On the search for representative characteristics of PV systems: Data collection and analysis of PV system azimuth, tilt, capacity, yield and shading

Sven Killinger, David Lingfors, Yves-Marie Saint-Drenan, Panagiotis Moraitis, Wilfried van Sark, Jamie Taylor, Nicholas Engerer, Jamie Bright

On the search for representative characteristics of PV systems Data collection and analysis of PV system azimuth, tilt, capacity, yield and shading

Article *in* Solar Energy · August 2018

DOI: 10.1016/j.solener.2018.08.051

CITATIONS
8

READS
266

8 authors, including:

Sven Killinger
Fraunhofer Institute for Solar Energy Systems ISE
47 PUBLICATIONS **174** CITATIONS
SEE PROFILE

David Lingfors
Uppsala University
33 PUBLICATIONS **289** CITATIONS
SEE PROFILE

Yves-Marie Saint-Drenan
MINES ParisTech, PSL Research University
36 PUBLICATIONS **191** CITATIONS
SEE PROFILE

Panagiotis Moraitis
Utrecht University
14 PUBLICATIONS **44** CITATIONS
SEE PROFILE

On the search for representative characteristics of PV systems: Data collection and analysis of PV system azimuth, tilt, capacity, yield and shading

Sven Killinger[a,b], David Lingfors[c], Yves-Marie Saint-Drenan[d], Panagiotis Moraitis[e], Wilfried van Sark[e], Jamie Taylor[f], Nicholas A. Engerer[a,2,**], Jamie M. Bright[a,*]

[a] *Fenner School of Environment and Society, The Australian National University, 2601 Canberra, Australia*
[b] *Fraunhofer Institute for Solar Energy Systems ISE, 79100 Freiburg, Germany*
[c] *Department of Engineering Sciences, Uppsala University, Lgerhyddsvgen 1, 752 37 Uppsala, Sweden*
[d] *MINES ParisTech, PSL Research University, O.I.E. Centre Observation, Impacts, Energy, 06904 Sophia Antipolis, France*
[e] *Copernicus Institute of Sustainable Development, Utrecht University, 3508 TC Utrecht, The Netherlands*
[f] *Sheffield Solar, University of Sheffield, Hicks Building, Hounsfield Road, Sheffield S3 7RH, UK*

Abstract

Knowledge of PV system characteristics is needed in different regional PV modelling approaches. It is the aim of this paper to provide that knowledge by a twofold method that focuses on (1) metadata (tilt and azimuth of modules, installed capacity and specific annual yield) as well as (2) the impact of shading.

Metadata from 2,802,797 PV systems located in Europe, USA, Japan and Australia, representing a total capacity of 59 GWp (14.8% of installed capacity worldwide), is analysed. Visually striking interdependencies of the installed capacity and the geographic location to the other parameters tilt, azimuth and specific annual yield motivated a clustering on a country level and between systems sizes. For an eased future utilisation of the analysed metadata, each parameter in a cluster was approximated by a distribution function. Results show strong characteristics unique to each cluster, however, there are some commonalities across all clusters. Mean tilt values were reported in a range between 16.1° (Australia) and 35.6° (Belgium), average specific annual yield values occur between 786 kWh/kWp (Denmark) and 1,426 kWh/kWp (USA South). The region with smallest median capacity was the UK (2.94 kWp) and the largest was Germany (8.96 kWp). Almost all countries had a mean azimuth angle facing the equator.

PV system shading was considered by deriving viewsheds for ≈ 48,000 buildings in Uppsala, Sweden (all ranges of solar angles were explored). From these viewsheds, two empirical equations were derived related to irradiance losses on roofs due to shading. The first expresses the loss of beam irradiance as a function of the solar elevation angle. The second determines the view factor as a function of the roof tilt including the impact from shading and can be used to estimate the losses of diffuse and reflected irradiance.

Keywords: PV system characteristics, Metadata, Shading, Data analysis

1. Introduction

With 402.5 GW of installed photovoltaic (PV) capacity globally (IEA, 2018), the integration of the large amounts of energy generated by the numerous distributed solar power systems into the electricity supply system is an issue ever gaining in importance. Modelling of the power generated by those decentralised solar systems is of utmost importance for several issues ranging from energy trading to network flow control. The estimation and forecast of PV power is made difficult by the fact that only a minority of systems continuously report their generation and are publicly accessible.

Different strategies have been proposed to overcome the lack of reporting (e.g. upscaling approaches or power simulations based on satellite derived irradiance); an extensive literature overview is provided in Bright et al. (2017b). Within this paper, the estimation of the aggregated power generated in a given region by a fleet of unknown PV systems is referred to as regional PV power modelling. Knowledge of PV system characteristics is required in the different regional PV modelling approaches to reconstruct the missing power measurements (Lorenz et al., 2011; Saint-Drenan et al., 2016). Some studies assign simplified assumptions of the PV system characteristics. This can result in over-exaggerated grid impacts (Bright et al., 2017a). Unfortunately in most cases, characteristics from PV systems are either unknown or only accessible for a small number of stakeholders (inverter manufacturers, monitoring solutions providers, etc.). As a result, progress in the area of regional PV power estimation or forecasting can be considered sub-optimal as potential contributors like universities or small companies are partially excluded from access to larger datasets of measurements or metadata. This is still the case despite grid integration of solar energy being considered a strategic societal issue. Therefore, it is the aim of this paper to offer any stakeholders the possibility to develop activities on this research field by collecting, analysing and disseminating metadata on millions of PV systems installed worldwide. To begin, we must establish which metadata are the most important.

Saint-Drenan (2015) carried out a sensitivity analysis and found that the four most influential characteristics impacting PV output generation are: (1) tilt angle and (2) azimuth angle of PV modules, (3) installed capacity and (4) total efficiency (represented herein as the specific annual yield). Furthermore, (5) shading is of crucial influence on the PV power generation but is not accessible from PV system metadata. The impact of shading can only be accessed with considerable effort, e.g. simulations that consider digital elevation models (DEM) including buildings, trees and other obstacles, by analysing PV power profiles or even weekly performance ratios (see Paulescu et al. (2012); Freitas et al. (2015); Lingfors et al. (2018); Tsafarakis et al. (2017) for further reading). Due to its significant influence, a shading analysis complements the focus of this study.

These five identified characteristics are the cen-

*Corresponding author
**Co-corresponding author
Email addresses: nicholas.engerer@anu.edu.au (Nicholas A. Engerer), jamie.bright@anu.edu.au (Jamie M. Bright), jamiebright1@gmail.com (Jamie M. Bright)

tral focus of this paper because of their general importance for regional PV modelling approaches. The overall aim of this paper is to achieve a full reproducibility of the five characteristics so that they can be used in regional PV power modelling applications such as nowcasting or forecasting, but also in power simulations that are used for energy system analysis, studying the grid impact, defining the PV power potential etc.

1.1. Related work

The relevant literature for this research has three prominent categories: (1) metadata analysis with intention to improve regional PV power simulations, (2) PV performance due to specific yield, and (3) models that consider shading analysis.

Category 1: Examples of literature using metadata to improve regional PV power simulations. Schubert (2012) provides a useful guidebook for the simulation of PV power that sketches important parts of the simulation chain and delivering assumptions for characteristics. An overview of different characteristics of tilt, azimuth, the module and installation type are given together with suggested weights. However, these weights seem to be assumptions with no datasets being cited as an empirical basis and so using these weights in PV simulations raise questions of trust.

Datasets are used by Lorenz et al. (2011), who evaluated the representativeness of a set of reference PV systems to predict regional PV power by analysing the orientation and module types of $\approx 8,000$ systems in Germany. The authors note that their dataset seem to have a disproportionate share of large PV systems and so do not fully represent a larger portfolio.

The problem of poor representativeness was bypassed in Saint-Drenan (2015); Saint-Drenan et al. (2017) by feeding a PV model with metadata statistics from a larger sample of PV systems as opposed to a smaller and unrepresentative subset. They derived joint probabilities of azimuth and tilt from 35,000 systems and clustered them by their system size and geographic location. These empiric distributions where then used to estimate the characteristics of all 1,500,000 PV systems installed in Germany at that time. Saint-Drenan et al. (2018) complemented their earlier research by reproducing it for more European countries using statistical distributions from 35,000 PV systems in Germany and 20,000 in France. This demonstrates the significant potential of generating representative statistical distributions with intended use in regional PV power simulations.

Kühnert (2016, pp. 80-85) followed a similar approach and derived statistical distributions for tilt and azimuth from $\approx$ 1,300 PV systems in Germany. Based on this portfolio, the author evaluated the representativeness should PV systems be clustered into different geographic regions and system sizes. The authors quantitatively derived recommendations between the two extremes of (1) a portfolio covering all PV systems and (2) a high number of subclasses with a very small number of PV systems. From this, we observe that there must be a well considered clustering approach in order to derive representative subclasses.

Killinger et al. (2017c) detailed a regional PV power upscaling approach which estimated the

power of ≈ 2,000 target PV systems based on 45 continuously measured PV systems in Freiburg, Germany. Whereas the azimuth and tilt of the 45 measured systems were known in their case, both parameters were derived through a geographic information system (GIS) based approach for the target PV systems.

Furthermore, Pfenninger and Staffell (2016) use PV power measurements and incorporate metadata from 1,029 systems in 25 European countries to derive empirical correction factors for PV power simulations. A comparison between the analysed tilt and latitude showed an trend towards steeper angles at higher latitudes, indicating that metadata might vary with the geographic location.

PV system metadata is thus used to successfully improve regional PV power upscaling across Europe in Pfenninger and Staffell (2016); Killinger et al. (2017c); Saint-Drenan (2015); Saint-Drenan et al. (2017, 2018); Kühnert (2016). These works applied information of azimuth, tilt, installed capacity and the geographic location from PV systems to estimate the power output of a larger PV fleet for similar geographies and different countries. They stand as an powerful and excellent example for how representative metadata distribution statistics can be employed. It is these examples that guide the first usage of our vast dataset towards deriving representative metadata distributions.

Category 2: Excerpts of literature that analyse the performance of PV systems. Performance is more complex than just tilt and azimuth as it is inherently influenced by other components, such as soiling and meteorology.

Nordmann et al. (2014) found a positive correlation between specific annual yield and incoming irradiance, as well as an observed negative correlation between system performance and ambient temperature. Their data was obtained via web-scraping of Solar-Log (2,914 systems in the Netherlands, Germany, Belgium, France and Italy) and collected by participants of the IEA task (>60,000 systems in the USA).

Moraitis et al. (2015) observed an increasing yield with decreasing latitude from ≈ 20,000 systems in Netherlands, Germany, Belgium, France and Italy, also achieved using web-scraping techniques. We therefore expect to observe geographical differences due to latitude and climate.

Taylor (2015) explored the generation of 4,369 distributed systems in the UK to derive the performance ratio and degradation rate. To allow reproducibility, the analysis of the performance ratio was enriched by approximating it with distribution functions. We intend to extend this style of analysis to PV system metadata.

Leloux et al. (2012a) examined data from residential PV systems in Belgium; Leloux et al. (2012b) focused on France. In Belgium, specific annual yield was analysed for 158 systems in 2009 and normalised by a factor which compared the incoming irradiance in this year to a 10 year average. The mean value was 836 kWh/kWp. The same approach led to a mean value of 1,163 kWh/kWp for 1,635 systems in 2010 in France. Weibull distributions were used throughout both papers to approximate the specific yield and performance indicators; Weibull distributions were selected for visual similarity and not for robustness of fit — we aim to use a more statistically rigorous approach to distribution

type selection. Furthermore, a relative distribution was provided for combinations of tilt and azimuth. Additionally, the installed capacity was analysed in France, showing a high number of systems with 3 kWp or slightly less. The reason for this is due to tax credits being denied for system sizes > 3kWp and a strongly increased VAT for such system sizes (Leloux et al., 2012b). The legal framework can thus have a strong influence on characteristics of PV systems. Further studies exist which analyse the specific energy of PV systems. However, most of these studies are limited to a particular region and less of them propose a parametric approximation of the data studied.

Category 3 — the impact of shading in many articles is only considered in a highly simplified manner, e.g. by setting irradiance values zero above a certain solar zenith angle (Lingfors and Widén, 2016), restricting simulations and analyses to time steps with certain solar zenith angles (Elsinga and van Sark, 2015; Elsinga et al., 2017; Jamaly et al., 2013; Killinger et al., 2016; Saint-Drenan et al., 2017; Yang et al., 2014; Bright et al., 2015), applying constant losses (Mainzer et al., 2017) or assuming a linear decrease in the PV power values (Schubert, 2012). Several authors expect improvements in their results, when the influence of shading is better represented (Bright et al., 2017a,b; Pareek et al., 2017)

1.2. Contribution

Considering the lessons and outcomes of the different studies described in our literature review, we see a clear need for the production of a representative set of distributions to appropriately represent PV system metadata. Currently, further advancements in regional PV power models in the absence of significant knowledge of metadata is hindered due to several reasons Firstly, implementation is hindered due to lack of access to PV system datasets. Empirically derived distributions of these PV system parameters could replace this need, though currently are only provided for performance indicators (Taylor, 2015) and the specific annual yield (Leloux et al., 2012a). Secondly, with exception of a few studies (e.g., Saint-Drenan et al. (2015)), the issue of sample representativeness is often omitted. This is a major omission, for example, a studied dataset including a majority of roof-mounted PV system has to be generalised in order to represent a fleet of systems encompassing a lot of rack-mounted PV systems. Thirdly, most of the identified studies focused on particular PV system characteristics; an integrated analysis encompassing all five key characteristics is required. Furthermore, the influence of shading is in most articles excessively simplified or more commonly excluded. Lastly, studies are mostly limited to a specific country and it is currently difficult to make comparisons between countries to assess applicability. A holistic overview of important parameters of metadata for multiple countries is clearly missing.

The objective of this paper is to address the aforementioned limitations by following the goals below:

1. To collect and process as many data sources as feasible of four identified key metadata parameters (tilt, azimuth, installed capacity and specific annual yield) for PV systems installed worldwide (section 2),

2. To explore the characteristics of these key parameters and their associated interdependencies (section 3.1),
3. To propose a a clustering approach to allow representative generalisation of our datasets (section 3.2),
4. To provide an eased access to the characteristics of each key parameter by fitting distribution functions to the observed probabilities (section 4),
5. To propose a method that evaluates the impact of shading (section 5.1) and which derives generalised findings for improved consideration and implementation (section 5.2).

The influence of meteorological conditions, panel degradation and soiling are not considered within this research, beyond those losses that are inherently and statically contained within the specific annual yield. Whilst they are highly interesting topics and research avenues that could be explored, we are more keenly interested in comparisons and parametrisations of PV system metadata and reserve such topics for future research, more ideas of which are presented in section 6. A summary of the paper is then given in section 7. In the Appendix A, the forms of the distributions used in this paper are defined and their fitted variables provided.

2. Collection and processing of PV system metadata

An intensive effort has been conducted to identify, collect and prepare good sources of PV system metadata. Some of the major monitoring companies and inverter manufacturers have been contacted. In parallel, free information on several solar portals have also been used to gather our dataset either by downloading or web-scraping techniques. Ultimately, we obtained a dataset containing 2,802,797 PV systems located in Europe, USA, Japan and Australia, which represents a total capacity of 59 GWp (14.8% of installed capacity worldwide). Every system in our records reported an installed capacity. However, the other parameters were not always reported. The systems in our database that reported a valid tilt/azimuth only have a relative share from the worldwide installed capacity of 1.7%. Geographic position was almost as often reported as installed capacity and the relative share is 14.5%. The specific annual yield has a relative share of 11%. Further detail of the parameter shares and subsequent quality filtering are found in Table A.5.

An overview of the regions covered by our study, the characteristics of the datasets and their sources are provided in Table 1. For some countries, data is derived from multiple sources. It shouldn't be ruled out that systems could be listed multiple times, leading to duplicates in the analysis. Due to the nature of reporting, a single PV system may not have the same metadata in different datasets and so it is accepted that this is an inherent error. The inhomogeneous nature of the datasets motivated us to apply some preprocessing operations to ensure that only valid system measurements are considered in our analysis and all datasets are in a consistent format. Some of these operations act as quality filters. They were developed based on our empiric experiences with the datasets and are shortly justified where presented.

Table 1: Regions, parameters and data sources. "Rest of Europe" contains different European countries not already listed with less than 1,000 systems each. The cumulated capacity is given in MWp and, where available, as a relative share of the total installed capacity in a region (own calculations based on IEA (2018) with data from 2016 and National Grid UK (2018) in case of UK with data from 2018).

Region –	No. systems –	Tilt & azi. (°)	Capacity (kW/kWp)	Spec. ann. yield (kWh/kWp)	Cumulated Capacity (MWp) / % of total	Source –
Australia	4,055	✓	✓	×	30 / 0.42	pvoutput.org
Austria	385	✓	✓	2012-2016	4 / 0.33	solar-log.com
	280	✓	✓	×	2 / 0.14	suntrol-portal.com
	268	✓	✓	2015-2017	2 / 0.17	sonnenertrag.eu
	112	✓	✓	2015-2017	1 / 0.04	pvoutput.org
Belgium	4,535	✓	✓	2012-2016	149 / 3.93	solar-log.com
	3,365	✓	✓	2015-2017	17 / 0.45	bdpv.fr
	541	✓	✓	2015-2017	12 / 0.32	sonnenertrag.eu
Denmark	933	✓	✓	2012-2016	7 / 0.80	solar-log.com
	630	✓	✓	×	4 / 0.42	suntrol-portal.com
	542	✓	✓	2015-2017	2 / 0.27	pvoutput.org
France	20,935	✓	✓	2015-2017	93 / 1.17	bdpv.fr
	558	✓	✓	2012-2016	8 / 0.10	solar-log.com
Germany	1,664,967	×	✓	2012-2016	41,478 / 98.76	bundesnetzagentur.de
	23,536	✓	✓	2012-2016	547 / 1.30	solar-log.com
	6,561	✓	✓	×	124 / 0.29	suntrol-portal.com
	6,447	✓	✓	2015-2017	112 / 0.27	sonnenertrag.eu
Italy	2,506	✓	✓	2012-2016	30 / 0.15	solar-log.com
	1,068	✓	✓	2015-2017	9 / 0.04	pvoutput.org
	358	✓	✓	2015-2017	11 / 0.06	sonnenertrag.eu
Japan	5,233	✓	✓	2012-2017	42 / 0.09	jyuri.co.jp
Netherlands	7,180	✓	✓	2015-2017	31 / 1.08	pvoutput.org
	1,115	✓	✓	2012-2016	14 / 0.49	solar-log.com
	917	✓	✓	2015-2017	9 / 0.31	sonnenertrag.eu
	290	✓	✓	×	2 / 0.07	suntrol-portal.com
Rest of Europe	1,191	✓	✓	2012-2016	23 / –	solar-log.com
	566	✓	✓	2015-2017	5 / –	pvoutput.org
	133	✓	✓	2015-2017	2 / –	sonnenertrag.eu
UK	18,543	✓	✓	2015-2016	58 / 0.45	microgen-database.sheffield.ac.uk
	2,286	✓	✓	2015-2017	9.5 / 0.07	pvoutput.org
USA	1,020,585	✓	✓	2017	16,521 / 32.39	openpv.nrel.gov
	2,176	✓	✓	2015-2017	17 / 0.03	pvoutput.org

Longitude and latitude: In cases where this information was not provided, the geographical coordinates were derived from OpenStreetMap using other given information such as the zip-code, city name, state name, etc. Erroneous locations outside the specific region are set NA. For confidentiality reasons geographic information was provided separately from the other parameters in case of the 18,543 systems from Sheffield Solar. The derived longitude and latitude are not required to be highly accurate to suit the needs of this paper as they are purely used for trend analysis when studying rough relationships to other parameters and for visualization purposes. The ability to allocate a PV system to a specific country is certain in all cases.

Tilt and azimuth: Unfortunately, this important metadata is not available from all sources. In case of Australia the provided data was imprecise (45° steps in the azimuth) and thus estimated by the approach described in Killinger et al. (2017b) and improved in Killinger et al. (2017a) as the PV power data was available. As Australia is on the southern hemisphere, azimuth angles were transformed to normalise the angles expressed for both hemispheres. Within this paper, we consider −90° to be east, 90° to be west, with 0° representing south in the Northern hemisphere and north in the Southern hemisphere. Multi-array systems are not considered in this paper. In a few of the listed datasets, an excessive amount of tilt values with 0° or 1° and azimuth values of −180° are reported. E.g. the Australian dataset reported 36% of all systems having a tilt angle $\leq 1°$. Visual inspections based on aerial images and results from the aforementioned parametrisation, however, showed that such small tilt angles were very rare and regularly incorrectly reported. From previous work with various datasets, we know that such boundary values are sometimes used as a default when data is missing. As we have no quality control measures on the data, the validity of the data at these boundary values is in question and so are removed from consideration. Tilt $\leq 1°$ or $> 89°$ and azimuth $< -179°$ or $> 179°$ are thus set NA.

Specific annual yield: There are many instances of systems reporting a specific annual yield of 0 kWh/kWp. Without further information from the datasets, it is not possible to distinguish whether this is a default value for missing data or a valid measurement. We expect that both cases regularly occur and so we must remove any input of 0 kWh/kWp from our analysis. Furthermore, the specific yield of a system is set NA if it was installed within the year of consideration to ensure that a full year of generation is the basis for the annual yield. In order to compensate annual meteorological fluctuations within a dataset of a country, all values within a year are divided by the ratio between the mean value from all systems in this year and the average of the mean values from all reported years. A similar approach is applied in Leloux et al. (2012a). In datasets reporting a continuous time series, the specific annual yield was derived by the summation of the normalised PV power values. Only systems which have less than 10 days / $\approx 2.7\%$ of missing time steps in their generation data are considered. The vast minority of systems in the datasets reported specific annual yield values that significantly exceed any meteorological potential. We believe that such values are either erroneous reports

of either yield or the installed capacity, as the latter is used in some datasets to derive the former through division. Whereas Taylor (2015) applied a statistical based upper limit for outliers, a fixed limit of 2,000 kWh/kWp was used in this paper. The fixed value was chosen to ensure a reliable filtering even though the quality and range of values may differ for the various datasets. A threshold value of 2,000 kWh/kWp acknowledges the increasing risk of erroneous data beyond this value and is a very cautious limit with the aim of avoiding any erroneous filtering. In fact, this limit was only exceeded in 2.65% of all systems that reported a yield value from openpv.nrel.gov where we observed the largest values within the study and only for 0.084% of all systems in this study.

Please note that, regarding the installed capacity, no pattern was recognised that led us to believe that there were any systematic quality issues. The same applies for the other parameters that were only available for some datasets, such as information about the network connection for Germany as visualised in Figure 2. Hence, the data was taken on an as-is basis in these cases.

A summary of the impact of our proposed quality control criteria is provided in the appendix in Table A.5 where percentages of removed data are presented. Data from pvoutput.org were strongly affected by the filtering of the low tilt values and justify the need of such a quality control. There is a significantly higher share of systems filtered by $< -179°$ when compared to the filter for azimuths $> 179°$. This is because south is defined as 180° in some datasets and are therefore transformed by subtracting 180°. Invalid entries in the same datasets were defined as 0° and subsequently filtered post-transformation by the lower threshold value for azimuth values. Insufficient information was given to derive the exact location for PV systems, mostly from pvoutput.org. All valid parameter entries that have passed the quality control are used for the analysis in the next two sections.

3. Analysis of PV system metadata

3.1. Analysis of parameters and dependencies

The datasets presented in section 2 are very inhomogeneous with large differences in the number of systems in each region and the availability of parameters. Before starting to explore individual clusters, typical ranges of these parameters and potential dependencies between them shall be studied on a global dataset. The general principle of the global dataset is that every region has the same weight. Consider Table 1, should all data be used to make global statistics, the results would be biased towards the countries with more data (USA and Germany). Therefore, a normalisation method must be employed to weight countries equally. Its derivation follows the following procedure:

(1) Specific annual yield is the only parameter that exists multiple times for each PV system. To evenly weight all systems, only one normalised specific yield value per system is considered by randomly selecting a year. This procedure was preferred to others such as e.g. taking the mean value for all values of a system in order to conserve system specific variability between years. (2) For each combination of two parameters (e.g. tilt and specific annual yield) the algorithm counts the number

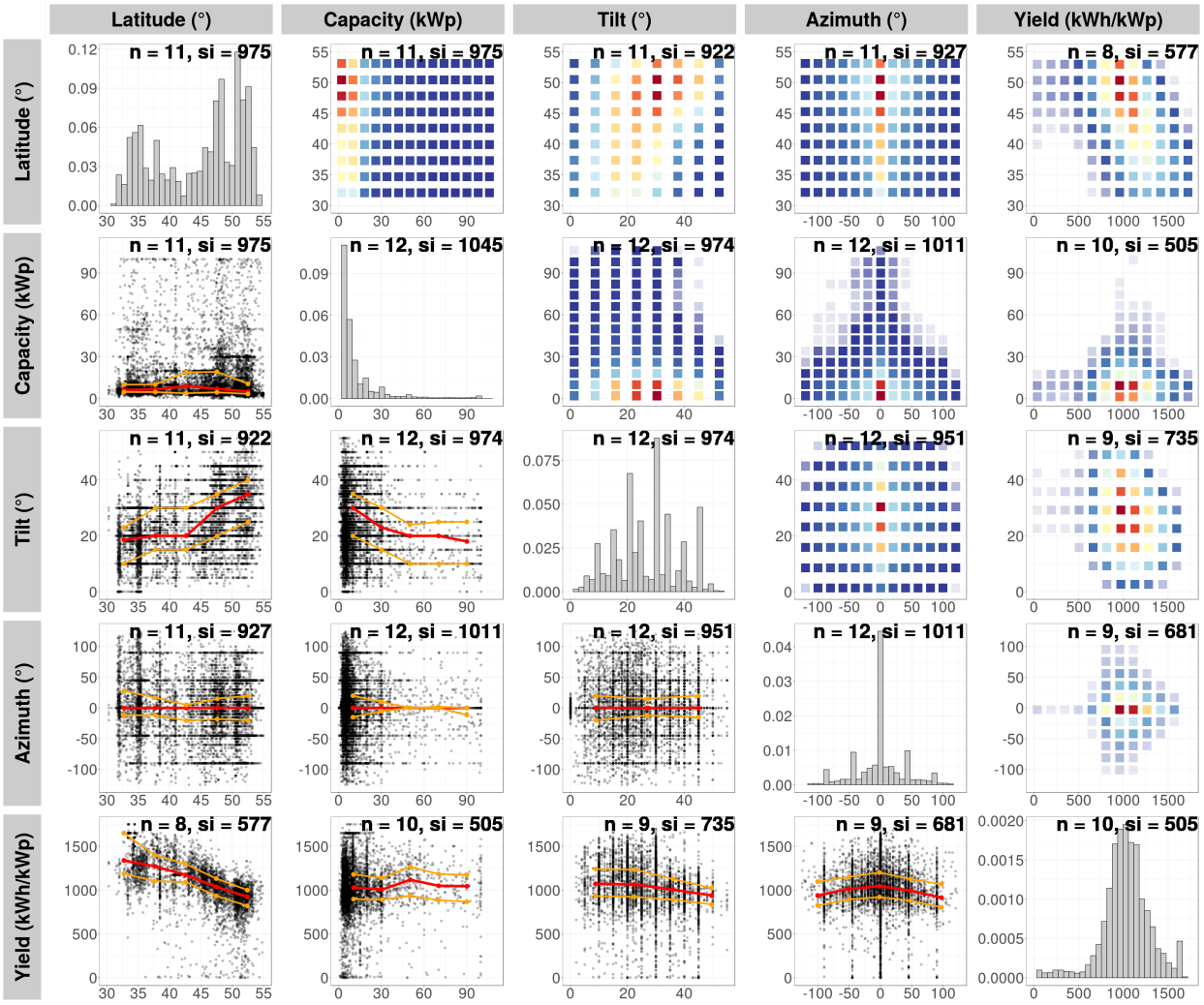

Figure 1: Hybrid graphic with plots of the different parameter pairs from the global dataset. The plots below the diagonal are scatter plots with the 25%, 50% and 75% quantiles as coloured lines. Plots on the diagonal are 1D-histograms of that parameter. Plots above the diagonal are 2D-histograms of the parameter pairs; the change in colour from white to red is an indication of probability and its distribution. The 2D-histograms and scatter plots have the parameter of their column on the x-axis and the parameter of their row on the y-axis. Note that each scatter and 2D-histogram pair have opposite axes but are identical data. 1D-histograms are the only exception with having displayed the density on the y-axis. For reasons of simplification the absolute value of the latitude is taken in these plots to make results from the northern and southern hemisphere comparable. The bold number in each plot shows the number of countries (**n**) which are considered in the plot as well as the sampling size (**si**).

of couples per region that have valid reports in both parameters that have passed the quality control in section 2. Only regions with a sample size of at least 500 complete couples are considered to ensure statistical relevance. (3) The smallest number of complete couples from all regions is taken to define the sample size for the global dataset. This way, the same number of complete couples is taken

from each region. Therefore, all of the data is considered for the region with the smallest number of complete couples. In all other regions, the same number of couples is randomly selected. To avoid under-representation of larger systems, the selection probability is linearly weighted with installed capacity, not frequency.

The significant advantage of this procedure is that regional characteristics are evenly weighted and the availability for each pair of parameters is individually considered. The disadvantage is that many systems are randomly banned due to the region with least availability. We applied different methods of sub-sampling the data, however, the resulting global data was quite insensitive to different sampling procedures indicating the robustness of our approach.

Results from the global dataset are displayed in Figure 1 and the following observations can be made:

Latitude: To have a robust quantity of data, PV systems in latitudes between 30° and 55° are studied. Latitude does not show any obvious influence to the installed capacity or azimuth angle. The tilt angle shows a tendency to increase with an increasing latitude, corroborating the same observation by Pfenninger and Staffell (2016) between the latitude ranges in the study. This finding agrees with studies showing that systems should have a smaller tilt closer to the equator in order to optimise their annual yield (Šúri et al., 2007). It is still surprising since many systems in our analysis are installed on roofs and strongly depend on the roofs' inclination. It can be suggested that the roof pitch has a tendency to be steeper at higher latitudes in our datasets and agrees with similar observations in Europe (McNeil, 1990, p. 883). Furthermore, a linear decline in the specific annual yield is observed with an increasing latitude. This occurs in accordance to the tendency of a higher solar potential in regions closer to the equator (Šúri et al., 2007).

Installed capacity: Within the plot, only systems < 100 kWp are displayed for ease of visualisation. Even though the sampling weights in preference of larger systems, there is a clear concentration of smaller system sizes. There is a visual trend towards smaller tilts with an increasing installed capacity. Furthermore, there is a clear observation that larger capacity systems are consistently oriented towards the equator whereas smaller systems have a much broader range of orientation. A dependency between the installed capacity and the specific annual yield cannot be observed in the global dataset. Despite that, we would expect that the efficiency of larger systems is usually higher and systems better maintained. Most likely, this trend is invisible here since data from many geographic regions were sampled. This hypothesis is checked in section 4. The finding that PV system size has interdependencies on the other parameters can be reaffirmed with everyday observations; smaller systems are in most cases mounted on the roof of residential buildings, medium systems are typically found on farming houses or industrial buildings, and large systems are mounted on a rack on the ground.

Tilt: Tilt in the dataset mainly occurs in a range up to 50° and is often reported in steps of 5°, though reporting steps of 10° are also common. No discernible trend between tilt and azimuth is ob-

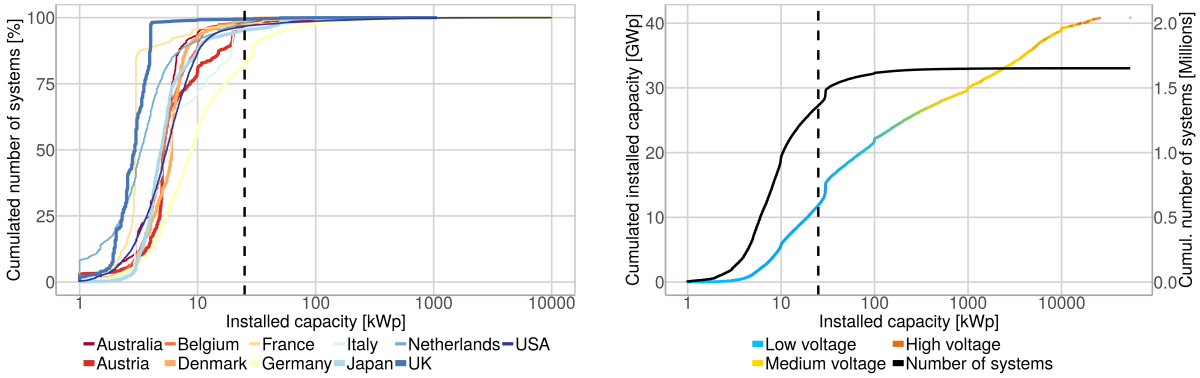

Figure 2: The installed capacity and its relationship to the relative share of systems for different countries (left). The line width and colours vary to simplify the differentiation. The cumulative installed capacity in case of Germany is shown in the right plot represented by the coloured line (colouration indicating the network connection) whereas the black line represents the cumulative number of systems. The dashed line indicates 25 kWp, which is used to sub-categorise the data in section 3.2.

served, however the 2D-histogram shows a significant density peak around the most frequent combination of azimuth and tilt with a radially decreasing probability, this was also observed by Saint-Drenan (2015); Killinger et al. (2017c). A decrease in the specific annual yield can be seen with an increasing tilt. This might be caused by the finding that tilt is usually smaller for decreasing latitudes which occur in combination with an increased specific annual yield.

Azimuth: There is a significant peak of azimuth angles pointing south (north in Australia). It is probable that this distinct peak is due to the targeted approach of solar installers who favour equatorial orientated rooftops due to performance benefits. Indeed, azimuth angles tend towards reaching a higher specific annual yield with systems oriented towards 0°. In general, outliers reach a range of +/- 100° with discrete reporting intervals being visible in the 1D-histogram and scatter plot, e.g. databases only requiring azimuth reported to nearest 15°.

Specific annual yield: The 1D-histogram of yield shows the most distinct shape of all parameters with a peak around 1,000 kWh/kWp. Furthermore, there is a small peak at 1,650 kWh/kWp which is caused by PV systems in southern regions of the USA. It is not possible with the limited latitude study area to infer that the regression of specific yield with latitude will extend towards the equator; climatic regions are expected to be far more influential on the specific yield whereby around the equator there is a significant presence of clouds, and around the tropics there tends to be desert. It is probable that the secondary peak above 1,650 kWh/kWp is for systems installed in particularly arid regions found in southern USA, however, climatic influence is outside the scope of this paper and is reserved for future study. The specific annual yield has the most visually recognisable trends to all other parameters, demonstrating the strong inter-relationship. There is a need for a more de-

tailed multi-variable analysis between specific annual yield and the other parameters. However, due to its extra complexity it falls outside the scope of this paper.

3.2. Representativeness of clusters

In the previous section, important characteristics of PV systems and their dependencies were derived. With exception of the annual specific yield and installed capacity in Germany, metadata of all the installed systems within the different regions is not known (e.g. we have access to 4,055 systems from Australia when there are an estimated 1.8 million installed). This restriction questions the representativeness and re-usability of our observations when using the statistics of a subset of systems to infer the statistics of the remainder because some characteristics could be over- or underrated in our datasets. To achieve representation, a solution is to sub-categorise metadata from the PV systems into smaller and more homogeneous clusters. By doing so, an end-user can use the statistics of the clusters and weight them individually by the probability of occurrence. Prior to an approximation of metadata in section 4, it is the objective within this section to define groupings or clusters of PV system that allow the derivation of representative characteristics.

The interdependency analysis reveals two dominating parameters which show multiple dependencies to others: (1) The installed capacity and (2) a geographical influence (c.f. absolute latitudes are used to account for hemispheres). These two findings are in accordance with Kühnert (2016); Saint-Drenan (2015); Saint-Drenan et al. (2017) who analysed azimuth and tilt for different classes of installed capacity and multiple regions. Such a separation has the benefit to acknowledge the impact from these two dominating parameters on others, while still allowing us to derive meaningful statistics within a chosen cluster. As Kühnert (2016) evaluates, a balance must be found between the number and size of the clusters, in order to guarantee that each class includes a sufficient number of data to be representative.

The left plot in Figure 2 provides further insights into the system size and its relative share for different countries in this paper. Differences can be observed between countries but all show a heightened concentration towards small scale systems with a relative share between 60% (Germany) and 99% (UK) of systems < 10 kWp. Whereas most datasets only cover a selection of systems within a country, the dataset in case of Germany (bundesnetzagentur.de) covers the vast majority of systems and is detailed on the right plot. Almost one million out of the 1.6 millions German systems are smaller than 10 kWp but in total, with an aggregated capacity of $\approx$ 5 GWp, they only represent $\approx$ 12% of the installed capacity. Another 650,000 systems occur in a range between 10 and 100 kWp and cover additional $\approx$17.5 GWp. Only 35,000 systems are $>$ 100 kWp yet are responsible for half of the total installed capacity.

On the search of a threshold value to split the datasets into representative clusters, a system size of 25 kWp was chosen by considering: (1) An installed capacity of 25 kWp is an adequate size between typical roof mounted systems and larger plants, particularly as larger capacities are linked to larger physical space requirements. (2) So even

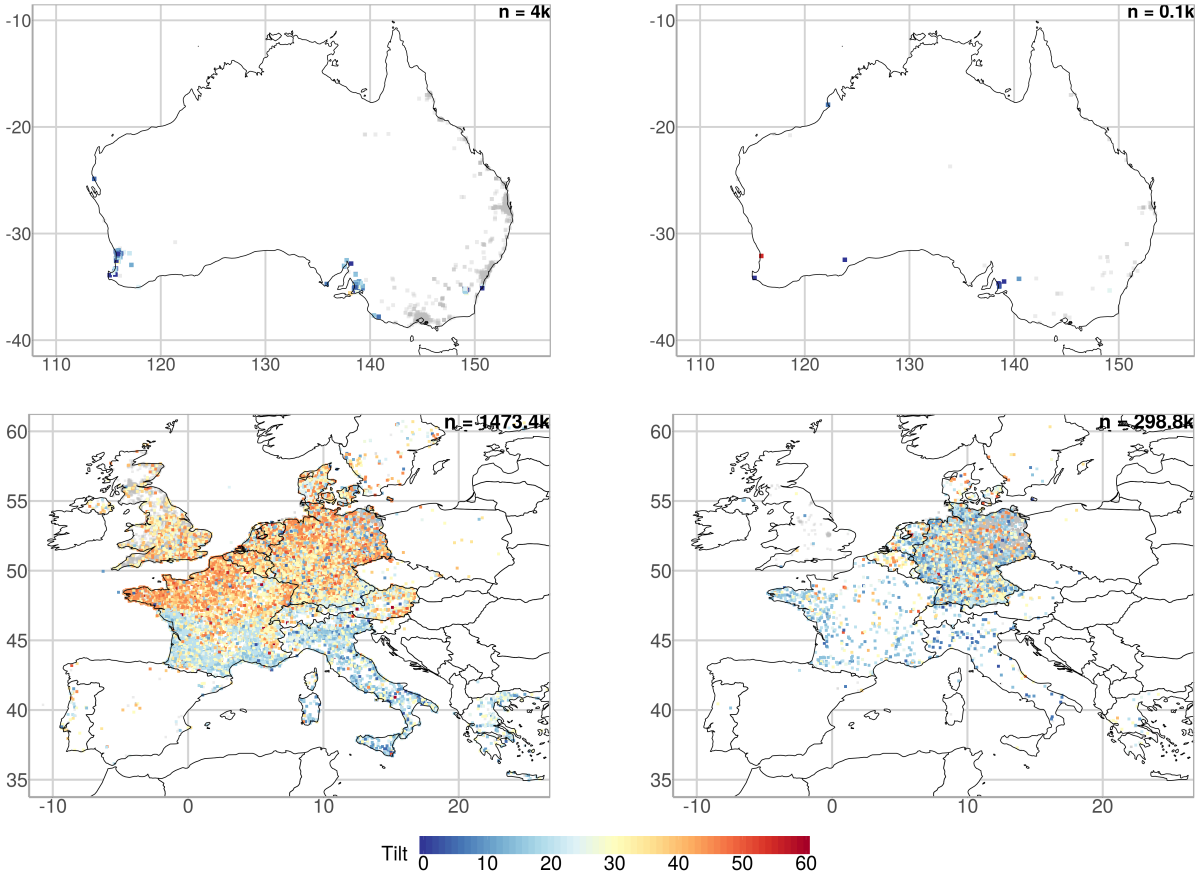

Figure 3: Maps for Australia (top) and Europe (bottom). The left column shows systems $\leq$ 25 kWp and the right column systems $>$ 25 kWp. Systems which do not report tilt are in grey colour.

though the number of larger systems is rather low in most countries, their strong contribution to the total power generation and the knowledge that characteristics change with the system size justify a consideration in a separate cluster. The threshold value of 25 kWp is displayed as a dashed vertical line in Figure 2. If the threshold value were higher, only a small number of systems would be left in the upper cluster and the derivation of representative statistics impeded. (3) Several threshold values were trialled in our analysis. A value of 25 kWp was finally decided upon as it satisfied the aforementioned criteria and passed visual inspection by producing distinct distribution curves.

Both the impact of system size and the geographical influence can be studied in respect to the tilt angle of systems in Figure 3 and Figure 4. All regions show a tendency towards smaller tilt angles for system sizes $>$ 25 kWp. Especially for systems $\leq$ 25 kWp in Europe, the dependency between latitude and tilt can be observed by an increasing tilt angle from Italy to Denmark. However, it should be noted that the spatial influence is not only limited to a pure geographical relationship; the spatial impact depends on regulations and incentives which often occur on a national level. The policy situation

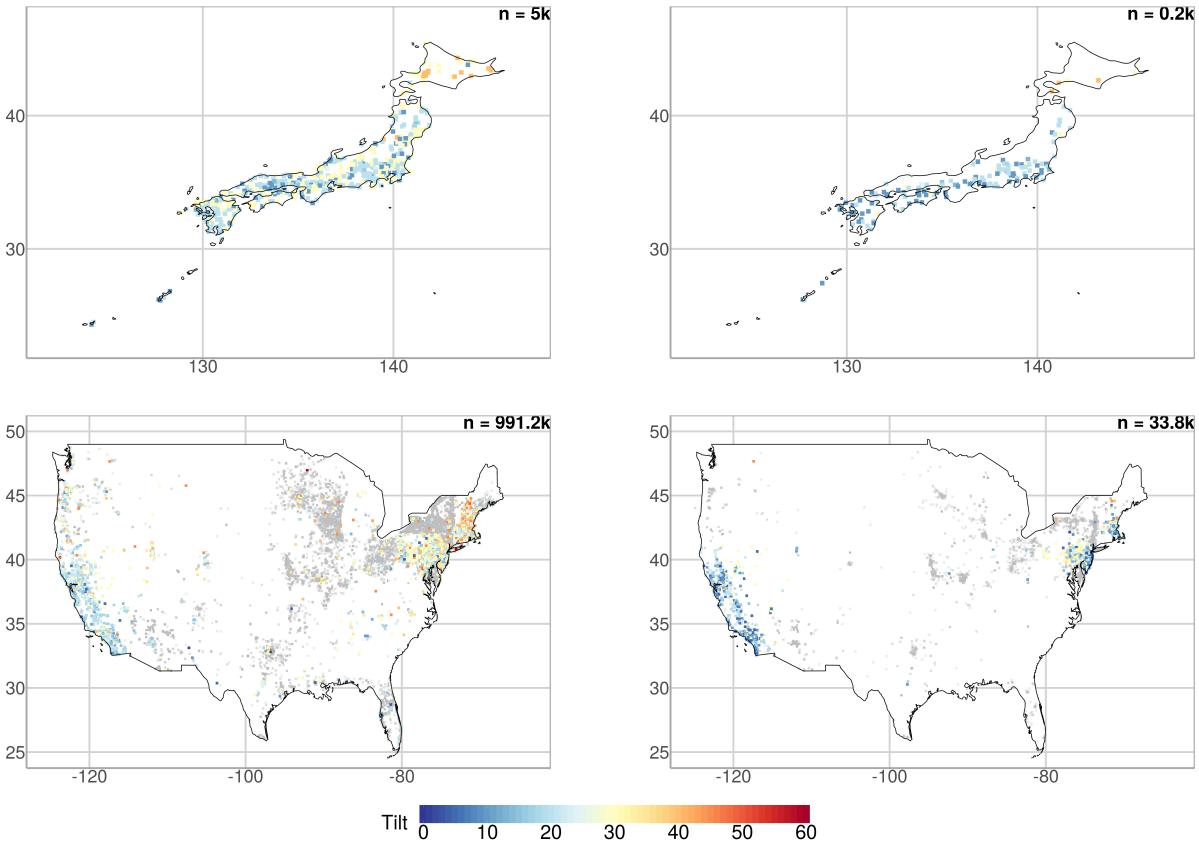

Figure 4: Maps for Japan (top) and the USA (bottom). The left column shows systems $\leq$ 25 kWp and the right column systems $>$ 25 kWp. Systems which do not report tilt are in grey colour.

in France leads to a high number of 3 kWp systems (see Leloux et al. (2012b) in section 1.1). In Germany, there are changing regulations and feed-in tariffs for systems $>$ 30 kWp resulting in an increase in the black line of the right plot in Figure 2. Furthermore, the UK had a higher feed-in tariff for systems $\leq$ 4 kWp up until January 2016 and has since moved to $\leq$ 10 kWp (ofgem, 2018). These are such examples of significant policy-specific regional influence that can impact upon the characteristics of PV systems.

There are many opportunities as to how we sub-categorise the data into clusters. Many of which could be explored in order to derive meaningful information depending on the approach. Options include separating by climatic region or grouping by policy similarities. However with respect to the aforementioned aspects, a clustering at a country level seems advisable for the following reasons: (1) National regulations and incentives have a visually evident impact on the occurrence of different system sizes which may itself influence other metadata. (2) A geographical influence was observed on multiple parameters. Countries limit this influence by their size. The only exception of this strategy is the USA. The enormous geographic area of this country

results in a inhomogeneous pattern of the specific annual yield. This is a direct consequence of the heterogeneity of the solar resource within a country. The USA was thus split at 37.5° N into a northern and a southern component. The same approach could be applied to Australia, however, the sample size of available data is too low. Further subdivisions e.g. by the latitude for systems $\leq$ 25 kWp in France (see tilt in Figure 4), could be considered but exceed the scope of this paper and is a focus of future work. (3) There is a certain convenience to clustering by countries. Many of the studies previous focused mainly on a single country, this is indicative of a researchers interests and data availability. We feel that, whilst there are many options of clustering that can be explored, a preliminary study at a country level is of most interest.

The region ``Rest of Europe" is not be considered further due to its inhomogeneous portfolio of systems across different countries in Europe. The clusters, defined by their belonging to a region and system size, are used in the next section to derive representative distributions for the metadata.

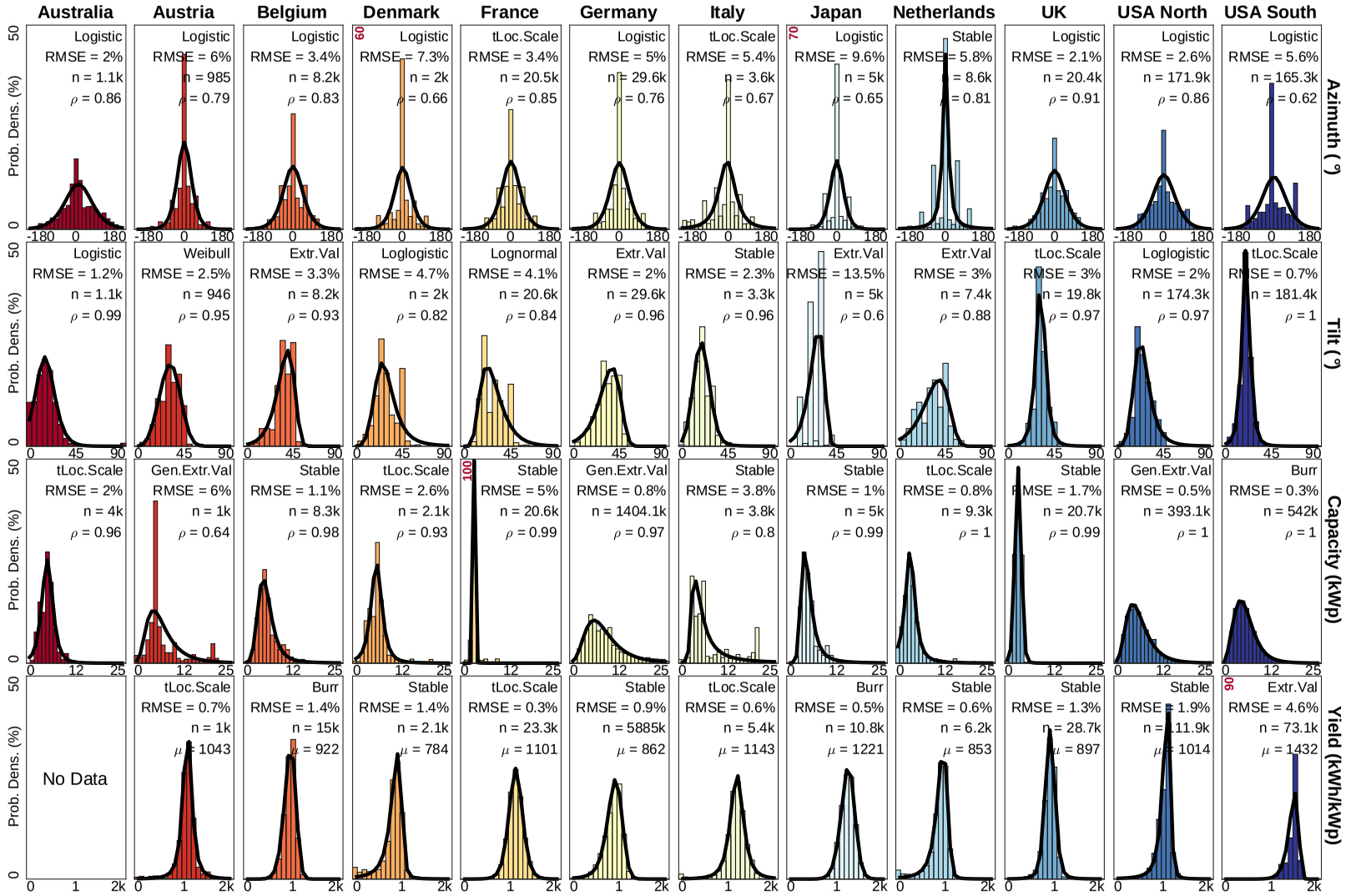

Figure 5: Histograms of real data (bar) with approximated probability distributions (line) for the different clusters (columns) and parameters (rows), where capacity is the installed capacity and yield is the specific annual yield. All systems reported within this figure have a capacity $\leq$ 25 kWp, see appendix for the same plot for $>$ 25 kWp. Within each of the axes is reported the name of the best fitting distribution type (see section 4.1 for detail), the root mean squared error (RMSE) between the scatter of real data in the histogram against the fitted probability density distribution, the number of data points considered for that cluster (n), and the Pearson correlation coefficient (ρ) of the linear regression. The mean value μ is shown in place of ρ for Yield. All y-axes are scaled between 0% and 50% probability except where a bold red value is assigned to the individual axis.

4. Approximation of parameters in clusters

The intentions of parametrisation are twofold. Firstly, we want to discover whether or not the parameters (tilt, azimuth, capacity and yield) can be represented with simple parametric distributions. Secondly, we want to explore the relative differences between clusters through comparison between probability distributions. We concede that simple distribution fitting has weaknesses such as not appropriately capturing a more complex relationship offered by non-parametric fitting, however, reproducibility of the statistics is encumbered with added complexity. For our first presentation of the substantial volume of PV system data collected, we focus on simple distribution fitting as interesting comparisons and individual cluster insights can be drawn, and we are able to comment on the ability for these complex parameters to be represented as such.

4.1. Methodology of fitting the distributions

In order to enable the utilisation of the aggregated statistics of each cluster (defined in section 3.2) and for each parameter (defined in section 3.1), individual distributions are fitted to the real-world probability density histograms. The results are presented in Figure 5 ($\leq$ 25 kWp) and Figure A.10 ($>$ 25 kWp). The total number of available data varies between clusters and parameters; there is no further processing beyond the criteria described in section 2; all possible data available is used. Differences in data within a cluster are due to some PV sites not reporting one or more parameters. There are up to 6 years of reported specific annual yield (2012-2017). The normalised value within each year is taken as an individual sample and so there are up to $6n$ more samples for this parameter.

For each cluster and for each parameter, many different distribution types were fitted to the probability density. Distributions were fit using the inbuilt FITDIST function of the software Matlab®(Matlab, 2018). There are 23 parametric distribution types available, of which all are fitted to the data. Where distribution types require only positive values (for parameters with negative bins) or values between 0 and 1, the data is scaled to satisfy the distribution requirements and allowed to rescale so that as many distributions could be tested; note that no distribution requiring this treatment was found to be best fitting, and so no further discussion is made regarding this normalising process. Probability density functions are then scaled to only exist between the x-axes limits as indicated in the figure, for example, the tilt distributions are only relevant between 0° and 90°. This means that the sum of all probabilities between the prescribed x-axis range must be equal to 1. This is important as some distributions facilitate values way outside of the bin limits resulting in the sum of probabilities between the bins of interest $\neq 1$, and so would not fit the histogram. The disclaimer is, therefore, that these distributions *must* be scaled before use and not be extrapolated beyond the specified bin ranges else risk persisting an under/overestimation about the scaling factor, defined as

$$s = \frac{1}{\sum_{a}^{b} p_{a:b}}, \tag{1}$$

where s is the scaling factor, p is the probability at each bin between the lower and upper bin limit, a and b, respectively. The resultant fitted distribution is then plotted against the real probability density and tested for linear fit; the root mean squared error (RMSE, percentage) and Pearson correlation coefficient (ρ, dimensionless) are derived. A perfectly fitted distribution would result in $y = x$ with $\rho = 1$ and an RMSE= 0%. The distribution type with the lowest RMSE was selected for the plot. Should there be more than one distribution type that has the same RMSE, then the type with lowest ρ is selected. Should there still be more than one distribution type after this, one of the remaining types is selected at random.

The exact parameterisation for each distribution presented in Figure 5 and Figure A.10 are detailed in Table A.4. Each distribution has up to 4 coefficients and are employed using different equations, not all 23 parametric distributions are detailed, only those that featured within the study. Whilst Table A.4 details the parameterisation of the coefficients, it is Table A.3 that explains how to use those values to form the distribution. Furthermore, the mean or median values of the whole dataset, exclusive of the 25 kWp separation, are presented in Table 2.

Table 2: Mean or median value extracted from entire data set (without separation by capacity size) for each of the parameters of tilt angle, azimuth angle, system capacity and specific annual yield.

Country	Tilt (°) mean	Azi. (°) mean	Cap. (kWp) median	Yield (kWh/kWp) mean
Australia	16.1	8.58	5.00	-
Austria	31.1	-0.34	5.15	1,040
Belgium	35.6	-1.69	5.20	921.5
Denmark	30.0	0.48	6.00	786.0
France	28.7	-0.28	2.96	1,101
Germany	31.6	-2.46	8.96	870.2
Italy	19.8	-15.9	5.88	1,142
Japan	23.8	-1.20	4.92	1,222
Netherlands	32.5	0.77	3.30	855.2
UK	31.8	-1.07	2.94	896.7
USA North	25.2	0.42	5.81	1,005
USA South	19.9	9.33	5.26	1,426

4.2. Discussion of the distributions

The following discussion about clusters and distributions mainly refers to Figure 5 with systems $\leq$ 25 kWp unless explicitly noted otherwise. The reason for this is that the vast majority of systems are within the $\leq$ 25 kWp category, and so are of most interest. However, important differences to system sizes $>$ 25 kWp are mentioned and can be observed in Figure A.10.

It is important to note that only rough dependencies between parameters, regions and system sized can be considered with this clustering approach. The more intricate and established interdependencies have not been explored within this paper as it is beyond the scope of the initial objective. The authors reserve this for future work.

4.2.1. Azimuth angle

The most noticeable feature of the azimuth observations is the significant probability of an equatorial facing PV system. This is unsurprising as it offers the best annual specific yield by receiving maximum system efficiency at peak solar position. The topic of extreme probability of an equatorial orientated system was discussed in section 3.1; the prevalence of 0° is true of all sites for both $<$ 25 and $>$ 25 kWp. The Japan and Netherlands clusters

have exaggerated angles of -45° or +45°, assumed to be a result of overly simplified reporting.

The distributions could not capture the probability of 0° with exception of the Netherlands where a Stable distribution fitted best. Even with large sample sizes for the USA North and south clusters, a distribution could not be fitted that satisfied the observed probability for an azimuth angle of 0°. Perhaps a more complex or bespoke distribution type is needed to suitably express the probability distribution of azimuth angle with reproducible accuracy. This large proportionality of 0° was also observed by Saint-Drenan et al. (2018), who fitted a normal distribution in similar magnitudes to the logistical distribution fitted in this article. That said, there is an argument that the significant 0° azimuth feature is exaggerated when considering the UK cluster. The majority of the data within the UK cluster is from Sheffield Solar. Their users report the system metadata, however, there is a feedback to the user reporting system performance analysis on a monthly basis, inclusive of a nearest-neighbour performance analysis of a system of similar metadata. Users are encouraged to verify their reported metadata and is often double checked with satellite imagery; the result is much more accurate reporting of metadata for the UK cluster leading to the smoothness of distribution fit. With improved PV system metadata reporting, we see a wider spread of azimuth about 0°.

4.2.2. Tilt angle

The tilt angle across all clusters is rather unique per cluster with 7 different distribution types being found as the best fitting among 10 clusters. We previously discussed the gentle increase of tilt angle with latitude. Solar installers can mount the PV panels with a steeper tilt angle to that of the roof at higher latitudes through arrangement of the mounting brackets; this is not expected to be overly common practise. The predominant factor for smaller roof integrated systems is expected to be the physical roof angle, which is influenced by local architectural styles. We suspect this is the case, particularly when considering France in Figure 3 where there exists a distinct change in the tilt for the ≤ 25 kWp systems at roughly 47.5° latitude. Note that France and Denmark have similar distributions despite France having a significant number of systems south of that 47.5° roof tilt feature. Furthermore, Belgium and the Netherlands share similar climate and latitude yet feature distinctive distributions.

Interestingly, the Australian cluster consisting of the second lowest number of observations has the second most accurate fit after USA South. This is in part due to the smoothness of the distributions and accuracy of method in which the tilt is obtained (see section 2). The USA cluster has excellently fitted distributions suggesting accurate measurement, particularly for the USA South cluster where the tilt distribution is fitted with $\rho = 1$ and RMSE=0.7%. The Japan cluster evidently suffers from reporting to the nearest 10°, and so we suggest to avoid using a best-guess approach to collecting metadata as it leads to biased distributions. The tendency for larger system sizes having smaller tilt angles, introduced in section 3.1, can be confirmed when comparing Figure 5 and Figure A.10. The only exception is Denmark, which reports only a small number of systems > 25 kWp.

Within the distributions, the smallest mean tilt angles were reported in Australia (16.07°), Italy (19.81°) and USA South (19.89°). The largest mean tilt values were reported in Belgium (35.58°) and closely followed by the UK, Germany and Austria (31°).

4.2.3. Installed capacity

The most obvious observation from the installed capacity is the extreme peak within the French cluster. Of all 20.6k systems ($\leq$ 25 kWp and $>$ 25 kWp), 73.74% of them report an installed capacity of 3 kWp when rounded to nearest integer, though note that the French dataset reported to a high decimal precision. The best fitting distribution cannot appropriately represent this extreme despite a very high $\rho = 0.99$; the RMSE value of 5% is indicative of the Stable distribution assigning 100% probability to 3 kWp. This distribution is, therefore, very limited even if it does most accurately capture the data for France. As discussed when defining the clusters, this peak in capacity is a direct response to regulations within that country. This is further observed in the UK database, with the vast majority of systems being $\leq$ 5 kWp due to the nature of the feed-in tariff rate. The north and south USA clusters demonstrate the power of a larger and consistent sample size reporting RMSE 0.5% and 0.3%, respectively, with both reporting $\rho = 1$. Interestingly, the distribution type between USA clusters are distinct from each other, with a slightly increased probability of smaller systems in the south. This is expected to be a result of more rooftop solar in the sunnier States, though this is speculation.

The shape of the distribution functions for system sizes $>$ 25 kWp (Figure A.10) differ to systems $\leq$ 25 kWp and show an heightened concentration of systems $<$ 50 kWp. Australia is an exception and reports many systems with an installed capacity of $\approx$ 100 kWp.

The mean values of capacity are too heavily influenced by the presence of large systems (cf. 2b), and so the median value is reported to reduce bias. From the distributions, the country with smallest capacity median is the UK (2.94 kWp) and the largest is Germany (8.96 kWp). The fact that the German data reveals such a high median is reflective of the thorough nature of data collection whereby nearly all systems are reported; we have very few large systems reported from the UK as the database is primarily used for rooftop solar and so this statistic is not overly representative.

4.2.4. Specific annual yield

The most noticeable detail of the specific annual yield distribution fits is the smoothness of the histograms of raw data. This is perceived to be of two reasons. Firstly, the sample size is typically much larger (n = 5.885m for the German cluster). Secondly, the data is digitally recorded and not reliant on human reporting. The mean μ is presented in place of the correlation coefficient so as not to over busy the plot, though for completeness, all sites reported $\rho \geq 0.98$ except USA South with $\rho = 0.93$. Recall that the specific annual yield is normalised for inter-annual differences and so we can directly compare clusters. Each cluster exhibits reasonably unique subtle traits, it is expected that the larger the share of equatorial orientated systems with more optimal tilts, the larger the specific yield,

however, it is also a function of local meteorology and climate local to the systems and not just latitude, orientation and tilt. More questions can be derived from these distributions than are really answered. For example, consider the German cluster. There is a substantial tail towards lower specific annual yields that is not observed in other clusters. Germany is known to be a mature market when it comes to PV, and so is this tail indicative of ageing systems, or perhaps all clusters would present this pattern given as large a sample size? The USA South cluster has an unexplainable peak at exactly 1650 kWp/kWh. The only other country within our study that is comparable to southern USA in terms of climate and land availability is Australia, alas, we have no data for this cluster to gain insight to this peak. We would expect to observe much higher yields in Australia akin to southern USA.

Leloux et al. (2012a) found that of 158 systems in Belgium,the mean specific annual yield was 836 kWh/kWp. Our analysis of 15k specific annual yields finds the mean to be 921.5 kWh/kWp. Leloux et al. (2012b) applied the same approach for 1,635 systems in France resulting in a mean of 1,163 kWh/kWp. Our analysis of 23.3k systems places the mean value at 1,101 kWh/kWp. When comparing Figure 5 and Figure A.10, the expected trend towards higher specific annual yield values for larger systems (see section 3.1) can be confirmed for Denmark, France, Germany, Japan, Netherlands, UK and USA North. The other four countries instead show a decrease in the mean specific annual yield for systems > 25 kWp.

From the distributions, we find the smallest mean specific annual yield is in Denmark (786.0 kWh/kWp) and largest in USA south (1,426 kWh/kWp).

4.3. Using the distributions

As we have observed that each cluster and parameter can be generally represented by a probability distribution, our discussion can shift towards the usage of these statistics in regional PV power modelling approaches. Generally, the cited publications in section 1.1 (category 1) not only emphasize the practical relevance of statistical distributions in regional PV power modelling approaches, but also sketch the procedure of how these statistics can be used and therefore serve as good examples. One of these publication is Saint-Drenan et al. (2018), which provides detailed information about how fitted distributions can be applied in a regional power simulation and therefore serves as a good example.

We foresee that the fitted distributions from the previous section can be used to randomly sample the desired metadata of a portfolio of systems in a specific cluster. As pointed out in section 3.2, this cluster can then be weighted individually by the probability of occurrence (as can be derived e.g. from Figure 2. To reproduce the distributions, one must extract the appropriate distribution variables from Table A.4, apply them in the expression from Table A.3, and scale the result according to Eq. (1). We state that the data we have is not representative enough to derive global distributions as there are too many features that can influence the PV system characteristics from regions we do not have access to. The derived distribution functions should only be used for their specific clusters, or for clusters with particularly similar climates and policies.

The usage of the fitted distributions is sketched in a practical example, assuming that the PV power generation in Germany is of interest.

1) In Germany, the installed capacity, geographic location and specific annual yield of all PV systems is known (see section 2) and should not been sampled if one wishes realistic outputs, though may be sampled for theoretical purposes.

2) For each PV system, tilt and azimuth is assigned by sampling the fitted distribution from the relevant cluster. E.g. a system with 10 kWp will use the distributions from the cluster for systems $\leq$ 25 kWp. In this example we can extract the data from Table A.4 such that the German cluster has a Logistic distribution for azimuth with location coefficient of -0.1366 and scale parameter of 20.6455. The distribution can be recreated using the logistic function defined in Table A.3. Please note that there is a high risk that the sampled characteristics won't accurately predict the metadata for a specific PV system. It is the objective to use these distributions to simulate larger PV portfolios and we expect to derive representative characteristics for that application.

3) The direct usage of yield is more complicated for two reasons. Firstly, in contrary to azimuth, tilt and installed capacity, it is not an input parameter in the simulation chain but instead indicates the power generation, which is the typical output of a simulation. Secondly per definition, the specific annual yield sums the PV power generation over a whole year. In many application however, simulations may cover a different time span. How can yield be used in simulating the regional PV power generation then? When making the simplifying assumption that the meteorological conditions are relatively similar within a cluster, the observed specific annual yield can be interpreted as a measure that expresses relative performance differences between PV systems. For instance, a PV system with a yield of 600 kWh/kWp can be said to be less efficient than a PV system with a yield of 800 kWh/kWp. For usage in a simulation, a conversion from yield into an performance factor is therefore necessary. A potential method of conversion is to take the range of yields (0 to 2000 kWh/kWp) and align it to typical ranges of the performance ratio, though taking care to centre the mean yield against the mean performance ratio (said to have a wide distribution centred about 0.74 as derived from 5,000 systems in the Netherlands (Tsafarakis et al., 2017; Reich et al., 2012)). A direct linear conversion could then be applied. For example, a system in Germany with yield 870.2 kWh/kWp (the mean for this cluster) could be assigned a performance ratio of 0.74. This performance ratio can then be applied as a correction factor to either the output power or the system capacity, and therefore facilitating representative differences between systems.

4) The individual PV power generation for each system can be simulated by considering the sampled system characteristics as well as the known installed capacity, efficiency of the specific system and its geographic location. Within such a simulation, other inputs will be needed such as the local irradiance or ambient temperature. Please note, if considering such a large number of systems is too computational intense, instead a smaller number could be randomly chosen and then used within an upscaling

approach.

5) Finally, the total power in Germany can then be derived by aggregating the simulated power from all systems.

5. Shading on roofs

5.1. Methodology of the shading analysis

An objective in this paper is to derive generalised findings of how to consider the impact of shading. We aim to achieve that in a more sophisticated manner than the overly simplified manner presented in category 3 from section 1.1. It is, however, not the aim to study differences of shading in rural or urban areas all over the world in this paper. This shading analysis is performed on $\approx$ 48,000 buildings in the city of Uppsala, Sweden (N59.9°,E17.6°). Uppsala provides a variety of different buildings (44% residential, 2.2% industries, 5.7% commercial and services, 49% other) and therefore allows studying differences in the impact of shading. The average height of the buildings studied here are 6.4 ± 4.1 m, which may be compared to a study on 12 US cities of various size (29,498 to 1,066,354 buildings) with average building height ranging from 4.1 to 9.7 m (Schläpfer et al., 2015). The analysis is not limited to any country specific influences because all combination of solar angles are considered; climate does not influence this shading study. The above reasons emphasise the general representativeness of Uppsala and were reason for its selection.

The shading analysis is realised by using the method in Lingfors et al. (2017), which was cross-validated in Lingfors et al. (2018). Inputs to the model are low-resolution LiDAR data (0.5-1 pts/m^2) and building footprints, provided by the Swedish Land Survey (2015, 2016). The model does the following:

1. Finds a simple roof shape from a template of roof types using linear regression on LiDAR data;
   (a) within the footprint of the examined building and,
   (b) within building footprints of similar shape in its proximity,
2. Each roof now consists of 1-4 facets depending on the roof type (1 for flat or shed, 2 for gabled and 4 for hipped or pyramidal). If LiDAR data are insufficient, the roof type cannot be determined and the building is excluded from further analysis. The number of roof facets are >90,000 (cf. number of buildings). However, around 1,000 facets which are > 20 m above ground are excluded, as there is an increased risk of these roofs being misrepresented due to noise in the LiDAR data.
3. After some filtering of the LiDAR data surrounding the building, a triangulated irregular network (TIN) is produced representing objects, predominantly trees and other buildings that may shade the roof.
4. Using the TIN as input, a viewshed (a map showing what parts of the sky are visible from the perspective of a point on the roof) at every 0.5 m × 0.5 m section of the roof is calculated to determine whether there are objects blocking the direct solar path. The resolution of the viewshed is limited to solar elevation angles,

α_s, of 2.5, 7.5,..., 87.5°, and solar azimuth angles, γ_s, of -180, -170,..., 170° where 0° is due south. Since the sky sectors are angular-equal, they are not equal in size (see dotted lines in Figure 6). Hence, the contribution of diffuse irradiance from each sky sector depends on its size and the angle-of-incidence of the irradiance from the sky sector onto the plane.

5. For each combination of α_s and γ_s the mean shading of the whole roof facet is calculated, noting that roofs can be partially shaded. This is illustrated in the *viewsheds* of Figure 6. For a discrete point of the roof, each element of the sky would be either shaded or not shaded corresponding to black or white (0 or 1), respectively, in the left panel of Figure 6. However, if the mean of all points of the roof are considered, the viewshed would be blurred (grey) as illustrated in the right panel of Figure 6. The mean viewshed displayed on the right of Figure 6 is only for illustrative purposes and can be considered to gain understanding as to how the beam, diffuse and reflected irradiance subcomponents are affected by shading for the general region of all facets within this study. Results of the shading analysis are presented in section 5.2.

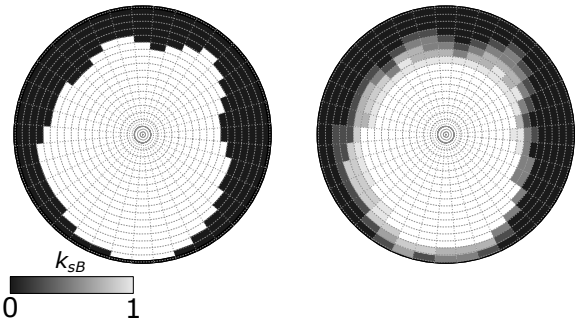

Figure 6: Polar diagrams of viewsheds, where the displayed angles represent the azimuth angle, and the radius the elevation angle. Left) illustrates the viewshed of a single point on the roof. Right) the mean viewshed of all the points on the roof is illustrated. The dotted lines mark the sky sectors for which the viewshed analysis was conducted from their respective centre points. Note that the right plot is purely illustrative and not used within any of the modelling stages.

5.2. Deriving a simplified shading model

The main results from the shading analysis on $\approx$ 48,000 buildings in Uppsala, Sweden, are presented in Figure 7. The colour of each bin in the left panel of Figure 7 represents the average ratio of all roof facets being visible to a sky sector, defined by the corresponding solar elevation angle and solar azimuth intervals. This visibility is here referred to as the *beam shade index*, $k_{sB} \in [0, 1]$ (see Figure 6), where 0 means the roof facet is fully shaded. The dashed lines illustrate the solar path for Uppsala, Sweden. However, the corresponding solar path could be over-layered for an arbitrary site to visualise the implication of shading for that site. From the left panel of Figure 7 it is also clear that the solar azimuth has very little importance, which is logical as shading should be as likely from any direction when a large portfolio of buildings is considered.

In the right panel of Figure 7, the average (×-marked) and percentiles (dashed) of k_{sB} for all roofs are presented as a function of only the solar elevation angle, hence it differs from the left panel by not considering the azimuth angle. The thin red lines

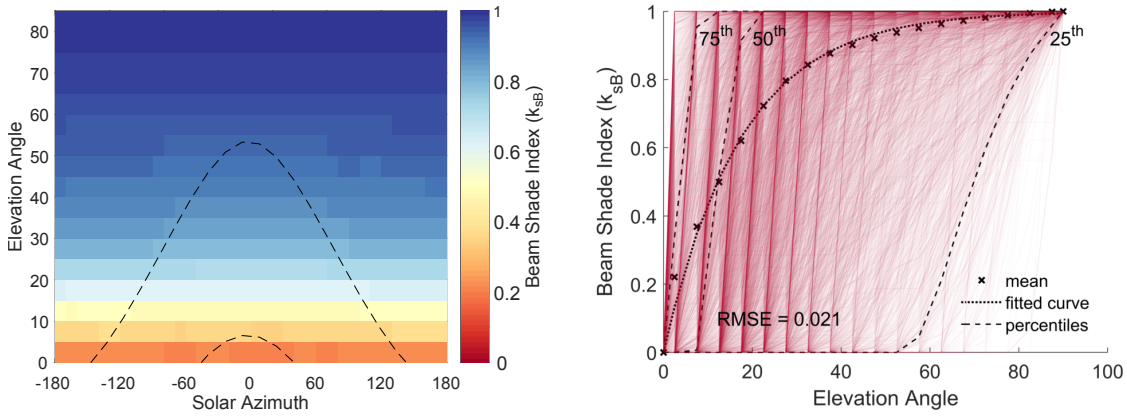

Figure 7: To the left, the mean shade index of all studied roof facets are presented at bins of a viewshed defined by the solar elevation angle and solar azimuth. To the right, the shade index is plotted against the solar elevation angle. Every shade index profile from each studied facet are indicated with a red solid line. The mean shade index of all facets is indicated with crosses, with fitted curve represented by a dotted line as presented in Eq. (2). The RMSE between the means and fitted curve is 0.021. The dashed lines represent the different percentiles.

represent 10,000 individual roofs. Many of these lines jumps from 0 to 1 when going from one elevation angle to the next, meaning that from being entirely obscured, the roof becomes entirely visible when the elevation angle is increased by 5°. The mean beam shade index, $\bar{k}_{sB}$, of all the roof facets can be represented by a fitted curve (dotted), derived as a function of the solar elevation angle, α_s:

$$\bar{k}_{sB} = 1 - e^{-\alpha_s/17.5}. \quad (2)$$

The average beam irradiance that will fall on a tilted roof, if shading is considered, could then be calculated as:

$$B_T = \bar{k}_{sB} \frac{\cos\theta}{\cos\theta_Z} B_H, \quad (3)$$

where B_H is the unshaded beam irradiance on the horizontal plane, θ is the angle between the incident irradiance and the normal of the roof plane and θ_Z is the solar zenith angle.

Assuming similar shading properties, i.e., vegetation and urban density, as in Uppsala, this function may be used in any area to determine the impact of shading on roofs as a function of the solar elevation angle. It gives a better estimation than solely assuming a cut-off solar elevation angle for the beam irradiance, which is a method commonly used for PV potential studies.

On the other hand, the red lines of figure in the right panel, representing individual buildings, reveals the variation in shading among the buildings. Thus, studies of higher detail where, for instance, the implications in a low-voltage grid due to shading on PV modules are studied require a method that reproduces these variations.

If the global irradiance on a shaded roof is of interest, one also needs to consider the diffuse (D_T) and reflected (R_T) irradiance subcomponents on the tilted plane, which both depend on the view factor (visible fraction of the sky, $f_{sky} \in [0, 1]$) from

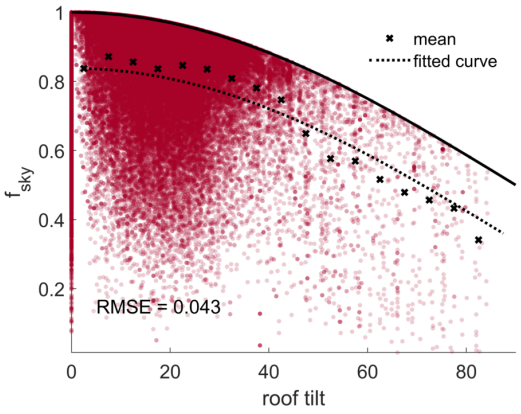

Figure 8: The mean f_{sky} as a function of the roof tilt (marked with x) and a fitted curve (dotted) presented in Eq. (4). The solid line represents the first term of Eq. (4)

the perspective of the roof. I.e., the viewshed of the roof should be considered (see Figure 6). The view factor f_{sky} represents the ratio of the isotropic diffuse irradiance from the sky hemisphere and reaches a value of 1 for a horizontal surface if there are no shading objects.

The values of f_{sky} for $> 90,000$ studied roof facets are presented in Figure 8. A fitted curve (with an RMSE of 0.043 with respect to the means, indicated by black crosses) was derived with the function:

$$\bar{f}_{sky} = \frac{1 + \cos(\beta)}{2} - C, \tag{4}$$

where the first term is the view factor for a free sky and C is a constant representing the contribution from shading objects, here found to be 0.162. This equation may be used to calculate the diffuse and reflected irradiance following equations (6) and (10) in Lingfors et al. (2017), respectively.

In Figure 9 the losses due to shading are presented for the three irradiance subcomponents (calculated individually for each roof facet), sorted with respect to decreasing diffuse irradiance losses. In this analysis, hourly instantaneous Global Horizontal Irradiance (GHI) and Direct Normal Irradiance (DNI) data from 2014 for Uppsala were used (SMHI, 2015). B_T was calculated through Eq. (3), if we let $\bar{k}_{sB}$ here represent the mean value of k_{sB} for all points on the individual roof facet. D_T, as well as R_T, were calculated through equations (6) and (10) in Lingfors et al. (2017), respectively, using the f_{sky} derived for each roof facet. These equations, adapted for the conventions used in the present paper, can be expressed as:

$$D_T = D_H \left[(1 - A_i) f_{sky} + \frac{\cos\theta}{\cos\theta_Z} A_i \right], \tag{5}$$

and

$$R_T = (B_H + D_H)\rho(1 - f_{sky}), \tag{6}$$

where A_i is the anisotropy index and ρ is the surface albedo, here assumed to be 0.2 for all surfaces (i.e., ground, trees, buildings etc.).

Hence, Eqs. (2) and (4) in the present paper were not used here, but could be valuable in future studies where, for instance, the level of detail of the building topography in a city is unknown or the time for making detailed simulations is limited, yet the impact of shading on solar power generation is of interest." The in-fold figure illustrates the negative correlation between diffuse (D_T) and reflected (R_T) irradiance. The diffuse irradiance decreases (i.e., the losses increase) with a decreasing f_{sky}, while instead the reflected irradiance increases (i.e., negative losses in Figure 9). From Figure 8, it is clear that f_{sky} decreases with an increasing roof tilt,

leading to a higher contribution of reflected irradiance for a highly tilted roof. The mean losses due to shading (expressed in relation to the unshaded global irradiance) for the whole building portfolio are 7.3%, 3.6%, 6.3% and −2.7% for the global, beam, diffuse and reflected irradiance, respectively, where the minus sign is indicative of an added contribution to the total irradiance, since trees, buildings etc. adds to the total reflective area seen by the roof when shading is considered. Hence, diffuse losses contribute the most for Uppsala, which has an annual clear-sky index of 0.63 (calculated as the global horizontal irradiation for 2014 divided by the clear-sky irradiation for the same period (Ineichen and Perez, 2002)). One should also remember that all roofs in Uppsala were considered. If only roofs with installed PV systems on them were considered, the losses would most likely be lower. It is likely that the present method over-estimates the reflected irradiance at clear conditions as all trees and buildings seen by a roof could also be themselves shaded, therefore, offering reduced reflected irradiance. To consider this is a complex matter and needs extensive research. For instance ray-tracing could be incorporated in the model but at a computational cost.

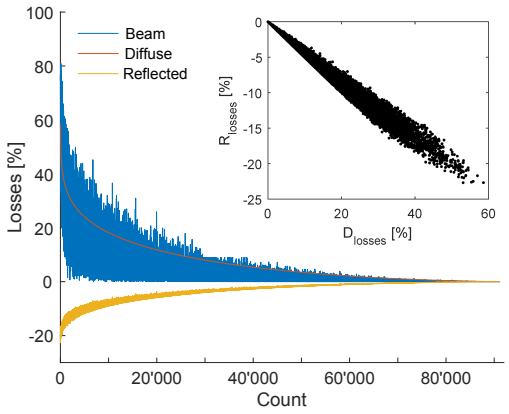

Figure 9: Beam and diffuse irradiance losses and the added contribution from reflected irradiance to the global irradiance considering shading on the $> 90,000$ studied roof facets. The in-folded figure illustrates the negative correlation between the losses of diffuse and reflected irradiance when shading is considered.

6. Future advancements beyond the scope of this work

The main objective of the paper was to fit distributions to selected metadata and approximate functions that describe the impact of shading. This enables replication of these characteristics and allows a usage in regional PV power modelling approaches with suitable representativeness. The underlying basis for the approximations are numerous datasets with metadata and simulated results from a model in the case of shading. Background information and references for further reading are provided in the related sections. Furthermore, the level of accordance of the fitted distributions and functions with the original data is expressed by error metrics and limitations of the procedure are critically discussed. Naturally, with such a considerable and detailed database of information, we cannot cover all aspects in a single paper. We have opted to present an overview in a manner that enables the user to engage with the findings. That said, we have identified several interesting topics during our work that we would like to study in more detail, however are beyond the scope of this paper's objectives.

- **Focus on specific parameters:** The whole

dataset offers so much information that it is impossible to evaluate all specific parameters in detail within one paper. This data could potentially be used to study various performance indicators, e.g. by including irradiance information in specific regions and the age of the systems (provided for most systems).

- **Dependencies between parameters:** In this paper we have qualitatively discussed pairwise dependencies between parameters. Furthermore, we have applied a quantitative approach to individual parameters by fitting distribution functions. The next step will be to quantitatively incorporate dependencies between multiple (two and more) parameters, e.g. by joint distributions, multivariate models, etc. By that, the complex relations should be better represented.

- **Complex distributions:** The azimuth angle presented irregularities with a wide base and tall 0° peak and on occasion presented a trimodality that is certainly not-able to be captured by standard parametric distribution types with satisfaction. Whilst we dispute the validity of much of the measured data due to reporting simplifications, there is scope to analyse the distributions in a more statistically rigorous manner. There is scope to combine distributions and to enable multi-modal, non-parametric definition of the non-conformal parameters, notably the tilt and azimuth. We intend to make available the actual probability distribution for the reader to draw their own conclusions, see our invitation for collaboration below.

- **Cluster refinement:** Influence of climatic region may influence certain parameters, particularly the specific annual yield. It is probable that the specific annual yield is not only a function of latitude (as we have demonstrated with a general regression between 30° and 50° of latitude), however, it is a function of the climatic region where those sites are situated. There is a lack of data within the 0° to 30° latitude band with which to successfully analyse this hypothesis. Further steps could be to replace the clusters by country with clusters by climate region using maps such as the Köppen-Geiger classifications, or perhaps by mean irradiance using a dataset such as NASA SSE.

- **Shading:** As mentioned in section 5.2 the results on shading from the present study can be used on a large portfolio of buildings, while for smaller areas one may want to produce realistic viewsheds for a few buildings to study the impact from shading. One simple approach would be to provide a database of viewsheds such as the one produced in this study, from which samples could be randomly drawn. To avoid the need of a database, another approach could be to design a model that can reproduce the distribution of shading profiles, perhaps stochastically using Markov chains to create statistically appropriate skylines. While the solar elevation angle is probably the most influential parameter, other factors such as the type of roof or height of the building would most likely also have an impact. Hence, a set

of building specific parameters would satisfy as inputs to such a model.

- **Invitation for collaboration and data access:** This work would not have been possible without support from many sides mentioned in the acknowledgements. Gathering this data has proven difficult at times, and finding the correct person to approach was not straight forward. Therefore, we would like to extend an invitation to the reader. Should you have good ideas of how to use this data, or have large data itself, particularly in countries that we have not detailed, we encourage you to get in contact with either Sven Killinger (sven.killinger@ise.fraunhofer.de, svnkllngr@gmail.com), Jamie Bright (jamie.bright@anu.edu.au, jamiebright1@gmail.com) or Nicholas Engerer (nicholas.engerer@anu.edu.au). Much of the data is confidential and so we cannot share it, however, the aggregated statistics are available. Should you wish to have access to the real distributions presented in the work, they are available on request, should they be publicly released, all communications will be made through our ResearchGate project. You are encouraged to follow that project for updates and communications (Bright et al., 2018).

7. Summary

Knowledge of PV system characteristics is needed in the different regional PV modelling approaches but are either unknown or only accessible for a small number of stakeholders. The aim of this paper was to provide knowledge of PV system characteristics through data collection, analysis and distribution fitting of PV system characteristics. The structure presented was twofold and focused on (1) metadata (tilt and azimuth of modules, installed capacity and specific annual yield) as well as (2) the impact of shading.

We considered 2,802,797 PV systems located in Europe, USA, Japan and Australia, which represented a total capacity of 59 GWp (14.8% of installed capacity worldwide). Interdependencies of the installed capacity and the geographic location to the other parameters tilt, azimuth and specific annual yield were observed. To acknowledge the impact from these two dominating parameters (installed capacity and geographic location) on others and to allow a derivation of meaningful statistics, a clustering of systems on a country-basis with additional separation by systems sizes $\leq$ 25 kWp and $>$ 25 kWp was introduced. For eased future utilisation of the analysed metadata, each parameter in a cluster was approximated by a distribution function. Results show strong characteristics unique to each cluster, however, there are some commonalities across all clusters. The smallest mean tilt values were reported in Australia (16.1°), USA South and Italy (19.8 and 19.9°, respectively). The largest mean tilt values were reported in Belgium (35.6°), the UK (31.8°) and Germany (31.6°) . We find the smallest mean specific annual yield is in Denmark (786.0 kWh/kWp) and largest in south USA (1,426 kWh/kWp), this corresponds well to the climatic differences between 30 and 50° latitude within the study. The region with smallest me-

dian capacity was UK (2.94 kWp) and the largest was Germany (8.96 kWp). Almost all countries had a mean azimuth angle normal to the equator. The number of equatorially-orientated systems was significantly higher than any other orientation, such that no distribution type could appropriately capture this characteristic. That said, it is expected that the number of systems with azimuth of 0° are exaggerated due to lacking precision of PV system metadata reporting, and perhaps the statistical distributions are more realistic than the data suggests, particularly when considering the reduced peak from higher accuracy metadata, such as that from the UK. Capacity demonstrated the most cluster-unique characteristics. As each cluster represented a country, it also captures national policy incentives that clearly influence the overall capacity distributions. The feed-in tariffs of France, Germany and the UK have clear impact on the PV system size. The shape of the distributions of specific annual yield offered the most similarity between clusters, with the location/mean being primarily a function of climate through latitude. Dissemination of clusters by climate may reveal more insightful differences. All of the distributions that are presented in the paper can be obtained from the tables in the appendix.

Shading was considered by computing the viewshed of individual roof facets of ≈ 48,000 buildings in Uppsala, Sweden, which meant that $> 90,000$ facets were analysed. Two empirical equations were derived and presented. The first represents the beam irradiance subcomponent, describing the mean ratio of a roof that is shaded as a function of the solar elevation angle. The second determines the view factor as a function of the roof tilt including the impact from shading and can be used to estimate the losses of diffuse and reflected irradiance. These equations are believed to better take shading into consideration than the coarse estimates used today. For the specific meteorological conditions of Uppsala, we also showed in this study that losses of diffuse irradiance due to shading are higher than that of beam on an annual basis and should not be neglected for sites of similar cloudiness as in Uppsala (annual clear-sky index of 0.63).

Several interesting research topics beyond the scope of this paper were sketched and the offer for future collaborations expressed.

8. Conflict of interest

The authors declare no conflicts of interest.

Acknowledgement

The authors would like to thank the Australian Renewable Energy Agency (ARENA) for supporting this work (Research and Development Programme Funding G00854). Part of this work is performed within the framework of the IEA-PVPS-Task 13 “Performance and Reliability of Photovoltaic Systems” (IEA-PVPS, 2018). We would also like to acknowledge the Jyukankyo Research Institute Inc. and particularly Mr Takahiro Tsurusaki, Managing Director, COO, for the collaborative spirit shown and for the provision of data for Japan. The authors would like to thank those researchers that were approached who warmly invited the call for collaboration, however, were unable to contribute directly. Dr Joao Gari da Silva

Fonseca Jr, Assistant Professor at the University of Tokyo, was invaluable in guiding us to the Japanese dataset. Rodrigo Palma-Behnke, Director and Assistant Professor of Solar Energy Research Center, Universidad de Chile, was able to provide the international collaborative effort with large scale PV systems information within Chile that was unfortunately outside the scope of this project. Dr Jan Kleissl, Professor at UCSD, USA, for directing us to databases within the USA. We would like to thank Dr Rodrigo Alonso Suárez of UDELAR, Uruguay, and Ricardo Bessa of INESC TEC, Portugal, who demonstrated collaborative spirit but were not able to obtain the data relevant for the study. David Trebosc and the bdpv platform team are thanked for their engagement, high value work and their exemplary policy of openness, which highly contributed to present work. Daniel Decker from the Federal Network Agency in Germany is thanked for the provision of the dataset and his kind support. We would like to thank Mr Ruppel and Mr Ausburg (SMA) for their efforts in gathering plant information while preserving the anonymity of the plants of the sunny portal. Lastly, we are mostly appreciative to those websites that allowed us to freely download their data: openpv.nrel.gov, pvoutput.org, solar-log.com, sonnenertrag.eu, suntrol-portal.com. Another site not used within the study that is an excellent source is https://www.californiadgstats.ca.gov/, however, the USA data collected from openpv.nrel.gov was feared to already incorporate this data and so potential duplication was avoided.

References

Bright, J.M., Babacan, O., Kleissl, J., Taylor, P.G., Crook, R., 2017a. A synthetic, spatially decorrelating solar irradiance generator and application to a LV grid model with high PV penetration. Solar Energy 147, 83–98. doi:10.1016/j.solener.2017.03.018.

Bright, J.M., Killinger, S., Lingfors, D., Engerer, N.A., 2017b. Improved satellite-derived PV power nowcasting using power data from real-time reference PV systems. Solar Energy doi:10.1016/j.solener.2017.10.091.

Bright, J.M., Killinger, S., van Sark, W., Saint-Drenan, Y.M., Moriatis, P., Taylor, J., Engerer, N.A., 2018. ResearchGate project: Characteristics and statistics of PV system metadata . URL: https://www.researchgate.net/project/Characteristics-and-statistics-of-PV-system-metadata.

Bright, J.M., Smith, C.J., Taylor, P.G., Crook, R., 2015. Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data. Solar Energy 115, 229–242. doi:10.1016/j.solener.2015.02.032.

Elsinga, B., van Sark, W., 2015. Spatial power fluctuation correlations in urban rooftop photovoltaic systems. Progress in Photovoltaics: Research and Applications 23, 1390–1397. doi:10.1002/pip.2539.

Elsinga, B., van Sark, W., Ramaekers, L., 2017. Inverse photovoltaic yield model for global horizontal irradiance reconstruction. Energy Science & Engineering 5, 1–14. doi:10.1002/ese3.162.

Freitas, S., Catita, C., Redweik, P., Brito, M.C., 2015. Modelling solar potential in the urban environment: State-of-the-art review. Renewable and Sustainable Energy Reviews 41, 915–931. doi:10.1016/j.rser.2014.08.060.

IEA, 2018. 2018: Snapshot of global photovoltaic markets: Report IEA PVPS T1-33:2018. URL: http://www.iea-pvps.org/fileadmin/dam/public/report/statistics/IEA_PVPS-A_Snapshot_of_Global_PV-1992-2017.pdf.

IEA-PVPS, 2018. Performance and reliability of photovoltaic systems. URL: http://www.iea-pvps.org/index.php?id=57.

Ineichen, P., Perez, R., 2002. A new airmass independent

formulation for the Linke turbidity coefficient. Solar Energy 73, 151–157. doi:10.1016/S0038-092X(02)00045-2.

Jamaly, M., Bosch, J., Kleissl, J., 2013. A power conversion model for distributed PV systems in California using SolarAnywhere irradiation. URL: http://calsolarresearch.ca.gov/images/stories/documents/Sol1_funded_proj_docs/UCSD/perf_model_v11.pdf.

Killinger, S., Braam, F., Müller, B., Wille-Haussmann, B., McKenna, R., 2016. Projection of power generation between differently-oriented PV systems. Solar Energy 136, 153–165. doi:10.1016/j.solener.2016.06.075.

Killinger, S., Bright, J.M., Lingfors, D., Engerer, N.A., 2017a. A tuning routine to correct systematic influences in reference PV systems' power outputs. Solar Energy 157, 1082–1094. doi:10.1016/j.solener.2017.09.001.

Killinger, S., Engerer, N., Müller, B., 2017b. QCPV: A quality control algorithm for distributed photovoltaic array power output. Solar Energy 143, 120–131. doi:10.1016/j.solener.2016.12.053.

Killinger, S., Guthke, P., Semmig, A., Müller, B., Wille-Haussmann, B., Fichtner, W., 2017c. Upscaling PV Power Considering Module Orientations. IEEE Journal of Photovoltaics 7, 941–944. doi:10.1109/JPHOTOV.2017.2684908.

Kühnert, J., 2016. Development of a photovoltaic power prediction system for forecast horizons of several hours. PhD Thesis. Universität Oldenburg. Oldenburg, Germany.

Leloux, J., Narvarte, L., Trebosc, D., 2012a. Review of the performance of residential PV systems in Belgium. Renewable and Sustainable Energy Reviews 16, 178–184. doi:10.1016/j.rser.2011.07.145.

Leloux, J., Narvarte, L., Trebosc, D., 2012b. Review of the performance of residential PV systems in France. Renewable and Sustainable Energy Reviews 16, 1369–1376. doi:10.1016/j.rser.2011.10.018.

Lingfors, D., Bright, J.M., Engerer, N.A., Ahlberg, J., Killinger, S., Widén, J., 2017. Comparing the capability of low- and high-resolution LiDAR data with application to solar resource assessment, roof type classification and shading analysis. Applied Energy 205, 1216–1230. doi:10.1016/j.apenergy.2017.08.045.

Lingfors, D., Killinger, S., Engerer, N.A., Widén, J., Bright, J.M., 2018. Identification of PV system shading using a LiDAR-based solar resource assessment model: an evaluation and cross-validation. Solar Energy 159, 157–172. doi:10.1016/j.solener.2017.10.061.

Lingfors, D., Widén, J., 2016. Development and validation of a wide-area model of hourly aggregate solar power generation. Energy 102, 559–566. doi:10.1016/j.energy.2016.02.085.

Lorenz, E., Scheidsteger, T., Hurka, J., Heinemann, D., Kurz, C., 2011. Regional PV power prediction for improved grid integration. Progress in Photovoltaics: Research and Applications 19, 757–771. doi:10.1002/pip.1033.

Mainzer, K., Killinger, S., McKenna, R., Fichtner, W., 2017. Assessment of rooftop photovoltaic potentials at the urban level using publicly available geodata and image recognition techniques. Solar Energy 155, 561–573. doi:10.1016/j.solener.2017.06.065.

Matlab, 2018. MathWorks: Documentation: Fitdist: Distribution Names . URL: https://mathworks.com/help/stats/fitdist.html#btu538h-distname.

McNeil, I., 1990. An Encyclopaedia of the history of technolology. Routledge reference, Routledge, London, England.

Moraitis, P., Kausika, B., van Sark, Wilfried G. J. H. M., 2015. Visualization of operational performance of grid-connected PV systems in selected European countries, in: 42nd IEEE Photovoltaic Specialists Conference (PVSC), New Orleans, USA. doi:10.1109/PVSC.2015.7355613.

National Grid UK, 2018. Demand data 2018. URL: https://www.nationalgrid.com/uk/electricity/market-operations-and-data/data-explorer.

Nordmann, T., Clavadetscher, L., van Sark, Wilfried G. J. H. M., Green, M., 2014. Analysis of Long-Term Performance of PV Systems: Different Data Resolution for Different Purposes. Report IEA-PVPS T13-05:2014.

ofgem, 2018. Feed-In Tariff (FIT) rates. URL: https://www.ofgem.gov.uk/environmental-programmes/fit/fit-tariff-rates.

Pareek, S., Chaturvedi, N., Dahiya, R., 2017. Optimal interconnections to address partial shading losses in solar photovoltaic arrays. Solar Energy 155, 537–551. doi:10.1016/j.solener.2017.06.060.

Paulescu, M., Paulescu, E., Gravila, P., Badescu, V., 2012.

Weather modeling and forecasting of PV systems operation. Green Energy and Technology, Springer, Dordrecht, Germany.

Pfenninger, S., Staffell, I., 2016. Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. Energy 114, 1251–1265. doi:10.1016/j.energy.2016.08.060.

Reich, N.H., Mueller, B., Armbruster, A., van Sark, Wilfried G. J. H. M., Kiefer, K., Reise, C., 2012. Performance ratio revisited: is PR > 90% realistic? Progress in Photovoltaics: Research and Applications 20, 717–726. doi:10.1002/pip.1219.

Saint-Drenan, Y.M., 2015. A probabilistic approach to the estimation of regional photovoltaic power generation using meteorological data: Application of the Approach to the German Case. Ph.D. thesis. University of Kassel. Kassel, Germany. URL: https://kobra.bibliothek.uni-kassel.de/bitstream/urn:nbn:de:hebis:34-2016090550868/3/DissertationYMSaintDrenan.pdf.

Saint-Drenan, Y.M., Fritz, R., Jost, D., 2015. Auswertung des Effekts der Sonnenfinsternis vom 20.03.2015 auf das deutsche Energieversorgungssystem. URL: http://www.energiesystemtechnik.iwes.fraunhofer.de.

Saint-Drenan, Y.M., Good, G.H., Braun, M., 2017. A probabilistic approach to the estimation of regional photovoltaic power production. Solar Energy 147, 257–276. doi:10.1016/j.solener.2017.03.007.

Saint-Drenan, Y.M., Good, G.H., Braun, M., Freisinger, T., 2016. Analysis of the uncertainty in the estimates of regional PV power generation evaluated with the upscaling method. Solar Energy 135, 536–550. doi:10.1016/j.solener.2016.05.052.

Saint-Drenan, Y.M., Wald, L., Ranchin, T., Dubus, L., Troccoli, A., 2018. An approach for the estimation of the aggregated photovoltaic power generated in several European countries from meteorological data. Advances in Science and Research 15, 51–62. doi:10.5194/asr-15-51-2018.

Schläpfer, M., Lee, J., Bettencourt, L.M.A., 2015. Urban Skylines: building heights and shapes as measures of city size. URL: http://arxiv.org/abs/1512.00946, arXiv:1512.00946.

Schubert, G., 2012. Modellierung der stündlichen Photovoltaik- und Windstromeinspeisung in Europa, in: 12. Symposium Energieinnovation, Graz, Austria.

SMHI, 2015. STRÅNG - a mesoscale model for solar radiation. URL: http://strang.smhi.se/.

Šúri, M., Huld, T.A., Dunlop, E.D., Ossenbrink, Heinz A., 2007. Potential of solar electricity generation in the European Union member states and candidate countries. Solar Energy 81, 1295–1305. doi:10.1016/j.solener.2006.12.007.

Swedish Land Survey, 2015. Produktbeskrivning: Laserdata [Product description LiDAR data] ver. 2.2. Technical Report 12. Swedish Land Survey. Gävle. URL: http://www.lantmateriet.se/globalassets/kartor-och-geografisk-information/hojddata/produktbeskrivningar/laserdat.pdf.

Swedish Land Survey, 2016. Produktbeskrivning GSD-Fastighetskartan, vektor [Product description Property Map, vectorized]. Technical Report. Swedish Land Survey. Gävle. URL: https://www.lantmateriet.se/globalassets/kartor-och-geografisk-information/kartor/produktbeskrivningar/fastshmi.pdf.

Taylor, J., 2015. Performance of distributed PV in the UK: a statistical analysis of over 7000 systems, in: 31st European Photovoltaic Solar Energy Conference and Exhibition, Hamburg, Germany.

Tsafarakis, O., Moraitis, P., Kausika, B.B., van der Velde, H., 't Hart, S., de Vries, A., de Rijk, P., de Jong, M.M., van Leeuwen, H.P., van Sark, W., 2017. Three years experience in a Dutch public awareness campaign on photovoltaic system performance. IET Renewable Power Generation 11, 1229–1233. doi:10.1049/iet-rpg.2016.1037.

Yang, D., Dong, Z., Reindl, T., Jirutitijaroen, P., Walsh, W.M., 2014. Solar irradiance forecasting using spatio-temporal empirical kriging and vector autoregressive models with parameter shrinkage. Solar Energy 103, 550–562. doi:10.1016/j.solener.2014.01.024.

Appendix A. Approximation of parameters in clusters: continued

The appendix is a continuation of section 4 and provides histograms together with fitted distribu-

tions of azimuth, tilt, specific annual yield and installed capacity for system sizes > 25 kWp in Figure A.10. Table A.4 presents the coefficients of the fitted distributions, while distinguishing between systems ≤ 25 kWp and > 25 kWp. All distribution functions are defined in Table A.3 and can be replicated with help of the parametrised coefficients. It is important that the user reads carefully section 4 in order to appropriately use the distributions. A summary of the impact of our proposed quality control criteria from section 2 is provided in the appendix in Table A.5 where percentages of removed data are presented.

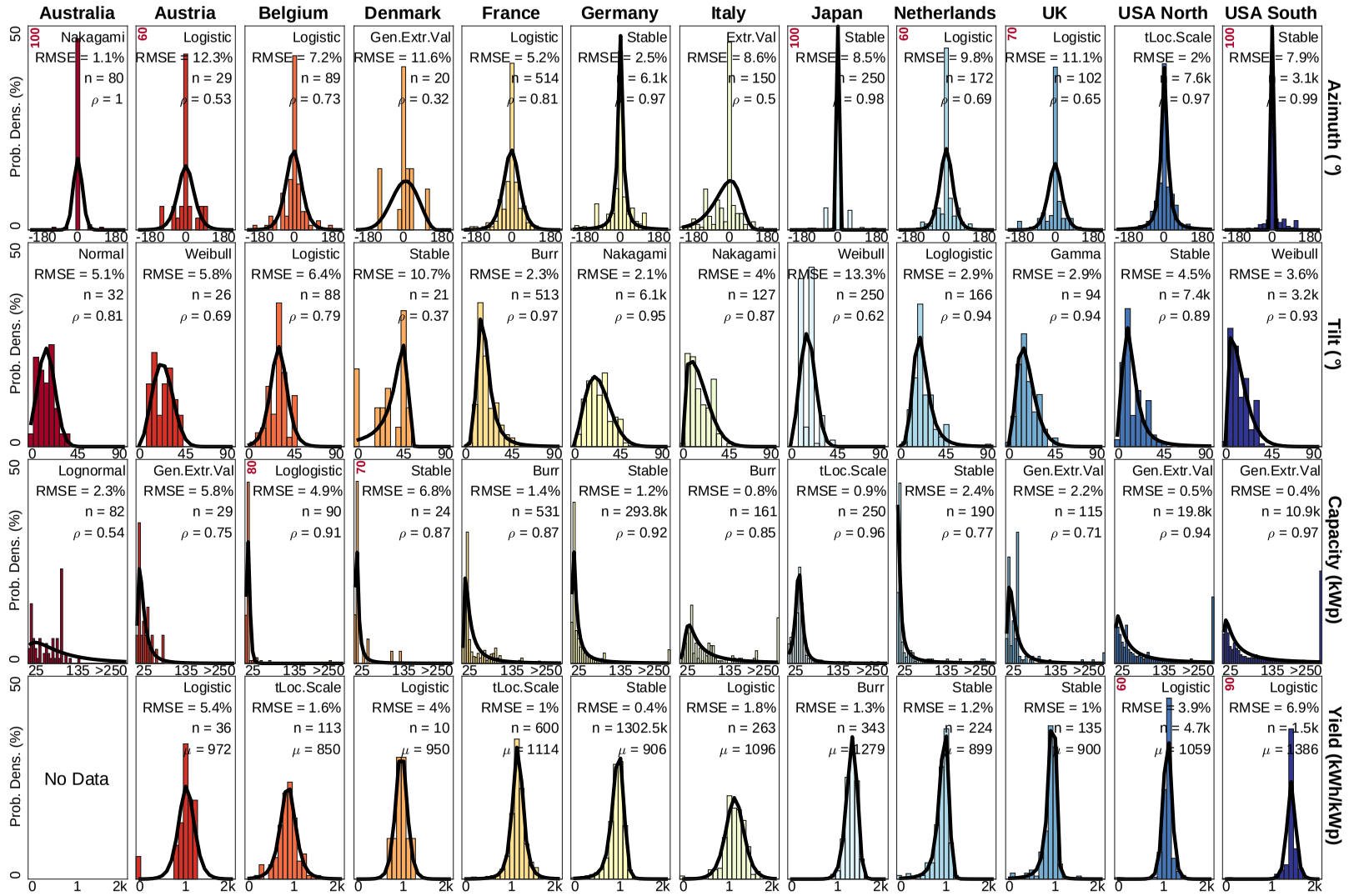

Figure A.10: Histograms of real data (bar) with approximated probability distributions (line) for the different clusters (columns) and parameters (rows), where capacity is the installed capacity and yield is the specific annual yield. All systems reported within this figure have a capacity > 25 kWp, see section 4 for complete derivation. Within each of the axes is reported the name of the best fitting distribution type, the root mean squared error (RMSE) between the scatter of real data in the histogram against the fitted probability density distribution, the number of data points considered for that cluster (n), and the Pearson correlation coefficient (ρ) of the linear regression. The mean value μ is shown in place of ρ for Yield. All y-axes are scaled between 0 and 50% probability except where a bold red value is assigned to the individual axis.

Table A.3: Definition of the probability density distributions used in the research. The coefficients correspond to those presented in Table A.4. The distribution name corresponds to the same Matlab®distribution names and readers are encourage to read the detailed descriptions at WWW.MATHWORKS.COM/HELP/STATS/CONTINUOUS-DISTRIBUTIONS.HTML. Each coefficient is defined. The equation is provided from the Matlab®documentation. Note that the Stable distribution is not explicitly a probability density function, but a characteristic function.

Distribution Name	Coeff. 1	Coeff. 2	Coeff. 3	Coeff. 4	Probability Density Function, $f(x\|...)$
Burr type XII	α Scale	c Shape 1	k Shape 2	-	$f(x\|\alpha,c,k)=\frac{\frac{kc}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1}}{\left(1+\left(\frac{x}{\alpha}\right)^{c}\right)^{k+1}}$
Extreme Value	μ Location	σ Scale	-	-	$f(x\|\mu,\sigma)=-\sigma^{-1}\exp\left(\frac{x-\mu}{\sigma}\right)\exp\left(-\exp\left(\frac{x-\mu}{\sigma}\right)\right)$
Gamma	a Shape	b Scale	-	-	$f(x\|a,b)=\frac{1}{b^{a}\Gamma(a)}x^{a-1}\exp\left(\frac{-x}{b}\right)$
Generalized Extreme Value	k Shape	σ Scale	μ Location	-	$f(x\|k,\mu,\sigma)=$ $\left(\frac{1}{\sigma}\right)\exp\left(-\left(1+k\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{k}}\right)\left(1+k\frac{(x-\mu)}{\sigma}\right)^{-1-\frac{1}{k}}$
Inverse Gaussian	μ Scale	λ Shape	-	-	$f(x\|\mu,\lambda)=\sqrt{\frac{\lambda}{2\pi x^{3}}}\exp\left(-\frac{\lambda}{2\mu^{2}x}(x-\mu)^{2}\right)$
Logistic	μ Location	σ Scale	-	-	$f(x\|\mu,\sigma)=\frac{\exp\left(\frac{x-\mu}{\sigma}\right)}{\sigma\left(1+\exp\left(\frac{x-\mu}{\sigma}\right)\right)^{2}}$
Loglogistic	μ Log Loc.	σ Log Scale	-	-	$f(x\|\mu,\sigma)=\frac{1}{\sigma}\frac{1}{x}\frac{\exp\left(\frac{\log(x)-\mu}{\sigma}\right)}{\left(1+\exp\left(\frac{\log(x)-\mu}{\sigma}\right)\right)^{2}}\frac{1}{\Gamma(\mu)}x^{2\mu-1}$
Lognormal	μ Log Loc.	σ Log. Scale	-	-	$f(x\|\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(\frac{-(\ln x-\mu)^{2}}{2\sigma^{2}}\right)$
Nakagami	μ Shape	ω Scale	-	-	$f(x\|\mu,\omega)=2\left(\frac{\mu}{\omega}\right)^{\mu}\frac{1}{\Gamma(\mu)}x^{(2\mu-1)}\exp\left(\frac{-\mu x^{2}}{\omega}\right)$
Normal	μ Location	σ Scale	-	-	$f(x\|\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^{2}}{2\sigma^{2}}\right)$
Stable	α Shape 1	β Shape 2	γ Scale	δ Location	$E(e^{itX})=$ $\exp\left(-\gamma^{\alpha}\lvert t\rvert^{\alpha}(1+i\beta\mathrm{sgn}(t)\tan\frac{\pi\alpha}{2}\left((\gamma\lvert t\rvert)^{1-\alpha}-1\right)i\delta t\right)$
t Location-Scale	μ Location	σ Scale	ν Deg. of Freedom	-	$f(x\|\mu,\sigma,\nu)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\nu+\left(\frac{x-\mu}{\sigma}\right)^{2}}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$
Weibull	a Scale	b Shape	-	-	$f(x\|a,b)=\frac{b}{a}\left(\frac{x}{a}\right)^{b-1}\exp\left(-\left(\frac{x}{a}\right)^{b}\right)$

Table A.4: The distribution coefficients corresponding to the definitions in Table A.3. The left side is for capacities ≤ 25 kWp and the right side is for > 25 kWp.

Cluster	Parameter	Dist. Type	Coeff. 1	Coeff. 2	Coeff. 3	Coeff. 4	Dist. Type	Coeff. 1	Coeff. 2	Coeff. 3	Coeff. 4
Australia	Azimuth	Logistic	8.3472	34.2603			Nakagami	28.0028	0.2537		
	Tilt	Logistic	15.0891	6.0136			Normal	14.6875	8.4183		
	Capacity	tLocationScale	4.9355	1.4855	3.2434		Lognormal	4.3491	0.8328		
Austria	Azimuth	Logistic	0.4566	17.6159			Logistic	0.6107	19.9366		
	Tilt	Weibull	34.6425	3.6381			Weibull	26.2683	2.7083		
	Capacity	GeneralizedExtremeValue	0.1633	2.9111	4.8713		GeneralizedExtremeValue	0.4337	8.5691	33.5998	
	FLH	tLocationScale	1.0668	0.0971	2.0439		Logistic	1.0144	0.1090		
Belgium	Azimuth	Logistic	-1.2365	24.3969			Logistic	-3.5828	19.3034		
	Tilt	ExtremeValue	39.8100	7.8703			Logistic	29.6932	5.1091		
	Capacity	Stable	1.4734	0.9578	1.3688	4.8772	Loglogistic	3.3900	0.0905		
	FLH	Burr	0.5814	9.2775	5.8277		tLocationScale	0.8570	0.1623	3.8904	
Denmark	Azimuth	Logistic	1.5526	20.6455			GeneralizedExtremeValue	-0.3684	49.6733	-13.0930	
	Tilt	Loglogistic	-1.1441	0.2256			Stable	0.9410	-1.0000	6.5572	41.6962
	Capacity	tLocationScale	5.7414	1.5014	2.3622		Stable	0.9767	1.0000	2.8140	28.6026
	FLH	Stable	1.2818	-0.8035	0.0948	0.8767	Logistic	0.9500	0.0778		
France	Azimuth	tLocationScale	0.3346	34.5161	5.4215		Logistic	-2.2831	19.1336		
	Tilt	Lognormal	-1.2139	0.4144			Burr	16.5673	5.0314	0.6793	
	Capacity	Stable	0.4000	0.5250	0.0345	3.0118	Burr	27.9585	19.1798	0.0934	
	FLH	tLocationScale	1.1011	0.1416	7.6129		tLocationScale	1.1194	0.1111	2.9295	
Germany	Azimuth	Logistic	-0.1366	23.0048			Stable	1.0309	-0.0002	9.9430	-0.2506
	Tilt	ExtremeValue	38.0648	9.6547			Nakagami	1.1890	0.0887		
	Capacity	GeneralizedExtremeValue	0.1143	3.5745	6.2413		Stable	0.7373	1.0000	5.2331	30.8089
	FLH	Stable	1.6611	-0.9423	0.1152	0.9086	Stable	1.5803	-0.8322	0.0902	0.9530
Italy	Azimuth	tLocationScale	-5.2371	32.9151	2.1254		ExtremeValue	4.4388	46.3326		
	Tilt	Stable	1.8917	1.0000	5.5359	19.0462	Nakagami	0.6147	0.0448		
	Capacity	Stable	0.9088	0.9411	1.5334	4.4130	Burr	37.2635	8.0048	0.1480	
	FLH	tLocationScale	1.1774	0.1415	2.5624		Logistic	1.1096	0.1246		
Japan	Azimuth	Logistic	-0.7254	15.9481			Stable	0.4045	0.0010	0.2121	0.0007
	Tilt	ExtremeValue	27.6791	6.4206			Weibull	18.9055	2.6245		
	Capacity	Stable	1.3094	0.9774	1.0384	4.4144	tLocationScale	48.0538	7.9913	1.4353	
	FLH	Burr	1.3730	11.1707	2.7755		Burr	1.4833	13.7688	5.1025	
Netherlands	Azimuth	Stable	1.0309	0.0001	11.2285	0.0294	Logistic	-0.3912	15.7031		
	Tilt	ExtremeValue	38.8465	11.6153			Loglogistic	3.0459	0.2311		
	Capacity	tLocationScale	3.2381	1.3462	1.6898		Stable	0.6078	1.0000	3.7179	28.2348
	FLH	Stable	1.3553	-0.8354	0.0893	0.9334	Stable	1.4819	-0.9477	0.0760	0.9577
UK	Azimuth	Logistic	0.2461	26.0886			Logistic	-0.8287	16.4894		
	Tilt	tLocationScale	31.5952	4.6591	3.3818		Gamma	3.9484	4.5939		
	Capacity	Stable	1.8937	-0.0544	0.5898	2.9504	GeneralizedExtremeValue	0.5210	10.8742	35.6490	
	FLH	Stable	1.6851	0.0000	0.0776	0.9035	Stable	1.3744	-0.8038	0.0638	0.9395
USA North	Azimuth	Logistic	0.4600	28.1468			tLocationScale	0.1686	12.6681	1.2295	
	Tilt	Loglogistic	-1.2944	0.2061			Stable	1.3001	1.0000	4.8192	10.6480
	Capacity	GeneralizedExtremeValue	0.0868	2.5493	4.6450		GeneralizedExtremeValue	1.1806	31.0335	45.3525	
	FLH	Stable	1.3201	-0.7411	0.0666	1.0628	Logistic	1.0679	0.0575		
USA South	Azimuth	Logistic	8.5676	29.3711			Stable	0.4000	0.4410	0.2829	0.0822
	Tilt	tLocationScale	20.4150	3.8106	3.0306		Weibull	0.1647	1.3876		
	Capacity	Burr	0.3092	2.5724	2.2109		GeneralizedExtremeValue	1.2834	38.6717	49.6038	
	FLH	ExtremeValue	1.4898	0.0963			Logistic	1.4035	0.0579		

Table A.5: Percentages of reported (Rep.) and filtered data due to the quality control procedure sketched in section 2 for tilt, azimuth, installed capacity (Capa.), geographic location (Loca.) and specific annual yield. Numbers are given as percentage and in relation to the total number of systems that were available in each dataset.

Region	Tilt			Azimuth			Capa.	Loca.	Yield			Source
	Rep.	$\leq 1°$	$> 89°$	Rep.	$< -179°$	$> 179°$	Rep.	Rep.	Rep.	=0	> 2000	
Australia	26.95	0.00	0.00	26.95	0.00	0.00	100.00	96.40	0.00	0.00	0.00	pvoutput.org
Austria	100.00	2.34	0.00	100.00	2.60	0.00	100.00	100.00	27.64	0.00	0.00	solar-log.com
	100.00	0.71	1.07	100.00	0.00	0.71	100.00	100.00	0.00	0.00	0.00	suntrol-portal.com
	100.00	1.87	0.00	100.00	1.12	0.00	100.00	100.00	95.65	39.93	0.00	sonnenertrag.eu
	83.04	28.57	0.89	93.75	6.25	0.00	100.00	67.86	24.11	0.60	0.30	pvoutput.org
Belgium	100.00	1.65	0.44	100.00	0.99	0.00	100.00	95.77	49.36	0.00	0.00	solar-log.com
	100.00	0.48	0.06	100.00	0.48	0.00	100.00	100.00	98.08	67.74	0.03	bdpv.fr
	100.00	1.48	0.00	100.00	0.74	0.00	100.00	99.63	97.41	41.47	0.00	sonnenertrag.eu
Denmark	100.00	2.25	0.11	100.00	2.57	0.00	100.00	100.00	25.70	0.00	0.00	solar-log.com
	100.00	0.63	0.00	100.00	0.00	1.27	100.00	100.00	0.00	0.00	0.00	suntrol-portal.com
	93.54	11.62	0.00	98.34	4.43	0.00	100.00	68.45	54.98	0.74	0.00	pvoutput.org
France	100.00	0.25	0.10	99.99	0.61	0.17	100.00	100.00	97.38	60.23	0.16	bdpv.fr
	100.00	4.84	0.36	100.00	5.02	0.00	100.00	63.44	23.26	0.00	0.00	solar-log.com
Germany	0.00	0.00	0.00	0.00	0.00	0.00	100.00	99.78	90.75	0.26	0.01	bundesnetzagentur.de
	100.00	2.30	0.49	100.00	2.77	0.10	100.00	100.00	39.31	0.00	0.00	solar-log.com
	100.00	0.76	0.06	100.00	0.00	1.01	100.00	100.00	0.00	0.00	0.00	suntrol-portal.com
	100.00	1.72	0.09	100.00	1.13	0.03	100.00	99.97	98.89	39.36	0.00	sonnenertrag.eu
Italy	100.00	3.63	0.56	100.00	4.71	0.00	100.00	86.91	33.49	0.00	0.00	solar-log.com
	86.42	26.12	0.09	96.35	4.59	0.00	100.00	78.84	28.40	1.03	0.06	pvoutput.org
	100.00	1.40	0.00	100.00	0.56	0.00	100.00	100.00	98.14	41.15	0.00	sonnenertrag.eu
Japan	100.00	0.00	0.00	100.00	0.32	0.00	100.00	100.00	35.56	0.00	0.00	jyuri.co.jp
Netherlands	92.08	18.44	0.10	97.09	7.08	0.00	100.00	55.06	13.43	0.35	0.03	pvoutput.org
	100.00	1.26	0.45	100.00	1.79	0.09	100.00	100.00	43.14	0.00	0.00	solar-log.com
	100.00	0.55	0.22	100.00	0.55	0.00	100.00	100.00	96.04	52.02	0.00	sonnenertrag.eu
	100.00	0.69	0.34	100.00	0.00	0.34	100.00	100.00	0.00	0.00	0.00	suntrol-portal.com
Rest of	100.00	7.30	1.76	100.00	13.18	0.00	100.00	72.88	18.19	0.00	0.00	solar-log.com
Europe	88.16	32.33	0.00	96.29	21.38	0.00	100.00	62.37	21.61	3.06	0.12	pvoutput.org
	100.00	3.76	0.00	100.00	5.26	0.00	100.00	100.00	97.24	36.09	0.00	sonnenertrag.eu
UK	99.15	0.06	0.00	100.00	0.08	0.00	100.00	100.00	100.00	0.21	0.38	microgen-database.sheffield.ac.uk
	81.71	16.23	0.00	97.16	4.64	0.00	100.00	62.07	4.33	0.28	0.00	pvoutput.org
USA	37.56	0.05	0.01	35.59	0.00	0.03	100.00	94.51	20.03	0.00	0.53	openpv.nrel.gov
	93.57	20.63	0.00	98.30	7.54	0.00	100.00	65.53	7.23	0.44	0.12	pvoutput.org