Renewable Energy 219 (2023) 119554

Contents lists available at ScienceDirect

Renewable Energy

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

Reconstruction of narrowband solar radiation for enhanced spectral selectivity in building-integrated solar energy simulations

Chenshun Chen [a], Qiuhua Duan [b], Yanxiao Feng [c], Julian Wang [a,*], Neda Ghaeili Ardabili [a], Nan Wang [a], Seyed Morteza Hosseini [d], Chao Shen [e]

[a] Department of Architectural Engineering, Pennsylvania State University, USA
[b] Department of Civil, Construction, and Environmental Engineering, The University of Alabama, USA
[c] School of Applied Engineering and Technology, New Jersey Institute of Technology, USA
[d] Department of Architecture, Design & Media Technology, Aalborg University Copenhagen, Denmark
[e] Harbin Institute of Technology, China

ARTICLE INFO

Keywords:
Solar irradiation
Solar envelope design and simulation
Machine learning
Spectral selective
Building integrated photovoltaics
Transparent photovoltaics

ABSTRACT

Solar radiation is a critical factor in advanced envelope design and solar energy's building integration, necessitating a shift from broadband solar radiation analyses towards more precise narrowband or spectrum-focused approaches. Understanding the performance potential of spectral-selective materials or structures requires accurate solar spectral information at specified locations, a feature often overlooked by conventional modeling tools. This work presents an innovative solar decomposing model capable of differentiating key solar irradiation components—visible and infrared—from broadband solar irradiance, without the need for expensive spectrum measurements. Our approach employs the extreme boosting regression tree method and leverages existing or easily derivable data from typical weather files. An exploratory analysis of the importance and interaction of different features in predicting solar irradiation components is also conducted. The results show that the proposed algorithm has an R^2 of 0.981 and 0.990, RMSE of 18.280 and 18.390, and MAE of 7.989 and 8.011, for predicting VIS and NIR amount in DNI, respectively (for the strongest model using all the predictors within the dataset). This research offers an added layer of practicality by including case studies demonstrating how the solar decomposition models serve in real-world applications, especially in the integration of wavelength-selective devices like window systems and transparent solar cells into advanced envelope designs. Such real-world testing has verified the presence of a disparity between the power output calculation of NIR-selective transparent photovoltaics upon the broadband solar radiation data and the suggested narrowband solar radiation data, potentially resulting in a maximum deviation of 15.7 %. The decomposing models developed empower researchers and designers to generate new weather files comprising narrowband solar irradiance data, thereby enhancing their capacity to examine the influence of spectral-selective materials on a building's solar performance using existing solar simulation programs. The proposed method will be further improved by including more data from different climate zones and weather characteristics in the training model and validating through field measurements.

1. Introduction

Solar radiation is a pivotal element in the sphere of advanced envelope design, particularly for those designs that incorporate solar cells or similar devices. This focus is aimed at harnessing solar power to cultivate energy-efficient building structures. The use of computational analysis tools such as EnergyPlus and Comfen has been adopted to dissect the thermal and optical properties of solar-based architectural design. Typically, a full year's weather data is assimilated into a simulation program to emulate a building's daylight environment, energy consumption, and other related conditions. In this context, a reliable solar radiation dataset becomes indispensable to ensure superior predictive performance for solar projects. Both the variability of solar resources, as mirrored in historical solar data, and the precision of the dataset, hold immense significance in the accurate estimation of the projects [1]. There are several widely available datasets: 1) the National Solar Radiation Data Base (NSRDB), developed by the National

* Corresponding author. 104 Engineering Unit A, Department of Architecture Engineering, The Pennsylvania State University, University Park, PA, 16802, USA.
E-mail address: julian.wang@psu.edu (J. Wang).

https://doi.org/10.1016/j.renene.2023.119554
Received 18 July 2023; Received in revised form 13 October 2023; Accepted 30 October 2023
Available online 31 October 2023

Nomenclature

NIR	Near-Infrared Light
UV	Ultraviolet Light
GHI	Global Horizontal Irradiance, W/m^2
DHI	Diffuse Horizontal Irradiance, W/m^2
VIS	Visible Light
DNI	Direct Normal Irradiance, W/m^2
CART	Classification and Regression Tree
RMSE	Root Mean Square Error
MBE	Mean Bias Error
TMY	Typical Meteorological Year
AOD	Aerosols Optical Depth
PWV	Perceptible Water Vapor
I_0	Extraterrestrial Solar Irradiance, W/m^2
TPV	Transparent Photovoltaic
SZA	Solar Zenith Angle
K_b	Normal Clearness Index
AM	Air Mass
SKC	Total Sky Cover, 1/tenth
Opqcld	Opaque Sky Cover, 1/tenth
T_{cld}	Cloud Transmittance
Dry	Dry Bulb Temperature, °C
Dew	Dew Point Temperature, °C
RH	Relative Humidity
Wdir	Wind Direction
Wspd	Wind Speed, m/s
DT	Decision Tree
R^2	R-squared
PCE	Power Conversion Efficiency

Renewable Energy Laboratory (NREL) and Sandia National Laboratory, 2) the Canadian Weather Energy and Engineering Datasets (CWEEDS), which are available through Environment Canada, and 3) SolarGIS, which merges ground observations, satellite data, and an atmospheric patterns database to deliver highly accurate global solar radiation data. The typical meteorological year (TMY) data files were developed from these datasets, which are also the most popular weather files in building envelope simulations. Each TMY data file consists of a full year of data constructed from 12 months chosen as most typical from the years that made up the database. Three broadband solar components including global horizontal irradiance (GHI), direct normal irradiance (DNI), and diffuse horizontal irradiance (DHI) are available in the TMY data and are mostly all needed for a solar analysis or energy simulation. Notably, all three solar irradiance types (GHI, DNI, DHI) in the typical weather data files are broadband and represent the total amount of ultraviolet (UV), visible light (VIS), and near-infrared radiation (NIR), three major components of the solar spectrum.

In recent years, growing evidence from the building and solar simulation perspective has demonstrated the necessity of separating broadband solar radiation into narrowband or even specific spectrum-focused design and analytics. For instance, of these three solar components, solar VIS radiation is beneficial to building electrical lighting energy savings and indoor health related to its circadian stimulation [2], but it may also lead to negative impacts on indoor visual comfort, such as glare issues. While the solar NIR transmission into the building could help to reduce the overall heating load in cold climates, it is undesirable in hot climates. Similarly, it has also recently been reported that solar VIS and NIR transmitted through glazing have different effects on the user's thermal comfort near window zones [3]. From the solar device or system design perspective, with the recent development of spectral-selective materials, independent spectral band modulation of solar radiation has become increasingly viable as a potential means of improving building energy efficiency and maintaining indoor visual comfort. For example, metallic nanoparticle-based nanocomposites have been recently studied and developed to decouple the modulation of solar VIS and NIR by using their plasmonic resonance effect at specific infrared wavelengths. Jahid et al. proposed reversible photothermal windows based on nanoscale solar infrared-induced plasmonic photothermal effects, which can modulate solar heat, independent of visible light conditions [4]. Shen et al. explored the potential of using silver nanorods (AgNRs) for energy-saving applications, with an adjustable plasma resonance band from the visible light to the infrared, they can ensure higher luminous transmittance than 50 %, while blocking solar radiation by about 80 % [5]. Forrest et al. conducted a comprehensive review of the recent advancements in semitransparent organic photovoltaics for various building applications, including power windows. The unique feature of narrow and intense absorption spectra exhibited by organic materials presents an exciting opportunity for the development of highly efficient organic photovoltaic devices. These devices can maintain semitransparency in the visible spectral range while exhibiting strong absorption in the ultraviolet and infrared spectral bands that are invisible to the human eye [6]. Other researchers also developed some light/heat splitting materials by designing spectral transmittance/absorptance of glazing materials in different solar spectra [7–10].

As such, to meet the above-mentioned needs of separating solar spectral components and understand the potential performance of the spectral-selective materials/structures, accurate solar spectral information in weather files of the selected locations is necessary. If solar simulations only take broadband solar radiation into consideration, the potential performance of the aforementioned materials/structures in building envelopes may not be fully understood. Even worse, it may yield misleading or erroneous results. For example, if one studies a spectral-selective glazing material (e.g., low-solar-heat-gain low-e coatings - low transmittance in the NIR region but high transmittance in VIS region), normal broadband simulations would get erroneous results because they lack the corresponding solar components to be multiplied with VIS or NIR transmittance. Nonetheless, the measurement of solar spectral irradiation is a challenging and costly process. Current predictive tools are limited to forecasting broad-spectrum solar radiation by relying on historical patterns through empirical methods like Autoregressive Integrated Moving Average (ARIMA) or Artificial Neural Network models (ANN). Alternatively, they can estimate specific, narrowband solar radiation using physical models such as SMARTS and MODTRAN, which are based on the physics of solar radiation. Consequently, conventional weather files typically lack the inclusion of spectral distribution or narrowband data regarding solar irradiation.

To address this research gap, a reconstruction algorithm decomposing solar visible and infrared irradiance from broadband solar radiation and weather data for building simulations needs to be developed. The solar radiation data used for building simulations include GHI, DHI, and DNI, so all these three solar components need to be decomposed to narrowband data. In our previous work, we have already developed a predicting model using the CART algorithm for decomposing VIS and NIR components in GHI based on the typical weather file [11]. Accordingly, ***this work focuses on building spectral solar radiation models for DNI and DHI*** based on extracted and added features from readily available weather files without adding new measurements and sensors. By leveraging these models, it becomes feasible to develop a web-based application that can effectively break down conventional, broad-spectrum weather files into distinct, narrowband weather files. Furthermore, as part of this research, a case study focused on the analysis of (semi)transparent solar cell performance is conducted using the developed models.

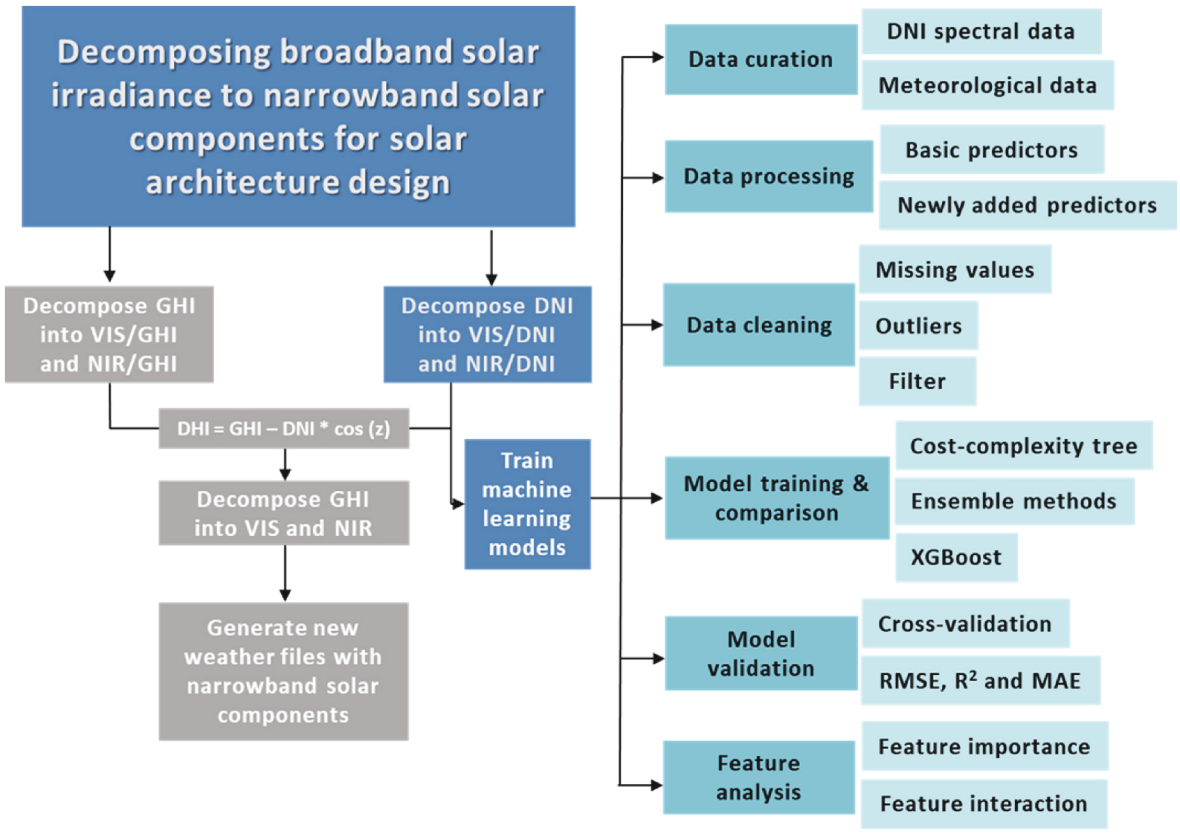

Fig. 1. The workflow of generating new weather files with narrowband solar components and the schema of decomposing DNI components by using machine learning algorithms.

One uniqueness of this research is to leverage the typical meteorological data from the ground-based weather station, such as dew-point temperature, relative humidity, and broadband solar irradiance into the decomposition models without additional measures. In other words, it deploys accurate, efficient machine-learning algorithms to decompose solar spectral bands, with easy-to-get weather files. This method enables more comprehensive and precise building performance simulations, especially with respect to building elements and products that have spectral-selective features. ***Second***, through comprehensive feature importance and feature interaction analyses, certain underlying and intrinsic relationships between the narrowband solar irradiance (VIS or NIR) and meteorological features are first known. This presents some evidence for the rough or simplified estimation of spectral features of solar radiation based on weather features, which can be useable for designers and engineers in solar architecture areas for rapid in situ decision-making. ***Last***, this research is not merely theoretical or limited to the realms of scientific investigation - it has practical implications and real-world applicability, as demonstrated through a case study. This case study functions as a concrete illustration of the application's intended purpose and how it can be deployed in real-life scenarios, particularly concerning advanced building envelopes with solar energy utilization and/or generation.

2. Related work

Since the 1940s, many spectral irradiance models have been developed. There are five types of spectral irradiance models including empirical models, rigorous and sophisticated codes, simple transmittance parameterizations, semiempirical models, and reconstruction models. Empirical models such as Moon's spectral radiation curve, Leckner's model, Brine and Iqbal's model, and SOLAR2000 usually combine historically measured weather and other solar irradiance data. Some rigorous and sophisticated codes including BRITE and FLASH, LOWTRAN 7, MODTRAN 6, SEA, and SOLMOD models can consider the physical characteristics of the atmosphere and vertical profiles of gaseous and aerosol constituents. The simple transmittance parameterization models such as the SPECTRAL2 and the SMARTS models proposed by the National Renewable Energy Laboratory (NREL) simplify the atmosphere's vertical profile. Semiempirical models involve both physical and statistical modeling processes while reconstruction models focus on modeling solar spectral irradiance variability. However, these five representative solar spectral irradiance models cannot be applied to building energy efficiency analysis because of the need for additional measurement and data input, implementation complexities, and wavelength range limitations [12].

A more practical way of integrating solar spectral irradiance in the application of building energy analysis is to develop models that can decompose major solar spectral components, such as UV, VIS, and NIR within GHI, DNI, and DHI. The most previous work is Bird's solar spectral model for direct-normal and diffuse horizontal irradiance [13]. It uses simple mathematical expressions and tabulated look-up tables to generate direct normal and diffuse horizontal irradiance for cloudless days (0.3–4 μm wavelength range). Nann et al. pushed this work further by developing a semiempirical model, called SEDESI, to predict the solar spectral irradiance under clear and cloudy skies, with only a few inputs including global and diffuse broadband irradiance measurements, precipitable water-vapor data, and the solar position [14]. With the improvement of ground and satellite-based solar spectral irradiance measurement, more accurate solar irradiance decomposition methods can be derived by calibrating with the measured spectral data, for example, Tatsiankou et al. developed a decomposition algorithm for deriving the broadband DNI and DHI irradiances from 1-min spectral global horizontal irradiance measurements performed by the SolarSIM-G. Spectral clearness indexes of different sky conditions were

derived based on measured/calculated atmospheric parameters (including solar position, Rayleigh scattering, and the transmittances, spectral AOD, total column ozone at 600 nm, precipitable water vapor at 940 nm, cloud transmittance with a spectral cloud correction, etc.), which were then used to decompose broadband solar irradiance. The algorithm was calibrated and validated at five stations, and it showed promising results: For the DNI estimation, the root mean square error (RMSE) ranged from 26 W/m^2 to 48 W/m^2, while the largest mean bias error (MBE) was 4 W/m^2. For the DHI estimation, the RMSE ranged from 14 W/m^2 to 27 W/m^2, while the largest MBE was 3 W/m^2 [15]. Kosmopoulos et al. looked into AOD-solar irradiance interactions in the eastern Mediterranean area and found that under extreme dust events (i.e., when AOD reaches 3.5), GHI could attenuated by 40–50 %, and DNI could decrease even more (80–90 %), with spectrally attenuated 37 % in the UV region, 33 % in the visible and around 30 % in the infrared [16]. Another characteristic research done by Charuchittipan et al. aims to estimate diffuse solar NIR radiation in Thailand using ground- and satellite-based data for mapping applications, three major atmospheric parameters were analyzed and used as predictors in the model: cloud effect, solar zenith angle, and precipitable water. This semiempirical model showed reasonable agreement with independent diffuse NIR, giving an RMSD and MBD of 16.7 % and 1.5 %, respectively [17]. There are also researchers trying to implement machine learning/deep learning algorithms in solving solar irradiance spectra or modeling solar radiation. Taylor et al. utilized neural network radiative transfer solvers for the generation of high-resolution solar irradiance spectra parameterized by cloud and aerosol parameters [18]. Hassan et al. explored the potential of tree-based ensemble methods in solar radiation modeling. But most of these works were either trying to model the solar irradiance spectral with a very high resolution (continuous spectrum), for meteorological and astronomical studies [19–23], or trying to predict the total solar radiation for renewable energy applications [24–30]. A relatively simpler way of implementing solar spectral irradiance in solar architecture design and energy simulation is to decompose the broadband solar irradiance into individual solar components/fractions. For instance, Szeicz verified that 0.5 is a better approximating ratio of the visible energy to the total received by the photosynthetically active part of the spectrum, his study indicated the VIS fraction of total global irradiance is closely associated with two factors: the presence of clouds and scattering caused by aerosol [31]. The NIR fraction is closely related to the total amount of column water vapor [32].

In summary, all previous works mentioned above studied solar spectral irradiance from meteorological or astronomical perspectives, or for the applications of photovoltaic and agriculture. For the applications of solar architecture design and building energy simulation in terms of all three solar components with good accuracy, few of these works seem applicable. Meanwhile, these studies unveiled the key impacting factors for solar spectral irradiance, including clearness index, precipitable water vapor, aerosols, solar position, and cloud level, which provides us with a valuable foundation for our work. With the aforementioned meteorological/astronomical feature, we can build an ease-of-manipulation tool by using newly developed machine learning algorithms, which are compatible with most of the current building energy simulation software or plug-ins, so that a more accurate, detailed solar radiance model could be built with this software.

3. Methodology

As shown in Fig. 1, firstly, we built two estimation models for decomposing solar components (VIS and NIR) from the broadband GHI and DNI, respectively. The GHI decomposing model and the framework of the data portal have been presented in our prior work [10] The scope of this paper focuses on the procedure for decomposing models from the broadband DNI, which consists of six major steps from data collection to cleaning, feature selecting, model training, comparing, and validating. The best model was selected based on its performance. Based on these two parts (GHI and DNI), one can compute the VIS and NIR components in DHI via the transposition equation (shown in Fig. 1) [33]. Subsequently, one can import their original weather files (TMY), and then export processed weather files (TMY) with individual solar components in replace of broadband GHI, DNI, and DHI information.

3.1. Solar and weather data curation

In this study, all the datasets were obtained from the BMS database of the NREL Solar Radiation Research Laboratory. All the datasets were based in Golden, Colorado with latitude of 39.742° N, longitude of 105.18° W, and elevation of 1828.8 m AMSL [34]. In general, two major datasets were curated: weather datasets (including broadband solar irradiance and other typical meteorological data) and spectral solar irradiance data from multiple sources.

1) Weather dataset

This dataset includes three components. The major component was based on and included most independent variables for modeling, such as GHI, DNI, DHI, cloud coverage, and dry-bulb temperature. The HMM is hourly data by averaging the value of all measurements taken from the 1-h interval and was collected from Jan 1, 2016, to December 31, 2019. In addition, based on the prior studies, both aerosol optical depth and precipitable water vapor parameters are often found in typical weather stations' data collection and are important to the solar spectra, while they are not included in HMM datasets. As such, in this work, we obtained these two atmospheric data from the NREL BMS AOD and PWV (GPS-based PWV) database. As a result, the curated weather dataset has identical variables and formats to the TMY weather file that has been widely applied to solar radiation and building energy performance simulations. All these weather data including broadband solar irradiance were used as the predicting variable candidates in this work.

2) Spectral solar dataset

The corresponding spectral solar components (VIS and NIR irradiance) in GHI were extracted from the outdoor solar spectra data (WISER), and the components in DNI were extracted from the outdoor solar spectra data (PGS-100). The WISER dataset was measured by EKO WISER spectroradiometer MS-710, MS-711, and MS-712 from 2016 to 2019. The MS-710, MS-711, and MS-712 instruments have 4 nm, 5 nm, and 6.5 nm spectral bandwidth respectively. Their measurement range is 300 nm–1100 nm, 300 nm–1100 nm, and 900 nm–1700 nm, respectively [35]. The WISER dataset has a higher resolution measurement for both wavelengths and time intervals (typically 5 min, but occasionally 1 min). The PGS-100 dataset was measured through a LICOR LI-1800 spectroradiometer. The instrument has a 3.6 nm spectral bandwidth and the useable spectral range of the instrument is 350 to 1,050 nm. Data was taken at approximately 0.7 nm intervals (slightly variable, differs for each serial number) every 5 min [32]. The hourly spectrum data were calculated by averaging the 5-min interval data for each hour, The average value of all measured points each hour is defined as the value for the time-stamp at the end of the 1-h interval. All GHI and DNI spectral data were then integrated over the specific bandwidths (VIS: 400 nm–700 nm, NIR: 800 nm–2500 nm) to compute the overall VIS and NIR irradiance within the broadband solar irradiation, following the trapezoidal rule.

$$VIS/NIR_{total} = \sum_{i=1}^{n-1} \frac{E(\lambda_i) + E(\lambda_{i+1})}{2} \times (\lambda_{i+1} - \lambda_i) \quad (1)$$

Where $E(\lambda_i)$ represented the irradiance at a given wavelength λ_i, and n was the number of data points.

3.2. Data processing

The raw dataset contained 35,059 data points, including the response variables VIS and NIR, predicting variables Albedo, Wind speed (Wspd), Wind direction (Wdr), Pressure, Relative humidity (RH), Dew-point temperature (Dew), Dry-bulb temperature (Dry), Opaque cloud cover (Opqcld), Total sky cover (SKC), Diffuse horizontal irradiance (DHI), Direct normal irradiance (DNI) and Global horizontal irradiance (GHI). After collecting the data, feature processing and data cleaning procedures were conducted to select predictors and clean useless data, before feeding the data to machine learning algorithms.

3.2.1. Pre-processing for features used in the model

In addition to the existing weather parameters, such as dry-bulb temperature, dew-point temperature, and relative humidity, several new predictors were generated and added based on the domain knowledge 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. These predictors are strongly correlated with solar irradiance and therefore used in our model.

1) Extraterrestrial solar irradiance I_0:

The hourly average extraterrestrial solar irradiance I_0 is determined using the equation [36].

$$I_0 = I_{sc}\left(\frac{R_{av}}{R}\right)^2 \tag{2}$$

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 Sun-Earth distance is:

$$\left(\frac{R_{av}}{R}\right)^2 = 1.00011 + 0.034221\cos(\beta) + 0.001280\sin(\beta) + 0.000791\cos(2\beta) + 0.000077\sin(2\beta) \tag{3}$$

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

2) Solar zenith and azimuth angle:

The solar zenith angle z is the angle between the solar and the vertical, and solar azimuth angle az is the angle between the sun and north cardinal direction. A Python package called Pysolar (developed by Brandon et al.) was used in this study. Pysolar is a collection of Python libraries for simulating the irradiation of any point on earth by the sun, based on longitude, latitude, time of the day, and date. It includes code for extremely precise ephemeris calculations and more [37].

3) Normal clearness index:

The clearness index b of direct beam irradiation is the ratio of the direct normal irradiance to the corresponding radiation incident on a horizontal plane at the top of the atmosphere. It depends on the altitude, solar zenith angle, and parameters describing the optical state of the atmosphere related to aerosols, water vapor, and other gases [38].

$$K_b = \frac{DNI}{I_0} \tag{4}$$

4) Air mass:

The air mass AM calculation was given by Kasten and Young [39], as

Table 1
List of models, alongside with predicting features each model used.

Model	Predicting features
V1, N1	Albedo, Wspd, Wdr, Pressure, RH, Dew, Dry, Opqcld, SKC, DHI, DNI, GHI, AM, I_0, K_b, Tcld, AOD, PWV, Zenith, Azimuth
V2, N2	Wspd, Wdr, Pressure, RH, Dew, Dry, DHI, DNI, GHI, AM, I_0, K_b, PWV, Zenith, Azimuth

$$AM = \frac{1}{\cos z + 0.50572(96.07995 - z)^{-1.6364}} \tag{5}$$

5) Aerosol Optical Depth:

As a basic optical parameter, aerosol optical depth (AOD) is a measure of the extinction effect of atmospheric aerosols and is widely used as a key parameter for assessing the degree of air pollution. Moreover, aerosols have been analyzed in research on climate change and atmospheric radiation balance [40]. We collected AOD_{500} data from the NREL BMS database, which is obtained from a 7-channel Prede POM-01 Photometer (at 500 nm). Raw data collected by this photometer was sent to the European Sky Radiometers (ESR) Network and processed using two methods: the SkyNet SkyRad Method and the ESR SunRad Method [41].

6) Precipitable water:

PWV (precipitable water vapor) is the total water vapor content of a unit area in the atmospheric column (unit: kg/m^2), which is equal to the liquid water content at the same height (unit: mm), and is related to the integrated wet profile above the station [42]. We collected Jan 1, 2016, to Dec 31, 2019, PWV data from the NREL GPS-based PWV database, with a time interval of 1hr, and the location at 39.7423° North, 105.1785° West (Elevation: 1828.8 AMSL). PWV can also be estimated by surface meteorological observations if measurements are unavailable, as shown in Equation (6), where e_s is the saturation water vapor in hPa, T is the dry-bulb temperature in °C, e is the actual vapor pressure and g is the acceleration due to pressure (9.81 m/s^2):

$$e_s(T) = 6.112 \times e^{\frac{17.67\times T}{T+243.5}}\ e = RH \times e_s(T) PWV \approx 0.1 \times \frac{e}{g} \tag{6}$$

7) Cloud transmittance:

The cloud transmittance is formed by using its correlation with sky cover:

$$T_{cld} = \frac{(1 - 0.1 * Opqcld)(1 - 0.1 * SKC + 0.1 * Opqcld)}{1 - 0.05 * Totcld} = \frac{(1 - 0.1 * Opqcld)(1 - 0.1 * Trn)}{1 - 0.05 * Totcld} \tag{7}$$

where Opqcld is the opaque sky cover, Totcld is the total sky cover, and Trn is the translucent sky cover, in which $Trn = SKC - Opqcld$ (This is based on the assumption that the direct beam transmittance of opaque clouds is zero and only valid for a large number of hours having the same opaque cloud cover) [11,43].

After the feature-selecting procedure, 20 parameters have been chosen as predicting features. These include the original meteorological parameters and newly added parameters, such as Air mass (AM), Extraterrestrial solar radiation (I_0), Normal clearness index (K_b), Cloud transmittance (Tcld), Aerosol optical depth (AOD), Precipitable water vapor (PWV), Solar zenith angle (Zenith) and Solar azimuth angle (Azimuth), as shown in *V1* (decomposing VIS light) and *N1* (decomposing NIR light) models in Table 1. Since some weather stations would not record certain parameters such as AOD, SKC, Opqcld, and Albedo, a

Table 2
Detailed description of the raw dataset.

	count	mean	std	min	25 %	50 %	75 %	max
GHI	35059	194.3923	278.2403	0	0	17	337	1134
DNI	35059	233.3288	349.5303	0	0	0	447	1079
DHI	35059	66.1971	98.28129	0	0	15	95	691
VIS	35059	100.5387	150.3429	**−41.7373**	**−0.09455**	0.8431	203.3645	1095.435
NIR	35059	123.7763	192.5793	**−228.87**	**−0.1741**	0.3977	225.8207	1065.757
Opqcld	35059	0.376993	105.7841	**−9900**	0	0.3	1.8	9.8
Dry	35059	7.834893	183.6196	**−9900**	3.7	11.4	19.1	35.9
Dew	35059	−5.91199	183.2599	**−9900**	−8.4	−2.7	3.7	17.2
RH	35059	42.1731	109.078	**−9900**	23.9	36.7	59.6	100.2
Pressure	35059	814.703	99.41184	**−9900**	813	817	820	831
Wdr	35059	191.7905	115.9286	0	80	245	284	360
Wspd	35059	3.156904	2.15405	0	1.7	2.7	4.1	22
Albedo	35059	0.130126	0.169756	0	0	0.12	0.2	2
Zenith	35059	89.67112	36.16765	17.6665	61.1429	89.6538	**118.2184**	**162.3577**
Azimuth	35059	179.9827	98.23221	7.5602	90.33525	176.0089	269.0277	352.3895
AM	35059	2.212897	4.91397	**−1**	**−1**	1.1766	2.4813	22.0847
AOD	35059	0.721639	1.604419	0	0	0.0503	**0.4249**	**16.1398**
PWV	35059	−55.4384	2560.761	−99999	5.2	8.6	14.15	36.95
I0	35059	1367.176	33.11951	1321.327	1334.042	1366.225	1400.161	1414.951
Kb	35059	0.171208	0.256278	0	0	0	0.330197	0.784086
Tcld	35059	0.875956	0.195337	0.039515	0.83497	0.979592	**1**	**1.997984**
SKC	35059	1.001549	105.8108	**−9900**	0	0.4	3.2	9.9

smaller dataset containing only 15 predicting features was also built to train a relatively simpler model, as shown in *V2* and *N2* models. A preliminary data analysis was conducted, as shown in Table 2. It was noticeable that some values (***bold***) were erroneous, for example, Zenith greater than 90°, VIS or NIR less than 0, etc. Thus, a data cleaning procedure needed to be performed after the feature selection.

3.2.2. Data cleaning

After building up the datasets with the predicting and responding features described above, the quality of the raw data was enhanced by a data cleaning process to filter out any anomaly data points. Basically, all the anomaly data due to measurement error exceeded reasonable range, data recorded during the night, low solar-elevation and heavily overcast sky conditions, and combinations of narrowband data conflicted with broadband data were eliminated. The following criteria were adopted in the data-cleaning process:

- If the values within (SKC, PWV, Pressure, RH, Dew, Dry, Opqcld) variables were equal to −9900 or −99999, the corresponding data entries would be considered meaningless and omitted to remove any possible outliers.
- If data values exceed their reasonable range (e.g., Albedo>1, RH>100 %), the corresponding data entries would be discarded.
- If the meteorological parameters AOD and pressure represent extremely rare conditions (e.g., AOD>3, Pressure<780 hPa), Tukey's Fences method was employed to eliminate potential outliers that fall beyond the lower outer fence ($Q_1 - 3 * IQR$) or upper outer fence ($Q_3 + 3 * IQR$), in which Q_1 and Q_3 respectively are the first and third quantile of the dataset, $IQR = Q_3$-Q_1 is the interquartile range [44].
- If the VIS or NIR value is less than 0, then the corresponding data entries would be discarded.
- Irradiance measurements were calculated for solar zenith angles ranging from 17.5° to 85.5° (the range of available solar zenith angles throughout the year at SRRL, excluding data near sunrise and sunset). So if the solar zenith angle was greater than 85.5°, or less than 17.5°, the corresponding data entries would be discarded (i.e., AM>11.5 would also be emitted) [2].

Table 3
Detailed description of the dataset after cleaning anomaly observations.

	count	mean	std	min	25 %	50 %	75 %	max
GHI	11862	500.0664	259.6076	50	286	489	702	1087
DNI	11862	632.696	310.9543	1	382	717	911	1079
DHI	11862	138.6929	100.0174	21	64	100	191	627
VIS in DNI	11862	270.725	131.0182	0.2514	166.3186	312.1841	375.5786	810.1841
NIR in DNI	11862	342.8875	176.04	0.0518	194.221	378.8343	501.4796	959.0474
Opqcld	11862	2.276125	2.046922	0.2	0.7	1.3	3.4	9.8
Dry	11862	16.42209	9.331261	−18.4	9.6	17.1	23.8	35.9
Dew	11862	−1.8829	7.894198	−32.2	−7.7	−1.9	4.3	16.2
RH	11862	31.74821	17.00421	1	18.8	27.8	41.7	99.9
Pressure	11862	816.2946	5.937008	792	813	817	821	831
Wdr	11862	141.2805	107.5435	0	48	109	256	360
Wspd	11862	3.192649	2.143738	0	1.8	2.7	3.9	22
Albedo	11862	0.241721	0.127577	0.08	0.19	0.21	0.24	0.96
Zenith	11862	55.74573	17.19004	17.6665	43.21168	58.7589	69.68938	84.418
Azimuth	11862	177.0164	61.28535	66.7883	124.8887	173.3167	229.6181	292.6278
AM	11862	2.500321	1.727307	1.0493	1.372925	1.93755	2.915475	11.4932
AOD	11862	0.662684	0.872402	0.0128	0.1056	0.2201	0.89135	3.9369
PWV	11862	10.22316	6.06198	0.15	5.35	8.9	14.5	33.9
I0	11862	1361.798	32.68458	1321.327	1329.538	1356.183	1392.952	1414.951
Kb	11862	0.464505	0.227526	0.000753	0.280863	0.526908	0.668704	0.784086
Tcld	11862	0.814888	0.166532	0.039588	0.719896	0.892737	0.939894	0.984975
SKC	11862	3.325923	2.659018	0.3	1.2	2.1	5.3	9.9

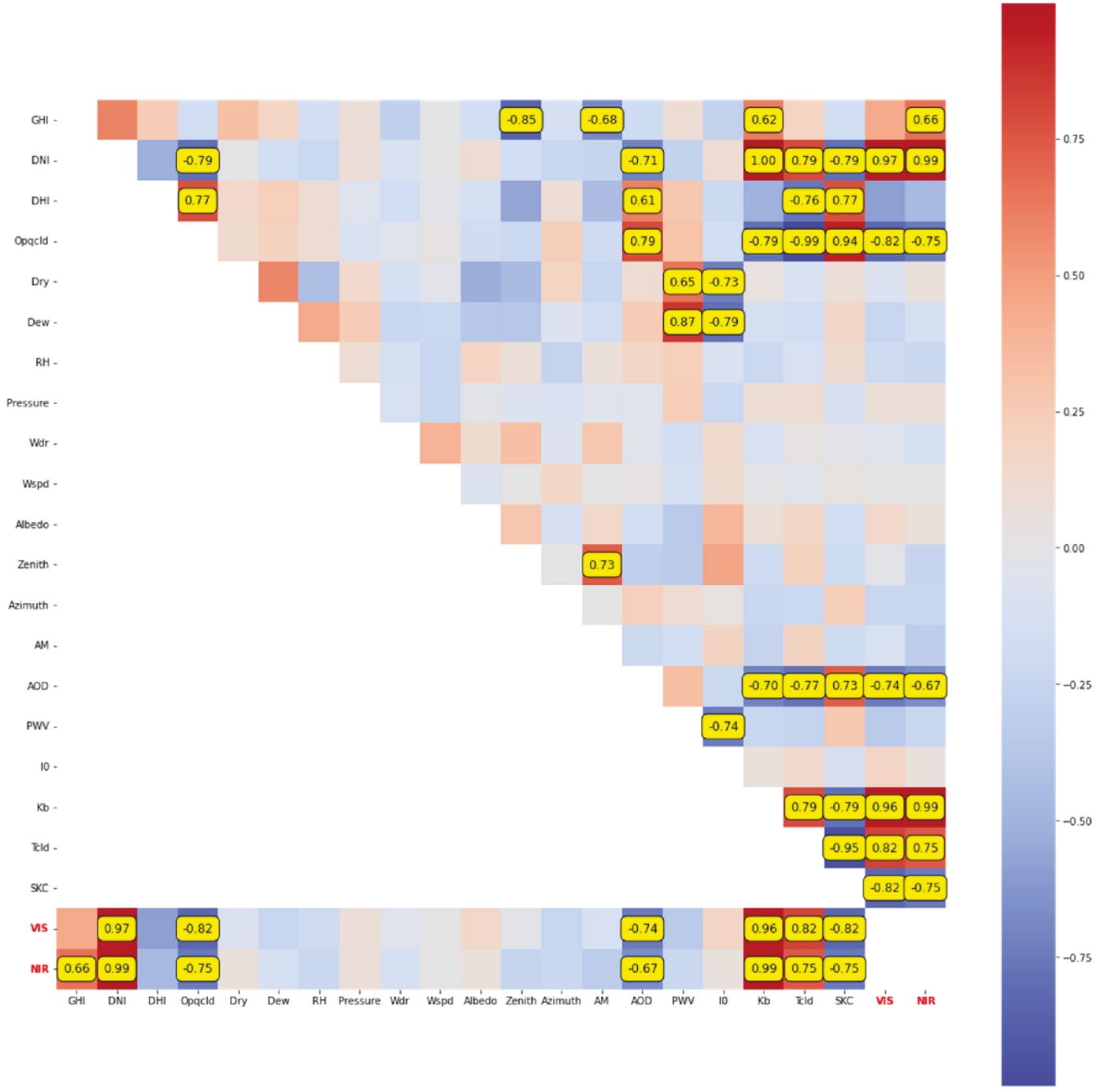

Fig. 2. Correlation heatmap displaying the correlations between predictors, predictors and reponses.

- If the GHI value is less than 50W/m^2, the corresponding data would be disregarded [2].
- If the fractions of VIS/DNI or NIR/DNI were greater than 1, the corresponding data entries would be neglected.
- If the sum of the ratios – VIS/DNI and NIR/DNI was greater than 1 or less than 0.9, then the corresponding data entries would be discarded [45].

The original datasets contain all hourly meteorological and narrowband solar spectrum data from Jan 1st, 2016, to Dec 31st, 2019. There are 35,059 data entries in total. Within the raw dataset, 507 data points were discarded due to measurement or documentation errors, 20,621 data points were discarded due to nighttime or low solar elevation, 1450 data points were discarded due to narrowband-broadband conflicts, 619 data points were discarded due to possible outliers within meteorological parameters. After the data cleaning process, the finalized dataset contains 11,862 observations in total. Detailed data description after cleaning is shown in Table 3.

3.2.3. *Exploratory data analysis after feature processing and data cleaning*

An exploratory data analysis was conducted to explore the correlations between each predictor, alongside predictors and responses. The correlation heatmap in Fig. 2 illustrated that both VIS and NIR components were strongly, and positively correlated with DNI, K_b, Tcld, and negatively correlated with SKC, Opqcld, and AOD. Other noteworthy findings include the positive correlations between DHI and Opqcld,

Opqcld and AOD, Dew, and PWV, and negative correlations between Dry/Dew and I_0, Opqcld and K_b, PWV and I_0, AOD and K_b, AOD and Tcld. These findings might be helpful to evaluate the contributions of different variables in predicting VIS and NIR solar components. Nevertheless, the interactive effects between predictors and responses have yet to be uncovered and will be examined after the model has been trained.

3.3. Data modeling techniques

Many machine-learning algorithms have been implemented in predicting global solar radiation (GSR) so far. For example, Hassan et al. explored the potential of tree-based ensemble methods, such as gradient boosting, bagging, and random forest for estimating global, diffuse, and normal radiation components in daily and hourly time scales [25]. Kisi et al. utilized a dynamic evolving neural-fuzzy inference system model for modeling solar radiation based on a univariable air temperature scheme [46]. Other machine learning algorithms include artificial neural network (ANN), support vector regression (SVR), adaptive neurofuzzy inference system (ANFIS), etc [47–49]. In this study, tree-based, supervised machine learning algorithms were implemented because of their robustness, parallelizability, and capability to handle non-linear relationships. The best tree-based model was selected based on their training, validation, and independent-testing scores, such as RMSE, MAE, and R^2.

1) Decision tree

Decision trees (i.e., DT) are simple but powerful methods for modeling. These involve stratifying or segmenting the predictor space into a number of simple regions. Unlike the generalized linear regression (GLM) model that pre-specifies the relationship between predictors and responses, the DT method utilizes heuristic regression techniques to partition along the predictor axes into subsets with homogeneous values for the response variables. To find the best split, DT takes a top-down, greedy approach that is known as recursive binary splitting, in which it begins at the top of the tree, and then successively splits the predictor space into the regions that lead to the greatest possible reduction in RSS [50]. This process may produce good predictions on the training set but is likely to overfit the data, leading to poor test set performance. To address this issue, tree pruning techniques are required to produce the best-fit subtree. In this study, a tree-pruning method called Cost-complexity pruning is implemented. Rather than considering every possible subtree, only a sequence of trees indexed by a non-negative tunning parameter α is considered. For each value of α, there corresponds a subtree $T \subset T_0$ such that equation (8) is as small as possible. Here $|T|$ indicates the number of terminal nodes of the tree T, and T_0 indicates the number of terminal nodes of the full-grown tree. R_m is the rectangle (i.e. the subset of predictor space) corresponding to the mth terminal node, and $\widehat{y}_{Rm}$ is the predicted response associated with R_m (i.e. the mean of the training observations in R_m). The tuning parameter α controls a trade-off between the subtree's complexity and its fit to the training data. When $\alpha = 0$, the subtree T will simply equal T_0, and then equation (8) just measures the training error. However, as α increases, there is a price to pay for having a tree with many terminal nodes, and so equation (8) will tend to be minimized for a smaller subtree.

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \widehat{y}_{Rm})^2 + \alpha|T| \tag{8}$$

2) Model tree

Like the traditional DT method, the model tree grows the tree the same way as the regression tree, except it has built the multiple linear regression model at each leaf (i.e., rather than make predictions based on the average value of examples that reach a leaf, it builds a linear regression model within each leaf to predict the values of the samples that reach that leaf). The model tree [51] takes one more step forward compared to the traditional decision trees since it builds a piecewise linear model to handle more complex nonlinear relationships while preserving interpretability at the same time. Yong Wang et al. introduced an algorithm called M5′ tree which embodied the model tree's idea of growing the tree with linear models at each leaf. M5′ tree takes three steps to train a model tree: the first step is to build the initial tree, similar to the DT method, M5′ tree splits the predictor space into small regions based on the greatest reduction in error, but unlike the DT method using constant, averaged predicting values at each leaf, M5′ tree makes predictions by using linear regression models at each leaf. The splitting criterion of the M5′ tree also differs from the DT method. It calculates SDR (standard deviation reduction) as its splitting criterion:

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \tag{9a}$$

After growing the initial tree, M5' tree uses a different approach to pruning the tree. The pruning process makes use of an estimate of the expected error that will be experienced at each node for test data. The absolute difference between the predicted value and the actual value is averaged for each training example that reaches that node, and this average is multiplied by the factor $(n+v)/(n-v)$ (where n is the number of the training examples that reach the node and v is the number of the parameters in the model) to compensate the underestimated expected error for the unseen case. M5 computes a linear model (standard regression) for each interior node of the unpruned tree, the resulting linear model is simplified by dropping terms to minimize the estimated error calculated using the above multiplication factor. Finally, once a linear model is in place for each interior node, the tree is pruned back from the leaves [51].

3) Random Forests

Despite the advantages such as easily-interpretable and non-parametric that tree methods possess, they are not competitive with the best-supervised learning approaches in terms of prediction accuracy. Single regression or model trees could be highly unstable due to their high variances. To address this issue, ensemble methods that produce multiple trees which are then combined to yield a single consensus prediction were also tried in this paper, such as Random Forests (RF). The consensus prediction is the mean of the outputs of all DTs:

$$R(x) = \frac{1}{n}\sum_{i=1}^{n} r_i(x) \tag{10a}$$

Where $R(x)$ is the predicted output for input x, and $r_i(x)$ is the prediction of the ith tree. The construction of an RF involves two primary sources of randomness: bootstrap sampling and random feature selection.

4) Extreme Gradient Boosting

The most recent ensemble method is Extreme Gradient Boosting (XGBoost), proposed by *Tianqi Chen* (2016) [52]. XGBoost is an optimization of gradient boosting techniques applied to DTs, which iteratively add new learners (typically decision trees) to an ensemble, where each new learner attempts to correct the mistakes of the combined ensemble of existing learners. Mathematically, if $f_m(x)$ is the prediction of the ensemble at step m, XGBoost seeks to add $h(x)$ such that the new prediction, $f_{m+1}(x) = f_m(x) + h(x)$, minimizes the following objective function:

$$L(f_{m+1}) = \sum_{i=1}^{N} l(y_i, f_{m+1}(x_i)) + \Omega(h) \tag{11a}$$

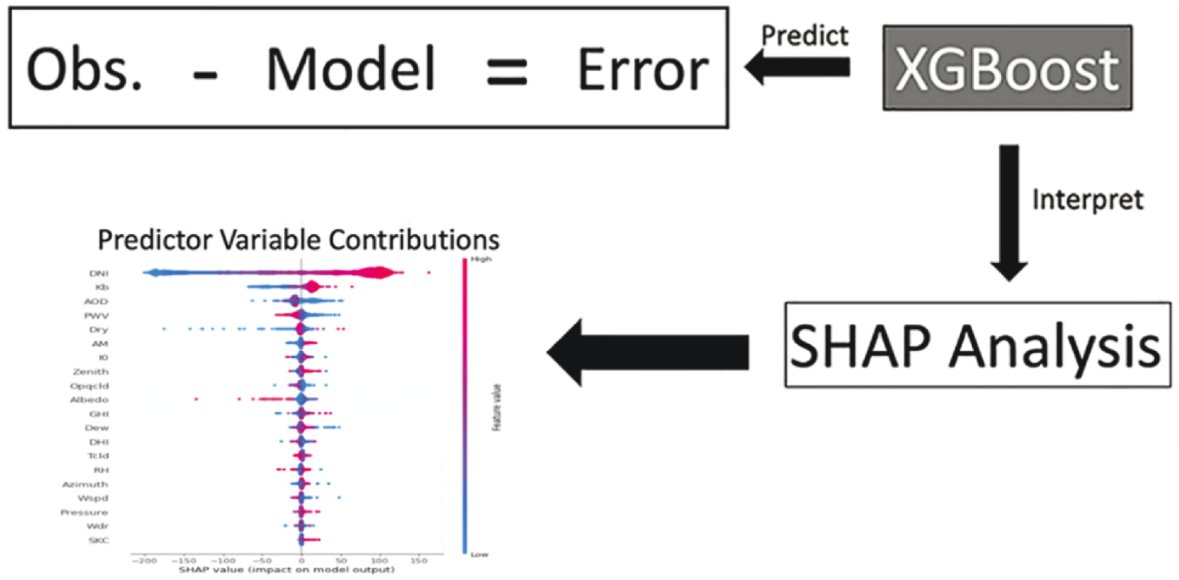

Fig. 3. SHAP method explanation.

Where l is a differentiable loss function that measures the difference between the predictions and the true data labels, y_i, for N data points. The term $\Omega(h)$ represents a regularization term that controls the complexity of the tree model h, preventing overfitting. By employing a second-order Taylor expansion of the loss function and introducing novel techniques for tree pruning, XGBoost provides computational advantages and promotes more robust predictions. Furthermore, the algorithm also supports parallel processing, handling missing values, and offers built-in cross-validation capabilities.

3.4. SHAP method for feature importance and interaction analysis

To evaluate the importance of the predictive features (i.e., the contribution of each feature to the model's predicting accuracy), the Shapley Additive exPlanations (SHAP) package in Python [53] was used to calculate the feature importance scores. SHAP is a game theory approach employed to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values [54] from game theory and related extensions. Shapley values provide a way of fairly distributing the payouts among the feature values, overcoming the issue of payout dependence on the sequence of features. It assigns each feature an importance value for a particular prediction.

As shown in Fig. 3, once a Shapley value is assigned to a single output/prediction, how each feature contributes to the prediction value can be known (i.e., is it positively or negatively correlated with the response). The overall feature importance is illustrated by a standard bar plot that takes the mean absolute value of the SHAP values for each feature [55].

4. Results and discussion

4.1. VIS and NIR components in DNI

In this study, we used the *sklearn, m5py*, and *XGBoost* packages in Python to build regression models for predicting VIS and NIR from broadband DNI. The entire dataset D was split into a training dataset (80 % of D) and a testing dataset (20 % of D). Different models were tuned by using cross-validation to find the best training parameters so that each individual model was optimized to fit the training set. After the parameter tuning process, a performance evaluation was conducted to find the best model. The optimal model was further investigated by analyzing its predictions (a simple logistical analysis) and feature importance (using the Python package *SHAP*). The following sections provide the detailed results of this study, beginning with the model performance evaluation.

Table 4
Random forest parameter tuning.

Parameter	Model			
	V1	V2	N1	N2
n_estimators	1788	1577	1577	1155
min_samples_split	5	5	5	5
min_samples_leaf	2	1	2	2
max_features	'sqrt'	'sqrt'	'sqrt'	'sqrt'
max_depth	15	15	21	21
bootstrap	'False'	'False'	'False'	'False'

4.1.1. Model tuning

Before evaluating the models' performances, a parameter tuning process needed to be conducted to ensure that each algorithm reached its optimal ability to predict the response.

For the single regression tree, in order to avoid overfitting, *cost-complexity pruning* was implemented. Then, the best α was chosen, based on the cross-validation score (k-fold cross-validation, with default k = 10). The best α was determined to be 1.214 for modeling V1, 1.012 for modeling V2, 1.3890 for modeling N1, and 0.7741 for modeling N2. For the M5' tree, the use_smoothing and use_pruning parameters were set to be true in the M5Prime function (Python m5py package), so that the algorithm would smooth the conjunction between two neighboring regions; this increased accuracy and pruned the tree to avoid overfitting.

With regards to ensemble methods, the RandomizedSearchCV (n_iter = 100, cv = 10) was used for the random forest parameter tuning and GridSearchCV (cv = 10) for the bagging and XGBoost regressor parameter tuning. In the random forest regressor, n_estimators (i.e., the number of trees in the forest), max_features (i.e., the number of features to consider when looking for the best split), max_depth (i.e., the maximum depth of the tree), min_samples_split (i.e., the minimum number of samples required to split an internal node), min_samples_leaf (i.e., the minimum number of samples required to be at a leaf node), and bootstrap (i.e., whether bootstrap samples were used when building trees) [56] were the parameters that typically needed to be considered for tuning. There were 100 candidates (500 fits) for the potential

Table 5
XGBoost parameter tuning.

Parameter	Model			
	V1	V2	N1	N2
max_depth	6	6	6	6
min_child_weight	5	0	4	2
Gamma	0.1	0.1	0.4	0.3
subsample	1.0	1.0	1.0	0.9
colsample_bytree	0.9	1.0	0.9	0.8
reg_alpha	200	100	200	300

parameters grid in total. After an exhaustive search, the best parameters grid was found, as shown in Table 4. As for the XGBoost regressor, max_depth (i.e., maximum tree depth for base learners), min_child_weight (i.e., the minimum sum of instance weight (hessian) needed for a child), gamma (i.e., minimum loss reduction required to make a further partition on a leaf node of the tree), subsample (i.e., the subsample ratio of the training instance), colsample_bytree (i.e., the subsample ratio of columns when constructing each tree), and reg_alpha (i.e., the L1 regularization term on weights) [57] were the parameters tuned for optimal model training. The results are shown in Table 5. Note that in the XGBoost parameters tuning, not all parameters were tuned together because of limited computational power, which may have had a certain influence on the final results.

4.1.2. Performance evaluation

To evaluate the prediction models' performances, we performed two tests. First, prediction accuracies were measured by three widely accepted statistical indictors: root mean square error (RMSE), mean absolute error (MAE), and R-squared (R^2). Second, the reliability levels of the regression models were screened by performing a heteroscedasticity test (i.e., residual analysis); if the residuals became more spread out at higher fitted values (i.e., the residuals were not equally distributed at each predictor level), then there was a high probability that heteroscedasticity (i.e., non-linearity, unequal error variances, and outliers) might be present in the regression model (i.e., the regression model was unreliable).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{n}} \tag{9b}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{10b}$$

Table 6
Model comparison for VIS/DNI.

Model	Algorithm	Training			Cross-Validation			Testing		
		RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
V1	DT	15.932	11.672	0.986	24.400	14.043	0.967	20.929	13.030	0.976
	M5′ tree	14.671	7.639	0.988	19.911	10.537	0.977	19.014	10.108	0.981
	RF	7.139	2.987	0.997	18.575	10.023	0.981	18.307	9.701	0.982
	XGBoost	0.001	0.0006	0.999	18.796	9.715	0.981	18.080	9.192	0.983
V2	DT	16.001	11.774	0.986	25.357	13.826	0.973	27.416	14.105	0.959
	M5′ tree	15.296	8.125	0.988	18.589	10.801	0.979	20.977	10.690	0.976
	RF	0.352	0.148	0.999	18.540	10.385	0.985	21.183	12.062	0.980
	XGBoost	0.001	0.0008	0.999	19.681	10.575	0.976	18.985	10.181	0.980
N1	DT	16.405	12.073	0.992	23.177	13.692	0.984	23.425	13.634	0.983
	M5′ tree	14.159	8.015	0.994	18.589	10.666	0.989	21.331	10.828	0.986
	RF	6.583	1.926	0.999	17.809	9.768	0.990	19.905	9.804	0.988
	XGBoost	7.361	4.836	0.998	16.980	8.597	0.991	15.868	8.383	0.993
N2	DT	15.350	11.334	0.993	23.323	13.353	0.984	24.555	13.436	0.982
	M5′ tree	15.231	8.458	0.993	18.589	11.047	0.989	21.115	10.728	0.986
	RF	6.908	2.285	0.998	18.452	10.370	0.990	20.231	10.331	0.987
	XGBoost	8.067	5.315	0.998	17.733	9.732	0.989	17.606	9.506	0.990

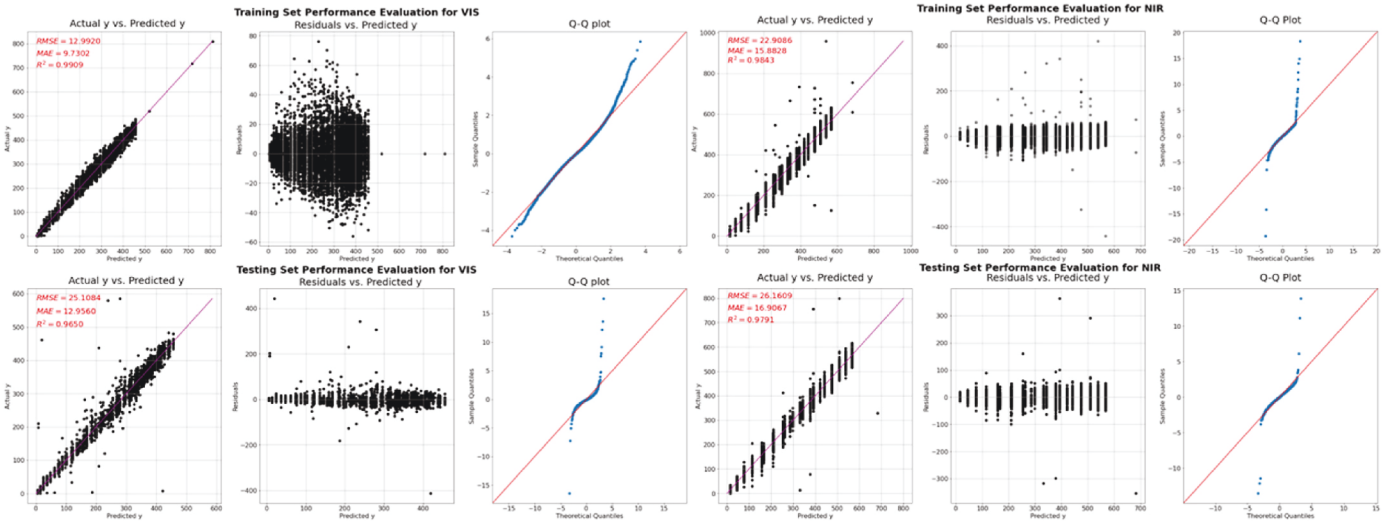

Fig. 4. *Cost complexity tree* residuals analysis (left: VIS model; right: NIR model).

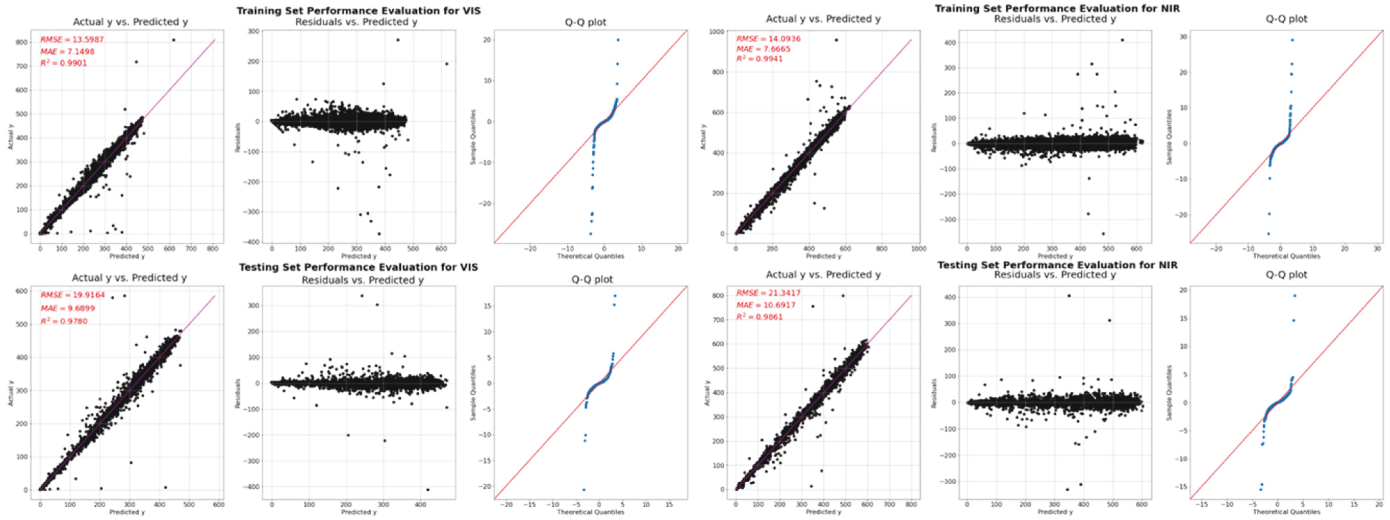

Fig. 5. *M5′ tree* residual analysis (left: VIS model; right: NIR model).

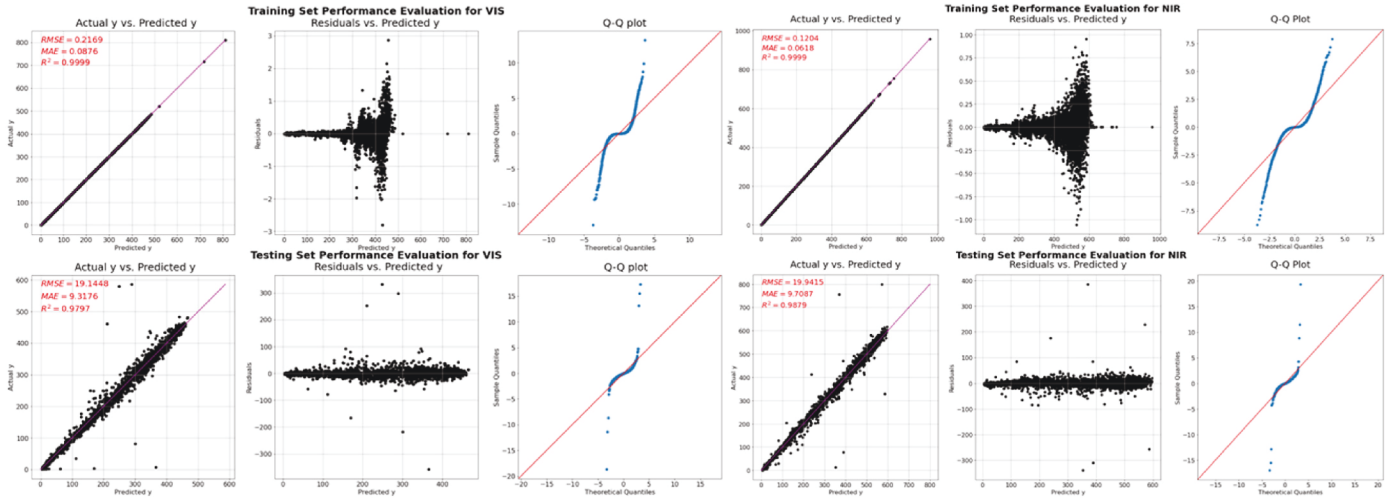

Fig. 6. *Random forest* residual analysis (left: VIS model; right: NIR model).

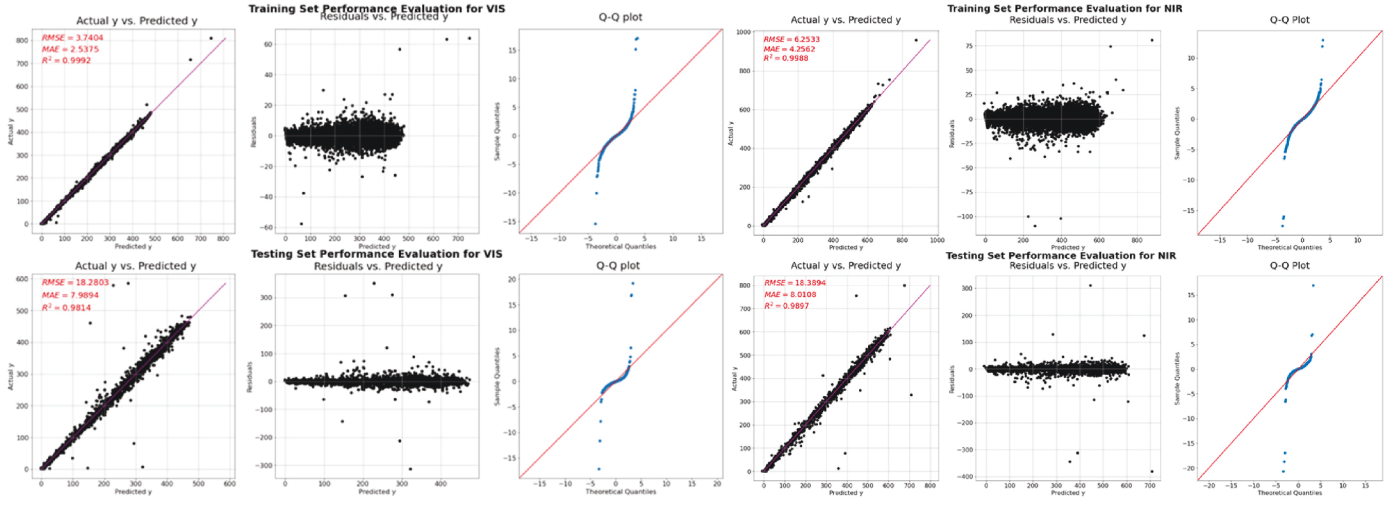

Fig. 7. *XGBoost* residual analysis (left: VIS model; right: NIR model).

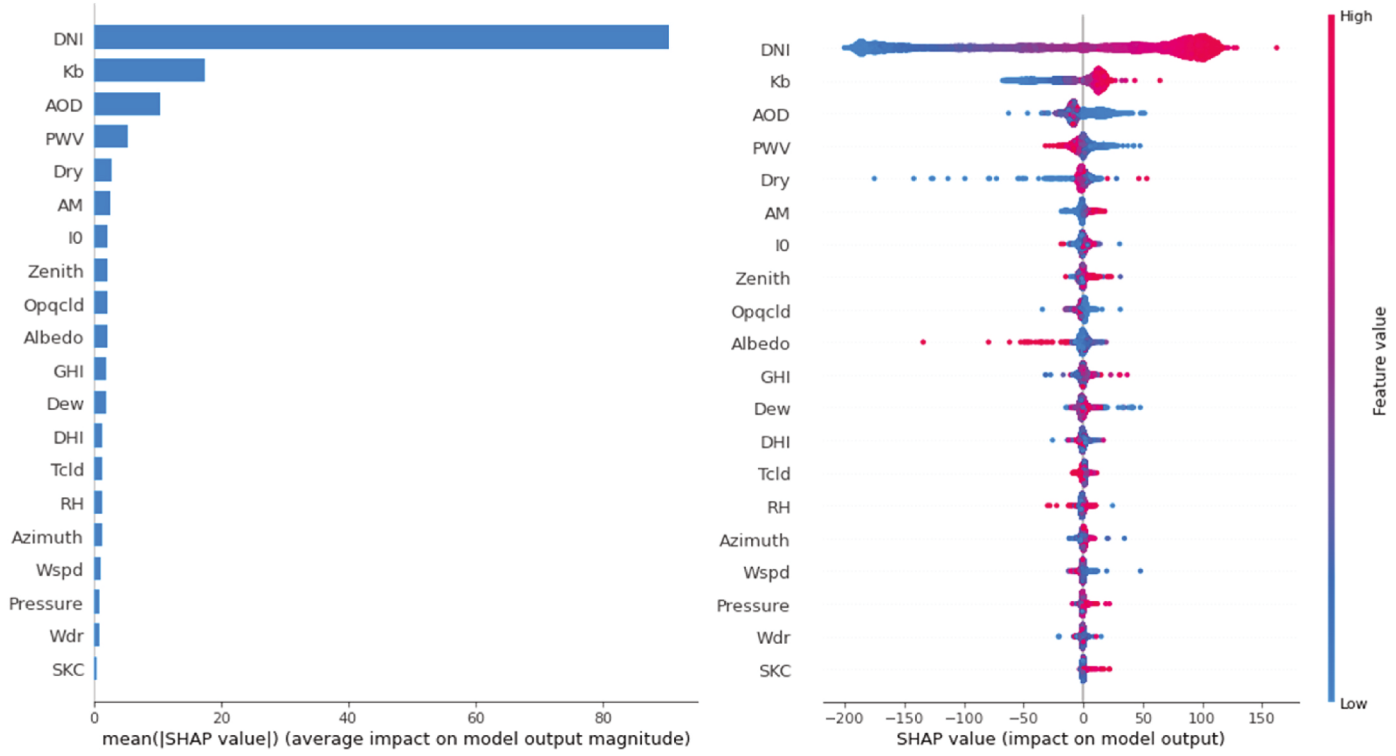

Fig. 8. XGBoost feature importance for V1.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y_i})^2} \tag{11b}$$

Table 6 summarizes all models' prediction accuracies for the training and testing sets (training set accuracies were evaluated using 10-fold cross-validation). Among all the regression methods, XGBoost had the lowest RMSE and MAE values and the highest R^2 value for the training datasets, followed by random forest, M5′ tree, and cost complexity tree. For the testing datasets, XGBoost also produced the best results, in which the RMSE equaled 18.280 and 18.390 for VIS and NIR, respectively. The MAE values were 7.990 for VIS and 8.011 for NIR. The R^2 equaled 0.981 for VIS and 0.990 for NIR. The least accurate regressor was the cost complexity tree, as it produced nearly 1.6 times the RMSE value than did XGBoost. It should be noted that when cross-validation was not performed, the random forest regressor could have extremely high accuracy (wt. 0.217 and 0.120 RMSE for VIS and NIR, respectively) for the training dataset, though this was not the case for the testing dataset. For the models' reliability tests (see Figs. 3–7), it was obvious that the random forest regressor had conical-shaped residuals versus the prediction plots for both the VIS and NIR models, indicating that heteroscedasticity existed. For the M5' tree and XGBoost regressors, the residuals "bounced randomly" around the zero-residual line and roughly formed a horizontal band. The sampled quantiles were similar to the theoretical quantiles throughout the whole prediction space, suggesting that the variances in the error terms were equal and errors were normally distributed. Thus, these two regressors were reliable for predictions. To summarize, although the random forest regressor could produce highly accurate results for the training dataset, it might have been over-fitted so that it was untrustworthy when used for independent testing. XGBoost, however, produced the least errors in predicting the validation datasets and showed its reliability through the heteroscedasticity test. Thus, it was trustworthy and selected to serve as our model algorithm.

After evaluating the model performances and determining the best model, there was one more step before the process could be used for DNI decomposition: checking the predicted VIS and NIR values in case of errors. To do so, a simple logistics test was performed, obeying the following conditions: (1) whether predicted VIS and NIR values were positive; (2) whether the sum of predicted VIS and NIR values at each timestamp was smaller than the DNI value at that timestamp; and (3) whether the sum of the predicted VIS and NIR values at each timestamp was greater than 90 % of the DNI value at that timestamp. The results showed that there were 11 data points that had negative predicted VIS values and 18 with negative NIR values in the training dataset. In the validation dataset, six data points had less-than-zero predicted VIS values, and three had less-than-zero predicted NIR values. Most of the negative predictions occurred when the DNI values were small. These non-positive predictions needed to be fixed to satisfy the real-world situation. For the second condition, 548 data points disagreed with Condition 2 (i.e., the sum of the predicted VIS and NIR values being greater than the corresponding DNI value) in the training dataset, with an average deviation of 11.70 %. A total of 186 data points disagreed with Condition 2 in the testing dataset, with an average deviation of 9.80 %. A large deviation always arose with a small DNI value (lower than 10 Wh/m^2). Sometimes, a DNI value of 1 Wh/m^2 had a deviation of nearly 468 %, as compared to the predicted VIS + NIR at that timestamp. This indicates the model's deficiency at lower DNI values. It is also worth noting that in these extreme cases, NIR predictions are always strongly overestimated, while VIS predictions are underestimated. As for the third condition, there were 128 data points that predicted VIS + NIR values less than 90 % of the corresponding DNI in the training datasets. In the validation datasets, 31 data points had predicted VIS + NIR values less than 90 % of the total corresponding DNI. Any predictions that disobeyed Condition 3 needed to be fixed so that all VIS + NIR predictions met the greater than 90 % DNI rule since DNI will normally have greater than 90 % energy stored in the VIS and NIR spectrums.

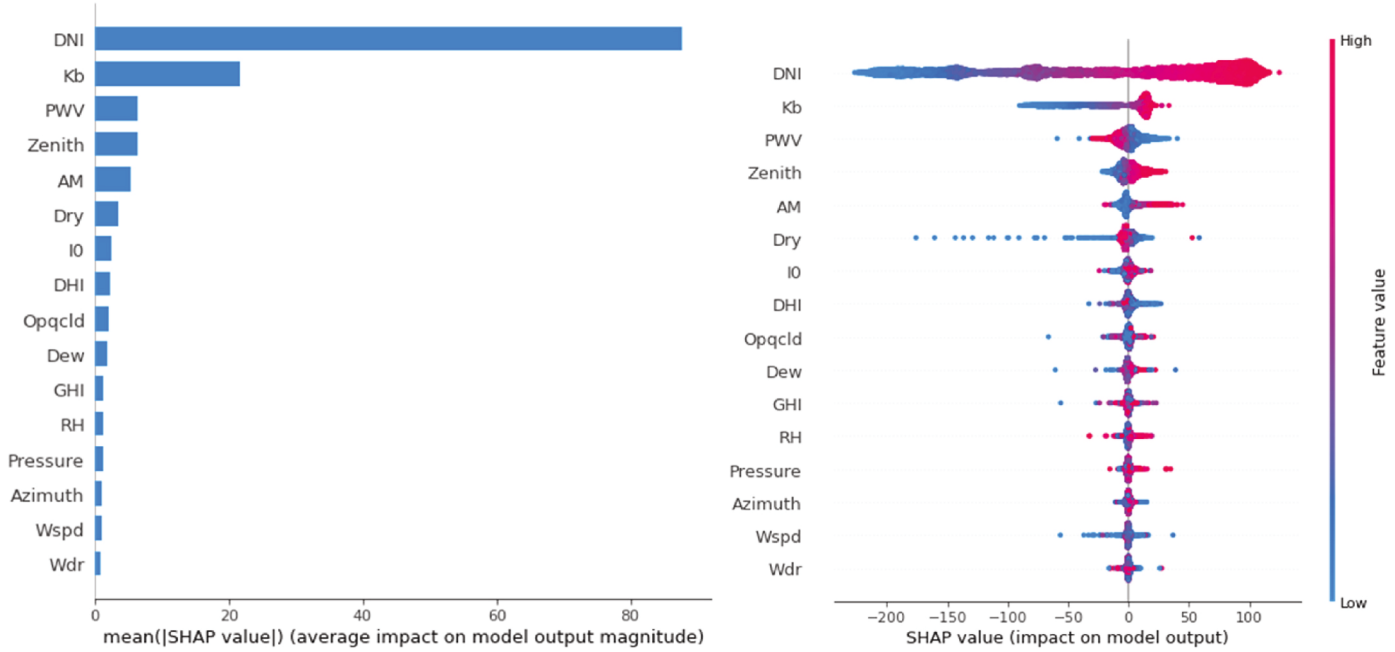

Fig. 9. XGBoost feature importance for V2.

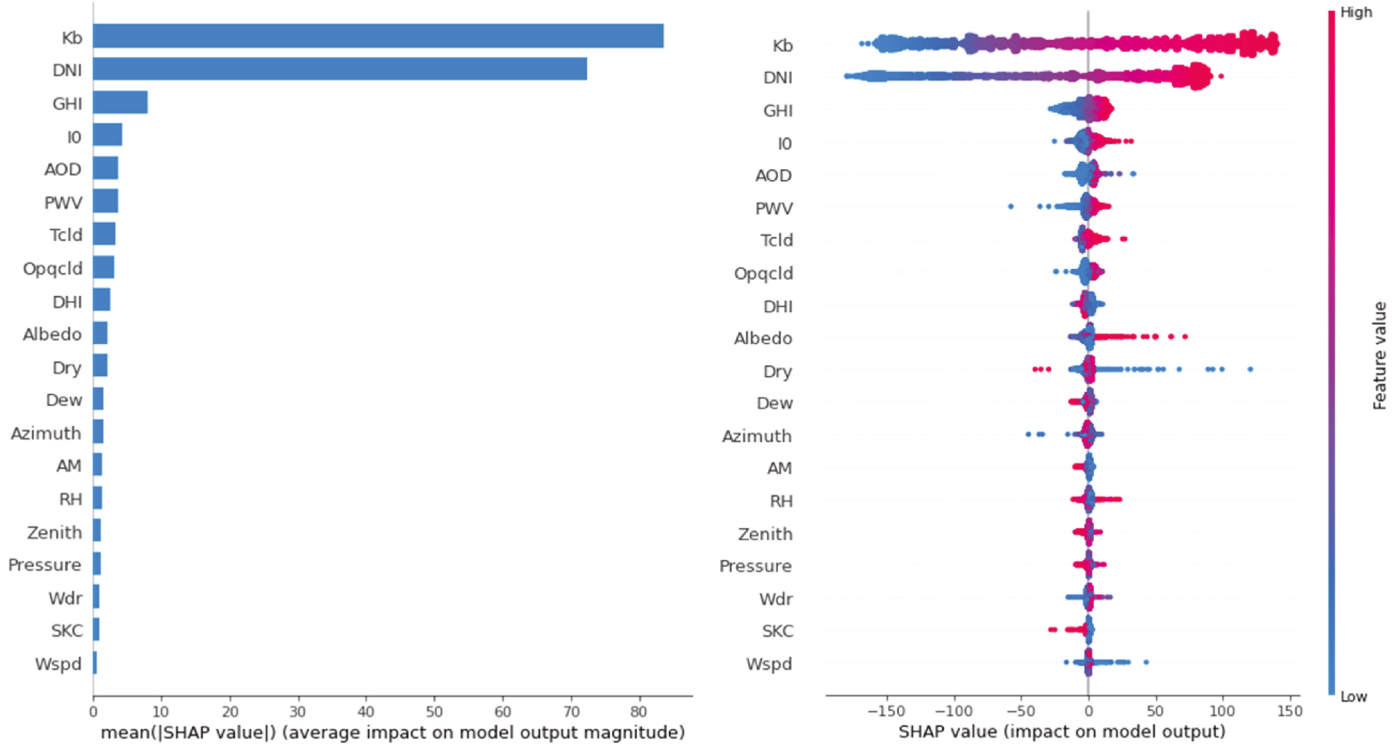

Fig. 10. XGBoost feature importance for N1.

4.2. Discussion

4.2.1. Feature importance

From the model performance evaluations, XGBoost was clearly the best algorithm for decomposing broadband DNI. However, the mechanism behind which weather affects narrowband DNI remains unclear. We tried to determine this through feature importance and interaction analysis based on the XGBoost model. For the feature importance analysis, the SHAP values were calculated for all sampled features on each training point and then averaged to obtain the overall feature importance score. Features with higher absolute SHAP values contributed more to the output and thus were relatively more important. Negative SHAP values meant that the feature negatively affected the output (i.e., the output was lower than average) while positive values had a counteraction effect on the output. As illustrated in Figs. 8–11, DNI and K_b were two of the most important features for both the VIS and NIR decomposition models. AOD played a more important role in the VIS model as compared to the NIR model, while GHI affected NIR more

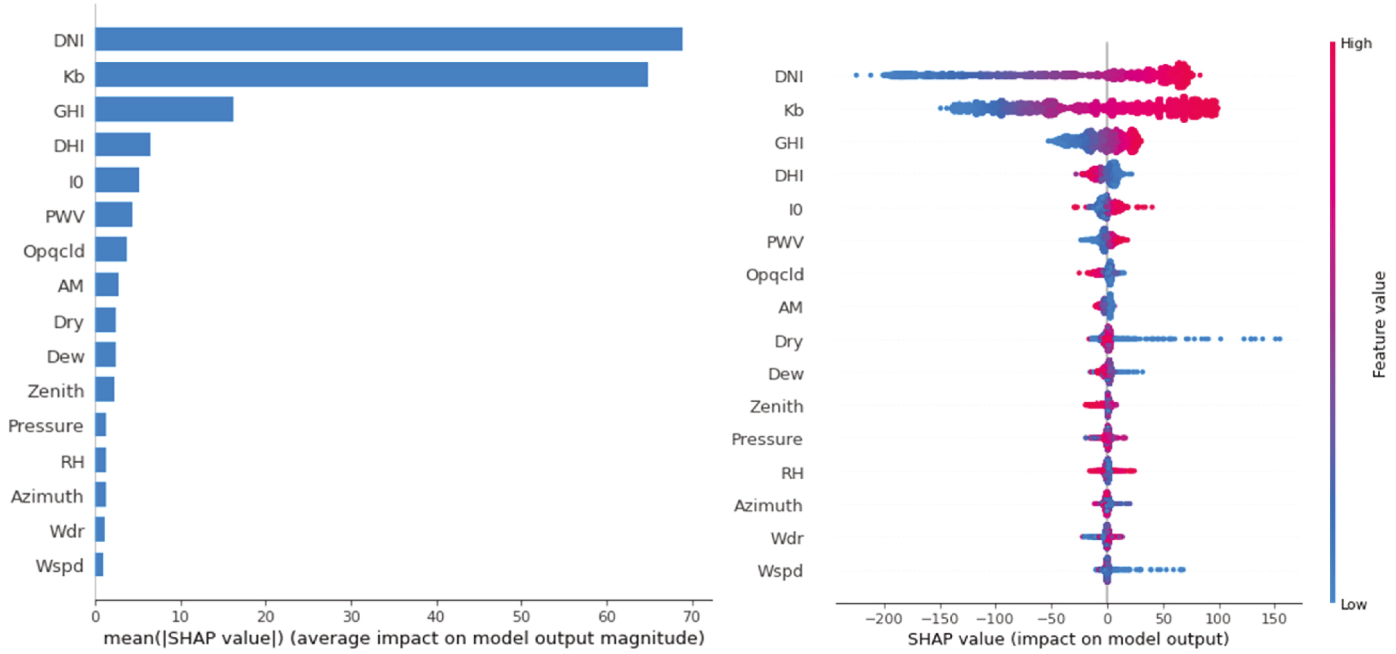

Fig. 11. XGBoost feature importance for N2.

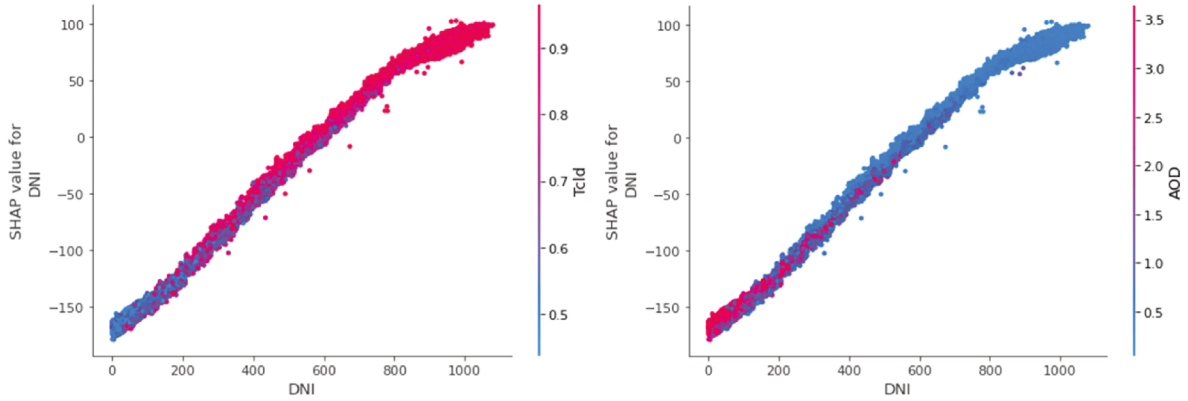

Fig. 12. DNI dependence plots (left: DNI vs. Tcld plot in VIS decomposition; right: DNI vs. AOD plot for NIR decomposition).

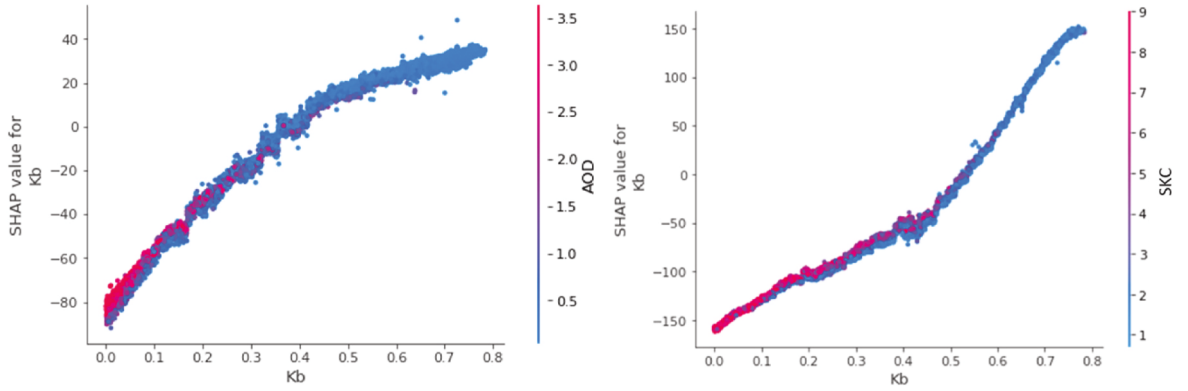

Fig. 13. K_b dependence plots (left: K_b vs. AOD in VIS decomposition; right: K_b vs. SKC in NIR decomposition).

than VIS. PWV and Opqcld also exhibited a certain level of feature importance in the prediction of the VIS and NIR components of DNI. Relatively speaking, the typical meteorological parameters (e.g., dry bulb temperature, wind, and relative humidity) often considered in solar architecture design and analysis have less important impacts (i.e., lower SHAP values) on prediction.

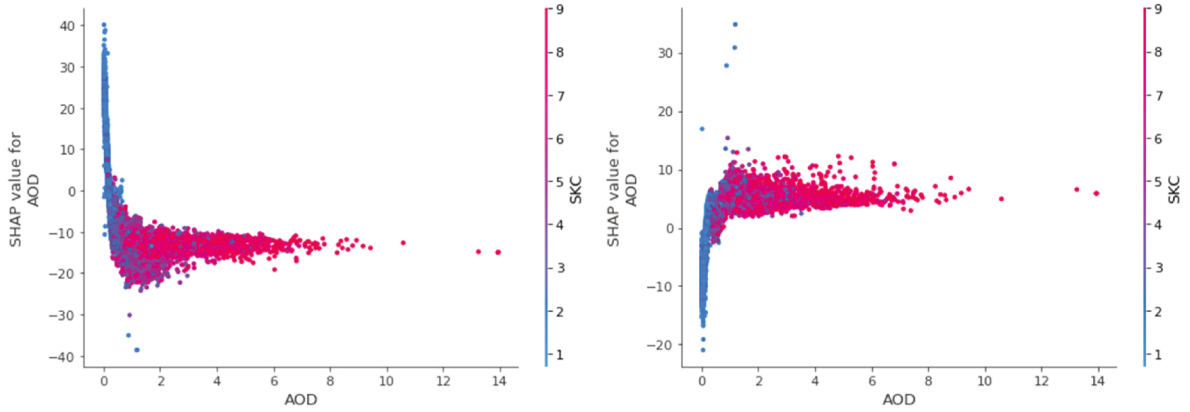

Fig. 14. AOD dependence plots with SKC as the interaction feature (left: AOD vs. SKC in VIS decomposition; right: AOD vs. SKC in NIR decomposition).

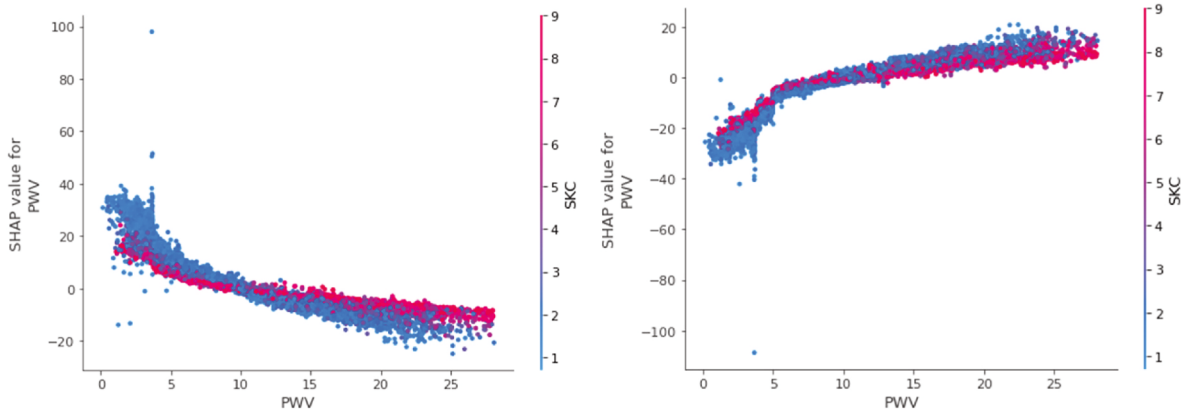

Fig. 15. PWV dependence plots with SKC as the interaction feature (left: PWV vs. SKC in VIS decomposition; right: PWV vs. SKC in NIR decomposition).

4.2.2. Interaction analysis

After identifying the key influential factors for DNI solar decomposition, the next step was to determine their relationships to the prediction of and correlations with other parameters. To achieve this, SHAP dependence plots for the key influential factors (identified above) with the interaction parameters were produced and are described below.

- It can be inferred from Figs. 12 and 13 that both DNI and K_b were positively and linearly correlated with VIS and NIR in DNI. Comparatively, the linear relationship and impacts were slightly stronger in the VIS prediction than in the NIR prediction. In theory, K_b and DNI have strong interactions with sky conditions, as represented by features including SKC, Opqcld, AOD, and Tcld. SKC, Opqcld, and Tcld are inter-correlated with one another, as well. Usually, a higher Opqcld means a higher SKC and thus lower Tcld. These sky-cover parameters affect DNI and Kb in a negative way. With a higher sky cover ratio (i.e., lower cloud transmittance), DNI and K_b will be smaller. These interaction characteristics can also be seen in these two figures.
- Fig. 14 presents the SHAP dependence plot for AOD with the interaction features related to sky cloud coverage. In the left portion referencing VIS, the steep slope when AOD approaches zero indicates that AOD had a strong negative linear relationship to the VIS prediction, meaning that the increase in AOD significantly reduced the

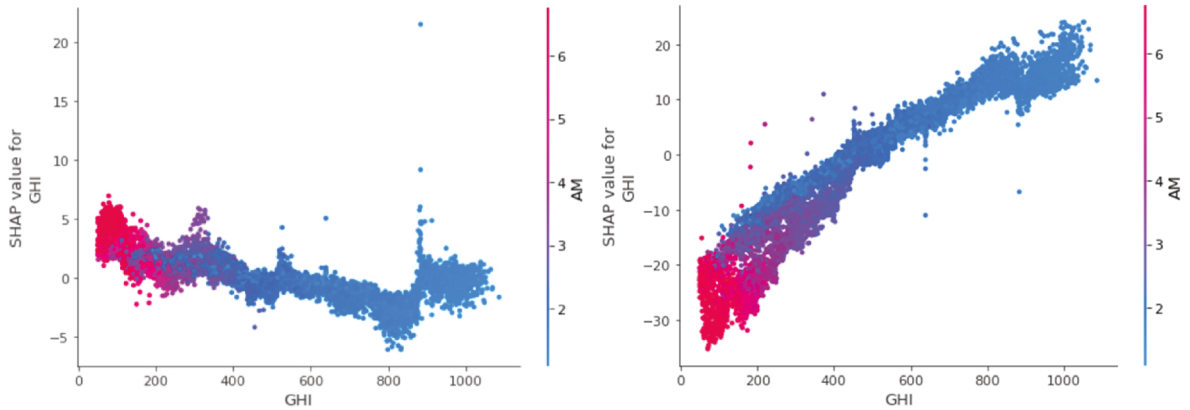

Fig. 16. GHI dependence plots with AM as the interaction feature (left: GHI vs. AM in VIS decomposition; Right: GHI vs. AM in NIR decomposition).

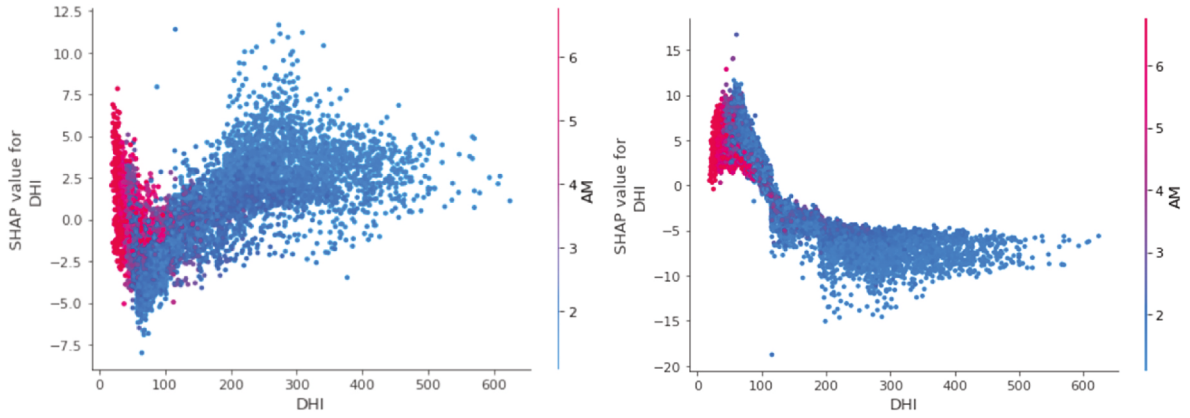

Fig. 17. DHI dependence plots with AM as the interaction feature (left: DHI vs. AM in VIS decomposition; right: DHI vs. AM in NIR decomposition).

VIS quantity. Regarding AOD's effect on NIR (see Fig. 14, right), AOD seems to have had little impact on the infrared. AOD's influence became much weaker as it got even larger. In such situations, the sky cloud coverage is usually also high. In other words, when more portions of the sky are covered by clouds, a greater amount of aerosol may interfere with both visible and infrared solar light, but there is no clear relationship in-between.

- PWV is another important feature of both VIS and NIR. It has a negative relationship with VIS but a positive relationship with NIR (see Fig. 15). PWV is the depth of water in the atmospheric column when all the water in that column is condensed and precipitated. In theory, an increase in PWV should reduce both the VIS and NIR from the sun, but this is different from the relationships exhibited in Fig. 15. With the interactive feature of the sky cover ratio (SKC), the relationships became clearer. When SKC was high (see the red dots), referring to an overall cloudy sky, PWV had little influence on the VIS and NIR of DNI. However, such relationships were more significant when the SKC variable was at a low level (see the blue dots), in which a clear sky was the dominant situation. In general, column water vapor content in the clear sky is significantly lower than that in the cloudy sky. Thus, it seems that the only reason causing the PWV increase was related to the formation of thick partial cloud cover. This type of cloud tends to have a strong reflective capacity for VIS and causes its' reduction in DNI [58]. However, PWV's enhancement effect on NIR was unexpected, as past studies have shown that longer wavelengths (i.e., near-infrared light) show a reduction when PWV increases [29]. This phenomenon may be due to "cloud enhancement," in which scattering from the clouds around the position of the sun disk may be enhanced by 20 %–30 % from solar radiation [59]. This unexpected feature related to the PWV-NIR relationship requires further investigation in the meteorology domain.
- Figs. 16 and 17 present the influence of GHI and DHI on decomposing components, respectively. Overall, both GHI and DHI had a greater influence on NIR in DNI than VIS. Also, there existed a positive, linear correlation between GHI and the amount of NIR in DNI, while DHI exhibited a non-linear but generally negative correlation with NIR in DNI. This meant that DNI tended to contain more NIR when GHI was high, while NIR tended to be less in DNI when DHI was high. This notion is basically aligned with the fact that most NIR comes from direct solar radiation, rather than reflected and scattered solar radiation. Conversely, these two features had less of an impact on the VIS portion of DNI, as shown on the left sides of Figs. 16 and 17. Comparatively, GHI had a relatively clearer negative relationship with VIS. In other words, with a higher GHI, the VIS portion decreased and the NIR portion increased. Additionally, it can be seen from these two figures that GHI and DHI both had a negative correlation with AM, as AM represented the distance of the atmospheric depth through which the solar irradiation had to travel. Thus, the greater the air mass is, the less solar irradiation (such as GHI and DHI) will reach the ground.

5. Application on performance analysis of transparent solar cells

The GHI, DNI, and DHI solar decomposing models have been seamlessly integrated into an executable web application using the Python *streamlit* package. Once users select their desired solar component, they will be prompted to upload a weather file (e.g., TMY file) and specify its location. SolarDecomp will then autonomously employ a pre-trained XGBoost model to perform the decomposition process. The result will be a modified weather file where the original broadband GHI, DNI, and DHI values have been replaced with the corresponding narrowband solar components (VIS or NIR). The final output will consist of two new weather files labeled TMY_VIS and TMY_NIR. With such new weather files, advanced envelope researchers, designers, and engineers are now able to incorporate narrowband solar components when simulating building solar devices.

The introduction section highlighted the recent emergence of transparent solar cells, a technology that promises to harness solar energy without disrupting the aesthetics or functionality of building envelopes. Unlike traditional opaque solar cells, transparent ones, however, typically demonstrate lower power conversion efficiency (PCE) at present. This phenomenon arises from the inherent compromise between transparency and light absorption; transparent solar cells must delicately balance the need to allow visible light to pass through while simultaneously capturing sufficient solar energy. The materials and design paradigms currently in use for transparent solar cells are still in the developmental stages. Still, researchers are engaged in vigorous efforts to boost their efficiency. While the PCE of transparent solar cells currently falls short when compared to opaque alternatives, there is a promise of significant enhancement in their efficiency as advancements are made in the realms of materials, device architecture, and manufacturing techniques. When it comes to evaluating the performance of transparent solar cells, the standard approach involves the use of rated PCE under Standard Testing Conditions (STC: 1000 W/m^2, 25 °C, spectral distribution according to IEC 60904-3). Yet, the accuracy of such assessments may be susceptible to compromise, given the wavelength selectivity feature of transparent solar cells in practical applications (i.e., they absorb light selectively within specific wavelength ranges while permitting other wavelengths, particularly visible light, to pass through). Some researchers have studied the impact of solar spectral irradiance on the yield of different PV technologies, for instance, Daniela et al. proposed a method to gauge the energetic spectral impact by calculating the spectral mismatch factor (MM) between the actual, measured spectral response of various PV technologies and their standard, referenced spectral response in Freiburg im Breisgau,

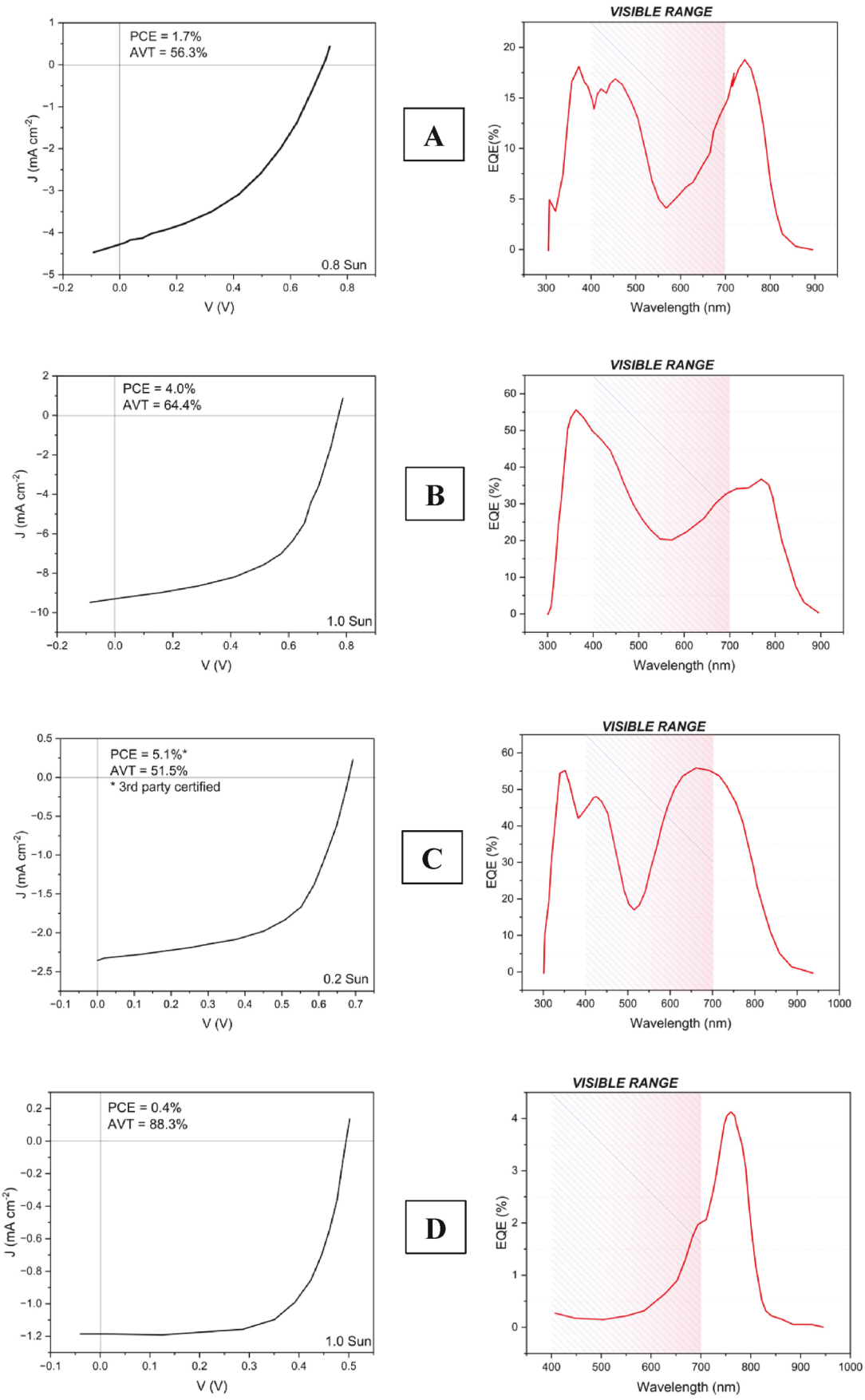

Fig. 18. Wavelength-selective TPV and LSC with >50 % AVT

Table 7
PCE values in broadband, VIS, and NIR.

TPV Device	$PCE_{overall}$	PCE_{VIS}	PCE_{NIR}
A	1.7 %	1.9 %	1.4 %
B	4.0 %	6.5 %	3.5 %
C	5.1 %	7.0 %	4.6 %
D	0.4 %	0.28 %	0.58 %

Germany [60]. Jessen et al. proposed new standardized spectra for DNI and Global Tilted Irradiance (GTI), in which different climatic and atmospheric conditions were considered, to address the spectral mismatch of current solar devices [61]. Despite the high accuracy level these works possessed, they require detailed, local spectral irradiance data to derive the accurate spectral mismatch factor, which is hard to get and lacks applicability to building solar simulation software. In light of this, this research puts forward a new approach to generating narrowband solar components to bolster the performance evaluation of spectral-selective solar cells, without the need to measure the actual spectral irradiance in specific locations. This advancement offers more convenient and effective analytics, thereby aiding in the informed selection, deployment, and integration of solar cells on building envelopes.

In this study, we demonstrate the usage of our model by employing NIR-selective transparent photovoltaic (TPV) solar cells in a building-integrated photovoltaic (BIPV) system. Fig. 18 illustrates the implementation case study, featuring four types of NIR-selective TPVs: (a) a TPV device based on small molecules (12-cells, series-integrated mini-module), (b) a high-efficiency single-junction polymer-based wavelength-selective TPV, (c) a series-integrated TPV module, and (d) an efficient, NIR-absorbing transparent LSC [62]. The PCE of these TPV solar cells ranges from 0.4 % to 5.1 %, with an average visible transmittance exceeding 50 %. It is important to note that the PCE measurements/computations of these solar cells were conducted under the AM1.5 standardized solar spectrum and subsequently evaluated and reported for potential applications using simulations with broadband solar radiation data. Instead of relying solely on an overall PCE value to calculate the final power output, we derived separate PCE values for the visible (VIS) range and the near-infrared (NIR) range. These separate PCE values were obtained using the listed functions and the reported spectral PCE values of the four solar cell types, where J_{sc} is the short-circuit current, $\varnothing(\lambda)$ is the photon flux at wavelength λ, q is the elementary charge, *Pmax* is the maximum power, *Pin* is the input power, *Voc* is the open-circuit voltage and *FF* is the fill factor for solar cells. Table 7 reported the PCE_{VIS} and PCE_{NIR} for all four kinds of TPV solar cells.

$$J_{sc-range} = \int (EQE(\lambda) * \varnothing(\lambda) * q) d\lambda \tag{10c}$$

$$P_{\max-\text{range}} = J_{sc-range} * V_{oc} * FF \tag{11c}$$

$$PCE_{range} = P_{\max-\text{range}} / P_{in-range} \tag{12}$$

5.1. Solar cell performance analysis between the broadband and narrowband solar data

A building solar irradiation analysis was carried out in ClimateStudio, Rhino, using the new TMY_VIS and TMY_NIR as imported weather files. Fig. 19 compares traditional broadband solar irradiation analysis, conducted with the original weather file, with two narrowband radiance analyses performed using the newly-created weather files. The total solar irradiation on all surfaces amounted to 883 kWh/m^2-yr for the broadband solar irradiation analysis. In contrast, the solar irradiation values in the VIS range were 436 kWh/m^2-yr and in the NIR range 358 kWh/m^2-yr. Subsequently, the power outputs for four different TPV solar cells were computed by multiplying the solar irradiation value with the respective PCE in the associated ranges. The energy output of these four solar cells in the broadband, VIS, and NIR ranges is summarized in Table 8. The re-calculated overall output, obtained by summing the output of VIS and NIR, is also included in the last column.

From this table, it can be inferred that when using the newly developed weather files featuring narrowband solar radiation data, the total energy output of all four solar devices differed (ranging from 4.3 to 15.7 %) from the energy output calculated using traditional weather files with broadband solar radiation data. Notably, for the TPV solar cell – B, the difference percentage was approximately 15.7 %. This can be attributed to the significant variance in its PCE between the VIS and NIR regions (6.5 % vs. 3.4 %). In essence, basic solar simulations using conventional weather files, which only include broadband solar radiation data, might either underestimate or overestimate the energy output. This discrepancy can result in significant errors, especially if the wavelength-selectivity feature of the solar devices is more pronounced.

5.2. Solar cell selection and envelope integration consideration

The integration of TPV solar cells into the building envelope is a complex process that demands a thorough analysis from two primary perspectives - visible light transmittance and energy output. This complexity arises from the functional and aesthetic demands placed on modern architectural design, where energy efficiency must coexist with a pleasant visual environment and electrical lighting for energy-saving purposes. In general, TPV solar cells are often compatible with fenestration systems due to their suitable visible light transmittance. In essence, their semi-transparent nature allows for the passage of a significant portion of visible light, contributing positively to the indoor lighting environment. This factor is crucial in the evaluation of potential envelope surfaces for TPV solar cell integration.

A primary condition for integration is that the envelope surface receives strong incident VIS and NIR solar radiation. Surfaces that meet this requirement are more applicable for integration, as they provide both aesthetic and functional benefits. Visible light that passes through can be used for indoor daylighting, enhancing the visual environment and contributing to lighting energy savings. Simultaneously, the NIR component can be harnessed and converted into electrical energy, contributing to the building's overall energy efficiency.

In this case study, the rooftop area, as denoted by the blue circles (shown in Fig. 20), meets these criteria. This area can, therefore, be considered a potential zone for the integration of both skylights and TPV solar cells, maximizing the benefits of natural lighting and solar energy conversion. Secondary envelope surfaces, such as the south and west facades, which receive both moderate and comparably equal levels of VIS and NIR radiation (also denoted by green circles), can also be considered. This balanced incident radiation could facilitate a steady energy output and a stable visual environment throughout the day. Contrarily, if an envelope surface area receives minimal NIR radiation, the integration of TPV solar cells might not be beneficial. The NIR absorption property of these cells is crucial for energy production. Thus, low incident NIR would translate to significantly reduced energy output. In this case study, the north-facing envelopes serve as examples of such surfaces (denoted by red circles).

As such, narrowband solar radiation analysis, involving separate solar simulations with VIS and NIR, can provide more precise and insightful information for the integration of TPV solar cells. Compared to a broadband solar radiation analysis, this approach enables more nuanced decision-making, taking into account the unique spectral response of TPV cells and the irradiation characteristics of the building envelope. This aids in optimizing both the energy efficiency and visual comfort provided by the building envelope design.

6. Conclusion

This work demonstrated the feasibility and excellent prediction

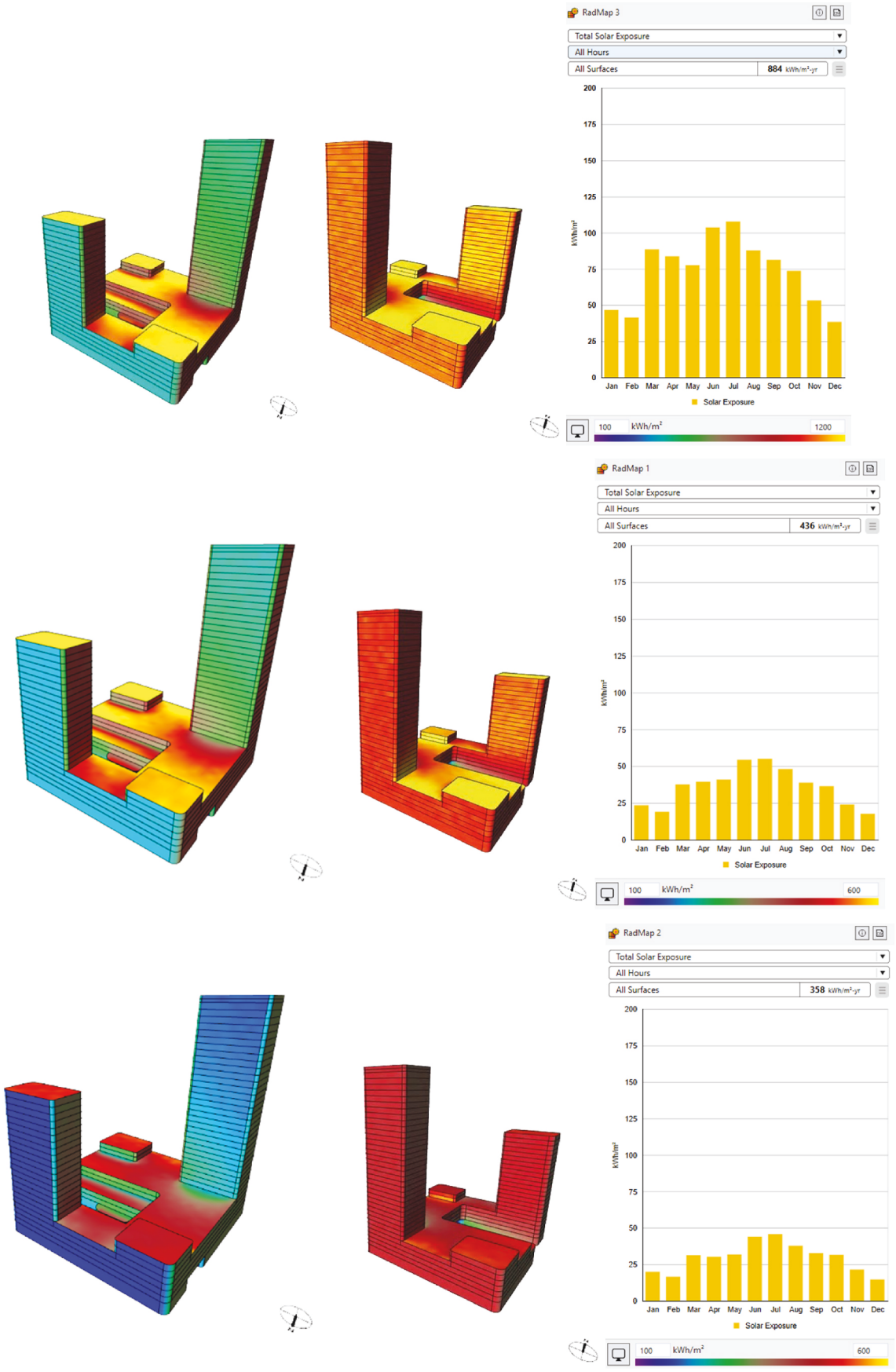

Fig. 19. Traditional broadband radiance analysis (top) vs. Modified narrowband radiance analysis (middle: VIS; bottom: NIR).

Table 8
Power outputs for the four NIR-selective TPVs in broadband, VIS range, and NIR range.

TPV Device	Output in Broadband kWh/m^2-yr	Output in VIS kWh/m^2-yr	Output in NIR kWh/m^2-yr	Sum (VIS + NIR) kWh/m^2-yr	Difference %
A	15.011	8.284	5.012	13.296	**11.4 %**
B	35.320	28.340	12.530	40.870	**15.7 %**
C	45.033	30.520	16.468	46.988	**4.3 %**
D	3.532	1.221	2.076	3.297	**6.7 %**

accuracy of using machine learning methods to decompose the VIS and NIR components of broadband DNI. To develop the decomposition model, several databases of the NREL Solar Radiation Research Laboratory were curated. After the data cleaning process, typical missing data, outliers, and physically erroneous information were removed. The finalized dataset contained 11,862 observations. Easy-to-obtain meteorological parameters were employed to build a generalized model for different locations' solar decomposition. There were 20 predictive variables used in the full estimation model and 15 predictive variables used in the simplified model (with a deficiency of hard-to-get meteorological variables). Four commonly used supervised, non-parametric machine learning algorithms (i.e., Decision tree, M5' tree, random forest, and XGBoost) were used to train the models. The algorithms were compared with one another to evaluate their accuracy and reliability. The results showed that ***XGBoost*** was the most accurate and reliable technique for DNI decomposition, in both full and simplified estimation models, with the lowest RMSE test (18.282 for VIS and 18.389 for NIR) and MAE (7.922 for VIS and 8.011 for NIR) and highest R^2 (0.981 for VIS and 0.990 for NIR). Among all the prediction variables, DNI and K_b were the most important features for both VIS and NIR decomposition. Some atmospheric parameters were also helpful when making predictions, such as AOD, PWV, and total and opaque sky cover. Through feature importance and interaction analyses, some hidden correlations between predictors and other predictors and predictors and responses became clear, which will help in determining the key factors affecting solar spectral irradiance on the ground.

This research significantly contributes to the advancement of solar technology by leveraging machine learning algorithms to transform conventional weather files, which contain broadband solar data, into more specific files incorporating direct normal solar spectral components. This innovative approach negates the need for complex and expensive solar spectra measurements. Instead, it leverages easily accessible and readily available meteorological data, providing a more efficient and cost-effective approach. Importantly, this methodology has a substantial potential for enhancing the design and efficiency of wavelength-selective solar devices such as window systems and TPV solar cells. By providing a more detailed analysis of solar spectral components, our approach can enable a more accurate prediction of these devices' performance, leading to more informed decisions in the design and integration of these systems into building envelopes. The implication is a potential improvement in both the energy efficiency and aesthetic value of buildings.

Nevertheless, despite the significant strides made in this research, there are areas yet to be fully explored. One such area is solar spectra analyses on tilted and vertical surfaces. Existing solar modeling algorithms, such as the Perez and Liu-Jordan models, could potentially be employed for computing incident solar spectra radiation on these surfaces. We aim to delve into these possibilities in future research. Additionally, our model has not been tested under varying climatic conditions due to the unavailability of long-term solar spectral data from different locations. Current models could also confront some limitations, such as the lack of certain predicting variables (e.g., AOD and PWV may not be available in some weather files). Consequently, further development and validation tests under different geographical locations, climate zones, and solar spectra are essential. This will aid in establishing the universality and adaptability of our work, thereby broadening its potential application. By filling these research gaps, we anticipate contributing more comprehensively to the field of solar technology, pushing the boundaries of energy-efficient design, and paving the way for a more sustainable future.

Data Availability

Datasets related to this article can be found at https://midcdmz.nrel.gov/apps/sitehome.pl?site=BMS an open online data repository hosted by the National Renewable Energy Laboratory.

CRediT authorship contribution statement

Chenshun Chen: Writing – original draft, Data Modeling, Discussion, Data curation, Data, Formal analysis, Analysis, Investigation, Data, Visualization. **Qiuhua Duan:** Data curation, Data Modeling. **Yanxiao Feng:** Formal analysis, Discussion. **Julian Wang:** Concept, Methodology, Project administration, Data Interpretation, Research Discussion. **Neda Ghaeili Ardabili:** Data curation, Data, Formal analysis, Analysis. **Nan Wang:** Data curation, Data, Formal analysis, Analysis. **Seyed Morteza Hosseini:** Research Discussion. **Chao Shen:** Research Discussion.

Declaration of competing interest

The authors declare that they have no known competing financial

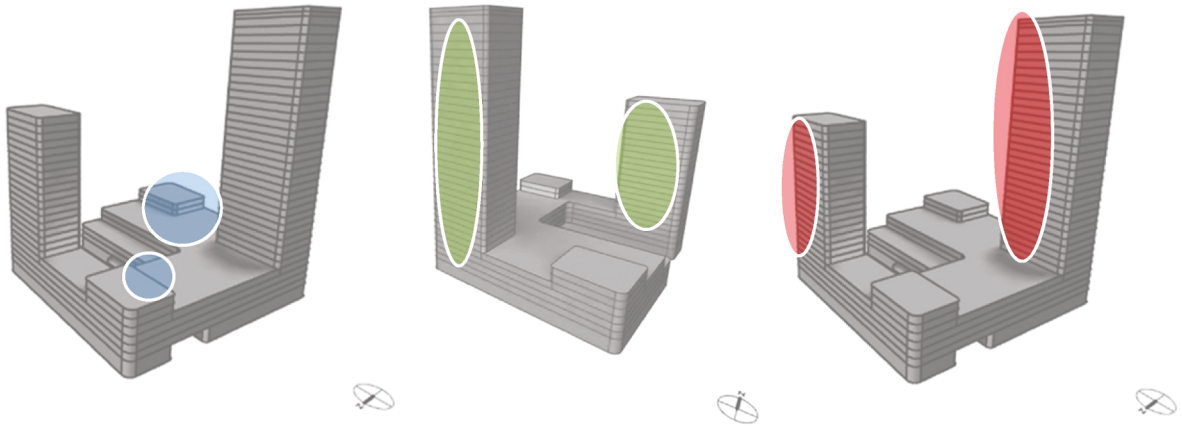

Fig. 20. Strategies of integrating TPVs in the example building.

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project is supported by the NSF awards: # 1953004 and # 2001207.

References

[1] F. Vignola, C. Grover, N. Lemon, A. McMahan, Building a bankable solar radiation dataset, Sol. Energy 86 (8) (Aug. 2012) 2218–2229, https://doi.org/10.1016/j.solener.2012.05.013.
[2] J.E. Roberts, Visible light induced changes in the immune response through an eye-brain mechanism (photoneuroimmunology), J. Photochem. Photobiol., B 29 (1) (1995) 3–15, https://doi.org/10.1016/1011-1344(95)90241-4. Jul.
[3] N. Wang, J. Wang, A spectrally-resolved method for evaluating the solar effect on user thermal comfort in the near-window zone, Build. Environ. 202 (Sep. 2021), 108044, https://doi.org/10.1016/j.buildenv.2021.108044.
[4] M. Anwar Jahid, J. Wang, E. Zhang, Q. Duan, Y. Feng, Energy savings potential of reversible photothermal windows with near infrared-selective plasmonic nanofilms, Energy Convers. Manag. 263 (2022), 115705, https://doi.org/10.1016/j.enconman.2022.115705. Jul.
[5] J. Pu, C. Shen, S. Yang, C. Zhang, D. Chwieduk, S.A. Kalogirou, Feasibility investigation on using silver nanorods in energy saving windows for light/heat decoupling, Energy 245 (Apr. 2022), 123289, https://doi.org/10.1016/j.energy.2022.123289.
[6] Y. Li, X. Huang, H. Sheriff, S. Forrest, Semitransparent organic photovoltaics for building-integrated photovoltaic applications, Nat. Rev. Mater. 8 (Dec. 2022) 1–16, https://doi.org/10.1038/s41578-022-00514-0.
[7] W.-C. Du, J. Xie, L. Xia, Y.-J. Liu, H.-W. Yang, Y. Zhang, Study of new solar film based on near-infrared shielding, J. Photochem. Photobiol. Chem. 418 (Sep. 2021), 113410, https://doi.org/10.1016/j.jphotochem.2021.113410.
[8] L.V. Besteiro, X.-T. Kong, Z. Wang, F. Rosei, A.O. Govorov, Plasmonic glasses and films based on alternative inexpensive materials for blocking infrared radiation, Nano Lett. 18 (5) (May 2018) 3147–3156, https://doi.org/10.1021/acs.nanolett.8b00764.
[9] Q. Gao, X. Wu, T. Huang, Novel energy efficient window coatings based on in doped CuS nanocrystals with enhanced NIR shielding performance, Sol. Energy 220 (1–7) (May 2021), https://doi.org/10.1016/j.solener.2021.02.045.
[10] Q. Xu, et al., Cs0.33WO3 as a high-performance transparent solar radiation shielding material for windows, J. Appl. Phys. 124 (19) (Nov. 2018), 193102, https://doi.org/10.1063/1.5050041.
[11] Q. Duan, Y. Feng, J. Wang, Clustering of visible and infrared solar irradiance for solar architecture design and analysis, Renew. Energy 165 (Mar. 2021) 668–677, https://doi.org/10.1016/j.renene.2020.11.080.
[12] Y. Song, Q. Duan, Y. Feng, E. Zhang, J. Wang, S. Niu, Solar infrared radiation towards building energy efficiency: measurement, data, and modeling, Environ. Rev. 28 (4) (2020) 457–465, https://doi.org/10.1139/er-2019-0067. Dec.
[13] R.E. Bird, A simple, solar spectral model for direct-normal and diffuse horizontal irradiance, Sol. Energy 32 (4) (Jan. 1984) 461–471, https://doi.org/10.1016/0038-092X(84)90260-3.
[14] S. Nann, C. Riordan, Solar spectral irradiance under clear and cloudy skies: measurements and a semiempirical model, J. Appl. Meteorol. Climatol. 30 (4) (Apr. 1991) 447–462, https://doi.org/10.1175/1520-0450(1991)030<0447.SSIUCA>2.0.CO;2.
[15] V. Tatsiankou, K. Hinzer, H. Schriemer, R. Beal, Improved global irradiance decomposition by sky condition classification from measured spectral clearness indices, 47th IEEE Photovolt. Spec. Conf. (PVSC) (2020) 72–76, https://doi.org/10.1109/PVSC45281.2020.9300629. Jun. 2020.
[16] P.G. Kosmopoulos, et al., Dust impact on surface solar irradiance assessed with model simulations, satellite observations and ground-based measurements, Atmos. Meas. Tech. 10 (7) (2017) 2435–2453, https://doi.org/10.5194/amt-10-2435-2017. Jul.
[17] D. Charuchittipan, P. Choosri, S. Janjai, S. Buntoung, M. Nunez, W. Thongrasmee, 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 (Mar. 2018) 175–183, https://doi.org/10.1016/j.renene.2017.10.045.
[18] M. Taylor, P.G. Kosmopoulos, S. Kazadzis, I. Keramitsoglou, C.T. Kiranoudis, Neural network radiative transfer solvers for the generation of high resolution solar irradiance spectra parameterized by cloud and aerosol parameters, J. Quant. Spectrosc. Radiat. Transf. 168 (Jan. 2016) 176–192, https://doi.org/10.1016/j.jqsrt.2015.08.018.
[19] R. Hulstrom, R. Bird, C. Riordan, Spectral solar irradiance data sets for selected terrestrial conditions, Sol. Cell. 15 (4) (Dec. 1985) 365–391, https://doi.org/10.1016/0379-6787(85)90052-3.
[20] C.A. Gueymard, Interdisciplinary applications of a versatile spectral solar irradiance model: a review, Energy 30 (9) (2005) 1551–1576, https://doi.org/10.1016/j.energy.2004.04.032. Jul.
[21] 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, 1, J. Appl. Meteorol. Climatol. 25 (Jan. 1986) 87–97, https://doi.org/10.1175/1520-0450(1986)025<0087:SSSMFD>2.0. CO;2.
[22] C.G. Justus, M.V. Paris, A model for solar spectral irradiance and radiance at the bottom and top of a cloudless atmosphere, J. Appl. Meteorol. Climatol. 24 (3) (Mar. 1985) 193–205, https://doi.org/10.1175/1520-0450, 1985)024<0193:AMFSSI>2.0.CO;2.
[23] I. Ermolli, et al., Recent variability of the solar spectral irradiance and its impact on climate modelling, Atmos. Chem. Phys. 13 (8) (2013) 3945–3977, https://doi.org/10.5194/acp-13-3945-2013. Apr.
[24] C.A. Gueymard, J.A. Ruiz-Arias, Validation of direct normal irradiance predictions under arid conditions: a review of radiative models and their turbidity-dependent performance, Renew. Sustain. Energy Rev. 45 (May 2015) 379–396, https://doi.org/10.1016/j.rser.2015.01.065.
[25] M.A. Hassan, A. Khalil, S. Kaseb, M.A. Kassem, Exploring the potential of tree-based ensemble methods in solar radiation modeling, Appl. Energy 203 (Oct. 2017) 897–916, https://doi.org/10.1016/j.apenergy.2017.06.104.
[26] D.R. Myers, Solar radiation modeling and measurements for renewable energy applications: data and model quality, Energy 30 (9) (2005) 1517–1531, https://doi.org/10.1016/j.energy.2004.04.034. Jul.
[27] F.O. Hocaoğlu, Stochastic approach for daily solar radiation modeling, Sol. Energy 85 (2) (Feb. 2011) 278–287, https://doi.org/10.1016/j.solener.2010.12.003.
[28] M.A. Behrang, E. Assareh, A. Ghanbarzadeh, A.R. Noghrehabadi, The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data, Sol. Energy 84 (8) (Aug. 2010) 1468–1480, https://doi.org/10.1016/j.solener.2010.05.009.
[29] D.R. Myers, Solar Radiation: Practical Modeling for Renewable Energy Applications, first ed., CRC Press, 2017 https://doi.org/10.1201/b13898.
[30] S. Samadianfard, A. Majnooni-Heris, S.N. Qasem, O. Kisi, S. Shamshirband, K. Chau, Daily global solar radiation modeling using data-driven techniques and empirical equations in a semi-arid climate, Eng. Appl. Comput. Fluid Mech. 13 (1) (Jan. 2019) 142–157, https://doi.org/10.1080/19942060.2018.1560364.
[31] G. Szeicz, Solar radiation for plant growth, J. Appl. Ecol. 11 (2) (1974) 617–636, https://doi.org/10.2307/2402214.
[32] 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 (3) (Mar. 2009) 299–309, https://doi.org/10.1016/j.apenergy.2008.04.013.
[33] F. Vignola, GHI Correlations with DHI and DNI and the Effects of Cloudiness on One-Minute Data, 2012.
[34] S. Achleitner, A. Kamthe, T. Liu, A.E. Cerpa, SIPS: solar irradiance prediction system, in: IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks, Apr. 2014, pp. 225–236, https://doi.org/10.1109/IPSN.2014.6846755.
[35] "MIDC: SRRL BMS instruments.", Accessed: Dec. 15, 2022. [Online]. Available: https://midcdmz.nrel.gov/srrl_bms/instruments.html.
[36] "UO SRML: solar radiation basics.", Accessed: Aug. 12, 2022. [Online]. Available: http://solardat.uoregon.edu/SolarRadiationBasics.html.
[37] B. Stafford, Note: right now, the latest commits of Pysolar don't work with Python 2.x, Aug. 12Accessed: Aug. 12, 2022. [Online]. Available: https://github.com/pingswept/pysolar/blob/18b319facc691d6e9eb843bd604ed398a7d46141/doc/index.rst, 2022.
[38] P. Blanc, et al., Direct normal irradiance related definitions and applications: the circumsolar issue, Sol. Energy 110 (Dec. 2014) 561–577, https://doi.org/10.1016/j.solener.2014.10.001.
[39] F. Kasten, A.T. Young, Revised optical air mass tables and approximation formula, Appl. Opt. 28 (22) (Nov. 1989) 4735–4738, https://doi.org/10.1364/AO.28.004735.
[40] J. Li, X. Ge, Q. He, A. Abbas, Aerosol optical depth (AOD): spatial and temporal variations and association with meteorological covariates in Taklimakan desert, China, PeerJ 9 (Jan. 2021), e10542, https://doi.org/10.7717/peerj.10542.
[41] EuroSkyRad - European SkyRad users network, Accessed: Nov. 17, 2022. [Online]. Available: http://www.euroskyrad.net/.
[42] P. Benevides, J. Catalao, P.M.A. Miranda, On the inclusion of GPS precipitable water vapour in the nowcasting of rainfall, Nat. Hazards Earth Syst. Sci. 15 (12) (Dec. 2015) 2605–2616, https://doi.org/10.5194/nhess-15-2605-2015.
[43] E.L. Maxwell, METSTAT—the solar radiation model used in the production of the National Solar Radiation Data Base (NSRDB), Sol. Energy 62 (4) (Apr. 1998) 263–279, https://doi.org/10.1016/S0038-092X(98)00003-6.
[44] H. Beyer, John W. Tukey, Exploratory data analysis. Addison-wesley publishing company reading, mass. — menlo park, cal., London, Amsterdam, DonMills, ontario, sydney, XVI, 688 S.," Biom. J 23 (4) (1977) 413–414, https://doi.org/10.1002/bimj.4710230408, 1981.
[45] A. Muscio, The solar reflectance index as a tool to forecast the heat released to the urban environment: potentiality and assessment issues, Climate 6 (1) (2018) 12, https://doi.org/10.3390/cli6010012. Feb.
[46] 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 (May 2019) 184–195, https://doi.org/10.1016/j.apenergy.2019.03.089.
[47] S. Belaid, A. Mellit, Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate, Energy Convers. Manag. 118 (Jun. 2016) 105–118, https://doi.org/10.1016/j.enconman.2016.03.082.
[48] I.A. Ibrahim, T. Khatib, A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm, Energy Convers. Manag. 138 (Apr. 2017) 413–425, https://doi.org/10.1016/j.enconman.2017.02.006.
[49] L. Zou, L. Wang, L. Xia, A. Lin, B. Hu, H. Zhu, Prediction and comparison of solar radiation using improved empirical models and Adaptive Neuro-Fuzzy Inference

Systems, Renew. Energy 106 (Jun. 2017) 343–353, https://doi.org/10.1016/j.renene.2017.01.042.
[50] F. Sohil, M.U. Sohali, J. Shabbir, "An introduction to statistical learning with applications in R: by gareth james, Daniela witten, trevor hastie, and robert tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7,", Stat. Theor. Relat. Field. 6 (1) (Jan. 2022) 87, https://doi.org/10.1080/24754269.2021.1980261, 87.
[51] Y. Wang, I. Witten, Induction of model trees for predicting continuous classes, in: Induction Model Trees Predict, Contin. Cl., Jan. 1997.
[52] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, Aug. 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785. KDD '16.
[53] S. Lundberg, slundberg/shap, Aug. 17Accessed: Aug. 17, 2022. [Online]. Available: https://github.com/slundberg/shap, 2022.
[54] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowl. Inf. Syst. 41 (3) (Dec. 2014) 647–665, https://doi.org/10.1007/s10115-013-0679-x.
[55] GitHub - slundberg/shap, A game theoretic approach to explain the output of any machine learning model.", Accessed: Nov. 17, 2022. [Online]. Available: https://github.com/slundberg/shap.
[56] "sklearn.ensemble.RandomForestRegressor," scikit-learn, Accessed: Oct. 14, 2022. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.
[57] Python API Reference — xgboost 1.6.2 documentation, Accessed: Oct. 14, 2022. [Online]. Available: https://xgboost.readthedocs.io/en/stable/python/python_api.html.
[58] Clouds & radiation fact sheet, Accessed: Jan. 18, 2023. [Online]. Available: https://earthobservatory.nasa.gov/features/Clouds.
[59] G. Pfister, R.L. McKenzie, J.B. Liley, A. Thomas, B.W. Forgan, C.N. Long, Cloud coverage based on all-sky imaging and its impact on surface solar irradiance, 10, J. Appl. Meteorol. Climatol. 42 (Oct. 2003) 1421–1434, https://doi.org/10.1175/1520-0450(2003)042<1421:CCBOAI>2.0. CO;2.
[60] D. Dirnberger, G. Blackburn, B. Müller, C. Reise, On the impact of solar spectral irradiance on the yield of different PV technologies, Sol. Energy Mater. Sol. Cells 132 (Jan. 2015) 431–442, https://doi.org/10.1016/j.solmat.2014.09.034.
[61] W. Jessen, et al., Proposal and evaluation of subordinate standard solar irradiance spectra for applications in solar energy systems, Sol. Energy 168 (2018) 30–43, https://doi.org/10.1016/j.solener.2018.03.043. Jul.
[62] C.J. Traverse, R. Pandey, M.C. Barr, R.R. Lunt, Emergence of highly transparent photovoltaics for distributed applications, Nat. Energy 2 (11) (Oct. 2017) 849–860, https://doi.org/10.1038/s41560-017-0016-9.