STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, February 2026, pp 991–999.
Published online in International Academic Press (www.IAPress.org)

Optimizing Photovoltaic Performance Prediction Using Machine Learning: Analysing the Impact of Environmental Variables in Marrakesh

Mustapha Ezzini[1,*], Raja Mouachi[2], Mohammed Ennejjar[3], Abdelali El gourari[2], Mohammed Boukendil[1], Mustapha Raoufi [4]

[1] *Cadi Ayyad University, UCA, Faculty of Sciences Semlalia, LMFE Unit affiliated to CNRST (URL-CNRTST N° 16), Department of Physics*
[2]*LAMIGEP, EMSI, Moroccan School of Engineering, Marrakesh, Morocco*
[3] *Cadi Ayyad University, UCA, Faculty of Sciences Semlalia, LISI Laboratory, I2SP team, Bd. Prince My Abdellah, Marrakech, Morocco*
[4]*Physics Department, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakesh, Morocco*

Abstract This study focuses on the optimization of photovoltaic (PV) prediction using machine learning (ML) models by analyzing the impact of environmental variables in Marrakech. The research compares two types of meteorological data from satellites and ground stations to assess their respective contributions to forecast accuracy. The results show that global solar irradiance (G), air temperature (Ta) and wind speed (Wv) are the most influential parameters on energy production, whatever the data source. However, forecasts based on ground-measured data showed slightly higher accuracy, with an R^2=0.98 for measured data versus 0.86 for stalietes data, underlining the importance of localized measurements. Of the scenarios tested, Scenario 1 (all inputs) achieved the highest accuracy, with an R^2 of 0.98 and an RMSE of 91.39. Scenarios 2 (without Wv) and 4(without DNI) also delivered acceptable levels of accuracy, albeit slightly lower than Scenario 1. These results highlight the importance of integrating localized weather data to improve the accuracy of PV power generation forecasts.

Keywords Photovoltaics, Machine Learning, Energy production prediction, Panel temperature, Environmental factors, PV efficiency, Regression models, Energy performance.

DOI: 10.19139/soic-2310-5070-2547

1. Introduction

Electricity production accounts for a significant share of global greenhouse gas (GHG) emissions, which increased by 62% between 1990 and 2022. In light of the growing challenges of climate change, solar photovoltaic (PV) energy plays a critical role in the energy transition. In 2022, PV energy production reached a record high of 1,300 TWh, marking a 26 % increase compared to the previous year. By 2023, this technology accounted for three-quarters of the new renewable capacity installed worldwide, emphasizing its strategic importance in meeting global energy demands while reducing carbon emissions. The development of renewable energy, particularly solar photovoltaic power, represents a major challenge in the global energy transition. As the demand for clean energy continues to grow, solar and wind power additions are expected to more than double by 2028 compared to 2022, reaching nearly 710 GW [1]. Photovoltaic panels, which convert solar energy into electricity, play a key role in this transition by reducing GHG emissions and reliance on fossil fuels.However, the performance of photovoltaic panels is influenced by various environmental factors, including surface temperature, which has a direct impact on their efficiency [2]. Indeed, it is generally observed that the performance of photovoltaic panels

*Correspondence to: Mustapha Ezzini (Email: m.ezzini.ced@uca.ac.ma). Physics Department, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakesh, Morocco).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

degrades with increasing temperature, presenting a significant challenge to maximizing their efficiency [3]. This variability requires an understanding of the key factors affecting energy production, with the aim of optimizing the output of photovoltaic systems under variable climatic conditions. This study focuses specifically on modeling the energy production of monocrystalline silicon photovoltaic panels, a technology renowned for its high efficiency. Through this modelling, we seek to establish a relationship between variable panel surface environments and energy yield, while identifying the key variables that influence this relationship. To do this, we integrate artificial intelligence approaches [4], including machine learning algorithms, to predict the energy yield and temperature of monocrystalline panels. These techniques enable us to fine-tune the performance of photovoltaic systems in various environments and optimize solar power generation strategies[5]. The integration of these modeling and machine-learning technologies offers a promising route to improving the efficiency of photovoltaic panels, guaranteeing a significant contribution to the transition to sustainable, renewable energy sources [6].

2. Materials and methods

In this work, we studied the photovoltaic installation with a capacity of 2.04 kWp, which was installed on the roof of the Semlalia Faculty of Sciences at Cadi Ayyad University in Marrakech, Morocco. The system consists of eight mono crystalline (m-Si) panels, each with a power of 255 Wp, connected in series to obtain the total power [7]. The installation was placed at a fixed 30° south-facing angle of inclination, thus optimizing solar exposure in the hot semi-arid climate of Marrakesh. The data for this project were collected over a period from 29 April 2016 to 7 May 2017.

Figure 1. Photovoltaic panels installation.

Figure 2 presents the temporal variations of the main environmental parameters that influence the production of photovoltaic energy from monocrystalline panels, derived both from satellite and terrestrial measurements. The first figure shows the variation of the global horizontal irradiance (G) over time, highlighting fluctuations in the solar energy received on a horizontal surface, which is essential for assessing solar potential. The second graph shows the variation of direct normal irradiance (DNI) over time, reflecting the solar radiation received by a surface perpendicular to the sun's rays. The third graph depicts wind speed variations, showing atmospheric movement fluctuations that impact ventilation and renewable energy systems. Lastly, the fourth graph illustrates changes in ambient temperature (Ta) over time, which significantly influence energy production and climatic conditions[8]. Peaks in irradiance around 400 to 800 W/m² and wind speeds mostly below 2 m/s are observed, along with ambient temperature variations crucial for understanding photovoltaic performance. This figure consolidates the distribution of key environmental variables measured both on the ground and by satellite.
This figure 3 illustrates a typical scientific workflow for the construction and evaluation of a machine learning model applied to data prediction, potentially for photovoltaic panels. The process begins with the acquisition of raw data, followed by a data cleaning phase to remove outliers and fill in missing values, thus ensuring better

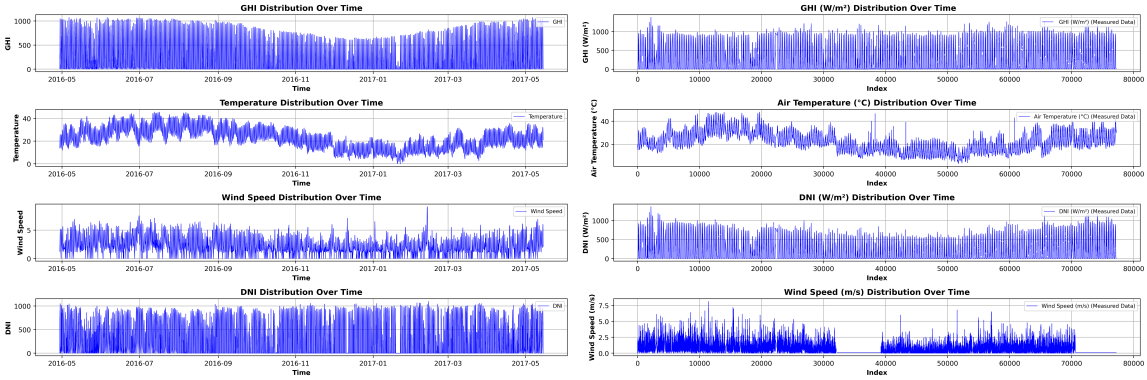

Figure 2. Variations in environmental parameters based on satellite and measured data

quality input data [9].
Next, a normalization step is applied to bring the data to the same scale, which is necessary to improve the convergence of machine learning algorithms.
After data preparation, a model is selected, which will be trained according to four different scenarios, each corresponding to a specific configuration of parameters or hyperparameters. The learning loop is regulated by a stopping criterion, represented by , which symbolizes an error threshold to be reached to guarantee acceptable performance. If the model fails to reach this threshold, it continues its training to improve its performance . Once the shutdown criterion is met, the optimal scenario is selected based on a set of evaluation parameters such as mean square error, mean square error or mean absolute error. The resulting model can then be used to make predictions about new data that provide a better predictive model.

2.1. Multilayer Perceptron (MLP)

Figure 4 shows a multi-layer Perceptron (MLP), a model commonly used in machine learning [10]. our MLP model consists of an input layer, two hidden layers and one output layer. The network inputs correspond to the explanatory variables of the problem studied. Each hidden layer contains 100 neurons and is fully connected to the previous layer, which means that each neuron in a layer is connected to all neurons in the second layer. Neurons in the hidden layers apply activation functions to introduce non-linearities, allowing the model to learn complex relationships in the data. Finally, the output layer provides prediction of the energy production model.

2.2. Model Architecture and Min-Max Normalization

The prediction model employs a Multi-Layer Perceptron (MLP) with optimized hyperparameters, including two hidden layers. Input features are normalized using min-max scaling to ensure numerical stability during training:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad \text{where} \quad X \in \mathbb{R}^{n \times m}$$

The Rectified Linear Unit (ReLU) activation function ($\sigma(z) = \max(0, z)$) is adopted for its efficiency in mitigating vanishing gradients. The forward propagation is mathematically expressed as:

$$\hat{y} = W_2 \sigma(W_1 X_{\text{norm}} + b_1) + b_2$$

where W_i and b_i represent the weight matrices and bias terms, respectively. Hyperparameters are tuned via 5-fold cross-validation, with the learning rate ($\eta = 0.001$) . This architecture demonstrates consistent performance across seasonal variations while maintaining computational efficiency.

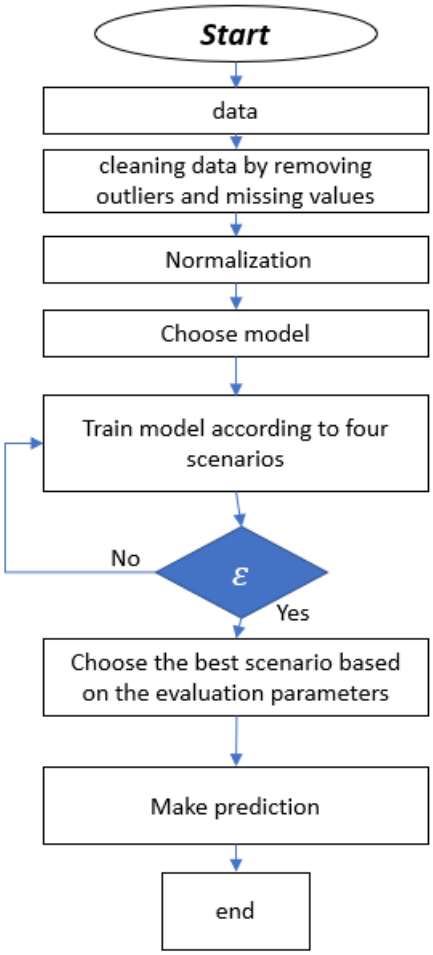

Figure 3. Workflow for the implemented methodology.

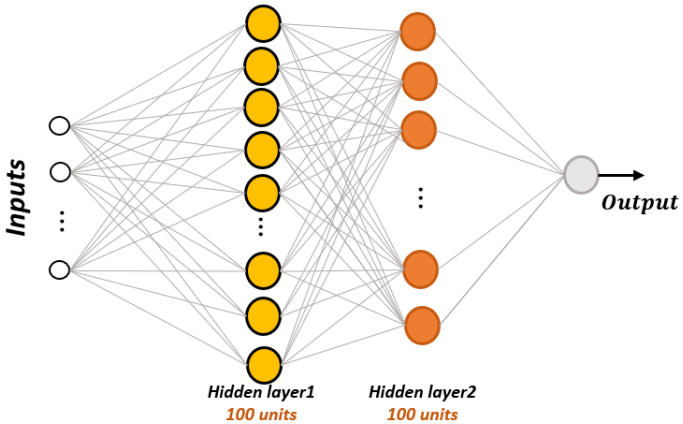

Figure 4. Suggested model structure.

In this context, the input variables introduced into the MLP are G (W/m 2), h (W/m 2), Wv (m/s) and Ta (°C), which respectively represent horizontal and inclined solar irradiance, wind speed and ambient temperature. These inputs are key to accurately predict the output, which in this case is the Pdc1 (W), the output power of the monocrystalline photovoltaic module. The role of the MLP is to learn how these variables influence the performance of the PV module and make accurate predictions based on new data [11]. Table 1: Different scenarios using a variable set of measured or satellite data.

Table 1. Different scenarios using a variable set of measured or satellite data

Scenario	G (W/m²)	h (W/m²)	Wv (m/s)	Ta (°C)
S1	×	×	×	×
S2	×	×		×
S3	×	×	×	
S4	×		×	×
S5		×	×	×

The table 1 presents different study scenarios using a variable set of measured or satellite data. Each scenario includes specific variables such as horizontal overall irradiance (G), inclined or normal irradiance on PV surface (h), wind speed (Wv) and air temperature (Ta). The selection of variables in each scenario is essential for the accuracy of the analysis. For example, scenario S1, which uses all available variables, provides a comprehensive analysis, taking into account the combined impact of irradiance, wind and temperature on the performance of photovoltaic panels. In contrast, scenario S2 excludes the variable Wv, which may reduce accuracy, particularly in wind speeds. Scenario S3 focuses on G, h and Wv, omitting temperature, which may be appropriate in environments where Ta has a limited effect. The choice of measured or satellite data directly influences the quality of predictions and the reliability of results. Therefore, a rigorous selection of variables is essential to obtain accurate models that are appropriate for the specific conditions of the study area regardless of the type of data.

To evaluate the results obtained, we used RMSE, MAPE and R2 as evaluation metrics [12].

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

$$\text{MAPE} = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100 \tag{2}$$

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{k=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

Where:

- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.
- $\bar{y}$ is the average of power production.
- n is the number of observations.

3. Results and discussion

Figure 5 presents a comprehensive comparison of prediction performance through scatter plots of the coefficient of determination (R^2) and error distributions across five scenarios using measured data (blue) and satellite data (yellow). The first row of graphs illustrates the R^2 values, while the second row displays the corresponding error distributions.

In Scenario 1, measured data achieves a high R^2 of 0.98 with an RMSE of 91.39 W, indicating strong predictive accuracy. However, for satellite data, the R^2 decreases to 0.86, while the RMSE significantly increases to 234.22 W, suggesting greater discrepancies between predictions and observations.

In Scenario 3 (without T_a), measured data maintains a strong R^2 of 0.95, while satellite data shows an RMSE of 255.4 W. Additionally, the error distribution exhibits significant widening, reaching extreme values between -650 W and +650 W, indicating increased variability in predictions.

In Scenario 5, the difference is even more pronounced. The R^2 for measured data remains relatively high at 0.79, but for satellite data, it drops sharply to 0.55, indicating a substantial loss in explanatory power. Similarly, the RMSE rises from 179.8 W to 389 W, reflecting increased prediction errors.

These results suggest that satellite data (yellow) introduces additional uncertainties compared to measured data (blue), particularly in scenarios with complex variations or missing key parameters such as T_a. While satellite data provides an alternative source for modeling, its predictive accuracy appears lower. The observed discrepancies highlight the need for improved satellite data processing techniques to enhance reliability in performance predictions.

This figure 6 highlights the influence of some specific input variables on the accuracy of photovoltaic power predictions in five scenarios. The analysis shows the effects of removing one or more variables: global horizontal irradiance (G), direct normal irradiance (h), wind speed (Wv) and ambient temperature (Ta) - using both measured and satellite data.

In scenario 1 (S1), all input variables are included, resulting in the most accurate forecast of 0.98 for R2 with a high consistency between measured and expected power. This scenario serves as a reference and demonstrates the importance of a complete data set by input to Scenario 2 (S2) excluding wind speed (Wv), resulting in minor deviations especially during high power periods and giving R2 0.97.

In the same sense, wind speed plays a secondary role, its absence slightly reduces the accuracy of the MLP. Scenario 3 (S3) excludes ambient temperature (Ta), resulting in some inaccuracies of our model during peak and low power periods. Temperature is essential to capture the thermal effects on the performance of the photovoltiac installation, and its absence affects the performance of preflight with RMSE increase from 91 to 97 W.

In scenario 4 (S4), the normal irradiance (h) is eliminated , which has a significant impact on the model's ability to predict power accurately. Direct irradiance is essential to understand power variations under a clear sky, and its absence leads to well-demonstrable deviations , especially in satellite data gives results similar to S2 salt. Finally, scenario 5 (S5) excludes overall horizontal irradiance (G), which has the most severe effects on the performance of our model. As the main driver of photovoltaic power generation, the absence of G leads to significant prediction errors, which underlines its critical importance with R2 of 0.91 measured data and 0.55 for satellite data .

Overall, the figure shows the important role of irradiance variables (G and h) in accurate power prediction especially on second type of data, while the exclusion of secondary variables such as Wv and Ta leads to less pronounced but noticeable effects. This study demonstrates the need for a complete inclusion of variables for a reliable forecast of photovoltaic power with a minimum error.

Figure 7 compares prediction performance using measured data (left) and satellite data (right), with emphasis on RMSE and R2 in five scenarios (S1 to S5). Each scenario eliminates one of the specific entries: wind speed (Wv), ambient temperature (Ta), direct normal irradiance (DNI or h) or global irradiance (g). The results show clearly that the global irradiance (g) is a key variable to increase accuracy predictions, since its exclusion considerably reduces performance. Ambient temperature (Ta) improves accuracy when included, whereas DNI and Wv may influence predictions but have a smaller impact than Ta and G. Measured data consistently show higher R2 and lower RMSE, This highlighted the importance of using measured inputs to achieve reliable forecasts. The coloured stars represent the different scenarios, highlighting the role of each feature in clarifying the model.

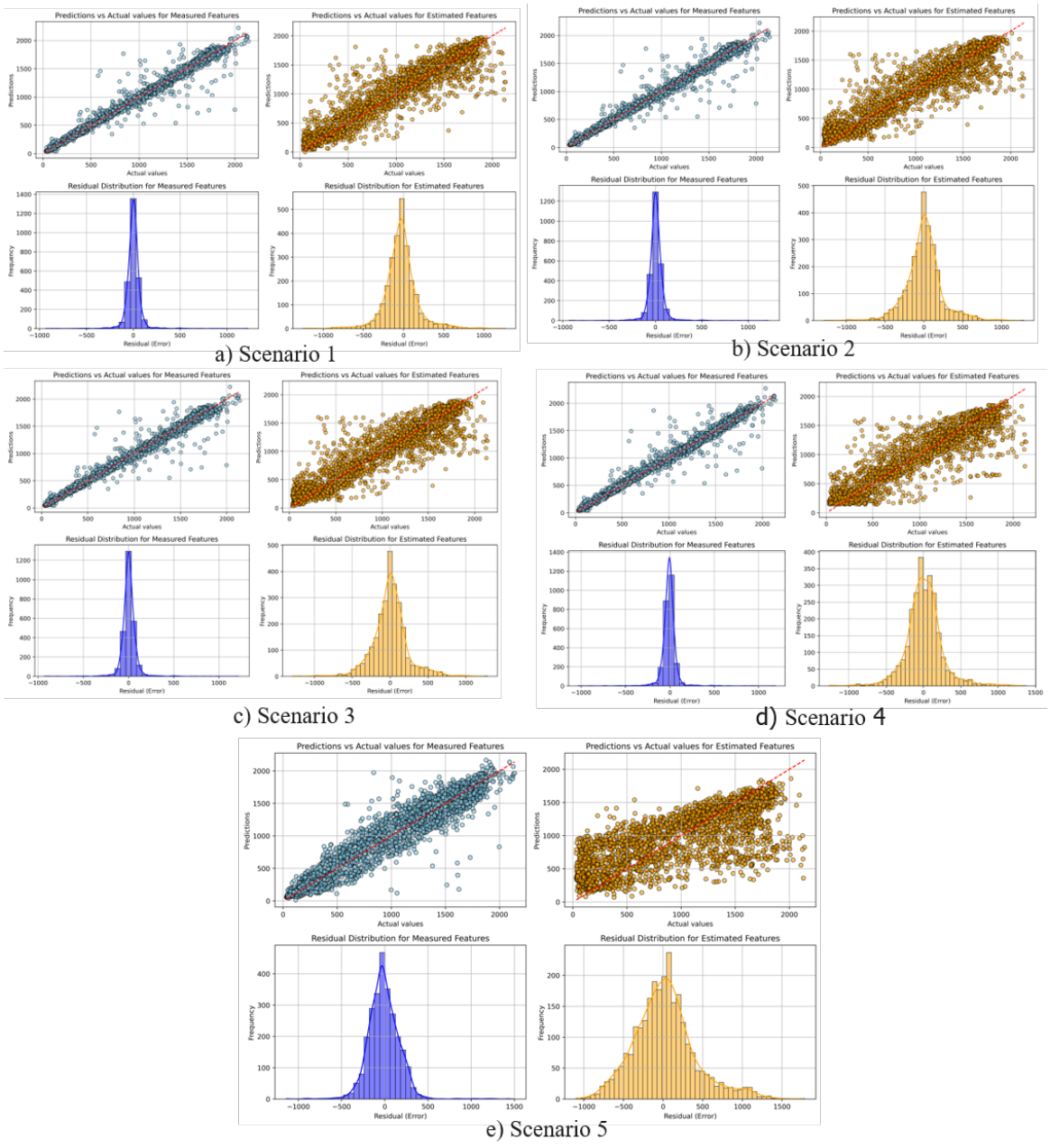

Figure 5. Scatter plot and error distribution for the five scenarios using both measured and satellite data.

4. Conclusion

This trail shows the importance of complete data (all inputs) and solid models for accurate prediction of photovoltaic power generation. The use of satellite data as a basis for prediction resulted in an R2 of 0.86 and an RMSE of 286 W for complete data. on the other hand, the exclusion of key variables, mainly global horizontal irradiance (GHI), in scenario 5 led to a significant performance decline, with an R2 falling to 0.55. These results demonstrate the limitations of using satellite data for local variations, and highlight the critical and important role of

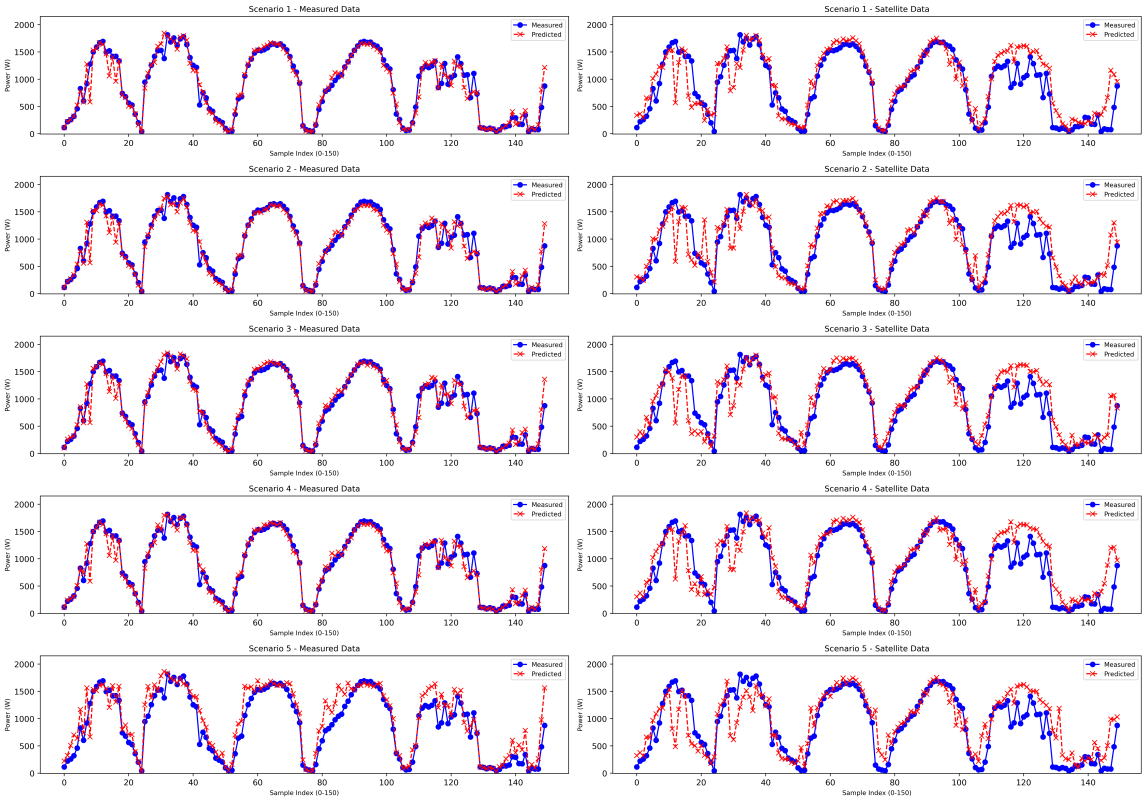

Figure 6. Comparison between measured and predicted data in the test phases accross different scenarios.

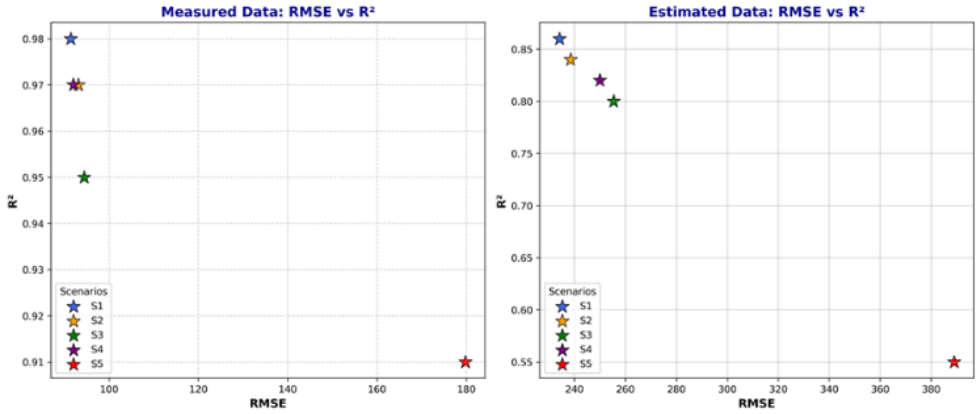

Figure 7. Comparison between the obtained performances by the five scenarios in term of R^2 and RMSE.

irradiation in predictive models. On the other hand, the inclusion of all key variables (global horizontal irradiance, direct normal irradiance, wind speed and ambient temperature) in Scenario 1 produced the best performance, with an R2 of 0.98 and an RMSE of 91.39. The current exclusion of ambient temperature shows senificative increase of RMSE towards 96.5 wIt underlines the need for precise local measurements to obtain accurate forecasts of photovoltaic power. The results highlight the value of comprehensive and high-quality data sets as well as advanced prediction techniques to optimize the performance of photovoltaic systems, particularly in warm and variable climate regions, improving energy efficiency and system reliability.

REFERENCES

1. *International Energy Agency (IEA),* , nternational Energy Agency (IEA)," 2024. [Online]. Available: https://www.iea.org/energy-system/electricity.
2. G. Singh, *Solar power generation by PV (photovoltaic) technology: A review*, Energy, vol. 53, pp. 1-13, 2013.
3. F. Mavromatakis, *Estimating the maximum power of a photovoltaic array*, Institut pédagogique de Crète, Département des sciences.
4. G .Priya ,S. Rhythm, *PV power forecasting based on data-driven models: a review.*, International Journal of Sustainable Engineering, 2021, vol. 14, no 6, p. 1733-1755.
5. Iheanetu, K. J. *Solar photovoltaic power forecasting: A review.*, Sustainability, (2022), 14(24), 17005.
6. Id Omar Nour-eddine, Boukhattem Lahcen, Oudrhiri Hassani Fahd, Bennouna Amin, *Power forecasting of three silicon-based PV technologies using actual field measurements*, vol. 43, no. 100915, 2021.
7. N. Aarich, A. Bennouna, N. Erraissi, M. Raoufi and A. Asselman, *Assessment the long-term performance ratio maps of three grid-connected photovoltaic systems in the Moroccan climate*, Energy for Sustainable Development, vol. 79, no. 101388, avril 2024.
8. Kim, J., Obregon, J., Park, H., Jung, J. Y. *Multi-step photovoltaic power forecasting using transformer and recurrent neural networks.*, Renewable and Sustainable Energy Reviews, (2024),200, 114479.
9. Ezzini, M., Mouachi, R., Ennejjar, M., El Gourari, A., Boukendil, M., and Raoufi, M. (2024, December), *Comparative Analysis of Monocrystalline and Polycrystalline Photovoltaic Performance in Arid Climates Using Random Forest for Missing Data Completion*, In 2024 International Conference on Decision Aid Sciences and Applications (DASA) (pp. 1-5). IEEE.
10. Cristian-Dragos Dumitru,Adrian Gligor, Calin Enachescu *Solar Photovoltaic Energy Production Forecast Using Neural Networks*, Procedia Technology, vol. 22, pp. 808-815, 2016.
11. N. C. Niranjan Singh Baghel, *Performance comparison of mono and polycrystalline silicon solar photovoltaic modules under tropical wet and dry climatic conditions in east-central India*, Vols. Clean Energy, Volume 6, Issue 1, Pages 165–177, February 2022.
12. M. Ennejjar, S. Chabaa, A. Khabba, S. Ibnyaich and A. Zeroual, *Completing a dataset for a passive house using a Feedforward Neural Network with Adam optimization*, in In 2024 International Conference on Global Aeronautical Engineering and Satellite Technology (GAST) (pp. 1-6). IEEE., Marrakech, (2024, April).