Journal of Renewable and Sustainable Energy

RESEARCH ARTICLE | AUGUST 24 2023

A novel data gaps filling method for solar PV output forecasting

Ian B. Benitez ; Jessa A. Ibañez ; Cenon D. Lumabad, III ; Jayson M. Cañete ; Francisco N. De los Reyes ; Jeark A. Principe

J. Renewable Sustainable Energy 15, 046102 (2023)
https://doi.org/10.1063/5.0157570

Journal of Renewable and Sustainable Energy ARTICLE pubs.aip.org/aip/rse

A novel data gaps filling method for solar PV output forecasting

Cite as: J. Renewable Sustainable Energy **15**, 046102 (2023); doi: 10.1063/5.0157570
Submitted: 9 May 2023 · Accepted: 6 August 2023 ·
Published Online: 24 August 2023

Ian B. Benitez,[1,a] Jessa A. Ibañez,[1,b] Cenon D. Lumabad III,[1,c] Jayson M. Cañete,[1,d] Francisco N. De los Reyes,[2,e] and Jeark A. Principe[3,f]

AFFILIATIONS

[1]National Engineering Center, University of the Philippines, Diliman Quezon City, Philippines
[2]School of Statistics, University of the Philippines, Diliman, Quezon City, Philippines
[3]Department of Geodetic Engineering, University of the Philippines, Diliman, Quezon City, Philippines

[a]Electronic mail: ibbenitez@up.edu.ph
[b]Electronic mail: jaibanez@up.edu.ph
[c]Electronic mail: cenonlumabadiii@gmail.com
[d]Electronic mail: jmcanete@up.edu.ph
[e]Electronic mail: fndelosreyes@up.edu.ph
[f]Author to whom correspondence should be addressed: japrincipe@up.edu.ph

ABSTRACT

This study proposes a modified gaps filling method, expanding the column mean imputation method and evaluated using randomly generated missing values comprising 5%, 10%, 15%, and 20% of the original data on power output. The XGBoost algorithm was implemented as a forecasting model using the original and processed datasets and two sources of solar radiation data, namely, Shortwave Radiation (SWR) from Advanced Himawari Imager 8 (AHI-8) and Surface Solar Radiation Downward (SSRD) from ERA5 global reanalysis data. The accuracy of the two sets of forecasted power output was evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Results show that by applying the proposed gap filling method and using SWR in forecasting solar photovoltaic (PV) output, the improvement in the RMSE and MAE values range from 12.52% to 24.30% and from 21.10% to 31.31%, respectively. Meanwhile, using SSRD, the improvement in the RMSE values range from 14.01% to 28.54% and MAE values from 22.39% to 35.53%. To further evaluate the accuracy of the proposed gap-filling method, the proposed method could be validated using different datasets and other forecasting methods. Future studies could also consider applying the said method to datasets with data gaps higher than 20%.

I. INTRODUCTION

Missing observations in a time series dataset are among the most common hurdles encountered in forecasting. Most of the time, collecting missing data are constrained by time and cost. To perform a complete time series analysis, blunders must be identified and corrected or removed whenever an attempt is made to fill the missing observations. In statistics, imputation is the process of replacing missing data with substituted values. Two general approaches for time series imputation include deterministic or numerical analysis modeling, and stochastic or time series modeling.[1,2]

Deterministic modeling uses numerical analysis techniques to model the time series data. It assumes that time series data follows a polynomial function and examines the time interval containing the missing values. Techniques in deterministic modeling include least squares approximations (polynomial curve fitting) and interpolation methods (linear, spline, and Stineman).[1] Meanwhile, stochastic modeling for time series uses random variables to forecast the probability of the outcome in the time series data. Included in this approach are stationary models, non-stationary models, and Box–Jenkin models.[3] Box–Jenkin model could be further classified into autoregressive models (AR models), moving average models (MA models), autoregressive moving average models (ARMA models), autoregressive integrated moving average (ARIMA) models, least-square principle, and interpolation. Iwueze *et al.* presented an exhaustive review on estimating missing values in time series.[2]

Imputation is one of the most common methods of filling the data gaps on a time series data. Imputation is the process of filling in data gaps on a time series using previously collected information.

There are several types of imputation methods that can be used such as linear interpolation (LI),[4–6] mean/mode imputation (MI),[7] regression imputation (RI),[8] row and column mean imputation (RMI & CMI),[9] optimally weighted average imputation,[10] multivariate imputation by chained equations (MICE),[11–13] and k-nearest neighbors (k-NN).[14] Meanwhile, Demirhan and Renwick compared the accuracy of different imputation methods in solar irradiance data.[4] Performance of various imputation methods was tested by creating an artificial gap based on actual solar irradiance data. Results of the study showed that the accuracy of imputation methods is highly dependent on the length of data gaps, location of data gaps over the series, and irregularity in the occurrence pattern of data gaps. Finally, Iwueze *et al.* proposed methods of estimating values in time series, including the column mean imputation (CMI) method, under the assumption that the time series is arranged such that the number of rows is greater than the number of columns. CMI obtains the missing value of a variable by computing the average of the remaining observations from the same column as shown in the following equation:

$$CMI = \frac{1}{m-1}\left[\sum_{u=1}^{i-1} X_{(u-1)s+j} + \sum_{u=i+1}^{m} X_{(u-1)s+j}\right], \quad (1)$$

where m refers to rows, s refers to columns, and X_t is the observation at time t, which is identified with row i and column j.

This study expanded the proposed gaps-filling method of Iwueze *et al.*,[2] where instead of simply scanning the dataset and averaging the variable column with missing values, our proposed method accounts for specified conditions discussed and disregards the values that do not meet them.

This study aims to evaluate the performance of the proposed data gaps filling method and identify which of the remotely sensed solar irradiance data yields better performance for the proposed gaps filling method and identifies the effects of the proposed method to solar PV output forecasting accuracy in terms of RMSE and MAE. Due to data availability, this study is limited to using the ERA5 weather parameters as additional variables for the proposed gaps filling method.

II. MATERIALS AND METHODS

A. Data

The solar PV power output data used in this study were obtained from one of the registered solar power plants in Tagum City, Davao of the Department of Energy (DOE). A total of 8784 data points for 2020 with hourly temporal resolution were obtained.

The Himawari-8 is a Japanese satellite designed to acquire data for weather analysis and forecasting. With a resolution of 500 m every 10 min, its primary instrument, the Advanced Himawari Imager, can take visible and infrared images of a large portion of Asia and the Pacific. The raw data are processed by the Japan Meteorological Agency (JMA) and Japan Aerospace Exploration Agency (JAXA) into a variety of products that are utilized for many different purposes, such as estimating the potential power of solar photovoltaic systems in the Asia-Pacific region.[15] This study utilized the L3 shortwave radiation (SWR) product of the Advanced Himawari Imager-8 (AHI-8).

The fifth generation of ECMWF's reanalysis of the world's climate and weather (ERA5) provides hourly estimates for a wide range of atmospheric, oceanic, and land-surface parameters. It applies the rules of physics to combine model data with global observations to produce a globally complete and consistent dataset using information going back to 1950 and replaces the ERA-Interim reanalysis.[16] The data also include monthly mean averages that have been precalculated for convenience. Uncertainty estimates were sampled by a 10-member ensemble at three-hourly intervals. For this study, ERA5 data on wind speed (ws), wind direction (wd), ambient temperature (t2m), relative humidity (rh), precipitation (tp), and surface solar radiation downward radiation (SSRD) were utilized. The choice of weather parameters was based on the following solar PV output studies.[17–19]

B. Data pre-processing

Outliers were checked using Grubb's test, the appropriate test to be used when it is unknown whether the datasets contain outliers.[20] Once outliers are removed, Pearson correlation is employed to all selected weather parameters with respect to solar PV output, eliminating the less significant variables.[21–24] The remaining variables were then used in the proposed data gaps-filling method. The choice of the 20% limit is based on the general practice that when the dataset has 20% data gaps, it must not be used for analysis. Furthermore, a 20% dataset gap can introduce uncertainty and jeopardize analyses or forecasts. This limit is applied to the solar PV output data only. The complete workflow for the data pre-processing is shown in Fig. 1(a) (data pre-processing).

C. Data gaps filling

Several studies had been conducted using different imputation methods for solar PV output data.[6,13,25–29] However, most of these studies did not consider different weather factors which affect solar PV output. As such, the proposed imputation method in this study aims to include the said parameters. The conceptual design for the proposed data gaps filling process is shown in Fig. 1(b) (data gaps filling).

Output data from the solar PV facility and its corresponding weather data were compared. If the solar PV output data are missing for a particular observation period, its corresponding weather data were used to estimate the missing value. The imputation approach begins by scanning for irradiance observations of the same hour from the entire dataset. Then, the next phase takes into account the scanned irradiance observations, identifying if they fall within the range of solar irradiance from the missing PV plant output data. If they fall outside the range, the solar PV output data are rejected. The same steps are taken for each weather parameter, where the process loop continues until the gap is filled. This process loop uses a limiting multiplier, which has an increment between −10% and 10% [Fig. 1(b), STEP 2a]. Once the solar PV output from all datasets passes the specified condition [Fig. 1(b), STEP 2b], they are averaged and utilized as a stand-in for the missing PV plant output data [Fig. 1(b), STEP 3]. The dataset with gaps filled using SWR was used for the forecasting model with SWR as solar irradiance, while the dataset with gaps filled using SSRD was used for the forecasting model with SSRD as solar irradiance.

D. Data analysis and evaluation

The general workflow for data analysis is shown in Fig. 1(c), where each forecasting model only utilizes one solar irradiance, either SWR or SSRD. The accuracy of the proposed data gaps filling method and forecasts are evaluated using different error metrics. The collected

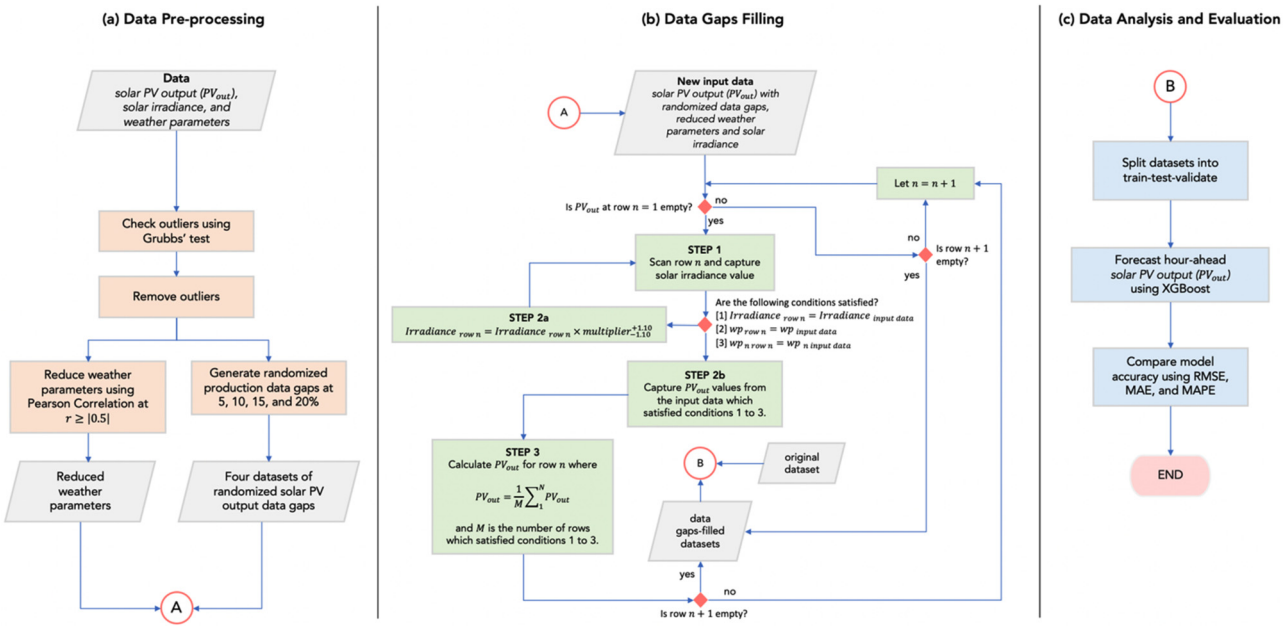

FIG. 1. Overall methodology in analyzing the improvement of solar PV output forecasting using the proposed data gap filling method.

datasets are divided into two parts: test and train. The total data points collected for the solar PV output was 8784, divided into 80% for training and 20% for testing. An additional 24 h was used for the validation part of the model. The forecast horizon is 24 h, resolution at 1 h, and lead time at 5 min.

XGBoost was utilized to forecast solar PV output. XGBoost is a distributed gradient boosting toolkit that builds on the basic Gradient Boosting framework with system and algorithmic improvements, making it extremely effective, adaptable, and portable.[30] XGBoost builds new trees in an iterative process to anticipate faults from earlier trees. Final forecasts are then made by combining these trees. Scores, referred to as feature importance, are assigned to features within the trees to indicate how crucial they are for making predictions. The value of a characteristic increases with usage, making it easier to comprehend how important it is for making predictions. The model configurations used are the default values of the hyperparameters for the XGBoost Models, and are shown in Table I.

TABLE I. Summary of the default XGBoost model configurations used in the study.

Hyperparameter	Value
lr	0.001
n_estimators	500
subsample	1
col sample byttee	1
col sample bylevel	1
min_child weight	1
max_depth	6
obj	reg squared error

To compare the accuracy of the proposed gap filling method and the XGBoost forecasts, this study adapted the top two error metrics used in evaluating the effectiveness of various forecasting models according to Ref. 31: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as shown in Eqs. (2) and (3), respectively.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}}, \tag{2}$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}, \tag{3}$$

where $\boldsymbol{n}$ is the number of forecasts while $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ are the actual and projected values, respectively. They are computed using only the daytime hours of the time series, which spans from 6:00 AM to 5:00 PM, based on the operating hours of the power plant. The resulting RMSE and MAE values were then compared in terms of percent change (%), as shown in Eqs. (4) and (5) to determine whether the dataset with filled data gaps performed better than the dataset with unfilled data gaps in forecasting solar PV output. It should be noted that a positive value of ΔRMSE and ΔMAE indicates an improvement in forecasting model accuracy. For the unfilled data gaps, the whole row with missing (unfilled) solar PV output was removed since the forecasting algorithm of the XGBoost requires a continuous time series input,

$$\Delta RMSE\,(\%) = \frac{RMSE_u - RMSE_f}{RMSE_u} \times 100, \tag{4}$$

$$\Delta MAE\,(\%) = \frac{MAE_u - MAE_f}{MAE_u} \times 100, \tag{5}$$

where u and f represent datasets with unfilled and filled data gaps, respectively.

TABLE II. Summary of datasets used in the study.

Source	Data	Unit	Temporal resolution
DOE	Solar PV Output	kW	Hourly
Himawari-8	SWR	$W \cdot m^{-2}$	Hourly
ERA5	SSRD	$W \cdot m^{-2}$	Hourly
	ws	$m \cdot s^{-2}$	Hourly
	wd	cardinal direction	Hourly
	t2m	°C	Hourly
	rh	%	Hourly
	tp	mm	Hourly

III. RESULTS AND DISCUSSION

A. Data

For the data evaluation, this study utilized several datasets from various sources as presented in detail in Table II. From Fig. 2, the solar PV output of Tagum follows a similar pattern with SWR, while its values mostly run below the solar irradiance values.

This indicates that the solar PV output and solar irradiance are correlated and as supported by the Pearson correlation results presented on Subsection III B–III D.

B. Data pre-processing

One of the key components of forecasting is data pre-processing, which ensures that the input data are as clean as possible and devoid of outliers. For this study, no outliers in the dataset were found using Grubbs' test. A Pearson correlation (r) was also used to identify the significant parameters for the proposed data gaps filling method and solar PV output forecasting. Parameters with r values above $|0.5|$ are considered in this study. Based on the results, SWR (0.831), SSRD (0.787), and ambient temperature (0.73) have a positive correlation to solar PV output, while relative humidity (−0.702) has a negative correlation. A high correlation between solar PV output and solar irradiance has also been reported by Gupta and Singh.[19]

Test cases were also established and gaps on the original dataset were generated randomly. Table III shows the summary of test cases

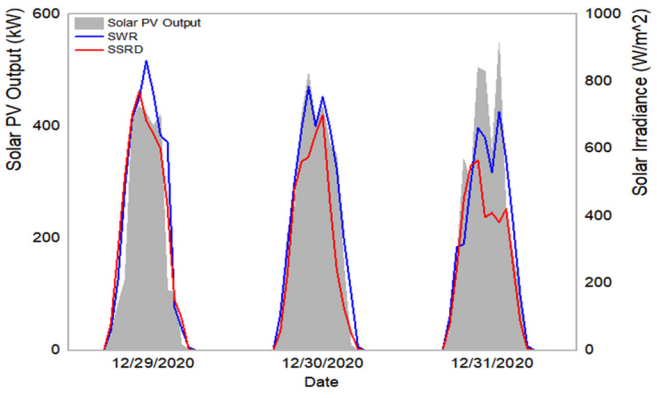

FIG. 2. Comparison between two remotely sensed solar irradiance data (SWR and SSRD) juxtaposed with the solar PV output of Tagum for December 29–31, 2020.

TABLE III. Summary of test cases and the number of missing data points for each randomly generated data gap used in the study.

Test case	Data gap (%)	Number of Datapoints
1	5	439
2	10	878
3	15	1318
4	20	1757

and their estimated percent (%) number of data gaps. These gaps were generated to test the accuracy of the proposed gap filling method and its impact on forecasting solar PV output. Meanwhile, SWR and SSRD were used on each test case to compare, which dataset produces better forecast and gap filling results.

C. Data gaps filling

Implementing the proposed data gaps filling method utilized two solar irradiance data, SWR and SSRD, to compare which performs better in filling the missing values. Table IV summarizes the proposed data gaps filling method results based on 5%, 10%, 15%, and 20% (per cent) increments. Results showed that SSRD perform better in terms of filling the gaps compared to SWR. The number of data gaps significantly decreased using the SSRD datasets, the gaps were reduced from 5%, 10%, 15% and 20% to 1.37%, 2.39%, 1.82%, and 1.99%, for test cases 1–4, respectively. Meanwhile, using the SWR dataset, the data gaps were reduced in all cases to 1.82%, 3.30%, 4.32%, and 5.18% for cases 1 to 4, respectively. In this study, the unfilled datagaps after running the proposed filling method were excluded in the implementation of XGBoost.

D. Data analysis and evaluation

Using XGBoost to forecast solar PV output, the RMSE and MAE values were compared for SWR and SSRD models across all test cases,

TABLE IV. Summary of data gaps filled for each test case at 5%, 10%, 15%, and 20% increment using SWR and SSRD data for solar irradiance.

	SWR				SSRD			
% Increment	1	2	3	4	1	2	3	4
1	110	201	275	367	101	200	255	345
2	124	262	355	445	158	330	451	549
3	80	147	258	324	83	162	261	369
4	40	94	145	194	40	73	167	197
5	29	56	98	124	17	39	69	118
6	14	31	57	72	13	17	39	61
7	16	27	27	53	9	14	25	36
8	10	18	27	40	7	13	13	22
9	6	8	9	31	2	5	4	12
10	2	5	10	16	3	4	10	13
Filled data	431	849	1261	1666	433	857	1294	1722
Unfilled data	8	19	57	91	6	21	24	35

as summarized in Table IV. Based on the SWR models for test cases 1–4, RMSE values improved from 62.97, 71.68, 84.51, and 84.16 (unfilled) to 55.08, 59.98, 63.97, and 66.58 (filled), respectively. The MAE values in the SWR models also follow the same trend, from 36.80, 43.28, 51.97, and 51.6 (unfilled), to 29.04, 32.23, 35.70, and 37.53 (filled), respectively. Meanwhile, for SSRD models, RMSE values improved in test cases 1–4, from 62.73, 71.84, 85.60, and 83.73 (unfilled), to 53.94, 58.60, 61.17, and 60.85 (filled), respectively. The same improvement trend was also observed for the MAE values in the SSRD models, from 36.80, 43.28, 51.97, and 51.6 (unfilled), to 28.13, 31.88, 33.89, and 33.55 (filled), respectively. Equation (4) was used to compare the solar PV output forecasting performance of SWR and SSRD models across all test cases using the gaps-filling method proposed in this paper. Results show that datasets with filled data gaps have better forecasting performance than datasets with unfilled data gaps. Improvement in forecasting accuracies also follows an increasing trend from test cases 1 to 4 for both RMSE and MAE, as presented in Table V. Furthermore, Table VI shows the improvement in forecast accuracies (%Change) for the SWR models are as follows: ΔRMSE = 12.52% (min) to 24.30% (max); ΔMAE = 21.10% (min) to 31.31% (max). Meanwhile, for the SSRD models, the improvement ranges from 14.01% to 28.54% for ΔRMSE, and 22.39% to 35.53% for ΔMAE.

For the unfilled dataset, the forecasting accuracy using SWR for test case 3 (data gaps at 15%) yielded higher errors than test case 4 (data gaps at 20%) (see Table VI). This contrasts the filled dataset, where test case 4 produced less accurate results than test case 3. When using SSRD in the forecasting models, test case 3 for filled and unfilled datasets produced higher errors than test case 4. The irregularity of the trend may be attributed to the number of unfilled data in the dataset for each case, as presented in Table IV.

IV. CONCLUSIONS AND RECOMMENDATIONS

One of the common problems in dealing with time series data are missing observations. In this study, a data gaps filling method was proposed to address this problem. The study carried out test cases at 5%, 10%, 15%, and 20% data gaps. Using the XGBoost algorithm, forecasting accuracy was computed by comparing the RMSE and MAE values for each test case. Results show that filling the gaps using SSRD outperformed SWR in forecasting solar PV output. At the same time, improvement of the XGBoost model accuracies follow the same trend across all test cases, where datasets with filled data gaps have better forecasting performance (max ΔRMSE = 28.54%, max ΔMAE = 35.53%) than datasets with unfilled data gaps. Moreover, an increasing trend in model improvement has been observed as the number of gaps filled increases (i.e., the fewer the number of remaining gaps in the dataset after applying the filling method, the higher the accuracy of the model). Finally, the proposed data gaps filling method has significantly improved the accuracy of the forecasting models with less than 20% data gaps.

To further evaluate the accuracy of the proposed method in this work, the proposed method is recommended to be validated using various solar PV output data from different locations, exogenous variables, time series lengths, and forecasting methods. Future studies could also consider applying this method to datasets with data gaps higher than 20%.

TABLE V. Summary of solar PV output forecasting accuracies in terms of RMSE and MAE values for each test case with filled and unfilled data gaps.

		Filled		Unfilled	
Irradiance	Test case	RMSE (kW)	MAE (kW)	RMSE (kW)	MAE (kW)
SWR	1	55.08	29.04	62.97	36.80
	2	59.98	32.23	71.68	43.28
	3	63.97	35.70	84.51	51.97
	4	66.58	37.53	84.16	51.63
SSRD	1	53.94	28.13	62.73	36.25
	2	58.60	31.88	71.84	43.12
	3	61.17	33.89	85.60	52.57
	4	60.85	33.55	83.73	51.29

TABLE VI. Summary of percent (%) change of the RMSE and MAE values for each test case used in the study. Positive values indicate improvement in the model accuracy.

		% Change	
Irradiance	Test case	ΔRMSE	ΔMAE
SWR	1	12.52	21.10
	2	16.33	25.52
	3	24.30	31.31
	4	20.89	27.31
SSRD	1	14.01	22.39
	2	18.42	26.07
	3	28.54	35.53
	4	27.33	34.58

ACKNOWLEDGMENTS

This study was implemented under the OutSolar component of Project SINAG (Solar PV Resource and Installation Assessment Using Geospatial Technologies). The authors would also like to acknowledge the Philippine Department of Science and Technology (DOST) as the funding agency of Project SINAG and the Department of Energy (DOE) for providing the Solar PV output data used in this study. Finally, the authors would like to acknowledge Mr. Lheander G. Gerna for his contribution on the initial conceptualization of the proposed method.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Ian Bartican Benitez: Conceptualization (lead); Data curation (lead); Formal analysis (equal); Methodology (lead); Resources (lead); Software (supporting); Writing – original draft (equal). **Jessa Alesna**

Ibanez: Conceptualization (supporting); Data curation (supporting); Formal analysis (equal); Methodology (supporting); Visualization (lead); Writing – original draft (equal). **Cenon De Vera Lumabad III:** Data curation (supporting); Software (supporting). **Jayson Masalihit Canete:** Data curation (supporting); Resources (supporting); Software (lead). **Francisco N. de los Reyes:** Validation (equal); Writing – review & editing (supporting). **Jeark Armingol Principe:** Funding acquisition (lead); Validation (equal); Writing – review & editing (lead).

DATA AVAILABILITY

The data that support the findings of this study are available from Saint Louis University. Access to the code for the data gaps filling methodology is available and can be requested through the corresponding author.

REFERENCES

[1]D. Fung, "Methods for the estimation of missing values in time series," M.S. thesis (Edith Cowan University, 2006).
[2]I. S. Iwueze, E. C. Nwogu, V. U. Nlebedim, U. I. Nwosu, and U. E. Chinyem, "Comparison of methods of estimating missing values in time series," Open J. Stat. **8**(2), 390–399 (2018).
[3]G. Box, "Box and Jenkins: Time series analysis, forecasting and control," in *A Very British Affair: Six Britons and the Development of Time Series Analysis during the 20th Century*, edited by T. C. Mills (Palgrave Macmillan, London, UK, 2013), pp. 161–215.
[4]H. Demirhan and Z. Renwick, "Missing value imputation for short to mid-term horizontal solar irradiance data," Appl. Energy **225**, 998–1012 (2018).
[5]I. Laña, I. (Iñaki) Olabarrieta, M. Vélez, and J. Del Ser, "On the imputation of missing data for road traffic forecasting: New insights and novel techniques," Transp. Res., Part C **90**, 18–33 (2018).
[6]T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting," Appl. Sci. **9**(1), 204 (2019).
[7]D. Sovilj *et al.*, "Extreme learning machine for missing data using multiple imputations," Neurocomputing **174**, 220–231 (2016).
[8]C. A. W. Glas, "Missing data," in *International Encyclopedia of Education*, 3rd ed., edited by P. Peterson, E. Baker, and B. McGaw (Elsevier, Oxford, 2010), pp. 283–288.
[9]M. Jamshidian and M. Mata, "2-advances in analysis of mean and covariance structure when data are incomplete this research was supported in part by the National Science Foundation Grant DMS-0437258," in *Handbook of Latent Variable and Related Models*, edited by S.-Y. Lee (North-Holland, Amsterdam, 2007), pp. 21–44.
[10]J. Peppanen, X. Zhang, S. Grijalva, and M. J. Reno, "Handling bad or missing smart meter data through advanced data imputation," in IEEE Power and Energy Society Innovative Smart Grid Technologies Conference (ISGT), 2016.
[11]M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?," Int. J. Methods Psychiatr. Res. **20**(1), 40–49 (2011).
[12]S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," J. Stat. Software **45**, 1–67 (2011).
[13]Q.-T. Phan, Y.-K. Wu, Q.-D. Phan, and H.-Y. Lo, "A study on missing data imputation methods for improving hourly solar dataset," in *8th International Conference on Applied System Innovation (ICASI)* (IEEE, 2022), pp. 21–24.
[14]G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," Appl. Artif. Intell. **17**(5–6), 519–533 (2003).
[15]R. Frouin and H. Murakami, "Estimating photosynthetically available radiation at the ocean surface from ADEOS-II global imager data," J. Oceanogr. **63**(3), 493–503 (2007).
[16]H. Hersbach *et al.*, "The ERA5 global reanalysis," Q. J. R. Meteorol. Soc. **146**(730), 1999–2049 (2020).
[17]T. Alquthami, M. A. Magzoub, and A. M. Osman, "A day ahead prediction of solar PV power output using ensemble neural network," in *22nd International Middle East Power Systems Conference (MEPCON)*. (IEEE, 2021), pp. 20–25.
[18]B. Kim, D. Suh, M.-O. Otto, and J.-S. Huh, "A novel hybrid spatio-temporal forecasting of multisite solar photovoltaic generation," Remote Sens. **13**(13), 2605 (2021).
[19]A. K. Gupta and R. K. Singh, "Short-term day-ahead photovoltaic output forecasting using PCA-SFLA-GRNN algorithm," Front. Energy Res. **10**, 1029449 (2022).
[20]F. E. Grubbs, "Sample criteria for testing outlying observations," Ann. Math. Stat **21**(1), 27–58 (1950).
[21]I. Jebli, F.-Z. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by pearson correlation using machine learning," Energy **224**, 120109 (2021).
[22]T. M. Al-Jaafreh and A. Al-Odienat, "The solar energy forecasting by Pearson correlation using deep learning techniques," EARTH Sci. Hum. Constr. **2**, 158–163 (2022).
[23]J. Guo, X. Xu, W. Lian, and H. Zhu, "A new approach for interval forecasting of photovoltaic power based on generalized weather classification," Int. Trans. Electr. Energy Syst. **29**(4), e2802 (2019).
[24]J. Zhong, L. Liu, Q. Sun, and X. Wang, "Prediction of photovoltaic power generation based on general regression and back propagation neural network," Energy Procedia **152**, 1224–1229 (2018).
[25]M. Shen, H. Zhang, Y. Cao, F. Yang, and Y. Wen, "Missing data imputation for solar yield prediction using temporal multi-modal variational auto-encoder," in *Proceedings of the 29th ACM International Conference on Multimedia*, New York, NY (ACM, 2021), pp. 2558–2566.
[26]I. de-Paz-Centeno, M. T. García-Ordás, Ó. García-Olalla, and H. Alaiz-Moretón, "Imputation of missing measurements in PV production data within constrained environments," Expert Syst. Appl. **217**, 119510 (2023).
[27]M. T. E. Seddik, O. Kadri, and M. R. Abdessemed, "Imputation as service using support vector regression: Application to a photovoltaic system in Algeria," in *1st National Conference of Materials Sciences and Engineering (MSE'22)* (HAL Open Science, 2022), available at https://hal.science/hal-03815846.
[28]T. Shireen, C. Shao, H. Wang, J. Li, X. Zhang, and M. Li, "Iterative multi-task learning for time-series modeling of solar panel PV outputs," Appl. Energy **212**, 654–662 (2018).
[29]I. P. Panapakidis, A. S. Bouhouras, and G. C. Christoforidis, "A missing data treatment method for photovoltaic installations," in IEEE International Energy Conference (ENERGYCON) (2018).
[30]T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY (ACM, 2016), pp. 785–794.
[31]A. Alcañiz, D. Grzebyk, H. Ziar, and O. Isabella, "Trends and gaps in photovoltaic power forecasting with machine learning," Energy Rep. **9**, 447–471 (2023).