System of Automatic Detection of PV Mounting Configuration

Oliver Osvald, Tomáš Hruška, Artur Skoczek, Tomáš Cebecauer, Markéta Hulík Jansová

Solargis s.r.o., Bratislava, 811 09, Slovakia

***Abstract* — A complete set of reliable system configuration parameters is a prerequisite for the holistic analysis of photovoltaic (PV) power production systems. However, in practice, it is common to have an incomplete or incorrect configuration on record. We propose an automatic system for determining basic mounting parameters using PV yield time series and location data. The system consists of quality control and inference modules, which we describe in this paper. The core inference module, based on solving an optimisation problem, is illustrated. Lastly, we demonstrate the utility of the proposed system with an evaluation exercise in the context of PV yield forecasting over a real asset portfolio.**

I. INTRODUCTION

Reliable PV system configuration parameters are crucial for a multitude of applications designed to ensure optimal operation, proactive maintenance, and accurate performance of PV installations. Such applications include but are not limited to, robust monitoring platforms and fault detection mechanisms designed to diagnose system malfunctions and anomalies, comprehensive performance evaluation procedures for benchmarking and optimising system efficiency, and dependable energy forecasting tools for grid integration and energy management purposes. However, direct practical experience within *Solargis* has revealed a concerning problem: issues related to the accurate definition of the mounting configuration of PV sites are surprisingly commonplace. We saw cases with undefined, partial, or even complete but significantly inaccurate parameter sets, as also reported in [1] and other works.

There are also some more subtle cases, where the proposed system is needed. One such case is when the system is not accurately represented in the provided PV configuration due to the technical complexity of the site. Another instance may be a site that is built on complicated topography. In those scenarios, the measured PV production behaves according to a set of '*effective*' configuration parameters which may be useful for some of the applications mentioned above.

Recognising the need for more accurate PV system configuration metadata, we set out to develop a system capable of ensuring access to a reliable and consistent set of these parameters. There are existing methods described in the literature, e.g. [1], [2], [3], or [6], which address this need. Our goal was to create a solution that is directly linked to and validated against actual PV production measurements provided as a time series. More specifically, our objective was to design a detection system that would take as its input the geographical coordinates of a PV installation and a corresponding time series of its energy output. Upon processing this input data, the system would determine and return the following set of fundamental system configuration parameters:

- **Mounting geometry type** (structural arrangement of the PV array, e.g. fixed-tilt or 1-axis tracking system),
- **Azimuth angle** (horizontal orientation of the PV array, measured in degrees clockwise from true north),
- **Tilt angle** (the angle at which the PV array is inclined relative to the horizontal plane),
- **Installed power capacity** (total rated power output of all the PV modules in the system),
- **Relative (row or column) spacing** (spacing between rows or columns of a PV array relative to its width; inverse of *ground coverage ratio*),
- **Rotation limits** (the limits of rotational movement for tracking mounting types).

II. APPROACH

At *Solargis*, we developed proprietary assets that enable us to pursue this effort: A satellite-based solar resource model, a state-of-art PV simulation engine, and a modular data quality control pipeline for measured time series data. With these instruments at hand, the core idea behind the detection algorithm is the formulation of the *optimisation problem* [2], [3], [6]: Given a range of feasible simulated configurations, identify one which has the *best* fit to the reference (measured) time series.

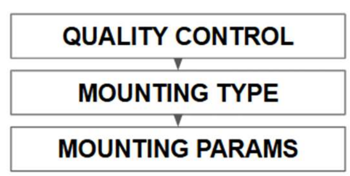

Diagram 2.1: Mounting configuration detection system.

For this approach to work well, there is another essential requirement – the fitting process must be performed on valid and relevant data. The validity of measured data ensures the accuracy and reliability of the fit, preventing the models from fitting on erroneous or inaccurate information. Consequently, the complete solution is comprised of a series of distinct components organised into a data pipeline (Diagram 2.1). The separation of the solution into modular components within a pipeline also offers advantages in terms of maintainability, scalability, and the ability to adapt individual stages as needed without affecting the entire system.

III. METHODOLOGY

A. Quality Control and Filtering

In the real world, PV production data is often affected by at least some of the numerous possible issues that may occur in measured time series. These issues can come from environmental, technical, or systemic sources, and may include missing data, duplicates (redundant or repeated values), values exceeding physical limits, time reference and data resolution inconsistencies, hardware malfunction, communication failures, string/inverter failures, shutdowns, power clipping, soiling, shading, seasonal changes in ground albedo, curtailment due to grid constraints, and maintenance outages. As [1] and [4] also highlight, it is vital to ensure that the quality of the measured time series is controlled before feeding it into the more advanced inference algorithms. For this purpose, we developed a quality control (QC) data pipeline consisting of several automatic quality assessment methods, some of which were inspired by notions described in [5] and/or were adapted from QC methods intended for irradiance measurements, which are considered strong precedent for PV QC [1]. These are included as preprocessing steps in the proposed system and, where applicable, include checks for time reference, invalid values, physical limits, logger issues, static values, statistical outliers, inverter clipping, and shading.

After the QC steps, the time series is additionally filtered prior to detection to ensure the inference is made from relevant data. Due to high variability during cloudy days and low solar elevations, fixed minimum thresholds are imposed on modelled clearness index (K_t as defined in [5]) and solar elevation (SE), such that $K_t > 0.6$ and SE > 5°. Similar filtering approach using K_t can be found in [6].

The importance of this step is easy to gloss over. However, in our experience, the removal of data points which are affected by unexpected and anomalous effects or likely to produce a short scale mismatch with modelled reference is crucial to ensuring the accuracy and consistency of the predictive system.

B. Mounting Type Identification

This module aims to use measured time series data to distinguish between three types of mounting geometry: *fixed* (with a single angle of tilt), 1-axis *tracked*, and 2-axis *tracked* mounting. Identification of mounting type is a prerequisite to selecting the applicable configuration parameters in the detection algorithm.

For this task, two binary classification models were trained using a sample of 1406 sites with PV time series and confirmed mounting types. This number was boosted by introducing *synthetic* sites – randomly chosen locations with randomly assigned configuration and a corresponding simulated PVOUT in random granularity. This was a convenient way to diversify the sample across geographies and configurations. The classifier model input features included values drawn from the (pre-filtered) measured time series at chosen quantile levels and features relating measured production to modelled solar parameters (GHI, DNI, DIF). The first classifier, a *Random Forest* algorithm, uses the extracted features to determine whether the target is a *fixed* or *tracked* PV system. The second classifier, a *Gradient Boosting* algorithm, was similarly structured and trained to determine if a *tracked* system has a *2-axis* mounting geometry. Both models were trained to a high standard, with high F1-score metrics (nearing 1.0) and only a handful of misclassified cases. In both cases, visual inspection of the false predictions showed that some datasets in question still contained severe data quality issues despite the QC procedures applied. As a result, the final module combines the two classifiers in an ensemble and provides the capability to infer a PV system's mounting type with high accuracy and reliability.

C. Detection of Mounting Configuration Parameters

Given solar resource time series at a geographic location, it is possible to calculate the global tilted irradiance (GTI) and simulate the energy yield (PVOUT) time series corresponding to a specific mounting configuration using a simulation engine. This capability extends beyond a single configuration, making it feasible to simulate a whole range of configurations. These configurations can be conceptualised as forming a multidimensional parameter search space. Within this space, each dimension corresponds to a specific configuration parameter (e.g. azimuth or tilt) which influences PV system performance. An example of a three-dimensional sub-space of this complex parameter search space is visualised in Fig. 3.1.

The next step is to identify the configuration that best represents the measured time series. There are many well-known metrics suitable for assessing the similarity between two time series, such as RMSE, MAPE, or the Pearson Correlation Coefficient. However, in the context of comparing a measured time series to a simulation, it is expected that the measured time series rarely exceeds the simulation significantly, whilst potentially underperforming frequently due to various practical circumstances (system failures, shading, etc.). Therefore, an RMSE-like error metric E_{SIGN} emerges, in which the positive and negative contributions are weighted unequally, such that

$$E_{SIGN}(\boldsymbol{p}^{MEAS}, \boldsymbol{p}^{SIM}) = \sqrt{\frac{w^+}{N}\sum_{\substack{k=1 \\ \Delta p_k > 0}}^{N} (\Delta p_k)^2 + \frac{w^-}{N}\sum_{\substack{k=1 \\ \Delta p_k < 0}}^{N} (\Delta p_k)^2} \quad (1)$$

where:

$$\Delta p_k = p_k^{MEAS} - p_k^{SIM}, \qquad k \in 1, 2, \dots, N$$

$$\boldsymbol{p}^{MEAS} = (p_1^{MEAS}, p_2^{MEAS}, \ldots, p_N^{MEAS})$$

$$\boldsymbol{p}^{SIM} = (p_1^{SIM}, p_2^{SIM}, \ldots, p_N^{SIM}).$$

The weights w^+ and w^- are restricted such that

$$w^+, w^- \in [0,1], \qquad w^+ + w^- = 1, \qquad w^+ > w^-,$$

so that the negative error contributions are neglected. The precise values of the weight parameters are determined empirically from the sample data. The search space is thus fully evaluated using the E_{SIGN} metric against the measured series and minimised to find a preliminary configuration candidate. This is further illustrated in Fig. 3.1, which shows the interplay of the three key configuration parameters and their impact on the E_{SIGN} metric, i.e., the similarity between simulated and measured, where lower E_{SIGN} represents a closer fit. By systematically exploring this search space using optimisation techniques, it becomes possible to identify the best-fit PV system configuration corresponding to the measured production.

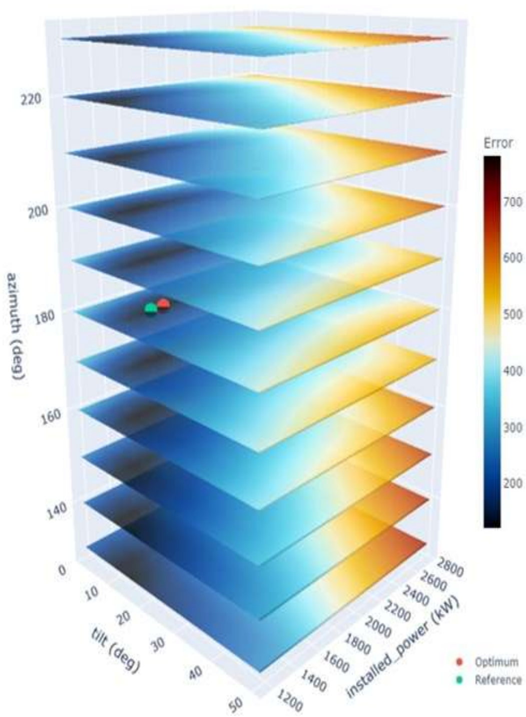

Figure 3.1: Search space evaluated with error metric E_{SIGN} in a sub-space spanned by PV configuration parameters azimuth, tilt, and installed power capacity.

IV. EVALUATION

A. Primary evaluation

The end-to-end mounting configuration detection system underwent a comprehensive evaluation using datasets from 86 operational photovoltaic sites. These sites, representing a diverse group of real PV installations, feature *fixed* mounting systems with known mounting configuration parameters and are geographically distributed across various regions. The minimum eligibility criteria for each site were known tilt, azimuth, installed power, and at least 6 months of measured time series. It is important to note that the parameter metadata was validated to the best of our knowledge and in some cases included our best estimates based on available evidence. The evaluation set encompassed various time series sampling frequencies, ranging from 5-minute intervals to more aggregated hourly measurements, reflecting the heterogeneity found when working with real production datasets.

		Region				
PV Parameter [unit]	Statistic of absolute difference (reference vs. predicted)	Central Europe (30)	South Europe (33)	Southeast Asia (20)	Indian Ocean (3)	All
Tilt [deg]	Average	2.40	4.00	2.05	6.00	3.06
	Median	2.00	2.00	1.50	6.00	2.00
	Std deviation	1.87	5.49	2.04	4.00	3.86
Azimuth [deg]	Average	7.00	4.58	16.25	15.33	8.51
	Median	5.50	3.00	15.00	16.00	5.00
	Std deviation	6.03	7.13	13.41	4.04	9.73
Installed Power [%*]	Average	6.18%	6.57%	6.56%	9.56%	6.53%
	Median	6.33%	5.00%	5.63%	8.47%	5.79%
	Std deviation	1.96%	5.05%	4.21%	4.02%	3.95%

*Table 4.1: Results of primary evaluation of key mounting parameters on 86 sites from Europe and Asia. * Installed power deviations are represented as a % of the reference (declared) value.*

The results of this evaluation exercise are shown in Table 4.1. The table summarizes the typical deviations between the predicted and declared parameters across the regions and in total. Notice that the deviations varied across regions. This was due to varying data quality, resolution and other factors. Sites closer to the equator typically employ lower *tilt* angles that in turn make the system production less sensitive to deviations in *azimuth*.

The individual sites in the primary evaluation exercise were also visually inspected for fit between the measured production

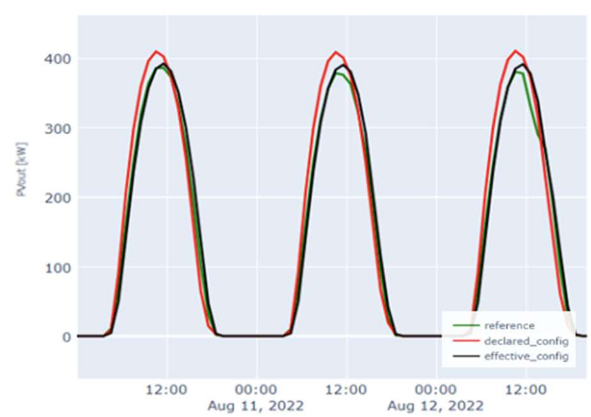

Figure 4.1: Forecasted PV yield using declared and effective (estimated) mounting configuration vs reference historical yield.

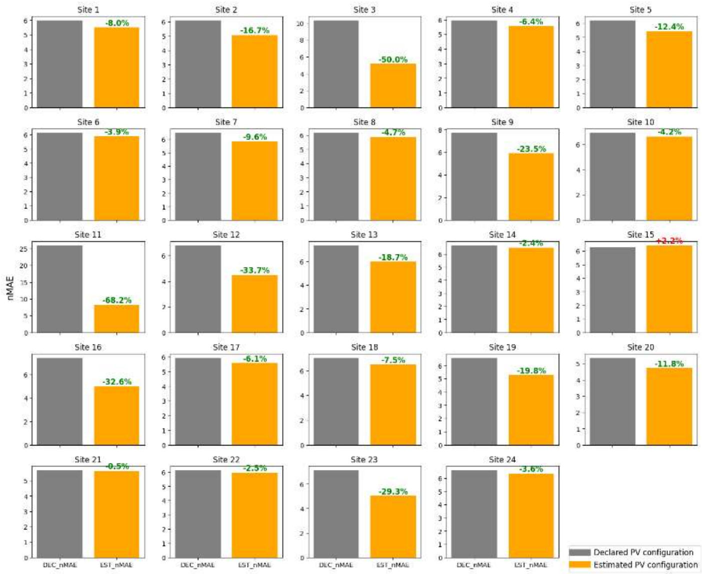

Figure 4.2: 24 sites in central Europe showing yearly comparison of normalised mean absolute error (nMAE) when using declared (grey) vs. estimated 'best fit' (orange) PV configuration.

time series and the best-fit simulation. This step was crucial as it confirmed our fitting was accurate and excluded uncertainty in the reference metadata. On most of the examined sites, we observed a good fit, as displayed in Fig. 4.1 where the curve of 'effective config' corresponds to the best-fit simulation.

B. Case study: Application in PV Yield Forecasting

The quality of PV energy yield forecasts depends on the quality of forecasted solar irradiance (global horizontal irradiance - GHI) by numerical weather prediction (NWP) models, the PV simulation accuracy, and provided PV system configuration. Regardless of the quality of the NWPs and the PV simulation, incorrect metadata may lead to systematic deviations between forecasted and real PV power output. In these cases, the proposed detection system can help us leverage historical yields data to reduce the final forecast uncertainty.

Our study compared forecast accuracy when using *declared* and *estimated* PV configurations on a portfolio of 24 operational PV installations in the central Europe region. The results, illustrated in Fig. 4.2, show improvement in forecast accuracy in 23 out of 24 sites when forecasting with estimated PV configuration. Overall statistics for the forecasted portfolio are shown in Table 4.2. This table aims to provide an overview of the typical size of deviations between declared and estimated parameters. The difference in median and average metrics indicates the presence of significant differences in a handful of sites where the estimated parameters deviated substantially (e.g. sites 3, 11, 12) from the declared and resulted in a severe reduction in forecast normalised mean absolute error (nMAE). The rest of the sites showed small-to-medium improvements except for site 15, where the difference in nMAE was +2.2% higher compared

Relative difference (estimated vs declared PV configuration)				
Statistic	**Installed power [%]***	**Tilt [deg]***	**Azimuth [deg]***	**Forecast nMAE**
Median	5.00%	5	10	-8.80%
Average	7.62%	7.13	15.13	-15.58%

Table 4.2: Comparative statistics for using estimated vs. declared PV configuration metadata in the portfolio of 24 forecasted sites.

to a forecast using declared settings. Since the observed increase is relatively small and the occurrence is not frequent, it is hard to conclude anything tangible from this single observation. There is certainly room for improvement and seeing more of such cases in the future can give us clues for optimisation.

It could be argued that results from a single portfolio may not generalise to others – after all, the improvement in forecast accuracy is enabled by the quality of the provided metadata. However, what we can conclude from the example here is that there exist portfolios of PV power plants where the application of a PV configuration detection system makes a significant difference to the final forecast. In practice, it is therefore beneficial to have this predictive solution available to mitigate the uncertainty associated with reference metadata within the PV yield forecasting process.

V. LIMITATIONS

Typically, the PV configuration detection system relies on at least a year of measurements so that the nuances of annual seasonality at a given site are well represented in the sample. Since the algorithm depends on the presence of clear sky conditions in the sample, the ideal amount will vary based on site climatology and the amount of data affected by issues and anomalies (and their temporal distribution). Minimum data resolution is hourly or more granular. During our evaluation and testing, there were cases in which the system performed significantly worse when the time series was aggregated to hourly (compared to the original 15-minute resolution), hence we recommend using the highest granularity available. From our general experience with this system, which now includes both testing and production use, 15-minute granularity is

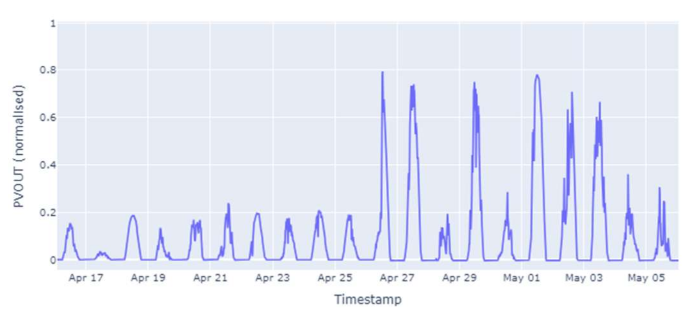

Figure 5.1: Example of a measured time series with time-dependent PV configuration.

generally granular enough to capture most of the required detail.

The configuration parameters *relative spacing* and tracker *rotation limits* are detected in a posterior search, allowing a tighter fit to the measured time series. Overall, these parameters are less influential than tilt, azimuth, and installed power, and it became apparent that their inference is sensitive to the quality of fit of the three mentioned. The effect of spacing and rotation limits manifests at lower sun elevations and other features such as shading also impact the prediction quality. Moreover, sites with reliable validation metadata including these two parameters were scarce, which prevented us from performing any extensive validation at this time.

Moreover, in the implementation presented here, it is assumed that the evaluated PV system's configuration does not change over time and the detection system is not able to provide interval-based evaluation. This could occur, for example, when the PV installed capacity at the site increases by connecting additional strings or segments. Such a scenario is illustrated in Fig. 5.1. In such cases, the proposed algorithm typically does not find a good fit to the full measured time series. Instead, we recommend manually (or otherwise) splitting the time series into multiple intervals and treating each interval as a separate instance of PV configuration. In such a way, the user could determine the time-dependent PV configuration on a piece-wise basis or simply focus on the latest.

Lastly, it is important to mention the detected configuration represents an *effective* best fit with respect to the PV simulation engine used and pre-configured with system specifications like module material type and age, degradation rate, hardware loss coefficients, DC/AC ratio, and others. Moreover, the process of PV simulation employed here may not be able to perfectly capture the environmental or technical complexity of the real site, e.g. a complex site terrain or contributions from multiple power plant segments with different configurations. Hence, we have sometimes observed that, despite a good fit between measured and simulated time series found within our framework, the real physical on-site configuration still differs. This may or may not matter to the final user depending on their intended use. In applications using the best-fit PV simulation as a reference, e.g. monitoring and forecasting, the exact resulting PV configuration values (corresponding to the best fit) might not be that interesting if the overall reference provides a good representation of the real system. On the other hand, if the final aim is the verification of real metadata, a discrepancy between estimated and real metadata (despite a good fit) poses further questions. To reconcile, it may be necessary to review the PV simulation settings, and the PV system's real physical settings, and to consider other factors that may not be well represented in the modelling. Ultimately, such reconciliation can lead to a better holistic understanding of the PV power production system's behaviour and performance.

VI. SUMMARY AND RELEVANCE

We proposed an automatic computational system for the detection of PV configuration. The components of the proposed system were described, and its core inference module was explained. The detection system was evaluated both independently and in an applied setting using a case study in PV yield forecasting accuracy. In the case study, the proposed detection system improved forecast accuracy in most of the evaluated portfolio of sites. In some cases, there were dramatic improvements in order of tens of % of installed capacity when compared to a forecast conducted with declared configuration. The tool failed to improve forecast accuracy in only one of the 24 test cases.

Accurate solar power forecasts are crucial for electrical grid operators and PV asset managers. These forecasts enable proactive planning, leading to the elimination or mitigation of logistical challenges such as managing energy storage, scheduling maintenance, and ensuring sufficient power supply to meet demand fluctuations. Furthermore, accurate predictions minimize financial risks associated with energy imbalances, grid penalties, and inefficient resource allocation. The result is a more stable and reliable electrical grid.

While numerical weather prediction (NWP) models, which form a primary basis for solar forecasting, are continuously refined and improved, the proposed solution targets one of the other critical factors influencing forecast accuracy - reliability and accuracy of PV system configuration parameters. Inaccuracies or inconsistencies in this metadata can introduce significant errors in the forecasting process, regardless of the level of sophistication of the NWP models employed.

The value proposition of the PV configuration detection system extends beyond simply improving forecast accuracy. Uncovering and reconciling the effective and up-to-date configuration parameters provides valuable feedback to the source of the questionable metadata. This could include PV asset managers responsible for maintaining system records, installation companies who initially provided the data or other entities involved in the lifecycle management of the PV asset. Armed with this feedback, stakeholders can investigate the identified discrepancies and implement appropriate resolutions, such as updating their databases, correcting installation records, or conducting site inspections to verify the actual system configuration. This feedback loop creates a continuous improvement cycle, leading to more reliable metadata over time and consequently, more accurate and dependable solar power forecasts and system diagnostics.

ACKNOWLEDGEMENT

An initial part of this work was developed as a deliverable for the *Serendi-PV* project, a research & development consortium funded by the European Union's Horizon 2020 programme. The project focused on innovations in the areas of PV simulation, modelling, monitoring, quality control, and forecasting. We wish to thank our *Serendi-PV* partners for the fruitful collaboration and exchange of datasets and insights. Our gratitude also goes to many internal collaborators at *Solargis* who provided feedback during the development, namely the customer data, consultancy, and forecast teams; and to multiple other teams who had developed the prerequisite tools enabling this work to succeed.

REFERENCES

[1] S. Killinger, N. Engerer, B. Müller, "QCPV: A quality control algorithm for distributed photovoltaic array power output" *Solar Energy* 143, 120-131 (2017)

[2] Y.M. Saint-Drenan, S. Bofinger, R. Fritz, S. Vogt, G.H. Good, and J. Dobschinski, "An empirical approach to parameterizing photovoltaic plants for power forecasting and simulation" *Solar Energy* 120, 479-493 (2015)

[3] B. Meng, R.C.G.M. Loonen, and J.L.M. Hensen, "Data-driven inference of unknown tilt and azimuth of distributed PV systems" *Solar Energy* 211, 418-432 (2020)

[4] A. Rivera, M. Theristis, E. Koumpli, S. Theocharides, G. Makrides, J. Sutterlueti, J.S. Stein, and G.E. Georghiou, "Data processing and quality verification for improved photovoltaic performance and reliability analytics" *Progress in Photovoltaics, Wiley,* 10.1002/pip.3349 (2020)

[5] M. Sengupta, A. Habte, S. Wilbert, C. Gueymard, J. Remund, E. Lorenz, W. van Sark, and A. R. Jensen, "Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Fourth Edition" *Golden, CO: National Renewable Energy Laboratory* NREL/TP-5D00-88300 (2024)

[6] V.D. Ruelle, M. Jeppesen, M. Brear, "Rooftop PV Model Technical Report", *The University of Melbourne <https://aemo.com.au/-/media/files/electricity/nem/planning_and_forecasting/demand-forecasts/nefr/2016/uom-rooftop-pv-model-technical-report.pdf>. (*2016)