User Guide for *Tracking the Sun* Public Data File

Version Date: September 10, 2019

Background

Lawrence Berkeley National Laboratory (Berkeley Lab) collects project-level data on residential and non-residential photovoltaic (PV) systems for its annual *Tracking the Sun* report. The data are sourced primarily from state agencies and utilities that administer PV incentive programs, solar renewable energy credit registration systems, or interconnection processes. In order to leverage this dataset for broader use, Berkeley Lab has issued a public data file, which can be downloaded at http://trackingthesun.lbl.gov.[1] The public project-level dataset will be updated once annually with data from the previous calendar year, and may also be updated on an interim basis as improvements to the data cleaning methodology and supplementary data fields are developed.

What is Included in the Public Data File?

The data file includes only grid-connected residential and non-residential PV systems, defined to consist of rooftop systems, regardless of size, and ground-mounted systems up to 5 MW_{AC}. Ground-mounted projects larger than 5 MW_{AC} are considered utility-scale and are not included in this dataset.

The current version of the public data file includes more than 1,500,000 PV systems installed through year-end 2018. The file includes more than 60 data fields describing key attributes of each system, which are listed and described in the table below. Note, though, that most fields are incomplete for most systems. The public data file is a single zip file, which contains two CSV files as well as this User Guide. The data has been separated into two csv files (part 1 and part 2) so that each can be opened in Excel.

What Data Cleaning Operations Are Performed?

The data collected for *Tracking the Sun* undergo extensive cleaning and quality control. Some elements of those operations are described in the table below. For additional information, please refer to Section 2 of the latest *Tracking the Sun* report ("Data Sources, Methods, and Sample Description") and to Appendix A.

One important convention should be noted: Missing data are coded in the database as -9999. Any operations performed on the data should therefore treat such values accordingly.

Who to Contact with Questions?

Questions or comments specifically about the *Tracking the Sun* public data file may be directed to either Naïm Darghouth (ndarghouth@lbl.gov) or Galen Barbose (glbarbose@lbl.gov).

[1] The public data file excludes any data provided under confidentiality agreements as well as other sensitive information that data providers requested to be withheld.

Data Fields in the Public Data File

Data Field Name	Units	Description, Key Notes, and Potential Data Quality Issues
Data Provider	n/a	The entity that supplied the data, generally a utility or PV incentive program administrator
System ID (from first Data Provider)	n/a	This is the system or application ID within the raw data file from the first data provider, if applicable.
System ID (from second Data Provider, if applicable)	n/a	This is the system or application ID within the raw data file from the second data provider, if applicable. A second system ID indicates that PV system data has been merged from two data providers.
System ID (Tracking the Sun)	n/a	This is the system or application ID created within Berkeley Lab's Tracking the Sun database. This ID number will be stable in future releases of Tracking the Sun.
Installation Date	date	For some data providers, the installation date may be based on the best available proxy, such as the date that an incentive claim was submitted or when the inspection was performed.
System Size	kw (DC)	The total rated direct-current (DC) output of the module arrays at standard test conditions. These data are generally reported directly by the data provider, but in some cases must be estimated, for example, based on the module model and quantity or based on reported alternating-current (AC) capacity.
Total Installed Price	dollars (nominal)	The total installed price for the system, prior to receipt of any incentives, as reported by the installer, host customer, or other incentive applicant. For third-party owned systems, the data may represent one of two things. If the third-party owner procured the system from an independent installation contractor, then the reported installed price likely refers to the intermediate sale price between the installation contractor and the third-party owner. If the third-party owner instead installed the system itself, then the reported installed price likely represents an appraised value. The installed price data may be subject to any number of other reporting inconsistencies, which may or may not be readily identifiable. In addition, the data may suffer simply from self-reporting errors, and the level of verification vary across data providers.
Appraised Value Flag	n/a	A flag used to indicate whether the reported installed price is likely to represent an appraised value. Caution should be used in relying on appraised values for analysis or benchmarking purposes, as such data do not represent a transaction price.
Sales Tax Cost	dollars (nominal)	The calculated cost of sales taxes. This is estimated based on average sales tax rates for the given state and year, accounting for any sales tax exemptions that may exist for PV systems. Sales taxes, if applicable, are assumed to be levied only on hardware costs, which are assumed to represent 55% of the total installed price.
Rebate or Grant	dollars (nominal)	The pre-tax value of any up-front rebate or grant provided by the entity supplying the data

Data Field Name	Units	Description, Key Notes, and Potential Data Quality Issues
Performance-Based Incentive (Annual Payment)	dollars (nominal)	Data reported by data providers generally consists of either the estimated annual PBI payment of the nominal sum of PBI payments over the full incentive term. In some cases, only the PBI rate ($/kWh) and PBI term are available, in which case the annual PBI payment is calculated based on estimated insolation levels and first-year energy production. PBI payment amounts are reported on a pre-tax basis.
Performance-Based Incentives (Duration)	years	Number of years that PBI payments are disbursed.
Feed-in Tariff (Annual Payment)	dollars (nominal)	The estimated pre-tax annual feed-in tariff (FIT) payment received in the first year of the FIT contract term. The calculation procedure mirrors that described above for PBI payments.
Feed-in Tariff (Duration)	years	Contract term length.
Customer Segment	n/a	Data on customer segment is mapped to one of six general types: RES, COM, SCHOOL, GOV, NON-PROFIT, and NON-RES, the last one being used only if more-specific information on non-residential customer type is unavailable.
New Construction	n/a	Indicates if the system was installed at the time of building construction. Data generally available for only those states or utilities that have separate programs or incentive rates for new construction vs. retrofits.
Tracking	n/a	Indicates if the system includes tracking equipment
Ground Mounted	n/a	Indicates if the system is ground-mounted (which may include pole-mounted systems). PV systems consisting of a combination of rooftop and ground-mounted arrays are coded as ground-mounted.
Battery System	n/a	Indicates if the system includes batteries
Zip Code	n/a	Host customer zip code
City	n/a	Host customer city. Spellings have not been corrected or standardized.
State	n/a	Host customer state
Utility Service Territory	n/a	The electric utility service territory, when reported directly by the data provider; this data has not been cleaned or standardized.
Third-Party Owned	n/a	Indicates if the system is third-party owned; that is, owned by an entity other than the site host and either leased or sold under a power purchase agreement to the site host.
Installer Name	n/a	Reported data, prior to being cleaned, is particularly "messy" given the complex spellings of models. These data have been cleaned and the spellings standardized to the extent feasible.
Self-Installed	n/a	Indicates if the system was installed by the site-host.
Azimuth #1	degrees	The horizontal direction of the array, where 180 degrees defines South facing PV orientation. Azimuth data reported by data providers was, in some cases, modified to adhere to this convention. Data fields are provided for up to three array orientations, though some systems may consist of a larger set of distinct orientations.
Azimuth #2	degrees	
Azimuth #3	degrees	
Tilt #1	degrees	The vertical tilt of the array, where zero degrees corresponds to a flat array. As with the azimuth data, fields are provided for
Tilt #2	degrees	

Data Field Name	Units	Description, Key Notes, and Potential Data Quality Issues
Tilt #3	degrees	up to three array orientations, though some systems may consist of a larger set of distinct orientations.
Module Manufacturer #1	n/a	Reported data, prior to being cleaned, is particularly "messy" given the complex spellings of models. These data have been cleaned and the spellings standardized to the extent feasible.
Module Model #1	n/a	
Module Manufacturer #2	n/a	
Module Model #2	n/a	
Module Manufacturer #3	n/a	
Module Model #3	n/a	
Additional module model	n/a	Indicates whether there are more than three module models.
Module Technology #1	n/a	Identifies the module technology type. This is determined by cross-referencing module manufacturer and model names against equipment specification data available through solarhub.com and the California Energy Commission's list of eligible equipment.
Module Technology #2	n/a	
Module Technology #3	n/a	
BIPV Module #1	n/a	Indicates if the modules are building integrated photovoltaics (BIPV). This is determined by cross-referencing module manufacturer and model names against equipment specification data available through solarhub.com and the California Energy Commission's list of eligible equipment.
BIPV Module #2	n/a	
BIPV Module #3	n/a	
Module Efficiency #1	percent	Identifies the energy conversion efficiency of the modules. This is determined by cross-referencing module manufacturer and model names against equipment specification data available through solarhub.com and the California Energy Commission's list of eligible equipment.
Module Efficiency #2	percent	
Module Efficiency #3	percent	
Inverter Manufacturer #1	n/a	Reported data, prior to being cleaned, is particularly "messy" given the complex spellings of models. These data have been cleaned and the spellings standardized to the extent feasible.
Inverter Manufacturer #2	n/a	
Inverter Manufacturer #3	n/a	
Inverter Model #1	n/a	
Inverter Model #2	n/a	
Inverter Model #3	n/a	
Additional inverter models	n/a	Indicates whether there are more than three inverter models.
Microinverter #1	n/a	Indicates if the inverter modules identified are micro-inverters. This is determined by cross-referencing inverter manufacturer and model names against equipment specification data available through solarhub.com and the California Energy Commission's list of eligible equipment.
Microinverter #2	n/a	
Microinverter #3	n/a	
System Inverter Capacity	kW (AC)	The total rated alternating-current (AC) output of the inverter(s). These data are sometimes reported directly by the data provider, but in most cases must be estimated based on the inverter model and quantity.
DC Optimizer	n/a	Indicates if the system uses any DC Optimizers, based on the inverter manufacturer names. All systems using SolarEdge inverters are assumed to also include a DC optimizer. Systems using DC optimizers manufactured by other companies (e.g., Tigo) cannot be identified based on the inverter manufacturer; as such, the DC Optimizer field is coded as unknown for all systems with string inverters manufactured by companies others than SolarEdge.

Data Field Name	Units	Description, Key Notes, and Potential Data Quality Issues
Inverter Loading Ratio	n/a	The Inverter Loading Ratio is the ratio of DC module capacity to AC inverter capacity. This is calculated from the reported or estimated values for System Size and System Inverter Capacity, described above.

Disclaimer