energies

MDPI

Article

Outdoor PV System Monitoring—Input Data Quality, Data Imputation and Filtering Approaches

Sascha Lindig [1,2,*], Atse Louwen [1], David Moser [1] and Marko Topic [2]

[1] Institute for Renewable Energy, EURAC Research, Viale Druso 1, 39100 Bolzano, Italy; atse.louwen@eurac.edu (A.L.); david.moser@eurac.edu (D.M.)
[2] Faculty of Engineering, University of Ljubljana, Trzaska Cesta 25, 1000 Ljubljana, Slovenia; marko.topic@fe.uni-lj.si
* Correspondence: sascha.lindig@eurac.edu; Tel.: +39-0471-055-715

Received: 27 August 2020; Accepted: 23 September 2020; Published: 30 September 2020

Abstract: Photovoltaic monitoring data are the primary source for studying photovoltaic plant behavior. In particular, performance loss and remaining-useful-lifetime calculations rely on trustful input data. Furthermore, a regular stream of high quality is the basis for pro-active operation and management activities which ensure a smooth operation of PV plants. The raw data under investigation are electrical measurements and usually meteorological data such as in-plane irradiance and temperature. Usually, performance analyses follow a strict pattern of checking input data quality followed by the application of appropriate filter, choosing a key performance indicator and the application of certain methodologies to receive a final result. In this context, this paper focuses on four main objectives. We present common photovoltaics monitoring data quality issues, provide visual guidelines on how to detect and evaluate these, provide new data imputation approaches, and discuss common filtering approaches. Data imputation techniques for module temperature and irradiance data are discussed and compared to classical approaches. This work is intended to be a soft introduction into PV monitoring data analysis discussing best practices and issues an analyst might face. It was seen that if a sufficient amount of training data is available, multivariate adaptive regression splines yields good results for module temperature imputation while histogram-based gradient boosting regression outperforms classical approaches for in-plane irradiance transposition. Based on tested filtering procedures, it is believed that standards should be developed including relatively low irradiance thresholds together with strict power-irradiance pair filters.

Keywords: photovoltaics; photovoltaic system performance; photovoltaic system data; data quality; data imputation; data filtering

1. Introduction

With the transition from being a niche energy source to becoming mainstream, photovoltaics (PV) have to compete from not only an economic but also a reliability point of view with established energy production techniques. Reliable system operation addresses the whole life cycle of PV systems from the production of high-quality solar cell material to the development and construction of solar plants until decommissioning and recycling of the plant and its individual components. Therefore, the by far longest phase in the life cycle is the operation and maintenance (O&M) part, which ensures smooth day-to-day system operation. O&M activities as well as related PV performance studies rely on high-quality PV data measurements. Modern system performance checks, which are based on measurements of the electrical parameter of a system, often in combination with weather-related measurements such as temperature and irradiance, are sometimes constantly carried out. Although this

is becoming mainstream for large scale PV plants, the performance of most installed PV systems is evaluated either in certain time intervals or just in case of detected production issues. In the meantime data are often recorded without quality checks. The study *Analytical Monitoring of Grid-connected Photovoltaic Systems* [1] of the IEAs PVPS (International Energy Agency's Photovoltaic Power Systems Programme) Task 13 provides a broad overview of explaining and visualizing field measured performance data. Van Sark et al. [2] summarize key aspects of PV system monitoring and characterization. 3E provides with their sensor check data service advanced diagnostic checks for solar radiation sensors [3]. Therefore, they rely on satellite-based irradiance data for comparison. Established research institutions, focusing on statistical analyses of monitoring data, have developed automatized algorithms to detect faulty and missing data in PV monitoring datasets. Well known examples are Pecos [4] from Sandia National Laboratories or a technical report by NREL [5]. Another example was presented by Killinger et al. [6], where they use clear sky and PV power models to identify faulty power output data. A comprehensive overview of PV system monitoring and fault detection approaches is given by Livera et al. [7], with clear guidelines on the required measurement parameters and their maximum uncertainties. Most of these studies discuss to some extent discuss the requirements in terms of measured performance and weather parameters. Unfortunately, however, all the listed approaches fail to provide guidelines of what to do with filtered data nor agree on a common filtering approach. A study by Koubli et al. [8] proposed back-filling algorithms of missing electrical and meteorological data using an electrical model [9] and synthetic climate data. Statistical or machine learning approaches, possibly from other scientific fields, might open up new possibilities for accurate data imputation.

In this work, we want to provide an overview of occurring problems when working on PV monitoring data, give useful examples of what faulty data look like, provide solutions on how to fill larger gaps of missing measured temperature and irradiance data and discuss the necessity of applying specific filter for statistical performance analyses. Monitoring data are commonly used and aggregated to provide information on the performance of the PV system in question. Performance evaluation studies can range from a simple yield comparison between inverters in the same PV field, to year-to-year comparison of yield measurements over performance loss rate calculations, where the yearly decline of PV system performance is quantified, to yield estimations or remaining-useful-lifetime (RUL) studies. The RUL of a PV system is a date at which a pre-defined power output cannot be reached anymore, and the system reached the economic end of its lifetime.

All these performance evaluation studies have a similar structure including input data treatment and data filtering. These points are discussed in greater detail in this work. Therefore, the following topics are covered. After introducing an experimental PV plant whose monitoring data are used for data visualization, Section 3 provides simple quality check of monitoring data based on visualized measurement data. First, guidelines of monitoring data acquisition according to standard the first part of IEC 61724:2017 [10] are presented. After, module temperature data are discussed together with common module temperature models and data filling techniques based on statistical models. In Section 3.2, the complexity of in-plane irradiance is introduced, and data imputation approaches are discussed. Afterwards, proposed visualized quality checks for PV power, in-plane irradiance and key performance indicators (KPI) are presented. Therefore, different categories of faulty or missing data are defined. In order to get reliable results of these types of studies high-quality input data are required and tailored data filter should be applied. In Section 4, general filtering approaches for different data evaluation studies are discussed.

2. Experimental PV System

In this work, monitoring data quality tests will be performed on a PV system, which is installed and operated at the Bolzano airport (ABD). This system is well suited for the study as it is an experimental PV installation under close surveillance with high-quality data. Longitude and latitude of the system are 46.4625° N and 11.3299° E respectively and the ground has an elevation of 240 m above

sea-level. The system was installed in 2010, and is part of a larger experimental PV plant. According to the Köppen–Geiger classification, the climate in Bolzano is categorized as a temperate climate with warm summers and without dry seasons (Cfb) [11].

The system consists of 16 mono-crystalline (mc-Si) modules connected in series, and has an installed capacity of 1.98 kW. The system is ground mounted with a fixed tilt of 30° and an orientation of 8.5° west of south. Additionally, a weather station is installed in close proximity to the test site. Here, the ambient temperature, various irradiance parameters such as the in-plane irradiance (G_{POA}) and the wind speed are recorded. On the rear side of each system the module temperature is measured. The sensors are systematically cleaned and periodically calibrated to comply with the first part of standard IEC 61724:2017 [10]. The in-plane irradiance is recorded with a pyranometer of the model Kipp and Zonen cmp11 with an estimated uncertainty of below 2%.

The weather data are recorded with a measurement frequency of one minute. Since the electrical parameters are measured at a resolution of 15 min, all values are averaged to the same time interval. A period of nine years is evaluated ranging from February 2011 until February 2020. The period of observation is not equal to the operation time. The systems began operating in August 2010, so roughly six months before the observation time starts. The delayed start of observation was set due to a delayed start of recording weather data.

3. Monitoring Data Acquisition

Monitoring data is usually provided from two sources. These are electrical parameters from the PV plant, e.g., the power in the maximum power point (P_{mpp}), and climate related data from a nearby weather station or satellite data. The first part of standard IEC 61724:2017 [10], *Photovoltaic system performance—monitoring*, covers the guidelines on how to correctly monitor PV plants. Therefore, monitoring systems are categorized into three classes based on the selection of measured variables and types of carried out performance assessments. The choice of monitoring system is usually dependent on the PV system size.

Table 1 presents the mandatory measured parameter per monitoring class including stated uncertainties according to part one of standard IEC 61724:2017. Here, Class A monitoring corresponds to the highest level of data monitoring and Class C to the lowest.

In this work, we focus on irradiance, temperature and power measurements. These are the most commonly measured and used data in performance evaluation studies of PV systems. In-plane irradiance G_{POA} is measured in the same plane of the PV modules. Irradiance for PV applications is measured either with thermopile pyranometers or photovoltaic reference devices. The usage of these devices, calibration interval and guidelines as well as possible measurement corrections are stated in standard IEC 60904:2015 [12] and in IEC 61724:2017 [10]. If on-site measurements are difficult to realize, irradiance datasets can be acquired via clear-sky modeling or satellite-derived data. Usually, such measurements are subject to higher uncertainties and do often deviate considerably from ground measurements. Another problem when using satellite data is data consistency. Although satellite data quality consistently improves the data retrieved in the past might have a different accuracy from data used today.

Temperature data are measured with temperature sensors, usually thermo-couple or resistance-based sensors such as a Pt100 (Pt for platinum). Ambient air temperature sensors must be well ventilated and shielded from solar radiation. PV module temperature sensors are attached at the back of the module. Here, it is very important to ensure a good adhesion between the sensor and the module to provide accurate readings. Thermally conductive adhesive should be used appropriately for prolonged outdoor usage. Measurement uncertainties are supposed to be below 2 °C. The number of sensors per PV plant is dependent on the plant size and is defined in part one of standard IEC 61724:2017 [10].

Table 1. Measured monitoring parameter divided by meteorological and electrical parameter [10].

Parameter	Class A	Class B	Class C
Irradiance parameter			
In-plane irradiance	x	x	x
Global horizontal irradiance	x	x	
Direct normal irradiance			
Diffuse irradiance			
Max. uncertainties	3%	8%	Any
Environmental parameter			
PV module temperature	x	x	
Ambient air temperature	x	x	x
Wind speed	x	x	
Wind direction	x		
Rainfall	x	x	
Humidity	x	x	
Electrical parameter			
DC array voltage & current	x		
DC array power	x		
AC array voltage & current	x	x	
AC array power	x	x	x
Output energy	x	x	x
Output power factor	x	x	x
Max. uncertainties	2%	3%	Any

Figure 1 shows the necessary steps of monitoring data preparation for any kind of performance evaluation. These include reading of monitoring data, possible cases of corrupted data and commonly used parameters which are subject to filtering. Therefore, the second and third part are closely related with one another as data filtering is used to address some of the common data issues. Commonly found monitoring data sources are listed in the figure as well as in Table 1.

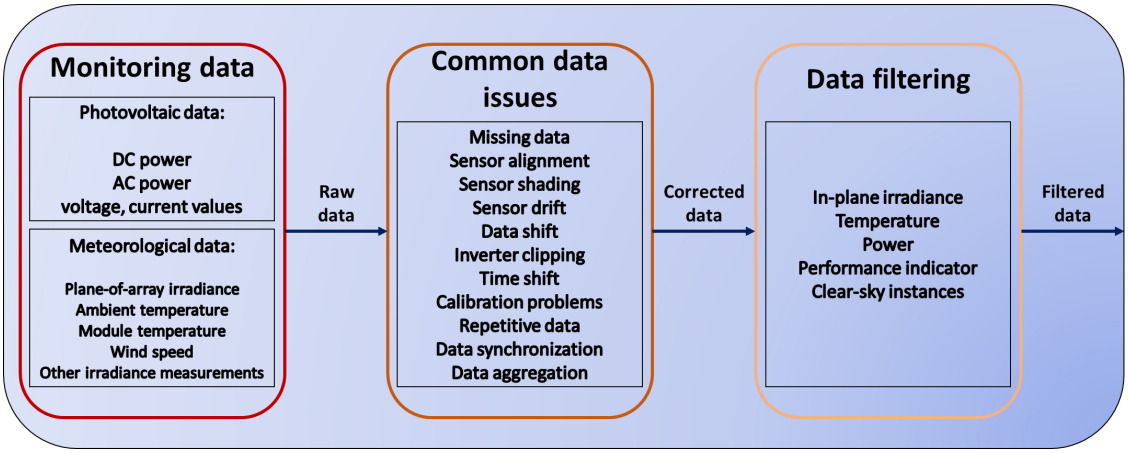

Figure 1. General steps of monitoring data preparation.

The first task of performance evaluation studies is to collect available data and perform a first data quality check. Very important data issues, which are not often discussed, are data synchronization and data aggregation. It is visible in Figure 1 that data usually come from different sources. A check of the timestamp synchronization is vital to ensure the calculation of high-quality KPIs. The knowledge of the type of data aggregation is thereby also an important information. Usually data are collected periodically (e.g., 1 min, 15 min). In order to provide consistent data, the way of recording high-resolution values is crucial, i.e., are they averaged over the collection period or is it an instantaneous value. Plotting the time series of single day observations helps to verify if data synchronization and aggregation is guaranteed.

Standardized quality checks include the deletion of invalid readings and treatment of missing data [10]. Therefore, it is recommended to identify faulty data entries and to apply realistic thresholds as well as statistical outlier tests. In the case of missing data, it is recommended to assess whether filling (imputation) of missing data is reasonably possible, and what kind of approach needs to be used. Many different approaches for data imputation exist, including using different types of interpolation, Kalman filtering, auto-regression or moving averages [13,14]. For high-resolution data (minutely or hourly) and short gaps, interpolation is a reasonable approach, but with larger gaps or lower resolution data (daily, weekly) other approaches yield better results [13]. Depending on the availability of other measured parameters (for instance satellite-based irradiance measurements or peered irradiance sensors in different locations), multi- or univariate regression or machine learning models can also be applied [14,15]. In the PV community there does not seem to be a consensus of how data filling should be performed and will often depend on the amount of data to be filled and the size of data gaps. In all situations where data has been imputed, it is recommended to label filled values to remain identifiable and to document the imputation approach applied.

In the process of data quality checking and data correction, temperature should be considered separately from irradiance and power data. Thresholds and outlier detection using statistical tests, which will be presented in the following section, are important to improve the quality of a dataset but not always sufficient to detect faulty measurements. By analyzing the time series of the raw data, significant measurement errors can already be identified. They can either be connected to faulty readings or stem from problems during measurement acquisition. Common problems are for example the shadowing of irradiance sensors by an object for a certain amount of time or the detachment of module temperature sensors from the module. Such events can easily be identified when visualizing the data at hand. If small parts of the datasets are affected data imputation can be used to recover the faulty/missing data. In case of longer data outages other sources should be used to retrieve the data, for example satellite data for irradiance or temperature models for the module temperature. In the following, temperature data on one hand, and irradiance and PV power data on the other hand are discussed with application examples in terms of data imputation. The examples stem from issues we were facing while analyzing the data for further treatment.

3.1. Temperature Data

The module temperature T_{mod} is a function of several solar irradiance related parameters such as ambient temperature, wind speed and direction, mounting configuration, thermal behavior and efficiency of the module, and system level parameters such as soiling or shading conditions. Usually, plotting ambient and module temperature over time provides a fairly good estimation of the measurement quality. If multiple module temperature sensors are available an inter-comparison is suggested. With such figures, strong outliers are easy to detect and can be taken care of. If module temperature readings are showing unexpected trends, module temperature models could be applied, and the measured values be compared to modeled values. If the discrepancy between both is too high, measured values should be replaced with modeled ones. The choice of the model will depend on the

availability of other climate data from that specific side. The simplest model hereby is the Nominal Operating Cell Temperature (*NOCT* Model) equation [16]:

$$T_{\mathrm{mod}}[^{\circ}\mathrm{C}] = T_{\mathrm{amb}} + \frac{G_{\mathrm{POA}}}{800\,\mathrm{W/m^2}}(NOCT - 20\,^{\circ}\mathrm{C}). \quad (1)$$

NOCT is the normal operating cell temperature and is determined for a 45° south-facing module with incident irradiance of 800 W/m^2, an ambient temperature of 20 °C and a wind speed of 1 m/s. G_{POA} is the measured in-plane irradiance and T_{amb} the measured ambient temperature. The *NOCT* variable is mostly provided in the datasheet of the respective model. The inclusion of wind speed, if available, usually improves the accuracy of the module temperature estimation. A well behaving model is the Sandia module temperature model (SMTM) [17]:

$$T_{\mathrm{mod}}[^{\circ}\mathrm{C}] = G_{\mathrm{POA}}(e^{a+b\cdot WS}) + T_{\mathrm{amb}} \quad (2)$$

WS is the measured wind speed and *a* and *b* empirical parameters depending on mounting configuration, module backside material as well as solar cell material. More sophisticated models including more empirical coefficients, the transmittance of the module cover and the absorption coefficient of the cell were formulated by Skoplaki et al. [18] and Mattei et al. [19]. The choice of the model, and therefore the accuracy of the modeled temperature, will always depend on the available input parameters.

Plotting module temperature over time already gives an impression whether the measurements are realistic. The following figure shows ambient and module temperature of the PV plant under investigation. When looking at Figure 2 it is obvious that the module temperature readings are faulty at the beginning of recording. For roughly one year the values are very similar to the ones of the ambient temperature sensor. The module temperature sensor was detached from the module because of the usage of unsuitable tape and glue. After reattaching using adequate adhesion material in April 2012, the readings are stable throughout the time of observation.

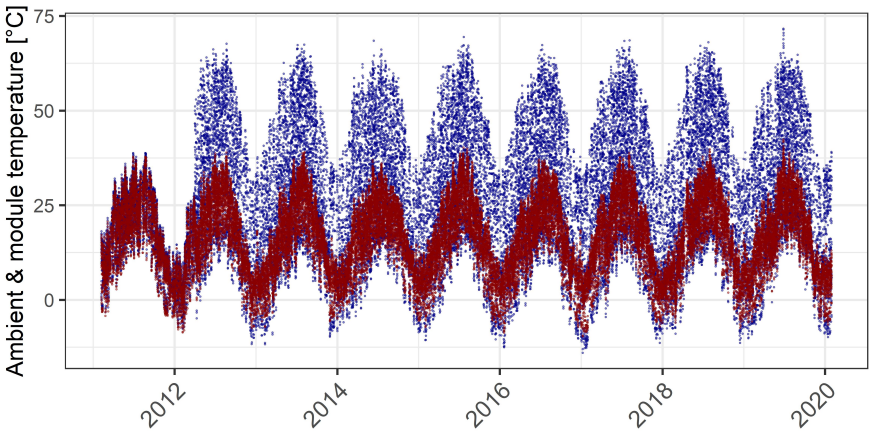

Figure 2. Original back of the module temperature measurements of mc-Si PV system; red—ambient temperature; blue—module temperature.

Several regression models have been tested to replace the faulty data with modeled module temperature values. Two of them, namely multivariate regression and multivariate adaptive regression splines (MARS), are discussed further as the better performing models. Therefore, the data were subject to light outlier filters according to the third part of IEC 61724:2016 [20], which are listed in Section 4. The initial motivation for using a regression model was to have a simple model which does not require any metadata of the PV system. Regression models rely only on measured data.

As explained before, to use the SMTM one must know the kind of mounting type and which backside material is used in the modules.

The faulty data in Figure 2 account for 11% of the overall dataset. Therefore, the strongly correlated values of ambient temperature, in-plane irradiance and wind speed were used to model T_{mod}. The remaining dataset was used to test the regression models. 20% of the remaining, trustful, data were used as test set and 80% of the data as training data. In order to rate the regression models, this and the two established models, *NOCT* and SMTM, haven been tested on the test set and the results yield the model parameter presented in Table 2.

The difference between the modeled module temperature using MARS regression and the correctly measured temperature can be seen in Figure 3. The relationship between measured and modeled data should be nearly linear. R^2 between model and measured values is 0.92 and the root-mean-square-error (RMSE) is 4.26 °C. These values show that a trained regression model is performing slightly better compared to the SMTM, provided that a sufficient amount of the measured data is trustful and can be used as training data. Furthermore, the modeling results in Table 2 show that the SMTM yields more accurate results compared to the *NOCT* model and is always preferable if wind speed measurements and metadata are available.

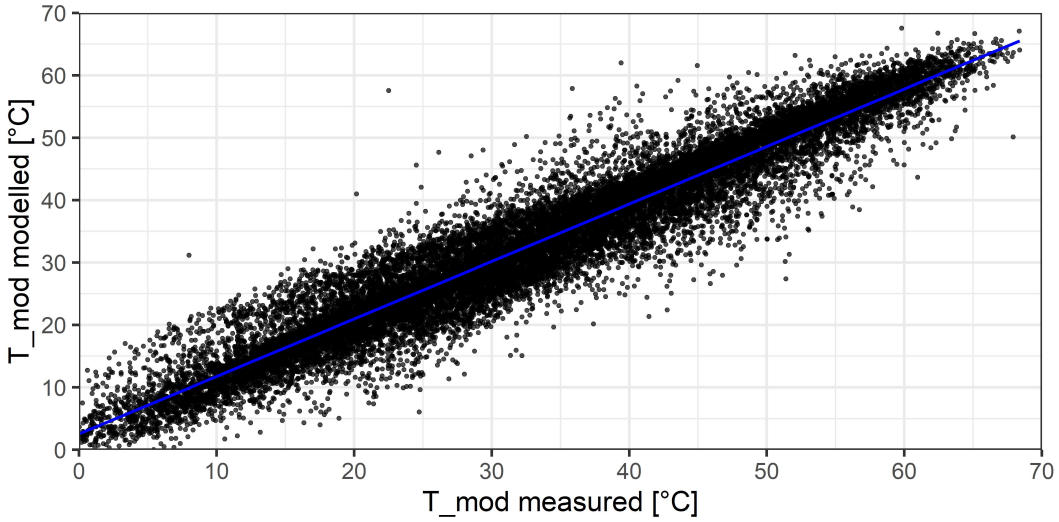

Figure 3. Measured versus modeled (MARS regression model) module temperature of mc-Si PV system.

Table 2. Modeling parameter R^2 and RMSE of module temperature data replacement.

Model	R^2	RMSE
Multivariate regression	0.91	4.60 °C
MARS	0.92	4.26 °C
NOCT Model	0.89	5.15 °C
SMTM	0.91	4.61 °C

3.2. In-Plane Irradiance Imputation

Irradiance determination is a very complex topic. In the best case, the in-plane irradiance (G_{POA}) is measured with an irradiance measurement device installed in the same plane as the investigated PV system. If no in-plane irradiance sensor is installed, a transposition of the G_{POA} from global horizontal irradiance *GHI* must be calculated. Therefore, irradiance data can be categorized into different accuracy classes based on data availability [21]:

(a) High accuracy: G_{POA} is measured on-site

(b) Medium accuracy: horizontal irradiance GHI is measured on-site and G_{POA} is estimated using decomposition and transposition approaches

(c) Low accuracy: G_{POA} is estimated using decomposition and transposition approaches from extracted GHI, which is taken from one of the following sources: interpolated (weighted regression) using peered data of different weather stations in relatively close proximity to the test side, satellite or re-analysis-based datasets, clear-sky modeled datasets

The order of accuracy corresponds to increasing uncertainties in the datasets. Although measured irradiance values can have uncertainties as low as 2% [22], the introduction of decomposition and transposition approaches as well as the estimation of GHI introduces additional, partially very high, uncertainties.

In general, ground measurements are always preferred because of higher accuracy, both for in-plane and horizontal irradiance. If weather stations are used the spatial resolution might be not high enough leading to high uncertainties. Datasets from satellite data might be inaccurate because of their low spatial (and possibly temporal) resolution and treatment of clouds, snow or aerosol. Many different clear-sky models are available. In the simplest case, they are based on geometrical calculations. More advanced clear-sky models take into account different measurable atmospheric parameters such as ozone, aerosols and precipitable water. The problem is that these data must be measured and provided as model inputs.

To get useful results, the best of these options must be selected for each case and carefully evaluated.

GHI is the sum of diffuse and direct irradiance and is defined by:

$$GHI[\mathrm{W/m^2}] = DHI + DNIcos(\theta_Z). \tag{3}$$

Here, DHI is the diffuse share of horizontal irradiance, which comes from all directions and DNI is the direct normal irradiance. θ_Z is the solar zenith angle.

G_{POA} depends on several factors such as the sun position, orientation of the system, individual irradiance components, albedo and shading. It can be expressed as the sum of the in-plane beam component of irradiance $G_{\mathrm{b_{POA}}}$ and the in-plane diffuse irradiance components, which include an in-plane ground-reflected component $G_{\mathrm{g_{POA}}}$ and a sky-diffuse component in the plane of array $G_{\mathrm{d_{POA}}}$:

$$G_{\mathrm{POA}}[\mathrm{W/m^2}] = G_{\mathrm{b_{POA}}} + +G_{\mathrm{g_{POA}}} + G_{\mathrm{d_{POA}}}. \tag{4}$$

If no measured in-plane irradiance is available, the individual components are calculated from GHI, provided through one of the scenarios mentioned above. The separation of the individual irradiance parts is necessary because the diffuse irradiance component is very complex to model. A comprehensive discussion on models can be found in [23] and a comparison and rating of different models in [24,25]. Figure 4 shows a simplified model structure to calculate the in-plane irradiance [26].

If instead G_{POA} sensors are available at the side, they should be used to ensure the highest possible data accuracy. Usually, the readings are much more precise and accurate. To ensure smooth and reliable operation of the sensors, they must be systematically cleaned and calibrated in order to comply with part one of standard IEC 61724:2017 [10]. Possible problems are measurement errors, sensor alignment issues or sensor drifts and are discussed in greater detail in the next section.

As discussed before, faulty or missing measurements for a limited amount of time could be replaced by data imputation. However, if a measurement is faulty or missing for longer periods of time, filling these values with simple approaches such as interpolation are no longer valid or possible.

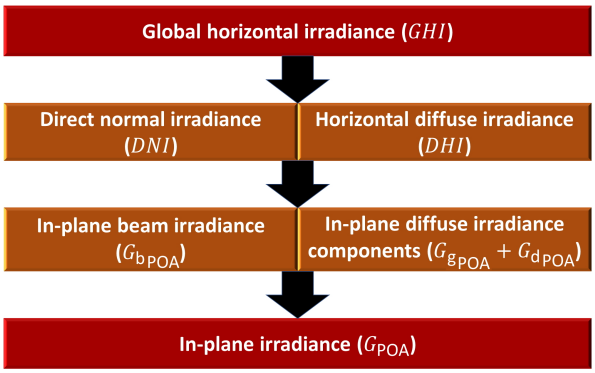

Figure 4. Simplified modeling steps from global horizontal to in-plane irradiance.

In a recent study, we encountered such an issue when examining a dataset [27] needed for PV performance loss calculations. We found that our input dataset was missing in-plane irradiance measurements for a period of four years at the beginning of the dataset, and had smaller gaps of hours to days in other years. Aside from in-plane irradiance, the dataset also contained other irradiance measurements (global horizontal irradiance GHI and diffuse horizontal irradiance DHI) and measurements of other parameters such as relative humidity RH. These other measurements were not missing for the first four years of the dataset.

The aim of the study was to fill the missing G_{POA} measurements using the available GHI, DHI and RH measurements. To replace the missing data, we compared several classical irradiance transposition models (implemented in the python software package PVLIB [28]) with several machine learning-based models [29]. We compared the isotropic [30], Klucher [31], Hay–Davies [32], Reindl [33,34], King (As discussed by the authors of PVLIB, the King model is not well documented nor is there published documentation [28]) and Perez [35] classical models, and random forest [36], extra trees [37], gradient boosting [38] and histogram-based gradient boosting (The implementation of histogram-based gradient boosting in scikit-learn is based on the LightGBM framework [39]) machine learning regression models as implemented in the python library scikit-learn [29]. We added solar position parameters (solar zenith, solar azimuth and solar elevation) to our input dataset of GHI, DHI and RH, and removed all measurements of solar elevation $\leq 0°$. Using a random subsample (n = 50,000) of our complete training dataset (n = 275,000), we performed hyperparameter optimization to determine optimal values for the modeling parameters for both the classical transposition models as well as the machine learning models. The models were subsequently run (transposition models) and trained and run (machine learning models) using the full training dataset, and cross-validated using a 0.75/0.25 train-test split. Although testing all considered methods for modeling estimation G_{POA}, we found that the machine learning-based models clearly outperformed the transposition-based models, with average RMSEs of around 30 W/m^2 and 70 W/m^2. The highest accuracy was found for the histogram-based gradient boosting regressor (RSME of 29.8 W/m^2). Using this regressor, we estimated and filled the missing G_{POA} values.

In summary, ground measurements are always preferred if available, possibly with necessary corrections. For certain locations, other methods may yield good results. Retrieving accurate irradiance data of locations with complicated shading conditions and a high amount of diffuse light due to regular fog, mist or cloud cover, is still an open issue.

3.3. In-Plane Irradiance and Power

A thorough check of power and in-plane irradiance raw data is absolutely necessary to perform any kind of data analysis. The quality of the datasets depends on various factors and can be compromised for many reasons. In this section, simple visual quality checks and examples of common data issues are presented. Therefore, Figure 5 is based on the measured data of the experimental PV

system introduced in Section 2, while Figure 6 presents data from various other plants installed across the globe which are kept anonymous.

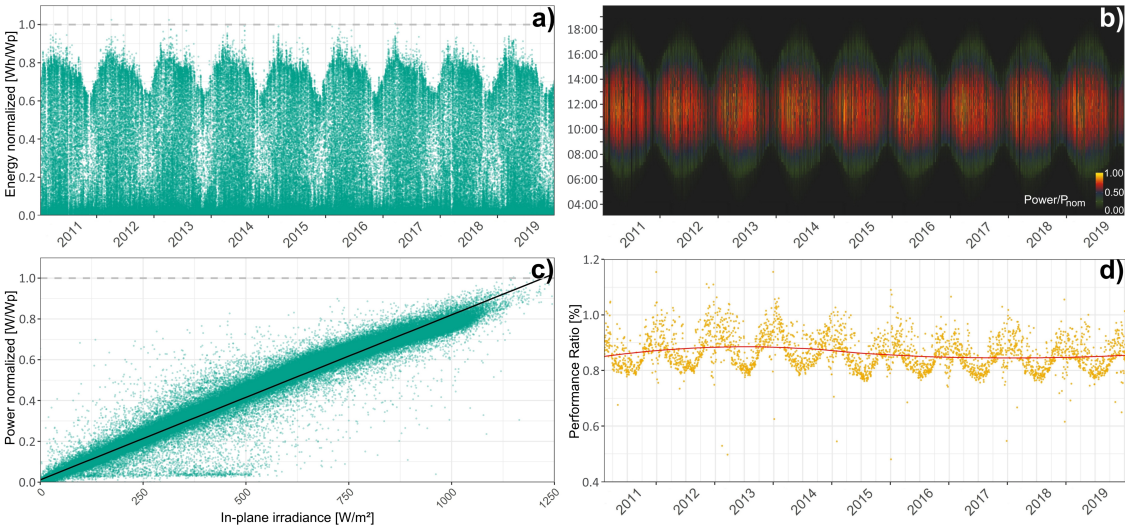

Figure 5. Data quality check figures for mc-Si PV system: (**a**) Normalized 1h energy values vs. time; (**b**) Normalized power-density plot—time of the day vs. day of the year; (**c**) Normalized 15 min power vs. in-plane irradiance; (**d**) Daily Performance Ratio vs. time.

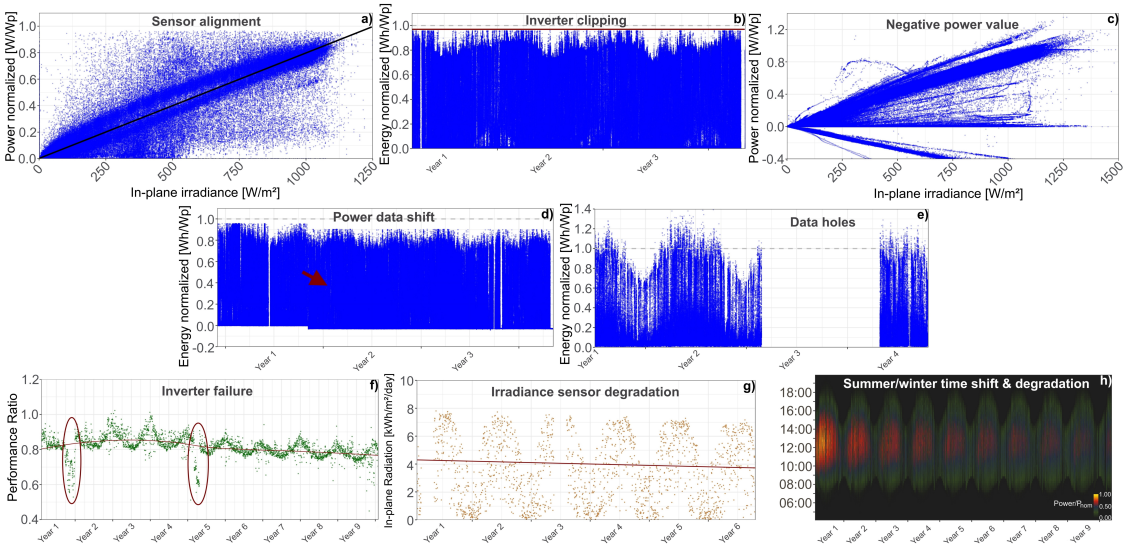

Figure 6. Exemplary data quality issues: (**a**) Imprecise irradiance sensor alignment (P vs. G_{POA}); (**b**) Inverter clipping (E vs. time); (**c**) Negative power values (P vs. G_{POA}); (**d**) Power data shift (E vs. time); (**e**) Data hole (E vs. time); (**f**) Inverter failure (daily PR vs. time); (**g**) Irradiance sensor degradation (daily in-plane radiation vs. time); (**h**) Summer/Winter time shift and strong degradation (Normalized power-density plot).

Since irradiance and power time series are supposed to behave in a similar fashion and are directly proportional over a large irradiance interval, similar checks can be performed. Figure 5 shows recommended visualizations for a basic data quality check.

The following relations are depicted for the example PV plant: (a) energy vs. time, (b) power-density plot, (c) power vs. in-plane irradiance and (d) daily performance ratio (PR) over time. All measured data can be evaluated in a similar fashion. The Performance Ratio describes

the relation between incoming irradiation and power produced by the PV system. It is a unit-less parameter and is calculated using the following formula [10]:

$$PR = \frac{Y_f}{Y_{ref}} = \frac{E_{AC}/P_{nom}}{H_{POA}/G_{STC}}, \quad (5)$$

$$PR_{DC} = \frac{Y_a}{Y_{ref}} = \frac{E_{DC}/P_{nom}}{H_{POA}/G_{STC}} \quad (6)$$

The first equation holds for AC power and the second for DC power. For AC, the final yield Y_f is divided by the reference yield Y_{ref}. If DC power is evaluated, Y_f is replaced by the array yield Y_a. Yield values are given in [kWh/m^2]. The yields itself are ratios of normalized energy E and normalized in-plane irradiation H_{POA}. Energy values are aggregated power measurements while irradiation values are aggregated irradiance instances (from W to Wh). The energy E is normalized by the nominal power of the respective system P_{nom} obtained under Standard Test Conditions [40] (ambient temperature T_{STC} = 25 °C, in-plane irradiance G_{STC} = 1000 W/m^2, air mass AM 1.5) and H_{POA} by STC irradiance G_{STC}. In Figure 5, all values have been normalized to STC. By doing this, an easier inter-comparison between different PV systems can be performed.

By examining the figures, it is apparent that this particular plant operates without any profound problems. The only visible issue is a very slight power loss over time, especially in the last two to three years, which can be seen in Figure 5a,d. Time-dependent system performance degradation has to be expected, and is not necessarily an issue, as long as the degradation is within acceptable margins. If strong system degradation is detected, it is recommended to evaluate the performance of individual modules to trace back the root causes of the observed degradation. Aside from detecting issues with system performance, Figure 5c,d are valuable to evaluate the alignment of the irradiance sensor in plane with the PV system and to rate the irradiance measurement quality. As said before, power and irradiance are nearly proportional. Therefore, high-quality data are characterized by a linear relationship, visualized in Figure 5c. Here, instantaneous measurement data are depicted. A higher number of outliers might suggest certain synchronization issues. In Figure 5c, some outlying values are visible where irradiance values of up to 500 W/m^2 are measured while no power is produced. That is because the system is installed in a valley of a mountainous region. Under low sun inclination in the morning, the irradiance sensor is already irradiated while the PV system is still in the shadow, and thereby not producing any power, while higher irradiance values are correctly measured. Figure 5d could also depict other aggregation time steps (daily/weekly/monthly sums/averages) for PR, power or irradiance values. Like this, time-dependent trends could be evaluated, which are assumed to be periodical and without strong degradation patterns.

Another helpful way to verify data quality is to look at a heatmap plot of instantaneous measurement values, as shown in Figure 5b for normalized power. Here, 15 min power data of the mc-Si system are colored according to normalized power and plotted as a function of the time of the day and the day of the year. Higher power values for a longer duration of the day are detected in summer because of higher irradiation and longer days in summertime. From such a plot one can detect longer system outages, timestamp issues, shading instances or also exceptionally strong degradation. If similar plots are shown on a weekly or monthly scale, additionally cloudy days can be detected. The data of the system are converted to UTC (Universal Time Coordinated) to remove daylight-saving time shifts in the density plot. It is visible that the data follow a fairly stable pattern and can be considered being of high quality.

In general, strong outliers, missing data, sensor measurement issues, inverter clipping instances or other common data issues can be identified from the plots above. In Figure 6, examples of such data issues are presented from several PV systems not belonging to the experimental PV installation introduced above. All data are anonymized and normalized.

In the figures, blue has been used to represent power values plotted on the y-axis, orange for irradiance values and green for PR values. In each subplot, the corresponding problem is depicted.

Figure 6a shows a typical sensor alignment issue. Many data points are away from the linear trend-line between power and irradiance. Furthermore, two distinct lines can be recognized. These issues can stem from various problems, Usually they are connected to an offset between timestamps, an incident which decreases the power output at some point, a change in irradiance readings or the sensor with a different tilt/orientation as the solar modules. In such situations it is recommended to investigate the power and irradiance data over smaller time scales to possibly detect certain performance impairing issues. Furthermore, the irradiance data must be thoroughly filtered or possibly replaced.

Figure 6b shows inverter clipping of a PV system, which occurs when their AC power rating is lower than the total installed PV module capacity and the output power is limited. This is actually not an error or issue, but a way to increase the reliability of PV systems. Nevertheless, it is important to be detected and taken into account for further data treatment. Using undersized inverters has the benefits of saving money for cheaper components, producing more power under low light conditions and the fact that PV systems degrade naturally over time (a high rated inverter power might not be needed anymore). Thus, inverter clipping is common practice in modern PV plants.

In Figure 6c, the system under investigation apparently produced negative power values. From a physical standpoint, that is not possible. It is more likely that the polarity has been switched. Filtering negative values is a common procedure defined in standard IEC 61724:2016 [20].

The system data, presented in Figure 6d, show inverter clipping and a power data shift after one year of operation. Since the source and reason of the shift is not known, it is advisable to omit the first year of operation to ensure realistic measurement conditions.

The data holes in Figure 6e stem from calibration activities of the sensors at the measurement site. In order to avoid these issues related to maintenance of the measurement system, it is recommended to have redundancy in sensors, and to have a proper calibration plan to avoid losing large amounts of measurements.

In Figure 6f, the daily aggregated PR of a PV system is seen. Two instances of inverter failures were recorded, marked with red ellipses. Inverter failures are preceded by distinct losses in performance. The performance drops and the PR values deviate clearly from their normal patterns. Either the inverter breaks down completely or the deviation is detected beforehand, and the inverter is repaired or exchanged. Inverter failures can be categorized as reversible performance losses.

Figure 6g depicts the daily in-plane radiation measured with an irradiance sensor. A simple approach to detect possible sensor drifts or sensor degradation (if solar cell material is used) is to perform a linear regression of the irradiance time series. A clear trend change of the regression line over time would indicate a possible drift in the sensor readings. Certain trend variations could also be explained by inter-annual irradiance variability (especially for shorter time series) and appearing global brightening effects in recent year. These effects are observed since the late 1980s and are attributed to reductions in aerosol content in the atmosphere and cloud cover leading to higher transmission of sunlight [41]. In this particular case a decrease in measured irradiance over time can be seen. If this irradiance sensor measurements were to be used for constructing PR or other KPI time-series, the PR would artificially increase, provided the corresponding power time series is fairly stable over time. After enquiring about this sensor, it was reported that it is an amorphous silicon reference cell. It is expected that the active solar material in the reference cell degraded, resulting in decreasing irradiance measurement values while not being re-calibrated. To evaluate the PV system performance corresponding to this irradiance sensor, it is necessary to use another source of irradiance measurement to ensure realistic readings.

Figure 6h shows a power heatmap for a thin-film PV system. The data are measured and plotted in central European time including summertime, therefore a 1-h shift in March and October every year can be seen. It is visible that this shift interrupts the periodic structure of the figure and should therefore be removed by converting the timestamp to UTC. Furthermore, this particular system is subject to an unusual high degradation, visible in the decreasing color intensity over time.

4. Data Filtering

Once the raw data are evaluated, data filters are used to provide stable measurement conditions and to extract the data of interest for specific applications. Therefore, the choice of filters will, among other things, depend on the findings based on performed quality checks. Standard IEC 61724:2016 part three provides guidelines of initial minimum filters for monitored high-resolution (15 min data) variables [20]. The most relevant ones are listed in Table 3 below:

Table 3. Recommended filter of part three from IEC 61724:2016 [20].

$-6\ W/m^2$	<	irradiance	<	$1500\ W/m^2$
$-30\ °C$	<	ambient temperature	<	$50\ °C$
$0\ m/s$	<	wind speed	<	$32\ m/s$
$-0.01 \times P_{nom}$	<	AC power	<	$1.02 \times P_{nom}$

Here, P_{nom} is the rated power of the system. The standard also suggests testing for inverter clipping, irradiance sensor shading, calibration drift and other malfunctions. Recommended visualized tests are presented above. Furthermore, standard IEC 61724:2017 part one [10] recommends an in-plane irradiance threshold of $20\ W/m^2$ ensuring measurements during daylight hours.

Next to the common usage of these standard filters, filtering will always depend on the purpose of the study. In the following, filters are discussed with respect to performance loss rate calculations, a field where filtering can have tremendous effects on the final outcome. No standards or guidelines are available which suggest best practices in terms of filtering and thus chosen filters mostly depend on the preferences and experiences of the individual research group. Often, filtering is performed without the required level of detail. Filtering approaches are partly selected in such a way to achieve desired results in later analyses and often not discussed properly. That is why some common thoughts about filtering are presented below.

Irradiance and power filters are already suggested by standard IEC 61724:2016 but are mostly extended. Furthermore, clear-sky or *PR* filters are often used to find representative power-irradiance pairs. In the following, filter approaches are discussed.

Irradiance threshold filter: Irradiance filters are usually deployed in form of maximum and minimum thresholds. They are intended to remove nighttime values and measurement errors. In some cases, narrow irradiance filters (e.g., $800\ W/m^2$–$1100\ W/m^2$) are used to limit the measured data to narrow bands with similar conditions compared to STC or NOCT. The downside of such approaches is that a large amount of the data are removed, and it is questionable if the remaining data are representative for the overall system performance under real operating conditions.

Power threshold filter: Power filters are first and foremost used to remove power measurement errors and system outages, since the extraction of data of interest is usually performed with an irradiance filter. The thresholds for the filter are not universally applicable compared to irradiance thresholds because the power output varies among systems. A simple way to account for that is to filter in relation to the maximum rated power.

Statistical Performance Metric filter: Power (*P*) and Performance Ratio (*PR*) are the most common performance metrics used. The *PR* describes the direct relation between irradiance and power and is a quality indicator for the performance of PV systems. It is very well suited to detect inconsistencies between power and irradiance measurements. Statistical Performance Metric filters are used to ensure that power-irradiance pairs match and remove sensor shading instances, measurements with strong soiling or other reasons which cause one-sided measurement deviations. In order to detect and remove such deviations, statistical approaches are used. Therefore, filter intervals are created and applied based on statistical averaging.

Clear Sky filter: Clear-sky filters are approaches which filter measured data for clear-sky instances and remove data periods where cloud cover or partial shading (to a certain degree) was prevalent. Clear-sky measurements have the advantage of being consistent throughout the time of observation

and are therefore well comparable. That is why measured clear-sky instances are preferred to be used to be compared with modeled irradiance, to evaluate the irradiance measurement quality, but also to rate different irradiance sources. Clear-sky models are developed and deployed in the PV performance libraries PVLIB [42] and RdTools [43], which are available for Python.

Inverter saturation: This filter type corresponds to instances of high DC/AC ratios of PV systems causing inverter clipping. Here, the produced DC power of the PV modules exceeds the rated AC power of the inverter and system output power will thus be limited to the rated AC level. To account for these instances a threshold filter at the saturation bound of the AC power can be set, usually 99% of the rated AC power is selected.

No matter which filters are applied, it is of utmost importance to ensure that the filtered data are trustful and of high quality. Furthermore, it is important to report the applied filters to be able to trace back any data issues one might experience at a later stage of the calculations, and to ensure reproducibility of the study.

Below, three different example filters are applied for the experimental PV plant introduced in Section 2 and the filtered data are presented in power vs. in-plane irradiance plots as well as aggregated daily *PR* plots.

Figure 7 illustrates the effect of different filters not just on the raw data, but also on *PR* time series, which deviate slightly among the chosen filters. Below the figures, the applied filter as well as the filter ratios (after removing negative values) are depicted. The filters represent three commonly used approaches for a subsequent PV performance analysis such as performance loss rate calculations. It has to be kept in mind that data filtering is location dependent as in some cases it is possible to apply stricter filters without losing too much data (especially approaches such as irradiance threshold or clear-sky filter).

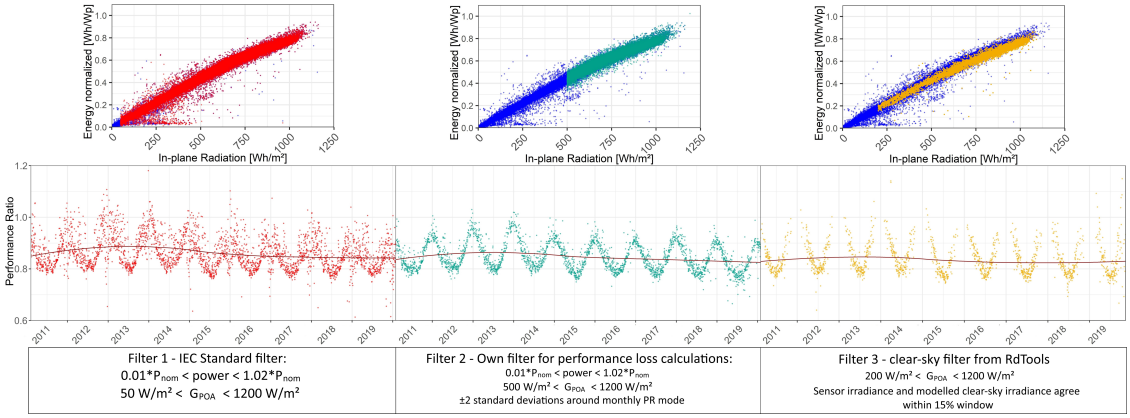

Figure 7. *P* vs. G_{POA} and daily *PR* vs. time plots for various filters applied to mc-Si PV system data—Filter 1: standard IEC 61724:2017 filter [10,20]; Filter 2: own filter used for performance loss analysis; Filter 3: clear-sky filter used in RdTools [43].

Filter 1 is used to only remove obvious measurement errors and nighttime values. The sole application of this filter is only recommended if someone is interested in the yearly produced power or yearly energy yield of the system since almost all data of importance are kept but no outlier removal has taken place. The daily *PR* of filter 1 plot omits numerous daily *PR* values below 0.6 for better visualization. The other two filters are removing a greater amount of raw data by applying stricter threshold and statistical filtering. Both filters provide efficient outlier removal using different approaches. Although filter 2 is based on a high irradiance threshold filter combined with a statistical power-irradiance outlier removal, filter 3 overlaps the measurements with a clear-sky model. Filter 2 requires an irradiance data quality check, while clear-sky filters (filter 3) account for that. The drawback of filter 3 is the high amount of filtered data. It is visible that only 3.8% of the initial raw data are kept. This is also visible in the daily *PR* figure, where just a few data points are

remaining during winter months. Although the share of removed data is also fairly high for filter 2, the remaining data still account for roughly 80% of the produced power from the system. Nevertheless, the high irradiance threshold may exclude performance affecting effects at lower irradiance levels. It is visible that the choice of the filter directly affects the shape of the *PR* time-series (or another KPI time-series), which is often used for further data evaluation. Consequently, the choice of the filter will alter the results of the evaluation.

It is important to consider that all three filtering approaches used in this study are deployed for performance analysis by analysts, depending on their preferences. That is why this topic warrants a serious discussion, and a way towards standardization should be explored.

It is clear that data filtering is a complex topic and the choice of filter and the ratio between raw and filtered data will influence all subsequent analyses steps. Usually, filtered data are aggregated to a desired time resolution such as days or months. If the filter ratio is not too high, aggregated time series should be created without large gaps. Small holes are either ignored or filled using common models such as moving average, interpolation or extrapolation. If large holes are present in the time series, filling algorithms should be used with care and possibly another raw data quality check is advisable. From experience, up to 20% of missing data can be usually recovered if they are spread along the timeline and the KPI follows a steady pattern.

There are certain filter guidelines one should obey, but unfortunately no universal filtering procedure for performance loss analyses has been developed yet apart from the minimal filter guidelines listed in standard IEC 61724:2017. We believe guidelines/standards should include the usage of lower irradiance thresholds (in the order of 100–200 W/m^2) and stringent power-irradiance pair filters such as relaxed clear-sky or *PR* filters. Lower irradiance thresholds ensure the inclusion of performance behavior along all irradiance levels. Power-irradiance filter should be applied to remove outlier, which would otherwise affect the analysis results. These outlier filters should be based on dynamic intervals to account for different climate and installation conditions and varying data quality.

5. Discussion and Summary

In this paper, different data visualization tools for measured PV system data have been presented for an example PV plant, together with an overview of common data filtering approaches. Furthermore, data imputation algorithms for module temperature and in-plane irradiance have been discussed. PV system data measurements are primarily used for yield predictions, remaining-useful-lifetime estimates, performance loss calculations or other performance related studies. Independent of the purpose, a good understanding and validation of data quality is crucial to perform any kind of reliable analyses.

Usually, measured data for PV system studies come from two different sources. Electrical parameters are collected directly from the PV system in question and meteorological data are recorded on-site or in close proximity to the test site. If no weather-related data are available, other sources must be used. The temporal visualization of a KPI, such as power or performance ratio, is recommended to get a sense of the data quality at hand. Furthermore, the nearly linear relation between power and irradiance can be used to detect data outliers or sensor related issues. After data issues have been detected and addressed, the raw data must be filtered depending on the purpose of the analysis. In general, four common filtering types are applied: in-plane irradiance threshold, power threshold, statistical filter around averaged performance ratio and clear-sky instance filter. Therefore, different thresholds and interval windows are set. A standardized set of outlier filters is given in the standard IEC 61724:2017. Unfortunately, tailored filters for analyses such as remaining-useful-lifetime estimates or performance loss studies are not available at that point and the choice of filtering method is up to the individual research group. We believe that the development towards standardized filtering guidelines would improve performance analysis inter-comparisons and the reliability of reported results. In general, low irradiance thresholds (100–200 W/m^2) together with strict power-irradiance pair filtering seem to yield a good compromise between keeping a sufficient amount of data and at the

same time ensuring high-quality power-irradiance data pairs. Low irradiance thresholds include a wider system performance range, which provides a more complete picture of the system performance.

In the first part of this paper, data imputation techniques have been tested to impute relatively small amounts (up to 20%) in-plane irradiance data and module temperature data. The motivation for this study was the investigation of PV datasets where a limited amount of data measurements were missing or faulty. Machine learning algorithms have been compared to classical irradiance and temperature models. Strongly correlated field measurements of other parameters have been used for this purpose. Module temperature data have been recovered using multivariate adaptive regression splines which yielded very similar results to the widely used Sandia module temperature model. The advantage of the algorithm over the Sandia model is that it does not require any metadata of the plant. For in-plane irradiance data imputation, several machine learning algorithms have been investigated and compared to models available in the PV performance modeling software PVLIB. Histogram-based gradient boosting regression outperformed classical irradiance models as well as other machine learning approaches for this purpose.

This work aims to be a step towards a more conscientious approach of measurement data handling. Although this topic is very complex, respecting certain guidelines and keeping a rigorous data correction and filter track will help to lower the bias in filtered data and ultimately lead to more reliable data analysis results. Furthermore, aside from visual inspections of the datasets, some form of automatized analysis performed in a regular interval as part of the measurement system could signal faulty measurements at an early stage, so data losses can be prevented and longer, continuous time series will be ensured. We recommend that authors in this field clearly present and discuss the chosen filters and how they affect the analysis and results, which will benefit standardization efforts in e.g., performance loss studies.

In the future, PV system data must be better understood to ensure comparable performance loss studies or end-of-life estimates. Therefore, paving the way towards standardized data handling and filtering practices will be crucial for any kind of PV performance investigations.

Author Contributions: S.L., A.L. and D.M. collected and provided the set of data for the study. S.L. and A.L. worked on data processing and analyses and wrote the paper. S.L., A.L., D.M. and M.T. discussed the results and commented on the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 721452-Project Solar-Train, and from the European Regional Development Fund (ERDF) of the Autonomous Province Bolzano—South Tyrol under project number FESR1128-Project PV4.0.

Acknowledgments: The authors would like to thank BayWa r.e. Operation Services S.r.l./Italy, Matevz Bokalic from the University of Ljubljana and the University of Utrecht for providing PV system datasets. Furthermore, data have been used from the US Dept. of Energy Regional Test Center Project [44].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Woyte, A.; Richter, M.; Moser, D.; Reich, N.; Green, M.; Mau, S.; Beyer, H.G. *Analytical Monitoring of Grid-Connected Photovoltaic Systems*; Report IEA-PVPS T13-03:2014; International Energy Agency: Paris, France, 2014.
2. Van Sark, W.; Louwen, A.; Tsafarakis, O.; Moraitis, P. PV System Monitoring and Characterization. In *Photovoltaic Solar Energy*; John Wiley and Sons, Ltd.: Hoboken, NJ, USA, 2017; Chapter 11.4, pp. 553–563. [CrossRef]
3. 3E Solar Irradiance Sensor Check. Available online: https://www.3e.eu/data-services/sensor-check/ (accessed on 9 June 2020).
4. Klise, K.A.; Stein, J.S. *Automated Performance Monitoring for PV Systems Using Pecos*; Sandia Technical Report; Sandia National Laboratories: Albuquerque, NM, USA, 2016. [CrossRef]

5. Kurtz, S.; Newmiller, J.; Kimber, A.; Flottemesch, R.; Riley, E.; Dierauf, T.; McKee, J.; Krishnani, P. *Analysis of Photovoltaic System Energy Performance Evaluation Method*; NREL Technical Report; NREL: Golden, CO, USA, 2013. [CrossRef]
6. Killinger, S.; Engerer, N.; Müller, B. QCPV: A quality control algorithm for distributed photovoltaic array power output. *Sol. Energy* **2017**, *143*, 120–131. [CrossRef]
7. Livera, A.; Theristis, M.; Makrides, G.; Georghiou, G.E. Recent advances in failure diagnosis techniques based on performance data analysis for grid-connected photovoltaic systems. *Renew. Energy* **2019**, *133*, 126–143. [CrossRef]
8. Koubli, E.; Palmer, D.; Rowley, P.; Betts, T.; Gottschalg, R. Assessment of PV System Performance with Incomplete Monitoring Data. In Proceedings of the 31st European Photovoltaic Solar Energy Conference and Exhibition, Hamburg, Germany, 14–18 September 2015. [CrossRef]
9. Huld, T.; Friesen, G.; Skoczek, A.; Kenny, R.P.; Sample, T.; Field, M.; Dunlop, E.D. A power-rating model for crystalline silicon PV modules. *Sol. Energy Mater. Sol. Cells* **2011**, *95*, 3359–3369. [CrossRef]
10. IEC61724-1:2017. *Photovoltaic System Performance—Part 1: Monitoring*; International Electrotechnical Commission: Geneva, Switzerland, 2017.
11. Kottek, M.; Grieser, J.; Beck, C.; Rudolf, B.; Rubel, F. World Map of the Köppen-Geiger Climate Classification Updated. *Meteorol. Z.* **2006**, *15*, 259–263. [CrossRef]
12. IEC60904:2015. *Photovoltaic Devices—Part 2: Requirements for Photovoltaic Reference Devices*; International Electrotechnical Commission: Geneva, Switzerland, 2015.
13. Demirhan, H.; Renwick, Z. Missing value imputation for short to mid-term horizontal solar irradiance data. *Appl. Energy* **2018**, *225*, 998–1012. [CrossRef]
14. Turrado, C.; López, M.; Lasheras, F.; Gómez, B.; Rollé, J.; Juez, F. Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions. *Sensors* **2014**, *14*, 20382–20399. [CrossRef]
15. Yagli, G.M.; Yang, D.; Gandhi, O.; Srinivasan, D. Can we justify producing univariate machine-learning forecasts with satellite-derived solar irradiance? *Appl. Energy* **2020**, *259*, 114122. [CrossRef]
16. Ross, R.G. Flat-Plate Photovoltaic Array Design Optimization. In Proceedings of the 14th IEEE Photovoltaic Specialists Conference, San Diego, CA, USA, 7–10 January 1980.
17. Kratochvil, J.A.; Boyson, W.E.; King, D.L. *Photovoltaic Array Performance Model*; Sandia Technical Report; Sandia National Laboratories: Albuquerque, NM, USA, 2004. [CrossRef]
18. Skoplaki, E.; Boudouvis, A.; Palyvos, J. A simple correlation for the operating temperature of photovoltaic modules of arbitrary mounting. *Sol. Energy Mater. Sol. Cells* **2008**, *92*, 1393–1402. [CrossRef]
19. Mattei, M.; Notton, G.; Cristofari, C.; Muselli, M.; Poggi, P. Calculation of the polycrystalline PV module temperature using a simple method of energy balance. *Renew. Energy* **2006**, *31*, 553–567. [CrossRef]
20. IEC61724-3:2016. *Photovoltaic System Performance—Part 3: Energy Evaluation Method*; International Electrotechnical Commission: Geneva, Switzerland, 2016.
21. Ascencio-Vasquez, J.; Bevc, J.; Reba, K.; Brecl, K.; Jankovec, M.; Topic, M. Advanced PV Performance Modelling Based on Different Levels of Irradiance Data Accuracy. *Energies* **2020**, *13*, 2166. [CrossRef]
22. Mariottini, F.; Belluardo, G.; Bliss, M.; Isherwood, P.; Cole, I.; Betts, T. Assessment and improvement of thermoelectric pyranometer measurements. In Proceedings of the 36th European Photovoltaic Solar Energy Conference and Exhibition, Marseille, France, 30 October 2019.
23. Sengupta, M.; Habte, A.; Kurtz, S.; Dobos, A.; Wilbert, S.; Lorenz, E.; Stoffel, T.; Renné, D.; Gueymard, C.; Myers, D.; et al. *Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications*; NREL Technical Report; NREL: Golden, CO, USA, 2015.
24. Gracia Amillo, A.M.; Huld, T. *Performance Comparison of Different Models for the Estimation of Global Irradiance on Inclined Surfaces*; JRC Technical Report; Publications Office of the European Union: Ispra, Italy, 2013. [CrossRef]
25. Toledo, C.; Amillo, A.; Bardizza, G.; Abad, J.; Urbina, A. Evaluation of Solar Radiation Transposition Models for Passive Energy Management and Building Integrated Photovoltaics. *Energies* **2020**, *13*, 702. [CrossRef]
26. Pearsall, N. *The Performance of Photovoltaic (PV) Systems*; Woodhead Publishing: Cambridge, UK, 2017; ISBN 978-1-78242-336-2.
27. DKASC. Alice Springs | DKA Solar Centre. Available online: http://dkasolarcentre.com.au/locations/alice-springs (accessed on 8 November 2019).

28. Holmgren, W.; Hansen, C.; Mikofski, M. Pvlib python: A python package for modeling solar energy systems. *J. Open Source Softw.* **2018**, *3*, 884. [CrossRef]
29. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
30. Hottel, H.; Woertz, B. The Performance of Flat Plate Solar-Heat Collectors. *Trans. ASME* **1942**, *64*, 64–91.
31. Klucher, T. Evaluation of models to predict insolation on tilted surfaces. *Sol. Energy* **1979**, *23*, 111–114. [CrossRef]
32. Hay, J.; Davies, J. Calculations of the solar radiation incident on an inclined surface. In Proceedings of the First Canadian Solar Radiation Data Workshop, Toronto, ON, Canada, 17–19 April 1978; Ministry of Supply and Services: Ottawa, ON, Canada, 1980.
33. Reindl, D.; Beckmann, W.; Duffie, J. Diffuse fraction correlations. *Sol. Energy* **1990**, *45*, 1–7. [CrossRef]
34. Reindl, D.; Beckmann, W.; Duffie, J. Evaluation of hourly tilted surface radiation models. *Sol. Energy* **1990**, *45*, 9–17. [CrossRef]
35. Perez, R.; Seals, R.; Ineichen, P.; Stewart, R.; Menicucci, D. A new simplified version of the Perez diffuse irradiance model for tilted surfaces. *Sol. Energy* **1987**, *39*, 221–232. [CrossRef]
36. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
38. Friedman, J. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
39. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3149–3157.
40. IEC60904-3:2016. *Photovoltaic Devices—Part 3: Measurement Principles for Terrestrial Photovoltaic (PV) Solar Devices with Reference Spectral Irradiance Data*; International Electrotechnical Commission: Geneva, Switzerland, 2016.
41. Wild, M.; Gilgen, H.; Rösch, A.; Ohmura, A.; Long, C.N.; Dutton, E.G.; Forgan, B.; Kallis, A.; Russak, V.; Tsvetkov, A. From Dimming to Brightening: Decadal Changes in Solar Radiation at Earth's Surface. *Science* **2005**, *308*, 847–850. [CrossRef] [PubMed]
42. Reno, M.J.; Hansen, C.W. Identification of periods of clear sky irradiance in time series of GHI measurements. *Renew. Energy* **2016**, *90*, 520–531. [CrossRef]
43. Deceglie, M.G.; Jordan, D.C.; Nag, A.; Deline, C.A.; Shinn, A. RdTools: An Open Source Python Library for PV Degradation Analysis. In Proceedings of the PV Systems Symposium, Albuquerque, NM, USA, 1–3 May 2018; NREL: Golden, CO, USA, 2018.
44. Sandia National Laboratories; United States Department of Energy Office of Scientific and Technical Information. *The US DOE Regional Test Center Program: Driving Innovation Quality and Reliability*; United States Department of Energy Office of Energy Efficiency and Renewable Energy: Washington, DC, USA, 2015.