



PHOTOVOLTAIC GENERATION FORECASTING CONSIDERING HEATWAVES VIA XGBOOST WITH ENHANCED TEMPERATURE AND IRRADIANCE ESTIMATION

Cristhian A. Lopez Castillo¹, Rubén D. Cardona Ruiz, Maria C. Valverde Brambila¹, Joel David Melo Trujillo¹

¹ Federal University of ABC, UFABC – anderson.lopez@ufabc.edu.br, ruben.cardona@ufabc.edu.br, maria.brambila@ufabc.edu.br, joel.melo@ufabc.edu.br

ABSTRACT

Heat waves, which are increasingly frequent due to climate change, negatively impact photovoltaic generation by reducing module efficiency. This paper presents an enhanced forecasting method based on the XGBoost (Extreme Gradient Boosting) algorithm to improve the estimation of air temperature and solar irradiance during extreme heat events. A binary indicator, derived from historical thresholds, is integrated into the model to explicitly capture the effects of heat waves on photovoltaic performance. Using meteorological data from Santo André, Brazil, results show a reduction in mean absolute error for temperature from 1.71 °C to 1.64 °C and for irradiance from 16.99 W/m² to 16.53 W/m² compared to a baseline model without the indicator. These findings highlight the potential of incorporating extreme climate indicators to enhance forecasting accuracy and support more reliable energy planning in the face of climate variability.

Keywords: XGBoost; Solar irradiance; Temperature; Heat waves.

1 INTRODUCTION

In recent years, the intensification of heat waves —characterized by persistently high temperatures over several days —has become an increasingly frequent phenomenon due to climate change. In the electricity sector, these extreme events create a dual impact: they increase electricity demand due to the intensive use of cooling systems and reduce the efficiency of generation technologies, such as photovoltaic solar power (STAFFELL; PFENNINGER; JOHNSON, 2023). In particular, elevated temperatures can significantly raise the operating temperature of photovoltaic modules, reducing their performance and altering the effective irradiance received (SEGBEFIA, 2023; HUANG et al., 2024). In this context, the central challenge of this work is to improve the estimation of critical variables, such as air temperature and solar irradiance, during heatwave episodes, to increase the accuracy of photovoltaic generation forecasting. To address this challenge, this paper presents a method based on the XGBoost algorithm, capable of efficiently modeling the behavior of these critical variables (ZHANG et al., 2024). The technique incorporates a binary heatwave indicator, calculated from historical thresholds, which enables the model to explicitly capture the effects of such extreme events on generation. Additionally, the proposal incorporates advanced techniques for preprocessing meteorological data, transforming cyclic variables, and standardization, thereby ensuring greater robustness in the results. The model's performance is evaluated through consolidated metrics such as MAE (Mean Absolute Error), RMSE (Root Mean Square Error), and R² (coefficient of determination), reinforcing the effectiveness of the approach in supporting more resilient energy planning in the face of climate variability and the increasing frequency of extreme events.



2 THE PROPOSED METHOD

A. Algorithm XGBOOST

The proposed method uses XGBoost (*Extreme Gradient Boosting*), recognized for modeling complex nonlinear relationships, resisting overfitting, and being computationally efficient (LI et al., 2022; XU et al., 2024; ZHANG, 2024). The algorithm estimates air temperature (T_a) and solar irradiance (G) to improve photovoltaic generation forecasting. The model, composed of decision trees trained sequentially, corrects the errors of previous trees and allows the prediction of T_a and G in independent models while sharing the same input structure (Eq. 1).

$$\hat{\mathcal{Y}}_i = \begin{bmatrix} \hat{T}_{a,i} \\ \hat{G}_i \end{bmatrix} = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F} \quad \text{Eq. (1)}$$

Where $\hat{\mathcal{Y}}_i$ is the prediction of air temperature and solar irradiance for the i -th observation, x_i contains the input variables (meteorological data), K is the number of trees in the ensemble, and $\mathcal{F}$ is the set of possible trees that XGBoost uses to minimize error and improve accuracy even under extreme conditions.

The objective function of XGBoost minimizes the error between the observed values and the model predictions, incorporating regularization terms to avoid overfitting, as shown in Eq. (2):

$$\mathcal{L}(\phi) = \sum_{i=1}^n l(\mathcal{Y}_i, \hat{\mathcal{Y}}_i) + \sum_{k=1}^K \Omega(f_k) \quad \text{Eq. (2)}$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad \text{Eq. (3)}$$

Where l is a differentiable loss function that measures the error between the observed value $\mathcal{Y}_i$ and the prediction $\hat{\mathcal{Y}}_i$, and $\Omega(f_k)$ is the regularization term for the k -th tree. T is the number of leaves of the tree, ω_j is the weight associated with leaf j , γ is the tree complexity regularization parameter, and λ controls the strength of regularization by the magnitude of the weights.

B. Integration of the Heat Wave (HW) Indicator into the Model

To address the impacts of heat waves (HW), XGBoost incorporates a binary HW indicator (I_{OC}), which takes the value 1 when the observation x_i corresponds to a heatwave period and 0 otherwise (Eq. 5). The threshold is defined by the historical average maximum temperature $T_{max,OC}$ (Eq. 6), where N_{OC} represents the number of days classified as HW. By including I_{OC} in the predictions, the model adjusts $\hat{\mathcal{Y}}_i$ for air temperature and solar irradiance, accounting for HW effects and improving the accuracy of photovoltaic generation forecasting under extreme events (Eq. 7).

$$I_{OC}(x_i) = \begin{cases} 1, & \text{if } x_i \text{ corresponds to a heatwave period} \\ 0, & \text{otherwise} \end{cases} \quad \text{Eq. (5)}$$

$$T_{max,OC} = \frac{1}{N_{OC}} \sum_{i \in OC} T_{max,i} \quad \text{Eq. (6)}$$

$$\hat{\mathcal{Y}}_i = \sum_{k=1}^K f_k(x_i, I_{OC}(x_i)) \quad \text{Eq. (7)}$$



C. Processing of Variables

Temporal information, namely month and hour, was encoded cyclically to capture periodic patterns throughout the year and the day (Eqs. 8 and 9). Wind speed was logarithmically transformed to stabilize its variance (Eq. 10), while wind direction, being cyclic, was represented in sine and cosine components to preserve angular continuity (Eq. 11). Atmospheric pressure (p) was standardized with zero mean (μ) and unit standard deviation (σ_p), ensuring comparability among the variables (Eq. 12).

$$Mes_{sin} = \sin\left(\frac{2\pi \cdot Mes}{12}\right), Mes_{cos} = \cos\left(\frac{2\pi \cdot Mes}{12}\right) \quad \text{Eq. (8)}$$

$$Hora_{sin} = \sin\left(\frac{2\pi \cdot Hora}{24}\right), Hora_{cos} = \cos\left(\frac{2\pi \cdot Hora}{24}\right) \quad \text{Eq. (9)}$$

$$v_{transf} = \log(v + 0.1) \quad \text{Eq. (10)}$$

$$dir_{sin} = \sin\left(\frac{2\pi \cdot dir_v}{360}\right), dir_{cos} = \cos\left(\frac{2\pi \cdot dir_v}{360}\right) \quad \text{Eq. (11)}$$

$$p_{std} = \frac{p - \mu_p}{\sigma_p} \quad \text{Eq. (12)}$$

D. Model Evaluation

The performance of the models was assessed using the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). R^2 indicates the proportion of variance explained by the model, MAE measures the average magnitude of prediction errors, and RMSE emphasizes larger deviations. These metrics together provide a clear evaluation of model accuracy and reliability.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad \text{Eq. (13)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \text{Eq. (14)}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad \text{Eq. (15)}$$

where $\bar{y}$ is the mean of the observed values and n is the total number of samples.

E. Photovoltaic System

The Photovoltaic System is characterized based on panel and inverter data, considering irradiance, temperature, and operational limits of grid connection. The DC power (P_{dc}) generated is obtained from the nominal module power (P_{mpp}), adjusted by irradiance and the temperature curve, as shown in Eq. (16), while the AC power (P_{ac}) injected corresponds to the DC power adjusted by the inverter efficiency curve (Eq. 17)

$$P_{dc} = P_{mpp} \times irradiance \times PTcurve(Temperature) \quad \text{Eq. (16)}$$

$$P_{ac}(t) = P_{dc}(t) \times EffCurve(P_{dc}(t)) \quad \text{Eq. (17)}$$

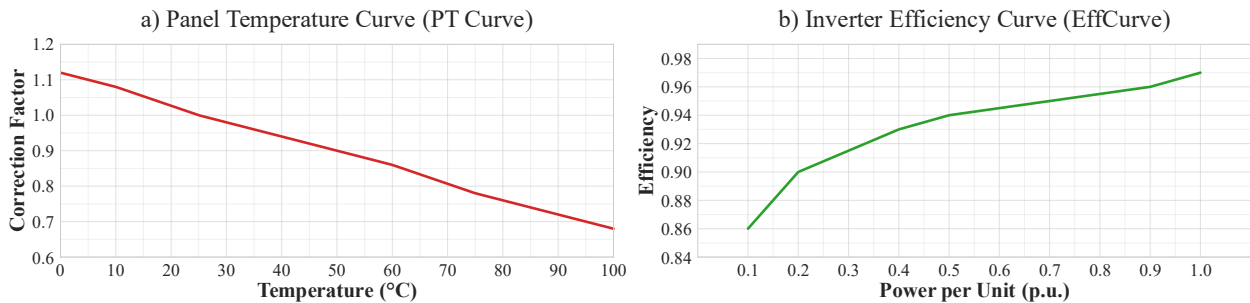


3 RESULTS AND DISCUSSION

A. Input Parameters

For training the estimation models, the Python XGBoost library was used, with a learning rate of 0.1, a maximum depth of 6, 200 trees, subsampling, and column fraction of 80%, and a fixed seed (42) to ensure reproducibility. The meteorological dataset from the UFABC solarimetric station, covering the period from January 7, 2023, to June 30, 2025, included 70,176 records with a 15-minute resolution. For validation, the data were split into 80% for training and 20% for testing, while preserving temporal sequence. For the case study analysis, the UFABC photovoltaic system was considered, focusing on a single inverter with $P_{mpp} = 55 \text{ kW}$, whose characteristic curves are shown in Figure 1: Figure 1a presents the module temperature correction curve (PT_{curve}) and Figure 1b shows the inverter efficiency curve (Eff_{Curve}).

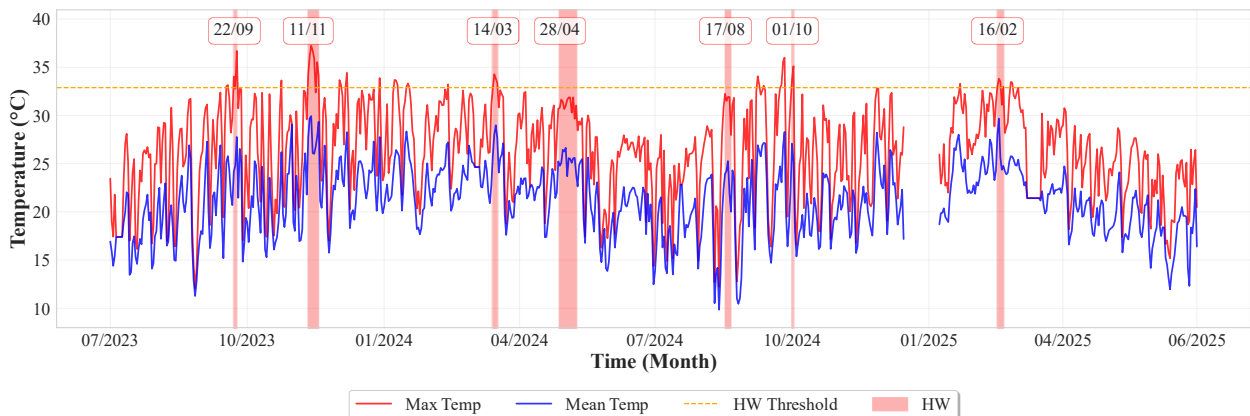
Figure 1 – Photovoltaic System Curves



B. Analysis of Heatwave Events

Between July 1, 2023, and June 30, 2025, historical HW events were identified based on climatological bulletins from Santo André (Valverde, 2024). The temperature threshold for each hardware event was determined from historical data, enabling the classification of critical periods. During this period, 7 events were recorded, totaling 39 days with an average maximum temperature of 32.89 °C, providing the basis for assessing the impact of extreme conditions on solar generation and meteorological variables.

Figure 2 – Historical Daily Temperature Curve

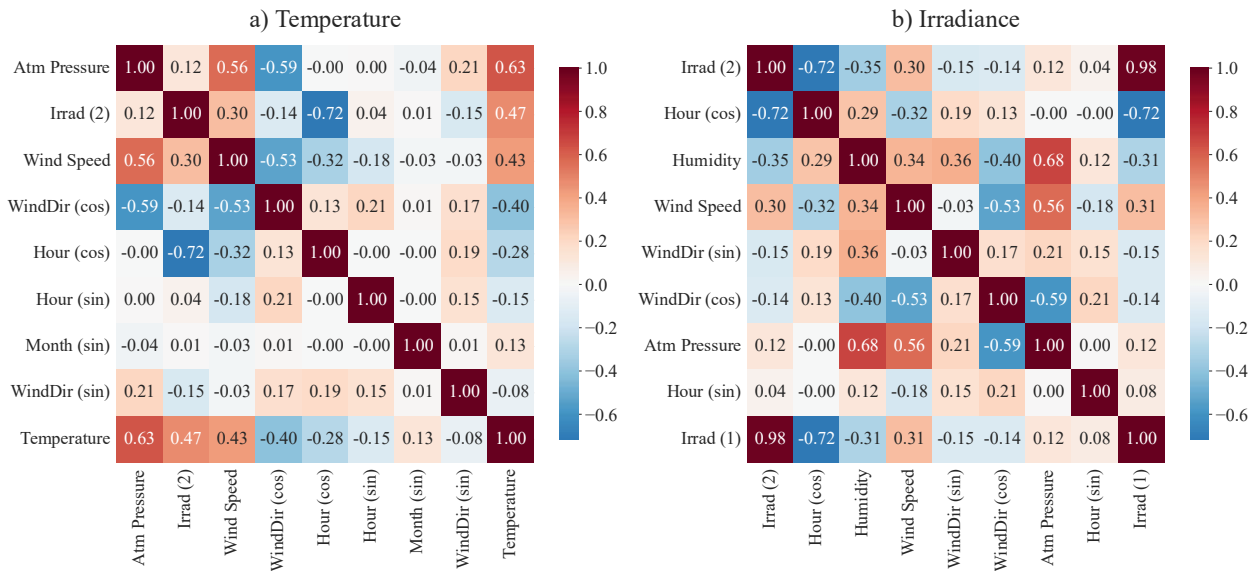




C. Analysis of Correlations of Meteorological Variables

Temporal variables (Eqs. 8 and 9) and meteorological variables (Eqs. 10–12) were transformed, generating eleven features for the correlation matrix of the most relevant ones. In Figure 3a, air temperature showed a negative correlation with relative humidity ($r = -0.72$) and with atmospheric pressure ($r = -0.55$), and a positive correlation with irradiance ($r = 0.52$), in agreement with heating processes. In Figure 3b, irradiance presented high consistency between sensors ($r = 0.98$), a negative correlation with humidity ($r = -0.53$) and with the hourly cycle ($r = -0.72$), as well as a positive correlation with wind speed ($r = 0.30$), reflecting its dependence on the diurnal cycle and on clear atmospheric conditions.

Figure 3 – Correlation Matrix of Temperature and Irradiance Variables



D. Evaluation of the Study Models

The proposed XGBoost model was developed and compared with a baseline version, resulting in two configurations: Model 1, without the heatwave indicator (I_{OC}), and model 2, which is part of the proposed method and includes it as an additional variable (Eq. 7). The performance evaluation for solar irradiance and ambient temperature estimation, summarized in Table 1, shows that Model 2 outperforms Model 1, with more evident improvements in temperature forecasting, highlighting the positive impact of the heatwave indicator.

Table 1 – Comparative Summary of the Study Models

Variable	Model	Training			Validation		
		RMSE	R ²	MAE	RMSE	R ²	MAE
Air Temperature	Model 1	0.9379	0.9771	0.6667	2.1469	0.7251	1.7154
	Model 2	0.9121	0.9783	0.6495	2.1680	0.7197	1.6406
Solar Irradiance	Model 1	12.1701	0.9981	6.7067	30.7783	0.9872	16.9921
	Model 2	11.8285	0.9982	6.2711	29.5407	0.9882	16.5340

The results demonstrate that both models exhibit excellent predictive capability, with coefficients of determination (R^2) higher than 0.97 for all variables in the training set. For air temperature estimation, Model 2 shows consistent

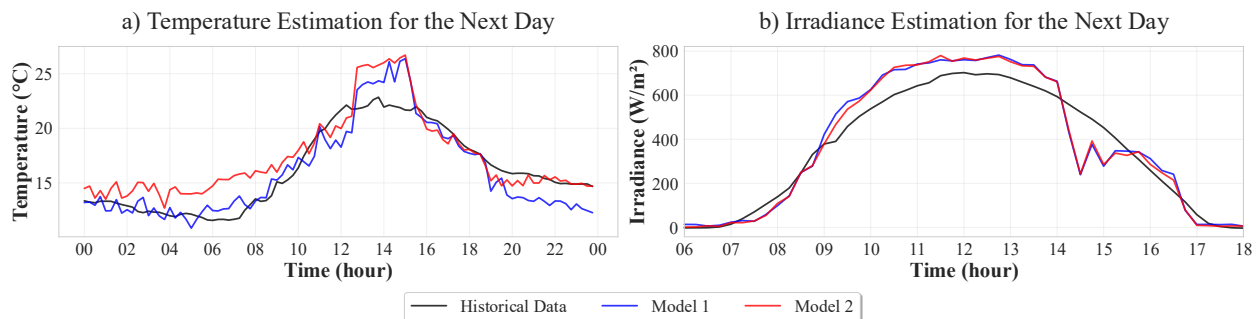


improvements across all evaluation metrics in training, with reductions of 2.75% in RMSE and 2.58% in MAE compared to Model 1. In validation, although R^2 is slightly lower (0.7197 vs. 0.7251), Model 2 maintains a significant 4.37% reduction in MAE, indicating greater accuracy in predicting individual values. For solar irradiance, the improvements of Model 2 are more subtle yet consistent, with reductions of 2.81% in RMSE and 6.50% in MAE during training, and improvements of 4.00% in RMSE and 2.69% in MAE during validation. The incorporation of the heatwave indicator into Model 2 yields a positive contribution to predictive capability, particularly for thermal variables, while maintaining good generalization, as evidenced by the high R^2 values observed in validation.

E. Estimation of Irradiance and Temperature

This section also compares the previously proposed Models 1 and 2. During the estimation of variables for the third week of July, the XGBoost model, including the heatwave indicator, showed variable-specific effects. For air temperature, R^2 slightly decreased from 0.7251 to 0.7197 and RMSE increased marginally from 2.15 to 2.17, while MAE improved from 1.72 to 1.64. This indicates a reduction in the average error without a significant change in explanatory power, likely because recorded temperatures during this period were relatively low. For solar irradiance, slight improvements were observed: R^2 increased from 0.9872 to 0.9882, and RMSE and MAE decreased to 29.54 and 16.53, respectively, showing consistent albeit small adjustments. As illustrated in Figure 4, the estimation curves for the first day reflect these trends. In Figure 4a, Model 2 ($17.35^\circ\text{C} \pm 3.70^\circ\text{C}$) closely follows the historical temperature trend ($16.21^\circ\text{C} \pm 3.64^\circ\text{C}$), particularly at peak values, compared to Model 1 ($15.97^\circ\text{C} \pm 4.11^\circ\text{C}$). In Figure 4b, solar irradiance estimates show that both models slightly exceed the historical peak of 702.4 W/m^2 , reaching 782.0 W/m^2 (Model 1) and 780.0 W/m^2 (Model 2), with average values slightly above the observed data.

Figure 4 – Estimation of the subsequent week for temperature



F. Generation Estimation

Models 1 and 2, as previously defined, were used to estimate photovoltaic generation for the period from July 15 to July 21, 2025, in Santo André, considering only solar hours (06:50–17:30). Daily production remained consistent between the models, with minor variations in average hourly power (-1.4 to $+0.6\text{ kW}$), indicating a limited effect of heatwaves. Total energy was $1,517.1\text{ kWh}$ for Model 1 and $1,509.6\text{ kWh}$ for Model 2, with daily averages of 216.7 and 215.7 kWh/day , respectively, resulting in a -0.5% difference (-7.5 kWh). Figure 5 shows the generation curves for the analyzed week, reflecting the close alignment of both models.



Figure 6 – Photovoltaic Generation Estimation

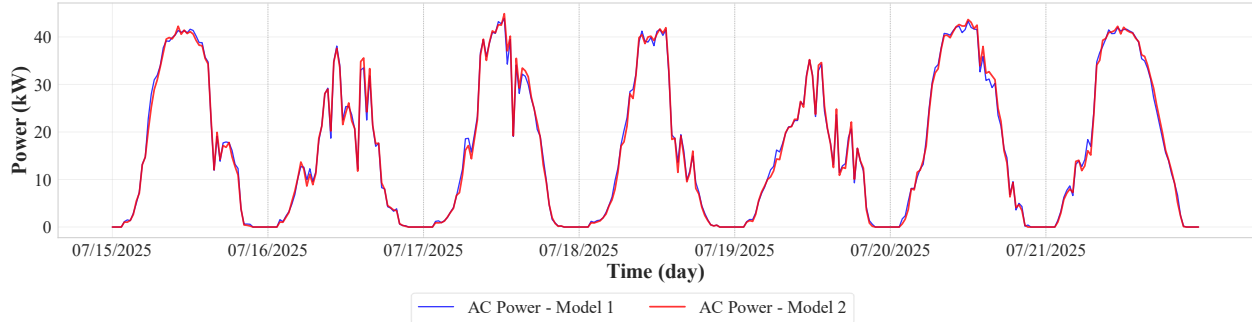


Table 2 presents the daily photovoltaic generation estimate for the analyzed period, with a maximum on July 20 (260.2 kWh) and a minimum on July 15 (165.2 kWh). Model 1 ranged from 165.2 to 258.8 kWh, while Model 2 varied from 166.1 to 260.2 kWh. The daily differences were small, ranging from -1.9% to 0.6% , indicating that the inclusion of the heatwave indicator did not significantly alter the estimated energy, demonstrating high consistency between the models.

Table 2 – Results of Photovoltaic Generation Estimation

Data	Model 1 (kWh)	Model 2 (kWh)	Difference	%Change
July 15, 2025	248.6	244.5	-4.2	-1.7%
July 16, 2025	165.2	166.1	0.9	0.6%
July 17, 2025	218.1	217.7	-0.4	-0.2%
July 18, 2025	194.2	190.5	-3.7	-1.9%
July 19, 2025	173.9	172.5	-1.3	-0.8%
July 20, 2025	258.8	260.2	1.3	0.5%
July 21, 2025	258.2	258.1	-0.2	-0.1%

G. Discussion

The incorporation of the heatwave indicator into Model 2 produced moderate improvements in the predictive capability of air temperature and solar irradiance. For temperature, Model 2 achieved reductions of 2.75% in RMSE and 2.58% in MAE during training, and a 4.37% improvement in MAE in validation, although R^2 decreased slightly (0.7197 vs. 0.7251). For irradiance, the improvements were smaller but consistent, with reductions of up to 6.50% in MAE. The correlations among variables remained in accordance with expected physical patterns: negative correlation of temperature with humidity ($r = -0.72$) and pressure ($r = -0.55$), and strong consistency of irradiance between sensors ($r = 0.98$). During the third week of July, Model 2 more accurately reflected temperature trends, while photovoltaic generation showed minimal differences (-1.9% to 0.6%), indicating that the indicator improves precision without affecting model consistency.

The differences observed in July 2025 may be associated with the atypical climatic conditions of that month, marked by the predominance of cold air masses, low temperatures (minimum of $9.1\text{ }^{\circ}\text{C}$), and extreme phenomena such as cold waves. This highlights the importance of evaluating model performance under infrequent situations or extreme events, such as cold waves, to ensure robustness and reliability



4 CONCLUSIONS

The results demonstrate that the inclusion of the heatwave indicator in the XGBoost model significantly enhances predictive capability, particularly for ambient temperature, with a reduction of 2.75% in RMSE and 2.58% in MAE during training, in addition to a 4.37% improvement in MAE during validation. This indicates greater accuracy in estimating individual values under extreme thermal conditions, which translates into more reliable forecasts for the sizing of cooling systems and energy demand management. For solar irradiance, although the improvements are more subtle, consistent reductions of 2.81% in RMSE and 6.50% in MAE were observed during training, with improvements of 4.00% in RMSE and 2.69% in MAE during validation. These results demonstrate an enhanced ability to anticipate variations in PV generation during extreme events, a critical aspect for the stability of the power grid. The developed methodology achieves its main objective of explicitly incorporating the impact of heatwaves into critical meteorological variables, becoming essential for energy operators who require accurate forecasts under adverse conditions. This approach effectively contributes to resilient energy planning strategies and climate risk mitigation in a context of intensifying extreme weather phenomena, where the precise anticipation of extreme conditions is essential to prevent systemic failures and ensure the continuity of energy supply.

ACKNOWLEDGMENTS

The authors would like to thank the Federal University of ABC (UFABC) for institutional support. This work was financially supported by the São Paulo Research Foundation (FAPESP), under grants #2024/14894-8, #2025/03963-1, #2025/07372-8 and #2025/02554-0; by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Funding Code 001; and by the National Council for Scientific and Technological Development (CNPq), through the Universal Project no. 407244/2023-9 and 309013/2025-9.

REFERENCES

- HUANG, Z. et al. Response of sustainable solar photovoltaic power output to summer heatwave events in Northern China. *Sustainability*, v. 16, n. 12, p. 5254, 2024. DOI: <https://doi.org/10.3390/su16125254>
- LI, X. et al. Probabilistic solar irradiance forecasting based on XGBoost. *Energy Reports*, v. 8, p. 1087-1100, 2022. DOI: <https://doi.org/10.1016/j.egy.2022.02.251>
- SEGBEFIA, O. K. Temperature profiles of field-aged photovoltaic modules affected by optical degradation. *Heliyon*, v. 9, n. 9, e19566, 2023. DOI: <https://doi.org/10.1016/j.heliyon.2023.e19566>
- STAFFELL, I.; PFENNINGER, S.; JOHNSON, N. A global model of hourly space heating and cooling demand under climate change scenarios. *Nature Energy*, v. 8, 2023. DOI: <https://doi.org/10.1038/s41560-023-01341-5>
- VALVERDE, M. C. *Boletim Climatológico de Santo André – Universidade Federal do ABC*. v. 5, ed. 1–12, 2024. Disponível em: <<https://portais.santoandre.sp.gov.br/defesacivil/boletim-climatologico-mensal/>>. Acesso em: 18 ago. 2025.
- XU, W. et al. Short-Term Photovoltaic Output Prediction Based on Decomposition and Reconstruction and XGBoost under Two Base Learners. *Energies*, v. 17, n. 4, p. 906, 2024. DOI: <https://doi.org/10.3390/en17040906>
- ZHANG, X. et al. Prediction of photovoltaic power generation based on a hybrid model. *Frontiers in Energy Research*, 2024.