatlas.info/area/Brazil. (accessed 21 may 2021).

HAN, S.; YANG, Y.; LIU, Y. **The Comparison of BP Network and RBF Network in Wind Power Prediction Application**. 2007 Second International Conference on Bio-Inspired Computing: Theories and Applications. **Anais**...IEEE, set. 2007.

HAYES, L.; STOCKS, M.; BLAKERS, A. Accurate long-term power generation model for offshore wind farms in Europe using ERA5 reanalysis. **Energy**, v. 229, p. 120603, 2021.

INMET. Governo Federal Brasileiro. Instituto Nacional de Meteorologia. Available at: https://portal.inmet.gov.br/>. (accessed 21 may 2021).

IPAKCHI, A.; ALBUYEH, F. Grid of the future. **IEEE Power and Energy Magazine**, v. 7, n. 2, p. 52–62, mar. 2009.

IRENA. International Renewable Energy Agency. Available at: https://www.irena.org/Data/View-data-by-topic/Capacity-and-Generation/Country-Rankings. (accessed 13 aug 2023).

JIANG, Y. et al. Evaluation of hrcldas and era5 datasets for near-surface wind over hainan island and south china sea. **Atmosphere**, v. 12, n. 6, 2021.

JOURDIER, B. Evaluation of ERA5, MERRA-2, COSMO-REA6, NEWA and AROME to simulate wind power production over France. **Advances in Science and Research**, v. 17, p. 63–77, 2020.

JUNG, C.; SCHINDLER, D. Wind speed distribution selection – A review of recent development and progress. **Renewable and Sustainable Energy Reviews**, v. 114, p. 109290, oct. 2019.

KHATIBI, A.; KRAUTER, S. Validation and performance of satellite meteorological dataset merra-2 for solar and wind applications. **Energies**, v. 14, n. 4, 2021.

KHODABUX, K. et al. **Mathematical approach to modelling sigmoidal power curves for Wind Energy Conversion Systems**. 2022 7th International Conference on Environment Friendly Energies and Applications (EFEA). **Anais**...IEEE, 14 dez. 2022.

KUSIAK, A.; ZHENG, H.; SONG, Z. Models for monitoring wind farm power. **Renewable Energy**, v. 34, n. 3, p. 583–590, mar. 2009.

LÁZARO, R.; YÜRÜŞEN, N. Y.; MELERO, J. J. Wind Turbine Power Curve Modelling using Gaussian Mixture Copula, ANN Regressive and BANN. Journal of Physics: Conference Series. Anais...Institute of Physics, 2 jun. 2022.

LI, H. et al. A multi-step ahead point-interval forecasting system for hourly PM2.5 concentrations based on multivariate decomposition and kernel density estimation. **Expert Systems with Applications**, v. 226, p. 120140, sep. 2023.

LINS, D. R. et al. Comparison of the performance of different wind speed distribution models applied to onshore and offshore wind speed data in the Northeast Brazil. **Energy**, v. 278, 1 sep. 2023.

LYDIA, M. et al. A comprehensive review on wind turbine power curve modeling techniques. **Renewable and Sustainable Energy Reviews**, v. 30, p. 452–460, feb. 2014.

MAÇAIRA, P. M. et al. Including wind power generation in Brazil's long-term optimization model for energy planning. **Energies**, v. 12, n. 5, 2 mar. 2019.

MACQUEEN, JAMES. Some methods for classification and analysis of multivariate observations. Em: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. [s.l: s.n.]. v. 1p. 281–297. 1967.

MARVUGLIA, A.; MESSINEO, A. Monitoring of wind farms' power curves using machine learning techniques. **Applied Energy**, v. 98, p. 574–583, oct. 2012.

MATHEW, S. **Wind Energy**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

MILAD, S. et al. Wind Turbine Modeling Using Wind Speed Measurement Data. 2023 22nd International Symposium on Power Electronics (Ee). Anais...IEEE, 25 out. 2023.

MOLINA, M. O.; GUTIÉRREZ, C.; SÁNCHEZ, E. Comparison of ERA5 surface wind speed climatologies over Europe with observations from the HadISD dataset. **International Journal of Climatology**, v. 41, n. 10, p. 4864–4878, 2021.

MURALIDHARAN, S. et al. A Comparative Study of Machine Learning Techniques for Wind Turbine Performance Prediction. **E3S Web of Conferences**, v. 387, p. 04011, 15 may 2023.

MURCIA, J. P. et al. Validation of European-scale simulated wind speed and wind generation time series. **Applied Energy**, v. 305, n. February 2021, p. 117794, 2022.

NEFABAS, K. L. et al. Modeling of ethiopian wind power production using era5 reanalysis data. **Energies**, v. 14, n. 9, p. 1–17, 2021.

OLAUSON, J. ERA5: The new champion of wind power modelling? **Renewable Energy**, v. 126, p. 322–331, 2018.

OLAUSON, J.; BERGKVIST, M. Modelling the Swedish wind power production using MERRA reanalysis data. **Renewable Energy**, v. 76, p. 717–725, 2015.

ONS. **Histórico da Operação**. Available at: http://www.ons.org.br/paginas/resultados-da-operacao/historico-da-operacao. (accessed 21 may 2021).

PERINI DE SOUZA, N. B. et al. Long-range correlations of the wind speed in a northeast region of Brazil. **Energy**, v. 243, 2022a.

PERINI DE SOUZA, N. B. et al. Wind mapping using the mesoscale WRF model in a tropical region of Brazil. **Energy**, v. 240, p. 122491, 2022b.

PIERROT EI, M. **The Wind Power**. Available at: https://www.thewindpower.net/. (accessed 25 may 2023).

PRYOR, S. C.; LETSON, F. W.; BARTHELMIE, R. J. Variability in wind energy generation across the contiguous United States. **Journal of Applied Meteorology and Climatology**, v. 59, n. 12, p. 2021–2039, 2020.

R CORE TEAM. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2018. Available at: https://www.r-project.org/.

RABBANI, R.; ZEESHAN, M. 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. **Renewable Energy**, v. 154, p. 1240–1251, 2020.

REN, G. et al. Characterization of wind resource in China from a new perspective. **Energy**, v. 167, p. 994–1010, 2019.

REZVANI, A. et al. Intelligent hybrid power generation system using new hybrid fuzzy-neural for photovoltaic system and RBFNSM for wind turbine in the grid connected mode. **Frontiers in Energy**, v. 13, n. 1, p. 131–148, 2 mar. 2019.

RYBERG, D. S. et al. The future of European onshore wind energy potential: Detailed distribution and simulation of advanced turbine designs. **Energy**, v. 182, p. 1222–1238, 2019.

SHERIDAN, L. M. et al. Validation of wind resource and energy production simulations for small wind turbines in the United States. **Wind Energy Science**, v. 7, n. 2, p. 659–676, 2022.

SHETTY, R. P.; SATHYABHAMA, A.; PAI, P. S. Comparison of modeling methods for wind power prediction: a critical study. **Frontiers in Energy**, v. 14, n. 2, p. 347–358, 1 jun. 2020.

SHOKRZADEH, S.; JAFARI JOZANI, M.; BIBEAU, E. Wind Turbine Power Curve Modeling Using Advanced Parametric and Nonparametric Methods. **IEEE Transactions on Sustainable Energy**, v. 5, n. 4, p. 1262–1269, oct. 2014.

SIGA ANEEL. Sistema de Informações de Geração da Agência Nacional de Energia Elétrica. Available at: <https://dados.gov.br/dataset/siga-sistema-de-informacoes-de-geracaodaaneel>. (accessed 2 may 2022).

SIGEL ANEEL. Sistema de Informações Georreferenciadas do Setor Elétrico da Agência Nacional de Energia Elétrica. Available at: https://sigel.aneel.gov.br/portal/home/>. (accessed 9 apr. 2023).

SILVA, G. R. Características de vento da região nordeste: Análise, Modelagem e Aplicações para Projetos de Centrais Eólicas. Dissertação (Mestrado em Engenharia Mecânica) – Pós-Graduação em Engenharia Mecânica, Universidade Federal de Pernambuco. Pernambuco. 2003.

SOHONI, V.; GUPTA, S. C.; NEMA, R. K. A Critical Review on Wind Turbine Power Curve Modelling Techniques and Their Applications in Wind Based Energy Systems. **Journal of Energy**, v. 2016, p. 1–18, 2016.

STAFFELL, I.; GREEN, R. How does wind farm performance decline with age? **Renewable Energy**, v. 66, p. 775–786, 2014.

STAFFELL, I.; PFENNINGER, S. Using bias-corrected reanalysis to simulate current and future wind power output. **Energy**, v. 114, p. 1224–1239, 2016.

THAPAR, V.; AGNIHOTRI, G.; SETHI, V. K. Critical analysis of methods for mathematical modelling of wind turbines. **Renewable Energy**, v. 36, n. 11, p. 3166–3177, nov. 2011.

ÜSTÜNTAŞ, T.; ŞAHIN, A. D. Wind turbine power curve estimation based on cluster center fuzzy logic modeling. **Journal of Wind Engineering and Industrial Aerodynamics**, v. 96, n. 5, p. 611–620, may 2008.

VAN RAVENZWAAIJ, D.; CASSEY, P.; BROWN, S. D. A simple introduction to Markov Chain Monte–Carlo sampling. **Psychonomic Bulletin and Review**, v. 25, n. 1, p. 143–154, 1 feb. 2018.

WACKER, B.; SEEBASS, J. V.; SCHLÜTER, J. CHR. A modular framework for estimating annual averaged power output generation of wind turbines. **Energy Conversion and Management**, v. 221, p. 113149, oct. 2020.

WAHBAH, M. et al. Wind speed probability density estimation using roottransformed local linear regression. **Energy Conversion and Management**, v. 199, p. 111889, nov. 2019.

WANG, Y. et al. Copula-based model for wind turbine power curve outlier rejection. **Wind Energy**, v. 17, n. 11, p. 1677–1688, 27 nov. 2014.

WANG, Y. et al. Approaches to wind power curve modeling: A review and discussion. **Renewable and Sustainable Energy Reviews**, v. 116, p. 109422, dec. 2019.

WEN, J.; ZHENG, Y.; DONGHAN, F. A review on reliability assessment for wind power. **Renewable and Sustainable Energy Reviews**, v. 13, n. 9, p. 2485–2494, dec. 2009.

WORTON, B. J. Using Monte Carlo Simulation to Evaluate Kernel-Based Home Range Estimators. **The Journal of Wildlife Management**, v. 59, n. 4, p. 794, oct. 1995.