TECHNICAL REPORT 2023

IM SCIENCE

A comparative study of time-series forecasting models for solar irradiance in Minas Gerais

(Submitted – conference – KDMILE)

06

TECHNICAL REPORT | 2023

A comparative study of time-series forecasting models for solar irradiance in Minas Gerais

Citation details:
Ricardo H. Guedes Furiati, João Nazareth, Felipe Domingos da Cunha, Simon Malinowski, Cristiana Brasil Maia, Silvio Jamil F. Guimarães (2023). A comparative study of time-series forecasting models for solar irradiance in Minas Gerais. Technical Report Im-Science/PUC Minas #06/2023.

06

Technical report – ImScience 06/2023

A comparative study of time-series forecasting models for solar irradiance in Minas Gerais

Ricardo H. Guedes Furiati · João Nazareth · Felipe Domingos da Cunha · Cristiana Brasil Maia · Silvio Jamil F. Guimarães

June 8, 2023

Abstract Hydropower generation accounts for over 60% of energy production in Brazil, and despite being a renewable source, the high dependency of the country on this kind of power generation poses a risk to the sustainability of energy production, given that pollutant power generation (like coal and oil) are activated in - increasingly common- periods of drought to compensate for the decrease in the output of hydropower. One of the alternatives to diversify the Brazilian energy matrix towards other renewables is the use of solar power generation, given its scalability and the large territorial extension of Brazil, despite having an increase of 40% in generation capacity in the last couple of years, solar generation accounts for less than 2% of the total energy produced. One of the main setbacks of the installation of solar equipment (solar panels and distribution infrastructure) is that they are directly tied to the study of solar characteristics of a given region. Currently, the study of patterns in solar behavior and its absolute values are made on a case-to-case basis, demanding a lot of resources. To mitigate that problem, the use of machine learning computational models can be of great help in facilitating the expansion of solar generation in the country. The present study aims to compare three forecasting models found in the literature (Sarima, Holt-Winters, and LSTM) to predict solar irradiance in the Brazilian state of Minas Gerais. A past study was conducted using the Holt-Winters forecasting model with data collected by the National Institute of Meteorology (INMET) ground stations. Given the inconsistency in the data (missing values and different periods of observations for the stations), this study uses the most expensive dataset found in the literature, with solar behavior collected by NASA satellites over 35 years. The main goal of this study is to define the best model (with the lower prediction error) to predict solar irradiance as well as understand the impact that the amount of training data has on each model. Experimental analysis pointed out that the Holt-Winters model presented the lower error for predicting solar irradiance, and the LSTM model has the most consistent variation with the amount of training data.

Keywords Forecasting · Holt-Winters · LSTM · SARIMA · Solar Irradiance · Time-Series.

1 Introduction

The globaThe global energy matrix primarily comprises nonrenewable sources, such as coal and oil derivates, accounting for up to 61% of total production [7]. On the other hand, the Brazilian energy matrix has its primary source of energy provided by renewable sources, accounting for 84% of the energy produced [8] in which hydropower generation accounts for most of the production, reaching over 63% of total energy production. Despite having most of its energy derived from renewable sources, the high dependence on hydropower generation poses a risk to the sustainability of energy production in the country, given that this kind of power plant significantly loses its total output during drought. To compensate for the decrease in energy production and to maintain the availability of power in the grid, pollutant power plants, such as coal and oil, are activated, furthering carbon emissions in this period, which are becoming increasingly common.

Given the efforts to diversify Brazil's renewable sources, as one of the goals of the 2030 National Energy Plan [5] proposed by the Ministério de Minas e Energia, solar power generation presents itself as one of the best alternatives given its scalability and the territorial extension of Brazil.

Silvio Jamil F. Guimarães
ImScience/PUC-Minas – Belo Horizonte 31980-110, Brazil
E-mail: sjamil@pucminas.br

Solar generation accounts for less than 2% of total production in 2020 [8], and despite having a 40% increase in generation capacity in the past couple of years [?], there is still a long way to go to enable solar power generation as a consistent source of energy.

The installation of solar generation equipment, such as photovoltaic panels, as well as distribution infrastructure, is aimed towards regions with greater solar incidence to obtain better results. Currently, the solar irradiance characteristics, absolute values, patterns, and future behavior are analyzed on a case-to-case basis, demanding a lot of time and resources, which could hold back further investments in solar infrastructure, delaying the diversification of renewable in the country. To mitigate that problem, using machine learning techniques to analyze solar irradiance data and predict future values can significantly help the decision-making process and facilitate feasibility studies for equipment installation. In [4], a method based on Holt-Winters was proposed for predicting solar irradiance in Minas Gerais. Here, we have extended this proposal in order: (i) to cope with missing values problems in the dataset; and (ii) to better understand the behavior of the forecasting methods. Thus, in [4] was used solar irradiance data collected by ground stations of the National Institute of Meteorology (INMET) from 2015 to 2020 to create a heat map of the predictions to better visualize the method's forecasting capabilities. Still, it was observed that the dataset used had a lot of missing values and inconsistency amongst the stations. To mitigate these shortcomings, the present study uses the most expansive dataset found in the literature, which contains data from 1984 to 2020 collected by NASA satellites.

Moreover, we also include two more forecasting methods, SARIMA and LSTM, to predict solar irradiance and compare the results. Preliminary experimental analysis showed that, although all three models can be used to predict solar irradiance in the state, the LSTM model is the most accurate. Furthermore, experimental analysis was able to wield a rough estimate of the optimal amount of training data for each model. To the best of our knowledge, no other work in the literature predicts average monthly solar irradiance in the State of Minas Gerais over an extensive period. Despite using standard methods found in the literature (SARIMA, Holt-Winters, and LSTM), this work hopes to fill the gap in analyzing how each model's prediction capabilities vary over a significant period.

This work is organized as follows. Section 2 describes a literature review regarding time-series forecasting models and the state of their use for solar irradiance prediction. In Section 3, we explain some fundamental concepts of time series and solar irradiance variation. In Section 4, the methodology used for developing the work. Some experiments and discussions are drawn in Section 5. Finally, Section 6 presents the conclusion and future work.

2 Related Work

The literature presents a few approaches to using time-series forecasting to predict solar irradiance. [11] present a comparative analysis between two deep learning models, Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). The study compares the two models to a naive approach, which uses the most recent value of solar irradiance as the next prediction. The study found that it is not possible to define a better model, given that the error found in both is equivalent. This study used hourly values of solar irradiance collected by a weather station in Ajaccio-France, from January 1998 to December 2007 to forecast one hour into the future. The present work uses monthly average solar irradiance values to predict one year into the future.

Another approach to time-series forecasting is the implementation of statistical techniques to incorporate historical single-variable data in statistical models. [2] compares the most common family of statistical forecasting models, ARMA and ARIMA, to forecast up to a year into the future. [2] differentiate the use cases of each of the models, where the Autorregressive Moving Average (ARMA) can only be used with static data, that is, a time series whose statistical properties (mean and variance) are constant over time, the Autorregressive Integrated Moving Average (ARIMA) can be used with non-stationary time-series data, attributing more flexibility. The author points out that the ARIMA model has a lower error than the ARMA model in all test cases. This can be explained by the nature of the solar irradiance data, which is not very stationary, given the overall seasonality of the data.

[1] presents a complete overview of predicting solar irradiance in Minas Gerais, comparing computational models (using machine learning technics) to traditional empirical models. The study used data collected by all the National Institute of Meteorology stations available in the state of Minas Gerais and applied two machine learning algorithms, Artificial Neural Network (ANN) and Multivariate Adaptive Regression Spline (MARS), to predict daily solar radiation. Given the robustness of the selected machine learning models, multiple variables, such as relative humidity and atmospheric pressure, enhanced the predictions' precision. To compare the results [1] used performance profiles that, combined with regular metrics (RMSE, MAE, R2), present a clear superiority of the computational models over the classical empirical models in almost all test cases.

Accordingly, our work compares three computational models to predict monthly average solar irradiance in the state. Furthermore, given the inconsistency of the data regarding the period in which the amount of data available varies with each station, the present work uses the same period of historical data for all the stations.

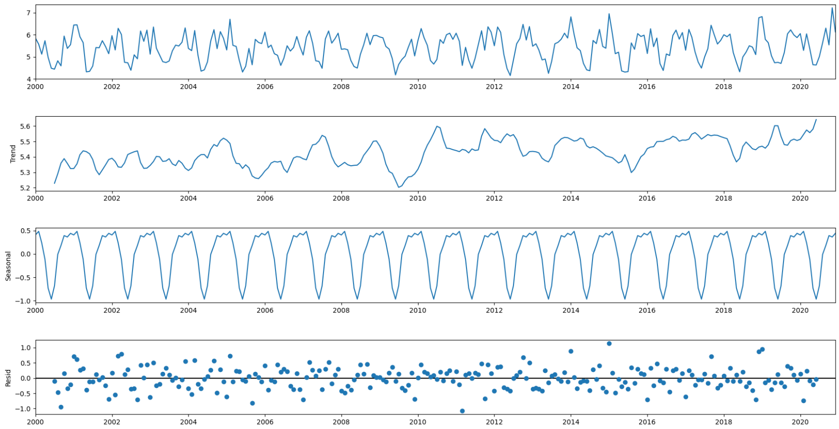


Fig. 1: Seasonal decompose of solar irradiance data.

3 Fundamental concepts

This section aims to address the main concepts used to develop this work. We start by introducing time-series and their main components, followed by information about the standard year and the predicted models used in this work.

3.1 Time-series

Time-series, as illustrated in the top of Figure 1, can be defined as a series of observations made along a period that can be analyzed and studied to understand better the characteristics of a problem, such as how it changes along the period as well as find patterns. A time series has three main components which illustrates how the data evolves over time:

1. Trend: Characterizes the overall tendency of the data
2. Seasonality: Describes regular patterns in the data along the time-period
3. Residual: The noise attributed to different variations in the data

Regarding time-series analysis in the context of solar irradiance, the component with the most significant impact on the forecasting models is seasonality that, in this case, presents a repeating pattern of solar irradiance every 12 months, as shown in Figure 1, attributed to the translation movement of the Earth. Although the overall pattern is recognizable in this period, the absolute values of solar radiance can vary given that the amount of solar irradiance that reaches Earth can be affected by climate patterns and variations in the sun's activity.

3.2 Typical Meteorological Year

Solar radiation is a measure that is constantly changing, as several factors such as atmospheric conditions, including cloud cover and relative humidity, influence it. For a more cohesive analysis, we decided to create a standard year with a reliable basis for comparisons between results [12]. The standard year is defined by the month closest to the monthly average of all years. Each month during the selected data period is calculated, and the one closest to the average is selected.

Figure 2 presents the results that compare the actual values of solar irradiance for Belo Horizonte in 2020 obtained from NASA database and the Typical Meteorological Year data. Although they follow the same general behavior, the absolute values are different from each other. This behavior was already expected since solar irradiance varies over the years.

4 Methodology for predicting solar irradiance

Studying solar irradiance is fundamental for significantly helping the decision-making process and facilitating feasibility studies for installing equipment in which the prediction of the solar incidence may increase their efficiency. Here, we

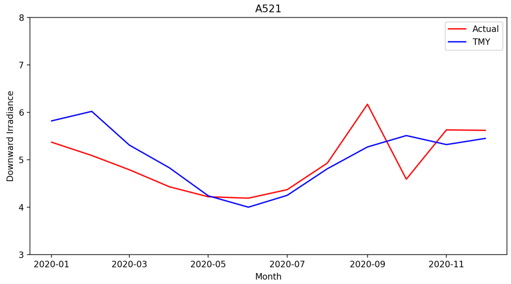


Fig. 2: Typical meteorological year and actual solar irradiance.

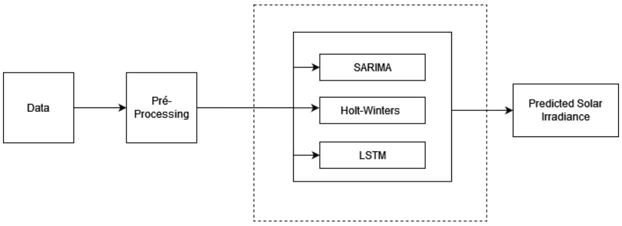


Fig. 3: Methodology for predicting solar irradiance.

have studied strategies for forecasting solar irradiance according to the methodology outlined in Figure 3. Moreover, we have decided to study 70 spot points in the Brazilian state of Minas Gerais using three different models and to compare the results.

In the following sections, we present more details about the dataset studied, the information of the ground stations analyzed, and the forecasting models applied in our work.

4.1 Dataset

The dataset used for the work consists of average monthly solar irradiance - All Sky Surface Shortwave Downward Irradiance ($kW-hr/m^2/day$) - collected by NASA satellites during 36 years (1984 - 2020). The data is selected through single point coordinates (latitude and longitude) and uses several attributes. For this study, only the irradiance was considered. To better understand how solar irradiance predictions change, the data was selected to involve all observations available in NASA's access viewer [10].

4.2 Stations

The coordinates used for the data are based on the location of 70 ground stations of the Brazilian Institute of Metheoroly [9] in Minas Gerais. Given the vastness of the territorial extension of the state, five stations with different characteristics (climate, biome) were selected (see Table 1) to exemplify the forecasting process in different locations of the state. To represent the predictions for the remaining stations, the average solar irradiance of all stations was used.

Station ID	City	Latitude	Longitude
A509	Monte Verde	-22.861	-46.043
A519	Campina Verde	-19.539	-49.518
A521	Belo Horizonte	-19.883	-43.969
A522	Serra dos Aimores	-17.798	-40.249
A526	Montalvania	-14.408	-44.404

Table 1: Baseline stations selected to exemplify the forecasting analysis.

4.3 Forecasting models

Following, we present the forecasting models that were used to predict solar irradiance:

SARIMA: Autoregressive Integrated Moving Average model [3] (ARIMA) is one of the leading models for predicting future values in a time series which only considers the series' past values, disregarding the data's seasonality. To account for that, the Seasonal Autoregressive Integrated Moving Average (SARIMA) model attributes significant relevance to the seasonality of the data. Given that solar irradiance data has a well-defined pattern that repeats every 12 months, the SARIMA model was selected.

Holt-Winters: The Holt-Winters method [13] uses a triple smoothing approach, in which there is a weight attributed to all observations used in the algorithm's training, where the more recent data has a more significant impact on the predictions.

LSTM: The Long Short Term Memory machine learning model [6] is one of the most versatile approaches when it comes to predicting future values in a time series, given that its hyperparameters (number of neurons, layers, and optimizer) can be tuned to better suit the context of the problem. The model used in this study uses one hidden layer with 64 neurons with the Relu activation function and the Adam optimizer.

The most critical parameters in the forecasting process are the number of training years, the amount of data used for each prediction, and the prediction window, which dictates how far into the future each model will forecast.

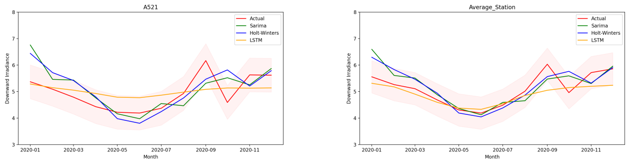
To better understand the impact that the number of training years has on the quality of the predictions, the amount of training data fed to each model varies within the limits of the available data. The minimum number of training years is two, which is increased to the maximum amount of training data: 35 years. Given that the absolute values of solar irradiance vary each year, and to ensure the comparison of the forecasts, all the predictions were made for 2020, with the training data starting in 2019 and going back sporadically till the start of the data (1984).

The absolute values of the predicted solar irradiance for each model are then compared using three metrics (Mean Absolute Error - MAE, Mean Squared Error - MSE, and Root Mean Squared Error - RMSE) calculated based on the predictions for 2020 using all the sets of training years.

5 Experimental Analysis

The main goal of the experimental analysis is to: (i) define the best model for predicting solar irradiance, and (ii) understand the impact of the amount of training data on the precision of each model.

For the first analysis, each model predicted solar irradiance for the test stations using five years of training data to find the model with the lower error. Predictions with five years of training data for station A521 are shown in Fig. 4-(a). The metrics presented in Tables 2 and 3 point out that the Holt-Winters model presented the lowest error, followed by the LSTM and SARIMA models with similar metrics. The Holt-Winters model presented the best results for the majority of the stations, with the LSTM model presenting a lower error in some stations.



(a) Belo Horizonte (b) Average Irradiance

Fig. 4: Predictions using 5 years of training data.

Alternatively, the three models maintained their prediction capabilities when comparing the results to the Typical Meteorological Year, as shown in Fig. 5a. Furthermore, it is worth pointing out the similarity of the Holt-Winters model predictions to the TMY, observed in Fig. 5b, which can be explained by the triple smoothing used in the model, which is very similar to the flatter curve represented by the average used in the TMY.

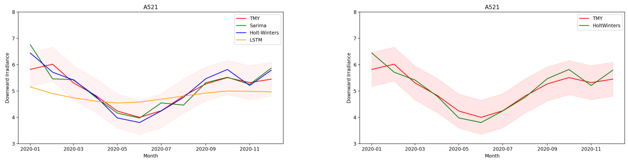
Table 2: Belo Horizonte metrics

Model	MAE	MSE	RMSE
SARIMA	0.2482	0.1341	0.3661
HoltWinters	0.2112	0.0719	0.2682
LSTM	0.2573	0.0885	0.2975

Table 3: Average Irradiance metrics

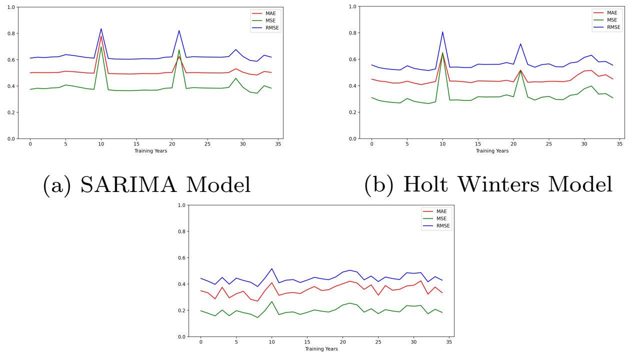
Model	MAE	MSE	RMSE
SARIMA	0.2022	0.0848	0.2912
HoltWinters	0.1366	0.0341	0.1847
LSTM	0.2667	0.0871	0.2952

For the second analysis, each model predicted solar irradiance with varying training data from 2 to 35 years, and its metrics were plotted to visualize their change over time. Before this analysis, a naive assumption would be to assume the lower error for the models would occur with the most significant amount of training data, which was not the case. Fig. 6 shows that the error metrics stayed within a limit regardless of the amount of training data. This behavior can be explained by the nature of solar irradiance, given that there is a lot of variation from year to year regarding absolute values. The peaks for each model's error metrics can be attributed to the inflection points where the added training data has absolute values with a more significant deviation from the predicted year's actual values, making the models over-correct for data that does not represent the absolute values for that year.



(a) Belo Horizonte (b) Holt-Winters

Fig. 5: Comparrison with the Typical Meteorological Year.



(a) SARIMA Model (b) Holt Winters Model

(c) LSTM Model

Fig. 6: Metrics variation for Belo Horizonte.

Furthermore, the optimal amount of training data (resulting in the lowest Mean Squared Error) was used to predict solar irradiance for 2020 for the test stations to help establish a pattern. Table 4 shows that the optimal amount of years worth of training data varies significantly with each station and method, with the LSTM model presenting lower amount of data required to reach the lowest error.

Table 4: Optimal number of training years for each station.

	Station					
Model	A509	A519	A521	A522	A526	Average
SARIMA	4	35	34	36	34	34
Holt Winters	28	26	10	3	9	6
LSTM	11	7	6	3	7	2

In conclusion, despite the three methods presenting satisfactory results of solar irradiance predictions, the Holt-Winters model presented the lowest overhaul error in the test cases and the LSTM model had the most consistent precision with the various amount of data used for each prediction.

6 Conclusions and Further works

The main goal of this work is to compare the prediction of solar irradiance in Minas Gerais using three different models, in addition to better understanding the impact that the amount of training data has on the efficiency of each model. To achieve that goal, each model was trained using increasing data to find the number of training data that would yield the lower overall error in each model. Despite the difference in the absolute values of the predictions compared to the actual values of solar irradiance, the models could forecast values compatible with past patterns. They can be used to estimate the solar irradiance in other regions.

Despite using the most expansive dataset available in the literature, when it comes to the period of observations, this work uses only single-variate models to predict solar irradiance. Given that the solar irradiance that reaches a certain point in the globe can vary with a series of factors, including other variables in the models - such as humidity, pressure, and sky clearness index- can further improve the precision of the predictions. Furthermore, given that the models are trained for each one of the stations used, the development of a more robust model capable of establishing a relation between solar irradiance and geographical factors (like biome and climate region) can help generalize the predictions to the other areas in Brazil and the globe.

References

1. Basílio, S.d.C.A., Putti, F.F., Cunha, A.C., Goliatt, L.: An evolutionary-assisted machine learning model for global solar radiation prediction in minas gerais region, southeastern brazil. Earth Science Informatics pp. 1–19 (2023)
2. Belmahdi, B., Louzazni, M., El Bouardi, A.: One month-ahead forecasting of mean daily global solar radiation using time series models. Optik **219**, 165,207 (2020)
3. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time series analysis: forecasting and control. John Wiley & Sons (2015)
4. Diniz Augusto, R.A., Maia, C., GUIMARAES, S.: Development of a solar radiation map using artificial intelligence techniques. Procceedings of the 19th Brazilian Congress of Thermal Sciences and Engineering (2021)
5. de Minas e Energia, M.: Plano nacional de energia 2030. https://www.gov.br/mme/pt-br/assuntos/secretarias/spe/publicacoes/plano-nacional-de-energia/plano-nacional-de-energia-2030/relatorio-final/plano-nacional-de-energia-2030-pdf.pdf/view (2007)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997). DOI 10.1162/neco.1997.9.8.1735
7. IEA: Energy statistics data browser. https://www.iea.org/data-and-statistics/data-tools/energy-statistics-data-browser?country=WORLD&fuel=Energy%20supply&indicator=ElecGenByFuel (2021)
8. IEA: Energy statistics data browser. https://www.iea.org/data-and-statistics/data-tools/energy-statistics-data-browser?country=BRAZIL&fuel=Energy%20supply&indicator=ElecGenByFuel (2021)
9. INMET: Catálogo de estações automáticas. https://portal.inmet.gov.br/paginas/catalogoaut (2023)
10. NASA: Prediction of worldwide energy resource (power), data access viewer enhanced (dave). https://power.larc.nasa.gov/beta/data-access-viewer/ (2023)
11. Sorkun, M.C., Paoli, C., Incel, Ö.D.: Time series forecasting on solar irradiation using deep learning. In: 2017 10th international conference on electrical and electronics engineering (ELECO), pp. 151–155. IEEE (2017)
12. Western, B.: The estimation of solar radiation in new zealand. Solar Energy **45**(3), 121–129 (1990). DOI https://doi.org/10.1016/0038-092X(90)90046-F. URL https://www.sciencedirect.com/science/article/pii/0038092X9090046F
13. Winters, P.R.: Forecasting sales by exponentially weighted moving averages. Management science **6**(3), 324–342 (1960)