Worldwide Climate Modeling of Diffuse Solar Fraction from One-minute Irradiance

Cristian M. Barni[1], Allan R. Starke[1], Leonardo F. L. Lemos[1], José M. Cardemil[2], John Boland[3] and Sergio Colle[1]

[1] LEPTEN - Laboratory of Energy Conversion Engineering and Energy Technology, Department of Mechanical Engineering, Federal University of Santa Catarina (UFSC), Florianopolis (Brazil)

[2] Department of Mechanical Engineering, Universidad de Chile. Beauchef 851, Santiago (Chile)

[3] Centre for Industrial and Applied Mathematics, University of South Australia, Mawson Lakes Boulevard, Mawson Lakes (Australia)

Abstract

This work presents a new one-minute diffuse fraction model based on the BRL-minute model, with a new predictor added to it and the use of a robust nonlinear regression method that can minimize the effects of outliers on the dataset without removing them. The presented model is also regressed using the BSRN and BOM data from many stations belonging to the same climate according to the Köppen-Geiger classification. A formal error analysis is also performed comparing the new model with the BRL and Engerer models.

Keywords: Diffuse and direct irradiance, BSRN, Köppen-Geiger climate, Robust regression, Separation models

1. Introduction

Diffuse fraction models, or separation models, have been extensively studied since the initial developments in the 1960s by Liu & Jordan (1960). In general, these models are based on measured data from different stations in different climatic conditions, use different temporal scales (e.g. monthly, daily, hourly, minute) and make use of one or several predictors (independent variables) (Aler, Galván, Ruiz-Arias, & Gueymard, 2017).

The developments in this area are driven by the need of accurate values of the three irradiance components, namely the global horizontal irradiance (G), diffuse horizontal irradiance (G_d) and direct normal irradiance (G_b), which are fundamental for the correct design and performance assessment of solar energy systems. However, measuring these three components requires complex tracking devices and significant operational efforts, which lead to high costs. As a result, there are several stations measuring only global irradiance, mainly because it requires less operational efforts and costs. A common solution to overcome the absence of measurements of the other two components is to use some sort of separation model to estimate both diffuse and direct components from global irradiance observations. Several separation models have been developed using hourly irradiance data, however such models do not appropriately describe fast transient episodes in solar irradiance, which happen at time scales much smaller than an hour. As a result, these models are not adequate to meet the current demands of the industry (Gueymard, 2017a, 2017b; Gueymard & Ruiz-Arias, 2016).

Gueymard and Ruiz-Arias (2016) reported an extensive validation study with 140 separation models using one-minute data and reported that cloud enhancement events (CEE) and high-albedo effects intensify the errors in irradiance estimates of separation models. Based on new criteria to evaluate the robustness of a separation model and comparing each model's performance within different climate zones, the authors recommended the development of separation models for each climate zone. The authors also concluded that models that consider variability and clear-sky irradiance as predictors tend to perform better.

Recently, Starke et al. (2018) presented a logistical model for one-minute irradiance based on the BRL model. The proposed model, named the BRL-minute model, uses a piecewise function, defined by two sub-domains,

doi:10.18086/swc.2019.42.01 Available at http://proceedings.ises.org

one for CEE and the other for non-CEE. The authors proposed a model for Brazil and another for Australia, using one-minute data from four BSRN stations and four stations of the Australian Government Bureau of Meteorology (BOM).

The main goal of this study is to present the further developments of the BRL-minute model, as one-minute irradiance data from many stations spread worldwide were used to develop diffuse fraction models for climate zones from the Köppen-Geiger climate classification, following the suggestion by Gueymard and Ruiz-Arias (2016). Also, a new predictor was added to the model and a robust nonlinear regression method was used to account for the data heteroscedasticity and possible outliers. Finally, to assess the performance of the new models, a formal error analysis was performed, based on comparisons between this work and the models of Ridley et al. (2010) and Engerer (2015), which figured among the best performing models selected in the work by Gueymard and Ruiz-Arias (2016).

2. Methodology

The model proposed here is based on the work of Starke et al. (2018), which proposed a piecewise logistical model that can predict the fast transient phenomena observed in 1-min data. The division of the model in two parts aimed to separate diffuse fraction values that correspond to cloud enhancement from those that do not. As proposed by Starke et al. (2018), the variables k_T and K_{CSI} were used to determine the domain breakpoint, i.e. to identify whether a data point is a cloud enhancement event or not,. In this way, variables with physical interpretation are used to identify cloud enhancement events, thus the value of the domain breakpoint is not arbitrarily defined, nor does it need to be defined by the regression process.

A few changes were made to the original model, as the hourly clearness index (K_T) was added as a model predictor. The reason of adding this predictor is based on the same logic used at the BRL model when the daily clearness index ($\overline{K_T}$) was used as predictor – if it seems logical to use daily clearness index to classify similar days on an hourly model, therefore it also seems logical to use hourly clearness index to classify similar hours on an one-minute model. It is worth mentioning that there is no additional complexity of adding this predictor, since it is easy to obtain the hourly clearness index using the predictors already present on the model – one minute and daily clearness indexes. The subdomain intervals were also reorganized in a clearer form, where the first equation models the cloud enhancements events and the other equation models the other events. The model proposed on this work is given by,

$$\hat{d} = \begin{cases} \frac{1}{1+e^{(\beta_0+\beta_1 k_T+\beta_2 AST+\beta_3\alpha+\beta_4 K_T+\beta_5\psi+\beta_6 CSI+\beta_7 K_T)}}, & K_{CSI} \geq 1.05 \; and \; k_T > 0.65 \\ \frac{1}{1+e^{(\beta_8+\beta_9 k_T+\beta_{10} AST+\beta_{11}\alpha+\beta_{12}\overline{K_T}+\beta_{13}\psi+\beta_{14} CSI+\beta_{15} K_T)}}, & Otherwise \end{cases} \quad \text{(eq. 1)}$$

where $\hat{d}$ is the modelled diffuse fraction, k_T is the clearness index at minute basis, defined as the ratio of the global horizontal irradiance (G) to the horizontal extra-terrestrial irradiance at the top of the atmosphere (G_o). AST is the apparent solar time in hours, α is the solar altitude in degrees, ($\overline{K_T}$) is the daily clearness index and ψ is a persistence factor defined by Ridley et al. (2010). CSI is the clear-sky irradiance in Wm^{-2}, K_T is the hourly clearness index and K_{CSI} is the ratio of the measured G and the CSI. It is worth mention that the values of the breakpoint were maintained the same as the ones presented by Starke et al. (2018). Moreover, to avoid any confusion and misunderstanding, the CSI inputs to the Eq. (1) are now given in Wm^{-2}, unlike the originally proposed by Starke et al. (2018), which were in MJ h^{-1} m^{-2}.

2.1. Irradiance Data

The quality of any separation model depends on the limitations and experimental error of the measured irradiance, depending on the radiometer's performance, radiometer calibration, station maintenance and instrument cleaning (Gueymard & Ruiz-Arias, 2016). To mitigate the impact of these factors on our model, only data from research-class stations were used.

The one-minute database used contains the three irradiance components measured with thermopile radiometers. Most of the stations (51) belong to the Baseline Surface Radiation Network (BSRN) and one station belongs to the Australian Government Bureau of Meteorology (BOM). A summary of all stations considered herein is presented in Table 1. It is worth mentioning that only data after the year 2000 is being considered, as most stations do not have one-minute irradiance measurements before that.

2.2. Climate classification

The Köppen-Geiger climate classification was used to classify the different stations considered on the present work. To do so, data provided on the work by Chen and Chen (2013) were used. The authors used global temperature and precipitation observations over the period of 1901–2010 to build the Köppen classification dataset on the interannual, interdecadal, and 30-year time scales. The result is a classification indicated by a letter code, where the first letter indicates the main climate type (A for tropical, B for dry, C for temperate, D for continental and E for polar), and the following letters denote the specific climate type. The reader is referred to the work by Chen and Chen (2013) for a more detailed description of the letters in the Köppen classification scheme. In the present work, the 30-year time scale was used, with data from 1981-2010, comprising most of the measurement period of the stations used. The Köppen climate of each station is also presented in Table 1.

2.3. Clear-sky model

The clear-sky irradiance (CSI) values needed in this work were calculated by the broadband simplified analytical version of the Solis model (Ineichen, 2008). This model estimates clear-sky irradiance at the evaluated site by multiplying the extraterrestrial irradiance (calculated using site latitude and longitude) by a correction factor that is a function of aerosol optical depth (AOD), atmosphere water vapor content (W) and site altitude. AOD and W values were taken from the CAMS Reanalysis dataset, which provides atmospheric composition data from 2003 up to 2017. For stations with data extending beyond 2017 (or before 2003), the information of the last (first) year is used as an approximation of the following (previous) years, as Starke et al. (2018) have shown that this approximation has a minor impact on the model performance.

Some of the stations used in this work are placed in areas where the site elevation is much higher than the average pixel elevation for the dataset used, which caused inconsistencies when computing the CSI estimate. In order to address this problem, an altitude correction was implemented as described by Bright and Gueymard (2019) to equalize the reference altitude for all datasets and it is calculated using the equation:

$$AOD(h) = AOD(h_0) \times \exp\left[\frac{(h-h_0)}{H_a}\right] \quad \text{(eq. 2)}$$

Where $AOD(h)$ is the AOD at the station altitude h, $AOD(h_0)$ is the AOD estimate from the CAMS dataset with reference altitude h_0, which was obtained using the geopotential measurement for the considered pixel (Bright & Gueymard, 2019b), also available on the CAMS dataset, and H_a is the scale height, whose value was adopted as a constant of 2100m, the value suggested by the referred authors for coastal sites. The same procedure was used to correct atmospheric water vapor content, by using W instead of AOD in the equation and the same scale height.

2.4. Quality control

The first quality checks applied to the solar data were the ones used by Gueymard and Ruiz-Arias (2016) in their extensive review. These checks ensure that the measured irradiance value is physically possible and that the three components of solar radiation are mutually coherent. Further quality checks were made based on the quality control methodology described by Lemos et al. (2017), such as the “Rayleigh limit test”, which guarantees that the measured diffuse irradiance is not below a minimum physically possible value. The “overcast test”, also listed by the referred authors, was applied as a lower allowable limit to G.

The number of valid data points (i.e. qualified data) for each station is presented in Table 1. Stations with less than 262800 qualified data points (roughly the sunlight minutes within one year) were removed from the pool.

Table 1 – General information on the 52 stations whose data were available for this study, including, in order of appearance, a three letter code, the station complete name, latitude and longitude in degrees, elevation in meters above sea level, the data source, the Köppen-Geiger climate and the number of valid data points after the quality control.

	Code	Station	Lat.	Long.	Elevation	Source	Climate	Data points
1	ADL	Adelaide	-34,929	138,601	61,7	BOM	C	444557
2	ALE	Alert	82,490	-62,420	127	BSRN	E	860461
3	ASP	Alice Springs	-23,798	133,888	547	BSRN	B	3414376
4	BAR	Barrow	71,323	-156,607	8	BSRN	E	571991
5	BER	Bermuda	32,267	-64,667	8	BSRN	C	727767
6	BIL	Billings	36,605	-97,516	317	BSRN	C	1326227
7	BOS	Boulder	40,125	-105,237	1689	BSRN	C	443931
8	BOU	Boulder	40,050	-105,007	1577	BSRN	C	685784
9	BRB	Brasilia	-15,601	-47,713	1023	BSRN	A	960528
10	CAB	Cabauw	51,971	4,927	0	BSRN	C	2101994
11	CAM	Camborne	50,217	-5,317	88	BSRN	C	1700709
12	CAR	Carpentras	44,083	5,059	100	BSRN	C	3230394
13	CLH	Chesapeake Light	36,905	-75,713	37	BSRN	C	2579937
14	CNR	Cener	42,816	-1,601	471	BSRN	C	1362716
15	COC	Cocos Island	-12,193	96,835	6	BSRN	A	1795630
16	DAA	De Aar	-30,667	23,993	1287	BSRN	B	1046750
17	DAR	Darwin	-12,425	130,891	30	BSRN	A	2221262
18	DOM	Concordia Station	-75,100	123,383	3233	BSRN	E	1134106
19	DWN	Darwin Met Office	-12,424	130,893	32	BSRN	A	1797528
20	E13	Southern Great Plains	36,605	-97,485	318	BSRN	C	1839161
21	EUR	Eureka	79,989	-85,940	85	BSRN	E	601728
22	FLO	Florianopolis	-27,605	-48,523	11	BSRN	C	737731
23	FUA	Fukuoka	33,582	130,376	3	BSRN	C	1247060
24	GCR	Goodwin Creek	34,255	-89,873	98	BSRN	C	303879
25	GOB	Gobabeb	-23,561	15,042	407	BSRN	B	1227162
26	GVN	Georg von Neumayer	-70,650	-8,250	42	BSRN	E	1820132
27	ISH	Ishigakijima	24,337	124,164	5,7	BSRN	C	1369626
28	IZA	Izaña	28,309	-16,499	2372,9	BSRN	C	1736789
29	KWA	Kwajalein	8,720	167,731	10	BSRN	A	581434
30	LAU	Lauder	-45,045	169,689	350	BSRN	C	2479489
31	LER	Lerwick	60,139	-1,185	80	BSRN	C	1285996
32	LIN	Lindenberg	52,210	14,122	125	BSRN	C	2191444
33	LRC	Langley Research Center	37,104	-76,387	3	BSRN	C	579563
34	MAN	Momote	-2,058	147,425	6	BSRN	A	2121929
35	MNM	Minamitorishima	24,288	153,983	7,1	BSRN	A	1654583
36	NAU	Nauru Island	-0,521	166,917	7	BSRN	A	1602723
37	NYA	Ny-Ålesund	78,925	11,930	11	BSRN	E	1786803
38	PAL	Palaiseau	48,713	2,208	156	BSRN	C	1918094
39	PAY	Payerne	46,815	6,944	491	BSRN	C	1568043
40	PSU	Rock Springs	40,720	-77,933	376	BSRN	D	363090
41	PTR	Petrolina	-9,068	-40,319	387	BSRN	B	1199345
42	REG	Regina	50,205	-104,713	578	BSRN	D	1875990
43	SAP	Sapporo	43,060	141,329	17,2	BSRN	D	1284085
44	SBO	Sede Boqer	30,860	34,779	500	BSRN	B	788147
45	SMS	São Martinho da Serra	-29,443	-53,823	489	BSRN	C	1040359
46	SON	Sonnblick	47,054	12,958	3108,9	BSRN	D	431596
47	SOV	Solar Village	24,910	46,410	650	BSRN	B	473645
48	SPO	South Pole	-89,983	-24,799	2800	BSRN	E	1126294
49	SYO	Syowa	-69,005	39,589	18	BSRN	E	1656897
50	TAM	Tamanrasset	22,790	5,529	1385	BSRN	B	1595877
51	TAT	Tateno	36,058	140,126	25	BSRN	C	3220673
52	TOR	Toravere	58,254	26,462	70	BSRN	D	1894408

2.5. Regression method

Boland and Ridley (2008) demonstrated the procedures for the construction of the logistical separation model, while Ridley et al. (2010) presented the method for building the BRL model using a nonlinear least-squares (NLS) regression method. However, the presence of outliers in the data can severely affect the results when the separation model equation is adjusted to measured data through a least-squares fit. Outlier removal methods have been presented by Younes et al. (2005) and Lemos et al. (2017), which consist of creating an envelope around plausible data, using specified functions. However, these methods require the individual inspection of each station's data and site-specific knowledge of the climate and irradiance behavior. Therefore, building an envelope-based outlier removal procedure for a heterogeneous database such as the BSRN is an infeasible task.

One solution for building the separation models without removing the outliers in advance is to use robust statistics, i.e. a robust nonlinear regression method. The methods of the robust approach can be employed to produce reliable parameter estimates when the data follow an arbitrary, non-normal distribution, that is, when the data is heteroscedastic. As highlighted by Riazoshams et al. (2018), in real-life data, the homoscedastic assumption might not be correct for every dataset. This can happen because of the natural behavior of the data, or because the data is affected by some discrepancy on the observations – the outliers.

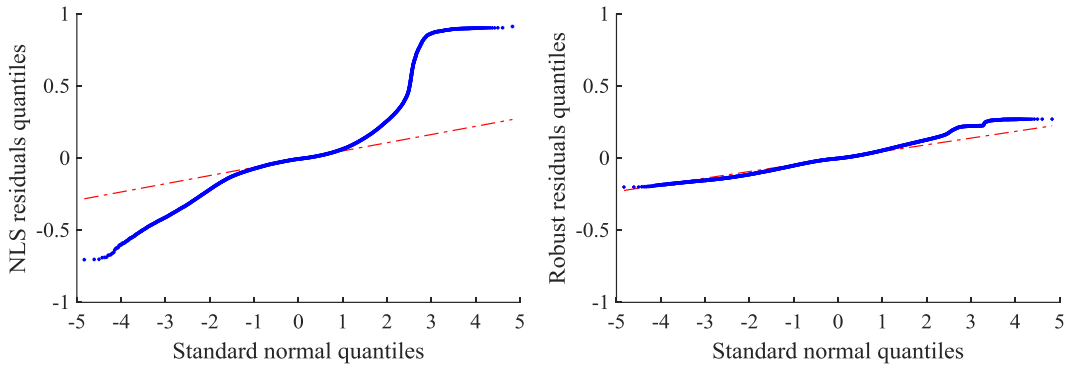

Fig. 1 – Q-Q plots of the residuals of the regressions for Florianopolis/Brazil using (a) a NLS regression method and (b) a robust nonlinear regression method; The red dashed lines represent a normal distribution for each regression, while the blue markers represents the distribution for the residuals of each regression.

In the case of the data from Florianopolis/Brazil, Fig. 1(a) shows that the residuals of the adjusted model that uses a NLS regression method , deviates a lot from a normal distribution, while Fig. 1(b) shows that when a robust least-square method is used, this discrepancy becomes a lot smaller, guaranteeing that the statistical indicators will be closer to the exact ones. Because of that, a robust nonlinear regression method, available in the MathWorks (2018) software MATLAB, was adopted to build the diffuse fraction models of the present study, i.e. determine the β_i coefficients. An iterative reweighted least squares algorithm (Dumouchel & O'Brien, 1991; Holland & Welsch, 1977) is used in the software, which recalculates the weights at each iteration based on the residual of the observations of the last iteration. This approach reduces the influence of the outliers on the fit at each iteration, where the process continues until the weights converge. A weight function and a tuning constant need to be defined for the robust fitting. After testing different equations for the weight's estimation, we decided to use the following logistic function,

$$w_i = \frac{\tanh(r_i)}{r_i} \qquad \text{(eq. 3)}$$

where w_i is the robust weight for the residual of the regression r_i on the observation i. The tuning constant used was 1.205, the default given by the MATLAB function. Fig. 2 shows the estimated weights for the Florianopolis/Brazil dataset and it can be seen that the majority of high weight values are on a region comparable to the envelopes proposed by Younes et al. (2005), which is the reason for the robust method to

work without the need of an outlier removal procedure. Even though this method does not completely remove the influence of obvious outliers, it assures that even values that may be improbable but are rightfully measured are considered on the diffuse fraction model.

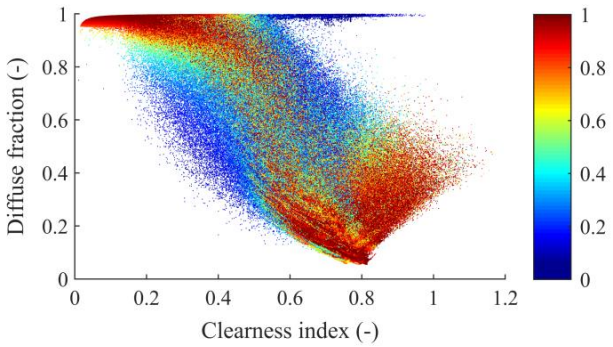

Fig. 2 – Observed clearness index (k_T) and diffuse fraction (d) correlation of 1-min irradiance data from Florianopolis/Brazil; The colormap denotes the estimated weight w_i for each point.

Also, since the proposed model is a piecewise equation, the regression method is used to estimate the model coefficients of each part of the domain. That is, for every dataset, the datapoints were split in two classifications, CEE or non-CEE, and each subset was individually submitted to the regression method in order to estimate the coefficients for each part of the final model and the weights for each datapoint.

2.6. Building local and climate-specific models

After the qualification procedure, the regression method was applied to the data from each individual station to create a locally optimized model, valid for that station. On the other hand, to create a model valid for a region (climate, country or "universal") the data from individual stations had to be merged to a single data set, and then analyzed with the defined regression method to create a regional model. In both cases, two thirds of the local data were randomly selected to feed the model, while the remaining third was used to evaluate it.

To build the climate models, we randomly selected the same amount of datapoints from each station belonging to the same broad climate zones in the Köppen classification (A, B, C, D and E), making the contributions from each station to the climate dataset to be equal. To each of these datasets a robust regression was made using the new proposed model and generating a set of coefficients for each, available on Table 2.

2.7. Statistical indicators of model performance

Gueymard (2014) presented a complete review of performance indicators that can be used in radiation models for validations purposes. Among those, three statistical indicators were considered for the formal error analysis: normalized RMSE, normalized MBE, and the KSI. In the present work, two of those indicators have been slightly modified to be able to use weighted residuals, as in the following expressions,

$$nRMSE = \frac{100}{\bar{d}}\sqrt{\frac{\sum_{i=1}^{n} {r_i}^2}{n}} \quad \text{(eq. 4)}$$

$$nMBE = \frac{100}{\bar{d}}\frac{\left[\sum_{i=1}^{n} r_i\right]}{n} \quad \text{(eq. 5)}$$

where $\bar{d}$ is the mean value of the diffuse fraction of the measurements data set, and r_i is the residual of the measurement i, of a set of n measurements, which, when using a robust regression method, can be defined as

$$r_i = \sqrt{w_i}\left(\hat{d}_i - d_i\right) \quad \text{(eq. 6)}$$

where d_i is the actual value of the diffuse fraction calculated, $\hat{d}_i$ is the estimated value of the diffuse fraction

for the point i, and w_i is the weight estimated using the robust method for the point i.

When using data without removing the outliers, the use of weighted residuals is a logical way to calculate the goodness of fit (GoF) of the regressed model, that is, how well the model can predict the main trend of the dataset it was created with, reducing the penalty for predicting a different value for an outlier. However, when using these indicators to evaluate the robust generated model performance on a different dataset, or another model where the weights are not available, setting the weights as w_i=1 for all points reduces Eq. (6) to the standard residuals of the NLS method:

$$r_i = \left(\widehat{d}_i - d_i\right) \quad \text{(eq. 7)}$$

The third statistical indicator is the Kolmogorov–Smirnov test Integral (KSI), which was proposed by Espinar et al. (2009) as a measure to compute the differences between CDFs, and is calculated as follows,

$$KSI = \frac{100}{\alpha_c}\int_{x_{min}}^{x_{max}} D_n dx \quad \text{(eq. 8)}$$

where x_{min} and x_{max} are the values that limit the independent variables, D_n is the difference between the CDFs of the measurements and the estimated values, while α_c is determined by $\alpha_c = V_c(x_{max} - x_{min})$. The critical value V_c is given by $V_c = 1.63/\sqrt{n}$,, for a 99% level of confidence, on a population size of $n \geq 35$. As the KSI becomes closer to zero, the distributions can be considered equivalent, and for large values, the model is considered non-fitting to the dataset. It is important to note that two different large datasets may rarely be considered equivalent, as small differences between them can become huge when added. Also, for large populations, like the ones used in this work, the critical value tends to zero, in consequence α_c tends to a small value and KSI will become a large value, possibly larger than 100 %.

2.8. Other separation models

In order to assess the performance of this new model, an error analysis is performed which is based on comparisons between our work and the models of Ridley et al. (2010) and Engerer (2015). The models selected for comparison are well known and are often referred to in the literature. Indeed, these models are among the ones selected in the review by Gueymard and Ruiz-Arias (2016) as the best performing separation models.

3. Local adjusted models

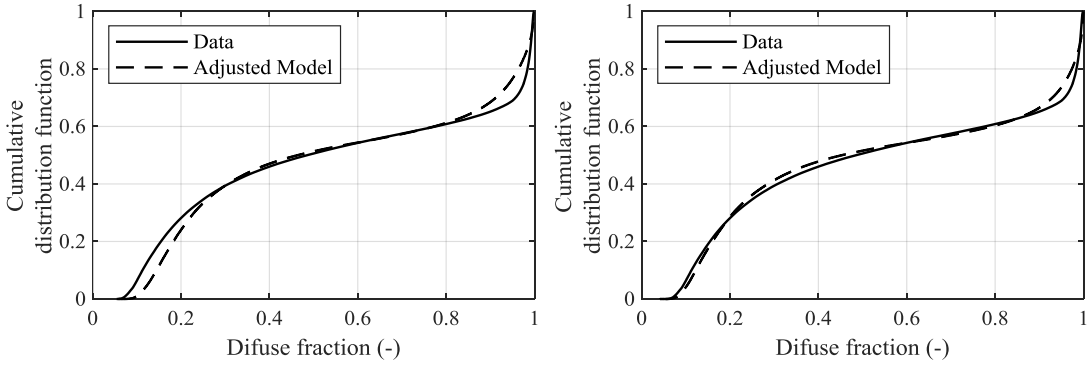

Fig. 3 – CDFs for the diffuse fraction in Florianopolis/Brazil. The continuous line represents the actual dataset, while the dashed line represents the results from the model generated using a NLS (left) and a robust (right) regression methods.

In order to assess the improvements of the model when regressed using a consistent dataset, two regressions for each station on Table 2 were made, one using the NLS method and another using the robust method. Both regressions produce similar results, but the robust method suffers less influence from the points that do not follow the main trend, like occasional outliers. The results for Florianopolis/Brazil are presented on Fig. 3,

where the robust method generates a model whose CDF resembles the data points CDF more than the NLS adjusted model, especially on the clear sky region.

We also performed a formal error analysis using the statistical indicators presented on section 2.7 for each model. The benefits of the proposed methodology – a robust regression method – can be seen on Fig. 4 on the top, which depicts the KSI indicators of the estimated data created by each local adjusted model. The proposed methodology is consistently better than the NLS method, resulting in lower values of KSI, therefore, the models created by the robust regression provide estimates with better similitude to the measured data. This is achieved because this method reduces the influence of outliers, providing a set of coefficients that gives better estimates.

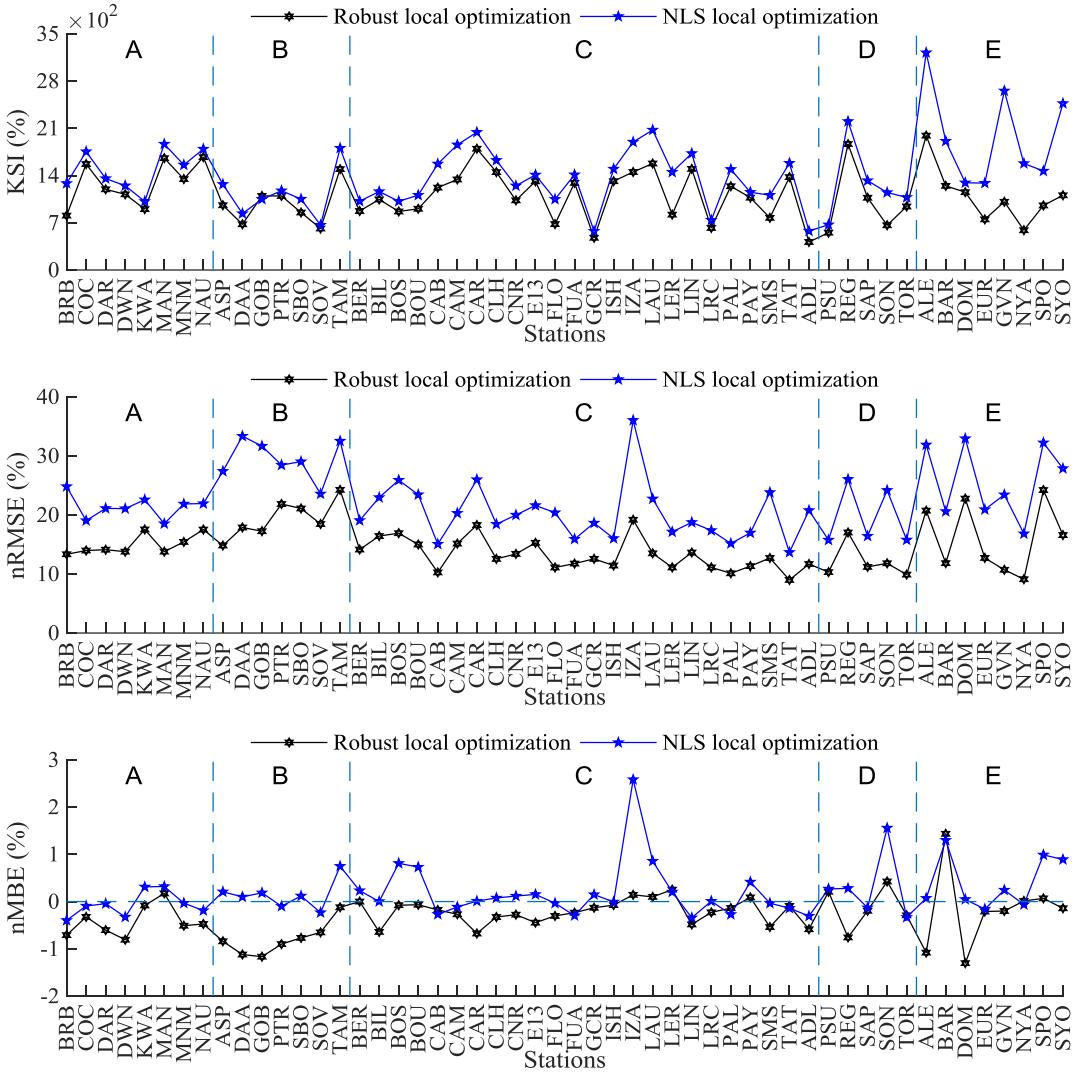

Fig. 4 – Comparison between KSI (top), nRMSE (middle) and nMBE (bottom) error indicators for each station when using a NLS or a robust regression method. The black markers represent the robust method and the blue markers represent the NLS method. The stations are grouped by climate. The lines between each marker are presented just to ease the visualization.

The middle and bottom plots on Fig. 4 present the goodness of fit of the regressions performed by the robust and NLS methods, in terms of the nRMSE and nMBE. As mentioned before, when using the robust method these indicators were calculated using the weighted residuals. Thus, the influence of the large values of residuals

caused by the outliers are reduced, thereby providing a "real" measure of accuracy of the models.

The nRMSE values calculated by the models created using the robust method are significantly lower than the ones calculated by the models created using the NLS method, obtaining results lower than 20%, with some exceptions from climate B and E, and reaching less than 10% in some cases. It is important to note that if no weights are employed on the calculation of the nRMSE of the models created using the robust approach, it will result in values similar to the ones observed for the models created by the NLS method, which demonstrates that the differences between the nRMSE values obtained with the two methodologies are due to the existence of the outliers on the dataset, and that the proposed methodology is a feasible solution to create diffuse fraction models without having to deal with the problems caused by them.

Fig. 4 also shows on the bottom plot the nMBE calculated using the models created with the two methodologies, resulting in larger biases for the robust approach than for the NLS. This behavior is expected, because the models created with the robust method are created to be biased to the main trend, allowing it to ignore the outliers, while the models created with the NLS tends to fit the outliers. The weighted residuals tend to reduce the influence of outliers on the nMBE, but it does not remove it completely. On the other hand, it also can be observed that the robust approach tends to remove some atypical behavior – large values of nMBE observed at some station with the models created with the NLS regression method (e.g. TAM, BOS, BOU, IZA, SON, SPO, SYO) – which may be due to the large amount of outliers on the dataset. In general, the proposed methodology produces models with bias lower than one percent.

4. Climate zone models

As the main objective of this work, we created diffuse fraction models for each major Köppen-Geiger climate. As described on section 2.6, we merged the data for each major climate and regressed the model coefficients to each of these datasets. The coefficients generated by the robust regression are available on Table 2.

Table 2 – Set of coefficients generated for each climate model

Coefficients	Climate model				
	A	B	C	D	E
β_0	0,42564	-0,34423	0,36991	1,04274	1,20864
β_1	-3,59398	-3,40409	-3,45296	-4,02595	-5,76996
β_2	-0,00165	0,00870	0,00401	0,00033	-0,00067
β_3	-0,01270	-0,03086	-0,02946	-0,03336	-0,08691
β_4	1,76296	2,22625	1,43024	1,44008	1,63921
β_5	0,78023	0,71716	0,80691	0,76189	0,60128
β_6	0,00230	0,00317	0,00314	0,00341	0,00611
β_7	1,15743	1,84879	1,63014	1,49343	3,30922
β_8	-6,97014	-6,52494	-7,22087	-7,73779	-11,23789
β_9	6,05413	6,38235	6,80208	7,27602	10,07770
β_{10}	-0,00197	0,01248	0,00263	-0,00033	0,00711
β_{11}	0,00004	-0,01927	0,01965	0,09402	0,50304
β_{12}	2,97457	1,95447	2,41274	2,06922	1,77769
β_{13}	1,07312	0,76189	0,81176	0,99505	0,30251
β_{14}	-0,00093	0,00044	-0,00278	-0,00784	-0,03284
β_{15}	2,79782	2,57401	3,30092	3,76271	5,93868

After that, we reapplied each model for all the data in each station that belonged to that climate to perform a formal error analysis. Since we applied each climate model to each station and compared the performance with other models, the error indicators were calculated without any weights. Therefore, the errors presented herein

consider the outliers present in the datasets.

Fig. 5 shows the formal error analysis of the proposed climate models, BRL model, Engerer model, and GoF of the locally adjusted models, presented here as the best estimation for that station. Significant improvements for climate B, D and E can be observed on Fig 5 at the top and middle plots, where the proposed climate models perform significantly better than the BRL and Engerer models, two "universal" models.

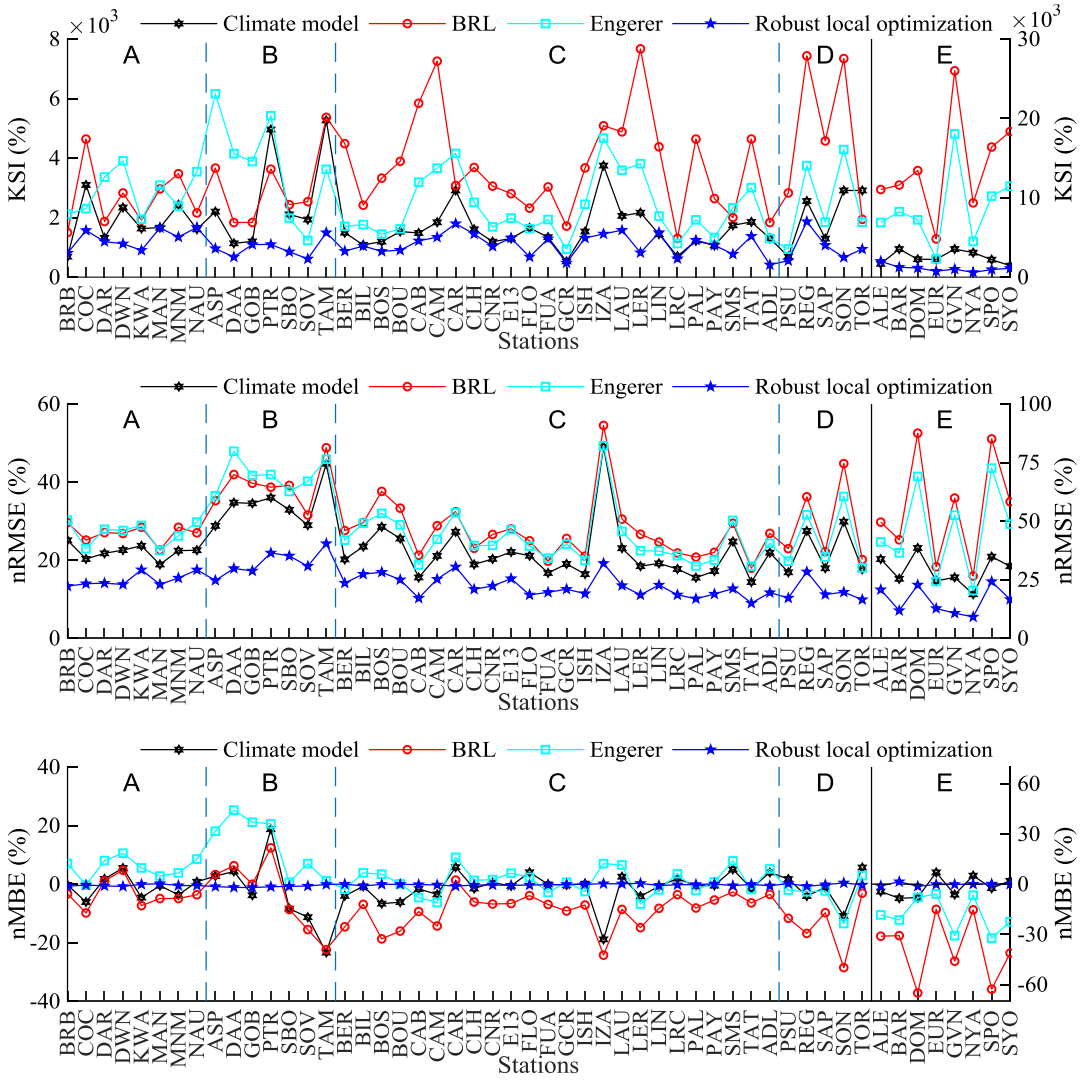

Fig. 5 – Comparison between the KSI (top), nRMSE (middle) and nMBE (bottom) error indicators for the climate (black), BRL (red), Engerer (lighter blue) and locally optimized (darker blue) models. The stations are grouped by climate. The right axis represents the values for climate E only. The lines between each marker are presented just to ease the visualization.

In the middle of Fig. 5 the climate models proposed produce nRMSE values lower than the values observed on the other models on most occasions. It can be observed that the proposed models perform significantly better for climate B and E, showing that our model can behave much better than any of the other compared models.

Regarding the nMBE, the climate models present lower bias when applied in each station, especially for climate C, showing nMBE values closer to the ones observed on the local adjusted models, when compared to the BRL and Engerer models. However, there are some stations where the climate models presented unusually high

nMBE (e.g. IZA, PTR, TAM), which could indicate that the station doesn't belong to this climate or doesn't behave like the other stations on this climate. This result indicates that there may exist a better way to classify stations with similar behavior, perhaps using cluster analysis or machine learning techniques.

The main advantage of the proposed methodology is depicted on the bottom plot of Fig. 5, in terms of the KSI. It can be observed that, in general, the climate models have KSI values like the ones observed on the local adjusted model, which shows that the climate models estimate diffuse fraction values with the same similitude observed on real data. Unfortunately, some exceptions can also be observed, which can reinforce the idea that a particular station doesn't behave like the other stations on the proposed climate.

When comparing the KSI of the proposed model and the one for the BRL model, it is expected to see a significant improvement, as the BRL model was developed using hourly data, which does not capture satisfactorily the CEE. However, even when compared to a minute model as the Engerer, it still holds better results in most cases, mainly on stations from climate B and E.

5. Conclusions

Even though information about the three components of solar irradiance is extremely important for designing and assessing the performance of solar energy systems, measuring them is expensive, and requires significant operational resources and efforts. In order to obtain this information, a common practice is to use some sort of separation model to estimate both diffuse and direct components from global horizontal irradiance. Having this in mind, this study provides a separation model based on a logistical function derived using 1-min data that is reliable and accurate worldwide. To do so, we propose a new version of the BRL-min model, adding K_T as a new predictor. Also, instead of generating one set of parameters for a "universal" model, we propose that for each climate a different set of parameters should be used. Lastly, we introduce the use of a robust nonlinear regression method in order to adjust the coefficients to the irradiance data, reducing the effects of outliers on the dataset, while preserving all the already qualified data for each station. Other regression methods could have been used in this case, e.g. a support vector machine regression, which is also capable of handling outliers and could bring up promising results.

The results presented show that the proposed model performs equally well or better than the original BRL and Engerer models for most stations, figuring as a strong alternative for the separation models which can be considered as the best available today.

6. References

Aler, R., Galván, I. M., Ruiz-Arias, J. A., & Gueymard, C. A. (2017). Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting. Solar Energy, 150, 558–569. https://doi.org/10.1016/j.solener.2017.05.018

Boland, J., & Ridley, B. (2008). Models of diffuse solar fraction. Modeling Solar Radiation at the Earth's Surface: Recent Advances, (1999), 193–219. https://doi.org/10.1007/978-3-540-77455-6_8

Bright, J. M., & Gueymard, C. A. (2019a). Climate-specific and global validation of MODIS Aqua and Terra aerosol optical depth at 452 AERONET stations. Solar Energy, 183(January), 594–605. https://doi.org/10.1016/j.solener.2019.03.043

Bright, J. M., & Gueymard, C. A. (2019b). Climate-specific and global validation of MODIS Aqua and Terra aerosol optical depth at 452 AERONET stations. Solar Energy, 183(January), 594–605. https://doi.org/10.1016/j.solener.2019.03.043

Chen, D., & Chen, H. W. (2013). Using the Köppen classification to quantify climate variation and change: An example for 1901–2010. Environmental Development, 6(1), 69–79.

https://doi.org/10.1016/j.envdev.2013.03.007

Dumouchel, W., & O'Brien, F. (1991). Computing and Graphics in Statistics. In A. Buja & P. A. Tukey (Eds.) (pp. 41–48). New York, NY, USA: Springer-Verlag New York, Inc.

Engerer, N. A. (2015). Minute resolution estimates of the diffuse fraction of global irradiance for southeastern Australia. Solar Energy, 116, 215–237. https://doi.org/10.1016/j.solener.2015.04.012

Espinar, B., Ramírez, L., Drews, A., Beyer, H. G., Zarzalejo, L. F., Polo, J., & Martín, L. (2009). Analysis of different comparison parameters applied to solar radiation data from satellite and German radiometric stations. Solar Energy, 83(1), 118–125. https://doi.org/10.1016/j.solener.2008.07.009

Gueymard, C. A. (2014). A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. Renewable and Sustainable Energy Reviews, 39, 1024–1034. https://doi.org/10.1016/j.rser.2014.07.117

Gueymard, C. A. (2017a). Cloud and albedo enhancement impacts on solar irradiance using high-frequency measurements from thermopile and photodiode radiometers. Part 1: Impacts on global horizontal irradiance. Solar Energy, 1–14. https://doi.org/10.1016/j.solener.2017.05.004

Gueymard, C. A. (2017b). Cloud and albedo enhancement impacts on solar irradiance using high-frequency measurements from thermopile and photodiode radiometers. Part 2: Performance of separation and transposition models for global tilted irradiance. Solar Energy, 153, 766–779. https://doi.org/10.1016/j.solener.2017.04.068

Gueymard, C. A., & Ruiz-Arias, J. A. (2016). Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. Solar Energy, 128, 1–30. https://doi.org/10.1016/j.solener.2015.10.010

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. Communications in Statistics - Theory and Methods, 6(9), 813–827. https://doi.org/10.1080/03610927708827533

Ineichen, P. (2008). A broadband simplified version of the Solis clear sky model. Solar Energy, 82(8), 758–762. https://doi.org/10.1016/j.solener.2008.02.009

Lemos, L. F. L., Starke, A. R., Boland, J., Cardemil, J. M., Machado, R. D., & Colle, S. (2017). Assessment of solar radiation components in Brazil using the BRL model. Renewable Energy, 108, 569–580. https://doi.org/https://doi.org/10.1016/j.renene.2017.02.077

Liu, B. Y. H., & Jordan, R. C. (1960). The interrelationship and characteristic distribution of direct, diffuse and total solar radiation. Solar Energy, 4(3), 1–19. https://doi.org/10.1016/0038-092X(60)90062-1

MathWorks. (2018). MATLAB R2018a. Retrieved March 22, 2019, from https://www.mathworks.com/help/stats/fitnlm.html

Riazoshams, H., Midi, H., & Ghilagaber, G. (2018). Robust Nonlinear Regression: with Applications using R. John Wiley & Sons.

Ridley, B., Boland, J., & Lauret, P. (2010). Modelling of diffuse solar fraction with multiple predictors. Renewable Energy, 35(2), 478–483. https://doi.org/10.1016/j.renene.2009.07.018

Starke, A. R., Lemos, L. F. L., Boland, J., Cardemil, J. M., & Colle, S. (2018). Resolution of the cloud enhancement problem for one-minute diffuse radiation prediction. Renewable Energy, 125, 472–484. https://doi.org/10.1016/j.renene.2018.02.107

Younes, S., Claywell, R., & Muneer, T. (2005). Quality control of solar radiation data: Present status and proposed new approaches. Energy, 30(9 SPEC. ISS.), 1533–1549. https://doi.org/10.1016/j.energy.2004.04.031