



Open-source quality control routine and multi-year power generation data of 175 PV systems

Cite as: J. Renewable Sustainable Energy **14**, 043501 (2022); <https://doi.org/10.1063/5.0100939>
Submitted: 27 May 2022 • Accepted: 27 July 2022 • Accepted Manuscript Online: 02 August 2022 •
Published Online: 26 August 2022

 Lennard R. Visser, Boudewijn Elsinga,  Tarek A. AISkaif, et al.

COLLECTIONS

 This paper was selected as Featured

 This paper was selected as Scilight



View Online



Export Citation



CrossMark

ARTICLES YOU MAY BE INTERESTED IN

[Sharing data from residential solar installations](#)

Scilight **2022**, 351111 (2022); <https://doi.org/10.1063/10.0013913>

[Short-term wind speed forecasting with regime-switching and mixture models at multiple weather stations over a large geographical area](#)

Journal of Renewable and Sustainable Energy **14**, 043305 (2022); <https://doi.org/10.1063/5.0098090>

[Mechanism study of flow characteristics on small HAWT blade surfaces based on airfoil concavity under yaw conditions](#)

Journal of Renewable and Sustainable Energy **14**, 043306 (2022); <https://doi.org/10.1063/5.0095690>

APL Machine Learning

Open, quality research for the networking communities

Now Open for Submissions

LEARN MORE



Open-source quality control routine and multi-year power generation data of 175 PV systems

Cite as: J. Renewable Sustainable Energy **14**, 043501 (2022); doi: 10.1063/5.0100939

Submitted: 27 May 2022 · Accepted: 27 July 2022 ·

Published Online: 26 August 2022



View Online



Export Citation



CrossMark

Lennard R. Visser,^{1,a)}  Boudewijn Elsinga,² Tarek A. AlSkaif,³  and Wilfried G. J. H. M. van Sark¹ 

AFFILIATIONS

¹Copernicus Institute of Sustainable Development, Utrecht University, Princetonlaan 8a, 3584 CB Utrecht, The Netherlands

²W/E Consultants, Jan van Hooffstraat 8E, 5611 ED Eindhoven, The Netherlands

³Information Technology Group, Wageningen University and Research, Droevendaalsesteeg 4, 6708 PB Wageningen, The Netherlands

^{a)} Author to whom correspondence should be addressed: l.r.visser@uu.nl

ABSTRACT

In this study, we introduce an open-source dataset holding power measurements of 175 residential photovoltaic (PV) systems that are distributed throughout the province of Utrecht, the Netherlands. The dataset features power measurements with a high temporal resolution, i.e., 1 min, for the period January 2014 until December 2017 (over 260×10^6 data points). Spatial information of the PV systems is mapped through latitude and longitude grids, with a resolution up to 150×150 m. In addition, we develop and publish a quality control routine that can be applied to validate and filter PV power measurements. Finally, we propose a method to estimate the rated DC capacity of a PV system based on the power measurements. We have deposited five files into the Zenodo repository [Visser *et al.* (2022). Zenodo, V. 0.0.1, Dataset <https://doi.org/10.5281/zenodo.6906504>], which are publicly available. Four numerical datasets are enclosed, holding unfiltered power measurements, filtered power measurements at two different stages and metadata. The latter includes information on the tilt angle, azimuth angle, the estimated DC and AC capacity, and location. Finally, a Python package featuring the quality control routine developed to validate and filter PV power measurements is published.

Published under an exclusive license by AIP Publishing. <https://doi.org/10.1063/5.0100939>

I. INTRODUCTION

The installed capacity of solar photovoltaic (PV) systems is rapidly growing on a global scale, ensuing increasing shares of solar PV power in the electricity system in many regions.¹ Since the power output of PV systems is variable and difficult to predict, the increasing share of electricity that is produced by PV systems forms a challenge to both transmission and distribution system operators in operating the electricity grid. To this end, system operators have to invest in additional measures to allow for a high PV penetration rate, e.g., increasing grid transport capacity and having additional balancing capacity available. Consequently, the (marginal) costs for hosting PV capacity in the electricity grid will increase with a growing PV penetration rate. Alternatively, system operators may set a maximum allowable PV capacity, due to technical or economic constraints.¹

A better understanding of the PV power output, variability, and impact on the electricity grid can lower these barriers and is needed to support the large-scale adoption of PV systems. This can be obtained through extensive research into a variety of topics, related to PV ramp rates, solar power forecasting, PV-battery systems, model predictive

control, grid simulations, and more. Nevertheless, at the moment, there is a lack of high-resolution data that is publicly available to the research community, limiting the ability to complete such studies.² Some exceptions include Bright *et al.*³ and Silwal *et al.*² Kapoor *et al.*⁴ present a recent overview of additional available datasets. Yet, all these datasets typically feature a coarse temporal resolution (≥ 15 min) or present data for a short period (< 1 month).

The lack of public datasets raises two additional issues. First of all, since many studies rely on nonpublic datasets, published results cannot be verified or reproduced by other researchers. Second, a lack of publicly available data preclude the ability for researchers and other interested parties to conduct research or compare obtained results among several studies.

Before any dataset can be used, the quality of the data must be validated. The objective of this quality control routine is to detect erroneous values. In the present literature, only few studies were found to deal with the quality control of PV power measurements,^{5,6} where a standardized quality control routine was as far as the authors are aware of only proposed by Killinger *et al.*⁷ Therefore, this field is still in a

pre-mature stage and less developed compared to, e.g., quality control of solar irradiance measurements.⁸ The absence of standardized (public) quality control routines for PV power data is a concern as the obtained results in any study may be highly affected by erroneous data. Yet, if any at all, most studies that consider PV power measurements report very limited details on the process followed to quality control the data that was used. In addition, the results of such checks are rarely disclosed, e.g., number of values filtered. As a consequence, there is a high need for publicly available standardized routines that can be widely adopted to quality control PV power measurements.

Both open-source data and code as well as standardized routines form a key element in realizing progress in any field of study. Given the current state of open-source data and code related to PV power measurements, more attention should be put on publishing data and developing and publishing quality control routines. In this paper, we present an open-source dataset holding PV power measurements of 175 PV systems with a 1-min temporal resolution for January 2014 until December 2017. This constitutes a dataset of over 260×10^6 quality-controlled data points. With this, we provide access to a high-quality dataset that can be used for research experiments on several topics related to PV power and the integration of PV in the electricity grid. Moreover, as the dataset with unfiltered power values is enclosed, it can be used to test adapted and/or new developed quality control routines, and verify and compare results. In addition, with this study, we strive to set a next step into developing a standardized routine for quality control of PV power measurements. By publishing a Python package holding the functions to conduct the quality control routine along with the dataset,⁹ we hope to stimulate others to adopt and improve the routine of quality control and work toward a widely adopted standardized routine. Finally, in this study, we present a novel approach to estimate the rated DC capacity of a PV system given the PV power measurements, tilt angle, azimuth angle, and weather data.

The paper is organized as follows. Section II presents the data records. Section III discusses the research methods and describes the approach followed to obtain the DC and AC capacity of a PV system as well as the quality control routine. Section IV shows the results of this study. The conclusions and recommendations follow in Secs. V and VI.

II. DATA RECORDS

A. PV systems

The dataset presented in this study features power measurements of rooftop mounted residential PV systems, which all comprise c-Si panels. The power records of these PV systems were initially collected in context of the Solar Forecasting and Smart Grids research project, see Elsinga.¹⁰ Furthermore, parts of this dataset were used in former projects for the development of (peer-to-peer) solar forecasting methods.^{11,12} The PV systems are located in the province of Utrecht, the Netherlands that covers an area from $52^{\circ}30' \text{ N}$ to $51^{\circ}68' \text{ N}$ and $4^{\circ}79' \text{ E}$ to $5^{\circ}62' \text{ E}$, see Fig. 1. Due to privacy concerns, the location of the PV systems is enclosed in the form of mapped grids that are formed by two latitude and two longitude lines, which present the East, West, North, and South boundaries. This anonymization procedural step is in accordance with proposed GDPR-compliant anonymization techniques.¹³

Figure 2 visualizes the characteristics of the PV systems included in the dataset as well as their annual production. The rated DC capacity of the PV systems is estimated as explained in Sec. III A 1. Additionally, the annual yield per PV system is determined for the entire period, January 2014 until December 2017, where missing values are filled by considering weather measurements at a local weather station¹⁴ and a PV model. The PV model is discussed in Sec. III A 1.

B. Power measurements

1. Measurements device

Every PV system is directly connected on the AC side of the inverter with a data logger that measures the AC power output. The power measurements are recorded at 0.5 Hz (i.e., takes approximately one power measurement per 2 s time interval) and 0.7 W resolution.¹⁵ After collection, the power measurements are via an internet connection instantly send to the remote server where the data are also stored.

2. Missing data

The unfiltered dataset contains missing values throughout the period. These values are reported as NaN. As can be observed in the unfiltered data file enclosed (unfiltered_pv_power_measurements.csv), the missing values are usually limited to a single system at a time. These are caused by either a temporal power outage or disconnection of the data logger. During some periods data are missing for all PV systems, which is due to a disconnection of the remote server, mostly due to failure or power outage of the local WiFi router. An overview of the data availability per system is given in Appendix B.

III. METHODS

A. AC and DC capacity estimation

1. DC capacity estimation

Although the rated DC capacity of the PV systems was initially reported by the owners of the system, several mistakes were found. Therefore, we developed a universal method that estimates the DC capacity by scaling the simulated power output of a PV system on a selection of clear sky days. An overview of the procedure is presented in Fig. 3 and considers the following steps (the numbers in Fig. 3 correspond to the steps discussed next). First, daily weather information regarding the level of cloudiness is retrieved from the closest weather station, de Bilt (located in the province of Utrecht, $52^{\circ}10' \text{ N}$, $5^{\circ}18' \text{ E}$, see Fig. 1) for the period 2014–2017. In its dataset, the Netherlands Royal Meteorological Institute (KNMI)¹⁶ reports the daily level of cloudiness using a classification system holding nine classes (octa's) from clear sky to complete overcast. From this, all clear sky days are obtained in the period 2014–2017. As a second step, through visual inspection of the daily power measurements of 40 PV systems, a sub-selection is made, where days are dropped in case the PV power measurements of at least 25% of the systems is inconsistent with the expected power output on the clear sky day. As a result, the final selection contains 18 clear sky days, see Table I. These 18 days form the base selection for the next step. Yet, in another round of visual inspection, a sub-selection of these days is made per system, where additional days may be dropped from the assessment in case part of the power

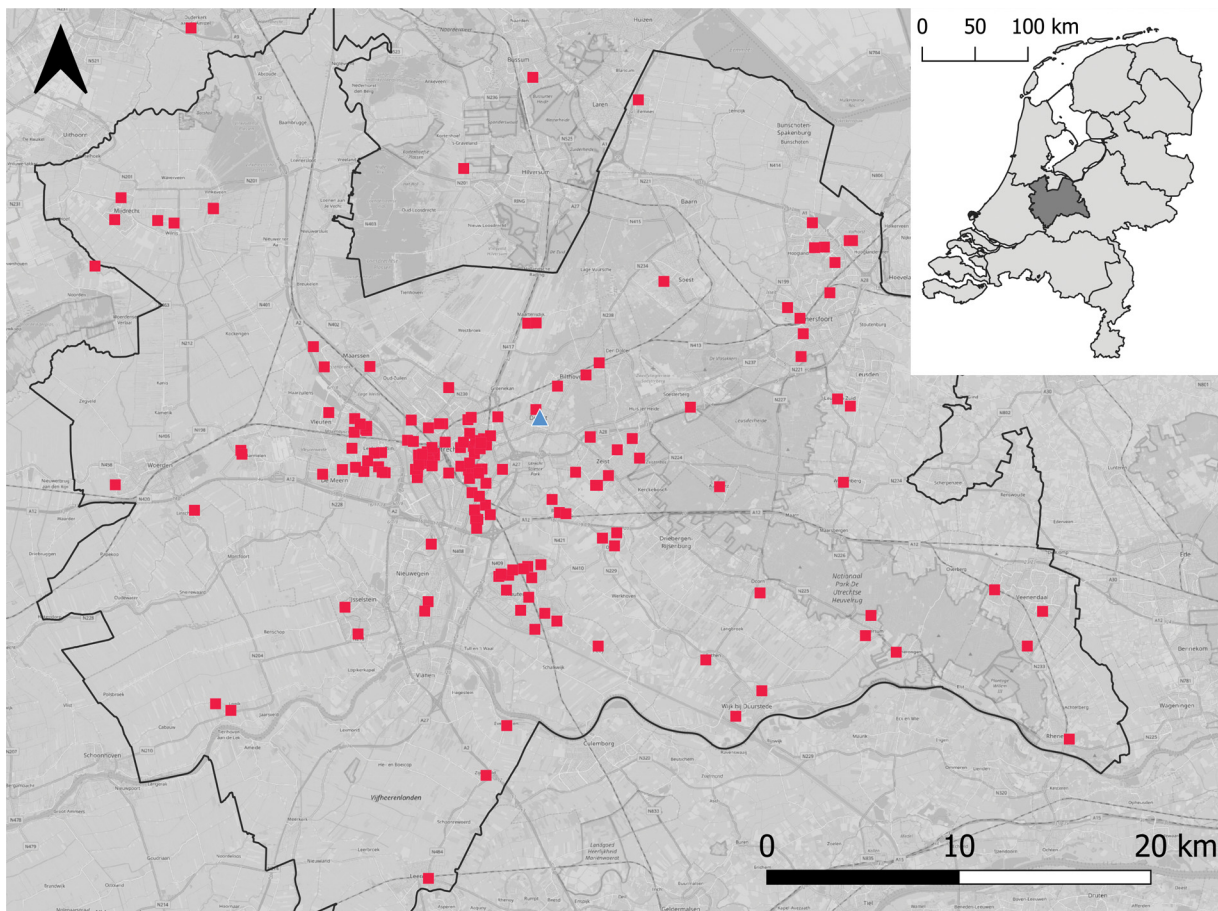


FIG. 1. Distribution of PV systems (red squares) and weather station De Bilt (blue triangle), Utrecht, the Netherlands. The location marks the center of the mapped grids.

measurements are missing or the reported power measurements are inconsistent with the expected production during clear sky conditions.

In the third step, the power measurements on these clear sky days are aggregated to 15-min values by averaging. As a fourth step, the hourly measurements reported by the weather station, including surface pressure, temperature, wind speed, and global horizontal irradiance (GHI),¹⁴ are up-sampled to 15-min samples by means of linear interpolation.

In the fifth step, the power output of each PV system is simulated based on the weather data, and the azimuth and tilt angle (as reported in the file metadata.csv). To this end, a number of sub-steps are to be taken. First, the Erbs model is used to estimate the direct normal and diffuse horizontal irradiance (DNI, DHI) from the reported GHI values.¹⁷ Next, the Perez model is used to calculate the in-plane irradiance components, given the azimuth and tilt angles of the PV system.¹⁸ In addition to the system characteristics and weather variables, the Perez model requires the input of the relative airmass, which is estimated with the Kasten–Young model.¹⁹ The effective in-plane irradiance is then obtained by applying the angle of incidence losses using a physical model.²⁰ As a last sub-step, the cell temperature is calculated by the Sandia Array Performance Model.²⁰ Finally, the power output of the

PV system is simulated with the PVWatts model.²¹ Since we do not know the DC capacity per PV system at this moment, a normalized PV power output profile is simulated. All the required models are available in the Python package pvlib.²² An overview of other relevant parameters and assumptions is given in Table II.

In the sixth step, the normalized simulated PV power output profile that is created for each system i individually ($p_{ns,i}$, expressed as W/W_p) is scaled to the observed power output (p_m) for the selection of clear sky days. The scaling is done by considering one adjustable parameter, i.e., the estimated DC capacity (S_{DC}). Hence, the simulated PV power output of a PV system is obtained by ($p_s = p_{ns} * S_{DC}$). Scaling is performed by means of least squares minimization of the residual between the PV power measurements and simulations, see Eq. (1), for each system i . An example is provided in Fig. 4. To exclude the potential impact of inverter clipping, we limit the estimation approach to times t where the power output is smaller than the inverter capacity (S_{AC} , which is obtained as explained in Sec. III A 2). To limit the impact of shading of surrounding objects and because of higher uncertainties of retrieving the DNI and DHI for high zenith angles, the estimation process is also limited to timestamps where the solar zenith angle (θ_z) is below 65° , which is given as follows:

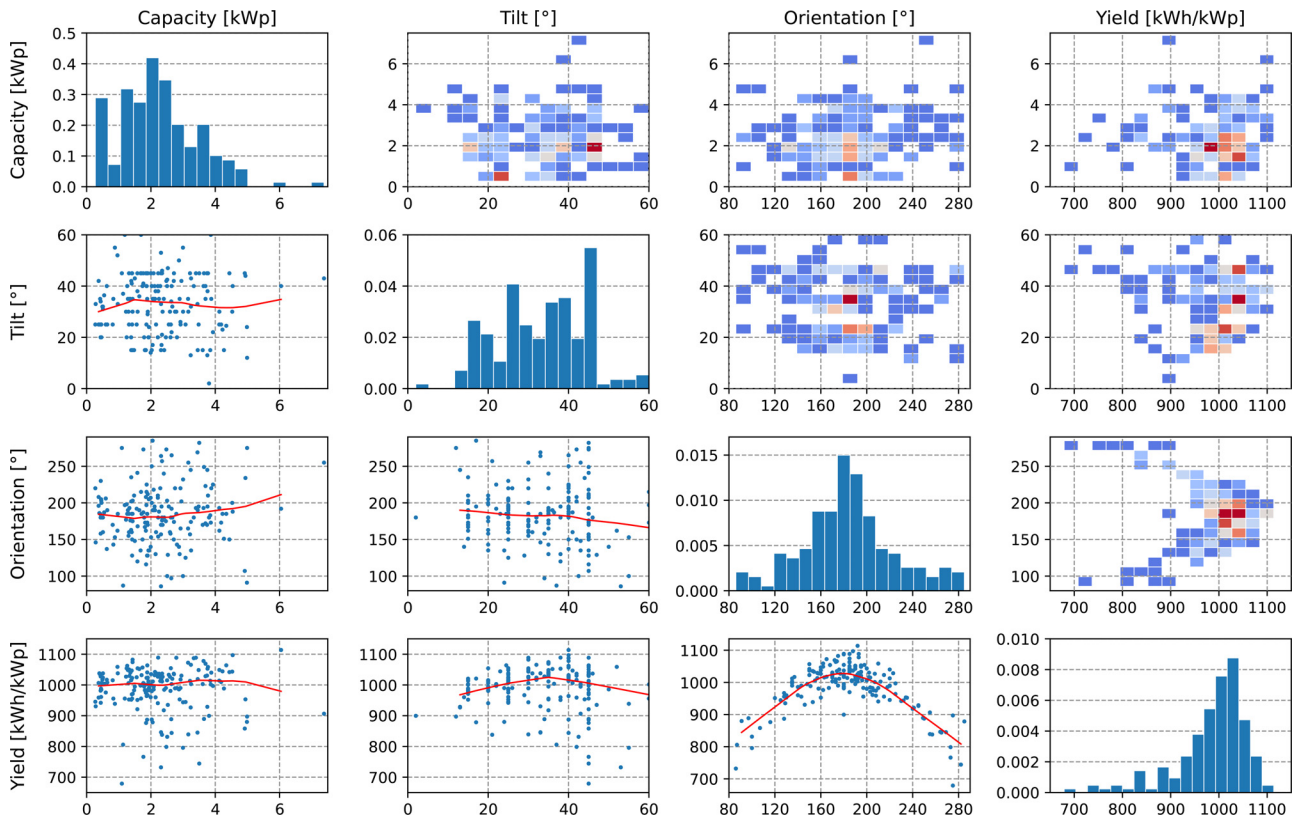


FIG. 2. An overview of the characteristics of the PV systems included in the dataset as well as the annual yield for the entire period 2014–2017. The figure should be read as a matrix, where the sub figures in the diagonal present the probability distribution of the characteristics among the PV systems. The sub figures to the left are scatter plots, whereas the red line shows the trend, i.e., a best fit line $[f(x)]$ fit to minimize $\sum_{i=1}^N (f(x_i) - y)^2$, where y is the true value of the PV system characteristic for N systems. The FIGS to the right (of the diagonal) present a heat map, where the color marks the density of systems found from low (blue) to high (red).

$$\text{minimize } \sum_{t=1}^T (p_m(t) - (p_{ns,i}(t) * S_{DC}))^2 \dots \forall t \quad (1)$$

where $(\theta_z < 65^\circ) \wedge (p_m(t) < S_{AC})$.

The ability of the estimation algorithm to obtain the DC capacity per PV system is evaluated by considering the relative Root Mean Square Error (RMSE) and bias of the residuals between the PV power simulations and measurements for the selection of clear sky days. The

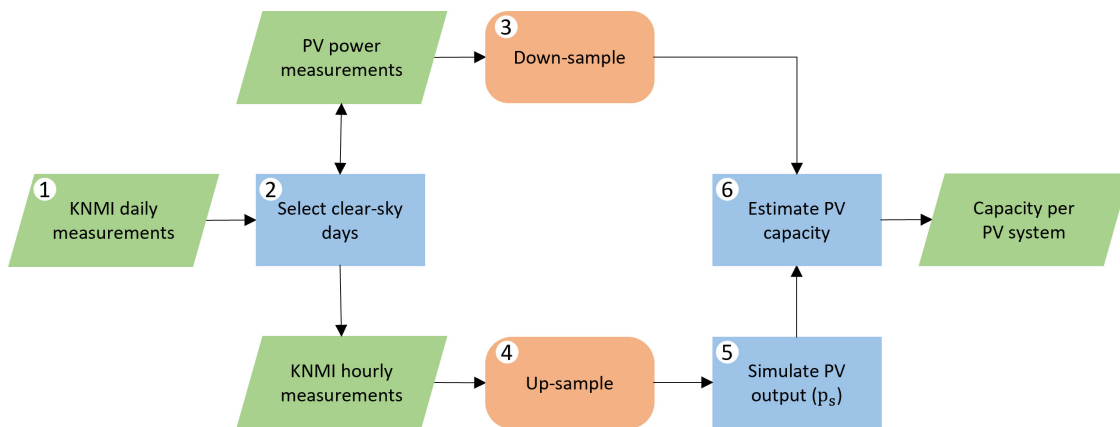


FIG. 3. Flowchart of the procedure followed to estimate the DC capacity per PV system. Numbers denote steps that are discussed in the text.

TABLE I. The clear sky days selected and used to estimate the rated DC capacity of each PV system.

Year	Date	Count
2014	March 9, 12, 13; and April 16	4
2015	March 12; April 9; June 4, 30; and October 2	5
2016	May 8; July 19; August 17; and September 13, 14	5
2017	February 13; April 9; May 26; and October 15	4

TABLE II. Parameter settings for power output calculations of the PV systems. DC capacity is expressed per square meter and inferred from typical DC capacities of PV modules as listed in the CEC²² and Sandia²⁰ module databases.

DC capacity (Wp/m ²)	180 ^{20,22}
Inverter efficiency (%)	96.0 ²¹
Temperature coefficient of power	-0.0035 ²¹

RMSE and bias are given in as follows, where $\frac{1}{S_{DC}}$ is added to obtain the relative error values:

$$RMSE = \frac{1}{S_{DC}} \sqrt{\frac{1}{T} \sum_{t=1}^T (p_m(t) - (p_{s,i}(t)))^2}, \quad (2)$$

$$Bias = \frac{1}{S_{DC}} \frac{1}{T} \sum_{t=1}^T (p_m(t) - (p_{s,i}(t))). \quad (3)$$

2. AC capacity estimation

Similar to the DC capacity, multiple mistakes were found in the initial documentation of the inverter or AC capacity. In order to infer the rated AC capacity per PV system, we followed a simple and effective approach, where we empirically evaluate the capacity. To this end, per PV system we performed a visual inspection on the cumulative distribution function of all power measurements (see Fig. 5) and on the time series of the PV power measurements for April, May, and June, which hold the highest PV power output values in the Netherlands. An example of the latter is shown in Fig. 6. In case a PV system experiences inverter clipping, this is characterized by a cut off power value, which is rarely exceeded. This cutoff value is expressed as the horizontal part of the cumulative distribution curve in Fig. 5 and the maximum power output value observed in Fig. 6. The cut off value per PV system (marked by the red lines in Figs. 5 and 6) is set equal to the AC capacity of the PV system. Note that a larger ratio DC to AC capacity is accompanied by a more significant cut off value.

B. Quality control routine

1. Quality control criteria for single PV systems

An overview of the quality control routine developed in this study is given in Fig. 7. The routine can be subdivided into a system

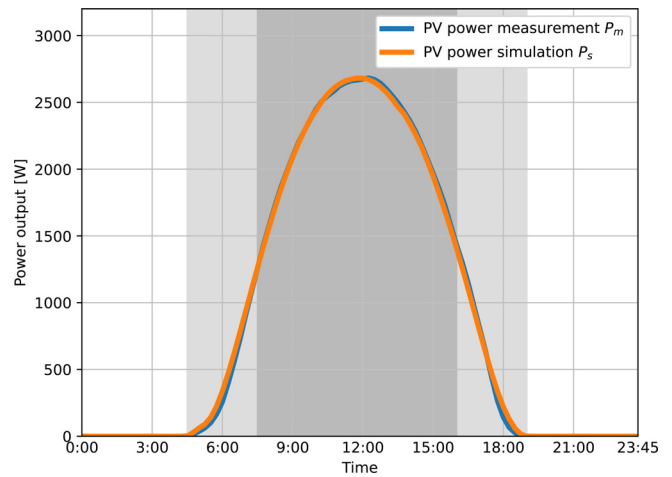


FIG. 4. An example of the simulated (P_s) and measured (P_m) power output of PV system ID128 for August 17, 2016 (a clear sky day from the selection presented in Table I). The light gray area indicates the period from sunrise to sunset. The dark gray area indicates the period where the zenith angle is smaller than 65° , which period is selected to estimate the rated DC capacity.

specific and an inter-system part, referred to as single and across. The former indicates quality control criteria (i.e., filters) that rely only on the power measurements of the specific system, the latter executes quality control criteria that require input from neighboring systems.

The first filter in the quality control routine checks the daily data availability rate of the power measurements, i.e., it ensures a minimal daily data availability. A complete day is removed from the dataset in case the daily power measurements fall below 50% of the number of timestamps, considering a 1-min resolution, for $\theta_z < 85^\circ$.

Second, a night filter is applied that sets all power output values during nighttime to zero, see Eq. (4). The filter considers nighttime as those timestamps where the expected power output of a PV system for clear sky conditions is smaller than or equal to zero, given the system's

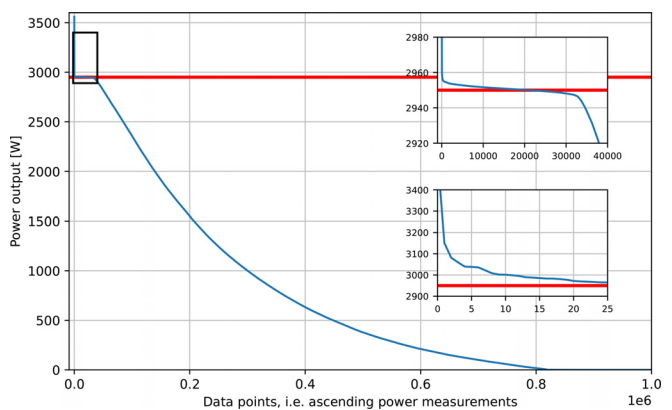


FIG. 5. Cumulative distribution curve (ascending) of the PV power output values for system ID077. The red horizontal line presents the estimated AC capacity, i.e., 2950 W. Note that the sub figures present an enlargement of the area marked with a black square.

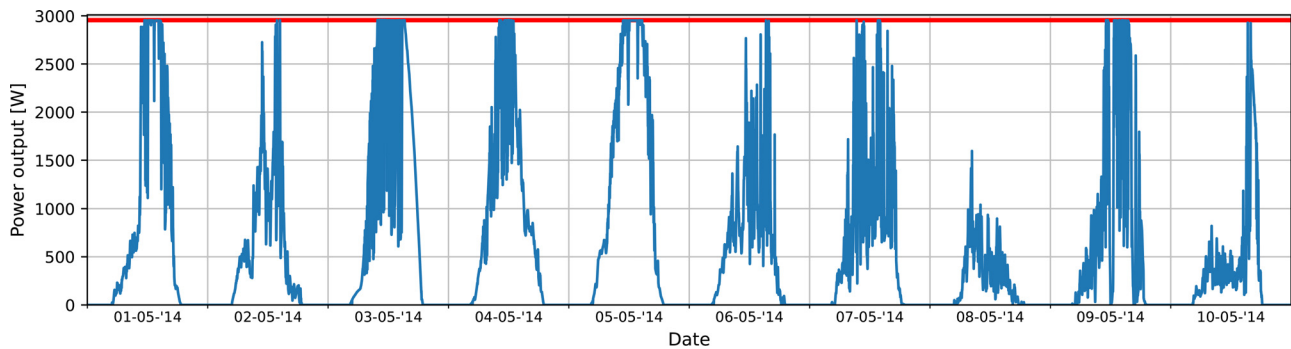


FIG. 6. Time series of the unfiltered power measurements of PV system with ID077, the red horizontal line presents the estimated AC capacity, i.e., 2950 W.

characteristics (i.e., tilt, azimuth, DC and AC capacity). The production of a PV system for clear sky conditions is simulated following the same procedure as discussed in Sec. III A 1, where the GHI is replaced by the clear sky irradiance ($p_{cs}(t)$).²³ Some data loggers were found to register negative power output values during sunrise and sunset. These negative values arise as a consequence of the orientation of the PV system, low GHI during sunrise and sunset, and shading caused by surrounding obstacles, e.g., buildings or trees. As a result, the night filter

also sets these values to zero. This is limited to a maximum time deviation of 30-min around sunrise and sunset,

$$p_{cs}(t) \leq 0 \Rightarrow p_m(t) = 0, \tag{4}$$

$$p_m(t) < 0, \quad \forall t \leq (\text{sunrise} + 30 \text{ min}) \vee t \geq (\text{sunset} - 30 \text{ min}) \Rightarrow p_m(t) = 0. \tag{5}$$

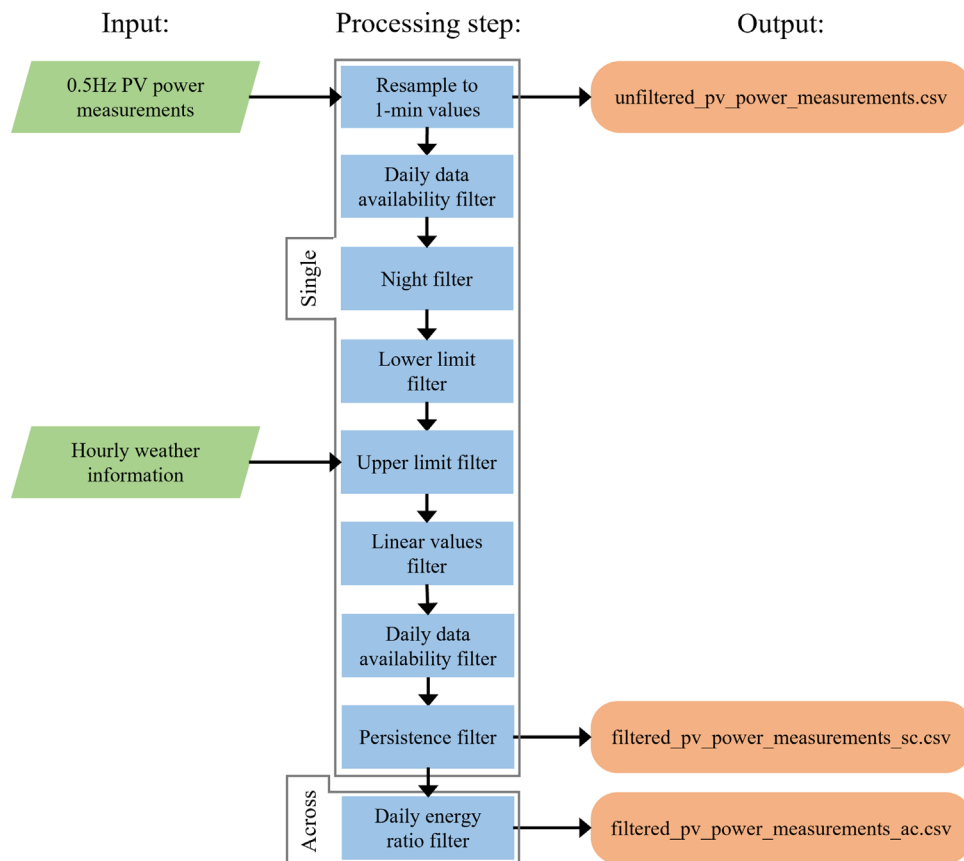


FIG. 7. Flowchart of the criteria (i.e., filters) included in the quality control routine for PV power measurements. The Python source code and output files are made available at the following repository. [Associated dataset available at <https://doi.org/10.5281/zenodo.6906504>] (Ref. 9).

Third, similar to Killinger *et al.*,⁷ a lower limit filter is applied. The lower limit filter simply assigns NaN values to all negative values. Given that this filter is applied after the night filter, the lower limit filter will only affect daytime values,

$$p_m(t) < 0 \Rightarrow p_m(t) = \text{NaN}. \quad (6)$$

An upper limit filter is applied to identify unrealistic high values. The filter sets power output values that exceed the upper limit to NaN. This filter exists of two components. The first component assigns NaN values to all values that exceed the AC capacity of the PV system, see Eq. (7). Here, v is a constant as previously it was identified that the inverter capacity can be exceeded for short periods up to 10-min.²⁴ The constant is set to 1.025, i.e., considers a temporary AC power output increase in 2.5%. Unrealistic high values can still occur as PV power output measurements may still exceed the PV power output for clear sky conditions.⁷ A second component is added to identify these instances. Here, the upper limit is defined by simulating the power output of a PV system under clear sky conditions, as also discussed in the night filter. Yet, given the high temporal resolution, cloud enhancement effects may occur.²⁴ In addition, estimated GHI for clear sky conditions as simulated by the clear sky model comes with uncertainty. As a result of the simplifications in the clear sky model (e.g., related to turbidity), the measured GHI as reported by Ref. 14 is occasionally found to exceed the simulated clear sky irradiance. Therefore, we consider a threshold value (k_{PV}) of 1.4. For small values of p_{cs} , i.e., high zenith angles, the upper limit can be exceeded,⁷ since the clear sky model and subsequently the simulated power output come with greater uncertainty for high ($\theta_z > 80^\circ$) solar zenith angles. Therefore, we adapt the upper limit filter for high zenith angles, see Eqs. (8)–(10). An example is presented in Fig. 8.

$$p_m(t) > S_{AC} \Rightarrow p_m(t) = \text{NaN}, \quad (7)$$

$$p_m(t) \geq p_{cs}(t) * k_{PV} \dots \forall t \quad \text{where } \theta_z < 80^\circ \Rightarrow p_m(t) = \text{NaN}, \quad (8)$$

$$(p_m(t) \geq p_{cs}(t) * k_{PV}) \wedge (p_m(t) \geq p_m(t) * 0.125 * S_{DC}), \dots, \forall t, \quad (9)$$

where $\theta_z \geq 80^\circ \wedge < 85^\circ \Rightarrow p_m(t) = \text{NaN},$

$$(p_m(t) \geq p_{cs}(t) * k_{PV}) \wedge (p_m(t) \geq p_m(t) * 0.075 * S_{DC}), \dots, \forall t, \quad (10)$$

where $\theta_z \geq 85^\circ \Rightarrow p_m(t) = \text{NaN}.$

This is followed by a linear data filter that is introduced to detect temporary failure of the data logger. This filter was developed as some data loggers were found to temporary report a constant, a linear increasing or decreasing value,

$$\frac{\partial p_m(t)}{\partial t} = C, \quad \forall t = \{t, t-1, \dots, t-19\} \wedge \theta_z < 85^\circ \Rightarrow p_m(t) = \text{NaN}, \quad (11)$$

where C represents a constant, which is 0 for constant power, or negative (positive) for a linear decrease (increase) of power. Equation (11) guarantees that the power measurements are set to NaN in case the filter is violated for at least 20 consecutive time steps, i.e., a period of 20 min or more. Figure 9 shows an example of the implementation of this filter.

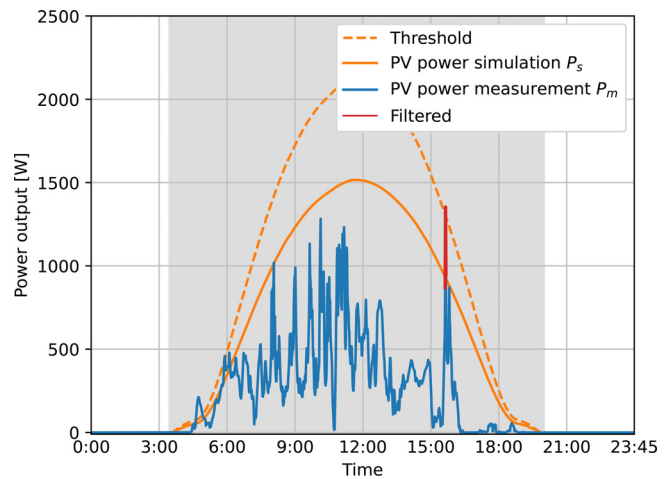


FIG. 8. An example of the operation of the upper limit filter. The figure presents the threshold value, simulated (P_s) and measured (P_m) power output of PV system ID082 for July 6, 2014. The filter identifies the values (red) that exceed the threshold values and sets these instances to NaN.

Subsequently, the minimal daily data availability filter is repeated in order to remove dates where the daily availability rate of power measurements for $\theta_z < 85^\circ$ fell below 50% as a result of the filters applied above.

Finally, a persistence filter was adopted to filter spurious data [see Eq. (12)]. The filter has the ultimate objective to flag and remove days where persistent or highly fluctuating power values are reported, resulting in minimum or extreme variability of the reported values. The persistence filter was first introduced by Journée and Bertrand,²⁵

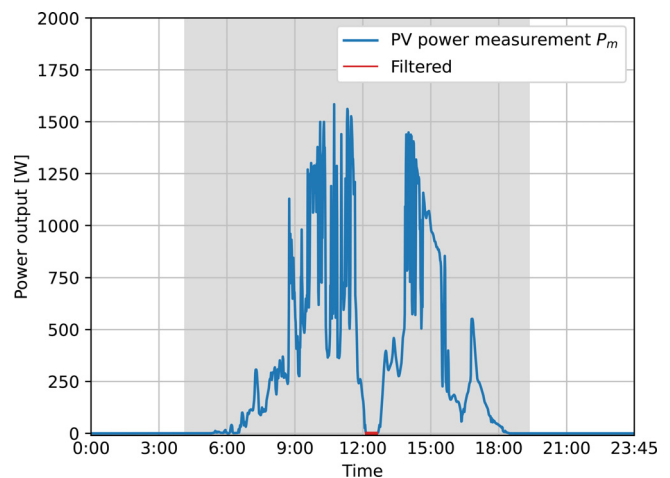


FIG. 9. An example of the operation of the linear data filter. The figure presents the measured (P_m) power output of PV system ID082 for August 7, 2014. Values around noon (red) are identified and set to NaN as the marked values present a linear line with a constant value of 0, which is caused by temporary failure of the data logger.

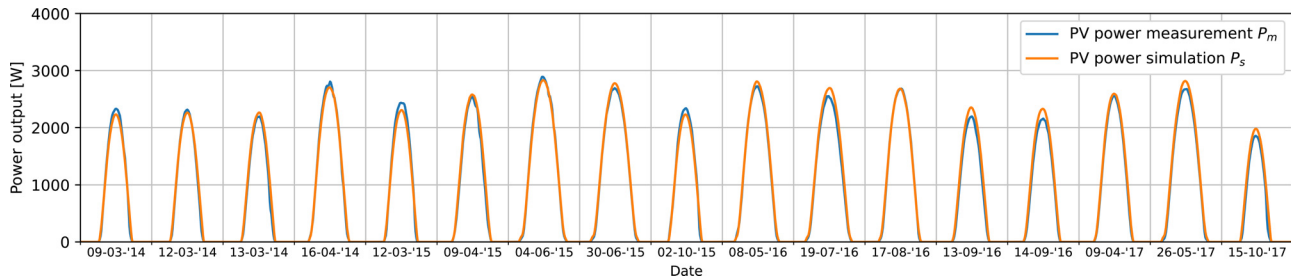


FIG. 10. The simulated (p_s) and measured (p_m) power output of a PV system for all selected clear sky days (see Table I) for system with ID128.

where it was meant to filter spurious GHI measurements. In Killinger *et al.*,⁷ it was adapted for the purpose of filtering PV power measurements.

$$\frac{1}{8} \mu \left(\frac{p_m / S_{DC}}{E_{ext}} \right) \leq \sigma \left(\frac{p_m / S_{DC}}{E_{ext}} \right) \leq \tau, \quad (12)$$

where μ and σ represent the mean and standard deviation, E_{ext} is the extraterrestrial solar radiation obtained with the pvlb package, and τ is a constant, which is set as 0.35.⁷

2. Quality control criteria across PV systems

The final step in the quality control routine concerns the application of the across system daily energy ratio filter, which is adopted from Killinger *et al.*⁷ The filter utilizes the information available from all PV systems by benchmarking the daily production values, as similar daily production values can be expected for neighboring systems. In the across filter, the daily energy ratio (r_{de}) [Eq. (13)] of each system is compared to the mean daily energy ratio of all PV systems, \bar{r}_{de} [Eq. (14)]. Consequently, for each PV system, the across system filter enables to flag and remove every day (d) with significant deviating power production values.

$$r_{de}(d) = \frac{\sum_{t=1min}^{T=24hr} p_m(t)}{\sum_{t=1min}^{T=24hr} p_{cs}(t)}, \quad (13)$$

$$-\lambda \bar{r}_{de}(d) \leq r_{de,i}(d) \leq \lambda \bar{r}_{de}(d), \quad (14)$$

where λ presents a constant, which was empirically set to 0.25. The across system filter can only be applied to quality controlling the power output values of PV systems installed in the same region.

IV. RESULTS

A. DC capacity estimation

Since the reported DC capacity values were deemed unreliable for some of the PV systems, the performance of the DC capacity estimation algorithm is evaluated empirically by comparing the simulated and measured PV power output for clear sky days, see Fig. 10. Where Fig. 4 already demonstrated an example of the ability of the algorithm to estimate the DC capacity by simulating the PV power output for

one PV system (ID128) on a single day, Fig. 10 presents the power measurements (p_m) and simulations (p_s) for the same PV system on all selected clear sky days. The minimal difference (i.e., residuals) between both time series in Fig. 10 proves the ability of the algorithm to estimate the DC capacity. The residuals for all 175 PV systems on the selected clear sky days are summarized in terms of RMSE and bias in Fig. 11. The figure shows an average RMSE of below 4%, and three quarters of the systems obtain an error of under 5%. Also, a mean bias of less than -0.2% is observed.

B. Quality control routine

The results of the quality control routine are summarized in Table III. The table shows the absolute and relative filtered and residual power measurements for each filter applied in the quality control routine. Here, the residual values present the power values in the dataset after each filter in the quality control routine is

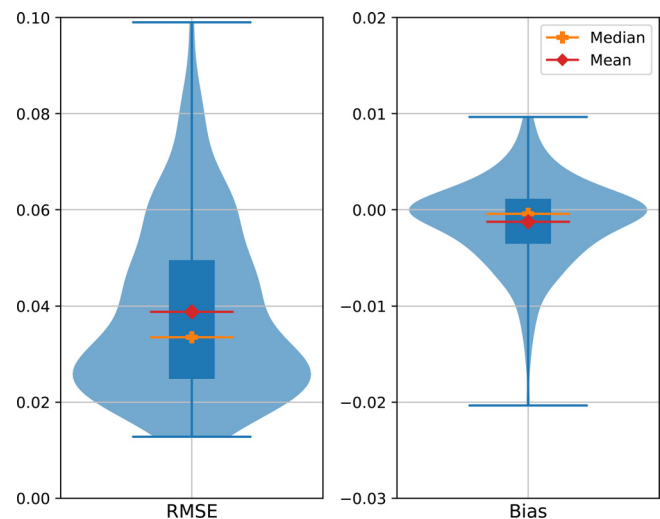


FIG. 11. Violin plot showing the distribution of the PV system DC capacity estimation errors that consider the difference between the simulated and measured power output for all clear sky days. The RMSE and bias are expressed relative to the DC capacity [and calculated by Eqs. (2) and (3)].

TABLE III. Overview of the number of residual and filtered values per filter in the quality control routine. Absolute values consider the sum of all 175 PV systems. Relative values present the mean of the share per system.

Filter	Number of values residual		Number of values filtered	
	Absolute (mln)	Relative (%)	Absolute (mln)	Relative (%)
Summary of unfiltered values				
Unfiltered values total	324.3	100
Unfiltered values true	275.0	85.1
Unfiltered values NaN	49.2	14.9
Single system filters				
Daily availability filter #1	267.4	82.9	7.60	2.28
Night filter	275.3	85.3	-7.90	-2.44
Lower limit filter	271.0	84.0	4.35	1.34
Upper limit filter	269.4	83.5	1.53	0.47
Linear values filter	269.1	83.3	0.38	0.13
Data availability filter #2	266.7	82.6	2.39	0.73
Persistence filter	266.5	82.5	0.16	0.05
Across systems filter				
Daily energy ratio filter	262.4	81.5	4.08	1.12

applied. The filtered values present the values that are flagged and set to NaN or removed per filter. The absolute values present the sum of all 175 PV systems, e.g., in total 4.35×10^6 values are filtered from the dataset in the lower limit filter. The relative values consider the mean of the share of values filtered per system. For example, the daily availability filter 1 is on average responsible for the removal of 2.28% of the power measurements per system. The value associated with the night filter is negative because the night filter sets values to zero, including values that were formerly reported as NaN.

Overall, as discussed in Sec. III B, the quality control routine delivers two filtered datasets. The first dataset concerns the filtered power values after only single system criteria are applied. Up to this point, a total of 10.7×10^6 power measurements were filtered

from the dataset, resulting in a quality-controlled dataset holding 264.3×10^6 power measurements, which is presented in filtered_pv_power_measurements_sc.csv. The second output also concerns the across systems criteria, where an additional 4×10^6 power measurements were filtered and removed from the dataset. Hence, this dataset holds 260.3×10^6 power measurements and is given in filtered_pv_power_measurements_ac.csv. An overview of the published datasets is given in Table IV.

C. Selection of PV systems

The original dataset from which the power measurements of the 175 enclosed PV systems are obtained holds data of 202 PV systems.

TABLE IV. Description of the published files.⁹ The four numerical data files are in comma separated values (csv) format. Missing and filtered data entries are marked as NaN, as discussed in Sec. III B.

File	Description
metadata.csv	Metadata of PV systems (tilt, azimuth, DC and AC capacity, location, begin and end of recording period).
unfiltered_pv_power_measurements.csv	Unfiltered power measurements of all PV systems with 1-min resolution.
filtered_pv_power_measurements_sc.csv	Quality controlled power measurements of all PV systems with 1-min resolution, single criteria (i.e., filters) applied only.
filtered_pv_power_measurements_ac.csv	Quality controlled power measurements of all PV systems with 1-min resolution, single and across criteria (i.e., filters) applied.
qcpv.py	Python package containing all filter functions to perform the quality control routine of the PV power measurements.
example.py	Example Python code to call the qcpv package and use the filter functions.

Yet, 27 of these systems were removed from the dataset for two main reasons. First, systems with data records for less than one year, either before or after applying the quality control routine, were excluded. Second, systems for which the DC capacity estimation algorithm was deemed unreliable were removed. This holds solely for PV systems that experience significant shadowing during the day. An example is presented in [Appendix A](#). These systems were identified through visual inspection.

V. CONCLUSIONS

Public datasets are essential in reaching a mature stage of all research related to PV power output, including but not limited to solar forecasting, model predictive control, PV-battery systems, and simulation of micro and smart-grids. In addition, standardized quality control routines form a fundamental element of progress in all research fields that rely on such data. Considering the minimal availability of public datasets that feature high-resolution PV power measurements, we present an open-source dataset holding 1-min power measurements of 175 PV systems located in Utrecht for a period of four years. In addition, we include an open-source quality control routine that can be applied to filter erroneous PV power measurements in the form of the Python package `qcpv`. The numerical datasets are made available at <https://doi.org/10.5281/zenodo.6906504>,⁹ holding unfiltered power measurements (`unfiltered_pv_power_measurements.csv`), quality-controlled power measurements after single system filters are applied (`filtered_pv_power_measurements_sc.csv`), and quality-controlled power measurements after all filters are applied (`filtered_pv_power_measurements_ac.csv`). These datasets are accompanied by metadata (`metadata.csv`), presenting the system's azimuth and tilt angle, the estimated DC and AC capacity, and location. We also developed and presented a novel approach that estimates the DC capacity of a PV system from power measurements.

VI. RECOMMENDATIONS

We highly encourage researchers to take advantage of the data and code presented in this study and use it in future work. Therefore, we like to inform those interested in using the data on the Dutch climate, interesting periods, data availability, and complementary data. According to the Köppen classification system, the climate in the Netherlands is characterized as oceanic or maritime, which typically feature mild summers and cool winters.²⁶ An optimal PV system is oriented south (i.e., an azimuth angle of 180°) and a tilt angle of 37°.²⁷ Optimally installed PV systems generate most power during April–July. Future work that aims to assess (extreme) PV power ramp rates, cloud enhancement, spatial smoothing, highly variable days, and/or high yield days should consider April and May. These months are particularly interesting as they feature a mix of overcast, highly variable and clear sky days, with high irradiation and compared to summer relative cold temperatures. An overview of the data availability per PV system is shown in [Appendix B](#). The highest availability of data for all systems is found from January 2014 until August 2015. For a substantial amount of systems, data availability is also high from September 2016 until December 2017. Complementary data that may be of interest include weather measurements and forecasts. KNMI^{14,28} provides weather measurements for various locations near the PV

systems including de Bilt and Cabauw. For Cabauw, solar irradiance measurements with a temporal resolution of 1-min are also available.²⁹ ECMWF³⁰ provides weather forecasts and re-analysis data for Utrecht.

ACKNOWLEDGMENTS

This work is part of the Energy Intranets (NEAT: ESI-BiDa 647.003.002) project, which is funded by the Dutch Research Council NWO in the framework of the Energy Systems Integration & Big Data programme. The authors would especially like to thank the PV owners who volunteered to take part in the measurement campaign.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Lennard Visser: Conceptualization (lead); Data curation (equal); Formal analysis (lead); Investigation (lead); Methodology (lead); Software (equal); Validation (lead); Visualization (lead); Writing – original draft (lead); Writing – review and editing (lead). **Boudewijn Elsinga:** Data curation (equal); Writing – original draft (supporting); Writing – review and editing (supporting). **Tarek Alsaikif:** Writing – original draft (supporting); Writing – review and editing (equal). **Wilfried G.J.H.M. van Sark:** Funding acquisition (lead); Writing – original draft (supporting); Writing – review and editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.6906504>, Ref. 9.

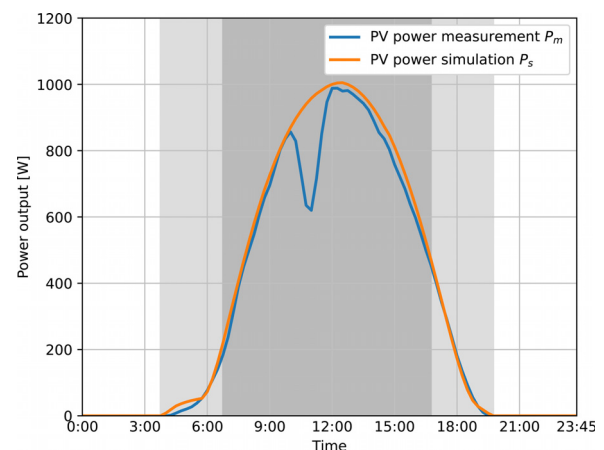


FIG. 12. Unfiltered power measurements and manually adjusted best estimate of simulated clear sky production of a single PV system on July 19, 2016. The PV system was excluded from all datasets, due to shading effects.

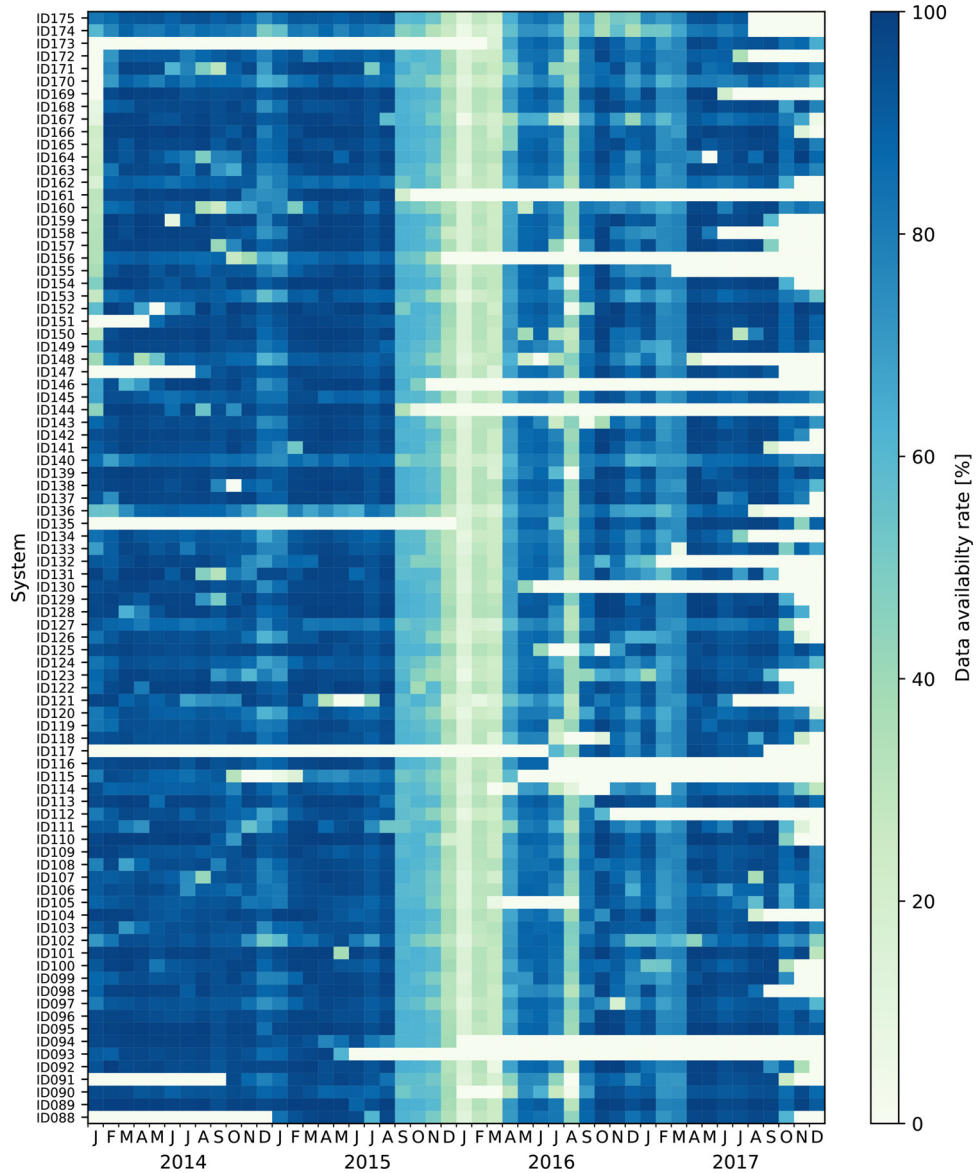


FIG. 13. Data availability rate of the quality-controlled power measurements for PV systems ID088–ID175.

APPENDIX A: EXAMPLE OF PV SYSTEM THAT EXPERIENCES SHADOW

Figure 12 shows the power time series of a single PV system on one of the selected clear sky days, July 19, 2016. The dip in the power output is caused by an object in the immediate vicinity of the PV system, possibly a chimney. The observed power measurements in the time series affect the ability of the algorithm presented in Sec. III A 1 to estimate the system’s DC capacity. Therefore, by means of visual inspection, these systems were removed from the dataset.

APPENDIX B: DATA AVAILABILITY

Figures 13 and 14 present an overview of the monthly availability rate of power measurements per PV system. The data availability rate indicates the share of quality-controlled power measurements (as reported in filtered_pv_power_measurements_ac.csv) during daylight hours as

$$\frac{1}{M} \sum_{t=1}^M p_m(t) \neq \text{NaN},$$

$$M = T \dots \forall t, \text{ where } (t \leq \text{sunrise} + 30 \text{ min} \vee t \geq \text{sunset} - 30 \text{ min}).$$

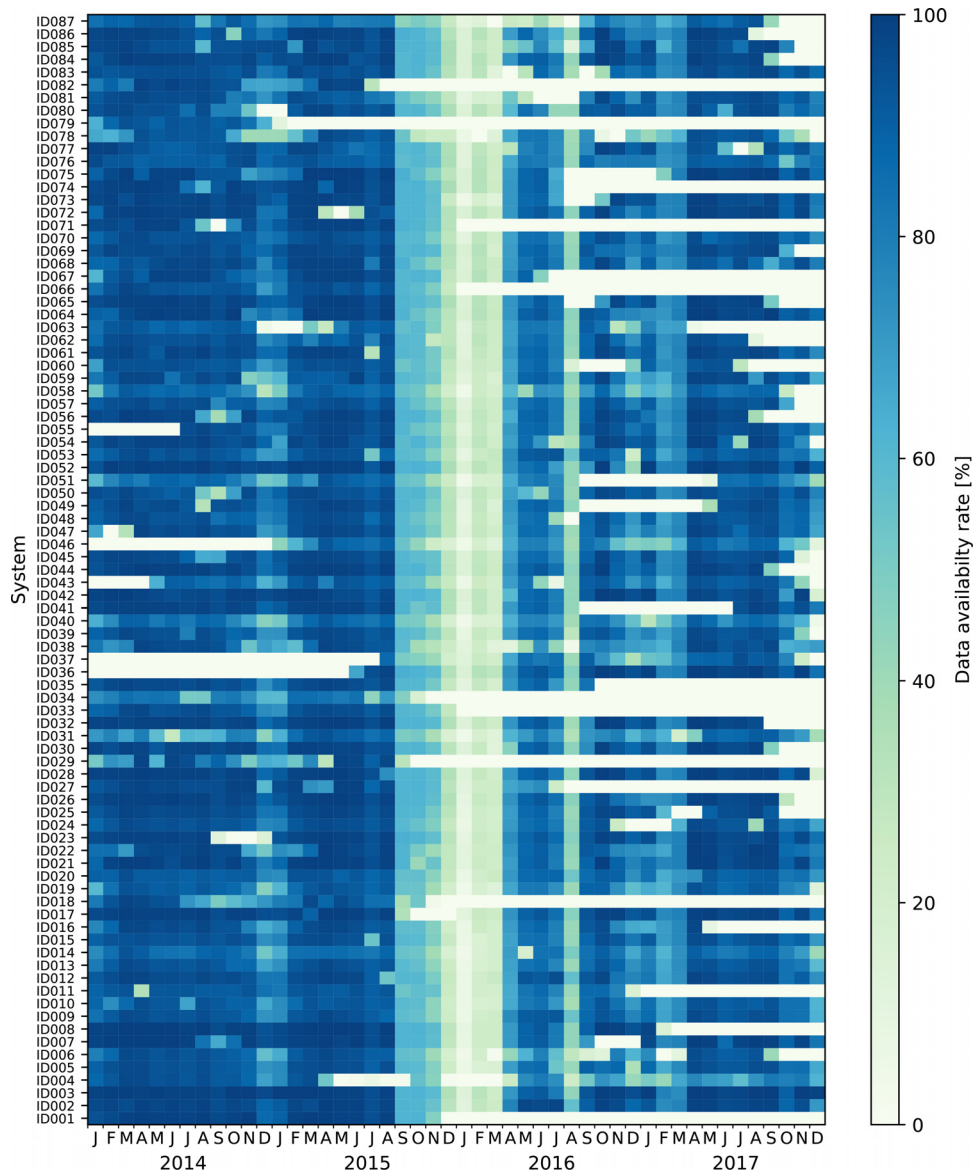


FIG. 14. Data availability rate of the quality-controlled power measurements for PV systems ID001–ID087.

REFERENCES

- ¹L. R. Visser, E. Lorenz, D. Heinemann, and W. G. J. H. M. van Sark, "Solar power forecasts," in *Photovoltaic Technology*, 2nd ed., edited by V. Fthenakis and W. G. J. H. M. van Sark (Elsevier, United Kingdom, 2022), Chap. 11, pp. 213–233.
- ²S. Silwal, C. Mullican, Y.-A. Chen, A. Ghosh, J. Dilliot, and J. Kleissl, "Open-source multi-year power generation, consumption, and storage data in a microgrid," *J. Renewable Sustainable Energy* **13**, 025301 (2021).
- ³J. M. Bright, S. Killinger, and N. A. Engerer, "Data article: Distributed PV power data for three cities in Australia," *J. Renewable Sustainable Energy* **11**, 035504 (2019).
- ⁴S. Kapoor, B. Sturmberg, and M. Shaw, "A review of publicly available energy datasets," *Wattwatchers' My Energy Marketplace (MEM)* (The Australian National University, Canberra, Australia, 2020).
- ⁵S. Lindig, A. Louwen, D. Moser *et al.*, "Outdoor PV system monitoring—input data quality, data imputation and filtering approaches," *Energies* **13**, 5099 (2020).
- ⁶A. Livera, M. Theristis, E. Koumpli, S. Theocharides, G. Makrides, J. Sutterlueti, J. S. Stein, and G. E. Georghiou, "Data processing and quality verification for improved photovoltaic performance and reliability analytics," *Prog. Photovoltaics* **29**, 143–158 (2021).
- ⁷S. Killinger, N. Engerer, and B. Müller, "QCPV: A quality control algorithm for distributed photovoltaic array power output," *Sol. Energy* **143**, 120–131 (2017).
- ⁸R. Urraca, A. Sanz-Garcia, and I. Sanz-Garcia, "BQC: A free web service to quality control solar irradiance measurements across Europe," *Sol. Energy* **211**, 1–10 (2020).

- ⁹L. Visser, B. Elsinga, T. AlSkaif, and W. van Sark (2022). “Open-source quality control routine and multi-year power generation data of 175 pv systems,” Zenodo, V. 0.0.1, Dataset .
- ¹⁰B. Elsinga *et al.*, “Chasing the clouds: Irradiance variability and forecasting for photovoltaics,” Ph.D. thesis (University Utrecht, 2017).
- ¹¹B. Elsinga and W. Van Sark, “Short-term peer-to-peer solar forecasting in a network of photovoltaic systems,” *Appl. Energy* **206**, 1464–1483 (2017).
- ¹²L. Visser, T. AlSkaif, and W. van Sark, “Operational day-ahead solar power forecasting for aggregated PV systems with a varying spatial distribution,” *Renewable Energy* **183**, 267–282 (2022).
- ¹³Article 29 Data Protection Working Party, “Opinion 05/2014 on anonymization techniques (wp216)” (2014); available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.
- ¹⁴KNMI (2021). “Koninklijk Nederlands Meteorologisch Instituut, KNMI,” Uurgegevens van het weer in Nederland, Dataset, <https://www.knmi.nl/nederland-nu/klimatologie/uurgegevens>.
- ¹⁵Upp Energy, “Upp Energy” (2021); available at <http://www.uppenergy.nl/product/details/1f>.
- ¹⁶KNMI (2021). “Koninklijk Nederlands Meteorologisch Instituut, KNMI,” Daggegevens van het weer in Nederland, Dataset, <https://www.knmi.nl/nederland-nu/klimatologie/daggegevens>.
- ¹⁷D. Erbs, S. Klein, and J. Duffie, “Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation,” *Sol. Energy* **28**, 293–302 (1982).
- ¹⁸R. Perez, P. Ineichen, R. Seals, J. Michalsky, and R. Stewart, “Modeling daylight availability and irradiance components from direct and global irradiance,” *Sol. Energy* **44**, 271–289 (1990).
- ¹⁹F. Kasten and A. T. Young, “Revised optical air mass tables and approximation formula,” *Appl. Opt.* **28**, 4735–4738 (1989).
- ²⁰D. L. King, J. A. Kratochvil, and W. E. Boyson, *Photovoltaic Array Performance Model* (Department of Energy, 2004).
- ²¹A. P. Dobos, “PVWatts version 5 manual,” Technical Report No. NREL/TP-6A20-62641 [National Renewable Energy Laboratory (NREL), Golden, CO, 2014].
- ²²W. F. Holmgren, C. W. Hansen, and M. A. Mikofski, “pvlib python: A python package for modeling solar energy systems,” *J. Open Source Software* **3**, 884 (2018).
- ²³P. Ineichen, “A broadband simplified version of the solis clear sky model,” *Sol. Energy* **82**, 758–762 (2008).
- ²⁴F. P. Kreuwel, W. H. Knap, L. R. Visser, W. G. van Sark, J. V.-G. de Arellano, and C. C. van Heerwaarden, “Analysis of high frequency photovoltaic solar energy fluctuations,” *Sol. Energy* **206**, 381–389 (2020).
- ²⁵M. Journée and C. Bertrand, “Quality control of solar radiation data within the RMIB solar measurements network,” *Sol. Energy* **85**, 72–86 (2011).
- ²⁶M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel, “World map of the Köppen-Geiger climate classification updated,” *Meteorol. Z.* **15**, 259–263 (2006).
- ²⁷A. Louwen, R. E. Schropp, W. G. van Sark, and A. P. Faaij, “Geospatial analysis of the energy yield and environmental footprint of different photovoltaic module technologies,” *Sol. Energy* **155**, 1339–1353 (2017).
- ²⁸KNMI (2022). “Koninklijk Nederlands Meteorologisch Instituut, KNMI,” KNMI Data Platform, Dataset, <https://dataplatform.knmi.nl/>.
- ²⁹A. Forstinger, S. Wilbert, A. Jensen, B. Kraas, C. Fernández Peruchena, C. A. Gueymard, D. Ronzio, D. Yang, E. Collino, J. Polo *et al.*, “Expert quality control of solar radiation ground data sets,” in *Proceedings of the ISES Solar World Congress* (ISES Solar World Congress, 2021).
- ³⁰ECMWF (2021). “European centre for medium-range weather forecasts, ECMWF,” ECMWF data archive, Dataset, <https://www.ecmwf.int/en/forecasts/datasets/archive-datasets>.