Quantifying Error in Photovoltaic Installation Metadata

Preprint

Kirsten Perry, Quyen Nguyen, and Robert White

National Renewable Energy Laboratory

Presented at the 52nd IEEE Photovoltaic Specialists Conference (PVSC52)
Seattle, Washington
June 9-14, 2024

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

Conference Paper
NREL/CP-5K00-90318
July 2024

Quantifying Error in Photovoltaic Installation Metadata

Preprint

Kirsten Perry, Quyen Nguyen, and Robert White

National Renewable Energy Laboratory

Suggested Citation
Perry, Kirsten, Quyen Nguyen, and Robert White. 2024. *Quantifying Error in Photovoltaic Installation Metadata: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-5K00-90318. https://www.nrel.gov/docs/fy24osti/90318.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

Conference Paper
NREL/CP-5K00-90318
July 2024

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 Number 38258. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. 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 U.S. Government purposes.

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.

Quantifying Error in Photovoltaic Installation Metadata

Kirsten Perry, Quyen Nguyen, Robert White
NREL, Golden, CO, 80401, USA

***Abstract*—In this research, we quantify the level of metadata error for a fleet of 2860 photovoltaic (PV) systems, using metadata values provided by fleet owners. Using satellite imagery and time series analysis techniques available in open-source Python packages Panel-Segmentation and PVAnalytics, respectively, we evaluate the accuracy of PV system metadata such as location, azimuth, tilt, and mounting configuration (fixed tilt vs. tracking). We find that approximately 75% of provided latitude-longitude coordinates are within 190 meters of the actual solar installation. We were unable to link 7.8% of latitude-longitude coordinates to any solar installation via satellite imagery analysis. We evaluate the level of error in owner-provided mounting configuration data (fixed tilt vs. single-axis tracking), finding only 8 systems with incorrect mounting configurations. When evaluating azimuth and tilt parameters, we find that approximately 64% of the data is correct, with data for 860 systems (approximately 30%) not provided by system owners.**

To illustrate the importance of having correct solar metadata, we evaluate how incorrect metadata affects solar performance estimates by modeling system AC energy output at ground-truth vs. incorrect latitude-longitude coordinates, mounting configurations, and azimuth-tilt configurations. Energy output estimates can vary significantly if incorrect metadata parameters are used, with incorrect mounting configuration leading to the largest discrepancy with over 20% variation in expectedv AC energy output.

***Index Terms*—photovoltaic, satellite imagery, metadata error, azimuth, tilt, location, mounting configuration, satellite imagery, time series analysis**

I. Introduction

The United States solar industry has expanded rapidly over the past decade, with no indication of slowing down. In Q3 2023 alone, the US solar industry installed over 6.5 gigawatts of capacity, representing a 35% year-over-year increase, bolstered by the passing of the Inflation Reduction Act (IRA) [1]. This massive rise in solar installations begets new challenges for solar owners and operators; particularly, in estimating overall fleet performance and monitoring overall system health. Correct metadata such as system latitude-longitude coordinates, azimuth, tilt, and mounting configuration are critical for accurately estimating system energy outputs, as well as system degradation.

It has been anecdotally discussed in the solar industry that solar metadata can be incorrect. This issue is especially apparent during acquisitions, where site information may be lost during transference between owners [2]. However, the amount of error associated with PV Fleet metadata has never been quantified, so the extent of the problem is unknown.

In this research, we validate 2860 commercial- and utility-scale solar sites in the NREL PV Fleet Data Initiative database, all of which were provided by private-industry solar partners. During this validation, we evaluate each site's given latitude-longitude coordinates, mounting configuration, azimuth and tilt. We quantify the level of error between the partner-given metadata and the actual site parameters, using a semi-automated verification pipeline that leverages satellite imagery analysis and time series analysis. In addition to quantifying the level of metadata error, we generate multiple PVWatts simulations [3] at various locations across the United States to illustrate how using incorrect metadata can lead to inaccurate solar performance modeling estimates.

II. Methods

A. Data Set

For this analysis, PV sites from the NREL PV Fleet Data Initiative were used. The PV Fleet Performance Data Initiative is a US Department of Energy-funded project focused on the collection of fielded PV performance data into a centralized cloud database for the purpose of large-scale degradation analysis of the US fleet [4]. The associated PV Fleets database contains over 60 billion rows of time series data associated with over 6500 sites across the United States. Sites range from residential to utility-scale. For this particular research, only 2860 commercial- and utility-scale sites were analyzed; residential sites were omitted.

A map of the sites used in this research is shown in Figure 1, with each site represented as a single purple dot. Most systems are concentrated in the southwestern, southeastern, and northeastern parts of the United States.

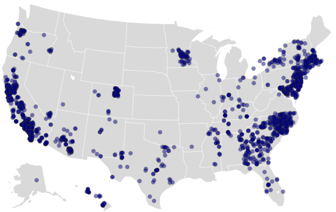

Fig. 1. United States map of solar systems evaluated from the NREL PV Fleets Initiative.

Figure 2 shows a breakdown of system size by percent of total fleet. Over 50% of the systems are between 100 kW and 1 MW in size. The vast majority of the fleet is commercial-scale, when using a 1MW cutoff for utility-scale systems. Only 1 percent of systems are over 100MW in size.

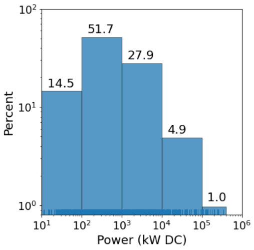

Fig. 2. Histogram showing system size (in kW DC) of the fleet analyzed, by percent.

B. Metadata Verification Process

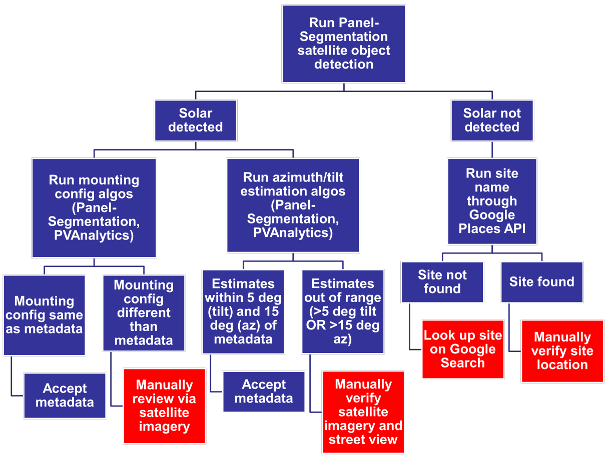
Run Panel-Segmentation satellite object detection
Solar detected
Solar not detected
Run mounting config algos (Panel-Segmentation, PVAnalytics)
Run azimuth/tilt estimation algos (Panel-Segmentation, PVAnalytics)
Run site name through Google Places API
Mounting config same as metadata
Mounting config different than metadata
Estimates within 5 deg (tilt) and 15 deg (az) of metadata
Estimates out of range (>5 deg tilt OR >15 deg az)
Site not found
Site found
Accept metadata
Manually review via satellite imagery
Accept metadata
Manually verify satellite imagery and street view
Look up site on Google Search
Manually verify site location

Fig. 3. Decision tree highlighting the metadata verification process, using software tools such as PVAnalytics and Panel-Segmentation. Boxes in red indicate a manual review step.

The metadata verification process used for this analysis is open-source and has been made public on Github [5]. A decision tree illustrating this process is show in Figure 3.

Several pre-existing Python software packages are leveraged for when automatically assessing site metadata, including PVAnalytics [6] and Panel-Segmentation [7]. First, each site's latitude-longitude coordinates are used to automatically generate a Google Maps image. This image is run through a deep learning object detection algorithm, available in the Panel-Segmentation package, to determine if solar is present in the associated satellite imagery [2]. If no solar installation is found, the site name is run through the Google Places API, and the associated outputs are manually reviewed.

If a solar installation is detected in the associated satellite image, then the azimuth and mounting configuration of the solar install are automatically estimated, using methodologies described in [8] and [2]. This process leverages deep learning models present in the Panel-Segmentation package, which mask the solar installation and classify its mounting type (fixed/tracking, ground/roof/carport). Additionally, if time series data is present for the site, all associated AC power data streams are run through pre-existing PVAnalytics package routines to estimate azimuth and tilt [9], as well as mounting configuration [6]. Running this routine for individual AC power data streams also helps to identify the azimuth and tilt on an inverter-level for systems with multiple azimuth-tilt configurations.

TABLE I
LATITUDE-LONGITUDE COORDINATE ERROR

Distance from Given Lat-Long Coordinates	Number Systems	System Percentage
<190m	2143	74.9
190m to 1 mile	417	14.6
1 mile to 10 miles	59	2.1
10 mile to 50 miles	10	0.3
>50 miles	7	0.2
Unknown/Not Found	224	7.8

The azimuth and mount estimates from Panel-Segmentation and azimuth, tilt, and mount estimates for PVAnalytics are compiled and compared to the partner-given values to check for discrepancies. Each azimuth-tilt pair generated in PVAnalytics is compared to the partner-given azimuth-tilt, and cases where the estimated azimuth is more than 15 degrees different, or tilt is more than 5 degrees different, are flagged for manual review. This same logic for azimuth is applied to Panel-Segmentation. Additionally, any mounting configuration (fixed/tracking) outputs from PVAnalytics or Panel-Segmentation that indicate a different mounting configuration than the partner designation are flagged for manual review.

When manually reviewing sites, the authors relied heavily on Google Earth satellite imagery and Google Street View for the verification of sites' azimuth, tilt, and mounting configuration. For sites with incorrect latitude-longitude coordinates, the authors double checked the satellite imagery for the area surrounding the given latitude-longitude coordinates to determine if any solar installations were present. Additionally, historical imagery was examined to determine if the site had been decommissioned and removed, and was consequently absent from recent satellite images. Google Street View was used to verify system tilt specifically, as it provided a view of the installation from its side.

In this analysis, 860 systems lacked critical metadata, including mounting configuration, azimuth, and tilt. These systems will be discussed further in the following sections, but the metadata validation pipeline described here was leveraged to automatically estimate these values for the 860 systems.

III. RESULTS

A. Latitude-Longitude Coordinate Accuracy

Table I gives a breakdown of the error in site latitude-longitude coordinates, following satellite imagery validation. 2143 systems, or approximately 80% of the fleet, were within 190 meters of their given latitude-longitude coordinates, and an additional 417 systems were within a mile of the given coordinates. Although these results are encouraging, we were still unable to locate the associated solar installation for 224 systems, or 7.8% of the fleet.

TABLE II
MOUNT LABEL ERROR: FIXED TILT VS. TRACKING

Status	Number Systems	System Percentage
Correct	1988	69.5
Incorrect	8	0.3
Unknown/Couldn't Verify	4	0.1
Not Provided	860	30.1

B. Mounting Configuration Accuracy

Each site was evaluated to determine if its mounting configuration data was correct, i.e. if the site was correctly labeled as fixed tilt or tracking. The results from this analysis are shown in Table II.

As previously mentioned, multiple data partners in the Fleets Initiative did not have any mounting information available for their sites. These sites are marked as 'Unknown/Not Given' in Table II. For some partners, this data was unavailable as the sites had recently been acquired from another company, and associated site metadata was not provided in the sale.

For the cases where data was provided, most partners provided the correct mounting configuration for their associated systems. Only 8 systems were flagged as incorrect, and 1988 systems had the correct mounting configuration designation.

C. Azimuth-Tilt Accuracy

Tables III and IV show the error results for azimuth and tilt, respectively. Approximately 64% of the azimuth data in the data set was verified as correct, and 63% of the tilt data was verified as correct. Azimuth had overall one of the highest error rates in the metadata; 94 systems had ground-truth azimuths more than 15 degrees different than the partner-provided azimuths (3.2%). An additional 72 systems were not verifiable via satellite imagery, as the solar system in question was not locatable.

Ground-truth tilt data was significantly more difficult verify when compared to all other metadata types. Flagged solar sites were manually examined via Google Street View to verify if tilt values were correct. Because Google Street View did not always align with the solar array, many flagged systems were unverifiable. Additionally, because some systems were unidentifiable via satellite imagery analysis, the tilt values for these systems could not be verified. In total, there were 185 systems in the dataset in the Unknown/Unverifiable category (6.5%). For the verifiable systems, 11 systems had a ground-truth tilt more than 5 degrees different than the partner-provided tilt value.

As previously mentioned, 860 systems were provided by data partners without azimuth or tilt information (30.1%).

D. Technology Type

It is currently difficult to validate system technology type using remote sensing techniques such as satellite imagery analysis, so we were unable to independently assess the quality of the module information provided. Instead, we were totally reliant on the module technology information provided by our partners. Figure 4 shows a breakdown of supplied module technology type for the 2860 systems assessed. The largest category is 'Unknown', representing approximately 57% of systems. This number is large because multiple fleet partners provided no information on module technology type. Solar partners may have this data available internally and it was just not provided; or this data was not machine-readable. We include these results to illustrate the level of uncertainty in module technology for fielded solar systems.

TABLE III
AZIMUTH ERROR

Abs. Degrees Difference from Ground Truth	Number Systems	System Percentage
<15	1834	64.1
15-30	30	1.0
30-45	13	0.5
>45	51	1.8
Unknown/Couldn't Verify	72	2.5
Not Provided	860	30.1

TABLE IV
TILT ERROR

Abs. Degrees Difference from Ground Truth	Number Systems	System Percentage
<5	1804	63.1
5-10	3	0.1
10-15	3	0.1
>15	5	0.2
Unknown/Couldn't Verify	185	6.5
Not Provided	860	30.1

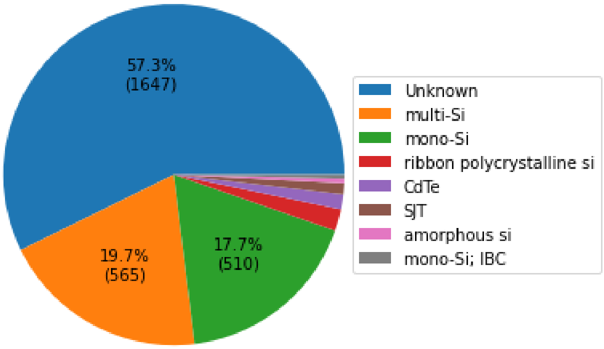

Fig. 4. Pie chart showing the breakdown of the fleet, by module technology type.

E. Metadata Verification Pipeline

In addition to quantifying error in the PV site metadata, this research introduces a framework that semi-automates the metadata verification process. The proposed methodology relies on pre-existing Python packages to do this [[6], [2]]. Although the functions in this process have been benchmarked independently, we provide statistics on the precision of the systems flagged for issues. Table V gives a summary

TABLE V
METADATA PIPELINE PRECISION: ERRONEOUS SYSTEMS ONLY

Metadata Type	Systems Flagged	Systems with Issues	Precision
Location	510	493	.967
Azimuth	684	94	0.137
Tilt	261	11	0.042
Mounting Config	54	8	0.148

of these results. As reference, unverifiable systems have been removed from the "flagged systems" count as their metadata could not be assessed as correct or incorrect.

Because of the scale of this analysis, only systems flagged for issues were manually reviewed, so overall F1 score for the process is not available. However, precision scores, or the number of flagged systems with actual erroneous metadata, were calculated for each metadata type. The location algorithm had by far the highest precision, with approximately 97% of the flagged systems not in their correct location. The tilt algorithm performed the worst, with only 4% of the flagged systems verified with wrong tilt data. Azimuth and mounting configuration (fixed vs. tracking) algorithms performed similarly, with approximately 14% of the systems flagged having the associated metadata issues. Although these are lower precision scores, the bounds for this analysis (<15 degrees difference in azimuth, and <5 degrees difference in tilt) are deliberately conservative, as flagging and manually reviewing the metadata for systems is better than missing systems with actual incorrect metadata. Previous literature states the median absolute error for the azimuth-tilt algorithm in PVAnalytics as 5.19 degrees and 1.21 degrees, respectively [9]. Given that these are median scores, some of the individual azimuth-tilt estimations will have higher error rates.

F. Error Trends

Following quantifying error for different metadata parameters, we investigated what type of systems were most likely to have incorrect metadata issues. In particular, we looked at whether metadata error was more common in commercial systems vs. utility-scale systems. The results of this analysis are shown in Figure 5. Systems were broken into three categories:

- <200 kW
- 200 kW-1 MW
- >1 MW

Commercial systems overwhelming have the most erroneous metadata. Commercial systems, which we will define as less than 1 MW in size, compose 66.2% of all systems validated in this study, but make up 100% of the systems with tilt errors and approximately 84% of the systems with azimuth errors. Only 60% of systems with mounting configuration errors are commercial-scale, which is more in-line with the total percentage of commercial-scale systems in the data set. Most utility-scale systems with mounting configuration issues were incorrectly labeled as fixed tilt when they were single-axis tracking.

The low rate of azimuth-tilt metadata error in utility-scale systems intuitively makes sense. These systems are much more closely monitored than commercial-scale systems. They also have far more consistent azimuth-tilt configurations than commercial-scale systems; most are single-axis tracking with a 180 degree azimuth. Conversely, commercial-scale systems can have incredibly unique mounting configurations with diverse azimuth-tilts combinations. This is particularly true for rooftop-mounted installations, where installations are heavily dependent on roof pitch and azimuth.

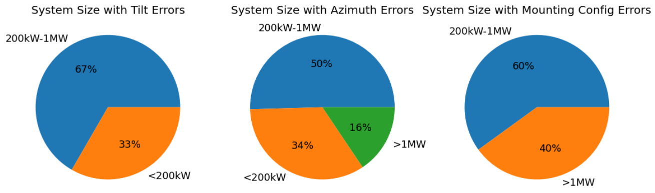

Fig. 5. Pie chart showing the size of systems with tilt, azimuth, and mount errors, respectively.

G. Modelling System Performance

In addition to quantifying the level of error/uncertainty in PV installation data, we investigated how this error impacts solar performance modeling. For simplicity, we used the PVWatts V8 model [3] to assess how estimated solar performance varied using ground-truth metadata as inputs vs. incorrect metadata as inputs.

1) Incorrect Latitude-Longitude Coordinates: For the first experiment, we assessed how system performance estimates vary with incorrect latitude-longitude coordinates. In particular, we looked at how estimates vary 1 mile, 10 miles, 50 miles, and 100 miles from a set of ground-truth coordinates at a variety of locations in the United States. We took the average error across cases directly north, south, east, and west of the ground truth coordinates.

The difference in yearly AC energy output was compared between the ground-truth estimate and the estimates at the incorrect latitude-longitude coordinates. Random locations across the United States were used as examples, and include the following:

- Greeley, CO
- Palmdale, CA
- Atlantic City, NJ
- Minneapolis, MN
- Marietta, GA

These locations were selected because they represent varying climates across the country, and all contained at least one solar site from the fleet analyzed.

A 100kW fixed-tilt system was modeled with 14% losses, a standard crystalline silicon module type, and a default azimuth and tilt of 180 and 20 degrees. The results for the varying latitude-longitude coordinate experiment are shown in Figure

6. These results show that error varies based on location and climate. All locations did experience AC energy estimate variation, particular when latitude-longitude coordinates were more than 50 miles away of the ground-truth location (mean 1.1% error). If latitude-longitude coordinates were within 1 mile of actual location, error was generally negligible (mean 0.22% error). These results highlight the importance of using correct latitude-longitude coordinates for solar sites, especially if the coordinates are egregiously off.

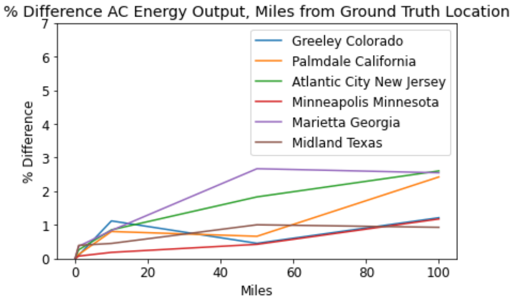

Fig. 6. Chart displaying the percent difference in AC energy estimates compared to ground-truth estimates, based on how far away the latitude-longitude coordinates are from the ground truth site.

2) Incorrect Mounting Configuration: We evaluated how estimates varied when a system was incorrectly identified as tracking when it was actually fixed tilt. Once again, the locations described above were run through a PVWatts simulation with the same input parameters described in the latitude-longitude experiment, except the mounting configuration was varied as either fixed-tilt or single-axis tracking. The percentage difference between the single-axis tracking estimate and the fixed-tilt estimate was then quantified. The results of this experiment are shown in Figure 7. These results are striking as average error is over 20%. These error metrics especially highlight the importance of using the correct mounting configuration when modeling solar performance.

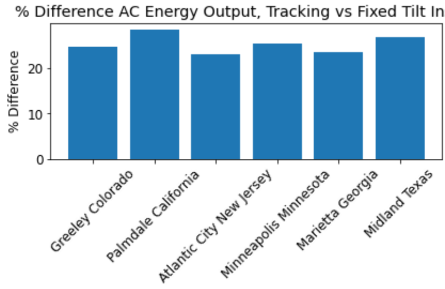

Fig. 7. Bar chart showing the percent difference in yearly AC energy estimates between a single-axis tracking system and a fixed tilt system at six locations across the United States.

3) Incorrect Azimuth-Tilt Configuration: We evaluated how varying azimuth and tilt changed AC energy outputs, when compared to a baseline system with an azimuth and tilt of 180 and 20 degrees, respectively. System parameters were held constant throughout the experiment, with only azimuth and tilt varied. The system was modelled in Palmdale, CA, with the same input parameters as described in the latitude-longitude experiment. Azimuth and tilt were varied by 5 degree increments in the experiment, with azimuth ranging between 110 to 250 degrees, and tilt ranging from 0 to 40 degrees. A heat map representing absolute error when compared to the ground truth configuration is shown in Figure 8, with a red dot representing the ground truth configuration, where AC energy output error is 0%.

The results shown in Figure 8 illustrate how AC energy estimates vary by azimuth-tilt configuration, with differences of over 15% for extreme cases. Median error for all combinations tested was 6.15%. The heat map illustrates that several azimuth-tilt combinations can lead to similar AC energy outputs as the ground truth estimate, but generally using the wrong azimuth-tilt inputs can lead to significant over- or under-estimates.

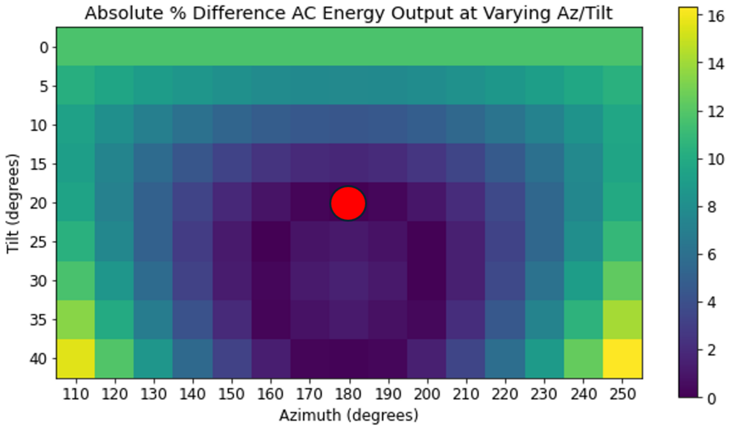

Fig. 8. Heat map showing the percent difference in AC energy outputs at varying azimuth and tilt configurations. The red dot in the center delineates the 'ground-truth' AC energy output that all other estimates are compared to.

IV. Key Takeaways & Future Work

The main intent of this research was to quantify the level of error in solar metadata, and reinforce why having correct metadata is important for evaluating system performance. The results of this work clearly highlight how prevalent incorrect solar site metadata is. The authors hope that this work motivates solar owners and operators to audit their site metadata to ensure its accuracy.

Additionally, we introduce a semi-automated framework to speed up metadata validation, based on open-source packages PVAnalytics and Panel-Segmentation. Using this framework, we were able to fill in metadata gaps for 860 systems without mount, azimuth, and tilt metadata in the NREL PV Fleets Initiative, which were provided by third-party data partners.

In terms of future work, we plan to apply this validation methodology to the open-source PVDAQ data set [10]. This data set is heavily used by both industry and academic researchers for analyzing PV performance time series data, but its associated metadata has not yet been rigorously validated. Applying this semi-automated methodology will help to identify and correct any serious metadata errors in the data set.

Acknowledgment

This work was authored in part by Alliance for Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory 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 Number 38258.

References

[1] Us solar market insight. https://www.seia.org/us-solar-market-insight. Accessed: 2024-01-22.

[2] Kirsten Perry and Christopher Campos. Panel segmentation: A python package for automated solar array metadata extraction using satellite imagery. *IEEE Journal of Photovoltaics*, 2023.

[3] Aron P Dobos. Pvwatts version 5 manual. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2014.

[4] Dirk C. Jordan, Kevin Anderson, Kirsten Perry, Matthew Muller, Michael Deceglie, Robert White, and Chris Deline. Photovoltaic fleet degradation insights. *Progress in Photovoltaics: Research and Applications*, 2022.

[5] Kirsten Perry and Quyen Nguyen. Pv metadata verification, 6 2024. URL: https://github.com/kperrynrel/pv_metadata_verification.

[6] Kirsten Perry, William Vining, Kevin Anderson, Matthew Muller, and Cliff Hansen. Pvanalytics: A python package for automated processing of solar time series data. URL: https://www.osti.gov/biblio/1887283.

[7] Ayobami Edun, Kirsten Perry, Christopher Deline, USDOE Office of Energy Efficiency, and Renewable Energy. Panel-segmentation, 11 2020. URL: https://www.osti.gov//servlets/purl/1726007, doi:10.11578/dc.20201130.12.

[8] Ayobami S. Edun, Kirsten Perry, Joel B. Harley, and Chris Deline. Unsupervised azimuth estimation of solar arrays in low-resolution satellite imagery through semantic segmentation and hough transform. *Applied Energy*, 2021.

[9] Kirsten Perry, Bennet Meyers, Kevin Anderson, and Matthew Muller. A reproducible validation of algorithms for estimating array tilt and azimuth from photovoltaic power time series. In *2023 IEEE 50th Photovoltaic Specialists Conference (PVSC)*, 2023.

[10] Chris Deline, Kirsten Perry, Michael Deceglie, Matthew Muller, William Sekulic, and Dirk Jordan. Photovoltaic data acquisition (pvdaq) public datasets. 12 2021. URL: https://www.osti.gov/biblio/1846021, doi:10.25984/1846021.