Retrieval of Cloud Fraction using Machine Learning Algorithms based on FY4A AGRI observations.

Jinyi Xia[1] Li Guan[1]

[1]China Meteorological Administration Aerosol-Cloud and Precipitation Key Laboratory, Nanjing University of Information Science and Technology, Nanjing 210044, China

Correspondence to: Li Guan liguan@nuist.edu.cn

Abstract

Cloud fraction as a vital component of meteorological satellite products plays an essential role in environmental monitoring, disaster detection, climate analysis, and other research areas. Random Forest(RF) and Multilayer Perceptron(MLP) algorithms were used in this paper to retrieve the cloud fraction of AGRI (Advanced Geosynchronous Radiation Imager) onboard FY-4A satellite based on its full-disc level-1 radiance observation. Corrections has been made subsequently to the retrieved cloud fraction in areas where solar glint occurs using a correction curve fitted with sun-glint angle as weight. The algorithm includes two steps: the cloud detection is conducted firstly for each AGRI field of view to identify whether it is clear sky, partly cloudy or overcast within the observation field. Then the cloud fraction is retrieved for the scene identified as partly cloudy. The 2B-CLDCLASS-LIDAR cloud fraction product from Cloudsat& CALIPSO active remote sensing satellite is employed as the truth to assess the accuracy of the retrieval algorithm. Comparison with the operational AGRI level 2 cloud fraction product is also conducted at the same time. The results indicate that both the Random Forest (RF) and Multi-Layer Perceptron (MLP) cloud detection models achieved high accuracy, surpassing that of operational products. However, both algorithms demonstrated weaker discrimination capabilities for partly cloudy conditions compared to clear sky and overcast situations. Specifically, they tended to misclassify fields of view with low cloud fractions (e.g., cloud fraction = 0.16) as clear sky and those with higher cloud fractions (e.g., cloud fraction = 0.83) as overcast. Between the two models, RF exhibited higher overall accuracy. Both RF and MLP models performed well in cloud fraction retrieval, showing lower mean error (ME), mean absolute error (MAE), and root mean square error (RMSE) compared to operational products. The ME for both RF and MLP cloud fraction retrieval models was close to zero, while RF had slightly lower MAE and RMSE than MLP. During daytime, the high reflectance in sun-glint areas led to larger retrieval errors for both RF and MLP algorithms. However, after correction, the retrieval accuracy in these regions improved significantly. At night, the absence of visible light observations from the AGRI instrument resulted in lower classification accuracy compared to daytime, leading to higher cloud fraction retrieval errors during nighttime.

Key words: Cloud detection; cloud fraction retrieval; FY-4A AGRI; CloudSat &

CALIPSO; machine learning; deep learning.

Introduction

Clouds occupy a significant proportion within satellite remote sensing data acquired for Earth observation. According to the statistics from the International Satellite Cloud Climatology Project (ISCCP), the annual average global cloud coverage within satellite remote sensing data is around 66% with even higher cloud coverage in specific regions (such as the tropics) (Zhang, et al., 2004). The impact of clouds on the radiation balance of the Earth's atmospheric system is influenced by the optical properties of clouds. Cloud detection, as a vital component of remote sensing image data processing, is considered a critical step for the subsequent identification, analysis, and interpretation of remote sensing images. Therefore, accurately determining cloud coverage is essential in various research domains, such as environmental monitoring, disaster surveillance and climate analysis.

Fengyun-4A (FY-4A) is a comprehensive atmospheric observation satellite launched by China in 2016. The uploaded AGRI (Advanced Geosynchronous Radiation Imager) has 14 channels and captures full-disk observation every 15 minutes. In addition to observing clouds, water vapor, vegetation and the Earth's surface, it also possesses the capability to capture aerosols and snow. Moreover, it can clearly distinguish different phases and particle size of clouds and obtain high- to mid-level water vapor content. It is particularly suitable for cloud detection due to its simultaneous use of visible, near-infrared, and long-wave infrared channels for observation with 4km spatial resolution.

Numerous cloud detection algorithms have been provided based on observations from satellite-borne imagers. The threshold method has been widely employed by researchers, including the early ISCCP (International Satellite Cloud Climatology Project) method (Rossow, 1993) and the proposed threshold methods based on different spectral features or underlying surfaces (Kegelmeyer,1994; Solvsteen,1995; Baum and Trepte,1996). However, there is a significant subjectivity in selection of thresholds whether it is the single and fixed threshold in the early days, multiple thresholds, dynamic thresholds, or adaptive thresholds. The selection of thresholds is influenced by season and climate. Surface reflectance varies significantly between different seasons, such as increased reflectance from snow in winter and vegetation flourishing in summer affecting reflectance. As a result, changes in surface features during different seasons lead to variations in the distribution of grayscale values in images, requiring adjustments to thresholds based on seasonal characteristics. Climate conditions like cloud cover, atmospheric humidity, etc., impact the distinguishability of clouds and other features. For instance, in humid or cloudy climates, the reflectance of the surface and clouds may be similar, necessitating stricter thresholds for differentiation. Therefore, climate conditions also influence threshold selection.

The other category of cloud detection algorithms is based on statistical probability theory. For example the principal component discriminant analysis and quadratic discriminant analysis methods were used to SEVIRI (Spinning Enhanced Visible and Infrared Imager) cloud detection (Amato et al., 2008). The cloud detection algorithm for Thermal Infrared (TIR) sensor was based on the Bayesian theory of total probability (Merchant et al., 2010) and the naive Bayes algorithm for AGRI (Yan , et al., 2022). The unsupervised clustering cloud detection algorithms for MERIS (Medium Resolution Imaging Spectrometer) (GomezChova , et al., 2007) and the fuzzy C-means clustering algorithms for MODIS (Pan, et al., 2009) all have achieved high accuracy in cloud detection.

More and more machine learning algorithms are being utilized by researchers in cloud detection studies with the development of machine learning. For instance, the probabilistic neural networks, especially radial basis function networks was used for AVHRR cloud detection (Zhang, et al., 2001). The utilization of convolutional neural network methods (Hu, et al., 2020) offers important perspectives for cloud detection research.

Currently, there is limited research literature on cloud detection and cloud fraction retrieval algorithms for FY-4A/4B AGRI. The operational cloud fraction product of FY-4A AGRI utilized a threshold method with 4 km spatial resolution. Differences in climatic and environmental factors lead to varying albedo and brightness temperature observations for the instrument at different times and locations. Therefore, the choice of thresholds is easily influenced by factors such as season, latitude and land surface type (Gao and Jing, 2019). Using multiple sets of thresholds for discrimination would significantly slow down the cloud detection process. Moreover, most algorithms focus solely on cloud detection, which classified the observed scenes into cloud or clear-sky without providing the specific cloud fraction information for the scenes. The use of active remote sensing instruments carried by Cloudsat & Calypso is not influenced by thresholds when retrieving cloud fraction, enabling a more accurate cloud fraction retrieval. However, due to Cloudsat & Calypso being polar-orbiting satellites, the cloud fraction over the full disk cannot be obtained. Utilizing the Cloudsat & Calypso Level 2 product 2B-CLDCLASS-LIDAR as the reference truth, a random forest model trained based on FY4A AGRI full disk radiation data can address the shortcomings of threshold methods and achieve a high accuracy of cloud fraction over the full disk.

In summary, this paper established cloud detection and cloud fraction retrieval models using a Multi-Layer Perceptron (MLP) and Random Forest (RF), based on FY-4A AGRI full-disk level 1 observed radiance data. The cloud fraction from the CloudSat & CALIPSO level 2 product 2B-CLDCLASS-LIDAR was used as the label. The results were compared with the 2B-CLDCLASS-LIDAR product and the official AGRI operational products for validation.

1 Research Data and Preprocessing

1.1 FY-4A data

FY-4A was successfully launched on December 11, 2016. Starting from May 25, 2017, FY-4A drifted to a position near the main business location of the Fengyun geostationary satellite at 104.7 degrees east longitude on the equator. Its successful launch marked the beginning of a new era for China's next-generation geostationary meteorological satellites as an advanced comprehensive atmospheric observation satellite. The Advanced Geosynchronous Radiation Imager (AGRI), one of the main payloads of the Fengyun-4 series geostationary meteorological satellites, can perform large-disk scans and rapid regional scans at a minute level. It has 14 observation channels in total with the main task of acquiring cloud images. The channel parameters and main uses of AGRI are detailed in Table 1 (https://www.nsmc.org.cn/nsmc/cn/instrument/AGRI.html). The first six visible light channels have no values at night, meaning that channels with a central wavelength less than or equal to 2.225μm are unavailable during nighttime. FY-4A AGRI data was downloaded from the official website of the China national satellite meteorological center (http://satellite.nsmc.org.cn), including level-1 full disk radiation observation data preprocessed through quality control, geolocation and radiation calibration as well as level-2 cloud fraction product (CFR). The spatial resolution of these data is all 4 km at nadir and the temporal resolution is 15 minutes.

Table 1 FY-4A AGRI channel parameters

Channel Number	Band Range /μm	Central Wavelength /μm	Spatial resolution/km	Main Applications
1	0.45 ~ 0.49	0.47	1	clouds, dust, aerosols
2	0.55 ~ 0.75	0.65	0.5	clouds, sand dust, snow
3	0.75 ~ 0.90	0.825	1	vegetation
4	1.36 ~ 1.39	1.375	2	cirrus
5	1.58 ~ 1.64	1.61	2	clouds、snow
6	2.10 ~ 2.35	2.225	2	cirrus、aerosols
7	3.50 ~ 4.00	3.75H	2	fire point, the intense solar reflection signal
8	3.50 ~ 4.00	3.75L	4	low clouds, fog
9	5.80 ~ 6.70	6.25	4	upper-level water vapor
10	6.90 ~ 7.30	7.1	4	mid-level water vapor
11	8.00 ~ 9.00	8.5	4	subsurface water vapor
12	10.30 ~ 11.30	10.8	4	surface and cloud-top temperatures
13	11.5 0~ 12.50	12.0	4	surface and cloud-top temperatures
14	13.2 ~ 13.8	13.5	4	cloud-top height

1.2 CloudSat & Calipso Cloud Product

CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations) is a satellite jointly launched by NASA and CNES (the French National Center for Space Studies) in 2006. It is a member of the A-Train satellite observation system. CALIPSO is equipped with three payloads, among which CALIOP (the Cloud and Aerosol Lidar with Orthogonal Polarization) is a primary observational instrument. Observing with dual wavelengths (532 nm and 1064 nm) CALIOP can provide high-resolution vertical profiles of clouds and aerosols with 30 m vertical resolution. As the first satellite designed to observe global cloud characteristics in a sun-synchronous orbit

CloudSat is also among NASA's A-Train series satellites. The CPR (Cloud Profile Radar) installed on it operates at 94 GHz millimeter-wave and is capable of detecting the vertical structure of clouds and providing vertical profiles of cloud parameters. The scanning wavelengths of CPR and CALIOP are different. CALIOP is capable of observing the top of mid-to-high level clouds, whereas CPR can penetrate optically thick clouds. Combining the strengths of these two instruments enables the acquisition of precise and detailed information on cloud layers and cloud fraction.

The joint level 2 product 2B-CLDCLASS-LIDAR is mainly utilizing in this study. It provides the cloud fraction at different heights with horizontal resolution 2.5 km (along-track) × 1.4 km (cross-track) through combining the observations from CPR and CALIOP. Since the two instruments have different spatial domain such as vertical resolution, spatial resolution and spatial frequency, the spatial domain of the output products is defined in terms of the spatial grid of the CPR. In the algorithm, the cloud fraction is calculated using a weighted scheme based on the spatial probability of overlap between the radar and lidar observations. The calculation of the lidar cloud fraction within a radar footprint is represented by the equation 1(Mace, G. G., et al, 2007):

$$C_l = \frac{\sum_{i=1}^{\#\,of\,lidar\,obs} w_i \delta_i}{\sum_{i=1}^{\#\,of\,lidar\,obs} w_i} \tag{1}$$

Where:

C_l represents the lidar cloud fraction within a radar footprint.

w_i is the spatial probability of overlap for a particular lidar observation.

δ_i indicates the lidar hydrometeor occurrence, where a value of 1 signifies the presence of hydrometeor and 0 indicates the absence.

i counts the lidar profile in a specific radar observational domain.

This calculation considers the contributions of multiple lidar observations within a radar resolution volume to determine the cloud fraction within that volume.The CloudSat product manual (Wang, 2019) can be referred for more detailed information on 2B-CLDCLASS-LIDAR. The data used is available to download from the ICARE data and services center (https://www.icare.univ-lille.fr/data-access/data-archive-access/).

1.3 Establishment of Training Data

The crucial aspect of establishing a training data in machine learning algorithms is how to obtain the cloud fraction values (ground truth) as labels. The error in cloud fraction retrieved solely from passive remote sensing instruments is significant. Using active remote sensing data can provide more accurate cloud fraction information in the vertical direction. Therefore, the spatiotemporally matched 2B-CLDCLASS-LIDAR cloud fraction are utilized as output labels in this paper.

The FY-4A AGRI and 2B-CLDCLASS-LIDAR data with a spatial difference between fields of view within 1.5 km and a time difference within 15 minutes are spatiotemporal matched. To make the 2B-CLDCLASS-LIDAR cloud fraction data collocated within AGRI pixels more effective, at least two 2B-CLDCLASS-LIDAR pixels are required within each AGRI field of view. The cloud fraction average of these pixels is used as the cloud fraction for that AGRI pixel. However, the errors in the matched dataset are unavoidable. The AGRI scanning method operates from left to right and top to bottom. Each complete scan of the full disk takes 15 minutes and generates a dataset. It is impossible to determine the exact moment of a specific point within the full disk. This limits the time range for matching datasets to within 15 minutes. However, in areas with higher wind speeds, clouds can move a significant distance within that 15-minute window. Therefore, errors arising from timing issues cannot be avoided.

Cloud detection and cloud fraction label generation for 2B-CLDCLASS-LIDAR are as follows. There may be multiple layers of clouds in each field of view. If there is at least one layer cloud with cloud fraction of 1 in the 2B-CLDCLASS-LIDAR profile, then the scene is labeled as overcast with a cloud fraction of 1. If all layers in the profile are cloud-free, the scene is labeled as clear sky. The scene between the above two situations is labeled as partly cloudy and the cloud fraction is the average of cloud fractions at different layers.

The algorithm includes two steps: the cloud detection is conducted firstly for each AGRI field of view to identify whether it is clear sky, partly cloudy or overcast within the observation field. Then the cloud fraction is retrieved for the scene identified as partly cloudy. So the training data include A dataset used for cloud detection and B dataset for cloud fraction retrieval. The input variables in A dataset are the FY-4A

AGRI level-1 radiative observations from 14 channels and the output variable is the temporally and spatially matched 2B-CLDCLASS-LIDAR cloud detection label. The output is categorized into three types: overcast, partly cloudy and clear sky with values 1, 2 and 3 respectively. The cloud fraction product from 2B-CLDCLASS-LIDAR consists of discrete values: 0, 0.16, 0.33, 0.50, 0.66, 0.83, and 1. Here, 0 indicates clear sky, values from 0 to 1 represent varying cloud fractions for partly cloudy conditions, and 1 signifies overcast. To ensure the balance and representativeness of the samples, the proportions of different cloud fraction samples in dataset A are set at 5:1:1:1:1:1:5. Regarding the samples for partly cloudy type in dataset A, the collocated 2B-CLDCLASS-LIDAR cloud fraction products serve as output labels for cloud fraction retrieval model B. The input of training dataset B remains the FY-4A AGRI level-1 radiative observations.

Due to the instrument's limited lifespan, only 2B-CLDCLASS-LIDAR data up to August 2019 can be obtained. The sample time range used in this paper is from August 2018 to July 2019. Five days were randomly selected each month as daytime samples and five days as nighttime samples. A total of 120 days of time and space matched FY-4A AGRI full-disk observations and 2B-CLDCLASS-LIDAR data were used as training and testing samples. Among them, 80% of the data was used for training, and 20% was used for testing. The total number of daytime samples in dataset A is 91,073, while dataset B contains 30,358 samples. The total number of nighttime samples in dataset A is 95,493, and dataset B includes 31,831 samples.

Although the model was trained and tested using data from 2018 to 2019, to test the universality of the algorithm, it was applied to real-time observations from FY-4A and FY-4B AGRI in 2023.

2 Algorithms

Our preliminary experiments involved multiple algorithms, including LibSvm, MLP, BP neural network, and Random Forest. These experiments highlighted that, among the baselines, Random Forest and MLP achieved the highest overall accuracy. For this reason, we selected them to perform additional experiments.

2.1 Random Forest (RF)

This algorithm integrates multiple trees based on the Bagging idea of ensemble learning, with the basic element being the decision tree (Breiman, 1999). When building a decision tree, N sets of independent and dependent variables are randomly sampled with replacement from the original training samples to create a new training sample set; m variables are randomly sampled without replacement from all independent variables, the dependent variable data is split into two parts using the selected variables, and the purity of the subsets is calculated for each split method. The variable utilized by the split method with the highest purity is used to partition the data, completing the decision at that node. This process of binary splitting continues to grow the decision tree until stopping criteria are met, completing the construction of a single decision tree. These steps are repeated Ntree times to build a random forest model consisting of Ntree decision trees (Breiman, 2001). Random Forest adopts ensemble algorithms, with the advantage of high accuracy. It can handle both discrete and continuous data, without the need for normalization, making it more efficient compared to other algorithms.

2.2 Multilayer Perceptron (MLP)

This algorithm consists of a fully connected artificial neural network(Duda, et al., 2001). The classifier/regressor takes feature vectors or tensors as input. The input is mapped through multiple fully connected hidden layers containing hidden weights, which produce classifications/regressions at the output layer. A nonlinear activation function (such as sigmoid or rectified linear unit (ReLU)) is applied in each hidden layer to facilitate a nonlinear model. For classifiers, the output of the final hidden layer is combined and passed through a softmax function to generate class predictions. The model's weights are trained in a supervised manner, utilizing stochastic gradient descent and backpropagation to achieve the desired classification/regression.

2.3 Hyperparameters

In this paper, a total of eight models were established, including daytime/nighttime random forest classification/regression models and daytime/nighttime MLP classification/regression models. For the random forest, we first conducted experiments using the following Hyperparameters ranges: Trees: [200, 300, 400, 500, 600,700], minleaf: [1, 2, 5, 10], criterion: [Gini, entropy]. Ultimately, the best selections were: (1) Daytime RF classification model: Trees=500, minleaf=1, criterion=gini; (2) Nighttime RF classification model: Trees=600, minleaf=1, criterion=gini; (3) Daytime RF regression model: Trees=400, minleaf=1, criterion=gini; (4) Nighttime RF regression model: Trees=500, minleaf=1, criterion=gini.

For the MLP, experiments were conducted using the following hyperparameter ranges: Hidden layer size: [2,3,4,5,6,7,8,9], Hidden layer neuron count: [8,16,32,64,128], Activation hyperparameter: [logistic, tanh, relu], MaxEpochs: [30,50,100], MiniBatchSize: [300,400,...,1500,1600], Solver hyperparameter: [lbfgs, sgd, adam]. The optimal parameters found are as follows: (1) MLP classification model for daytime: hidden layer size = 5, MiniBatchSize = 1500. (2) MLP classification model for nighttime: hidden layer size = 7, MiniBatchSize = 800. (3) MLP regression model for daytime: hidden layer size = 4, MiniBatchSize = 600. (4) MLP regression model for nighttime: hidden layer size = 6, MiniBatchSize = 500. All four models have hidden layer neuron count = 64, activation = relu, MaxEpochs = 50, solver = adam, InitialLearnRate = 0.01, LearnRateSchedule = piecewise, LearnRateDropFactor = 0.1, LearnRateDropPeriod = 10.

3 Results and Analysis

To assess the accuracy and stability of the retrieval model, two types of validation methods are utilized. One way involves a direct comparison from images, qualitatively comparing the model's retrieval results and official cloud fraction products with AGRI observed cloud images. Another approach uses 2B-CLDCLASS-LIDAR as the ground truth and introduces five parameters for quantitative comparison: recall, false alarm rate (FAR), mean error (ME), mean absolute error (MAE), and root mean square error (RMSE). To evaluate the ability of operational products, RF, and MLP cloud detection models to distinguish overcast, partly cloudy, and clear sky, the recall is calculated using the formula POD=TP/(TP+FN), and the false alarm rate is calculated using the formula FAR=FP/(TP+FP). Taking the overcast scene as an example, TP represents the number of correctly identified overcast conditions, FN represents the number of overcast conditions misidentified as partly cloudy or clear sky, and FP represents the number of clear sky or partly cloudy conditions misidentified as overcast. When assessing the accuracy of operational products and cloud fraction models for the cloud fraction retrieval results of partly cloudy scenes, mean error (ME), mean absolute error (MAE), and root mean square error (RMSE) are used.

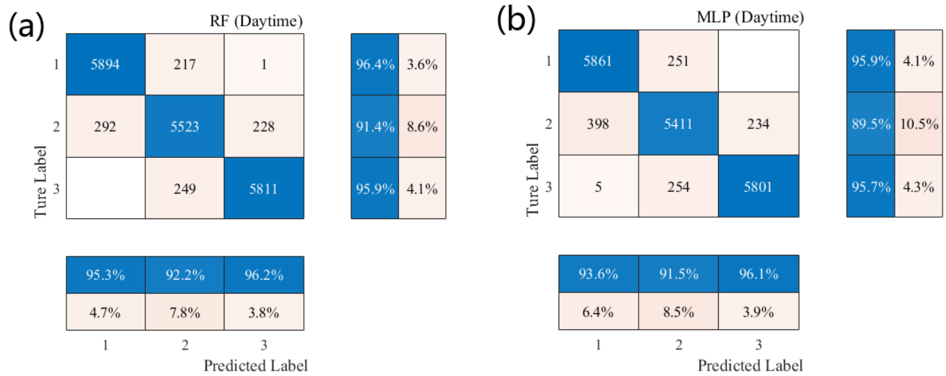
3.1 Objective Analysis of Cloud Fraction Retrievals

First, using the 2B-CLDCLASS-LIDAR cloud fraction product as the ground truth, we calculated the accuracy of the operational cloud detection products. The results are shown in Table 2. The samples used for this statistic are the same as those for testing the model below (20% of dataset A).

Table 2: Recall Rate and FAR of Operational Cloud Detection Products

	Sky Classification	Daytime Product	Nighttime Product
	Clear Sky	0. 6359	0.5781
POD	Partly cloudy	0.7174	0.7449
	Overcast	0.7736	0.7384
	Clear Sky	0.1778	0.0934
FAR	Partly cloudy	0.1819	0.2117
	Overcast	0.2499	0.2683

Based on the cloud detection model trained above, cloud detection experiments were conducted using the test samples from Dataset A. The time-space matched 2B CLDCLASS-LIDAR cloud fraction product served as the ground truth to assess the accuracy of cloud detection. Figure 1 shows the results: (a) Random Forest model results during the day, (b) MLP model results during the day, (c) Random Forest model results during the night, and (d) MLP model results during the night. The x-axis represents the model predictions, while the y-axis represents the ground truth. A value of 1 on both axes indicates clear skies, 2 indicates partly cloudy, and 3 indicates overcast. The blue area on the right side of each plot shows the recall rate for each type, while the light-colored area at the bottom represents the False Alarm Rate (FAR). During the day, the Random Forest model achieved an overall accuracy of 94.2%, while the MLP model had an overall accuracy of 93.4%. The Random Forest model exhibited slightly higher recall rates for clear skies, partly cloudy, and overcast conditions compared to the MLP model, and its FAR was lower as well. Both models performed poorly in recognizing partly cloudy conditions, as the models tended to classify true cloud fractions of 0.16 as clear skies and those of 0.83 as overcast. At night, the Random Forest model achieved an overall accuracy of 89.4%, while the MLP model had an accuracy of 87.7%. The Random Forest model had higher recall rates for clear skies and partly cloudy conditions compared to the MLP, while the recall rates for overcast conditions were similar for both models. The FAR for the Random Forest model was lower than that of the MLP. Overall, both the Random Forest and MLP models showed higher classification accuracy for clear skies, partly cloudy, and overcast conditions compared to operational products, with the Random Forest model performing better.



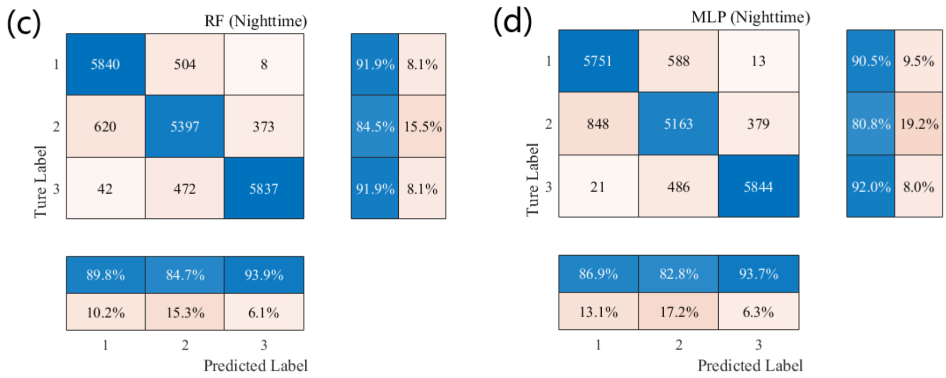


Figure 1 Model Cloud Detection Accuracy: (a) Daytime RF, (b) Daytime MLP, (c) Nighttime RF, (d) Nighttime MLP (In the axis, 1 represents clear sky, 2 represents partly cloudy, and 3 represents overcast.)

Based on the previous model's assessment of the field of view as partly cloudy, the cloud fraction in this AGRI field of view is retrieved using the cloud fraction model established earlier. For model evaluation, both the operational product and the 2B-CLDCLASS-LIDAR cloud fraction product are classified as partly cloudy, with the matched 2B-CLDCLASS-LIDAR cloud fraction product considered as the ground truth. The average error, mean absolute error, and root mean square error for both daytime and nighttime operational products (Table 3) and cloud fraction model retrieval (Table 4) are calculated. It can be observed that the average errors of both models are close to 0 during both daytime and nighttime. The errors are smaller during the day than at night, with the RF model exhibiting lower errors than the MLP model. In summary, the errors of both models are smaller than those of the operational products, and the RF model performs better in the cloud fraction retrieval task.

Table 3: Errors of Operational Product Cloud Fraction

	Daytime Operational Product	Nighttime Operational Product
ME	0.1987	0.2121
MAE	0.2279	0.2441
RMSE	0.2776	0.2938

Table 4:Model Retrieval Error

	Daytime RF	Daytime MLP	Nighttime RF	Nighttime MLP
ME	0.0006	-0.0009	-0.0028	-0.0032
MAE	0.1011	0.1053	0.1221	0.1322
RMSE	0.1285	0.1332	0.151 0	0.1623

Based on the experiments mentioned above, the performance of RF in cloud detection and cloud fraction retrieval slightly outperforms that of MLP. Therefore, subsequent experiments will utilize the RF algorithm.

3.2 Cloud fraction correction in sun glint regions

Sun glint refers to the bright areas created by the reflection of sunlight to the sensors of observation systems (satellites or aircrafts). This phenomenon usually occurs on extensive water surfaces, such as oceans lakes or rivers. This specular reflection of sunlight will cause an increase in the reflected solar radiation received by onboard sensors, manifested as an enhancement of white brightness in visible images. The increase in visible channel observation albedo will affect various subsequent applications of data, including cloud detection and cloud cover retrieval, etc.

The position of Sun glint area can be determined using the SunGlintAngle value in the FY-4A GEO file. SunGlintAngle is defined as the angle between the satellite observation direction or reflected radiation direction and the mirror reflection direction on a calm surface (horizontal plane). It is generally accepted that the range of SunGlintAngle < 15° is easily affected by sun glint (Kay S, et al., 2009). The positions of the SunGlintAngle contour lines at 5 and 15° are marked in Figure 1(a). It can be observed that the edge of sun glint in Figure 1(a) essentially overlaps with the position of SunGlintAngle = 15°. Thus, the region where SunGlintAngle < 15° is defined as the sun glint range in this paper and only the cloud fraction within this range will be adjusted in the subsequent correction.

To correct the cloud fraction in the sun-glint areas, we first identified the fields of view (FOVs) where sun-glint occurred during FY-4A AGRI observations from August 2018 to July 2019, totaling 1,476 FOVs. Subsequently, a direct least squares fitting was conducted between the retrieved cloud fraction and the collocated 2B-CLDCLASS-LIDAR cloud fraction (ground truth). The scatter plot is illustrated in Figure 2(b), where x-axis is the 2B-CLDCLASS-LIDAR cloud fraction and y-axis is the model-retrieved cloud fraction. The blue line represents the curve (namely Eq.2) fitted by the least squares method between the retrievals and the truths. The thin dash line is the x=y line. It is evident that the retrieved cloud fraction is generally slightly overestimated.

Taking observations at 04:00 on 5 June 2019 as an example, Figure 2(c) presents the distribution of SunGlintAngle and the flight trajectory of the Cloudsat&Calypso satellite. White circles denote the sun glint region with SunGlintAngle < 15° and the white line represents the satellite flight track. As depicted in the figure, the majority of Cloudsat&Calypso flight trajectories do not pass through the central position of sun glint area but instead traverse locations with larger SunGliantAngle values. The intensity of sun glint effect decreases with the increase of SunGliantAngle. This suggests that the true values for spatial and temporal matching mostly do not fall within the strongest sun glint region. From Figure 2(d), it can be seen that the impact of sun glint becomes stronger as SunGlintAngle decreasing, which results in a higher observation albedo. This further leads to the overestimated cloud fraction values in the retrieval. It is evident that the cloud fraction error is related to the value of SunGlintAngle and this influence is not considered in Eq. (2). Directly applying equation (2) to correct the cloud fraction retrievals would result in a too small correction intensity for the FOVs near the center of sun glint and an excessively large correction intensity for the FOVs in the Sun-glint edge region (even erroneous clear sky may appear). Considering this, a correction formula (3)-(4) using SunGlintAngle as weight is introduced, where W_i represents the angle weight for a certain pixel *i* in the sun glint region, n is the number of pixels within the SunGlintAngle < 15° range, yi is the initial model retrieval of cloud cover for the field of view *i* and x_i is the final corrected cloud fraction.

$$x = (y - 0.2441)/0.8092 \tag{2}$$

$$W_i = \frac{glintangle_i}{\frac{1}{n}\sum_{i=0}^{n} glintangle_i} \tag{3}$$

$$x_i = W_i \left(\frac{y_i - 0.2441}{0.8092}\right) \tag{4}$$

Figure 2(d) shows the distribution of errors with respect to SunGlintAngle, where the blue dots represent the error distribution corrected using formula (2), and the orange dots represent the error distribution corrected using formula (4). It can be seen from Figure 2(d) that after correction by formula (4), the errors in the smaller range of SunGlintAngle are significantly reduced.

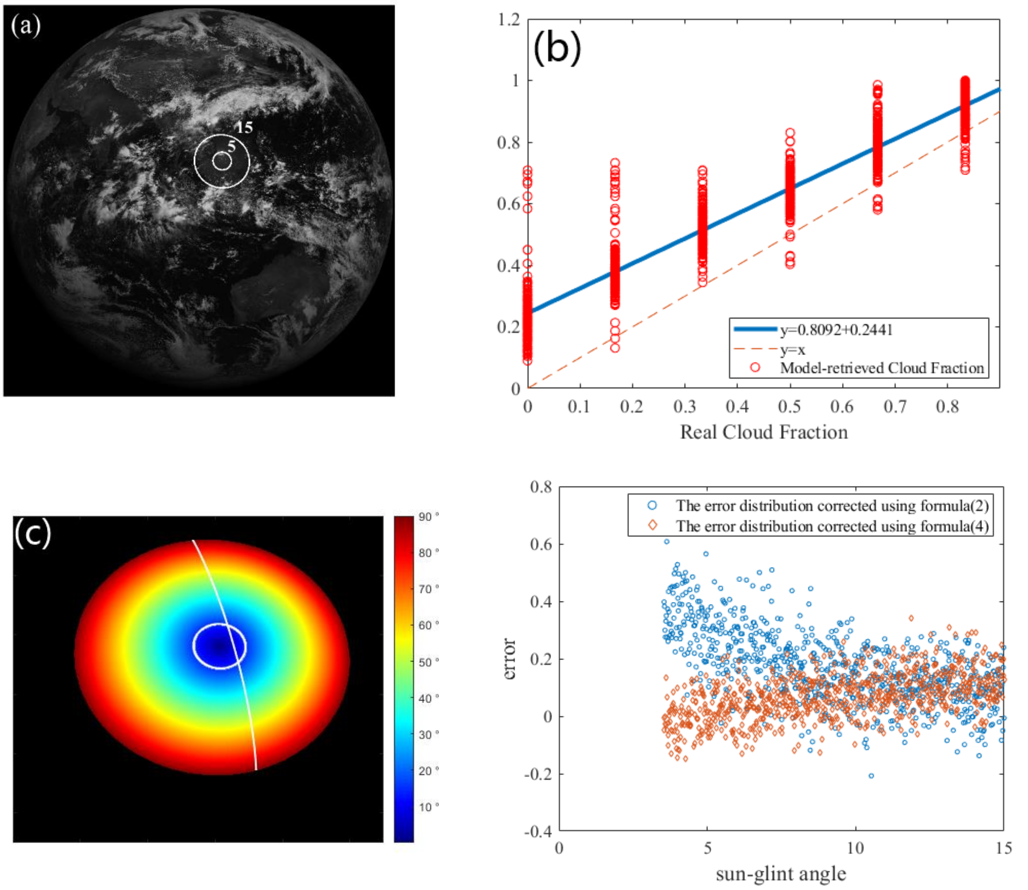


Figure 2 (a) albedo image of 0.67μm channel (the circles are the contours of the sun-

glint angle), (b) Scatter plot of cloud fraction in sun glint region (The blue line represents the curve (namely Eq.2) fitted by the least squares method between the retrievals and the truths.), (c) Distribution of SunGlintAngle and satellite flight track of CloudSat & Calypso at 4:00 on June 5, 2019, (d) Distribution of cloud fraction retrieval error with sun-glint angle.

3.3 Algorithm universal applicability testing

Although the retrieval model in this article was built based on data from May 2019 due to the limited lifespan of the instrument, how effective is it in real-time FY-4A AGRI observations and even subsequent FY-4B AGRI applications? The algorithm's universal applicability was tested using real-time observations from FY-4A and FY-4B AGRI in 2023.

Taking the full-disk observation of FY-4A AGRI at 04:00 (UTC, the same below) on 1 June 2023 as an example, the radiance observations from 14 channels are initially fed into the random forest cloud detection model to determine the sky classification (overcast, partly cloudy or clear sky) in each AGRI field. The random forest cloud fraction retrieval model is utilized to retrieve the cloud fraction in scenes identified as partly cloudy. Figure 3(a) is the observed albedo at 0.67 μm, where the circles represent the contours of the sunglint angle, (b) is the cloud fraction retrievals from random forest algorithm, (c) is the official operational cloud fraction product and (d) is random forest cloud fraction retrievals with sun-glint correction. It can be seen from Figure 3 that many clear-sky scenes are erroneously identified as cloudy by the operational product and the cloud fraction is generally overestimated with many scenes having a cloud fraction of 1. The random forest algorithm identifies more regions as clear skies or partly cloudy than the operational products, matching better with the observations in the 0.67 μm albedo image. Brighter regions in the visible image correspond to cloud cover areas and darker areas represent clear sky conditions. The sun glint region in the central South China Sea (the circled area in Figure 3(a)) is depicted in Figure 3(b), where the clear-sky scenes over the ocean are misidentified as partly cloudy by random forest algorithm due to the increase in observed albedo. Although operational product in this area also suffers from the impact of unremoved sun glint, it identifies more clear-sky scenes and the cloud fraction is relatively low. Thus, it is evident that the random forest algorithm exhibits significant cloud detection and cloud fraction errors in these sun glint regions. Correction is necessary for the cloud fraction retrievals in the sun glint region.

Figure 3(d) shows the cloud fraction distribution after correction using equation (9) in the sun glint region., The correction eliminates the influence of sun glint comparing to the cloud fraction in sun glint area before correction in Figure 3(b). The scenes misjudged as partly cloudy are corrected to clear sky and match well with the actual albedo observations in 3(a), which accurately restores the true cloud coverage over the South China Sea.

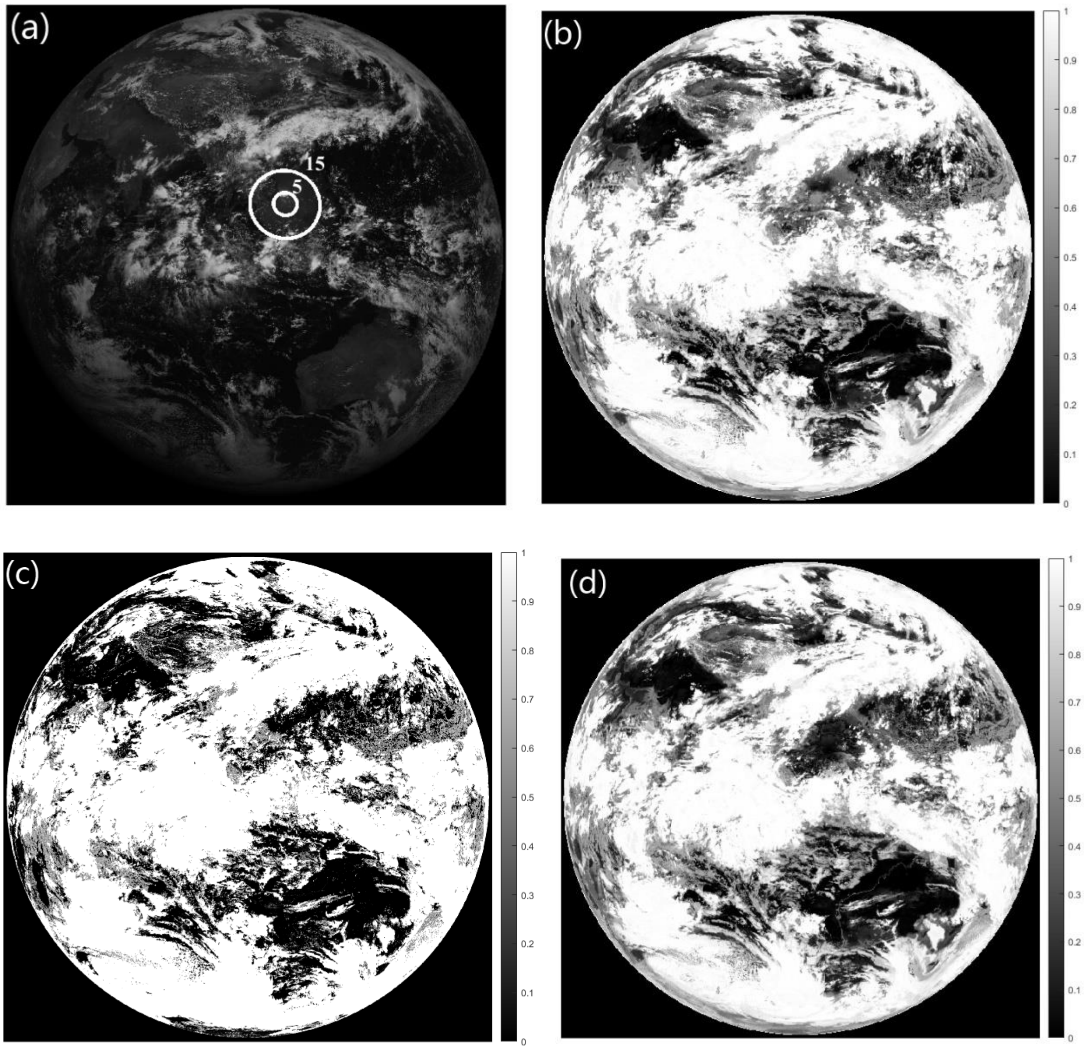


Figure 3 FY-4A AGRI at 04:00 on 1 June 2023 (a) albedo image of 0.67μm channel (the circles are the contours of the sun-glint angle), (b) random forest cloud fraction retrieval without sun-glint correction, (c) operational cloud fraction product, (d) random forest cloud fraction retrieval with sun-glint correction.

Statistical analysis was conducted on the correction effect using samples with sun glint in the training data. The POD and FAR in sun glint area is listed in table 5 and the error is in table 6. It can be seen that after correcting for cloud fraction, the POD for clear skies has increased from 0.0987 to 0.9023. The FAR for partly cloudy has decreased from 0.7943 to 0.0276. Both ME, MAE, and RMSE show significant reductions, and the results after correction outperform operational products.

Table 5 POD and FAR of Cloud Detection in sun glint area

	Sky Classification	Operational Product	RF	RF after Correction
POD	Clear Sky	0.4120	0.0987	0.9023
	Partly cloudy	0.7371	0.9663	0.9587
	Overcast	0.8856	0.9845	0.9845
FAR	Clear Sky	0.1229	0.1633	0.0938
	Partly cloudy	0.3332	0.7943	0.0276
	Overcast	0.2983	0.1321	0.1321

Table 6 cloud fraction Errors in sun glint area

	Operational Product	RF Retrievals	RF after Correction
ME	0.2354	0.1741	0.0670
MAE	0.2511	0.1820	0.0849
RMSE	0.2771	0.2166	0.1041

FY-4B launched in 2021 has a total of 15 channels with an additional low-level water vapor channel at 7.42 μm compared to FY-4A. Taking the full-disk observation of FY-4B AGRI at 17:00 on April 18, 2023, as an example, The radiance observation data of the remaining eight channels (near-infrared and infrared channels) except for the 7.42 μm channel and the visible light channels were input into the random forest cloud detection model. Figure 4 (a) shows the brightness temperature distribution observed in the 10.8 μm channel of FY-4B AGRI, (b) represents the operational cloud fraction product for FY-4B AGRI and (c) shows the cloud fraction retrieved by this algorithm. Figure 4 illustrates that the random forest algorithm identifies more regions as clear skies or partly cloudy than the operational products, aligning better with the brightness temperature observations in 10.8 μm. Especially in high latitude regions of the southern hemisphere and areas with strong convection near the equator, the cloud cover provided by operational products is too high and even misjudged. It can be seen that the random forest algorithm is also suitable for cloud fraction retrieval of FY-4B AGRI.

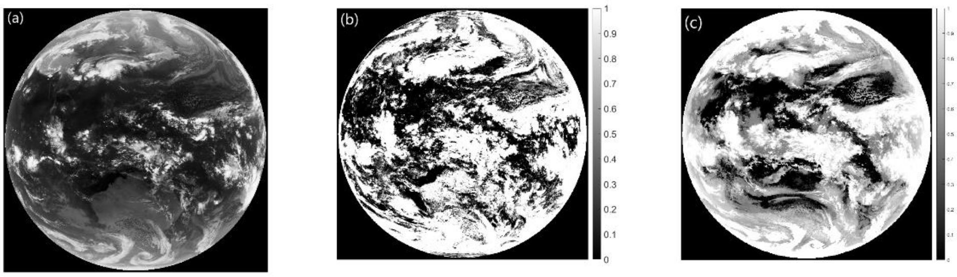


Figure 4 FY-4B AGRI at 17:00 on 18 April 2023, (a) brightness temperature of 10.8μm channel, (b) operational cloud fraction product, (c) random forest cloud fraction retrieval.

4 Conclusion

This paper used the random forest and multi-layer perceptron (MLP) algorithms to retrieve cloud fraction from FY-4A AGRI full-disk Level-1 radiance observation data, and verified the accuracy of the algorithms using the Cloudsat & Calypso active remote sensing satellite's 2B CLDCLASS-LIDAR cloud fraction product. The following conclusions were drawn:

(1) The random forest and MLP algorithms performed well in cloud detection and cloud fraction retrieval tasks, and their accuracy was higher than that of operational products. The accuracy of cloud detection can reach over 93%, and the error of cloud fraction retrieval is close to zero. Compared with the MLP algorithm, the RF algorithm has a slightly higher accuracy in cloud detection, and a slightly lower error in cloud fraction retrieval, showing better performance.

(2) At night, the classification accuracy is lower than during the day due to the lack of observations in the visible channel of AGRI, resulting in higher cloud fraction errors at night.

(3) The accuracy of identifying partly cloudy scenes is lower than that of identifying clear sky and overcast scenes for both RF and MLP algorithms. Scenes with very low cloud fraction (0.16) are often misclassified as clear sky, while scenes with high cloud fraction (0.83) are often misclassified as overcast.

(4) The sun-glint area cloud fraction correction curve, fitted with SunGlintAngle

as the weight, greatly improves the accuracy of cloud fraction retrieval and reduces the misclassification rate of clear sky scenes as partly cloudy or partly cloudy scenes as overcast due to increased reflectance.

Data availability

FY-4A AGRI data is available at http://satellite.nsmc.org.cn and the 2B-CLDCLASS-LIDAR data at https://www.icare.univ-lille.fr/data-access/data-archive-access/

Author contributions

JX: Formal analysis, Methodology, Software, Visualization and Writing – original draft preparation. LG: Conceptualization, Data curation, Funding acquisition, Supervision, Validation and Writing – review & editing.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Disclaimer

Acknowledgements

Funding: This work was supported by the National Natural Science Foundation of China under grant no. 41975028.

We acknowledge the High Performance Computing Center of Nanjing University of

Information Science & Technology for their support of this work.

References

Amato, U., Antoniadis, A., Cuomo, V., Cutillo, L., Franzese, M., Murino, L., Serio, C.: Statistical cloud detection from SEVIRI multispectral images, *Remote Sensing of Environment*, 112, 750–766, https://doi.org/10.1016/j.rse.2007.06.004, 2008.

Baum, B., Trepte Q.: A Grouped Threshold Approach for Scene Identification in AVHRR Imagery, *Journal of Atmospheric & Oceanic Technology*, 16, 793-800, https://doi.org/10.1175/1520-0426(1999)016<0793:AGTAFS>2.0.CO;2, 1999.

Breiman L.1999. Random Forests-Random Features [J]. Machine Learning.45(1): 5-32.

Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001). doi.org/10.1023/A:1010933404324

Merchant, C.J., Harris, A.R., Maturi, E., Maccallum S.: Probabilistic physically based cloud screening of satellite infrared imagery for operational sea surface temperature retrieval, *Quarterly Journal of the Royal Meteorological Society*, 131, 2735-2755, https://doi.org/10.1256/qj.05.15, 2005.

Gao, J., Jing, Y.: Satellite Remote Sensing Cloud Detection Method Based on Fully Convolutional Neural Network,*Infrared Technology*, 41, 607-615, 2019.

Gomez-Chova, L., Camps-Valls, G., Amoros-Lopez, J., Guanter, L., Alonso, L., Calpe, J., Moreno, J.: New Cloud Detection Algorithm for Multispectral and Hyperspectral Images: Application to ENVISAT/MERIS and PROBA/CHRIS

Sensors*, IEEE International Symposium on Geoscience and Remote Sensing,* 2757–2760, doi:10.1109/igarss.2006.709, 2006.

Hu, J.: Research on Cloud Detection Algorithm of Remote Sensing Image Based on Convolution Neural Network, *Nanjing University of Information Science and Technology*. doi:10.27248/d.cnki.gnjqc.2020.000625, 2020.

Kay, S., Hedley, J., Lavender, S.: Sun Glint Correction of High and Low Spatial Resolution Images of Aquatic Scenes: a Review of Methods for Visible and Near-Infrared Wavelengths, Remote Sensing, 1, 697-730, https://doi.org/10.3390/rs1040697, 2009.

Kegelmeyer, W.P.J.: Extraction of cloud statistics from whole sky imaging cameras,1994.

Kong, Y.-L., Huang, Q., Wang, C., Chen, J., Chen, J., & He, D. (2018). Long Short-Term Memory Neural Networks for Online Disturbance Detection in Satellite Image Time Series. *Remote Sensing*, 10(3), 452. doi:10.3390/rs10030452

Mace, G. G., R. Marchand, Q. Zhang, et al. (2007). CloudSat Project: Level 2 Radar-Lidar GEOPROF product process description and interface control document. Jet Propulsion Laboratory.

Pan, C., Xia B., Chen, Y.: Research on MODIS Cloud Detection Algorithms Based on Fuzzy Clustering, *Microcomputer Information*, 25, 124-125+131, 2009.

Yan J, Guo X, Qu J, Han M. An FY-4A/AGRI cloud detection model based on the naive Bayes algorithm. Remote Sensing for Natural Resources, 34(3): 33-42. doi:

10.6046/zrzyyg.2021259. 2022

Rossow, W. B., Leonid, C.G.: Cloud detection using satellite measurements of infrared and visible radiances for ISCCP. *Journal of Climate*, 12, 2341-2369, https://doi.org/10.1175/1520-0442(1993)006<2341:CDUSMO>2.0.CO;2, 1993.

R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, New York: John Wiley & Sons, 2001, pp. xx + 654, ISBN: 0-471-05669-3. Journal of Classification 24, 305–307 (2007). https://doi.org/10.1007/s00357-007-0015-9

Solvsteen, C.: Correlation based cloud-detection and an examination of the split-window method, *Proceedings of SPIE - The International Society for Optical Engineering,* 86-97, 1995.

Wang, Z.: CloudSat Project: CloudSat 2B-CLDCLASS-LIDAR product process description and interface control document, *Jet Propulsion Laboratory*, 2019.

Yan, J., Guo, X., Qu, J.: An FY-4A/AGRI cloud detection model based on the naive Bayes algorithm, *Remote Sensing for Natural Resources*, 34, 33-42, 2022.

Zhang, W., He, M., Mak, M.W.: Cloud detection using probabilistic neural networks, *Geoscience and Remote Sensing Symposium*, IEEE 2373-2375, 2001.

Zhang, Y., William, B. R., Andrew, A. L., Valdar, O., Michael, I. M.: Calculation of radiative fluxes from the surface to the top of atmo- sphere based on ISCCP and other global data sets: Refine- ments of the radiative transfer model and the input data, *Journal of Geophysical Research Atmospheres*, 109, 1-27, https://doi.org/10.1029/2003JD004457, 2004.