Modeling the hourly solar diffuse fraction in Taiwan

Chia-Wei Kuo a, Wen-Chey Chang a, Keh-Chin Chang b,*

a Energy Research Center, National Cheng Kung University, Tainan, Taiwan
b Department of Aeronautics & Astronautics, National Cheng Kung University, Tainan, Taiwan

A R T I C L E   I N F O

Article history:
Received 25 March 2013
Accepted 28 November 2013
Available online

Keywords:
Diffuse fraction
Multiple regression
Model assessment
Mathematical modeling

Abstract

Using the data for global and diffuse radiation in Tainan, Taiwan, for the years of 2011 and 2012, respectively, four correlation models with five predictors: the hourly clearness index (k t), solar altitude, apparent solar time, daily clearness index and a measure of persistence of global radiation level, are constructed to relate the hourly diffuse fraction on a horizontal surface (d) to the clearness index. Two models use a single logistic equation for all k t values, Eqs. (6) and (7), and the other two models use a set of piece-wise linear equations for four k t intervals, Eqs. (8) and (9). The proposed models are compared respectively with the fourteen models available in the literature, in terms of the four statistical indicators: the mean bias error, the root-mean-square error, the t-statistic and the Bayesian Information Criterion, using the out-of-sample dataset for Tainan, Taiwan. It is concluded from the analysis that the proposed piece-wise linear models perform well in predicting the diffuse fraction, while the performances of the proposed logistic models are more case-dependent. Among those fourteen models considered in this study, the models developed by Erbs et al., Chandrasekaran and Kumar, and Boland et al. have competitive performances as the proposed piece-wise linear models do, when applying to the prediction of diffuse fraction in Tainan, Taiwan.

© 2013 Published by Elsevier Ltd.

1. Introduction

Global radiation (I global) consists of two parts: diffuse radiation (I diffuse) and beam radiation (I beam). Information about hourly diffuse fraction (d), defined in Eq. (1), is prerequisite to evaluate the performances of concentrating solar thermal systems. Since the measurements of diffuse or beam radiation are not frequently possible on a site of interest, it is necessary to find a model of diffuse fraction correlating the diffuse radiation to the global radiation which is usually available in the reports from the meteorological stations.

\[ d = \frac{I_{\text{diffuse}}}{I_{\text{global}}} \]  (1)

An early model developed by Liu and Jordan [1] estimates d in terms of the hourly sky clearness index (k t), which is an indicator to measure how clear the skies are, given by

\[ k_t = \frac{I_{\text{global}}}{H_o} \]  (2)

where H o is the hourly extraterrestrial radiation and can be theoretically determined by specifying the site latitude, the day of each year and the hour of each day [2]. The idea of such early model is rather simple: when k t is large (clear sky), diffuse fraction is small due to less obstruction by small droplets and particulates suspended in the atmosphere; when k t is small (cloudy sky), diffuse fraction is large because of the greater scattering frequency. Following the same idea, quite a few other models in terms of the single predictor k t [3–12] were developed later, each of which successfully fit the data of diffuse fraction within a specified region, to varying degrees.

Studies that include multiple predictors in modeling to achieve a better data fit were also conducted. For example, Reindl et al. [13] suggested that the solar altitude (a), ambient temperature (T a) and relative humidity (RH) are the other three effective ones, by performing a statistical analysis of twenty-eight possible predictors, using the data from cities worldwide.

Recently, Boland and his coworkers [14,15] demonstrated that the performance of the time-dependent model was better than the other and corroborated the significance of time (apparent solar
time) as an extra model predictor. On the basis of these investigations, they recently proposed the Boland—Ridley—Lauret model (abbreviated as the BRL model hereafter) [16] using a total of five predictors: \( k_t, a, \) the apparent solar time \((t)\), the daily clearness index \((\psi)\) and a measure of persistence of global radiation level \((\psi)\). The last two predictors are defined as follows.

\[
K_T = \frac{\sum_{i=1}^{24} f_{\text{global}}}{\sum_{i=1}^{24} H_0}
\]

\[
\psi = \begin{cases} 
\frac{k_{t+1} + k_{t-1}}{2} & \text{for sunrise} < t < \text{sunset} \\
k_{t+1} & \text{for } t = \text{sunset} \\
k_{t-1} & \text{for } t = \text{sunrise}
\end{cases}
\]

Despite the abundance of models, such as those summarized in Tables 1 and 2, there is still a need for developing correlation models for the Taiwan area. According to the comparative studies of model performance made by Jacobsides et al. [11], Torres et al. [17] and Dervishi and Mahdavi [18], most of the existing correlation models are not applicable to all geographical regions, without modification. Therefore, the developments of applicable but accurate correlation models for a specific region, which account for the geographical and climatic conditions, are still preferable.

This study correlates four models with two different mathematical formats: logistic and piecewise linear equations, using the two available yearly sets of data in Tainan, Taiwan. The logistic models, which are based on the recent study of Ridley et al. [16], use a single logistic equation for all values of \( k_t \). In contrast, the other relatively simple models use a set of piece-wise linear equations for different intervals of \( k_t \) that use the same group of predictors as the logistic models. The models listed in Table 1 are also tested using the available database in Taiwan, for the purpose of comparison with the proposed models developed in this study.

### 2. Experimental set-up and database

Because of the lack of diffuse radiation data in daily reports from all meteorological stations of the Taiwan Central Weather Bureau, this modeling work uses the in situ data for global and diffuse radiation, measured at the Kuei-Jen campus of the National Cheng Kung University, Tainan, Taiwan (23° N 120° E), from 1 January 2011 to 31 December 2012.

Two sets of devices from Eppley Laboratory, Inc., each of which included a pyranometer (Model PSP) without and with a shadow band stand (Model SBS), were used to measure global and diffuse radiation, as shown in Fig. 1(a) and (b), respectively. The sampling rate was of 1 Hz. The method of data checking followed partially the ideas of Reindl et al. [13], to identify the missing data and data which violated physical limits. The missing data for individual seconds mostly occurred while transmitting were filled using a linear interpolation of the neighboring data in the time sequence. After all of the missing data had been filled, the data for individual seconds were converted into an hourly value by integration with respect to time.
Any hourly data that violated a physical limit, including those with negative values for radiation or with a value for global radiation that exceeded the extraterrestrial radiation, were then eliminated from the dataset. The final datasets that had passed the quality control checks mentioned produced 3621 and 3382 hourly data points, for the yearly databases of 2011 and 2012, respectively.

3. Model development

It is noted [16] that the BRL model using a single logistic equation for the entire range of $k_t$ instead of a set of piecewise linear equations, could evaluate the diffuse fraction satisfactorily. Thus, the approach of the BRL model is firstly used for the modeling work. As expressed in Eq. (5), the other predictors used in the BRL model include solar altitude ($\alpha$, in rad), apparent solar time ($t$, in h), daily clearness index ($K_t$, defined in Eq. (3)) and a measure of the persistence of the global radiation level ($\psi$, defined in Eq. (4)).

$$d = 1/[1 + \exp(\gamma_1 + \gamma_2k_t + \gamma_3\alpha + \gamma_4t + \gamma_5K_t + \gamma_6\psi)]$$  \hspace{1cm} (5)

where $\gamma_i$ are coefficients to be determined. Using the hourly $d-k_t$ databases collected in 2011 and 2012, respectively, together with four other corresponding predictors, and the coefficients of the BRL model (see Table 1) as the initial guessed values, the optimal forms of the modified BRL models for the Taiwan area are given as

Model 1 (with fitting dataset of 2011)

$$d = 1/[1 + \exp(-4.5274 + 5.6956k_t - 0.0814\alpha - 0.0464t + 2.4162K_t + 1.0125\psi)]$$  \hspace{1cm} (6)

Model 2 (with fitting dataset of 2012)

$$d = 1/[1 + \exp(-4.5312 + 5.7627k_t - 0.0882\alpha - 0.0391t + 1.9998K_t + 1.1521\psi)]$$  \hspace{1cm} (7)

As seen from the models compiled in Table 1, most of the existing models split the data into sub-regions defined for different $k_t$ values, before any regression is attempted. Two relatively simple piece-wise linear models that use the same five predictors as the previous modified BRL models are suggested as Eqs. (8) and (9). The models use four sub-regions, instead of the two or three sub-regions commonly used in the existing piece-wise models (Table 1), in order to produce a better fit for the present database. Note that the variable form of “$\sin(\alpha)$” instead of “$\alpha$” used in Eqs. (6) and (7), is used in Eqs. (8) and (9). This follows the method used in the model of Reindl et al. [13].

Model 3 (with fitting dataset of 2011)

![Fig. 1. (a) The pyranometer and (b) the pyranometer + shadow band stand used to measure global and diffuse radiation, respectively.](image)
4. Assessment of model performance

Fig. 2(a)–(d) shows graphical comparisons between the predicted diffuse fraction and the measured data versus the sky clearness index. The comparisons reveal that the predictions made with Eqs. (6)–(9) match the data well. Three commonly used statistical indicators: the mean bias error (MBE), the root-mean-square error (RMSE) and the t-statistic defined in Eqs. (10)–(12), respectively, are firstly used to evaluate the performances of the proposed models against those summarized in Table 1, except for the model of Reindl et al. which uses four predictors \((k_t, a, T_a, RH)\), due to lack of the ambient temperature \((T_a)\) and relative humidity \((RH)\) data in the present measurements. In other words, a total of 14 models from Table 1 are compared here.

Model 4 (with fitting dataset of 2012)

\[
\begin{align*}
    d &= 0.9896, \quad 0 \leq k_t < 0.2 \\
    d &= 1.0874 - 0.3936k_t + 0.0359 \sin(a) + 0.0035T_a - 0.1899T - 0.1253, \quad 0.2 \leq k_t < 0.3 \\
    d &= 1.4188 - 1.2191k_t + 0.0150 \sin(a) + 0.0063T_a - 0.3403T - 0.2125, \quad 0.3 \leq k_t < 0.75 \\
    d &= 0.2775, \quad k_t \geq 0.75
\end{align*}
\]
The tests for MBE and RMSE respectively provide information about the long-term and short-term performance of any. In general, the smaller the absolute value of the MBE and the RMSE are, the better the model performs. However, there possibly appear conflicting results such as simultaneously having a large value of RMSE and a small value of MBE, or vice versa. Thus, the t-statistic is introduced to give an additional indicator to measure the statistical significance of a model at a specified confidence level [19]. To fulfill this purpose, the critical t value (t_{crit}) at a particular level of significance n−1 degrees of freedom must be determined, which can be obtained from standard statistical tables. The greater the degree to which t_{stat} is smaller than t_{crit}, the better is the model's performance.

Next, to study the trade-off between the goodness of fit and the complexity of the models, the Bayesian Information Criterion (BIC), which takes the fitting errors and the number of parameters (k) into account, is calculated using the following equation, as with the studies of Ridley et al. [16] and Torres et al. [17]. When comparing the performance of the model, a lower value for the BIC indicates a better model.

\[
\text{BIC} = n \cdot \ln \left( \frac{1}{n} \sum_{i-1}^{n} (d_{\text{est}} - d_{\text{mea}})^2 \right) + k \cdot \ln(n)
\]

To evaluate the performances of the four proposed models, Eqs. (6)–(9), the out-of-sample data are used. In other words, Models 1 and 3, which were constructed by using the training dataset of 2011, use the dataset of 2012 as the out-of-sample data for their model assessments, and vice versa for Models 2 and 4. Tables 3 and 4 compare the statistical performances and the BIC values of the proposed models of two different formats, i.e. Models 1 and 3 as well as Models 2 and 4, against the 14 existing models considered in Table 1, using the collected dataset of 2011 and 2012, respectively, in Fig. 2.

As seen from Tables 3 and 4, the proposed piece-wise linear models (Models 3 and 4) perform well in terms of the value of MBE, RMSE, and BIC among all the models considered, while their t_{stat} values exceed the t_{crit} value slightly. In contrast, the modified BRL models (Models 1 and 2) show more case-dependent performances, particularly, for the evaluation of t_{stat}, as compared to the proposed piece-wise linear models. Model 2 performs slightly better than Model 4 by using the dataset of 2011 for the out-of-sample data, whereas Model 1 performs worse than Model 3, particularly with respect to the t_{stat}, by using the dataset of 2012 for the out-of-sample data. Another better performing models concluded from the comparison results in Tables 3 and 4 include two piece-wise higher-order polynomial models with a single \( k_i \) predictor, respectively developed by Erbs et al. [4] as well as Chandrasekaran and Kumar [9], and one time-independent logistic model with a single \( k_i \) predictor, developed by Boland et al. [15].

It is agreed that, to avoid the use of the extreme case of weather conditions, the best choice of the dataset for modeling work is the one based on a typical meteorological year, which is constructed of each representative month in a year over a long period (usually 10 or 15 years) weather records. However, due to lack of available long-period database in Tainan, Taiwan, it is hard to make a definite assessment of the model performance at the present time.

As noted from Table 2, the model of Boland et al. was constructed with data from cities worldwide and the model of Erbs et al. was developed with data from cities in USA. In contrast, the model of Chandrasekaran and Kumar used the data from a single city (Madras, India) which is the same situation as for development of the present models (Tainan, Taiwan). Thus, the models of Erbs et al. and Boland et al. possess more potential to be applied to other places in the world in comparison with the presently proposed models.

### 5. Conclusions

Four correlation models with five predictors: \( k_0, a, t, K_f \) and \( \psi \), as suggested in the BRL model of Ridley et al. [16], are constructed.
respectively by using two different yearly datasets to relate the hourly diffuse fraction on a horizontal surface to the clearness index for the Taiwan area. Two models, Eqs. (6) and (7), which follow the mathematical format of the BRL model, use a single logistic equation for all values of $k_0$, and the other two models, Eqs. (8) and (9), use a set of piece-wise linear equations for four intervals of $k$. The proposed models are compared respectively with the fourteen models available in the literature using the four widely used statistical indicators: the MBE, the RMSE, the $t$-statistic and the BIC, using the out-of-sample databases for Tainan, Taiwan. From the analyses, it is found that the proposed piece-wise linear models perform well in diffuse-fraction predictions in Tainan, Taiwan, while the proposed logistic (or modified BRL) models show more case-dependent performance. Another better performing models among the others considered in Table 1, include two piece-wise higher-order polynomial models respectively developed by Erbs et al. [4] as well as Chandrasekaran and Kumar [9], and one time-independent logistic model developed by Boland et al. [15], when applying to the estimation of diffuse fraction in the Taiwan area.

**Acknowledgments**

This work was financially supported by the Bureau of Energy, the Ministry of Economic Affairs, Taiwan under Grant No. 102-D0304.

**Nomenclature**

- $d$: hourly horizontal diffuse fraction
- $d_{est}$: estimated value of $d$
- $d_{mea}$: measured data of $d$
- $H_0$: extraterrestrial radiation (kJ/h m$^2$)
- $I_{beam}$: horizontal beam radiation (kJ/h m$^2$)
- $I_{diffuse}$: horizontal diffuse radiation (kJ/h m$^2$)
- $I_{global}$: horizontal global radiation (kJ/h m$^2$)
- $K_T$: daily sky clearness index
- $k$: number of coefficients to be estimated in the BIC test
- $k_1$: hourly sky clearness index
- $n$: number of data points
- RH: relative humidity
- $T_a$: ambient temperature (°C)
- $t$: apparent solar time (h)
- $t_{crit}$: critical value of $t$-stat.
- $t_{stat}$: value of the $t$-statistic
- $\alpha$: solar altitude (degree)
- $\psi$: measure of persistence of global radiation level

**References**