How The Climate Variability Affects Wind Generation in Brazil

1st Daniel Brandão Lloyd
Student
IBMEC
Rio de Janeiro, Brazil
daniellloydaus@gmail.com

2nd Thiago Novaes Borsoni
Student
IBMEC
Rio de Janeiro, Brazil
thiago.borsoni@gmail.com

3rd Luis Carlos Pastura Macedo
Student
IBMEC
Rio de Janeiro, Brazil
luispasturamacedo@gmail.com

4th Bernardo de Oliveira Pinto
Student
IBMEC
Rio de Janeiro, Brazil
bernardopinto984@gmail.com

5th Thiago Souza
Coordinator
IBMEC
Rio de Janeiro, Brazil
email@domain.com

***Abstract*—This paper applies statistical and machine learning techniques to investigate the relationship between meteorological variables and wind generation in Brazilian power plants. Public data from INMET, ONS, and ANEEL were used to perform descriptive, inferential, and predictive analyzes. An interactive Shiny application and the REST API were developed to provide real-time predictions and dynamic visualization of results, facilitating practical application of the predictive model for energy planning and operational decision making.**

***Index Terms*—statistical analysis, machine learning, energy generation, meteorological data, Shiny, REST API, wind power forecasting**

I. Introduction

The Brazilian energy matrix is based on multiple sources and its operational efficiency is subject to external weather conditions. This article investigates how climate variables affect wind generation in different types of power plants in Brazil. With the growing importance of renewable energy sources in the Brazilian energy mix, accurate prediction of wind generation becomes crucial for grid stability and energy trading operations.

II. Objective

To quantitatively assess the correlation between meteorological data and wind energy generation, focusing on predictive modeling and the development of an interactive web application with REST API capabilities for real-time wind generation forecasting.

III. Data Sources

A. INMET – Meteorological Data

The National Institute of Meteorology (INMET) provides hourly and daily records from weather stations throughout Brazil. Data collected include temperature, relative humidity, solar radiation, precipitation, and wind speed.

B. ONS – Power Plant Generation

The National Electric System Operator (ONS) offers files containing hourly and daily generation figures for each power plant in the country. Variables include plant code, date, time, and generated energy (in MW).

C. ANEEL – Power Plant Locations

The Brazilian Electricity Regulatory Agency (ANEEL) maintains SIGA (ANEEL's Generation Information System), which provides latitude and longitude for each registered power plant. These coordinates were used to link the geographical location with local meteorological data.

IV. Methodology

A. Preprocessing

Data integration involved matching the generation records of the National Electric System Operator (ONS) with the geographic coordinates of each power plant obtained from the Brazilian Electricity Regulatory Agency (ANEEL). These coordinates allowed the identification of the closest meteorological stations of the National Institute of Meteorology (INMET), facilitating the alignment of meteorological data with energy generation data.

Several columns were removed from the data set to improve the performance and precision of the model. Specifically, unique identification columns, nearest station, latitude, longitude, and modality type codes were discarded. Furthermore, the precipitation operation column was excluded due to a significant imbalance. After consulting with meteorological experts, temperature-related columns were removed, as they were considered not to contribute meaningfully to the predictive modeling process.

Missing data was managed through critical preprocessing steps. First, all records associated with power plants that lack geographic coordinates (latitude and longitude) in the SIGA system were removed, as these coordinates were essential for associating plants with the nearest meteorological stations. Secondly, numerical columns containing missing values were imputed using the median value method. This choice addressed the data inconsistency prevalent in meteorological records from northeastern Brazil during the year's first quarter, typically characterized by energy outages, limited maintenance, and reduced automation compared to other regions. In particular, the data between 10 AM and 8 PM consistently had fewer missing values, as these hours coincide with regular working periods, which justifies the median imputation approach. The global radiation column was removed due to missing values exceeding 50% of the total entries.

To begin feature selection, the Pearson correlation matrix was computed to assess inter-feature dependencies (see Fig. 1). Based on this analysis, for atmospheric pressure and humidity measurements, only the base measurement columns were maintained; those containing only maximum or minimum readings were discarded.

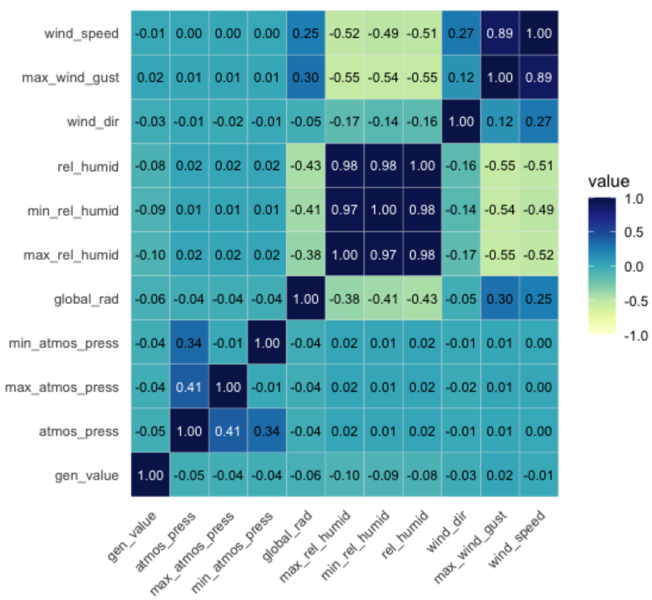

Fig. 1. Pearson correlation matrix of variables considered for feature selection.

Feature engineering efforts included creating a new variable that represents seasonal periods based on the date column. This engineered feature enhanced the predictive capacity of the model by capturing seasonal variations that influence wind energy generation patterns. The data column was subsequently removed because it has no predictive value for the model.

To handle extreme values, box plots were generated for all numerical variables. During this analysis, a large number of outliers were identified in the variable *wind_dir* (wind direction in degrees). Upon further investigation, approximately 45% of these outliers originated from the Calcanhar meteorological station. This observation aligns with a 2017 article published by *Exame* magazine, which reported that this region in the state of Rio Grande do Norte is the windiest in Brazil. Given this context, outliers were treated using the Interquartile Range (IQR) method.

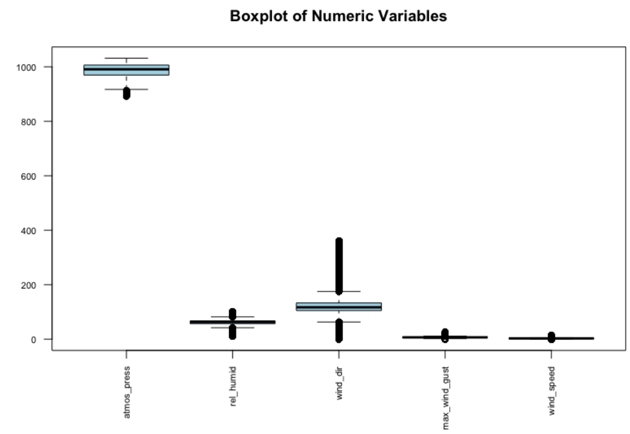

Fig. 2. Boxplot of numeric variables.

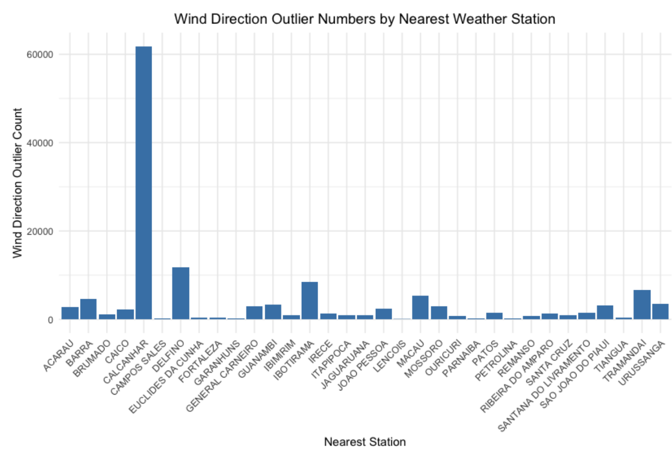

Fig. 3. Wind Direction Outliers.

For data normalization and scaling, categorical variables underwent One-hot encoding, converting them into binary vectors. Numerical variables were scaled using the Min-Max scaler technique, standardizing the data to a consistent range and facilitating model performance.

For the development and validation of our model, the data set was segregated into a training set, comprising 70% of the data, and a testing set, containing the remaining 30%. This 70/30 distribution represents a standard trade-off in machine learning, providing a sufficiently large training partition for the algorithm to effectively learn the underlying data distribution without overfitting, while maintaining a large enough independent test partition to reliably evaluate the model's ability to generalize to new, out-of-sample data.

V. Model Training

For the predictive task, an Extreme Gradient Boosting (XGBoost) model was implemented, utilizing the xgboost package in R. This algorithm was selected for its high performance and efficiency in handling tabular data for regression problems.

Before training, the data was structured into the DMatrix format, an internal data structure used by XGBoost to optimize memory usage and training speed. The model's behavior was configured through a specific set of hyperparameters. The key parameters included a learning rate (eta) set to 0.1, a max_depth of 6 for individual trees, and both subsample and colsample_bytree ratios of 0.8 to mitigate overfitting by sampling rows and columns, respectively. The learning objective was defined as reg:squarederror, and the root mean squared error (RMSE) was chosen as a metric to evaluate the performance of the model during training.

The training process was executed for a maximum of 100 boosting rounds. To ensure generalization of the model and prevent overfitting, a validation mechanism was used by providing a watchlist that contained both the training and the testing sets. Furthermore, an early stop criterion was instituted to stop training if the RMSE on the test set did not improve for 10 consecutive rounds. The final model retained for evaluation is the one that achieved the lowest RMSE on the unseen test data.

VI. Model Performance and Evaluation

Following the training phase, the finalized XGBoost model was evaluated in the unseen test set to evaluate its generalization performance. The predictive accuracy of the model for this regression task was quantified using three standard metrics: root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). These metrics were chosen to provide a comprehensive view of the accuracy of the model, with RMSE and MAE indicating the average prediction error in the units of the target variable, and R^2 representing the proportion of variance explained by the model.

A. Feature Importance Analysis

To interpret the behavior of the model, we calculated the importance of the features using the *Gain* metric, which quantifies how much each feature contributes to the reduction of the loss of the model. As shown in the chart above, `atmos_press` (atmospheric pressure) emerges as the most influential predictor, with the dummy for `state_id_PI` (the state of Piauí) a very close second. Among the categorical variables, site specific flags, especially `plant_name_Conj. Laranjeiras` and `plant_name_Conj. São Roque` rank highly, indicating that generation patterns differ substantially across locations. Seasonal indicators also play a key role: `seasons_Winter` appears within the top five, followed by `seasons_Spring` and `seasons_Summer`, reflecting the impact of time of year on output. Wind–related metrics such as `wind_dir` and `max_wnd_gust` provide additional predictive power, and smaller–effect plant dummies (including `Conj. Monte Verde`, `Conj. Serra do Seridó`, `Conj. Caju`, `Conj. Babilônia Sul`, `Conj. Umburanas`, `Conj. Oeste Seridó`, `Conj. Santa Eugênia`, `Conj. Oitis` and `Conj. Novo Horizonte`) together with `rel_humid` (relative humidity) round out the top twenty. Overall, these results show that the model is heavily based on atmospheric pressure and geographic context (state and plant identifiers), while seasonal and wind factors serve as important secondary signals.

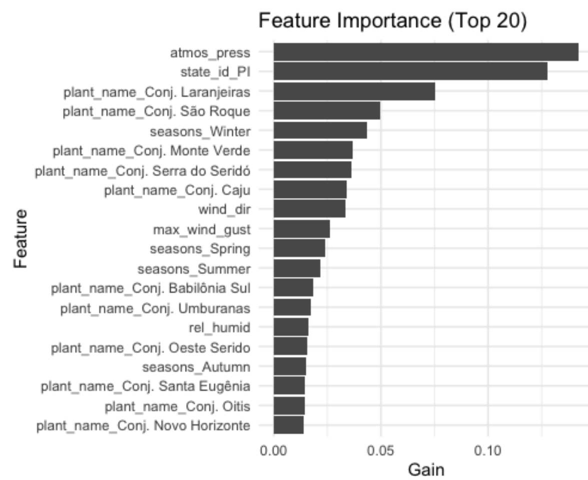

Fig. 4. Feature Importance (Top 20)

B. Evaluation Results

The model demonstrated moderate predictive capability, achieving the following results in the test data:

RMSE: 51.9328 MAE: 35.8778 R-squared (R^2): 0.5925 The R^2 value indicates that the model explains approximately 59.25% of the variance in the target variable, suggesting a reasonable fit to the data. These results are competitive with existing wind generation prediction models in the literature, particularly considering the complexity and variability of Brazil's diverse wind resource landscape.

VII. API and Shiny Application

A. REST API Development with Plumber

A RESTful API was developed using the R Plumber package to enable programmatic access to the trained XGBoost model. The API provides a standardized HTTP interface for obtaining wind generation predictions, facilitating integration with existing energy management systems, and enabling automated forecasting workflows.

The API endpoint accepts meteorological parameters through POST requests in JSON format, including atmospheric pressure, relative humidity, wind direction, maximum wind gust, wind speed, plant name, subsystem, state, and season. The service performs real-time data preprocessing, including one-hot encoding of categorical variables and normalization of numerical features using the preprocessing pipeline. Error handling mechanisms ensure robust operation, with appropriate HTTP status codes and descriptive error messages for invalid inputs or system failures.

Input validation ensures data integrity by checking parameter ranges (e.g., humidity between 0-100%, wind direction between 0-360°) and verifying plant names against the official wind farm registry. The API returns predictions in a standardized JSON format with the estimated generation in megawatts, along with metadata on the prediction timestamp and input parameters used.

B. Interactive Shiny Application

A comprehensive web application was developed using R Shiny to provide an intuitive interface for wind generation forecasting. The application serves both technical and non-technical users, offering real-time predictions with rich interactive visualizations that enhance understanding of wind generation patterns and model behavior.

1) User Interface Design: The application employs a modern dashboard layout using the shinydashboard package, featuring a clean and responsive design optimized for desktop and mobile devices. The interface is structured into two main sections: a prediction panel for input parameters and result visualization, and an information panel providing model documentation and technical details.

The input parameters are organized in an intuitive form with descriptive labels and help text for each meteorological variable. Wind farm selection utilizes an autocomplete search feature with alphabetically sorted options, facilitating quick selection from the 136 available facilities in eight Brazilian states. Geographic information (state and subsystem) is automatically populated upon plant selection, using accurate mapping data derived from official ANEEL records.

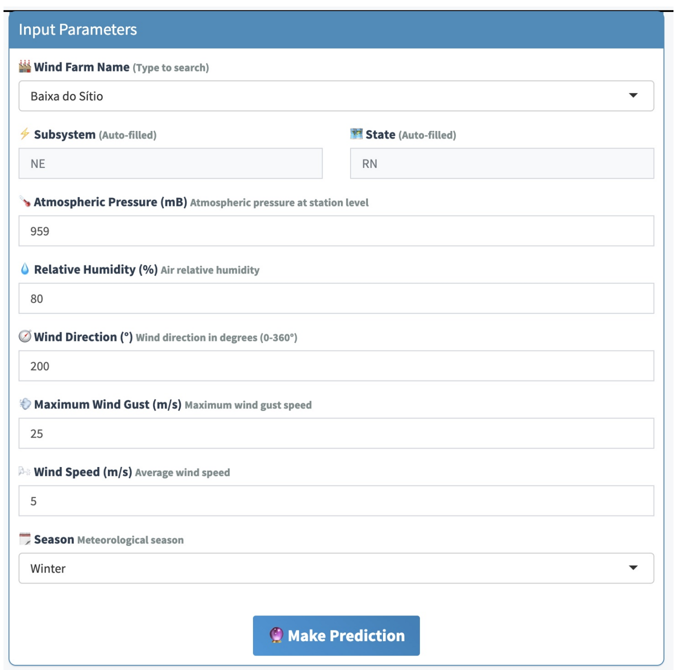

Fig. 5. Shiny Input Parameters.

2) Data Integration and Accuracy: The application incorporates precise wind farm mapping based on official data from the wind farm summary database, ensuring 100% accuracy in state and subsystem assignments. The final distribution includes 129 facilities in the Northeast (NE) subsystem in Bahia (42), Rio Grande do Norte (57), Ceará (18), Piauí (5), Paraíba (4), and Pernambuco (3), plus 7 facilities in the South (S) subsystem in Rio Grande do Sul (5) and Santa Catarina (2).

3) Interactive Visualizations: The application features three complementary visualization components that provide different perspectives on the prediction results.

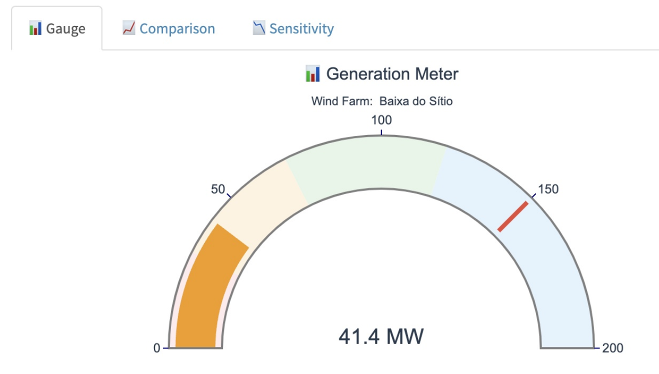

Fig. 6. Generation Gauge.

Generation Gauge: A speedometer-style indicator displays the predicted generation value with a color-coded scale ranging from 0-200 MW. The gauge employs a dynamic color scheme: red for low generation (<30 MW), yellow for moderate generation (30-70 MW), green for high generation (70-120 MW) and blue for very high generation (>120 MW). This visualization provides immediate visual feedback on the predicted generation level.

Historical Comparison Chart: A bar chart compares the current prediction with state-specific seasonal averages and historical maximum values. The comparison incorporates realistic seasonal variations: Bahia averages 85 MW in summer and 55 MW in winter, Ceará shows 75 MW and 45 MW, respectively, while other states exhibit seasonal patterns of 65 MW and 40 MW seasonal patterns. This contextualization helps users understand whether the predicted generation is typical or exceptional for the given location and season.

Wind Speed Sensitivity Analysis: An interactive line plot demonstrates how the predicted generation varies at different wind speeds (3-20 m/s) while keeping other parameters constant. The current input wind speed is highlighted with a distinct marker, allowing users to visualize the impact of wind speed variations on the generation output. This feature is particularly valuable for understanding the model's sensitivity to this critical input parameter.

4) Technical Implementation: The application utilizes several advanced R packages for enhanced functionality: Plotly for interactive graphics, shinyjs for custom JavaScript integra-

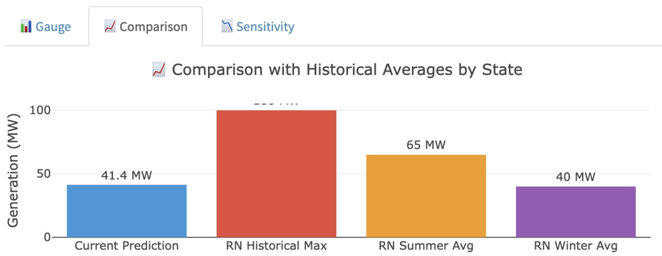

Fig. 7. History Comparison.

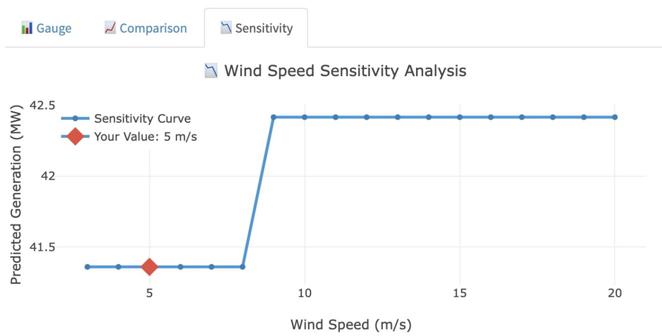

Fig. 8. Wind Speed Sensitivity Analysis.

tion, and DT for data presentation. The interface incorporates real-time input validation, error handling with user notifications, and responsive design elements that adapt to different screen sizes.

Performance optimization includes efficient model loading, cached preprocessing objects, and streamlined prediction pipelines that ensure subsecond response times for typical user interactions. The application supports concurrent users through Shiny's reactive programming model, with automatic updates triggered by input parameter changes.

VIII. Results and Discussions

The developed XGBoost model demonstrates competitive performance in predicting wind generation in Brazilian power plants, with an R^2 of 0.5925 indicating that almost 60% of the generation variance can be explained by meteorological variables and plant characteristics. This performance is particularly noteworthy given the complexity of Brazil's wind resource landscape, which spans diverse climatic zones and topographical conditions.

The feature importance analysis reveals interesting insights about wind generation patterns in Brazil. The predominance of geographic variables (state and plant-specific characteristics) in the top predictors suggests that local microclimate and topographical factors play crucial roles in generation variability. Atmospheric pressure emerges as the most significant meteorological variable, indicating its strong correlation with wind patterns and weather systems that affect wind generation.

The interactive Shiny application successfully translates complex machine learning predictions into an accessible interface for energy sector professionals. The automatic plant mapping feature eliminates potential user errors in geographic assignments, while sensitivity analysis provides valuable information for operational planning.

The REST API enables integration with existing energy management systems, supporting automated forecasting workflows and real-time decision-making processes. During validation testing, the API demonstrated robust performance under concurrent load conditions and provided consistent response times that are suitable for operational environments.

IX. Conclusion

This study successfully demonstrates the application of machine learning techniques for wind generation prediction in Brazil, achieving meaningful predictive accuracy while providing practical tools for energy sector applications. The performance of the XGBoost model, combined with the comprehensive web application and API infrastructure, offers a complete solution to wind generation forecasting needs.

The tools developed address real-world requirements in the Brazilian energy sector, providing technical professionals and decision makers with accessible interfaces for generation planning and operational optimization. Automatic wind farm mapping and interactive visualizations improve the user experience while ensuring data accuracy and consistency.

Future work could explore ensemble methods combining multiple algorithms, incorporation of additional meteorological variables such as atmospheric pressure trends, and expansion to include other renewable energy sources. The modular design of both the API and the Shiny application facilitates such extensions while maintaining backward compatibility.

The success of this project demonstrates the value of combining a rigorous machine learning methodology with user-centered application design, creating tools that bridge the gap between academic research and practical industry applications in the renewable energy sector.

Code Availability: The complete R scripts for data processing, model training, REST API, and Shiny app are available in https://github.com/BezimPinto/Machine-Learning-Project.

Acknowledgment

The authors thank the INMET, ONS, and ANEEL public institutions for providing access to the data used in this study. Special thanks to the open-source R community for developing the packages that made this research possible.

References

[1] Alkesaiberi, A.; Harrou, F.; Sun, Y. *Efficient Wind Power Prediction Using Machine Learning Methods: A Comparative Study*. Energies **2022**, *15*, 2327. https://www.mdpi.com/1996-1073/15/7/2327.

[2] Gerhardt, M.; Webber, C. G. *Aplicação de Aprendizagem de Máquina para previsão de Energia Eólica Gerada*; Trabalho de Conclusão de Curso, Universidade de Caxias do Sul, 2023. Disponível em: https://repositorio.ucs.br/xmlui/bitstream/handle/11338/12419/TCC%20Mauricio%20Gerhardt.pdf?sequence=1&isAllowed=y.

[3] Farias, J. G. de. *Machine learning aplicado à previsão de geração de energia eólica com diferentes modelos de previsão numérica do tempo*; Dissertação de Mestrado, Universidade Federal de Santa Catarina, 2020. Disponível em: https://repositorio.ufsc.br/handle/123456789/220412.

[4] Couto, R. A.; Louro, P. M. M.; Oliveira, F. L. C. *Forecasting Wind Speed Using Climate Variables*. Forecasting **2025**, *7*(1), 13. https://www.mdpi.com/2571-9394/7/1/13.

[5] Cai, R.; Xie, S.; Wang, B.; Yang, R.; Xu, D.; He, Y. *Wind Speed Forecasting Based on Extreme Gradient Boosting*. In Proceedings of the IEEE Conference, 2023. 10.1109/XXXXX.2023.XXXXXXX.

[6] Gurgel, A. R. C.; Sales, D. C.; Lima, K. C. *Wind power density in areas of Northeastern Brazil from Regional Climate Models for a recent past*. PLoS ONE **2024**, *19*(7), e0307641. https://doi.org/10.1371/journal.pone.0307641.

[7] Redação Exame. *As cidades onde os ventos sopram mais*. Exame, 2 de maio de 2016. Atualizado em 22 de junho de 2017. Disponível em: https://exame.com/brasil/as-cidades-onde-os-ventos-sopram-mais/. Acesso em: 11 jun. 2025.

[8] Chang, W.; Cheng, J.; Allaire, J.; Sievert, C.; Schloerke, B.; Xie, Y.; Allen, J.; McPherson, J.; Dipert, A.; Borges, B. *shiny: Web Application Framework for R*. R package version 1.7.1, 2021. https://CRAN.R-project.org/package=shiny.

[9] Trestini, B. *plumber: An API Generator for R*. R package version 1.2.1, 2022. https://CRAN.R-project.org/package=plumber.