Aerosol and Air Quality Research, 16: 1509–1522, 2016

ISSN: 1680-8584 print / 2071-1409 online
doi: 10.4209/aaqr.2015.05.0375

A Method to Improve MODIS AOD Values: Application to South America

Bethania L. Lanzaco, Luis E. Olcese*, Gustavo G. Palancar, Beatriz M. Toselli

INFIQC - CONICET / CLCM / Departamento de Fisicoquímica. Facultad de Ciencias Químicas. Universidad Nacional de Córdoba. Ciudad Universitaria, 5000 Córdoba, Argentina

ABSTRACT

We present a method to correct aerosol optical depth (AOD) values taken from Collection 6 MODIS observations, which resulted in values closer to those recorded by the ground-based network AERONET. The method is based on machine learning techniques (Artificial Neural Networks and Support Vector Regression), and uses MODIS AOD values and meteorological parameters as inputs.

The method showed improved results, compared with the direct MODIS AOD, when applied to nine stations in South America. The percentage of improvement, measured in terms of R^2, ranged from 2% (Alta Floresta) to 79% (Buenos Aires). This improvement was also quantified considering the percentage of data within the MODIS expected error, being 91% for this method and 57% for direct correlation.

The method corrected not only the systematic bias in temporal data series but also the outliers. To highlight this ability, the results for each AERONET station were individually analyzed.

Considering the results as a whole, this method showed to be a valuable tool to enhance MODIS AOD retrievals, especially for locations with systematic deviations.

Keywords: Support Vector Regression; Artificial Neural Networks; AOD satellite retrieval; MODIS AOD bias correction; AERONET.

INTRODUCTION

The importance of aerosols has been increasingly stressed in recent decades. As aerosols scatter and absorb solar radiation, changes in their atmospheric concentrations and their chemical and physical properties can alter the transmission of the radiation through the atmosphere and impact climate change (IPCC, 2007). Regardless of the large number of studies dealing with the radiative properties of aerosols, their net effect on global climate is still unknown and represents one of the major uncertainties in the understanding of Earth's climate system (e.g., Bond and Bergstrom, 2006). An accurate estimation of the radiative effects of aerosols requires knowledge of their optical and physical properties (Tripathi *et al.*, 2005). However, as a consequence of their short lifetime, aerosols exhibit a strong spatial and temporal variation in their properties and concentrations.

To retrieve or infer the optical properties of aerosols, different procedures are used, mainly based on the interaction between particles and radiation (e.g., Giles *et al.*, 2012). Radiation measurements can be made at the surface or at different platforms such as satellites. Ground-based observations are measurements at a single point, and therefore cannot account for spatial variations unless a large network is operational. This gap can be filled by satellite sensors, which have the great advantage of covering the whole globe in a rather homogeneous manner. However, satellite measurements are based on important assumptions about the aerosol and surface properties. Surface measurements have not such constraint and therefore provide the aerosol optical properties with high accuracy. This is the reason why aerosol optical properties retrieved from satellites are usually validated against surface measurements (e.g., Estellés *et al.*, 2012).

The earth viewing sensor MODerate resolution Imaging Spectroradiometer (MODIS) aboard the NASA Earth Observing System Terra and Aqua satellites provides aerosol products in near-real time for monitoring and forecasting the aerosol transport (Kaufman *et al.*, 1997). The evaluation of any retrieved data product is of critical relevance to establish its quality and suitability to be used in radiative, weather, and air pollution models.

On a global scale, several studies have compared MODIS based aerosol optical depth (AOD) measurements with ground based measurements retrieved by AERONET (Holben *et al.*, 1998) for different areas showing, in general, a good

* Corresponding author.
Tel.: +54-351-5353866; Fax: +54-351-4334188
E-mail address: lolcese@fcq.unc.edu.ar

correlation (e.g., Remer *et al.*, 2008; More *et al.*, 2013). Similar studies have been performed using the Multi-angle Imaging SpectroRadiometer (MISR) aboard Terra satellite, obtaining comparable correlations (Zhang and Reid, 2006; Khan *et al.*, 2010). However, on a global comparison made by Mishchenko *et al.* (2010), the agreement is far less favorable than what has been obtained in previous studies.

The number of studies investigating the satellite against ground based measurements correlation in South American stations is much more limited and they focus mostly on the biomass burning season. In one of these studies, Castro Videla *et al.* (2013) investigated the relative contribution of different South American biomass burning zones to the continental aerosol load. A comprehensive study for most of South American stations during the biomass burning season (August–October) has been published by Hoelzemann *et al.* (2009). In this work, they present a 2001–2007 comparison between AERONET AOD observations and the MODIS AOD product from the Collection 5, finding that MODIS systematically underestimates the low AOD values and overestimates the high ones. MODIS products present some limitations over South America and some other regions around the globe. These limitations are related to the fact that the absorption and scattering properties of the aerosols display a significant variability on local or regional scales (Hoelzemann *et al.*, 2009). In addition, most algorithms that retrieve AOD from satellite observation are derived from forward-simulation models according to the domain knowledge of the aerosol physical properties. AOD is retrieved by matching observed reflectance with the simulated values stored in the Look-Up Tables. One of the main problems of these tables is that they are generated using the aerosol properties from better characterized regions. Thus, they frequently have differences with the local aerosol properties. These algorithms are periodically tuned by domain scientists after validating the AOD values against the AERONET retrievals. Despite these drawbacks, satellite-based observations help us to improve our knowledge on the geographical and temporal variation of aerosol properties, and for this reason, they represent an essential complement of the spatially limited surface measurements when seen from a large-scale perspective. Therefore, it would be useful to have a way to enhance the accuracy of MODIS, mainly for those cases where it has been demonstrated to fail. An efficient strategy is to build a model based on the collocated satellite and ground-based observations using satellite-based observations as inputs and ground-based observations as outputs. This model can therefore be used to predict AOD from satellite observations where ground-based retrievals are no longer available (temporally or permanently). If sufficient amount of training data is available, this model would be flexible to different retrieval scenarios and more accurate than the deterministic algorithms. Machine learning methods fulfill these requirements and have therefore been widely used in the aerosol science field in the last years (e.g., Hirtl *et al.*, 2014). These studies use global or regional data sets to correct MODIS biases at these scales. However, they do not allow improving the retrievals in regions with low density of AERONET sites. Also, they do not account for the behavior of individual AERONET stations, making them reliable only on a global scale.

In this work, a method to improve MODIS AOD (AOD_M), using AERONET AOD (AOD_A) as a reference, is presented. The method was applied to South American stations for the whole MODIS measurement period (2000–2014). We used the recently released MODIS Collection 6 (C6), which improves the algorithm from Collection 5, and corrects calibration errors of the MODIS sensor onboard Terra satellite (Levy *et al.*, 2013). The method is based on machine learning techniques that significantly improve the results of the direct (AOD_M against AOD_A) correlation. The models were trained separately for every AERONET station in South America, seeking to improve the retrievals at each site and to provide a simple and reliable way to correct local bias and outliers. In order to make the method applicable to raw data, the outliers were not removed from the data set. To keep the method requirements at a minimum, only meteorological variables and two new variables (related to the average variation of the meteorological conditions throughout the year) were added as possible inputs.

DATA

AERONET

AERONET (AErosol RObotic NETwork) is a remote sensing aerosol monitoring network of CIMEL sun-sky photometers, established and maintained by NASA and LOA-PHOTONS (CNRS) (Holben *et al.*, 1998). It provides a long-term public database of aerosol optical properties in strategic sites all over the world.

The photometer makes three solar extinction measurements in eight spectral bands 30 seconds apart, creating a triplet observation for each wavelength. These triplet observations are made every 15 minutes, and are then used to compute AOD. Sharp discontinuities among the triplets or between consecutive triplet averages allow cloud screening because of the different time variation presented by clouds compared to the aerosols.

In this study, level 2.0 AODs (cloud screened and quality assured for instrument calibration), processed with the algorithm version 2 were used. As the AOD value at 550 nm (measured by MODIS) is not provided by AERONET, the available AOD values at all the other wavelengths were adjusted to a log-log quadratic regression at a given time (Eck *et al.*, 1999).

Only stations having at least a thousand days of measurements were included in this work: Cordoba-CETT, CEILAP-BA, Arica, Sao_Paulo, CUIABA-MIRANDA, Campo_Grande_SONDA, Alta_Floresta, Rio_Branco, and SANTA_CRUZ (Fig. 1 and Table 1). In addition, ground-based meteorological data have to be available within 40 km from the AERONET site. However, some of the stations that satisfied those criteria were not used. For example, La Paz, CASLEO, and Trelew were not used because the AOD_A values were very low (within the MODIS sensitivity).

MODIS

The MODIS (Moderate Resolution Imaging

Fig. 1. Geographical location of the used AERONET stations and the South American largest cities.

Table 1. AERONET stations with their location, number of MODIS data, and average AOD_A (interpolated at 550) for the measurement period.

Name	Location	Long; Lat [degrees]	Elevation [masl]	Number of MODIS data Terra / Aqua	Average AOD_A 550 nm
Alta_Floresta	Alta Floresta, Brazil	–56.10; –9.87	277	468 / 356	0.27
Arica	Arica, Chile	–70.31; –18.47	25	320 / 428	0.22
Campo_Grande_SONDA	Campo Grande, Brazil	–54.54; –20.44	677	635 / 472	0.13
CEILAP-BA	Buenos Aires, Argentina	–58.50; –34.57	10	999 / 1023	0.092
Cordoba-CETT	Córdoba, Argentina	–64.46; –31.52	730	765 / 602	0.084
CUIABA-MIRANDA	Cuiabá , Brazil	–56.02; –15.73	210	836 / 597	0.23
Rio_Branco	Río Branco, Brazil	–67.87; –9.96	212	490 / 213	0.25
SANTA_CRUZ	Santa Cruz de la Sierra, Bolivia	–63.18; –17.80	442	250 / 218	0.18
Sao_Paulo	São Paulo, Brazil	–46.73; –23.56	865	367 / 208	0.22

Spectroradiometer) instruments were launched aboard the Terra and Aqua satellites in the years 2000 and 2002, respectively, to make global observations of the earth in a wide wavelength range (Kaufman *et al.*, 1997). These measurements are used to derive several aerosol parameters, based on the fact that the aerosol contribution is low at 2.1 μm. This fact allows determining the surface reflectance at this wavelength and estimating its contribution in the spectral visible range. The top-of-atmosphere and surface reflectances are used as inputs of dynamical aerosol models to retrieve AOD values. Two algorithms are applied to retrieve AOD over land: dark target (Kaufman *et al.*, 1997), developed to be used over dense and dark vegetation, and deep blue (Hsu *et al.*, 2004), that provides coverage over bright surfaces such as deserts.

In this work, we used the recently released Atmosphere level 2, C6, product MOD04_L2 for Terra, and MYD04_L2 for Aqua. Details of the modifications introduced in C6 are explained by Levy *et al.* (2013). All data available for each satellite up to the year 2014 were included. Results for deep blue algorithm are not presented in this work (except for CEILAP-BA site) due to a variety of reasons. For Amazon rainforest-influenced stations, it is known (Sayer *et al.*, 2014) that deep blue algorithm provides around 20% less matchups than dark target. As Sao_Paulo exhibits a similar yearly aerosol variation to the Amazonian stations, dark target was also the chosen algorithm. For the other sites, both algorithms were evaluated in all the cases in order to find out the best one to train the models. Dark target had the best correlation with AERONET in all the sites, except for Aqua satellite in CEILAP-BA site, where deep blue produced better results. It was noticeable that in the Arica site, a better performance of the deep blue algorithm was expected due to the proximity of the site to the desert, but its R^2 was much lower than dark target (0.1 vs. 0.7).

AOD at 550 nm was retrieved for both ocean and land only if it had the best quality data (QA Confidence Flag = 3). The valid range for this parameter goes from –0.05 to 5.00. Negative values appear because MODIS does not have sensitivity over land to retrieve AOD values with accuracy better than ±0.05. These values are indicative of very clean conditions and they are allowed in order to avoid an artificial bias in statistics.

AOD 550 values are available from the Level 1 and Atmosphere Archive and Distribution System (LAADS) website (http://ladsweb.nascom.nasa.gov) and are delivered in Hierarchical Data Format (HDF), with a pixel resolution of 10 km × 10 km (at nadir).

METHODOLOGY

AOD Spatiotemporal Collocation

Comparison between MODIS and AERONET values is not direct: even though the measurement times were coincident (which rarely happens), the single data obtained by the sun-photometer at a given time is not equivalent to the geographically averaged AOD retrieved over the MODIS pixel. Hence, to validate the MODIS retrievals, it is necessary to employ a spatiotemporal approach to ensure a proper comparison between both measurements. To do so, the spatial average of the pixels that fall within a 22.5 km radius from the AERONET station is computed and correlated with the temporal average of the AERONET measurements that fall within ±30 minutes of MODIS overpass (Petrenko *et al.*, 2012). In this work, the geographical and temporal averages were calculated. To evaluate the temporal averages, at least two data points were required. This widely used (e.g., Remer *et al.*, 2008) spatiotemporal approach is available at MAPSS webpage (http://giovanni.gsfc.nasa.gov/mapss/) for each AERONET station over the world.

Machine Learning Methods

Machine learning is a subfield of artificial intelligence which attempts to develop algorithms that can empirically learn from the behavior or the properties of a given data set. The methods used in this work were Artificial Neural Networks (ANN) and Support Vector Machines (SVM). In these methods, a training data set is provided separated into "inputs" and "outputs", and the algorithm tries to find a connection between them. The outputs are the variables to be predicted, while the inputs are the variables which will feed the algorithm. A portion of the dataset was not used during the training; it was reserved only for validation purposes.

In this work, the goal is to obtain an AOD value from MODIS as close as possible to the one obtained from AERONET. AOD_M (collocated with AOD_A as previously described) with the addition of variables representing the day of the year and meteorology were used as inputs, and AOD_A was used as output.

Machine Learning Methods - Artificial Neural Network

ANNs are a kind of machine learning method inspired in biological nervous systems (Basheer and Hajmeer, 2000). ANNs are composed of several processing units called neurons, working in parallel. In the structure of a typical ANN, the inputs are connected to one or several layers of neurons which, in turn, are connected to the outputs. The weights of these connections are the parameters adjusted during an ANN training. The training process is iterative: in each loop the root mean square error, rmse, is evaluated and, according to the result, the connection weights are modified until the network can correctly predict the outputs. This way of training is called supervised learning because the truthful outputs are provided.

In this work, the training was carried out using the Matlab® Neural Network Toolbox (version 2011Rb). The data set was randomly separated into three sub-sets: 70% of the data were used for training the ANN, 15% of the data were used for validation (by evaluating the rms error), and 15% of the data were used for testing after the training was completed. The ANNs were trained using the Levenberg-Marquardt back-propagation algorithm, changing both the number of neurons (only one layer was used) and the transfer function in order to find the best network architecture for each data set.

Machine Learning Methods -Support Vector Regression

SVMs were first introduced and developed for

classification and regression problems (Vapnik, 1995). SVMs are widely accepted in the machine learning community. They have been used for a broad range of applications because of their ability to generalize (e.g., Hirtl *et al.*, 2014). One of its important characteristics is that when a data set is not linearly separable, the SVM method uses a kernel representation to project the data onto a high dimensional feature space where the linear separation is possible.

Briefly, the goal in the Support Vector Regression technique (SVR) is to find a function $f(x)$ that has at most ε deviation from the outputs of the training data set and, at the same time, is as flat as possible to control the system complexity and the training error simultaneously. However, not all data pairs can be adjusted with ε deviation and need to be captured by slack variables ξ_i, ξ_i^* ($\xi_i^{(*)}$ refers to both of them), which are penalized in the function by introducing a regularization constant C. In this work, the method υ-SVR was used (Schölkopf *et al.*, 2000), which introduces a constant $\upsilon \in (0, 1)$ that moderates the value of ε. The size of ε is a tradeoff between model complexity and slack variables.

Given a simple training data set $\{(x_1, y_1), \ldots, (x_l, y_l)\}$ the linear function can be written in terms of a weight vector w and a bias b:

$$f(\mathrm{x}) = \langle \mathrm{w}, \mathrm{x} \rangle + b \,. \tag{1}$$

Function flatness can be achieved seeking a small w, and one way to ensure this is to minimize the norm ($\|\mathrm{w}\|^2 = \langle w, w \rangle$). The problem can be written as a convex optimization problem, which allows finding the global minimum, i.e.,

$$\text{Minimize } \frac{1}{2}\|\mathrm{w}\|^2 + C \cdot \left(\upsilon\varepsilon + \frac{1}{l}\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \right), \text{ subject to}$$

$$\begin{cases} \langle \mathrm{w}, \mathrm{x} \rangle + b - y_i \le \varepsilon + \xi_i \\ y_i - \langle \mathrm{w}, \mathrm{x} \rangle - b \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^*, \varepsilon \ge 0 \end{cases} . \tag{2}$$

This optimization problem can be solved more easily in its dual formulation, introducing multipliers for the constraints ($\alpha_i^{(*)}$, $\eta_i^{(*)}$, β) and constructing the Lagrangian In order to minimize expression (2), it is necessary to find the saddle point of the Lagrangian function, which yields to Eq. (3) and the support vectors.

$$w = \sum_i \left(\alpha_i^* - \alpha_i\right) x_i \tag{3}$$

To make the SV algorithm nonlinear, the training data pairs are mapped into some feature space with higher dimensionality using a kernel function $k(x, y)$ to carry out all computations in that space but without using directly the function Φ that maps into that space

$$k(x, y) = (\Phi(x) \cdot \Phi(y)) \tag{4}$$

Finally, the υ-SVR optimization problem is to maximize:

$$W\left(\alpha^{(*)}\right) = \sum_{i=1}^{l}\left(\alpha_i^* - \alpha_i\right) y_i - \frac{1}{2}\sum_{i,j=1}^{l}\left(\alpha_i^* - \alpha_i\right)\left(\alpha_j^* - \alpha_j\right) k\left(x_i, x_j\right), \text{ subject to}$$

$$\begin{cases} \sum_{i=1}^{l}\left(\alpha_i^* - \alpha_i\right) = 0 \\ \alpha_i^{(*)} \in \left[0, \frac{C}{l}\right] \\ \sum_{i=1}^{l}\left(\alpha_i^* - \alpha_i\right) \le C \cdot \upsilon \end{cases} . \tag{5}$$

In this work, the υ-SVR provided by the library LIBSVM (Chang and Lin, 2011) was used. The data set was randomly separated into two equally sized sub-sets. One sub-set was used for training and the other one for testing. The histogram of the output values was uniform for both sub-sets. For each data set, the cross validation accuracy for different combinations of C and γ (a kernel parameter) was evaluated to find the best combination for each particular case. The value of υ was changed only for one data set, observing no significant improvement in the correlation. Therefore, a fixed value of υ (0.5), which is the default value in LIBSVM, was used.

Inputs Design

With the purpose of finding the best set of inputs for each station, all the variables were collocated in space and time as described before. Day of the year (DOY) could be used as a parameter to represent meteorological conditions. However, to make the DOY representative, it was necessary to introduce some modifications, given the fact that January 1st and December 31st have nearly the same average conditions. This is because they are placed at the ends of the DOY scale (1 and 365, respectively). Therefore, the DOY was modified to create a different, but related variable: the modified DOY (MDOY) (Olcese *et al.*, 2015). The MDOY is equal to 1 on January 1st, reaches its highest value (183) on July 1st and July 2nd, and then decreases to be again equal to 1 on December 31st (Eq. (6)), so it represents the annual average trend of meteorological conditions. In addition, to differentiate MDOY with the same values, a complementary variable called MDOY direction was introduced.

$$MDOY = \begin{cases} DOY, DOY \le 183 \\ |DOY - 365|, DOY > 183 \end{cases}$$

$$MDOY\ direction = \begin{cases} 0, DOY \le 183 \\ 1, DOY > 183 \end{cases} \tag{6}$$

In some cases (Table 2), actual meteorological data needed to be included in order to improve the prediction capability. In these cases, meteorological variables (temperature, relative humidity, wind speed and direction)

Table 2. Input variables for the best model trained at each AERONET station.

Variable	Alta_Floresta	Arica	Campo_Grande_SONDA	CEILAP-BA	Cordoba-CETT	CUIABA-MIRANDA	Rio_Branco	SANTA_CRUZ	Sao_Paulo
MODIS AOD value	Terra	Terra	Terra	Terra	Terra	Terra	Terra	Terra	Terra
	Aqua	Aqua	Aqua	Aqua	Aqua	Aqua	Aqua	Aqua	Aqua
MDOY + MDOY direction	Terra		Terra	Terra	Terra	Terra	Terra	Terra	Terra
	Aqua	Aqua	Aqua	Aqua	Aqua	Aqua		Aqua	Aqua
Temperature				Terra	Terra				Terra
	Aqua	Aqua	Aqua	Aqua	Aqua	Aqua	Aqua		Aqua
Relative humidity					Terra		Terra		
	Terra	Terra							
			Aqua	Aqua	Aqua		Aqua	Aqua	Aqua
Wind speed			Terra		Terra			Terra	
	Aqua	Aqua	Aqua					Aqua	
x and y wind directions		Terra	Terra	Terra			Terra		Terra
		Aqua		Aqua	Aqua				

were taken from weather stations located no further away than 40 km from the AERONET stations. Wind direction was considered as a vector with (x, y) components. The meteorological variables included here were the only ones found to be relevant to our model. As the meteorological data were available as one hour averages, some considerations about the correspondence between the MODIS passing time and the meteorological data were necessary. If the MODIS passing time over the site was during the first half of a given hour, then the average meteorological data from that particular hour were used. Otherwise, values for the next hour were used. It should be noted that not all the AOD values have their corresponding meteorological data, which implies that not all the AOD data were included in the finally selected data set. To verify that this loss of data does not introduce an artificial bias, the R^2 and the slope from both linear correlations between AOD_M and AOD_A (with and without meteorology) were compared, without observing significant differences.

Because every site is different, it is not possible to know, a priori, which is the best combination of variables producing the best correlation for each machine learning method. Therefore, after all inputs were spatial and temporally collocated, each of the 32 possible combinations among AOD_M and up to five of the other variables (MDOY and MDOY direction/temperature/relative humidity/wind speed/ x and y wind directions) was evaluated in order to determine the best set. Every combination was used for training both ANN and SVR.

Best Model Selection

A comprehensive procedure was carried out to evaluate the best combination of ANN/SVM architecture and the best combination of inputs, referred as models from now on.

Once the models were trained, a linear regression between the model output and AOD_A was performed. To evaluate model performance, the coefficients of determination (R^2) for the training, testing and the whole data set were computed. The best model and input variables for each site and satellite (Terra and Aqua) were selected according to the highest R^2 for the unseen testing set only if the difference between the highest and the lowest R^2 (for all the sub-sets) was less than 0.07. This constraint assured that the model worked well with the whole data set (seen and unseen values), ensuring also that there was no overtraining/memorization.

Choosing which machine learning method was the best for each site was not always straightforward because only in a few sites a method performed much better than the other one. The ideal method should yield the highest R2 for the whole data set, have a slope close to 1, and predict values as close as possible to AOD_A. In the sites where the results from ANN and SVR were very similar, additional conditions were considered: the method should include the lowest number of inputs, predict correctly the highest AOD values, and reproduce properly its annual variation.

RESULTS AND DISCUSSION

The direct correlation between AOD_A and AOD_M

measurements and the results for the model considered as the best one (ANN or SVR) are shown in Fig. 2 (Terra) and Fig. 3 (Aqua). In these figures, the equations for the linear fits are included in each panel. In general, these figures show several improvements with respect to the direct correlation. One of the most noticeable improvements was that almost all the outliers were corrected. The other one was that both the systematic under and overestimations existing at some sites were not longer present.

Fig. 4(a) depicts the R^2 values for the direct and improved correlations for the Terra satellite. Fig. 4(b) shows the slopes and Fig. 4(c) shows the fraction of data falling within the expected error for the dark target algorithm, which has been determined to be $\pm(0.05 + 0.15 \times AOD_A)$ (Levy *et al.*, 2013). Figs. 4(d)–4(f) shows the equivalent results, but for Aqua satellite. Fig. 5 shows the percentage of improvement of the previous variables compared to the AOD_M. The improvement in the slope is calculated considering its closeness to 1. Table 2 shows the combinations of input variables used to obtain the best results at each site.

In order to show the overall performance of the methods for all the South American stations, histograms of the difference between AOD_A and the model results, as well as the same histogram, but for the direct correlation are presented in Fig. 6. It can be seen that in 81% and 80% of the cases (Terra and Aqua, respectively) the results fall within the sensitivity of MODIS over land (±0.05), which represents a large improvement over the 42% for Terra and 42% for Aqua for the direct correlation.

However, an analysis of the continental improvement

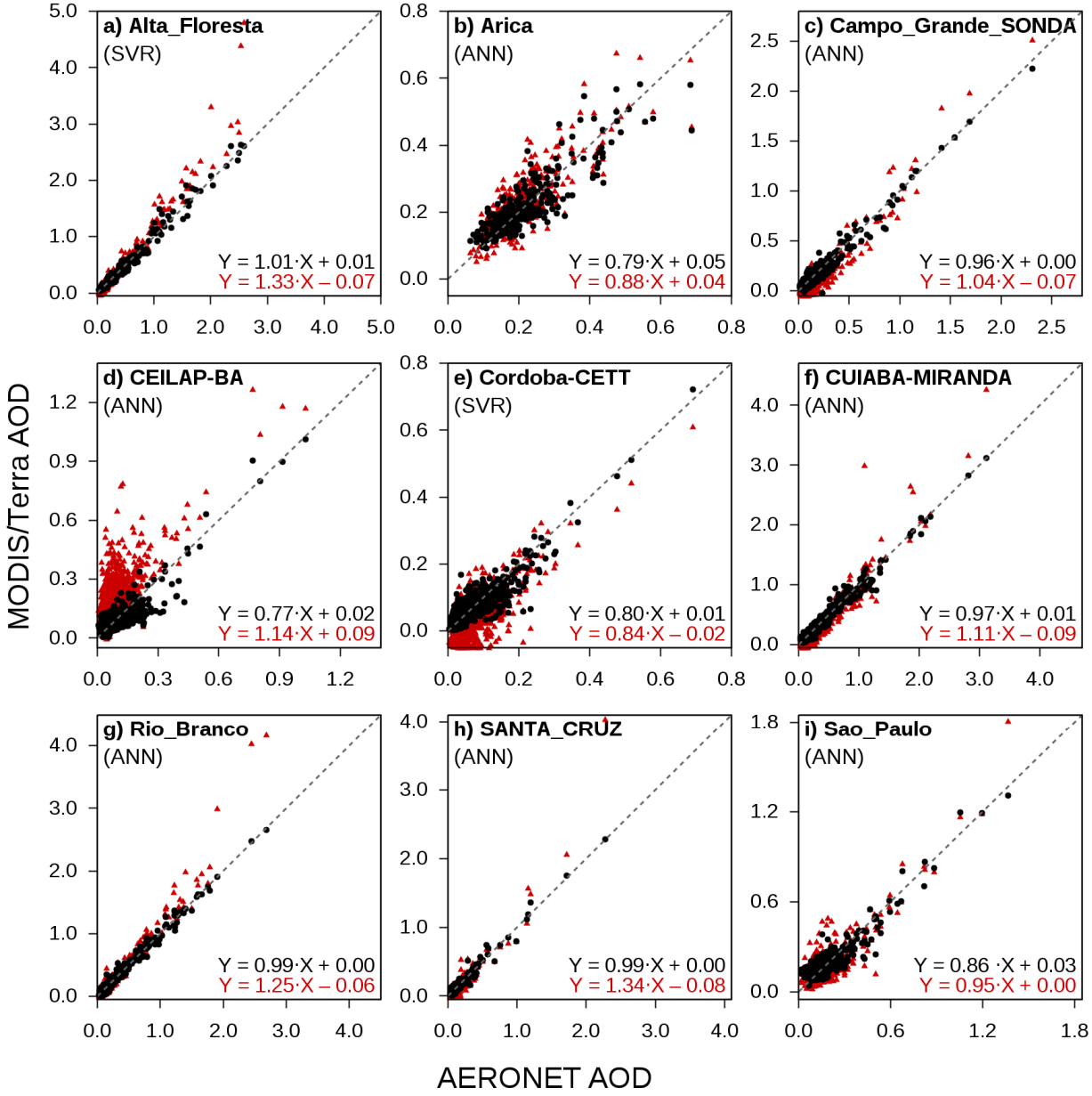

Fig. 2. Direct (red points) and improved (black points) Terra AOD_M vs. AOD_A correlation, the method used to obtain it, and the corresponding linear fit equations.

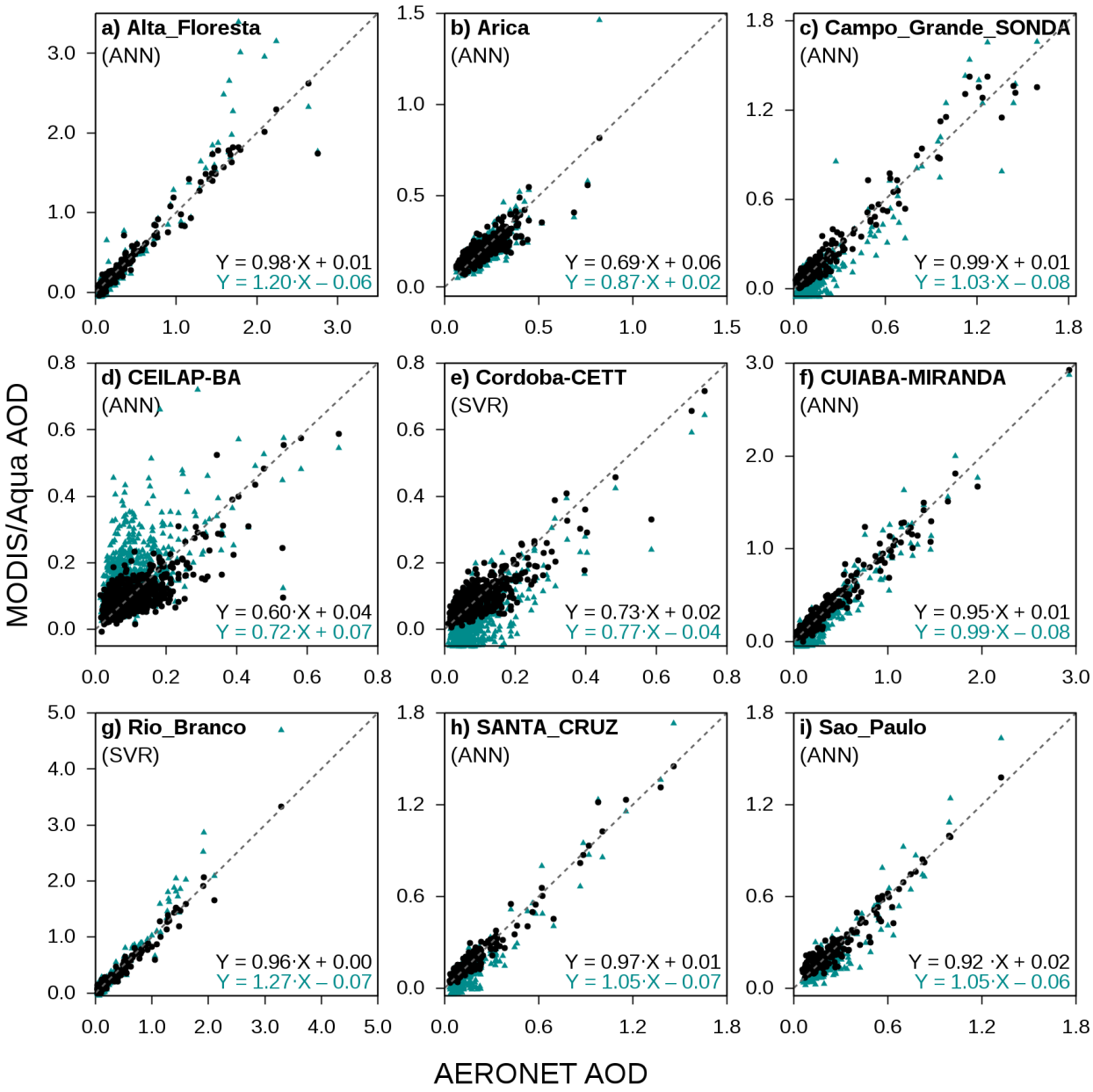

Fig. 3. Direct (green points) and improved (black points) Aqua AOD_M vs. AOD_A correlation, the method used to obtain it, and the corresponding linear fit equations.

does not reflect the MODIS drawbacks and the method performance for each particular station. Therefore, a comprehensive analysis of the results for each site is presented in the following sections.

Buenos Aires (CEILAP-BA)

The characteristics of pollution detected in Buenos Aires city are very particular. Although it is one of the South American megacities, with a population of nearly 13 million inhabitants, it does not have an aerosol load as high as it would be expected. This is mainly because of the flatness of the terrain and its coastal location at the edge of the Río de la Plata River (50 km wide at that point). These characteristics lead to a quick dilution and ventilation of the air pollutants, mainly because of the winds blowing from the river (Arkouli *et al.*, 2010).

The direct correlation between AOD_A and AOD_M values shows a R^2 equal to 0.52 and 0.33 for Terra and Aqua satellites, respectively. These low correlation values were expected, given that it is a known fact that the MODIS aerosol algorithm cannot provide accurate retrievals in coastal zones due to surface inhomogeneity and/or sub-pixel water contamination (Chu *et al.*, 2002). The regression not only is low but also shows a strong overestimation of the AOD_A values, as can be seen more clearly in Figs. 2 and 3.

This is the only site where the deep blue product from the Aqua satellite was used. A large discrepancy was observed in the AOD retrievals from both satellites (although both were low), mainly in the R^2 values. The reason for this large difference is still not clear. One possible explanation

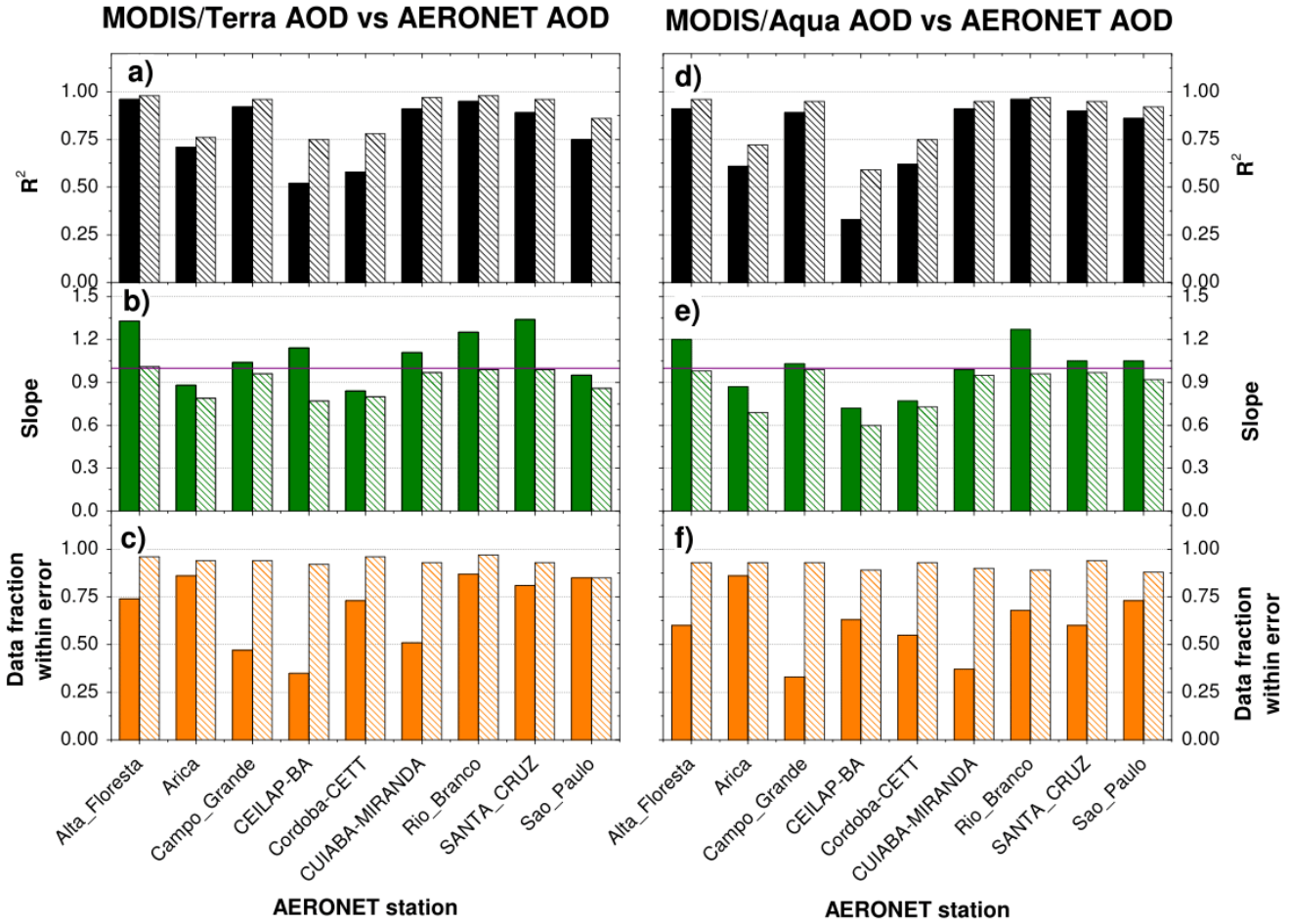

Fig. 4. R^2, slope and fraction within error values for the linear fit between AOD_M and AOD_A (solid) and between the best model and AOD_A (dashed).

is related to the effect of different aerosol patterns in the morning (Terra) and in the afternoon (Aqua). In addition, the low average aerosol load (AOD_A values are 0.092 and 0.089 at the time of Terra and Aqua overpass, respectively) coupled with the MODIS uncertainties (±0.05 units) can produce large errors in the determinations.

Regarding the correlations using machine learning methods, both SVR and ANN showed an important improvement over the direct correlation. Using ANN, the R^2 values improved 0.75 for Terra and 0.59 for Aqua; similar values were obtained using SVR. Here, it is important to note that the use of the ANN allowed discriminating which of the high AOD_M values corresponded to the actual high AOD values and which ones were consequences of errors in the MODIS algorithm. As expected, and despite the noticeable improvement in the correlation, the discrepancy in the retrievals from both satellites was not removed (in both cases there was a difference of about 0.25 units in the R^2 values).

The systematic overestimation was noticeably improved by the method. A histogram of the data (not shown), similar to Fig. 6, shows that in the improved method for Terra, 92% of all the points fall within the expected error of the MODIS algorithm (compared to 35% for the direct correlation). For Aqua, these values are 89% and 63%, respectively. Temporal series of the AOD_A minus AOD_M are presented in Figs. 7(c)–7(d), where it can be seen how the model corrects not only the systematic overestimation, but also the outliers frequently found in this station.

Córdoba (Cordoba-CETT)

The AERONET station is located in a rural area, 20 km to the west of Córdoba City (around 1.5 million inhabitants). The meteorology in this region is characterized by dry winters and rainy summers. During the dry season (April–September) the surrounding hills and mountains are prone to fires, especially during winter–spring time. Monthly AOD variation shows a spring peak with the maximum value around September–October, similarly to what has been observed for other sites in South America. In this period, the low humidity, the dryness of the soil, and the strong winds favor the relatively high aerosol loading. The opposite situation (low wind speed and frequent precipitations) is observed during summer and fall leading to the low AOD values (Olcese *et al.*, 2014).

The direct correlation between AOD_A and AOD_M values showed a R^2 of 0.58 and 0.62 for Terra and Aqua satellites, respectively. AOD values are underestimated by MODIS

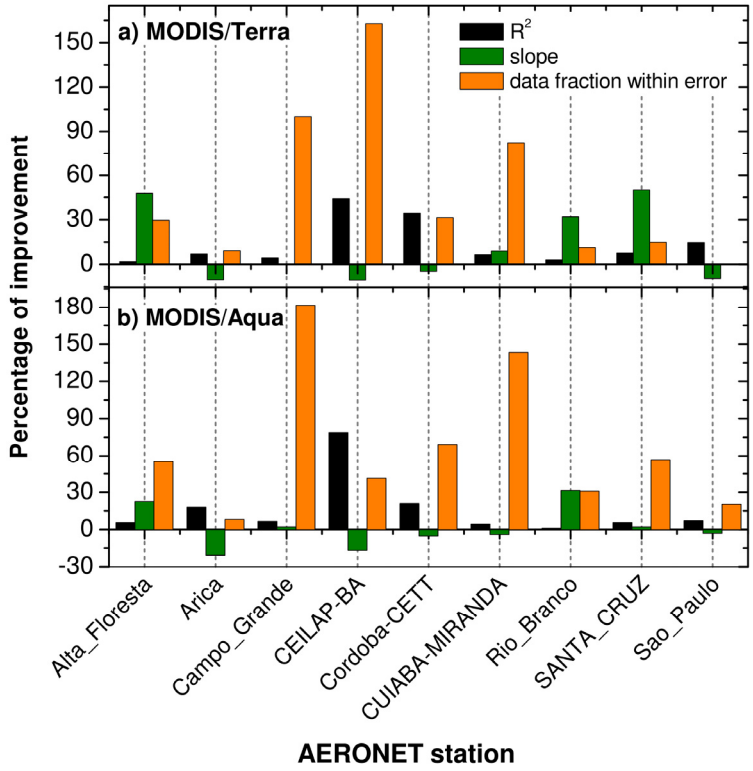

Fig. 5. Percentage of improvement of R^2, slope and data fraction within error when applying the best method compared to the direct correlation.

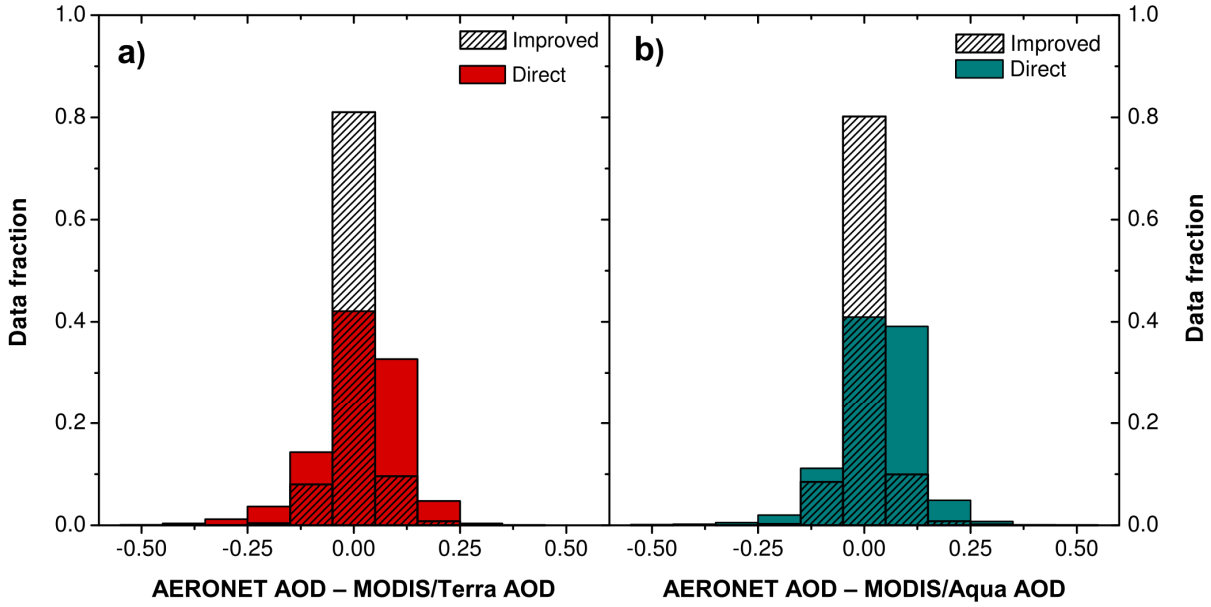

Fig. 6. Histogram of AOD_A minus AOD_M for all South American stations, and for the direct and the improved values with the chosen method.

in most of the cases, probably due to an incorrect characterization of the local aerosols and the predominantly low AOD values observed. Although the underestimation is systematic throughout the year, it is within the MODIS sensitivity. This problem is particularly noticeable for the monthly average values (not shown).

The method improved these correlations by 35 and 21% for Terra and Aqua, respectively. The underestimation problem was corrected, and in the few cases where the AOD values were high, the predictions are now much closer to the ground-truth AERONET values.

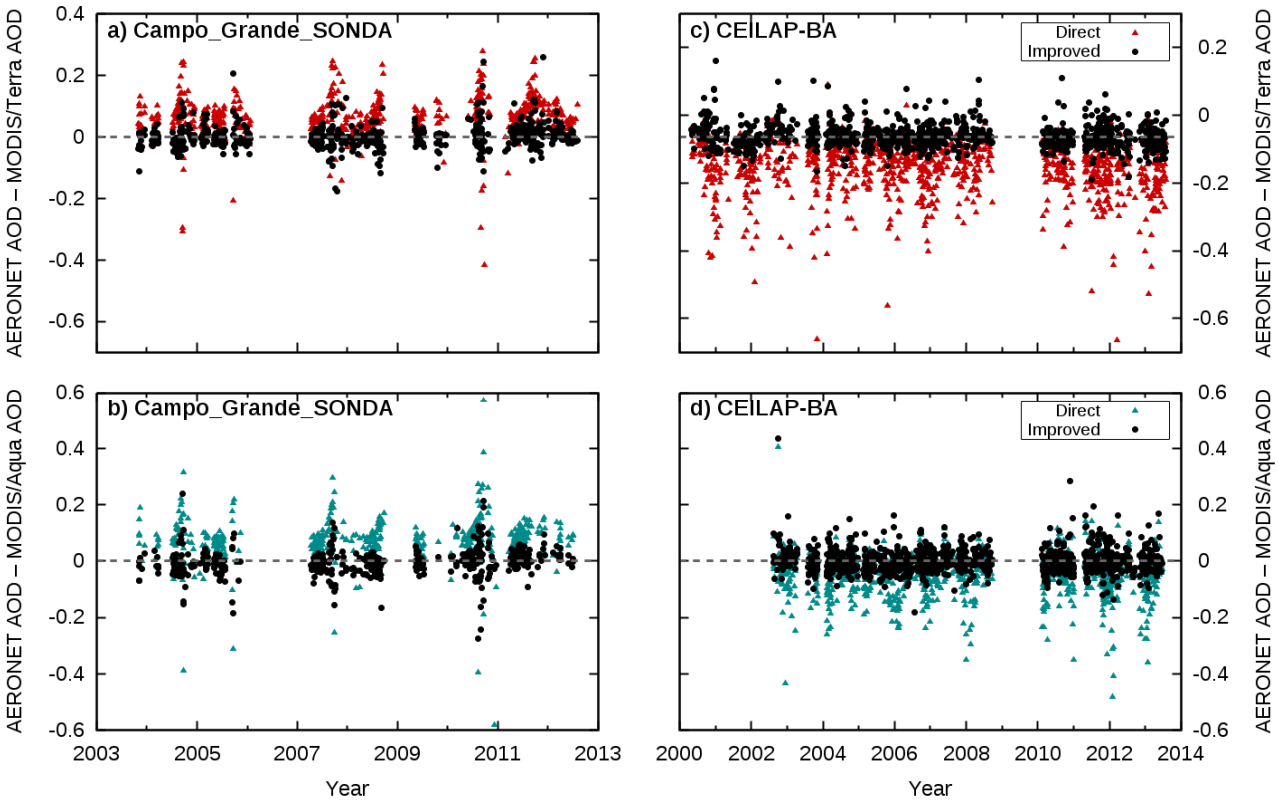

Fig. 7. Temporal series of AOD_A minus AOD_M in CEILAP-BA and Campo_Grande_SONDA stations for the direct and the improved values with the chosen method.

Arica (Arica)

The AERONET station is located in the coastal city of Arica, Chile (about 200,000 inhabitants). The city is surrounded by the Atacama Desert, and the monthly variation of the AOD (not shown here) is remarkably different from all the other stations in South America. The main source of aerosols in this site is the dust coming from the desert, identifiable even by satellite. There is no contribution of urban or biomass burning related aerosols (Mélin *et al.*, 2010). Although the site is close to the Amazon rainforest (around 400 km), its location on the west side of the Andes mountain range prevents the transport of aerosols produced by the biomass-burning activities in that region.

The yearly average AOD is 0.22 with almost no seasonal dependence. There is a small peak in August, probably related to a slight change in the weather conditions, mainly due to an increase in the relative humidity.

The direct correlation gave R^2 values of 0.71 and 0.61 for Terra and Aqua, respectively. This correlation is good for a coastal station, specially compared to Buenos Aires. The machine learning algorithms produced R^2 values of 0.76 for Terra and 0.72 for Aqua, which represent an improvement of 7% and 18%, respectively, although the slope values were further away from 1. This was the only site where the benefits of the utilization of the method over the direct correlation were only marginal.

São Paulo (Sao_Paulo)

The AERONET site in São Paulo City (20 million inhabitants in the metropolitan area) is located about 60 km away from the coast so that the satellite measurements are not influenced by the water surface as in Buenos Aires. Although both São Paulo and Buenos Aires are megacities, their monthly variation patterns of AOD are different. São Paulo shows a strong AOD peak in September, mainly related to biomass burning emission from the Amazon rainforest and the Cerrado region, located in the southeast of the Amazon (Landulfo *et al.*, 2003). In addition, fires related to agricultural activities and sugar cane harvesting inside and close to the São Paulo state during the dry season (June–October) contribute to these high AOD values (Hoelzemann *et al.*, 2009). During the rest of the year, the monthly average values of AOD are around 0.15, which is twice the value recorded in Buenos Aires (0.07). The cause of this difference may be related to the higher emissions and a stable planetary boundary layer, producing frequent thermal inversion layers, thus resulting in unfavorable conditions for the dispersion of pollutants (de Almeida Albuquerque *et al.*, 2011). The direct correlation showed a R^2 coefficient of 0.75 for Terra and 0.86 for Aqua. In the case of Terra and Aqua measurements, there was a slight, but systematic overestimation of values in the September–March period, as well as a slight, but systematic underestimation of AOD values in the May–August period. The slope of the linear fit is 0.95, but this is caused by the compensation between under and overestimated measurements. Thus, this number cannot be used as an indicator of a good estimation.

The increase in the R^2 coefficient was 15% for Terra and

7% for Aqua. More importantly, the Terra and Aqua under/over estimation was corrected.

Amazon Rainforest-Influenced Stations (Alta_Floresta, CUIABA-MIRANDA, Rio_Branco, Campo_Grande_SONDA, SANTA_CRUZ)

The Rio_Branco and Alta_Floresta stations are located in the Amazon rainforest. Instead, CUIABA-MIRANDA, SANTA_CRUZ and Campo_Grande_SONDA are located in its vicinity (about 200, 400 and 600 km, respectively). The aerosol regimes of all these stations are heavily influenced by the rainforest emissions (Hoelzemann *et al.*, 2009). The average monthly variation for these stations shows low AOD values from January to July, then a steep increase, reaching the peak in September, and finally a decrease up to the minimum value in November. The main factor controlling this pattern is the biomass burning, as there is a strong correlation between the burned area and AOD levels (Bevan *et al.*, 2009). This pattern is valid even for the more urban influenced site located in Santa Cruz de la Sierra. The AOD values recorded in these stations were the highest in South America, with many values higher than 1.

The general agreement between AOD_M and AOD_A in the region is very good, being all the R^2 higher than 0.89. Nevertheless, MODIS underestimates the lower values and overestimates the larger values, which may indicate an incorrect parameterization of particle absorption (Hyer *et al.*, 2011), although this overestimation has been partially corrected in the C6.

As an example, the results from Campo_Grande_SONDA site are analyzed in detail. In this site, MODIS underestimates AOD values lower than one. The opposite behavior is found at the CEILAP-BA station. In addition, the AOD values higher than one are sometimes overestimated, which is particularly notorious when plotting the temporal series of AOD_A minus AOD_M (Figs. 7(a)–(7b)). Both deviations were corrected by the model. The overestimation of the higher values did not perceptibly change the R^2 because the correlation between the points with lower AOD values is very good. Nevertheless, this drawback would prevent the use of the highest AOD values from MODIS in case the AERONET station is not operative.

Even though the direct correlations for the Amazon rainforest-influenced sites are very good, there is still room for some improvements. First, the results of the machine learning methods showed a small improvement in the correlation. Second, the overestimations were corrected, which resulted in a significant improvement of the regression slopes. Lastly, the fraction of data within the expected error was strongly increased, going from 62% and 57% to 94% and 91% (Terra and Aqua respectively).

Going a step further in this subject, and given the important number of AOD values higher than 1 found at these stations, a more detailed study about the overestimation problem was carried out using this particular subset of data (Figs. 8(a)–8(b)). In this case, the R^2 values increased from 0.77 to 0.95 for Terra satellite and from 0.62 to 0.80 for Aqua. The slopes of the fitted line showed values of 0.97 instead of 1.42 for Terra and 0.80 instead of 1.20 for Aqua.

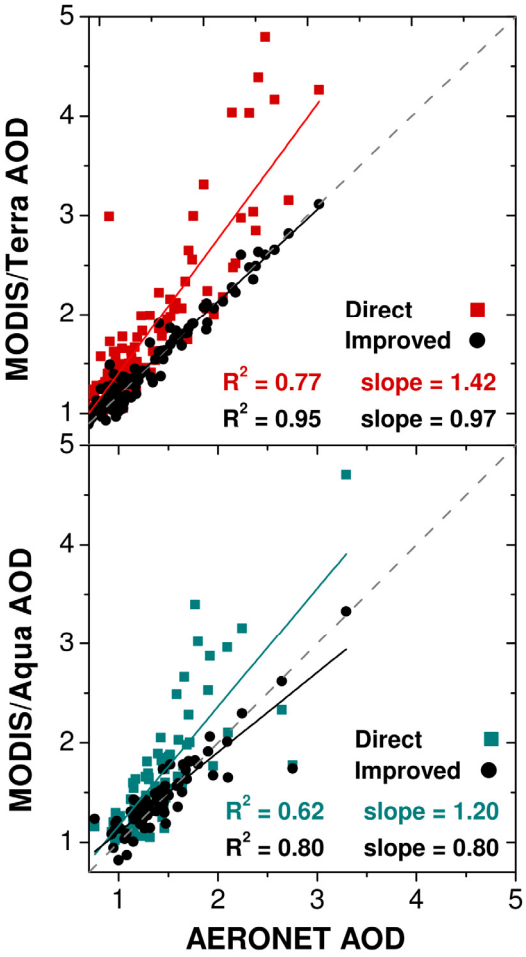

Fig. 8. AOD_M vs. AOD_A correlation (for AOD_M values larger than 1) in Amazon rainforest-influenced stations for the direct and improved correlations. Linear fit is also showed.

SUMMARY AND CONCLUSIONS

In this work, a method to correct AOD values measured by the MODIS instruments was presented. The method is based on machine learning techniques using AOD_M values and meteorological parameters as inputs. The method was applied to nine stations in South America, located in a variety of environments, showing significant improvements in the correlations between the recently released MODIS Collection 6 and AOD_A values. The improvement was particularly noticeable since it corrected the outliers of high AOD values, the underestimation of low AOD values measured at the stations in the Amazon region, and the systematic overestimation and outliers in Buenos Aires City.

In theory, SVR should be better to generalize the relationship between the pairs of inputs-outputs without memorizing them (thus being able to predict new cases). However, ANN and SVR methods performed similarly well at estimating the AOD, measured in terms of the coefficient of determination (R^2). Based on this, there is not a simple way to determine which method will produce the best results; the choice of the technique will depend on the user preference. Computationally, the search for the optimum

conditions to run the model (network topology in ANN and parameters in SVR) is time and resource consuming. However, once the conditions are found, the implementation of both methods is fast and straightforward.

The input variables required to estimate the AOD values, in addition to AOD_M, were the MDOY and the MDOY direction for most of the sites. This result reinforces the idea that the simple addition of MDOY is a reasonable way to represent the annual average variability in the weather conditions. All the other variables (temperature, RH, wind speed and wind direction) were used in approximately the same number of sites (about half of them). No clear correlation has been observed among the required meteorological variables, their ranges and which of them were used in each site. Other input variables, such as planetary boundary height, aerosol composition, etc., were not included because they are not available for all the sites of the study.

This method can also be applied to other regions of the world, especially in zones where the aerosols are poorly characterized, and thus the MODIS algorithm is less reliable. Those zones are located mostly in Africa, Asia and Australia (Shi *et al.*, 2011). The only requirement to train a model is to have an AERONET station and meteorological data.

It would be interesting to extend the applicability of the method to regions where ground-based data were never available. One possible way to accomplish that is to define an area surrounding each station in which the obtained model would still be valid. Once this area is defined, the AOD values from MODIS can be corrected for this region. Elbern *et al.* (2007) used this approach to estimate the shape of these areas for several chemical species in Europe, and Hoelzemann *et al.* (2009) used a similar approach to find the area of influence of the AERONET stations in South America during the biomass burning season. This kind of studies can be improved by using the correction of the MODIS values proposed in this work.

Estimations of $PM_{2.5}$ and PM_{10} based on satellite measurements are a topic of great relevance to the atmospheric sciences (e.g., Luo *et. al*, 2015). By using local measurements of particulate matter, and applying a similar methodology to that described in this work, it would be possible to obtain better values of $PM_{2.5}$ and PM_{10} from satellite measurements. Although the use of this method to estimate particulate matter would require other input variables measured by MODIS or by local stations, it can be useful for underdeveloped zones, where particulate matter is not routinely measured and the only sources of data are short or sparse campaigns.

ACKNOWLEDGMENTS

We thank CONICET (PIP 2013-2015 grant 1120120100004CO) and SeCyT-UNC (grant 05/C275) for partial support of the work reported here. We thank the AERONET principal investigators Brent Holben, Paulo Artaxo, and Enio B. Pereira and their staff for establishing and maintaining the South America sites used in this investigation. We thank the science and support teams of MODIS for their data. Bethania L. Lanzaco thanks CONICET for a graduate fellowship.

REFERENCES

Arkouli, M., Ulke, A.G., Endlicher, W., Baumbach, G., Schultz, E., Vogt, U., Müller, M., Dawidowski, L., Faggi, A., Wolf-Benning, U. and Scheffknecht, G. (2010). Distribution and temporal behavior of particulate matter over the urban area of Buenos Aires. *Atmos. Pollut. Res.* 1: 1–8.

Basheer, I.A. and Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *J. Microbiol. Methods* 43: 3–31.

Bevan, S.L., North, P.R.J., Grey, W.M.F., Los, S.O. and Plummer, S.E. (2009). Impact of atmospheric aerosol from biomass burning on amazon dry-season drought. *J. Geophys. Res.* 114: D09204.

Bond, T.C. and Bergstrom, R.W. (2006). Light absorption by carbonaceous particles: An investigative review. *Aerosol Sci. Technol.* 40: 27–67.

Castro Videla, F., Barnaba, F., Angelini, F., Cremades, P. and Gobbi, G.P. (2013). The relative role of Amazonian and non-Amazonian fires in building up the aerosol optical depth in South America: A five year study (2005–2009). *Atmos. Res.* 122: 298–309.

Chang, C.C. and Lin, C.J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2: 27:1–27:27.

Chu, D.A., Kaufman, Y.J., Ichoku, C., Remer, L.A., Tanré, D. and Holben, B.N. (2002). Validation of MODIS aerosol optical depth retrieval over land. *Geophys. Res. Lett.* 29: MOD2–1.

De Almeida Albuquerque, T.T., de Fátima Andrade, M. and Ynoue, R.Y. (2012). Characterization of atmospheric aerosols in the city of São Paulo, Brazil: comparisons between polluted and unpolluted periods. *Environ. Monit. Assess.* 184: 969–984.

Eck, T., Holben, B., Reid, J., Dubovik, O., Smirnov, A., O'neill, N., Slutsker, I. and Kinne, S. (1999). Wavelength dependence of the optical depth of biomass burning, urban, and desert dust aerosols. *J. Geophys. Res.* 104: 31333–31349.

Elbern, H., Strunk, A., Schmidt, H., and Talagrand, O. (2007). Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys.* 7: 3749–3769.

Estellés, V., Campanelli, M., Utrillas, M.P., Expósito, F. and Martínez-Lozano, J.A. (2012). Comparison of AERONET and SKYRAD4.2 inversion products retrieved from a Cimel CE318 sunphotometer. *Atmos. Meas. Tech*. 5: 569–579.

Giles, D.M., Holben, B.N., Eck, T.F., Sinyuk, A., Smirnov, A., Slutsker, I., Dickerson, R.R., Thompson, A.M. and Schafer, J.S. (2012). An analysis of AERONET aerosol absorption properties and classifications representative of aerosol source regions. *J. Geophys. Res.* 117: D17203.

Hirtl, M., Mantovani, S., Krüger, B.C., Triebnig, G., Flandorfer, C., Bottoni, M. and Cavicchi, M. (2014).

Improvement of air quality forecasts with satellite and ground based particulate matter observations. *Atmos. Environ*. 84: 20–27.

Hoelzemann, J.J., Longo, K.M., Fonseca, R.M., do Rosário, N.M.E., Elbern, H., Freitas, S.R. and Pires, C. (2009). Regional representativity of AERONET observation sites during the biomass burning season in South America determined by correlation studies with MODIS Aerosol Optical Depth. *J. Geophys. Res.* 114: D13301.

Holben, B., Eck, T., Slutsker, I., Tanre, D., Buis, J., Setzer, A., Vermote, E., Reagan, J., Kaufman, Y., Nakajima, T., Lavenu, F., Jankowiak, I. and Smirnov, A. (1998). AERONET - A federated instrument network and data archive for aerosol characterization. *Remote Sens. Environ.* 66: 1–16.

Hsu, N.C., Tsay, S.C., King, M.D. and Herman, J.R. (2004). Aerosol properties over bright-reflecting source regions. *IEEE Trans. Geosci. Remote Sens.* 42: 557–569.

Hyer, E.J., Reid, J.S. and Zhang, J. (2011). An over-land aerosol optical depth data set for data assimilation by filtering, correction, and aggregation of MODIS collection 5 optical depth retrievals. *Atmos. Meas. Tech*. 4: 379–408.

IPCC (2007). *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Core Writing Team, Pachauri, R.K. and Reisinger, A. (Eds.), IPCC, Geneva, Switzerland.

Kahn, R.A., Gaitley, B.J., Garay, M.J., Diner, D.J., Eck, T.F., Smirnov, A. and Holben, B.N. (2010). Multiangle imaging spectroradiometer global aerosol product assessment by comparison with the aerosol robotic network. *J. Geophys. Res.* 115: D23209.

Kaufman, Y.J., Tanré, D., Remer, L.A., Vermote, E.F., Chu, A. and Holben, B.N. (1997). Operational remote sensing of tropospheric aerosol over land from EOS moderate resolution imaging spectroradiometer. *J. Geophys. Res.* 102: 17051–17067.

Landulfo, E., Papayannis, A., Artaxo, P., Castanho, A.D.A., de Freitas, A.Z., Souza, R.F., Vieira, N.D, Jorge, M.P.M.P., Sánchez-Ccoyllo, O.R. and Moreira, D.S. (2003). Synergetic measurements of aerosols over São Paulo, Brazil using LIDAR, sunphotometer and satellite data during the dry season. *Atmos. Chem. Phys*. 3: 1523–1539.

Levy, R.C., Mattoo, S., Munchak, L.A., Remer, L.A., Sayer, A.M., Patadia, F., and Hsu, N.C. (2013). The collection 6 MODIS aerosol products over land and ocean. *Atmos. Meas. Tech.* 6: 2989–3034.

Luo, N., Wong, M.S., Zhao, W., Yan, X., and Xiao, F. (2015). Improved aerosol retrieval algorithm using Landsat images and its application for PM_{10} monitoring over urban areas. *Atmos. Res.* 153, 264–275.

Mélin, F., Clerici, M., Zibordi, G., Holben, B.N. and Smirnov, A. (2010). Validation of SeaWiFS and MODIS aerosol products with globally distributed AERONET data. *Remote Sens. Environ.* 114: 230–250.

Mishchenko, M.I., Liu, L., Geogdzhayev, I.V., Travis, L.D., Cairns, B. and Lacis, A.A. (2010). Toward unified satellite climatology of aerosol properties. 3. MODIS versus MISR versus AERONET. *J. Quant. Spectrosc. Radiat. Transfer* 111: 540–552.

More, S., Pradeep Kumar, P., Gupta, P., Devara, P.C.S. and Aher, G.R. (2013). Comparison of aerosol products retrieved from AERONET, MICROTOPS and MODIS over a tropical urban city, Pune, India. *Aerosol Air Qual. Res.* 13: 107–121.

Olcese, L.E., Palancar, G.G. and Toselli, B.M. (2014). Aerosol optical properties in central argentina. *J. Aerosol Sci*. 68: 25–37.

Olcese, L.E., Palancar, G.G. and Toselli, B.M. (2015). A Method to estimate missing AERONET AOD values based on artificial neural networks. *Atmos. Environ*. 113: 140–150.

Petrenko, M., Ichoku, C. and Leptoukh, G. (2012). Multi-sensor Aerosol Products Sampling System (MAPSS). *Atmos. Meas. Tech*. 5: 913–926.

Remer, L.A., Kleidman, R.G., Levy, R.C., Kaufman, Y.J., Tanré, D., Mattoo, S., Martins, J.V., Ichoku, C., Koren, I., Yu, H. and Holben, B.N. (2008). Global aerosol climatology from the MODIS satellite sensors. *J. Geophys. Res.* 113: D14S07.

Sayer, A.M., Munchak, L.A., Hsu, N.C., Levy, R.C., Bettenhausen, C. and Jeong, M.J. (2014). MODIS Collection 6 aerosol products: Comparison between Aqua's e-Deep Blue, Dark Target, and "merged" data sets, and usage recommendations. *J. Geophys. Res.* 119: 13965–13989.

Schölkopf, B., Smola, A.J., Williamson, R.C. and Bartlett, P.L. (2000). New support vector algorithms. *Neural Comput.* 12: 1207–1245.

Shi, Y., Zhang, J., Reid, J.S., Hyer, E.J., Eck, T.F., Holben, B.N. and Kahn, R.A. (2011). A critical examination of spatial biases between MODIS and MISR aerosol products – Application for potential AERONET deployment. *Atmos. Meas. Tech.* 4: 2823–2836.

Tripathi, S.N., Dey, S., Chandel, A., Srivastava, S., Singh, R.P. and Holben, B.N. (2005). Comparison of MODIS and AERONET Derived aerosol optical depth over the ganga basin, India. *Ann. Geophys*. 23: 1093–1101.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Zhang, J. and Reid, J.S. (2006). MODIS aerosol product analysis for data assimilation: Assessment of over-ocean level 2 aerosol optical thickness retrievals. *J. Geophys. Res.* 111: D22207.

Received for review, May 29, 2015
Revised, September 14, 2015
Accepted, November 13, 2015