






Solar Data Tools: a Python library for automated analysis of unlabeled PV data

Sara A. Miskovich¹[¶], Bennet E. Meyers¹[¶], Elpiniki Apostolaki-Iosifidou¹, Claire Berschauer¹, Chengcheng Ding³, Aramis Dufour², David Jose Florez Rodriguez³, Jonathan Goncalves¹, Alejandro Londono-Hurtado¹, Victor-Haoyang Lian³, Tristan Lin³, Junlin Luo³, Xiao Ming³, Duncan Ragsdale¹, Derin Serbetcioglu¹, Shixian Sheng³, Jose St Louis³, Tadatoshi Takahashi¹, Nimish Telang¹, Mitchell Victoriano¹, Haoxi Zhang³, and Nimish Yadav³

¹ SLAC National Accelerator Laboratory, Menlo Park, CA, 94025, USA ² Stanford University, Stanford, CA, 94305, USA ³ Carnegie Mellon University, Pittsburgh, PA 15213, USA [¶] Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

Effectively processing and leveraging the growing volume of photovoltaic (PV) system performance data is essential for the operation and maintenance of PV systems globally. However, many distributed rooftop PV systems suffer from lower data quality, are difficult to model, and lack access to reliable environmental data.

Solar Data Tools is an open-source Python library designed for automated data quality and loss factor analysis of *unlabeled* PV time-series data, i.e. without requiring a system model, meteorological data, or performance indices. Solar Data Tools empowers PV system operators and fleet owners to better understand their system's performance using only basic power output data.

Solar Data Tools is user-friendly, requiring minimal setup, and is compatible with all types of PV systems, from small rooftop generators to large utility power plants. Using advanced signal decomposition techniques (B. E. Meyers & Boyd, 2023), the library enables the performance and reliability analysis of large volumes of PV power time-series data across various formats and quality levels. It eliminates the need for site-specific meteorological inputs or pre-defined system models, simplifying the analysis process. This library can be valuable for a wide range of users that work with unlabeled solar power data, including professionals in the private solar industry or utility companies, researchers and students in the solar energy field, community solar owners, and rooftop PV system owners.

Solar Data Tools is developed openly on GitHub (B. Meyers & others, 2024e) under the permissive BSD-2 license and is available to install via the Python Package Index (PyPI) (B. Meyers & others, 2024c) and the conda-forge repository (B. Meyers & others, 2024b). The library is actively maintained and contributions from the community are welcome. The documentation is hosted on Read the Docs (B. Meyers & others, 2024a). More detailed information about Solar Data Tools, its algorithms and its features can be found in the PVSC 2020 (B. E. Meyers et al., 2020) and PVSC 2024 (Miskovich & Meyers, 2024) papers.

Statement of need

With the growing number of photovoltaic (PV) systems worldwide, it is crucial to have tools that can process and analyze data from systems of all sizes and configurations. The data typically consists of time-series measurements of real power production, reported as average power over intervals ranging from one minute to one hour, spanning several years and possibly containing missing entries.

Historically, PV data analysis tools have focused on data combined with local meteorological measurements and system configuration information, such as those from large power plants. Data cleaning tasks have largely been manual, and analyses have relied on metrics like the performance index (Townsend et al., 1994), which require accurate site models and meteorological data. For smaller, distributed rooftop PV systems, meteorological information is often lacking, making accurate system modeling difficult. For such systems, insights must be derived from just the PV power data in isolation (referred to as *unlabeled* data), for which forming a performance index is difficult or impossible. Given that distributed rooftop PV systems accounted for over 40% of the installed capacity in 2020 (Davis et al., 2021), there is a clear need for automated and model-free data processing and analysis tools that enable remote monitoring of system health and optimization of operations and maintenance activities of these systems.

Solar Data Tools (SDT) (B. E. Meyers et al., 2020; B. Meyers & others, 2024d) is an open-source Python library designed for the automatic processing and analysis of unlabeled PV data signals. SDT automates the cleaning, filtering, and analysis of PV power data, including loss factor estimation, eliminating the need for user configuration or “babysitting” regardless of data quality or system configuration. It is suitable for a wide range of applications and system types, from large utility-scale trackers to small, multi-pitch rooftops. SDT provides practical tools for both small and fleet-scale PV performance analyses without requiring the calculation of performance indices for each system. It may be used to fully automate a quality and loss factor estimation pipeline, or it may be used to onboard, visualize, and explore new data or prepare data for a custom analysis.

The software has been used by researchers at Stanford University (Ogut et al., 2024), Case Western Reserve University (Pierce et al., 2024), and LBNL (Li et al., 2023, 2024) to prepare data in the development data-driven performance models for the solar PV domain. In addition, the software has been used in loss factor estimation intercomparison studies by NREL (Perry et al., 2023) and the IEA (Lindig et al., 2021).

Two other libraries offer similar data analysis tools for solar applications: PVAnalytics (Perry et al., 2022) and RdTools (Deceglie & others, 2024). Unlike SDT, these libraries are model-driven and require users to define their own analyses. PVAnalytics focuses on preprocessing and quality assurance, while RdTools specializes in loss factor analysis. SDT, on the other hand, provides both data quality and loss factor analysis, operates *automatically* with minimal setup, and is **model-free**, requiring no weather or other external information. To our knowledge, no other open-source software provides flexible model-free automated analysis for unlabeled PV data. SDT is particularly suited for users who need a pre-defined pipeline to analyze complex systems that cannot be easily modeled and lack meteorological data—a common scenario for small, distributed systems.

Acknowledgements

This work is supported by the U.S. Department of Energy’s Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technologies Office Award Numbers 34368 and 38529.

References

- 85
- 86 Davis, M., Smith, C., White, B., Goldstein, R., Sun, X., Cox, M., Curtin, G., Manghani, R.,
87 Rumery, S., Silver, C., & Baca, J. (2021). *U.S. Solar market insight executive summary,*
88 *2020 year in review.* Wood Mackenzie; Solar Energy Industries Association.
- 89 Deceglie, M., & others. (2024). *NREL/rdtools: Version 3.0.0-alpha.6* (Version v3.0.0).
90 <https://doi.org/10.5281/zenodo.13356337>
- 91 Li, B., Chen, X., & Jain, A. (2024). Power modeling of degraded PV systems: Case studies
92 using a dynamically updated physical model (PV-pro). *Renewable Energy*, 236, 121493.
93 <https://doi.org/https://doi.org/10.1016/j.renene.2024.121493>
- 94 Li, B., Karin, T., Meyers, B. E., Chen, X., Jordan, D. C., Hansen, C. W., King, B. H., Deceglie,
95 M. G., & Jain, A. (2023). Determining circuit model parameters from operation data for PV
96 system degradation analysis: PVPRO. *Solar Energy*, 254, 168–181. <https://doi.org/https://doi.org/10.1016/j.solener.2023.03.011>
- 97
- 98 Lindig, S., Moser, D., Curran, A. J., Rath, K., Khalilnejad, A., French, R. H., Herz, M.,
99 Müller, B., Makrides, G., Georghiou, G., Livera, A., Richter, M., Ascencio-Vásquez, J.,
100 Iseghem, M. van, Meftah, M., Jordan, D., Deline, C., Sark, W. van, Stein, J. S., ... Luo,
101 W. (2021). International collaboration framework for the calculation of performance loss
102 rates: Data quality, benchmarks, and trends (towards a uniform methodology). *Progress*
103 *in Photovoltaics: Research and Applications*, 29(6), 573–602. <https://doi.org/https://doi.org/10.1002/pip.3397>
- 104
- 105 Meyers, B. E., Apostolaki-Iosifidou, E., & Schelhas, L. T. (2020). Solar data tools: Automatic
106 solar data processing pipeline. *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*,
107 0655–0656. <https://doi.org/10.1109/PVSC45281.2020.9300847>
- 108 Meyers, B. E., & Boyd, S. P. (2023). Signal decomposition using masked proximal operators.
109 *Foundations and Trends® in Signal Processing*, 17, 1–78. [https://doi.org/10.1561/](https://doi.org/10.1561/20000000122)
110 [20000000122](https://doi.org/10.1561/20000000122)
- 111 Meyers, B., & others. (2024a). *Solar data tools documentation.* [https://solar-data-tools.](https://solar-data-tools.readthedocs.io/)
112 [readthedocs.io/](https://solar-data-tools.readthedocs.io/)
- 113 Meyers, B., & others. (2024b). *Solar data tools on anaconda.* [https://anaconda.org/](https://anaconda.org/conda-forge/solar-data-tools)
114 [conda-forge/solar-data-tools](https://anaconda.org/conda-forge/solar-data-tools)
- 115 Meyers, B., & others. (2024c). *Solar data tools on PyPi.* [https://pypi.org/project/](https://pypi.org/project/solar-data-tools/)
116 [solar-data-tools/](https://pypi.org/project/solar-data-tools/)
- 117 Meyers, B., & others. (2024d). *Slacgismo/solar-data-tools: v1.2.2* (Version v1.2.2). Zenodo.
118 <https://doi.org/10.5281/zenodo.10888385>
- 119 Meyers, B., & others. (2024e). *Solar data tools* (Version v1.2.2). [https://github.com/](https://github.com/slacgismo/solar-data-tools)
120 [slacgismo/solar-data-tools](https://github.com/slacgismo/solar-data-tools)
- 121 Miskovich, S. A., & Meyers, B. (2024). Automated analysis of unlabeled PV data with solar
122 data tools software: Overview and feature updates. *2024 IEEE 52nd Photovoltaic Specialist*
123 *Conference (PVSC)*, 1728–1730. <https://doi.org/10.1109/PVSC57443.2024.10749597>
- 124 Ogut, G., Meyers, B., Dufour, A., & Boyd, S. (2024). Time dilated Bundt cake analysis
125 of PV output. *2024 IEEE 52nd Photovoltaic Specialist Conference (PVSC)*, 877–883.
126 <https://doi.org/10.1109/PVSC57443.2024.10749393>
- 127 Perry, K., Meyers, B., & Muller, M. (2023). *Survey of time shift detection algorithms for*
128 *measured PV data.* National Renewable Energy Laboratory (NREL), Golden, CO (United
129 States). <https://www.osti.gov/biblio/1990039>
- 130 Perry, K., Vining, W., Anderson, K., Muller, M., & Hansen, C. (2022). *PVAnalytics: A python*

- 131 *package for automated processing of solar time series data.* [https://www.osti.gov/biblio/](https://www.osti.gov/biblio/1887283)
132 [1887283](https://www.osti.gov/biblio/1887283)
- 133 Pierce, B. G., Wieser, R. J., Ciardi, T. G., Yao, A. D., French, R. H., Bruckman, L. S., &
134 Li, M. (2024). Comparison of empirical and data driven digital twins for a PV+battery
135 fleet. *2024 IEEE 52nd Photovoltaic Specialist Conference (PVSC)*, 1391–1397. [https:](https://doi.org/10.1109/PVSC57443.2024.10749422)
136 [//doi.org/10.1109/PVSC57443.2024.10749422](https://doi.org/10.1109/PVSC57443.2024.10749422)
- 137 Townsend, T., Whitaker, C., Farmer, B., & Wenger, H. (1994). New performance index for
138 PV system analysis. *Conference Record of the IEEE Photovoltaic Specialists Conference, 1*,
139 1036–1039. <https://doi.org/10.1109/wcpec.1994.520138>

DRAFT