CDL: A Cloud Detection Algorithm Over Land for MWHS-2 Based on the Gradient Boosting Decision Tree

Shuxian Liu, Yan Yin, Zhigang Chu, and Shuai An

***Abstract*—This article presents a standalone cloud detection algorithm over the land (CDL) for microwave humidity sounder-2 (MWHS-2), which is characterized by the first operational satellite sensor measuring 118.75 GHz. The CDL is based on the advanced machine learning algorithm gradient boosting decision tree, which achieves the state-of-the-art performance on tabular data with high accuracy, fast training speed, great generalization ability, and weight factor ranking of predictors (or features). Given that the new-generation weather radar of China (CINRAD) provides improved cloud information with extensive temporal-spatial coverage, the observations from CINRAD are used to train the algorithm in this study. There are four groups of radiometric information employed to evaluate the CDL: all frequency ranges from MWHS-2 (all-algorithm), the humidity channels near 183.31 GHz (hum-algorithm), the temperature channels near 118.75 GHz (tem-algorithm), and the window channels at 89 and 150 GHz (win-algorithm). It is revealed that the tem-algorithm (around 118.75 GHz) has a superior performance for CDL along with the optimal values of most evaluation metrics. Although the all-algorithm uses all available frequencies, it shows inferior ability for CDL. Followed are the win-algorithm and hum-algorithm, and the win-algorithm performs better. The analysis also indicates that the latitude, zenith angle, and the azimuth are the top-ranking features for all four algorithms. The presented algorithm CDL can be applied in the quality control processes of assimilating microwave radiances or in the retrieval of atmospheric and surface parameters for cloud filtering.**

***Index Terms*—Cloud detection, ground-based radar, machine learning (ML), microwave humidity sounder-2 (MWHS-2).**

I. Introduction

IN CONTRAST to visible and infrared satellite observations, which can only sense the radiation from the top of clouds, microwave (MW) sounders can propagate through most nonprecipitating clouds and have a better ability to sense the cloud particles [1], [2]. Moreover, MW temperature sounder (MWTS) and microwave humidity sounder (MWHS) can acquire multiple channels (CHs) of brightness temperatures (BTs), providing rich information for profiling atmospheric temperature and moisture. However, the measured BT is affected by many factors, including the water vapor, cloud and rain contamination, surface emissivity, and so on.

Over the ocean, a number of cloud detection methods for MW observations have been developed. The amount of cloud liquid water path can be obtained from the two window CHs (at 23.8 and 31.4 GHz) [3], [4]. Two other window CHs of 89 and 150 GHz can be used to retrieve the cloud scattering index [5] and cloud ice water path [6]. Using the dual oxygen absorption bands (at 50–60 and 118.75 GHz) in MWTS and MWHS, several pairs of oxygen CHs can be applied to compute the cloud emission and scattering index at different height levels [7]. Buehler *et al*. [8] developed a cloud filter method based on the BT differences (BTDs) of CHs around the water vapor line at 183.31 GHz.

Nonetheless, cloud detection over the land (CDL) is still challenging because the impact of the land surface on the BT is much larger than that of the clouds and the surface emissivity changes spatially and temporally with the latitude and seasons. Despite these difficulties, several approaches from physically based detection methods to quite a few statistical techniques aided by various cloud products have been proposed. Actually, cloud detection is well suited for machine learning (ML) techniques [9]–[13], as it is a type of classification that involves multivariate analysis and has a complex nonlinear relationship between the variables.

Therefore, the quantities of ML-based cloud detection methods have been developed on different ranges of frequencies available onboard the MW instruments. The temperature CHs (between 50 and 60 GHz), window CHs (at 23.8, 31.4, 89, and 150 GHz), and humidity CHs (around 183.31 GHz) have already been employed in the cloud classification model for advanced microwave sounding units A and advanced microwave sounding units B, by using a neural network (NN) method trained with the cloud classification products of Meteosat second-generation spinning-enhanced visible and infrared imager (SEVIRI) [1]. Furthermore, the window CHs from 19 to 91 GHz have been proved to perform better than the humidity CHs near 183.31 GHz over the land in cloud detection [14] based on the Naïve Bayes classifier for special sensor MW imager/sounder. Favrichon *et al*. [15] use the NN model trained on the SEVIRI cloud products,

Manuscript received May 11, 2020; revised July 6, 2020; accepted July 25, 2020. Date of publication August 4, 2020; date of current version August 21, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 41590873 and in part by the National Key R&D Program of China under Grant 2018YFC1506603. *(Corresponding author: Yan Yin.)*

Shuxian Liu, Yan Yin, and Zhigang Chu are with the Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Key Laboratory for Aerosol-Cloud-Precipitation of China Meteorological Administration, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: liushuxian@nuist.edu.cn; yinyan@nuist.edu.cn; chuzhigang@nuist.edu.cn).

Shuai An is with the Smart Supply Chain Y BU, JD.COM, Beijing 100000, China (e-mail: anshuai1@jd.com).

Digital Object Identifier 10.1109/JSTARS.2020.3014136

and the analysis show that the cloud index confidence increases with the number of CH frequencies, with the accuracy more than 70% in detecting cloud contamination.

However, it is noted that the CHs near 118.75 GHz have not yet been evaluated or compared with the humidity and window CHs in ML-based cloud detection methods. As previous studies indicate, the BTs at 118.75 GHz show an extremely strong dependence on cloud particles [16]. Thus, it is imperative to incorporate the radiometric information of 118.75 GHz into the ML algorithm and assess its performance for cloud detecting. Gradient boosting decision tree (GBDT) algorithm achieves state-of-the-art performance on the tabular data (data size less than 1000, more than 10 000, or even more than 1 million) instead of the image data, and the ensemble properties of GBDT may avoid the overfitting. Radar measurements are widely recognized as relatively accurate labels in ML-based cloud classifier models [17]. Compared with the cloud radar, ground-based weather radar has a wider temporal and spatial coverage, thus having obvious superiority in collocating with polar-orbiting satellite measurements. Here, the observations from China's new-generation weather radar (CINRAD) are used to train the GBDT classifier.

The purpose of this study is to develop a standalone cloud detection algorithm over the land for MWHS-2 in order to identify the cloudy scenes prior to assimilating the radiances into the numerical weather prediction (NWP) models or retrieving the atmospheric and surface parameters. And the CDL model will be developed for multiple frequency ranges in order to compare the results of 118.75 GHz with the humidity and window CHs.

The rest of this article is organized as follows. The satellite and radar datasets are described in Section II. The cloud detection algorithm is presented in Section III. In Section IV, the results are presented focusing on the effects of CDL for MWHS-2. Finally the conclusion is summarized in Section V.

II. Data

The MWHS-2 onboard FY-3C is a 15-CH cross-track scanning MW radiometer with 8 CHs in the oxygen absorption line (118.75 GHz), 5 CHs in the water vapor absorption line (183.31 GHz), and 2 window CHs at 89 and 150 GHz. The characteristics of MWHS-2 are illustrated in Table I, containing the peak weighting functions and the central frequencies for the 15 CHs. As also given in Table I, CHs 2–9 are for temperature sounding from 20 to 1000 hPa, CHs 11–15 are for humidity sounding from 450 to 800 hPa, and CHs 1 and 10 are two window CHs.

In this study, we select eight *S*-band weather radars located in east China from the CINRAD *S*-band A-type with an effective distance of ∼230 km (see Fig. 1). The quality control of radar data includes fuzzy logic clutter filter,[1] median filter, and reflectivity bias correction [18], [19]. The 2-D composite reflectivity is generated by using the severe weather automatic nowcast system [20], which is developed by the China Meteorological Administration. Then, the value of composite reflectivity, whose measure time is closest to the time when MWHS-2 passes East China and the measurement range is in the area of MWHS-2's filed-of-view (FOV), is selected to calculate the average value in each FOV.

[1][Online]. Available: http://www.weather.gov/code88d/

TABLE I
Instrument Characteristics of MWHS-2

Channel	Center Frequency (GHz)	Peak Weighting Function (hPa)	Sounding
1	89.0	surface	window
2	118.75 ± 0.08	20	temperature
3	118.75 ± 0.2	60	temperature
4	118.75 ± 0.3	100	temperature
5	118.75 ± 0.8	250	temperature
6	118.75 ± 1.1	300	temperature
7	118.75 ± 2.5	700	temperature
8	118.75 ± 3.0	surface	temperature
9	118.75 ± 5.0	surface	temperature
10	150.0	surface	window
11	183.31 ± 1	450	humidity
12	183.31 ± 1.8	500	humidity
13	183.31 ± 3	600	humidity
14	183.31 ± 4.5	700	humidity
15	183.31 ± 7	800	humidity

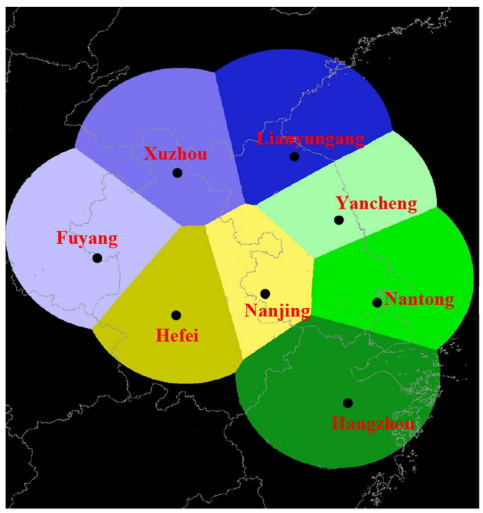

Fig. 1. Spatial distribution of the radar network in East China.

Radar reflectivity factor is often used for cloud detection [21]–[25]. In this study, the scenes are flagged as cloudy when the radar reflectivity exceeds 5 dBZ. To assess the sensitivity of each CH to clouds, the probability density function (PDF) of BTs is examined for clear and cloudy scenes, as shown in Fig. 2. It seems that the presence of clouds tends to decrease the observed BTs over the land. The mean value and the standard deviation in cloudy scenes are greater than that in clear scenes, especially for the window CHs 1 and 10, temperature-sounding CHs 7–9, and humidity-sounding CHs 13–15. The rather different distributions for these CHs are quite promising for detecting the cloud contamination.

III. Algorithm for CDL

A. GBDT Algorithm

The GBDT [26] is an iterative decision tree algorithm and is also known as multiple additive regression tree. Due to its

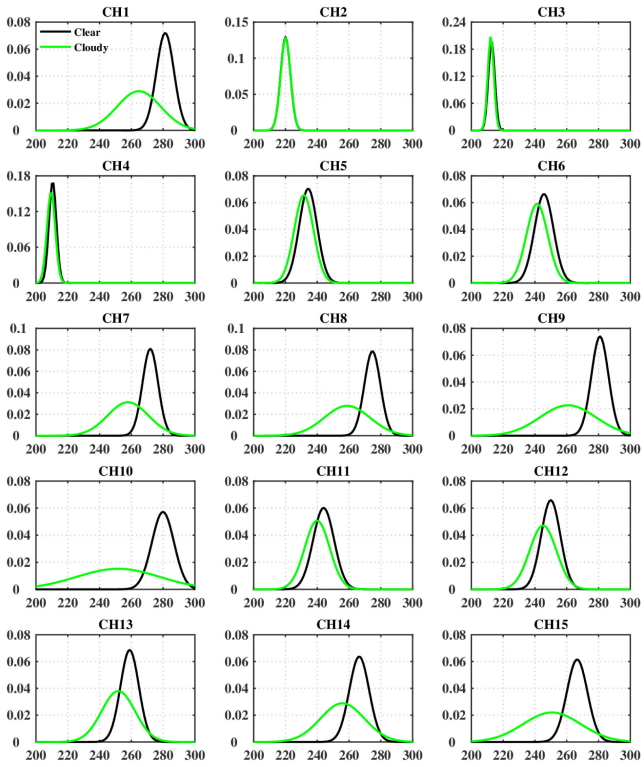

Fig. 2. PDF of BT under clear (black solid line) and cloudy (green solid line) scenes for 15 CHs of MWHS-2.

strong generalization ability, the GBDT has been widely used in many ML scenarios, such as click-through rate estimation, search ranking, and commodity sales forecasting [27]–[30].

The GBDT is based on a boosting strategy, which is a primary method of ensemble learning [31]. It constructs a set of weak learners (trees) and accumulates the results of multiple decision trees as the final predicted output. Specifically, the new decision tree learns the error residuals of all previous trees in each iteration, thereby generating a stronger base model. The boosting tree model can be expressed as an additive model of the decision tree

$$f_M(x) = \sum_{m=1}^{M} T(x, \Theta_m) \tag{1}$$

where $T(x, \Theta_m)$ is the base decision tree model; x is the feature vector; Θ_m is the parameter of a decision tree; and M is the number of trees. The boosting tree model algorithm proceeds as follows

Initialize the first base model

$$f_0(x) = 0. \tag{2}$$

For $m = 1$–M, calculate the error residual

$$r_{mi} = y_i - f_{m-1}(x_i), \quad i = 1, 2, \ldots, N \tag{3}$$

where N is the sample size, and y is the label of x.

Fit the error residual r_{mi} to learn a regression tree and obtain $T(x, \Theta_m)$. Update

$$f_m(x) = f_{m-1}(x) + T(x, \Theta_m). \tag{4}$$

TABLE II
SUMMARY OF TUNING PARAMETERS AND THEIR DYNAMIC RANGES

Parameter and dynamic range
1. max number of leaves in one tree (num_leaves) [50, 100, 150]
2. maximum depth of the tree (max_depth) [7, 10, 15]
3. max number of bins (max_bin) [100, 150, 200, 255]
4. minimal number of data in one leaf (min _leaf) [100, 150, 200]
5. fraction of features randomly selected on each tree (feature_fraction) [0.6, 0.8, 1.0]
6. specifies the fraction of data to be used for each iteration and is generally used to speed up the training and avoid overfitting (bagging_fraction) [0.6, 0.8, 1.0]
7. shrinkage rate (learning_rate) [0.01, 0.02, 0.05, 0.1, 0.2, 0.5]
8. number of boosting iteration (n_estimators) [100, 200, 300, 400, 500]

In the above algorithm, the most important step is to calculate the error residual r_{mi}. For various loss functions L, GBDT uses the idea of steepest descent, that is, it uses the negative gradient of the loss function to approximate the residuals, thus obtaining a general framework

$$-\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right] f(x) = f_{m-1}(x), \tag{5}$$

LightGBM [32] is the most successful and advanced gradient boosting framework that uses the GBDT-based learning algorithm. It is designed to be distributed and efficient with many advantages, such as good accuracy, fast training speed, high efficiency, and low memory usage. In many cases, although the performance of LightGBM is problem dependent, this algorithm has been found to be more accurate and faster than the other GBDT tools, including the Scikit-learn[2] and the XGboost [33]. In this article, we use LightGBM to learn the cloud classification model.

Generally, we should tune the parameters of GBDT because it has many parameters that may affect the performance of the model. In this study, eight tuning parameters are used, and their dynamic ranges are summarized in Table II. The grid search is used to find the optimal parameter combination, which means the algorithm needs to be tuned iteratively 29 160 ($3 \times 3 \times 4 \times 3 \times 3 \times 3 \times 6 \times 5$) times, as indicated in Table II.

B. Training and Testing Datasets for CDL

The training dataset in this study is 53 460 for 61 days from July to August 2016, and the testing dataset is 6571 for 7 days in September 2016. Particularly, the optimal parameters are estimated with five-fold cross validation using the original training data for the robustness. Fig. 3 shows the PDFs for training (blue solid line) and testing (red solid line) datasets, respectively. It can be seen that the PDFs of training and testing datasets are similar with respect to the radar reflectivity (label).

We primarily use four algorithms to train the CDL prediction model based on different frequency ranges, including the window CHs (CHs at 89 and 150 GHz) algorithm (win-algorithm),

[2][Online]. Available: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

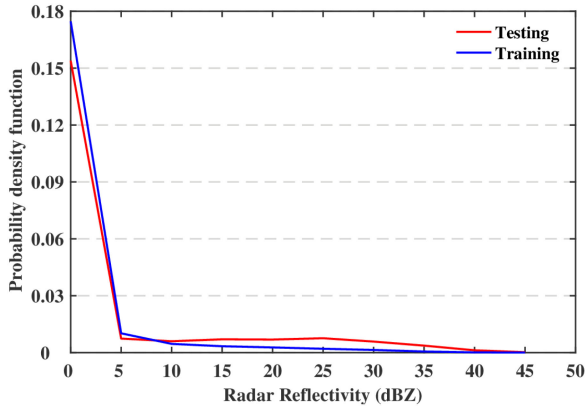

Fig. 3. PDFs of training (blue solid line) and testing (red solid line) datasets with respect to the radar reflectivity.

TABLE III
FEATURES FOR THE FOUR ALGORITHMS

Features	Win-algorithm	Hum-algorithm	Tem-algorithm	All-algorithm
BT(89.0 GHz)	✓			✓
BT(150.0 GHz)	✓			✓
BTD(150.0 − 89.0 GHz)	✓			✓
BT(183.31±3 GHz)		✓		✓
BT(183.31±4.5 GHz)		✓		✓
BT(183.31±7 GHz)		✓		✓
BTD(183.31±3 − 183.31±1 GHz)		✓		✓
BTD(183.31±7 − 183.31±1 GHz)		✓		✓
BTD(183.31±7 − 183.31±3 GHz)		✓		✓
BT(118.75±2.5 GHz)			✓	✓
BT(118.75±3.0 GHz)			✓	✓
BT(118.75±5.0 GHz)			✓	✓
BTD(118.75±3.0 − 118.75±2.5 GHz)			✓	✓
BTD(118.75±5.0 − 118.75±2.5 GHz)			✓	✓
BTD(118.75±5.0 − 118.75±3.0 GHz)			✓	✓
latitude	✓	✓	✓	✓
zenith angle	✓	✓	✓	✓
azimuth	✓	✓	✓	✓

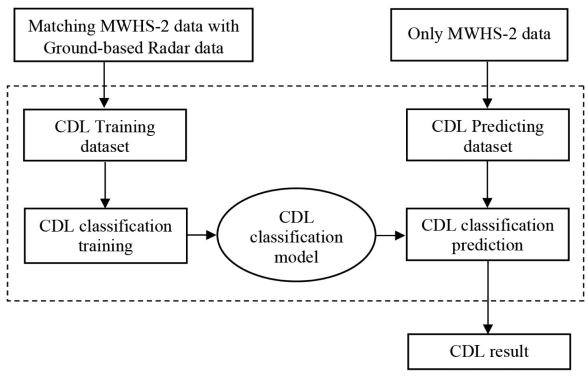

Fig. 4. Strategy and flowchart of the CDL algorithm.

humidity CHs (CHs around 183.31 GHz) algorithm (hum-algorithm), temperature CHs (CHs around 118.75 GHz) algorithm (tem-algorithm), and all frequency ranges (all-algorithm), as indicated in Table III. Besides the BT and BTD, the latitude, zenith angle, and azimuth are also considered into the

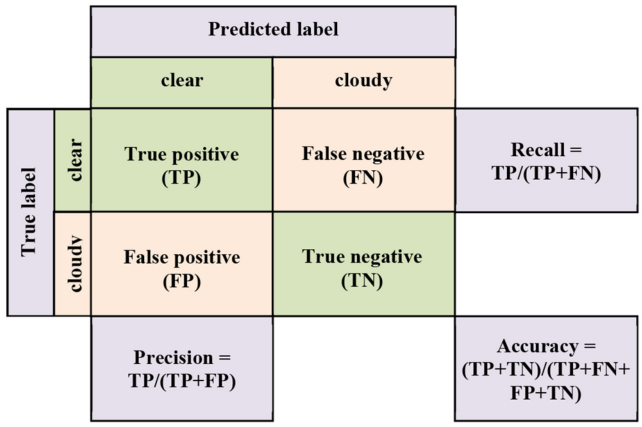

Fig. 5. Confusion matrix and evaluation metrics.

algorithms. Fig. 4 displays the processing flowcharts of CDL training and prediction.

C. Model Performance Metrics

The algorithms are evaluated quantitatively based on the accuracy, precision, recall, area under the curve (AUC), and log loss. The accuracy, precision, and recall metrics are calculated according to the confusion matrix, as shown in Fig. 5. Furthermore, the F_1 score is also calculated in order to find the optimum balance or harmonic mean between the precision and recall

$$F_1 = \frac{2{*}\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad . \tag{6}$$

AUC is the area covered by the receiver operating characteristic curve. The meaning of AUC is the probability that the prediction result of the positive sample is greater than that of the negative sample. Therefore, AUC represents the capability of the classifier to sort the samples. The larger the AUC is, the better the classification effect is.

Log loss is defined based on the probability estimates. And it is also called the logistic regression loss or cross-entropy loss. Regarding the binary classification with a probability estimate $p = \Pr(y = 1)$, the log loss per sample is calculated with the negative log-likelihood of the classifier when the true label $y \in \{0, 1\}$ is given

$$\begin{aligned} L_{\log}\ (y, p) &= -\log \Pr\ (y|p) \\ &= -\left(y \log\left(p\right) + (1-y) \log\left(1-p\right)\right). \end{aligned} \tag{7}$$

The closer the log loss is to 0, the better the model performs.

D. Optimal Prediction Model Parameter Selection

For each algorithm, the GBDT model has been tuned for 29 160 $\times$ basing on the eight parameters of num_leaves, bagging_fraction, learning_rate, *n*_estimators, max_depth, max_bin, min_leaf, and feature_fraction (see Table II). And the optimal model is decided when log loss, which is widely used as an effective reference, gets the minimum value. Finally, the selection of optimal parameters for the four algorithms is indicated in Table IV.

TABLE IV
OPTIMAL PREDICTION MODEL PARAMETERS SELECTION FOR THE FOUR ALGORITHMS

	num_leaves	max_depth	max_bin	min_leaf	feature_fraction	bagging_fraction	learning_rate	n_estimators
Win-algorithm	100	15	100	100	0.8	0.6	0.05	300
Hum-algorithm	50	15	100	100	0.6	0.6	0.02	500
Tem-algorithm	100	15	100	100	0.6	0.6	0.02	500
All-algorithm	100	15	100	100	0.8	0.6	0.02	500

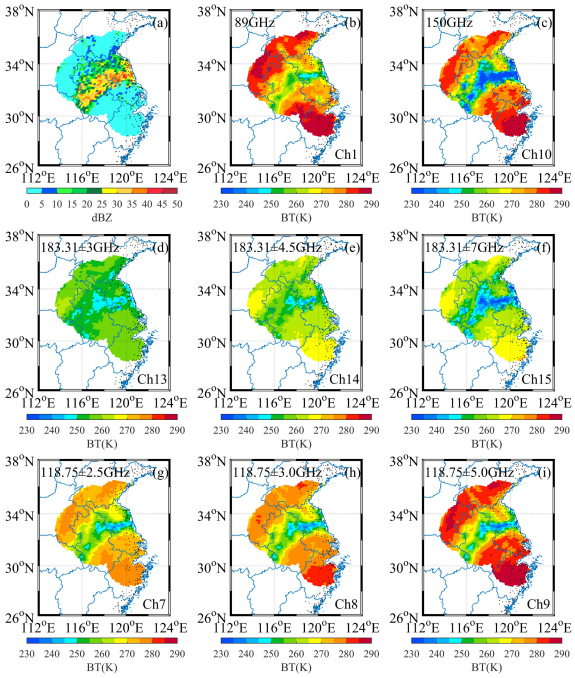

Fig. 6. Spatial distributions of (a) radar reflectivity and MWHS-2 BT observations for (b) CH 1, (c) CH 10, (d)–(f) CHs 13–15, and (g)–(i) CHs 7–9 on July 4, 2016.

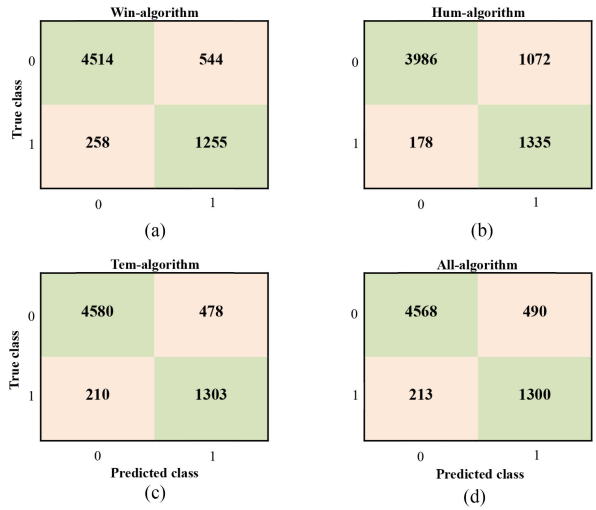

Fig. 7. Confusion matrix of the four algorithms for the testing dataset. 0 denotes the clear scene and 1 denotes the cloudy scene. Green boxes represent the number of pixels that are correctly classified for each of the classes and light orange boxes represent the false classifications.

IV. RESULTS AND DISCUSSION

A. Observations of the BT and Radar Reflectivity

The spatial distributions of radar reflectivity and BT observations for CH 1, CHs 7–10, and CHs 13–15 have been displayed in Fig. 6. It can be seen that the spatial distribution of BT corresponds well with the radar reflectivity. In clear scenes (0–5 dBZ), the BTs are ∼285 K for CHs 1, 9, and 10, ∼275 K for CHs 7 and 8, and between 255 and 265 K for CHs 13–15. Generally, BT in the south region is ∼5 K larger than that in the north region, which can be explained by the effect of latitude, and the fact that the temperature is higher in the region closer to the equator in summer. In cloudy regions (>5 dBZ), the depressions of BTs increase significantly with the radar reflectivity due to the scattering effect by the cloud hydrometeors. And the BTs are almost as low as ∼230 K when the radar reflectivity is above 25 dBZ.

B. Evaluation of the CDL Results for MWHS-2

The confusion matrices (see Fig. 7) present the results of CDL, demonstrating for each true class (y-axis) versus the predicted class (x-axis), where 0 denotes the clear scene and 1 denotes the cloudy scene. In the figure, each 2×2 matrix shows the overall class statistics. The diagonal of the confusion matrix exhibits the correctly classified numbers for clear (true positive, TP) and cloudy (true negative, TN) scenes. It shows that the tem-algorithm outperforms the other algorithms with the relatively high TP value of 4580 and the TN value of 1303. The hum-algorithm performs worst and has the wrongly classified numbers of clear scenes (false negative, FN) value as high as 1072, which is about twice that of the other algorithms. This may be attributed to the fact that 183.31 GHz CHs are responsive to both water vapor and cloud hydrometeors [8], [35]. So, it is difficult for hum-algorithm to explicitly distinguish the high humidity in clear scenes or the cloud hydrometeors in cloudy scenes. A more detailed evaluation of the four algorithms is described as follows.

TABLE V
EVALUATION METRICS OF FOUR ALGORITHMS FOR THE TESTING DATASET

Algorithm name	F_1	Accuracy	AUC	Log loss
Tem-algorithm	0.930	0.895	0.883	0.36
All-algorithm	0.928	0.893	0.881	0.28
Win-algorithm	0.918	0.878	0.861	0.32
Hum-algorithm	0.864	0.810	0.835	0.36

Table V summarizes the evaluation metrics for the CDL prediction model, divided on the basis of the win-algorithm, hum-algorithm, tem-algorithm, and all-algorithm. Obviously, the tem-algorithm model shows superiority over the other three

TABLE VI
SPLIT NUMBERS OF FEATURES IN FOUR ALGORITHMS OF THE OPTIMAL CDL MODEL AND THEIR CORRESPONDING RANKING

Features	Win-algorithm		Hum-algorithm		Tem-algorithm		All-algorithm	
	Split number	Ranking	Split number	Ranking	Split number	Ranking	Split number	Ranking
BT (89.0 GHz)	10848	4					3345	8
BT (150.0 GHz)	10801	6					2629	13
BTD (150.0 − 89.0 GHz)	10651	5					3740	5
BT (183.31 ± 3 GHz)			3373	9			2972	10
BT (183.31 ± 4.5 GHz)			3473	8			2761	12
BT (183.31 ± 7 GHz)			4432	5			2559	14
BTD (183.31 ± 3 − 183.31 ± 1 GHz)			4368	6			3898	4
BTD (183.31 ± 7 − 183.31 ± 1 GHz)			3768	7			3186	9
BTD (183.31 ± 7 − 183.31 ± 3 GHz)			5943	2			3605	7
BT (118.75 ± 2.5 GHz)					1073	7	2267	18
BT (118.75 ± 3.0 GHz)					1442	6	2364	16
BT (118.75 ± 5.0 GHz)					1827	4	2315	17
BTD (118.75 ± 3.0 − 118.75 ± 2.5 GHz)					973	9	3724	6
BTD (118.75 ± 5.0 − 118.75 ± 2.5 GHz)					1034	8	2545	15
BTD (118.75 ± 5.0 − 118.75 ± 3.0 GHz)					1575	5	2910	11
latitude	12635	1	5665	4	2088	3	5903	1
zenith angle	11712	2	6586	1	2299	1	4204	3
azimuth	11053	3	5770	3	2214	2	4398	2

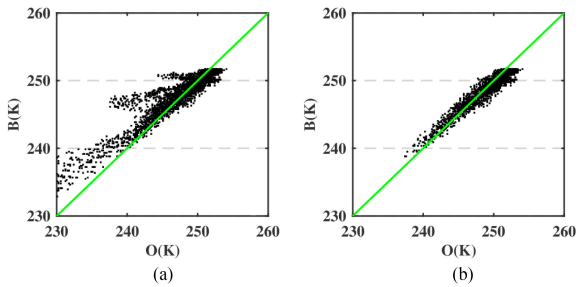

Fig. 8. Comparison between the observed BTs (O) (x-label) and simulated BTs (B) (y-label) (a) before and (b) after the cloud filtering by the CDL. Additionally, B is calculated using the CRTM in clear sky scenarios.

models with the maximum values of F_1 (0.93), accuracy (0.895), and AUC (0.883). It should be noted that the performance of the all-algorithm is suboptimal to the tem-algorithm regarding the evaluation metrics of F_1, accuracy, and AUC, even though the all-algorithm has the maximum features. This might stem from the fact that the temperature-sounding CHs around 118.75 GHz are most sensitive to cloud properties and outperform the other CHs, whereas the log loss of all-algorithm has a minimum value of 0.28. In addition, the hum-algorithm shows inferior prediction capability when compared with the win-algorithm.

Fig. 8 displays the comparison between the observed BTs (O) and simulated BTs (B), which are calculated by the community radiative transfer model (CRTM) in clear sky scenarios for MWHS-2. Before cloud filtering, there are many outliers existing between O and B, as indicated in Fig. 8(a). However, the number of outliers has been significantly reduced after removing the cloudy radiances identified by the CDL. Furthermore, the linear correlation coefficient between O and B increases from 0.86 to 0.92, and the maximum of Z-score, which is calculated with the biweight mean and standard deviation [35], [36], is reduced from 6.98 to 3.01 after the cloud filtering.

Table VI lists the split number of the features for the win-algorithm, hum-algorithm, tem-algorithm, and all-algorithm, respectively. The split number can be calculated after the GBDT model fits as a weight factor of every feature. It is apparent that the latitude ranks the first for the win-algorithm and all-algorithm, and the zenith angle ranks the first for the hum-algorithm and tem-algorithm. The strong latitude dependence may be attributed to the effect of surface emissivity and season, which can be evidenced from Fig. 6. The significant dependence of CDL on the zenith angle can be explained by the fact that the length of the optical path for a cross-track scanning radiometer varies with the zenith angle, which is also called the limb effect. Besides, the ranking of azimuth is also very high, showing a significant connection to CDL.

Apart from the latitude, zenith angle, and azimuth, the BT and BTD are also important features for CDL. For the all-algorithm, the ranking of the split number for BTD is higher than that for BT in general. Specifically, the BTD (183.31 ± 3 − 183.31 ± 1 GHz), BTD (150.0 − 89.0 GHz), and BTD (118.75 ± 3.0 − 118.75 ± 2.5 GHz) rank 4, 5, and 6, respectively. For the hum-algorithm, BTD (183.31 ± 3 − 183.31 ± 1 GHz) and BT at 183.31 ± 7 GHz are relatively top factors, and the BTs are of minor importance in comparison with the BTDs, with the average rankings for the BTDs and BTs of 5 and 7, respectively. Regarding the tem-algorithm, the weighting factor for BT (118.75 ± 5.0 GHz) and BTD (118.75 ± 5.0 − 118.75 ± 3.0 GHz) ranks 4 and 5, respectively. For the win-algorithm, the BT at 89 GHz contributes more than BTD (150.0 − 89.0 GHz) and BT at 150 GHz.

V. SUMMARY

The less sensitivity to clouds for MW radiation when compared with the visible-infrared radiation, and the complex and variable surface emissivity make the CDL much more

challenging than that over the ocean. Since the assimilation of MW measurements in the operational NWP model is independent of the other instrument measurements, a standalone algorithm CDL for MWHS-2 is proposed in this study. It is based on the GBDT algorithm and is trained on the CINRAD observations. The CDL has been investigated by employing the win-algorithm, hum-algorithm, tem-algorithm, and all-algorithm. The model has been tuned iteratively 29 160 times for each algorithm to find the optimal prediction model parameters.

The CDL estimates are reasonably comparable with the cloud mask from CINRAD measurements. It is recommended that the new temperature-sounding CHs around 118.75 GHz are very important for fitting a prediction model for CDL, as significant improvements have been found for tem-algorithm over the land. Specifically, the evaluation metrics for the tem-algorithm are optimal among the four algorithms, including the F_1 score (0.93), accuracy (0.895), and AUC (0.883). It is noticeable that the tem-algorithm even outperforms the all-algorithm that has the maximum number of features, and this is contradictory to the study by Chai *et al.* [12]. Compared with the algorithm for window CHs, the algorithm for CHs around 183.31 GHz shows inferior cloud property prediction capability over the land, which is consistent with the results from Chen *et al*. [11]. For all four algorithms, the latitude, zenith angle, and azimuth are the top-ranking features and have relatively high split numbers. Additionally, after removing the cloudy radiances identified by the CDL, the linear correlation coefficient between O and B can reach 0.96 and the maximum of Z-score is reduced to 3.01, which is promising for satellite data assimilation in the forthcoming work.

Although the ML-based CDL algorithm is able to detect cloud contamination with high accuracy, the disadvantages of the CDL algorithm are still evident. To some extent, the spatial and temporal distribution of the sample in this study is limited. What is more, the weather radar measurements that are used as the label in this algorithm are unable to detect the thin clouds under nonprecipitating conditions accurately. However, the important point is that the MW radiation can penetrate most nonprecipitating clouds and shows extremely weak sensitivity to thin clouds [37]. Thus, it is not worth taking them into account in cloud detection because this cloud information has a little impact on the MW measurements [11]. Future work will focus on multiclass and multilayered cloud classification. Overall, the ML-based cloud detection method presented in this study can be used in the quality control processes of assimilating the MW radiances in the NWP center and be applied for performing MW retrievals of atmospheric and surface parameters over the clouds. The CDL algorithm is also suitable for other MW sounders, such as FY-3D.

References

[1] F. Aires, F. Marquisseau, C. Prigent, and G. Sèze, "A land and ocean microwave cloud classification algorithm derived from AMSU-A and -B, trained using MSG-SEVIRI infrared and visible observations," *Monthly Weather Rev.*, vol. 139, pp. 2347–2366, 2011.

[2] H. Han *et al.*, "Microwave sounder cloud detection using a collocated high-resolution imager and its impact on radiance assimilation in tropical cyclone forecasts," *Monthly Weather Rev.*, vol. 144, pp. 3937–3959, 2016.

[3] N. Grody, J. Zhao, R. Ferraro, F. Weng, and R. Boers, "Determination of precipitable water and cloud liquid water over oceans from the NOAA-15 advanced microwave sounding unit," *J. Geophys. Res. Atmos.*, vol. 106, pp. 2943–2953, 2001.

[4] F. Weng, L. Zhao, R. R. Ferraro, G. Poe, X. Li, and N. C. Grody, "Advanced microwave sounding unit cloud and precipitation algorithms," *Radio Sci.*, vol. 38, no. 4, 2003, Art. no. 8068.

[5] R. Bennartz, A. Thoss, A. Dybbroe, and D. B. Michelson, "Precipitation analysis using the advanced microwave sounding unit in support of nowcasting applications," *Meteorol. Appl.*, vol. 9, no. 2, pp. 177–189, 2002.

[6] L. Zhao and F. Weng, "Retrieval of ice cloud parameters using the advanced microwave sounding unit (AMSU)," *J. Appl. Meteorol.*, vol. 41, pp. 384–395, 2002.

[7] Y. Han, X. Zou, and F. Weng, "Cloud and precipitation features of super typhoon Neoguri revealed from dual oxygen absorption band sounding instruments on board FengYun-3C satellite," *Geophys. Res. Lett.*, vol. 42, pp. 916–924, 2015.

[8] S. A. Buehler *et al.*, "A cloud filtering method for microwave upper tropospheric humidity measurements," *Atmos. Chem. Phys.*, vol. 7, pp. 5531–5542, 2007.

[9] H. Ishida, Y. Oishi, K. Morita, K. Moriwaki, and T. Y. Nakajima, "Development of a support vector machine based cloud detection method for MODIS with the adjustability to various conditions," *Remote Sens. Environ.*, vol. 205, pp. 390–407, 2018.

[10] F. Xie, M. Shi, Z. Shi, J. Yin, and D. Zhao, "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3631–3640, Aug. 2017.

[11] Y. Chen, R. Fan, M. Bilal, X. Yang, J. Wang, and W. Li, "Multilevel cloud detection for high-resolution remote sensing imagery using multiple convolutional neural networks," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 5, 2018, Art. no. 181.

[12] D. Chai, S. Newsam, H. K. Zhang, Y. Qiu, and J. Huang, "Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks," *Remote Sens. Environ.*, vol. 225, pp. 307–316, 2019.

[13] Y. Chen *et al.*, "Cloud and cloud shadow detection based on multiscale 3D-CNN for high resolution multispectral imagery," *IEEE Access*, vol. 8, pp. 16505–16516, 2020.

[14] T. Islam *et al.*, "CLOUDET: A cloud detection and estimation algorithm for passive microwave imagers and sounders aided by naïve Bayes classifier and multilayer perceptron," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 9, pp. 4296–4301, Sep. 2015.

[15] S. Favrichon, C. Prigent, C. Jimenez, and F. Aires, "Detecting cloud contamination in passive microwave satellite measurements over land," *Atmos. Meas. Techn.*, vol. 12, no. 3, pp. 1531–1543, 2019.

[16] P. Bauer and A. Mugnai, "Precipitation profile retrievals using temperature-sounding microwave observations," *J. Geophys. Res.*, vol. 108, 2003, Art. no. 4730.

[17] T. Islam, P. K. Srivastava, Q. Dai, and M. Gupta, "Ice cloud detection from AMSU-A, MHS, and HIRS satellite instruments inferred by cloud profiling radar," *Remote Sens. Lett.*, vol. 5, no. 12, pp. 1012–1021, 2014.

[18] J. Han *et al.*, "The establishment of optimal ground-based radar datasets by comparison and correlation analyses with space-borne radar data," *Meteorol. Appl.*, vol. 25, no. 1, pp. 161–170, 2018.

[19] Z. Chu *et al.*, "Mitigating spatial discontinuity of multi-radar QPE based on GPM/KuPR," *Hydrology*, vol. 5, no. 3, 2018, Art. no. 48.

[20] T. Wu, Y. Wan, W. Wo, and L. Leng, "Design and application of radar reflectivity quality control algorithm in SWAN," *Meteorol. Sci. Technol.*, vol. 41, no. 5, pp. 809–817, 2013.

[21] S. Liu, Z. Chu, Y. Yin, and R. Liu, "Evaluation of MWHS-2 using a co-located ground-based radar network for improved model assimilation," *Remote Sens.*, vol. 11, 2019, Art. no. 2338.

[22] B. E. Martner and K. P. Moran, "Using cloud radar polarization measurements to evaluate stratus cloud and insect echoes," *J. Geophys. Res.*, vol. 106, no. D5, pp. 4891–4897, 2001.

[23] G. Hong, G. Heygster, J. Miao, and K. Kunzi, "Detection of tropical deep convective clouds from AMSU-B water vapor channels measurements," *J. Geophys. Res.*, vol. 110, 2005, Art. no. D05205, doi: 10.1029/2004JD004949.

[24] S. Manandhar, F. Yuan, Y.-H. Lee, and Y. S. Meng, "Weather radar to detect and differentiate clouds from rain events," in *Proc. USNC-URSI Radio Sci. Meeting*, 2016, pp. 103–104.

[25] Z. Wang, Z. Wang, X. Cao, and F. Tao, "Comparison of cloud top heights derived from FY-2 meteorological satellites with heights derived from ground-based millimeter wavelength cloud radar," *Atmos. Res.*, vol. 199, pp. 113–127, 2018.

[26] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
[27] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, pp. 367–378, 2002.
[28] M. Richardson, E. Dominowska, and R. Ragno, "Predicting clicks: Estimating the click-through rate for new ads," in *Proc. 16th Int. Conf. World Wide Web Conf.*, 2007, pp. 521–530.
[29] L. Ping, "Robust logitboost and adaptive base class (ABC) logitboost," in *Proc. 26th Conf. Uncertain. Artif. Intell.*, 2010, pp. 302–311.
[30] M. Min, J. Li, F. Wang, Z. Liu, and W. P. Menzel, "Retrieval of cloud top properties from advanced geostationary satellite imager measurements based on machine learning algorithms," *Remote Sens. Environ.*, vol. 239, 2020, Art. no. 111616.
[31] G. T. Dietterich, "Ensemble learning," in *The Handbook of Brain Theory and Neural Networks*, vol. 2. Cambridge, MA, USA: MIT Press, 2002, pp. 110–125.
[32] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 3146–3154, 2017.
[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.
[34] O. Bobryshev, S. A. Buehler, V. O. John, M. Brath, and H. Brogniez, "Is there really a closure gap between 183.31-GHz satellite passive microwave and *in situ* radiosonde water vapor measurements?" *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2904–2910, May 2018.
[35] J. R. Lanzante, "Resistant, robust and nonparametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data," *Int. J. Climatol.*, vol. 16, pp. 1197–1226, 1996.
[36] X. Zou and Z. Zeng, "A quality control procedure for GPS radio occultation data," *J. Geophys. Res., Atmos.*, vol. 111, no. D2, 2006, Art. no. D02112.
[37] F. Weng, "Advances in radiative transfer modeling in support of satellite data assimilation," *J. Atmos. Sci.*, vol. 64, pp. 3799–3807, 2007.

Shuxian Liu received the B.S. degree in atmospheric physics, in 2016 from the Nanjing University of Information Science and Technology, Nanjing, China, where she is currently working toward the Ph.D. degree with the School of Atmospheric Physics.

Her research interests include satellite data assimilation and cloud detection.

Yan Yin received the Ph.D. degree in atmospheric physics from Tel Aviv University, Tel Aviv, Israel, in 1999.

From 1999 to 2004, he was with the University of Leeds, U.K., and from 2004 to 2005, he was a Lecturer with the University of Wales, Aberystwyth, U.K. He is currently a Professor of Atmospheric Physics and Atmospheric Environment with the Nanjing University of Information Science and Technology, Nanjing, China. His main research works are focused on cloud and precipitation physics, aerosol properties and their effects on environment and climate, and weather modification. He has authored or coauthored more than 250 scientific papers, including 150 published in SCI journals and about 50 in EI journals, and six books.

Zhigang Chu received the M.S. degree in atmospheric sounding and the Ph.D. degree in atmospheric physics from the Nanjing University of Information Science and Technology, Nanjing, China, in 2009 and 2013, respectively.

He is currently a Lecturer with the School of Atmospheric Physics, Nanjing University of Information Science and Technology. His research interests include remote sensing for understanding and quantifying weather and cloud/precipitation microphysics.

Shuai An received the M.S. degree in computer science from Nankai University, Tianjin, China, in 2018. He is currently working toward the Ph.D. degree in database theory at the School of Informatics, University of Edinburgh, Edinburgh, U.K.

From 2018 to 2020, he was an Algorithm Engineer with JD.COM, Beijing, China. His research interests include database theory and machine learning.