Solar Energy 228 (2021) 12–26

Contents lists available at ScienceDirect

Solar Energy

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

An observational method for determining daily and regional photovoltaic solar energy statistics

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

[a] Statistics Netherlands, Henri Faasdreef 312, 2492JP The Hague, the Netherlands
[b] Ministry of Economic Affairs and Climate Policy, Bezuidenhoutseweg 73, 2594AC The Hague, the Netherlands
[c] Copernicus Institute, Utrecht University, Princetonlaan 8A, 3584 CB Utrecht, the Netherlands

ARTICLE INFO

Keywords:
PV Systems
Citizen science
Representativeness

ABSTRACT

This paper presents a classical estimation problem for calculating the energy generated by photovoltaic solar energy systems, on a daily, annual, regional and national basis. Our methodology relies on two data sources: PVOutput, an online portal with solar energy production measurements, and modelled irradiance data available for large parts of Africa and Europe, from the Royal Netherlands Meteorological Institute. Combining these, we obtain probability functions of observing energy production, given the irradiation. These are applied to a PV systems database, using *Monte Carlo sampling*, allowing daily and annual solar energy production to be calculated. These are, in turn, used to calculate solar energy production per municipality. As a case study, we apply this methodology to one country in particular, namely the Netherlands. By examining the variation in our estimates as a result of taking different subsets of PVOutput systems with certain specifications such as azimuth, tilt and inverter loading ratio, we obtain specific annual energy yields in the range of 877 −946 kWh/kW$_p$ and 838 −899 kWh/kW$_p$ for 2016 and 2017 respectively. The current method used at Statistics Netherlands assumes this to be 875 kWh/kW$_p$, irrespective of irradiation, meaning the yields were underestimated in 2016 and overestimated in 2017. In the case of the Netherlands, this research demonstrates that an irradiation based measure of solar energy generation is necessary. More generally, this research shows that different types of open data sources may be combined to develop models that calculate the energy production of PV system populations.

1. Introduction

Over the past decade, photovoltaic (PV) systems have seen an explosive growth around the world (IEA, 2020). In the European Union, for example, the installed capacity has increased from 29 GW$_p$ in 2010 (EUobserver, 2011) to 131 GW$_p$ in 2019 (EUobserver, 2020). With policy measures, targeted at drastically increasing the renewables share, starting to take effect, attention to detailed and accurate measurements of solar energy production on a highly frequent and regional basis are becoming ever more important for policy makers. To achieve this, extra information may be acquired from a wealth of openly available data sources such as citizen science projects (e.g. Kirk et al., 2021) or satellite data (e.g. Malof et al., 2016). Through combining such data sources with more traditional data sources, new, simple statistical methods may be devised and applied, providing fast and timely ways of estimating local and daily production from PV systems.

1.1. Past research

A vast number of different research papers, concerned with nowcasting or forecasting regional energy production, is available. Here, we highlight some of the methods, relevant to our research, that are comprehensively presented in Bright et al. (2018) and Saint-Drenan et al. (2019). Often, so–called upscaling methods are used to extrapolate information from a small set of reference PV systems to a set of target PV systems, hence circumventing the general penury of largely available open measurement data. Bright et al. (2018) identify three different subcategories within the realm of upscaling: the first is to use various methods to extrapolate reference systems' measurements to target system information through e.g. inverse distance weighting or a different form of calibration (e.g. Schierenbeck et al., 2010; Bessa et al., 2015). The second technique includes performing quality controls to account for variability in the measurement and meta data as well as other systematic effects, before subsequently applying the first approach, (e.g. Killinger et al., 2017). The last category of upscaling methods is more

* Corresponding author.
E-mail address: bpm.laevens@cbs.nl (B.P.M. Laevens).

https://doi.org/10.1016/j.solener.2021.08.077
Received 19 April 2021; Received in revised form 24 August 2021; Accepted 30 August 2021
Available online 23 September 2021
0038-092X/© 2021 International Solar Energy Society. Published by Elsevier Ltd. All rights reserved.

Nomenclature

Abbreviations

PV	Photovoltaic
pc4/6	Postal code 4 or 6 area in the Netherlands
KNMI	Royal Netherlands Meteorological Institute
SN	Statistics Netherlands

Symbols

ϕ	Azimuth angle of a PV system (°)
θ	Tilt angle of a PV system (°)
$\in$	Inverter loading ratio of a PV system (P_i/P)
P	Total power of PV system (W_p)
P_p	Panel power of a PV system (W_p)
P_i	Inverter size of a PV system (W)
Y_{inst}	Instantaneous energy production of a PV system (kWh)
Y_{cum}	Cumulative energy production of a PV system (kWh)
Y_a	Annual energy production (GWh)
Y_d	Daily energy production (GWh)
$Y_{s,a}$	Specific annual yield (kWh/kW_p)
$Y_{s,d}$	Specific daily yield (kWh/kW_p)
G	Modelled global horizontal irradiance (kW/m^2)
H_a	Modelled annual global horizontal irradiation (kWh/m^2)
H_d	Modelled daily global horizontal irradiation (kWh/m^2)
l	Longitude (°)
b	Latitude (°)
d	Installation date of a PV system in the PV Systems database
N_d	Number of systems in the PV Systems database on date d

involved: reference systems' power is converted into irradiance. Subsequently the irradiance is interpolated to the target system locations and then translated back into power (e.g. Marion and Smith, 2017; Killinger et al., 2017). Besides upscaling methods, other methods exist which use meteorological data as input to a model which predicts the output at various locations as a function of PV system parameters (e.g. Junior et al., 2015; Saint-Drenan et al., 2017). Various other methods exist and are explored in more depth in Bright et al. (2018).

1.2. Current method at Statistics Netherlands

In this paper, we report on a new, general, statistical method, which we validate by applying it to the situation of the Netherlands. Statistics Netherlands (SN) is the body responsible for publishing solar energy production statistics in the Netherlands (SN, 2020). Like in the EU, the recorded capacity has substantially increased from 0.09 GW_p in 2010 versus 10.2 GW_p in 2020 (SN, 2020). While small households made up the bulk of the PV systems in the past, as of 2019 just over half of the capacity was made up of large PV systems ($> 15\,kW_p$) such as solar parks (SN, 2019). The foundation for these statistics is a traditional database, containing almost all Dutch PV systems, detailing their location, installation date and system size amongst others. While subsidy schemes for large PV systems have meant that their monthly energy production is recorded (CertiQ, 2020), this is not the case for small households, with the solar energy production unfortunately unknown.

A framework for measuring the total energy production from household PV systems was introduced to combat the paucity of measurement data and is outlined in the Netherlands Enterprise Agency's protocol of renewable energy (RVO et al., 2015). At the centre of this framework is an Eq. (1) that translates the average installed PV capacity in a specific year, into an average annual energy production (Y_a), using the specific annual yield, currently fixed at $Y_{s,a} = 875\,kWh/kW_p$ (van Sark et al., 2014):

$$Y_a = \left(\frac{\sum_{n=1}^{N_1} P_n + \sum_{n=1}^{N_{365}} P_n}{2} \right) Y_{s,a}, \tag{1}$$

where P_n is the system size of one PV system in the database, N_1 and N_{365} the number of PV systems in the database on the first and last days of the year respectively. This calculation is currently performed by SN on a national basis.

The $Y_{s,a} = 875\,kWh/kW_p$ figure originates from a previous analysis of two datasets of PV systems' performances in 2012 and 2013. For both years in question $Y_{s,2012} = 877 \pm 140$ and $Y_{s,2013} = 874 \pm 140\,kWh/kW_p$ were determined (van Sark et al., 2014). A disadvantage of using Eq. (1) is the assumption that $Y_{s,a} = 875$ kWh/kW_p does not vary between years. The total annual irradiation was $H_{2012} = 989$ and $H_{2013} = 1003$ kWh/m^2, both of which are similar to the 30 year average of 986 kWh/m^2. The past four years (2016–2019) have been substantially sunnier with $H_{2016} = 1040, H_{2017} = 1020, H_{2018} = 1137, H_{2019} = 1099$ kWh/m^2 (KNMI, 2017; KNMI, 2018; KNMI, 2019; KNMI, 2020). It is therefore very likely that $Y_{s,a} > 875$ kWh/kW_p for these years.

PV energy production is not just a function of H. While the aforementioned research pinpointed a value for $Y_{s,a}$ in those two years, it could be improved upon in terms of corrections for lack of representativeness. It is known from the literature that various aspects of PV systems are influential in determining $Y_{s,a}$ or Y_a (Reinders et al., 2016). When extrapolating $Y_{s,a}$ to the whole country, for instance, it is important to account for the true distributions of azimuth (ϕ), tilt (θ) and inverter loading ratio ($\in = P_i/P$) of the entire PV systems population. Similarly, it is also imperative to account for the geographical density of PV systems through their longitude (l) and latitude (b) since weather varies locally. It is known that Western parts of the Netherlands receive up to 10% more H, on an annual basis, compared to Eastern parts (Litjens et al., 2017). Western parts of the country have the highest population density, implying high density of household PV systems, whereas sparser Northern and Eastern areas allow for larger PV systems such as solar parks. Finally, technological advancement must also be accounted for.

1.3. Objectives and paper outline

In light of the discussion we have set out in the previous sections, we have developed a method to solve a classical statistical estimation problem: a population of PV systems exists for which we wish to estimate Y_d and Y_a on a national and regional basis, but we only have a non–probability sample of measurements which we need to extrapolate to the whole population. Through the application of simple *Monte Carlo sampling* techniques, we may achieve this. Finally, we investigate the sensitivity of our results for Y_d and Y_a by selecting different priors relating to different distributions of PV systems characteristics (ϕ, θ and $\in$). While this method was born out of a necessity to improve Dutch solar energy statistics, it can be performed in any national context, should three different types of data sources be present around which our method revolves: measurement data of some reference PV systems, irradiance data and a PV systems database. In our case, we used `PVOutput` and modelled irradiance data from the Royal Netherlands Meteorological Institute as well as our own database of PV systems at SN.

In Section 2 we briefly provide more detail of the three aforementioned data sources and the variables they contain. Section 3 details our methodology, followed by a refinement to our method in light of representativeness in Section 4. Daily and annual, national and regional

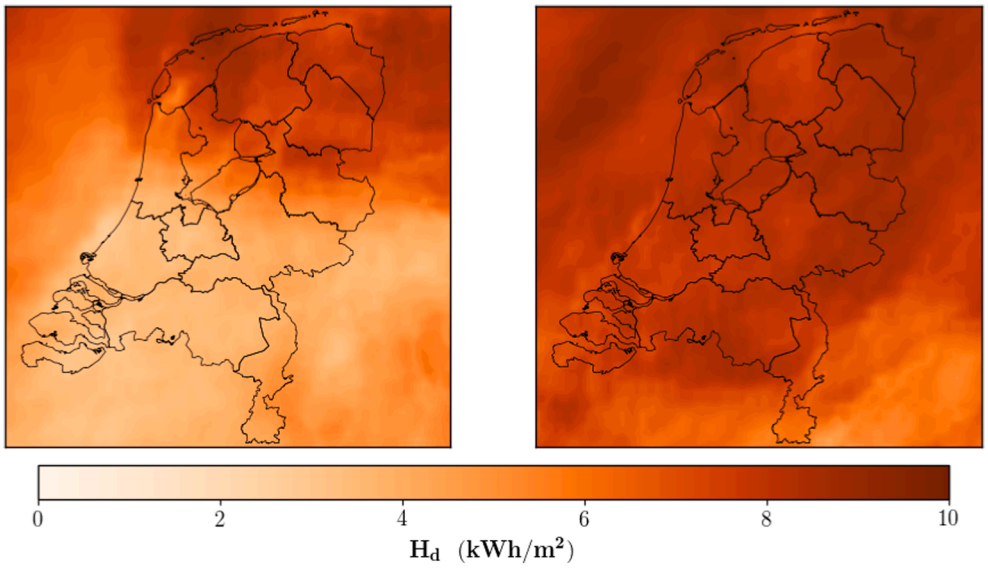

Fig. 1. H_d per surface area (kWh/m^2) for two different days in June 2016: 21 (left) and 22 (right).

estimates are presented in Section 5: we provide discussion and contrast our results with other literature values (where known), before concluding and summarising in Section 6.

2. Data

A brief overview of the three data sources is provided here, with appendix A detailing the cleaning and processing of these data sources.

2.1. PVOutput

`PVOutput` is an Australian online portal with near real–time information of PV energy generation at various locations throughout the world, with substantial registered capacity in Australia, USA, the Netherlands, Italy, Germany, UK and Belgium. As of October 2019, the Dutch capacity stands shy of 50 MW (PVOutput, 2020), which is $\sim 0.9\%$ of the total Dutch capacity and amounting to some 5600 PV systems in 2016 and 2017. The following information is available in the metadata and measurement data[1]:

- power: Number of panels (N_p), panel power (P_p), total system size (P), inverter size (P_i) and number of inverters (N_i).
- geometry: azimuth (ϕ) and tilt (θ).
- brand: panel and inverter brands.
- location: Postal code 4 (pc4) area[2], longitude (l) and latitude (b).
- time: installation date of the system (d).
- comments: explanatory comments can be added by the user.
- energy: instantaneous (Y_{inst}) and cumulative energy measurements (Y_{cum})
- date and time of energy measurements.

2.2. Dutch meteorological weather data

The *Koninklijk Nederlands Meteorologisch Instituut* or Royal Netherlands Meteorological Institute (hereafter KNMI) is the body responsible for measuring variables related to the weather in the Netherlands (KNMI, 2020). Besides the 30 different ground–based weather stations, which measure various different meteorological variables, the institute has also developed a physics based empirical model to calculate variables, such as irradiance amongst others, for large parts of Africa and Europe (Deneke et al., 2008; Greuell et al., 2013).[3] This model uses input data from the *Spinning Enhanced Visible and Infrared Imager* (SEVIRI) instrument (Schmid, 2019) on board the *Meteosat Second Generation* satellites, located in a geostationary orbit at 36,000 km. SEVIRI observes properties of the atmosphere every 15 min and has a resolution of 3x3 km^2 at nadir (KNMI, 2019). Due to projection effects, this results in a resolution of 3x6 km^2 for the Netherlands. It is important to note that the irradiance data are only available at times when the Sun's elevation exceeds 12° (Greuell et al., 2013). The modelling does not work well outside this regime due to 3D cloud effects. These data contain the following variables:

- irradiance (G) in units of kW/m^2
- grid cell centre l and b

Fig. 1 shows H_d for two consecutive days in June 2016, nicely illustrating the effect of different weather conditions.

2.3. PV Systems Database

Statistics Netherlands constructs its own database of PV systems based on different data sources. The biggest and most important of these is the *Product Installation Register* or PIR (SN, 2019), provided by the network operators. This data source is incomplete and is estimated to contain around $\sim 85\%$ of small household systems. This is why, in recent years, additional systems, not present in the PIR, have been identified using data from the Dutch tax authority. People are incentivised to purchase PV systems by registering a VAT return for the cost and installation of the panels. Although this is not obligatory, it is assumed that this can entice a lot of people to do so. Even after adding tax returns the register is still incomplete. This data source contains the following variables:

- power: total system size (P).
- time: installation date (d).
- location: house number and postal code.

[1] For more details regarding these data, please see PVOutput (2020)

[2] Postal codes in the Netherlands consist of four digits (pc4), followed by two letters (pc6).

[3] Please consult https://msgcpp.knmi.nl/ for the interactive viewer.

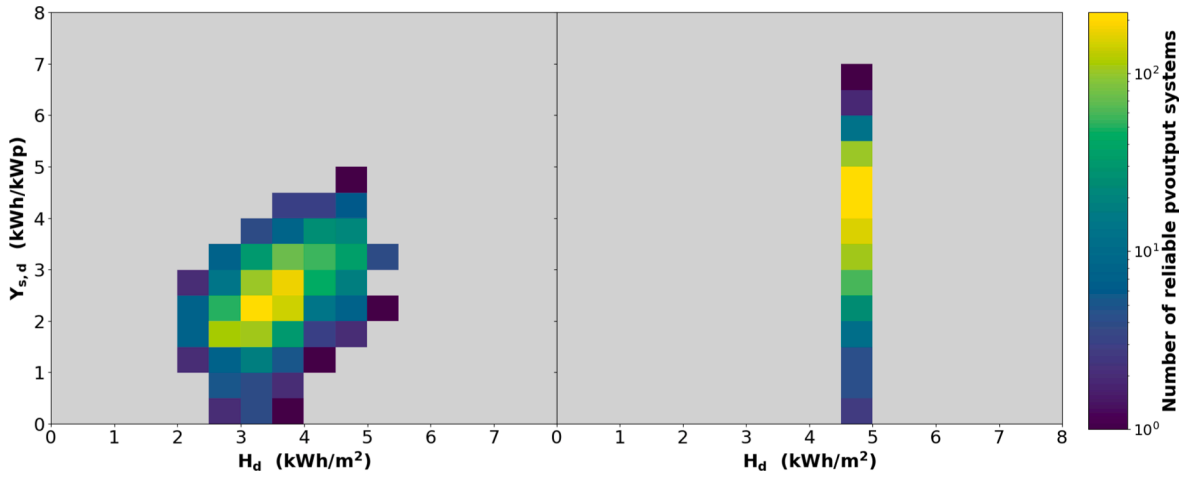

Fig. 2. H_d vs. $Y_{s,d}$ for all PVOutput systems on 13/06/2016 (left) and 13/09/2016 (right).

As mentioned in the Introduction, Statistics Netherlands has administrative data on P and monthly energy production (Y_m) from large PV systems (e.g. solar farms) from the government–led certification process for renewable energy as executed by CertiQ (CertiQ, 2020), a 100% subsidiary of TenneT, the European electricity transmission system operator for the Netherlands (TenneT, 2019). We do not use the CertiQ measurement data because our model is aimed at computing Y_d (rather than Y_m), as a function of different PV system parameters such as ϕ, θ and ϵ, all of which we also do not know for the CertiQ systems.

3. Methods

We outline our new method for determining Y_d and Y_a for the Netherlands in Section 3.1. An overview, in formal notation, is presented in Sections 3.1.1, with the different aspects of the method expanded upon in Sections 3.1.2. In Section 3.2, we describe our method to convert the national estimates into regional estimates.

3.1. National solar energy production

3.1.1. Procedure

Our main aim is to calculate $p_d(Y)$: the distribution of Y_d emanating from the population of PV systems in the Netherlands, which we can aggregate to Y_a. The relationship between these three quantities are defined by Eqs. 2 and 3:

$$Y_d = \int p_d(Y) dY, \tag{2}$$

$$Y_a = \sum_{d=1}^{d=365} Y_d. \tag{3}$$

Eq. (3) may be re–expressed as a function of the weather and more specifically H, which is denoted by Eq. (4). De–constructing $p_d(Y,H)$[4] into two separate functions $p_d(Y|H)$ and $p_d(H)$ and integrating this over the number of PV systems in the database (N_d) gives Eq. (5):

$$p_d(Y) = \int p_d(Y,H) dH, \tag{4}$$

$$p_d(Y) = \int^{N_d} p_d(Y|H) p_d(H) dH = \int^{N_d} p_d(Y_s|H) p_d(P) p_d(H) dH, \tag{5}$$

where in the final step, we have re–expressed $p_d(Y)$ in terms of $p_d(Y_s)$ and $p_d(P)$, the specific yield and power of the systems respectively. Evaluating $p_d(H)$ is trivial: the distribution of H of all database locations is obtained by matching each database system to its nearest irradiance grid cell. Evaluating $p_d(Y_s|H)$ is more complex since we construct this with our non–probability sample `PVOutput`, which we use as a proxy for the population. We can evaluate $p_d(Y_s|H)$ as the marginal likelihood, in terms of the PV system characteristics $x = \{\phi, \theta, \epsilon\}$, given by Eq. (6):

$$p_d(Y_s|H) = \int p_d(Y_s|x) p_d(x|H) dx. \tag{6}$$

By selecting a certain prior $p_d(x|H)$, we obtain a realisation of the `PVOutput` data: $p_d(Y_s|H)$, satisfying the aforementioned prior. Since for an observed value of H, there is a corresponding spread in Y_s, there is no way of knowing which Y_s to allocate to a database location. We therefore randomly draw all the systems in the database and assign them a value for Y_s, such that the ensemble of Y_s values assigned to the database locations, respects the proportions observed in `PVOutput`. This procedure of *Monte Carlo sampling* may be repeated a number of times, allowing us to repeatedly evaluate Eq. (5) and, in turn, Eq. (2). The result of this is an estimate of the mean energy production (μ_{Y_d}) and standard deviation (σ_{Y_d}).

Our calculations for Eq. (5) will strongly depend on the choices we make for the prior, relating to the distribution of $x = \{\phi, \theta, \epsilon\}$, in Eq. (6). By exploring various scenarios, i.e. different choices of our prior, we can explore the margins of our estimates. We will further expand these ideas and equations in Sections 3.1.2 and 4.

3.1.2. Evaluating Y_d with Monte Carlo sampling

We have seen in Section 3.1.1 that our method relies on evaluating Eq. (5). Before being able to do so, it is necessary to construct the functions $p_d(Y_s|H)$ and $p_d(H)$. Both of these are obtained by linking either `PVOutput` locations, in the case of $p_d(Y_s|H)$, or database PV systems, in the case of $p_d(H)$, to H_d grid cells, following a simple procedure set out in Appendix B. Once linked, $p_d(Y_s|H)$ may be easily constructed according to the following procedure: for any given day d and `PVOutput` location i, we can use the cumulative energy measurement Y_{cum} (or Y_d in other words) to obtain $Y_{s,d}$, according to: $Y_{s,d\ i} = Y_{d,i}/P_i$. The system sizes are those quoted in the `PVOutput` metadata. We refer the reader to Appendix A for a short explanation on how we compute H_d from the quarter hourly G values. Fig. 2 shows two examples

[4] Please note that when we write $p_d(Y,H)$ or any other example with p_d, the subscript is also implied for the quantities in the function i.e. Y_d and H_d. We drop these subscripts to avoid cluttering.

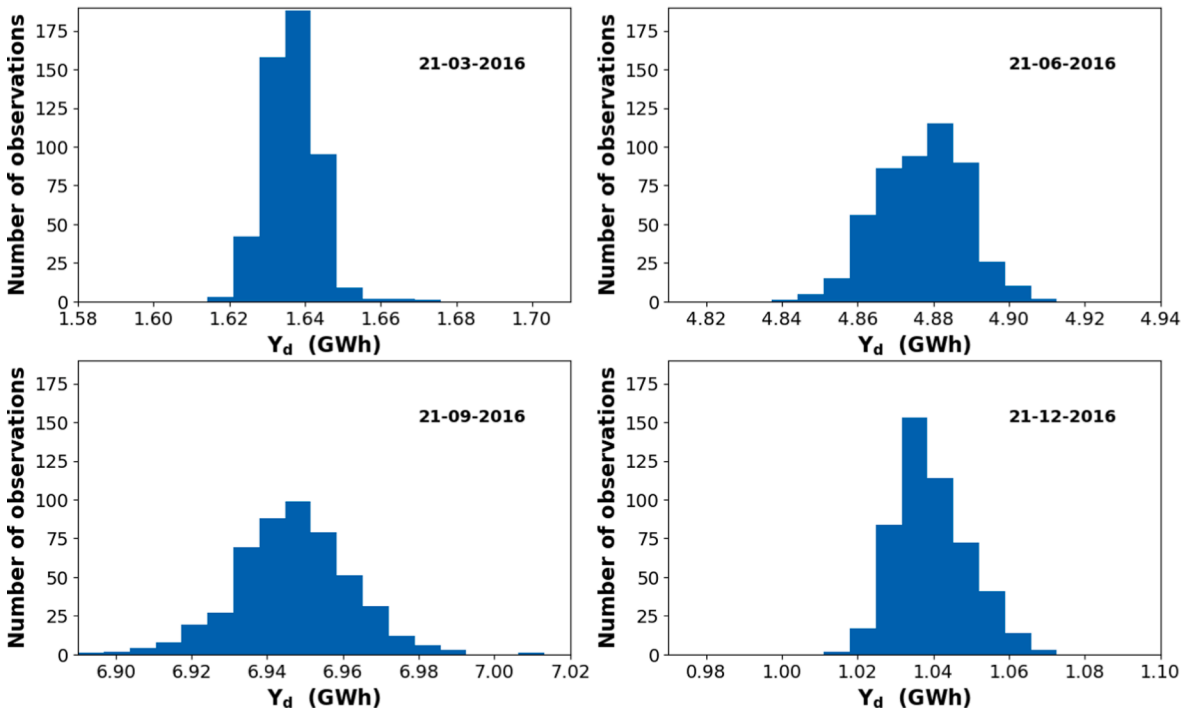

Fig. 3. Distributions of Y_d for all database PV systems (according to scenario 1) on the Summer and Winter solstices and the Spring and Autumn equinoxes. These energy production distributions are calculated using Monte Carlo sampling, using Eqs. (7)–(10).

of the number density of reliable[5] `PVOutput` systems in the $H_d - Y_{s,d}$ plane. The left panel (13/06/2016) shows a day with large variations in H_d and $Y_{s,d}$, whereas the right panel (13/09/2016) shows an exceptionally clear day over the whole country, nevertheless producing a wide spread in $Y_{s,d}$, due to different efficiencies of `PVOutput` systems which are a function of parameters such as ϕ and θ.

With all the necessary elements in hand, we evaluate Eq. (5) by re–writing the integral as a sum. This is more intuitive in light of the graphical representation of $p_d(Y_s|H)$ we discussed in Fig. 2. We can define that for a number of irradiation (N_{H_d}) and energy production (N_{Y_d}) bins, the probabilities of $Y_{s,d}$ being observed must sum to one (Eq. (7)). Using $p_d(H)$, we compute the number of systems N_l per bin l, which must satisfy Eq. (8), where N_d is the number of systems in the database. We can use this information to compute the number of systems per bin N_{kl} (Eq. (9)). Keeping track of which systems fall in bin l, we can perform a random sample of N_{kl} systems from a total of N_l systems for bin k, l and insert this into Eq. (10), where P_m is the system power of one system in that N_{kl} sample of systems.

$$\sum_{k=1}^{N_{Y_d}} \sum_{l=1}^{N_{H_d}} p_{kl}(Y_{s,d}) = 1, \tag{7}$$

$$N_d = \sum_{l=1}^{N_{H_d}} N_l, \tag{8}$$

$$N_{kl} = \frac{p_{kl}}{\sum_{k=1}^{N_{Y_d}} p_{kl}} \frac{N_l}{N_d}, \tag{9}$$

$$Y_d = \sum_{k=1}^{N_{Y_d}} \sum_{l=1}^{N_{H_d}} N_{kl} Y_{s\ kl} \sum_{m=1}^{N_{kl}} P_m. \tag{10}$$

[5] For an explanation on what reliable means, we refer the reader to Appendix A for further reading regarding data cleaning.

We can perform *Monte Carlo sampling* by repeatedly evaluating Eq. (10), resulting in a probability density function, indicating the mean (μ_{Y_d}) and standard deviation (σ_{Y_d}) of our estimates for Y_d. Fig. 3 shows what these functions look like for four different days (Spring and Autumn equinoxes, Summer and Winter solstices). These functions were constructed by performing the *Monte Carlo sampling* 500 different times, allowing for uncertainty margins to be quantified.

Our decision to bin the data in 0.5kWh/m^{2}x0.5kWh/kW$_p$ bins, as can be seen in Fig. 2 is motivated by practical concerns. We want to produce a simple, easy and intuitive model allowing us to easily read off $p_{kl}(Y_s)$. The size of our bins is chosen in such a way that the resolution is high enough such that meaningful differences in Y_s may be discerned, while at the same time keeping the resolution low enough, increasing the chances that each value of H at a location in the database is also observed in $p_d(Y_s|H)$ from `PVOutput`. For those systems which have a value H that is not observed in $p_d(Y_s|H)$, we use our estimate of $Y_{s,d}$ for all the present systems and multiply this by the systems' P. We note that the fraction of such systems is always lower than 1%.

3.2. Regional solar energy production

3.2.1. Procedure

Estimating regional solar energy production is a lot less work now that we have set out the framework in Section 3.1. Up until now we grouped database systems together in 0.5kWh/m^2 bin sizes such that we could construct $p_d(Y_s|H)$ and hence estimate Y_d and Y_a (see Fig. 2). This of course means that many systems were placed together in bins and assigned the same value Y_s, even though their observed H could differ by as much as the size of the bin, i.e. 0.5kWh/m^2. By returning to each database location, we can re–determine Y_d using the exact measurement of H_d.

First we convert μ_{Y_d} into $\mu_{Y_{s,d}}$, using the total system size of the PV systems population in the database on that day. Then, we make the assumption that the mean specific yield $\mu_{Y_{s,d}}$ must correspond to the mean irradiation (μ_{H_d}) observed on that day in the Netherlands. Therefore, for a system at a location j, we can read off once again $H_{d,j}$ and

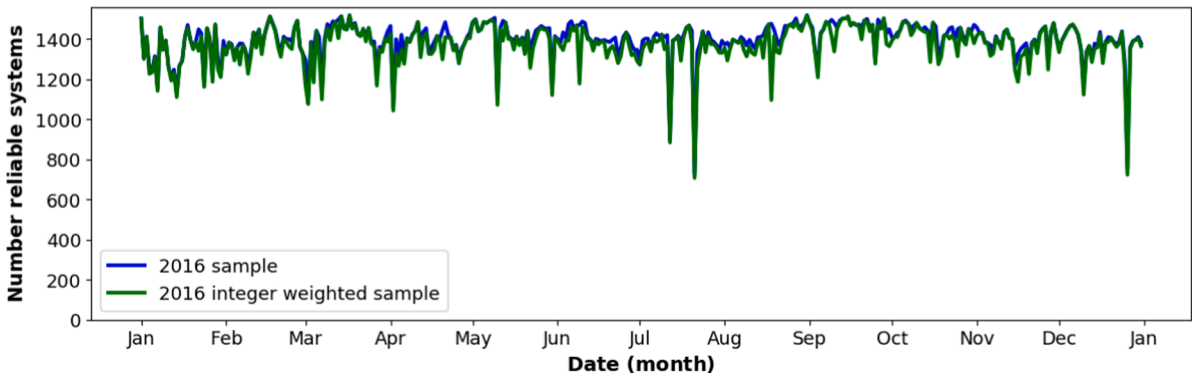

Fig. 4. Timeseries of the number of reliable PV systems (after data cleaning is performed) for 2016 (blue), along with an integer weighted set (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

calculate its offset relative to μ_{H_d} and offset $Y_{s,d\ j}$ accordingly (Eq. (11)). It is then possible to aggregate each system to a regional level of our choosing such as a municipality, given by Eq. (12). We discuss our regional results in Section 5.2.

$$Y_{s,d\ j} = \mu_{Y_{s,d}} + \mu_{Y_{s,d}} \left(\frac{H_{d,j}}{\mu_{H_d}} - 1 \right), \tag{11}$$

$$Y_{d,mun} = \sum_{j=1}^{N_{d,mun}} Y_{s,d\ j} P_j. \tag{12}$$

4. Representativeness

From the description of our method so far, it is clear that any results we compute for Y_d and Y_a will be strongly driven by the composition of the `PVOutput` sample on any given day i.e. $p_d(Y_s|H)$. For example: in appendix A we described that our daily `PVOutput` samples fluctuate daily, i.e. each day does not contain the same number or type of systems (see blue line in Fig. 4). This is due to the data cleaning process: very few systems reliably deliver data 100% of the time. We are therefore estimating Y_d on the basis of populations that are not directly comparable. Were we to construct a method that could make these samples very similar over all the days of the year, we would still be faced with another issue, namely deciding what the true population of PV systems looks like and adapting the `PVOutput` samples accordingly. We already briefly explored this in Section 3.1.1, when talking about our decision for the prior $p_d(x|H)$.

In this section, we therefore propose two refinements to our method allowing us to address these challenges. In Section 4.2 we outline a sensitivity analysis: what is the effect on the computation of Y_d and Y_a when selecting different priors (or subpopulations)? Before doing so, we first define a method in Section 4.1 which allows us to re–weight $p_d(Y_s|H)$ such that the same populations are contained within each $p_d(Y_s|H)$ of the entire year. To make any progress on either of these refinements, we need to determine which variables characterise the population. In Section 1.1 we already highlighted that $x = \{\phi, \theta, \epsilon\}$ and (l, b) of the PV systems will influence the determination of Y_d and Y_a.

4.1. Re–sampling PVOutput

New PV systems are continually placed throughout the year, with a peak in the Spring and Summer months(SN, 2019). This means, at least in theory, that the distributions of system variables ϕ, θ and ϵ can vary as a function of time. In the absence of these distributions, we argue that they can be assumed to remain constant in time. This we motivate through large number statistics: if a large population already exists with some distribution $p(x)$, and a small – relative to the total already present – new number of systems is continually placed, then it seems probable that these effects will average or smooth out. An obvious exception to this would be a large solar park which started generating energy from one day to the next and had a set up capable of heavily skewing $p(x)$. While our assumption would seem to hold over the course of different days, it is less obvious whether this should be the case over several years. Past research has also investigated the issue of representativeness e.g. Killinger et al. (2018) derive probability distribution functions for ϕ, θ, capacity and Y. Unfortunately, in the case of the Netherlands, this is based largely on `PVOutput` data (75% of PV systems) supplemented with other smaller data sources.

4.1.1. Re–sampling ϕ, θ and ϵ

Making days consistent with each other, in terms of x can be achieved by choosing a ground truth for $p(x)$ and adjusting $p_d(x)$ accordingly. Rather than inventing $p(x)$, we can choose it to be $p_1(x)$: the distribution on the first day of the year:[6]

$$p_d(\phi, \theta, \epsilon) \sim \mathrm{p}_1(\phi, \theta, \epsilon), \quad \text{where} \quad 1 < \mathrm{d} \leqslant 365. \tag{13}$$

Integer weighting can be used to satisfy Eq. (13): certain systems are randomly selected more than once, while others may be dropped. This process is repeated until Eq. (13) is satisfied. Our choice for integer weighting, as opposed to floating point weighting, is motivated by the fact that PV systems are discrete quantities. If we adopted floating point weights, this would produce odd situations whereby systems can be partially duplicated or discarded. To satisfy Eq. (13), we can bin a `PVOutput` variable x according to Eq. (16). Then, summing all these bins for variable x, we satisfy Eq. 17, which must always equal one. Finally x_i and x_{i+1} in Eq. (16) are defined by Eq. 15, which does nothing more than defining the lower and upper limits of bin i for a given bin size Δx. Finally, the number of bins is defined by Eq. (14).

$$N_{bins} = \frac{x_{max} - x_{min}}{\Delta x}, \tag{14}$$

$$x_i = \sum_{i=1}^{N_{bins}} x_{min} + i\Delta x, \tag{15}$$

$$p_i(x) = \frac{1}{N} \sum_{n=1}^{N} \begin{cases} x_i < x_n < x_{i+1} & 1 \\ \text{else} & 0, \end{cases} \tag{16}$$

$$p_d(x) = \sum_{i=1}^{N_{bins}} p_i(x) = 1. \tag{17}$$

[6] In the case of 2016 this is in fact 366 since this is a leap year. This also applies to Eq. (18)

Table 1
Minima, maxima and bin sizes for the four different parameters: ϕ, θ, ϵ and H.

parameter	min	max	Δ parameter
ϕ	0°	360°	45°
θ	0°	90°	15°
ϵ	−1	1	1
H	min(H)	max(H)	0.5kWh/m^2

4.1.2. Re–sampling H

Now we must account for one final variable: $p_d(l,b)$. With the aim of making $p_d(Y_s|H)$ as accurate as possible, the geographic number density of systems in PVOutput should match the density observed in the database. If, for example, 40% of the PVOutput systems, used to construct $p_d(Y_s|H)$, lie in the West of the country on day d, while in the database this is 20%, then our estimation of Y_d could end up being too optimistic or pessimistic depending on the weather on that day.

In practice, it's very difficult to apply integer weighting to $p_d(l,b)$, since one would have to agree on bin sizes for (l,b). Such aggregation bin sizes would have to change daily depending on the weather (or H): on a perfectly sunny day over the whole country (e.g. 13 September in Fig. 2), it could make sense to define one bin for (l,b) encompassing the whole country, whereas on a day with a lot of local weather effects, a different aggregation level would be necessary. Since the sample size of PVOutput is too small in any case to split it up into smaller portions, we can use a proxy for (l, b), which is H itself. We can apply integer weighting to H observed in PVOutput such that its distribution satisfies the distribution in the database D, given by Eq. (18)):

$$p_d(H) \sim D_d(H), \quad \text{where} \quad 1 \leqslant d \leqslant 365. \tag{18}$$

4.1.3. Integer Weighting: Discussion

Table 1 shows the minima, maxima and bin sizes for ϕ, θ, ϵ and H. It should be noted that bins for ϵ do not correspond to anything physically: 0 means $\epsilon = 1$, -1 means $\epsilon > 1$ and 1 that $\epsilon < 1$. Our choice for the other bin sizes is motivated by practical limitations: $\Delta\phi = 45°$ because PVOutput only allows the input of four cardinal and four intercardinal directions. We choose $\Delta\theta = 15°$ such that we retain a statistically significant number ($\sim 100-200$) of PV systems per bin. We choose $\Delta H = 0.5\text{kWh/m}^2$ since this is what we already decided earlier on in Section 3.4 when combining H_d and Y_d (see Fig. 2).

We draw the reader's attention to the fact that our earlier decision for integer weighting means we cannot exactly satisfy Eqs. 13 and 18. This is why we allow for a leeway of 1.5% when trying to satisfy these equations. The choice for this number is a pragmatic one: it is lenient enough to allow us to efficiently implement our procedure, but strict enough that the equations are almost exactly satisfied. Fig. 4 shows the effect of integer weighting: On some days, the overall number of systems increases while on most days the set decreases. Fig. 5 shows $p_{1/6/16}(\phi, \theta, \epsilon)$ and $p_{1/6/16}(H)$, which have been re–sampled so they satisfy Eqs. 13 and 18.

As stated above, the procedure of integer weighting has the consequence that the two constraints in practice will never both be satisfied exactly. By making these constraints 'softer', i.e. allowing an interval around an exact match, they become probabilistic in nature, so that it is advisable to generate multiple realisations of the distributions, all within that small allowed interval. Given that each realisation is itself a sample of between 800 and 1400 instances, a modest number of 50 realisations of the distributions is sufficient to ensure that an average over that ensemble of distributions can be used for this analysis.

4.2. Choosing different priors

We now return to our second refinement. In Eq. (6) of Section 3.1.1, we saw that it is possible to evaluate $p(Y_s|H)$ as the marginal likelihood, with a prior $p(x|H)$. We can now decide to make different selections for $p(x|H)$ and propagate these through in our calculations. For example, what will Y_d and Y_a be if only all South–facing systems are selected? Experimenting with different choices of $p(x|H)$ will give us a sense of how much our current estimates at SN could vary depending on what the true specifics are of the Dutch PV system population.

4.3. Schematic summary of method

We summarise our methodological framework from Sections 3 and 4 for the reader:

1. Make a choice for the prior $p(x|H)$, e.g. all systems face South.
2. Select all systems in PVOutput on 1 January that satisfy $p(x|H)$.

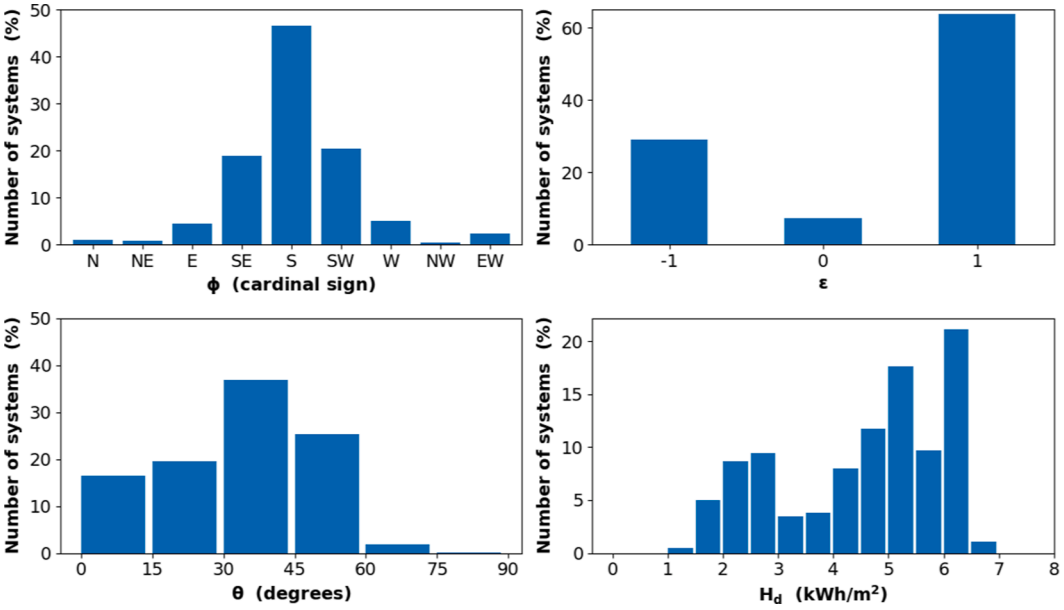

Fig. 5. $p_{1/6/16}(\phi, \theta, \epsilon)$ and $p_{1/6/16}(H)$, integer weighted w.r.t. $p_{1/1/16}(\phi, \theta, \epsilon)$ and $p_{1/1/16}(H)$, according to scenario 1. Integer weighting is performed by satisfying Eq. (13) (ϕ, θ and N_p) and Eq. (18) (H).

Table 2
$Y_{s,2016}$, Y_{2016}, $Y_{s,2017}$ and Y_{2017} (with 1σ uncertainty margins) for 15 different scenarios.

Sc	ϕ	θ	ϵ	$Y_{s,2016}$ (kWh/kW$_p$)	$Y_{s,2017}$ (kWh/kW$_p$)	Y_{2016} (GWh)	Y_{2017} (GWh)
1	$\forall$	$\forall$	$\forall$	910 ± 0.14	868 ± 0.19	1632 ± 0.26	2131 ± 0.49
2	$= 180°$	$\forall$	$\forall$	946 ± 0.16	899 ± 0.21	1697 ± 0.30	2209 ± 0.52
3	$\{135°..225°\}$	$\forall$	$\forall$	927 ± 0.14	882 ± 0.20	1663 ± 0.26	2168 ± 0.49
4	$\{90°..180°\}$	$\forall$	$\forall$	929 ± 0.14	885 ± 0.20	1668 ± 0.26	2176 ± 0.49
5	$\{180°..270°\}$	$\forall$	$\forall$	921 ± 0.15	877 ± 0.19	1652 ± 0.28	2154 ± 0.48
6	$\neq 180°$	$\forall$	$\forall$	877 ± 0.16	838 ± 0.21	1573 ± 0.29	2059 ± 0.51
7	$\forall$	$\{0°..30°\}$	$\forall$	895 ± 0.18	860 ± 0.22	1605 ± 0.32	2113 ± 0.55
8	$\forall$	$\{30°..90°\}$	$\forall$	920 ± 0.14	872 ± 0.20	1652 ± 0.26	2143 ± 0.50
9	$\forall$	$\{30°..45°\}$	$\forall$	923 ± 0.16	875 ± 0.21	1656 ± 0.29	2150 ± 0.51
10	$\forall$	$\forall$	$\{1\}$	903 ± 0.15	860 ± 0.21	1619 ± 0.27	2114 ± 0.52
11	$\forall$	$\forall$	$\{-1\}$	921 ± 0.20	875 ± 0.25	1652 ± 0.36	2150 ± 0.62
12	$\forall$	$\forall$	$\{0,1\}$	905 ± 0.14	864 ± 0.20	1624 ± 0.26	2122 ± 0.48
13	$\forall$	$\forall$	$\{-1,0\}$	922 ± 0.18	879 ± 0.24	1654 ± 0.33	2158 ± 0.59
14	$\{135°..225°\}$	$\{30°..45°\}$	$\{1\}$	938 ± 0.17	889 ± 0.22	1684 ± 0.31	2183 ± 0.54
15	$\{135°..225°\}$	$\{30°..45°\}$	$\forall$	943 ± 0.15	893 ± 0.21	1695 ± 0.28	2195 ± 0.51

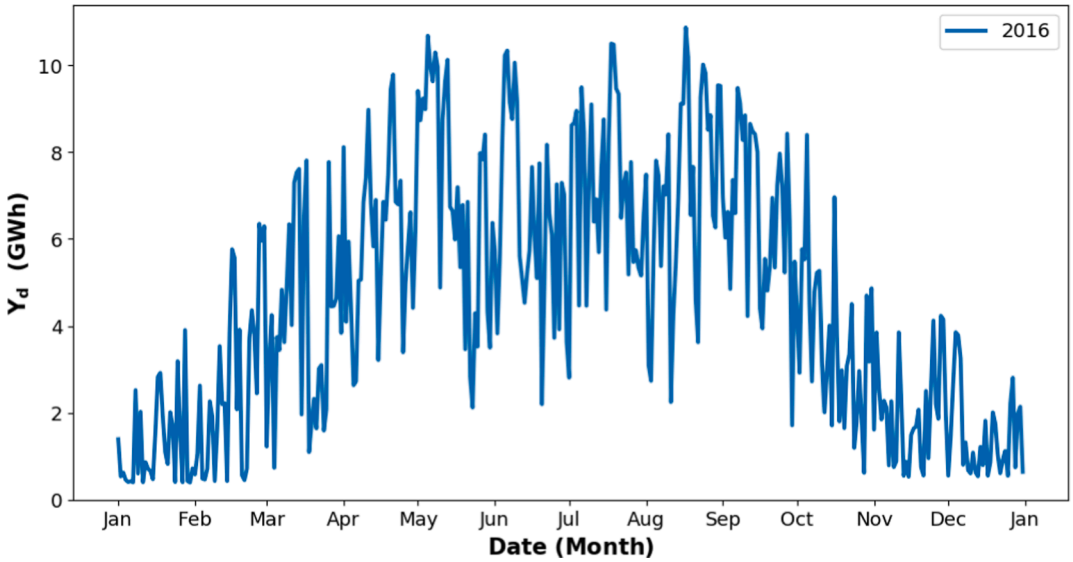

Fig. 6. Y_d in 2016, according to scenario 2.

3. Re–sample such that $p_d(\phi,\theta,\epsilon) \sim \mathrm{p_l}(\phi,\theta,\epsilon)$ for day $1 < d \leqslant 365$ and $p_d(H) \sim D_d(H)$ for day $1 \leqslant d \leqslant 365$.
4. Repeat step 3, 50 times.
5. Randomly select one of the 50 'realisations' of $p_d(Y_s|x)$ and insert into $p_d(Y_s|H)$.
6. Calculate $p_d(Y)$.
7. Repeat the last two steps 500 times.
8. Estimate μ_{Y_d} and σ_{Y_d}.
9. Aggregate to time frame of choice: daily or annually.
10. Convert national estimates to regional estimates, for any given choice of $p(x|H)$, by comparing local H to μ_{H_d} and offsetting local $Y_{s,d}$ accordingly.
11. Aggregate to spatial frame of choice: municipality.

5. Results and Discussion

5.1. Daily and annual national energy production

We present the results of the various different 'scenarios' – or different choices for our prior: $p(x|H)$ – in Table 2. The choice for these specific scenarios is driven by their feasibility: `PVOutput` samples are not very large and so we are limited to those scenarios which retain enough systems. For example, determining Y_d and Y_a by selecting only North–facing systems ($\phi = 0°$) would only leave a handful of systems. With the exception of scenario 1, which takes $p(\phi,\theta,\epsilon)$ as a given, the other scenarios explore different choices for ϕ (scenarios 2–6), θ (scenarios 7–9) and ϵ (scenarios 10–13). Scenarios 14 and 15 explore combinations of all three different parameters.

Scenario 2, closely followed by 15, shows the slargest $Y_{s,a}$, which is consistent with the expectation that South–facing systems outperform other set–ups. Restricting θ to a more optimal angle range (scenario 15) at the expense of relaxing ϕ, gives similar results though. Scenario 15 outperforms scenario 14 by a small margin, suggesting that a smaller P_i could be beneficial. Scenario 6 delivers the poorest Y_a, which is unsurprising, given that all South–facing systems are excluded. A final noteworthy mention is scenario 7: restricting PV systems to near–flat or fully flat systems, produced the second lowest Y_a, presumably due to low solar elevation angles in Winter. The Y_a of the other scenarios lie in between these extrema. Fig. 6 shows Y_d for 2016 according to the best performing scenario 2. This visualisation nicely demonstrates what the differences are in Y_d in Summer compared to Winter. It is worth noting that a high–irradiation Winter's day delivers relatively high Y_d

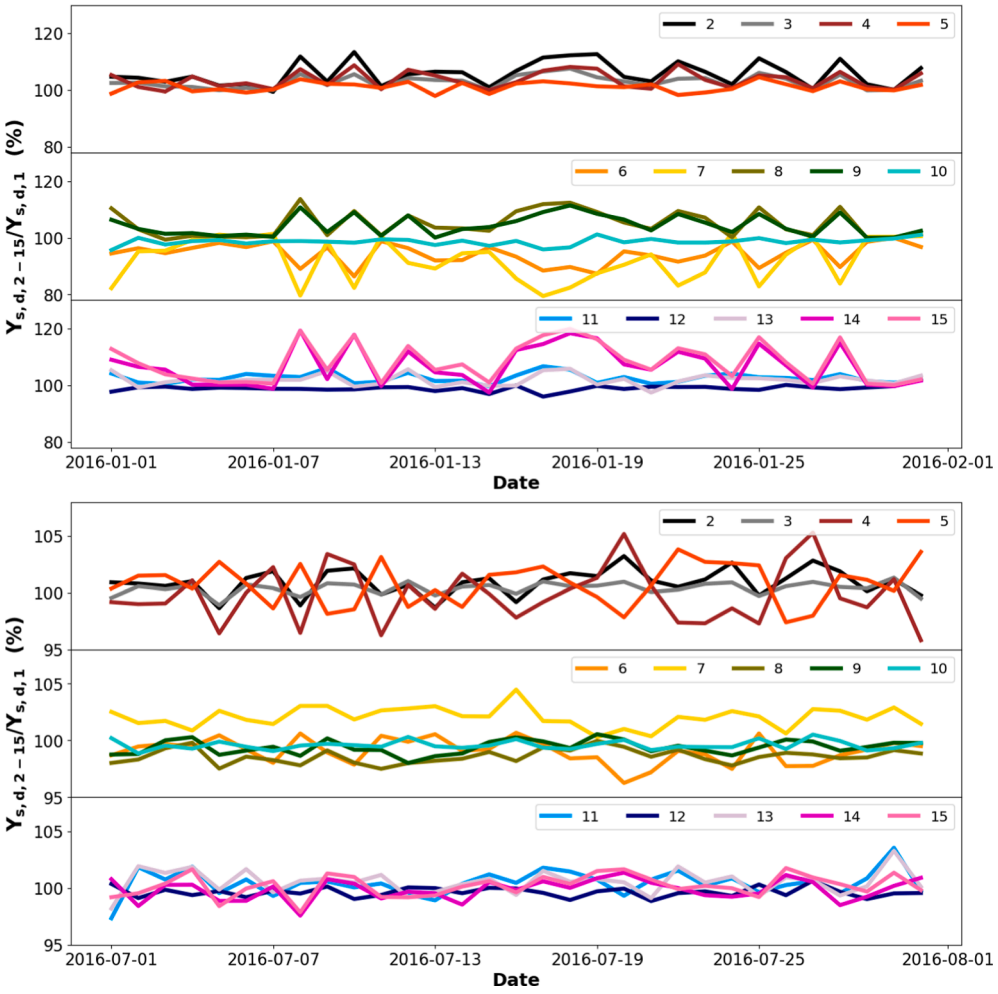

Fig. 7. The ratio of $Y_{s,d,2-15}/Y_{s,d,1}$ for January (top) and July (bottom).

Table 3
Comparison of monthly to annual energy production ratios (Y_m/Y_a) in 2016 for scenario 1 and the large PV systems in Certiq.

2016	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
scenario 1	2.5	4.9	8.1	11.6	13.9	11.9	13.2	12.5	10.6	5.8	2.9	2.2
CertiQ	2.5	4.7	8.0	11.5	14.2	12.5	13.5	12.4	10.1	5.6	2.9	2.2

Table 4
$Y_{s,a}$ according to SolarCare (Solar Magazine, 2020), SN and our research.

Year	SolarCare (kWh/kW$_p$)	SN (kWh/kW$_p$)	This research (kWh/kW$_p$)
2012	900	875	n.a.
2013	890	875	n.a.
2016	920	875	910
2017	880	875	868

especially when comparing to a low–irradiation Summer's day.

Fig. 7 shows Y_d, during the months of January and July, for scenarios 2–15, normalised w.r.t. Y_d from scenario 1. The y–axis is thus an index where a number higher than 100, means that particular scenario led to a higher Y_d than scenario 1 and vice versa. We include this figure because it says something about the validity of our model. We highlight one of many interesting patterns from this figure to illustrate our point. Scenario 7 can be contrasted with the other scenarios, for the two months in questions. It can be seen that this set–up leads to lower and higher Y_d in January and July respectively, which is wholly consistent with the fact that these are low tilt systems. The less than optimal θ of these systems acts differently in the Summer with the negative effects in the Winter mitigated by a bonus in the Summer. We draw the reader's attention to the fact that for some scenarios, one can indirectly deduce the weather on that day: when contrasting scenarios 2, 3 and 4 with each other in July, one can see whether the weather was better in the morning, in the afternoon or the same. These observations support the soundness of the presented model.

The uncertainty margins $\sigma_{Y_{s,a}}$ and σ_{Y_a} are on the order of 0.05%. We see four possible explanations for these small margins: the `PVOutput` sample is quite small and therefore the variation between the 50 'realisations' of the data for any choice of $p(x|H)$ is small. Secondly, the large size of the $H_d - Y_d$ bins (0.5kWh/m^{2}x0.5kWh/kW$_p$) when evaluating $p_d(Y|H)$, can cause a lot of smaller scale effects to smooth out. Future work could include examining by how much σ_{Y_a} and σ_{Y_d} change as a function of the bin resolution. Thirdly, differences on a micro level can average out due to the very large size of the PV systems database, which contains hundreds of thousands of entries. This is further compounded by the fact that differences in Y_d will average out anyway when aggregating to Y_a. This is confirmed by the observation that σ_{Y_d} is an order of magnitude higher than σ_{Y_a}. Finally, we would like to remind the reader

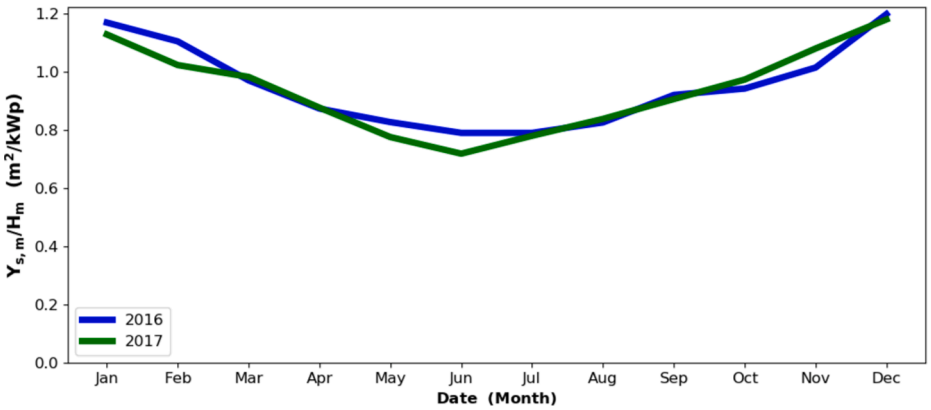

Fig. 8. The performance ratio ($Y_{s,m}/H_m$) for 2016 (blue) and 2017 (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

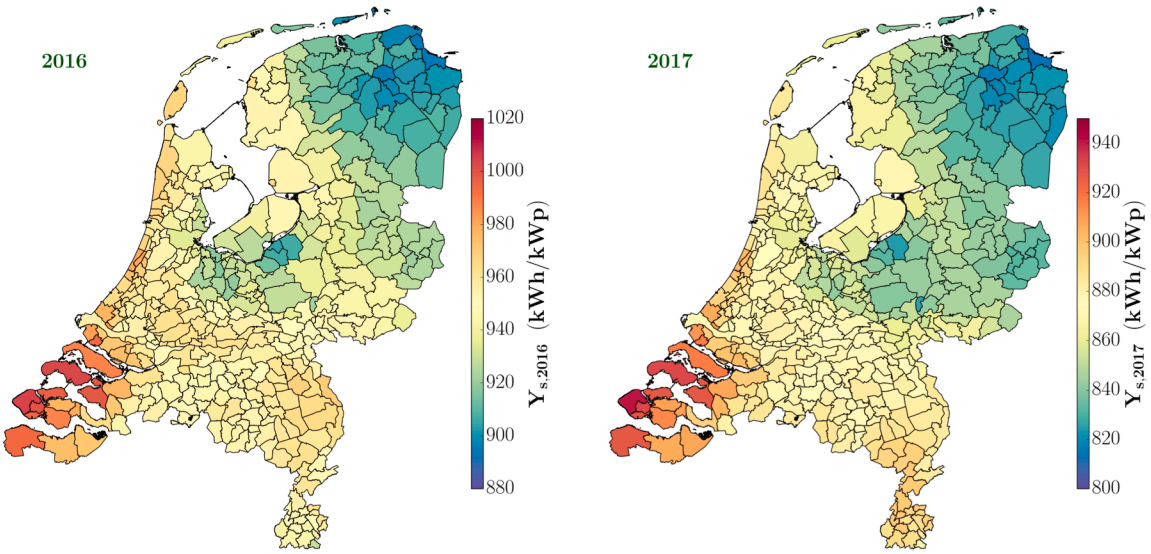

Fig. 9. $Y_{s,2016}$ (left) and $Y_{s,2017}$ (right) per Dutch municipality according to scenario 2.

that a bias is possible due to the choices we made when we cleaned the data.

Now we turn to the underlying assumptions that have been made throughout this paper w.r.t. *Monte Carlo sampling*. This type of sampling assumes that observations are independently and identically distributed (*i.i.d.*). It is possible that on the micro level *i.i.d.* is not fully satisfied due to correlations that can arise e.g. in a street with terraced houses facing the same way, it is probable that $Y_{s,d}$ is quite similar for all of the PV systems (assuming ϵ and other factors are negligible). Our method does not take these correlations into account. We would argue that the degree to which such correlations matter, depends on what aggregation levels one selects for Y_d. Since it is not our goal to present results for Y_d on a micro–level, we argue that this effect must average out on a national level, where hundreds of thousands of database PV systems are generating energy. In this regime, $p(Y_s|H)$ and $p(H)$ will be the far more dominant factors when determining μ_{Y_d}. If, for example, we are missing an important subpopulation of PV systems in $p(Y_s|H)$, this will have a rather larger impact. We briefly return to *i.i.d.* in Section 5.2, when discussing our regional results.

5.1.1. Comparison with Statistics Netherlands figures

Our results for 2016 are consistently higher than those currently measured by SN: $877 \leqslant Y_{s,2016} \leqslant 946 \text{kWh/kW}_\text{p}$ and $1605 \leqslant Y_{2016} \leqslant 1697$GWh. We remind the reader that these should be contrasted with SN's estimate of $Y_{s,2016} = 875 \text{kWh/kW}_\text{p}$ and $Y_{2016} = 1602$ GWh (SN, 2020). The result is even more significant, given that the lower range of $877 \text{kWh/kW}_\text{p}$ corresponds to an unrealistic set up: no South–facing panels (scenario 6). The picture is more mixed for 2017: $838 \leqslant Y_{s,2017} \leqslant 899 \text{kWh/kW}_\text{p}$ and $2059 \leqslant Y_{2016} \leqslant 2209$ GWh, with $Y_{s,a}$ lying somewhere within our estimated range.

5.1.2. Comparison with CertiQ and SolarCare

In Section 2.3 we mentioned that SN also has Y_m measurements for $\sim$ 1800 large PV systems (CertiQ). Unfortunately the specifics such as ϕ, θ and ϵ are unknown. Indeed this is the reason we chose not to include the data source in our method, preferring to use it as a form of validation. Table 3 shows our calculations for Y_m/Y_a for `PVOutput` (according to scenario 1) and CertiQ. While the Winter months seem to be spot on, there are small discrepancies for the Summer months, with `PVOutput` showing lower Y_m, where June and September have the most striking offset. We see two possible explanations for this. Firstly, the configuration of the large PV systems may be more optimal (e.g. ϕ), thus delivering better Y_m in the Summer months. Secondly, panel temperatures are likely to be higher on roofs than in fields, resulting in lower conversion efficiencies (typically 5–10% lower), and thus Y_m (Drews et al., 2007). Finally, we note that $Y_{s,2016} = 904 \text{kWh/kW}_\text{p}$, which we obtain from CertiQ data, is in close agreement with `PVOutput`'s scenario 1: $Y_{s,2016} = 910 \text{kWh/kW}_\text{p}$.

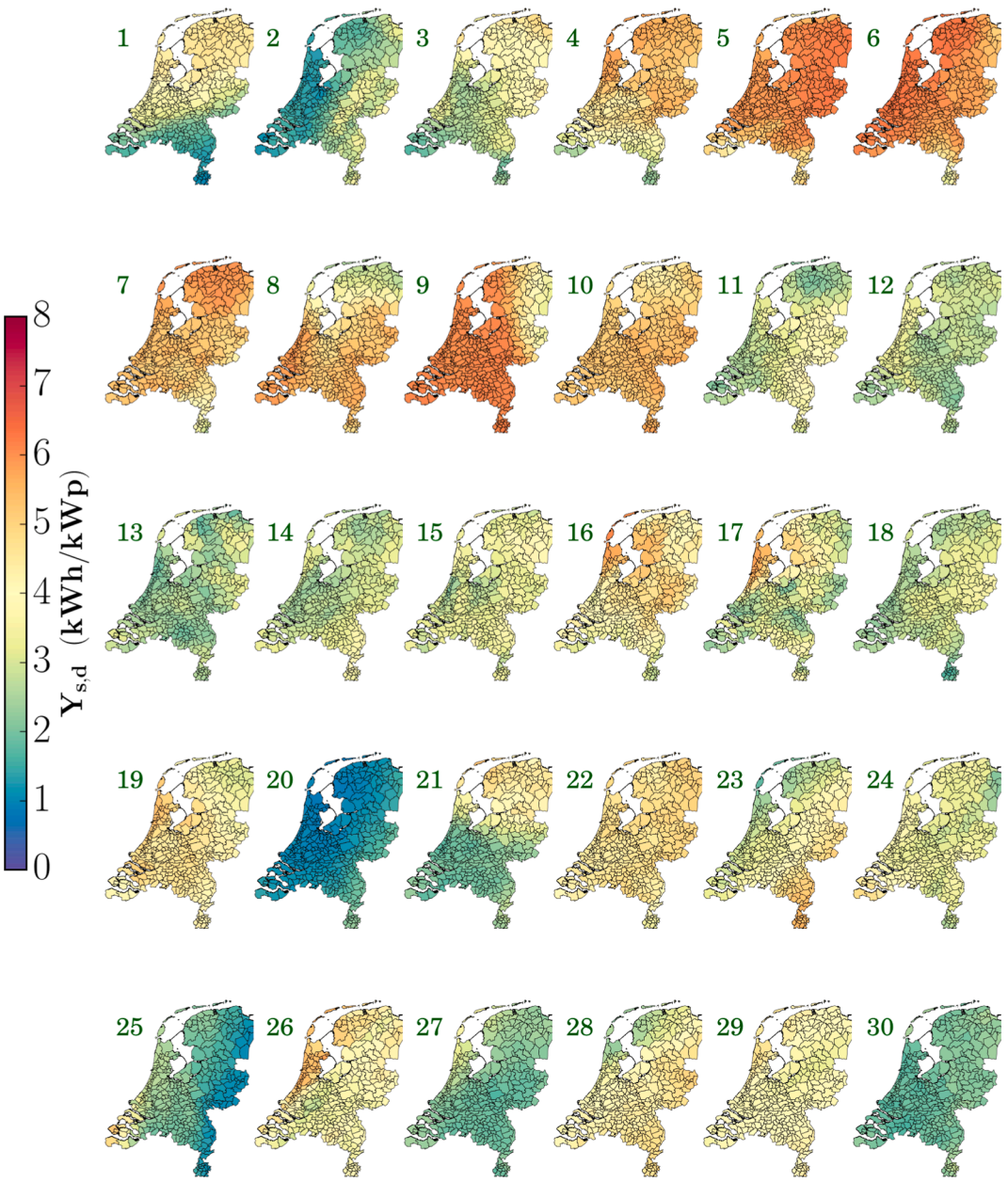

Fig. 10. $Y_{s,d}$ per Dutch municipality for each day in June 2016.

SolarCare (Solar Magazine, 2020) publish results for $Y_{s,a}$ and are summarised in Table 4. These figures are based on some 2500 PV systems across the whole country, making their estimates very robust and worthy of comparison. We notice a small offset for 2016 and 2017 of $\sim$ 10kWh/kW$_p$ between SolarCare and our results, with our results a little more conservative. Our results are therefore broadly in line with those recorded in CertiQ and SolarCare.

5.1.3. Comparison between 2016 and 2017

Our measurement of $Y_{s,2016}$ seems to be consistent with the – higher than 30 year average – measurement of $H_{2016} = 1039\text{kWh/m}^2$, as recorded by the KNMI's central weather station: De Bilt (please see Section 1.1 for more information regarding the 30 year average) (KNMI, 2017; KNMI, 2013; KNMI, 2014). This cannot be said for $Y_{s,2017}$, which, despite a higher than average irradiation (1020kWh/m^2) (KNMI, 2018), remains rather average when comparing to SN's $Y_s = 875\text{kWh/kW}_p$. Fig. 8 shows the monthly performance ratio $PR_m = Y_{s,m}/H_m$, using H_m as was measured at De Bilt, for 2016 (blue) and 2017 (green), according to scenario 1 (see e.g. Reich et al. (2012) for more information regarding the performance ratio). In this figure, it is striking that PR_{may} and PR_{jun} are lower in 2017 than 2016, suggesting the conversion of H to Y seems to have been less efficient. We note that June 2017, in particular, was an abnormally warm month with eight Summer days ($T_{max} > 25°C$) and two tropical days ($T_{max} > 30°C$). From the weather records, it appears that June 2017 was the warmest June month on record KNMI (2017). June 2016 contained five Summer days, consistent with the average number expected in June. From the literature it is known that panel efficiency drops with about 0.4% per °C. For a sunny day with high ambient temperature a panel on a roof can reach 80°C, thus leading to a relative reduction of 25% in efficiency (see e.g. Figure 5 in Drews et al. (2007)). We therefore hypothesise that a possible temperature effect was responsible for the decrease in Y_{2017}.

5.2. Daily and annual regional energy production

Fig. 9 shows $Y_{s,mun}$ for 2016 and 2017, as a result of calculating Eqs.

11 and 12. These figures validate our model because it shows patterns one expects: coastal regions enjoy more sunshine and therefore also higher $Y_{s,a}$ or Y_a. The lower $Y_{s,2017}$ compared to $Y_{s,2016}$ is also entirely consistent with other results (see Section 5.1). The overall pattern of $Y_{s,a}$ looks similar between 2016 and 2017. Finally Fig. 10 shows the $Y_{s,d,mun}$ for every day in June 2016, showing a large variation in the daily patterns.

While our method is constructed in such a way that we produce Y_d per database location, we would like to emphasise that this does not mean that we can accurately predict Y_d produced at any location in the Netherlands. Rather, the point of our method is that, when aggregated to sufficiently large enough areas, we expect $Y_{d,mun}$ on those levels to come close to reflecting reality, should the PV systems be installed according to the scenario set–ups as specified in Table 2. Even then, it should be noted that if the relative share of e.g. a large solar park is high compared to household PV systems in a given municipality, then this could influence the measurement of $Y_{d,mun}$. In light of the earlier discussion on *i.i.d.*, it should be noted that re–computing the energy production locally in this fashion, accounts to some degree for the correlations we mentioned earlier.

We remind the reader again of the fact that G is only provided when the solar elevation is higher than 12°. This means that the offsets that are calculated in Eq. (11) become more uncertain the closer we get to the Winter solstice: the Sun's highest elevation angle keeps reducing, meaning that larger portions of the day will not be captured in the totals of H_d. It is, for example, possible that $Y_{s,d\ j} \approx \mu_{Y_{s,d}}$, but that there was a lot more $H_{d,j}$ than average in the early morning and late afternoon, which would increase $Y_{d,j}$. This is not captured in our method. Future work could include supplementing these irradiance data with ERA5 data (Hersbach et al., 2019) for low elevations periods, thus obtaining complete irradiance measurements. Finally it should be noted that the calculation of Eq. (11) assumes $H_{d,j}$ acts on $Y_{d,j}$ in the same way, regardless of the location. This also need not be true, since other local weather factors, especially temperature, could have an effect on $Y_{d,j}$.

6. Conclusion and summary

We have presented a new method, in the form of a classical estimation problem, to determine the daily, annual, national and regional energy production generated by photovoltaic systems. This new method is validated by applying it to the situation of the Netherlands. We combined information from two different types of data sources. The first was a non–probability sample of solar energy production measurements in the form of citizen science on the online portal `PVOutput`. The second was high resolution irradiance data, derived from satellite based observations, obtainable from the Royal Netherlands Meteorological Institute. By matching these two data sources on their location, we generated daily probabilistic functions indicating the most likely specific daily yield, as observed in `PVOutput`, given the irradiation. These functions were applied to our database containing almost all PV systems in the Netherlands, producing daily and national energy production estimates, whose uncertainties were estimated through *Monte Carlo sampling*. The national figures were converted to regional estimates, by comparing the local and mean irradiation observed and offsetting the energy production proportionally.

The effect of choosing different priors, relating to the distributions of azimuth, tilt and inverter loading ratio of the systems, was explored for the daily and hence annual energy production estimates. For 2016, we found specific annual yields in the range of $877-946\text{kWh/kW}_\text{p}$, which is consistently higher than $875\text{kWh/kW}_\text{p}$ used in the current method. For 2017, Statistics Netherlands may have overestimated the energy production, since we found specific yields in the range of $838-899\text{kWh/kW}_\text{p}$. These results highlight, in the case of the Netherlands, the need for a specific yield to be determined on a daily and annual basis which is a function of the irradiation. More generally, our developed method may be applied in a variety of different national and regional contexts, provided similar or the same data sources are present. The outcomes of this method can be of great use for policy making at a regional level, where efforts are often undertaken to stimulate renewable energy initiatives.

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

We thank Alex Priem, Dick Windmeijer, Jurriën Vroom, Reinoud Segers, Anne Miek Kremer, André Meurink, Otto Swertz, Lyana Curier, Sofie De Broe, Michael Maseda and Nicolas Martin for their help and discussions. The authors also wish to thank colleagues at other European statistical offices which provided information on the calculation of the solar production in their respective countries. W.G.J.H.M. van Sark acknowledges financial support from the Ministry of Economic Affairs and Climate Policy through Topsector Energy funding for the TKI-Urban Energy project PV Observatory. Finally, B.P.M. Laevens thanks his colleagues at the ministry of Economic Affairs and Climate Policy for their support.

Appendix A. Data Cleaning and Processing

A.1. PVOutput

We identified two data cleaning challenges for `PVOutput`. The first was uniformising and correcting the metadata such that apparently similar quantities, inputted by different owners, mean the same thing. The second involved analysing the real–time `PVOutput` data to determine how realistic the data is, also in relation to the metadata.

Uniformising and correcting the metadata. While most variables are reasonably consistent (e.g. ϕ is limited to choosing a cardinal or intercardinal sign), the reliability of some other variables is less clear e.g. P_p, N_p and P. If inputted correctly, then $P = N_p \cdot P_p$; however, for about 10% of all PV systems, this was not the case. To resolve this, we chose P to be the more reliable variable, on the assumption that `PVOutput` owners were more likely to make a mistake concering N_p or P_p.

Since `PVOutput` contains two types of co–ordinates: pc4 and (l,b), we check the reliability of both of these by comparing them to each other. Postal codes in the Netherlands consist of four digits (pc4), followed by two letters. The pc4 provided in `PVOutput`, narrows the locations down to a spatial

area of $\sim 1-8km^2$. The (l, b) allow for pinpointing an exact location. For each `PVOutput` location, we calculated the distances between (l, b) and the pc4 centroids. For about 5% of these systems the distance is more than 5 km. Manual inspection of the pc4 – (l, b) distances for a few municipalities shows an interesting pattern: half of the households took the effort to indeed specify (l, b) nearby their dwelling, whereas the other half seems to have chosen a generic (l, b) based on the municipality. These findings led us to retain the pc4 centroid as the location at the expense of losing – in some cases – more accurate information from (l,b), where we assume that owners are more likely to know what their pc4 is. This level of precision is also sufficient for our research goals. Fig. 11 displays the pc4 – (l, b) distances for the reliable set (see next section) of systems in 2016.

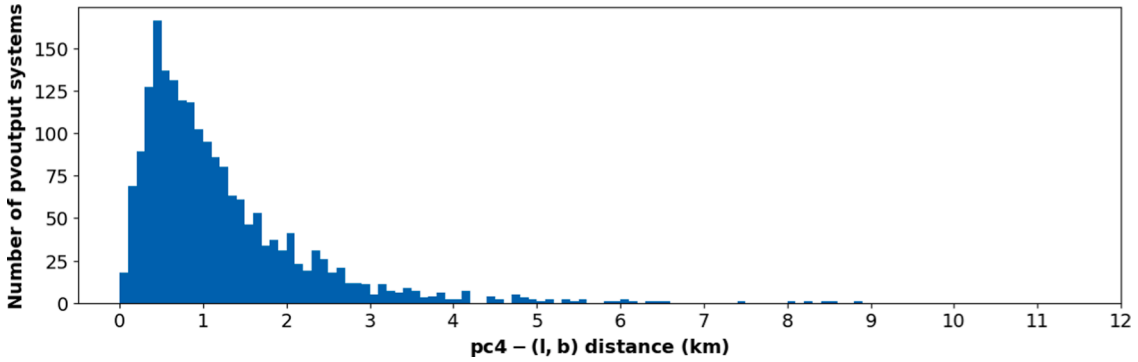

Fig. 11. PC4 centroid–lat/lon distances for the reliable set of PV systems in 2016.

Constructing a reliable set of measurements per day. Having corrected the metadata, we now turn to the measurements. Two examples of `PVOutput` Y-profiles can be seen in Fig. 12 (blue lines). We devise four different quality criteria, which are performed per PV system and day. Daily measurements are deemed reliable if they pass all four checks.

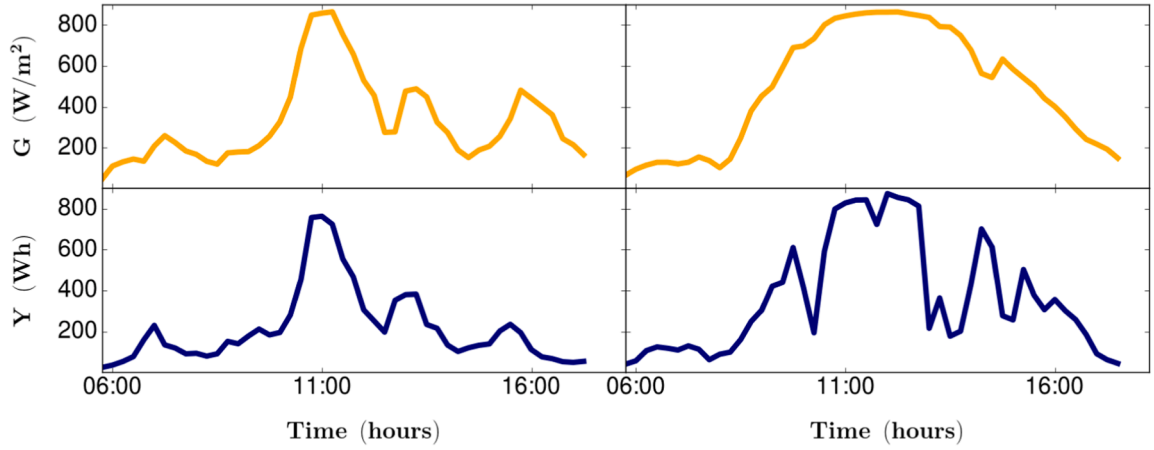

Fig. 12. G (orange) and Y (blue) at two different locations on 19/05/2016. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In Section 3.1.1 we saw that two different energy measurements are provided by `PVOutput`: Y_{inst} and Y_{cum}. The reliability of the latter can be checked by re–calculating the cumulative measurements ($Y_{cum,calc.}$), using the former (Eq. (A.1)). A measurement is deemed acceptable if it satisfies Eq. (A.2), i.e. allowing for a 10% relative difference with $Y_{cum,calc}$. The second check involves identifying the peak energy per day of a system ($P_{inst,peak}$) and checking it makes sense when comparing to P quoted in the metadata (Eq. (A.3)), allowing for a value exceeding P by 20%. The second criterion could be too harsh on some days when cloud–induced superirradiance is capable of temporarily producing $P_{inst,peak}$ which is higher than P (Zhang et al., 2018). We decide not to take this into account as it is difficult for us to ascertain when this local effect may be occurring.

$$Y_{cum,calc.} = \sum_{n=1}^{N} Y_{inst}\Delta t \tag{A.1}$$

$$0.9Y_{cum} < Y_{cum,calc.} < 1.1Y_{cum} \tag{A.2}$$

$$P_{inst,peak} \leqslant 1.2P \tag{A.3}$$

The third quality check (Eq. (A.4)) examines the time intervals (Δt) between each Y_{inst} on a given day. Some PV systems suffer from measuring gaps e.g. a system may measure Y_{inst} consistently, during several hours, with $\Delta t = 5$min, before recording a gap e.g. $\Delta t = 2$ h. This check, in turn, influences quality check number one (Eq. (A.1)). The final check sees if there is a measurement for each day of the year for a given PV system (Eq. (A.5)).

$$\Delta t \leqslant 15\text{min} \tag{A.4}$$

$$N_{meas} > 0 \tag{A.5}$$

It should be noted that the decision to remove PV systems for certain days resulting from Eqs. A.4 and A.5 is not straightforward. We cannot ascertain whether the gaps or missing days are as a result of Wi–Fi stability issues and/or malfunctioning software or whether the PV system is really not producing any energy or has been turned off. The blue line in Fig. 4 shows the number of systems per day for 2016 after data cleaning is performed.

A.2. Dutch meteorological weather data

We do not perform data cleaning on the KNMI irradiance data, since these have been thoroughly checked by KNMI itself. Having downloaded the quarterly hour G_k data, we aggregate G_k to daily irradiation totals H_d (Eq. (A.6)), following the same procedure as Eq. (A.1).

$$H_d = \sum_{k=1}^{K} G_k \Delta t \tag{A.6}$$

where $\Delta t = 15\text{min}$. We notice that, occasionally, an G_k is missing and we therefore adapt Δt in Eq. (A.6) accordingly. We decided not to compare modelled G data with weather station G data because this has already been done in (Greuell et al., 2013), who showed that discrepancies between the two are minimal and negligible. Fig. 1 shows H_d for two consecutive days in June 2016, nicely illustrating the effect of different weather conditions.

Appendix B. Linking PV systems to irradiance grid cells

`PVOutput` systems or database systems are linked to irradiance grid cells in exactly the same way. Firstly, all addresses within a pc4 area are approximated to the centroid of that area in geographic co–ordinates $(l, b)_k$ (Eq. (A.7)). Using the Haversine formula (Rios et al., 1797), the $(l, b)_k$ and $(l, b)_j$ (cell centroids of H) distances may be calculated, thus identifying the closest cell (Eq. (A.8)).

$$pc4_k \sim (l, b)_k, \tag{A.7}$$

$$d_k = \min_{j=1}^{N_{cells}} |(l, b)_k - (l, b)_j|, \ \forall k \in [1, N_{pc4}]. \tag{A.8}$$

This approach makes two assumptions. Firstly, the assignment to the nearest grid cell is correct. Secondly, weather behaves according to the resolution size of a grid cell H. The first will not always hold since pc4 areas can range from ~ 1 to 8 km^2 in size (Kaal and Vanderveen, 2007), because these are effectively a proxy for population density. Since the grid cells of H are 3x6 km^2 in size, it is conceivable that locations in larger pc4 areas are not always linked correctly. The second assumption will not always hold because weather can be more local than the resolution of a grid cell H. This is illustrated by Fig. 12, where at the first location (left panels) Y closely follows H. At the second location (right panels) this is also broadly the case, but it is also obvious that more local effects can be seen in Y, which are not captured by H.

References

Bessa, R., Trindade, A., Silva, C.S., Miranda, V., 2015. Probabilistic solar power forecasting in smart grids using distributed information. International Journal of Electrical Power & Energy Systems 72, 16–23. https://www.sciencedirect.com/science/article/pii/S0142061515000897, doi: 10.1016/j.ijepes.2015.02.006. the Special Issue for 18th Power Systems Computation Conference.

Bright, J.M., Killinger, S., Lingfors, D., Engerer, N.A., 2018. Improved satellite-derived pv power nowcasting using real-time power data from reference pv systems. Solar Energy 168, 118–139. https://www.sciencedirect.com/science/article/pii/S0038092X17309714, doi: 10.1016/j.solener.2017.10.091. advances in Solar Resource Assessment and Forecasting.

CertiQ, 2020. Our Mission. CertiQ, Arnhem http://www.certiq.nl/en/we-are/mission.

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

Drews, A., de Keizer, A.C., Beyer, H.G., Lorenz, E., Betcke, J., van Sark, W.G.J.H.M., Heydenreich, W., Wiemken, E., Stettler, S., Toggweiler, P., Bofinger, S., Schneider, M., Heilscher, G., Heinemann, D., 2007. Monitoring and remote failure detection of grid-connected PV systems based on satellite observations. Sol. Energy 81, 548–564. https://doi.org/10.1016/j.solener.2006.06.019.

EUobserver, 2011. Photovoltaic Barometer 2011. EUobserver, Brussels https://www.eurobserv-er.org/photovoltaic-barometer-2011/.

EUobserver, 2020. Photovoltaic Barometer 2020. EUobserver, Brussels https://www.eurobserv-er.org/photovoltaic-barometer-2020/.

Fonseca Junior, J.G.d.S., Oozeki, T., Ohtake, H., Takashima, T., Ogimoto, K., 2015. Regional forecasts of photovoltaic power generation according to different data availability scenarios: a study of four methods. Progress in Photovoltaics: Research and Applications 23, 1203–1218. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.2528, doi:https://doi.org/10.1002/pip.2528, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/pip.2528.

Greuell, W., Meirink, J.F., Wang, P., 2013. Retrieval and validation of global, direct, and diffuse irradiance derived from SEVIRI satellite observations. Journal of Geophysical Research (Atmospheres) 118, 2340–2361. https://doi.org/10.1002/jgrd.50194.

Hersbach, H., Bell, W., Berrisford, P., Horányi, A., J., M.S., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C., Dee, D., 2019. Global reanalysis: goodbye era-interim, hello era5, 17–24URL https://www.ecmwf.int/node/19027, doi:10.21957/vf291hehd7.

Kaal, H.L., Vanderveen, G.N.G., 2007. Hoe groot is jouw buurt? Nemo, Amsterdam https://www.nemokennislink.nl/publicaties/hoe-groot-is-jouw-buurt.

IEA, 2020. Solar PV. International Energy Agency, Paris https://www.iea.org/reports/solar-pv.

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

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

Killinger, S., Lingfors, D., Saint-Drenan, Y.M., Moraitis, P., van Sark, W., Taylor, J., Engerer, N.A., Bright, J.M., 2018. 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, 1087–1106. https://doi.org/10.1016/j.solener.2018.08.051.

Kirk, P.J., Clark, M.R., Creed, E., 2021. Weather Observations Website. Weather 76, 47–49. https://doi.org/10.1002/wea.3856.

KNMI, Koninklijk Nederlands Meteorologisch Instituut., 2013. Jaaroverzicht van het weer in Nederland – 2012. Koninklijk Nederlands Meteorologisch Instituut, De Bilt. https://cdn.knmi.nl/knmi/map/page/klimatologie/gegevens/mow/jow_2012.pdf.

KNMI, Koninklijk Nederlands Meteorologisch Instituut., 2014. Jaaroverzicht van het weer in Nederland – 2013. Koninklijk Nederlands Meteorologisch Instituut, De Bilt. https://cdn.knmi.nl/knmi/map/page/klimatologie/gegevens/mow/jow_2013.pdf.

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

KNMI, Koninklijk Nederlands Meteorologisch Instituut., 2017b. KNMI - Juni 2017. Koninklijk Nederlands Meteorologisch Instituut, De Bilt. https://www.knmi.nl/nederland-nu/klimatologie/maand-en-seizoensoverzichten/2017/juni.

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

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

KNMI, Koninklijk Nederlands Meteorologisch Instituut., 2019b. Meteosat satellietmetingen van bewolking, zonnestraling en neerslag. Koninklijk Nederlands Meteorologisch Instituut, De Bilt. https://www.knmi.nl/kennis-en-datacentrum/achtergrond/meteosat-satellietmetingen-van-bewolking-zonnestraling-en-neerslag.

KNMI, Koninklijk Nederlands Meteorologisch Instituut., 2020a. Jaaroverzicht van het weer in Nederland – 2019. Koninklijk Nederlands Meteorologisch Instituut, De Bilt. https://cdn.knmi.nl/knmi/map/page/klimatologie/gegevens/mow/jow_2019.pdf.

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

Litjens, G.B.M.A., Worrell, E., van Sark, W.G.J.H.M., 2017. Influence of demand patterns on the optimal orientation of photovoltaic systems. Sol. Energy 155, 1002–1014. https://doi.org/10.1016/j.solener.2017.07.006.

Malof, J.M., Bradbury, K., Collins, L.M., Newell, R.G., 2016. Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. Appl. Energy 183, 229–240. https://doi.org/10.1016/j.apenergy.2016.08.191 https://www.sciencedirect.com/science/article/pii/S0306261916313009.

Marion, B., Smith, B., 2017. Photovoltaic system derived data for determining the solar resource and for modeling the performance of other photovoltaic systems. Sol. Energy 147, 349–357. https://doi.org/10.1016/j.solener.2017.03.043 https://www.sciencedirect.com/science/article/pii/S0038092X17302049.

PVOutput, 2020. Help. PVOutput. https://pvoutput.org/help.html#overview.

PVOutput, 2020. Latest Outputs. PVOutput. https://pvoutput.org.

Reich, N.H., Mueller, B., Armbruster, A., van Sark, W.G.J.H.M., Kiefer, K., Reise, C., 2012. Performance ratio revisited: is pr?>?90% realistic? Prog. Photovoltaics Res. Appl. 20, 717–726. https://doi.org/10.1002/pip.1219. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.1219, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/pip.1219.

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

Rios, J.D.M.Y., Banks, J., Cavendish, H., 1797. Recherches Sur Les Principaux Problemes de l'Astronomie Nautique. Par Don Josef de Mendoza y Rios, F.R.S. Communicated by Sir Joseph Banks, Bart. K.B.P.R.S. Philosophical Transactions of the Royal Society of London Series I 87, 43–122.

RVO, Netherlands Enterprise Agency and Statistics Netherlands., SN, 2015. Protocol monitoring hernieuwbare energie. Rijksdienst voor Ondernemend Nederland, Utrecht. https://www.rvo.nl/sites/default/files/Protocol%20Monitoring%20HE%20Interactief%20V3.pdf.

Saint-Drenan, Y., Good, G., Braun, M., 2017. A probabilistic approach to the estimation of regional photovoltaic power production. Sol. Energy 147, 257–276. https://doi.org/10.1016/j.solener.2017.03.007 https://www.sciencedirect.com/science/article/pii/S0038092X17301676.

Saint-Drenan, Y.M., Vogt, S., Killinger, S., Bright, J.M., Fritz, R., Potthast, R., 2019. Bayesian parameterisation of a regional photovoltaic model – application to forecasting. Sol. Energy 188, 760–774. https://doi.org/10.1016/j.solener.2019.06.053 https://www.sciencedirect.com/science/article/pii/S0038092X19306279.

Schierenbeck, S., Graeber, D., Semmig, A., Weber, A., 2010. Ein distanzbasiertes hochrechnungsverfahren für die einspeisung aus photovoltaik. Energiewirtschaftlichetagesfragen 60, 60–64.

Schmid, J., 2019. The SEVIRI instrument. European Organisation for the Exploitation of Meteorological Satellites, Darmstadt edisp/pdf_ten_msg_seviri_instrument.pdf.

SN, Statistics Netherlands, 2019. Business solar capacity now exceeds residential. Statistics Netherlands, The Hague https://www.cbs.nl/en-gb/news/2020/25/business-solar-capacity-now-exceeds-residential.

SN, Statistics Netherlands., 2019. Smart ways of monitoring solar power. Statistics Netherlands, The Hague https://www.cbs.nl/en-gb/our-services/innovation/project/smart-ways-of-monitoring-solar-power.

SN, Statistics Netherlands., 2020. StatLine - Renewable electricity; production and capacity. Statistics Netherlands, The Hague https://opendata.cbs.nl/statline/#/CBS/en/dataset/82610ENG/table.

Solar Magazine, 2020. Solar Magazine - SolarCare: opbrengsten zonnepanelen in 2019 0,92 kilowattuur per wattpiek. De Duurzame Uitgeverij, Uden https://solarmagazine.nl/nieuws-zonne-energie/i20334/solarcare-opbrengsten-zonnepanelen-in-2019-0-92-kilowattuur-per-wattpiek.

TenneT, 2019. Our key tasks - TenneT. TenneT, Arnhem https://www.tennet.eu/our-key-tasks.

van Sark, W.G.J.H.M., Bosselaar, L., Gerrissen, P., Esmeijer, K., Moraitis, P., van den Donker, M., Emsbroek, G., 2014. Update of the Dutch specific yield for determination of PV contribution to renewable energy production: 25% more energy! Proceedings of the 29th European Photovoltaic Solar Energy Conference 112, 3131–3141. doi:10.1016/j.rse.2008.03.012.

Zhang, J., Watanabe, K., Yoshino, J., Kobayashi, T., Hishikawa, Y., Doi, T., 2018. Physical process and statistical properties of solar irradiance enhancement observed under clouds. Jpn. J. Appl. Phys. 57, 08RG11. https://doi.org/10.7567/jjap.57.08rg11 https://doi.org/10.7567/%2Fjjap.57.08rg11.