See the Light: Modeling Solar Performance using Multispectral Satellite Data

Akansha Singh Bansal and David Irwin
akanshasingh@umass.edu,deirwin@umass.edu
University of Massachusetts Amherst

ABSTRACT

Developing accurate solar performance models, which infer solar output from widely available external data sources, is increasingly important as the grid's solar capacity rises. These models are important for a wide range of solar analytics, including solar forecasting, resource estimation, and fault detection. The most significant error in existing models is inaccurate estimates of clouds' effect on solar output, as cloud formations and their impact on solar radiation are highly complex. In 2018 and 2019, respectively, the National Oceanic and Atmospheric Administration (NOAA) in the U.S. began releasing multispectral data comprising 16 different light wavelengths (or channels) from the GOES-16 and GOES-17 satellites every 5 minutes. Enough channel data is now available to learn solar performance models using machine learning (ML). In this paper, we show how to develop both local and global solar performance models using ML on multispectral data, and compare their accuracy to existing physical models based on ground-level weather readings and on NOAA's estimates of downward shortwave radiation (DSR), which also derive from multispectral data but using a physical model. We show that ML-based solar performance models based on multispectral data are much more accurate than weather- or DSR-based models, improving the average MAPE across 29 solar sites by over 50% for local models and 25% for global models.

CCS CONCEPTS

• **Applied computing** → *Environmental sciences*; • **Information systems** → Geographic information systems; **Data analytics**.

KEYWORDS

Solar Modeling and Analysis, Satellite Data, Black-Box Models

ACM Reference Format:
Akansha Singh Bansal and David Irwin. 2020. See the Light: Modeling Solar Performance using Multispectral Satellite Data. In *The 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '20), November 18–20, 2020, Virtual Event, Japan.* , 10 pages. https://doi.org/10.1145/3408308.3427610

BuildSys '20, November 18–20, 2020, Virtual Event, Japan

ACM ISBN 978-1-4503-8061-4/20/11…$15.00
https://doi.org/10.1145/3408308.3427610

1 INTRODUCTION

Grid-connected solar capacity continues to grow exponentially at roughly a 20-30% increase per year [8]. This is in line with Swanson's law, which observes that the price of solar photovoltaic (PV) modules tends to drop 20% for every doubling of production volume [32]. This increase in solar capacity is expected to continue for the foreseeable future with solar power expected to satisfy 25% of global electricity demand by 2050 [7]. Of course, solar's potential is much higher as enough sunlight strikes the Earth's surface in only 1.5 hours to satisfy the world's annual energy consumption [20]. This dramatic increase in solar power is expected to place increasing stress on the electric grid, which must continue to balance supply and demand despite large potential fluctuations in solar power generation that are geographically distributed.

The underlying reason is the mismatch in activation time between solar modules and conventional thermal generators. While solar modules are always active and ramp power up and down nearly instantaneously as clouds pass by, conventional generators may take anywhere from tens of seconds to days to activate depending on their size. As a result, under large solar penetrations, utilities keep many conventional generators active as spinning reserve to quickly offset any dips in solar power. This is both expensive and highly energy-inefficient, and akin to indefinitely maintaining an idling car that is only driven periodically for short distances. While batteries can mitigate some of this inefficiency, they are unlikely to eliminate it at the grid level in the near future due to both high cost and the limited supply of lithium on Earth [25].

A complementary approach to improving grid operations is to improve the accuracy of current and projected solar power output. Solar performance models infer one or more sites' solar output based on their physical and environmental characteristics, and are a basis for a range of solar analytics, including short- and long-term forecasting [19], resource estimation [4], fault detection [3, 6], and disaggregation [22, 28]. In general, solar power is a well-known function of a module's physical characteristics, e.g., type, wiring topology, inverter, tilt, orientation, location, elevation, etc., and its environment, primarily the time of day, day of year, temperature, and cloud cover. There are many "white box" modeling frameworks, such as PVlib [14] and the U.S. Department of Energy's System Advisor Model (SAM) [24], that enable users to configure their physical and environmental characteristics to estimate solar output. There has also been recent work on "black box" data-driven modeling, such as Solar-TK [16, 21], which automatically derive physical characteristics from data, and uses them to estimate solar output based on current or forecasted environmental characteristics.

Unfortunately, the accuracy of these frameworks in estimating solar power is only as good as their input. In general, white-box approaches, such as PVlib and SAM, assume the components of

ground-level solar radiation are well-known, e.g, global horizontal irradiance (GHI), direct normal irradiance (DNI), and diffuse horizontal irradiance (DHI). Solar performance modeling is highly accurate given accurate estimates of ground-level solar radiation, derived from either a pyranometer [14] or a nearby solar site. Unfortunately, high fidelity solar radiation data is not widely available at most sites. Thus, prior black-box approaches have estimated ground-level solar radiation using cloud cover estimates commonly reported by weather stations [21, 22]. Unfortunately, the frequency, resolution, and spatial coverage of these cloud cover estimates are coarse and imprecise, which results in significant inaccuracy. Importantly, this inaccuracy in estimating the effect of cloud cover on ground-level solar radiation is *by far the largest source of error* in solar performance models that estimate solar output.

An alternative approach for inferring cloud effects is to use data from satellites. For example, the Heliosat family of algorithms were first introduced in the late 1980s and have been updated since then [17]. These algorithms analyze satellite images in the visible light spectrum, and estimate a "cloud index" by comparing a pixel's actual value with the value it would have under a clear sky. These algorithms generally use physical models that are calibrated from empirical observations of a location, and have grown increasingly more complex as satellite sensors have grown more sophisticated. In particular, the latest generation of U.S. satellites (GOES-16 and GOES-17) include a sensor—the Advanced Baseline Imager (ABI)—that takes images of the Earth with 16 spectral bands, including two visible channels, four near-infrared channels, and ten infrared channels. The ABI is capable of imaging the entire continental U.S. (CONUS) at resolutions ranging from $0.5 - 2$km every 5 minutes.

The U.S. National Oceanic and Atmospheric Administration (NOAA) began releasing both raw data and derived data products from the GOES-16 and GOES-17 in early 2019. As a result, there is now enough raw channel data available to learn solar performance models using machine learning (ML). In this paper, we show how to develop both local and global solar performance models using ML on multispectral satellite data. Local solar performance models are trained on data from a specific solar site where the input features include multispectral data and the output is solar power generation, while global models are trained on normalized data from many solar sites. As we show, local models are more accurate, but require local data from each new site for training, while global models are less accurate, but do not require any local data for training.

We compare our ML models above with existing calibrated physical models using both ground-level weather readings and NOAA's estimates of downward shortwave radiation (DSR). The latter estimates also derive from multispectral data, but using a physical model, and represent the state-of-the-art for physical modeling of surface radiation from satellite data. Our work differs from prior work on estimating solar radiation in that we focus on end-to-end solar performance models that estimate the solar power generation of a particular site (at a specific location and time) using widely available environmental data. Most prior work, including PVlib and SAM, instead decouples estimating surface solar radiation from estimating solar power output based on its physical characteristics, e.g., efficiency, tilt/orientation, shading, temperature coefficient, etc., given surface solar radiation. We focus on end-to-end modeling because it is simpler, and there is less need for decoupling when using ML, as ML training is capable of jointly learning the solar radiation and the effect of a site's physical characteristics.

Our hypothesis is that training ML-based solar performance models on new multispectral satellite data can yield higher accuracy than existing physical models that use either multispectral satellite data or ground-level cloud cover readings. In evaluating our hypothesis, we make the following contributions.

Analyzing Multispectral Satellite Data. We analyze existing multispectral satellite data, and its derived data products, from GOES-16 and GOES-17 that are being made publicly available. We compile a dataset composed of solar generation every 5m-1hr from 29 solar sites at known locations, along with the value of the 16 spectral bands every 5m-1hr, DSR estimate, temperature, and ground-level cloud cover reading, e.g., clear, scattered, broken, overcast, etc.

ML-based Solar Performance Models. We develop approaches for training both local and global ML models using multispectral satellite data, and compare them with prior approaches that use calibrated physical models. The local models are simple, and trained on a dataset that includes multispectral data as input features and solar generation as the dependent output variable. Instead, the global model requires normalizing each site's solar output to enable training a consistent model across multiple sites.

Implementation and Evaluation. We implement both our ML-based models and existing models and evaluate them on up to 2 years of multispectral data (the maximum that has been released) from the 29 sites. We show that ML-based solar performance models based on multispectral data are much more accurate than weather or DSR-based models, improving the average MAPE across 29 solar sites by over 50% for local models and 25% for global models.

2 BACKGROUND ESTIMATES

Our problem is to develop a solar performance model that infers solar power output for a specific location, time-of-day, and day-of-year given historical solar power output, and the location's environmental data at the same time. Below, we discuss prior approaches that use physical models, but with different types of environmental data as input. We then detail the characteristics of the data sets we use for our learning approach, including the raw channel data gathered by the GOES-16 and GOES-17 satellites.

2.1 Prior Approaches

White-box Modeling. Solar performance modeling is a mature area with detailed physical models available that can accurately estimate the power output of a solar system. These models describe how the system's environment (e.g., due to temperature, cloud cover, etc.), physical characteristics (e.g., wiring topology, conversion efficiency, conversion losses, etc.), and location (e.g., time-of-day, day-of-year, elevation, shading, etc.) affect solar power. White-box modeling frameworks, such as PVlib [14] and SAM [24], require users to configure virtual solar systems that include these details—down to the type of hardware and surface irradiance—and then uses the available physical models to provide a solar estimate. Since solar systems are often highly complex, and surface-level irradiance measurements are often not available, recent work has

ABI Band	*Central Wavelength* (μm)	*Spatial Resolution* (km)	*Type*
1	0.47	1	Visible
2	0.64	0.5	Visible
3	0.86	1	Near-Infrared
4	1.37	2	Near-Infrared
5	1.6	1	Near-Infrared
6	2.2	2	Near-Infrared
7	3.9	2	Infrared
8	6.2	2	Infrared
9	6.9	2	Infrared
10	7.3	2	Infrared
11	8.4	2	Infrared
12	9.6	2	Infrared
13	10.3	2	Infrared
14	11.2	2	Infrared
15	12.3	2	Infrared
16	13.3	2	Infrared

Table 1: Wavelength for 16 channels of GOES-16 and -17 [9].

also explored data-driven approaches to learning the parameters of the physical models solely from historical solar power data [21, 22].

Data-driven Modeling. Data-driven solar modeling approaches estimate surface irradiance by combining well-known clear sky models [10, 29] with simple cloud cover models [21, 27]. Clear sky models accurately estimate surface irradiance based on the Sun's position in the sky, which is deterministic for a given location at a given time-of-day and day-of-year. Simple cloud cover models then translate basic weather station readings of cloud cover, which are made available by the National Weather Service (NWS) for every location in the U.S. These cloud cover readings are coarse observations in units of *oktas*, where 1 okta represents one-eighth of the sky being covered by clouds. The measurements are typically made by placing a circular sky mirror divided into eight slices on the ground, such that any slice that reflects a cloud is 1 okta. Okta-based measurements are typically reported as common string values, such as "clear," "scattered," "broken," and "overcast." Simple data-driven solar modeling uses the reported oktas to estimate a cloud index, which captures the percentage reduction in the clear sky irradiance due to cloud cover. Clearly, measuring cloud cover using oktas is highly imprecise, and thus represents the largest source of error in simple data-driven solar modeling.

Clouds formations are highly complex, and have different impacts on solar radiation depending on their height in the sky and composition. In addition, other atmospheric properties, such as water vapor and aerosol particles, can affect the absorption and scattering of solar radiation. These complex effects simply cannot be captured by okta measurements that only range from 1-8.

Satellite-based Modeling. An alternative approach is to estimate a similar cloud index, and surface radiation, using satellite images taken from space. Even early satellite-based imagers were capable of more precision than okta-based measurements. The Heliosat method was first introduced in the late 1980s [17, 18, 30] to estimate a similar cloud index from visual images of the Earth's surface, and has been improved upon multiple times as satellites have improved. The basic idea is that the more light clouds reflect back in satellite images, the less light reaches the surface. Thus, in visual images, darker pixels represent higher surface irradiance, and lighter pixels represent lower irradiance. Of course, the physical models are highlly complex, as different locations have different ground reflectivities, which can change over time, e.g., due to foliage, snow, roadways, etc. Thus, Heliosat, and methods derived from it, use complex physical models to translate satellite measurements into a surface radiation estimate. The latest methods go well beyond using simple clear sky models, and account for atmospheric changes even under clear skies, such as the Linke turbidity factor. However, many of the latest methods are proprietary, as large-scale solar radiation data is becoming an increasingly valuable commodity for a wide range of applications beyond solar energy modeling [12, 13].

Satellite-based methods also necessarily change as new satellites are launched with new and more advanced sensors. The GOES-16 and GOES-17 are the latest generation of weather satellites launched by the U.S. GOES-16 and GOES-17 became operational in 2017 and 2018, respectively, and cover different regions. GOES-16 covers the east region of the U.S. and GOES-17 covers the west coast and much of the Pacific ocean. The GOES satellites record 16 spectral bands (or channels) for the continental U.S. every 5 minutes at a high spatial resolution (1-2km depending on the channel). By contrast, Heliosat was originally developed for visual images in a single spectral band, while the previous generation of satellites recorded only 5 spectral bands at 15 minutes resolution. Importantly, NOAA makes the satellite data publicly available for download [11]. In addition, NOAA is developing numerous higher-level data products based on the raw spectral data. In particular, the Downward Shortwave Radiation (DSR) product represents the state-of-the-art in estimating surface radiation from satellite data using physical models. The theoretical basis and algorithm for the DSR physical model is described in a 125-page white paper released by NOAA [2]. These DSR estimates are made publicly available hourly, and can be used as input directly into white-box or data-driven models for solar output (or by using it to compute a cloud index). Unfortunately, DSR's physical model cannot often not compute its value under overcast sky conditions, and thus often has missing data points [2].

2.2 Spectral Data Characteristics

Table 1 shows the centerpoint of spectral bands recorded by GOES-16 and GOES-17, and their wavelengths. Solar cells generate power from wavelengths in the range 0.38μm to 0.75μm, which are described by channels 1 and 2 and part of channel 3. However, while not directly relevant, the other channels may also embed important information about the characteristics of clouds and the atmosphere that could indirectly reveal information about solar generation. A full description of the channels, and what they are capable of sensing, is outside the scope of this paper. Since we focus on using ML rather than physical modeling for developing our models, the precise meaning of the different channels is not as significant. We simply treat channel data as a "black box" for learning.

Figure 1 gives some intuitive sense of the relationship between channel values and a site's solar output on a sunny day with no clouds. The left y-axis shows a site's solar output over a day normalized by the maximum solar output over the day. The right y-axis then shows the channel values for the same location (within 1km^2 area), which are in Wm$^{-2}\mu$m^{-1}. As shown the channel values follow a similar trend. The relationship with higher channels does not

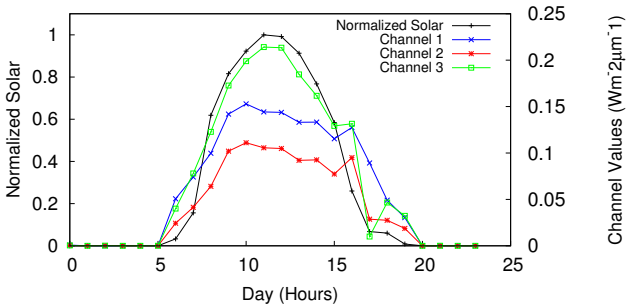

Figure 1: Relationship between solar generation at a site, and the values of channels 1, 2, and 3 at the same location.

exactly follow the trend, since they measure longer wavelengths of irradiance that capture properties not reflected in solar output.

3 ML-BASED SOLAR MODELS

We present both local and global ML solar performance models using GOES satellite multi-spectral data. Local models are trained for each individual solar site using its own data, and only apply to that one site. In contrast, global models are trained on data from many solar sites, and are applicable to any new solar site even if data for the site is not available in the training set.

3.1 Local ML Models

Our local ML solar performance model is simple: the input features are time-series data of the 16 channel values for a particular solar site location and the location's ambient temperature, while the dependent output variable is the average power generated by the solar site over the same time intervals. The channel values indirectly quantify the surface irradiance, while the temperature is necessary because solar cell conversion efficiency varies with the cell temperature. In general, for every 1°C increase in cell temperature, the efficiency of converting solar irradiance to electrical energy decreases by $\sim 0.5\%$. While we assume a solar site's location is known, and used to determine the associated channel values and temperature, prior work shows how to extract an accurate location directly from solar power data at one-hour or less time resolution [23]. We discuss in detail the process of extracting the 16 channel values for a location from the GOES satellite's NetCDF-formatted data files in Section 4. We simply retrieve the temperature from Weather Underground, a popular online weather website. Finally, we retrieve solar power data remotely from web-based solar monitoring systems. We have archived multiple years worth of 5 to 15 minute resolution average solar generation data from 29 solar sites.

For each solar site, we curate a training dataset with the timestamp, 16 channel values, temperature, and average solar power generation. We then train an ML model using a support vector machine (SVM). While we could use any ML regression model for training, SVMs have been used for solar modeling in the past, and shown to have higher accuracy than other regression models. SVM is well-suited for a variety of reasons. In particular, it is a non-linear model and our input variables do not have a linear relationship, and we have multiple input features with different magnitudes, units, and ranges. For our SVM, we select a specific error range by setting the margins and a radial function. We define a tolerance margin (ϵ), a regularization co-efficient C, and use the radial basis function (RBF) as the kernel. The tolerance ϵ and co-efficient C are estimated using 5-fold cross-validation on the training data in the following range: $\epsilon \in \{0.005, 0.01, 0.05, 0.1, 0.2\}$ and $C \in \{1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\}$. For our basic model, we perform 5-fold SVM regression for each site individually to evaluate its performance. We use 5-fold evaluation to get a more robust estimate of the performance. We call this a local model because the model is trained separately for each solar site.

A key benefit of our approach above compared to both prior ML-based solar models and prior physical models is its simplicity. We apply a standard ML regression model to a simplistic dataset composed of only three input feature types—timestamp, channel values, and temperature—without using any domain-specific knowledge. As a result, the approach is purely a "black box" that requires only gathering and curating the datasets for training. In contrast, prior ML approaches to solar modeling are much more complex, and not pure black box approaches, because they lack the data necessary to directly infer surface level irradiance [26, 31]. As a result, these approaches must use time and cloud conditions to indirectly estimate surface irradiance. These methods also often mix ML with numerous physical models that describe the effect of temperature, solar geometry, location, and time to improve accuracy. While doing so improves accuracy and reduces the training data necessary to learn the model, it also increases model complexity.

As we show in Section 5, our simple local ML solar performance models, which include no physical models, are significantly more accurate than these prior approaches. The largest source of error in prior approaches stems from the inaccurate measurements of the effect of clouds, which satellite data improves upon. Our ML approach is also able to learn the effects of the physical models above, which mitigates the advantage of using these models.

3.2 Global ML Models

The local ML models above must be trained for each individual solar site, which requires acquiring sufficient training data to learn the model. In general, roughly one year of training data that captures all of the Sun's positions in the sky across the year is necessary to learn an accurate model. As a result, local ML models have some significant practical limitations. To address this problem, we also develop a global ML model that uses satellite data. Global models, once trained, can be applied to any solar site without retraining the model. As we show in Section 5, these models are less accurate than the local ML models above, but still more accurate than prior approaches that do not use GOES satellite data for estimating the effect of clouds on surface irradiance. The primary reason for the degraded accuracy is that global ML models can conflate the effects of many characteristics that are unique to each solar site when training, including each site's unique shading behavior, geometry (i.e., tilt and orientation), temperature coefficient, wiring topology, inverter type, conversion efficiency, etc. However, as mentioned above, since the effect of these differences is small compared to the effect of inaccurate cloud cover estimates, global ML models are able to maintain higher accuracy than prior approaches.

To develop our global models, we train our models not from data from a single site, but using data from many sites. That is, we combine the training sets for individual sites above together into a large dataset. The only change we make is to the dependent output variable, which in the local models is the solar power output in watts. Since solar sites have different sizes, and thus different solar outputs, we must normalize this power by the maximum power a site is capable of producing *at the given time*. To do so, we adopt an approach from prior work [22] that bounds a site's solar power curve using the solar irradiance curve from a clear sky model. Prior work shows that this method requires few datapoints, and yields an accurate model of a solar site's maximum solar output under clear skies at any time. Dividing solar power by the maximum generation yields a normalized output across sites in the range $[0, 1]$. We then train the global ML model using the same approach as above.

3.3 ML Model Variants

In addition to defining the basic local and global ML models above, we also experimented with many model variants, which we present the results of in Section 5. We describe these variants below.

Varying Resolutions. The spectral satellite data is made available every 5 minutes, enabling us to train models at any resolution greater than 5-minutes. By contrast, DSR and ground-level cloud cover observations are typically reported only every hour. Thus, for comparison, we train our ML models at multiple different time resolutions, including 5 minutes, 15 minutes, and 1 hour. Similarly, for comparison, we increase the resolution of DSR and ground-level cloud cover readings by simply assuming that every 5 or 15 minute interval within an hour has the same value.

Varying Channels. We compare the accuracy of using different numbers of channels. While in our basic model, we use all 16 channels, we also examine the accuracy of using only the first 3 channels that corresponds to the visible region and directly senses the wavelengths converted to solar power.

Multi-Satellite Models. While GOES-16 targets the eastern portion of the U.S. and GOES-17 targets the west coast and Pacific ocean, they both capture data from the entire continental U.S. from different angles. Thus, we augment our basic models above, which primarily use GOES-16 data since most of our sites are in the eastern part of the U.S., to use both GOES-16 and GOES-17 data. This provides data from two different vantage points in space for the same location. To do so, we simply augment our model above to also include the 3 channels of data from GOES-17.

4 IMPLEMENTATION

We implemented our satellite-based ML models using multispectral data in python, along with two competing approaches that apply physical models to ground-level cloud cover readings and DSR. We summarize these competing approaches more below. We used python's *scikit-learn* ML library to train the SVM and other regression models. We collect hourly temperature and ground-level cloud cover readings from Weather Underground, a popular online weather site. For the physical modeling approaches, we use the pysolar [1] library to derive a site's clear sky irradiance based on its location and time. We collect solar power data from 29 sites remotely via their energy meter API. We initially gathered data from 75 sites and filtered sites where we could not verify the solar array in satellite imagery, e.g., from Google, did not have minute-level solar data available, or did not have 2 years worth of solar data. This left us with the 29 sites across U.S. which we analyze.

Data	*Units*	*Time Resolution*	*Source*
GOES-16 Channel	$\mathrm{Wm}^{-2}\mu\mathrm{m}^{-1}$	5 Minutes	NOAA
GOES-17 Channel	$\mathrm{Wm}^{-2}\mu\mathrm{m}^{-1}$	5 Minutes	NOAA
DSR	Wm^{-2}	60 Minutes	NOAA
Okta Cloud Cover	Percentage	60 Minutes	Weather Underground
Solar Generation Data	kW	1 - 60 Minutes	Energy Meter
Temperature	Celsius	60 Minutes	Weather Underground
Clear Sky Irradiance	Wm^{-2}	Minutes	Pysolar

Table 2: Summary of data sources, units, and resolution.

The GOES-16 and GOES-17 multispectral data is made available by NOAA as netCDF files downloaded from Amazon S3 buckets. We use a script to recursively download the data for specific dates each year along with the description of the ABI product, bucket, and the satellite name. The size of each 5 minute netCDF file is in the range of ~75MB, which requires nearly 16 terabytes to store two years of data from one satellite. Each 5 minute file includes data for all locations. To minimize storage requirements, we filter each file as we download it to extract only the channel data for the locations we are interested in, and discard the rest. The DSR data is also made available by NOAA in the form of netCDF files, but using a different mechanism, which currently requires manually submitting a request and then receiving an FTP link for download. Table 2 provides a summary of these data sources, their units, and their maximum resolution. In our experiments, we compare accuracy of the models at resolutions coarser than the maximum.

The netCDF files for multispectral and DSR data require some processing to filter out the data for the location of interest. Specifically, our python module reads the *goes_imager_projection* variable to convert (x, y) degree coordinates for latitude and longitude to radians. We then search the file for the latitude-longitude pair that is closest to our location of interest. Since these are geostationary satellites, their rotation matches that of the Earth, enabling us to look at the same part of the file each time. Thus, we read a file and first create a list of closest latitude longitude pair using the Vincenty formula [33], which calculates the distance between two points on the surface of a spheroid. This is done to reduce the computational resources so that the process of finding the nearest location is not repeated for each 5 minute file.

4.1 Physical Modeling Approaches

For comparison, we implemented two physical modeling approaches discussed in Section 2. These approaches are distinguished by the input data they use to estimate the effect of cloud cover on surface irradiance. We summarize these approaches below.

Okta-based Approach. This approach is described in prior work [21] and uses ground-level cloud cover readings in oktas to capture the atmospheric and cloud effects on solar power output. In particular, the approach uses a simple formula originally developed by Kasten-Gzeplank [27] to translate an okta-based cloud cover reading into a cloud index, which quantifies the percentage reduction in surface irradiance due to clouds. The approach then essentially multiplies this cloud index by a solar site's estimated

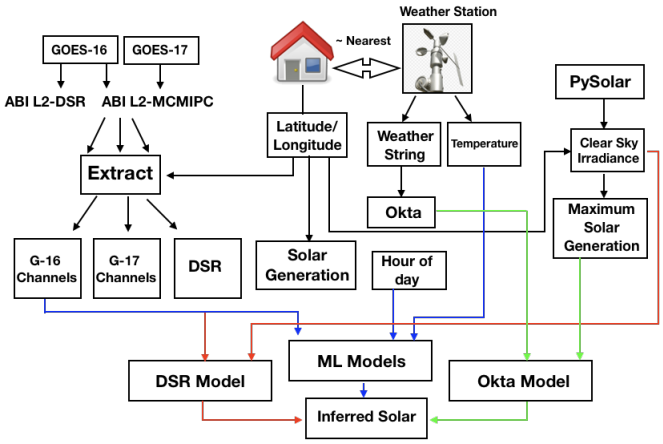

Figure 2: Diagram of data inputs for different solar performance modeling approaches we implement and evaluate.

maximum output at a given time. This maximum output is modeled in the same way as in our global model by finding the tightest upper bound on the data using a parameterized solar curve from the clear sky model. The tightest upper bound is used instead of a best fit, since the maximum solar generation is dictated by the clear sky model's solar curve. The parameters applied to the solar curve include an efficiency constant, which captures the solar site's conversion efficiency at 25°C, a temperature coefficient constant, which captures the effect of temperature on efficiency, and constants that capture the solar geometry (e.g., tilt and orientation angles). The approach searches for the parameters that dictate the tight upper bound, which provides an accurate model of a site's maximum solar output. More details are available in prior work [21, 22].

DSR Approach. This approach uses the same physical model as above, but instead of using oktas to compute a cloud index uses the DSR value computed from the GOES-16 satellite. As mentioned in Section 2, this DSR value is computed from the channel data using a sophisticated physical model [2], which yields an estimate of the surface radiation. We divide this DSR estimate for a location by the clear sky irradiance to yield a similar cloud index as above, and apply it in the same way. Note that DSR is often not made available, as certain conditions prevent it from being computed accurately, especially under overcast skies. In addition, the DSR technical report [2] evaluates its accuracy for estimating surface radiation and highlights that its accuracy degrades as the cloud cover increases, which are, unfortunately, exactly the times when solar performance modeling is most important.

Both approaches above are deterministic physical models that require calibration, e.g., by fitting known model function parameters to data, and do not require black-box ML training of unknown models. Calibrating parameters for well-known physical models is an advantage compared to using ML to learn these models. Figure 2 captures the different inputs, and data processing steps for our ML model and these two physical models.

5 EVALUATION

We evaluate our ML-based multispectral model and compare it to the physical models in the previous section on 29 sites across two years, which is currently the maximum data available from the satellites. We use two primary metrics in our evaluation: the Mean Absolute Percentage Error (MAPE) and the Capacity Error Percentage (CEP). The MAPE is computed as below, where S_t and P_t are the ground truth and model-inferred solar generation, respectively, at time t, and n is total number of temporal data points. The MAPE quantifies the average percentage error across time.

$$MAPE = \frac{1}{n}\sum_{t=0}^{n} |\frac{S_t - P_t}{S_t}|$$

We use MAPE because it is an intuitive metric that is comparable across solar sites of different sizes and configurations. However, MAPE is highly sensitive to periods of low absolute solar generation. For example, if solar generation for a 10kW site is only 10W early in the morning, and our model infers 40W, we record a 300% error, even though the 30W error is only 0.3% of the site's capacity. As a result, MAPE can be significantly affected by these large percentage errors that are actually small and insignificant absolute errors. Thus, we pair the MAPE metric with an absolute error metric, called CEP, that places less weight on these small absolute errors. We define CEP below as the absolute difference in watts between the actual (S_t) and inferred solar generation (P_t) divided by a site's maximum observed capacity (S_{max}). We use CEP instead of other absolute metrics, such as the Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE), because it is expressed as a percentage and thus is still comparable across solar sites of different sizes.

$$CEP = \frac{1}{n}\sum_{t=0}^{n} |\frac{S_t - P_t}{S_{max}}|$$

5.1 Performance of Local Models

We use 5-fold validation in all the experiments. This splits the entire data for each site into 5 sets each of which is used in turn as a test set and the remaining data as the training set. For each site, we compute the average MAPE and CEP for all 5 of the test sets and report the average with standard deviation. We use data from 2018-2019 for all evaluations except where specified.

Analysis on Individual Sites. We first study the performance of the proposed ML model using multispectral satellite data on all 29 solar sites. For this analysis we consider only summer months, May-September, and the middle of the day time period, 10am-3pm. This targets evaluation for sunny periods which are prone to less fluctuations and is the common time period used in evaluation in prior work [16, 21]. The results are shown in Figure 3, which shows the MAPE for all the individual sites in inferring solar generation over 15 and 60 minutes time resolution. Note that Okta is only available at a 60 minute resolution and the 15 minute resolutions are obtained by treating the Okta value as constant over the 60 minute period. On the x-axis, we have the different sites in the order of increasing MAPE under 60 minutes resolution, and the y-axis is the MAPE. For each site on the x-axis, we have four comparisons showing the performance of the Channel and Okta model under

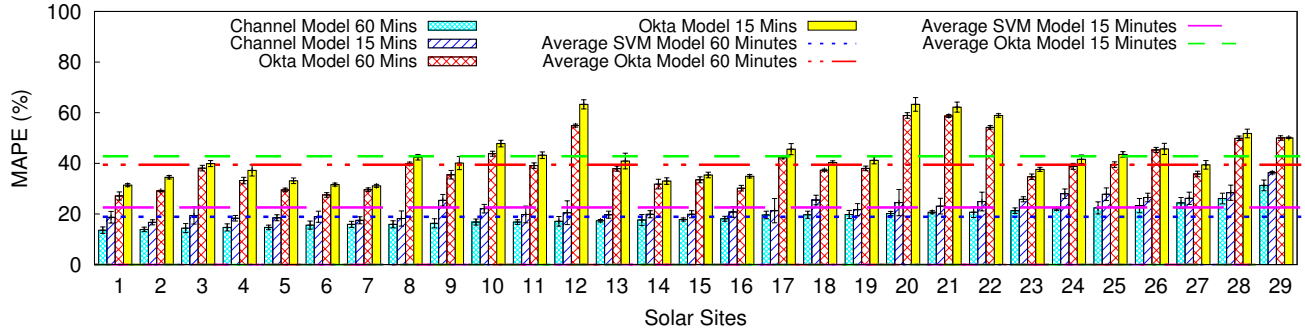

Figure 3: Performance comparison across the 29 sites. Data consists of only summer months and the middle of the day. Channel data shows consistently better performance.

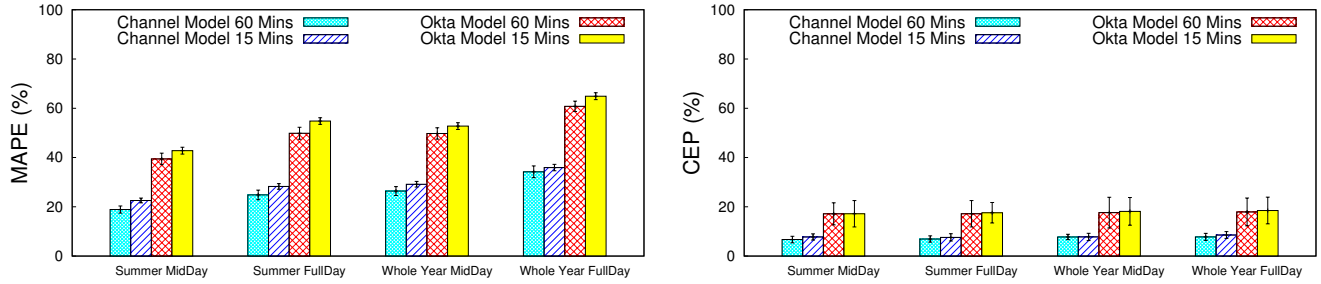

Figure 4: Performance comparison of local model under different time-periods. Average over 29 sites is shown. Left shows comparison using MAPE and right using CEP.

15 and 60 minutes time resolution. The average for the local model across all these sites are shown as flat lines for all the 4 cases.

We can see that the performance of these models is consistent across all these sites with the percentage error being the lowest for our Channel model. At 60 minutes resolution, which is the minimum resolution for Okta, we observe significantly better solar inference using the Channel model. On average across the sites at 60 minutes resolution, the Channel model gives a MAPE of 18.9% compared to 39.4% from Okta model. At 15 minutes resolution, the Okta performance worsens as expected. The Channel model still performs well with a slightly higher MAPE compared to the performance at 60 minutes resolution. Averaged across the sites at 15 minutes resolution, the Channel model yields a MAPE of 22.6% compared to 42.8% from Okta.

CEP and Performance under Different Conditions. We now analyze the performance for the models under different time periods throughout the year. The averaged result across the sites is shown in Figure 4. Summer refers to May-September and mid-day refers to 10am to 3pm. Again, we consider both 15 and 60 minute time resolutions. The trend is the same with 60 minutes Channel Model showing the lowest error and 15 minutes Okta Model showing the highest error under all scenarios. We can also see that the errors are lowest in the case of summer months and middle of the day time period because this eliminates the period of low solar generation, to which MAPE is sensitive. This is followed by summer months and full day time-period. This period includes data from sunrise to sunset for these summer months, thus eliminating the possibility of snow but still keeping the rainy and cloudy time periods. Furthermore, we can see that when we include the data for the whole year the MAPE further increases. Note that these are the averages across all the sites and contain the mix of sites with snow periods and sites with no snow throughout the year. In all these cases we observe that the Channel Model performs better by almost 50% in comparison to the Okta Model.

Since MAPE is very sensitive to periods of low solar generation, which will be frequent when the entire year is considered, we also analyze the performance with respect to the more balanced Capacity Error Percentage (CEP) metric in Figure 4 on the right. Since CEP is normalized with respect to the maximum solar generation for each site, we see that the percentage numbers are considerably lower and performance is comparable for summer mid-day and whole year. For instance, Channel model yields a CEP of 6.7% in the best case sunny scenario and 7.8% across the entire year. Moreover, Channel still leads to substantial error reduction compared to Okta.

DSR Comparison. To evaluate the performance of DSR with the Okta and Channel model, we show the comparison across 2 years worth of data in Figure 5. We have again used summer months and middle of the day to show the average results across sites under each model. The number of data points used in this evaluation differ from the other graphs because of the low availability of DSR data. The time periods where DSR value was not available were dropped so as to have the comparison of Okta and Channel model on the

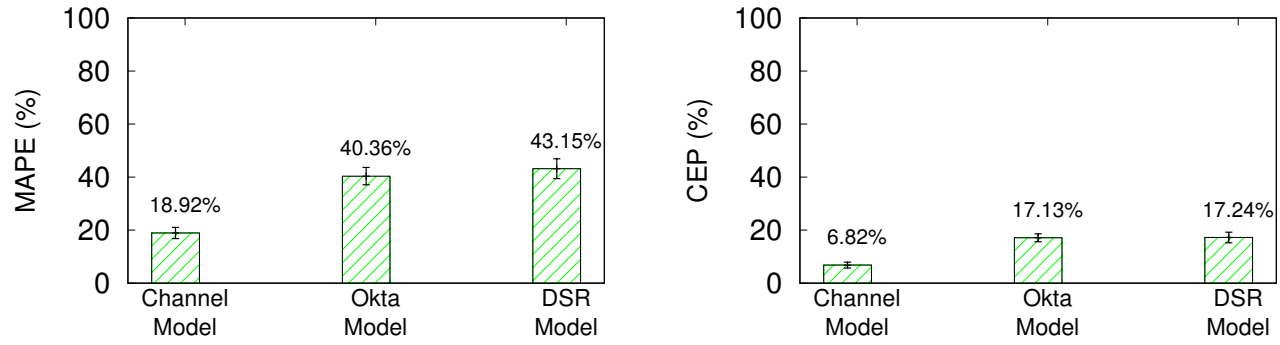

Figure 5: Performance comparison of Okta, DSR and Channels. Average across 29 solar sites over summer months and middle of the day is shown. Left shows comparison using MAPE and right using CEP.

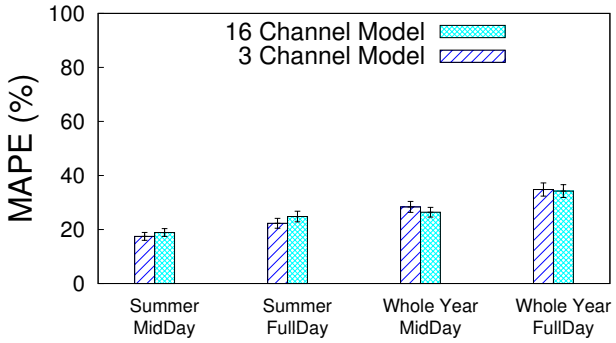

Figure 6: Performance comparison of first 3 channels versus 16 channels from GOES-16. Average MAPE across sites is shown.

same set of points as the DSR. DSR data is only 40% available as compared to the data from the weather station. Interestingly, we find that the DSR performs comparable to the Okta model. DSR analysis has been studied in the past [15].

Varying Channels. We also evaluate the performance of using the first three channels from GOES-16 versus using all 16 channels. The first three channels correspond to blue, red and veggie (green) bands and together form the visible region. Also we have seen from Figure 1, how the first three bands relate to the actual solar output at any given time. The higher channels correspond to higher wavelengths and indirectly contribute to the solar output by embedding information about cloud cover, water vapor, etc. For example, channel 6 and channel 11 have information about cloud particle size and cloud-top phase. Thus, it can be important to look at all the channels while modeling the solar output. In Figure 6, we have shown this comparison across different time periods. We can see here that the performance of 3 channels is slightly better than 16 channels when only summer months are analyzed. However, when considering the full year data, 16 channels perform slightly better. This is because summer months only capture the peak of solar generation data while under whole year we have different weather conditions ranging from snow to rain and cloudy periods. At those time periods, using all the16 channels gives better performance.

Multi-Satellite Models. We also analyze the performance when data from both GOES-16 and GOES-17 satellites are combined. As discussed before, the two satellites provide different view points of the same location from space and can provide complimentary information. Figure 7 shows the results, comparing both MAPE (on left) and CEP (on right). We have used year 2019 data for this evaluation as GOES-17 data is only available since 2019 [5]. We also evaluate here the performance for the models at 5 minute resolution on this data as the energy meter at the solar sites only store minute-level data for the most recent year. We see that combining the GOES-16 and GOES-17 data further improves the performance for the Channel model indicating that the two satellites provide complimentary information. For instance at 60 minutes resolution, using the combination of satellites improves MAPE from 19.3% to 13.7%. Moreover, even at 5 minute resolution, we observe good performance from the Channel models compared to the Okta model.

Comparing Different Regression Models. While we used SVM regression model in all of the analysis, we also compared different ML based models like decision tree regression, a simple linear regression, and SVM regression. Additional parameters, like the decision tree depth, were also estimated using 5-fold validation on the training data. Figure 8 shows the performance of these models. In this case, we perform the evaluation for 15 minute and 60 minute resolution. We can see from the graph that on both 15 and 60 minutes, SVM is performing best and has the highest accuracy, i.e., lowest MAPE. It is also evident that even a simple regression model performs generally well indicating a strong direct relationship between the channel data and the solar generation.

5.2 Performance of Global Models

We now move to the evaluation of the global model. While the local Channel model performs significantly better, it has the downside that it requires at least a year's worth of site-specific data for training. The global model, discussed in Section 3, overcomes this limitation and builds a general model that is applicable to any new site as long as we have one day's worth of data to calculate the site's physical parameters for maximum solar generation profile [21, 22]. We again use 5-fold validation for all evaluations. This now splits the entire data *based on the sites*, so that each fold consists of 1/5th of the sites. Each fold is then used in turn as a test set and the

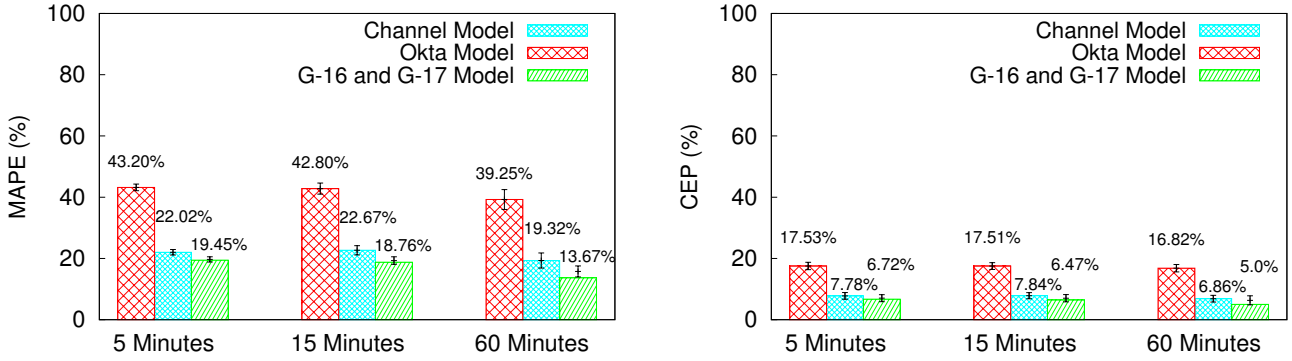

Figure 7: Combination of GOES-16 and GOES-17. Average MAPE (left) and CEP (right) across sites using 2019 data over summer months and middle of the day. Performance of the models at 5 minutes resolution is also shown.

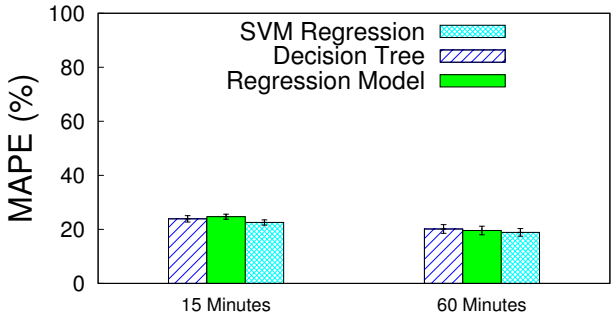

Figure 8: Performance comparison of different machine learning models.

remaining sites as the training set. For each fold, we compute the average MAPE and CEP across the sites in that fold and report the average with standard deviation.

In Figure 9 (left) we have shown our results for the global models. The data for these models are again at 60 minutes resolution for the summer months and middle of the day. The left graph covers only the GOES-16 satellite and hence hourly data of two years. We can see on the y-label the error percentages in the form of MAPE while the x-axis contains the local and global model results. The local model is the individual site-specific model evaluated previously and reproduced here for comparison. Note that the local model error is the average across all sites, where some data for each site was used in training the local model while the global model error is on new sites whose data was not used for training the global model. From the graph on the left, we can see that the Okta model has same performance under both local and global setting since it does not learn anything from the data. Comparing the global models, the Channel model still outperforms Okta by a large margin. The error of the global Channel model is higher than the local model. This is expected as the global model does not use any data from the test sites during training and hence will miss site-specific physical parameters, such as shading and location, which are modeled implicitly by the local model.

In Figure 9 (right), we have compared the performance of using data from only GOES-16 with the performance of using the combination of GOES-16 and GOES-17 satellites. Note that this only uses year 2019 data. We can see again that the combination model performs better but improvement is not as high as in the case of local model.

6 RELATED WORK

Our work differs from this prior work in multiple ways. First, we use data from the latest generation of U.S. satellites (GOES-16 and GOES-17) launched within the last two years, which includes 16 spectral bands instead of a single band in the visible spectrum. These 16 spectral bands include 3 bands that directly translate to solar generation. This data has much higher temporal and spatial resolution than prior satellites. Second, we do not use any physical models as part of our approach, and instead learn black-box machine learning models. As a result, our approach is highly accessible to those outside atmospheric sciences, which has generally been the domain of solar forecasting. We also use publicly available data from NOAA, so replicating our approach is possible for other researchers. We compare our ML models with a physical model of surface radiation provided by NOAA as a higher level data product called Downward Shortwave Radiation (DSR) [2]. Finally, our work also differs from prior work in that we focus on end-to-end modeling of solar power, rather than decoupling models of solar irradiance and solar power generation (given the irradiance). This approach is also more accessible, as real-time solar radiation estimates are not widely available, even though they are required as input to popular modeling frameworks, such as PVlib and SAM.

7 CONCLUSION AND FUTURE WORK

In this paper, we evaluate the performance of using multispectral channel data for solar modeling from the new range of GOES-r series of satellite(s) that cover the east and west CONUS domain of North America. We compare the performance of these channels with conventional ground-level okta-based measurements and a secondary satellite product, DSR, at different time granularities and

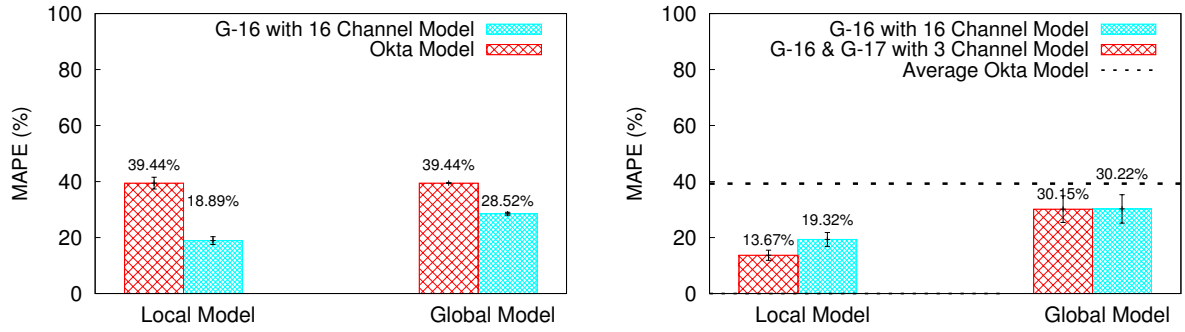

Figure 9: Performance comparison of global model. Average MAPE on year 2018-2019 data from GOES-16 (left) and year 2019 data from both GOES-16 and GOES-17 (right). Note that the global model performance is on new sites which are not part of training of the regression model.

at different times of the year. Our results show that the multispectral channel data performs better as compared to okta-based cloud measurements and DSR-based approaches by over 50% for local models and 25% for global models. Prior approaches were compared at one hour time granularity and only during sunny conditions whereas we compare our models at finer granularities of 5 and 15 minutes under different conditions throughout the year, with improved results. We also show the merits of combining data from GOES-16 and GOES-17 satellites.

Overall, our results show a strong correlation between satellite data and solar output, and lays a foundation for future work on using multispectral data for solar performance modeling. In future, this opens up avenues to explore satellite data for better forecasting of solar generation at minute-level resolutions, studying the effect of unmodeled parameters such as shading to further improve global models using multispectral data, and constructing hybrid models which incorporate both satellite and ground-level measurements for improved performance.

Acknowledgements. This research was supported by NSF grant #1645952.

REFERENCES

[1] 2007. PySolar. http://pysolar.org/.
[2] 2018. *GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Document for Downward Shortwave Radiation (Surface), and Reflected Shortwave Radiation (TOA).* Technical Report. NOAA NESDIS Center for Satellite Applications and Research.
[3] 2018. SolarClique: Detecting Anomalies in Residential Solar Arrays. In *COMPASS*.
[4] 2020. Extend: A Framework for Increasing Energy Access by Interconnecting Solar Home Systems. In *COMPASS*.
[5] 2020. SolarAnywhere. https://www.solaranywhere.com/2020/leveraging-goes-17-for-more-accurate-solar-data/.
[6] 2020. SunDown: Model-driven Per-Panel Solar Anomaly Detection for Residential Arrays. In *COMPASS*.
[7] 2020. The SunShot Initiative. https://www.energy.gov/eere/solar/sunshot-initiative.
[8] 2020. World now has 583.5 GW of operational PV. https://www.pv-magazine.com/2020/04/06/world-now-has-583-5-gw-of-operational-pv/
[9] Accessed 2020. ABI Technical Summary Chart. https://www.goes-r.gov/spacesegment/ABI-tech-summary.html.
[10] Accessed 2020. Bird Simple Spectral Model. http://rredc.nrel.gov/solar/models/spectral/.
[11] Accessed 2020. Registry of Open Data on AWS. https://registry.opendata.aws/noaa-goes/.
[12] Accessed 2020. SoDa - Solar Radiation Data. http://www.soda-pro.com/gl/.
[13] Accessed 2020. SolarAnywhere. https://www.solaranywhere.com/.
[14] R.W. Andrews, J.S. Stein, C. Hansen, and D. Riley. 2014. Introduction to the Open Source pvlib for Python Photovoltaic System Modelling Package. In *IEEE Photovoltaic Specialist Conference*.
[15] Akansha Singh Bansal and D. Irwin. 2020. Exploiting Satellite Data for Solar Performance Modeling. In *SmartGridComm*.
[16] N. Bashir, D. Chen, D. Irwin, and P. Shenoy. 2019. Solar-TK: A Data-driven Toolkit for Solar PV Performance Modeling and Forecasting. In *MASS*.
[17] H.G. Beyer, C. Costanzo, and D. Heinemann. 1996. Modifications of the Heliosat Procedure for Irradiance Estimates from Satellite Images. *Solar Energy* 56, 3 (1996).
[18] D. Cano, J. Monget, M. Albuisson, H. Guillard, N. Regas, and L. Wald. 1986. A Method for the Determination of the Global Solar Radiation from Meteorological Satellite Data. *Solar Energy* 37 (1986).
[19] P. Chakraborty, M. Marwah, M. Arlitt, and N. Ramakrishnan. 2012. Fine-grained Photovoltaic Output Prediction using a Bayesian Ensemble. In *AAAI*.
[20] David L. Chandler. 2011. Phys.org, Vast amounts of solar energy radiate to the Earth, but tapping it cost-effectively remains a challenge. https://phys.org/news/2011-10-vast-amounts-solar-energy-earth.html.
[21] D. Chen, J. Breda, and D. Irwin. 2018. Staring at the Sun: A Physical Black-box Solar Performance Model. In *BuildSys*.
[22] D. Chen and D. Irwin. 2017. SunDance: Black-box Behind-the-Meter Solar Disaggregation. In *e-Energy*.
[23] D. Chen, S. Iyengar, D. Irwin, and P. Shenoy. 2016. SunSpot: Exposing the Location of Anonymous Solar-powered Homes. In *BuildSys*.
[24] J.M. Freeman, N.A. DiOrio, N.J. Blair, T.W. Neises, M.J. Wagner, P. Gilman, and S. Janzou. 2018. *System Advisor Model (SAM) General Description (Version 2017.9. 5).* Technical Report. National Renewable Energy Lab.(NREL), Golden, CO (United States).
[25] Tam Hunt. 2015. GreenTech Media, Is There Enough Lithium to Maintain the Growth of the Lithium-Ion Battery Market? https://www.greentechmedia.com/articles/read/Is-There-Enough-Lithium-to-Maintain-the-Growth-of-the-Lithium-Ion-Battery-M
[26] S. Iyengar, N. Sharma, D. Irwin, P. Shenoy, and K. Ramamkritham. 2014. SolarCast - A Cloud-based Black Box Solar Predictor for Smart Homes. In *BuildSys*.
[27] F. Kasten and G. Gzeplak. 1980. Solar and Terrestrial Radiation Dependent on the Amount and Type of Cloud. *Solar Energy* 24, 2 (1980).
[28] R. Mohan, T. Cheng, A. Gupta, V. Garud, and Y. He. 2014. Solar Energy Disaggregation using Whole-House Consumption Signals. In *NILM Workshop*.
[29] D. Myers. 2006. Cloudy Sky Version of Bird's Broadband Hourly Clear Sky Model. In *Annual conference of the American Solar Energy Society (SOLAR)*.
[30] C. Rigollier, M. Lefevre, and L. Wald. 2004. The Method Heliosat-2 for Deriving Shortwave Solar Radiation Data from Satellite Images. *Solar Energy* 77, 2 (2004).
[31] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy. 2011. Predicting Solar Generation from Weather Forecasts Using Machine Learning. In *SmartGridComm*.
[32] R. Swanson. 2006. A Vision for Crystalline Silicon Photovoltaics. *Progess in Photovoltaics: Research and Appliations* 14, 5 (August 2006).
[33] Thaddeus Vincenty. 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey review* 23, 176 (1975), 88–93.