

# Machine Learning and Bias Correction of MODIS Aerosol Optical Depth

D. J. Lary, L. A. Remer, D. MacNeill, B. Roscoe, and S. Paradise

**Abstract**—Machine-learning approaches (neural networks and support vector machines) are used to explore the reasons for a persistent bias between aerosol optical depth (AOD) retrieved from the MODerate resolution Imaging Spectroradiometer (MODIS) and the accurate ground-based Aerosol Robotic Network. While this bias falls within the expected uncertainty of the MODIS algorithms, there is room for algorithm improvement. The results of the machine-learning approaches suggest a link between the MODIS AOD biases and surface type. MODIS-derived AOD may be showing dependence on the surface type either because of the link between surface type and surface reflectance or because of the covariance between aerosol properties and surface type.

**Index Terms**—Aerosol optical depth (AOD), machine learning, neural networks, support vector machines (SVMs).

## I. INTRODUCTION

AEROSOL and cloud radiative effects remain the largest uncertainties in our understanding of climate change [1]. Over the past decade, observations and retrievals of aerosol characteristics have been conducted from space-based sensors, from airborne instruments, and from ground-based samplers and radiometers. Much effort has been directed at these data sets to collocate observations and retrievals and to compare results. Ideally, when two instruments measure the same aerosol characteristic at the same time, the results should agree within well-understood measurement uncertainties. When interinstrument biases exist, we would like to explain them theoretically from first principles. One example of this task is the comparison between the aerosol optical depth (AOD) retrieved by the MODerate resolution Imaging Spectroradiometer (MODIS) and the AOD measured by the Aerosol Robotic Network (AERONET). While progress has been made in understanding the biases between these two data sets, we still have an imperfect understanding of the root causes. Thus, in this letter, we

examine the efficacy of empirical machine-learning algorithms for bias correction.

## II. PREVIOUS STUDIES

The MODIS instruments are aboard both the Aqua and Terra satellites, launched on May 4, 2002 and December 18, 1999, respectively. The MODIS instruments collect data over the entire globe in two days. The AOD is retrieved using dark target methods in bands at 550, 670, 870, 1240, 1630, and 2130 nm, over ocean, and at 470, 550, and 670 nm over land [2], [3]. Other wavelengths are also used in the retrieval, for instance, short-wave infrared wavelengths for the land algorithm. Previous MODIS aerosol validation studies have compared the Aqua and Terra MODIS-retrieved AOD with the ground-based AERONET observations [2]. AERONET is a global system of ground-based sun and sky scanning sun photometers that measure AOD in various channels, depending on individual instrument, but usually include measurements at 340, 380, 440, 500, 675, 870, and 1020 nm [4]. Measurements are taken every 15 min during daylight hours. AERONET Level 2 quality assured AOD observations are accurate to within 0.01 for wavelengths of 440 nm and higher.

These previous studies concluded that MODIS AOD agreed with AERONET observations to within MODIS expected uncertainties, on a global basis. AERONET is only available for land locations, although some sites are in coastal regions.

However, the correlation for the MODIS ocean algorithm was much better than the agreement for the MODIS land algorithm, in the Collection 4 data set. Revision and implementation of a new land algorithm and reprocessing of the data resulted in much improvement to the retrieved MODIS AOD over land [3]. Even so, there remains a small overprediction of the AOD for low values and underprediction at high AOD values [3], [5].

In previous studies, we intercompared the normalized difference vegetation indices (NDVIs) from different sensors [6]. We have found that machine-learning algorithms are able to effectively perform interinstrument cross-calibration. Here, we extend this approach to consider AOD. In our previous intercomparison of NDVIs, we found that the surface type played a key role in explaining a significant fraction of the interinstrument differences. In this letter, we wanted to investigate if the same was true for AOD.

Xiao *et al.* [7] have examined the difference between AODs retrieved from the Multiangle Imaging Spectro-Radiometer (MISR) and MODIS over mainland Southeast Asia. They found that, although the difference between MISR and MODIS should be small and randomly distributed over space, the difference

Manuscript received February 18, 2009; revised March 16, 2009 and April 22, 2009. First published July 7, 2009; current version published October 14, 2009. This work was supported in part by the National Aeronautics and Space Administration (NASA) through the Awards NNG06GB78G, NNX06AG04G, NNX06AF29G, and NNX07AD49G and in part by the NASA Goddard Space Flight Center student DEVELOP Program.

D. J. Lary is with the Joint Center for Earth Systems Technology, University of Maryland, Baltimore County, MD 21228 USA, and also with the Software Integration and Visualization Office, NASA Goddard Space Flight Center, Greenbelt, MD 20771 USA.

L. A. Remer is with the NASA Goddard Space Flight Center, Greenbelt, MD 20771 USA.

D. MacNeill and B. Roscoe are with the NASA Goddard Space Flight Center DEVELOP Program, Greenbelt, MD 20771 USA.

S. Paradise is with the NASA Jet Propulsion Laboratory, Pasadena, CA 91109 USA.

Digital Object Identifier 10.1109/LGRS.2009.2023605

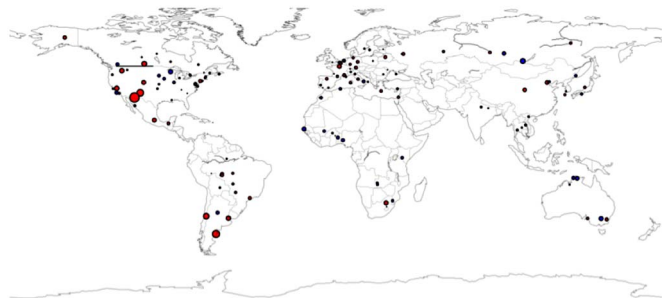


Fig. 1. MODIS bias with respect to AERONET [8]. Computed as a regression with intercept at the origin. Red indicates that MODIS is higher; blue indicates that AERONET is higher. The size of the circle is proportional to the slope of the regression for slope  $> 1$  (where MODIS is higher) and to the inverse of the slope for slope  $< 1$ .

actually has a strong negative relationship with MODIS AODs and tends to be spatially clustered. They concluded that further research is needed to fully understand the spatial dependence in these differences. The machine-learning approach outlined here is also relevant to the MISR comparison in [7].

### III. DATA DESCRIPTION

We use the global 10-km MODIS Collection 5 AOD product, over land and ocean, and all the available AERONET version 2.0 data. The AERONET program provides a long-term, continuous, and readily accessible public domain database of aerosol optical properties. The network imposes standardization of instruments, calibration, processing, and distribution. The location of individual sites is available from the AERONET web site <http://aeronet.gsfc.nasa.gov/>.

We first identify all MODIS overpasses of the AERONET sites throughout the lifetime of the two MODIS missions. We use the single green band MODIS AOD (550 nm) in the geographic grid point that contains the AERONET site. AERONET AOD measurements within 30 min of the MODIS observation are averaged. AERONET data are interpolated (in log-log space) to the green band where they are missing. We found a strong correlation between geographic location and bias. For example, there is a negative bias (MODIS underestimation relative to AERONET) over vegetated Western Africa (from Liberia to Nigeria) and a positive bias over the Southwestern U.S. The spatial dependence of the differences between AERONET and MODIS is shown in Fig. 1 [8].

### IV. AOD INTERCOMPARISON

Fig. 2(a) and (b) shows the scatter diagram comparisons of AOD from AERONET ( $x$ -axis) and MODIS ( $y$ -axis) as green circles overlaid with the ideal case of perfect agreement (blue line). The left-hand column of plots is for MODIS Aqua, and the right-hand column of plots is for MODIS Terra. These comparisons between AERONET and MODIS are for the entire period of overlap between the MODIS and AERONET instruments from the launch of the MODIS instrument to the present and include all possible collocations from all AERONET stations. We note that MODIS has a high bias relative to

AERONET (the slope is not one), there is substantial scatter, and there are correlation coefficients of 0.86 and 0.84 for MODIS Aqua and MODIS Terra, respectively. The bias and scatter indicate that the agreement between AERONET and MODIS may be dependent on some factors not completely accounted for in the retrievals. Note that the plots include both land and ocean retrievals.

In an exploratory data analysis study, we examined whether this bias could be explained by a variety of factors, including surface type, soil type, cultivation type, cloud reflectivity, and total ozone column, to name just a few. In other words, we constructed a comprehensive set of as many variables as possible and determined which of these variables was correlated with the AOD bias between AERONET and MODIS. It was found that the surface type could explain much of the difference between MODIS and AERONET. The surface classification we used was the global landcover classification for the year 2000 (GLC2000) at a resolution of  $(1/8)^\circ \times (1/8)^\circ$  (<http://www-gem.jrc.it/glc2000/>). Before using the surface classification in our machine-learning bias correction (described hereinafter), we reordered the surface types such that their annual mean area weighted albedos are in ascending order. The reordering was done as follows: When we use the surface type as an input for the machine-learning algorithms, it is, in effect, being treated as a quasi-continuous variable. As the surface reflectivity is one of the most important properties of each surface type for this problem, we want a surface-type classification which is monotonic in surface reflectivity.

When we augmented the surface type with variables available within the MODIS AOD HDF files (MOD04 and MYD04), we found that the machine-learning algorithms were able to further improve their bias correction. In the results presented in Fig. 2, the variables we used in explaining the AOD bias between MODIS and AERONET were the surface type, the solar zenith angle, the solar azimuth angle, the sensor zenith angle, the sensor azimuth angle, the scattering angle, and the reflectance at 550 nm.

#### A. Machine Learning

Machine learning is a subfield of artificial intelligence that is concerned with the design and development of algorithms that allow computers to empirically learn the behavior of data sets. A major focus of machine-learning research is to automatically produce (induce) models from data. In this letter, we have applied two types of machine learning to the correction of the bias between MODIS and AERONET, i.e., neural networks and support vector machines (SVMs).

For each of these machine-learning approaches, we used two training data sets, i.e., one for MODIS Aqua and one for MODIS Terra. These training data sets include all contemporaneous measurements of the MODIS instruments and AERONET made from launch to the present that were within 30 min of each other, within a great circle distance of  $0.25^\circ$ , and within a solar zenith angle of  $0.1^\circ$ . For MODIS Aqua, this gave us a training record of 7543 points, and for Terra, 13 034 points.

The purpose of training a machine-learning algorithm is to construct a mapping between a set of input variables and an

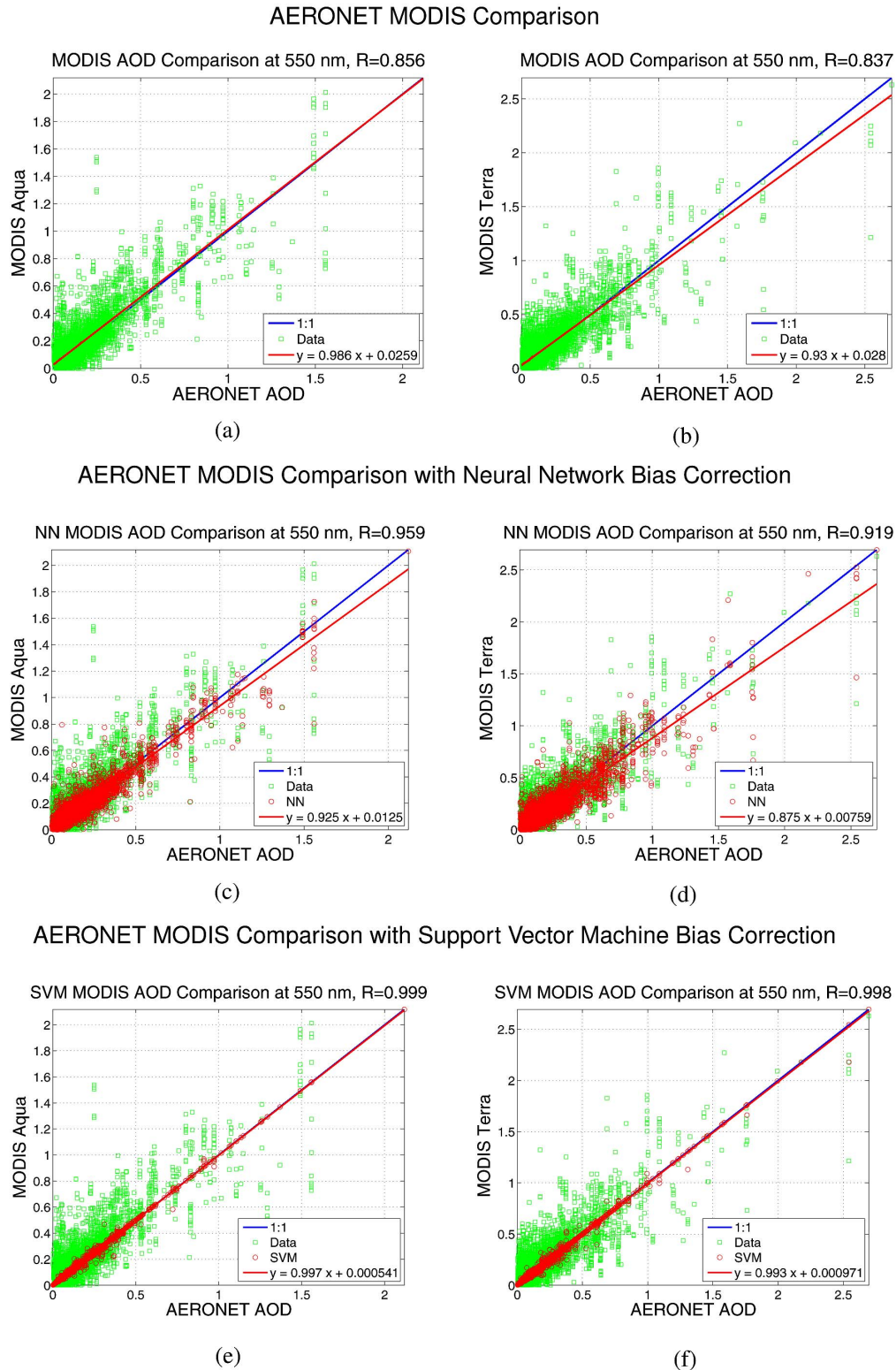


Fig. 2. Scatter diagram comparisons of AOD from AERONET ( $x$ -axis) and MODIS ( $y$ -axis) as green circles overlaid with the ideal case of perfect agreement (blue line). The measurements shown in the comparison were made within half an hour of each other, with a great circle separation of less than  $0.25^\circ$  and with a solar zenith angle difference of less than  $0.1^\circ$ . The left-hand column of plots is for MODIS Aqua, and the right-hand column of plots is for MODIS Terra. The first row shows the comparisons between AERONET and MODIS for the entire period of overlap between the MODIS and AERONET instruments from the launch of the MODIS instrument to the present. The second row shows the same comparison overlaid with the neural network correction as red circles. We note that the neural network bias correction makes a substantial improvement in the correlation coefficient with AERONET. An improvement from 0.86 to 0.96 for MODIS Aqua and an improvement from 0.84 to 0.92 for MODIS Terra. The third row shows the comparison overlaid with the SVM correction as red circles. We note that the SVM bias correction makes an even greater improvement in the correlation coefficient than the neural network correction. An improvement from 0.86 to 0.99 for MODIS Aqua and an improvement from 0.84 to 0.99 for MODIS Terra.



output variable (i.e., a multivariate nonlinear nonparametric fit). For each data set, the inputs were the surface type, the solar zenith angle, the solar azimuth angle, the sensor zenith angle, the sensor azimuth angle, the scattering angle, the reflectance, and the MODIS AOD. For each data set, the output was the AERONET AOD at 550 nm.

1) *Neural Networks*: Neural networks are multivariate nonparametric “learning” algorithms [9], [10] that are ideally suited to learning, and correcting for, interinstrument biases.

When training a neural network, we randomly split the training data set into three portions of 80%, 10%, and 10%. The 80% of the data are used to train the neural network weights. This training is iterative, and on each iteration, we evaluate the current rms error of the neural network. The rms error is calculated by using the second 10% of the data that were not used in the training. We use the rms error and the way it changes with training iteration (epoch) to determine the convergence of our training. When the training is complete, we use the final randomly chosen 10% as a validation data set. These 10% of the data were randomly chosen and not used in either the training or rms evaluation. We only use the neural network if the validation scatter diagram, which plots the actual data from validation portion against the neural network estimate, yields a straight-line graph with a slope of one. This is a stringent and independent validation. The validation is global as the data were randomly selected over all temporal and spatial data points available. The neural network algorithm used was a feedforward back-propagation network with 20 hidden nodes. The training was done by the Levenberg–Marquardt back-propagation algorithm provided by the Matlab neural network toolbox (<http://www.mathworks.com/products/neuralnet/>).

Fig. 2(c) and (d) shows the result of performing a neural network bias correction. We see that the neural network is able to make a substantial improvement in the correlation coefficient with AERONET: an improvement from 0.86 to 0.96 for MODIS Aqua and an improvement from 0.84 to 0.92 for MODIS Terra.

When we perform linear regression on the scatter diagram of AERONET AOD versus the MODIS AOD corrected by the neural network fit, we see that the intercept (bias) is considerably reduced, from 0.03 to 0.01 for both Aqua and Terra. However, the slope of the neural network fit is not close to one.

2) *SVMs*: SVMs were initially used for classification and are based on the concept of decision planes that define decision boundaries and were first introduced by Vapnik [11], [12]. SVMs have subsequently been extended by others to include regression, i.e., support vector regression (SVR) [13], [14]. In this letter, we use the SVR provided by LIBSVM [15], [16].

Fig. 2(e) and (f) shows the result of performing an SVR bias correction. The SVR makes an even greater improvement than the neural network correction, improving the correlation coefficient from 0.86 to 0.99 for MODIS Aqua and from 0.84 to 0.99 for MODIS Terra.

When we perform linear regression on the SVM fit, we see that the intercept (bias) is considerably reduced, from 0.03 to 0.0005 for Aqua and from 0.03 to 0.0001 for Terra. In addition, the slope of the SVM fit is almost 1 (0.99) for both Aqua and Terra.

3) *Why the Improvement*: Why did the SVM model outperform the neural networks? SVMs use a kernel function to map the data into a different space. The concept of a kernel mapping function is very powerful. The SVM model algorithmic process utilizes higher dimensional space to achieve superior predictive power.

The SVM algorithmic process offers an important advantage compared with neural network approaches. Specifically, neural networks can suffer from multiple local minima; in contrast, the solution to an SVM is global and unique. This characteristic may be partially attributed to the development process of these algorithms; SVMs were developed in the reverse order to the development of neural networks. SVMs evolved from the theory to implementation and experiments; neural networks followed a more heuristic path, from applications and extensive experimentation to theory.

4) *Factor Analysis*: As we have seen, there is a suite of variables available that can be used collectively to empirically “correct” the MODIS AOD to better agree with AERONET. However, some of these variables “overlap” in the sense that groups of them are interdependent. We can determine if this is so by using factor analysis. Factor analysis is a well-established statistical method used to explain the variability among a set of observed variables in terms of fewer unobserved variables called factors [17]–[19]. The observed variables are modeled as linear combinations of these underlying factors, plus error terms. Factor analysis determines that surface type is the variable that best explains the bias in the MODIS AOD data. Trends are similar for both Aqua MODIS and Terra MODIS.

In addition, the MODIS solar zenith and scattering angles also have a weak correlation (correlation coefficients between 0.1 and 0.2) with the AOD difference between MODIS and AERONET.

## V. SIGNIFICANCE

MODIS-derived AOD may show dependence on surface type either because of the link between surface type and surface reflectance or because of the covariance between aerosol properties and surface type. Different surface types (e.g., forests, croplands, pastures, bare rock, or soil) exhibit varying reflectance properties. For example, deciduous forests in full foliage are dark, with reflectances in the range of 0.03–0.10 in the visible portion of the solar spectrum. Bare soil or rock is bright, with reflectances that can be as high as 0.3–0.4. The MODIS algorithm needs to extract an atmospheric aerosol signal from the combined surface–atmosphere reflectances measured by the satellite sensor. The separation of atmosphere from surface reflectance is based on assumptions concerning spectral properties of the surface [3]. These surface spectral properties are determined empirically and are dependent on sun–satellite geometry and an atmospherically resistant vegetation index ( $NDVI_{SWIR}$ ) [3], [20]. The results of the neural network exercise suggest a residual dependence on surface type in the assumptions of surface reflectance that is not already parameterized by the vegetation index. Note that, in the development of the current Collection 5 MODIS aerosol algorithm over land, surface type was explored as a possible influential factor

before vegetation index was chosen as the parameter. Vegetation index was chosen over surface type because no unique linear relationship was found between surface reflectance and wavelength, contingent upon surface type. The neural network analysis provides a nonlinear relationship that otherwise could not have been found.

On the other hand, the reason for dependence between the MODIS AOD and surface type may have nothing to do with surface reflectance, but instead be linked to aerosol optical properties found in different places of the world. For example, we expect to find a dominance of dust aerosol over bare or desert surfaces and urban/industrial pollution over urban surfaces. Other relationships may not be so obvious, but could be revealed by the nonlinear neural network analysis. The MODIS retrieval algorithm requires assumptions of aerosol properties in order to retrieve aerosol loading. Assuming dust when the aerosol is actually urban pollution will result in a significantly large error in the AOD retrieval. The assumptions of aerosol properties are based on a cluster analysis of AERONET retrieval data that are fixed seasonally and geographically [21]. While this distribution should represent typical values, it will introduce errors whenever the actual aerosol properties differ from the expected. The neural network analysis may represent an adjustment to the algorithm's global and seasonal distribution of assumed aerosol properties, resulting in collocated retrievals closer to AERONET observations.

Overall, the machine-learning results show us that there is opportunity in the MODIS aerosol algorithm to improve the accuracy of the AOD retrieval, as compared with AERONET, and that this improvement is linked to surface type. We can use information from AERONET, from other satellite sensors such as MISR, and from detailed field experiments to continue to test and refine the assumptions in the MODIS algorithm. The results from the machine-learning analysis that point to surface type as the missing piece of information will allow us to focus the refinement procedure where it will help most.

## VI. CONCLUSION

Machine-learning algorithms were able to effectively adjust the AOD bias seen between the MODIS instruments and AERONET. SVMs performed the best, improving the correlation coefficient between the AERONET AOD and the MODIS AOD from 0.86 to 0.99 for MODIS Aqua and from 0.84 to 0.99 for MODIS Terra. Key in allowing the machine-learning algorithms to "correct" the MODIS bias was provision of the surface type and other ancillary variables that explain the variance between MODIS and AERONET AOD.

## REFERENCES

- [1] R. Pachauri and A. Reisinger, Eds., *Climate Change 2007 Synthesis Report*. Geneva, Switzerland: IPCC, UNEP, 2007.
- [2] L. A. Remer, Y. J. Kaufman, D. Tanre, S. Mattoo, D. A. Chu, J. V. Martins, R. R. Li, C. Ichoku, R. C. Levy, R. G. Kleidman, T. F. Eck, E. Vermote, and B. N. Holben, "The MODIS aerosol algorithm, products, and validation," *J. Atmos. Sci.*, vol. 62, no. 4, pp. 947–973, Apr. 2005.
- [3] R. C. Levy, L. A. Remer, S. Mattoo, E. F. Vermote, and Y. J. Kaufman, "Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of Moderate Resolution Imaging Spectroradiometer spectral reflectance," *J. Geophys. Res.—Atmos.*, vol. 112, no. D13, p. D13 211, Jul. 13, 2007.
- [4] B. Holben, T. Eck, I. Slutsker, D. Tanre, J. Buis, A. Setzer, E. Vermote, J. Reagan, Y. Kaufman, T. Nakajima, F. Lavenu, I. Jankowiak, and A. Smirnov, "AERONET—A federated instrument network and data archive for aerosol characterization," *Remote Sens. Environ.*, vol. 66, no. 1, pp. 1–16, Oct. 1998.
- [5] L. A. Remer, R. G. Kleidman, R. C. Levy, Y. J. Kaufman, D. Tanre, S. Mattoo, J. V. Martins, C. Ichoku, I. Koren, H. Yu, and B. N. Holben, "Global aerosol climatology from the MODIS satellite sensors," *J. Geophys. Res.—Atmos.*, vol. 113, no. D14, p. D14 S07, Jul. 29, 2008.
- [6] M. E. Brown, D. J. Lary, A. Vrieling, D. Stathakis, and H. Mussa, "Neural networks as a tool for constructing continuous NDVI time series from AVHRR and MODIS," *Int. J. Remote Sens.*, vol. 29, no. 24, pp. 7141–7158, Dec. 2008.
- [7] N. Xiao, T. Shi, C. A. Calder, D. K. Munroe, C. Berrett, S. Wolfenbarger, and D. Li, "Spatial characteristics of the difference between MISR and MODIS aerosol optical depth retrievals over mainland Southeast Asia," *Remote Sens. Environ.*, vol. 113, no. 1, pp. 1–9, Jan. 15, 2009.
- [8] S. Paradise, A. Braverman, N. Cressie, R. Kahn, and B. Wilson, "Long-term global comparisons of aerosol optical depth from MISR, MODIS and AERONET using AMAPS," presented at the Fall AGU, San Francisco, CA, 2007.
- [9] C. Bishop, *Neural Networks for Pattern Recognition*. London, U.K.: Oxford Univ. Press, 1996.
- [10] S. Haykin, *Kalman Filtering and Neural Networks*. New York: Wiley-Interscience, 2001.
- [11] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
- [12] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998, ser. Adaptive and Learning Systems for Signal Processing, Communications, and Control.
- [13] B. Scholkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, May 2000.
- [14] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Stat. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.
- [15] R. E. Fan, P. H. Chen, and C. J. Lin, "Working set selection using second order information for training support vector machines," *J. Mach. Learn. Res.*, vol. 6, pp. 1889–1918, Dec. 2005.
- [16] P. H. Chen, R. E. Fan, and C. J. Lin, "A study on SMO-type decomposition methods for support vector machines," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 893–908, Jul. 2006.
- [17] K. G. Joreskog, "Some contributions to maximum likelihood factor analysis," *Psychometrika*, vol. 32, no. 4, pp. 443–482, Dec. 1967.
- [18] D. N. Lawley and A. E. Maxwell, *Factor Analysis as a Statistical Method*, 2nd ed. New York: American Elsevier, 1971.
- [19] H. H. Harman, *Modern Factor Analysis*, 3rd ed. Chicago, IL: Univ. Chicago Press, 1976.
- [20] A. Karnieli, Y. J. Kaufman, L. A. Remer, and A. Wald, "AFRI—Aerosol free vegetation index," *Remote Sens. Environ.*, vol. 77, no. 1, pp. 10–21, Jul. 2001.
- [21] R. C. Levy, L. A. Remer, and O. Dubovik, "Global aerosol optical properties and application to Moderate Resolution Imaging Spectroradiometer aerosol retrieval over land," *J. Geophys. Res.—Atmos.*, vol. 112, no. D13, p. D13 210, Jul. 13, 2007.