Non–supervised classification of aerosol mixtures for ocean color remote sensing

Lydwine Gross[a], Robert Frouin[a], Christophe Pietras[b], and Giulietta Fargion[b]

[a]Scripps Institution of Oceanography, UCSD, La Jolla, California, USA.
[b]Sciences Applications International Corporation, SIMBIOS–NASA/GSFC, Greenbelt, USA.

ABSTRACT

Satellite ocean-color algorithms generally use aerosol-mixture models to estimate and remove the atmospheric contribution to the measured signal. These models, based on aerosol samples, may or may not be realistic. In atmospheric correction, we are more interested in the optical behavior of the aerosols through the entire atmosphere. Comparisons of SeaWiFS-derived and measured aerosol optical thickness have revealed a systematic underestimation of the Angström coefficient, suggesting that the reference models may not be representative of actual conditions. To investigate the adequacy of the models and ultimately to improve atmospheric correction, we analyze atmospheric optics data collected by the AERONET project under a wide range of aerosol conditions at coastal and island sites. Using non-supervised classification techniques (self-organized mapping, hierarchical clustering), we determine the natural distribution of retrieved aerosol properties of the total atmospheric column, i.e., the volume size distribution function and the refractive index, and more importantly identify clusters in this distribution. These clusters may be used as new aerosols mixtures in radiative transfer algorithms. We compare the clusters with the SeaWiFS reference models and, through application examples, conclude about their potential to improve atmospheric correction of satellite ocean color.

Keywords: Ocean color remote sensing, atmospheric correction, aerosols, classification, neural networks.

1. INTRODUCTION

Ocean color algorithms generally use aerosol mixture models firstly to evalutate the atmospheric contribution to the signal (atmospheric correction) and secondly derive the oceanic content, indexed by chlorophyll–a concentration. Indeed, the accuracy of ocean color retrievals from SeaWiFS, POLDER, OCTS, MODIS, GLI, MERIS, etc., relies on assumptions about the optical properties associated with each aerosol type. Gordon and Wang[1] used nine reference aerosol models, namely the Shettle and Fenn[2] oceanic, maritime, coastal, tropospheric models with a humidity variation of the aerosol optical properties, and introduced a coastal aerosol model, actually a mixture of the maritime and tropospheric models.

These models may or may not be realistic. Shettle and Fenn[2] developed their models using aerosol samples for which they derived the optical characteristics. In atmospheric correction, however, we are more interested in the optical behavior of the aerosols through the entire atmosphere. Comparisons of SeaWiFS-derived and measured aerosol optical thickness,[3] on the other hand, have revealed a systematic underestimation of the Angström coefficient. This might be evidence that the reference models are not representative of actual conditions, although it is not excluded that the discrepancy might be due to the procedure to select the models or to errors in the radiometric calibration.

To provide answers to the above questions (i.e., representation of the models, origin of atmospheric correction errors), and ultimately improve atmospheric correction, one needs to analyse atmospheric optics data under varied aerosol conditions, i.e., encountered over the world oceans. The Aerosol Robotic Network project

(Send correspondence to L. Gross)
L.Gross: lidwine@polaris.ucsd.edu, tel.: 1 858 822 1416, R. Frouin: rfrouin@ucsd.edu, tel.: 1 858 534 6243, C. Pietras: pietras@simbios.gsfc.nasa.gov, tel.: 1 301 286 9892, G. Fargion: gfargion@simbios.gsfc.nasa.gov, tel.: 1 301 286 0744,
Address: Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive, La Jolla, California, 92093-0221, USA.

Ocean Remote Sensing and Applications, Robert J. Frouin, Yeli Yuan, Hiroshi Kawamura, Editors,
Proceedings of SPIE Vol. 4892 (2003) © 2003 SPIE · 0277-786X/03/$15.00

(AERONET),[4] with CIMEL radiometers operating continuously at many island and coastal sites, the maturity of the CIMEL data processing procedures and inversion algorithms,[5] allows us to make global statistics on aerosols mixtures.

A non-supervised classification of the retrieved aerosol properties of the total atmospheric column, i.e., the volume size distribution function and the refractive index, may allow us to determine their natural distribution and more importantly to identify clusters in this distribution. We use here a probabilistic self-organizing map (PRSOM) to approximate the distribution of the data, followed by a hierarchical clustering to identify the more encountered geophysical conditions in the data base. From this classification, 23 separate "statistical mixtures" were created, and then compared to the SeaWiFS models. We assessed their effect on the atmospheric correction by comparing the selection of phase function and single scattering albedo during the process.

2. DATA SET DESCRIPTION

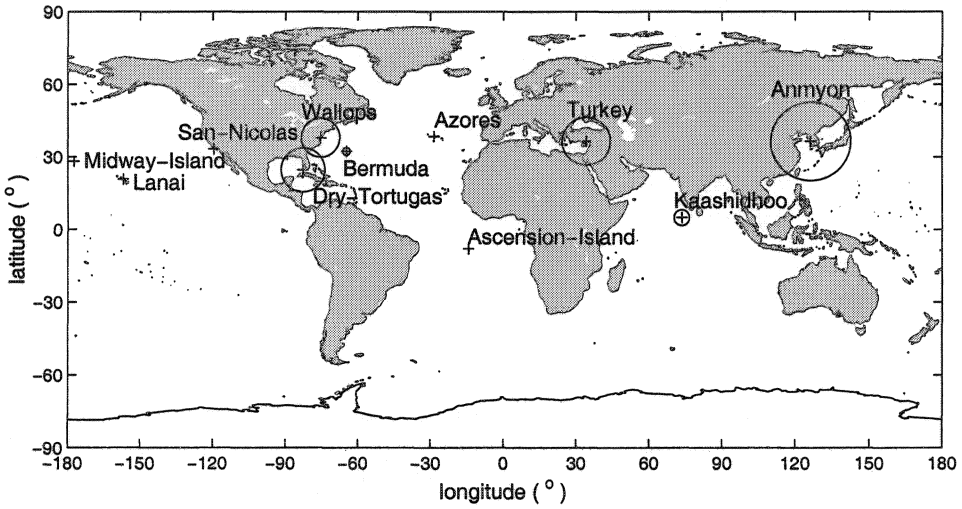

Figure 1. Locations of the AERONET retrievals used for the classification. The circle diameters are proportional to the amount of data in each site (in %): Azores, 0.14, San Nicolas, 0.18, Midway Island, 0.59, Ascension Island, 1.08, Lanai, 1.63, Bermuda, 3.21, Kaashidhoo, 6.47, Wallops, 16.15, Dry Tortugas, 18.27, Turkey, 19.81, Anmyon, 32.47.

2.1. AERONET retrievals from CIMEL radiometers

We classified cloud–screened retrievals of the AERONET inversion algorithm[5,6] on islands and coastal sites, i.e. columnar particle volume size distributions $dV/d\ln r$ for radii ranging from 0.05 to 15 μm, associated to refractive indices $\tilde{m}(\lambda) = n(\lambda) + i\ k(\lambda)$ at wavelengths 440, 670, 870 and 1020 nm. The AERONET inversion algorithm is a statistically optimized code which simultaneously computes $dV/d\ln r$ and $\tilde{m}$ from Sun radiance and angular distribution of sky radiance in the solar almucantar ($\theta = \theta_s$, where θ is the viewing zenith angle and θ_s is the solar zenith angle). The inversion is designed as a search for the best fit of all considered data by a theoretical model. It takes into account the different levels of accuracy of the fitted data, allowing the most accurate solution in the presence of random errors. The optimized solution is the minimum of a constrained quadratic form Ψ which includes *a priori* knowledge on the smoothness of the retrieved characteristics.

According to Ref. 7 where the robustness of the AERONET algorithm is tested when dealing with errors due to experimentation or physical assumptions, and as suggested in Ref. 8, we kept AERONET retrievals when Ψ_{min} is less that 7%, when the aerosol optical thickness $\tau(440)$ is greater than 0.15 and when the solar zenith angle θ_s is greater than 45°. These conditions allow us to deal with all types of aerosol mixture while minimizing errors on the retrievals. Moreover, we eliminated retrievals for which fine particles mode radius is

lesser than 0.1 μm, and for which large particles mode radius is larger than 7 μm as these rare cases cannot be well retrieved by the algorithm.

Figure 1 displays the geographic distribution of the data set after the selection described above (2211 vectors). More than 87% of the data are dispersed on four coastal sites: Anmyon, Turkey, Dry Tortugas and Wallops. This means that coastal aerosol mixtures are much more representated in the data set than marine aerosols. This however should not biais the results as our classification is able to separate rare occurences from the others, as long as these occurences present a distinghuishing pattern. In other words, even if the data set contains only a few cases of marine aerosols, they should be separated from other mixtures.

Practically all retrieved size distributions have bimodal structure with quite wide local minimum around $r = 0.6\ \mu m$ and shapes visually close to lognormal curves. Consequently size distributions may be approximated by a classical bimodal lognormal function[2,9]:

$$\frac{dV(r)}{d\ln r} = \sum_{i=1}^{2} \frac{C_{v,i}}{\sqrt{2\pi}\sigma_{v,i}} \exp\left[-\frac{(\ln r - \ln r_{v,i})^2}{2\sigma_{v,i}^2}\right], \tag{1}$$

where $i = 1$ for the fine mode, $i = 2$ for the coarse mode, $C_{v,i}$ is defined as the particle volume concentration of the mode i, $r_{v,i}$ its median radius, and $\sigma_{v,i}$ its standard deviation. These parameters are computed from the retrieved $dV/d\ln(r)$ using the formulas:

$$C_v = \int_{r_{min}}^{r_{max}} \frac{dV(r)}{d\ln r} d\ln r, \tag{2}$$

$$r_v = \exp\left[\int_{r_{min}}^{r_{max}} \ln r \frac{dV(r)}{d\ln r} d\ln r \Big/ \int_{r_{min}}^{r_{max}} \frac{dV(r)}{d\ln r} d\ln r\right], \tag{3}$$

$$\sigma_v = \sqrt{\int_{r_{min}}^{r_{max}} (\ln r - \ln r_v)^2 \frac{dV(r)}{d\ln r} d\ln r \Big/ \int_{r_{min}}^{r_{max}} \frac{dV(r)}{d\ln r} d\ln r}, \tag{4}$$

and defining all particles with radius smaller than 0.6 μm as belonging to the fine mode and all particles larger than 0.6 μm as belonging to the coarse mode. This transformation reduces the information contained in $dV/d\ln r$ to 6 parameters and eliminates noisy oscillations on $dV/d\ln r$, initially computed for 22 radii between 0.05 to 15 μm.

Aerosol size distribution and refractive index (mainly the imaginary part) influence the spectral dependence of aerosol optical thickness τ. Since each retrieval of size distribution and refractive index is associated to a measured τ spectrum, we used this information to constrain the classification. The relationship between $\ln\tau$ and $\ln\lambda$ is classicaly approximated by a linear fit[10]: $\ln\tau = -\alpha\ln\lambda + \beta$. The Angström parameter α gives a general description of the size distribution.[11] However, natural size distributions may induce a spectral curvature in the $\ln\tau$ versus $\ln\lambda$ relationship, and a high order fit may be more adapted.[12–14] More generally, the Angström parameter can be defined as the negative of the first derivative of the relationship: $\alpha = -d\ln\tau/d\ln\lambda$. The negative of the second derivative (denoted α'), which quantifies the amount of curvature, gives information about the relative dominance of fine versus coarse mode.[13–15] We made second–order polynomial fits on the measurements of τ spectra at the wavelength 440, 500, 670 and 870 nm, and then obtained a wavelength dependent $\alpha(\lambda)$ and a constant α' for each measurement. Second–order fits allow us to avoid overfitting as we kept only four wavelengths, chosen for their relative high confidency. Besides, a second order fit envelops most of the spectral variations of τ detectable by a Sun photometer.[14] In the classification we use the Angström parameter α at 670 nm, since current satellite sensors use spectral information in the red part of the spectrum.

2.2. Basic Statistics

One data vector (denoted $\mathbf{x}$) is composed by 16 variables: the Angström exponent $\alpha(670)$ and its derivative α', the 6 parameters describing the size distribution (Eq. 1, 2, 3, 4), 4 spectral values of the real part of the refractive index $n(\lambda)$, and 4 spectral values of the imaginary part $k(\lambda)$. By computing the correlation matrix

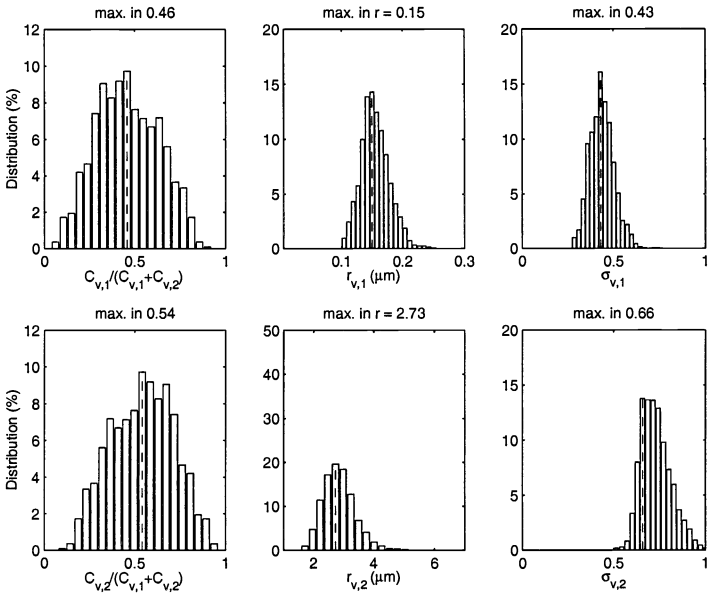

Figure 2. Histograms of the parameters describing aerosol size ditribution (Eq. 1, 2, 3, 4).

of the data set, we find that the information related to $dV/d\ln(r)$ (i.e. the first 8 dimensions of $\mathbf{x}$), is generally not correlated to the information related to the refractive index. Major correlations are seen between the real part and the imaginary part of the refractive index, and between the parameters $\alpha(670)$ and $C_{v,2}$, which are correlated with a factor of -60% (or $\alpha(670)$ and $\log_{10}(C_{v,2})$ are correlated with a factor of -74%).

Figure 2 displays histograms of the parameters describing size distribution. The median radius of the fine mode, $r_{v,1}$, is ranging from 0.1 to 0.2 μm, while in comparison the coarse mode radius, $r_{v,2}$ is ranging from 2 to 4 μm. The standard deviations $\sigma_{v,1}$ and $\sigma_{v,2}$ indicate that fine mode is generally narrower than coarse mode. The relative importance of fine mode versus coarse mode $C_{v,1}/(C_{v,1}+C_{v,2})$ is visualized in the upper left corner plot. All the situations are present, from a fine mode dominating situation like biomass burnings or urban soot, to a coarse mode dominating situation like dust or maritime aerosols. Most aerosol mixtures, however, present two separate modes. Correspondingly, histograms of Angström exponent $\alpha(670)$ and its derivative α' are shown in Figure 3. The parameter $\alpha(670)$ is ranging from 0.07 to 2.46, but most data have an $\alpha(670)$ around 1.3, due to the large amount of coastal data in our ensemble. As a comparison, an histogram of $\alpha(670)$ obtained by SIMBAD radiometers on oceanic cruises (http://polaris.ucsd.edu/ simbad/) exibits a maximum around 0.7, and so does recent statistics made in oceanic conditions with CIMEL radiometers on island sites.[16,17] We can thus conclude that oceanic conditions are well representated in our data set. The parameter α' is ranging from -2 to 2; most data, however, have a very slightly negative α', which means that their spectra are almost linear.

There is no particular correlation between aerosol optical thickness and the Angström parameter. For a standard aerosol loading ($\tau(870) = 0.1$), $\alpha(670)$ is ranging from 0.5 to 2.5. However, when the aerosol loading is increasing or disminishing, the possible range of $\alpha(670)$ proportionally decreases. For low $\tau(870)$, $\alpha(670)$ is restrained to high values (> 1.5), while for high $\tau(870)$, $\alpha(670)$ has a tendency to be confined to low values (< 1).

3. ASSESSING THE DISTRIBUTION OF AEROSOL VOLUME SIZE DISTRIBUTIONS AND REFRACTIVE INDICES

Our classification is performed using two steps.[18,19] First, the information contained in the data set is summarized by referent vectors (denoted $\mathbf{r}$), produced by a Probabilistic Self Organizing Map (PRSOM). This operation allows us to reduce 2211 row data to an affordable number of statistical vectors. Second, the referent

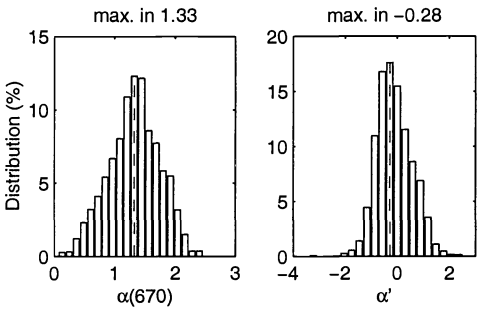

Figure 3. Histograms of Angström exponent $\alpha(670)$ and its derivative α'.

vectors are closely studied by the mean of a hierarchical clustering, which determines natural breaks between clusters in the data space. If some referent vectors have been found to form a cluster, they are declared to belong to the same class and are combined using a ponderated mean. At the end of the process, we obtain N classes, each one containing 1 particle volume size distribution and 1 refractive index.

3.1. Step 1: PRSOM

Self-Organizing Maps (SOM) are neural models, which were first introduced by Kohonen,[20] for visualising and clustering n-dimensional observations. SOM models have two-layers (Fig. 4, A), the input layer where the number of neurons is equal to the dimension of the data space (here 16) and the topological map layer which is a discrete lattice of neurons, connected to the neurons of the input layer by ponderated links (the neurons weights $\mathbf{r}$). In this study, the topological map is a 6 by 6 lattice (36 neurons) with a quadratic neighbourhood. The SOM algorithm minimizes a cost function which depends on the output values of the neurons located on the map, their weights and their topological relationship. Once the SOM is calibrated, each neuron j of the topological map gathers a subset of c_j similar data (c_j is usually called the cardinality of the neuron). The neurons weights stand for referent vectors characterising the gathered data, and are expressed in the data dimensions. They can be considered as a summary of the observation set under study. The topological relationship between the different neurons means that close neurons on the map represents close data in the original data set.

In the present study, we use an advanced SOM model, the Probabilistic Self Organising Map,[21] which deals with a probabilistic formalism. Each neuron of the topological map is associated with a spherical Gaussian density function defined by its mean and its covariance matrix. The PRSOM algorithm approximates the probability distribution of the data with a mixture of normal distributions. As in the SOM formalism, each neuron represents a subset of observations, but this subset is now summarized by both a referent vector $\mathbf{r}_j$ and a measurement of its dispersion $\mathcal{V}_j$. The referent vector (the weights of the neuron) is the center of the neuron's Gaussian function, and $\mathcal{V}_j$, its variance-covariance matrix. We took hyperspherical gaussians described by a a unique variance v_j for each dimension. As for SOM, these parameters are estimated during a learning phase by minimizing a cost function, but this latter is more sophisticated.

3.2. Step 2: hierarchical clustering (HC)

The referent vectors $\mathbf{r}$, which summarize the data set, are submitted to a cluster analysis. Indeed, we cannot directly use the referent vectors as statistical aerosol models since we had to choose *a priori* the number of referent vectors made by the PRSOM (36). Some referent vectors may be very close to one others and thus carry redundant information. A classical binary, hierarchical cluster tree may help us to detect natural groupings among the referent vectors.[22]

We first calculate the distance δ between every pair of referent vectors using the Ward dissimilarity:

$$\delta(\mathbf{r}_p, \mathbf{r}_q) = \frac{c_p\, c_q \sum_{k=1}^{m} (r_{pk} - r_{qk})^2}{c_p + c_q}, \tag{5}$$

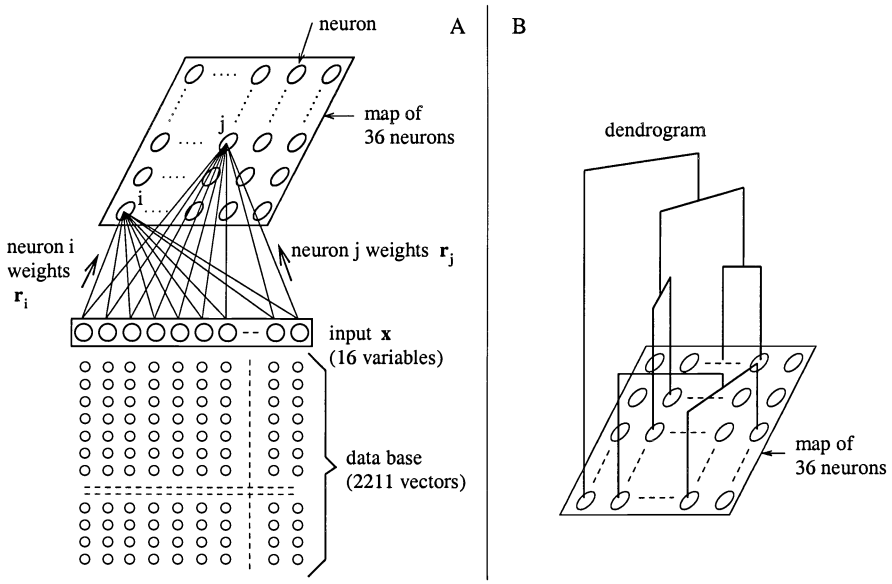

Figure 4. A: Architecture of the self organizing map: the input layer receives the 16–dimension data set, which is spread on the topological map layer (here 36 neurons on a quadratic lattice). B: Example of cluster tree (dendrogram) of the the topological map referent vectors.

where $m = 16$ is the dimension of the data; c_p and c_q (cardinalities) are the numbers of actual data represented by the vectors $\mathbf{r}_p$ and $\mathbf{r}_q$, respectively. Then we group the two closest referent vectors, creating a new class referred by its gravity center $\mathbf{r}_{pq}$ and its weight c_{pq}:

$$\mathbf{r}_{pq} = \frac{c_p\,\mathbf{r}_p + c_q\,\mathbf{r}_q}{c_p + c_q}\,, \tag{6}$$

$$c_{pq} = c_p + c_q\,. \tag{7}$$

The HC algorithm consists in computing a new dissimilarity matrix between this class and the remaining referent vectors and to iterate the process until only one class remains. The Ward dissimilarity allows us to minimize the increase of intra-class variance while building the cluster tree. The dendrogram of the referent vectors (Fig. 4, B) illustrates the HC result: the height of a link z_k is proportional to the distance between the linked classes.

One way to determine the natural cluster divisions of the referent vectors is to compare the height of each link in the dendrogram with the heights of neighbouring links below it. If the height of a link differs from neighboring links, it indicates that there are dissimilarities between the objects at this level in the cluster tree. This link is said to be inconsistent with the links around it. In cluster analysis, inconsistent links can indicate the border of a natural division in a data set. To compute the inconsistency coefficient I_k of each link k, we use the formula:

$$I_k = (z_k - \bar{z})/\sigma_z\,, \tag{8}$$

where z_k is the height of the link in the dendrogram, $\bar{z}$ is the mean of the z values for k and three links above k, and σ_z is the z standard deviation for the same links.

Once the cluster tree is cut, the l referent vectors of each cluster i are combined by a ponderated mean, taking into account the number and the variance of actual data attached to each referent:

$$\mathbf{R}_i = \sum_{j=1}^{l}(c_j/v_j)\mathbf{r}_j \Big/ \sum_{j=1}^{l}(c_j/v_j)\,. \tag{9}$$

Finally we obtain N classes $\mathbf{R}_i$, each one containing 1 normalized particle volume size distribution and 1 refractive index.

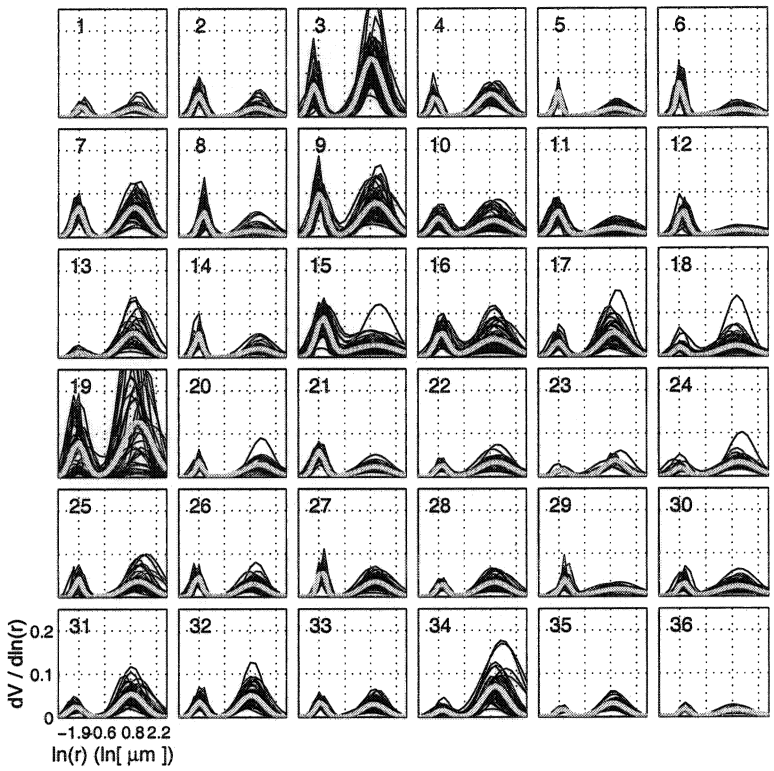

Figure 5. The calibrated PRSOM 36 neurons: volume size distribution $dV/d\ln r$. In each plot, the data gathered by the corresponding neuron are plotted in black, while their referent vector is plotted in gray.

3.3. Results

Step 1: the referent vectors

for the PRSOM, 1 data is 1 particle volume size distribution function attached to 1 refractive index, 1 Angström parameter and its derivative. But because these informations do not have the same units, we cannot present the results on the same figure. Fig. 5 shows the result of the PRSOM algorithm only for the information related to the particle volume size distribution function $dV/d\ln r$. On this figure, each plot stands for 1 neuron of the calibrated PRSOM (36 neurons on the whole). In each plot we display the data gathered by the neuron (in black) and the corresponding referent vector which represents them in the further steps of the classification (in gray). Similar results are obtained for the refractive index and the Angström parameters.

This series of plots allows us to visually estimate the quality of the summary obtained by PRSOM. We can see for example that neuron 19 has failed to gather similar data, and may be considered as a "dumpster" neuron. The quality of each neuron j may also be estimated by the study of c_j and v_j, its cardinality and its variance, respectively (see Section 3.1). The greater c_j and the lesser v_j the better. Except neuron 19, the neuron variances are ranging from 2 to 4% of the total variance, which is acceptable since a perfect spreading of the data on 36 neurons would lead to 2.8% of variance per neuron. Note that the scatter of the data around the referent vectors may be explained by the fact that we classified together a lot of different uncorrelated information (see Section 2.2). The cardinality is ranging from 0.5 to 5.4% of the total amount of data (2211). This means that neurons 23 and 36, for example, are much less important than neurons 5 and 26, and that is why we use Ward dissimilarity to classify the referent vectors in the following. Using this criteria, weak neurons are more likely grouped with others than strong neurons.

Step 2: the hierarchical clustering of the referent vectors

Since we could not use the referent vector of neuron 19, we removed it from our ensemble and performed

our hierarchical clustering on the remaining referent vectors $\mathbf{r}_j$. We chose to cut the cluster tree where link inconsistencies I_k (Eq. 8) were above 0.7. This led to $N = 23$ clusters which gravity centers $\mathbf{R}_i$ (Eq. 9) are shown in Fig. 6. Applying Mie theory to these $\mathbf{R}_i$, we computed corresponding theoretical aerosol optical thicknesses τ_i for four wavelength (with a height scale of 1 km) and a new Angström exponent $\alpha_i(670)$ like described in section 2.1 (the results are given in Table 1). We chose the numbering so that $\alpha(670)$ increases with the cluster number. Cluster no. 1 has a $\alpha(670)$ equal to 0.72, and Cluster no. 23 has a $\alpha(670)$ equal to 1.98. Data with lower and higher $\alpha(670)$ do not appear in the classification result since they were too rare in the data base (see Fig. 3). The theoretical aerosol optical thickness is ranging from 0.06 to 0.26 at 865 nm. The description of each cluster is detailed on Table 1. Except for cluster no. 19 which is composed by data from three coastal sites (Wallops, Turkey, Anmyon), each cluster gathers data from both coastal and island sites. From a statitical point of view, the best clusters are clusters no. 4, 8, 12, 17 and 22, as they all gather more than 155 data while having a low variance. Note that from Clusters no. 1 to 11, the coarse mode dominates, while starting from Cluster no. 12, the fine mode dominates.

4. COMPARISON BETWEEN CIMEL-DERIVED AEROSOL MIXTURES AND SEAWIFS AEROSOL MODELS

Using Mie theory, the 23 gravity centers $\mathbf{R}_i$ obtained in the previous section were transformed into phase functions $P_i(\chi, \lambda)$ (where χ is the scattering angle) and single scattering albedo $\omega_i(\lambda)$, leading to 23 "statistical mixtures". We now compare these CIMEL-derived mixtures to the 12 SeaWiFS models,[1] and assess their impact on atmospheric correction and estimation of oceanic chlorophyll content.

4.1. Phase functions and single scattering albedo

Figure 7 displays the phase functions and single scattering albedos of the SeaWiFS models (left) and statistical mixtures (right). The statistical phase functions present less variability than the SeaWiFS phase functions, however, the statistical single scattering albedos are much more variable and induce more absorption than the SeaWiFS albedos. The SeaWiFS atmospheric correction is based on the parameter $\epsilon(\chi, \lambda, \lambda_0)$[23]:

$$\epsilon(\chi, \lambda, \lambda_0) \equiv \rho_{as}(\lambda)/\rho_{as}(\lambda_0)\,, \tag{10}$$

where ρ_{as} is the top of the atmosphere single–scattering reflectance due to aerosols, and λ_0 a wavelength of reference ($\lambda_0 = 865\ nm$). The atmospheric correction assumes that in the red and the near infrared the ocean contribution to the top of the atmosphere signal is negligible, and thus, the parameter $\epsilon(\chi, 670, 865)$ can be estimated directly from the satellite reflectance corrected from the Rayleigh scattering. The value of the parameter ϵ for a given geometry allows to determine which aerosol model is to be used for the atmospheric correction. Figure 8, left, shows the relationships that could be established between ϵ and the aerosol phase function, using the statistical or the SeaWiFS phase functions, for four geometries corresponding to χ values of 0, 30, 60 and 90° (θ and θ_s are taken less than 60°). Although the statistical mixtures and SeaWiFS models do not cover the same range of ϵ, we can see that the relationship $P = f(\epsilon)$ is very different for the two sets of

Table 1. Description of each cluster i: the neurons j which compose it, the number of data it gathers, n_i, the percentage of total variance it represents, v_i, the corresponding Angström parameter $\alpha_i(670) \times 10$, and aerosol optical thickness at 865 nm for a height scale of 1 km, $\tau_i(865) \times 10^2$.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
j	3	34	13	4	17	7	25	24	26	28	16	9	35	15	11	18	5	29	14	1	8	2	12
	31			32	23	10			33			21			22		27	30	20		36	6	
n_i	133	69	44	157	61	114	70	53	217	52	57	155	39	57	104	83	189	94	78	40	109	159	31
v_i	5	5	4	3	5	4	5	5	3	4	5	4	4	6	4	6	3	5	4	4	5	4	4
α_i	7	8	9	9	12	12	13	13	14	14	14	14	14	15	15	16	16	17	18	18	19	19	20
τ_i	21	13	14	13	14	18	09	10	09	11	23	18	09	26	13	11	09	13	09	06	11	09	13

models. The same observation can be made for the relationships $\omega = g(\epsilon)$ and $\alpha = h(\epsilon)$, which are also necessary to model the aerosol reflectance.

4.2. Impact of aerosol model selection on chlorophyll concentration retrieval

To evaluate the potential impact of aerosol model selection on the retrieval of phytoplankton pigment concentration, let us assume that the actual aerosol model is the average CIMEL-derived model characterized by $\epsilon_1(\chi, \lambda, 865)$. The atmospheric correction algorithm will select the Shettle and Fenn model characterized by $\epsilon_2(\chi, \lambda, 865)$, and we have $\epsilon_1(\chi, 670, 865) = \epsilon_2(\chi, 670, 865)$. The error $d\rho_{as}$ on ρ_{as} due to selecting the Shettle and Fenn model instead of a CIMEL-derived model is approximately expressed as:

$$d\rho_{as}(\lambda) \approx [\epsilon_2(\chi, \lambda, \lambda_0) - \epsilon_1(\chi, \lambda, \lambda_0)]\rho_{as}(\lambda_0) . \quad (11)$$

Table 2 displays the $d\rho_{as}$ values in the blue (443 nm) and green (555 nm) for typical solar and viewing angles and $\epsilon(\chi, 670, 865)$ values of 1.6 (scattering angle of 0 degree) and 1.3 (scattering angles of 30, 60, and 90 degrees). The aerosol optical thickness is assumed to be 0.1 at 865 nm. The errors are large in magnitude for scattering angles of 0 and 90 degrees, equal or above the maximum absolute value of 0.002 permitted in the blue (e.g., Gordon[23]). In the open ocean, relative errors are larger in the green, where marine reflectance is generally low (a few 0.001). Values are more negative (i.e., marine reflectance is more overestimated) as scattering angle decreases to zero. The resulting effect on phytoplankton pigment concentration is presented in Fig. 8, right. For the scattering angles of 0 and 90 degrees, the errors are strongly dependent on pigment

Table 2. Error on retrieved aerosol reflectance, ρ_{as}, due to aerosol model selection for several geometric conditions. Aerosol optical thickness is 0.1 at 865 nm. The actual aerosol model is deduced from the CIMEL observations, and the selected model is the Shettle and Fenn model having the same $\epsilon(\chi, 670, 865)$. The $\epsilon(\chi, 670, 865)$ values considered are 1.6 for the first case ($\chi = 0$), and 1.3 for the others.

θ_s	θ	χ	$d\rho_{as}(443)$	$d\rho_{as}(555)$
15	15	0	-0.0123	-0.0028
30	30	30	-0.0020	-0.0004
45	45	60	-0.0002	≈ 0
60	60	90	0.0051	0.0015

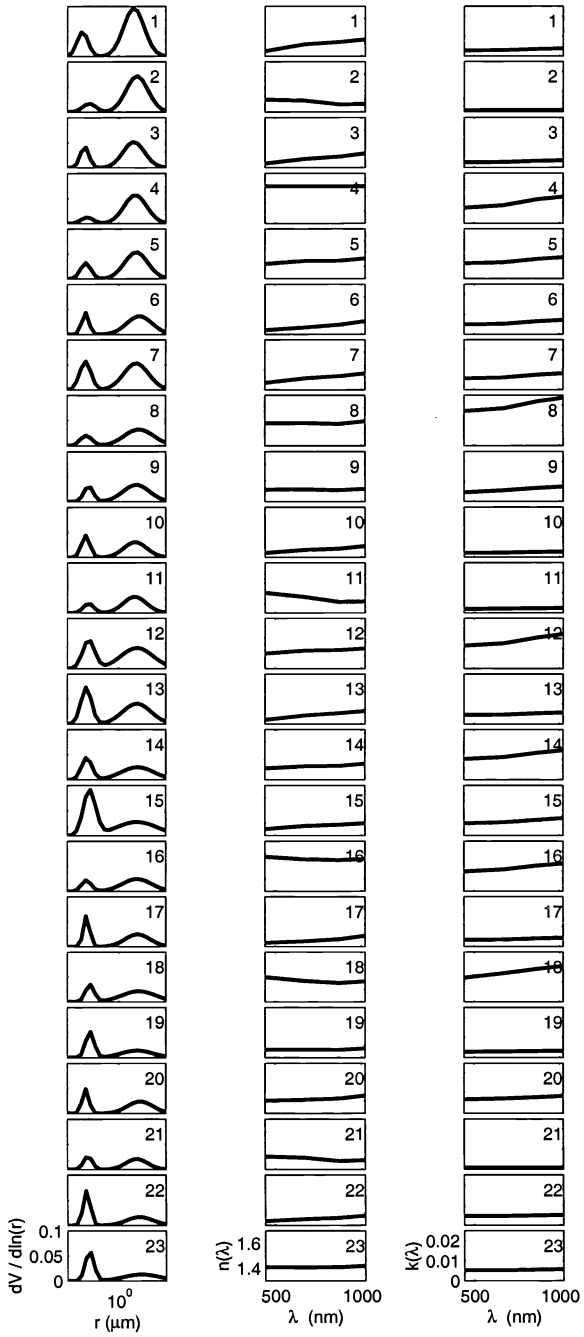

Figure 6. Gravity centers of the $N = 23$ clusters obtained by the classification. First column is $dV/d\ln r$, second column is n, and third is k.

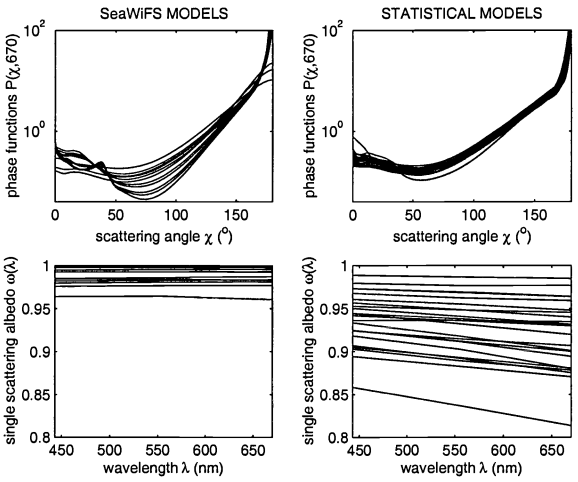

Figure 7. Phase functions and single scattering albedo of the 12 SeaWiFS models and the 23 statistical models. The Angström parameter $\alpha(670)$ is ranging from -0.08 to 1.49 for the SeaWiFS models, and from 0.72 to 1.98 for the statistical models.

concentration, and become very large in magnitude ($> 100\%$ for 0^o) at pigment concentrations above 1 $mg\ m^{-3}$. Even at scattering angles of 30 and 60 degrees, where the $d\rho_{as}$ values are smaller, errors may reach over 20% and 80% in magnitude, respectively.

5. DISCUSSION AND CONCLUSION

This global statistical study shows that CIMEL-derived aerosol mixtures are very different from the classical Shettle and Fenn models. If we assume that AERONET provides representative conditions for ocean color remote sensing, two important conclusions have to be emphasized. First, oceanic conditions, corresponding to low $\alpha(670)$, large coarse mode in the aerosol size distribution, and low absorption are more likely to be close to $\alpha(670) \approx 0.7$ like in our cluster no. 1, or in the statistics of Smirnov et al.[16,17] The Shettle and Fenn models allow us to process lower $\alpha(670)$ values, but these conditions may not be the most encountered, even in the open ocean. Second, the differences between the phase functions and single scattering albedos of the two sets of models may have a large impact on the atmospheric correction, and thus on the retrieval of chlorophyll–a concentration.

Large errors in SeaWiFS-derived chlorophyll–a concentration, however, have not been generally reported in evaluation studies. It was shown, on 2849 samples, that retrieved and measured values agree to about 0.22 on a logarithmic scale.[24–26] This good performance may be explained, at least partly, by the vicarious radiometric calibration method used for SeaWiFS, which forces agreement between estimates and measurements of water-leaving radiance at the MOBY site.[27] The statistical models based on CIMEL measurements, which take into account the entire atmospheric column, appear as a suitable alternative to the Shettle and Fenn models. Their application to SeaWiFS, however, would require re-adjusting the calibration coefficients of the satellite instrument. This would provide, in particular, a more realistic spectral dependence of the aerosol reflectance in the red and near-infrared.

ACKNOWLEDGMENTS

This work was supported by the SIMBIOS project (NASA). We thank A. Smirnov, O. Dubovik and M. Wang (from NASA Goddard Space Flight Center) for stimulating discussions. CIMEL data were gathered by the AERONET project.