

# The Impact of Meteorological Data Quality on Accurate Forecast of Solar Power Plant Production

Ana Stojkić<sup>1</sup>, Donata Borić<sup>1</sup>, Leila Luttenberger Marić<sup>1</sup>, Vladimir Špišić<sup>1</sup>

<sup>1</sup> KONČAR-Digital Ltd. For Digital Services, Fallerovo šetalište 22, HR-10000, Zagreb, Croatia

E-mail: [ana.stojkic@koncar.hr](mailto:ana.stojkic@koncar.hr), [donata.boric@koncar.hr](mailto:donata.boric@koncar.hr), [leila.luttenberger@koncar.hr](mailto:leila.luttenberger@koncar.hr), [vladimir.spisic@koncar.hr](mailto:vladimir.spisic@koncar.hr)

**Abstract** — As we strive to mitigate the impact of global warming, RESs (Renewable Energy Source) have become indispensable in shaping a sustainable future. However, as RESs play an increasingly significant role, power grids are facing greater volatility in maintaining a power system stability. Among the most important renewable energy sources is solar power, whose production is highly contingent upon weather conditions and whose prevalence continues to grow. The forecast accuracy is directly linked to the quality of meteorological data, which must be precisely measured and reflect specific local conditions. Usually, relying on a single data source alongside with different data analysis approaches and feature selection methods can yield varying forecast outputs. This research aims to conduct a comparative analysis of how meteorological data from various sources, combined with models ranging from simple linear regression to advanced deep learning neural networks, can influence the accuracy of forecasts, ultimately enhancing the reliability of solar power production.

**Keywords:** *data sources, forecasting accuracy, power grid, solar power*

## I. INTRODUCTION

In the collective effort to achieve agreed global warming reduction targets [1] and decrease reliance on conventional energy sources, RESs (Renewable Energy Sources) play an indispensable role in shaping a sustainable future. With the increasing integration of RESs, power grids are becoming more complex, making it increasingly challenging to maintain system stability, i.e. the balance between energy production and consumption. Among these RESs, solar energy stands out due to the rapid installation of solar panels and hence its significant contribution to energy production. Predictions for 2050 estimate that the installed capacity of solar power will exceed 7,000 TWh [2]. According to the International Energy Agency, solar and wind energy are forecasted to double by 2028 compared to 2024, reaching nearly 710 GW [3]. Due to the unpredictability and volatility of solar energy production, achieving more accurate predictions of solar power plant output power is essential. These predictions offer several key benefits, including better integration of SPPs (solar power plant) into the power system, optimal electricity usage, increased participation in the energy and ancillary services

markets, and a positive impact on the environment [4]. To achieve these goals, multiple forecasting models are developed, followed by calculated metrics for each model using various sources of meteorological data. According to [5] the most significant impact on solar generation have meteorological data such as: GHI (Global Horizontal Irradiance), ambient temperature, WS (wind speed), air pressure etc. Different data sources can have varying levels of accuracy, which can significantly impact the precision of the forecasts. While meteorological stations installed at solar power plants should provide relevant local data, online sources data can be useful in cases where local meteorological stations are unavailable, their data is incomplete, or they are too costly. It has been established that mathematical algorithms for evaluating features provide accurate estimates for the above-mentioned meteorological variables.

## II. APPROACHES TO FORECASTING SOLAR POWER PLANT OUTPUT POWER

Given the goals of reducing the carbon footprint and increasing the share of renewable energy usage [1], the importance of high-quality models for RESs output power forecasting is indisputable. Developing accurate solar power generation prediction model requires high-quality meteorological data, in addition to a well-defined process for building and evaluating the model. A typical data analysis process involves several steps, including data acquisition, data preprocessing, algorithm selection and modeling, followed by the evaluation of performance metrics. Data acquisition involves determining data sources, collecting data from them, and storing it for the purpose of training and testing machine learning algorithms [6]. Some SPPs have nearby meteorological stations that measure weather features which significantly influence solar power generation. If no meteorological station is available near the SPP, open-source data can be employed.

During the preprocessing stage of model construction, apart from cleaning the data, highly correlated features should

be carefully considered to improve model accuracy. However, including too many features can lead to overfitting, which reduces the overall quality of the model. After data preprocessing, selecting the appropriate model becomes the next critical step. For solar power generation predictions, XGBoost and LSTM (Long Short-Term Memory) models, often combined with other types of NNs (Neural Network) have proven to be highly effective [7]. XGBoost represents an advanced optimized distributed gradient boosting library known for its high efficiency, flexibility, and great performance speed [8]. LSTM, a type of neural network used in supervised learning, effectively captures feature dependencies and patterns within the data, hence appears as valuable choice for predicting output power of SPP [7].

Among stochastic models, ARIMA (Autoregressive Integrated Moving Average) remains one of the most widely used approaches for solar generation forecasting [9].

The final accuracy of the above-mentioned models is significantly influenced by the selected hyperparameters. However, there is limited research on how the quality of data affects model accuracy. By comparing data sources and modelling techniques, this research offers practical guidelines for enhancing forecast reliability, thereby supporting power grid stability, and contributing to a more resilient energy system.

## III. CASE STUDY

Accurately forecasting the SPP output power remains a challenging task, initially due to the variability in the quality of meteorological data. To address this, the following case study explores the impact of different meteorological data sources on the accuracy of prediction models. The analysis is based on the data from the Desert Knowledge Australia Solar Centre at the Yulara SPP in Northern Territory, Australia, located at coordinates  $-25.3392^\circ$  latitude and  $131.0325^\circ$  longitude [10]. The power plant has an installed capacity of 326.7 kW, and the data spans from January 2017 to December 2019, with a 5-minute sampling interval. The Yulara solar

system includes a meteorological station, whose data will be utilized as real meteorological data in this study. Additionally, open-source meteorological data used in this study were obtained from open-meteo.com [11], the NSRDB (National Solar Radiation Database) [12], and the PVGIS (Photovoltaic Geographical Information System) [13]. These platforms leverage the geographical coordinates of the solar power plant and employ mathematical algorithms to approximate key meteorological parameters mandatory for the analysis.

Figures 1,2, and 3 illustrate the comparison between real meteorological data from associated meteorological station and the data obtained from PVGIS. A significant disparity is observed, particularly in the wind speed values, where the measurements provided by the PVGIS differ heavily. Based on these deviations, models utilizing open-source meteorological data, especially wind speed data, are expected to demonstrate lower accuracy.

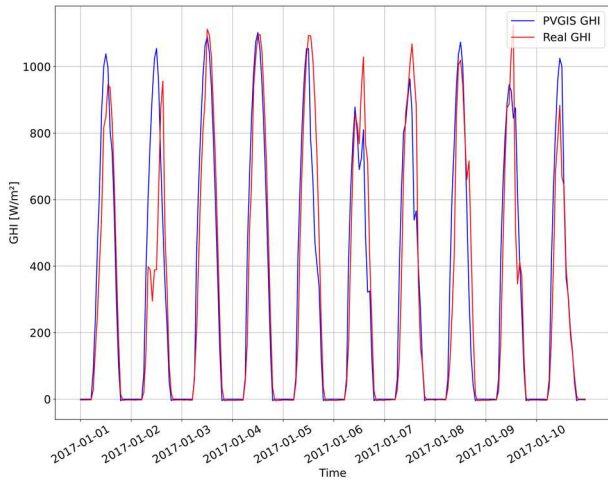


Figure 1. The difference between the GHI from real meteorological data and PVGIS

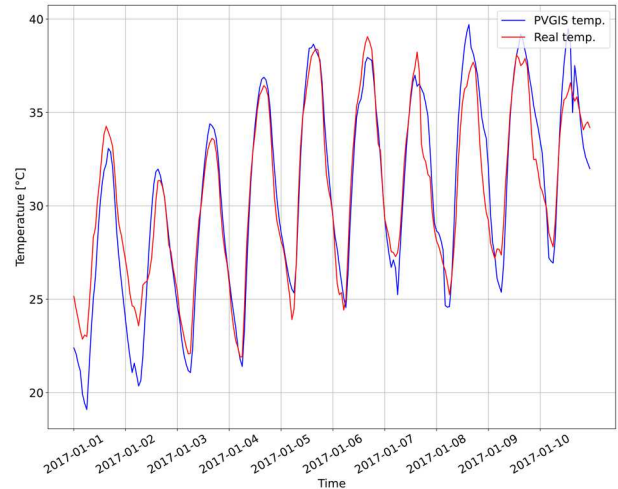


Figure 2. The difference between the temperature from real meteorological data and PVGIS

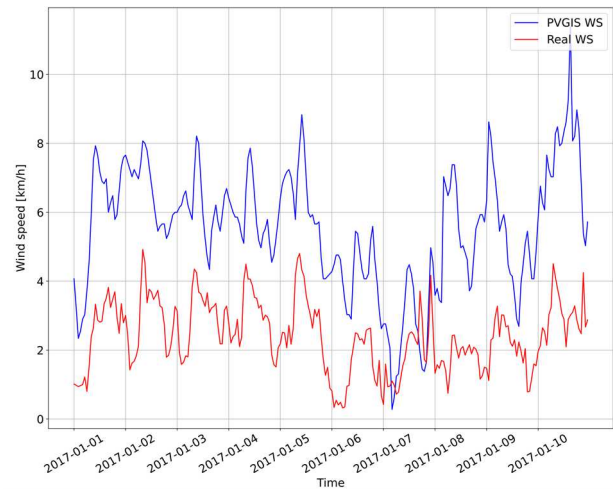


Figure 3. The difference between the wind speed from real meteorological data and PVGIS

During the preprocessing stage, missing values were addressed using a relatively new approach known as MICE (Multiple Imputation by Chained Equations) [14]. This method involves an initial imputation strategy to fill in missing values, followed by iterative refinements. The data was aggregated to an hourly resolution, resulting in 26,280 observations spanning three years. Pearson’s and Spearman’s correlation coefficients were used to identify the variables with the most significant impact on the output power of the SPP. Based on the analysis, three key features were selected: GHI, ambient temperature, and wind speed (WS). To enhance the efficiency and speed of

model training, the features were scaled using *StandardScaler*. The prediction of SPP output power was executed using ARIMA, XGBoost, and a combination of LSTM and CNN (Convolutional Neural Network) models.

Consequently, the analysis compares actual vs. predicted values of SPP output power under different feature combinations: (1) using only GHI, (2) using GHI and ambient temperature, and (3) using all three features: GHI, ambient temperature, and WS. Figures 4 and 5 illustrate comparison between predictions based on real meteorological data and open-source data from NSRDB. As expected, these figures clearly demonstrate that more accurate predictions were obtained using original meteorological data.

The models have been evaluated using MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) and  $R^2$  (Coefficient of Determination) metrics. Among the tested models, XGBoost achieved the best results across all three evaluation metrics, as shown in Tables 1 and 2.

XGBoost model with included all three features from real meteorological data accomplished the best result with MAE = 5.16 kW and  $R^2 = 0.99$ . Incorporating all three features from the NSRDB into the combined LSTM/CNN model yields the best performance metrics across all open-source meteorological data applications, as shown in Tables 1 and 2 (MAE = 6.71 kW,  $R^2 = 0.98$ ).

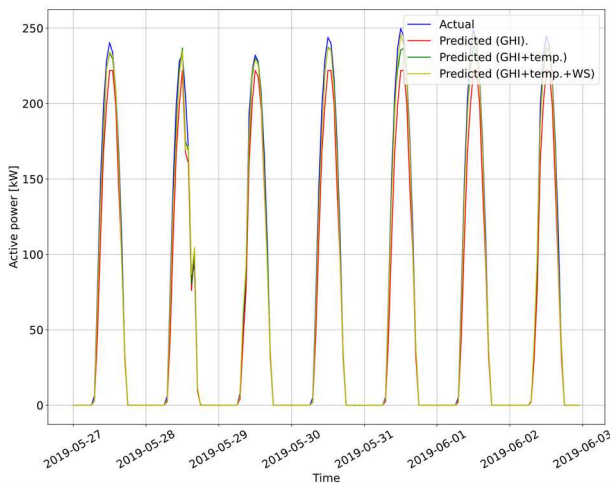


Figure 4. XGBoost model with original meteorological data

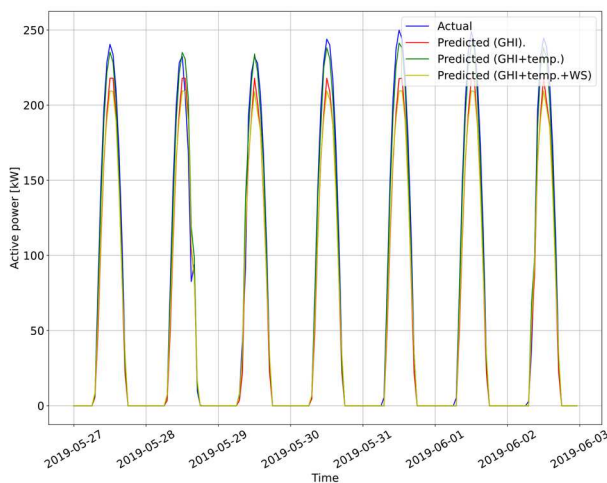


Figure 5. XGBoost model with NSRDB meteorological data

Table 1. MAE metric for all used models

		<i>MAE metric [kW]</i>				
<i>Model</i>	<b>Source of meteorological data</b>				<b>Feature</b>	
	original	open- meteo	NSRDB	PVGIS		
<i>XGBoost</i>	7.98	18.82	9.29	19.31	GHI	
	5.18	11.00	6.89	17.24	GHI+T2m	
	5.16	10.96	8.78	17.44	GHI+T2m+WS	
<i>ARIMA</i>	12.61	17.87	12.86	13.30	GHI	
	12.86	29.73	21.74	33.48	GHI+T2m	
	12.77	28.57	20.88	30.94	GHI+T2m+WS	
<i>LSTM</i>	8.09	18.93	9.63	19.78	GHI	
	5.26	11.61	6.91	17.91	GHI+T2m	
	5.49	11.41	6.71	17.73	GHI+T2m+WS	

Table 2. R2 metric for all used models

*R<sup>2</sup> metric*

Through the analysis and visualization of results across all models, feature combinations, and meteorological data sources, it is noticeable that the best predictive performance is achieved with the data utilized from the Yulara meteorological station. However, this study demonstrates a viable alternative; when such a station is unavailable, open-source meteorological data can still be leveraged to develop models with strong performance metrics. While proper data preprocessing and interpretability are essential, this approach provides a substantial groundwork for building high-quality predictions.

IV. CONCLUSION

This study demonstrates the critical impact of meteorological data quality on the accuracy of solar power plant output power prediction. By comparing multiple forecasting models (LSTM/CNN, ARIMA, and XGBoost) using both real and online-sourced meteorological data we highlight the importance of selecting reliable data sources. While the results indicate that models trained on real meteorological data achieve the highest accuracy, predictions based on open-source data sources yield comparable performance, particularly for LSTM and XGBoost models. This underscores the potential for using high-quality estimated meteorological data when on-site measurements are unavailable.

These findings emphasize the need for improving the accessibility and accuracy of meteorological data sources for solar power forecasting. For future research endeavors, integrating real-time weather updates, and refining predictive models should be achieved.

Model	Source of meteorological data				Feature
	original	open-meteo	NSRDB	PVGIS	
XGBoost	0.98	0.87	0.97	0.86	GHI
	0.99	0.94	0.98	0.87	GHI+T2m
	0.99	0.94	0.97	0.87	GHI+T2m+WS
ARIMA	0.96	0.89	0.95	0.88	GHI
	0.96	0.84	0.92	0.82	GHI+T2m
	0.96	0.86	0.93	0.84	GHI+T2m+WS
LSTM	0.97	0.86	0.97	0.85	GHI
	0.99	0.94	0.98	0.86	GHI+T2m
	0.99	0.96	0.98	0.86	GHI+T2m+WS

REFERENCES

- [1] "Directive - EU - 2023/2413 - EN - Renewable Energy Directive - EUR-Lex." Available: <https://eur-lex.europa.eu/eli/dir/2023/2413/oj/eng>
- [2] K. Iheanetu, "Solar Photovoltaic Power Forecasting: A Review," *Sustainability*, vol. 14, p. 17005, Dec. 2022, doi: 10.3390/su142417005.
- [3] Post, Share, Post, Print, Email, and License, "Renewables growth puts COP28 goals within reach, but acceleration is needed: IEA," Utility Dive. Available: <https://www.utilitydive.com/news/renewable-energy-cop28-growth-economics-global/704446/>
- [4] T.-Z. Ang, M. Salem, M. Kamarol, H. S. Das, M. A. Nazari, and N. Prabakaran, "A comprehensive study of renewable energy sources: Classifications, challenges and suggestions," *Energy Strategy Rev.*, vol. 43, p. 100939, Sep. 2022, doi: 10.1016/j.esr.2022.100939.
- [5] "Effects of different environmental and operational factors on the PV performance: A comprehensive review - Hasan - 2022 - Energy Science & Engineering - Wiley Online Library." Available: <https://scijournals.onlinelibrary.wiley.com/doi/full/10.1002/ese3.1043>
- [6] "What is Data Acquisition in Machine Learning?," GeeksforGeeks. Available: <https://www.geeksforgeeks.org/what-is-data-acquisition-in-machine-learning/>
- [7] I. Benitez, J. Ibañez, C. Lumabad, J. Cañete, and J. Principe, "Day-Ahead Hourly Solar Photovoltaic Output Forecasting Using SARIMAX, Long Short-Term Memory, and Extreme Gradient Boosting: Case of the Philippines," *Energies*, vol. 16, p. 7823, Nov. 2023, doi: 10.3390/en16237823.

- [8] “ML | XGBoost (eXtreme Gradient Boosting),” GeeksforGeeks. Available: <https://www.geeksforgeeks.org/ml-xgboost-extreme-gradient-boosting/>
- [9] C. Tarmanini, N. Sarma, C. Gezezin, and O. Ozgonenel, “Short term load forecasting based on ARIMA and ANN approaches,” *Energy Rep.*, vol. 9, pp. 550–557, May 2023, doi: 10.1016/j.egy.2023.01.060.
- [10] “Data Download | DKA Solar Centre.” Available: <https://dkasolarcentre.com.au/download?location=yulara>
- [11] P. Zippenfenig, *Open-Meteo.com Weather API*. (Dec. 31, 2024). Zenodo. doi: 10.5281/ZENODO.7970649.
- [12] “NSRDB.” Available: <https://nswrdb.nrel.gov/>
- [13] “JRC Photovoltaic Geographical Information System (PVGIS) - European Commission.” Available: [https://re.jrc.ec.europa.eu/pvg\\_tools/en/](https://re.jrc.ec.europa.eu/pvg_tools/en/)
- [14] “MICE imputation - How to predict missing values using machine learning in Python - Machine Learning Plus.” Available: <https://www.machinelearningplus.com/machine-learning/mice-imputation/>