Next Energy 8 (2025) 100320

Contents lists available at ScienceDirect

Next Energy

journal homepage: www.sciencedirect.com/journal/next-energy

Research article

Advancements in solar spectral irradiance modelling for photovoltaic systems: A machine learning approach utilising on-site data

Haoxiang Zhang[a,*], Sunny Chaudhary[a], Carlos D. Rodríguez-Gallegos[b], Tasmiat Rahman[a]

[a] School of Electronics and Computer Science, University of Southampton, Southampton, Hampshire, SO17 1BJ, United Kingdom
[b] Solar Energy Research Institute of Singapore (SERIS), National University of Singapore (NUS), Singapore 117574, Singapore

ARTICLE INFO

Keywords:
Photovoltaics
Solar irradiance
Machine learning
Measuring instruments

ABSTRACT

Energy yield estimation for photovoltaics (PV) plays a crucial role in the growth of renewable energy. To reduce uncertainty in these estimations, having a spectral resolved irradiance is key. In the field of PV, radiative transfer models (RTMs) and spectroradiometers are commonly utilised to determine spectral solar irradiance, which is crucial for assessing spectral effects. However, these methodologies have inherent limitations; RTMs require precise and complex inputs of aerosol and meteorological data, while spectroradiometers entail significant costs. With the advancement of machine learning (ML) techniques, a data-driven spectral irradiance model is proposed in this study, which only requires the global horizontal irradiance (*GHI*) measured by pyranometer and the reference cell as input. Spectral data and meteorological data collected by Solar Energy Research Institute of Singapore (SERIS) at four sites across three continents are used for the training and testing of our models. We examined the viability on spectra modelling of three ML techniques including Long Short-Term Memory networks (LSTM), Random Forest (RF) algorithms and Extreme Gradient Boost (XGBoost). XGBoost achieves relatively good accuracy; additionally, the computational cost is much lower compared to LSTM and RF. The proposed ML model shows an overall R^2 of 0.974 in comparison with 0.646 of the SMARTS model in the spectrum range 350.4–1052.4 nm. The ML models outperform the SMARTS model particularly under intermediate and overcast conditions. We have also shown that a model trained on data from a specific site cannot be effectively applied to other locations.

1. Introduction

Solar irradiance is one of the key factors determining the performance of photovoltaic devices (PV). Most modern PV devices are developed and tested under the AM1.5 STC conditions defined in ASTM G173–23 (ASTM, 2013) [3], which is modelled using SMARTS2 with 1976 U.S Standard Atmosphere [15]. However, the solar irradiance at the location of the actual deployment could be significantly different compared to AM1.5 depending on the composition of the atmosphere, the position of the sun and the angle of incidence (AOI). This leads to a range of spectral effects on PV system components and applications. Dirnberger et al. [10] analysed the impact of solar spectral irradiance on different cell technologies with measured data from Fraunshofer ISE in Germany, demonstrating that the spectral gain varies from 0.6% for copper indium gallium selenide solar cell (CIGS) cell to 3.4% for a-Si cell. In a similar study, [11] with spectra data modelled using SPCTRAL2 [30], reported an efficiency difference ranging from −10 to 15% between different seasons for a-Si cells and 4% for c-Si cells. For more recent tandem solar cells and perovskites solar cells, which show a very different spectral response, the research from Hörantner and Snaith [23] shows that their optimisations and band-gap tuning can be related to regional spectral distribution. It is evident that access to the spectral solar irradiance data is of great importance.

In this research, we use global horizontal irradiance *(GHI)* as the target to measure the performance of the model, since it is the fundamental input for solar decomposition models [34], transposition models [40], or empirical models [4].

The acquisition of solar spectra falls into two categories: field measurement with spectroradiometers and modelling with radiative transfer models. The ground measured data are considered more accurate compared with radiative transfer models. However, the availability of measured spectral irradiance is very limited due to the high cost of these devices [39].

Therefore, site-specific spectra data are commonly generated using radiative transfer models (RTMs), which generate the spectra based on the location-specific information. A range of RTMs have been developed and

* Corresponding author.
E-mail address: haoxiang.zhang@soton.ac.uk (H. Zhang).

https://doi.org/10.1016/j.nxener.2025.100320
Received 31 December 2024; Received in revised form 6 May 2025; Accepted 12 May 2025

have been proven to have precision in atmospheric research, such as SBDART [32], SMARTS [15], Lowtran [24], Hitran [33], LibRadTran [12], FAScode [36], SPCTRAL2 [5]. Generally, the SMARTS model is preferred, as it requires much simpler input and offers fast execution speed, while maintaining 1–2% typical difference from the reference model and 5% overall experimental error when compared with the spectrally resolved measurements [16]. In addition, it can be used to calculate multiple different output variables (43), including direct normal irradiance (*DNI*), *GHI*, diffused horizontal irradiance (*DHI*), albedo, which are commonly used for PV applications. However, like any other RTM, it relies on accurate aerosol and meteorological data input, which can be difficult to obtain. In addition, the SMARTS model is developed and validated for cloudless skies, which generates additional uncertainty when applying under overcast conditions.

In recent years, the advancement of machine learning (ML) techniques is widely applied in solar irradiance predictions. Antonanzas et al. [1] reviews over 70 different solar irradiance prediction techniques, where 53% of the research utilises machine learning, namely, Artificial - (Neural Networks)(A-(NNs) - 24%), k-Nearest Neighbours (k-NN - 18%), Support Vector Machines (SVM - 18%) and Random Forests (RF - 5%). Gupta et al. [17] used auto-seasonal autoregressive integrated moving average (auto-SARIMA) to perform long-term estimation on GHI for the next 5 years. In a more recent publication from Ramadhan et al. [31], for estimating solar irradiance, a comparison between the physical model and the machine learning model has been made utilising RNN, LSTM and Grated Recurrent Unit (GRU), demonstrating that machine learning models outperform physical models; LSTM and GRU were recommended for modelling time-series data with high data volume capacity. Gupta et al. [19] shows that the feature selection (FS) method gives better predictive performance than the feature combination (FC) method for GHI estimation. Successive work [18] shows that extra-trees (ET) outperforms Decision Tree (DT), RF, Gradient Boost (GB), Light Gradient Boost Machine (LGBM) when predicting GHI using FS. Gupta et al. [20] proposed a stacking ensamble (SE) model that integrates a range of models to ET model. It shows that the SE model works better than any individual base models.

However, few studies focus on wavelength-resolved irradiance. Zhang et al. [42] built a low-cost spectrophotometer using data from a few irradiance sensors and then processed using an ANN ML model for a wavelength range of 360–790 nm. In a separate study, Le et al. [26] emphasised the correlation between irradiance at different wavelengths and proposed principal component analysis (PCA) and NN models to recreate the spectra with the irradiance value of representative wavelengths. The success of ML in predicting and analysing data is the motivation for the current study.

Solar spectra estimation using ML techniques is comparatively less explored while utilising multiple spectral wavelengths. Del Rocco et al. [9] demonstrates a method for estimating spectral irradiance using a machine learning regression model and high dynamic range (HDR) sky images, incorporating computer vision techniques, where the target spectrum range is between 350 and 1780 nm. A range of ML regression models were tested and presented, including ET, RF, KNN and linear regression (LNR). It was shown that tree-based model performed better, subsequently, ETR was determined to be the better model. Implementing such models requires complicated training data sets, where a sky scanner is required for spectral data and an all-sky HDR camera is required for sky images. As the trained model is not universal; it needs to be fine-tuned for the site of interest.

In a separate study, del Campo-Ávila et al. [7] developed a data mining system that estimates solar global spectral irradiance in the range of 350–900 nm. In the reported model, the spectrum distribution is first determined using a cluster selection procedure, then a normalisation factor is calculated and applied; RF was used for both clustering and normalisation. This model requires meteorological and atmospheric input and can be deployed under various conditions. However, the model accuracy for air masses greater than 2.1 is not validated. Additionally, exploratory data analysis on input features is lacking; therefore, the requirement of multiple different features is not justifiable.

A more recent study from Chen et al. [8] demonstrates a machine learning approach to decompose broadband solar irradiance into visible (VIS) and infrared (NIR) components. Their study shows that XGBoost is the most accurate and reliable model for VIS and NIR decomposition. Easily accessible meteorological data have been used as predictors for the solar spectral components, which is a challenge to perform a wavelength-resolved spectral irradiance estimation study.

This work aims to provide further understanding on the interaction between meteorological data and spectral irradiance using a data-driven statistical method; additionally, determining which machine learning model out of the selected ones suits the spectral data best. In this study, three different ML models are investigated, XGBoost, LSTM, and RF. XGBoost algorithm was first developed by Friedman in 2001 [13]. With good regularisation in regression, it can show good tolerance against the variance and noise in long time series data. XGBoost also comes with good non-linear relationship handle and scalability for larger data sets. RF was first designed by Breiman [6]. RF is a set of decision trees that can be helpful in avoiding overfitting. It also shows good ability in handing missing data. LSTM was developed by Hochreiter and Schmidhuber [21] in 1977 and is a type of Recurrent Neural Network (RNN), famous for handling sequential time-series data. Published work has shown that LSTM behaves better in processing time series data [25,38,41].

Therefore, the aim of the present study is to investigate and understand the interaction between meteorological data and spectral irradiance through data-driven statistical methods, identifying which machine learning model from XGBoost, LSTM, or RF is the most effective for spectral data estimation. By developing an accurate approach for predicting spectral solar irradiance using readily available meteorological input, we address the limitations of SMARTS model and ML methods constraints.

2. Model tunings and comparisons

2.1. Data sources

The training and testing sets for the presented ML models are based on site-specific data obtained from the Solar Energy Research Institute of Singapore (SERIS), with a temporal resolution of one minute for Germany, Australia, Singapore and China. The data set includes spectral *GHI* in Wm^{-2} nm, broadband *GHI* in W m^{-2}, broadband *DNI* in W m^{-2}, global tilted broadband irradiance (*GTI*) in W m^{-2} and meteorological parameters such as relative humidity (*RH*) in percentage, ambient temperature (T_{air}) in degree Celsius, wind speed in ms^{-1}, wind direction in degrees, and rainfall depth in mm. Broadband *GHI* and *GTI* are measured using both pyranometers and reference cells, covering a wide range of spectrum. The coordinates of the measurement site, the spectral range of the spectroradiometer, the tilt angle of the *GTI* sensors are presented in Table 1.

For a fair comparison, it is crucial to use the accuracy of the SMARTS model as a reference. To generate spectra with the same minute-by-minute temporal resolution and maintain accuracy, we used *RH* and T_{air} data on site from SERIS data sets, and aerosol optical depth at 500 nm (AOD500) and ozone abundance (AbO3) from Aerosol Robotic Network (AERONET) [22] as input variables. AERONET data are Level 2.0, validated with manual quality assurance, post-field calibration, and are recommended for publications by Holben et al. [22]. Additional fixed inputs for the site are listed in Table 2.

2.2. Selections of features

The data from SERIS contain a series of data measured simultaneously with the spectral irradiance data, including global horizontal irradiance measured by pyranometer (*GHI*), reference solar cells (GHI_{Si})

Table 1
Data from SERIS

Location	Germany	Australia	Singapore	China
Coordinate (°)	51.77, 11.77	−23.76, 133.88	1.28, 103.87	32.14, 114.03
Phase (Year)	2017–2023	2017–2023	2019–2023	2020–2023
Bandwidth (nm)	350.4–1052.4	350.4–1055.6	349.0–1053.0	363.7–1138.6
GTI angle (°)	35	24	10	32

GTI = global tilted broadband irradiance; SERIS = Solar Energy Research Institute of Singapore.

Table 2
Fixed inputs for SMARTS

Parameter	Value
Bandwidth (nm)	300.0–1200.0
Altitude (km)	0.08
Height (km)	0.002
Solar constant (W m^{-2})	1366.1 [2]
Aerosol model	S&F_RURAL [29]
Ground material	Soil

and UV sensor (GHI_{UV}), global tilted horizontal irradiance measured by pyranometer (*GTI*) and reference solar cell (GTI_{Si}), direct normal irradiance measured by pyrheliometer (*DNI*), the temperature of the *GHI* reference cell (T_{GHI}) and *GTI* reference cell (T_{GTI}), rain depth, wind speed, wind direction and ambient humidity. Exploratory data analysis was performed for statistical measure of the different parameters to better use computational resources, reduce complexity, and gain a better understanding of the feature-to-target relations.

Initially, linear regression was performed for all possible combinations of features with the 0.8 training ratio to obtain the baseline mean squared error (MSE). Then $MSE_{\%}$ relative to the target variance was calculated, given by Eq. (1a) below:

$$MSE_{\%} = \left(\frac{MSE}{\sigma_y^2}\right) \times 100\%, \tag{1a}$$

where σ_y^2 is given by Eq. (1b):

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2, \tag{1b}$$

where n is the number of timestamps in the dataset, y_i is the average of the irradiance value. Based on $MSE_{\%}$, the occurrence of each feature in the top 100 combinations with the lowest MSE percentage was calculated, presented in Fig. 1.

It is evident that *GHI* has a significant relationship with spectral irradiance. The UV irradiance, *DNI*, and temperature of the reference cells demonstrate a high correlation. However, these were not selected as training features due to limited data availability. As a primary aim of this study, we seek to minimise the number of meteorological inputs required for the determination of spectral irradiance. For this purpose, *GHI* data from reference cells and pyranometers are selected as key training features in the models because they are widely available across observation sites. Due to the sequential nature of the dataset, it is essential to include cyclic temporal features. However, this cyclical nature is not well represented by raw timestamps; the numerical integer difference between 23:00 and 00:00 is large, but they are only one hour apart in reality. Using sine and cosine transformations allows smooth transitions between the end and the beginning of cycles. For instance, the transformation ensures that the transition from December (12th month) to January (1st month) is consistent. Eqs. (2a) and (2b) are used for the transitions of the time series index:

$$n_{sin} = \sin\left(\frac{2\pi n}{N}\right), \tag{2a}$$

$$n_{cos} = \cos\left(\frac{2\pi n}{N}\right), \tag{2b}$$

Eqs. (2a) and (2b) are applied to all cyclical variables, including month, date, day, hour and minute, where n is the cyclical variable and N is the upper limit of n.

2.3. Model tunings

To evaluate the performance of different models on spectral irradiance, 18 wavelengths between 350.4 and 1052.4 nm were selected. To find the optimal parameters for different models, the model-related parameters were hyperparameter tuned. The hyperparameters used and their corresponding ranges, if applicable, for each model are presented in Table 3, Table 4 and Table 5.

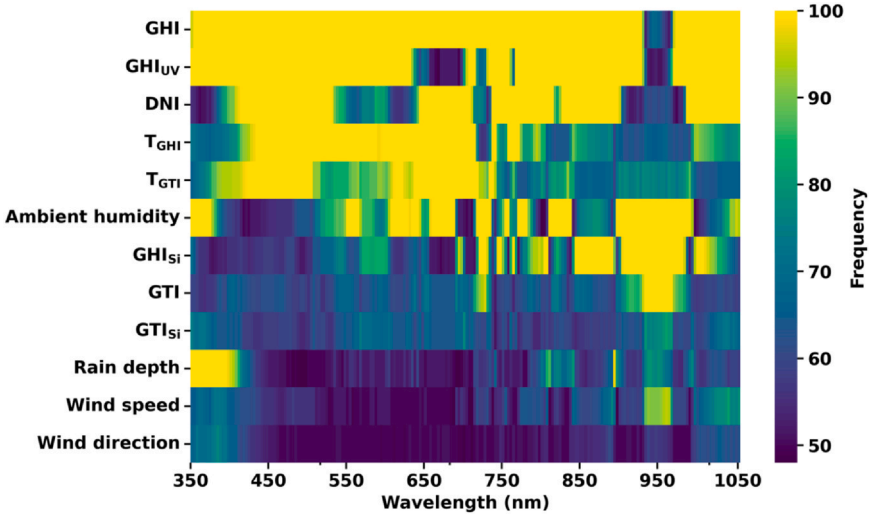

Fig. 1. Feature frequency as a function of wavelength.

Table 3
LSTM hyperparameters

Hyperparameter	Range
Hidden dimensions	50, 100, 150, 200, 250, 300
Learning rate	10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}
Batch size	32, 64
Number of layers	1, 2
Dropout rate	0.2

LSTM = long short-term memory.

Table 4
XGBoost hyperparameters

Hyperparameter	Range
Max depth	3, 6, 9
Learning rate	10^{-3}, 10^{-2}, 10^{-1}
Number of estimators	100, 200
Subsample	0.7, 0.9
Colsample per tree	0.7, 0.9

Table 5
RF hyperparameters

Hyperparameter	Range
Number of estimators	100, 700
Max depth	50, 300
Minimum sample split	2, 9
Number of layers	1, 9

RF = random forest.

3. Results and discussions

After performing hyperparameter tuning for LSTM, XGBoost and RF for each of the selected wavelength using the aforementioned hyperparameter parameter ranges, the model with the lowest root mean squared error (*RMSE*) is selected as the best performance model. To further increase the accuracy of the model, two stacking models, namely *XGB_LSTM* and *RF_LSTM*, were investigated, which are structured using the best performance model of its own class. XGB is used as an abbreviation of XGBoost here. The stacking was defined in Eqs. (3a) and (3b):

$$XGB\_LSTM = a \times XGB(X) + (1 - a) \times LSTM(X), \tag{3a}$$

$$RF\_LSTM = a \times RF(X) + (1 - a) \times LSTM(X), \tag{3b}$$

where X is the input feature, and a is a number between 0 and 1, which will be looped through to optimise the stacking model with the lowest *RMSE*.

After constructing models from different classes for each wavelength, metrics of each model class is represented using the average value of the metrics of the best performance model on each selected wavelength, using Eqs. (4a), (4b) and (4c):

$$MAE_{avg} = \frac{\sum_{i=1}^{n} MAE_{\lambda_i}}{n}, \tag{4a}$$

$$RMSE_{avg} = \frac{\sum_{i=1}^{n} RMSE_{\lambda_i}}{n}, \tag{4b}$$

$$R^2_{avg} = \frac{\sum_{i=1}^{n} R^2_{\lambda_i}}{n}, \tag{4c}$$

where λ_i represents selected wavelengths, n is the total number of selected wavelengths. The performance of each model class on the selected wavelengths is shown in Fig. 2. The 2018 data are divided into 80% training and 20% testing/validation, and all data from the 2019 data were used for the unseen test.

As shown in Fig. 2, the performance of all models in 2018 and 2019 shows similar trends, helping to prove that the model is not overfitting. For nonstacking models, the LSTM model performs the worst for all three metrics. RF performs noticeably better than XGBoost, with 0.3% higher R^2, 4.8% lower *mean absolute error* (*MAE)* and 4.9% lower *RMSE*. However, considering the significantly higher computational resources and the longer training time, XGBoost is in general a better alternative. Considering the case of stacking models, no significant variations were observed. When stacking RF and LSTM, the stacked model shows 0.05% higher R^2, 0.7% lower *MAE* and 1. 1% lower *RMSE* compared to the RF model. When stacking XGBoost and LSTM, the stacked model shows 0.05% higher R^2, 0.3% lower *MAE* and 1% lower *RMSE*. XGBoost is chosen for its relatively good accuracy and shorter training time.

To better evaluate the performance of different models, the data sets were classified into clear, overcast, and intermediate using the clearness index (*Kt*), where *Kt* below 0.3 is classified as overcast, *Kt* between 0.3 and 0.78 is classified as intermediate, *Kt* above 0.78 is considered clear by Mercier et al. [28]. *Kt* is calculated using Eq. (5a):

$$K_t = \frac{GHI}{G_0}, \tag{5a}$$

where G_0 is the extraterrestrial solar irradiance given by Eq. (5b):

$$G_0 = I_0 \cos(\theta_z)\left(1 + 0.034 \cos\left(\frac{360n}{365}\right)\right), \tag{5b}$$

where I_0 is the solar constant (1366.1 W m^{-2}), and θ_z is the solar zenith angle calculated using Eq. (5c):

$$\cos(\theta_z) = \sin(\phi)\sin(\delta) + \cos(\phi)\cos(\delta)\cos(H), \tag{5c}$$

where θ and H are given by Eqs. (5d) and (5e):

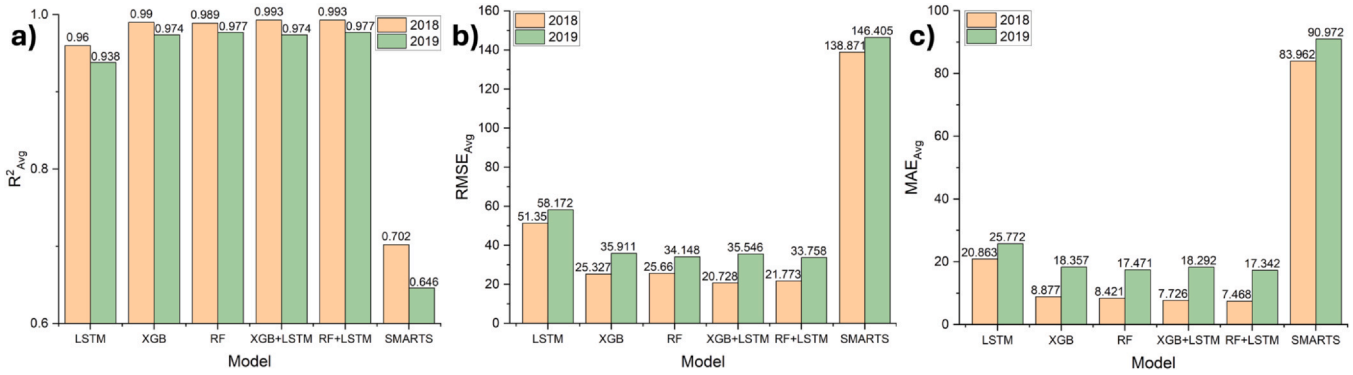

Fig. 2. Model comparisons using different evaluation metrics: (a) R^2, (b) *RMSE*, (c) *MAE*. *R^2 = R-squared; MAE = mean absolute error*; RMSE = root mean squared error.

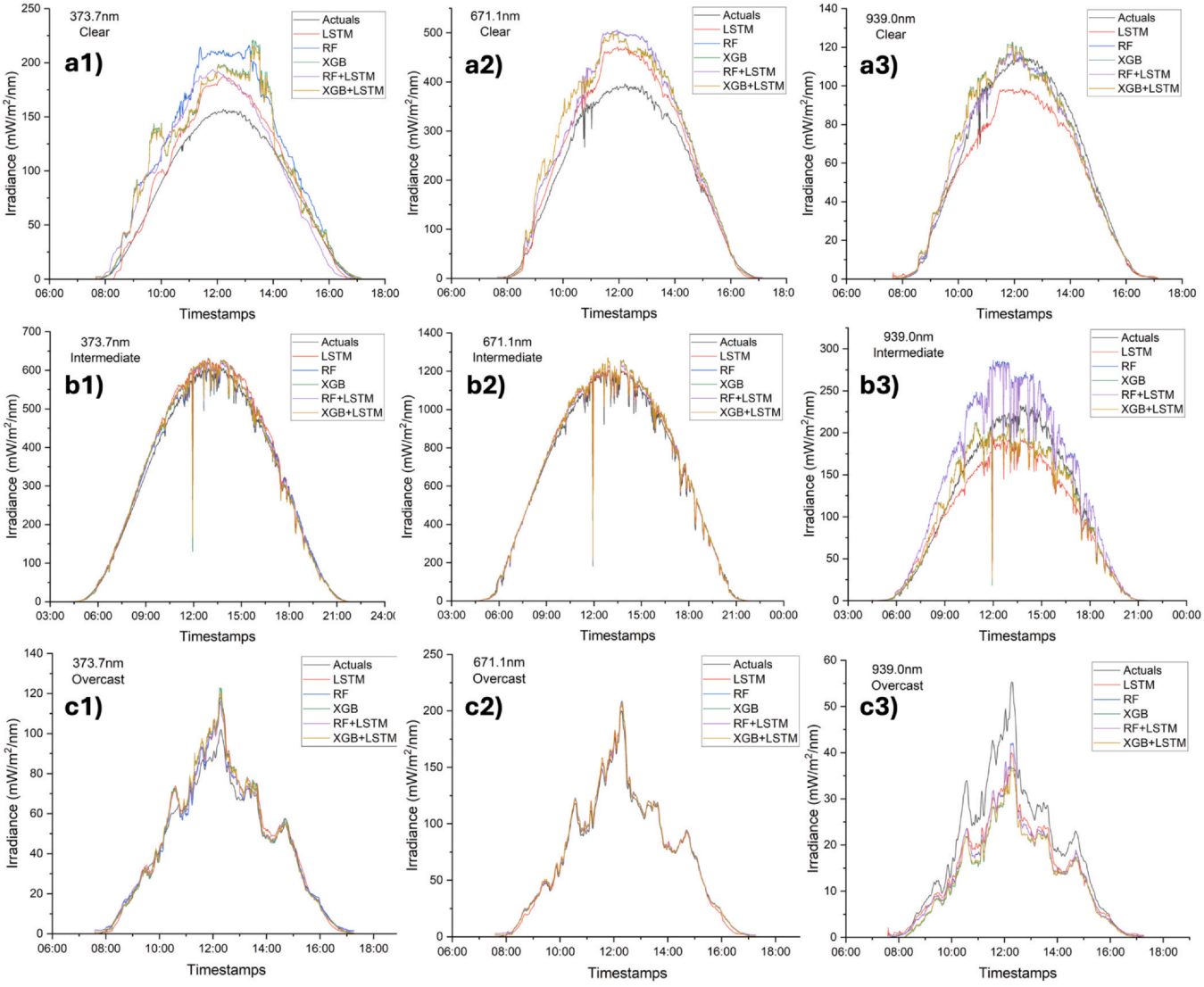

Fig. 3. Model predictions for wavelengths 373.2 nm, 671.1 nm, and 939.0 nm: (a1)–(a3) Clear day predictions, (b1)–(b3) Intermediate day predictions, (c1)–(c3) Overcast day predictions.

$$\theta = 23.45^{\circ}\sin\left(\frac{360}{365}(DOY - 81)\right), \tag{5d}$$

$$H = 15^{\circ}(Time_{solar} - 12), \tag{5e}$$

where *DOY* refers to day of the year.

To show the variations between each model, finer hourly resolution was investigated, and three wavelengths are picked, including 373.7, 671.1, and 939 nm. The predictions against timestamps for each wavelength under clear, intermediate, and overcast conditions are plotted in Fig. 3.

From Figs. 3(a1) − (a3), it can be concluded that all models effectively capture the trend of underlying changes in irradiance over time. The difference between a standalone model and a stacking model is minor on a minute-by-minute scale, which aligns with observations with overall model performance. Categorised by sky type, it is noticeable that ML models work particularly well under overcast conditions and intermediate conditions, under which RTMs cannot accurately estimate or require additional meteorological data to process. It can also be seen that the ML models capture the trend worse at a longer wavelength from Figs. 3(a3), (b3) and (c3), which resulted from the lack of highly dependent humidity data at longer wavelength.

Stacking models make minor differences compared to standalone models; scatter plots of the XGBoost model at 3 wavelengths are shown in Fig. 4, where solar zenith angles that are over 85 degrees are trimmed as they involve artificial errors when calculating extra-terrestrial solar irradiance.

The subgroups of overcast data are well clustered along the trend line, showing the least number of outliers. Data under intermediate conditions are more scattered, followed by the most discrete data under clear conditions. Similarly to the findings from the timestamp plots, model predictions globally also deviate further from the actual values at longer wavelengths, as irradiance outside the visible spectrum experiences more absorption.

Given that XGBoost provides relatively good accuracy and computational efficiency, we performed a full-spectrum exploratory data analysis to demonstrate the impact of different input features on each wavelength. As shown in Fig. 5, we plotted the *R-squared* (R^2) at each wavelength for the highlighted input features identified in the preliminary linear regressions in Fig. 1. The models here were hyperparameter tuned using the same hyperparameter range as in Table 3. From the plot, it is evident that using *GHI* alone results in poorer performance across the spectrum. By adding *GHI* measured by a reference cell, the model performance improves significantly, especially at longer wavelengths in the NIR zone. Incorporating *GHI* measured by UV sensors leads to a significant improvement at shorter wavelengths. The improvements from introducing *DNI* and humidity are minor. In particular, the addition of temperature results in

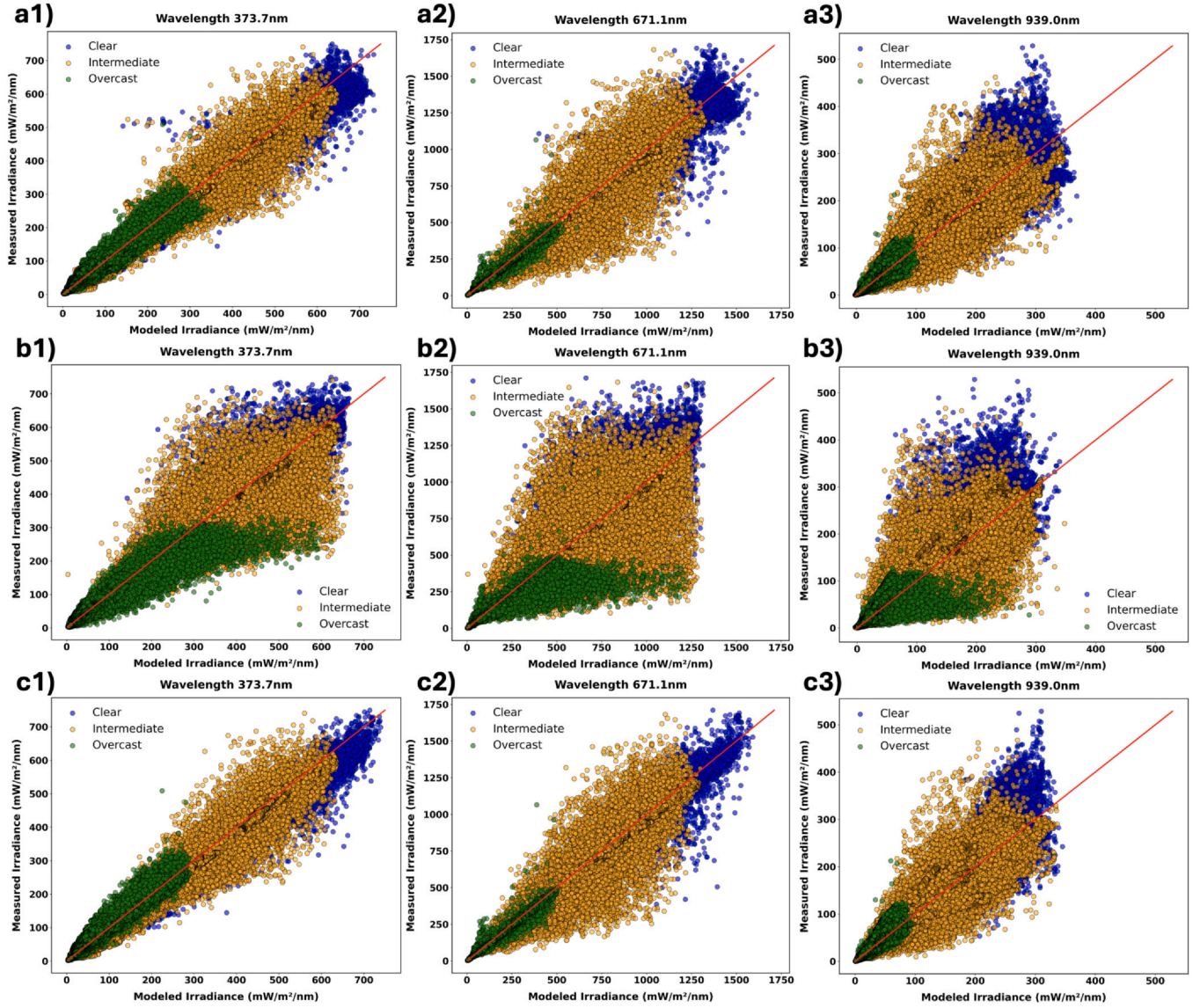

Fig. 4. Scatter plots for three wavelengths 373.2 nm, 671.1 nm, and 939.0 nm using different models: (a1)–(a3) XGBoost Model, (b1)–(b3) LSTM Model, (c1)–(c3) RF Model. LSTM = long short-term memory; RF = random forest; XGBoost = extreme gradient boost.

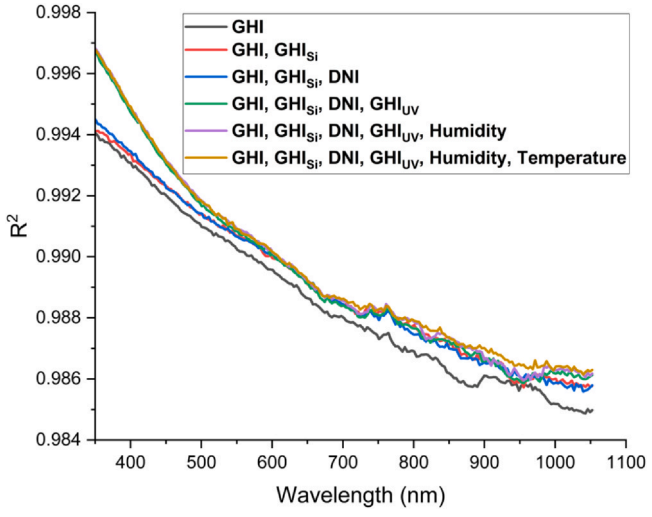

Fig. 5. Exploratory data analysis of input features for XGBoost. XGBoost = extreme gradient boost.

slight improvements in the NIR zone. Apart from that, it is noticeable that the accuracy of the model gradually drops towards NIR. At longer wavelengths, the sensor sensitivity typically declines, atmospheric absorption intensifies, and the signal-to-noise ratio worsens. Additionally, calibration drift or limited data coverage in this spectral region can hamper learning. Combined with potential complexities in infrared interactions, these conditions can reduce the accuracy of the model compared to lower wavelengths. To gain deeper insights into the role of each predictor, SHapley Additive exPlanations (*SHAP*) analysis was conducted on 10 evenly spaced wavelengths within the measurement spectrum range. The Shapley value, originally formulated in cooperative game theory as a unique solution for fairly distributing payoff among coalition members [35], has become a cornerstone for explaining complex machine-learning (ML) models [14]. By enumerating all possible feature coalitions, it quantifies the marginal contribution of each input while preserving axioms of efficiency, symmetry and additivity. Lundberg and Lee [27] introduced SHAP, an efficient framework that embeds Shapley values in additive feature-attribution models, enabling fast, model-agnostic explanations even for deep networks and ensemble learners. In photovoltaic (PV) research, SHAP analyses are increasingly used to uncover the physical drivers behind predictions of solar irradiance, module temperature and power output.

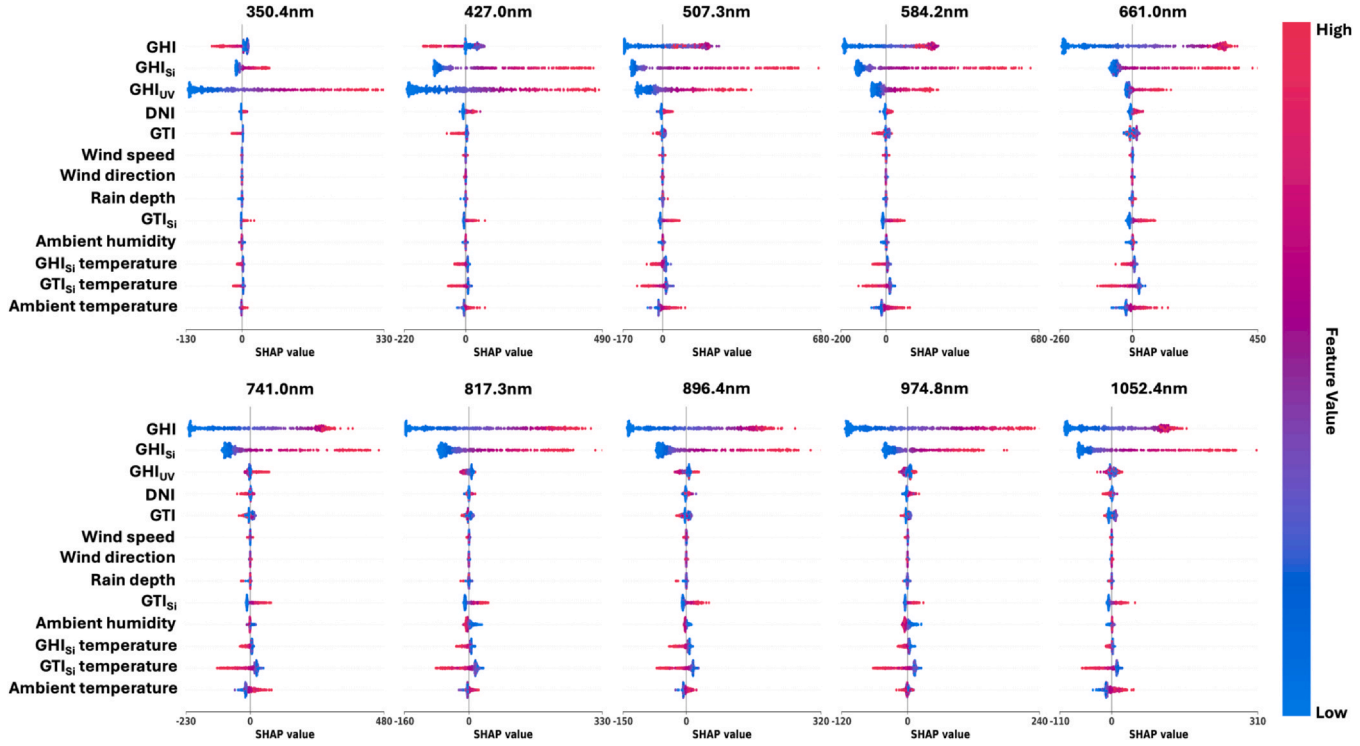

Fig. 6. Feature importance for XGBoost at 10 evenly spaced wavelengths. XGBoost = extreme gradient boost.

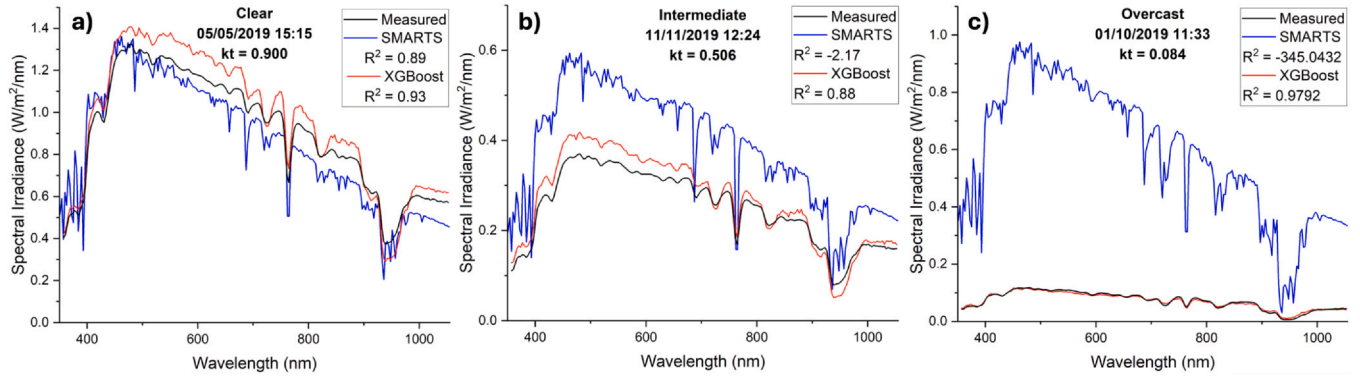

Fig. 7. Comparison between SMARTS and XGBoost models under different sky conditions: (a) Clear, (b) Intermediate, (c) Overcast. XGBoost = extreme gradient boost.

Tree-based ensembles coupled with SHAP have clarified how aerosols, cloud fraction and solar geometry influence global radiation forecasts [37]. Chen et al. [8] used SHAP to explain the contributions from meteorological data on reconstructing the narrow-band solar spectra using XGBoost. In our SHAP analysis, XGBoost models for each wavelength have been hyperparameter tuned based on Table 3 with a 0.8 training ratio. As shown in Fig. 6, we observe that *GHI* and GHI_{Si} have the highest impact across the spectrum. In the UV zone, a higher value of GHI corresponds to an increased prediction, while this turns to the opposite in the VIS and NIR zones. It is noticeable that the impact of the lower *GHI* increases from UV to VIS zone. Across the spectrum, the lower GHI_{Si} value has less influence on the models' output compared to the higher GHI_{Si} value. Another noticeable pattern is how GHI_{UV} affects the model's output in UV and near VIS, which can be understood from both a data-driven and a physical perspective. The UV sensor is specifically designed to respond to ultraviolet radiation, so its prominence in UV predictions aligns well with its intended sensing function. This suggests that the model not only captures statistical correlations, but also utilises physically meaningful features. The SHAP analysis supports the selection of *GHI* and GHI_{Si} as important features, highlighting their strong contribution to the predictions of the models.

As shown in Fig. 7, comparisons of spectral irradiance between SMARTS and proposed XGBoost models have been made under 3 timestamps, representing clear sky conditions, intermediate conditions, and overcast conditions. Under clear sky conditions, both the SMARTS and XGBoost models perform well, showing a R^2 value of 0.89 and 0.93 respectively. However, as clarity increases, the performance of the SMARTS model drops dramatically with R^2 values of -2.17 and -345.04 for intermediate conditions and overcast conditions, reflecting its limited application scenarios under clear conditions only. In contrast, the XGBoost model demonstrates precise estimation of the spectrum under 3 conditions, especially for overcast conditions, which agrees with the conclusion drawn from the scatter plots.

The model derived from Germany data is evaluated at other sites, including Australia in 2020, China in 2021 and Singapore in 2020. As shown in Fig. 8, the performance of the models is significantly worse at some absorption bands, indicating that the models are site specific.

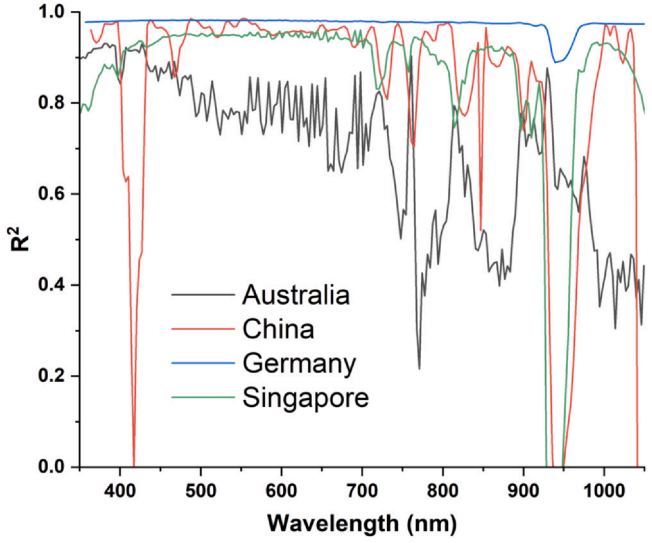

Fig. 8. Versatility of the XGBoost model across different locations: Australia, China, Germany, and Singapore. XGBoost = extreme gradient boost.

4. Summary and conclusions

This study has demonstrated that machine learning (ML) models can effectively estimate solar spectra based on *GHI* measured only by a pyranometer and a solar reference cell. The models employed in this research, including LSTM, RF and XGBoost, have been shown to be more precise than the SMARTS model with 45.2%, 51.2%, 50.8% lower R^2 respectively.

Although model stacking provided minor performance improvements compared to standalone models, it was observed that LSTM models performed significantly worse than RF and XGBoost models with 4% and 3.6% lower R^2 respectively. Among the models, RF slightly outperformed XGBoost but at the cost of significantly higher computational resources, making XGBoost an efficient alternative.

To evaluate the performance of the model under varying sky conditions, the data were classified into clear, intermediate, and overcast skies. All ML models demonstrated robust adaptability throughout the day in different sky conditions and wavelengths. In particular, LSTM models exhibited fewer abnormal peaks, likely due to their tolerance for overfitting, whereas RF and XGBoost models provided better fits over extended time-spans.

Compared to the SMARTS model, the machine learning (ML) models generally exhibited better fitting accuracy under overcast conditions compared to intermediate and clear conditions. Specifically, LSTM models tended to underestimate spectral irradiance under overcast skies and overestimate it under clear skies. This accounts for their relatively poorer performance compared to the RF and XGBoost models. The tendency for overestimation under clear conditions, particularly at longer wavelengths, may be due to the lack of additional input, such as humidity. From the exploratory data analysis conducted for XGBoost, it can be concluded that *GHI* measured by UV sensors can significantly improve model performance in the UV zone. *GHI* measured by a reference cell improves performance throughout the spectrum. Including temperature can slightly improve model performance in the NIR zones.

This research underscores the viability of ML techniques for solar spectra estimation, revealing significant improvements over the widely used radiative transfer model, SMARTS. However, ML methods appear to be site-specific. Future works will focus on improving the general applicability of ML models.

Declaration of Competing Interest

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

Acknowledgements

T. Rahman acknowledges support from EPSRC grant EP/X033333/1.

This work was supported by the Solar Energy Research Institute of Singapore (SERIS) at the National University of Singapore (NUS). SERIS is supported by NUS, the National Research Foundation Singapore (NRF), the Energy Market Authority of Singapore (EMA), and the Singapore Economic Development Board (EDB).

References

[1] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez-de Pison, F. Antonanzas-Torres, Review of photovoltaic power forecasting, Sol. Energy 136 (2016) 78–111, https://doi.org/10.1016/j.solener.2016.06.069.
[2] ASTM, 2006. Solar constant and zero air mass solar spectral irradiance. Annual Book of ASTM Standards 15.03, pp. 1-16.
[3] ASTM, 2013. Standard tables for reference solar spectral irradiances: direct normal and hemispherical on 37° tilted surface. ASTM 03.
[4] F. Besharat, A.A. Dehghan, A.R. Faghih, Empirical models for estimating global solar radiation: a review and case study, Renew. Sustain. Energy Rev. 21 (2013) 798–821, https://doi.org/10.1016/j.rser.2012.12.043.
[5] R.E. Bird, C. Riordan, Simple solar spectral model for direct and diffuse irradiance on horizontal and tilted planes at the earth's surface for cloudless atmospheres, J. Clim. Appl. Meteorol. (1986) 25, https://doi.org/10.1175/1520-0450(1986)025<0087:SSSMFD>2.0.CO;2.
[6] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[7] J. del Campo-Ávila, M. Piliougine, R. Morales-Bueno, L. Mora-López, A data mining system for predicting solar global spectral irradiance. performance assessment in the spectral response ranges of thin-film photovoltaic modules, Renew. Energy 133 (2019) 828–839, https://doi.org/10.1016/j.renene.2018.10.083.
[8] C. Chen, Q. Duan, Y. Feng, J. Wang, N. Ghaeili, N. Wang, M. Hosseini, C. Shen, Reconstruction of narrowband solar radiation for enhanced spectral selectivity in building-integrated solar energy simulations, Renew. Energy 219 (2023) 119554, https://doi.org/10.1016/j.renene.2023.119554.
[9] J. DelRocco, P.D. Bourke, C.B. Patterson, J.T. Kider, Real-time spectral radiance estimation of hemispherical clear skies with machine learned regression models, Sol. Energy 204 (2020) 48–63, https://doi.org/10.1016/j.solener.2020.04.006.
[10] 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 (2015) 431–442, https://doi.org/10.1016/j.solmat.2014.09.034.
[11] R. Eke, T.R. Betts, R. Gottschalg, Spectral irradiance effects on the outdoor performance of photovoltaic modules, Renew. Sustain. Energy Rev. 69 (2017) 429–434, https://doi.org/10.1016/j.rser.2016.10.062.
[12] C. Emde, R. Buras-Schnell, A. Kylling, B. Mayer, J. Gasteiger, U. Hamann, J. Kylling, B. Richter, C. Pause, T. Dowling, L. Bugliaro, The libradtran software package for radiative transfer calculations (version 2.0.1), Geosci. Model Dev. 9 (2016) 1647–1672, https://doi.org/10.5194/gmd-9-1647-2016.
[13] J. Friedman, R. Tibshirani, T. Hastie, Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), Ann. Stat. 28 (2000) 337–407, https://doi.org/10.1214/aos/1016120463.
[14] A. Ghorbani, J. Zou, Data Shapley: Equitable Valuation of Data for Machine Learning, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, pp. 2242–2251.
[15] C.A. Gueymard. 1995. Smarts2, a simple models of the atmospheric radiative transfer of sunshine. Florida Solar Energy Center/University of Central Florida, p. 84.
[16] C.A. Gueymard, Prediction and validation of cloudless shortwave solar spectra incident on horizontal, tilted, or tracking surfaces, Sol. Energy 82 (2008) 260–271, https://doi.org/10.1016/j.solener.2007.04.007.
[17] R. Gupta, A.K. Yadav, S. Jha, P.K. Pathak, Long term estimation of global horizontal irradiance using machine learning algorithms, Optik 283 (2023) 170873, https://doi.org/10.1016/j.ijleo.2023.170873.
[18] R. Gupta, A.K. Yadav, S. Jha, P.K. Pathak, Composition of feature selection techniques for improving the global horizontal irradiance estimation via machine learning models, Therm. Sci. Eng. Prog. 48 (2024) 102394, https://doi.org/10.1016/j.tsep.2024.102394.
[19] R. Gupta, A.K. Yadav, S. Jha, P.K. Pathak, Predicting global horizontal irradiance of north central region of india via machine learning regressor algorithms, Eng. Appl. Artif. Intell. 133 (2024) 108426, https://doi.org/10.1016/j.engappai.2024.108426.
[20] R. Gupta, A.K. Yadav, S. Jha, P.K. Pathak, A robust regressor model for estimating solar radiation using an ensemble stacking approach based on machine learning, Int. J. Green. Energy 21 (2024) 1853–1873, https://doi.org/10.1080/15435075.2023.2276152.
[21] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[22] B.N. Holben, T.F. Eck, I. Slutsker, D. Tanré, J.P. Buis, A. Setzer, E. Vermote, J.A. Reagan, Y.J. Kaufman, T. Nakajima, F. Lavenu, I. Jankowiak, A. Smirnov, Aeronet—a federated instrument network and data archive for aerosol characterization, Remote Sens. Environ. 66 (1998) 1–16, https://doi.org/10.1016/S0034-4257(98)00031-5.

[23] M.T. Hörantner, H.J. Snaith, Predicting and optimising the energy yield of perovskite-on-silicon tandem solar cells under real world conditions, Energy Environ. Sci. 10 (2017) 1983–1993, https://doi.org/10.1039/c7ee01232b.
[24] F.X. Kneizys, Laboratory, U.S.A.F.G. 1988. Users Guide to Lowtran 7. Air Force Geophysics Laboratory.
[25] P. Kumari, D. Toshniwal, Deep learning models for solar irradiance forecasting: A comprehensive review, J. Clean. Prod. (2021) 318, https://doi.org/10.1016/j.jclepro.2021.128566.
[26] T. Le, C. Liu, B. Yao, V. Natraj, Y.L. Yung, Application of machine learning to hyperspectral radiative transfer simulations, J. Quant. Spectrosc. Radiat. Transf. (2020) 246, https://doi.org/10.1016/j.jqsrt.2020.106928.
[27] S.M. Lundberg, S. Lee, 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30, Curran Associates, Inc. pp. 4765-4774.
[28] T.M. Mercier, A. Sabet, T. Rahman, Vision transformer models to measure solar irradiance using sky images in temperate climates, Appl. Energy 362 (2024) 122967, https://doi.org/10.1016/j.apenergy.2024.122967.
[29] D.R. Myers, C.A. Gueymard, Description and availability of the smarts spectral model for photovoltaic applications, Org. Photovolt. V. 5520 (2004) 56, https://doi.org/10.1117/12.555943.
[30] S. Nann, K. Emery, Spectral effects on pv-device rating, Sol. Energy Mater. Sol. Cells 27 (1992) 189–216, https://doi.org/10.1016/0927-0248(92)90083-2.
[31] R.A. Ramadhan, Y.R. Heatubun, S.F. Tan, H.J. Lee, Comparison of physical and machine learning models for estimating solar irradiance and photovoltaic power, Renew. Energy 178 (2021) 1006–1019, https://doi.org/10.1016/j.renene.2021.06.079.
[32] P. Ricchiazzi, S. Yang, C. Gautier, D. Sowle, Sbdart: a research and teaching software tool for plane-parallel radiative transfer in the earth's atmosphere, Bull. Am. Meteorol. Soc. (1998) 79, https://doi.org/10.1175/1520-0477(1998)079<2101:SARATS>2.0.CO;2.
[33] L.S. Rothman, R.R. Gamache, R.H. Tipping, C.P. Rinsland, M.A.H. Smith, D.C. Benner, V.M. Devi, J.M. Flaud, C. Camy-Peyret, A. Perrin, A. Goldman, S.T. Massie, L.R. Brown, R.A. Toth, The hitran molecular database: Editions of 1991 and 1992, J. Quant. Spectrosc. Radiat. Transf. 48 (1992) 469–507, https://doi.org/10.1016/0022-4073(92)90115-K.
[34] J. Sarkar, S. Bhattacharyya, Application of graphene and graphene-based materials in clean energy-related devices minghui, Arch. Thermodyn. 33 (2012) 23–40, https://doi.org/10.1002/er.
[35] L.S. Shapley, A Value for n -Person Games, in: H.W. Kuhn (Ed.), Contributions to the Theory of Games II. Princeton University Press, Princeton, 1953, pp. 307–317.
[36] H. Smith, D. Dube, M. Gardner, S. Clough, F. Kneizys, 1978. Fascode - fast atmospheric signature code (spectral transmittance and radiance), p. 143.
[37] Z. Song, S. Cao, H. Yang, An interpretable framework for modeling global solar radiation using tree-based ensemble machine learning and shapley additive explanations methods, Appl. Energy 364 (2024) 123238, https://doi.org/10.1016/j.apenergy.2024.123238.
[38] S. Srivastava, S. Lessmann, A comparative study of lSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data, Sol. Energy 162 (2018) 232–247, https://doi.org/10.1016/j.solener.2018.01.005.
[39] J. Troscianko, Osprad: an open-source, low-cost, high-sensitivity spectroradiometer, J. Exp. Biol. (2023) 226, https://doi.org/10.1242/jeb.245416.
[40] D. Yang, Solar radiation on inclined surfaces: corrections and benchmarks, Sol. Energy 136 (2016) 288–302, https://doi.org/10.1016/j.solener.2016.06.062.
[41] Y. Yu, J. Cao, J. Zhu, An lSTM short-term solar irradiance forecasting under complicated weather conditions, IEEE Access 7 (2019) 145651–145666, https://doi.org/10.1109/ACCESS.2019.2946057.
[42] Y. Zhang, L.O. Wijeratne, S. Talebi, D.J. Lary, Machine learning for light sensor calibration, Sensors (2021) 21, https://doi.org/10.3390/s21186259.