Exploiting Satellite Data for Solar Performance Modeling

Akansha Singh Bansal and David Irwin
University of Massachusetts Amherst

***Abstract*—Developing accurate solar performance models, which infer solar power output in real time based on the current environmental conditions, are an important prerequisite for many advanced energy analytics. Recent work has developed sophisticated data-driven techniques that generate customized models for complex rooftop solar sites by combining well-known physical models with both system and public weather station data. However, inferring solar generation from public weather station data has two drawbacks: not all solar sites are near a public weather station, and public weather data generally quantifies cloud cover—the most significant weather metric that affects solar—using highly coarse and imprecise measurements.**

In this paper, we develop and evaluate solar performance models that use satellite-based estimates of downward shortwave (solar) radiation (DSR) at the Earth's surface, which NOAA began publicly releasing after the launch of the GOES-R geostationary satellites in 2017. Unlike public weather data, DSR estimates are available for every 0.5km^2 area. As we show, the accuracy of solar performance modeling using satellite data and public weather station data depends on the cloud conditions, with DSR-based modeling being more accurate under clear skies and station-based modeling being more accurate under overcast skies. Surprisingly, our results show that, overall, pure satellite-based modeling yields similar accuracy as pure station-based modeling, although the relationship is a function of conditions and the local climate. We also show that a hybrid approach that combines the best of both approaches can also modestly improve accuracy.

I. Introduction

Solar energy generation has grown at nearly an exponential rate over that past 30 years, and is now cheaper than the retail price of electricity in many locations [1]. The goal for the U.S. Department of Energy's SunShot initiative is for solar to satisfy 14% of U.S. electricity demand by 2030 and 27% by 2050 [2], or a factor of 10× and 20×, respectively, greater than the 1.4% it satisfied in 2018 [3]. Reaching these targets will require improving solar performance models, which infer solar power output in real time based on current environmental conditions. These models are a prerequisite for a wide range of energy analytics, including solar forecasting, energy disaggregation, and grid simulations, that are necessary for grid operations and planning to accommodate higher solar penetrations.

To address the problem, recent work develops sophisticated data-driven modeling techniques that automatically derive a solar performance model for small-scale sites from public weather data, and thus are more scalable than prior manual approaches [4]–[7]. Once built, the model estimates a site's solar output at any time given the current weather conditions. Such data-driven models are highly accessible and useful for modeling any solar site in the U.S., as they rely only on well-known physical models of solar generation and public weather station data that is released in real-time for every location in the U.S. by the National Weather Service (NWS).

Unfortunately, using public weather station data has two primary drawbacks: not all solar sites are near a public weather station, and public weather station data generally quantifies cloud cover—the most significant metric that affects solar—using highly coarse and imprecise measurements [8], [9]. This measurement is in *oktas* and is often taken using a circular sky mirror placed on the ground that divides the sky into eight equal slices, such that the number of slices that contain a cloud translates to the number of oktas. The NWS then quantifies cloud cover using textual descriptions that map to a specific range of oktas. For example, "scattered clouds" maps to 3-5 oktas [9]. The imprecision of cloud cover measurements is by far the largest source of inaccuracy in large-scale data-driven solar performance modeling. Of course, while more accurate cloud cover measurements are possible using a pyranometer, which directly measures solar irradiance at the Earth's surface, only a few public weather stations include pyranometers.

Recently, National Oceanic and Atmospheric Agency (NOAA) in the U.S. has begun, as of 2018, releasing data products derived from a new generation of remote sensing geostationary satellites—the GOES-R series [10]. One of the secondary data products is the Downward Shortwave Radiation (DSR) that is incident at the Earth's surface, which estimates both the direct and diffuse solar radiation. Thus, DSR estimates the solar radiation available at the surface to generate solar power [11]. DSR is derived from the raw satellite data using a state-of-the-art algorithm that analyses the reflectance measurements of GOES-R's Advanced Baseline Imager (ABI) [12]. These DSR estimates account for cloud albedo, or the solar radiation reflected by clouds, and atmospheric conditions, and are available for any 0.5-2km^2 area within the satellite's view.

In contrast, the distance between a solar site and the nearest public weather station varies widely, and can be up to dozens of kilometers. In addition, unlike coarse okta measurements, DSR is a fine-grained measurement. Thus, using satellite data for solar performance modeling has the potential to address the drawbacks of public weather station data. However, satellite data also has drawbacks. While public weather station measurements are taken at the surface and represent ground truth, satellite measurements are taken from geostationary orbit, which is 35,800km above the Earth's surface. Satellites also

978-1-7281-6127-3/20/$31.00 ©2020 IEEE

can only measure the solar radiation reflected by the top of clouds, but cannot accurately assess cloud depth, height, or temperature, all of which affect the radiation that reaches the Earth's surface. As a result, unlike public weather station data, satellite data does not represent ground truth. Thus, while oktas provide coarse but direct measurements of surface radiation, satellites provide fine-grained but indirect measurements.

In this paper, we develop and evaluate a solar performance model that uses DSR, and compare it to a similar modeling framework that uses oktas. In doing so, we show how to integrate satellite data into an existing data-driven solar performance model from prior work [7], and examine multiple model variants that i) incorporate satellite data in lieu of public weather station data, and ii) use a combination of both. We have made our satellite-based modeling framework publicly available along with a solar and DSR dataset from nearly 50 sites that we have curated as part of our evaluation. To the best of our knowledge, this is the first use and evaluation of DSR for solar performance modeling, in part, because NOAA only began making this data product available in 2018.

Our work identifies strengths and weaknesses in using DSR satellite data for solar performance modeling. In particular, and contrary to our intuition, we find that using satellite-based DSR measurements *does not* improve the accuracy of solar performance models compared to using public weather station data. While DSR estimates provide slightly better accuracy during mostly clear skies, the estimates are much worse under overcast conditions. In most cases, DSR measurements are not even available during overcast periods due to these known limitations in accuracy under these conditions [12]. Thus, despite DSR's promise in other areas, especially long-term climate modeling, using public weather station data for solar performance modeling yields similar accuracy and is much more accessible. We do show that a hybrid approach that strategically uses satellite DSR data during mostly clear skies can modestly increase accuracy. In performing our data analysis, this paper makes the following contributions.

Satellite Data Background. We present background on the GOES-R series of satellites and the DSR data product, including its availability, accessibility, and ground truth accuracy. We also curate a new dataset that consists of hourly readings of solar generation, cloud cover in oktas, and DSR estimates for each of the 47 solar sites we analyze in our evaluation.

Exploiting DSR for Solar Performance Modeling. We show how to modify an existing data-driven solar performance model that uses cloud cover measurements from public weather stations to instead use satellite-based DSR measurements. We then illustrate salient differences between okta- and satellite-based measurements for a representative solar site. We define multiple model variants that combine satellite and okta data in different ways to understand their strengths and weaknesses.

Implementation and Evaluation. We implement the solar performance models above and evaluate them across the 47 solar sites in our dataset. Our evaluation shows that a physical model that uses okta-based measurements yields similar accuracy as using satellite DSR data, and that a hybrid approach can offer a modest improvement in accuracy.

II. Background

We provide background on measuring the impact of clouds using DSR and oktas, as well as on data-driven solar modeling.

A. Satellite-based DSR

There has been significant prior work on inferring solar irradiance incident at the Earth's surface using satellites. Much of this work, including the Heliosat family of algorithms [13]–[15], infers solar irradiance from visible satellite imagery, assuming a pixel's intensity is related to cloud cover. In contrast, GOES-R satellites include an Advanced Baseline Imager (ABI) that takes images of the Earth across 16 different spectral bands, which include two visible channels, four near-infrared channels, and ten infrared channels [16]. These 16 bands compare to only 5 bands from the previous generation of weather satellites and offer $4\times$ greater spatial and $5\times$ greater temporal resolution [17]. Specifically, the spatial resolution of contiguous U.S. (CONUS) is 0.5-2km^2 and the temporal resolution is every five minutes.

NOAA publicly releases the raw spectral data in near real time, as well as a large number of higher-level data products derived from this raw data. The raw data only began being released in 2018 (for GOES-16) and 2019 (for GOES-17) with higher-level data products being released later. Our work focuses specifically on a Level 2b+ data product that estimates the *downward shortwave radiation* (DSR) at the Earth's surface [11], which includes the ground-level direct and diffuse solar radiation in the visible, infrared, and near-infrared spectrums. Solar cells convert some fraction of DSR to electrical power based on their physical characteristics, e.g., power conversion efficiency, temperature coefficient, tilt, orientation, etc. DSR derives from a sophisticated physical model built on lower-level data products, e.g., for cloud optical depth, particle size, height, etc., that estimates cloud albedo and the atmosphere's composition, and represents the state-of-the-art in estimating radiation at the Earth's surface. That said, the DSR documentation quantifies its accuracy, which can vary widely depending on many factors, including the cloud characteristics, solar zenith angle, and latitude [11].

B. Ground-level Cloud Cover Measurements

Prior work on data-driven solar performance modeling combines clear sky solar irradiance models with ground-level cloud cover measurements in oktas, which are publicly available, to infer surface irradiance. Public weather stations typically report cloud cover as one of five weather strings, including clear skies (CLR), few clouds (FEW), scattered clouds (SCT), broken clouds (BKN), and overcast skies (OVC). These strings map directly to specific ranges of okta values [9]. Specifically, CLR maps to 0-1 oktas, FEW maps to 1-3 oktas, SCT maps to 3-5 oktas, BKN maps to 5-7 oktas, and OVC maps to 7-8 oktas. Prior work captures the relationship between cloud cover measured in oktas, and the clear sky index (CSI), which is the ratio between the actual irradiance at the surface divided by the irradiance at the surface under clear skies. For example,

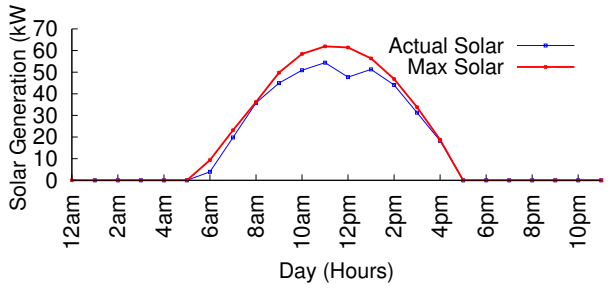

Fig. 1: Depiction of bounding solar generation using Equation 2.

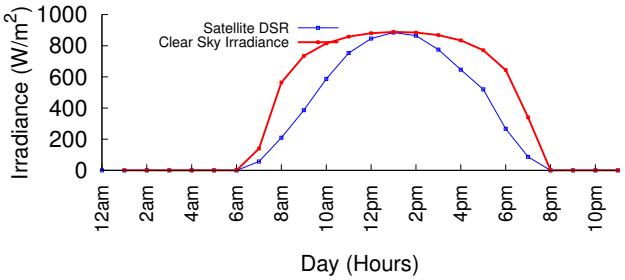

Fig. 2: Relationship between DSR and clear sky irradiance

in prior work, Kasten and Czeplak derived the empirical model below, which is widely used in textbooks [18].

$$CSI = 1 - 0.75 \times n^{3.4} \tag{1}$$

Here, n represents the fraction of cloud cover, e.g., by taking the midpoint of the okta range and dividing by 8. Chen et al. recently refined this empirical model using a much larger dataset [5]. Note that clear sky solar irradiance is a deterministic function of location, i.e., latitude and longitude, and time, and can thus be accurately estimated without any external inputs [19]. There are many software libraries, such as pysolar [20] and pvlib [21], that compute the clear sky solar irradiance given a location and timestamp. Thus, we can infer ground-level solar irradiance simply by multiplying the clear sky solar irradiance by the CSI from Equation 1 above, which is based on the cloud cover reported by public weather stations.

C. Data-driven Solar Performance Modeling

Our work builds on a simple data-driven solar performance modeling approach from prior work to quantify the accuracy of using satellite DSR estimates to infer solar generation [4], [5], [7]. We briefly summarize this approach, which we show how to modify in §III to incorporate DSR estimates. As input, the approach only requires a site's location and some historical generation data. The approach leverages the fact that a solar site's generation is always bounded by its maximum generation $P_{max}(t)$ described by the physical model below.

$$P_{max}(t) = I_{clearsky}(t) \times k \times (1 + c \times |T_{baseline} - T_{air}(t)|) \times [cos(90 - \Theta) \times sin(\beta) \times cos(\Phi - \alpha) + sin(90 - \Theta) \times cos(\beta)] \tag{2}$$

Here, $I_{clearsky}(t)$ is the clear sky solar irradiance at time t, and k is the solar site's efficiency parameter, which is a product of its size and solar conversion efficiency. Since solar conversion efficiency is a function cell temperature, the model multiplies k by an additional term. Here, c is the solar modules' temperature coefficient, while $T_{baseline}$ represents the baseline temperature when the conversion efficiency is k. Solar efficiency varies linearly with temperature, so the model multiplies the absolute value of the difference between the current temperature $T_{air}(t)$ and the baseline by the temperature coefficient. Typical values of c are $\sim$0.5%, such that efficiency increases this amount for every 1C drop in temperature. Finally, the lower term captures the impact of solar geometry: Θ and α represent the Sun's zenith and azimuth angles, respectively, while β and Φ represent the solar modules' tilt and orientation angles, respectively. These solar angles are a function of location and time, and can be computed using a library.

Prior work describes an efficient method for searching for values of k, c, $T_{baseline}$, β, and Φ in the equation above that yield the closest upper bound on the historical generation data [7]. The insight is that under clear skies, solar generation should conform to the model above (for some constant values of k, c, $T_{baseline}$, β, and Φ), while under cloudy skies, solar generation should be strictly less than the model above. Figure 1 depicts an example of bounding generation data using Equation 2.

After bounding the equation above to the data, we can compute $P_{max}(t)$ at any time t. The model then leverages the relationship below, which follows directly from Equation 2.

$$\frac{P_{actual}(t)}{P_{max}(t)} \sim \frac{I_{actual}(t)}{I_{clearsky}(t)} = CSI \tag{3}$$

To understand the relationship, observe that the only change in Equation 2 when computing actual solar output under cloudy skies is that we must replace $I_{clearsky}(t)$ with the actual solar irradiance under cloudy skies (assuming no change in temperature). All other parameters are independent of the cloudiness. As a result, when dividing $P_{actual}(t)$ by $P_{max}(t)$, everything on the right side of Equation 2 cancels out, which simply leaves $I_{actual}(t)/I_{clearsky}(t)$ or the clear sky index CSI.

Thus, given a model of $P_{max}(t)$ and the CSI, we can infer $P_{actual}(t)$ by simply multiplying the CSI by $P_{max}(t)$.

III. Satellite-based Solar Performance Modeling

We show how to use the data-driven solar modeling framework from the previous section to leverage both oktas and DSR, as well as a hybrid approach that uses both. Importantly, we use the same approach described in the previous section for each of these solar performance models, but where the clear sky index (CSI) is computed from different sources. In particular, satellite-based modeling computes the CSI using DSR data, while the okta-based approach computes it from the improved Kasten-Czeplank model [5]. As a result, any differences in modeling accuracy are only due to changing this input. In addition, since these models derive directly from the physical relationships in §II, they do not take into account any site-specific characteristics. Thus, for comparison, we also develop solar performance models using machine learning (ML) that can learn site-specific characteristics.

A. Satellite-based Model

Our satellite-based model is simple: to derive CSI, we take DSR directly from NOAA and divide it by a solar site's clear sky irradiance based on its location and time using a clear sky model. Note that this is a purely physical model that does not perform any regression to learn the relationship between DSR and solar output. Figure 2 shows the clear sky irradiance and DSR at a particular site. The graph shows DSR in watts per meter squared (W/m^2) and the corresponding clear sky irradiance over a representative clear day. The graph demonstrates that the clear sky irradiance is a strict upper bound on the DSR, such that the values are close when the sky is clear as expected. Interestingly, the values are nearly equal at solar noon, while the clear sky irradiance is slightly greater than DSR before and after solar noon.

Figure 3(left) shows the relationship between normalized DSR, or $DSR/I_{clearsky}$, and normalized solar generation, or P_{actual}/P_{max}, across many locations. The graph shows the normalized DSR on the x-axis and the normalized solar generation on the y-axis. As the graph shows, the the relationship is roughly linear, albeit noisy. This noise is largely due to inaccuracy in the DSR measurement, but may also result from unaccounted variables in our model, such as the presence of shading and topography at a solar site. We evaluate this relationship more fully for DSR in §V. A benefit of this approach, as discussed earlier, is that DSR is available every 0.5-2km^2, and thus provides more precise measurements than weather stations.

B. Oktas-based Model

In our oktas-based model, we compute the CSI using the ground-level cloud cover measurements provided by public weather stations. In this case, we use the mapping of each weather string—CLR, FEW, SCT, BKN, and OVC—to an oktas range. Since the NWS only specifies a coarse range for these values [9], we simply use the average of each range and map it to a number when computing the CSI. As explained above, we use this okta value for CSI in our data-driven solar model to infer the solar output. Again, this is a pure physical model that does not require learning a model from generation and weather data that is specific to a site. Figure 3(right) shows the relationship between oktas and the actual CSI for a particular location. Here, the x-axis is the ground-level cloud cover measurements (okta) and the y-axis is the actual CSI derived from the solar data (as discussed in §II). The graph shows that the okta-based measurements, while also noisy, do roughly follow the expected trend of the empirical models defined in prior work [5], [18]. In this case, the increased noise is largely due to the coarseness and imprecision of okta-based cloud cover measurements, which are derived from weather stations that are an unknown distance from each site.

C. Hybrid Model

As we show in §V, the satellite-based DSR model tends to be much less accurate than the oktas-based model when the cloud conditions are broken (BKN) or overcast (OVC), and slightly more accurate otherwise. The DSR documentation explicitly states this limitation of DSR, and, as a result, often does not even provide DSR readings when skies are cloudy [12]. To address this problem, we also design a hybrid solar performance model that leverages both oktas and DSR. This model uses the ground-level public weather station data as a filter by using the observed cloud cover to decide whether to use the satellite-based model or the oktas-based model. When the ground-level observation is CLR, FEW, or SCT, we use the satellite-based model (as described above), while we use the oktas-based model when the ground-level observation is BKN or OVC. This approach combines the best attributes of both models.

D. Machine Learning Models

The satellite-based, oktas-based, and hybrid models above all use physical models that are general and applicable to all solar sites. We also develop machine learning (ML) models that are specific to the characteristics of each site. ML models naturally capture unmodeled variables that are unique to each site, such as shading, which our physical models above cannot capture. However, one drawback of ML models is that they require sufficient data for training. Since some characteristics, such as shading, change throughout the year due to the seasons, this may require multiple years of data for solar models.

We train our ML models based on historical energy generation from each solar site. As with the hybrid model, these models combine DSR and oktas as input variables, as well as the clear sky irradiance and time-of-day/year. The dependent output variable is the site's actual solar generation under these conditions. We train our ML models using data from all of 2018, and use the data in 2019 for testing. Since DSR from the GOES satellites only recently became available, this is the maximum amount of training and test data that is available. In addition, since we train our ML models on each site individually, they implicitly incorporate site-specific physical characteristics that affect solar generation, which the physical models above do not, including the site-specific impact of non-ideal solar geometry (i.e., different panel tilts and orientations) and shading. The ML models are purely a black box and do not incorporate any of the physical models above in their training.

We evaluate two different common ML models: decision tree and support vector machines (SVMs). Decision trees are a flow chart-like structure where each internal node represents a test on a feature for classification and each leaf node represents a class label, while the branches represent features responsible for the class labels. In our decision tree, we used 10-fold cross-validation to select the tree depth from a maximum depth of 20 to avoid over-fitting. We also compare with SVMs, which attempt to fit as many datapoints with the kernel function while limiting margin violations. Under SVM with regression, we define a margin of tolerance (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 10-fold cross-validation in the following range: $\varepsilon \in \{0.01, 0.05, 0.1, 0.2\}$ and $C \in \{1, 10, 100, 1000\}$. For both ML models, we also add the hour of the day as an additional feature.

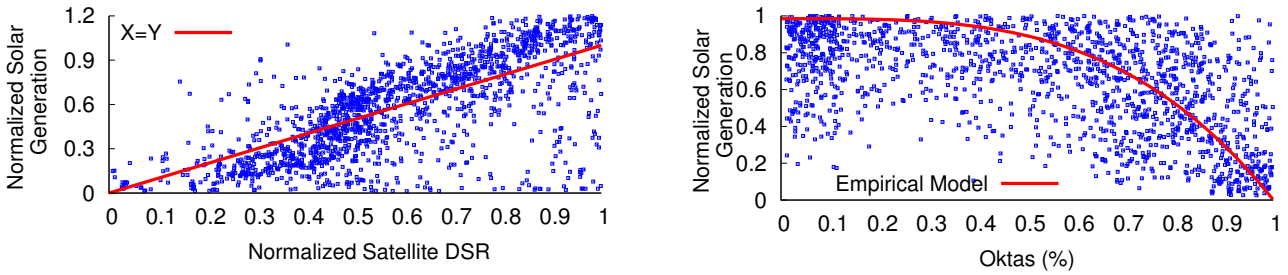

Fig. 3: Scatter plot of normalized solar generation versus normalized satellite DSR (left) and okta-based measurements (right) across many solar sites.

IV. Implementation

We implemented the solar performance models from the previous section as a python module, which we have publicly released as open source.[1] We use python's *scikit-learn* ML library to build the ML models in the previous section. We also used the numpy and pandas python packages to decode the NetCDF-formatted satellite data described below. Our module only requires a site's latitude and longitude as input, which it uses to compute clear sky irradiance using pysolar [20], a python library for simulating the solar irradiance at any point on Earth at any time. Similarly, our module programmatically fetches current and historical hourly temperature and cloud cover data from Weather Underground, a commonly used online weather website that maintains historical weather archives. Given a location, Weather Underground automatically determines the nearest weather station to that location. We also have access to two years of solar generation data from 47 homes. While we do not have physical access to all 47 homes, we can visibly observe many of their physical characteristics, e.g., size, shading, tilt, orientation, etc., in satellite imagery.

We use a web service provided by NOAA to access the satellite DSR data. Currently, users must download NetCDF-formatted files from an FTP server or via Amazon S3 buckets, as NOAA does not offer access to it via a web service with a programmatic interface. NetCDF is a common machine-independent data format for array-oriented scientific data. Users submit requests for data products, such as the ABI L2+ DSR product for the GOES-16 satellite, via NOAA's Archive Information Request System (AIRS) for up to 30 days. Once approved, NOAA sends the user a link via email to download the requested files (typically within an hour). Each file includes data for the entire contiguous U.S. for a single hour. As a result, our python module must decode the NetCDF data, and extract the DSR value for the sites of interest based on their latitude and longitude. Since extracting the DSR value for a site from the NetCDF file is non-trivial, we describe the process below.

Extracting Satellite Data. To extract a site's DSR, we must project the data file onto a geographic map. There is a summary option in each NetCDF file that gives all the variables available in the file. Specifically, the variable *goes_imager_projection* is essential for converting (x, y) coordinates for latitude and longitude in degrees to radians. Our python module uses this variable to extract the satellite sweep, longitude, and satellite height. The projected x and y coordinates equal the product of the scanning angle (in radians) and the satellite height. Following this projection, we can extract the latitude-longitude pairs in the form of a matrix from the NetCDF file. We calculate the nearest pair of coordinates from this matrix with our specified location using the Vincenty formula [22], which calculates the distance between two points on the surface of a spheroid. For the nearest computed location, we then extract the corresponding DSR value for the latitude-longitude pair.

As with the weather data, the satellite DSR is released hourly. Thus, we focus on solar performance modeling at an hourly resolution. Our python module combines the hourly temperature, cloud cover, satellite DSR, and solar generation for each location into a tabular format, e.g., a CSV file, with a corresponding timestamp for each reading. These data sources are stored in many different formats, particularly with different timestamps and time zones. As a result, our python module normalizes all timestamps and time zones to UTC time. Since our models currently do not account for snow, we focus on periods with no snow: May to October in 2018 and 2019. Incorporating snow is future work. Our primary metric is the Mean Absolute Percentage Error (MAPE) between our models and the ground truth, where a lower MAPE indicates less error.

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

Here S_t and P_t are the ground truth and model-inferred solar generation, respectively, at hour t, and n is total number of hourly data points. We use MAPE because it is an intuitive metric that is comparable across solar sites of different sizes. However, note that 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 400% error, even though the 30W error is only 0.3% of the site's capacity. Thus, when evaluating any single solar site, an absolute error metric, such as the Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) may be more appropriate. However, since our primary focus is comparing across sites with different sizes and characteristics, we continue to use MAPE, and mitigate its drawbacks by focusing on the 10am-3pm time period to eliminate periods that always have low absolute generation. Our primary focus is on the *relative* difference between the MAPEs of models in §III and not the absolute value.

[1] https://github.com/sustainablecomputinglab/satellite-dsr

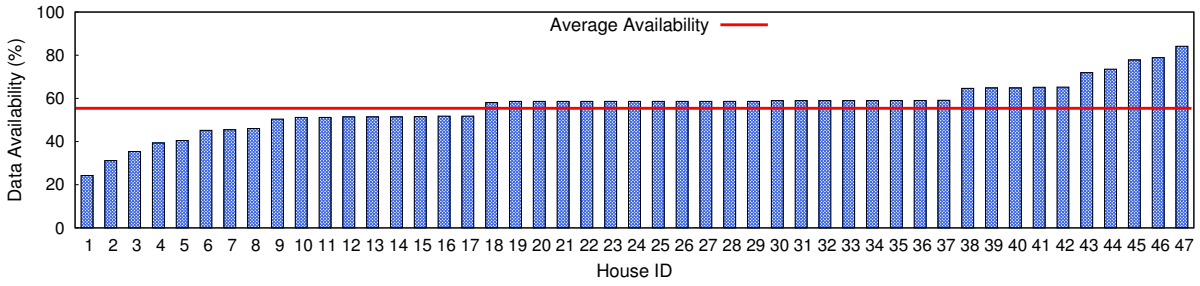

Fig. 4: Availability of DSR data product across our 47 solar sites.

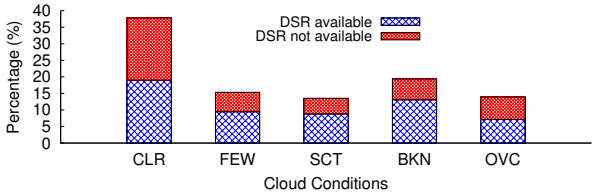

Fig. 5: DSR availability under different cloud conditions.

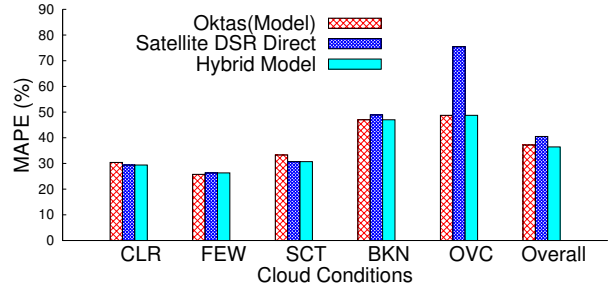

Fig. 6: MAPE for satellite-based, okta-based, and hybrid model 2018-19.

V. Evaluation

We evaluate our solar performance model using DSR on 47 solar sites. Unfortunately, however, DSR is unavailable during periods when its physical model is too uncertain [12]. On average, across our 47 sites, DSR is only available 55.4% of the time, although this differs across sites. Figure 4 shows the data availability across all 47 sites with a horizontal line at the 55.4% average. Figure 5 shows DSR's unavailability under different cloud conditions, and shows that this unavailability is higher during clear and overcast skies. This unavailability is currently a drawback to using DSR, especially during overcast skies as modeling solar performance is most important during these periods. Given this lack of availability, we restrict our analysis below to only those periods where DSR is available.

Physical Models. We first analyze the MAPE for our satellite-based, okta-based, and hybrid models from §III. Since these are physical models and do not require training, we can use the entire two-year dataset to evaluate their accuracy across all 47 sites. Figure 6 shows the overall results, as well as the MAPE under different cloud conditions. We find that, overall, the hybrid approach slightly outperforms the okta-based approach, and, surprisingly, the DSR approach performs the worst. As shown, the inaccuracy of the satellite-based DSR approach is due to its low accuracy during overcast conditions.

To emphasize the point, Figure 7 shows the MAPE under overcast cloud conditions for all 47 sites, and demonstrates that this performance for DSR is consistent across almost all of the sites with some sites reporting MAPEs in excess of 100% using DSR. However, as shown in Figure 5, since there are few overcast time periods where DSR is available, this inaccuracy does not contribute a significant amount to the overall results. Under all other cloud conditions, we observe a similar accuracy across the three techniques. Since our hybrid approach uses DSR when skies are not overcast and the okta-based approach otherwise, it slightly outperforms the pure okta-based approach. While our focus is on the relative difference between the models, the absolute MAPEs we find are similar to the okta-based models evaluated in prior work [5].

Machine Learning Models. For our ML models, we use 2018's data for training and 2019 for testing our decision tree (DT) and support vector machine (SVM) regression models. Figure 8 shows the overall MAPE for both our physical and ML models in 2019 under all cloud conditions, only overcast conditions, and all cloud conditions except overcast. We separate out overcast conditions since they are the most challenging conditions to model. We see that the ML models do not significantly improve upon our hybrid physical model, which does not require training. Overall, the hybrid model performs the best in all three cases, and is slightly better than the oktas-based model. The DT and SVM models actually perform worse than the satellite-based DSR model in overcast conditions. This poor performance may be due to the lack of training data in our dataset, as prior work uses multiple years of training data. Since DSR has only been available for two years, there is limited data available for training our models.

Key Point. The key takeaway point of our evaluation is that the current DSR data product released by NOAA, which represents the state-of-the-art in satellite-based estimates of surface irradiance, *does not* substantially improve solar performance modeling when compared with using okta-based measurements from weather stations. While DSR is slightly more accurate under non-overcast cloud conditions, it is significantly less accurate under overcast skies. In addition, DSR is also frequently unavailable, which is a significant drawback.

VI. Related Work

Solar performance modeling that infers a site's solar generation from its location, time, physical characteristics, and weather is a foundation for performing a wide range of solar analytics. There has been significant prior work on solar modeling and forecasting. Recent work on data-driven modeling develops techniques to automatically derive solar performance models for small-scale sites using public data, such as aerial imagery

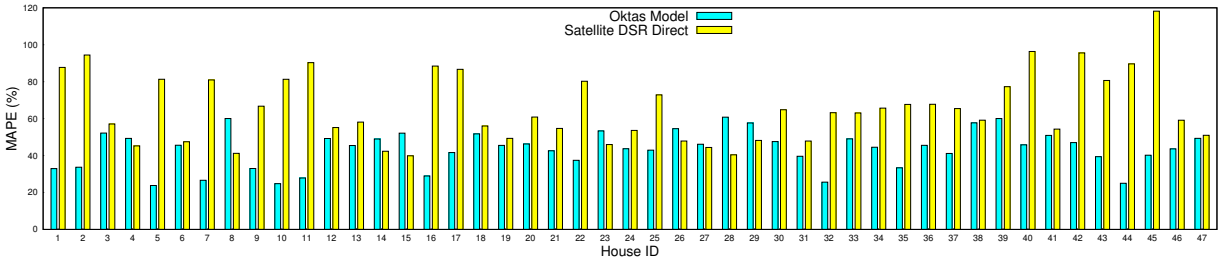

Fig. 7: MAPE for satellite-based and okta-based models under overcast cloud conditions.

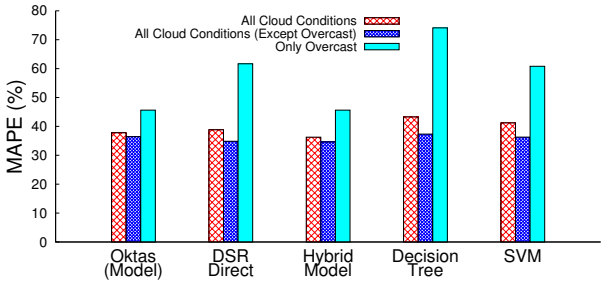

Fig. 8: MAPE for physical and ML models in 2019.

and weather data and thus are more scalable than prior manual approaches [4]–[7]. Using satellite data to infer ground-level irradiance has also been well-studied. For example, the Heliosat algorithm [23] is nearly 30 years old and uses visible satellite imagery to infer the global horizontal irradiance based on cloud cover. Our work differs from this and other work on this topic by specifically evaluating NOAA's DSR data product derived from the new generation of GOES satellites. These satellites were not launched until late 2017 and this data product did not become available until 2018. While recent work has compared DSR to ground-level irradiance measurements [24], we know of no work that has evaluated it for solar performance modeling.

VII. Conclusion and Future Work

This paper evaluates the use of DSR estimates from the new generation of GOES satellites for use in solar performance modeling. We show how to leverage DSR for solar performance modeling and compare it with okta-based and ML-based models. We show that the accuracy of satellite-based models depends on the cloud conditions. Surprisingly, our results show that pure satellite-based modeling yields similar accuracy as pure okta-based modeling with a hybrid approach that uses both showing only a modest improvement in accuracy. We also show that ML models are less accurate than physical models, although this may be due to limited training data. In future work, we plan to explore using the raw satellite data for solar performance modeling, rather than the secondary-level DSR data product, especially given DSR's high unavailability. By comparison, the raw hyperspectral satellite data is always available at a higher resolution (every 15 minutes).

Acknowledgements. This work is funded by NSF grant CNS-1645952.

References

[1] V. Shah and J. Booream-Phelps, "Deutsche Bank Markets Research, Crossing the Chasm," https://www.db.com/cr/en/docs/solar_report_full_length.pdf, February 27th 2015.

[2] "The SunShot Initiative," https://www.energy.gov/eere/solar/sunshot-initiative, Accessed February 3rd 2020.

[3] "U.S. Energy Information Administration, What is U.S. electricity generation by energy source?" https://www.eia.gov/tools/faqs/faq.php?id=427&t=3, Accessed February 3rd 2020.

[4] D. Chen and D. Irwin, "Black-box Solar Performance Modeling: Comparing Physical, Machine Learning, and Hybrid Approaches," in *Greenmetrics*, June 2017.

[5] D. Chen, J. Breda, and D. Irwin, "Staring at the Sun: A Physical Black-box Solar Performance Model," in *BuildSys*, November 2018.

[6] S. Lee, S. Iyengar, M. Feng, P. Shenoy, and S. Maji, "DeepRoof: A Data-driven Approach for Solar Potential Estimation Using Rooftop Imagery," in *KDD*, August 2019.

[7] N. Bashir, D. Chen, D. Irwin, and P. Shenoy, "Solar-TK: A Data-driven Toolkit for Solar PV Performance Modeling and Forecasting," in *MASS*, November 2019.

[8] "NASA, Make a Sky Mirror to Observe Clouds and Contrails," https://mynasadata.larc.nasa.gov/science_projects/make-a-sky-mirror-to-observe-clouds-and-contrails/, June 2019.

[9] "Weather.gov Terms," http://www.weather.gov/bgm/forecast_terms, 2018.

[10] "Geostationary Operational Environmental Satellites-R Series," https://www.goes-r.gov/, Accessed February 3rd 2020.

[11] "Data Products: Downward Shortwave Radiation (Surface)," https://www.goes-r.gov/products/baseline-DSR.html, Accessed February 3rd 2020.

[12] "GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Document for Downward Shortwave Radiation (Surface), and Reflected Shortwave Radiation (TOA)," NOAA NESDIS Center for Satellite Applications and Research, Tech. Rep., November 20th 2018.

[13] H. Beyer, C. Costanzo, and D. Heinemann, "Modifications of the Heliosat Procedure for Irradiance Estimates from Satellite Images," *Solar Energy*, vol. 56, no. 3, 1996.

[14] D. Cano, J. Monget, M. Albuisson, H. Guillard, N. Regas, and L. Wald, "A Method for the Determination of the Global Solar Radiation from Meteorological Satellite Data," *Solar Energy*, vol. 37, 1986.

[15] C. Rigollier, M. Lefevre, and L. Wald, "The Method Heliosat-2 for Deriving Shortwave Solar Radiation Data from Satellite Images," *Solar Energy*, vol. 77, no. 2, 2004.

[16] "Instruments: Advanced Baseline Imager (ABI)," https://www.goes-r.gov/spacesegment/abi.html, Accessed February 3rd 2020.

[17] "Instruments: ABI Improvements," https://www.goes-r.gov/spacesegment/abi-improvements.html, Accessed February 3rd 2020.

[18] F. Kasten and G. Gzeplak, "Solar and Terrestrial Radiation Dependent on the Amount and Type of Cloud," *Solar Energy*, vol. 24, no. 2, 1980.

[19] J. S. S. Matthew J. Reno, Clifford W. Hansen, "Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis," Sandia National Laboratories, Tech. Rep., March 2012.

[20] "PySolar," http://pysolar.org/, 2007.

[21] R. Andrews, J. Stein, C. Hansen, and D. Riley, "Introduction to the Open Source pvlib for Python Photovoltaic System Modelling Package," in *IEEE Photovoltaic Specialist Conference*, 2014.

[22] T. Vincenty, "Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations," *Survey review*, vol. 23, no. 176, pp. 88–93, 1975.

[23] R. H. Inman, H. T. Pedro, and C. F. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in energy and combustion science*, vol. 39, no. 6, pp. 535–576, 2013.

[24] S. Goodman, T. J. Schmit, J. M. Daniels, and R. J. Redmond, *The GOES-R series: a new generation of geostationary environmental satellites*. Elsevier, 2020.