The Journal of Open Source Software

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

Sara A. Miskovich[1¶], Bennet E. Meyers[1¶], Elpiniki Apostolaki-Iosifidou[1], Claire Berschauer[1], Chengcheng Ding[3], Aramis Dufour[2], David Jose Florez Rodriguez[3], Jonathan Goncalves[1], Alejandro Londono-Hurtado[1], Victor-Haoyang Lian[3], Tristan Lin[3], Junlin Luo[3], Xiao Ming[3], Duncan Ragsdale[1], Derin Serbetcioglu[1], Shixian Sheng[3], Jose St Louis[3], Tadatoshi Takahashi[1], Nimish Telang[1], Mitchell Victoriano[1], Haoxi Zhang[3], and Nimish Yadav[3]

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

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

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. The primary objective of Solar Data Tools is to empower PV system operators or fleet owners to 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 solar systems. Using advanced signal decomposition techniques, 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.

Statement of need

With the growing number of real-world installations 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 (**?**), 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. Meyers et al., 2020; Bennet Meyers & others, 2024) 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 systems, 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.

Two other libraries offer similar data analysis tools for solar applications: PVAnalytics (Perry et al., 2022) and RdTools (**?**). 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. 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. (cite tutorial here for more info?)

Figures

Add example plots (heatmaps, loss analysis, what else?)

Figures can be included like this: Caption for example figure. and referenced from text using section .

Figure sizes can be customized by adding an optional second parameter: Caption for example figure.

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 Number 38529.

References

Davis, M., Smith, C., White, B., Goldstein, R., Sun, X., Cox, M., Curtin, G., Manghani, R., Rumery, S., Silver, C., & Baca, J. (2021). *U.S. Solar market insight executive summary, 2020 year in review*. Wood Mackenzie; SEIA.

Meyers, B., Apostolaki-Iosifidou, E., & Schelhas, L. (2020). *Solar data tools: Automatic solar data processing pipeline* (pp. 0655–0656). https://doi.org/10.1109/PVSC45281.2020.9300847

Meyers, Bennet, & others. (2024). *Slacgismo/solar-data-tools: v1.2.2* (Version v1.2.2). Zenodo. https://doi.org/10.5281/zenodo.10888385

Perry, K., Vining, W., Anderson, K., Muller, M., & Hansen, C. (2022). *PVAnalytics: A python package for automated processing of solar time series data*. https://www.osti.gov/biblio/1887283