**Estimation of Cloud Fraction over the South-West Indian Ocean using BSRN Surface Observations**

Tina Herimino Andriantsalama1, Chao Tang1, Remy Ineza Mugenga1, Thierry Portafaix2, Mathieu Delsaut1, Alexandre Graillet1, Patrick Jeanty1, Girish Kumar Beeharry3, Roddy Lollchund3, Tyagaraja S. Modelly Cunden4, Lwidjy Baraka1, Béatrice Morel1

✉ Chao TANG

Email: [chao.tang@univ-reunion.fr](mailto:chao.tang@univ-reunion.fr)

1. ENERGY-lab, Université de La Réunion, La Réunion, France
2. LACy, Université de La Réunion, La Réunion, France
3. Department of Physics, Faculty of Science, University of Mauritius, Mauritius
4. Department of Electromechanical Engineering and Automation, Université des Mascareignes, Rose Hill, Mauritius

# Abstract

Cloudiness is the key to understand the variability of surface solar radiation, and to improve its predictability, which is particularly important for photovoltaic (PV) electricity production. This study estimates Cloud Fraction (CF) with observations from the BSRN (Baseline Solar Radiation Network) station at the island of La Réunion (55.5°E, 20.9°S) in the South-West Indian Ocean, using a XGBoost machine-learning model. Besides the statistical XGBoost model, a physical model known as the Automatic Partial Cloud Amount Detection Algorithm (APCADA) is also applied as a benchmark to estimate CF using LWD, surface air temperature and relative humidity data. CF data are obtained from collocated and coincident observations from a SkyCamVision all-sky camera within the UV-Indien network. After testing and validating the quality of the data with a set of quality control (QC) procedures derived from the BSRN recommendations, a XGBoost model is applied to estimate CF. A 2-year dataset (September 2019-February 2021) with a 5-min time step is used for this study, with 90% of it used for training and 10% for validation and test. After feature selections and model optimizations, a final score of 92.02% is achieved, which is comparable to the results from recent studies. Compared with APCADA, XGBoost performs better in terms of correlation (0.93 vs 0.85), Root Mean Square Error (0.12 vs 0.31) and Mean Absolute Error (0.08 vs 0.21). The statistical Machine Learning model, like XGBoost, could be potentially used to extend the temporal coverage of available CF observations, at a better accuracy compared to the traditional physical model such as APCADA.

Keywords: cloud fraction, longwave radiation, BSRN, XGBoost, Reunion Island, APCADA

# Abbreviation

1. Longwave Downward Radiation: lwd908
2. root mean square error (RMSE),
3. the mean absolute error (MBE),
4. and correlation efficient r (COR)

# Introduction

## Importance of cloud:

Clouds play an important role in Earth's energy balance by reflecting incoming solar radiation back to space and trapping outgoing longwave radiation from the Earth's surface. With the strong attenuation effect of Surface Solar Radiation (SSR), clouds can have a significant impact on the output of solar photovoltaic (PV) systems, and thus responsible for a large part of the intermittency of solar energy resource at various temporal and spatial scales ([Maimó-Far et al., 2020](#_ENREF_23)), which challenges one of the primary goals of electric utilities, which is to balance supply and demand.

Clouds also play an important role in the Earth's water cycle, as they transport and distribute water vapor around the globe. This can have important implications for the availability of water resources, as well as the frequency and intensity of precipitation events.

Understanding the formation, evolution of cloud and its complex meteorological and energetic effect of clouds is the fundamental for weather forecasting, climate change, and optimizing solar energy technologies, which calling the needs of cloud dataset of large spatial and long temporal coverage and of high quality.

## Cloud observation

### History of CF observation:

Cloudiness could be derived from human observation, meteorological satellites and surface remote observation and modeling. However, these methods have limitations that can affect the accuracy and precision of cloud cover estimation. For instance, human observation relies on the subjective judgment of the observer and irregular operating time (1 or 2 hours in daytime and much less during nighttime) and thus degrade the quality of the observation by high level of uncertainty. Meteorological satellites can complete the lack of objectivity of human observations, however, they either have a short comes of low spatial resolution (stationary satellite has typically 2km \* 2km) or low temporal resolution frequency (1 or 2 times a day, depending on the location) of the polar-orbiting satellites (such as MODIS). Additionally, ground-based remote sensing equipment includes ceilometers, lidar, and camera-based imagers ([Tapakis and Charalambides, 2013](#_ENREF_40)). Ground-based ceilometers and lidar can detect only in a fixed and limited part of the sky. Camera-based devices, on the other hand, have a number of useful benefits for detecting clouds, such as the observation area (coverage), image resolution, and observation period ([Kim et al., 2016](#_ENREF_17)). However, it is difficult to detect nighttime cloud with only visible channel information. Responding to the absence of visible light during nighttime, infrared sensors and filters have been developed to detect radiance from clouds (e.g., [Klebe et al., 2014](#_ENREF_18); [Wang et al., 2021](#_ENREF_42)). Nevertheless, they have the disadvantage of high equipment investment and extensive data processing. In the last, cloudiness could be simulated during day and night by cloud-resolving climate models ([Guichard and Couvreux, 2017](#_ENREF_14); [Satoh et al., 2019](#_ENREF_34)), however, in the cost of very expensive computational resources.

Therefore, many cloud cover detection algorithm from the proxy variables are developed. As an example a method of Automatic Partial Cloud Amount Detection Algorithm (APCADA) is developed by [Dürr and Philipona (2004)](#_ENREF_11) to estimate the cloud amount from longwave downward radiation (LWD), air temperature, and humidity, which could provide a simple and robust real-time cloud detection method available 24 hours a day. However, high clouds can barely be detected by this method, and the output values are discrete (in the unit of octas) but not continuous.

This study aims to improve the estimation of CF based on longwave radiation (LW) by applying a machine learning logarithm during daytime. This approach has the potential to be extended to nighttime and it can provide a more accurate and continuous estimation of CF compared to existing methods.

This paper is structured as follows: section presents the cloud climatology, the data used in this study and the methods employed to estimate CF. Then section is dedicated to the results and the comparison with reference method. Finally, conclusion and discussion are made in section

====================================================

To minimize human errors, ground- and aircraft-based measurements have been introduced to timely record 1-D cloud snapshots [Qian et al., 2012](#_ENREF_32). In the global observation trend, remote sensing has been applied to evaluate global cloud properties and estimate cloud cover parameters over years from visible and infrared radiations [Qian et al., 2012](#_ENREF_32).

Use of camera-based devices in cloud coverage observation improved spatiotemporal resolution, and image quality with a fish-eye lens (180o field of view), and ability to provide continuous valuable cloud datasets. Even though data quality depends on the capacity of the camera, cloud cover detection with sunlight showed great performance during the day.

Presence of long-term continuous ground-based measurements has attracted/created statistical and machine learning (ML) research interests. Forsythe

CF, short for fraction of area occupied by clouds, found to have a significant correlation with air temperature, precipitation, and topography [Didier, 2015](#_ENREF_7); [Forsythe et al., 2015](#_ENREF_13). After several studies of ML techniques in atmospheric radiative transfer studies, due to its ability to address nonlinear complexity problems, some studies were done in CF estimation (ref).

====================================================

# Context, data and methods

## context

Located in the South-West Indian Ocean, La Réunion island (21°S; 55.5°E) is mainly under an east-southeast trade wind flow ([Pous et al., 2014](#_ENREF_31)), which changes in direction according to the season: easterly in summer and more southeasterly in winter. The high altitude (~ 3 000 meters) of the island (Fig. 1), makes it an isolated and especially prominent obstacle to the atmospheric circulation in the SWIO. Hence, like most tropical islands with a marked topography, Reunion exhibits a strong climate contrast between its windward and leeward sides ([Mialhe et al., 2020](#_ENREF_24); [Morel et al., 2020](#_ENREF_27); [Réchou et al., 2019](#_ENREF_33); [Tang et al., 2023](#_ENREF_39)). The main cloud regimes over La Réunion are (a) orographic clouds generated by the local topography and (b) clouds caught on the island driven by synoptic systems ([Badosa et al., 2013](#_ENREF_2)). The vertical development of the clouds in these two regimes is limited by the thermal inversion, the height of which ranges from 1,700 to 3,100 m over Reunion, with an average value of ∼2,000 m (Bhugwant et al., 2000; Guilpart et al., 2017; [Réchou et al., 2019](#_ENREF_33)). Pauline .

The large variability of cloudiness over Reunion ([Badosa et al., 2013](#_ENREF_2); [Durand et al., 2021](#_ENREF_10); [Vérèmes et al., 2019](#_ENREF_41)) could challenge the development of solar energy, despite the high solar resource availability over the island ([Mialhe et al., 2020](#_ENREF_24);[Bessafi et al., 2018](#_ENREF_3); [Morel et al., 2021](#_ENREF_26)).

The BSRN station in the Saint-Denis campus of University of Reunion island (UR), our study of interest, is located close to the northern coast of the island. The diurnal and annual cycles of CF at the station are shown in Fig. 2 along with their respective data distributions within data variability. The lowest CF is found in the early morning, during 8 or 9 AM at Reunion local time. This is probably due to the dissimilation of cloud by solar heating at this time. And at seasonal scale, a mean CF about 0.6 is found from May to September (austral winter) and larger value about 0.7 from January to April (austral summer).

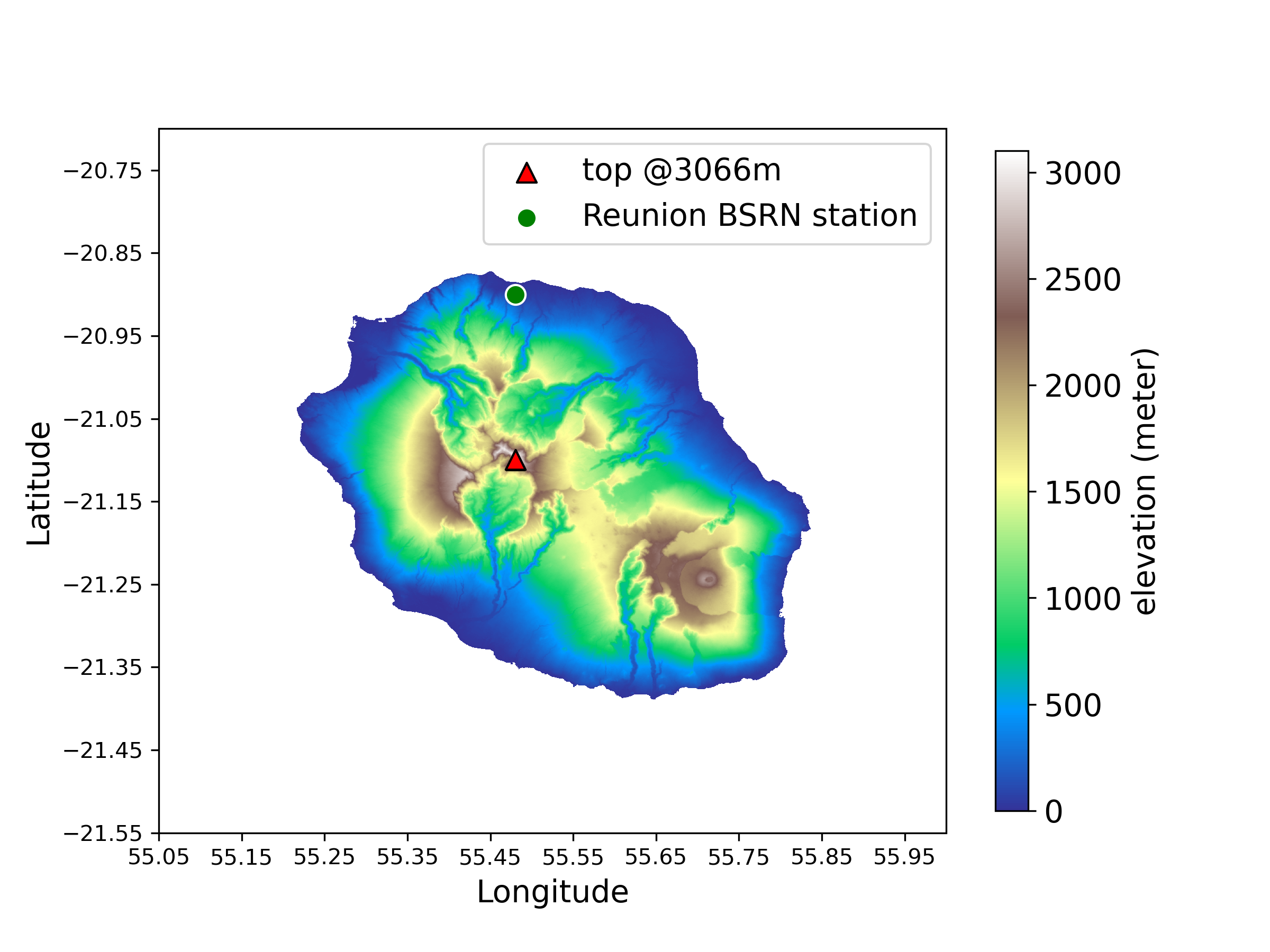


Fig. 1 Topography of Reunion. This map is based on the ASTER Global Digital Elevation Model from NASA Jet Propulsion Laboratory ([ASTER (2019)](#_ENREF_1). The highest summit, depicted by a red triangle on the map, is the inactive (since 20, 000 years ago) Piton des Neiges at 3066 m that forms the northwestern two thirds of the island. And the second highest summit is the very active Piton de La Fournaise volcano at 2560 m in the east. Between these two lays a 1500 m-high plateau. The Reunion BSRN station is located by in the northern coast as marked by the green dot.

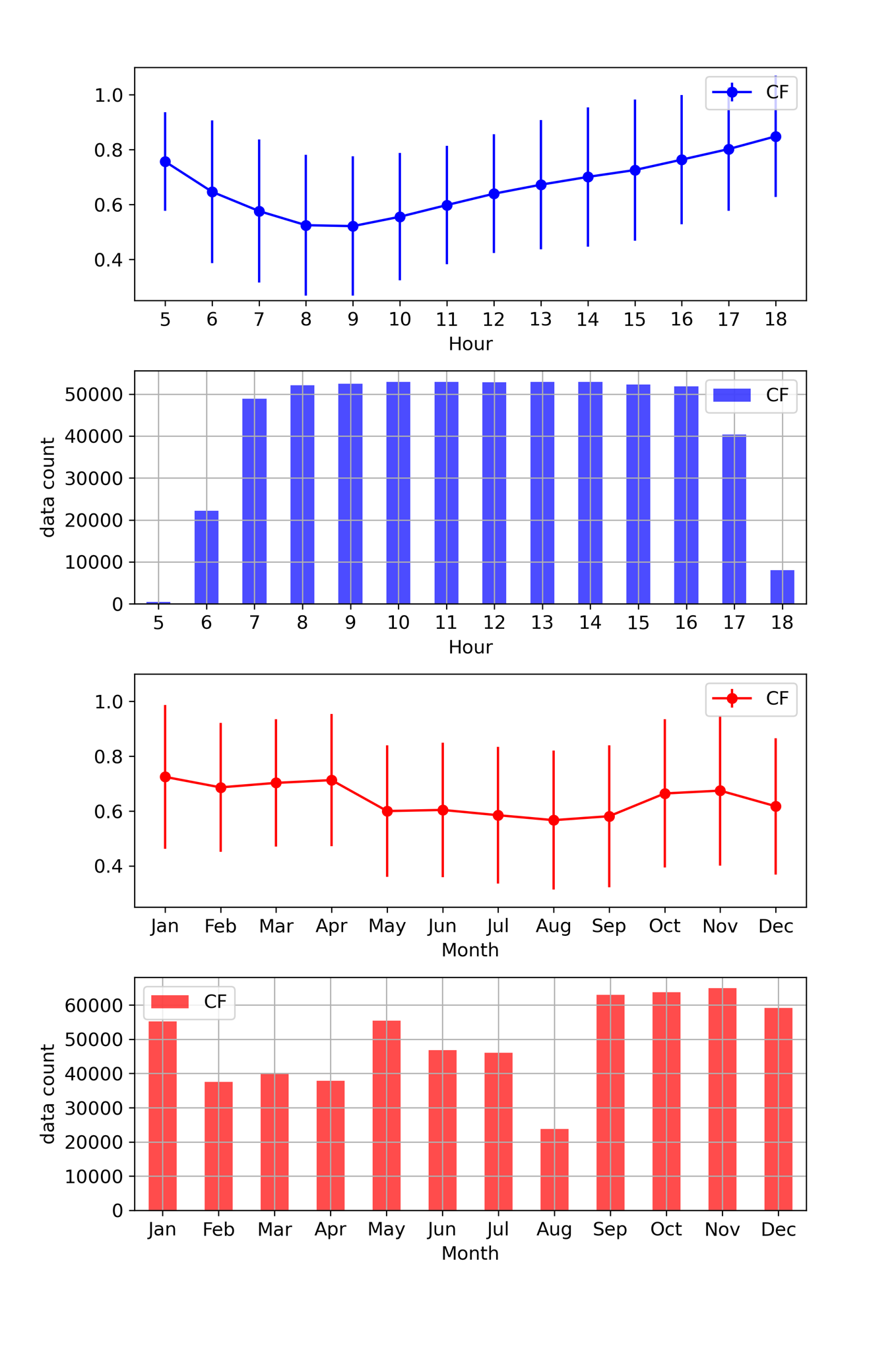


Fig. 2 Diurnal (1st plot) and annual cycle (3rd plot) of CF in Reunion LACy station. Data counts at frequency of one minute are below the respective diurnal and annual cycles. Data between 2019-09-13 and 2022-09-28, with missing values.

### LW correlation with CF

In 2013, while estimating daytime downward longwave radiation under both cloudy and clear skies, found a high correlation between CF and longwave.

Cloud are found important while estimating downward longwave radiation.

Several studies have been done estimating longwave with cloud fractions as input ([Carmona et al., 2014](#_ENREF_5); [Duarte et al., 2006](#_ENREF_9); [Heidinger and Cox, 1996](#_ENREF_15); [Silber et al., 2019](#_ENREF_36); [Sugita and Brutsaert, 1993](#_ENREF_38); [Yang and Cheng, 2020](#_ENREF_44); [Yeo et al., 2018](#_ENREF_45)), but no work done yet the other way around with ML.

* Price of LW : Carmona et al. 2014
* Cloud relationship with LWdn: Yang and Cheng 2020.

However, to our best knowledge, no previous study has dealt with CF estimation based on LW.

In this paper, we estimated CF with the Reunion new station of the Baseline Surface Radiation Network (BSRN). Two methods were used: XGBoost, a ML model that uses a tree and a sequential neural network, and a physical model: Automatic Partial Cloud Amount Detection Algorithm (APCADA). The results were compared by using Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and coefficient of correlation (r).

## Data

### Meteorological observation from Reunion BSRN-station

As predictors of local CF, several meteorological variables are taken from Reunion BSRN station, as one of the BSRN’s global network, characterized by high-quality measurements of surface solar radiation ([Ohmura et al., 1998](#_ENREF_29)). The station is managed by the ENERGY-lab and is located on the main campus of the University of Reunion Island in the city of Saint-Denis, situated on the northern coast of the island (as depicted in Fig. 1).

Several variables recorded at 1-minute interval are used in this study after passing an automatic ([Long and Shi, 2008](#_ENREF_22)) and visual quality control (QC). They includes: the global and diffuse horizontal shortwave irradiance (GSW and SWDif) measured by unshaded and shaded CMP22 pyranometer respectively, direct normal shortwave irradiance (SWDir) measured by a CHP11 pyrheliometer, downward longwave irradiance (LWD) by a CGR4 pyrgeometer and temperature (T), relative humidity (RH) and pressure (P) are measured by a WXT530 weather transmitter (Table 1).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Full name** | **unit** | **data source (device)** | **XGBoost** | **APACADA** |
| T | near surface air temperature | ° C | BSRN station | predictor | input |
| (WXT530 weather transmitter) |
| GSW | global horizontal shortwave irradiance | W/m2 | BSRN station | predictor | - |
| (unshaded CMP22 pyranometer) |
| SWDir | direct normal shortwave irradiance | W/m2 | BSRN station | predictor | - |
| (CHP11 pyrheliometer) |
| SWDif | diffuse horizontal shortwave irradiance | W/m2 | BSRN station | predictor | - |
| (shaded CMP22 pyranometer) |
| LWD | downward longwave irradiance | W/m2 | BSRN station | predictor | input |
| (CGR4 pyrgeometer) |
| RH | relative humidity | % | BSRN station | predictor | input |
| (WXT530 weather transmitter) |
| P | pressure | hPa | BSRN station | predictor | - |
| (WXT530 weather transmitter) |
| CF | cloud fraction | % | UV-Indien network | predictand | output |
| (allsky camera) |

Table 1 Summary of data used in this study. The temporal resolution of all these variables is 5-minute.

In addition to the QCs, the LWD data are also compared to the hourly satellite-based CERES (Clouds and Earth’s Radiant Energy Systems) product (i.e., SYN1deg-1Hour; [Smith et al., 2011](#_ENREF_37)), as shown in Fig. 2. The LWD of CERES is retrieved from the nearest pixel of its 1°x1° grid. Data in the period of June 2019 - February 2021 was used for the comparison at hourly, daily, and monthly timescales. Statistics, including the root mean square error (RMSE), the mean absolute error (MBE), and correlation efficient r (COR) are calculated.

As shown in Fig. 3 small errors and good correlations are found between LWD of the BSRN station and CERES. The corresponding statistics are comparable to previous works on CERES LWD data, including either the regional studies ([Yan et al., 2011](#_ENREF_43), [Sheng et al., 2009](#_ENREF_35), [dos Santos Nascimento et al., 2018](#_ENREF_8)) or the global ones (e.g., [NASA, 2021](#_ENREF_28)).

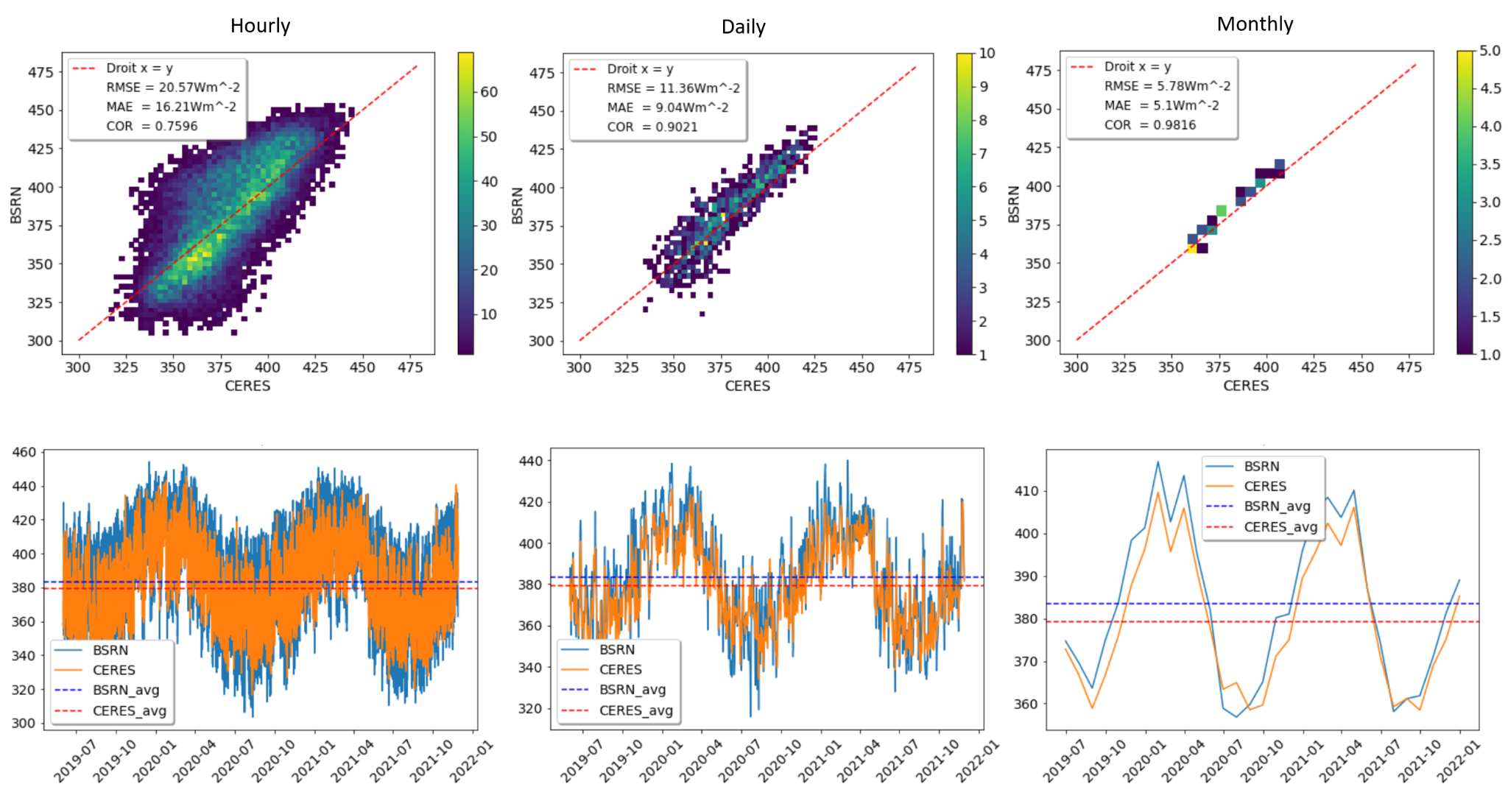


Fig. Comparison of LWD from Reunion BSRN station and CERES, at timescales of hourly, daily, and monthly timescales in the period of June 2019 - February 2022.

### CF measurements from the UV-Indien network

CF data are obtained from a SkyCamVision all-sky camera (<http://www.reuniwatt.com/>) within the UV-Indien network ([Lamy et al., 2021](#_ENREF_19)). This camera is located less than 100 meters from the Reunion BSRN station. It equipped with a fisheye-lens CMOS Sensor of 1600 x 1200 pixels resolution and acquires 1-minute hemispherical images in the visible range (380-440 nm). Image pixels are classified into clear sky; sun; thick cloud; or thin cloud with cloud segmentation algorithm. CF is computed from pixels with clear and cloudy sky, geometrically calibrated image ([Cadet et al., 2020](#_ENREF_4)). Due to the absence of infrared sensors on the camera, the BSRN station is unable to record cloud fraction (CF) observations during nighttime. And thus, only daytime CF observations in the period of 2019-09-13 to 2022-09-28 are used in this study. The training dataset comprises data from 2019-09-13 to 2021-09-12, whereas the testing dataset consists of two distinct periods of data (because of data availability): the summer months from 2021-10-01 to 2022-01-14, characterized by relatively higher cloud cover, and the winter months from 2022-05-01 to 2022-09-28, with relatively lower cloud cover (Fig. 2). Probability density function (PDF) and monthly count-bar of these two datasets are shown in Fig. 4. First, the PDF of training dataset, with missing data in 2021-07-17 – 2021-08-31 is quite similar to that of data in the period from 2019-09-13 to 2020-09-12 nearly without missing values (not shown), which implies the 1.5 months missing doesn’t change much of the annual cloud cover profile (PDF), partly due to the two-seasons annual cycle, with early 6 months for each season (Fig. 2). Then testing dataset is chosen to peak around 0.3 and 1.0 in its PDF, along with the two values of maximum occurrence (blue line in Fig. 4), thereby providing strong representative of the annual profile of CF.

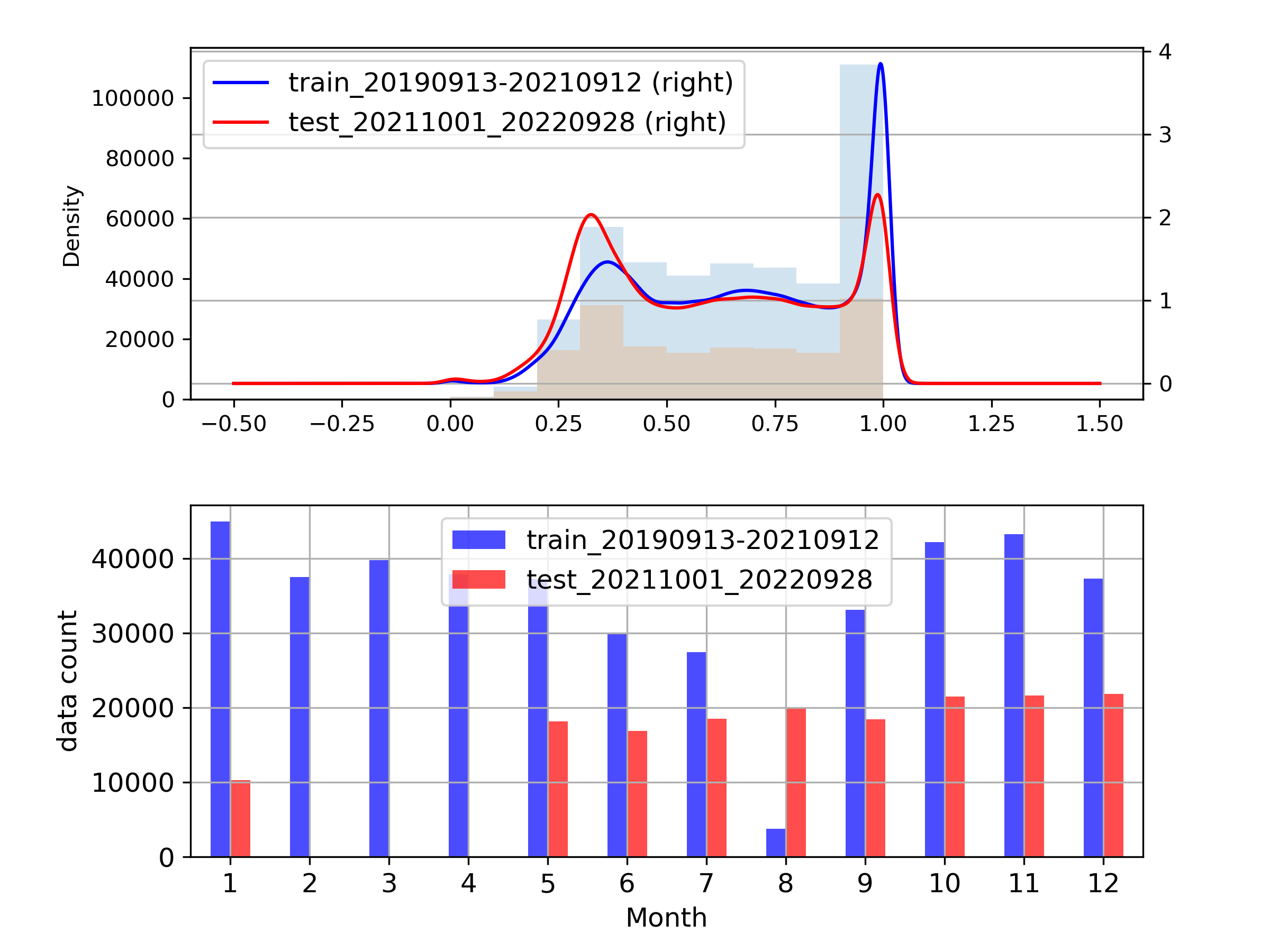


Fig. Probability density function (up plot) and monthly count-bar (bottom plot) of training dataset and testing dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| CF | LW down | Train | Test |
| 2019-09-13 – 2021-07-16 | 2019-06-present | 2019-09-13 – 2021-09-13 |  |
| 2021-07-17 – 2021-08-31 |  |  |
| 2021-09-01 – 2022-01-14 |  |  | 2021-10-01 - 2022-01-14 higher period |
| 2022-01-15 – 2022-04-31 |  |  |  |
| 2022-05-01 – 2022-09-28 |  |  | 2022-05-01 – 2022-08-31 lower period |

#### To-do: seasonal data test.

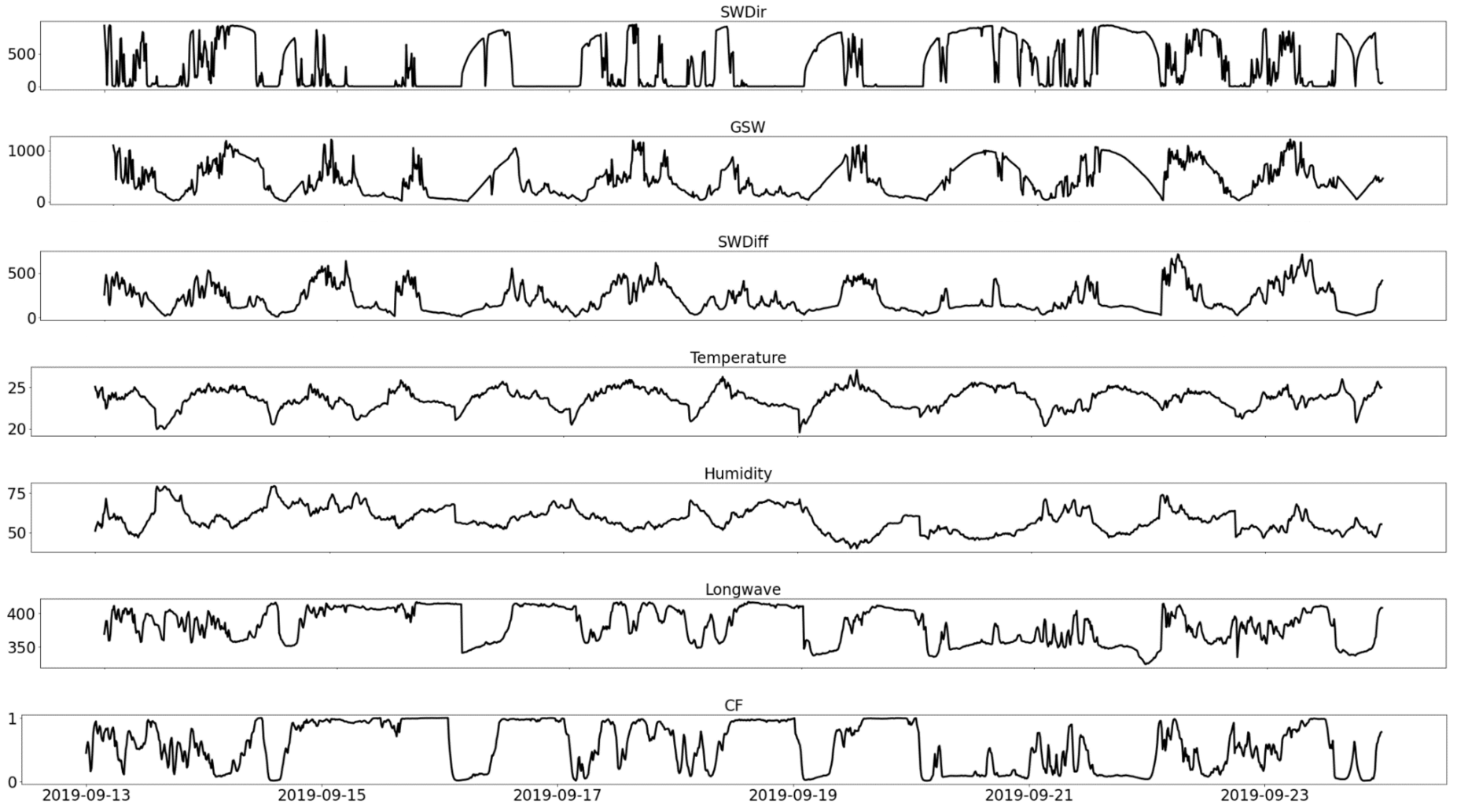


Fig. 5 Examples of one-week time series of the measurements used in this study.

## Automatic Partial Cloud Amount Detection Algorithm (APCADA)

A method of Automatic Partial Cloud Amount Detection Algorithm (APCADA) is developed by [Dürr and Philipona (2004)](#_ENREF_11) to estimate the cloud amount without high clouds from longwave downward radiation (LWD), air temperature, and humidity.

The determination of partial cloud amount according to APCADA is based on two parameters: the cloud-free index (CFI) and the variability of longwave downward radiation (STD LWD). The CFI, a ratio of apparent emittance of the sky to the empirical apparent cloud free emittance, is used to distinguish between clear or cloudy conditions, calculated as:

Eq. 1

where is the Stefan–Boltzmann constant, is the air temperature in Kelvin, and the emissivity of a cloud-free sky, which is defined as:

Eq. 2

where is altitude-dependent emittance of dry atmosphere, is a location-dependent coefficient, is air temperature (in K), and is water vapor pressure (Pa). In this study the emissivity of a cloud-free sky, i.e., is defined by fitting Eq. 2 in the clearsky condition selected by the method of [Long and Ackerman (2000)](#_ENREF_21) in Reunion during the period of September 2019 - February 2021 (Fig. 4).

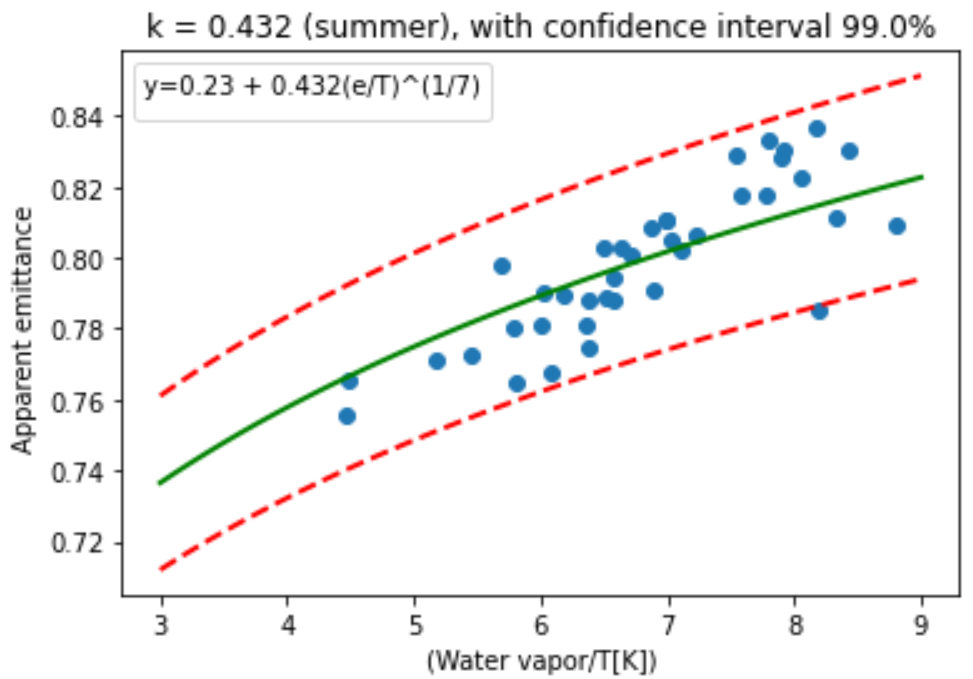


Fig. 6 Apparent emittance vs ratio of water vapor pressure and temperature at Reunion during the period of September 2019 - February 2021.

According to [Dürr and Philipona (2004)](#_ENREF_11), besides CFI, variability of LWD is also used, as it allows the distinction between cloud fraction types: broken clouds strongly influence the variability signal, while overcast and cloudless skies lead to a low variability. Since LWD measured at the Earth’s surface is marginally affected by high clouds because of large distance and cold emittance temperature, as a drawback, APCADA can detect only total cloud amounts without high clouds (hereinafter referred to as partial cloud amount, PCA). A dataset on frequency of 5-min from September 2019 to February 2021 is used in this study, according to the availability of CF measurements.

## Machine learning method: the XGBoost model

Extreme gradient boosting (XGBoost) regression is an ensemble machine learning algorithm widely used in data mining with excellent performance ([Chen et al., 2019](#_ENREF_6)). In contrast to some other machine learning models such as RF, XGBoost has a more complex structure and the introduced regularization items in loss function to control against overfitting so that it can better handle complex data. As a boosting method, XGBoost proceeds by iteratively adding new trees that predict the residuals or errors of prior trees, then all trees are combined to make the final prediction.

### Variable selection

See:

Math of XGBoost see: [Zhang et al., 2018](#_ENREF_46).

In this study, the XGBoost is applied on a 2-year dataset (September 2019-February 2022) with a 5-min temporal resolution, where 90% for training and 10% for validation and test. The predictors in the model are then the LWD and SWD fluxes and the additional meteorological parameters measured at the BSRN station (Table 1), and the predictand is the CF measurements from the UV-Indien network. Cross validation is performed for training. Learning curves based on RMSE are used to evaluate the model during the training and the validation.

see [Fan et al., 2020](#_ENREF_12) and [Liu et al., 2022](#_ENREF_20).

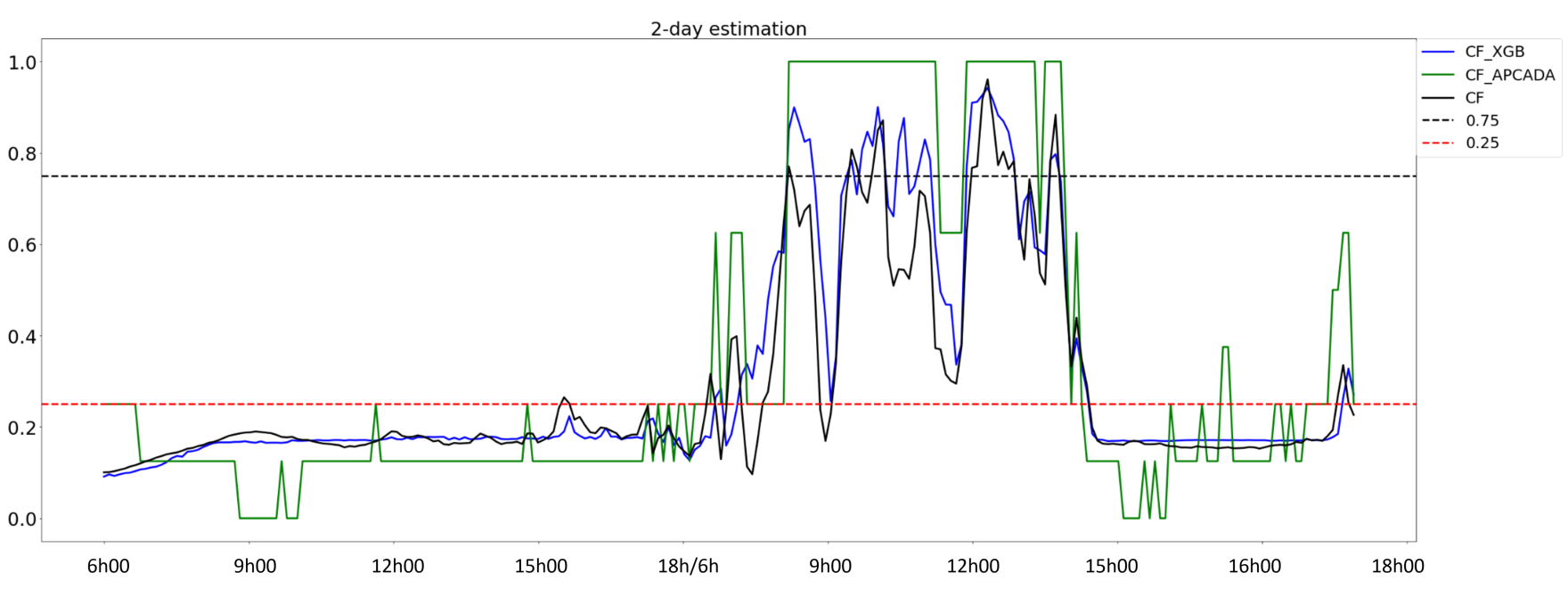
## Error metrics

To evaluate the performance of XGBoost, APCADA, root mean square error (RMSE), mean absolute error (MAE), and correlation coefficient (r) were calculated.

# Results

## APCADA performance

APCADA-estimated PCA (in octas) was first converted to CF (continuous values) and then compared with observations in Reunion from September 2019 to February 2021. APCADA was able to estimate dynamics of the observed CF, a correlation of 0.84 is achieved, with RMSE of 0.23 and MAE of 0.31.



|  |  |  |  |
| --- | --- | --- | --- |
| Model \ level of CF | 0.00 - 0.25 | 0.25 - 0.75 | 0.75 - 1.00 |
| XGBoost | 0.08 | 0.13 | 0.14 |
| APCADA | 0.22 | 0.31 | 0.14 |

Table 2 Model performance in estimating CF different levels (low, medium, and high level of CF)

Comment: All values are greater than 80%, except for KWA, where the minimum score is about 70%.The lowest score rates are found for KWA ([Table 5](https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2003JD004182#jgrd11026-tbl-0005)). Investigations showed that the percentage of broken PCA (2–6 octas) in KWA is over 60% compared to about 30% in PAY, which decreases the precision of APCADA. This is a small island next to Australia, [link](https://www.google.com/maps/place/Kwajalein+Atoll,+RMI/@8.716688,167.698314,13z/data=!4m5!3m4!1s0x65aeb9f09b73c9a5:0xacb7ba68a691d2a3!8m2!3d8.716667!4d167.733333?hl=en-GB)

## XGBoost performance

See turning of XGBoost in [Mo et al., 2019](#_ENREF_25).

### Variable selection

see [Huang et al., 2021](#_ENREF_16)

The variable selection step is important in constructing machine learning models. We used the random forest algorithm to select data variables (Zeng et al., 2020). Normalized daily data were used to construct and train the random forest model and to calculate the model’s importance. The data preprocessing experiment was intended to verify the importance of variables in a given model and to analyze the impact of changes in the variables on the model’s predictive performance. The experiment proceeded as follows:

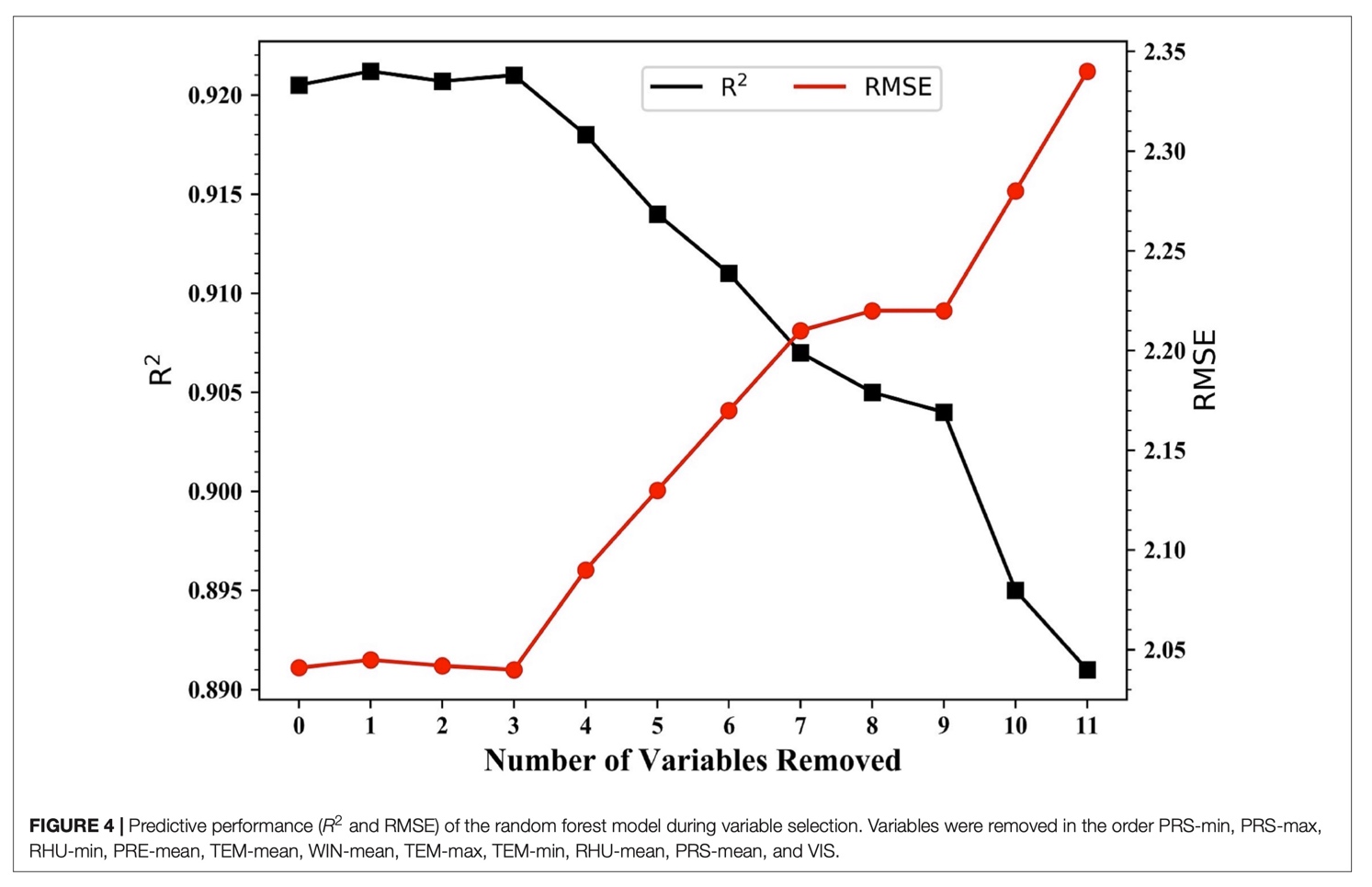
(1) divide the dataset into a training set and test set after completing the data quality control process;

(2) use the training set to train and save the model, then calculate the correlation coefficient (*R*2) and the root mean square error (RMSE) of the saved model;

(3) based on the order of importance of the variables in the model, eliminate the least important variable;

(4) repeat steps(2) and (3) until only two variables remain (the minimum required for calculation).

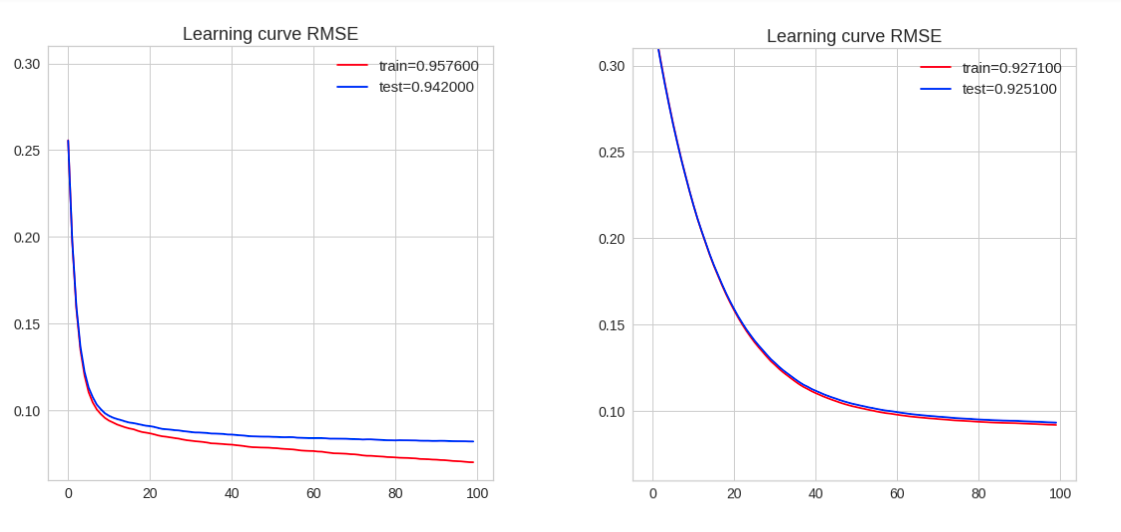
**Figure 4** shows that when the model contained <10 variables, *R*2 tended to decrease and the RMSE tended to increase. Between 12 and 10 variables, *R*2 reached 0.921 and the RMSE was 2.042 MJ/m2. With four variables, *R*2 decreased sharply from 0.904 to 0.895 and the RMSE decreased from 2.19 to 2.28 MJ/m2. Therefore, the prediction of solar radiation can achieve the best performance when using 10 variables, then the subsequent model experiments were trained with these 10 variables.



The first model fitting with the XGBoost default parameter values returns a model with an overfitting issue. To overcome this problem, several tests were performed with different optimization methods such as Principal Component Analysis (PCA), Recursive Feature Elimination (RFE) and feature importance. Each applied method was followed by tuning of hyperparameters. Results after PCA are not satisfactory as the model performance degrades. In contrast, in the case of RFE, results in three of the tests performed show an increase in the model’s performance. Finally, feature importance gives results similar to the best results with RFE. The figure (\ref3)) shows learning curves before and after tuning hyperparameters. And the table (\ref2)) shows optimal hyperparameters.

Chao: I suppose in the table you will show the final optimized parameters. while the optimization is not ONLY to handle overfitting. The final purpose of ML is generalization, there should be other considerations when turning the hyperparameters. So better to make people think the turning is not only for overfitting.

After feature selections and model optimizations, a final score of 92.02% is achieved, which is comparable to the results from recent studies ([Dürr and Philipona, 2004](#_ENREF_11)) do we have other similar studies). Figure (\ref5)) shows CF estimated time series with observations, with RMSE and MAE of 0.12 and 0.08 respectively (Table(\ref2))). We find that the surface pressure is not necessary for this application since it makes the XGBoost model more complex, Figure(\ref4)).



(a) (b)

Fig 3: Learning curve (RMSE) before (a) and (b) after tuning hyper-parameters of XGBoost model.

Table 2: XGBoost model optimized hyperparameters

|  |  |
| --- | --- |
| Hyper-parameter | Values |
| N estimators | 100 |
| Learning rate | 0.05 |
| Max depth | 6 |
| Subsample | 1 |
| Colsample bylevel | 0.8 |
| Colsample bytree | 1.0 |
| Colsample bynodes | 0.8 |
| Min child weight | 4 |

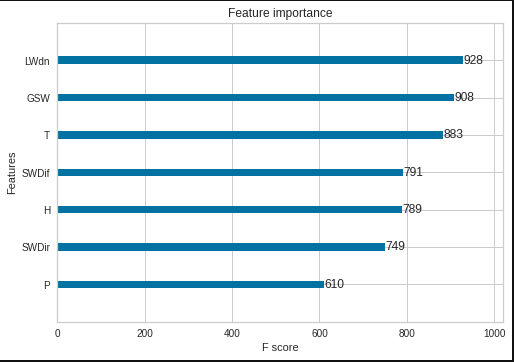


Fig 4: F-score of XGBoost model features to CF estimation.

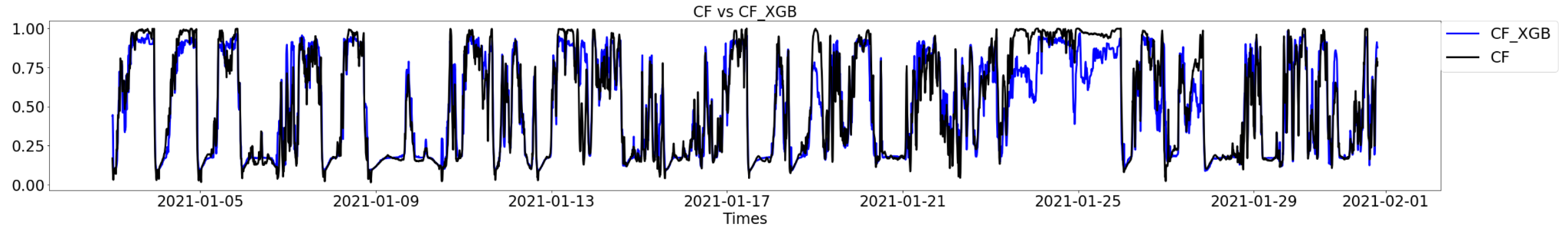


Fig 5: XGBoost CF estimation and observations at Reunion for January 2021 - February 2021 period. (plot modification)

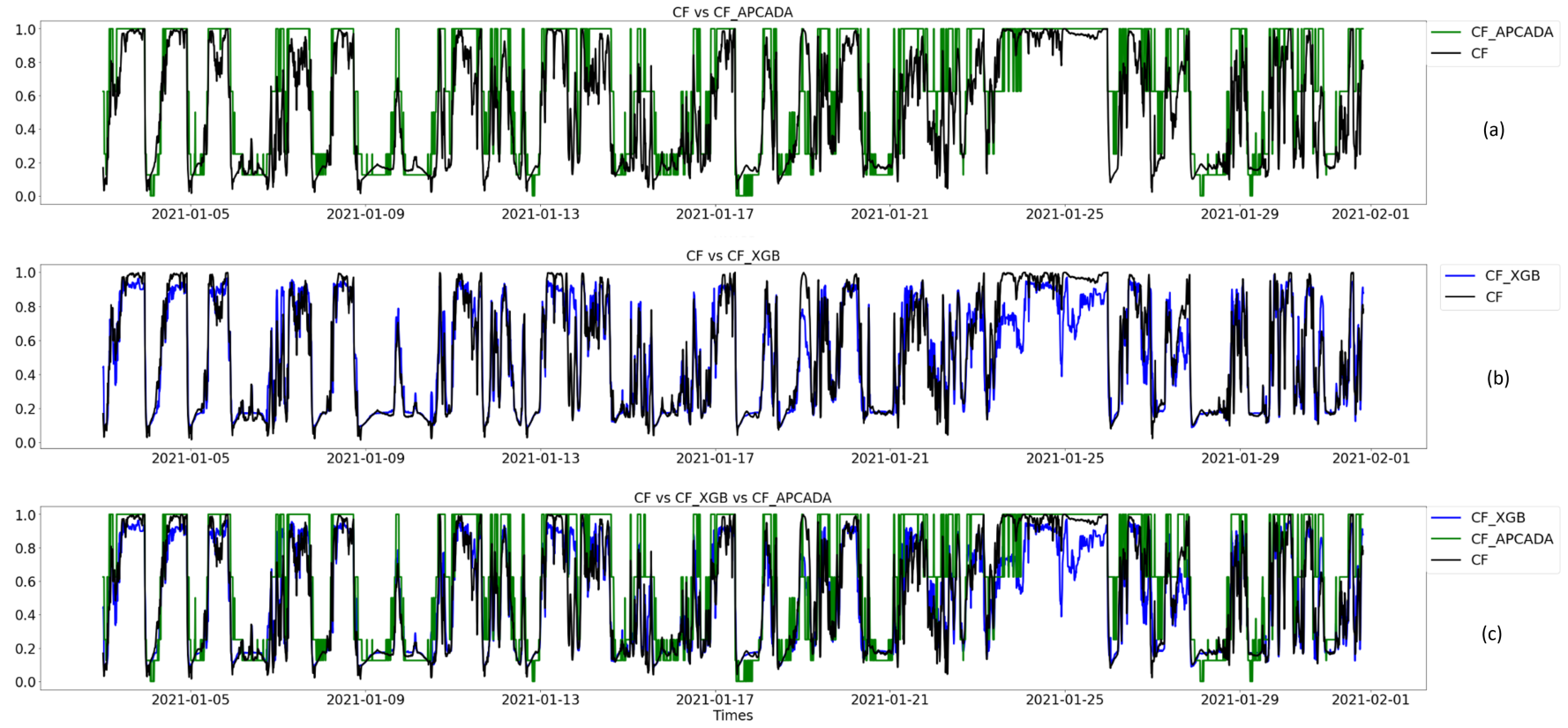


Fig 6: Comparison of observed and estimated CF, (a) observation, (b) XGBoost, and (c) APCADA together with CFI.

## CF estimations

Statistical and physical methods were compared, Table (\ref2), and \ref3)). The correlation of 85.9% was observed between XGBoost and APCADA estimations. In addition, RMSE and MAE of 0.27 and 0.21 were observed between the two methods.

Chao: is the r Pearson correlation? if yes, 1) usually, the correlation coefficient varies between +1 and -1. 2) The Pearson correlation is less appreciated if the two variables (CF\_APCADA, converted from octas for example) are not gaussian, and need to be checked. 3) it’s better that we keep the consistency of the precision of the matrix (r, RMSE, etc) 1 -> 1.00, 0.931 -> 0.93, that’s important but not easy to explain by text here.

Table 3: Correlation matrix of the estimated CF by XGBoost, and APCADA with observed CF.

|  |  |  |  |
| --- | --- | --- | --- |
| r | CF | CF\_XGB | CF\_APCADA |
| CF | 1.00 | 0.93 | 0.84 |
| CF\_XGB | 0.93 | 1. | 0.89 |
| CF\_APCADA | 0.84 | 0.89 | 1. |

Table 4: Error metrics between CF estimations and observation

|  |  |  |
| --- | --- | --- |
| Estimations | RMSE | MAE |
| CF vs CF\_XGB | 0.12 | 0.08 |
| CF vs CF\_APACADA | 0.23 | 0.16 |
| CF\_XGB vs CF\_APCADA | 0.21 | 0.16 |

Discussion and perspectives

This study shows the possibility to estimate cloud fraction by a machine learning method from meteorological variables, where the longwave radiation plays a crucial role because of its strong relationship with clouds. The XGBoost model is successfully applied to estimate CF by LW radiation (along with other meteorological variables). It could be better to highlight the role of LW for the XGBoost model, we could add this later. The performance of XGBoost is proven superior to the traditional physical empirical method (e.g., APCADA). While it is worth noting that, XGBoost estimates CF in continuous values, whereas APCADA produces discrete data, which may potentially impact the statistical comparison of these two methods. High-quality and abundant CF measurements is the key to apply a Machine Learning approach (e.g., XGBoost), a future application could benefit from the increasing availability of CF data.

Only daytime CF, from the visible camera, is used to train the XGBoost model in this study. Similar experiments could be done during the night with nighttime CF measurement such as that from an infrared camera. With performed accuracy compared to traditional physical methods, the XGBoost model could be applied to extend the temporal coverage of existing CF observations from about 1.5 year in this study to more than 2.5 years with available meteorological variables.

## Limitations of APCADA at Reunion

Presence of dynamic low-level clouds in ocean regions mainly from the near surface moisture hindered the performance of APCADA. For the same reason, a correlation of below 70% was observed at Kwajalein (8.43°N, 167.44°E) (durr2004automatic)). other information on this?

* Underestimation of lower cloud fraction
* few clear sky days over the region → dynamics of cloud formation over oceanic region
* number of clear sky days tends to decrease from morning to evening
* data dependence estimation

## Limitation of XGBoost model

Chao, I don’t have an idea on this part

* underestimating high values, [link](https://www.mdpi.com/2072-4292/13/9/1848)
* Overestimation?
* huge amount of data dependence

# Conclusion

nd its representation in Numerical Weather Prediction (NWP) models still has a long way to go (heidinger1996finite)). The current uncertainties lie from misrepresentation of brokenness (inhomogeneity) of clouds in atmospheric radiative transfer calculations, which has been studied over 20 years for solar radiation in general and longwave radiation in particular (aida1977reflection), and ellingson1982effects)).

In addition to emissions, cloud brokenness effects to longwave radiation consider the side face shading of the broken cloud blocks. (more information on CF in NWP \*11 May\*)

Even though clouds play a key role in the Earth's radiation budget, there is still a gap in cloud fraction (CF) estimation for renewable energy (RE), photovoltaic (PV) production in particular. Different studies showed a strong radiative cooling effect from low level clouds, and negligible effect from the high clouds corresponding to the distance and emittance temperature ([Dürr and Philipona, 2004](#_ENREF_11), [Park and Shin, 2019](#_ENREF_30)). Presence of these low-level clouds, with abundant water droplets, is high in ocean regions mainly from near surface moisture. In addition to cloud vertical structure, and cloud optical depth, CF as a fraction of cloud occupied area has been used as a measure of cloud quantification and in the Earth’s energy budget ([Qian et al., 2012](#_ENREF_32)), and [Zhang et al., 2011](#_ENREF_47)). With the rise of CF measurement, different CF estimation methodologies have been developed.

As CF ground-based measurement is getting to maturity stage, radiation transfer uncertainties in general circulation models (GCMs), and physical models have started being addressed. Accuracy of long-term ground-based measurements have been hindering CF simulation and estimation with GCMs and physical models (qian2012evaluation)). In different regions of the world, use of both ground-based cameras and aircraft measurements produces 1-D cloud snapshots. For global observations, with remote sensing, global cloud properties and cloud cover parameters have been developed over years from visible and infrared radiations ([Qian et al., 2012](#_ENREF_32)). With increase of data, statistical studies have been conducted and found significant correlations of CF with precipitation, air temperature, and

With recent technology of using aerosol lidar and thermographic cameras, continuous valuable cloud datasets are being produced and welcoming machine learning methods.

Estimation of the cloud brokenness has been using different approximation methods, which requires a significant computational effort. Statistical methods - correlation analysis - have been used on a large amount whereas significant correlations have been found between cloud fraction and precipitation, air temperature, and topography (didier2015comparison), and forsythe2015detailed)). With the recent rise of machine learning and availability of data, different studies showed a large correlation between cloud fraction and longwave radiation while estimating the downward longwave (ref). (more information on relationship of CF with other variables -physically and statistically \*12 May\*)

have been used to evaluate the results, and comparison of the longwave data with Clouds and the Earth’s Radiant Energy System (CERES) data. And to evaluate the machine learning models, learning curves were used.

Before using the model, we need to check if the model work well. By using leaning curve we can evaluate the performance of the model. Learning curve show 2 curves, one is for training and the other for validation. Overall, during the training both curves are supposed to decrease together and the generalization gap between both curves is necessary low. Generalization gap is the difference between both curves. Three cases can be appears: over-fitting, under-fitting and good fitting. Over-fitting mean when the model is good during the training, but bad for the validation. Under-fitting mean the model doing great for the validation and bad for the training. Good fitting when the generalization gap is lower and both curves are nearly overlapping. The tuning hyper-parameters are necessary to solve over-fitting or under-fitting issues.

The next figures show the learning curves of the model before and after tuning hyper-parameteres:

Feature importance is an usefull method used to enhence a model machine learning, it allows to select the most important feature to the model. The feature importance result is presented to the following figure :

Taking into account the result of feture iumportance, several test have been carried out to improve the performance of the model. The results are shown in the following figures :

Result about CF estimation with XGBoost

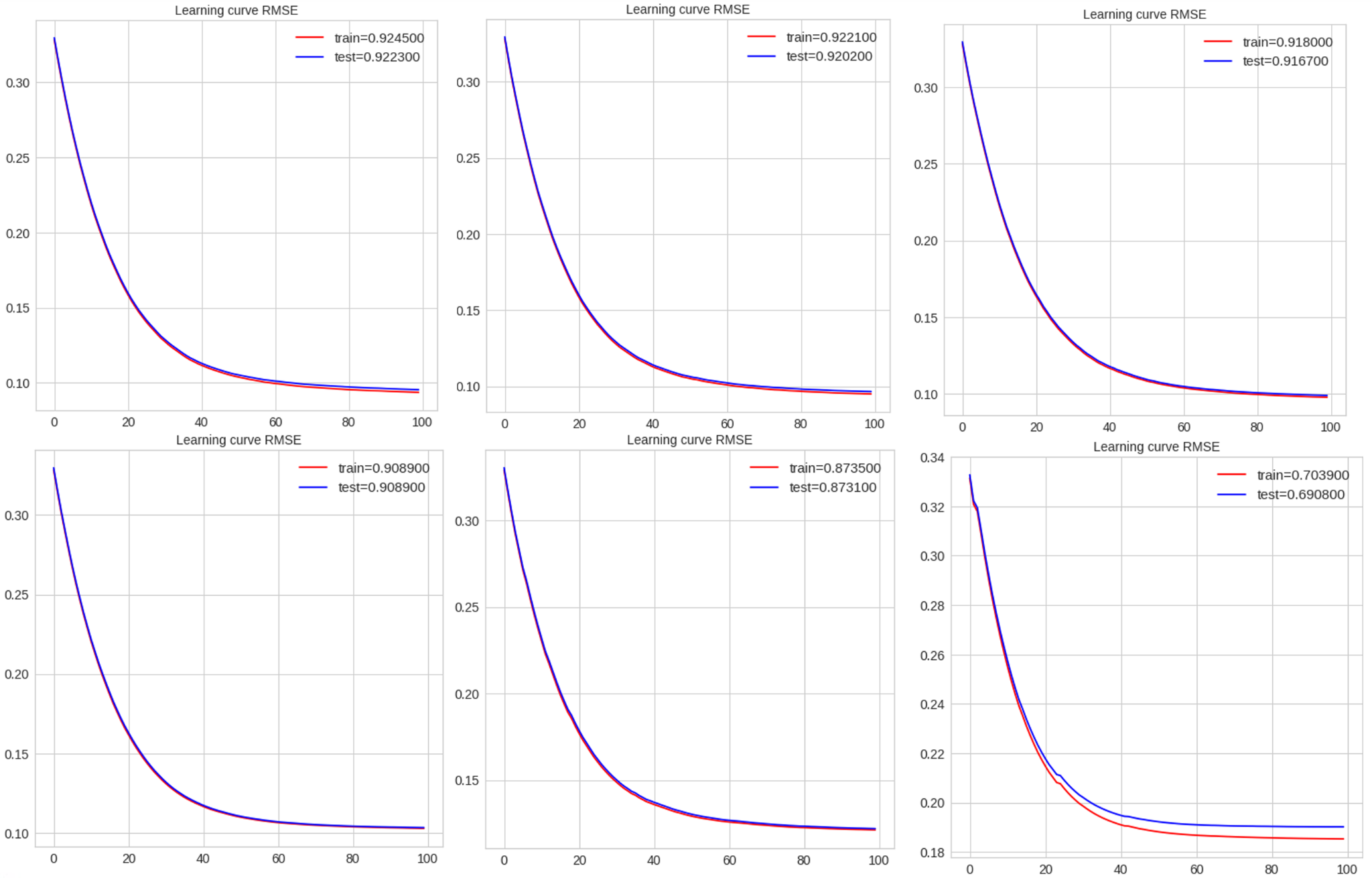
An estimation of the cloud fraction was made after selecting the best characteristics for the model and the following figure is a sample of the result:

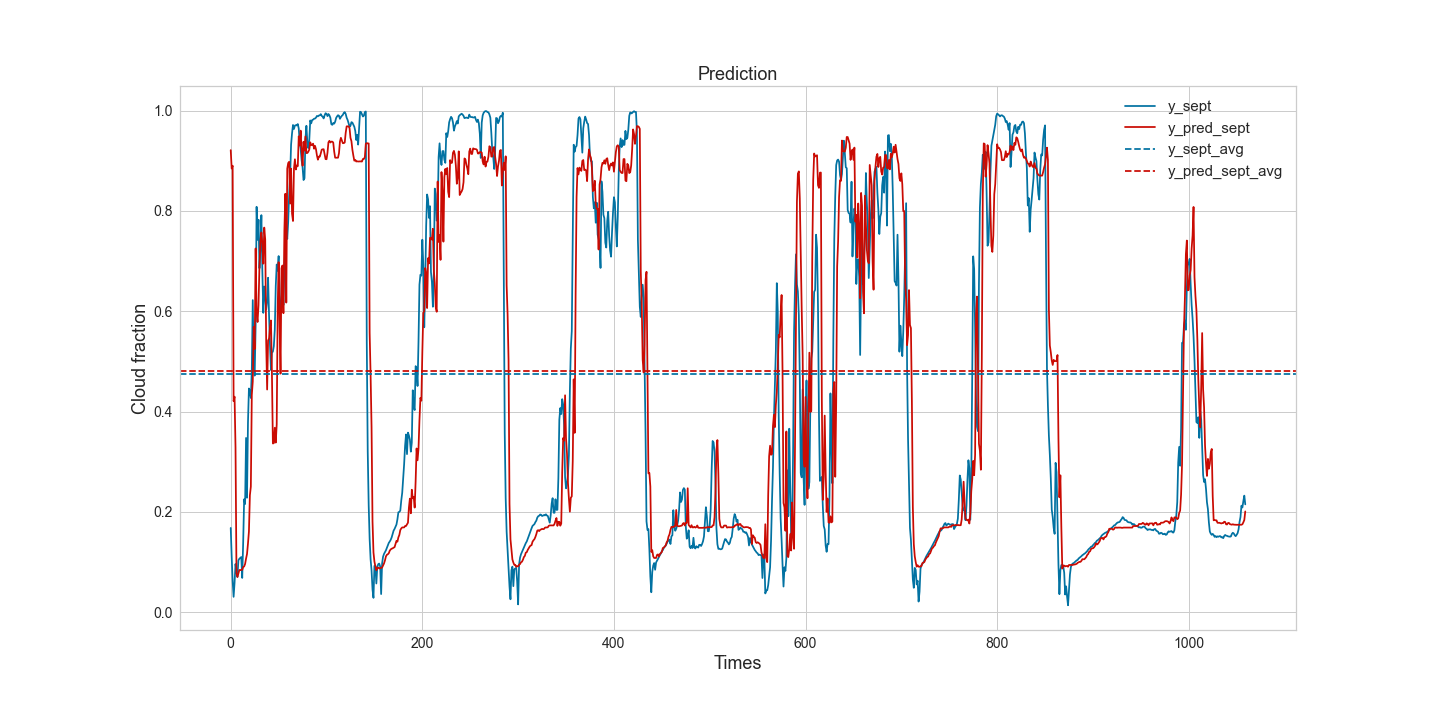
APCADA results :

Physical model called APCADA was used as a reference to assess the results obtained with the model XGBoost. In the following figure, CF\_octat is the real observation, CF\_XGB\_octat represent the result from the model XGBoost, PCA is the partial cloud amount in octat from APCADA, PCA/8 represente the result from the model physique, CFI cloud fraction index, CF\_XGB represent the cloud fraction fom XGBoost model and CF is the real observation.

|  |  |  |  |
| --- | --- | --- | --- |
| r | RMSE | MAE | Corr |
| CF\_octat vs PCA | 2.47 | 1.77 | 0.77 |
| CF\_octat vs CF\_XGB\_octat | 1.06 | 0.66 | 0.92 |
| PCA vs CF\_XGB\_octat | 2.13 | 1.65 | 0.84 |

|  |  |  |  |
| --- | --- | --- | --- |
| Model XGBoost | | | |
| Data source | Input | Output | Period |
| BSRN and UV-Indien network | GSW, SWDif, SWDir, LWD, T, RH, P | CF\_XGB, CF | 2019-09-01 to 2021-03-01 |
| Model APCADA | | | |
| Data source | Input | Output | Period |
| BSRN Reunion, LACy | LWD, T, RH | CF\_APCADA, CF | 2019-11-01 to 2021-05-31 |





# References

ASTER, 2019. ASTER Global Digital Elevation Model V003.

Badosa, J., Haeffelin, M., Chepfer, H., 2013. Scales of spatial and temporal variation of solar irradiance on Reunion tropical island. Solar Energy 88, 42-56.

Bessafi, M., Mihailović, D.T., Malinović-Milićević, S., Mihailović, A., Jumaux, G., Bonnardot, F., Fanchette, Y., Chabriat, J.-P., 2018. Spatial and Temporal Non-Linear Dynamics Analysis and Predictability of Solar Radiation Time Series for La Reunion Island (France). Entropy 20(12), 946.

Cadet, J.-M., Portafaix, T., Bencherif, H., Lamy, K., Brogniez, C., Auriol, F., Metzger, J.-M., Boudreault, L.-E., Wright, C.Y., 2020. Inter-Comparison Campaign of Solar UVR Instruments under Clear Sky Conditions at Reunion Island (21°S, 55°E). International Journal of Environmental Research and Public Health 17(8), 2867.

Carmona, F., Rivas, R., Caselles, V., 2014. Estimation of daytime downward longwave radiation under clear and cloudy skies conditions over a sub-humid region. Theoretical and Applied Climatology 115(1), 281-295.

Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C.H., Liu, R., 2019. XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. IEEE Access 7, 13149-13158.

Didier, N., 2015. Comparison of spatial and temporal cloud coverage derived from CloudSat, CERES, ISCCP and their relationship with precipitation over Africa. American Journal of Remote Sensing 3(2), 17-28.

dos Santos Nascimento, G., Ruhoff, A., Maurício Munar, A., Cavalcanti, J.R., da Motta Marques, D., Roberti, D.R., Ribeiro da Rocha, H., 2018. Validation of CERES earth radiation budget estimations, EGU General Assembly Conference Abstracts. p. 10837.

Duarte, H.F., Dias, N.L., Maggiotto, S.R., 2006. Assessing daytime downward longwave radiation estimates for clear and cloudy skies in Southern Brazil. Agricultural and Forest Meteorology 139(3), 171-181.

Durand, J., Lees, E., Bousquet, O., Delanoë, J., Bonnardot, F., 2021. Cloud Radar Observations of Diurnal and Seasonal Cloudiness over Reunion Island. Atmosphere 12(7), 868.

Dürr, B., Philipona, R., 2004. Automatic cloud amount detection by surface longwave downward radiation measurements. Journal of Geophysical Research: Atmospheres 109(D5).

Fan, Z., Zhan, Q., Yang, C., Liu, H., Bilal, M., 2020. Estimating PM2.5 Concentrations Using Spatially Local Xgboost Based on Full-Covered SARA AOD at the Urban Scale. Remote Sensing 12(20), 3368.

Forsythe, N., Hardy, A.J., Fowler, H.J., Blenkinsop, S., Kilsby, C.G., Archer, D.R., Hashmi, M.Z., 2015. A Detailed Cloud Fraction Climatology of the Upper Indus Basin and Its Implications for Near-Surface Air Temperature. Journal of Climate 28(9), 3537-3556.

Guichard, F., Couvreux, F., 2017. A short review of numerical cloud-resolving models. Tellus A: Dynamic Meteorology and Oceanography 69(1), 1373578.

Heidinger, A.K., Cox, S.K., 1996. Finite-Cloud Effects in Longwave Radiative Transfer. Journal of Atmospheric Sciences 53(7), 953-963.

Huang, L., Kang, J., Wan, M., Fang, L., Zhang, C., Zeng, Z., 2021. Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. Frontiers in Earth Science 9, 596860.

Kim, B.-Y., Jee, J.-B., Zo, I.-S., Lee, K.-T., 2016. Cloud cover retrieved from skyviewer: A validation with human observations. Asia-Pacific Journal of Atmospheric Sciences 52(1), 1-10.

Klebe, D.I., Blatherwick, R.D., Morris, V.R., 2014. Ground-based all-sky mid-infrared and visible imagery for purposes of characterizing cloud properties. Atmos. Meas. Tech. 7(2), 637-645.

Lamy, K., Portafaix, T., Brogniez, C., Lakkala, K., Pitkänen, M.R.A., Arola, A., Forestier, J.B., Amelie, V., Toihir, M.A., Rakotoniaina, S., 2021. UV-Indien network: ground-based measurements dedicated to the monitoring of UV radiation over the western Indian Ocean. Earth Syst. Sci. Data 13(9), 4275-4301.

Liu, W., Liu, Y., Zhang, T., Han, Y., Zhou, X., Xie, Y., Yoo, S., 2022. Use of physics to improve solar forecast: Part II, machine learning and model interpretability. Solar Energy 244, 362-378.

Long, C.N., Ackerman, T.P., 2000. Identification of clear skies from broadband pyranometer measurements and calculation of downwelling shortwave cloud effects. Journal of Geophysical Research: Atmospheres 105(D12), 15609-15626.

Long, C.N., Shi, Y., 2008. An automated quality assessment and control algorithm for surface radiation measurements. The Open Atmospheric Science Journal 2(1).

Maimó-Far, A., Tantet, A., Homar, V., Drobinski, P., 2020. Predictable and Unpredictable Climate Variability Impacts on Optimal Renewable Energy Mixes: The Example of Spain. Energies 13(19), 5132.

Mialhe, P., Pohl, B., Morel, B., Trentmann, J., Jumaux, G., Bonnardot, F., Bessafi, M., Chabriat, J.-P., 2020. On the determination of coherent solar climates over a tropical island with a complex topography. Solar Energy 206, 508-521.

Mo, H., Sun, H., Liu, J., Wei, S., 2019. Developing window behavior models for residential buildings using XGBoost algorithm. Energy and Buildings 205, 109564.

Morel, B., Jeanty, P., Delsaut, M., Hassambay, N., Graillet, A., Chabriat, J.-P., 2021. Surface measurement of total solar and ultraviolet irradiance and ancillary meteorological data at the South-West Indian Ocean Solar network (IOS-net) stations. Data in Brief 37, 107136.

Morel, B., Pohl, B., Tang, C., \*, 2020. Regionalizing Surface Solar Radiation at Very High Resolution over La

2 Reunion Island Using a Regional Climate Model.

NASA, 2021. CERES\_SYN1deg\_Ed4A Data Quality Summary (4/8/2021).

Ohmura, A., Dutton, E.G., Forgan, B., Fröhlich, C., Gilgen, H., Hegner, H., Heimo, A., König-Langlo, G., McArthur, B., Müller, G., Philipona, R., Pinker, R., Whitlock, C.H., Dehne, K., Wild, M., 1998. Baseline Surface Radiation Network (BSRN/WCRP): New Precision Radiometry for Climate Research. Bulletin of the American Meteorological Society 79(10), 2115-2136.

Park, S., Shin, J., 2019. Heuristic estimation of low-level cloud fraction over the globe based on a decoupling parameterization. Atmos. Chem. Phys. 19(8), 5635-5660.

Pous, S., Lazure, P., André, G., Dumas, F., Halo, I., Penven, P., 2014. Circulation around La Réunion and Mauritius islands in the south‐western Indian Ocean: A modeling perspective. Journal of Geophysical Research: Oceans 119(3), 1957-1976.

Qian, Y., Long, C.N., Wang, H., Comstock, J.M., McFarlane, S.A., Xie, S., 2012. Evaluation of cloud fraction and its radiative effect simulated by IPCC AR4 global models against ARM surface observations. Atmos. Chem. Phys. 12(4), 1785-1810.

Réchou, A., Flores, O., Jumaux, G., Duflot, V., Bousquet, O., Pouppeville, C., Bonnardot, F., 2019. Spatio-temporal variability of rainfall in a high tropical island: Patterns and large-scale drivers in Réunion Island. Quarterly Journal of the Royal Meteorological Society 145(720), 893-909.

Satoh, M., Stevens, B., Judt, F., Khairoutdinov, M., Lin, S.-J., Putman, W.M., Düben, P., 2019. Global Cloud-Resolving Models. Current Climate Change Reports 5(3), 172-184.

Sheng, G., Shunlin, L., Lin, L., 2009. Validation of surface radiation data provided by the CERES over the Tibetan Plateau, 2009 17th International Conference on Geoinformatics. pp. 1-6.

Silber, I., Verlinde, J., Wang, S.-H., Bromwich, D.H., Fridlind, A.M., Cadeddu, M., Eloranta, E.W., Flynn, C.J., 2019. Cloud Influence on ERA5 and AMPS Surface Downwelling Longwave Radiation Biases in West Antarctica. Journal of Climate 32(22), 7935-7949.

Smith, G.L., Priestley, K.J., Loeb, N.G., Wielicki, B.A., Charlock, T.P., Minnis, P., Doelling, D.R., Rutan, D.A., 2011. Clouds and Earth Radiant Energy System (CERES), a review: Past, present and future. Advances in Space Research 48(2), 254-263.

Sugita, M., Brutsaert, W., 1993. Cloud effect in the estimation of instantaneous downward longwave radiation. Water Resources Research 29(3), 599-605.

Tang, C., Mialhe, P., Pohl, B., Morel, B., Wild, M., Koseki, S., Abiodun, B., Bessafi, M., Lennard, C., Beeharry, G., Loollchund, R., Cunden, T.S.M., Singh, S., 2023. Intraseasonal and synoptic modulation of diurnal surface solar radiation over the South-West Indian Ocean (reviewing). Solar Energy.

Tapakis, R., Charalambides, A.G., 2013. Equipment and methodologies for cloud detection and classification: A review. Solar Energy 95, 392-430.

Vérèmes, H., Listowski, C., Delanoë, J., Barthe, C., Tulet, P., Bonnardot, F., Roy, D., 2019. Spatial and seasonal variability of clouds over the southwest Indian Ocean based on the DARDAR mask product. Quarterly Journal of the Royal Meteorological Society 145(725), 3561-3576.

Wang, Y., Liu, D., Xie, W., Yang, M., Gao, Z., Ling, X., Huang, Y., Li, C., Liu, Y., Xia, Y., 2021. Day and Night Clouds Detection Using a Thermal-Infrared All-Sky-View Camera. Remote Sensing 13(9), 1852.

Yan, H., Huang, J., Minnis, P., Wang, T., Bi, J., 2011. Comparison of CERES surface radiation fluxes with surface observations over Loess Plateau. Remote Sensing of Environment 115(6), 1489-1500.

Yang, F., Cheng, J., 2020. A framework for estimating cloudy sky surface downward longwave radiation from the derived active and passive cloud property parameters. Remote Sensing of Environment 248, 111972.

Yeo, H., Park, S.-J., Kim, B.-M., Shiobara, M., Kim, S.-W., Kwon, H., Kim, J.-H., Jeong, J.-H., Park, S.S., Choi, T., 2018. The observed relationship of cloud to surface longwave radiation and air temperature at Ny-Ålesund, Svalbard. Tellus B: Chemical and Physical Meteorology 70(1), 1-10.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., Si, Y., 2018. A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost. IEEE Access 6, 21020-21031.

Zhang, T., Stackhouse, P.W., Gupta, S.K., Cox, S.J., Mikovitz, J.C., Srb, N.G., 2011. THE EFFECT OF CLOUD FRACTION ON THE RADIATIVE ENERGY BUDGET: The Satellite-Based GEWEX-SRB Data vs. the Ground-Based BSRN Measurements. pp. A13B-0261.