# Experience with a Physics-Based Anomaly Detection Algorithm to Detect DC Faults at Utility PV Plants

Scott Sheppard[1], Steven Koskey[1], Corson Teasley[1], Christopher Perullo[1], Daniel Fregosi[2], Wayne Li[2]

[1]Turbine Logic, Atlanta, GA, 30332, USA
[2]Electric Power Research Institute, Charlotte, NC, 28262, USA

**Abstract — String, combiner, and tracker faults comprise a significant reduction in PV plant capacity. A method is presented to detect these faults using data typically monitored at utility-scale plants. The detection algorithm has been validated using aerial infrared data and the accuracy is presented across a range of sites, conditions, and scenarios, including automated versus manual model configuration and satellite versus ground-based weather data.**

## I. Introduction

PV plants are susceptible to electrical and mechanical faults that gradually reduce the capacity of a plant over time, if left unfixed. It has been observed that 2% of system capacity, on average, is lost due to combiner, string, and tracker faults [1]. Another study found that system capacity is reduced at a rate of 0.1% per year, on average, due to recoverable faults [2]. The value of this potentially recoverable energy is evident as many owners pay for aerial infrared imaging at an annual frequency or higher to detect and locate faults to be repaired.

While these faults are relatively easy to fix, detection is difficult due to the large geographic spread of plants. Manual testing of each string is infeasible. Aerial infrared imaging, while expensive, is commonly used.

In this work, a method to detect string and tracker faults using data points typically monitored through SCADA systems was demonstrated. The value of this approach is that it enables faster repairs to increase energy production as faults can be detected in near-real time, compared to annually with aerial infrared. Furthermore, the cost is low as the method uses inverter-level monitoring data typically available through SCADA systems.

A unique aspect of this work is that fault detection has been applied to 9 utility scale plants and the accuracy has been validated using aerial infrared data. This paper overviews the algorithm and presents the validation results across a variety of conditions and scenarios including satellite versus ground-based weather data, automated versus manual model configuration, and across seasonal and weather differences.

## II. Fault Detection Algorithm

### *A. Physics-Based Model*

The modeling approach used in this work uses a combination of intelligent data filtering, machine learning (ML)-based site setup, physics-based models, and ML-based anomaly detection to model PV site performance and determine anomalous combiner box behavior. The model uses Sandia National Lab's pylib-python package to perform physics-based calculations of expected plant output based on knowledge of PV plant architecture and hardware and using real-time weather data inputs [3]. The algorithm compares combiner box performance to both modeled performance and peer combiners within a given inverter. These outputs are featurized and ML techniques are used to identify when a particular combiner box signal is operating anomalously. Greater details of this algorithm are provided in [4].

Data filtering and cleanup is a necessary step before the models can be executed. Filterers are used to remove data points with known bad values, low plane of array irradiance measurements, and low solar elevation angle. Additional filters are applied to specific data channels based on values and point-to-point changes that are physically unreasonable [5].

In addition to these filters, a cloud detection routine is executed using the irradiance data from the site. This routine uses a seven-day rolling window and statistical filtering to construct a clear sky irradiance signal using only site-measured data. The irradiance data is then compared to that calculated clear sky signal to determine the presence of clouds, and to filter out any data where clouds were present. Additional details of this routine can be found in [4].

Originally, the physics-based models were constructed after detailed review of site drawings during which information was manually extracted from the drawings to use in the models. This information included PV module and inverter hardware, number of modules per string, number of strings per combiner box, and number of combiner boxes per inverter. At 20+ MW plants, these last two architectural details often varied throughout the plant, resulting in inverter-level specifications throughout each model. To reduce the amount of time required to construct each individual model, DuraMAT's PV-Pro tool was used to estimate modules per string and strings per combiner box from each combiner box current signal [6]. Deployment of this package results in estimates of the quantities on a *per signal* basis, rather than the *per inverter* basis that occurs at a plant. Machine learning techniques were used to estimate a single value of the number of modules per string and a few unique values of strings per combiner for the

site. Additional details on this automated model building application are described in [7].

### *B. Validation*

The physics-based modeling approach has been validated by comparing anomaly detection results to faults found via aerial infrared scans. First, models were constructed for nine PV plants distributed across the United States from two power generation utilities. Each site model was then executed for a time period that was concurrent with the performance of an aerial scan. One of the partner utilities was able to provide two scans for two sites, resulting in a total of 11 validations datasets by which to evaluate the models. Anomaly detection results were collected for the same day as each aerial scan, and the aerial scans were treated as "ground truth" for the evaluation of true and false positive rates. Overall fault detection results for the algorithm are summarized in Table I.

TABLE I
OVERALL DETECTION SUMMARY

| Sensitivity | TPR | FPR | F1-score |
|---|---|---|---|
| High | 0.65 | 0.20 | 0.57 |
| Medium | 0.5 | 0.10 | 0.54 |
| Low | 0.37 | 0.07 | 0.46 |

The aerial scans performed are capable of identifying a variety of module level faults, string outages, and tracker faults. It was assumed that, using combiner box level data, fault magnitudes at the string outage level, or more severe, were required for the fault to be detected. The alarms raised by the detection algorithm were also analyzed by fault type; these detection metrics are shown in Table II.

TABLE II
DETECTION SUMMARY BY FAULT TYPE

| Fault Classification | Sensitivity | TPR |
|---|---|---|
| String Outage | High | 0.66 |
| | Medium | 0.50 |
| | Low | 0.37 |
| Tracker Fault | High | 0.53 |
| | Medium | 0.38 |
| | Low | 0.35 |

## III. IMPLEMENTATION

### *A. Automating Model Setup*

The models were originally constructed using information extracted from site drawings by manual review. This process is highly time consuming, and, to reduce future efforts was automated to increase ease of deployment of the algorithm. The automated model building does not perfectly match what is known about the site from the as-built drawings, and as a result there is a tradeoff between model setup time and detection capabilities. The data used for the automated model setup will also impact detection capabilities. Fig. 1 shows how the F1 score is impacted by the choice of training data when constructing the site model automatically, for six of the 11 validation sets. Though the detection impacts are site specific, automated model building with summer data often led to significant reduction in detection capabilities compared to the manual setup and winter data. A significant driver of this difference in impact was found to be the greater presence of both curtailment and inverter clipping in the summer data.

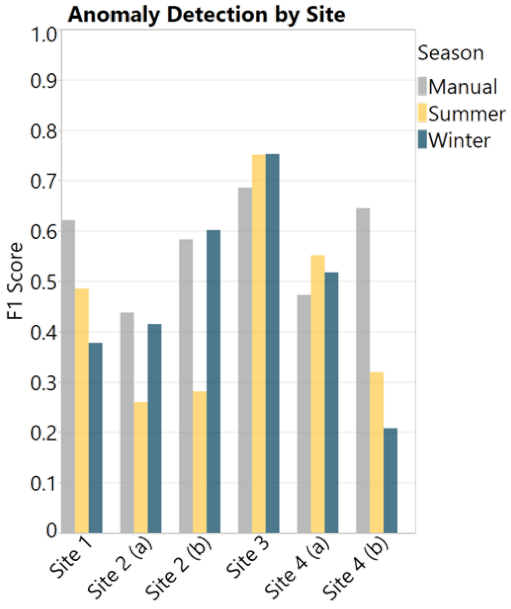


Fig. 1. Variation in F1 score at different sites comparing manual model building to automatic model building with data from different times of year.

### *B. Site-to-site Variation*

While Table I shows overall detection metrics for the algorithm, there is considerable variability between the validation sets. Fig. 2 shows the relationship between TPR and FPR for each validation dataset. Based on conversations with the utility partners, an FPR greater than 0.1 was considered highly undesirable, so the sensitivity of the algorithm has been adjusted to cover this range.

There are several reasons for this site-to-site variability. First, differences in hardware and architecture all have an impact on fault detectability. For sites with a higher string count per combiner box, the impact of a string outage on combiner box output will be lower. Also, sites with large numbers of faults make it more challenging for the algorithm to determine which hardware is faulted when a large portion of hardware appears faulted. Weather quality during the validation period also has very clear impacts on fault detectability.

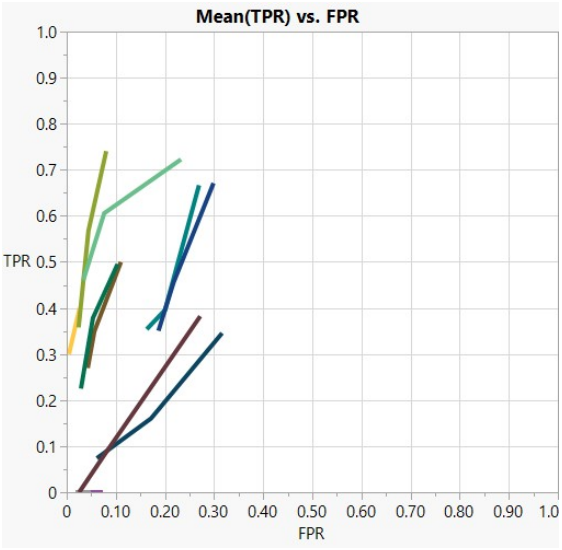


Fig. 2. TPR and FPR for each of the validation datasets used during development of the anomaly detection algorithm.

### C. Weather Variations

For three of the sites, faults of varying magnitudes were intentionally introduced at combiner-level hardware to determine the threshold of fault detectability. Two of these sites experienced prolonged cloudiness through the validation period, severely impacting the algorithm's ability to detect the introduced faults. These results were analyzed to better understand how weather quality impacted the likelihood of detection of these faults, and the results are shown in Fig. 3. The impacts of weather are significant; probability of detection of a fault magnitude of 5% can be reduced from approximately 90% to less than 10% by weather alone.

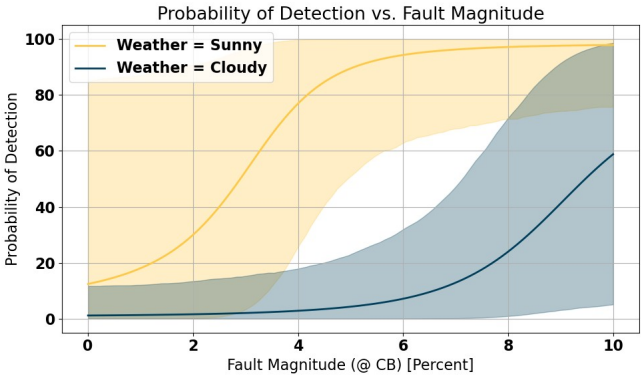


Fig. 3. Impact of weather quality on detection ability over three sites with intentionally introduced faults.

### D. Alternative Weather Sources

When constructing a new site, there is always the question of what sensor suites to deploy. There is a point of interest in understanding whether onsite weather sensors are necessary, or if using a third-party source is sufficient. To aid in understanding the impacts of this choice, satellite-based weather was acquired for each of the validation datasets and used in the anomaly detection algorithm. These results are summarized in Fig. 4. While the external weather sources did not have a major impact on false positive rates, each satellite source did see a reduction in true positive rates. This result is driven by the relatively lower spatial resolution that satellite sources will incur compared to a sensor located at the site.

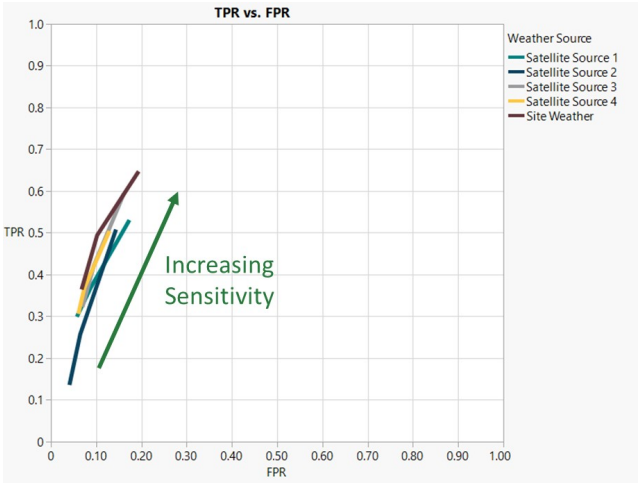


Fig. 4. Impacts of third-party satellite weather data sources to site-based weather data.

## ACKNOWLEDGEMENTS

This material is based upon work supported by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Solar Energy Technologies Office under Award Number DE-EE-0008976.

## REFERENCES

[1] kWh Analytics, "Solar Risk Assessment: 2022 – Quantitative Insights from the Industry Experts," San Francisco, CA: kWh Analytics.

[2] kWh Analytics, "Solar Risk Assessment: 2019 – Quantitative Insights from the Industry Experts," San Francisco, CA: kWh Analytics.

[3] William F. Holmgren, Clifford W. Hansen, and Mark A. Mikofski. "pvlib python: a python package for modeling solar energy systems." Journal of Open Source Software, 3(29), 884, (2018). https://doi.org/10.21105/joss.00884.

[4] S. Sheppard, T. Cook, D. Fregosi, C. Perullo and M. Bolen, "Field Experience Detecting PV Underperformance in Real Time Using Existing Instrumentation," 2022 IEEE 49th Photovoltaics Specialists Conference (PVSC), Philadelphia, PA, USA, 2022, pp. 0307-0313, doi: 10.1109/PVSC48317.2022.9938522.

[5] "Photovoltaic systems performance – Part 3: Energy evaluation method," IEC Technical Specification, IEC TS 61724-3, ISBN 978-2-8322-3531-7.

[6] DuraMAT, Berkeley, CA. 2022. *PV Production Tools (PV-Pro)*, ver 0.0.4.

[7] S. Koskey, S. Sheppard, C. Teasley, C. Perullo, J. Kee, D. Fregosi, W. Li, "Photovoltaic Site Architecture Estimation Using Performance Data," 2023 IEEE 50th Photovoltaic Specialists Conference (PVSC), San Juan, PR, USA, 2023, pp. 1-6, doi: 10.1109/PVSC48320.2023.10359804.