

A Simple Model for Estimating the Diffuse Fraction of Solar Irradiance from Photovoltaic Array Power Output

N. A. Engerer^a and Y. Xu^a

^a*Fenner School of Environment and Society, The Australian National University, Frank Fenner Building,
Canberra, ACT, Australia, 2601
Email: nicholas.engerer@edu.au*

Abstract: Given the rapid proliferation of data recording equipment for distributed photovoltaic (PV) arrays globally, there exists a new opportunity to use the power output from these systems for the purpose of surface solar radiation assessment. Direct measurements of the beam and diffuse irradiance represent the best methods for producing such assessments, however the equipment required for these observations are expensive and require routine maintenance, which therefore mean that the measurements are quite sparse globally. Satellite derived solar radiation estimates, meanwhile, have global coverage with increasingly fine resolution, but still require surface measurements of radiation in order to assess the performance of their solar radiation estimation algorithms (e.g. Heliostat). Therefore, it is global horizontal irradiance measurements recorded by a pyranometer, which have become the most common measurement of surface radiation. Pyranometers provide accurate surface radiation observations and are relatively inexpensive. As such, models which separate the diffuse and beam components in a global measurement have been discussed and developed vigorously in recent decades, with many modern models now accepted as the state of the art. This paper posits that the power output from PV systems is not altogether different from that recorded by pyranometers, and could be used in place of, or in supplement to, radiation observation equipment. This would greatly increase the density of the surface radiation measurement network, allowing for the many millions of PV systems reporting power output measurements globally to be applied to this purpose. PV system power output has a first order relationship with incoming solar radiation, but is confounded by additional second order interactions such as losses related to temperature, module efficiency, DC-AC conversion, soiling and shading, etc. Recently, research work by the first author has demonstrated that the individual nuances of PV systems can be accommodated through normalisation of their power output to their simulated clear sky performance. This normalised variable is termed the clear sky index for photovoltaics, K_{PV} . We use this value as the primary input to a logistical regression model in place of the traditional input, the clearness index K_t , and explore the use of additional predictor variables to optimise accuracy. PV power output was collected from 18 sites in two Australian cities (Adelaide and Melbourne) in which Bureau of Meteorology solar radiation measurement stations are deployed. This allowed us to fit and test K_t and K_{PV} based models to the observed diffuse radiation, and directly compare these approaches. Surprisingly, initial results suggest a K_{PV} based model has nearly equivalent performance to that of the traditional, pyranometer based K_t model. This paper will explore this relationship more fully, and provide the first simple model available for this purpose.

Keywords: Solar radiation, solar energy, photovoltaics, DNI, diffuse fraction

1 INTRODUCTION

The rate of solar photovoltaic (PV) installation is climbing quickly worldwide. The International Energy Agency estimates that 16% of the world's total electricity generation will be generated by solar PV by 2050, with more than half of the installations being small-scale, "rooftop" solar arrays. There is also large growth in the industry of PV data logging technologies that allow for end users to record the power output from their PV system and report it to a central server. Given the growing global reach of these PV power output databases, the opportunity to leverage this data for other scientific purposes grows along with it. Yet, there are many difficulties with the data recorded by PV systems, mainly that they each have their own unique layouts, orientations, module and inverter types and experience unique de-rating events such as soiling, shading, degradation and wiring inefficiencies. Fortunately, recent advances in distributed PV power output processing have made it possible to directly remove, or correctly account for such system level nuances. First, the work of Engerer and Mills [2014] has made it possible to remove the individual fixed system properties from the PV power output time series, through normalisation of the power output to the estimated clear sky power output from the given system. The resulting value is known as the "clear-sky index for photovoltaics", K_{PV} , and is further detailed in Section 3. Secondly, and more recently, the QCPV routine has been developed [Engerer, 2015b], which is a quality control routine for distributed PV data, which uses physical limits and across-systems testing to remove erroneous data while also providing methods for determining the tilt and azimuth of a PV system and its respective level of de-rating from soiling and shading. With these tools at one's disposal, it may become possible to use the extensive number of PV systems recording their power output data for scientific purposes.

One such scientific purpose is the estimation of the direct (E_b - also known as "beam") and diffuse (E_d) components of irradiance using PV power output data as the primary input. Knowledge of these components at the Earth's surface are relevant to a wide variety of applications, including for example, both solar thermal and photovoltaics, photosynthesis[Spitters et al., 1986] and in the design of buildings Lam and Li [1996]. Yet, direct measurements of these components are quite expensive (\$30K USD installation) and require routine maintenance. This makes alternative methods of determining these two components very attractive. To date, using PV power output to estimate E_b and E_d has not been discussed anywhere in the literature, but it has been shown that PV system power output can be used to detect biases in pyranometer measurements [Tan et al., 2014]. The subject of estimating E_b and E_d from global horizontal irradiance (E_g), however, has been discussed extensively over the past four decades [Liu and Jordan, 1960]. Models used for this purpose are still being developed and improved, with recent advances in this field including vigorous testing of the more relevant models, and progression of these models to high-resolution (e.g. 1 minute) datasets [Gueymard, 2012; Engerer, 2015a]. This manuscript suggests that the global horizontal irradiance, as measured by a pyranometer, is not fundamentally different from the "global" radiation measurement recorded by a PV system. Thus, whereas E_g has thus far been the primary input to the various models which estimate E_b and E_d (termed "decomposition" or "separation" models), it may now be possible for PV power output to be used in a similar way. This will be tested by fitting observed data to an established generalised logistic function [Engerer, 2015a] whose primary input is either E_g or PV power output and comparing the results.

2 DATA

In order to test this theory, a number of data sources are required. First, knowledge of the surface solar radiation is required through measurements of E_g , E_b and E_d . These are available at several high quality radiation sites, maintained by the Australian Bureau of Meteorology, at temporal resolutions up to 1 minute. Second, PV power output measurements are also required, but with the caveat that they must be located nearby a high quality radiation measurement site. Using the website, PVOutput.org, a public webpage where PV system owners report their power output data, 18 sites were identified within 10 km of BoM radiation monitoring sites in Adelaide (13) and Melbourne (5). Radiation observations and PV system power output measurements were extracted for these two locations for the period between January 2013 and June 2014, and then averaged to hourly values (post quality control) for the purposes of this study. Detailed information about the PV systems used is presented in Table 1.

2.1 Quality-Control

Both the radiation data and PV power output were subject to strict quality control algorithms. The radiation data was processed through the well-established QCRad routine of Long and Shi [2006], which imposes physically possible limits, performs across observations testing and creates limits based on climatological statistics to flag and remove erroneous data. In order to process the PV system data, the methods of Engerer

Table 1: PV systems used in this study. Information on the system was provided by the owner and made available on PVOutput.org. Actual tilt and orientation was derived via the QCPV algorithms [Engerer, 2015b]

Site	Rating (W)	Module Rating (W)	# of Modules	Tilt ($^{\circ}$)	Azimuth
<i>Adelaide</i>					
1	2500	250	10	21	NNW
2	3040	190	16	26	NNW
3	2200	110	20	22	N
4	6000	250	24	15	N
5	5000	250	20	18	N
6	3040	190	16	23	NNE
7	5700	190	30	31	N
8	4840	220	22	20	N
9	4230	235	18	16	NE
10	1480	185	8	26	NE
11	4560	190	14	20	NNW
12	3800	190	20	22	N
13	3500	250	14	20	N
<i>Melbourne</i>					
14	3040	190	16	22	NW
15	3885	185	21	22	NE
16	4560	190	24	24	N
17	2100	175	12	27	N
18	3885	185	21	25	W

[2015b] were utilised. These include determination of the azimuth and tilt of a given PV system based on its historical power output, the application of the QCPV routine (physical limits and across-systems testing) using the standard recommended parameters, and the determination of de-ratings arising from shading, soiling and degradation.

3 METHODS

In order to test the capability of a PV power output based model for estimating E_b and E_d , it is most suitable to also fit and test a model based on E_g observations for the same period of observation data, so that they can be directly compared. Here, we adopt the “diffuse fraction” approach, which is the classical method first suggested by the seminal work of Liu and Jordan [1960]. In this format, a regression is fit to values of the diffuse fraction, K_d , which is:

$$K_d = \frac{E_d}{E_g} \quad (1)$$

using the clearness index K_t as the primary input. The clearness index is expressed as:

$$K_t = \frac{E_g}{E_{ext_h}} \quad (2)$$

where E_{ext_h} is the horizontal component of extraterrestrial radiation.

Herein, a generalised logistic function format is used to build K_t and K_{PV} based models. This approach has been used successfully in several studies [Boland et al., 2008; Ridley et al., 2010; Engerer, 2015a] and provides a smooth, continuous function:

$$Y(x) = C + (A - C) / (1 + \beta_0 * \exp(\beta_1 + \beta_2 * x)) \quad (3)$$

where C is the value of the lower asymptote, A the value of the upper asymptote, x the independent variable and the coefficients $\beta_0, \beta_1, \beta_2$ are determined by regression. This study will use four prediction variables in the format first proposed by Engerer [2015a]: K_t , the zenith angle (θ_z), apparent solar time (AST) and the deviation of the observed K_t from the clear sky value of the clearness index K_{tc} , which is

$$\Delta K_{tc} = K_{tc} - K_t \quad (4)$$

In accordance with [Engerer, 2015a] K_{tc} is computed by the REST2 clear sky model [Gueymard, 2008].

In the case of the PV based model, the equivalent ΔK_{PVc} is used and is calculated according to

$$\Delta K_{PVc} = K_{PVc} - K_{PV} \quad (5)$$

where

$$K_{PVc} = \frac{PV_{clr}}{E_{ext} \cdot \cos(\theta_{aoi})} \quad (6)$$

where E_{ext} is the normal component of extraterrestrial radiation, θ_{aoi} is the angle of incidence and K_{PV} is the clear sky index for photovoltaics, which is calculated as:

$$K_{PV} = \frac{PV_{MEAS}}{PV_{CLR}} \quad (7)$$

where PV_{MEAS} is the measured power output from a system and PV_{CLR} is the simulated clear-sky power output.

The forms of the multi-parameter models are then:

$$K_d = C + (1 - C)/(1 + \exp(\beta_0 + \beta_1 * K_t + \beta_2 * AST + \beta_3 * \theta_z + \beta_4 * \Delta K_{tc})) \quad (8)$$

$$K_d = C + (1 - C)/(1 + \exp(\beta_0 + \beta_1 * K_{PV} + \beta_2 * AST + \beta_3 * \theta_z + \beta_4 * \Delta K_{PVc})) \quad (9)$$

and the corresponding single parameter models:

$$K_d = C + (1 - C)/(1 + \exp(\beta_0 + \beta_1 * K_t)) \quad (10)$$

$$K_d = C + (1 - C)/(1 + \exp(\beta_0 + \beta_1 * K_{PV})) \quad (11)$$

These models are then fit by non-linear least squares regression on a random selection of approximately 50% of the data. The coefficients from this fitting process are presented in Tables 2 and 3, for the single parameter and multi-parameter models, respectively. The remaining 50% of data is then used to test the model through the use of three performance measures: relative Mean Bias Error (rMBE), relative Root Mean Square Error (rRMSE) and the coefficient of determination (R^2). These error measures are computed as:

$$rMBE = \frac{1}{n\bar{O}} \sum_{i=1}^n P_i - O_i \quad (12)$$

$$rRMSE = \frac{1}{\bar{O}} \sqrt{\frac{1}{n} \sum_{i=1}^n (P_i - O_i)^2} \quad (13)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \quad (14)$$

where P_i is a given model estimate, O_i is the observed value at time i , n is the total number of observations and \bar{O} is the mean over all data points. The validation is performed on estimates and predictions of E_d .

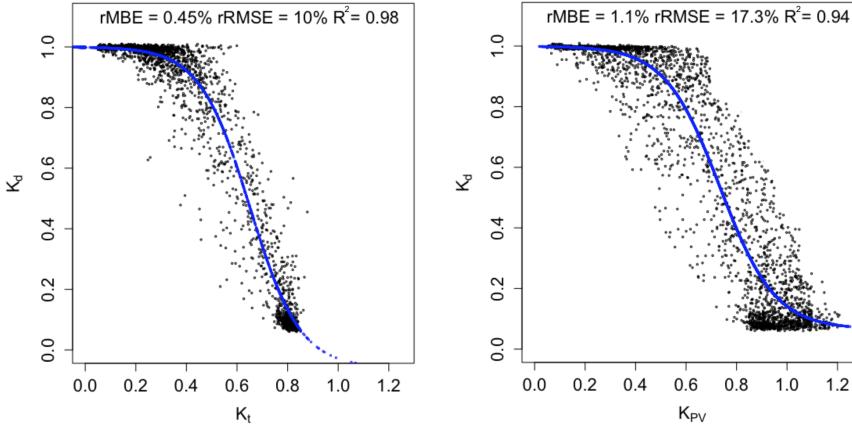


Figure 1: At left, the single parameter $K_d - K_t$ model, at right the $K_d - K_{PV}$ model, for comparison. Black points represent the observations and blue, the respective model predictions. Error measures rMBE and rRMSE are presented along with the coefficient of determination.

Table 2: Coefficients for the single parameter $K_d - K_t$ and $K_d - K_{PV}$ models.

Coefficient by model	C	β_0	β_1
K_t	-0.05879	-6.5672	10.0877
K_{PV}	0.07198	-6.8172	9.2997

4 RESULTS AND DISCUSSION

4.1 Single Parameter Models

In Figure 1, we present both the classic $K_d - K_t$ and the newly proposed $K_d - K_{PV}$ relationship. In both instances a fall in diffuse fraction is observed with increasing values of K_t and K_{PV} . This decrease in K_d follows the classically observed pattern observed in many other studies (as previously discussed). There is a concentration of high K_d values at lower values of overall clearness and low values of K_d at high values of clearness. For moderate values of clearness ($0.4 < K_t < 0.7$ and $0.4 < K_{PV} < 0.9$), observed K_d values are more diverse and less well-correlated with K_t and K_{PV} values. This is understood to be caused by the wide variety of possible cloud cover conditions that partial sky cover represents.

For the single parameter models, we observe a very similar model prediction from the two respective methods. Very little bias (rMBE) is noted, with values of 0.45% and 1.1% for the K_t and K_{PV} methods, respectively, indicating that the overall fit of the model is well-matched to the observed data. Owing to the inability of the single parameter model to estimate more than one value of K_d for a given value of K_t or K_{PV} , the models receive elevated rRMSE values of 10% and 17.3%, for the $K_d - K_t$ and $K_d - K_{PV}$ models, respectively. Observed overall correlation is relatively strong, with R scores of 0.98 and 0.94. Overall, similar performance is achieved. There is a clear benefit in using the K_t based model, which is unsurprising given the professional grade radiation equipment utilised in its computation. We note that the K_{PV} based model performs nearly as well, but suffers notably from reduced accuracy as compared to the K_t model equivalent.

The coefficients from these two models are presented in Table 2.

4.2 Four Parameter Models

It is well-established that multi-parameter diffuse fraction methods achieve better modelling results than their single parameter equivalents, and that the level of additional skill achieved is highly dependent on choosing those additional input variables wisely. Herein the four parameter model structure of Engerer [2015a], has been adopted, and the resultant $K_d - K_t$ and $K_d - K_{PV}$ models presented in Figure 2. Most immediately, it is obvious that these multi-parameter models are capable of estimating a wider number of possible K_d values for a given value of clearness, which is a particularly important step towards increasing model accuracy. To that end, the $K_d - K_t$ model rRMSE value falls from 10% to 9%, while the $K_d - K_{PV}$ model reduces its

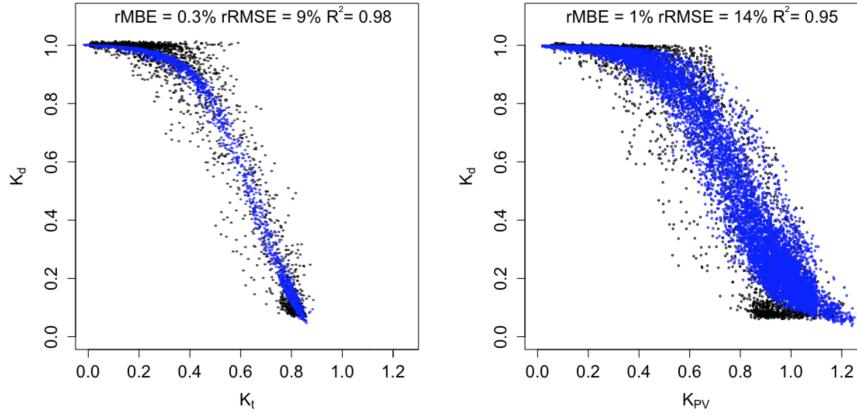


Figure 2: At left, the four parameter $K_d - K_t$ model, at right the $K_d - K_{PV}$ model. Black points represent the observed and blue, the respective model predictions. Error measures rMBE and rRMSE are presented along with the coefficient of determination, as in Figure 1.

Table 3: Coefficients for the four parameter $K_d - K_t$ and $K_d - K_{PV}$ models.

Coefficient by model	C	β_0	β_1	β_2	β_3	β_4
K_t	-0.05325	-7.7026	11.2782	0.01523	0.1358	1.3454
K_{PV}	0.01659	-9.4242	11.2719	0.2230	-3.1457	3.7497

rRMSE from 17% to 14%. The K_{PV} based model also strengthens its overall correlation (R^2 score rises to 0.95).

To further explore the modelling results, predicted versus observed K_d values are plotted in Figure 3. We note that there is a better correlation in the $K_d - K_t$ model, as indicated by the R^2 value of 0.98, but find that the $K_d - K_{PV}$ model displays remarkably similar behaviour. There is increased spread in the results in the K_{PV} based approach, but the R^2 value of 0.95 is comparable to that in the K_t based model.

The resulting coefficients from these two models are presented in Table 2. Please note that we suggest further validation of the K_{PV} model to be undertaken to determine suitability of this model for operational use, but do believe it will work well for the two regions in which the model was developed, provided the input PV data is well-behaved.

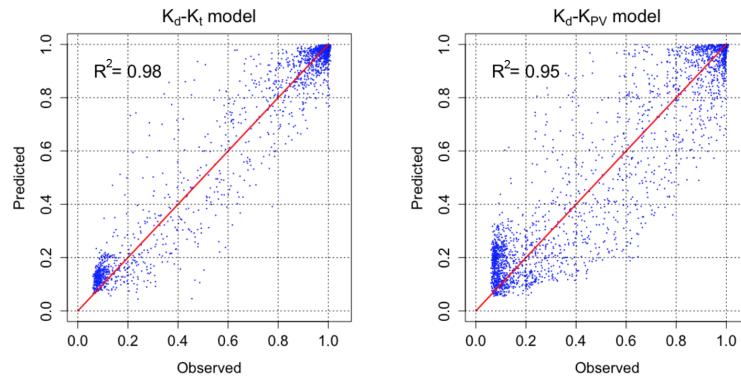


Figure 3: Predicted versus observed K_d for the four parameter $K_d - K_t$ and $K_d - K_{PV}$ models with the identity line shown in red and the overall coefficient of determination (R^2) score provided.

5 CONCLUSIONS

Given the remarkably similar performance between the $K_d - K_t$ and $K_d - K_{PV}$ models, we conclude that this simple approach to estimating the diffuse fraction through PV array power output via a generalised logistic function is a viable and promising approach. This suggests that PV systems, and their power output, may be used as method for deriving the diffuse and direct components of radiation, given that they come from quality controlled, well-curated data with minimal shading impacts. Hourly diffuse fraction estimates may be obtained in this manner with approximately a 5% increase in rRMSE values over methods using a pyranometer measurement. Overall, this model achieves the requisite levels for a “good” radiation model as laid out in Gueymard and Myers [2008] which recommends an rMBE of less than 5% and an rRMSE less than 15%.

In future, this simple approach may be improved upon through the adjusting the model fit through additional data, particularly from locations with more varied climatological conditions. There is also the potential, as suggested in Engerer [2015a] to split the model, using separate coefficients for clear and cloudy sky conditions. There is also opportunity to explore additional input variables to the PV model, such as module temperature. This is a promising first step towards using PV systems as radiation sensors, in applications such as resource assessment or other related scientific modelling.

REFERENCES

- Boland, J., B. Ridley, and B. Brown (2008). Models of Diffuse Solar Radiation. *Renewable Energy* 33, 575–584.
- Engerer, N. A. (2015a). Minute Resolution Estimates of the Diffuse Fraction of Global Irradiance for South-eastern Australia. *Solar Energy* 116, 215–237.
- Engerer, N. A. (2015b). QCPV: A Quality Control Algorithm for Distributed Photovoltaic Array Power Output. *IEEE Transactions on Sustainable Energy (Submitted)* August.
- Engerer, N. A. and F. P. Mills (2014). KPV: A Clear-Sky Index for Photovoltaics. *Solar Energy* 105, 679–693.
- Gueymard, C. (2008). REST2: High-Performance Solar Radiation Model for Cloudless-Sky Irradiance, Illuminance, and Photosynthetically Active Radiation: Validation with a Benchmark Dataset. *Solar Energy* 82, 272–285.
- Gueymard, C. (2012). Progress in Direct Irradiance Modeling and Validation. In *SOLAR 2012 Conference Proceedings*.
- Gueymard, C. and D. Myers (2008). Validation and Ranking Methodologies for Solar Radiation Models. In *Modeling Solar Radiation at the Earth's Surface*, pp. 479–509. Springer.
- Lam, J. C. and D. H. Li (1996). Correlation between Global Solar Radiation and its Direct and Diffuse Components. *Building and Environment* 31(6), 527–535.
- Liu, B. and R. Jordan (1960). The Interrelationship and Characteristic Distribution of Direct, Diffuse and Total Solar Radiation. *Solar Energy* 4, 1–19.
- Long, C. N. and Y. Shi (2006). The QCRad Value Added Product : Surface Radiation Measurement Quality Control Testing, Including Climatology Configurable Limits. Technical Report September, U.S. Department of Energy.
- Ridley, B., J. Boland, and P. Lauret (2010). Modelling of Diffuse Solar Fraction with Multiple Predictors. *Renewable Energy* 35, 478–483.
- Spitters, C., H. Toussaint, and J. Goudriaan (1986). Separating the Diffuse and Direct Component of Global Radiation and its Implications for Modeling Canopy Photosynthesis Part I: Components of Incoming Radiation. *Agricultural and Forest Meteorology* 38, 217–229.
- Tan, J., N. A. Engerer, and F. P. Mills (2014). Estimating Hourly Energy Generation of Distributed Photovoltaic Arrays : a Comparison of Two Methods. In *Proceedings of the 52nd Annual Conference, Australian Solar Energy Society (Australian Solar Council)*.