SunDown: Model-driven Per-Panel Solar Anomaly Detection for Residential Arrays

Menghong Feng
University of Massachusetts Amherst
mfeng@umass.edu

Noman Bashir
University of Massachusetts Amherst
nbashir@umass.edu

Prashant Shenoy
University of Massachusetts Amherst
shenoy@cs.umass.edu

David Irwin
University of Massachusetts Amherst
deirwin@umass.edu

Dragoljub Kosanovic
University of Massachusetts Amherst
kosanovic@umass.edu

ABSTRACT

Solar arrays often experience faults that go undetected for long periods of time, resulting in generation and revenue losses. In this paper, we present SunDown, a sensorless approach for detecting per-panel faults in solar arrays. SunDown's model-driven approach leverages correlations between the power produced by adjacent panels to detect deviations from expected behavior, can handle concurrent faults in multiple panels, and performs anomaly classification to determine probable causes. Using two years of solar data from a real home and a manually generated dataset of solar faults, we show that our approach is able to detect and classify faults, including from snow, leaves and debris, and electrical failures with 99.13% accuracy, and can detect concurrent faults with 97.2% accuracy.

CCS CONCEPTS

• **Hardware → Renewable energy**; • **Computing methodologies → Anomaly detection**.

KEYWORDS

Solar anomaly detection; data-driven modeling; machine learning

ACM Reference Format:
Menghong Feng, Noman Bashir, Prashant Shenoy, David Irwin, and Dragoljub Kosanovic. 2020. SunDown: Model-driven Per-Panel Solar Anomaly Detection for Residential Arrays. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '20), June 15–17, 2020, , Ecuador*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3378393.3402257

1 INTRODUCTION

Recent technological advances and falling prices has led to a significant increase in deployments of both large utility-scale and smaller residential solar arrays. Large utility-scale solar farms tend to be instrumented with sensors for monitoring real-time generation to identify production issues. Due to cost reasons, smaller residential-scale systems lack such sensing and instrumentation and may only have coarse-grain monitoring capabilities, at best, to detect system-level faults. Thus, it is not uncommon for residential solar arrays to encounter power anomalies or other local faults that go undetected for long periods, resulting in generation and revenue losses.

To address these challenges, we present *SunDown*, a sensorless approach for detecting per-panel faults in small-scale solar arrays. Prior work on per-panel solar anomaly detection are based on time series [15] or statistical [3, 30] analysis of a panel's output or use of sensors such as a pyranometer [12] to detect faults. In contrast, our approach uses the actual output from *other* nearby panels to estimate each panel's expected output and find anomalous deviations from this estimate. Our model-driven approach is based on machine learning and, similar to [15], can detect physical anomalies, such as snow, leaves, and electric faults at panels. In designing, implementing, and evaluating our SunDown system we make the following contributions.

1. We present a model-driven approach that leverages correlations in the generated output between adjacent panels to predict the expected output of a particular panel and flags anomalies when the model predictions deviate from the expected values. Further, our approach can handle and detect multiple concurrent faults in the system, a key challenge that has not been addressed by prior work. We present a random forest-based classification technique to classify the probable cause of the observed fault.

2. We construct a real-world labelled dataset of solar anomalies that we release to the community. Using this dataset, we show that SunDown has a MAPE of 2.98% when predicting per-panel output, demonstrating the efficacy of using nearby panels to perform model-driven predictions. Furthermore, SunDown is able to detect and classify faults such as snow cover, leaves, and electrical failures with 99.13% accuracy for single faults and is able to handle concurrent faults in multiple panels with 97.2% accuracy.

2 BACKGROUND

This section presents background on solar anomaly detection.

Residential Solar Arrays. Our work primarily focuses on residential solar arrays that are typically small-scale installations with capacities of 10kW or less and comprise a few to a few dozen solar panels (see Figure 1). We assume that the power generation of the array can be monitored at a per panel level. This is a reasonable assumption since many residential arrays are equipped with micro-inverters (e.g. Enphase micro-inverters [1]) or DC power optimizers [2]. As shown in Figure 1, such systems provide real-time per-panel

COMPASS '20, June 15–17, 2020, , Ecuador

ACM ISBN 978-1-4503-7129-2/20/06…$15.00
https://doi.org/10.1145/3378393.3402257

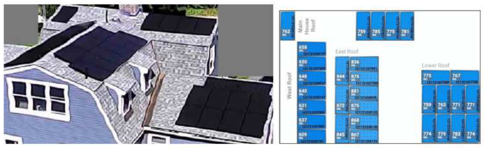


Figure 1: A residential solar array (left) with 31 panels deployed on four roof planes, and per-panel power data (right).

generation data, which is essential for our approach. We seek to develop a sensor-less approach for per-panel solar anomaly detection and do not assume any other sensor or instrumentation.

Solar Generation. Solar generation at any site depends upon on its location, time, physical characteristics, cloud cover, and temperature [6, 18]. For the purpose of this work, we assume that per-panel solar generation on any given day can be reduced to two factors: *transient*, which consists of factors that temporarily impact power output, and *faults* which consist of factors that have a prolonged negative impact on output. Transient factors include weather conditions such as cloud cover, wet panels caused by rain or dew, as well as site specific factors, such as shading caused by nearby trees or other structures. We classify transient factors into two classes—common transient factors that affect all the panels on a site or local transient factors that impact only a subset of panels on that site.

Solar Faults. Anomalies in our case are defined to be factors that cause a persistent drop in production but can be rectified by the owner of the site. We are particularly interested in the following three types of faults (1) snow cover on one or more panels, (2) partial occlusions such as dust or leaves on a panel, (3) electric faults such as module failure, short circuits or open circuits. These faults cause either a reduction in output or zero output for a particular panel or a subset of panels. Due to their close proximity, multiple panels in a residential array may experience the same fault—for example, snow may cover multiple adjacent panels (or even the entire system), resulting in concurrent faults. Of course, a site may also suffer a full system outage, which is also a fault but is easier to detect.

Problem Statement. Consider a solar array with N solar panels. We assume that the panels are mounted on a residential roof across one or more roof planes. Given such a setup, our problem is to design a technique that monitors the power output of each panel and the entire system, and labels the observed output in each time interval (e.g. a day) as normal or abnormal. Further, our technique should identify specific solar panels in the system that are experiencing faults and also determine the possible cause of the fault (e.g. snow, partial occlusion, or electric fault).

3 PER-PANEL SOLAR ANOMALY DETECTION

In this section, we describe our model-driven approach for per-panel solar fault detection.

3.1 Basic Idea

Consider a solar installation with N panels. Suppose that k panels are experiencing an anomaly that results in a reduction, or loss, of output from those panels. Initially, let us assume $k = 1$ (only one panel out of N is faulty). Since all N panels are mounted on the same roof in close proximity to each other, it follows that they experience highly correlated weather conditions, and produce similar output. Thus, our "sensorless" approach first constructs a model to predict the expected output of a panel from n neighboring panels ($n \geq 1$). For example, a simple predictor is one that uses the mean output of n neighboring panels to estimate a particular panel's output. Under normal conditions, the model prediction will match the observed output of that panel with high accuracy.

When a panel experiences an anomaly, however, the model predictions will continue to estimate the "normal case" output of that panel, while the observed output will deviate from this normal case. If the deviation is "large" and persists over an extended period of time, it is indicative of a fault, rather than an error in the model prediction. The cause of the fault can be separately determined by analyzing the amount of loss or the power pattern exhibited by the panel. Such a model-driven approach only uses the panels' observed output to detect anomalies—no other instruments or sensors are needed for anomaly detection unlike some approaches [5].

3.2 Model-Based Predictions

We now present two model-driven techniques for predicting the power output of an individual panel using neighboring panels.

3.2.1 Linear Regression-Based Model. Since the power generated by solar panels in close proximity of one another are highly correlated, we can use regression to predict the output of a panel given the observed output of neighboring panels.

Let P_i denote the observed power output of panel i at time instant i. Let us assume we wish to predict the output of panel i using n other panels. A linear regression model allows us to estimate the output of desired panel as a linear function of the others:

$$P_i = w_i P_{i1} + w_2 P_{i2} + w_3 P_{i3} + \ldots + w_n P_{in} + \epsilon_i \quad (1)$$

where $X = \{i_1, i_2, ..., i_n\}$ is the set of n panels used to model the output of the i^{th} panel. We can use linear regression to estimate the weight w_i that minimizes the error term ϵ_i.

Such an approach yields N distinct regression models, one for each panel in the system, where each model makes a prediction using the observed output of n other panels. To determine if a panel has a fault, we compare the model prediction $P_i(t)$ at time t with the observed value $\hat{P}$. If the difference between the model's prediction and observed value is large and persists over a period of time (e.g., a day or multiple days), the approach flags that panel as faulty.

3.2.2 Graphical Model and Half-Sibling Regression. Our second model is based on a recently proposed machine learning technique called half-sibling regression that uses a Bayesian approach to remove the effects of confounding variables [28]. This approach is based on our prior work on SolarClique[17] that predicted the output of an entire array using nearby solar arrays. We draw inspiration from the half-sibling regression method [28] and SolarClique [17] for SunDown's *per-panel* anomaly detection. Additional details of our the approach, which is summarized below, can be found in [11]. Using the Bayesian approach, our algorithm to estimate the amount of production loss due to anomalies is as follows.

We first use regression to estimate the power output of a particular panel, denoted by a random variable P, using the power output of n other panels in the system, denoted by a random variable X (a vector of size n). The regression yields $E[P|X]$ - an estimate of P given the observed output of n neighboring panels that constitute

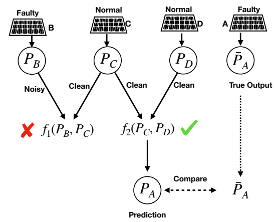


Figure 2: A forecasting model is used to ensure non-noisy inputs to our Bayesian model.

X. Since P itself is observed, subtracting $E[P|X]$ from P yields an estimate of the output loss $\hat{L}$ due to transient factor and anomalies. A key difference between the linear regression model of Section 3.2.1 and here is that we use bootstrapping to construct multiple regression models by subsampling the data (instead of a single regression model) and use an ensemble method based on Random Forest that uses the mean of multiple models to estimate $E[P|X]$.

Next, since $\hat{L}$ contains effects of transient factors such as shade on panels as well anomalies, we must remove the impact of transient factors to obtain the "true" anomalies. We can use time series decomposition to extract the seasonal component that represents the shading effects that occur daily at set time periods and remove it from $\hat{L}$ [17]. The remainder of $\hat{L}$ then represents production loss at that panel due to any anomalies. Under normal operation $\hat{L}$ will be close to zero (no anomalies and no loss of output). When $\hat{L}$ is significant and persistent over a period of time, our model-driven approach flags an anomaly in the panel.

3.3 Handling Multiple Concurrent Faults

Next we consider the case where $k > 1$ and multiple panels are faulty. To handle this case, we construct *multiple models for each panel* by choosing different subsets of n panels out of N for each model. Any model that uses faulty panels as input will have higher errors while a model that uses all non-faulty inputs will continue to provide good predictions. Our goal then is to construct multiple models, and then choose one of these models at each instant that uses non-faulty inputs. To distinguish between faulty and non-faulty inputs, we use a solar forecasting approach that predicts the output of the solar panel based on weather forecasts [6, 18]. Using the forecasting model, we label panels as "normal" or "noisy" if the model predicted power is close to the observed power or deviates significantly, respectively. Any model that uses one or more noisy panels as an input should be eliminated from consideration for anomaly detection purposes. Figure 2 illustrates the process, where a model based on B and C panels is discarded as B is noisy. A model based on panels with normal output, C and D, is used to make a prediction for panel A. Note that, forecasting models cannot be directly used for anomaly detection as they exhibit high error leading to higher false positive as compared to the Bayesian approach that uses the actual panel output.

4 CLASSIFYING SOLAR ANOMALIES

Given anomalies detected by our Bayesian model we use a random forest classifier to label the possible cause of the fault for each panel that is faulty. The classifier needs to distinguish between three types of faults: snow, partial occlusion and open circuit. Note that partial snow over a panel and partial occlusion faults both result in diminished, but non-zero output. Full snow cover on a panel and open circuit faults both yield zero output. To distinguish between these cases, we first sample 40 randomly chosen points over an entire day and compute the percentage reduction in power output when compared to the model's predictions for each of these points. This power loss vector is a key feature of our classifier. We also use two other features: month of the year and snow depth values from NOAA's weather service. We train our random forest classifier using a training dataset of real snow and synthetic anomalies. Depending on the season (winter versus other seasons) and the observed power loss over a period of time, our classifier can label the probable cause of the fault for each panel. Our approach can also label system-wide faults, caused either by a system-wide electrical failure or full snow cover, both of which cause near total loss of power output.

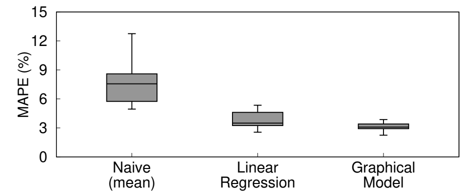


Figure 3: Machine Learning Model

5 EXPERIMENTAL EVALUATION

We evaluate SunDown by quantifying (1) the accuracy of model-based power inference where we infer the output of a single panel using nearby panels and (2) the accuracy of our anomaly classification. We quantify the accuracy of predicting a panel's output using Mean Absolute Percentage Error (MAPE) between the inferred output and the actual solar generation, as below.

$$MAPE = \frac{1}{m}\sum_{t=1}^{m}\left|\frac{P_O(t) - P_I(t)}{\bar{P_O}}\right| \quad (2)$$

Here, m is the number of samples, $P_O(t)$ is the observed solar power at time t, $P_I(t)$ is the inferred power at time t, and $\bar{P_O}$ is the mean of observed power generation. We use three different metrics to quantify different aspects of the classification task: accuracy, sensitivity, and specificity.

Solar Anomaly Open Dataset. Since there are no datasets of solar faults available for research use, we constructed a labelled dataset using two arrays: a 31-panel production residential site, and a 20-panel ground-mounted site where we introduced anomalies, such as dust, leaves, and electrical faults, to mimic real-world faults. Our dataset is available at http://traces.cs.umass.edu and details of our dataset construction can be found in [11].

5.1 Prediction Model Accuracy

We begin by evaluating the accuracy of predicting the power output of an individual panel using neighboring panels.

5.1.1 Machine Learning Model. To evaluate the accuracy of model inference, we choose test data only from the days where the site experiences no anomaly. We then use the normal days of the home dataset to train our linear regression and graphical model. We also compare their performance with a naive approach that infers the power output of a panel as the mean output of n other panels. As

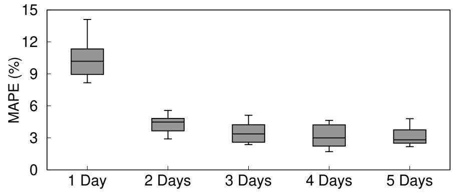
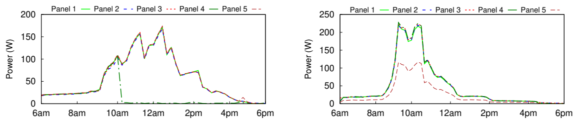


Figure 4: Size of training data required



(a) Synthetic Open Circuit (b) Synthetic Leaf Occlusion

Figure 5: Synthetic fault injection of different types.

shown in Figure 3, the MAPE values for the Bayesian model, linear regression, and naive approach are 3%, 4%, and 8.6%, respectively. The naive approach has the worst accuracy since it assumes all panels produce similar output, which is not true in many cases due to panel level variations. Linear regression works well when the output of different panels is highly correlated, which is not true when some of the panels experience partial shading. Our graphical model is able to capture non-linear relationships, and yields the highest accuracy and a tight confidence interval. We use the graphical model for our subsequent experiments, unless stated otherwise.

5.1.2 Impact of Training Data Size . Next, we evaluate model accuracy for different amounts of training data. If a model requires a lot of training data for good accuracy, it can hinder its use for solar sites that have been recently deployed or for the sites where long-term panel level data is not available. We vary the training data size and evaluate its accuracy for predicting output using a test dataset. Figure 4 demonstrates that our model can achieve reasonable accuracy and a 10% MAPE with only one day of per panel data. If the number of days is increased to 4, the MAPE drops to 3.5% and stays almost constant beyond four days.

5.2 Anomaly Classification Accuracy

We next evaluate the accuracy of our model-driven approach and classifier in detecting and classifying anomalies, respectively. The common anomalies we consider include snow fault, open circuit, and partial occlusions due to leaves.

Our home dataset already includes real snow faults that are labelled and we evaluate the accuracy of our classifier on identifying these snow faults. We then use the synthetic faults from our solar anomaly datatset and inject them into the home data set by introducing synthetic single panel faults as well as concurrent fault and evaluate the accuracy of our classifier. Figure 5 presents per-panel data for a typical day when an electric fault or object covering anomaly has been injected into one or many panels.

5.2.1 Snow Fault Detection. We first evaluate the ability of our classifier in detecting snow faults in the home dataset. We extract the features from daily power output, which include Pearson's correlation coefficient, ratio of maximum observed power and the nominal panel capacity, and weather data such as snow and cloud cover and use them as inputs to our random forest classifier. Table

Classification	Accuracy	Specificity	Sensitivity
System level	98.13%	95.12%	100%
Single, panel-level	98.78%	97%	100%
Multiple panel-level	97.2%	97.06%	97.26%

Table 1: Classification Metrics

1 shows that our approach is able identify system-level snow faults with 99.13% accuracy, sensitivity of 100%, and specificity of 95.12%.

5.2.2 Single and Concurrent Fault Classification. We next show that our approach is capable of fine-grain anomaly detection and classification of a single fault and it is also capable of detecting concurrent faults in a subset of the panels. To do so, we use our solar anomaly dataset and choose the partial occlusion and open circuit anomaly from the dataset and inject these faults into a single, randomly chosen, panel of the array; different panels have faults injected into them on different days. We use our model to detect the presence of the fault and our random forest classifier to identify the type of fault. We next inject multiple concurrent faults of all types (snow, occlusion, open circuit) into the array using a similar methodology and attempt to detect and classify each fault using our model and classifier. Note that, in this case, we need to use our concurrent fault detection approach. Table 1 shows our model can classify single faults with an accuracy of 98.78%, specificity of 97%, and sensitivity of 100%. For concurrent faults, the model obtains accuracy of 97.2%, specificity of 97.06%, and sensitivity of 97.26%.

6 RELATED WORK

There has been significant work on predicting power output for solar sites [4, 6, 10, 23, 24, 27, 29]. All of these studies predict only system level output and generally report 20-30% error. These high errors and inability to predict panel level output would cause their prediction for all panels to be the same, and limit their ability to detect and classify anomalies. There is also significant prior work on anomaly detection and classification in solar photovoltaic systems, which can be broadly classified into model-based approaches [9, 13, 16, 19, 20] and machine learning based [7, 8, 12, 14, 21, 22, 25, 26, 31, 32] approaches. Some of these studies use power data from nearby solar sites [17, 30] to detect and classify anomalies. In [30], authors compare the performance of different solar arrays at the same site, but do not do anomaly classification. Our work uses the output of other nearby panels to predict a panel's output for detecting faults and can classify various types of faults, i.e. snow, object covering, and electrical faults, on a single or multiple panels.

7 CONCLUSIONS

In this paper, we proposed SunDown, a sensorless approach to detecting per-panel anomalies in residential solar arrays. We take a model-driven approach that leverages correlations between the power produced by adjacent panels to detect deviations from expected behavior. We constructed and released an open dataset of solar anomaly faults for experimental use. Finally, we showed that our approach can predict panel level output with a MAPE of 2.98% and can correctly classify anomalies with >97% accuracy.

Acknowledgements This research was supported by NSF grant 1645952, DOE grants DE-EE0007708, DE-EE0008277, and MA DOER.

REFERENCES

[1] 2020. Enphase Microinverters. https://enphase.com/en-us/products-and-services/microinverters.

[2] 2020. SolarEdge Power Optimizer. https://www.solaredge.com/products/power-optimizer.

[3] 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 *Proc. IEEE 29th International Conference on Advanced Information Networking and Applications*.

[4] 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*.

[5] A. Arenella, A. Greco, A. Saggese, and M. Vento. 2017. Real Time Fault Detection in Photovoltaic Cells by Cameras on Drones. In *International Conference on Image Analysis and Recognition*. Springer.

[6] N. Bashir, D. Chen, D. Irwin, and P. Shenoy. 2019. Solar-TK: A Data-driven Toolkit for Solar PV Performance Modeling and Forecasting. In *IEEE International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*.

[7] M.D. Benedetti, F. Leonardi, F. Messina, C. Santoro, and A. Vasilakos. 2018. Anomaly Detection and Predictive Maintenance for Photovoltaic Systems. *Neurocomputing* (2018).

[8] W. Chine, A. Mellit, V. Lughi, A. Malek, G. Sulligoi, and A.M. Pavan. 2016. A Novel Fault Diagnosis Technique for Photovoltaic Systems Based on Artificial Neural Networks. *Renewable Energy* (2016).

[9] M. Dhimish, V. Holmes, and M. Dales. 2017. Parallel Fault Detection Algorithm for Grid-connected Photovoltaic Plants. *Renewable Energy* (2017).

[10] N.A. Engerer and F.P. Mills. 2014. Kpv: A Clear-sky Index for Photovoltaics. *Solar Energy* 105 (July 2014).

[11] M. Feng, N. Bashir, P. Shenoy, D. Irwin, and B. Kosanovic. 2020. SunDown: Model-driven Per-Panel Solar Anomaly Detection for Residential Arrays. arXiv:2005.12181 [eess.SP]

[12] P.X. Gao, L. Golab, and S. Keshav. 2015. What's Wrong with my Solar Panels: a Data-Driven Approach. In *EDBT/ICDT Workshops*.

[13] E. Garoudja, F. Harrou, Y. Sun, K. Kara, A. Chouder, and S. Silvestre. 2017. Statistical Fault Detection in Photovoltaic Systems. *Solar Energy* (2017).

[14] F. Harrou, A. Dairi, B. Taghezouit, and Y. Sun. 2019. An Unsupervised Monitoring Procedure for Detecting Anomalies in Photovoltaic Systems using a One-class Support Vector Machine. *Solar Energy* (2019).

[15] Bo Hu. 2012. Solar Panel Anomaly Detection and Classification. MS Thesis, University of Waterloo, Canada.

[16] Y. Hu, B. Gao, X. Song, G.Y. Tian, K. Li, and X. He. 2013. Photovoltaic Fault Detection using a Parameter Based Model. *Solar Energy* (2013).

[17] S. Iyengar, S. Lee, D. Sheldon, and P. Shenoy. 2018. Solarclique: Detecting Anomalies in Residential Solar Arrays. In *ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–10.

[18] S. Iyengar, N. Sharma, D. Irwin, P. Shenoy, and K. Ramamritham. 2017. A Cloud-Based Black-Box Solar Predictor for Smart Homes. *ACM Transactions on Cyber-Physical Systems* (2017).

[19] B.K. Kang, S.T. Kim, S.H. Bae, and J.W. Park. 2012. Diagnosis of Output Power Lowering in a PV Array by Using the Kalman-filter Algorithm. *IEEE Transactions on Energy Conversion* (2012).

[20] K.A. Kim, G.S. Seo, B.H. Cho, and P.T. Krein. 2015. Photovoltaic Hot-spot Detection for Solar Panel Substrings using AC Parameter Characterization. *IEEE Transactions on Power Electronics* (2015).

[21] G. Liu and W. Yu. 2017. A Fault Detection and Diagnosis Technique for Solar System Based on Elman Neural Network. In *IEEE Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. IEEE.

[22] G. Liu, W. Yu, and L. Zhu. 2018. Condition Classification and Performance of Mismatched Photovoltaic Arrays via a Pre-filtered Elman Neural Network Decision Making Tool. *Solar Energy* (2018).

[23] E. Lorenz, D. Heinemann, H. Wickramarathne, H.G. Beyer, and S. Bofinger. 2007. Forecast of Ensemble Power Production by Grid-connected PV Systems. In *20th European PV Conference*. Milano.

[24] E. Lorenz, J. Hurka, D. Heinemann, and H.G. Beyer. 2009. Irradiance Forecasting for the Power Prediction of Grid-connected Photovoltaic Ssystems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2009).

[25] H. Mekki, A. Mellit, and H. Salhi. 2016. Artificial Neural Network-based Modelling and Fault Detection of Partial Shaded Photovoltaic Modules. *Simulation Modelling Practice and Theory* (2016).

[26] J. Pereira and M. Silveira. 2018. Unsupervised Anomaly Detection in Energy Time Series Data using Variational Recurrent Autoencoders with Attention. In *IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE.

[27] R. Perez, J. Schlemmer, S. Kivalov, J. Dise, P. Keelin, M. Grammatico, T. Hoff, and A. Tuohy. 2018. A New Version of the SUNY Solar Forecast Model: A Scalable Approach to Site-Specific Model Training. In *2017 IEEE 45th Photovoltaic Specialists Conference (PVSC)*.

[28] B. Schölkopf, D.W. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.J. Simon-Gabriel, and J. Peters. 2016. Modeling Confounding by Half-sibling Regression. *Proceedings of the National Academy of Sciences* (2016).

[29] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy. 2011. Predicting Solar Generation from Weather Forecasts using Machine Learning. In *IEEE International Conference on Smart Grid Communications (SmartGridComm)*.

[30] S. Vergura. 2018. Hypothesis Tests-based Analysis for Anomaly Detection in Photovoltaic Systems in the Absence of Environmental Parameters. *Energies* (2018).

[31] Y. Zhao, Q. Liu, D. Li, D. Kang, Q. Lv, and L. Shang. 2018. Hierarchical Anomaly Detection and Multimodal Classification in Large-Scale Photovoltaic Systems. *IEEE Transactions on Sustainable Energy* (2018).

[32] H. Zhu, L. Lu, J. Yao, S. Dai, and Y. Hu. 2018. Fault Diagnosis Approach for Photovoltaic Arrays Based on Unsupervised Sample Clustering and Probabilistic Neural Network Model. *Solar Energy* (2018).