Introducing the spectral characteristics index: A novel method for clustering solar radiation fluctuations from a plant-ecophysiological perspective

Siriwardana, Amila Nuwan
Faculty of Agriculture, Kyushu University

Kume, Atsushi
Faculty of Agriculture, Kyushu University

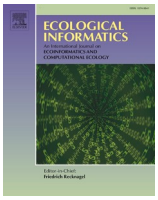
出版情報：Ecological Informatics. 85, pp.102940-, 2025-03. Elsevier
バージョン：

Ecological Informatics 85 (2025) 102940

Contents lists available at ScienceDirect

Ecological Informatics

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



Introducing the spectral characteristics index: A novel method for clustering solar radiation fluctuations from a plant-ecophysiological perspective

Amila Nuwan Siriwardana, Atsushi Kume *

Faculty of Agriculture, Kyushu University, 744 mottooka, Nishi-ku, Fukuoka 819-0395, Japan

ARTICLE INFO

Keywords:
Solar radiation clustering
Diffuse radiation
Machine learning
Spectral distribution
Spectral characteristics index

ABSTRACT

Solar radiation (SR) is a critical environmental factor influencing plant ecophysiology and ecosystem dynamics, not merely as an energy source but through its spectral characteristics, including critical wavelength ratios (CWRs) that trigger photomorphogenic responses in plants, the diffuse fraction (DF), that affect light distribution within canopies, and the variability of SR. This study presents the Spectral Characteristics Index (SCI), a novel method that integrates spectral quality and energy flux to classify daily SR conditions.

Data were collected using a rotating shadow-band spectroradiometer. The study applied agglomerative hierarchical clustering (AHC) based on cumulative Euclidean distance matrices and identified five SR clusters ranging from clear (SCI-01) to overcast (SCI-05) conditions, with spectral shifts from red to blue. Significant differences in DF, global solar irradiance (GSI), and CWRs were observed across clusters ($p < 0.0001$, $F > 27$).

Given the challenges in obtaining comprehensive spectral data in certain regions, machine learning models replicated SCI clustering using easily accessible environmental variables (DF, GSI, variability, airmass, and vapor pressure). The support vector machine (SVM) model achieved 88.03 % validation accuracy and 94.29 % test accuracy, providing a practical alternative where spectral measurements are not available. While long-term data collection across various climatic zones could improve the validity and adaptability of SCIs to different geographical locations, the current model demonstrates high accuracy and efficiency. This innovative approach enhances the understanding of SR dynamics and advances ecological research on plant responses and ecosystem functions.

1. Introduction

Solar radiation (SR) dynamics across days and seasons (Becker and Boyd, 1957) provide the energy essential for photosynthesis and various other metabolic processes (Durand et al., 2021; Xiao et al., 2020). This variability considerably affects plant ecophysiology, ecosystems, and agricultural productivity (Gürel et al., 2023), thus necessitating a comprehensive understanding of SR dynamics. Depending on light signals, photomorphogenesis responses may produce physiological and morphological changes in plants. For instance, in sweet basil, the β-Carotene concentration based on the leaf dry mass and leaf area at the fourth node increased with decreasing red (R) (λ 620–680 nm)/ blue (B) (λ 450–495 nm) ratio, regardless of the photosynthetically active radiation (PAR) (λ 400–700 nm) photon flux density (Hikosaka et al., 2021). Moreover, the R/far-red (FR) (λ 700–750 nm) ratio influences several processes including shade avoidance responses such as increased stem length, internode length, petiole length, and leaf length under low R/FR ratios (Demotes-Mainard et al., 2016).

The prediction of the SR characteristics in a region is of considerable importance in many fields and industries such as climate research and solar industry, respectively (Nematchoua et al., 2022). Recently, with the evolution of technology, artificial intelligence has gained momentum in SR prediction and clustering (Ağbulut et al., 2021; Nematchoua et al., 2022). Identifying specific SR characteristics under various weather conditions and classifying days into different categories have enabled us to comprehend the quantity and quality of available solar energy. Furthermore, such clustering may improve the understanding of natural climate trends and global climate changes, enabling the more accurate prediction of SR. Several attempts have been made to cluster (Fortin et al., 2008), model, and predict SR under various weather conditions (Hartmann, 2020). Some studies have employed

* Corresponding author.
E-mail address: akume@agr.kyushu-u.ac.jp (A. Kume).

https://doi.org/10.1016/j.ecoinf.2024.102940
Received 18 June 2024; Received in revised form 2 December 2024; Accepted 2 December 2024
Available online 11 December 2024

Nomenclature

R^2	Coefficient of determination
r	Correlation coefficient
ANOVA	Analysis of variance
nRMSE	Normalized root mean square error
RMSE	root mean square error
μ	Mean
JST	Japan standard time
SR	Solar radiation
PV	Photovoltaic
SSI	Spectral solar irradiance
GSI	Global solar irradiance
SCI	spectral characteristics index
DF	Diffuse fraction
15-min MSG	15-min moving standard deviation of global solar radiation
DTW	Dynamic time wrapping
AHC	Agglomerative hierarchical clustering
ML	Machine learning
KNN	k-nearest neighbors
ANN	Artificial neural network
VP	Water vapor pressure
AM	Airmass
RSB	Rotating shadow band
UV	Ultraviolet
IR	Infrared
CWRs	Critical wavelength ratios
PAR	Photosynthetically active radiation
NEE	Net ecosystem exchange
SVM	Support vector machine

machine learning (ML) models such as artificial neural networks (ANNs) and support vector machines (SVMs) to cluster and predict SR. However, these studies have only focused on the energy flux of SR (Ağbulut et al., 2021). Furthermore, previous studies have not much focused on specific solar signals or the diffuse fraction (DF) of SR in clustering (Nematchoua et al., 2022) that considerably affects plant ecophysiology and ecology. Some clustering methods involve grouping days based on SR, possibly due to the natural and clear definition of a day from sunrise to sunset (Kang and Tam, 2013).

Despite the existence of numerous clustering methods that cluster the energy flux of SR, a comprehensive understanding of the effects on SR characteristics remains elusive (Hartmann, 2020), primarily owing to the absence of a clustering method that can effectively utilize and describe the qualitative and quantitative characteristics of SR. In natural settings, plants and ecosystems experience the comprehensive dynamics of SR, not merely the changing energy flux. There is a paucity of research that categorizes days based on SR in this context. Consequently, the comprehensive effects of SR on ecosystems and plant ecophysiology are yet to be fully elucidated. To address this gap, the present study aimed to introduce the spectral characteristics index (SCI), a novel approach for categorizing days based on the qualitative and quantitative aspects of SR.

The critical wavelength ratios (CWRs) perceivable by plants as light signals serve as essential external factors that mediate various physiological and developmental processes in plants, rendering these ratios crucial for ecological and plant physiological research (Xiao et al., 2021). Further, our findings indicate a considerable seasonal fluctuation in plant-perceivable light signals, which could impact plant growth and development as well as the overall ecosystem in the temperate monsoon region (Siriwardana and Kume, 2024). Moreover, recent years have witnessed a notable shift in scientific interest from the total solar energy flux to its spectral distribution (Kaskaoutis et al., 2008), highlighting the need for a more detailed characterization of the state of the sky to enhance our understanding of SR (Kang and Tam, 2013). Therefore, herein, we acknowledge the importance of plant-perceivable light signals, which rely on solar spectral irradiance (SSI), in clustering days based on SR. For instance, the R and B light regions are essential for photosynthesis and employed to optimize light conditions in precision agriculture and horticulture (Smith, 1982). Furthermore, infrared (IR) light contributes to heat flux and affects the microclimates and energy balance models used for predicting crop yields and forest dynamics (Monteith and Unsworth, 2013). Variations in ultraviolet (UV) radiation are important in studies related to skin cancer, atmospheric ozone depletion, and photochemical smog formation (McKenzie et al., 2007). Accordingly, the categorization of SR based on spectral characteristics provides a more comprehensive understanding regarding the interaction of the energy flux and spectral characteristics of SR and their contribution to various ecological processes. This study aimed to cluster daily SR characteristics into distinctive categories that can improve the precision of the models used for ecosystem assessment, climatological data analysis, and agricultural management.

Furthermore, if models fail to accurately cluster SR based on the energy accessible to plants, comparing plant responses and the spectral characteristics of SR would be meaningless (Xiao et al., 2020). The energy accessible to plants is considerably influenced by the path of light through the canopy, which is affected by the DF, a characteristic of light that is primarily dependent on sky cloudiness. On cloudy days, a higher DF results in a more homogeneous light profile in the canopy than on clear-sky days (Li and Yang, 2015), thereby increasing the energy accessible to deeper canopy layers and consequently enhancing net ecosystem productivity (Han et al., 2020). Despite the substantial importance of plant ecophysiology, its consideration during SR data clustering is often overlooked. In our approach for clustering days based on SR, we emphasize the importance of the DF. Partially cloudy and overcast weather conditions may cause considerable short-term fluctuations in SR (Lou et al., 2019). Empirical evidence indicates that these fluctuations could reduce biomass production compared with constant-light conditions (Chiang et al., 2020), thereby affecting the functions and morphology of the associated plants (Schumann et al., 2017). Studies have emphasized the importance of considering short-term fluctuations in SR and recommended their inclusion in future experiments aimed at understanding plant ecophysiology under natural conditions (Vialet-Chabrand et al., 2017). Consequently, this study acknowledges the importance of SR fluctuations in clustering days based on SR.

Unlike most studies that adopt either a qualitative or quantitative perspective (Moreno-Tejera et al., 2017), our approach incorporates both perspectives. The novel SCI category introduced by our study acknowledges the GSI, DF, and short-term fluctuation of SR as well as the SSI, thus providing a comprehensive understanding of SR in natural environments.

To address the uncertainties due to Citrus clouds in SR research, often associated with satellite data, herein, we used high-precision ground monitoring of SR (Kaskaoutis et al., 2008). Moreover, we used a shadow-band spectroradiometer with a rotating arm, which has the advantage of collecting data under all weather conditions. While diverse methods, including ML models, radiative transfer models, and remote sensing models, are available (Xiao et al., 2020), we employed ML for clustering owing to its capacity to effectively discern trends and patterns (Huang et al., 2021).

Our methodology commenced by eliminating uncertain data collected at 5-min intervals, as described in the methodology section.

Subsequently, we constructed variables such as the GSI, SSI, and the 15-min moving standard deviation of global SR (15-min MSG). We used Euclidean distance to determine similarity among non-time-series data, and dynamic time warping (DTW) for time-series data (Laperre et al., 2020). DTW facilitates temporal alignment and improves our understanding of specific features, such as trends, patterns, and anomalies, across multiple time series (Franses and Wiemann, 2020).

Furthermore, we applied agglomerative hierarchical clustering (AHC) to the cumulative sum of the normalized Euclidean distance matrices to cluster days according to the similarity in terms of SR. This approach facilitates the simultaneous consideration of the qualitative and quantitative characteristics of SR for clustering, unlike the approaches used in previous studies that solely processed the GSI using AHC for clustering SR data (Maldonado-Salguero et al., 2022). The iterative adjustment of variables guarantees optimal clustering, despite the complexity and time-consuming nature of the task. This iterative process is crucial because enhancing clustering for one criterion may potentially compromise another criterion owing to the simultaneous consideration. The optimal number of clusters was determined by employing the silhouette function that measures the extent of similarity of an object to its own cluster. We have denoted this clustering method as the SCI because it effectively describes daily spectral characteristics of SR comprehensively.

Our previous study has revealed that the DF, airmass (AM), and atmospheric water vapor pressure (VP) considerably affect plant-perceivable light signals (Siriwardana and Kume, 2024). Building upon this finding, the present study aimed to replicate these clustering outcomes using advanced ML models. We leveraged readily available SR data, comprising daily means of short-term fluctuation the GSI, GSI, and DF, supplemented with additional environmental variables such as daily mean VP and AM. Employing various sophisticated ML methods, including weighted K-nearest neighbor (KNN), fine KNN, support vector machine (SVM) (Ali Ou Salah et al., 2021), and ANN, we identified the best-performing ML model for reproducing the SCI from the given data. This approach enables the application of the SCI across diverse areas where comprehensive data collection may be challenging. Each SCI category will provide a comprehensive description of the average SR characteristics within that category. The obtained SCI will enhance our understanding of the natural dynamics of SR with respect to weather and seasons, which is crucial for developing innovative plant-ecophysiological and ecological response models. Furthermore, this approach could be useful in various other research and practical applications, such as solar Photovoltaic (PV) electricity production (Polo et al., 2015), human health, biological research, atmospheric and environmental sciences, and agriculture. (Miller et al., 2008).

2. Methods

2.1. Data collection and data cleaning

The data was collected at the Ito Campus of Kyushu University, located on the Itoshima Peninsula in the western part of Fukuoka Prefecture on Kyushu island, Japan (33.594 N, 130.214 E; 90 m asl.). The selected location is characterized by a warm temperate climate (Cfa) and is considerably influenced by monsoon weather patterns. For precise radiometric measurements, an EKO MS-711 spectroradiometer and EKO MB-22 rotating shadow band (RSB) were used (Fig. 3b). The spectroradiometer, equipped with a temperature-controlled detector core, yielded accurate irradiance data within the spectral range of 300–1100 nm with an optical resolution full width at half maximum of < 7 nm, a wavelength interval of 0.3–0.5 nm, a wavelength accuracy of ± 0.2 nm, and a field of view of 180°. Moreover, the RSB, a commercially available prototype since 2020, enabled the calculation of distinct irradiance components through its four operational positions. The RSB positions included the initial position (IRR1) beyond the field of view of the instrument; the second position (IRR2) at −5° from the solar disk; the third position (IRR3) covering the solar disk for diffuse horizontal irradiance measurement; and the fourth position (IRR4) at +5° from the solar disk, which facilitated the evaluation of the irradiance components (GSI = Direct solar irradiance (DIR) + diffuse solar irradiance (DFR)).

Our methodology involved an initial examination of spectral radiation data acquired from a spectroradiometer. To ensure the reliability and accuracy of the data, rigorous data cleaning was implemented, wherein any data with considerable discrepancies in IRR2 and IRR4 were discarded. Additionally, data points with a DF of > 1.01 were excluded from further analysis. For data points with DF in the range $1.0 < \mathrm{DF} \leq 1.01$, the value of DF was adjusted to 1.0. In this cleaning process, 3.01 % of the total data was eliminated. Simple calculations were used to calculate the DF, diffuse radiation, and direct radiation from the measurements (Eqs. 1, 2, and 3).

$$DF = \frac{2 \times IRR3}{IRR2 + IRR4} \tag{1}$$

$$Diffuse\ radiation = GSI \times DF \tag{2}$$

$$Direct\ radiation = GSI - Diffuse\ radiation \tag{3}$$

Observations were made at 5-min intervals from sunrise to sunset each day during the whole year of 2021. To ensure data consistency, we employed a second-class pyranometer FMP3 to collect the whole-day GSI. Regular maintenance of the radiometers was performed to minimize instrumental uncertainty. In addition to the spectral data, we obtained other environmental data such as daily average temperature (T), VP, and AM.

2.2. Euclidean distances

In this study, we utilized the Euclidean distance (ℓ_2) matrix to quantify the similarity of daily SR data, also known as the root sum of squared differences (Keogh and Pazzani, 2000). When computing the Euclidean distance, which is defined in Eq. 4, if **X** and **Y** are both K-dimensional signals, then matric prescribes $d_{mn}(\mathbf{X}, \mathbf{Y})$, the distance between the m^{th} sample of **X** and n^{th} sample of **Y** (Senin, 2008). We primarily calculated Euclidean distances in non-time-series data, such as the SSI, solar spectral DF, and ultraviolet (UV)/infrared (IR) ratio.

$$d_{mn}(\mathrm{X}, \mathrm{Y}) = \sqrt{\sum_{k=1}^{k} \left(\mathrm{x}_{k,m} - \mathrm{y}_{k,n}\right) * \left(\mathrm{x}_{k,m} - \mathrm{y}_{k,n}\right)} \tag{4}$$

2.3. DTW algorithm

We used the DTW algorithm to find the distance matrix of time-series data. The DTW algorithm has gained widespread recognition owing to its efficiency in measuring time-series similarity. It minimizes the effects of shifting and distortion of time series by enabling the "elastic" transformation of time series to detect similar shapes with different phases (Fig. 1a) (Senin, 2008).

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{K,1} & x_{K,2} & \cdots & x_{K,M} \end{bmatrix}$$

and

$$Y = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \cdots & Y_{1,M} \\ Y_{2,1} & Y_{2,2} & \ldots & Y_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{K,1} & Y_{K,2} & \cdots & Y_{K,M} \end{bmatrix}$$

have M and N samples, respectively. For given d_{mn} (X, Y), the

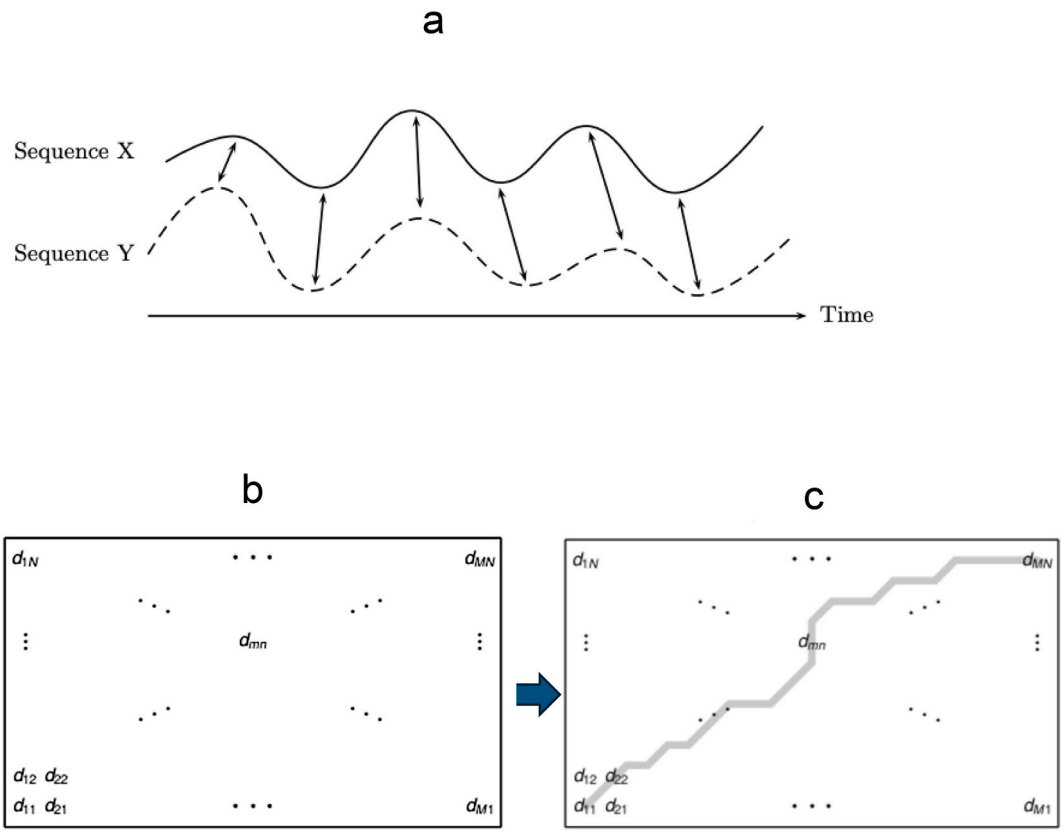


Fig. 1. **(a)** Illustration of DTW alignment. This figure provides a simplified visualization of the application of DTW in aligning two distinct signals, sequence X and Y. Unlike strict pair-wise correspondence, DTW optimizes the alignment between two sequences to minimize their overall distance, providing a more flexible and accurate sequence alignment. Furthermore, using DTW facilitates examining the temporal alignment and similarity across time series, enhancing understanding of data features across multiple time series (Franses and Wiemann, 2020). It is a well-established technique for finding optimal alignment between two given (time-dependent) sequences under certain restrictions (Tanaka and Takahashi, 2017). When considering a pair of signals with K-dimensions X and Y. **(b and c)** Example of DTW for optimal sequence alignment. This illustrates the DTW process in aligning two K-dimensional signals, X and Y, each with M and N samples, respectively. **(b)** The process begins by arranging all possible values of the distance between the m^{th} sample of X and n^{th} sample of Y, denoted as $d_{mn}(X, Y)$, into a single lattice. **(c)** DTW identifies two sequences of the same length, iX, and iY, which minimize the cumulative distance, d. The optimal path, which starts and ends diagonally without skipping samples or turning back, is illustrated.

distance between the m^{th} sample of X and n^{th} sample of Y, DTW arranged all possible values of d_{mn} (X, Y) into a single lattice (Fig. 1b) (Keogh and Pazzani, 2000). DTW was then parameterized using two sequences of the same length (iX & iY) that yield the minimum distance (d), and the path of d continued with certain conditions, such as the wrapping path moving diagonally from start to end (Eq. 5) without skipping samples and turning back (Fig. 1c) (Keogh and Pazzani, 2000). Therefore, in this study, we employed DTW as a methodological tool to compute the distance matrix for our time-series data that encompasses the daily fluctuation of variables such as the GSI, DF, and 15-min MSG.

$$d = \sum_{mn \in ix \in iy} d_{mn}(X, Y) \tag{5}$$

2.4. Agglomerative hierarchical clustering (AHC)

Furthermore, we employed AHC to process the cumulative sum of the Euclidean distance matrices, thereby clustering days based on SR. We tested the data and distance matrix with and without normalization to determine the most effective variable combination. Utilizing the cumulative sum of the Euclidean distance matrix allowed us to consider all variables simultaneously. Unlike previous research, which used AHC for SR data clustering based solely on the variation of GSI data (Maldonado-Salguero et al., 2022), this study incorporated multiple SR characters, including daily variation of the GSI, daily variation of the DF, fluctuations of the SSI, and daily mean UV/IR ratio.

Cluster analysis classifies objects into clusters based on their degree of similarity or dissimilarity. The fundamental principle of this method lies in constructing clusters wherein the objects assigned to a particular cluster exhibit shared characteristics. Clustering is often more suitable than supervised learning approaches, such as classification and regression, under unavailable extensive knowledge or expertise similar to our study (Bouguettaya et al., 2015). Among various clustering methods for clustering SR data, AHC is widely adopted owing to its broad applicability across diverse data types, including SR data (Bouguettaya et al., 2015). Although AHC is computationally complex, it holds an advantage over k-means clustering because it eliminates the need for predetermined parameters. This feature is particularly advantageous when dealing with natural environmental data, where identifying appropriate computational parameters can be challenging (Bouguettaya et al., 2015). AHC analysis follows a bottom-up approach: clustering begins with the number of observations/days (N) equal to the number of clusters (k). The process involves merging two clusters to obtain $N - 1$ ($k = N - 1$) clusters, and this merging is continued until all observations form a single cluster ($k = 1$) (Teichgraeber and Brandt, 2018). Once the required k is determined, clustering can be established, as shown in Fig. 2 (Tanaka and Takahashi, 2016).

2.5. Silhouette coefficient

To achieve optimal clustering across all considered aspects, the variables (such as daily mean GSI, daily mean DF, etc.) were iteratively

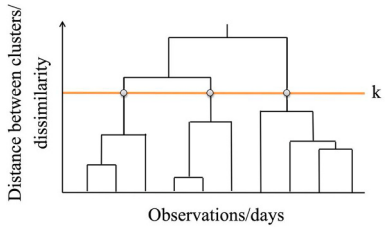


Fig. 2. Agglomerative hierarchical clustering (AHC) is depicted through a dendrogram, illustrating the agglomerative process and optimal cluster determination. This visual representation delineates the hierarchical clustering process, wherein observations or days are grouped based on their similarity. The y-axis denotes the dissimilarity or distance between clusters. A specific cut-off (k) is highlighted, signifying the optimal number of clusters formed.

adjusted. Optimal k was determined using the silhouette function (Sup. 1) proposed by Rousseeuw, 1987, as utilized in previous studies (Watanabe et al., 2021); subsequently, clusters corresponding to k were selected among four to ten clusters. The silhouette coefficient is used to calculate the consistency within data clusters and is a measure of how similar an object is to its cluster compared with other clusters. The silhouette coefficient for each object, denoted as s(i), is defined as shown in Eq. 6 (Rousseeuw, 1987):

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{6}$$

Here, $a(i)$ denotes the average dissimilarity of the i^{th} object to all other objects in the same cluster and $b(i)$ denotes the minimum average dissimilarity of i^{th} object compared with all objects in any other clusters except the one it belongs to. This minimum average dissimilarity is computed as shown in Eq. 7, which represents the dissimilarity of the object $b(i)$ to the second-best cluster (Rousseeuw, 1987):

$$b(i) = mind(i, C) \tag{7}$$

where $d(i, C)$ denotes the average dissimilarity of the object (i) to all objects in C, which refers to all clusters except the cluster that (i) belongs (Watanabe et al., 2021). We identified five clusters based on the analysis of the Silhouette coefficient, which reached its optimum value, ensuring a robust clustering structure. These clusters were subsequently named as the Spectral Characteristics Index (SCI) categories, as they effectively encapsulate the comprehensive dynamics of solar radiation (SR) characteristics over a given day. These characteristics encompass the SSI, fluctuations of the GSI and DF, short-term GSI variability measured as the 15-min MSG, and plant-perceivable light signals measured as CWRs. Simultaneous consideration of all variables in optimizing cluster numbers and clustering variables introduces complexity as enhancing one criterion may inadvertently compromise another criterion. This phenomenon is inherent to multidimensional data analysis. Despite this challenge, the aim of this study was to maximize intracluster similarity and intercluster dissimilarity across all cluster evaluation criteria when introducing the SCI.

2.6. Criteria for model performance

During cluster evaluation, we employed the ANOVA test, followed by multiple comparison tests, to observe statistically significant differences among clusters. Model performances were evaluated using the coefficient of determination (R^2_{adj}; Eq.8), the root mean square error (*RMSE*; Eq. 9), and normalized root mean square error (*nRMSE*) that could be calculated as *RMSE*/ range (maximum value - minimum value of the data set).

$$R^2 = \frac{\left(\sum_{i=1}^{n}(E_i - \overline{E}_i)(M_i - \overline{M}_i)\right)}{\left(\sum_{i=1}^{n}(E_i - \overline{E}_i)^2 \sum_{i=1}^{n}(M_i - \overline{M}_i)^2\right)} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(E_i - M)^2}{n}} \tag{9}$$

where n represents the number of data points and Ei and Mi represent the simulated and measured values, respectively. Then, we trained ML models to replicate the same clustering this approach provides a simple and efficient way to cluster days based on SR and generate SCI categories without using the previously mentioned complicated data analyzing process (2.2 to 2.6), thereby rendering the model readily available for use for new data. This not only saves computational time but also increases the efficiency of the process, thus enhancing the practical applicability of this study.

2.7. Training and selection of ML models

Building upon the framework outlined in the introduction section, this study aimed to reproduce the clustering results using advanced ML models, leveraging readily available environmental data, including the GSI, DF, VP, and AM. Various sophisticated ML models were employed, including naive Bayes (Kwon et al., 2019), decision trees (Singh et al., 2022), discriminant analysis (Gastón-Romeo et al., 2011), ensemble methods (Sevas et al., 2024), neural networks (Mfetoum et al., 2024), Kernel methods (Zhang et al., 2021), KNN (Zhang et al., 2021), SVM (Ayodele et al., 2019), and efficient logistic regression (Ghimire et al., 2022).

The aforementioned models were trained using different variables, and the best-performing model was selected based on R^2_{adj}, *RMSE* and *nRMSE* metrics. Upon evaluating the performance of these ML models, SVM emerged as the most effective model and was selected for further use in this study.

2.8. Support vector machine (SVM)

SVM is a supervised learning model (Vapnik, 2000) and is widely utilized in SR clustering and modeling owing to its higher generalization capability (Chapelle et al., 1999) and superior performance compared to traditional statistical methods and other ML models like decision trees and linear regression, especially when handling complex and nonlinear data (Ali Ou Salah et al., 2021). Among our other trained ML models, SVM exhibited higher accuracy; therefore, we employed SVM as our ML model. Moreover, our model was found to perform better for SVM with a quadratic kernel function. Thus, we used SVM with a quadratic function to replicate the SCI. The SVM model configuration included an automatic kernel scale, a box constraint level of 1, and a one-vs-one multiclass coding strategy. Data standardization was applied to enhance the model's robustness across different features, ensuring consistency in predictions. This approach facilitates the application of the SCI where comprehensive SSI data collection may not be feasible. Moreover, it simplifies the prediction of comprehensive SR conditions. Intricate data processing (Fig. 3) was accomplished using MATLAB R2023b software.

3. Results

3.1. Comparison of clusters

Significant variations were observed among the SCI categories with respect to the GSI, DF, and 15-min MSG (with $p = 0$ and *F-statistic* values of 3155, 6508, and 417, respectively; Table 1; Fig. 4; Sup. 2). The highest mean GSI value was observed for SCI-01, while the lowest GSI values were observed SCI-04 and SCI-05 (Fig. 4a; Table 2). Moreover,

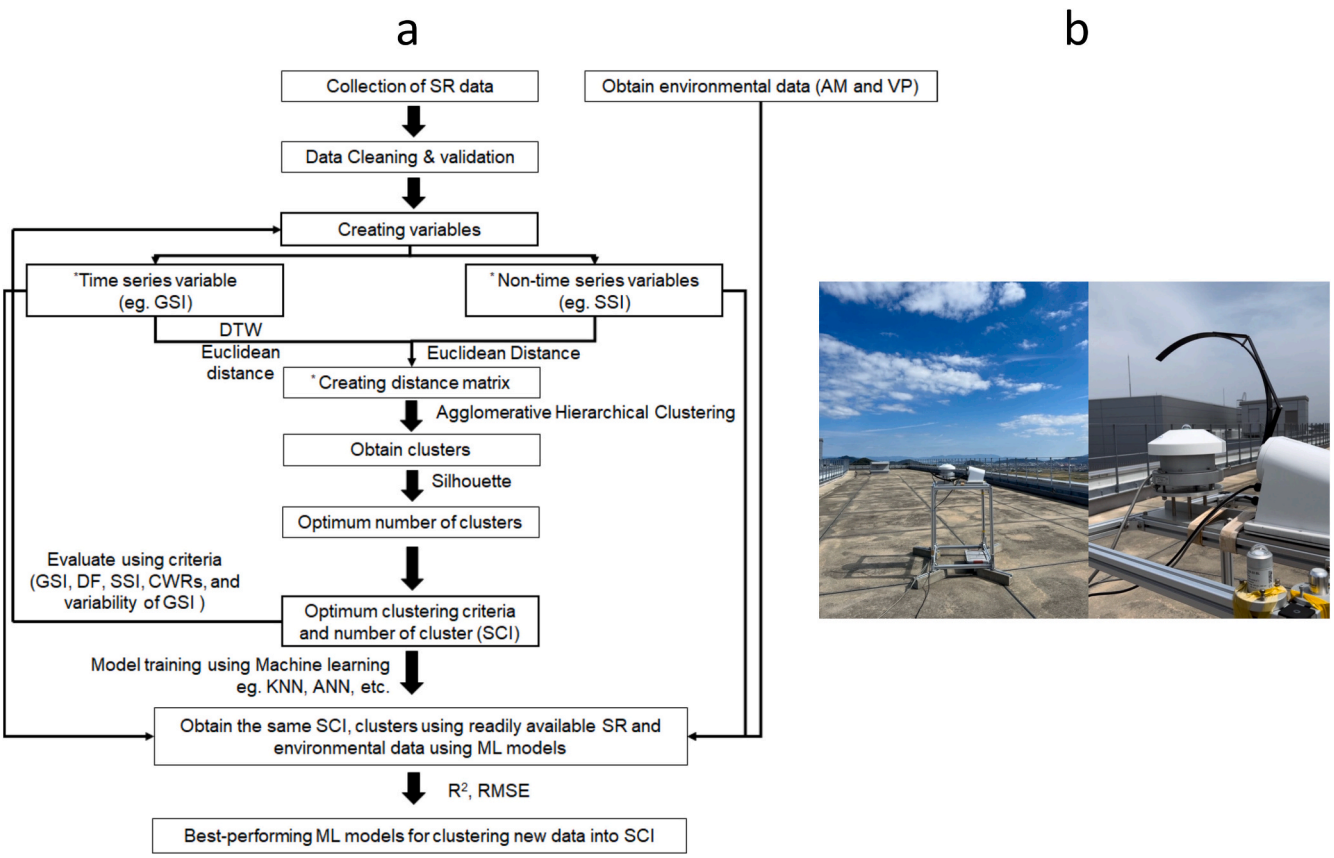


Fig. 3. **(a)** Methodological flowchart of day clustering by solar radiation characteristics using agglomerative hierarchical clustering: a machine learning approach. In this figure, the symbols convey the following meanings: "*" signifies data that has been tested both with and without normalization; "R^{2*}" represents the coefficient of determination; "*RMSE*" stands for root mean square error; "SCI" refers to the spectral characteristics index, and "ML" denotes the machine learning component of the methodology. **(b)** instrument setup in the field, arrangement of radiometric measurements EKO MS-711 spectroradiometer, and EKO MB-22 rotating shadow band (RSB).

Table 1
Analysis of variance (ANOVA) test statistics (GSI, DF, 15-min MSG) and daily means quantum flux of important wavelength, wavelength ranges, and plant-perceivable CWR among SCI categories.

Description	Wavelength/ Wavelength Range (nm)	ANOVA test statistics
R-630	630	109***
B-470	470	103***
UV-B	300–315	690***
UV-A	315–400	9950***
UV	300–400	6346***
B-light	450–495	11188***
G-light	500–600	25680***
R-light	620–680	15960***
FR	700–750	12720***
Near IR	750–1000	35872***
PAR	400–700	68287***
GSI	300–1100	255***
R/B ratio		226***
R/G ratio		185***
B/G ratio		268***
R/FR ratio		51***
UV-A/UV-B ratio		27***
UV-B/B ratio		43***
UV/PAR ratio		232***
UV/GSI ratio		203***
15 min MSG		159***
DF		6508***

*** $p < 0.0001$.

the lowest mean DF value was observed for SCI-01, while the mean DF values of SCI-04 and SCI-05 observed high compared to others SCI categories (Fig. 4b; Table 2). The highest mean 15-min MSG was observed for SCI-05, followed by SCI-03 and SCI-4 (Fig. 4c, Table 2). The second lowest mean 15-min MSG value was observed for SCI-01, while the lowest mean 15-min MSG value was observed for SCI-02 (Fig. 4c; Table 2). Additionally, the highest average SSI of clusters was observed for SCI-01, while the lowest was observed for SSI-05 (Fig. 4d). Generally, from SCI-01 to SCI-05, the daily mean GSI and SSI decreased, while the daily mean DF increased. A considerable change in GSI variability was detected among clusters as the 15-min MSG changed. The daily mean DF and GSI exhibited clear variations among SCI categories (Fig. 5a, b).

Our analysis revealed significant variations in the wavelength spectra encompassing R-light, B-light, green (G) -light (λ 500–530 nm), UV-A-light, UV-B-light, UV-light, IR-light, FR-light, PAR-light, and GSI-light among the SCI categories($F > 690$, $p = 0.000$; Table 1; Sup. 3). Furthermore, the specific wavelengths R-630 nm and B-470 nm demonstrated considerable differences among the SCI categories ($F > 100$, $p < 0.0001$; Table 1). This highlights the marked variability in the spectral characteristics across the SCI categories.

The plant-perceivable CWRs of R/FR, UV-A/UV-B, and UV-B/B exhibited significant differences among the SCI categories ($F > 25$, $p < 0.0001$; Table 1). Moreover, the R/G ratio demonstrated significant variation among the SCI categories ($F > 160$, $p < 0.0001$; Table 1). Likewise, the R/B ratio, UV/PAR, and UV/GSI ratio displayed considerable differences across the SCI categories ($F > 200$, $p = 0.0001$; Table 1). These findings collectively highlight the substantial variability in spectral characteristics, particularly in CWRs across the SCI

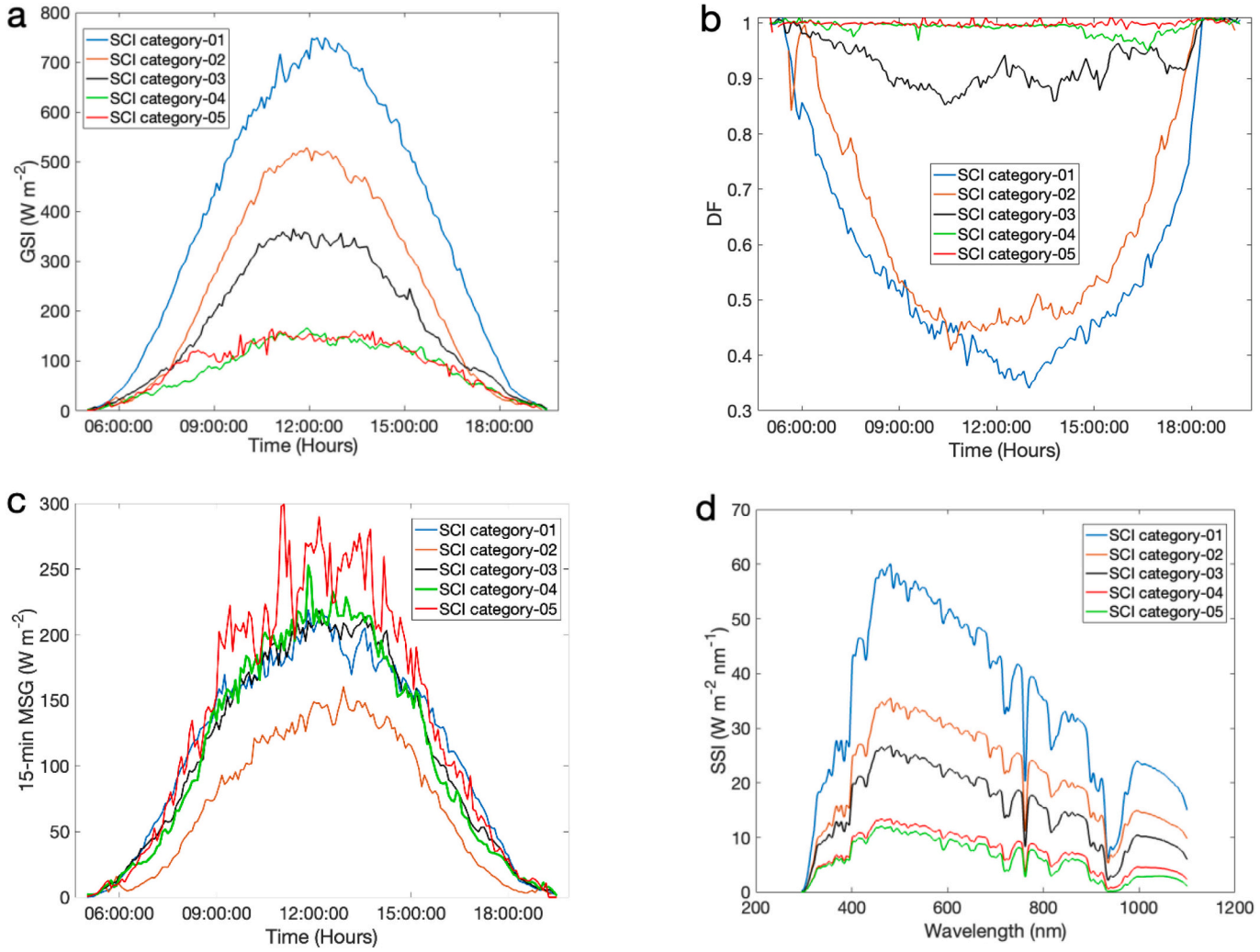


Fig. 4. Comparative analysis of key solar radiation parameters across SCI categories: **(a)** daily variation in GSI, **(b)** daily variation in DF, **(c)** daily variation of 15-min MSG, and **(d)** daily variation of SSI.

Table 2
Daily average and maximum GSI, 15-min MSG, and average daily DF.

SCI category	scale	GSI (W m^{-2})			scale	15-min MSG (W m^{-2})		scale	DF	
		median	mean	max		mean	max		mean	min
1	a	393.06	377.49	749.62	c	112.48	219.14	d	0.594	0.34
2	b	244.92	251.56	529.06	d	73.36	160.40	c	0.650	0.41
3	c	152.67	174.71	365.89	b	108.49	218.46	b	0.931	0.85
4	d	86.64	83.86	166.40	b	109.05	252.86	a	0.992	0.95
5	d	101.28	92.43	164.81	a	133.81	299.32	a	0.999	0.98

In the post hoc test results, we observed significant differences among the clusters. These differences were clearly illustrated using the Likert scale ratings "a" through "e," with "a" signifying the highest mean value and "e" the lowest. This scale was instrumental in emphasizing the significant variations among the clusters.

categories. Notably, SCI-02 exhibited the highest daily mean UV-A/UV-B ratio ($\mu = 49.61$) and lowest UV-B/B ratio ($\mu = 0.0109$) (Fig. 6a, b; Table 3), distinguishing itself from all other clusters.

Considering other CWRs, it is evident that the R/B ratio was the highest for SCI-02 ($\mu = 1.5769$) and the lowest for SCI-05 ($\mu = 1.3727$). A similar trend was observed for the R/G ratio, which was the highest for SCI-02 and the lowest for SCI-05 ($\mu = 0.6005$). Furthermore, the B/G and R/FR ratios ($\mu = 0.4085$ and $\mu = 1.2608$, respectively) were the lowest for SCI-02, whereas these ratios ($\mu = 0.4448$ and $\mu = 1.5144$, respectively) were the highest for SCI-05. In general, these findings reveal a distinct pattern across the different SCI categories.

Furthermore, the UV/PAR and UV/GSI ratios ($\mu = 0.0986$ and $\mu = 0.0514$, respectively) were the highest for SCI-05, whereas these ratios ($\mu = 0.0770$ and $\mu = 0.0337$, respectively) were the lowest for SCI-01. Generally, from SCI-01 to SCI-05, an ascending trend was observed for UV/PAR, UV/GSR, B/G, and R/FR, whereas a descending trend was observed for R/B and R/G. However, no prominent trend was observed for UV-A/UV-B and UV-B/B, although a contrasting behavior was observed for these ratios.

3.2. SCI category description

SCI-01: This SCI category demonstrated the highest average GSI at $\mu = 393.06$ W m^{-2}, peaking at approximately 12:00 JST with a peak value

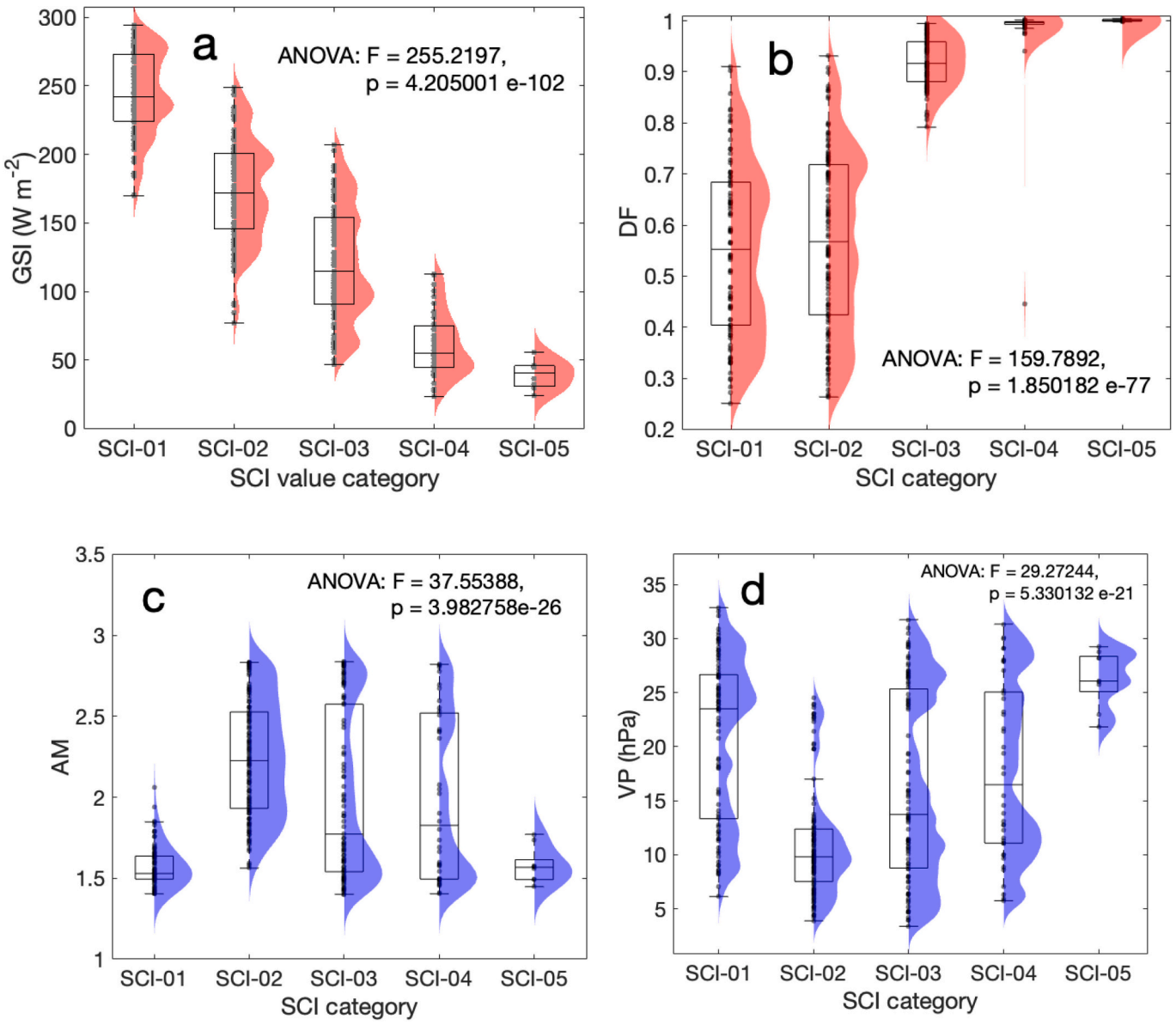


Fig. 5. Variation of daily mean SR (GSI and DF), as well as other environmental parameters such as AM and VP across SCI categories, was observed using raincloud plots with box plots. The plots were constructed to observe the variation among SCI categories for the following parameters: **(a)** daily mean GSI **(b)** daily mean DF **(c)** daily mean AM, and **(d)** daily mean VP.

of $\mu = 750$ W m^{-2}. The 15-min MSG of this category was moderate ($\mu =$ 112.48 W m^{-2}), indicating high variability in the GSI (Fig. 4c; Sup. 2c; Table 2). The average DF remained below 0.6 from 8:00 to 17:00 JST, with the mean DF reaching 0.594. Notably, the R/B ratio ($\mu = 1.5371$), R/G ($\mu = 0.6345$), and UV-B/B ($\mu = 0.0141$) were the second highest in this SCI category. Conversely, the UV/PAR ratio (μ = 0.0817), UV/GSI ratio ($\mu = 0.369$), B/G ratio ($\mu = 0.4129$), and R/FR ratio ($\mu = 1.3212$) ranked the second lowest in this SCI category. However, the UV-A/UV-B ratio ($\mu = 39.4792$) was the lowest for this SCI category. In general, a high GSI with moderate variation was observed for this category, as indicated by the low 15-min MSG. Further spectral analysis revealed a bias for red light than those of blue light comparison to other SCI categories. This was substantiated by the obtained R/B, R/G, and B/G ratios.

SCI 2: This SCI category demonstrated a high average GSI at $\mu =$ 244.92 W m^{-2}, peaking at 10:30–13:30 JST with a peak value of $\mu =$ 529.06 W m^{-2}. The lowest average 15-min MSG ($\mu = 73.36$ W m^{-2}) was recorded for this category, indicating low variability in the GSI. The average DF remained below 0.6 from at 8:30–16:00 JST, with the mean DF at 0.650 (Fig. 4b). The UV-A/UV-B ratio at $\mu = 49.6073$, R/B ratio at $\mu = 1.5769$, and R/G ratio at $\mu = 0.6439$ were the highest for this category. Conversely, the UV-B/B ratio at $\mu = 0.0109$, UV/PAR ratio at $\mu = 0.0770$, Fig. 6c, UV/GSI ratio at $\mu = 0.033$; Fig. 6d), B/G ratio at $\mu =$ 0.4085, and R/FR ratio at $\mu = 1.2608$ were the lowest for this category. Generally, this category was characterized by a high GSI with minimal variability, as demonstrated by the low 15-min MSG. Further spectral analysis revealed a tendency toward red light than those of blue light comparison with SCI-03 to SCI-05 (Fig. 6; Sup. 4, 5). This was substantiated by the obtained R/B, R/G, and B/G ratios.

SCI 3: This SCI category demonstrated a moderate GSI at $\mu = 152.67$ W m^{-2}, peaking at 10:30–14:00 JST with a peak value of $\mu = 365.89$ W m^{-2} (Fig. 4a). Moreover, it exhibited a high average 15-min MSG at $\mu =$ 108.49 W m^{-2}, indicating high variability in the GSI (Fig. 4c). The average DF remained below 0.91 from 8:30 to 17:50 JST, with the mean DF being 0.931 (Fig. 4b). Notably, the UV-A/UV-B ratio at $\mu = 44.5886$, UV/GSI ratio at $\mu = 0.0391$, UV/PAR ratio at $\mu = 0.0865$, R/B ratio at μ

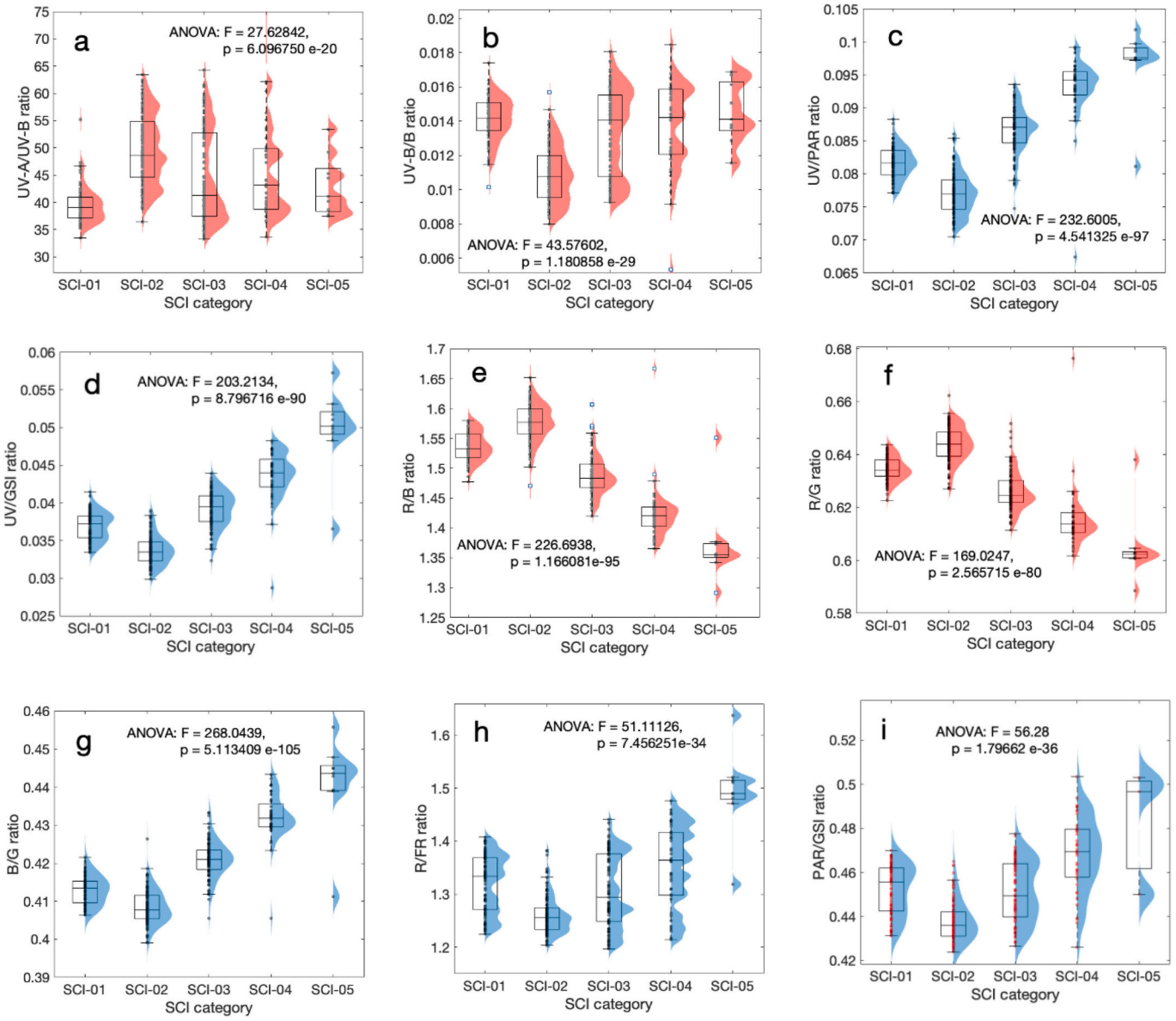


Fig. 6. Variation of daily mean photon flux ratios of UV ratios R, B, G, and FR ratios was observed across SCI categories using raincloud plots with box plots. The plots were constructed to observe the variation among SCI categories for the following ratios: **(a)** UV-A/UV-B ratio, **(b)** UV-B/B ratio, **(c)** UV/PAR ratio, **(d)** UV/GSI ratio, **(e)** R/B ratio, **(f)** R/G ratio, **(g)** B/G ratio, **(h)** R/FR ratio, **(i)** PAR/GSI ratio.

Table 3
UV-A/UV-B ratio, UV-B/B ratio, UV/GSI ratio, UV/PAR ratio, R/B ratio, R/G ratio, B/G ratio, and R/FR ratios among SCI categories.

SCI Category	UV-A/UV-B ratio		UV-B/B ratio		UV/GSI ratio		UV/PAR ratio		R/B ratio		R/G ratio		B/G ratio		R/FR ratio	
	scale	mean	scale	mean	scale	mean	scale	mean	scale	mean	scale	mean	scale	mean	scale	mean
01	d	39.4792	a	0. 0141	d	0.0369	d	0.0817	b	1.5371	b	0. 6345	d	0.4129		1.3212
02	a	49.6073	b	0. 0109	e	0.0337	e	0.0770	a	1.5769	a	0. 6439	e	0.4085		1.2608
03	b, c	44.5886	a	0. 0134	c	0.0391	c	0.0865	c	1.4880	c	0. 6263	c	0.4210		1.3060
04	b, c	46.0903	a	0. 0137	b	0.0436	b	0.0930	d	1.4252	d	0. 6157	b	0.4322		1.3541
05	c, d	43.0235	a	0. 0146	a	0.0514	a	0.0986	e	1.3727	e	0. 6005	a	0.4448		1.5144

In the post hoc test results, we observed significant differences among the clusters. These differences were clearly illustrated by the Likert scale ratings "a" through "e," with "a" signifying the highest mean value and "e" the lowest. This scale was instrumental in emphasizing the significant variations among the clusters.

$= 1.4880$, R/G ratio at $\mu = 0.6263$, B/G ratio at $\mu = 0.4210$, and R/FR ratio at $\mu = 1.3060$ were moderate compared with those for other SCI categories. However, the UV-B/B ratio at $\mu = 0.0134$ was the second lowest among those for all the SCI categories. In general, this category was characterized by a moderate DF and a moderate GSI that exhibited considerable variability. Further spectral analysis revealed a more balanced spectrum compared to other SCI categories.

SCI 4: This SCI category demonstrated a low GSI at $\mu = 86.64$ W m^{-2}, peaking at 09:30–15:35 JST with a peak value of $\mu = 166.40$ W m^{-2} (Fig. 4a). It also exhibited a high average 15-min MSG at 109.05 W m^{-2}.

The DF was below 0.9 only during the morning and evening, while it remained above 0.9 for most of the day. The mean DF for this SCI category was 0.992. The daily means of the UV-A/UV-B ratio at μ = 46.0903, UV/GSI ratio at $\mu = 0.0436$, UV/PAR ratio at $\mu = 0.0930$, B/G ratio at μ = 0.4322, and R/FR ratio at μ = 1.3541 were the second highest among the SCI categories. Conversely, the R/B ratio at μ = 1.4252 and R/G ratio at $\mu = 0.6157$ were the second lowest among those for all the SCI categories. The UV-B/B ratio at μ = 0.0137 was intermediate among those for all the SCI categories. In general, this category was characterized by a remarkably high DF and a low GSI that exhibited significant variability. Further spectral analysis revealed a tendency to have higher levels of blue light than red light compared to the SCI-01 to SCI-03 categories. This was substantiated by the R/B, R/G, and B/G ratios (Fig. 6e, f, g).

SCI 5: This SCI category demonstrated a lower GSI at μ = 101.28 m^{-2}, peaking at 08:00–16:30 JST with a peak value of $\mu = 164.81$ W m^{-2} (Fig. 4a). It also exhibited the highest average 15-min MSG at μ = 133.81 W m^{-2}. All days within this category showed considerably high DF values, exceeding 0.9, with a mean DF of 0.999, the highest among all the SCI categories. The daily mean values of UV-B/B ratio at μ = 0.0146, UV/GSI ratio at $\mu = 0.0514$, UV/PAR ratio $\mu = 0.0986$, B/G ratio $\mu = 0.4448$, and R/FR ratio $\mu = 1.5144$ (Fig. 6h) were the highest for this SCI category compared with those for the other categories. In contrast, the R/B ratio ($\mu = 1.37270$) and R/G ratio ($\mu = 0.6005$) were the lowest for this category. The UV-A/UV-B ratio (μ = 43.0235) was the second lowest among those for all the SCI categories. Overall, this category was characterized by a remarkably high DF throughout the day and a low GSI that exhibited significant variability. Further spectral analysis revealed a tendency to have higher levels of blue light than red light compared to SCI-01 to SCI-03 categories. This was substantiated by the R/B, R/G, and B/G ratios (Fig. 6e, f, g). The PAR/GSI ratio was increased from SCI-01 to SCI-05 (Fig. 6i).

When considering other environmental parameters, the AM and its variability were low for SCI-01 and SCI-05. Conversely, other SCI categories showed a bias toward high AM values with notable variability, particularly SCI-02, which had a considerably high AM value (Fig. 5c). The ANOVA test (F = 37.5538, $p < 0.001$; Fig. 5c) confirmed that the differences in the AM across different SCI categories were statistically significant. Regarding daily mean VP values among the SCI categories, SCI-01 and SCI-05, exhibited the highest median value, unlike the AM (Fig. 5c, d). Other SCI categories also showed a bias toward low VP values with high variability, contrasting with AM. The ANOVA test revealed significant differences in VP among the SCI categories (F = 29.27244, $p < 0.001$; Fig. 5d).

3.3. Model description

In this study, we utilized a variety of readily obtainable environmental and SR data at different intervals (hourly and 5-min) and daily mean data to replicate clustering. Several ML models that have previously been used to cluster SR data were trained for this purpose. Further, optimum cluster replication was achieved via the SVM-based ML models using daily means of the AM, VP, GSI, DF, and 15-min MSG (Sup. 6, 7).

The models were trained using 80 % of the available data, while the remaining 20 % of the data was used for evaluating the performance of the models in replicating the clusters. Performance assessment of each model was based on accuracy and total cost for both validation and test datasets, employing 5-fold cross-validation, a robust method for assessing the generalizability of ML models. In our comparison, the SVM-quadratic model achieved the highest accuracy (94.29 % on the test dataset, 88.03 % on validation) and the lowest error metrics ($nRMSE = 0.0797$, $RMSE = 0.3605$, $MSE = 0.1299$) among the tested ML models (Sup. 7). The quadratic kernel effectively captured nonlinear patterns in SR data, crucial for differentiating SCI clusters. Moreover, the SVM-quadratic model showed high computational efficiency (~6300 predictions/s) and compactness (58 kB), making it practical applications compared to other ML models. Moreover, our model demonstrated strong predictive power, as indicated by high coefficients of determination. The confusion matrices for validation and test datasets offered a comprehensive overview of the performance of the SVM model and demonstrated its effectiveness in clustering days according to their SR characteristics (Fig. 7a, b). The robustness of the SVM model was further substantiated by analyzing receiver operating characteristic (*ROC*) curve results. The area under the curve values, a key performance metric, were notably high for all five clusters, exceeding the threshold of 0.85 (Fig. 7c).

4. Discussion

Understanding the variation in the GSI is crucial as its variability considerably affects terrestrial climate (Vázquez, 2004). Our clustering approach categorized natural SR conditions into five SCI categories (Sup. 1), which grouped the GSI and DF into four considerably distinct groups (Fig. 4a, b). The stochastic nature of cloud cover caused short-term GSI fluctuations (Tomson and Tamm, 2006), which we identified using 15-min MSG data. The results section details SR characteristics in each SCI category, aligning with our aim to facilitate understanding of these categories. Clouds modulate both shortwave (100–3000 nm) and longwave (> 3000 nm) radiation, influencing plant productivity and water use efficiency (Hughes et al., 2024). The SCI methodology offers a novel approach to clustering SR fluctuations, with significant implications for plant ecophysiology under varying cloud conditions.

Our clustering analysis revealed that categories with a high DF predominantly exhibit a high 15-min MSG (Fig. 4b, c), indicating significant GSI fluctuations. Our observations suggest that rainy days, characterized by the stratus and nimbostratus clouds, contribute to the high variability in the GSI. This high variability is primarily attributed to considerable fluctuations in diffuse radiation. This dynamic light condition can induce plant stress and ecophysiological variation compared with those under static light (Vialet-Chabrand et al., 2017). These findings emphasize the importance of considering the quantitative and qualitative aspects of SR to understand its impact on plant stress and overall ecosystem productivity.

The DF plays a critical role in net ecosystem exchange (NEE) (Urban et al., 2007). Our results indicate that under natural conditions, from clear-sky SCI-01 to overcast-sky SCI-05 the DF increases in contrast to decreasing GSI. Consequently, NEE may not necessarily decrease with decreasing GSI and may be compensated with diffuse radiation due to enhanced light penetration into the canopy. This enhanced light penetration can reduce light saturation and photoinhibition in the upper leaves while increasing photosynthesis in the lower leaves (Zhao et al., 2023) facilitating canopy photosynthetic production.

Considering the UV signals, particularly UV/PAR and UV/GSI ratios, SCI categories differentiate days into five clusters; these ratios increase from SCI-01 to SCI-05. Literature evidence suggests that the negative effect of UV-B irradiation on plants can be reduced by high PAR intensity (Jadidi et al., 2023). This indicates the potential for inducing UV stress, even under low radiation levels, in higher SCI categories. Additionally, SCI effectively clustered days into three distinct UV-A/UV-B groups (Fig. 6a; Sup. 4a) and two UV-B/B groups (Fig. 6b; Sup. 4b). Thus, the SCI categories enabled us to consider UV stress along with the DF, GSI, and their variations simultaneously. These variations potentially lead to substantial changes in environmental conditions, public health, and ecosystem dynamics. Understanding such variations in daily SR, in addition to GSI, may be crucial for the development of many UV-related models.

We observed considerable variation in plant-perceivable light signals among the SCI categories. From SCI 01 to SCI-05, the R/B and R/G ratios decreased, whereas the B/G and R/FR ratios increased (Fig. 6g, h). Such pronounced variation among CWRs may trigger differential solar signals crucial for photomorphogenesis, plant ecophysiology, ecosystem, and animal behavior. Our results have demonstrated the effectiveness of

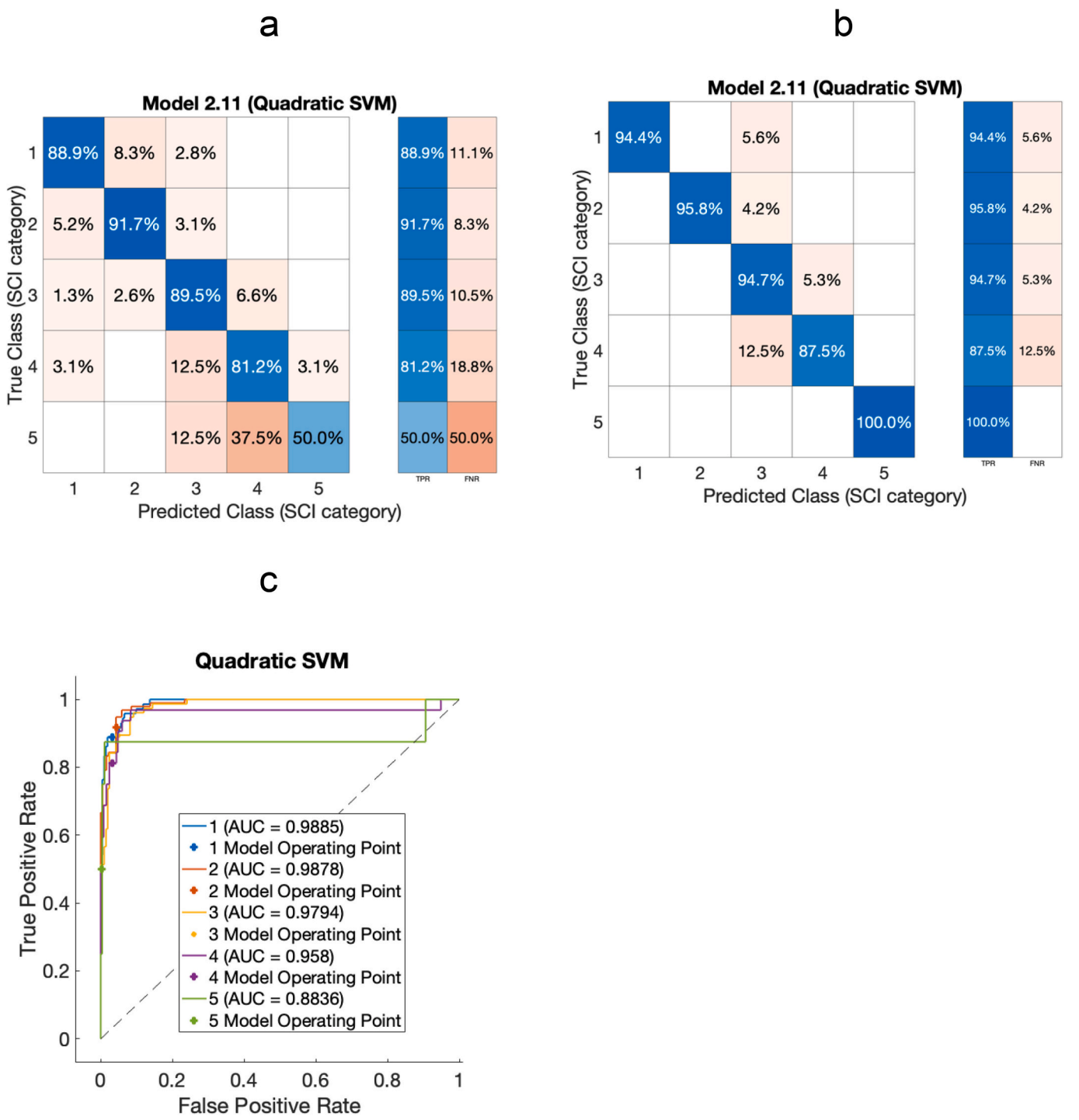


Fig. 7. Comprehensive assessment of quadratic SVM model performance in clustering SR according to SCI categories: **(a)** confusion matrices illustrating classification accuracy quadratic SVM with validation data, **(b)** confusion matrices displaying classification accuracy quadratic SVM with test data, and **(c)** ROC curve analysis for model robustness and the area under the curve (AUC) values.

using SCI categories to identify differences in CWRs (Fig. 6). Overall, distinct patterns across different SCI categories exhibit unique biologically important spectral characteristics, paving the way for incorporating light signals, alongside energy flux and the DF, into new models for ecosystem dynamics and plant ecophysiology.

SCI-01 represents high red light levels typical of clear-sky days, with peak solar irradiance and minimal cloud interference (Navrátil et al., 2007). SCI-02 also indicates clear-sky conditions but with lower, more consistent GSI and distinct spectral differences (CWRs), likely due to seasonal variations, as only a few days in the SCI-02 category belong to summer. Compared with summer, during other seasons, the northern hemisphere leans from the Sun, resulting in a decrease in solar altitude and an increase in AM, which may explain the aforementioned discrepancies in SR characteristics. SCI-03 is characterized by partially cloudy-sky, while SCI-04 features thick cloud cover and limited rainfall (Fig. 8). SCI-05 is associated with stormy weather, high DF, low GSI, and significant rainfall. The main difference between SCI-04 and SCI-05 spectral balance might be a result of cloud type and morphological properties of the cloud as they influence the SR penetration and surface SR characteristics (Pfister et al., 2003).

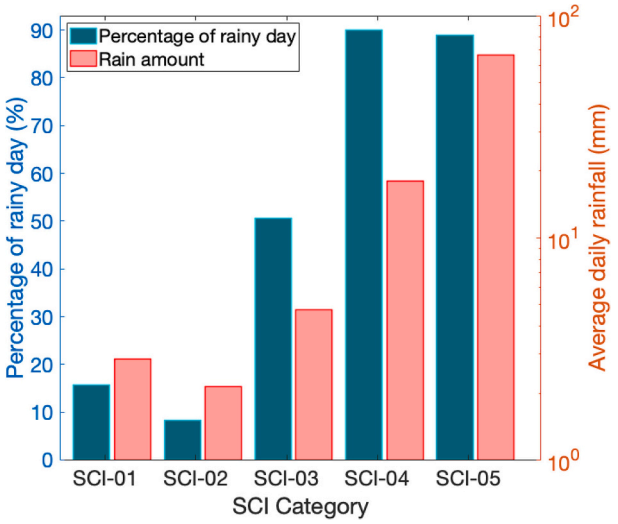


Fig. 8. Comparative rainfall patterns, the percentage of rainy days, and average daily rainfall across different SCI categories.

Our study aligns with recent findings that cloud-induced variations in light quality can lead to substantial changes in photosynthetic activity and overall plant growth (Hughes et al., 2024).

The SR environment is influenced by weather conditions and/or cloud distribution; for example, clear skies, partly cloudy, cloudy, or rain as well as specific cloud types such as cirrus, cumulus, stratus, and nimbus. Differences in weather and cloud cover considerably affect terrestrial SR characteristics. However, the relationship between daily weather conditions and the SR environment has rarely been quantitatively evaluated. This study demonstrates that the relationship between the daily atmospheric environment and SR characteristics.

This study shows significant variations in key environmental factors, such as AM and VP, as well as SR characteristics across SCI categories (Fig. 5). Among the tested ML models, the SVM-quadratic model exhibited the highest accuracy and lowest error metrics, demonstrating its ability to effectively capture nonlinear patterns in addition to its balanced performance and explaining its consistent accuracy across SCI categories (Fig. 7). Its computational efficiency and compact design further establish it as the optimal choice for replicating SCI clusters (Sup. 7), by utilizing readily available SR and environmental data without relying on complex SSI measurements. Thus, the model is particularly suited for scenarios where direct spectral measurements are impractical. Although the SVM-quadratic model demonstrates high accuracy and efficiency, expanding its application with long-term datasets from diverse climatic zones could enhance its generalizability and adaptability to other regions. Despite this limitation, the model's robustness makes it a valuable tool for clustering days based on SR characteristics and generating SCI categories efficiently.

The SCI categories offer a valuable framework for integrating complex SR dynamics into plant-ecophysiological, ecological, and agricultural models, offering potential advancements in research and practical applications. Combining SCI clusters with plant growth data enhances our understanding of SR variations and their influence on plant development, ecosystem interactions, and agricultural productivity. Additionally, it provides opportunities to optimize lighting conditions in agriculture, potentially boosting crop yields and improving ecological modeling.

5. Conclusion

Our research successfully achieved the clustering of SR data based on both energy flux and quality by introducing five SCI categories. The ANOVA test confirmed significant differences in the DF, GSI, and plant-perceivable critical wavelength ratios across clusters ($p < 0.0001$; $F > 27$). The SCI categories provided a simplified yet comprehensive representation of SR conditions. The SCI categories ranged from clear-sky conditions (SCI-01) to overcast-sky conditions (SCI-05). From SCI-01 to SCI-05, the SR energy flux decreased, while the DF and UV ratios increased. Additionally, spectral shifts from red to blue occurred from SCI-01 to SCI-05. Moreover, this study successfully introduced an SVM-based ML model to replicate clustering patterns using readily available SR data (daily means of the DF, GSI, and 15-min MSG) and supplementary environmental data (AM and VP), thus circumventing the need for more difficult to obtain solar spectral irradiance (SSI) data. The model demonstrated 94 % accuracy (test data) with improved efficiency. A major limitation of this study is that it is based on one year of data from a single geographical location. Although the findings of this study reflect considerable advancements in SR and ecological research, future long-term data collection might fine-tune the SCI categories and geographical variations in terms of SCI applicability. In addition to the ecological and plant-ecophysiological perspectives, extending the use of SCI categories to real-world applications such as precision agriculture and urban planning could further validate the practicality of SCI in the future. This SCI categories address the demand for more precise SR classifications for enhance modeling in ecosystem dynamics, plant ecophysiology, and photovoltaic applications.

CRediT authorship contribution statement

Amila Nuwan Siriwardana: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Atsushi Kume:** Writing – review & editing, Methodology, Investigation, Funding acquisition, Conceptualization.

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.

Acknowledgment

Tomoko Kawaguchi Akitsu provided technical advice on the spectroradiometer used in this study. The Faculty of Agriculture at Kyushu University facilitated the installation of the spectroradiometer. This study was also supported by the Japan Society for the Promotion of Science, KAKENHI Grant Number JP18H02511.

Data availability

The data supporting the findings of this study are publicly available in the data repository under the DOI: https://doi.org/10.48708/7324454.

References

Ağbulut, Ü., Gürel, A.E., Biçen, Y., 2021. Prediction of daily global solar radiation using different machine learning algorithms: evaluation and comparison. Renew. Sust. Energ. Rev. 135, 110114. https://doi.org/10.1016/J.RSER.2020.110114.

Ali Ou Salah, H., Oukarfi, B., Bahani, K., Ramdani, M., Moujabbir, M., 2021. Forecasting hourly solar radiation using a novel hybrid technique based on machine learning models. In: Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning, 1, pp. 135–142. https://doi.org/10.5220/0010729800003101.

Ayodele, T.R., Ogunjuyigbe, A.S.O., Amedu, A., Munda, J.L., 2019. Prediction of global solar irradiation using hybridized k-means and support vector regression algorithms. Renew. Energy Focus 29, 78–93. https://doi.org/10.1016/J.REF.2019.03.003.

Becker, C.F., Boyd, J.S., 1957. Solar radiation availability on surfaces in the United States as affected by season, orientation, latitude, altitude and cloudiness. Sol. Energy 1, 13–21. https://doi.org/10.1016/0038-092X(57)90050-6.

Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., Song, A., 2015. Efficient agglomerative hierarchical clustering. Expert Syst. Appl. 42, 2785–2797. https://doi.org/10.1016/J.ESWA.2014.09.054.

Chapelle, O., Haffner, P., Vapnik, V.N., 1999. Support vector machines for histogram-based image classification. IEEE Trans. Neural Netw. 10, 1055–1064. https://doi.org/10.1109/72.788646.

Chiang, C., Bånkestad, D., Hoch, G., 2020. Reaching natural growth: the significance of light and temperature fluctuations in plant performance in indoor growth facilities. Plants (Basel) 9, 1–18. https://doi.org/10.3390/PLANTS9101312.

Demotes-Mainard, S., Péron, T., Corot, A., Bertheloot, J., Le Gourrierec, J., Pelleschi-Travier, S., Crespel, L., Morel, P., Huché-Thélier, L., Boumaza, R., et al., 2016. Plant responses to red and far-red lights, applications in horticulture. Environ. Exp. Bot. 121, 4–21. https://doi.org/10.1016/J.ENVEXPBOT.2015.05.010.

Durand, M., Murchie, E.H., Lindfors, A.V., Urban, O., Aphalo, P.J., Robson, T.M., 2021. Diffuse solar radiation and canopy photosynthesis in a changing environment. Agric. For. Meteorol. 311, 108684. https://doi.org/10.1016/J.AGRFORMET.2021.108684.

Fortin, J.G., Anctil, F., Parent, L.É., Bolinder, M.A., 2008. Comparison of empirical daily surface incoming solar radiation models. Agric. For. Meteorol. 148, 1332–1340. https://doi.org/10.1016/J.AGRFORMET.2008.03.012.

Franses, P.H., Wiemann, T., 2020. Intertemporal similarity of economic time series: an application of dynamic time warping. Comput. Econ. 56, 59–75. https://doi.org/10.1007/s10614-020-09986-0.

Gastón-Romeo, M., Leon, T., Mallor, F., Ramírez-Santigosa, L., 2011. A morphological clustering method for daily solar radiation curves. Sol. Energy 85, 1824–1836. https://doi.org/10.1016/J.SOLENER.2011.04.023.

Ghimire, S., Nguyen-Huy, T., Deo, R.C., Casillas-Pérez, D., Salcedo-Sanz, S., 2022. Efficient daily solar radiation prediction with deep learning 4-phase convolutional neural network, dual stage stacked regression and support vector machine CNN-REGST hybrid model. Sustain. Mater. Technol. 32, e00429. https://doi.org/10.1016/J.SUSMAT.2022.E00429.

Gürel, A.E., Ağbulut, Ü., Bakır, H., Ergün, A., Yıldız, G., 2023. A state of art review on estimation of solar radiation with various models. Heliyon 9, e13167. https://doi.org/10.1016/J.HELIYON.2023.E13167.

Han, J., Guo, C., Ye, S., Zhang, L., Li, S., Wang, H., Yu, G., 2020. Effects of diffuse photosynthetically active radiation on gross primary productivity in a subtropical coniferous plantation vary in different timescales. Ecol. Indic. 115, 106403. https://doi.org/10.1016/J.ECOLIND.2020.106403.

Hartmann, B., 2020. Comparing various solar irradiance categorization methods – a critique on robustness. Renew. Energy 154, 661–671. https://doi.org/10.1016/J.RENENE.2020.03.055.

Hikosaka, S., Moriyama, F., Goto, E., 2021. Effects of photosynthetic photon flux density and red/blue light ratio on the leaf shape and concentrations of functional and aromatic compounds in sweet basil (*Ocimum basilicum* L.). Hortic. J. 90, 357–364. https://doi.org/10.2503/hortj.UTD-273.

Huang, L., Kang, J., Wan, M., Fang, L., Zhang, C., Zeng, Z., 2021. Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. Front. Earth Sci. 9, 596860. https://doi.org/10.3389/feart.2021.596860.

Hughes, N.M., Sanchez, A., Berry, Z.C., Smith, W.K., 2024. Clouds and plant ecophysiology: missing links for understanding climate change impacts. Front. For. Glob. Chang. 7, 1330561. https://doi.org/10.3389/ffgc.2024.1330561.

Jadidi, M., Mumivand, H., Nia, A.E., Shayganfar, A., Maggi, F., 2023. UV-A and UV-B combined with photosynthetically active radiation change plant growth, antioxidant capacity and essential oil composition of Pelargonium graveolens. BMC Plant Biol. 23, 1–14. https://doi.org/10.1186/S12870-023-04556-6/TABLES/5.

Kang, B.O., Tam, K.S., 2013. A new characterization and classification method for daily sky conditions based on ground-based solar irradiance measurement data. Sol. Energy 94, 102–118. https://doi.org/10.1016/J.SOLENER.2013.04.007.

Kaskaoutis, D.G., Kambezidis, H.D., Kumar Kharol, S., Badarinath, K.V.S., 2008. The diffuse-to-global spectral irradiance ratio as a cloud-screening technique for radiometric data. J. Atmos. Sol. Terr. Phys. 70, 1597–1606. https://doi.org/10.1016/j.jastp.2008.04.013.

Keogh, E.J., Pazzani, M.J., 2000. Scaling up dynamic time warping for datamining applications. In: Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/347090.347153.

Kwon, Y., Kwasinski, A., Kwasinski, A., 2019. Solar irradiance forecast using naïve bayes classifier based on publicly available weather forecasting variables. Energies 12. https://doi.org/10.3390/EN12081529.

Laperre, B., Amaya, J., Lapenta, G., 2020. Dynamic time warping as a new evaluation for DST forecast with machine learning. Front. Astron. Space Sci. 7, 540918. https://doi.org/10.3389/fspas.2020.00039.

Li, T., Yang, Q., 2015. Advantages of diffuse light for horticultural production and perspectives for further research. Front. Plant Sci. 6, 704. https://doi.org/10.3389/fpls.2015.00704.

Lou, S., Li, D.H.W., Chen, W., 2019. Identifying overcast, partly cloudy and clear skies by illuminance fluctuations. Renew. Energy 138, 198–211. https://doi.org/10.1016/J.RENENE.2019.01.080.

Maldonado-Salguero, P., Bueso-Sánchez, M.C., Molina-García, Á., Sánchez-Lozano, J.M., 2022. Spatio-temporal dynamic clustering modeling for solar irradiance resource assessment. Renew. Energy 200, 344–359. https://doi.org/10.1016/J.RENENE.2022.09.113.

McKenzie, R.L., Aucamp, P.J., Bais, A.F., Björn, L.O., Ilyas, M., 2007. Changes in biologically-active ultraviolet radiation reaching the Earth's surface. Photochem Photobiol Sci 6, 218–231. https://doi.org/10.1039/B700017K.

Mfetoum, I.M., Ngoh, S.K., Molu, R.J.J., Nde Kenfack, B.F., Onguene, R., Naoussi, S.R.D., Tamba, J.G., Bajaj, M., Berhanu, M., 2024. A multilayer perceptron neural network approach for optimizing solar irradiance forecasting in Central Africa with meteorological insights. Sci. Rep. 14, 3572. https://doi.org/10.1038/s41598-024-54181-y.

Miller, D.G., Rivington, M., Matthews, K.B., Buchan, K., Bellocchi, G., 2008. Testing the spatial applicability of the Johnson–Woodward method for estimating solar radiation from sunshine duration data. Agric. For. Meteorol. 148, 466–480. https://doi.org/10.1016/J.AGRFORMET.2007.10.008.

Monteith, J.L., Unsworth, M.H., 2013. Microclimatology of Radiation: (iii) Interception by Plant Canopies and Animal Coats. Principles of Environmental Physics. 111–133. https://doi.org/10.1016/B978-0-12-386910-4.00008-1.

Moreno-Tejera, S., Silva-Pérez, M.A., Ramírez-Santigosa, L., Lillo-Bravo, I., 2017. Classification of days according to DNI profiles using clustering techniques. Sol. Energy 146, 319–333. https://doi.org/10.1016/J.SOLENER.2017.02.031.

Navrátil, M., Špunda, V., Marková, I., Janouš, D., 2007. Spectral composition of photosynthetically active radiation penetrating into a Norway spruce canopy: the opposite dynamics of the blue/red spectral ratio during clear and overcast days. Trees 21, 311–320. https://doi.org/10.1007/s00468-007-0124-4.

Nematchoua, M.K., Orosa, J.A., Afaifia, M., 2022. Prediction of daily global solar radiation and air temperature using six machine learning algorithms; a case of 27 European countries. Ecol. Inform. 69, 101643. https://doi.org/10.1016/J.ECOINF.2022.101643.

Pfister, G., McKenzie, R.L., Liley, J.B., Thomas, A., Forgan, B.W., Long, C.N., 2003. Cloud coverage based on all-sky imaging and its impact on surface solar irradiance. J. Appl. Meteorol. 42, 1421–1434.

Polo, J., Gastón, M., Vindel, J.M., Pagola, I., 2015. Spatial variability and clustering of global solar irradiation in Vietnam from sunshine duration measurements. Renew. Sust. Energ. Rev. 42, 1326–1334. https://doi.org/10.1016/J.RSER.2014.11.014.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Schumann, T., Paul, S., Melzer, M., Dörmann, P., Jahns, P., 2017. Plant growth under natural light conditions provides highly flexible short-term acclimation properties toward high light stress. Front. Plant Sci. 8, 681.

Senin, P., 2008. Dynamic Time Warping Algorithm Review. University of Hawaii at Manoa, Honolulu, USA [Internet]. [accessed 2024 Apr 22]. https://www.researchgate.net/publication/228785661_Dynamic_Time_Warping_Algorithm_Review.

Sevas, M.S., Sharmin, N., Santona, C.F.T., Sagor, S.R., 2024. Advanced ensemble machine-learning and explainable ai with hybridized clustering for solar irradiation prediction in Bangladesh. Theor. Appl. Climatol. 155, 5695–5725. https://doi.org/10.1007/S00704-024-04951-5.

Singh, N., Jena, S., Panigrahi, C.K., 2022. A novel application of decision tree classifier in solar irradiance prediction. Mater. Today Proc. 58, 316–323. https://doi.org/10.1016/J.MATPR.2022.02.198.

Siriwardana, A.N, Kume, A., 2024. Annual variability in diffuse ratio and spectral characteristics of solar radiation: Cloud effects in a temperate monsoon region. bioRxiv. https://doi.org/10.1101/2024.10.14.618107.

Smith, H., 1982. Light Quality, Photoperception, and Plant Strategy. Annu Rev Plant Biol 33 (1), 481–518. https://doi.org/10.1146/ANNUREV.PP.33.060182.002405.

Tanaka, Y., Takahashi, M., 2016. Dynamic time warping-based cluster analysis and support vector machine-based prediction of solar irradiance at multi-points in a wide area. Stochastic Syst. Theor. Appl. (SSS) 2016, 210–215. https://doi.org/10.5687/SSS.2016.210.

Tanaka, Y., Takahashi, M., 2017. SVM-based multi-point prediction system of solar irradiance using DTW-based cluster analysis. Trans. ISCIE 30, 305–313.

Teichgraeber, H., Brandt, A.R., 2018. Systematic comparison of aggregation methods for input data time series aggregation of energy systems optimization problems. Comput. Aid. Chem. Eng. 44, 955–960. https://doi.org/10.1016/B978-0-444-64241-7.50154-3.

Tomson, T., Tamm, G., 2006. Short-term variability of solar radiation. Sol. Energy 80, 600–606. https://doi.org/10.1016/J.SOLENER.2005.03.009.

Urban, O., Janouš, D., Acosta, M., Czerný, R., Marková, I., Navrátil, M., Pavelka, M., Pokorný, R., Šprtová, M., Zhang, R., et al., 2007. Ecophysiological controls over the net ecosystem exchange of mountain spruce stand. Comparison of the response in direct vs. diffuse solar radiation. Glob. Chang. Biol. 13, 157–168. https://doi.org/10.1111/j.1365-2486.2006.01265.x.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory (Second Edition). The Nature of Statistical Learning Theory. https://doi.org/10.1007/978-1-4757-3264-1.

Vázquez, M., 2004. Solar variability and climate change. Environ. Chall. Mediterr. 2000–2050, 29–46. https://doi.org/10.1007/978-94-007-0973-7_2.

Vialet-Chabrand, S., Matthews, J.S.A., Simkin, A.J., Raines, C.A., Lawson, T., 2017. Importance of fluctuations in light on plant photosynthetic acclimation. Plant Physiol. 173, 2163–2179. https://doi.org/10.1104/PP.16.01767.

Watanabe, T., Oka, K., Hijioka, Y., 2021. Assessment of characteristics of surface solar irradiance on consecutive days using a self-organizing map and clustering methods. Meteorol. Appl. 28. https://doi.org/10.1002/MET.1984.

Xiao, M., Yu, Z., Cui, Y., 2020. Evaluation and estimation of daily global solar radiation from the estimated direct and diffuse solar radiation. Theor. Appl. Climatol. 140, 983–992. https://doi.org/10.1007/s00704-020-03140-4.

Xiao, Y., Chu, L., Zhang, Y., Bian, Y., Xiao, J., Xu, D., 2021. HY5: a pivotal regulator of light-dependent development in higher plants. Front. Plant Sci. 12, 800989. https://doi.org/10.3389/fpls.2021.800989.

Zhang, Z., Wang, C., Peng, X., Qin, H., Lv, H., Fu, J., Wang, H., 2021. Solar radiation intensity probabilistic forecasting based on K-means time series clustering and Gaussian process regression. IEEE Access. 9, 89079–89092. https://doi.org/10.1109/ACCESS.2021.3077475.

Zhao, P., Bai, Y., Zhang, Z., Wang, L., Guo, J., Wang, J., 2023. Differences in diffuse photosynthetically active radiation effects on cropland light use efficiency calculated via contemporary remote sensing and crop production models. Ecol. Inform. 73, 101948. https://doi.org/10.1016/J.ECOINF.2022.101948.