Cloud Detection from a Hyperspectral Infrared Atmospheric Sounder Using a Machine-Learning Model

First author: Huaxiang Shi [1,a]
Organization:[1] College of Meteorology and Oceanography, National University of Defense Technology, Changsha, China
[a]e-mail: shihuaxiang19@nudt.edu.cn

Co-first author: Yi Yu [2,b]
Organization:[2] College of Meteorology and Oceanography, National University of Defense Technology, Changsha, China
[b]e-mail: yuyi2019@nudt.edu.cn

Third author: Weimin Zhang [3,c*]
Organization:[3] College of Meteorology and Oceanography, National University of Defense Technology, Changsha, China
[*] Corresponding author: [c]weiminzhang@nudt.edu.cn

Fourth author: Gang ma [4,d]
Organization:[4] ational Satellite Meteorological Center, China Meteorological Administration (NSMC/CMA), Beijing, China
[d]-mail: magang@cma.gov.cn

Fifth author: Qi Zhang [5,e]
Organization:[5] College of Meteorology and Oceanography, National University of Defense Technology, Changsha, China
[e]e-mail: zhangq@nudt.edu.cn

Sixth author: Tengling Luo [6,f]
Organization:[6] College of Meteorology and Oceanography, National University of Defense Technology, Changsha, China
[f]e-mail: luotengling@nudt.edu.cn

Huaxiang Shi and Yi Yu are co-first authors of the article

Seventh author: Qunbo Huang [7,g]
Organization:[7] 93110 Troops of People's Liberation Army of China (PLA), Beijing 100843, China
[g]e-mail: hqb09@163.com

***Abstract*— Cloud detection is an essential step in the application of hyperspectral infrared (HIR) data. In this paper, a new cloud detection method using a LightGBM algorithm based on principal component (PC) space is proposed for HIR data from the High Spectral Infrared Atmospheric Sounder (HIRAS). Considering the difference of infrared radiation between ocean and land, day and night, this paper respectively builds the LightGBM day-land, day-sea, night-land and night-sea models. The truth cloud fraction of the HIRAS field of view (FOV) is determined by the collocated cloud masks of the Medium Resolution Spectral Imager-II (MERSI) onboard the FY-3D satellite. The HIRAS infrared channels were transformed into the leading K PCs as predictors through principal component analysis (PCA), with the advantages of reducing the correlation of infrared channels, accelerating model convergence and prediction. The cloudy FOVs in global datasets are randomly down-sampling to alleviate the impact of dataset imbalance. The validation experiments have shown that the LightGBM model has high accuracy for the completely cloudy and completely clear-sky FOVs. However, the partially cloudy FOVs are sometimes detected as the clear-sky. It may be because these partially cloudy FOVs have prominent clear-sky radiation properties. The cloud detection performance of the LightGBM model (accuracy=0.93, HSS=0.85) for land HIR data is better than that of sea HIR data (accuracy=0.89, HSS=0.65). It may be because there are more partially cloudy FOVs in the ocean. In addition, the imbalance of the ocean dataset significantly reduced its Heidke skill score (HSS=0.65). Compared with the HIRAS L1c cloud cover product, the accuracy of the LightGBM model for land (sea) HIR data is increased by about 0.22 (0.07). The time complexities of the algorithms have shown that the cloud detection speed of the LightGBM model is approximately 670 times that of the HIRAS/MERSI collocation cloud detection method. The higher cloud detection accuracy and faster efficiency are helpful to the operational application of the LightGBM cloud detection method.**

Keywords-component; High Spectral Infrared Atmospheric Sounder (HIRAS); cloud detection; Machine Learning; LightGBM; Principal component analysis (PCA); FengYun 3D (FY-3D)

I. INTRODUCTION

The hyperspectral infrared (HIR) atmospheric sounder has thousands of detection channels, providing high resolution three-dimensional atmospheric temperature and humidity structure information with approximately 1km vertical resolution [1,2]. It is currently one of the most critical observational data sources for global numerical weather prediction (NWP) operations centres such as the European Centre for Medium-Range Weather Forecasts (ECMWF). Currently, the hyperspectral infrared atmospheric sounders on operational applications include the Atmospheric Infrared Sounder (AIRS) [3], the Infrared Atmospheric Sounding Interferometer (IASI) [4], the Cross-track Infrared Sounder (CrIS) [5] etc. The HIR data from these sensors significantly impacts global numerical weather prediction (NWP) [6].

978-1-6654-0692-5/21/$31.00 ©2021 IEEE
DOI 10.1109/CISAI54367.2021.00028

Clouds significantly impact infrared radiation, which is a far more substantial effect than that from the uncertainty of atmospheric temperature and composition profiles [7,8]. Cloud detection is a significant data preprocessing step that still is the largest source of uncertainty for the assimilation of HIR data in NWP [9]. The High Spectral Infrared Atmospheric Sounder (HIRAS) is an essential hyperspectral infrared (HIR) sounder onboard the FengYun 3D (FY-3D) satellite. It represents a significant improvement in the Chinese operational infrared sounding capability [10]. It is necessary to develop an accurate and efficient cloud detection algorithm to promote the application of HIRAS data.

There have been many cloud detection methods for HIR data. Goldberg et al. used the satellite observations and model simulation data to determine empirical thresholds used for the clear-sky detection of AIRS over the ocean and land [11]. McNally et al. selected the clear channels that are not affected by clouds based on the difference between the sounder measurements and cloud-free model simulations [7,8]. Both of these cloud detection schemes are affected by model simulation. Lin et al. developed a double CO2 band cloud detection method for CrIS to reduce the influence of the model simulation [9]. However, the double CO2 method cannot detect the partially cloudy FOV well. In addition, cloud detection of HIR data is usually done by collocating the cloud mask of the high-spatial-resolution imager. Li. et al. determined cloud information of the IASI through collocated cloud products of the Moderate Resolution Imaging Spectroradiometer (MODIS) [12]. Eresmaa acquired the cloud information of the IASI FOV based on observation of collocated Advanced Very High-Resolution Radiometer (AVHRR) pixel [13]. A similar method to extend cloud-clearing of CrIS is to use the collocated cloud products and observations from the Visible Infrared Imaging Radiometer Suite (VIIRS) [14]. The imager-based method can provide subpixel cloud information for infrared sounder and not depend on the model's background field error [15]. However, the collocation method relies on the cloud products precision of other sensors and requires high computational resources. Machine learning provides new ideas to improve the accuracy and efficiency of cloud detection from HIR data.

In recent years, machine learning has effectively used cloud detection in satellite remote sensing data [16]. Luo et al. trained a cloud detection model based on logistic regression for IASI data, with four spectral channels as predictors [17]. Then, Zhang. et al. [18] constructed a cloud detection model based on machine learning for HIR data from Geostationary Interferometric Infrared Sounder (GIIRS). Luo and Zhang's machine learning model has high detection accuracy for completely cloudy and completely clear-sky FOVs. However, their algorithm cannot accurately detect the partially cloudy FOVs. The more general partially cloudy FOVs' detection still needs further research. Liu et al. proposed an artificial neural network (ANN) cloud detection model based on principal component space for CrIS HIR data [19]. Liu's ANN model has an overall accuracy of 93% for the cloud detection of CrIS HIR data, while the detection effect of clear-sky FOVs in the unbalanced dataset is not further analyzed. The number of cloudy FOVs is far more than the number of clear-sky FOVs in global HIR data. The imbalance of the dataset will reduce the detection effect of the minority category (the clear-sky FOVs). However, the detection of clear-sky FOVs is more critical for the application of HIR data. The detection effect of the clear-sky FOV in the unbalanced HIR dataset needs to be further improved.

The LightGBM is a widely used ensemble learning algorithm with strong feature extraction capabilities and generalization performance in many classification and regression tasks. Therefore, this paper will build a fast cloud detection model based on the LightGBM algorithm to improve the cloud detection accuracy and efficiency of HIRAS HIR data.

The paper is structured as follows. The HIRAS and MERSI data are presented in Section 2. The cloud detection methods are described in Section 3. The validations of algorithms are examined in Section 4. Section 5 summarize this study.

II. HIRAS AND MERSI DATA

HIRAS measures the IR radiation of the earth-atmosphere system in three spectral bands: the long-wave IR bands covering 650-1135cm^{-1}, the middle-wave IR bands covering 1210-1750cm^{-1}, and the short-wave IR bands covering 2155-2550cm^{-1} with a spectral resolution of 0.625 at full spectral resolution. A cross-track scan has 29 fields of regard (FOR) with 4 fields of view (FOV) on the earth's surface. HIRAS is a cross-track scanning instrument with a maximum scanning angle of 50.4°. The radiometric accuracy of HIRAS has been assessed by comparing the HIRAS observations to radiance simulations and CrIS measurements, showing a high measurement accuracy with low radiation noise [20,21]. The cloud cover products of HIRAS L1c are generated by matching the cloud mask products of MERSI, which are used to verify the accuracy of the cloud detection method based on machine learning.

The MERSI provides radiation observations from 25 spectral bands covering 0.412 μm to 12.0 μm. Its spatial resolution is about 1 km near the sub-satellite point with a maximum scanning angle of 55.1° [22]. The MERSI has higher spatial resolution and larger spatial coverage than HIRAS. The MERSI L2 cloud masks (CLM) have four confidence levels: confidently clear, probably clear, probably cloudy, and confidently cloudy, which are collocated to determine the cloud fraction of HIRAS FOV.

III. METHOD

A. HIRAS/MERSI collocation cloud detection method

A collocation algorithm proposed by Wang et al. [23] is adapted to collocate the HIRAS and MERSI data. The line-of-sight (LOS) is defined as the vector from the sensor location to the measurement pixels position on the Earth surface. In the East, North, Up (ENU) coordinate system, the LOS vectors can be determined as ($x_{ENU}, y_{ENU}, z_{ENU}$),

$$LOS_{ENU} = \begin{pmatrix} x_{ENU} \\ y_{ENU} \\ z_{ENU} \end{pmatrix} = \begin{pmatrix} L \sin\phi \sin\psi \\ L \sin\phi \cos\psi \\ L \cos\phi \end{pmatrix} \quad (1)$$

where ϕ is sensor zenith angle, ψ is sensor azimuth angles, L is satellite range that is the distance between the satellite and the observation point. L can be approximated as,

$$L = H/\cos\phi \quad (2)$$

where H is the satellite orbit height. Since the coordinate origin of ENU varies with the observation position, the LOS_{ENU} is further converted into the Earth-Centered, Earth-Fixed (ECEF) coordinate. The rotation equation is expressed as,

$$LOS_{ECEF} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} -\sin\upsilon & -\cos\upsilon\sin\mu & \cos\upsilon\cos\mu \\ \cos\upsilon & \sin\upsilon\sin\mu & \sin\upsilon\cos\mu \\ 0 & \cos\mu & \sin\mu \end{pmatrix} \begin{pmatrix} x_{ENU} \\ y_{ENU} \\ z_{ENU} \end{pmatrix} \quad (3)$$

where υ and μ are geodetic longitude and latitude. In ECEF, the successful matching of HIRAS and MERSI LOS vectors follow the criterion as,

$$\frac{LOS_{HIRAS} \cdot LOS_{MERSI}}{\|LOS_{HIRAS}\| \, \|LOS_{MERSI}\|} > \cos\left(\frac{1}{2}\eta\right) \quad (4)$$

where the LOS_{HIRAS} and LOS_{MERSI} are LOS vectors of HIRAS and MERSI, the η is FOV angle of HIRAS.

Through the above matching algorithm, the MERSI pixels in each HIRAS FOV can be determined. There are four values for MERSI L2 cloud mask products: confidently clear, probably clear, probably cloudy, and confidently cloudy. To simplify the problem, the confidently clear and probably clear pixels are labelled as the clear, and the probably cloudy and confidently cloudy pixels are marked as the cloudy. The cloud masks of the matched MERSI pixels are used to generate the cloud cover of each HIRAS FOVs, defined as the ratio of the number of cloudy pixels to the total number of pixels. The HIRAS FOV will be classified as clear-sky or cloudy by an appropriate threshold of cloudiness. The threshold varies with different application scenarios. For example, when the proportion of collocated VIIRS cloudy pixels in CrIS FOV is greater than 5%, Wang et al. (2017) and Liu et al. (2020) flagged the CrIS FOV as partially cloudy. A smaller threshold corresponds to a stricter clear condition. Since the spatial resolution of HIRAS FOV (16 km at nadir) is lower than CrIS (14km at nadir), the cloudiness threshold of HIRAS FOV is set to 10%. When the threshold of cloudiness is greater than 10%, the HIRAS FOV is marked as cloudy (label=0). Otherwise, the HIRAS FOV is flagged as clear-sky (label=1). The labels of HIRAS FOV generated by the collocation mechanism are used as the truth for model training and accuracy validation.

B. The machine learning cloud detection method based on PCA

1) Principal component analysis

Due to the strong correlation between the spectral channels of infrared detectors, all spectral channels as input features will cause information redundancy. In addition, the measurement and transmission process of infrared radiations are often affected by instrument noise. Principal component analysis (PCA) is widely used for compression reconstruction and feature transformation of HIR data[24]. The PCA converts the dependent variables into linearly independent principal components (PC) that describe the main variance of the original data. The instrument noises are nearly evenly distributed on each PCs [25]. Therefore, PCA has the advantages of feature transformation, dimensionality reduction and noise reduction.

The HIRAS HIR dataset is represented by N×d matrix R, where N is the number of observation samples and d is the number of infrared spectrum channels. Firstly, to eliminate the influence of the magnitude of radiation, R is normalized:

$$X = (R - \lambda)/\sigma \quad (5)$$

where λ and σ are the mean and standard deviation of each channel of R, respectively. Then the covariance matrix S of the normalized X is computed by:

$$S = \frac{1}{N} X X^T \quad (6)$$

Where T represents the matrix transposition operation. Furthermore, the singular value decomposition (SVD) of covariance matrix S is performed by:

$$S = U \Lambda U^T \quad (7)$$

where the dimensions of the matrices S, Λ, and U are all d×d. The Λ is a diagonal matrix composed of d eigenvalues ($\lambda_1 > \lambda_2 > \cdots > \lambda_d$) of the covariance matrix S. The U is a matrix composed of eigenvectors corresponding to eigenvalues of S, and each eigenvector constitutes the orthogonal unit basis of the principal component space.

After that, the eigenvectors corresponding to the first K (K<d) eigenvalues form the linear transformation matrix U_K. The first K PCs of HIR dataset R are given by:

$$R^K = U_K^T X \quad (8)$$

The PC dataset R^K with a larger K represents more original data information, and the reduction of instrument noise and redundant information is less obvious.

2) The Optimal machine learning algorithm

No machine learning algorithm can be applied to all problems. This paper selects the optimal algorithm from five classic machine learning algorithms for cloud detection of HIRAS data. The five algorithms are logistic regression [26], random forest [27], K-nearest neighbour[28], gradient boosting decision tree (GBDT) [29], and LightGBM [30]. They are widely used in various classification and regression tasks. Only 1% (about 5033 samples) of HIRAS global data are used to select an optimal algorithm to reduce memory and time consumption. The ratio of training and test data is set to 8:2.

All infrared spectrum channels and optical path (1/cos (sensor)) zenith angle)) are used as predictors. The sample label is generated by the HIRAS/MERSI collocation cloud detection method in section 3.1. The Hyperopt optimization algorithm [31] is used to adjust the parameters (as TABLE I) of the five machine learning algorithms to achieve their best cloud detection performance. Hyperopt is a Bayesian optimization method that heuristically searches for the optimal parameters of a machine learning algorithm from a larger parameter space. Hyperopt is often more effective than manual tuning, grid search [32] and random search [33].

TABLE I. PARAMETERS AND SEARCH RANGE OF FIVE MACHINE LEARNING MODELS

Algorithm	**parameters and searching range**			
K Neighbors Classifier	Number of neighbors (‘n_neighbors’): range(1,50)		Algorithm to find neighbors (‘algorithm’): (‘auto’, ’ball_tree’, ’kd_tree’, ’brute’)	
Logistic Regression	The norm of penalization (penalty):(‘l1’,’l2’)		Regularization coefficient (‘C’) uniform (0.01, 5)	
Random Forest	Number of tree (‘n_estimators’): range(50,150,10)	Minimum samples at split node (‘min_samples_split’): range(10,400,10)	minimum samples at leaf node (‘min_samples_leaf’): range(4,20,2)	Maximum depth of tree (‘max_depth’): range(100,300,10)
GBDT	The number of boosting (‘num_iterations’): range(10,250,5)	(‘min_samples_split’): range(10,50,2)	Shrinkage rate (’learning_rate’): uniform(0.01,1)	(‘max_depth’): range(8,60,2)
LightGBM	(‘num_iterations’): range(10,400,10)	Maximum leaves of tree (‘num_leaves’): range(10,400,10)	(‘learning_rate’): uniform(0.001, 0.8)	(‘max_depth’): (10,400,10)

range(start,stop,step): an ordinal sequence between start and stop with step as the interval. uniform(start,stop): the uniform distribution between start and stop.

The cloud detection performance of machine learning algorithms is measured by accuracy (ACC) and average training time (Time, unit: s). The ACC represents the ratio of the number of cloudy and clear-sky samples correctly detected by the machine learning model to the total number of test samples. T means the time of building and optimizing the machine learning model. The cloud detection performance of the five machine learning algorithms under the optimal parameters is shown in TABLE II. The LightGBM algorithm has the highest accuracy (ACC=0.86), and the model training time (T=124s) is relatively short. Therefore, this paper will build a HIRAS global cloud detection model based on the LightGBM algorithm.

TABLE II. THE CLOUD DETECTION PERFORMANCE OF FIVE ALGORITHMS UNDER OPTIMAL HYPERPARAMETERS.

Algorithm	**Best parameters**	**POD**	**FAR**	**ACC**	**T (s)**
K Neighbors Classifier	‘algorithm’=’ball_tree’, ‘n_neighbors’=24	0.80	0.20	0.81	57
Logistic Regression	‘penalty’=’l2’, ‘C’=1.91,	0.83	0.19	0.82	3205
Random Forest	‘n_estimators’=200, ’max_depth’=260, ‘min_samples_leaf’=6, ’min_samples_split’=12	0.83	0.17	0.84	74
GBDT	‘num_iterations’=125, ’learning_rate’=0.38, ‘min_samples_split’=36 ’max_depth’=18	0.84	0.16	0.85	1532
LightGBM	‘num_iterations’=140, ’learning_rate’=0.03, ‘num_leaves’=160, ’max_depth’=100	0.86	0.15	0.87	124

The LightGBM algorithm is an efficient implementation form of the traditional gradient boosting tree (GBDT) algorithm [30]. Compared with GBDT, LightGBM is more suitable for classification and regression tasks of high-dimensional large datasets. LightGBM innovatively proposed Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) algorithms to improve model efficiency. The GOSS and EFB reduce the number of training samples and feature dimensions, respectively, so that the model can significantly improve training efficiency while maintaining accuracy. LightGBM uses the Histogram algorithm [34] to discretize the floating-point value of the feature into M integers to find the optimal split node. The histogram algorithm can reduce memory occupation and computational consumption, and coarser split nodes also have the effect of regularization. LightGBM uses a leaf-wise growth decision tree, whose leaf nodes split from the leaf with the largest split gain. The Leaf-wise growth method can quickly improve the fitting ability of the decision tree, and it is also prone to over-fitting problems. LightGBM can reduce the decision tree's complexity by limiting the max depth ('max_depth') and the maximum number of leaves ('num_leaves') of the tree. In addition, the over-fitting problem is also dealt with by adjusting the minimal samples at a leaf node ('min_data_in_leaf') and using L1 and L2 regularization. Increasing the number of decision trees ('num_iterations') can improve the fitting ability of the model. Learning rate ('learning_rate') shrinks the contribution of each tree. ("Bagging_fraction") and ("feature_fraction") respectively represent the random sampling ratio of samples and features in each iteration.

3) Training mechanism and model configuration

The training and test datasets are composed of the HIRAS and MERSI collocation data randomly selected from each

month of 2020 (January 15, February 6, March 13, April 30, May 12, June 5, July 10, August 28, September 16, October 2, November 5, and December 23) with an 8:2 ratio. The sampling interval is one scan line and three FOVs to ensure that the samples come from different atmospheric and ground conditions. For the reason that the qualities of HIRAS infrared radiation and MERSI cloud mask products in the polar regions (60-90° N, 60-90° S) are affected by the reflection of snow on the ground, the polar samples are excluded from the training and test datasets.

HIR radiations are significantly affected by cloud cover and cloud phase, which is the physical basis of HIR data cloud detection. However, the surface type (sea or land), solar radiation (day or night), optical path (1/cos (sensor zenith angle)) and other factors also affect the properties of infrared radiation. Therefore, HIRAS global HIR data are divided into four datasets: day-land, day-sea, night-land, and night-sea (TABLE III), used to build LightGBM cloud detection models separately. The true labels of the HIRAS FOVs (cloud=0, clear sky=1) are generated by the HIRAS/MERSI collocation cloud detection method in section 3.1.

The cloudy samples are approximately 8 (3) times as large as the clear-sky samples in the ocean (land) dataset. The imbalance between cloudy and clear-sky samples may make the LightGBM model learn more cloudy sample features and reduce the detection performance of clear-sky samples [35,36]. The clear-sky samples are the more concerned category in cloud detection. The cloud samples in the dataset are randomly downsampled to alleviate the impact of data imbalance, controlling the ratio of cloudy and clear-sky samples about 2:1. Finally, the constructed HIRAS global training dataset and test dataset are shown in TABLE III. A total of 402,620 samples are used to train LightGBM model, and 10,0657 samples are used for validation.

TABLE III. TRAINING AND TEST DATASETS

Model	Train dataset		Test dataset	
	cloudy	*clear-sky*	*cloudy*	*clear-sky*
Day land	54149	30351	13538	7588
Day sea	77520	38760	19380	9690
Night land	61943	39033	15486	9758
Night land	67243	33621	16811	8406

In this paper, a sensitivity experiment is used to select appropriate principal components as the predictor of the model. Since HIRAS short-wave infrared radiations are affected by the solar stray light, the PCs of daytime land (sea) models are calculated by the long-wave and medium-wave infrared radiations. And the PCs of nighttime land (sea) models are computed by the long-wave, medium-wave and short-wave infrared radiations.

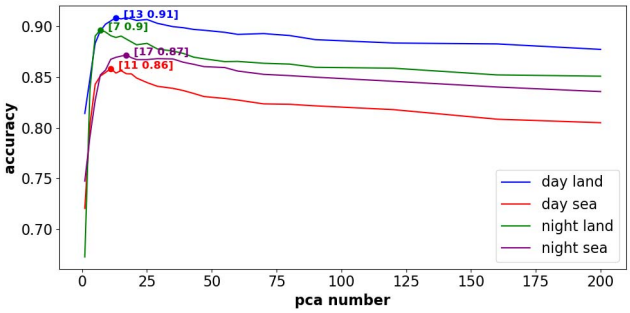

Figure 1. sensitivity experiment of the number of PCs

As the number of PCs increases, the cloud detection accuracy of LightGBM increases rapidly (Figure 1). It is because more PCs provide more effective spectral information. The accuracy drops slightly after reaching the "knee point" corresponding to the optimal number of PCs. It may be that the instrument noise gradually affects the adequate spectral information. Therefore, the leading K PCs of datasets (the day-land dataset is 13, the day-sea dataset is 11, the night-land dataset is 7, the night-sea dataset is 17) are respectively used as the predictors of the corresponding LightGBM model. The Hyperopt tuner is used to select the optimal hyperparameter combination of the LightGBM model. Finally, the trained LightGBM model is shown in TABLE IV.

TABLE IV. THE LIGHTGBM CLOUD DETECTION MODEL

Model	num iterations	Num leaves	Max depth	Learning rate	Min data in leaf
Day land	290	280	170	0.12	60
Day sea	270	300	290	0.10	55
Night land	270	280	130	0.09	40
Night sea	290	380	260	0.09	40

IV. VALIDATION

A. *The accuracy validation of collocation algorithm*

The MERSI 25 band (12.0 μm) detecting the temperature of land and water cloud is fully covered by the HIRAS's long-wave IR band (8.8-15.38 μm), which is used to compare with HIRAS radiances. The spectral convolution (formula 9) of HIRAS hyperspectral radiances is performed to match the MERSI's band radiance [23]. In formula 9, the $I(v)$ is HIRAS hyperspectral radiance at wavenumber v, and the F(v) is the spectral response function (SRF) of MERSI 25 band. The R is HIRAS-convolved MERSI band radiance at bandpass limits of v_1 and v_2.

$$R = \frac{\int_{v_1}^{v_2} I(v)F(v)dv}{\int_{v_1}^{v_2} F(v)dv} \quad (9)$$

The radiances of MERSI within HIRAS FOV are averaged and then compared with the HIRAS-convolved MERSI 25

band radiances. The differences between the spectral-convolved bright temperatures (BTs) of HIRAS and the spatial-averaged BTs of MERSI can reflect the accuracy of the collocation algorithm. Because the spatial-averaged MERSI BT is sensitive to matched MERSI pixels, the smaller the BT differences, the higher the accuracy of the collocation algorithm. The clear-sky measurements of HIRAS and MERSI on the global ocean are used as validation data, selected by collocation algorithm (the threshold of cloudiness is 10%). Figure 2 are the BT difference map (a) and the probability density function (PDF) of differences between the HIRAS-convolved and spatial-averaged BTs for the MERSI 25 band. Although the selected clear FOVs have diverse atmospheric and ocean surface conditions, the BT differences between HIRAS and MERSI are mainly among 0-0.5 K. Moreover, the BT differences conform to the normal distribution with a mean value of 0.32 K and a standard deviation of 0.32. The spectral-convolved BTs of HIRAS are warmer than the spatial-averaged BTs of MERSI as a whole. The presence of a small number of clouds in some HIRAS clear-sky FOVs may cause a BT deviation of 0.32K. In general, the HIRAS/MERSI collocation algorithm has shown high accuracy.

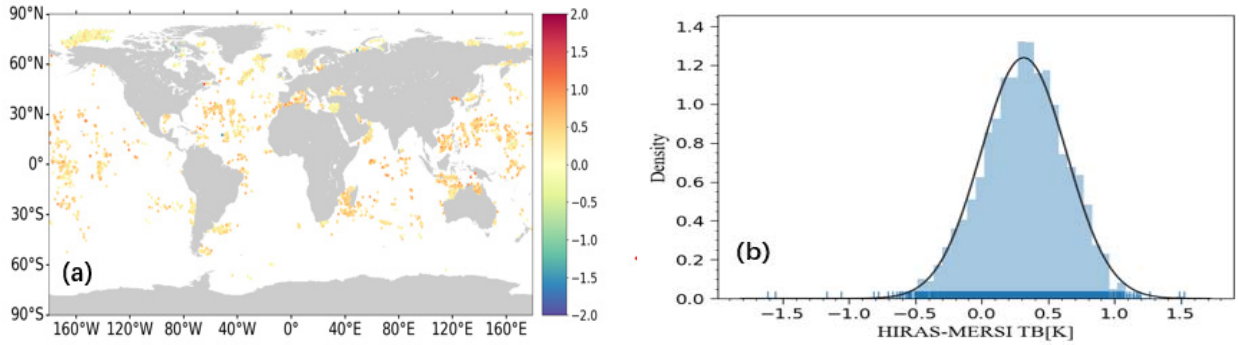

Figure 2. HIRAS-MERSI BT difference map (a) and its PDF (b) on global ocean clear-sky FOVs on June 5, 2020. The black line represents the fitted normal distribution curve.

B. validation of LightGBM model

1) validation using MERSI cloud mask product

Four classical metrics [37] are used to evaluate the cloud detection performance of the LightGBM model, including the probability of detection (POD), false alarm ratio (FAR), accuracy (ACC) and Heidke skill score (HSS) [38].

$$POD = a/(a+c) \quad (10)$$

$$FAR = b/(a+b) \quad (11)$$

$$ACC = (a+d) / (a+b + c+d) \quad (12)$$

$$HSS = 2(ad-bc) / [(a+c)(a+d)+(a+b)(b+d)] \quad (13)$$

Where a is the number of samples that both the matching method and LightGBM model are classified as clear-sky, b is the number of samples identified as cloudy by the matching method but labelled as clear-sky by the LightGBM model, c is the number of samples identified as clear-sky by the matching method but classified as cloudy by the LightGBM model, d is the number of samples that both the matching method and LightGBM model are identified as cloudy. The labels generated by the matching method are regarded as the reference truth. POD represents the ratio of the number of clear-sky samples correctly detected by the LightGBM model to the total number of clear-sky samples. A larger POD denotes that more clear-sky samples are correctly detected (optimal=1). FAR represents the ratio of the number of clear-sky samples incorrectly detected by the LightGBM model to the number of all predicted clear-sky samples. FAR represents the false detection rate of the LightGBM model for clear-sky samples (optimal = 0). ACC represents the ratio of clear-sky and cloudy samples correctly identified by the LightGBM model to the total samples. ACC signifies the overall cloud detection accuracy of the LightGBM model (optimal = 1). The HSS is a commonly used robust performance metric that eliminates the cloud detection accuracy obtained due to random chance (optimal = 1).

TABLE V. CLOUD DETECTION SCORES OF THE LIGHTGBM MODEL IN THE TEST DATASET

Model	POD	FAR	ACC	HSS
Day land	0.93	0.09	0.92	0.84
Day sea	0.87	0.20	0.88	0.65
Night land	0.92	0.09	0.92	0.83
Night sea	0.87	0.20	0.87	0.65

The HIRAS global datasets are randomly divided into 20% as the test datasets (TABLE III) to verify the generalization performance of the trained LightGBM models for unknown samples. The cloud detection performances of the LightGBM models in the test datasets are shown in TABLE V. The LightGBM day and night models have similar cloud detection performance, with an ACC of 0.9. It may be because the separate training of the day and night models alleviated the influence of solar radiation on cloud detection. The ACC of the LightGBM land model can reach 0.92, and its HSS is 0.84. The LightGBM sea models have an ACC of 0.89 and an HSS of 0.65. The land models have better cloud detection performance

than the sea models. It may be because the ocean scene has more partially cloudy FOVs than the land scene (as shown in fig 2.a and 3.a). However, the partially cloudy FOVs are "hard cases" in cloud detection classification because their infrared radiation characteristics are between the completely cloudy FOVs and the clear-sky FOVs. The partially cloudy FOVs are easily incorrectly detected as the clear-sky FOVs, resulting in the FAR of the ocean model (FAR=20%) is much higher than that of the land model (FAR=9%). In addition, the infrared radiation of ocean surface and low-level water clouds have a certain similarity, which also increases the difficulty of cloud detection. In general, the LightGBM models have high accuracy and good generalization performance for cloud detection from HIRAS HIR data.

2) Test case analysis

To further verify the performance of the LightGBM cloud detection model, one day of HIRAS and MERSI collocation data (September 18, 2020) were randomly selected as a test case. The data is outside the training dataset. Figure 3 (Figure 4) compares cloud detection results of observation data during the day (night). For HIRAS L1c cloud cover products, when the cloud cover is less than 10%, the HIRAS FOV is identified as clear-sky (label=1). When the cloud cover is greater than 10%, the HIRAS FOV is classified as cloudy (label=0). For the LightGBM cloud detection method, observation data are first classified into day-land, day-sea, night-land, and night-sea, and then call the corresponding trained LightGBM models for cloud detection.

TABLE VI. CLOUD DETECTION SCORES OF LIGHTGBM MODEL AND HIRAS L1C CLOUD COVER PRODUCTS IN TEST CASES

Model	Product	POD	FAR	ACC	HSS
Day land	LightGBM	0.93	0.08	0.93	0.85
Day land	HIRAS/L1c	0.64	0.36	0.68	0.28
Day sea	LightGBM	0.82	0.24	0.89	0.57
Day sea	HIRAS/L1c	0.59	0.43	0.81	0.16
Night land	LightGBM	0.93	0.08	0.93	0.85
Night land	HIRAS/L1c	0.68	0.30	0.71	0.38
Night sea	LightGBM	0.81	0.24	0.88	0.56
Night sea	HIRAS/L1c	0.59	0.43	0.81	0.15

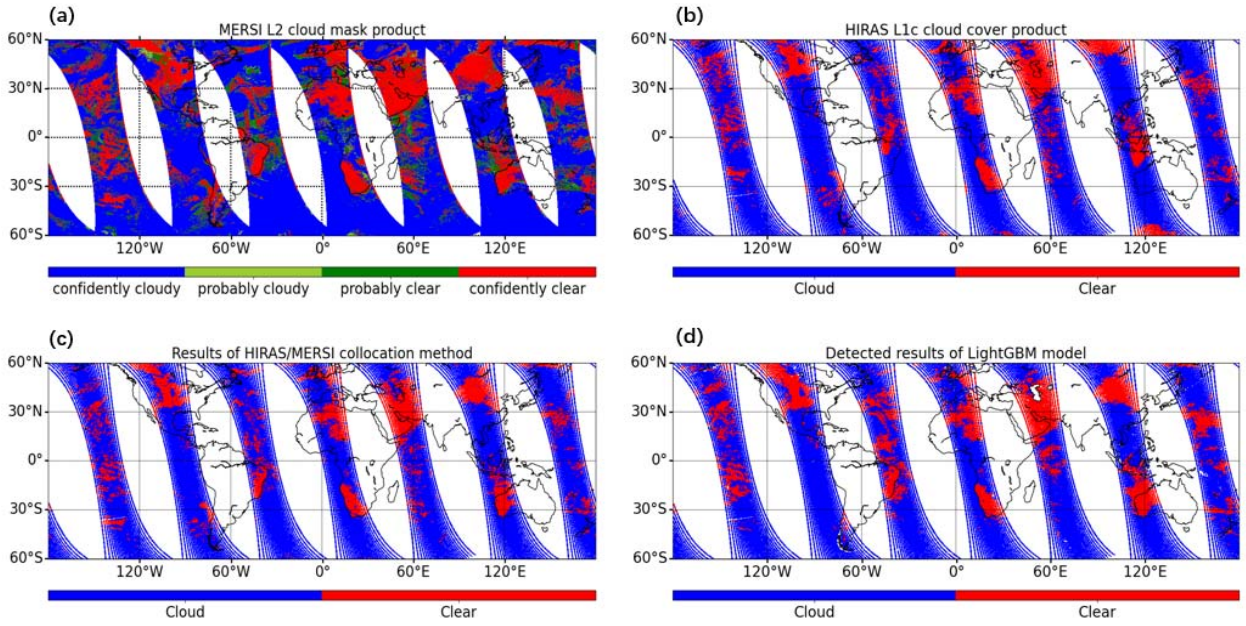

Figure 3. The cloud detection comparison of observation data during the day on September 18, 2020.（a）MERSI L2 cloud mask product; (b) HIRAS L1c cloud cover product；(c) the cloud detection results of HIRAS/MERSI collocation method; (d) the cloud detection results of LightGBM model.

For the daytime observation data of the test cases (Figure 3), the cloud detection results of the HIRAS/MERSI collocation algorithm (Figure 3.c) and the MERSI L2 cloud masks (Figure 3.a) have high consistency, showing adequate cloud detection accuracy. Therefore, the matching cloud detection results are used as the reference truth. There are some deviations between the HIRAS L1c cloud cover product (Figure 3.b) and the MERSI cloud mask, such as western Australia (30°S, 120°E), the central Pacific (5°N, 140°W) and other regions. The cloud detection accuracies of HIRAS L1c cloud cover products on land and sea areas are only 0.68 and 0.81, respectively (TABLE VI). It may be due to systematic errors in HIRAS L1c cloud cover products. As shown in Figure 3.d, many cloudy and clear-sky FOVs are correctly classified by the LightGBM model, with an accuracy of 0.93 for land data and an accuracy of 0.89 for sea data (TABLE VI). The partially cloudy FOVs in a sizeable clear-sky area, such as the central Pacific Ocean (5°S, 140°W), are easily classified as clear-sky FOVs by the LightGBM model. It may be that these FOVs have prominent clear-sky radiation characteristics. Because there are more partially cloudy FOVs on the ocean, the FAR of sea model (FAR=0.24) is much larger than the land model (FAR=0.08).

Some clear-sky FOVs in the high-latitude ocean, such as the South Pacific (40°S, 130°W), are identified as cloudy by the LightGBM model. It may be due to the low water temperature of the ocean at high latitudes, and its infrared radiations are similar to that of water clouds. In addition, the imbalance between the cloudy and clear-sky FOVs in the ocean (the cloudy FOVs: the clear-sky FOVs = 8:1) further increases the difficulty of cloud detection for the sea HIR data. Therefore, the cloud detection performances of the LightGBM model on land (ACC=0.93, HSS=0.86) are better than that of the ocean (ACC=0.93, HSS=0.57).

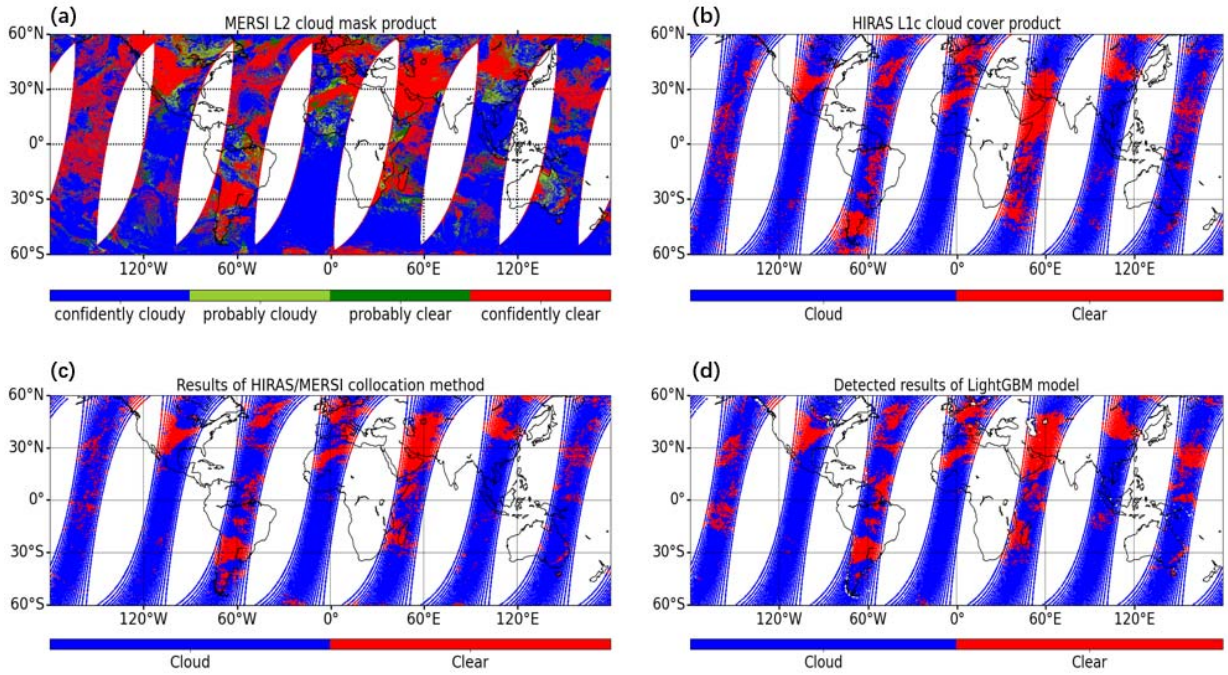

Figure 4. Cloud detection comparison of night observation data on September 18, 2020.（a）MERSI L2 cloud mask product; (b) HIRAS L1c cloud cover product；(c) the cloud detection results of HIRAS/MERSI collocation method; (d) the cloud detection results of LightGBM model.

As shown in Figure 4, the LightGBM model also has similar conclusions for the night observation data of the test cases. Compared with HIRAS L1c cloud cover products, the cloud detection performance of the LightGBM model has been dramatically improved (the ACC of the land model has increased by about 22%, and the ACC of the sea model has increased by about 7%). It is also worthy to note that the cloud detection scores of the LightGBM model in the test case are similar to the performance in the test set, showing superior generalization performance.

C. Time complexity of the algorithm

For HIRAS/MERSI collocation cloud detection algorithm (Section 2.1), each HIRAS FOVs need to search for the overlapping MERSI pixels based on the LOS from a space of about 100*100. The time complexity of the HIRAS/MERSI collocation algorithm is O(N*M), where N is the number of HIRAS FOV and M is the number of MERSI pixels to be matched (about 10000 times of N). Therefore, the time complexity is simplified to O(N2). The trained LightGBM model is a known functional relationship. LightGBM consists of multiple decision trees, and each decision tree includes multiple branch nodes. The time complexity is related to the structure of the model (the number of decision trees, etc.). However, these structural parameters are constants. Therefore, the time complexity of the LightGBM cloud detection algorithm is O(N), where N is the number of HIRAS FOVs. Thus, the time cost of the HIRAS-MERSI collocation algorithm increases faster than the LightGBM algorithm with the number of input samples. LightGBM cloud detection algorithm has significant computational efficiency.

The HIR data from HIRAS on September 18, 2020, are used to compare further the actual running time of the two cloud detection algorithms. The experimental environment is the same 8GB intel i5 computer. The average time cost, defined as the average cloud detection time of a HIRAS's L1 file, is used to measure the cloud detection efficiency of the algorithm. The experimental results show that the average time cost of the HIRAS/MERSI collocation algorithm is about 302.77s, and the average time cost of the LightGBM model is about 0.45s. Therefore, the cloud detection speed of the LightGBM model is approximately 670 times that of the HIRAS/MERSI collocation algorithm, showing a superior cloud detection efficiency.

V. Conclusion

Cloud detection is a crucial step in applying HIR data, such as data assimilation, the profile inversion of temperature and humidity, etc. The traditional physical cloud detection methods need to set a series of empirical thresholds. The collocation imager method and the dual-channel method rely on the fusion of multi-source data. The clear channel cloud detection method

based on radiative transfer model simulation is affected by the error of the model background field. Therefore, this paper aims to build a fast cloud detection algorithm based on machine learning, which directly identifies the cloud labels through the infrared radiation characteristics of the HIRAS FOVs.

At present, the cloud detection algorithms based on machine learning from HIR data are only well applied to the completely cloudy FOVs and the completely clear-sky FOVs. To improve the detection effect of the machine learning model for the partially cloudy FOVs, this paper has made the following attempts:

- For machine learning algorithms. The LightGBM algorithm with the best cloud detection performance is selected from five classic machine learning algorithms. The LightGBM adopts the boosting integration, focusing on the misclassified samples in model training, with strong learning ability and generalization performance. In addition, the LightGBM also has apparent training efficiency.
- For data processing. The HIRAS global observation data are divided into day-land, day-sea, night-land, and night-sea datasets to control the influencing factors of infrared radiation. The four datasets are used to train the machine learning models separately. The HIRAS infrared radiations are transformed into K principal components by PCA technology, used as predictors. The PCA has the advantages of reducing redundant information and sensor noises, accelerating model convergence and prediction. The cloudy samples in the datasets are randomly downsampled to improve the detection performance of the LightGBM model on the clear-sky FOVs.

The constructed LightGBM model showed satisfactory cloud detection performance in the randomly selected test dataset and test cases. The LightGBM model has a higher accuracy for the large clear-sky and cloudy areas than the partially cloudy areas. It may be because the partially cloudy FOV has the radiation properties of both the cloudy and clear-sky FOVs. The LightGBM model has a similar cloud detection performance for HIRAS day and night HIR data. The LightGBM model has an accuracy of 0.93, an HSS of 0.85 for land HIR data and an accuracy of 0.89, an HSS of 0.65 for sea HIR data. The LightGBM model has better cloud detection performance on land than on sea. It may be because the ocean has more partially cloudy FOVs, and the infrared radiation of high-latitude oceans and low-level water clouds have a certain similarity. In addition, the imbalance of ocean data sets further increases the difficulty of cloud detection. Compared with HIRAS L1c cloud cover products, the cloud detection accuracy of the LightGBM model has been improved by about 22% (7%) for land (ocean) HIR data. The cloud detection speed of the LightGBM model is about 670 times that of the HIRAS/MERSI collocation algorithm, showing well cloud detection efficiency. The improvement of cloud detection accuracy and efficiency will benefit the operational application of the LightGBM cloud detection method. Future work also needs to improve the detection effect of the partially cloudy FOVs on the ocean. In addition, the influence of the LightGBM cloud detection method on the application of HIR data needs to be further studied.

ACKNOWLEDGMENT

This research was funded by the National Natural Science Foundation of China with the grant number of 41675097, 41375113, 42075149, and 42005120.

REFERENCES

[1] Joo, S., Eyre, J., Marriott, R. (2013) The impact of MetOp and other satellite data within the Met Office global NWP system using an adjoint-based sensitivity method. Monthly weather review 141: 3331–3342.

[2] Cucurull, L., Anthes, R., Tsao, L.L. (2014) Radio occultation observations as anchor observations in numerical weather prediction models and associated reduction of bias corrections in microwave and infrared satellite observations. Journal of Atmospheric and Oceanic Technology 31: 20–32.

[3] Aumann, H.H., Miller, C.R. (1995) Atmospheric infrared sounder (AIRS) on the earth observing system. Advanced and Next-Generation Satellites. International Society for Optics and Photonics, 1995, Vol. 2583, pp. 332–343.

[4] Clerbaux, C., Hadji-Lazaro, J., Turquety, S., George, M., Coheur, P.F., Hurtmans, D., Wespes, C., Herbin, H., Blumstein, D., Tourniers, B., others. (2007) The IASI/MetOp1 Mission: First observations and highlights of its potential contribution to GMES2. Space Research Today 168: 19–24.

[5] Smith, A., Atkinson, N., Bell, W., Doherty, A. (2015) An initial assessment of observations from the Suomi-NPP satellite: data from the Cross-track Infrared Sounder (CrIS). Atmospheric Science Letters 16: 260–266.

[6] Hilton, F., Atkinson, N.C., English, S.J., Eyre, J.R. (2009) Assimilation of IASI at the Met Office and assessment of its impact through observing system experiments. Quarterly Journal of the Royal Meteorological Society 135: 495–505.

[7] McNally, A.P., Watts, P.D. (2003) A cloud detection algorithm for high-spectral-resolution infrared sounders. Quarterly Journal of the Royal Meteorological Society 129: 3411–3423.

[8] Mcnally, A.P., Watts, P.D., A. Smith, J., Engelen, R., Kelly, G.A., Thépaut, J.N., Matricardi, M. (2006) The assimilation of AIRS radiance data at ECMWF. Quarterly Journal of the Royal Meteorological Society 132: 935–957.

[9] Lin, L., Zou, X., Weng, F. (2017) Combining CrIS double CO2 bands for detecting clouds located in different layers of the atmosphere. Journal of Geophysical Research: Atmospheres 122: 1811–1827.

[10] Qi, C., Wu, C., Hu, X., Xu, H., Lee, L., Zhou, F., Gu, M., Yang, T., Shao, C., Yang, Z., Zhang, P. (2020) High Spectral Infrared Atmospheric Sounder (HIRAS): System Overview and On-Orbit Performance Assessment. IEEE Transactions on Geoscience and Remote Sensing 58: 4335–4352.

[11] Goldberg, M.D., Qu, Y., McMillin, L.M., Wolf,W., Lihang Zhou., Divakarla, M. (2003) AIRS near-real-time products and algorithms in support of operational numerical weather prediction. IEEE Transactions on Geoscience and Remote Sensing 41: 379–389.

[12] Li, J., Menzel,W.P., Sun, F., Schmit, T.J., Gurka, (2004) J. AIRS Subpixel Cloud Characterization Using MODIS Cloud Products. Journal of Applied Meteorology 43: 1083 – 1094.

[13] Eresmaa, R. (2014) Imager-assisted cloud detection for assimilation of Infrared Atmospheric Sounding Interferometer radiances.Quarterly Journal of the Royal Meteorological Society 140: 2342–2352.

[14] Wang, P., Li, J., Li, Z., Lim, A.H.N., Li, J., Schmit, T.J., Goldberg, M.D. (2018) The Impact of Cross-track Infrared Sounder (CrIS) Cloud-Cleared Radiances on Hurricane Joaquin (2015) and Matthew (2016) Forecasts. Journal of Geophysical Research: Atmospheres 122: 13,201–13,218.

[15] Wang, P., Li, J., Li, J., Li, Z., Schmit, T.J., Bai, W. (2014) Advanced infrared sounder subpixel cloud detection with imagers and its impact on

radiance assimilation in NWP. Geophysical Research Letters 41: 1773–1780.

[16] Mahajan, S., Fataniya, B. (2019) Cloud detection methodologies: Variants and development—A review. Complex & Intelligent Systems pp: 1–11.

[17] Luo, T., Zhang, W., Yu, Y., Feng, M., Duan, B., Xing, D. (2019) Cloud detection using infrared atmospheric sounding interferometer observations by logistic regression. International Journal of Remote Sensing 40: 6530–6541.

[18] Zhang, Q., Yu, Y., Zhang,W., Luo, T., Wang, X. (2019) Cloud Detection from FY-4A's Geostationary Interferometric Infrared Sounder Using Machine Learning Approaches. Remote Sensing 11.

[19] Liu, Q., Xu, H., Sha, D., Lee, T., Duffy, D.Q., Walter, J., Yang, C. (2020) Hyperspectral Infrared Sounder Cloud Detection Using Deep Neural Network Model. IEEE Geoscience and Remote Sensing Letters pp: 1–5.

[20] Carminati, F., Xiao, X., Lu, Q., Atkinson, N., Hocking, J. (2019) Assessment of the Hyperspectral Infrared Atmospheric Sounder (HIRAS). Remote Sensing 11.

[21] Wu, C., Qi, C., Hu, X., Gu, M., Yang, T., Xu, H., Lee, L., Yang, Z., Zhang, P. (2020) FY-3D HIRAS Radiometric Calibration and Accuracy Assessment. IEEE Transactions on Geoscience and Remote Sensing 58: 3965–3976.

[22] Xu, N., Niu, X., Hu, X., Wang, X., Wu, R., Chen, S., Chen, L., Sun, L. Yang, Z., Zhang, P. (2018) Prelaunch calibration and radio-metric performance of the advanced MERSI II on FengYun-3D. IEEE Transactions on Geoscience and Remote Sensing 56: 4866-4875.

[23] Wang, L., Tremblay, D., Zhang, B., Han, Y. (2016) Fast and Accurate Collocation of the Visible Infrared Imaging Radiometer Suite Measurements with Cross-Track Infrared Sounder. Remote Sensing 8.

[24] Fan, S., Han,W., Gao, Z., Yin, R., Zheng, Y. (2019) Denoising algorithm for the FY-4A GIIRS based on principal component analysis. Remote Sensing 11: 2710.

[25] Lee, L., Zhang, P., Qi, C., Hu, X., Gu, M. (2019) HIRAS noise performance improvement based on principal component analysis. Appl. Opt. 58: 5506–5515.

[26] Feng, J., Xu, H., Mannor, S., Yan, S. (2014) Robust logistic regression and classification. Advances in neural information processing systems 27: 253–261.

[27] Wang, C., Platnick, S., Meyer, K., Zhang, Z., Zhou, Y. (2020) A machine-learning-based cloud detection and thermodynamic-phase classification algorithm using passive spectral observations. Atmospheric Measurement Techniques 13: 2257–2277.

[28] Goldberger, J., Hinton, G.E., Roweis, S., Salakhutdinov, R.R. (2004) Neighbourhood components analysis. Advances in neural information processing systems 17.

[29] Anghel, A., Papandreou, N., Parnell, T., De Palma, A., Pozidis, H. (2018) Benchmarking and optimization of gradient boosting decision tree algorithms. arXiv preprint arXiv:1809.04559.

[30] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y. (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30: 3146–3154.

[31] Bergstra, J., Yamins, D., Cox, D. (2013) Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. International conference on machine learning. PMLR, pp: 115–123.

[32] Liashchynskyi, P., Liashchynskyi, P. (2019) Grid search, random search, genetic algorithm: a big comparison for NAS. arXiv preprint arXiv:1912.06059.

[33] Bergstra, J., Bengio, Y. (2012) Random search for hyper-parameter optimization. Journal of machine learning research 13.

[34] Chen, T., Guestrin, C. (2016) Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp: 785–794.

[35] Liu, X.Y., Wu, J., Zhou, Z.H. (2008) Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39: 539–550.

[36] Jeatrakul, P., Wong, K.W., Fung, C.C. (2010) Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. International Conference on Neural Information Processing. Springer pp: 152–159.

[37] Nurmi, P. Recommendations on the verification of local weather forecasts 2003.

[38] Mecikalski, J.R., Bedka, K.M., Paech, S.J., Litten, L.A. (2008) A statistical evaluation of GOES cloud-top properties for nowcasting convective initiation. Monthly Weather Review 136: 4899–4914.