SolarClique: Detecting Anomalies in Residential Solar Arrays

Srinivasan Iyengar, Stephen Lee, Daniel Sheldon, Prashant Shenoy
University of Massachusetts Amherst

ABSTRACT

The proliferation of solar deployments has significantly increased over the years. Analyzing these deployments can lead to the timely detection of anomalies in power generation, which can maximize the benefits from solar energy. In this paper, we propose SolarClique, a data-driven approach that can flag anomalies in power generation with high accuracy. Unlike prior approaches, our work neither depends on expensive instrumentation nor does it require external inputs such as weather data. Rather our approach exploits correlations in solar power generation from geographically nearby sites to predict the expected output of a site and flag anomalies. We evaluate our approach on 88 solar installations located in Austin, Texas. We show that our algorithm can even work with data from few geographically nearby sites (>5 sites) to produce results with high accuracy. Thus, our approach can scale to sparsely populated regions, where there are few solar installations. Further, among the 88 installations, our approach reported 76 sites with anomalies in power generation. Moreover, our approach is robust enough to distinguish between reduction in power output due to anomalies and other factors such as cloudy conditions.

CCS CONCEPTS

•**Computing methodologies** **Anomaly detection;** *Supervised learning by regression;* •**Social and professional topics** *Sustainability;*

KEYWORDS

anomaly detection; renewables; solar energy; computational sustainability

ACM Reference format:
Srinivasan Iyengar, Stephen Lee, Daniel Sheldon, Prashant Shenoy. 2018. SolarClique: Detecting Anomalies in Residential Solar Arrays. In *Proceedings of ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS), Menlo Park and San Jose, CA, USA, June 20–22, 2018 (COMPASS '18),* 10 pages.
DOI: 10.1145/3209811.3209860

1 INTRODUCTION

Technological advances and economies of scale have significantly reduced the costs and made solar energy a viable renewable alternative. From 2010 to 2017, the average system costs of solar have dropped from $7.24 per watt to $2.8 per watt, a reduction of approximately 61% [13]. At the same time, the average energy cost of producing solar is 12.2¢ per kilo-watt and is approaching the retail electricity price of 12¢ per kilo-watt [1]. The declining costs have spurred the adoption of solar among both utilities and residential owners.

Recent studies have shown that the total capacity of small-scale residential solar installations in the US reached 7.2 GW in 2016 [3]. Unlike large solar farms, residential installations are not monitored by professional operators on an ongoing basis. Consequently, anomalies or faults that reduce the power output of residential solar arrays may go undetected for long periods of time, significantly reducing their economic benefits. Further, large solar farms have extensive sensor instrumentation to monitor the array continuously, which enables faults or anomalous output to be determined. In contrast, residential installations have little or no sensor instrumentation beyond displaying the total power of the array, making sensor-based monitoring and anomaly detection infeasible in such contexts. Adding such instrumentation increases the installation costs and is not economically feasible in most cases.

In this paper, we seek to develop a data-driven approach for detecting anomalous output in small-scale residential installations *without* requiring any sensor information for fault detection. Our key insight is that the solar output from other nearby installations are correlated, and thus these correlations can be used to identify anomalous deviations in a specific installation. We note that the solar output from multiple sites within a city or region is available. For instance, Enphase, an energy company, provides access to power generation information of more than 700,000 sites across different locations[1]. Similarly, other sites[2] share their energy generation data from tens of thousands of solar panels through web APIs. Thus, the availability of such datasets makes our approach feasible. However, the primary challenge in designing such an application is to handle intrinsic variability of solar and site-specific idiosyncrasies.

Several factors affect the output of a solar panel — such as weather conditions, dust, snow cover, and shade from nearby trees or structures, temperature, etc. We refer to such factors as transient factors since they temporarily reduce the output of the solar array. For instance, a passing cloud may briefly decrease the power output of the panel but doesn't reduce the solar output permanently. Similarly, shade from nearby buildings or trees can be considered transient factors as they reduce the output temporarily and may occur only at certain periods of the day.

Interestingly, some transient factors, such as overcast conditions, impact the output of all arrays in a geographical neighborhood, while other factors such as shade from a nearby tree impact the output of only a portion of the array. In addition, factors such as malfunctioning solar modules or electrical faults also reduce the output of a solar array, and we refer to them as *anomalies* — since human intervention is needed to correct the problem. Prior

COMPASS '18, Menlo Park and San Jose, CA, USA

978-1-4503-5816-3/18/06…$15.00
DOI: 10.1145/3209811.3209860

[1] Enphase: https://enphase.com/
[2] PVSense: http://pvsense.com/

studies have shown that such factors can significantly reduce the power output by as much as 40% [6, 11, 14]. In our work, we need to distinguish between the output fluctuations from transient and anomalous factors. Further, site specific idiosyncrasies (such as shade, tilt/orientation of panels) need to be considered when exploiting the correlation between solar arrays in a region.

Naive approaches such as labeling a solar installation as anomalous whenever its power output remains "low" for an extended period do not work well. Since drops in power output may be caused due to cloudy conditions, depending on the weather, the solar output may remain low for days. Labeling such instances as anomalies may result in many false positives. Since the challenge lies in differentiating drops in power output due to transient factors (i.e., factors that impact power output temporarily) and those that are anomalies (i.e., factors that may require human intervention), we propose SolarClique, a new approach for detecting solar anomalies using geographically nearby sites. In designing, implementing and evaluating our approach, we make the following contributions:

- We demonstrate how geographically nearby sites can be used to detect anomalies in a residential solar installation. In our algorithm, we present techniques to account for and remove transient seasonal factors such as shade from nearby trees and structures.
- Our approach doesn't require any sensor instrumentation for fault/anomaly detection. Rather, it only requires the production output of the array and those of nearby arrays for performing anomaly detection. Since power output of geographically nearby sites are readily available, our approach can be easily applied to millions of residential installations that are unmonitored today with little added expense.
- We implement and evaluate the performance of our algorithm. Our results show that the power output of a candidate site can be predicted using geographically nearby sites. Moreover, we can achieve high accuracy even when few geographically nearby sites (>5 sites) are available. This indicates that our approach can be used in sparsely populated locations, where there are few solar installations.
- We ran our anomaly detection algorithm on solar installations located in Austin, Texas. SolarClique reported power generation anomalies in 76 sites, many of which see a solar output reduction for weeks or months. Moreover, our approach can identify different types of anomaly — (i) no production, (ii) underproduction, and (ii) gradual degradation of solar output over time, thereby exhibiting the real-world utility of our approach.

2 BACKGROUND

Our work focuses on detecting anomalous solar power generation in a residential solar installation using information from geographically nearby sites. Unlike power generation from traditional mechanical generators (e.g., diesel generators), where power output is constant and controllable, the instantaneous power output from a PV system is inherently intermittent and uncontrollable. The solar power output may see sudden changes, with energy generation at peak capacity at one moment to reduced (or zero) output in the next period (see Figure 1). The change in the power output can be attributed to a number of factors and our goal is to determine whether the drop in power can be attributed to anomalous behavior in the solar installation.

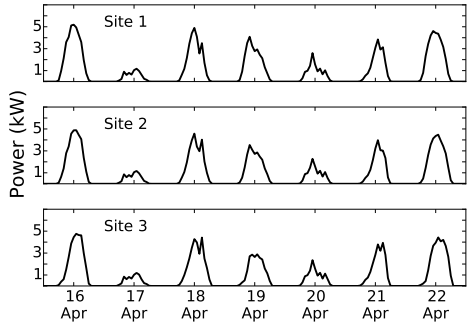


Figure 1: Power generation from three geographically nearby solar sites. As shown, the power output is intermittent and correlated for solar arrays within a geographical neighborhood.

2.1 Factors affecting solar output

A primary factor that influences the power generation of a solar panel is the solar irradiance, i.e., the amount of sunlight that is incident on the panel. The amount of sunlight a solar panel receives is dependent on many factors such as time of the day and year, dust, temperature, cloud cover, shade from nearby buildings or structures, tilt and orientation of the panel, etc. These factors determine the amount of power that is generated based on how much light is incident on the solar modules.

However, a number of other factors, related to hardware, can also reduce the power output of a solar panel. For instance, the power output may reduce due to defective solar modules, charge controllers, inverters, strings in PV, wired connections and so on. Clearly, there are many factors that can cause problems in power generation. Thus, factors affecting output can be broadly classified into two categories: (i) *transient* — factors that have a temporary effect on the power output (such as cloud cover); and (ii) *anomalies* — factors that have a more prolonged impact (e.g., solar module defect) on the power output.

The transient factors can further be classified into *common* and *local* factors. The common factors, such as weather, affect the power output of all the solar panels in a given region. Moreover, its effect is temporary as the output changes with a change in weather conditions. For instance, overcast weather conditions temporarily reduce the output of all panels in a given region. The local factors, such as shade from nearby foliage or buildings, are usually site-specific and do no affect the power output of other sites. These local factors may be recurring and reduce the power output at fixed periods in a day. In contrast, anomalous factors, such as bird droppings or system malfunctions, reduces power output for prolonged periods and usually require corrective action to restore normal operation of the site. Note that both transient and anomalous factors may reduce the power output of a solar array. Thus, a key challenge in designing

a solar anomaly detection algorithm is to differentiate the reduction in power output due to transient factors and anomalies.

2.2 Anomaly detection in solar installations

Prior approaches have focused on using exogenous factors to predict the future power generation [8, 20, 28]. A simple approach is to use such prediction models and report anomaly in solar panels if the power generated is below the predicted value for an extended period. However, it is known that external factors such as cloud cover are inadequate to accurately predict power output from solar installations [17]. Thus, prediction models may over- or under-predict power generation, and such an approach may not be sufficient for detecting anomalies.

Prediction models can be improved using additional sensors but can be prohibitively expensive for residential setups [21]. For instance, drone-mounted cameras can detect occlusions in a panel but are expensive and require elaborate setup. Other studies use an ideal model of the solar arrays to detect faults [5, 12]. These studies rely on various site-specific parameters and assume standard test condition (STC) values of panels are known. However, site-specific parameters are often not available. Thus, most large solar farms usually depend on professional operators to continuously monitor and maintain their setup to detect faults early[3,4]. Clearly, such elaborate setups may not be economically feasible in a residential solar installation. Below, we present our work that focuses on a data-driven and cost-effective approach for detecting anomalies in a solar installation.

3 ALGORITHM DESIGN

We first introduce the intuition behind our approach to detect anomalous power generation in a solar installation. Our primary insight is that other geographically nearby sites can predict the solar output potential, which can then reveal issues in a given site. Since factors such as the amount of solar irradiance (e.g., due to cloudy conditions) are similar within a region, the power output of solar arrays in a geographical neighborhood is usually correlated. This can be seen in the power output from three different solar installation sites in the same geographical neighborhood (see Figure 1). As seen, the solar arrays tend to follow a similar power generation pattern. So we can use the output of a group of sites to predict the output of a specific site and flag anomalies if the prediction significantly deviates.

We hypothesize that predicting the output using geographically nearby sites can "remove" the effects of confounding factors (i.e., common factors). By accounting for confounding factors, the remaining influence on power generation can be attributed to local factors in the solar installation. The local factors may include both transient local factors and anomalies. Thus, any irregularity in power generation, after accounting for confounding and transient local factors, must be due to anomalies in the installation. For example, cloudy or overcast conditions in a given location have a similar impact on all solar panels and will reduce the power output of all sites. However, a malfunctioning solar module in a site (a local event) will observe a higher drop in power output than others. If the drop in power due to cloudy conditions (a confounding factor) along with transient local factors is accounted for, any further drop in power can be attributed to anomalies. Our approach follows this intuition to detect anomalies in a solar installation.

[3]ESA Renewables: http://esarenewables.com/

[4]Affinity Energy: https://www.affinityenergy.com/

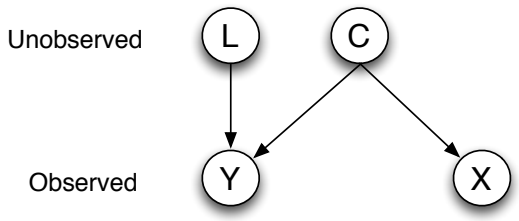


Figure 2: Graphical model representation of our setup.

The rest of the section is organized as follows. We present a graphical model representation for our setup that models the confounding variables. Next, we discuss how our algorithm removes the confounding factors and detects anomalies in solar generation.

3.1 Graphical model representation

Our work is inspired by a study in astronomy, wherein *Half-Sibling Regression* technique was used to remove the effects of confounding variables (i.e., noises from measuring instruments) from observations to find exoplanets [27]. We follow a similar approach to model and detect anomalies in a solar installation.

Let C, L, X and Y be the random variables (RVs) in our problem. Here, Y refers to the power generated by a candidate solar installation site. X represents the power produced by each of the geographically nearby solar installations (represented in a vector format). While C represents the confounding variables that affect both X and Y, the variable L represents site-specific local factors affecting a candidate site. These local factors include both transient factors and anomalies that affect a candidate site. In our setup, both X and Y are observed variables (as power generation of a site can be easily measured), whereas C and L are latent unobserved variables. Figure 2 depicts a directed graphical model (DAG) that illustrates the relationship between these observed and unobserved random variables.

We are interested in the random variable L which represents anomalies at a given site. As seen in the figure, since both L and C affect the observed variable Y, without the knowledge of C it is difficult to calculate the influence of L on Y. Clearly, X is independent of L as variable L impacts only Y. However, we note that C impacts X and when conditioned on Y, Y becomes a *collider*, and the variables X and L become dependent [23]. This implies that X contains information about L and we can recover L from X.

To reconstruct the quantity L, we impose certain assumptions on the type of relationship between Y and C. Specifically, we assume that Y can be represented as an additive model denoted as follows:

$$Y = L + f(C) \tag{1}$$

where f is a nonlinear function and its input C is unobserved. Since L and X are independent, variable X cannot account for the influence of L on Y. However, X can be used to approximate $f(C)$, as C also affects X. If X exactly approximates $f(C)$, then $f(C) = E[f(C)|X]$, and we can show that L can be recovered completely using (1). Even if X does not exactly approximate $f(C)$, in our case, X is sufficiently

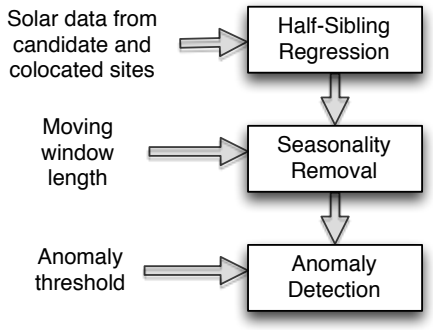


Figure 3: An overview of the key steps in the SolarClique algorithm.

large to provide a good approximation of $E[f(C)|X]$ up to an offset. A more detailed description of the approach is given in [27]. Thus, using X to predict Y (i.e., $E[Y|X]$), $f(C)$ can be approximated and removed from (1) to estimate $\hat{L}$ as follows:

$$\hat{L} := Y - E[Y|X] \tag{2}$$

where $\hat{L}$ is an estimate of the local factors that may include both transient local factors and anomalies.

3.2 SolarClique Algorithm

We now present our anomaly detection algorithm called *SolarClique*. Figure 3 depicts an overview of the different steps involved in the *SolarClique* algorithm. First, we use the *Half-Sibling Regression* approach to build a regression model that predicts the solar generation of a candidate site using power output from geographically nearby sites. Next, we remove any seasonal component from the above regression model using time series decomposition. Finally, we detect anomalies by analyzing the deviation in the power output. Below, we describe these three steps in detail.

3.2.1 Step 1: Remove confounding effects. The first step is to build a regression model that predicts the power generation output Y of a candidate site using X, a vector of power generation values from geographically nearby solar installations. As mentioned earlier, the regression model estimates $E[Y|X]$ component in the additive model shown in (2). Since Y is observed, subtracting the $E[Y|X]$ component determines the L component.

Standard regression techniques can be used to build this regression model. The regression technique learns an estimator that best fits the training data. Instead of constructing a single regression model, we use *bootstrapping* — a technique that uses subsamples with replacement of the training data — which gives multiple regression models and the properties of the estimator (such as standard deviation). We use an ensemble method, wherein the mean of the regression models is taken to estimate the $E[Y|X]$ in the testing data. Finally, we remove the confounding component $E[Y|X]$ from Y to obtain an estimate of $\hat{L}_t$ $\forall t \in T$ in the testing data. The final output of this step is an estimate $\hat{L}_t$ and the standard deviation (σ_t) of the estimators.

3.2.2 Step 2: Remove seasonal component. As discussed earlier, the solar output of a site is affected by both common (i.e., confounding) and local factors. Using the *Half-Sibling Regression* approach, we can account for the *transient* confounding factors such as weather changes. However, we also need to account for *transient* local factors, such as shade from nearby trees, which may temporarily reduce the power output at a specific time of the day. Since variable $\hat{L}_t$ include both transient local factors and anomalies, we need to remove the local factors to determine the anomaly $\hat{A}_t$.

We note that the time period of such occlusions (those from nearby trees or structures) may not vary much on a daily basis. This is because the maximum elevation of the sun in the sky varies by less than 2° over a period of a week[5] on average. Using time series decomposition techniques over short time intervals (e.g. one week), such seasonal components (i.e the pattern occurring every fixed period) can be removed. Thus, we perform a time series decomposition to account for transient local factors as follows. We compute the seasonal component and remove it from $\hat{L}_t$ only when $\hat{L}_t$ is outside the confidence interval 4σ and on removal of the seasonal component, $\hat{L}_t$ doesn't go outside the confidence interval. After removal of the seasonal component, if any, we obtain $\hat{A}_t$ from $\hat{L}_t$ as our final output.

3.2.3 Step 3: Detect Anomalies. We use the output $\hat{A}_t$ (from Step 2) and the standard deviation σ_t (from Step 1) to detect anomalies in a candidate site. Specifically, we flag the day as anomalous when three conditions hold. First, the deviation of $\hat{A}_t$ should be significant, i.e., greater than four times the standard deviation. Second, the anomaly should occur for at least k contiguous period. Finally, when the period t is during the daytime period (not including the twilight). Thus, an anomaly can be defined as follows:

$$anomaly = (\hat{A}_t < -4\sigma_t) \wedge ... \wedge (\hat{A}_{t+k} < -4\sigma_t) \qquad \forall t \in T \tag{3}$$

where T denotes the time during the daytime period.

Based on our assumption that $\hat{A}_t$ is Gaussian, it follows that the odds of an anomaly are very high when the deviation is more than 4σ. These anomalous values belong to the end-tail of the normal distribution. The second condition (i.e., contiguous anomaly period) ensures that the drop in power output is for an extended period. In practice, depending on the data granularity, the contiguous period can range from minutes to hours. Clearly, we would like to detect anomalies during the period when sunlight is abundant. During the night or twilight, the solar irradiation is very low to provide any meaningful power generation. Thus, we choose the daytime period in our algorithm for anomaly detection.

4 IMPLEMENTATION

We implemented our SolarClique algorithm in python using the SciPy stack [4]. The SciPy stack consists of efficient data processing and numeric optimization libraries. Further, we use the regression techniques in the scikit-learn library to learn our models [24]. The scikit-learn library comprises various regression tools, which takes a vector of input features and learn the parameters that best describe

[5]The sun directly faces the Tropic of Cancer (+23.5°) on the summer solstice. Whereas, it faces the Tropic of Capricorn (-23.5°) on the winter solstice. Thus, over half the year (26 weeks) the maximum elevation of the sun changes by ≈47°, i.e., ¡2° per week.

Number of solar installations	88
Solar installation size (kW)	0.5 to 9.3
Residential size (sq. ft.)	1142 to 3959
Granularity	1 hour
Year	2014, 2015

Table 1: Key characteristics of the dataset.

the relationship between the input and the dependent variable. Additionally, we use Seasonal and Trend decomposition using Loess (STL) technique to remove the seasonality component [9]. The STL technique performs a time series decomposition on the input and deconstructs it into trend, seasonal, and noise components.

5 EVALUATION

5.1 Dataset

For evaluating the efficacy of SolarClique, we use a public dataset available through the Dataport Research Program [2]. The dataset contains solar power generation from over hundred residential solar installations located in the city of Austin, Texas. The power generation from these installations are available at an hourly granularity. Table 1 shows the key characteristics of the dataset. For our case study, we selected those homes that have contiguous solar generation data, i.e., no missing values, for an overlapping period of at least two years. Based on this criteria, we had 88 homes for our evaluation in the year 2014 and 2015.

5.2 Evaluation Methodology

We partitioned our dataset into training and testing period. We used the first three months of data to train the model, and the remaining dataset for testing (21 months). Further, for bootstrapping, we sample our training dataset by randomly selecting 80% of the training samples with replacement. These samples are then used to build the estimator, and we repeated this step 100 times to learn the properties of the estimator. To build our model, we used five popular regression techniques namely Random Forest (RF), k-Nearest Neighbor (kNN), Decision Trees (DT), Support Vector Regression (SVR), and Linear Regression (LR). Finally, we selected the contiguous period as $k = 2$ (see Step 3 of our algorithm) since our data granularity is hourly. Unless stated otherwise, we use all homes in our dataset for our evaluation.

5.3 Metrics

Since the installation capacity can be different across solar panels, it may not be meaningful to use a metric such as Root Mean Squared Error (RMSE). This is because the magnitude of the error may be different across predictions. Thus, we use Mean Absolute Percentage Error (MAPE) to measure the regression model's accuracy in predicting a candidate's power generation. MAPE is defined as:

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{y_t - p_t}{\bar{y}_t} \right| \quad (4)$$

where y_t and p_t are the actual and predicted value at time t respectively. $\bar{y}_t$ represents the average of all the values and n is the number of samples in the test dataset.

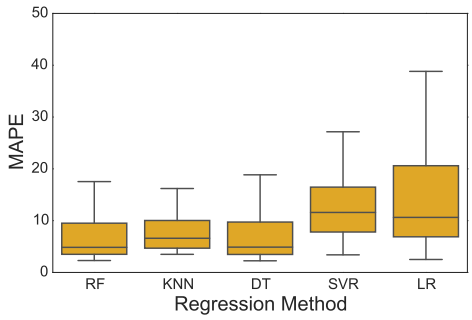


Figure 4: Performance of different regression techniques used to predict the power generation of a site.

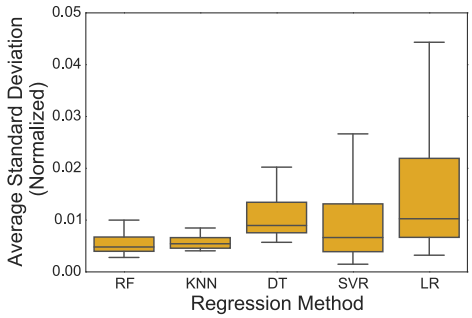


Figure 5: Mean standard deviation of predictions for different regression techniques

5.4 Results

Below, we summarize the results of using SolarClique on the Dataport dataset.

5.4.1 Prediction performance using geographically nearby sites. We compare the accuracy of the five regression techniques used to predict the power generated at a candidate site (Y) using the data from nearby sites (X). Figure 4 shows the spread of the MAPE values for the regression techniques used for all the 88 sites. Random Forest and Decision Trees show the best performance closely followed by k-NN with average MAPE values of approximately 7.81%, 7.87%, and 8.94% respectively. Linear Regression, on the other hand, shows poor accuracy with an average MAPE of 19%.

As discussed earlier, our approach uses bootstrapping to generate the standard deviation values for each prediction. Note that a small standard deviation means tighter confidence interval and indicates that the regression technique has a consistent prediction across runs. Figure 5 shows the mean value of standard deviation over all the testing samples normalized by the size of the solar installation. We observe that RF and k-NN have tight confidence intervals, while LR has considerably wider bounds. In particular, we observe that the average standard deviation of RF and k-NN is 0.0032 and 0.0059 using all the sites, respectively. In comparison, the average standard deviation of LR is 0.0078. Since RF performs better than other regression techniques, we use RF for the rest of our evaluation.

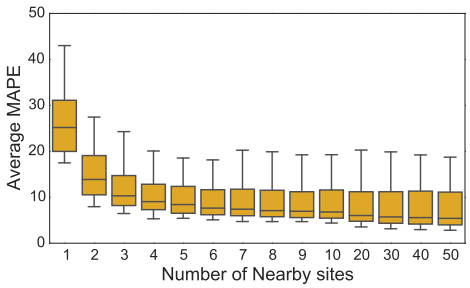


Figure 6: Average MAPE diminishes with increase in the number of geographically nearby sites.

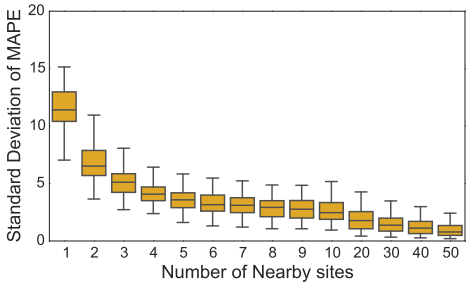


Figure 7: Standard deviation of MAPE diminishes with increase in the number of geographically nearby sites.

5.4.2 Impact due to the number of geographically nearby sites. We now focus on understanding the minimum number of geographically nearby sites to accurately predict the power generated at the candidate site. As discussed earlier, the power output of geographically nearby sites are used as input features to build the regression model. Since in this experiment we are not interested in analyzing the confidence intervals, we use the entire training data to build the model (i.e., no bootstrapping). We vary the number of geographically nearby sites from 1 to 50 and for each value, we build 100 different models learnt from choosing random combinations of nearby sites.

Figure 6 shows the spread of average MAPE values as we vary the number of geographically nearby sites used for all 88 sites. We use the Random Forest regression technique to build the model. As expected, the average MAPE value reduces when more number of geographically nearby sites are used to predict the output. Note that as the nearby sites increase, the variations in nearby sites cancel out, which provides a more robust regression model. This suggests that an increase in the nearby site can improve the accuracy of the power generation model of a candidate site. We also note that the reduction in MAPE diminishes as the number of geographically nearby sites increases. With at least five randomly chosen geographically nearby sites, we observe that the MAPE is around 10%. This indicates that our algorithm can be effective in sparsely populated regions such as towns/villages, having few solar installations.

Next, we analyze the variability in performance of the different models as the number of geographically nearby sites increases. Figure 7 shows the spread of the standard deviation of the 100 models with increasing number of geographically nearby sites. As shown in the figure, we observe that the variability reduces when the number of nearby sites increases. However, unlike the previous result, the variability continues to reduce — albeit at a slower rate — even when the number of nearby sites is greater than five. Thus, the performance of the learned models is closer to its average.

5.4.3 Detection of anomalies. We illustrate the different steps involved in our algorithm using Figure 8. In the top subplot of the figure, the blue line depicts the power generation trace from a solar installation for over a week in August, 2015. The red marker shows the prediction from the RandomForest regression technique with data from the remaining 87 sites as features. While the prediction (i.e., red marker) closely follows the actual power output (i.e., blue line), there is a significant difference in the actual and predicted after 14th August. As seen, there is a sharp drop in the actual power generated in the late morning of 14th August. The drop in power is significant, and there is no output recorded in the site for an extended period until October (not shown in the figure). However, the regression model forecasts a non-negative power output for the given site.

The second subplot shows the residual, i.e., the difference between the actual and the predicted values (i.e., the black line) along with the confidence interval (i.e., the gray shaded region). The confidence interval, which is within $\pm 4\sigma$, is calculated using the pointwise standard deviation obtained from the bootstrap process. In this figure, we observe that the residual sometimes lie outside the confidence interval at the same time of the day across multiple days — which indicates a fixed periodic component.

On removing the seasonal component using our approach, we observe that the residual always lies within the confidence interval, except when there is an anomaly in power generation. This is shown in the third subplot of the figure, where the black line (i.e., residual) lie within the gray shaded region (i.e., the confidence interval). Finally, the last subplot depicts our anomaly detection algorithm in action. We observe that our algorithm accurately flags periods of no output as an anomaly (depicted by the red shaded region).

6 CASE-STUDY: ANOMALY DETECTION ANALYSIS

In this case study, we use the solar installations in the Dataport as they represent a typical setup within a city. We ran our SolarClique algorithm on the generation output from all solar installations and obtained the anomalous days in the dataset. Below, we present our analysis.

6.1 Anomalies in solar installations

Figure 9 shows the total number of anomalous days in each solar installation site. We observe that our SolarClique algorithm found anomalous days in 76 solar installations, out of the 88 sites in the dataset. As seen in the figure, the total number of anomalous days span from a day to several months. Together, all the installation sites had a total of 1906 anomalous days. This indicates a significant loss of renewable power output. Specifically, we observe that 17 of

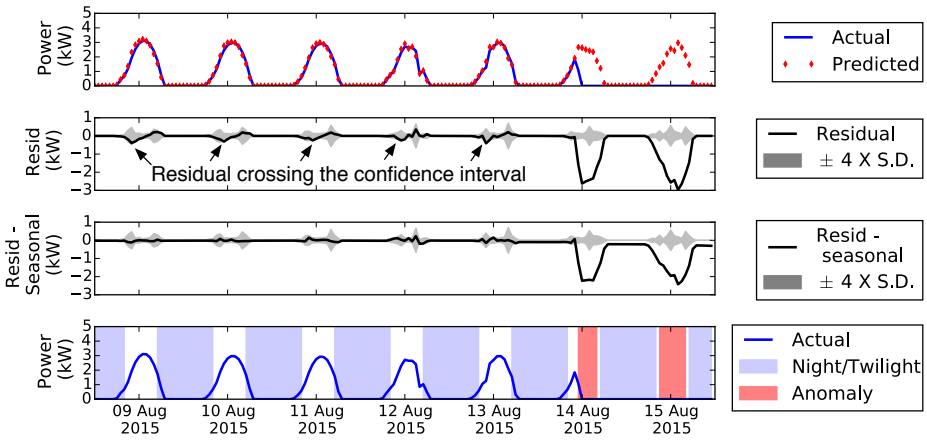


Figure 8: An illustrative example that depicts the data-processing and anomaly detection steps in SolarClique.

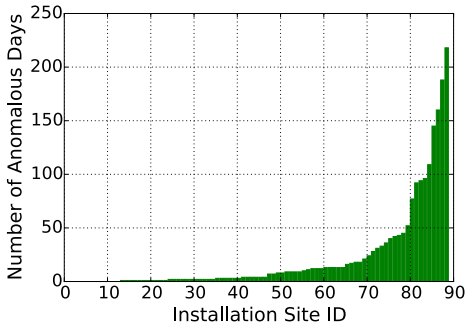


Figure 9: Number of anomalous days for each site. Installation sites are plotted in ascending order of anomalous days.

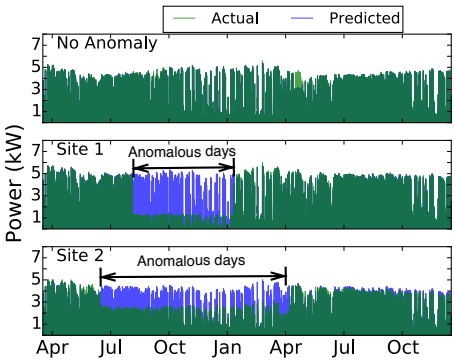


Figure 10: Under-production of solar detected using our algorithm.

the 88 (around 20%) installations had anomalous power generation for at least a total of one month that represents more than 5% of the overall 640 days in the testing period. Anomalies from these installations account for nearly 80% of all the anomalous days.

To better understand the anomalous periods, we group them into short-term and long-term periods. The short-term periods have less than three contiguous anomalous days, while the long-term periods have consecutive anomalous days for at least three days. Our results show the dataset has 587 occurrences of short-term periods spread over 683 days. Further, we observe 123 occurrences of long-term periods spread over 1223 days. We also observe that the maximum contiguous anomalous period found in a site was approximately five months (i.e., 158 days), with no power output during that period. Clearly, such high number of long-term anomalous periods demonstrate the need for early anomaly detection tools. Additionally, we note that long-term anomalies are relatively easier to detect than short-term anomalies. While long-term anomalies represent serious issues that may need immediate attention, short-term anomalies may be minor problems, if unattended, could become major problems in future. The advantage of our approach is we can detect both short-term and long-term anomalies.

6.2 Analysis of anomalies detected

Note that the reduction in power output depends on the severity of an anomaly. This is because some electrical faults (e.g., short-circuit of a panel) may have localized impact on a solar array, which can marginally reduce the power output, while other faults (e.g., inverter faults) may show significant power reduction or completely stop power generation.

SolarClique detects anomalous days when there is no solar generation and also when an installation under produces power. Our algorithm reported 1099 and 807 anomalous days with under production and no solar generation, respectively. Since no solar generation days are trivially anomalous, we specifically examine cases of solar under production. Figure 10 shows the power output from three different sites. The top plot shows the power output (depicted by the blue line) with no anomalous days, the subplots below show sites that have anomalous days (depicted by the red marker). Our results show that the SolarClique algorithm detects anomalies even when a site under produces solar power. Note that the site with no anomaly, which is exposed to the same solar irradiance as other

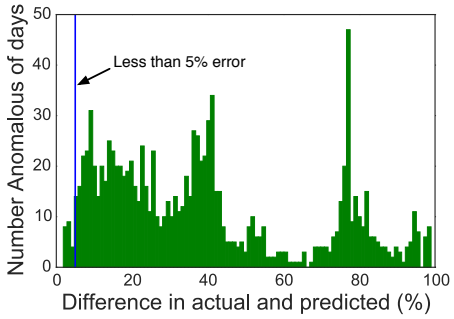
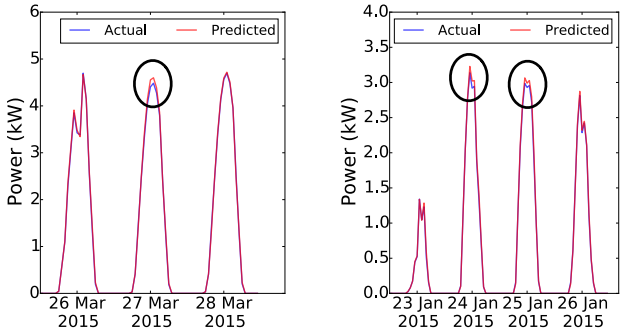


Figure 11: Distribution of the difference in actual and predicted on underproducing anomalous days.



(a) Site A (b) Site B

Figure 12: Anomalies detected in two sample sites where the difference in actual and predicted was less than 5%. The figure shows a good fit on all days except the anomalous period highlighted in the circle.

sites, continues to produce solar output. However, we observe a drop in power output for an extended period in the anomalous sites. Specifically, we observe the drop in power output is around 75% and 40% in Site 1 and Site 2, respectively — presumably due to factors such as line faults in the solar installation. Usually, anomalies such as line faults can cause a significant drop in the power output. In particular, a 75% drop in Site 1 can be attributed to faults in three fourth of the strings (i.e., connected in series).

We further examine the reduction in power output in the under-production cases. Figure 11 shows the distribution of the difference in actual and predicted power output for anomalous days. Out of the1099 under production days, our algorithm reported 23 days when the difference in percentage was less than or equal to 5%. Typically, more than 5% drop in power output is considered significant. This is because malfunctioning of a single panel in a solar array with 20 panels[6] will result in a 5% reduction. Thus, we investigate anomalous days wherein the difference is less than or equal to 5%. Figure 12 compares the regression fit of anomalous days with two normal days (adjacent to the anomalous days) from two sample sites where the difference was less than 5%. Note that the figure shows a good fit for most periods except during the anomalous period highlighted in the circle. In comparison to other periods, we observe a drop in power during the anomalous period, occurring during the mid-day. Even though the difference in percentage is small, it represents a relatively significant drop since the power output is at its peak during the mid-day.

[6]Typically, a 5kW installation capacity has 20 panels, each panel having 250W capacity.

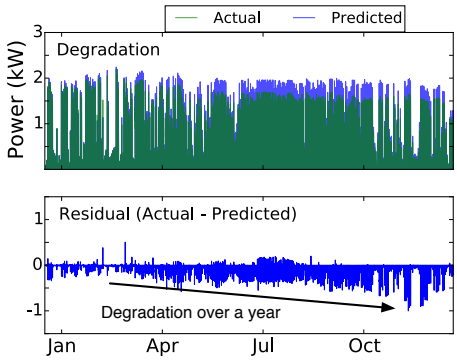


Figure 13: Accelerated degradation in the power output of a solar site.

Anomaly Type	#Sites	#Days	Avg. power reduction(%)
Single No Production	5	515	98.87
Multiple No Production	3	295	98.65
Single Under Production	2	348	60.22
Multiple Under Production	4	164	43.63
Severe Degradation	3	179	30.67

Table 2: Types of anomaly in sites having more than a month of anomalous days.

We observe that our approach also detects anomalies due to degradation in the power output, which usually spans over an extended time period. Since the drop in power output over the time period may be small, such changes are more subtle and harder to identify. Figure 13 shows the degradation in power output of an anomalous site. Our algorithm reports an increase in the frequency of anomalous days in the installation site over the year, with more anomalous days in the latter half. To understand the increase in anomalous days, we plot the difference between the actual and predicted (seen in the bottom subplot). We observe that the difference between the actual and predicted value steadily increases over time. It is known that the power output of solar installations may reduce over time due to aging [22] at a rate of around 1% a year. However, the accelerated degradation seen in Figure 13 is presumably due to occurrences of hot-spots or increased contact resistance due to corrosion. Early detection of such conditions can help homeowners take advantage of product warranties available on solar panels.

We now examine the types of anomalies in the top 17 sites with more than a month of anomalous days. The power output of anomalous days can be categorized into three types — (i) no production,

(ii) under production, and (iii) degradation over a period. Table 2 summarizes the different types occurring over a period in these sites. The single period represents a single contiguous period of anomaly, while the multiple period represents more than one contiguous period. We observe that the average power reduction during anomalous periods may range from 98.8% to 30.6%. We classify "no production days" as days with no power output for the majority of the period. Overall, we observe that there are 810 no production days — a significant loss in renewable output. Although the average power reduction due to severe degradation is 30%, it is likely to grow over time.

7 FUTURE EXTENSIONS TO SOLARCLIQUE

As mentioned earlier, several third-party sites exist that host solar generation data for rooftop installations. While in our approach, we use power to determine the existence of anomalies in power generation, several other electrical characteristics such as voltage and current are available that carry much richer information about the type of anomaly. This information can be leveraged to further infer the exact type of anomaly in power generation. For example, a line fault (broken string) will reduce the current produced by the overall setup, but the voltage will remain unchanged. Conversely, covering of dust/bird droppings can impact both the voltage and the current. Thus, our algorithm can be extended to use multi-modal data (e.g., voltage, current, and power) to further diagnose the exact cause of the anomaly.

Our approach can also be extended to a single solar installation for detecting anomalies. With the proliferation of micro-inverters in residential solar installations, power generation data from individual panels are available. Power output from these colocated panels can also be used to detect faults in the PV setup, as they can predict the power output with higher fidelity. This can be used in remote locations where data from other solar installations are not easily available. As part of future work, we plan to use SolarClique algorithm to discover faults in a single panel by comparing power generated with others in the same setup.

8 RELATED WORK

There has been significant work on predicting the solar output from solar arrays [7, 8, 16, 20, 28]. While some studies have used site-specific data such as panel configuration [8, 20] for building the prediction model, others have used external data such as weather or historical generation data [17, 28]. Such models can provide short-term generation forecast (e.g., an hour) to long-term forecast (e.g., days or weeks). Although these studies can predict the reduction in power output, a limitation in these studies is that they cannot attribute the reduction to anomalies in the solar installation.

Prior work has also focused on anomaly detection in PV panels [14, 15, 22, 25, 30–32]. These studies propose methods to model the effects of shades/covering [14, 19], hot-spots [18], degradation [22, 30] or short-circuit and other faults [15]. However, these methods require extensive data (such as statistics on different types of anomalies) [29] or do not focus on hardware-related issues [14]. For instance, [29] proposes a solution to determine probable causes of anomalies but require detailed site-specific information along with pre-defined profiles of anomalies. Unlike prior approaches, our approach doesn't require such extensive data or setup and relies instead on power generation from co-located sites. Thus, it provides a scalable and cost-effective approach to detect anomalies in thousands of solar installation sites.

The idea behind our approach is similar to [26, 27]. However, the authors use the approach in the context of an astronomy application, wherein systematic errors are removed to detect exoplanets. In this case, the systematic errors are confounding factors due to telescope and spacecraft, which influences the observations from distant stars. In contrast, our solution uses inputs from other geographically nearby sites to detect anomalies in solar. As discussed earlier, today, such datasets are easily accessible over the internet, which makes our approach feasible. Further, using regression on the data from neighbors has been studied earlier [10]. However, the main focus of this work was in the context of quality control in climate observations by imputing missing values. In our case, we use the learned regression model to find anomalous solar generation.

9 CONCLUSION

In this paper, we proposed SolarClique, a data-driven approach to detect anomalies in power generation of a solar installation. Our approach requires only power generation data from geographically nearby sites and doesn't rely on expensive instrumentation or other external data. We evaluated SolarClique on the power generation data over a period of two years from 88 solar installations in Austin, Texas. We showed how our solar installation regression models are accurate with tight confidence intervals. Further, we showed that our approach could generate models with as few as just five geographically nearby sites. We observed that out of the 88 solar installations, 76 deployments had anomalies in power generation. Additionally, we found that our approach is powerful enough to distinguish between reduction in power output due to anomalies and other factors (such as cloudy conditions). Finally, we presented a detailed analysis of the different anomalies observed in our dataset.

Acknowledgment This research is supported by NSF grants IIP-1534080, CNS-1645952, CNS-1405826, CNS-1253063, CNS-1505422, CCF-1522054 and the Massachusetts Department of Energy Resources.

REFERENCES

[1] 2016. When Will Rooftop Solar Be Cheaper Than the Grid? https://goo.gl/h1Ayy5. (2016). Accessed March, 2018.
[2] 2017. Dataport dataset. https://dataport.cloud/. (2017).
[3] 2017. EIA adds small-scale solar photovoltaic forecasts to its monthly Short-Term Energy Outlook. https://www.eia.gov/todayinenergy/detail.php?id=31992. (2017). Accessed March, 2018.
[4] 2018. SciPy Stack. http://www.scipy.org/stackspec.html. (Accessed March 2018).
[5] Mohamed Hassan Ali, Abdelhamid Rabhi, Ahmed El Hajjaji, and Giuseppe M Tina. 2017. Real Time Fault Detection in Photovoltaic Systems. *Energy Procedia* (2017).
[6] Rob W Andrews, Andrew Pollard, and Joshua M Pearce. 2013. The effects of snowfall on solar photovoltaic performance. *Solar Energy* 92 (2013), 84–97.
[7] Yona Atsushi and Funabashi Toshihisa. 2007. Application of recurrent neural network to short-term-ahead generating power forecasting for photovoltaic system. In *Power Engineering Society General Meeting*. Tampa, Florida, USA.
[8] Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. 2009. Online short-term solar power forecasting. *Solar Energy* (2009).
[9] Robert B Cleveland, William S Cleveland, and Irma Terpenning. 1990. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics* (1990).
[10] Christopher Daly, Wayne Gibson, Matthew Doggett, Joseph Smith, and George Taylor. 2004. A probabilistic-spatial approach to the quality control of climate

observations. In *Proceedings of the 14th AMS Conference on Applied Climatology, Amer. Meteorological Soc., Seattle, WA*.

[11] Chris Deline. 2009. Partially shaded operation of a grid-tied PV system. In *Photovoltaic Specialists Conference (PVSC), 2009 34th IEEE*. IEEE, 001268–001273.

[12] Mahmoud Dhimish, Violeta Holmes, and Mark Dales. 2017. Parallel fault detection algorithm for grid-connected photovoltaic plants. *Renewable Energy* (2017).

[13] Ran Fu, David J Feldman, Robert M Margolis, Michael A Woodhouse, and Kristen B Ardani. 2017. *US solar photovoltaic system cost benchmark: Q1 2017*. Technical Report. National Renewable Energy Laboratory (NREL), Golden, CO (United States).

[14] Peter Xiang Gao, Lukasz Golab, and Srinivasan Keshav. 2015. What's Wrong with my Solar Panels: a Data-Driven Approach.. In *EDBT/ICDT Workshops*. 86–93.

[15] Elyes Garoudja, Fouzi Harrou, Ying Sun, Kamel Kara, Aissa Chouder, and Santiago Silvestre. 2017. Statistical fault detection in photovoltaic systems. *Solar Energy* (2017).

[16] Rui Huang, Tiana Huang, Rajit Gadh, and Na Li. 2012. Solar generation prediction using the ARMA model in a laboratory-level micro-grid. In *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*. IEEE.

[17] Srinivasan Iyengar, Navin Sharma, David Irwin, Prashant Shenoy, and Krithi Ramamritham. 2017. A Cloud-Based Black-Box Solar Predictor for Smart Homes. *ACM Transactions on Cyber-Physical Systems* 1, 4 (2017), 21.

[18] Katherine A Kim, Gab-Su Seo, Bo-Hyung Cho, and Philip T Krein. 2016. Photovoltaic hot-spot detection for solar panel substrings using ac parameter characterization. *IEEE Transactions on Power Electronics* (2016).

[19] Alexander Kogler and Patrick Traxler. 2016. Locating Faults in Photovoltaic Systems Data. In *International Workshop on Data Analytics for Renewable Energy Integration*. Springer.

[20] Elke Lorenz, Johannes Hurka, Detlev Heinemann, and Hans Georg Beyer. 2009. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. *IEEE Journal of selected topics in applied earth observations and remote sensing* (2009).

[21] Ricardo Marquez and Carlos FM Coimbra. 2013. Intra-hour DNI forecasting based on cloud tracking image analysis. *Solar Energy* (2013).

[22] Ababacar Ndiaye, Cheikh MF Kébé, Pape A Ndiaye, Abdérafi Charki, Abdessamad Kobi, and Vincent Sambou. 2013. A novel method for investigating photovoltaic module degradation. *Energy Procedia* (2013).

[23] Judea Pearl. 2009. *Causality*. Cambridge university press.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011).

[25] M Sabbaghpur Arani and MA Hejazi. 2016. The comprehensive study of electrical faults in PV arrays. *Journal of Electrical and Computer Engineering* 2016 (2016).

[26] Bernhard Schölkopf, David Hogg, Dun Wang, Dan Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. 2015. Removing systematic errors for exoplanet search via latent causes. In *International Conference on Machine Learning*.

[27] Bernhard Schölkopf, David W Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. 2016. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences* (2016).

[28] Navin Sharma, Pranshu Sharma, David Irwin, and Prashant Shenoy. 2011. Predicting solar generation from weather forecasts using machine learning. In *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*. IEEE, 528–533.

[29] S Stettler, P Toggweiler, E Wiemken, W Heydenreich, AC de Keizer, WGJHM van Sark, S Feige, M Schneider, G Heilscher, E Lorenz, and others. 2005. Failure detection routine for grid-connected PV systems as part of the PVSAT-2 project. In *Proceedings of the 20th European Photovoltaic Solar Energy Conference & Exhibition, Barcelona, Spain*. 2490–2493.

[30] Ali Tahri, Takashi Oozeki, and Azzedine Draou. 2013. Monitoring and evaluation of photovoltaic system. *Energy Procedia* (2013), 456–464.

[31] Patrick Traxler. 2013. Fault detection of large amounts of photovoltaic systems. In *Proceedings of the ECML/PKDD 2013 Workshop on Data Analytics for Renewable Energy Integration*.

[32] Achim Woyte, Mauricio Richter, David Moser, Stefan Mau, Nils Reich, and Ulrike Jahn. 2013. Monitoring of photovoltaic systems: good practices and systematic analysis. In *Proc. 28th European Photovoltaic Solar Energy Conference*.