Solar Energy 230 (2021) 122–130

Contents lists available at ScienceDirect

Solar Energy

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

Data article

A photovoltaic power output dataset: Multi-source photovoltaic power output dataset with Python toolkit

Tiechui Yao [a,b], Jue Wang [a,b,*], Haoyan Wu [c], Pei Zhang [c,d], Shigang Li [e], Yangang Wang [a,b], Xuebin Chi [a,b], Min Shi [f]

[a] Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[b] University of Chinese Academy of Sciences, School of Computer Science and Technology, Beijing, China
[c] Beijing Jiaotong University, School of Computer and Information Technology, Beijing, China
[d] East China Jiaotong University, School of Electrical and Automation Engineering, Nanchang, Jiangxi, China
[e] ETH Zurich, Department of Computer Science, Switzerland
[f] State Grid Hebei Electric Power Co., Ltd, Shijiazhuang, Hebei, China

ARTICLE INFO

Keywords:
Photovoltaic power output dataset
Open source
Python toolkit
Data article

ABSTRACT

The power output of photovoltaic (PV) systems is chiefly affected by climate and weather conditions. In that, PV farm requires accurate weather data, particularly, solar irradiance, in order to predict its power output as a means to improve solar energy utilization. Nevertheless, publicly available datasets which consist both power and weather data are exceptionally few. This may be a combined effect of data propriety and cumbersome collection procedure. And the rarity of such data greatly hinders the progress of solar PV research. Indeed, most solar energy meteorology applications, such as solar forecasting or PV performance evaluation, can benefit from multi-source high-quality datasets. In view of that, we release a PV power output dataset (PVOD), which contains metadata, numerical weather prediction data, and local measurements data from 10 PV systems located in China. In PVOD, a Python toolkit with basic functions for data access and preprocessing is provided. Additionally, a case study on PV power output estimation is depicted to demonstrate the potential usage of the dataset.

1. Introduction

Solar energy has the potential to become the largest contributor to the world's future energy mix. Photovoltaic (PV) is currently the most common type of solar energy conversion technology, due to its low levelized cost of electricity, high deployability, and mature market structure. However, insofar as collection of PV data is concerned, the amount of available data seems to fall short of its popularity, in that, publicly available databases containing *research-grade* PV data are exceedingly rare. The reason is thought two-fold. First, as PV systems are mainly owned by private entities, who do not seem to have any motivation for sharing the data to the research community, most PV data are kept as proprietary. Second, PV power output data, on its own, has little research value, for it needs to be paired with other forms of solar data, such that various energy meteorology research activities can be conducted.

Indeed, data that facilitate PV energy meteorology research can be categorized into four main types (Sengupta et al., 2021). There are those ground-based measurements, collected by high-quality radiometers, which measure surface solar irradiance. This type of data is very accurate when measured right, and thus is able to provide essential support in situations where ground-truth is needed. Nonetheless, one set of research-grade ground-based radiation monitoring equipment costs hundreds of thousands of dollars (Yang and Liu, 2020). Clearly, it would not be possible to deploy such monitoring equipment at all locations where PV system is (to be) installed. In fact, there are only tens of high-quality radiation monitoring stations in the public domain, such as those consisted in the Baseline Surface Radiation Network (BSRN; Driemel et al., 2018) and Surface Radiation Budget Network (SURFRAD; Augustine et al., 2005, 2000). To that end, satellite-derived irradiance, which constitutes the second type of solar data, is often used as an alternative source of data.

Example satellite-derived irradiance databases include the National Solar Radiation Data Base (NSRDB; Sengupta et al., 2018), Clouds and the Earth's Radiant Energy System (CERES; Wielicki et al., 1996), and Copernicus Atmosphere Monitoring Service Radiation Service (CAMS-RAD; Qu et al., 2017). Satellite-derived irradiance is gridded, and is available for all locations on Earth between ±60° latitudes. However, its

* Corresponding author at: Computer Network Information Center, Chinese Academy of Sciences, Beijing, China.
E-mail address: wangjue@sccas.cn (J. Wang).

https://doi.org/10.1016/j.solener.2021.09.050
Received 23 August 2021; Received in revised form 10 September 2021; Accepted 18 September 2021
Available online 15 October 2021

accuracy is generally lower than ground-based measurements. Nonetheless, some recent evidence suggests that, with the latest-generation satellite-to-irradiance conversion technology, there may be interchangeability between satellite-derived irradiance and ground-based measurements, for certain energy meteorology applications such as forecasting (Yagli et al., 2020), forecast post-processing (Yang, 2019b), and forecast verification (Yang and Perez, 2019; Perez et al., 2016). In exceptional cases, satellite-derived irradiance can even detect calibration drift of ground-truth stations (Perez et al., 2017) and augment the conventional data quality control (Urraca et al., 2017).

The third type of solar data comes from dynamical weather models, which can be subdivided into numerical weather prediction (NWP) data and reanalysis data (Randles et al., 2017; Gelaro et al., 2017; Dee et al., 2011). NWP and reanalysis are also gridded, but they have an even lower accuracy than satellite-derived irradiance. However, these data contain not only information on radiation, but virtually all variables pertaining to our understanding about the Earth's atmosphere. In recent years, a steady accumulation of scientific knowledge in regard to atmospheric physics has led to a quite revolution of dynamical weather models, which allow better prediction at high spatio-temporal resolutions (Bauer et al., 2015). Indeed, some NWP models, such as the High-Resolution Rapid Refresh, have a spatial resolution of 3 km, which is comparable to that of satellite-derived irradiance.

Last but not least, PV data generally refers to those information related to a PV system, which include mainly the system's power output and the site's metadata. As mentioned earlier, power output data, by itself, does not offer much value to the operation and control of the system. This is because the amount of power generated by any PV system is chiefly tied to the climate and weather conditions. In that, the goodness of a PV system must be evaluated with reference to the conditions under which the system is operated, which in turn requires collocated and temporally aligned weather data. Furthermore, because radiation and cloud processes are physical and spatio-temporal by nature, single-location PV power output data has only marginal predictive ability, so far as capturing the cloud dynamics is concerned (Yang, 2019a). Hence, any PV system requires by default weather information, in order to predict its power output as a means to improve solar energy utilization.

But soliciting all above-mentioned data types is no easy task. Even if such data can be identified, data downloading, cleaning, and quality control are often viewed as a time-consuming step and require specialized skills (Urraca et al., 2017; Yang et al., 2018c), preventing researchers from leveraging the best-possible data practices, resulting in a growingly divided literature with highly inconsistent interpretations on what constitutes the state-of-the-art. Moreover, when data is opaque (or proprietary, as some authors often prefer to claim), the works are non-reproducible, which basically render them completely useless. In this regard, *open research* (a synonym for *reproducibility*) has been identified as one of the most important aspects of energy research by many leading researchers (Hong et al., 2020; Yang, 2019a; Kezunovic et al., 2020).

As reproducibility has become an iconic characteristic of high-quality data-oriented energy research, many journals, such as Solar Energy or Journal of Renewable and Sustainable Energy, have come up policies which facilitate data publication (Yang et al., 2018a). The overarching goal of *Data Article* is to provide standardized datasets, such that researchers around the world are able to benefit from those clean, representative, and holistic datasets. On top of that, another defining characteristic of reproducible data-oriented research is computer code. In today's academia where algorithms and frameworks have become increasingly intricate, the effort to reproduce a state-of-the-art work without computer code would be monumental. To that end, many authors have recognized the importance of reproducibility, and responded positively in order to support the Data Article initiative.

For instance, Yang (2018, 2019c) presented a package called `SolarData`, for the R programming environment. The package allows easy access of many popular solar databases, including the aforementioned BSRN, SURFRAD, and NSRDB. Similarly, Bright et al. (2020) presents a Python package for the downloading and post-processing the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) reanalysis data, which is essential for computing the clear-sky models using the REST2 algorithm (Gueymard, 2008). In fact, besides REST2, the same group of authors also implemented almost all available clear-sky models to date, and compiled them into several R scripts, facilitating the worldwide uptake of their results (Sun et al., 2019, 2021). Contribution of a liked nature include Feng et al. (2019), Holmgren et al. (2018), Lamigueiro (2012); to save space, the reader is simple referred the original works for details.

Whereas the above-mentioned works all focus on code, there are other Data Articles that focus on releasing datasets. For example, Pedro et al. (2019) released a comprehensive dataset facilitating solar forecasting of various kinds; the dataset contains ground-based irradiance, sky camera and satellite imagery, and NWP output. The authors released alongside with the article some benchmarking forecasting methods, which, up to now, have already received much echo (Yang et al., 2020b; Kezunovic et al., 2020). Peterson and Vignola (2020) discussed a new comprehensive format for radiometry measurement; the work also points the reader to the world's longest-history radiation monitoring station, namely, the Eugene station, operated by Frank Vignola from the Solar Radiation Monitoring Laboratory, University of Oregon. Silwal et al. (2021), on the other hand, provided a multi-year power generation, consumption, and storage dataset in a microgrid context, allowing various energy modeling research works to be conducted. Similarly, Stowell et al. (2020) releases solar PV installations datasets across the UK for small-scale solar panel installations.

Clearly, none of the above-mentioned dataset overlaps with one another in terms of their intended purposes, dataset covered, or code released. Nonetheless, datasets that contains both PV and NWP data are still rare. In consideration of that, an open-sourced PV power output dataset (PVOD) containing local measurements of PV power stations and numerical weather prediction (NWP) is released in this paper, and to facilitate its uptake, a lightweight and extensible Python toolkit is developed for this dataset.

This Data Article should serve the purpose of introducing the PVOD dataset and the accompanied Python toolkit. In this version, the toolkit consists of two parts: (1) basic function library, which provides data access and preprocessing functions, and (2) user-defined sample code. As we should wish to seek constant expansion and improvements of the PVOD dataset, this version is label as version 1.0. The remaining part of the paper is organized as follows. The data description is presented in Section 2. The potential usage of this dataset is discussed in Section 3. Subsequently, the computer code and a case study, depicting the PV output estimation procedure, are expounded in Section 4. Conclusions follow at the end.

2. Data description

The PVOD dataset consists of a metadata file and data files of 10 PV sites, in comma-separated values (CSV) format, which can be easily browsed in Microsoft Office Excel or Notepad. As shown in Table 1, which describes the contents of these files, metadata.csv covers the basic information of all PV sites, whereas the files station*.csv contain the details of meteorological data and *in situ* measurements.

Technical specifications of PV panels (e.g., capacity, area, number, and orientation) and site location, as described in metadata.csv, are essential if one wishes to convert irradiance to PV power, using *model chain* (Mayer and Gróf, 2021). On the other hand, the temporal resolution of weather data need to be consistent with that of PV power output—otherwise one has to make a resolution change when those data are integrated, which is often considered to be an extremely challenging task (Yang and van der Meer, 2021; Yang et al., 2019). Similarly, the amount of data points ought to be substantial, such that

Table 1
Detailed description of the files in PV output dataset (PVOD).

File	Name	Description	Units
Metadata	Station_ID	The ID of stations is numbered from 0 to 9	–
	Capacity	The installed capacity of the power station	kW
	PV_Technology	The material type of PV panel	–
	Panel_Size	The size of a PV panel	m^2
	Module	The module information of the PV panel	–
	Inverter	The solar inverters information of the PV system	–
	Panel_Number	Total number of PV panels laid in for the station	1
	Array_Tilt	Angle of PV panels	degree
	Pyranometer	The pyranometers information of the station	–
	Longitude	The longitude of the station	degree
	Latitude	The latitude of the station	degree
Station data [0-9]	Date_time	Format: Year-Month-Day Hour-Min	–
	nwp_globalirrad	Global irradiance of NWP	W/m^2
	nwp_directirrad	Direct irradiance of NWP	W/m^2
	nwp_temperature	10-meter dry-bulb temperature of NWP	°C
	nwp_humidity	10-meter relative humidity of NWP	%
	nwp_windspeed	10-meter wind speed of NWP	m/s
	nwp_winddirection	10-meter wind direction of NWP, zero north clockwise	degree
	nwp_pressure	Atmospheric pressure of NWP	hPa
	lmd_totalirrad	Global irradiance of LMD	W/m^2
	lmd_diffuseirrad	Diffuse irradiance of LMD	W/m^2
	lmd_temperature	Temperature of LMD	°C
	lmd_pressure	Atmospheric pressure of LMD	hPa
	lmd_winddirection	Wind direction of LMD	degree
	lmd_windspeed	Wind speed of LMD	m/s
	Power	PV output of the station	MW

training routines of various kinds, in regard to predictive modeling, can be efficient and effective. On this point, PVOD has a total of 271,968 records and contains NWP output at a 15-min temporal resolution, which is the same as that of local measurement data (LMD) from PV sites.

The NWP model used to generate the PVOD data is a version of the Weather Research and Forecasting (WRF) model (Michalakes et al., 2001), or more specifically, the Advanced Research WRF (ARW) version 3.9.1 modeling system. ARW is a fully compressible, Eulerian, and non-hydrostatic model that uses a terrain-following hydrostatic-pressure vertical coordinate and an Arakawa C-grid staggering spatial discretization. The model runs once a day, operationally, with a horizontal grid resolution of 4 km, 45 terrain-following (Eta) vertical levels from the surface to the top at 70 hPa. The model is initialized with 3-hourly, 0.125°×0.125° global-scale NWP forecasts, disseminated operationally at 12 Coordinated Universal Time (UTC) by the European Centre for Medium-Range Weather Forecasting (ECMWF), which is commonly regarded as the most accurate global NWP today. The physics schemes and parameterization used in ARW include the New Thompson microphysics scheme, Grell–Freitas cumulus parameterization, Mellor–Yamada–Janjic planetary boundary layer, and RRTMG scheme for short and longwave radiation. Since these choices are well-documented in WRF manual, they are not reiterated herein.

Insofar as this paper is concerned, the NWP variables are extracted over the horizon ranging from 28 to 54 h, with a 15-min resolution. A total 7 features, namely, global horizontal irradiance (GHI), direct normal irradiance (DNI), 10-m temperature, 10-m humidity, 10-m wind speed and wind direction, as well as pressure, are extracted, because these are the variables that are thought most relevant to PV power modeling and forecasting (Hong et al., 2016). Local measurement data corresponds to NWP output, in that, a total of 7 variables, namely, the GHI, diffuse horizontal irradiance (DHI), temperature, pressure, wind direction, wind speed, and PV output, are included. One should take note that although DNI is not included in LMD, it can be in fact calculated via the closure relationship, i.e., $\mathrm{GHI} = \mathrm{DNI}\cos Z + \mathrm{DHI}$, where Z is zenith angle which can be calculated using a solar positioning algorithm with latitude, longitude and time as inputs.

All 10 PV sites are located in Hebei Province, China (see Fig. 1), which is situated between 36.64403–39.5155° N in latitude, 113.6419–117.4572° E in longitude. The overall dataset covers over 300 days

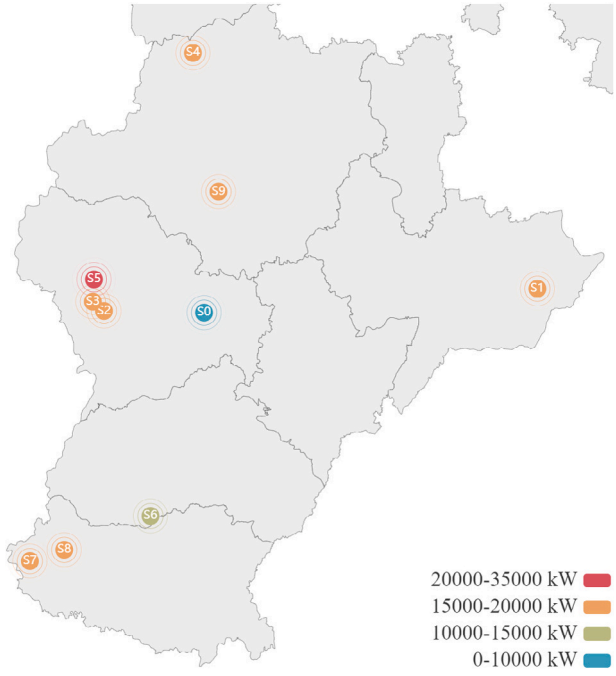

Fig. 1. Geographical distribution and installed capacity of the power stations in PVOD.

and is arranged in chronological order, from 2018-07-01 to 2019-06-13. It should be noted that this data range seems a bit short, but given the resolution of the data, it might be sufficient for a variety of studies, as discussed in Section 3. Moreover, when more data becomes progressively available in the future, the present authors are happy to update the PVOD dataset. In any case, the timestamp, which is in Coordinated Universal Time (UTC), is formatted into corresponding string (e.g., year-month-day hour-minute). The PV output, from two sample stations, over a 9-day period, is shown in Fig. 2. Last but not least, to facilitate the uptake, the sample code is programmed in Python, which is the most popular programming language today, offering basic usage demos and user-defined functions. All data and sample code can be

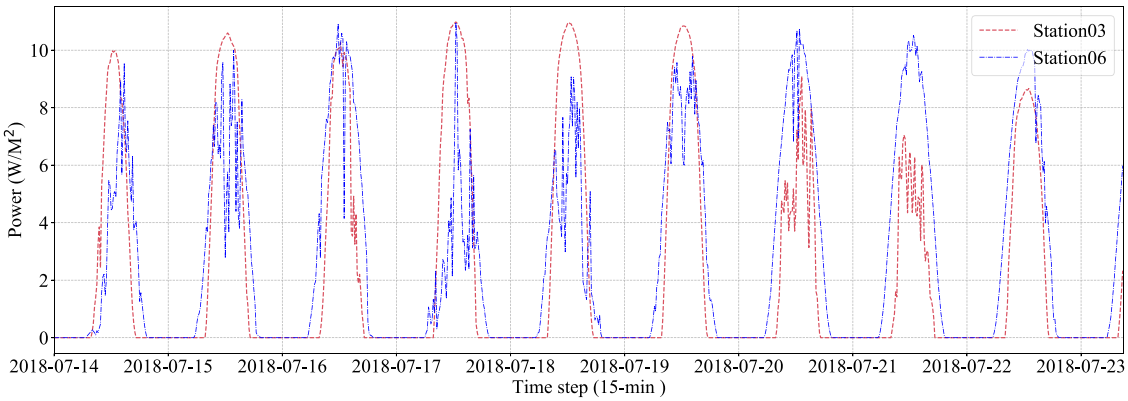

Fig. 2. Sample PV power output from two (randomly chosen) power stations, over a 9-day period from July 14–23, 2018.

downloaded as supplementary material. More updates can be found on Github (https://github.com/yaotc/PVODataset).

3. Potential usage

Limited by the fact that the PVOD is released for the first time, there is a lack of published previous papers that used this dataset. Despite of that, the application of datasets similar to PVOD in photovoltaic power forecasting is of reference significance. Wolff et al. (2016) proposed a physical modeling approach using a private PV dataset[1] (e.g., measurements, NWP data). Besides the PV power measurements, information on location, tilt, and installed capacity of these PV systems, were used to model and forecast PV power, as well as the NWP that were partly offered by ECMWF. Raza et al. (2018) combined historical PV power output, humidity, temperature, wind speed, and solar irradiance as inputs, from the dataset at University of Queensland St. Lucia and Gatton campuses[2], for neural network ensemble scheme of PV power output. Wiese et al. (2019) introduced an open power system data platform that provided five data packages for electricity system modeling, e.g, conventional power plants, renewable power plants, time series, and weather data. In the time series package[3], all variables (e.g., electricity prices, electricity consumption, wind power generation, solar power generation, capacities) were well collected and provided in the hourly resolution, whereas some data processing codes were not designed as functions.

Furthermore, the potential applications of this dataset are numerous, and only a few are enumerated in what follows. Firstly, the most iconic application that is supported by this dataset is PV power output modeling. Generally speaking, the modeling of PV power output can be categorized into two types. There are those direct methods and indirect methods. On one hand, direct methods seek to regress PV power output onto various predictors, such as GHI, DNI, temperature or wind speed. On the other hand, indirect methods (also known as model chain) exploit the physical relationship of various meteorological predictors with PV power output. Comparing the two categories of methods, the latter appears, *a priori*, to be more advantageous. However, at the moment, owing to the lack of attention on this research topic, there is not any consensus nor concrete evidence suggesting which category of methods is more accurate. Hence, the current dataset supports investigation of that sort.

Next, a related application is to study the error propagation from NWP forecasts to PV power forecasts. In essence, the errors of PV power forecasts come from two main sources: (1) the error of NWP forecasts, and (2) the error of model chain. Currently, how these two error sources interact seems opaque. The simultaneously availability of NWP output and LMD may provide some new evidence. For instance, one can first perform PV power output modeling using LMD, whose uncertainties are much lower than that of NWP forecasts; this would nevertheless result in some errors due to the imperfect model chain. Subsequently, when NWP forecasts are used as input instead, the additional errors seen in the PV power output can be solely attributed to NWP forecasts. Since such an investigation is only possible when both NWP output and LMD contain the same variables, which is the case for PVOD, the present dataset is valuable in that respect.

The third application which we should wish to mention is geographical smoothing. In short, geographical smoothing describes the phenomenon that when multiple PV farms are situated within a wide area, their combined power output is smoother (or less variable) than that of any single PV farm (Klima et al., 2018; Lave et al., 2012, 2013). Put it simply, the smoothing is caused by different cloudiness at different locations, such that the peaks and valleys of solar power production within an area are evened out. Owing to this effect, geographical smoothing has a profound implication on the power system stability and control, which is becoming more and more important with higher and higher penetration level of solar energy (Hasan et al., 2019). Most generally, in order to investigate the smoothing effect, geographically dispersed PV systems are needed. Given the fact that the PVOD dataset contains PV systems spanning a provincial grid, it can be deemed appropriate.

Last but not least, the present dataset supports spatio-temporal solar forecasting. Of course, to produce PV power forecasts, one can opt to use only endogenous data, which basically means to use the past power records from the PV plant to forecast the future (e.g., Rana et al., 2016; Cheng et al., 2021). However, due to the lack of spatio-temporal information, forecasting methods trained using single-location data are unable to detect incoming clouds, which is commonly regarded as the most influential factor to PV power output. Indeed, because of this grave disadvantage, in almost every review on solar forecasting (e.g., Antonanzas et al., 2016; Yang et al., 2018b; Yang, 2019a), forecasting methods that uses exogenous variables are advocated and recommended. Exogenous variables for the present dataset come from two ways: (1) NWP data is considered to be one, and its usefulness has been made fully evident earlier; (2) PV power output and weather data from neighboring locations is the other. In the latter case, due to the fact that PV power output from neighboring stations are often correlated, forecasting using multiple stations is almost always better than that using a single station (Agoua et al., 2019; Yang et al., 2015). With that, we conclude this section.

4. Toolkit usage

4.1. Functions

The sample code of PVOD is programmed using the object-oriented programming method, and it follows the open-closed principle and

[1] http://www.meteocontrol.com/.
[2] http://www.uq.edu.au/solarenergy/.
[3] https://data.open-power-system-data.org/time_series/.

single responsibility principle. The object-oriented programming enables the sample code of this dataset to implement query, retrieval, calculation, and modification operations, without writing redundant functions. The open-closed principle makes sure that all entities in our sample code (classes, modules, functions, etc.) are open for extension, but closed for modification. In other words, users are allowed to extend their functions without modifying the sample code. The single responsibility principle protects that each function plays a specific role, which makes the user feel free to access this dataset and upgrade their code.

We provide a class PVODataset that supports basic functions, such as reading original data, summarizing information, slicing data according to date and time (UTC and UTC+08:00), obtaining common data of multiple power stations, dividing training set and test set, or user-defined extension. For example, the dataset can be loaded through a directory path and listed via **show_files** function.

```
# load PV output Dataset
# Optional: timezone="UTC", or "UTC+8", Asia/Shanghai.
pvod = PVODataset(path="../datasets/", timezone="UTC+8")
# show all files
files = pvod.show_files()
files
```

```
## Welcome to PVODataset (PVOD).
##
## ['metadata.csv', 'station00.csv', 'station01.csv',
## 'station02.csv', 'station03.csv', 'station04.csv',
## 'station05.csv', 'station06.csv', 'station07.csv',
## 'station08.csv', 'station09.csv']
```

The keywords in station*.csv (cf. Table 1) and the amount of each PV station are presented by function **info**.

```
# show basic information of PV output Dataset
pvod.info()
```

```
## PVOD provides 1 metadata file and 10 PV station data files.
## The header of station files is : 'date_time', 'nwp_globalirrad',
## 'nwp_directirrad', 'nwp_temperature', 'nwp_humidity',
## 'nwp_windspeed', 'nwp_winddirection', 'nwp_pressure',
## 'lmd_totalirrad', 'lmd_diffuseirrad', 'lmd_temperature',
## 'lmd_pressure','lmd_winddirection', 'lmd_windspeed', 'power'.
## -->Records of Station_0 are 28896.
## -->Records of Station_1 are 33408.
## -->Records of Station_2 are 30432.
## -->Records of Station_3 are 14688.
## -->Records of Station_4 are 33408.
## -->Records of Station_5 are 9696.
## -->Records of Station_6 are 31104.
## -->Records of Station_7 are 32928.
## -->Records of Station_8 are 33120.
## -->Records of Station_9 are 24288.
## -> Total 271968 records.
```

We also provide a simple way to access metadata and PV station data via functions **read_metadata** and **read_ori_data**. However, one should note that the ID number needs to be passed to the function in order to read PV data from a specific station.

```
# load metadata
metadata = pvod.read_metadata()
print(metadata)
# load selected station original data.
ori_data = pvod.read_ori_data(station_id=3)
print(ori_data)
```

```
## Station_ID Capacity PV_Technology ... Array_Tilt Longitude Latitude
## station00 6600  Poly-Si ... South 33° 114.951390 38.047780
## station01 20000 Poly-Si ... South 33° 117.457220 38.183060
##  ... ... ... ... ... ... ... ...
## station08 20000 Poly-Si ... South 33° 113.899990 36.707610
## station09 20000 Poly-Si ... South 31° 115.059855 38.731417
## 10 rows x 8 columns
##
## date_time nwp_globalirrad ... lmd_windspeed power
## 2019-01-12 00:00:00 0.0 ... 1.0 0.0
## 2019-01-12 00:15:00 0.0 ... 1.0 0.0
## ... ... ... ... ... ...
## 2019-06-13 23:30:00 0.0 ... 0.0 0.0
## 2019-06-13 23:45:00 0.0 ... 2.4 0.0
## 14688 rows x 15 columns
```

Statistical indicators (e.g., average, standard deviation, maximum, and minimum) are necessary for data interpretation, preprocessing, and normalization. In this regard, the function **station_info** can output these indicators of each feature in this dataset.

```
# show station information
station01_info = pvod.station_info(station_id=1)
station01_info
```

```
##     nwp_globalirrad nwp_directirrad ... lmd_windspeed power
## count 33408.000000 33408.000000 ... 33408.000000 33408.000000
## mean 158.834149 138.337171 ... 1.040493 3.677611
## std  235.866804 217.017108 ... 1.217974 5.553049
## min  0.000000 0.000000 ... 0.000000 0.000000
## 25%  0.000000 0.000000 ... 0.000000 0.000000
## 50%  0.000000 0.000000 ... 0.700000 0.000000
## 75%  278.115000 232.375000 ... 1.600000 6.464543
## max  936.420000 879.530000 ... 11.300000 19.997459
## 8 rows x 14 columns
```

Filtering data within a specified date range is essential for research. The function **select_daterange** supports slicing the entire dataset based on different power stations, start dates, and end dates. It should be noted that the end date must be later than the start date.

```
# select data within date range
t1 = "2019/3/05 08:00"
t2 = "2019/5/20 17:00"
slice_data = pvod.select_daterange(station_id=5, start_date=t1,
      end_date=t2)
slice_data
```

```
## date_time nwp_globalirrad ... lmd_pressure lmd_winddirection
lmd_windspeed power
## 2019-03-05 08:00:00 160.15 ... 994.000000 187 0.7 0.00000
## 2019-03-05 08:15:00 217.53 ... 994.299988 192 0.7 0.00000
## ... ... ... ... ... ... ... ... ... ...
## 2019-05-20 16:45:00 462.04 ... 984.400024 41 4.1 12.47291
## 2019-05-20 17:00:00 424.80 ... 984.200012 42 1.6 10.43153
## 7333 rows ÃŮ 15 columns
```

When the user decides to study the tasks involving multiple power stations, the function **date_intersection** can be used to generate the intersectional timestamp between power stations. Subsequently, the corresponding data fragments can be obtained via the function **select_daterange**.

```
# data intersection between 2 PV stations
start, end = pvod.date_intersection(station_id_a=3,
      station_id_b=9)
start, end
```

```
## Station_3: start:2019-01-12 00:00:00, end:2019-06-13 23:45:00
## Station_9: start:2018-09-26 00:00:00, end:2019-06-13 23:45:00
## intersec : start:2019-01-12 00:00:00, end:2019-06-13 23:45:00
## ('2019-01-12 00:00:00'), ('2019-06-13 23:45:00')
```

Should the user be interested in dividing the entire dataset according to a certain proportion, using the **split_data** function will be an ideal choice. This function divides the dataset into two parts in chronological order based on a given ratio. If the user's task is a regression problem (e.g., irradiance prediction, PV power forecasting), the training set and the test set can be easily derived.

```
# split Train dataset and Test dataset
ori_data = pvod.read_ori_data(station_id=3, timezone="UTC+8")
train_data, test_data = pvod.split_data(xy=ori_data, mode="
      end_order", ratio=0.8)
print(f"train :\n {train_data} \n test:\n {test_data}, \n\
       len_train: {len(train_data)}, len_test: {len(test_data)}
            ")
```

```
## train :
  date_time  nwp_globalirrad  nwp_directirrad  nwp_temperature \
## 2019-01-12 00:00:00  0.00  0.00         -0.07
## 2019-01-12 00:15:00  0.00  0.00         -0.13
## ...      ...      ...   ...          ...
## 2019-05-14 09:00:00  609.95  561.95       25.15
## 2019-05-14 09:15:00  647.76  599.06       25.68
##
## ...  lmd_pressure  lmd_winddirection  lmd_windspeed    power
## ...   988.500000            48         1.0  0.000000
## ...   988.500000           138         1.0  0.000000
## ...   ...          ...          ...     ...      ...
## ...   975.099976           267         0.3  7.786116
## ...   975.099976            46         1.7  5.740129
##
## [11750 rows x 15 columns]
##
## test:
  date_time  nwp_globalirrad  nwp_directirrad  nwp_temperature \
## 2019-05-14 09:30:00  663.30  613.56         26.15
## 2019-05-14 09:45:00  681.97  622.12         26.56
##...      ...      ...     ...      ...
## 2019-06-13 23:30:00  0.00  0.00          23.95
## 2019-06-13 23:45:00  0.00  0.00          23.82
##
## ...  lmd_pressure  lmd_winddirection  lmd_windspeed    power
## ...   975.099976           257         0.9  9.164315
## ...   975.099976           265         1.6  11.352380
## ...   ...          ...          ...     ...      ...
## ...   972.799988           104         0.0  0.000000
## ...   972.799988           163         2.4  0.000000
##
## [2938 rows x 15 columns],
## len_train: 11750, len_test: 2938
```

For more advance usage of the dataset, the user-defined class UDFClass can be freely developed based on class PVODataset. Following the demo code below, users can implement various functions on this dataset, e.g., correlation analysis and data normalization, mathematical modeling, or machine-learning modeling. For instance, if users expect to calculate the total area of the PV panels of the first power station, they need to implement function **calculation**. The input is the station ID (1), the area of a single PV panel ("Panel_Size", keyword in metadata.csv), and the number of PV panels ("Panel_Number", keyword in metadata.csv). Next, all they need to do is write one line of code (Line 14), and the total area just returns.

```
1  # User-Defined Functions Class (demo code)
2  class UDFClass(PVODataset):
3      """
4      The usage of Inheritance grammar.
5      """
6      def __init__(self, path=../datasets/, params=0):
7          super(UDFClass, self).__init__(path)
8          pass
9
10     # user-defined demo0
11     def calculation(self, station_id, param0, param1):
12         meta_id = self.metadata.loc[station_id]
13         value0, value1 = meta_id[param0], meta_id[param1]
14         area = value0 * value1
15         print(f"area of station {station_id} = {area} m^2.")
16         return area
17
18     # users-defined demo1
19     def get_id_metadata(self, station_id):
20         return self.metadata.loc[station_id]
21
22     # users-defined demo2
23     def norm_dataframe(self, xy, mode="minmax"):
24         data = xy.copy()
25         if mode == "std":
26             xy = (data - data.mean()) / data.std()
27             return np.array(xy)
28         elif mode == "minmax":
29             xy = (data - data.min()) / (data.max() - data.min())
30             return np.array(xy)
31         else:
32             raise "norm Error."
33
34 # usage demo
35 users_func = UDFClass()
```

```
36 users_func.calculation(station_id=1, param0="Panel_Size",
       param1="Panel_Number")
37 users_func.get_id_metadata(1)
```

```
## Welcome to PVODataset (PVOD).
## area of station 1 = 123099.0 m^2.
##
## Station_ID                    station01
## Capacity                        20000
## PV_Technology                   Poly-Si
## Panel_Size                      1.6635
## Module          products types:LW6P60-2...
## Inverters      products types:TC500KH\nMa...
## Layout          modules per string:22\ns...
## Panel_Number                     74000
## Array_Tilt                    South 33°
## Pyranometer    GHI: \nproducts types: TBQ...
## Longitude                      117.457
## Latitude                       38.1831
```

4.2. Case study

To demonstrate the potential usage of this dataset, a toy example of is given. More specially, a PV power output modeling example, for station 7 (S7) and power station 8 (S8), is depicted. It should be noted that in many applications, such as solar forecasting, the algorithm is often not directly applied to irradiance, instead, it is applied on a normalized quantity known as the clear-sky index (Yang et al., 2018b; Yang, 2019a). In the case of PV power, the normalized quantity is know as K_{PV}, which is the clear-sky index of PV power, which is calculated as:

$$K_{PV} = \frac{P_{MEAS}}{P_{CLR}}, \tag{1}$$

where P_{MEAS} is the measured PV power, and P_{CLR} is the expected PV power output under a cloud-free, i.e., clear, sky condition (Engerer and Mills, 2014). Clearly, to calculate P_{CLR}, one needs to know the clear-sky irradiance, then, the clear-sky irradiance can be converted to clear-sky power output via a model chain.

Like many other classes of radiation models, there are nearly a hundred clear-sky irradiance models available in the literature (Sun et al., 2019, 2021). The performance of these models varies largely, depending on how much physics is involved during modeling. Whereas aerosol and water vapor are two main causes of attenuation of incoming irradiance, other atmospheric particulates also have some effect. To that end, the models that explicitly consider these are generally superior to those do not. In a recent study on the choice of clear-sky models in solar forecasting, Yang (2020) has argued that the McClear clear-sky model (Lefevre et al., 2013; Gschwind et al., 2019), due to its physical modeling and ease of access, is a good option. This view has subsequently confirmed by an international panel of experts (Yang et al., 2020a). To that end, the McClear clear-sky irradiance is used in this paper; it can be downloaded at the SoDa website.[4]

After clear-sky GHI, DHI, and DNI for stations S7 and S8. These components are passed onto a model chain, using the `pvlib` package in Python (Holmgren et al., 2018). The first step of the model chain is to convert the horizontal irradiance components to the plane-of-array (POA) irradiance; this is known as transposition modeling. There are about 30 well-known transposition models in the literature, but the 1990 version of the Perez model (Perez et al., 1990) has hitherto been the most choice, owing to its universality and its asymptotic level of optimization (Yang, 2016). Hence, it is used here. Besides transposition model, for other components of the model chain, such as loss model, DC model, or AC model, the `pvlib` default choices are used with our system specifications as follows:

[4] http://www.soda-pro.com/web-services/radiation/cams-mcclear.

```
# import packages
from pvlib.temperature import TEMPERATURE_MODEL_PARAMETERS
from pvlib.pvsystem import PVSystem
from pvlib.modelchain import ModelChain
...
tem_para = TEMPERATURE_MODEL_PARAMETERS["sapm"]["
      open_rack_glass_glass"]
# Parameters of PV module
module_parameters = pvsystem.retrieve_sam("CECMod")["
      Yingli_Energy__China__YL250P_29b"]
# Parameters of inverter
# redefine the inverter parameters based on a similar one because
      of the missing record of inverter product
inverter_para = pvsystem.retrieve_sam("cecinverter")
["Advanced_Energy_Industries__Solaron_500kW...480V_"]
inverter_para["Pdco"]      = 567000
inverter_para["Vdco"]      = 315
inverter_para["Vdcmax"]    = 1000
inverter_para["Idcmax"]    = 1134
inverter_para["Mppt_low"]  = 460
inverter_para["Mppt_high"] = 950
# init PV system
system = PVSystem(surface_tilt=33, surface_azimuth=180,
                  module_parameters=module_parameters,
                  inverter_parameters=inverter_para,
                  modules_per_string=20,
                  strings_per_inverter=100,
                  temperature_model_parameters=tem_para)
...
# init ModelChain
# Parameters of ModelChain
mc = ModelChain(system, site,
                transposition_model="perez",
                solar_position_method="nrel_numpy",
                orientation_strategy="south_at_latitude_tilt",
                aoi_model="physical",
                spectral_model="no_loss")
...
mc.run_model(self.read_weather())
```

At this moment, PV_{CLR} is obtained. For visualization, the daily transients of PV_{MEAS} and PV_{CLR} are shown in Fig. 3(a), for March 6, 2019, as an example. Using Eq. (1), K_{PV} is calculated and depicted in Fig. 3(b). It can be seen that March 6 is a day without cloud, and that is why the PV_{MEAS} and PV_{CLR} agree with each other well, while K_{PV} during daytime appears to be flat. Similarly, as shown on the left side of Fig. 4(a) and Fig. 4(b), due to the partly cloudy weather condition on March 17, the K_{PV} in the front part is relatively flat, and the latter part is heavily fluctuates. More intuitively, as shown on the right side, K_{PV} continues to fluctuate throughout the day because of the overcast situation on March 18.

5. Conclusion

Inspired by the recent wave of promoting open research in solar engineering (Yang, 2019c; Bright et al., 2020), we released this PV power output dataset (PVOD). This dataset comes from two sources (NWP and local measurements), and include 14 columns of features and timestamps. Furthermore, a Python toolkit is provided for easy usage of this dataset. We expect that this released dataset, together with the Python toolkit, can help promote the corresponding research in the field of solar energy. In the future, we should wish to continue to expand this dataset, by adding more power stations and updating the Python toolkit. Additionally, the present authors also would consider linking satellite-derived irradiance from Fengyun-4 or Himarari-8, in order to further enhance the value of this dataset.

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.

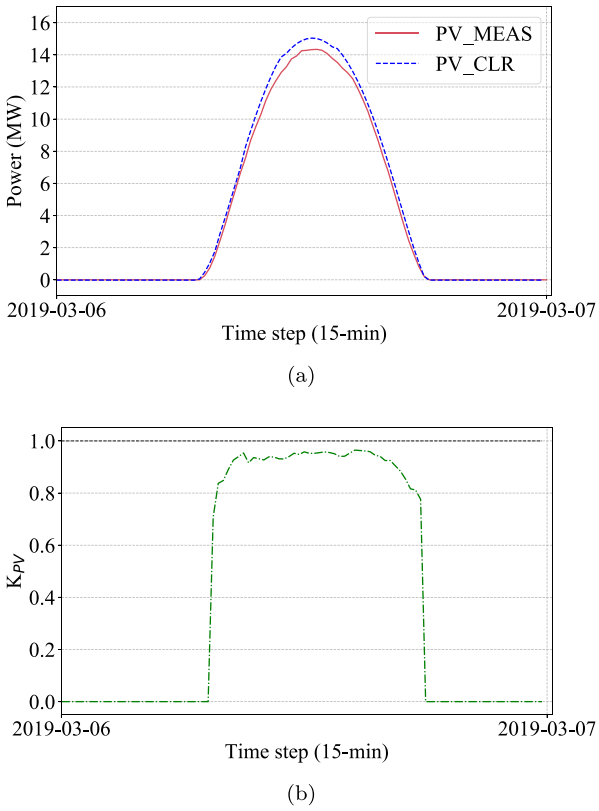

Fig. 3. (a) PV_{CLR} calculations and PV_{MEAS} from March 6, 2019, for station S7. (b) K_{PV} for the same day.

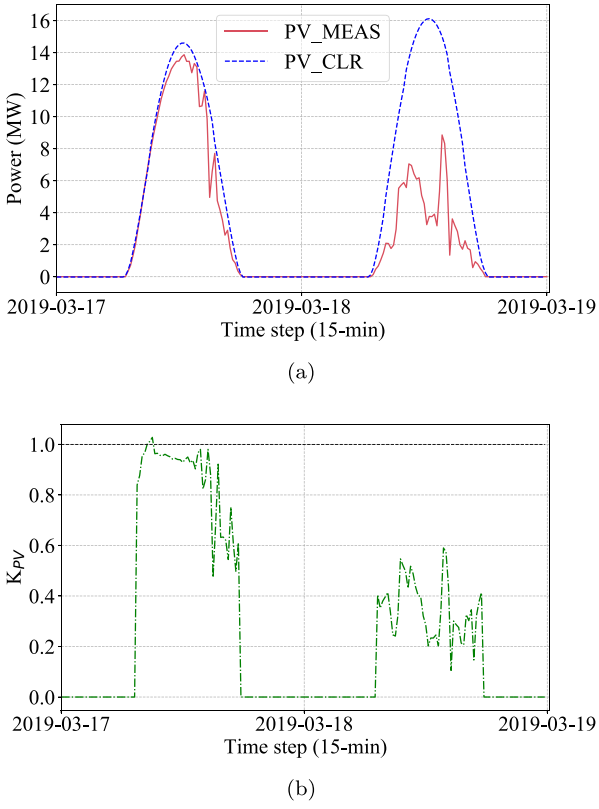

Fig. 4. (a) PV_{CLR} calculations and PV_{MEAS} from March 17–19, 2019, for station S7. (b) K_{PV} for the same days.

Acknowledgments

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (grant number XDA27000000).

Appendix A. Some issues with the PVOD dataset

- Timestamp alignment: Since the timestamps of each power station are not completely consistent, please pay attention to the alignment of the timestamps, especially when using the data of multiple power stations for joint experiments.
- Missing record of products: Some modules of PV plane and inverters products are missing in the CEC database.[5] If these products are used in other applications, it may be necessary to manually complete the relevant parameters according to the metadata.

Appendix B. Supplementary material

- Data access: This dataset can be also publicly accessed (Yao et al., 2021) through Science Data Bank (http://www.dx.doi.org/10.11922/sciencedb.01094).
- Code support: Researchers are welcome to use this code on GitHub (https://github.com/yaotc/PVODataset), we also provide the code in the form of *Jupyter Notebook*.

References

Agoua, X.G., Girard, R., Kariniotakis, G., 2019. Probabilistic models for spatio-temporal photovoltaic power forecasting. IEEE Trans. Sustain. Energy 10 (2), 780–789. http://dx.doi.org/10.1109/TSTE.2018.2847558.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de Pison, F.J., Antonanzas-Torres, F., 2016. Review of photovoltaic power forecasting. Sol. Energy 136, 78–111.

Augustine, J.A., DeLuisi, J.J., Long, C.N., 2000. SURFRAD—A national surface radiation budget network for atmospheric research. Bull. Am. Meteorol. Soc. 81 (10), 2341 — 2358, URL: https://doi.org/10.1175/1520-0477(2000)081<2341:SANSRB>2.3.CO;2.

Augustine, J.A., Hodges, G.B., Cornwall, C.R., Michalsky, J.J., Medina, C.I., 2005. An update on SURFRAD—The GCOS surface radiation budget network for the continental united states. J. Atmos. Ocean. Technol. 22 (10), 1460–1472. http://dx.doi.org/10.1175/JTECH1806.1.

Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction. Nature 525 (7567), 47–55. http://dx.doi.org/10.1038/nature14956.

Bright, J.M., Bai, X., Zhang, Y., Sun, X., Acord, B., Wang, P., 2020. irradpy: Python package for MERRA-2 download, extraction and usage for clear-sky irradiance modelling. Sol. Energy 199, 685–693. http://dx.doi.org/10.1016/j.solener.2020.02.061, URL: https://www.sciencedirect.com/science/article/pii/S0038092X20301894.

Cheng, L., Zang, H., Ding, T., Wei, Z., Sun, G., 2021. Multi-meteorological-factor-based graph modeling for photovoltaic power forecasting. IEEE Trans. Sustain. Energy.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137 (656), 553–597. http://dx.doi.org/10.1002/qj.828.

Driemel, A., Augustine, J., Behrens, K., Colle, S., Cox, C., Cuevas-Agulló, E., Denn, F.M., Duprat, T., Fukuda, M., Grobe, H., Haeffelin, M., Hodges, G., Hyett, N., Ijima, O., Kallis, A., Knap, W., Kustov, V., Long, C., Longenecker, D., Lupi, A., Maturilli, M., Mimouni, M., Ntsangwane, L., Ogihara, H., Olano, X., Olefs, M., Omori, M., Passamani, L., Pereira, E.B., Schmithüsen, H., Schumacher, S., Sieger, R., Tamlyn, J., Vogt, R., Vuilleumier, L., Xia, X., Ohmura, A., König-Langlo, G., 2018. Baseline Surface Radiation Network (BSRN): structure and data description (1992–2017). Earth Syst. Sci. Data 10 (3), 1491–1501. http://dx.doi.org/10.5194/essd-10-1491-2018, URL: https://www.earth-syst-sci-data.net/10/1491/2018/.

[5] https://github.com/NREL/SAM/tree/develop/deploy/libraries.

Engerer, N., Mills, F., 2014. KPV: A clear-sky index for photovoltaics. Sol. Energy 105, 679–693. http://dx.doi.org/10.1016/j.solener.2014.04.019, URL: https://www.sciencedirect.com/science/article/pii/S0038092X14002151.

Feng, C., Yang, D., Hodge, B.M., Zhang, J., 2019. OpenSolar: Promoting the openness and accessibility of diverse public solar datasets. Sol. Energy 188, 1369–1379. http://dx.doi.org/10.1016/j.solener.2019.07.016, URL: https://www.sciencedirect.com/science/article/pii/S0038092X19306693.

Gelaro, R., McCarty, W., Suárez, M.J., Todling, R., Molod, A., Takacs, L., Randles, C.A., Darmenov, A., Bosilovich, M.G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A.M., Gu, W., Kim, G.K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J.E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S.D., Sienkiewicz, M., Zhao, B., 2017. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). J. Clim. 30 (14), 5419–5454. http://dx.doi.org/10.1175/JCLI-D-16-0758.1.

Gschwind, B., Wald, L., Blanc, P., Lefèvre, M., Schroedter-Homscheidt, M., Arola, A., 2019. Improving the McClear model estimating the downwelling solar radiation at ground level in cloud-free conditions–McClear-v3. Meteorol. Z. 28 (2).

Gueymard, C.A., 2008. REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation – Validation with a benchmark dataset. Sol. Energy 82 (3), 272–285. http://dx.doi.org/10.1016/j.solener.2007.04.008, URL: http://www.sciencedirect.com/science/article/pii/S0038092X07000990.

Hasan, K.N., Preece, R., Milanović, J.V., 2019. Existing approaches and trends in uncertainty modelling and probabilistic stability analysis of power systems with renewable generation. Renew. Sustain. Energy Rev. 101, 168–180.

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

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J., 2016. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. Int. J. Forecast. 32 (3), 896–913. http://dx.doi.org/10.1016/j.ijforecast.2016.02.001, URL: https://www.sciencedirect.com/science/article/pii/S0169207016000133.

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., Zareipour, H., 2020. Energy forecasting: A review and outlook. IEEE Open Access J. Power Energy 7, 376–388. http://dx.doi.org/10.1109/OAJPE.2020.3029979.

Kezunovic, M., Pinson, P., Obradovic, Z., Grijalva, S., Hong, T., Bessa, R., 2020. Big data analytics for future electricity grids. Electr. Power Syst. Res. 189, 106788. http://dx.doi.org/10.1016/j.epsr.2020.106788, URL: https://www.sciencedirect.com/science/article/pii/S0378779620305915.

Klima, K., Apt, J., Bandi, M., Happy, P., Loutan, C., Young, R., 2018. Geographic smoothing of solar photovoltaic electric power production in the Western USA. J. Renew. Sustain. Energy 10 (5), 053504.

Lamigueiro, O.P., 2012. solaR: solar radiation and photovoltaic systems with R. J. Stat. Softw. 50 (9), 1–32.

Lave, M., Kleissl, J., Arias-Castro, E., 2012. High-frequency irradiance fluctuations and geographic smoothing. Sol. Energy 86 (8), 2190–2199. http://dx.doi.org/10.1016/j.solener.2011.06.031, Progress in Solar Energy 3. URL: https://www.sciencedirect.com/science/article/pii/S0038092X11002611.

Lave, M., Kleissl, J., Stein, J.S., 2013. A wavelet-based variability model (WVM) for solar PV power plants. IEEE Trans. Sustain. Energy 4 (2), 501–509. http://dx.doi.org/10.1109/TSTE.2012.2205716.

Lefevre, M., Oumbe, A., Blanc, P., Espinar, B., Gschwind, B., Qu, Z., Wald, L., Schroedter-Homscheidt, M., Hoyer-Klick, C., Arola, A., et al., 2013. McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmos. Meas. Tech. 6 (9), 2403–2418.

Mayer, M.J., Gróf, G., 2021. Extensive comparison of physical models for photovoltaic power forecasting. Appl. Energy 283, 116239. http://dx.doi.org/10.1016/j.apenergy.2020.116239, URL: https://www.sciencedirect.com/science/article/pii/S0306261920316330.

Michalakes, J., Chen, S., Dudhia, J., Hart, L., Klemp, J., Middlecoff, J., Skamarock, W., 2001. Development of a next-generation regional weather research and forecast model. In: Developments in Teracomputing. World Scientific, pp. 269–276.

Pedro, H.T.C., Larson, D.P., Coimbra, C.F.M., 2019. A comprehensive dataset for the accelerated development and benchmarking of solar forecasting methods. J. Renew. Sustain. Energy 11 (3), 036102. http://dx.doi.org/10.1063/1.5094494.

Perez, R., Ineichen, P., Seals, R., Michalsky, J., Stewart, R., 1990. Modeling daylight availability and irradiance components from direct and global irradiance. Sol. Energy 44 (5), 271–289. http://dx.doi.org/10.1016/0038-092X(90)90055-H, URL: https://www.sciencedirect.com/science/article/pii/0038092X9090055H.

Perez, R., Schlemmer, J., Hemker, K., Kivalov, S., Kankiewicz, A., Dise, J., 2016. Solar energy forecast validation for extended areas & economic impact of forecast accuracy. In: 2016 IEEE 43rd Photovoltaic Specialists Conference. PVSC, pp. 1119–1124. http://dx.doi.org/10.1109/PVSC.2016.7749787.

Perez, R., Schlemmer, J., Kankiewicz, A., Dise, J., Tadese, A., Hoff, T., 2017. Detecting calibration drift at ground truth stations a demonstration of satellite irradiance models' accuracy. In: 2017 IEEE 44th Photovoltaic Specialist Conference. PVSC, pp. 1104–1109. http://dx.doi.org/10.1109/PVSC.2017.8366469.

Peterson, J., Vignola, F., 2020. Structure of a comprehensive solar radiation dataset. Sol. Energy 211, 366–374. http://dx.doi.org/10.1016/j.solener.2020.08.092, URL: https://www.sciencedirect.com/science/article/pii/S0038092X20309312.

Qu, Z., Oumbe, A., Blanc, P., Espinar, B., Gesell, G., Gschwind, B., Klüser, L., Lefèvre, M., Saboret, L., Schroedter-Homscheidt, M., Wald, L., 2017. Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method. Meteorol. Z. 26 (1), 33–57. http://dx.doi.org/10.1127/metz/2016/0781.

Rana, M., Koprinska, I., Agelidis, V.G., 2016. Univariate and multivariate methods for very short-term solar photovoltaic power forecasting. Energy Convers. Manage. 121, 380–390.

Randles, C., da Silva, A., Buchard, V., Colarco, P., Darmenov, A., Govindaraju, R., Smirnov, A., Holben, B., Ferrare, R., Hair, J., Shinozuka, Y., Flynn, C., 2017. The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation. J. Clim. 30 (17), 6823–6850. http://dx.doi.org/10.1175/JCLI-D-16-0609.1.

Raza, M., Mithulananthan, N., Li, J., Lee, K., Gooi, H., 2018. An ensemble framework for day-ahead forecast of pv output power in smart grids. IEEE Trans. Industrial Informatics 15, 4624–4634.

Sengupta, M., Habte, A., Wilbert, S., Gueymard, C., Remund, J., 2021. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications. Technical Report, National Renewable Energy Lab.(NREL), Golden, CO (United States).

Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G., Shelby, J., 2018. The national solar radiation data base (NSRDB). Renew. Sustain. Energy Rev. 89, 51–60. http://dx.doi.org/10.1016/j.rser.2018.03.003, URL: http://www.sciencedirect.com/science/article/pii/S136403211830087X.

Silwal, S., Mullican, C., Chen, Y.A., Ghosh, A., Dilliott, J., Kleissl, J., 2021. Open-source multi-year power generation, consumption, and storage data in a microgrid. J. Renew. Sustain. Energy 13 (2), 025301. http://dx.doi.org/10.1063/5.0038650.

Stowell, D., Kelly, J., Tanner, D., Taylor, J., Jones, E., Geddes, J., Chalstrey, E., 2020. A harmonised, high-coverage, open dataset of solar photovoltaic installations in the UK. Sci. Data 7 (1), 1–15.

Sun, X., Bright, J.M., Gueymard, C.A., Acord, B., Wang, P., Engerer, N.A., 2019. Worldwide performance assessment of 75 global clear-sky irradiance models using Principal Component Analysis. Renew. Sustain. Energy Rev. 111, 550–570. http://dx.doi.org/10.1016/j.rser.2019.04.006, URL: http://www.sciencedirect.com/science/article/pii/S1364032119302187.

Sun, X., Bright, J.M., Gueymard, C.A., Bai, X., Acord, B., Wang, P., 2021. Worldwide performance assessment of 95 direct and diffuse clear-sky irradiance models using principal component analysis. Renew. Sustain. Energy Rev. 135, 110087. http://dx.doi.org/10.1016/j.rser.2020.110087, URL: https://www.sciencedirect.com/science/article/pii/S1364032120303786.

Urraca, R., Gracia-Amillo, A.M., Huld, T., Martinez-de Pison, F.J., Trentmann, J., Lindfors, A.V., Riihelä, A., Sanz-Garcia, A., 2017. Quality control of global solar radiation data with satellite-based products. Sol. Energy 158, 49–62.

Wielicki, B.A., Barkstrom, B.R., Harrison, E.F., Lee, R.B., Smith, G.L., Cooper, J.E., 1996. Clouds and the earth's radiant energy system (CERES): An earth observing system experiment. Bull. Am. Meteorol. Soc. 77 (5), 853–868. http://dx.doi.org/10.1175/1520-0477(1996)077<0853:CATERE>2.0.CO;2.

Wiese, F., Schlecht, I., Bunke, W., Gerbaulet, C., Hirth, L., Jahn, M., Kunz, F., Lorenz, C., Mühlenpfordt, J., Reimann, J., et al., 2019. Open power system data–frictionless data for electricity system modelling. Applied Energy 236, 401–409.

Wolff, B., Kühnert, J., Lorenz, E., Kramer, O., Heinemann, D., 2016. Comparing support vector regression for pv power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. Solar Energy 135, 197–208.

Yagli, G.M., Yang, D., Gandhi, O., Srinivasan, D., 2020. Can we justify producing univariate machine-learning forecasts with satellite-derived solar irradiance? Appl. Energy 259, 114122. http://dx.doi.org/10.1016/j.apenergy.2019.114122, URL: https://www.sciencedirect.com/science/article/pii/S0306261919318094.

Yang, D., 2016. Solar radiation on inclined surfaces: Corrections and benchmarks. Sol. Energy 136, 288–302. http://dx.doi.org/10.1016/j.solener.2016.06.062, URL: https://www.sciencedirect.com/science/article/pii/S0038092X16302432.

Yang, D., 2018. SolarData: An R package for easy access of publicly available solar datasets. Sol. Energy 171, A3–A12. http://dx.doi.org/10.1016/j.solener.2018.06.107, URL: http://www.sciencedirect.com/science/article/pii/S0038092X18306583.

Yang, D., 2019a. A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES). J. Renew. Sustain. Energy 11 (2), 022701. http://dx.doi.org/10.1063/1.5087462.

Yang, D., 2019b. Post-processing of NWP forecasts using ground or satellite-derived data through kernel conditional density estimation. J. Renew. Sustain. Energy 11 (2), 026101. http://dx.doi.org/10.1063/1.5088721.

Yang, D., 2019c. SolarData package update v1.1: R functions for easy access of Baseline Surface Radiation Network (BSRN). Sol. Energy 188, 970–975. http://dx.doi.org/10.1016/j.solener.2019.05.068, URL: https://www.sciencedirect.com/science/article/pii/S0038092X19305493.

Yang, D., 2020. Choice of clear-sky model in solar forecasting. J. Renew. Sustain. Energy 12 (2), 026101. http://dx.doi.org/10.1063/5.0003495.

Yang, D., Alessandrini, S., Antonanzas, J., Antonanzas-Torres, F., Badescu, V., Beyer, H.G., Blaga, R., Boland, J., Bright, J.M., Coimbra, C.F.M., David, M., Frimane, Â., Gueymard, C.A., Hong, T., Kay, M.J., Killinger, S., Kleissl, J., Lauret, P., Lorenz, E., van der Meer, D., Paulescu, M., Perez, R., Perpiñán-Lamigueiro, O., Peters, I.M., Reikard, G., Renné, D., Saint-Drenan, Y.M., Shuai, Y., Urraca, R., Verbois, H., Vignola, F., Voyant, C., Zhang, J., 2020a. Verification of deterministic solar forecasts. Sol. Energy 210, 20–37. http://dx.doi.org/10.1016/j.solener.2020.04.019, URL: https://www.sciencedirect.com/science/article/pii/S0038092X20303947.

Yang, D., Gueymard, C.A., Kleissl, J., 2018a. Editorial: Submission of Data Article is now open. Sol. Energy 171, A1–A2. http://dx.doi.org/10.1016/j.solener.2018.07.006, URL: http://www.sciencedirect.com/science/article/pii/S0038092X18306698.

Yang, D., Kleissl, J., Gueymard, C.A., Pedro, H.T.C., Coimbra, C.F.M., 2018b. History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. Sol. Energy 168, 60–101. http://dx.doi.org/10.1016/j.solener.2017.11.023, Advances in Solar Resource Assessment and Forecasting. URL: https://www.sciencedirect.com/science/article/pii/S0038092X17310022.

Yang, D., Liu, L., 2020. Solar project financing, bankability, and resource assessment. In: Sustainable Energy Solutions for Remote Areas in the Tropics. Springer International Publishing Cham, pp. 179–211.

Yang, D., Perez, R., 2019. Can we gauge forecasts using satellite-derived solar irradiance? J. Renew. Sustain. Energy 11 (2), 023101. http://dx.doi.org/10.1063/1.5046711.

Yang, D., van der Meer, D., 2021. Post-processing in solar forecasting: Ten overarching thinking tools. Renew. Sustain. Energy Rev. 140, 110735. http://dx.doi.org/10.1016/j.rser.2021.110735, URL: https://www.sciencedirect.com/science/article/pii/S1364032121000307.

Yang, D., van der Meer, D., Munkhammar, J., 2020b. Probabilistic solar forecasting benchmarks on a standardized dataset at Folsom, California. Sol. Energy 206, 628–639. http://dx.doi.org/10.1016/j.solener.2020.05.020, URL: https://www.sciencedirect.com/science/article/pii/S0038092X20305090.

Yang, D., Wu, E., Kleissl, J., 2019. Operational solar forecasting for the real-time market. Int. J. Forecast. 35 (4), 1499–1519. http://dx.doi.org/10.1016/j.ijforecast.2019.03.009, URL: https://www.sciencedirect.com/science/article/pii/S0169207019300755.

Yang, D., Yagli, G.M., Quan, H., 2018c. Quality control for solar irradiance data. In: 2018 IEEE Innovative Smart Grid Technologies - Asia. ISGT Asia, pp. 208–213. http://dx.doi.org/10.1109/ISGT-Asia.2018.8467892.

Yang, D., Ye, Z., Lim, L.H.I., Dong, Z., 2015. Very short term irradiance forecasting using the lasso. Sol. Energy 114, 314–326. http://dx.doi.org/10.1016/j.solener.2015.01.016, URL: https://www.sciencedirect.com/science/article/pii/S0038092X15000304.

Yao, T., Wang, J., Wu, H., Zhang, P., Li, S., Wang, Y., Chi, X., Shi, M., 2021. Data from:PVOD v1.0 : A photovoltaic power output dataset. http://dx.doi.org/10.11922/sciencedb.01094, http://www.dx.doi.org/10.11922/sciencedb.01094.