Machine Learning-Based Solar Radiation Forecasting for Green Hydrogen Production

Mateus Vasconcelos Albuquerque[1] and Wallace Casaca[2(✉)]

[1] ICMC, University of São Paulo, São Carlos, Brazil
mateus.vasconcelos@outlook.com.br
[2] IBILCE, São Paulo State University, São José do Rio Preto, Brazil
wallace.casaca@unesp.br

Abstract. The transition to renewable energy sources has generated increased interest in accurate forecasting methods for green hydrogen production. This work aims to predict green hydrogen production from solar energy generation using machine and deep learning models. The proposed approach includes data preprocessing from different sites, implementing AI-driven techniques, running hyperparameter optimization, and extrapolating these parameters for training with real data from other sites. In addition, the Time Delay Embedding technique is applied to capture the temporal dependencies of the data for supervised learning. The methods Random Forest, Support Vector Regression, Extreme Gradient Boosting, and Long Short-Term Memory are trained and properly tuned. The results demonstrate that Extreme Gradient Boosting model achieves the highest accuracy, with all models adapting well to data from the distinct stations analyzed. The extrapolation of optimized hyperparameters proved efficient, reducing computational costs without compromising accuracy. In conclusion, the proposed approach is robust and viable for predicting the production of green hydrogen at different locations, making it a scalable solution for supporting clean energy planning.

Keywords: Green Hydrogen · Photovoltaic Generation · Machine Learning · Time Series Forecasting

1 Introduction

The urgent need for global decarbonization to limit temperature rise to 1.5 °C, as highlighted in the 2023 UN Climate Change Conference (COP23) report [25], has intensified the focus on renewable energy, including the green hydrogen production [2]. Green hydrogen, produced via electrolysis powered by renewable sources, is emerging as a key player in reducing carbon emissions across various industries, from replacing fossil fuels to serving as a critical feedstock in ammonia production and metal refining [13]. Furthermore, it can act as an energy storage medium, mitigating fluctuations in renewable power generation.

One of the primary challenges in green hydrogen production is the stochastic nature of renewable energy sources, such as wind [20] and solar power [22], which

O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15650, pp. 35–52, 2025.
https://doi.org/10.1007/978-3-031-96962-1_3

essentially depend on environmental variables like airspeed and solar irradiance. As a result, developing predictive models for these variables, particularly Global Horizontal Irradiance (GHI), has become a prominent research focus in solar-driven hydrogen production.

To tackle this challenge, statistical modeling and time series analysis techniques, such as Auto-Regressive Integrated Moving Average (ARIMA), coupled with machine learning approaches like Support Vector Machines (SVM) and Deep Neural Networks (DNN), have proven effective in accurately predicting renewable energy generation variables, including GHI [13,22]. These methods have demonstrated strong predictive performance over different forecasting horizons, ranging from short-term (10 min) to daily predictions. Therefore, enhancing the accuracy of solar generation forecast is crucial for improving the efficiency and scalability of green hydrogen generation.

To further explore the potential of these techniques for clean hydrogen production, this paper implements, fine-tunes and compares different machine learning models for solar irradiance forecasting across distinct locations. The predicted solar radiation values are taken as input to estimate the potential hydrogen production capacity in these locations, utilizing publicly available meteorological data. In addition, this study investigates whether hyperparameters optimized in one location can be effectively extended to others, ensuring scalability and model reusability. Our empirical evaluation assesses the adaptability of the proposed methodology and the consistency of model performance across multiple locations. The central research question is whether these models can effectively support hydrogen production planning in diverse geographic regions using only publicly available meteorological data. The main challenge lies in handling the stochastic and site-specific nature of solar irradiance, which affects both predictive accuracy and model generalizability.

This paper is structured as follows: Sect. 2 presents a review of the literature on solar irradiance forecasting, green hydrogen production, and relevant machine learning methodologies. Section 3 details our methodology, including data sources, modeling techniques, and evaluation metrics. Section 4 discusses the obtained results, while Sect. 5 concludes the study with key findings and future research directions.

2 Related Work

This section provides an in-depth review of the specialized literature on prediction methods for green hydrogen production, focusing on the application of machine learning and statistical algorithms to forecast sustainable hydrogen generation, solar irradiation, photovoltaic energy generation, and their interrelationships. The growing importance of this field is reflected in the increasing number of publications since 2019. However, the limited availability of reliable, operational historical data, due to the fact that only about 15% of green hydrogen-related projects identified by the International Energy Agency (IEA) are operational with most being pilot or demonstration projects [15], poses significant challenges in using hydrogen production as a direct target variable for prediction models.

Given the aforementioned issue, recent research has prioritized forecasting variables that serve as inputs into physical models of hydrogen production, such as solar irradiation and photovoltaic generation. In fact, the generation of renewable hydrogen from solar or wind energy is intrinsically dependent on the energy produced by photovoltaic panels, which, in turn, are influenced by stochastic quantities like GHI and wind speed [12,20,22].

When it comes to predicting solar irradiation for hydrogen-based applications, conventional approaches can be classified as follows [11]: (i) persistence models, which rely on the assumption that irradiation values remain constant over short periods; (ii) physical models, such as those based on Numerical Weather Prediction (NWP), grounded in physical principles and differential equations; (iii) statistical models, including methods like ARIMA and SARIMA; and (iv) AI-based models, which employ machine learning and DNN architectures. Notice that other taxonomies are also commonly adopted in the specialized literature, particularly those that classify forecasting methods according to the time horizon they are designed to address [1].

Comparative analyses, such as the one conducted by Allal et al. [3], have assessed the performance of various machine learning models for GHI forecasting, highlighting the superior accuracy of the Multilayer Perceptron (MLP) and the Random Forest (RF) algorithms. In a similar vein, Belmahdi et al. [5] evaluated several models, including MLP, ARIMA, and RF, for short-term prediction tasks, and found that the RF method outperformed others across diverse climatic and seasonal scenarios. In parallel, recent efforts in the field of eXplainable Artificial Intelligence (XAI) have focused on understanding the influence of input features within machine learning models. A representative example is the study by Sevas et al. [23], which integrates LightGBM and SHAP-driven analyzes into interactive platforms designed for solar irradiation forecasting.

In the context of green hydrogen production forecasting, Cheng et al. [10] and Haider et al. [13] applied machine learning models, specifically SVM and Prophet, to predict solar irradiation, a critical factor in hydrogen yield estimation. While Cheng et al. found SVM to be more effective, Haider et al. reported better performance with Prophet, highlighting how model suitability can vary across datasets and distinct contexts.

DNNs have also started to play a significant role in creating predictive models for green hydrogen production. For example, Alhussan et al. [2] proposed the use of Recurrent Neural Networks (RNN) with metaheuristic optimization techniques to forecast solar radiation in a location in Hawaii, with the goal of producing green hydrogen for the following season. Nikulins et al. [19] evaluated the effectiveness of Convolutional Neural Network (CNN) architectures for short-term wind speed forecasting, which serves as input for green hydrogen production. Sareen et al. [22] developed an algorithm based on Bidirectional Long Short-Term Memory (BiDLSTM) to forecast global horizontal irradiance and create an atlas of green hydrogen potential in India.

In contrast to most studies that take solar irradiation or wind speed data, some works have explored the use of real photovoltaic production data to train

machine learning models. For example, Asiedu et al. [4] tested various data-driven models, including Linear Regression (LR), Ridge and Lasso Regression, DNN, RF, AdaBoost and Extreme Gradient Boosting (XGBoost), observing that DNN stood out in short-term predictions, while XGBoost and RF were more effective in longer time horizons. Buonanno et al. [8] combined historical photovoltaic production data with predictive data from NWP models, concluding that linear models were the most effective in their approach.

3 Methodology

This section presents the methodology implemented in this study, covering data collection, physical modeling of hydrogen production from solar energy, machine and deep learning approaches for the forecasting task, and evaluation metrics. Figure 1 illustrates the methodology adopted, which is partially based on the CRISP-DM framework [27]. Section 3.1 describes the data collection and preprocessing steps, detailing the meteorological datasets used. Section 3.2 focuses on the physical modeling of hydrogen production based on solar energy conversion and electrolysis, while Sect. 3.3 explores the machine and deep learning techniques used in our investigation. Section 3.4 describes the Time Delay Embedding technique, which is applied to adapt supervised learning methods for time series forecasting, and Sect. 3.5 presents the model evaluation metrics.

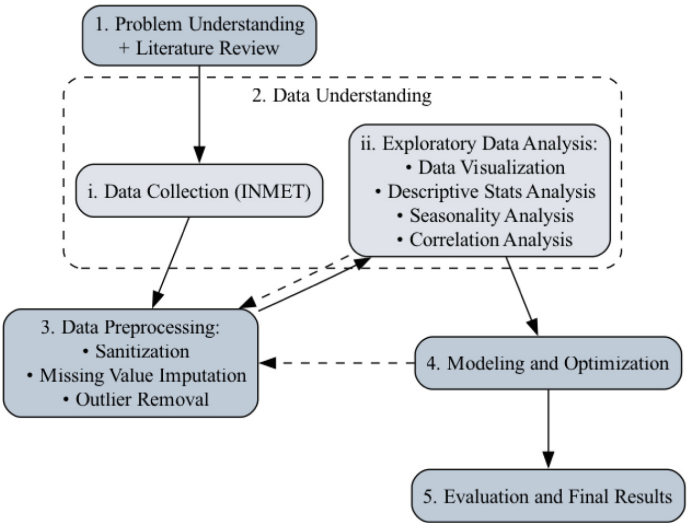

Fig. 1. Flowchart of the proposed methodology.

3.1 Data Collection and Pre-processing

The meteorological data used in this study were obtained from the Meteorological Database provided by the Brazilian Institute of Meteorology (INMET) portal, which gathers data across various locations in Brazil. These data are available as historical datasets, separated by weather station, covering hundreds of stations throughout the country. The data are recorded at an hourly resolution, enabling a detailed analysis of meteorological conditions over the study period. This level of granularity is essential for capturing daily and seasonal variations in input variables, which are crucial for accurately modeling green hydrogen production.

Initially, data were collected from 16 automatic weather stations located in the state of Ceará, Brazil, covering a period of over nine years, from September 1, 2015, to August 31, 2024. The dataset comprised 20 key meteorological variables relevant to solar energy modeling and hydrogen production. To ensure data consistency, preprocessing steps were applied before selecting the stations, including filtering erroneous measurements and correcting inconsistent or unrealistic values. The percentage of missing data was then analyzed, leading to the selection of four stations based on data quality and geographic distribution. This selection aimed to capture the diverse meteorological patterns present across the entire study area. As shown in Fig. 2, the selected stations are well distributed throughout the examined area, with some located inland and one near the coast, providing a comprehensive representation of varying climatic conditions.

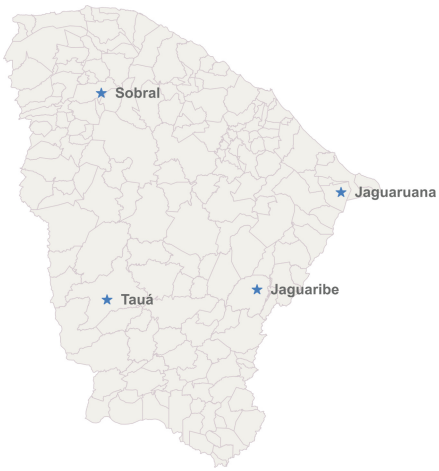

Fig. 2. Map of selected meteorological stations within the state of Ceará, Brazil.

After selecting the most representative weather stations, the data were refined through specific preprocessing steps, including handling missing and inconsistent values, detecting and treating outliers, and discarding irrelevant features, following a procedure similar to that of [17]. More specifically, the data preprocessing

involved several steps to ensure quality, consistency, and integrity for analysis. Erroneous values were filtered using predefined physical and meteorological constraints. Missing data for highly correlated variables were imputed using linear regression, while missing precipitation and nighttime global radiation data were set to zero. Maximum and minimum values of relevant variables over a 1-hour period were used to impute the target variable, such as estimating average temperature from hourly temperature extremes. Features weakly correlated with the target variable, Global Radiation, were discarded, ensuring only the most relevant variables were retained, as shown in Table 1.

Table 1. Meteorological variables and their database identifiers.

Variable	Identifier (in Portuguese)
Total Precipitation	`precipitacao_total_horario`
Atmospheric pressure at the site	`pressao_atmosferica_nivel_estacao_horaria`
Air Temperature	`temperatura_ar_bulbo_seco_horaria`
Relative Humidity	`umidade_relativa_ar_horaria`
Wind Direction	`vento_direcao_horaria`
Maximum Wind Gust	`vento_rajada_maxima`
Wind Speed	`vento_velocidade_horaria`
Global Radiation	`radiacao_global`

Outlier detection was also performed for each variable using the well-known IQR method, rejecting data points outside the $1.5 \times$ IQR range, with the threshold adjusted if more than 1% of data was rejected. No outliers were detected for global radiation. Missing data were imputed based on gap duration: seasonal averages for gaps over 24 h, interpolation for gaps under 6 h, and variable-specific methods for intermediate gaps. This preprocessing ensured data readiness for model development, with key variables like pressure, temperature, humidity, wind speed, and global radiation used in further analysis.

3.2 Hydrogen Production Modeling

In our approach, a solar-powered electrolysis system is taken so that the electricity generated by the photovoltaic plant is used to operate the electrolysis process and produce green hydrogen. This system consists of two modules: the energy generation module and the electrolysis module.

Photovoltaic Generation. In the energy generation module, the photovoltaic (PV) system converts global radiation (GHI) into electricity, with the captured radiation transformed into DC electricity based on system efficiency and operational conditions to power the water electrolysis system. Predicted GHI values are employed to estimate solar energy generation.

The energy produced by the photovoltaic system can be defined as [21]:

$$E_{PV} = GHI \cdot \eta_{PV} \cdot \eta_{PC}\,, \tag{1}$$

where GHI corresponds to the predicted Global Horizontal Irradiance, η_{PV} represents the efficiency of the photovoltaic module, and η_{PC} corresponds to the efficiency of the power conditioning system (PC). For the latter, a reasonable value of 85% was adopted, as reported in the specialized literature [7], although this parameter can reach values of up to 97%.

The so-called BiHiKu7 CS7N-670MB-AG photovoltaic module from *Canadian Solar* [9] was selected for this study. This module operates with power outputs of up to 670 W, excluding the bifacial gain of the system. Table 2 presents technical information about the selected module.

Table 2. Technical and performance parameters of selected photovoltaic module [9].

Specification	Value
Model	Canadian Solar BiHiKu7 CS7N-670MB-AG
Warranty	12 years
Electrical Characteristics	
Maximum Nominal Power (P_{max})	670 Wp
Optimal Operating Voltage (V_{mp})	38.7 V
Optimal Operating Current (I_{mp})	17.32 A
Open-Circuit Voltage (V_{oc})	45.8 V
Open-Circuit Current (I_{oc})	18.55 A
Module Efficiency (η_{PV})	21.6%
Power Degradation (first year)	<2%
Nominal Operating Temperature ($NMOT$)	$41 \pm 3°C$
Physical Characteristics	
Cell Type	Monocrystalline
Cell Arrangement	132 [2 x (11 × 6)]
Dimensions	2384 × 1303 x 33 mm
Weight	37.8 kg
Glass	2 mm Tempered Glass
Frame	Anodized Aluminum Alloy

Electrolysis System. In the electrolysis module, electrical energy from the photovoltaic system is used to split water molecules into hydrogen and oxygen in an electrolyzer. The hydrogen produced is proportional to the electricity supplied and the system's efficiency, with the model considering electrical variables and the electrolysis system's technical properties, as described by [6]:

$$\eta_{ele} = \frac{HHV_{H_2} \cdot \dot{n}_{H_2}}{P_{DC}}\,. \tag{2}$$

Rewriting in terms of the amount of hydrogen produced in kg/km^2, the following expression is derived:

$$M_{H_2} = \frac{E_{PV} \cdot \eta_{ele}}{HHV_{H_2}}, \tag{3}$$

where η_{ele} denotes the efficiency of the electrolytic cell, serving as an indicator of the effectiveness of the materials and design used. HHV_{H_2} refers to the *Higher Heating Value* of hydrogen gas, which is 285.83 kJ/mol. M_{H_2} represents the estimated hydrogen production in kg/km^2, while P_{DC} and E_{PV} correspond to the DC power supplied by the photovoltaic system and the energy obtained from Equation (1), respectively, expressed in the same units as global radiation.

The efficiency η_{ele} can be derived from operational parameters, following [10]:

$$\eta_{ele} = \frac{N_{H_2,out} \cdot LHV_{\eta_{H_2}}}{E_{PV} + Q_{PEM}\left(1 - \frac{T_0}{T_X}\right) + Q_m\left(1 - \frac{T_0}{T_s}\right)}, \tag{4}$$

where N_{H_2} corresponds to the hydrogen output flow, and $LHV_{\eta_{H_2}}$ is the Lower Heating Value (LHV) of H_2. Q_{PEM} refers to the thermal energy supplied to the electrolyzer, while Q_m denotes the thermal energy supplied to heat the working fluid of the heat exchanger at the end of the process. T_0 represents the ambient temperature of the system's heat source, and T_s is the external temperature of this process. Finally, T_x accounts for the external heat source temperature.

The specific electrical energy consumption for producing one normal cubic meter of hydrogen under nominal operation is typically provided by electrolyzer manufacturers in various forms of efficiency. In our approach, the Elyzer P-300 model of the *Siemens Energy* [24] was selected. Table 3 presents the technical specifications of the chosen electrolysis system.

Table 3. Technical and Performance Parameters of the Electrolysis System.

Specification	Value
Model	Elyzer P-300
Electrolysis Type	Atmospheric PEM
Stack Design	Optimized for 80 kW
Power Demand	17.5 MW
Hydrogen Production per Hour	335 kg
Outlet Pressure	100 mbar
Plant Efficiency (ϵ)	>75.5%
Minimum Load	At least 40%
Demineralized Water Consumption	<10 L/kg of hydrogen
Hydrogen Quality	Up to 99.999%
Startup Time	<1 min

Through the described modeling and manufacturer parameters, global radiation is transformed from the target variable to an intermediate variable, shifting the focus to hydrogen production as the primary variable, allowing direct prediction of hydrogen output from meteorological inputs.

3.3 Machine Learning and Deep Learning Models

As previously mentioned, this research employs supervised Machine Learning techniques, including traditional models, Ensemble Learning, and Deep Learning, to predict global radiation. Our methodology consists of two main stages: first, a traditional Machine Learning model, two Ensemble Learning models, and one Deep Learning model based on RNNs are trained, with hyperparameter optimization performed via Random Search strategy. For traditional and Ensemble Learning models, the Time Delay Embedding technique is applied to generate features that represent the lags of the original variables. Once optimal hyperparameters are computed, they are generalized to train models for other sites, ensuring consistency and comparability. Finally, these forecasts are then used to estimate hydrogen production, as detailed in Sect. 3.2.

Support Vector Regression. Support Vector Machines (SVM) are generalizations of models known as maximal margin classifiers [16], which employ kernels to find a hyperplane that best separates training data into different classes. These models can be extended to regression problems, becoming known as Support Vector Regression (SVR) [26]. Their fundamental principle consists of fitting a function that minimizes the prediction error while keeping the predicted values within an acceptable error margin.

The SVR algorithm takes a loss function that allows for error tolerance, where predictions within a predefined range, ϵ, are considered acceptable. Kernels enable the algorithm to operate in high-dimensional spaces by calculating the inner product of the data, projecting it into higher dimensions to capture complex relationships. This approach allows the regression function to be adjusted so that most points fall within the tolerable range without overfitting.

Random Forest. The Random Forest (RF) is an ensemble machine learning algorithm in which several independent (or decorrelated) decision trees are built, trained on a random sample with replacement, or bootstrap, of the training data. These trees are then combined to generate a prediction of interest, such as predicting a value in a time series prediction problem [16].

In this method, at each split point of the decision tree, m predictors are randomly selected from the total of p available predictors, and the best among them is used to perform the split. This process is repeated independently for each tree in the random forest.

By randomly selecting these m predictors, the influence of any single predictor that dominates the algorithm is reduced, thus lowering the variance. At the end of the process, an average of the outputs from the generated regressor trees is used to predict the final value.

Extreme Gradient Boosting. In contrast to a random forest, gradient-boosting algorithms model and arrange decision trees in series, with each subsequent tree in the series seeking to correct the errors made by the previous tree. Usually, these individual trees have a very small size (weak learners), which optimizes the algorithm's performance.

The eXtreme Gradient Boosting (XGBoost) is an optimized implementation of gradient boosting method that includes regularization to control decision tree complexity, a second-order approximation for the loss function (instead of linear approximation), predictor sampling in the internal nodes (similar to random forests), and advanced numerical methods for optimized scalability [18]. These enhancements make XGBoost a popular choice in machine learning competitions and industry applications due to its accuracy and scalability.

Long Short-Term Memory Networks. Long Short-Term Memory (LSTM) networks are a special recurrent neural network architecture, addressing the long-term dependency issue of traditional RNNs. LSTMs are crucial for sequential data prediction, introduced by [14].

An LSTM cell (Fig. 3) includes: c_{t-1} (input), x_t (vector input) and H_t (output to both network output and the next hidden layer).

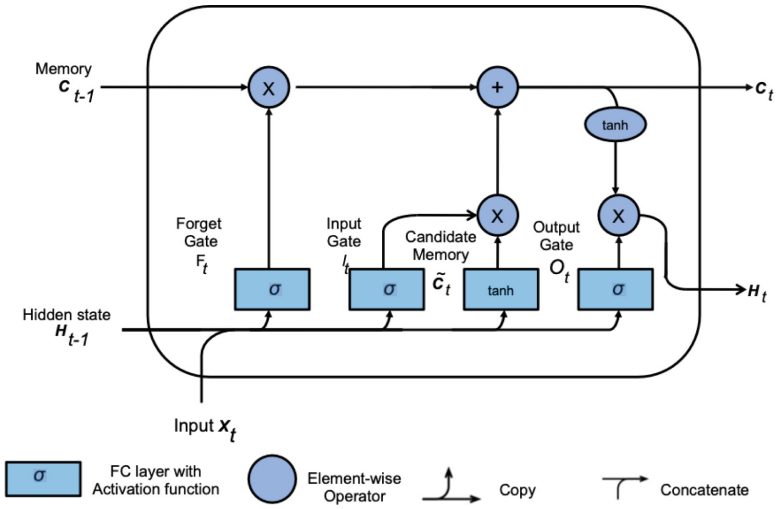

Fig. 3. Visual representation of an LSTM network [18].

H_t and c_t represent short and long-term memory, respectively. Three gates control the LSTM: Output Gate (O_t), Input Gate (I_t), and Forget Gate (F_t) [18]. These gates are calculated as follows:

$$\mathbf{O_t} = \sigma(\mathbf{X_t W_{xo}} + \mathbf{H_{t-1} W_{ho}} + \mathbf{b_o}) \tag{5}$$

$$\mathbf{I_t} = \sigma(\mathbf{X_t W_{xi}} + \mathbf{H_{t-1} W_{hi}} + \mathbf{b_i}) \tag{6}$$

$$\mathbf{F_t} = \sigma(\mathbf{X_t W_{xf}} + \mathbf{H_{t-1} W_{hf}} + \mathbf{b_f}) \tag{7}$$

The memory cell is updated according to the following equation:

$$c_t = \mathbf{F_t} \odot c_{t-1} + \mathbf{I_t} \odot \tilde{c}_t \,, \tag{8}$$

while the hidden state is computed as:

$$H_t = \mathbf{O_t} \odot \tanh(c_t) \,. \tag{9}$$

With this structure, the LSTM can carefully control the input, output, and removal of information, enabling more efficient learning and long-term memory retention.

3.4 Time Delay Embedding

Time Delay Embedding is a method that reconstructs a time series by using previous values of the target variable and its predictors. By capturing temporal relationships within the data, this approach enriches the feature space with time-dependent patterns, enabling machine learning models to achieve more accurate predictions.

For example, given a target variable y and two predictors x_1 and x_2, the original dataset can be written as follows:

t	y_t	$x_{1,t}$	$x_{2,t}$
1	y_1	$x_{1,1}$	$x_{2,1}$
2	y_2	$x_{1,2}$	$x_{2,2}$
3	y_3	$x_{1,3}$	$x_{2,3}$
4	y_4	$x_{1,4}$	$x_{2,4}$
5	y_5	$x_{1,5}$	$x_{2,5}$

By applying Time Delay Embedding with lags of 1 and 2, the resulting transformed matrix for predicting y_{t+1} becomes:

t	y_t	y_{t-1}	y_{t-2}	$x_{1,t}$	$x_{1,t-1}$	$x_{2,t}$	$x_{2,t-1}$	y_{t+1}
3	y_3	y_2	y_1	$x_{1,3}$	$x_{1,2}$	$x_{2,3}$	$x_{2,2}$	y_4
4	y_4	y_3	y_2	$x_{1,4}$	$x_{1,3}$	$x_{2,4}$	$x_{2,3}$	y_5

In this structure, y_{t+1} is the target value to be predicted, while the features include current and lagged values of the target (y_t, y_{t-1}, y_{t-2}) and predictors $(x_{1,t}, x_{1,t-1}, x_{2,t}, x_{2,t-1})$.

3.5 Evaluation Metrics

The performance of the predictive models was assessed using the following evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R-squared coefficient (R^2).

Mean Absolute Error (MAE). It measures the average of the absolute errors between the model's predictions and the actual values. MAE is calculated using the following equation:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,, \tag{10}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n represents the total number of observations.

Root Mean Square Error (RMSE). Provides a measure of mean squared errors, placing higher penalties on larger errors. The equation for calculating RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\,. \tag{11}$$

R-squared Coefficient (R^2). It is an evaluation metric that verifies how well a model explains the variability of the observed data. Its formula is given by:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}\,, \tag{12}$$

where $\bar{y}$ is the mean of the observed values. R^2 values range from 0 to 1, with higher values indicating better explanatory power.

4 Experimental Design, Results and Discussion

First, an Exploratory Data Analysis (EDA) was performed to examine statistical patterns of the variables, including their correlations, as shown in Fig. 4. Next, the variables were normalized using Min-Max scaling, and the dataset was split into 70% training and 30% testing. The Time Delay Embedding technique was then applied to both traditional machine learning and ensemble learning models, as described in Sect. 3.4.

After applying the Time Delay Embedding technique, a feature selection step was performed based on feature importance scores obtained from training a Random Forest model on the training data. An optimal set of 40 features was selected, as adding more did not significantly enhance accuracy relative to computational cost. The most important features selected were the following:

- `radiacao_global` at lags: `t-0, t-1, t-2, t-3, t-4, t-5, t-6, t-7, t-8, t-11, t-12, t-17, t-18, t-19, t-20, t-21, t-22, t-23, t-24`
- `umidade_relativa_ar_horaria` at lags: `t-0, t-19`
- `temperatura_ar_bulbo_seco_horaria` at lags: `t-0, t-1, t-19`
- `vento_direcao_horaria` at lags: `t-0, t-1, t-2, t-3, t-4, t-5, t-6, t-17, t-18, t-19, t-20, t-21, t-22, t-23, t-24`
- `vento_rajada_maxima` at lag: `t-0`

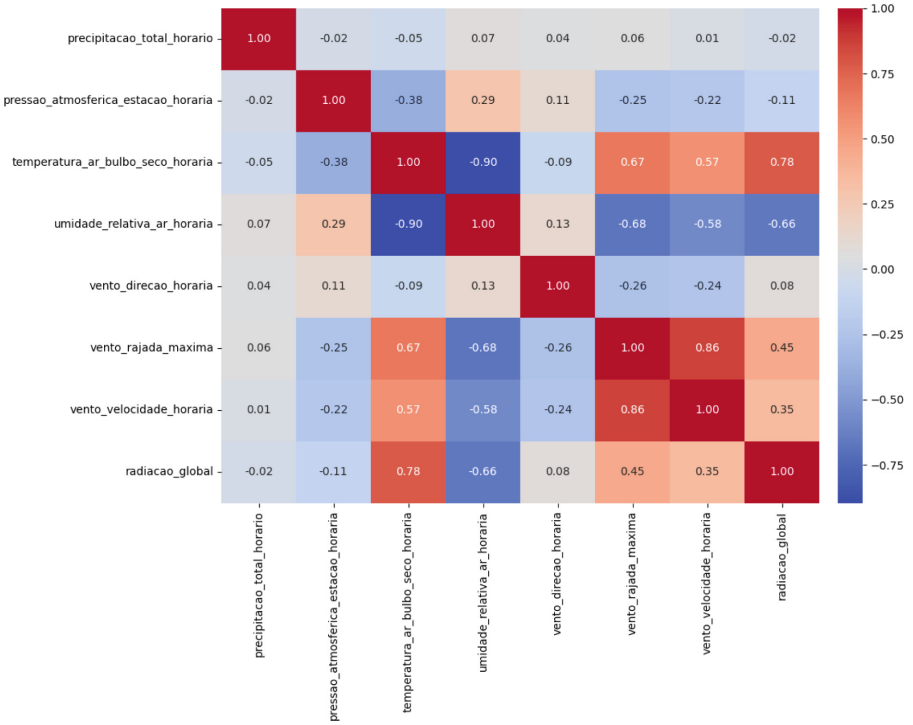

Fig. 4. Correlation matrix after preprocessing for Sobral station.

4.1 Hyperparameters Optimization

In order to optimize the hyperparameters of each model, the Random Search strategy was employed via `RandomizedSearchCV` from `scikit-learn` for the Sobral site, taking computational efficiency into account. This method performs more efficiently than Grid Search strategy by randomly sampling combinations from a defined hyperparameter space. The evaluation metric used was `neg_mean_absolute_error`, suitable for regression. Cross-validation was performed with `TimeSeriesSplit` to preserve the temporal order of data, using five splits for most models and three folds for the LSTM model. The Random Search included 50 iterations for most models and 20 for the LSTM. Table 4 lists the hyperparameters tested for each model, along with the optimal values for the Sobral site, computed by the Random Search method.

4.2 Comparative Performance

Table 5 summarizes the evaluation metrics of the best forecast models, assessed on the test dataset. This comparison highlights the performance of each model after hyperparameter optimization, providing insight into their relative effectiveness at the Sobral site – our first study location.

Table 4. Hyperparameter optimization results for different models at Sobral site.

Hyperparameter	Tested Values	Best Value Found
RF		
n_estimators	100, 200, 500, 1000	1000
max_depth	None, 10, 50, 100	100
min_samples_split	2, 5, 10, 15	2
min_samples_leaf	1, 5, 10, 20	1
max_features	`sqrt`, `log2`	`sqrt`
XGBoost		
n_estimators	100, 200, 300, 400, 500	100
learning_rate	0.01, 0.05, 0.1, 0.2	0.1
max_depth	3, 5, 7, 10	7
subsample	0.6, 0.7, 0.8, 0.9	0.9
colsample_bytree	0.6, 0.7, 0.8, 0.9	0.7
gamma	0, 1, 2, 3	0
min_child_weight	1, 3, 5, 7	7
reg_alpha	0, 0.1, 0.5, 1, 5	0
reg_lambda	0.1, 0.5, 1, 2, 5	0.1
SVR		
C	0.1, 1, 10, 100, 1000	10
epsilon	0.001, 0.01, 0.1, 0.5, 1.0	0.001
kernel	poly, rbf, sigmoid	rbf
gamma	scale, auto	auto
LSTM		
neurons	32, 64, 128, 256, 512	32
layers	1, 2	2
dropout	0.1, 0.2, 0.3, 0.5	0.2
batch_size	16, 32, 64	32
learning_rate	10^{-4}, 10^{-3}, 10^{-2}	10^{-3}
epochs	50, 100, 200	50

The comparative performance of the models, as shown in Table 5, reveals key differences in their predictive capabilities. Among the machine learning models, XGBoost outperformed the others with the lowest MAE (0.065) and RMSE (0.129), while achieving a R^2 of 0.946, indicating strong predictive accuracy. In contrast, the LSTM model, despite providing a reasonable R^2 of 0.919, exhibited higher error metrics (MAE of 0.086 and RMSE of 0.158), suggesting it was less effective at the Sobral site compared to the tree-based models.

Table 5. Performance metrics for different models at Sobral site.

Metric	Value
RF	
Mean Absolute Error	0,067 mol/m^2
Root Mean Squared Error	0,132 mol/m^2
R^2	0,944
XGBoost	
Mean Absolute Error	0,065 mol/m^2
Root Mean Squared Error	0,129 mol/m^2
R^2	0,946
SVR	
Mean Absolute Error	0,067 mol/m^2
Root Mean Squared Error	0,132 mol/m^2
R^2	0,944
LSTM	
Mean Absolute Error	0,086 mol/m^2
Root Mean Squared Error	0,158 mol/m^2
R^2	0,919

Next, the four ML-based models were trained using data from three other sites: Jaguaruana, Jaguaribe and Tauá, with hyperparameters previously optimized for the Sobral site using the Random Search method. This approach aimed to assess the feasibility of reusing hyperparameter optimization without the need for exhaustive retraining at each new location.

Performance evaluation, using MAE, RMSE and R^2, revealed that XGBoost outperformed the other models across all evaluation metrics, demonstrating its superior capability in time series prediction tasks. In contrast, although the LSTM model achieved a reasonable R^2 of 0.919, its higher error metrics suggest that it was less efficient than expected at the Sobral site. We acknowledge that further architectural improvements could enhance its performance in future iterations. Following closely behind, RF ranked second, showcasing strong performance, while the SVRl ranked third. These findings are visually represented in Fig. 5, highlighting the robustness of the models across different sites.

The evaluation metrics presented only slight differences across sites, which highlights the reliability of the preprocessing method and how well the optimized hyperparameters from Sobral worked at other locations. The current approach proves to be a practical solution for predicting green hydrogen production at nearby sites, saving both time and computational resources by avoiding the need to search for new hyperparameters.

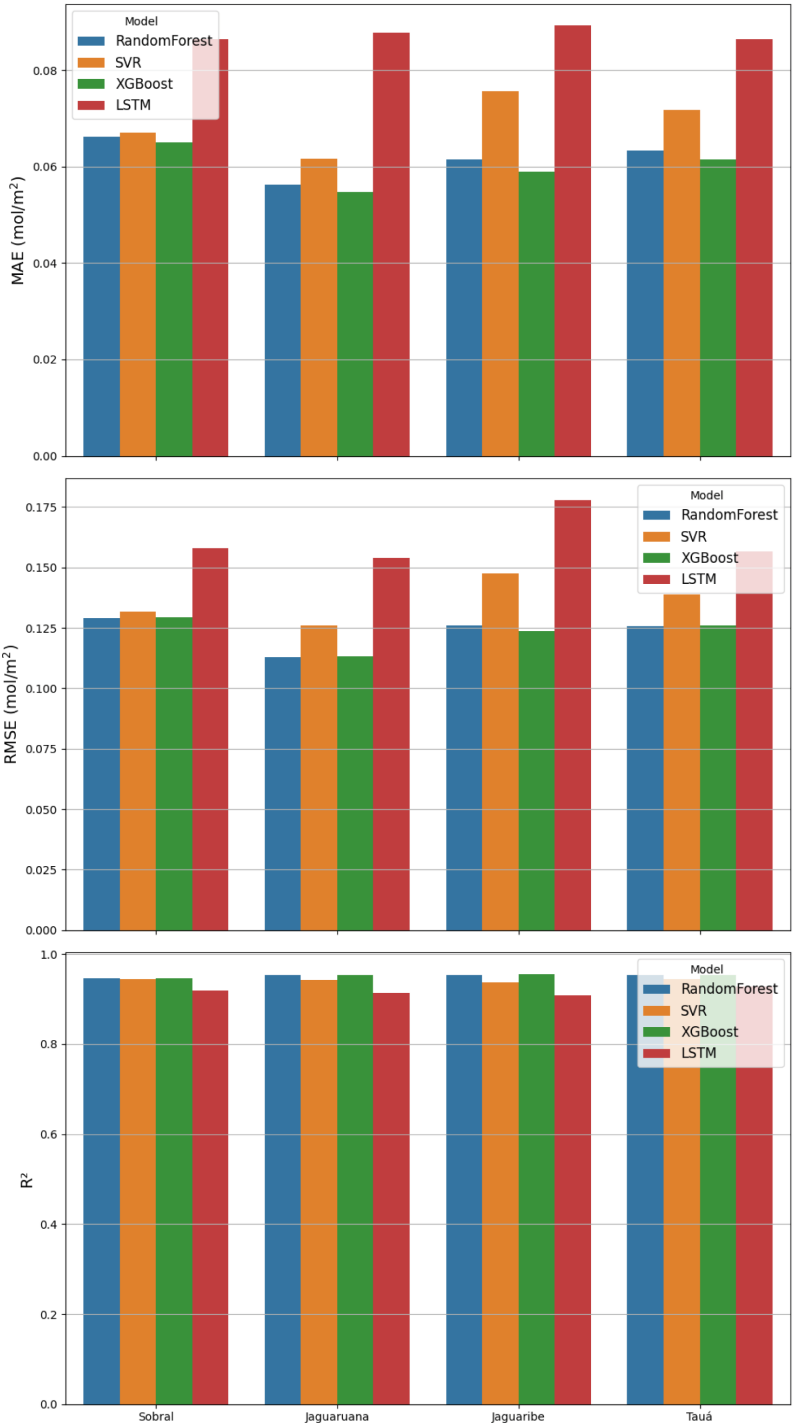

Fig. 5. Evaluation metrics across stations and predictive models.

5 Conclusion and Future Direction

This study introduced a comprehensive, data-driven approach to forecasting green hydrogen production from solar energy, by utilizing machine and deep learning models and hyperparameter optimization techniques. Our approach was applied to data from four different sites, including Sobral, and it was shown that transferring the optimized hyperparameters from Sobral to the other sites effectively reduced computational costs while preserving strong model performance. All implemented models delivered reliable and consistent results, with XGBoost outperforming the others in terms of MAE, RMSE and R^2.

In conclusion, our methodology provides an effective and accurate computational framework for predicting green hydrogen production from solar energy, showcasing the potential of machine learning in optimizing renewable energy generation. The results also pave the way for future advancements that could improve the applicability and scalability of this technology in real-world settings.

For future work, a key direction is to refine the LSTM model by exploring architectural adjustments and conducting additional configuration tests. Another important step is to expand the analysis to a broader range of geographical locations with varying climatic conditions, in order to evaluate the model's generalizability and ensure its effectiveness across different regions.

Acknowledgment. The authors would like to thank the São Paulo Research Foundation (FAPESP – grants #2013/07375-0 and #2023/14427-8) and the National Council for Scientific and Technological Development (CNPq – grant #316228/2021-4) for providing resources that contributed to the development of this research.

References

1. Ahmed, R., Sreeram, V., Mishra, Y., Arif, M.: A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renew. Sustain. Energy Rev. **124**, 109792 (2020)
2. Alhussan, A.A., et al.: Green hydrogen production ensemble forecasting based on hybrid dynamic optimization algorithm. Front. Energy Res. **11**, 1221006 (2023)
3. Allal, Z., Noura, H.N., Chahine, K.: Machine learning algorithms for solar irradiance prediction: a recent comparative study. e-Prime-Adv. Electr. Eng., Electron. Energy **7**, 100453 (2024)
4. Asiedu, S.T., Nyarko, F.K., Boahen, S., Effah, F.B., Asaaga, B.A.: Machine learning forecasting of solar pv production using single and hybrid models over different time horizons. Heliyon **10**(7) (2024)
5. Belmahdi, B., BOUARDI, A.E.: Short-term solar radiation forecasting using machine learning models under different sky conditions: evaluations and comparisons. Environ. Sci. Pollut. Res. **31**(1), 966–981 (2024)
6. Bessarabov, D., Wang, H., Li, H., Zhao, N.: PEM electrolysis for hydrogen production: principles and applications. CRC press (2016)
7. Boudries, R., Dizene, R.: Potentialities of hydrogen production in Algeria. Int. J. Hydrogen Energy **33**(17), 4476–4487 (2008)

8. Buonanno, A., et al.: Machine learning and weather model combination for PV production forecasting. Energies **17**(9), 2203 (2024)
9. Canadian Solar: Canadian solar bihiku7 cs7n-670mb-ag datasheet (2024). https://static.csisolar.com/wp-content/uploads/2020/10/06153525/CS-Datasheet-BiHiKu7_CS7N-MB-AG_v2.4_EN.pdf. Accessed 10 2024
10. Cheng, G., et al.: Analysis and prediction of green hydrogen production potential by photovoltaic-powered water electrolysis using machine learning in china. Energy 129302 (2023)
11. Chodakowska, E., Nazarko, J., Nazarko, Ł, Rabayah, H.S.: Solar radiation forecasting: a systematic meta-review of current methods and emerging trends. Energies **17**(13), 3156 (2024)
12. Donti, P.L., Kolter, J.Z.: Machine learning for sustainable energy systems. Annu. Rev. Environ. Resour. **46**(1), 719–747 (2021)
13. Haider, S.A., Sajid, M., Iqbal, S.: Forecasting hydrogen production potential in Islamabad from solar energy using water electrolysis. Int. J. Hydrogen Energy **46**(2), 1671–1681 (2021)
14. Hochreiter, S.: Long short-term memory. Neural Computation MIT-Press (1997)
15. IEA: Hydrogen production and infrastructure projects database (2024). https://www.iea.org/data-and-statistics/data-product/hydrogen-production-and-infrastructure-projects-database, iEA, Paris. Licence: CC BY 4.0
16. James, G., Witten, D., Hastie, T., Tibshirani, R., et al.: An introduction to statistical learning, vol. 112. Springer (2013)
17. Leme, J.V., Casaca, W., Colnago, M., Dias, M.A.: Towards assessing the electricity demand in brazil: Data-driven analysis and ensemble learning models. Energies **13**(6), 1407 (2020)
18. Murphy, K.P.: Probabilistic machine learning: an introduction. MIT press (2022)
19. Nikulins, A., et al.: Deep learning for wind and solar energy forecasting in hydrogen production. Energies **17**(5), 1053 (2024)
20. Paula, M., et al.: Predicting energy generation in large wind farms: a data-driven study with open data and machine learning. Inventions **8**(5), 126 (2023)
21. Rahmouni, S., Negrou, B., Settou, N., Dominguez, J., Gouareh, A.: Prospects of hydrogen production potential from renewable resources in Algeria. Int. J. Hydrogen Energy **42**(2), 1383–1395 (2017)
22. Sareen, K., Panigrahi, B.K., Shikhola, T., Nagdeve, R.: Deep learning solar forecasting for green hydrogen production in India: a case study. Int. J. Hydrogen Energy **50**, 334–351 (2024)
23. Sevas, M.S., Sharmin, N., Santona, C.F.T., Sagor, S.R.: Advanced ensemble machine-learning and explainable AI with hybridized clustering for solar irradiation prediction in Bangladesh. Theor. Appl. Climatol. **155**(7), 5695–5725 (2024)
24. Siemens Energy: Siemens energy Elyzerp-300 PV module product datasheet (2024). https://p3.aprimocdn.net/siemensenergy/9c60ee43-311e-473f-8148-b19a008958a2/Brochure-Electrolyzer_16zu9_240617-pdf_Original%20file.pdf. Accessed 10 2024
25. Smeeth, L., Haines, A.: Cop 28: Ambitious climate action is needed to protect health (2023)
26. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**, 199–222 (2004)
27. Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. vol. 1, pp. 29–39. Manchester (2000)