Received: 29 April 2020 | Revised: 10 August 2020 | Accepted: 12 September 2020

DOI: 10.1002/pip.3349

RESEARCH ARTICLE

PROGRESS IN PHOTOVOLTAICS WILEY

Data processing and quality verification for improved photovoltaic performance and reliability analytics

Andreas Livera[1] | Marios Theristis[2] | Elena Koumpli[3] | Spyros Theocharides[1] | George Makrides[1] | Juergen Sutterlueti[4] | Joshua S. Stein[2] | 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, NM, 87185, USA

[3]SolarCentury, 90 Union Street, London, SE1 0NW, UK

[4]Gantner Instruments GmbH, Montafonerstraße 4, Schruns, 6780, Austria

Correspondence
Andreas Livera, PV Technology Laboratory, FOSS Research Centre for Sustainable Energy, Department of Electrical and Computer Engineering, University of Cyprus, Nicosia 1678, Cyprus.
Email: livera.andreas@ucy.ac.cy

Funding information
European Regional Development Fund and the Republic of Cyprus through the Cyprus Research & Innovation Foundation (RESTART 2016 - 2020 PROGRAMMES), Grant/Award Number: PV-ANALYTIC project (P2P/SOLAR/0818/0012); U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technologies Office Award Number 34364

Abstract

Data integrity is crucial for the performance and reliability analysis of photovoltaic (PV) systems, since actual in-field measurements commonly exhibit invalid data caused by outages and component failures. The scope of this paper is to present a complete methodology for PV data processing and quality verification in order to ensure improved PV performance and reliability analyses. Data quality routines (DQRs) were developed to ensure data fidelity by detecting and reconstructing invalid data through a sequence of filtering stages and inference techniques. The obtained results verified that PV performance and reliability analyses are sensitive to the fidelity of data and, therefore, time series reconstruction should be handled appropriately. To mitigate the bias effects of 10% or less invalid data, the listwise deletion technique provided accurate results for performance analytics (exhibited a maximum absolute percentage error of 0.92%). When missing data rates exceed 10%, data inference techniques yield more accurate results. The evaluation of missing power measurements demonstrated that time series reconstruction by applying the Sandia PV Array Performance Model yielded the lowest error among the investigated data inference techniques for PV performance analysis, with an absolute percentage error less than 0.71%, even at 40% missing data rate levels. The verification of the routines was performed on historical datasets from two different locations (desert and steppe climates). The proposed methodology provides a set of standardized analytical procedures to ensure the validity of performance and reliability evaluations that are performed over the lifetime of PV systems.

KEYWORDS
analytics, data fidelity, data inference, data quality, invalid values, performance, photovoltaics

1 | INTRODUCTION

High-quality data are of utmost importance for monitoring and facilitating advanced performance analytics of photovoltaic (PV) systems.[1] For the rapidly evolving PV industry, the benefits of increasing and improving operation and maintenance (O&M) practices through data-driven monitoring approaches are evident. In this sense, the quality and validity of the acquired data coupled with underlying data-driven

performance and reliability analytics are prerequisites for maintaining optimal performance over the lifetime of a system.

With respect to data integrity, invalid data (i.e., missing and outlying values), caused by power outages, equipment/component faults, communication failures, or interruption for maintenance reasons, are a commonly exhibited problem in PV monitoring systems. The processes and techniques applied to mitigate invalid data can potentially introduce noticeable bias that obscures underlying PV performance and reliability analyses. To this end, invalid datasets have to be detected and processed with appropriate mitigation tools, before commencing with data analytics.

Even though data quality constitutes the foundational block for performance and reliability analytics, only few reference guidelines and reports address issues that focus on data processing and quality control tools. Existing guidelines and reports are mainly limited to the requirements of monitoring systems with respect to data acquisition, outlier detection, and data processing for performance assessment of PV plants.[2–9]

In particular, several preprocessing data quality checks that include invalid data detection and filtering are outlined in the International Electrotechnical Commission (IEC) 61724 standard.[2–4] The standard recommends the application of a stage filter to ensure the presence of data during daylight hours (in-plane irradiance ≥ 20 W/m^2) followed by identification of gaps, duplicates, missing, and erroneous data points, which are also filtered out. The recommended methods for identifying invalid measurements include the application of threshold ranges (minimum and maximum parameter bounds), limits on the maximum rate of change between successive data points, statistical methods (not defined), comparisons among different sensors (if available), and clear-sky models to identify outliers. Furthermore, error codes signaled by sensors and data acquisition devices (DAQ) are recorded, and the timestamps are checked to identify gaps or duplicates in a given dataset. The identified invalid data may be discarded or treated by replacement with modeled or estimated values (from the valid data points recorded before or after the missing time step) or with averaged values (from the available data at that time period) in case of partial unavailability. The main disadvantage of the IEC 61724 falls in the qualitative description, which does not offer a case-specific approach that could enable reproducible and unbiased results.

Similarly, the European Joint Research Centre (JRC) guidelines recommend that all processed data are checked for consistency and gaps in order to identify data anomalies.[5,6] Reasonable ranges are set for each recorded parameter, and data points that fall outside these ranges or are otherwise inconsistent are filtered out. Other metrics such as the total time of monitoring activity and outage fraction are recommended; however, these guidelines also fail to provide a universal and quantitative approach for data quality.

A technical report from the National Renewable Energy Laboratory (NREL) highlighted the importance and challenges of obtaining high-quality data through periodic data quality checks.[7] The proposed data quality assurance checks include the identification of missing and erroneous values, inconsistencies in the frequency of data collection, filtering of nighttime measurements, identification of duplicate records, detection of underperformance (by comparing outputs of similar subarrays), detection of outlying and poor data from equipment malfunction (based on nearby sensor data or clear-sky models), and treatment of missing values (e.g., with averaged values or modeled data). Even though the data processing procedure holistically provides information on how to detect invalid data, it does not provide details on treating the identified invalid datasets, nor does it explicitly define how to handle missing data.

An open-source tool for PV monitoring (Pecos) was developed by Sandia National Laboratories (SNL).[8] This tool is designed to perform quality control checks on time series datasets in order to identify a wide range of anomalous conditions within a dataset. It leverages an initial time filter used to eliminate data points that fall outside specific time intervals (e.g., time filter between 3 a.m. and 9 p.m.) and subsequently applies quality control tests to diagnose missing data points by searching for blank (empty) cells, Not a Number ('NaN'), and Not Available ('NA') values within the dataset. Erroneous values are identified by setting physical limits (bounds) and checking that the difference between consecutive data values is within an expected range. Additionally, corrupt data points are detected by searching for specific values (i.e., some dataloggers will record '−999' or '999' values in the dataset), while duplicate timestamp records are detected by checking the time index for nonmonotonic, duplicate, and missing indexes. The tool (Python package) can automatically run a series of quality control tests and generate customized performance and monitoring reports.[10] This library has been validated using raw data from existing PV systems; however, similar to the aforementioned reports, it does not offer specific suggestions on how to detect invalid data from PV systems. Nonetheless, it could be used to implement such guidance.

A quality control routine for detecting invalid power measurements was presented by Killinger *et al*.[9] The algorithm identifies invalid power output data by setting physical limits, comparing measurements against sky models and indexes (e.g., extraterrestrial irradiance, clear-sky index, and PV power models) and also through the application of system statistics (e.g., variability check of measured power). This quality control routine is focused on the detection part without any information on how to handle the detected invalid data points.

Although the PV-related literature on handling invalid data is limited,[11–14] studies from other disciplines (e.g., mathematics and computer science) can provide insights into how to deal with data anomalies. More specifically, outliers (which are replaced by NA values) and missing values are dealt based on the missing data pattern (also called missing data mechanism). A study by Batista and Monard[15] demonstrated the presence of three main types of missing data patterns: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Missing data classified as MCAR occur when there is no specific mechanism of missingness, while MAR data occur when other variables affect the existence of missing values and the likelihood of having a missing value is independent of the value itself.[14] On the contrary, if the likelihood of having a

missing value is associated with the missing value itself, the exhibited missing data pattern is NMAR.[14] Identifying the type of the exhibited missing data pattern is important as it determines which treatment method is appropriate. This is a challenging task because the application of data deletion and inference techniques is strongly dependent on the missing data pattern and requires careful examination of the dataset in order to avoid the introduction of bias.

Existing data quality assurance guidelines analyze either the available valid measurements, excluding the invalid data/periods (by applying either listwise or pairwise deletion) or replace the missing data with modeled or estimated values.[16] In the case of listwise deletion, all rows with at least one missing data point are excluded from the analysis.[17] In pairwise deletion, only the missing values are removed. As such, bias may be introduced in the analysis depending on the missingness rate, deletion method, and so on causing false performance alarms and unnecessary maintenance activities.[11] In order to correct for this bias, missing data could be inferred (either imputed or estimated by a model); however, there is no PV-related investigation available to support this hypothesis. For the MCAR case, missing values can be ignored or inferred without knowing the reason the data are missing. In contrast, the missingness pattern must be thoroughly evaluated for the MAR and NMAR cases, before determining the most appropriate data quality routine.

A previous study demonstrated that missing data rates (defined as the ratio of missing values to the total number of data points) of less than 1% pose negligible impact on performance metrics, whereas missing data rates over 5% necessitate data inference techniques to yield accurate analytical results.[15] Different data inference techniques have been proposed based on statistical procedures, parametric models, empirical, and machine learning approaches.[11,12,15,18,19] Numerous studies propose data imputation with simulated data,[11,20] mean or median imputation,[21] optimally weighted average imputation,[22] multivariate imputation by chain equations (MICEs),[23] linear interpolation (LI),[14] *k*-nearest neighbors (*k*-NN) imputation,[15] last observation carried forward (LOCF), seasonal decomposition (SD),[24] bootstrapping,[21] and random forest (RF).[23] Dataset reconstruction is important for ensuring that the significant features of the time series are preserved and not lost due to reductions in the dimensionality.

A standardized and/or universally applicable mechanism of data quality control for PV performance and reliability analyses is not available. Hence, a complete and quantitative PV data processing methodology is proposed in this work for bridging the qualitative–quantitative gap that exists in current practices. The proposed methodology builds on quantifiable criteria from IEC 61724 and other PV data quality reports and minimizes existing process gaps that are presented in an ambiguous and/or qualitative manner. Such gaps can be translated into different ways depending on the PV performance analyst and can be one of the main sources of bias and inconsistency. Therefore, data quality routines (DQRs) that operate on measurements were developed, and each step of the methodology is described in a quantitative manner based on detailed analyses and not arbitrary assumptions. The aim is that the DQRs will become (or contribute to) an open-source library enabling the analysis of bulk PV data and, hence, benefitting the PV industry and research community. Functionality with other packages (such as PVLIB[25] and RdTools[26]) will also be investigated. The paper is organized in a way to present the complete methodology in the form of a table, where each quantifiable or decision-making step is justified. Finally, data visualization steps are included in the Appendix for completeness.

2 | METHODOLOGY

The methodology (Figure 1) builds on quantifiable criteria/steps from IEC 61724 standard[2–4] and other PV data quality reports.[5–9] It is a pipeline of sequentially structured DQRs that include the application of initial statistics, consistency examination, filtering, detection of invalid values and data rates, treatment of invalid data, and aggregation at different granularities.

The main focus of this work is on Step 6 of Figure 1. In order to provide a specific approach of handling invalid values, a number of parametric studies have been conducted aiming to provide answers to the following questions:

1. How should missing and outlying data be detected and visualized (see Appendix)?
2. When should data be filtered out (at different random and continuous missing data rates and sequences), by how much and by which method (i.e., listwise or pairwise)?
3. When should missing data be inferred (at different random and continuous missing data rates and sequences), by how much and by which method (i.e., statistical method or physical model)?
4. How should seasonal effects be treated?
5. What is the effect of the sampling method (i.e., instantaneous vs. averaged values)?
6. Is the methodology system- and location-independent?

In order to answer these questions, reference datasets were constructed using module and system measurements from two different locations. Both averaged and instantaneous measurements were utilized, and PV performance and reliability metrics were extracted. Artificially 'invalid' datasets were also generated by introducing missing data at different rates and sequences to enable a comparative analysis. The DQRs were then applied by detecting and treating the invalid datasets using different methods of deletion and inference. Each step of this parametric analysis was compared against the reference values in order to optimize the DQRs methodology.

2.1 | Experimental apparatus

The developed DQRs were validated against data of different sampling and geographical locations. The field measurements were acquired from a well-maintained test PV module installed at the outdoor test facility (OTF) of Gantner Instruments (GI) in Arizona, US (Köppen-Geiger-Photovoltaic climate classification *BK*; desert climate

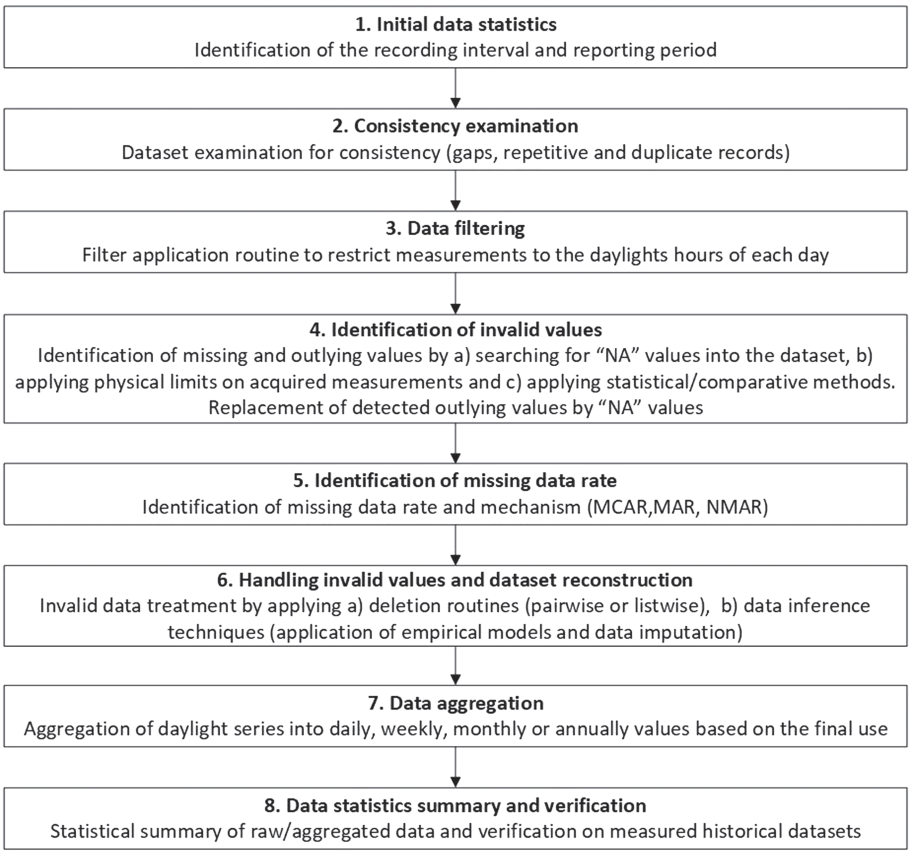

FIGURE 1 Flowchart of data processing and quality verification methodology

with very high irradiation)[27] and a test PV system at the OTF of the UCY in Nicosia, Cyprus (Köppen-Geiger-Photovoltaic climate classification *CH*; steppe climate with high irradiation)[27]; system availability was higher than 98%. The polycrystalline silicon (poly-c-Si) PV module was installed in an open-field mounting structure at the GI OTF, and it is rated at 220 W_p nominal power. The test PV system at the UCY OTF includes 5 poly-c-Si PV modules (rated at 205 W_p) that were connected in series to form a string of nominal power capacity 1.025 kW_p at the input of a string inverter.[28] The PV system is installed in an open-field mounting arrangement.

The electrical performance of the test PV module and system along with the prevailing irradiance and environmental conditions was recorded according to the requirements set by the IEC 61724 standard and stored with the use of a measurement monitoring platform.[2] The monitoring systems at both locations include solar irradiance (pyranometers), wind (anemometer and wind vane), temperature (thermocouples), and electrical (current shunts and transducers, voltage divider, and transducers) sensors connected to a central DAQ system that stores data at every second. Both outdoor test sites collect: in-plane irradiance (G_I), ambient air temperature (T_{amb}), wind speed (W_s), and direction (W_a). The PV measurements include module temperature (T_{mod}), array current (I_A), and voltage (V_A)—multiplied together to calculate the DC power (P_A) and the AC output power (P_{out}). Additional yields and performance metrics such as the final PV system yield (Y_f) and the reference yield (Y_r), the monthly performance ratio (*PR*), and the monthly temperature-corrected performance ratio (PR_{TC}), were also calculated.[2]

2.2 | Reference datasets

In order to demonstrate the effectiveness of the data processing and quality verification methodology, three reference datasets were generated to serve as a baseline for comparison and benchmarking in respect to (1) location-independence, (2) system-independence, (3) sampling method (instantaneous or averaged). The three reference datasets are (a) 1-year dataset containing 60-min instantaneous measurements from the test PV module in Arizona, (b) 1-year dataset with 60-min average measurements from the test PV system in Cyprus, and (c) 5-year dataset with 60-min average measurements from the PV system in Cyprus. The reference datasets include daylight measurements only (without any gaps, missing and outlying values).[2]

2.3 | Artificially invalid datasets

In an attempt to examine the impact of missing data points (indicated with NA) on PV performance and reliability analyses, different invalid datasets were generated by inserting artificial missing data points in the reference datasets.

More specifically, the artificially invalid datasets were generated by first selecting a missing data rate from 1% to 40% in whole number increments. Then, data records from the reference datasets were randomly selected and replaced with NA until the target missing data rate was reached. The selected data points were designated as MCAR in order to create artificial missing periods,[29] and then apply any data

treatment method to the missing data without the risk of introducing bias.[14,15] This process was repeated 50 times for each missing data rate (1%, 2%, ..., 40%), resulting in 2,000 invalid datasets per reference dataset. The record selection was performed in two specific ways:

- Random: missing data were randomly added by iteratively sampling a random number from 1 to 4,100 (daylight hours in a year) and assigning that hour as missing (NA) until the target missing data rate was reached.
- Continuous: the method for generating missing data assumed that the missing data were due to a sensor or system outage that resulted in continuous and consecutive missing data. To generate realizations for this case, a single random number from 1 to 4,100 was sampled, which was assumed to be the start time of the outage. For each target missing data rate (from 1% to 40% in whole numbers), the number of hours after the start time were all assigned as missing (NA). For example, in the case of a 10% missing data rate, 410 hours would be marked as NA beginning at the random start time. A constraint was added to ensure that the start time is early enough so that the entire missing period fits into the year.

A structured sampling method was used to ensure that random and continuous missing data points were evenly distributed in different months in order to capture seasonality. It should be noted that current O&M best practices in Europe restrict the continuous missing data rate to <10% for electrical and irradiance data.[7] When the rate is higher, the whole period is discarded.[7] However, this might not be the case worldwide, and therefore, this analysis extends to up to 40% of missing data rates for completeness.

During the investigation, different scenarios of invalid data points were considered, reflecting real PV system monitoring test cases of data loss:

1. Random and continuous invalid power measurements (irradiance measurements are available)
2. Random and continuous invalid irradiance measurements (power measurements are available)
3. Random and continuous invalid module temperature measurements (power and irradiance measurements are available)

Test Cases 1 and 2 represent PV installations which operate independent meteorological and electrical monitoring systems, while Test Case 3 represents the faulty operation of temperature sensors. Table 1 summarizes the different cases of invalid datasets that were investigated in this study.

TABLE 1 Summary of the different test cases that were considered in this study

Test case	Sequence of missing measurements	Missing measurements
1a	Random	P_A
1b	Continuous	P_A
2a	Random	G_I
2b	Continuous	G_I
3a	Random	T_{mod}
3b	Continuous	T_{mod}

Abbreviations: G_I, in-plane irradiance; P_A, array power (DC); T_{mod}, module temperature.

2.4 | Data handling techniques

The invalid datasets were analyzed by (a) discarding the missing measurements and analyzing only the available measurements (pairwise deletion method) and (b) discarding the missing periods with invalid measurements and analyzing only the periods with all data values available (listwise deletion method). To mitigate the effects of missing measurements, the invalid datasets were treated using data inference techniques that back-fill the missing measurements with estimated values from statistical or empirical models. In this work, the invalid data points were treated with two univariate imputation methods (RF and bootstrap), multiple imputation by predictive mean matching (PMM)[23] and empirical models. Univariate data imputation techniques (i.e., imputation of missing data in one parameter/column based on nonmissing observations from the same parameter/column) were selected when meteorological or electrical measurements were not available and for datalogging communication errors or other failures that affect all the recorded measurements/columns, and hence, imputation is only possible from pre- and post-observations.[21]

The employed RF method uses a RF trained on the observed values of the dataset to predict the missing values, while the bootstrap method relies on sampling from the posterior distributions and replacing the missing values with the sampled ones. The bootstrap algorithm then re-evaluates and resamples the posterior distribution and replaces the missing data with the newly sampled values.[21] When additional recorded measurements were available (i.e., electrical and/or meteorological measurements), statistical approaches based on multiple imputation and empirical models were applied for the inference process. For this purpose, imputation by PMM was employed. For each missing value, the PMM method creates a small set of candidate donors from the complete dataset that have predicted values closest to the predicted value of the missing data point.[30] In parallel, a simplified version of the Sandia PV Array Performance Model (SAPM) was also used to predict the power output of the PV modules/systems.[31,32] The electrical model (Equation 1) was selected because of the smaller number of input parameters (only the in-plane irradiance and module temperature measurements are required), its training capability and its high prediction accuracy for a range of PV module technologies under different climatic conditions (e.g., clear-sky, cloudy, and partly cloudy conditions).[11,32–34]

$$P_A = \frac{G_I}{G_{STC}}\left(P_{STC} + k_1 \ln\left(\frac{G_I}{G_{STC}}\right) + k_2 \ln\left(\frac{G_I}{G_{STC}}\right)^2 + k_3(T_{mod} - T_{STC})\right.$$
$$\left. + k_4(T_{mod} - T_{STC})\ln\left(\frac{G_I}{G_{STC}}\right) + k_5(T_{mod} - T_{STC})\ln\left(\frac{G_I}{G_{STC}}\right)^2 + k_6(T_{mod} - T_{STC})^2\right), \quad (1)$$

where G_{STC}, P_{STC}, and T_{STC} are the standard test conditions (STCs) irradiance, power and temperature, respectively, and $k_1 - k_6$ are empirical fit coefficients. In order to define the best set of empirical coefficients and capture the local climatic conditions, at least 40–50 days are required for the training process.[34] Similarly, module temperature was calculated from the in-plane irradiance and ambient temperature using the Ross thermal model[35]:

$$T_{mod} = T_{amb} + G_I\, k, \quad (2)$$

where k is the Ross coefficient extracted from the graphical representation of $T_{mod} - T_{amb}$ against G_I using the valid available measurements.

For installations that include wind speed measurements, it is possible to offset the influence of wind on the module temperature by using the Sandia module temperature model (SMTM)[31]:

$$T_{mod} = T_{amb} + G_I\left(e^{a + bW_s}\right), \quad (3)$$

where a and b are empirical coefficients to establish the upper limit for module temperature at low wind speeds and high solar irradiance and to account for forced convection by wind, respectively.

Finally, for locations under transient climatic conditions, module temperature can be calculated using the weighted-moving-average temperature model[36]:

$$T_{mod\,A,i} = \frac{\sum_{i=2}^{t_i \le 1200}\left(T_{SS,i}\, e^{-Pt_i}\right)}{\sum_{i=2}^{t_i \le 1200}\left(e^{-Pt_i}\right)}, \quad (4)$$

where i is the index of a number of prior timesteps, t_i is the number of seconds in the past for each timestep, $T_{SS,i}$ is the steady-state temperature prediction at t_i seconds in the past (°C), and $T_{mod\ A,i}$ is the moving-average model temperature prediction for the current timestep (°C).

2.5 | PV performance and reliability metrics

The *PR* was selected as the performance metric in this investigation because it is a normalized parameter and a key performance indicator (KPI), typically used to characterize PV plant performance for acceptance and operations testing.[37] The reliability of the PV modules was evaluated based on the performance loss rate (*PLR*), which can either be linear or nonlinear.[38–41] In this analysis, a constant *PLR* over time was assumed and estimated by applying linear regression with ordinary least squares (OLS) on the 5-year reference dataset of the test PV system in Cyprus.[42] The absolute *PLR* was calculated as follows:

$$PLR = a \cdot t, \quad (5)$$

where a is the slope and t is a conversion factor between the timestamp and years (e.g., 12 or 365 for monthly or daily aggregation, respectively). For a reliable evaluation of the *PLR*, at least a 5-year *PR* time series should be available to yield credible results that are not influenced by seasonal performance variations.[43,44] In order to compare the *PLR* obtained using the reference dataset against the *PLR* values obtained from the artificially invalid datasets (constructed in Section 2.3), the absolute percentage error (*APE*) was used:

$$APE = \left|\frac{A_t - P_t}{A_t}\right| 100, \quad (6)$$

where A_t is the actual value and P_t is the predicted value. For this analysis, A_t was set as the average monthly *PR* (or the reference *PLR* for reliability analysis) from the reference dataset with no missing data, while P_t was set as the average monthly *PR* of the 2,000 invalid datasets.

3 | RESULTS

3.1 | Data deletion methods

The influence of missing data on the *PR* estimates of the test PV module was evaluated by examining PR results from random (Test Case 1a) and continuous (Test Case 1b) sequences. The boxplots depicted in Figure 2 demonstrate the average monthly *PR* values over the 1-year evaluation period, obtained after reconstructing the 50 invalid datasets that included random or continuous missing power measurements at missing data rates up to 40%, using either the pairwise or listwise deletion methods. The results showed that the sequence of invalid measurements (random or continuous) along with the dataset reconstruction method significantly bias the performance analysis and can provide misleading results.

More specifically, the comparison between the *PR* of the reference and reconstructed (using pairwise deletion method) time series exhibited deviations in the range of 0.04–0.46 at missing data rates between 1% and 40% for randomly invalid datasets (Figure 2A). The results further showed that for a missing data rate of 40%, the *PR* that was reconstructed with pairwise deletion (0.517 ± 0.029) was not in agreement with the *PR* of the reference dataset (0.866 ± 0.034). Higher spread of the average monthly *PR* (in the range of 0.01–0.74) and lower deviations from the mean (0.553 ± 0.236 at 40% missing data rate) were observed when the continuous invalid datasets were reconstructed using the pairwise deletion method (Figure 2B).

The effect of random missing power measurements was mitigated by listwise deletion, even for a 40% missing data rate since the calculated *PR* of the invalid datasets (0.866 ± 0.033) agreed with the *PR* of the reference dataset (0.866 ± 0.034), as shown in Figure 2C. The results in Figure 2C, demonstrate that the effect of random missing power measurements (Test Case 1a) was successfully mitigated by listwise deletion, exhibiting an *APE* up to 0.12%, even at 40% missing data rates.

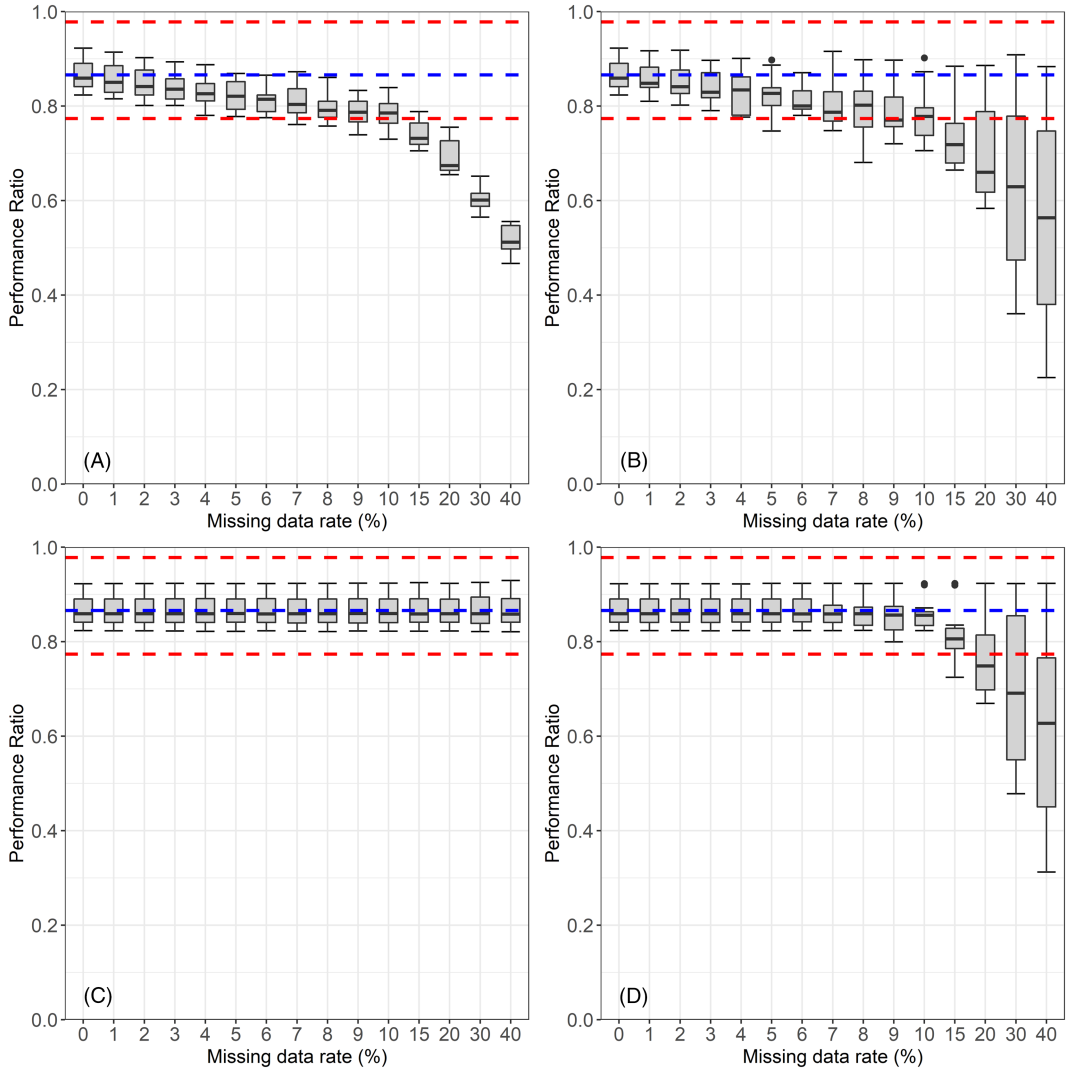

FIGURE 2 Boxplot of the average monthly performance ratio (PR) of the poly-c-Si PV module for (A) random missing power datasets (Test Case 1a) reconstructed using the pairwise deletion method, (B) continuous missing power datasets (Test Case 1b) reconstructed using the pairwise deletion method, (C) random missing power datasets (Test Case 1a) reconstructed using the listwise deletion method, and (D) continuous missing power datasets (Test Case 1b) reconstructed using the listwise deletion method. The horizontal red dashed lines indicate the upper and lower limits (±6% uncertainty on the calculated PR), while the blue dashed line indicates the average monthly PR of the reference yearly dataset. The upper and lower limits were calculated as the ±6% deviation from the highest and lowest PR values of the reference dataset [Colour figure can be viewed at wileyonlinelibrary.com]

The application of listwise deletion proved to be an effective method of handling continuous missing power measurements of up to 10% missing data rate by providing a maximum *APE* of 0.92% on the *PR*. Conversely, at higher missing data rates (in the range of 15% to 40%), the application of listwise deletion was not optimal since *APE* values up to 62.01% were obtained on the *PR*, as shown in Figure 2D. The large deviations observed at missing data rates higher than 10% (and for the worst-case scenario of Test Case 1b—whole month missing in a yearly dataset) signify the need of using other mitigation routines (i.e., application of data inference techniques to back-fill missing measurements).

Another important outcome of this investigation was that the application of listwise deletion was effective in reconstructing the time series for all invalid datasets of the investigated test cases (i.e., missing irradiance measurements and module temperature measurements in the case of PR_{TC}) at levels of up to 40% random and 10% continuous missingness.

3.2 | Data inference

3.2.1 | Data inference on missing power measurements

The resulting 2,000 invalid time series of Test Case 1b were reconstructed by applying different inference techniques (bootstrap, RF, multiple imputation by PMM, and back-filled using the variant of the SAPM model).[32] The average monthly *PR* of the reconstructed time series is shown in Figure 3, illustrating that the application of data

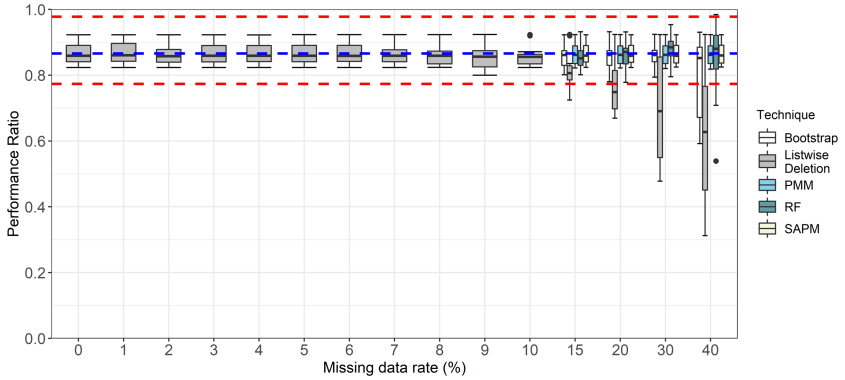

FIGURE 3 Boxplot of the average monthly performance ratio (PR) of the poly-c-Si PV module for continuous missing power datasets (Test Case 1b) reconstructed using different data imputation techniques and the Sandia Array Performance Model (SAPM). The horizontal red dashed lines indicate the ±6% uncertainty on the calculated PR, while the blue dashed line indicates the average PR of the reference dataset [Colour figure can be viewed at wileyonlinelibrary.com]

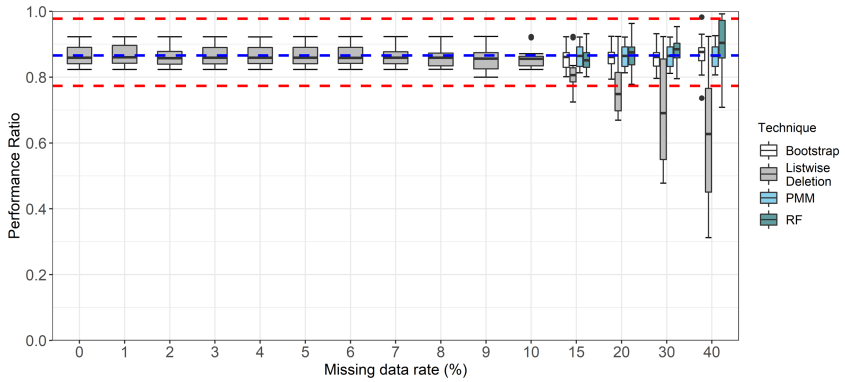

FIGURE 4 Boxplot of the average monthly performance ratio (PR) of the poly-c-Si PV module for continuous missing irradiance datasets (Test Case 2b) reconstructed using different data imputation techniques. The horizontal red dashed lines indicate the ±6% uncertainty on the calculated PR, while the blue dashed line indicates the average PR of the reference dataset [Colour figure can be viewed at wileyonlinelibrary.com]

inference techniques provided improved accuracies when compared to the application of listwise deletion. The maximum *APE* exhibited for the *PR* between the reference and reconstructed with listwise deletion time series was 37.91% and 62.01% at missing data rates of 30% and 40%, respectively. On the contrary, at 30% missing data rates, the maximum *APE* between the calculated *PR* of the reference and time series reconstructed with bootstrap and RF was lower than 13.92%. Moreover, when applying the PMM data inference technique, an *APE* of 1.2% in the *PR* was observed at a missing data rate of 40%.

The data inference evaluation demonstrated that time series reconstruction by applying SAPM yielded the lower error among the investigated data inference techniques, since the obtained *APE* of the *PR* was 0.71% even at 40% missing data rate levels. The results further showed that even at 40% of missingness, the *PR* calculated by SAPM (0.866 ± 0.034) was equal to the *PR* of the reference dataset (0.866 ± 0.034), as shown in Figure 3. This demonstrates the suitability of the SAPM as a data inference model.

Finally, the analysis conducted to investigate the effect of time series reconstruction on the *PR* calculation showed that data fidelity can be ensured with the application of data inference techniques that treat invalid datasets.

3.2.2 | Data inference on missing irradiance measurements

A benchmarking exercise was carried out by inferring the missing irradiance datasets (Test Case 2b) of the 2,000 invalid time series, and the boxplots depicted in Figure 4 show the average monthly *PR* values calculated from the reconstructed time series by employing univariate imputation by bootstrap and RF, and multiple imputation by PMM. The *PR* results showed that the application of data inference techniques for time series reconstruction yielded more accurate performance estimates when compared to the application of listwise deletion. More specifically, the maximum *APE* exhibited for the *PR* between the reference and reconstructed with bootstrap and RF time series was 4.89% and 13.82% at 30% missing data rates, respectively. In addition, multiple imputation by PMM yielded the highest accuracy among the investigated data inference techniques, demonstrating the lowest *APE* of 1.97% between the *PR* of the reference and inferred time series at 40% missing data rate.

3.2.3 | Data inference on missing module temperature measurements

In the case that module temperature measurements are missing in a continuous sequence (Test Case 3b), the evaluation of the PR_{TC} was performed by back-filling the missing measurements using the PMM imputation and the SMTM (thermal model selected for relatively steady climatic conditions).[31] For PR_{TC} estimates, knowledge of temperature is necessary.

Overall, the application of data inference techniques resulted in PR_{TC} estimates of lower *APE*, when compared to the listwise deletion technique (Figure 5). The PR_{TC} *APE* between the reference and reconstructed with listwise deletion time series was 49.47% and 67.01% at missing data rates of 30% and 40%, respectively. On the contrary, the *APE* between PR_{TC} of the reference and the reconstructed by PMM time series was less than 1.36% for a missing data rate of 40%.

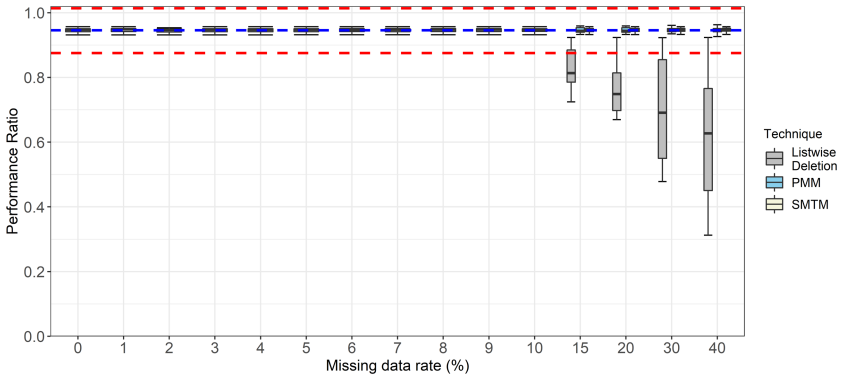

FIGURE 5 Boxplot of the average monthly temperature-corrected performance ratio (PR_{TC}) of the poly-c-Si PV module for continuous missing module temperature datasets (Test Case 3b) reconstructed using imputation by predictive mean matching (PMM) and Sandia module temperature model (SMTM). The horizontal red dashed lines indicate the ±6% uncertainty on the calculated PR, while the blue dashed line indicates the average PR of the reference dataset [Colour figure can be viewed at wileyonlinelibrary.com]

Finally, the SMTM exhibited the lowest *APE* of the PR_{TC} up to 0.28% even for a 40% missing data rate, among the investigated data mitigation techniques. The average monthly PR_{TC} constructed by SMTM (0.947 ± 0.007) was in very close agreement to the average monthly PR_{TC} of the reference dataset (0.946 ± 0.007), as shown in Figure 5. This demonstrates the suitability of the SMTM as a data inference model.

3.2.4 | Data deletion and inference verification

The investigation was also applied to the 2,000 invalid datasets of the test PV system installed in Cyprus in order to verify the location, system and sampling independence of the proposed DQRs. Table 2 shows close agreement to the results obtained when analyzing the time series of the test PV module in Arizona, verifying the performance of the proposed DQRs methodology independent of system, location, and sampling method.

3.3 | Data integrity effect on *PLR* analysis

In order to test the sensitivity of the *PLR* to invalid data points, an analysis was conducted by comparing the reference PLR_{ref} values against the estimates from the 2,000 invalid datasets of Case 1b.

The *PLR* analysis demonstrated that the annual *PLR* (calculated by applying OLS to the monthly *PR* time series of the PV system in Cyprus) was sensitive to the amount of continuous invalid power datasets (Test Case 1b), even at 1% of missing rates (Figure 6). The

TABLE 2 Summary of results of the proposed data quality routines applied for Test Cases 1–3 on invalid datasets from the PV system in Cyprus

	Test cases 1a, 2a, 3a		Test case 1b		Test case 2b		Test case 3b	
Missing data rate (%)	Mitigation technique	*APE* (%)	Mitigation technique	*APE* (%)	Mitigation technique	*APE* (%)	Mitigation technique	*APE* (%)
1–10	Listwise	0.15	Listwise	0.96	Listwise	0.96	Listwise	0.96
15	Listwise	0.15	SAPM	0.45	PMM	1.85	SMTM	0.19
20	Listwise	0.15	SAPM	0.47	PMM	1.85	SMTM	0.22
30	Listwise	0.15	SAPM	0.73	PMM	1.85	SMTM	0.28
40	Listwise	0.15	SAPM	0.81	PMM	2.01	SMTM	0.35

Abbreviations: *APE*, absolute percentage error; PMM, predictive mean matching; SAPM, Sandia array performance model; SMTM, Sandia module temperature model.

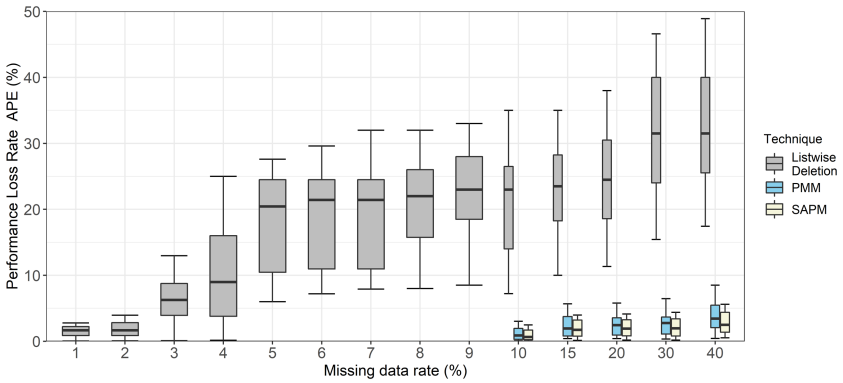

FIGURE 6 Boxplot of the performance loss rate absolute percentage error (APE) of the 2000 emulated invalid datasets with continuous missing power measurements (Test Case 1b) calculated by applying ordinary least squares (OLS) for 1% to 40% missing data rate. The invalid datasets were reconstructed using imputation by predictive mean matching (PMM) and Sandia array performance model (SAPM). For 1% to 9% missing data rate, the boxplots of the datasets reconstructed using data inference techniques were omitted because the exhibited low error (maximum APE of 2.96%) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Data processing and quality verification framework

1. Initial data statistics	
a. Identification of the recording interval and reporting period	• Identify the recording interval and reporting period • Identify the number of rows and columns
2. Consistency examination	
a. Dataset examination for consistency (timestamp gaps, repetitive and duplicate records, and synchronization issues)	• Check for repetitive, duplicate and 'NA' timestamp records • Remove repetitive and duplicate timestamp records • Verification against known (simulated) timestamp series • Time series synchronization between meteorological and electrical data
3. Data filtering	
a. Filter application	• Apply a daylight filter by: ◦ An irradiance filter (e.g., G_I > 20 W/m^2),[2] or a time filter (sunrise < time < sunset) or a sun position filter (e.g., solar elevation > 10°)
4. Identification of invalid values[a]	
a. Identification of outlying values (i.e., values that are out of range)	• Identify outliers by: a) Physical limits (threshold ranges)[4,45]: ◦ 0 W/m^2 < G_I < 1,300 W/m^2 ◦ 0 W < P_{out} < 1.02 × AC inverter power rating W ◦ 0 V < V_A < 1.3 × V_{OC} of the array V ◦ 0 A < I_A × 1.5 × I_{SC} A ◦ −40°C < T_{amb} < 60°C ◦ −40°C < T_{mod} < 100°C for open rack mounted ◦ −40°C < T_{mod} < 120°C for roof-mounted and building-integrated systems ◦ 0 m/s < W_s < 32 m/s ◦ 0% < PR < 110% b) Comparison of measurements from different/multiple sensors, sky models, and indices (e.g., clear-sky and PV power models)[9] c) Maximum change between successive data points (applicable only for up to 15-min time interval)[4]: ◦ G_I > 800 W/m^2 ◦ P_{out} > 80% rating W ◦ T_{amb} > 4°C ◦ T_{mod} > 4°C ◦ W_s > 10 m/s d) Visual inspection of scatter plots[47] e) Apply statistical and comparative tests (local outlier factor, sigma rule, Hampel identifier, boxplot rule, etc.)[2,7] • Replace outliers by 'NA' values
b. Identify missing values	• Search for 'NA' or 'NaN' values and blank cells
5. Identification of missing data rate	
a. Identification of missing data mechanism and rate	• Identify missing data mechanism (MCAR, MAR, or NMAR) by applying a visualization method (see Appendix A) • Identify missing data rate and missingness rate for every recorded field measurement
6. Handling invalid values and dataset reconstruction	
a. Invalid data treatment	• Missing data rates lower than 10%: ◦ Discard the missing period (listwise deletion) or ◦ Infer the missing measurements for (a) a whole month missing for a yearly performance analysis and (b) providing robust degradation and performance loss rate estimates • Missing data rates higher than 10%: ◦ If meteorological data are available, infer the missing data using empirical models a) Back-fill missing power measurements for c-Si PV modules using the SAPM from the Python Library (PVLIB),[25] or a variant of the model[32,48] b) Back-fill missing module temperature measurements using the SMTM[31] (or the Ross thermal model[35] when only G_I and T_{amb} are available). Other options include the open-source Faiman module temperature model and the transient weighted moving-average model[36] (for transient climatic conditions) from the Python Library (PVLIB)[25]

TABLE 3 (Continued)

	◦ If meteorological and satellite data are not available, impute the missing values using univariate data imputation techniques[b]: a) Impute missing power measurements using the bootstrapping univariate data imputation technique b) Impute missing module temperature measurements using the bootstrapping univariate data imputation technique ◦ If satellite data are available, infer the missing meteorological measurements using satellite observations[c] ◦ If electrical data are available, impute the missing data using multiple imputation techniques: a) Impute missing meteorological measurements using multiple imputation by PMM
7. Data aggregation	
a. Data aggregation	• Data aggregation (daily, weekly, monthly, or annually). The aggregation step depends on the final use of the measured PV system data
8. Data statistics summary	
a. Statistics summary	• Final data validity statistical summary (mean, median, min and max values, NAs, and correlation matrix)

Note: Quantitative recommendations that are adapted from the IEC 61724 standard and/or other guidelines are cited accordingly.
Abbreviations: G_I, in-plane irradiance; I_A, array current; I_{SC}, short circuit current; MAR, missing at random; MCAR, missing completely at random; NA, not available; NaN, not a number; NMAR, not missing at random; P_A, array power (DC); PMM, predictive mean matching; P_{out}, AC output power; *PR*, performance ratio; PV, photovoltaic; SAPM, Sandia array performance model; SMTM, Sandia module temperature model; T_{amb}, ambient air temperature; T_{mod}, module temperature; V_A, array voltage; V_{OC}, open circuit voltage; W_s, wind speed.
[a]Recommendations provided at Step 4, may be adjusted based on the PV system design and local conditions.
[b]Further investigations are required to verify the applicability of univariate data imputation methods in cloudy locations.
[c]Further analysis is required.[49,50]

higher *APE* deviations (in the range of 6.11%–35.01% at missing rates of 5%–10%) provide evidence that the application of listwise data deletion introduces a bias in the calculation of the *PLR* consistently at increasing missing data rates. Furthermore, the data inference using SAPM and PMM yielded more accurate *PLR* results when compared to the *PLR* estimates obtained by listwise deletion. In particular, for a missing data rate of 10%, the maximum *APE* of the *PLR* was 35.01% by listwise deletion, while when using the SAPM and multiple imputation by PMM, the obtained *APE* of the *PLR* was less than 3.03%. For a missing data rate of 40%, the maximum *APE* of the *PLR* calculated by listwise deletion was 48.91%, whereas data inference techniques demonstrated an *APE* lower than 8.50% (average *APE* of 3.86%). The SAPM yielded robust *PLR* estimates; for a missing data rate level of 40%, the maximum exhibited *APE* of the calculated *PLR* was 5.69% (average *APE* of 2.82%).

4 | OUTLINE OF DATA PROCESSING AND QUALITY VERIFICATION FRAMEWORK

The proposed methodology is a pipeline of sequentially structured DQRs that include the application of initial statistics, consistency examination, filtering, detection of invalid values and missing data rates, treatment of invalid data, and aggregation at different granularities.

The initial step (Step 1) uses data statistics to determine the recording interval (time between two consecutive time records) and the reporting period. For PV performance and reliability analyses, the reporting period should be long enough to provide representative PV operational data and ambient conditions (i.e., minimum of 1-year of continuous monitoring for outdoor PV performance evaluation).[45] The fidelity of the time series is then examined to find timestamp gaps, repetitive entries, duplicate records, and synchronization issues between meteorological and electrical data (Step 2). After removing the repetitive and duplicate timestamp records, the time series is verified and resampled against known (i.e., simulated) timestamp series. In case of mismatches between timestamps of different dataloggers (e.g., weather and electrical measurements are logged separately), a data time series synchronization is performed. A daylight filter (G_I > 20 W/m^2) is then applied to the dataset (Step 3).[2] Alternative daylight filters may include the use of time or sun elevation filters.

The missing values are then identified by searching for NA or NaN values within the examined dataset. Outliers are detected by imposing physical limitations on the recorded data, applying variation limits between successive data points methods and statistical and comparative methods (e.g., Local Outlier Factor, Sigma rule, Hampel identifier, boxplot rule, and rolling mean).[4,7,46] The detected outlying values are then replaced by NA values. At this point, the missing data rate is calculated. The next step (Step 5) is to identify the missing data mechanism (MCAR, MAR, and NMAR) by applying a suitable data visualization method (e.g., heatmaps, aggregation, scatter, and spine plots). Once the invalid data points are detected, the dataset is reconstructed based on the calculated missing data rate and mechanism (Step 6).[15] Thus, missing values are either treated by data deletion (pairwise or listwise) or inference techniques (e.g., empirical models, multiple, and univariate data imputations) in the form of dataset reconstruction routines.

In the case of missing data rates lower than 10%, the missing periods should be discarded from the dataset (listwise deletion). To

mitigate the effects of missing measurements, data inference techniques should be considered for missing data rates higher than 10%. Data inference with SAPM is recommended for missing power measurements (since this model has already been tested for a range of PV module technologies and climatic conditions), whereas missing temperature measurements can be inferred with a model such as SMTM, or the Ross thermal model (if no wind speed data is available), or the transient weighted moving-average model (for transient climatic conditions). In addition, multiple imputation techniques (e.g., multiple imputation by PMM) should be used to infer the missing irradiance measurements.

Once the dataset is treated and reconstructed, aggregation is applied depending on the final use (Step 7). Final data statistics are then recorded based on the reconstructed dataset (Step 8). Table 3 summarizes the data processing and quality verification steps, with all quantifiable metrics and specifications of each DQR in the form of guidelines.

5 | CONCLUSIONS

A unified methodology for PV data processing, quality verification, and reconstruction is presented in an attempt to reduce bias and enable reproducible PV performance, degradation, and *PLR* analyses. The methodology is a pipeline of sequentially structured DQRs that include the application of initial data statistics, consistency examination, filtering, detection of invalid data and missing data rate, invalid data treatment, and dataset aggregation. More specifically, the DQRs operate on the measurements (either instantaneous or averages) and detect and reconstruct invalid data by applying data deletion and inference techniques.

The results demonstrated that PV performance assessment and *PLR* studies are sensitive to the invalid or missing data rate. Missing data rates lower than 10% can be reconstructed using the listwise deletion method producing results in close agreement to the actual (maximum *APE* of 0.92% on the *PR*). At higher missing data rates (between 15% and 40%), the application of listwise deletion is not recommended, since *APE* up to 62.01% was observed. For missing data rates higher than 10%, data inference techniques are recommended. Data inference with SAPM is recommended for missing power measurements whereas inference with SMTM is advised for temperature data (or inference with the weighted moving-average model for transient climatic conditions). PMM should be used for inferring irradiance measurements.

The proposed methodology is relatively straightforward and can become a concrete block for feeding high-quality and refined data to automated data-driven performance functionalities of monitoring systems.

ACKNOWLEDGEMENTS

This work was funded through the PV-ANALYTIC project (P2P/SOLAR/0818/0012) which was co-financed by the European Regional Development Fund and the Republic of Cyprus through the Cyprus Research & Innovation Foundation (RESTART 2016 – 2020 PROGRAMMES). The work of Marios Theristis and Joshua S. Stein was 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.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

ORCID

Andreas Livera https://orcid.org/0000-0002-3732-9171
Marios Theristis https://orcid.org/0000-0002-7265-4922
Spyros Theocharides https://orcid.org/0000-0003-2164-6081
Joshua S. Stein https://orcid.org/0000-0002-9422-1976
George E. Georghiou https://orcid.org/0000-0002-5872-5851

REFERENCES

1. Livera A, Theristis M, Makrides G, Georghiou GE. Recent advances in failure diagnosis techniques based on performance data analysis for grid-connected photovoltaic systems. *Renew Energy*. 2019;133:126-143. https://doi.org/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. Blaesser G, Munro D. 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. Blaesser G, Munro D. 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. Kurtz S, Newmiller J, Dierauf T, Kimber A, Mckee J. Flottemesch R. *Analysis Photovoltaic Sys Enrgy Perform Eval Meth*. November 2013; 2013:1-64.
8. Klise KA, Stein JS. Performance monitoring using Pecos, *SANDIA Report SAND2016–3583*; 2016.
9. Killinger S, Engerer N, Müller B. QCPV: a quality control algorithm for distributed photovoltaic array power output. *Sol Energy*. 2017;143:120-131. https://doi.org/10.1016/j.solener.2016.12.053
10. Klise KA, Stein JS. Automated performance monitoring for PV systems using Pecos. In: 43rd IEEE Photovoltaic Specialist Conference (PVSC); 2016. https://doi.org/10.1109/PVSC.2017.8366806
11. Koubli E, Palmer D, Rowley P, Gottschalg R. Inference of missing data in photovoltaic monitoring datasets. *IET Renew Power Gener*. 2016;10(4):434-439. https://doi.org/10.1049/iet-rpg.2015.0355

12. Panapakidis IP, Bouhouras AS, Christoforidis GC. A missing data treatment method for photovoltaic installations. *IEEE Int Energy Conf*. 2018; 2018:1-6. https://doi.org/10.1109/ENERGYCON.2018.8398780
13. Turrado CC, López MDCM, Lasheras FS, Gómez BAR, Rollé JLC, Juez FJDC. Missing data imputation of solar radiation data under different atmospheric conditions. *Sensors (Switzerland)*. 2014;14(11): 20382-20399. https://doi.org/10.3390/s141120382
14. Demirhan H, Renwick Z. Missing value imputation for short to mid-term horizontal solar irradiance data. *Appl Energy*. 2018;225: 998-1012. https://doi.org/10.1016/j.apenergy.2018.05.054
15. Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. *Appl Artif Intell*. 2010;17(5–6): 519-533. https://doi.org/10.1080/713827181
16. Rubin RJ, Little DB. *Statistical Analysis with Missing Data*. New York: John Wiley and Sons; 1987.
17. Peugh JL, Enders CK. Missing data in educational research: a review of reporting practices and suggestions for improvement. *Rev Educ Res*. 2004;74(4):525-556. https://doi.org/10.3102/00346543074004525
18. Li D, Deogun J, Spaulding W, Shuart B. Towards missing data imputation: a study of fuzzy K-means clustering method. In: Rough Sets and Current Trends in Computing, 4th International Conference (RSCTC 2004); 2004.
19. Grzymala-Busse, JW Goodwin LK. Handling missing attribute values in preterm birth data sets. In: Rough Sets and Current Trends in Computing, 4th International Conference (RSCTC); 2004. https://doi.org/10.1007/11548706
20. Platon R, Martel J, Woodruff N, Chau TY. Online fault detection in PV systems. *IEEE Trans Sustain Energy*. 2015;6(4):1200-1207. https://doi.org/10.1109/TSTE.2015.2421447
21. Phinikarides A, Makrides G, Georghiou GE. Estimation of the degradation rate of fielded photovoltaic arrays in the presence of measurement outages. In: 32nd European Photovoltaic Solar Energy Conference and Exhibition (EU PVSEC); 2016.
22. Peppanen J, Zhang X, Grijalva S, Reno MJ. Handling bad or missing smart meter data through advanced data imputation. *2016 IEEE Power Energy Soc Innov Smart Grid Technol Conf ISGT*. 2016;2016:1–5. https://doi.org/10.1109/ISGT.2016.7781213
23. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1-67. https://doi.org/10.18637/jss.v045.i03
24. Livera A, Phinikarides A, Makrides G, Georghiou GE. Impact of missing data on the estimation of photovoltaic system degradation rate. *44th IEEE Photovolt Spec Conf*. 2018;1954-1958. https://doi.org/10.1109/pvsc.2017.8366442
25. Holmgren WF, Hansen CW, Mikofski MA. Pvlib Python: a python package for modeling solar energy systems. *J Open Source Softw*. 2018;3(29):884. https://doi.org/10.21105/joss.00884
26. Jordan DC, Deline C, Kurtz SR, Kimball GM, Anderson M. Robust PV degradation methodology and application. *IEEE J Photovoltaics*. 2018; 8(2):525-531. https://doi.org/10.1109/JPHOTOV.2017.2779779
27. Ascencio-Vásquez J, Brecl K, Topič M. Methodology of Köppen-Geiger-photovoltaic climate classification and implications to worldwide mapping of PV system performance. *Sol Energy*. 2019;191: 672-685. https://doi.org/10.1016/j.solener.2019.08.072
28. Theristis M, Venizelou V, Makrides G, Georghiou GE. Chapter II-1-B—energy yield in photovoltaic systems. In: Kalogirou SA, ed. *McEvoy's Handbook of Photovoltaics*. Third ed. Academic Press; 2018: 671-713.
29. Luengo J, García S, Herrera F. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowl and Inform Syst*. 2012;32(1):77-108. https://doi.org/10.1007/s10115-011-0424-2
30. Little RJA. Missing-data adjustments in large surveys. *J Bus Econ Stat*. 1988;6(3):287-296. https://doi.org/10.1080/07350015.1988.10509663
31. King DL, Boyson WE, Kratochvil JA. Photovoltaic array performance model, SANDIA Report SAND2004-3535. 2004;8. https://doi.org/10.2172/919131
32. Huld T, Friesen G, Skoczek A, et al. A power-rating model for crystalline silicon PV modules. *Sol Energy Mater sol Cells*. 2011;95(12): 3359-3369. https://doi.org/10.1016/j.solmat.2011.07.026
33. Dittmann S, Friesen G, Williams S, et al. Results of the 3rd modelling round Robin within the European project "PERFORMANCE"—comparison of module energy rating methods. In: 25th European Photovoltaic Solar Energy Conference (EU PVSEC); 2010. https://doi.org/10.4229/25thEUPVSEC2010-4AV.3.109
34. Koumpli E. Impact of data quality on photovoltaic (PV) performance assessment, Doctoral Thesis University of Loughborough; 2017.
35. Ross RG. Interface design considerations for terrestrial solar cell modules. In: *12th IEEE Photovoltaic Specialist Conference (PVSC)*; 1976.
36. Prilliman M, Stein JS, Riley D, Tamizhmani G. Transient weighted moving-average model of photovoltaic module back-surface temperature. *IEEE J Photovoltaics*. 2020;10(4):1053-1060. https://doi.org/10.1109/JPHOTOV.2020.2992351
37. Dierauf T, Growitz A, Kurtz S, Hansen C. Weather-corrected performance ratio National Renewable Energy Laboratory (NREL) Technical Report NREL/TP-5200-57991.
38. Jordan DC, Silverman TJ, Sekulic B, Kurtz SR. PV degradation curves: non-linearities and failure modes. *Prog Photovoltaics Res Appl*. 2017; 25(7):583-591. https://doi.org/10.1002/pip.2835
39. Phinikarides A, Kindyni N, Makrides G, Georghiou GE. Review of photovoltaic degradation rate methodologies. *Renew Sustain Energy Rev*. 2014;40:143-152. https://doi.org/10.1016/j.rser.2014.07.155
40. Theristis M, Livera A, Micheli L, et al. Modeling nonlinear photovoltaic degradation rates. In: *47th IEEE Photovoltaic Specialist Conference (PVSC)*; 2020.
41. Theristis M, Livera A, Jones CB, Makrides G, Georghiou GE, Stein JS. Nonlinear photovoltaic degradation rates: modeling and comparison against conventional methods. *IEEE J Photovoltaics*. 2020;10(4):1112-1118. https://doi.org/10.1109/JPHOTOV.2020.2992432
42. Makridakis S, Wheelwright S, Hyndman R. *Forecasting: Methods and Applications*. 3rd ed. New York: John Wiley & Sons; 1998.
43. Osterwald CR, Adelstein J, Cueto JA, Kroposki B, Trudell D, Moriarty T. Comparison of degradation rates of individual modules held at maximum power. In: *4th World Conference on Photovoltaic Energy Conference (WCPEC)*. Waikoloa, Hawaii; 2006:2085-2088. https://doi.org/10.1109/WCPEC.2006.279914.
44. Makrides G, Theristis M, Bratcher J, Pratt J, Georghiou GE. Five-year performance and reliability analysis of monocrystalline photovoltaic modules with different backsheet materials. *Sol Energy*. 2018;171: 491-499. https://doi.org/10.1016/j.solener.2018.06.110
45. Copper J, Bruce A, Spooner T, Calais M, Pryor T, Watt M. Australian technical guidelines for monitoring and analysing photovoltaic systems. *Australian PV Institute*. 2013;1(1).
46. Zhao Y, Balboni F, Arnaud T, Mosesian J, Ball R, Lehman B. Fault experiments in a commercial-scale PV laboratory and fault detection using local outlier factor. In: 40th IEEE Photovoltaic Specialist Conference (PVSC); 2014. https://doi.org/10.1109/PVSC.2014.6925661
47. Woyte A, Richter M, Moser D, Green M, Mau S, Beyer HG. Analytical monitoring of grid-connected photovoltaic systems: good practice for monitoring and performance analysis, IEA International Energy Agency, IEA PVPS Task 13, Subtask 2 Report IEA-PVPS T13–03: 2014; 2014. https://doi.org/10.13140/2.1.1133.6481
48. Ransome S, Sutterlueti J, Instruments G, Solutions E. A systematic comparison of 12 empirical models used for energy yield prediction VS PV technology. In: 33rd European Photovoltaic Solar Energy Conference (EU PVSEC); 2017.
49. Drews A, Beyer HG, Rindelhardt U. Quality of performance assessment of PV plants based on irradiation maps. *Sol Energy*. 2008;82(11): 1067-1075. https://doi.org/10.1016/j.solener.2008.04.009

50. Palmer D, Koubli E, Cole I, Betts T, Gottschalg R. Satellite or ground-based measurements for production of site specific hourly irradiance data: which is most accurate and where? *Sol Energy*. 2018;165:240-255. https://doi.org/10.1016/j.solener.2018.03.029

How to cite this article: Livera A, Theristis M, Koumpli E, et al. Data processing and quality verification for improved photovoltaic performance and reliability analytics. *Prog Photovolt Res Appl*. 2021;29:143–158. https://doi.org/10.1002/pip.3349

APPENDIX A.

A.1 | DQRs visualization

A dataset from a PV system installed at the OTF of the UCY was utilized to visualize the steps of the data processing and quality verification methodology.

Step 1:

The recording interval is 15-minute, and the reporting period is 365 days, which is in line with the minimum reporting period requirement of 1 year for PV performance assessment. In addition, the dataset consisted of 35,040 rows (records) and 8 columns (recorded field measurements); *Date/Time*, in-plane irradiance (G_I), ambient air temperature (T_{amb}), module temperature (T_{mod}), array current (I_A), voltage (V_A) and power (P_A) at the DC side, and AC output power (P_{out}).

Step 2:

The dataset was examined for consistency; 5 repetitive/duplicate and NA timestamp (*Date/Time*) records were identified. After removing the repetitive and duplicate timestamp records, the time series was checked and verified against known (simulated) timestamp series.

Step 3:

An irradiance filter (G_I > 20 W/m^2) was then applied to the dataset to yield daylight time series. As a result, the number of rows was further reduced to 15,820.

Step 4:

Missing values were identified by searching for NA values in the dataset, while the outlying data points were detected by imposing range limits on the data and visually inspecting scatter plots. The developed DQRs identified 61 outlying (erroneous) in-plane irradiance and AC output power data points through the application of physical limitations. Visual inspection of the irradiance-power diagnostic plot (Figure A1) was deemed sufficient for detecting outlying values. Observations close to the irradiance-power linear relationship line are assumed to be normal, while observations far from this line are considered as outliers.

Similarly, the dataset was also examined for global outliers using automated approaches (i.e., Sigma rule method, Hampel identifier, or standard boxplot rule). The outlier detection routines (ODRs) identified 31 outlying values for the recorded AC power measurements (the ODRs results for the AC and DC power measurements are depicted in Figure A2).

A comparison of measurements from two identical pyranometers installed nearby was performed to interpret the working condition of the irradiance sensor. The irradiance measurements acquired from the pyranometers were plotted on the scatter diagram of Figure A3, showing their linear relationship. The extracted determination coefficient (R^2) was 0.998, indicating that the system irradiance sensor was operating properly.

The detected outlying values were replaced by NA.

Step 5:

The missing data rate was calculated. The portion of missingness for each recorded measurement is depicted in Figure A4 (where black color indicates missing values while grey color represents available measurements).

Additionally, in order to identify the missing data mechanism, an aggregation plot (Figure A5) was used to visualize the data and expose the relationship between available and missing data points. Figure A5A shows a bar for the recorded measurements where the bar height corresponds to the proportion of missing values. Figure A5B shows the missingness pattern for the variables. Closer inspection of Figure A5B reveals no links between the missing values for the acquired measurements. Thus, the missing data mechanism is MCAR.

Step 6:

The DQRs provide algorithms for handling missing invalid values and dataset reconstruction. Since the missing data rate was less than 10%, the identified invalid data points are treated by listwise deletion.

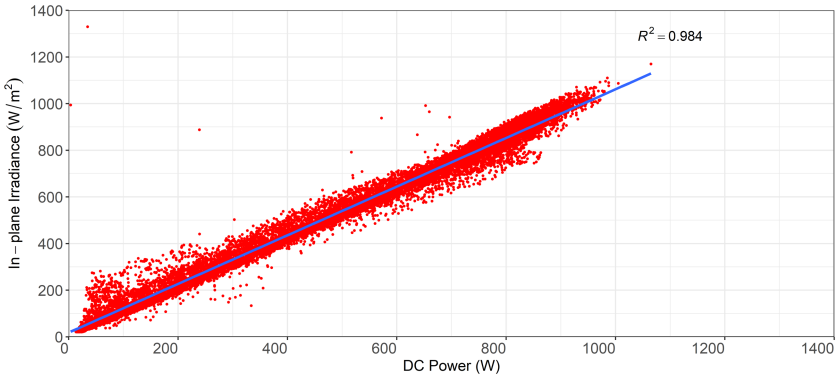

FIGURE A1 Performance monitoring scatter plot of in-plane irradiance versus array power [Colour figure can be viewed at wileyonlinelibrary.com]

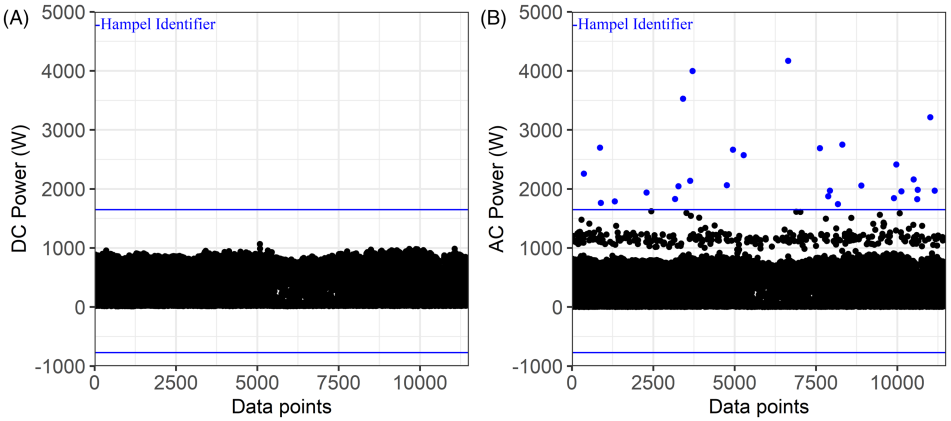

FIGURE A2 Outlier detection rules (ODRs) results for (A) DC power and (B) AC output power measurements. The upper and lower limits for the Hampel identifier rule are depicted by blue lines [Colour figure can be viewed at wileyonlinelibrary.com]

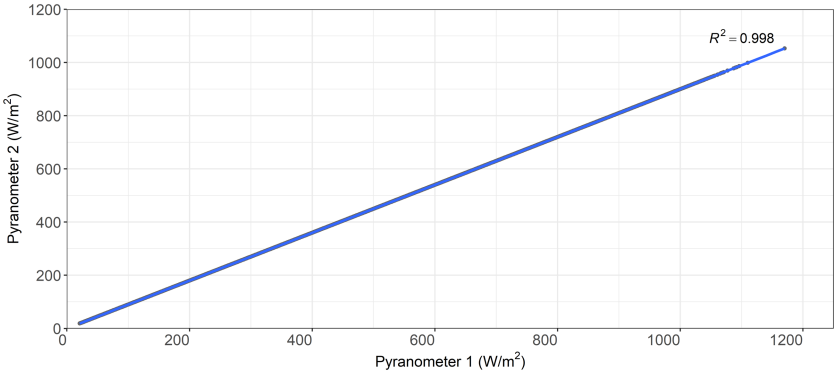

FIGURE A3 Comparison of the irradiance measurements obtained from two pyranometers installed nearby [Colour figure can be viewed at wileyonlinelibrary.com]

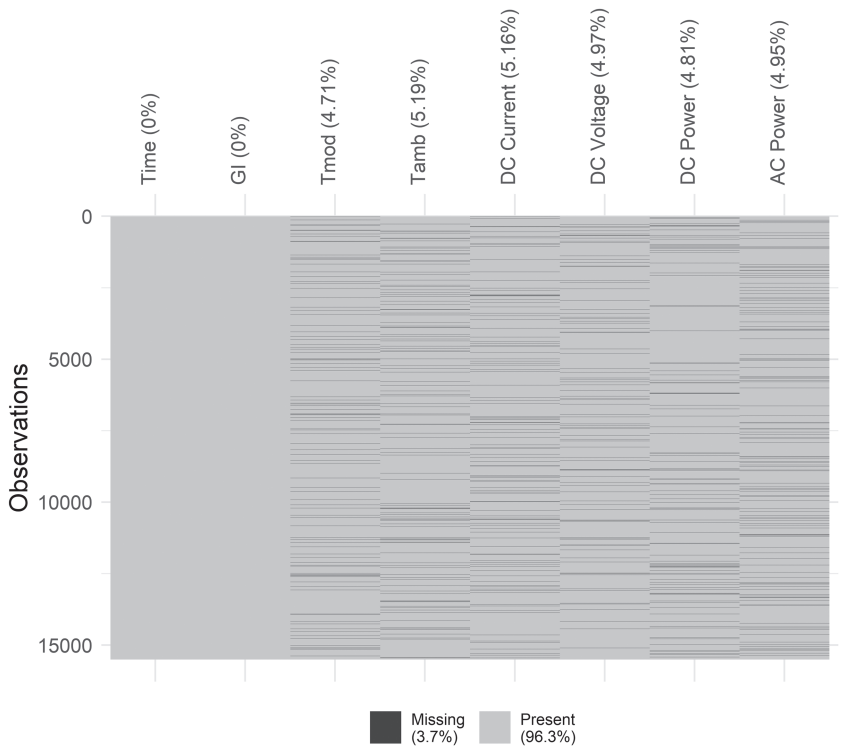

FIGURE A4 Portion of missingness for the recorded measurements in the tested dataset. The black color indicates missingness while grey represents available measurements

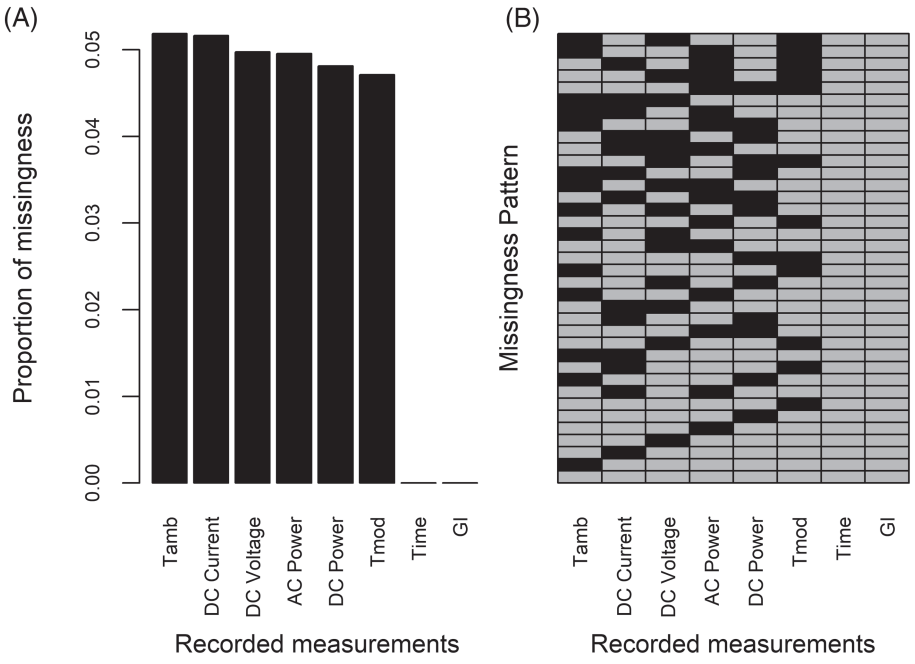

FIGURE A5 Amount of missing values and missingness pattern in the tested dataset. The black color indicates missingness while grey represents available measurements

TABLE A1 Data statistics validity of the reconstructed dataset

Variable	Time	G_I	T_{mod}	T_{amb}	I_A	V_A	P_A	P_{out}
Min	01-06-2016 05:15	21.0	0.0	0.7	0.0	142.3	3.5	0.0
First quartile	—	183.0	24.9	16.7	0. 8	179.3	146.1	126.0
Median	—	506.0	35.6	22.3	2.5	185.0	468.8	437.0
Mean	—	485.6	35.5	22.9	2.4	183.9	447.9	423.5
Third quartile	—	768.0	46.3	29.8	3.9	190.0	731.5	686.0
Max	31-05-2017 18:15	1170.0	67.0	40.0	5.8	237.0	1064.8	1265.0
NAs	0	0	0	0	0	0	0	0

Abbreviations: G_I, in-plane irradiance; I_A, array current; NA, not available; P_A, array power (DC); P_{out}, AC output power; T_{amb}, ambient air temperature; T_{mod}, module temperature; V_A, array voltage.

Table A1 summarizes the final data statistics of the reconstructed dataset and provides the mean, median, 25th percentile (first quartile) and 75th percentile (third quartile), minimum (min) and maximum (max) values and the amount of NA values per variable. There are no missing data points in the reconstructed dataset and the minimum and maximum values per recorded measurement are within the physical limits (predefined set threshold ranges).