Atmospheric Environment 113 (2015) 140–150

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

A method to estimate missing AERONET AOD values based on artificial neural networks

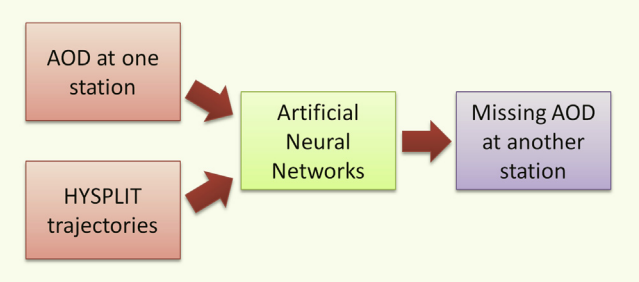
Luis E. Olcese*, Gustavo G. Palancar, Beatriz M. Toselli

Departamento de Físico Química, INFIQC, CONICET, Facultad de Ciencias Químicas, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina

HIGHLIGHTS

- Missing AODs are estimated using AOD at other station, ANN and air mass trajectories.
- The method is particularly useful for cloudy days.
- It was validated in the Eastern US and Iberian Peninsula regions.
- It is suitable for estimating other aerosol optical properties.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 November 2014
Received in revised form
5 May 2015
Accepted 8 May 2015
Available online 9 May 2015

Keywords:
AOD prediction
Eastern US region
Iberian Peninsula region
HYSPLIT

ABSTRACT

In this work, we present a method to predict missing aerosol optical depth (AOD) values at an AERONET station. The aim of the method is to fill gaps and/or to extrapolate temporal series in the station datasets, i.e. to obtain AOD values under cloudy sky conditions and in other situations where there is a temporary or permanent lack of data. To accomplish that, we used historical AOD values at two stations, air mass trajectories passing through both of them (calculated by using the HYSPLIT model) and ANN calculations to process all the information. The variables included in the neural network training were the station numbers, parameters representing the annual average trend of meteorological conditions, the number of hours and the distance traveled by the air mass between the stations, and the arrival height of the air mass. The method was firstly applied to predict AOD at 440 nm in 9 stations located in the East Coast of the US, during the years 1999–2012. The coefficient of determination r^2 between measured and calculated AOD values was 0.855, which show the good performance of the method. Besides, this result represents a remarkable improvement compared to three simple approaches. To further validate the method, we applied it to another region (Iberian Peninsula) with different characteristics (lower density of AERONET stations, different meteorology, and lower wind field spatial resolution). Although the results are still good (r^2 = 0.67), the performance of the method was affected by these characteristics. Considering the obtained results, this method can be used as a powerful tool to predict AOD values in several conditions. The methodology can also be easily adapted to predict AOD values at other wavelengths or other aerosol optical properties.

1. Introduction

Aerosols are liquid or solid particles suspended in the atmosphere. Hence, they modify the radiation balance of the Earth by

* Corresponding author. Departamento de Físico Química, INFIQC, CONICET, Facultad de Ciencias Químicas, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina.
E-mail address: lolcese@fcq.unc.edu.ar (L.E. Olcese).

http://dx.doi.org/10.1016/j.atmosenv.2015.05.009
1352-2310/

scattering and absorbing radiation, leading to two opposite effects: cooling the atmosphere by backscattering the solar radiation to space, and warming it by absorbing the terrestrial radiation (Brooks and Legrand, 2000; Lelieveld et al., 2002). The magnitude of these effects depends, primarily, on the aerosol load (usually quantified through the Aerosol Optical Depth, AOD), chemical composition and size distribution. The balance between these effects is still under discussion, with no unanimous opinion (IPCC, 2007). In addition, the uncertainties in their optical properties and size distribution at local and regional scales are one of the main factors affecting the confidence in the predictive ability of the regional and general circulation models (IPCC, 2007).

The properties of tropospheric aerosols varies widely in time and space due to the different kind of sources, their short residence time (2–7 days), their dependence on relative humidity and their variability in size, shape and chemical composition. Furthermore, aerosols are ubiquitous in the troposphere and can be transported long distances by the wind. Thus, their effects on the atmosphere show a strong temporal and spatial variability. To study the aerosols and to overcome their complexity, both direct measurements and transport models are currently used. One of main sources is the AERONET (AErosol RObotic NETwork) network (Holben et al. 1998, 2001), deployed to retrieve information about local aerosol optical properties in more than 300 sites over the globe in the year 2013 and more than 800 sites since its beginning in 1993. The other relevant source is the satellite measurements, being one of the most complete ones the Moderate-Resolution Imaging Spectroradiometer (MODIS), deployed on board the Terra and Aqua satellites (Salomonson et al., 1992). MODIS has the great advantage of covering the whole globe in a rather homogeneous manner. That is why satellite data are being increasingly used as the main source of AOD data. However, due to uncertainties introduced by surface reflectivity and clouds, aerosol optical properties retrieved from satellites must be validated against surface measurements. Consequently, in this work we used AERONET as the source of data even though its spatial coverage is limited. Regarding the aerosol transport, it can be considered by analyzing the wind fields available from regional and global databases. The most widely used tool to do that is the HYSPLIT (HYbrid Single-Particle Lagrangian Integrated Trajectory) model (Draxler and Hess, 1997, 1998), which is capable of calculating an air mass trajectory forward or backward in time. Both HYSPLIT trajectories and AERONET measurements can be very useful to study different aspects of air masses containing aerosols, but by itself, none of them allows to obtain and correlate information about the evolution of the aerosol properties along a trajectory.

Although AERONET sunphotometers are set up to measure every 15 min, it is very common to have gaps of hours, days or even weeks in the temporal series of optical properties. Beyond the data removal because of the AERONET version 2.0 constraints to assure the data quality, there are several other reasons by which the temporal series can be incomplete, including the instrument going temporarily offline or being relocated, or, mainly, due to the presence of clouds (in many cases during several days). Under these circumstances, the hourly aerosol optical properties values are not available, and the radiative transfer and air quality models that require those parameters as inputs have to be run using average values or crude estimations. Besides, these gaps can lead to a distortion in the statistics of the variables, and bias the long-term analysis and trends. Thus, the aim of this work is to predict a missing AOD value at a site by using the AOD value of a nearby station, provided the trajectory of an air mass passes through both stations. This is accomplished by using neural network calculations that, in turn, use as inputs HYSPLIT trajectories, AOD values at a nearby station and parameterized variables that represent the annual meteorological average trend. In this work, AERONET stations located in the surroundings of Washington, DC, USA (called here Eastern US region) were used as the case study. The zone was chosen because it has high-resolution meteorological databases and presents a high density of AERONET stations. In order to verify the applicability of the method to other regions with different characteristics, we used stations located in Spain, Portugal and Algeria (called the Iberian Peninsula region). This zone is characterized by a lower resolution in the wind fields, and stations sparsely located.

Although the method is used here to predict AOD values measured at 440 nm, it can be also extended to predict values at other wavelengths or different aerosol optical properties. In addition, it can be used not only to fill temporary gaps, but also to extrapolate the AOD values for a period after or before an AERONET station was operative.

2. Databases and tools

In this work, we integrated an aerosol optical properties database (AERONET), with an air mass trajectory model (HYSPLIT) and a predictive tool (Artificial Neural Networks, ANN) to calculate the AOD at a given site. All of them are described in the following sections.

2.1. AERONET

AERONET is a federated international network of radiometers, widely used to retrieve information about local aerosol optical properties (e.g. Putaud et al., 2014; Rahul et al., 2014). The network is coordinated by the NASA Goddard Space Flight Center, which maintains an historical record of over 800 automatic sun/sky CIMEL photometers worldwide. The principle of operation of the CIMEL instrument is to acquire aureole and sky radiance measurements every 15 min, considering that valid measurements are done only when the sun is visible. Sun and sky measurements are performed in seven spectral bands (340, 380, 440/441, 500, 670, 870 and 1020 nm), from which the AOD, Ångström coefficient, size distribution and single scattering albedo, among others, are derived. A detailed description of the instruments and data acquisition procedure was given by Holben et al. (1998, 2001). In this work, only Level 2.0 data were used (cloud screened and quality-assured), even though this largely reduces the number of available values. An accuracy assessment of the AERONET retrievals, as well as the algorithms used to obtain the inversion products, can be found in the work of Dubovik et al. (2000). The results presented in this work are based on the AOD values at 440/441 nm but, for simplicity, they will be referenced as AOD. Tests performed for other wavelength and for Ångström coefficient (not shown here) yields comparable results to those obtained for AOD at 440/441 nm. Thus, the method described here is suitable to be used at any wavelength or for any other aerosol optical properties.

2.1.1. AERONET sites description

The surroundings of Washington, DC are a region with a high number of AERONET stations. We selected an area of about 130 × 300 km (longitude–latitude) that includes nine stations operative in at least 3 months in any year during the period 2004–2012. This area covers the state of Delaware, Washington, DC, and the eastern portions of the states of Virginia and Maryland, totaling about 15 million inhabitants. A map including the locations of the AERONET stations is shown in Fig. 1.

Table 1 contains information about latitudes, longitudes, heights, and period of measurements of all the stations used in this work, together with the number of measurements and the average AOD value.

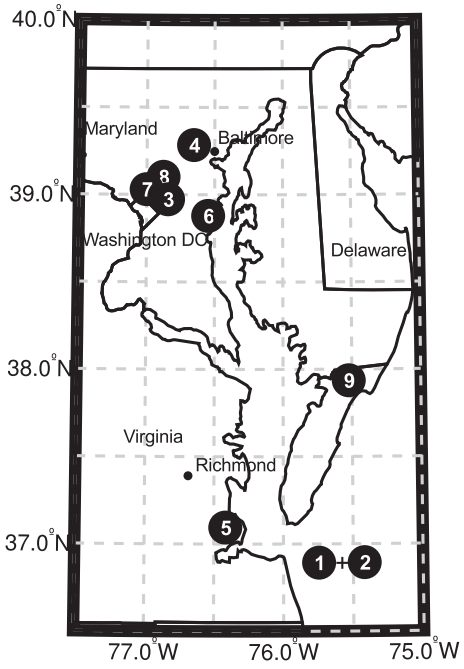


Fig. 1. Map of the Eastern US region.

Note that COVE and COVE_SEAPRISM (sites 1 and 2) are located in the same platform over the ocean and thus considered, in this work, as a single site.

2.2. HYSPLIT model

The HYSPLIT model was designed by the Bureau of Meteorology of Australia and the ARL-NOAA of USA (Draxler and Hess, 1997, 1998) and is commonly used to model a wide range of scenarios related to the regional or long-range transport, dispersion, and deposition of aerosols and air pollutants. It uses a hybrid approach, in which the calculation employs puff distribution in the horizontal direction and particle dispersion in the vertical one under a Lagrangian framework. The transport and dispersion of a parcel is calculated by assuming the release of a single puff that will expand, and then it will split into several puffs when its size exceeds the meteorological grid cell spacing (40 × 40 km in the Eastern US region, see below). The trajectory calculation is obtained by the time integration of the position of an air parcel as it is transported across the wind field. HYSPLIT can also be used to calculate the position of an air parcel by moving backward in time, ending at a receptor at a particular time. The results of the HYSPLIT analysis of forward and backward trajectories have been used in a large variety of aerosol and pollutants related studies (e.g. Valenzuela et al., 2012; Wang et al., 2011; Guo et al. 2010).

The meteorological field datasets used in HYSPLIT comes from a variety of sources, with different spatial and temporal resolutions. The one used for the Eastern US region was the Eta Data Assimilation System (EDAS) (Black, 1994) from National Centers for Environmental Prediction (NCEP), which has a 40 × 40 km Lambert Conformal Grid (185 by 129 cells) and 26 pressure levels, up to 50 hPa, covering the continental United States with a temporal resolution of 3 h. Intermediate points in the spatial and temporal grids are interpolated internally by the HYSPLIT model. As the NCEP dataset is only available for North America, for the Iberian Peninsula region we used the Global Data Assimilation System dataset (GDAS) (NOAA, 2003) from NCEP, which uses a global 1° × 1° grid (360 by 181 cells) and 12 pressure levels, up to 50 hPa, with a temporal resolution of 3 h.

2.3. Neural networks

Artificial neural networks are a powerful tool that can be used in a wide variety of complex problems, particularly in the fields of association, classification and prediction. It has been shown that neural networks can solve almost any problem more efficiently than the traditional modeling and statistical methods (Hornik, 1993). A neural network has the capacity to recognize patterns and make predictions from new data, that is to say, to generalize the observed behavior, rather than simply to memorize a given training data set. In the particular case of the atmospheric sciences, several works have been published in the last two decades (e.g. Olcese and Toselli, 2004; Miller and Emery, 1997; Fontes et al., 2014; Guo et al. 2009).

Typically, an artificial neural network is composed of a set of neurons grouped in two or more layers: one input layer with as many neurons as input variables, zero or more hidden layers with a variable number of neurons, and one output layer, with as many neurons as output variables. There are no predetermined rules to define neither how many hidden layers should be included in a neural network, nor how many neurons should contain each of these layers. Some authors indicate different numbers of hidden neurons as a function of the number of inputs, but the quantity and quality of the input data are a key factor in the final decision (Wang et al., 1993). A similar situation occurs in terms of the size of the dataset needed to obtain the best training results from the network, which is usually dependent on the complexity of the problem.

Multilayer Perceptron (MLP) is the most common and successful neural network architecture with feed-forward network topology. Each of the previously described layers uses a linear combination function to interact with the neurons in the next layer. The inputs

Table 1
Eastern US AERONET stations characteristics.

Name	#	Location	Lat; long [degrees]	Elevation [masl]	Measurement dates	Number of days with measurements	Number of hours with measurements	Average AOD
COVE	1	Chesapeake Lighthouse ocean platform, Virgina	36.90; −75.71	37	Oct-1999 – Jan-2008	1598	10,990	0.245
COVE_SEAPRISM	2	Chesapeake Lighthouse ocean platform, Virgina	36.90; −75.71	24	Apr-2005 – Dec-2012	37	275	0.248
GSFC	3	Greenbelt, Maryland	38.99; −76.84	87	May-1993 – Dec-2012	4061	27,586	0.224
MD_Science_Center	4	Baltimore, Maryland	39.28; −76.62	15	Sep-1999 – Dec-2012	2573	16,995	0.227
NASA_LaRC	5	Hampton, Virginia	37.10; −76.38	5	Nov-2004 – Dec-2012	26	151	0.097
SERC	6	Annapolis, Maryland	38.88; −76.50	5	Nov-1994 – Dec-2012	1669	11,310	0.244
USDA-BARC	7	Beltsville, Maryland	39.03; −76.93	46	Dec-2004 – May-2006	216	1274	0.174
USDA-Howard	8	Beltsville, Maryland	39.05; −76.88	52	Jul-2006 – Apr-2007	164	1035	0.196
Wallops	9	Wallops Island, Virginia	37.94; −75.48	10	Jul-1993 – Dec-2012	2038	14184	0.235

are fully connected to the hidden layer, which is fully connected to the outputs. These networks are used to create a model and, in that way, map the input to the output using historical data. These networks are called supervised networks because they need a desired output to learn (supervised training). For examples of MLP applications in the atmospheric sciences, see the work of Gardner and Dorling (1998). The most common supervised training algorithm is the so-called backpropagation (Haykin, 1994). With backpropagation, the input data are repeatedly presented to the neural network, the output is compared to the desired output, and an error is computed. This error is then fed back (backpropagated) to the neural network and used to adjust the weights such that the error decreases with each iteration and the neural model gets closer to the known output. This process is known as "training". This kind of training is relatively easy and offers good support for prediction applications. All the Neural Networks processing presented in this work have been carried out using the Neural Network Toolbox Matlab R2011B (Mathworks®).

3. Methodology

The steps to calculate a missing AOD value at an AERONET station are briefly summarized here, and described with more detail in the following subsections.

1) HYPLIT trajectories: one 120-h HYSPLIT back trajectory at each height is generated for each hour with a valid AOD measurement for every site. As they are back trajectories, these sites are labeled as ending stations (Section 3.1).
2) Trajectories selection: for each of the previously calculated back trajectories, a test is performed to check if it passes through another station (labeled as starting stations) and if it complies with the established criteria. The combinations of the different options for each criterion will lead to 18 scenarios (i.e. datasets) (Section 3.2).
3) Dataset generation: each of the 18 input/output datasets for ANN processing is generated. Each element of these sets comprises the parameters obtained from one trajectory connecting the starting and ending AERONET stations, the AOD values recorded at both stations, and parameters associated with meteorological conditions (Section 3.3).
4) Neural networks calculations: each of the 18 datasets was independently used to train 56 neural networks with different topologies. The network and the dataset with the best coefficient of determination (r^2) were chosen to calculate the missing AOD values (Section 3.4).

3.1. HYSPLIT trajectories

One 120-h HYSPLIT backward trajectory was calculated at every hour on the hour, provided there is an AOD measurement in the ±30 min range of the corresponding hour. That was done for every station in the selected years. In the case of several values reported during that interval, only the nearest one to the hour on the hour was selected.

Different air masses, with different trajectories, can reach the same point at the same time, but at different heights. Nevertheless, aerosol optical depth is defined as the integrated extinction coefficient over a vertical column of unit cross section. Thus, in order to account for the aerosol masses that can be present at different heights, the previously described backward trajectories have also been calculated starting at different heights above ground level (0, 500, and 1000 m).

3.2. Trajectories selection

The next step was to determine if each of the previously calculated trajectories pass through another AERONET station. In order to do that, it is necessary to define the maximum horizontal distance, between the measurement site and the trajectory path, which will be considered as acceptable. A large distance could result in large uncertainties while a short distance will lead to a small number of trajectories in the dataset. Thus, three different maximum horizontal distances have been considered as options (5, 10, and 15 km).

Similarly, it is necessary to define a minimum travel time to consider a trajectory as valid. For simplicity, in this method the starting and ending points of a backward trajectory are set at the hour on the hour (the closest one to the measurement value). Consequently, for a 1-h trajectory (the minimum travel time in this method), the time difference between the measurements at the two AERONET stations could correspond to a real time between 1 and 119 min, what certainly could lead to a large relative error in the travel time parameter. On the other hand, longer minimum times will reduce the number of eligible trajectories as many stations are relatively close together. Thus, we considered as options for the minimum travel time 1, 2, and 3 h.

Regarding the height of the ending point of the trajectory we considered two options: one including only backward trajectories starting at ground level and the other one considering all the three different heights (0, 500, and 1000 m above ground level).

Once established that a trajectory is eligible according to the previously described criteria, it is checked if there is an AOD measurement for the starting station at the air mass arrival time (again, in a ±30 min range). If this is the case, the date, the AERONET stations numbers, the trajectory arrival altitude, the AOD values at both stations, the travel time, and the distance traveled by the air mass (not the linear distance between both sites) are recorded, constituting one element of a dataset, which will feed the ANN.

The combination among all the options allowed for the three mentioned criteria (three maximum site-trajectory distances, three minimum travel times, and two height options at the ending point) lead to a total of 18 scenarios (Table 2). The datasets generated for each of these cases have been used in the ANN training to determine the best option for each criterion (see Section 3.4).

3.3. Dataset generation

Certainly, the meteorological variables (mainly wind speed and direction, temperature, relative humidity and rain/snow episodes) have a large effect on the measured AOD values and its temporal variation (Kim et al., 2007). However, because aerosols can travel a long time and under different weather conditions (some trajectories span over several days) it is very difficult to appropriately consider these effects. One way to get through this difficulty in the ANN calculations is using the Julian Day (JD), as it is reasonable to assume that those parameters have similar variations along each year. The mayor drawback of using the JD is that although January 1st and December 31st have, statistically, the same weather conditions, their JDs are in the opposite ends of the scale (i.e. 1 and 365). To overcome this fact, a Modified Julian Day (MJD) and a Modified Season (MS) are defined, based on the proximity of a given day to the middle days of summer and winter (in the Northern Hemisphere). Considering these days as the typical weather conditions for summer and winter, respectively, the midsummer day (Julian day 217) will be represented by the maximum MJD value (91), whereas the midwinter day (Julian day 35) will have the minimum MJD value (−91). The variation of the MJD values in between those numbers will represent the variation

Table 2
Scenarios used for the ANN calculation.

Heights [m.a.g.l.]	Site radius [km]	Minimum time [h]	Number of cases	Training r^2	Validation r^2	Testing r^2	All r^2
0	5	1	1260	0.885	0.762	0.729	0.845
0	5	2	946	0.762	0.716	0.504	0.711
0	5	3	728	0.904	0.563	0.590	0.787
0	10	1	5097	0.885	0.815	0.787	0.856
0	10	2	3883	0.709	0.642	0.598	0.681
0	10	3	2964	0.712	0.646	0.607	0.684
0	15	1	11,210	0.667	0.621	0.658	0.659
0	15	2	8682	0.692	0.646	0.667	0.681
0	15	3	6678	0.857	0.766	0.769	0.830
0/500/1000	5	1	3015	0.618	0.598	0.527	0.601
0/500/1000	5	2	2056	0.924	0.787	0.704	0.861
0/500/1000	5	3	1473	0.776	0.746	0.716	0.760
0/500/1000	10	1	11,964	0.848	0.805	0.783	0.832
0/500/1000	10	2	8352	0.845	0.787	0.819	0.830
0/500/1000	10	3	5952	0.846	0.790	0.692	0.814
0/500/1000	15	1	26,157	0.870	0.848	0.837	0.863
0/500/1000	**15**	**2**	**18,567**	**0.867**	**0.828**	**0.835**	**0.856**
0/500/1000	15	3	13,255	0.828	0.824	0.781	0.819

Different scenarios used to find the best conditions for AOD calculation, together with the ANN coefficients of determination for the different groups. Bolded line indicates the chosen scenario.

in the weather conditions for the transition from summer to winter and vice versa. In addition, MS will be equal to +1 in the days when the MJD increases, and equal to −1 when it decreases (Fig. 2). The reason behind this choice is that it is expected that the MJD and the MS be correlated to the meteorology along the year. Thus, in this method, MJD and MS are used to parameterize the effects of the meteorology on the AOD calculation. Clearly, these variables do not represent the short-term (weekly, monthly, or even seasonal) variations of the meteorology, but only its annual average trend. They can be calculated as:

$$\begin{array}{lll} \text{if } JD \leq 35 & : & MJD = -56 - JD \quad MS = -1 \\ \text{if } JD > 35 \text{ and } JD \leq 217 & : & MJD = -126 + JD \quad MS = +1 \\ \text{if } JD > 217 & : & MJD = 309 - JD \quad MS = -1 \end{array}$$

As the stations are located on different environments, and the different land uses and topography under the path of the air mass have direct influence on the AOD value, a number (1–9) was assigned to each station and these numbers have also been included as variables in the input dataset. Assuming that there is a more frequent path between two stations (given by the dominant wind fields in the region), it is reasonable to associate the effects of the topography and the land use with the numbers of the two stations connected by a trajectory (i.e. 56 possible combinations, as stations 1 and 2 are grouped).

Thus, a total of nine inputs and one output variable were used for the training purpose. Table 3 shows the summary and statistical information about the variables.

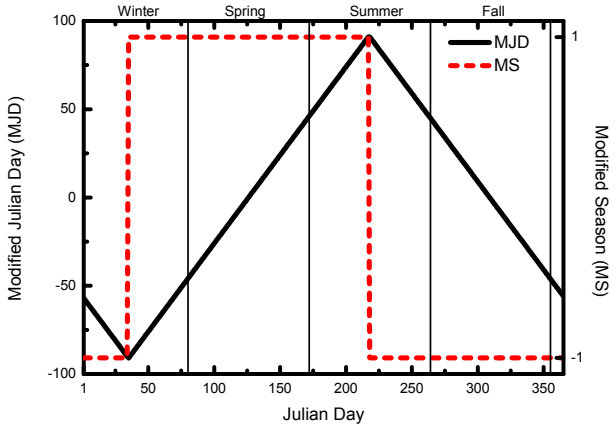


Fig. 2. MJD and MS values as a function of Julian Day.

3.4. Neural networks calculations

In this work, the used neural network architecture was the feed forward, multi-layer perceptron (MLP), widely considered as capable to approximate any function, given an acceptable number of data and the right number of neurons. The dataset was randomly divided into three subsets: training, validation and testing. Following Perez and Reyes (2002), the training of the neural networks was conducted using 70% of the data set, leaving 15% for the validation and 15% for testing purposes. All the input and output variables were separately normalized between −1 and +1. A distribution frequency comparison among all subsets has been carried out to ensure their representativeness. The validation set was used to determine the end point for the training process, in order to improve their generalization ability. The independent testing set was ignored in all steps of the training process and was solely used for the statistical comparison of the obtained results against unseen data.

The results obtained through neural network simulations depend on the network topology and activation function. In order to identify the configuration that gives the best results, different number of hidden layers and neurons have been used, as well as different kind of activation functions. A total of 56 different

Table 3
Variables used in the ANN calculation.

Variable	Type	Range of values
Ending station	Input	1–9
Starting station	Input	1–9
Trajectory altitude at ending station	Input	0, 500 and 1000 m
MJD at ending station	Input	−91–92
MS at ending station	Input	−1/1
Hours elapsed between stations	Input	1–120 h
Distance covered between stations	Input	<5–4523 km
AOD recorded at ending station	Input	0.01–2.02
AOD recorded at starting station	Output	0.01–2.12

combinations have been tested. The parameters that were modified and its possible values are:

- Hidden layers: one or two.
- Activation functions: Tan-Sigmoid or Linear, but in each layer all the activation functions have to be the same.
- Number of neurons in the hidden layer(s): from 4 to 60 total neurons, distributed in one or two layers.

For each of the 18 scenarios described in Section 3.2, an ANN was trained for each of the 56 network topologies. After training, the average coefficient of determination (calculated vs. measured AOD) for the testing dataset of each one of the 56 networks was used to choose the best configuration, which resulted in two hidden layers (the first one with 40 neurons and the second one with 6) and all the activation functions set to Tan-Sigmoid.

4. Results

In Section 4.1, we present an analysis of the direct correlation between AERONET AOD measurements at two stations and another correlation incorporating HYSPLIT trajectories. Then, in Section 4.2, we show the improvement reached by using the ANN simulation in the Eastern US region. Finally, in Section 4.3, we validate the method by applying it to the Iberian Peninsula region.

4.1. Direct comparison

One of the frequently used approaches to estimate a missing daily AOD value is to use the average value (AOD_{av}) for this JD calculated by using all the other years with measurements. To assess the errors associated to this approach, we compared the AOD value for every day and every year against AOD_{av}. As an example, the results of this approach for station GSFC are shown in Fig. 3. Fig. 3a shows the absolute AOD difference between every day for all the years and the corresponding AOD_{av}. Here, it can be observed that the usage of this approach leads frequently to absolute errors of up to ±0.7 units. These errors are especially noticeable during the months of July and August, which shows a seasonal variation probably associated to the different weather conditions on different years. Fig. 3b shows the relationship between AOD_{av} and the real AOD value for every day and every year. The diagonal bands are a result of the way of calculating the average values. The r^2 for the linear fit is 0.29. Similar plots are observed for all the other stations, with r^2 values even lower (not shown). Overall, the results show that this approach is not a reliable way to estimate missing AOD values.

Another simple way to estimate a missing AOD value at a given station is by using simultaneous historical records at this and at another station to obtain a correlation function between them.

The 2D histogram for all the pairs of AOD values (77,914) that were simultaneously recorded (in a ±30 min interval) at any two stations is shown in Fig. 4. This correlation gives a linear coefficient of determination equal to 0.69. Although this could be considered a reliable fit, note that this correlation is mainly driven by the relationship between nearby stations, as it is shown in Fig. 5. Each of the r^2 values in this figure corresponds to the linear fit for all the data of a given pair of stations, provided there are more than 10 values. As seen, the data fit an exponential decay function ($r^2 = 0.67$), showing that beyond about 200 km the r^2 values decrease rapidly, making the approach unreliable for stations separated by relatively long distances. Here, it is important to remember that the Eastern US region has an unusually large density of AERONET stations. In most regions of the world, the usual distance between AERONET stations is of several hundred

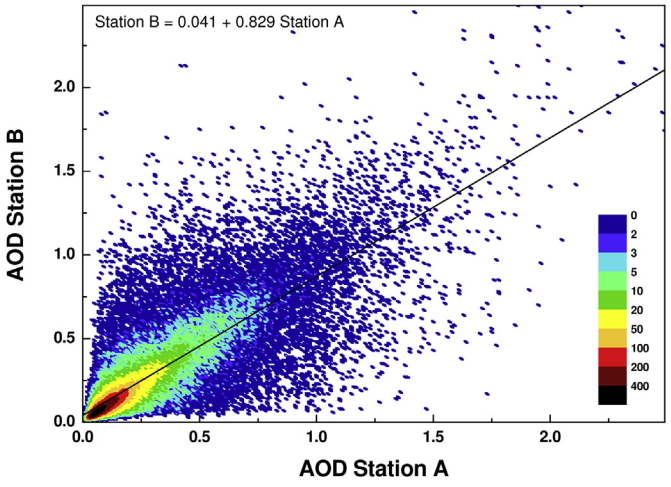


Fig. 4. Relationship between AOD values measured at two AERONET stations at the same time and its linear fit.

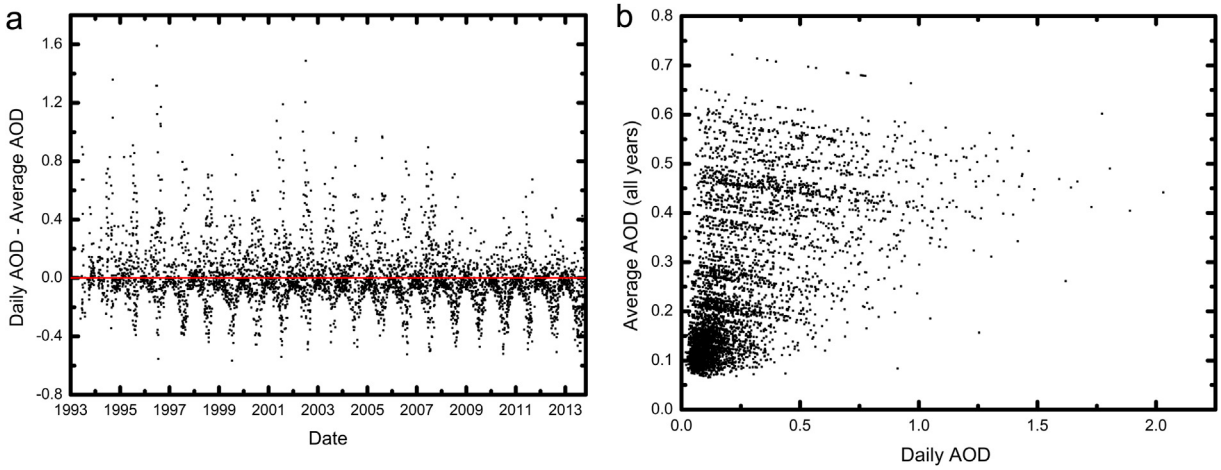


Fig. 3. a) Absolute differences between every daily AOD value and the AOD average of the corresponding Julian Day of all the other years. b) Correlation between daily AOD values and the average AOD value for the same Julian Day measured in all the other years. Both plots correspond to GSFC AERONET station (Station 3).

kilometers. On the other hand, it is reasonable to assume that if it is cloudy at one station, which is one of the main reasons for the lack of data in AERONET stations, the same condition will prevail at all the nearby stations. Thus, we conclude that, if the cause of the lack of data is a cloudy weather, this approach is not applicable even if the stations are very close to each other.

Another approach to estimate AOD values is to assume that the AOD measured at a station is a consequence, at least in part, of the aerosol transport. Thus, a possible way to estimate AOD values is to correlate the ΔAOD between two stations and the elapsed time used by the air mass to travel between them calculated through HYSPLIT back trajectories starting at 0, 500 and 1000 m (trajectories were calculated using 1 h as minimum air mass travel time and 15 km minimum distance). Fig. 6 shows the frequency distribution of the ΔAOD vs. time plot. The AOD value at a site is determined by both the local sources and the transport. The fact of using a ΔAOD, instead of the absolute AOD value, takes also into account (statistically) not only the local sources and the emissions but also the dispersion and the deposition along the trajectory. The significantly lower amount of data at 12, 36, 60, etc. hours are due to the almost absence of trajectories with AERONET measurements at those intervals. At first sight, these values cannot be fitted to any function. The data only show a slight increase in the ΔAOD as a function of time at the same time that they become sparser. In order to quantify this trend, a linear fit was done, showing a low coefficient of determination of $r^2 = 0.13$. Note that, by far, the largest concentration of point is located for trajectories of 1 h and ΔAOD equal to zero. This result clearly means that this way of estimating AOD values based on the travel time of the air mass trajectories is not reliable.

Despite this poor result to predict AOD values, some interesting aspects arise from the procedure followed to build this figure. As an example, Fig. 7 shows the number of trajectories from and to Site 3, as this is the location with the largest number of trajectories (12,125, including 6882 from and 5243 to the site). Fig. 8 shows the average ΔAOD between Station 3 and all the other stations. The numbers on top of each box represent the average ΔAOD for air masses going from Station *n* to Station 3, and the bottom numbers are the average ΔAOD for air masses starting at Station 3 and arriving to Station *n*. No data are shown for station 5, as the number of trajectories is not enough to obtain a reliable statistic (See Fig. 7). In all the cases shown in Fig. 8, the average variation of the AOD is

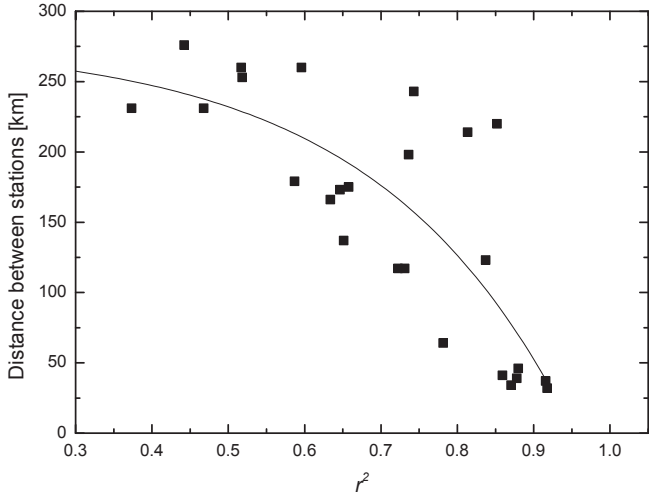


Fig. 5. Correlation between distance and coefficient of determination for all AOD values measured simultaneously in the pairs of stations.

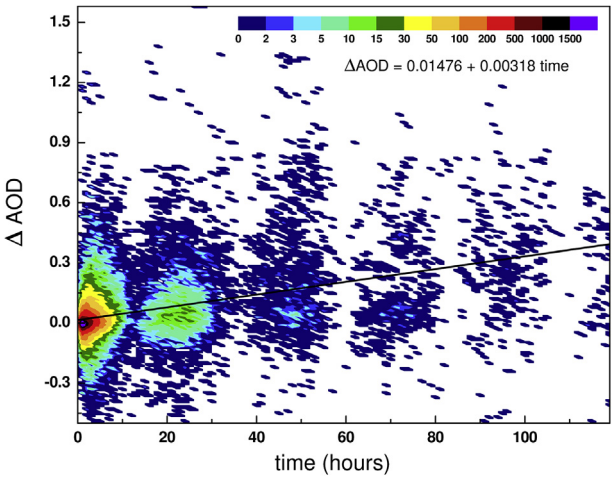


Fig. 6. Scatter plot of all the observed ΔAOD as a function of the time used for the air mass to travel between two stations.

positive, meaning that the aerosol loading increases over time. The largest variations are for air masses ending in sites 1 + 2 and 9, a result that could have many and different causes. Thus, for the sake of brevity, a detailed analysis will not be performed here.

It should be noted that other approaches in spatio-temporal modeling for filling missing values, have been proposed (e.g. Liang and Kumar, 2013; Oleson et al., 2012). However, they will not be applied here.

4.2. ANN results

Being the results of the two previously described approaches questionable to predict AOD values, the method combining AERONET measurements, HYSPLIT trajectories and ANNs was applied. In Section 3.2, 18 scenarios were presented, representing different options of maximum site-trajectory distances, minimum travel times, and height options at the ending station. As usual in ANNs, the results that must be analyzed are the r^2 values for the testing dataset. In addition, to select the best scenario to apply the method, it is necessary to take into account not only the value of the coefficient of determination but also the number of elements in the dataset. Thus, the chosen scenario includes trajectories arriving at the three heights, a radius of 15 km around the site, and a minimum

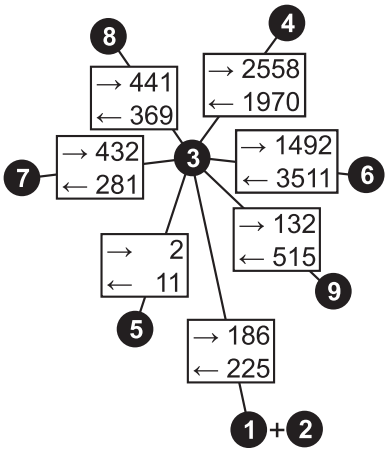


Fig. 7. Number of trajectories from Site 3 to each site (upper number) and from each site to Site 3 (lower number). The distances between sites are not in scale.

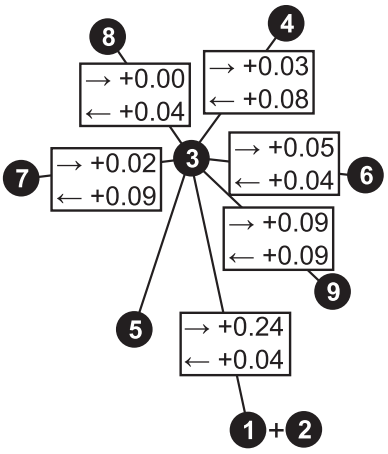


Fig. 8. Average ΔAOD for trajectories from Site 3 to each site (upper number) and from each site to Site 3 (lower number). The distances between sites are not in scale.

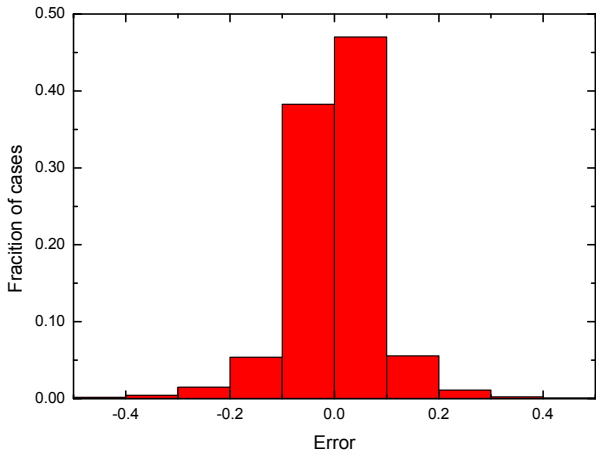


Fig. 10. Histogram of residuals (predicted minus calculated AOD values) for all the points in the Eastern US dataset.

of 2 h of travel time between stations (Table 2). This dataset scenario included 18,567 cases and will be used, from now on, to illustrate the application of the proposed method. Fig. 9 shows the predicted vs. measured AOD values for each of the three groups in which the dataset was divided (training, validation, and testing) plus their r^2 coefficients, slopes and interception values and their associate errors. The coefficient of determination for the whole dataset was 0.855. This value is significantly better than the one calculated by using a simple correlation between the AOD measurements recorded simultaneously at two stations (Fig. 4). Besides, unlike the first approach, only a few points deviate significantly from the fitting line, mostly cases where the neural network predicts a small negative instead of a small but positive AOD value. Many of these few predicted negative values have been observed to be associated to long trajectories (more than 4 days). The cause of these negative values has not been fully analyzed in this work, but they might be related to the inherent uncertainties of the method. Fig. 10 shows a histogram of the residuals of the prediction, calculated as the predicted minus the measured value for all the points. This plot shows that in more than 85% of the cases, the predicted AOD values fall within the ±0.1 range of the measured ones. For the whole dataset, the average relative error, compared to the measured AOD values, was 25% (with 45% of the values having a relative error of less than 10%).

In order to assess the effect of the meteorology in the ANN prediction, a calculation using the same dataset but without using the meteorological-related variables (MJD and MS) was performed. The result for this calculation is that the r^2 coefficient for the whole dataset and for the same scenario decreased from 0.855 to 0.611 (note that, as two variables have been removed, a new best neural network topology had to be selected for this case). This result

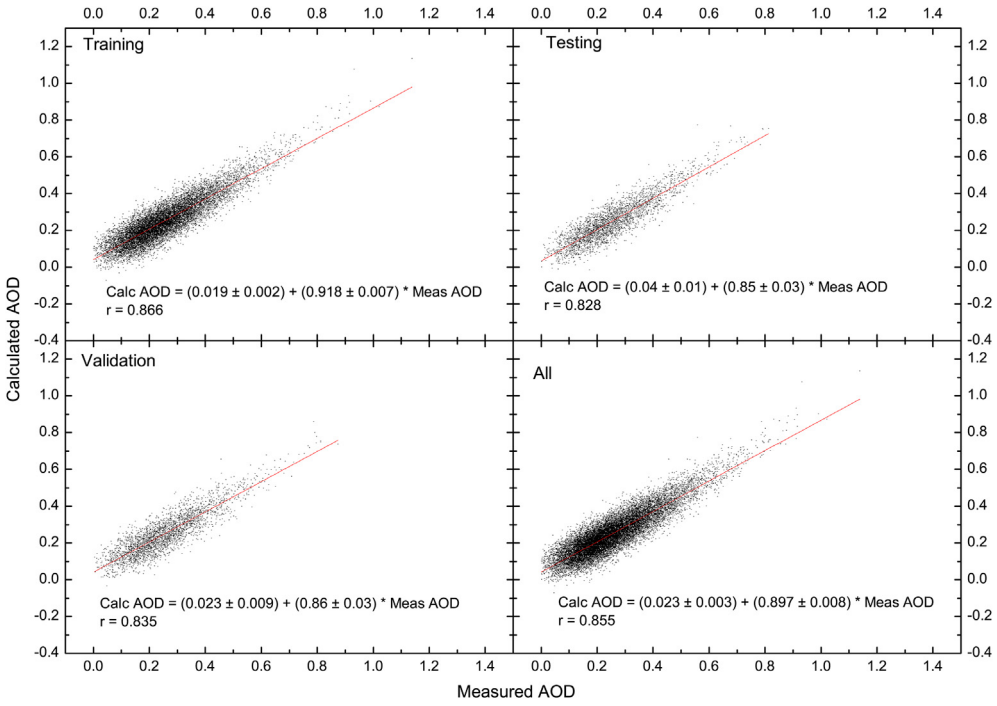


Fig. 9. Scatter plot of calculated vs. measured AOD for the different groups in which the dataset was divided.

reinforces the importance of taking into account the seasonal variability of the meteorological variables, even though it is incorporated by using two simple parameters.

Similarly, in order to test the effects of including actual meteorological data, a calculation using temperature and relative humidity at the ending stations as input variables (instead of MJD and MS) has been performed. A new ANN topology was selected and the calculations were performed using the same procedure. The r^2 of the new calculation was 0.882, compared to 0.855 of the calculation using MJD and MS. Besides, due to the lack of meteorological data, the number of cases decreased from 18,567 to 16,321. Thus, as the improvement in the r^2 was only 3%, with the added complexity in the meteorological data retrieval and the loss of data, we consider that, in this case, the usage of actual meteorological variables is not justified.

Finally, we identified a gap in an AOD dataset and used the method to predict the missing AOD values. The selected period spans over September and October 2007 at the Wallops station (Station 9). Fig. 11 shows the daily-averaged predicted values, together with the values for the same dates measured for the years 1993–2012. Values for August and November 2007 are also included. The average value for the two-month period for all the years was (0.2 ± 0.2), while the predicted average value was (0.3 ± 0.1). As seen, the predicted daily values fall within the same range for the other years. Besides, the behavior showed by the predicted points follows the general trend found by using the previous and next months. A linear fit for the predicted values has approximately the same slope than the trend for the four months shown.

4.3. Iberian Peninsula region

The analysis presented in the previous section showed that the method yields good results for the Eastern US region. However, very few regions in the world have a similar density of AERONET stations and a meteorological HYSPLIT database with the same geographic resolution. This could potentially lead to lower coefficients of determination, because HYSPLIT trajectories will have larger errors and because fewer trajectories will pass through two sites. In order to verify the validity of the proposed method, we applied it to another region with different characteristics in terms of domain size, geography and density of AERONET stations, the Iberian Peninsula.

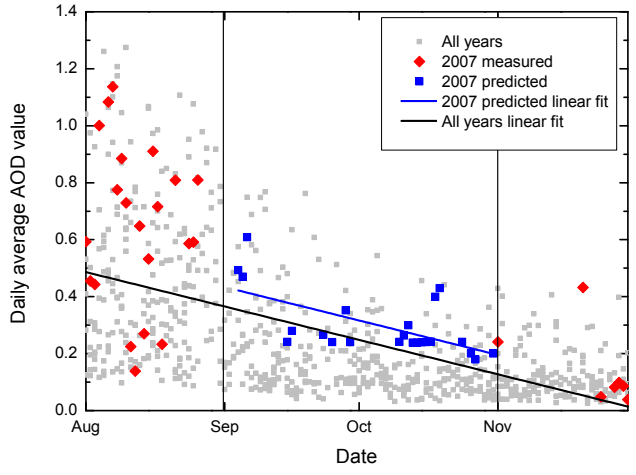


Fig. 11. Daily AOD values for the period 1993–2012 at Wallops AERONET station (Station 9), together with the predicted values for the months of September and October 2007 with their corresponding linear trends.

Similarly to the procedure followed for the Eastern US region, a simple comparison between values simultaneously recorded at two stations was performed first (Section 4.1). Then, the calculations using ANN were carried out and the results were compared.

The zone under study (from now on Iberian Peninsula region) includes stations operative in the period 2003–2012 in Spain (9 stations), Portugal (2 stations), and the northern portion of Algeria (1 station), as shown in Fig. 12. AERONET station locations and AOD measurement description are summarized in Table 4. This region has a lower resolution in the wind fields and the stations are more sparsely located than in the previous case. It spans on 15° × 8° (about 1300 × 900 km) and the approximate population is 100 million inhabitants. The meteorological database used was the Global Data Assimilation System (GDAS) (NOAA, 2003) from National Centers for Environmental Prediction (NCEP), which uses a global 1° × 1° grid (360 by 181 cells) and 12 pressure levels, up to 50 hPa, with a temporal resolution of 3 h.

The simple relationship between AOD values simultaneously recorded at two stations has a $r^2 = 0.106$ (53,351 values, plot not shown). One of the main reasons for this poor correlation is the large separation between the stations, confirming the results observed in Fig. 5.

To further verify the prediction ability of the proposed method, we carried out the calculations using ANN, following the same procedure described in Section 3. In this case, the selected ANN topology was two hidden layers (the first one with 10 neurons, the second one with 30 neurons) and all the activation functions set to Tan-Sigmoid. The chosen scenario included trajectories arriving at the three heights, a radius of 15 km around the site, and a minimum of 2 h of travel time between stations (i.e. same scenario as for Eastern US), resulting in a dataset with 9830 elements. The coefficients of determination for the three groups in which the dataset was divided (training, validation, and testing) were (0.73 ± 0.01), (0.66 ± 0.02) and (0.61 ± 0.02), respectively. The coefficient of determination for the whole dataset was (0.67 ± 0.01) with 77% of the data falling in the ±0.1 range of the measured AOD. The average relative error, compared to the measured AOD values, was 33% (with 56% of the values having a relative error of less than 10%).

These results represent a substantial improvement over the simple AOD correlation. In addition, although this r^2 value is lower than the one obtained for the Eastern US region, is still good enough to make reasonable AOD estimations. The reason of the difference in the coefficients of determination between the two regions is both the lower spatial resolution of the meteorological wind field

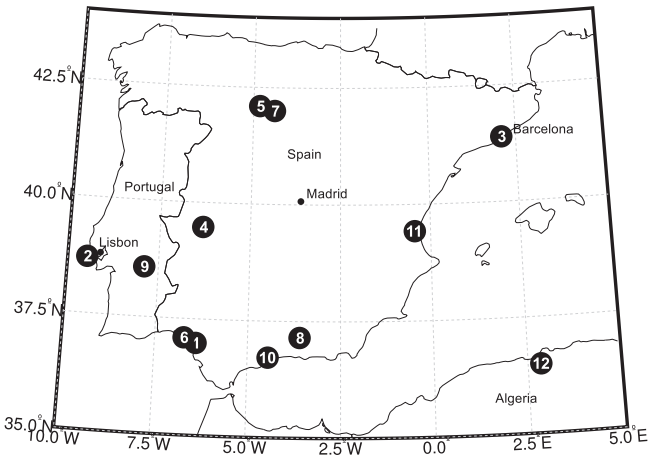


Fig. 12. Map of the Iberian Peninsula region.

Table 4
Iberian Peninsula AERONET stations characteristics.

Name	#	Location	Lat; long [degrees]	Elevation [masl]	Measurement dates	Number of days with measurements	Number of hours with measurements	Average AOD
Huelva	1	Huelva, Spain	37.02; −6.57	25	Mar-2010 – Dec-2012	223	1961	0.148
Cabo_da_Roca	2	Cabo da Roca, Portugal	38.78; −9.50	140	Dec-2003 – Dec-2012	1052	7117	0.165
Barcelona	3	Barcelona, Spain	41.39; 2.12	125	Dec-2004 – Dec-2012	1854	13,558	0.208
Caceres	4	Cáceres, Spain	39.48; −6.34	397	Jul-2005 – Jun-2012	1421	11,450	0.129
Palencia	5	Palencia, Spain	41.99; −4.51	750	Jan-2003 – Dec-2012	1670	12,607	0.141
El_Arenosillo	6	El Arenosillo, Spain	37.11; −6.73	0	Feb-2000 – Mar-2010	2404	19,681	0.167
Autilla	7	Autilla del Pino, Spain	42.00; −4.60	873	Sep-2007 – Dec-2012	384	2887	0.114
Granada	8	Granada, Spain	37.16; −3.61	680	Dec-2004 – Dec-2012	1493	12,780	0.179
Evora	9	Evora, Portugal	38.57; −7.91	293	Jul-2005 – Dec-2012	2201	18,054	0.145
Malaga	10	Malaga, Spain	36.72; −4.47	40	Feb-2009 – Dec-2012	1133	9528	0.171
Burjassot	11	Burjassot, Spain	39.51; −0.42	30	Apr-2007 – Dec-2012	1482	12,034	0.176
Blida	12	Blida, Algeria	36.50; 2.88	230	Oct-2003 – Mar-2012	1644	11,726	0.252

data and the lower density of AERONET stations (which results in larger uncertainties in the air mass trajectory calculations).

4.4. Uncertainties

Beyond the high coefficient of determination obtained for the AOD prediction using this method, here it is important to identify the sources of uncertainties, which affect the results.

- AERONET measurements: AOD measurements are affected by errors about ±0.01. In addition, AOD values lower than 0.3 may have increasingly larger errors (Dubovik et al., 2000).
- HYSPLIT trajectories: trajectory models are subject to uncertainties arising from the interpolations of sparse meteorological data, assumptions regarding to vertical transport, observational errors, sub-grid-scale phenomenon, turbulence, convection, evaporation and condensation (Polissar et al., 1999).
- ANN configuration: given the inherent limitations of the neural networks (to reach a local instead of a global minimum, infinite possible topologies, etc.), the chosen configuration might not be the best possible one (Stanley and Miikkulainen, 2002).
- Method assumptions:
  - Trajectory distance to the starting station: considering that a trajectory passes through an AERONET site if the distance is less than 15 km (Section 3.2) we are assuming that the AOD value measured at the station will be the same at 15 km from it. The validity of this assumption will depend on the local sources and transport in the region.
  - Trajectory heights at the ending station: for simplicity, only three heights were considered for the trajectory ending point. The intermediate heights are assumed to be represented by these three heights.
  - Simplified meteorology: by using a parameterization to consider the meteorology (MJD and MS), we are neglecting the influence of the actual meteorological conditions on the AOD values not only at both stations but also during the trajectory of the air mass.
  - Measurement-trajectory time differences: the difference between the time of the AERONET measurement and the time of the arrival/starting of the trajectory could be up to 30 min. Adding the differences at both the arrival and starting points, this could lead to a maximum absolute error of up to 59 min, which leads to a large uncertainty for the shortest trajectories. However, note that the duration of most of the trajectories (Fig. 6) makes this error negligible.

5. Summary and conclusions

The aim of this work was to predict a missing AOD value at an AERONET station. To accomplish this we firstly assessed the applicability of three simple approaches, finding that none of them produced good results under all conditions. To overcome these limitations, we proposed a method based on the historical AOD values at two given stations, air mass trajectories passing through both stations (calculated by using the HYSPLIT model) and ANN calculations to relate/process/link all the information. After carefully selecting the ANN topology and the best scenario (minimum time of a trajectory, maximum distance between the trajectory and the AERONET site, and height(s) of the trajectory) the method was applied to two regions with different characteristics (density of AERONET stations, meteorology, and wind fields spatial resolution): Eastern US and Iberian Peninsula.

The method demonstrated to be robust yielding r^2 values of 0.855 and 0.67 for the Eastern US and Iberian Peninsula cases, respectively, and showing a remarkable improvement over all simple approaches. Although the average error for the Eastern US region is 25%, almost half of the data showed relative errors less than 10%. The application of the method to predict the AOD values for a two-month gap in an AERONET station dataset showed consistent results in terms of both the absolute values and trend along the year.

This method also showed the importance of including a representation of the meteorology, even in a simple way (here, as a parameterization based on the Julian Day).

Although in this work we used AOD measured at 440 nm as the predicted variable, the method is applicable to the prediction of AOD at other wavelengths or to other aerosol optical properties (Ångström coefficient, asymmetry factor, etc.). In addition, even though this method was thought to predict missing AOD values (filling gaps or extrapolating data series) using backtrajectories, the availability of HYSPLIT forecast trajectories gives the possibility of using it to forecast AOD values (using forward trajectories), which would constitute a valuable tool to feed air quality or radiative transfer models.

Acknowledgments

We thank CONICET (11220120100004CO), FONCYT (Préstamo BID PICT 309) and SeCyT (UNC) for partial support of the work reported here.

We thank the principal investigators of the AERONET network and the NASA AERONET groupware is appreciated for establishing and maintaining the 21 sites used in this work/research.

References

Black, T.L., 1994. The new NMC mesoscale Eta model: description and forecast examples. Weather Forecast. 9, 265–278.

Brooks, N., Legrand, M., 2000. Dust variability and rainfall in the Sahel. In: McLaren, S., Kniveton, D. (Eds.), Linking Climate Change to Land-surface Change. Kluwer Academic Publishers, pp. 1–25.

Draxler, R.R., Hess, G.D., December, 1997. Description of the HYSPLIT 4 Modelling System, NOAA Technical Memorandum ERL ARL-224. Available from:. National Technical Information Service, 5285 Port Royal Road, Springfield,VA 22161]. Web address. http://www.arl.noaa.gov/ready/hysplit4.html.

Draxler, R.R., Hess, G.D., 1998. An overview of the HYSPLIT_4 modeling system of trajectories, dispersion, and deposition. Aust. Meteorol. Mag. 47, 295–308.

Dubovik, O., Smirnov, A., Holben, B.N., King, M.D., Kaufman, Y.J., Eck, T.F., Slutsker, I., 2000. Accuracy assessments of aerosol optical properties retrieved from AERONET sun and sky-radiance measurements. J. Geophys. Res. 105, 9791–9806.

Fontes, T., Silva, L.M., Silva, M.P., Barros, N., Carvalho, A.C., 2014. Can artificial neural networks be used to predict the origin of ozone episodes? Sci. Total Environ. 1 (488–489), 197–207.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multi-layer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. 32, 2627–2636.

Guo, J.-P., Zhang, X.-Y., Che, H.-Z., Gong, S.-L., An, X., Cao, C.-X., Guang, J., Zhang, H., Wang, Y.-Q., Zhang, X.-C., Xue, M., Li, X.-W., 2009. Correlation between PM concentrations and aerosol optical depth in eastern China. Atmos. Environ. 43, 5876–5886.

Guo, J., Zhang, X., Cao, C., Che, H., Liu, H., Gupta, P., Zhang, H., Xu, M., Li, X., 2010. Monitoring haze episodes over the Yellow Sea by combining multisensor measurements. Int. J. Remote Sens. 31, 4743–4755.

Haykin, S., 1994. Neural Networks: a Comprehensive Foundation. IEEE Press, Macmillan.

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

Holben, B.N., Tanré, D., Smirnov, A., Eck, T.F., Slutsker, I., Abuhassan, N., Newcomb, W.W., Schafer, J., Chatenet, B., Lavenue, F., Kaufman, Y.J., Van de Castle, J., Setzer, A., Markham, B., Clark, D., Frouin, R., Halthore, R., Karnieli, A., O'Neill, N.T., Pietras, C., Pinker, R.T., Voss, K., Zibordi, Z., 2001. An emerging ground-based aerosol climatology: aerosol optical depth from AERONET. J. Geophys. Res. 106, 12067–12097.

Hornik, K., 1993. Some new results on neural network approximation. Neural Netw. 6, 1069–1072.

IPCC, 2007. IPCC Fourth Assessment Reports (AR4): Working Group I Report: Climate Change 2007, the Physical Basis (WMO/UNEP Report).

Kim, S.-W., Yoon, S.-C., Kim, J., Kim, S.-Y., 2007. Seasonal and monthly variations of columnar aerosol optical properties over east Asia determined from multi-year MODIS, LIDAR, and AERONET sun/sky radiometer measurements. Atmos. Environ. 41, 1634–1651.

Lelieveld, J., Berresheim, H., Borrmann, S., Crutzen, P.J., Dentener, F.J., Fischer, H., Feichter, J., Flatau, P.J., Heland, J., Holzinger, R., Korrmann, R., Lawrence, M.G., Levin, Z., Markowicz, K.M., Mihalopoulos, N., Minikin, A., Ramanathan, V., De Reus, M., Roelofs, G.J., Scheeren, H.A., Sciare, J., Schlager, H., Schultz, M., Siegmund, P., Steil, B., Stephanou, E.G., Stier, P., Traub, M., Warneke, C., Williams, J., Ziereis, H., 2002. Global air pollution crossroads over the Mediterranean. Science 298, 794–799.

Liang, D., Kumar, N., 2013. Time-space Kriging to address the spatiotemporal misalignment in the large datasets. Atmos. Environ. 72, 60–69.

Miller, S.W., Emery, W.J., 1997. An automated neural network cloud classifier for use over land and ocean surfaces. J. Appl. Meteorol. 36, 1346–1362.

NOAA (National Oceanic and Atmospheric Administration), 2003. Website for Documentation and Data. Document. http://www.emc.ncep.noaa.gov/gmb/gdas/. ftp://arlftp.arlhq.noaa.gov/pub/archives/gdas1. data.

Olcese, L.E., Toselli, B.M., 2004. A method to estimate industrial stack parameters based on neural networks. Chemosphere 57, 691–696.

Oleson, J.J., Kumar, N., Smith, B.J., 2012. Spatiotemporal modeling of irregularly spaced aerosol optical depth data. Environ. Ecol. Stat. 20, 297–314.

Perez, P., Reyes, J., 2002. Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile. Atmos. Environ. 36, 4555–4561.

Polissar, A.V., Hopke, P.K., Paatero, P., Kaufmann, Y.J., Hall, D.K., Bodhaine, B.A., Dutton, E.G., Harris, J.M., 1999. The aerosol at Barrow, Alaska: long-term trends and source locations. Atmos. Environ. 33, 2441–2458.

Putaud, J.P., Cavalli, F., Dos Santos, M., Dell'Acqua, A., 2014. Long-term trends in aerosol optical characteristics in the Po Valley, Italy. Atmos. Chem. Phys. 14, 9129–9136.

Rahul, P.R.C., Sonbawne, S.M., Devara, P.C.S., 2014. Unusual high values of aerosol optical depth evidenced in the Arctic during summer 2011. Atmos. Environ. 94, 606–615.

Salomonson, V.V., Barnes, W.L., Maymon, P.W., Montgomery, H.E., Ostrow, H., 1992. MODIS: advanced facility instrument for studies of the earth as a system. IEEE Trans. Geosci. Remote Sens. 145–153.

Stanley, K.O., Miikkulainen, R., 2002. Evolving neural networks through Augmenting topologies. Evol. Comput. 10, 99–127.

Valenzuela, A., Olmo, F.J., Lyamani, H., Antón, M., Quirantes, A., Alados-Arboledas, L., 2012. Classification of aerosol radiative properties during African desert dust intrusions over southeastern Spain by sector origins and cluster analysis. J. Geophys. Res. Atmos. 117, 6.

Wang, Y., Stein, A.F., Draxler, R.R., de la Rosa, J.D., Zhang, X., 2011. Global sand and dust storms in 2008: observation and HYSPLIT model verification. Atmos. Environ. 45, 6368–6381.

Wang, C., Judd, S.J., Venkatesh, S.S., 1993. When to stop: on optimal stopping and effective machine size in learning. In: Proc. Conf. On Neural Information Processing Systems, Colorado.