SAND2020-9272C

GUIDELINES FOR ENSURING DATA QUALITY FOR PHOTOVOLTAIC SYSTEM PERFORMANCE ASSESSMENT AND MONITORING

Andreas Livera[1], Marios Theristis[2], Elena Koumpli[3], George Makrides[1], Joshua S. Stein[2] and George E. Georghiou[1]
[1]PV Technology Laboratory, FOSS Research Centre for Sustainable Energy,
Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, 1678, Cyprus
[2]Sandia National Laboratories, Albuquerque, New Mexico, 87185, USA
[3]SolarCentury, 90 Union Street, London SE1 0NW, UK

ABSTRACT: High-quality datasets are crucial for the performance and reliability analysis of photovoltaic (PV) systems. With respect to data integrity, invalid data are a common problem exhibited in PV monitoring systems. A data pipeline approach was recently introduced aiming to support reproducible results in PV performance. The methodology is expanded in this study by examining further outlier observations in respect to detection techniques, impact and treatment methods. The outlier detection results demonstrated that the standard boxplot rule yielded the highest detection rate of 95.3% (by taking a moving data window) at 40% of outlying data points and the effect of random outlying data points was mitigated by listwise deletion. The comparative analysis of outlying data treatment demonstrated that back-filling with the Sandia Array Performance Model (SAPM) yielded more accurate degradation rate (R_D) estimates (absolute percentage error, APE, of up to 0.36% at 40% of outlying data) compared to filtering out the outlying data points (APE of up to 2.53% with listwise deletion).

Keywords: data quality, outliers, system performance, reliability, photovoltaics.

1 INTRODUCTION

Ensuring data quality is of utmost importance for the performance and reliability analysis of photovoltaic (PV) systems [1]. Actual in-field measurements commonly exhibit invalid data (i.e. gaps, missing data, erroneous and outlying values) caused by power outages, equipment/component faults, communication failures or interruption for maintenance reasons that can significantly bias the results of the data-based analysis. For this reason, invalid data should be detected and addressed appropriately before proceeding to any analyses.

Even though it is very common to have missing and outlying (erroneous) observations in a given PV dataset, only few reference guidelines and reports on data quality checks for PV monitoring studies have been reported in the literature. The existing guidelines and reports mainly focus on data processing requirements for PV performance assessment [2]–[9]. The reported studies, however, do not provide a specific data treatment approach that would support reproducible results.

A complete methodology describing how to handle (and ensure quality of) large high-resolution datasets acquired from PV systems was recently presented in our previous work [10]. The detailed methodology comprises of a framework of sequentially structured Data Quality Routines (DQRs) that operate on acquired PV system and meteorological measurements. The DQRs methodology comprises of algorithms that detect data anomalies and reconstruct invalid PV datasets through a sequence of data quality checks, filtering stages, data deletion and inference techniques. This study expands on this methodology by examining outliers in respect to detection, impact and treatment using PV performance and reliability metrics as criteria.

2 METHODOLOGY

The data quality assurance methodology (depicted in Fig. 1) comprises of a data pipeline procedure that includes the application of initial data statistics (Step 1), consistency examination (Step 2), filtering (Step 3), invalid data detection (Step 4), determination of missing data mechanism and rate (Step 5), invalid data treatment (Step 6), aggregation at different granularities (Step 7), final data validity and statistics summary (Step 8) [10].

The initial step includes the preliminary application of data statistics to the PV dataset in order to gain insights of the dimensionality by identifying the recording interval (time between two consecutive time records) and the reporting period (i.e. the minimum of 1-year of continuous monitoring for outdoor PV performance evaluation) [11].

The fidelity of the dataset was then examined by verifying the consistency of the series (timestamp gaps, repetitive rows, duplicate timestamp records and synchronization issues) and detecting data inconsistencies. After removing the repetitive and duplicate timestamp records, the dataset was checked against a known timestamp series and was finally synchronised and resampled (reconstructed).

A filter was then applied to the dataset in order to restrict the measurements to daylight hours and remove night-time effects (i.e. irradiance filter > 20 W/m^2).

Subsequently, the missing values were identified by searching for Not a Number (NaN) and Not Available (NA) values into the dataset. Outliers (or erroneous values) were detected by a) manual approaches; imposing physical limitations on the recorded data [4], [9], visually inspecting scatter plots [12], [13] and applying variation limits between successive data points methods [4] and b) automated approaches [14]. Automated approaches include the use of statistical and comparative tests (e.g. Sigma rule method and standard boxplot rule, etc.), density-, deviation- and distance-based approaches (e.g. the local outlier factor, rolling mean, etc.) [14], [15]. Since there is no standardised method for detecting outliers for PV assessment analyses, a performance comparison between different common automated methods used in PV field applications was conducted to identify the optimum identification technique. The detected outlying values were then replaced by "NA" values and treated in the same manner as missing data.

At this point, the missing data rate (portion of

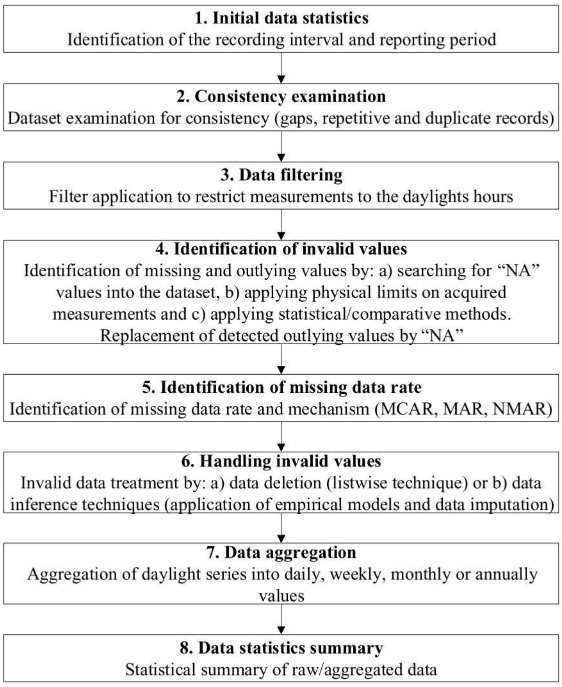
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

missing values to the total number of data points) was estimated. An essential next step performed was the identification of the missing data mechanism/pattern (MCAR - Missing Completely At Random, MAR - Missing At Random and NMAR - Not Missing At Random) by applying a suitable data visualization method (e.g. heatmaps, aggregation, scatter and spine plots) [16]. Identifying the type of the exhibited missing data pattern is important as it determines which treatment method is appropriate.

In Step 6, the appropriate treatment method (data deletion or inference) was determined based on the missing data rate and mechanism [10]. This is a challenging task because the application of data deletion and inference techniques is strongly dependent on the missing data rate and pattern and requires careful examination of the dataset in order to avoid introducing bias to the performed analysis.

In particular, missing values were treated by data deletion (listwise deletion) for missing data rates lower than 10% and by data inference techniques (application of empirical models, multiple and univariate data imputation) for missing data rates higher than 10% (applicable only for the MCAR case) [10]. For the MAR and NMAR cases, the reason of missingness was further examined before determining an approach to handle the invalid values.

The final step of the methodology was to aggregate the acquired daylight measurements into daily, weekly, monthly, or annual values (depending on the final analytical use of the acquired PV system data) and to provide the final statistical summary.



1. Initial data statistics
Identification of the recording interval and reporting period

2. Consistency examination
Dataset examination for consistency (gaps, repetitive and duplicate records)

3. Data filtering
Filter application to restrict measurements to the daylights hours

4. Identification of invalid values
Identification of missing and outlying values by: a) searching for "NA" values into the dataset, b) applying physical limits on acquired measurements and c) applying statistical/comparative methods. Replacement of detected outlying values by "NA"

5. Identification of missing data rate
Identification of missing data rate and mechanism (MCAR, MAR, NMAR)

6. Handling invalid values
Invalid data treatment by: a) data deletion (listwise technique) or b) data inference techniques (application of empirical models and data imputation)

7. Data aggregation
Aggregation of daylight series into daily, weekly, monthly or annually values

8. Data statistics summary
Statistical summary of raw/aggregated data

Figure 1: Flowchart of data quality processing methodology. Figure obtained from Livera *et al.* [10].

To facilitate the verification process of the outlier's identification routines, a baseline (reference) PV dataset was constructed by utilizing the acquired measurements from a test PV system installed in Nicosia, Cyprus (Köppen-Geiger-Photovoltaic climate classification CH: Steppe climate with high irradiation) [17]. In order to enable the comparative analysis, artificially "invalid" PV datasets were also generated by introducing outlying measurements at different data rates. Outlier detection routines (ODRs) were then applied to the invalid PV datasets in order to detect data anomalies and derive the optimum automated identification technique.

2.1 Experimental setup

At the outdoor test facility (OTF) of the University of Cyprus (UCY), grid-connected PV systems of different technologies and approximately of 1 kW_p capacity each, were installed and commissioned in June 2006 (see Fig. 2). The performance of each PV system and the prevailing irradiance and meteorological conditions are recorded according to the requirements set by the IEC 61724 standard [2] and stored with the use of a measurement monitoring platform. The monitoring platform comprises of solar irradiance, wind, temperature and electrical operation sensors and stores data at every second. The recorded meteorological measurements include the in-plane irradiance (G_I) measured with a pyranometer, ambient temperature (T_{amb}), module back-surface temperature (T_{mod}), relative humidity (*RH*), wind speed (W_s) and direction (W_a). The electrical data include the array current (I_A), voltage (V_A) and power (P_A) at the DC side. Additional yields and performance parameters such as the PV array energy yield (Y_A), the final PV system yield (Y_f), the reference yield (Y_r) and the monthly DC PR (*PR*) were also calculated [18].

In order to derive the optimum outlier identification technique, historical field measurements acquired from a test PV system installed at the UCY OTF were utilised. The test PV system is well maintained (at an availability of higher 99% during the reporting period) and comprises of a PV array of 5 poly-crystalline Silicon (poly-c Si) PV modules, each of nominal power 205 W_p, as depicted by the manufacturer's datasheet. The PV modules are connected in series to form one string, at the input of a grid-connected inverter. The system is installed in an open-field mounting arrangement and an inclination angle of 27.5° due South.

Figure 2: OTF of the UCY in Nicosia, Cyprus.

2.2 Baseline PV datasets

In order to assess the performance of the ODRs, a 1- and 5-year datasets (defined as the baseline PV datasets) containing 15-minute average measurements (and calculated performance parameters) acquired from the test PV system at a resolution of 1 second were used. The PV datasets were initially examined for consistency and then filtered to restrict measurements to daylight hours of each day.

2.3 Generation of outlying data points in the baseline PV dataset – Artificial invalid PV datasets

An automated approach was employed, in order to create datasets with artificial invalid data points (defined as the invalid PV datasets). These datasets were used to examine the impact of outlying data points on the PV performance and reliability analyses and the robustness of the dataset reconstruction routines.

In particular, the method introduces artificially global outliers (randomly distributed) into the PV baseline dataset from 1% to 40% invalid data rate (in whole number increments) for the recorded P_A measurements. The process was repeated 50 times for each invalid data rate, resulting in 2000 invalid datasets with outliers. The investigation focused on global outliers since this type of outliers is commonly exhibited in PV monitoring systems [19].

2.4 Detection of outliers

In order to derive the optimum identification technique, a comparative analysis was performed between three automated methods that are commonly used in PV field applications for outlier's detection; the 3-Sigma rule method, Hampel identifier and standard boxplot rule.

The 3-Sigma rule classifies any data point above and below the ±3 standard deviation (±3σ) from the mean (μ) as an outlier. The normal range defined by the 3-Sigma rule is the closed interval [$\mu - 3\sigma$, $\mu + 3\sigma$] [20]. Similarly, the Hampel identifier uses the sample median ($\tilde{\boldsymbol{x}}$) and the median absolute deviation (MAD) to define the data range for normal points [14]. Data points that fall outside the closed interval [$\tilde{\boldsymbol{x}}$ – 3*MAD, $\tilde{\boldsymbol{x}}$ + 3*MAD] are classified as outliers. Finally, the standard boxplot rule is based on the lower quartile (Q1, 25th percentile), the upper quartile (Q3, 75th percentile) and the inter-quartile distance (IQR= Q3-Q1) and its nominal data range is the closed interval [Q1 – 1.5 * IQR, Q3 + 1.5 * IQR] [14].

The outlier detection algorithms were applied on the whole length of the tested variable and were also applied on a moving data window by taking 1% increments of the length of the variable.

2.5 Treatment of invalid data points

Existing data quality assurance guidelines analyse either the available valid measurements, excluding the invalid periods (by listwise deletion) or replace the missing data with estimated values (using data inference techniques) [21]. Livera *et al.* [10] demonstrated that for missing data rates lower than 10%, periods with continuous missing measurements can be discarded from the analysis (listwise deletion) [10]. On the other hand, data inference techniques should be employed for missing data rates higher than 10% [10].

2.6 PV performance and reliability metrics

The analysis of the PV performance was based on the monthly *PR* time series constructed from the outdoor field measurements [22]. The reliability of the PV modules was evaluated based on degradation rate (R_D) [23] assuming linearity [24], [25]. The R_D was estimated by applying the conventional statistical method of linear regression with ordinary least squares (OLS) on the 5-year baseline dataset of the test PV system [24].

In order to compare the R_D obtained using the baseline dataset against the R_D values obtained from the artificially invalid datasets, the absolute percentage error (APE) metric was used [10].

In parallel, to assess the accuracy of the outlier identification techniques, the detection rate performance metric was used. The detection rate (units of %) is defined as the ratio between the detected invalid data points to the total number of invalid data points.

3 RESULTS

3.1 Outliers detection

A comparison of three different outlier identification methods was performed and the average detection accuracy results for different rates of outliers (inserted randomly to the tested PV dataset) are reported in Table I. The obtained results showed that the standard boxplot rule was the most successful global outlier identification technique (among the investigated algorithms) achieving an average detection rate of 95.3% at 40% of outlying data points. In addition, the Hampel identifier achieved a detection rate of 94.8% at 40% of outlying data points.

Table I: Detection rate of ODRs during different invalid data rates.

Invalid data	**Detection rate (%)**		
Rate (%)	3-Sigma rule	Hampel identifier	Boxplot
10	90.3	96.0	97.3
20	89.2	95.4	96.8
30	87.9	95.0	96.7
40	86.7	94.8	95.3

For demonstration purposes, a one-week period is depicted in Fig. 3 when applying the ODRs on the whole length of the tested variable, while Fig. 4 depicts the boxplot rule application on a moving data window.

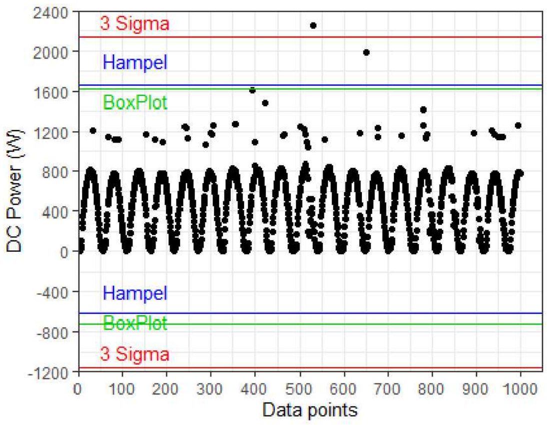


Figure 3: ODRs application on the recorded DC power measurements. The limits were calculated by applying the ODRs on the whole length of the tested variable. The upper and lower these limits for the 3-Sigma rule, Hampel identifier and boxplot rule are depicted by red, blue and green lines, respectively.

By comparing Fig. 3 with Fig. 4, it can be concluded that the standard boxplot rule proved to be more effective in detecting outliers by applying the technique on a moving data window (i.e. by taking increments of the length of the variable).

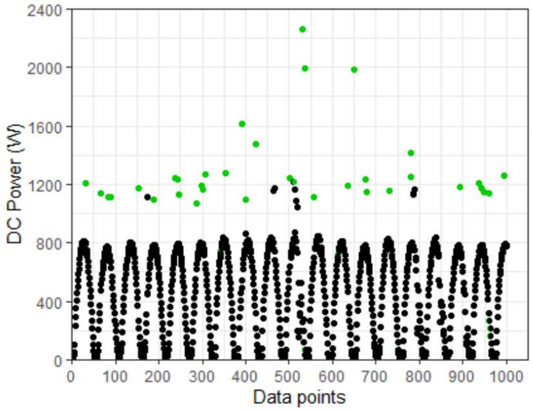


Figure 4: Boxplot rule application on the recorded DC power measurements. The boxplot rule was applied on 1% increments of the length of the tested variable and the detected outliers are colored in green.

3.2 Impact of outliers on PV performance analysis and outlier's treatment

The detected outliers were replaced by "NA" values and treated as random missing data points (i.e. the data points were distributed randomly in the time series). The 2000 invalid PV datasets were reconstructed by listwise deletion and the average *PR* results are depicted in Fig. 5. As shown in Fig. 5, the effect of random missing data points was mitigated by listwise deletion, even for a 40% invalid data rate (exhibiting an *APE* less than 0.7%).

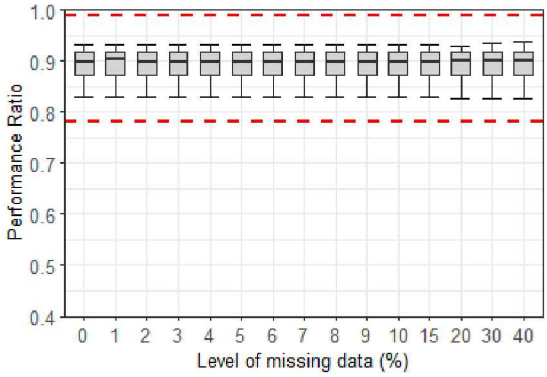


Figure 5: Boxplot of the monthly average *PR* of the test PV system for random missing power measurements reconstructed by listwise deletion. The horizontal lines (coloured in red) indicate the ±6% uncertainty on the calculated *PR* [26].

3.3 Impact of outliers on the estimation of the linear R_D of PV systems

The performed reliability investigation demonstrated that the R_D estimates (calculated by applying the linear OLS method to model the trend) were slightly biased in the presence of random missing data points. In particular, for 40% of random missingness, the maximum *APE* of the linear R_D estimated by applying the listwise deletion was 2.53% (see Fig. 6).

Data inference using the Sandia Array Performance Model (SAPM) to back-fill the random missing data points yielded more accurate R_D estimates compared to the listwise deletion, since for 40% random missing data, the *APE* of the annual R_D, was lower than 0.36% (see Fig. 6).

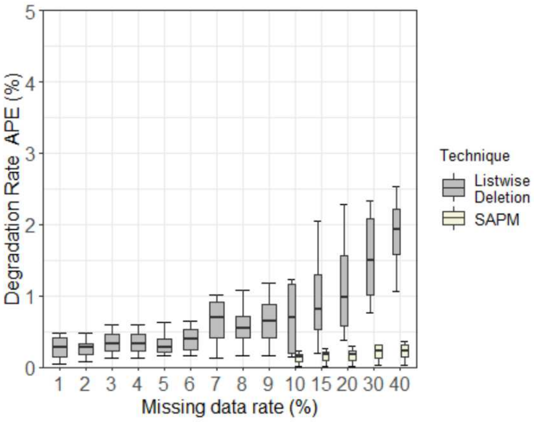


Figure 6: Boxplot of the average R_D APE of the 2000 invalid datasets estimated with OLS for 1-40% level of random missing data. The invalid datasets were reconstructed by the SAPM.

4 CONCLUSIONS

DQRs that operate on acquired field measurements were recently developed to ensure data validity. The proposed DQRs reconstruct invalid datasets through a sequence of data processing, quality checks, initial filtering stages, data deletion and inference techniques.

Expanding on this line of work, this study examined the application of ODRs and the impact of outlying data points (that are randomly distributed in the time series) on PV performance and reliability analyses. The ODRs results showed that the standard boxplot rule yielded the highest detection rate of 95.3% (by taking a moving data window) at 40% of outlying data points. In addition, the Hampel identifier achieved a detection rate of 94.8% at a level of 40% outlying data points. Therefore, the boxplot rule and Hampel identifier are recommended for detecting global outliers in PV performance datasets.

The results of the PV performance investigation showed that the effect of outlying data points (random missing data points) on the *PR* analysis was mitigated by listwise deletion. Furthermore, the annual R_D estimates were slightly biased in the presence of random missing data points. For 40% random missing data, the maximum *APE* of the linear R_D estimated by applying the listwise deletion was 2.53%.

Finally, the application of the SAPM for inferring the missing data yielded more accurate estimates when compared to estimates by listwise deletion, since for 40% random missing data, the *APE* of the annual R_D was less than 0.36%.

5 ACKNOWLEDGEMENTS

This work was funded through the SMARTER-PV project (BILATERAL/ISRAEL/0518/0017). The project is part of the RESTART 2016-2020 Programmes for Research, Technological Development and Innovation and is funded by the Research and Innovation Foundation (RIF) in Cyprus and the Israel Innovation Authority (IIA). This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technologies Office Award Number 34364. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

6 REFERENCES

[1] A. Livera, M. Theristis, G. Makrides, and G. E. Georghiou, "Recent advances in failure diagnosis techniques based on performance data analysis for grid-connected photovoltaic systems," *Renew. Energy*, vol. 133, pp. 126–143, 2019, doi: 10.1016/j.renene.2018.09.101.

[2] IEC 61724-1:2017, "Photovoltaic system performance - Part 1: Monitoring," 2017.

[3] IEC 61724-2:2016, "Photovoltaic system performance - Part2: Capacity evaluation method," 2016.

[4] IEC 61724-3:2016, "Photovoltaic system performance - Part 3: Energy evaluation method," 2016.

[5] G. Blaesser and D. Munro, "Guidelines for the assessment of photovoltaic plants Document A photovoltaic system monitoring. Commission of the European Communities, Joint Research Centre, Ispra, Italy, EUR 16338 EN, Issue 4.2, 1995."

[6] G. Blaesser and D. Munro, "Guidelines for the assessment of photovoltaic plants Document B Analysis and presentation of monitoring data, Commission of the European Communities, Joint research Centre, Ispra, Italy, EUR 16339 EN, 1995."

[7] S. Kurtz, J. Newmiller, T. Dierauf, A. Kimber, J. Mckee, and R. Flottemesch, "Analysis of Photovoltaic System Energy Performance Evaluation Method," no. November 2013, pp. 1–64, 2013.

[8] K. A. Klise and J. S. Stein, "Performance Monitoring using Pecos, SANDIA Report SAND2016-3583," 2016.

[9] S. Killinger, N. Engerer, and B. Müller, "QCPV: A quality control algorithm for distributed photovoltaic array power output," *Sol. Energy*, vol. 143, no. February, pp. 120–131, 2017, doi: 10.1016/j.solener.2016.12.053.

[10] A. Livera *et al.*, "Data processing and quality verification for improved photovoltaic performance and reliability analytics," *Prog. Photovoltaics Res. Appl.*, 2020 [Under Revision].

[11] M. C. Jessie Copper, Anna Bruce, Ted Spooner, "Australian Technical Guidelines for Monitoring and Analysing Photovoltaic Systems," *Aust. PV Institue*, vol. 1, no. 1, p. 44, 2013.

[12] R. Platon, J. Martel, N. Woodruff, and T. Y. Chau, "Online Fault Detection in PV Systems," *IEEE Trans. Sustain. Energy*, vol. 6, no. 4, pp. 1200–1207, 2015, doi: 10.1109/TSTE.2015.2421447.

[13] A. Woyte, M. Richter, D. Moser, M. Green, S. Mau, and H. G. Beyer, "Analytical Monitoring of Grid-connected Photovoltaic Systems," 2014.

[14] Y. Zhao, B. Lehman, R. Ball, J. Mosesian, and J.-F. De Palma, "Outlier Detection Rules for Fault Detection in Solar Photovoltaic Arrays," *28th IEEE Annu. Appl. Power Electron. Conf. Expo.*, 2013.

[15] Y. Zhao, F. Balboni, T. Arnaud, J. Mosesian, R. Ball, and B. Lehman, "Fault experiments in a commercial-scale PV laboratory and fault detection using local outlier factor," in *40th IEEE Photovoltaic Specialist Conference (PVSC)*, 2014, doi: 10.1109/PVSC.2014.6925661.

[16] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 519–533, 2010, doi: 10.1080/713827181.

[17] J. Ascencio-Vásquez, K. Brecl, and M. Topič, "Methodology of Köppen-Geiger-Photovoltaic climate classification and implications to worldwide mapping of PV system performance," *Sol. Energy*, vol. 191, no. August, pp. 672–685, 2019, doi: 10.1016/j.solener.2019.08.072.

[18] M. Theristis, V. Venizelou, G. Makrides, and G. E. Georghiou, "Chapter II-1-B – Energy yield in photovoltaic systems," in *Kalogirou, S.A. (Ed.), McEvoy's Handbook of Photovoltaics, third ed. Academic Press*, 2018, pp. 671–713.

[19] F. Mallor, T. León, L. De Boeck, S. Van Gulck, M. Meulders, and B. Van der Meerssche, "A method for detecting malfunctions in PV solar panels based on electricity production monitoring," *Sol. Energy*, vol. 153, pp. 51–63, 2017, doi: 10.1016/j.solener.2017.05.014.

[20] A. Livera, G. Makrides, J. Sutterlueti, and G. E. Georghiou, "Advanced Failure Detection Algorithms and Performance Decision Classification for Grid-connected PV Systems," in *33rd European Photovoltaic Solar Energy Conference and Exhibition (EU PVSEC)*, 2017.

[21] R. J. Rubin and D. B. Little, *Statistical Analysis with Missing Data*. John Wiley and Sons, New York, 1987.

[22] T. Dierauf, A. Growitz, S. Kurtz, and C. Hansen, "Weather-Corrected Performance Ratio," *National Renewable Energy Laboratory (NREL) Technical Report NREL/TP-5200-57991*, 2013. .

[23] A. Phinikarides, N. Philippou, G. Makrides, and G. E. Georghiou, "Performance loss rates of different photovoltaic technologies after eight years of operation under warm climate conditions," *29th Eur. Photovolt. Sol. Energy Conf. Exhib. (EU PVSEC)*, no. June, pp. 2664–2668, 2014, doi: 10.4229/EUPVSEC20142014-5BV.1.27.

[24] M. Theristis, A. Livera, C. B. Jones, G. Makrides, G. E. Georghiou, and J. S. Stein, "Nonlinear Photovoltaic Degradation Rates: Modeling and Comparison Against Conventional Methods," *IEEE J. Photovoltaics*, vol. 10, no. 4, pp. 1112–1118, 2020, doi: 10.1109/JPHOTOV.2020.2992432.

[25] M. Theristis *et al.*, "Modeling nonlinear photovoltaic degradation rates," in *47th IEEE Photovoltaic Specialist Conference (PVSC)*, 2020.

[26] M. Richter, J. Kalisch, T. Schmidt, E. Lorenz, and K. De Brabandere, "Best Practice Guide On Uncertainty in PV Modelling," *Public Rep. Perform. Plus WP2 Deliv. 2.4*, no. 308991, pp. 1–37, 2015.