

# Cloud classification through machine learning and global horizontal irradiance data analysis

Anabela Rocío Lusi<sup>1,2</sup>  | Pablo Facundo Orte<sup>1,2</sup> | Elian Wolfram<sup>3</sup> | José Ignacio Orlando<sup>4</sup>

<sup>1</sup>Departamento de Investigaciones en Láseres y Aplicaciones (DEILAP), UNIDEF (CITEDEF-CONICET), IRL-3351-IFAECI-CNRS, Villa Martelli, Argentina

<sup>2</sup>Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina

<sup>3</sup>Servicio Meteorológico Nacional (SMN), Dirección de redes de Observación, Buenos Aires, Argentina

<sup>4</sup>Yatiris Group, Instituto PLADEMA, CONICET, UNCPBA, Tandil, Argentina

## Correspondence

Anabela Rocío Lusi, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires (C1425), Argentina.

Email: [anabelalusi@gmail.com](mailto:anabelalusi@gmail.com)

## Funding information

Agencia I+D+i, Grant/Award Number: PICT-2020-SERIEA-026 01; CONICET PIP 2021–2023, Grant/Award Numbers: 11220200101558CO, 11220200102472CO

## Abstract

Cloud observations and characterization are crucial owing to their influence on energy balance, climate, and weather. Their particular effects on radiation vary depending on different cloud parameters, such as cloud base or top height, water content, and cloud optical thickness, all of them closely related to the specific cloud type. Cloud classification therefore becomes a crucial task in meteorology, although it remains challenging for weather services worldwide owing to the intensive associated labor and cost. In this study we introduce a new low-cost method for automating cloud classification based on a combination of ground-based global horizontal irradiance (GHI) measurements, a clear-sky model, and machine learning. Based on the hypothesis that different cloud types have their own GHI signatures, we trained different supervised learning algorithms using GHI data manually labeled by meteorological observers from time-synchronized all-sky images. Multiple time windows were extracted from each GHI series, with eight features defined in each case to characterize the sequence. The best outcome was achieved using an XGBoost model on features extracted on time windows of 33 min, obtaining an accuracy of 0.88 and a Cohen's kappa of 0.84 in a held-out test set. The development presented in this study has the ability to provide low-cost cloud classification from ground-based observations, which is a challenge for weather services worldwide owing to intensive labor and cost.

## KEY WORDS

all-sky images, cloud classification, ground-based measurements, solar irradiance, supervised learning algorithms

## 1 | INTRODUCTION

Clouds exert a significant influence on Earth's energy balance, influencing both incoming solar radiation and outgoing long-wave radiation (Stephens & L'Ecuyer, 2015). Equilibrium is pivotal for Earth's climate and weather dynamics. A critical parameter within this context is global horizontal irradiance (GHI), denoting the total

short-wave radiation received from above by a surface horizontal to the ground. The role of clouds in solar irradiance modulation involves processes such as reflection, absorption, and scattering of radiation by cloud particles. Their impact on GHI is intricately linked to factors such as volume, shape, thickness, and composition (Deneke *et al.*, 2008; Kokhanovsky, 2004; Ododo *et al.*, 1996; Orsini *et al.*, 2002). Furthermore, the effect of a solitary cloud

on solar irradiance contrasts with that of multiple clouds or overcast skies. This is because the decrease in irradiance depends on whether the sun is obscured by clouds and on the extent of cloud coverage. If a cloud is positioned between the sun and the observer, it can block the sun, leading to a decrease in direct normal irradiance. However, under cloudy conditions, diffuse irradiance can increase because of the reflection of solar irradiance from the underside of clouds and the scattering of direct irradiance by cloud particles (Tapakis & Charalambides, 2013). Therefore, the presence and movement of clouds can cause large short-term variations in the surface GHI, which can have profound implications, particularly in fields such as solar energy, affecting energy production, especially when interrupted for short periods (Abdellatif *et al.*, 2015).

The World Meteorological Organization (WMO) classifies clouds into 10 different types, using a classification system based on genera, species, and varieties (WMO, 2017). Clouds are organized into 10 primary categories, referred to as genera, with each observed cloud belonging to a single genus. Apart from this classification, the WMO further categorizes clouds based on their altitude, dividing the troposphere into three vertical levels: high, middle, and low. The determination of each level is based on the vertical height range at which specific cloud types are typically observed. These levels are identified as high-level clouds (cirrus, Ci; cirrocumulus, Cc; cirrostratus, Cs), mid-level clouds (altocumulus, Ac; altostratus, As; nimbostratus, Ns), and low-level clouds (stratocumulus, Sc; stratus, St; cumulus, Cu; cumulonimbus, Cb).

Cloud classification can contribute to different fields. It can provide valuable information for activities significantly affected by weather and cloud conditions, such as aviation, agricultural, maritime navigation, and warning systems due to potential weather hazards. On the other hand, cloud classification can contribute to several practical applications, such as data collection for meteorological agencies, photovoltaic power prediction, and weather forecasting (Duchon & O'Malley, 1999; Li *et al.*, 2023; Zhu *et al.*, 2022).

The accurate classification of clouds is crucial for predicting their motion and how they affect the amount of solar radiation that passes through them. First, different cloud types obstruct varying amounts of solar radiation. This depends on the type of cloud and can be affected by their size and opacity (Matuszko, 2012). In addition, ice crystals and water droplets have different effects on the absorption and scattering of solar radiation (Sun & Shine, 1994). Second, because different types of clouds are located at different altitudes, they can be influenced by different air currents, causing them to move in varying directions at different speeds (Zhang *et al.*, 2005). Finally, weather conditions at different altitudes can affect

cloud development over time, potentially leading to rainfall (Tapakis & Charalambides, 2013).

Over the past few decades, surface cloud information has been recorded through human observations, where the amount of clouds covering the sky (octas) and the types of clouds have been distinguished based on their appearance (WMO, 2017). However, this data collection method is expensive and subjective (Luiz *et al.*, 2018; Martínez-Chico *et al.*, 2011; Utrillas *et al.*, 2022). Therefore, many researchers have reported more objective approaches for estimating the amount of clouds in the sky using different instruments. Some of these alternatives use downward long-wave radiation along with other meteorological parameters acquired at the surface (Dürr & Philipona, 2004; Marty & Philipona, 2000). Other studies have investigated the cloud cover conditions based on satellite data (Escríg *et al.*, 2013; Jang *et al.*, 2016; Liang & Yuan, 2016; Nespoli *et al.*, 2022; Perez *et al.*, 2002) and/or on all-sky camera images (Kassianov *et al.*, 2005; Long *et al.*, 2006; Pfister *et al.*, 2003; Sabburg & Long, 2004). Given the difficulties associated with visual classification, particularly in terms of subjectivity and cost, researchers have sought to develop more objective, cost-effective methodologies for cloud detection.

Duchon and O'Malley (1999), on the other hand, proposed a method to discriminate the cloud type based on a pyranometer in areas with restricted sky observations. They used irradiance data and performed statistical analyses to determine the cloud types. However, their approach showed agreement with human observations only 45% of the time. These differences were mainly attributed to the pyranometer's focus on clouds crossing the sun's path and its susceptibility to aerosol interference. Our work is inspired by their approach and serves as a basis for the development of our own methodology.

Considerable progress has been made in the development of machine-learning (ML) methodologies and practical applications in Earth science (e.g., crop disease detection, air pollution estimation, and precipitation forecasting; Lary *et al.*, 2018; Zhou *et al.*, 2021). These types of models are flexible enough to be applied in different scenarios, as they are able to automatically learn patterns and relationships between their inputs and outputs. Several researchers have used these techniques in tasks such as clear-sky detection (Hollstein *et al.*, 2016; Liu *et al.*, 2021; Mommert, 2020) or to detect different cloud conditions (Heinle *et al.*, 2010; Jang *et al.*, 2016; Lee *et al.*, 2004; Taravat *et al.*, 2015). Taravat *et al.* (2015), for instance, developed a method that automatically classifies clouds in ground-based whole-sky images using neural networks and support vector machines. The pixel values of the red, green, and blue bands in the images served as inputs for the models, and the outputs provided pixels classified in

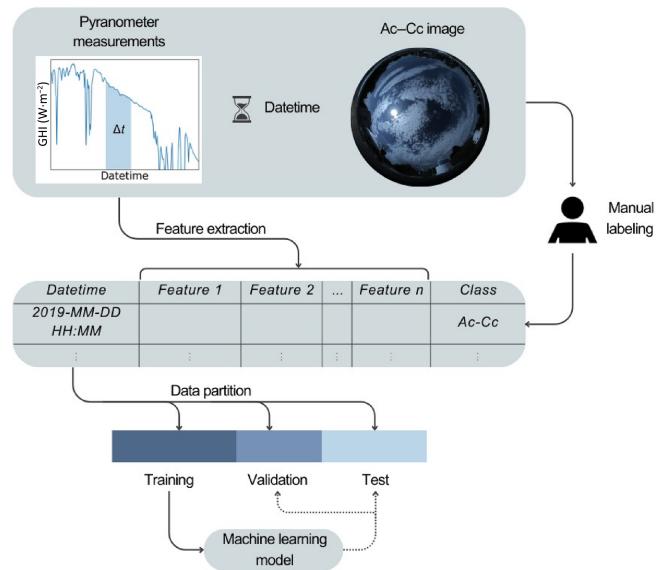
terms of cloud coverage or others (cloud-free pixels and sun). Both models achieved an accuracy of around 95% in detecting clouds in a given region of an image. These studies have shown the great potential of ML in identifying periods of clear and cloudy skies. However, working with all-sky camera images presents its own challenges in terms of computational and complexity cost.

In this article we introduce a novel ML-based classification approach for discriminating different cloud types using ground-based GHI observations and a clear-sky model that estimates the GHI under clear skies (McCleary model; Lefèvre *et al.*, 2013). The labels for training the supervised models were derived from the classification of all-sky images conducted by meteorological observers using the classes proposed by Loyber (2022), which was based on the classification delineated by Liu *et al.* (2019). This approach aims to enhance the accuracy and efficiency of cloud classification, contributing to a more robust understanding of atmospheric conditions based on GHI data and clear-sky modeling. This is the first ML development that uses this specific data to classify cloudiness, according to our literature review.

The remainder of the article is organized as follows: Section 2 summarizes our data acquisition and annotation process and the design and implementation process of our cloud classification approach, including details about the GHI-based features and the ML models. Results are presented and discussed in detail in Sections 3 and 4 respectively. Lastly, conclusions are summarized in Section 5.

## 2 | MATERIALS AND METHODS

A schematic representation of our experimental design is shown in Figure 1. Labels were assigned to determine the class in each all-sky image at a specific datetime through visual assessment by meteorological observers (Section 2.1). Simultaneously, a pyranometer measurement window was extracted to obtain the relevant GHI features (Section 2.3). We performed experiments with time windows ranging from 9 to 41 min, with increments of 2 min in window size. The resulting dataset was subsequently partitioned into training, validation, and testing sets as detailed in Section 2.2. We developed various ML classification models to discern among different cloud classes (Section 2.4) that correspond to six groups of different cloud types, following the classification proposed by Liu *et al.* (2019) and Loyber (2022): (1) Ac and Cc; (2) Ci and Cs; (3) clear sky (Clear-sky); (4) Cb and Ns; (5) Cu; (6) Sc, St, and As (see Figure 2). From here on, these classes will be referred to as “not mixed”. Although “Clear-sky” is not technically a cloud class, but rather a condition of the sky, we refer to all categories collectively as “cloud classes”



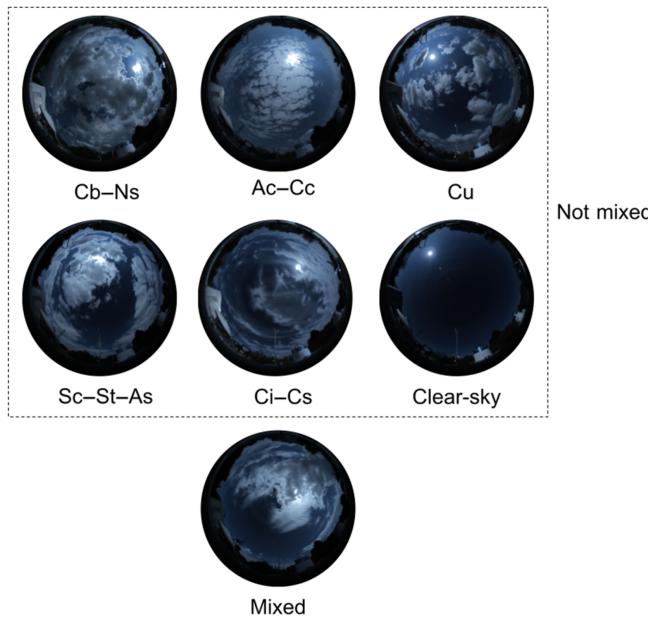
**FIGURE 1** Experimental design. Manual labeling was performed by meteorological observers from the all-sky images. Features were extracted from pyranometer (global horizontal irradiance, GHI) measurements data at the same datetime. Data partitioning (training, validation, test) was applied to train, calibrate, and evaluate the respective machine-learning models. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

to ensure consistency and ease of comprehension throughout the article. On the other hand, cases with cloud types belonging to more than one group were labeled as “mixed”.

### 2.1 | Data acquisition and manual labeling

The data used in this study were retrospectively collected from Earth-based observation made in Villa Martelli ( $34.58^{\circ}\text{S}$ ,  $58.48^{\circ}\text{W}$ , 25 m a.s.l.) in 2019. Villa Martelli is situated in the northern zone of Greater Buenos Aires, a province located in the central-eastern region of Argentina. One-minute GHI data were obtained using a Kipp & Zonen CMP-21 pyranometer in the total range (285–2800 nm) belonging to the Saver-Net Argentine network (Orte *et al.*, 2022). These instruments typically report an uncertainty of  $\pm 2\%$  for daily totals (OTT HydroMet 2023). The measurements have exhaustive quality control, through visual inspection of the data series and graphical tools, in order to meet the criteria established by the Baseline Surface Radiation Network for GHI (Ohmura *et al.*, 1998). The defective or shadowed periods were removed from the data series.

At the same time points, all-sky images were recorded using a SONA 201D V6.1 camera (Sieltec Canarias S.L.). These cameras, equipped with an array of sensors and a fisheye lens, enable monitoring sections of the



**FIGURE 2** Examples of all-sky images used for producing manual cloud-type labels for our experiments. Cb-Ns: cumulonimbus and nimbostratus; Ac-Cc: altocumulus and cirrocumulus; Cu: cumulus; Sc-St-As: stratocumulus, stratus, and altocumulus; Ci-Cs: cirrus and cirrostratus; Clear-sky, clear sky. Notice that the class “mixed” corresponds to those images exhibiting a combination of different cloud class (the example exhibits Ac and St, which belong to Ac-Cc and Sc-St-As classes). [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

atmosphere through a narrow field of view. Thanks to the resulting panoramic view of the sky-dome, they can therefore capture, store, and process images of the celestial vault, covering a maximum angle of around 180° field of view.

A total of 7,025 images were collected for the corresponding time period and then manually labeled by meteorological observers from the Argentine Servicio Meteorológico Nacional (<https://www.smn.gob.ar/>), who assigned different cloud classes (Figure 2). All meteorological observers were trained based on the competencies of Annex 5.A of the WMO *Guide to Instruments and Methods of Observation* Vol. 5 (WMO, 2018). The labeling process involved several iterations until achieving sets verifiable by all persons involved. Additionally, there were “administrator” profiles overseeing the task and simultaneously validating the accuracy of the classifications. The manual labeling is comprehensively explained in Loyber (2022). As the model was trained considering the six not-mixed classes, mixed cases were clustered and a specific test was conducted as outlined in Section 2.5. Since the Cb-Ns class had very few samples, we decided to augment this subset by retrospectively collecting new cases into validation/test partitions

under the assumption that lightning only occurs in the presence of Cb clouds, following the WMO cloud identification guide (Holzworth, 2023; Kaplan & Lau, 2021; Rawlins, 1982; WMO, 2017).

The detection of lightning was determined using the information from the electrical activity records obtained from the World Wide Lightning Location Network, a network of very low frequency (VLF) radio lightning sensors operated by the University of Washington in Seattle (USA) (Holzworth, 2023). Most ground-based observations in the 3–30 kHz VLF band are dominated by impulsive signals from lightning discharges called “sferics”. Significant radiated electromagnetic power exists from a few hertz to several hundred megahertz, with the bulk of the energy radiated at VLF.

Hence, we examined instances where lightning was detected within a radius of 14 km around the site, which is approximately the same perspective achievable with an all-sky camera, and finally added these cases to the Cb-Ns class.

## 2.2 | Data preparation

We split our dataset into 80% and 20% for training and testing respectively. The 80% saved for training was further divided into 80% for training the models and 20% for calibrating their hyperparameters (validation set). To guarantee the independence and identical distribution of samples, the sets were produced using a season-based approach. Specifically, we divided the data year into four groups corresponding to the beginning standard dates of equinoxes and solstices for the Southern Hemisphere, which occur around December 21 for summer, March 21 for autumn, June 21 for winter, and September 21 for spring). The same proportion of samples from each season was preserved in the training, validation, and test sets. Each seasonal set was treated independently, and the samples were separated into training, validation, and testing sets, considering that there were no samples close together in time. Subsequently, to minimize temporal correlations in samples within the validation and test sets, we removed those samples that fell within a ±3 hr time-frame. Finally, for the remaining validation and test samples, the cloud-type labels were reviewed to ensure their accuracy. All these steps helped us avoid any biases or patterns in the data that could affect the model’s performance. As a result, we obtained training, validation, and test sets with 3,031, 161 and 161 samples respectively. Figure 3 illustrates the different types of clouds in each data partition, and Table 1 lists the number of samples for each cloud type in the dataset.



**FIGURE 3** Distribution of samples for each cloud class in the training, validation and test sets. Ac-Cc: altocumulus and cirrocumulus; Cb-Ns: cumulonimbus and nimbostratus; Ci-Cs: cirrus and cirrostratus; Cu: cumulus; Sc-St-As: stratocumulus, stratus, and altocumulus. [Colour figure can be viewed at wileyonlinelibrary.com]

**TABLE 1** Sample numbers of six cloud types in the dataset.

Cloud type	No. of samples
Clear-sky	931
Ci-Cs	647
Ac-Cc	558
Cu	458
Sc-St-As	373
Cb-Ns	386

Abbreviations: Ac-Cc, altocumulus and cirrocumulus; Cb-Ns, cumulonimbus and nimbostratus; Ci-Cs, cirrus and cirrostratus; Cu, cumulus; Sc-St-As, stratocumulus, stratus, and altocumulus.

## 2.3 | Feature extraction

Images classified by the observers were matched with the corresponding GHI time stamp. A time window  $\Delta t$  was extracted from each GHI data series, centered around the specific time stamp of the corresponding image, to characterize the period of irradiance associated with the observed cloud. The clear-sky GHI ( $\text{GHI}_{\text{CS}}$ ) was also calculated within the time window using the McClear model (Lefèvre *et al.*, 2013), which was evaluated in previous research at this site (Lusi *et al.*, 2023). We evaluated different features, all of which were selected considering the research hypothesis, which is that different cloud types introduce characteristic short-term variability in GHI. To mitigate potential redundancies in data, the

combination of features was determined by analyzing their Pearson correlation in the training set, as suggested by Assegie *et al.* (2021). This procedure allows identifying and removing redundant features that were highly correlated, thereby reducing the dimensionality of the dataset. In instances in which a high correlation was detected, we conducted an in-depth examination of the characteristics and selected the most representative features. This made the model more efficient, potentially improved its performance, and mitigated the risk of overfitting by preventing the model from overemphasizing the importance of these features. Moreover, this process resulted in a model that was easier to interpret, owing to the reduced number of features. Notably, all decisions affecting the model were made using only the training set to avoid data leakage, ensuring that no information from the test set was inadvertently introduced into the model. This comprehensive approach effectively mitigated potential redundancies in our data. Finally, eight features were extracted:

1. Standard deviation of the variability index,  $\text{std}(\text{VI})$ .
2. Root-mean-square deviation (RMSD) between GHI and  $\text{GHI}_{\text{CS}}$ .
3. Standard deviation of the modified clearness index,  $\text{std}(k_t^*)$ .
4. Mean value of the modified clearness index,  $\bar{k}_t^*$ .
5. Maximum value of the modified clearness index,  $\max(k_t^*)$ .
6. Skewness index of the modified clearness index,  $\text{skew}(k_t^*)$ .
7. Kurtosis indices of the modified clearness index,  $\text{kurt}(k_t^*)$ .
8. Average variability  $\sigma$ .

All features implicitly incorporate adjustments relative to the clear-sky model or extraterrestrial irradiance pattern. They were standardized based on their own mean and standard deviation values to account for differences in their ranges. Statistics for standardizations were obtained from the training set to avoid data leakage. Each feature is explained in detail in the following and summarized in Table A1 of the Appendix.

### 2.3.1 | Standard deviation of variability index

The variability index (VI) is the normalized line length of irradiance, as presented in (Stein *et al.*, 2012). It is a simple measure of irradiance variability (unitless) over a period of time. VI is calculated following Equation (1), where GHI and  $\text{GHI}_{\text{CS}}$  are vectors of length  $n$  of observed and clearsky

(modeled) GHI averaged at some time interval in minutes  $\Delta t$ , respectively. The VI is calculated for each minute over  $\Delta t$  and then the standard deviation is applied:

$$\text{VI} = \frac{\sum_{k=2}^n \sqrt{(\text{GHI}_k - \text{GHI}_{k-1})^2 + \Delta t^2}}{\sum_{k=2}^n \sqrt{(\text{GHI}_{\text{CS}_k} - \text{GHI}_{\text{CS}_{k-1}})^2 + \Delta t^2}}. \quad (1)$$

### 2.3.2 | RMSD between GHI and $\text{GHI}_{\text{CS}}$

The RMSD ( $\text{W}\cdot\text{m}^{-2}$ ) measures the dispersion of the deviation. The RMSD feature used here is an analogy of the root-mean-square error commonly used to determine statistical errors but calculated between the GHI measurement and model for clear-sky conditions. In other words, it indicates how far GHI values fall from  $\text{GHI}_{\text{CS}}$  values. It is calculated as

$$\text{RMSD} = \frac{\sqrt{\sum_{i=1}^n (\text{GHI}_i - \text{GHI}_{\text{CS}})^2}}{n}, \quad (2)$$

where the  $n$  value is the length of the time interval considered.

### 2.3.3 | Standard deviation, mean and maximum of the modified clearness index $k_t^*$

The clearness index  $k_t$  is the ratio between the GHI on the surface of the Earth and extraterrestrial radiation at the top layer of the atmosphere. It represents the proportion of extraterrestrial solar radiation that reaches the surface. The GHI is influenced by clouds, which modulate the amount of solar radiation reaching the surface. By analyzing these variations, we aim to identify different cloud types and their impacts. Perez *et al.* (1990) modified  $k_t$  to avoid dependence on solar elevation and used it as a more reliable indicator of sky conditions, known as the modified clearness index  $k_t^*$ . This index is calculated from the relationship between  $k_t$  and the optical air mass  $m$  (Kasten & Young, 1989):

$$k_t^* = \frac{k_t}{1.031 \times \exp \left[ \frac{-1.4}{0.9 + (9.4/m)} \right] + 0.1}. \quad (3)$$

The  $k_t^*$  (unitless) for each minute of measurement within the time interval is calculated, and then the period is represented by the standard deviation, mean, and maximum value.

### 2.3.4 | Skewness and kurtosis indices of the modified clearness index $k_t^*$

Skewness and kurtosis (unitless) are two statistical measures used to describe the shape and distribution of a dataset (Garcia-Gutierrez *et al.*, 2022). Skewness is a measure of the asymmetry of the data distribution, indicating whether the distribution is symmetric or skewed to one side. Hence, if the skewness value is close to zero then the distribution is approximately symmetric. Kurtosis, on the other hand, is a measure of the concentration of data in the tail of a distribution, indicating whether the distribution has a concentration of data at the center or the tail. Thus, a kurtosis value close to zero indicates that the distribution has a concentration of data similar to the normal distribution. These indices were calculated in the  $k_t^*$  time series to characterize the period.

### 2.3.5 | Average variability

The variability metric for a single location is defined as the standard deviation of the change in  $k_t^*$  values, within a 4-min moving window:

$$\sigma = \text{std}(\Delta k_t^*). \quad (4)$$

This variability is directly proportional to the change in the clear-sky index using the specified time interval  $\Delta t$  (Hoff & Perez, 2010; Perez & Hoff, 2013). Once the variability metric, or sigma  $\sigma$ , was calculated in minute intervals, an average was computed.

## 2.4 | ML methods

To understand the influence of different ML algorithms on the results, we evaluated a series of popular supervised learning methods for multiclass classification. In particular, we studied one-vs.-rest (Whitaker *et al.*, 2021) and softmax logistic regression (Deng *et al.*, 2023) and random forest (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016). We chose these approaches owing to their linear and their nonlinear natures respectively, which allows us to analyze the influence on these assumptions in the results.

All models were trained, calibrated, and evaluated using the same training, validation and test partitions, obtained as explained in Section 2.2. To optimize their individual hyperparameters, we used random search on the validation set, and the optimal configuration was then used for testing. We focused on optimizing the model

**TABLE 2** Parameters setting to determine the optimal hyperparameters for the models.

Model	Hyperparameter	Threshold									
One-vs.-rest logistic regression	$C$	-5	-4	-3	-2	-1	0	1	2	3	4
	Penalty	l1	l2								
Softmax regression	$C$	-3	-2	-1	0	1	2	3	4	5	
Random forest	n_estimator	5	10	25	50	100	200				
	max_depth	3	5	7	9	11	None				
XGBoost	n_estimator	100	150	200							
	max_depth	3	4	5	7						
	learning_rate	0.01	0.03	0.1							

by using a held-out test set to ensure its performance and generalizability. It is worth noting that the softmax loss function was utilized for XGBoost, and all configurations explored for the models are summarized in Table 2.

## 2.5 | Evaluation metrics

Models were evaluated in terms of accuracy, Cohen's kappa, and  $F_1$  score. Accuracy is a metric for multiclass classification defined as the number of correct predictions divided by the total number of predictions. Cohen's kappa score is also used for evaluating multiclass classifiers, and it measures the numerical rating of the degree of agreement between a detection result and a ground truth labeling. Additionally,  $F_1$  score, another metric commonly used for evaluating binary classifiers, is robust enough to be applied in datasets with imbalanced classes. In this context, we applied it in each class individually, rather than providing a macro average. This approach is effective in handling multiclass problems, offering insights into the performance of the model across different classes. We also included precision and recall values on these analyses, to comprehensively understand the degree of false positives and negatives when detecting individual classes.

On the other hand, we utilized a confusion matrix to provide a detailed analysis of the predictive capabilities of the model by comparing the predicted labels against the ground-truth labels.

The discrimination of mixed clouds was conducted by means of an entropy analysis to discriminate between mixed and non-mixed clouds using receiver operating characteristic (ROC) analysis. The ROC curves display the sensitivity or true-positive rate versus the false-positive rate (or 1 – Specificity) and show a possible trade-off

between sensitivity and specificity. The ideal ROC curve has an area under the curve (AUC) value of 1.

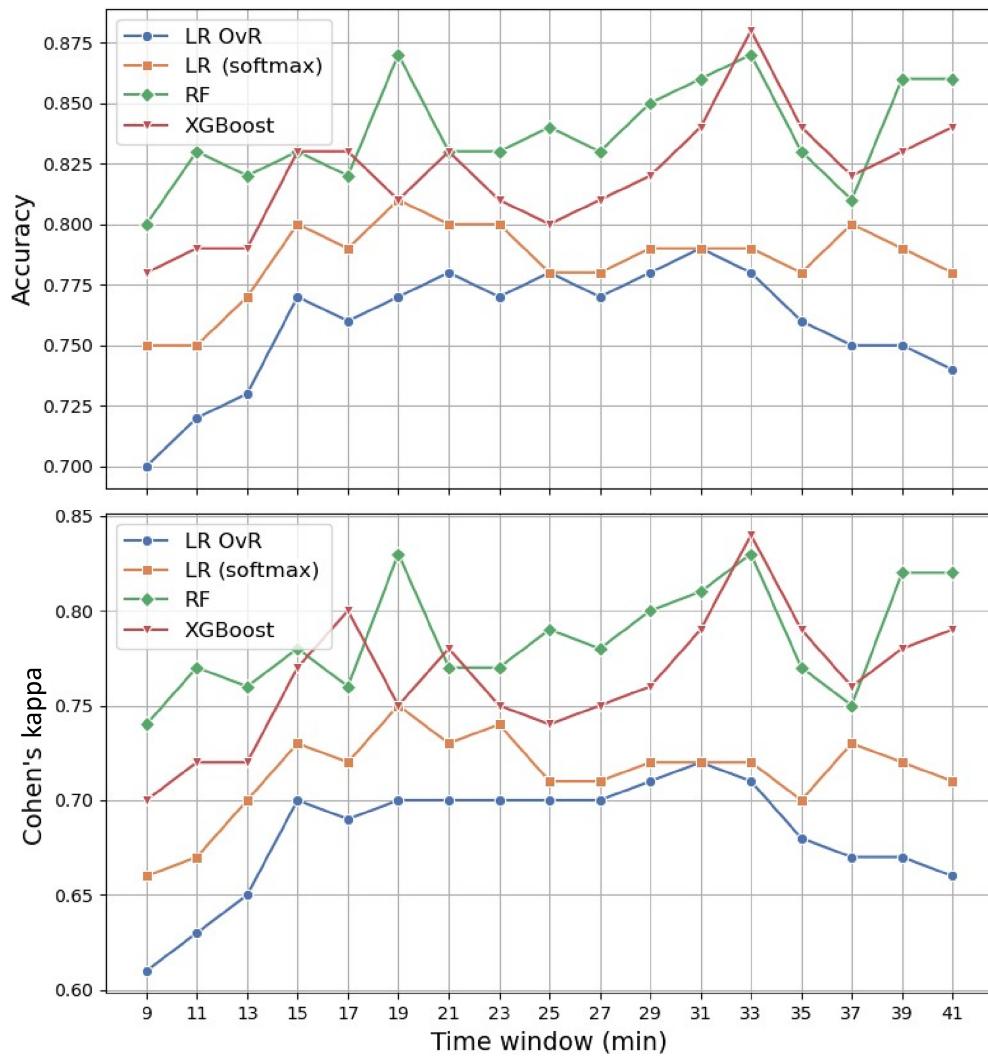
## 3 | RESULTS

### 3.1 | Models' performance with respect to $\Delta t$ in the test set

In order to determine the optimal  $\Delta t$ , an extensive range of widths were tested for characterizing cloudiness. We trained our algorithm using  $\Delta t$  ranging from 9 to 41 min, with a 2 min increment (i.e., we tested 17 widths for each model). These ranges were chosen because a window shorter than 9 min provides a point-in-time measurement, particularly if a single cloud persists between the pyranometer and the direct beam within the window width. Conversely, a much longer window, such as 50 min, is also inappropriate because it may encompass different types of cloudiness (Duchon & O'Malley, 1999). Figure 4 shows the accuracy and Cohen's kappa values obtained for each  $\Delta t$  within the test set. It is observed that the performance of all models showed enhancement with increasing  $\Delta t$  until  $\sim 31$  or 33 min, with the nonlinear models demonstrating superior scores. In particular, the XGBoost model obtained its best performance with a  $\Delta t$  of 33 min, achieving an accuracy of 0.88 and Cohen's kappa of 0.84. On the other hand, the random forest model also performed optimally with a  $\Delta t$  of 19 min, attaining an accuracy of 0.87 and Cohen's kappa of 0.83.

### 3.2 | Model performance with respect to cloud type in the test set

Figure 5 displays the  $F_1$  score achieved for detecting each specific cloud type by the random forest and XGBoost models, using different  $\Delta t$ . It can be observed that for



**FIGURE 4** Accuracy and Cohen's kappa for the logistic regression (LR) one-vs.-rest (OvR), LR softmax, random forest (RF), and XGBoost models in each time interval in the test set. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

both models the performance of the clear-sky class remained largely unaffected by varying the  $\Delta t$  widths. This consistent pattern is also evident in the Cb-Ns and Cs-Ci classes, with scores of around 0.9 and 0.8 respectively. Notably, within the Cb-Ns class, random forest exhibited a more consistent and higher level of performance. Furthermore, for the Ac-Cc class we obtained  $F_1$  scores close to 0.6 with both models. The Cu class showed the best  $F_1$  score at 19 min in the random forest model. In contrast, the XGBoost model exhibited a different behavior, with the performance improving as the  $\Delta t$  increased, reaching its maximum at 31 min. Conversely, the Sc-St-As class started with a low  $F_1$  score and decreased notably as the  $\Delta t$  expanded.

### 3.3 | Top-performing model analysis in the test set

The XGBoost model outperformed all other approaches in the validation set, with its best configuration being the one

with 200 estimators and a  $\Delta t$  of 33 min. Hence, we evaluated this particular model on the test set. Figure 6 shows the resulting confusion matrix, and Table 3 presents the evaluation metrics obtained for each individual cloud type. Notice that the normalized  $(i, j)$  values in the matrix represent the proportion of instances of a class  $i$  classified as  $j$  with respect to the total number of samples. Although these values do not directly correspond to recall or precision, the values on the diagonal do represent recall for each class, as they reflect the proportion of true positives correctly classified. We found that the clear sky was notably predicted, showing an  $F_1$  score of 99%. The Cb-Ns class also performed well, with an  $F_1$  score of 97%. The Ac-Cc, and Ci-Cs classes were classified with an accuracy of 70% and 87% respectively, and  $F_1$  scores of 68% and 88% respectively. Approximately 82% of the predictions in the Cu class were aligned with the ground-truth label. On the other hand, the model faced challenges in accurately classifying the Sc-St-As class, achieving a 64% accuracy rate and  $F_1$  score of 67%. The model often confused this class with Ac-Cc, and Cu, as observed in the confusion matrix.

**FIGURE 5**  $F_1$  score for random forest (RF) and XGBoost in different time intervals in the test set. Ac, altocumulus; As, altostratus; Cb, cumulonimbus; Cc, cirrocumulus; Ci, cirrus; Cs, cirrostratus; Cu, cumulus; Ns, nimbostratus; Sc, stratocumulus; St, stratus. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

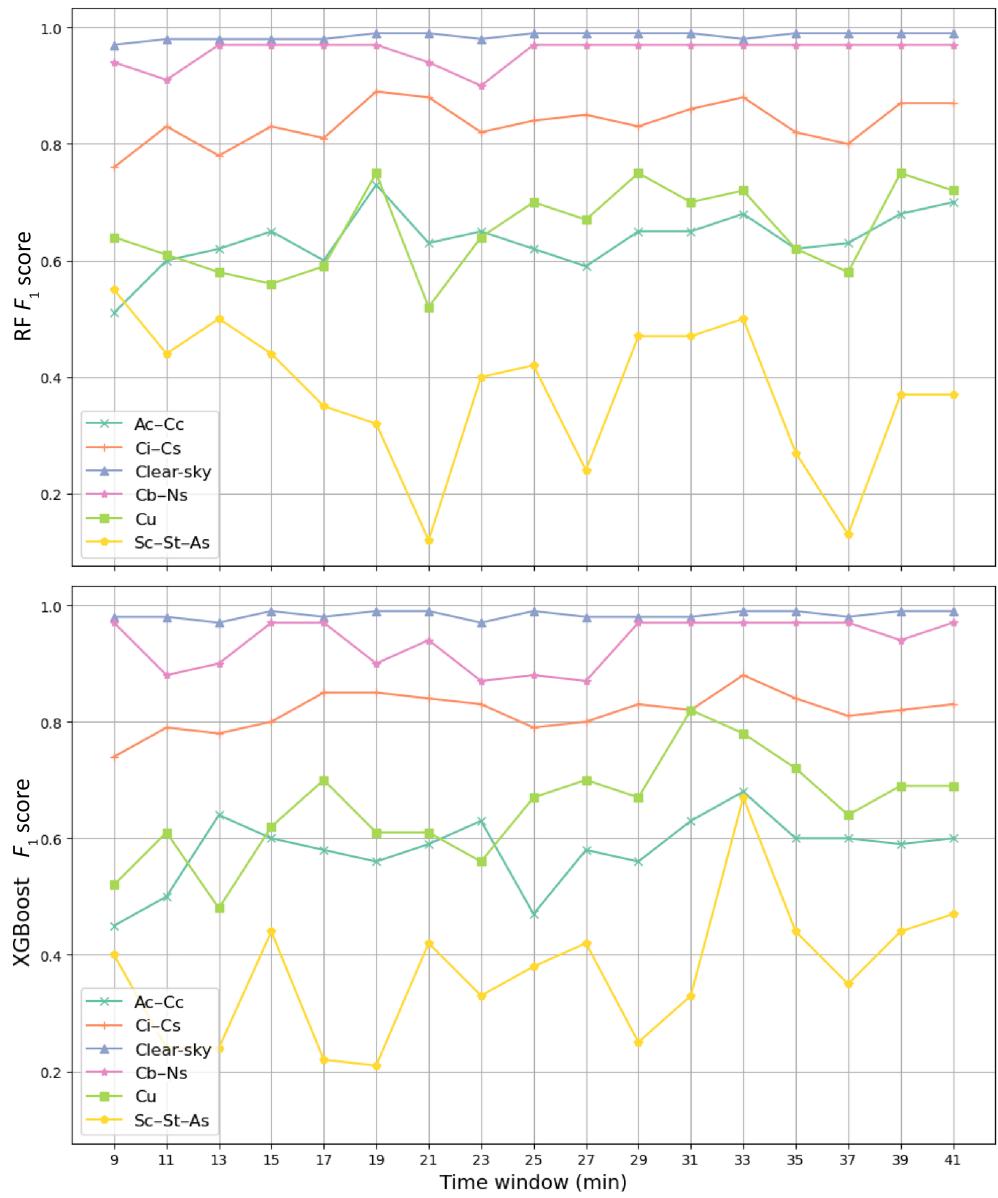


Figure 7 presents the feature importance of XGBoost. The analysis revealed that the most significant importance was associated with the RMSD, the average variability  $\sigma$  and the standard deviation of the  $k_t^*$ .

### 3.4 | Managing multiple cloud classes

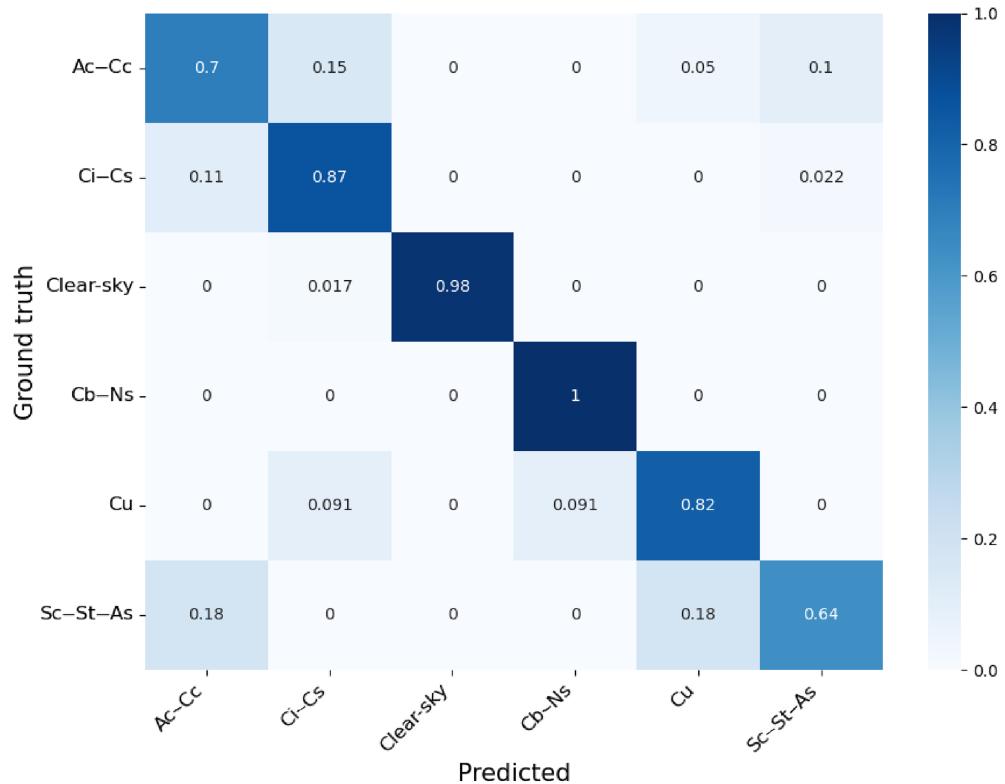
To demonstrate the capability of our algorithm in distinguishing scenarios in which multiple types of cloud classes coexist, we conducted an uncertainty analysis by calculating the Shannon entropy of the output probabilities. Our hypothesis was that the model should exhibit high uncertainty (and therefore high entropy) in those cases labeled as mixed cloud, as it should not be able to discriminate a specific class over the others. To confirm or discard this first, we computed the entropy on images labeled as mixed

class or not and evaluated the resulting value as a score for discriminating the mixed class from not-mixed. Figure 8 shows the resulting ROC curve and the boxplots of the entropies between not-mixed and mixed classes.

The ROC curve presented an AUC of 0.83, whereas the boxplot reflects a notable difference between the entropies of the not-mixed and mixed classes.

## 4 | DISCUSSION

In this study we presented a novel ML model for cloud classification based on GHI features. This was rendered possible thanks to the temporal matching of all-sky images with the measurements collected using the pyranometer, which allows the manual labels of the cloud types to be directly associated with the signals. Despite the potential



**FIGURE 6** Confusion matrix of the predicted value over ground truth for XGBoost on the test set. Ac, altocumulus; As, altostratus; Cb, cumulonimbus; Cc, cirrocumulus; Ci, cirrus; Cs, cirrostratus; Cu, cumulus; Ns, nimbostratus; Sc, stratocumulus; St, stratus. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

**TABLE 3** Precision, recall, and  $F_1$  score achieved by XGBoost on the test set.

Category	Precision	Recall	$F_1$
Clear-sky	1	0.98	0.99
Ci-Cs	0.89	0.87	0.88
Ac-Cc	0.67	0.7	0.68
Cu	0.75	0.82	0.78
Sc-St-As	0.7	0.64	0.67
Cb-Ns	0.94	1	0.97

Abbreviations: Ac, altocumulus; As, altostratus; Cb, cumulonimbus; Cc, cirrocumulus; Ci, cirrus; Cs, cirrostratus; Cu, cumulus; Ns, nimbostratus; Sc, stratocumulus; St, stratus.

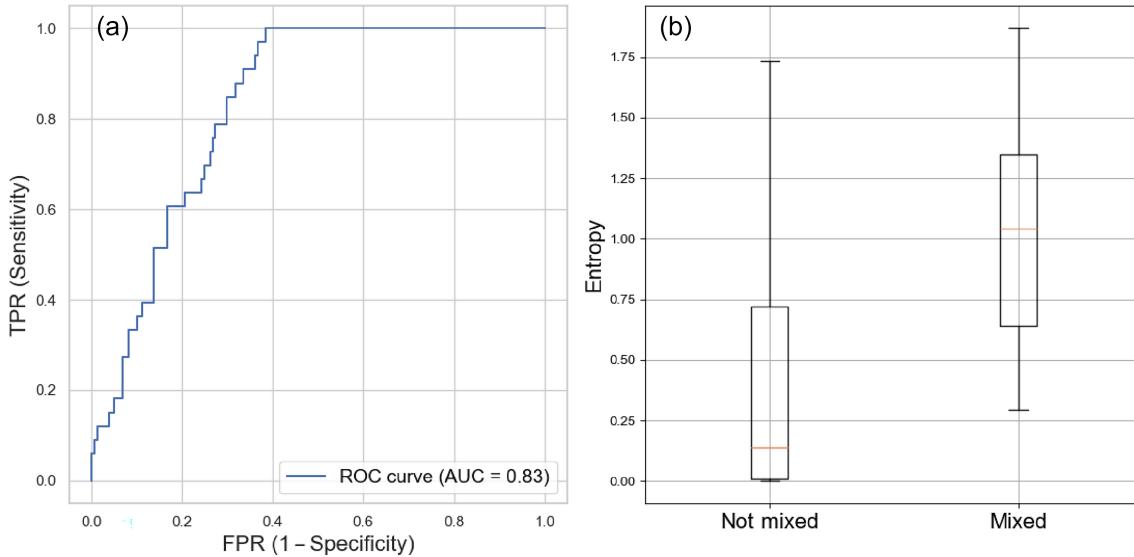
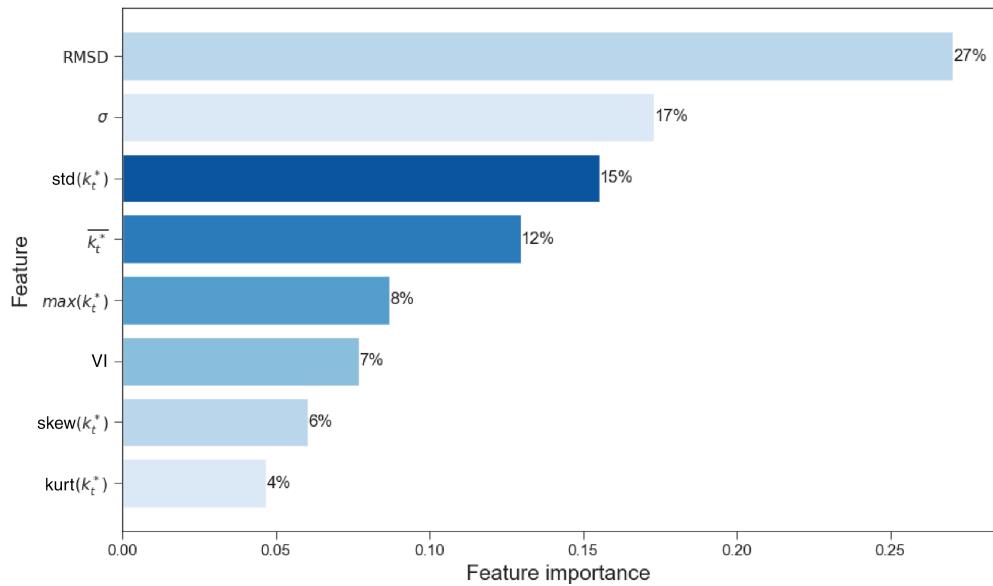
limitations associated with the subjective nature of human observations and variations in interpretation during the labeling process, the results demonstrate that our model is capable of predicting cloud behavior comparable to that of a meteorological observer from a sky image. With an achievement of 0.88 accuracy, our XGBoost model relies on a cost-effective, low-maintenance pyranometer signal and is fully automatic.

All models considered showed variations in performance depending on the size of the time window considered for analysis (see Figure 4). Although the persistence of specific cloud types in the sky, observed from a particular geographical location, is inherently variable owing

to factors such as cloud type, atmospheric conditions, and wind, determining a suitable time window was necessary for this study. Various researchers, such as Reno and Hansen (2016), opted for moving windows of 10 min, whereas others, such as Duchon and O’Malley (1999), chose a 21-min window. In our analysis, we explored time periods ranging from 9 to 41 min and found the best performance at 33 min. A trend in the performance was observed when the time window was increased. It also presented significant instability, which is mainly attributed to the fluctuations of the Sc-St-As class in its performance. This class was the most challenging to predict for the algorithm and its high unstable behavior in both nonlinear models can be attributed to its status as the most diverse class in the dataset. Notably, for the well-performing classes (Clear-sky, Cb-Ns, Ci-Cs), variations in  $\Delta t$  did not significantly impact their  $F_1$  score performance.

Experimentally, we observed that XGBoost and random forest performed much better than the logistic regression counterparts, with XGBoost exhibiting the largest accuracy and Cohen’s kappa values in the validation and test sets. This superior performance can be attributed to the complex nonlinear relationships within the data being effectively captured by the ensemble nature of tree-based models like XGBoost and random forest, allowing for a more accurate representation of the patterns. The literature supports our findings, with

**FIGURE 7** XGBoost feature importance. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



**FIGURE 8** (a) Receiver operating characteristic (ROC) curve illustrating the model's discrimination performance between mixed and not-mixed cloud scenarios with an area under the curve (AUC) value of 0.83. FPR: false-positive rate; TPR: true-positive rate. (b) Boxplot showing the distribution of entropy values between mixed and not-mixed classes. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

various studies highlighting the effectiveness of XGBoost in outperforming other models (Asselman *et al.*, 2023; Osman *et al.*, 2021; Kaushik *et al.*, 2023).

Once we confirmed XGBoost as the top ranked model, we proceeded to evaluate it on the test set. Our results showed the XGBoost outperformed other models in terms of accuracy and Cohen's kappa values, showing the agreement between predicted and actual class labels beyond what would occur by chance. The robustness of XGBoost played a pivotal role in its success. Though the differences with random forest may not be stark, the performance superiority of XGBoost, especially in comparison with logistic regression models, highlights its capacity to excel

in scenarios where nonlinear relationships and complex patterns play a crucial role.

Regarding feature importance analysis, the RMSD emerged as a crucial predictor, signifying those changes in this feature substantially affected the model's predictions. In particular, its importance value was almost seven times greater than the kurtosis of  $k_t^*$ , which represented the lowest feature importance value. The higher importance of RMSD would lie in its direct measurement of the high dispersion of GHI into the time window with respect to the clear-sky GHI. High short-term variability in GHI could be related to clouds with low horizontal developments or broken clouds such as Cu, which may

attenuate and subsequently increase the GHI above the  $GHI_{CS}$  (in cases of enhancement) in short periods of time (on the order of minutes), which also depends on the vertical development (Mol *et al.*, 2023). Similarly, the  $\sigma$  and standard deviation of the  $k_t^*$  also played significant roles, with importance values of 17% and 15% respectively, giving less weight to extreme cases. Unlike RMSD, both these features provide information about GHI variability relative to the clear-sky model or extraterrestrial irradiance pattern within the chosen time window. However, they assign less weight to extreme cases than RMSD does. Combined with the mean value of  $k_t^*$ , these features could offer valuable information to determine GHI patterns related to different cloud types. For example, clouds with low standard deviation and low mean  $k_t^*$  could explain the presence of homogeneous clouds with high horizontal extension with a relatively low vertical development such as St.

When discriminating the results per individual cloud class, we observed that the class with the highest confusion rates was Sc-St-As (see Figure 6). Though Sc and St are low-level clouds, As is characterized by being middle-level cloud. Thus, this class includes diverse cloud formations that manifest at different altitudes and displays distinct visual patterns. For example, whereas Sc clouds are flat, layered, and possess a smooth texture, the St clouds are recognized for their flat and grayish aspects, often spreading across expansive portions of the sky. Therefore, the optical properties among the three types included in this class are different, leading to different patterns that can affect the model's ability to group them into a single class.

The differences between clouds within the same class underscore the challenges in identifying features that accurately represent them. In this study, we adopted the classification from prior work; however, future research should consider cloud grouping when defining classes to mitigate such errors. One solution could be addressed by relabeling the dataset and associating these types with different classes. Moreover, the confusion between these three cloud types and Cu clouds could stem from the distortion present in the all-sky images. Cu clouds exhibit a heaped-up, puffy shape reminiscent of a cauliflower. However, the fisheye view from an all-sky camera introduces distortion and has systematic detection errors such as misdetection of thin clouds, leading to the potential misidentification of clouds during the labeling process by meteorological observers (Calbó *et al.*, 2005). It is important to note that this distortion affects all cloud types, but its effect is particularly pronounced for cumuliform clouds. In addition, when labelling from a two-dimensional all-sky image, meteorological observers lose the three-dimensional perspective they would have when determining cloud types through direct sky observation. This loss of perspective may hinder the identification

of cloud attributes such as height, which plays a key role in determining the types of clouds. As anticipated, different cloud types have distinct implications for the amount of irradiance reaching the Earth's surface. For instance, low-level clouds such as Sc-St, owing to their extensive coverage and high optical depth, tend to reflect a significant portion of incoming solar radiation back to space, thereby reducing the amount of irradiance at the surface (ISCCP, n.d.; Rossow & Schiffer, 1999; Wood, 2015). In addition, middle-level clouds, such as As, while also reflecting some radiation, can allow a larger fraction of irradiance to penetrate through, especially when they are thin or semi-transparent. Sc-St-As were grouped in the present study. However, it is evident that the impacts on GHI vary among these cloud types. This variation could be another potential cause of the poor performance observed in this class. Therefore, accurately grouping cloud types is a challenge for future research.

Except for those classes that are inherently more difficult to classify, our model demonstrates a good performance across most classes and did not show evidence of overfitting, even in classes with few samples in the test set. To further validate the model, we also performed cross-validation on the combined training and validation data and compared the results with our initial configuration. Despite a slight variation in the hyperparameters, the performance showed minimal differences, reinforcing our confidence in the model's generalization and stability.

Regarding the analysis of the ability to discriminate between mixed and not-mixed, we found the ROC curve (Figure 8a) shows that the entropy of the output probabilities has a high predictive value to determining the presence of multiple types of clouds simultaneously. This behavior is more evident in Figure 8b, where the differences between the mixed and not-mixed cloud sets are marked. The high AUC value of 0.83 demonstrates that the model exhibits high uncertainty on these scenarios, which is consistent with the behavior expected when analyzing the all-sky image. Through the preceding analysis, we can conclude that, when predicting a specific class, there is significant confidence that our model avoids any potential confusion with the mixed class.

Finally, this study should be considered in the context of its own limitations. In particular, the subjective nature of human observation and potential variations in interpretation during the labeling process contribute to discrepancies in the labeled ground truth. To address these challenges, it is crucial to refine the training of observers and standardize labeling procedures. Exploring alternative methods for labeling data, such as using lidar measurements to determine cloud base height (Lu *et al.*, 2021; Pal *et al.*, 1992) and introducing a cloud altitude classification, could also be instrumental in enhancing the accuracy

of cloud detection models. One detail to keep in mind is that, in this work, we used the same database for training the algorithms as in Loyber (2022), which labeled the data using the criterion of (Liu *et al.*, 2019), enabling the integration of this set with the Multimodal Ground-based Cloud Database (Liu, Li, *et al.*, 2020; Liu, Mei, *et al.*, 2020) to train more robust algorithms, as in (Liu *et al.*, 2019). In this sense, it is important to consider that the internal proportion of each cloud type in the target groups with which we work may affect the model's ability to discern more minority types. Nonetheless, our framework is sufficiently general to be applied to datasets with any type of labeling. Although the model is unable to distinguish between the 10 different types of clouds defined by the WMO, limiting some practical application, this development represents an advance in this sense by categorizing them into six broader categories. On the other hand, the classification of cloud types into single categories (e.g., grouping two cloud genera into one) could represent an oversimplification. Refining this division to consider the intrinsic characteristics of each cloud type is a promising direction for future work. Despite these limitations that restrict the applicability of the algorithm to problems such as detailed cloud type analysis, this development could contribute to areas such as operational meteorology, aviation, and solar energy management, among others.

Additionally, the number of Cb samples in the test and validation sets was very low after the partition. To address this issue, we implemented a novel criterion based on the presence of lightning, which allowed us to quickly include more samples in these evaluation sets. This additional criterion might affect the trade-off between variance and bias. However, since the training set did not include Cb samples collected using the lightning criterion, we did not expect the model to be affected by this selection procedure. Nevertheless, including these samples in the validation set might have introduced a certain bias, as the hyperparameters were calibrated using this set. We invite the community to validate and further explore this aspect using our model, which is publicly released at <https://github.com/anabelalusy/cloud-classification>.

Another limitation of the model could be the classification of high aerosol loading events (such as volcanic eruptions, biomass burning, among others). The algorithm may misidentify classes during these events because it was not specifically trained under such conditions. Furthermore, it is important to acknowledge the inherent differences between our pyranometer-based approach and human observer classification. The pyranometer method primarily considers clouds that intersect the sun's path, whereas human observers have a panoramic view of clouds across the entire sky. These differences underscore the complexity of our task in the labeling of data and the

need for continued consideration of multiple data sources in cloud-classification algorithms. Moreover, suboptimal performance when grouping heterogeneous cloud types indicates an area for enhancement in future research.

It is also worth mentioning that our model does not feature any engineering tools to control and prevent errors in outliers, meaning that it could suffer from typical problems of ML algorithms such as data drifting (Žliobaitė *et al.*, 2014). Although this might be prevented by, for example, providing to the user the Shannon entropy as a surrogate of the uncertainty of the model, we have experimentally observed a relationship between this metric and the presence of mixed types of clouds. One potential solution to avoid using an input sample that is actually an outlier might be to train an anomaly detection model, such as isolation forests (Liu *et al.*, 2008), using the original training data and apply it in test time on any new input sample. This would allow it to identify deviations from the original distribution of data used for training, therefore informing the user about the correctness of the output in an unsupervised setting. We encourage future research in this direction.

Finally, notice that all our experiments were conducted using data from a single meteorological station. Although our approach exhibited good performance for discriminating clouds under this setting, future research should be focused on studying if this behavior could be extrapolated to, for example, regions with similar climate conditions or across other meteorological and geographical settings.

## 5 | CONCLUSION

A low-cost ML model for cloud classification was developed using GHI time-series obtained from pyranometer measurements and a clear-sky model as input. This information was used to obtain the model features in a specific time interval. The classes were obtained by labeling all-sky images by meteorological observers. Linear and nonlinear models were trained, with experimental results showing that nonlinear models outperform linear models. Between the two nonlinear models evaluated, the XGBoost algorithm stands out as the best performer, achieving high efficiency reaching an accuracy of 0.88 and a Cohen's kappa of 0.84 for an optimal time window of 33 min. In particular, the algorithm presents an excellent performance to determine Clear-sky cases with accuracy and  $F_1$  score above 0.99. These results demonstrate that the characterization of the variability of GHI through the proposed features is related to the cloud classes; that is, the GHI is a valuable parameter to classify clouds.

This novel approach addresses the constraints faced by sites with limited equipment and provides a valuable solution for cloud classification using readily available

measurements. The ability of the algorithm to classify cloudiness based on irradiance measurements alone opens up possibilities for improved forecasting and analysis in areas where comprehensive meteorological data or experts are limited. In addition, our results pave the way towards combining pyranometer measurements with additional multimodal data—that is, the height of the cloud base as measured using lidar—to further improve classification results. We expect that, by combining these additional data sources with GHI measurements, our algorithm might achieve improvements, particularly for classes that have proven to be more challenging to predict from these signals.

Overall, owing to the critical considerations we made during the data preprocessing phase, our algorithm achieved good experimental results, positioning it as an efficient, cost-effective method for cloudiness classification and widespread availability at most meteorological sites. Our model is publicly released for research purposes at <https://github.com/anabelalusi/cloud-classification>.

## ACKNOWLEDGEMENTS

We would like to express our gratitude to the Japan International Cooperation Agency (JICA) and Japan Science and Technology Agency (JST) for their financial support for the Saver-Net project. Additionally, we thank institutions Servicio Meteorológico Nacional and DEILAP (CITEDEF-UNIDEF), for operating and maintaining the Saver-Net network. Special acknowledgments are extended to Raúl D'Elia for his technical support and to Pablo José Loyber for the information provided regarding the labeled dataset. We thank all meteorological observers who labeled the all-sky images: Victor Oviedo, Gustavo Gutierrez, Pablo Díaz, Nidia Garcia, and Hernán Palmisano, with special appreciation to Gustavo Gutierrez for his invaluable assistance in resolving queries and aiding in the interpretation of the images. We extend our gratitude to the World Wide Lightning Location Network (<http://wwlln.net>), a collaborative effort of numerous universities, institutes, and researchers worldwide, for providing the lightning location data utilized in this study. We kindly acknowledge support in the data analysis of electrical activities to Dr Gabriela Nicora and Dr Daiana Baissac, both experts in the field of geophysics. We also thank the reviewers and QJRMS Editor for their insightful comments that improved the article. Anabela Lusi is a graduate student in the Environmental and Health Applied Sciences Doctoral Program (DCAAS) at National University of the Center of Buenos Aires Province (UNCBA), Argentina.

## FUNDING INFORMATION

This research was supported by the Agencia I + D + i into the framework of the PICT-2020-SERIEA-026 01

project and by two CONICET PIP 2021–2023 grants (codes 11220200101558CO and 11220200102472CO).

## CONFLICT OF INTEREST STATEMENT

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

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

## ORCID

Anabela Rocío Lusi  <https://orcid.org/0000-0002-8528-2961>

## REFERENCES

- Abdellatif, Y., Alsalameh, A., Muslih, I. & Alshduifat, A. (2015) Cloud effect on power generation of grid connected small PV systems. *International Journal of Electrical and Computer Engineering*, 9(9), 1054–1059. Available from: <https://doi.org/10.5281/zenodo.1108671>
- Assegie, T.A., Sushma, S.J., Bhavya, B.G. & Padmashree, S. (2021) Correlation analysis for determining effective data in machine learning: detection of heart failure. *SN Computer Science*, 2, 213. Available from: <https://doi.org/10.1007/s42979-021-00617-5>
- Asselman, A., Khalidi, M. & Aammou, S. (2023) Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, 31, 3360–3379. Available from: <https://doi.org/10.1080/10494820.2021.1928235>
- Breiman, L. (2001) Random Forests. *Machine Learning*, 45, 5–32. Available from: <https://doi.org/10.1023/A:1010933404324>
- Calbó, J., Pages, D. & González, J.-A. (2005) Empirical studies of cloud effects on UV radiation: a review. *Reviews of Geophysics*, 43, RG2002. Available from: <https://doi.org/10.1029/2004RG000155>
- Chen, T. & Guestrin, C. (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD'16*. New York, NY: Association for Computing Machinery, pp. 785–794. Available from: <https://doi.org/10.1145/2939672.2939785>
- Deneke, H.M., Feijt, A.J. & Roebeling, R.A. (2008) Estimating surface solar irradiance from METEOSAT SEVIRI-derived cloud properties. *Remote Sensing of Environment*, 112, 3131–3141. Available from: <https://doi.org/10.1016/j.rse.2008.03.012>
- Deng, Y., Li, Z. & Song, Z. (2023) Attention scheme inspired softmax regression.
- Duchon, C.E. & O'Malley, M.S. (1999) Estimating cloud type from pyranometer observations. *Journal of Applied Meteorology and Climatology*, 38, 132–141. Available from: [https://doi.org/10.1175/1520-0450\(1999\)038<0132:ECTFPO>2.0.CO;2](https://doi.org/10.1175/1520-0450(1999)038<0132:ECTFPO>2.0.CO;2)
- Dürr, B. & Philipona, R. (2004) Automatic cloud amount detection by surface longwave downward radiation measurements. *Journal of Geophysical Research: Atmospheres*, 109, D05201. Available from: <https://doi.org/10.1029/2003JD004182>

- Escríg, H., Batlles, F.J., Alonso, J., Baena, F.M., Bosch, J.L., Salbidegoitia, I.B. et al. (2013) Cloud detection, classification and motion estimation using geostationary satellite imagery for cloud cover forecast. *Energy*, 55, 853–859. Available from: <https://doi.org/10.1016/j.energy.2013.01.054>
- Garcia-Gutierrez, L., Voyant, C., Notton, G. & Almorox, J. (2022) Evaluation and comparison of spatial clustering for solar irradiance time series. *Applied Sciences*, 12, 8529. Available from: <https://doi.org/10.3390/app12178529>
- Heinle, A., Macke, A. & Srivastav, A. (2010) Automatic cloud classification of whole sky images. *Atmospheric Measurement Techniques*, 3, 557–567. Available from: <https://doi.org/10.5194/amt-3-557-2010>
- Hoff, T.E. & Perez, R. (2010) Quantifying PV power output variability. *Solar Energy*, 84, 1782–1793. Available from: <https://doi.org/10.1016/j.solener.2010.07.003>
- Hollstein, A., Segl, K., Guanter, L., Brell, M. & Enesco, M. (2016) Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in Sentinel-2 MSI images. *Remote Sensing*, 8, 666. Available from: <https://doi.org/10.3390/rs8080666>
- Holzworth, R. (2023) World Wide Lightning Location Network. <http://wwlln.net/> [Accessed 24th September 2023]
- Osman, A.I., Najah Ahmed, A., Chow, M.F., Feng Huang, Y. & El-Shafie, A. (2021) Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal*, 12, 1545–1556. Available from: <https://doi.org/10.1016/j.asej.2020.11.011>
- ISCCP. Definition of Cloud Types, International Satellite Cloud Climatology Project. <https://isccp.giss.nasa.gov/cloudtypes.html> [Accessed 22nd November 2023]
- Jang, H.S., Bae, K.Y., Park, H.-S. & Sung, D.K. (2016) Solar power prediction based on satellite images and support vector machine. *IEEE Transactions on Sustainable Energy*, 7, 1255–1263. Available from: <https://doi.org/10.1109/TSTE.2016.2535466>
- Kaplan, J.O. & Lau, K.H.-K. (2021) The WGLC global gridded lightning climatology and time series. *Earth System Science Data*, 13, 3219–3237. Available from: <https://doi.org/10.5194/essd-13-3219-2021>
- Kassianov, E., Long, C. & Ovchinnikov, M. (2005) Cloud sky cover versus cloud fraction: whole-sky simulations and observations. *Journal of Applied Meteorology*, 44, 86–98. Available from: <https://doi.org/10.1175/JAM-2184.1>
- Kasten, F. & Young, A.T. (1989) Revised optical air mass tables and approximation formula. *Applied Optics*, 28, 4735–4738. Available from: <https://doi.org/10.1364/AO.28.004735>
- Kaushik, B., Vijayvargiya, A., Uppal, J. & Gupta, A. (2023) Comparative analysis of machine learning approaches for classifying erythema-squamous skin diseases. In: Kadry, S. & Prasath, R. (Eds.) *Mining intelligence and knowledge exploration, lecture notes in computer science*. Switzerland, Cham: Springer Nature, pp. 67–77. Available from: [https://doi.org/10.1007/978-3-031-44084-7\\_7](https://doi.org/10.1007/978-3-031-44084-7_7)
- Kokhanovsky, A. (2004) Optical properties of terrestrial clouds. *Earth Science Reviews*, 64, 189–241. Available from: [https://doi.org/10.1016/S0012-8252\(03\)00042-4](https://doi.org/10.1016/S0012-8252(03)00042-4)
- Lary, D., Zewdie, G., Liu, X., Wu, D., Levetin, E., Allee, R. et al. (2018) Machine learning applications for earth observation. *Earth Observation Open Science and Innovation*, 15, 165–218. Available from: [https://doi.org/10.1007/978-3-319-65633-5\\_8](https://doi.org/10.1007/978-3-319-65633-5_8)
- Lee, Y., Wahba, G. & Ackerman, S.A. (2004) Cloud classification of satellite radiance data by multiclass support vector machines. *Journal of Atmospheric and Oceanic Technology*, 21, 159–169. Available from: [https://doi.org/10.1175/1520-0426\(2004\)021<0159:CCOSRD>2.0.CO;2](https://doi.org/10.1175/1520-0426(2004)021<0159:CCOSRD>2.0.CO;2)
- Lefèvre, M., Oumbe, A., Blanc, P., Espinar, B., Gschwind, B., Qu, Z. et al. (2013) McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. *Atmospheric Measurement Techniques*, 6, 2403–2418. Available from: <https://doi.org/10.5194/amt-6-2403-2013>
- Li, Z., Hoiio, K. & Chang-Seng, W. (2023) Neural network-based identification of cloud types from ground-based images of cloud layers. *Applied Sciences*, 13, 4470. Available from: <https://doi.org/10.3390/app13074470>
- Liang, X. & Yuan, C. (2016) Derivation of 3D cloud animation from geostationary satellite images. *Multimedia Tools and Applications*, 75, 8217–8237. Available from: <https://doi.org/10.1007/s11042-015-2738-7>
- Liu, F.T., Ting, K.M. & Zhou, Z.-H. (2008) Isolation Forest. In: 2008 eighth IEEE international conference on data mining, Pisa, Italy, pp. 413–422. Available from: <https://doi.org/10.1109/ICDM.2008.17>
- Liu, M., Xia, X., Fu, D. & Zhang, J. (2021) Development and validation of machine-learning clear-sky detection method using 1-min irradiance data and sky imagers at a polluted suburban site, Xianghe. *Remote Sensing*, 13, 3763. Available from: <https://doi.org/10.3390/rs13183763>
- Liu, S., Duan, L., Zhang, Z. & Cao, X. (2019) Hierarchical multimodal fusion for ground-based cloud classification in weather station networks. *IEEE Access*, 7, 85688–85695. Available from: <https://doi.org/10.1109/ACCESS.2019.2926092>
- Liu, S., Li, M., Zhang, Z., Xiao, B. & Durrani, T.S. (2020) Multi-evidence and multi-modal fusion network for ground-based cloud recognition. *Remote Sensing*, 12, 464. Available from: <https://doi.org/10.3390/rs12030464>
- Liu, S., Mei, L., Zhong, Z., Xiaozhong, C. & Tariq, S. (2020) Ground-based cloud classification using task-based graph convolutional network. *Geophysical Research Letters*, 47, e2020GL087338. Available from: <https://doi.org/10.1029/2020GL087338>
- Long, C., Sabburg, J., Calbó, J. & Pages, D. (2006) Retrieving cloud characteristics from ground-based daytime color all-sky images. *Journal of Atmospheric and Oceanic Technology*, 23, 633–652. Available from: <https://doi.org/10.1175/JTECH1875.1>
- Loyer, P.J. (2022) *Entrenamiento de un modelo de IA para el procesamiento de imágenes todo cielo y clasificación de nubes*. Argentina: Instituto Tecnológico de Buenos Aires, ITBA.
- Lu, X., Mao, F., Rosenfeld, D., Zhu, Y., Pan, Z. & Gong, W. (2021) Satellite retrieval of cloud base height and geometric thickness of low-level cloud based on CALIPSO. *Atmospheric Chemistry and Physics*, 21, 11979–12003. Available from: <https://doi.org/10.5194/acp-21-11979-2021>
- Luiz, E.W., Martins, F.R., Costa, R.S. & Pereira, E.B. (2018) Comparison of methodologies for cloud cover estimation in Brazil – a case study. *Energy for Sustainable Development*, 43, 15–22. Available from: <https://doi.org/10.1016/j.esd.2017.12.001>
- Lusi, A., Orte, F., Alonso-Suárez, R., D'Elía, R. & Wolfram, E. (2023) Evaluación de los Modelos de Radiación Solar Global Heliosat-4 y McClear en dos Sitios de Argentina. *Anales (Asociación Física*

- Argentina)*, 34, 76–81. Available from: <https://doi.org/10.31527/analesafa.2023.34.3.76>
- Martínez-Chico, M., Batlles, F.J. & Bosch, J.L. (2011) Cloud classification in a mediterranean location using radiation data and sky images. *Energy*, 36, 4055–4062. Available from: <https://doi.org/10.1016/j.energy.2011.04.043>
- Marty, C. & Philipona, R. (2000) The clear-sky index to separate clear-sky from cloudy-sky situations in climate research. *Geophysical Research Letters*, 27, 2649–2652. Available from: <https://doi.org/10.1029/2000GL011743>
- Matuszko, D. (2012) Influence of the extent and genera of cloud cover on solar radiation intensity. *International Journal of Climatology*, 32, 2403–2414. Available from: <https://doi.org/10.1002/joc.2432>
- Mol, W.B., Van Stratum, B.J.H., Knap, W.H. & Van Heerwaarden, C.C. (2023) Reconciling observations of solar irradiance variability with cloud size distributions. *Journal of Geophysical Research: Atmospheres*, 128, e2022JD037894. Available from: <https://doi.org/10.1029/2022JD037894>
- Mommert, M. (2020) Cloud identification from all-sky camera data with machine learning. *Astronomy Journal*, 159, 178. Available from: <https://doi.org/10.3847/1538-3881/ab744f>
- Nespoli, A., Niccolai, A., Ogliari, E., Perego, G., Collino, E. & Ronzio, D. (2022) Machine learning techniques for solar irradiation nowcasting: cloud type classification forecast through satellite data and imagery. *Applied Energy*, 305, 117834. Available from: <https://doi.org/10.1016/j.apenergy.2021.117834>
- Ododo, J.C., Agbakwuru, J.A. & Ogbu, F.A. (1996) Correlation of solar radiation with cloud cover and relative sunshine duration. *Energy Conversion and Management*, 37, 1555–1559. Available from: [https://doi.org/10.1016/096-8904\(96\)86837-3](https://doi.org/10.1016/096-8904(96)86837-3)
- Ohmura, A., Dutton, E.G., Forgan, B., Fröhlich, C., Gilgen, H., Hegner, H. et al. (1998) Baseline surface radiation network (BSRN/WCRP): new precision radiometry for climate research. *Bulletin of the American Meteorological Society*, 79, 2115–2136. Available from: [https://doi.org/10.1175/1520-0477\(1998\)079<2115:BSRNBW>2.0.CO;2](https://doi.org/10.1175/1520-0477(1998)079<2115:BSRNBW>2.0.CO;2)
- Orsini, A., Tomasi, C., Calzolari, F., Nardino, M., Cacciari, A. & Georgiadis, T. (2002) Cloud cover classification through simultaneous ground-based measurements of solar and infrared radiation. *Atmospheric Research*, 61, 251–275. Available from: [https://doi.org/10.1016/S0169-8095\(02\)00003-0](https://doi.org/10.1016/S0169-8095(02)00003-0)
- Orte, F., Wolfram, E., Luccini, E., D'Elia, R., Lusi, A., Pallotta, J. et al. (2022) Red de Monitoreo de la Irradiancia Solar UV-Total en Argentina "Saver-Net". *Meteorológica*, 47, e016. Available from: <https://doi.org/10.24215/1850468Xe016>
- OTT HydroMet. (2023) CMP Series Pyranometers, Operational Manual. <https://s.campbellsci.com/documents>
- Pal, S.R., Steinbrecht, W. & Carswell, A.I. (1992) Automated method for lidar determination of cloud-base height and vertical extent. *Applied Optics*, 31, 1488–1494. Available from: <https://doi.org/10.1364/AO.31.001488>
- Perez, R. & Hoff, T.E. (2013) Chapter 6 – solar resource variability. In: Kleissl, J. (Ed.) *Solar energy forecasting and resource assessment*. Boston: Academic Press, pp. 133–148. Available from: <https://doi.org/10.1016/B978-0-12-397177-7.00006-1>
- Perez, R., Ineichen, P., Moore, K., Kmiecik, M., Chain, C., George, R. et al. (2002) A new operational model for satellite-derived irradiances: description and validation. *Solar Energy*, 73, 307–317. Available from: [https://doi.org/10.1016/S0038-092X\(02\)00122-6](https://doi.org/10.1016/S0038-092X(02)00122-6)
- Perez, R., Ineichen, P., Seals, R. & Zelenka, A. (1990) Making full use of the clearness index for parameterizing hourly insolation conditions. *Solar Energy*, 45, 111–114. Available from: [https://doi.org/10.1016/0038-092X\(90\)90036-C](https://doi.org/10.1016/0038-092X(90)90036-C)
- Pfister, G., McKenzie, R., Liley, J., Thomas, A., Forgan, B. & Long, C. (2003) Cloud coverage based on all-sky imaging and its impact on surface solar irradiance. *Journal of Applied Meteorology*, 42, 1421–1434. Available from: [https://doi.org/10.1175/1520-0450\(2003\)042<1421:CCBOAI>2.0.CO;2](https://doi.org/10.1175/1520-0450(2003)042<1421:CCBOAI>2.0.CO;2)
- Rawlins, F. (1982) A numerical study of thunderstorm electrification using a three dimensional model incorporating the ice phase. *Quarterly Journal of the Royal Meteorological Society*, 108, 779–800. Available from: <https://doi.org/10.1002/qj.49710845804>
- Reno, M.J. & Hansen, C.W. (2016) Identification of periods of clear sky irradiance in time series of GHI measurements. *Renewable Energy*, 90, 520–531. Available from: <https://doi.org/10.1016/j.renene.2015.12.031>
- Rossow, W.B. & Schiffer, R.A. (1999) Advances in understanding clouds from ISCCP. *Bulletin of the American Meteorological Society*, 80, 2261–2287. Available from: [https://doi.org/10.1175/1520-0477\(1999\)080<2261:AIUCFI>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<2261:AIUCFI>2.0.CO;2)
- Sabbagh, J.M. & Long, C.N. (2004) Improved sky imaging for studies of enhanced UV irradiance. *Atmospheric Chemistry and Physics*, 4, 2543–2552. Available from: <https://doi.org/10.5194/acp-4-2543-2004>
- Stein, J.S., Hansen, C. & Reno, M.J. (2012) The variability index: a new and novel metric for quantifying irradiance and PV output variability. Presented at the World Renewable Energy Forum, United States.
- Stephens, G.L. & L'Ecuyer, T. (2015) The Earth's energy balance. *Atmospheric Research*, 166, 195–203. Available from: <https://doi.org/10.1016/j.atmosres.2015.06.024>
- Sun, Z. & Shine, K.P. (1994) Studies of the radiative properties of ice and mixed-phase clouds. *Quarterly Journal of the Royal Meteorological Society*, 120, 111–137. Available from: <https://doi.org/10.1002/qj.49712051508>
- Tapakis, R. & Charalambides, A.G. (2013) Equipment and methodologies for cloud detection and classification: a review. *Solar Energy*, 95, 392–430. Available from: <https://doi.org/10.1016/j.solener.2012.11.015>
- Taravat, A., Del Frate, F., Cornaro, C. & Vergari, S. (2015) Neural networks and support vector machine algorithms for automatic cloud classification of whole-sky ground-based images. *IEEE Geoscience and Remote Sensing Letters*, 12, 666–670. Available from: <https://doi.org/10.1109/LGRS.2014.2356616>
- Utrillas, M.P., Marín, M.J., Estellés, V., Marcos, C., Freile, M.D., Gómez-Amo, J.L. et al. (2022) Comparison of cloud amounts retrieved with three automatic methods and visual observations. *Atmosphere*, 13, 937. Available from: <https://doi.org/10.3390/atmos13060937>
- Whitaker, T., Beranger, B. & Sisson, S.A. (2021) Logistic regression models for aggregated data. *Journal of Computational and Graphical Statistics*, 30, 1049–1067. Available from: <https://doi.org/10.1080/10618600.2021.1895816>
- WMO. (2017) Manual on the Observation of Clouds and Other Meteors – International Cloud Atlas. Annex I to the WMO Technical Regulations, WMO-No. 407, Basic Documents No. 2. <https://library.wmo.int/idurl/4/66297>
- WMO. (2018) Guide to Instruments and Methods of Observation, WMO-No. 8. <https://library.wmo.int/idurl/4/41650>

- Wood, R. (2015) CLOUDS AND FOG | Stratus and Stratocumulus. In: North, G.R., Pyle, J. & Zhang, F. (Eds.) *Encyclopedia of atmospheric sciences*, Second edition. Oxford: Academic Press, pp. 196–200. Available from: <https://doi.org/10.1016/B978-0-12-382225-3.00396-0>
- Zhang, M.H., Lin, W.Y., Klein, S.A., Bacmeister, J.T., Bony, S., Ced erwall, R.T. et al. (2005) Comparing clouds and their seasonal variations in 10 atmospheric general circulation models with satellite measurements. *Journal of Geophysical Research: Atmospheres*, 110, D15S02. Available from: <https://doi.org/10.1029/2004JD005021>
- Zhou, Y., Liu, Y., Wang, D., Liu, X. & Wang, Y. (2021) A review on global solar radiation prediction with machine learning models in a comprehensive perspective. *Energy Conversion and Management*, 235, 113960. Available from: <https://doi.org/10.1016/j.enconman.2021.113960>
- Zhu, W., Chen, T., Hou, B., Bian, C., Yu, A., Chen, L. et al. (2022) Classification of ground-based cloud images by improved combined convolutional network. *Applied Sciences*, 12, 1570. Available from: <https://doi.org/10.3390/app12031570>
- Žliobaitė, I., Bifet, A., Pfahringer, B. & Holmes, G. (2014) Active learning with drifting streaming data. *IEEE Transactions on Neural Networks and Learning Systems*, 25, 27–39. Available from: <https://doi.org/10.1109/TNNLS.2012.2236570>

**How to cite this article:** Lusi, A.R., Orte, P.F., Wolfram, E. & Orlando, J.I. (2024) Cloud classification through machine learning and global horizontal irradiance data analysis. *Quarterly Journal of the Royal Meteorological Society*, 1–17. Available from: <https://doi.org/10.1002/qj.4880>

## APPENDIX A

TABLE A1 Equations and references used for feature extraction.

	Equation form	Equation number	Reference
VI	$\frac{\sum_{k=2}^n \sqrt{(GHI_k - GHI_{k-1})^2 + \Delta t^2}}{\sum_{k=2}^n \sqrt{(GHI_{CS_k} - GHI_{CS_{k-1}})^2 + \Delta t^2}}$	(1)	Stein et al. (2012)
RMSD	$\sqrt{\frac{\sum_{i=1}^n (GHI_i - GHI_{CS})^2}{n}}$	(2)	
$k_t^*$	$\frac{k_t}{1.031 \times \exp\left[\frac{-1.4}{0.9 + (9.4/m)}\right] + 0.1}$	(3)	Perez et al. (1990)
$\sigma$	$\text{std}(\Delta k_t^*)$	(4)	Hoff and Perez (2010), Perez and Hoff (2013)