Renewable Energy 165 (2021) 668–677

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

Clustering of visible and infrared solar irradiance for solar architecture design and analysis

Qiuhua Duan , Yanxiao Feng , Julian Wang *

Department of Architectural Engineering, Pennsylvania State University, 108 Engineering Unit A, University Park, 16802, PA, United States

ARTICLE INFO

Article history:
Received 10 June 2020
Received in revised form
9 September 2020
Accepted 14 November 2020
Available online 17 November 2020

Keywords:
Solar radiation
Solar architecture design
Classification trees
Prediction model
Solar energy
Building energy

ABSTRACT

Incoming solar radiation is a key factor influencing solar architecture design. It determines the thermal and optical regime of the building envelope and affects the solar heat and light transfer between the indoors and outdoors. Computational analysis is an essential tool in solar architecture design. Usually, an entire year's weather data in a conventional weather file can be imported into such computational analyses. Solar irradiance data used in a conventional solar architecture design analytics are broadband (the total of UV, VIS, and NIR); however, these three components play different roles in building energy efficiency. So, analyzing individual solar components separately can be desirable. This research is to develop estimation models of the VIS and NIR components that can be captured efficiently from readily available datasets collected from the ground weather stations; such a model can then be conveniently implemented into current solar architecture design and research. We explored and tested classification-based modeling methods for decomposing hourly broadband global horizontal solar irradiance data in conventional weather files into hourly global horizontal solar VIS and NIR components. Furthermore, a workflow of how to implement these models in solar architecture design and analysis has been developed and discussed herein.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Solar architecture is an architectural approach that makes the best possible use of locally available solar energy by employing both passive and active measures [1]. The first solar architecture in America was proposed by Tod Neubauer in the 1950s [2]. Research in this field has addressed the theoretical background, simulation techniques, and experimental testing. Computational analysis in solar architecture design has been described and discussed widely in recent decades [3]. Usually an entire year's weather data are imported in a conventional format (e.g., TMY, WYEC2 BLAST) into an energy simulation program to calculate the energy consumption of a building. Solar irradiance data in a complete weather file also include global horizontal irradiation (GHI), diffuse horizontal irradiation (DHI), and direct normal irradiation (DNI). Regardless of the three solar irradiance types noted above, the solar irradiance data are broadband and represent the total of ultraviolet (UV), visible light (VIS), and near-infrared radiation (NIR), three components of the solar spectrum.

With two known solar data variables, the other variable can be calculated via the mathematical relations among them. However, these three components play different roles in solar architecture design. Of these three major components, VIS always provides benefits to indoor building energy savings (e.g., electrical lighting), while solar NIR is beneficial to building energy savings in winter but undesirable in summer [4]. Similarly, the COVID-19 pandemic has heightened interest in the solar UV component and its potential impact on the spread and seasonality of disease. Therefore, in some in-depth building environment performance analyses, especially building energy simulation work, separate analyses focusing on each solar radiation component are desirable. With recent discoveries and engineering solutions emerging related to nanomaterials and nanostructures, independent band modulation of solar radiation on building envelopes (including glazing systems) has become increasingly viable as a potential means of improving building energy savings and indoor visual comfort. However, the meteorological data in conventional weather files do not normally include the spectral power distribution data of incident solar light because measuring the narrowband spectral distribution of sunlight is much more difficult and expensive than measuring broadband

* Corresponding author.
E-mail addresses: qfd5026@psu.edu (Q. Duan), yxf5136@psu.edu (Y. Feng), julian.wang@psu.edu (J. Wang).

https://doi.org/10.1016/j.renene.2020.11.080
0960-1481/© 2020 Elsevier Ltd. All rights reserved.

radiation (e.g., using pyranometers). As a consequence, there is a pressing need for reliable performance estimations of spectral solar radiation control and response on a building scale. To assess this, we need solar spectral irradiance source data, or at least band (i.e., VIS and NIR) solar irradiance data as input.

To address this research gap and the practical need for solar architecture design, this work has developed an estimation model for VIS and NIR components that can be captured efficiently from readily available datasets without the addition of new measurements and associated sensors; this can then be conveniently implemented into current solar architecture design and research. In particular, several research questions have been answered, including how to yield a reliable model with readily available weather data such as dew point temperature, relative humidity, and broadband solar irradiance information; how the first principles of solar radiation and building physics domains should be combined, and what new (readily available) parameters can be incorporated to facilitate the modeling procedure; and how to simplify the manipulation process for building energy modelers to apply new models in energy simulations. In this research, we explored and tested classification-based modeling methods for decomposing hourly broadband global horizontal solar irradiance data in conventional weather files into hourly global horizontal solar VIS and NIR components, yielding two accurate models of the VIS and NIR fractions of the overall solar irradiance (or GHI). Furthermore, a data conversion workflow of how to implement this in solar architecture design and analysis processes was developed and is described herein. The methodology established in this work presents a new, efficient, and accurate method based on readily available weather data documented in conventional weather files, enabling more comprehensive and precise building energy and performance-related analyses, especially with respect to building elements and products that have features of spectral selectivity. The uniqueness of this model is that the model development in this work is targeting the solar system application or passive solar strategies in building engineering and energy efficiency, so the solar spectral bands defined in this work are consistent with the requirements in solar architecture design and analysis. Another novelty is that we only used the most basic meteorological elements, such as humidity, temperature, etc., which are normally documented in ground weather stations, combined with several new parameters based on solar radiation physics. This could support an ease-of-manipulation for building simulation by using conventional weather files in the architecture, construction, and engineering industry.

2. Related work

Different spectral irradiance models have been proposed since the 1940s. Moon's spectral radiation curve [5], Leckner's model [6], Brine and Iqbal's model [7], and SOLAR2000 [8] are empirical models based on an understanding of solar spectral irradiance combined with historically measured weather and other solar irradiance data. The BRITE and FLASH [9], LOWTRAN 7 [10], MODTRAN 6 [11], SEA [12], and SOLMOD models [13] consider the physical characteristics of the atmosphere and use references or measured vertical profiles of gaseous and aerosol constituents; they are typically rigorous and sophisticated codes. The National Renewable Energy Laboratory (NREL) provides the Bird Simple Spectral Model (SPCTRAL2) [14] and the SMARTS model [15] that simplify the atmosphere's vertical profile and facilitate solar technology integration. Reconstruction models usually model solar spectral irradiance variability by a linear combination of indicators of solar activity [16]. Although these spectral irradiance models are available and effective for the estimation of detailed spectra, the approaches and resultant models are not suitable for building energy efficiency analysis due to wavelength range limitations, the need for additional measurement and data input, implementation complexities, etc. [17].

A relatively simpler method of integrating solar spectra into application areas is to develop models of major solar spectral components such as solar UV, VIS, and NIR irradiance. Most previous studies on this topic have determined simple representative fractions for VIS and NIR. Comparatively, NIR/GHI has been less frequently investigated than VIS/GHI. The NIR/GHI fraction was reportedly around 46.5% in Brazil and 51.8% on the Tibetan plateau [18,19]. Some studies have argued that these fractions could vary significantly in different weather and atmospheric situations. For instance, Szeicz verified that ratio of the visible energy to the total received by the photosynthetically active part of the spectrum 0.5 is a better general approximation according to a theoretical model and an experiment, his study indicated the VIS/GHI fraction is closely associated with two factors: the presence of clouds and scattering caused by aerosol [20]. The NIR/GHI fraction is closely related to the total amount of column water vapor [21]. Few studies have attempted to developed regression models of these fractions. The most representative work was done by Escobedo et al., who established monthly and hourly fraction models for the UV, VIS, and NIR solar components in Brazil [21]. In that work, they have found that the clearness index (ratio of the global-to-extraterrestrial solar radiation) of sky conditions can be a determinant factor to develop the simple linear regression models for the hourly and daily factions of UV and GHI [21]. Comparatively, the linear regression models derived to estimate the NIR and VIS components may be obtained without sky condition conditions. However, it is worth mentioning that the developed linear regression models in that work were based on the specific variations or features of the sky condition in the selected site, which was with a maximum variation of 8%. In other words, the models may not be effective for other situations with larger sky variations. Another characteristic research done by Charuchittipan et al. was to estimate the diffuse NIR radiation from satellite- and ground-based data including atmospherical reflectivity, precipitable water, relative humidity, and air temperature. This semi-empirical model is in reasonable agreement with independent diffuse NIR data, giving an RMSD and MBD of 16.7% and 1.5%, respectively [22]. The satellite data are necessary for the estimation in this model, which seems suitable to the mapping application purposes, presenting NIR data on satellite images. However, such satellite data are not typically available or utilized in the domain of solar buildings. Similar regression modeling works using ground and/or satellite measurements were also conducted by other researchers [23–30], but most of these works focus on the VIS part of solar radiation and the agriculture applications.

In summary, in these prior studies, we understand that various atmospheric variables including clearness index, water vapor pressure, ozone column, aerosol optical depth, air relative humidity, etc. may significantly affect these spectral components; however, they are not always available in typical weather files complied from the measurements in ground weather stations. Meanwhile, it is consistent among the above works that the sky clearness index plays a very significant role in classifying the UV, VIS, and NIR solar components, providing a valuable foundation for our work. As mentioned above, the objectives of this work are different from these previous studies in two key aspects. First, the model development in this work is targeting the solar system application or passive solar strategies in building engineering and energy efficiency, so that the spectral band coverages are not exactly the same with the ones in the previous works focusing on agriculture, forestry, oceanography, or general atmospherical studies. Second,

this work aims to achieve an ease-of-manipulation for building simulation purposes by using basic meteorological elements collected in typical ground weather stations and used in conventional weather files in the architecture, construction, and engineering industry. Such concerns have to date not been addressed.

3. Methodology

Fig. 1 shows the research framework and workflow of this study. It illustrates that we first built a precise estimation modeling of VIS and NIR components from hourly global solar radiation and hourly meteorological parameters. This procedure consisted of five major steps from data collection to processing, cleaning, classification, regression trees (CART) technique application for modeling, and model validation. After validation of the developed solar spectral models, we proposed and designed a workflow to introduce how to incorporate the models into solar architecture design and analysis. Next, we will provide the details of each step shown in this diagram.

3.1. Data collection

Two major datasets, hourly meteorological measurements (HMM) and outdoor solar spectra data (WISER), in the location (Latitude: 39.742° North, Longitude: 105.18° West, Elevation: 1828.8 m AMSL) were selected from the SRRL BMS database of the NREL Solar Radiation Research Laboratory for the modeling done in this study [31]. The HMM dataset was used to retrieve and process the independent variables, including GHI, DNI, DHI, cloud coverage, dry-bulb temperature, dewpoint, relative humidity, and wind speed, while the key dependent variables (i.e., solar VIS and NIR irradiance) were calculated from the WISER dataset [31].

The HMM dataset for 2018 and 2019 was used in this project. It describes the basic solar radiation and meteorological elements with hourly timestamps, which has identical variable types and formats with the TMY weather file. Building upon this TMY format-compliant dataset enables us to perform the conversion from the hourly broadband solar irradiance to spectral components based on typical weather files in the future. Note that the average value of all measured points each hour is defined as the value for the timestamp at the end of the 1-h interval [32]. For example, the value at timestamp 08:00 in the HMM dataset equals the average value of all measurements taken from 07:00 to 08:00. This dataset is well-organized and has been used widely to simulate the solar radiation and building energy performance in the architecture, engineering, and construction industries.

The WISER measurement database is formed from two spectroradiometers (i.e., MS-711 and MS-712) that are combined to measure global horizontal spectral solar irradiance data [31]. MS-711 covers the measurement range from 300 nm to 1100 nm, and MS-712 focuses on the NIR range from 900 nm to 1700 nm [31]. We selected data from the same period: 2018 and 2019. The WISER database has a higher resolution measurement for both wavelengths (0.41 nm and 1.6 nm resolutions for the MS-711 and MS-712, respectively) and time intervals (typically 5 min, but occasionally 1 min). To coordinate these two solar datasets from different sources, the 5-min interval data were processed using R software. The hourly spectrum data were calculated by averaging the 5-min interval data for each hour, following the criterion of timestamp calculation regulated in the HMM dataset. The day-of-year time format was also modified to fit the time format of UTC (Coordinated Universal Time), as it was the same format used in the HMM.

3.2. Data processing

First, to obtain the solar VIS and NIR components, we summed the spectral data for the corresponding wavelength ranges of 380 nm–780 nm and 781 nm to 1700 nm for VIS and NIR, respectively, based on the International Standards Organization's spectral band definitions [33] and the spectroradiometer measurement ranges in this work. We obtained the fractions of *VIS*/*GHI* and *NIR*/*GHI* by using the VIS and NIR values calculated from the WISER dataset and GHI values calculated from the HMM dataset.

Second, to potentially enhance modeling accuracy, we generated several additional predictor variables. The primary principles applied when adding these predictors were obtained from the knowledge and theory of solar radiation and building physics, with a focus on calculations that did not require new sensors and measurements and demanded a minimum amount of computation.

1) Extraterrestrial solar radiation I_0

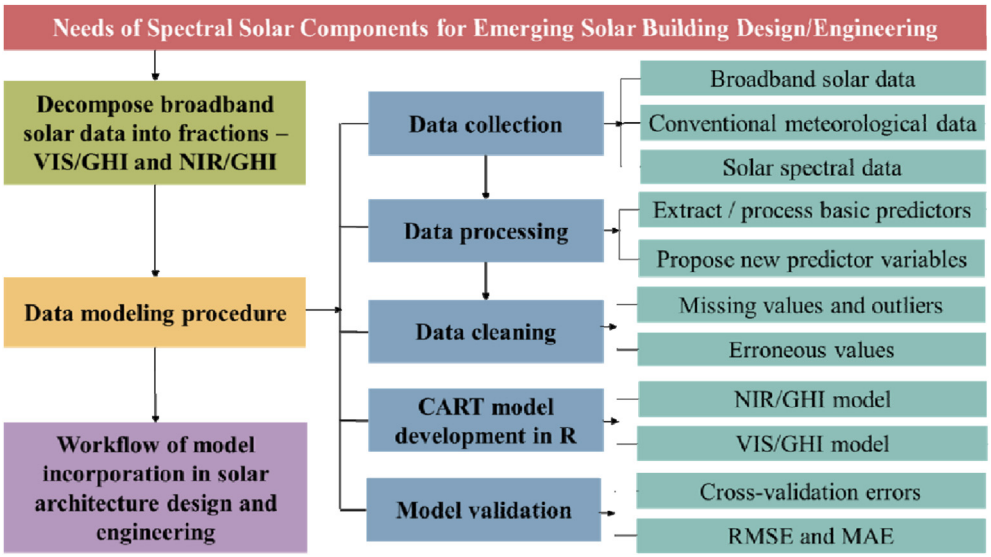

Fig. 1. Research framework and workflow.

The hourly average extraterrestrial solar radiation on the horizontal surface I_0 is determined using the following Equation [34].

$$I_0 = I_{sc}(R_{av}/R)^2 \tag{1}$$

where I_{sc} is a solar constant (1367 W/m^2), R_{av} is the mean Sun-Earth distance, and R is the actual Sun-Earth distance depending on the day of the year. An approximate equation for the effect of the Sun-Earth distance is:

$$(R_{av}/R)^2 = 1.00011 + 0.034221\cos(\beta) + 0.001280\sin(\beta) + 0.000719\cos(2\beta) + 0.000077\sin(2\beta)$$

where $\beta = 2\pi n/365$ radians and n is the day of the year.

2) Solar zenith angle ζ

The solar zenith angle ζ is the angle between the solar and the vertical. We used AstroCalc4R, developed by Jacobson et al. in R statistical software, to calculate the solar zenith angles based on latitude, longitude, time of day, and date [35].

3) Clearness index K_t

The clearness index K_t is the ratio of the horizontal global irradiance to the corresponding irradiance available outside the atmosphere. It may be considered an attenuation factor of the atmosphere and can be calculated by the following Equation [36].

$$K_t = \frac{GHI}{I_0\cos(\zeta)} \tag{2}$$

where GHI is the horizontal global irradiance, I_0 is extraterrestrial solar radiation on the horizontal surface, and ζ is the solar zenith angle.

4) Cloud transmittance T_{cld}

We formed a new parametric cloud transmittance T_{cld} based on our understanding of the physical behavior of solar irradiance transmission. T_{cld}, defined as:

$$T_{cld} = \frac{(1-0.1T_{opq})(1-0.1T_{tot}+0.1T_{opq})}{1-0.05T_{tot}} = \frac{(1-0.1T_{opq})(1-0.1T_{trn})}{1-0.05T_{tot}} \tag{3}$$

where T_{opq} is the opaque sky cover transmittance, T_{tot} is the total sky cover transmittance, and T_{trn} is the translucent sky cover transmittance $T_{trn} = T_{tot} - T_{opq}$.

5) Air mass AM

The relative air mass AM was given by Kasten as [14].

$$AM = \frac{1}{\cos(\zeta) + 0.15(93.885-\zeta)^{-1.253}} \tag{4}$$

3.3. Data cleaning

After building up the datasets, including the original date, calculated data, and additional data described above, the quality of the raw data was enhanced by a data cleaning process that filtered it for any uncertainties or errors [37]. The following requirements were taken into account for quality control of the datasets:

1) If there were missing data regarding global horizontal solar radiation, diffuse horizontal solar radiation, pressure, relative humidity, or dew and/or dry bulb temperatures, such data for that hour were omitted.
2) If the ratios of NIR/GHI or VIS/GHI were greater than 1, the corresponding data entries were omitted.
3) The clearness index K_t was calculated and if the solar zenith ζ was greater than 85.5°, the corresponding data were disregarded [38].
4) If the GHI was smaller than 50 W/m^2, the corresponding data were disregarded.

After the above data cleaning process, the finalized dataset included 7583 observations.

3.4. Classification method

Classification and regression trees (i.e., CART) are a simple but powerful technique for modeling. Unlike the generalized linear regression model (GLM) that typically pre-specifies and tests the relationship between the response and predictor, CART does not develop a prediction relationship. It constructs a set of decision rules for the predictor variables [39]. The data are partitioned along the predictor axes into subsets with homogeneous values for the dependent variable. The best split is chosen for all of the predictors by an exhaustive search procedure. An analysis of variance (ANOVA) is conducted to select the splits, which maximizes the homogeneity of the two resulting groups with respect to the response variable. The output is a tree diagram with the branches determined by the splitting rules and a series of terminal nodes that contain the mean response. The procedure initially grows full trees and then uses a cross-validation process to prune the overfitted tree to an optimal size [40]. CART modeling has several disadvantages compared to conventional regression modeling, including it being very close to a simple linear relationship when the size of the tree is small; also, the predictions are unstable due to high variance single regression trees. That is, small changes in data can produce substantially different trees [41]. However, CART analysis also has clear advantages over classical statistical methods, effectively uncovering structures in data with hierarchical or nonadditive variables. CART also provides the possibility of interactions and nonlinearities among variables and has been found to be very interpretable. It also makes it easy to understand a variable's importance in making predictions, and is quick to use because there are no complicated calculations [42]. Such methods have been useful in solar radiation modeling applications, including both prediction and estimation [43–45].

3.5. Tree selection

CART uses a technique known as binary recursive partitioning and outputs four indicators: the complexity parameter *cp*, relative error *rel error*, cross-validation error *xerror*, and standard error *xstd*. The *rel error* is the ratio of the sum of the squared differences of the observed and predicted values and the original variance. The indicator *rel error* is the observations, while *xerror* and *xstd* are errors from cross-validation of the data [39,46]. The indicator *xerror* is related to the predictive residual sum of squares (PRESS) statistic. If it is assumed that the dataset is partitioned into i regions (R_i), the actual response is y_i and the predicted constant is c_i, so the residual sum of square error SSE of the subtrees can be expressed as:

$$SSE = \sum_{i \in R_i} (y_i - c_i)^2 \tag{5}$$

The *xerror* indicator is the *SSE* from the cross-validation data. The *cp* indicator is illustrated below.

Employing these indicators, two methods are normally used to assess and select the tree structure and avoid overfitting the data.

1) Minimal cost complexity method

In a simple ANOVA, the objective at each node is to minimize the *SSE* or maximize the between-group sum-of-squares. However, minimizing the *SSE* is not a good measure for selecting a subtree because it always prefers a bigger tree. The complexity parameter *cp*, tree size *T*, and *SSE* of the tree with no splits ($SSE(T_1)$) is then used as a penalty term to measure the cost complexity of the tree; the objective function for pruning the tree is shown in Equation (6). If $cp = 0$, then the biggest tree will be chosen because the complexity penalty term is essentially dropped. As *cp* approaches infinity, the Size 1 tree will be selected.

$$minimize\{SSE(T) + cp \cdot |T| \cdot SSE(T_1)\} \tag{6}$$

The optimal size of the tree is the fewest branches that still minimize all errors. Typically, we evaluate multiple models across a spectrum of *cp* and use cross-validation to identify the optimal size, and thus the optimal subtree that best generalizes to the data. If the cost of adding another variable to the decision tree of the current node is above the threshold, then tree building does not continue and the threshold of complexity parameter *cp* is reported.

2) One-standard-error (1-SE) rule

An alternative rule for post-pruning the tree model is the one-standard-error (1-SE) rule. The 1-SE rule is based on cross-validation estimates of the error of the subtrees in the initially grown tree, together with the *xstd* of these estimates. This uses the first level where the *xerror* falls into the ±1 *xstd* range of min (*xerror*) that is calculated based on a defined *cp* (e.g., 0.01), which is expressed as follows:

$$xerror \leq \min(xerror) + xstd$$

The level at which the *xerror* is at or below horizontal is displayed as a red dotted line in the cross-validation error plots. Then, the simplest model (i.e., the smallest tree size) is chosen. This method takes into account the variability of *xerror* resulting from cross-validation because in most practices, the plot of *xerror* has an initially sharp drop, followed by a relatively flat plateau and then a slow rise. In other words, the minimum cross-validation error rate is no guarantee that the cross-validation error is a random quantity.

4. Results and discussion

In this study, we used the *rpart* package in R software to build regression trees for *VIS/GHI* and *NIR/GHI*. We split the entire dataset *D* into a training dataset (90% of *D*) and a test dataset (10% of *D*). The *rpart* implementation first fit a fully grown tree onto the training dataset with *N* terminal nodes. Then, it pruned the fully grown tree by *k*-fold cross-validation (default $k = 10$).

4.1. CART results for the VIS/GHI fraction

1) Cross-validation error plot

Fig. 2 shows the cross-validation error plot for the *VIS/GHI* tree. The vertical axis represents the relative cross-validation *SSE*(*xerror*). From this figure, we can see that when $cp = 0.014$, the Size 7 regression tree has the minimum cross-validation error. This tree model is shown in Fig. 3. The red dotted line in Fig. 2 refers to where the cross-validation error is just smaller than the sum of the minimum cross-validation relative estimates error *xerror* and the cross-validation standard error *xstd* at that tree (i.e., the 1-SE rule).

The CART procedure generated a regression tree with a minimum cross-validation error containing seven terminal nodes for *VIS/GHI* (see Fig. 3). The fraction of *VIS/GHI* ranged from 0.287 to 0.609 in these seven groups, among which the fraction 0.547 represented the major (33.5%) training observations that belonged to that node. Comparatively, the visible solar radiation occupies about 49% (or 0.49) extraterrestrial solar radiation. So, in other words, the fraction, *VIS/GHI*, should tend to 0.49 when the sky condition is clear. On the contrary, under the cloudy situation, water vapor in the Earth's atmosphere may significantly absorb the sunlight, especially in the infrared region, which leads to a relatively higher fraction of visible solar radiation. The tree result in Fig. 3 confirms these physical explanations and shows that the first variable selected for splitting was the clearness index K_t. If $K_t < 0.415$, the group was further split according to *RH* and *Dew*. In the other major branch of this regression tree, if $K_t \geq 0.415$, the parameters of *RH* and *Dew* were also used to form the groups further. The fractions in the low K_t regions were relatively larger than the ones in the high K_t regions. In the meanwhile, both *RH* and *Dew* were found significant to determine the terminal tree nodes in the two major branches, which also complies with the fact that the content of water in the air plays an essential role to affect the fractions of. *VIS/GHI*.

The red dotted line in Fig. 2 represents the highest cross-validation error minus the minimum cross-validation relative estimates error *xerror*, plus the cross-validation standard error *xstd* at that tree (via the 1-SE rule). A reasonable choice of *cp* for pruning is often the leftmost value, where the mean is less than the horizontal line. As shown in Fig. 4, in this case, the optimal size of the tree contained only three terminal nodes for *VIS/GHI*. The percentage of *VIS/GHI* ranged from 26.3% to 40.2% in these three groups. The first variable selected for splitting was the clearness index K_t. If $K_t < 0.415$, no further split was observed for Group 1: 26.3% of *VIS/GHI* data, with a mean value of 0.567. If $K_t \geq 0.415$, the group was further split according to $Dew < -1.05°C$ (Group 2: 40.2% of *VIS/GHI*

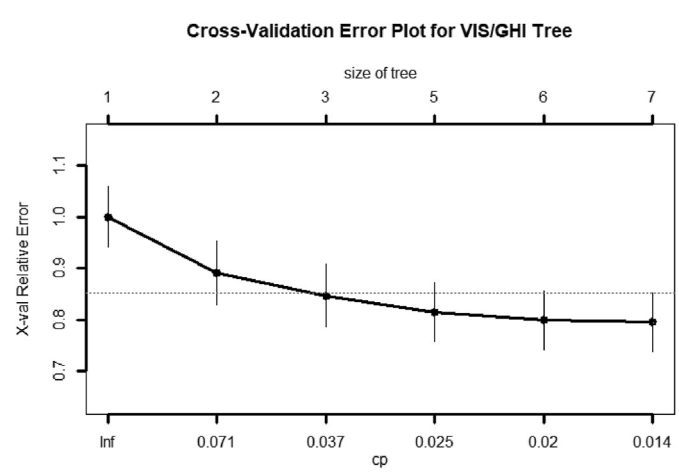

Fig. 2. Cross-validation error plot for the *VIS/GHI* tree.

2) Regression tree with minimum cross-validation error

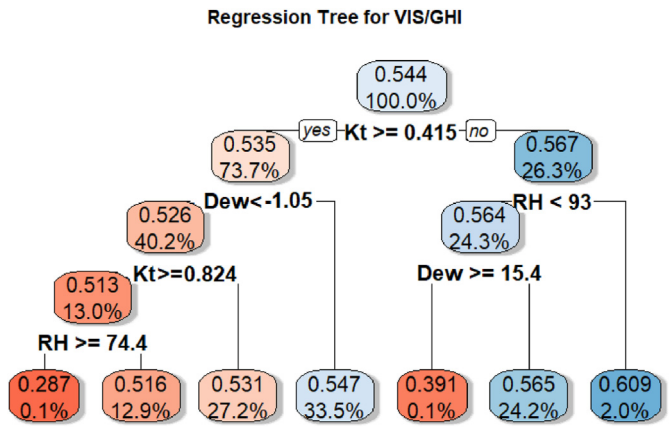

Fig. 3. Regression tree model for *VIS/GHI*.

3) Regression tree with the 1-SE rule

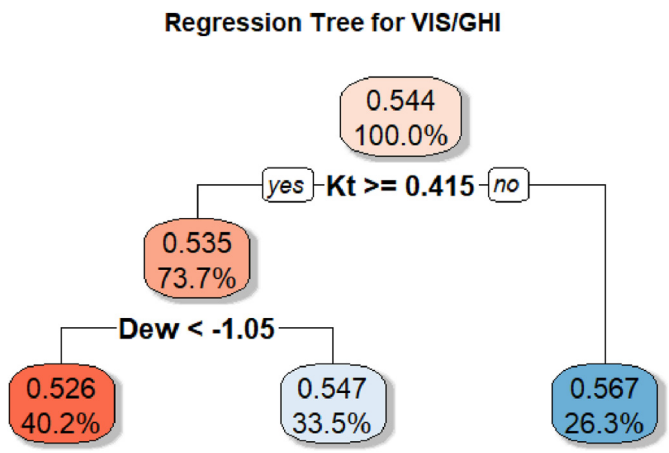

Fig. 4. Pruned regression tree model for *VIS/GHI*.

data, with a mean value of 0.526) or *Dew* $\geq$ -1.05°*C* (Group 3: 33.5% of *VIS/GHI* data, with a mean value of 0.547). Similarly, the fraction values' differences depending on the sky clearness levels can still be found in this resultant tree. Furthermore, although both parameters - dewpoint temperature and relative humidity are related to the water content in the air, the result demonstrated that the *Dew* parameter seemed more determinant to estimate *VIS/ GHI*.

4.2. CART results for the NIR/GHI fraction

1) Cross-validation error plot

Fig. 5 shows the cross-validation error plot for the *NIR/ GHI* tree. From this figure, we can see that when $cp = 0.01$, the Size 10 regression tree has the minimum cross-validation error. This tree model is shown in Fig. 6.

The CART procedure generated a tree containing 10 terminal nodes for *NIR/GHI* (see Fig. 6). The fraction of *NIR/GHI* ranged from 0.358 to 0.743 in these ten groups, among which the fraction 0.414 represented the primary (33.5%) training observations that belonged to that node. As discussed above, the sky clearness conditions, or cloudiness situations, are important to affect the transmitted solar radiation features in terms of the visible and infrared components through the atmospherical layer. Therefore, we could find a similar but reversal relationship in the regression tree for

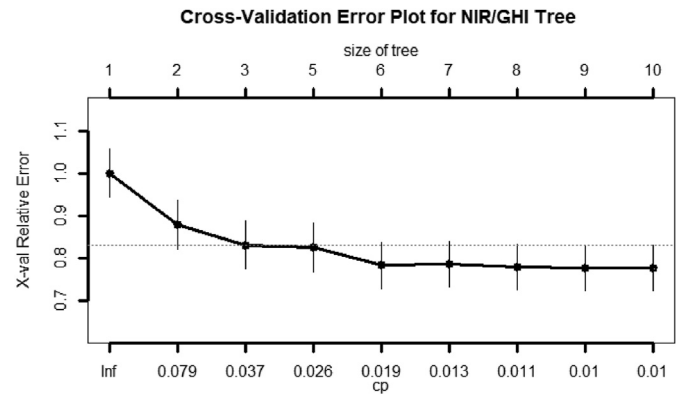

Fig. 5. Cross-validation error plot for the *NIR/GHI* tree.

2) Regression tree with minimum cross-validation error

(The red dotted line refers to the simplest tree, following the 1-SE rule). . (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

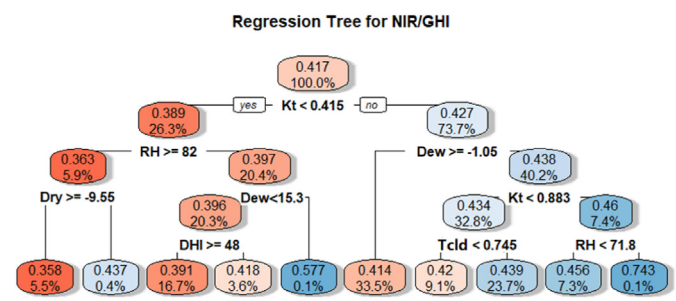

Fig. 6. Regression tree model for *NIR/GHI*.

3) Regression tree with the 1-SE rule

NIR/GHI. The clearness index K_t was still the most determinant parameter selected by the CART procedure, while most higher fractions *NIR/GHI* existed in the higher K_t regions because less near-infrared solar radiation was absorbed by the atmosphere in such situations. However, compared to the regression tree *VIS/GHI*, the regression tree for *NIR/GHI* seemed more complicated. On the one hand, a more scattered percentage could be found in the terminal nodes. On the other hand, more predictors were involved in the pruned regression tree model. In addition to *RH* and *Dew*, dry bulb temperature *Dry*, diffused horizontal solar irradiance *DHI*, and cloud transmittance T_{cld} that was newly proposed in this work were used to form the terminal groups.

The dashed red line in Fig. 5 shows the position of the 1-SE rule with the minimum *xerror* + *xstd*; Fig. 7 shows that the pruned tree using the 1-SE rule for *NIR/GHI* contained three terminal nodes. The percentage of *NIR/GHI* ranged from 26.3% to 40.2% in these three groups. The first variable selected for splitting was the clearness index K_t. If $K_t <$ 0.415, no further split was observed for Group 1: 26.3% of *NIR/GHI*, with a mean value of 0.389. If $K_t \geq$ 0.415, the group was further split according to *Dew* $\geq$ -1.05°*C* (Group 2: 33.5% of *NIR/GHI*, with a mean value of 0.414) or *Dew* < -1.05°*C* (Group 3: 40.2% of *NIR/GHI*, with a mean value of 0.438). In this model, K_t and *Dew* are important parameters used to estimate *NIR/GHI*, which is consistent with the regression tree for the visible component.

4.3. Estimation performance evaluation

The resultant tree models in Figs. 3 and 4 and Figs. 6 and 7 are

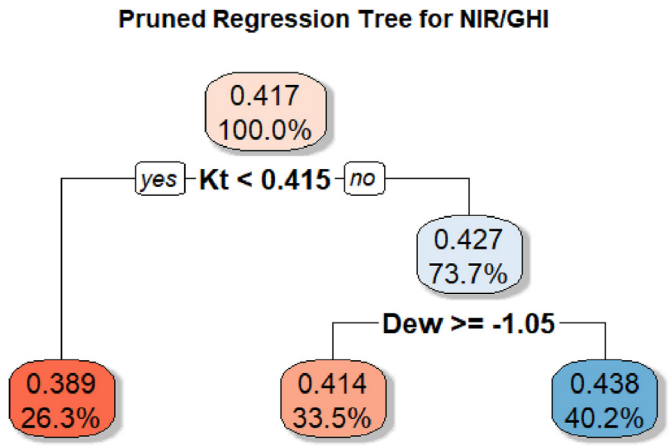

Fig. 7. Pruned regression tree model for *NIR/GHI*.

named Model 1, Model 2, Model 3, and Model 4, respectively. To further understand each model's estimation performance, we calculated the root mean squared error (RMSE) and the mean absolute error (MAE) of these four tree models on the test dataset with 758 observations. The $\hat{y}_j$ variable was the prediction.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \tag{8}$$

From Table 1, we can see that the RMSE decreased as the tree size decreased, but the MAE increased as the tree size decreased. Comparing Models 1 and 2, the RMSE decreased by 0.24% and the MAE increased by 6.6%. Comparing Models 3 and 4, the RMSE decreased by 0.77% and the MAE increased by 1.4%. Regarding the changes in RMSE, since the errors were squared before they were averaged, larger errors receive a relatively higher weight. This means that the RMSE is more useful when significant errors are particularly undesirable. However, the RMSE did not necessarily increase with the variance of the errors. The RMSE increased with the variance of the frequency distribution of error magnitudes Fig. 8 indicates the agreement level between the predicted data from the four models and actual value in the validation tests. Based on the information shown in Fig. 8 and Table 1, we can find the accuracy level differences among the models were negligible in this work. Both Models 1 and 2 had excellent prediction performances for *VIS/GHI*, and Models 3 and 4 had outstanding prediction performances for *NIR/GHI*. This offers the opportunity to simplify the computation process if the weather data are insufficient.

Table 1
Comparison of RMSE and MAE by model.

Regression Tree	VIS/GHI		NIR/GHI	
	Model 1	Model 2	Model 3	Model 4
Tree size	7	3	10	3
RMSE	0.0425	0.0424	0.0391	0.0388
MAE	0.0225	0.0241	0.0213	0.0216

5. Workflow of model incorporation in solar architecture design and analysis

The tree models developed were written via a programming language to form an executable file that could then be used to modify the weather file (e.g., TMY); the result was two new weather files labeled Weather_VIS and Weather_NIR. Note that the HMM data following the TMY weather file's format was selected in this study, but not necessarily to be the input file. As long as the required input data are available in any typical weather files, the decomposition computation can be processed. Table 2 summarizes the required input variables for each model. In brief, *Kt* and *Dew* are the two most important variables to determine both trees, and adding the parameter, *RH*, may slightly increase the accuracy for the models. Comparatively, to get the most accurate model for *NIR/GHI*, some other parameters would be needed, such as *Dry* and T_{cld}.

If the weather data variables were complete, Models 1 and 3 were adopted for Weather_VIS and Weather_NIR file generation, respectively. If some variables were missing, the simpler models (i.e., Models 2 and 4) were applied, as those variables are normally available or computable in most standard weather files. For instance, the cloud coverage data may not be recorded in some weather files; then, Model 2, rather than Model 3, would be applied to calculate the NIR. The input files do not have to be serially complete or comprised of an entire year of 8760 h. A yearly file of daylight hours, monthly file of daylight hours, or just a few hours of data can be used for the computation. In the two files generated, the original GHI data were replaced with the solar VIS and NIR components calculated in each, which were based on the resultant classification tree models and input of the original weather file.

Solar architecture designers can now use existing solar isolation calculation engines embedded in design platforms, such as the Solar Analysis plugin for Revit and Solar Exposure plugin for Sketchup to calculate the solar insolation on building forms. These new separate VIS and NIR solar analysis results will provide more comprehensive and accurate quantities for designers during the early design stage, guiding window placement, window-to-wall ratios, and essential solar heat utilization or blockage. Fig. 9 presents a schematic diagram of this model and examples of applications. Window energy performance includes both optical and thermal aspects. The optical aspect is correlated to VIS and determines the indoor daylighting benefits and electrical lighting energy savings, while the NIR plays an important role in the thermal aspect, especially for transparent NIR reflecting, blocking, or photovoltaic window products, determining the indoor heating and cooling energy use or electric power generated [47]. As more and more spectrally controllable independent building elements emerge, such simulation ability will enable designers and engineers to perform more accurate and comprehensive analyses at the early design stage.

6. Conclusion

This work demonstrated the feasibility and excellent prediction performance of regression tree models for hourly *VIS/GHI* and *NIR/GHI*. The two-year solar spectra and TMY format-compliant hourly weather data obtained from the SRRL BMS database of the NREL Solar Radiation Research Laboratory were utilized for model development. The solar spectra data, ranging from 300 nm to 1700 nm, was used in this study. After the data cleaning process, based on the typical removal of missing and outlier values and erroneous data upon physics-based calculations, the finalized dataset included 7583 observations. To build a more generalizable model for different locations, we intentionally incorporated ten local meteorological parameters, such as humidity, temperature,

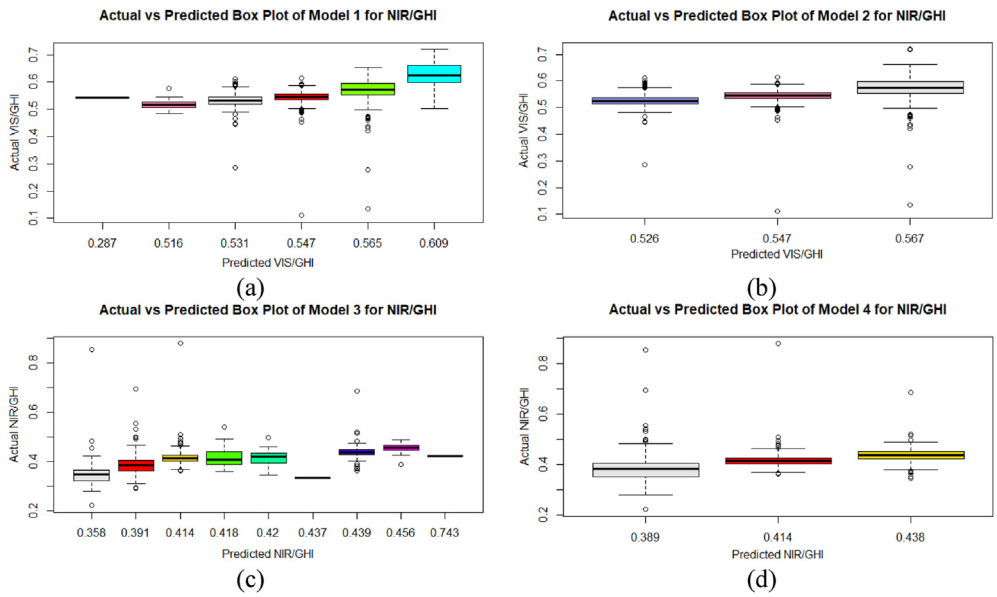

Fig. 8. Actual vs. predicted value box plot of the four models.

Table 2
Required input variables.

VIS/GHI		NIR/GHI	
Model 1	Model 2	Model 3	Model 4
K_t, *Dew*, *RH*	K_t, *Dew*	K_t, *Dew*, *RH*, *Dry*, *DHI*, T_{cld}	K_t, *Dew*

etc., into the modeling procedure. Based on the knowledge and theory of solar radiation and building physics, we also developed five new predictor parameters such as cloud transmittance T_{cld} that can be derived by cloudiness values in typical weather files. In total, 15 predictor variables were used to build estimation models for hourly *VIS*/*GHI* and *NIR*/*GHI*. This research yielded models capable of converting the broadband solar irradiance data in weather files into two separate solar components, VIS and NIR, for building energy and performance-related studies in which independent solar spectra products are examined, such as analyses of spectrally selective glazing, transparent photovoltaic panels, etc. Solar components, especially NIR, are significantly affected by atmospheric parameters, but those parameters are not very well documented observationally and dependent on local geographic and climatic features. In general, the clearness index K_t and dew point temperature *Dew* were the most important variables for clustering the two fractions of VIS and NIR. Adding the new parameter T_{cld} was relatively effective in enhancing the model accuracy when it comes to the NIR solar component. Additionally, the validation tests indicated the MAE (2.13%–2.41%) and RMSE (3.88%–4.25%), demonstrating the reliable performance of the classification tree models developed. To briefly illustrate the importance of this study, we can take the measured solar radiation data in 2019 in Boulder, Denver, as an example. We can get the annual solar resource,1.65 MWh/m [2], on a horizontal surface based on the measured broadband solar data. If we take the simple fractions 40% and 51% to represent the visible and infrared components [48], respectively, we could get 0.66 MWh/m^2 annual solar energy in the visible region and 0.84 MWh/m^2 annual solar energy in the infrared region. However, applying the models developed in this work to the broadband solar data can yield two very different numbers: 0.89 MWh/m^2 and 0.69 MWh/m^2 for solar visible and infrared energy, respectively. Such differences inform us of the potential biases or errors in analyzing building energy performance when it comes to spectrally selective materials or devices if we lack the data of solar spectral components in our analysis.

The major contribution of this work is to provide an easy-of-use tool that can transform the conventional weather files with broadband solar data into the weather files with solar spectral components. Furthermore, this transformation procedure does not require costly solar spectral measurements but rather the typical and readily accessible meteorological data. Combined with computational solar analytic approaches in the current design and engineering platforms, the clustering of solar visible and infrared irradiance can provide a foundation for solar architecture design and analysis. However, a variety of solar modeling algorithms (e.g., Perez model, Liu-Jordan model) used for calculating the incident solar radiation on tilted or vertical building surfaces in different solar building design and analysis programs. These embedded algorithms determine how to retrieve and process the solar radiation data (i.e., GHI, DHI, DNI) and other related weather data (e.g., cloudiness, dry bulb temperature) for computing the incident solar radiation. Therefore, the question of how to fully utilize the solar spectral weather data generated by the models in this work in various design- and simulation-based programs has not been addressed in this work. We plan to select several representative programs in our future work and then carry out an in-depth investigation of their inner solar analytic algorithms, incorporate the solar spectral models, and finally validate the simulation results compared with the actual solar spectral measurement data. Additionally, validation tests based on different locations and solar spectra data will also be conducted in our future work to demonstrate the generalizability of the present research.

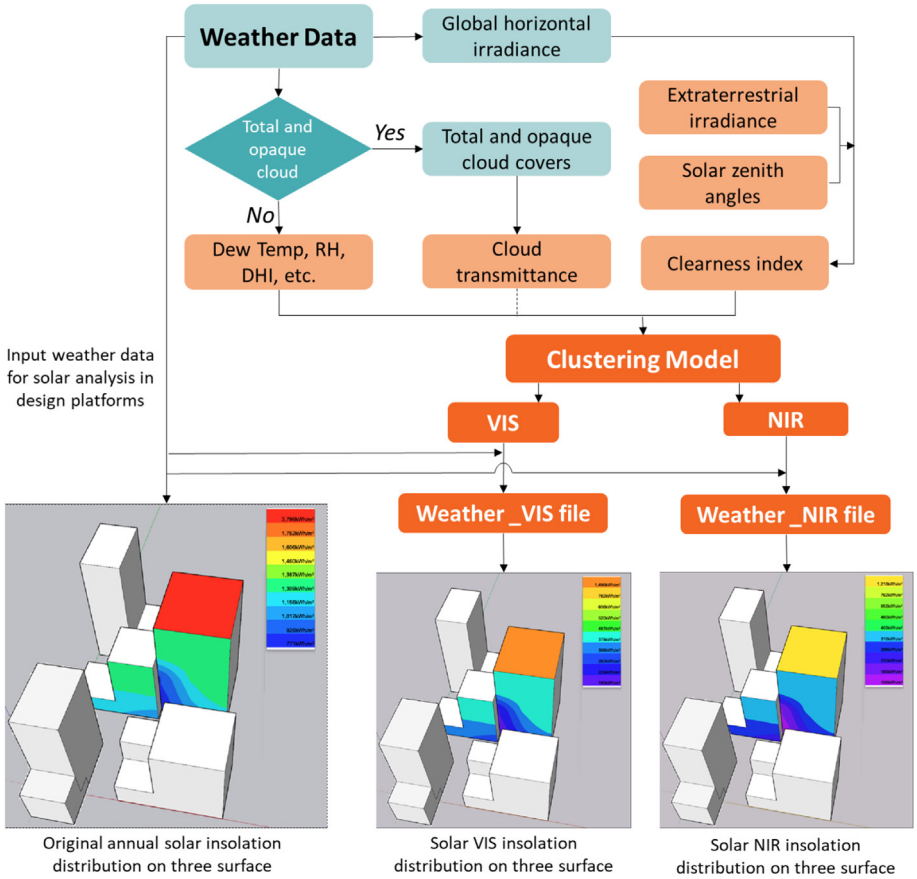

Fig. 9. Schematic diagram of the use of the clustering models.

CRediT authorship contribution statement

Qiuhua Duan: Investigation, Formal analysis, Data curation, Validation, Visualization, Writing - original draft. **Yanxiao Feng:** Data curation, Resources. **Julian Wang:** Conceptualization, Methodology, Writing - original draft, Visualization, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by the National Science Foundation (2001207): CAREER: Understanding the Thermal and Optical Behaviors of the Near Infrared (NIR) - Selective Dynamic Glazing Structures and the USDA Natural Resources Conservation Service, United States (NR203A750008G006): Spectrally-Selective Solar Films for Operational Energy Savings of Urban Greenhouses.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.renene.2020.11.080.

References

[1] C. Schittich, Solar Architecture: Strategies, Visions, Concepts, Walter de Gruyter, 2012.
[2] J. Perlin, Let it Shine: the 6,000-year Story of Solar Energy, New World Library, 2013.
[3] T. Kisilewicz, Computer simulation in solar architecture design, Archit. Eng. Des. Manag. 3 (2007) 106–123.
[4] J. Wang, D. Shi, Spectral selective and photothermal nano structured thin films for energy efficient windows, Appl. Energy (2017) 83–96, https://doi.org/10.1016/j.apenergy.2017.10.066.
[5] C.A. Gueymard, D. Myers, K. Emery, Proposed reference irradiance spectra for solar energy systems testing, Sol. Energy 73 (2002) 443–467.
[6] B. Leckner, The spectral distribution of solar radiation at the earth's surface-elements of a model, Sol. Energy (1978), https://doi.org/10.1016/0038-092X(78)90187-1.
[7] D.T. Brine, M. Iqbal, Diffuse and global solar spectral irradiance under cloudless skies, Sol. Energy 30 (1983) 447–453.
[8] W.K. Tobiska, et al., The SOLAR2000 empirical solar irradiance model and forecast tool, J. Atmos. Solar-Terrestrial Phys. 62 (2000) 1233–1250.
[9] R.E. Bird, R.L. Hulstrom, L.J. Lewis, Terrestrial solar spectral data sets, Sol. Energy 30 (1983) 563–573.
[10] R.E. Bird, C. Riordan, Simple solar spectral model for direct and diffuse irradiance on horizontal and tilted planes at the earth's surface for cloudless atmospheres, J. Clim. Appl. Meteorol. 25 (1986) 87–97.
[11] W.T. Ball, N.A. Krivova, Y.C. Unruh, J.D. Haigh, S.K. Solanki, A new SATIRE-S spectral solar irradiance reconstruction for solar cycles 21-23 and its implications for stratospheric ozone, J. Atmos. Sci. (2014), https://doi.org/10.1175/JAS-D-13-0241.1.
[12] J. Lean, Evolution of the sun's spectral irradiance since the maunder minimum, Geophys. Res. Lett. (2000), https://doi.org/10.1029/2000GL000043.
[13] A.I. Shapiro, et al., A new approach to long-term reconstruction of the solar irradiance leads to large historical solar forcing 67 (2011) 1–8.
[14] R.E. Bird, C. Riordan, Simple solar spectral model for direct and diffuse irradiance on horizontal and tilted planes at the earth's surface for cloudless atmospheres, J. Clim. Appl. Meteorol. (1984), https://doi.org/10.1175/1520-0450(1986)025<0087:SSSMFD>2.0.CO;2.
[15] C.A. Gueymard, Parameterized transmittance model for direct beam and

circumsolar spectral irradiance, Sol. Energy 71 (2001) 325–346.
[16] K.L. Yeo, N.A. Krivova, S.K. Solanki, EMPIRE: a robust empirical reconstruction of solar irradiance variability, J. Geophys. Res. Sp. Phys. 122 (2017) 3888–3914.
[17] Q. Duan, Y. Song, Y. Feng, E. Zhang, J. Wang, S. Niu, Solar infrared radiation towards building energy efficiency: measurement, data, and modeling, Environ. Rev. 999 (2020) 1–9. http://www.nrcresearchpress.com/doi/abs/10.1139/er-2019-0067.
[18] A.R. Pereira, E.C. Machado, M.B.P. de Camargo, Solar radiation regime in three cassava (Manihot esculenta Crantz) canopies, Agric. Meteorol. (1982), https://doi.org/10.1016/0002-1571(82)90053-X.
[19] X. Zhang, Y. Zhang, Y. Zhoub, Measuring and modelling photosynthetically active radiation in Tibet Plateau during April-October, Agric. For. Meteorol. (2000), https://doi.org/10.1016/S0168-1923(00)00093-9.
[20] G. Szeicz, Solar radiation for plant growth, J. Appl. Ecol. 11 (1974) 617–636.
[21] J.F. Escobedo, E.N. Gomes, A.P. Oliveira, J. Soares, Modeling hourly and daily fractions of UV, PAR and NIR to global solar radiation under various sky conditions at Botucatu, Brazil, Appl. Energy 86 (2009) 299–309.
[22] D. Charuchittipan, et al., A semi-empirical model for estimating diffuse solar near infrared radiation in Thailand using ground-and satellite-based data for mapping applications, Renew. Energy 117 (2018) 175–183.
[23] F. Ferrera-Cobos, J.M. Vindel, R.X. Valenzuela, J.A. González, Analysis of spatial and temporal variability of the PAR/GHI ratio and PAR modeling based on two satellite estimates, Rem. Sens. 12 (2020) 1262.
[24] T.J. Rossi, et al., Global, diffuse and direct solar radiation of the infrared spectrum in Botucatu/SP/Brazil, Renew. Sustain. Energy Rev. 82 (2018) 448–459.
[25] L. Wang, et al., Measurement and estimation of photosynthetically active radiation from 1961 to 2011 in Central China, Appl. Energy 111 (2013) 1010–1017.
[26] X. Yu, Z. Wu, W. Jiang, X. Guo, Predicting daily photosynthetically active radiation from global solar radiation in the Contiguous United States, Energy Convers. Manag. 89 (2015) 71–82.
[27] L.J.G. Aguiar, et al., Modeling the photosynthetically active radiation in South West Amazonia under all sky conditions, Theor. Appl. Climatol. 108 (2012) 631–640.
[28] C.P. Jacovides, J. Boland, D.N. Asimakopoulos, N.A. Kaltsounides, Comparing diffuse radiation models with one predictor for partitioning incident PAR radiation into its diffuse component in the eastern Mediterranean basin, Renew. Energy 35 (2010) 1820–1827.
[29] C.P. Jacovides, F.S. Tymvios, J. Boland, M. Tsitouri, Artificial Neural Network models for estimating daily solar global UV, PAR and broadband radiant fluxes in an eastern Mediterranean site, Atmos. Res. 152 (2015) 138–145.
[30] I. Foyo-Moreno, I. Alados, L. Alados-Arboledas, A new conventional regression model to estimate hourly photosynthetic photon flux density under all sky conditions, Int. J. Climatol. 37 (2017) 1067–1075.
[31] A. Andreas, T. Stoffel, NREL Solar Radiation Research Laboratory (SRRL): Baseline Measurement System (BMS), 1981, https://doi.org/10.5439/1052221.
[32] Solargis. Data format, Available at: https://solargis.com/docs/product-guides/time-series-and-tmy-data/data-format, 2019. Accessed: 20th May 2020.
[33] International Standars Organization (ISO), ISO 20473-2007, Optics and Photonics — Spectral Bands, ISO, Geneva, Switzerland, 2007, 2007.
[34] University of Oregon, UO SRML: solar radiation basics, Available at: http://solardat.uoregon.edu/SolarRadiationBasics.html. Accessed: 20th May 2020.
[35] L. Jacobson, A. Seaver, J. Tang, AstroCalc4R: Software to Calculate Solar Zenith Angle; Time at Sunrise, Local Noon, and Sunset; and Photosynthetically Available Radiation Based on Date, Time, and Location, 2019. Available at: https://nefsc.noaa.gov/publications/crd/crd1114/. Accessed: 20th May 2020.
[36] B.Y.H. Liu, R.C. Jordan, The interrelationship and characteristic distribution of direct, diffuse and total solar radiation, Sol. Energy 4 (1960) 1–19.
[37] N. Persaud, D. Lesolle, M. Ouattara, Coefficients of the Angström-Prescott equation for estimating global irradiance from hours of bright sunshine in Botswana and Niger, Agric. For. Meteorol. 88 (1997) 27–35.
[38] E.L. Maxwell, A Quasi-Physical Model for Converting Hourly Global Horizontal to Direct Normal Insolation, Solar Energy Research Inst., Golden, CO (USA), 1987.
[39] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC press, 1984.
[40] T.M. Therneau, E.J. Atkinson, An Introduction to Recursive Partitioning Using the RPART Routines, 1997.
[41] A.M. Prasad, L.R. Iverson, A. Liaw, Newer classification and regression tree techniques: bagging and random forests for ecological prediction, Ecosystems 9 (2006) 181–199.
[42] D.M. Moore, B.G. Lees, S.M. Davey, A new method for predicting vegetation distributions using decision tree analysis in a geographic information system, Environ. Manage. 15 (1991) 59–71.
[43] O. Kisi, S. Heddam, Z.M. Yaseen, The implementation of univariable scheme-based air temperature for solar radiation prediction: new development of dynamic evolving neural-fuzzy inference system model, Appl. Energy 241 (2019) 184–195.
[44] A. Torres-Barrán, Á. Alonso, J.R. Dorronsoro, Regression tree ensembles for wind energy and solar radiation prediction, Neurocomputing 326 (2019) 151–160.
[45] B. Keshtegar, C. Mert, O. Kisi, Comparison of four heuristic regression techniques in solar radiation modeling: kriging method vs RSM, MARS and M5 model tree, Renew. Sustain. Energy Rev. 81 (2018) 330–341.
[46] I. Janssen, J.H. Stebbings, Prediction of Radiation Levels in Residences: A Methodological Comparison of CART (Classification and Regression Tree Analysis) and Conventional Regression, Argonne National Lab., IL (USA), 1990.
[47] Yanxiao Feng, Qiuhua Duan, Julian Wang, Stuart Baur, Julian Wang, Approximation of building window properties using in situ measurements, Build. Environ. 169 (2020), https://doi.org/10.1016/j.buildenv.2019.106590.
[48] S.C. Bhatia (Ed.), Advanced Renewable Energy systems,(Part 1 and 2), CRC Press, 2014, p. 32.