Solar Energy 255 (2023) 171–179

Contents lists available at ScienceDirect

Solar Energy

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

Data article

SKIPP'D: A SKy Images and Photovoltaic Power Generation Dataset for short-term solar forecasting

Yuhao Nie [a,1], Xiatong Li [b,1], Andea Scott [a], Yuchi Sun [a,2], Vignesh Venugopal [a,2], Adam Brandt [a,*]

[a] Department of Energy Science and Engineering, Stanford University, Stanford, CA 94305, USA
[b] Department of Civil and Environmental Engineering, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Keywords:
Solar forecasting
PV output prediction
Fish-eye camera
Sky images
Deep learning
Computer vision

ABSTRACT

Large-scale integration of photovoltaics (PV) into electricity grids is challenged by the intermittent nature of solar power. Sky-image-based solar forecasting using deep learning has been recognized as a promising approach to predicting the short-term fluctuations. However, there are few publicly available standardized benchmark datasets for image-based solar forecasting, which limits the comparison of different forecasting models and the exploration of forecasting methods. To fill these gaps, we introduce SKIPP'D—a SKy Images and Photovoltaic Power Generation Dataset. The dataset contains three years (2017–2019) of quality-controlled down-sampled sky images and PV power generation data that is ready-to-use for short-term solar forecasting using deep learning. In addition, to support the flexibility in research, we provide the high resolution, high frequency sky images and PV power generation data as well as the concurrent sky video footage. We also include a code base containing data processing scripts and baseline model implementations for researchers to reproduce our previous work and accelerate their research in solar forecasting.

1. Introduction

Solar PV is rapidly becoming a significant source of power generation. Fluctuations in solar power generation due to short-term events (like moving clouds) can have large impacts in areas with high solar PV penetration. Images captured by ground-based fish-eye cameras contain a wealth of information about the sky, but this information is challenging to extract and use for reliable predictions. In the past five years, using emerging deep learning models to "read" the sky and make forecasts of PV power generation (or solar irradiance) has shown promising performance, These deep learning models are mainly based on convolutional neural networks (CNNs), either solely using CNNs (Sun et al., 2018a, 2019; Nie et al., 2020; Feng and Zhang, 2020; Feng et al., 2022) or hybridizing CNNs with recurrent neural networks (RNNs), such as LSTM (Zhang et al., 2018; Paletta et al., 2021a,b).

However, two major challenges have been identified in this fast growing area. First, prior work is hard to compare as the models are developed using different datasets with different specifications. There is a lack of standardized datasets for benchmarking deep-learning-based solar forecasting models. Secondly, deep learning models are data hungry. To make deep learning models generalize well, it often requires massive and diversified training data. Several different parties have been contributing to addressing the data accessibility issues. The first type of party is comprised of national labs and research organizations which have multi-year efforts in data collection and publication of archived datasets to the public. For example, Solar Radiation Research Laboratory (SRRL) (Stoffel and Andreas, 1981) from NREL developed a Baseline Measurement System (BMS), which open sources multi-years of 10 min high resolution sky images from two different sky cameras together with 1 min radiation and meteorological measurements in Golden, Colorado. However, the low temporal resolution of the image data can hardly satisfy the need of short-term solar forecasting due to the variation and volatility of the clouds. Another example is the National Surface Radiation (SURFRAD) Budget Network developed by National Oceanic and Atmospheric Administration (NOAA) (Augustine et al., 2000). It provides an archived dataset of sky images and radiation and meteorological sensor measurement data from multiple sites across United States for public use, but the sky images are usually logged in low-resolution (e.g., 288 × 352) and low temporal frequency (e.g., 1 h). The second type of party is research groups sharing the solar forecasting datasets used in their publications. One of the most comprehensive datasets for solar forecasting was compiled by Pedro et al. (2019), which includes 3-years of high-resolution sky images (1536 × 1536)

* Corresponding author.
E-mail address: abrandt@stanford.edu (A. Brandt).
[1] The first two authors have equal contribution.
[2] Current affiliation: Energy and Environmental Economics, Inc., San Francisco, CA 94104, USA.

https://doi.org/10.1016/j.solener.2023.03.043
Received 25 October 2022; Received in revised form 6 March 2023; Accepted 21 March 2023
Available online 28 March 2023

and irradiance measurements in 1 min frequency collected at Folsom, California, together with overlapping data from satellite imagery and Numerical Weather Prediction forecasts, as well as the features extracted from the imagery and irradiance data. Other efforts include *SkyCam* by Ntavelis et al. (2021) which contains 1-year of sky images (600 × 600) collected from three different locations in Switzerland together with the overlapping irradiance measurements both logged in 10-second frequency, and *Girasol* by Terrén-Serrano et al. (2021) which provides high temporal frequency sky images from both visible and infrared cameras and irradiance measurements data for 244 days, while the visible images only have one intensity channel and the infrared images are of low resolution. The datasets released by Dissawa et al. (2021) and Bassous et al. (2021) both provide sky images and PV power output data, but the sizes are pretty small as they only include a few days/months of data. Finally, a third type of party is researchers developing open-source tools for easy access of publicly available datasets, most of which are from national labs or research organizations as described previously. Such efforts include Yang (2018)'s *SolarData* and Feng et al. (2019)'s *OpenSolar*.

Despite the various efforts we mentioned above, there are limited publicly available datasets that are compiled with high quality data and are suitable for deep-learning-based or computer-vision-based short-term solar forecasting research. To fill these gaps, we introduce SKIPP'D — a SKy Images and Photovoltaic Power Generation Dataset, collected and compiled by the Environmental Assessment and Optimization (EAO) Group at Stanford University. The dataset contains the following two levels of data which distinguishes it from most of the existing open-sourced datasets and makes it especially suitable for deep-learning-based solar forecasting research:

1. Benchmark dataset: 3 years of processed sky images (64 × 64) and concurrent PV power generation data with a 1 min interval that are ready-to-use for deep learning model development;
2. Raw dataset: Overlapping high resolution sky video footage (2048 × 2048) recorded at 20 frames per second, and sky image frames (2048 × 2048) and history PV power generation data logged with a 1 min frequency that suit various research purposes.

In addition, we provide a code base containing data processing scripts and baseline model implementations for researchers to quickly reproduce our previous work and accelerate solar forecasting research. We hope that this dataset will facilitate the research of image-based solar forecasting and contribute to a standardized benchmark for evaluating and comparing different solar forecasting models. Besides, we also encourage the users to explore on solar forecasting related areas with this dataset, such as sky image segmentation and cloud movement forecasting.

2. Data sources

Our research group started the data collection from March 2017 at Stanford University's campus, located in the center of the San Francisco Peninsula, in California. According to the Köppen climate classification system, Stanford has a warm-summer Mediterranean climate, abbreviated Csb (C = temperate climate s = dry summer b = warm summer) on climate maps (Kottek et al., 2006). In terms of cloud coverage, Stanford is featured by long summers with mostly clear skies and short winters with partly cloudy skies. Two major categories of data are collected and logged: sky images and PV power generation, which are detailed in Sections 2.1 and 2.2, respectively. Over the past five years, our lab has collected over two terabytes of data, which have enabled multiple published solar forecasting studies, covering a wide range of topics, including nowcasting (Sun et al., 2018a), forecasting (Sun et al., 2018b, 2019), data fusion (Venugopal et al., 2019), sky-condition-specific submodels (Nie et al., 2020), data augmentation (Nie et al., 2021), transfer learning (Nie et al., 2022b) and survey of open-source sky image datasets (Nie et al., 2022a). Besides, there is ongoing work across the world based on our dataset.

Table 2.1
Specifications table.

Subject area	Solar forecasting; Computer vision; Deep learning
More specific subject area	PV power generation prediction; Sun tracking; Cloud detection; Cloud movement prediction
Data collection period	March 2017 to December 2019
Type of data	1. Processed (benchmark) data: 1 min 64 × 64 sky images (*.npy*) and PV power generation (*.npy*) pairs, partitioned into model development set (training+validation) and test set, and further structured and stored as *.hdf5* format. 2. Raw data: 2048 × 2048 sky videos recorded at 20 frames per second (*.mp4*), 1 min 2048 × 2048 sky images (*.jpg*) and PV power generation (*.csv*)
How data was acquired	The sky videos/images are acquired by the camera installed on top of the Green Earth Sciences Building (37.427°, −122.174°) at Stanford University. The PV power generation data are from PV panel approximately 125 m away from the camera on the roof of the Jen-Hsun Huang Engineering Center at Stanford University, which are logged by Stanford Utility and shared to us.
Data accessibility	The processed (benchmark) data is available at https://purl.stanford.edu/dj417rh1007 and the raw data is deposit separately by each year given its large size. The 2017 raw data is available at https://purl.stanford.edu/sm043zf7254 and the links to 2018 and 2019 data can be found in the "Related items" elsewhere on the same web page.
GitHub repository	Code base of data processing and baseline model are available at the GitHub repository https://github.com/yuhao-nie/Stanford-solar-forecasting-dataset.

In this release, we open-source the data from March 2017 to December 2019.[3] Here, we provide two levels of data to suit the different needs of researchers: (1) A processed dataset consists of 1 min downsampled sky images (64 × 64) and PV power generation pairs, which is intended for fast reproducing our previous work and accelerating the development and benchmarking of deep-learning-based solar forecasting models; (2) A raw dataset consisting of high resolution sky images (2048 × 2048) and PV power generation data, as well as the source sky video footage, which is intended for customizing data extraction and exploring other related areas of solar forecasting such as cloud segmentation and cloud movement forecasting. The specifications of the data are summarized in Table 2.1 and a full description of the data files can be found in Appendix A.

In future releases, we will open source the data from 2020 and beyond of the Stanford dataset and include two additional data sources[4]: sky images and PV power generation data from a solar farm in Oregon collected by our research group and sky images from cameras set up by NREL which correspond to solar irradiance data collected by them. The solar farm in Oregon is a 3.08 MW-DC ground mounted solar array located outside of Sheridan, a rural community. Sheridan is located in the Willamette Valley which, like Stanford, has a warm-summer Mediterranean climate. However, unlike Stanford, the winters are long and typically characterized by cloudy skies and periods of rain. Our team set up a fish-eye camera (Hikvision DS-2CD6365GOE-IVS) at this location at the end of 2021 to start collecting data. In addition to sky images from Oregon, our group started collecting sky

[3] The dataset suffers from some interruptions due to the water intrusion, wiring and/or electrical failure of the camera, as well as daylight-saving adjustment failure of the camera in 2017 and 2018, which is back to normal for 2019 and beyond.

[4] The updated information will be released on our dataset GitHub repository: https://github.com/yuhao-nie/Stanford-solar-forecasting-dataset.

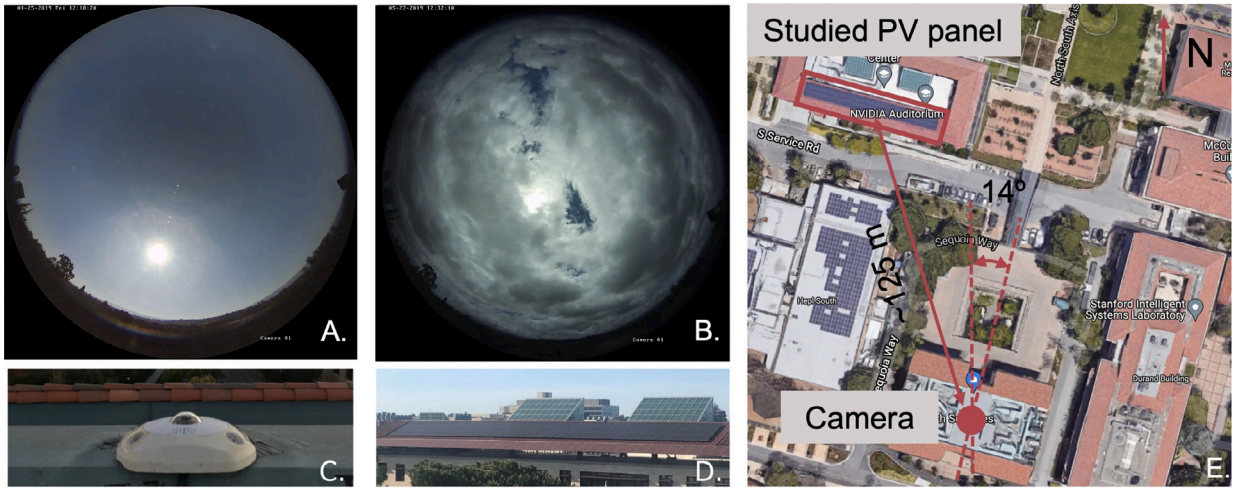

Fig. 2.1. Photos of Sky Images and Research Equipment. (A. Sky image captured on a clear day at 12:18:20 pm, January 25, 2019. B. Sky image of a cloudy day captured at 12:32:10 pm May 27, 2019. C. Fish-eye camera used for sky imaging. D. Studied PV panels. E. Locations of the camera and studied solar panels.)
Source: Adapted from Sun et al. (2018a), used with permission.

images in 2021 from NREL's Solar Radiation Research Laboratory in Golden, Colorado via their website,[5] which provides a live view of the sky image every 60 s. Unfortunately, NREL is currently only storing images every 10 min, so our team is collecting and storing the minutely images in order to have a dataset that is useful to the short-term solar forecasting community. These images can be paired with NREL's minutely irradiance data which is available via their website.

2.1. Sky images

The sky images are frames from videos recorded during daytime (6:00 AM ~ 8:00 PM[6]) by a 6-megapixel 360-degree fish-eye camera (Hikvision DS-2CD6362F-IV2[7]) located on top of the Green Earth Sciences Building (37.427°, −122.174°) at Stanford University and oriented towards 14° south by west. Compared with other high-end commercial sky imaging systems, this off-the-shelf network camera is more accessible and affordable for sky monitoring. The camera holds constant camera aperture, white balance and dynamic range and captures video with a resolution of 2048 × 2048 pixels at 20 frames per second (fps). The images (*.jpg*) are extracted from the video at 1 min sampling frequency and are down-sampled to a resolution of 64 × 64 pixels, which is found to be acceptable for PV output forecast while retaining reasonable model training time (Sun et al., 2018a). Fig. 2.1 gives examples of sky images in different weather conditions, and shows the locations of the camera and PV panels used in this collection.

2.2. PV power generation

The PV output data are collected from solar panel arrays approximately 125 m away from the camera, situated on the top of the Jen-Hsun Huang Engineering Center at Stanford University, with and elevation angle of 22.5° and an azimuth angle of 195°. The PV panels are manufactured with poly-crystalline technology and the system is rated at 30.1 kW-DC. The PV output generation data are logged with 1 min frequency and are minutely averaged. The forward average is applied, e.g., value at 8:00:00 am representing the average PV generation from 8:00:00 to 8:00:59 am.

3. Data processing

To support the flexibility of research, we also open source high-resolution, high-frequency raw data, and the users can customize their own data processing pipeline and process the data based on their needs. Here, we provide a reference by going over the data processing steps we used to generate the processed (benchmark) data. The processing steps were largely used in our previous published work except some minor modifications.[8] We also provide the code base of data processing for users' reference and the descriptions of the code files can be found in Appendix B.

The data processing basically includes the following four major steps. More details on the data processing steps can be found in the PhD dissertation by Sun (2019).

1. Obtain raw high-resolution image frames (2048 × 2048) by snapshotting the video footage at a designated frequency. While 1 min sampling frequency was used in the dataset, it is freely adjustable.
2. Process raw PV power generation history, which includes the following two sub-steps:
   (a) Interpolate PV data to every 10 s in preparation for matching with images in Step 3 with irregular time stamps, e.g., 08:20:10 (raw PV data are logged regularly as 08:00:00, 08:01:00, 08:02:00, etc.)
   (b) Filter out the PV data that are abnormally recorded (e.g., data logger repeatedly logs one value for a certain time period), negative (e.g., night time) or have missing records larger than 1 h.
3. Process raw high resolution images and pair the processed images with the concurrent processed PV generation data. The image processing includes the following two sub-steps:

[5] NREL Solar Radiation Laboratory Baseline Management System website: https://midcdmz.nrel.gov/apps/sitehome.pl?site=BMS.

[6] Data were recorded based on the local time zone, which is either Pacific Standard Time (PST) or Pacific Daylight Time (PDT). In US, PDT starts on the second Sunday in March and ends on the first Sunday in November.

[7] The camera model Hikvision DS-2CD6362F-IV is discontinued and is replaced by a new model Hikvision DS-2CD6365GOE-IVS. We replace the old model with the new model on April 29, 2022 due to aging.

[8] This dataset have provided the minutely average raw PV data, so users do not need to take rolling average during data processing to get the minutely average data. This is the case in previous published work (Sun et al., 2018a, 2019) as we used the instantaneous raw PV data.

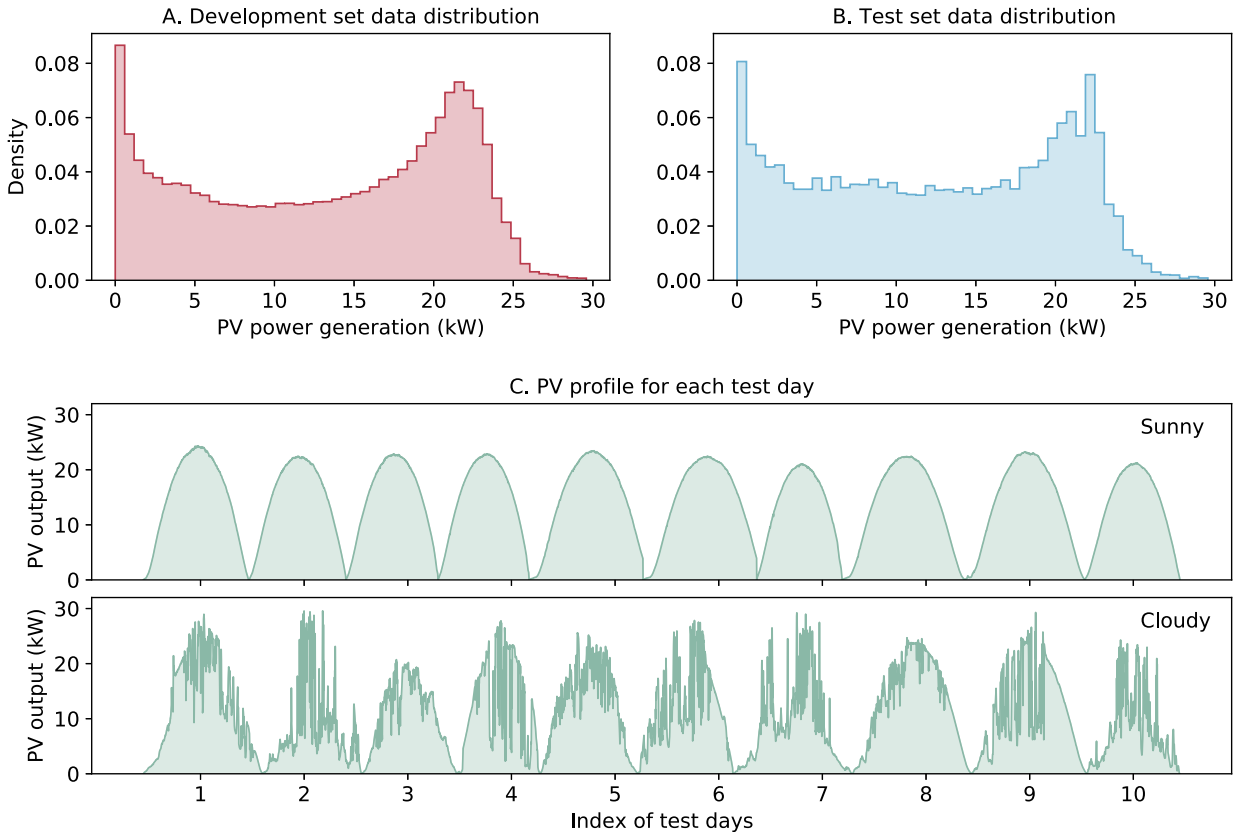

Fig. 4.1. The PV power generation data distribution of the benchmark dataset: A. development set PV data distribution; B. test set PV data distribution; and C. the PV power generation profiles of the 10 sunny days and 10 cloudy days used in the test set: upper panel shows the sunny days, and the lower panel is for the cloudy days.

(a) Down-sample the high resolution image frames into low resolution (64 × 64) for sake of saving training time. Down-sampling rate is adjustable.
(b) Filter out erroneously repeating images caused by the occasional abnormal behavior of OpenCV video decoder FFmpeg.

We further partition the processed dataset into development set (including data for training and validating the model) and test set. The test set includes 10 sunny days and 10 cloudy days selected manually across 2017 to 2019; the rest of the data goes into the development set. The number of samples contained in the model development set and the test set are 349,372 and 14,003, respectively, roughly a 96%:4% split. It should be noted that we have updated the test days from our previous publications given that missing data points were observed in certain days of the old test set. For comparison with the results from our previous publications, we encourage the users to re-organize the data by themselves. We hope this processed dataset can be used as a standardized benchmark for model training, evaluating and comparison in the solar forecasting community.

4. Benchmark dataset for image-based solar forecasting using deep learning

The benchmark dataset ($\mathscr{D}$) contains the model development set ($\mathscr{D}_d$) and test set ($\mathscr{D}_t$) obtained from the data processing steps described in 3. The samples of the benchmark dataset are organized as aligned pairs of sky images ($\mathcal{I}$) and PV power generation ($\mathcal{P}$), i.e., $\mathscr{D} = \{(\mathcal{I}_i, \mathcal{P}_i) \mid i \in \mathbb{Z} : 1 \le i \le N\}$, where $N = 363,375$ is the total number of samples in the benchmark dataset. Fig. 4.1 shows the distribution of the PV power generation data for the development set and test set and the PV power generation profiles of the 20 days in the test set. The statistics of the 20 test days are listed in Table 4.1.

Table 4.1
Statistics of the 10 sunny and 10 cloudy days used in the test set.

Date	Index	Mean (kW)	Max (kW)	Std (kW)
2017–09–15	Sunny_1	15.23	24.35	7.83
2017-10-06	Sunny_2	14.65	22.44	7.03
2017-10-22	Sunny_3	15.17	22.86	6.97
2018-02-16	Sunny_4	15.28	22.89	6.86
2018-06-12	Sunny_5	14.69	23.45	7.60
2018-06-23	Sunny_6	14.30	22.48	7.32
2019-01-25	Sunny_7	14.08	21.07	6.26
2019-06-23	Sunny_8	13.31	22.45	7.80
2019-07-14	Sunny_9	13.72	23.27	7.98
2019-10-14	Sunny_10	13.56	21.22	6.67
2017-06-24	Cloudy_1	12.39	28.95	8.30
2017-09-20	Cloudy_2	7.55	29.57	6.88
2017-10-11	Cloudy_3	10.64	20.72	6.04
2018-01-25	Cloudy_4	12.39	27.77	8.08
2018-03-09	Cloudy_5	12.45	25.60	7.07
2018-10-04	Cloudy_6	11.83	27.82	6.76
2019-05-27	Cloudy_7	8.62	29.22	7.39
2019-06-28	Cloudy_8	13.25	24.70	7.68
2019-08-10	Cloudy_9	11.14	29.28	7.48
2019-10-19	Cloudy_10	7.71	24.28	6.19

5. Sample uses of the benchmark dataset

In this section, we demonstrate two use cases of the benchmark dataset based on our published works. Our group has developed a specialized convolutional neural network (CNN) model named SUNSET (Stanford University Neural Network for Solar Electricity Trend) for

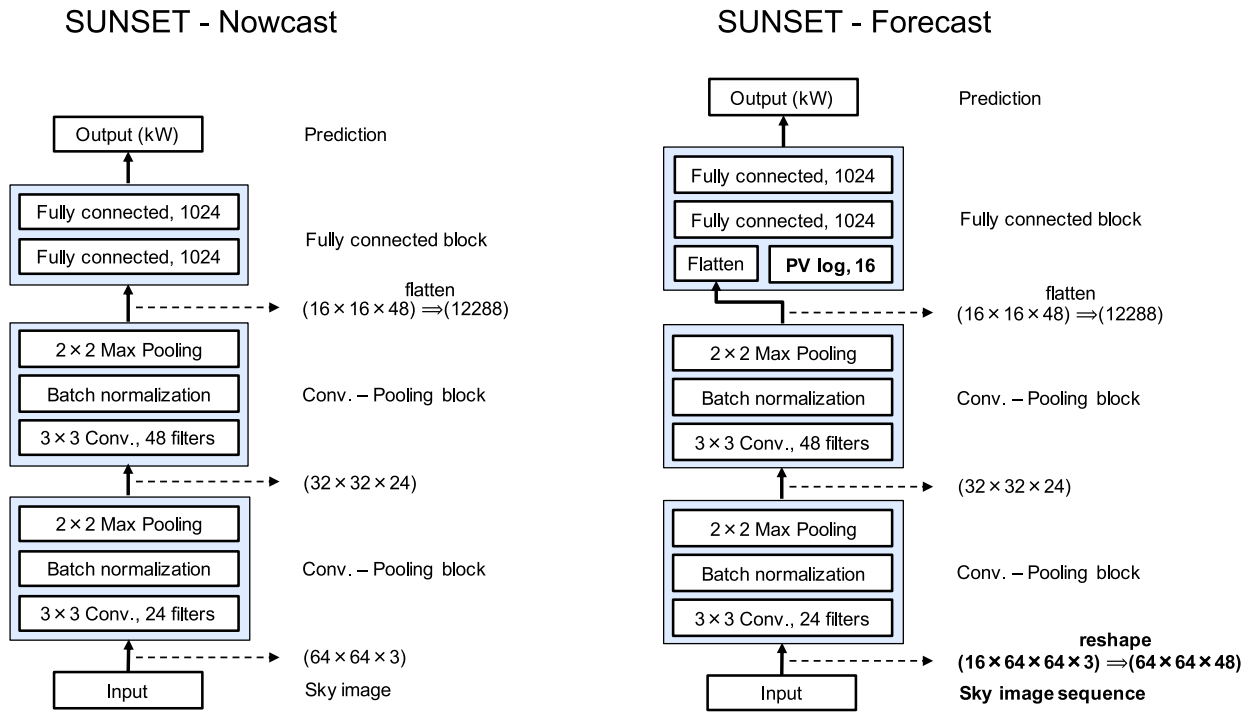

Fig. 5.1. The model architectures for PV power generation nowcast and forecast (differences between the two model architectures are highlighted in bold font). *Source:* Adapted from Nie et al. (2021), used with permission.

PV power generation prediction. Two specific prediction tasks were investigated based on SUNSET, including (1) PV power generation nowcast (Sun et al., 2018a), i.e., given a sky image, predicting the contemporaneous PV output; and (2) PV power generation forecast (Sun et al., 2019), given sky images and PV output for the past 15 min on 1 min resolution, predicting PV output 15 min ahead into the future. Fig. 5.1 shows the model architectures for the nowcast and forecast tasks. The details of these two models can be found in the corresponding published papers, and we demonstrate the use cases based on the setups described in these papers. We implemented these two deep learning models using TensorFlow 2.x, which is an update from the TensorFlow 1.x code base used in our previous publications. The new code base is included along with the dataset as described in Appendix B.

The performance of each model is evaluated by the following metrics: mean absolute error (MAE), root mean squared error (RMSE), and additionally, forecast skill (FS) for evaluating the SUNSET forecast model, which is computed based on the RMSE and is relative to the persistence model. The detailed method for calculating the FS can be found in Sun et al. (2019). The codes for model evaluation are included in the model implementation codes (see Appendix B). It should be noted that the goal here is to demonstrate the value of the data by providing some sample use cases, rather than evaluating specific methodologies. Therefore, our analysis on the sample results will mostly be high-level.

5.1. PV power generation nowcast

The PV power generation nowcast task is essentially learning a mapping (f_N) from sky images to the simultaneous PV power generation. An analog one can think of is the computer vision task estimating the age of people based on their facial images.

$$f_N : \mathcal{I}_i \mapsto \mathcal{P}_i, \text{where } \{\mathcal{I}_i, \mathcal{P}_i\} \in \mathcal{D} \tag{1}$$

We developed the nowcast model based on $\mathcal{D}_d$ and evaluated the model based on $\mathcal{D}_t$. During the model development phase, ten-fold cross-validation is employed, and during the test phase, the prediction is represented by the ensemble mean of the predictions from the 10 submodels. Table 5.1 shows the performance of the SUNSET nowcast model on the test set evaluated by RMSE and MAE. Fig. 5.2 shows the predictions of the SUNSET nowcast model on each of the test days compared with the ground truth.

Table 5.1
The performance of SUNSET nowcast and forecast models evaluated by common error metrics.

Model	Test set	RMSE (kW)	MAE (kW)	Forecast skill (%)
SUNSET Nowcast	Sunny days	0.80	0.66	–
	Cloudy days	3.34	2.34	–
	Overall	2.43	1.50	–
SUNSET Forecast	Sunny days	0.61	0.50	−45.80
	Cloudy days	4.27	2.95	17.03
	Overall	3.03	1.71	16.44

The results show that the SUNSET nowcast model can effectively extract the information in the sky images and correlate it with the local PV panel generation. It can well approximate the sun angle equations in the sunny days with clear sky conditions and reasonably estimate the states of PV power generation under different cloudy conditions. The potential use of the nowcast model is to serve as an alternative to the traditional sensor measurement of solar irradiance or PV panel power output, which is generally expensive. A similar study by Jiang et al. (2020) has also examined the potential of using an end-to-end CNN model to estimate the state of solar irradiance.

5.2. PV power generation forecast

The PV power generation forecast task can be mathematically described as learning a mapping (f_F) from historical sky image and PV power generation sequences to the future PV power generation.

$$f_F : (\mathcal{I}, \mathcal{P})_{t_i-H:\delta:t_i} \mapsto \mathcal{P}_{t_i+T}, \text{where } \{(\mathcal{I}, \mathcal{P})_{t_i-H:\delta:t_i}, \mathcal{P}_{t_i+T}\} \in \tilde{\mathcal{D}} \tag{2}$$

Here, we define forecast dataset $\tilde{\mathcal{D}}$ with $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ and $\tilde{\mathcal{D}} = \{[(\mathcal{I}, \mathcal{P})_{t_i-H:\delta:t_i}, \mathcal{P}_{t_i+T}] \mid i \in \mathbb{Z} : 1 \le i \le M\}$, where H is the length

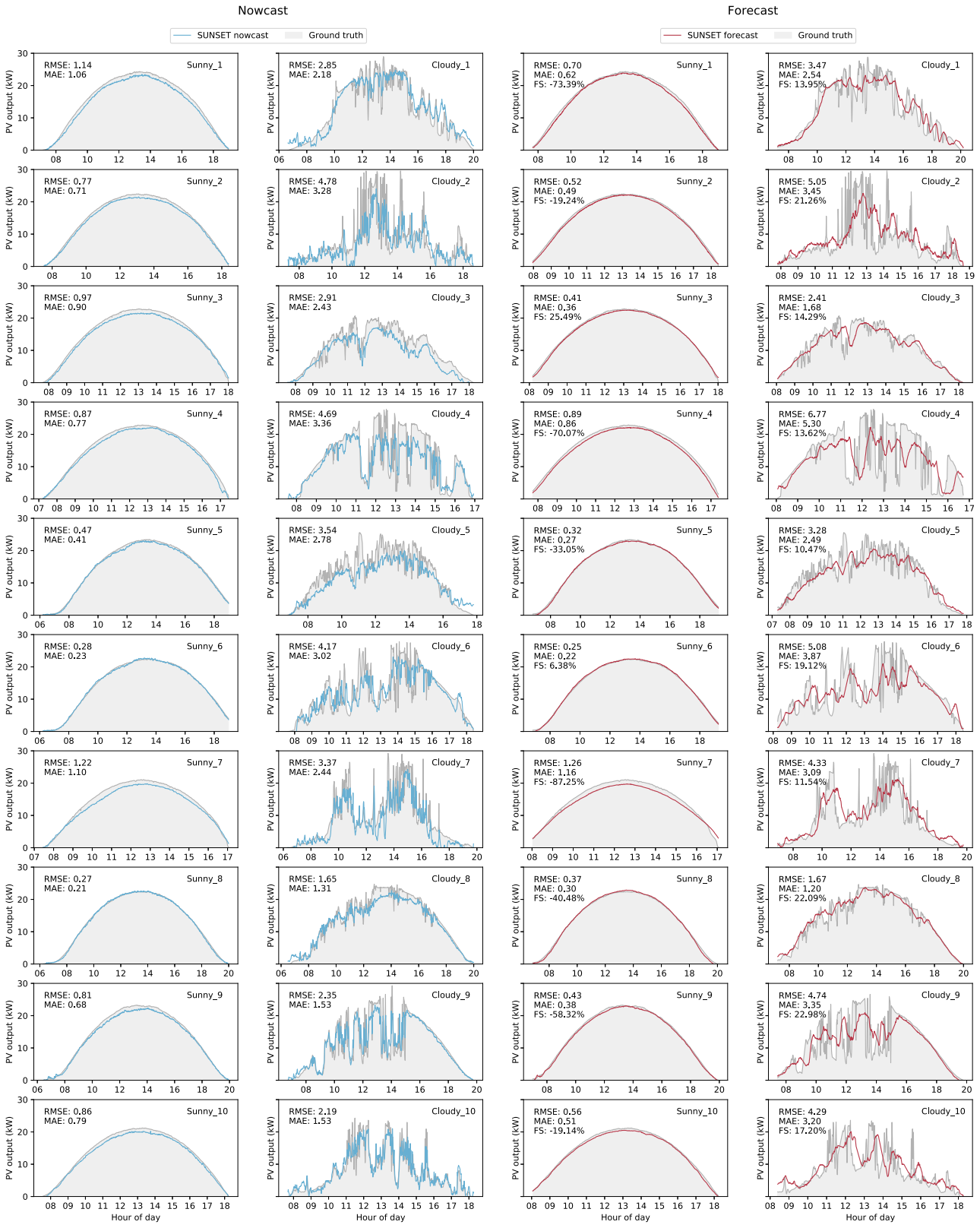

Fig. 5.2. SUNSET nowcast and forecast predictions on each of the test days. The first and second columns represent the nowcast prediction, and the third and fourth columns represent the forecast prediction. The test days' indices and the error metrics are shown on the right and left of each panel respectively.

of historical terms, δ is the interval between historical terms, T is the forecast horizon, and M is the total number of samples in $\tilde{\mathscr{D}}$. To build the forecast dataset $\tilde{\mathscr{D}}$ for demonstration, we follow the parameters used in the baseline of Sun et al. (2019), namely, $H = T = 15$ min, $\delta = 1$ min, and a sampling frequency of 2 min (i.e., $t_{i+1} - t_i = 2$) to obtain valid samples from $\mathscr{D}$, which leads to 161,928 samples in total for the forecast dataset. Following the definition by Yang (2019, 2020) to formalize the description of a solar forecasting task,

Table A.1
List of available files in the data repository.

File	Type	Size	Description
2017_2019_images_pv_processed.hdf5	Benchmark data	4.16 GB	A file-directory like structure consisting of two groups: ''trainval'' and ''test'', for storing model development set and test set, respectively, with each group containing two datasets: ''images_log'' and ''pv_log'', which stores the processed images stored as 8-bit RGB (256 color levels) data and PV generation data in the unit of kW from all three years stored in Python NumPy array format.
times_trainval.npy	Benchmark data	10.8 MB	Python NumPy array of time stamps stored in Python *datetime* format corresponding to development set in *.hdf5* file.
times_test.npy	Benchmark data	434 KB	Python NumPy array of time stamps stored in Python *datetime* format corresponding to test set in *.hdf5* file.
{Year}_{Month}_videos.tar	Raw data	2017: 500 GB; 2018: 640 GB; 2019: 449 GB	Tar archives with daytime 2048 × 2048 sky videos (*.mp4*) recorded at 20 frames per second for each month from 2017/03 to 2019/12.
{Year}_{Month}_images_raw.tar	Raw data	2017: 28.2 GB; 2018: 50.1 GB; 2019: 55.3 GB	Tar archives with daytime 2048 × 2048 sky images (*.jpg*) captured at 1 min intervals for each month from 2017/03 to 2019/12.
{Year}_pv_raw.csv	Raw data	2017: 16.7 MB; 2018: 16.6 MB; 2019: 13.7 MB	One-min PV generation data for the year 2017, 2018 and 2019. The unit of PV generation data is kW.

Table B.1
List of available code files.

File	Type	Description
data_preprocess_snapshot_only.ipynb	Data processing code	Jupyter Notebook used to capture images from the video stream at designated frequency.
data_preprocess_pv.ipynb	Data processing code	Jupyter Notebook used to process the raw PV power generation history.
data_preprocess_nowcast.ipynb	Data processing code	Jupyter Notebook used to down-sample the image frames, filter out the invalid frames and match images with the concurrent PV data, and partition model development and testing sets.
data_preprocess_forecast.ipynb	Data processing code	Jupyter Notebook used to generate valid samples for the forecast task.
SUNSET_nowcast.ipynb	Model code	Jupyter Notebook used to implement the SUNSET nowcast model to correlate PV output to contemporaneous images of the sky, including model training, validation and testing.
SUNSET_forecast.ipynb	Model code	Jupyter Notebook used to implement the SUNSET forecast model to predict 15 min ahead minutely-averaged PV output, including model training, validation and testing.
Relative_op_func.py	Helper function code	Helper functions for calculating theoretical PV power output under clear sky condition and the clear sky index.

ours can be described as $\{S^{1\,\mathrm{min}}, R^{1\,\mathrm{min}}, L^{15\,\mathrm{min}}, U^{2\,\mathrm{min}}\}$,[9] where S is forecast span, R is forecast resolution, and L is forecast lead time and U is forecast submission update rate. A graphic illustration of these temporal parameters can be found in Yang et al. (2019). For consistent comparisons with each other's model, we encourage the users to clearly state their forecasting setups.

The forecast dataset $\tilde{\mathscr{D}}$ is further separated into model development set $\tilde{\mathscr{D}}_d$ and test set $\tilde{\mathscr{D}}_t$ based on the same partition as the benchmark development set $\mathscr{D}_d$ and test sets $\mathscr{D}_t$. The processing codes we used are also open-sourced (see Appendix B) and users can either modify these parameters in our processing code based on their model input and output configurations or develop their own processing codes to build the forecast dataset. We trained the forecast model with 10-fold cross-validation, and during the test phase, the prediction is represented by the ensemble mean of the predictions from the 10 submodels. Table 5.1 shows the performance of the SUNSET forecast model on the test set evaluated by RMSE, MAE, and FS. Fig. 5.2 shows the predictions of the SUNSET forecast model on each of the test days compared with the ground truth.

The results show that the SUNSET forecast model can well predict the 15 min ahead future power generation on the sunny days but it

[9] For training and validation of the forecast model, we used a forecast submission update rate of 2 min ($U^{2\,\mathrm{min}}$) to speed up the model development while retaining a good model performance based on the findings by Sun et al. (2019). [Note that the same term is called sampling frequency in the work by Sun et al.] For model testing, we used a forecast submission update rate of 1 min ($U^{1\,\mathrm{min}}$).

struggles somehow on the cloudy days. In cloudy conditions, SUNSET tends to avoid large fluctuations by predicting conservatively towards the average. It helps the model in capturing the general trend, however, it usually fails in predicting the right magnitudes of the peaks and dips, which leads to one of the major prediction errors. Another type of error is associated with temporal lags in prediction. The model has a larger tendency of following the trend of the past observations than actively anticipating future events. These two aspects of errors direct to the following future research directions: (1) generating probabilistic prediction rather than point prediction to deal with the uncertainty in the sudden changes of PV power generation; (2) developing models for better prediction of the cloud movement to deal with the systematic temporal lags in forecasting. Compared with the persistence model, SUNSET outperforms it by 16.4% overall in FS on all of the test days and by 17% on cloudy days, although it under-performs the persistence model by 46% on sunny days. To this end, using different types of models for different sky conditions could boost the accuracy of the forecast (Nie et al., 2020).

6. Conclusion

We introduced a curated dataset named SKIPP'D with the goal of providing a standardized benchmark for the solar forecasting community to evaluate and compare different solar forecasting models as well as to facilitate the research of image-based solar forecasting using deep learning and other related areas such as sky image segmentation, cloud type classification and cloud movement forecasting. Two levels of data are provided in this dataset including a processed benchmark dataset containing 1 min down-sampled sky images and PV output pairs and a raw dataset containing high-resolution sky videos, images and PV power generation data. We have also detailed the data processing steps used to obtain the benchmark dataset and demonstrated two use cases based on the benchmark dataset for users' reference and future studies.

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.

Acknowledgments

This research was supported by the Dubai Electricity and Water Authority (DEWA) through their membership in the Stanford Energy Corporate Affiliates (SECA) program. The authors thank Jacques de Chalendar from Stanford University who helps provide access to the PV power generation history data. The authors would also like to acknowledge the Stanford Research Computing Center for providing the computational resources to conduct the experiments in this study. The authors are also grateful to Amy Hodge from Science and Engineering Resource Group at Stanford Libraries for facilitating the datasets depositing.

Appendix A. Data repository

The benchmark data is available at https://purl.stanford.edu/dj417rh1007and the raw data is deposit by year given its large size. The 2017 raw data is available at https://purl.stanford.edu/sm043zf7254 and the links to 2018 and 2019 data can be found in the "Related items" elsewhere on the same web page. All datasets are licensed under CC-BY-4.0. Table A.1 shows the list of all data files included in this dataset. Although every effort was made to ensure the quality of the data, no guarantees are implied by the authors of the dataset.

Appendix B. Code base

As part of the data release, we include the data processing and baseline model implementation code written in Python 3.6, which can be found on the Github Repository https://github.com/yuhao-nie/Stanford-solar-forecasting-dataset. All code files are licensed under the MIT license. Table B.1 shows the list of all the code files included along with this dataset. Users can either use the reference codes we provided here or customize their own data processing pipeline.

References

Augustine, John A., DeLuisi, John J., Long, Charles N., 2000. SURFRAD–A national surface radiation budget network for atmospheric research. Bull. Am. Meteorol. Soc. 81 (10), 2341–2358.

Bassous, Guilherme Fonseca, Calili, Rodrigo Flora, Barbosa, Carlos Hall, 2021. Development of a low-cost data acquisition system for very short-term photovoltaic power forecasting. Energies 14 (19), 6075.

Dissawa, Lasanthika H, Godaliyadda, Roshan I, Ekanayake, Parakrama B, Agalgaonkar, Ashish P, Robinson, Duane, Ekanayake, Janaka B, Perera, Sarath, 2021. Sky image-based localized, short-term solar irradiance forecasting for multiple PV sites via cloud motion tracking. Int. J. Photoenergy 2021.

Feng, Cong, Yang, Dazhi, Hodge, Bri Mathias, Zhang, Jie, 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.

Feng, Cong, Zhang, Jie, 2020. SolarNet: A sky image-based deep convolutional neural network for intra-hour solar forecasting. Sol. Energy 204 (April), 71–78. http://dx.doi.org/10.1016/j.solener.2020.03.083.

Feng, Cong, Zhang, Jie, Zhang, Wenqi, Hodge, Bri Mathias, 2022. Convolutional neural networks for intra-hour solar forecasting based on sky image sequences. Appl. Energy 310, 118438. http://dx.doi.org/10.1016/J.APENERGY.2021.118438.

Jiang, Huaiguang, Gu, Yi, Xie, Yu, Yang, Rui, Zhang, Yingchen, 2020. Solar irradiance capturing in cloudy sky days–A convolutional neural network based image regression approach. IEEE Access 8, 22235–22248. http://dx.doi.org/10.1109/ACCESS.2020.2969549, URL https://ieeexplore.ieee.org/document/8970273/.

Kottek, Markus, Grieser, Jürgen, Beck, Christoph, Rudolf, Bruno, Rubel, Franz, 2006. World map of the Köppen-Geiger climate classification updated. Meteorol. Z. 15 (3), 259–263. http://dx.doi.org/10.1127/0941-2948/2006/0130.

Nie, Yuhao, Li, Xiatong, Paletta, Quentin, Aragon, Max, Scott, Andea, Brandt, Adam, 2022a. Open-source ground-based sky image datasets for very short-term solar forecasting, cloud analysis and modeling: A comprehensive survey. arXiv preprint arXiv:2211.14709.

Nie, Yuhao, Paletta, Quentin, Scotta, Andea, Pomares, Luis Martin, Arbod, Guillaume, Sgouridis, Sgouris, Lasenby, Joan, Brandt, Adam, 2022b. Sky-image-based solar forecasting using deep learning with multi-location data: training models locally, globally or via transfer learning? arXiv preprint arXiv:2211.02108.

Nie, Yuhao, Sun, Yuchi, Chen, Yuanlei, Orsini, Rachel, Brandt, Adam, 2020. PV power output prediction from sky images using convolutional neural network: The comparison of sky-condition-specific sub-models and an end-to-end model. J. Renew. Sustain. Energy 12 (4), 046101. http://dx.doi.org/10.1063/5.0014016, URL http://aip.scitation.org/doi/10.1063/5.0014016.

Nie, Yuhao, Zamzam, Ahmed S., Brandt, Adam, 2021. Resampling and data augmentation for short-term PV output prediction based on an imbalanced sky images dataset using convolutional neural networks. Sol. Energy 224 (May), 341–354. http://dx.doi.org/10.1016/j.solener.2021.05.095.

Ntavelis, Evangelos, Remund, Jan, Schmid, Philipp, 2021. SkyCam: A dataset of sky images and their irradiance values. http://dx.doi.org/10.48550/ARXIV.2105.02922.

Paletta, Quentin, Arbod, Guillaume, Lasenby, Joan, 2021a. Benchmarking of deep learning irradiance forecasting models from sky images – An in-depth analysis. Sol. Energy 224, 855–867. http://dx.doi.org/10.1016/j.solener.2021.05.056, URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X21004266.

Paletta, Quentin, Hu, Anthony, Arbod, Guillaume, Lasenby, Joan, 2021b. ECLIPSE : Envisioning cloud induced perturbations in solar energy. URL http://arxiv.org/abs/2104.12419.

Pedro, Hugo T.C., Larson, David P., Coimbra, Carlos 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.

Stoffel, T., Andreas, A., 1981. NREL Solar Radiation Research Laboratory (SRRL): Baseline Measurement System (BMS); Golden, Colorado (Data). National Renewable Energy Lab.(NREL), Golden, CO (United States).

Sun, Yuchi, 2019. Short-term Solar Forecast Using Convolutional Neural Networks with Sky Images (Ph.D. thesis). Stanford University, URL http://purl.stanford.edu/fm704js1179.

Sun, Yuchi, Szűcs, Gergely, Brandt, Adam R., 2018a. Solar PV output prediction from video streams using convolutional neural networks. Energy Environ. Sci. 11 (7), 1811–1818. http://dx.doi.org/10.1039/C7EE03420B, URL http://xlink.rsc.org/?DOI=C7EE03420B.

Sun, Yuchi, Venugopal, Vignesh, Brandt, Adam R., 2018b. Convolutional neural network for short-term solar panel output prediction. In: 2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC)(a Joint Conference of 45th IEEE PVSC, 28th PVSEC & 34th EU PVSEC). IEEE, pp. 2357–2361.

Sun, Yuchi, Venugopal, Vignesh, Brandt, Adam R., 2019. Short-term solar power forecast with deep learning: Exploring optimal input and output configuration. Sol. Energy 188, 730–741. http://dx.doi.org/10.1016/j.solener.2019.06.041, URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X19306164.

Terrén-Serrano, Guillermo, Bashir, Adnan, Estrada, Trilce, Martínez-Ramón, Manel, 2021. Girasol, a sky imaging and global solar irradiance dataset. Data in Brief 35, 106914.

Venugopal, Vignesh, Sun, Yuchi, Brandt, Adam R., 2019. Short-term solar PV forecasting using computer vision: The search for optimal CNN architectures for incorporating sky images and PV generation history. J. Renew. Sustain. Energy (ISSN: 1941-7012) 11 (6), 066102. http://dx.doi.org/10.1063/1.5122796, URL http://aip.scitation.org/doi/10.1063/1.5122796.

Yang, Dazhi, 2018. SolarData: An r package for easy access of publicly available solar datasets. Sol. Energy 171, A3–A12.

Yang, Dazhi, 2019. A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES). J. Renew. Sustain. Energy (ISSN: 19417012) 11 (2), http://dx.doi.org/10.1063/1.5087462.

Yang, Dazhi, 2020. Comment: Operational aspects of solar forecasting. Sol. Energy (ISSN: 0038092X) 210 (February), 38–40. http://dx.doi.org/10.1016/j.solener.2020.04.014.

Yang, Dazhi, Wu, Elynn, Kleissl, Jan, 2019. Operational solar forecasting for the real-time market. Int. J. Forecast. (ISSN: 01692070) 35 (4), 1499–1519. http://dx.doi.org/10.1016/j.ijforecast.2019.03.009.

Zhang, Jinsong, Verschae, Rodrigo, Nobuhara, Shohei, Lalonde, Jean François, 2018. Deep photovoltaic nowcasting. Sol. Energy 176 (September), 267–276. http://dx.doi.org/10.1016/j.solener.2018.10.024.