Generating Synthetic Time Series Photovoltaic Data with Real-World Physical Challenges and Noise for Use in Algorithm Test and Validation

Matthew Muller, Kevin Anderson, and Michael Deceglie

National Renewable Energy Laboratory

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-5K00-86459
September 2023

Generating Synthetic Time Series Photovoltaic Data with Real-World Physical Challenges and Noise for Use in Algorithm Test and Validation

Matthew Muller,[1] Kevin Anderson,[1] and Michael Deceglie[1]

National Renewable Energy Laboratory

Suggested Citation

Muller, Matthew, Kevin Anderson, and Michael Deceglie. 2023. *Generating Synthetic Time Series Photovoltaic Data with Real-World Physical Challenges and Noise for Use in Algorithm Test and Validation*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-5K00-86459. https://www.nrel.gov/docs/fy23osti/86459.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-5K00-86459
September 2023

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Numbers 34348. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via www.OSTI.gov.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

Preface

The PV Fleet Data Initiative and other projects seek to develop algorithms for automated analysis of photovoltaic (PV) time series data. This analysis can be used to extract statistical information and other parameters from the data, such as degradation rates, soiling loss information, tracker performance, clipping or curtailment, and system availability. Although there is a vast body of PV data available for application of these extraction algorithms, it is difficult to validate these algorithms, because the true parameters to be extracted are not known. Synthetic data has been widely used in the literature for algorithm validation, but this synthetic data is typically very bounded by the problem or topic at hand. The PV Fleet Data Initiative has demonstrated that real time series PV data almost always includes a host of data quality and physical problems that, in reality, any automated PV abstraction algorithm must handle appropriately. For this reason, this work describes the development of a complex synthetic PV time series data set that includes data quality and physical problems that have been experienced in real-world PV data. The quality and physical problems are documented in the synthetic data so that users can test the validity of various existing PV extraction algorithms as well as develop new algorithms to solve problems this data set can support. The synthetic dataset used in this report can be accessed on the Durable Module Materials Consortium Data Hub at https://doi.org/10.21948/1999772.

List of Acronyms

CDF	cumulative distribution function
GHI	global horizontal irradiance
GW	gigawatts
IAM	incidence angle modifier
ILR	inverter loading ratio
ISO	International Organization for Standardization
NREL	National Renewable Energy Laboratory
NSRDB	National Solar Radiation Database
PI	performance index
POA	plane of array
PSM3	Physical Solar Model version 3
PV Fleets	NREL PV Fleet Data Initiative
PV	photovoltaic
QA	quality assurance
Rd	rate of degradation

Table of Contents

List of Figures

List of Tables

Introduction

As photovoltaic (PV) capacity has grown in recent decades, so has the body of time series data associated with these installations. As a result of this data growth, it has become unreasonable to expect that individual human analysts can examine each PV system's time series data set to study performance or other system characteristics. For example, the National Renewable Energy Laboratory's (NREL's) PV Fleet Data Initiative (PV Fleets) has ingested 8.4 gigawatts (GW) of PV systems, resulting in more than 36.8 billion rows of time series data. This data set contains valuable information about the U.S. PV fleet, such as degradation rate statistics, tracker performance, soiling losses, system availability, and other factors. However, extracting this information requires the development of automated algorithms that have been validated against their intended purpose. Although much progress has been made in algorithm development and validation, one of the biggest challenges is the lack of ground truth data sets in the validation phase [1]–[5]. Where possible, NREL has employed expert analysts to label time series data with information like inverter clipping, cleaning events, and stuck tracker events, and has made these labeled data sets publicly available [6]–[9]. However, analyst-labeled data sets are not ideal, as they are time-consuming to create and are often not broadly used because they leave room for ambiguity and debate. For this reason, many works in the literature have created focused synthetic data sets geared toward validation of a specific algorithm or model. These synthetic data sets have been quite valuable, but the PV Fleets real-world data set indicates that these synthetic data are often oversimplified. Real-world data presents countless compounding challenges that automated algorithms must be able to navigate if valid information is to be extracted from large-scale PV time series data. Therefore, through this work, we seek to develop and publish a complex PV time series data set that includes both data quality and physical problems that have been experienced in the U.S. PV fleet through the PV Fleets project. The paper is divided as follows. First, a section on real-world data provides background on the challenges and physical problems found in the PV Fleets data set. Then, a section on modeling and methodology provides the process for generating the synthetic data set. Next, a results section provides representative examples of the synthetic data, and finally, brief conclusions are made.

The synthetic dataset used in this report can be accessed on the Durable Module Materials Consortium Data Hub at https://doi.org/10.21948/1999772.

1 Real-World Data

1.1 Data Quality Issues

PV data is often considered through a perfect lens; in other words, the time series data for irradiance, power, temperature, and other parameters is measured without error and passed to an analyst fully intact. Furthermore, there is an assumption that what the data represent is clearly communicated, and that all the necessary metadata are available to develop an in-depth performance model for a given system. The reality is far from this, with countless problems being the norm and not the exception. In the PV Fleets project, significant efforts are made in the data ingestion step to validate individual data stream labels, units, time zone and timestamp conventions, and system metadata. Because problems are still likely, all data is then subjected to an automated quality assurance (QA) analysis, where data streams that don't meet required thresholds are rejected for further analysis [10]. Some common problems are: (1) data is mislabeled (for example, plane of array (POA), irradiance, and global horizontal irradiance (GHI) are swapped), (2) units are incorrect or shift somewhere in the data stream, (3) daylight savings shifts are present although not expected, (4) time frequency changes or left versus right time labeling are unclear, (5) large unexpected shifts occur within an individual time series, (6) significant portions of the data are missing, and (7) system orientation (fixed azimuth, tilt, or tracking) is incorrect or missing. Although QA algorithms are intended to correct or remove the poorest data, problems will persist in data subjected to degradation, soiling, availability, or other desired analysis. We closely examined the daily performance index (PI) for 47 systems that passed QA and found that three challenges persisted: (1) entire days or periods with missing data, (2) residual seasonality in the PI that had an amplitude of a few percent to as high 40 percent, and (3) obvious data shifts that still occurred on the order of tens of percent. Note that in this report, we use PI loosely to refer measured PV output divided my modeled PV output. For PV Fleets data the PI is generated on an aggregated daily basis using the RdTools™ workflow while for synthetic data a PI is generated on a 15-minute basis using methods described in section 2. Analysis of the PV Fleets PI data showed that data was missing for a mean length of 3.1 days per outage, with a standard deviation of 10.5 days, and there were a mean of 8.6 outages per year, with a standard deviation of 8.3 outages. Although residual seasonality can occur for a number of potentially compounding reasons, the most common reason is suspected to be due to misalignment between the array orientation and irradiance measurements (whether the misalignment is real or due to mislabeling of a data stream or array metadata). PI data shifts can also occur for a number of reasons, such as incorrect calibration, mis-entry of calibration coefficients, or physical reasons, such as string outages or plant curtailment. Although the PV Fleets data set does not contain sufficient information to identify the cause of such data shifts, their occurrence can have obvious impacts on degradation and soiling analysis, and therefore they are worth considering in synthetic data generation.

1.2 Physical Data Issues

Here, "physical data issues" mean physical occurrences that directly alter the performance of a PV system, either rapidly or over time. Physical issues can at times be difficult to distinguish from data quality issues. For example, seasonality in the PI signal can be due to a data quality issue, such as incorrect labeling of an irradiance sensor, or a physical issue, such as inaccurate alignment between the irradiance sensor and the array plane, seasonal shading, temperature variation, or spectral effects. When creating a representative synthetic PV data set, it is not critical that all issues be successfully separated into data quality versus physical issues, but rather that common issues be included in the final time series. The following is a list of common physical issues seen in the PV Fleets data that were taken into consideration when generating synthetic data. Snow coverage of PV panels is another physical issue, but it is not included here because the current synthetic data set is for the southern United States, where snow is an infrequent issue.

- Soiling in dry environments
- Pollen or bio-soiling in rainy environments
- String outages
- Utility curtailment
- AC inverter clipping
- Permanent PV system degradation
- Tracker stalls
- Misalignment between array and irradiance sensor
- System self-shading.

1.3 Data Noise

Noise is often considered a corruption or distortion to the underlying true data signal or measurement. In most synthetic PV data sets that have been reported in the literature, noise is either ignored or the noise affecting each data point is randomly sampled from a normal or skewed distribution [11]–[14]. The skewed distribution is justified in that real PV PI data shows noise biased to underperformance rather than overperformance. Observations from the PV Fleets data set suggest that what is being deemed noise is not random and that further consideration is needed to appropriately account for noise. When looking at PV Fleets data to quantify noise, we calculate it on a daily basis. N_{daily} is the difference between the daily PI and the 14-day rolling median of the PI. This daily metric for noise allows us to account for trends in the data that may occur due to soiling losses or other unknown issues that would not be accounted for by a basic performance model. Figure 1 (system 5007), Figure 2 (system 7306), and Figure 3 (system 7316) show different time series plots of PI data (in blue), the 14-day rolling median of PI (in black), and daily rainfall (in green) for real PV systems. System 5007 is a site with regular rainfall, and it shows a PI that appears evenly scattered along the time series with a minor bias toward underperformance or negative N_{daily}. System 7306 has intermittent rainfall; during long dry periods, there are downward soiling trends that have minimal N_{daily}, whereas in late winter, when rain does occur, there is negative N_{daily} (often reaching −40%). System 7316 also has intermittent rainfall, but N_{daily} is significantly less than in system 7306 across the data set—although N_{daily} of system 7316 is still the most negative in the late winter and early spring. Figure 4 shows the N_{daily} for the 47 systems previously mentioned plotted against the normalized daily insolation total. It is clear from this graph that the range of noise decreases as the daily insolation totals increase.

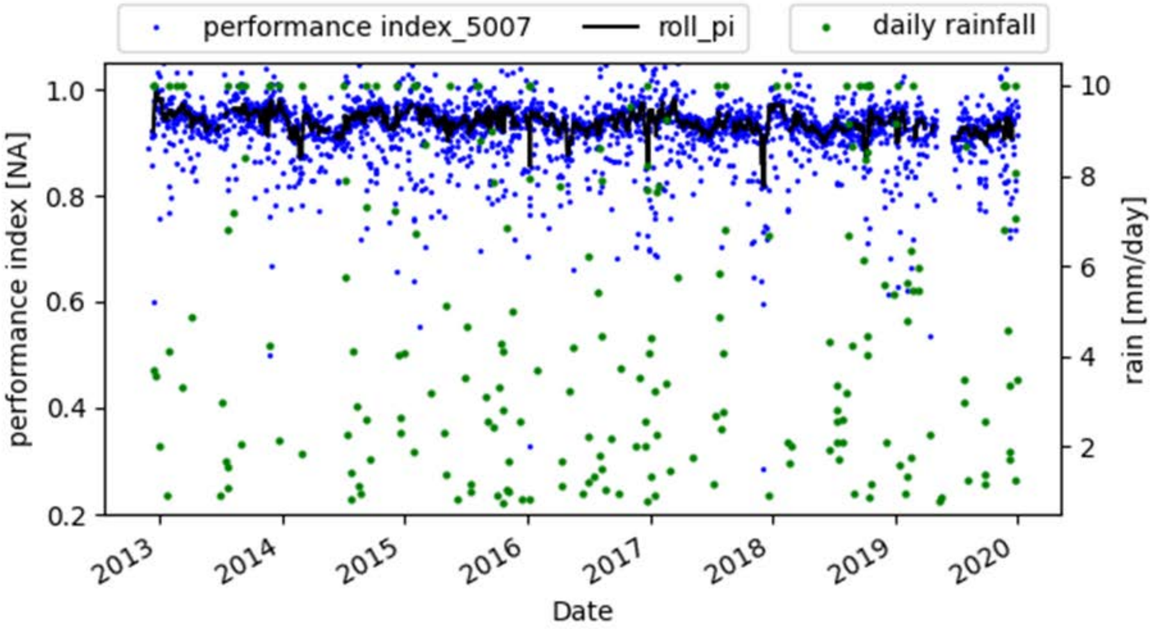

Figure 1. For site 5007, N_{daily} (the difference between the blue and black signals) is relatively consistent across the time series. This site has regular rainfall (cloudier weather) across the time series (shown in green).

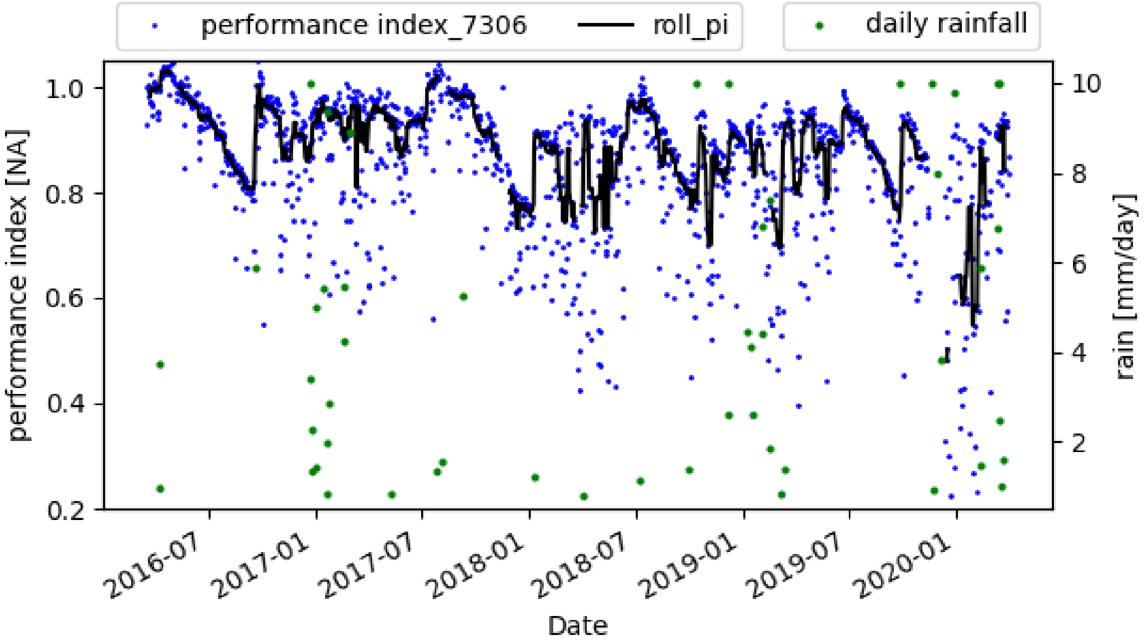

Figure 2. For site 7306, N_{daily} (the difference between the blue and black signals) is nearest to zero in late summer and fall, when little rainfall occurs and downward soiling trends can be seen in the data. By contrast, N_{daily} reaches extremes (~−40%) in the winter and spring months, when rainfall is more likely to occur (cloudier weather).

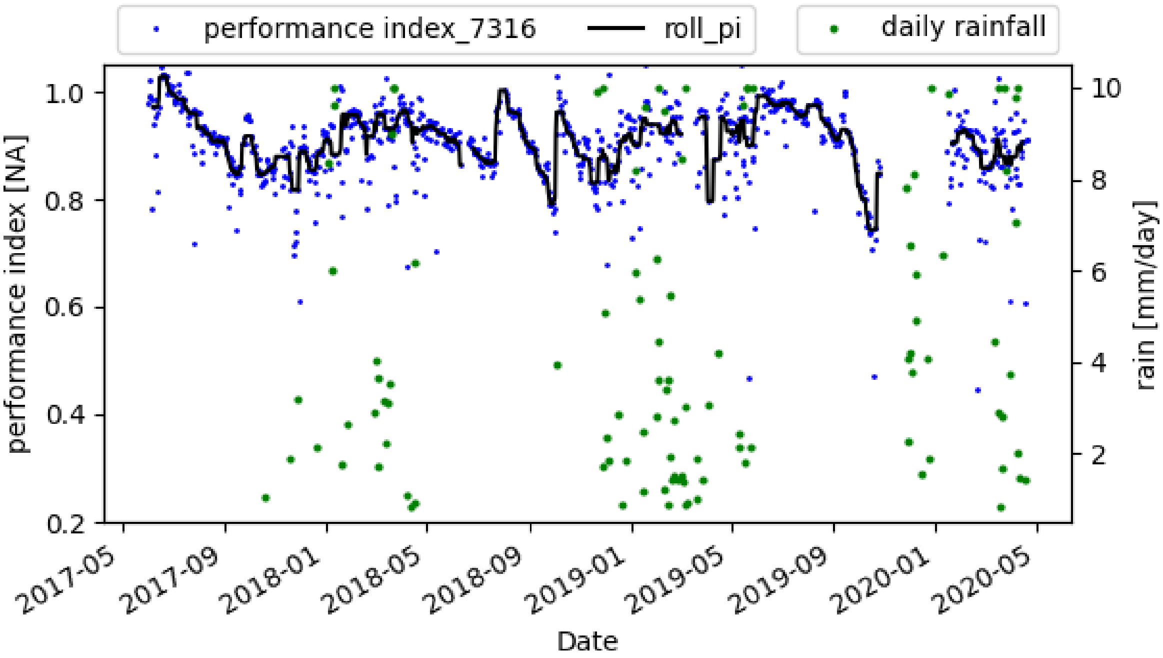

Figure 3. Compared to site 7306, site 7316 has a PI with less scatter and a lower overall N_{daily} (the difference between the blue and black signals). Both sites are in the Southwest United States and have intermittent rainfall. N_{daily} is closest to zero in the long dry soiling periods and typically has more negative extremes in the late winter and early spring, when rain is more common.

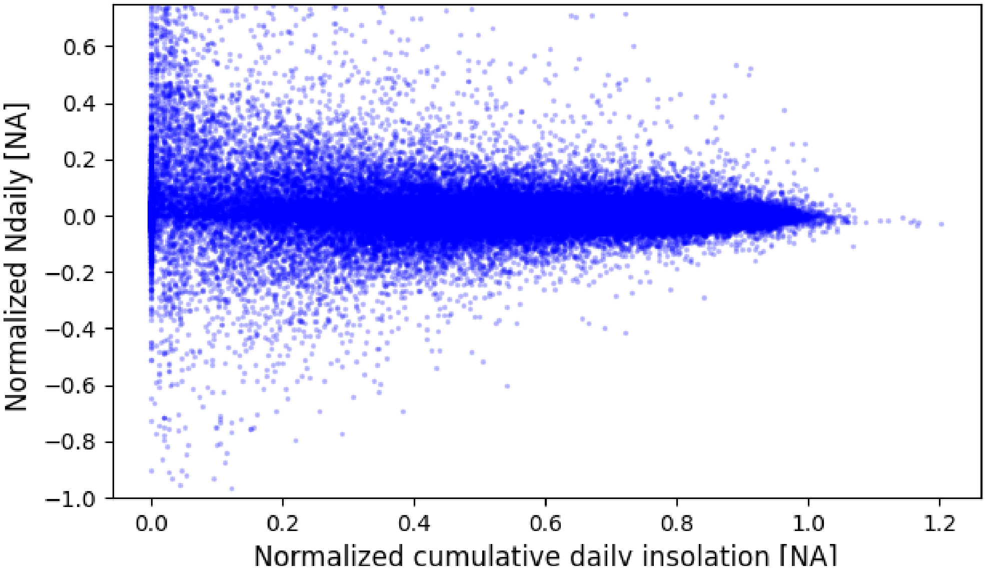

Figure 4. N_{daily} from 47 PV Fleets systems.

In an effort to better understand the possible relationship between PI noise and insolation, one of NREL’s high-fidelity systems was examined against 5-minute National Solar Radiation Database (NSRDB) Physical Solar Model version 3 (PSM3) data [15]. Figure 5 plots noise for the 15-minute normalized PV performance, N_{15min} (where N_{15min} is the deviation between the measured and modeled system performance for each 15-minute increment), versus irradiance level and variability. The irradiance level is quantified as the mean irradiance within the 15-minute window, and the variability is quantified as the base 10 logarithm of the standard deviation of the irradiance values within the 15-minute window. Irradiances above 800–900 W/m^2 show low variability within the 15-minute window and noise that is less than a few percent. On the other hand, at lower irradiances and high variability, the noise can be ±50%. The findings in Figure 4 and Figure 5 laid the groundwork for generating noise within the synthetic data set based on both irradiance levels and irradiance variability, as described in Section 2.4.

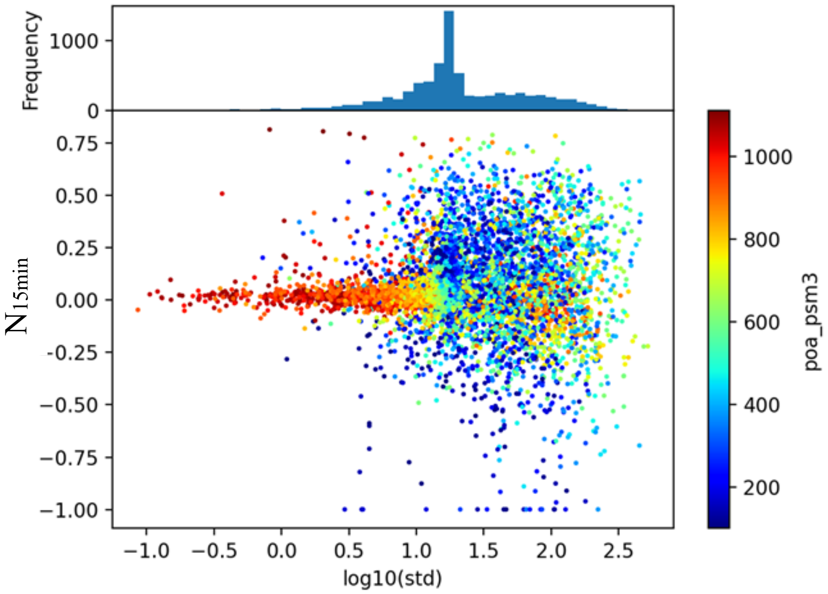

Figure 5. N_{15min} plotted against the base 10 logarithm of the standard deviation of the three 5-minute irradiance values within the 15-minute window and color coded by irradiance level.

2 Modeling and Methodology

Modern PV performance modeling stands on the shoulders of decades of progress in model development and validation, resulting in sophisticated and accurate software tools for simulating realistic system performance time series. Some of the performance issues of interest in this work (soiling, clipping, curtailment) can be modeled directly with these tools. However, more exotic issues, like sensor misalignment, tracker failures, and pollen-based soiling, require custom modeling. Therefore, instead of using one of the many familiar PV modeling packages, we have developed a customized simulation workflow for this work, the details of which are described in the following sections.

2.1 Input Data

As our intent is to build a synthetic data set that captures real-world challenges, we connect each time series generation to a specific latitude and longitude. The 5-minute irradiance, temperature, and wind speed in NSRDB PSM3 version 3.2.2 serve as inputs to the PV model, and PRISM Climate Group (PRISM) [15], [16] daily rainfall is the primary input to the soiling loss modeling. The model code can be used for any location where input data are available, but the initial data set is based on 24 locations in the Southwest United States and 14 locations in the Southeast United States (see Figure 6; more explanation of the location choices is provided within the soiling losses subsection). The simulated data set spans the four years from 2018–2021, the current extent of NSRDB's 5-minute data set.

2.2 Modeling Flow

2.2.1 Generation of 15-Minute PV Output Data

For each 5-minute data point, the following are calculated in the lead-up to "true" PV output: solar position, array orientation (if tracking and not in a stuck orientation), solar transposition to POA (using the Hay-Davies diffuse sky model), cell temperature with transience (using the Sandia PV Array Performance Model and the Prilliman model [17]), incidence angle modifier (IAM), losses (using a single-slab optical transmission model [18]), self-shading losses (using a simplified nonlinear [19] model), degradation losses (see model variants section), and DC PV power (using PVWatts® [18] while including the previously described DC losses). Soiling losses and string outages (see model variants section) are applied to the resulting 5-minute DC PV power before applying inverter clipping and efficiency losses per a generic inverter efficiency curve. If relevant, utility curtailment is applied to the AC inverter output, and the result of all these steps is the "true" 5-minute PV output. Five-minute data are then averaged to 15 minutes for the output data set. Artificial noise, as discussed in Section 2.4, is applied to each 15-minute data point based on the variability of the three 5-minute data points and the PSM3 evaluation of the data as clear sky. Finally, data outages are applied as sampled from normal distributions, with mean and standard deviations provided in Section 2.1. The 15-minute output data with noise and missing data periods represents the final synthetic PV system measured data stream.

2.2.2 Generating a Performance Index (PI)

A synthetic PI (synthetic generated power divided by modeled power) is generated for each 15-minute data point, as prescribed by the RdTools sensor workflow [20], [21]. First, the synthetic power data are normalized, and both power and POA data are filtered for erroneous data. Note that here, POA data is a 15-minute data stream generated through transposition of PSM3 data, and it can be misaligned with the PV array, as discussed further in the model variants section. Synthetic power data are filtered to remove inverter clipping per the RdTools logic clip filter. PVWatts inputs wind speed, ambient temperature, POA, location, and orientation information and outputs normalized modeled power, which enables the final synthetic PI calculation. Note that the synthetic PI calculation has no knowledge of string outages, POA misalignment, soiling, or other physical data issues; these issues can be included as described in the model variants section.

2.3 Soiling Losses

Typical PV soiling models assume that soiling follows a sawtooth pattern: linear soiling during dry periods followed by abrupt recovery or cleaning during rainfall events. These assumptions have been demonstrated for a number of PV systems in dry, dusty climates similar to the Southwest United States [2], [4], [5], [22]. Due to the frequency of rainfall in the eastern United States, sawtooth soiling models estimate near zero soiling losses in this region. By contrast, through work with system owner/operators per NREL's PV Fleet project, it has become clear that there are systems in the Southeast United States that have soiled as high as 10% and are not recovering with regular rainfall. The current hypothesis is that sticky soiling due to pollen or other biological sources is deposited on PV panels in the spring but is persistent against rapid rainfall removal. For this reason, we consider two regions for soiling losses and both a sawtooth and a pollen soiling model. As previously mentioned, the initial data set is based on 24 locations in the Southwest United States and 14 locations in the Southeast United States (see Figure 6). A sawtooth soiling loss model is applied to all 38 locations, but pollen soiling is also applied to the Southeastern sites.

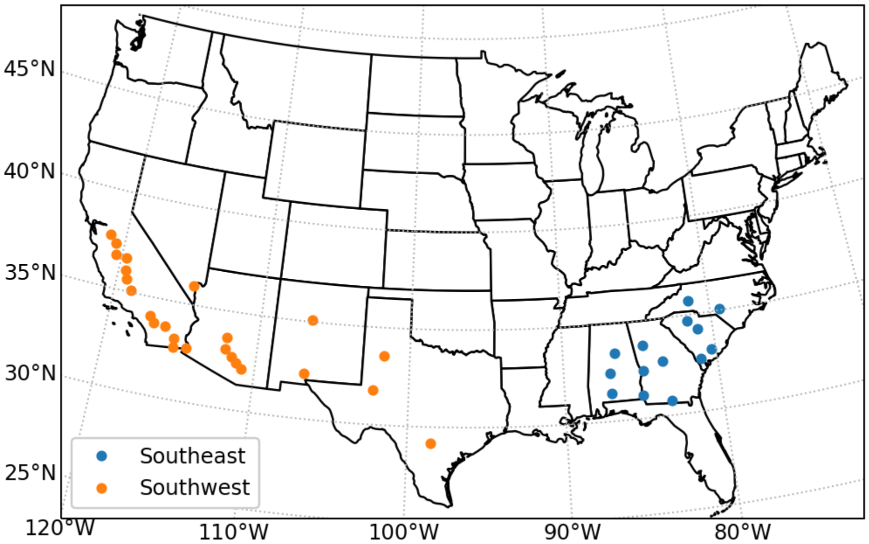

Figure 6. Locations of the 38 sites chosen for generating synthetic data using historical irradiance and weather data.

2.3.1 Sawtooth Soiling Details

Days where rainfall occurs per PRISM data (≥0.5 mm/day) are used to separate each location's time series into dry soiling periods. Linear soiling rates for each dry period in the Southwest United States are selected from a chopped normal distribution (no positive rates) with a mean of −0.14%/day and a standard deviation of −0.11%/day. The parameters of this distribution are established based on a distribution of soiling rates from the Southwestern United States provided by NREL's online soiling map [23]. Similarly, rates for the Southeastern United States are selected from a chopped normal distribution (no positive rates) with a mean of −0.05%/day and a standard deviation of −0.025%/day. For daily rainfall totals between 0.5 and 3 mm/day, the recovery or cleaning is not assumed to be perfect and is sampled from a range between 50% and 100% recovery. For rainfall greater than 3 mm/day, 100% recovery is assumed.

2.3.2 Pollen Soiling Details

In the absence of published models for predicting the performance loss associated with pollen-based soiling accumulation, we use a simple empirical model that produces a soiling profile qualitatively similar to what we have observed in performance data from real systems in the Southeastern United States. The empirical characteristics we seek to recreate are a rapid decline in performance over a few weeks followed by a gradual partial recovery over the subsequent months. For this purpose, we use a cubic Hermite spline fit to three anchor points: the date of soiling onset, the point of maximum performance loss, and the leveling off point of the gradual partial recovery. The choice of cubic Hermite spline was motivated by the ability to set the derivative of the resulting polynomial to zero at each of these three anchor points, making it possible to prevent the produced soiling ratio from exceeding 1.0 while maintaining a smooth curve. Our chosen anchor points, intended to produce a severe but plausible curve based on our experience, are defined in Table 1. The resulting profile, with and without a simulated manual cleaning on June 1, is shown in Figure 7.

Table 1. Pollen Soiling: Cubic Hermite Spline Anchor Point Definitions

Description	Date	Soiling Ratio [-]	Derivative [1/day]
Soiling onset	March 1	1.0	0.0
Maximum loss	April 12	0.85	0.0
Leveling off	September 30	0.95	0.0

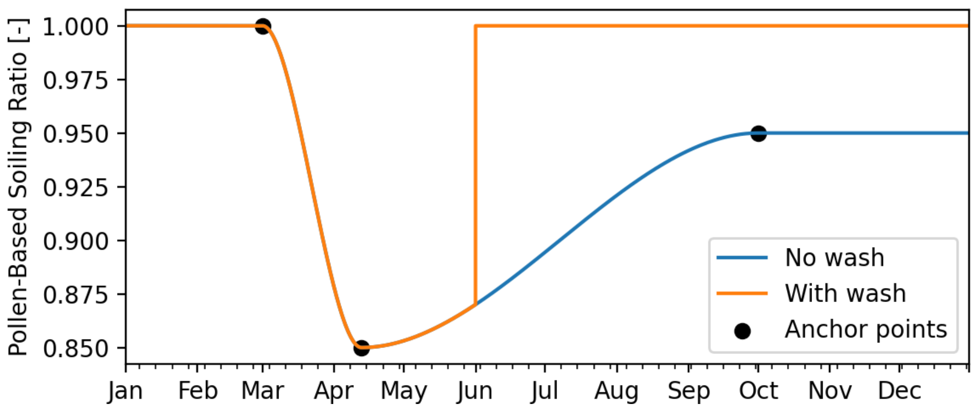

Figure 7. The empirical pollen-based soiling profiles (with and without manual washing), assuming a maximum performance loss of 15% and a residual loss of 5%.

Because pollen soiling appears to be a highly localized phenomenon that does not affect all systems in the Southeastern United States, simulations in this region are chosen at random (with a 50% chance) to be affected by pollen soiling. Pollen soiling is never applied to locations in the Southwest. The "with wash" profile shown in Figure 7 is used for all simulations where pollen soiling is applied, with no year-to-year or site-to-site variation.

2.4 Artificial Noise

As noted in Section 1.3, the noise observed in realistic PI data sets varies between systems and depends on irradiance conditions. Additionally, it isn't necessarily well-represented by Gaussian or other convenient distributions. In order to replicate these traits in the synthetic data sets, we calculate the empirical PI noise statistics observed in real system data sets in a way that preserves the relationship with irradiance conditions and the variation between systems. Table 2 lists the systems used for N_{15min} characterization.

Table 2. Systems Used for N_{15min} Characterization

System Name	Region	Year
RTC NV	Southwest	2021
PV Fleets 7336	Southwest	2021
NIST Ground	Southeast	2018
PV Fleets 8241	Southeast	2019

These systems were chosen because their performance is mostly unaffected by the other performance effects (soiling, clipping, etc.) already included in our model. For each of these four systems, empirical noise (N_{15min}) was defined as the relative difference between measured AC power and (noiseless) expected power, calculated using the approach described in Section 2.2.1. This noise was then divided into clear- and cloudy-sky subsets (using the location's PSM3 sky classification for each timestamp) and characterized as described below.

2.4.1 Clear-Sky N_{15min} Characterization

We observe that timestamp-by-timestamp PI deviation from a longer-term average is more stable in clear-sky conditions and might be better described as "bias" than "noise." Therefore, we evaluate clear-sky noise at the daily level rather than at the level of 15-minute timestamps. Taking the clear-sky subset of each day, and throwing out days without at least 10 remaining values, we calculate the median noise value for each day. This forms an overall distribution of clear-sky biases, which is characterized with a cumulative distribution function (CDF) by calculating its 1st, 10th, 20th, 40th, 50th, 60th, 80th, 90th, and 99th percentiles. Daily biases range from ±2% to ±5% depending on the system. Figure 8 shows the clear-sky bias CDFs for the four systems listed in Table 2.

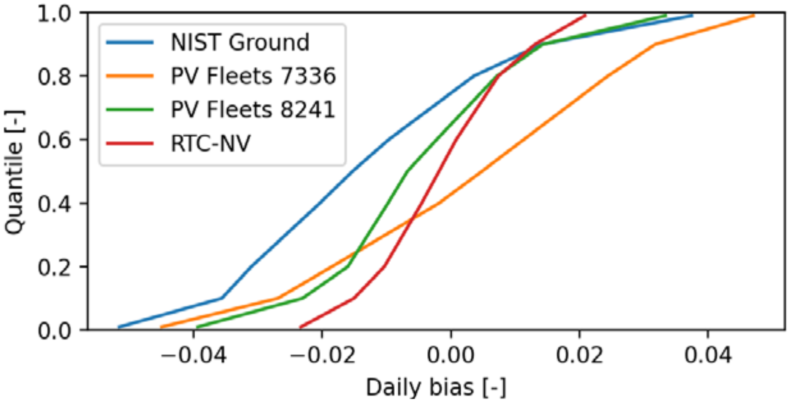

Figure 8. Discretized daily clear-sky bias distributions for the four noise characterization systems.

2.4.2 Cloudy-Sky N_{15min} Characterization

In contrast to the clear-sky biases being evaluated at a daily scale, cloudy-sky noise is characterized for each 15-minute timestamp and partitioned according to irradiance level and variability. Irradiance level is quantified as the mean irradiance within the 15-minute window, and variability is quantified as the base 10 logarithm of the standard deviation of the irradiance values within the 15-minute window. The cloudy-sky noise values are then binned in two dimensions according to these irradiance statistics. Bins are taken in steps of 100 W/m^2 for irradiance level and 0.5 for variability. Finally, within each bin, an approximate CDF is calculated using the same percentiles used for the clear-sky CDFs. This forms a two-dimensional lookup table that provides a noise CDF for any combination of irradiance level and variability. Figure 9 shows these distributions for one system.

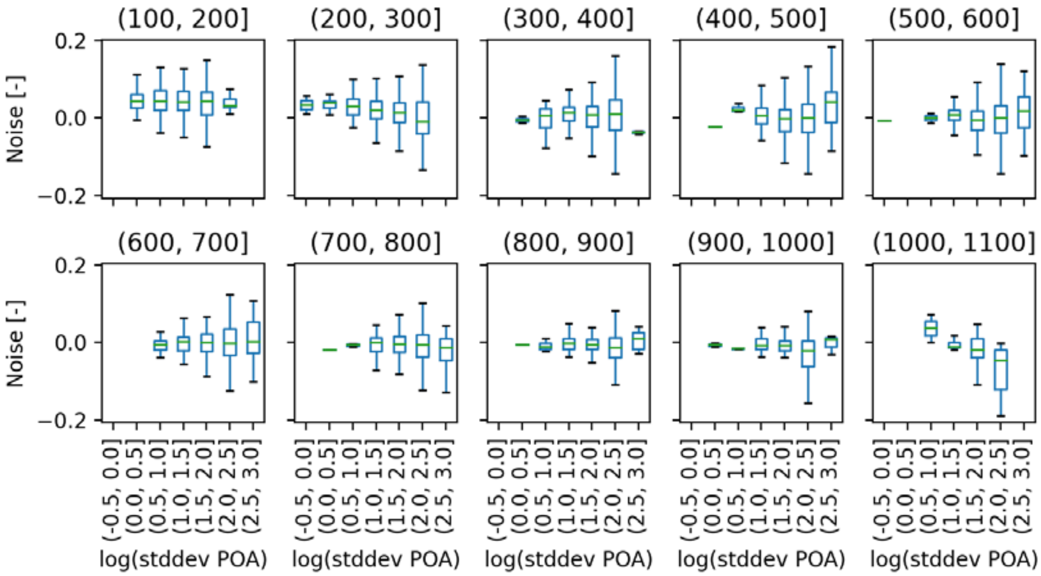

Figure 9. Observed N_{15min} for the PV Fleets 8241 system, partitioned by irradiance level and variability. Irradiance level varies across subfigures, with the irradiance range given in W/m^2 as the title. Irradiance variability—quantified as the base 10 logarithm of the standard deviation of the three 5-minute values within each 15-minute period—varies within subfigures.

2.4.3 Noise Generation

The artificial noise applied to simulated data is generated by sampling the empirical clear- and cloudy-sky noise distributions for a system chosen randomly from Table 2 according to region. As mentioned earlier, clear-sky biases apply to the clear-sky portions of entire days, whereas cloudy-sky noise is sampled independently for each 15-minute timestamp according to the CDF corresponding to the timestamp's irradiance level and variability. In the rare cases where a timestamp's irradiance level or variability falls outside the ranges covered by the cloudy-sky CDF lookup table, the timestamp is assigned zero noise.

This approach is limited in that it does not account for autocorrelation of noise during cloudy periods or seasonal variation in noise (outside of what can be captured with the base irradiance statistics). However, when the synthetic time series data were visually examined against PV Fleets data, the results were reasonable and showed improvement upon simply sampling from a normal distribution (for example, compare Figure 4 to Figure 12). Future work may include more methods for noise characterization and corresponding synthesis models.

2.5 Model Variants

To create both complexity and controls for testing within the synthetic data set, we generate a matrix of simulation variants as given in Table 3. For every variant in Table 3, the following are also applied per sampling from criteria in Table 4: (1) data outages, (2) utility curtailment, (3) dry soiling rates, and (4) permanent linear DC system degradation. Stuck tracking is also applied to all tracking variants, as applied per sampling in Table 4. Note that in the current model, all trackers are either stuck or tracking, whereas in real field scenarios, it is common for only a fraction of trackers to be stuck. This was done for simplicity; future variants may include such partially stuck trackers. All variants are given a ground coverage ratio of 0.4. Fixed-tilt systems have one module in portrait, and self-shading occurs as dictated by the system geometry. Tracking systems are assumed to backtrack within a flat field, and therefore no self-shading occurs. Each variant for Southeastern sites is randomly selected (50% chance of selection) for application of pollen soiling in addition to dry soiling. Each variant is also randomly assigned one of the two regional systems for assigning artificial noise.

Table 3. Model Variants for Synthetic Data Generation

		Orientation Variants		
		East-West Single-Axis Tracking	**Fixed South 10° Tilt**	**Fixed South 25° Tilt**
		POA Misalignments in Tilt		
Electrical Builds	***String Outages***			
DC/AC ratio = 1.0	Yes	None GHI	None POA + 1° POA + 5° GHI	None POA + 1° POA + 5° GHI
	No	None GHI	None POA + 1° POA + 5° GHI	None POA + 1° POA + 5° GHI
DC/AC ratio = 1.3	Yes	None GHI	None POA + 1° POA + 5° GHI	None POA + 1° POA + 5° GHI
	No	None GHI	None POA + 1° POA + 5° GHI	None POA + 1° POA + 5° GHI

The grey section of Table 3 demonstrates that there are 40 unique variants generated for each of the 38 sites. In Sections 1.1 and 1.2, we discussed that within the PV Fleets project, residual seasonality often occurs in the PI. Residual seasonality can occur for various reasons, but one of the most common is misalignment between the POA irradiance sensor and the PV array plane. Figure 10 shows that irradiance sensor misalignment of 1° can cause seasonal errors on the order of 2%; misalignment of 5° can produce as much as 8% error; and misalignment of 20° can result in more than 30% error. The POA misalignments in Table 3 (“none” indicates 0° misalignment; POA + 1° is a 1° misalignment; POA + 5° is a 5° misalignment; and GHI means GHI was mislabeled as POA and therefore actual misalignment depends on the system orientation) are included as separate variants. This allows the POA misalignments to represent the range of PI seasonality observed in PV Fleets while also providing signals with minimal seasonality to serve as a baseline. Note that stuck trackers impact PV power output but do not impact POA misalignments (in other words the POA measurement is assumed to be on a tracker that is fully functional).

Table 4. Criteria for Determining Statistical Variation in Model Variants

Application	Determination
Data outages	The number of data outages is randomly selected from a normal distribution with a mean of 8.6 outages per year and a standard deviation of 8.3 outages. The length of the outage is selected from a normal distribution with a mean length of 3.1 days per outage with a standard deviation of 10.5 days per outage.
Soiling rates	Soiling rates are randomly selected from a chopped normal distribution (no positive rates) for each dry period. For sites in the Southwest, the distribution mean is −0.14%/day and the standard deviation is −0.11%/day. For sites in the Southeast, the distribution mean is −0.05%/day and the standard deviation is −0.025%/day.
Stuck trackers	Trackers stall at horizontal four times within the data set and the stall lasts three days. The times of stalled tracker implementation are randomly selected.
Utility curtailment	AC capacity output is curtailed to 20% of the nominal AC output. Curtailment occurs 40 times in the entire data set for a 3-hour interval each time. The times of implementation are randomly selected.
String outages	There are three outages per data set, with a mean reduction in power of 10% and a mean length of 21 days of outage. Both the loss level and the length are randomly selected to be between 0.5x and 2x of the mean.
Linear degradation rate	The linear degradation rate (Rd) is randomly selected from the PV Fleets published degradation cumulative distribution function (see Figure 11) [1].
Pollen Soiling	The pollen soiling profile (as described in Section 2.3.2) is applied to all variants per site in the Southeast when randomly selected for application to that site.

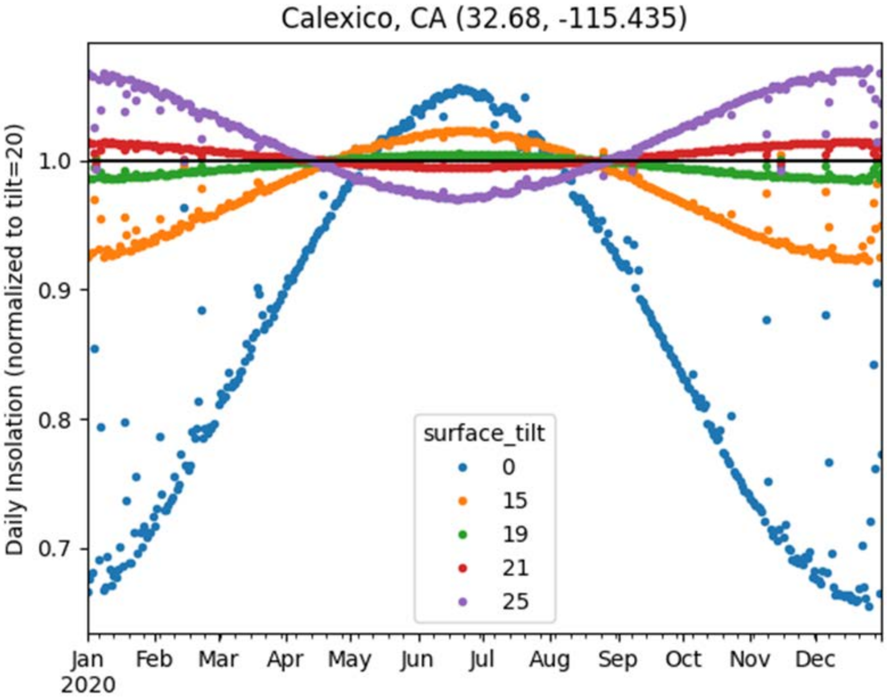

Figure 10. Seasonal impact on normalized insolation when there is misalignment between actual and desired insolation. All results are normalized to a tilt of 20 degrees. For example, the blue curve, which is at 0 tilt, represents a misalignment of 20 degrees, whereas the purple curve represents a misalignment of 5 degrees.

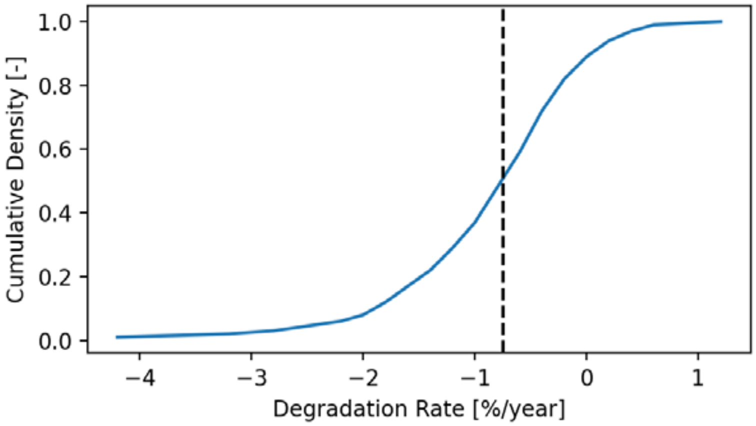

Figure 11. Distribution of degradation rates from the PV Fleets 2022 results. The dashed line indicates the median of −0.75% per year [1].

2.6 Output Data

2.6.1 Simulation Metadata File

The key inputs for each of the 1,520 simulation variants are recorded in a simulation metadata file. This file is tabular, with one row per simulation and columns as described in Table 5.

Table 5. Column Descriptions for the Simulation Metadata File

Column	Type	Unit	Description
id_number	numeric	-	Serial integer identifying the simulation; matches the time series data file name
lat	numeric	decimal degrees	Latitude coordinate of the simulation's location
lon	numeric	decimal degrees	Longitude coordinate of the simulation's location
name	char	-	Name of the simulation's location
is_dry	boolean	-	Whether the simulation is treated as western or eastern for soiling and noise assumptions (if True, the system is treated as western)
ilr	numeric	-	Inverter loading ratio (aka DC/AC ratio)
rd_percent_per_year	numeric	%/year	Assumed degradation rate
tracking	boolean	-	Whether the simulated system is tracking or fixed tilt (if True, the system is treated as tracking)
array_tilt	numeric	degrees	For fixed-tilt systems, the array tilt from horizontal (blank when tracking is True)
sensor_tilt	numeric	degrees	For fixed-tilt systems, the irradiance sensor tilt from horizontal (blank when tracking is True)
calendar_years	list	-	List of calendar years spanned by the data set
string_outages	boolean	-	Whether the simulation includes any string outages (if True, string outages are included)
pollen_soiling	boolean	-	Whether the simulation includes the effect of pollen soiling (if True, pollen soiling is included)
noise_parameters	char	-	Name of the system from Table 2 used for artificial noise

2.6.2 15-Minute Files

The 15-minute simulated power time series—along with various auxiliary time series—is stored in wide time series CSV form. The first column is a time-zone-localized timestamp in International Organization for Standardization (ISO) 8601 format (except the "T" is replaced with a space). The data columns are described in Table 6. The synthetic PI with all challenges included for the given variant is calculated as "actual power" divided by "pexp (misaligned tilt)."

Table 6. Column Descriptions for the 15-Minute Data Files

Column	Type	Unit	Description
actual power	numeric	W	Inverter AC power including the complicating effects
reference power	numeric	W	Inverter AC power without the complicating effects
pexp (array tilt)	numeric	W	Simple RdTools-style expected power, assuming the correct irradiance sensor orientation
pexp (misaligned tilt)	numeric	W	Simple RdTools-style expected power, assuming the misaligned sensor orientation
poa irradiance (array tilt)	numeric	W/m^2	Plane-of-array irradiance, assuming the correct irradiance sensor orientation
poa irradiance (misaligned tilt)	numeric	W/m^2	Plane-of-array irradiance, assuming the misaligned irradiance sensor orientation
ambient temperature	numeric	°C	Ambient air temperature from PSM3
ghi	numeric	W/m^2	Global horizontal irradiance from PSM3
wind speed	numeric	m/s	Wind speed from PSM3
data outage	boolean	-	Missing data indicator (1 = missing, 0 = not missing)
pollen soiling	numeric	-	Soiling ratio for pollen-based soiling (1.0 = no soiling loss)
conventional soiling	numeric	-	Soiling ratio for sawtooth soiling (1.0 = no soiling loss)
string outages	numeric	-	Fraction of online DC capacity (1.0 = no string outage)
curtailment	numeric	-	Fraction of allowed AC capacity (1.0 = no curtailment)
tracker stall	boolean	-	Tracker stall indicator (1 = tracking, 0 = stall)
clipping	boolean	-	Inverter clipping indicator according to the geometric clipping filter (1 = clipping, 0 = not clipping)

2.6.3 Daily Files

For the convenience of analyses that operate on daily aggregated performance values, the simulated data sets are also provided at the daily scale, throwing out timestamps flagged as corresponding to inverter clipping. These files contain all the columns from the 15-minute files (taken as daily averages of the 15-minute values) plus some additional columns listed in Table 7.

Table 7. Column Descriptions for the Daily Data Files

Column	Type	Unit	Description
poa insolation Wh/m^2 (array tilt)	numeric	Wh/m^2	Daily integrated POA irradiance, filtered to remove times determined as clipping, assuming the correct irradiance sensor orientation
poa insolation Wh/m^2 (misaligned tilt)	numeric	Wh/m^2	Daily integrated POA irradiance, filtered to remove times determined as clipping, assuming a misaligned irradiance sensor orientation
total rainfall mm	numeric	mm	Daily rainfall from PRISM

3 Synthetic Data Results and Discussion

We reviewed the graphical daily time series results from the 1,520 variants described in Table 3 and found that the synthetic results were within expectations. In an attempt to compare the noise in the synthetic data with PV Fleets, Figure 12 plots synthetic N_{daily} versus normalized cumulative daily insolation for all variants except those with GHI used as the irradiance sensor. In comparison to the 47 PV Fleets systems shown in Figure 4, the trend is similar for the majority of data points, but the PV Fleets data shows additional sparsely scattered data above 0.2 and below −0.2.

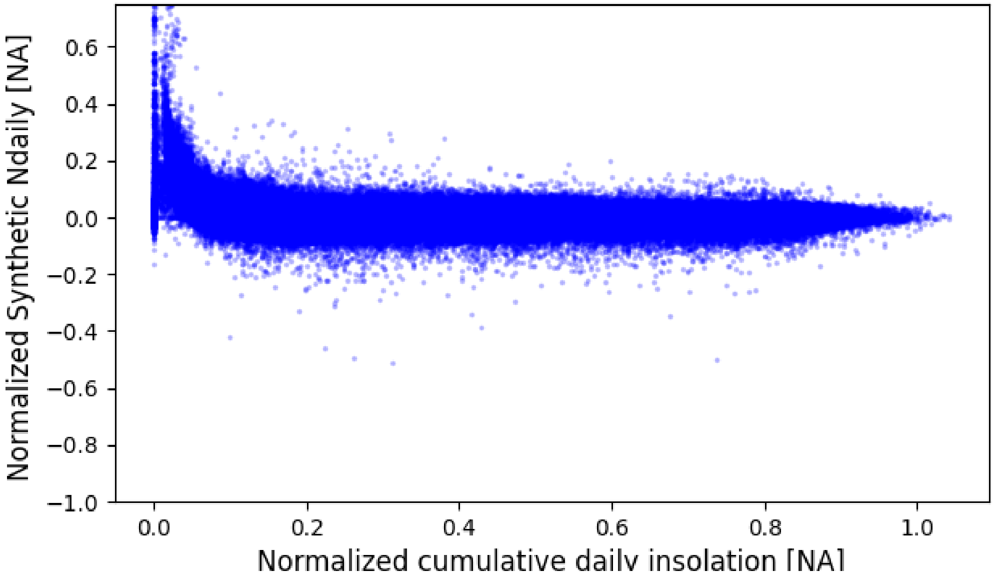

Figure 12. N_{daily} for time series data from all synthetic variants except those with GHI used as the irradiance sensor.

Figure 13 presents a representative synthetic time series for the variants embedded in Figure 12. Although the results are reasonable, compared to Figure 1 through Figure 3, the PI shows less scatter. This suggests that the noise applied to the 15-minute data may still underrepresent the level of noise typically seen in PV Fleets field data.

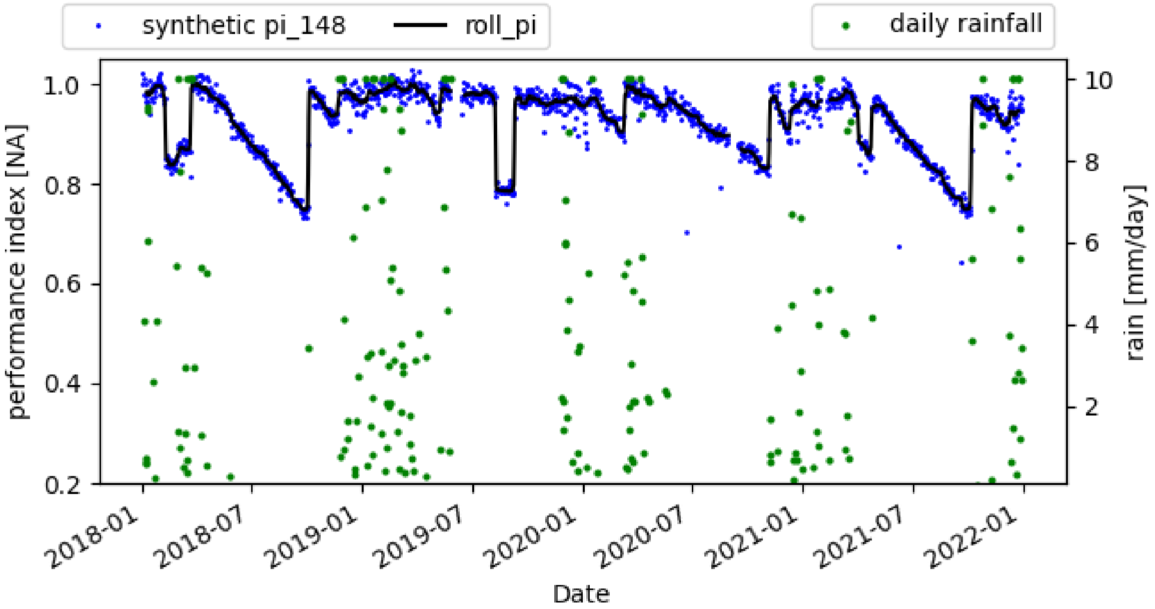

Figure 13. Typical time series plot associated with the data in Figure 12. This variant is as follows: inverter loading ratio (ILR) = 1.3, Rd = −0.80, fixed tilt of 10°, POA = 11°, string outages included, site is in the Southwest.

Figure 14 plots the same data as Figure 12 but also includes 20 tracked systems that have GHI used as the irradiance sensor. Although these results show additional scatter, as is seen in the PV Fleets data set, we can make no claim that the scatter in Figure 4 is due to incorrect irradiance sensors. Figure 15 and Figure 16 are time series for some of the additional variants embedded in Figure 14. These time series show similar features to those in the PV Fleets data (Figure 1–Figure 3)—for example, trends in N_{daily}, seasonality, and soiling.

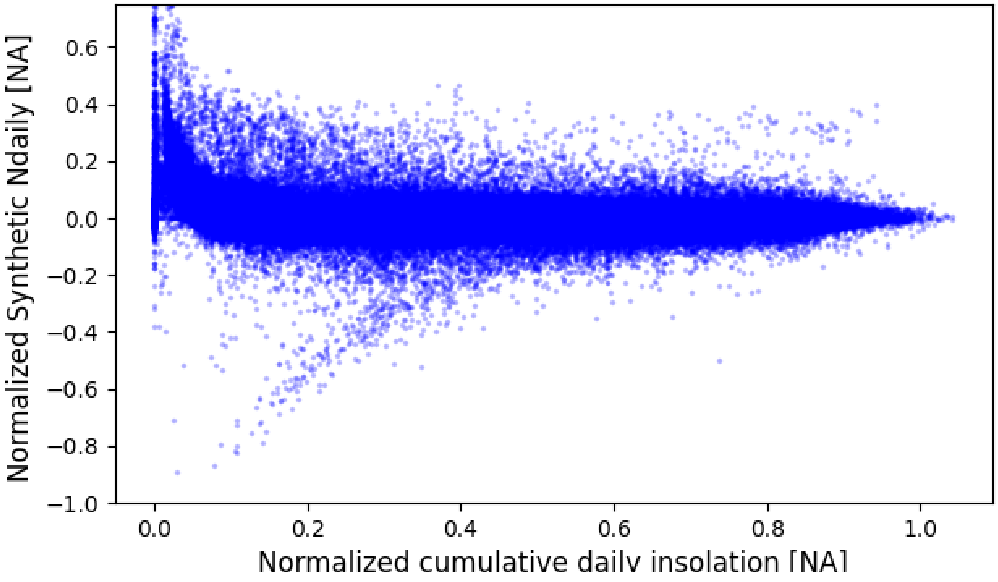

Figure 14. N_{daily} data from Figure 12 plus 20 synthetic variants where GHI was used as the irradiance for a tracking system.

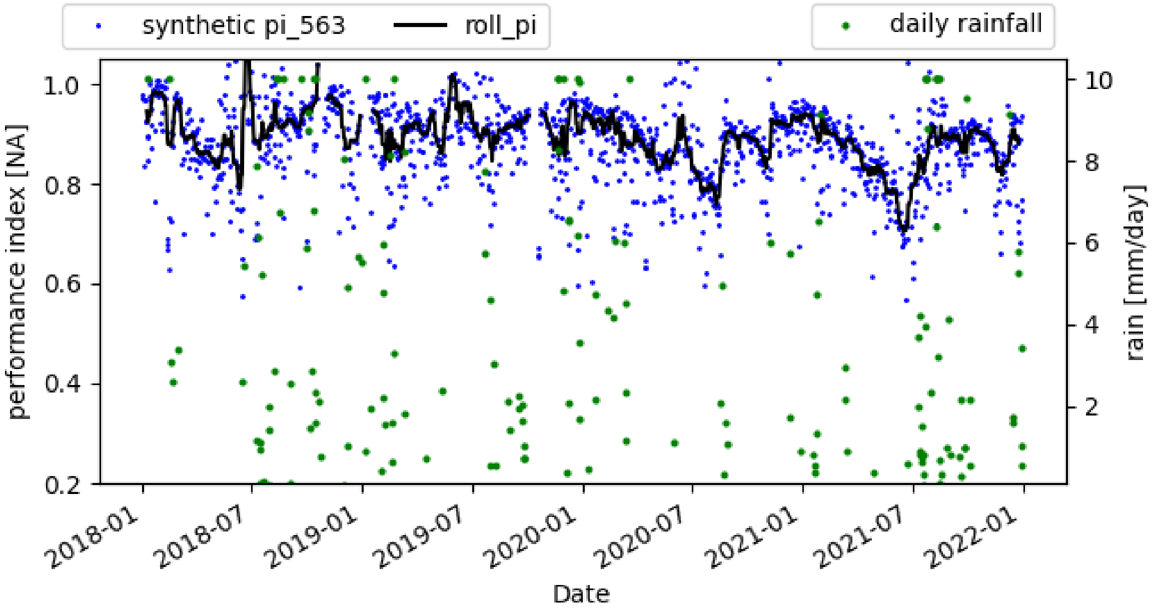

Figure 15. Typical N$_{daily}$ for a tracking system where GHI has been mislabeled as POA. Variant details are ILR = 1, Rd = −1.29, tracking, POA = GHI, no string outages, site is in the Southwest, seasonality is ~6%.

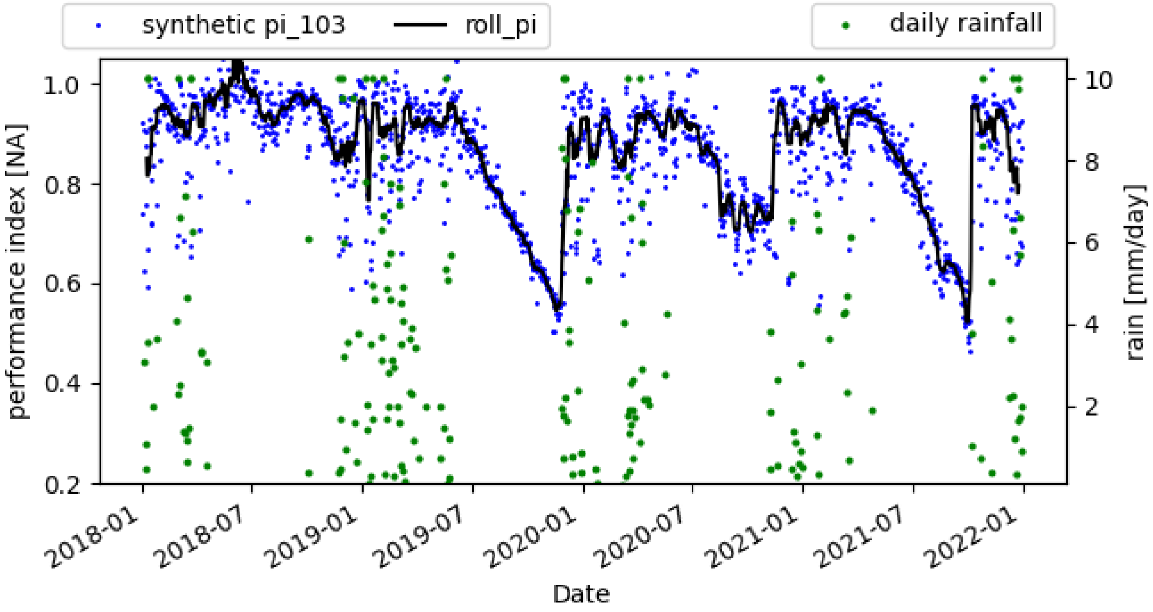

Figure 16. Variant details are ILR = 1.3, Rd = −0.72, tracking, POA = GHI, no string outages, site is in the Southwest. This example shows minimal N$_{daily}$ during long soiling periods and more extreme N$_{daily}$ during the rainy season.

Figure 17–Figure 22 each show unique variants from the synthetic data set that cover both the Southwest and Southeast United States. Each figure caption provides the details of the variant as well as how that particular time series shows issues that can be valuable in algorithm development. The last two of these figures are examples of implementing pollen soiling as described in the modeling section. Pollen and other bio-soiling trends are currently under investigation by the authors, and therefore the pollen model is expected to be updated as dictated by the evidence.

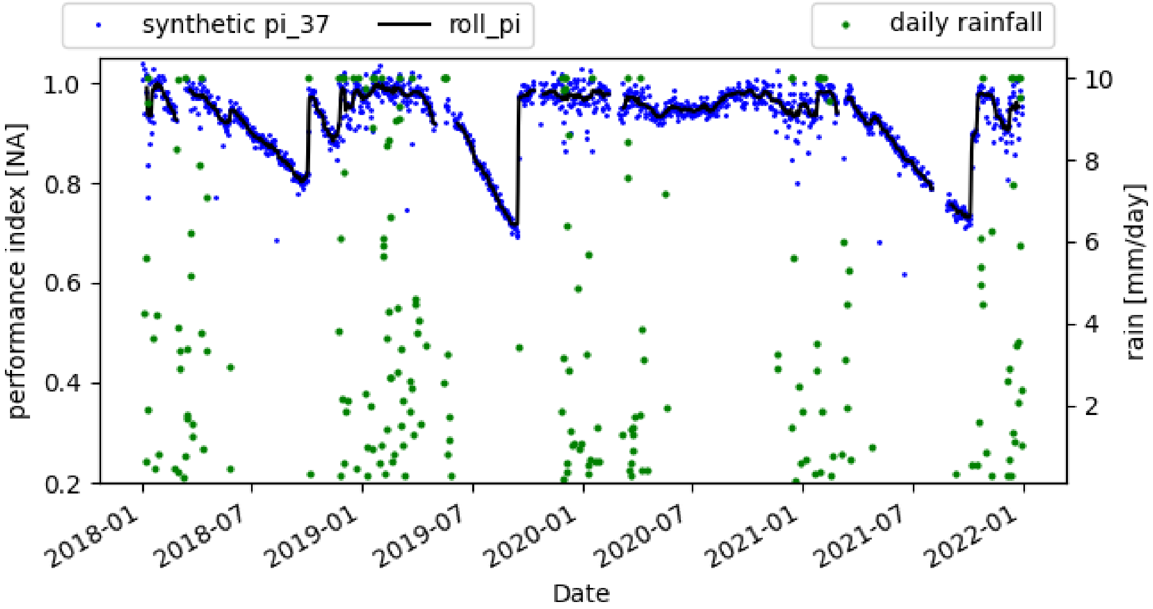

Figure 17. Variant details are ILR = 1.3, Rd = −0.57, fixed tilt = 25°, POA = 26°, no string outages, site is in the Southwest, seasonality = ~5%. This time series provides an example of how soiling trends make it difficult to distinguish the underlying seasonality.

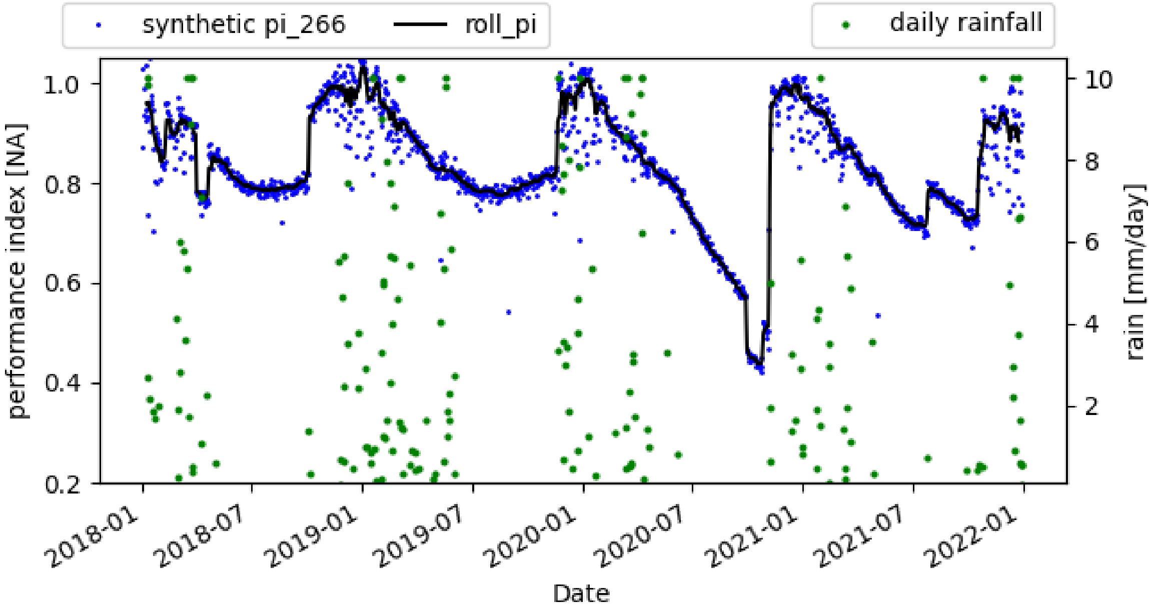

Figure 18. Variant details are ILR = 1.3, Rd = −1.69, fixed tilt = 10°, POA = GHI, yes string outages, site is in the Southwest, seasonality = ~25%. This time series provides an example of winter seasonal decline coinciding with dry periods with varying degrees of soiling loss. This is an ideal test case for soiling algorithms that can account for both seasonality and soiling.

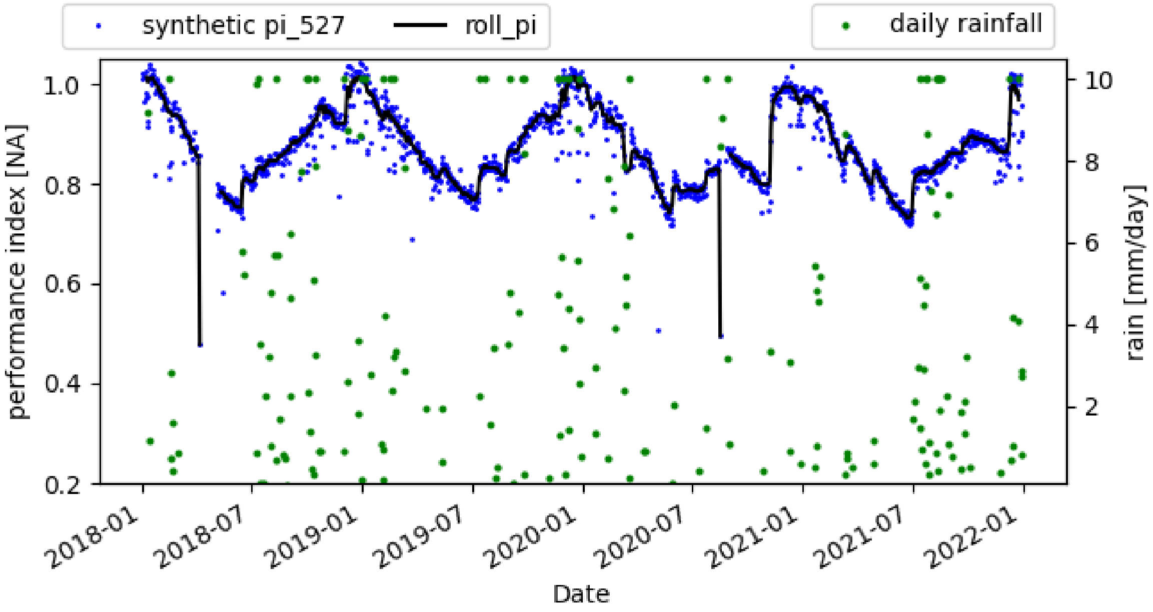

Figure 19. Variant details are ILR = 1, Rd = −0.60, fixed tilt = 10°, POA = GHI, no string outages, site is in the Southwest, seasonality = 25%. This time series provides an example where soiling occurs throughout the year but is the smaller signal compared to high seasonality.

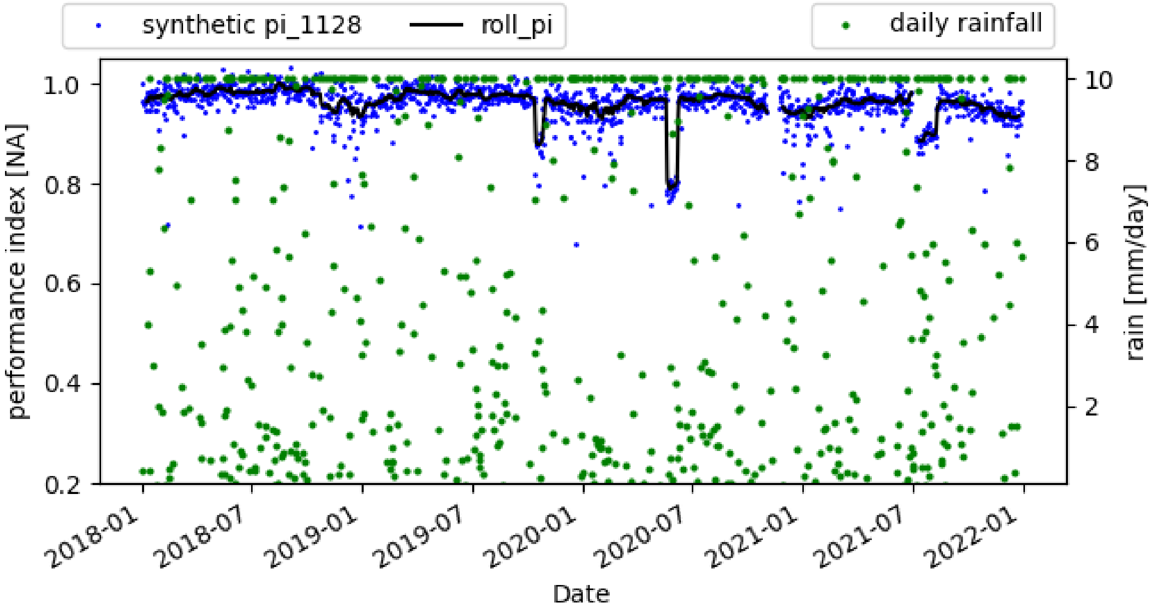

Figure 20. Variant details are ILR = 1, Rd = −0.66, fixed tilt = 10°, POA = 11°, string outages included, site is in the Southeast, seasonality = ~0. This time series provides a typical example of a Southeastern site with no pollen soiling. N_{daily} is slightly more extreme in the winter months, and there are a few drops in the PI due to string outages.

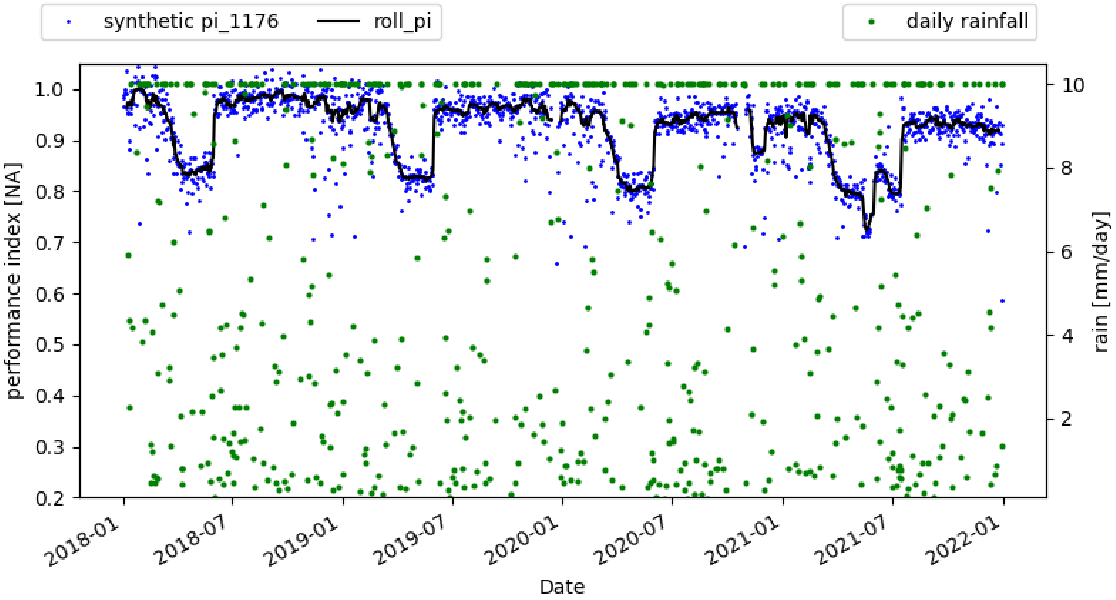

Figure 21. Variant details are ILR = 1, Rd = −1.66, fixed tilt = 25°, POA = 26°, pollen soiling included, string outages included, site is in the *S*outheast, seasonality = ~5%. This time series provides a typical example of the pollen soiling trend being implemented each spring*,* followed by a manual cleaning in the summer.

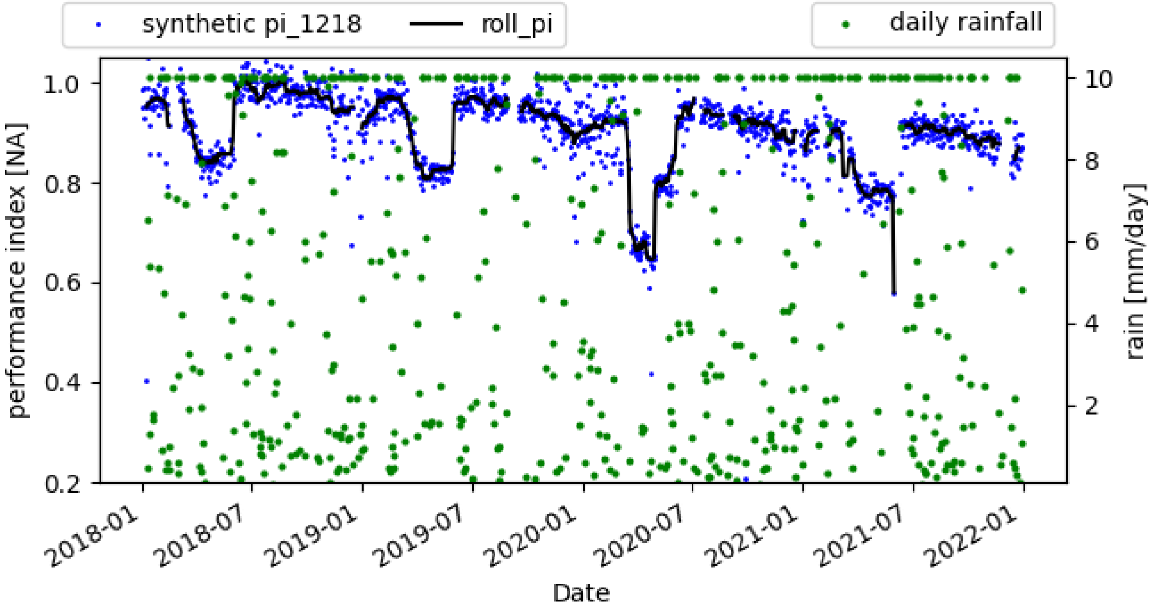

Figure 22. Variant details are ILR = 1, Rd = −2.61, fixed tilt = 25°, POA = 30°, pollen soiling included, string outages included, site is in the Southeast, seasonality = ~0. This time series shows how string outages can coincide with other events (in this case pollen soiling) and can make the underlying signal more complex to interpret.

4 Conclusions

Years of work with gigawatts of PV Fleets time series data have demonstrated the countless ways that PV field data can have data quality (QA) issues or physical occurrences that directly alter the performance of a PV system either rapidly or over time (physical issues). Both QA and physical issues hinder efforts to develop algorithms to extract important PV characteristics—such as degradation rates, soiling losses, tracker availability, or other key metrics—because these issues are not included in the metadata. Synthetic PV time series data can be a key asset in both algorithm development and validation if the synthetic data sufficiently includes real-world QA and physical issues that each algorithm might encounter. In this work, we have demonstrated a model flow to generate synthetic data with a wide range of QA and physical issues, while relying on historic weather and irradiance data from 38 locations spread across the southern United States. We have made available 1,520 variants of synthetic time series PV data and a metadata file provides all the associated problems with each variant. This allows the data user to work with a subset of variants that are relevant to the algorithm under development. The authors note that several physical issues, such as snow on PV panels, partially stuck trackers, and pollen soiling, are either not included in the current variants or will need to be improved with future research. As time and research allow, we expect to publish addendums to this work to improve the model flow and expand on the available synthetic data.

References

[1] D. C. Jordan *et al.*, “Photovoltaic fleet degradation insights,” *Prog. Photovolt. Res. Appl.*, vol. 30, no. 10, pp. 1166–1175, 2022, doi: 10.1002/pip.3566.

[2] M. G. Deceglie, M. Muller, Z. Defreitas, and S. Kurtz, “A scalable method for extracting soiling rates from PV production data,” in *2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC)*, Jun. 2016, pp. 2061–2065. doi: 10.1109/PVSC.2016.7749992.

[3] C. Deline *et al.*, “PV Fleet Performance Data Initiative: Performance Index-Based Analysis,” NREL/TP--5K00-78720, 1766838, MainId:32637, Feb. 2021. doi: 10.2172/1766838.

[4] L. Micheli and M. Muller, “An investigation of the key parameters for predicting PV soiling losses,” *Prog. Photovolt. Res. Appl.*, vol. 25, no. 4, pp. 291–307, 2017, doi: 10.1002/pip.2860.

[5] L. Micheli, M. T. Muller, M. G. Deceglie, and D. Ruth, “Time Series Analysis of Photovoltaic Soiling Station Data: Version 1.0, August 2017,” NREL/TP--5J00-69131, 1390775, Sep. 2017. doi: 10.2172/1390775.

[6] K. Perry, M. Muller, and K. Anderson, “Performance Comparison of Clipping Detection Techniques in AC Power Time Series,” in *2021 IEEE 48th Photovoltaic Specialists Conference (PVSC)*, Jun. 2021, pp. 1638–1643. doi: 10.1109/PVSC43889.2021.9518733.

[7] M. Muller, K. Perry, L. Micheli, F. Almonacid, and E. F. Fernández, “Automated detection of photovoltaic cleaning events: A performance comparison of techniques as applied to a broad set of labeled photovoltaic data sets,” *Prog. Photovolt. Res. Appl.*, vol. 30, no. 5, pp. 567–577, 2022, doi: 10.1002/pip.3523.

[8] K. Anderson, C. Downs, S. Aneja, and M. Muller, “A Method for Estimating Time-Series PV Production Loss From Solar Tracking Failures,” *IEEE J. Photovolt.*, vol. 12, no. 1, pp. 119–126, Jan. 2022, doi: 10.1109/JPHOTOV.2021.3123872.

[9] K. Perry and M. Muller, “Automated Shift Detection in Sensor-Based PV Power and Irradiance Time Series,” National Renewable Energy Lab. (NREL), Golden, CO (United States), NREL/PO-5K00-83200, Aug. 2022. Accessed: Apr. 07, 2023. [Online]. Available: https://www.osti.gov/biblio/1883388

[10] S. Vogt, J. Schreiber, and B. Sick, “Synthetic Photovoltaic and Wind Power Forecasting Data,” 2022, doi: 10.48550/ARXIV.2204.00411.

[11] M. Theristis *et al.*, “Comparative Analysis of Change-Point Techniques for Nonlinear Photovoltaic Performance Degradation Rate Estimations,” *IEEE J. Photovolt.*, vol. 11, no. 6, pp. 1511–1518, Nov. 2021, doi: 10.1109/JPHOTOV.2021.3112037.

[12] Y. Tang, J. W. M. Cheng, Q. Duan, C. W. Lee, and J. Zhong, “Evaluating the variability of photovoltaics: A new stochastic method to generate site-specific synthetic solar data and applications to system studies,” *Renew. Energy*, vol. 133, pp. 1099–1107, Apr. 2019, doi: 10.1016/j.renene.2018.10.102.

[13] A. Skomedal and M. G. Deceglie, “Combined Estimation of Degradation and Soiling Losses in Photovoltaic Systems,” *IEEE J. Photovolt.*, vol. 10, no. 6, pp. 1788–1796, Nov. 2020, doi: 10.1109/JPHOTOV.2020.3018219.

[14] “Physical Solar Model (PSM) v3 API | NREL: Developer Network.” https://developer.nrel.gov/docs/solar/nsrdb/psm3-download/ (accessed Apr. 08, 2023).

[15] “PRISM Climate Group at Oregon State University.” https://prism.oregonstate.edu/explorer/ (accessed Apr. 08, 2023).

[16] W. F. Holmgren, C. W. Hansen, and M. A. Mikofski, “pvlib python: a python package for modeling solar energy systems,” *J. Open Source Softw.*, vol. 3, no. 29, p. 884, Sep. 2018, doi: 10.21105/joss.00884.

[17] A. Dobos, “PVWatts Version 5 Manual,” NREL/TP-6A20-62641, 1158421, Sep. 2014. doi: 10.2172/1158421.

[18] K. Anderson, “Maximizing Yield with Improved Single-Axis Backtracking on Cross-Axis Slopes,” in *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, Jun. 2020, pp. 1466–1471. doi: 10.1109/PVSC45281.2020.9300438.

[19] D. C. Jordan, C. Deline, S. R. Kurtz, G. M. Kimball, and M. Anderson, “Robust PV Degradation Methodology and Application,” *IEEE J. Photovolt.*, vol. 8, no. 2, pp. 525–531, Mar. 2018, doi: 10.1109/JPHOTOV.2017.2779779.

[20] “API reference — RdTools 2.1.4+0.g996f843.dirty documentation.” https://rdtools.readthedocs.io/en/stable/api.html (accessed Apr. 08, 2023).

[21] A. Kimber, L. Mitchell, S. Nogradi, and H. Wenger, “The Effect of Soiling on Large Grid-Connected Photovoltaic Systems in California and the Southwest Region of the United States,” in *2006 IEEE 4th World Conference on Photovoltaic Energy Conference*, May 2006, pp. 2391–2395. doi: 10.1109/WCPEC.2006.279690.

[22] “Photovoltaic Module Soiling Map.” https://www.nrel.gov/pv/soiling.html (accessed Apr. 08, 2023).