ASSESSMENT OF REANALYSIS DATA FOR SOLAR PV OUTPUT FORECASTING IN THE PHILIPPINES: CASE OF PANGASINAN, NEGROS OCCIDENTAL, AND DAVAO DEL NORTE

C. J. A. Gavina[1], J. A. Ibañez[2], I. B. Benitez [2,3], C. D. Lumabad III [2,4], J. A. Principe [5*]

[1] Department of Electrical Engineering, Pamantasan ng Lungsod ng Maynila - cjagavina2020@plm.edu.ph
[2] National Engineering Center, University of the Philippines Diliman, Quezon City - jaibanez1@alum.up.edu.ph,
[3] Electrical Engineering Department, FEU Institute of Technology, City of Manila, ibbenitez@feutech.edu.ph
[4] Department of Computer, Electronics, and Electrical Engineering, Cavite State University, Don Severino Delas Alas Campus Indang, Cavite, Philippines - cdlumabad@cvsu.edu.ph
[5] Department of Geodetic Engineering, University of the Philippines Diliman, Quezon City - japrincipe@up.edu.ph

KEY WORDS: ERA5, XGBoost, PCA, Nested-Cross Validation, Solar PV Output Forecasting, Weather Parameters

ABSTRACT:

The sustainable energy transition in the Philippines requires accurate forecasting of solar PV output to optimize energy efficiency and grid management. While existing studies have emphasized the positive correlation between solar irradiance and PV production, this study aims to explore whether forecasting improves with the inclusion of weather data. This research conducts a comparative analysis between relying solely on solar irradiance against integrating various weather parameters to enhance solar PV output forecasting. The study focuses on three distinct locations (Pangasinan, Negros Occidental, and Davao Del Norte) and employs two models per each site: Model 1 (M1), which relies only on solar irradiance as predictors, and Model 2 (M2), which incorporates solar irradiance and weather parameters. Using Fifth Generation ECMWF Reanalysis (ERA5) Data, Principal Component Analysis (PCA) is conducted on the significant weather parameters. Extreme Gradient Boosting (XGBoost) with 5-fold nested cross-validation is applied for solar PV output forecasting. Models are assessed using Mean Absolute Percentage Error (MAPE) and skill scores. Results showthat while solar irradiance alone suffices for predicting solar PV output in Negros Occidental, incorporating weather parameters improves forecasting accuracy in Davao Del Norte and Pangasinan. This paper recommends caution in generalizing the findings to different regions with varying weather patterns, as the forecasting performance of the models is influenced by data quality, specific location, and prevailing weather conditions.

1. INTRODUCTION

The Philippine government's focus on sustainable energy solutions, driven by the need to adopt renewables and reduce carbon emissions, has highlighted the importance of accurate solar PV output forecasting for efficient energy utilization and grid management. In recent years, the country has made remarkable strides in renewable energy generation, marking it as one of the emerging economies and photovoltaic (PV) markets worldwide (Farias-Rocha et al., 2019). Bertheau (2020) emphasized the potential of renewable energy projects to stimulate economic growth and contribute to the realization of sustainable development goals (SDGs). This has spurred intensified research efforts to understand the intricate relationship between solar irradiance, other weather parameters, and solar PV output. Previous studies have revealed a strong positive correlation between solar irradiance and PV production (Das et al., 2018). This highlights the feasibility of using solar irradiance as the sole predictor for forecasting PV output. However, it is widely recognized that solar power production is influenced by various weather parameters, including temperature, humidity, wind speed and direction, cloud cover, and precipitation (Alcañiz et al., 2023; Benitez et al., 2022).

Furthermore, recognizing the pressing need to address climate change while promoting sustainable growth, the renewable energy sector has witnessed notable expansion with a simultaneous focus on optimizing its efficiency by applying artificial intelligence (AI) techniques, as observed by Hannan et al. (2021). In response to the high demand for accurate short-term solar power forecasts, various machine learning methods have gained widespread adoption. Nevertheless, the challenge lies in effectively selecting the most suitable machine learning models and relevant data features. Extreme Gradient Boosting (XGBoost) has gained prominence in solar power forecasting because of its superior performance, as demonstrated by studies conducted by Grzebyk et al. (2023), Dimitropoulos et al. (2021), and Zhong & Wu (2020), showcasing its effectiveness compared with other machine learning techniques. Additionally, Munawar et al. (2020) determined that the combination of the XGBoost method with feature selection via Principal Component Analysis (PCA) surpasses other approaches in terms of predictive accuracy.

This study aims to fill research gaps by exploring the accuracy of relying solely on solar irradiance predictions versus incorporating other weather parameters to improve solar PV output forecasting in the Philippines. A comparative analysis using the Fifth Generation ECMWF Reanalysis (ERA5) Data for solar PV output forecasting was conducted and evaluated the model performance of Extreme Gradient Boosting (XGBoost). By evaluating the forecast accuracy of the models incorporating the two predictor sets, the study seeks to determine the most suitable and effective model to enhance the accuracy and reliability of solar PV output predictions within the context of the Philippine energy landscape. The study's findings hold significant implications for informed decision-making in sustainable energy planning, facilitating the optimal adoption of solar PV systems, and advancing renewable energy initiatives in the country.

* Corresponding author

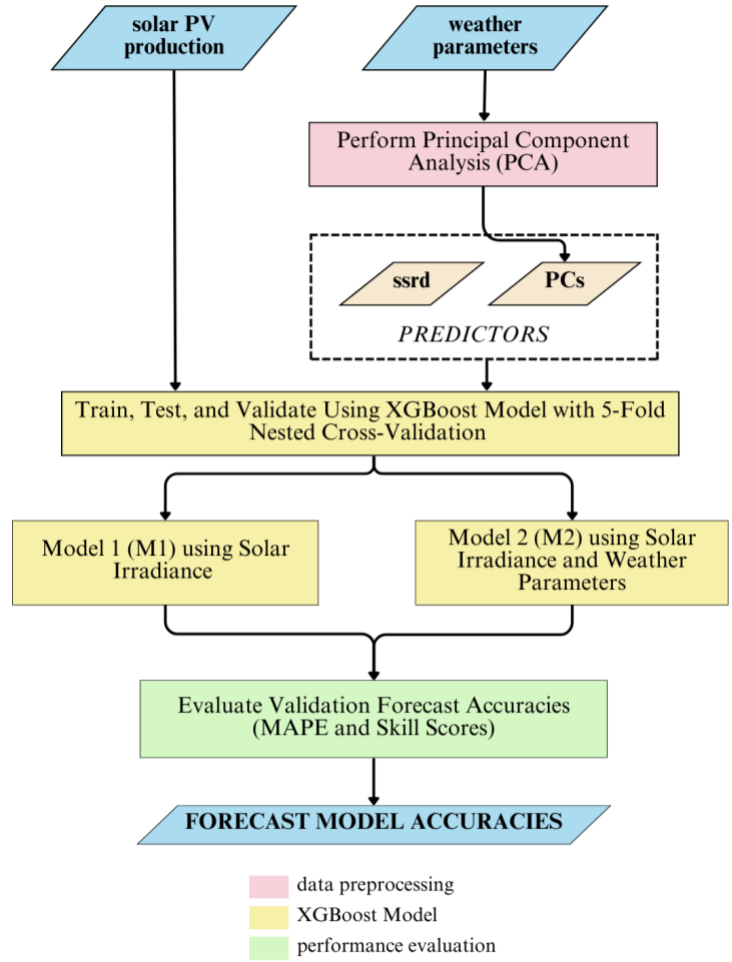

Figure 1. Overview of Methodology.

2. METHODOLOGY

Figure 1 illustrates an overview of the methodology. Utilizing the collected data, weather parameters underwent Principal Component Analysis. After selecting the initial components accounting for the most variation, two predictors were established: solar irradiance alone and the principal components. XGBoost forecasting was implemented using a 5-fold nested cross-validation approach. This iterative process encompassed model training, hyperparameter optimization, model testing, and validation, resulting in the final optimized models: M1 utilizing solar irradiance and M2 integrating solar irradiance and weather parameters. The evaluation of performance involved mean absolute percentage error and skill scores, producing forecast model accuracies as the output.

2.1 Data Collection and Model Setup

This study adopted a comparative approach to evaluate solar photovoltaic output forecasting in the Philippines, incorporating ERA5 data. Table 1 shows the summary of datasets used in the study. The time series length encompassed January 2020 to December 2021 (8:00 am to 5:00 pm, UTC+8), chosen to ensure consistent data availability across all sites. The study examined two models: Model 1 (M1) utilizing solar irradiance solely as predictors, and Model 2 (M2) incorporating both solar irradiance and weather parameters as predictors. These models provided a comprehensive framework to evaluate the predictive efficiency of XGBoost when employing distinct sets of predictors for solar PV output forecasting.

2.2 Data Preprocessing

To mitigate multicollinearity among the key weather parameters—namely, solar irradiation (ssrd), wind speed (ws), wind direction (wd), relative humidity (rh), temperature at 2 meters (t2m), high cloud cover (hcc), low cloud cover (lcc), medium cloud cover (mcc), total cloud cover (tcc), and total precipitation (tp)—Principal Components Analysis (PCA) was performed. This statistical technique reduced the dimensionality of the dataset while preserving the essential information contained within these variables (Bro & Smilde, 2014). The initial stage encompassed the standardization of all variables, evaluation of its covariance matrix, and transformation of the original data values.

PCA then employed eigenvectors and eigenvalues to decompose the covariance matrix. The eigenvectors established orthogonal components, known as principal components, whereas the corresponding eigenvalue quantified the variance (Alskaif et al., 2020). The principal components

Source	Data	Unit	Resolution	
			Temporal	*Spatial*
ERA5	Solar Irradiance (ssrd)	W/m^2	Hourly	27 km^2
	Wind Speed (ws)	m/s		
	Wind Direction (wd)	Cardinal Direction		
	Relative Humidity (rh)	%		
	Temperature at 2m (t2m)	°C		
	High Cloud Cover (hcc) Low Cloud Cover (lcc) Medium Cloud Cover (mcc) Total Cloud Cover (tcc)	octa		
	Total Precipitation (tp)	mm		
Plant (Pangasinan, Negros Occidental, Davao Del Norte)	Production	kW	Hourly	In-situ

Table 1. Summary of Datasets

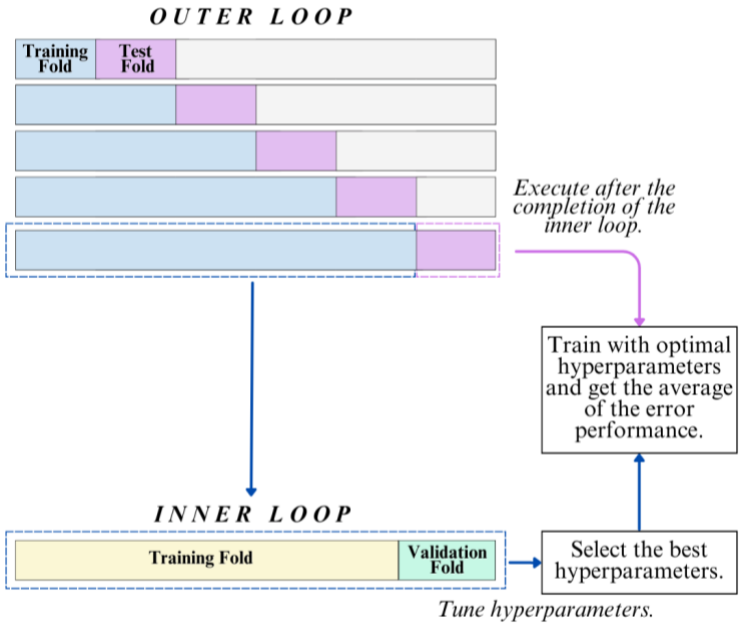

Figure 2. XGBoost Model with 5-Fold Nested Cross-Validation.

were ranked by its eigenvectors in descending order, creating a list of components that was arranged from highest to lowest variance. Using only the initial components that accounted for the most variation allowed for the retention of most of the original information while simplifying the dataset's complexity (Jolliffe & Cadima, 2016). The cumulative proportion of variance explained by these principal components was carefully evaluated, ensuring that they retained the most critical information while eliminating multicollinearity. PCA's characteristics improved the interpretability and accuracy of weather parameter influence analyses, notably in the context of solar PV output forecasting (Chahboun & Maaroufi, 2022).

2.3 XGBoost Model

The application of XGBoost forecasting with 5-fold nested cross-validation is shown in Figure 2. The implementation involved the utilization of the TimeSeriesSplit() function within the outer loop, where a 5-fold configuration was employed. This procedural choice inherently encompassed the automated management of index slicing. The method adopted herein adhered to a walk-forward paradigm, wherein the dimensions of the training dataset were progressively augmented with the progression of each successive fold. Within the inner loop, an allocation scheme of 80% and 20% was assigned for training and validation purposes, respectively. For illustrative purposes, in the initial fold, 20% of the data was reserved for the training set of the outer loop, while an equivalent proportion was designated for testing. Within this 20% training set of the outer loop, a further division ensued, partitioning it into an 80-20 ratio to accommodate hyperparameter tuning.

Within each fold of the outer loop, the model was trained with a predefined number of estimators, utilizing the Outer Train dataset. Then, each resultant model, in conjunction with its corresponding validation errors, was evaluated on the Outer Test dataset. This iterative process continued for each fold until the entirety of the folds had been completed. Afterwards, the scores obtained from each individual test were subjected to averaging. The hyperparameter denoted as "n-estimator," associated with the lowest computed average test score, was identified as the optimal value. This optimal value then served as the basis for the refinement of the subsequent set of hyperparameters.

Hyperparameter	Pangasinan		Negros Occidental		Davao Del Norte	
	M1	***M2***	***M1***	***M2***	***M1***	***M2***
objective	reg:squarederror	reg:squarederror	reg:squarederror	reg:squarederror	reg:squarederror	reg:squarederror
n_estimators	100	100	100	100	100	100
max_depth	2	1	1	1	1	1
min_child_weight	50	50	5	5	50	50
gamma	0	0	1	1	1	1
subsample	1	1	0.9	0.6	1	1
colsample_bytree	0.4	0.5	1	0.5	1	1
colsample_bylevel	1	0.9	0.9	1	1	1
learning_rate / eta	0.1	0.3	0.2	0.1	0.3	0.3

Table 2. Summary of XGBoost Model Hyperparameters.

The optimal configuration of the XGBoost forecasting models for solar PV output was determined through a systematic exploration of various hyperparameter combinations. This process was crucial because it enabled fine-tuning of the models, enhancing its predictive accuracy and robustness.

In Table 2, a thorough summary of the XGBoost model hyperparameters used in this study was presented, encompassing variations across different geographical regions and model iterations (M1 and M2). The chosen combinations cater to the unique characteristics of each location and model, balancing the need for accuracy, while preventing overfitting.

2.4 Performance Evaluation

Each model was assessed using Mean Absolute Percentage Error (MAPE) presented in Eq. (1) and computation of skill scores based on Yang (2019) presented in Eq. (2):

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{actual\ output_i - estimated\ output_i}{actual\ output_i}\right| \quad (1)$$

$$\text{skill score} = 1 - \frac{error_{proposed}}{error_{reference}} \quad (2)$$

where n is the total number of data points, $error_{proposed}$ is the MAPE value of M2 per fold for each location, and $error_{reference}$ is the MAPE of M1 per fold for each location. Skill scores ranging between 0.3 and 0.5 denote improvement, while higher skills such as 0.5-0.7 can be achieved but not over an extended validation period such as a year (Yang, 2019).

3. RESULTS AND DISCUSSIONS

This paper examines and compares the solar PV output forecasting performance of two models: one that used only solar irradiance (M1) and another which combined solar irradiance with weather parameters (M2). The application of Principal Components Analysis (PCA) to streamline the selection of weather variables obtained from ERA5 for inclusion in M2 is presented in Table 3. Through this analysis, the original weather parameters are condensed into five principal components, each accounting for a cumulative proportion of 80% or greater of the total variance in the data. These principal components, labeled as PC1 to PC5, play a crucial role in simplifying the complexity of the weather data.

The values within Table 3 represent the correlations or loadings, between the original weather variables and the five principal components across three locations: Pangasinan, Negros Occidental, and Davao del Norte. These loadings indicate both the strength and direction of the relationship between each weather variable and principal components. A high magnitude indicates that the variable has a substantial impact on the principal component, while values near zero suggest minimal influence. The sign (- or +) of a loading denotes whether there exists a positive or negative correlation between a variable and a principal component. For instance, in Pangasinan, the first principal component exhibits high loadings for high cloud cover (0.4), total cloud cover (0.422), and relative humidity (0.446), while wind speed has the least influence (-0.012). Also, PC1 demonstrates a negative correlation with wind speed, temperature at 2m (-0.312), and solar irradiation (-0.339). This dimensionality reduction simplifies the weather data analysis and provides valuable insights into meteorological dynamics across various geographic locations.

Moreover, in Table 4, the MAPE values recorded from the validation fold form the basis for assessing the improvements achieved by Model 2 (M2) over Model 1 (M1) in terms of predictive accuracy. These values reflect the actual forecasting errors for each fold and location, allowing for a direct comparison between the two models. Analyzing the MAPE results reveals that in Pangasinan, M2 (25.951%) outperforms M1 (29.850%), yielding lower MAPE values on average. However, in Negros Occidental, M2 (31.051%) exhibits slightly improved average MAPE values compared to M1 (31.683%), with relatively minor overall improvements. In contrast, Davao Del Norte experiences significant improvements when using M2, consistently yielding superior performance to M1 across all folds.

Table 5 introduces the skill scores computed based on the MAPE values in Table 4, employing Equation (2) to assess the extent of improvement, fold by fold, for each location. Pangasinan's case is marked by a mix of improvements, highlighting a varied response to the inclusion of weather variables. As indicated by the skill scores, the forecast accuracy exhibits improvements in specific instances, reflecting the potential advantages of incorporating weather-related insights for this location. On the other hand, Negros Occidental demonstrates low overall improvements, as indicated by skill scores close to zero. The skill scores suggest that, for this site, the inclusion of weather parameters may not lead to a substantial improvement in

predictive accuracy. Davao Del Norte, on the other hand, showcases gain in each fold. The varying impact becomes more apparent when the skill scores across different locations are compared. While Pangasinan witnesses a notable improvement at skill score of 0.356 (Fold 3) and Davao Del Norte at skill score of 0.398 (Fold 1) with the incorporation of weather data, this improvement is not observed for Negros Occidental. This distinction underlines that the solar irradiance factor alone sufficiently captures the underlying dynamics of solar PV output for this installation in Negros Occidental, a conclusion supported by relatively stable skill scores.

These skill scores underscore the location-specific impact of weather parameters on solar PV output forecasting, emphasizing the necessity for tailored modeling approaches that consider the unique weather dynamics of each region. The results highlight that although weather parameters can significantly enhance forecasting accuracy in some locations, its effect can vary significantly, reinforcing the importance of location-specific modeling in renewable energy planning

Variable	Pangasinan					Negros Occidental					Davao del Norte				
	PC1	***PC2***	***PC3***	***PC4***	***PC5***	***PC1***	***PC2***	***PC3***	***PC4***	***PC5***	***PC1***	***PC2***	***PC3***	***PC4***	***PC5***
ssrd	-0.339	0.51	0.008	-0.061	-0.033	-0.458	-0.263	0.124	-0.072	-0.076	-0.459	-0.209	-0.065	-0.062	0.197
ws	-0.012	0.082	-0.592	-0.048	0.737	-0.319	-0.157	-0.136	0.481	-0.032	-0.206	-0.085	-0.461	-0.5	-0.128
wd	0.149	-0.008	-0.346	-0.84	-0.192	0.11	-0.267	0.048	-0.782	-0.085	0.272	0.209	-0.094	-0.274	0.825
rh	0.446	-0.298	0.048	-0.067	-0.11	0.514	0.138	-0.127	0.019	-0.048	0.506	0.179	-0.018	0.01	-0.145
t2m	-0.312	0.507	0.076	-0.15	-0.167	-0.476	-0.234	0.199	-0.159	0.037	-0.488	-0.225	0.07	0.056	0.18
tp	0.243	0.257	-0.398	0.262	-0.238	0.021	-0.326	-0.522	-0.153	0.559	0.093	-0.278	-0.392	0.579	0.383
lcc	0.271	0.21	-0.298	0.098	-0.457	0.049	-0.334	-0.476	0.026	-0.762	0.14	-0.1	-0.641	-0.28	-0.135
mcc	0.309	0.244	-0.184	0.375	0.04	0.104	-0.378	-0.297	0.201	0.28	0.203	-0.381	-0.246	0.35	-0.194
hcc	0.4	0.318	0.368	-0.162	0.264	0.306	-0.418	0.441	0.163	0.074	0.222	-0.532	0.329	-0.235	0.086
tcc	0.422	0.342	0.324	-0.127	0.205	0.28	-0.475	0.356	0.188	-0.068	0.244	-0.554	0.192	-0.277	0.02

Table 3. Summary of Principal Components.

Location	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5		Average	
	M1	**M2**	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**
Pangasinan	21.857	20.605	32.723	30.081	51.318	33.027	16.049	17.959	27.572	28.081	29.850	25.951
Negros Occidental	24.682	24.121	31.638	31.291	30.135	29.801	38.477	37.097	33.484	32.946	31.683	31.051
Davao Del Norte	37.026	22.284	34.603	32.917	44.928	43.756	39.264	35.743	48.485	38.913	40.861	34.723

Table 4. Summary of Validation MAPE Results

Location	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Pangasinan	0.045	0.081	0.356	-0.119	-0.018
Negros Occidental	0.023	0.011	0.011	0.036	0.016
Davao Del Norte	0.398	0.049	0.026	0.090	0.197

Table 5. Summary of Skill Scores.

4. CONCLUSIONS AND RECOMMENDATIONS

The study examines solar PV output forecasting accuracy in the Philippines, focusing on the incorporation of Fifth Generation ECMWF Reanalysis (ERA5) data. The investigation involved a comparative evaluation of two models: Model 1 (solar irradiance) and Model 2 (solar irradiance and weather parameters), using the XGBoost toolkit. Results showed that, in the case of the solar PV installations considered in this study, relying solely on solar irradiance as a predictor was sufficient to predict solar PV output for installations in Negros Occidental. However, for Davao Del Norte and Pangasinan, incorporating weather parameters improved the accuracy of solar PV output forecasting. These results emphasize the need for a tailored approach to solar PV output forecasting, considering the unique weather patterns and conditions of each region in the Philippines.

A limitation of this study is that the forecasting performance of the models is also dependent on the quality of the data, the specific location, and the prevailing weather conditions. These factors can introduce uncertainties and variability that may affect the accuracy of solar PV output predictions. Therefore, it is imperative to note that the study's findings are specific to the dataset used and the locations considered, and should not generalize the results and apply the same to other regions of the country or periods with distinct weather patterns.

As the Philippines continues to explore and expand its renewable energy sector, harnessing the full potential of solar power necessitates the use of forecasting methodologies that account for the dynamic nature of weather patterns across different regions.

ACKNOWLEDGEMENT

This study was implemented under the OutSolar component of Project SINAG (Solar PV Resource and Installation Assessment Using Geospatial Technologies). Project SINAG is funded by the Philippine Department of Science and Technology (DOST).

REFERENCES

Alcañiz, A., Grzebyk, D., Ziar, H., Isabella, O., 2023. Trends and gaps in photovoltaic power forecasting with machine learning. Energy Reports 9, 447–471. https://doi.org/10.1016/j.egyr.2022.11.208

AlSkaif, T., Dev, S., Visser, L., Hossari, M., van Sark, W., 2020. A systematic analysis of meteorological variables for PV output power estimation. Renewable Energy 153, 12–22. https://doi.org/10.1016/j.renene.2020.01.150

Benitez, I., Gerna, L., Ibañez, J., Principe, J., De Los Reyes, F., 2022. Use of SARIMAX Model for Solar PV Power Output Forecasting in Baguio City, Philippines, in: 2022 International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE). Presented at the 2022 International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE), pp. 1–7. https://doi.org/10.1109/ICUE55325.2022.10113538

Bertheau, P., 2020. Assessing the impact of renewable energy on local development and the Sustainable Development Goals: Insights from a small Philippine island. Technological Forecasting and Social Change 153, 119919. https://doi.org/10.1016/j.techfore.2020.119919

Bro, R., Smilde, A.K., 2014. Principal component analysis. Anal. Methods 6, 2812–2831. https://doi.org/10.1039/C3AY41907J

Chahboun, S., Maaroufi, M., Chahboun, S., Maaroufi, M., 2022. Principal Component Analysis and Artificial Intelligence Approaches for Solar Photovoltaic Power Forecasting, in: Advances in Principal Component Analysis. IntechOpen. https://doi.org/10.5772/intechopen.102925

Das, U.K., Tey, K.S., Seyedmahmoudian, M., Mekhilef, S., Idris, M.Y.I., Van Deventer, W., Horan, B., Stojcevski, A., 2018. Forecasting of photovoltaic power generation and model optimization: A review. Renewable and Sustainable Energy Reviews 81, 912–928. https://doi.org/10.1016/j.rser.2017.08.017

Dimitropoulos, N., Sofias, N., Kapsalis, P., Mylona, Z., Marinakis, V., Primo, N., Doukas, H., 2021. Forecasting of short-term PV production in energy communities through Machine Learning and Deep Learning algorithms, in: 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA). Presented at the 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA), pp. 1–6. https://doi.org/10.1109/IISA52424.2021.9555544

Farias-Rocha, A.P., Hassan, K.M.K., Malimata, J.R.R., Sánchez-Cubedo, G.A., Rojas-Solórzano, L.R., 2019. Solar photovoltaic policy review and economic analysis for on-grid residential installations in the Philippines. Journal of Cleaner Production 223, 45–56. https://doi.org/10.1016/j.jclepro.2019.03.085

Grzebyk, D., Alcañiz, A., Donker, J.C.B., Zeman, M., Ziar, H., Isabella, O., 2023. Individual yield nowcasting for residential PV systems. Solar Energy 251, 325–336. https://doi.org/10.1016/j.solener.2023.01.036

Hannan, M.A., Al-Shetwi, A.Q., Ker, P.J., Begum, R.A., Mansor, M., Rahman, S.A., Dong, Z.Y., Tiong, S.K., Mahlia, T.M.I., Muttaqi, K.M., 2021. Impact of renewable energy utilization and artificial intelligence in achieving sustainable development goals. Energy Reports 7, 5359–5373. https://doi.org/10.1016/j.egyr.2021.08.172

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374, 20150202. https://doi.org/10.1098/rsta.2015.0202

Munawar, U., Wang, Z., 2020. A Framework of Using Machine Learning Approaches for Short-Term Solar Power Forecasting. J. Electr. Eng. Technol. 15, 561–569. https://doi.org/10.1007/s42835-020-00346-4

Yang, D., 2019. A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES). Journal of Renewable and Sustainable Energy 11, 022701. https://doi.org/10.1063/1.5087462

Zhong, Y.-J., Wu, Y.-K., 2020. Short-Term Solar Power Forecasts Considering Various Weather Variables, in: 2020 International Symposium on Computer, Consumer and Control (IS3C). Presented at the 2020 International Symposium on Computer, Consumer and Control (IS3C), pp. 432–435. https://doi.org/10.1109/IS3C50286.2020.00117