Correlating Photovoltaic Clear-Sky and Actual Output to Sky Cover Measurements

David Daily[1], David Holmberg[2], and John Baras[1]

1) University of Maryland, College Park, MD, 20742, USA
2) National Institute of Standards and Technology, Gaithersburg, MD, 20899

***Abstract* — As distributed energy resources are added to building systems, and variable priced electricity tariffs are introduced, facility owners may profit from understanding the expected performance of their building systems. Hourly, forecast energy generation data from photovoltaic sources allows owners to manage energy storage and consumption for their optimal benefit. This paper documents an empirical method which uses publicly–available sky cover observations to estimate the output of a photovoltaic (PV) array. Two functions are developed which together relate PV array output to sky cover observations reported by the National Oceanic and Atmospheric Administration's National Weather Service (NWS) for a nearby airport. Results are evaluated by comparison of estimated PV array output with actual output. Results show this approach can produce useful estimates of hourly and daily PV system output which may be applied to forecast weather data in order to manage PV array outputs. Although some large deviations are observed between forecast and actual PV output, the method presented uses data that would be available to a building owner with insufficient on-site instrumentation beyond a meter.**

***Index Terms* — buildings, correlation, photovoltaic, sky cover, sky condition.**

I. Background

One to 24 hours in-advance, facility owners may need to estimate the power that will be produced from the photovoltaic array(s) on their roofs. This estimation capability allows better planning for charging of thermal storage, and for management of loads to minimize both demand peaks and overall energy usage during peak electric price periods. The facility owner benefits from knowing power output throughout the day, ideally for tomorrow and some estimate of totals for the days to follow.

A commercial or industrial facility may have photovoltaic arrays, thermal storage, variable loads, a variable price electric tariff, and alternative fuels. The facility owner has the potential to save money by careful management of energy streams and scheduling of energy intensive processes outside of peak price periods. The facility owner may switch energy sources to fuel oil or natural gas for some processes when the electricity price is high. The owner may charge thermal storage when electricity prices are low (such as at night). Some processes might be shifted during the day to take advantage of periods of lower electric prices.

If the weather forecast for tomorrow calls for high temperatures, then it is likely that facility cooling load will be higher. However, if the sky is cloudy, moderating temperatures, this will reduce the facility cooling load, but also reduce any PV system output. On the other hand, if the day is very hot, but a weather front brings clouds in mid-afternoon, it is likely that cooling load will remain high for some time even as the PV generation is reduced. Meanwhile, the loss of PV generation coincides with the electrical price peak resulting in more expensive cooling. The facility owner would do well to prepare for this by having thermal storage fully charged at the time of the arrival of the clouds. This preemptive action then allows the facility owner to reduce the electricity used to cool the facility, shutting off compressors.

It follows that in the short-term, hourly and day in-advance forecasting is important for improving building operations. For PV system performance, many existing tools give results based on averaged, historical weather data (for a Typical Meteorological Year) and in large time steps (daily/monthly/annually) [3]. A popular tool, PVWATTS, performs an hourly simulation of in situ weather parameters, but reports out monthly [4]. A tool that predicts actual PV system performance based on forecast weather for the next 24 hours would be quite useful for planning building system operations.

This paper develops an approach for empirical correlation of Sky Cover (**SK**) to PV output. **SK** was chosen as it is a commonly reported parameter [2], and correlates strongly to irradiance, which translates into PV array output [6]. The NWS provides historically measured values of **SK** from various quality controlled weather stations [9]. They also supply forecasts of **SK** at grid points [7] which emphasizes the usefulness of this prediction method for facility owners.

The method has two steps: (1) create a clear-sky PV output model that is a function of Time of Day (**ToD**) and Day of Year (**DoY**), and (2) create a normalization model to account for **SK**. The end results are two values whose product yields an estimated PV output which is a function of only three parameters: **SK**, **ToD**, **DoY**. While the idea of normalizing the data to the clear sky value stems from work done by [Perez], the method's clear-sky PV model output is unique in its 3D nature and simple variables for the given application. Similar work has been accomplished using multi-variable inputs into a radial basis function network for forecasted PV output [1]. While this previous work acknowledges that an empirical approach is well suited to an existing installation where a facility owner has no instrumentation besides the electric meter, and no information about PV system components, their correlations are on-site sensor data based.

II. Clear-Sky PV Output

Before a facility owner can account for the impacts of the forecasted sky conditions on PV output, he/she must first have an estimate of what the PV output would be on any given day if the sky were clear. Call this the Clear Sky output, $\mathbf{E_{CS}}$, in kilowatt hours (kWh), over some measured time interval. The traditional approach for estimating $\mathbf{E_{CS}}$ is to use an available sky and PV system model that includes many factors such as: position of the sun, panel temperature, panel orientation, type of PV panel, type of inverter, and various degradation factors. These models can then be tuned to best match actual output [3]. On the other hand, the complexity to use such models is multiplied by the number of arrays, and becomes more difficult for existing installations where system details (characteristics of panels and inverters, orientation, shading) may not be readily available or overly burdensome to collect.

The approach implemented here for estimating output is strictly empirical. If one only sampled PV output data during clear sky daytime conditions, and then fit a 3D surface to a year's worth of this data, the result would be an empirical correlation between time and $\mathbf{E_{CS}}$ for every day of the year. This correlation would be subject to change over time due to system degradation and component replacement, but would be useful to provide the typical maximum output that could be expected from a PV array (or collection of facility PV arrays) under clear sky conditions.

For the typical facility, there may be no on-site sensor to measure sky conditions, which could be used to filter PV system power output for clear-sky versus cloudy conditions. Different ways exist to deal with this. One approach might be to plot measured PV output as a function of **ToD** and **DoY** and use some algorithm that can effectively "throw a blanket" over the data points such that the resulting surface fit is to the highest points (sunny conditions) while ignoring the low (cloudy condition) points. Alternatively, one might find a published source of sky condition data at a nearby location that could be used to filter out measurements taken under cloudy conditions. Both of these methods are computationally challenging.

For the current research, **SK** data was taken from measurements at the nearest (NWS) observation site: the Leesburg, VA, airport approximately 18 miles (29 km) southwest from the National Institute of Standards and Technology (NIST) Gaithersburg, MD, site where the PV array is located. Observed **SK** is reported [9] as shown in column one of Table 1, with the "Okta value" representing the amount of the sky dome that is covered by clouds in eighths, 0/8 being clear sky and 8/8 being completely overcast. Following the approach given in [Perez], numerical values were assigned to the different **SK** levels as shown in column 3 of Table 1. The **SK** measurements were reported in 20 minute intervals.

PV Output data from a NIST Administration Building PV array were reported as total energy produced (kWh) in 15 min increments. Since **SK** was reported in 20 minute intervals, the two datasets were normalized to 1 hour time steps. The PV Output data, as used for correlation purposes, took the sum of the energy produced over the hour while the **SK** data took the average of the Sky Cover value over the hour.

TABLE I
RELATIONSHIP USED FOR CONVERTING **SK** DATA TO NUMERICAL VALUES.

METAR Sky Cover Code	Okta Value	Associated Value
CLR	0	0.00
FEW	1 - 2	0.125
SCT	3 - 4	0.4375
BKN	5 - 7	0.75
OVC	8	1.00

In order to estimate the PV output under clear sky conditions, hourly measurements were filtered based on whether observed **SK** for a given hour was "CLR" (Table 1) for the entire hour. Night hours were also removed. This filtering resulted in clear sky values, although only an approximation due to the location difference between the **SK** sensor and the PV array. This will be discussed more below.

The MATLAB Curve Fitting Toolbox (Mention of specific commercial products in this paper does not imply endorsement by NIST or imply that these are the best tools for the job.) was used to fit a polynomial curve to the clear sky data. The curve has 2 degrees of freedom (X and Y) with both degrees to the second power.

Fig. 1 shows the resulting filtered data and surface plot. The equation for the surface is in the form:

$$\begin{aligned} E_{CS}(ToD, DoY) = {} & A - B*ToD - \\ & C*DoY - D*ToD^2 + \\ & E*ToD*DoY + F*DoY^2 \end{aligned} \quad (1)$$

Where $\mathbf{E_{CS}}$ is the PV Output during clear sky conditions [kWh], **ToD** is the time of the day, and **DoY** is the day of the year beginning on June 28th.

The resulting coefficients (A through F) are valid only over the measurement months (June 28th through September 18th), and only for the output from this array with **SK** observations measured at the Leesburg, VA airport (KJYO). In addition, any changes to PV array components plus degradation over time will impact this correlation.

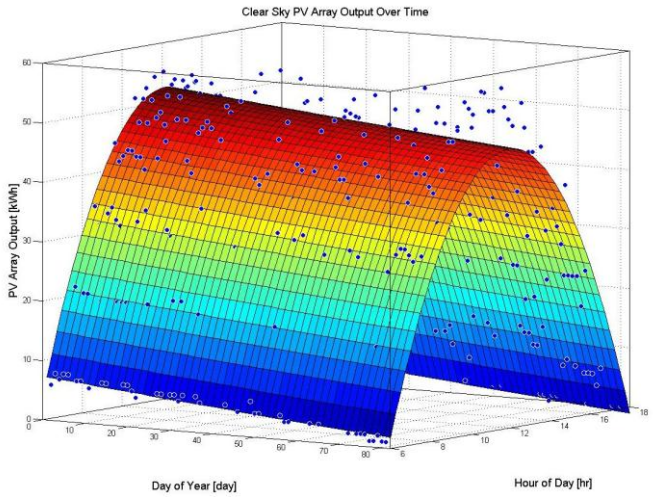

Fig. 1. Surface curve of E_{CS} along with clear sky PV array output data points ($\mathbf{E_{obs}}$). Curve is fit over the hour of the day and day of the year starting from June 28.

In some cases, the actual observed output ($\mathbf{E_{obs}}$) fell significantly below the predicted curve ($\mathbf{E_{CS}}$) seen in Fig. 1. This result is believed to be due mainly to the geographical distance between the **SK** sensor and the PV array. An abnormally low output value may occur when cloud cover is present above the PV array while clear sky is reported at the **SK** sensor location. To account for this, a Least Absolute Residuals method was implemented to reduce the impact of outliers. This method uses the absolute difference rather than the squared differences of the residuals to calculate the best curve. [5]

The opposite case where it is cloudy over the **SK** sensor and cloudless over the PV array is not applicable to this curve as any data with non-zero **SK** values were filtered out.

III. Estimating PV Output from SK Data

The next step is to perform a simple polynomial curve fit between the observed **SK** values and a corresponding normalization factor ($\boldsymbol{\mu}$) which is the observed PV array output divided by the expected clear sky array output:

$$\mu(ToD, DoY) = \frac{E_{obs}}{E_{CS}(ToD, DoY)} \qquad (2)$$

Figure 2 shows the resulting normalization factor for all hours in the summer data set along with the resulting correlation function. The equation for the function is a 4th degree polynomial in the form:

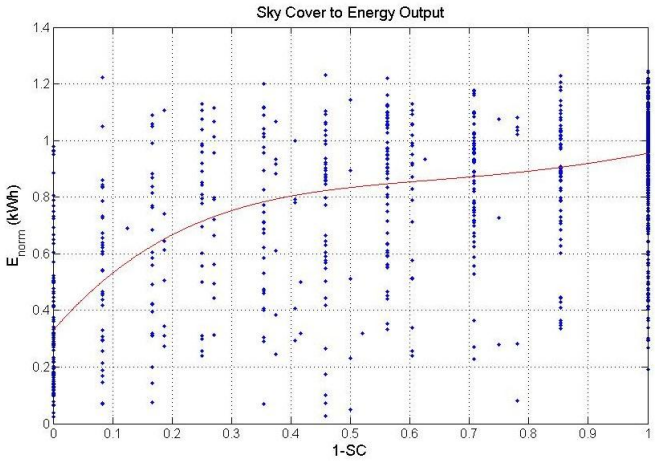

Fig. 2. Relation of sky cover to normalization factor.

$$\mu_{est}(SC) = A*SC^4 + B*SC^3 + C*SC^2 + D*SC + E \qquad (3)$$

where **SC** is 1-**SK**. **SC** is used rather than **SK** to give a positive correlation. A robust fit with bisquare weights [5], which increases the weight of each point based on its proximity to the function, is used to decrease the influence of outliers while strengthening the points closer to the trend.

For the data shown in Fig. 2, the resulting adjusted R^2 value was 0.41 with root mean squared error (RMSE) = 0.23. It may be expected that the goodness of fit would improve if the sky cover measurement were co-located with the PV array. Nonetheless, it provides some useful indication of PV array production that might be used together with SK forecasts and equation (1).

The normalization factor, $\boldsymbol{\mu}$, was found to be unreliable at the fringe hours (6am and 6pm) due to very small values of $\mathbf{E_{CS}}$ in the denominator of Eqn. 2. For this reason, data from these hours were not used in the calculation of $\boldsymbol{\mu_{est}}$.

Finally, predicted energy output ($\mathbf{E_{est}}$) may be found by multiplying the results of equations (1) and (3). This estimate can then be compared to the corresponding $\mathbf{E_{obs}}$.

$$E_{est}(SC, ToD, DoY) = \mu_{est}(SC) * E_{CS}(ToD, DoY) \qquad (4)$$

IV. Results

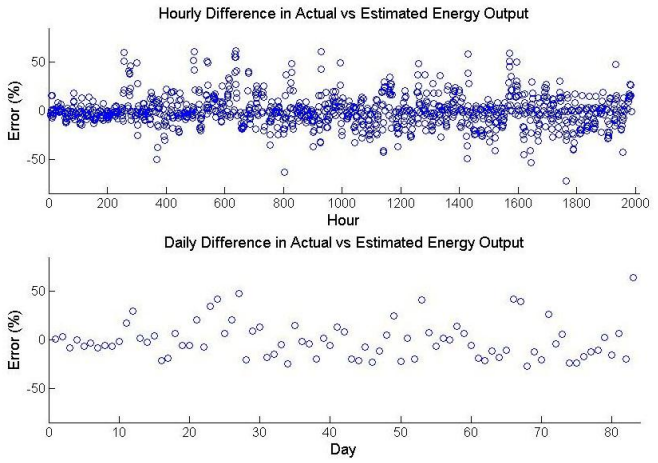

Fig. 3. Normalized error plots in the form of $(E_{est} - E_{obs}) / E_{CSpeak-hour}$ for hourly and $(E_{est} - E_{obs}) / E_{CSday}$ for daily time steps.

Fig. 3 shows the normalized error, $\mathbf{(E_{est} - E_{obs}) / E_{CSpeak-hour}}$ for hourly and $\mathbf{(E_{est} - E_{obs}) / E_{CSday}}$ for daily. A zero point indicates that the model matched the observed output perfectly. The hourly error is the difference in estimated and observed output normalized by the peak clear sky output (seen to be approximately 50 kWh in Fig. 1 for the period of year used in this analysis).Using the peak value for normalization provides a consistent reference point and also reduces the influence of the morning and evening hour errors. At these fringes, the percentage error between estimated and observed values can be very high if the normalizing value (either estimated or observed) is very close to zero. The daily normalization factor was, similarly, taken as the total production on a clear sky day. This can be seen as the area under the Fig. 1 curve for a given day (approximately 400 kWh for the period of year used in this analysis).

The RMSE for the normalized hourly differences (Fig. 3 top) is 16.2 %, versus 18.9% for the daily RMS error (Fig. 3 bottom). Despite the significant variation seen in the data around the SK curve fit in Fig. 2, the resulting error seen in Fig. 3 may be low enough such that the facility owner can have some useable estimate of hourly and daily PV array output based on a nearby SK sensor measurement.

Positive error percentages indicate $\mathbf{E_{est} > E_{obs}}$. Large positive errors seen in Fig. 3 can be attributed to the imprecision of the curve fit in Fig. 2 (data scatter) combined with the situation where sunnier skies are present to the west (SK sensor site), while overcast skies are present across peak hours at the PV array site. For example, array output may be reduced almost to zero by the presence of clouds (for example, a thunderstorm) while the SK sensor may measure clear sky. Conversely, the situation where a cloud passes over the SK sensor, with no cloud present over the PV array, yields a negative error. In reference to Fig. 2, some error is attributable to the inherent mismatch between observed cloud cover (where no distinction is made based on opacity of clouds) and irradiance, as evidenced by, for example, the results of [8].

VI. Conclusion

Results show it is possible to take historical PV output data with historical **SK** data, and develop a meaningful correlation to predict PV array output. This development is achieved through a 3D empirical model for estimating clear sky PV output as a function of date and time, in addition to a correlation of **SK** to normalized PV output. Conceivably, a facility owner could be given packaged code that takes as input **SK** data and historical PV array power to produce a correlation providing forecast power production as a function of forecast **SK** with meaningful results, using available NWS forecasts.

Additional uncertainty of a forecast has not been considered in the analysis presented here, but is addressed in [2]. The usefulness of this approach for estimating future solar array power output for any given time window will diminish as the forecast time moves farther out into the future. On the other hand, correlation could likely be improved significantly with the addition of a co-located **SK** sensor.

In future work, data may be collected from a new weather station installed on the NIST campus. Similar methods may be used to produce empirical correlations of PV array energy production to parameters such as global horizontal irradiance (GHI) and direct normal irradiance (DNI). Results may also be compared for three different PV arrays on campus. Finally, the accuracy of the correlation may be compared for weather observations versus NWS forecasts.

References

[1] Chen, Changsong, Shanxu Duan, Tao Cai, Bangyin Liu, "Online 24-h solar power forecasting based on weather type classification using artificial neural network," Solar Energy, 2011, pp. 2856-2870.

[2] Kim, Sean Hay, Godfried Augenbroe, "Using the National Digital Forecast Database for model-based building controls," Automation in Construction, Volume 27, November 2012, Pages 170-182, ISSN 0926-5805,10.1016/j.autcon.2012.05.012. <http://www.sciencedirect.com/science/article/pii/S0926580512000854>

[3] Klise, Geoffrey T., and Joshua S. Stein. "Models Used to Assess the Performance of Photovoltaic Systems." Rep. no. SAND2009-8258. Sandia National Laboratories, Dec. 2009. Web. 25 Jan. 2013. <http://energy.sandia.gov/wp/wp-content/gallery/uploads/098258.pdf>.

[4] Marion, B., M. Anderberg, R. George, P. Gray-Hann, and D. Heimiller, "PVWATTS Version 2 – Enhanced Spatial Resolution for Calculating Grid-Connected PV Performance," in Proceedings of the NCPV Program Review Meeting, October 14-17 2001 <http://www.nrel.gov/docs/fy02osti/30941.pdf>

[5] MATLAB, MATLAB Curve Fitting Toolbox User's Guide R2012b, The MathWorks, Inc. 2012. <http://www.mathworks.com/help/pdf_doc/curvefit/curvefit.pdf>

[6] Maxwell, E. , R. George, and S. Wilcox, "A Climatological Solar Radiation Model," in Proceedings of the 1998 Annual ASES Conference, June 14-17, 1998, pp. 505-510 <http://www.nrel.gov/gis/pdfs/proceedings_solar98.pdf>

[7] NDFD, the National Digital Forecast Database, National Weather Service, NOAA, 2004. Washington DC. <http://www.nws.noaa.gov/ndfd/>

[8] Perez, Richard, Kathleen Moore, Steve Wilcox, David Renné, Antoine Zelenka, "Forecasting solar radiation – Preliminary evaluation of an approach based upon the national forecast database," Solar Energy, Volume 81, Issue 6, June 2007, Pages 809-812, ISSN 0038-092X, 10.1016/j.solener.2006.09.009. <http://www.sciencedirect.com/science/article/pii/S0038092X06002404>

[9] QCLCD, the Quality Controlled Local Climatological Data, National Weather Service, NOAA, 2005. Washington DC. <http://cdo.ncdc.noaa.gov/qclcd/QCLCD>.