IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 June 2025, accepted 20 June 2025, date of publication 15 July 2025, date of current version 22 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3589275

RESEARCH ARTICLE

Artificial Intelligence in Data Science: Evaluating Forecasting Models for Solar Energy in the Amazon Basin

ANDRÉ LUIS FERREIRA MARQUES[1], RICARDO SBRAGIO[2], PEDRO LUIZ PIZZIGATTI CORRÊA[1], AND MARCELO RAMOS MARTINS[2]

[1]BigData Lab C216, Computer Engineering Department, Polytechnic School of the University of São Paulo, São Paulo 05508-010, Brazil
[2]Laboratory of Analysis, Evaluation and Risk Management (LabRisco), Naval Architecture and Ocean Engineering Department, Polytechnic School of the University of São Paulo, São Paulo 05508-010, Brazil

Corresponding author: André Luis Ferreira Marques (alfmarques@alumni.usp.br)

This work was supported in part by Brazilian National Council for Scientific and Technological Development (CNPq) under Grant 141469/2024-2 and Grant 303908/2022-0; in part by the Human Resources Program of Brazilian National Agency for Petroleum, Natural Gas and Biofuels (PRH-ANP); in part by São Paulo Research Foundation (FAPESP), Brazil, under Grant 2024/10537-6, Grant 2024/12694-1, and Grant 2024/12663-9; in part by the Research Centre for Greenhouse Gas Innovation (RCGI) hosted by the University of São Paulo (USP) and sponsored by FAPESP and Shell Brasil under Grant 2014/50279-4, Grant 2020/15230-5, and Grant 2022/07974-0; and in part by ANP through the Research and Development Regulations.

ABSTRACT Forecasting models employing machine learning (ML) and deep learning (DL) have become fundamental for assessing the technical feasibility of renewable energy systems. Among these, solar energy stands out as a renewable energy option, particularly relevant for supporting the preservation of the Amazon rainforest. This study introduces a novel approach using ML and DL methods—integrated with Universal Kriging and Holt-Winters (time series) models — to forecast solar irradiance (kWh/m^2) in cities across the state of Amazonas. The analysis is grounded in the Data Science cycle, with input data sourced from both ground stations and satellite products. Forecasting performance was evaluated for short-term horizons (one to three days ahead) across three representative cities. The hybrid SARIMAX-CNN-LSTM, SARIMAX-CNN-Transformer, and SARIMAX-TCN models achieved MAPE values ranging from 18.1% to 26.6% for the different forecast horizons and cities. These results are consistent with existing literature and reinforce the suitability of advanced ML/DL approaches for solar energy forecasting in highly variable and challenging environments such as the Amazon Basin.

INDEX TERMS Deep learning, long short-term memory, multi-layer perceptron, data science, Amazon Basin, solar energy, kriging metamodel.

LIST OF MAIN ABBREVIATIONS

The main abbreviations used throughout this work are listed in Table 1.

I. INTRODUCTION

This research explores the intersection of renewable energy and Data Science (DS), focusing on the use of Machine Learning (ML) and Deep Learning (DL) techniques to forecast solar energy use. As the impacts of climate change, such as widespread wildfires, floods, and prolonged droughts, become more pronounced, a range of actions have been initiated. These include enhanced atmospheric monitoring through satellite networks, drones, and ground stations, along with broader efforts in data and information dissemination.

A key area of focus in these efforts is the monitoring of greenhouse gases (GHGs), particularly those emitted from the use of fossil fuels, which are widely recognized as the primary drivers of global temperature rise. As a result, renewable energy sources have gained significant attention and development due to their lower carbon footprint and reduced Carbon Dioxide (CO_2) emissions [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

TABLE 1. List of main abbreviations.

Abbreviation	Term / Meaning
AWS	Amazon Web Services
CNN	Convolutional Neural Network
D+1, D+2, D+3	Forecast horizon: Day +1, Day +2, Day +3
DL	Deep Learning
DS	Data Science
DT	Decision Tree
DWT	Discrete Wavelet Transform
EDA	Exploratory Data Analysis
EMD	Empirical Mode Decomposition
GHG	Greenhouse Gas
GRU	Gated Recurrent Unit
ISO	International Organization for Standardization
INMET	Brazilian National Institute of Meteorology
K-means	K-means Clustering Algorithm
Kriging	Kriging Metamodel (Geostatistical Interpolation)
LSTM	Long Short-Term Memory (Neural Network)
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MODWPT	Maximal Overlap Discrete Wavelet Packet Transform
NASA	National Aeronautics and Space Administration
NN	Neural Network
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit (activation function)
RMSE	Root Mean Square Error
SARIMAX	Seasonal AutoRegressive Integrated Moving Average with Exogenous Regressors
STL	Seasonal-Trend decomposition via Loess
TCN	Temporal Convolutional Network
VMD	Variational Mode Decomposition

Among these renewable energy sources, solar energy has seen rapid adoption, largely due to its relatively quick installation process and ease of operation and maintenance.

The rainforests present a unique challenge, with the Amazon Basin experiencing an alarming rate of deforestation, despite increasing environmental monitoring and preservation efforts [2]. Solar energy has been introduced in the Amazon Basin to enhance the quality of life for local communities while simultaneously reducing reliance on fossil fuels in the region.

From a technical perspective, the Amazon Basin is characterized by a challenging environment, with air humidity often exceeding 90%, temperatures ranging from 25^{o}C to 40^{o}C (or above), dense vegetation, and a rich diversity of fauna and flora [3]. These conditions complicate the management and collection of data from ground stations or the use of drones, even with the increasing intensity of scientific research in the area. Additionally, existing ground stations have encountered technical difficulties, often requiring human intervention in most of the surrounding cities, due to the recent fires in the rainforest.

A. RELATED WORKS

Data Science (DS) has increasingly leveraged mathematical algorithms and tools, including Time Series (TS) analysis, ML, DL, and alternative methods such as Kriging metamodels [4], [5], for a variety of applications. The growing utilization of these tools is largely due to advancements in computing power and cloud infrastructure. A review of methods used in solar irradiance forecasting, as conducted by [6], offers valuable insights into this research area. Data evaluations have demonstrated correlations between solar radiation and atmospheric/meteorological variables, such as air temperature and cloud cover.

For instance, local temperature significantly influences solar panel performance due to material engineering factors. Consequently, forecasting solar energy levels has garnered increased attention, particularly in technical and feasibility assessments, including those related to the economics of renewable energy options. Typically, forecasting time spans cover shorter intervals, ranging from minutes to hours, due to the nature of solar phenomena. However, there are also instances where forecasts are made for day-ahead (short-term), monthly (medium-term), and yearly (long-term) periods, serving various objectives in energy market decision-making processes.

The literature review revealed a limited number of references focusing on forecasting solar irradiance using the Holt-Winters method. The research by [7] explores short-term solar power forecasting for photovoltaic (PV) systems, identifying the number of days and the weight parameter as critical factors, with analysis conducted using Matlab. Similarly, [8] addresses PV power generation through TS analysis and the application of the Holt-Winters method. This work builds upon insights from [9], which evaluates TS forecasting applications. Additionally, [10] applied the Holt-Winters method in their research. Moreover, [11] compared the Holt-Winters and Seasonal Variation methods for forecasting PV generation using real data from isolated rural areas in Ecuador. The study concluded that while both methods performed well in forecasting, the Holt-Winters model demonstrated superior accuracy.

On the other hand, due to the variability of the data involved, Artificial Intelligence (AI), particularly Neural Networks (NN), has offered a range of options, allowing for the integration of diverse data sources and even the combination of different network architectures. Given the nature of the physical phenomena involved, the Long Short-Term Memory (LSTM) neural network has been widely adopted. General insights into this field are provided by [12], [13], [14], [15], and [16], contributing significantly to the understanding of solar energy forecasting. Neural networks can be configured in various ways, with several key types being particularly relevant to solar energy forecasting. Table 2 provides a summary of the main models used.

Optimization techniques are also frequently applied in neural networks, including methods such as Dropout [29] to prevent overfitting, Adam (an optimized algorithm for stochastic gradient descent), Rectified Linear Unit (ReLU), greedy algorithms (incremental solutions), cross-validation, adaptive and colony algorithms, Grid Search, Maximal Overlap Discrete Wavelet Packet Transform (MODWPT) [30], Nesterov-accelerated Adaptive Moment Estimate (Nadam), and backward elimination algorithms, among others.

TABLE 2. Different configurations of neural networks.

Model Type	Description	Reference
Bidirectional Neural Networks	Use two types of hidden layers: one for processing data in the forward direction and another for handling backward data flow.	[17]
Deep Recurrent Neural Network (DRNN)	The output of one step is used as input for the next. It employs a hidden state that retains information from previous steps.	[18]
Gated Recurrent Unit (GRU)	Similar to RNNs/DRNNs, includes features like a forget gate. GRUs are akin to LSTMs and often used in signal processing.	[19]
Convolutional Neural Network (CNN)	A feedforward neural network that learns feature engineering through filter optimization.	[20]
Multi-Layer Perceptron (MLP)	Uses a non-linear activation function and consists of at least three layers. Effective for non-linearly separable data.	[21]
Deep Belief Network + Feedforward Neural Network (DBN + FNN)	Combines generative graphical models with multiple layers of latent variables.	[22]
LSTM + Deep Generative Model (DGM)	Hybrid model using unsupervised methods to describe phenomena based on data within a neural network.	[23]
Discrete Wavelet Transform (DWT) + LSTM	Uses signal decomposition for neural networks, with the decomposed signal forming multiple time series within a frequency band.	[24]
Stationary Wavelet Transform (SWT) + LSTM + DNN	Similar to DWT + LSTM, but addresses the lack of translation invariance in DWT.	[25]
Wavelet Packet Decomposition (WPD) + LSTM	Related to DWT but uses additional data filters for improved decomposition.	[26]
DBN + Autoregressive Integrated Moving Average (ARIMA)	Combines a neural network with a time series model using moving average techniques.	[27]
Genetic Algorithm Network (GAN) + CNN	Applies Darwinian selection-inspired algorithms to optimize neural network weights, offering an alternative to backpropagation.	[28]

Another option for forecasting solar irradiance is the Kriging metamodel. Kriging is a machine learning technique commonly used in the Design and Analysis of Computer Experiments (DACE) as a surrogate model to approximate computer experiment data [4], [5], [31], [32]. Originally developed for geostatistical analysis [33], Kriging has been applied across various fields, including the simulation and optimization of engineering problems [34], [35], [36], [37], [38], [39], cost estimation of engineering systems [40], risk assessment and reduction [41], spatial interpolation of chemical concentrations [42], water quality monitoring [43], estimation of geological features [44], and precision agriculture applications [45], among others. In this research, Kriging is applied to simulate an engineering and environmental problem related to solar radiation forecasting.

In a study by [46], the performance of decomposition models for diffuse solar radiation and variogram models for the Kriging process was analyzed. This investigation involved downscaling satellite data to characterize the spatial dependence of solar radiation in a region with complex topography. Studies by [47] and [48] employed the ordinary Kriging method to analyze available data on the temporal and spatial distribution of solar radiation, enabling interpolation at a specific location of interest. The research conducted by [49] applied Co-Kriging to enhance the accuracy of solar insolation predictions derived from meteorological network and satellite imagery datasets.

In [50], the potential of the Kriging method for spatial prediction of solar irradiance is demonstrated. The research by [51] quantified the short-term variability of solar irradiance at a specific area of a photovoltaic power plant using a Kriging method and applied it to optimize the number of sensors required for reliable prediction. The study by [52] applied Kriging for spatio-temporal forecasting of solar irradiance using input data from a limited number of monitoring stations, noting that Kriging has the advantage of providing interpolated spatial irradiance information not readily available from other forecasting methods. Similarly, [53] utilized spatio-temporal Kriging for very short-term irradiance forecasts, as this method can interpolate irradiance across both space and time. In another study, [54] employed spatio-temporal Kriging to determine the threshold distance for monitoring stations, beyond which the accuracy of solar irradiance forecasting does not improve.

Regardless of the mathematical technique employed — whether it be a neural network, machine learning ensemble, or another method — various preprocessing tools are typically used when handling large datasets, as observed in similar applications. These tools include data normalization, outlier or anomaly removal/treatment, changing data resolution, data augmentation, correlation analysis, and the use of Virtual Machine Systems (VMS) [55]. Other advanced techniques include Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) [56], Variational Mode Decomposition (VMD) [57], K-means clustering, Principal Component Analysis (PCA), Graph Construction, and Empirical Mode Decomposition (EMD) [58]. It is worth noting that only a few cases involve working with simple raw data [59], [60].

B. RESEARCH CONTRIBUTION AND ORGANIZATION

This study presents a novel approach for forecasting solar irradiance in the Amazon Basin by integrating advanced deep learning models with statistical and geospatial techniques. The methodology combines component-wise time series decomposition with hybrid neural architectures (CNN-LSTM and CNN-Transformer), tailored specifically to handle the high variability of solar data. Additionally, a SARIMAX model is used for the stable components of the signal, and a Kriging metamodel offers a complementary spatial forecasting framework. Forecasting horizons extend up to three days ahead, using multivariate weather inputs such as temperature, wind, and cloud cover. The geographic focus begins with twelve cities in the Amazonas state and is

refined to three representative locations through clustering. This multistage process supports the evaluation of renewable energy potential in remote regions with limited infrastructure, providing a technical basis for replacing fossil-fuel-based generation with solar alternatives.

By addressing both the modeling complexity and regional limitations in data availability, this research contributes a practical and scalable framework to advance renewable energy planning and environmental sustainability in underrepresented areas like the Amazon.

The remainder of this work is organized as follows: Section II covers the materials and follows the steps of a DS workflow, along with the mathematical methods employed in this research. Section III presents the results, and Section IV provides a discussion of the findings. Finally, the conclusions are presented in Section V.

II. MATERIALS AND METHODS

A. DATA SCIENCE CYCLE

The identification of insights remains a key contribution of DS, involving tasks such as prediction, classification, clustering, and more. DS employs various mathematical tools, including TS analysis, Decision Trees (DT), Ensemble techniques, Neural Networks (NN), or combinations of these in hybrid models.

The DS cycle begins with accessing and verifying the metadata associated with the data. Initial processing may be necessary, such as handling invalid entries and substituting missing values. Exploratory Data Analysis (EDA) then provides insights into data distribution through statistical parameters, clustering, and outlier identification. Trends and patterns are identified, guiding the selection of features for mathematical modeling. The subsequent steps involve data preparation, algorithm application, and the generation of initial insights through classification or regression. Iterative loops may be required to refine results until acceptable outcomes are achieved. The final step involves comparing the outcomes to standards or specifications, leading to initial insights and potential solutions to the problem or question. As a cycle, it accepts several reviews of the steps according to analysis and acceptance criteria.

B. DATA WORK SCOPE

This research focuses on data from the state of Amazonas, the largest state in Brazil, where the primary landscape is the rainforest. Data from 12 cities within the Amazon were collected to analyze solar incidence. These cities are all located near major rivers and are surrounded by rainforests, each at varying stages of deforestation, human occupation, and development indexes. Based on the exploratory data analysis (EDA), the study identified spatial and contextual patterns that justified the use of clustering techniques to optimize computational efficiency. The methodology involved selecting three representative groups of cities, ensuring geographic diversity by including at least one city from the northern, southern, and central regions, as well as one from the eastern and western sectors. Notably, a single city could fulfill multiple regional criteria simultaneously, thereby enhancing the representativeness and efficiency of the clustering process.

TABLE 3. Technical variables for solar incidence forecast.

Variable	Unit	Description
ALLSKY_SFC_SW_DWN	kW-hr/m^2	The total solar irradiance incident (direct plus diffuse) on a horizontal plane at the surface of the earth under all sky conditions.
ALLSKY_KT	dimensionless	A fraction representing clearness of the atmosphere and all-sky insolation transmitted through the atmosphere to strike the surface of Earth divided by the average of top of the atmosphere total solar irradiance incidents.
WS10M	m/s	The average of wind speed at 10 meters above the surface of the earth.
WD10M	degrees	The average of the wind direction at 10 meters above the surface of the earth.
PS	kPa	The average of surface pressure at the surface of the earth.
PRECTOTCORR	mm/day	The bias corrected average of total precipitation at the surface of the earth in water mass (includes water content in snow).
RH2M	dimensionless	The ratio of actual partial pressure of water vapor to the partial pressure at saturation, expressed in percent (%), at 2 meters above the surface of the earth.
T2M	Celsius	The average air (dry bulb) temperature at 2 meters above the surface of the earth.
T2M_MIN	Celsius	The minimum day air (dry bulb) temperature at 2 meters above the ground surface in the period of interest.
T2M_MAX	Celsius	The maximum day air (dry bulb) temperature at 2 meters above the ground surface in the period of interest.

To achieve this, the K-means algorithm [61] was applied in conjunction with Principal Component Analysis (PCA), using key variables such as latitude, longitude, incident solar energy, and population. Table 3 shows the variables considered, along with their respective definitions [62]. Thus, the research considered the spatial distribution of the cities, taking cases from each cardinal spot (e.g. N, S, E and W) and central regions of the Amazonas State, with strong relations to the weather data collection. This approach confirmed the cities of Labrea, Manaus and São Gabriel da Cachoeira (SGC) as the set to be considered in the computing effort.

Table 4 provides initial data, including latitude, longitude, altitude, and population [63], adhering to the ISO 6709 standard for representing latitude and longitude, which simplifies

FIGURE 1. Location of cities Labrea, Manaus, and São Gabriel da Cachoeira, in Amazonas State. (Grid dataset provided by Google Maps).

TABLE 4. Data from the cities.

City	Latitude (^{o})	Longitude (^{o})	Altitude (m)	Population
Barcelos	-0.97	-60.92	40.00	27,638
Benjamin Constant	-4.38	-70.03	65.00	44,873
Coari	-4.08	-63.13	46.00	86,713
Codajas	-3.84	-62.06	32.00	29,691
Eirunepe	-6.67	-69.87	104.00	36,121
Iaurete	0.61	-69.18	120.00	3,000
Labrea	-7.25	-64.83	61.00	47,685
Manaus	-3.10	-60.02	61.25	2,255,903
Manicore	-5.82	-61.30	50.00	57,405
Parintins	-2.63	-56.73	29.00	116,439
São Gabriel da Cachoeira	0.13	-67.06	90.00	47,031
Tefe	-3.83	-64.70	47.00	59,250

mathematical operations. Most of the cities in this study are classified as 'small,' except for Manaus, which is categorized as 'large.'

From the clustering procedure, the three representative cities were: Labrea, Manaus, and São Gabriel da Cachoeira (SGC), shown in the map of Fig. 1.

For solar incidence, the primary data source comes from ground stations equipped with specialized instruments to measure solar irradiance [64]. The local variables include solar radiation incidence, relative humidity, wind speed and direction, mean temperature, and rainfall precipitation. Initially, we sourced data from the Brazilian National Institute of Meteorology (INMET) website. However, upon reviewing the data, it became apparent that the ground stations exhibited inconsistent data profiles, making continuous and straightforward use by the mathematical tools challenging. This imposes a main limitation for the current research. Consequently, we shifted to using NASA data products (satellite sources), specifically the CERES and MERRA2 models [65], supplemented by ground station data for validation. The selected timeframe spans from January 1, 2013, to October 31, 2023, aligning with another research project in the region [66]. The dataset was obtained in comma-separated values (CSV) format, with daily time aggregation.

In this study, the target variable was solar energy incidence, forecasted for three future horizons: one, two, and three days ahead (D+1, D+2, and D+3). The choice of these horizons comes from the feasible time to store electric energy, using batteries, for that region, in a worst case. The input variables, such as temperature, wind speed, and cloud cover observed on the current day (D), were aligned with the corresponding target values for each forecast horizon. This structure allowed the model to learn the relationship between present weather conditions and future solar energy output. Additionally, data curation included an outlier detection step using the Isolation Forest algorithm with a 5% contamination factor. This process improved the quality of the input data, contributing to more reliable and robust model performance across all forecasting methods.

The distributions of the variables are presented in Fig. 2 for the city of Manaus, serving as a typical case. The plot of the data distribution for all variables in this figure helps in understanding the covariance behavior among two or more variables, which is consistent across other cities as well. The target variable has a mean value close to 5 kWh/m^2, with a non-Gaussian distribution. The input variable *ALLSKY_KT* exhibits a flatter distribution, with a mean value of 0.45, indicating most of the time overcast weather, which is a challenge for the solar energy option. The mean temperature is approximately 26.5^{o}C, also displaying a non-Gaussian distribution. The relative humidity has a mean value close to 90%, while the wind speed averages 1.25 m/s. The local pressure follows a Gaussian distribution, with a mean value around 100.4 kPa.

Thinking of the time series approach, Fig. 3 presents the boxplot with the annual statistics. This figure provided insights into the order of magnitude of the variables and their

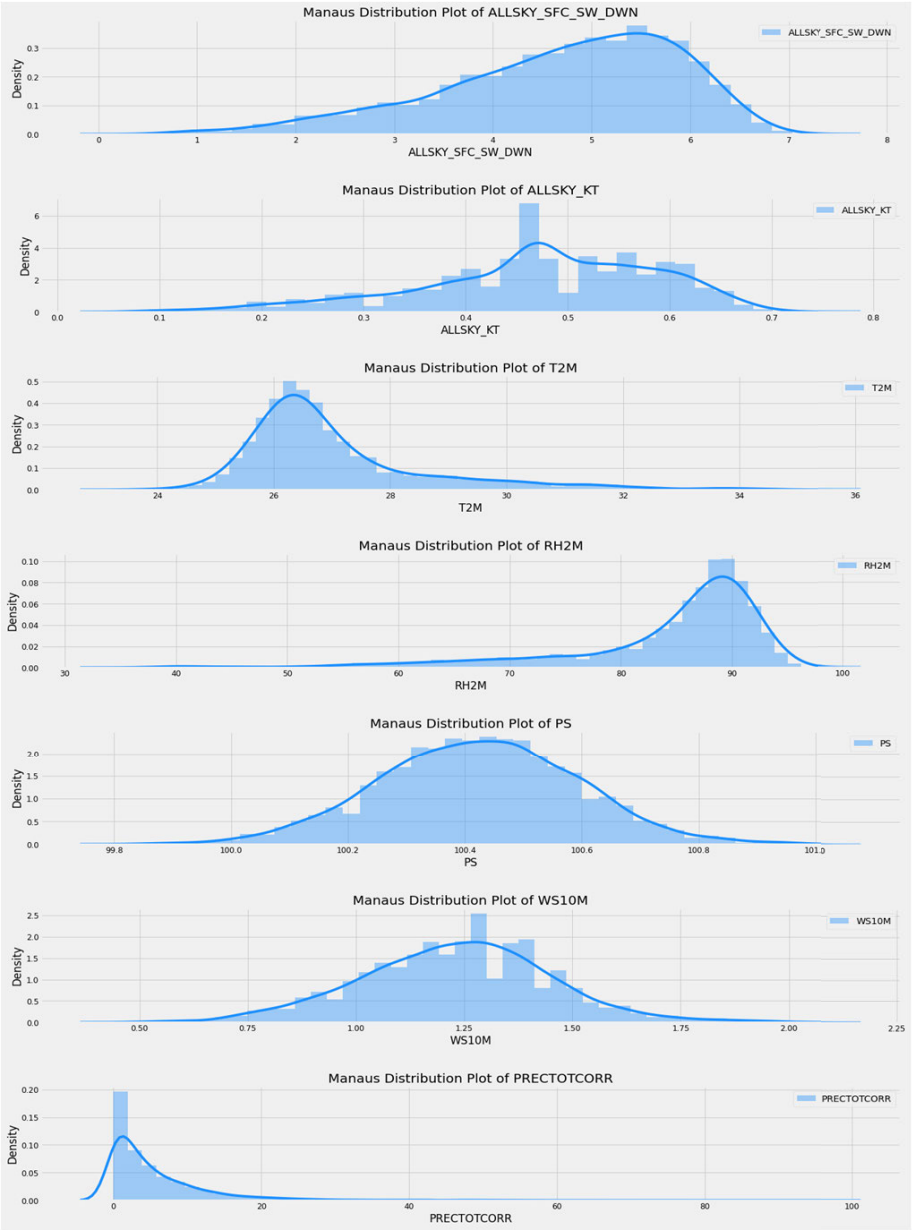

FIGURE 2. **Data distributions related to manaus.**

relationship with time. This task is crucial for the tuning phase of the forecasting algorithms.

This study used a total of 3,956 observations. For the neural network models, the data were split into three sets — training, validation, and testing — while maintaining the chronological order of the time series. The training set contained 2,963 observations (75% of the total), covering the period from January 1, 2013, to February 25, 2021. The test set included 989 observations (25% of the data), spanning from February 26, 2021, to October 31, 2023, and was used to evaluate the forecasting performance.

The data variance is a critical factor to assess due to its influence on the performance of the forecasting models. For the training set, the input variables exhibited a variance close to 0.03, while the target variable had a variance of 1.204. A similar profile was observed in the validation set, where the input variables again showed a variance around 0.03, and the target variable's variance was approximately 1.38.

However, the test set presented contrasting figures, with the input variables having a variance around 0.167 and the target variable a variance close to 0.012. This suggests that the performance of the algorithms may exhibit different behavior when comparing the train/validation sets against the test set.

C. FEATURE IMPORTANCE

Prior to developing mathematical models, the research carried out a comprehensive assessment of the relevance of each

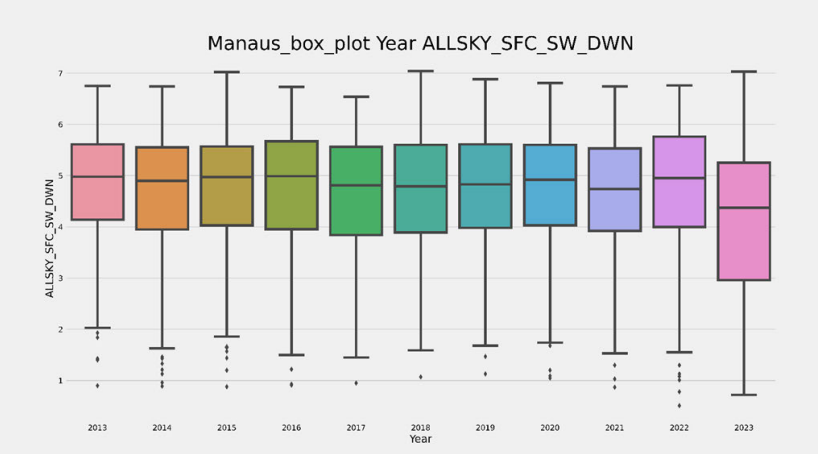

FIGURE 3. Year boxplot of variable *ALLSKY_SFC_SW_DWN* for manaus.

input variable. A correlation matrix was used to evaluate the strength of the relationship between each input and the target variable. For the city of Manaus, the variable *ALLSKY_KT* emerged as the most influential, with a correlation coefficient of 0.9, a pattern consistently observed in the other cities as well. This result is intuitive, as *ALLSKY_KT* reflects the degree of cloud cover, which directly affects the amount of solar radiation reaching the surface.

To complement this analysis, the relative importance of variables was further examined using Decision Tree-based methods, specifically the Extra Trees Regressor. The modeling process began with the full set of input variables and systematically reduced them through iterative testing. This approach ultimately identified *ALLSKY_KT* as the most significant predictor, yielding the best performance metrics when used as the sole input variable. Some seasonal aspects are covered by the relative humidity (RH2M) variation, such as its derivative, which was considered in feature engineering.

To guide the selection of input variables for the forecasting models, a correlation-based filtering method was applied. The correlation matrix (visualized as a heatmap) was used to evaluate the linear relationships between candidate input variables and the target variable. Variables were retained if their Pearson correlation coefficient with the target fell within the range of 0.35 to 0.85, ensuring moderate to strong association while avoiding potential target leakage. Additionally, to reduce multicollinearity among inputs, only variable pairs with mutual correlation coefficients below 0.5 were considered initially. This dual-criteria strategy helped preserve the predictive relevance of individual inputs while maintaining model robustness and computational efficiency. It is worth noting that lagged variables were also used, considering different periods, based on the signal decomposition in trend, seasonal and noise parts.

D. K-MEANS METHOD

K-means is a widely used unsupervised clustering algorithm, often applied to classification and signal processing tasks. The algorithm partitions a dataset into c clusters by assigning each observation to the nearest cluster centroid, using a defined distance metric, typically the Euclidean distance, though the Mahalanobis distance may be adopted in specific cases [67]. K-means is valued for its simplicity, computational efficiency, and robustness.

In this work, K-means was used to cluster data from 12 cities (see Table 4), thereby reducing computational complexity by grouping them into at most three clusters. The algorithm operates by minimizing the objective function $O(D)$, as determined by (1), (2) and (3).

$$O(D) = \sum_{i=1}^{c} K_i, \tag{1}$$

$$K_i = \sum_{j=1}^{c_i} D_{ij}, \tag{2}$$

$$D_{ij} = |c_i - v_j|^2, \tag{3}$$

where c_i represents the number of observations linked to a chosen cluster i; c denotes the number of cluster centroids (initially assumed); v_j signifies the position of the assumed centroid; D_{ij} stands for the distance between the observation and the cluster centroid v_j, which can be defined in several ways; K_i represents the sum of all distances from the observations to a specific centroid j; and $O(D)$ is the objective function to be minimized.

The K-means procedure consists of the following iterative steps:

1) Select the number of clusters c;
2) Compute the distance D_{ij} between each observation and all centroids;
3) Assign each observation to the nearest centroid;
4) Update centroids according to (4):

$$v_i = (1/c_i) \sum_{i=1}^{c_i} x_i, \tag{4}$$

where c_i is the number of observations x_i in a specific cluster, and v_i is the new cluster center.

5) Repeat steps 2-4 until the objective function O(D) converges to a minimum.

This iterative process continues until cluster assignments stabilize or a predefined convergence criterion is met, ensuring that the total within-cluster distance is minimized.

E. HOLT-WINTERS (TIME SERIES)

The Holt-Winters method is a classical univariate time series forecasting approach based on exponential smoothing, which progressively down-weights older observations to smooth abrupt variations. It models three main components: trend, seasonality and noise. In signal processing terms, Holt-Winters works as a low-pass filter, attenuating high-frequency noise. The method allows for both additive and multiplicative seasonality, selected according to the data characteristics. Its simplicity and computational efficiency also make it a suitable baseline for significance hypothesis testing against more complex models. The method is defined by (5), (6), (7), and (8):

$$S_t = a(Y_t/I_{t-1}) + (1-a)(S_{t-1} + b_{t-2}), \quad (5)$$

$$b_t = g(S_t - S_{t-1}) + (1-g)b_{t-1}, \quad (6)$$

$$I_t = b(Yt/S_t) + (1-b)I_{t-L}, \quad (7)$$

$$F_{t+m} = (S_t + mb_t)I_{t-L+m}, \quad (8)$$

where t is the time instant of the observation; Y represents the observation of the original time series; S denotes the smoothed time series, due the application of the Holt-Winters method; b is the trend or slope factor; I signifies the seasonal index; F is the forecasted value in a future time; a, b and g are constants determined by minimizing the performance metric (error); L stands for the number of time periods for seasonal components; and m is the number of periods related to seasonality or periodicity.

F. NEURAL NETWORKS FORECASTING METHODOLOGY: HYBRID NEURAL AND STATISTICAL MODELING

To enhance the accuracy and interpretability of solar irradiance forecasting, this research used a hybrid modeling framework that integrates classical statistical techniques with advanced neural network architectures, as seen in the references. The work followed a step-by-step process, beginning with exploratory modeling using recurrent networks and evolving into a component-wise forecasting strategy based on time series decomposition.

Initial experiments used Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, both well-suited for time series forecasting. Each had four hidden layers (30-40 units), 20% dropout, and used the Adam optimizer. Training spanned 50 to 100 epochs. Although results were promising, these models struggled to capture the high variability in solar irradiance data, leading the research to use hybrid approaches combining the method Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX) and neural networks architectures. To address the observed variance in the data and better tailor models to specific characteristics, the time series was decomposed using Seasonal-Trend decomposition via Loess (STL). By isolating each component, the forecasting task became more manageable and suitable for specialized model configurations. In brief, the original signal is separated into three components: Trend or long-term evolution of the series; Seasonality or recurring patterns or cycles; and Noise (Residuals) or also the short-term irregularities or random fluctuations. Given that the Noise/Residuals exhibited high variability and posed the greatest challenge to forecasting accuracy, advanced deep learning models were used for this segment, combining types of NN:

- CNN-LSTM: A convolutional neural network (CNN) to extract local temporal features from the noise, followed by LSTM layers to capture sequential dependencies. This hybrid model combines the pattern-recognition power of CNNs with the memory capabilities of LSTMs.
- CNN-Transformer: As an alternative, the LSTM was replaced with a Transformer decoder, which applies self-attention mechanisms to learn long-range dependencies within the noise. This configuration was particularly beneficial in capturing abrupt fluctuations that LSTMs might overlook.
- TCN (Temporal Convolutional Network): In addition to hybrid models, TCNs were implemented, employing causal and dilated convolutions to capture both short- and long-term temporal dependencies within the noise component.

The trend and seasonal parts were forecasted with SARIMAX and the noise by NN. All models were optimized using the 'Optuna' optimization framework. Specifically for NN, it allowed automated tuning of hyperparameters such as sequence length, learning rate, dropout, number of hidden units and activation function. The performance of the models was evaluated using Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) on the test dataset.

Tables 5 and 6 show the technical parameters considered in the models.

Figure 4 presents the data workflow related to the algorithms. For each city and horizon, the data is checked in terms of outliers and consistency (data curation). Then, the data is decomposed into trend, seasonal and noise parts, according to the best period (1, 2, 7, 14 days, etc.). The trend and seasonal parts are added together. The SARIMAX method will handle this sum, while the noise will be treated by the NN. Before the application of the forecasting method, the feature engineering applies, with feature importance analysis, including lagged variables. After each

TABLE 5. Hyperparameter ranges for SARIMAX model.

Hyperparameter	Range	Notes
p (AR order)	0 to 3	Autoregressive component
d (I order)	0 to 2	Degree of differencing
q (MA order)	0 to 3	Moving average component

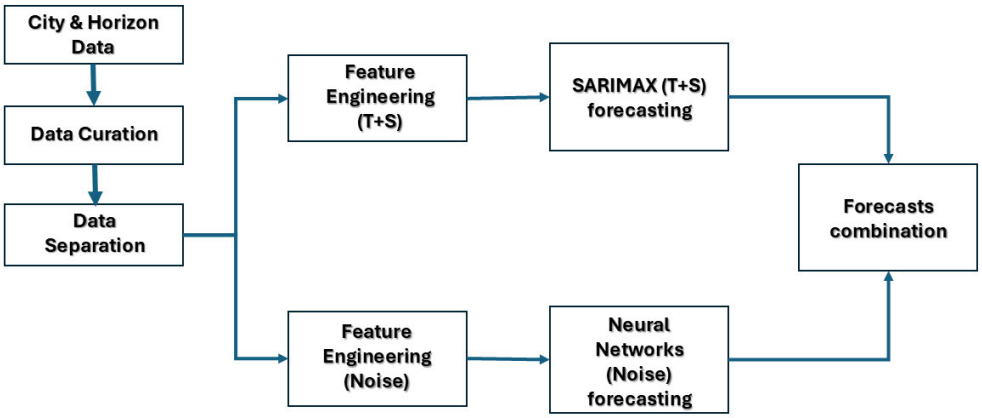

FIGURE 4. Data workflow related to the algorithms.

TABLE 6. Hyperparameter ranges for CNN-LSTM, CNN-transformer and TCN.

Hyperparameter	Search Range / Options	Notes
cnn_out_channels	16 to 128 (step = 16)	Number of output channels in the CNN layer
cnn_kernel_size	[3, 5, 7]	Kernel size for CNN convolution
dropout_rate	0.1 to 0.5 (step = 0.1)	Dropout to reduce overfitting
lr (learning rate)	1e-5 to 1e-2 (log scale)	Used in Adam optimizer
batch_size	[16, 32, 64]	Batch size for training
lstm_hidden_size	32 to 128 (step = 16)	Only in CNN-LSTM models
nhead	[2, 4, 8]	Number of attention heads in Transformer
d_model	[32, 64, 128, 256] (divisible by nhead)	Transformer model dimension
num_encoder_layers	1 to 4	Number of Transformer encoder layers

forecasting application, the outcomes from SARIMAX and NN are combined, making the final forecast figure, to be compared against the test phase data.

G. KRIGING METAMODEL

The Kriging metamodel applied in this research follows the theory and algorithms presented by [32]. The metamodel acts as an interpolator that relates the inputs and outputs of m known experiments. Each input variable is a vector composed of n variables, and each output is a vector composed of q results. In this study, the input variables are: $ALLSKY\_KT, T2M, RH2M, WD10M, WS10M, T2M\_MIN$, and $T2M\_MAX$, corresponding to $n = 7$ variables (as listed in Table 3). The output variable is the total solar irradiance ($ALLSKY_SFC_SW_DWN$) with $q = 1$. The $m = 2963$ experiments were obtained from daily satellite data of selected cities between the years 2013 and 2021, and were used to train the metamodel.

The interpolation of any input data x is achieved through a function combining a regression model $\mathcal{F}(\beta, x)$ and a stochastic process $z(x)$, as defined in (9), resulting in a deterministic response $\hat{y}(x)$. The coefficients β are the regression parameters obtained through a least squares fit of the m experiments during metamodel training.

$$\hat{y}(x) = \mathcal{F}(\beta, x) + z(x). \tag{9}$$

The regression model $\mathcal{F}$ is a quadratic model defined by a linear combination of the functions $f(x)$, shown in (10), with coefficients β.

$$f(x) = 1, x_1, \ldots, x_n, x_1^2, x_1x_2, \ldots, x_1x_n, x_2^2, x_2x_3, \ldots, x_2x_n, \ldots, x_n^2. \tag{10}$$

The stochastic process $z(x)$ follows a normal distribution with a zero mean and covariance expressed by (11):

$$cov[z(w), z(x)] = \sigma^2\mathcal{R}(\theta, w, x), \tag{11}$$

where σ^2 represents the process variance; θ is an optimization parameter used to minimize the mean squared error of the metamodel response; w and x are the inputs composed of the n variables of the metamodel; and $\mathcal{R}(\theta, w, x)$ is the correlation matrix that can be selected by the user. In this research, the Spherical correlation (12) was chosen for modeling the solar irradiance of the selected cities. This decision was based on the best accuracy obtained during the validation phase of the metamodels.

$$\mathcal{R}(\theta, w, x) = \prod_{j=1}^{n}(1 - 1.5\xi_j + 0.5\xi_j^3), \text{ with}$$
$$\xi_j = min(1, \theta_j|w_j - x_j|). \tag{12}$$

The validation of the Kriging metamodel was conducted using separate experimental data with the same input and output variables as the training experiments. The validation set consisted of 989 daily data points from 2021 to 2023.

For more details on other regression models, correlation models, or the complete development of the Kriging method, refer to [32].

H. ALGORITHM PERFORMANCE METRICS

There are several methods to evaluate the performance of a specific prediction algorithm. In this research, we focused on error metrics, which involve comparing the real values against the predicted ones, with the latter not being used to either prepare or train the algorithm. Depending on the research objective, the Mean Absolute Percentage Error (MAPE) was selected as the primary metric to compare the outputs of the prediction algorithms. However, we also considered other error metrics to facilitate comparison with other studies, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The equations and units for each metric are provided in (13), (14), and (15).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^*| \qquad (kWh/m^2), \tag{13}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{|y_i - y_i^*|}{y_i}\right] \qquad (dimensionless), \tag{14}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2} \qquad (kWh/m^2), \tag{15}$$

where y_i is the target value input from the data source (NASA satellite models); y_i^* is the forecasted value generated by the forecast algorithm; and n is the number of observations.

I. COMPUTING RESOURCES

In this research, the authors utilized Amazon Web Services (AWS) SageMaker and Google Colab notebooks and Python libraries. The computing environment consisted of 4 GB of memory and two vCPUs (ml.t3.medium), and the total computational work amounted to 40 hours.

III. RESULTS

The algorithms were executed independently for each city, as there is no interdependence among them. The objective was to forecast solar energy incidence for three future time horizons: Day +1, Day +2, and Day +3.

The Holt-Winters Exponential Smoothing method, with additive version, was used as a baseline and to capture key time series components trend, seasonality, and residual noise. The research applied the Diebold-Mariano (DM) test using the Holt-Winters forecast as the reference. The DM statistics and their corresponding p-values were computed for each forecasting method. In all cases, the p-values (0.1 - 0.9%) were below the 5% significance threshold, confirming that the proposed neural network models delivered statistically significant improvements over the Holt-Winters baseline.

The forecast accuracy, measured by the Mean Absolute Percentage Error (MAPE), is reported for each method in Table 7.

IV. DISCUSSION

After reviewing the initial use of GRU, MLP and LSTM methods, the use of hybrid techniques get the outcomes within the same order of magnitude as shown in Table 7. The city of Labrea has shown the smallest forecast performances and the city of Manaus the highest, within a difference range of 40%. The smallest MAPE was observed with the Kriging method in Labrea D+1 case, within a 2% difference in comparison to the Sarimax-CNN-LSTM method. The hybrid methods also showed the same order of magnitude for the 3 types of time horizons, in each city, indicating coherence.

TABLE 7. Forecasting performance of the hybrid models and Kriging.

City	Forecast Horizon	MAPE NN	Method NN	MAPE Kriging
Labrea	D+1	18.145%	SARIMAX-CNN-LSTM	17.787%
	D+2	18.195%	SARIMAX-CNN-Transformer	18.289%
	D+3	18.696%	SARIMAX-CNN-LSTM	18.451%
Manaus	D+1	26.581%	SARIMAX-CNN-LSTM	26.169%
	D+2	26.230%	SARIMAX-TCN	26.769%
	D+3	26.575%	SARIMAX-CNN-Transformer	27.389%
SGC	D+1	22.741%	SARIMAX-CNN-LSTM	22.500%
	D+2	22.498%	SARIMAX-TCN	22.680%
	D+3	22.774%	SARIMAX-CNN-LSTM	22.994%

TABLE 8. Overall MAPE performance for 5 cities in Bangladesh [68].

Method	MAPE
RNN	21.88%
LSTM	24.94%
GRU	19.28%

In assessing the results above, this research sought to benchmark its findings against existing studies whenever possible. Nevertheless, no directly comparable cases were identified in terms of region, time span, or forecast interval. Thus, references were selected based on the use of neural network models for solar energy forecasting with a forecasting horizon of one or more days, incorporating weather variables as input features. For instance, a study involving five cities in Bangladesh [68] reported the MAPE performance of various neural network architectures, as presented in Table 8. This analysis considered multiple weather variables in time series format, covering the period from 2014 to 2019, with data collected hourly (14 observations per day).

When compared to the study involving five cities in Bangladesh, the MAPE values in this research are slightly lower for Labrea (Day+1), which is noteworthy given the similar boundary conditions to the Amazon Basin.

Considering the stability of the MAPE figure, Fig. 5 presents the test phase and forecast values for the city of Labrea (D+1) case. For the sake of readability, the period chosen was July-2023, and it represents the general behavior of other test phase time slots.

The forecast method (Sarimax-CNN-LSTM) tries to follow the test data profile, exhibiting coherence but not catching the full data variance. Fig. 6 shows the Absolute Percentage Error for the same period.

As seen, there are spikes in the error over time profile, suggesting the data variance demands further effort to tune the method to generate a more stable MAPE profile.

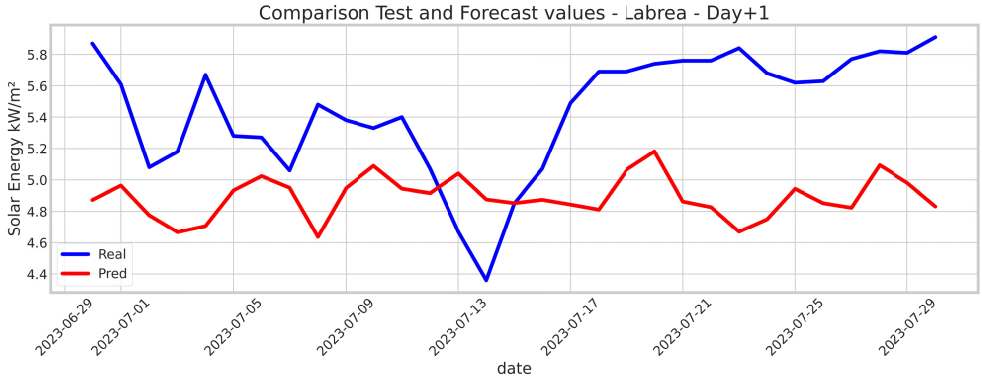

FIGURE 5. Test and forecast data comparison - Labrea (D+1).

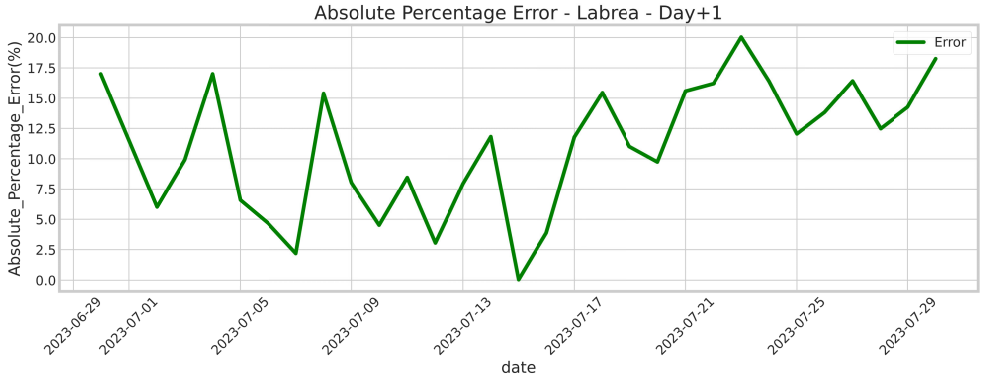

FIGURE 6. Absolute percentage error.

Nonetheless, in the period addressed, the error stays below 20% and with a mean value of 10%.

V. CONCLUSION

This study evaluated the performance of several machine learning methods—including hybrid approaches—and a Kriging metamodel for short-term solar energy forecasting in the Amazon Basin (up to three days ahead). The Holt-Winters method was employed as a baseline and for hypothesis significance testing. The analysis initially included twelve Amazonian cities, but this number was later reduced to three representative cities selected through a clustering procedure.

Regarding neural network (NN) applications for solar energy forecasting, the results of this study are in line with previous research, even considering the significant variability in boundary conditions. Overall, the cases analyzed here yielded MAPEs around 20%. High data variance consistently challenged model performance across all three cities. Nonetheless, several aspects should be further explored in future work:

1) **Seasonal Adjustments:** It may be advantageous to develop separate NNs for distinct climatic periods, such as the wet and dry seasons, to better account for seasonal effects.
2) **El Niño influence [69]:** This atmospheric phenomenon, which notably alters weather variables, was present during three periods covered by this research (2015-2016, 2018-2019, and 2023-2024). El NiÃ±o impacts data distribution and modifies the amplitude of daily temperatures, relative humidity, and other factors.
3) **Benchmarking Against Decision Tree Algorithms:** Comparing the MAPE results of NNs with those obtained from decision tree-based machine learning models, as discussed in [70] and [71], could help validate the forecasts under similar boundary conditions.

This future work shall consider updated references such as [73], [74], [75], and [76].

DATA AVAILABILITY

The data and Jupyter notebooks used in this article are available on Zenodo.org at https://zenodo.org/records/10565479 [72].

ACKNOWLEDGMENT

André Luis Ferreira Marques would like to thank the technical support from the Amazon Basin Protection System-Ministry of Defense.

REFERENCES

[1] K. Saidi and A. Omri, "The impact of renewable energy on carbon emissions and Economic growth in 15 major renewable energy-consuming countries," *Environ. Res.*, vol. 186, Jul. 2020, Art. no. 109567, doi: 10.1016/j.envres.2020.109567.

[2] S. Margulis. (2003). *Causes of Deforestation of the Brazilian Amazon*. [Online]. Available: http://documents.worldbank.org/curated/en/758171468768828889

[3] J. L. G. Silva, V. B. Capistrano, J. A. P. Veiga, and A. L. Brito, "Regional climate modeling in the Amazon basin to evaluate fire risk," *Acta Amazonica*, vol. 53, no. 2, pp. 166–176, Jun. 2023, doi: 10.1590/1809-4392202201881.

[4] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, Nov. 2005, doi: 10.7551/mitpress/3206.001.0001.

[5] J. P. C. Kleijnen, "Kriging metamodeling in simulation: A review," *Eur. J. Oper. Res.*, vol. 192, no. 3, pp. 707–716, Feb. 2009, doi: 10.1016/j.ejor.2007.10.013.

[6] D. S. Kumar, G. M. Yagli, M. Kashyap, and D. Srinivasan, "Solar irradiance resource and forecasting: A comprehensive review," *IET Renew. Power Gener.*, vol. 14, no. 10, pp. 1641–1656, Jul. 2020, doi: 10.1049/iet-rpg.2019.1227.

[7] W. Kanchana and S. Sirisukprasert, "PV power forecasting with holt-winters method," in *Proc. 8th Int. Electr. Eng. Congr. (iEECON)*, Chiang Mai, Thailand, Mar. 2020, pp. 1–4, doi: 10.1109/iEECON48109.2020.229517.

[8] E. Kim, M. S. Akhtar, and O.-B. Yang, "Designing solar power generation output forecasting methods using time series algorithms," *Electric Power Syst. Res.*, vol. 216, Mar. 2023, Art. no. 109073, doi: 10.1016/j.epsr.2022.109073.

[9] B. Billah, M. L. King, R. D. Snyder, and A. B. Koehler, "Exponential smoothing model selection for forecasting," *Int. J. Forecasting*, vol. 22, no. 2, pp. 239–247, Apr. 2006, doi: 10.1016/j.ijforecast.2005.08.002.

[10] L. Hakim, Y. A. Kurniawan, Khairudin, H. Gusmedi, and U. Hasanudin, "Modelling of hourly solar irradiance from field measurements in bandar Lampung," *AIP Conf. Proc.*, vol. 2563, Oct. 2022, Art. no. 070001, doi: 10.1063/5.0103174.

[11] M. Rodriguez, H. Cisneros, D. Arcos-Aviles, and W. Martinez, "Forecast of photovoltaic generation in isolated rural areas of Ecuador using holt-winters and seasonal variation methods," in *Proc. 48th Annu. Conf. IEEE Ind. Electron. Soc.*, Oct. 2022, pp. 1–6, doi: 10.1109/IECON49645.2022.9968817.

[12] G. Alkhayat and R. Mehmood, "A review and taxonomy of wind and solar energy forecasting methods based on deep learning," *Energy AI*, vol. 4, Jun. 2021, Art. no. 100060, doi: 10.1016/j.egyai.2021.100060.

[13] A. Michiorri, A. M. Sempreviva, S. Philipp, P. Perez-Lopez, A. Ferriere, and D. Moser, "Topic taxonomy and metadata to support renewable energy digitalisation," *Energies*, vol. 15, no. 24, p. 9531, Dec. 2022, doi: 10.3390/en15249531.

[14] D. Chakraborty, J. Mondal, H. B. Barua, and A. Bhattacharjee, "Computational solar energy-ensemble learning methods for prediction of solar power generation based on meteorological parameters in eastern India," *Renew. Energy Focus*, vol. 44, pp. 277–294, Mar. 2023, doi: 10.1016/j.ref.2023.01.006.

[15] N. Azizi, M. Yaghoubirad, M. Farajollahi, and A. Ahmadi, "Deep learning based long-term global solar irradiance and temperature forecasting using time series with multi-step multivariate output," *Renew. Energy*, vol. 206, pp. 135–147, Apr. 2023, doi: 10.1016/j.renene.2023.01.102.

[16] Y. Gao, S. Miyata, and Y. Akashi, "Multi-step solar irradiation prediction based on weather forecast and generative deep learning model," *Renew. Energy*, vol. 188, pp. 637–650, Apr. 2022, doi: 10.1016/j.renene.2022.02.051.

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.

[18] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," 2013, *arXiv:1312.6026*.

[19] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D, Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306, doi: 10.1016/j.physd.2019.132306.

[20] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, "Fundamental concepts of convolutional neural network," in *Recent Trends and Advances in Artificial Intelligence and Internet of Things*. Springer, Nov. 2019, pp. 519–567, doi: 10.1007/978-3-030-32644-9_36.

[21] A. Khosravi and S. Syri, "Modeling of geothermal power system equipped with absorption refrigeration and solar energy using multi-layer perceptron neural network optimized with imperialist competitive algorithm," *J. Cleaner Prod.*, vol. 276, Dec. 2020, Art. no. 124216, doi: 10.1016/j.jclepro.2020.124216.

[22] L. Lei, W. Chen, B. Wu, C. Chen, and W. Liu, "A building energy consumption prediction model based on rough set theory and deep learning algorithms," *Energy Buildings*, vol. 240, Jun. 2021, Art. no. 110886, doi: 10.1016/j.enbuild.2021.110886.

[23] M. Liu, J. Shi, K. Cao, J. Zhu, and S. Liu, "Analyzing the training processes of deep generative models," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 77–87, Jan. 2018, doi: 10.1109/TVCG.2017.2744938.

[24] B. Liu, L. Zhang, Q. Wang, and J. Chen, "A novel method for regional NO_2 concentration prediction using discrete wavelet transform and an LSTM network," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, Apr. 2021, Art. no. 6631614, doi: 10.1155/2021/6631614.

[25] K. Yan, W. Li, Z. Ji, M. Qi, and Y. Du, "A hybrid LSTM neural network for energy consumption forecasting of individual households," *IEEE Access*, vol. 7, pp. 157633–157642, 2019, doi: 10.1109/ACCESS.2019.2949065.

[26] S. Zhou, Y. Lu, and D. Bao, "Fault diagnosis of PEMFC systems based on wavelet packet energy decomposition and long short-term memory neural network," in *Proc. 5th Int. Conf. Power Energy Appl. (ICPEA)*, Guangzhou, China, Nov. 2022, pp. 585–591, doi: 10.1109/ICPEA56363.2022.10052339.

[27] T. Hirata, T. Kuremoto, M. Obayashi, S. Mabu, and K. Kobayashi, "Time series prediction using DBN and ARIMA," in *Proc. Int. Conf. Comput. Appl. Technol.*, Matsue, Japan, Aug. 2015, pp. 24–29, doi: 10.1109/CCATS.2015.15.

[28] E. Vetrimani, M. Arulselvi, and G. Ramesh, "Building convolutional neural network parameters using genetic algorithm for the croup cough classification problem," *Measurement: Sensors*, vol. 27, Jun. 2023, Art. no. 100717, doi: 10.1016/j.measen.2023.100717.

[29] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: https://www.jmlr.org/papers/v15/srivastava14a.html

[30] N. H. M. M. Shrifan, M. F. Akbar, and N. A. Mat Isa, "Maximal overlap discrete wavelet-packet transform aided microwave nondestructive testing," *NDT E Int.*, vol. 119, Apr. 2021, Art. no. 102414, doi: 10.1016/j.ndteint.2021.102414.

[31] A. Booker, "Design and analysis of computer experiments," *Stat. Sci.*, vol. 4, no. 4, pp. 409–423, Aug. 1998. [Online]. Available: http://www.jstor.org/stable/2245858

[32] S. N. Lophaven, H. B. Nielsen, and J. Søndergaard, "DACE–A MATLAB Kriging toolbox," Tech. Univ. Denmark, Kongens Lyngby, Denmark, Tech. Rep., 2002. [Online]. Available: http://www2.compute.dtu.dk/pubdb/pubs/3213-full.html

[33] M. A. Oliver and R. Webster, "A tutorial guide to geostatistics: Computing and modelling variograms and kriging," *CATENA*, vol. 113, pp. 56–69, Feb. 2014, doi: 10.1016/j.catena.2013.09.006.

[34] A. Booker, J. E. Dennis, P. D. Frank, D. B. Serafini, and V. Torczon, "Optimization using surrogate objectives on a helicopter test example," in *Proc. Comput. Methods Optim. Design Contro, Prog. Syst. Control Theory*, Jan. 1998, pp. 49–58, doi: 10.1007/978-1-4612-1780-0_3.

[35] J. Wang, N. Vlahopoulos, Z. P. Mourelatos, O. Ebrat, and K. Vaidyanathan, "Probabilistic and sensitivity analyses for the performance characteristics of the main bearings in an operating engine due to variability in bearing properties," *Int. J. Vehicle Design*, vol. 40, no. 4, p. 265, Feb. 2006, doi: 10.1504/ijvd.2006.009072.

[36] G. G. Wang and S. Shan, "Review of metamodeling techniques in support of engineering design optimization," *J. Mech. Design*, vol. 129, no. 4, pp. 370–380, Apr. 2007, doi: 10.1115/1.2429697.

[37] K. Maki, R. Sbragio, and N. Vlahopoulos, "System design of a wind turbine using a multi-level optimization approach," *Renew. Energy*, vol. 43, pp. 101–110, Jul. 2012, doi: 10.1016/j.renene.2011.11.027.

[38] J. Eason and S. Cremaschi, "Adaptive sequential sampling for surrogate model generation with artificial neural networks," *Comput. Chem. Eng.*, vol. 68, pp. 220–232, Sep. 2014, doi: 10.1016/j.compchemeng.2014.05.021.

[39] R. Sbragio, O. R. Filho, and M. R. Martins, "Methodology for the estimation of an oil spill origin: Analysis of the 2019 Brazilian coast oil spill," *Mar. Pollut. Bull.*, vol. 197, Dec. 2023, Art. no. 115676, doi: 10.1016/j.marpolbul.2023.115676.

[40] C. G. Hart, Z. He, R. Sbragio, and N. Vlahopoulos, "An advanced cost estimation methodology for engineering systems," *Syst. Eng.*, vol. 15, no. 1, pp. 28–40, Mar. 2012, doi: 10.1002/sys.20192.

[41] J. R. Nelson and T. H. Grubesic, "A repeated sampling method for oil spill impact uncertainty and interpolation," *Int. J. Disaster Risk Reduction*, vol. 22, pp. 420–430, Jun. 2017, doi: 10.1016/j.ijdrr.2017.01.014.

[42] N. Abdalla, S. Banerjee, G. Ramachandran, M. Stenzel, and P. A. Stewart, "Coastline kriging: A Bayesian approach," *Ann. Work Exposures Health*, vol. 62, no. 7, pp. 818–827, Aug. 2018, doi: 10.1093/annweh/wxy058.

[43] K. Chen, M. Ni, M. Cai, J. Wang, D. Huang, H. Chen, X. Wang, and M. Liu, "Optimization of a coastal environmental monitoring network based on the Kriging method: A case study of quanzhou bay, China," *BioMed Res. Int.*, vol. 2016, pp. 1–12, Sep. 2016, doi: 10.1155/2016/7137310.

[44] G. Erdogan Erten, M. Yavuz, and C. V. Deutsch, "Combination of machine learning and Kriging for spatial estimation of geological attributes," *Natural Resour. Res.*, vol. 31, no. 1, pp. 191–213, Jan. 2022, doi: 10.1007/s11053-021-10003-w.

[45] M. A. Oliver, *Geostatistical Applications for Precision Agriculture*. Cham, Switzerland: Springer, 2010, doi: 10.1007/978-90-481-9133-8.

[46] M. Bessafi, V. Oree, A. Khoodaruth, and J.-P. Chabriat, "Impact of decomposition and Kriging models on the solar irradiance downscaling accuracy in regions with complex topography," *Renew. Energy*, vol. 162, pp. 1992–2003, Dec. 2020, doi: 10.1016/j.renene.2020.10.018.

[47] M. Jamaly and J. Kleissl, "Spatiotemporal interpolation and forecast of irradiance data using kriging," *Sol. Energy*, vol. 158, pp. 407–423, Dec. 2017, doi: 10.1016/j.solener.2017.09.057.

[48] D. Palmer, I. Cole, T. Betts, and R. Gottschalg, "Interpolating and estimating horizontal diffuse solar irradiation to provide U.K.-wide coverage: Selection of the best performing models," *Energies*, vol. 10, no. 2, p. 181, Feb. 2017, doi: 10.3390/en10020181.

[49] V. D'Agostino and A. Zelenka, "Supplementing solar radiation network data by co-Kriging with satellite images," *Int. J. Climatol.*, vol. 12, no. 7, pp. 749–761, Nov. 1992, doi: 10.1002/joc.3370120707.

[50] D. Yang, "Spatial prediction using Kriging ensemble," *Sol. Energy*, vol. 171, pp. 977–982, Sep. 2018, doi: 10.1016/j.solener.2018.06.105.

[51] S. H. Monger, E. R. Morgan, A. R. Dyreson, and T. L. Acker, "Applying the Kriging method to predicting irradiance variability at a potential PV power plant," *Renew. Energy*, vol. 86, pp. 602–610, Feb. 2016, doi: 10.1016/j.renene.2015.08.058.

[52] D. Yang, C. Gu, Z. Dong, P. Jirutitijaroen, N. Chen, and W. M. Walsh, "Solar irradiance forecasting using spatial–temporal covariance structures and time-forward kriging," *Renew. Energy*, vol. 60, pp. 235–245, Dec. 2013, doi: 10.1016/j.renene.2013.05.030.

[53] A. W. Aryaputera, D. Yang, L. Zhao, and W. M. Walsh, "Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging," *Sol. Energy*, vol. 122, pp. 1266–1278, Dec. 2015, doi: 10.1016/j.solener.2015.10.023.

[54] D. Yang, Z. Dong, T. Reindl, P. Jirutitijaroen, and W. M. Walsh, "Solar irradiance forecasting using spatio-temporal empirical Kriging and vector autoregressive models with parameter shrinkage," *Sol. Energy*, vol. 103, pp. 550–562, May 2014, doi: 10.1016/j.solener.2014.01.024.

[55] Y. Li, W. Li, and C. Jiang, "A survey of virtual machine system: Current technology and future trends," in *Proc. 3rd Int. Symp. Electron. Commerce Secur.*, Nanchang, China, Jul. 2010, pp. 332–336, doi: 10.1109/ISECS.2010.80.

[56] L. Hu, L. Wang, Y. Chen, N. Hu, and Y. Jiang, "Bearing fault diagnosis using piecewise aggregate approximation and complete ensemble empirical mode decomposition with adaptive noise," *Sensors*, vol. 22, no. 17, p. 6599, Sep. 2022, doi: 10.3390/s22176599.

[57] X. Zhou, C. Liu, Y. Luo, B. Wu, N. Dong, T. Xiao, and H. Zhu, "Wind power forecast based on variational mode decomposition and long short term memory attention network," *Energy Rep.*, vol. 8, pp. 922–931, Nov. 2022, doi: 10.1016/j.egyr.2022.08.159.

[58] X. Guo, W.-J. Li, and J.-F. Qiao, "A self-organizing modular neural network based on empirical mode decomposition with sliding window for time series prediction," *Appl. Soft Comput.*, vol. 145, Sep. 2023, Art. no. 110559, doi: 10.1016/j.asoc.2023.110559.

[59] J. F. Barraza, L. G. Bräuning, R. B. Perez, C. B. Morais, M. R. Martins, and E. L. Droguett, "Deep learning health state prognostics of physical assets in the oil and gas industry," *P. I. Mech. Eng. O-J Ris.*, vol. 236, no. 4, pp. 598–616, Dec. 2020, doi: 10.1177/1748006x20976817.

[60] L. Guarda, J. E. Tapia, E. L. Droguett, and M. Ramos, "A novel capsule neural network based model for drowsiness detection using electroencephalography signals," *Expert Syst. Appl.*, vol. 201, Sep. 2022, Art. no. 116977, doi: 10.1016/j.eswa.2022.116977.

[61] Y. Li and H. Wu, "A clustering method based on K-Means algorithm," *Phys. Proc.*, vol. 25, pp. 1104–1109, Jan. 2012, doi: 10.1016/j.phpro.2012.03.206.

[62] NASA. *POWER Data Access Viewer*. Accessed: Oct. 1, 2023. [Online]. Available: https://power.larc.nasa.gov/data-access-viewer/

[63] IBGE. *Brasil/Amazonas*. Accessed: Jun. 1, 2023. [Online]. Available: https://cidades.ibge.gov.br/brasil/am/panorama

[64] INMET. *Tabela De Dados Das Estações*. Accessed: Jun. 1, 2023. [Online]. Available: https://tempo.inmet.gov.br/TabelaEstacoes/

[65] X. Zhang, N. Lu, H. Jiang, and L. Yao, "Evaluation of reanalysis surface incident solar radiation data in China," *Sci. Rep.*, vol. 10, no. 1, Feb. 2020, doi: 10.1038/s41598-020-60460-1.

[66] P. Artaxo et al., "The green ocean Amazon experiment (GoAmazon2014/5) observes pollution affecting gases, aerosols, clouds, and rainfall over the rain forest," *Bull. Amer. Meteorological Soc.*, vol. 98, no. 5, pp. 981–997, May 2017, doi: 10.1175/bams-d-15-00221.1.

[67] H. Ghorbani, "Mahalanobis distance and its application for detecting multivariate outliers," *Facta Universitatis, Series: Math. Informat.*, vol. 34, no. 3, pp. 583–595, Oct. 2019, doi: 10.22190/fumi1903583g.

[68] A. N. M. F. Faisal, A. Rahman, M. T. M. Habib, A. H. Siddique, M. Hasan, and M. M. Khan, "Neural networks based multivariate time series forecasting of solar radiation using meteorological data of different cities of Bangladesh," *Results Eng.*, vol. 13, Mar. 2022, Art. no. 100365, doi: 10.1016/j.rineng.2022.100365.

[69] IPAM. *El Niño*. Accessed: Jan. 4, 2024. [Online]. Available: https://ipam.org.br/glossario/el-nino-2/

[70] A. L. Ferreira Marques, M. José Teixeira, F. Valencia de Almeida, and P. L. P. Corrêa, "Application of data science in the prediction of solar energy for the Amazon basin: A study case," *Clean Energy*, vol. 7, no. 6, pp. 1344–1355, Dec. 2023, doi: 10.1093/ce/zkad065.

[71] A. L. F. Marques, M. J. Teixeira, F. V. De Almeida, and P. L. P. Corrêa, "Neural networks forecast models comparison for the solar energy generation in Amazon basin," *IEEE Access*, vol. 12, pp. 17915–17925, 2024, doi: 10.1109/ACCESS.2024.3358339.

[72] A. L. F. Marques. *Solar Energy Forecast Algorithms Comparison for the Amazon Basin*. Accessed: Jan. 25, 2024. [Online]. Available: https://zenodo.org/records/10565479

[73] E. Chodakowska, J. Nazarko, Ł. Nazarko, and H. S. Rabayah, "Solar radiation forecasting: A systematic meta-review of current methods and emerging trends," *Energies*, vol. 17, no. 13, p. 3156, Jun. 2024, doi: 10.3390/en17133156.

[74] S. Murugesan, M. Mahasree, F. Kavin, and N. Bharathiraja, "Solar energy forecasting with performance optimization using machine learning techniques," *Electric Power Compon. Syst.*, vol. 2024, pp. 1–13, Feb. 2024, doi: 10.1080/15325008.2024.2316245.

[75] N. Kushwaha, V. K. Yadav, and R. Saha, "Advancing solar energy through artificial intelligence: A focus on optimization and forecasting," in *Proc. 1st Int. Conf. Adv. Comput. Sci., Electr., Electron., Commun. Technol. (CE2CT)*, Nainital, India, Feb. 2025, pp. 639–644, doi: 10.1109/ce2ct64011.2025.10939626.

[76] J. Rajarajeswaran, "Applications of artificial intelligence and machine learning for accurate forecasting and optimization of renewable energy generation," in *Proc. Int. Conf. Electron. Renew. Syst. (ICEARS)*, Tuticorin, India, Feb. 2025, pp. 1886–1889, doi: 10.1109/icears64219.2025.10940813.

ANDRÉ LUIS FERREIRA MARQUES received the bachelor's degree in naval engineering (specializing in propulsion and hydrodynamics) and the master's degree in nuclear engineering from the University of São Paulo, in 1990 and 1995, respectively, and the master's degree in mechanical engineering and the Engineering degree in nuclear engineering from Massachusetts Institute of Technology, in 1999. He is currently pursuing the Ph.D. degree in computer engineering with the University of São Paulo, focusing on the application of data science in the energy matrix transition with renewable sources. After serving in Brazilian Navy, his development and research of nuclear systems, dealing with experimental work, and hardware construction and operation, for 40 years. He is a Data Scientist.

RICARDO SBRAGIO received the bachelor's degree in naval engineering and the master's degree from the University of São Paulo (USP), in 1989 and 1995, respectively, and the master's degree, the Professional degree in nuclear engineering, and the Ph.D. degree in naval architecture from the University of Michigan, in 1997, 1998, and 2001, respectively. In 2023, he completed a postdoctoral research in naval engineering at the University of São Paulo. He was a Research Scholar at the University of Michigan on projects related to acoustics and structural simulations, and for Brazilian Navy in the research and development of nuclear systems, naval propellers, and hydrodynamic test facilities. He is currently a Visiting Scholar at the Human Resources Program of Brazilian National Agency for Petroleum, Natural Gas and Biofuels (PRH-ANP), Laboratory of Analysis, Evaluation and Risk Management (LabRisco), USP.

PEDRO LUIZ PIZZIGATTI CORRÊA received the bachelor's and master's degrees in computer science from the University of São Paulo, in 1987 and 1992, respectively, and the Ph.D. degree in electrical engineering from the Polytechnic School of the University of São Paulo, Brazil, in 2002. He completed a postdoctoral fellowship in data science at the University of Tennessee, in 2015. Since 2017, he has been an Associate Professor at the Computer Engineering and Digital Systems Department (PCS), Polytechnic School of the University of São Paulo. His current research interests include distributed databases, data science, computer system modeling, distributed system architecture, computing in biodiversity, agricultural automation, and electronic government.

MARCELO RAMOS MARTINS received the degree, master's, and Ph.D. degrees in naval and oceanic engineering from the Escola Politécnica da USP, in 1992, 1996, and 1999, respectively, the Postdoctoral degree from the University of Maryland, in 2010, and the Postdoctoral degree from the University of California at Los Angeles, in 2019. He is currently a Full Professor at the Polytechnic School of the University of São Paulo (USP), the Coordinator of the Postgraduate Program in Naval and Oceanic Engineering at USP, and the Coordinator of the Human Resources Training Program (PRH06) of the National Agency of Petroleum and Biofuels. He is the Director and a Main Researcher at the Laboratory of Analysis, Evaluation and Risk Management (LabRisco), USP. He has coordinated research and development projects financed both by funding agencies (CAPES, CNPq, FINEP, and FAPESP) and by companies in the industrial sector, among them Petrobras, Transpetro, Repsol, and Vale. He has experience in probabilistic analysis of risk, reliability, maintenance, and security, and design of ships and oceanic systems and system dynamics. He is a member of the Scientific Committee of the International Conference on Ocean, Offshore and Arctic Engineering of Ocean, Offshore & Arctic Engineering Division of ASME, and the Founding Partner and a member of Brazilian Risk Association (ABRisco). He is a Regular Reviewer of several journals, such as *Reliability Engineering and System Safety*, *International Journal of Quality, Statistics, and Reliability*, *Journal of Risk Analysis*, *Journal of the Operational Research Society*, and *Journal of Loss Prevention in the Process Industries*.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - ROR identifier: 00x0ma614