Signal Processing on PV Time-Series Data: Robust Degradation Analysis without Physical Models

Bennet Meyers, Michael Deceglie, Chris Deline, and Dirk Jordan

Stanford University and SLAC National Accelerator Laboratory, Menlo Park, CA, USA
National Renewable Energy Laboratory, Golden CO, USA

Abstract—A novel unsupervised machine learning approach for analyzing time-series data is applied to the topic of photovoltaic (PV) system degradation rate estimation, sometimes referred to as energy-yield degradation analysis. This approach only requires a measured power signal as an input—no irradiance data, temperature data, or system configuration information. We present results on a data set that was previously analyzed and presented by NREL using RdTools, validating the accuracy of the new approach and showing increased robustness to data anomalies while reducing the data requirements to carry out the analysis.

Index Terms—photovoltaic systems, distributed power generation, computer aided analysis, data analysis, statistical learning.

I. INTRODUCTION

A large amount of research has been published on methods for estimating PV module and system degradation rates (for an overview, see [1]). This work focuses on the estimation of system-level degradation from historical time-series power data. Recent advances in this area include the year-on-year (YOY) estimation method [2], using clear sky models for robustness to irradiance sensor issues [3], and the development of RdTools to standardize and automate the process of estimating system degradation rates from historical time-series data [4].

Typically, the process for estimating the degradation rate of PV systems requires three steps: normalization, filtering, and data analysis. A standard implementation of this process is described in detail in [4]. The normalization process requires the use of a physical model to estimate the expected power output of the system. The model can be based on a simple performance ratio calculation or a detailed DC performance model (such as the Sandia Array Performance Model [5]). Even with the standardization provided by RdTools, there are many decisions that must be made by an analyst attempting to implement this process: what performance model to use, what source of irradiance data to use (on-site measurements, satellite-based measurements, a clear sky model, etc.), how to estimate plane-of-array irradiance from available data, and how to estimate the cell temperature of the system, to name a few. In addition, all methods also require knowledge of the system configuration including some combination of site location, mounting configuration (tilt and azimuth, if fixed-tilt), and module technology. This general approach works very well for analyzing utility-scale PV systems, which tend to be well characterized, data rich, and of sufficient size to justify hand-cleaning of data and hand-tuning of models.

Unfortunately, many real-world data sets of distributed rooftop PV systems lack correlated irradiance and meteorological data and system configuration parameters. Additionally, analysis of distributed rooftop systems must be done at a scale that precludes any hand-cleaning or tuning on a site-by-site basis. In light of this, we present an alternative method for estimating the bulk degradation of PV systems that requires no additional information, data, or models. Based on a method for estimating clear sky system power directly from power data (presented at WCPEC-7/PVSC-45 [6]), this novel approach utilizes recent results from the area of optimization and unsupervised machine learning, implementing a domain-specific form of generalized low rank models (GLRM) [7].

We refer to the clear sky estimation methodology as Statistical Clear Sky Fitting (SCSF). Building on this approach, we develop a method for estimating YOY system degradation as part of the GLRM fitting procedure.

This paper presents three topics: the methodology of including a degradation rate estimation in the model fitting procedure, a discussion on model tuning parameters and their impact on degradation estimation, and a fleet-scale application of SCSF. We apply SCSF to a data set of over 500 PV systems in the United States, which was previously analyzed using RdTools [8]. We present a comparison between the model-less SCSF approach and the RdTools-based approach. We find that the model-less approach agrees very well with RdTools on average over all sites, while having a lower variance around that average and being more resilient to common data errors. The algorithm is available as open-source software here: https://github.com/slacgismo/StatisticalClearSky

II. METHODOLOGY

A. Background on SCSF

For PV researchers and professionals, SCSF can be thought of as an abstract function that takes in measured power data and returns an estimate of the clear sky power output of the system, as shown in Figure 1. As described in Section II-B, the previously presented methodology has been extended to include estimation of YOY system degradation.

SCSF is comprised of two parts: (1) a mathematical model of PV power data over time and (2) an algorithm for optimally
fitting this model to observed data [6]. This approach exploits approximate periodicity in the PV power signal on daily and yearly time-scales to help separate the clear sky behavior of the system from all other dynamics. At a high level, the procedure involves forming the time-series data into a matrix and then finding a low-rank approximation of that matrix [7]. Fitting the model requires solving a non-convex optimization problem (called the "SCSF Problem"), and the algorithm presented in [6] provides an effective heuristic that approximately solves the problem by solving a series of convex optimization problems, given a reasonable starting point.

B. Estimation of Degradation

We expect the shape of the clear sky signal to repeat year-over-year, but we expect and do observe that the energy output reduces over time. In other words, a degradation signal is present in the power data. This means that fitting a clear sky model to multi-year data sets requires relaxing exact yearly periodicity so that the energy content changes over time.

Estimating this degradation, however, is mathematically difficult due to the non-convex nature of YOY degradation. As stated previous, the SCSF Problem is non-convex; however, it has a particular structure known as biconvexity [9]. This structure is what allows for the algorithm based on solving a series of convex optimization problems. Naively introducing a YOY degradation constraint on the problem would break the biconvexity of the problem. In this section, we describe how the non-convexity arises and how the issue is solved mathematically.

Let \( d_i \) be the total estimated clear sky energy produced by the system on day \( i \) in the data set, and let \( T \) be the total number of days in the data set. Enforcing strict yearly periodicity in clear sky daily energy would mean including the following equality constraint to the optimization problem:

\[
d_{i+365} - d_i = 0, \quad \text{for } i = 1 \ldots T - 365, \tag{1}
\]

where each \( d_i \) is a decision variable. This is a linear constraint, so if the original subproblems were convex, the new problem with this constraint added would also yield convex subproblems. The most natural way to introduce degradation to the problem (removing strict year-over-year periodicity) would be to introduce an additional decision variable, \( \beta \):

\[
d_{i+365} - d_i = \beta, \quad \text{for } i = 1 \ldots T - 365, \tag{2}
\]

which allows pairs of days a year apart from each other to differ in daily energy. All pairs of days must differ by the same value, which is found by solving the optimization problem. This relaxation is still a linear constraint. However, YOY degradation is not defined in terms of the difference in energy output but in terms of the percent change in energy output. In other words, we would actually like to introduce the following constraint to the problem:

\[
\frac{d_{i+365} - d_i}{d_i} = \beta, \quad \text{for } i = 1 \ldots T - 365. \tag{3}
\]

This equation contains a ratio of problem variables and is therefore not convex. If we were to naively add this constraint to the SCSF Problem, we would break the biconvex structure. We overcome this difficulty by exploiting the fact that we are using an iterative algorithm to solve the problem. We use bootstrapping to linearize Equation 3 as follows: Let \( d^{(j)} \in \mathbb{R}^T \) be the estimate of the clear sky energy output for all days after iteration \( j \) of the algorithm, and \( d^{(j)}_i \in \mathbb{R} \) is the estimate of day \( i \) after iteration \( j \). So, the bootstrap-linearized constraint at iteration \( j \) becomes

\[
\frac{d^{(j)}_{i+365} - d^{(j)}_i}{d^{(j)}_i} = \beta, \quad \text{for } i = 1 \ldots T - 365. \tag{4}
\]

The denominator is the estimate from the previous iteration and is not a problem variable during this iteration. Equation 4 is a linear constraint in \( d^{(j)}_i \) and \( \beta \), so the convexity of the problem is maintained. In the limit \( k \to \infty \), \( d^{(j)}_i = d^{(j-1)}_i \), so this \( \beta \) converges to the YOY degradation rate in the clear sky response of the system.

C. Tuning parameters

The SCSF problem contains four tuning parameters that control the fit and behavior of the model, as summarized in Table I. The functional form of SCSF is therefore:

\[
\hat{p}_{\text{clear sky}} = f(p_{\text{measured}}; k, \tau, \mu_L, \mu_R). \tag{5}
\]

It is necessary to correctly set the four model parameters for the function to generate reasonable and useful estimates of clear sky power. The recommended values given in Table I are based on the parameter sensitivity study presented in Section III-B.

| Param. | Description                                    | Value   |
|--------|------------------------------------------------|---------|
| \( k \) | Rank of the matrix factorization               | 6       |
| \( \tau \) | Approximate quantile of the data to be fit     | 0.85    |
| \( \mu_L \) | Weight of the smoothing term on the left matrix | 500     |
| \( \mu_R \) | Weight of the smoothing term on the right matrix| 1000    |

III. RESULTS

The algorithm is applied to a data set of 573 residential PV systems. This data was previously compiled by NREL for an analysis [8] that used RdTools to normalize, filter, and analyze the data. For that work, 387 of the 573 systems were selected in a data down-selection process that removed sites with less than two years of data. The NREL study used hourly satellite data for normalization, a major source of uncertainty in the analysis. In Sections III-A and III-B, we reference a
Fig. 2. Top: measured time-series data for ES1, viewed as a heatmap. Bottom: estimated clear sky response (with degradation).

Fig. 3. Comparison of summed daily energy, calculated from the observed power signal and from the two estimates of the clear sky response. The same set of parameter values is used for both model implementations. The orange dots indicate roughly clear days in the data set. Note that the summed daily energy in the clear sky models creates an approximate upper envelope fit of the measured data.

Fig. 4. Comparison of the residuals of the SCSF model with a degradation term and without. A simple linear fit is shown for both. The residuals of the fit without a degradation term have a strong dependence on day number, indicating that the model is missing a time-dependent term.

Fig. 5. Comparison of the residuals of the SCSF model with a degradation term and without. A simple linear fit is shown for both. The residuals of the fit without a degradation term have a strong dependence on day number, indicating that the model is missing a time-dependent term.

A. Fitting the degradation rate

We illustrate the need to fit a degradation rate in the SCSF model by examining Example System 1 (ES1). The power production data and its SCSF baseline are shown in Figure 2. This example was chosen for its relatively large degradation rate to emphasize the necessity of including degradation in the SCSF model. ES1 has a degradation rate between $-2.4\%$ and $-5.6\%$ according to the RdTools analysis. The SCSF analysis yields a degradation rate of $-2.6\%$. Figure 3 shows a comparison of the measured daily energy from ES1 to the estimated clear sky daily energy with and without fitting a degradation rate. The model with the degradation rate is observed to more closely follow the upper envelope of the measured daily energy trend. Figure 4 shows the residuals for the SCSF model with and without a degradation term. Residuals are calculated as the actually daily energy minus the estimated clear sky daily energy, on approximately clear days only (orange dots in Figure 3). The residuals for the model without a degradation term show a strong dependence on day number, showing that the model is missing a time-dependent factor.

B. Tuning parameter sensitivity

We explore the sensitivity of model fitting to the parameters described in Section II-C by examining Example System 2 (ES2), whose data is shown in Figure 5. This system had a high degree of uncertainty in the RdTools analysis, likely due to the large number of cloudy days. The RdTools analysis yields a degradation rate of $-2.6\%$. Figure 3 shows a comparison of the measured daily energy from ES1 to the estimated clear sky daily energy with and without fitting a degradation rate. The model with the degradation rate is observed to more closely follow the upper envelope of the measured daily energy trend. Figure 4 shows the residuals for the SCSF model with and without a degradation term. Residuals are calculated as the actually daily energy minus the estimated clear sky daily energy, on approximately clear days only (orange dots in Figure 3). The residuals for the model without a degradation term show a strong dependence on day number, showing that the model is missing a time-dependent factor.

| Param. | Low Val. | Mid Val. | High Val. |
|--------|----------|----------|-----------|
| $k$    | 4        | 6        | 8         |
| $\gamma$ | .8      | .85     | .9        |
| $\mu_L$ | 1e2     | 5e2     | 1e3       |
| $\mu_R$ | 5e2     | 1e3     | 5e3       |
Fig. 5. Measured power data for ES2. This system was subjected to a grid search sensitivity study across the four tuning parameters.

Fig. 6. The distribution of degradation rates for ES2 from the grid search study. The red dashed line corresponds to the parameter settings given in Table I.

Degradation estimate for this system is \(-2.2\%\), with the 95th percentile range of \([-4.5\%\) to \(+0.4\%\).

A grid search was performed over the values summarized in Table II, resulting in \(3^4 = 81\) SCSF runs on ES2. The average solve time for these runs was 8.7 minutes, for a total linear time of 11.7 hours of computation. This was parallelized over four processors, so the study took about three hours to complete.

The distribution of the degradation rates obtained from this study is presented in Figure 6. The estimation of degradation is quite stable over these range of values, and the selected values in Table II correspond to a distinct center peak in the distribution. We interpret these results to mean that the degradation rate for ES2 is \(-1.4\%\) with an uncertainty of \(\pm 0.1\%\).

Qualitatively, we found that setting \(\mu_L\) and \(\mu_R\) too low or too high adversely affected the quality of the clear sky baseline, but did not have a strong effect of the estimated degradation rate. We also found that the rank parameter, \(k\), must be sufficiently large such that the model is expressive enough to capture the dynamics in the clear sky signal. Setting \(k\) to be too large does not impact the model fitting, but simply adds more variables that must be solved for in the SCSF Problem. Setting \(k = 6\) was expressive enough for all systems we observed.

Finally, the sensitivity on \(\tau\) is more significant and more subtle. A more detailed sensitivity study, focusing on \(\tau\) was performed for 13 sites from the data set. For these sites, all other parameters were fixed to those given in Table I and \(\tau\) was varied between 0.8 and 0.9 in steps of 0.05 (21 SCSF runs per site). Over this range of \(\tau\), SCSF produces reasonable clear sky baseline estimates, while outside this range the model ceases to yield reasonable estimates. The results of this study are summarize in Figure 8. Of the sites included in this study, 6 sites showed variability over this range of less than 0.1%, and 12 of the 13 showed variability of less than 0.25%. Site 09 showed the largest variability, with a total range of 0.3%.

Interestingly, all 13 sites exhibit one of four behaviors over this range of \(\tau\): (1) monotonic increase, (2) monotonic decrease, (3) a maximum value at around \(\tau = 0.85\), and (4) a minimal value around \(\tau = 0.85\). More formally, we find that using \(\tau = 0.85\) as a normalization point (as in Figure 8) approximately minimizes the overall variation in the data. This is illustrated in Figure 9 and provides the argument for using \(\tau = 0.85\) as a default value for fleet-scale analysis.

C. Fleet-scale analysis

The SCSF algorithm was applied “blindly” to the NREL data set using the parameters summarized in Table I. The complete data set contains data for 573 unique systems from across the continental United States. The RdTools approach rejected 186 sites, while the SCSF methodology rejected 22 sites. Only three sites were rejected by both methods. 368 sites were included in both analyses. Table III summarizes these results.
Fig. 8. The dependence of the degradation rate on $\tau$ across 13 sites. Because the differences between sites is larger than the variation within a site, the trends are normalized by subtracting the nominal value, i.e., the degradation rate at $\tau = 0.85$. All systems show one of the following: monotonic increase, monotonic decrease, a maximum at $\tau \approx 0.85$, or a minimum at $\tau \approx 0.85$.

Fig. 9. The residual variability in the degradation rate estimates, after normalizing to different values of $\tau$. This function has a minimum at $\tau \approx 0.85$.

Figure 10 shows the distribution of YOY degradation rates across all systems in the NREL data set from both the RdTools approach and the SCSF approach. The median values of two methods are in very close agreement, validating that the two methods are measuring the same fundamental quantity, in expectation. After excluding outliers (bottom plot), the standard deviation of the SCSF distribution is 1.0% versus 1.4% for RdTools, indicating a reduction in uncertainty of the fleet-scale analysis. An outlier here is defined based on Tukey’s definition [10]: If the quartiles of the data set are $Q_1$ for the first quartile and $Q_3$ for the third quartile, then outliers are points that fall outside the following interval:

$$[Q_1 - 1.5 \cdot (Q_3 - Q_1), Q_3 + 1.5 \cdot (Q_3 - Q_1)]$$

![Table III](image)

**TABLE III**

| Summary of site selection for both methods |
|---------------------|---------------------|---------------------|
| Included Sites      | Excluded Sites      | Unique Excluded Sites |
|---------------------|---------------------|---------------------|
| SCSF                | 551                 | 22                  |
| RdTools             | 387                 | 186                 |

Fig. 10. Comparison of YOY system degradation for all systems, as estimated by RdTools and SCSF. The top plot includes all 368 sites included in both analysis, and the bottom excludes outliers, defined as 1.5 times the interquartile range (332 sites).

Fig. 11. Comparison of SCSF estimates of degradation to RdTools for sites included by both methodologies, with outliers removed (332 sites). The red dashed line is where the two methods agree. Sites that have anomalous, positive estimates from RdTools but negative estimates from SCSF are colored orange.

Figure 11 compares the SCSF estimates to the RdTools estimates. Outliers have been removed, defined as 1.5 times the interquartile range. On a site-by-site basis, the two methods can differ on the order of 2%. Each RdTools estimate should be considered the P50 degradation rate in the context of an associated confidence interval [8], and 85.6% of the sites have an SCSF estimate that is within the confidence bounds for the RdTools approach. There is a significant population of sites corresponding to positive RdTools estimates and negative SCSF estimates, colored in orange. Note that there is not a corresponding population in quadrant 2 (positive SCSF, negative RdTools). The SCSF analysis resulted in positive degradation rates for 20 sites, while the RdTools analysis found positive degradation for 50 sites. For crystalline silicon based systems, such as those considered here, energy output of a system is generally not expected to increase over time, so
positive “degradation” rates typically indicate an error with the analysis or a lack of robustness to data anomalies.

An interesting case study for contrasting the SCSF and RdTools approaches is Example System 3 (ES3). The measured power and fit clear sky model for this system are shown in Figure 12. This system shows a different apparent capacity in the first year than the subsequent four years. This type of data anomaly is quite common and can be caused by someone updating a scale factor in the data logger program or by simply installing more panels (the cause in this case is unknown). Figure 13 illustrates the robustness of the SCSF method to this data anomaly. The clear sky estimator completely ignores the data in the first year, focusing the fit on the later data. Notably, the RdTools approach estimated the degradation rate for this system to be +11.8%, whereas SCSF estimate is approximately zero which appears to be a more reasonable estimate of YOY degradation for this system.

IV. CONCLUSION

We present an extension to Statistical Clear Sky Fitting \[6\] that allows for efficient estimation of system YOY degradation rate. This approach to estimating degradation requires no model of the site nor any estimates of irradiance or weather data. For this reason, it lends itself naturally to fleet-scale analysis of heterogeneous PV systems, where such supplementary data may be missing or incorrect.

We show that fitting a periodic data model to measured PV power data requires the inclusion of a degradation term, and we demonstrate the sensitivity of that degradation term to model tuning parameters. The SCSF approach to degradation estimation is at least as accurate, on average, as the RdTools approach. In addition, the analysis of this fleet of PV systems suggests that the SCSF approach yields improved accuracy by reducing the variance of the degradation rates across all systems and by reducing the number of systems with non-negative degradation rates.

The ability to perform model-less estimation of bulk PV system degradation rates presents a large opportunity for evaluating the value proposition of fleet-scale collections of distributed rooftop systems. This algorithm enables degradation analysis for a broad set of systems where data to supplement the system power signals is unavailable, unreliable, or expensive to procure. In addition, the SCSF method requires no engineering time to model the system, merge PV power and weather data sets, or filter data. The approach is naturally robust to missing and bad power data, handling common anomalies such as missing data and changes in apparent system capacity.

The SCSF algorithm provides not only an estimate of system degradation but also an estimate of the clear sky response of the system, which can be thought of as an optimal baseline capturing the cyclostationarity of the observed time-series power signal. In addition, the biconvex structure of the SCSF approach lends itself to efficient computation and scalability. SCSF represents a fundamentally new way of modeling and understanding PV power data sets, exploiting signal processing techniques and periodicity rather than physical models to understand system behavior. The software implementation of SCSF is available at https://github.com/slacgismo/StatisticalClearSky Data cleaning and preprocessing was automated with software available at https://github.com/slacgismo/solar-data-tools.

ACKNOWLEDGMENTS

Thank you to Laura Schelhas and Elpiniki Apostolaki Iosifidou, who collaborate on the PVInsight project, and to the rest of the Gismo team at SLAC National Accelerator Laboratory. Additional thanks to Prof. Stephen Boyd at Stanford University for the support and inspiration on this work. This material is based upon work supported by the U.S. Department of Energys Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Numbers 34911 (“PVInsight”), 30311, and 34348. This work was coauthored by Alliance for
Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory for the US Department of Energy (DOE) under Contract No. DEAC3608GO28308. The views expressed in the article do not necessarily represent the views of the DOE or the US Government. The US Government retains and the publisher, by accepting the article for publication, acknowledges that the US Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for US Government purposes.

REFERENCES

[1] A. Phinikarides, N. Kindyni, G. Makrides, and G. E. Georghiou, “Review of photovoltaic degradation rate methodologies,” Renewable and Sustainable Energy Reviews, vol. 40, pp. 143–152, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.rser.2014.07.155

[2] E. Hasselbrink, M. Anderson, Z. Defreitas, M. Mikofski, Y. C. Shen, S. Caldwell, A. Terao, D. Kavulak, Z. Campeau, and D. Degraaff, “Validation of the PVLife model using 3 million module-years of live site data,” Conference Record of the IEEE Photovoltaic Specialists Conference, pp. 7–12, 2013.

[3] D. C. Jordan, C. Deline, S. R. Kurtz, G. M. Kimball, and M. Anderson, “Robust PV Degradation Methodology and Application,” IEEE Journal of Photovoltaics, vol. 8, no. 2, pp. 525–531, 2018.

[4] “RdTools v1.2.2.” [Online]. Available: https://github.com/NREL/rdtools

[5] D. L. King, W. E. Boyson, and J. A. Kratochvil, “Photovoltaic array performance model,” Sandia Report No. 2004-3535, vol. 8, no. December, pp. 1–19, 2004. [Online]. Available: http://prod.sandia.gov/techlib/access-control.cgi/2004/043535.pdf

[6] B. Meyers, M. Tabone, and E. C. Kara, “Statistical Clear Sky Fitting Algorithm,” Conference Record of the IEEE Photovoltaic Specialists Conference, pp. 1–6, 2018. [Online]. Available: http://www.wcpec7.org/eWCPEC/manuscripts/MeWCPEC930_0521112519.pdf

[7] M. Udell, C. Horn, R. Zadeh, and S. Boyd, “Generalized Low Rank Models,” Foundations and Trends in Machine Learning, vol. 9, no. 1, pp. 1–118, 2016. [Online]. Available: https://web.stanford.edu/~boyd/papers/glrm.html

[8] M. Deceglie, D. C. Jordan, A. Nag, A. Shinn, and C. Deline, “Fleet-scale energy-yield degradation analysis applied to hundreds of residential and nonresidential photovoltaic systems,” IEEE Journal of Photovoltaics, vol. PP, pp. 1–7, 01 2019.

[9] J. Gorski, F. Pfeuffer, and K. Klamroth, “Biconvex sets and optimization with biconvex functions: A survey and extensions,” Mathematical Methods of Operations Research, vol. 66, no. 3, pp. 373–407, 2007.

[10] J. W. Tukey, Exploratory data analysis, ser. Addison-Wesley series in behavioral sciences. Reading, MA: Addison-Wesley Pub. Co., 1977.