Developing a Comprehensive, System-Wide Forecast to Support High-Penetration Solar

Task 4 Report

DER Production Database

November 2018

Prepared for

California Energy Commission

Prepared by

Clean Power Research, LLC

Clean Power Research®

Legal Notice from Clean Power Research

This report was prepared by Clean Power Research. This report should not be construed as an invitation or inducement to any party to engage or otherwise participate in any transaction, to provide any financing, or to make any investment.

Any information shared prior to the release of the report is superseded by the report. Clean Power Research owes no duty of care to any third party and none is created by this report. Use of this report, or any information contained therein, by a third party shall be at the risk of such party and constitutes a waiver and release of Clean Power Research, its directors, officers, partners, employees and agents by such third party from and against all claims and liability, including, but not limited to, claims for breach of contract, breach of warranty, strict liability, negligence, negligent misrepresentation, and/or otherwise, and liability for special, incidental, indirect, or consequential damages, in connection with such use.

Contents

Executive Summary

Clean Power Research (CPR) was commissioned by the California Energy Commission (Contract EPC-17-003) to perform a broad study of solar forecasting, entitled "Developing a Comprehensive, System-Wide Forecast to Support High-Penetration Solar." This report covers work performed under Task 4 of the contract, focused on developing a library of Distributed Energy Resource (DER) production data and the refinement of DER growth projections in the Distribution Resources Plans (DRPs) developed in 2015 by the three independently owned utilities (IOUs) in California.

The *DER production database* covers a target period from January 1, 2011 through December 31, 2016 and contains solar PV production data by zip code at 15-minute intervals. The library is built on measured historical production data from 414 of 504 systems that were monitored under the California Solar Initiative (CSI) program. The remaining 90 of 504 systems were discarded due to data issues. To produce a complete and continuous data set, gaps in the measured data and invalid data needed to be filled and replaced with simulated data reflecting the characteristics of the underlying systems. It was therefore necessary to obtain accurate and detailed PV system specifications that could be used for modeling production and improving the measured data.

CPR used the available measured data to infer system specifications. Both inferred specifications and specifications reported by installers under CSI were used in conjunction with SolarAnywhere® satellite-derived irradiance data to model PV production. CPR evaluated the error in modeled production data versus measured data for both the inferred specificiations and the reported specifications to determine, system by system, which produced the lowest error. Then CPR combined the measured data with the best modeled production data. The result was an improved hybrid data set with continuous data for the entire six year target period. CPR then aggregated the production data by zip code.

Figure ES-1 charts the distribution of relative mean absolute error for the 414 systems using inferred and reported system specifications. Production data modeled using the inferred specs proved to be the source with the lowest error 96% of the time.

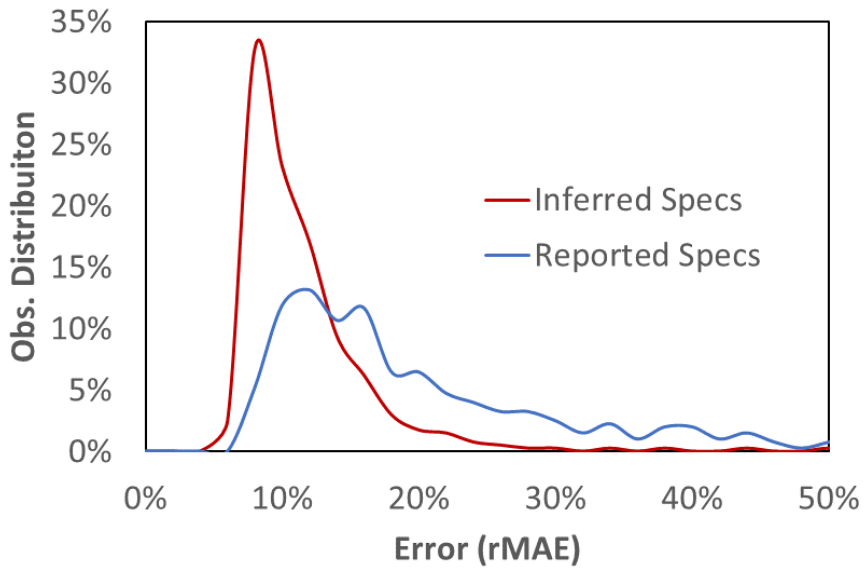

Figure ES-1. Error distribution using inferred and reported specifications

At the fleet level, where production from all systems is combined for each hour, production modeled using inferred specifications closely matched the measured data. Figure ES-2 shows fleet production for a single clear day. Figure ES-3 compares the modeled production for each hour with the measured production for that hour. With an hourly relative mean absolute error (rMAE) of just 4.3%, the fleet output begins to approach the typical error in the underlying satellite-derived irradiance data.

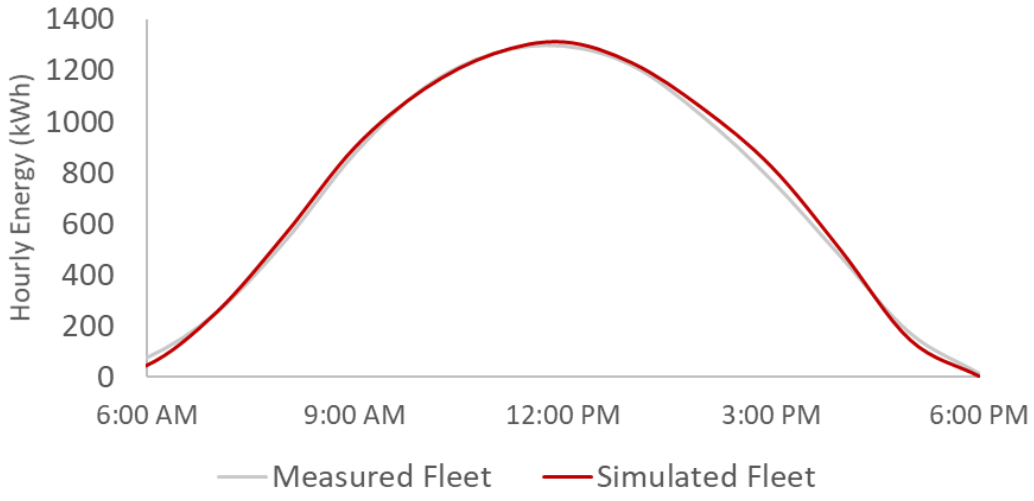

Figure ES-2. Fleet production on August 23, 2015

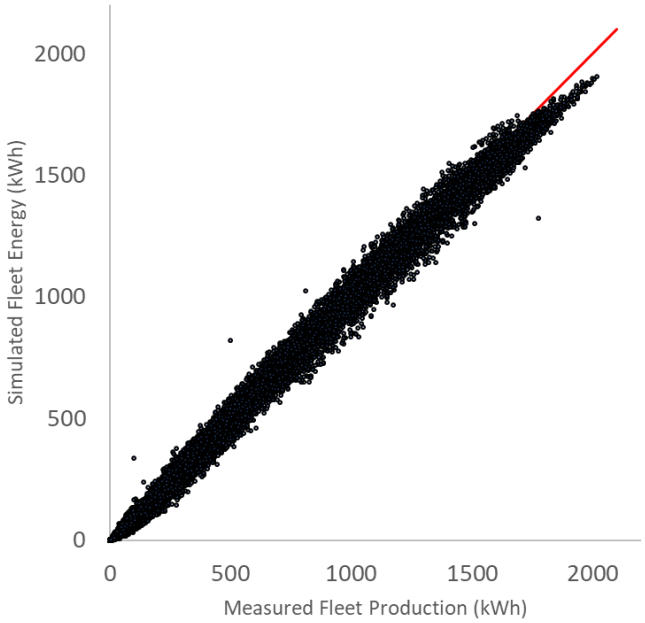

Figure ES-3. Fleet output for measured versus simulated using inferred specifications

CPR also used the solar PV net energy metering (NEM) Currently Interconnected Data Set (CIDS) to evaluate historic DER capacity growth and compared it to the capacity growth estimated in the Distribution Resources Plan (DRP) prepared by each of California's three Investor Owned Utilities (IOUs): Pacific Gas & Electric Company (PG&E), Southern California Edison Company (SCE), and San Diego Gas & Electric Company (SDG&E). As the final part of Task 4, CPR investigated methodologies for collecting information about other installed DER technologies to gain additional insight into their effect on California's electric grid.

Overview

Clean Power Research (CPR) was commissioned by the California Energy Commission (Contract EPC-17-003) to perform a broad study of solar forecasting, entitled "Developing a Comprehensive, System-Wide Forecast to Support High-Penetration Solar." This report covers work performed under Task 4 of the contract, focused on developing a library of Distributed Energy Resource (DER) production data and the refinement of DER growth projections in the Distribution Resources Plans (DRPs) developed in 2015 by the three independently owned utilities (IOUs) in California.

The *DER production database* covers a target period from January 1, 2011 through December 31, 2016 and contains production data for each zip code at 15-minute intervals. The library is built on measured historical production data from 414 of 504 systems that were monitored under the California Solar Initiative (CSI) program. The remaining 90 of 504 systems were discarded due to data issues. CPR collected, processed, and analyzed the measured data, analyzed the error in modeled versus measured data, improved the measured data by combining it with modeled production data, and aggregated the data by zip code.

CPR also used the solar PV net energy metering (NEM) Currently Interconnected Data Set (CIDS) to evaluate historic DER capacity growth and compared it to the capacity growth estimated in the Distribution Resources Plan (DRP) prepared by each of California's three Investor Owned Utilities (IOUs): Pacific Gas & Electric Company (PG&E), Southern California Edison Company (SCE), and San Diego Gas & Electric Company (SDG&E). As the final part of Task 4, CPR investigated methodologies for collecting information about other installed DER technologies to gain additional insight into their effect on California's electric grid.

DER Production Database Creation

Overview

To create the database of DER production, CPR first collected, analyzed, and processed the existing measured PV production data from CSI. As expected from a data acquisition effort of this type, CPR found significant amounts of missing or erroneous data. Therefore, to provide a complete data set for the target period, CPR had to fill in the missing data with simulated (modeled) data. CPR performed simulations using its SolarAnywhere® database of historical irradiance and temperature from the locations and time periods of interest.

CPR obtained design specifications for each system (system attributes required for modeling output) from two alternate sources: (1) specifications reported by installers through the CSI program; and (2) specifications "inferred" using the periods of good measured data. The inference of system specifications constituted one of the major efforts of this task.

CPR used the specifications—from either source depending on relative quality—to model PV production, then compared the accuracy of the modeled production data to the measured. Finally, CPR combined the measured and simulated production data and aggregated it by zip code to produce the final database.

Data Collection, Pre-processing, and Analysis

As part of the program's measurement and evaluation effort, the CSI program authorized its program contractor Itron, Inc. to install production meters on 504 CSI systems in 2010. CPR obtained 15-minute interval data for these systems from the California Distributed Generation Statistics web site.[1] This measured PV production data formed the core of the *DER production database*.

The interval data is comprised of three compressed comma-separated values (CSV) files—one for each of California's IOUs. Each of the three files contained all measurements for all systems metered by Itron in the respective utility's service territory.

With as many as 36 million rows per file, CPR began by separating the measurements in the files by system to create smaller files that were easier to manage and, along the way, creating a set of statistics and a list of the CSI application identifiers (IDs) that CPR would process. CPR used the application IDs to identify the applications in CIDS and in the database used by CSI for tracking applications while the program was active.

Measured Data Statistics

CPR confirmed that the set of three files contained data from 504 systems. Measurements did not start or end on the same date for all systems. First measurement dates for a given system ranged from 6/16/2010 to 3/14/2014 and last measurement dates ranged from 6/11/2011 to 12/31/2016, with the total period covered ranging from 271 days (just under 9 months) to more than 2,391 days (6 ½ years).

In addition to looking at the total period between the first and last measurement dates, CPR used flags included by Itron for each row indicating invalid data and identifying blocks with no missing or invalid data. Except for a single system for which there were no valid measurements, the largest block of continuous valid data was over 3 ½ years, while the smallest was just over 15 hours. More than 80% of the systems contained one year or more of continuous data that Itron had not identified as invalid.

In examining the data in more detail for some of the systems, CPR discovered that some systems had large blocks of negative energy recorded, as if the PV system were using rather than producing energy.[2] Eight systems had 50 or more blocks of eight hours or longer with more than 280 Watts of continuous

[1] https://www.californiadgstats.ca.gov/downloads/

[2] Although PV systems may have negative energy flow during the night when the inverter draws power, it is typically negligible.

power draw. In addition, two of the systems had some periods with unrealistically high energy production (more than 82 GWh in a 15-minute period). CPR treated these as invalid data.

Obtaining System Specifications

To provide the most complete and accurate DER production data possible, CPR planned to replace the missing or invalid measured data for each system with modeled PV production for the target period. Accurate historical PV system modeling requires time-correlated irradiance data at the system's location. In addition, this project required aggregation of system output at the zip code level, so CPR needed to identify the zip code for each system.

The interval data identified systems only by CSI application ID, so CPR used the March 31, 2018 CIDS to determine each system's zip code.[3] CPR downloaded the file from the California Distributed Generation Statistics web site.[4] CPR obtained the latitude and longitude for each system from PowerClerk® - the software used by CSI for program management.

Eleven of the application IDs indicated in the interval data were not listed in the CIDS, but were found in the Low-Income Solar PV Data (available on the same web site) from which zip code data was retreived.

Once the zip code was identified for each system, CPR determined that the 504 systems were located in 327, or about 20%, of the 1,601 zip codes where the CIDS-listed PV systems were installed. The total capacity of the measured systems represented 0.1% of the total CIDS capacity. Sixty-five percent of the zip codes for which there was measured data had only a single system in that zip code.

System Specification Sources

In addition to location-specific weather data, accurate modeling of PV production requires detailed and accurate PV system specifications that include azimuth, tilt, system rating, inverter efficiency and power rating, electrical losses, and the elevation of any solar obstructions. Although the CIDS has columns for azimuth, tilt, and system rating, it does not contain information on solar obstructions, module rating, inverter efficiency or inverter maximum power rating. For all but 11 systems, the database used by CSI consistently contains all the information needed, except losses and solar obstructions. Also, CSI applicants generally self-reported the system specifications, and specifications may have not been independently verified.

Inferring Specifications from Measured Production Data

[3] CIDS files are updated monthly. The March 31, 2018 data became available in May 2018.

[4] https://www.californiadgstats.ca.gov/download/interconnection_nem_pv_projects/NEM_CurrentlyInterconnectedDataset_2018-03-31.zip

Prior to this project, CPR had developed a method to infer system specificiations using measured production data. The method determines the specifications which, when used to simulate production, results in the lowest hourly error levels compared to the reference measured time series. The method had been developed using a spreadsheet format for a single system. For this project, a one-by-one use of the spreadsheet was impractical for 504 systems, so CPR encoded the method in software for batch processing.

By way of background, one approach to inferring PV specifications from measured data is to model all possible combinations of tilt, azimuth, tracking, solar obstructions, inverter rating, and losses for a one-year period and choose the one that results in the lowest error compared to measured. However, even if azimuth is limited to values between 90˚ and 270˚ (180 possibilities), tilt angle is limited to values between 0˚ and 90˚ (90 possibilities), and obstructions are modeled as opaque features in seven different 30˚ azimuthal bins (e.g., 75˚ to 105˚, 105˚ to 135˚, …, 255˚ to 285˚) with the obstruction's elevation angle for each azimuth bin between 0˚ and 50˚ (50 possibilities), this would require simulation of 12 quadrillion candidate systems (i.e., 180 x 90 x 50^7). With today's computing performance, this approach is not feasible.

Instead of employing such a "brute force" approach, CPR's method instead employs a golden section search to identify, one PV system attribute at a time, the values that minimize error. The method works through the attributes in a specific order, selecting a coarse approximation initially, then refining some of the values in a second iteration. For example, in the initial modeling CPR made use of a "constant horizon,"—solar obstructions at the same elevation angle for every azimuth—to approximate diffuse and direct losses from obstructions. Elevation angles for azimuth-specific obstructions were selected only after the values for other attributes had been determined.

The method also stores results of previous simulations to avoid modeling with the same specifications multiple times. Using this approach, CPR modeled a maximum of 431 different candidate specifications for each system for the period covered by the measured data and inferred specifications for 414 of the 504 systems. Using a combination of automated filters and manual inspection, CPR decided to eliminate ninety systems due to bad or missing data.

Modeling PV Production

Two alternate sets of system specifications were therefore available for use in modeling missing periods: specifications reported by installers and specifications inferred using the method described above. CPR modeled every system over the target period of January 1, 2011 through December 31, 2016 using both sets of specifications. CPR then calculated hourly relative mean absolute error[5] (rMAE) for every system using all periods with valid measured data and selected the specification set that resulted in the lowest

[5] Hoff, T.E., Kleissl, J., Perez, R., Renne, D., Stein, J. (2012 Proceedings of the American Solar Energy Society). Reporting of Irradiance Model Relative Errors, https://www.cleanpower.com/wp-content/uploads/Reporting-of-Irradiance-Model-Relative-Errors.pdf

rMAE. This helped ensure the best possible complete data in the *DER production database*. The procedure was as follows. CPR performed PV production modeling for the 414 systems using installer-reported specifications plus 414 systems using inferred specifications for the six-year period. These simulations made use of the SolarAnywhere® solar simulation APIs, which in turn use 30-minute, 1 km x 1 km SolarAnywhere satellite-derived irradiance data, temperature and wind speed.

SolarAnywhere simulations use a power model that calculates the direct and diffuse irradiance striking the tilted plane of a PV system's modules, taking into account any solar obstructions, module efficiency changes due to temperature and other DC losses, as well as the inverter's maximum power rating and its varying efficiency at different power levels. The model is capable of simulating fixed, single-axis tracking and dual-axis tracking systems, as well as systems with multiple arrays with different orientations and different DC to AC ratios. This capability was used for the simulations done using reported specifications. However, the code that inferred specifications from measured data did not attempt to discern whether a system had multiple arrays with different orientations. Therefore, the inferred specifications were always single-array systems whose output best approximates the measured data.

Comparing Results from Inferred and Reported System Specifications

When calculating error in modeled data, it is desirable to compare it against measured data that is known to be accurate and complete. Because of previously-identified invalid data periods, CPR applied filters to the measured data to eliminate:

1. Any period marked as invalid by Itron
2. Any period with a value less than 0 or greater than 1.35 x the 99th percentile value for that system
3. All nighttime periods
4. Any daytime period with a value less than 0.00001 (1/100 of a Watt hour in 30 minutes)[6]

After filtering the measured data, CPR calculated the hourly rMAE of the measured production data compared to data modeled using reported specifications and data modeled using inferred specifications. The rMAE for data from inferred specifications ranged from 6.3% to 50.8% with a median rMAE of 10.1%, while the rMAE for data from reported specifications ranged from 7.2% to 127.6% with a median rMAE of 16.7%.

For 96% of the time where we had both inferred and reported specs, energy production data modeled using inferred specs resulted in lower error when compared to measured data than the production data modeled using reported specs. Only 17 systems had lower error using reported specs rather than inferred specs.

Figure 1 compares the distribution of error (as measured using rMAE) for production data modeled using inferred and reported specifications.

[6] CPR identified daylight periods using sun position data provided by SolarAnywhere.

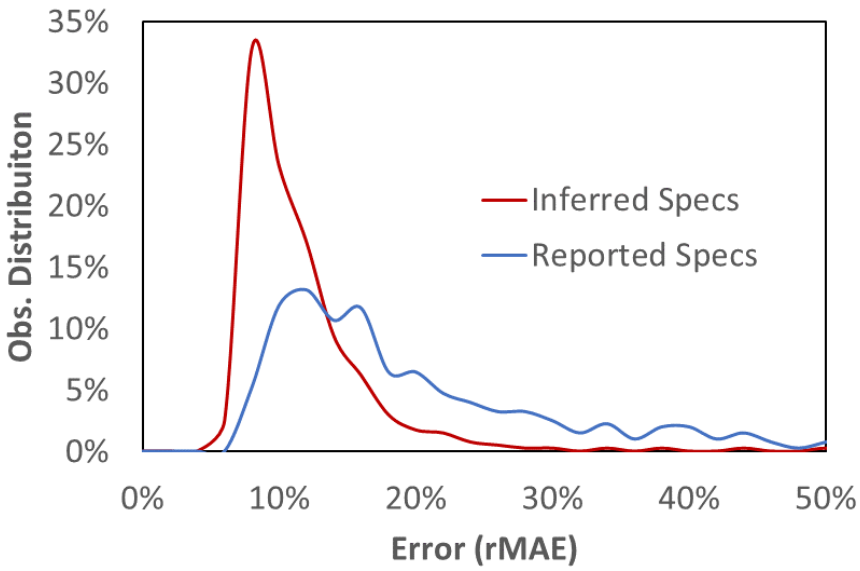

Figure 1 – Error measured vs. modeled using inferred and reported specifications

Another way to look at this is using the cumulative distribution. Figure 2 presents the cumulative distribution. As an example, the figure shows that, using the inferred specs approach, 95% of the systems have less than 18% error. On the other hand, 95% of the systems have less than 46% error when using the reported specs. CPR selected data from the inferred specs 96% of the time as the source to be merged with measured data for the *DER production database*.

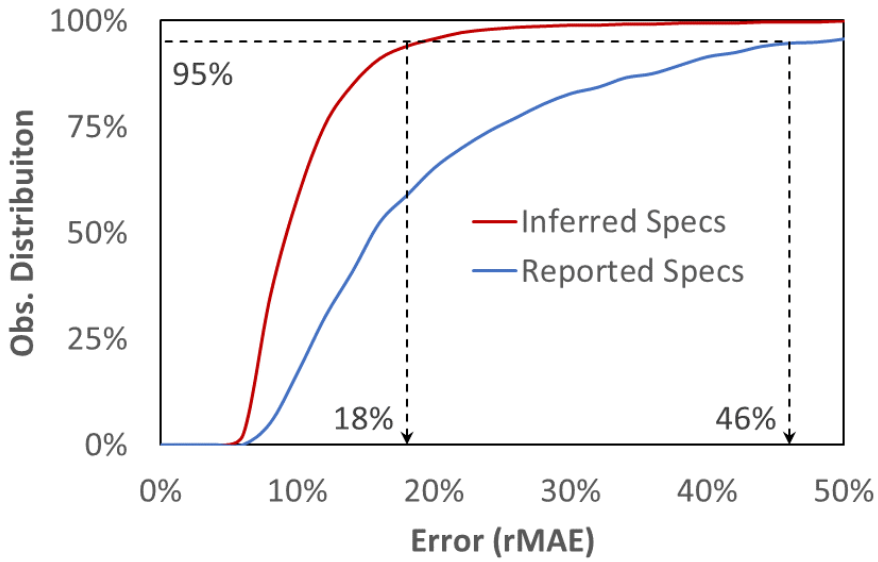

Figure 2 - Cumulative distribution of errors

Figure 3 provides a glimpse of the diversity of installation types reported for the 414 systems. Production data from fixed, single orientation systems are simpler to handle when inferring system specifications, while tracking systems can be more difficult. CPR did not attempt to specifically identify systems with multiple orientations. Overall, the inferred specs correctly identified systems as fixed or tracking 95% of the time and produced a false positive for a tracking system only once.

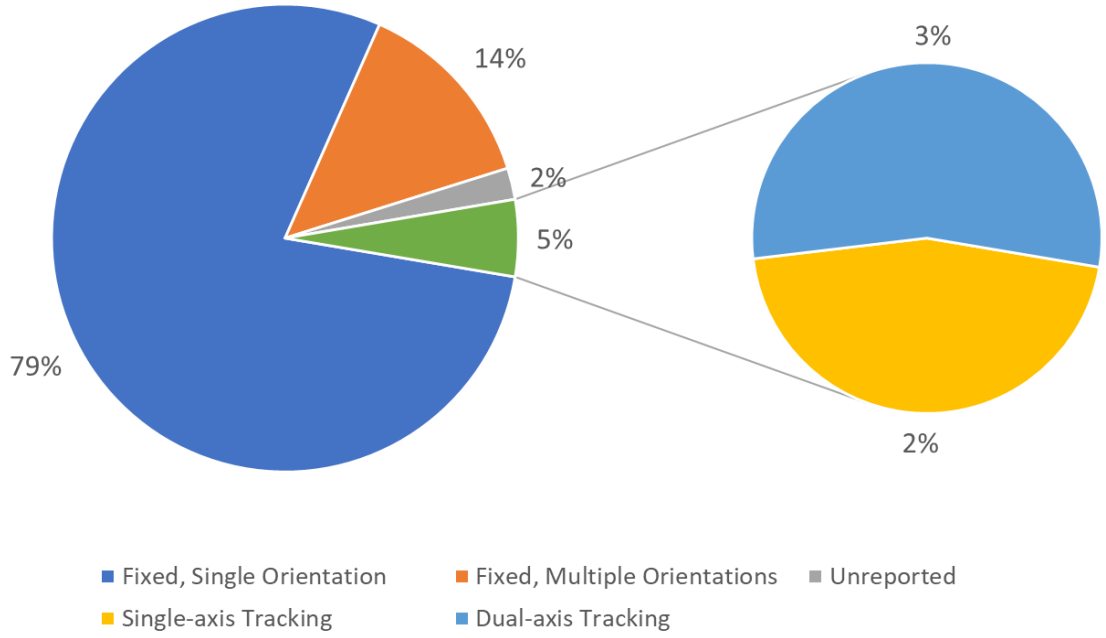

Figure 3 – Reported types of installations

Overall, inferred specifications yielded excellent results. Although production data using reported specs resulted in higher error most of the time, the results were still acceptable. There were, however, systems for which simulated data did not provide a close match with the measured data. Figure 4 and Figure 6 show results for the systems with the lowest error using inferred and reported specs, respectively. Figure 5 and Figure 7 shows results from systems using inferred and reported specs, respectively, that resulted in typical (median) error. In some cases, errors can reflect problems with either the measured data or the physical system itself. Degradation and soiling on some systems caused an increase in error and although CPR worked to identify system degradation rates and soiling patterns, more research is needed before that work yields a significant reductions in error.

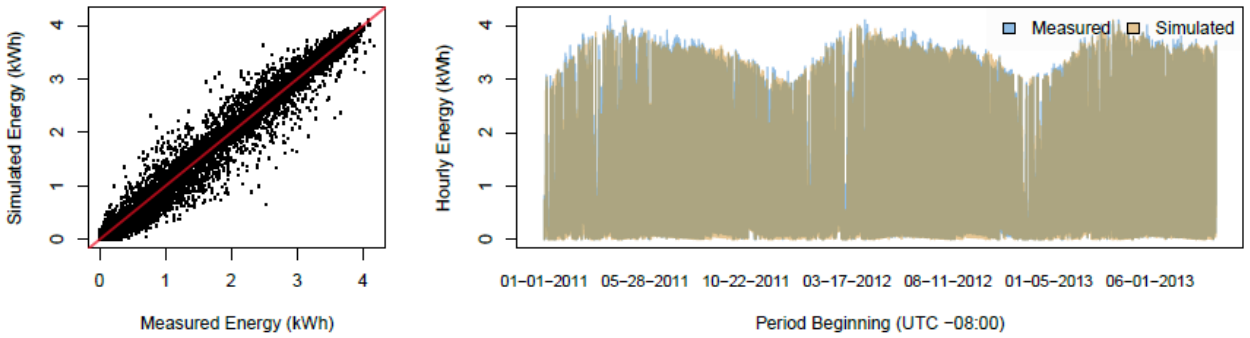

Figure 4 - Inferred specs with lowest error (PGE-CSI-24017 at 6.3% hourly rMAE)

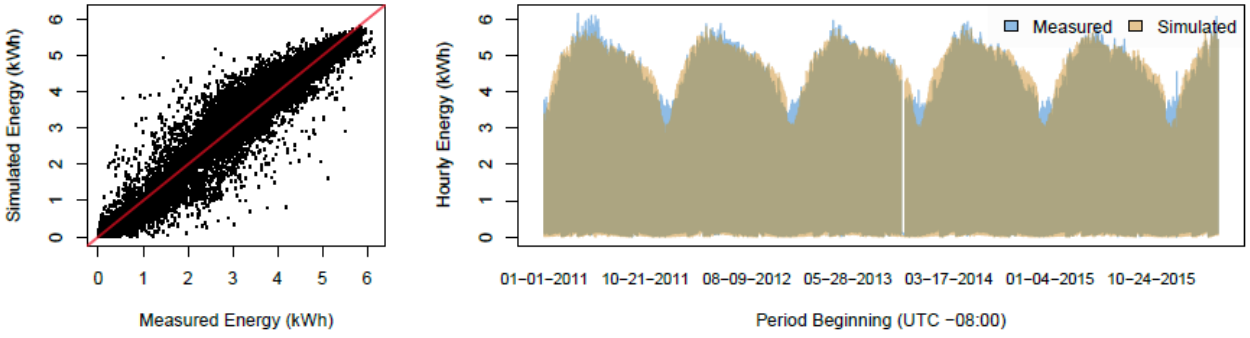

Figure 5 - Inferred specs with median error (SCE-CSI-13299 at 10.1% hourly rMAE)

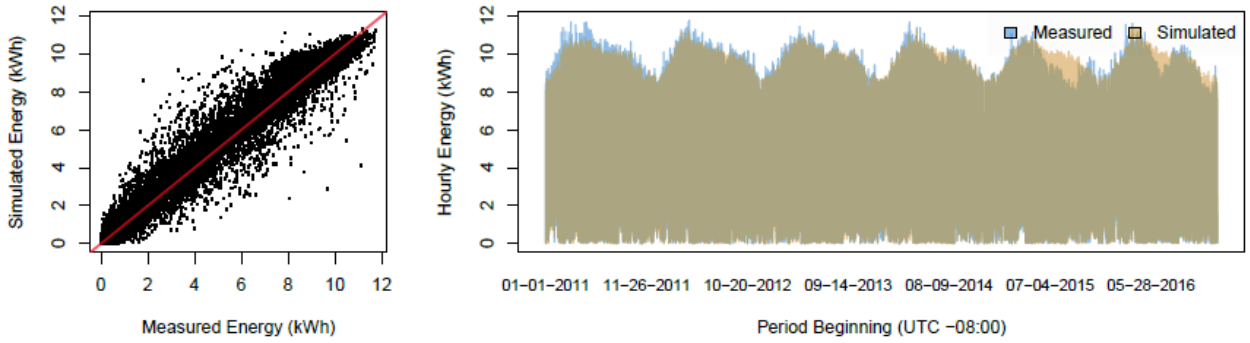

Figure 6 - Reported specs with lowest error (SCE-CSI-09966 at 7.2% hourly rMAE)

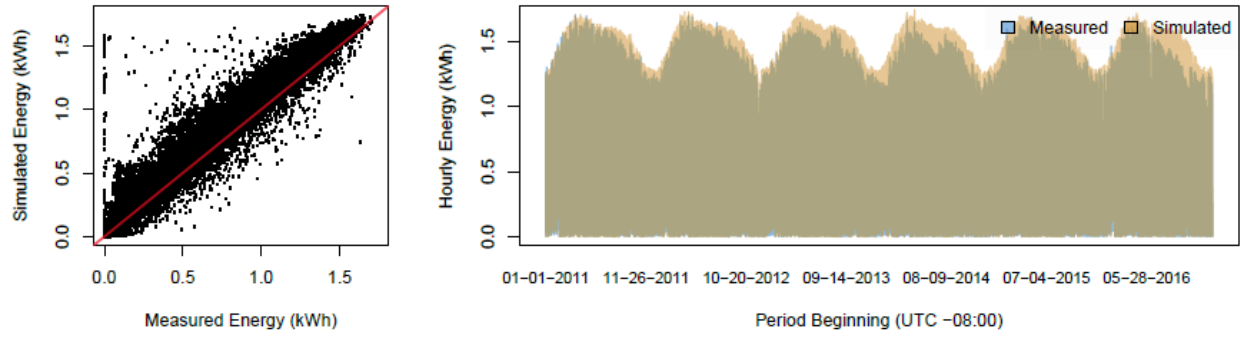

Figure 7 - Reported specs with median error (SCE-CSI-06793 at 16.7% hourly rMAE)

Although it makes sense to use all valid measured data when calculating error, it can also be instructive to compare production for individual days as a way of spot checking the data. Figure 8 and Figure 9 each show a single day of production for PGE-CSI-24017—the system whose inferred specs had the lowest overall rMAE at 6.3%. On a sunny day, such as May 5, 2011, results match quite closely, as illustrated in Figure 8. There are, however, some times when results are less exact. Production on January 19, when there was considerable cloud cover in the morning, is plotted in Figure 9. The reported specs for PGE-CSI-24017, with 8.5% rMAE overall, yielded similar results, albeit with slightly lower peak production on both days.

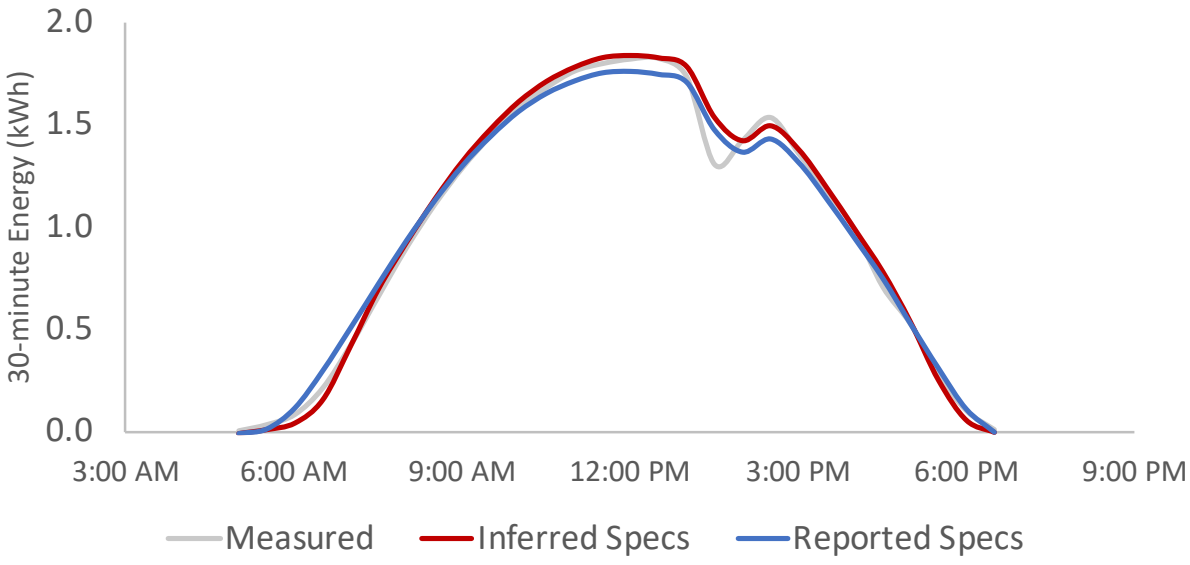

Figure 8 - PGE-CSI-24017 production comparison for May 5, 2011

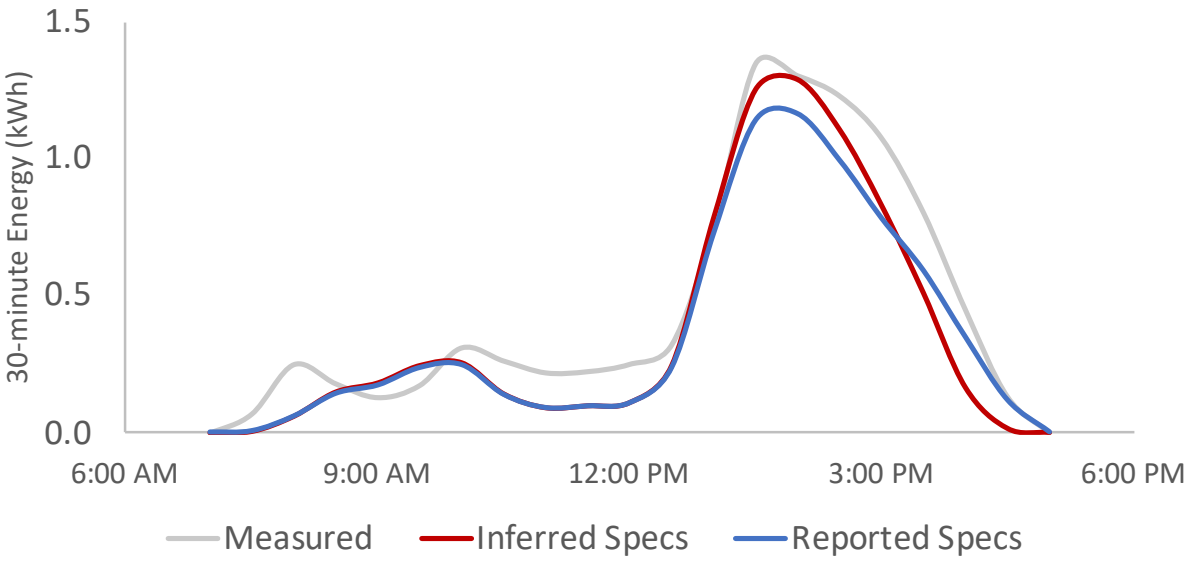

Figure 9 - PGE-CSI-24017 production comparison for January 19, 2011

So far, this report has focused on hourly error for individual systems. The distributed nature of these systems and the diversity in system orientation means that the aggregated “fleet” output as a whole is likely to have a significantly lower level of error as well as much less variability in output from one moment to the next.

Figure 10 shows fleet production on a clear day, when one might expect system output to be relatively smooth, but because the fleet is spread out geographically and has systems with a variety of orientations, this curve is wider and flatter than one would see with a single system with the same rating. Figure 11 shows fleet output on a day with variable cloudiness and illustrates variability in fleet output is reduced when compared to a single system on the same day, as shown in Figure 12. Because output for most zip codes in the *DER production database* will come from a single system, the output should not be considered representative of the entire fleet of systems installed in that zip code.

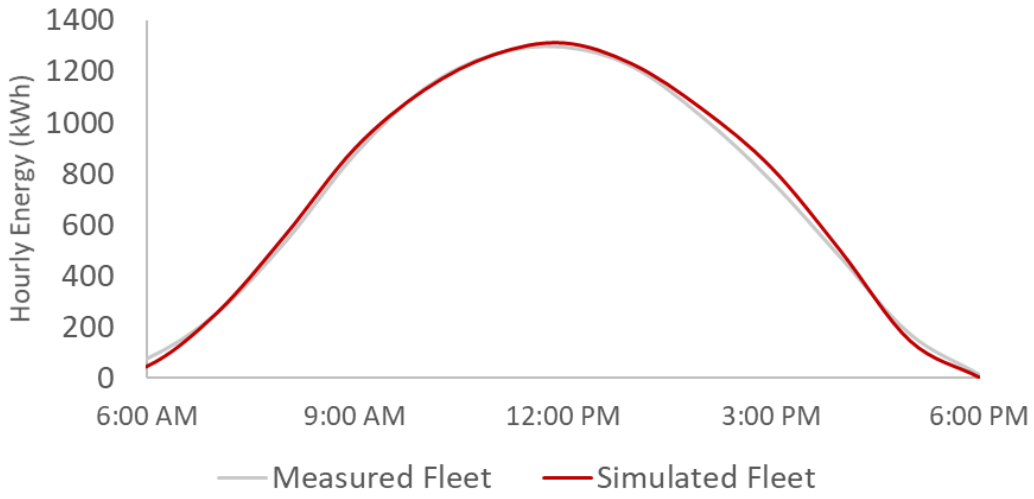

Figure 10 - Fleet production August 23, 2015

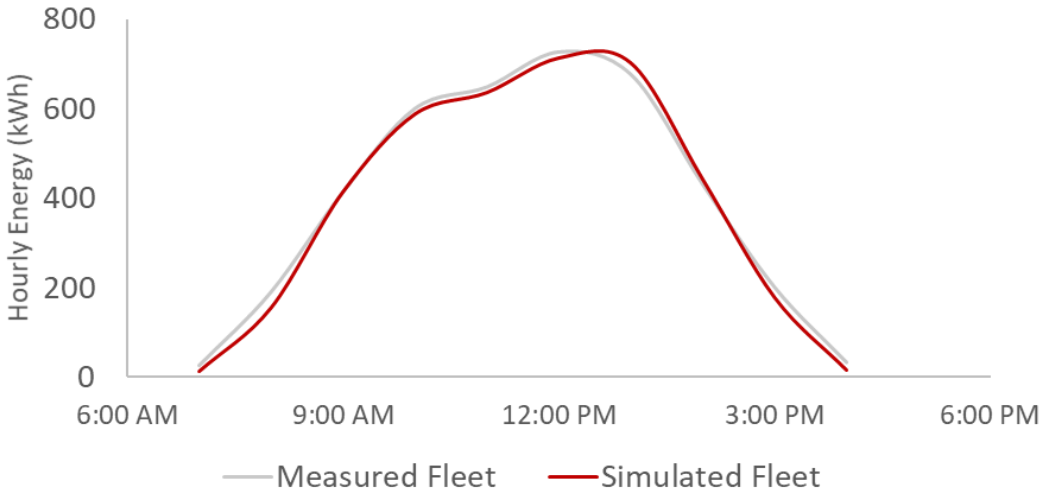

Figure 11 - Fleet production January 8, 2016

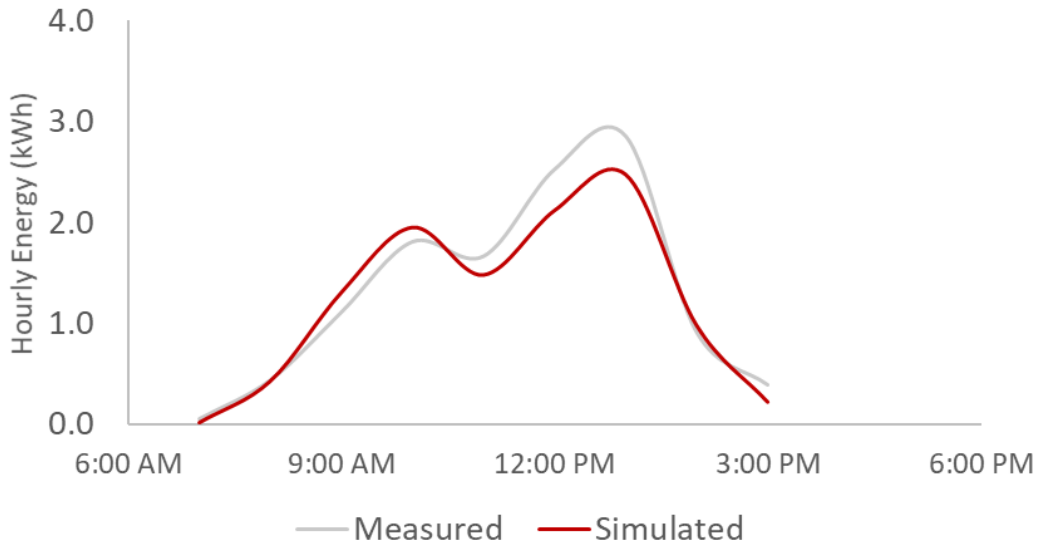

Figure 12 - PGE-CSI-00221 production January 8, 2016

Figure 13 compares simulated PV fleet output versus measured fleet output (combined output for every hour of all systems with valid measured data). With an hourly rMAE of just 4.3%, the fleet output begins to approach the error in the underlying satellite-derived irradiance data.

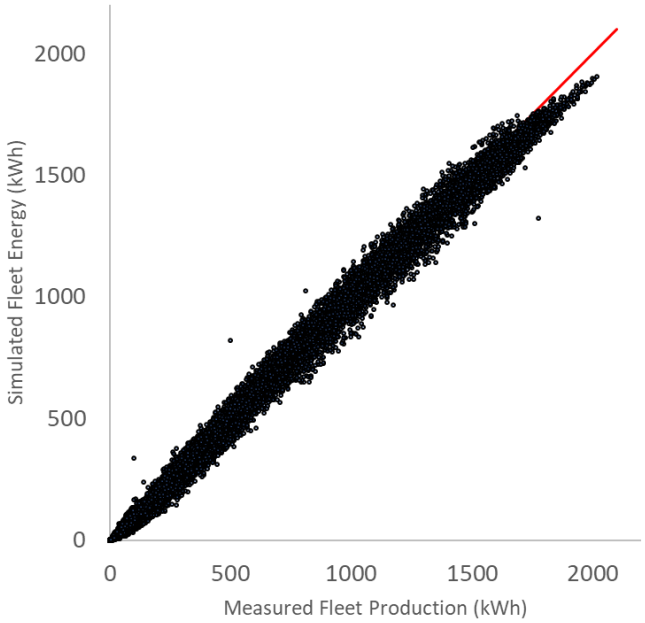

Figure 13 – Fleet output, measured vs simulated using inferred specs, 4.3% hourly rMAE

Future Research

Although the results from the inferred system specifications can be quite good, particularly with good sets of measured data to use for inference, there are several questions that may be worth exploring further. For example, how much of the error in the data from the reported specs was due to the assumed 90% general derate or the lack of good data on solar obstructions? Would a blended approach, where inferred solar obstructions are used in conjunction with reported orientation yield a much lower error? What is the effect on error of using a constant horizon with reported specs instead of azimuth-specific solar obstructions? What techniques could be employed to more effectively identify invalid measured data?

Construction of the DER Production Data Base

Replacing Data for Individual Systems

As part of the previously described process of comparing simulated data and selecting the replacement source with the lowest error, CPR filtered the measured data to remove periods that contained data considered invalid. The result was measured production data for individual systems that had fewer periods with invalid data, but also potentially had more missing periods. After identifying the best source for filling in the periods needed to create a complete continuous production dataset for the target period from January 1, 2011 through December 31, 2016, CPR created a custom tool to fill in all

missing periods in the filtered measured data for each system with the modeled production data for those periods from the selected source (inferred specs or reported specs).

Combining System Production by Zip Code

The CIDS contains a list of all interconnected solar PV (NEM) systems within PG&E, SCE and SDG&E service territories. Those CIDS systems are located in the 1,601 distinct zip codes indicated by green dots on Figure 14. The 504 systems in the CSI Measured Production Set are located in the 327 zip codes indicated by red dots in Figure 14. The per zip code capacity of the measured systems represented anywhere from 0.01% to 78.1% of the corresponding CIDS zip code capacity, with more than 98% of the zip codes having less than 10% of the CIDS capacity represented.

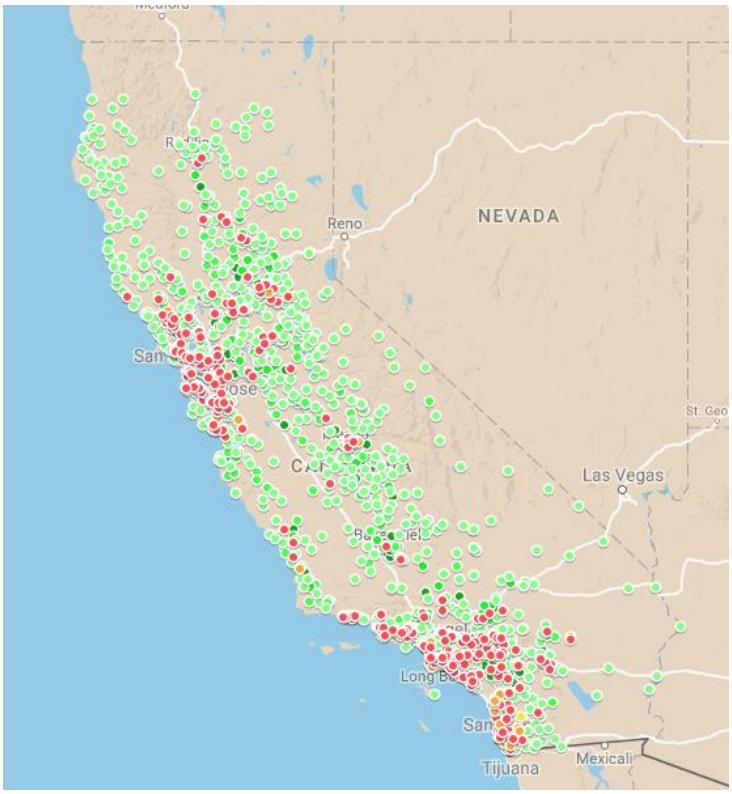

Figure 14 - Zip Code centroids for CIDS systems (green) and CSI measured data systems (red)

After CPR filled in the missing periods in the filtered measured data for individual systems, the individual system data was aggregated by zip code. After eliminating systems with an excessive amount of invalid or missing data, CPR had 414 systems in 288 zip codes on which to base the *DER production database*.

Using the AC rating listed in the CIDS for each system, CPR calculated the total AC capacity in each zip code and for each period divided the sum of the energy from all systems in the zip code by the total AC

zip code capacity to get normalized energy (kWh per kW AC). The per-zip code aggregate normalized energy production was then saved as a CSV file named for the zip code.

For example, there are two systems in the zip code 90254: SCE-CSI-03325, rated at 4.9 kW AC, and SCE-CSI-11519, rated at 27.7 kW AC, so the total rating for 90254 is 32.6 kW AC. The combined output from these two systems for the period beginning 8/1/2016 11:30 AM was 6.76 kWh, therefore the normalized energy for that period is 6.76 kWh ÷ 32.6 kW AC = 0.207 kWh per kW AC. All of the production data for this zip code was saved to a file named "90254.csv."

Production for systems of various sizes can be calculated by multiplying the normalized energy production value for each period by the system's AC rating.[7] However, it should not be assumed that this scaled production would be representative of a large number of systems with diverse orientations and locations.

DER Production Database

The DER Production Database was delivered as a set of 288 CSV files. Each file is named according to the zip code and contains two columns and 210,433 rows. The first row of each file is a header with the label, "Period Beginning (UTC -08:00)," in the first column and, "Energy (kWh)," in the second column. Rows two through 210,433 contain data.

The first column of each data row contains the time stamp of the beginning of each period in the format MM/dd/yyyy H:mm:ss, where MM is the two-digit month number from 01 to 12, dd is the two-digit day number from 01 to 31, yyyy is the four digit year from 0000 to 9999, H is the one or two-digit hour on a zero-based 24-hour clock, from 0 to 23, mm is the two-digit minute from 00 to 59, and ss is the two-digit second from 00 to 59. For this data set, the year value is always from 2011 to 2016, the minute value is always 00, 15, 30, or 45, and the second value is always 00. The period beginning at midnight (12:00 AM) on January 1, 2011 and ending at 12:15 AM would appear as, "01/01/2011 0:00:00." As indicated by the column header, all time stamps are in Pacific Standard Time or Coordinated Universal Time minus eight hours (UTC -08:00). The first time stamp in each file is 01/01/2011 0:00:00 and the last time stamp in each file is 12/31/2016 23:45:00. Time stamps can be converted to daylight saving time by adding one hour during the period when daylight saving time is in effect.

The second column of each data row contains the 15-minute normalized energy production, in kWh per kW AC, for the systems in that zip code. Values range from 0 to 1. To convert these values from energy to average power (kW), multiply each 15-minute energy value by four.

Aggregated Data

[7] AC rating here refers to the PTC rating of an array times the CEC-listed average inverter efficiency.

Aggregate output from a large number of PV systems can have a very different shape than output from a small number of systems. This occurs because a) larger groups of systems are more geographically disperse and therefore experience different weather conditions at different times and b) larger groups of systems have more orientational diversity which makes it more likely that some systems will be producing more energy at times when others are producing less energy. In general, fewer systems in closer geographic proximity to each other will result in higher variability in output. Figure 11 shows output for a fleet of several hundred systems on a day with variable cloudiness, while Figure 12 shows the increased variability in output from a single system. However, in this case, this is the only system in zip code 94040. Therefore, output for that zip code's fleet, even when scaled to match actual capacity for the entire zip code, would overstate variability for an actual fleet of systems.

Consider, for example, four systems located in the same zip code. One day of normalized (AC capacity weighted) 15-minute energy from these systems is shown in Figure 15 where the differences in output are clear, even though the output has been normalized and the systems are all located in the same zip code. When the aggregated output is normalized, as shown by the dark line in Figure 16, one can begin to see a broader flatter curve appear. This simple example uses only four systems. The effect becomes even more pronounced with more systems.

Since 86% of the zip codes covered by the *DER Production Database* contain data for only one or two systems, the data for a single zip code should not be used as a proxy for the combined output of all systems within that zip code, as the combined output would have a significantly different profile in most cases.

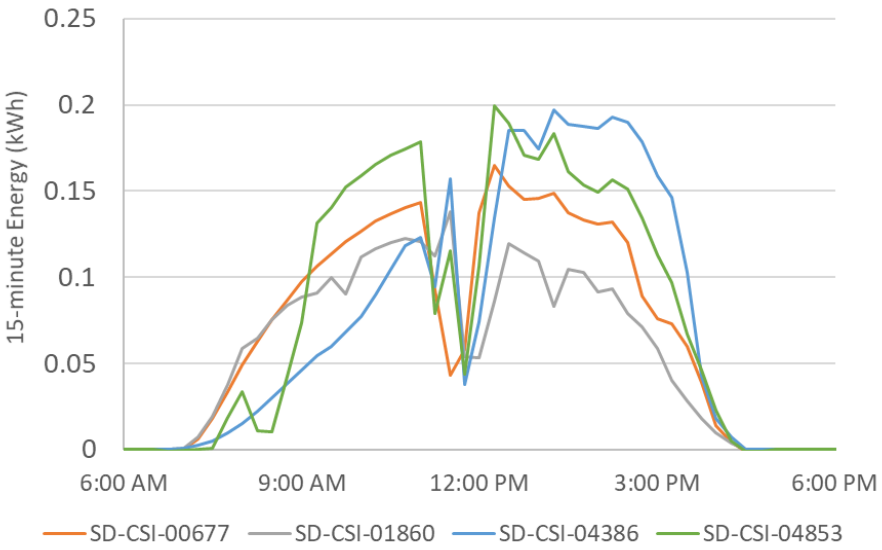

Figure 15 - Normalized production from four systems in the 92064 Zip Code

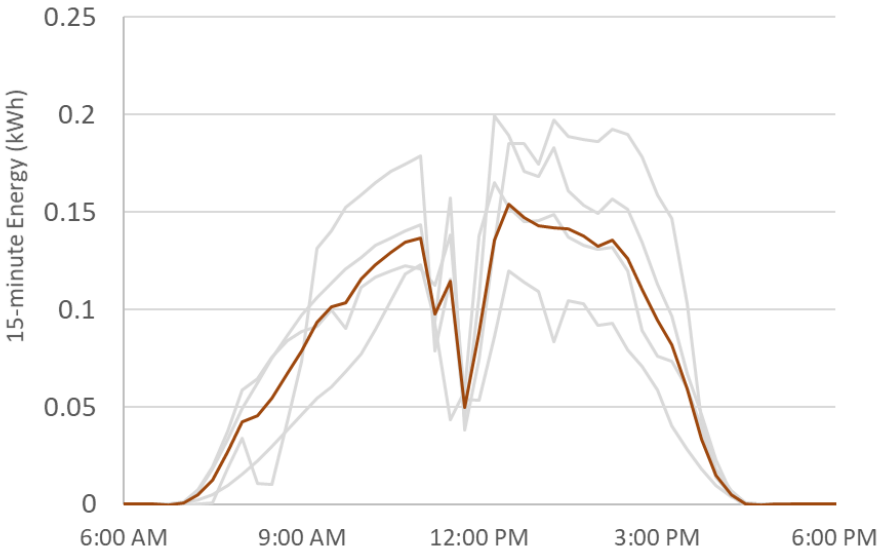

Figure 16 – Aggregate normalized PV production shown in red

DRP Mid-Term Growth Projections

CIDS data is used by CPR[8] as the definitive public record of installed DER capacity. It is updated monthly by the California IOUs which have the best available information about DER interconnections. While its spatial resolution in this public dataset is limited to zip code-level boundaries to protect customer privacy, the capacity of each system is published. Therefore, CIDS can be reliably used as a metric of as-built DER installed capacity.

System ratings and installation dates in CIDS may be used to refine the mid-term growth projections found in the IOUs' Distribution Resources Plans (DRPs). These plans, filed with the CPUC in mid-2015 and published on the CPUC website,[9] were developed in accordance with the final guidance of the Assigned Commissioner Ruling on February 6, 2015. Guidance related to the content and structure of the DRPs. DRPs address the IOUs' existing and future electric distribution infrastructure and planning procedures as they pertain to the incorporation of DERs into the planning and operation of the systems.

Given that the DRPs were filed in mid-2015 and the CIDS is updated each month, the CIDS capacity metrics could be used to update the DRP projected DER impacts. Data from 2015, 2016, and 2017 could be used to update the mid-term adoption forecasts. Unfortuanely, the DRPs do not indicate installed PV capacity, but instead report the impacts of projected PV capacity. Specifically, the DRPs provide mid-term projections of the change in peak load.

CPR therefore approached the problem by asking the question, "How could the CIDS data be used to project installed PV capacity, given the additional capacity data published since the time of the DRP filings?" The method described here could be used by the IOUs as an input to the peak load impact calculation, and it could be used in future years as well.

The growth patterns of the three IOUs followed a consistent pattern. In Table 1, the cumulative installed capacity C, in MW, is shown for each IOU in each year from 2013 to 2017. This is shown to increase each year. Also, the rate of change of the installed capacity $\Delta C/\Delta t$, in MW/year, is positive each year. The interesting pattern is the second derivative $\Delta^2 C/\Delta t^2$ shown in the third row.

[8] Monforte, et.al., *Improving Solar & Load Forecasts: Reducing the Operational Uncertainty behind the Duck Chart*, final report to CEC, May 2018, EPC-14-001.
[9] http://www.cpuc.ca.gov/General.aspx?id=5071

Table 1. CIDS capacity by IOU and derivative values

PG&E	2013	2014	2015	2016	2017
C	921	1229	1732	2359	2938
$\frac{\Delta C}{\Delta t}$		308	503	627	579
$\frac{\Delta^2 C}{\Delta t^2}$			196	123	-48

SCE	2013	2014	2015	2016	2017
C	590	815	1178	1618	2031
$\frac{\Delta C}{\Delta t}$		225	363	440	413
$\frac{\Delta^2 C}{\Delta t^2}$			138	77	-27

SDG&E	2013	2014	2015	2016	2017
C	211	314	483	685	830
$\frac{\Delta C}{\Delta t}$		103	169	202	145
$\frac{\Delta^2 C}{\Delta t^2}$			66	33	-57

In all three cases, the second derivative declines each year and turns negative in 2017. In other words, the rate of installations of new PV capacity has increased each year until 2017, the first year of declining rates.

The capacity data is plotted in Figure 17, where the solid lines show the reported cumulative PV capacity from the CIDS database. The phenomenon is shown as an inflection point around 2016—before this year the curve is concave and after this year the curve is convex.

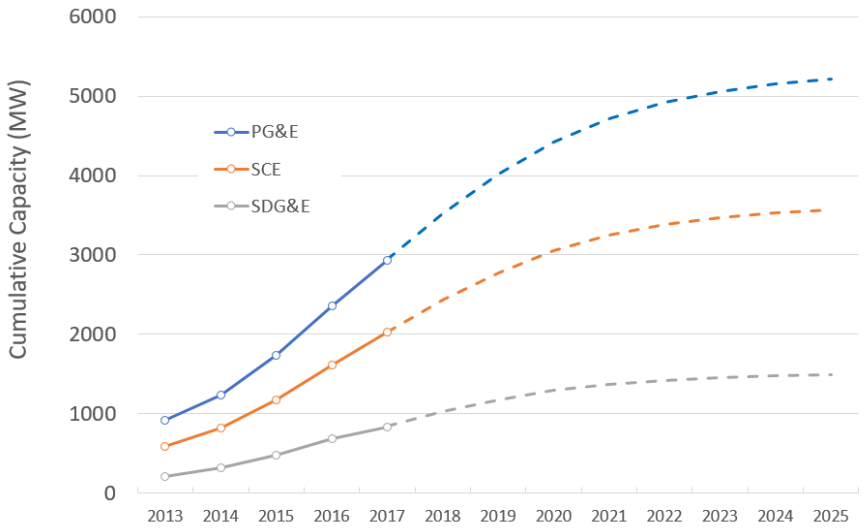

Figure 17. Cumulative capacity by IOU and mid-term projection

This behaviour suggests that the adoption of PV is following a logisitic function, sometimes described as a the Bass Diffusion Model.[10] The projected capacity (dotted lines) is a fitted logistic curve that best matches the original five points and extends to 2025, the final year of the DRP mid-term projections.

This curve is of the form

$$C(t) = \frac{L}{1 + e^{-k(t-t_0)}}$$

where *C(t)* is the capacity in year t, *L* is the maximum technical potential for the IOU, *k* is a constant, and t_0 is the year of the curve's inflection point.

Mathematically, t_0 is the mid-point, where half the ultimate capacity is reached. In the future, deviations from this curve are possible due to changes in market pricing, policy, and technology. With this caveat, the installed capacity in 2016 could be considered an approximate estimate of half the installed capacity. Using this measure, the ultimate amount of behind-the-meter PV to be installed in the future would be approximately $2 \times C(2016)$, or 4700 MW at PG&E, 3200 MW at SCE, and 1400 MW at SDG&E.

[10] See, for example, https://cleantechnica.com/2017/04/14/simple-model-predict-future-solar-pv-adoption/

Next Steps: Future DER Production Database

The *DER production database*, in its current form, is limited to only a single technology (PV) and is based on only a limited sample of systems, typically one or two, within a given zip code. Also, not all zip codes are included, limiting analysis use cases to those zip codes in which measured data was availabe.

This begs the question, "For the future, how can we design a DER Production Database that would represent DER production profiles broadly with other DER technologies?"

To design such a database, it is useful to first consider how the database would be used. CPR proposes the following "strawman" use cases for a future database:

- **Developing California energy goals**. Developing statewide energy goals requires an understanding of grid impacts. For example, in setting targets for electric vehicle adoption or 100% renewable sourcing, the Database could be used to predict what the hourly load profile would look like.

- **Examining differential growth rate scenarios**. Suppose load grows at a 1% constant rate, PV capacity grows using the logistic functions described above, and EV grows at 5%? The Database should be able to support such a study using these input assumptions.

- **Defining technology requirements to meet energy objectives**. Suppose that a study sought to determine how much energy storage capacity (in MW and MWh) would be needed by 2030 in order to keep the peak load at a constant level? This question could be answered with appropriately scaled hourly production profiles.

- **Designing incentive programs cost-effectively**. Incentives could be used to target regions, step amounts, and rollout timing based on future growth scenarios, and this would be supported by the Database.

- **Impacts of new technologies**. Suppose new technologies, whether supply-side or demand-side, came into the mix. The Database should be usable to evaluate its impact on the electric grid.

The above use cases illustrate the value of having such a Database, so long as it is designed to support these types of analysis. This will require separation of production by technology, e.g., the PV production time series should be independently scalable from the energy storage charge/discharge time series. There should also be customer load data to inform dispatch analysis and test energy and peak demand response. For simplicity, the Database would not cover multiple years, but rather a single, typical year, so the individual time series could be scaled to represent future year impacts. Finally, there should be a clear mapping between the Database and the source data used to develop it, namely, the CIDS database used to chart installations at the customer level and the metered data (similar to the Itron data used for this work).

The relationship between CIDS, the metered data, and the DER Production Database is illustrated in Figure 18. CIDS is updated by the IOUs and includes the (non-confidential) customer-level resource

attributes. These attributes should contain sufficient information to model each attribute, such as PV rating and orientation, DR rating, and so on. The Sample Interval Meter Data, a generalized version of the Itron PV dataset used for this project, would include all DER interval meter measurements, mapped to the CIDS system record. The metering data would include not only the DER itself, but other DERs at that customer location and the net customer load. This dataset would span multiple years.

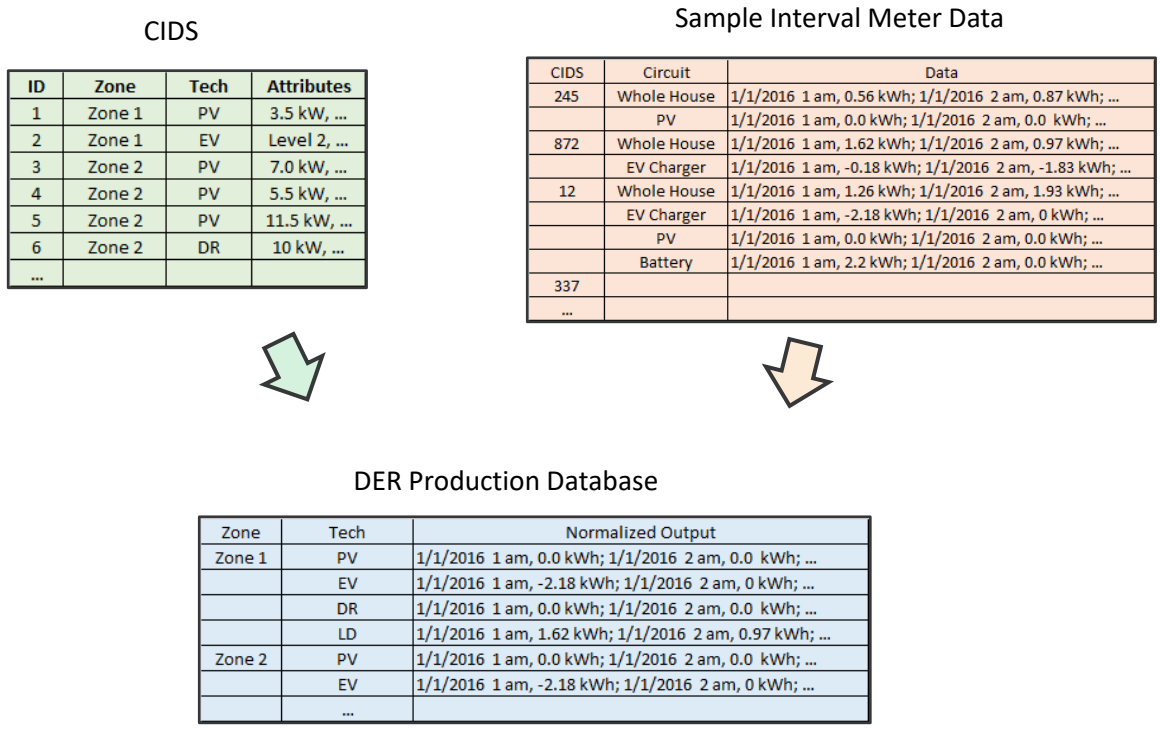

Figure 18. Illustration of CIDS, metered data, and DER Production Database

The DER Production Database would be produced from the other two sources. As in this project, erroneous or missing data from the metered database would be identified and filled in through modeling based on the CIDS dataset. The details of modeling would be unique to each technology and technologies such as PV or building thermal load shifting may require location-dependent meteorological data. The DER Production Database would also be normalized to give unit output by zone (e.g., output for a 1 kW PV system in each zone), and would represent a typical, base year of 8760 hours. Finally, zones would not need to be zip codes, but could alternatively be defined by electrical circuits or substations, provided that the above mapping approach is observed.

The resulting dataset could be scaled according to analysis need. If in year *t*, 500 EVs are assumed for a particular zone, then the profile for that zone's EV time series is multipled by 500. Then if in year *t+1*, 550 EVs are assumed, then the same base curve is used, but the scaling factor of 550 is used. Scaling factors would be customizable by technology, zone, and year, depending on the nature of the study.

A few words are in order related to the creation of the above set of databases. First, the CIDS data would have to expanded in scope, requiring coordination between the CEC, the CPUC, and the utilities. Details of the above design would have to be worked out, such as the specific technology attributes to be included in CIDS, the definition of zones, and the metering intervals (hourly, 15-minute, etc). Additional study would be required to determine the number of metered systems required for a representative sample in each zone. Each technology would have to be clearly defined, such as what is meant by non-dispatchable demand response and what are its attributes. Models would have to be developed and agreed to for filling in missing data.