A Hybrid Algorithm for Mineral Dust Detection Using Satellite Data

Peichang Shi
Dept. Information Systems
U. of Maryland, Baltimore County
pshi1@umbc.edu

Qianqian Song
Dept. Physics
U. of Maryland, Baltimore County
cd11735@umbc.edu

Janita Patwardhan
Dept. Mathematics and Statistics
U. of Maryland, Baltimore County
janita1@umbc.edu

Zhibo Zhang
Dept. Physics
U. of Maryland, Baltimore County
zzbatmos@umbc.edu

Jianwu Wang
Dept. Information Systems
U. of Maryland, Baltimore County
jianwu@umbc.edu

Aryya Gangopadhyay
Dept. Information Systems
U. of Maryland, Baltimore County
gangopad@umbc.edu

***Abstract*—Mineral dust, defined as aerosol originating from the soil, can have various harmful effects to the environment and human health. The detection of dust, and particularly incoming dust storms, may help prevent some of these negative impacts. In this paper, using satellite observations from Moderate Resolution Imaging Spectroradiometer (MODIS) and the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation Observation (CALIPSO), we compared several machine learning algorithms to traditional physical models and evaluated their performance regarding mineral dust detection. Based on the comparison results, we proposed a hybrid algorithm to integrate physical model with the data mining model, which achieved the best accuracy result among all the methods. Further, we identified the ranking of different channels of MODIS data based on the importance of the band wavelengths in dust detection. Our model also showed the quantitative relationships between the dust and the different band wavelengths.**

***Index Terms*—hybrid dust detection, data mining, physical model, satellite data, feature importance**

I. Introduction

In arid and dry regions with high velocity winds, soil particles are lifted into the atmosphere, becoming mineral dust. It is one of the most abundant types of aerosol in the atmosphere with the Saharan desert as the largest contributor. Mineral dust aerosols affect the Earth's energy budget through several ways. It has a direct radioactive effect by scattering and absorbing solar radiation. By acting as cloud nucleation nuclei, mineral dust can indirectly impact the global radiation balance. High levels of mineral dust results a significant decrease in the air quality, negatively affecting our health. Inhalation of large quantities of mineral dust can lead to lung fibrotic diseases (where damage occurs to the lung tissue) as well as an increase in hospital admissions due to aggravated asthma, chronic bronchitis, and other respiratory illnesses [7]. Unfortunately, the amount of dust in the atmosphere and its direct impact is unknown largely due to errors in the methods of retrieval.

Many of the methods for dust detection rely upon the usage of satellite data. The more accurate data has been from the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation Observation (CALIPSO). While CALIPSO is more accurate at dust detection, it has multiple drawbacks like only gathering data from a smaller swath of the Earth's surface. Researchers have shifted towards using data from Moderate Resolution Imaging Spectroradiometer (MODIS), which is a passive sensor. However, MODIS is unable to directly detect mineral dust. Thus various algorithms have been developed combining physical knowledge of mineral dust and the data captured by MODIS to calculate the probability of dust [4], [1], [5], [6], [8]. Unfortunately, many of these algorithms have a lower detection rate or accuracy rate than desired.

In this paper, we want to leverage accurate dust detection capability of CALIPSO satellite dataset to predict dust for the large volume of MODIS satellite dataset which covers the whole earth. To do it, we mainly tackled the following two challenges. First, how to best integrate existing physics model based dust detection approach and data-driven machine learning algorithms for the best results? Second, how to identify the importance of different channels at the MODIS instrument for dust detection, which could provide important recommendations for future dust detection instrument construction?

In summary, the contributions of this paper are three folds.

- We compared several machine learning algorithms to traditional physical models and evaluated their performance regarding mineral dust detection for MODIS data.
- Based on the comparison results, we proposed a hybrid algorithm to combine machine learning techniques with a physics background to develop an algorithm with around 90% accuracy rate, which is the best accuracy result among all available methods.
- Based on the Lasso variable selection approach, we identified the ranking of different channels/bands of MODIS data based on the importance of the band wavelengths in dust detection.

The remainder of the paper is organized as follows. In

Section II, we discuss related studies and their differences from our proposed approach. Section IV discusses and compares different physics-based algorithms and machine learning algorithms for dust detection. Section V explains our channel importance ranking approach. Our main hybrid dust detection algorithm is presented in Section VI and evaluated in Section VII. Last, we conclude in Section VIII with some ideas for future work.

II. Related Work

Dust detection using MODIS data has been reported in many papers [17], [18], [20], [25], [27]. The commonly used algorithms include Support Vector Machine (SVM), Artificial neural network (ANN), decision tree, random forests, multiple regression etc. Logistic regression is seldom reported. In most papers, they compared physical models to their machine learning approach. Some of them tried satellite image analysis for dust analysis [17], [18], [20], [25], [27]. But none of these papers mentioned the feature engineering, one reason is that people used the same variables from the physical algorithms; another reason is that some machine learning methods don't require feature selection, such as random forests. One draw back for most data mining methods is that they are black-box models, such as neural network. With a number of hidden layers, it is hard to interpret the relationship between the input variables and outcome, and the whole process works like a black box.

Different from the above approaches, we proposed a hybrid algorithm by combining physics-based model and data-driven models. Specifically, our uniqueness lies as follows. 1) We quantitatively demonstrated the relationship between input variables and dust outcome. Our approach not only shows the importance of each variable, but also shows their contribution to the dust. For example, the coefficient for band 20 is 0.85, which means one unit increase for band 20 would lead to the probability of being dust 2.3 times higher than not being dust; 2) Through feature engineering, we identified the importance of input variables based on their entering steps during variable selection. For example, in Lasso selection, band 32 entered the model first, which indicates that band 32 is the most important variable in dust detection. The importance of band 32 was also reported in several other literature [1], [33]; 3) Combination of physical algorithm and data mining approach. Besides the pure data driven results from logistic regression, we added physical understanding from physical algorithm, which could strengthen the model interpretation.

III. Satellite Data

MODIS is a passive sensor onboard the Terra satellite since 1999 and the Aqua satellite since 2002 launched by NASA. With a viewing swath of 2,330 km, it images the entirety of the earth at most every two days. It measures data in 36 spectral bands, ranging from 0.045 to 14.385 μm, at three different spatial resolutions, 250m, 500m, and 1km. The data can be accessed at various levels, depending on the information requested. In this study, we use MODIS level-1 data, which contains calibrated and geolocated radiance observation in the 36 spectral bands. The information is stored in HDF files, each containing approximately 5 minutes of MODIS data referred to as a $granule$. The MODIS observation have been applied to derive a large variety of remote sensing products, from land vegetation coverage to sea surface temperature, from ice and snow extend to aerosol and cloud distributions. Here in this study, of particular interest in the detection of air-borne mineral dust aerosols using MODIS observations. In the past, a number of algorithms have been developed based on physical principles to detect dust aerosols based on satellite observations. Most of these algorithms are based on the analysis of the reflectance of sunlight in the visible bands and the brightness temperature of the thermal emission in the infrared region. For example, Zhao et al. [8] developed a physically-based algorithm which tests a variety of the optical properties of the target, i.e., brightness, color, and temperature to determine if it is dust aerosol. Such algorithms face several challenges. First of all, it is difficulty to validate and evaluate the detection results. Secondly, the algorithms relies on the test of multiple threshold, which often lead to artificial effects, such as abrupt discontinuity. Furthermore, the development of such algorithm often involves large amounts of trial-error testing and fine-tuning, which is tedious and time consuming. We wanted to be able to validate our results from the MODIS data using observations from CALIPSO. As both CALIPSO and Aqua are among the international satellites along the same orbital track called the A-Train, we decided to use MODIS data from Aqua.

The CALIPSO satellite, which is a joint venture between NASA and its French counterpart CNES, has been recording data as a part of the A-train as of 2006. Among its three instruments, it has a lidar sensor, called Cloud-Aerosol Lidar with Orthogonal Polarization. Different from the above-mentioned MODIS, CALIPSO is an active senor. It measures the reflection, refraction, and scattering of its own transmitted lidar signals by the Earth's surface and atmosphere. CALIPSO measures the strength of the reflected lidar signals in two bands, the 532 nm and 1064 nm bands. In addtion, it also measures the so-called depolarization ratio of the lidar signal in the 532 nm band. If the aerosols particles are spherical, such as sulphate and smoke aerosols, then their scattered lidar signals have near zero depolarization ratio. In contrast, the scattering of the non-spherical aerosol particles, such as dust, have significant depolarization. Therefore, using the observed lidar depolarization, it is easy to detect the dust aerosols distinguish them from other types of aerosols. Through this use of depolarization, it is able to better detect clouds and dust aerosols. However, as seen in Figure 1, it covers much less area than MODIS, which is why we would like to use MODIS data to detect aerosol.

In the first stage of our work, we used MODIS and CALIPSO data at the same location using the collocation algorithm in [32]. With the MODIS data, we were able to predict dust, which was then compared against the results from CALIPSO. We were fortunate to have access to already

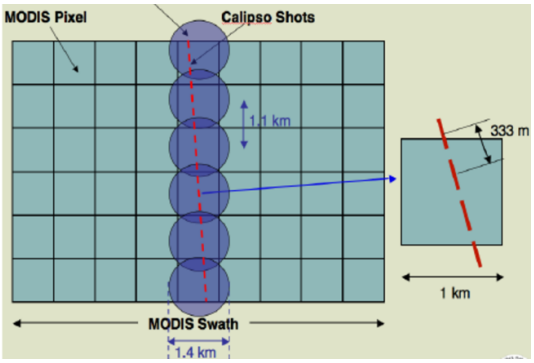

Fig. 1. Comparison of MODIS granule and CALIPSO track.

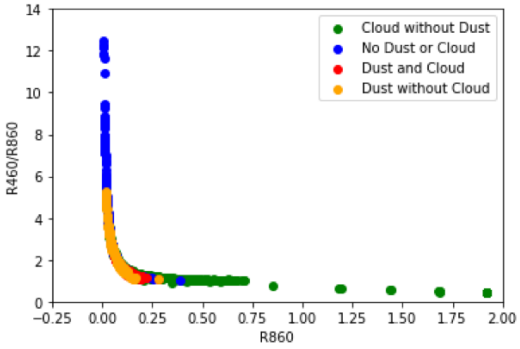

Fig. 2. The color ratio R_{460} nm/R_{860} nm as it depends on the reflectance at 860 nm, classified into the four cases: Cloud without Dust, No Dust No Cloud, Dust and Cloud, and Dust without Cloud.

collocated data for MODIS Level-2 and CALIPSO. This allowed us to determine the correct MODIS Level-1 files corresponding to the CALIPSO data. An important difference between the two data sets was the spatial resolution; CALIPSO has dust detection for every 5 km while the data utilized from MODIS was over 1 km. We decided to average over 5 pixels (each 1 km) for the MODIS data so that the data sets would correspond.

IV. Comparison of Physical Algorithms Methods and Machine Learning Methods for Dust Detection

A. Physical Algorithms

For this part of study, we use MODIS and CALIPSO collocated data to develop an algorithm for dust aerosol detection. In the collocated data, CALIPSO provides robust information of dust identification, MODIS provides radiances or emittance for up to 36 spectral bands. By using those pixels with both MODIS and CALIPSO observations and based on the knowledge of physical properties of mineral dust aerosols and previous studies on dust detection, we tried several methods to separate MODIS pixels with and without dust aerosols.

1) Color Ratio Algorithm over Ocean: Considering clear sky over ocean is much darker than dust and clouds, the reflectance at visible wavelengths for clear sky should be much smaller than the other two cases. Moreover, we know that dust aerosols are yellowish and clouds are usually white in color. Therefore, we expect that the color ratio defined as R_{460} nm/R_{860} nm may be different among clear, dusty and cloudy sky. To determine the ratios corresponding to each case, we plotted the color ratio as a function of its reflectance at 860 nm. As seen in Figure 2, strict classifications were not found. Thus, we were unable to proceed with the use of the color ratio in dust detection.

2) Reflectance and Emittance Ratio Algorithm over Ocean: Clouds are usually more reflective than yellowish dust aerosols and dark ocean at visible wavelengths. In contrast, in the thermal infrared such as $11\mu m$, ocean surface emits more than dust aerosols and clouds due to the higher temperature of ocean surface. Therefore, we investigated the relation among reflectance at 859nm, emittance at 11 μm and R_{859} nm/E_{11} μm, which is shown in Figure 3 and Figure 4. We can see that dust aerosols are not able to be separated from other cases by using R_{859} nm and E_{11} μm. Hence, we decided to investigate other methods for a physical algorithm.

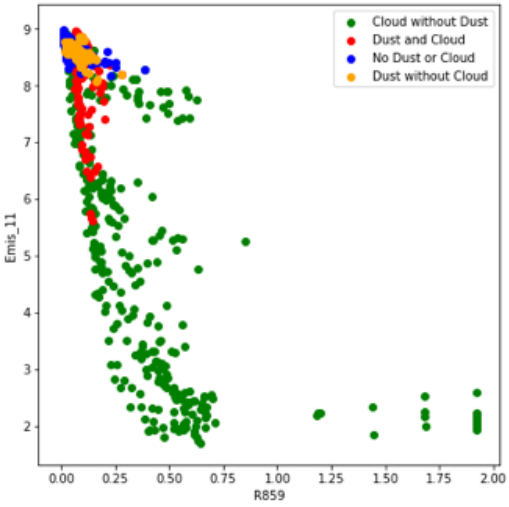

Fig. 3. The emissivity at 11 μm as a function of the reflectance at 859 nm, classified by the 4 different possible outcome.

3) Infrared Algorithm: Through observation and modeling studies, Ackerman [8] showed that brightness temperature difference (BTD), defined as the difference between the brightness temperature at 11 μm and 12 μm, of dust is smaller than that of clouds. In this algorithm, we first find a BTD threshold distinguishing between the dust and cloud cases. If BTD is smaller than the threshold, the pixel is classified as dust. In order to determine the BTD threshold, we first applied different thresholds for MODIS data along CALIPSO track and then calculated detection accuracy for different BTD threshold using CALIPSO dust detection as reference. We achieved the highest accuracy between 60% and 70% with the BTD threshold at 0.8. Using this threshold, we wrote an algorithm to detect dust aerosols over the entire MODIS granule.

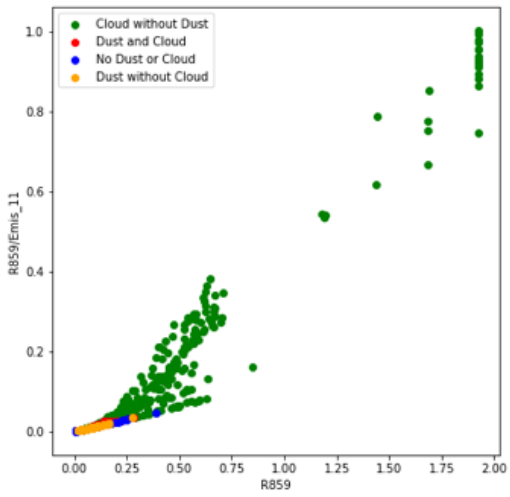

Fig. 4. The ratio of the reflectance at 859 nm to emissivity at 11 μm as a function of the reflectance at 859 nm, classified by the 4 different possible outcome.

B. Machine Learning Methods

Machine learning has been widely used in science and engineering fields, such as medical image analysis and it also has been proved to be very useful for remote sensing data including crop disease detection, new product creation etc [9]. The most commonly used data mining methods include artificial neural networks (ANN), support vector machines (SVM), decision trees, also some ensemble methods, such as random forests. We explored different machine learning methods for our dust detection in our study.

1) Logistic Regression: Logistic regression is one simple but powerful method, especially for binary outcome. One key component is the logistic function, which could convert the multivariate input into the probability of the outcome between 0 and 1. Among all the machine learning algorithms, logistic regression has multiple advantages. First, no assumption is needed such as normal distribution of independent variables; Second , no assumption is needed about linear relationship between outcome and covariates. Most importantly, it is easy to understand and interpret the results [10], [11]. In our logistic regression model, we used the glm package in R and SAS for variable selection.

2) Artificial Neural Network (ANN): There has been considerable applications of ANN in remote sensing data. The basic structure of the ANN includes input layer, output layer and some hidden layers. The input layer is composed of input variables, the output layer is the number of outcomes. The hidden layers could be one or multiple layers. With zero hidden layers, we can consider the neural network as one simple logistic regression model. Through controlling the number of hidden layers and number of nodes within each layer, ANN could be built for non-linear and complex relationships, which is important for dealing with real life problem. Like logistic regression, it also does not need any distribution assumption for the input variables, output variables. Another important advantage is that ANN could infer new relationships on unseen data, and thus make the model more generalized for new unknown data [10], [11].

3) Support Vector Machine (SVM): SVM is another popular machine learning algorithm based on statistical learning theory. The SVM algorithm is to find a decision boundary which could maximize the distance between the two closest classes. The biggest advantage for SVM is that it could model non-linear decision boundary.It has multiple kernel functions and it is pretty robust against over fitting. However one disadvantage to this algorithm is that SVM is very memory intensive and may not scale well to large datasets [12].

4) Random forests: Random forests are considered as one of the most accurate machine learning methods, which are an ensemble classifier and proved to be the top winner in several data competitions. Random forests consist of many decision trees and combine the result from the individual trees. The attractive benefits using random forests lie in the following facts: 1) random forests could handle thousands of input variables without variable selection, which is heavy burden for logistic regression; 2) through large number of decision trees within random forest, it could produce an unbiased estimate of the generalization error; 3) it may allow large portion of missing data [13].

5) Ensemble learning: The purpose of ensemble methods is trying to use multiple learning methods to achieve better predictive performance than single method [14]. There are different types of ensembles, in this paper, we used stacking ensemble learning. In stacking, several basic learning methods were applied to the datasets, and then another model could be built from the outputs from each individual models. It has been reported that stacked ensemble models could boost predictive accuracy. For this approach, we basically took logistic regression, ANN, SVM and random forests models as the base learner and logistic regression as meta learner. Our machine learning methods and final model are done through Weka 3.8.0.

C. Comparison of Physical Algorithms and Machine Learning Methods

1) Results from Infrared Physical Algorithm: Then we make use of the threshold to detect dust aerosols over the entire MODIS granule and compare with the RGB image to check how good our infrared dust detection algorithm is. We selected two dust storm cases over Atlantic ocean, the RGB images from MODIS observation of those two dust storms are shown in Figure 5. From the above RGB figures, we could easily tell white clouds and dust aerosols, which are yellowish.

Then we use 0.8 as BTD threshold to detect dust aerosols. If BTD (11-12 μm) of a MODIS pixel is smaller than the threshold, then the pixel is identified as dust-loading pixel. We apply this algorithm to the entire MODIS granule to detect dust aerosols. Figure 7 and Figure 8 show that the infrared BTD algorithm could detect dust aerosols to some extent, but still it may mistake clouds as dust aerosols.

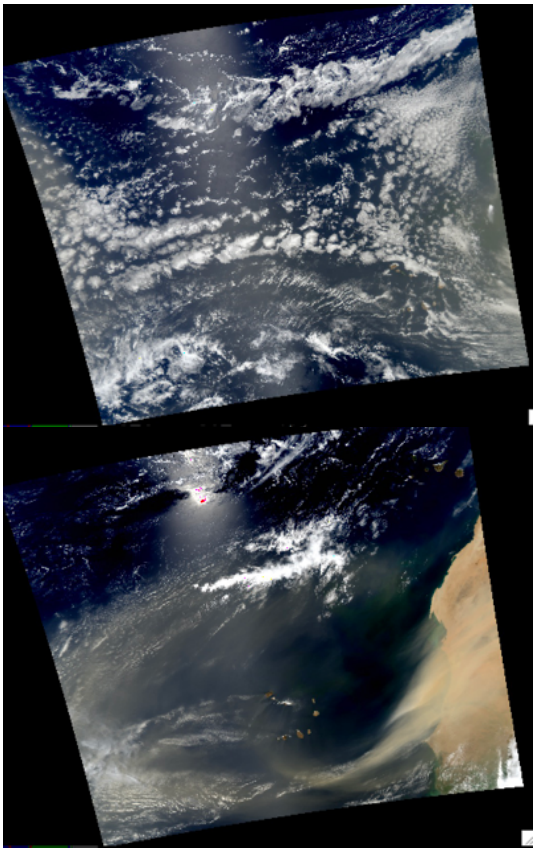

Fig. 5. RGB images of two dust storms from MODIS observations, the above is 06/22/2009, the below is 07/15/2007.

2) Results from Machine Learning Algorithms: To decide which machine learning method is better, we compared the performance among different approaches. Since some data mining approaches are very time consuming with large data, we used data on 07/15/2007 (3,335 data points with 1,510 dust and 1,825 non dust) to predict the data on 06/22/2009 (3,335 data points with 1,915 dust points and 1,410 non dust points) for our model selection analysis. The predictor variables include all 38 band values. We used 3 measure metrics to compare those model performance: The area under the curve (AUC), accuracy and Youden index.

AUC is one of the most important evaluation metrics to check model performance. It can tell how much the model is capable of distinguishing between classes. Higher AUC means better model. A poor model has AUC near to 0, and if model could predict a perfect outcome, AUC will be 1 [31].

Accuracy is one very intuitive measure for model performance. The accuracy is defined as the sum of correct prediction for dust and not dust divided by total data points. Youden's index is another popular model performance metric , which is simply calculated as $sensitivity + specificity - 1$ and can be used for the optimal cut-point. Sensitivity is defined as the true positive rate, which is the proportion of correctly predicted dust. Specificity is defined as the true negative rate, which is the proportion of correctly predicted not dust. We tried logistic regression (LR), Random forest (RF), SVM, ANN and one stacking classifier. In the stacking classifier, the base classifiers are the four individual classifiers (Random forests, Logistic regression, ANN and SVM), and the meta classifier is still logistic regression.

From the comparison result at Table II, we can see logistic regression model has the best accuracy, AUC and Youden index values compared to other machine learning methods, also, logistic regression needs little specification and is convenient for implementation. We decide to choose logistic regression as our further analysis.

TABLE I
PERFORMANCE COMPARISON AMONG DIFFERENT LEARNING METHODS: DUST DETECTION ALONG CALIPSO TRACK.

Method	Accuracy	AUC	Youden index
Random Forest	67.2%	0.765	0.436
Logistic regression	82.0%	0.864	0.654
ANN	69.8%	0.833	0.429
SVM	59.6%	0.648	0.376
Stacking classifiers (RF, LR, ANN, SVM)	63.7%	0.68	0.370

V. VARIABLE IMPORTANCE AND SELECTION

Overfitting is one common issue in regression analysis, which is normally caused by extraneous predictors in the model. When this occurs, the coefficients may have inflated magnitude, and then the R square will be large too. To reduce this effect, feature selection becomes very important [23]. Stepwise, backward or forward variable selection methods are traditional variable selection procedures, however the drawback is that they can not deal with large number of covariates very well, and may lead to highly bias in parameter estimates. Lasso selection is one alternative selection method, which was developed to overcome the limitation of traditional variable selection when the number of covariates is large [23], [24]. For Lasso variable selection, a widely used penalized measure of model fit is the Schwarz Bayes criterion (SBC). The formula is $SBC = -2 * LL + log(n) * K$, where LL is log likelihood of the logistic model, k is the degree of freedom in the model and n is the sample size. The smaller SBC, the better. Since $log(n) * K$ will increase with increasing number of variables, thus more variables in the model will be penalized and increase the SBC value [19]. Akaike information criteria (AIC) is also helpful for comparing models regarding the model fit and model complexity.

We used the *hpgenselect* package in SAS 9.4 for our variable selection [30]. The variable selection process can be found at Table I. After Lasso selection, the variables remain in the models are column 2 in Table I, and the other variables are dropped due to insignificant effect in the model.

Our selected variables are very consistent with Lee et al.'s finding in [32], where the authors claimed that their input variables are sixteen brightness temperatures (Band 20 - Band 36). Another paper [29] reported B20, B29, B30, B31, B32 as their important band variables.

TABLE II
VARIABLE SELECTION AND SELECTION STEPS USING LASSO.

Lasso Selection Details			
Step	Description	AIC	SBC
0	Initial model without covariates	23751	23758.8
1	Band 32 (11.770-12.270) μm)	23248	23271.2
2	Solar azimuth	22340.6	22363.9
3	Solar zenith	20573.4	20604.4
4	Band 36 (14.085-14.385) μm)	20124.8	20171.3
5	Band 31 (10.780-11.280) μm)	19824.3	19878.5
6	Band 24 (4.433-4.498) μm)	19653.7	19723.5
7	Sensor azimuth	19653.7	19723.5
8	Band 28 (7.175-7.475) μm)	19475.5	19560.7
9	Band 29 (8.400-8.700) μm)	19475.5	19560.7
10	Band 27 (6.535-6.895) μm)	19295.5	19396.2
11	Band 35 (13.785-14.085) μm)	19126.4	19227.1
12	Band 21 (3.929-3.989) μm)	18976.8	19085.2
13	Band 33 (13.185-13.485) μm)	18846.5	18962.8
14	Band 22 (3.939-3.989) μm)	18732.2	18856.2
15	Band 25 (4.482-4.549) μm)	18630.2	18761.9

VI. A HYBRID APPROACH FOR DUST DETECTION

Our hybrid algorithm is a combination of physical algorithms and logistic regression model (see Figure 6 for the flow chart of our hybrid algorithm).

1) Through Lasso automatic selection to identify the most important variables without human input. Only when variables meet the minimum requirement of SBC, the variables stay in the model.
2) Add variables from all physical algorithms to the model and do another round of Lasso selection.
3) Check the model parameter estimates and remove the non significant variables from the model and determine the final model.

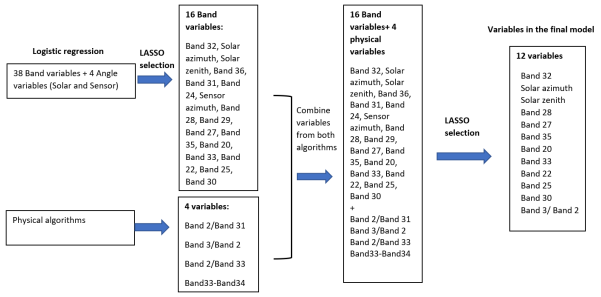

Fig. 6. Flowchart of variable selection procedure for our hybrid algorithm.

VII. EVALUATION ON OUR HYBRID ALGORITHM BASED DUST PREDICTION

We have two types of prediction tasks. One is using data in CALIPSO region data to predict the data in CALIPSO region. All those data have accurate labels: dust or not dust. Another type of prediction is to use CALIPSO region data as training data, and predict the data outside the CALIPSO area. For these prediction, since we don't have labels, and can only approximately validate the prediction accuracy through visually checking the predicted images against raw RGB images.

The data sets used in our experiments include the following dates: 1) 03/13/2007, 05/09/2007, 07/15/2007, 03/31/2008, 06/22/2009 and 04/22/2010 at northwest coast of Africa near Atlantic Ocean and 2) 11/05/2009 and 11/11/2009 at the coast of Arabian peninsula near Arabian Sea. Dust is defined as dust and clouds or dust without clouds, and not dust is defined as no dust and no clouds or clouds without dust, so the study outcome became one binary classification question. The total data points are 17,058, there are 8,248 dust points and 8,810 not dust points. We include all band wavelengths and 4 geometry variables in our analysis.

A. Results for Predicting CALIPSO Region Data

1) Parameter Estimate Comparison between our Hybrid Model and Pure Logistic Regression Model: The parameter estimates (shown in Table III and Table IV) from pure logistic regression and our hybrid model are pretty consistent regarding whether the variables have the positive or negative effect on the dust prediction, though the values are different. For example, band 20 has coefficient -0.58 compared to -0.86 in our hybrid model.

TABLE III
PARAMETER ESTIMATES FROM LOGISTIC REGRESSION MODEL

	Estimate	Std Error	P value
(Intercept)	-27.9684	5.639937	7.09E-07
Band 20 (3.660-3.840) μm	-0.57661	0.047896	2.22E-33
Band 22 (3.939-3.989) μm	0.949023	0.077796	3.15E-34
Band 25 (4.482-4.549) μm	1.557671	0.051672	1.22E-199
Band 27 (6.535-6.895) μm	0.376049	0.017946	1.73E-97
Band 28 (7.175-7.475) μm	-0.76633	0.025158	8.54E-204
Band 31 (10.780-11.280) μm	-4.19596	0.116641	2.20E-283
Band 32 (11.770-12.270) μm	2.858426	0.126925	2.61E-112
Band 33 (13.185-13.485) μm	2.183724	0.119588	1.71E-74
Band 35 (13.785-14.085) μm	-2.35714	0.087024	1.43E-161
Solar azimuth	-0.10573	0.005056	4.35E-97
Solar zenith	0.01164	0.001417	2.15E-16

TABLE IV
PARAMETER ESTIMATES FROM OUR HYBRID MODEL

	Estimate	Std Error	P value
(Intercept)	31.33543	6.21873	4.68E-07
Band 20 (3.660-3.840) μm	-0.85731	0.051076	3.15E-63
Band 22 (3.939-3.989) μm	1.149253	0.081038	1.19E-45
Band 25 (4.482-4.549) μm	1.047929	0.055049	8.55E-81
Band 27 (6.535-6.895) μm	0.477713	0.019291	2.23E-135
Band 28 (7.175-7.475) μm	-0.85452	0.026784	2.38E-223
Band 31 (10.780-11.280) μm	-2.96282	0.121345	1.14E-131
Band 32 (11.770-12.270) μm	1.853966	0.131862	6.70E-45
Band 33 (13.185-13.485) μm	2.275425	0.124293	7.29E-75
Band 35 (13.785-14.085) μm	-2.31015	0.090737	5.50E-143
Solar azimuth	-0.14228	0.005403	7.82E-153
Solar zenith	0.003036	0.001472	0.03915294
Band 3(459-479 nm)/ Band 2 (841-876 nm)	-0.68804	0.02672	3.24E-146

2) Model Performance Comparison among Physical models, Logistic Regression Model and our Hybrid Approach:

We applied 10-fold cross validation for our model with all data points to evaluate our model performance. In our analysis, the original sample is randomly partitioned into 10 equal size subgroups. Among the 10 groups, 9 groups are used for model development, and the remaining single group is used for testing the model. The cross validation process will be repeated 10 times so that each of the 10 subgroups will be used exactly once as the testing data. The averaged results are shown in Table VI.

Table V and Table VI showed the performance differences under different conditions, where we can find that the hybrid approach gave the best accuracy, AUC and Youden index values.

TABLE V
PERFORMANCE COMPARISON: USING DATA 07/15/2007 AS TRAINING, DATA 06/22/2009 AS TESTING

Algorithms	Accuracy	AUC	Youden index
Hybrid approach	0.839	0.907	0.682
Logistic regression	0.784	0.832	0.567
Physical algorithm: Infrared	0.701	0.741	0.406
Physical algorithm: Color ratio	0.414	0.424	NA
Physical algorithm: Reflectance ratio	0.423	0.5	NA

Note: Due to the poor performance, the Youden index is not available for physical algorithms: color ratio and reflectance ratio.

TABLE VI
PERFORMANCE COMPARISON: USING ALL DATA POINTS WITH 10-FOLD CROSS VALIDATION

Algorithms	Accuracy	AUC	Youden index
Hybrid approach	0.886	0.949	0.774
Logistic regression	0.847	0.933	0.695
Physical algorithm: Infrared	0.674	0.750	0.325
Physical algorithm: Color ratio	0.655	0.561	0.318
Physical algorithm: Reflectance ratio	0.633	0.453	0.322

B. Results for Predicting MODIS Region Data

The biggest challenge for dust detection for MODIS region (2,748,620 data points) is that we do not have any labels for MODIS region, which means we do not know whether the prediction is correct or not. We can only visually compare the RGB images to raw images. We tried to predict the dust of the whole MODIS region using our hybrid approach and produced the RGB images based on the predicted probabilities. The images produced by our hybrid approach (Figure 7 and Figure 8, below) look better than the ones produced by physical algorithm (infrared) compared to the raw images in Figure 5.

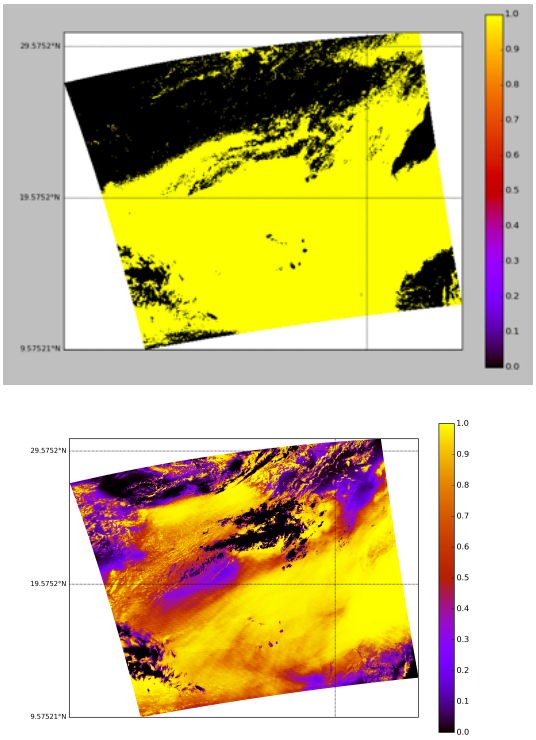

Fig. 7. Dust prediction for MODIS region using infrared physical algorithm (above) and our hybrid approach (below), the image date is 07/15/2007.

VIII. CONCLUSIONS

In our study, we tried both physical algorithms and several data mining approaches for dust detection. Our results showed that pure machine learning methods could significantly improve the prediction accuracy compared to pure physical algorithm (around 85% vs 67% for all data prediction, and 78% vs 70% for different day prediction see Tables V and VI), which could greatly enhance our capability for future dust detection. Meanwhile we also tried to combine physical algorithms with machine learning approach and the combined approach provided even better results (89% and 84% , see Tables V and VI).

For future work, we plan to investigate the relationship between the variables from data mining approach and variables from the physical algorithm for further variable selection and composite variable creation. We would also like to expand our research to land dust detection, which requires slightly different methods for analysis and increase our data points from the coast off North Africa to the whole world and include multiple time periods.

ACKNOWLEDGMENT

This work is supported by the grant CyberTraining: DSE: Cross-Training of Researchers in Computing, Applied Mathematics and Atmospheric Sciences using Advanced Cyberin-

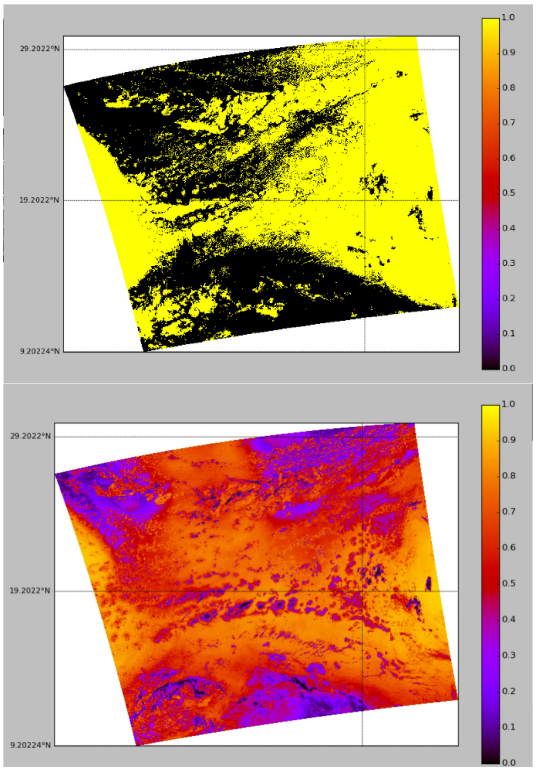

Fig. 8. Dust prediction for MODIS region using infrared physical algorithm (above) and our hybrid approach (below), the image date is 06/22/2009.

frastructure Resources from the National Science Foundation (grant no. OAC1730250)

REFERENCES

[1] Madhavan, Sriharsha, John J. Qu, and X. Hao. "Saharan dust detection using multi-sensor satellite measurements." Heliyon 3.2 (2017): e00241.

[2] Lee, Sang-Sam, and Byung-Ju Sohn. "Dust detection and quantification from MODIS IR bands using Artificial Neural Network (ANN) model." 13th Conference on Aviation, Range and Aerospace Meteorology. 2008.

[3] Bursac, Zoran, et al. "Purposeful selection of variables in logistic regression." Source code for biology and medicine 3.1 (2008): 17.

[4] Cho, Hyoun-Myoung, et al. "Detection of optically thin mineral dust aerosol layers over the ocean using MODIS." Journal of Atmospheric and Oceanic Technology 30.5 (2013): 896-916.

[5] Kaufman, Yoram J., Arnon Karnieli, and Didier Tanr. "Detection of dust over deserts using satellite data in the solar wavelengths." IEEE Transactions on Geoscience and Remote Sensing 38.1 (2000): 525-531.

[6] Legrand, Michel, Michel Desbois, and Kwami Vovor. "Satellite detection of Saharan dust: Optimized imaging during nighttime." Journal of climate 1.3 (1988): 256-264.

[7] Prospero, Joseph M. "Long-range transport of mineral dust in the global atmosphere: Impact of African dust on the environment of the southeastern United States." Proceedings of the National Academy of Sciences 96.7 (1999): 3396-3403.

[8] Zhao, Tom X-P., Steve Ackerman, and Wei Guo. "Dust and smoke detection for multi-channel imagers." Remote Sensing 2.10 (2010): 2347-2368.

[9] Lary, David J., et al. "Machine learning in geosciences and remote sensing." Geoscience Frontiers 7.1 (2016): 3-10.

[10] Tu, Jack V. "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes." Journal of clinical epidemiology 49.11 (1996): 1225-1231.

[11] Dreiseitl, Stephan, and Lucila Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review." Journal of biomedical informatics 35.5-6 (2002): 352-359.

[12] Mountrakis, Giorgos, Jungho Im, and Caesar Ogole. "Support vector machines in remote sensing: A review." ISPRS Journal of Photogrammetry and Remote Sensing 66.3 (2011): 247-259.

[13] Pal, Mahesh. "Random forest classifier for remote sensing classification." International Journal of Remote Sensing 26.1 (2005): 217-222.

[14] Zhang, Cha, and Yunqian Ma, eds. Ensemble machine learning: methods and applications. Springer Science Business Media, 2012.

[15] Ackerman, Steven A. "Remote sensing aerosols using satellite infrared observations." Journal of Geophysical Research: Atmospheres 102.D14 (1997): 17069-17079.

[16] Li, Xian, and Weidong Song. "Dust storm detection based on Modis Data." Proceedings of the International Conference on Geo-spatial Solutions for Emergency Management and the 50th Anniversary of the Chinese Academy of Surveying and Mapping, Beijing, China. 2009.

[17] Jose, Subin, et al. "Satellitebased shortwave aerosol radiative forcing of dust storm over the Arabian Sea." Atmospheric Science Letters 17.1 (2016): 43-50.

[18] Ciren, Pubu, and Shobha Kondragunta. "Dust aerosol index (DAI) algorithm for MODIS." Journal of Geophysical Research: Atmospheres 119.8 (2014): 4770-4792.

[19] Cohen, Robert A. "Introducing the GLMSELECT procedure for model selection." Proceedings of the Thirty-First Annual SAS Users Group International Conference. 2006. Atmospheres 119.8 (2014): 4770-4792.

[20] Ciren, Pubu, and Shobha Kondragunta. "Dust aerosol index (DAI) algorithm for MODIS." Journal of Geophysical Research: Atmospheres 119.8 (2014): 4770-4792.

[21] Mazzoni, Dominic, et al. "An operational MISR pixel classifier using support vector machines." Remote Sensing of Environment 107.1-2 (2007): 149-158.

[22] Unal, Ilker. "Defining an optimal cut-point value in roc analysis: an alternative approach." Computational and mathematical methods in medicine 2017 (2017).

[23] McNeish, Daniel M. "Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences." Multivariate Behavioral Research 50.5 (2015): 471-484.

[24] Lu, Miao, et al. "Application of penalized linear regression methods to the selection of environmental enteropathy biomarkers." Biomarker research 5.1 (2017): 9.

[25] Albugami, Sarah, et al. "Evaluating MODIS dust-detection indices over the Arabian Peninsula." Remote Sensing 10.12 (2018): 1993.

[26] Souri, Amir Hossein, and Sanaz Vajedian. "Dust storm detection using random forests and physical-based approaches over the Middle East." Journal of Earth System Science 124.5 (2015): 1127-1141.

[27] Sorek-Hamer, M., et al. "Classification of dust days by satellite remotely sensed aerosol products." International journal of remote sensing 34.8 (2013): 2672-2688.

[28] Jafari, Reza, and Mansoureh Malekian. "Comparison and evaluation of dust detection algorithms using MODIS Aqua/Terra Level 1B data and MODIS/OMI dust products in the Middle East." International Journal of Remote Sensing 36.2 (2015): 597-617.

[29] Shahrisvand, M., and M. Akhoondzadeh. "A Comparison Of Empirical And Inteligent Methods For Dust Detection Using Modis Satellite Data." International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 1 (2013): W3.

[30] Lund, Bruce, and Magnify Analytic Solutions. "Logistic Model Selection with SAS PROCs LOGISTIC, HPLOGISTIC, HPGENSELECT."

[31] Japkowicz, and Stan Szpakowicz. "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." Australasian joint conference on artificial intelligence. Springer, Berlin, Heidelberg, 2006.

[32] Wang, Likun, et al. "Fast and accurate collocation of the visible infrared imaging radiometer suite measurements with cross-track infrared sounder." Remote Sensing 8.1 (2016): 76.

[33] Li, Xian, and Weidong Song. "Dust storm detection based on Modis Data." Proceedings of the International Conference on Geo-spatial Solutions for Emergency Management and the 50th Anniversary of the Chinese Academy of Surveying and Mapping, Beijing, China. 2009.