Solar Energy 265 (2023) 112132

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

A Markov Chain Monte Carlo approach for the estimation of photovoltaic system parameters

Benjamin P.M. Laevens [a,*], Frank P. Pijpers [b], Harm Jan Boonstra [a], Wilfried G.J.H.M. van Sark [c], Olav ten Bosch [b]

[a] Statistics Netherlands, CBS-weg 11, 6412EX Heerlen, The Netherlands
[b] Statistics Netherlands, Henri Faasdreef 312, 2492JP The Hague, The Netherlands
[c] Copernicus Institute, Utrecht University, Princetonlaan 8A, 3584CB Utrecht, The Netherlands

ARTICLE INFO

Keywords:
PV systems
Azimuth
Tilt
Markov Chain Monte Carlo

ABSTRACT

Knowledge of the installation parameters of photovoltaic systems is essential in the context of grid management: by relating these parameters to performance data, forecasting models may be optimised to improve the management of power flow into the grid. In the case of small residential systems, these parameters are often not available. We present a novel method for determining the azimuth (ϕ), tilt (θ) and rated power (P) of photovoltaic systems, using openly available data over the course of 2016–2018 of 12 photovoltaic systems in `PVOutput`. This method consists of two steps: firstly we identify a candidate list of clear days by computing descriptive statistics of a larger set of 80 `PVOutput` system profiles. In the second step we compare the observed clear-day profiles, of the aforementioned 12 systems, with modelled clear-sky profiles from the `PVLib` library. The fits are performed employing a Markov Chain Monte Carlo (MCMC) approach, implemented with the `Emcee` package: the most favoured parameters and their associated uncertainties, for any given day, are obtained by sampling from the posterior assuming a Gaussian sampling distribution. The results for our 12 systems are in good agreement with the `PVOutput` metadata.

1. Introduction

The global installed photovoltaic (PV) capacity reached the terawatt level in early 2022 [1], and present annual markets of around 200 GW_p are expected to increase further [2]. Such market growth is deemed necessary for tens of TW_p of PV power by mid-century in support of a 100% renewables powered society [2–6]. In the current fossil-fuel dominated electrical power generation system, centralised large power plants are well monitored and controlled, which allows proper electricity demand following by means of various electricity market mechanisms. The future, decentralised electrical power generation system will rely on millions of PV systems, ranging from small-sized (1 – 5 kW_p) to large-sized (up to several GW_p) PV parks, complemented with wind parks of 10+ MW each. For large size PV and wind parks, proper generation data are available and are used in electricity markets. Operators are well aware of the proper functioning of these systems. In contrast, about half of the grid-connected PV systems in Europe are small-size (<15 kW_p) residential systems [2], for which power generation data is not centrally available. Owners are typically not aware to what extent their PV system is functioning as it should. Analyses of many PV systems have been performed to reveal that overall, these small-sized systems are functioning properly [7–9]. In addition, these studies have revealed that so-called metadata, i.e., system size (P), azimuth (ϕ) and tilt (θ) are either missing or, when available, impacted by various sources of uncertainty: differences in definition regarding North, large rounding errors or imputed dummy information where 0° is often a placeholder. As a consequence, prediction of power (energy) generation cannot be done correctly to compare with actual power (energy) generation in order to assess PV system health. It is therefore necessary to develop methods which automatically derive these metadata directly from power (energy) generation time series.

1.1. Previous research

There is a large amount of literature concerned with the derivation of installation parameters of PV systems. Meng et al. [10] provide an excellent overview of the various methods that have been developed over the years. More specifically, they identify four broad categories, the last of which we will explore in more detail since our method belongs to this category:

* Corresponding author.
E-mail address: blaevens@ukaachen.de (B.P.M. Laevens).

https://doi.org/10.1016/j.solener.2023.112132
Received 13 February 2023; Received in revised form 26 May 2023; Accepted 17 October 2023
Available online 1 November 2023

Nomenclature

Abbreviations

KNMI	Royal Netherlands Meteorological Institute
MCMC	Markov Chain Monte Carlo
MeAE	Median Absolute Error
NSI	National Statistics Institute
pc4	Postal code 4 area in the Netherlands
PV	Photovoltaic
RMSE	Root Mean Squared Error
TDP	Typical Daily Profile
mod	Modelled
obs	Observed
pred	Predicted
ref	Reference

Symbols

ϕ	Azimuth angle of PV system (°)
θ	Tilt angle of PV system (°)
N_p	Number of panels of PV system
P	Total power of PV system (W_p)
P_p	Power of PV panel (W_p)
P_i	Inverter size of PV system (W)
E_{inst}	Instantaneous energy production of PV system (kWh)
E_{cum}	Cumulative energy production of PV system (kWh)
N	Number of measurement points per energy profile of a PV system
l	Longitude (°)
b	Latitude (°)
k_t	Clearness Index
GHI	Global horizontal irradiance ($W\,m^{-2}$)
DHI	Direct horizontal irradiance ($W\,m^{-2}$)
β	Model parameters (ϕ, θ, N_p) in the MCMC simulation
σ	Nuisance parameter in the MCMC simulation
$p(\beta\|D)$	Posterior probability in Bayes' rule
$p(\beta)$	Prior probability in Bayes' rule
$p(D\|\beta)$	Likelihood in Bayes' rule
$p(D)$	Marginal probability in Bayes' rule
σ_β^2	Variance of the MCMC samples
N_{iter}	Number of iterations of the MCMC samples after burn in
K	Number of retained draws from the posterior distribution
$\langle f(\beta)\rangle$	Monte Carlo approximation to the posterior mean
τ	Integrated autocorrelation time
σ_c	Standard deviation in a sigma clipping procedure
E_n	Energy measurement of a PV profile for index n
$E_{n_{max}}$	Peak (daily) energy measurement for index n_{max}
E_{morn}	All PV energy measurements before $E_{n_{max}}$
E_{aft}	All PV energy measurements after $E_{n_{max}}$
ΔE_{morn}	Ensemble of gradients between successive points before $E_{n_{max}}$
ΔE_{aft}	Ensemble of gradients between successive points after $E_{n_{max}}$
Q_{morn}	Fraction of positive gradient points before $E_{n_{max}}$
Q_{aft}	Fraction of negative gradient points after $E_{n_{max}}$
C_i	Binary coefficient to calculate Q_{morn} or Q_{aft}

1. Manual measurement: parameters are obtained using various apparatus such as bubble levels.
2. Digital elevation models: roof parameters, which serve as proxies for the PV system parameters, are identified with 3D spatial representations.
3. Text extraction: metadata files from various monitoring projects may be used to extract parameters.
4. Parametrisation: Observed PV system profiles can be combined with other information e.g. modelled profiles, thus obtaining the parameters [10, and references therein].

The parametrisation approach is the most successful method for a variety of reasons which we briefly explain by contrasting this method with the downsides of the first three. Manual measurement is not appealing since it requires the physical presence of someone measuring the PV system parameters at various sites. Not only is it costly because of the labour involved, it can only be scaled up with even more labour. The digital elevation models are computationally intensive and depend greatly on high resolution scales for the methods to work. While text extraction can provide answers fairly quickly, there are often question marks as to the reliability of these data. Killinger et al. [8] found that disproportionately high numbers of flat roofs exist in metadata, when in reality those same roofs appear not to be flat. Seemingly, when θ is unknown, flat roofs become the default.

Two works that are highly relevant in the context of our new method are those by Saint-Drenan et al. [11] and Ruelle et al. [12]. Both of these research papers have in common that they use modelled and observed profiles to determine PV system parameters. Saint-Drenan et al. determine PV system parameters such as ϕ and θ by selecting the modelled profile which best characterises the observed data or profile. They achieve this by finding the global minimum of a cost function. The modelled profiles are constructed by combining GHI with module parameters (ϕ and θ) and an angular loss coefficient to predict what the effective irradiation would be. Look-up tables, which quantify the combination of real-life effects of PV systems are used to calculate what the power output ought to be as a function of the irradiation and temperature measurements, which are also used in the calculation model.

Ruelle et al. similarly compare modelled and observed profiles, but in a different way. The modelled profiles here rely on the Python software package `PVLib` [13], which is also the basis for modelled profiles in our research work. First, the authors use weather data to filter out all the cloudy days. For this they rely on clearness indices. Then, they remove those parts of the days that suffer from shading. They do this by analysing the shapes of the observed profiles. Finally, they retain the moments of the day that produce maximum power. This is filtered in to an optimisation algorithm which calculates the normalised mean absolute error between the observed peak power and the modelled profile, which is a function of various parameters such as ϕ, θ and P.

Other methods include that described by Meng et al. One of the advantages of their method is that it does not rely on weather data, obtained at the site location. Instead, hourly off-site irradiance data from nearby weather stations are used to calculate the plane of array irradiance as a function of different ϕ and θ, using the Perez et al. [14] model. Over 30 000 time series are produced for any combination of $0° \leq \phi < 360°$ and $0° \leq \theta \leq 90°$ with $\Delta\phi, \Delta\theta = 1°$. Thus curve matching is performed between all these modelled and observed profiles, through calculating the root mean squared error (RMSE). For each day of the year (2017), the daily diffuse fraction is calculated which is the ratio of

diffuse to total irradiance. The procedure is subsequently performed on the clearest day of each month. Per day, 1% of the best fitting curves with low RMSE are retained. This figure is adopted to allow seasonal as well as secondary effects to be taken into account. The favoured set of parameters for a PV system is determined by comparing the overlap between the 12 top 1% parameter combinations. The maximum number of overlaps is determined and thus all associated parameter combinations are retained. A mean value is determined as a point estimate of the azimuths and tilts of PV systems. We will return to Meng et al.'s method in Section 6.2, when we benchmark our results.

Brecl et al. [15] use typical daily profiles (TDP's) [16] to determine azimuth and tilt angles in Slovenia. In this particular case, TDP's are constructed (per system) by superimposing energy yield measurement data (with hourly resolutions) for every single day of a particular month (in this case June). Per hour bracket, the measurements are ranked and the 90th percentile is selected, where the 90th percentile is assumed to represent energy yields corresponding to clear-sky conditions. Next, irradiance profiles are constructed for all combinations of azimuth and tilt, in 5° steps for a particular PV system's location, which is known from ZIP codes. For each of these profiles, the hour is identified at which maximum irradiance occurs for that particular combination of azimuth and tilt. The peak of these profiles is compared to noon and a time deviation is computed between the two. A first filtering takes place by retaining only those irradiance profiles which have a time deviation that is equal to the time deviation of the TDP's profile from noon. Finally, taking into account shading and temperature, these remaining profiles are compared to the TDP. More specifically the azimuth and tilt angle is chosen that minimises the difference between the irradiance and TDP profiles. For further information on other existing methods, we refer the reader to references quoted in the overview of Meng et al.

1.2. Solar energy case studies at Statistics Netherlands

Our motivation for developing a new method to determine PV system parameters comes from a slightly different perspective when comparing to the reasons set out in Section 1. Over recent years, national statistics institutes (NSI's) have seen a big change in the way their statistics are compiled due to a wealth of new and readily available data sources. Whereas in the past, the production of statistics was contingent on survey data in the form of e.g. questionnaires, now these have been steadily replaced by databases, e.g. registers, compiled or obtained by these NSI's. When the initiative of collecting such data was not instigated by the NSI, but rather, the data were obtained from a third party, the question of metadata reliability inevitably emerges. This could be for a variety of reasons. For example, the original purpose of data collection by said third party was different to that of compiling statistics. Therefore there may have been a reduced incentive to validate the reliability of certain metadata fields. NSI's are, hence, often tasked with developing new techniques which can verify or deduce certain metadata. An example, in the context of solar energy, is a recent initiative at Statistics Netherlands to use machine learning techniques to identify PV systems [17] and compare these to a PV systems database containing almost all photovoltaic systems in the Netherlands [18].

Another reason we wish to develop a method to determine PV system parameters is related to ongoing efforts at our NSI to improve national statistics with regards to generated solar energy from PV systems. In a previous paper, Laevens et al. [19], we presented a new method which combines meteorological data with a non-probability sample of `PVOutput` data [20], allowing a model to be constructed which determined the most likely yield given the irradiance observed on any given day. This information was subsequently applied to our PV systems database at Statistics Netherlands, allowing solar energy production per day and municipality to be determined. The computation of these yields depended on the accurate imputation of ϕ, θ and P by `PVOutput` owners. The system size was important when determining daily specific yields of the sample which were subsequently used to extrapolate to the whole population of PV systems in the database, whereas ϕ and θ were employed to assess the sensitivity of results when selecting different subpopulations e.g. all systems facing southwards. By selecting different subpopulations, we obtained annual specific yields that ranged from 877 $\mathrm{kWh\,kW_p^{-1}}$ to 943 $\mathrm{kWh\,kW_p^{-1}}$ in 2016. While θ and P are recorded in the metadata with a high accuracy (e.g. for θ this is 1°), ϕ is limited to a choice of four cardinal and four intercardinal directions, which translates into an accuracy of 45°. It is therefore conceivable that azimuths are not always accurately recorded in `PVOutput`.

1.3. Research objectives and paper outline

The objectives of this research are to devise a new method that will perform parameter estimation for the following three PV system variables: azimuth (ϕ), tilt (θ) and the number of panels (N_p). In this paper we demonstrate the practical application of this method for 12 different `PVOutput` systems in the Netherlands. By selecting these 12 systems, the relevant parameter space is adequately covered and remains computationally tractable. Further details are in Section 4.1. This new method takes into account two different challenges that have been identified in Section 1.1. The first challenge pertains to the availability of different data sources whereas the second challenge is related to the type of (statistical) method employed. We explore these two elements in the following two paragraphs, while highlighting what sets apart our method in terms of improvements to solve these aforementioned challenges.

The feasibility, reliability and therefore success of any new method (besides that it is statistically sound) will depend largely on accessibility to different data sources. An ideal situation is to have the least number of data sources without any (or a small amount of) trade-off in the method itself. It is generally difficult to come by data sources, such as on-site irradiance and temperature data. On top of this, these sources can vary wildly from country to country, making a generalised method more difficult to obtain. Even when general data sources are available in many countries, such as weather station data or modelled irradiance data, these come with practical limitations. In the Netherlands, for example, weather station irradiance data are limited to some thirty different station locations [21] around the country and need not be representative of a PV system location. Additionally, these data are only available on an hourly basis. While irradiance data, derived from satellite data [22,23], are available on a higher temporal (every 15 mins) and spatial (3×6 km^2) frequency, these are not available when the solar elevation is lower than 12° [23].

Secondly, a new technique must probe parameter space efficiently such that convergence of a solution may be obtained in a fast way. What many of the past curve matching techniques have in common is that they must be agnostic about the guess parameters of θ, ϕ and P (or N_p, the number of panels). As such, all possible solutions are considered as likely as each other and the methods consequently consider all parts of parameter space as equally likely. Concretely this means, in the example of Meng et al. [10], that modelled profiles need to be obtained for all angles of ϕ and θ for the analysis to work. It should be noted that Brecl et al. go some way to solving this issue by performing a filtering process to eliminate unlikely PV system configurations for a given observed profile.

Our new method addresses the issue of data availability by relying only on two different data sources: observed PV generation data, from `PVOutput`, and synthetic modelled profiles from an easily accessible Python library called `PVLib`. For clear-sky days, we determine the PV system parameters by comparing observed and modelled profiles through the application of a Markov Chain Monte Carlo approach in a Bayesian context. The fact that our approach is Bayesian allows us to avoid probing all of parameter space equally. As mentioned earlier, we are interested in finding an optimal trade-off between data sources and method. An operational preference of an NSI, to have to rely on as few

data sources as feasible, is thereby achieved – we do not use irradiance data – with the possible expense of a small loss of accuracy: the parameters are determined on a daily basis, rather than pooling information. What we lose in accuracy here, we gain by analysing several days where the systems perform in different (Summer) conditions, allowing us to obtain an overall reliable picture of the PV system parameters. In Section 2 we propose our methodology for detecting clear days as well as determining PV system parameters using an MCMC algorithm. In Section 3, we present, in some detail, the Python `Emcee` package which was used for the MCMC sampling. The data source `PVOutput` and the `PVLib` models are detailed in Section 4. In Sections 5 and 6, we outline and discuss our results respectively, before summarising and concluding in Section 7.

2. Methodology

Our proposed methodology consists of two parts: in Section 2.1 we delineate our approach to detect clear condition days over the course of 2016–2018. The MCMC algorithm, designed to estimate the PV system parameters (ϕ, θ and N_p) and their uncertainties through comparison of observed and modelled profiles, is laid out in Section 2.2.

2.1. Identifying clear days

2.1.1. Data cleaning

The `PVOutput` metadata and measurement data are corrected and cleaned using the same data cleaning approach as described in appendix A1 of Laevens et al. Five quality criteria are performed per `PVOutput` system, which are all aimed at checking the reliability of the information provided: the first two uniformise/correct the metadata whereas the final three analyse the measurement data to check how realistic these data are, also in relation to the metadata information provided.

The first two criteria have the same effect in mind: check that two different metadata fields, that are largely equivalent to each other, are consistent with one another. The first quality criterion looks at the system sizes in the metadata (P, N_p and P_p) and checks the following equation is satisfied: $P = N_p \cdot P_p$. If this is not the case, then we assume that the total system size (P) will be more reliable. The second criterion compares two different geographical fields: the postal code and the geographical co-ordinates provided in longitude and latitude. It turns out that these are not always consistent and we assume that postal code is the more reliable field as people are less likely to get this wrong.

The third quality check looks at the instantaneous and cumulative yield measurements: if these have been measured correctly, then the cumulative yield, measured by summing the instantaneous measurements, ought to equal the total daily yields. If the calculated daily yield falls within a bandwidth of 90%–110% of the quoted daily yield, the measurement for that day is considered reliable. The next criterion compares the peak instantaneous power with the system size. If these are consistent, then peak power of the system at any given time may not be substantially higher than the quoted system size. We designate reliable measurements as those where the peak power did not exceed the quoted system size by more than 20%. Finally, for our fifth and final check, we exclude PV system measurements on days where large measurement gaps are apparent. For more extensive information on this data cleaning procedure, we refer the reader once again to appendix A1 of Laevens et al.

Data over the course of 2016 and 2017 were used in [19], whereas in this research we have an extra year (2018) available. We use the outcomes of the data cleaning in the aforementioned work to identify systems of which at least 360 days in each year (2016 and 2017) are deemed to be reliable and which also have observed data in 2018. Of approximately 6050[1] `PVOutput` systems registered on the website for 2018, the total number of systems passing these (strict) criteria in the Netherlands is approximately 80.

2.1.2. Clear-sky day selection: Descriptive statistics

The rationale for identifying such highly reliable systems is that in this next step we want to compute statistics which make use of the properties of the profiles and their expected shapes on clear days. By determining the index (n_{max}) of the moment in time at which maximum energy ($E_{n_{max}}$) occurs, given by Eq. (1), it is possible to subdivide the profiles into two separate portions. This is shown in Eqs. (2) and (3). For simplicity we refer to these as the morning and afternoon parts of the profile.[2] On clear days and for systems with a single azimuth, most if not all consecutive points before n_{max} ought to have a positive gradient, whereas all consecutive points after n_{max} should have a negative gradient. Thus, we can compute the gradients between each two successive points of E_{morn} and E_{aft}, given by Eqs. (4) and (5).

$$E_{n_{max}} = \max_{n=1}^{N}\{E_1, \dots, E_n\}, \tag{1}$$

$$E_{morn} = \{E_1, \dots, E_{n_{max}}\}, \tag{2}$$

$$E_{aft} = \{E_{n_{max}}, \dots, E_N\}, \tag{3}$$

$$\Delta E_{morn} = \{E_2 - E_1, \dots, E_{n_{max}} - E_{n_{max}-1}\}, \tag{4}$$

$$\Delta E_{aft} = \{E_{n_{max}+1} - E_{n_{max}}, \dots, E_N - E_{N-1}\}. \tag{5}$$

By calculating the fraction of positive points in the morning and negative points in the afternoon, given by Eqs. (6) and (7), we can determine how clear a day really is. The closer Q_{morn} and Q_{aft} are to one, the clearer the day. Allowing for some margin and corroborated by visual inspection of a handful of systems, we consider a clear day to be one that satisfies the criteria expressed in Eqs. (8)–(10).

$$Q_{morn} = \left(\frac{1}{n_{max}}\right) \cdot \sum_{i=1}^{n_{max}} C_i \qquad \text{where } C_i = \begin{cases} 1 & \text{if } \Delta E_{morn,i} > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

$$Q_{aft} = \left(\frac{1}{N - n_{max}}\right) \cdot \sum_{i=n_{max}+1}^{N} C_i \qquad \text{where } C_i = \begin{cases} 1 & \text{if } \Delta E_{aft,i} < 0 \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

$$Q_{morn} > 0.85, \tag{8}$$

$$Q_{aft} > 0.85, \tag{9}$$

$$0.9 < \frac{Q_{morn}}{Q_{aft}} < 1.1. \tag{10}$$

Two different approaches are possible to construct a list of reliable dates. The first would be to calculate these descriptive statistics for each of the 12 chosen `PVOutput` systems individually (see Section 4.1 for more information regarding selection of these systems). An alternative to this would be to use data from more systems to construct a global clear day list over the region surveyed (the Netherlands). Arguments can be made for either approach. The former allows for more clear days to be identified per system, as it only relies on data per system. At the same time, this approach is more susceptible to including more days that – while satisfying Eqs. (8)–(10) – are not entirely clear. This can occur because the statistics are not precise enough to eliminate days that e.g. had a brief period where clouds appeared. The second approach allows – at the expense of eliminating some days – for data to be pooled across systems, which in general constructs a list of dates with a higher degree of confidence. Because of this, we decided to use the second approach: we pooled information from approx. 80 `PVOutput` systems, as stated in Section 2.1.1.

[1] Please note that in practice, the number of systems that actually record data on the website is around 40%–50% of this figure.

[2] Assigning morning and afternoon labels is, strictly speaking, not always true given different ϕ of PV systems.

2.2. Determining installation parameters

2.2.1. Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a numerical method, often applied in a Bayesian context [24], aimed at determining parameters of a model by sampling from the posterior distribution. *Monte Carlo* refers to the fact that samples are taken in a randomised manner. *Markov Chain*, on the other hand, refers to the process which is followed to draw those samples: a drawn sample, say at iteration k, is only dependent on the sample that was drawn before it at iteration $k-1$ and not $k-2$ etc. These techniques have grown in importance in recent years due to their low(er) computational expense when calculating high dimensional integrals. In official statistics, MCMC methods are increasingly used for the estimation of complex models. This is often necessary to combine data from an ever increasing diversity of data sources and to fulfil the need for more detailed statistics e.g. [25]. Arguably the most famous of the MCMC methods is the Metropolis-Hastings algorithm [26,27].

2.2.2. Bayes' rule

We can introduce our model using Bayes' rule [28]. This theorem states that our prior knowledge of some event occurring may be updated in light of some new evidence into a revised posterior probability of that event. Given some observed data D and model parameters β, the posterior probability, $p(\beta|D)$, which is the probability of a set of model parameters given the data, may be expressed by Eq. (11):

$$p(\beta|D) = p(\beta) \cdot \frac{p(D|\beta)}{p(D)}, \tag{11}$$

where $p(\beta)$ represents our prior knowledge regarding the possible values that β can take on before we were presented with some extra information. The probability of an event occurring, in other words our evidence (our data D) is represented by $p(D)$. This normalisation constant $p(D)$ is of no great concern because, in general, MCMC methods are not dependent on it. This is good news, since this is in many cases not-trivial to compute. The likelihood of observing data D given a set of model parameters β is represented by $p(D|\beta)$. The ratio of these last two terms may be viewed as an index that updates our prior knowledge depending on how informative our observation is. For more information on this and any other Bayesian related information, we refer the reader to Jaynes [29].

2.2.3. Proposed model

We propose a model which fits for the azimuth (ϕ), tilt (θ) and system size (P) of PV systems. In practice, this last parameter is a function of two other variables: the number of panels (N_p) and the individual panel power (P_p). For the purposes of our model, we make two simplifying assumptions. We assume that each panel, for the PV system in question, has the same power P_p. Since the degradation effects over time are unknown, we do not include this in our model. Please see Section 6.4, *2. Explore PV system set-ups*, for further improvements that could be made to our model concerning module degradation. We shall see in Section 4, when the `PVLib` models are presented, that the modelled profiles depend on choices that are made for P_p and N_p. It therefore becomes more useful to think of our model as being a function of N_p, rather than P. Our model parameters are therefore: $\beta = \{\phi, \theta, N_p\}$.

Returning to the notation in Section 2.2.2, we need to introduce the three different terms of Bayes' rule so we can perform our MCMC simulations. The selection of the prior $p(\beta)$ is constrained by the physical interpretation of the β-parameters. We know that $0° \leq \theta \leq 90°$, $0° \leq \phi < 360°$. In theory, the number of panels can be any number, but in practice it is good to constrain this to a sensible number since we are dealing with small PV systems, hence we opt for $1 < N_p \leq 100$. If there is some prior knowledge as to β, then the priors can be adapted accordingly, allowing for the MCMC algorithm to be sped up and converge faster towards a solution. For our problem we consider the priors to be uniform within these limits: all values are equally likely to occur within these prior ranges. As noted in Section 2.2.2, the normalisation $p(D)$ is not necessary, which leaves us with the final element: the likelihood function. Since we adopt uniform priors, the posterior function becomes proportional to the likelihood function. If we assume that our underlying data were generated following a normal distribution, we can represent the likelihood function in terms of a Gaussian function:

$$p(D|\beta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-[obs_i - mod_i(\beta)]^2}{2\sigma^2} \right\}. \tag{12}$$

Here, we are formulating that when the difference between an observed profile and a modelled profile is minimal, the contribution to the (log-) likelihood and (log-) posterior are at their maximum. Through this notation, we have also introduced a fourth parameter, namely σ. While this (nuisance) parameter is of secondary interest, due to the fact that it does not represent something physical with regards to the PV system set-up, it is important to account for σ in the calculations as it can ultimately influence the uncertainties and values we derive for our three other main parameters. This parameter, σ, can be interpreted as the joined effect of the model and measurement uncertainty. Just like the other three parameters, σ is allowed to vary, but is also constrained to a sensible range: $0 < \sigma < 1000$. Please see Table B.3 for an impression of typical values for σ. A common and indeed advisable trick is to compute the natural logarithm of the likelihood, when performing MCMC algorithms. There are a variety of reasons for this, one of which is floating-point precision: when evaluating $p(D|\beta)$ one quickly runs out of decimal places, which is why the $\log\ p(D|\beta)$ is calculated. This leads to Eq. (13)

$$\log\ p(D|\beta) = -N\ \log(\sigma) - \sum_{i=1}^{N} \frac{[obs_i - mod_i(\beta)]^2}{2\sigma^2} + C, \tag{13}$$

where C is a constant. Please note that, since a uniform prior is used on the model parameters, maximising the posterior distribution would be equivalent to maximising the (log-)likelihood, i.e. minimising the mean squared error. Once MCMC sampling has occurred, the parameter values, and their associated uncertainties, may be approximated by histograms of the K–samples, where K refers to the number of (retained) draws from the posterior distribution generated by the MCMC simulation. Hence, Eq. (14) is the Monte Carlo approximation to the posterion mean of $f(\beta)$:

$$\langle f(\beta) \rangle = \int f(\beta) p(\beta|D) d\beta \sim \frac{1}{K} \sum_{j=1}^{K} f(\beta_j). \tag{14}$$

For more information regarding various different MCMC methods, we refer the reader to MacKay [30].

3. `Emcee`

The MCMC algorithm is carried out using the Python library `Emcee`[3] [31], which is an implementation of the Goodman and Weare [32] Affine Invariant Markov Chain Monte Carlo ensemble sampler. This library has been successfully applied, predominantly in the area of astrophysics with large success rates. We opt for this package due to its speed and efficiency, which we expand on below. A lot of information presented in this section originates from the two references cited at the beginning of this paragraph.

3.1. Affine invariance

One of the central aims of the Goodman and Weare MCMC implementation is the optimisation of computational efficiency when sampling from the probability or target distribution. What many other MCMC methods have in common e.g. Metropolis-Hastings is that they

[3] Version 3.0.2.

become less tractable as the complexity of the problem increases. More specifically, the Goodman and Weare implementation speeds up convergence when sampling from e.g. anisotropic distributions, especially when combined with a large dimensionality. The idea here is to simplify the problem by using an affine transformation which converts the target distribution to something which is more easy to sample from. The affine transformation ensures that the performance of the algorithm is the same regardless of the linear transformation applied to the distribution. We refer the reader to Goodman and Weare's motivating example, which demonstrates the necessity for their method and contains further information. More specifically, the original target distribution is shown in Goodman and Weare's Eq. (3) and Fig. 1 while the proposed affine transformation is detailed by their Eq. (4).

The Goodman and Weare MCMC algorithm can be understood as an extension of the traditional Metropolis-Hastings algorithm in the following sense. The Goodman and Weare algorithm makes use of an ensemble of particles or chains, but unlike the independent parallel chains often used in more traditional MCMC set-ups, the update in each chain can depend not only on the previous state of the same chain but also on the current states of the other chains. The proposal distribution therefore depends on other chains and as a consequence is adaptive. The acceptance-rejection steps are the same as in the ordinary Metropolis-Hastings algorithm. The ensemble approach and adaptive nature of the proposal distribution of the Goodman and Weare algorithm allows the target distribution to be explored more efficiently. The pseudo-code (algorithms 1, 2 and 3) in Foreman-Mackey et al. is a useful tool to better understand the differences between Metropolis-Hastings and Goodman and Weare. Within this framework there are different affine invariant flavours. The simplest one is the *Stretch Move*, where the next proposed position of a walker k is dependent on the position of one other walker l. The *Walk Move*, on the other hand, defines proposals for a walker k, which are dependent on the mean position of a subset of walkers, say l, m and n for example. We opt for the latter option as this leads to a quicker convergence of the β-parameters.

3.2. Walker initialisation and number

Initial guess values for the parameters, i.e. $\beta_{i=0,k}$ for the zeroth iteration i and k walkers, need to be assigned in order to commence the MCMC sampling. Since we are dealing with several different walkers, there are a couple of possibilities as to how we initialise them. The first option is to start off the walkers in a "Gaussian ball" around initial parameter values that *a priori* appear most 'reasonable' [31]. This means that all the walkers are initialised at values which only deviate slightly from each other e.g. $\beta_{0,1} = \{190°, 40°, 16\}$, $\beta_{0,2} = \{190.05°, 39.95°, 16\}$ etc. Another possibility is to set off the walkers with very different starting values. To continue the analogy, an example of this could be: $\beta_{0,1} = \{190°, 40°, 16\}$, $\beta_{0,2} = \{110°, 60°, 4\}$ etc. Both options have their limitations. The former can lead to some walkers getting stuck in a local maximum and not capable of escaping, which would lead us to erroneously conclude that the algorithm had converged upon the correct set of preferred β. This is especially true if all the walkers happen to be quite close to each other. The latter can lead to very low acceptance rates[4] due to the fact that the walkers can very often be situated on local likelihood peaks, resulting in steps nearby not being accepted due to likelihood troughs. We experiment with both options and concluded empirically that for this application the first option was most beneficial for convergence.

The recommendation of the `Emcee` developers regarding the number of walkers may be summarised with: "the more, the better", with a number of at least 100 walkers quoted. While the `Emcee` library is very fast, most of the overhead of our algorithm is the continual generation of `PVLib` profiles for any given choice of $\beta_{i,k}$. This means that for each extra walker, it will take longer to compute a number of iterations. We therefore decide to, firstly, use a lower number of walkers and, secondly, allow the number of walkers to vary during the two different periods of the MCMC sampling, i.e. the burn-in phase and the sampling phase. The MCMC is initialised with 40 different walkers, allowing for parameter space to be quickly explored in the burn-in phase. After every 50 iterations, we compute two statistics given by Eqs. (15) and (16). The former computes the mean posterior $\log\ p(\beta|D)_k$ for each walker k, whereas the latter computes the variance of the posterior of all these walkers. These statistics are computed based solely on the preceding 50 iterations.

$$\mathrm{Mean}[\log\ p(\beta|D)_k] = \frac{1}{M}\sum_{i=0}^{M=50} \log\ p(\beta|D)_{i,k}, \tag{15}$$

$$\mathrm{Var}[\log\ p(\beta|D)_k] = \frac{1}{M}\sum_{i=0}^{M=50} \{\log\ p(\beta|D)_{i,k} - \mathrm{Mean}[\log\ p(\beta|D)_k]\}^2. \tag{16}$$

If $\mathrm{Var}[\log\ p(\beta|D)_k] > 5$, then a k-means clustering algorithm is applied, which clusters the walkers into one or two different groups. If the clustering algorithm finds two different groups, the walkers in the group with higher $\log\ p(\beta|D)_k$ are retained, while dropping the second group. If one group is found, all walkers are retained. We now motivate why we do this: even though we apply the *Walk Move* and set off the MCMC in a "Gaussian ball", we notice that occasionally a few of the walkers do get stuck in less ideal parts of parameter space and take a long time leaving those parts. This clustering approach, inspired by Hou et al. [34], allows for this problem to be solved. Once $\mathrm{Var}[\log\ p(\beta|D)_k] \leq 5$ for 5 consecutive calculations (250 iterations), we consider the burn-in phase complete. After the burn-in phase, the number of walkers may be heavily reduced to the 8 walkers with the highest $\log\ p(\beta|D)_k$. We can do this as we have now found the optimal part of parameter space. Fig. 1 shows the first 600 iterations for system 10 on 07/05/2016. These 600 iterations correspond to the complete burn-in phase. As shown in the Figure, the first stage (iterations 1–350) show that $\mathrm{Var}[\log\ p(\beta|D)_k] > 5$ and therefore k-means clustering is taking place. This is why some of the 40 walkers disappear from the Figure e.g. the chain showing $\theta_{pred} \sim 85°$ at iteration 300. The last 250 iterations of the burn-in phase show $\mathrm{Var}[\log\ p(\beta|D)_k] \leq 5$ and that the remaining walkers have found a stable part of parameter space. At iteration 600 the 8 walkers with highest $\log\ p(\beta|D)_k$ are retained and parameter estimation may start.

3.3. Uncertainty estimates and the auto-correlation time

A good measure of the convergence of the MCMC algorithm is the integrated auto-correlation time. This metric quantifies how many iterations (or number of calculations of the posterior function) are required such that independent samples may be drawn which correctly represent the target density in question. The equation for the integrated auto-correlation function requires some derivation and lengthy explanations which distract from the main research topic of this paper. This is why, for this particular quantity, we prefer to refer the reader to eq. (11–13) in Foreman-Mackey et al. as well as the *Autocorrelation analysis & convergence* section in the tutorial of the `Emcee` documentation.[5]

In Section 2.2.2, we formulated Eq. (14), which was the Monte Carlo approximation to the posterior mean $f(\beta)$. This equation allowed us to

[4] The acceptance rate refers to the proportion of times that the newly proposed parameter values in the next step of the Markov Chain are accepted or rejected. Usually an acceptance rate between 0.25 and 0.5 is quoted as being optimal [33], since this is a good balance between efficiency of the chain (accepting proposal steps), while at the same time not accepting too many such that other parts of parameter space may be explored.

[5] https://emcee.readthedocs.io/en/stable/.

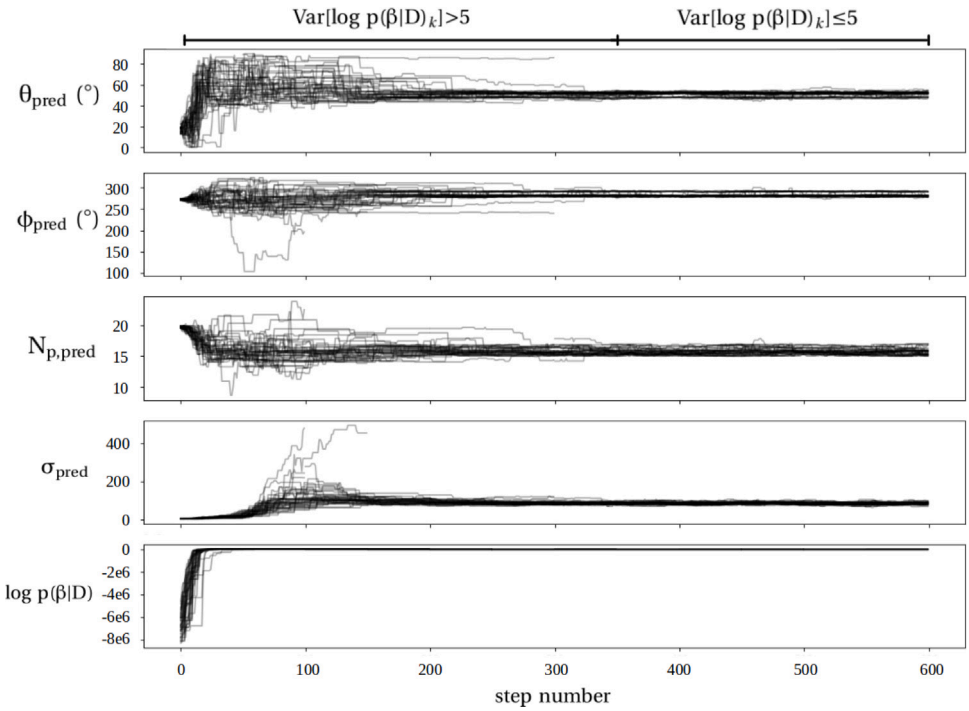

Fig. 1. The complete burn-in phase for system 10 in Table B.3 on 07/05/2018. The two different stages of burn-in, corresponding to the evaluation of the variance criterion (see Eq. (16)), are indicated at the top of the Figure. Please also consult paragraph 3 of Section 3.2 for more information.

estimate our model parameters. From the expected value in Eq. (14), it follows that the MCMC simulation variance of the posterior mean of the function of parameters, $f(\beta)$, can be expressed as σ_β^2 in Eq. (17):

$$\sigma_\beta^2 = \frac{\mathrm{Var}[f(\beta)]}{K}. \tag{17}$$

This equation is an important diagnostic tool because, ideally, σ_β^2 should be small compared to $\mathrm{Var}[f(\beta)]$ i.e. the posterior variance of $f(\beta)$, as larger samples are generated. This relationship, however, holds for samples that were drawn independently from each other. But as presented earlier, we know this not to be true since chains of walkers are interdependent on each other. Therefore σ_β^2 in Eq. (17) can be modified slightly such that it is expressed in function of the integrated autocorrelation time, given by Eq. (18)

$$\sigma_\beta^2 = \frac{\tau}{K}\,\mathrm{Var}[f(\beta)]. \tag{18}$$

In other words, if one has a way of estimating the autocorrelation time, an indication is available as to how many samples need to be generated, hence rendering the fraction $\frac{\tau}{K}$ small and thus also σ_β^2. For further general reading about this, Foreman-Mackey et al. refer the reader to Sokal [35]. Based on the aforementioned sources, we decide to calculate two metrics every 50 iterations of our MCMC simulations. Both of these involve the auto-correlation time and must be satisfied for the algorithm to have converged:

1 The length of the chain must be, at least, $50 \cdot \tau$
2 The auto-correlation time calculated at e.g iteration $i = 200$ cannot deviate by more than 1% compared to iteration $i = 150$.

For further information on the computation of the integrated autocorrelation time or other matters related to `Emcee`, we refer the reader to the online documentation (see footnote 5) of `Emcee` [31].

4. Observed and modelled energy profiles

The successful execution of this MCMC method relies on comparing observed with modelled data. In Section 4.1 we provide more detail regarding the `PVOutput` [20] data, whereas in Section 4.2 we detail the modelled profiles from the Python package `PVLib` [13].

Fig. 2. Locations of the 12 chosen `PVOutput` systems.

4.1. Observed profiles: `PVOutput`

`PVOutput` is an Australian online portal with near real-time information of PV energy generation at various locations throughout the world. In an earlier study [19], the authors obtained data for the whole of the Netherlands for 2016 and 2017, which was later supplemented with

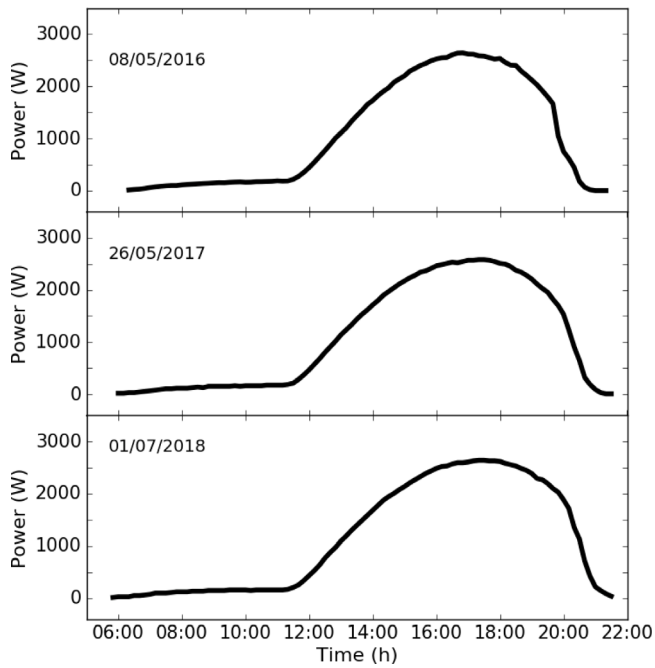

Fig. 3. Profiles of three different clear days for system 10 in Table 2. The metadata of this system is $\phi_{ref} = 315°$, $\theta_{ref} = 46°$, $P_{ref} = 3060$ W_p.

data from 2018. The following parameters are available in the metadata (1–4) and measurement data (5–6):

1. Power: number of panels ($N_{p,ref}$), panel power ($P_{p,ref}$), total system size (P_{ref}), inverter size ($P_{i,ref}$) and number of inverters ($N_{i,ref}$).
2. Geometry: azimuth (ϕ_{ref}) and tilt (θ_{ref}).
3. Panel and inverter brands.
4. Location: postal code 4 (pc4) area, longitude (l_{ref}) and latitude (b_{ref}).
5. Energy: instantaneous (E_{inst}) and cumulative energy measurements (E_{cum})
6. Date and time.

Note that we have introduced the notation of ref, referring to the metadata values, allowing us to later contrast with β_{pred}, which are the predicted values.

In Section 2.1.2 we saw that approx. 80 `PVOutput` systems pass our data cleaning criteria on at least 360/365 days in 2016 and 2017. For our case study, we select a subset of 12 systems for which we perform MCMC sampling. The time resolution of these measurement data is either 5 or 10 min. The 12 systems are chosen such that a good balance is obtained between enough systems present for the effect and accuracy of the MCMC algorithm to be apparent and not too many systems such that the results remain tractable. These 12 systems have also been selected on the basis of their metadata information: a broad variety of systems is included, hence spanning the ranges of our three spatial parameters: $\beta_{ref} = \{\phi_{ref}, \theta_{ref}, N_{p,ref}\}$. This is more easily seen when consulting Fig. 9, which shows these systems occupying various different parts of parameter space. A final point of consideration is that it is desirable that these systems are spread around the country so we also obtain results in different geographical settings. The systems' locations may be seen in Fig. 2, whereas the metadata are displayed in the first six columns of Table 2. We shall see in Section 4.2 that the location of the PV systems will be an input parameter of the modelled profiles. In paragraph 2 of Section 2.1.1, we already motivated our choice for using the postal code 4[6] variable as opposed to using the geographic co-ordinates available in `PVOutput`. We therefore re-assign longitude and latitude co-ordinates based on the centroids of these postal code 4 areas. Examples of three clear-day profiles, as determined in Section 2.1 may be seen for one `PVOutput` system in Fig. 3

4.2. Modelled profiles: `PVLib`

`PVLib`[7] [13] is a Python library, which contains many different routines geared towards the simulation of PV system performance, developed by the Sandia National Laboratories in the United States. Several different functions were used to generate synthetic profiles. The intrinsic properties of the system, such as the surface tilt and azimuth, the module and inverter parameters, the number of modules per string and the number of strings per inverter, are defined with the *PVSystem* function. *Location* on the other hand takes location dependent parameters as an input: l, b, altitude and timezone. These two functions are combined and submitted to the *ModelChain* routine, which generates the synthetic profiles. More specifically this is done by invoking the *get_clearsky* function, using the *Ineichen* clear sky model, where we adopted the default Linke Turbidity settings supplied with `PVLib`. For more information, please consult the `PVLib` documentation. We motivate our choice for the *Ineichen* model by the fact that this model has been found to perform very well compared to other models [36].

Of the six intrinsic properties, we are trying to solve for three of them. Using the notation in Section 2.2, these are $\beta = \{\phi, \theta, N_p\}$, where N_p is what was referred to as the number of modules per string in the previous paragraph (or the number of panels). Assumptions need to be made regarding the other three: a standard situation of one inverter for the whole system set-up is assumed for the number of strings per inverter. As for the module and inverter parameters, we choose *SunPower SPR 230 WHT 2007 E* from the Sandia module database and *AE Solar Energy AE5 0 208V* from the SAPM inverter database. We choose Sun Power as this brand is representative of monocrystalline silicon solar panels. Often, the choice for $P_p = 230$ W does not coincide with the $P_{p,ref}$ as recorded in the `PVOutput` metadata. This is done intentionally, since we would like to remain agnostic as to the power of the panel when performing these calculations. This ought not matter since we can always compare the total predicted and reference system sizes P, using the simple relation: $P_{pred} = N_{p,pred} \cdot P_{p,pred}$. For the location parameters, we use the centroid co-ordinates (l, b) of the postal code 4 area, as discussed in Section 4.1. As the Netherlands is mostly a flat and low-lying country, we assume an altitude of zero metres. This can, of course, be adapted to suit other country requirements, using local GIS maps. Examples of such synthetic profiles may be seen in Fig. 4. We conclude by noting that, at low tilt values (e.g. $\theta < 15°$), the determination of correct azimuth is challenging. This can be explained by the fact that the POA irradiance is very close to GHI, at low tilt angles.

5. Results

5.1. Identifying clear days

The candidate list of dates that we identify as highly reliable clear days for our MCMC method may be seen in Table 1. Note that we only apply eqs. (1–8), as set out in Section 2.1.2, to months where solar noon elevation angles are at their highest, which in the Northern Hemisphere is May, June, July and August.[8] We do this in order to maximise the

[6] In the Netherlands postal codes consist of 4 numbers, followed by two letters. A postal code 4 area does not contain the last two digits.

[7] Version 0.8.0.

[8] For Amsterdam these angles range from a minimal 46° in August to 61° in June, compared to maximum angles of 23° in November and 14° in December.

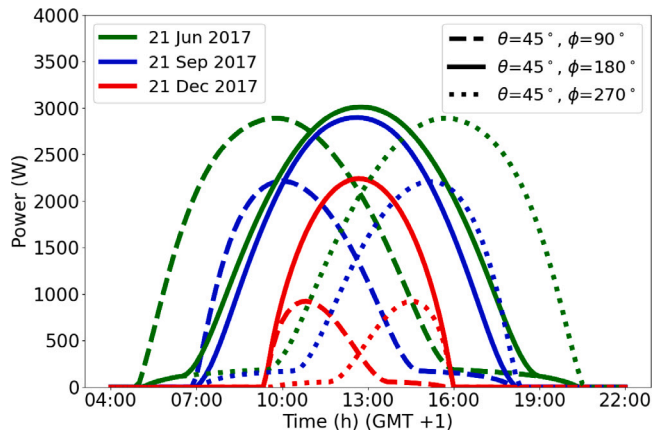

Fig. 4. Clear-sky PVLib profiles in Amsterdam on three different days (solstices and autumnal equinox) and three different PV system configurations ($\theta = 45°$, $\phi = \{90°, 180°, 270°\}$), assuming $P_p = 230$ W and $N_p = 16$.

Table 1
Candidate list of clear days based on number of systems that pass the quality criteria set out in Eqs. (8)–(10).

Date (2016)	n_{2016}	Date (2017)	n_{2017}	Date (2018)	n_{2018}
08/05/2016	54	26/05/2017	54	08/05/2018	61
24/08/2016	50	19/06/2017	42	06/05/2018	60
17/08/2016	47	25/05/2017	32	07/05/2018	55
19/07/2016	45	10/05/2017	28	05/05/2018	54
25/08/2016	43	14/06/2017	22	02/07/2018	53
		27/05/2017	21	01/07/2018	52
				04/05/2018	52
				29/06/2018	49

efficiency of our MCMC method. In these months it is less likely that PV profiles are affected by various objects that cast shade over the panels or block the sunlight altogether. However, in those months, high ambient temperatures in combination with high irradiance will lead to high panel temperatures, which reduces PV efficiency, thus lowering peak power at noon (typically). The results in Table 1 are in concordance with the weather statistics of those years: $H_{2016} = 1040$ kWh m^{-2} [37] and $H_{2018} = 1137$ kWh m^{-2} [38] were significantly higher than the thirty-year average of $H_{30yr} = 986$ kWh m^{-2}, which might naïvely imply a higher number of clear-sky days for 2016 and 2018. By contrast, $H_{2017} = 1020$ kWh m^{-2}, indicating 2017 was fairly average [39]. Since the number of clear days in 2017 was unexpectedly low and we wish to have a reasonable spread in number of days over the three years, we allow for a somewhat higher proportion of less good days in 2017 such that the number of reliable days in 2016, 2017 and 2018 are roughly the same number.

We check these results with another method for determining clear-sky days. A clearness index may be defined such that $k_{t,hr} = \text{GHI}/\text{GHI}_{\text{ext}}$, where GHI is the measured global horizontal irradiance on the Earth's surface at, for example, a weather station and GHI_{ext} is the extra-terrestrial GHI i.e. the irradiance just before it enters the atmosphere, corrected for the solar elevation angle [40]. Using hourly irradiance data from the KNMI (Royal Netherlands Meteorological Institute) weather station in De Bilt[9] [21] and combining this with the PVLib *Irradiance Clearness Index* function, clearness indices are computed for April–August for 2016, 2017 and 2018. From a graphical point of view k_t should reach high values and look smooth to be considered a clear day. When we filter all our days, in 2016–2018, in terms of $k_{t,day} > 0.7$, we recuperate all the clear days from Table 1. Nevertheless some other days are included when performing this cut. An impression of hourly and daily clearness indices may be seen for the first fifteen days of May 2018 in Fig. 5. Here, it can be seen that in addition to 4–8 May, 2 and 3 May also pass the criterion (though are not present in Table 1). However, when visually inspecting these profiles, it is obvious that 4–8 May are entirely clear, whereas 2 and 3 May are not so. This is also the case for other months and years. Though 14 and 27 May 2017 and 10 June 2017 pass the $k_{t,day} > 0.7$ criterion, based on visual inspection of the profiles, we find that we may not necessarily have included these as clear days. This agrees with Table 1, which also shows lower numbers of PVOutput systems experiencing clear days on those dates, hinting at good weather being more localised on those days. Similarly, we also find that many more days could have been selected in 2018 e.g. 30/07/2018 or 03/08/2018. Our method also found these days to be very favourable in terms of clearness, but we decided to truncate the number of days for 2018, to have a reasonable spread of clear days over the three years in question, as discussed in the first paragraph of this section.

5.2. Installation parameters

A comprehensive overview of all MCMC simulations for parameters $\beta_{pred} = \{\phi_{pred}, \theta_{pred}, N_{p,pred}\}$ and nuisance parameter (σ_{pred}) of the 12 chosen PVOutput systems may be consulted in Table B.3. Alongside these four parameters, we present the computed total system size P_{pred} through scaling $N_{p,pred}$ with the chosen PVLib panel power of $P_p = 230$ W, as discussed in Section 4.2. We use the posterior median values as point estimates and compute the uncertainties of β_{pred} by taking the 16th and 84th percentile of the MCMC samples respectively. This is equivalent to $\pm 1\sigma$ of a Gaussian distribution. We choose this interval because this allows for easier interpretation in cases where the posterior distribution does not look like a Gaussian function: the same probability mass is contained within these intervals and these uncertainties are easier to interpret than quoting $\pm 1\sigma$. Since the uncertainties for $N_{p,pred}$ are always zero, we do not quote any uncertainty measures for P_{pred}. The reason for these zero uncertainties is that N_p is discrete: any uncertainty is 'hidden' in P_p and θ, until a deviation becomes so large that a different N_p becomes more likely. In that case N_p 'jumps' but the uncertainty is once again 'hidden' in P_p and θ. A graphical example of the distribution of the posteriors can be seen in Fig. 6 (once again for system 10 on 07/05/2018). A strong negative correlation between θ_{pred} and ϕ_{pred} may be seen in the first panel on the second row of Fig. 6. Because the system is westward facing, the profile is characterised by a long, gradual increase in energy, followed by a sharper decline in energy after maximum energy. Please consult Fig. 7 to see the profile shape of this system. Hence, there are more observed profile points on the ascending side than on the descending side. Any combinations of ϕ_{pred} and θ_{pred} will be favoured that minimise the difference between the observed and modelled profiles overall. This means that modelled profiles cannot deviate too much from the shape of the observed profile on the ascending side given the proportion of total ascending points relative to the total observed points. This means that combinations of azimuth and tilt that are negatively correlated are favoured. The opposite effect (positive correlation) is seen for Eastward facing systems whereas little correlation is seen for e.g. Southward facing systems. Using the favoured values from Fig. 6, a synthetic PVLib profile may be generated and overlayed on top of the observed PVOutput profile to check that these simulations did indeed work. This can be seen in Fig. 7 for the same system and day.

Besides the parameter results, Table B.3 contains three more variables. The first two are related to MCMC metrics, which readers may find insightful. The N_{iter} column refers to the number of iterations

[9] This weather station is considered to be representative of the weather as a whole in the Netherlands.

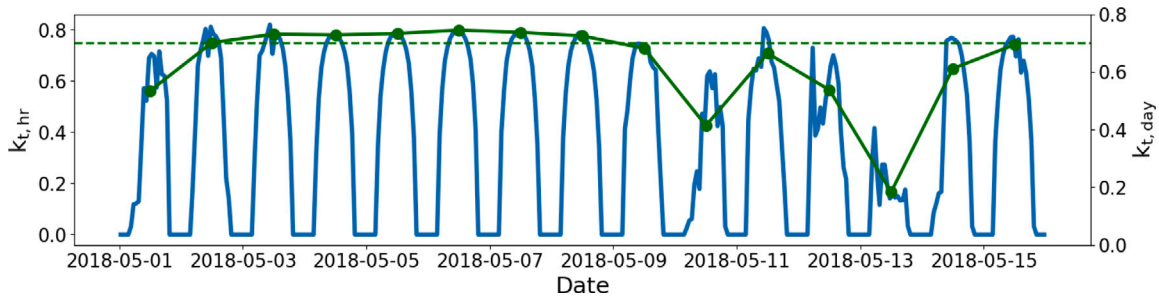

Fig. 5. The hourly (blue) and daily (green) clearness indices for the first 15 days of May 2018, based on irradiance measurements at De Bilt in The Netherlands. The green dashed line shows the $k_{t,day} > 0.7$ line which is used to filter out clear days. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

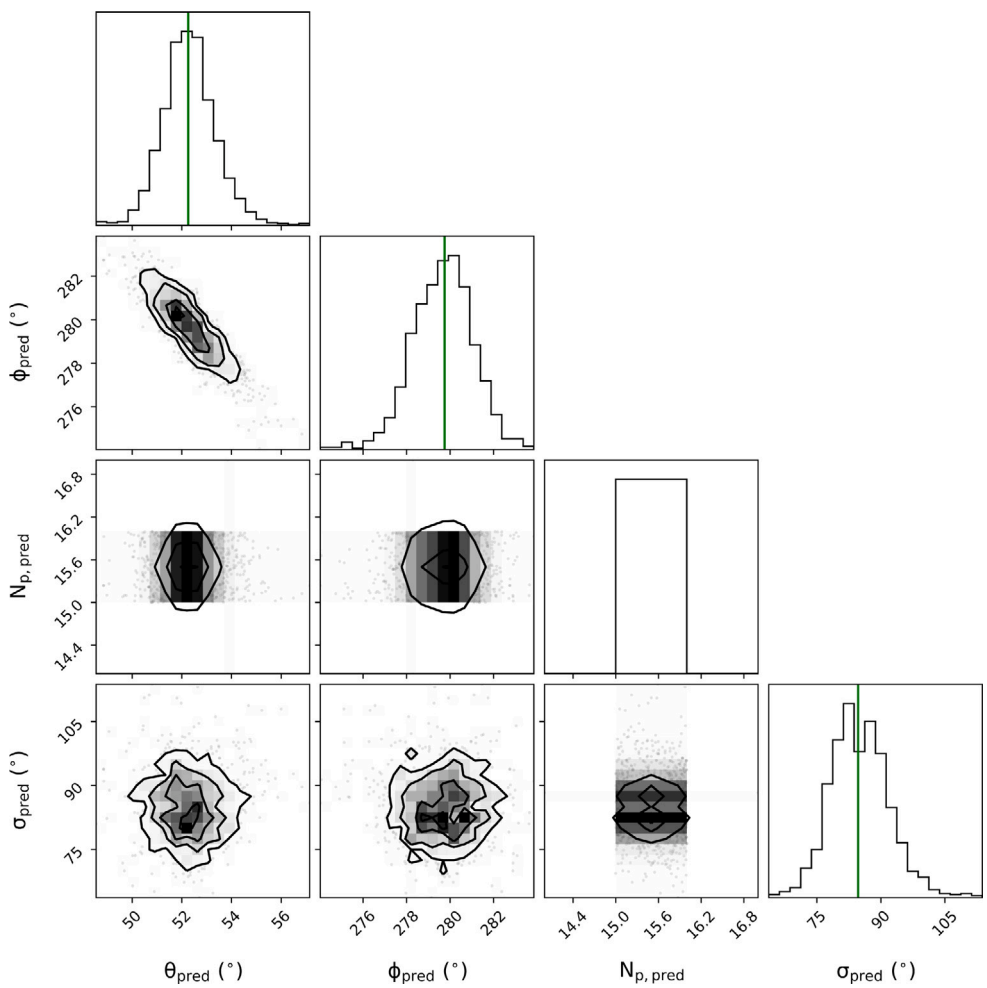

Fig. 6. The posterior pdfs, as set out in Eq. (14), for ϕ_{pred}, θ_{pred}, $N_{p,pred}$ and σ_{pred} (nuisance parameter) for system 10 on 07/05/2018.

necessary for the MCMC simulations to converge. Conditions for convergence were discussed in Section 3.3. In most cases, less than 3000 iterations are necessary for convergence, barring a few exceptions. The acceptance rates are also quoted and lie in the region of 0.37–0.42 (once again with a couple of exceptions). This is closer to the upper end of the acceptable 0.25–0.5 acceptance rate, as was quoted earlier in Section 3.2. The final column of Table B.3 contains the outcome of performing a simple sigma-clipping (σ_c) procedure per system, where 0 means the result is an outlier and 1 means the system survived the sigma clipping. Sigma clipping is a method that may be applied to identify outliers within a distribution of results. This process is iterative: data points are continually rejected that lie a certain number of σ_c away from the mean of the distribution until no more outliers can be removed. After experimenting with $2, 3$ and $4\sigma_c$, we adopt $2\sigma_c$, which agrees intuitively with the results which appear to be outliers in Table B.3. The number of iterations necessary to remove all outliers is system-dependent; however, with the exception of system 11 (8 iterations), 5 or fewer iterations were sufficient to remove outliers. In Section 6, we will use the sigma-clipped results when discussing our results and comparing them with the literature.

We would like to draw the reader's attention to the fact that the results in Table B.3 are days that passed two different criteria: $\sigma_{pred} < 100$ and $0.25 <$ acceptance < 0.5. The former is applied to avoid including MCMC results for unclear days: we saw in Section 2.1 that a list of clear days is constructed, but that this did not mean that this was universally the case for each day at a given location. The cut-off point

Table 2
Metadata of chosen PVOutput systems, along with the inferred results using (i) MCMC methods, (ii) Meng et al. method applied to 2017 and (iii) Meng et al. method applied to Summer months of 2017.

	PVOutput metadata					MCMC inferred results (May–Aug 2016, 2017, 2018)				Meng et al. [10]			
										Meng 1		Meng 2	
id	ϕ_{ref} (°)	θ_{ref} (°)	$N_{p,ref}$	$P_{p,ref}$ (W)	P_{ref} (W)	$\overline{\phi}_{pred}$ (°)	$\overline{\theta}_{pred}$ (°)	$\overline{N}_{p,pred}$ (W)	$\overline{P}_{pred}$ (W)	ϕ (°)	θ (°)	ϕ (°)	θ (°)
1	135	50	15	235	3525	140	56	15	3450	147	68	147	67
2	225	37	12	250	3520	220	35	14	3220	218	32	222	33
3	135	15	15	250	3750	163	16	16	3680	139	12	143	11
4	315	15	9	195	1755	268	16	6	1380	239	44	237	22
5	135	45	18	250	4500	138	44	20	4600	136	43	131	39
6	180	45	16	245	3920	178	43	17	3910	173	50	178	48
7	180	40	22	250	5500	183	28	23	5290	178	38	178	30
8	180	0	18	250	4500	191	39	18	4140	192	39	192	39
9	315	13	4	275	1100	340	20	5	1150	287	8	312	11
10	315	46	18	170	3060	292	50	16	3680	278	52	259	65
11	45	13	6	170	2040	351	8	10	2300	9	2	170	0
12	0	7	10	250	2500	14	15	12	2760	46	9	38	9

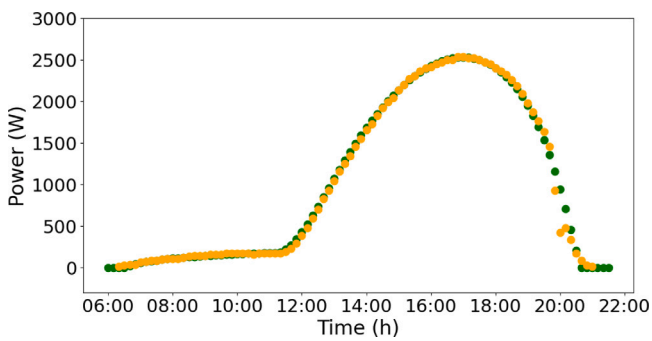

Fig. 7. The PVLib profile (green) compared with the observed profile (orange) for system 10 as a result of performing an MCMC simulation for 07/05/2018. The PVLib profile is constructed using the favoured parameters, as quoted in Table B.3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of 100 was determined by visually inspecting the observed profiles and comparing them to σ_{pred}. For profiles that were not completely clear, e.g. some cloudy moments are present, the value for σ_{pred} quickly became very large. We note that all MCMC samples with acceptance rates lower than 0.25, produce bimodal[10] results, which is why we adopt the second criterion.

To make our results more insightful to the reader, we have summarised the results from Table B.3 into columns 7–10 of Table 2 and Fig. 8. The former displays mean values of all posterior medians that passed the σ_c procedure. Please note that the mean system size was computed according to $\overline{P}_{pred} = \overline{N}_{p,pred} \cdot P_p$, where $P_p = 230$ W, as discussed in Section 4.2. The latter – inspired by similar, yet different figures in Meng et al. – is a graphical representation of all sigma-clipped results of θ_{pred} and ϕ_{pred}. Per system, we have taken the 2-D distributions of θ_{pred} and ϕ_{pred} (an example of which is shown in the first panel on the second row of Fig. 6) and normalised them such that each day has the same weight. This is essential because the number of iterations necessary to converge on every given day fluctuates (see N_{iter} in Table B.3). This means that some distributions contain more chains than others. A final distribution per system is constructed by cumulating the individual (daily) distributions. Finally, this new cumulative distribution is normalised once more such that each bin value sums to one.

6. Discussion

6.1. Comparison with PVOutput

Let us now turn to discussing the results in Table B.3 and start with some general statements that are true for all systems, before making more specific observations related to subgroups of these systems. All systems (except system 4) have in common that ϕ_{pred} is better constrained than θ_{pred}. What we mean by this is that the outcome of the azimuth is relatively insensitive to the 'good' day chosen. This is evidenced by the median ranges[11]: $\Delta\phi_{pred,\frac{1}{2}} = 8°$ and $\Delta\theta_{pred,\frac{1}{2}} = 5°$. Given that the range of ϕ_{pred} (360°) is four times that of θ_{pred} (90°), $\Delta\phi_{pred,\frac{1}{2}}$ is considerably smaller, when taking into account a factor of 4 in difference. An explanation for this could be that correct determination of azimuth is strongly dependent on the daily pattern of PV generation, whereas differences in tilts arise from seasonal patterns. Computing PV parameters on a daily basis therefore probably better constrains ϕ_{pred} compared to θ_{pred} (though please also see our discussion below in point 2). The median range in the number of panels is $\Delta N_{p,pred,\frac{1}{2}} = 2$, corresponding to $\Delta P_{pred,\frac{1}{2}} = 460$ W.

Let us now make some specific observations with regards to subgroups of these 12 systems:

1. *Azimuth*. South facing systems are more accurately constrained than other azimuths. Systems 6, 7 and 8 have spreads of $\Delta\phi_{pred} = 2° - 3°$ (taking into account outlier elimination through sigma clipping). This observation is not necessarily surprising given that these azimuths mean that peak performance of the system coincides with solar noon. For eastern, western and northern facing systems, this moment will be less pronounced.
2. *The relationship between tilt and system size*. Part of the reason a wider spread is observed for θ_{pred} is that $N_{p,pred}$ is a discrete number: the choice for $N_{p,pred}$ in the MCMC sampling will subsequently drive the choice for θ_{pred}. Since each day is different with regards to weather conditions, it is quite conceivable that profiles will differ to such an extent that fluctuations arise in the results for $N_{p,pred}$. This will, in turn, affect θ_{pred}.

[10] Here, we mean that the posterior density functions do not produce one clearly defined peak in parameter space. Rather, two or more peaks are identified as likely outcomes.

[11] This is obtained by calculating per system i: $\Delta\phi_i = \phi_{max,i} - \phi_{min,i}$ and $\Delta\theta_i = \theta_{max,i} - \theta_{min,i}$. A median is taken of these 12 results, producing $\Delta\phi_{pred,\frac{1}{2}}$ and $\Delta\theta_{pred,\frac{1}{2}}$. We compute these for the sigma-clipped results in Table B.3.

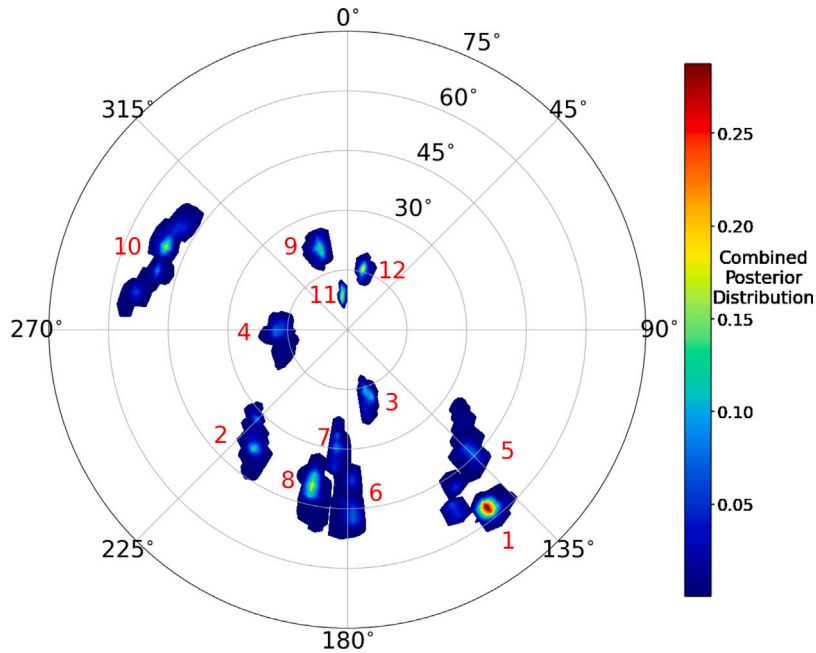

Fig. 8. Combined posterior distributions (ϕ_{pred} and θ_{pred}) for systems 1–12, with θ_{pred} running radially from 0° in the centre to 75° at the outer circle. Please note that the two blue blobs near systems 1 and 5 belong to system 5.

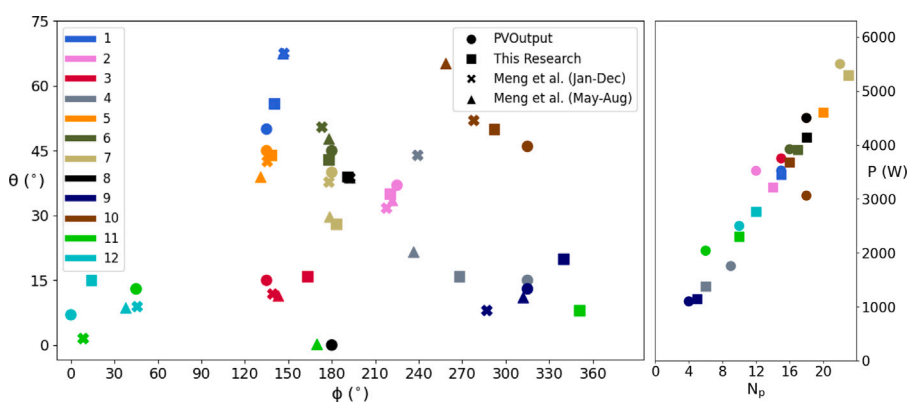

Fig. 9. A comparison between the metadata and three different methods for the variables ϕ and θ (left hand side) and N_p and P (right hand side). Circles represent the metadata whereas the inferred results, using three different methods are denoted by squares (MCMC), crosses (Meng 1) and triangles (Meng 2). Each colour denotes one of the 12 different `PVOutput` systems. Please note that only MCMC results are shown for N_p and P (right hand side) because Meng et al. do not calculate this in their method. Please also note that systems 5 and 8 have identical system sizes, according to the `PVOutput` metadata, which is why it looks like system's 5 metadata values are missing (right hand side). This figure is essentially a visualisation of the results in Table 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. *Accuracy of metadata*. The reason for developing methods to determine parameters of PV systems (as opposed to solely relying on metadata) is most obvious when considering system 8. According to the metadata $\theta_{ref} = 0°$, whereas $\overline{\theta}_{pred} = 39°$. This system looks to be an example of the subpopulation of flat systems that is disproportionately large, as mentioned by Killinger et al.. A slightly less striking offset may be seen for system 7, where all sigma-clipped measurements put the system at $\overline{\theta}_{pred} = 28°$, 12° lower than θ_{ref} (though please also see the benchmarking discussion in Section 6.2). Improvements in ϕ_{pred} can be observed for systems 3, 9 and 11. In the case of the first two, it may be seen that these systems appear to be equidistant between the two nearest values of ϕ.[12] In the case of system 11, an incorrect ϕ_{ref} appears to have been registered. As discussed in Laevens et al., but also in the second paragraph of Section 2.1.1, another mistake is sometimes present in the metadata, namely that: $P_{ref} \neq P_{calc}$ where $P_{calc} = N_{p,ref} \cdot P_{p,ref}$. In the method we developed in Laevens et al., we assumed that

[12] We remind the reader that `PVOutput` owners can only select 8 different orientations, corresponding to cardinal and intercardinal directions.

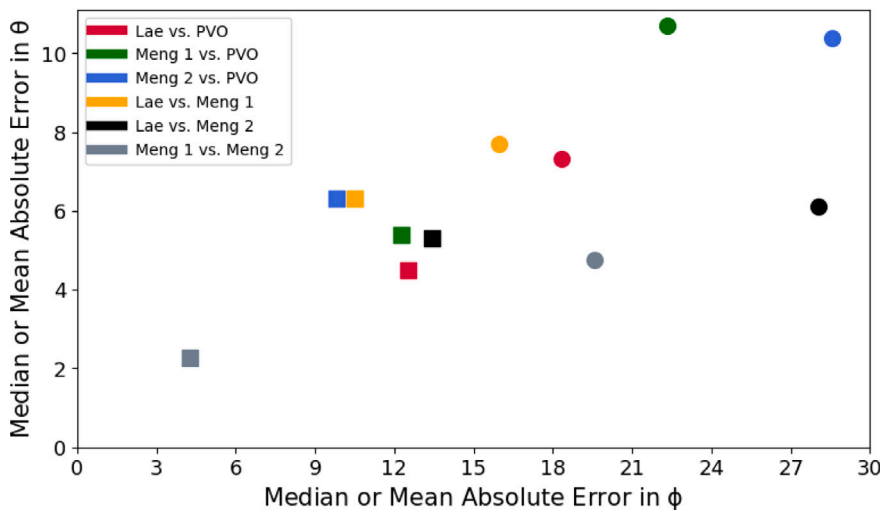

Fig. 10. Median (squares) or mean (circles) absolute errors for ϕ and θ. The colours denote the comparisons between the different methods (or metadata). Our research, Meng 1 and Meng 2 are compared with the `PVOutput` metadata in red, green and blue respectively. A comparison between our method and Meng 1 is shown in orange, whereas black denotes a comparison between our method and Meng 2. Finally, the performance of Meng 1 and Meng 2 are compared in grey. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

`PVOutput` owners would be more likely to get P_{ref} right rather than $N_{p,ref}$ and $P_{p,ref}$. It was therefore of some interest to us to include two systems (2 and 11) in which the aforementioned equation is not satisfied. In either case it appears that neither P_{ref} nor $N_{p,ref}$ and $P_{p,ref}$ have been imputed correctly in the metadata. For system 2 it appears that $P_{calc} < P_{pred} < P_{ref}$, whereas for system 11, $P_{pred} > P_{ref} > P_{calc}$. Systems 4, 8 and 10 have system sizes that are $P_{pred} > P_{ref}$. A possible explanation for this is that extra panels were purchased after signing up to `PVOutput`.

6.2. Comparison with the literature

To evaluate the performance of our proposed method, we decide to apply Meng et al.'s method to our sample of `PVOutput` systems. There are a number of reasons why we should and would like to do this. Firstly, our method has assumed that the four "Summer" months should be sufficient for an adequate determination of the system parameters. Consequently, a relatively low number of days is used per year to achieve this. Secondly, applying another method may shed further light on whether some of the recorded `PVOutput` metadata is (in)correct. Finally, we would like to compare our method's efficiency with another method to find out whether our claim about computational efficiency, in the Introduction, is justified. We have chosen to benchmark with Meng et al. for a number of substantive and practical reasons: (i) The methodology in both studies relies on just two data sources. In addition to this, both studies share similarities in that they evaluate fits between observed and modelled profiles, and they determine a most likely profile given the observed data. In the case of Meng et al., parameter values are estimated by computing RMSE's (curve matching) for all possible combinations of azimuth and tilt, without any prior assumptions. Mean point estimates are obtained from the top 1% of best fits. Our research, by contrast, constructs the complete posterior distribution of the model parameters, which allow the parameters and uncertainties to be estimated. A flat prior is used in this process. By constructing Markov Chains with Monte Carlo methods, the next set of parameter values in the chain is informed by the previous set (in classical MCMC). In this fashion, the chains meander through parameter space in search of regions where the posterior distribution is at maximum which is equivalent to the mean squared error being at a minimum (for this particular research). (ii) Both studies use real observed data from a dozen or so systems, in the same country (The Netherlands). (iii) Meng et al.'s research concentrates on 2017, which is one of the three years in our dataset. It therefore follows that the benchmarking we do here is dependent on 2017 data only. (iv) The method employed by Meng et al. is open source and may therefore also be easily applied. We refer the reader back to the Introduction and the Meng et al. publication itself for further information regarding their method.

To assess the three assumptions or reasons mentioned at the start of the previous paragraph, we apply Meng et al.'s method in two different ways, which we will denote as Meng 1 and Meng 2 henceforth. The former is the Meng et al. method, as described in their paper, applied to the totality of 2017. Using Meng 1 will provide a good overview as to how this method, when applied integrally, compares with our method applied only to Summer months. The latter (Meng 2) is the same method as Meng 1, but applied only to the months of May–August 2017. This second method allows us once again to compare and contrast with Meng 1 as well as the MCMC method. If a selection effect were present as a result of only picking Summer months, then it may be hypothesised that this will have an effect on Meng et al.'s method. The results of this can be viewed in the final 4 columns of Table 2 and the left hand side panel of Fig. 9. We will discuss the outcome of this benchmarking, in light of the aforementioned three assumptions in the following three subsections.

6.2.1. Summer selection effect

From these comparative analyses, we see little evidence of a selection effect by using the months of May, June, July and August in our method. Besides the information already mentioned in Table 2, this is more obvious when considering Fig. 10, which shows the median (and mean) absolute errors between the results obtained through our MCMC method, Meng 1 and Meng 2. We consider it more useful to compare the median absolute errors (MeAE) from here on as these are more robust to outliers. The MeAE between Meng 1 and Meng 2 is 2.3° and 4.3° for θ and ϕ respectively, indicating very little difference between the two methods. The performance of Meng 1 and Meng 2, compared

to this research are quite similar, with our results being slightly more consistent with Meng 1 than Meng 2, but only by a small margin: the median absolute errors (compared to this research) are $(10.5°, 6.3°)$ and $(13.4°, 5.3°)$ for Meng 1 and Meng 2 respectively. Meng et al. were also interested in ascertaining whether their method could be applied with less than one year's worth of data. They noted that there are indeed seasonal influences present, but that the Summer months and more specifically May, June and July of 2017, show the smallest deviations between the known and computed orientations (see their Section 5.4). They explain that this is likely due to the systems operating closest to standard conditions in these months. Their findings therefore reinforce our decision to only select Summer months.

When considering certain systems in Table 2, it is nonetheless tempting to see the presence of a Summer selection effect. If we zoom in on specific systems, there appears to be a lowering of θ for systems 4 and 7: it is striking that Meng 2 produces θ values that comes quite close to our MCMC inferred result, whereas Meng 1 does not. On closer inspection, it also turns out that these two systems are the most poorly constrained systems by either the Meng 1 or Meng 2 methods. As explained earlier, the Meng et al. method determines the most likely parameter values by finding the maximum overlap between the top 1% RMSE's for each month. The number of overlapping months, according to Meng 1, is 2 and 3 for systems 4 and 7 respectively. According to Meng 2, there are 2 overlapping months for both systems in question. We see that for many of our other systems the Meng 1 method returns a significantly higher number of overlapping months, with 5, 6 or even 7 months being quite common for Meng 1. This would suggest that we may be less confident about the Meng results for systems 4 and 7.

6.2.2. `PVOutput` *Metadata accuracy re-visited*

An interesting aspect of applying another method is that it can confirm or infirm some of the observations that were made in *Accuracy of the metadata* in Section 6.1. Globally speaking, the three methods are quite similar in effectivity with tilt and azimuth MeAE's of $(4.5°, 12.5°)$, $(5.4°, 12.3°)$ and $(6.3°, 9.8°)$ for our MCMC method, Meng 1 and Meng 2. Our study performs slightly better for tilt angles, possibly due to the inclusion of N_p in our model. Azimuths, on the other hand, seem better constrained in the Meng 1 and Meng 2 methods (as can be seen in Fig. 10). It should be noted that the median and mean absolute errors will be inflated for ϕ. This is because it is only possible to record azimuths down to an accuracy of 45° in the `PVOutput` metadata. Therefore there is reason to think that the azimuth results could be even more accurate than they appear.

It is striking that there appears to be less agreement between the different methods when it comes to determining systems' azimuths in the range of $270° - 45°$ e.g. systems 4, 9, 10, 11 and 12. At the same time, it should be noted that in fact four of these systems have low θ, making azimuth increasingly difficult to pinpoint. Please see the last two sentences of Section 4.2 for more information regarding this. A reasonable hypothesis for the discrepancy could be the differing time resolution between our MCMC methods (5 or 10 min) and the Meng et al. method (1 h).

6.2.3. *MCMC method speed*

It is not entirely straightforward to make a direct comparison between the two methods in terms of their speed as they both operate differently and have distinct goals. Taking the methods at face value and running them each, we see that our method performs slower than Meng et al.. It took us around 55 min and 1 CPU to analyse 12 different days' worth of data to obtain parameter estimates for one of our systems, using the Meng 1 method. The MCMC method needs an average of 10 min and 8 CPU's[13] to obtain parameters for one system on one day. Extrapolating this to 1 CPU,[14] this means that around 60–70 min are necessary to obtain just one day's estimate.

One of the main reasons our method is slower is we employ a 4D model, whereas Meng et al. use a 2D model. Our method has been designed expressly to operate fast and efficiently with a higher number of dimensions, while remaining relatively slow with a lower number of dimensions. This is because most of the overhead in our method comes from continually generating the `PVLib` profiles. Whether we do this for 10 or 100 measured points and for 2 or 3 parameters does not change much as a profile needs to be generated anyway: the computation of the likelihood is negligible in comparison to the time taken to generate a profile. An option could be to compute profiles in advance, but we decide against this for a number of reasons. Firstly, we would like this method to be easily applied in other countries as well. While it is possible to construct a family of `PVLib` profiles in advance using a central longitude and latitude of the Netherlands, this is not possible for larger countries and decisions would need to be taken accordingly. A more fundamental reason, however, is the framework of our method which allows for uncertainty estimation of the parameters. This relies on profiles being generated with many significant digits for ϕ_{pred}, θ_{pred} and σ_{pred}. Generating profiles in advance with this precision would be very challenging as there would be too many possible combinations[15] of the 4D parameters. Decisions would have to be made to manage this e.g. azimuths and tilts rounded to 1 degree. Given the uncertainties of these parameters are on the order of a degree or more, this approach will not be able to produce reliable uncertainty estimates.

It is difficult to make a statement on what were to happen to the speed of the Meng et al. model should it include an additional parameter (such as our $N_{p,pred}$ or σ_{pred}), since it was not designed to include an extra parameter. Nevertheless our methods do overlap in the sense that they both rely on invoking `PVLib` to generate a modelled profile. In the case of Meng et al. both observed and modelled profiles are normalised and are assumed to be one on one related to each other. This approach makes a lot of sense as modelled irradiance is being compared with power, whereas we are comparing modelled and observed power. Because of this normalisation, the system size is effectively taken out of the equation, rendering it a 2D problem for Meng et al.. Nevertheless, if a similar prescription (comparison of modelled and observed power) were used to ours, but without an MCMC approach, it is very clear that any method that cycles through all possible outcomes in 2D will be slowed down by a factor P in 3D, where P is the number of values that need to be cycled through in the extra third dimension. For a 4D problem this will be $P \cdot Q$, with Q the number of possibilities in the fourth dimension. Given that, in our set-up, N_p is allowed to be any value between 1 and 100, this means that any algorithm that intends to find the system size too would be slowed down by a factor 100. This is even without considering σ, though this is of course not a PV system parameter and therefore of secondary interest. In such a scenario, integer values for all these variables would need to be used to find the parameter estimates. The MCMC is not constrained by this as it can take on any floating point value. It would therefore seem that our method is, in fact, already rather fast, given its dimensionality and time resolution.

Finally, it should be noted that the convergence criteria we imposed when implementing the MCMC method are designed for statistical robustness. We saw in Section 3.3 that convergence is an important factor in the overall functioning of the method and that great lengths are taken to halt the MCMC method only after convergence is reached.

[13] As mentioned earlier, we use different walkers for the MCMC method and therefore run the code in parallel.

[14] We assume a factor of 6 or 7 speed up for 8 CPU's and not a factor of 8. This comes from the fact that running code in parallel with x CPU's does not give you an exact speed up of x.

[15] Please also see the end of the next paragraph for a discussion about parameter combinations, in a different context.

Furthermore, a main objective of MCMC is to approximate the posterior distribution, and not just to obtain a point estimate. In other words, we obtain a sense of uncertainty about the parameter estimates, under the assumption that the model is correct. We think it is insightful to point out that the overall time taken for the algorithm to reach the most favoured part of parameter space happens in the space of around 1 min (using the aforementioned set-up). An overwhelmingly large fraction of the MCMC simulation time ($\sim$ 90%) is spent meandering around the already correctly determined part of parameter space until convergence is obtained, by satisfying the 50τ criterion, as mentioned in Section 3.3. If just an initial impression were necessary as to where the system lies in parameter space, then this should already be obvious very quickly.

6.3. Practical application perspective

We now return to the two reasons why we undertook this research, as mentioned in the Introduction: (i) test an MCMC method to see whether this can be used for assessing metadata reliability at our NSI and (ii) test assumptions that were made in Laevens et al.: the accuracy of ϕ, θ and P in the `PVOutput` metadata drive some of the conclusions of this previous publication. From this small case study we can conclude that we have successfully demonstrated that an MCMC method may be used to verify metadata information. We hope that this method can be used in other settings at our NSI. Secondly, we also see that most calculated parameters appear to be in reasonable agreement with the `PVOutput` metadata. Nevertheless, some specific examples exist that highlight the necessity for developing these new methods. This was discussed in Sections 6.1 and 6.2.2. A sample size of 12 is large enough to demonstrate the concept and be reasonably confident in the results, also of those reported in Laevens et al.. Notwithstanding, it might be worthwhile, for large-scale production of national statistics, to expand to a much larger set of systems. This would further minimise the likelihood of having missed influential system types that are very dissimilar to these. This is why we explore the practical application perspective of this method in the remainder of this section.

Although quite some effort was undertaken to quantify the efficiency of our method, through our comparison with Meng et al., we identify the following 4 challenges on the route to implementation of this method more generally:

1. *Repeat method on systems with known ground truth*: Even more confidence in our MCMC method may be obtained by repeating the method on a larger sample of systems for which we know that the metadata is completely accurate and reliable.
2. *Test method in other countries:* Another good validation test is to apply the MCMC method in other countries to see how well it performs in a different context. While we do not expect there to be a difference in performance, it is important to find out whether there are situations we have not yet conceived of that could adversely affect this method.
3. *Further exploration of modelled profiles*: We saw in Section 4.2 that the generation of `PVLib` profiles relies on the selection of a clear-sky model, which we chose to be *Ineichen*. Nevertheless, other choices are possible, such as *Simplified Solis* or *Haurwitz*, which can affect the amplitudes of the modelled profiles and, therefore, the favoured parameters which are returned by the MCMC algorithm. Within these different models, it is possible to experiment with elements such as the *Linke Turbidity* factor, which defines the atmosphere's optical thickness: water vapour and aerosols can cause absorption and scattering in the atmosphere. In *Explore PV system set-ups* in Section 6.4, we detail further refinements that could be made to define the system, in terms of its set-up and location, which also affect the modelled profiles.
4. *Explore alternative methods for clear-day assessment*: Besides the approach described in Sections 2.1.1 and 2.1.2 for identifying clear days, other approaches could be envisaged. Firstly, the list of clear days could be re-computed for our existing approach, but on the basis of hourly data, rather than 5 or 10 min resolution data. This would make profiles less susceptible to short fluctuations, which can cause certain days to be discarded. Secondly, the alternative approach, detailed in Section 2.1.2 could be implemented instead: descriptive statistics could be calculated for systems individually, rather than pooling information, allowing a heterogeneous lists of good days to be compiled that is system-dependent. This would benefit systems 1, 5, 7 and 10 which had a low number of results in 2016 and/or 2017. The reliability of such a list could be further increased, by combining it with information from another approach such as the calculation of clearness indices, as described in Section 5.1. This would have the drawback of adding more overhead to our method though. Finally, other methods from the literature could be consulted to determine clear days.
   Despite a low number of clear days per year, we notice that the estimates of the system parameters do not fluctuate greatly between different years. This leads us to suggest that a good indication of PV system parameters may already be obtained after MCMC simulations are performed for a handful of clear-sky days in the months of May, June, July or August of any one given year. A possible explanation for the lack of variability between years could be our strict criteria for identifying clear-sky days, which result in a low, yet very reliable, number of clear-sky days over the different years. We therefore suggest that the implementation of this method does not require three years' worth of data before metadata may be inferred. The procedure, such as that described in Section 2.1.1 (last paragraph), could therefore be performed for individual years. This makes the threshold for this method's implementation a lot lower and allows it to be easily applied in any year, also after 2018. As touched upon in the previous paragraph, four systems could have benefited from a higher number of results per year. This is why further research should be conducted to see how much the estimation of PV system parameters varies, when expanding the list of clear-sky days to include system-dependent clear-sky days. Finally, given that this method was tested in a country which is not known for its plentiful sunshine, we think this method may be easily implemented elsewhere.

6.4. Future work

There is a variety of other things that could be investigated but which we think are of lesser importance for the route to implementation. We identify the following investigatory points:

1. *Adding or constraining model parameters:* It would be interesting to investigate whether further dimensionality could be added to the model such that it operates without even knowing where any of the systems lie. This could be achieved by adding l_{pred} and b_{pred} to the model. Simple techniques using the Equations of Time, as described in Haghdadi et al. [41] could be used to pinpoint l_{pred}. This could subsequently be a fixed parameter in the MCMC model, but with b_{pred} allowed to vary, rendering this a 5 or 6D model. In the 4D case, it could be worthwhile investigating what happens to the 3D case when ϕ_{pred} is fixed. A reason for fixing this variable is that we have seen that it is more easily pinpointed than θ_{pred}. Therefore determining ϕ_{pred} only for a few clear days and subsequently fixing it for other days would allow us to see how much this influences the values of θ_{pred} and $N_{p,pred}$ (and σ_{pred}). Finally, results from a handful of ultra reliable days, could be used as input for a constrained

prior, which could be adopted for slightly less reliable days. We remind the reader that the results in Table B.3 were obtained using the following uniform priors: $0° \leq \phi < 360°$, $0° \leq \theta \leq 90°$ and $1 < N_p \leq 100$.

2. *Explore PV system set-ups:* When defining PVLib profiles, it is assumed that $P_i = P$, which need not be the case. Sometimes a smaller inverter is preferred because this allows PV systems to perform more optimally in low-light conditions such as on cloudy days or mornings and evenings. This has the disadvantage of clipping profiles around midday. These subtleties could be incorporated when choosing P_i for the corresponding inverter model in PVLib. Secondly, different choices for P_p could be made, when generating PVLib profiles, adding further precision to θ_{pred}, $N_{p,pred}$ and P_{pred}. Thirdly, a combination of different azimuths could be accounted for in the PV system set up. This could be an area where the MCMC method could excel greatly for reasons already mentioned in Section 6.2.3: once more permutations are possible, the advantages of MCMC methods become even more apparent. PV system degradation over time, which is typically 0.5% per year [42], could also be accounted for in the model if the installation date of a PV system is known. Finally, the robustness of our method could be further tested by incorporating the effects of shading into the modelled profiles.
3. *Lower resolution data:* One could experiment with reducing the 5 or 10 min intervals in the 12 chosen PVOutput profiles to e.g. 15, 30 min or even 1 h intervals. In some cases, this would make the profiles also less sensitive to small fluctuations between measurements on a 5 or 10 min interval. Finally, coarser data would also make the fitting less susceptible to problems when fitting the beginnings and endings of Summer days, where, after the Sun has set, there is still a time period when DHI plays an active role and causes a bump in the energy profile.
4. *Include temperature effect for clear days:* Temperature data of the nearest weather station for each PV system could be researched and compared to the results, to see whether a temperature effect (e.g. Reich et al. [43]) may be observed in the results. If this is the case, then days with moderate temperatures could be retained and hotter days eliminated.

7. Summary and conclusion

We have presented a method to determine the azimuth, tilt and system size of photovoltaic systems. This method was applied to 12 different PVOutput systems. We first identified a list of 19 highly reliable days over the course of 2016–2018 using descriptive statistics of the profiles from a larger set of approximately 80 PVOutput systems. For each of the 12 chosen systems, and on every reliable day, we run a Markov Chain Monte Carlo algorithm, implemented with the Emcee Python package, which compares synthetic PVLib profiles with observed PVOutput profiles. Hence, the best fit is found, and the corresponding parameters that best fit the observations. The results are broadly in line with the PVOutput metadata. To obtain even more confidence in our new method, we benchmark it with another method in the literature: Meng et al. We do this to test several assumptions we made, such as relying solely on Summer days to determine the PV system parameters. We find that our results are in good agreement with the aforementioned method. Nevertheless the results also demonstrate the necessity for such techniques: sometimes the metadata is inaccurate e.g. the azimuth is not very precise or incorrect. The same can be said for the tilt and system size. We identify further challenges that are necessary to address on the route to implementation. Further testing on systems for which the ground truth is known is desirable. Finally, we identified a variety of other "nice to have" refinements, such as the addition of extra parameters to the model and experimentation with less granular data, amongst others.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

B.P.M. Laevens would like to thank Nicolas Martin and Yvonne Gootzen for discussions regarding the application of MCMC methods. W.G.J.H.M. van Sark acknowledges financial support from the Ministry of Economic Affairs and Climate Policy, The Netherlands through Topsector Energy funding for the TKI-Urban Energy project PV Observatory.

Appendix A. Source code

This section contains additional information regarding the implementation of the Markov Chain Monte Carlo algorithm to determine azimuth, tilt and system size of PV systems. Please access the Python source code at the following: https://github.com/blaevens/PV_mcmc. The repository contains a Jupyter-notebook which gives users the option to run the MCMC code on synthetic or real data. For the former, artificial data is first generated as part of the routine. The underlying functions used in the Jupyter-Notebook are subdivided into five different scripts. These are:

- synthetic_data_4D.py: various functions allowing artificial data to be generated based on a metadata file containing $\beta = \{\phi, \theta, N_p\}$ specifications for three systems, with the option of adding noise to the profiles.
- data_processing_4D.py: functions that perform different data preprocessing steps to enable the MCMC algorithm to run smoothly.
- mcmc_functions_4D.py: these functions are central to the method. This file contains important functions such as the likelihood, prior as well as the MCMC burn in and sampling functions.
- clustering_functions_4D.py: functions performing a clustering analysis on the MCMC walkers to maximise the MCMC efficiency.
- plot_functions_4D.py: different routines designed to produce the MCMC plots such as Figs. 1, 6, 7, amongst others.

A requirements file is provided detailing all the different packages and their dependencies that were used for this project. Two of the most critical packages are:

- Emcee: Package which performs MCMC with multiple walkers following the Goodman and Weare implemenation. For more information, please consult: https://emcee.readthedocs.io/en/stable/
- PVLib: Package which allows modelled profiles to be generated which are used for the MCMC sampling. We refer the reader to https://pvlib-python.readthedocs.io/en/stable/ for additional information.

Appendix B. Results of Markov Chain Monte Carlo samples

This section contains the results for $\beta_{pred} = \{\phi_{pred}, \theta_{pred}, N_{p,pred}\}$ and the nuisance parameter σ_{pred} for the twelve chosen PVOutput systems. Please consult Section 5.2 regarding all the information contained within these tables.

Table B.3
Results for $\beta_{pred} = \{\phi_{pred}, \theta_{pred}, N_{p,pred}\}$ and σ_{pred} for the twelve chosen `PVOutput` systems.

id	ϕ_{ref} (°)	θ_{ref} (°)	$N_{p,ref}$	$P_{p,ref}$ (W_p)	P_{ref} (W)	Date	ϕ_{pred} (°)	θ_{pred} (°)	$N_{p,pred}$	σ_{pred}	P_{pred} (W)	accept.	N_{iter}	Clip
1	135	50	15	235	3525	19/07/2016	$144.9^{+0.3}_{-0.3}$	$60.7^{+0.3}_{-0.3}$	15^{+0}_{-0}	$56.0^{+3.2}_{-2.8}$	3450	0.40	2200	0
						26/05/2017	$140.6^{+0.4}_{-0.4}$	$54.9^{+0.4}_{-0.4}$	15^{+0}_{-0}	$56.4^{+3.4}_{-2.8}$	3450	0.40	2450	1
						27/05/2017	$140.2^{+0.4}_{-0.4}$	$56.6^{+0.4}_{-0.4}$	15^{+0}_{-0}	$62.7^{+3.3}_{-3.1}$	3450	0.42	1700	1
						14/06/2017	$141.0^{+0.5}_{-0.5}$	$55.4^{+0.5}_{-0.5}$	15^{+0}_{-0}	$85.4^{+4.6}_{-4.5}$	3450	0.40	2050	1
						19/06/2017	$142.9^{+0.2}_{-0.2}$	$58.6^{+0.2}_{-0.2}$	15^{+0}_{-0}	$48.2^{+2.7}_{-2.5}$	3450	0.42	1500	0
						29/06/2018	$140.2^{+0.3}_{-0.3}$	$56.7^{+0.3}_{-0.3}$	15^{+0}_{-0}	$47.4^{+2.6}_{-2.3}$	3450	0.41	3150	1
						01/07/2018	$139.3^{+0.3}_{-0.3}$	$56.8^{+0.3}_{-0.3}$	16^{+0}_{-0}	$56.2^{+2.9}_{-2.7}$	3680	0.42	2350	1
						02/07/2018	$139.8^{+0.3}_{-0.3}$	$58.2^{+0.3}_{-0.3}$	16^{+0}_{-0}	$62.0^{+3.6}_{-3.3}$	3680	0.39	2700	1
2	225	37	12	250	3520	08/05/2016	$218.7^{+0.8}_{-0.7}$	$35.5^{+0.8}_{-0.7}$	14^{+0}_{-0}	$59.3^{+3.3}_{-3.1}$	3220	0.40	2450	1
						19/07/2016	$223.1^{+0.9}_{-0.8}$	$32.6^{+0.7}_{-0.7}$	14^{+0}_{-0}	$46.9^{+2.7}_{-2.2}$	3220	0.41	2150	1
						17/08/2016	$245.7^{+1.0}_{-1.1}$	$23.3^{+0.4}_{-0.4}$	17^{+0}_{-0}	$86.2^{+4.5}_{-4.5}$	3910	0.40	1700	0
						26/05/2017	$225.6^{+0.8}_{-0.8}$	$31.3^{+0.6}_{-0.6}$	14^{+0}_{-0}	$46.8^{+2.6}_{-2.5}$	3220	0.39	3150	1
						27/05/2017	$219.5^{+0.6}_{-0.6}$	$35.1^{+0.5}_{-0.5}$	14^{+0}_{-0}	$54.7^{+3.0}_{-2.9}$	3220	0.40	3450	1
						19/06/2017	$218.2^{+0.6}_{-0.6}$	$36.6^{+0.6}_{-0.6}$	14^{+0}_{-0}	$48.7^{+2.8}_{-2.4}$	3220	0.41	2450	1
						04/05/2018	$215.7^{+0.8}_{-0.8}$	$36.9^{+0.9}_{-0.9}$	14^{+0}_{-0}	$63.8^{+3.6}_{-3.5}$	3220	0.41	3550	1
						05/05/2018	$217.1^{+1.0}_{-1.0}$	$37.2^{+1.1}_{-1.2}$	14^{+0}_{-0}	$87.9^{+5.1}_{-4.4}$	3220	0.41	1850	1
						06/05/2018	$218.8^{+1.0}_{-0.9}$	$35.4^{+0.9}_{-0.9}$	14^{+0}_{-0}	$72.1^{+4.1}_{-3.7}$	3220	0.40	2600	1
						07/05/2018	$218.8^{+0.8}_{-0.7}$	$35.3^{+0.8}_{-0.8}$	14^{+0}_{-0}	$53.6^{+3.1}_{-3.0}$	3220	0.41	1800	1
						08/05/2018	$215.4^{+1.8}_{-1.7}$	$38.0^{+2.2}_{-2.2}$	14^{+0}_{-0}	$55.3^{+3.2}_{-2.6}$	3220	0.41	2300	1
						29/06/2018	$225.3^{+0.8}_{-0.8}$	$31.2^{+0.6}_{-0.6}$	14^{+0}_{-0}	$49.3^{+2.6}_{-2.5}$	3220	0.41	2250	1
						01/07/2018	$230.0^{+1.0}_{-1.1}$	$29.9^{+0.6}_{-0.6}$	15^{+0}_{-0}	$55.7^{+3.0}_{-2.7}$	3450	0.41	1600	0
						02/07/2018	$236.5^{+1.8}_{-1.9}$	$26.8^{+0.9}_{-0.8}$	15^{+0}_{-0}	$65.3^{+3.8}_{-3.3}$	3450	0.40	1800	0
3	135	15	15	250	3750	08/05/2016	$163.9^{+0.4}_{-0.4}$	$19.8^{+0.4}_{-0.4}$	16^{+0}_{-0}	$35.0^{+2.0}_{-1.8}$	3680	0.40	2600	1
						17/08/2016	$159.1^{+0.4}_{-0.4}$	$16.7^{+0.3}_{-0.3}$	16^{+0}_{-0}	$35.4^{+2.1}_{-1.9}$	3680	0.41	2500	1
						24/08/2016	$138.9^{+1.3}_{-1.4}$	$9.6^{+0.3}_{-0.2}$	16^{+0}_{-0}	$56.2^{+3.0}_{-3.0}$	3680	0.40	2450	0
						25/08/2016	$155.0^{+0.8}_{-0.8}$	$16.1^{+0.4}_{-0.4}$	15^{+0}_{-0}	$55.4^{+3.3}_{-3.0}$	3450	0.41	4650	0
						10/05/2017	$161.5^{+0.5}_{-0.5}$	$14.2^{+0.3}_{-0.3}$	16^{+0}_{-0}	$33.6^{+1.8}_{-1.7}$	3680	0.39	2850	1
						26/05/2017	$163.0^{+0.5}_{-0.5}$	$17.0^{+0.4}_{-0.4}$	16^{+0}_{-0}	$38.0^{+2.1}_{-1.8}$	3680	0.42	1300	1
						27/05/2017	$159.4^{+1.0}_{-0.9}$	$17.1^{+0.7}_{-0.7}$	15^{+0}_{-0}	$68.8^{+3.8}_{-3.6}$	3450	0.40	2400	1
						14/06/2017	$165.6^{+0.7}_{-0.7}$	$19.8^{+0.7}_{-0.7}$	15^{+0}_{-0}	$68.6^{+3.8}_{-3.3}$	3450	0.41	3300	1
						19/06/2017	$159.1^{+0.7}_{-0.8}$	$15.2^{+0.5}_{-0.5}$	15^{+0}_{-0}	$44.2^{+2.6}_{-2.2}$	3450	0.40	1450	1
						04/05/2018	$164.9^{+0.5}_{-0.5}$	$14.7^{+0.3}_{-0.3}$	16^{+0}_{-0}	$39.1^{+2.1}_{-2.0}$	3680	0.40	2150	1
						05/05/2018	$164.8^{+0.4}_{-0.4}$	$17.9^{+0.3}_{-0.3}$	16^{+0}_{-0}	$33.2^{+1.9}_{-1.8}$	3680	0.41	1700	1
						06/05/2018	$165.3^{+0.3}_{-0.4}$	$14.7^{+0.2}_{-0.2}$	16^{+0}_{-0}	$28.7^{+1.6}_{-1.4}$	3680	0.41	2750	1
						07/05/2018	$162.9^{+0.4}_{-0.4}$	$14.3^{+0.2}_{-0.2}$	16^{+0}_{-0}	$27.6^{+1.5}_{-1.5}$	3680	0.40	2550	1
						08/05/2018	$147.1^{+1.0}_{-1.0}$	$8.3^{+0.2}_{-0.2}$	16^{+0}_{-0}	$38.0^{+2.1}_{-2.1}$	3680	0.41	2750	0
						29/06/2018	$164.6^{+0.6}_{-0.6}$	$15.2^{+0.5}_{-0.5}$	16^{+0}_{-0}	$43.3^{+2.3}_{-2.1}$	3680	0.40	1600	1
						01/07/2018	$162.2^{+0.8}_{-0.9}$	$14.0^{+0.5}_{-0.5}$	17^{+0}_{-0}	$51.8^{+2.9}_{-2.6}$	3910	0.41	2600	1
						02/07/2018	$153.0^{+1.6}_{-1.8}$	$10.5^{+0.6}_{-0.6}$	17^{+0}_{-0}	$67.4^{+3.6}_{-3.5}$	3910	0.41	3100	0
4	315	15	9	195	1755	08/05/2016	$257.8^{+1.9}_{-2.0}$	$18.4^{+0.4}_{-0.4}$	6^{+0}_{-0}	$46.5^{+3.0}_{-2.5}$	1380	0.41	3250	0
						19/07/2016	$266.5^{+2.5}_{-2.5}$	$14.8^{+0.4}_{-0.4}$	6^{+0}_{-0}	$42.2^{+2.5}_{-2.2}$	1380	0.41	3400	1
						17/08/2016	$312.2^{+1.2}_{-1.2}$	$19.6^{+0.3}_{-0.3}$	8^{+0}_{-0}	$55.6^{+3.2}_{-3.0}$	1840	0.42	2500	0
						24/08/2016	$266.5^{+1.9}_{-1.9}$	$17.7^{+0.5}_{-0.5}$	6^{+0}_{-0}	$50.9^{+3.2}_{-3.1}$	1380	0.39	2800	1
						25/05/2017	$266.7^{+2.6}_{-3.0}$	$16.0^{+0.5}_{-0.4}$	6^{+0}_{-0}	$52.8^{+3.2}_{-2.9}$	1380	0.41	1450	1
						26/05/2017	$271.7^{+2.8}_{-2.9}$	$15.6^{+0.4}_{-0.4}$	6^{+0}_{-0}	$53.5^{+3.4}_{-3.1}$	1380	0.40	2500	1
						14/06/2017	$252.4^{+5.5}_{-5.6}$	$15.5^{+1.0}_{-0.9}$	6^{+0}_{-0}	$81.4^{+4.8}_{-4.4}$	1380	0.41	2400	1
						04/05/2018	$265.3^{+1.5}_{-1.6}$	$17.0^{+0.3}_{-0.3}$	6^{+0}_{-0}	$39.9^{+2.8}_{-2.3}$	1380	0.41	1600	1
						05/05/2018	$273.0^{+1.7}_{-1.6}$	$15.9^{+0.3}_{-0.3}$	6^{+0}_{-0}	$40.4^{+2.5}_{-2.3}$	1380	0.42	2050	1
						06/05/2018	$270.7^{+1.8}_{-1.8}$	$16.6^{+0.3}_{-0.3}$	6^{+0}_{-0}	$44.1^{+3.1}_{-2.7}$	1380	0.40	1700	1
						07/05/2018	$276.6^{+1.8}_{-1.9}$	$16.0^{+0.3}_{-0.3}$	6^{+0}_{-0}	$44.3^{+3.0}_{-2.6}$	1380	0.40	3200	1
						08/05/2018	$221.9^{+1.6}_{-1.6}$	$29.0^{+1.3}_{-1.2}$	5^{+0}_{-0}	$44.1^{+2.8}_{-2.4}$	1150	0.41	2200	0
						29/06/2018	$237.2^{+3.9}_{-4.0}$	$20.2^{+1.5}_{-1.3}$	6^{+0}_{-0}	$65.1^{+4.2}_{-3.6}$	1380	0.41	2650	0
						01/07/2018	$227.6^{+2.5}_{-2.5}$	$27.2^{+1.5}_{-1.3}$	6^{+0}_{-0}	$62.6^{+3.9}_{-3.9}$	1380	0.41	2300	0
						02/07/2018	$236.7^{+3.6}_{-3.6}$	$22.4^{+1.4}_{-1.3}$	6^{+0}_{-0}	$65.3^{+4.4}_{-3.6}$	1380	0.40	2350	0

(*continued on next page*)

Table B.3 (*continued*).

5	135	45	18	250	4500	08/05/2016	$147.6^{+0.3}_{-0.3}$	$52.6^{+0.4}_{-0.4}$	20^{+0}_{-0}	$59.8^{+3.4}_{-3.0}$	4600	0.41	2400	1
						19/07/2016	$135.5^{+0.6}_{-0.7}$	$40.6^{+0.6}_{-0.6}$	19^{+0}_{-0}	$70.1^{+3.9}_{-3.6}$	4370	0.41	2400	1
						26/05/2017	$144.1^{+0.3}_{-0.4}$	$46.5^{+0.4}_{-0.5}$	19^{+0}_{-0}	$66.1^{+3.4}_{-3.3}$	4370	0.41	2500	1
						27/05/2017	$146.7^{+0.2}_{-0.2}$	$50.8^{+0.3}_{-0.3}$	20^{+0}_{-0}	$58.3^{+3.2}_{-3.0}$	4600	0.41	2450	1
						06/05/2018	$133.0^{+0.9}_{-0.9}$	$41.9^{+1.0}_{-0.9}$	20^{+0}_{-0}	$69.9^{+3.6}_{-3.6}$	4600	0.40	2050	1
						07/05/2018	$127.0^{+1.8}_{-1.6}$	$36.1^{+1.5}_{-1.2}$	20^{+0}_{-0}	$87.0^{+4.9}_{-4.8}$	4600	0.39	2100	1
						08/05/2018	$134.8^{+0.8}_{-0.7}$	$42.5^{+0.8}_{-0.8}$	19^{+0}_{-0}	$83.6^{+4.7}_{-4.4}$	4370	0.40	2450	1
						29/06/2018	$134.4^{+0.5}_{-0.6}$	$44.0^{+0.5}_{-0.5}$	19^{+0}_{-0}	$66.6^{+3.6}_{-3.3}$	4370	0.41	2550	1
						01/07/2018	$138.9^{+0.3}_{-0.4}$	$49.8^{+0.4}_{-0.4}$	21^{+0}_{-0}	$67.2^{+3.4}_{-3.3}$	4830	0.43	1900	0
						02/07/2018	$135.4^{+0.5}_{-0.6}$	$45.5^{+0.6}_{-0.5}$	20^{+0}_{-0}	$73.2^{+3.8}_{-3.6}$	4600	0.40	2700	1
6	180	45	16	245	3920	08/05/2016	$177.1^{+0.1}_{-0.1}$	$45.2^{+0.3}_{-0.4}$	17^{+0}_{-0}	$36.6^{+2.1}_{-1.8}$	3910	0.40	3450	1
						19/07/2016	$178.8^{+0.2}_{-0.2}$	$42.4^{+0.4}_{-0.4}$	16^{+0}_{-0}	$46.9^{+2.5}_{-2.3}$	3680	0.41	3350	1
						17/08/2016	$178.9^{+0.2}_{-0.2}$	$35.6^{+0.6}_{-0.6}$	18^{+0}_{-0}	$38.2^{+2.2}_{-1.9}$	4140	0.41	3000	1
						24/08/2016	$177.4^{+0.3}_{-0.3}$	$21.3^{+0.3}_{-0.3}$	18^{+0}_{-0}	$37.9^{+2.3}_{-2.0}$	4140	0.40	2750	0
						25/08/2016	$177.5^{+0.2}_{-0.2}$	$36.2^{+0.8}_{-0.8}$	17^{+0}_{-0}	$39.3^{+2.4}_{-2.2}$	3910	0.41	1450	1
						26/05/2017	$178.8^{+0.2}_{-0.2}$	$46.8^{+0.3}_{-0.3}$	17^{+0}_{-0}	$45.2^{+2.2}_{-1.9}$	3910	0.41	2700	1
						14/06/2017	$179.0^{+0.2}_{-0.2}$	$46.2^{+0.3}_{-0.3}$	17^{+0}_{-0}	$49.3^{+2.8}_{-2.4}$	3910	0.40	2450	1
						19/06/2017	$179.1^{+0.3}_{-0.3}$	$44.4^{+0.5}_{-0.5}$	16^{+0}_{-0}	$70.9^{+3.6}_{-3.4}$	3680	0.42	2750	1
						04/05/2018	$178.1^{+0.1}_{-0.1}$	$42.8^{+0.5}_{-0.4}$	17^{+0}_{-0}	$34.7^{+1.9}_{-1.8}$	3910	0.41	2400	1
						05/05/2018	$178.4^{+0.2}_{-0.1}$	$41.6^{+0.4}_{-0.4}$	17^{+0}_{-0}	$38.8^{+2.3}_{-2.0}$	3910	0.40	2900	1
						06/05/2018	$178.0^{+0.1}_{-0.1}$	$43.4^{+0.4}_{-0.4}$	17^{+0}_{-0}	$37.6^{+2.2}_{-2.0}$	3910	0.41	2200	1
						07/05/2018	$177.4^{+0.2}_{-0.2}$	$47.3^{+0.5}_{-0.5}$	17^{+0}_{-0}	$54.9^{+3.1}_{-2.8}$	3910	0.41	3850	1
						08/05/2018	$178.0^{+0.2}_{-0.2}$	$38.0^{+0.5}_{-0.5}$	16^{+0}_{-0}	$44.5^{+2.5}_{-2.2}$	3680	0.40	2550	1
						01/07/2018	$178.9^{+0.2}_{-0.2}$	$49.2^{+0.4}_{-0.3}$	18^{+0}_{-0}	$72.2^{+4.0}_{-3.7}$	4140	0.41	1750	1
						02/07/2018	$178.5^{+0.2}_{-0.2}$	$43.7^{+0.4}_{-0.4}$	17^{+0}_{-0}	$67.7^{+3.4}_{-3.3}$	3910	0.42	2250	1
7	180	40	22	250	5500	08/05/2016	$183.9^{+0.3}_{-0.3}$	$29.4^{+0.8}_{-0.9}$	24^{+0}_{-0}	$74.0^{+4.2}_{-3.8}$	5520	0.41	2600	1
						19/07/2016	$182.1^{+0.2}_{-0.2}$	$28.6^{+0.5}_{-0.5}$	23^{+0}_{-0}	$59.6^{+3.0}_{-2.9}$	5290	0.42	2000	1
						24/08/2016	$189.4^{+0.9}_{-0.8}$	$7.7^{+0.2}_{-0.2}$	28^{+0}_{-0}	$63.8^{+3.8}_{-3.5}$	6440	0.41	3300	0
						25/08/2016	$184.2^{+1.0}_{-1.0}$	$9.0^{+0.3}_{-0.2}$	27^{+0}_{-0}	$85.3^{+5.0}_{-4.8}$	6210	0.40	3650	0
						25/05/2017	$182.9^{+0.3}_{-0.3}$	$32.5^{+1.0}_{-2.6}$	23^{+0}_{-0}	$91.6^{+5.0}_{-4.6}$	5290	0.34	2150	1
						26/05/2017	$182.4^{+0.2}_{-0.2}$	$32.3^{+0.5}_{-0.5}$	23^{+0}_{-0}	$58.2^{+2.9}_{-3.0}$	5290	0.41	3150	1
						04/05/2018	$183.7^{+0.3}_{-0.2}$	$24.2^{+0.6}_{-0.5}$	24^{+0}_{-0}	$60.2^{+3.5}_{-2.9}$	5520	0.41	2550	1
						05/05/2018	$183.5^{+0.2}_{-0.3}$	$26.6^{+0.8}_{-0.7}$	24^{+0}_{-0}	$72.9^{+4.1}_{-3.9}$	5520	0.41	1350	1
						06/05/2018	$182.6^{+0.2}_{-0.2}$	$24.7^{+0.6}_{-0.6}$	24^{+0}_{-0}	$59.5^{+3.3}_{-2.8}$	5520	0.42	3150	1
						07/05/2018	$182.2^{+0.2}_{-0.2}$	$28.7^{+0.5}_{-0.4}$	23^{+0}_{-0}	$55.8^{+3.1}_{-3.0}$	5290	0.40	2900	1
						08/05/2018	$183.1^{+0.2}_{-0.2}$	$24.5^{+0.5}_{-0.4}$	23^{+0}_{-0}	$52.6^{+2.9}_{-2.5}$	5290	0.42	1050	1
						29/06/2018	$183.4^{+0.2}_{-0.2}$	$29.8^{+0.4}_{-0.5}$	23^{+0}_{-0}	$56.3^{+3.1}_{-2.8}$	5290	0.41	3150	1
						01/07/2018	$182.8^{+0.2}_{-0.2}$	$35.4^{+0.4}_{-0.4}$	25^{+0}_{-0}	$70.1^{+3.7}_{-3.7}$	5750	0.41	3000	0
						02/07/2018	$182.3^{+0.2}_{-0.2}$	$32.1^{+0.4}_{-0.4}$	24^{+0}_{-0}	$61.1^{+3.3}_{-2.9}$	5520	0.41	3250	1
8	180	0	18	250	4500	08/05/2016	$191.1^{+0.3}_{-0.3}$	$40.5^{+0.6}_{-0.7}$	18^{+0}_{-0}	$64.0^{+3.8}_{-3.4}$	4140	0.40	2350	1
						17/08/2016	$193.4^{+0.7}_{-0.7}$	$35.7^{+1.7}_{-1.6}$	19^{+0}_{-0}	$93.1^{+5.5}_{-5.5}$	4370	0.41	1250	0
						24/08/2016	$192.1^{+0.6}_{-0.6}$	$37.3^{+1.5}_{-1.4}$	18^{+0}_{-0}	$95.0^{+5.5}_{-5.3}$	4140	0.41	3150	1
						25/08/2016	$191.7^{+0.6}_{-0.6}$	$36.7^{+1.8}_{-1.6}$	18^{+0}_{-0}	$78.2^{+4.6}_{-4.0}$	4140	0.42	2300	1
						10/05/2017	$191.3^{+0.7}_{-0.7}$	$37.7^{+1.5}_{-1.6}$	18^{+0}_{-0}	$72.6^{+8.9}_{-7.5}$	4140	0.41	1800	1
						25/05/2017	$192.4^{+0.9}_{-0.9}$	$33.6^{+1.5}_{-1.6}$	17^{+0}_{-0}	$82.2^{+9.7}_{-7.9}$	3910	0.40	3150	0
						26/05/2017	$190.4^{+0.6}_{-0.6}$	$44.5^{+1.1}_{-1.1}$	18^{+0}_{-0}	$81.0^{+9.5}_{-7.7}$	4140	0.41	2500	0
						14/06/2017	$190.0^{+0.6}_{-0.6}$	$41.3^{+1.0}_{-1.1}$	18^{+0}_{-0}	$88.9^{+9.9}_{-8.6}$	4140	0.42	3100	1
						19/06/2017	$190.5^{+0.4}_{-0.5}$	$39.5^{+0.7}_{-0.7}$	17^{+0}_{-0}	$50.9^{+6.1}_{-5.0}$	3910	0.40	3500	0
						04/05/2018	$191.2^{+0.7}_{-0.6}$	$39.0^{+1.5}_{-1.6}$	18^{+0}_{-0}	$67.9^{+8.5}_{-7.5}$	4140	0.41	1600	1
						05/05/2018	$191.2^{+0.9}_{-1.1}$	$37.4^{+6.0}_{-1.9}$	18^{+0}_{-0}	$83.0^{+10.3}_{-8.6}$	4140	0.31	5550	1
						06/05/2018	$191.4^{+0.8}_{-0.8}$	$39.8^{+1.8}_{-2.1}$	18^{+0}_{-0}	$80.3^{+9.6}_{-8.6}$	4140	0.40	1850	1
						07/05/2018	$188.9^{+0.6}_{-0.5}$	$45.4^{+1.2}_{-1.3}$	18^{+0}_{-0}	$72.6^{+8.1}_{-7.2}$	4140	0.41	1700	0
						08/05/2018	$191.1^{+0.8}_{-0.8}$	$37.6^{+1.7}_{-1.9}$	17^{+0}_{-0}	$87.9^{+10.9}_{-8.6}$	3910	0.40	3400	0
						29/06/2018	$190.8^{+0.6}_{-0.6}$	$43.6^{+0.9}_{-0.9}$	18^{+0}_{-0}	$76.0^{+8.5}_{-7.6}$	4140	0.41	2050	0
						01/07/2018	$192.3^{+0.8}_{-0.8}$	$38.1^{+1.2}_{-1.2}$	18^{+0}_{-0}	$88.2^{+10.0}_{-8.8}$	4140	0.41	1400	1
						02/07/2018	$192.1^{+0.8}_{-0.7}$	$39.2^{+1.0}_{-1.1}$	18^{+0}_{-0}	$83.5^{+9.9}_{-8.2}$	4140	0.40	2150	1

(*continued on next page*)

Table B.3 (*continued*).

9	315	13	4	275	1100	08/05/2016	$340.2^{+0.4}_{-0.5}$	$18.7^{+0.2}_{-0.2}$	5^{+0}_{-0}	$18.6^{+1.1}_{-0.9}$	1150	0.39	3200	1
						19/07/2016	$316.2^{+2.8}_{-3.2}$	$10.4^{+0.5}_{-0.4}$	4^{+0}_{-0}	$36.4^{+2.1}_{-1.9}$	920	0.40	2100	0
						17/08/2016	$339.7^{+0.5}_{-0.5}$	$21.7^{+0.2}_{-0.2}$	6^{+0}_{-0}	$22.5^{+1.3}_{-1.2}$	1380	0.41	2300	1
						24/08/2016	$342.9^{+0.4}_{-0.4}$	$22.6^{+0.1}_{-0.1}$	6^{+0}_{-0}	$18.7^{+1.1}_{-1.0}$	1380	0.40	3300	1
						25/08/2016	$335.2^{+0.5}_{-0.5}$	$16.7^{+0.1}_{-0.1}$	5^{+0}_{-0}	$14.7^{+0.9}_{-0.8}$	1150	0.40	2450	1
						10/05/2017	$335.7^{+1.8}_{-1.8}$	$20.4^{+0.6}_{-0.7}$	5^{+0}_{-0}	$79.1^{+4.2}_{-4.2}$	1150	0.41	2050	1
						25/05/2017	$340.9^{+0.6}_{-0.6}$	$21.7^{+0.3}_{-0.2}$	5^{+0}_{-0}	$28.6^{+1.6}_{-1.4}$	1150	0.41	2400	1
						26/05/2017	$340.3^{+0.6}_{-0.6}$	$21.0^{+0.2}_{-0.2}$	5^{+0}_{-0}	$25.9^{+1.5}_{-1.3}$	1150	0.40	2250	1
						27/05/2017	$327.3^{+5.3}_{-6.9}$	$8.2^{+1.1}_{-1.0}$	4^{+0}_{-0}	$68.8^{+4.2}_{-3.3}$	920	0.39	1250	0
						14/06/2017	$282.2^{+4.0}_{-3.9}$	$8.3^{+0.3}_{-0.3}$	4^{+0}_{-0}	$29.9^{+1.6}_{-1.4}$	920	0.41	3050	0
						19/06/2017	$318.4^{+1.5}_{-1.7}$	$11.4^{+0.3}_{-0.3}$	4^{+0}_{-0}	$22.8^{+1.3}_{-1.1}$	920	0.39	2850	0
						04/05/2018	$339.1^{+0.4}_{-0.3}$	$18.5^{+0.1}_{-0.1}$	5^{+0}_{-0}	$14.3^{+0.8}_{-0.8}$	1150	0.40	3250	1
						05/05/2018	$338.6^{+0.4}_{-0.4}$	$18.3^{+0.1}_{-0.1}$	5^{+0}_{-0}	$14.4^{+0.8}_{-0.8}$	1150	0.37	3550	1
						06/05/2018	$339.6^{+1.0}_{-1.0}$	$19.2^{+0.4}_{-0.4}$	5^{+0}_{-0}	$44.4^{+2.5}_{-2.1}$	1150	0.42	3000	1
						07/05/2018	$341.8^{+0.4}_{-0.4}$	$19.5^{+0.1}_{-0.1}$	5^{+0}_{-0}	$16.5^{+0.9}_{-0.9}$	1150	0.40	1900	1
						08/05/2018	$343.8^{+0.4}_{-0.4}$	$20.8^{+0.1}_{-0.2}$	5^{+0}_{-0}	$17.5^{+1.0}_{-0.9}$	1150	0.40	2900	1
						29/06/2018	$340.6^{+0.6}_{-0.7}$	$22.6^{+0.3}_{-0.3}$	5^{+0}_{-0}	$30.2^{+1.6}_{-1.5}$	1150	0.41	2100	1
						01/07/2018	$336.8^{+0.7}_{-0.7}$	$20.3^{+0.3}_{-0.3}$	5^{+0}_{-0}	$28.5^{+1.4}_{-1.4}$	1150	0.42	2150	1
						02/07/2018	$283.9^{+4.9}_{-5.5}$	$9.2^{+0.4}_{-0.4}$	4^{+0}_{-0}	$43.6^{+2.4}_{-2.2}$	920	0.41	2250	0
10	315	46	18	170	3060	08/05/2016	$286.4^{+0.7}_{-0.7}$	$48.8^{+0.5}_{-0.5}$	16^{+0}_{-0}	$58.1^{+4.6}_{-4.4}$	3680	0.40	1800	1
						19/07/2016	$301.7^{+0.9}_{-0.9}$	$47.8^{+0.5}_{-0.5}$	18^{+0}_{-0}	$93.5^{+7.5}_{-6.8}$	4140	0.41	2050	1
						26/05/2017	$292.8^{+0.6}_{-0.6}$	$48.7^{+0.3}_{-0.3}$	16^{+0}_{-0}	$46.8^{+3.7}_{-3.2}$	3680	0.40	2850	1
						07/05/2018	$279.8^{+1.2}_{-1.4}$	$52.3^{+1.0}_{-1.0}$	15^{+0}_{-0}	$84.6^{+6.9}_{-6.1}$	3450	0.38	1650	1
						08/05/2018	$264.8^{+2.9}_{-3.2}$	$60.8^{+2.7}_{-2.3}$	14^{+0}_{-0}	$76.5^{+6.0}_{-5.7}$	3220	0.39	1750	0
						29/06/2018	$294.5^{+0.8}_{-0.8}$	$49.2^{+0.4}_{-0.4}$	16^{+0}_{-0}	$59.8^{+4.5}_{-4.3}$	3680	0.41	1900	1
						02/07/2018	$295.8^{+0.9}_{-1.0}$	$50.2^{+0.5}_{-0.5}$	17^{+0}_{-0}	$77.6^{+6.0}_{-5.5}$	3910	0.39	3050	1
11	45	13	6	170	2040	08/05/2016	$350.4^{+0.7}_{-0.7}$	$7.2^{+0.1}_{-0.1}$	10^{+0}_{-0}	$22.3^{+1.3}_{-1.2}$	2300	0.40	3300	1
						19/07/2016	$355.8^{+0.6}_{-0.6}$	$10.8^{+0.2}_{-0.2}$	10^{+0}_{-0}	$27.8^{+1.6}_{-1.5}$	2300	0.40	2600	0
						17/08/2016	$354.1^{+0.3}_{-0.3}$	$16.1^{+0.1}_{-0.1}$	13^{+0}_{-0}	$25.5^{+1.4}_{-1.4}$	2990	0.42	2500	0
						24/08/2016	$355.4^{+0.3}_{-0.3}$	$17.7^{+0.1}_{-0.1}$	13^{+0}_{-0}	$22.6^{+1.4}_{-1.2}$	2990	0.40	3450	0
						25/08/2016	$355.2^{+0.4}_{-0.4}$	$14.6^{+0.1}_{-0.1}$	12^{+0}_{-0}	$24.3^{+1.3}_{-1.3}$	2760	0.39	2950	0
						26/05/2017	$351.8^{+0.7}_{-0.7}$	$9.6^{+0.2}_{-0.2}$	10^{+0}_{-0}	$27.0^{+1.6}_{-1.4}$	2300	0.40	2650	1
						27/05/2017	$184.5^{+4.0}_{-4.0}$	$2.7^{+0.4}_{-0.4}$	9^{+0}_{-0}	$40.5^{+2.2}_{-2.2}$	2070	0.41	2750	0
						14/06/2017	$353.3^{+1.3}_{-1.4}$	$7.9^{+0.3}_{-0.3}$	10^{+0}_{-0}	$44.2^{+2.5}_{-2.3}$	2300	0.41	2550	1
						19/06/2017	$332.5^{+6.8}_{-8.0}$	$1.6^{+0.3}_{-0.3}$	9^{+0}_{-0}	$34.6^{+2.0}_{-1.6}$	2070	0.40	2200	0
						04/05/2018	$347.5^{+0.6}_{-0.7}$	$7.4^{+0.1}_{-0.1}$	10^{+0}_{-0}	$21.3^{+1.2}_{-1.1}$	2300	0.41	2100	1
						05/05/2018	$345.9^{+1.6}_{-1.7}$	$7.2^{+0.3}_{-0.3}$	10^{+0}_{-0}	$52.9^{+3.1}_{-2.7}$	2300	0.41	2300	0
						06/05/2018	$349.3^{+0.8}_{-0.7}$	$7.0^{+0.1}_{-0.1}$	10^{+0}_{-0}	$23.4^{+1.3}_{-1.2}$	2300	0.40	3100	1
						07/05/2018	$352.7^{+0.6}_{-0.6}$	$8.0^{+0.1}_{-0.1}$	10^{+0}_{-0}	$22.5^{+1.2}_{-1.2}$	2300	0.42	1650	1
						08/05/2018	$354.3^{+0.7}_{-0.7}$	$9.4^{+0.2}_{-0.2}$	10^{+0}_{-0}	$30.2^{+1.6}_{-1.5}$	2300	0.39	2150	1
						29/06/2018	$353.3^{+0.8}_{-0.8}$	$9.0^{+0.2}_{-0.2}$	10^{+0}_{-0}	$29.0^{+1.6}_{-1.4}$	2300	0.39	3200	1
						01/07/2018	$340.1^{+1.8}_{-1.9}$	$4.3^{+0.2}_{-0.2}$	10^{+0}_{-0}	$29.5^{+1.6}_{-1.5}$	2300	0.41	1850	0
						02/07/2018	$349.3^{+0.9}_{-1.0}$	$6.4^{+0.2}_{-0.2}$	10^{+0}_{-0}	$25.2^{+1.4}_{-1.2}$	2300	0.41	2150	1
12	0	7	10	250	2500	19/07/2016	$19.8^{+0.4}_{-0.4}$	$14.0^{+0.1}_{-0.1}$	11^{+0}_{-0}	$25.9^{+1.5}_{-1.4}$	2530	0.40	2650	1
						17/08/2016	$11.8^{+0.3}_{-0.2}$	$16.3^{+0.1}_{-0.1}$	14^{+0}_{-0}	$21.4^{+1.2}_{-1.1}$	3220	0.40	3450	0
						24/08/2016	$10.4^{+0.2}_{-0.2}$	$18.6^{+0.1}_{-0.1}$	14^{+0}_{-0}	$20.6^{+1.6}_{-1.3}$	3220	0.41	3050	0
						25/08/2016	$9.3^{+0.6}_{-0.6}$	$21.4^{+0.2}_{-0.2}$	15^{+0}_{-0}	$52.3^{+4.0}_{-3.4}$	3450	0.41	2750	0
						25/05/2017	$14.8^{+0.4}_{-0.4}$	$13.4^{+0.1}_{-0.1}$	11^{+0}_{-0}	$22.2^{+1.2}_{-1.2}$	2530	0.39	1400	1
						26/05/2017	$15.8^{+0.3}_{-0.3}$	$13.0^{+0.1}_{-0.1}$	11^{+0}_{-0}	$19.1^{+1.1}_{-1.0}$	2530	0.40	3600	1
						14/06/2017	$16.3^{+1.0}_{-0.8}$	$14.2^{+0.3}_{-0.3}$	11^{+0}_{-0}	$56.5^{+3.2}_{-2.8}$	2530	0.42	2950	1
						04/05/2018	$11.3^{+0.3}_{-0.3}$	$14.7^{+0.1}_{-0.1}$	12^{+0}_{-0}	$21.3^{+1.3}_{-1.2}$	2760	0.40	2450	1
						05/05/2018	$12.5^{+0.3}_{-0.3}$	$14.6^{+0.1}_{-0.1}$	12^{+0}_{-0}	$20.2^{+1.3}_{-1.1}$	2760	0.41	2100	1
						06/05/2018	$11.9^{+0.8}_{-0.8}$	$16.2^{+0.3}_{-0.3}$	12^{+0}_{-0}	$63.5^{+4.5}_{-3.9}$	2760	0.40	3050	1
						07/05/2018	$14.3^{+0.3}_{-0.3}$	$15.3^{+0.1}_{-0.1}$	12^{+0}_{-0}	$18.6^{+1.1}_{-1.1}$	2760	0.42	2200	1
						08/05/2018	$19.8^{+0.4}_{-0.4}$	$11.8^{+0.1}_{-0.1}$	11^{+0}_{-0}	$21.3^{+1.4}_{-1.3}$	2530	0.41	1950	0
						29/06/2018	$19.0^{+1.6}_{-1.6}$	$11.2^{+0.4}_{-0.4}$	11^{+0}_{-0}	$46.2^{+4.4}_{-3.9}$	2530	0.41	1650	0
						01/07/2018	$14.8^{+0.6}_{-0.6}$	$14.8^{+0.2}_{-0.2}$	12^{+0}_{-0}	$27.7^{+2.8}_{-2.2}$	2760	0.41	2000	1
						02/07/2018	$13.5^{+0.5}_{-0.5}$	$15.2^{+0.2}_{-0.2}$	12^{+0}_{-0}	$24.2^{+2.5}_{-2.2}$	2760	0.39	1550	1

References

[1] J. Weaver, World has Installed 1TW of Solar Capacity, PV Magazine, 2022, https://www.pv-magazine.com/2022/03/15/humans-have-installed-1-terawatt-of-solar-capacity/.

[2] IEA-PVPS, Trends in Photovoltaic Applications 2022, Report IEA PVPS T1-43:2022, 2022, https://iea-pvps.org/trends_reports/trends-2022/.

[3] C. Breyer, D. Bogdanov, A. Gulagi, A. Aghahosseini, L.S. Barbosa, O. Koskinen, M. Barasa, U. Caldera, S. Afanasyeva, M. Child, J. Farfan, P. Vainikka, On the role of solar photovoltaics in global energy transition scenarios, Prog. Photovolt., Res. Appl. 25 (8) (2017) 727–745, http://dx.doi.org/10.1002/pip.2885.

[4] N.M. Haegel, H. Atwater, T. Barnes, C. Breyer, A. Burrell, Y.-M. Chiang, S.D. Wolf, B. Dimmler, D. Feldman, S. Glunz, J.C. Goldschmidt, D. Hochschild, R.

Inzunza, I. Kaizuka, B. Kroposki, S. Kurtz, S. Leu, R. Margolis, K. Matsubara, A. Metz, W.K. Metzger, M. Morjaria, S. Niki, S. Nowak, I.M. Peters, S. Philipps, T. Reindl, A. Richter, D. Rose, K. Sakurai, R. Schlatmann, M. Shikano, W. Sinke, R. Sinton, B. Stanbery, M. Topic, W. Tumas, Y. Ueda, J. van de Lagemaat, P. Verlinden, M. Vetter, E. Warren, M. Werner, M. Yamaguchi, A.W. Bett, Terawatt-scale photovoltaics: Transform global energy, Science 364 (6443) (2019) 836–838, http://dx.doi.org/10.1126/science.aaw1845.

[5] P.J. Verlinden, Future challenges for photovoltaic manufacturing at the terawatt level, J. Renew. Sustain. Energy (2020).

[6] M. Victoria, N. Haegel, I.M. Peters, R. Sinton, A. Jäger-Waldau, C. del Cañizo, C. Breyer, M. Stocks, A. Blakers, I. Kaizuka, K. Komoto, A. Smets, Solar photovoltaics is ready to power a sustainable future, Joule 5 (5) (2021) 1041–1056, http://dx.doi.org/10.1016/j.joule.2021.03.005.

[7] P. Moraitis, B.B. Kausika, N. Nortier, W. van Sark, Urban environment and solar PV performance: The case of the netherlands, Energies 11 (2018) 1333, http://dx.doi.org/10.3390/en11061333.

[8] S. Killinger, D. Lingfors, Y.-M. Saint-Drenan, P. Moraitis, W. van Sark, J. Taylor, N.A. Engerer, J.M. Bright, On the search for representative characteristics of PV systems: Data collection and analysis of PV system azimuth, tilt, capacity, yield and shading, Sol. Energy 173 (2018) 1087–1106, http://dx.doi.org/10.1016/j.solener.2018.08.051.

[9] J. Leloux, L. Narvarte, A. Desportes, D. Trebosc, Performance to peers (P2P): A benchmark approach to fault detections applied to photovoltaic system fleets, Sol. Energy 202 (2020) 522–539, http://dx.doi.org/10.1016/j.solener.2020.03.015.

[10] B. Meng, R. Loonen, J. Hensen, Data-driven inference of unknown tilt and azimuth of distributed pv systems, Sol. Energy 211 (2020) 418–432, http://dx.doi.org/10.1016/j.solener.2020.09.077.

[11] Y.M. Saint-Drenan, S. Bofinger, R. Fritz, S. Vogt, G.H. Good, J. Dobschinski, An empirical approach to parameterizing photovoltaic plants for power forecasting and simulation, Sol. Energy 120 (2015) 479–493, http://dx.doi.org/10.1016/j.solener.2015.07.024.

[12] V.D. Ruelle, M. Jeppesen, M. Brear, Rooftop PV model technical report, 2016, https://aemo.com.au/-/media/files/electricity/nem/planning_and_forecasting/demand-forecasts/nefr/2016/uom-rooftop-pv-model-technical-report.pdf.

[13] W.F. Holmgren, C.W. Hansen, M.A. Mikofski, pvlib python: a python package for modeling solar energy systems, J. Open Source Softw. 3 (29) (2018) 884, http://dx.doi.org/10.21105/joss.00884.

[14] R. Perez, P. Ineichen, R. Seals, J. Michalsky, R. Stewart, Modeling daylight availability and irradiance components from direct and global irradiance, Sol. Energy 44 (5) (1990) 271–289, http://dx.doi.org/10.1016/0038-092X(90)90055-H.

[15] K. Brecl, J. Ascencio-Vásquez, M. Topič, Performance of PV systems in Slovenia with the help of typical daily profiles and automatic detection of orientation and inclination angles, Sol. Energy 236 (2022) 870–878, http://dx.doi.org/10.1016/j.solener.2022.03.059.

[16] J. Ascencio-Vásquez, J.C. Osorio-Aravena, K. Brecl, E. Muñoz-Cerón, M. Topič, Typical daily profiles, a novel approach for photovoltaics performance assessment: Case study on large-scale systems in Chile, Sol. Energy 225 (2021) 357–374, http://dx.doi.org/10.1016/j.solener.2021.07.007.

[17] T. De Jong, S. Bromuri, X. Chang, M. Debusschere, N. Rosenski, C. Schartner, K. Strauch, M. Boehmer, L. Curier, Monitoring spatial sustainable development: semi-automated analysis of satellite and aerial images for energy transition and sustainability indicators, 2020, arXiv:2009.05738.

[18] Statistics Netherlands, Smart Ways of Monitoring Solar Power, Statistics Netherlands, The Hague, 2019, https://www.cbs.nl/en-gb/about-us/innovation/project/smart-ways-of-monitoring-solar-power.

[19] B.P. Laevens, O. ten Bosch, F.P. Pijpers, W.G. van Sark, An observational method for determining daily and regional photovoltaic solar energy statistics, Sol. Energy 228 (2021) 12–26, http://dx.doi.org/10.1016/j.solener.2021.08.077.

[20] PVOutput, Latest Outputs, PVOutput, 2020, https://pvoutput.org.

[21] Koninklijk Nederlands Meteorologisch Instituut, KNMI - About KNMI, Koninklijk Nederlands Meteorologisch Instituut, De Bilt, 2020, https://www.knmi.nl/over-het-knmi/about.

[22] H.M. Deneke, A.J. Feijt, R.A. Roebeling, Estimating surface solar irradiance from METEOSAT SEVIRI-derived cloud properties, Remote Sens. Environ. 112 (6) (2008) 3131–3141, http://dx.doi.org/10.1016/j.rse.2008.03.012.

[23] W. Greuell, J.F. Meirink, P. Wang, Retrieval and validation of global, direct, and diffuse irradiance derived from SEVIRI satellite observations, J. Geophys. Res. (Atmos.) 118 (2013) 2340–2361, http://dx.doi.org/10.1002/jgrd.50194.

[24] M. Bayes, M. Price, An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S., Philos. Trans. R. Soc. Lond. Ser. I 53 (1763) 370–418.

[25] Y. Si, J.S. Trangucci, J. Gabry, A. Gelman, Bayesian hierarchical weighting adjustment and survey inference, Surv. Methodol. 46 (2) (2020) 181–214, URL: https://www150.statcan.gc.ca/n1/pub/12-001-x/2020002/article/00003-eng.htm.

[26] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (6) (1953) 1087–1092, http://dx.doi.org/10.1063/1.1699114.

[27] W. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1) (1970) 97–109, http://dx.doi.org/10.1093/biomet/57.1.97.

[28] Pierre Simon Laplace, Memoir on the probability of the causes of events, Statist. Sci. 1 (3) (1986) 364–378, URL: http://www.jstor.org/stable/2245476.

[29] E. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, Cambridge, 2003, http://dx.doi.org/10.1017/CBO9780511790423,

[30] D.J.C. MacKay, Information Theory, Inference & Learning Algorithms, Cambridge University Press, USA, 2003.

[31] D. Foreman-Mackey, D.W. Hogg, D. Lang, J. Goodman, emcee: The MCMC hammer, Publ. Astron. Soc. Pac. 125 (925) (2013) 306, http://dx.doi.org/10.1086/670067.

[32] J. Goodman, J. Weare, Ensemble samplers with affine invariance, Commun. Appl. Math. Comput. Sci. 5 (1) (2010) 65–80, http://dx.doi.org/10.2140/camcos.2010.5.65.

[33] A. Gelman, G.O. Roberts, W.R. Gilks, Efficient Metropolis jumping rules, in: J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (Eds.), Bayesian Statistics, Oxford University Press, Oxford, 1996, pp. 599–608.

[34] F. Hou, J. Goodman, D.W. Hogg, J. Weare, C. Schwab, An affine-invariant sampler for exoplanet fitting and discovery in radial velocity data, Astrophys. J. 745 (2) (2012) 198, http://dx.doi.org/10.1088/0004-637X/745/2/198.

[35] A. Sokal, Monte Carlo methods in statistical mechanics: Foundations and new algorithms, in: C. DeWitt-Morette, P. Cartier, A. Folacci (Eds.), Functional Integration: Basics and Applications, Springer US, Boston, MA, 1997, pp. 131–192, http://dx.doi.org/10.1007/978-1-4899-0319-8_6.

[36] X. Sun, J.M. Bright, C.A. Gueymard, B. Acord, P. Wang, N.A. Engerer, Worldwide performance assessment of 75 global clear-sky irradiance models using principal component analysis, Renew. Sustain. Energy Rev. 111 (2019) 550–570, http://dx.doi.org/10.1016/j.rser.2019.04.006.

[37] Koninklijk Nederlands Meteorologisch Instituut, Jaaroverzicht van het weer in Nederland – 2016, Koninklijk Nederlands Meteorologisch Instituut, De Bilt, 2017, https://cdn.knmi.nl/knmi/map/page/klimatologie/gegevens/mow/jow_2016.pdf.

[38] Koninklijk Nederlands Meteorologisch Instituut, Jaaroverzicht van het weer in Nederland – 2018, Koninklijk Nederlands Meteorologisch Instituut, De Bilt, 2019, https://cdn.knmi.nl/knmi/map/page/klimatologie/gegevens/mow/jow_2018.pdf.

[39] Koninklijk Nederlands Meteorologisch Instituut, Jaaroverzicht van het weer in Nederland – 2017, Koninklijk Nederlands Meteorologisch Instituut, De Bilt, 2018, https://cdn.knmi.nl/knmi/map/page/klimatologie/gegevens/mow/jow_2017.pdf.

[40] B.Y. Liu, R.C. Jordan, The interrelationship and characteristic distribution of direct, diffuse and total solar radiation, Sol. Energy 4 (3) (1960) 1–19, http://dx.doi.org/10.1016/0038-092X(60)90062-1.

[41] N. Haghdadi, J. Copper, A. Bruce, I. MacGill, A method to estimate the location and orientation of distributed photovoltaic systems from their generation output data, Renew. Energy 108 (2017) 390–400, http://dx.doi.org/10.1016/j.renene.2017.02.080.

[42] A. Reinders, P. Verlinden, W. van Sark, A. Freundlich, Photovoltaic Solar Energy: From Fundamentals to Applications, Wiley & Sons, 2016, http://dx.doi.org/10.1002/9781118927496.

[43] N.H. Reich, B. Mueller, A. Armbruster, W.G.J.H.M. van Sark, K. Kiefer, C. Reise, Performance ratio revisited: is PR > 90% realistic? Prog. Photovolt., Res. Appl. 20 (6) (2012) 717–726, http://dx.doi.org/10.1002/pip.1219.