Journal of Renewable and Sustainable Energy

Satellite-augmented diffuse solar radiation separation models

Cite as: J. Renewable Sustainable Energy **11**, 023705 (2019); https://doi.org/10.1063/1.5087463
Submitted: 01 January 2019 . Accepted: 18 March 2019 . Published Online: 23 April 2019

Dazhi Yang, and John Boland

COLLECTIONS

Note: This paper is part of the Special Collection on Best Practices in Renewable Energy Resourcing and Integration.

This paper was selected as an Editor's Pick

View Online Export Citation CrossMark

ARTICLES YOU MAY BE INTERESTED IN

A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES)
Journal of Renewable and Sustainable Energy **11**, 022701 (2019); https://doi.org/10.1063/1.5087462

Can we gauge forecasts using satellite-derived solar irradiance?
Journal of Renewable and Sustainable Energy **11**, 023704 (2019); https://doi.org/10.1063/1.5087588

A hybrid MWOAL approach for fast and efficient maximum power point tracking in wind energy conversion systems
Journal of Renewable and Sustainable Energy **11**, 033302 (2019); https://doi.org/10.1063/1.5080784

AIP Author Services
English Language Editing

AIP Publishing

© 2019 Author(s).

Journal of Renewable and Sustainable Energy ARTICLE scitation.org/journal/rse

Satellite-augmented diffuse solar radiation separation models

Cite as: J. Renewable Sustainable Energy **11**, 023705 (2019); doi: 10.1063/1.5087463
Submitted: 1 January 2019 · Accepted: 18 March 2019 ·
Published Online: 23 April 2019

View Online Export Citation CrossMark

Dazhi Yang[1,a)] and John Boland[2]

AFFILIATIONS

[1]Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research (A*STAR), Singapore
[2]Centre for Industrial and Applied Mathematics, University of South Australia, Mawson Lakes, SA, Australia

Note: This paper is part of the Special Collection on Best Practices in Renewable Energy Resourcing and Integration.
[a)]**Author to whom correspondence should be addressed:** yangdazhi.nus@gmail.com. **Tel.:** +65 9159 0888.

ABSTRACT

Separation models predict diffuse horizontal irradiance from other meteorological parameters such as the global horizontal irradiance or zenith angle. From a mathematical point of view, the separation modeling problem is a regression, where the regressors are observed or computed variables and the regressand is the unobserved diffuse fraction. The most successful minute-resolution separation model prior to 2016 was proposed by Engerer, which is constructed using a trend component (cloud enhancement) and a main effect (logistic function). Subsequently, the Starke model published in 2018 further improved the Engerer model. It is herein shown that the logistic-function type of model, and many other separation models, can be transformed into a (condition-based) multiple linear regression problem. Under this transformation framework, two new models are proposed, which strictly dominate the performance of the Engerer model and the Starke model, at all 7 test sites across the continental United States, making them probably the most accurate separation models to date. The new models are also tested at 5 European sites with unseen data (i.e., not involved during model parameter fitting); their performance again dominates all benchmarking models. The new separation models leverage half-hourly satellite-derived diffuse fraction. Since satellite data are available globally, the satellite-augmented separation models have universal applicability. However, despite their good performance, empirical separation models suffer from a series of issues. Hence, models driven by atmospheric physics are the "true gems" that one should pursue.

Published under license by AIP Publishing. https://doi.org/10.1063/1.5087463

NOMENCLATURE

AM	airmass
AST	apparent solar time
D_h	diffuse horizontal irradiance, DHI
G_{cs}	clear-sky GHI
G_{ext}	extraterrestrial GHI
G_h	global horizontal irradiance, GHI
I	beam normal irradiance, BNI
k_{CSI}	clear-sky index, $k_{\mathrm{CSI}} = G_h/G_{cs}$
k_d	diffuse fraction, $k_d = D_h/G_h$
k_{de}	$k_{de} = \max(0, 1 - G_{cs}/G_h)$
k_t	clearness index, $k_t = G_h/G_{ext}$
K_T	daily average k_t
k_{tc}	clearness index of clear-sky GHI, $k_{tc} = G_{cs}/G_{ext}$
MLR	multiple linear regression
NLS	nonlinear least squares
NREL	National Renewable Energy Laboratory
NSRDB	National Solar Radiation Data Base
PSM	physical solar model
SURFRAD	Surface Radiation Budget Network
Z	solar zenith angle in degrees
Δk_{tc}	$\Delta k_{tc} = k_{tc} - k_t$
ψ	three-point moving average of k_t

I. INTRODUCTION

In its simplest form, separation modeling of solar irradiance is a class of models that estimates diffuse horizontal irradiance (DHI) from global horizontal irradiance (GHI). Subsequently, by using the closure equation, the beam normal irradiance (BNI) can be estimated. Mathematically, given the GHI at time t, $G_h(t)$, the one-parameter separation model is in the following form:

$$D_h(t) = f(G_h(t)), \tag{1}$$

where $D_h(t)$ is the DHI at time t. Since most of the separation models discussed in this paper do not consider temporal smoothing, the

Published under license by AIP Publishing

time indexing will be dropped hereafter, unless it is otherwise mentioned.

Historically, separation models were mostly developed using hourly data. Many models are in the form of piecewise polynomial regressions. Furthermore, to reduce the model dependency on the diurnal pattern of irradiance, the models are often constructed using the diffuse fraction and the clearness index. For instance, the ERBS model[1] is given by

$$k_d^{\text{ERBS}} = \begin{cases} 1.0 - 0.09k_t, & k_t \le 0.22; \\ 0.9511 - 0.1604k_t + 4.388k_t^2 - 16.638k_t^3 + 12.336k_t^4, & 0.22 < k_t \le 0.80; \\ 0.165, & k_t > 0.80, \end{cases} \tag{2}$$

where $k_d = D_h/G_h$ is the diffuse fraction and $k_t = G_h/G_{ext}$ is the clearness index—the ratio between GHI and extraterrestrial irradiance. Since the one-parameter separation modeling aims at estimating k_d from k_t, the constructed models are often visualized using a scatter plot of k_d against k_t. Figure 1 shows an example visualization with ERBS overlaid. In this paper, SMALL CAPS is used to name the separation models, following the convention of Gueymard and Ruiz-Arias[2] and Yang.[3]

It is evident that the ERBS one-parameter modeling has two deficiencies: (1) besides the piecewise modeling, there may be other mathematical functions that can provide a more flexible fitting and (2) the one-to-one correspondence between k_t and k_d cannot represent the non-injective nature of diffuse fraction, i.e., one k_t value can be mapped to multiple k_d values. To address the first issue, Boland *et al.*[4] considered the logistic function

$$k_d^{\text{BOLAND1}} = \frac{1}{1 + e^{\beta_0 + \beta_1 k_t}}, \tag{3}$$

where $\beta_0 = -4.686$ and $\beta_1 = 7.997$. This model is denoted as BOLAND1. BOLAND1 is plotted in Fig. 1. A fairly rigorous justification for using the logistic function is shown in the study by Boland *et al.*[5] For the second issue, additional meteorological parameters such as humidity, apparent solar time (AST), solar zenith angle, temperature, or airmass are used to estimate k_d based on its marginal distribution. In this way, the prediction curve is no longer a line, but a hyperplane in a space spanned by the meteorological parameters.

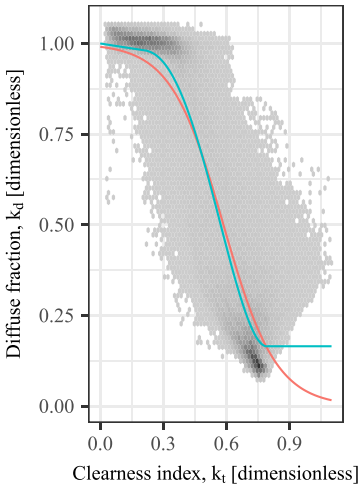

FIG. 1. Relationship between the clearness index, k_t, and the diffuse ratio, k_d. Hexagon binning is used for visualization, with darker gray indicating a higher count in the bin. ERBS is shown in turquoise, whereas BOLAND1 is shown in red.

Owing to the high flexibility in modeling, a large number of separation models have been proposed. Furthermore, with the growing needs of high-resolution environmental monitoring, the modeling regime has shifted from hourly irradiance to 1-min irradiance. Since the performance of different models varies across locations and climate conditions, choosing the "best" model for location-specific solar engineering applications has become tedious. Fortunately, Gueymard and Ruiz-Arias[2] performed an extensive validation of 140 separation models using 1-min data from 54 research-grade stations from 7 continents and recommended ENGERER2[6] as a quasi-universal model. To that end, the validation scope for future studies on separation modeling is greatly reduced.

A. ENGERER2 and the motivation for a new trend component

ENGERER2 is a five-parameter model constructed based on a logistic function and a cloud-enhancement variable. In other words, it assumes that the diffuse fraction is a linear combination of a "main effect" (logistic function) and a "trend component" (cloud-enhancement effect).[7] More specifically, ENGERER2 is given by

$$k_d^{\text{ENGERER2}} = C + \frac{1 - C}{1 + e^{\beta_0 + \beta_1 k_t + \beta_2 \text{AST} + \beta_3 Z + \beta_4 \Delta k_{tc}}} + \beta_5 k_{de}, \tag{4}$$

$$\Delta k_{tc} = k_{tc} - k_t = G_{cs}/G_{ext} - k_t, \tag{5}$$

$$k_{de} = \max(0, 1 - G_{cs}/G_h), \tag{6}$$

where G_{cs} is the clear-sky GHI and Z is the solar zenith angle in degrees. The values of the model parameters are fitted using Australia data; they are $C = 0.042336$, $\beta_0 = -3.7912$, $\beta_1 = 7.5479$, $\beta_2 = -0.010036$, $\beta_3 = 0.003148$, $\beta_4 = -5.3146$, and $\beta_5 = 1.7073$.

The use of the multipredictor main effect in ENGERER2 can be traced back to Boland *et al.*,[4] where the AST was used as an additional predictor on top of BOLAND1

$$k_d^{\text{BOLAND2}} = C + \frac{1 - C}{1 + e^{\beta_0 + \beta_1 k_t + \beta_2 \text{AST}}}, \tag{7}$$

Published under license by AIP Publishing

where $C = -0.039$, $\beta_0 = -8.769$, $\beta_1 = 7.325$, and $\beta_2 = 0.377$. As compared to BOLAND2, ENGERER2 considered two additional predictors, namely, Z and k_{tc}. The choice of predictors is more discussed in Sec. I B. As for the trend component, the effect of cloud-enhancement is accounted for using the ratio between clear-sky GHI and GHI, see Eq. (6). Equivalently, the equation can be written as $k_{de} = \max(0, 1 - k_{tc}/k_t)$. This formulation ignores the fact that cloud enhancement is not only unique to those $k_{tc} < k_t$ conditions. Hence, in theory, if a better trend model can be identified, the performance of such a separation model would be superior to ENGERER2.

The first objective of the current work is therefore to explore new approaches to detrend k_d. One way of removing the trend in k_d is by using satellite-derived diffuse fraction data, denoted by $k_d^{(s)}$. The satellite-derived diffuse fraction data usually come with a lower temporal resolution, such as 30 min or 1 h. Therefore, when $k_d^{(s)}$ is used for minute separation modeling, linear interpolation is needed. Moreover, any systematic bias in $k_d^{(s)}$ can be removed by adding a coefficient term and an intercept term.

B. STARKE and the motivation for a new predictor

As seen in Fig. 1, the logistic function describes the k_t–k_d relationship quite well. However, the predictors used in the logistic function vary across different separation models. Besides k_t, which is known to be relevant, it is in general unclear which predictors are more useful. For example, the coefficients β_2 and β_3 in ENGERER2 are small as compared to β_0, β_1, and β_4.

To understand the statistical significance of a predictor, the most well-accepted method is hypothesis testing. More specifically, the null hypothesis for a linear regression model is H_0: $\beta_i = 0$, where β_i is the coefficient of a predictor. Subsequently, the p-value for each term in the linear regression tests this hypothesis. A low p-value indicates that there is sufficient evidence to conclude that the coefficient is not equal to zero. However, if this approach is to be used in nonlinear regression—as in the case of BOLAND1, BOLAND2, and ENGERER2—the null hypothesis "for each parameter" needs to be redefined. Since the null hypothesis depends on the expectation function, the parameter's place in it, and the field of study, it is impossible to create a single test that suits all nonlinear models.

In order to circumvent the need to create specific hypothesis tests for each separation model, a variable transformation can be considered. Using the previously discussed BOLAND1 and BOLAND2 as examples, a variable k_d^* can be created that has a linear relationship with the predictors, i.e.,

$$k_d^{*\text{BOLAND1}} = \log\left(\frac{1}{k_d} - 1\right) = \beta_0 + \beta_1 k_t, \tag{8}$$

$$k_d^{*\text{BOLAND2}} = \log\left(\frac{1 - C}{k_d - C} - 1\right) = \beta_0 + \beta_1 k_t + \beta_2 \text{AST}, \tag{9}$$

or in the case of ENGERER2

$$k_d^{*\text{ENGERER2}} = \log\left(\frac{1 - C}{k_d - C - \beta_5 k_{de}} - 1\right) = \beta_0 + \beta_1 k_t + \beta_2 \text{AST} + \beta_3 Z + \beta_4 \Delta k_{tc}. \tag{10}$$

With the transformed variable, one can compute the p-values of the coefficients easily using the well-established methods developed for linear regression. This method has been documented in the studies by Boland *et al.*[4] and Ridley *et al.*[8]

Among the various logistic-function-based separation models, the most accurate model to date is perhaps STARKE,[9] which has been shown to outperform the quasi-universal ENERGER2 at multiple locations in Australia and Brazil. STARKE is given by

$$k_d^{\text{STARKE}} = \begin{cases} \dfrac{1}{1 + e^{\beta_7 + \beta_8 k_t + \beta_9 \text{AST} + \beta_{10} Z + + \beta_{11} K_T + \beta_{12} \psi + \beta_{13} G_{cs}/277.78}}, & k_{\text{CSI}} \geq 1.05 \text{ and } k_t > 0.65; \\ \dfrac{1}{1 + e^{\beta_0 + \beta_1 k_t + \beta_2 \text{AST} + \beta_3 Z + + \beta_4 K_T + \beta_5 \psi + \beta_6 G_{cs}/277.78}}, & \text{otherwise}, \end{cases} \tag{11}$$

where K_T is the daily clearness index, ψ is the moving average k_t with a window size of 3 min, and the number 277.78 converts the unit of G_{cs} from W/m^2 to MJ/h/m^2; this unit conversion was not explicitly stated in the study by Starke *et al.*[9] k_{CSI} is the clear-sky index, the ratio between G_h and G_{cs}. The two conditions $k_{\text{CSI}} \geq 1.05$ and $k_t > 0.65$ jointly model k_d under the cloud-enhancement event. STARKE is based on the BRL[12] model or otherwise known as RIDLEY2 according to Gueymard and Ruiz-Arias.[2] RIDLEY2 is a five-parameter model

$$k_d^{\text{RIDLEY2}} = \frac{1}{1 + e^{\beta_0 + \beta_1 k_t + \beta_2 \text{AST} + \beta_3 Z + + \beta_4 K_T + \beta_5 \psi}}. \tag{12}$$

The reader is referred to Rojas *et al.*[10] and Boland *et al.*[11] for the evolution of logistic-function-based models.

Although STARKE has superior performance compared to ENGERER2 and RIDLEY2, it is a seven-parameter model, which has more parameters than any of the 140 models considered by Gueymard and Ruiz-Arias.[2] Moreover, the two parameters K_T and ψ are smoothing parameters and rely on future values of k_t. Hence, if there are other real-time predictors that can provide a similar smoothing effect, the performance of such models would be comparable or even better than STARKE. On this point, using the satellite-derived diffuse fraction is thought to be appropriate.

C. Contribution of this work

Based on the above literature review on state-of-the-art separation modeling, a general modeling framework and two satellite-augmented separation models are herein proposed. The proposed framework provides high flexibility in terms of modeling and predictor selection. Furthermore, it is designed to allow hypothesis testing on the statistical significance of the chosen predictors. The two new satellite-augmented separation models have dominating performance—strictly better in terms of normalized root-mean-square error (nRMSE)—over all benchmarking models, at all 7 test sites across the continental United States. Due to the worldwide availability of satellite-derived irradiance data, the proposed models have universal applicability and are likely to be the most accurate separation models to date.[13]

Published under license by AIP Publishing

The remaining part of this paper is organized as follows. The separation modeling framework and the two specific satellite-augmented models are described in Sec. II. The experimental data used for the empirical study are described in Sec. III. Results and discussion are presented in Sec. IV. Conclusions follow at the end.

II. METHODOLOGY

A general separation model is first proposed

$$k_d = \mu(\boldsymbol{X}) + g \circ f(\boldsymbol{X}) + \varepsilon, \tag{13}$$

where $\boldsymbol{X}$ is a vector of predictors, $\mu(\cdot)$ is the trend component, $g(\cdot)$ is a (preferably) invertible function, $f(\cdot)$ is a linear or condition-based linear function, ε is the error component, and the symbol "$\circ$" denotes a composite function. Many previously proposed models can be written in the form of Eq. (13). Some examples are shown below.

A. The equivalent forms of some separation models

Erbs[1]

$$\begin{aligned} \boldsymbol{X} &= (1, k_t, k_t^2, k_t^3, k_t^4), \\ \mu(\boldsymbol{X}) &= 0, \\ f(\boldsymbol{X}) &= \begin{cases} 1.0 - 0.09k_t, & k_t \le 0.22; \\ 0.9511 - 0.1604k_t + 4.388k_t^2 - 16.638k_t^3 + 12.336k_t^4, & 0.22 < k_t \le 0.80; \\ 0.165, & k_t > 0.80, \end{cases} \\ g \circ f(\boldsymbol{X}) &= f(\boldsymbol{X}). \end{aligned} \tag{14}$$

Boland1[4]

$$\begin{aligned} \boldsymbol{X} &= (1, k_t), \\ \mu(\boldsymbol{X}) &= 0, \\ f(\boldsymbol{X}) &= \beta_0 + \beta_1 k_t \\ g \circ f(\boldsymbol{X}) &= \frac{1}{1 + e^{f(\boldsymbol{X})}}. \end{aligned} \tag{15}$$

Boland2[4]

$$\begin{aligned} \boldsymbol{X} &= (1, k_t, \mathrm{AST}), \\ \mu(\boldsymbol{X}) &= C, \\ f(\boldsymbol{X}) &= \beta_0 + \beta_1 k_t + \beta_2 \mathrm{AST}, \\ g \circ f(\boldsymbol{X}) &= \frac{1 - C}{1 + e^{f(\boldsymbol{X})}}. \end{aligned} \tag{16}$$

Ridley2[8]

$$\begin{aligned} \boldsymbol{X} &= (1, k_t, \mathrm{AST}, \alpha, K_T, \psi), \\ \mu(\boldsymbol{X}) &= 0, \\ f(\boldsymbol{X}) &= \beta_0 + \beta_1 k_t + \beta_2 \mathrm{AST} + \beta_3 \alpha + \beta_4 K_T + \beta_5 \psi, \\ g \circ f(\boldsymbol{X}) &= \frac{1}{1 + e^{f(\boldsymbol{X})}}. \end{aligned} \tag{17}$$

RuizArias[14]

$$\begin{aligned} \boldsymbol{X} &= (1, k_t, k_t^2, \mathrm{AM}, \mathrm{AM}^2), \\ \mu(\boldsymbol{X}) &= \beta_0, \\ f(\boldsymbol{X}) &= \beta_2 + \beta_3 k_t + \beta_4 k_t^2 + \beta_5 \mathrm{AM} + \beta_6 \mathrm{AM}^2, \\ g \circ f(\boldsymbol{X}) &= -\beta_1 e^{e^{f(\boldsymbol{X})}}. \end{aligned} \tag{18}$$

where AM is the airmass and the fitted parameters (for the globe) are $\beta_0 = 0.944$, $\beta_1 = 1.538$, $\beta_2 = 2.808$, $\beta_3 = -5.759$, $\beta_4 = 2.276$, $\beta_5 = -0.125$, and $\beta_6 = 0.013$.

Engerer2[6]

$$\begin{aligned} \boldsymbol{X} &= (1, k_t, \mathrm{AST}, Z, \Delta k_{tc}, k_{de}), \\ \mu(\boldsymbol{X}) &= C + \beta_5 k_{de}, \\ f(\boldsymbol{X}) &= \beta_0 + \beta_1 k_t + \beta_2 \mathrm{AST} + \beta_3 Z + \beta_4 \Delta k_{tc}, \\ g \circ f(\boldsymbol{X}) &= \frac{1 - C}{1 + e^{f(\boldsymbol{X})}}. \end{aligned} \tag{19}$$

Starke[9]

$$\begin{aligned} \boldsymbol{X} &= (1, k_t, \mathrm{AST}, \alpha, K_T, \psi, G_{cs}), \quad \mu(\boldsymbol{X}) = 0, \\ f(\boldsymbol{X}) &= \begin{cases} \beta_7 + \beta_8 k_t + \beta_9 \mathrm{AST} + \beta_{10}\alpha + \beta_{11} K_T + \beta_{12}\psi + \beta_{13} G_{cs}/277.78, & G_h/G_{cs} \ge 1.05 \text{ and } k_t > 0.65; \\ \beta_0 + \beta_1 k_t + \beta_2 \mathrm{AST} + \beta_3 \alpha + \beta_4 K_T + \beta_5 \psi + \beta_6 G_{cs}/277.78, & \text{otherwise}, \end{cases} \\ g \circ f(\boldsymbol{X}) &= \frac{1}{1 + e^{f(\boldsymbol{X})}}. \end{aligned} \tag{20}$$

Published under license by AIP Publishing

B. Some practical considerations on designing separation models

It is noted that most predictors used in separation models can be calculated on-the-fly, and the only measured quantity is GHI. On this point, all the above models can be used in real-time applications, except for STARKE and RIDLEY2, because one can only compute K_T and ψ when the day has passed.[15] Hence, during the design of separation models, one should choose an $\boldsymbol{X}$ that is easily accessible. Specific predictors that require additional measurements or future values of GHI should be avoided since they limit the uptake of the model. On this point, one should be aware that the high-performance clear-sky models, such as the REST2 model,[16] usually require many parameters that are not easily accessible. Hence, when a simpler clear-sky model is used, the performance of some of the separation models may suffer. In this article, the Ineichen–Perez clear-sky model[32] is used.

In all the above examples, $\mu(\boldsymbol{X})$ is a constant, except for ENGERER2. This is because D_h depends on a large number of factors that are governed by the intricate atmospheric scattering mechanisms, and isolating one or a few factors for trend modeling is difficult. However, as satellite-to-irradiance models and numerical weather prediction (NWP) output become increasingly accessible around the globe, they can be utilized in separation modeling. Although the temporal resolution of satellite-derived irradiance and NWP output is low, it is thought to be sufficient in providing a trend for k_d.

At the moment, there is not much research on discovering the physical laws for the variation in k_d. Perhaps the only physical separation model is HOLLANDS.[17,18] The reader is referred to the study by Gueymard and Ruiz-Arias[2] for more information regarding the lack of physical modeling. Whereas physical modeling of k_d is no doubt the preferred way, current research practice mostly focuses on empirical modeling. In this regard, physically based satellite-derived irradiance or NWP output is a step forward in including physics aspects in separation modeling.

C. Two new separation models

Based on the above discussion and the model construction of ENGERER2, two new separation models are proposed. The first model includes the new predictor, $k_d^{(s)}$, the satellite-derived diffuse fraction, in the trend component

$$\begin{aligned} \boldsymbol{X} &= \left(1, k_t, \mathrm{AST}, Z, \Delta k_{tc}, k_{de}, k_d^{(s)}\right), \\ \mu(\boldsymbol{X}) &= C + \beta_5 k_{de} + \beta_6 k_d^{(s)}, \\ f(\boldsymbol{X}) &= \beta_0 + \beta_1 k_t + \beta_2 \mathrm{AST} + \beta_3 Z + \beta_4 \Delta k_{tc}, \\ g \circ f(\boldsymbol{X}) &= \frac{L}{1 + e^{f(\boldsymbol{X})}}. \end{aligned} \tag{21}$$

The inclusion of the new predictor makes $\mu(\boldsymbol{X})$ a trend surface. Whereas k_{de} accounts for high-frequency fluctuations in D_h, $k_d^{(s)}$ describes the low-frequency variability. However, the explicit removal of the trend surface distorts the range[19] of the logistic-function part, see the left panel of Fig. 2, which plots the detrended k_d using the data shown in Fig. 1. A remedy to the problem is introducing another regression coefficient L, which models the maximum value of the new logistic function. This model is called YANG1, and it can be written in an equivalent form

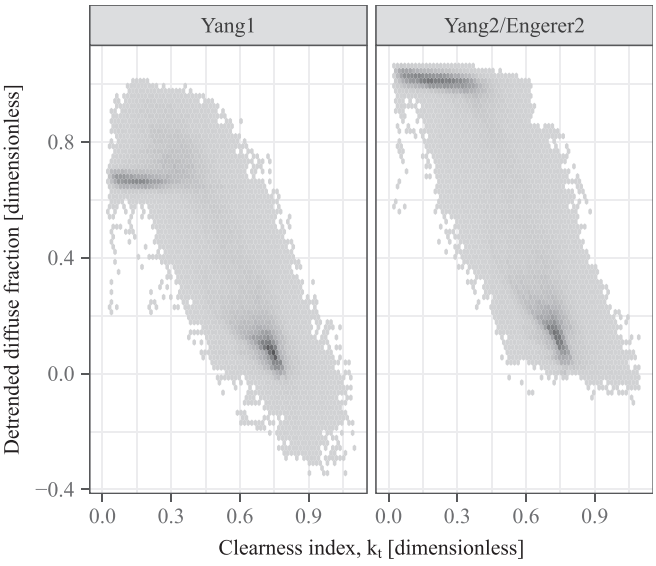

FIG. 2. The detrended diffuse fraction via YANG1 and YANG2 (the same as the detrending used in ENGERER2). One can compare these plots with Fig. 1 to see how the scatter evolves after detrending.

$$k_d^{\mathrm{YANG1}} = C + \frac{L}{1 + e^{\beta_0 + \beta_1 k_t + \beta_2 \mathrm{AST} + \beta_3 Z + \beta_4 \Delta k_{tc}}} + \beta_5 k_{de} + \beta_6 k_d^{(s)}. \tag{22}$$

Unlike YANG1, the second model adds $k_d^{(s)}$ to the main effect

$$\begin{aligned} \boldsymbol{X} &= \left(1, k_t, \mathrm{AST}, Z, \Delta k_{tc}, k_{de}, k_d^{(s)}\right), \\ \mu(\boldsymbol{X}) &= C + \beta_5 k_{de}, \\ f(\boldsymbol{X}) &= \beta_0 + \beta_1 k_t + \beta_2 \mathrm{AST} + \beta_3 Z + \beta_4 \Delta k_{tc} + \beta_6 k_d^{(s)}, \\ g \circ f(\boldsymbol{X}) &= \frac{1 - C}{1 + e^{f(\boldsymbol{X})}} \end{aligned} \tag{23}$$

or equivalently

$$k_d^{\mathrm{YANG2}} = C + \frac{1 - C}{1 + e^{\beta_0 + \beta_1 k_t + \beta_2 \mathrm{AST} + \beta_3 Z + \beta_4 \Delta k_{tc} + \beta_6 k_d^{(s)}}} + \beta_5 k_{de}. \tag{24}$$

The assumption in YANG2 that the cloud-enhancement events only affect DHI when GHI overshoots its clear-sky expectation is the same as ENGERER2. Once these "spikes" are removed from k_d, the smoothed version of k_d is modeled using a logistic function, see the right panel of Fig. 2. In this case, $k_d^{(s)}$ is used as an additional predictor for the logistic part of modeling, which has a similar functionality to parameters K_T and ψ in STARKE.

D. Parameter estimation

YANG1 has a total of 9 coefficients to be fitted, whereas YANG2 has 8. Since both the models are everywhere differentiable, the coefficients can be estimated via a Gauss–Newton-type nonlinear least squares (NLS) routine (see Sec. 2.2.1 in the study by Bates and Watts[20]). The particular choice of optimization routine is thought to be not so important; the default NLS function in the R package "stats" is herein used.

Published under license by AIP Publishing

TABLE I. Metadata of the SURFRAD network.

Abbrv.	Station	Latitude	Longitude	Elevation (m)	Time zone	# validation data points
BON	Bondville, Illinois	40.05192°N	88.37309°W	230	Central	198 669
TBL	Table Mountain, Boulder, Colorado	40.12498°N	105.23680°W	1689	Mountain	211 237
DRA	Desert Rock, Nevada	36.62373°N	116.01947°W	1007	Pacific	209 306
FPK	Fort Peck, Montana	48.30783°N	105.10170°W	634	Mountain	193 878
GWN	Goodwin Creek, Mississippi	34.2547°N	89.8729°W	98	Central	205 911
SXF	Sioux Falls, South Dakota	43.73403°N	96.62328°W	473	Central	204 851
PSU	Pennsylvania State University, Pennsylvania	40.72012°N	77.93085°W	376	Eastern	206 772

III. DATA

A total of four datasets are used in this article. The first dataset consists of research-grade ground-based measurements. The 7-station surface radiation (SURFRAD) budget network was established in 1993 and has been providing high-resolution radiation data since 1995. The metadata of SURFRAD are shown in Table I. Besides GHI, DHI, and BNI, SURFRAD also measures other meteorological parameters. However, in this work, only the three basic irradiance components are used. More specifically, 1-min GHI, DHI, and BNI from all 7 stations over a two-year period from 2015 to 2016 are selected for the case study. The data from the year 2015 are used for fitting, whereas the 2016 data are used for validation. The quality control of the SURFRAD data follows that in the studies by Yang *et al.*[21] and Yang,[22] and hence, it is not reiterated here. Furthermore, the R code to access SURFRAD and the quality-control routine are provided in the study by Yang.[22] The total number of data points used for the validation for each station is listed in Table I.

The second dataset is the National Solar Radiation Data Base (NSRDB) created and maintained by the National Renewable Energy Laboratory (NREL). NSRDB is modeled through physical solar model (PSM) version 3. PSM is a sophisticated satellite-derived product that leverages on five types of meteorological properties, namely, cloud properties, atmospheric profiles, aerosol properties, surface albedo, and snow albedo.[23] These properties are combined through a fast all-sky radiation model, called "FARMS."[24] The spatial resolution of the NSRDB is 0.04° in both the latitude and the longitude on a regular grid, covering most of America, whereas the temporal resolution is 30 min, spanning 1998–2016. NSRDB has been extensively studied in the literature for various solar engineering applications (e.g., Refs. 25 and 26). Its uncertainty has also been validated.[27] NSRDB is publicly available. Moreover, the R code for accessing the NREL PSM application programming interface is provided in the study by Yang,[22] whereas a python version is available at the NREL website. Since the NSRDB diffuse fraction—the ratio between NSRDB DHI and NSRDB GHI—will be used as a predictor in this work, i.e., $k_d^{(s)}$, a subset of NSRDB that is collocated and temporally aligned with the SURFRAD dataset is selected. Since $k_d^{(s)}$ computed from NSRDB is half-hourly, to align with the 1-min SURFRAD data, $k_d^{(s)}$ is linear interpolated at the 1-min timestamps.

Whereas the above two datasets will be used to fit and validate the new models, the remaining two datasets will be used to test the performance of these new models at unseen sites. Since the global parameters of Engerer2 and Starke were fitted using Australia and Brazil data, respectively, and Yang1 and Yang2 will be fitted using data from the continental US, European sites are chosen for a fair comparison. The third dataset comes from the Baseline Solar Radiation Network (BSRN).[28] BSRN is an initiative led by the World Meteorological Organization, which performs 1-min or 3-min research-grade solar radiation measurements at 59 stations worldwide. In this work, data from 5 European sites over the year 2016 are selected. Similar to Table I, the metadata for these sites are summarized in Table II. The quality-control routine used for SURFRAD is also used for BSRN.

The last dataset is another satellite-derived irradiance database, which will be used to provide the $k_d^{(s)}$ estimates for the European sites. More specifically, the Copernicus Atmosphere Monitoring Service radiation service (CAMS-RAD) developed by the European Centre for Medium-Range Weather Forecasts is considered. CAMS-RAD is an operational product released after the pre-operational versions called Monitoring Atmospheric Composition and Climate (MACC) and MACC-II/-III. The CAMS-RAD irradiance is produced using images from the Meteosat Second Generation satellites and the Heliosat-4 method, which uses a physical cloud retrieval process and a fast parameterization of radiative transfer.[29] The spatial extent of CAMS-RAD is −66° to 66° in both latitudes and longitudes, whereas its temporal coverage is from 2004–02-01 up to 2 days ago. The subset of

TABLE II. Metadata of the (partial) BSRN network.

Abbrv.	Station	Latitude	Longitude	Elevation	Time zone	# validation data points
CAB	Cabauw, The Netherlands	51.9711°N	4.9267°E	0	Europe/Amsterdam	205 985
CAR	Carpentras, France	44.0830°N	5.0590°E	100	Europe/Paris	215 660
CNR	Cener, Spain	42.8160°N	1.6010°W	471	Europe/Madrid	218 721
PAL	Palaiseau, France	48.7130°N	2.2080°E	156	Europe/Paris	207 026
PAY	Payerne, Switzerland	46.8150°N	6.9440°E	491	Europe/Zurich	209 631

Published under license by AIP Publishing

TABLE III. Model coefficients of YANG1 and YANG2 fitting using 1 year (2015) of 1-min data collected at 7 SURFRAD stations.

Model	C	β_0	β_1	β_2	β_3	β_4	β_5	β_6	L
YANG1	0.0369	−3.4986	7.9735	−0.0030	0.0031	−7.6836	1.0179	0.3505	0.6768
YANG2	0.0361	−0.5744	4.3184	−0.0011	0.0004	−4.7952	1.4414	−2.8396	…

CAMS-RAD that is collocated and temporally aligned with the BSRN dataset is selected.

IV. RESULTS AND DISCUSSION

A total of 10 separation models are considered in the empirical part of this paper. ERBS is selected to represent the class of "baseline" one-parameter models. Since ENGERER2 is the best performing model prior to 2016, it is also selected for benchmarking. It is well known that the performance of ENGERER2 can be improved by including locally fitted coefficients; in addition to the original set of "global" coefficients (fitted using data from three Australian sites), a set of re-estimated coefficients—using 2015 SURFRAD data—for each station is also tested. These two versions of ENGERER2 are denoted as ENGERER2 (Australia) and ENGERER2 (local), respectively. The next two separation models are two versions of STARKE, with parameters fitted using Brazil[9] and local data. They are denoted as STARKE (Brazil) and STARKE (local), respectively. The sixth model directly compares the interpolated NSRDB DHI with the measured DHI. Although it is not a separation model, it is necessary to understand the associated uncertainty of NSRDB data.[30] The remaining four models are the proposed pair of models, namely, YANG1 and YANG2. To make fair comparisons, both globally and locally fitted model coefficients are tested. The global fitting uses the 2015 data from all 7 SURFRAD stations, whereas the local fitting only uses data from the focal station. To facilitate future uptake of the proposed models, the global coefficients of YANG1 and YANG2 are given in Table III, whereas the local coefficients are omitted.

Two well-accepted error metrics, namely, normalized mean bias error (nMBE) and nRMSE, are used for evaluation. It is noted that although all separation models estimate k_d, the errors are computed based on $\hat{D}_h$, i.e., $\hat{k}_d * G_h$. Both metrics are expressed in percentage

$$\mathrm{nMBE} = \frac{\sum_{t=1}^{n}\left[\hat{D}_h(t) - D_h(t)\right]}{\sum_{t=1}^{n} D_h(t)} \times 100 \tag{25}$$

and

$$\mathrm{nRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left[\hat{D}_h(t) - D_h(t)\right]^2}}{\frac{1}{n}\sum_{t=1}^{n} D_h(t)} \times 100, \tag{26}$$

TABLE IV. The nRMSE and nMBE in percentage of 9 separation models validated at 7 SURFRAD stations over a period of 1 year (2016). The errors are computed between the predicted and measured 1-min diffuse horizontal irradiance. In addition, the satellite-derived diffuse fractions are interpolated at 1-min intervals to give perspective on the results of using these data directly without separation modeling. Boldface denotes the best-performing model in a row.

Station	ERBS	ENGERER2 (-)	ENGERER2 (local)	STARKE (Brazil)	STARKE (local)	Satellite interp.	YANG1 (US)	YANG1 (local)	YANG2 (US)	YANG2 (local)
nRMSE (%)										
BON	38.91	34.01	33.47	29.39	30.04	49.62	28.43	27.48	27.50	**26.50**
TBL	46.38	42.94	40.77	40.92	38.05	84.13	38.07	35.34	38.66	**34.45**
DRA	51.02	46.41	40.26	36.98	34.48	55.63	33.62	33.16	**32.07**	32.09
FPK	39.99	36.97	36.18	32.36	31.74	53.71	29.35	30.09	**28.55**	29.19
GWN	40.39	35.73	34.15	29.29	29.43	58.21	28.88	28.93	**28.39**	28.92
PSU	37.93	33.26	33.03	28.19	28.18	59.49	28.57	28.26	27.67	**27.37**
SXF	40.60	36.39	34.83	32.57	32.56	48.73	27.37	27.17	**26.06**	26.28
Overall	42.17	37.96	36.10	32.81	32.07	58.50	30.61	30.06	29.84	**29.26**
nMBE (%)										
BON	−4.43	3.88	3.57	−3.62	3.86	−10.49	−2.98	2.23	−3.18	**0.87**
TBL	−1.47	8.67	3.51	−2.61	**−0.47**	14.55	9.02	2.63	8.49	2.58
DRA	16.87	21.54	−1.32	−1.80	−3.96	**−0.24**	6.39	−3.55	4.01	−4.40
FPK	1.57	9.47	2.76	2.02	5.09	−9.18	1.67	**−0.70**	0.78	−1.14
GWN	2.32	9.73	3.99	3.05	4.44	−4.92	1.64	4.21	**1.33**	3.72
PSU	−4.07	4.55	4.80	**−0.63**	2.75	−3.06	1.54	4.76	1.38	4.59
SXF	**−0.12**	8.15	2.55	0.40	3.26	−6.74	0.68	3.60	0.47	3.06
Overall	1.52	9.43	2.84	**−0.46**	2.14	−2.87	2.57	1.88	1.90	1.33

Published under license by AIP Publishing

where n for each station, the number of time stamps used in validation, is listed in the last column of Table I.

The validation results for DHI prediction are shown in Table IV. By examining the results, one can quickly rule out the one-parameter ERBS and the satellite-interpolated DHI, due to their high errors. The one-to-one k_t–k_d relationship described by ERBS is insufficient to model the non-injective diffuse irradiance. On the other hand, the satellite DHI downscaling—converting the 30-min DHI to 1-min DHI—could not capture the high-frequency variation in the diffuse irradiance.

For the other more elaborated models, ENGERER2 (Australia) performs the worst. It is worth noting that this model was identified to be quasi-universal by Gueymard and Ruiz-Arias.[2] Although slight improvements can be made by using locally fitted parameters, i.e., ENGERER2 (local), its overall performance is still worse than the STARKE and YANG families of models. This observation aligns with that made by Starke *et al.*,[9] and one can thus confirm the superiority of STARKE over ENGERER2, at least for the locations in the continental United States.

Both YANG1 and YANG2 have exceptional performance. Owing to the physical solar model used in the NSRDB data, the (low-frequency) variability of DHI is better explained. This can be seen from Figs. 3 and 4, where YANG1 and YANG2 show best coverage of the gray scatter among all the models.

Besides DHI, the performance of various separation models is also evaluated in terms of BNI. BNI is denoted with I, which is the irradiance received by a plane normal to the sun ray (not to be confused with the beam irradiance received by a horizontal plane, or BHI). Predicted BNI is given by

$$\hat{I} = (G_h - \hat{D}_h)/\cos Z. \tag{27}$$

The nMBE and nRMSE are calculated using Eqs. (25) and (26), replacing all D_h with I. The results are tabulated in Table V. All previous discussions on DHI performance apply. Hence, the superiority of YANG1 and YANG2 is shown.

A. Discussion: Performance at unseen sites

Separation modeling is only useful when on-site diffuse measurements are not available or incorrect. Hence, the primary goal for separation modeling is to predict DHI accurately at any arbitrary locations, i.e., at unseen sites. In this subsection, the performance of ENGERER2, STARKE, YANG1, and YANG2 is evaluated with 1-min data from the 5 European sites listed in Table II. Since none of these models used data from Europe for parameter fitting, the comparison is fair. The procedure used for this case study is identical to the previous one.

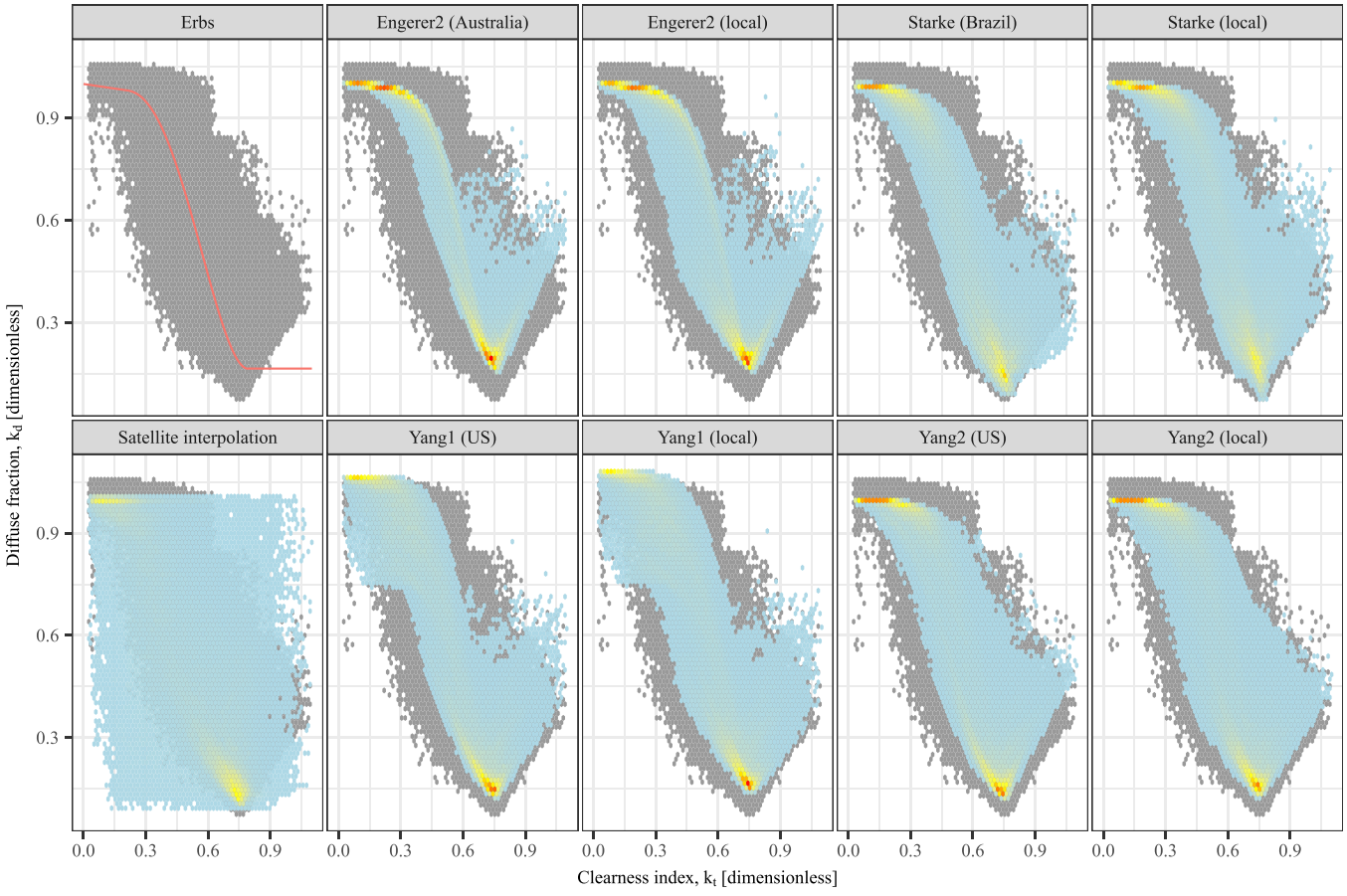

FIG. 3. Diffuse fraction data for the BON station (40.05192°N, 88.37309°W), overlaid with the results of various separation models. Hexagon binning is used for visualization, and the total number of data points in each plot is listed in Table I.

Published under license by AIP Publishing

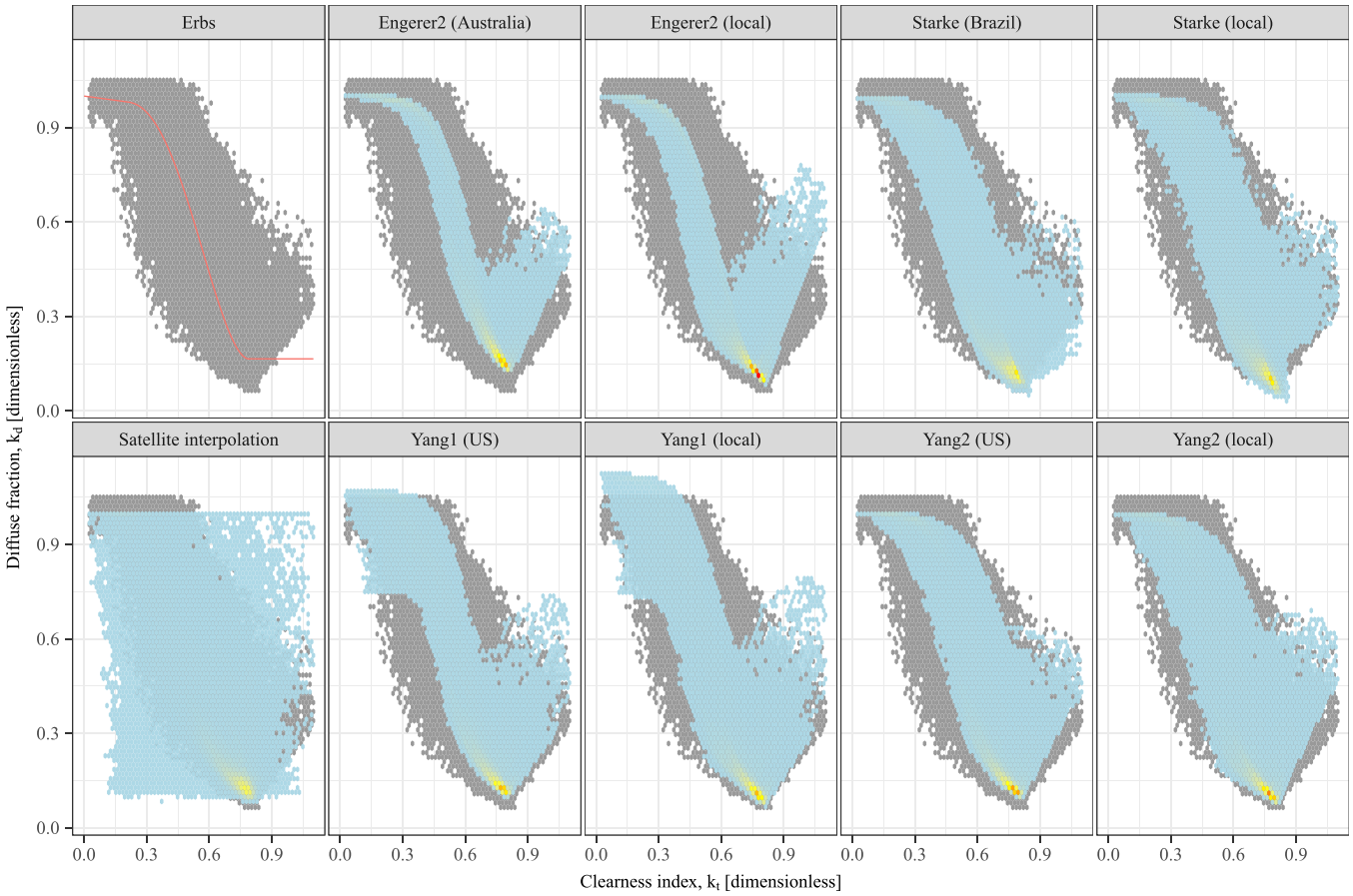

FIG. 4. The same as Fig. 3, but for DRA (36.62373°N, 116.01947°W), where the clear-sky condition predominates.

The error metrics for DHI and BNI are shown in Tables VI and VII, respectively. In addition, the k_t–k_d plots for two stations, namely, CAR and PAL, are shown in Fig. 5. It is observed that the new models outperform Engerer2 and Starke at all sites in terms of nRMSE, except for BNI prediction at PAY, where Starke is 0.6% better than Yang1. Hence, the superiority of the new models can be confirmed. To further enhance the performance of Yang1 and Yang2, more data—either more years or more locations—could be used in fitting. However, such a discussion will be postponed to a future paper.

B. Discussion: Choice of predictors

Both Yang1 and Yang2 are six-parameter models, depending on k_t, AST, Z, Δk_{tc}, k_{de}, and $k_d^{(s)}$. When new parameters such as K_T, ψ, or even temperature and humidity, are added into the model, the performance may or may not be better. To that end, solar engineers usually perturb the choice of model parameters and decide on the best choice based on the predictive performance of a model. The drawback of this approach is that the choices of parameter combination will never be exhaustive, and hence, pursuing the best empirical model is a never-ending game. Furthermore, the best model at one location may not be the best model elsewhere. There is in fact very little meaning to search for additional predictors that can only lead to a marginal improvement in performance. In other words, future studies should focus on understanding the atmospheric physics governing the diffuse radiation, rather than modifying the current regression-based model or specifying alternative predictor sets.

C. Discussion: The *p*-value problem

As mentioned earlier, with transformation of variables, one can rewrite many separation models into multiple linear regressions. Subsequently, hypothesis testing, such as the *t*-test or the partial *F*-test, can be used to "conclude" whether a predictor is significant or whether adding a predictor is significant. However, in most engineering applications, the effect of the large sample size is ignored during hypothesis testing. This type of analysis often leads to dangerous pitfall, known as "the *p*-value problem," which has long been known to statisticians.

Today, the coefficients of the separation models are fitting using several hundred thousands of samples or even millions. Such a large sample size will result in very small *p*-values, simply because the test statistics is often a function of sample size. Take the *t*-test for example, when the null hypothesis is H_0: $\beta = 0$, the absolute value of the observed *t*-statistics, $|T|$, is $\hat{\beta}/\hat{\sigma}_{\hat{\beta}}$. In a single independent variable case

Published under license by AIP Publishing

TABLE V. The same as Table IV, but the errors are computed between the predicted and measured 1-min beam normal irradiance.

Station	Erbs	Engerer2 (Australia)	Engerer2 (local)	Starke (Brazil)	Starke (local)	Satellite interp.	Yang1 (US)	Yang1 (local)	Yang2 (US)	Yang2 (local)
nRMSE (%)										
BON	30.56	25.42	24.09	22.29	22.38	32.93	20.38	20.07	19.90	**19.41**
TBL	24.54	20.97	19.98	19.76	19.02	36.97	18.04	**17.47**	18.51	17.69
DRA	17.31	14.99	12.68	11.27	11.37	16.21	10.96	10.72	10.64	**10.50**
FPK	30.22	25.77	24.88	22.71	22.98	37.29	**21.53**	21.72	21.58	21.96
GWN	25.77	21.48	19.48	17.29	17.02	29.68	16.69	16.58	**16.36**	16.43
PSU	33.64	27.49	26.38	24.35	23.77	43.36	22.81	22.65	22.30	**21.85**
SXF	30.11	24.98	22.85	23.29	22.94	31.29	18.77	17.74	18.17	**17.66**
Overall	27.45	23.01	21.48	20.14	19.92	32.53	18.45	18.14	18.21	**17.93**
nMBE (%)										
BON	**−0.12**	−4.58	−3.67	1.56	−4.03	7.52	1.50	−2.10	1.05	−1.94
TBL	**−0.02**	−4.51	−2.02	2.38	0.10	−5.55	−3.59	−1.00	−3.79	−1.44
DRA	−5.75	−6.63	**0.54**	1.21	1.28	0.58	−1.78	1.33	−1.20	1.50
FPK	−0.77	−5.00	**−0.56**	0.84	−2.09	9.62	1.22	3.04	1.20	2.63
GWN	−3.05	−6.02	−1.84	−1.31	−1.95	5.07	**0.01**	−1.34	−0.20	−1.54
PSU	3.51	−2.26	−1.63	2.70	**−0.21**	7.41	2.05	−0.97	1.33	−1.33
SXF	−0.42	−4.85	−1.28	1.26	−1.78	5.31	0.65	−1.08	**0.14**	−1.50
Overall	−0.95	−4.84	−1.49	1.24	−1.24	4.28	**0.01**	−0.30	−0.21	−0.52

$$\hat{\sigma}_{\hat{\beta}} = \frac{\sqrt{\text{MSE}}}{s_x\sqrt{n-1}}, \tag{28}$$

where MSE is the estimate of the error variance and s_x is the standard deviation of the independent variable x. One can see that when $n \to \infty$, $\hat{\sigma}_{\hat{\beta}} \to 0$ and $|T| \to \infty$. As a result, the cumulative t-distribution until $|T|$, i.e., $\Phi(df, |T|)$, approaches 1, and eventually, the p-value, namely, $2(1 - \Phi(df, |T|))$ goes to 0. The reader is referred to the study by Lin *et al.*[31] for a detailed discussion on the p-value problem.

Based on the above discussion, it should be clear that a small p-value in the large-sample problem is misleading. In other words, even an extremely small effect can become statistically significant. In this regard, despite the fact that Ridley *et al.*[8] and Boland *et al.*[4] tested the effect of predictors, such hypothesis testing is inconclusive. That said, in terms of predictive accuracy, there is no doubt that Boland2 is

TABLE VI. The nRMSE and nMBE in percentage of 4 separation models validated at 5 European BSRN stations over a period of 1 year (2016). The errors are computed between the predicted and measured 1-min diffuse horizontal irradiance. In addition, the satellite-derived diffuse fractions are interpolated at 1-min intervals to give perspective on the results of using these data directly without separation modeling.

Station	Engerer2 (Australia)	Starke (Brazil)	Satellite interp.	Yang1 (US)	Yang2 (US)
nRMSE (%)					
CAB	31.74	26.35	41.70	26.39	**24.98**
CAR	51.06	42.90	46.61	33.83	**31.72**
CNR	37.20	31.37	50.04	28.43	**26.88**
PAL	32.85	27.01	44.43	25.97	**24.47**
PAY	35.97	29.23	45.49	28.88	**27.99**
Overall	37.76	31.37	45.66	28.70	**27.21**
nMBE (%)					
CAB	**−0.36**	−2.74	−12.67	−4.53	−3.66
CAR	22.91	13.18	**0.03**	9.16	8.45
CNR	7.98	2.55	−6.64	**1.08**	1.40
PAL	1.82	**−0.18**	−8.75	−1.30	−1.02
PAY	5.45	**−0.70**	−12.33	−2.06	−1.06
Overall	7.56	2.42	−8.07	**0.47**	0.82

Published under license by AIP Publishing

TABLE VII. The same as Table VI, but the errors are computed between the predicted and measured 1-min beam normal irradiance.

Station	Engerer2 (Australia)	Starke (Brazil)	Satellite interp.	Yang1 (US)	Yang2 (US)
nRMSE (%)					
CAB	34.23	29.44	44.36	28.64	**27.34**
CAR	22.03	17.94	20.78	14.93	**14.28**
CNR	25.46	21.45	33.58	19.61	**19.11**
PAL	33.30	27.43	42.82	26.40	**25.32**
PAY	33.40	**25.73**	39.97	26.37	26.68
Overall	29.68	24.40	36.30	23.19	**22.55**
nMBE (%)					
CAB	**0.99**	4.96	15.36	6.84	4.76
CAR	−10.56	−5.06	**0.00**	−3.99	−4.16
CNR	−5.30	**−0.01**	5.20	0.08	−0.99
PAL	−1.74	1.69	10.07	2.87	**1.33**
PAY	−5.77	1.79	12.21	2.33	**0.13**
Overall	−4.48	0.67	8.57	1.62	**0.21**

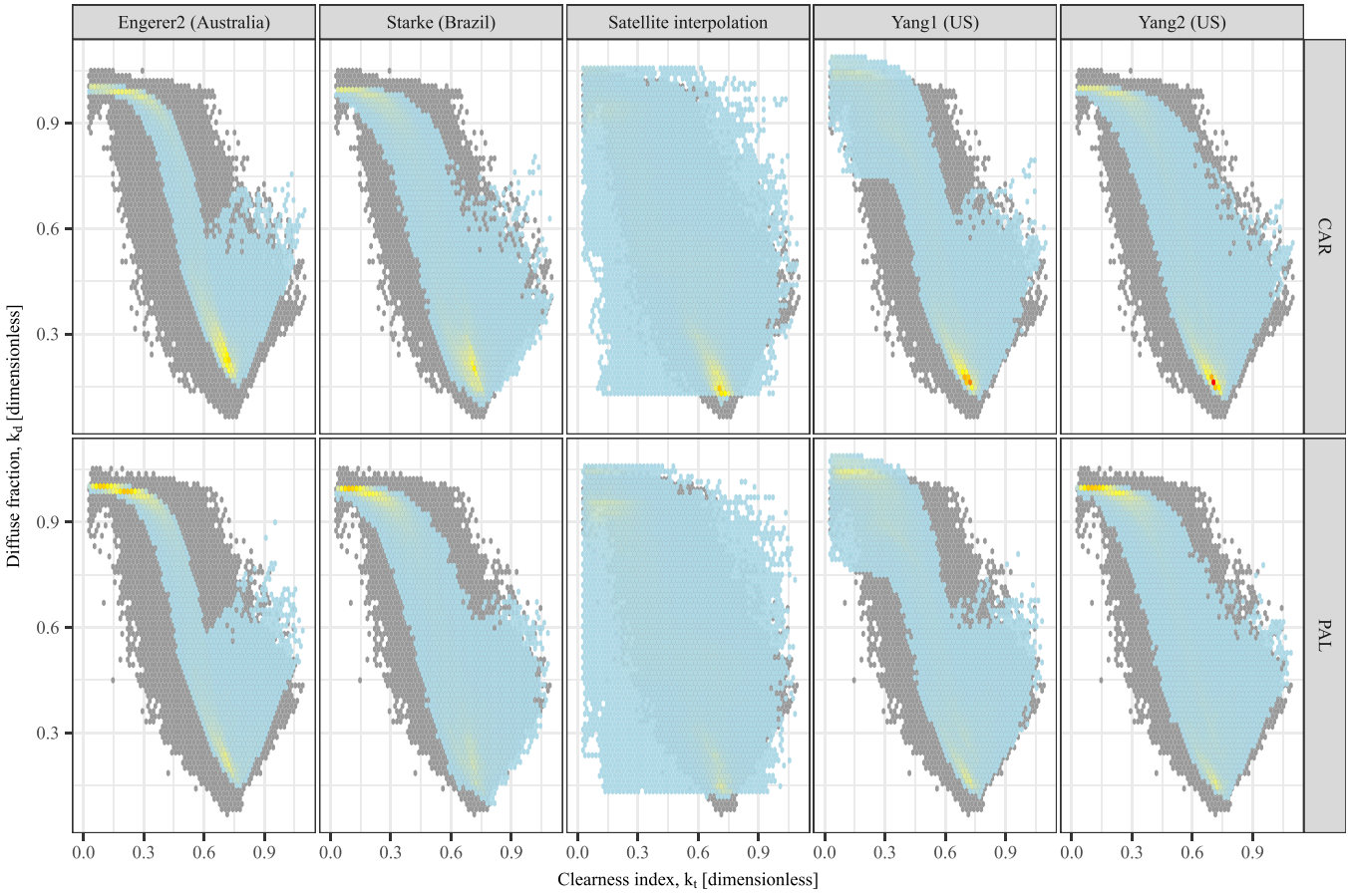

FIG. 5. Diffuse fraction data for the CAR (44.083°N, 5.059°E) and PAL (48.713°N, 2.208°E) stations, overlaid with the results of various separation models with global parameters. Parameters for Engerer2 and Starke are fitted using Australia and Brazil data, whereas the parameters for Yang1 and Yang1 are fitted using data from the continental US.

Published under license by AIP Publishing

better than BOLAND1, and so, such performance improvement has to be attributed to the additional AST. Notwithstanding, when the number of predictors gets large, the effect of each predictor in the empirical model becomes unclear. Physical approaches have to be used to expand the envelope of our knowledge regarding diffuse solar radiation.

V. CONCLUSION

Separation models estimate diffuse horizontal irradiance (DHI) from global horizontal irradiance (GHI) and other meteorological parameters. Due to the non-injective nature of DHI, i.e., a single GHI value can correspond to many DHI values, regression-based empirical models have clear limitations. Prior to 2016, ENGERER2 was identified as the quasi-universal separation model. Shortly after that, STARKE was shown to be superior compared to ENGERER2 using Australia and Brazil data. In this work, two satellite-augmented separation models are proposed, which outperform both STARKE and ENGERER2, resulting in the best separation models to date.

The importance of including satellite-based k_d estimates into separation modeling has been confirmed. Since satellite-based irradiance estimates are usually based on physical models, they are efficient and effective in explaining the low-frequency variability in DHI. Furthermore, due to the worldwide availability of such data, satellite-augmented models can be applied at all locations, except for the high-latitude regions where the satellite-image resolution is poor.

Empirical separation modeling has been the dominating practice in solar engineering. However, although empirical models can be fine-tuned to some extent as in this paper, it has limited improvement potential. Moving forward, the physically laws governing the DHI–GHI relationship should be emphasized.

SUPPLEMENTARY MATERIAL

See supplementary material for the instructions for reproducing partial results of Tables IV and VI and Fig. 3 presented in this article. The code is written in R, and several packages (and, of course, their dependencies, see http://stat.ethz.ch/R-manual/R-patched/library/tools/html/package.dependencies.html) need to be installed before the scripts can be executed.

- Package `dplyr` (https://cran.r-project.org/web/packages/dplyr/index.html) is a fast, consistent tool for working with data frame like objects, in both memory and out of memory.
- Package `ggplot2` (https://cran.r-project.org/web/packages/ggplot2/index.html) offers a powerful graphics language for creating elegant and complex plots.
- Package `lubridate` (https://cran.r-project.org/web/packages/lubridate/index.html) makes it easier to work with dates and times.
- Other packages include `insol` and `zoo`.

Data: Four data files, namely, `bon_2015-2016.txt`, `bon_psm30.txt`, `cab_2016.txt`, and `cab_cams.csv`, are included. The raw 1-min data from the BON station, over 2015–2016, are preprocessed with the basic solar-positioning routine. Hence, zenith angle, clear-sky, and extraterrestrial irradiance values can be found in `bon_2015-2016.txt`. On the other hand, `bon_psm30.txt` contains the corresponding NSRDB satellite-derived irradiance data obtained from the NREL website. Similarly, the 2016 data for one of the European sites, namely, CAB, are provided in `cab_2016.txt` (BSRN) and `cab_cams.csv` (CAMS-RAD). Finally, the fitted coefficients for the proposed separation model are provided as an RData file, namely, `pars.RData`.

Code: Two R scripts `bon_DHI.R` and `cab_DHI.R` reproduce partial DHI validation results shown in Tables IV and VI, and Fig. 3. To execute these scripts, the user only needs to change the working directory.

REFERENCES

[1]D. Erbs, S. Klein, and J. Duffie, "Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation," Sol. Energy **28**, 293–302 (1982).
[2]C. A. Gueymard and J. A. Ruiz-Arias, "Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance," Sol. Energy **128**, 1–30 (2016).
[3]D. Yang, "Solar radiation on inclined surfaces: Corrections and benchmarks," Sol. Energy **136**, 288–302 (2016).
[4]J. Boland, L. Scott, and M. Luther, "Modelling the diffuse fraction of global solar radiation on a horizontal surface," Environmetrics **12**, 103–116 (2001).
[5]J. Boland, B. Ridley, and B. Brown, "Models of diffuse solar radiation," Renewable Energy **33**, 575–584 (2008).
[6]N. Engerer, "Minute resolution estimates of the diffuse fraction of global irradiance for southeastern Australia," Sol. Energy **116**, 215–237 (2015).
[7]The word "effect" comes from the statistical modeling literature.
[8]B. Ridley, J. Boland, and P. Lauret, "Modelling of diffuse solar fraction with multiple predictors," Renewable Energy **35**, 478–483 (2010).
[9]A. R. Starke, L. F. Lemos, J. Boland, J. M. Cardemil, and S. Colle, "Resolution of the cloud enhancement problem for one-minute diffuse radiation prediction," Renewable Energy **125**, 472–484 (2018).
[10]R. G. Rojas, N. Alvarado, J. Boland, R. Escobar, and A. Castillejo-Cuberos, "Diffuse fraction estimation using the BRL model and relationship of predictors under Chilean, Costa Rican and Australian climatic conditions," Renewable Energy **136**, 1091–1106 (2019).
[11]J. Boland, J. Huang, and B. Ridley, "Decomposing global solar radiation into its direct and diffuse components," Renewable Sustainable Energy Rev. **28**, 749–756 (2013).
[12]This naming convention takes the initials from the authors' last names from Ref. 8.
[13]This conclusion is drawn based on the fact that nRMSEs of the new models are much smaller than those of STARKE (the most recent advance in separation modeling in 2018) and ENGERER2 (the quasi-universal model confirmed in 2016).
[14]J. Ruiz-Arias, H. Alsamamra, J. Tovar-Pescador, and D. Pozo-Vázquez, "Proposal of a regressive model for the hourly diffuse solar radiation under all sky conditions," Energy Convers. Manage. **51**, 881–893 (2010).
[15]Arguably, one can consider using KT from yesterday and y from the previous three time steps so that the model gains the real-time property. Nevertheless, the performance of such a model requires further verification.
[16]C. A. Gueymard, "REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation – Validation with a benchmark dataset," Sol. Energy **82**, 272–285 (2008).
[17]K. Hollands, "A derivation of the diffuse fraction's dependence on the clearness index," Sol. Energy **35**, 131–136 (1985).
[18]K. Hollands and S. Crha, "An improved model for diffuse radiation: Correction for atmospheric back-scattering," Sol. Energy **38**, 233–236 (1987).
[19]The word "range" is used in the context of "domain and range of a function."
[20]D. M. Bates and D. G. Watts, *Nonlinear Regression Analysis and Its Applications* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 1988).
[21]D. Yang, G. M. Yagli, and H. Quan, "Quality control for solar irradiance data," in *2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)* (2018), pp. 208–213.
[22]D. Yang, "SolarData: An R package for easy access of publicly available solar datasets," Sol. Energy **171**, A3–A12 (2018).
[23]M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (NSRDB)," Renewable Sustainable Energy Rev. **89**, 51–60 (2018).

Published under license by AIP Publishing

[24]Y. Xie, M. Sengupta, and J. Dudhia, "A fast all-sky radiation model for solar applications (FARMS): Algorithm and performance evaluation," Sol. Energy **135**, 435–445 (2016).
[25]D. Yang, "Kriging for NSRDB PSM version 3 satellite-derived solar irradiance," Sol. Energy **171**, 876–883 (2018).
[26]D. Yang, "Spatial prediction using kriging ensemble," Sol. Energy **171**, 977–982 (2018).
[27]D. Yang, "A correct validation of the national solar radiation data base (NSRDB)," Renewable Sustainable Energy Rev. **97**, 152–155 (2018).
[28]A. Driemel, J. Augustine, K. Behrens, S. Colle, C. Cox, E. Cuevas-Agulló, F. M. Denn, T. Duprat, M. Fukuda, H. Grobe, M. Haeffelin, G. Hodges, N. Hyett, O. Ijima, A. Kallis, W. Knap, V. Kustov, C. N. Long, D. Longenecker, A. Lupi, M. Maturilli, M. Mimouni, L. Ntsangwane, H. Ogihara, X. Olano, M. Olefs, M. Omori, L. Passamani, E. B. Pereira, H. Schmithüsen, S. Schumacher, R. Sieger, J. Tamlyn, R. Vogt, L. Vuilleumier, X. Xia, A. Ohmura, and G. König-Langlo, "Baseline surface radiation network (BSRN): structure and data description (1992–2017)," Earth Syst. Sci. Data **10**, 1491–1501 (2018).
[29]Z. Qu, A. Oumbe, P. Blanc, B. Espinar, G. Gesell, B. Gschwind, L. Klüser, M. Lefèvre, L. Saboret, M. Schroedter-Homscheidt, and L. Wald, "Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method," Meteorol. Z. **26**, 33–57 (2017).
[30]If the errors of satellite interpolation are comparable to separation models, one can directly use the downscaled satellite-derived DHI.
[31]M. Lin, H. C. Lucas, and G. Shmueli, "Too big to fail: Large samples and the p-value problem," Inf. Syst. Res. **24**, 906–917 (2013).
[32]R. Perez, P. Ineichen, K. Moore, M. Kmiecik, C. Chain, R. George, and F. Vignola, "A new operational model for satellite-derived irradiances: Description and validation," Sol. Energy **73**, 307–317 (2002).

Published under license by AIP Publishing