Staring at the Sun: A Physical Black-box Solar Performance Model

Dong Chen*
Florida International University

Joseph Breda, David Irwin
University of Massachusetts Amherst

ABSTRACT

Developing accurate solar performance models, which estimate solar output based on a deployment's unique physical characteristics and weather, is increasingly important as the aggregate energy generated from solar rises. Since manually developing "white box" physical models based on site-specific information requires expert knowledge and thus does not scale, recent research focuses on "black box" approaches that use training data to automatically learn a custom machine learning (ML) model. Unfortunately, this approach requires months-to-years of training data, and often does not incorporate well-known physical models of solar generation, which reduces its accuracy. To address the problem, we develop a physical black-box modeling approach that leverages many of the same fundamental properties as existing white-box models.

Rather than manually determining values for physical model parameters, our approach automatically calibrates them by finding values that best fit the data. This calibration requires much less data (as few as 2 datapoints) than training a ML model, as the physical model already embeds the complex relationship between the input parameters and solar output. In developing our approach, we isolate the effects of 10 different weather metrics on solar output from nearly 343 million hourly weather and solar readings, or 78,435 aggregate years, gathered from 11,205 solar sites. We show that our physical model accurately describes weather's effect on solar output at all sites, obviating the need for training custom ML models using weather metrics. Instead, we augment our physical model by applying ML to learn only the relationships that are unique to each site, specifically non-weather-based shading. We evaluate our approach on solar and weather data from 100 sites, and show it yields higher accuracy than current state-of-the-art ML approaches.

CCS CONCEPTS

•Computing methodologies → Model development and analysis;

KEYWORDS

Solar Modeling, Net Meter

*Work performed while at the University of Massachusetts Amherst.

BuildSys '18, Shenzhen, China

978-1-4503-5951-1/18/11…$15.00
DOI: 10.1145/3276774.3276782

ACM Reference format:
Dong Chen and Joseph Breda, David Irwin. 2018. Staring at the Sun: A Physical Black-box Solar Performance Model. In *Proceedings of The 5th ACM International Conference on Systems for Energy-Efficient Built Environments, Shenzhen, China, November 7–8, 2018 (BuildSys '18),* 11 pages.
DOI: 10.1145/3276774.3276782

1 INTRODUCTION

The increasing impact of solar on the grid has motivated a strong interest in developing custom performance models that estimate a deployment's solar output based on its unique location, physical characteristics, and weather conditions. Solar performance models are useful for a variety of solar analytics, including solar monitoring [25], forecasting [16, 41], "behind the meter" disaggregation [22, 30, 35], anonymous localization [23, 24], and fault detection [14, 27, 28]. Recent research focuses on learning "black box" models [21, 29], primarily in the context of forecasting [16, 41], using machine learning (ML). Black-box ML approaches are attractive because they require only historical solar and weather data for training. Historical and current weather data are freely available for nearly every location in the U.S. from the National Weather Service (NWS) and many websites, such as Weather Underground [12]. Thus, utilities and third-parties that remotely monitor tens of thousands of solar deployments can directly use black-box techniques to develop performance models at large scales without any detailed site-specific information, which is often not available.

Interestingly, these black-box ML approaches are often "off the shelf" and do not leverage well-known physical models of solar generation based on fundamental physical properties. Instead, prior work on physical modeling generally takes a "white box" approach that assumes detailed knowledge of a deployment, such as the number and type of inverters and solar modules, as well as their rated capacity, efficiency, tilt, orientation, nominal operating cell temperature, and wiring. To develop white-box physical models, experts gather and translate this information into the parameters the models require. The PV Performance Modeling Collaborative distills a series of ten white-box modeling steps [39] implemented as part of the open source PVlib library [15]. Unfortunately, while white-box approaches may yield high accuracy, gathering the necessary information to construct these models at large scales for millions of small-scale deployments is infeasible. Thus, white-box models are typically only developed for utility-scale solar farms.

While recent black-box ML approaches do not require such site-specific information, they also have significant drawbacks. In particular, they require months-to-years of training data to derive accurate models [16, 29, 38], and thus are not immediately applicable to new solar sites coming online, or those that have not archived their historical data. In addition, "off the shelf" ML approaches often do not incorporate well-known physical models of solar generation based on fundamental properties, which reduce their accuracy. To

address the problem, we develop an approach to physical black-box modeling that leverages many of the same fundamental properties as existing white-box models. However, rather than derive physical model parameters from a manual site inspection, our approach calibrates them by finding the values that best fit the data. Since the physical model embeds detailed information about the relationship between the input parameters and solar output, this calibration typically requires less data than ML model training.

Our physical black-box model leverages well-known physical models that describe how a solar array's size, efficiency, tilt, and orientation affects its output. However, a significant challenge is that the precise physical relationship between each weather metric, e.g., cloud cover, pressure, dew point, humidity, etc., reported by a weather station and solar output is not well-known. This is not an issue for ML models, which can automatically learn any unknown relationships from observed data. We address this challenge by isolating the effects of 10 different weather metrics on solar output from nearly 343 million hourly weather and solar readings, or 78,435 aggregate years, gathered from 11,205 solar sites.

Our analysis shows that only 2 weather metrics affect solar generation—temperature and cloud cover—and that their effect is universal and independent of time and location after normalizing for a deployment's physical characteristics. While temperature's linear effect on solar efficiency is well-known, our analysis shows that commonly-used models for estimating a location's global horizontal irradiance (GHI) based on cloud cover are inaccurate [15, 31, 32]. We improve these existing cloud cover models in developing our approach, which estimates a site's solar output at any time based on widely-available temperature and cloud cover readings. Unlike prior ML approaches, which require months-to-years of data to train accurate models, our model requires as few as 2 datapoints to calibrate a specific solar site. Thus, our approach immediately enables highly accurate solar performance models anywhere.

We evaluate our physical black-box model using solar and weather data from 100 sites, and show that it yields similar or higher accuracy than current state-of-the-art ML approaches. We also show that our model, which uses imprecise cloud cover measurements, has better accuracy than a performance model using GHI estimates derived from visible satellite imagery. Intuitively, our approach demonstrates that the effect of weather and many physical site characteristics on solar output are universal and common across all solar sites, and thus do not need to be "learned" separately at each site using ML. Of course, there are unique aspects of each solar site that do affect solar output, particularly non-weather-based shading, e.g., from nearby buildings, trees, and mountains, which our physical model does not capture.

Thus, we augment our approach by applying ML to learn only how much unique site-specific shading effects decrease the solar output expected by our physical model. As we discuss, these site-specific effects are largely a function of the Sun's azimuth and zenith angles. We show that our ML-enhanced physical black-box model further improves accuracy, especially at sites and during periods with significant shading from obstructions. While prior ML approaches may indirectly learn shading effects, e.g., by including time as one of their input features, they conflate them with other effects, such as weather, which are accurately described by physical models. In contrast, our approach distills the input features—the Sun's azimuth and zenith angle—that directly determine a site's unique shading effects on the solar output estimated by our physical model. As a result, our ML-extended physical model yields much higher accuracy than current state-of-the-art ML approaches across all 100 sites in our evaluation.

2 BACKGROUND

A solar performance model is simple: given a site's location and its current weather conditions over some time interval τ as input, it returns an estimate of average solar power generation as output over τ. The NWS and many websites publicly report current measurements for numerous weather metrics, including cloud cover, temperature, pressure, dew point, humidity, etc., at every location in the U.S. every hour. Historical weather archives at every location every hour are also available. Black-box approaches use only these historical weather archives, a site's location, and its solar generation data to derive a performance model. A variety of supervised ML techniques, such as deep neural nets (DNNs) and support vector machines (SVMs), are capable of learning a model from training data that maps weather metrics to solar output. However, since solar potential varies each day and over the year, this approach requires learning a separate model for each time period, which significantly increases the training data required to learn an accurate model, as each sub-model requires distinct training data [38].

To reduce the size of the training data, ML-based modeling can normalize the solar output based on time, such that it can use each datapoint to learn a single model [29, 35]. For example, prior work normalizes solar output by dividing the raw solar power output by the solar capacity, defined as the system's maximum generation over some previous interval, e.g., a year, which it calls the solar intensity [35]. In addition to weather metrics, the approach also adds the time of each datapoint to the set of input features, along with the time of sunrise and sunset. Including time as a feature enables the model to automatically learn the solar generation profile. For example, a time closer to sunrise or sunset will have a lower solar intensity, even in sunny clear sky conditions, compared to a time closer to solar noon. The approach then trains a model using a SVM with a Radial Basis Function (RBF) kernel, which is common in solar modeling, since it attempts to fit a Gaussian curve to solar data and solar profiles appear similar to Gaussian curves [17, 35, 38].

2.1 Modeling Physical Characteristics

The canonical ML approach above is completely data-driven and does not incorporate any physical models of solar generation, other than the insight that solar intensity varies over time under clear skies similar to a Gaussian function. However, detailed physical models exist that dictate solar potential under clear skies based on a deployment's location, time, size, efficiency, module tilt, and module orientation. For example, there are numerous clear sky irradiance models with varying levels of complexity, which accurately estimate the global horizontal irradiance (GHI), in Watts per meter squared (W/m^2), at any location on Earth at any time under clear skies [33]. These models are implemented by many open source libraries [1, 9, 15]. Of course, while the GHI represents the maximum solar power available to a solar module to convert to electricity, solar modules are not 100% efficient. Instead, their

efficiency varies based on the type of module, e.g., poly- versus mono-crystalline, their size and efficiency, as well as their orientation and tilt. The well-known equation below computes the power $P_s(t)$ a solar module generates at any time t based on its tilt (β) and orientation (ϕ) relative to the Earth's surface, and the Sun's zenith (Θ) and azimuth (α) angles, which are a well-known function of location and time [19]. $I_{incident}$ is the solar irradiance incident on the module, which under clear skies is determined by the clear sky model above, while k is a module-specific parameter that combines conversion efficiency (as a percentage) and module size (in m^2). A similar expression exists for modules that track the sun.

$$P_s(t) = I_{incident}(t) * k * [\cos(90 - \Theta) * \sin(\beta) * \cos(\phi - \alpha) + \sin(90 - \Theta) * \cos(\beta)] \quad (1)$$

White-box models can directly measure the module angles, size, and efficiency. While black-box models cannot directly measure these values, given the relationships above, they can search for these parameters via curve fitting, as shown in prior work [22]. This process sets the tilt and orientation to their ideal values (a tilt equal to the location's latitude and a south-facing orientation in the northern hemisphere), and then conducts a binary search for the k that both minimizes the Root Mean Squared Error (RMSE) with the observed data and represents a strict upper bound on the data, as we know generation should never exceed the maximum dictated by the clear sky GHI. As a result, the process is robust to degraded solar output under cloudy skies, as long as there are some datapoints under clear skies. After fitting k, the process then iterates by conducting a similar binary search for orientation and then tilt, since changing any parameter value affects the others. The iterative search terminates when the new value for each parameter does not significantly change. Prior work shows that this search results in highly accurate values for k and the orientation (ϕ) and tilt (β) angles [22]. Note that, in the limit, accurately determining these parameters requires only a single datapoint under clear skies, as only one datapoint is necessary to determine the upper bound.

2.2 Modeling Weather Effects

The model above accurately estimates a site's maximum solar power generation under clear skies by setting $I_{incident}$ in Equation 1 equal to the current clear sky GHI. However, numerous other factors, particularly the weather, also affect $I_{incident}$ and $P_s(t)$. Physical models also exist for important weather metrics, as described below.

Temperature Effects. The Nominal Operating Cell Temperature (NOCT) model describes the effect of temperature on solar cell efficiency [18]. Specifically, for every degree increase (or decrease) in T_{cell}, module efficiency drops (or rises) by roughly a constant percentage, which varies between modules, but is ~0.5% per degree Celsius. Thus, to account for temperature we can re-calibrate the model above by adjusting the parameter value k based on the temperature at each datapoint using the equation below, where $T_{baseline}$ is the temperature at the datapoint that is closest to the upper bound solar curve that minimizes RMSE found in the search [22].

$$k'(t) = k * (1 + c * (T_{baseline} - T_{air}(t))) \quad (2)$$

Note that, while efficiency varies strictly based on cell temperature, the cell temperature's relationship to the air temperature T_{air} differs only by an additive constant, which cancels out when subtracting two cell temperatures, enabling us to use ambient air temperatures in the equation. The baseline air temperature $T_{baseline}$ that bounds the curve will represent the coldest point in the dataset under a clear sky, since this is the most efficient operating point that maximizes the fraction of clear sky GHI converted to solar power. Again, the process conducts a binary search for the value of c that both minimizes the RMSE with the observed data, and is also a strict upper bound. In the limit, determining an accurate value of c requires only 2 datapoints under clear skies that exhibit a difference in temperature. Prior work shows that estimating the model's parameters using only 2 days of data yields a similar accuracy (under clear skies) to estimating them using a year of data [22].

Cloud Cover Effects. Using the temperature adjustment above, the physical models can estimate a site's maximum solar generation under clear skies at a given air temperature. However, skies are not always clear, such that the GHI at the Earth's surface is much less than the clear sky GHI. Cloud cover is the primary metric that dictates the fraction of the clear sky GHI that reaches a solar module. As above, there are well-known physical models that describe the effect of cloud cover on clear sky GHI [11, 31, 32, 36]. For example, PVlib implements both a linear cloud cover model, and the Liu-Jordan model (1960) [32]. The latter is, in part, a function of the Sun's zenith angle. In contrast, the Kasten-Czeplak model (1980), which is commonly-used in textbooks [11], is independent of the solar angle, i.e., time and location. The Kasten-Czeplak model is below—it was originally derived empirically based on hourly cloud cover observations and GHI measurements in Hamburg, Germany over a 10 year period (1964-1973) [31].

$$I_{incident}/I_{clearsky} = (1 - 0.75n^{3.4}) \quad (3)$$

Here, $I_{incident}$ represents the solar irradiance that reaches the module, $I_{clearsky}$ represents the GHI from the clear sky model, and n represents the fraction of cloud cover (0.0-1.0). This cloud cover (or sky condition) is measured in *oktas*, which represents how many eighths of the sky are covered in clouds, ranging from 0 oktas (completely clear sky) to 8 oktas (completely overcast). The cloud cover reported by the NWS and other websites translates directly to an okta range [13]. For example "Clear/Sunny" is <1 okta, "Mostly Clear/Mostly Sunny" is 1-3 oktas, "Partly Cloudy/Partly Sunny" is 3-5 oktas, "Mostly Cloudy" is 5-7 oktas, and "Cloudy" is 8 oktas. Okta measurements are typically taken using a circular sky mirror divided into eight slices, such that when the mirror is placed on the ground, the oktas are equivalent to the number of slices with a cloud present [3]. Thus, oktas are an imprecise measure of cloud cover, which does not account for cloud type or thickness.

Equation 3 enables us to adjust the physical model above by multiplying the solar output $P_s(t)$ in our temperature-adjusted model above by the fraction $I_{incident}/I_{clearsky}$. Note that, while Equation 3 is in terms of solar irradiance and not solar power, the ratio of observed solar power to the maximum solar generation estimated by the model after the temperature adjustment (from Equation 1) should be equivalent to Equation 3, since the effect of the physical characteristics (k, tilt, and orientation) all cancel out in the division, leaving only the ratio of the observed incident irradiance ($I_{incident}$) to the clear sky GHI ($I_{clearsky}$). We show this

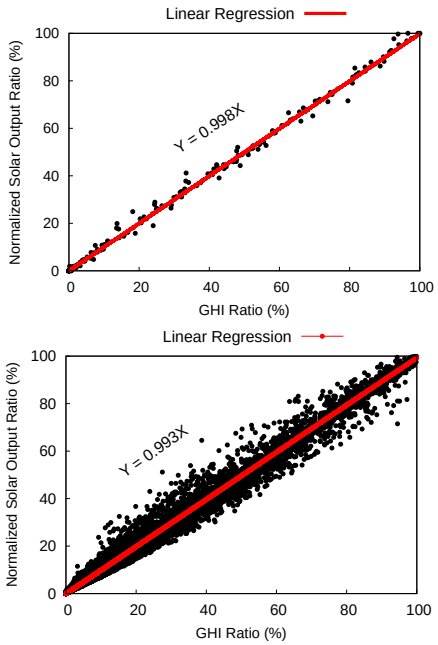


Figure 1: Normalized solar output as a function of GHI from a solar radiation sensor (top) and satellite imagery (bottom).

empirically in the next section. Thus, multiplying the temperature-adjusted $P_s(t)$ by $I_{incident}/I_{clearsky}$ yields a solar power estimate for a site with a specific size, module efficiency, tilt, and orientation under cloudy skies at a specific location, time, and temperature.

Other Weather Effects. The NWS and other weather sites, such as Weather Underground, report numerous other weather metrics, including dew point, humidity, visibility, pressure, precipitation intensity, precipitation probability, wind speed, and wind bearing. While some work has examined the effect of a few of these metrics on solar output [34, 37, 40], their effects are still not well understood and there are no commonly-used white-box physical models for them. Black-box ML approaches generally include these additional weather metrics as input features in case they do affect solar output.

Summary. The black-box model above is entirely based on well-known physical solar models that incorporate the effect of module size, efficiency, location, time, temperature, cloud cover, tilt, and orientation on solar output. The model is black-box, since it can determine the unique parameter values for each site, given its location, with a small amount of data for calibration. Of course, black-box ML models that are purely data-driven can potentially learn these relationships, given enough training data from a site. However, since these relationships are based on fundamental physical properties common across all solar sites, re-learning them at every site is wasteful and unnecessary. *Unfortunately, as we show in §6, the black-box physical model we describe above is highly inaccurate and performs significantly worse than black-box ML models.* This inaccuracy derives from either the physical models above being inaccurate, or from the effect of unmodeled physical parameters, such as the unmodeled weather metrics or shading from surrounding buildings. In the next section, we conduct a large-scale data analysis to determine the primary source of the inaccuracy by isolating the effects of 10 different weather metrics on solar output.

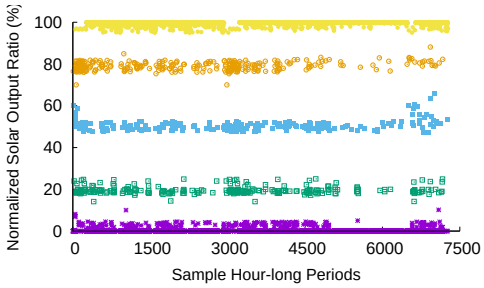


Figure 2: Normalized solar output ratio across many solar sites and hour-long time periods for 5 different clusters, indicated by the 5 colors, with the same weather.

3 LARGE-SCALE WEATHER DATA ANALYSIS

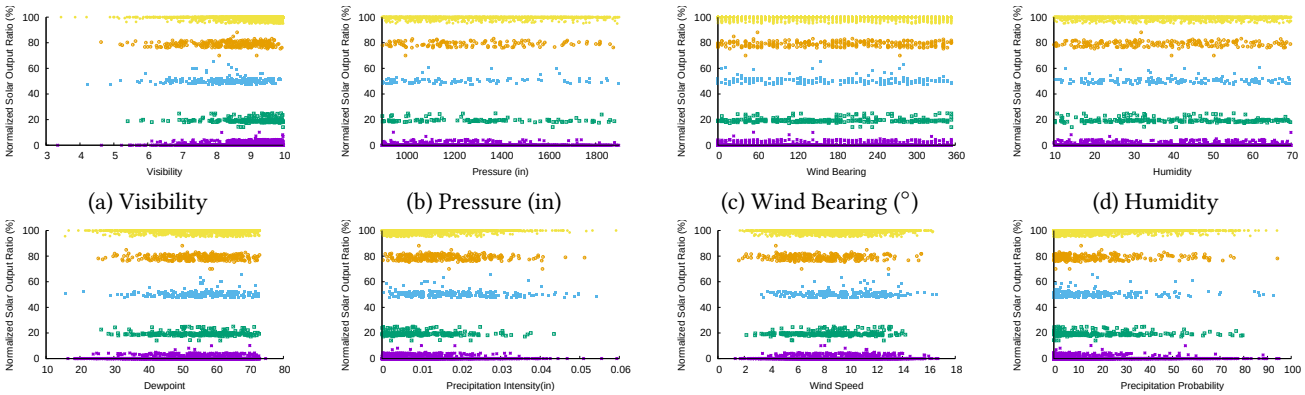
To isolate the effect of each weather metric on solar output, we must normalize for the effects of all the other variables. We use the models of physical characteristics from §2.1 to normalize for the effect of module/array location, size, efficiency, tilt, orientation, and the time of the day and year. This normalization divides the raw solar power observation by the maximum solar generation estimate under clear skies from our model, which, as we discuss in §2.2, should be the same as ratio of the observed GHI to the clear sky GHI, modulo any module shading or soiling effects. Figure 1 verifies this by plotting our normalized solar output ratio (on the y-axis) as a function of the GHI ratio (on the x-axis), which is also called the clear sky index. Figure 1(top) uses a solar radiation sensor deployed at a single unshaded solar site to determine the GHI ratio. However, since most sites do not have an on-site solar radiation sensor, Figure 1(bottom) computes the GHI ratio using the Heliosat-3 algorithm, which estimates it from visible satellite imagery [26]. Both figures show that our normalized solar power ratio using the physical model from §2.1, including the temperature effect, is close to the GHI ratio, with a linear regression line close to y=x.

Importantly, our normalization makes the solar output from many different sites directly comparable. Thus, our data analysis normalizes hourly solar output data from nearly 343 million hourly weather and solar readings, or 78,435 aggregate years, gathered from 11,205 solar sites. We gathered this data from public sources, including Pecan Street's DataPort [6] and PVoutput [8]. We gathered weather data from Weather Underground's API [12], which includes current and historical data from 180k weather stations in the U.S. To isolate the effect of 10 weather metrics on solar output, we examine clusters of datapoints where *all 10 weather metrics have nearly the same value.* One reason we use so much solar data is that finding a statistically significant number of hours where all 10 metrics are the same is difficult at a single site (or even a few sites), as the weather across these 10 metrics varies too much. Even in our massive-scale dataset, identifying sufficiently large clusters is challenging. As a result, our clusters only ensure that each of the 10 metrics is within a small range. We suspect this massive data requirement is one reason a similar analysis has not been conducted in prior work. In contrast, the Kasten-Czeplak cloud cover equation was derived from 10 years of hourly data at only a single site [31].

Figure 2 plots the normalized solar output ratio on the y-axis for 5 example clusters, which each include 7500 hourly datapoints on the x-axis. We use the multi-dimensional k-Means algorithm to identify

Metric	Visibility	Pressure (in)	Wind Bearing	Dewpoint (°F)	Precipitation Intensity(in)	Wind Speed(mph)	Humidity (%)	Precipitation Probability(%)	Cloud Cover(%)
Cluster 1	9.69	1216.51	189.25	44.83	0.008	7.81	48.15	33.13	94.79
Cluster 2	11.53	1211.06	203.90	43.31	0.019	9.05	48.36	31.21	77.86
Cluster 3	11.77	998.10	179.07	41.69	0.022	8.41	58.39	35.03	40.97
Cluster 4	9.58	1103.03	169.33	55.01	0.017	6.11	38.60	29.18	21.85
Cluster 5	11.47	1198.61	180.03	49.84	0.011	8.32	35.71	38.19	1.83

Table 1: The centroid for each of the 5 clusters of weather metrics.



(a) Visibility (b) Pressure (in) (c) Wind Bearing (°) (d) Humidity

(e) Dewpoint (°F) (f) Precipitation Intensity (in) (g) Wind Speed (mph) (h) Precipitation Probability

Figure 3: The isolated effect of 8 unmodeled weather metrics on the normalized solar output ratio for our 5 clusters.

these clusters, such that we define a threshold for the distance to each cluster's centroid to control the number of datapoints. For each cluster, we select the minimum distance threshold necessary to ensure 7500 datapoints. Table 1 shows the weather metric value at the centroid of each cluster. Note that these hourly datapoints are from different times from numerous solar sites with different locations and physical characteristics: the only commonality is the value of the 10 weather metrics within each cluster. We explicitly selected these illustrative clusters to yield different ratios on the y-axis. Importantly, since each cluster has the same normalized output ratio across all hour-long time periods, Figure 2 shows that the same weather conditions have the same percentage effect on the normalized solar output ratio at any site, regardless of the time, a site's location, or its other physical characteristics. That is, weather's effect is *universal* across all solar sites in these clusters.

We can isolate the effect of any single weather metric in our 5 example clusters by removing it from the cluster, and including all datapoints where the other 9 metrics are within the cluster, but the removed metric can take on any value. Doing so, empirically demonstrates the effect of only that single weather metric on the normalized solar output ratio (modulo any shading or soiling effects) across our massive dataset. Note that we have already applied the temperature adjustment to all these ratios based on each site's temperature coefficient using the physical model from §2.2. We empirically validated the linear NOCT physical model that describes the temperature effect using our data, but omit the graph due to space constraints. Figure 3 shows the results that isolate the effect of the 8 other weather metrics without physical models. The graphs show these weather metrics have no significant effect on solar output, as the normalized solar output ratio remains the same regardless of the metric's value, whether extremely high or low. That is, the lines are horizontal with a value equal to that from Figure 2. Thus, these metrics are not useful in estimating solar performance, and need not be included when training ML models.

In some cases, our data analysis seems counter-intuitive. For example, our analysis shows that visibility has no significant impact on solar generation after normalizing for cloud cover. In addition, our weather data shows there are periods where both visibility and cloud cover are low (i.e., no clouds but low visibility), and visibility and cloud cover are high (many clouds, but high visibility). Since our work focuses on large-scale empirical data analysis, we have not observed the physical properties that would yield such datapoints, and leave an examination of them to future research.

Unlike the other weather metrics, Figure 4(top) isolates cloud cover, which demonstrates a clear non-linear relationship with the normalized solar output ratio. At first glance, the relationship appears similar to the Karston-Czeplak model from §2.2. However, Figure 4(bottom) plots the normalized temperature-adjusted solar output ratio as a function of cloud cover for a larger random sampling of our entire dataset (and not just from the 5 clusters), as the entire dataset is too large to fit on a graph. The graph shows that, while imprecise, the datapoints follow the same trend as in Figure 4(top). The imprecision is not surprising, given the imprecision inherent to oktas. In addition, module shading and soiling, which can cause the ratio to be lower than expected based solely on weather, also contributes to the imprecision. We also graph the Kasten-Czeplak equation [31], as well as PVlib's models in their default configuration. In this case, for the Liu-Jordan model, we set the zenith angle to 45°, as our data normalizes for the zenith angle.

Similar to the linear model, the Liu-Jordan model is linear for any given zenith angle. As a result, both of PVlib's models are poor fits for the normalized data. The Kasten-Czeplak model is a better fit, but becomes increasingly imprecise as the cloud cover increases, with errors greater than 2× for cloud covers above 90%. Thus, we

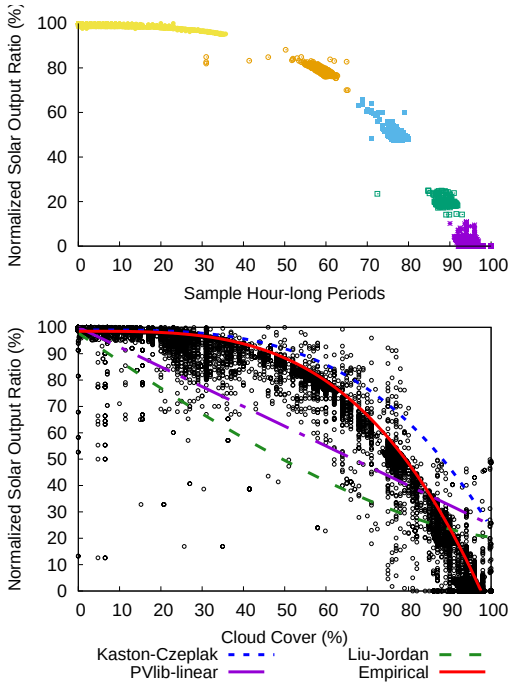


Figure 4: Isolated effect of cloud cover on normalized solar output ratio for 5 clusters (top). Normalized solar output versus cloud cover for a larger random sample of clusters, along with existing models and our new empirical one (bottom).

improve on Kasten-Czeplak by keeping the same model form, but finding parameters that provide a tight upper bound on the bulk of the datapoints. We assume the high outlier values are incorrect okta measurements, and use k-Means clustering to filter them out. As before, we fit the tightest upper bound that minimizes the RMSE with the data, which automatically filters out low values due to shading, soiling, or imprecise okta measurements. Our corrected empirical cloud cover equation is below and in Figure 4.

$$I_{incident}/I_{clearsky} = (0.985 - 0.984n^{3.4}) \quad (4)$$

4 INTEGRATING ML-BASED MODELING

The previous section i) demonstrates that weather's effect on a site's normalized solar output ratio is universal across solar sites, ii) shows that the only weather metrics that affect solar output are temperature and cloud cover, and iii) derives a new physical model to account for cloud cover's effect. The other physical models in §2 are also universal across solar sites, and account for module/array size, location, time, efficiency, tilt, and orientation. Since these models are universal, there is no need to separately learn them at each site using ML. However, ML is potentially useful for learning the effect of other unmodeled parameters that *are* unique to each site. One important unmodeled parameter is shade from surrounding buildings and trees. Thus, we enhance our physical model using ML to learn each site's unique shading effects from training data.

Since shading effects are a direct function of the position of the Sun in the sky—its azimuth and zenith angle—we use these angles as input features for our ML model. For the dependent output variable, we use the observed solar output divided by the solar output estimated by our physical model. Thus, with no effects from unmodeled parameters, the output variable should be 1, while with significant effects, the output variable should be <1. Note that the Sun's azimuth and zenith angles are a function of time at a given location, and many prior ML models include time as an input feature, enabling them to indirectly learn such shading effects. However, directly using solar angles as features makes all points comparable and eliminates the need for approximations, which enables a more accurate model.

5 IMPLEMENTATION

We implement our solar performance model using a mixture of python and C++. To build the model, we require a site's latitude and longitude, as well as some time-stamped solar generation data as input. We use the location to fetch historical hourly measurements of temperature and cloud cover at the time of solar generation from Weather Underground's API [12]. Once built, the model estimates solar output at any time t based on the weather at t. Our implementation requires a clear sky GHI model for calibration. We implement a clear sky GHI model from first principles using an open source C++ implementation of the PSA algorithm, which computes the Sun's azimuth and zenith angles to within 0.0083°.

Prior work describes how to compute the clear sky GHI given the solar angles, which are a function of location and time [7, 22]. We use the scikit-learn ML library in python to train our ML model based on solar angles, as well as the ML models we compare against in our evaluation [10]. Our approach in §4 is compatible with any ML modeling technique, such as SVMs or DNNs. Our current implementation uses SVM-RBF, similar to prior work on solar modeling [17, 35, 38]. We also use NumPy [4] and Pandas [5] libraries for weather and energy data processing. Our model implementation and data are available at the UMass Trace Repository.[1]

We compare our approach with multiple existing state-of-the-art approaches to black-box solar performance modeling.

Pure Physical. We implement a *pure physical* approach using the physical models in §2 including the Kasten-Czeplak cloud cover model. Note that we did not implement either PVlib cloud cover models, since they would result in significantly worse accuracy than Kasten-Czeplak based on our analysis in Figure 4.

Pure ML. We also implement the *pure ML* approach described at the beginning of §2 and in prior work [35], which uses all 10 weather metrics and each day's sunset and sunrise times as input features for training, and solar intensity as its output variable. Recall that solar intensity is raw solar output divided by a site's absolute solar capacity. The approach uses a SVM with a RBF kernel, and, importantly, leverages no physical models in its training.

Hybrid ML. We also implement a *hybrid ML* approach from prior work [21], which is similar to the pure ML approach above, but, instead of solar intensity, uses a normalized solar output ratio similar to that described at the beginning of §3 for its output variable. However, the model adjusts this ratio using the NOCT temperature and Kasten-Czeplak cloud cover models before training, and thus removes them as input features. As a result, the approach uses standard physical models to account for cloud cover and temperature effects, and then uses ML to learn the effect of other features.

[1] http://traces.cs.umass.edu

(a) Site #1 (b) Site #2 (c) Site #3 (d) Site #4 (e) Site #5 (f) Site #6

Figure 5: Satellite images (top) and Google Sunroof images (bottom) depicting 6 illustrative solar sites and their shading level. The site-specific shading level increases from left (a), with 0% shade, to right (f), with 60% shade.

Satellite Imagery. We implement a model that uses visible *satellite* imagery to estimate solar output instead of weather data. Geostationary satellites provide visible images of cloud cover every 15 minutes for nearly the entire world. The Heliosat algorithm [26] uses these images to estimate the effect of cloud cover on ground-level GHI. To estimate solar output, our satellite model uses the physical model from §2 to estimate a site's maximum solar output (including the temperature adjustment), and then multiplies this value by the GHI ratio from the satellite imagery. The GHI ratio is the ground-level GHI derived from the satellite imagery divided by the clear sky GHI derived from a clear sky model.

Empirical. Since we derive our physical model empirically, we label it as *empirical*. This model is equivalent to the pure physical model above, but instead uses our improved cloud cover equation.

Empirical ML. We evaluate our empirical model after enhancing it with ML, as described in §4, which we label as *empirical ML*.

We train all ML-based performance models on 5 years of solar data for each site using 20-fold cross-validation with a 70-30% split of training data to test data. For a fair comparison of accuracy, we calibrate the model of maximum solar output for the pure physical, satellite, and our empirical physical model using the same training data. We quantify model accuracy using the Mean Absolute Percentage Error (MAPE) between the ground truth solar power ($S(t)$) and the solar energy estimated by each model ($P_s(t)$) at each time t in our test set, which spans 13,140 hours for each solar site.

$$MAPE = \frac{100}{n} \sum_{t=0}^{n} \left| \frac{S(t) - P_s(t)}{S(t)} \right| \tag{5}$$

Lower MAPEs have higher accuracy with 0% being a perfect model. We only evaluate MAPE between sunrise and sunset. Note that MAPE is a conservative metric as small absolute errors can lead to large percentage errors at the beginning and end of each day, when solar output is low. To mitigate this effect, we also report MAPEs during the middle of each day (10am - 3pm).

6 EXPERIMENTAL EVALUATION

We evaluate and compare the accuracy of our solar performance model with the other models described in §5 on data from 100 solar sites at different locations with widely different physical and shading characteristics. We randomly select these 100 solar sites from 11,205 solar sites. Of course, the accuracy of the models varies across these sites. To better understand the attributes that affect model accuracy, we first examine in-depth the 6 homes pictured in Figure 5. This figure shows both a photograph from a satellite and from Google's Project Sunroof [2], which leverages LIDAR data to estimate a site's solar potential based on the roof tilt, orientation, and surrounding shading. Brighter colors indicate more solar potential. As the figure shows, some sites, such as (a), have little shade and are ideally positioned for solar generation, while other sites, such as (f), have non-ideal orientations and significant shading. We order the solar sites left to right from least shaded to most shaded.

Figure 6 then shows the accuracy in MAPE for all the performance models for each of these 6 solar sites. We order the performance models left to right from least accurate to most accurate. The top graph shows the MAPE over the entire day, while the middle graph shows the MAPE over just the middle of the day (10am - 3pm), where shading has the least effect, and the bottom graph shows the MAPE over just the beginning and end of each day (Sunrise - 10am and 3pm - Sunset), where shading has the most effect. All the graphs show the impact of the inaccurate Kasten-Czeplak cloud cover model (Equation 3) on the pure physical performance model, as it has by far the highest MAPE across all sites.

Surprisingly, the satellite-based performance model is the next least accurate across all sites and time periods. This likely demonstrates the limitations of using visible satellite imagery to estimate ground-level GHI. These limitations were also evident in the inaccuracy of Figure 1(bottom). Ultimately, visible satellite imagery can only detect the reflectivity of the tops of clouds, and cannot asses their thickness and the amount of solar radiation that ultimately reaches the ground. The pure ML and hybrid ML models are more accurate than the satellite-based model even though they use coarse measurements of cloud cover using oktas. However, while coarse, in contrast to satellite imagery, these measurements are taken at ground level near the solar site. For all sites and time periods, the hybrid ML model has a slightly higher accuracy than the pure ML model, likely because it uses feature engineering based on the physical models in §2 before training its ML model. Both the pure ML and hybrid ML models have decreased accuracy at the beginning and end of each day, where shading exhibits a greater effect, compared to the middle of the day. This indicates that both ML models have issues learning shading. The pure physical and

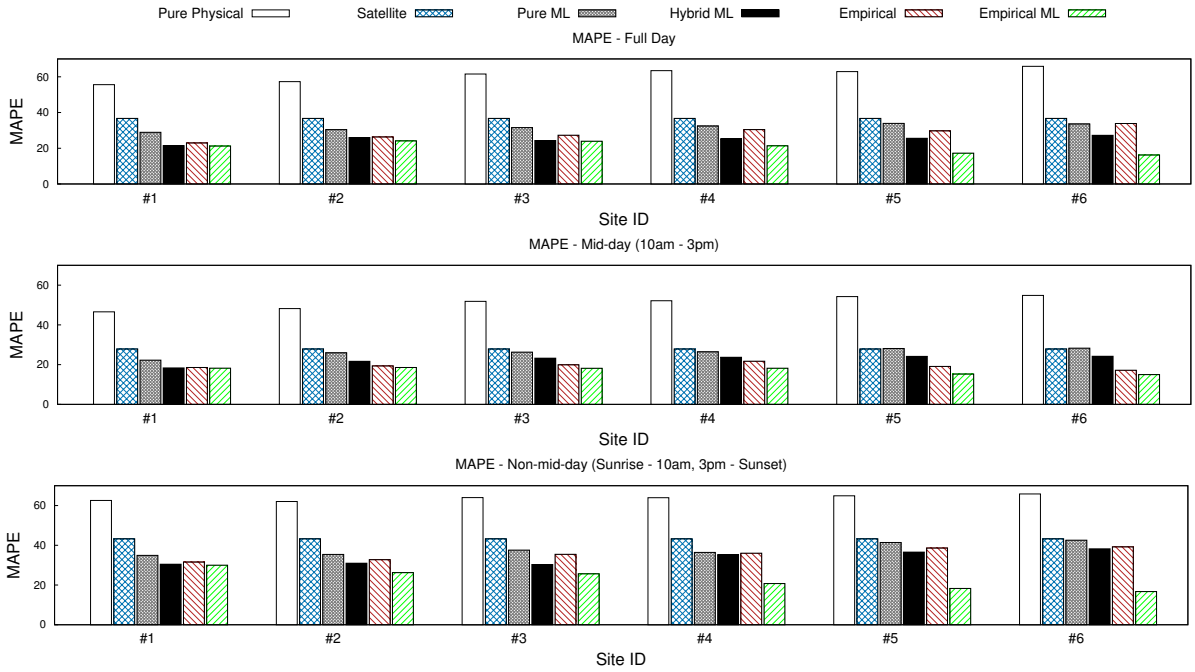


Figure 6: MAPE for 6 different black-box solar performance modeling techniques for the 6 solar sites in Figure 5.

satellite models have similar issues, as they also exhibit a lower accuracy at the beginning and end of each day.

Our empirical physical model, which is the same as the pure physical model but with an improved cloud cover equation, substantially increases the accuracy of the pure physical model. In all cases, our empirical model is also more accurate than the satellite model (even though it uses imprecise okta measurements for cloud cover) and the pure ML model (even though it does not incorporate shading effects). In contrast, the hybrid ML model is slightly more accurate than our empirical physical model over a full day, likely because the hybrid ML incorporates some physical models and indirectly accounts for shading effects during its training. However, over mid-day, where shading effects are minimal, our empirical physical model, which requires much less data for calibration, has equal or better accuracy across all sites. Further, our enhanced empirical ML physical model, which uses ML to account for unique site-specific shading effects, substantially improves the empirical physical model's accuracy. This improvement in accuracy increases as the site-specific shading effects increase from left to right, such that our empirical ML model reduces MAPE by ~2× for the most shaded site #6. Not surprisingly, the accuracy of our empirical ML model is similar to that of our empirical model over mid-day, where shading effects are minimal, but is significantly better at the beginning and end of each day, especially as site shading increases.

The only difference between the pure physical model and our empirical physical model is the cloud cover equation. To illustrate, Figure 7 shows a breakdown of accuracy based on the percentage of cloud cover for site #3. The pure physical model's accuracy becomes steadily worse as the cloud cover increases, due to increasing inaccuracy in the Kasten-Czeplak model, while our empirical model accuracy is consistent. The figure also shows how the empirical ML model improves upon the empirical model. Under minimal cloud

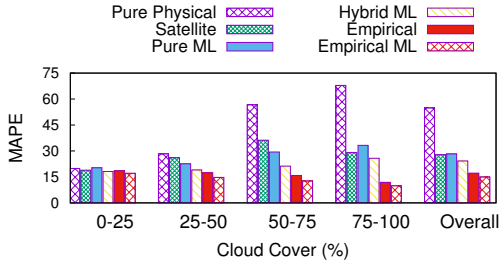


Figure 7: MAPE for mid-day solar generation during different weather conditions for solar site #3 from Figure 8.

cover, all the models exhibit similar accuracy (~15%). We suspect that some of the inaccuracy derives from okta and satellite measurement error and not model error, as indicated in Figure 1(bottom) and Figure 4(bottom), respectively. Imprecise measurements ultimately bound the accuracy of solar performance modeling.

Finally, Figure 8 shows results for all of the models across 100 rooftop solar deployments at different locations with various climates and shading levels. From top to bottom, the graphs show the pure physical, satellite, pure ML, hybrid ML, empirical, and empirical ML models described earlier. Note that the y-axis range is [0-80] with a dotted line at 20 as visual reference point for comparison. Since space constraints prevent us from including pictures of all 100 sites, we manually divide the deployments into different general shading levels and group them together. Within each shading level, we order sites based on their average cloud cover, such that less cloudier sites within a group have a lower ID (and are on the left). In general, our results from 100 sites echo our results from 6 sites. Interestingly, in all models except for empirical ML, the average accuracy slightly decreases as the shade increases. In contrast, the accuracy of the empirical ML model is more consistent across all shade levels, and even slightly better for sites with higher levels

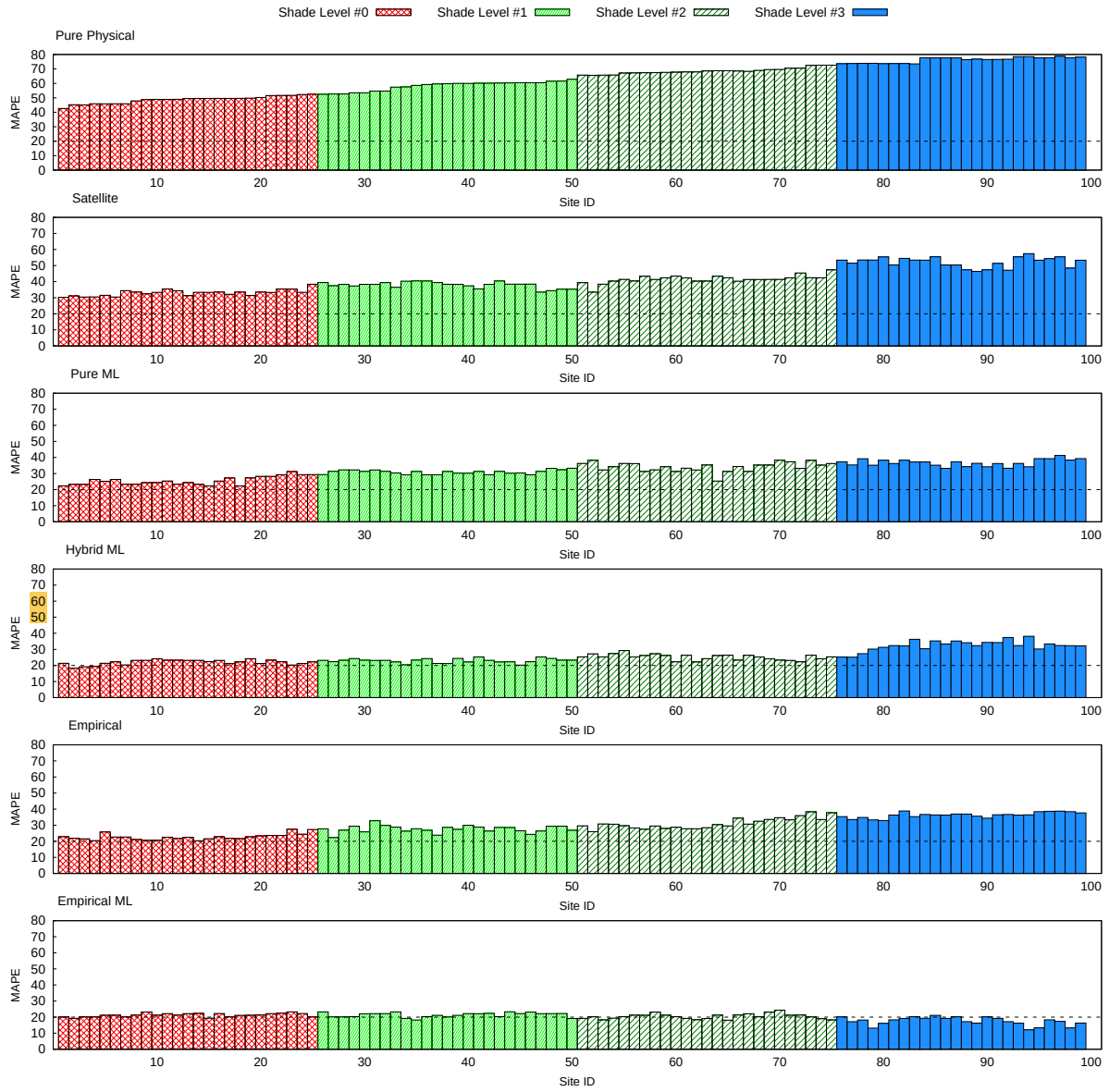


Figure 8: MAPE of 6 solar performance models across 100 solar sites over 1.5 years. Sites are grouped by their shade level.

of shade. Overall, across all 100 sites, the average MAPE of our empirical ML model was 20.7%, or 18% and 49% better than the average MAPE of the state-of-the-art hybrid ML model (25.3%) and the satellite model (40.6%), respectively.

7 APPLICATIONS AND RELATED WORK

Our approach combines the best aspects of white-box modeling with black-box modeling. We compare with numerous other black-box ML approaches in §6. We plan to compare our approach with PVlib's white-box modeling as part of future work [15]. However, comparing with PVlib is time consuming, and likely not feasible at large scales, since it requires deep infrastructure-level access to construct a white-box model of each solar site. Solar performance modeling is also a foundation for many solar analytics.

Solar Monitoring. Our solar performance model enables indirect solar monitoring if the sensors directly monitoring power generation fail, by simply replacing the sensor data with the model output. We can also easily adapt our performance model to enable us to infer a site's solar output from other nearby sites if weather data is not available, as in prior work [25]. In this case, we can simply multiply the normalized solar output ratio from a nearby site by our model's estimate of a site's maximum power generation.

Solar Forecasting. Our solar performance model also enables solar forecasting by providing as input a future time, and forecast for temperature and cloud cover at that time. Recent research focuses on black-box ML approaches to solar forecasting similar to ours [20, 29, 38]. These techniques are generally off-the-shelf, do not incorporate physical models, and require months-to-years of

training data to learn accurate models. As future work, we plan to apply our performance model to solar forecasting, compare its accuracy with existing black-box ML approaches, and quantify the weather forecast, modeling, and measurement error.

Solar Fault Detection. Our solar performance model enables anomaly and fault detection if its accuracy deviates from its expected accuracy. In addition, a recent approach uses a sophisticated black-box ML algorithm for detecting anomalies by analyzing and comparing with the solar output of nearby sites [28]. Our solar performance model enables a similar function by simply comparing the normalized solar output ratios of nearby sites after adjusting for shading effects. Since these ratios should be equal across sites experiencing the same weather, any divergence signals an anomaly.

Solar Disaggregation. Prior work on solar disaggregation, which separates solar generation from energy consumption in combined net meter data, implicitly leverages the insight that the effect of weather on solar generation is universal [22, 35]. That is, these approaches learn black-box ML models that infer solar output on labeled data from one set of solar sites for training, and then assume they can use these models to accurately infer solar output at other sites, where raw solar data is not available. To disaggregate, these approaches simply subtract the inferred solar output from the net meter data to infer energy consumption. This paper empirically verifies this assumed universal weather effect across data from tens of thousands of solar sites, demonstrates that temperature and cloud cover are the only weather metrics that affect solar output, and derives an improved equation between cloud cover and GHI.

8 CONCLUSION

This paper develops a physical black-box solar performance model, which requires minimal data for calibration, based on fundamental properties of solar generation. In particular, we leverage a large-scale data analysis across tens of thousands of years of solar and weather data in aggregate to i) demonstrate that weather's effect on a site's normalized solar output ratio is universal across all solar sites, ii) show that the only weather metrics that affect solar output are temperature and cloud cover, and iii) derive a new physical model to quantify cloud cover's effect on solar generation. We show that our physical black-box model yields similar accuracy as state-of-the-art black-box ML approaches. We then enhance our physical model with a ML model that learns each site's unique shading effects. We show that our enhanced model has significantly better accuracy than state-of-the-art ML models, as well as a model based on GHI estimates from visible satellite imagery.

Acknowledgements. This research is supported by NSF grants IIP-1534080, CNS-1645952, CNS-1405826, CNS-1253063, CNS-1505422, Army Research Lab Award W911NF-17- 2-0196, and the Massachusetts Department of Energy Resources.

REFERENCES

[1] 2018. Bird Simple Spectral Model. http://rredc.nrel.gov/solar/models/spectral/. (June 2018).
[2] 2018. Google Project Sunroof. https://www.google.com/get/sunroof. (June 2018).
[3] 2018. 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 2018).
[4] 2018. NumPy. http://www.numpy.org/. (June 2018).
[5] 2018. Pandas: Python Data Analysis Library. https://pandas.pydata.org/. (2018).
[6] 2018. Pecan Street. http://www.pecanstreet.org/. (June 2018).
[7] 2018. PVEducation.org. https://www.pveducation.org/pvcdrom/properties-of-sunlight/air-mass. (June 2018).
[8] 2018. PVoutput.org. http://pvoutput.org/. (June 2018).
[9] 2018. PySolar. http://pysolar.org/. (June 2018).
[10] 2018. scikit-learn: Machine Learning in Python. http://scikit-learn.org/stable/index.html. (June 2018).
[11] 2018. Solar Radiation Cloud Cover Adjustment Calculator. http://www.shodor.org/os411/courses/_master/tools/calculators/solarrad/. (June 2018).
[12] 2018. Weather Underground API. https://www.wunderground.com/weather/api/. (June 2018).
[13] 2018. Weather.gov Terms. http://www.weather.gov/bgm/forecast_terms. (2018).
[14] Y. Akiyama, Y. Kasai, M. Iwata, E. Takahashi, F. Sato, and M. Murakawa. 2015. Anomaly Detection of Solar Power Generation Systems Based on the Normalization of the Amount of Generated Electricity. In *AINA*.
[15] 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*.
[16] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez de Pison, and F. Antonanzas-Torres. 2016. Review of Photovoltaic Power Forecasting. *Solar Energy* 136 (October 2016).
[17] M. Benghanem and A. Mellit. 2010. Radial Basis Function Network-based Prediction of Global Solar Radiation Data: Application for Sizing of a Stand-alone Photovoltaic System at Al-Madinah. *Energy* 35, 9 (2010).
[18] R. Bharti, J. Kuitche, and M. TamizhMani. 2009. Nominal Operating Cell Temperature (NOCT): Effects of Module Size, Loading and Solar Spectrum. In *PVSC*.
[19] M. Blanco-Muriel, D.C. Alarcon-Padilla, T. Lopea-Moratalla, and M. Lara-Coira. 2001. Computing the Solar Vector. *Solar Energy* 70, 5 (August 2001).
[20] P. Chakraborty, M. Marwah, M. Arlitt, and N. Ramakrishnan. 2012. Fine-grained Photovoltaic Output Prediction using a Bayesian Ensemble. In *AAAI*.
[21] D. Chen and D. Irwin. 2017. Black-box Solar Performance Modeling: Comparing Physical, Machine Learning, and Hybrid Approaches. In *Greenmetrics*.
[22] D. Chen and D. Irwin. 2017. SunDance: Black-box Behind-the-Meter Solar Disaggregation. In *e-Energy*.
[23] D. Chen and D. Irwin. 2017. Weatherman: Exposing Weather-based Privacy Threats in Big Energy Data. In *BigData*.
[24] D. Chen, S. Iyengar, D. Irwin, and P. Shenoy. 2016. SunSpot: Exposing the Location of Anonymous Solar-powered Homes. In *BuildSys*.
[25] N.A. Engerer and F.P. Mills. 2014. KPV: A Clear-sky Index for Photovoltaics. *Solar Energy* 105 (July 2014), 670–693.
[26] A. Hammer, D. Heinemann, C. Hoyer, R. Kuhlemann, E. Lorenz, R. Muller, and H. Beyer. 2003. Solar Energy Assessment using Remote Sensing Technologies. *Remote Sensing of Environment* 86 (2003), 423–432.
[27] B. Hu. 2012. *Solar Panel Anomaly Detection and Classification*. Technical Report. University of Waterloo.
[28] S. Iyengar, S. Lee, D. Sheldon, and P. Shenoy. 2018. SolarClique: Detecting Anomalies in Residential Solar Arrays. In *ACM COMPASS*.
[29] 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*.
[30] E. Kara, M. Tabone, C. Roberts, S. Kiliccote, and E. Stewart. 2016. Poster Abstract: Estimating Behind-the-meter Solar Generation with Existing Measurement Infrastructure. In *BuildSys*.
[31] F. Kasten and G. Czeplak. 1980. Solar and Terrestrial Radiation Dependent on the Amount and Type of Cloud. *Solar Energy* 24, 2 (October 1980).
[32] B. Liu and R. Jordan. 1960. The Interrelationship and Characteristic Distribution of Direct, Diffuse and Total Solar Radiation. *Solar Energy* 4, 3 (July 1960).
[33] Joshua S. Stein Matthew J. Reno, Clifford W. Hansen. 2012. *Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis*. Technical Report. Sandia National Laboratories.
[34] S. Mekhilef, R. Saidur, and M. Kamalisarvestani. 2012. Effect of Dust, Humidity and Air Velocity on Efficiency of Photovoltaic Cells. *Renewable and Sustainable Energy Reviews* 16, 5 (2012).
[35] R. Mohan, T. Cheng, A. Gupta, V. Garud, and Y. He. 2014. Solar Energy Disaggregation using Whole-House Consumption Signals. In *NILM Workshop*.
[36] D. Myers. 2006. Cloudy Sky Version of Bird's Broadband Hourly Clear Sky Model. In *Annual conference of the American Solar Energy Society (SOLAR)*.
[37] C. Schwingshackl, M. Petitta, J.E. Wagner, G. Belluardo, D. Moser, M. Castelli, M. Zebisch, and A. Tetzlaff. 2013. Wind Effect on PV Module Temperature: Analysis of Different Techniques for an Accurate Estimation. *Energy Procedia* 40 (2013).
[38] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy. 2011. Predicting Solar Generation from Weather Forecasts Using Machine Learning. In *SmartGridComm*.
[39] J.S. Stein. 2012. The Photovoltaic Performance Modeling Collaborative (PVPMC). In *IEEE Photovoltaic Specialist Conference*.
[40] A. Vasel and F. Iakovidis. 2017. The Effect of Wind Direction on the Performance of Solar PV Plants. *Energy Conversion and Management* 153 (December 2017).
[41] C. Voyant, G. Notton, S. Kalogirou, M. Nivet, C. Paoli, F. Motte, and A. Fouilloy. 2017. Machine Learning Methods for Solar Radiation Forecasting: A Review. *Renewable Energy* 105 (May 2017).