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, Béatrice Morel1

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**

A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: (1) Background: place the question addressed in a broad context and highlight the purpose of the study; (2) Methods: describe briefly the main methods or treatments applied; (3) Results: summarize the article's main findings; (4) Conclusion: indicate the main conclusions or interpretations. The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.

**A 200 words 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.

APCADA scores, and a concluding sentence

**A submitted 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) by 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. Variables explored are the radiative fluxes measured at the station, namely the downward longwave radiation (LWD) from a CGR4 pyrgeometer, the global and diffuse components of the downward shortwave radiation (SWD) from two CMP22 pyranometers, the direct component of the SWD from a CH1 pyrheliometer, along with some additional meteorological parameters (surface pressure, surface air temperature, and relative humidity) from a WXT530 weather transmitter. CF data are obtained from collocated and coincident observations from a SkyCamVision all-sky camera within the UV-Indien network.

In the first step, a set of quality control (QC) procedures derived from the BSRN recommendations, is used to test the quality of the LWD measurements taken at the station of La Réunion. And because this study is the first one based on the LWD at this station, the LWD data are also compared to CERES (Clouds and Earth’s Radiant Energy Systems) satellite-based product. In the second step, a XGBoost model is applied to estimate CF. A 2-year dataset (September 2019-June 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. The predictors in the model are then the LWD and SWD fluxes and the additional meteorological parameters measured at the BSRN station, and the predictand is the CF measurements from the UV-Indien network. Cross validation is performed for training. Learning curves are used to evaluate the model during the training and the validation. Statistical metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Square Error (RMSE), are also calculated for the final evaluation and model comparison.

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. In this study, the 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. After feature selections and model optimizations, a final score of 92.02% is achieved, which is comparable to the results from recent studies. We find that surface pressure is not necessary for this application since it makes the XGBoost model more complex.

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. This algorithm is based on an enhanced version of the clear-sky index and provides a robust real-time cloud amount. The output of this algorithm can be compared with cloud cover measurement from other high temporal cloud detection systems. A comparison of the CF estimation with these two methods was performed.

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

**Introduction**

The Sustainable Development Goals (SDGs) of the United Nations (<https://sdgs.un.org/goals>) provide a powerful framework for international cooperation to achieve a sustainable future for the planet. The global goal on energy - SDG 7 - which aims at ensuring access to affordable, reliable, sustainable, and modern energy for all by 2030, implies a shift away from fossil-fuel-based sources toward renewable energy sources (RES; e.g., Gielen et al. 2019). However, variability and uncertainty inherent to RES technologies represent major challenges for the integration of RES in the energy system (Widén et al., 2015; Engeland et al., 2017; Verzijlbergh et al., 2017). This is particularly the case for solar energy (SE), one of the most widespread RES today. Surface solar radiation (SSR), which varies over a wide range of temporal and spatial scales, is affected by many factors, mostly the position of the sun and cloud cover (Perez et al., 2016). Because the presence of clouds adds unpredictable variations to SE (Maimó-Far et al., 2020), understanding the occurrence of clouds and their link to SSR is then important for SE management (Danso et al., 2020).

Our case study is the tropical island of La Réunion (21°S; 55.5°E) in the South West Indian Ocean (SWIO), where SE is one key in the transition to an electricity system 100% renewable by 2030 (Selosse et al., 2018). Despite the high solar resource availability over the island (Mialhe et al., 2020), the development of SE there, could be hindered by drawbacks associated with the large variability of cloudiness over the island (Badosa et al., 2013; Vérèmes et al., 2019; Durand et al., 2021).

Located in the South West Indian Ocean, La Réunion island is mainly under an east-southeast trade wind flow, which changes in direction according to the season: easterly in summer and more southeasterly in winter (Mialhe et al., 2020). Because of the topographical characteristics of the island (Fig. ?), this main flow is more a “flow around” than a “flow over '' regime. 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., 2015).

chao: shall we put some sentences of climate of Reunion somewhere (here or data, where we show the target variable from LACy (data) ?), especially the information linked to cloud, e.g., the wet dry seasons, etc. to interpret the results which may be season-dependent.

Information on local cloud coverage can be derived from human observation, meteorological satellites and surface remote observation data (\cite{kim2020cloud}). Human observation refers to a regular observer recording cloud cover data at either 1, 3, or 6 h interval. To minimize human errors, ground- and aircraft-based measurements have been introduced to timely record 1-D cloud snapshots (\cite{qian2012evaluation}). In the global observation trend, remote sensing has been applied to evaluate global cloud properties and estimate cloud cover parameters for over years from visible and infrared radiations (\cite{qian2012evaluation}). 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. Responding to the absence of visible light during night time, infrared sensors and filters have been developed to detect radiance from clouds (\cite{kim2020cloud}).

Presence of long-term continuous ground-based measurements has attracted/created statistical and machine learning (ML) research interests. CF, fraction of area occupied by clouds, found to have a significant correlation with air temperature, precipitation, and topography (\cite{didier2015comparison}, and \cite{forsythe2015detailed}). 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). In cloud-resolving models (CRMs), neural networks-based cumulus parameterization accurately diagnosed both precipitation and CF with little instability for multiple time steps prediction (\cite{brenowitz2018prognostic}). By learning from CRMs simulations, deep neural networks achieved a more accurate convection and rapid execution on accelerators like Graphical Processing Units (GPUs) (\cite{ukkonen2020accelerating}. Statistically, Yunfeng Cao (2022) - with random forests (RF), extremely randomized trees (ERT), and categorical boosting (CatBoost) - found a high dependence of atmospheric temperature and water vapour on the Clear-sky longwave downward radiation. In 2013, while estimating daytime downward longwave radiation under both cloudy and clear skies, Facundo Carmona found a high correlation between CF and longwave with multiple linear regression. Several studies have been done estimating longwave with cloud fractions as input, but no work done yet the other way around with ML.

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).

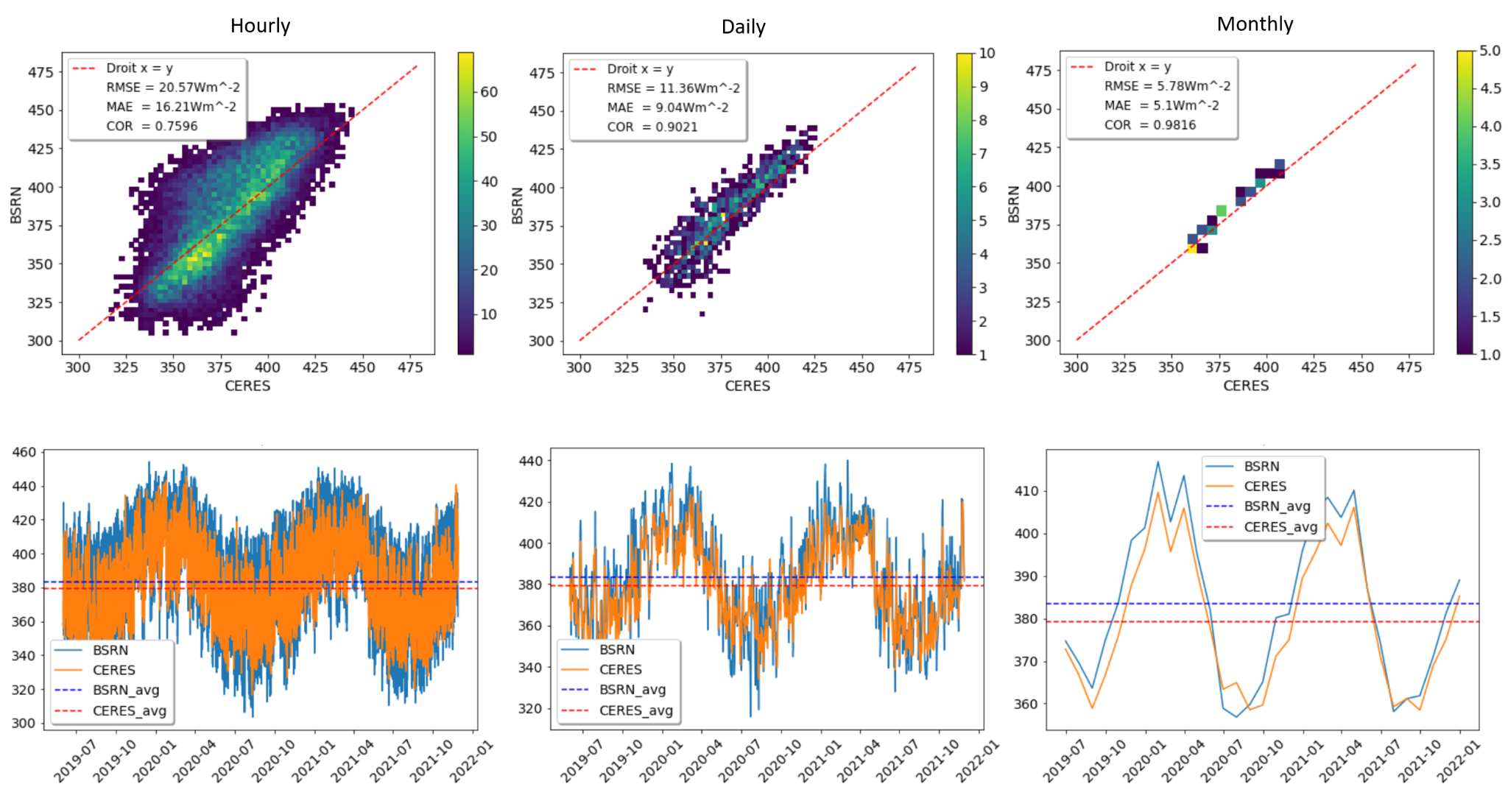
**Methods and data**

1. Data
   1. *RUN BSRN-station data*

The BSRN station of La Réunion (RUN) is managed by the ENERGY Lab. It is located on the main campus of the University of Reunion island (Moufia at 28.3° S, 16.5° W; xxx m a.s.l.), at Saint-Denis, on the northern coast of the island (Fig. ?).

BSRN station data in the island of La Réunion, as other BSRN stations, benefit from automatic quality control (AQC) and visual quality control (VQC) (\cite{long2008automated}). In addition, high precision sensors are used in all BSRN stations (do we need a BSRN map?) over the globe, calibrated every two years at the World Radiometric Reference and the World Radiation Centre (PMOD/WRC) in Davos. The station, at 1-minute intervals, records global horizontal shortwave irradiance (GSW, in W/m2), diffuse horizontal shortwave irradiance (SWDif, in W/m2), direct normal shortwave irradiance (SWDir, in W/m2), downward longwave irradiance (LWD, in W/m2), temperature (T, in °C), relative humidity (RH, in %), and pressure (P, in hPa). Unshaded and shaded CMP22 pyranometer measures GSW, and SWDif respectively, CHP11 pyrheliometer is used to measure SWDir, CGR4 pyrgeometer for LWdn, and WXT530 weather transmitter for T, RH, and P.

And because this study is the first one based on the LWD at this station, in addition to the QCs, the LWD data are also compared to CERES (Clouds and Earth’s Radiant Energy Systems) satellite-based product. All available data (June 2019 - February 2022) was used, and RMSE, MBE, r were used for comparison. Figure (\ref{1}) shows a correlation of 76, 90, and 98%; RMSE of 20.57, 11.36, and 5.76 W/m2 between LWD of the BSRN station and CERES at hourly, daily, and monthly intervals respectively.



( a ) ( b ) ( c )

Fig 1: Comparison of LWD from Reunion BSRN station and CERES, at timescales of (a) hourly, (b) daily, and (c) monthly for the period of June 2019 - February 2022.

* 1. CF measurements from the UV-Indien network

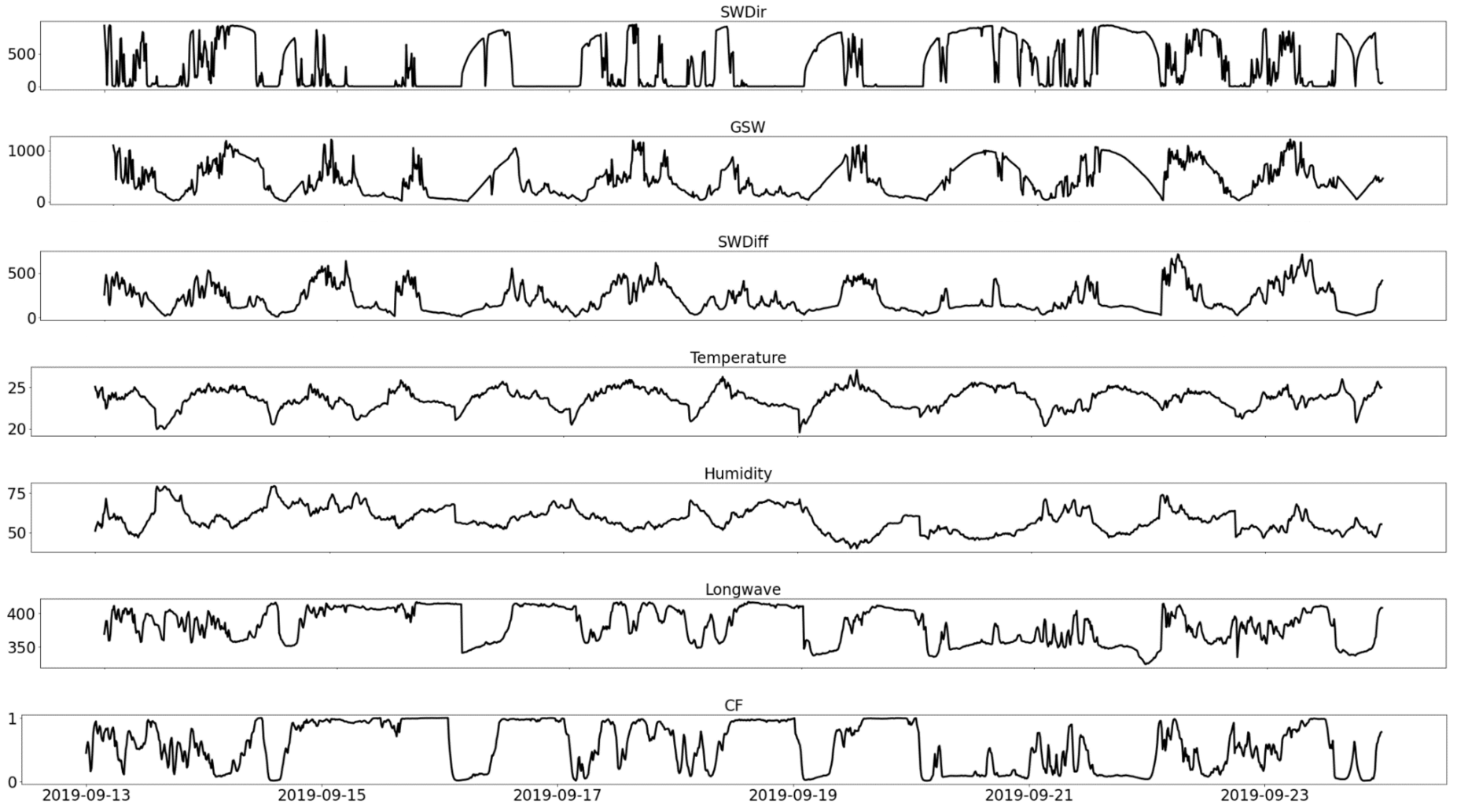
CF data are obtained from collocated and coincident observations from a SkyCamVision all-sky camera (<http://www.reuniwatt.com/>) within the UV-Indien network. This camera is equipped with a fisheye-lens CMOS Sensor of 1600 x 1200 pixels resolution, and acquires 1-minute hemispherical images in the visible range (380-440 nm). From images, pixels are classified into clear sky; sun; thick cloud; or thin cloud with cloud segmentation algorithm, inspired by Random Forest (RF). CF is computed from pixels with clear and cloudy sky, geometrically calibrated image (\cite{cadet2020inter}). As the camera is not mounted with infrared sensors, the station does not record night-time CF. In this study, daytime CF observations were used for the ‘2019-09-13 to 2021-02-01’ period, Figure (\ref{}).

Chao: some information for this station, at least the distance from our BSRN station.

Chao: for future, better to add another CF measurement for comparison.

Chao: should be mentioned that we have only daytime obs (sure), donc not estimation during night.

Shall we present the predictand by a plot, to show the variability of CF ? (In the performance section, we have only some days.)



1. Automatic Partial Cloud Amount Detection Algorithm (APCADA)

Influence of broken clouds on LWD has been explained by the observed relationship between LWD variability and partial cloud amount (PCA) (\cite{durr2004automatic}). In addition to the standard deviation of LWD (Stdev LWD), PCA estimation uses cloud free index (CFI). CFI is a ratio of apparent emittance of the sky to the empirical apparent cloud free emittance;

, with , and eq.1

Where is the Stephan-Boltzmann constant, T is air temperature (K), e is water vapor pressure (Pa), Ta is is altitude-dependent emittance of dry atmosphere, and k is a location-dependent coefficient.

Parameter estimation is based on the Cloud-free day detection, recently using global and diffuse solar radiation(\cite{garcia2018comparison}). In this study, we used Long and methods (\cite{long2000identification}).

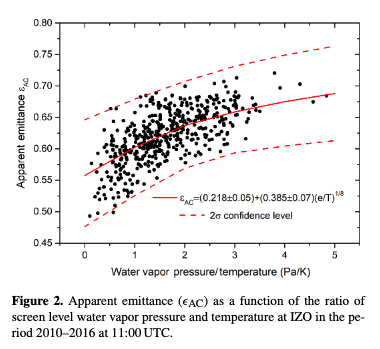
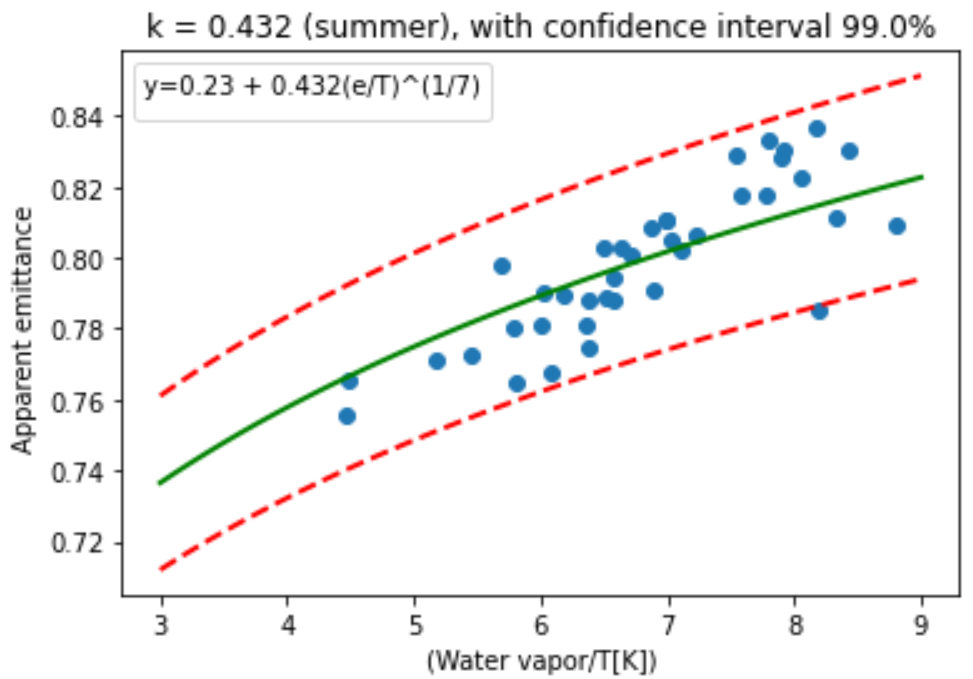


Fig 2: Apparent emittance vs ratio of water vapor pressure and temperature at Reunion for September 2019 - February 2021 period. (equation inside the figure?)

E\_ad = 0.23 and k = 0.432

And lastly, the Automatic PCA algorithm (APCADA) of Durr and Philipona (\cite{durr2004automatic}) which uses Stdev LWD, and CFI with heuristic rules of optimized thresholds was used to estimate the PCA in octas. In this study, a 5-min interval for the September 2019 - February 2021 period, based on CF availability.

1. XGBoost model

The XGBoost algorithm belongs to the Gradient Boosting Decision Tree (GBDT) algorithm, an iterative decision tree algorithm consisting of multiple decision trees (\cite{chen2016xgboost}). This scalable end-to-end tree boosting system is based on the idea of building a strong and high performance classifier from a series of weak and low accurate classifiers. Starting with one weak learner, it iteratively adds new weak learners to approximate functional gradients (\cite{chen2019xgboost}). With the use of parallel tree boosting, accuracy, and ability to handle various data, XGBoost has become one of the most popular ML algorithms.

In this study, we used the XGBoost library available on Github (<https://github.com/dmlc/xgboost>) about the used codes? A 2-year dataset (September 2019-February 2022) with a 5-min time step is used for this study, with 90% of it used 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.

1. Error metrics

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

eq. 2

eq. 3

eq. 4

Chao: the above table is not clear enough, what about the more simple one below. (maybe even remove the column of period if all data in the same period). solved

Table 1: Summary of the used data

| **Variable** | **Full name** | **unit** | **data source** | **XGBoost** | **APACADA** |
| --- | --- | --- | --- | --- | --- |
| T | near surface air temperature | ° C | BSRN station | predictor | input |
| GSW | global horizontal shortwave irradiance | W/m2 | BSRN station | predictor | - |
| SWDir | direct normal shortwave irradiance | W/m2 | BSRN station | predictor | - |
| SWDif | diffuse horizontal shortwave irradiance | W/m2 | BSRN station | predictor | - |
| LWD | downward longwave irradiance | W/m2 | BSRN station | predictor | input |
| RH | relative humidity | % | BSRN station | predictor | input |
| P | pressure | hPa | BSRN station | predictor | - |
| CF | cloud fraction | % | UV-Indien network | predictand | output |

**Results**

1. **APCADA performance**

APCADA-estimated PCA (in octas) was first converted to CF and then compared with observations at Reunion for September 2019 - February 2021 period. Figure (\ref{}) shows observed and estimated CF time series, APCADA was able to estimate dynamics of the observed CF. About 78.1% of CF estimations agree with observations. RMSE and MAE of 0.31, and 0.23 respectively were observed.

We need to add some other results to increase analysis: (1) showing the level of estimating three classes: CF [0 - 0.25, 0.25 - 0.75, and 0.75 - 1] → to ‘better’ evaluate under/overestimation. (2) plot 1-day results.

here

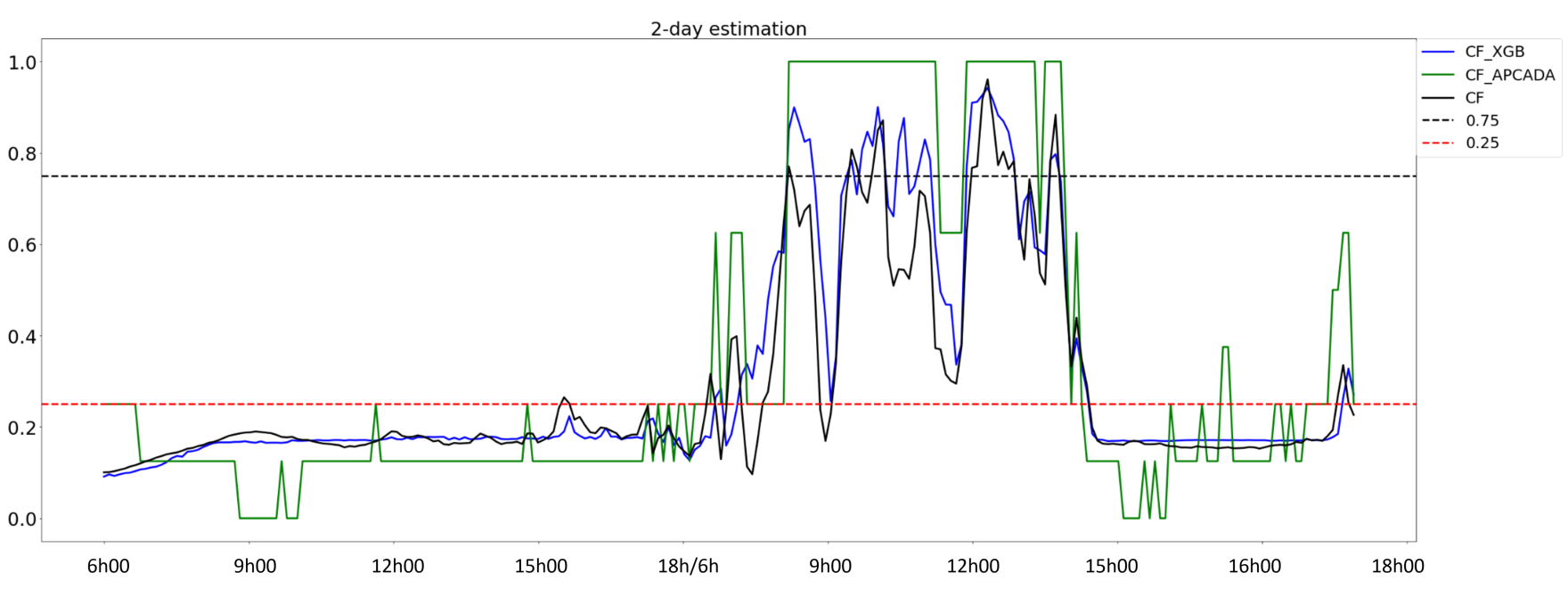


Table 2: Model performance in estimating CF different levels (low, medium, and high cloudy)

| Model\I | 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 |

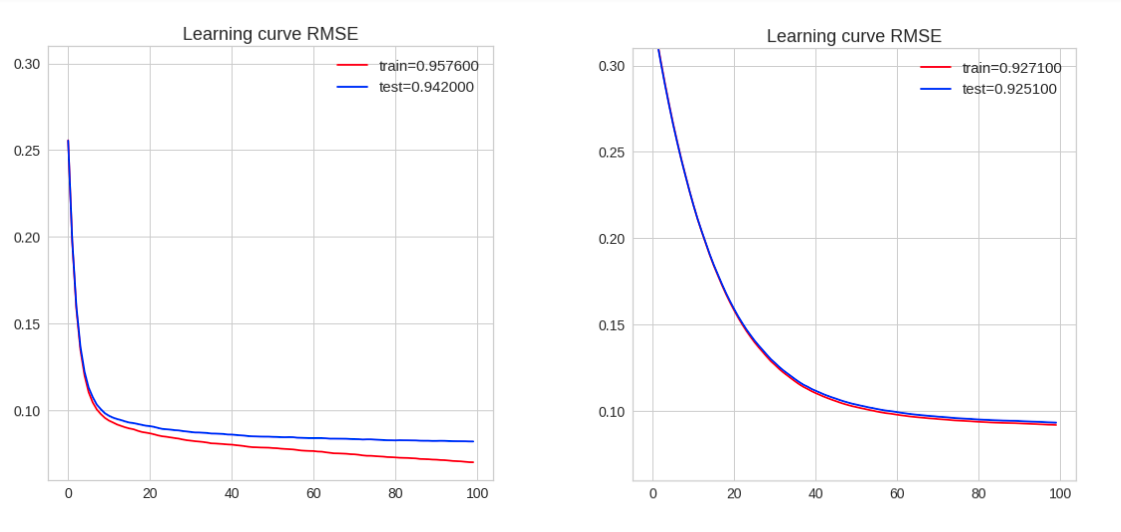
Comment: All values are greater than 80%, except for KWA, where the minimum score is about 70%.The lowest score rates are found for KWA ([Table 5](https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2003JD004182#jgrd11026-tbl-0005)). Investigations showed that the percentage of broken PCA (2–6 octas) 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)

1. **XGBoost performance**

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 (\ref{3}) shows learning curves before and after tuning hyperparameters. And the table (\ref{2}) 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 (\cite{durr2004automatic} do we have other similar studies). Figure (\ref{5}) shows CF estimated time series with observations, with RMSE and MAE of 0.12 and 0.08 respectively (Table(\ref{2})). We find that the surface pressure is not necessary for this application since it makes the XGBoost model more complex, Figure(\ref{4}).



( a ) ( b )

Fig 3: Learning curve (RMSE) before (a) and (b) after tuning hyper-parameteres 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 |

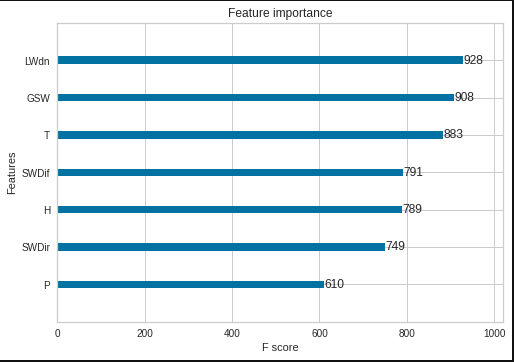


Fig 4: F-score of XGBoost model features to CF estimation.

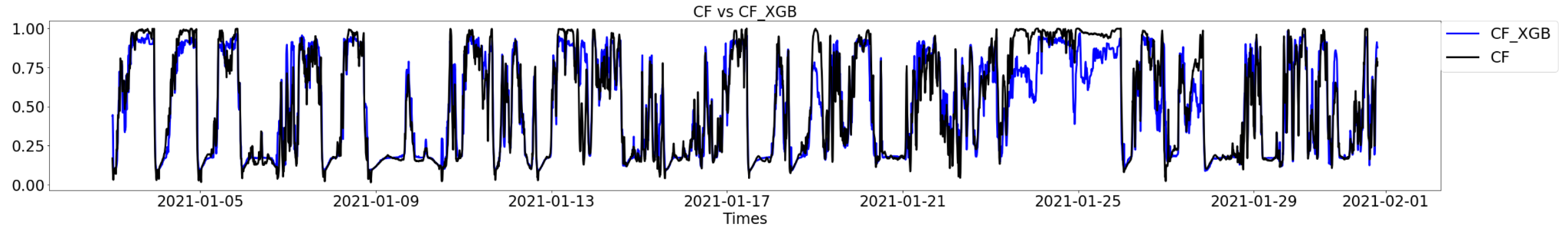


Fig 5: XGBoost CF estimation and observations at Reunion for January 2021 - February 2021 period. (plot modifi-cation)

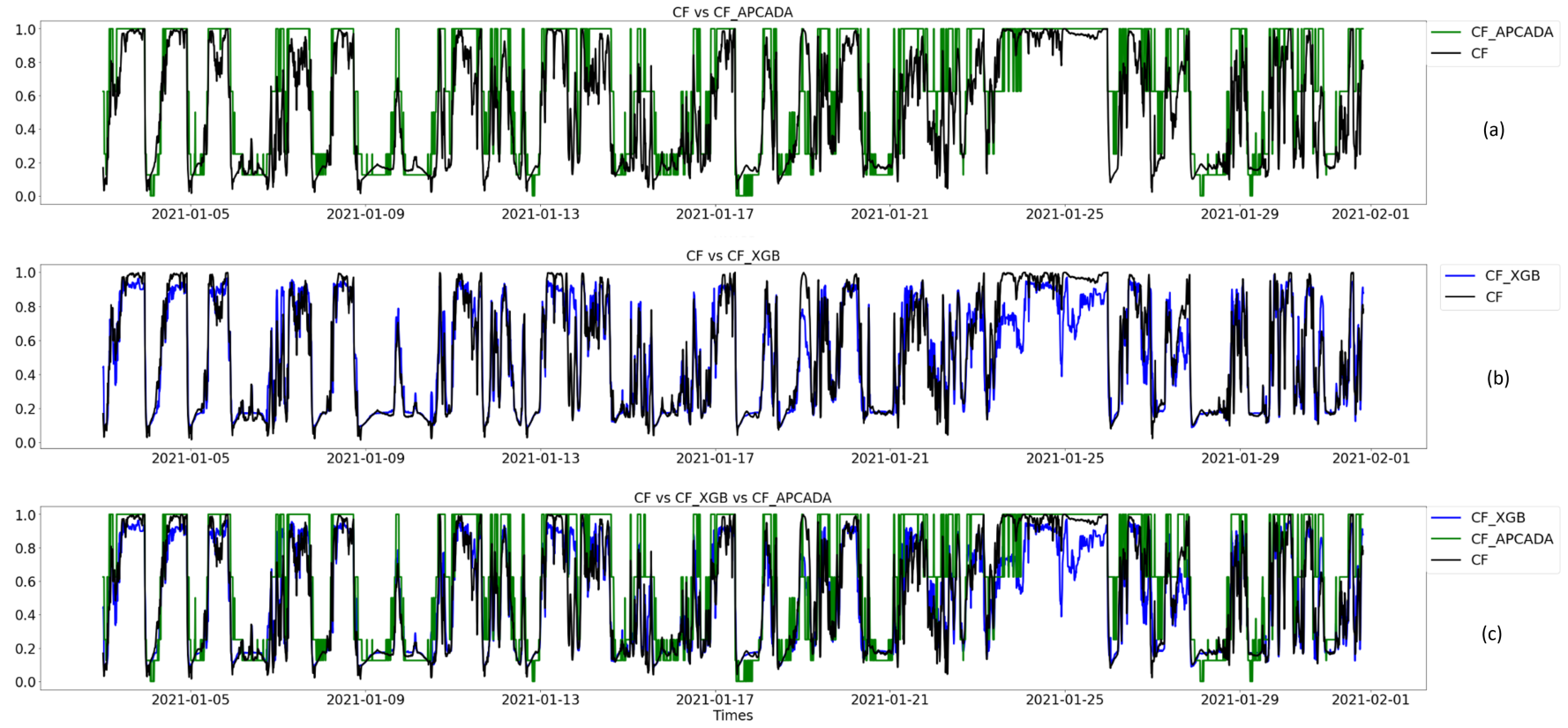


Fig 6: Comparison of observed and estimated CF, (a) observation, (b) XGBoost, and (c) APCADA together with CFI.

1. **CF estimations**

Statistical and physical methods were compared, Table (\ref{2}, and \ref{3}). 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.

1. **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) (\cite{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

1. **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**

Réserves

nd its representation in Numerical Weather Prediction (NWP) models still has a long way to go (\cite{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 (\cite{aida1977reflection}, and \cite{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 (\cite{durr2004automatic}, \cite{park2019heuristic}). 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 (\cite{qian2012evaluation}, and \cite{zhang2011effect}). 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 (\cite{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 (\cite{qian2012evaluation}). 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 (\cite{didier2015comparison}, and \cite{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

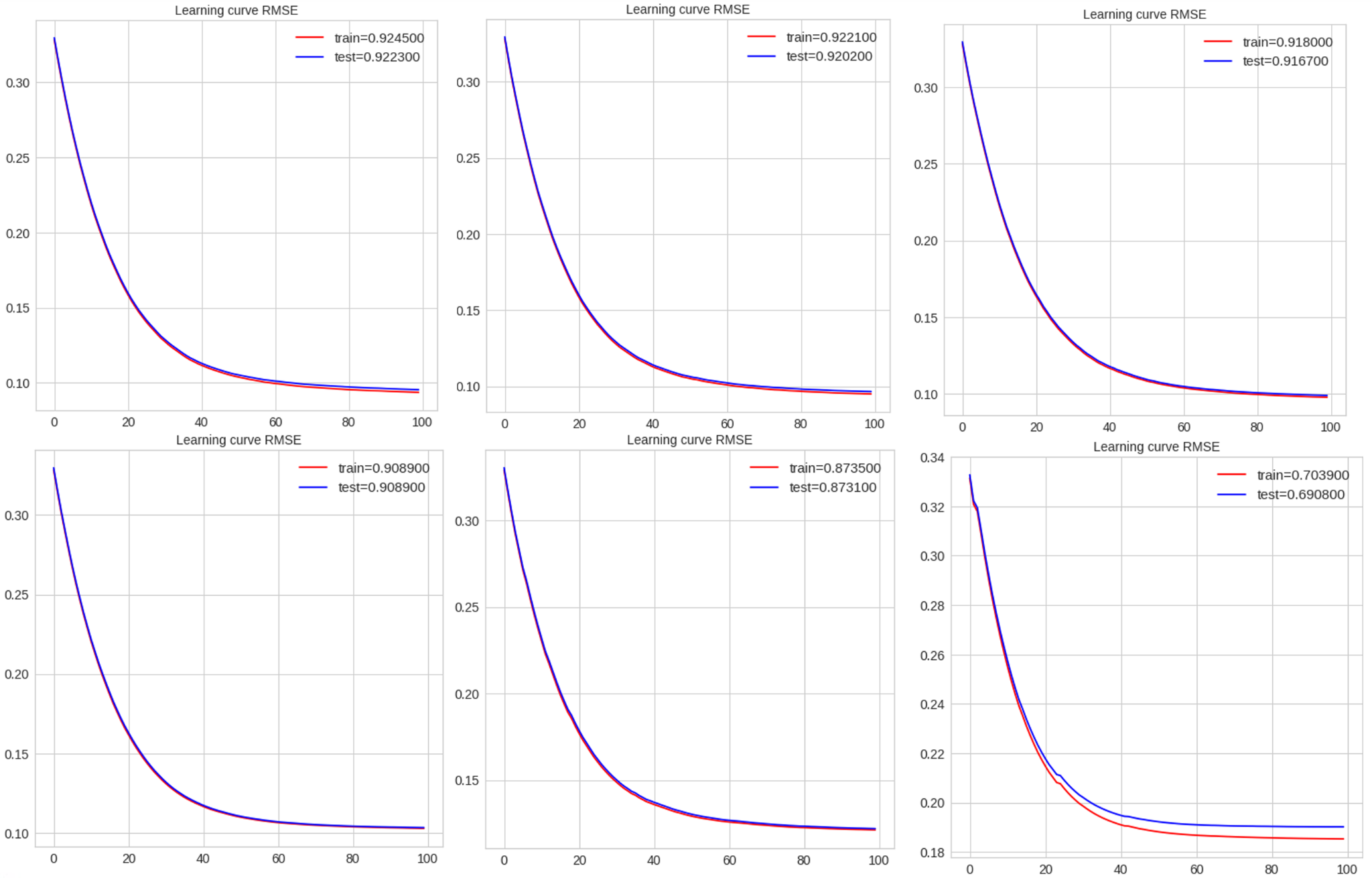
An estimation of the cloud fraction was made after selecting the best characteristics for the model and the following figure is a sample of the result:

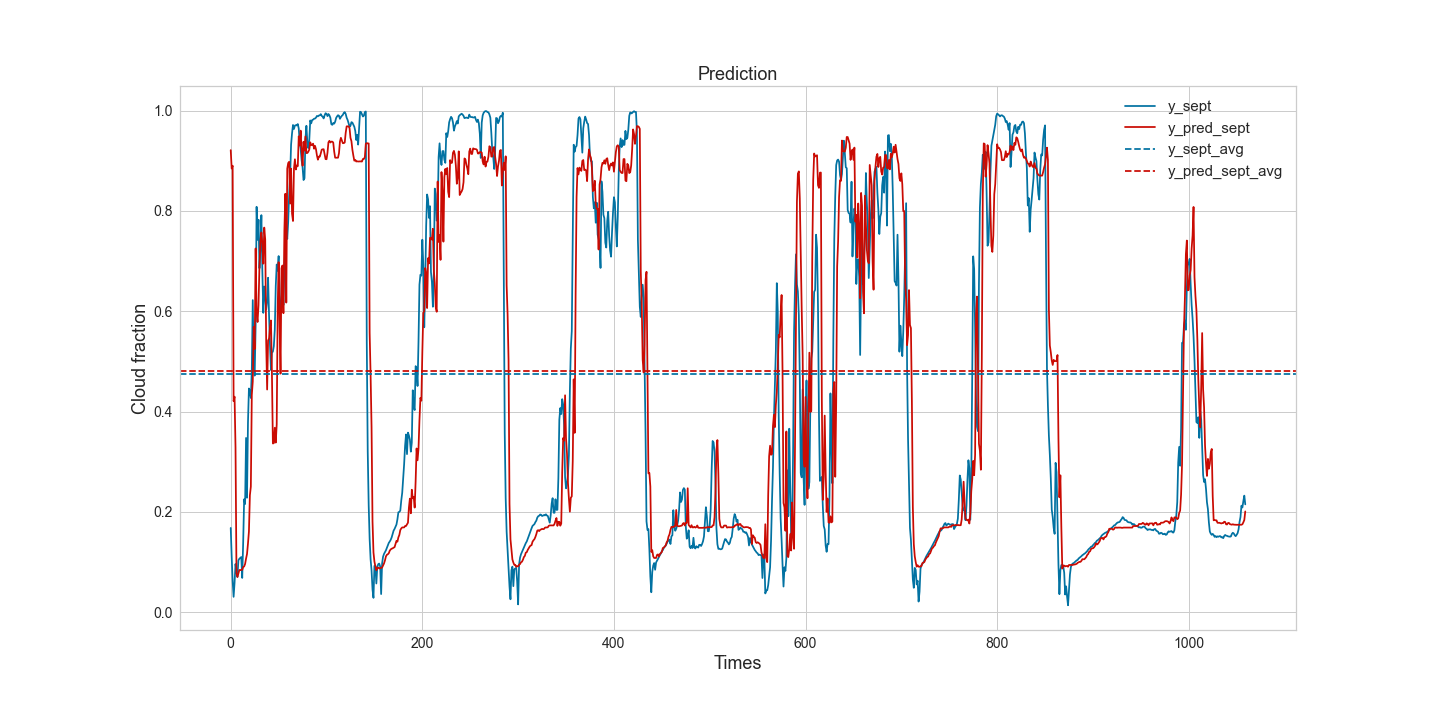
APCADA results :

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 |





@article{durr2004automatic,

title={Automatic cloud amount detection by surface longwave downward radiation measurements},

author={D{\"u}rr, Bruno and Philipona, Rolf},

journal={Journal of Geophysical Research: Atmospheres},

volume={109},

number={D5},

year={2004},

publisher={Wiley Online Library}

}

@article{park2019heuristic,

title={Heuristic estimation of low-level cloud fraction over the globe based on a decoupling parameterization},

author={Park, Sungsu and Shin, Jihoon},

journal={Atmospheric Chemistry and Physics},

volume={19},

number={8},

pages={5635--5660},

year={2019},

publisher={Copernicus GmbH}

}

@inproceedings{zhang2011effect,

title={The Effect of Cloud Fraction on the Radiative Energy Budget: The Satellite-Based GEWEX-SRB Data vs. the Ground-Based BSRN Measurements},

author={Zhang, T and Stackhouse, PW and Gupta, SK and Cox, SJ and Mikovitz, JC and Srb, Gewex},

booktitle={AGU Fall Meeting Abstracts},

volume={2011},

pages={A13B--0261},

year={2011}

}

@article{qian2012evaluation,

title={Evaluation of cloud fraction and its radiative effect simulated by IPCC AR4 global models against ARM surface observations},

author={Qian, Yun and Long, Charles N and Wang, Hailong and Comstock, Jennifer M and McFarlane, Sally A and Xie, Shaocheng},

journal={Atmospheric Chemistry and Physics},

volume={12},

number={4},

pages={1785--1810},

year={2012},

publisher={Copernicus GmbH}

}

@article{kim2020cloud,

title={Cloud Observation and Cloud Cover Calculation at Nighttime Using the Automatic Cloud Observation System (ACOS) Package},

author={Kim, Bu-Yo and Cha, Joo Wan},

journal={Remote Sensing},

volume={12},

number={14},

pages={2314},

year={2020},

publisher={Multidisciplinary Digital Publishing Institute}

}

@article{cadet2020inter,

title={Inter-Comparison Campaign of Solar UVR Instruments under Clear Sky Conditions at Reunion Island (21 S, 55 E)},

author={Cadet, Jean-Maurice and Portafaix, Thierry and Bencherif, Hassan and Lamy, K{\'e}vin and Brogniez, Colette and Auriol, Fr{\'e}d{\'e}rique and Metzger, Jean-Marc and Boudreault, Louis-Etienne and Wright, Caradee Yael},

journal={International Journal of Environmental Research and Public Health},

volume={17},

number={8},

pages={2867},

year={2020},

publisher={Multidisciplinary Digital Publishing Institute}

}

@inproceedings{chen2016xgboost,

title={Xgboost: A scalable tree boosting system},

author={Chen, Tianqi and Guestrin, Carlos},

booktitle={Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining},

pages={785--794},

year={2016}

}

@article{chen2019xgboost,

title={XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system},

author={Chen, Minghua and Liu, Qunying and Chen, Shuheng and Liu, Yicen and Zhang, Chang-Hua and Liu, Ruihua},

journal={IEEE Access},

volume={7},

pages={13149--13158},

year={2019},

publisher={IEEE}

}

@article{durr2004automatic,

title={Automatic cloud amount detection by surface longwave downward radiation measurements},

author={D{\"u}rr, Bruno and Philipona, Rolf},

journal={Journal of Geophysical Research: Atmospheres},

volume={109},

number={D5},

year={2004},

publisher={Wiley Online Library}

}

@article{long2000identification,

title={Identification of clear skies from broadband pyranometer measurements and calculation of downwelling shortwave cloud effects},

author={Long, Charles N and Ackerman, Thomas P},

journal={Journal of Geophysical Research: Atmospheres},

volume={105},

number={D12},

pages={15609--15626},

year={2000},

publisher={Wiley Online Library}

}

@article{garcia2018comparison,

title={Comparison of observed and modeled cloud-free longwave downward radiation (2010--2016) at the high mountain BSRN Iza{\~n}a station},

author={Garc{\'\i}a, Rosa Delia and Barreto, Africa and Cuevas, Emilio and Gr{\"o}bner, Julian and Garc{\'\i}a, Omaira Elena and G{\'o}mez-Pel{\'a}ez, Angel and Romero-Campos, Pedro Miguel and Redondas, Alberto and Cachorro, Victoria Eugenia and Ramos, Ramon},

journal={Geoscientific Model Development},

volume={11},

number={6},

pages={2139--2152},

year={2018},

publisher={Copernicus GmbH}

}