Photovoltaic Site Architecture Estimation Using Performance Data

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

[1] Turbine Logic, Atlanta, GA, 30308, USA

[2] Electric Power Research Institute, Charlotte, NC, 28262, USA

Abstract — **Most photovoltaic power generation sites schedule maintenance as a result of physical inspections and observations. For example, a site may use aerial infrared imaging to determine the fault status of individual combiner boxes and/or strings. However, the costs to perform aerial scans result in infrequent, typically annual, application. As a result, DC faults can often go unnoticed for months at a time. While this repetitive, expensive task is attractive for automation, the limited granularity of modern sensor suites makes it difficult. The authors' work has enabled continuous, real-time PV anomaly detection using existing, installed sensor suites. This relies on a detailed knowledge of the site layout to correctly predict expected performance. Previously this information was manually codified using available site drawings for each site. However, the manual review and codification of metadata is time-consuming, increasing the investment required for an M&D center to implement the code. This difficulty is exacerbated by the competitiveness of the PV market, which has led to leaner O&M investments. This work presents a new method to estimate the site architecture using performance data and a fraction of the metadata. The setup speed is accelerated by more than a factor of 15, while achieving similar anomaly detection quality to the previous work using manually codified site layouts.**

I. Introduction

Monitoring and diagnostics (M&D) play an increasingly important role supporting both business and engineering decisions across the power industry. Utility M&D centers leverage the advances associated with modern digitalization – affordable data storage and computational power, advanced analytics, and modern sensor suites – to characterize and improve power plant operations. While the trend towards more data-informed decision-making is universal, the specific goals and implementations vary significantly across power generation methods. In the case of photovoltaic (PV) power plants, M&D centers generally observe electrical properties and weather conditions. In essence, the weather conditions, particularly irradiance measurements, give the operators an idea of how the plant *should be* operating, and the electrical properties give them an idea of how it *is* operating.

Operators at PV plants perform both corrective and preventative maintenance. The former corrects issues which have already occurred, and the latter corrects issues which are expected to occur. Large and/or widespread problems often receive corrective maintenance priority as they impact a plant's power production capabilities more significantly. Preventative maintenance needs to balance the cost of maintenance with the risk of experiencing a future outage [1]. In this case, M&D centers' data can be used to train risk models that inform preventative maintenance plans.

One significant challenge for M&D centers to address is the scale and remoteness of PV plants. While M&D centers are uniquely situated to handle large amounts of data, due to the costs of new sensors, the need to maintain sensors, and the need to store the data collected by each sensor, it is often cost prohibitive to put sensors on each solar panel or each string of panels. When compared to traditional power generating assets, such as gas turbines, PV plants require one to two orders of magnitude *more* sensors, depending on the size and layout of the plant. Many PV plants collect and record data at the inverter and the combiner box level. Each site analyzed in this paper has between 250 and 2400 combiner boxes, approximately 16 strings per combiner, and approximately 20 modules per string. Instrumenting at the module level would increase the required number of sensors by 320-fold. Each sensor collects its measurements as a function of time, often sending minute-by-minute data back to the M&D center.

Standard sensor suites make it trivial to detect large problems, like inverter outages. Smaller scale sub-inverter issues in the DC collector field, such as string outages, are more difficult to detect even though they can account for approximately 2% of losses from a plant's nameplate capacity [2]. Recent work has attempted to automatically detect and localize these faults using available plant data [3]. Accurately identifying a fault's location within the site is vital because it enables the maintenance team to spend less time finding the faults and more time fixing them. Consider a case in which an anomaly detection routine detects a string outage within an inverter. The maintenance team can find the fault much more efficiently if they know which combiner is faulted, rather than inspecting all combiner boxes within the inverter [5].

If automated, a data-driven anomaly detection method can operate on real-time data to perform continuous fault monitoring. This differs from other fault diagnosis methods, which often involve manual panel inspection or aerial infrared inspections performed by flying an aircraft or drone above the site – both of which are too expensive to perform continuously. Aerial fault status measurements are generally performed as part of annual maintenance routines [1]. As a result, faults can go unnoticed for months on end. Continuous anomaly detection enables operators to prepare maintenance schedules using more up-to-date information.

While previous anomaly detection work has shown promising results in terms of accuracy and usability [3], it still requires a significant time investment to configure a model. The competitiveness of the PV market has resulted in lean O&M budgets, making significant time investments to configure new software more difficult. In general, model frameworks are created first, then they are calibrated to specific sites. The variation between sites introduces a tedious problem for the modeling team: configuration. One fundamental variation that all PV modelers have to consider is site layout – that is, *how many combiners are there in each inverter, how many strings are there in each combiner, and how many modules are there in each string*. Traditionally, modelers manually review the site as-built drawings. This is a laborious, time-consuming task that leaves room for user errors.

Third-party modelers, who often interact with data from various utilities, face another hurdle in the site configuration process: they need to translate the tags from the utility's naming convention to one that works across all of the datasets being modeled. Tags are names by which operators and modelers can reference data streams. The tag translation step is needed so third-party code can pull needed data from the source.

The present work introduces two new algorithms to automatically estimate the PV site layout, enabling automated configuration for PV modeling pipelines. The first method automatically translates the tag names used by a utility using a short list of patterns provided by the user. The translation enables modelers to easily access needed data streams regardless of any utility's tag-naming schema – third-party models that interact with data from various utilities benefit most from this functionality. The second method codifies the hierarchical architecture of a site down to the combiner box using the available tags, then estimates the number of strings and the number of modules per string for each combiner box using performance data. The algorithms write code-readable configuration files that can be used in arbitrary PV modeling pipelines.

II. Methodology

The anomaly detection method applied to this work and described in [3] requires an engineer to manually provide (i) tag mappings for each data stream, (ii) a detailed codification of the site architecture, and (iii) individual hardware component specifications. In the authors' experience from doing this for several sites, manual configuration takes at least eight human hours, depending on the size of the plant. The current work simplifies the configuration process by automatically generating the first two items from the list given a reduced set of metadata. The automatic configuration takes approximately 30 minutes and has three high-level steps.

First, the tags are mapped from the utility naming convention to a user-defined convention. Electrical and weather readings are then filtered to only retain clean maximum power point (MPP) data, which is compared with the modules' datasheet values to estimate the combiner-level layout information. Finally, the tag mapping and estimated per-combiner layout are combined into the necessary site configuration files for a fault detection algorithm.

A. Tag Mapping Standardization

Naming conventions vary greatly from utility to utility. For example, imagine a combiner box at hypothetical Site ABC. It may have the following tag: *ABC-1.D.4-A*. From this format, an analyst can infer that it contains amperage data coming from Site ABC, array 1, inverter D, combiner box 4. But an analyst from another utility, or a third-party modeler, can only guess what it means.

Due to the variation between tag naming conventions, it is helpful for the user to translate the tags to their own convention for each of the relevant information types. While relevant information types vary depending on the application, the anomaly detection method discussed in [3] is used to exemplify the process. It relies on the following information types (each as a function of time):

- Inverter current, power, and voltage
- Combiner box current
- Ambient temperature
- Plane of array irradiance
- Wind speed
- Tracker angles (if using solar trackers)

Previously an engineer had to manually map each tag from the site to its standard counterpart. This took an author approximately an hour per site; the specific duration scales significantly with the site's size. The new algorithm relies on user-input encodings to produce a one-to-one mapping between formats.

Table 1 shows a few examples of different tag names for various data typically collected at PV plants. As PV plants are highly modular, it is generally easy to identify regular tag name formats that have been used to identify similar types of information, such as inverter voltages. The new method enables an engineer to list out a much shorter dictionary of tag name formats, such as those in the "Pattern" column of Table 1. So long as the site tags have a regular pattern, that pattern can be used to efficiently translate the tags.

After the user provides a pattern for each information type, the algorithm uses regular expressions to create a mapping dictionary. It first determines which information type a tag contains and then searches each site tag for the components defined in brackets in the pattern dictionary – for example, array, inverter, etc. The algorithm can translate to any user-defined format, using pattern components as specified by the user. For this work, the authors extracted the full inverter ID (i.e., *{Array}.{Inverter}*), the combiner box ID, and the met-station ID for each data stream. After each component is extracted, it is composed into a user-defined standard format. Using this method, site data tags can be translated to the same format regardless of the source format.

TABLE 1
EXAMPLE TAG NAMES AND THEIR ARCHETYPICAL PATTERN

Information Type	Example Site Tags	Pattern
Inverter Voltage	ABC-1.D-V	{Site}-{Array}.{Inv.}-V
Inverter Current	ABC-1.D-A	{Site}-{Array}.{Inv.}-A
Combiner Current	ABC-1.D.1-A	{Site}-{Array}.{Inv.}.{Comb.}-A
Combiner Current	ABC-1.D.2-A	{Site}-{Array}.{Inv.}.{Comb.}-A
Ambient Temperature	ABC-1.met1-T	{Site}-{Array}.met{Station ID}-T

B. *Data Cleaning*

Once the tags have been standardized, the data is cleaned and the meteorological readings are used to determine when a panel can be expected to operate at its MPP. The data filtering described in [3] has been used as a first step to reduce noise in the dataset. Points with low irradiance and low solar elevation angles are removed. Users can provide a list of known bad values (such as historian timeout values) which will also be removed. The cloud detection strategy from [3] was used to remove off-MPP data from each dataset. Additional data quality filters as described in [4] have been used to filter for non-physical values and point-to-point changes in the data streams.

As the final estimation compares measured data to spec sheet data, one additional filtering step has been applied. The PVPRO package has a built-in function to estimate when PV current and voltage data are at the hardware's maximum power point. This functionality is used to filter the dataset to only those points in time.

C. *Codification of Site Architecture*

PV sites have hierarchical structures similar to Fig. 1. Working inwards, the components are array, inverter, combiner box, string, and module. This hierarchy is reflected in the standard tag mapping – that is, each combiner box tag contains information for the array and inverter under which it is housed, as indicated in Table 1. However, many models and analyses, such as the routine in [3], also rely on the number of strings per combiner box and the number of modules per string. Neither of these numbers is generally encoded in the combiner box tag names. As a result, the automated site setup routine infers these string and panel counts from measured performance data. Given the panel specifications, the current at max power (i_{MP}) and the voltage at max power (V_{MP}) can be estimated for each combiner box using PVPRO [6]. Once i_{MP} and V_{MP} have been calculated, the ratio is taken between the combiner box-level estimate and the single module datasheet values, as suggested in PVPRO's documentation. The ratios are as shown in (1) and (2). The current ratios estimate the number of strings per combiner box and the voltage ratios estimate the number of modules per string.

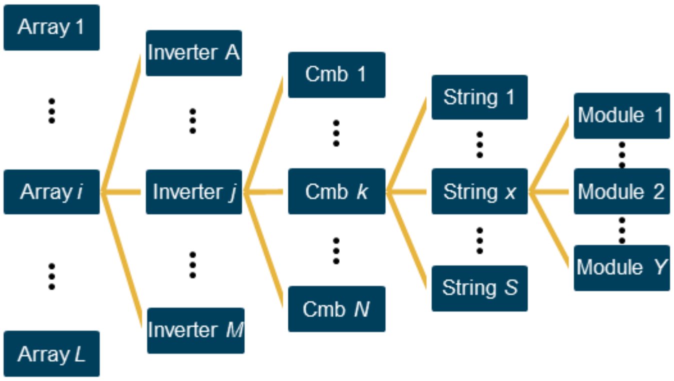

Fig. 1. The hierarchical structure of a PV site can be represented using arrays, inverters, combiner boxes, strings, and modules.

$$n_{strings} = \frac{i_{MP,estimate}}{i_{MP,datasheet}} \tag{1}$$

$$n_{modules} = \frac{V_{MP,estimate}}{V_{MP,datasheet}} \tag{2}$$

The PVPRO method results in a unique estimation of the string count for each combiner box and the number of modules per string for that combiner box. But the precise estimates are not integers, whereas the real sites have integer string and combiner counts. Furthermore, the sites reviewed in this work had a few distinct string counts, and the number of modules per string was constant across the entire site. To convert PVPRO's precise estimates to values more aligned with the real sites, the authors used two separate aggregation methods, one for string counts and the other for module counts. The module estimates are averaged to estimate a single value for each site. The string estimates, on the other hand, are clustered using a kernel density estimate (KDE) method like that described in [7].

Fig. 2(a) shows the essence of the string clustering process used in this work. In this example, the ground truth string counts (as defined by the site drawings) are shown at the top: each combiner box contains either 16, 19, or 22 strings. The combiner-by-combiner estimates are shown next, with the small, filled points. Because the i_{MP} and V_{MP} estimates from PVPRO depend on the actual site performance data, these estimates reflect the current state of the plant and capture any degradation that has grown since the construction of the plant. As a result, the ratio between the estimated and nameplate MPP production does not scale perfectly with the number of strings (or modules) and ratios do not yield whole numbers. The unfilled circles at the bottom visualize how the hypothetical estimates may be clustered by the KDE method. While the clustered estimates are close, they don't always match the blueprint values, as shown by the blue cluster.

Figure 2(b) shows the KDE fit for a real site. In essence, clustering using a kernel density estimate entails: (i) fitting a kernel density estimate to the data and (ii) thresholding using that function’s peaks and troughs. The dark blue line indicates the KDE function. The local minima (green) are then selected as bounding thresholds and the local maxima (gray) are selected as cluster centers. Each estimate (yellow) is bounded according to its adjacent thresholds and is assigned the value of the corresponding cluster center.

Interestingly, there is often a small cluster of PVPRO estimates around zero strings per combiner. While relatively unlikely, this can happen if the combiner box is entirely faulted. It is more likely that either the sensor has failed, or the data historian administrator set up combiner box tags that – for whatever reason – do not correspond to a real combiner box collecting data. The authors have observed this at a site with more than 2000 active combiner boxes.

Finally, some light post-processing is performed on both sets of data, such as rounding the estimates to whole numbers (numbers of strings and modules are discrete values). The results are written to configuration files compatible with an anomaly detection pipeline, like that in [3].

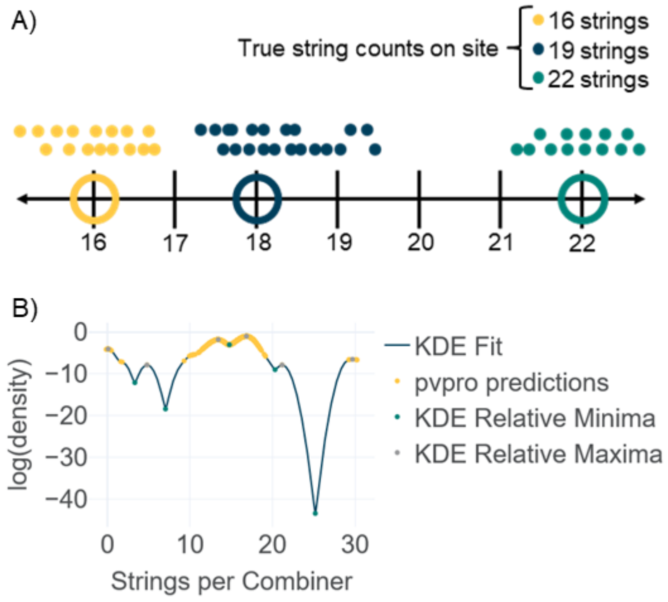

Fig. 2. Overview of the KDE clustering method as it is applied to the strings per combiner estimate clustering. (A) shows hypothetical estimates and true values for a site. (B) shows estimates and thresholds for a real site.

III. Results and Discussion

A. Configuration Speed

Manually configuring a site takes approximately eight human hours to complete. Using the automated method, configuration takes approximately 20 human minutes and 10 computer minutes. This is achieved by automating the most time-consuming, tedious tasks. Consider the following tasks needed to enable the anomaly detection routine discussed in [3]:

1) Translate 100+ to 1000+ tags (1 hour)
2) Extract site layout from blueprints (4 hours)
3) Create model configuration files (2 hours)
4) Review and fix mistakes (2 hours)

Each of these tasks is automated to some extent in the new method, leaving the following human tasks:

1) Specify ~10 tag name patterns (10 minutes)
2) Specify site specific metadata, like site location and module hardware specification (10 minutes)
3) Run code (10 computer minutes)

Note that the automated method does not require the user to spend nearly two hours finding and debugging mistakes. While repetitive and tedious tasks tend to tire people out, leading to mental lapses and errors, algorithms run precisely the same way each iteration. The automatic method automates each of the first three items in the manual method, making the fourth unnecessary.

B. Accuracy Relative to Blueprints

The automated setup simply translates and codifies existing tags, so the resulting configurations are as accurate as the naming nomenclature and the data collected at the site. In some cases, the automatic approach yields more accurate representations of the site layout (down to the combiner box level) than the as-built drawings. The manually generated configs rely on site drawings to infer the site layout, and while these drawings *should be* perfect representations of the site, supply chain and hardware sourcing issues often lead builders to deviate from the plans. Discrepancies can also result if site operators create tags that correspond to non-existent combiner boxes, which the authors have observed in one instance. In general, however, the two methods generate identical site layouts, down to the combiner box level.

There is more deviation between the two configuration methods for the string and module counts. This is because the automatic method relies on estimates, generated from site operations data, whereas the manual method uses the as-built site drawings. To compare the two values, an error metric is established that is simply the difference between the automatic estimates and the blueprint specifications. A negative error results when the automatic estimate is smaller than the blueprint value. The error – after aggregation – is shown in Figure 3.

Since the automated setup relies on performance data, which is significantly affected by the time-of-year, the configuration was performed using both summer and winter data. It is clear that the performance data seasonality affects the estimates. For three of the four sites investigated, the estimates based on winter data yielded string estimates more similar to what is listed in the as-built bill of materials. Winter data also yielded equal or better estimates for all four sites’ module estimates. Several factors impact the accuracy of these results. At high temperatures, PV module performance degrades away from their stated spec sheet performance. The cooler winter temperatures have a slight cooling effect on the modules, lessening the impact of this heat-based performance reduction. The lower overall POA irradiance values in the winter reduce the impact of inverter clipping on the array performance, which artificially shifts the plant away from MPP operation and leads to mischaracterization of the plant by the automatic layout generation algorithm. Finally, curtailment was frequently observed in the summer months. This would have a similar impact as inverter clipping, in which operation is artificially shifted away from the natural MPP of the system.

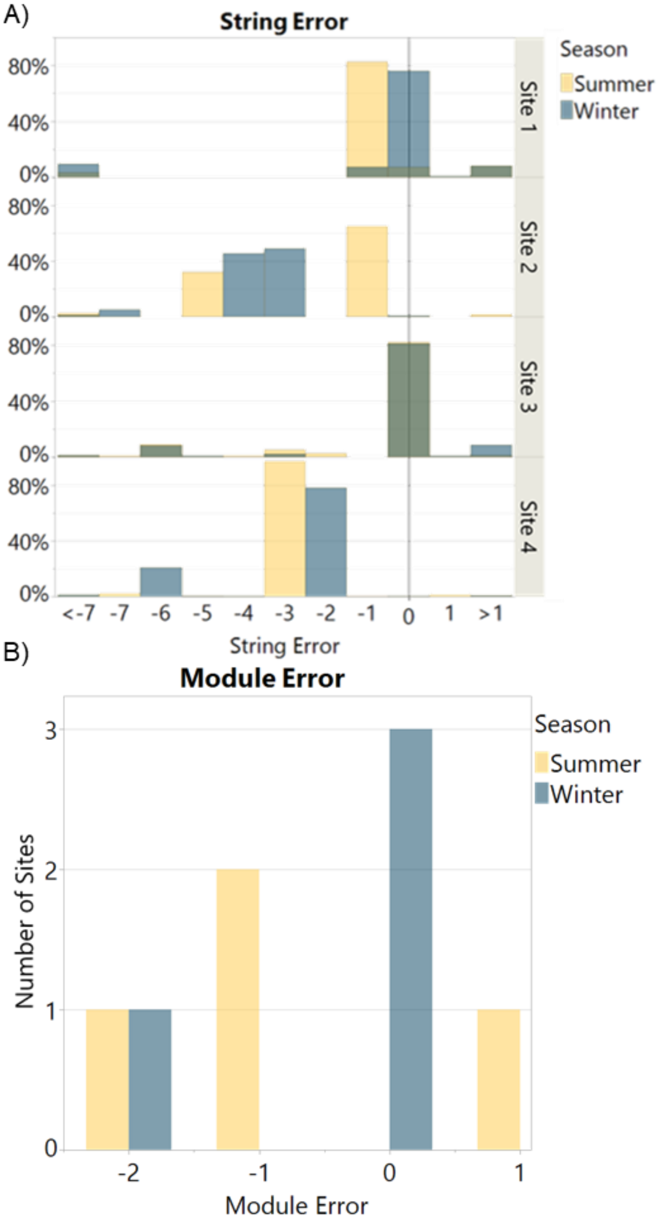

Fig. 3. Overview of the string count and module count estimation error at each site and for different training data.

Figure 3 also shows that the module estimates tend to be far more accurate than the string estimates. This can be explained by the relative stability of voltage signals compared with combiner box signals. Since voltage is measured at the inverter, partial shading will have less of an impact on its signal than it would for a single combiner box. Additionally, the respective aggregation methods likely impact the estimation accuracy. The string estimates rely on KDE clustering to automatically determine the number of unique string counts at a site and which group each combiner box belongs to, which could lead to misassignment for some pieces of hardware. Additionally, since the KDE method is reliant on the point density of the estimates, several outliers could lead to the creation of a highly erroneous cluster value. On the other hand, the module estimates are all clustered into a single group, represented by the average of all estimates. Essentially, the module estimates have more points to determine the single correct cluster-center, so outlier estimates have less of an effect.

Finally, while the bulk of the string estimate errors are within approximately five strings of the site drawing value, they do have a very wide range. This could happen for numerous reasons – for example, imperfections in filtering and modeling noisy data increases the estimates' spread, and sensors flatlining can lead to significant under-predictions. The automatic configuration method is meant to streamline the model configuration process rather than entirely automate it. To mitigate errors of this sort, the authors added a post-processing rule that flags unusually low PVPRO estimates for manual review.

C. *Effect on Anomaly Detection*

While it is useful to explore the specific differences between the automatic estimates and the as-built site drawings, it does not give a definitive answer about the method's ultimate usefulness. That is, it isn't clear whether methods using the automated site setup work. To this end, the authors assessed the performance of an anomaly detection model derived from that in [3] using both types of site configurations: (i) manual configuration created using the as-built site drawings and (ii) automatic configurations created using the methods in this work. The detection routine first generates features for each combiner box using a physics-based model. It leverages a clustering algorithm to identify the anomalous signals in these features. While the clustering algorithm has a tunable sensitivity, this work uses a single sensitivity to simplify comparisons between sites.

Aerial IR imaging scans from the sites were used to generate ground truth classifications for each site. The ground truth classifications were compared with the automatic workflow's results to calculate the true positive rate (TPR) and false positive rate (FPR) for each of the sites. In discussions with site operators, these are the two most important metrics – the former indicates how many faults the algorithm can catch, and the latter indicates how much time will be wasted inspecting non-faulted hardware. F1 score is a standard classification accuracy metric that incorporates TPR and FPR. Sites 2 and 4 each were each scanned in two separate years, so the routine was evaluated separately for each scan – denoted with (a) and (b).

The impact of the automated site set up on the anomaly detection routine's ability to detect string outages is shown in Fig. 4. This – along with the time to configure a site – is the most important metric of the process's utility. If an automatically configured site could never detect anomalies, the faster setup would be moot.

Generally speaking, automating the model setup slightly decreases anomaly detection performance, but the results vary depending on the site. Automatically configuring the site with winter data mostly performs on par with the manual setup, while the detection success was more noticeably reduced when summer data was used for the architecture estimation. As discussed in the prior section, winter calibration may lead to more accurate results because of the relative infrequency of curtailment and inverter clipping in the winter.

Taking a step back, the main takeaway from Fig. 4 is that the automatic configuration achieves detection rates similar to those achieved with the manual configuration. This demonstrates that any inaccuracies derived from automating the setup process do not jeopardize the overall value proposition of continuous monitoring.

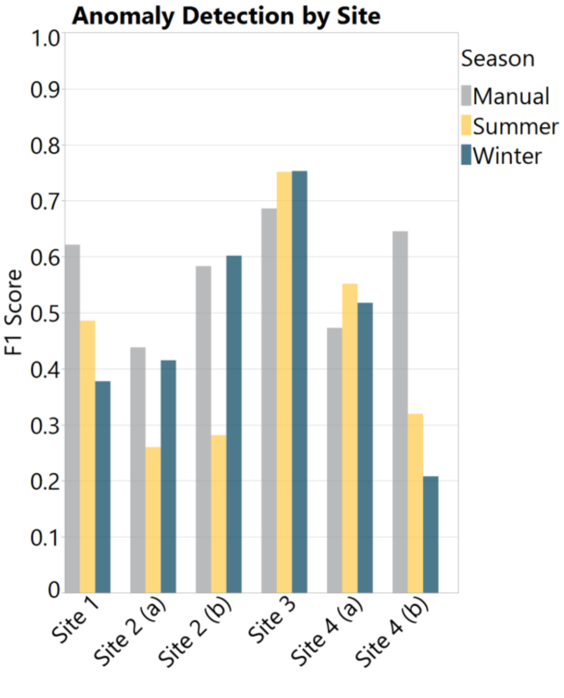

Fig. 4. Automatic configuration achieves detection rates similar to those with manual configuration. Winter mostly outperforms summer.

IV. Conclusion

This automated setup method allows simpler implementation of an anomaly detection workflow that detects many DC faults that currently go unnoticed. More broadly, the setup method accelerates model calibration and, as a result, model uptake throughout utility-scale PV applications. Due to the hierarchical nature of PV plants, models that rely on site performance data need to also know the site layout, down to the module level. Otherwise, they cannot scale the results properly. Since data generally is not collected at a finer granularity than the combiner box level, this work introduces a method which estimates the strings per combiner and the modules per string using combiner box level performance data. For higher levels, the tags are used to infer the relationships between inverters and combiner boxes. The tag translation method is particularly well-suited to the large-scale, hierarchical structure of PV sites (and data historians). It can be applied to translate the tags for any dataset, so long as they have a strictly followed schema.

The new workflow is evaluated by the speed to setup a new site, the similarity between automatically generated site layouts and manually generated ones, and the performance of an anomaly detection workflow using each method. Generally speaking, automating the setup leads to slightly different model configurations. Vitally, the anomaly detection pipeline provides usable results with both the manual and the automatic configurations. Continuous anomaly detection provides operators with more up-to-date information, and this algorithm enables widespread adoption by streamlining the setup process.

Acknowledgements

This work is funded by the U.S. Department of Energy Solar Technologies Office, under award numbers DE-EE-0008976.

References

[1] D. Tansy. "Best practices for operation and maintenance of photovoltaic and energy storage systems; 3rd Edition," Golden, CO, USA. 2018.

[2] N. Vadhavkar, E. Obropta, S. Carey. "Solar Risk Assessment: 2022," kWh analytics, Raptor Maps. 2022.

[3] S. Sheppard, T. Cook, D. Fregosi, C. Perullo, M. Bolen, "Field experience detecting PV underperformance in real time using existing instrumentation," in *2022 IEEE 49th Photovoltaics Specialists Conference (PVSC)*, 2022.

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

[5] A. Triki-Lahiani, A. Bennani-Ben Abdelghani, I. Slama-Belkhodja, "Fault detection and monitoring systems for photovoltaic installations: A review," *Renewable and Sustainable Energy Reviews*, vol. 82, part 3, pp. 2680-2692, 2018. doi: https://doi.org/10.1016/j.rser.2017.09.101.

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

[7] W.J. Wang, Y.X. Tan, J.H. Jiang, J.Z. Lu, G.L. Shen, R.Q. Yu, "Clustering based on kernel density estimation: nearest local maximum searching algorithm," *Chemometrics and Intelligent Laboratory Systems*, vol 72, iss. 1, pp. 1-8, 2004. Doi: https://doi.org/10.1016/j.chemolab.2004.02.006.