Mineral Dust Detection Using Satellite Data

CyberTraining: Big Data + High-Performance Computing + Atmospheric Sciences

Peichang Shi[1], Qianqian Song[2], & Janita Patwardhan[3],
Faculty Advisors: Dr. Zhibo Zhang[2] and Dr. Jianwu Wang[1]

[1]Department of Information Systems, University of Maryland, Baltimore County
[2]Department of Physics, University of Maryland, Baltimore County
[3]Department of Mathematics and Statistics, University of Maryland, Baltimore County

Technical Report HPCF-2018-13, hpcf.umbc.edu > Publications

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. We investigated both physical and machine learning algorithms of dust aerosols detection over the Atlantic Ocean using satellite observations from Moderate Resolution Imaging Spectroradiometer (MODIS) and the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation Observation (CALIPSO). We found that the machine learning algorithms achieved a higher accuracy rate compared to those of the physical algorithms. Through combining a logistic regression algorithm with our physical understanding of dust aerosols, we were able to reach the highest detection accuracy.

1 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 [12]. 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 [3, 5, 6, 8, 15].

Unfortunately, many of these algorithms have a lower detection rate or accuracy rate than desired. As we have access to a large collection of satellite data, we combined big data and machine learning techniques with a physics background to develop an algorithm with around 90% accuracy rate.

In Section 2, we discuss the data sets utilized in our study, delving a little deeper into the differences between MODIS and CALIPSO data. Sections 3 focuses on the algorithms investigated, from a few simple physical algorithms to the combined machine learning and physics algorithm. The results of our tested algorithms are outlined in Section 4, along with a comparison table. Lastly, we conclude in Section 5 with some ideas for future work.

2 Satellite Data

MODIS is a passive sensor onboard the Terra satellite since 1999 and the Aqua satellite since 2002. 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 is the least processed. The information is stored in HDF files, with a data point for every 5-minute measurement, called a granule. 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. As an active sensor, it measures the reflection, refraction, and scattering of its own transmitted signals by the Earth's surface and atmosphere. Through this use of depolarization, it is able to better detect clouds and dust aerosols. However, it requires more energy than a passive sensor and as seen in Figure 2.1, it covers much less area than MODIS, which is why we would like to use MODIS data to detect aerosol.

In the first stages of our work, we used MODIS and CALIPSO collocated data. 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 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.

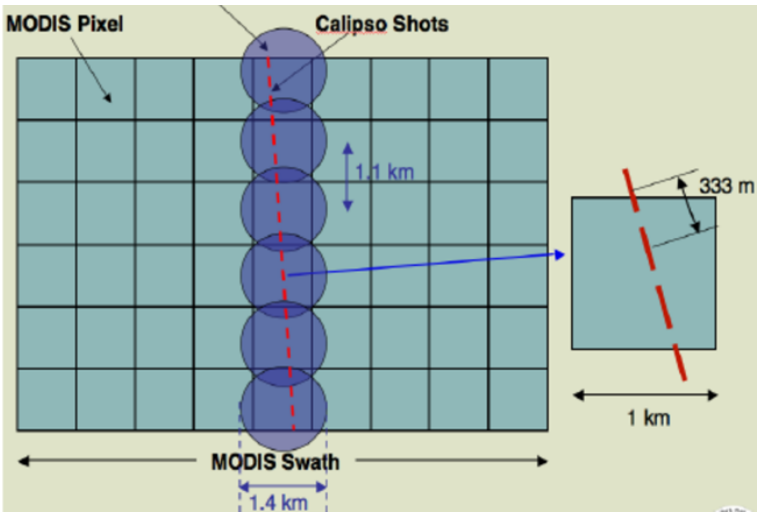

Figure 2.1: Comparison of MODIS granule and CALIPSO track

3 Methods

3.1 Physical Algorithm

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.

3.1.1 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 3.1, strict classifications were not found. Thus, we were unable to proceed with the use of the color ratio in dust detection.

3.1.2 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}

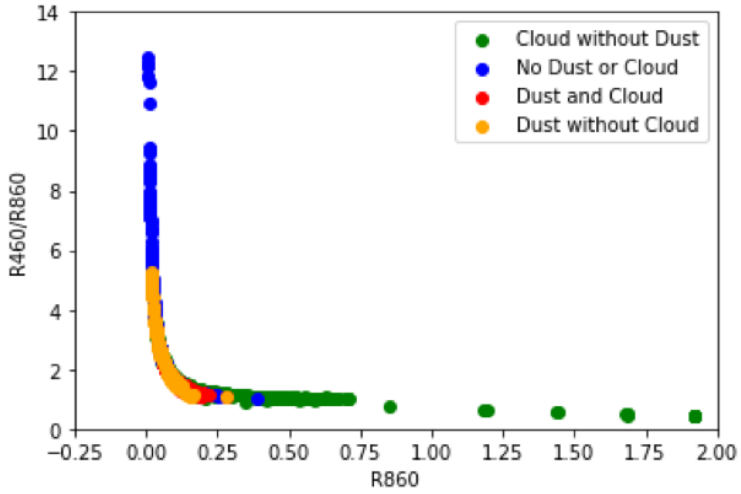

Figure 3.1: The color ratio R_{460} nm/R_{860} nm as it depends on the reflectance at 860 nm, classified into the four cases: Cloud without Dust, No Dust of Cloud, Dust and Cloud, and Dust without Cloud.

μm, which is shown in Figure 3.2. 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.

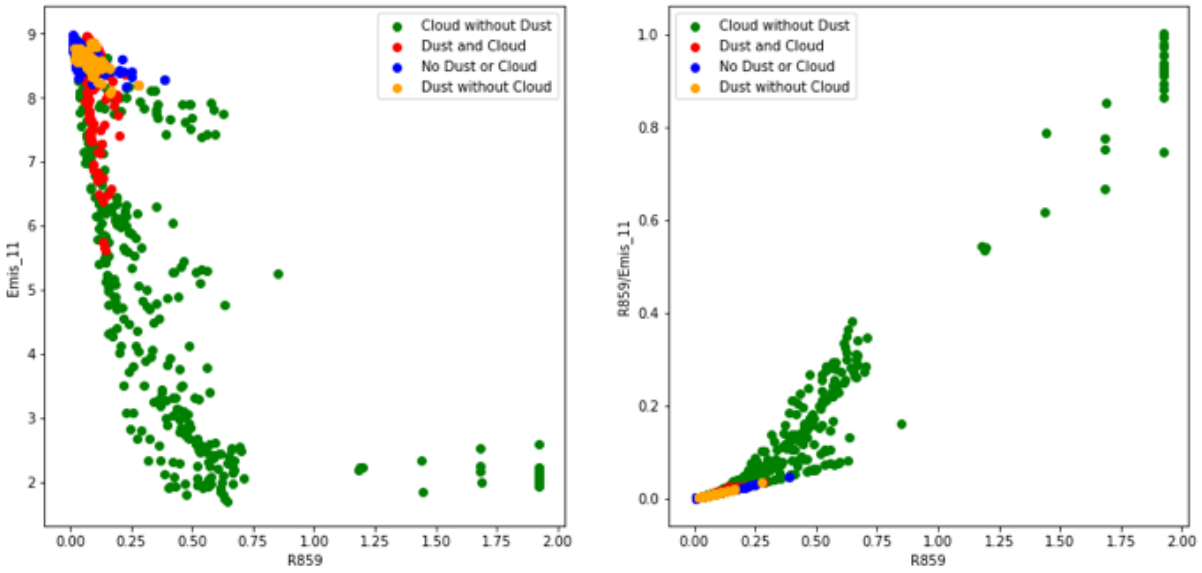

Figure 3.2: The emissivity at 11 μm (left) and the ratio of the reflectance at 859 nm to emissivity at 11 μm (right) as a function of the reflectance at 859 nm, classified by the 4 different possible outcome.

3.1.3 Infrared Algorithm

Through observation and modeling studies, Ackerman [1] 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.

3.2 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 [7]. 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 trees. For this study, we have explored different machine learning methods for our dust detection.

3.2.1 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 multi variate input into the probability of the outcome between 0 and 1. Among all the machine learning algorithms, logistic regression has multiple advantages. Firstly, no assumption is made regarding the dependent variables following a normal distribution. There is also no assumption about a linear relationship between outcome and covariates. Most importantly however, it is easy to understand and interpret the results [4, 13] . In our logistic regression model, we used the glm in R with stepwise selection function for variable optimization.

3.2.2 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 layer. The input layer is composed of input variables, the output layer is the number of outcomes. The hidden layers could be 1 or multiple layers. With 0 hidden layers, we can consider the neural network as one simple logistic regression model. There are multiple advantages of ANN. 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 variable. Another important advantage for ANN is that ANN

could infer new relationships on unseen data, and thus make the model more generalized for new unknown data [4, 13]. Neuralnet package in R was used for our ANN analysis. In the formula, we used 1 hidden layer with 5 nodes. The input layer includes all variables. We also tried 2 5 layers with varied nodes (5-100) using SparkR, strangely which took long time to converge with our high performance facility and without significant improvement. We finally decided to use R as our final analysis tool.

3.2.3 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 advantages for SVM is that it could model non-linear decision boundary; it has multiple kernel functions and it is pretty robust against over fitting [9, 11]. However one drawback to this algorithm is that SVM is very memory intensive and may not scale well to large datasets. SVM was run using R package "e1071" with a similar formula to logistic regression.

3.2.4 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 section, 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 [2, 10]. Random forest in R is pretty straightforward. In its setting, we took 2000 as the number of trees in the forest, and also set the importance to True. The node size was 10.

3.2.5 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 tried stacking ensemble learning. In stacking, several basic learning methods are applied to the datasets,and then another model could be build 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 the average of the predicted probability from previous logistic regression, ANN and random forests models.

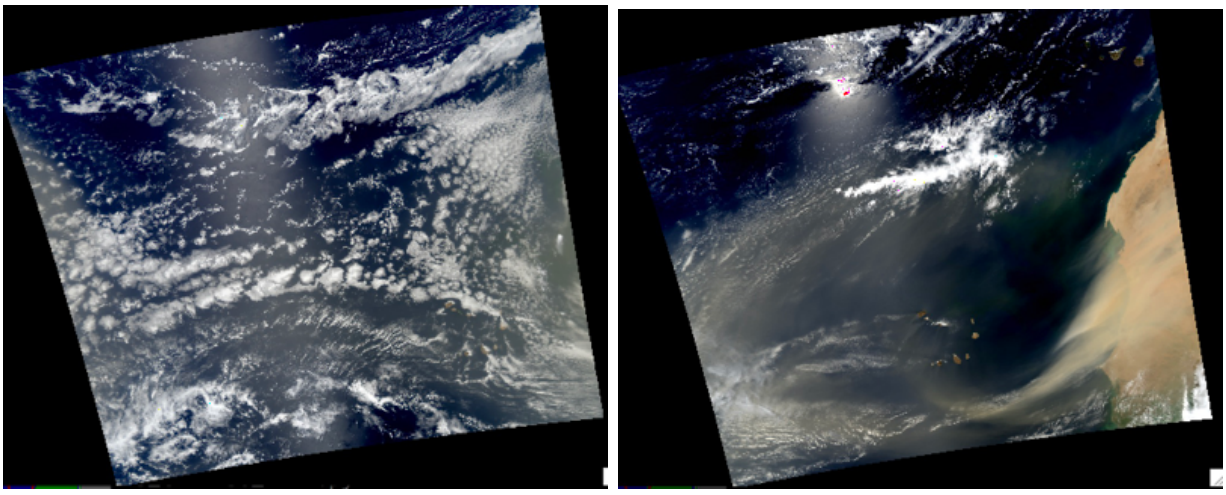

Figure 4.1: RGB images of two dust storms from MODIS observations

4 Results

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 label- dust or not. 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 validate the prediction accuracy through visually checking the predicted image against raw image.

4.1 Infrared Algorithm Results

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 3.3. 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 (11-12μm) 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. Below we show our dust detection results. Comparison between Figure 3.3 and Figure 3.4 shows that the infrared BTD algorithm could detect dust aerosols to some extent, but still it may mistake clouds as dust aerosols.

4.2 Results for predicting CALIPSO region data

To decide which machine learning method is better, we need to compare the performance among those approaches. We used data on July 15, 2007(3335 data points with 1510 dust and 1825 non dust) to predict the data on June 22, 2009 (3335 data points with 1915 dust points and 1410 non dust points). The predictor variables include all 38 band values.

The comparison metric is the accuracy rate, which is a simple ratio between the correctly predicted dust and the total dust. We tried logistic regression(LR), Random forest (RF), SVM , ANN and one stacking classifiers. The stacking classifier is basically the average of the probabilities from the 4 individual classifiers(Random forests, Logistic regression, ANN and SVM). From the comparison table 4.1,we can see logistic regression model could achieve the best accuracy 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 final model.

Table 4.1: Performance comparison among different learning methods:dust detection along CALIPSO track

Method	Accuracy
Random Forest	79.8%
Logistic regression	83.9%
ANN	64.7%
SVM	65.8%
Stacking classifiers(RF, LR, ANN,SVM)	75.6%

4.3 Variable selection for logistic regression and combination of physical algorithm and machine learning approach

The original data set has 38 variables,due to high multicollinearity among some variables, for example, variable band 30 and band 29 have correlation coefficient ¿0.9, We need feature engineering to identify the most important variables. We used step-wise function in logistic regression to end up with 16 variables out of 38 band variables (Band30, Band32, Band29, Band35, Band20, Band6, Band16, Band9, Band14, Band4, Band3, Band12, Band22, Band31, Band28, Band26, Band27, Band21, Band1). The variables based on physical algorithms are the division of band 2 value by band 32 or band 33 values, and also the difference between band 32 and band 33. To combine the physical component and machine learning approach, we simply add the 4 physical variables to the 16 variables selected by machine learning approach. In table 4.2, we showed the performance differences under different conditions, the model with combination of machine learning and physical algorithm gives the best accuracy result.

4.4 Results for predicting MODIS region data

The biggest challenge for dust detection for MODIS region(with 2748620 data points) is that we don't have any labels for MODIS region, which means we don't know whether the prediction is correct or not. We can only visually compare the RGB image to raw image. We applied our model based on machine learning and physical algorithm to predict the dust of the whole MODIS region. The image produced by our combined approach(figure 4.2,

Table 4.2: Performance comparison using different number of variables: dust detection along CALIPSO track

Models	Accuracy
July 15,2007 data: 70% for training, 30% testing	
Physical algorithm	0.554
All band variables	0.924
Selected 16 band variables based on machine learning	0.929
Selected 16 band variables + 4 variables based on physical algorithm	0.931
Selected 16 band variables + 4 sensor angle variables	0.925
July 15,2007 data for training , June 22,2009 data for testing	
Physical algorithm	0.423
All band variables	0.832
Selected 16 variables based on machine learning	0.820
Selected 16 variables + 4 variables based on physical algorithm	0.835
Selected 16 band variables + 4 sensor angle variables	0.809

right) looks better than the one produced by physical algorithm compared to the raw image in figure 4.1, right.

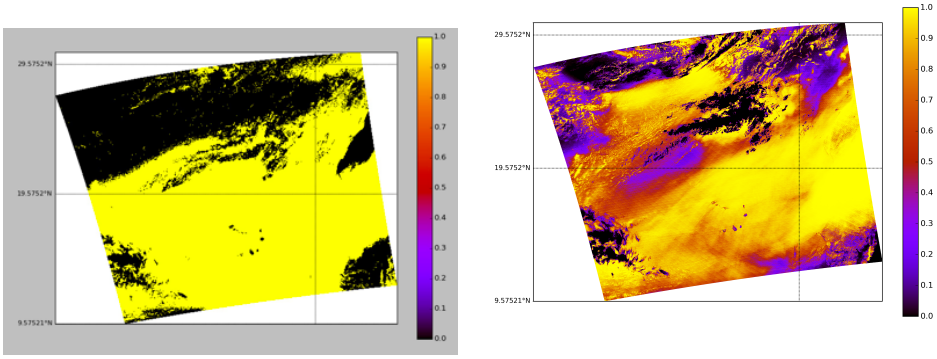

Figure 4.2: Dust prediction for MODIS region using infrared physical algorithm(left) and machine learning approach(right)

5 Conclusions

In our study, we tried both physical algorithms and several data mining approaches for dust detection. Our results showed that machine learning methods could significantly improve the prediction accuracy compared to pure physical algorithm (around 55% vs 93% for same day prediction, and 42% vs 80% for different day prediction), which could greatly enhance our ability to predict future dust detection. Meanwhile we also tried to combine physical

algorithms with machine learning approach, and due to time limit, we simply put the variables in the machine learning approach and variables from physical algorithm together. The combined approach provides even better results. Next step, we may need 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. In future study of dust detection, we would like to expand our research to land dust detection, which requires slightly different methods to analyze. Also, we would like to increase our data points from the coast off North Africa to the whole world and include multiple time periods. Given the increase of the data size, we also need to think about how to efficiently used our high performance facilities at UMBC to handle the big data. We already learned from our exercise that HPCF at UMBC could reduce the running time from around 30 hours to around 30 minutes based on the available computing clusters. Hopefully in the future we can build a real-time dust detection application using the global data and our HPCF.

Acknowledgment

Team 3 members gratefully acknowledge the NSF-funded CyberTraining program and all instructors for providing this chance for us to learn more about parallel computing. The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF). The facility is supported by the U.S. National Science Foundation through the MRI program (CNS–0821258, CNS–1228778, and OAC–1726023) and the SCREMS program (DMS–0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See hpcf.umbc.edu for more information on HPCF and the projects using its resources.

References

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

[2] Mariana Belgiu and Lucian Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016.

[3] Hyoun-Myoung Cho, Shaima L Nasiri, Ping Yang, Istvan Laszlo, and Xuepeng Tom Zhao. Detection of optically thin mineral dust aerosol layers over the ocean using modis. *Journal of Atmospheric and Oceanic Technology*, 30(5):896–916, 2013.

[4] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 2002.

[5] Xianjun Hao and John Jianhe Qu. Saharan dust storm detection using moderate resolution imaging spectroradiometer thermal infrared bands. *Journal of Applied Remote Sensing*, 1(1):013510, 2007.

[6] Yoram J Kaufman, Arnon Karnieli, and Didier Tanré. Detection of dust over deserts using satellite data in the solar wavelengths. *IEEE Transactions on Geoscience and Remote Sensing*, 38(1):525–531, 2000.

[7] David J Lary, Amir H Alavi, Amir H Gandomi, and Annette L Walker. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10, 2016.

[8] Michel Legrand, Michel Desbois, and Kwami Vovor. Satellite detection of saharan dust: Optimized imaging during nighttime. *Journal of climate*, 1(3):256–264, 1988.

[9] Giorgos Mountrakis, Jungho Im, and Caesar Ogole. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3):247–259, 2011.

[10] Mahesh Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005.

[11] Mahesh Pal and PM Mather. Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, 26(5):1007–1011, 2005.

[12] Joseph M Prospero. 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):3396–3403, 1999.

[13] Jack V Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231, 1996.

[14] Cha Zhang and Yunqian Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

[15] Tom X-P Zhao, Steve Ackerman, and Wei Guo. Dust and smoke detection for multi-channel imagers. *Remote Sensing*, 2(10):2347–2368, 2010.