Identifying High-Significance Latent Physical Anomalies in Solar Energy Systems

Kang Pu*, Yue Zhao†, John Gorman‡, and Philip Court‡
*Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794, USA
†Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794, USA
‡Ecogy Energy, Brooklyn, NY 11217, USA
Emails: {kang.pu, yue.zhao.2}@stonybrook.edu, {john.gorman, philip.court}@ecogyenergy.com

***Abstract*—Data-driven physical anomaly detection for solar energy systems is studied. A fully unsupervised learning approach based on traces of solar generation data and weather data is developed. The idea is that, without needing any anomaly labels, a) predictors that output expected solar generation can be trained based on generation data under normal system operations, and b) a variety of anomalies can be identified based on analyzing the deviations between the actual and predicted solar generation. This paper focuses on identifying physical anomalies that are a) significant and sustained over long periods of time, yet b) "latent", i.e., can be easily missed by asset managers in practice. Two types of predictors — weather-based predictors and cross-inverter predictors — are developed that can provide complementary information in identifying major anomalies. Furthermore, conditional probabilities of the prediction errors are estimated for accurate probabilistic evaluation and statistical interpretations of anomalies. As such, conditional log-error-probabilities are employed as error metrics. Comprehensive experiments are conducted based on rich real-world solar energy data sets that span over 4+ years and across different states. It is demonstrated that the developed method successfully identifies a variety of high-significance physical anomalies that evade asset managers' attention for sustained periods from weeks to years.**

***Index Terms*—Solar energy system, anomaly detection, unsupervised learning, data-driven, long-term anomalies, asset management, operations and maintenance**

I. Introduction

As an important component of our society's solutions to address climate change, a large and growing amount of solar energy has been installed worldwide. While significant progress has been made over the past decades in reducing the manufacturing cost of solar energy systems [1], another important cost factor is the operation and maintenance (O&M) cost. Reducing the O&M cost thus plays a key role in improving the affordability and efficiency of solar energy supplies.

A key aspect of reducing the O&M costs of solar energy systems is to effectively diagnose their underlying issues. This is because, more accurate and timely awareness of physical issues of solar energy systems can greatly improve the efficiency and effectiveness of O&M practices. For example, a) visits of maintenance personnel, a major component of O&M cost, can be scheduled more efficiently to address the detected underlying issues, b) on time maintenance of hardware can greatly prolong their life time and reduce the need of expensive replacement, and c) loss of energy production due to unaddressed system issues can be significantly reduced.

There has been a plethora of recent work on PV system diagnostics that utilize data-driven techniques. [2] develops a diagnostic model for estimating inverter degradation severity by learning from accelerated life testing data of full life cycles of inverters. [3] (among others) develops deep learning systems for thermal image analysis to detect PV cell anomalies. [4] develops a hierarchical method based on minute-level SCADA data, including individual strings' current data, to identify string-level anomalies in PV systems. [5] develops solar farm voltage anomaly detectors based on clustering of μPMU data. [6] develops a PV fault detection toolkit that utilizes self-organizing maps applied to SCADA data. [7] presents a variational recurrent autoencoder with attention for detecting short-term anomalies in time-series data of PV production. Additionally, [8] develops a prognostic model based on supervised learning that utilizes a variety of inverter measurements and alert records for detecting PV failures.

Notably, each of these methods relies on specialized or high-resolution data sources, such as thermal imaging, μPMU measurements, or extensive event logs, which are not universally available across solar energy systems. Furthermore, while some models excel in identifying short-term anomalous events, they often lack the capability to detect gradual, long-term anomalous patterns. These challenges highlight the need for more robust and data-efficient models that can operate across diverse data environments and effectively address short- and long-term fault conditions in PV systems.

In this work, we develop fully data-driven methods for identifying major underlying physical anomalies of solar energy systems that greatly impact their performance. The focus of this paper is on identifying relatively *long-term* anomalies that a) can persist for periods ranging from weeks to years, b) have high significance due to their impact on system performance, and yet c) are often latent and left unattended in practice, as they can still escape asset managers' attention for very long periods. Notably, instead of relying on any specialized or high-resolution data, this work utilizes only *hourly solar generation data* at the inverter level, assisted with publicly available weather data. As such, the developed methods are widely applicable in practice.

This work is supported by DOE EERE SETO under Award DE-EE0009883, by Ecogy Energy, and by the Center for Grid Innovation Development and Deployment (GrIDD) at Stony Brook University.

Specifically, we propose a fully *unsupervised* learning-based approach that trains predictors without relying on any anomaly labels or other human-labeled data. The idea is to leverage *self-supervised* learning to predict one part of the data using other parts and to identify major anomalies based on the prediction errors. In particular, two self-supervised learning problems are formulated: weather-based generation prediction and cross-inverter generation prediction. Notably, these trained predictors provide *complementary* information for identifying significant physical anomalies. Furthermore, we estimate the probabilities of the observed generation using conditional kernel density estimation. We then employ *log-error-probabilities* as the prediction error metrics, based on which a variety of informative curves are plotted for identifying anomalies. We conduct extensive experiments with very comprehensive real-world industry data sets to examine the effectiveness of the developed methods. Notably, a variety of high-significance, long-term physical anomalies have been discovered that were previously unknown to the industry asset managers responsible for the studied systems.

II. Problem Description

For a solar energy site potentially with multiple inverters, we consider that the hourly solar generation data from each inverter is monitored and collected. While other types of data, including event logs, could be available in practice, we recognize that their availability may not be universal or by any means complete. As such, our method does not assume the availability of any system monitoring data other than solar generation. In addition, publicly available weather data are also utilized. *We aim to develop predictors that, as solar generation and weather data are updated over time, continuously produce outputs that effectively aid in identifying major physical anomalies in solar energy systems.* In particular, our focus here is to identify physical anomalies that can persist for *long periods — from weeks to years — before being noticed, if at all, by asset managers in practice*. As such, we are interested in identifying such latent yet significant physical anomalies whose identification would allow very substantial savings in O&M costs in practice.

Importantly, while we introduce methods and techniques that are generally applicable, our study is *entirely grounded on comprehensive real-world data from the industry*. Thus, our method development needs to address all the challenges that working with raw, real-world data entails. As such, we aim to develop a complete pipeline of algorithms (starting from dealing with missing data) that proves to be effective in achieving our objective. While our method does not rely on any anomaly labels whatsoever, thanks to the rich, real-world data sets provided by our industry partners, we are indeed able to cross-reference our discoveries with some available event logs. Moreover, our discoveries are extensively vetted through the profound firsthand experiences of the asset managers responsible for the solar energy systems studied.

III. Methodology

To train detectors of major physical anomalies in a purely unsupervised manner, we develop a self-supervised learning approach that relies solely on production data from inverters, supplemented by publicly available weather data. The high-level idea is that, although physical anomalies can be of various types and can often lack ample and accurate anomaly labels in practice, normal operational data are typically abundantly available. By learning from and predicting expected inverter production under normal operations, our model can effectively detect anomalies when actual production deviates significantly from expected values.

Specifically, two complementary prediction strategies — weather-based generation prediction and cross-inverter generation prediction — are designed to identify anomalies:

- **Weather-based prediction:** For each inverter, we use weather and time-related features to model its expected production under normal operations.
- **Cross-inverter prediction:** At a solar energy site with multiple inverters, we leverage the correlations across inverters within the same site: for each inverter, we predict its production by using data from the other inverters.

We note that, in case there is only one inverter at a solar energy site, the method reduces to using weather-based prediction only. Indeed, weather-based prediction alone can already provide strong signals of abnormality in general. In practice, an issue that we observe is that a weather-based predictor can sometimes be trained to overfit the production data of an inverter, in effect normalizing the abnormality to some extent. Cross-inverter prediction thus offers a complementary approach that independently provides strong signals when a particular inverter behaves abnormally compared to others.

In Figure 1, we depict the complete anomaly detection pipeline developed, which includes data imputation, weather-based and cross-inverter predictions, and an error analysis phase leveraging log error probability plots computed based on conditional kernel density estimation.

A. Data Imputation

Data imputation is essential in our study because missing and outage data can significantly hinder the performance of data-driven unsupervised learning models. Notably, our approach addresses both missing data and outage data (i.e., instances where an inverter's production is zero). We implement a simple yet effective imputation method that leverages the fact that inverters at the same solar energy site typically share identical tilt and azimuth angles. Based on the physical models of solar generation [9], the ideal production values across different inverters at the same site would be approximately proportional to their capacity given their identical orientations. Thus, for an inverter i at a given timestamp t, we impute its missing or outage data by computing a weighted average of

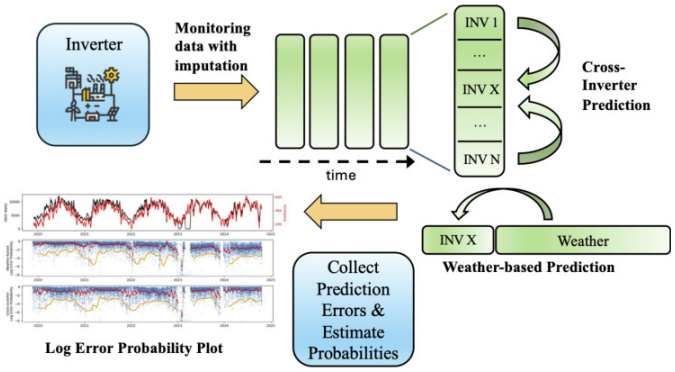

Fig. 1. Overall Pipeline for Identifying Significant Physical Anomalies.

the production values of other inverters on the same site that have non-missing and non-outage data:

$$\tilde{P}_{i,t} = \frac{C_i}{\sum_{j \in S(t)} C_j} \sum_{j \in S(t)} P_{j,t} \tag{1}$$

where $\tilde{P}_{i,t}$ is the imputed production for inverter i at time t, C_i is the capacity connected to inverter i, $S(t)$ denotes the set of inverters with available (non-missing, non-outage) production data at time t, $P_{j,t}$ is the production of inverter j at time t. We will provide the details of how the imputed data are used in training and testing in the remainder of this section.

B. Weather-based Generation Prediction

In this subsection, we develop predictors of inverter production under normal operations based on weather data and time information. For each inverter, we train a predictor based on the normal operation data as well as the imputed missing and outage data (as the expected normal data). In testing, in addition to normal operation data: a) Entries with outage data are utilized without imputation because it is the *ground truth* data (not the imputed one) to which we want the predicted values to compare, and b) Entries with missing data are skipped because there is no ground truth to compare with. Specifically, the input features of the predictors include weather-related variables, such as solar-irradiance-related features — Global Horizontal Irradiance (GHI), Diffuse Horizontal Irradiance (DHI), Direct Normal Irradiance (DNI), and zenith — as well as other weather conditions such as temperature, relative humidity, and wind speed. Time-related features, such as hour and month, are also included.

Input features are also adjusted based on site-specific factors, such as potential snowfall. For example, snow depth is included as a feature for systems in regions with snow (e.g., Washington D.C.) but omitted for regions like Florida where snow is not observed. Such adjustment enables the model to capture location-specific production patterns more effectively. For the training objective, Huber loss is used so that the predictor has some level of robustness to outliers.

$$L_\delta(P_i, \hat{P}_i) = \begin{cases} \frac{1}{2}(P_i - \hat{P}_i)^2 & \text{if } |P_i - \hat{P}_i| \le \delta, \\ \delta \cdot |P_i - \hat{P}_i| - \frac{1}{2}\delta^2 & \text{if } |P_i - \hat{P}_i| > \delta \end{cases} \tag{2}$$

where δ in Huber loss is set to be the standard deviation of the inverter's production.

Significant deviations of actual production $P_{i,t}$ from $\hat{P}_{i,t}$ may signal physical anomalies potentially caused by faults, degradation, etc. To precisely understand the levels of significance of the detected anomalies, we will propose a probabilistic evaluation framework later in Section III-D to estimate the likelihood of the detected anomalies.

C. Cross-Inverter Generation Prediction

In this subsection, we develop cross-inverter predictors to predict the production of an inverter at a solar energy site based on the production of other inverters at the same site. The predictor training is again based on the normal operation data as well as the imputed missing and outage data. Unlike weather-based predictor training where the imputed data are used as training labels, for cross-inverter predictor training, the imputed data are exclusively used as *input features*. The reason for not using it as training labels here is because our imputation method (1) is essentially a simple linear cross-inverter predictor model, and it does not make sense to learn a data-driven model from data that we impute with a known simple model. The use of imputed data as part of the input features is, however, particularly critical for cross-inverter predictor training. This is because, when there are many inverters at the same site, a missing/outage entry from any single inverter could render the entire data vector of that time unusable. Having imputed data would greatly increase the amount of training data that can be utilized. In testing, we again include the imputed data in the predictor input features.

For a given site with multiple inverters, the input features of the predictor for any inverter i include the production data (possibly imputed) from all the other inverters at the same site, along with time-related features. A significant deviation between actual production $P_{i,t}$ of inverter i and the predicted production $\hat{P}_{i,t}$ suggests potential faults or degradation primarily due to issues specific to Inverter i.

D. Conditional Probability and Kernel Density Estimation

With an observed solar generation P_i and its prediction $\hat{P}_i$, we define the prediction error as $e_i = P_i - \hat{P}_i$. As such, positive/negative error implies generation above/below expectation. For our purpose of physical anomaly detection, negative errors are our primary focus. While a variety of error metrics exist (e.g., squared error), we are particularly interested in error metrics that provide rich probabilistic information and statistically sound interpretations of anomalous behaviors. Accordingly, we would like to compute the cumulative distribution function (CDF) of any observed error e_i,

$$F(e_i) = \Pr(E_i \le e_i), \tag{3}$$

where E_i denotes the prediction error as a random variable before it is realized.

While trained predictor models are generally unbiased, however, a "regression to the mean" phenomenon typically exists [10]: Specifically, when the ground truth solar generation P_i is

high/low, the error e_i would have a distribution skewed toward negative/positive values. In other words, the predicted value $\hat{P}_i$ tends to be closer to the unconditional mean than the ground truth P_i. This phenomenon motivates us to refine our goal to further estimate the *conditional* CDF of e_i. Specifically, as solar generation heavily depends on irradiance, for an observed prediction error e_i, we estimate its conditional probability density function (PDF) *conditioned on the irradiance* denoted by irr, $f(e_i|irr)$, and the conditional CDF, $F(e_i|irr)$. By estimating the conditional probabilities, we explicitly capture the regression to the mean phenomenon and compute the likelihood of errors based on the skewed conditional probability distributions, which provide more accurate information than the unbiased unconditional probability distribution.

In practice, the exact error probability distribution is unknown and needs to be estimated empirically. We estimate these conditional probability distributions using non-parametric kernel density estimation (KDE) [11] with Gaussian kernel, applied across prediction errors from all the inverters at the same site:

$$\hat{f}(e_i|irr) = \frac{1}{nh\sigma}\frac{1}{\sqrt{2\pi}}\sum_{e_{irr,j}\in E(irr)} \exp\left(\frac{-(e_i - e_{irr,j})^2}{2h^2\sigma^2}\right)$$

$$E(irr) = \{e_{i,t}|irr - \mu <= irr(t) < irr + \mu\} \tag{4}$$

where $E(irr)$ represents all prediction errors obtained under irradiance within μ units of the irradiance of interest, irr. Here, σ is the standard deviation of the errors in $E(irr)$, and h is the bandwidth, determined using Scott's Rule [12],

$$h = n^{-\frac{1}{d+4}} \tag{5}$$

where $n = |E(irr)|$ and d is the dimension of the data. In particular, the CDF $F(e_i|irr) = \Pr(E_i \leq e_i|irr)$, capturing the likelihood of its occurrence. A low $F(e_{i,t}|irr)$ would quantify the rarity (and thus abnormality) of solar generation for Inverter i at time t.

Last but not least, we use *log-likelihood*, $\log(F(e_{i,t}|irr))$, as the error metric. For example, an extremely rare error would lead to a close to zero probability and hence a significantly negative log-likelihood. A major advantage of using log-likelihood is the following: Collectively evaluating multiple errors naturally corresponds to *multiplying* their error probabilities, which is equivalent to *summing* their log-likelihood. This implies that computing a *moving average* of the log-likelihood curve $\log(F(e_{i,t}|irr)), \forall t$ — a common method for smoothing out noise and capturing major trends — naturally provides a statistically meaningful metric.

IV. Experiment results

A. Dataset Description

We conduct extensive experiments on datasets collected from four solar energy system sites: Site A is located in Washington, D.C. with 8 inverters. Sites B, C and D are located in Florida with 7, 5, and 7 inverters, respectively. Across all these sites, the past 4+ years of hourly solar generation data of all inverters are collected [13]. Additionally, hourly weather data (cf. Section III-B) at all these sites are obtained from Solcast [14].

B. Training and Testing Procedures

Both the weather-based and cross-inverter predictors for all the inverters are trained and tested in a way similar to cross-validation. We partition the data into 5 folds. For each fold as the test set, we use the other 4 folds for training and validation. The testing is accordingly performed on *all* the folds.

For weather-based prediction, we employ a 4-layer Multilayer perceptron (MLP) with a residual block embedded between the first and last fully connected layers. Batch normalization and dropout are applied to mitigate overfitting. For cross-inverter prediction, we employ XGBoost [15] as the predictor model with hyperparameters optimized via a grid search. To compute the prediction error metrics, we compute $F(e_i|irr) = \Pr(E_i \leq e_i|irr)$ where Global Horizontal Irradiance (GHI) is used as irr. The parameter μ in (4) is set to 5. We then take the log of the probabilities.

C. Discovering High-Significance Physical Anomalies from Observing Log-Error-Probabilities

By observing the log probabilities of prediction errors, a variety of significant and informative physical anomalies of solar energy systems have been identified. Importantly, while these anomalies are a) statistically significant according to the log-probability metrics, they b) remain unnoticed by the asset managers in practice for long periods of times (from weeks to years). We summarize four major types of identified physical anomalies:

- Seasonal performance anomalies, e.g., abnormally low performance during winter or summer times.
- Performance abnormalities that precede system outages.
- Even without major failures, sustained low performance of the solar system for long periods of time.
- Long-term performance degradation, including increasingly frequent abnormally-low-performance instances.

Next, we present representative examples that demonstrate the above. (Other examples have to be left out due to space limits.) For clear and informative visualizations, we display a set of 3 plots for each inverter. The first plot shows the 10-day *causal* moving averages of the solar generation and GHI, allowing a direct comparison of generation with irradiance over time. We note that, the moving averages employed in this work are all causal: at time t, an average is computed over *a window of time right before* t. This ensures that such moving averages are feasible to compute in practice in real time. The second plot presents error metrics from the weather-based predictor, including the raw hourly log-error-probabilities (cf. the blue dots), the 30-day causal moving average of the hourly log-error-probabilities (cf. the red curve), and the 30-day causal moving average of the 4% quantile log-error-probabilities (cf. the orange curve). Specifically, at each time t, its 4% quantile log-error-probability (before moving average) is collected from the 30-day window right before t.

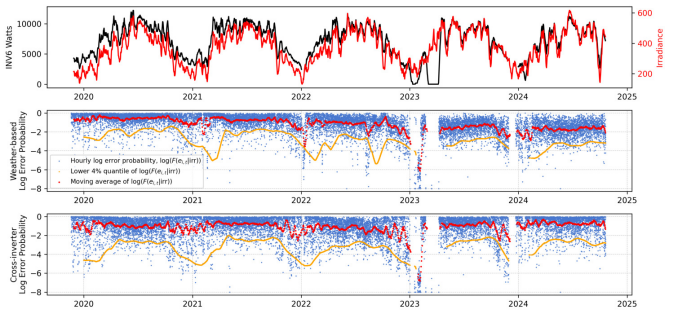

Fig. 2. Site A, inverter 6 in Washington, D.C.

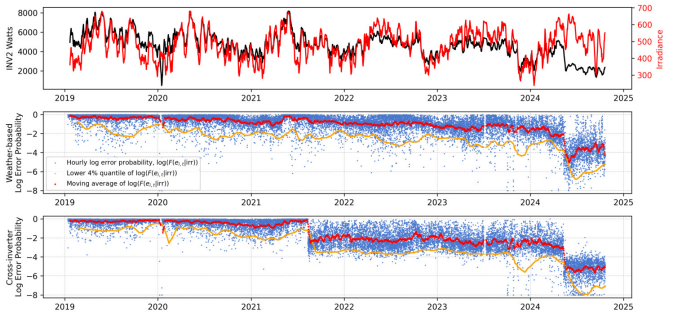

Fig. 3. Site B, inverter 2 in Florida.

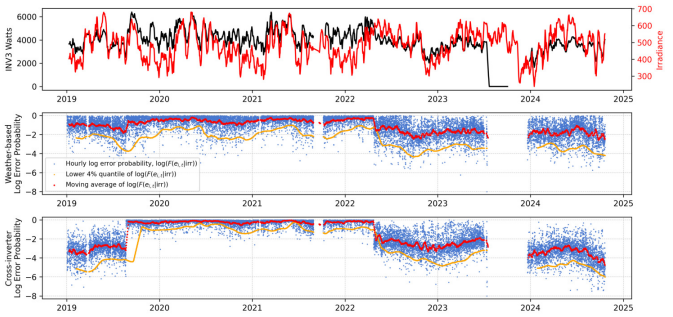

Fig. 4. Site C, inverter 3 in Florida.

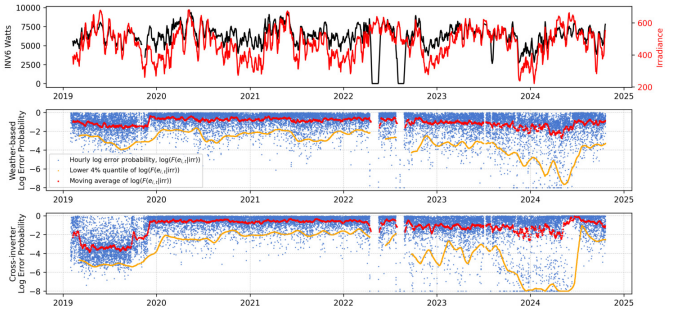

Fig. 5. Site D, inverter 6 in Florida

Notably, while the red curve keeps track of the likelihood of all the errors within a 30-day window, the orange curve keeps track of the occurrences of the more extreme errors over time. The third plot presents the same error metrics as the second plot but for the cross-inverter predictor, offering a comparative view of anomaly detection across the two different prediction methods. We note that there are gaps and valleys in these plots due to missing data or inverter outages.

The plots from 2020 to 2024 for Inverter 6 at Site A, Washington, D.C. are plotted in Figure 2. From both the second and third plots, we see that a) as indicated by the blue and orange curves, the system struggles with performance during *winter* months, and (b) as indicated by the blue and red curves, clear downward trends appear *leading up to major outages* (the blank periods) around early 2023 (and late 2023). We note that, these anomalous behaviors are not sporadic anomalies (due to, e.g., malfunctioning communication systems that we indeed sometimes observe) or noise; rather, they are significant and sustained *physical* anomalies of the solar energy systems.

The plots from 2019 to 2024 for Inverter 2 at Site B, Florida, are plotted in Figure 3. From the second plot, especially the blue dots, a) clear seasonal performance abnormalities during the *summer* months are observed, and (b) a clear sustained *long-term performance degradation* is observed (cf. the red curve), and c) a major performance drop is observed since mid-2024. The third plot does not indicate the above issues a) and b). However, it clearly indicates another *severe and sustained performance drop* since mid-2021 that is, however, not evident in the second plot.

The plots from 2019 to 2024 for Inverter 3 at Site C, Florida, are plotted in Figure 4. Multiple periods of sustained low performance are observed from both the second and third plots, notably throughout the majority of 2019 and ever since the spring of 2022. While the weather-based prediction errors do indicate some level of abnormality, the cross-inverter prediction errors confirmed such abnormality much more decisively. We note that, prior to the discoveries of this work, the low-performance issues since the spring of 2022 still have not been noticed in practice. The plots we generated thus can greatly reduce the time to discover such low-performance issues and save significant lost production from the asset.

The plots from 2019 to 2024 for Inverter 6 at Site D, Florida, are shown in Figure 5. From both the second and third plots, we observe: (a) a major performance anomaly in 2019, and (b) recurrent seasonal anomalies during the *summer* months are again captured, particularly in the second plot, and (c) from late 2022 onward, the system begins to show increasingly frequent signs of poor performance, as indicated by persistently deteriorating log error probabilities in both plots. This phenomenon is particularly pronounced throughout 2023, and is further corroborated by the raw data which reveal frequent abnormal short outage periods. We note that, in both Site C, inverter 3 and Site D, inverter 6, the major performance issues observed in 2019 were indeed addressed by maintenance events that successfully restored system performance by late 2019.

Notably, our observations highlight the important fact that, while the weather-based (second plot) and cross-inverter (third plot) predictions can often confirm each other's discoveries, they can also provide key *complementary* information, each with its own unique strengths, in identifying different types of anomalies.

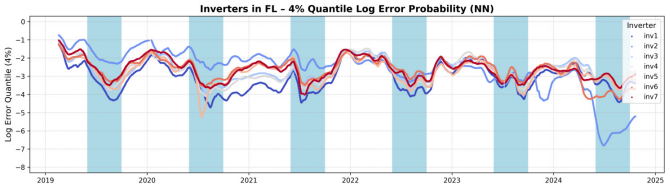

Fig. 6. Site B in Florida: all inverters experience similar abnormality during summer months.

For example, (a) In Figure 2, seasonal anomalies during the winter are captured by both the weather-based and cross-inverter log error probability plots, indicating that this inverter experiences more pronounced issues during winter months compared to its peer inverters at the site. (b) In Figure 3, the weather-based log error probability plot reveals seasonal anomalies during the summer. However, these anomalies are not detected by the cross-inverter log probability plot because *all inverters* at this site exhibit *similar seasonal issues* during the summer months. Indeed, we plot the log error probabilities of the weather-based predictors for all the inverters at this site in Figure 6, and observe their abnormality patterns occurring in unison during the summer months. As a result, the cross-inverter predictor interprets these patterns as normal. In such scenarios, weather-based predictor becomes essential, as it can identify site-wide seasonal problems that the cross-inverter predictor might overlook. On the other hand, the cross-inverter log error probability plot shows a clear performance drop beginning in mid-2021—an issue not apparent in the weather-based plot. This discrepancy is likely due to overfitting in the weather-based predictor, which may normalize abnormal conditions. (c) In Figure 4, performance issues are more effectively identified by the cross-inverter predictor. In this case, nearly half of the production data used by the weather-based predictor are affected by abnormal conditions, causing it to overfit to abnormality and diminishing its ability to detect anomalies. The cross-inverter predictor, by comparing relative performance across inverters, is more effective at revealing these issues.

Taken together, these examples underscore the complementary strengths of the weather-based and cross-inverter predictors in capturing a broad range of anomalies. Although each method has its limitations, one often performs well when the other falls short. By combining insights from both predictors, we can form a more comprehensive and reliable anomaly detection framework. This dual-perspective approach enhances the robustness of fault detection, helping to uncover subtle or prolonged performance issues that can otherwise go unnoticed by asset managers for weeks, months, or even years.

V. Conclusion

In this paper, we developed a fully unsupervised learning approach for identifying significant and sustained physical anomalies in solar energy systems. We employed two types of predictors of expected solar generation — weather-based predictors and cross-inverter predictors — that provide complementary information. Furthermore, we estimate the conditional prediction error probabilities, conditioned on solar irradiance, using kernel density estimation. Log error probabilities are employed as the error metrics, offering sound statistical interpretations of the prediction errors. We conduct extensive experiments of the developed methods based on rich real-world datasets. A variety of significant and informative physical anomalies have been successfully identified that a) have sustained for long periods of time yet b) have evaded asset managers' attention without being addressed. Further classification of anomalies with more detailed information — such as their points of origin (e.g., inverter, PV module, combiner box) and underlying causes (e.g., overheating, MPPT fault, shading) — is left for future work.

Acknowledgment

The authors thank Arun Veeramany, Meghana Ramesh, and Anastasios (Tassos) Golnas for their insightful discussions.

References

[1] NREL, "Documenting a decade of cost declines for PV systems," 2021.
[2] S. Karakaya, M. Yildirim, S. Zhao, F. Qiu, J. D. Flicker, B. Peters, and Z. Wang, "Leveraging high-fidelity sensor data for inverter diagnostics: A data-driven model using high-temperature accelerated life testing data," in *IEEE 50th PV Specialists Conf. (PVSC)*, 2023, pp. 1–7.
[3] R. Pierdicca, M. Paolanti, A. Felicetti, F. Piccinini, and P. Zingaretti, "Automatic faults detection of photovoltaic farms: solAIr, a deep learning-based system for thermal images," *Energies*, vol. 13, no. 24, 2020.
[4] Y. Zhao, Q. Liu, D. Li, D. Kang, Q. Lv, and L. Shang, "Hierarchical anomaly detection and multimodal classification in large-scale photovoltaic systems," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 3, pp. 1351–1361, 2019.
[5] M. Dey, S. P. Rana, C. V. Simmons, and S. Dudley, "Solar farm voltage anomaly detection using high-resolution μpmu data-driven unsupervised machine learning," *Applied Energy*, vol. 303, p. 117656, 2021.
[6] A. Betti, M. Tucci, E. Crisostomi, A. Piazzi, S. Barmada, and D. Thomopulos, "Fault prediction and early-detection in large pv power plants based on self-organizing maps," *Sensors*, vol. 21, no. 5, 2021.
[7] J. Pereira and M. Silveira, "Unsupervised anomaly detection in energy time series data using variational recurrent autoencoders with attention," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 1275–1282.
[8] L. Liu, Y. Luo, Z. Wang, F. Qiu, S. Zhao, M. Yildirim, and R. Roychowdhury, "Deep learning-based failure prognostic model for pv inverter using field measurements," *IEEE Transactions on Sustainable Energy*, vol. 15, no. 4, pp. 2789–2802, 2024.
[9] K. Pu and Y. Zhao, "An unsupervised similarity-based method for estimating behind-the-meter solar generation," in *2023 IEEE PES Innovative Smart Grid Technologies Conference (ISGT)*, 2023, pp. 1–5.
[10] K. Belitz and P. Stackelberg, "Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models," *Environmental Modelling & Software*, vol. 139, p. 105006, 2021.
[11] G. R. Terrell and D. W. Scott, "Variable kernel density estimation," *The Annals of Statistics*, pp. 1236–1265, 1992.
[12] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.
[13] J. W. Gorman, "Solarnetwork.net an open source platform for distributed generation," Master's thesis, University of Auckland, 2009.
[14] Solcast, "Global solar irradiance data and PV system power output data," 2019.
[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.