



Identification of clear-sky periods and cloudiness using diffuse-to-direct spectral ratios of hyperspectral solar irradiance and sky radiance measurements

Masoud Moradi¹ · Johan van der Molen² · Léon Serre-Fredj² · Behnaz Arabi³

Received: 21 December 2024 / Accepted: 2 August 2025 / Published online: 12 August 2025
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2025

Abstract

Accurate identification of clear-sky and cloudy periods from hyperspectral radiometric data is essential for atmospheric and remote sensing studies. Traditional approaches that rely on single-wavelength spectral ratios often fail under broken-clouds due to rapid irradiance fluctuations and aerosol interference. This study introduces a new method based on the spectrally integrated diffuse-to-direct irradiance ratio (*IDR*), which is derived from modeled hyperspectral components and provides a robust indicator of sky conditions. The model is developed using 15-year hyperspectral measurements of solar downwelling irradiance ($E_s(\lambda)$, $\text{mW m}^{-2} \text{ nm}^{-1}$) and sky radiance ($L_s(\lambda)$, $\text{mW m}^{-2} \text{ nm}^{-1} \text{ sr}^{-1}$) collected in a fixed coastal station located at the Marsdiep tidal inlet in the Dutch Wadden Sea. The model is validated against reference clear-sky periods identified using established clear-sky models and detection algorithms based on simultaneous Global Horizontal Irradiance (*GHI*) measurements. Results show that *IDR* thresholds of ≤ 0.35 for clear skies and ≥ 0.70 for overcast conditions achieve high agreement (95% and 92%, respectively) with reference sky classifications. The *IDR* effectively accounts for the influence of rapidly changing cloud cover in intermediate conditions, ensuring robust classification of clear-sky and overcast periods. The classification is further supported by spectral analysis using Principal Component Analysis, which confirm the stability of diffusive component to $E_s(\lambda)$ features and the interferences of $L_s(\lambda)/E_s(\lambda)$ ratio across different sky conditions. This work highlights the importance of the diffuse irradiance component role in hyperspectral sky classification. Future research should validate *IDR* with direct diffuse irradiance measurements to enhance the proposed model under diverse atmospheric conditions.

Keywords Sky classification · Hyperspectral radiometry · Clear-sky detection · Global horizontal irradiance · Wadden sea

Introduction

Clear-sky detection is a critical prerequisite for atmospheric studies and utilization of solar energy. However, clear-sky conditions are not directly measurable and must instead be inferred from broader all-sky datasets using surrogate indicators such as the clearness index, diffuse fraction, or modeled irradiance comparisons (Younes and Muneer 2007; Gueymard 2019a; Correa et al. 2022). This indirect approach poses significant challenges, particularly due to the absence of a universal definition or method for identifying clear skies (Gueymard et al. 2019; Sun et al. 2019). The lack of a standardized, universally applicable clear-sky classification remains a key limitation in solar resource modeling and radiometric quality control. A robust and

Communicated by Hassan Babaie

✉ Masoud Moradi
moradi_msd@yahoo.com

- ¹ Iranian National Institute of Oceanography and Atmospheric Science (INIOAS), Etemadzadeh St, No. 3, Tehran 14118 13389, Iran
- ² Department of Coastal Systems, Royal Netherlands Institute for Sea Research (NIOZ), PO-Box 59, Den Burg 1790 AB, The Netherlands
- ³ Geoinformatics- Spatial Big Data Research Group Faculty of Biology, Chemistry & Earth Sciences, University of Bayreuth, Universitätsstraße 30, 95440 Bayreuth, Germany

spectrally-informed detection method that can distinguish between clear, intermediate, and overcast skies—especially under transitional conditions—would significantly enhance the consistency and reliability of solar and atmospheric datasets used in climate studies and remote sensing (Forstinger et al. 2021; Gueymard 2014a). Generally, the absence of visible clouds across the sky indicates the clear sky condition (Younes and Muneer 2007). The situation is complicated when the finite field of view (FOV) of the sky-looking sensors is taken into account. Since the sky-looking sensors capture a narrow FOV, small cloud patches may obstruct the sensor's view of the sky. Hence, meteorological cloud cover observations associated with measurement platforms, if available, cannot be used directly to identify clear sky conditions. Furthermore, aerosols and precipitable water vapor can vary in time, influencing the transmittance of solar radiation through the atmosphere even when there are no clouds in the sky (Kim and Ramanathan 2008; Reno and Hansen 2016). Therefore, there is a need for clear sky detection methods that are able to accurately identify periods of 'cloudless clear sky' radiation or 'irradiance similar to clear sky'.

The well-known clear sky detection methods applied until now usually rely on Global Horizontal Irradiance (GHI , $W\ m^{-2}$) measurements to identify sky conditions. These methods include GHI -based models and statistical algorithms, each with distinct advantages and limitations. GHI -based models use broadband irradiance measurements to identify clear-sky periods by comparing observed GHI to modeled clear-sky GHI (Younes and Muneer 2007; Sun et al. 2019; Gueymard et al. 2019; Correa et al. 2022). Statistical algorithms employ time-series analysis of irradiance data to detect clear-sky periods, offering robustness to short-term fluctuations (Reno and Hansen 2016; Ruiz-Arias and Gueymard 2023). These clear-sky detection methods are usually validated against ground truth data such as satellite images, and/or ground-based sky images. While most clear-sky detection models have been developed for solar radiation measurements, their direct application to hyperspectral radiometric data remains limited. Hyperspectral measurements provide detailed spectral information across hundreds of narrow wavelength bands, making them highly sensitive to subtle atmospheric changes, including thin clouds, aerosols, and water vapor (Griffin and Burke 2003; Kuusk and Kuusk 2018). Accurate identification of clear-sky conditions from hyperspectral datasets is essential for applications such as satellite validation, atmospheric correction, and the parameterization of radiative transfer models (Garrison 1992; Jordan et al. 2022; Norgren et al. 2021). However, the high spectral resolution also makes these datasets more vulnerable to noise and misclassification under variable sky conditions. Therefore, dedicated clear-sky

detection strategies tailored to hyperspectral data are needed to ensure the spectral integrity and scientific utility of long-term measurements.

Unmanned automated hyperspectral time series measurements are an increasingly cost-effective technology for long-term atmospheric studies and for calibration and validation of satellite remote sensing data. A fixed station of radiometric measurements for atmospheric research consists of sensors for two essential components: downwelling solar irradiance ($E_s(\lambda)$, $mW\ m^{-2}\ nm^{-1}$) and sky radiance ($L_s(\lambda)$, $mW\ m^{-2}\ nm^{-1}\ sr^{-1}$). Simultaneous measurements of $E_s(\lambda)$ and $L_s(\lambda)$ are widely used for remote sensing applications (Jordan et al. 2022; Kutser et al. 2013), atmospheric correction models in remote sensing tasks (Zibordi et al. 2009; Arabi et al. 2016), and atmospheric studies (Griffin and Burke 2003; Kuusk and Kuusk 2018). Notably, most hyperspectral data are collected in unavoidable non-ideal atmospheric conditions which lead to significant uncertainties in the resulting time series datasets. Ideal conditions include a clear sky, low Solar Zenith Angle (SZA), no sun and sky glint, and low concentrations of aerosols. An optimal viewing geometry of the station minimizes the effects of sun and sky glints. Cloud cover significantly increases the sky glint contribution to measurable (ir)radiance, to the point that it can change the shapes and intensities of $E_s(\lambda)$ and $L_s(\lambda)$ spectra (Gege 2012; Mobley 2015). Hence, nearly all of the radiative transfer-based models for the simulation of solar irradiance [e.g., Gregg and Carder (1990), Gege (2012)], optical properties of water bodies (e.g., Mobley 2022), and remote sensing applications [e.g. Simis and Olsson (2013); Gueymard 2019a; Bi et al. (2023)] have been developed for clear sky conditions.

A daily time-series of GHI and E_s at 550 nm [$E_s(550)$] typically shows a bell-shaped probability distribution for clear skies (Pitarch et al. 2020; Lin et al. 2022). On a clear sky day, the $L_s(550)$ to $E_s(550)$ ratio increases from morning to noon and decreases in the afternoon. For cloudy conditions, the values of $1/E_s(\lambda)$ and $L_s(\lambda)$ show an increase in near infra-red (NIR) and a decrease in blue wavelengths (Gueymard 2019b; Ruddick et al. 2019). Accordingly, the temporal variation of $E_s(\lambda)$ for the clear sky conditions is relatively small on a daily time scale and depends on SZA, whereas for overcast or partially cloudy conditions there are significant variations in the magnitude and shape of $L_s(\lambda)$ (Ruddick et al. 2019). Therefore, variations in $L_s(\lambda)/E_s(\lambda)$ ratios at NIR and blue wavelengths are sensitive to sky conditions, which can be used for identification of sky cloudiness (Ruddick et al. 2006; Simis and Olsson 2013; Groetsch et al. 2017). While single-band spectral ratios, such as $L_s(750)/E_s(750)$ or $L_s(400)/E_s(400)$, provide simple method for sky condition classification, they have several important limitations: (i) simplification of atmospheric state: Skies

are rarely purely clear or overcast. Intermediate states (e.g., scattered clouds, thin cirrus, haze) are common, and single-band ratios fail to capture these nuances and may misclassify cloudy or optically thin clouds (Chesnoi et al. 2024; Nyamsi et al. 2024). (ii) Wavelength-specific sensitivity: the NIR bands (e.g., 750 nm) are sensitive to thick clouds and aerosols but less sensitive to thin clouds of fine aerosols. The blue bands (e.g., 400 nm) are highly sensitive to Rayleigh scattering, aerosols, and thin clouds but saturates quickly under thick clouds (Kim et al. 2024). (iii) Aerosol interference: high aerosol loads (e.g., dust, pollution) can elevate sky radiance, mimicking cloud signatures. For instance, heavy aerosols at 750 nm may trigger false overcast classifications (Choi et al. 2021). (iv) Ambiguity of thresholds: the single-band ratios use a threshold value for classification (please see Sect. 4.5). For example $L_S(750)/E_S(750) \leq 0.1$ for clear skies. The gap between clear and overcast thresholds leaves transitional skies unclassified. Values in this range require additional criteria for interpretation. In addition, the threshold values do not accurately represent the boundary between different atmospheric conditions. (v) Lack of spectral context: Clouds and aerosols have spectral signatures. Single-band ratios ignore valuable information from other wavelengths (e.g., water vapor bands) (Jiang et al. 2022). Hyperspectral data is underutilized with this approach. (vi) Sensitivity to sensor calibration: the single-band ratios amplify errors in calibration or sensor drift. At 400 nm, where the magnitude of E_S is low, noise can significantly affect L_S/E_S reliability (Müller 2014). (vii) Solar geometry dependence: the single-band ratios vary with SZA. At low SZA, Rayleigh/path scattering increases, potentially mimicking cloud effects even in clear skies. Fixed thresholds may thus fail at high/low sun angles (Zhang et al. 2015; Gómez-Uceda et al. 2021).

Given the limitations outlined above, there is a need for more robust methods that can integrate multiple wavelengths and/or combine spectral components with bandwidth-based metrics. An $E_S(\lambda)$ spectrum is composed of direct and diffusive irradiance components. The direct component is the radiation that reaches the sensor directly from the sun. The diffusive component is the radiation that reaches the sensor from the sky after scattering in the atmosphere, which also includes backscattered radiation that was reflected by the surrounding earth surface (Gege 2005; Vermote et al. 1997). Besides the prominent absorption peaks of water vapor (≈ 940 nm) and oxygen (≈ 760 nm), these irradiance components contribute additional prominent spectral features to the $E_S(\lambda)$ and $L_S(\lambda)$ spectra. Hence, the magnitude of $E_S(\lambda)$ and $L_S(\lambda)$ spectral features are related to SZA, intensity of solar irradiance, and atmospheric scattering and absorption by molecules, aerosols, and clouds (Kim and Ramanathan 2008; Ruddick et al. 2019). Clouds affect the $E_S(\lambda)$ and

$L_S(\lambda)$ by altering the intensity, shapes, temporal variability, and the balance between direct and diffuse components. The direct radiation is dominant in the clear skies, while the direct radiation is attenuated and the diffuse component becomes dominant under cloudy conditions (Gege 2005). These changes cause a reduction in the intensity of spectral features in the shorter wavelengths (blue light) due to the scattering effects of water droplets and ice crystals in the clouds (Liou 2002). Moreover, clouds can enhance specific absorption features, particularly in the Near-Infrared (NIR) region where water vapor absorption bands are prominent (Pérez-Ramírez et al. 2012). The increased diffuse radiation under cloudy skies tends to have a more uniform spectral distribution compared to direct sunlight, which is more spectrally selective (Ruddick et al. 2019).

In this study, we present a practical method using the direct and diffuse components derived from $E_S(\lambda)$ and $L_S(\lambda)$ measurements. By integrating spectral information across multiple wavelengths, this method provides a robust indicator of sky conditions, improving classification accuracy under variable cloud cover and aerosol conditions, and validated against reference methods. This research comprises the following steps: (i) modeling the clear sky GHI (hereafter GHI_{CS}) using existing optimal clear sky models. (ii) identification of clear sky periods from GHI and GHI_{CS} using existing Clear Sky Detection (CSD) methods. (iii) identification of thick and scattered cloudy periods using an existing model. (iv) modeling the hyperspectral solar irradiance components. (v) developing a method for detecting different sky conditions based on hyperspectral measurements. (vi) validate the new method against the reference sky conditions. The results of steps (ii) and (iii) are used as reference clear sky and overcast conditions, respectively. In addition, the previously known indices are evaluated. This assessment is conducted to demonstrate the performance of spectral ratios and statistical variations for identifying different weather conditions.

Datasets

Hyperspectral data

The Royal Netherlands Institute for Sea Research (NIOZ) owns a jetty station (NJS) in the Marsdiep tidal inlet of the Dutch Wadden Sea. The NJS has been equipped with a set of above-water hyperspectral sensors consisting of 6 TRIOS-RAMSES radiometers: one Ramses-ACC for $E_S(\lambda)$, one Ramses ACC-UV (ultraviolet), and two pairs of Ramses-ARC for $L_S(\lambda)$ looking southeast and southwest. The $E_S(\lambda)$ and $L_S(\lambda)$ spectra have been recorded every 10 min at 1 nm intervals from 350 nm to 950 nm. The ARC sensors have

been setup with viewing angle of $\theta_v = 35^\circ$ and direction of $\phi = 135^\circ$ (south-east) and $\phi = 225^\circ$ (south-west).

The optimal viewing geometry for sensors mounted in a fixed platform is $30^\circ \leq \theta_v \leq 50^\circ$ and $90^\circ \leq \phi \leq 135^\circ$, where θ_v is the zenith plane angle and ϕ is the azimuth angle of sensor viewing direction (Lee et al. 2002; Mobley 1999). The setup of ACC and ARC sensors was selected to optimize the capture of sky radiance under varying solar positions while minimizing sun glint and shading effects from the jetty structure, following recommendations for coastal radiometric measurements at NJS (Wernand 2002; Pitarch et al. 2020). The south-east and south-west directions ensure consistent sampling of sky conditions across the diurnal solar cycle, particularly under the prevailing westerly winds in the Wadden Sea (Moradi et al. 2024). A figure of the ARC sensor setup at NJS is presented in our recent study (Arabi et al. 2025) to illustrate the instrument configuration.

In this study, we used the spectra of $E_S(\lambda)$ and south-east looking $L_S(\lambda)$ ($\phi = 135^\circ$) from 2008 to 2022, with $90^\circ \leq \phi \leq 135^\circ$ and spanning from 6:00 a.m. to 6:00 p.m. UTC. The quality control tests were applied to remove invalid/suspicious spectra. The quality control (QC) procedures, adapted from Moradi et al. (2024), ensure the reliability of the hyperspectral measurements by removing invalid or distorted spectra. These procedures include five groups of criteria: (i) Basic checks for sensor malfunctions or invalid data (e.g., $\text{SZA} > 85^\circ$, negative $E_S(\lambda)$ or $L_S(\lambda)$); (ii) Low-light conditions with $E_S(\lambda_{\max}) \leq 20 \text{ mW m}^{-2} \text{ nm}^{-1}$; (iii) Distorted spectra detection (e.g., $E_S(480)/E_S(680) \leq 1.0$; $E_S(\lambda_{\max})/E_S(865) \leq 1.25$); (iv) High humidity/precipitation flagging, which indicates enhanced water vapor absorption in the near-infrared (NIR) region; and (v) Sun glint/surface perturbation detection ($L_S(800-950) / E_S(800-950) > 0.025 \text{ sr}^{-1}$, $L_S(850) / E_S(850) \geq 0.025 \text{ sr}^{-1}$). These thresholds, particularly for humidity and glint, were chosen to exclude spectra distorted by atmospheric conditions, as high humidity can significantly alter the spectral shape of $E_S(\lambda)$ and $L_S(\lambda)$, rendering the diffuse and direct components unreliable (Sect. 3.3.1). After applying these QC routines, A total of 144,460 quality-controlled spectra were selected for further analysis. The quality controlled hyperspectral data from 2011 to 2022 ($N=119,646$) were utilized for model development, and data from 2008 to 2010 ($N=24,814$) were employed for model assessment.

GHI and meteorological observations

The Royal Netherlands Meteorological Institute (KNMI) owns a network of 35 ground-based automatic weather stations. Details of these stations are available at KNMI website (<https://dataplatform.knmi.nl/dataset/waarneemstation-s-4>). At 30 of these stations, *GHI* is measured at 12-second

temporal resolution using pyranometers (CM11, Kipp & Zonen). The 10-min *GHI* data from the De Kooy station, located approximately 8 km southwest of NJS, collected during 2008–2022 were used to evaluate and validate the clear sky models. In addition, cloud cover (CC, coverage of the upper sky in eighths, 9=upper sky invisible), atmospheric pressure, and relative humidity were obtained simultaneously with the *GHI* observations from De Kooy station ($N=629,569$). Since all clear sky models have been tuned with *GHI* data at $\text{SZA} \leq 85^\circ$, data with $\text{SZA} > 85^\circ$ were excluded. Additionally, the *GHI* data were quality-controlled using the method described by Long and Shi (2006), called QCRAID (codes available at <https://github.com/pvlib/pvanalytics/>). After implementing the quality control routines, a total of 282,818 *GHI* data (44.9%) remained for further analysis.

Methods

An overview of our methodology is presented in Fig. 1. This flowchart is divided into two main parts: one illustrating the extraction of reference atmospheric correction, and the other outlining the methodology for identifying sky conditions using hyperspectral measurements. The details of the methodology are explained in the following sections.

Identification of reference clear sky periods

Here, existing well-known algorithms relying on *GHI* were used to identify the reference periods similar to those occurring during clear sky conditions. Accordingly, the time series of measured *GHI* and a clear sky model were used to identify the reference clear sky periods using a clear sky detection method (CSD). The well-known clear sky models and CSD methods were assessed, and the optimum ones were selected to identify reference clear sky periods.

Clear sky models

Identifying clear sky periods from *GHI* data depends on using a clear sky model, which estimates the *GHI* values at a given location and time (GHI_{CS}). Since clear sky irradiance depends on atmospheric parameters, validation of clear sky models should be conducted for the study area to ensure fairness in the testing procedure, appropriate selection of clear-sky periods, consideration of geographic location, and accurate input variables. Sun et al. (2019) have evaluated 75 clear sky models in five Koppen-Geiger climate regions [i.e., tropical (A), arid (b), temperate (C), cold (D), and polar (E)], as well as for overall global performance. Our study area is classified in class ‘C’ of Koppen-Geiger climatic

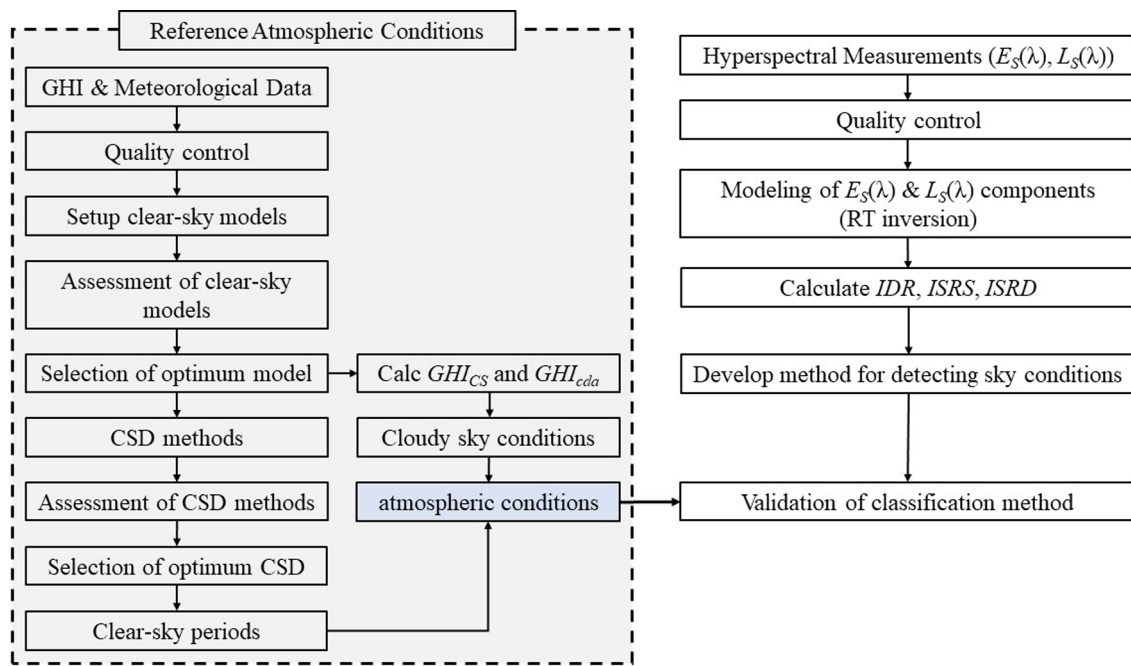


Fig. 1 Workflow of the proposed method for clear-sky detection from hyperspectral radiometric measurements

regions. Their ranking scores of clear sky models for class C ranged from -20.526 to 2.187 , with models scoring > 1.9 achieving predictions with a confidence level of 95%. Hence, we used the five models with the highest-ranking score > 1.9 for class C (their Table 6). The selected clear sky models included MAC2, Bashahu, REST2, Ineichen-Perez, and Yang, in order of ranking score. The ranking scores and the required input variables for the selected clear sky models are given in Table 1. For the formulation and description of models see Sun et al. (2019).

As explained in the introduction section, the measured GHI is not always indicative of a clear sky, even no clouds were observed ($CC=0$). As can be expected, Gueymard et al. (2019) found that the clear sky index, $r_{CS} = GHI_{CS} / GHI$, remained stable and close to one under cloudless conditions. However, $r_{CS} \approx 1$ did not indicate the cloudless sky because of the frequent occurrence of clouds, the albedo effect, and the presence of aerosols (Gueymard 2017). Therefore, a period during which GHI changes were very close to GHI_{CS} is assumed to correspond with clear sky conditions. For this study, conditions were identified as ‘clear sky’ if $CC=0$ AND when the hourly variations of GHI were within the 95% of confidence level of GHI_{CS} variations. This method indicated that the matchups of GHI_{CS} and measured GHI were representative of observed cloudless conditions and reconciled the GHI_{CS} to the measured GHI . The matchups were used to test the performance of the clear sky models.

Input variables of the clear sky models

The specifications and sources of the required input data for the clear sky models (Table 1) are explained below.

- Geographical coordinates, h (elevation, m), and p (air pressure, mbar) were obtained from the De Kooy station observational datasets.
- E_0 (solar constant, $W\ m^{-2}$): The value of E_0 assigned to each model has been determined by its developer. If E_0 was unspecified, it was set to a default value of $E_0 = 1361.1\ W\ m^{-2}$ (Muneer et al. 2000).
- SZA (degree): was calculated following the method of Greene et al. (2019).
- R_G (surface albedo, frac.), τ (aerosol optical depth at 550 nm), α (Angstrom exponent), β (Angstrom turbidity coefficient, i.e., aerosol optical thickness at $\lambda=550$ nm), u_{O3} (total ozone, atm-cm), and PWV (total precipitable water vapor, atm-cm):

Hourly timeseries of these parameters for the study site were obtained from the Modern-Era Retrospective Analysis for Research Applications version 2 (MERRA-2) reanalysis, obtained from the NASA Modeling and Assimilation Office (<https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/>) (Gelaro et al. 2017). Each of the MERRA-2 variables has been validated worldwide against AERONET, particularly the aerosol and column water vapor products (Holben et al. 1998; Mukkavilli et al. 2019). The MERRA-2 reanalysis has been suggested as an ideal candidate for applications of

Table 1 Selected clear sky models and their required inputs in order of ranking score from sun et al. (2019). The input variables are solar constant (E_0 , W m^{-2}), solar zenith angle (SZA), site elevation (h , m), surface albedo (R_G , frac.), air pressure (p , mbar), Linke turbidity factor (T_L), aerosol optical depth at 550 nm (τ), Angstrom exponent of aerosol scattering (α), atmospheric turbidity coefficient (β), total Ozone amount (u_{O_3} , atm-cm), total nitrogen dioxide amount (u_{NO_2} , atm-cm), and total precipitable water vapor (PWV , atm-cm)

Model	score	E_0	SZA	h	R_G	p	T_L	τ	α	β	u_{O_3}	u_{NO_2}	PWV	Ref.
MAC2	2.187	•	•	•	•	•	•	•	•	•	•	•	•	(1)
Bashahu	2.046	•	•	•	•	•	•	•	•	•	•	•	•	(2)
REST2	2.033	•	•	•	•	•	•	•	•	•	•	•	•	(3)
Ineichen-Perez	1.946	•	•	•	•	•	•	•	•	•	•	•	•	(4)
Yang	1.925	•	•	•	•	•	•	•	•	•	•	•	•	(5)

1 - Davies and McKay (1982); 2 - Bashahu and Laplace (1994); 3 - Gueymard (2008); 4 - Ineichen and Perez (2002); 5 - Yang and Koike (2005)

clear sky models (Antonanzas-Torres et al. 2019; Gueymard 2019a). Previous studies indicated that using the MERRA-2 parameters did not lead to significant changes in GHI_{CS} compared to using AERONET observations (Gueymard 2014b, 2019b; Sun et al. 2019). This suggests that using MERRA-2 reanalysis variables as input for clear sky models provides reliable GHI_{CS} datasets. The MERRA-2 dataset was linearly interpolated to 10-min temporal resolution.

u_{NO_2} (total nitrogen dioxide, atm-cm): Only the REST2 model required the u_{NO_2} . According to the REST2 literature, u_{NO_2} is extracted from NASA's Ozone Monitoring Instrument (OMI) product onboard the Aura polar orbiting satellite (<https://aura.gsfc.nasa.gov/omi.html>). The OMI presents a quality controlled daily dataset. Since the OMI instrument is a board a polar-orbiting satellite, it naturally results in missed recordings between flight paths. The OMI product (mol cm^{-2}) is converted to a columnar amount in atmosphere centimeters (atm-cm) by a division by 2.69×10^{16} . As REST2 only works within a range of $0.0 \leq u_{NO_2} \leq 0.03$ atm-cm, a default value of 0.0002 atm-cm is applied for gaps (Gueymard 2008). The OMI dataset was linearly interpolated to 10-min temporal resolution.

- T_L (Linke Turbidity): This factor represents the absorption by the water vapor and the aerosol particles scattering compared to a dry and clean atmosphere (Linke 1992). T_L can be correlated with PWV and β using several empirical methods (Dogniaux 1974; Grenier et al. 1994; Gueymard 1998; Ineichen 2008; Molineaux et al. 1995; Remund et al. 2003). Sun et al. (2019) showed that the Grenier formulation (Grenier et al. 1994) performs best for the Ineichen-Perez clear sky model. Hence the T_L was calculated here using the Grenier formulation:

$$T_L = \frac{\beta - \left(\begin{array}{l} -0.10545 - 0.02005 \times PWV \\ +0.0050689 \times PWV^2 \\ -0.0005202 \times PWV^3 \end{array} \right)}{0.073554 - 0.0029011 \times PWV + 0.00075553 \times PWV^2 - 0.0000783 \times PWV^3} \quad (1)$$

Clear sky detection (CSD) methods

A total of 21 available CSD methods have been reviewed by Gueymard et al. (2019). They recommend utilizing the multi-criteria CSD method and implementing a duration filter on the final method. Since no diffuse and direct horizontal irradiance ground measurements were available, we employed the six CSD methods that utilized only GHI ground measurements as input (Table 2). All the methods required GHI and GHI_{CS} as input variables. In Table 2, Local Solar Time ($LSoT$) is the time according to the sun's

Table 2 Selected clear Sky detection (CSD) methods and their required inputs. LSoT is the local solar time (Eq. 2)

CSD method	Input variables				Ref.
	SZA	GHI	GHI _{CS}	LSoT	
	E_0				
AliaMartinez	▪	▪	▪	▪	Alia-Martinez et al. (2016)
Ellis		▪	▪		Ellis et al. (2018)
Polo	▪	▪	▪	▪	Polo et al. (2009)
Quesada		▪	▪		Quesada-Ruiz et al. (2015)
Reno		▪	▪		Reno and Hansen (2016)
Zhang	▪	▪	▪		Zhang et al. (2018)

position in the sky relative to particular place on Earth. It was calculated following (Kittler and Darula 2013):

$$SLoT = LST + 4 \times (LL - LSTM) + [9.87 \times \sin(2B) - 7.35 \times \cos(B) - 1.5 \times \sin(B)] \quad (2)$$

$$B = 360 \times (doy - 81)/365$$

where LST is local standard time adjusted for daylight saving time, LL is the local longitude (degree), $LSTM$ is the local standard time meridian (degree), doy is the day of the year.

Each CSD method provided a binary classification for each measurement: 0=clear, 1=cloudy. The results of each clear sky model were tested against the clear sky matchups (Sect. 3.1.1) and the optimum CSD method was used as the reference of clear sky periods.

The clear-sky models (Sect. 3.1.1) and CSD methods were applied to identify reference clear-sky periods. These results were used in Sect. 4.1 and 4.2 to achieve the reference clear-sky periods, which then used for classifications and to characterize the spectral behavior of hyperspectral measurements under different sky conditions.

Identification of reference cloudy conditions

For step iii), we use the classification algorithm, CAELUS, developed by Ruiz-Arias and Gueymard (2023) for precise sky classification. They have used GHI data at 1-min resolution and a set of four input variables to classify sky conditions into six cloud classes: 'Cloud enhancement', 'Cloudless', 'Overcast', 'Thin clouds', 'Thick clouds', and 'Scattered clouds'. The CAELUS method requires SZA, GHI , and a good estimate of the associated GHI_{CS} and modeled GHI under cloudless, clean-and-dry atmosphere (i.e., without water vapor and aerosol) (hereafter GHI_{cda}). The result of classification is largely dependent on the accuracy of GHI_{CS} and GHI_{cda} . The thick and scattered cloudy conditions were considered periods of cloudy or intermediate

state between overcast and clear sky conditions. Some clear sky instances might be rejected by CAELUS Cloudless type, leading to biased results especially for low SZA and turbid skies. In addition, the results of CAELUS sky classification should be validated against well-known reference observations. Therefore, we followed the methods outlined in Sect. 3.1 to identify clear sky conditions to validate the results of the CAELUS model.

Detection of clear sky periods from hyperspectral measurements

The magnitude and spectral shape of $E_S(\lambda)$ and $L_S(\lambda)$ are very sensitive to SZA, backscattering, aerosol absorption, and optical state of the atmosphere (Simis and Olsson 2013; McNeill 2017). Hence, the direct and diffuse components of $E_S(\lambda)$ are necessary for investigating the spectral characteristics in different atmospheric conditions. Since no in-situ hyperspectral measurements were available for solar direct and diffuse components, these components were estimated using existing models (Sect. 3.3.1). Then, a new method is presented for identification of clear skies and cloudy conditions (Sect. 3.3.2).

Modeling of $E_S(\lambda)$ and $L_S(\lambda)$ components

In cloudless atmosphere, the $E_S(\lambda)$ spectrum is composed of direct ($E_{Sd}(\lambda)$, $\text{mW m}^{-2} \text{nm}^{-1}$) and diffuse ($E_{Ss}(\lambda)$, $\text{mW m}^{-2} \text{nm}^{-1}$) components (Gregg and Carder 1990). $E_{Sd}(\lambda)$ and $E_{Ss}(\lambda)$ represent the sun's radiation as a direct light source and radiation from sky, respectively. $E_{Ss}(\lambda)$ is a combination of Rayleigh scattering ($E_{Ssr}(\lambda)$, $\text{mW m}^{-2} \text{nm}^{-1}$) and aerosol (Mie) scattering ($E_{Ssa}(\lambda)$, $\text{mW m}^{-2} \text{nm}^{-1}$) (Gege 2005; Gregg and Carder 1990):

$$E_S(\lambda) = f_{dd} \cdot E_{Sd} + f_{ds} \cdot (E_{Ssr} + E_{Ssa}); E_{Ss} = E_{Ssr} + E_{Ssa} \quad (3)$$

where f_{dd} and f_{ds} are the relative intensities of the light source components $E_{Sd}(\lambda)$ and $E_{Ss}(\lambda)$ relative to conditions with undisturbed illumination, respectively. Similarly, $L_S(\lambda)$ is parametrized in cloudless skies as follows (Gege and Grötsch 2016):

$$L_S(\lambda) = g_{dd} \cdot E_{Sd}(\lambda) + g_{dsr} \cdot E_{Ssr}(\lambda) + g_{dsa} \cdot E_{Ssa}(\lambda) \quad (4)$$

where the g_{dd} , g_{dsr} , and g_{dsa} (in unit of sr^{-1}) represent the weights of relative intensity of each light source component. Since $E_{Sd}(\lambda)$, $E_{Ssr}(\lambda)$, and $E_{Ssa}(\lambda)$ indicate downwelling irradiance, the components of $L_S(\lambda)$ in Eq. 4 indicate the direct and diffusive sky radiance recorded by a sensor for SZA, θ_s , and ϕ coordinates. Hence, the two components of $g_{dd} \times E_{Sd}(\lambda)$ (hereafter L_{S-sun}) and $g_{dsr} \times E_{Ssr}(\lambda) + g_{dsa} \times E_{Ssa}(\lambda)$

(hereafter L_{S-sky}) are indicative of the sun light source and sky radiance, respectively (Gege 2005). A consistent formulation and parametrization of E_{Sd} , E_{Ssr} , and E_{dsa} were given by Gregg and Carder (1990). They expressed the $E_{S*}(\lambda)$ as a function of E_0 , SZA, PWV , α , and β , scale height Ozone (Hoz , atm-cm), atmospheric pressure, and relative humidity. More details about the formulation and description of the models are given by Gege (2012) and Groetsch et al. (2017).

In this study, the Gregg and Carder (1990) model is not used to simulate irradiance in absolute physical units under cloudy conditions. Rather, we invert its parameterized formulation to decompose $E_S(\lambda)$ into relative $E_{Sd}(\lambda)/E_{Ss}(\lambda)$, using only the observed spectral shapes and their statistical properties, not absolute magnitudes. This inversion is performed using a bounded genetic algorithm optimization, where the model acts as a generic spectral shape-fitting tool, rather than a strict physical radiative transfer model. The physical parameters retrieved in this inversion (i.e., f_{dd} , f_{ds}) represent effective scaling factors that best approximate the $E_S(\lambda)$ spectrum under the constraints of a radiative transfer-based decomposition (Eq. 3). These scaling factors are not interpreted directly as physically accurate quantities, but rather as statistical proxies that allow us to reconstruct the spectral partitioning of downwelling irradiance, which is our main classification metric. This approach is supported by Jordan et al. (2022), who conducted a similar inversion using direct hyperspectral measurements of diffuse and direct components from a novel hyperspectral pyranometer (HSP). Their findings showed that even when used outside its strict physical domain, the Gregg and Carder (1990) model — when inverted using spectral fitting techniques — can yield accurate ratios of $E_{Sd}(\lambda)/E_S(\lambda)$ and $E_{Ss}(\lambda)/E_S(\lambda)$ for sky condition assessment. Importantly, they found that the inversion remained robust for characterizing atmospheric state transitions, especially when paired with high-resolution, quality-controlled spectral observations — the same context in which we operate. Furthermore, our goal is not to retrieve radiometrically accurate fluxes or atmospheric constituents under cloudy skies, but to identify transitions in the spectral balance between direct and diffuse components that are diagnostic of sky conditions.

Consequently, the $E_{Sd}(\lambda)$, $E_{Ssr}(\lambda)$, and $E_{dsa}(\lambda)$ were modelled using inversion of Eq. 3 by applying the parameter bounds and initial guesses in Table 3. Similarly, the g 's (Eq. 4) were estimated using the inversion of $L_S(\lambda)$ and the parameters bound in Table 3. The min, max, and initial guesses of α , β , f_{dd} , and f_{ds} were adopted from Pitarch et al.

(2020) specified for NJS; for PWV and Hoz from (Groetsch et al. 2017); and for g_{dd} , g_{dsr} , and g_{dsa} from Gege and Grötsch (2016) in clear sky conditions and with a viewing direction avoiding sun glint reflection.

The inversion models were implemented in MATLAB (Mathworks, Inc.) using genetic algorithms to minimize residuals between measured [i.e., $E_S(\lambda)$ and $L_S(\lambda)$] and simulated spectra [i.e., $E_S^m(\lambda)$ and $L_S^m(\lambda)$] (Saraswat and Sharma 2013). A least-squares minimization was applied to minimize the residual-sum-of-squares between $L_S(\lambda)/E_S(\lambda)$ and $L_S^m(\lambda)/E_S^m(\lambda)$. The f_{dd} , f_{ds} , g_{dd} , g_{dsr} , and g_{dsa} were treated as fit parameters in inverse modeling. The mutation and crossover of the generic algorithm optimization were determined using MATLAB built-in functions. The initial guess of mutation was taken proportional to the standard deviation of $L_S(\lambda)$ and $E_S(\lambda)$, and it decreased in each optimization step as a function of the shrink parameter. For crossover parametrization, a half fraction of data was considered. The highest efficiency of the genetic algorithm was found with a crossover fraction of 0.4 and a reduction of the shrink parameter of 0.6 using 1001 iterations. The models were considered converged when the residuals between $E_S^m(\lambda)$ and $L_S^m(\lambda)$ and the corresponding observed spectra were less than 5%.

Clear sky and overcast identification model

Here, we propose a new method to identify clear sky and overcast conditions, based only on hyperspectral observations from a fixed station in combination with a modelled separation of the downwelling radiation into the direct contribution from the sun and the contribution by atmospheric scattering. The method uses thresholds for integrals over relevant parts of the spectra.

In general, the spectral shape of $E_S(\lambda)$ remains relatively similar under all atmospheric conditions (Sect. 4.6). In contrast, the ratios of $E_{Sd}(\lambda)/E_S(\lambda)$ and $E_{Ss}(\lambda)/E_S(\lambda)$ are very sensitive to atmospheric scattering and solar irradiance (Simis and Olsson 2013; Jordan et al. 2022). In clear sky conditions, $E_{Sd}(\lambda)$ and $E_{Ss}(\lambda)$ are entirely derived from sun downwelling irradiance and atmospheric scattering (Rayleigh and aerosol), respectively. Conversely, the magnitude of $E_{Ss}(\lambda)$ increases at the expense of $E_{Sd}(\lambda)$ when clouds start to build (Gregg and Carder 1990; Kim and Ramathan 2008). Therefore, the ratio of $E_{Sd}(\lambda)/E_S(\lambda)$ can indicate the presence of clouds, approaching one in overcast conditions and fluctuating rapidly in scattered-cloud conditions. This approach has been used for identification of clouds,

Table 3 Parameter bounds and initial guesses for above water model inversion at NJS

	β	α	PWV	Hoz	f_{dd}	f_{ds}	g_{dd}	g_{dsr}	g_{dsa}
Min	0.1	0.1	0.1	0.1	0.0	-0.02	0.001	0.01	0.01
Max	1.5	2.5	5.0	5.0	1.0	1.0	1.0	π	π
Init	1.0	1.5	2.5	0.4	0.1	0.1	0.02	$1/\pi$	$1/\pi$

estimation of aerosol optical depth, and characterization of atmospheric optical state using in situ measurements of diffuse downwelling irradiance (Norgren et al. 2021; Jordan et al. 2022).

Here, we adopted the spectrally integrated diffuse ratio (*IDR*) using the modeled diffuse irradiance, defined (Jordan et al. 2022):

$$IDR = \frac{\int_{\lambda_1}^{\lambda_2} E_{Ss}(\lambda) \cdot d\lambda}{\int_{\lambda_1}^{\lambda_2} E_S(\lambda) \cdot d\lambda} \quad (5)$$

where $\lambda_1 = 385$ nm and $\lambda_2 = 900$ nm. Following Simis and Olsson (2013), a convolutional third-order Savitzky-Golay smoothing filter with 70 nm width was applied on both $E_{Sd}(\lambda)$ and $E_S(\lambda)$ spectra to suppress high-frequency spikes without distorting the signal tendency. *IDR* ranged from zero for clear skies to one for overcast conditions. *IDR* values for intermediate cloudy conditions lay between clear and overcast values with higher variability for scattered clouds and lower variability for haze conditions. Figure 2 shows an illustrative example of *IDR*, $E_S(\lambda)$, and $L_S(\lambda)$ variability in different atmospheric conditions.

Notably, during clear skies where the $E_S(550)$ spectra show a regular bell shape in the daily time series, there may be high levels of Rayleigh aerosol scattering. In such cases, the modeled diffuse components of solar irradiance become dominant to the expense of the modeled direct components. For very high levels of atmospheric scattering, $E_{Ss}(\lambda)$ and $L_{S-sky}(\lambda)$ become saturated and *IDR* tends to higher values similar to those observed for overcast conditions, while $E_S(550)$ and *GHI* show a pattern of clear sky conditions (Fig. 3, top row). Furthermore, $L_{S-sun}(\lambda)$ comprises a relatively small component in the $L_S(\lambda)$ spectrum (less than 10% in this study) (Fig. 2, second row). Higher values indicate the sun glint, the spectra for which have been excluded during the quality control process. During overcast conditions with high and thick cloud cover, the absorption of downwelling irradiance begins to dominate (Gueymard et al. 2019; Norgren et al. 2021; Jordan et al. 2022). In turn, this reduces the $E_{Ss}(\lambda)$ and $L_{S-sky}(\lambda)$ to levels comparable to those observed in clear skies. Since, $E_S(\lambda)$ and $L_S(\lambda)$ are the sum of direct and diffuse components (Eqs. 3 and 4), $E_{Sd}(\lambda)$ and $L_{S-sun}(\lambda)$ tend to increase (Fig. 3, second row). Therefore, *IDR* tends towards lower values similar to those observed for clear skies.

To detect the aforementioned atmospheric effects and eliminate them from *IDR*, the spectrally integrated sky diffuse ratio (*ISRS*) and integrated sky direct ratio (*ISRD*) were used, which are defined as:

$$ISRS = \frac{\int_{\lambda_1}^{\lambda_2} L_{S-sky}(\lambda) \cdot d\lambda}{\int_{\lambda_1}^{\lambda_2} L_S(\lambda) \cdot d\lambda} \quad (6)$$

$$ISRD = \frac{\int_{\lambda_1}^{\lambda_2} L_{S-sun}(\lambda) \cdot d\lambda}{\int_{\lambda_1}^{\lambda_2} L_S(\lambda) \cdot d\lambda} \quad (7)$$

where $\lambda_1 = 385$ nm and $\lambda_2 = 900$ nm. The $L_S(\lambda)$, $L_{S-sky}(\lambda)$, and $L_{S-sun}(\lambda)$ spectra were filtered using the same third-order Savitzky-Golay smoothing filter as was applied to *IDR*. $ISRS < 1.0$ and $ISRD < 0.01$ indicated a normal atmospheric state without significant cloud absorption or saturation of downwelling diffuse components. Therefore, clear skies were identified using $IDR \leq X$ and $ISRD < 0.01$, and overcast conditions with $IDR \geq Y$ and $ISRS < 1.0$. Here, $X = 0.35$ and $Y = 0.70$ are thresholds calculated using the statistical distribution of *IDR* during reference sky periods (see Sect. 4.3).

Assessment of methods

A confusion matrix of the results of the CSD methods/model (Sect. 3.1.3) and the corresponding observed/reference data was used to evaluate their performance. Four classes were defined in the confusion matrix: False Negative (FN), False Positive (FP), True Negative (TN), and True Positive (TP). In this context, TP and TN denoted that the proposed method correctly classified the sky conditions. FP was used when a cloudy sky indicated as clear, and FN was used when a clear sky indicated as cloudy. High TP, low FP, and low FN values with a relaxed concern on TN indicated high performance.

However, in general, it is difficult to manipulate the results of the confusion matrix into a ranking score. Therefore, the four criteria above were summarized using the six most common combination metrics:

- **F-score** is a measure of predictive performance and defined as (Sasaki 2007):

$$F_Z = \frac{(1 + Z^2) \times TP}{[(1 + Z^2) \times TP + Z^2 \times FN + FP]} \quad (8)$$

where Z is constant that considers the weight of TP. Here, two values $Z = 1$ and 2 were considered, resulting F1 and F2 scores. The F2-score is a variation of the F1-score, placing more emphasis on FP and combining TP and FP into a single score. The maximum value of $F_Z = 1.0$ indicated perfect classification, while the minimum value of $F_Z = 0.0$ indicated complete absence of TP classifications.

- **Youden's J** statistics is a single statistic that gives equal weights to TP and TN, as (Powers 2020):

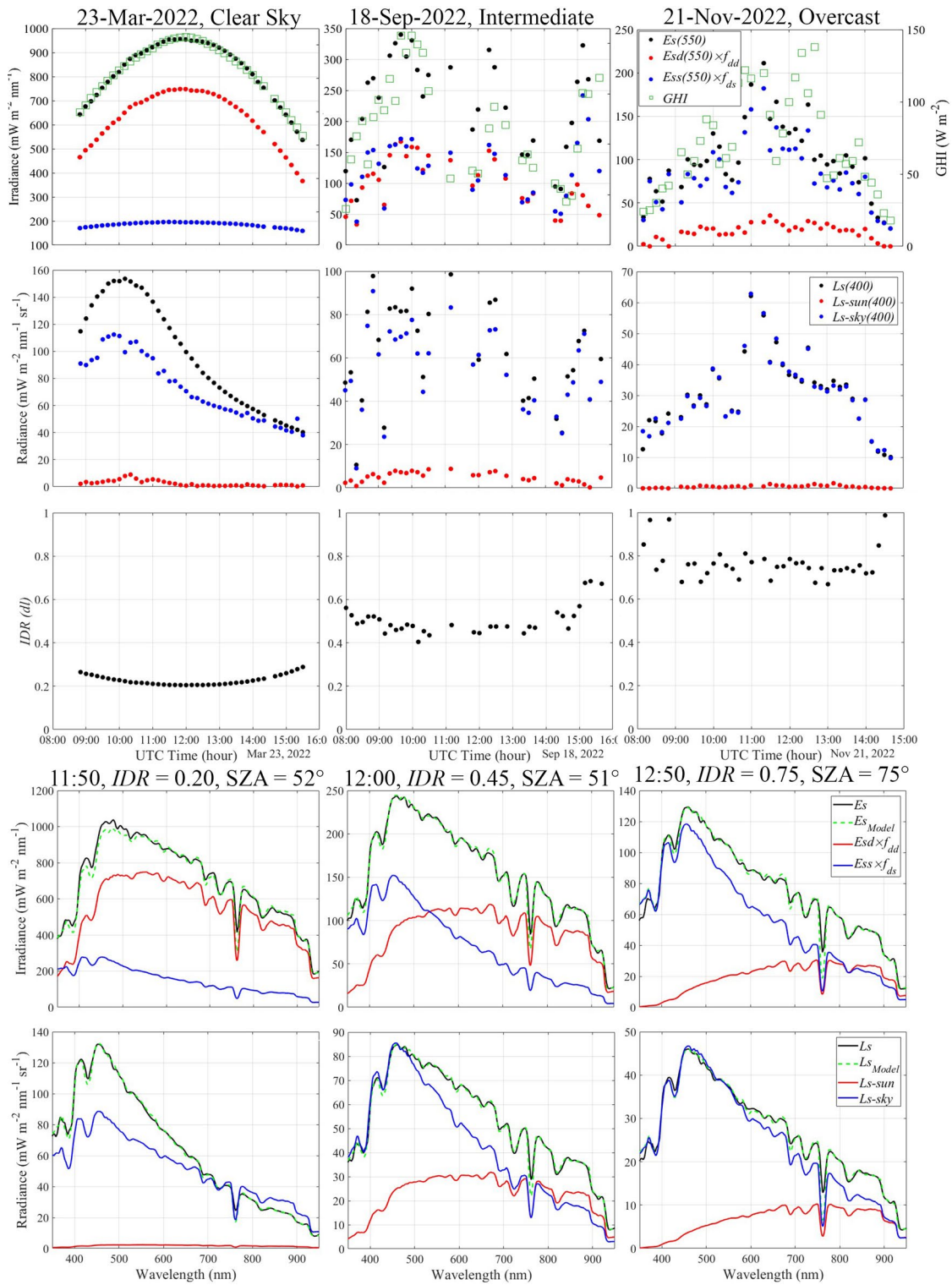


Fig. 2 Examples of daily time series of GHI & $E_{S^*}(\lambda)$ (top row), $L_{S^*}(\lambda)$ (second row), and IDR (third row) in clear skies (left column, $SZA \leq 69^\circ$), intermediate clouds (center column, $SZA \leq 74^\circ$), and over-

cast conditions (right column, $SZA \leq 79^\circ$). The rows 4 and 5 show the spectra of $E_{S^*}(\lambda)$ and $L_{S^*}(\lambda)$ at noon. The legend of each row is shown in the right column

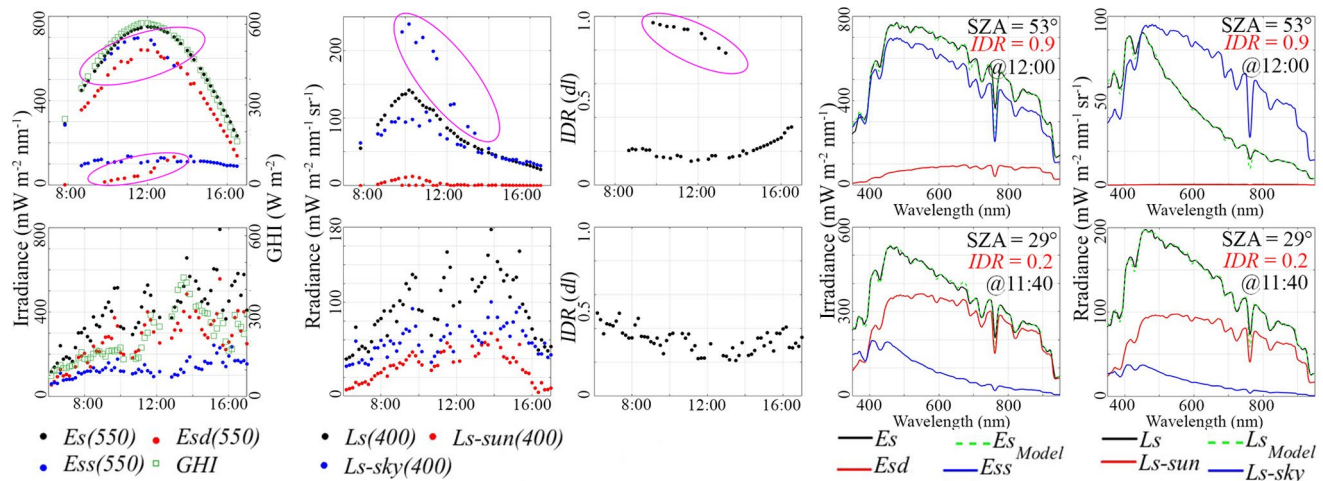


Fig. 3 **top row)** Example of daily time series of GHI , $E_s(\lambda)$, and $L_s(\lambda)$, 21-Mar-2012 (UTC time), $SZA \leq 73^\circ$. The ellipses show the position of a high level of atmospheric scattering that correspond to saturation of diffuse components and lead to high values of IDR (> 0.8), while they

have been recorded under clear skies. **bottom row)** Similar to the top row on 23-Jun-2021, under overcast conditions, $SZA \leq 67^\circ$. The $IDR \leq 0.48$, similar to those observed for clear skies. Please refer to the text

$$J = \frac{TP}{(TP + FN)} + \frac{TN}{(TN + FP)} - 1 \quad (9)$$

Youden's J values range from -1 to 1 , indicating completely false and perfectly true classification, respectively.

- **Matthew Correlation Coefficient (MCC)** is a balanced measure of the confusion matrix classes and takes all classes, especially when there is a significant difference in class sizes. It is calculated as (Matthews 1975):

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{[(TP + FP) \times (TP + FN)] \times [(TN + FP) \times (TN + FN)]}} \quad (10)$$

MCC values ranged from -1 to 1 , indicating the perfect classification of every case and bad agreement of the matrix, respectively.

- **Positive Predictive Value (PPV)** gives the portions of positive instances that were correctly indicated.

$$PPV = \frac{TP}{(TP + FP)} \quad (11)$$

- **Negative Predictive Value (NPV)** gives the portions of negative instances that were correctly indicated.

$$NPV = \frac{TN}{(FN + TN)} \quad (12)$$

- **Accuracy (ACC)** gives the portions of the true positive and negative instances that were correctly indicated.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (13)$$

These metrics were interpreted following standard performance thresholds, where values above 80% for accuracy and F1-score indicate strong agreement with reference classifications. Precision and recall were assessed to account for any class imbalance and to ensure the method performs consistently across sky types. These metrics collectively reflect the robustness and reliability of the classification scheme (Chandaka et al. 2009; Ruiz-Arias and Gueymard 2023).

Spectral variability and principal component analysis

To characterize the spectral variability of hyperspectral measurements across clear-sky, overcast, and intermediate conditions, Principal Component Analysis (PCA) was applied to the quality-controlled $E_s(\lambda)$, $L_s(\lambda)$, $L_s(\lambda)/E_s(\lambda)$, and $E_{ss}(\lambda)/E_s(\lambda)$. PCA was performed using builtin MATLAB 'pca' function. The spectra were first normalized by subtracting the mean and dividing by the standard deviation for each wavelength to ensure equal weighting across the spectral range. PCA was applied separately to each dataset, grouped by sky condition as classified by the IDR . The analysis extracted the first three principal components (PC1–PC3), which accounted for most of the variance in each dataset, to identify dominant spectral features and assess their stability across sky conditions. Eigenvectors were analyzed at prominent spectral feature to evaluate the influence

of atmospheric scattering and absorption. The method of extracting prominent spectral feature has been described in our recent work (Moradi et al. 2024).

Statistical metrics

To quantify the consistency between models and observations, the statistical parameters of coefficient of determination (R^2), normalized root mean square error (nRMSE), unbiased root mean square error (uRMSE), normalized mean absolute error (nMAE), and mean ratio (MR) were used.

$$nRMSE = \frac{\sqrt{\overline{(x_{mod} - x_{obs})^2}}}{\overline{x_{obs}}} \quad (14)$$

$$uRMSE = \frac{\sqrt{\overline{[(x_{mod} - x_{obs})^2]}}}{(\max(x_{obs}) - \min(x_{obs}))} \quad (15)$$

$$nMAE = \frac{|\overline{x_{mod} - x_{obs}}|}{\overline{x_{obs}}} \quad (16)$$

$$MR = \frac{\overline{x_{mod}}}{\overline{x_{obs}}} \quad (17)$$

where x_{mod} and x_{obs} represent the estimated and observed values. The overlines on parameters indicate the average values.

Additionally, the statistical parameters of Coefficient of Variation (CV), Inter-Quartile Range (IQR), and Mean Absolute Deviation (MAD) were used to evaluate the variability of data.

$$CV = \frac{STD(x)}{mean(x)} \quad (18)$$

$$IQR = Q3(x) - Q1(x) \quad (19)$$

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - mean(x)| \quad (20)$$

where STD is standard deviation; Q1 is the first quartile, the value below which 25% of the data falls, and Q3 is the third quartile, the value below which 75% of data falls.

Results

Assessment of existing clear sky models

Figure 4 shows the correlation between GHI_{CS} and GHI for all clear sky models (Table 1). Within the confidence level of 0.95% of the matchups of GHI_{CS} and GHI (Sect. 3.1.1), the values of nRMSE are between 4% and 17%, which indicates good performance of all models (Fig. 2f). However, the REST2 (Fig. 4c) and Bashahu models (Fig. 4b) underestimate the GHI_{CS} values slightly more than the other models ($MR \leq 0.9$). An $MR \approx 1$ and $uRMSE \leq 0.03$ are achieved by the Ineichen-Perez and MAC2 models. All models are aligned with $R^2=0.99$, of which the Ineichen-Perez model shows the highest consistency with measured GHI (slope=1.04 and intercept=-15.02) (Fig. 4d).

A total of 66,197 data points show the cloudless condition from meteorological observation ($CC=0$), but only 20,704 (31.3%) samples fall within the 95% confidence level of the Ineichen-Perez GHI_{CS} and GHI variations. The latter had $r_{CS} = 1 \pm 0.2$, and were colored in Fig. 4. The cloudless conditions outside this interval were indicated by grey points. 1,626 (2.4%) of the GHI data fall within the matchups of GHI_{CS} and GHI but are not characterized as observed cloudless ($CC>0$) so are excluded.

Assessment of clear sky detection methods

The CSD methods are evaluated against the occurrence of clear sky matchups (Fig. 4). Table 4 shows the confusion matrix classes and scores for the CSDs methods. The Ellis and Reno CSD methods exhibit the lowest values of FP and FN criteria. Meanwhile, the Reno CSD method shows the lowest FP and highest TP scores. Considering that 7.3% of the total periods are associated with clear sky matchups in this study (Fig. 4, colored points), the Reno ($ACC=0.92$) and Ellis ($ACC=0.88$) CSD methods display the best TP scores. The classification schema of Reno and Ellis show the highest score (PPV and NPV>0.85), while the Zhang method scores low. The classification metrics indicate that the Reno CSD method presents the highest metric values. The AliaMartinez method also shows a relatively good identification of clear skies (PPV and NPV>0.85, and $ACC=0.86$). Overall, the Reno CSD method is considered the optimum CSD method in this study according to the classification metrics and criteria of confusion matrix.

Thresholds and variability of IDR for different sky conditions

The Ineichen-Perez clear sky model is used to calculate the GHI_{CS} and GHI_{cda} . The GHI_{cda} is calculated using T_L with

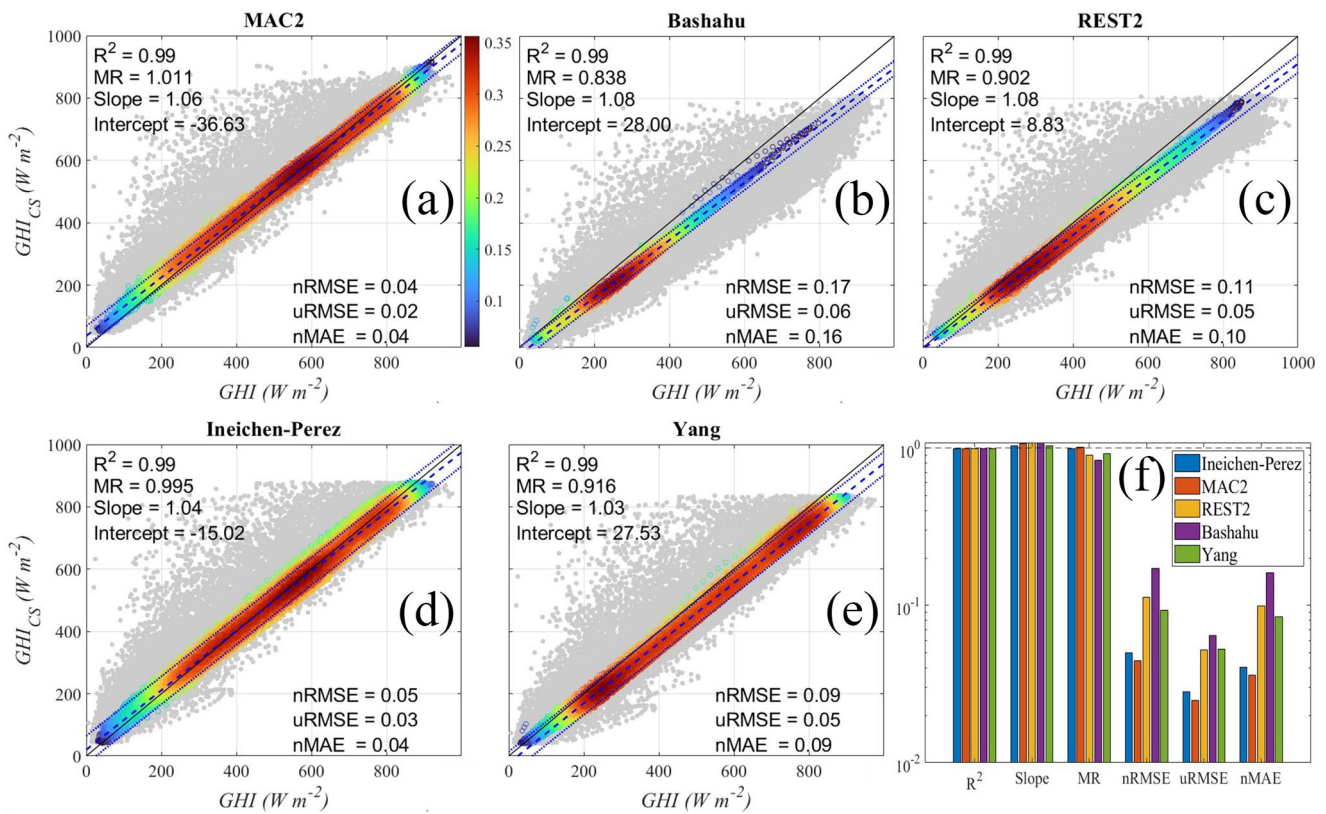


Fig. 4 Correlation between GHI and GHI_{CS} from different clear sky models (a–e). The grey points show samples with $CC=0$ that fall outside the matchups of GHI and GHI_{CS} (Sect. 3.1.1). The dashed and dotted lines show the best fit and the statistical bounds of the 95%

confidence level, respectively. The 1:1 correlation is shown using solid line. The colors represent the probability density of samples. **f**) Bar charts of the statistical parameters

Table 4 Confusion matrix classes (in percent) and scores for CSDs methods ($N=282,818$). The total period of cloudless matchups is observed for 20,704 timesteps (7.3%). Bold numbers indicate the best result in each column

CSD	Classes				Metrics						
	TP	TN	FP	FN	F1	F2	J	MCC	PPV	NPV	ACC
AliaMartinez	85.2	88.4	14.8	11.6	0.86	0.87	0.73	0.74	0.85	0.88	0.86
Ellis	88.6	87.4	11.4	12.6	0.88	0.87	0.76	0.77	0.89	0.87	0.88
Polo	76.1	87.6	23.9	12.4	0.81	0.83	0.64	0.65	0.76	0.87	0.81
Quesada	39.4	91.1	60.6	8.9	0.56	0.76	0.62	0.49	0.39	0.68	0.69
Reno	89.8	96.5	10.2	3.5	0.91	0.93	0.84	0.85	0.89	0.95	0.92
Zhang	33.1	89.5	66.9	10.5	0.46	0.60	0.33	0.27	0.33	0.79	0.61

$\beta=0$ and $PWV=0$ for a hypothetical cloudless-and-clean-and-dry atmosphere. The Reno CSD method is utilized to indicate the reference clear sky periods, and the reference overcast conditions are identified using CAELUS algorithm. The extracted clear sky periods with $ISRD \geq 0.01$ and overcast conditions with $ISRS \geq 1.0$ are excluded from the reference clear sky and overcast conditions, respectively (Sect. 3.3.3). A total of 3,702 periods of clear sky and 2,486 periods of overcast conditions are included in the reference sky conditions.

Figure 5 shows the probability distribution of the IDR and its classification in the different reference sky conditions based on the Ineichen-Perez clear sky with the

Reno CSD method and the CAELUS algorithm for overcast conditions. The resulting IDR ranges for clear sky and overcast conditions are $IDR \leq 0.37$ and $IDR \geq 0.68$, respectively. Slight overlaps in IDR values of 8% and 5% are observed for clear sky and overcast conditions with intermediate conditions, respectively (Fig. 5a). As a result, the ranges $IDR \leq 0.35 \pm 0.02$ and $IDR \geq 0.70 \pm 0.05$ fall within the 95% confidence level of clear sky and overcast conditions, respectively. Therefore, we refer to $IDR \leq 0.35$ as 'clear conditions', $0.35 < IDR < 0.70$ as 'intermediate conditions', and $IDR \geq 0.70$ as 'overcast conditions' hereafter.

Figure 5b shows the variability of hourly IDR values for the three sky conditions. The mean and STD are computed

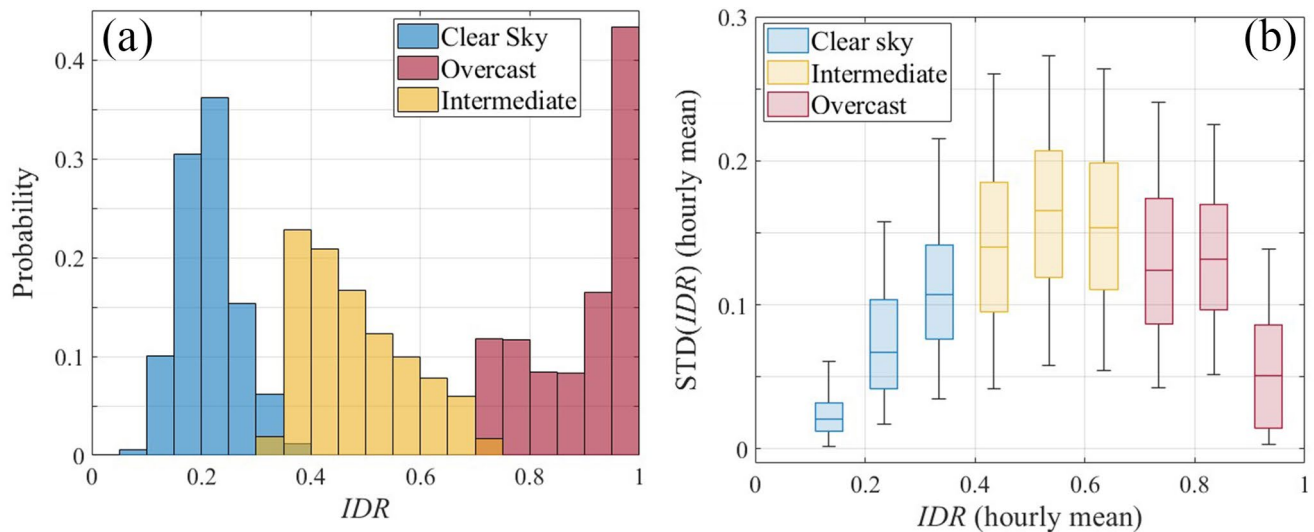


Fig. 5 (a) Distribution of the integrated diffuse ratio (IDR) for reference sky conditions. The IDR data are divided with a bin width of 0.05. (b) Standard deviations of hourly mean IDR in different sky condi-

tions. The upper and lower bounds of boxes correspond to the 25th and 75th percentile, and the center lines indicate the median values

Table 5 Confusion matrix classes (in percent) and scores for identifying of clear sky and overcast conditions using the provided model (Sect. 3.3) compared to reference Sky conditions during 2008–2010. Bold numbers indicate the best result in each column

Sky condition	Classes				Metrics						
	TP	TN	FP	FN	F1	F2	J	MCC	PPV	NPV	ACC
Clear	95.2	97.2	4.8	2.8	0.79	0.89	0.86	0.88	0.95	0.93	0.95
Overcast	94.6	98.8	5.4	1.2	0.75	0.84	0.77	0.78	0.92	0.95	0.92

using a moving time window of 60 min (six consecutive measurements). The results indicate the relationship between atmospheric conditions and the variability of IDR. The lowest variability of IDR is observed for clear sky conditions, followed by overcast, with the highest variability for intermediate conditions.

To assess the impact of excluding spectra due to QC processes, we analyzed a subset of excluded spectra (2008–2010, $N=5,237$) with high humidity ($>85\%$) but zero cloud cover ($CC=0$, based on meteorological observations from De Kooy station). Approximately 12% of these spectra showed IDR values slightly above the clear-sky threshold ($0.35\text{--}0.42$, compared to $IDR \leq 0.35$), suggesting a minor bias toward higher IDR values in humid conditions. However, as shown in Figs. 1 and 2, the IDR remains robust against distorted spectra, as the application of ISRS (<1.0) and ISRD (<0.01) indices effectively excludes periods with high atmospheric scattering or cloud absorption, ensuring that IDR thresholds for clear-sky ($IDR \leq 0.35$) and overcast ($IDR \geq 0.70$) conditions are not significantly affected by variable atmospheric conditions. A Kolmogorov-Smirnov test confirmed that the exclusion of high-humidity spectra did not significantly alter the overall IDR distribution for clear-sky periods ($p > 0.05$, Fig. 5a).

Model performance assessment

The performance of the presented model is evaluated using the observed datasets from 2008 to 2010. The methods outlined in Sect. 4.3 (the Ineichen-Perez clear sky model combined with the Reno CSD method for clear sky and the CAELUS algorithm for overcast conditions) are used to determine the reference sky conditions during 2008–2010. The sky cloudiness is also determined using the IDR threshold values. The reference sky conditions are then used to evaluate the performance of the provided model using the confusion matrix approach (Sect. 3.4). A total of 1,720 and 1,658 reference periods with clear sky and overcast conditions are identified, respectively. Table 5 shows the confusion matrix classes and scores of clear skies and overcast conditions compared with the reference sky conditions during 2008–2010. The high values of PPV and NPV (>0.9) indicates a very good classification schema of clear sky and overcast against intermediate conditions. The low percentage of FP and FN values indicate that the thresholds of $IDR \leq 0.35$ for clear sky and $IDR \geq 0.70$ for overcast conditions provide good model skill for classifying different sky conditions using the evaluation dataset. Maximum values of $IDR=0.44$ for clear skies and minimum values of $IDR=0.64$ for overcast conditions are observed, comprising the 4.8% and 5.4% (FP values) of the total number of observations

outside the prescribed *IDR* classification ranges. However, all metrics show high scores indicating that the presented limits of *IDR* values are good setting to classify the various sky conditions. Overall, the total accuracies for clear sky and overcast identification are 95% and 92%, respectively.

Comparison of IDR classification with previous studies

To our knowledge, the most dominant application of hyperspectral measurements for sky classification is restricted to single-wavelength spectral ratios. Furthermore, Jordan et al. (2022) used the direct measurements of $E_{ss}(\lambda)$ to calculate the *IDR* in Western Channel, which is the only study

comparable to this work. Here, we first compare the proposed method with two classification methods based on spectral ratio indices. Groetsch et al. (2017) and Ruddick et al. (2006) utilized the ratio of $L_S(750)/E_S(750)$ to identify clear skies ($L_S(750)/E_S(750) < 0.1$) and overcast conditions ($L_S(750)/E_S(750) \geq 0.3$). Simis and Olsson (2013) used that the ratio of $\pi L_S(400)/E_S(400)$, showing that $\pi L_S(400)/E_S(400) < 0.25$ corresponds with clear skies, and $\pi L_S(400)/E_S(400) > 0.8$ with predominantly clouded skies. We apply these indices to hyperspectral observations from 2011 to 2022.

Figure 6 shows scatterplots of $E_S(750)$ vs. $L_S(750)$ and $E_S(400)$ vs. $\pi L_S(400)$ for clear sky and overcast periods. About 88.6% of clear sky spectra have $L_S(750)/E_S(750) < 0.1$, and 13.8% have $\pi L_S(400)/E_S(400) < 0.25$. Moreover, about

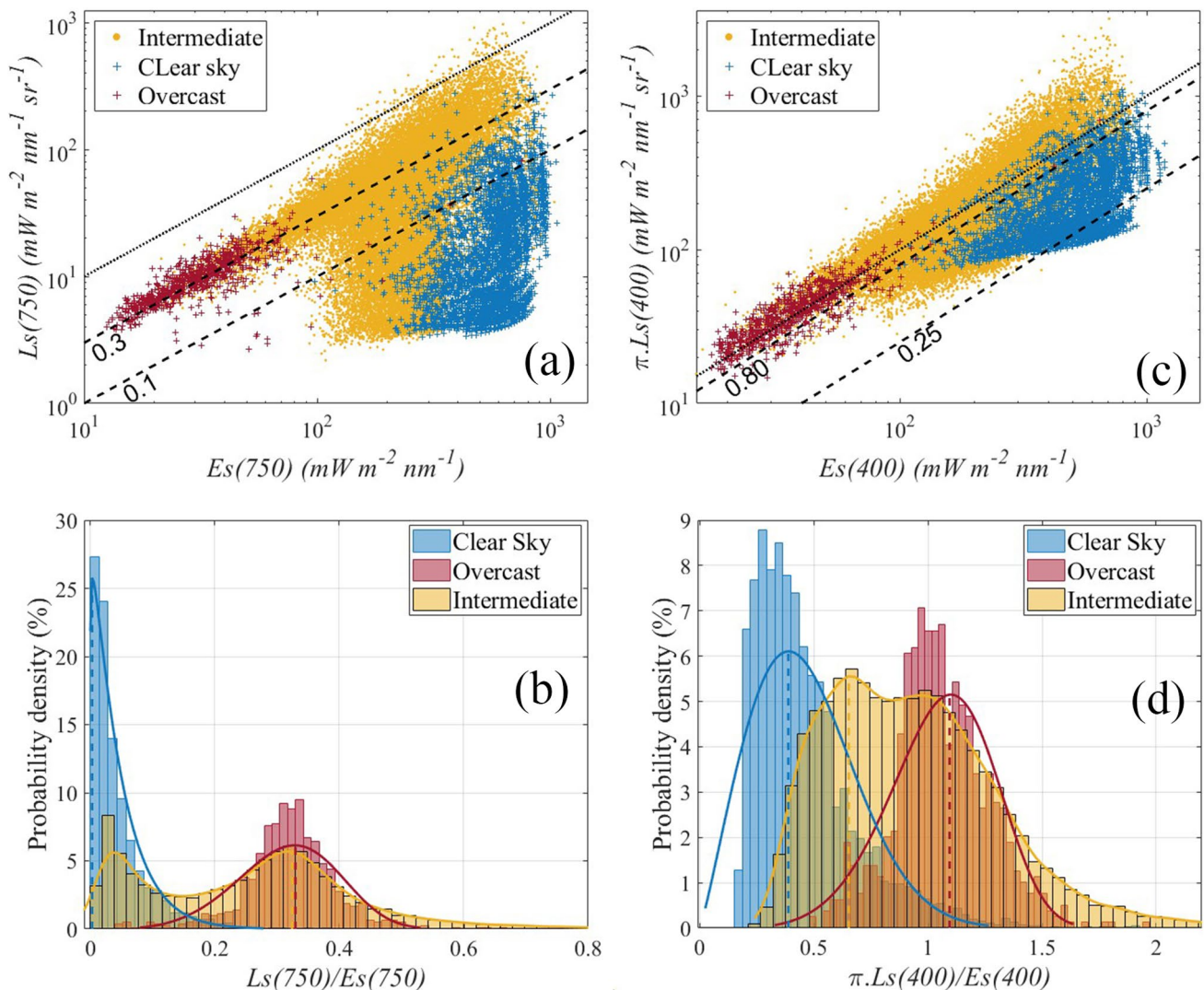


Fig. 6 (a) Scatterplot of $E_S(750)$ vs. $L_S(750)$. The colored points show the reference clear sky and overcast conditions. The dashed lines show the borders of clear sky (lower) and overcast (upper) conditions depicted by Groetsch et al. (2017) and Ruddick et al. (2006). The dotted lines show the 1:1 ratio. (b) Probability distributions of $L_S(750)/E_S(750)$ for different sky conditions as depicted in (a). The

thick lines show the kernel distribution of samples. The vertical dashed lines show the average values of each histogram. (c) Similar to (a), but for $E_S(400)$ vs. $\pi L_S(400)$. The borders of clear sky (lower) and overcast (upper) conditions (dashed lines) are depicted by Simis and Olsson (2013). (d) Similar to (b), but for $\pi L_S(400)/E_S(400)$ as depicted in (c)

65.3% of overcast spectra have $L_S(750)/E_S(750) \geq 0.3$, and 90.4% have $\pi L_S(400)/E_S(400) > 0.8$. Although the results may be relatively satisfactory (except for the second index of clear skies), about 24.7% and 41.2% of intermediate sky conditions fall in the clear sky and overcast boundaries of Fig. 6, respectively. Although the mean values of clear sky and overcast spectra show significant discrimination, there is a noticeable overlap between clear sky and Intermediate spectra (9% in Figs. 6b and 56% in Fig. 6d), as well as between overcast and Intermediate spectra (23% in Figs. 6b and 26% in Fig. 6d). Table 6 shows the confusion matrix classes and scores for clear skies and overcast conditions indicated by both indices, compared with the reference sky conditions. For clear sky classes, the ratio of $L_S(750)/E_S(750)$ represents a relatively good classification ($ACC=0.77$), but the result of $\pi L_S(400)/E_S(400)$ is not effective ($ACC=0.47$). For overcast conditions, the ratio of $\pi L_S(400)/E_S(400)$ shows a better result ($ACC=0.75$) than the ratio of $L_S(750)/E_S(750)$ ($ACC=0.64$). However, the total misclassification of clear sky and overcast conditions for both indices ($FP \sim 10\text{--}85\%$ and $FN \sim 35\%$) are significant, leading to overall weak scores (Table 6).

Figure 7 shows the dependence of $CV[L_S(\lambda)/E_S(\lambda)]$ at $\lambda=400$ nm and $\lambda=750$ nm on the variability of hourly mean IDR . The variability of IDR against $CV(IDR)$ indicates that the clear sky and overcast conditions have the lowest variability on an hourly scale (Fig. 7a). Conversely, intermediate conditions show high variability of IDR with median $CV(IDR) \approx 45\%$ and upper quartiles $\approx 100\%$ higher than clear sky and overcast conditions. The clear sky with $IDR < 0.2$ and overcast with $IDR > 0.9$ show the lowest variability of IDR with the median of $CV(IDR) \leq 0.1$. The $CV[L_S(400)/E_S(400)]$ show lower variability of intermediate conditions with median $CV(IDR) \approx 23\%$ and upper quartiles $\approx 32\%$ higher than clear sky and overcast conditions (Fig. 7b). In $\lambda=750$ nm, clear sky and overcast conditions have relatively similar $CV[L_S(750)/E_S(750)]$, while the intermediate conditions show a great variability of $CV[L_S(750)/E_S(750)]$ from 0.1 to 0.95 with median of 0.42 (Fig. 7c). In $\lambda=400$ nm and $\lambda=750$ nm a relatively similar variability of clear sky and overcast conditions are observed on an hourly scale, while the intermediate conditions show different variability. As a result, the $CV[L_S(400)/E_S(400)]$ and $CV[L_S(700)/E_S(700)]$ do not show clear dependence to atmospheric state (variability of IDR) with no significant correlation ($R^2 < 0.35$) between $CV(IDR)$ and $CV[L_S(\lambda)/E_S(\lambda)]$ on an hourly under different sky conditions.

The only available direct measurements of direct-diffuse solar irradiance have been outlined by Wood et al. (2017) in a marine environment. Jordan et al. (2022) utilized the measured direct-diffuse solar irradiance to calculate the IDR (Eq. 5) for the Western Channel. They used $IDR < 0.2$ for

Table 6 Confusion matrix classes (in percent) and scores for reference clear Sky and overcast conditions compared with the two existing indices

Index	Sky cond.	Classes				Metrics						
		TP	TN	FP	FN	F1	F2	J	MCC	PPV	NPV	ACC
$L_S(750)/E_S(750)$	Clear sky	88.6	66.5	11.4	33.5	0.79	0.75	0.58	0.56	0.88	0.66	0.77
	Overcast	65.3	63.9	34.7	36.1	0.65	0.64	0.29	0.29	0.65	0.64	0.64
$\pi L_S(400)/E_S(400)$	Clear sky	13.8	66.5	86.2	33.5	0.19	0.24	0.18	0.14	0.14	0.73	0.47
	Overcast	90.4	64.1	9.6	35.9	0.79	0.74	0.55	0.51	0.90	0.57	0.75

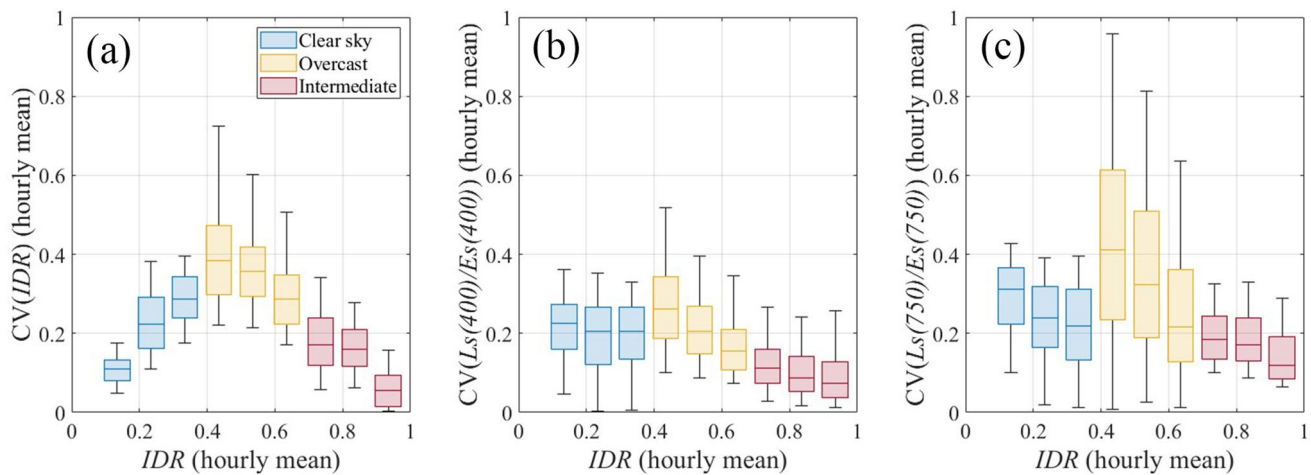


Fig. 7 (a) Relationship between mean IDR and the $CV(IDR)$ on an hourly scale in various sky conditions. (b) Atmospheric dependence of $CV[L_S(400)/E_S(400)]$ on an hourly scale. (c) similar to (a), but for $CV[$

$L_S(750)/E_S(750)]$. The upper and lower bounds of boxes correspond to the 25th and 75th percentile, and the center lines indicate the median values

Table 7 Comparison between existing and the proposed methods for identification of Sky conditions using hyperspectral measurements

Method	Basis	Advantages	Limitations
$L_S(750)/E_S(750)$	Single-band NIR ratio	Simple, sensitive to scattered clouds	Fails under scattered clouds (TP=65.3%, Table 6)
$\pi L_S(400)/E_S(400)$	Blue-band ratio	Simple, sensitive to aerosol scattering	Overlap in intermediate conditions (FP=86.2%, Table 6)
IDR (Jordan et al. 2022)	Relationship between STD and mean of IDR	High performance in clear-sky and overcasts	Fails in a high-level atmospheric scattering, leading to $IDR > 0.8$
This study	Calculate IDR , $ISRD$, and $ISRS$ using RT models	Adds $ISRD$ and $ISRS$ to exclude aerosol-saturated cases	The IDR thresholds for sky conditions may be site specific and not tested elsewhere

clear-sky and $IDR > 0.8$ for overcast conditions (their Fig. 9). In this study, $ISRD \geq 0.01$ and $ISRS \geq 1.0$ are excluded from IDR for clear-sky and overcast conditions (Sect. 4.3). Similar to the Jordan et al. (2022), if $ISRD$ and $ISRS$ are omitted from the IDR calculation, the values of $IDR \leq 0.22$ and $IDR \geq 0.81$ are obtained for clear-sky and overcast conditions, respectively. In this case, the ACC values for identifying of clear sky and overcast conditions compared to reference sky conditions during 2008–2010 will be 0.84 and 0.81, respectively. This finding indicates that the IDR values obtained in this study fall within the range reported by Jordan et al. (2022) when the effects of the high-level atmospheric

scattering (i.e., $ISRD$ and $ISRS$) are ignored; however, the accuracy of the IDR classification decreases significantly. The high variability of IDR in intermediate conditions (median $CV(IDR) \approx 45\%$, Fig. 7a) reflects the dynamic nature of scattered clouds, which cause rapid fluctuations in the diffuse-to-direct ratio. This variability is effectively managed by applying $ISRS$ and $ISRD$ thresholds, which exclude periods with saturated diffuse components or significant cloud absorption, ensuring accurate classification of clear-sky ($IDR \leq 0.35$) and overcast ($IDR \geq 0.70$) conditions (Fig. 5). Table 7 summarizes the comparative framework between the existing and proposed methods.

Variability of hyperspectral data in different sky conditions

The spectra of reference clear sky, overcast, and intermediate periods are selected to illustrate the shape, intensity, and variability of $E_S(\lambda)$ and $L_S(\lambda)$, and the ratios of $L_S(\lambda)/E_S(\lambda)$ and $E_{SS}(\lambda)/E_S(\lambda)$ (Fig. 8). The variability of spectra and selected ratios are shown in Table 8. The following results are obtained for different sky conditions:

- The magnitude of $E_S(\lambda)$ spectra are similar under all sky conditions, with a maximum absolute difference of 26% between clear sky and overcast conditions in the blue wavelengths.
- Clear skies are characterized by low variability of $E_S(\lambda)$, $L_S(\lambda)$, and $E_{SS}(\lambda)/E_S(\lambda)$ ratio ($CV < 0.2$, Range < 0.2 , IQR ≤ 0.05) with relatively high variability of $L_S(\lambda)/E_S(\lambda)$ ($CV > 0.25$, Range > 0.25 , IQR > 0.1). The boundary limits of $L_S(750)/E_S(750) < 0.1$ and $L_S(400)/E_S(400) < 0.2$, similar to existing indices (Sect. 4.5), are obtained. Both ratios present a significant overlap ($> 20\%$) with

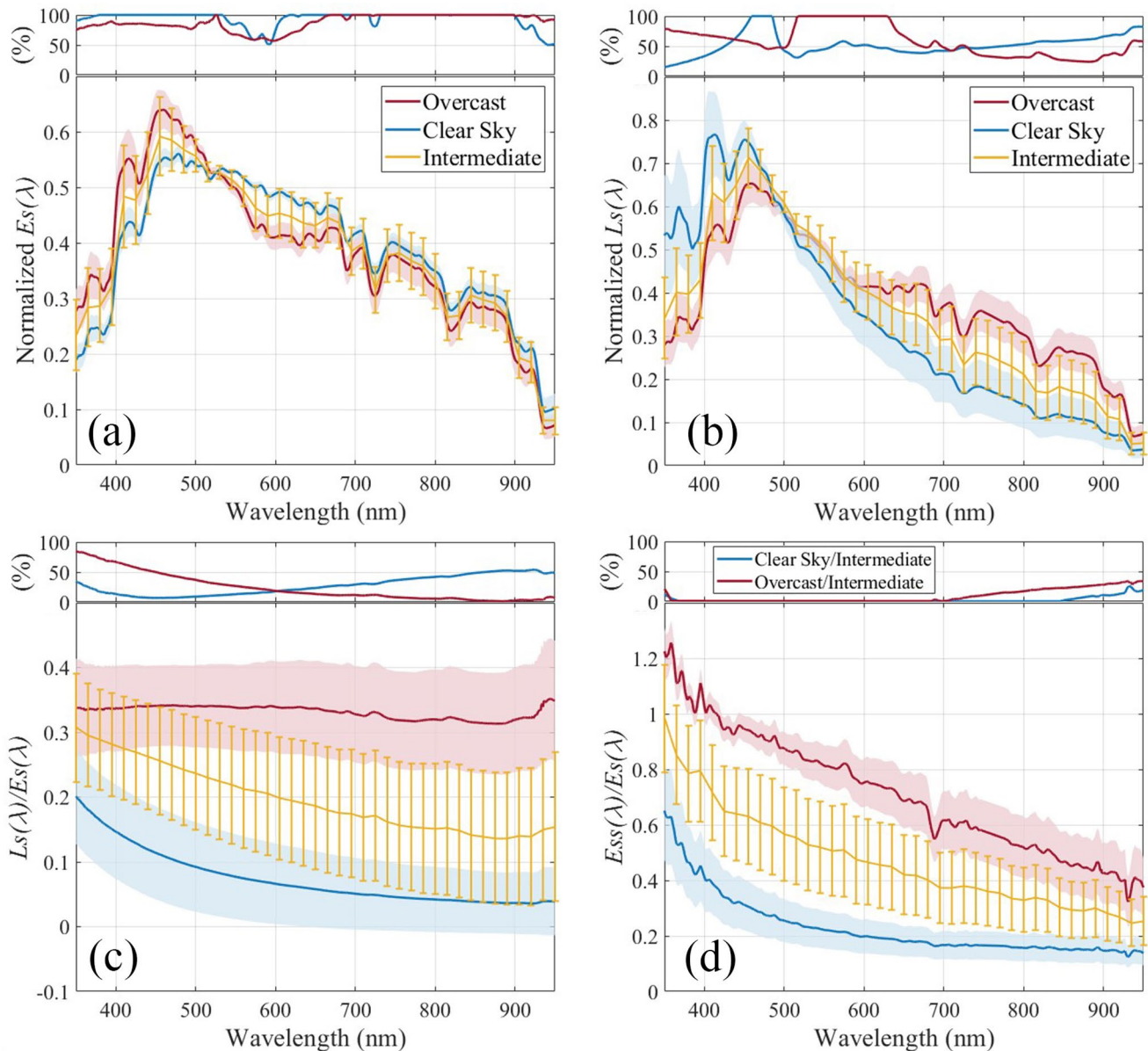


Fig. 8 (a) Spectral variability of normalized $E_s(\lambda)$ in different sky conditions. Thick lines show the average and shaded areas show ± 1 STD. The line chart above the main chart shows the percentage of False

intermediate conditions spectra (Table 8; Fig. 8c), leading to misclassification of clear sky from intermediate spectra (Table 6). Conversely, the overlaps between clear skies and intermediate conditions is less than 1% for the ratio of $E_{ss}(\lambda)/E_s(\lambda)$ in $\lambda \leq 840$ nm (Table 8; Fig. 8d).

- Overcast spectra show low variability of $E_s(\lambda)$, $L_s(\lambda)$, and $E_{ss}(\lambda)/E_s(\lambda)$ ratio, and relatively high variability of $L_s(\lambda)/E_s(\lambda)$, similar to clear skies (Table 8). The values of $L_s(750)/E_s(750) \geq 0.25$ and $L_s(400)/E_s(400) \geq 0.27$ indicates the boundary limits of overcast spectra. These boundary limits overlap significantly with intermediate spectra, with false positive rates exceeding 60% in blue

Positive of Intermediate conditions spectra that overlap with clear sky/overcast conditions. (b) similar to (a), but for normalized $L_s(\lambda)$. (c) similar to (a), but for $L_s(\lambda)/E_s(\lambda)$. (d) similar to (a), but for $E_{ss}(\lambda)/E_s(\lambda)$

wavelengths and about 10% in NIR wavelengths. Analogous to the clear skies, the ratio of $E_{ss}(\lambda)/E_s(\lambda)$ is wavelengths dependent and show relatively low variability. The overlaps of $E_{ss}(\lambda)/E_s(\lambda)$ ratio between overcast and intermediate spectra is less than 1% in $\lambda \leq 720$ nm and reaches up to 32% in NIR (Table 8; Fig. 8d).

The Principal Component Analysis (PCA) is used to characterize the spectral variability of $E_s(\lambda)$, $L_s(\lambda)$, $L_s(\lambda)/E_s(\lambda)$, and $E_{ss}(\lambda)/E_s(\lambda)$ across clear-sky, overcast, and intermediate conditions, as classified by the IDR method (Fig. 9). The PCA eigenvectors of the most prominent spectral absorption

Metrics	Clear skies				Overcast				Intermediate			
	CV	Range	MAD	IQR	CV	Range	MAD	IQR	CV	Range	MAD	IQR
$L_3/E_3(\lambda)$	0.05	0.17	0.01	0.02	0.09	0.20	0.01	0.03	0.14	0.22	0.03	0.05
$L_3(\lambda)$	0.16	0.18	0.04	0.05	0.11	0.19	0.02	0.04	0.24	0.29	0.04	0.09
$L_3(\lambda)/E_3(\lambda)$	0.26	0.27	0.12	0.11	0.22	0.28	0.14	0.18	0.52	0.49	0.19	0.22
$E_{38}(\lambda)/E_3(\lambda)$	0.18	0.19	0.09	0.08	0.15	0.21	0.11	0.10	0.32	0.42	0.16	0.14
	mean±STD	BL	FP _{in}	BL	mean±STD	BL	FP _{in}	BL	mean±STD	BL	mean±STD	BL
$L_3(L_3(750)/E_3(750))$ (sr ⁻¹)	0.045±0.05	<0.10	36.4	≥0.25	0.32±0.07	≥0.25	18.2	≥0.47 & ≥0.27	0.16±0.10	≥0.47 & ≥0.27	0.16±0.10	≥0.47 & ≥0.27
$L_3(L_3(400)/E_3(400))$ (sr ⁻¹)	0.15±0.07	<0.20	23.6	≥0.27	0.34±0.07	≥0.27	66.8	≥0.20 & ≥0.35	0.28±0.08	≥0.20 & ≥0.35	0.28±0.08	≥0.20 & ≥0.35
$E_{38}(400)/E_3(400)$	0.41±0.11	<0.52	<1	≥0.92	0.97±0.05	≥0.92	<1	≥0.60 & ≥0.88	0.74±0.28	≥0.60 & ≥0.88	0.74±0.28	≥0.60 & ≥0.88
$E_{38}(750)/E_3(750)$	0.17±0.05	≤0.20	<1	>0.47	0.58±0.11	>0.47	10.4	≥0.23 & ≥0.48	0.36±0.21	≥0.23 & ≥0.48	0.36±0.21	≥0.23 & ≥0.48

The total variances of $L_S(\lambda)/E_S(\lambda)$ spectra are dominated along the first two for 98.5%, 96.3%, and 97.8% in clear sky, overcast, and intermediate conditions, respectively. Figure 9c shows that the total variances of $L_S(\lambda)/E_S(\lambda)$ eigenvectors along the PC1 and PC2 are similar under different sky conditions. The biplot of PC1 vs. PC2 of $L_S(\lambda)/E_S(\lambda)$ eigenvectors indicates that the variability of spectra are mostly from $L_S(\lambda)$ and $E_S(\lambda)$ individually at $\lambda=410$ nm. The eigenvector of $\lambda=410$ nm in intermediate conditions has similar coefficients and influence (equal length) on the variability of the spectra in overcast conditions. The eigenvector of $\lambda=865$ nm has a negative coefficient along PC1 and a positive coefficient along PC2 in intermediate condition, which contributes significantly (equal lengths) to the variability of

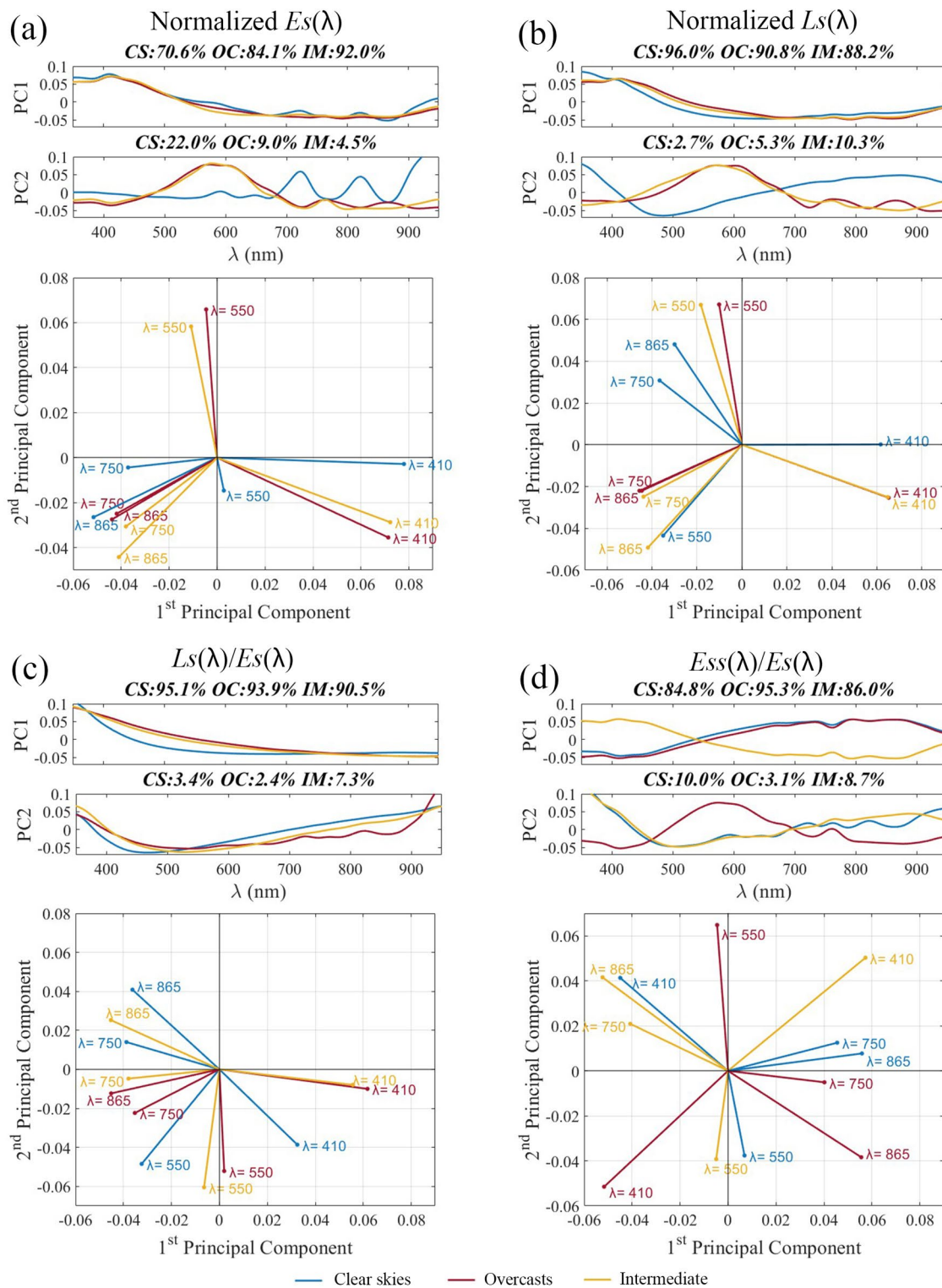


Fig. 9 Principal Component Analysis (PCA) of spectra in different sky conditions. **(a)** PCA of normalized $E_s(\lambda)$ spectra. The two top line charts show the amplitudes of PC1 and PC2. CS=Clear Sky, OC=Overcasts, and IM=Intermediate conditions. The numbers show the percentage of PC1 and PC2 in different sky conditions. The lower biplot shows

the PCA eigenvectors along the PC1 and PC2 components in different sky conditions for selected wavelengths. **(b)** similar to (a), but for normalized $L_s(\lambda)$ spectra. **(c)** similar to (a), but for $L_s(\lambda)/E_s(\lambda)$ spectra. **(d)** similar to (a), but for $E_{ss}(\lambda)/E_s(\lambda)$ spectra

$\lambda=865$ eigenvector in clear skies. Similarly, the eigenvector of $\lambda=750$ nm has a negative coefficient along PC1 and PC2 in intermediate condition, which controls the variability of the eigenvector in clear skies. These findings are consistent with Fig. 9c, indicating the influence of intermediate spectra on the variability of overcast in the blue region and clear skies in the NIR region (Fig. 9c). The total variances of $E_{ss}(\lambda)/E_s(\lambda)$ spectra include 97.9%, 99.1%, and 98.1% along the first three PCA components under clear, overcast, and intermediate conditions, respectively. Figure 9d shows that the variabilities of $E_{ss}(\lambda)/E_s(\lambda)$ spectra along PC1 under clear sky and overcast conditions are opposite to those under intermediate conditions. The eigenvectors of $\lambda=410$ nm have different coefficients and relatively equal lengths along PC1 and PC2 under different sky conditions (Fig. 9d). This indicates that the variabilities of $E_{ss}(\lambda)/E_s(\lambda)$ spectra in clear skies and overcasts are not influenced by the variability of this ratio in intermediate conditions. Similar findings are obtained from eigenvectors of $\lambda=750$ nm and $\lambda=865$ nm. A relatively neutral coefficients of $\lambda=550$ nm are observed in the variabilities of $E_{ss}(\lambda)/E_s(\lambda)$ spectra in different sky conditions. The findings indicate that 94.8% (amplitude of PC1+PC2) of $E_{ss}(\lambda)/E_s(\lambda)$ spectral variabilities under clear skies and 98.4% under overcast conditions in the blue and NIR regions are not influenced by intermediate spectra.

It should be noted that the PCA is applied to datasets classified by the IDR method, and thus serves to characterize — rather than independently validate — the spectral structure of the identified classes. The PCA results confirm that $E_{ss}(\lambda)/E_s(\lambda)$ exhibits consistent patterns within the clear and overcast groups, with reduced overlap from the intermediate group in selected spectral regions. This supports the internal coherence of the classification scheme, but does not establish external validation.

Discussion

$E_s(\lambda)$ and $L_s(\lambda)$ spectral shapes contain several narrow spectral features originating from gas absorption in Earth's atmosphere. The measured radiometric data illustrate the consistent presence of atmospheric gas absorption features at specific wavelength regions, including: the red-edge region in $\lambda=680\text{--}750$ nm due to the absorption edge of oxygen (Curran 1989; Kokaly et al. 2009), absorption edge of water vapor in the NIR region at $\lambda=940$ nm (Wild 2005; Vaquero-Martínez and Antón 2021), aerosols absorption in the visible region (400–700 nm) (Rozanov and Kokhanovsky 2005; Sayer et al. 2014), and atmospheric scattering in the NIR wavelengths (in addition to 940 nm) due to coarse particles such as dust, smoke, and other particulate matter in 710–740 nm and 810–840 nm (Chylek and Wong

1995; Smirnov et al. 2002). In general, these absorption features are distributed across the spectra of $E_s(\lambda)$ and $L_s(\lambda)$ under various sky conditions. Several atmospheric factors influence the spectral position and width of absorption features in $E_s(\lambda)$ and $L_s(\lambda)$ spectra. The most important factors include solar activity, increased ozone absorption, increased water vapor absorption, intense atmospheric reddish hue of dusk or dawn radiation, precipitation, high-humidity, cloud cover, and aerosol scattering, (Kopp and Lean 2011; Marchenko et al. 2016). Except for the last two items, the other factors are the subject of quality-control routines (Sect. 2.1) and the corresponding spectra are excluded in this study. Therefore, absorption features of $E_s(\lambda)$ and $L_s(\lambda)$ spectra are investigated here which are solely influenced by clouds and aerosols scattering. Our analysis reveals the consistent width and position of the most prominent absorption features of $E_s(\lambda)$ and $L_s(\lambda)$ spectra (Sect. 4.6, Fig. 9). Overall, the different absorption features show relatively equal prominence in the select quality-controlled samples with a spectral width of ± 10 nm. Hence, the range of 760 ± 10 nm is omitted entirely in our analysis to avoid the effect of the oxygen absorption band (Sect. 3.3). Meanwhile, the most prominent features in the blue, green, and NIR regions are selected to investigate the spectral variabilities (Sect. 4.6, Fig. 9). It should be noted that the absorption features of the selected spectra are distributed across the $E_s(\lambda)$ and $L_s(\lambda)$ spectra without significant spectral band-shifting or deformation (e.g., red-shifting and high NIR intensities) under different sky conditions (Fig. 8a–b). Conversely, the spectral shape of $E_s(\lambda)$ may be deformed and shifted towards longer wavelengths due to scattering in NIR region by coarse particles (Chylek and Wong 1995; Smirnov et al. 2002), fine aerosol absorption in the visible region (Rozanov and Kokhanovsky 2005; Sayer et al. 2014), absorption of water vapor in NIR region during precipitation or very high humid conditions (Vaquero-Martínez and Antón 2021; Wild 2005), and absorption of atmospheric gases in the blue region (Curran 1989; Kokaly et al. 2009). In such cases, the variabilities of $E_s(\lambda)$ and $L_s(\lambda)$ spectra, and the ratios of $L_s(\lambda)/E_s(\lambda)$ and $E_{ss}(\lambda)/E_s(\lambda)$ under different sky conditions may differ from those obtained in this study, and the presented methodology may fail to indicate sky conditions. Therefore, the results of this study are valid for spectra that do not show band-shifting or deformation in overcast and intermediate conditions compared to clear skies.

The spectral variabilities of the $L_s(\lambda)/E_s(\lambda)$ ratios reveal that these spectral ratios can indicate the sky conditions, particularly in the NIR region. The $L_s(750)/E_s(750)$ ratio achieves an accuracy of approximately 75% for identifying clear sky and overcast conditions when rapid changes of thick and scattered clouds are ignored. The presence of thick and scattered clouds introduces substantial uncertainty

in these measurements, leading to inaccurate identification of sky conditions (Table 6). Likewise, the spectral ratios in the blue region also exhibit significant variability in different sky conditions (Figs. 6 and 8). The findings show that although the spectral ratio in the blue region provides valuable information, the $L_S(400)/E_S(400)$ ratio does not offer a robust signal for distinguishing clear and cloudy conditions (Table 6). This suggests that in scattered cloudy skies, whether obscuring the sun being or patchy within the sensor's FOV, none of these indices are appropriate to detect clear sky conditions. Moreover, these indices do not show significant dependence on atmospheric state over short periods (e.g., hourly scale) under different sky conditions. This finding indicates that the $L_S(\lambda)/E_S(\lambda)$ ratios are not sensitive to rapidly changing cloud cover (Fig. 9). In contrast, the $E_{SS}(\lambda)/E_S(\lambda)$ variations show that thick and scattered clouds do not influence ~95% of the spectral variabilities in different sky conditions. In addition, the results indicate the dependence of *IDR* variability in a short period (hourly) on sky cloudiness (Fig. 7a). These findings indicate that the diffuse component of downwelling solar irradiance is less affected by the variability induced by clouds, and it is also sensitive to changing cloud cover. While the PCA analysis in Sect. 4.6 initially relies on *IDR*-classified data, which could introduce circularity, the PCA using independently classified data (Reno CSD and CAELUS) confirms the stability of the $E_{SS}(\lambda)/E_S(\lambda)$ ratio across sky conditions. This independent analysis mitigates concerns about circularity and reinforces the reliability of $E_{SS}(\lambda)/E_S(\lambda)$ as a robust metric for sky condition classification, as its spectral variability remains consistent regardless of the classification method used.

This study focuses on classifying sky conditions based on radiometric effects observed in hyperspectral irradiance data, without explicit retrieval of cloud optical properties. While this approach captures the radiative signature of cloud cover, it does not distinguish cloud types based on physical or microphysical characteristics. We estimate the diffuse component of downwelling solar irradiance using a radiative transfer inverse modeling (Sect. 3.3.1), since no direct measurements of diffuse component are available. Although the Gregg and Carder (1990) model was developed for clear-sky simulations, in this study it is used in an inversion context as a spectral decomposition framework to estimate the relative contributions of direct and diffuse irradiance from measured $E_S(\lambda)$ spectra. The retrieved parameters (i.e., f_{dd} , f_{ds}) serve as effective scaling terms, constrained by known physical bounds, that enable classification of sky conditions via *IDR*. Similar inversion-based applications have been demonstrated by Jordan et al. (2022), where this approach provided consistent and accurate diffuse-to-direct ratios even under intermediate or overcast skies. While the

absolute physical meaning of these parameters is uncertain outside clear-sky regimes, their role as proxies for irradiance partitioning is statistically stable and adequate for the classification goals of this study. A new generation of the hyperspectral pyranometer (HSP) provides field measurements of $E_{SS}(\lambda)$ and $E_S(\lambda)$. The HSP monitors the distribution of downwelling solar irradiance in global, diffuse, and direct components. Field measurements of diffuse irradiance enable us to adapt the RT model to the site-specific atmospheric conditions and investigate the variability of atmospheric state against measured *IDR*. The measure *IDR* then enables a comparative investigation of atmospheric models as a function of atmospheric state for estimation of aerosol optical depth (τ), angstrom exponent (α), aerosol optical thickness (β), and *PWV* (Sect. 3.2.1). It is recommended to evaluate the presented method and the results of this study using HSP measurements. Wood et al. 2017 describe the characteristics of HSP and its application for diffuse irradiance measurements.

The *IDR* thresholds ($IDR \leq 0.35$ for clear sky, $IDR \geq 0.70$ for overcast) were derived for our study region, accounting for local atmospheric conditions (e.g., aerosol load, water vapor, and typical cloud types). Jordan et al. (2022) utilized the HSP measured diffuse component of solar irradiance to calculate the *IDR* for the Western Channel. They referred to $IDR < 0.2$ as clear-sky, $0.2 < IDR < 0.8$ as intermediate conditions, and $IDR > 0.8$ as overcasts. Here, we considered the effect of saturation of diffuse irradiance component, which increases with a high level of atmospheric scattering, an issue that has not been addressed in previous studies (Sect. 3.3.2, Fig. 3). Thus, the observed differences between the *IDR* thresholds in this study and those reported in the Jordan et al. (2022) study may largely be attributed to the inclusion of this effect. The effect of high atmospheric scattering (aerosol-saturated) increases the values of $IDR > 0.8$ which tends to misclassification of scattered clouds (Sect. 4.5). Therefore, we believe that the *ISRD* and *ISRS* coefficients should essentially be excluded from the *IDR* calculation to minimize the effect of high-level scattering on *IDR* variability.

Additionally, the strict QC thresholds, particularly for high humidity ($\geq 85\%$), ensure high-quality hyperspectral data but may exclude certain clear-sky scenarios, such as humid but cloudless conditions common in coastal regions. Under high humidity, hyperspectral spectra can exhibit enhanced water vapor absorption in the NIR (e.g., 940 nm), which distorts $E_S(\lambda)$ and $L_S(\lambda)$ and renders $E_{SS}(\lambda)$ and $E_{Sd}(\lambda)$ components invalid (Moradi et al. 2024). To assess the impact of excluding such spectra, our analysis revealed that about 12% of *IDR* values tended to higher values of 0.35–0.50, suggesting a minor bias toward higher *IDR* values in humid conditions. To further improve inclusivity, future

studies could incorporate adaptive humidity thresholds or use ancillary data, such as precipitable water vapor (PWV), to better account for humid clear-sky conditions without compromising data quality. However, we acknowledge that the applicability of *IDR* for classifying atmospheric states in other regions depends on three key factors, including: **(i) Atmospheric Composition:** Regions with significantly higher aerosol loads (e.g., urban/desert areas) or humidity (e.g., tropics) may shift the *IDR* distribution due to enhanced scattering/absorption. **(ii) Sensor Geometry:** The thresholds assume a fixed viewing geometry ($\theta_v=35^\circ$, $\phi=135^\circ$). Changes in sensor angle or sun-glint contamination could alter the diffuse/direct balance. **(iii) Cloud optical properties** (e.g., thin cirrus vs. thick cumulus) affects the *IDR* distribution. Regions with predominant cloud types differing from our study (e.g., ice clouds) may require recalibration. As a result, we recommend the following for general applicability: **(a) Local Validation:** we suggest validating the *IDR* thresholds against local clear-sky models (Tables 1 and 2) using *GHI* measurements, which are routinely measured at weather stations. **(b) Dynamic Thresholds:** for automated systems, region-specific *IDR* distributions could be derived using short-term calibration data. **(c) Ancillary Data:** combining *IDR* with aerosol optical depth (*AOD*) or precipitable water vapor (*PWV*) measurements may improve adaptability.

The high variability of *IDR* in intermediate (scattered-cloud) conditions, as shown in Fig. 7a (median $CV(IDR) \approx 45\%$), reflects its sensitivity to rapid changes in cloud cover, particularly for scattered clouds. This variability, however, does not undermine the method's ability to classify clear-sky and overcast conditions, as the application of *ISRS* (<1.0) and *ISRD* (<0.01) thresholds effectively excludes periods with high atmospheric scattering or cloud absorption that could lead to misclassification (Fig. 3). This approach ensures that *IDR* remains a robust metric for distinguishing stable atmospheric states (clear and overcast) from transitional ones, with the observed variability in intermediate conditions indicating dynamic cloud effects rather than a limitation of the method.

While our method achieves strong agreement ($>90\%$) with reference classifications, these references themselves derive from modeled *GHI* thresholds (Reno CSD) and CAE-LUS outputs. Absolute accuracy relative to ground-truth sky imagery or lidar data remains to be validated in future work. Eventually, we believe that the presented model in this study is transferable, but thresholds should be adjusted regionally. Future work could explore a global *IDR* parameterization tied to atmospheric covariates.

Conclusion

This study presents a method for identifying sky conditions as a function of cloudiness using hyperspectral observations of solar downwelling irradiance and sky radiance. The diffuse and direct components of solar downwelling irradiance are used to identify periods of 'cloudless clear sky' radiation or 'irradiance similar to clear sky'. The Radiative Transfer models are used to estimate the direct and diffusive components of solar downwelling irradiance. Meanwhile, the $L_S(\lambda)/E_S(\lambda)$ ratio in the blue and NIR regions may be used as an index to distinguish clear sky and overcast conditions. The results indicate that this index is not sensitive to thick and scattered clouds, leading to increased uncertainty in sky condition identification. In contrast, the enhanced method of spectrally integrated diffuse and direct ratios exhibits superior performance, effectively identifies different sky conditions relative to reference model classifications, including those with thick and scattered clouds and rapidly changing cloud cover. Our findings demonstrate that diffuse irradiance is a crucial component for identification of sky cloudiness using hyperspectral measurements. Since in this study solar irradiance components are estimated using radiative transfer models, it is recommended to evaluate the presented method and results using measured diffuse irradiance.

Acknowledgements The authors express their gratitude to the Royal Netherlands Institute for Sea Research (NIOZ) and the Royal Netherlands Meteorological Institute (KNMI) for providing the data of this research. Eric Wagmaakers is acknowledged for his efforts in hyperspectral data collection and management.

Author contributions Masoud, Johan, Leon and Behnaz contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Funding The authors did not receive support from any organization for the submitted work.

Data availability No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

- Alia-Martinez M, Antonanzas J, Urraca R, Martinez-de-Pison FJ, Antonanzas-Torres F (2016) Benchmark of algorithms for solar clear-sky detection. *J Renew Sustain Energy*, 8(3)

- Antonanzas-Torres F, Urraca R, Polo J, Perpiñán-Lamigueiro O, Escobar R (2019) Clear Sky solar irradiance models: A review of Seventy models. *Renew Sustain Energy Rev* 107:374–387. <https://doi.org/10.1016/j.rser.2019.02.032>
- Arabi B, Salama MS, Wernand MR, Verhoef W (2016) MOD2SEA: a coupled atmosphere-hydro-optical model for the retrieval of chlorophyll-a from remote sensing observations in complex turbid waters. *Remote Sens* 8:722
- Arabi B, Moradi M, Hommersom A, van der Molen J, Serre-Fredj L (2025) Assessment of remote sensing reflectance Glint correction methods from fixed automated Above-Water hyperspectral radiometric measurement in highly turbid coastal waters. *Remote Sens* 17:2209. <https://doi.org/10.3390/rs17132209>
- Bi S, Hieronymi M, Röttgers R (2023) Bio-geo-optical modelling of natural waters. *Front Mar Sci* 10:1196352
- Chandaka S, Chatterjee A, Applications SM-ES, with (2009) U., 2009. Cross-correlation aided support vector machine classifier for classification of EEG signals. *Expert Systems with Applications* 36:1329–1336
- Chesnoiu G, Ferlay N, Chiappello I, Auriol F, Catalfamo D, Compiègne M, Elias T, Jankowiak I (2024) Influence of cloudy and clear-sky partitions, aerosols, and geometry on the recent variability in surface solar irradiance components in Northern France. *Atmos Chem Phys* 24:12375–12407
- Choi M, Sander SP, Spurr RJ, Pongetti TJ, van Harten G, Drouin BJ, Diner DJ, Crisp D, Eldering A, Kalashnikova OV (2021) Aerosol profiling using radiometric and polarimetric spectral measurements in the O2 near infrared bands: Estimation of information content and measurement uncertainties. *Remote Sens Environ* 253:112179
- Chylek P, Wong J (1995) Effect of absorbing aerosols on global radiation budget. *Geophys Res Lett* 22:929–931
- Correa LF, Folini D, Chtirkova B, Wild M (2022) Earth Space Sci 9:e2021EA002197. <https://doi.org/10.1029/2021EA002197>. A Method for Clear-Sky Identification and Long-Term Trends Assessment Using Daily Surface Solar Radiation Records
- Curran PJ (1989) Remote sensing of foliar chemistry. *Remote Sens Environ* 30:271–278
- Dogniaux R (1974) Représentations analytiques des composantes du Rayonnement lumineux solaire, conditions de Ciel Serein. Institut Royal Meteorologique de Belgique-
- Ellis BH, Deceglie M, Jain A (2018) Automatic detection of clear-sky periods using ground and satellite based solar resource Data, in: 2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC)(A Joint Conference of 45th IEEE PVSC, 28th PVSEC & 34th EU PVSEC). IEEE, pp. 2293–2298
- Forstinger A, Wilbert S, Jensen A, Kraas B, Peruchena CF, Gueymard C, Ronzio D, Yang D, Collino E, Martinez JP, Ruiz-Arias J, Hanrieder N, Blanc P, Saint-Drenan Y-M (2021) Expert quality control of solar radiation ground data sets. Presented at the ISES Solar World Congress
- Garrison JD (1992) Estimation of atmospheric precipitable water over Australia for application to the division of solar radiation into its direct and diffuse components. *Sol Energy* 48:89–96. [https://doi.org/10.1016/0038-092X\(92\)90037-B](https://doi.org/10.1016/0038-092X(92)90037-B)
- Gege P (2005) The Water Colour Simulator WASI-User manual for version 3
- Gege P (2012) Analytic model for the direct and diffuse components of downwelling spectral irradiance in water. *Appl Opt* 51:1407–1419
- Gege P, Grötsch P (2016) A spectral model for correcting sunglint and skyglint. *Proceedings of Ocean Optics XXIII* 2016 1–10
- Gelaro R, McCarty W, Suárez MJ, Todling R, Molod A, Takacs L, Randles CA, Darmenov A, Bosilovich MG, Reichle R (2017) The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *J Clim* 30:5419–5454
- Gómez-Uceda FJ, Moreno-García IM, Pérez-Castañeda Á, Fernández-Ahumada LM (2021) Study of the dependence of solar radiation regarding design variables in photovoltaic solar installations with optimal dual-axis tracking. *Appl Sci* 11:3917
- Greene CA, Thirumalai K, Kearney KA, Delgado JM, Schwanghart W, Wolfenbarger NS, Thyng KM, Gwyther DE, Gardner AS, Blankenship DD, Geochemistry (2019) *Geophys Geosyst*, 20, 3774–3781. <https://doi.org/10.1029/2019GC008392>
- Gregg WW, Carder KL (1990) A simple spectral solar irradiance model for cloudless maritime atmospheres. *Limnol Oceanogr* 35:1657–1675
- Grenier JC, De La Casinière A, Cabot T (1994) A spectral model of Linke's turbidity factor and its experimental implications. *Sol Energy* 52:303–313
- Griffin MK, Burke HK (2003) Compensation of hyperspectral data for atmospheric effects. *Linc Lab J* 14
- Groetsch PM, Gege P, Simis SG, Eleveld MA, Peters SW (2017) Validation of a spectral correction procedure for sun and Sky reflections in above-water reflectance measurements. *Opt Express* 25:A742–A761
- Gueymard CA (1998) Turbidity determination from broadband irradiance measurements: A detailed multicoefficient approach. *J Appl Meteorol Climatol* 37:414–435
- Gueymard CA (2008) REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation—Validation with a benchmark dataset. *Sol Energy* 82:272–285
- Gueymard CA (2014a) *Renew Sustain Energy Rev* 39:1024–1034. <https://doi.org/10.1016/j.rser.2014.07.117>. A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects
- Gueymard CA (2014b) Impact of on-site atmospheric water vapor Estimation methods on the accuracy of local solar irradiance predictions. *Sol Energy* 101:74–82
- Gueymard CA (2017) Cloud and albedo enhancement impacts on solar irradiance using high-frequency measurements from thermopile and photodiode radiometers. Part 1: impacts on global horizontal irradiance. *Sol Energy* 153:755–765. <https://doi.org/10.1016/j.solener.2017.05.004>
- Gueymard CA (2019a) Clear-sky radiation models and aerosol effects. Fundamentals and Applications, Solar Resources Mapping, pp 137–182
- Gueymard CA (2019b) Clear-Sky radiation models and aerosol effects. In: Polo J, Martín-Pomares L, Sanfilippo A (eds) Solar resources mapping, green energy and technology. Springer International Publishing, Cham, pp 137–182. https://doi.org/10.1007/978-3-319-97484-2_5
- Gueymard CA, Bright JM, Lingfors D, Habte A, Sengupta M (2019) A posteriori clear-sky identification methods in solar irradiance time series: review and preliminary validation using Sky imagers. *Renew Sustain Energy Rev* 109:412–427
- Holben BN, Eck TF, Slutsker Ial, Tanré D, Buis JP, Setzer A, Vermote E, Reagan JA, Kaufman YJ, Nakajima T (1998) AERONET—A federated instrument network and data archive for aerosol characterization. *Remote Sens Environ* 66:1–16
- Ineichen P (2008) Conversion function between the Linke turbidity and the atmospheric water vapor and aerosol content. *Sol Energy* 82:1095–1097
- Jiang X, Xue Y, Jin C, Bai R, Sun Y, Wu S (2022) A simple band ratio library (BRL) algorithm for retrieval of hourly aerosol optical depth using FY-4A AGRI geostationary satellite data. *Remote Sens* 14:4861
- Jordan TM, Simis SG, Grötsch PM, Wood J (2022) Incorporating a hyperspectral Direct-Diffuse pyranometer in an Above-Water reflectance algorithm. *Remote Sens* 14:2491

- Kim D, Ramanathan V (2008) Solar radiation budget and radiative forcing due to aerosols and clouds. *J Geophys Res* 113:2007JD008434. <https://doi.org/10.1029/2007JD008434>
- Kim SY, Kwon DY, Jang A, Ju YK, Lee J-S, Hong S (2024) A review of UAV integration in forensic civil engineering: from sensor technologies to geotechnical, structural and water infrastructure applications. *Measurement* 224:113886
- Kittler R, Darula S (2013) Determination of time and sun position system. *Sol Energy* 93:72–79. <https://doi.org/10.1016/j.solener.2013.03.021>
- Kokaly RF, Asner GP, Ollinger SV, Martin ME, Wessman CA (2009) Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. *Remote Sens Environ* 113:S78–S91
- Kopp G, Lean JL (2011) A new, lower value of total solar irradiance: evidence and climate significance. *Geophys Res Lett*, 38
- Kutser T, Vahtmäe E, Paavel B, Kauer T (2013) Removing Glint effects from field radiometry data measured in optically complex coastal and inland waters. *Remote Sens Environ* 133:85–89
- Kuusk J, Kuusk A (2018) Hyperspectral radiometer for automated measurement of global and diffuse Sky irradiance. *J Quant Spectrosc Radiative Transf* 204:272–280. <https://doi.org/10.1016/j.jqsrt.2017.09.028>
- Lee RG, Hooker S, Mobley CD, McLean S (2002) Above-water radiance and remote sensing reflectance measurement and analysis protocols. *Ocean optics protocols for satellite ocean color sensor validation*, revision 3, 171
- Lin J, Dall’Olmo G, Tilstone GH, Brewin RJ, Vabson V, Ansko I, Evers-King H, Casal T, Donlon C (2022) Derivation of uncertainty budgets for continuous above-water radiometric measurements along an Atlantic meridional transect. *Opt Express* 30:45648–45675
- Linke F (1992) Transmission-koeffizient und Trübungsfaktor. *Beitr Phys Atmos* 10:91–103
- Liou K-N (2002) An introduction to atmospheric radiation. Elsevier
- Long CN, Shi Y (2006) The QCRad value added product: Surface radiation measurement quality control testing, including climatology configurable limits. *Atmospheric Radiation Measurement Program Technical Report*
- Marchenko SV, DeLand MT, Lean JL (2016) Solar spectral irradiance variability in cycle 24: observations and models. *J Space Weather Space Clim* 6:A40
- Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta (BBA)-Protein structure*. 405:442–451
- McNeill VF (2017) Atmospheric aerosols: clouds, chemistry, and climate. *Annu Rev Chem Biomol Eng* 8:427–444. <https://doi.org/10.1146/annurev-chembioeng-060816-101538>
- Mobley CD (1999) Estimation of the remote-sensing reflectance from above-surface measurements. *Appl Opt* 38:7442–7455
- Mobley CD (2015) Polarized reflectance and transmittance properties of windblown sea surfaces. *Appl Opt* 54:4828–4849
- Mobley C (2022) The oceanic optics book. Dartmouth, NS, Canada, International Ocean Colour Coordinating Group (IOCCG) 942. <https://doi.org/10.25607/OBP-1710>
- Molineaux B, Ineichen P, Delaunay J-J (1995) Direct luminous efficacy and atmospheric turbidity—improving model performance. *Sol Energy* 55:125–137
- Moradi M, Arabi B, van der Hommersom A, Samimi C (2024) Quality control tests for automated above-water hyperspectral measurements: radiative transfer assessment. *ISPRS J Photogrammetry Remote Sens* 215:292–312
- Mukkavilli SK, Prasad AA, Taylor RA, Huang J, Mitchell RM, Troccoli A, Kay MJ (2019) Assessment of atmospheric aerosols from two reanalysis products over Australia. *Atmos Res* 215:149–164
- Müller R (2014) Calibration and verification of remote sensing instruments and observations. *Remote sensing*
- Muneer T, Gul MS, Kubie J (2000) Models for estimating solar radiation and illuminance from meteorological parameters. *J Sol Energy Eng* 122:146–153
- Norgren MS, Wood J, Schmidt KS, van Dienenhoven B, Stamnes SA, Ziemba LD, Crosbie EC, Shook MA, Kittelman AS, LeBlanc SE (2021) Above-aircraft cirrus cloud and aerosol optical depth from hyperspectral irradiances measured by a total-diffuse radiometer. *Atmospheric Measurement Techniques Discussions* 2021:1–36
- Nyamsi WW, Saint-Drenan Y-M, Augustine JA, Arola A, Wald L (2024) On the relationships between Clear-Sky indices in photosynthetically active radiation and broadband ranges in overcast and Broken-Cloud conditions. *Remote Sens* 16:3718
- Pérez-Ramírez D, Navas-Guzmán F, Lyamani H, Fernández-Gálvez J, Olmo FJ, Alados-Arboledas L (2012) Retrievals of precipitable water vapor using star photometry: assessment with Raman lidar and link to sun photometry. *J Geophys Research: Atmos*, 117
- Pitarch J, Talone M, Zibordi G, Groetsch P (2020) Determination of the remote-sensing reflectance from above-water measurements with the 3 C model: a further assessment. *Opt Express* 28:15885–15906
- Polo J, Zarzalejo LF, Martín L, Navarro AA, Marchante R (2009) Estimation of daily Linke turbidity factor by using global irradiance measurements at solar noon. *Sol Energy* 83:1177–1185
- Powers DMW (2020) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation
- Quesada-Ruiz S, Linares-Rodríguez A, Ruiz-Arias JA, Pozo-Vázquez D, Tovar-Pescador J (2015) An advanced ANN-based method to estimate hourly solar radiation from multi-spectral MSG imagery. *Sol Energy* 115:494–504
- Remund J, Wald L, Lefèvre M, Ranchin T, Page JH (2003) Worldwide Linke turbidity information. *ISES solar world Congress 2003. International Solar Energy Society (ISES)*, pp 13–p
- Reno MJ, Hansen CW (2016) Identification of periods of clear Sky irradiance in time series of GHI measurements. *Renewable Energy* 90:520–531
- Rozanov VV, Kokhanovsky AA (2005) On the molecular-aerosol scattering coupling in remote sensing of aerosol from space. *IEEE Trans Geosci Remote Sens* 43:1536–1541
- Ruddick KG, De Cauwer V, Park Y-J, Moore G (2006) Seaborne measurements of near infrared water-leaving reflectance: the similarity spectrum for turbid waters. *Limnol Oceanogr* 51:1167–1179
- Ruddick KG, Voss K, Banks AC, Boss E, Castagna A, Frouin R, Hieronymi M, Jamet C, Johnson BC, Kuusk J (2019) A review of protocols for fiducial reference measurements of downwelling irradiance for the validation of satellite remote sensing data over water. *Remote Sens* 11:1742
- Ruiz-Arias JA, Gueymard CA (2023) CAELUS: classification of Sky conditions from 1-min time series of global solar irradiance using variability indices and dynamic thresholds. *Sol Energy* 263:111895
- Saraswat M, Sharma AK (2013) Genetic algorithm for optimization using MATLAB. *Int J Adv Res Comput Sci* 4:155–159
- Sasaki Y (2007) The truth of the F-measure. *Teach Tutor Mater* 1:1–5
- Sayer AM, Munchak LA, Hsu NC, Levy RC, Bettenhausen C, Jeong M-J (2014) MODIS collection 6 aerosol products: comparison between aqua’s e-Deep blue, dark target, and merged data sets, and usage recommendations. *J Geophys Research: Atmos* 119:13–965
- Simis SG, Olsson J (2013) Unattended processing of shipborne hyperspectral reflectance measurements. *Remote Sens Environ* 135:202–212
- Smirnov A, Holben BN, Kaufman YJ, Dubovik O, Eck TF, Slutsker I, Pietras C, Halthore RN (2002) Optical properties of atmospheric aerosol in maritime environments. *J Atmos Sci* 59:501–523

- Sun X, Bright JM, Gueymard CA, Acord B, Wang P, Engerer NA (2019) Worldwide performance assessment of 75 global clear-sky irradiance models using principal component analysis. *Renew Sustain Energy Rev* 111:550–570
- Vaquero-Martínez J, Antón M (2021) Review on the role of GNSS meteorology in monitoring water vapor for atmospheric physics. *Remote Sens* 13:2287
- Vermote EF, Tanré D, Deuze JL, Herman M, Morcette J-J (1997) Second simulation of the satellite signal in the solar spectrum, 6S: an overview. *IEEE Trans Geosci Remote Sens* 35:675–686
- Wernand MR (2002) Guidelines for (ship-borne) auto-monitoring of coastal and ocean colour, in: *Proceedings of Ocean Optics XVI*
- Wild M (2005) Solar radiation budgets in atmospheric model inter-comparisons from a surface perspective. *Geophys Res Lett* 32
- Wood J, Smyth TJ, Estellés V (2017) Autonomous marine hyperspectral radiometers for determining solar irradiances and aerosol optical properties. *Atmos Meas Tech* 10:1723–1737
- Younes S, Muneer T (2007) Clear-sky classification procedures and models using a world-wide data-base. *Appl Energy* 84:623–645
- Zhang Y, Li X, Bai Y (2015) An integrated approach to estimate short-wave solar radiation on clear-sky days in rugged terrain using MODIS atmospheric products. *Sol Energy* 113:347–357
- Zhang W, Kleiber W, Florita AR, Hodge B-M, Mather B (2018) A stochastic downscaling approach for generating high-frequency solar irradiance scenarios. *Sol Energy* 176:370–379
- Zibordi G, Mélin F, Berthon J-F, Holben B, Slutsker I, Giles D, D'Alimonte D, Vandemark D, Feng H, Schuster G (2009) AERONET-OC: a network for the validation of ocean color primary products. *J Atmos Ocean Technol* 26:1634–1651

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.