# A comprehensive dataset for the accelerated development and benchmarking of solar forecasting methods

Cite as: J. Renewable Sustainable Energy **11**, 036102 (2019); https://doi.org/10.1063/1.5094494
Submitted: 02 March 2019 . Accepted: 21 May 2019 . Published Online: 24 June 2019

Hugo T. C. Pedro, David P. Larson, and Carlos F. M. Coimbra

Journal of Renewable and Sustainable Energy ARTICLE scitation.org/journal/rse

# A comprehensive dataset for the accelerated development and benchmarking of solar forecasting methods

Cite as: J. Renewable Sustainable Energy **11**, 036102 (2019); doi: 10.1063/1.5094494
Submitted: 2 March 2019 · Accepted: 21 May 2019 ·
Published Online: 24 June 2019



Hugo T. C. Pedro,[a] David P. Larson,[a] and Carlos F. M. Coimbra[b]

## AFFILIATIONS

Department of Mechanical and Aerospace Engineering and Center for Energy Research University of California San Diego, La Jolla, California 92093, USA

**Note:** This paper is part of the Special Collection on Best Practices in Renewable Energy Resourcing and Integration.
[a]**Contributions:** H. T. C. Pedro and D. P. Larson contributed equally to this work.
[b]**Author to whom correspondence should be addressed:** ccoimbra@ucsd.edu

## ABSTRACT

We describe and release a comprehensive solar irradiance, imaging, and forecasting dataset. Our goal with this release is to provide standardized solar and meteorological datasets to the research community for the accelerated development and benchmarking of forecasting methods. The data consist of three years (2014–2016) of quality-controlled, 1-min resolution global horizontal irradiance and direct normal irradiance ground measurements in California. In addition, we provide overlapping data from commonly used exogenous variables, including sky images, satellite imagery, and Numerical Weather Prediction forecasts. We also include sample codes of baseline models for benchmarking of more elaborated models.



## I. INTRODUCTION

Solar forecasting is an enabling technology for the integration of weather-dependent, variable solar power generation into an electric grid.[1–3] Therefore, it is unsurprising that there has been strong interest in the subject over the past decade. However, despite the rapid growth, there are few standardized datasets for the development and benchmarking of solar forecasting methods. The lack of such datasets limits comparative analysis of forecasting methods and inhibits the rate of research progress,[3] particularly for those without the resources to deploy and maintain their own solar and meteorological instruments.

Recent efforts try to address this issue by facilitating the access to public data. To that end, Yang[4] provided an excellent open-source tool for easy access of publicly available solar datasets. Another important source for solar data is the National Solar Radiation Database (NSRDB). The NSRDB is managed by the National Renewable Energy Laboratory (NREL) and provides half-hourly values of satellite-derived irradiance that covers most of the USA. A comprehensive list of modeled and measured, historical and current solar resource data sources is available in a recent "Best Practices Handbook" published by the National Renewable Energy Laboratory (NREL).[5] The Handbook combines the knowledge of foremost experts in solar energy meteorology and disseminates best practices in solar resource assessment and forecasting. This paper shares the same goals and attempts to create a comprehensive dataset that can be used for solar irradiance forecasting. The motivation for this work stems from the fact that the available datasets are not enough to recreate and benchmark many of the latest forecasting models without substantial effort in data acquisition and data quality control. For instance, the most complete dataset currently available for forecasting benchmarking can be acquired from the SURFRAD network. It provides 1-min irradiance and sky images (on request). However, the sky images are captured using a Total Sky Imager (TSI) that produces low-resolution images and does not allow for a total view of the sky dome due to the presence of a black sun-blocking stripe. In summary, to the best of our knowledge, there are no publicly available datasets that contain the following:

1. Multiple years of quality controlled 1-min irradiance and weather data;
2. Colocated high-resolution sky images for the same time period;
3. Satellite images for the same target area and time interval;
4. Numerical Weather Prediction (NWP) data for the same target area and time interval.

Thus, the goal of this data release is twofold. First is to provide data for a region of high interest for solar forecasting, in this case, California's Central Valley, which has experienced continuous growth in terms of both population and solar generation. In order to fulfill this goal, the dataset includes endogenous and exogenous data necessary to benchmark the state-of-the-art in solar forecasting for intrahour to day-ahead horizons. Second is to present guidelines and an invitation for other researchers to release their own solar forecasting datasets, to the benefit of interested parties. Together, the hope is that the solar forecasting community will soon have a diverse range of datasets to leverage in their own work. These data and code releases can generate accelerated progress, similar to what has occurred to image classification methods with the release of datasets such as MNIST and CIFAR.

This work is organized as follows. Section II discusses the data sources that are used to create the dataset. Section III details the processing applied to the various data sources, including how features are extracted for use as inputs to the forecasting models. Section IV presents sample forecasts for intrahour, intraday, and day-ahead horizons, while Appendix B describes the format, intended use, and the conditions for proper use of the datasets. Finally, Sec. V summarizes this work and provides recommendations for future data releases.

## II. DATA SOURCES

Our research group has deployed a range of solar irradiance and meteorological instruments at sites throughout the West Coast of the United States. The sites span from Bellingham, Washington to San Diego, California, including one site on the Hawaiian island of Oahu. At each site, we installed one or more irradiance sensors to measure both global horizontal irradiance (GHI) and direct normal irradiance (DNI) at sample rates of 1-min or faster. In addition, we installed colocated fish-eye lens cameras to provide ground-based sky images at several of the locations. Measurements from each sensor were logged locally and then automatically transferred to our private servers, where the data were stored in MySQL databases and regularly backed up to external storage media. Over the past ten years, our lab has collected over several tens of terabytes of data, which have enabled a multitude of published solar forecasting studies.[6–21]

For this data release, our choice of data was driven by a combination of factors. First, the data should be from areas of interest for solar forecasting, i.e., areas with large amounts of pre-existing or planned solar power generation. Second, the data should span two or more years sequentially, to enable both the training and testing sets which are at least a year long each. Third, all data sources for the site should be of high quality, with minimal intervals of missing data and with quality control issues. Fourth, the data should include the most common exogenous inputs for solar forecasting, such as sky images, satellite imagery, and NWP forecasts.

Based on the above criteria, we select the Folsom, CA site (38.642°, −121.148°) for this data release. Folsom is a city in Sacramento County, in the California Central Valley (see Fig. 1), with a Csa (C = temperate climate s = dry summer a = hot summer) classification in the Köppen climate scheme. The instruments were mounted on the south roof of the headquarter building of the California Independent System Operator (CAISO) in 2012 (see Fig. 2). The primary components of the system are a Rotating Shadowband Radiometer (RSR) for the measurement of GHI, DNI, and diffuse horizontal irradiance (DHI), a fish-eye lens camera for sky images, and a Campbell Scientific CR1000 datalogger. Data were recorded at 1-min average rates for all instruments, with their internal clocks automatically synchronized with an on-site Network Time Protocol (NTP) server to ensure consistency.

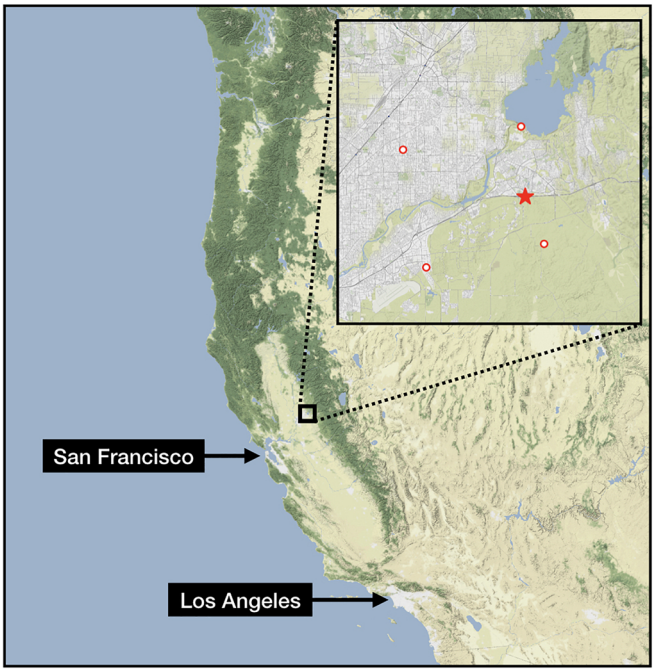


**FIG. 1.** Map of the West Coast of the United States, showing the area surrounding the Folsom, CA site (38.642°, −121.148°). The inset plot shows the location of the Folsom site (star marker), which lies directly south of Lake Folsom, as well as the four nearest NAM grid points (circle markers). The locations of San Francisco and Los Angeles are noted purely for the reference of the readers.

### A. Irradiance

The primary datasets for solar forecasting are the two main modes of solar irradiance, namely, GHI (global) and DNI (beam). These two variables are used to train the models and assess the forecasting performance.

The GHI and DNI data included in this data release are measured using a second-generation RSR (RSR-2) from Augustyn, Inc. The RSR-2 consists of a main shadowband head unit and two Licor LI-200SZ pyranometers, which have a typical error of ±5% compared to an Eppley Precision Spectral Pyranometer (PSP) (https://www.licor.com/). The first pyranometer provides a continuous measurement of GHI, while the second pyranometer and shadowband enable the measurement of DHI. DNI is computed directly from the GHI, DHI, and solar zenith angle ($\theta_z$). Comparisons against reference instrumentation over a 12-month period[22] show that the RSR-2 exhibits uncertainties ranging from −1.2% to 1.0% and −0.2% to 3.0% for GHI and DHI, respectively. The comparisons also showed that uncertainties increase for larger zenith angles and that differences above 5% are possible, especially in winter. This study concluded that the uncertainty levels for this instrument are in accordance with historical values reported in the literature for solar monitoring instruments.

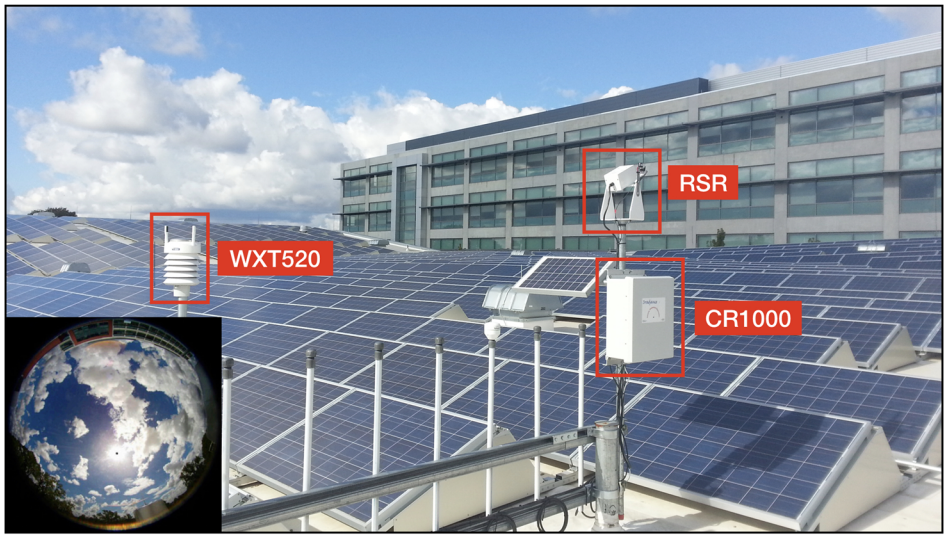


**FIG. 2.** Solar and meteorological instruments installed on the south roof of the CAISO headquarters in Folsom, CA, USA. The specific location of the instruments was chosen to prevent shading, with the building in the background located ∼60 m due north of the instruments and no obstructions located to the south. A RSR for GHI and DNI measurements, a Vaisala WXT520 for meteorological measurements (temperature, wind speed, etc.), and a Campbell Scientific CR1000 datalogger for storing the data locally prior to transmission to our servers at UC San Diego are shown. A sample sky image from the colocated fish-eye lens camera (out of frame) is shown in the bottom left.

### B. Sky images

Ground-based sky images are a standard exogenous input for intrahour forecasts. These images provide high resolution information (spatial and temporal) about clouds that determine the solar irradiance. Usually, sky images are explored under two distinct frameworks: physics-based models or data-driven models. The first framework is more popular[7,14,23,24] and typically follows a well-defined flow chart: (i) differentiation between clear-sky pixels and cloudy pixels, (ii) cloud classification and cloud optical depth determination, and (iii) determination of cloud motion and cloud advection. When multiple sky cameras are available,[25] the physical-based models may also include the calculation of the cloud height and cloud shadow tracking.

The data-driven approach relies on the extraction of image features that are then used as predictors in machine learning algorithms.[17,26] This strategy has seen an increase in popularity in recent times due to the maturity of tools such as convolution neural networks.[27]

As mentioned above, the absence of a common dataset for different developers makes it difficult to properly evaluate competing sky-image algorithms. Thus, in this paper, we provide sky images obtained using a sky camera colocated with the irradiance sensors. The sky camera captures Red-Green-Blue (RGB) color images at a medium resolution (1536 × 1536 pixels), at intervals of 1-min.

### C. Satellite imagery

Satellite imagery is helpful when forecasting over horizons of one to several hours ahead.[20,28,29] The Geostationary Operational Environmental Satellite (GOES) system is a set of geosynchronous satellites, denoted GOES-West and GOES-East, which provide a range of remote sensing measurements over the Western Hemisphere. Due to the location of the site, we are including images from GOES-15, which was operated as GOES-West from 2011 until February 2019, when it was superseded by GOES-17. The Earth-facing imager on GOES-15 has five spectral bands: one visible band centered at 0.63 $\mu$m and four infrared bands centered at 3.9, 6.5, 10.7, and 13.3 $\mu$m. Following the previous literature,[20,28,29] we include measurements from the visible band (VIS), which has a spatial resolution of 1.0 km and a temporal resolution of one image every 30 min. It should be noted that satellite-derived cloud and irradiance products such as those provided by the Clouds from the Advanced Very High Resolution Radiometer (AVHRR)—Extended (CLAVR-x) code package developed by the National Oceanic and Atmospheric Administration (NOAA) are also valuable predictors for irradiance forecasting.[30,31] However, here, we limit the data release to visible and infrared images since they are the basis of many cloud identification and cloud advection algorithms for solar irradiance forecasting.

### D. Numerical weather prediction

For day-ahead horizons, Numerical Weather Prediction (NWP) models are the preferred exogenous input for solar forecasting. We have chosen to include forecasts from the North American Mesoscale Forecast System (NAM) due to its extensive presence in solar irradiance and power forecasts.[21,32,33] Other commonly used NWP models include the Global Forecast System (GFS), the European Center for Medium-Range Weather Forecast (ECMWF), Integrated Forecasting System (IFS),[34] and the High-Resolution Rapid Refresh (HRRR).[35–37] NAM provides forecasts 1–84 h ahead on a 0.11° grid (∼12 km) for the Continental United States (CONUS), generated four times per day: 00Z, 06Z, 12Z, and 18Z. Although selecting the NAM grid point closest to the site is the obvious choice for solar forecasting, previous studies have shown forecast improvement from considering a set of grid points around the target site.[32,38] Therefore, we have included NAM forecasts from the four nearest grid points, measured by their physical distance from the site (see Table I). For each of the four grid points, we extracted a range of relevant variables, which are summarized in Table II. Additional NWP data, from NAM and other NWP models, can be obtained through multiple data archives, e.g., the NOAA Operational Model Archive and Distribution System (NOMADS).

### E. Weather data

The solar instrumentation used to collect irradiance is complemented by a weather station that records the following data beyond

**TABLE I.** The coordinates of the four nearest NAM grid points, along with their distance and direction from the target site. The relative location of the grid points can be seen in Fig. 1.

| Latitude (°) | Longitude (°) | Distance (km) | Direction |
|---|---|---|---|
| 38.599891 | −121.126680 | 5.0 | North |
| 38.704328 | −121.152788 | 6.9 | Southwest |
| 38.579454 | −121.260320 | 12.0 | Southeast |
| 38.683880 | −121.286556 | 12.9 | Northeast |

the shortwave values of GHI, DHI, and DNI: ambient temperature, relative humidity, pressure, wind speed, wind direction, maximum wind speed, and precipitation. All variables, except maximum wind speed, are 1-min averages. The maximum wind speed is the maximum value measured in each 1-min window. The weather data are included in this data release although these additional variables are not used in the forecast benchmarks presented below.

## III. FEATURE ENGINEERING

In the Sec. II, we described the primary data that are provided in this paper. With these datasets, one can replicate many of the studies presented in the solar energy literature. We could, at this point, conclude that the goal of providing a comprehensive dataset to solar forecasting has been achieved. However, we opt to provide a secondary dataset with data derived from primary sources. In this way, we illustrate common techniques for data preprocessing and feature extraction from time series data and sky images.

### A. Irradiance

Features engineered from irradiance data use the clear-sky index, thus removing deterministic daily and seasonal variations in the data. The clear-sky index time series is defined as $k_t = I/I_{cs}$, where $I$ denotes GHI or DNI and $I_{cs}$ is the respective clear-sky irradiance. The clear-sky model used in this case is the popular Ineichen and Perez model[39] that parameterizes irradiance in terms of the Linke turbidity. Linke turbidity is estimated from monthly climatological values[40] which were created based on the algorithm proposed by Remund *et al.*[41]

Once $k_t$ is computed, three features are engineered from the time series within a processing window that precedes the forecasting issuing time $t$

- Backward average for the clear-sky index time series: for a given time stamp $t$, this feature is given by the vector $\mathbf{B}(t)$ with components

$$B_i(t) = \frac{1}{N} \sum_{t \in ]t - i\delta - T, t - T]} k_t(t), \quad i = \{1, 2, \ldots, M\}. \tag{1}$$

- Lagged average values for the clear-sky index time series: this feature is given by the vector $\mathbf{L}(t)$ with components

$$L_i(t) = \frac{1}{N} \sum_{t \in ]t - i\delta - T, t - (i-1)\delta - T]} k_t(t), \quad i = \{1, 2, \ldots, M\}. \tag{2}$$

- The clear-sky index variability: this feature is given by the vector $\mathbf{V}(t)$ with components

$$V_i = \sqrt{\frac{1}{N} \sum_{t \in ]t - i\delta - T, t - T]} \Delta k_t(t)^2}, \quad i = \{1, 2, \ldots, 12\}, \tag{3}$$

where $\Delta k_t(t) = k_t(t) - k_t(t - \Delta t)$.

In these equations, $\delta$ is a minimum window size, $N$ is the number of data points in the processing window, $t - T$ is the rightmost edge of the processing window, and $M$ is the number of processing windows to consider. The parameters $\delta$, $T$, and $M$ depend on the forecast horizon: $\delta = \{5, 30, 60\}$ min, $T = \{0, 0, 8\}$ h, and $M = \{6, 6, 12\}$ for the intrahour, intraday, and day-ahead forecasts, respectively.

Figure 3 shows, in the left panel, the clear-sky index for GHI and DNI in a six-hour period on 2014–03-14. The panels on the right show the DNI features computed at four distinct instances indicated by the vertical bars in the left panel. These features encode information about the past behavior of the DNI time-series.

### B. Sky images

In this section, we describe features derived from sky images. The features used here are computed from all sky-dome pixels, that is, all pixels that do not correspond to ground or obstacles. The 8-bit color data from the selected pixels are then flattened into floating point vectors **r**, **g**, and **b**, for the red, green, and blue channels, respectively. Two additional vectors are computed from **r** and **b**: the red-to-blue ratio $\boldsymbol{\rho}$ with components $\rho_i = r_i/b_i$ and the normalized red-to-blue ratio $\boldsymbol{\eta}$ with components $\eta_i = (r_i - b_i)/(r_i + b_i)$.

For each one of these five vectors, three features are calculated

- Average

**TABLE II.** List of extracted NAM variables.

| Variable | NAM name | Description | Units |
|---|---|---|---|
| Pressure | `PRES: surface` | Surface pressure | Pa |
| Temperature | `TMP: surface` | Surface temperature | K |
| Relative humidity | `RH: 2 m above ground` | Relative humidity 2 m above ground | % |
| U-wind | `UGRD: 10 m above ground` | U-component of wind 10 m above ground | $ms^{-1}$ |
| V-wind | `VGRD: 10 m above ground` | V-component of wind 10 m above ground | $ms^{-1}$ |
| Precipitation | `APCP: surface` | Total precipitation | $kg/m^2$ |
| GHI | `DSWRF: surface` | Downward short-wave radiation flux | $W/m^2$ |
| Cloud cover | `TCDC: entire atmosphere` | Total cloud cover | % |

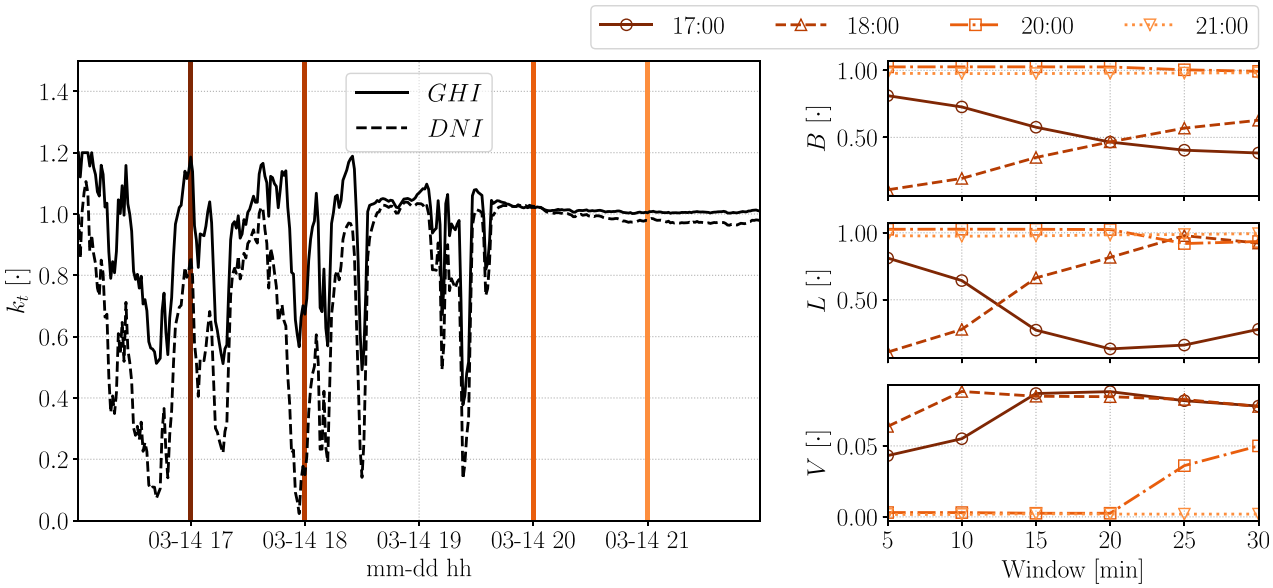


**FIG. 3.** Irradiance time series features. Left: $k_t$ values for GHI and DNI in a six-hour period in 2014-03-14 at Folsom, CA. Right: the three panels on the right show DNI features computed at the times indicated by the vertical lines on the left panel. From top to bottom: backward average, lagged data, and variability.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} v_i. \tag{4}$$

- Standard deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(v_i - \mu)^2}, \tag{5}$$

- and entropy

$$e = -\sum_{\substack{i=1 \\ p_i \neq 0}}^{N_B} p_i \log_2(p_i), \tag{6}$$

where $v_i$ represents one of the five vectors, $N$ is the number of elements in the vector, and $p_i$ is the relative frequency for the $i$th bin (out of $N_B = 100$ bins evenly spaced. These features are computed for all the images in the dataset, yielding a total of 15 features per image (three metrics × five color data vectors). The features are shown in Fig. 4 for the same seven-hour period as in Fig. 3. The figure's top panel shows nine sky images in this period, and the bottom three panels show the features for all images within the time frame.

### C. Satellite imagery

Satellite imagery can be processed using the algorithm described in Sec. III B. However, here, we present a simpler approach that has been used in previous studies.[20] In this case, for each image, we crop a $w \times w$ region (with $w = 10$ in this case), centered around the target site, and then flatten it into a vector of length $n = w^2$. The result is a time-series of $m$ samples, with $n$ variables per sample, excluding the timestamp.

## IV. SAMPLE FORECASTS

Finally, in this section, we discuss common techniques to make use of the primary and secondary data for the purpose of solar irradiance forecasting. We present sample forecasts for three common forecast horizons: intrahour, intraday, and day-ahead, using the GHI and DNI measurements from the RSR as the ground truth. Rather than an exhaustive study of forecast methods, we evaluate a subset of methods, which were chosen for their ease of implementation and interpretation. In addition, we fully expect future studies to achieve forecast performance beyond these reference models, which should be considered as lower bounds.

### A. Forecast models

In this case, the forecast follows the mathematical formulation

$$\hat{I}_\Delta(t_o + \tau) = r^m(t_o) I_{cs,\Delta}(t_o + \tau), \tag{7}$$

where $\Delta$ indicates the data aggregation, $\tau$ the forecasting horizon, and $t_o$ the forecasting issuing time. These parameters vary depending on the type of forecast as shown in Table III. Note that the day-ahead forecasts are issued once daily at 12:00 UTC (4 a.m. PST) to be compatible with the CAISO Day-Ahead Market (DAM) submission time.

The correction factor $r^m$ depends on the forecasting algorithm used. For the smart persistence model that uses the latest $k_t$ value available, this factor is given by

$$r^{sp}(t_o) = \frac{I_\Delta(t_o)}{I_{cs,\Delta}(t_o)} = k_{t,\Delta}(t_o). \tag{8}$$

In the forecasting implementation for intrahour and intraday, the latest $k_t$ value used in Eq. (8) is the one given by the backward average over the shortest window (5 and 30 min, respectively).

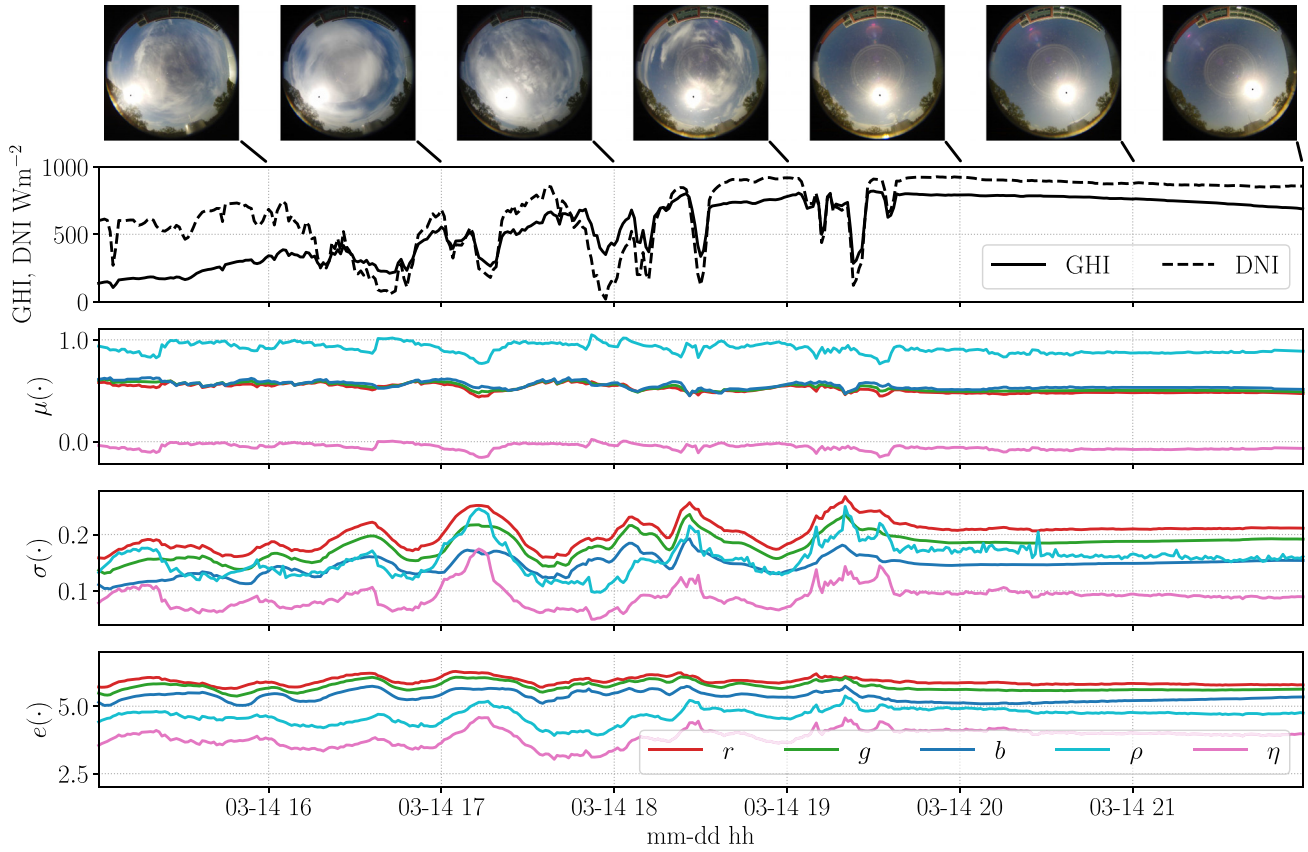


**FIG. 4.** Sky image features. The top panel shows seven sky images for a seven-hour period in 2014-03-14. The second panel shows the GHI and DNI data in the same period. The bottom panels show image features computed for all images in the period (not just the ones shown in the top panel). For visualization, the $\mu$ and $\sigma$ features for the $r$, $g$, and $b$ vectors are normalized by 256 (8-bit color).

Additionally, we consider three forecast methods based on linear regression: Ordinary Least-Squares (OLS), Ridge Regression, and Lasso. We select these models due to their high interpretability, ease-of-use, and widespread availability in statistical software. However, we note that these methods rarely provide state-of-the-art solar forecasting performance, due to the highly nonlinear nature of the solar resource. Instead, these methods provide a straight-forward approach to highlighting the value in our data release. Following Eq. (7), these models are trained to predict the clear-sky index irradiance for the horizons listed in Table III. The actual values for GHI and DNI are then computed by multiplying the predicted $k_t$ values with the respective clear-sky values averaged appropriately. For more details on the models, see Appendix A.

**TABLE III.** Forecasting experiments. For all three horizons, we only evaluate forecasts that apply during daylight hours, i.e., when the solar zenith angle ($\theta_z$) is less than 85°.

| | $\Delta$ | $\tau$ | $t_o$ |
|---|---|---|---|
| Intrahour | 5 min | $\{5, 10, \ldots, 30\}$ min | Every 5 min |
| Intraday | 30 min | $\{30, 60, \ldots, 180\}$ min | Every 30 min |
| Day-ahead | 1 h | $\{26, 27, \ldots, 39\}$ h | Daily at 12:00 UTC |

## B. Results

Here, we report and discuss the sample forecast results. For all horizons, we use 2014 and 2015 as the training set and 2016 as the testing, remove night values according to the solar zenith angle (night $:= \theta_z > 85°$), and select model hyperparameters using tenfold Cross-Validation (CV). However, as noted before, the goal of these sample forecasts is to illustrate the value of the data, rather than evaluating specific forecast methodologies. Therefore, our analysis will be brief and will focus on high-level trends, with standard forecast error metrics: mean absolute error (MAE), mean bias error (MBE), root mean square error (RMSE), and forecast skill, which we compute using RMSE.[1,42] For interested readers, we have included the Python code used to produce these results (see Appendix B for more details).

### 1. Intrahour

We consider intrahour forecasts with horizons 5–30 min, at a temporal resolution of 5-min, backward-averaged. For each horizon, we compare forecasts using only endogenous features and endogenous

**TABLE IV.** GHI forecast results broken down by horizon, model, and feature sets. For each error metric, we report the mean ± the standard deviation over all horizons. The forecast skill is computed relative to a Smart Persistence (Pers.) for the intrahour and intraday horizons and relative to the NAM forecasts for day-ahead. For each model, we consider endogenous-only features vs a combination of endogenous and exogenous features.

| Horizon | Model | Features | MAE ($W/m^2$) | MBE ($W/m^2$) | RMSE ($W/m^2$) | Skill (%) |
|---|---|---|---|---|---|---|
| Intrahour | Pers. | N/A | 32.3 ± 7.1 | 0.9 ± 0.5 | 73.2 ± 11.9 | N/A |
| | OLS | Endog. | 33.0 ± 6.9 | −1.9 ± 2.6 | 67.5 ± 9.8 | 7.5 ± 2.1 |
| | | + Sky images | 37.3 ± 8.3 | −8.0 ± 3.7 | 67.5 ± 10.0 | 7.5 ± 1.9 |
| | Ridge | Endog. | 33.0 ± 6.9 | −1.9 ± 2.6 | 67.5 ± 9.8 | 7.5 ± 2.1 |
| | | + Sky images | 37.3 ± 8.3 | −8.0 ± 3.7 | 67.5 ± 10.0 | 7.5 ± 1.9 |
| | Lasso | Endog. | 33.0 ± 6.9 | −1.9 ± 2.6 | 67.5 ± 9.8 | 7.5 ± 2.1 |
| | | + Sky images | 36.7 ± 8.2 | −7.4 ± 4.0 | 66.8 ± 9.8 | 8.4 ± 2.1 |
| Intraday | Pers. | N/A | 49.9 ± 13.7 | −8.0 ± 6.0 | 89.6 ± 20.3 | N/A |
| | OLS | Endog. | 50.1 ± 11.1 | −16.8 ± 14.4 | 89.2 ± 20.6 | 0.5 ± 0.7 |
| | | + Satellite | 47.8 ± 10.7 | −18.1 ± 12.4 | 83.1 ± 20.6 | 7.6 ± 2.3 |
| | Ridge | Endog. | 50.1 ± 11.1 | −16.8 ± 14.4 | 89.1 ± 20.6 | 0.6 ± 0.7 |
| | | + Satellite | 47.7 ± 10.8 | −18.3 ± 12.4 | 82.8 ± 20.7 | 8.0 ± 2.5 |
| | Lasso | Endog. | 50.0 ± 11.1 | −16.9 ± 14.5 | 89.1 ± 20.6 | 0.6 ± 0.8 |
| | | + Satellite | 47.8 ± 10.9 | −19.2 ± 12.5 | 82.6 ± 21.3 | 8.4 ± 3.3 |
| Day-ahead | NAM | N/A | 85.1 ± 21.6 | −20.5 ± 62.3 | 110.0 ± 29.3 | N/A |
| | OLS | Endog. | 72.0 ± 42.2 | 0.7 ± 7.9 | 101.0 ± 56.7 | 12.5 ± 38.6 |
| | | + NAM | 54.5 ± 27.3 | −2.8 ± 7.4 | 77.6 ± 37.6 | 31.5 ± 25.4 |
| | Ridge | Endog. | 70.4 ± 40.9 | 0.9 ± 8.2 | 98.5 ± 54.6 | 14.4 ± 37.4 |
| | | + NAM | 51.5 ± 25.0 | −2.1 ± 7.6 | 75.6 ± 35.9 | 33.2 ± 24.3 |
| | Lasso | Endog. | 70.9 ± 41.4 | 2.5 ± 9.3 | 96.9 ± 53.2 | 15.7 ± 36.6 |
| | | + NAM | 50.2 ± 23.9 | −1.5 ± 7.8 | 74.8 ± 35.3 | 33.8 ± 23.9 |

features together with sky image features. More specifically, the endogenous features are the backward-averaged, lagged, and variability features over the past 30 min of irradiance, in steps of 5-min, according to Sec. III A. For the sky images, we extract the average, standard deviation, and entropy features according to Sec. III B. These values are then averaged in 5-min bins for each forecasting issuing time. Following convention, our baseline intrahour forecast is Smart Persistence for both GHI and DNI. Tables IV and V show the forecast results for GHI and DNI, respectively. For all model and feature set combinations, we see positive mean forecast skill values of ∼7.5%–8.4% for GHI and ∼3.0%–4.4% for DNI. The small variation in forecast skill shows that the linear models are not well-suited to take advantage of the additional predictive information encoded in the sky image features. Hence, most intrahour forecast studies use sky image features together with nonlinear models, e.g., Artificial Neural Networks (ANNs).[7,8,11]

### 2. Intraday

For intraday horizons, we evaluate forecasts for 30–180 min ahead, at a temporal resolution of 30-min, which matches the sampling rate of the satellite imagery from GOES-15. The endogenous features are computed for the past 3 h, in steps of 30-min, while the satellite imagery is processed according to Sec. III C and used as exogenous features. As with the intrahour forecasts, we use Smart Persistence as the baseline forecast for both GHI and DNI. The addition of the exogenous features improves the forecasting skill for both GHI and DNI. The lack of regularization in the OLS models results in overfitting which decreases the test forecast skill, relative to the Ridge and Lasso models.

### 3. Day-ahead

As mentioned above, the day-ahead forecasts are generated to be compatible with the CAISO Day-Ahead Market (DAM), which requires forecasts be submitted by 10:00 a.m. Pacific for the following day. Based on the forecast schedule of NAM, we choose to evaluate the 12Z cycle, which corresponds to 4 a.m./5 a.m. Pacific. In this case, the baseline forecast will be the unprocessed NAM forecasts, specifically, the NAM DSWRF forecast for GHI and an estimate of DNI from the NAM DSWRF using the DISC model.[43,44] We compare these baseline forecasts against forecasts trained on endogenous features computed from the previous 8 to 20 h (the first 8 h are not used since they correspond to nighttime), in steps of 1 h, as well as the NAM DSWRF and TCDC forecasts as exogenous features. The addition of the exogenous features has a clear positive effect on forecast performance, with the significant improvement of the forecast skill. This matches the previous literature which showed that model output statistics (MOS) and related techniques can improve the forecast performance over the baseline NAM forecasts.[32,33]

## V. CONCLUSIONS

We introduce a comprehensive dataset with the goal of accelerating the development and benchmarking of solar resource forecasting methods for intrahour, intraday, and day-ahead horizons. The dataset

**TABLE V.** The same as Table IV, but for DNI. For the day-ahead horizons, the NAM DSWRF is used with the DISC model to estimate the DNI, which is then used as the reference model when computing the skill.

| Horizon | Model | Features | MAE ($W/m^2$) | MBE ($W/m^2$) | RMSE ($W/m^2$) | Skill (%) |
|---|---|---|---|---|---|---|
| Intrahour | Pers. | N/A | 56.7 ± 12.9 | 2.7 ± 1.6 | 129.0 ± 25.1 | N/A |
| | OLS | Endog. | 68.9 ± 17.2 | −3.7 ± 6.4 | 124.1 ± 22.9 | 3.6 ± 1.3 |
| | | + Sky images | 74.1 ± 18.2 | −13.8 ± 9.1 | 125.0 ± 23.3 | 3.0 ± 1.0 |
| | Ridge | Endog. | 68.9 ± 17.2 | −3.7 ± 6.4 | 124.1 ± 22.9 | 3.6 ± 1.3 |
| | | + Sky images | 74.0 ± 18.1 | −13.9 ± 9.2 | 124.9 ± 23.2 | 3.0 ± 1.0 |
| | Lasso | Endog. | 69.0 ± 17.2 | −3.7 ± 6.4 | 124.1 ± 22.9 | 3.6 ± 1.3 |
| | | + Sky images | 72.2 ± 18.4 | −14.7 ± 10.0 | 123.0 ± 22.8 | 4.4 ± 1.1 |
| Intraday | Pers. | N/A | 100.4 ± 25.3 | −14.4 ± 9.2 | 183.5 ± 39.2 | N/A |
| | OLS | Endog. | 125.3 ± 27.8 | −37.4 ± 34.2 | 189.2 ± 41.9 | −3.0 ± 2.2 |
| | | + Satellite | 117.3 ± 28.6 | −38.5 ± 32.4 | 178.1 ± 41.8 | 3.2 ± 2.9 |
| | Ridge | Endog. | 125.0 ± 27.7 | −37.6 ± 34.3 | 189.1 ± 41.8 | −3.0 ± 2.2 |
| | | + Satellite | 117.0 ± 28.7 | −38.7 ± 32.5 | 177.4 ± 41.9 | 3.6 ± 3.1 |
| | Lasso | Endog. | 125.1 ± 27.7 | −37.8 ± 34.5 | 189.2 ± 42.0 | −3.0 ± 2.3 |
| | | + Satellite | 116.8 ± 28.8 | −40.3 ± 32.7 | 176.4 ± 42.3 | 4.2 ± 3.6 |
| Day-ahead | NAM | N/A | 173.6 ± 58.1 | −11.9 ± 128.4 | 246.5 ± 36.8 | N/A |
| | OLS | Endog. | 209.2 ± 58.1 | 7.3 ± 18.7 | 257.9 ± 68.5 | −9.2 ± 37.7 |
| | | + NAM | 138.2 ± 17.2 | 11.6 ± 27.1 | 189.4 ± 30.5 | 21.0 ± 18.3 |
| | Ridge | Endog. | 208.3 ± 57.9 | 8.7 ± 19.8 | 254.7 ± 67.6 | −7.9 ± 37.3 |
| | | + NAM | 136.4 ± 16.8 | 13.2 ± 27.7 | 186.7 ± 29.6 | 22.1 ± 18.0 |
| | Lasso | Endog. | 208.9 ± 59.0 | 8.7 ± 21.1 | 252.2 ± 66.6 | −6.9 ± 36.9 |
| | | + NAM | 136.9 ± 16.7 | 12.8 ± 28.4 | 185.1 ± 28.2 | 22.8 ± 17.5 |

is of particular value to the development of statistical and hybrid forecasting methods that make use of multiple exogenous inputs, e.g., sky or satellite imagery. The dataset includes irradiance (GHI and DNI) measurements for three complete years (2014–2016) in California, a high-value region for solar forecasting studies. To complement the irradiance data, we also included a range of common endogenous and exogenous features derived from local telemetry, sky imagery, remote sensing, and NWP forecasts. Data are provided in a ready-to-use format, but we have detailed the preprocessing techniques used to derive both the endogenous and exogenous features from the original data sources. Additionally, we include sample intrahour, intraday, and day-ahead forecasting results and sample codes for the simplest methods to highlight the value of the data, as well as to provide baseline methods for future studies.

## ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement No. EE0008216.

## NOMENCLATURE

| | |
|---|---|
| $B, L, V$ | Irradiance time series features |
| DHI | Diffuse horizontal irradiance |
| DNI | Direct normal irradiance |
| DSWRF | Downward short-wave radiation flux |
| GHI | Global horizontal irradiance |
| GOES | Geostationary operational environmental satellite |
| $I, I_{cs}$ | Irradiance (GHI or DNI) and clear-sky irradiance |
| $\hat{I}$ | Forecasted irradiance |
| $k_t$ | Clear-sky index |
| MAE | Mean absolute error |
| MBE | Mean bias error |
| NAM | North American mesoscale forecast system |
| NWP | Numerical weather prediction |
| OLS | Ordinary least squares |
| PST | Pacific standard time |
| RGB | Red-Green-Blue |
| RMSE | Root mean square error |
| RSR | Rotating shadowband radiometer |
| $t_o$ | Forecasting issuing time |
| TCDC | Total cloud cover |
| UTC | Coordinated Universal Time |
| $\Delta$ | Data aggregation window |
| $\mu, \sigma, e$ | Sky image features |
| $\tau$ | Forecasting horizon |

## APPENDIX A: LINEAR FORECAST MODELS

Here, we discuss the mathematical details of the three considered linear forecast models: OLS, Ridge Regression, and Lasso. The three models can be formulated by solving the following optimization problems:

**TABLE VI.** List of files available in the data repository. The second column indicates the type of information contained in the file, where "Primary" refers to quality controlled data from the primary sources (solar sensor, sky imager, and NWP), "Secondary" refers to data obtained by processing the primary data, and "Code" refers to Python 3 scripts. All these files can be accessed at https://doi.org/10.5281/zenodo.2826939.

| File | Type | Description |
|---|---|---|
| `Folsom_irradiance.csv` | Primary | One-minute GHI, DNI, and DHI data. |
| `Folsom_weather.csv` | Primary | One-minute weather data. |
| `Folsom_sky_images_{YEAR}.tar.bz2` | Primary | `Tar` archives with daytime sky images captured at 1-min intervals for the years 2014, 2015, and 2016, compressed with `bz2`. |
| `Folsom_NAM_lat{LAT}_lon{LON}.csv` | Primary | NAM forecasts for the four nodes nearest the target location. `{LAT}` and `{LON}` are replaced by the node's coordinates listed in Table I. |
| `Folsom_sky_image_features.csv` | Secondary | Features derived from the sky images. |
| `Folsom_satellite.csv` | Secondary | 10 pixel by 10 pixel GOES-15 images centered in the target location. |
| `Irradiance_features_{horizon}.csv` | Secondary | Irradiance features for the different forecasting horizons (`{horizon}` = `{intrahour, intraday, day-ahead}`). |
| `Sky_image_features_intra-hour.csv` | Secondary | Sky image features for the intrahour forecasting issuing times. |
| `Sat_image_features_intra-day.csv` | Secondary | Satellite image features for the intraday forecasting issuing times. |
| `NAM_nearest_node_day-ahead.csv` | Secondary | NAM forecasts (GHI, DNI computed with the DISC algorithm, and total cloud cover) for the nearest node to the target location prepared for day-ahead forecasting. |
| `Target_{horizon}.csv` | Secondary | Target data for the different forecasting horizons. |
| `Forecast_horizon.py` | Code | Python script used to create the forecasts for the different horizons. |
| `Postprocess.py` | Code | Python script used to compute the error metric for all the forecasts. |

$$\underset{x}{\text{minimize}} \quad ||Ax-b||_2^2 \ \text{(OLS)},$$
$$\underset{x}{\text{minimize}} \quad ||Ax-b||_2^2+\lambda||x||_2 \ \text{(Ridge)},$$
$$\underset{x}{\text{minimize}} \quad ||Ax-b||_2^2+\lambda||x||_1 \ \text{(Lasso)},$$

where $x \in \mathbb{R}^n$ encodes the model parameters, $A \in \mathbb{R}^{m\times n}$ is the input data, and $b \in \mathbb{R}^m$ is the output data for $m$ samples. The key difference between the three is the regularization parameter or lack thereof. In practice, the $\ell$-2 regularizer ($||\cdot||_2$) prevents overfitting, whereas the $\ell$-1 regularizer ($||\cdot||_1$) promotes a sparse solution. In both cases, the strength of the regularization is controlled by a hyperparameter $\lambda \in \mathbb{R}$, which can be selected via cross-validation. Further information on these models, both on the theory and implementation details, can be found in any standard textbook on Machine Learning.

## APPENDIX B: DATA REPOSITORY AND SAMPLE CODE

This section introduces the steps to download the datasets and sample codes described previously.

### 1. Data repository

All datasets are available at the open-access repository at https://doi.org/10.5281/zenodo.2826939 under a Creative Commons (CC) license. Numerical data (e.g., irradiance time series and image features) are given in the comma separated values `CSV` format, and sky images are provided as `Tar` archives containing compressed `JPG` files. All the files available are described in Table VI. Note that the quality control of the data for different forecast horizons and issuing times yielded instances for which not all data entries are available (e.g., missing satellite images or sky images). In those instances, we left the offending timestamps in the data files, and the missing data are identified by the string `NaN`.

### 2. Sample code

As part of the data release, we are also including the sample code written in pure Python 3. The preprocessed data used in the scripts are also provided. The code can be used to reproduce the results presented in this work and as a starting point for future studies. Besides the standard scientific Python packages (numpy,[45] scipy,[46] and matplotlib[47]), the code depends on pandas[48] for time-series operations, pvlib[49] for common solar-related tasks, and scikit-learn[50] for Machine Learning models. All required Python packages are readily available on Mac, Linux, and Windows and can be installed via, e.g., pip. The scripts used to create the forecast and postprocess the results are listed in Table VI.

The usage of the datasets and sample codes presented here is intended for research and development purposes only and implies explicit reference to the present paper, as opposed to reference to the dataset DOI only. Although every effort was made to ensure the quality of the data, no guarantees or liabilities are implied by the authors or publishers of the data.

## REFERENCES

[1]R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra, "Solar forecasting methods for renewable energy integration," Prog. Energy Combust. Sci. **39**, 535–576 (2013).
[2]E. W. Law, A. A. Prasad, M. Kay, and R. A. Taylor, "Direct normal irradiance forecasting and its application to concentrated solar thermal output forecasting—A review," Sol. Energy **108**, 287–307 (2014).
[3]D. Yang, J. Kleissl, C. A. Gueymard, H. T. C. Pedro, and C. F. M. Coimbra, "History and trends in solar irradiance and PV power forecasting: A

preliminary assessment and review using text mining," Sol. Energy **168**, 60–101 (2018).
[4]D. Yang, "SolarData: An R package for easy access of publicly available solar datasets," Sol. Energy **171**, A3–A12 (2018).
[5]M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The National Solar Radiation Data Base (NSRDB)," Renewable Sustainable Energy Rev. **89**, 51–60 (2018).
[6]R. Marquez, V. Gueorguiev, and C. F. M. Coimbra, "Forecasting of global horizontal irradiance using sky cover indices," J. Sol. Energy Eng. **135**, 011017 (2013).
[7]R. Marquez and C. F. M. Coimbra, "Intra-hour DNI forecasting based on cloud tracking image analysis," Sol. Energy **91**, 327–336 (2013).
[8]Y. Chu, H. T. C. Pedro, and C. F. M. Coimbra, "Hybrid intra-hour DNI forecasts with sky image processing enhanced by stochastic learning," Sol. Energy **98**, 592–603 (2013).
[9]S. Quesada-Ruiz, Y. Chu, J. Tovar-Pescador, H. T. C. Pedro, and C. F. M. Coimbra, "Cloud-tracking methodology for intra-hour DNI forecasting," Sol. Energy **102**, 267–275 (2014).
[10]Y. Chu, L. Nonnenmacher, R. H. Inman, Z. Liao, H. T. C. Pedro, and C. F. M. Coimbra, "A smart image-based cloud detection system for intra-hour solar irradiance forecasts," J. Atmos. Oceanic Technol. **31**, 1995–2007 (2014).
[11]Y. Chu, H. T. C. Pedro, M. Li, and C. F. M. Coimbra, "Real-time forecasting of solar irradiance ramps with smart image processing," Sol. Energy **114**, 91–104 (2015).
[12]H. T. C. Pedro and C. F. M. Coimbra, "Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances," Renewable Energy **80**, 770–782 (2015).
[13]H. T. C. Pedro and C. F. M. Coimbra, "Short-term irradiance forecastability for various solar micro-climates," Sol. Energy **122**, 587–602 (2015).
[14]M. Li, Y. Chu, H. T. C. Pedro, and C. F. M. Coimbra, "Quantitative evaluation of the impact of cloud transmittance and cloud velocity on the accuracy of short-term DNI forecasts," Renewable Energy **86**, 1362–1371 (2016).
[15]Y. Chu, M. Li, and C. F. M. Coimbra, "Sun-tracking imaging system for intra-hour DNI forecasts," Renewable Energy **96**, 792–799 (2016).
[16]Y. Chu and C. F. M. Coimbra, "Short-term probalistic forecasts for direct normal irradiance," Renewable Energy **101**, 526–536 (2017).
[17]H. T. C. Pedro, C. F. M. Coimbra, M. David, and P. Lauret, "Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts," Renewable Energy **123**, 191–203 (2018).
[18]R. Marquez, H. T. C. Pedro, and C. F. M. Coimbra, "Hybrid solar forecasting method uses satellite imaging and ground telemetry as inputs to ANNs," Sol. Energy **92**, 176–188 (2013).
[19]L. Nonnenmacher and C. F. M. Coimbra, "Streamline-based method for intra-day solar forecasting through remote sensing," Sol. Energy **108**, 447–459 (2014).
[20]D. P. Larson and C. F. M. Coimbra, "Direct power output forecasts from remote sensing image processing," J. Sol. Energy Eng. **104**, 021011 (2018).
[21]D. P. Larson, L. Nonnenmacher, and C. F. M. Coimbra, "Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest," Renewable Energy **91**, 11–20 (2016).
[22]S. M. Wilcox and D. R. Myers, "Evaluation of radiometers in full-time use at the National Renewable Energy Laboratory Solar Radiation Research Laboratory," Technical Report No. NREL/TP-550-44627, 946331, 2008.
[23]C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, J. Kleissl, J. Shields, and B. Washom, "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed," Sol. Energy **85**, 2881–2893 (2011).
[24]V. Bone, J. Pidgeon, M. Kearney, and A. Veeraragavan, "Intra-hour direct normal irradiance forecasting through adaptive clear-sky modelling and cloud tracking," Sol. Energy **159**, 852–867 (2018).
[25]B. Nouri, P. Kuhn, S. Wilbert, N. Hanrieder, C. Prahl, L. Zarzalejo, A. Kazantzidis, P. Blanc, and R. Pitz-Paal, "Cloud height and tracking accuracy of three all sky imager systems for individual clouds," Sol. Energy **177**, 213–228 (2019).
[26]H. T. C. Pedro, C. F. M. Coimbra, and P. Lauret, "Adaptive image features for intra-hour solar forecasts," J. Renewable Sustainable Energy **11**, 036101 (2019).
[27]L. Ye, Z. Cao, and Y. Xiao, "DeepCloud: Ground-based cloud image categorization using deep convolutional features," IEEE Trans. Geosci. Remote Sens. **55**, 5729–5740 (2017).
[28]R. Perez, P. Ineichen, K. Moore, M. Kmiecik, C. Chain, R. George, and F. Vignola, "A new operational model for satellite-derived irradiances: Description and validation," Sol. Energy **73**, 307–317 (2002).
[29]R. Perez, S. Kivalov, J. Schlemmer, K. Hemker, Jr., D. Renné, and T. E. Hoff, "Validation of short and medium term operation solar radiation forecasts in the US," Sol. Energy **84**, 2161–2172 (2010).
[30]S. D. Miller, M. A. Rogers, J. M. Haynes, M. Sengupta, and A. K. Heidinger, "Short-term solar irradiance forecasting via satellite/model coupling," Sol. Energy **168**, 102–117 (2018).
[31]I. Bilionis, E. M. Constantinescu, and M. Anitescu, "Data-driven model for solar irradiation based on satellite observations," Sol. Energy **110**, 22–38 (2014).
[32]P. Mathiesen and J. Kleissl, "Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States," Sol. Energy **85**, 967–977 (2011).
[33]P. Mathiesen, C. Collier, and J. Kleissl, "A high-resolution, cloud-assimilating numerical weather prediction model for solar irradiance forecasting," Sol. Energy **92**, 47–61 (2013).
[34]ECMWF, *IFS Documentation CY45R1*, IFS Documentation (ECMWF, 2018).
[35]S. Sperati, S. Alessandrini, and L. Delle Monache, "An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting," Sol. Energy **133**, 437–450 (2016).
[36]R. Perez, E. Lorenz, S. Pelland, M. Beauharnois, G. Van Knowe, K. Hemker, Jr., D. Heinemann, J. Remund, S. C. Müller, W. Traunmüller, G. Steinmauer, D. Pozo, J. A. Ruiz-Arias, V. Lara-Fanego, L. Ramirez-Santigosa, M. Gaston-Romero, and L. M. Pomares, "Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe," Sol. Energy **94**, 305–326 (2013).
[37]S. E. Haupt, B. Kosović, T. Jensen, J. K. Lazo, J. A. Lee, P. A. Jiménez, J. Cowie, G. Wiener, T. C. McCandless, M. Rogers, S. Miller, M. Sengupta, Y. Xie, L. Hinkelman, P. Kalb, and J. Heiser, "Building the Sun4Cast system: Improvements in solar power forecasting," Bull. Am. Meteorol. Soc. **99**, 121–136 (2017).
[38]S. Pelland, G. Galanis, and G. Kallos, "Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model," Prog. Photovolt.: Res. Appl. **21**, 284–296 (2013).
[39]P. Ineichen and R. Perez, "A new airmass independent formulation for the Linke turbidity coefficient," Sol. Energy **73**, 151–157 (2002).
[40]See www.soda-pro.com for "Linke Turbidity factor."
[41]J. Remund, L. Wald, M. Lefèvre, T. Ranchin, and J. H. Page, "Worldwide Linke turbidity information," in Proceedings of ISES Solar World Congress (2003), Vol. 400.
[42]R. Marquez and C. F. M. Coimbra, "Proposed metric for evaluation of solar forecasting models," J. Sol. Energy Eng. **135**, 011016–1–011016–9 (2012).
[43]E. L. Maxwell, "A quasi-physical model for converting hourly global horizontal to direct normal insolation," Technical Report No. SERI/TR-215-3087, Solar Energy Research Institute, Golden, CO, 1987.
[44]A. Skarveit, J. A. Olseth, and M. E. Tuft, "An hourly diffuse fraction model with correction for variability and surface albedo," Sol. Energy **63**, 173–183 (1998).
[45]T. E. Oliphant, *A Guide to NumPy* (Trelgol Publishing, USA, 2006).
[46]T. Oliphant, "Python for scientific computing," Comput. Sci. Eng. **9**, 10–20 (2007).
[47]J. D. Hunter, "Matplotlib: A 2d graphics environment," Comput. Sci. Eng. **9**, 90–95 (2007).
[48]W. McKinney, "Data structures for statistical computing in python," in Proceedings of the 9th Python in Science Conference (2010), pp. 51–56.
[49]W. F. Holmgren, C. W. Hansen, and M. A. Mikofski, "pvlib Python: A python package for modeling solar energy systems," J. Open Source Software **3**, 884 (2018).
[50]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, V. D. Ron Weiss, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).