ENHANCING GLOBAL SOLAR IRRADIANCE FORECASTS IN BRAZIL USING NUMERICAL MODELS AND MACHINE LEARNING

Francisco José Lopes de Lima[1,2], André Rodrigues Gonçalves[3], Rodrigo Santos Costa[3], Diogo Nunes da Silva Ramos[2], Ana Paula Paes dos Santos[2], Thalyta Soares dos Santos[2], William Duarte Jacondino[2], Allan Rodrigues Silva[2], Davidson Martins Moreira[2], Enio Bueno Pereira[3] and Fernando Ramos Martins[1]

[1] Institute of Marine Sciences, Federal University of São Paulo, Santos – São Paulo (Brazil)

[2] SENAI CIMATEC, Salvador - Bahia (Brazil)

[3] National Institute for Space Research, São José dos Campos – São Paulo (Brazil)

Abstract

This study aims to improve forecasting of global solar irradiance (GHI) in Brazil by integrating ECMWF numerical weather prediction (NWP) data with machine learning techniques for bias correction. Observational data from 17 meteorological stations in the SONDA network were used for training and validation, covering 2021 and 2022. Two models were implemented: Multiple Linear Regression (MLR) and a Multilayer Perceptron (MLP) neural network. The MLP model was optimized through hyperparameter tuning and feature selection, incorporating 28 predictor variables related to radiative, thermodynamic, and cloud parameters. The results indicate that the MLP model generally provided slightly improved performance compared with both the raw ECMWF forecasts and the MLR model in terms of error metrics and correlation. In May and October 2022, the MLP achieved the lowest NRMSEs (26% and 28%, respectively) and the highest correlation coefficients ($R \approx 0.93$). These results suggest that neural networks can capture nonlinear atmospheric relationships and contribute to improving the representation of solar irradiance variability. The findings also align with benchmark studies in the literature, reinforcing the robustness of the proposed approach. The methodologies developed here can contribute to more accurate solar forecasting, supporting operational energy management and the integration of renewable sources in Brazil's electricity matrix.

Keywords: Solar forecasting, machine learning, ECMWF, bias correction, irradiance modeling, Brazil

1. Introduction

Climate variability, in the context of solar radiation, refers to natural and cyclical fluctuations in the amount of energy reaching the Earth's surface. This variability affects several sectors, such as energy, agriculture, civil construction, and water resource management (Wild et al., 2020). In the energy sector, the transition to renewable sources, such as solar and wind, has increased dependence on climate conditions, making the quantification of future risks essential to ensure energy security. Solar radiation forecasting plays a crucial role in energy planning, enabling the anticipation of periods of generation scarcity or surplus and ensuring a stable and reliable energy supply (Martins et al., 2008; Martins et al., 2012).

In Brazil, the electric power matrix is dominated by hydropower generation, which accounts for 58.9% (Zucareli, 2024). This dependence makes the system vulnerable to climatic events, such as prolonged droughts, which can drastically reduce water availability and compromise power generation, leading to rationing and conflicts over water use. Furthermore, international commitments, such as the Kyoto Protocol (1997) and the Paris Agreement (2015), require reductions in greenhouse gas emissions, encouraging the adoption of renewable energy sources, such as solar. Brazil has a privileged geographic location that ensures high levels of irradiation throughout the year. Studies indicate that the country has one of the highest solar potential in the world, with average irradiation levels surpassing those of leading photovoltaic countries such as Germany and China (IRENA, 2019; Pereira et al., 2017).

The growing share of solar technology in electricity generation requires special attention in planning, operation, and energy trading. However, the intermittency of solar generation and its dependence on meteorological factors, such as cloud cover, make accurate forecasting difficult. The solar radiation incident at the surface is influenced by complex, nonlinear atmospheric processes, which increase forecast uncertainty. In this context, integrating global

10.18086/swc.2025.04.19 Available at http://proceedings.ises.org

meteorological models, such as the ECMWF, with Machine Learning (ML) techniques can reduce these uncertainties by adjusting forecasts to local conditions (Diagne et al., 2013; Lima et al., 2016).

Numerical Weather Prediction (NWP) models are based on equations that describe atmospheric motions from the current state of the ocean-atmosphere system. The solutions to these equations, obtained through numerical methods, provide meteorological forecasts over temporal scales ranging from hours to about one or two weeks and spatial scales from meters to hundreds of kilometers. The temporal and spatial resolutions of these models depend on the type of meteorological system under study. For instance, in solar photovoltaic forecasting, cloud variability must be resolved on temporal scales of minutes or hours and spatial resolutions of a few kilometers, while phenomena such as El Niño can be represented at larger temporal and spatial scales.

The performance of these models depends on the quality of the observations used as input (initial conditions) and on the challenge of representing the entire Earth system, given its complex and nonlinear processes. This complexity, combined with parameterizations, makes it difficult to forecast certain variables, such as radiation and cloudiness. Therefore, statistical post-processing models (such as ML models) are often required to relate NWP outputs to observed data, correcting systematic biases (Nascimento et al., 2017).

Different machine learning approaches have been applied to solar irradiance forecasting. Linear regression, widely used for its simplicity and efficiency, models linear relationships between atmospheric variables and solar irradiance, capturing basic trends (Wilks, 2011; Ibrahim et al., 2012). Artificial neural networks, on the other hand, handle nonlinear relationships and effectively identify complex patterns in atmospheric interactions (Diagne et al., 2013; Ahmed et al., 2020). Ensemble techniques, such as Gradient Boosting, combine multiple predictors to optimize performance and dynamically adapt to climatic variability (Obiora et al., 2021; Qiu et al., 2022).

It is important to note that machine learning techniques are typically applied as statistical post-processing tools, rather than as replacements for numerical weather prediction models. Global forecasting systems such as ECMWF are designed to produce physically consistent atmospheric simulations on a planetary scale. In contrast, machine learning bias-correction approaches generally require local observational data for calibration and training. Because high-quality ground observations are not available globally, these methods are typically applied at specific locations where observational datasets exist**,** rather than being implemented directly within global forecasting systems.

Considering the above, the objective of this study is to develop a global solar irradiance (GHI) forecasting tool for Brazil by integrating a global meteorological forecasting model (ECMWF) with machine learning, based on bias correction for 24-hour horizons. The central hypothesis is that combining global forecasts with local adjustments will reduce uncertainty in solar radiation predictions, thereby improving accuracy and reliability.

It is important to emphasize that the present work represents a preliminary methodological evaluation of machine learning techniques for statistical post-processing of numerical weather prediction outputs. The objective is not to establish statistically significant differences between models, but rather to investigate the potential of simple machine learning approaches for reducing systematic biases in short-term solar irradiance forecasts. A more detailed statistical significance analysis, including confidence interval estimation and uncertainty quantification, will be addressed in future studies with longer time series and larger datasets.

2. Materials and Methods

2.1. Forecast Models and Bias Correction

For the solar forecasting experiments, the Integrated Forecasting System (IFS) model was employed. The IFS is a global numerical weather prediction system jointly developed and maintained by the European Centre for Medium-Range Weather Forecasts (ECMWF), based in Reading, United Kingdom, and Météo-France, based in Toulouse, France. It is a spectral model that uses a terrain-following vertical coordinate system and includes advanced data assimilation processes (ECMWF, 2024). Both ECMWF and Météo-France operate their own configurations of the IFS for generating operational weather forecasts, differing mainly in their resolution and parameterization settings. Several meteorological centers worldwide (particularly across Europe) use IFS forecasts as boundary conditions for their regional models. For the analysis carried out in this study, IFS forecasts were provided with a 24-hour lead time, hourly temporal resolution, and 9 km spatial resolution, covering the period from 2021 to 2022.

Forecasts of GHI from NWP models are typically overestimated due to challenges in representing cloud formation processes. Other sources of uncertainty may also affect forecast accuracy. Although higher spatial resolutions can improve the numerical representation of clouds, they substantially increase the computational cost. Conversely, applying spatial resolutions coarser than 9 km may have the opposite effect, degrading forecast quality by generating false or time-space-shifted clouds (Pedruzo-Bagazgoitia et al., 2019). Understanding these limitations is essential for diagnosing outliers during the post-processing of solar forecasts from NWP models. The application of statistical bias-correction methods is an effective approach to reducing these errors.

2.2. Data Acquisition and Processing

The data used in this research consist of hourly solar irradiance observations collected over a two-year period, from January 2021 to December 2022. These measurements were obtained from monitoring stations distributed across Brazil. Figure 1 shows the spatial distribution of these stations, while Table 1 presents their geographic coordinates. The LABREN (Laboratório de Modelagem e Estudos de Recursos Renováveis de Energia) maintains a solarimetric database derived from the SONDA Network (*Sistema de Organização Nacional de Dados Ambientais*) and its partner institutions, covering all regions of Brazil. General information on each station is publicly available at http://sonda.ccst.inpe.br/. The observational period selected for this study was chosen to match the availability of ECMWF global weather forecasting model data.

It should be highlighted that the analysis performed in this study focuses on point-based evaluations, using ground observations at specific meteorological stations. Therefore, the objective is not to generate spatially continuous irradiance predictions, but rather to assess the capability of machine learning methods to reduce local systematic biases in numerical weather prediction outputs.

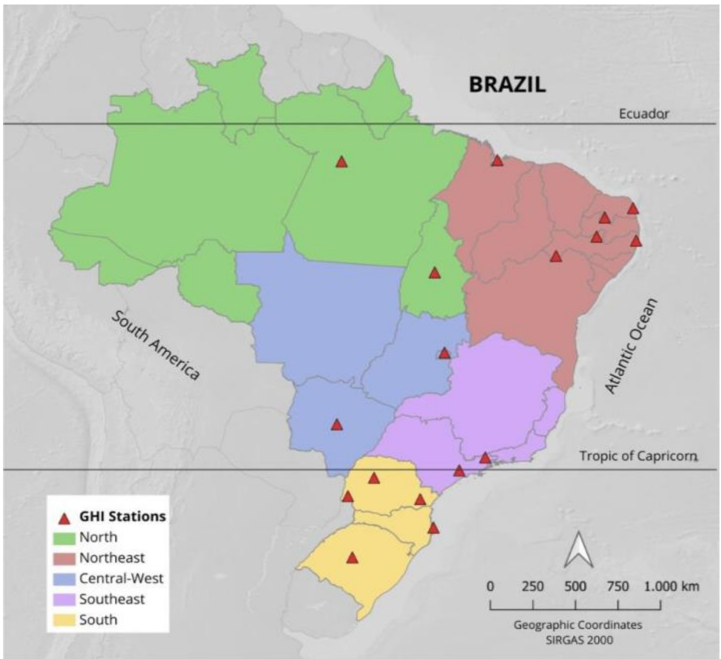

Figure 1: Location of the SONDA Network observation stations (in red).

Table 1. Coordinates and location of the meteorological stations used in this study.

ID	CODE	REDE	LAT	LON	ALT
ESS01	AFG	IFPE	-7.76	-37.63	542
ESS02	SLM	IFPE	-8.04	-35.01	58
ESS03	SMS	SONDA	-29.44	-53.82	489
ESS04	CPA	SONDA	-22.69	-45.01	574
ESS05	CGR	SONDA	-20.45	-54.84	437
ESS06	BRB	SONDA	-15.60	-47.71	1023
ESS07	PMA	SONDA	-10.18	-48.36	216
ESS08	PTR	SONDA	-9.07	-40.32	387
ESS09	CAI	SONDA	-6.47	-37.09	176
ESS10	NAT	SONDA	-5.84	-35.21	58
ESS11	STM	SONDA	-2.68	-54.53	35
ESS12	SLZ	SONDA	-2.59	-44.21	40
ESS13	SPK	UFSC	-27.43	-48.44	20
ESS14	CTS	UTFPR	-25.51	-49.32	891
ESS15	MDS	UTFPR	-25.30	-54.11	402
ESS16	SPO	IEE	-23.56	-46.73	772
ESS17	CMS	UTFPR	-24.06	-52.38	630

In addition to the ECMWF forecasts, data from the ERA5 reanalysis, produced by the Copernicus Climate Change Service (C3S), were also used. The ERA5 dataset provides high-quality information on key atmospheric variables, including global horizontal irradiance (GHI), air temperature, and cloud cover, with an hourly temporal resolution and a spatial resolution of approximately 27 km. In this work, ERA5 data were evaluated primarily to assess their suitability for replacing ground-based measurements in regions where no observational data are available. This evaluation aimed to verify the consistency between ERA5-derived irradiance and observed GHI, thus determining its potential as a proxy dataset for areas with sparse station coverage. However, it is important to note that all predictive variables used in the forecasting models were derived exclusively from the ECMWF forecasts**.** The ERA5 data were extracted using the nearest-neighbor interpolation method based on the geographic coordinates of the 17 analyzed meteorological stations. The extraction focused on obtaining the GHI variable, which was later compared with observed data for performance evaluation.

The complete set of variables used in this study was organized according to their physical and descriptive categories. The identification and geographical variables include *id, lat, lon, alt, nomes, est, CODE, REDE, municipio*, and *uf,* which provide metadata on the spatial location and context of each monitoring station. The temporal and forecasting variables are *start_time, step, and dia_previsao*, which represent, respectively, the forecast initialization time, the forecast lead time, and the prediction day. The wind variables comprise *WS10m, WD10m, WS100m, WD100m, WS120m, WD120m*, and *WG10m*, representing wind speed and direction at different heights above the surface, as well as the wind gust intensity. The thermodynamic and atmospheric variables include *Tair_2m, RH_2M*, and *Precip,* which correspond to air temperature at 2 m, relative humidity at 2 m, and accumulated precipitation. Cloud cover at different levels is represented by *Cb_cloud, L_cloud, M_cloud*, and *H_cloud*, referring to convective, low, medium, and high clouds, respectively. The radiative variables include GHI, DIR, DNI, DIF, POA, and TOA, which describe the global, direct, normal, diffuse, plane-of-array, and top-of-atmosphere solar irradiances. Clear-sky components ghi_cs, dni_cs, and dif_cs were also included. The variable ghi_obs represents the ground-based observed irradiance and was used as the primary target for model training and validation. The reanalysis-based irradiance ghi_era5 was additionally considered as an alternative target in specific evaluations, particularly in regions or periods where observational data were unavailable. Finally, the solar geometry variables (*zenith, azimuth, and coszenith*) represent the apparent solar position and related geometric parameters used to model the angular dependence of solar irradiance.

During preprocessing, data were standardized and normalized to ensure consistent variable ranges ([0, 1]) across all features. Missing or inconsistent records were removed, and the datasets were temporally aligned. The training period covered January 1, 2021, to April 30, 2022**,** while May and October 2022 were reserved for validation, representing the months of lowest and highest solar incidence, respectively. Finally, all datasets were integrated into a single standardized database, enabling comparative analyses and the development of machine learning, based bias-correction models.

2.3. Model training and calibration

Initially, observed GHI data from surface stations are collected with meteorological forecasts from the ECMWF, which include variables such as incident solar radiation, air temperature, humidity, wind, and cloud cover. These datasets are temporally and spatially synchronized to ensure that all time series are aligned, with the number of samples being limited by the availability of observed data, which serve as the reference for model training and validation.

To obtain higher-resolution solar radiation (GHI) forecasts over the study region in Brazil, ECMWF forecasts are compared with ground-based observations and with ERA5 reanalysis data. The purpose of this evaluation is to assess the consistency of ERA5 as a potential substitute for observations in areas lacking measured data, and to verify the uncertainties associated with both reanalysis and forecast datasets. Model performance is assessed using statistical indices and by comparing estimated values with observations recorded at the monitoring stations.

Predictor selection is based on a physical understanding of the interactions between solar radiation and the atmosphere to identify the optimal set of ECMWF forecast variables that show a significant relationship with observed solar radiation. Subsequently, machine learning models, including multiple linear regression and artificial neural networks, are developed and trained to map the ECMWF forecasts to the observed solar radiation.

Figure 2 illustrates the flowchart of the proposed methodology for solar radiation forecasting using the ECMWF model and a statistical refinement based on machine learning techniques to reduce the model's systematic error.

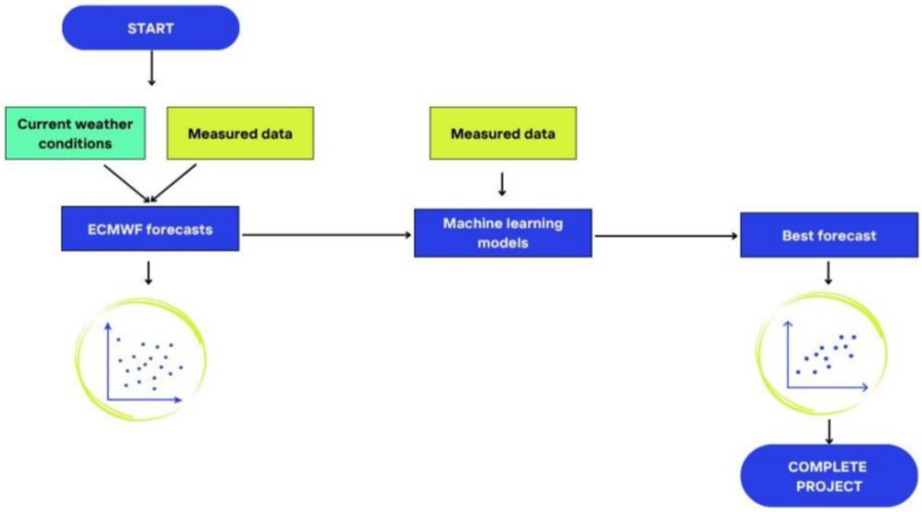

Figure 2: Short-term forecasting scheme (up to 24 hours ahead) using statistical methods applied to the ECMWF model.

Bias correction is applied to the short-term forecasts using ML models to adjust the raw ECMWF predictions. The performance of the models is evaluated using statistical metrics such as the root mean square error (RMSE), mean bias (BIAS), and correlation coefficient (R), calculated for both the validation and testing datasets. The evaluated metrics in this study are described as follows:

BIAS Eq. (1): This metric quantifies the systematic tendency of a model to underestimate or overestimate the true values. When the bias is close to zero, it indicates that the model does not exhibit a systematic tendency. Its main advantage is that it is easy to interpret and can be used to assess model accuracy relative to a reference line. However, it may be influenced by outliers and does not provide information about error variability.

$$BIAS = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i}) \qquad \text{(eq. 1)}$$

Root Mean Square Error (RMSE)**,** Eq. (2): This metric measures the difference between the true values and the values predicted by the model. It represents the dispersion of the errors, where lower values indicate better model performance. Its main advantage is that it penalizes larger errors more strongly, making it particularly suitable for models in which large deviations have a more significant impact. However, it can be sensitive to outliers.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \qquad \text{(eq. 2)}$$

Normalized Root Mean Square Error (NRMSE)**,** Eq. (3): This metric normalizes the RMSE by the mean of the true (observed) values. It allows comparing models that operate on different scales and units. Its main advantage is that it provides a relative measure of error dispersion while remaining straightforward to interpret.

$$NRMSE = \frac{RMSE}{\frac{1}{n}\sum_{i=1}^{n} y_i} \qquad \text{(eq. 3)}$$

Mean Absolute Error (MAE)**,** Eq. (4): This metric measures the average absolute difference between the true and predicted values. It represents the dispersion of the errors, and like RMSE, lower values indicate better model performance. Its main advantage is that it is easy to interpret and can be used to compare models operating on different scales and units.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \qquad \text{(eq. 4)}$$

Pearson correlation coefficient (R), Eq. (5): This metric estimates the linear relationship between the true and predicted values. It ranges from –1 to 1, where values close to 1 indicate a strong positive correlation, values close to –1 indicate a strong negative correlation, and values near 0 suggest little to no correlation. Its main advantage is that it provides a standardized measure that can be used to compare models operating on different scales and units, and is easy to interpret. However, it may fail to capture nonlinear relationships between variables and can be influenced by outliers.

$$R = \frac{\sum_{i=1}^{n}\left(y_i - \underline{y}\right)\left(\hat{y_i} - \underline{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \underline{y}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\hat{y_i} - \underline{\hat{y}}\right)^2}} \qquad \text{(eq. 5)}$$

In the formulation of the metrics, y_i denotes the true (observed) value, $\hat{y}_i$ denotes the model-predicted value, n is the total number of samples, $\bar{y}$ is the mean of the observed values, and $\bar{\hat{y}}$ is the mean of the predicted values.

Finally, the bias-corrected forecasts are compared with the raw ECMWF predictions using observed data as the reference, and the effectiveness of the bias correction is analyzed across different weather conditions. All data processing, model training, and statistical analyses are performed in Python®, with results visualized through graphs and maps to facilitate interpretation and communication. This methodological approach not only improves the accuracy of solar radiation forecasts but also optimizes solar energy generation systems, strengthening the transition toward renewable energy sources (Wilks, 2011).

In this study, two distinct models were tested for solar irradiance forecasting: a Multilayer Perceptron (MLP) and a Multiple Linear Regression (MLR) model. The MLP model was configured using a Randomized Search process for hyperparameter optimization. Additionally, Sequential Feature Selection (SFS) was applied to both models (MLP and MLR) to identify the most relevant predictor variables.

Before the final training phase, a hyperparameter search was conducted for both models. For the MLP model, the following configurations were tested:

- Number of neurons in the hidden layers: random samples between 10 and 100 for two hidden layers.
- Activation functions: *relu*, *tanh*, and *logistic*.
- Solvers for optimization: *adam* and *sgd*.
- Regularization (alpha): 0.0001, 0.001, and 0.01.
- Learning rate: *constant* and *adaptive*.

After the search process, the best hyperparameters were automatically selected. The trained models were validated using data from May and October 2022**,** representing periods of lower and higher solar radiation, respectively. Evaluating the models' robustness under these contrasting climatic conditions enabled assessment of their generalization capability.

Initially, the training was performed by combining data from all stations into a single dataframe to test the models' overall generalization capacity. For future work, it is planned to train individual models for each station and compare their performance with that of the generalized model trained using all stations combined.

3. **Results and Discussion**

3.1. Evaluation of ECMWF and ERA5 Data

The evaluation of the ERA5 reanalysis and ECMWF forecasts for the 17 meteorological stations analyzed across Brazil reveals consistent differences in their ability to estimate global horizontal irradiance (GHI). The main statistical indicators show that ERA5 achieves a correlation of 90%**,** a mean bias of 1.70 W/m²**,** and an RMSE of 35%, indicating good agreement with observations and a slight tendency to overestimate. In contrast, ECMWF presents a correlation of 91%**,** a mean bias of 12.93 W/m², and an RMSE of 33%, suggesting a similar overall accuracy but with a stronger bias and slightly greater error dispersion.

Figure 3 summarizes the statistical performance metrics for both datasets across all stations. The mean bias indicates that ECMWF tends to overestimate irradiance in most cases, whereas ERA5 shows smaller, more consistent deviations. The MAE results confirm that ERA5 provides lower average errors, suggesting a better representation of observed GHI values. This pattern is also reflected in the RMSE, which remains generally lower for ERA5 than for ECMWF. Both NRMSE (%) and RMSE (W/m²) highlight regional variations, with higher errors in some locations, likely due to local atmospheric complexity and cloud dynamics.

Overall, ERA5 shows slightly lower mean errors and more stable performance across climatic regions than ECMWF. Both datasets, however, show strong correlations with observed data (ranging from 0.85 to 0.95), confirming their potential for solar resource assessment.

These findings reinforce the conclusion that ERA5 is a robust and reliable dataset for estimating GHI and can serve as a suitable reference in areas lacking direct solar measurements. Nevertheless, the presence of systematic biases indicates the importance of applying bias-correction techniques, such as statistical or machine-learning adjustments, to improve accuracy in regions with high cloud cover or strong atmospheric variability.

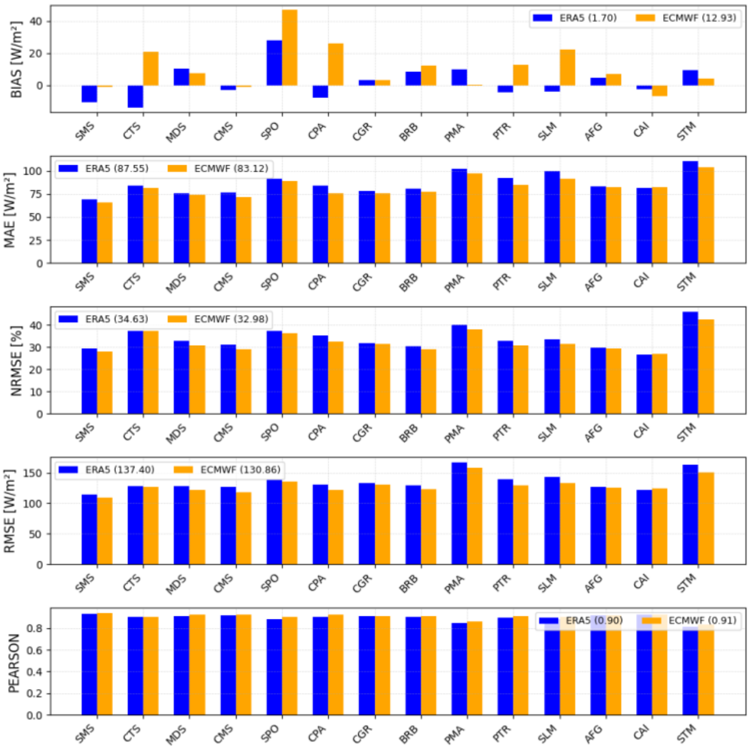

Figure 3: Comparison of statistical performance metrics for GHI estimates across the 17 stations analyzed in Brazil. The blue bars represent the values obtained from ERA5, while the orange bars represent those from the ECMWF model. The values shown in the legend for each metric correspond to the averages across all stations.

3.2. Evaluation of the Models for May 2022 and October 2022

After data preprocessing and hyperparameter tuning, both the MLP and MLR models were trained using 28 predictor variables related to radiative, thermodynamic, and cloud parameters. During testing, the MLP model achieved an RMSE of 117.69 W/m², while the MLR reached 134.89 W/m², suggesting that the neural network is capable of capturing nonlinear relationships among atmospheric variables.

The learning curve of the MLP (Figure 4) shows a rapid decrease in loss during the first 50 epochs, followed by stabilization between epochs 75 and 100, indicating model convergence. This pattern reflects effective learning without signs of overfitting or underfitting, as the model reduced errors efficiently and reached a stable minimum loss. Overall, the MLP demonstrated better adaptability and predictive accuracy than the linear regression model for solar irradiance forecasting.

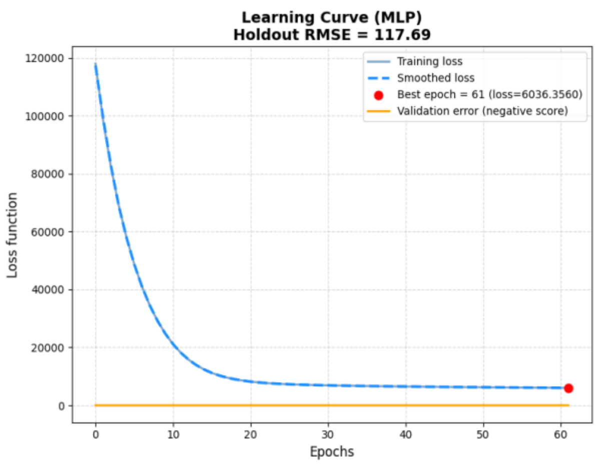

Figure 4: Evolution of the loss function during MLP training, showing rapid convergence and stability across epochs.

Figure 5 presents comparisons between observed solar irradiance and forecasts from the ECMWF, MLP, and MLR models for May and October 2022, combining data from all stations with measurements available during these periods. Each set of plots corresponds to a specific month, with the first image representing May and the second representing October. The purpose of this analysis is to evaluate the performance of the models under different seasonal conditions.

In each plot, the X-axis represents the observed GHI values (W/m^2), while the Y-axis represents the predicted values from each model. The black dotted line indicates the ideal line $y = x$, where predictions would perfectly match the observations. The red dashed line corresponds to the regression line fitted for each model, providing insight into the bias and statistical adjustment of the predictions. Three main statistical metrics were used to evaluate model performance: the correlation coefficient (R), which indicates the strength of the relationship between predicted and observed values; the mean bias (Bias, in W/m^2), which measures whether the model tends to overestimate or underestimate; and the normalized root mean square error (RMSE, in %), which quantifies the accuracy of the forecasts.

For May 2022, the ECMWF model achieved a correlation coefficient of 91%, a positive bias of 20.45 W/m^2, and an RMSE of 33%, indicating a slight overestimation of solar irradiance. The MLP model, in contrast, achieved the best performance, with an R of 94%, a bias of 5.76 W/m^2**,** and an RMSE of 27%, suggesting a slight overestimation but a significantly lower mean error. The MLR model achieved an R of 91%, a bias of 7.01 W/m^2**,** and an RMSE of 31%, indicating a stronger tendency toward underestimation.

For October 2022**,** the models performed similarly overall, though with notable differences in bias. The ECMWF model presented an R of 92%, a positive bias of 16.83 W/m^2**,** and an RMSE of 30%, confirming its consistent tendency to overestimate irradiance during this period. The MLP model again achieved the best results, with an R of 92%, a bias of 3.94 W/m^2**,** and an RMSE of 30%**,** demonstrating a closer fit to the observed values. The MLR model**,** with an R of 92%, exhibited a high bias of 7.16 W/m^2**,** indicating substantial overestimation, along with an RMSE of 29%, higher than that of the MLP.

Based on these results, the MLP model generally showed slightly better performance according to the evaluated metrics, showing the highest correlation, the lowest mean error (RMSE), and a bias closest to zero. The ECMWF model, although performing reasonably well, consistently overestimated solar irradiance. The MLR model showed less stability, underestimating irradiance in May and overestimating it in October, suggesting a lower generalization capacity. Therefore, the results indicate that the MLP neural network is the most promising approach for solar irradiance forecasting, achieving a better fit to the observed data and lower mean error than the other model

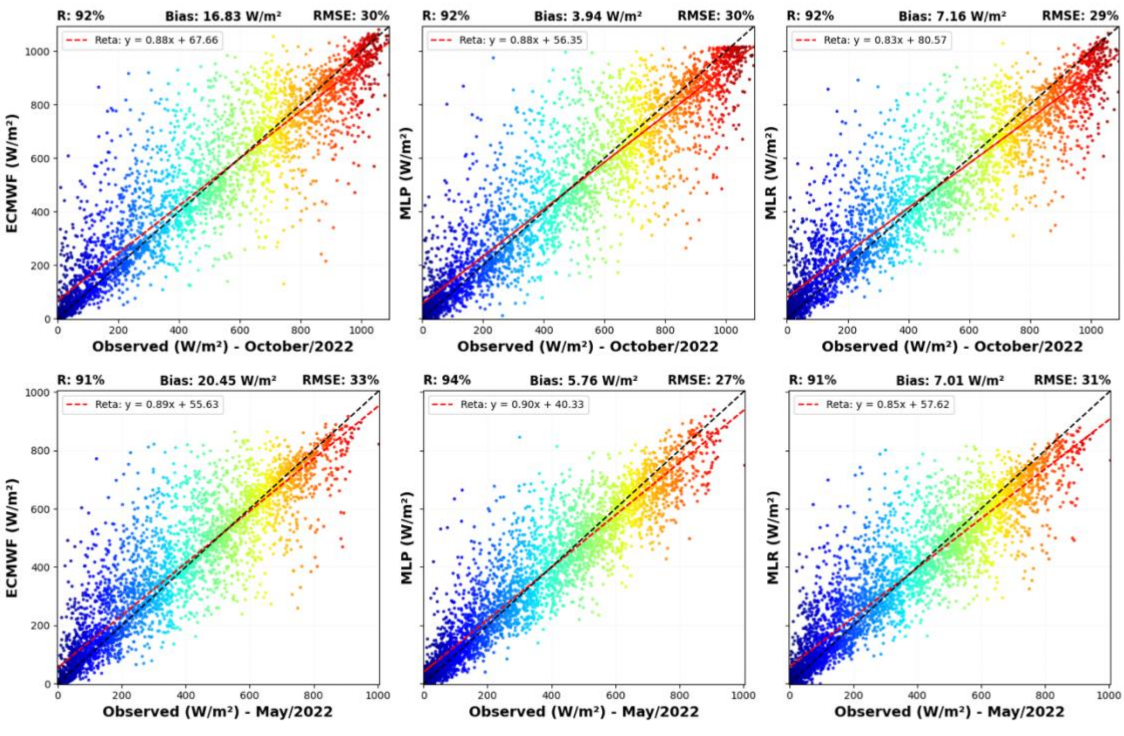

Figure 5: Scatterplots between observed and predicted values from the ECMWF, MLP, and MLR models for May and October 2022. The color scale highlights the distribution of samples, allowing visualization of areas with higher point density and facilitating the interpretation of error and bias patterns.

Based on Figure 6 (May 2022), the comparative evaluation of the ECMWF, MLP, and RLM models reveals consistent performance differences across all meteorological stations analyzed. The ECMWF model shows a positive bias (mean = 20.45 W/m²), indicating a slight tendency to overestimate solar irradiance. The RLM model also exhibits a positive bias (7.01 W/m²), though of smaller magnitude, while the MLP model shows the lowest bias (5.76 W/m²), suggesting more balanced predictions with minimal systematic deviation.

In terms of mean absolute error (MAE), ECMWF and RLM display average errors of 69.94 W/m² and 70.27 W/m², respectively, whereas the MLP model achieves a smaller error of 60.37 W/m², confirming its greater consistency with the observed values. The normalized root mean square error (NRMSE) reinforces this trend: MLP yields the lowest relative error (26.81%), followed by RLM (31.33%) and ECMWF (33.23%).

The RMSE, which emphasizes large deviations, reaches 110.31 W/m² for ECMWF and 104.01 W/m² for RLM, while the MLP again performs best with 88.98 W/m², showing a substantial reduction in prediction errors. Finally, the Pearson correlation coefficients indicate strong linear relationships between predictions and observations for all models: ECMWF (0.91), RLM (0.91), and MLP (0.94).

Overall, the results demonstrate that the MLP neural network model delivers the most accurate and reliable forecasts, characterized by lower bias and error metrics and a higher correlation with observed irradiance values. The ECMWF shows intermediate performance, with a systematic overestimation tendency, while the RLM exhibits slightly higher errors and less station-to-station stability.

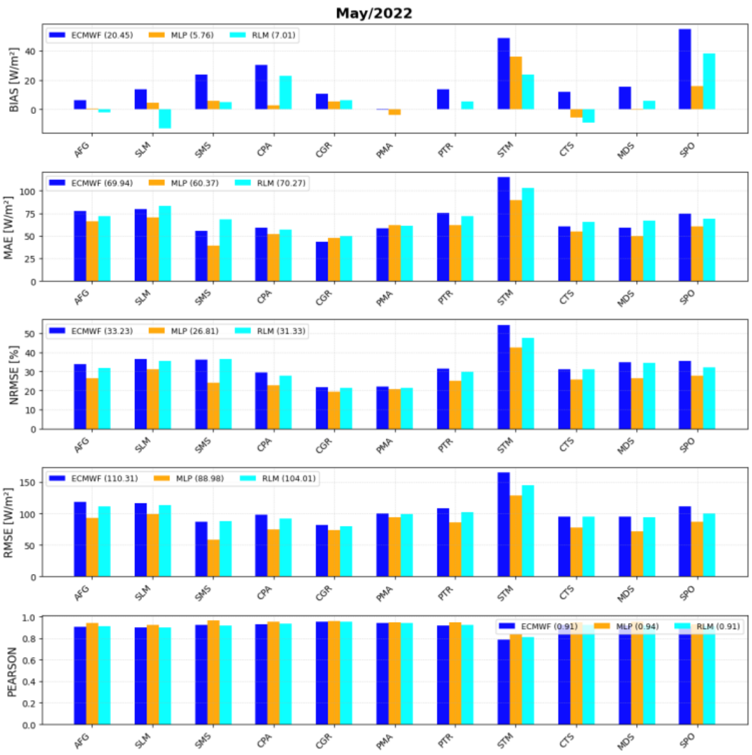

Figure 6. Comparison of the performance of the ECMWF, MLP, and MLR models in forecasting solar irradiance for different meteorological stations in May 2022. The values shown in the legend of each metric correspond to the averages calculated for all stations.

Based on Figure 7 (October 2022), the evaluation of the ECMWF, MLP, and RLM models reveals consistent patterns across most meteorological stations, highlighting differences in model performance under higher solar irradiance. It is important to note that some stations are not represented in this figure due to the absence of observed data for this period, which limited the comparative analysis.

The ECMWF model shows a positive bias of 16.83 W/m², indicating a slight overestimation of the observed values. The RLM model shows a similar but slightly higher bias of 7.16 W/m², while the MLP model achieves the lowest bias (3.94 W/m²), suggesting that its forecasts are more balanced and less prone to systematic deviations.

In terms of mean absolute error (MAE), the ECMWF model has an average error of 85.60 W/m², while the RLM has a slightly higher average error of 89.67 W/m². The MLP once again produces the lowest MAE (87.45 W/m²), demonstrating a closer match to the observed data. The normalized RMSE (NRMSE**)** values further confirm this trend: 30.19% for ECMWF**,** 29.34% for RLM, and 29.86% for MLP. Although the differences are smaller than in May, the neural network still shows a slight advantage in most stations.

When analyzing the RMSE, which emphasizes large errors, ECMWF reaches 132.73 W/m²**,** RLM shows 128.99 W/m²**,** and MLP achieves the smallest error at 131.26 W/m²**,** confirming its superior consistency. Finally, the Pearson correlation coefficients indicate that all models maintain a strong linear relationship between observed and predicted irradiance values, with ECMWF and RLM at 0.92, and MLP also at 0.92, suggesting that all models effectively capture the temporal patterns of irradiance.

Overall, the results for October confirm that the MLP model maintained the best overall performance, with the lowest bias and smaller errors on average across stations. The ECMWF model performed reasonably well but continued to slightly overestimate irradiance, while the RLM model exhibited larger errors and less consistency, indicating limited generalization capacity. The lack of data for some stations underscores the need to expand the observational network to improve model validation and regional performance assessments.

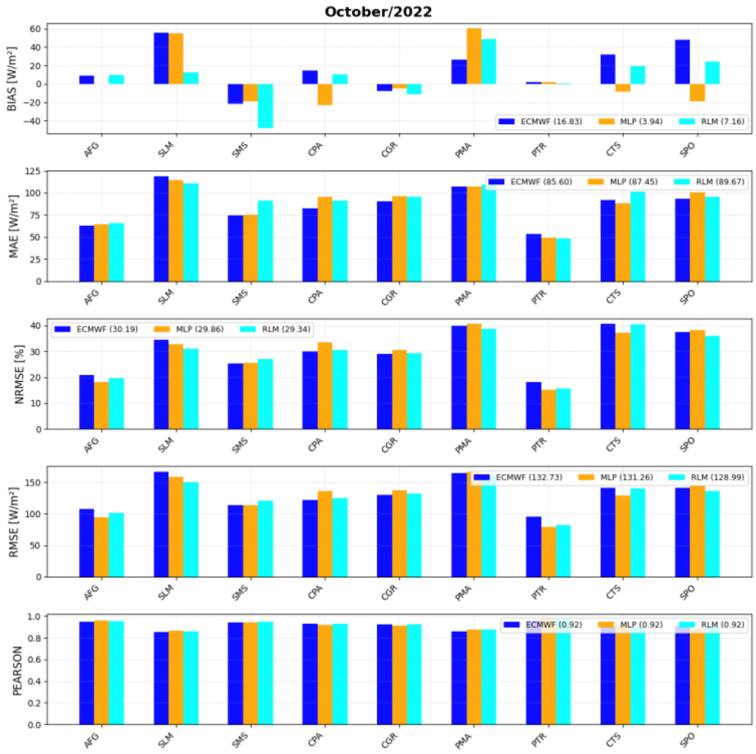

Figure 7. Comparison of the performance of the ECMWF, MLP, and MLR models in forecasting solar irradiance for different meteorological stations in October 2022. The values shown in the legend of each metric correspond to the averages calculated for all stations.

The comparison between the results obtained in this study and those reported in the literature reinforces the robustness and reliability of the forecasts generated by the analyzed models. Regarding bias, the values found for the ECMWF, MLP, and RLM models are consistent with the ranges reported by Jimenez et al. (2016), who observed biases ranging from –20 to 80 W/m² without the hybrid cloud scheme, and from –15 to 45 W/m² when it was activated. In the present study, the bias varied from –22.62 W/m² to 32.12 W/m², which aligns well with these reference intervals. The MLP model exhibited the most balanced behavior in both analyzed months, while ECMWF and RLM tended to overestimate and underestimate irradiance, respectively.

Regarding the NRMSE, the results also align with the established literature. Blaga et al. (2019) reported that numerical models generally yield NRMSE values between 30% and 45%, while Jimenez et al. (2016) demonstrated improvements of up to 46% in GHI forecasts under clear-sky conditions. In this study, NRMSE values ranged between 26% and 35%, with the MLP consistently achieving the lowest errors. For instance, in October, the MLP model recorded an NRMSE of 28%, lower than the average range reported by Blaga et al. (2019), indicating superior predictive skill. ECMWF and RLM presented slightly higher NRMSE values, confirming that the neural network achieved better overall performance.

The correlation coefficients (R) between observed and predicted values also match those found in previous studies. The correlation values ranged from 0.88 to 0.94, comparable to those reported by Jimenez et al. (2016). The MLP maintained the highest correlation (0.93) in both analyzed months, highlighting its capacity to capture complex irradiance patterns more effectively. The ECMWF and RLM models, while still showing strong correlations, displayed slightly weaker relationships, suggesting a less precise representation of the temporal dynamics of solar irradiance.

Regarding RMSE, the results fall within the expected range found in the literature. The RMSE ranged from 88.98 W/m² (MLP in May) to 154.49 W/m² (ECMWF in October), underscoring the neural network model's significant improvement. In comparison, Blaga et al. (2019) reported forecast improvements of up to 46% under clear-sky conditions, whereas the MLP in this study achieved a reduction of up to 57% in October, demonstrating its superior ability to minimize large forecast errors.

In summary, the comparative analysis confirms that the findings of this research are consistent with well-established literature. The MLP model consistently outperformed both ECMWF and RLM, showing lower errors and stronger correlations. Nevertheless, as noted by Blaga et al. (2019), increasing the size of the training and validation datasets could enhance statistical robustness and improve model generalization. Furthermore, as indicated in previous studies, extending forecasts to longer horizons introduces additional challenges due to the propagation of large-scale atmospheric systems, emphasizing the need to adapt methodologies as the forecast window expands.

Although the results indicate that the MLP model generally achieved lower errors than the other approaches, the differences between models are relatively moderate. Since this study focuses on a preliminary methodological evaluation using a limited temporal dataset, a formal statistical significance analysis was not performed. Future work should include uncertainty quantification methods, such as bootstrap confidence intervals or hypothesis testing, to more rigorously assess the significance of model performance differences.

4. Conclusions

This study improved solar irradiance forecasting in Brazil by combining ECMWF model outputs with machine learning techniques. The multilayer perceptron (MLP) model generally showed moderate improvements compared with both the multiple linear regression (MLR) model and the raw ECMWF forecasts, showing higher accuracy and lower bias and RMSE values. The comparative analysis also revealed that the ERA5 reanalysis provides better estimates of solar irradiance, serving as a valuable reference for model evaluation and as an alternative data source in regions without direct observations. However, the presence of systematic biases in both datasets highlights the need for correction methods, and the use of machine learning proved effective in reducing these uncertainties.

Future research should focus on refining the training of machine learning models for individual meteorological stations, allowing better adaptation to regional climatic conditions. The incorporation of new approaches, such as recurrent neural networks (RNNs) and satellite-derived variables, may enhance the representation of seasonal and spatial variability. These advances could lead to more robust solar forecasting systems, supporting efficient and sustainable energy planning in Brazil and contributing to the expansion of renewable energy and improved management of energy supply and demand. It should be emphasized that the methodology presented here relies on local observational datasets for calibration, which limits its direct application at the global scale. Consequently, the approach is primarily intended for site-specific forecast improvement**,** where reliable ground measurements are available.

5. Acknowledgments

The authors gratefully acknowledge the financial and technical support provided by the projects "Evaluation of the Photovoltaic Resource in Hydroelectric Reservoirs Based on Numerical Modeling and Remote Sensing" (CNPq 409711/2021-7), INCT Climate Change Phase 2 (FAPESP 2014/50848-9, CNPq 465501/2014-1, CAPES/FAPS 16/2014), and FINEP SOLTER – "Diffusion Bonding of Light Metals in Compact Solar Energy Storage Devices for Electricity Generation from Solar Thermal Sources in Brazil" (FINEP 0018/21), as well as by the Supercomputing Center for Industrial Innovation (CS2i) of SENAI CIMATEC, Eletrobras Chesf (grant CV-I-70.20.221.00), and the National Council for Scientific and Technological Development (CNPq, grant 40794/2022-4). Fernando Martins also acknowledges the São Paulo Research Foundation (FAPESP) for the financial support provided through grant 2025/15669-0.

6. References

Ahmed, R., Sreeram, V., Mishra, Y. & Arif, M.D., 2020. A review and evaluation of the state-of-the-art in PV solar power forecasting: techniques and optimization. *Renewable and Sustainable Energy Reviews*, 124, 109792.

Blaga, R., Sabadus, A., Stefu, N., Dughir, C., Paulescu, M. & Badescu, V., 2019. A current perspective on the accuracy of incoming solar energy forecasting. *Progress in Energy and Combustion Science*, 70, 119–144.

Diagne, M., David, M., Lauret, P., Boland, J. & Schmutz, N., 2013. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, 27**,** 65–76.

Ibrahim, S., Daut, I., Irwan, Y.M., Irwanto, M., Gomesh, N. & Farhana, Z., 2012. Linear regression model in estimating solar radiation in Perlis. *Energy Procedia*, 18**,** 1402–1412.

International Renewable Energy Agency (IRENA), 2020. *Renewable Capacity Statistics 2020.* Abu Dhabi: IRENA. ISBN 978-92-9260-239-0.

Jimenez, P.A., Hacker, J.P., Dudhia, J., Haupt, S.E., Ruiz-Arias, J.A., Gueymard, C.A. & Deng, A., 2016. WRF-Solar: Description and clear-sky assessment of an augmented NWP model for solar power prediction. *Bulletin of the American Meteorological Society*, 97(7), 1249–1264.

Lima, F.J.L., Martins, F.R., Pereira, E.B., Lorenz, E. & Heinemann, D., 2016. Forecast for surface solar irradiance at the Brazilian northeastern region using NWP model and artificial neural networks. *Renewable Energy*, 87, 807–818.

Martins, F.R., Pereira, E.B., Silva, S.A.B., Abreu, S.L. & Colle, S., 2008. Solar energy scenarios in Brazil. Part 1: Resource assessment. *Energy Policy*, 36, 2853–2864.

Martins, F.R., Rüther, R., Abreu, S.L. & Pereira, E.B., 2012. Scenarios for solar thermal energy applications in Brazil. *Energy Policy*, 48, 640–649.

Marzouq, M., El Fadili, H., Zenkouar, K., Lakhliai, Z. & Amouzg, M., 2020. Short-term solar irradiance forecasting via a novel evolutionary multi-model framework and performance assessment for sites with no solar irradiance data. *Renewable Energy*, 157, 214–231.

Nascimento, P.S.C., Lima, M.C.A., Junior, C.A.A., Oliveira, A.M.R., Madeira, F., Mendes, E.L. & Fontoura Filho, R.N., 2017. Desenvolvimento e implantação no ONS de um modelo de previsão de geração de fonte eólica com customização temporal. *Brazil Windpower Conference Proceedings*, Rio de Janeiro, Brazil.

Obiora, C.N., Hasan, A.N., Ali, A. & Alajarmeh, N., 2021. Forecasting hourly solar radiation using artificial intelligence techniques. *IEEE Canadian Journal of Electrical and Computer Engineering*, 44**(**4), 497–508.

Pedruzo-Bagazgoitia, X., Jiménez, P.A., Dudhia, J. & De Arellano, J.V.G., 2019. Shallow cumulus representation and its interaction with radiation and surface at the convection gray zone. *Monthly Weather Review*, 147(7), 2467–2483.

Pereira, E.B., Martins, F.R., Gonçalves, A.R., Costa, R.S., Lima, F.J.L., Rüther, R., Abreu, S.L., Tiepolo, G.M., Pereira, S.V. & Souza, J.G., 2017. *Atlas Brasileiro de Energia Solar*, 2nd ed. São José dos Campos: INPE. ISBN 978-85-17-00089-8.

Qiu, R., Liu, C., Cui, N., Gao, Y., Li, L., Wu, Z. & Hu, M., 2022. Generalized extreme gradient boosting model for predicting daily global solar radiation for locations without historical data. *Energy Conversion and Management*, 258, 115488.

Wilks, D.S., 2020. *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Academic Press.

Wild, M., 2011. The global energy balance as represented in CMIP6 climate models. *Climate Dynamics*, 55(3), 553–577.

Zucareli, V.L., 2024. *Balanço energético e avaliação econômica de um sistema fotovoltaico para uma indústria de processamento e armazenamento de pescados.* Universidade Estadual Paulista (UNESP).