Importance of Input Data and Uncertainty Associated with Tuning Satellite to Ground Solar Irradiation

James Alfi[1], Alex Kubiniec[2], Ganesh Mani[1], James Christopherson[1], Yiping He[1], Juan Bosch[3]

[1] EDF Renewable Energy, San Diego, CA, 92128, USA
[2] Clean Power Research, Kirkland, WA, 98003, USA
[3] Dept. Applied Physics, University of Granada, 18071, IISTA-CEAMA, Granada, Spain

***Abstract* — High quality satellite solar irradiation data is used throughout the solar industry to perform energy estimates. The uncertainty of the raw satellite data has been shown to be low. Ground data is often used to correct satellite data but determining the uncertainty of the final dataset could be challenging since the traditional statistical uncertainty and error calculation methods have proven to be unrepresentative. In this paper the limitations of traditional statistical methods are explored along with alternative approaches to calculate a more representative uncertainty value for a long term dataset resulting from ground corrected satellite data.**

I. Introduction

The use of satellite data has become more prevalent in the solar industry for energy estimates and continues to gain popularity. Integration of satellite data into public tools such as PVWatts and NREL SAM has helped shift the industry away from sparsely populated typical meteorological year (TMY) datasets to satellite based solar resource files. Clean Power Research (CPR) is a leading vendor of high quality satellite data "SolarAnywhere®," processed using the latest algorithms from Dr. Richard Perez. SolarAnywhere datasets do not have any ground corrections applied, which allows the end user to perform at their discretion. The reported U95 uncertainty of the raw satellite data is 5%, and by performing ground corrections there is an opportunity to remove local and seasonal biases and lower the uncertainty.

There are many methods used to perform a ground-satellite data correction. Simple linear regression can be used with either hourly or daily data to correct for bias present between the ground and satellite data. More complex non-linear methodologies such as those used by CPR attempt to correct the satellite data through multiple channels such as clear sky, cloudy conditions, and seasonality. Independent of the tuning method used to correct the satellite data, the resulting long term dataset must have a representative uncertainty reported. Ideally the method(s) used to calculate the uncertainty would be a statistically sound methodology that could be used in conjunction with any ground-satellite tuning process.

II. Limitations of Traditional Calculations

Traditional calculation methods to determine uncertainty have proved to be insufficient in capturing the uncertainty of the final long term dataset. For a ground-satellite correction based on least-squares regression, uncertainty is driven by residuals and the variability of the input dataset. While these methods typically produce accurate uncertainty results, they have been found to be insufficient for solar irradiation ground-satellite corrections for a number of reasons: 1) The resulting long term average of a ground-satellite correction is dependent on the time period that is being used for regression, thus simply looking at the residuals from the regression would not account for the uncertainty and error that is present from correlating higher than average, lower than average, or outlier years. 2) All traditional uncertainty methodologies assume that the relationship between X and Y is linear. The fact that the long term average is dependent on the time period used would imply that while the relationship is almost linear, there are nonlinear artifacts that affect the correlation. 3) Solar data is variable by nature. Traditional methods look at standard deviation or standard error relative to the mean of the dataset to estimate uncertainty, which works well for a manufacturing line setting or for calculating measurement uncertainty. Applying the same method to estimate uncertainty of solar data is not representative as solar irradiation varies constantly every day and hour by nature. While ground-satellite correlation on sites with higher variability are more uncertain due to the nature of correlation, estimating the uncertainty in solar data requires a different approach than traditional methods.

For ground-satellite corrections that are not based on least-squares regression, standard error or fit metrics are often used to estimate the uncertainty of the dataset. This study will attempt to quantify the uncertainty associated with those error statistics.

Independent of the tuning method, the nature of the input data itself can drive the final result from the tuning process. Beyond tuning uncertainty there is also uncertainty present due to the amount of input data, time of year of input data, and local climate region. Traditional calculations are not able to capture the uncertainty due to the error contributed by factors outside of the tuning process (e.g., input data), but the inherent ground data errors must be propagated in the final uncertainty of the corrected satellite dataset.

III. Importance of Input Data

The ground data itself could introduce uncertainty to the tuning process in three ways: 1) The length of the input data. More input data will usually have a lower uncertainty than less input data. A larger or longer set of input data will be a more representative sample of the population and lower the uncertainty. However, Figure 1 shows that this relationship is non-linear. 2) The specific time of year that the data covers. Different times of the year have different seasonal biases, which can either increase or decrease uncertainty. For example, data sampled in winter may have a higher uncertainty whereas data sampled in summer may have a lower uncertainty. 3) The local climate of the site. Different sites and climates have differing uncertainties due to the length, time of year that the data is sampled, and the variability in the climate itself.

CPR has developed a method for quantifying these uncertainties:

1. Data Selection: Minute level Ground Irradiance data from the SURFRAD and ISIS stations were used for the study. The choice to use these datasets was due to the long term nature of the measurements (10+ years), the quality of the sensors and measurements, and the geographic spread of the stations.
2. Data Quality Control: Minute level ground data was converted to hour ending averages and compared to hour ending SolarAnywhere data. Bad data and night time values were removed.
3. Varying input data length and time of year: Data from 2005 through 2015 was used for all sites. The data was broken up into 1-24 month segments through the 2005-2015 period. For each site there would be 132 one month segments, 131 two month segments, 130 three month segments, and so on.
4. Tuning: CPR's dual sliding window tuning [3] was performed and tuning parameters were determined for each month segment, and applied to the full 11-year time period. The mean bias error (MBE) between the ground and the newly tuned 11-year data was calculated. This step was repeated for all sites and each set of month segments (1-24 months).
5. Results: The MBE was calculated for each of the resulting tuned datasets and the standard deviation of the MBEs were calculated for all of the monthly segments and for each site. The resulting standard deviations of the MBEs were graphed.

The long term data was separated into rolling fixed length periods of time to determine the effects of using data from different times of year and different years on the tuning process. Varying the lengths of the fixed rolling periods allows for the impact associated with length of time to also be quantified.

From the resulting MBE distributions, the uncertainty can be quantified by looking at the standard deviation. Quantifying the uncertainty of the tuning process based on varying location, length of ground dataset, and time of year allows for the determination of the minimum number of months needed to achieve a certain level of uncertainty. Based on this it is possible for an uncertainty to be reported for a tuning process based on the region and the amount of ground data present for tuning.

Figures 1 and 2 show the results of the CPR tuning based methodology using rolling and varying lengths of input data. There is a clear trend indicating that more months of ground data used for the tuning process, the lower the associated standard deviation of the MBEs of the tuned data. At around 11 months of ground data the improvement in the standard deviation of MBE starts to become asymptotic. The relative decrease in standard deviation per added month of data becomes minimal after 12 months of ground data is used for tuning.

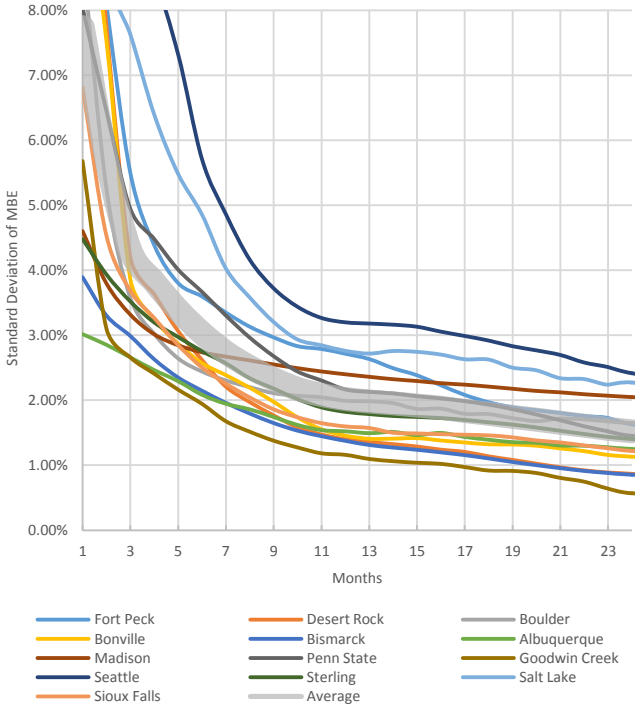

Figure 1: Standard Deviation of Tuned MBEs

Some sites require less ground data than others to reach a low level of uncertainty. For example, locations such as Madison, Bismarck, Desert Rock, Bonville, and Albuquerque all have standard deviations of MBE below 2% when using only 12 months of ground data for the tuning process. These results could be used to determine the ideal amount of ground data that needs to be collected based on proximity of the project site to the ground stations used in this study, or based on matching climate regions. For example, the Madison site has a very low standard deviation even at the 9-month sample-level. This would suggest that at locations near or similar to the Madison site, 9 months of input ground data for tuning may have an acceptable level of uncertainty. For a broader picture the average standard deviation of the 14 sites may be more useful. If 2% standard deviation is the maximum acceptable level of uncertainty associated with the tuning process, then 12 months of data on average has a standard

deviation of 2.02% and 13 months of data on average has a standard deviation of 1.98%. These results can be used to assign an uncertainty associated with tuning for a given length of ground data and sites with similar climate or sites located near one of the 14 stations used.

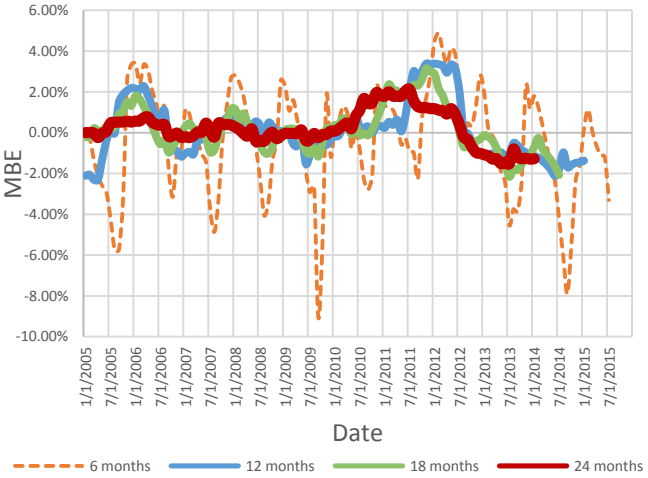

Figure 2: Tuning Results for Varying Rolling Periods at the SURFRAD station located in Desert Rock

Figure 2 illustrates the impact of different time periods on the tuning process. The MBE which is on the y-axis changes as the time period changes. Clearly tuning processes are sensitive to the time period, even if the length of time is the same. A clear seasonal signal is present, especially at the 6-month level. This is due to the misrepresentative nature of only sampling the summer or winter months. A sample with an exceptionally sunny or cloudy bias will cause the tuning to perform poorly as seen at Desert Rock 8/1/2009 at the 6-month sample level. As more data is used, the results have reduced variance because the sample becomes more representative of the entire dataset, implying a reduction in the uncertainty associated with tuning. Desert Rock is an ideal location to show limited uncertainty associated with tuning due to the relatively low variance in the irradiance at this site.

Plotting all the MBEs after tuning for all sites provides insight on the normality of the tuning process. Figure 3 is a histogram of MBEs for all sites using 12-24 month segments.

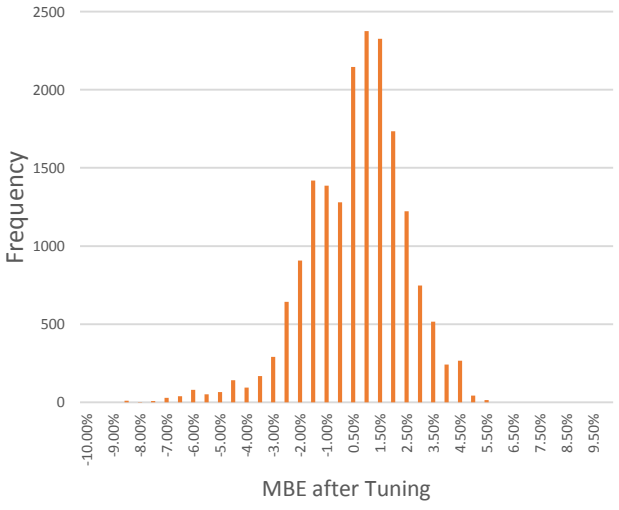

Figure 3: Histogram of MBEs for all sites 12-24 month segments

Month segments with lengths of 1-11 months were not included in this figure because those lengths of time were determined to be too sensitive to seasonal biases and only enhance the negative skew already present. The negative skew is due to the effect cloudy months, or month segments that include disproportionate amounts of winter months, have on the tuning and in turn the uncertainty. All MBEs for month segments 12-24 for all sites are present in Figure 3. This shows all the possible resultant MBEs after tuning. The MBEs greater than ±3% represent the less than ideal tuning scenarios for above average cloudy periods, periods that disproportionately sample summer or winter months, sites with more variable climates, or a combination of all three. However, 90.30% of MBEs fall below ±3% so the uncertainty for all sites with 12-24 months of ground data is low.

IV. Basis for a New Uncertainty Calculation

Noting the limitations of traditional uncertainty calculation method, and the need to have a representative uncertainty reported for each tuned dataset, an iterative Monte Carlo based simulation should provide the most representative uncertainty. By performing an adequate number of iterations, the results should converge and a distribution should be formed by the results. Assuming the resulting distribution (of annual kWh/m^2 in this case) is normal, relative standard error could be used to generate a more representative uncertainty value.

There are many ways that a Monte Carlo simulation could be applied to the data. Based on what is known about ground-satellite correlations, ideally the amount of data used in each iteration and the time period used for each iteration would change. To explore the applicability of a Monte Carlo approach, a site was chosen with 30 months of ground data. A total of 46,656 iterations were run consisting of a ground-satellite correction based on every possible combination of 30 months to form a 12 month period. The resulting distribution as seen in Figure 4 was tri-modal. A deeper look at the results revealed that each of the three distinct peaks in the distribution represented each of the three years that the data was sampled from.

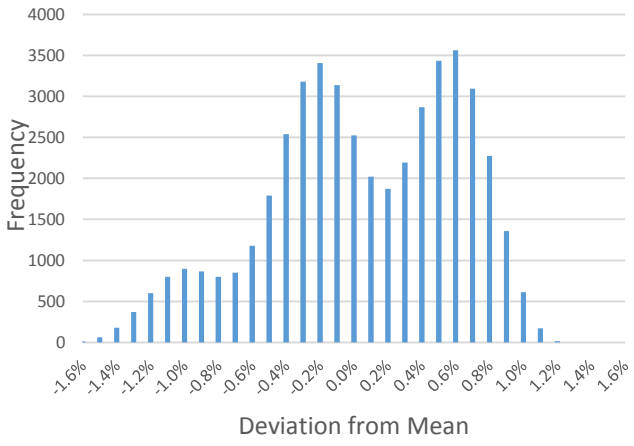

Figure 4: Histogram Based on 12 Month Combination Method

The Monte Carlo simulations for combinations of 12 month periods confirm that the result is dependent on the time period used. Also, this method is not adequate to calculate uncertainty because of a few reasons, 1) it ignores the impact of including more than 12 months of ground data in each iteration, 2) longer amounts of ground data will result in exponentially larger number of possible combinations and the computing limit of standard computers might be reached quickly, and 3) the resulting distribution might not be normal in some cases, which would make any uncertainty calculation difficult.

V. Proposed Methodology

The methodology developed to calculate uncertainty for ground-satellite correlations combines bootstrapping and a Monte Carlo based approach. Bootstrapping provides the ability to run a countless number of iterations with a limited amount of data. Input ground and satellite data can be daily sums or hourly data, in either case the same uncertainty calculation process can be used. However, hourly data is recommended since more data points are available. There is no requirement on size of the dataset, but according to the results shown in Section III at least one full year of ground data is highly recommended in most cases.

EDF Renewable Energy has developed the following method for quantifying the uncertainty:

1. Data pairs were randomly sampled from the available dataset. The amount of data to be sampled from each month was randomly varied for each iteration between 50% and 100% of all available data pairs. Linear weighting was used to heavily weight smaller sample sizes because there is no need to run 100% of the samples more than once, and less possibilities exist as the sample percentage increases.
2. All of the data pairs for each of the 12 months (including all available years of data) were aggregated. The number of data pairs sampled for each iteration was based on the sampling percentage that was randomly determined for the iteration in step 1.
3. The ground-satellite correction was performed based on the randomly sampled data. In regards to this study, the correction was performed based on linear regression of the concurrent ground and satellite data pairs. An ordinary least-square coefficients were generated for each of the twelve months, and also an annual least-square coefficients were generated based on all sampled data pairs. Monthly coefficients were used to correct the long term satellite data. If the coefficient of determination for a given month was poor, then the annual coefficient was used for the given month.
4. An annual P50 long term mean (GHI kWh/yr) was generated.
5. The process was repeated at least 50,000 times or until convergence occurs.
6. Output of each iteration is the annual P50 long term mean and the random sampling percentage which was used for the iteration.
7. Deviation of each iteration from the annual P50 long term mean generated using the entire dataset was calculated. A histogram was created with the results from all of the iterations.
8. The relative standard error (RSE) was calculated representing the standard uncertainty (U68) resulting from the tuning process.

$$RSE = \frac{\sigma}{\mu}$$

Implementation of the described methodology could be done using a scripting language, FORTRAN, or Matlab. EDF Renewable Energy has developed a proprietary FORTRAN program which performs the Monte Carlo simulation and also provides the user with additional inputs and outputs.

VI. Results and Discussion

The proposed uncertainty calculation methodology was evaluated with hourly solar irradiance data from EDF Renewable Energy owned solar meteorological stations and also the SURFRAD meteorological stations. From the SURFRAD stations, three years of good quality ground data without missing values were chosen for this study. All simulations were run for 50,000 times and convergence was seen at this number of iterations.

Summary of the results from the study is shown in Table 1.

Station Location	No. of Months of Ground Data	Calculated RSE of the Tuning Process	Total Uncertainty (U68)
Blythe, CA	40	0.13%	1.45%
Mojave, CA	33	0.10%	1.45%
Moore's Crossing, TX	27	0.28%	1.47%
Long Island, NY	17	0.60%	1.56%
Corcoran, CA	24	0.12%	1.45%
Boulder, CO*	36	0.24%	-
Desert Rock, CA*	36	0.11%	-
Fort Peck, MT*	36	0.22%	-
Goodwin Creek, MS*	36	0.17%	-

Table 1: Results of Uncertainty Calculation Using the Proposed Method
*SURFRAD Meteorological Stations

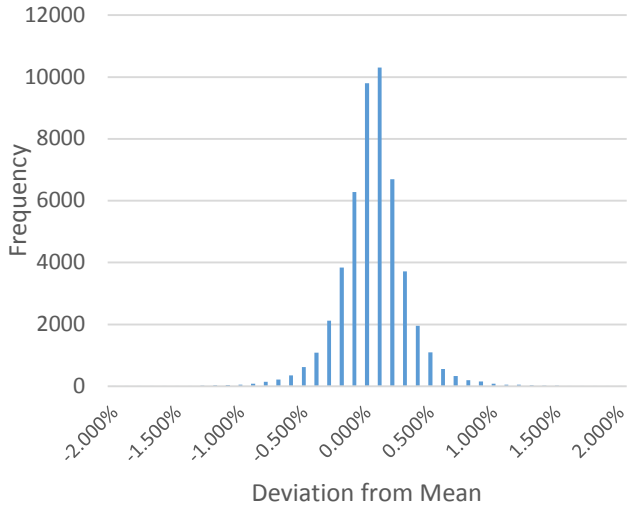

Figure 5: Histogram Based on the Monte Carlo Uncertainty Calculation Method

All sites had over a year of data and the results show a very low relative standard error and the distributions of the deviations from the mean for all sites were normal as shown in the sample graph. The results show that randomly choosing data pairs in random quantities for the tuning process will still converge to a long term mean, without much variance.

The relative standard error calculated based on the proposed uncertainty methodology is only an estimate of uncertainty of the tuning process. This uncertainty calculation method will be applicable for any tuning process regardless of the correction methodology used, such as EDF Renewable Energy's linear regression approach or the method used by Clean Power Research. The bootstrapping technique facilitates sampling data pairs randomly which are then used for each iteration with the chosen tuning method.

The underlying assumption for this methodology is that the ground data is true. However, while for the correction we assume that the ground data is accurate, the uncertainty in the measurement of the ground data has to be accounted for in the final uncertainty calculation. EDF Renewable Energy accounts for this uncertainty through various factors such as pyronometer accuracy, data logger measurement uncertainty, and installation and maintenance uncertainty (e.g. angle of pyranometers, cleanliness of pyranometers, etc.). Table 1 shows the total uncertainty of the tuning process based on EDF Renewable Energy's methodology. The uncertainty of the tuning is combined in quadrature with the uncertainty of the ground data to generate the total uncertainty of the long term P50 mean. Since the ground data uncertainty and the tuning uncertainty are independent they are combined in quadrature and not added. The total uncertainty in Table 1 for SURFRAD stations was not calculated because the ground data measurement uncertainty is unknown.

VII. Conclusion

For a solar project, the solar resource data is one of the largest and most important drivers to production estimates and the overall project economics. The reported uncertainty associated with the solar resource data is as important. Data without accurate uncertainty estimation limits the validity and confidence associated with the dataset. Thus, understanding the uncertainty of all input datasets is paramount to having an accurate production estimate.

The annual uncertainty of raw satellite data is typically reported as 5% (U95). With ground measurements there is an opportunity to remove the bias present in the satellite data and reduce the uncertainty even further. The importance of input data for ground satellite tuning is very apparent. As shown, at least 12 months of ground data will result in an acceptable level of uncertainty. By using more input data it was shown that the MBE was reduced, although the incremental benefit of having over 12 months of data becomes asymptotic. In general, for cloudier locations or locations with greater climate variability a larger input dataset is helpful to drive the uncertainty down further. Datasets of at least 12 months are necessitated due to the monthly and seasonal variation in the relationship between ground and satellite data.

Standard statistical metrics that are calculated relative to the mean of the dataset are not adequate enough in capturing the uncertainty of solar irradiance data because of the varying nature of the data. Therefore, a Monte-Carlo based method is proposed in this paper to randomly sample subsets of data from the available ground meteorological dataset to perform multiple iterations of the tuning process in order to calculate the uncertainty associated with the tuning process. This method estimates uncertainty of the ground-satellite data tuning regardless of the methodology used. The total uncertainty of the final long term corrected dataset must also include the uncertainty associated with the ground data. With the combined uncertainty of both the tuning process and uncertainty of the ground data, the combined U68 uncertainty is almost 50% lower than that of the raw satellite data. Besides having a lower uncertainty, any bias that exists between the satellite and ground data is removed.

References

[1] Manno, I. Introduction to the Monte Carlo Method. Budapest, Hungary: Akadémiai Kiadó, 1999.

[2] R. Perez, S. Kivalov, A. Zelenka, J. Schlemmer and K. Hemker Jr., (2010):Improving The Performance of Satellite-to-Irradiance Models using the Satellite's Infrared Sensors. Proc., ASES Annual Conference, Phoenix, Arizona.

[3] R. Perez, A. Kankiewicz, J. Dise, and E. Wu., (2014): Reducing Solar Project Uncertainty with and Optimized Resource Assessment Tuning Methodology. Proc. ASES Annual Conference, San Francisco, CA.

[4] Perez, R., J. Schlemmer, K. Hemker, Jr., S. Kivalov, A. Kankiewicz and C. Gueymard., (2015): Satellite-to-Irradiance Modeling – A new version of the SUNY Model. 42nd IEEE PV Specialists Conference, New Orleans, LA.