

Statistical analysis of absorption spectra of phytoplankton and of pigment concentrations observed during three POMME cruises using a neural network clustering method

Aymeric Chazottes,^{1,*} Michel Crépon,¹ Annick Bricaud,² Joséphine Ras,² and Sylvie Thiria¹

¹Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques (LOCEAN/IPSL),
4 place Jussieu 75252 Paris, France

²Laboratoire d'Océanographie de Villefranche, CNRS and Université Pierre et Marie Curie, Villefranche-sur-Mer, France

*Corresponding author: ayclod@locean-ipsl.upmc.fr

Received 10 November 2006; revised 12 March 2007; accepted 13 March 2007;
posted 16 March 2007 (Doc. ID 76985); published 31 May 2007

We present a neural network methodology for clustering large data sets into pertinent groups. We applied this methodology to analyze the phytoplankton absorption spectra data gathered by the Laboratoire d'Océanographie de Villefranche. We first partitioned the data into 100 classes by means of a self-organizing map (SOM) and then we clustered these classes into 6 significant groups. We focused our analysis on three POMME campaigns. We were able to interpret the absorption spectra of the samples taken in the first oceanic optical layer during these campaigns, in terms of seasonal variability. We showed that spectra from the PROSOPE Mediterranean campaign, which was conducted in a different region, were strongly similar to those of the POMME-3 campaign. This analysis led us to propose regional empirical relationships, linking phytoplankton absorption spectra to pigment concentrations, that perform better than the previously derived overall relation. © 2007 Optical Society of America

OCIS codes: 010.4450, 010.7340.

1. Introduction

The principal objective of analyzing large databases is to extract pertinent information, such as seasonal and regional characteristics [1,2,3] and to present this information in a suitable form to facilitate its interpretation. Recently Chazottes *et al.* [4] proposed an advanced neural network method for analyzing the Laboratoire d'Océanographie de Villefranche (LOV) data set which is a large set of phytoplanktonic absorption spectra. In the present paper we present a methodology derived from that of Chazottes *et al.* [4] and able to capture regional and seasonal information embedded in the same data set.

The LOV has gathered a large set of ocean water samples for which the absorption spectra of phyto-

plankton and the corresponding pigment concentrations have been measured. These samples were taken during several cruises covering different parts of the world ocean at different seasons and therefore present a wide variety of situations. Chazottes *et al.* [4] processed the phytoplankton absorption spectra with a sophisticated neural network method suitable for classifying complex phenomena, the so-called self-organizing map (SOM) proposed by Kohonen [5]. The aim was to compress the information embedded in the data set into a reduced number of classes, the so-called reference vectors, *rv*, which characterize the data set; this facilitates the subsequent analysis. Chazottes *et al.* [4] were thus able to retrieve well-known relationships among pigment concentrations and to display them on maps to facilitate their interpretation. They were also able to propose new empirical relationships linking absorption spectra and pigment concentrations. In Chazottes *et al.* [4], the

number of classes is quite large (10×10), making some analyses, such as that of spatial or temporal variability, quite difficult. The objective of the present paper is to extend the work of Chazottes *et al.* [4] by clustering the SOM classes into a small number of groups in order to facilitate the interpretation in terms of bio-optical considerations. We were able to reveal regional and seasonal structures for which specific empirical relationships between absorption spectra and pigment concentrations can be proposed for each group. Focus is placed on the analysis of the biogeochemical characteristics of several oceanic regions represented in the LOV database.

2. Data and Methods

The database we processed and the SOM algorithm have been fully described by Chazottes *et al.* [4]; we recall here its major characteristics.

A. Database and Data Sets

Water samples were collected during ten cruises, in various seasons and various areas of the world ocean, between 1990 and 2001 (Table 1). In this study, we consider phytoplankton absorption spectra as pigment concentrations corresponding exclusively to oceanic case-1 waters.

Methods employed for particulate and algal absorption measurements are described in detail by Bricaud *et al.* [6,7]. The absorption spectrum was measured every 2 nm from 400 to 700 nm. We then applied a triangular moving window of size 3 to filter the noisy part of each spectrum. The filtered spectra were then sampled every 10 nm. Each spectrum is therefore represented by a 31-dimension vector. The phytoplankton spectral absorption coefficients are represented by the symbol $a_{ph}(\lambda)$ where λ stands for the wavelength in nanometers (nm).

Pigment concentrations were measured by high-pressure liquid chromatography (HPLC), using the procedure described by Vidussi *et al.* [8]. All the pig-

ments were grouped into five main categories, according to their spectral similarities [6,7].

Owing to the large variation (covering several decades) in the absorption spectral values, $a_{ph}(\lambda)$, we subsequently used $\log_{10}(a_{ph}(\lambda))$ values rather than $a_{ph}(\lambda)$ values, and for analogous reasons we also used the log-transformed values of the pigment concentrations.

Each absorption spectrum is thus represented by a 31-component vector, whose first 30 components are the spectrum derivatives computed as the difference between the $\log_{10}(a_{ph}(\lambda))$ values for two consecutive wavelengths (i.e., $\log(a_{ph}(\lambda_{400+10i})) - \log(a_{ph}(\lambda_{400+10(i+1)}))$, where $i = 1 \dots 30$), and the last component is the maximum value of the absorption [4].

The whole data set, D , which is described in Table 1, comprises 3734 samples. In the present study the learning data set, L , which was used to estimate the parameters of the self-organizing map, comprised 2163 samples. These data are from the samples of various cruises, among which are those of the POMME 1, 2, and 3 cruises. To avoid a possible bias due to the large number of the POMME data, only 525 POMME samples were retained in the learning set, L . The remaining 1571 POMME samples constituted a validation set, denoted hereinafter V .

In the present study, we focused on the properties of water samples in the surface layer, the most accessible to ocean color sensors and the most active one in terms of seasonal biological activity. The "surface" layer has been defined here as having a thickness equal to the penetration depth (i.e. the depth above which 90% of the diffusely reflected irradiance originates), computed with respect to the photosynthetically available radiation (PAR). This penetration depth is approximately equal to $z_e/4.6$, where z_e represents the euphotic depth. This euphotic depth was either directly measured during the different cruises or computed from the chlorophyll profile according to Morel and Maritorena [9].

Table 1. Information Concerning the Cruises on which the Different Water Samples Were Collected

| Cruises | Location | Usual Trophic State | Date | Total Number of Samples |
|---------|---------------|---------------------------|---------------------|-------------------------|
| 1 | TOMOFRONT | Mesotrophic | April 1990 | 28 |
| 2 | EUMELI3 | Oligotrophic, mesotrophic | Oct. 1991 | 49 |
| 3 | FLUPAC | Oligotrophic | Sept.–Oct. 1994 | 80 |
| 4 | OLIPAC | Oligotrophic | Nov. 1994 | 183 |
| 5 | MINOS | Oligotrophic | May 1996 | 115 |
| 6 | ALMOFRONT2 | Mesotrophic | Dec. 1997 Jan. 1998 | 477 |
| 7 | PROSOPE (Med) | Oligotrophic | Sept.–Oct. 1999 | 554 |
| 8 | PROSOPE (upw) | Eutrophic | September 1999 | 52 |
| 9 | POMME 1 | Mesotrophic | Feb.–March 2001 | 187 + (561) |
| 10 | POMME 2 | Mesotrophic | March–May 2001 | 193 + (577) |
| 11 | POMME 3 | Oligotrophic | Aug.–Oct. 2001 | 145 + (433) |
| 12 | BENCAL | Mesotrophic, eutrophic | Oct. 2002 | 100 |

Table 2. Sample Repartition of the Surface Samples in the Different Data Sets and Groups

| | Learning Set: L_{surf} | | Validation Set: V_{surf} | | |
|------------|--------------------------|-------------|----------------------------|-------|-------|
| | Learning | PROSOPE-Med | POMME | POMME | POMME |
| | | | 1 | 2 | 3 |
| Group 1 | 50 | - | - | - | - |
| Group 2 | 84 | - | 65 | 61 | - |
| Group 3 | 24 | - | 3 | - | 1 |
| Group 4 | 62 | - | - | 4 | 12 |
| Group 5 | 141 | 102 | - | - | 71 |
| Group 6 | 16 | - | - | - | - |
| All groups | 377 | 102 | 68 | 65 | 84 |

We thus defined three other data sets restricted to the surface-layer samples: D_{surf} , (the complete data set for the first optical layer), L_{surf} (the learning data set for the first optical layer), V_{surf} (the POMME validation data set for the first optical layer). The number of samples in each data set is indicated in Table 2.

B. Self-Organizing Map

The SOM is an unsupervised classification method which extracts pertinent statistical information from the data. This method proposed by Kohonen [5] has been extensively described [10,11]. It is used for visualizing and clustering multidimensional data sets. The classification aims at summarizing the information contained in the learning data set, L , by producing a set of reference vectors rv (synthetic spectra) that are representative of the data: close neurons on the map represent similar subsets of data (classes presenting similarities).

C. Self-Organizing Map + Hierarchical Ascendant Classification Method

The SOM algorithm was used by Chazottes *et al.* [4]. These authors decomposed the LOV data set into 10×10 classes corresponding to specific statistical characteristics of the absorption spectra. This large number of classes allowed us to take into account the complexity of the data set but may have prevented us from synthesizing some information embedded in the data, such as regional or seasonal specificities. To counteract this difficulty, we decided to aggregate this large number of classes into a smaller number of groups based on their similarities. This was done by using a hierarchical ascendant classification (HAC), as explained by Niang *et al.* [10]. The SOM algorithm plus the HAC classification is hereafter designated as the SOM + HAC method. In the present work, we aggregated the 10×10 neurons into eight significant groups. The resulting clustering of the spectra associated with the neurons of the SOM is given in Fig. 1. We note that the SOM + HAC clustering is very coherent, since the groups represent clusters whose neurons are contiguous on the SOM.

The number of groups (8) was selected because it presents the most significant discriminative partition with respect to the full dendrogram of the HAC (not shown). Figure 2 shows the upper part of the dendrogram in which groups 7 and 8 have been left apart, owing to the low number of particular spectra that

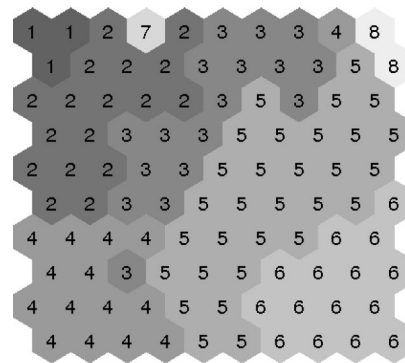


Fig. 1. Clustering of the neurons on the SOM. Following the HAC, based on the maximum amplitude and the slopes of the spectra, eight groups were retained. The group number resulting from the HAC is displayed for each neuron. The SOM + HAC clustering is coherent, since the groups represent clusters of contiguous neurons.

fell into them. Two groups connected by the same dendrite present more similarities than those connected by hierarchical dendrites.

In the following, we study the variability of the phytoplankton absorption spectra from samples corresponding only to the first optical depth (surface waters), which are those observed by satellite sensors. For each group, we computed the mean and standard deviation corresponding to the spectra retained by its neurons, the mean values of the spectrum derivatives (as defined previously), the mean pigment concentrations and their normalized ratios (i.e. pigment concentration:Tchl-a ratios). In Chazottes *et al.* [4], the neurons, which were clustered according to their Tchl-a concentration, were also found to be associated with pigment ratios relative to Tchl-a (Tchl-b/Tchl-a, etc.). Although Fig. 3 shows that the groups are ranked in decreasing order with respect to their Tchl-a concentration, the various group ranges partially overlap each other. Figure 4 displays the mean pigment concentrations relative to Tchl-a of the L data set for each group. We note the different patterns of groups 4, 5, and 6 with respect to that of groups 1, 2, 3 which present a lower TPPC/Tchl-a ratio, showing that SOM + HAC is able to cluster the water samples not only from the amplitude of their absorption spectra, but also from their

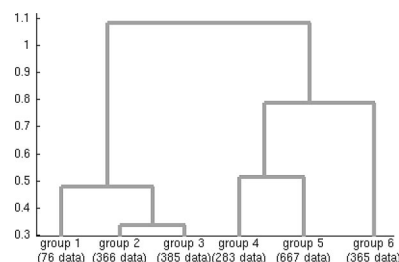


Fig. 2. Top of the dendrogram resulting from the HAC is presented for the six groups used in the present study. The number of data of L corresponding to each group is also known. The higher the node linking two groups the further apart they are in the data space.

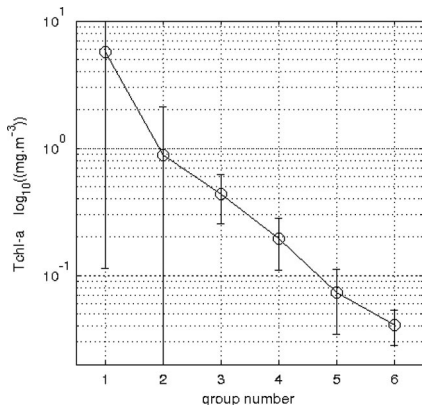


Fig. 3. Mean Tchl-a concentrations and their standard deviation with respect to the groups; the groups are ranked according to their Tchl-a concentration. The overlapping of the error bars suggests that the HAC contains some additional information on the data set.

pigment composition. In particular, Fig. 5 shows the partition of the L_{surf} samples for the different cruises in the six groups that cluster samples having similar spectral characteristics. We note that groups 2, 3, 4, 5, 6 gather samples coming from at least three cruises corresponding to different oceanic regions and different seasons.

D. Interest of the Method

Several questions may be raised on the above methodology. The first one concerns its usefulness. Could the following analysis have been conducted with a careful inspection of the database? The answer is yes, but with a lot of effort and laborious comparisons and tests. A major advantage of the SOM + HAC method is its highly efficient discrimination in producing a coherent first-order classification of the absorption spectra (amplitude of the spectra associated with the Tchl-a) as well as a second-order classification (pigment concentrations normalized by Tchl-a concentration i.e. pigmentary composition). A second question is the visualization associated with the SOM and the group clustering, allowing us to easily and quickly

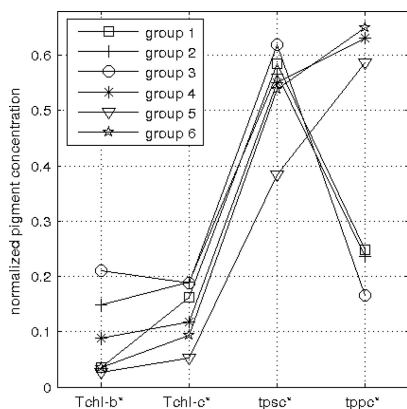


Fig. 4. Normalized pigment concentrations for the six significant groups; each group has different standardized pigment ratios. Discrepancies among the groups result from these various pigment combinations.

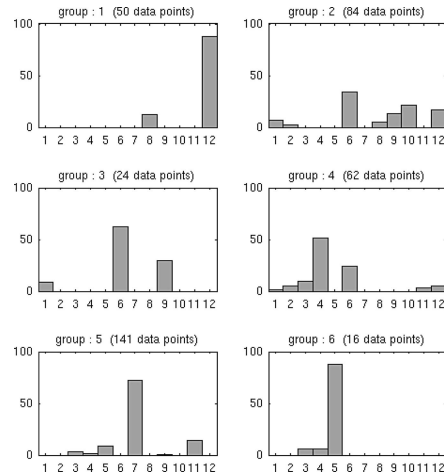


Fig. 5. Histogram of the distribution of the surface water samples from the 12 cruises in each of the six groups. The histogram is computed as the ratio of the number of samples of a given campaign in a given group to the total number of samples in that group. Basic information on the campaigns, including their respective numbers, is given in Table 1.

capture specific information embedded in the database.

Two theoretical questions remain concerning the HAC. First, could we have carried out the HAC directly on the data instead of on the rv of the SOM? It has been shown [12] that it is more efficient to do the HAC on the rv , which represent the most significant synthetic observations associated with the database than to apply it directly to the data which may be noisy. We could also have directly used a (3×2) SOM to partition the full data set into a small number of clusters. We tested such a SOM algorithm. The partition we obtained is less pertinent in terms of physical interpretation than the SOM + HAC method when looking at the different elements of the database clustered in the different classes. The reason is that the (3×2) SOM does not have a sufficient degree of liberty with respect to the (10×10) SOM + HAC algorithm to adequately extract the statistical information from the original database that can be noisy and to make a pertinent partition.

In the following we used SOM + HAC to analyze the POMME samples of the surface (V_{surf}) and some bio-optical relationships relating pigment concentrations to absorption spectra.

3. Analysis of Absorption Spectra from the POMME Cruises

We now try to characterize the biogeochemical variability of the POMME region by analyzing the phytoplanktonic absorption spectra with the SOM + HAC method. The POMME (Programme Océan Multidisciplinaire Méso Echelle) experiment took place in the North Atlantic Ocean ($21.33^\circ - 15.33^\circ W$, $38.00^\circ - 45.00^\circ N$) [13]; the region was extensively sampled between October 2000 and September 2001.

We processed the phytoplankton absorption spectra from the 217 V_{surf} surface samples from the

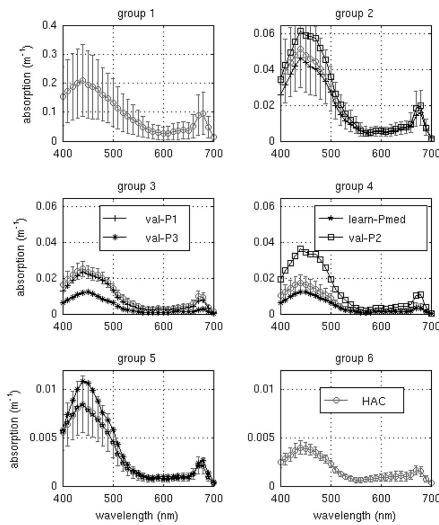


Fig. 6. Group clustering of absorption spectra in samples from the first optical layer. The mean absorption spectrum and its standard deviation computed from the surface learning data set are displayed for each group (open circles). The mean absorption spectra for the samples from each of the three POMME cruises (val-P1, val-P2 and val-P3) and for the PROSOPE-Med cruise (learn-Pmed) have been superposed.

POMME 1, 2, and 3 cruises, using the SOM + HAC method described in Section 2. Table 2 shows the distribution of the data among the various groups for the learning surface data set L_{surf} and the three POMME campaigns (V_{surf}). In Fig. 6, for each of the six groups, we show the mean and the standard deviation of the spectra (estimated using L_{surf}) as well as the mean spectra for each of the three POMME

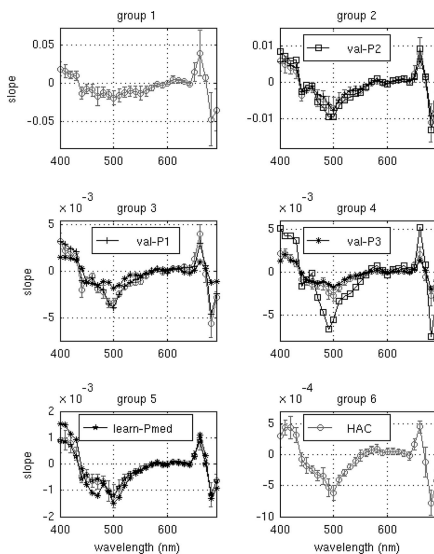


Fig. 7. Group clustering of the derivative of the absorption spectra in samples from the first optical layer. The mean absorption spectrum slope and its standard deviation computed from the surface learning data set are displayed for each group (open circles). The mean absorption spectrum slopes for the samples from each of the three POMME cruises (val-P1, val-P2 and val-P3) and the PROSOPE-Med cruise (learn-Pmed) have been superposed.

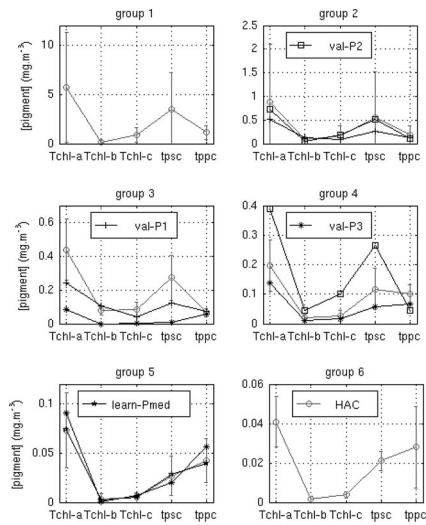


Fig. 8. Group clustering of the mean pigment concentration in samples from the first optical layer. The mean pigment concentrations of the samples from the three POMME cruises (val-P1, val-P2 and val-P3) and the PROSOPE-Med cruise (learn-Pmed) are displayed. We have also displayed the mean pigment concentrations and their standard deviation computed from the learning data set for each group (open circles).

cruises (estimated from the samples of V_{surf}). The corresponding slopes of the spectra are displayed in Fig. 7.

POMME-1 and POMME-2 spectra mainly belong to group 2, whereas POMME-3 data mostly belong to group 5 (Table 2). This denotes a strong difference in phytoplankton optical properties between POMME-1 and POMME-2 with respect to POMME-3.

In the following we used the SOM + HAC visualization in order to analyze the result of the classification for each campaign.

The amplitude of the mean absorption spectrum of POMME-1 is lower (Fig. 6) than that of group 2 surface water samples. This is in agreement with the low mean pigment concentrations of POMME-1 with respect to that of surface waters of group 2, as presented in Fig. 8. In Fig. 9, we plotted the four pigment concentrations relative to Tchl-a for the six groups. The POMME-1 Tchl-b/Tchl-a ratio is particularly high compared to that of the other groups, in agreement with Bricaud *et al.* [14]. Figure 10 shows that the amplitude of the mean group 2 absorption spectrum for the full learning data set, L , is smaller than that of the L_{surf} data set and close to that for POMME-1.

We should recall that the POMME-1 cruise was conducted from January to March, just before the phytoplankton spring bloom [13,15]. During that period the depth of the mixed layer exceeded 120 m [13,16]. This led us to tentatively interpret the behavior described above according to the following scenario: since the mixed layer was deep, the POMME-1 phytoplankton species are probably deep species, as shown by the high Tchl-b/Tchl-a ratio. (It should be noted that the Tchl-b index also contains divinyl-

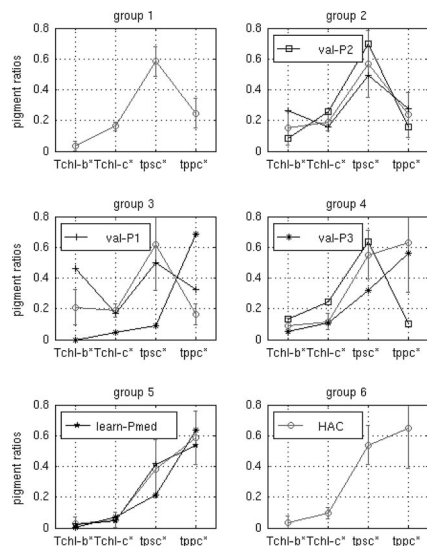


Fig. 9. Group clustering of the mean normalized pigment concentration in samples from the first optical layer. The mean normalized pigment concentrations of the samples from the three POMME cruises (val-P1, val-P2 and val-P3) and the Prosopé-Med cruise (learn-Pmed) are displayed. We have also displayed the mean normalized pigment concentrations and their standard deviation computed from the learning data set for each group (open circles).

chlorophyll-b which is the indicator of prochlorophytes, a not necessarily deep species.) The low amplitude of the absorption spectrum and the pigment concentration is due to the fact that POMME-1 was conducted before the spring bloom, when the phytoplankton concentration was low.

The POMME-2 cruise was conducted from March to May, just after the spring bloom. The mixed layer during that period was much shallower than during POMME-1. The fact that the POMME-1 and POMME-2 spectra are both clustered in group 2 shows that the POMME-2 spectrum is similar to that of POMME-1 (Fig. 6). This is confirmed by the spectrum derivatives that are both closely related to the group 2 reference vector (Fig. 7). The simi-

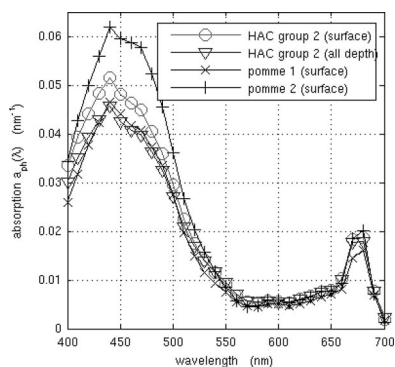


Fig. 10. Mean absorption spectrum of the group 2 surface waters (o), of the group 2 water column including deep and surface waters, (∇), of the POMME-1 (×) and POMME-2 (+) cruises. The POMME-1 spectrum is close to the mean spectrum of the group 2 which includes deep samples.

ilarity between the POMME-1 and the POMME-2 spectra and spectrum derivatives suggests that the POMME-2 phytoplankton results from the blooming of the POMME-1 phytoplankton population. The major difference between the two spectra and the two pigment means is the amplitude difference as seen in Figs. 6 and 8. This is in agreement with a higher phytoplankton concentration observed during the POMME-2 cruise than during POMME-1.

The POMME-3 cruise took place in September and October. The POMME-3 mean spectrum, which mainly belongs to group 5 (Table 2), is very different from those of POMME-1 and POMME-2 (Fig. 6 for the amplitude; Fig. 7 for the derivatives). Furthermore, the POMME-3 normalized pigment concentrations (Fig. 9), being also very different from those of POMME-1 and POMME-2, we can argue that the phytoplankton species were different during POMME-3. In fact, the POMME-3 cruise was held a sufficient time after the spring bloom, which apparently occurred between the POMME-1 and the POMME-2 cruises, to be uncorrelated with them [13,15].

Since SOM + HAC allows us a pertinent biogeophysical interpretation of results from new cruises whose samples were not included in the learning set, we propose to use the clustering into 6 groups to give a regionalized account of various relations linking phytoplankton absorption spectra and pigment concentrations.

4. Determination of Regional Bio-optical Relations

POMME-3 and PROSOPE Mediterranean surface data were captured by the same neurons of the SOM that belong to the group 5 (Table 2). This is due to the fact that the mean POMME-3 and the PROSOPE phytoplankton absorption spectra (Fig. 6) are close to that of the mean group 5, suggesting that the absorption properties of the phytoplankton in the POMME-3 and PROSOPE Mediterranean surface waters were very similar. The pigment values and their normalized ratio (Figs. 8 and 9) were also closely related.

Based on this affirmation, we considered the 6 groups as a regionalization in the data space which is expected to provide better-fitted relationships linking phytoplankton absorption spectra and pigment con-

Table 3. For Each Group, Performance Evaluation for the SOM + HAC Regression Lines as Well as the s^2 and rmse of the Bricaud *et al.* [14] Regression Line

| Group | HAC | | | | | Bricaud <i>et al.</i> [14] | |
|-------|-----------|-------|----------------|----------------|-------|----------------------------|-------|
| | Intercept | Slope | R ² | s ² | rmse | s ² | rmse |
| 1 | -1.121 | 0.604 | 0.731 | 0.108 | 0.106 | 0.115 | 0.113 |
| 2 | -1.204 | 0.465 | 0.438 | 0.112 | 0.112 | 0.131 | 0.130 |
| 3 | -1.470 | 0.327 | 0.721 | 0.051 | 0.049 | 0.156 | 0.151 |
| 4 | -1.303 | 0.637 | 0.753 | 0.084 | 0.083 | 0.100 | 0.099 |
| 5 | -1.315 | 0.647 | 0.770 | 0.072 | 0.072 | 0.083 | 0.083 |
| 6 | -2.160 | 0.176 | 0.046 | 0.091 | 0.085 | 0.243 | 0.227 |

Table 4. Global (on the Whole Data Set) Performance Evaluation for the SOM + HAC Regression Lines and for the Bricaud *et al.* [14] Regression Line

| HAC | | | Bricaud <i>et al.</i> [14] | | |
|----------------|----------------|-------|----------------------------|----------------|-------|
| R ² | s ² | rmse | R ² | s ² | rmse |
| 0.961 | 0.092 | 0.092 | 0.934 | 0.113 | 0.113 |

centrations at regional scale. Indeed, initially, the concept behind regionalization was to adapt the parameters of some algorithm to a specific region of the world ocean. Recently, Alvain *et al.* [17] proposed a “regionalization” of the OC4V4 SeaWiFS algorithm based on classes of normalized radiance. Similarly, our regionalization is based on resemblances of phytoplankton absorption spectra linked to pigment con-

centrations. In the following subsection we revisit some absorption:concentration relationships using it.

A. Relation between $a_{ph}(440)$ and Tchl-a

Bricaud *et al.* [14] proposed a widely used relationship allowing computation of the phytoplankton absorption coefficient at 440 nm from the Tchl-a concentration (for the first optical depth waters), which is of the form

$$a_{ph}(440) = 0.0654Tchl-a^{0.728}$$

(with R² = 0.934, N = 596). (1)

We estimated a similar power-law relationship for each SOM + HAC group. The regression lines for the 6 groups are presented in Fig. 11, concomitant with the Bricaud *et al.* [14] regression line. The group

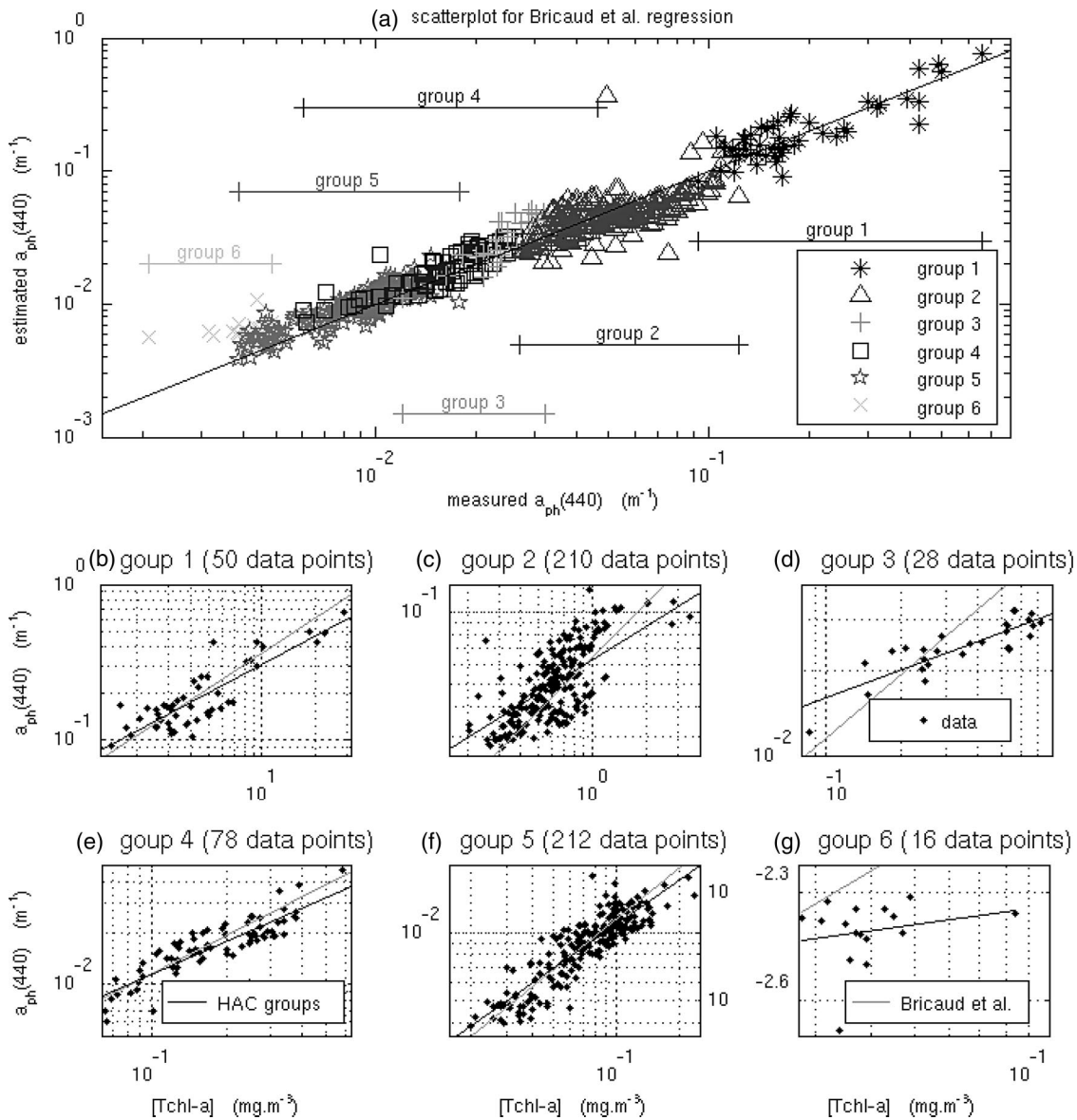


Fig. 11. (a) Scatterplot of the Bricaud *et al.* [14] regression and (b)–(g) Bricaud *et al.* [14] and HAC-group relationships linking absorption to Tchl-a for the six groups.

regression lines better fit the data than the Bricaud *et al.* [14] regression does. This is confirmed by the fact that the rms computed for each group regression line is much smaller than that computed using the Bricaud *et al.* [14] regression line (Tables 3 and 4). These six regressions constitute a piecewise fitting of the relationship. The R^2 estimator of this piecewise fitting was 0.961, which is larger than the corresponding Bricaud *et al.* [14] value of 0.934, showing that the Bricaud *et al.* [14] relationship is valid at the first-order level but is unable to fit second-order non-linearity. Furthermore, Fig. 11 clearly shows that the group 2 relationship, which has a wavelike shape, is not a power law and is governed by a more complex relationship. The partition of the data set into groups allowed us to refine the exploration and the understanding of the behavior of that data set.

B. Relation between the Derivative of $a_{ph}(640)$ and the Tchl-b/Tchl-a Ratio

Chazottes *et al.* [4] proposed a new empirical relationship between the Tchl-b/Tchl-a ratio and the derivative of the absorption spectrum at 640 nm (whenever Tchl-b is not zero). This relationship has been estimated on the global learning data set L . It is of the form

$$\begin{aligned} \text{Tchl-b/Tchl-a} &= 0.090 (a_{650}/a_{640})^{6.838} \\ &\text{(with } \mathbf{R}^2 = 0.531 \text{ and } \mathbf{s} = 0.246). \end{aligned} \quad (2)$$

We have computed a similar relationship using the samples from the first optical layer. It is of the form

$$\begin{aligned} \text{Tchl-b/Tchl-a} &= 0.0684 (a_{650}/a_{640})^{6.624} \\ &\text{(with } \mathbf{R}^2 = 0.431 \text{ and } \mathbf{s} = 0.240). \end{aligned} \quad (2')$$

We estimated similar relationships calibrated for the surface samples of each group. The objective was to obtain specific regional regressions that should be more accurate in each group than the overall regression. The specific regressions are much better than the overall regression for groups 1 and 2 ($\mathbf{R}^2 = 0.656$ and $\mathbf{s} = 0.174$ and $\mathbf{R}^2 = 0.747$ and $\mathbf{s} = 0.142$, respectively). The goodness of fit drops dramatically for the other groups ($\mathbf{R}^2 = 0.36$ for group 3, $\mathbf{R}^2 = 0.38$ for group 4, $\mathbf{R}^2 = 0.26$ for group 5 and $\mathbf{R}^2 = 0.002$ for group 6). This means that the relationship (2') is mainly driven by data belonging to groups 1 and 2. The goodness of fit decreases with the group number. This can be due to the fact that the pigment concentrations strongly decrease with the group number, the relation is valid only above a given concentration threshold.

As mentioned above, POMME-1 and POMME-2 absorption spectra belong to group 2, while POMME-3 spectra belong to group 5. Figure 12 shows the new relationships calibrated on the group 2 L_{surf} data set and on the group 2 surface samples, respectively. We note that the group 2 (POMME-1 and POMME-2) data better fit the new regression calibrated on group

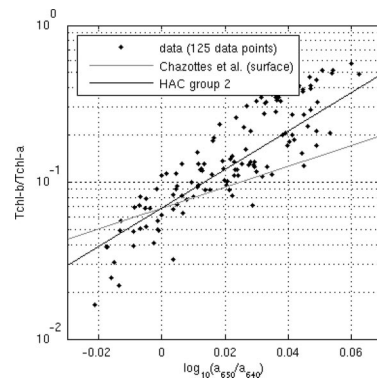


Fig. 12. Plot of the relation given by Eq. (2') on the group 2 data for samples from the first optical layer. The group 2 data displayed are from the validation data set, V_{surf} (POMME-1 and POMME-2).

2 surface data than the regression line calibrated on L_{surf} .

C. Relationship between the Derivative of $a_{ph}(500)$ and the Fucoxanthin/Tchl-a Ratio

We did a similar study for the fucoxanthin/Tchl-a ratio and the derivative of the absorption spectrum at 510 nm. As in the previous section, we reestimated the regression proposed by Chazottes *et al.* [4] on the samples in the first optical layer (L_{surf}), which is of the form

$$\begin{aligned} \text{Fucoxanthin/Tchl-a} &= 1.311 (a_{510}/a_{500})^{6.74} \\ &\text{(with } \mathbf{R}^2 = 0.46, \mathbf{s} = 0.264). \end{aligned} \quad (3)$$

We estimated a similar power-law relationship for each SOM + HAC group, using its specific surface data. Only the group 2 regression presents a better goodness of fit ($\mathbf{R}^2 = 0.685$ and $\mathbf{s} = 0.193$) than that estimated from the learning surface data set, L_{surf} . Figure 13, which is similar to Fig. 12, shows the new relationships, together with the surface data from the POMME-1 and POMME-2 cruises. The POMME-1 and POMME-2 surface data better fit the new regression calibrated on group 2 surface data than the re-

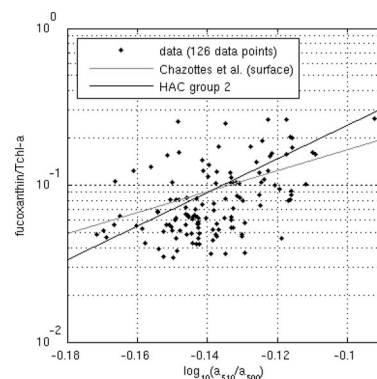


Fig. 13. Plot of the relation given by Eq. (3) on the group 2 data for samples from the first optical layer. The group 2 data displayed are from the validation data set, V_{surf} (POMME-1 and POMME-2).

gression line calibrated on the global learning surface data set, L_{surf} .

5. Summary and Conclusions

By processing the absorption spectra of the LOV database with the SOM + HAC algorithm, we were able to extract pertinent information from the algal absorption spectra. Focus was placed on the POMME experiment. The SOM algorithm reduced the absorption spectra set to a number (10×10) of reference absorption spectra considering specific statistical characteristics without any *a priori* information. As it is not always easy to extract relevant information from such a large number of classes, the SOM + HAC method aggregates them into a small number of groups (6 in the present study) according to their spectral similarities. We were then able to extract some knowledge from the database thanks to the ability of the group decomposition to extract information embedded in the database. In particular this analysis led us to propose regional empirical relationships, linking phytoplankton absorption spectra to pigment concentrations, that are better than the previously derived global ones.

The clustering was done with respect to the spectrum amplitude (Fig. 6) and spectrum shape (Fig. 7). The efficiency of the clustering is evident in the spectrum derivatives (Fig. 7). The clustering also coherently clustered the various phytoplankton pigment concentrations and normalized pigments, as shown in Figs. 4 and 8.

The data of the different POMME cruises (which were not learned and constituted a validation set) were assigned to coherent groups. We were able to show the different behavior of POMME-1 and POMME-2 with respect to POMME-3, not only in terms of optical properties, but also of phytoplankton pigment concentrations. Most of the POMME-1 and POMME-2 data were clustered in group-2, whereas the POMME-3 data are in group-5. POMME-1 presents a very high Tchl-b/Tchl-a ratio, as is usually found in deep water samples. The TPSC/Tchl-a ratio is higher than the TPPC/Tchl-a for POMME-1 and POMME-2, whereas the opposite holds for POMME-3 and PROSOPE-Med and more generally for samples in groups 4, 5, and 6. We showed the influence of physical parameters associated with the season, such as the depth of the thermocline or the phytoplankton maturity on the different POMME cruises. But we were not able to explain the similarities between POMME-3 and PROSOPE-Med in oceanographic terms.

We revisited the empirical relationship proposed by Bricaud *et al.* [14] linking the absorption at 440 nm to Tchl-a by computing a specific regression for each group. These different regressions constituted a piecewise fitting of the relationship, which better fits the data than does the Bricaud *et al.* [14] regression, showing that there exists some second-order nonlinearity which the Bricaud *et al.* [14] regression does not account for. This nonlinearity is evident for the group 2 data presented in Fig. 11.

We also revisited the Chazottes *et al.* [4] relationships linking the Tchl-b/Tchl-a ratio to the derivative of the absorption spectrum at 640 nm and the fucoxanthin/Tchl-a ratio to the derivative of the absorption spectrum at 510 nm. The aim was to propose an improved specific relationship for each group. We found that these two relationships were significantly better for group 2 data only and that there was a somewhat improved relationship for the Tchl-b/Tchl-a ratio for group-1 data. Interpretation may be sought in the decrease in the pigment concentrations as the group number increases and consequently an implicit threshold effect on the pigment concentration, the relationships being valid above a certain concentration only. Another reason might be the phytoplankton diversity associated with the different groups as shown by the analysis presented in Appendix A.

Appendix A

We divided the samples into three water types according to Tchl-a concentrations, based on their Tchl-a concentrations. For the sake of simplification, we will hereafter call “oligotrophic” those waters with $[\text{Tchl-a}] < 0.2 \text{ mg.mg.m}^{-3}$, “mesotrophic” those with $[\text{Tchl-a}]$ between 0.2 mg.m^{-3} and 1 mg.m^{-3} , and “eutrophic” those with $[\text{Tchl-a}] > 1 \text{ mg.m}^{-3}$. Applying the SOM – HAC algorithm gave the number of samples, for each water type, captured by each group: group 1 contained eutrophic waters only; group 2, and 3 were predominantly mesotrophic waters. These three groups are contiguous on the SOM (Fig. 14). Group 4 comprised a mixture of mesotrophic and oligotrophic waters, and groups 5 and 6 contained oligotrophic water only.

In Fig. 4, the normalized total photosynthetic carotenoids (TPSCs) are much larger than the nor-

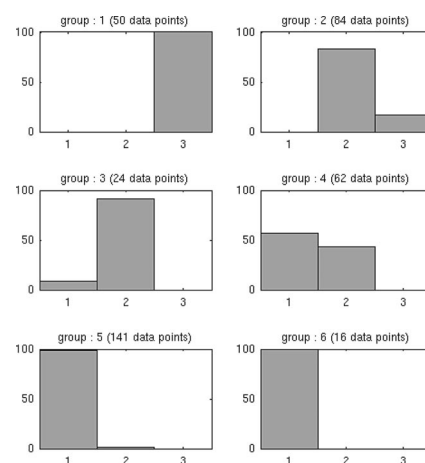


Fig. 14. Histogram of the distribution of the surface-water samples from the water types in each of the six groups. The histogram is computed as the ratio of the number of samples of a given water type in a given group to the total number of samples in that group. The number of water samples is given for each group. 1, 2, and 3 stand for the different water types, respectively, oligotrophic, mesotrophic, and eutrophic.

malized total photoprotectant carotenoids (TPPCs) for the first three groups (1, 2, 3), which is a characteristic of eutrophic and mesotrophic surface waters; the contrary holds for the other three groups (4, 5, 6) and is a characteristic of deep oligotrophic waters. Besides, the group-1 data came mainly from the Bencal campaign, which was conducted in a highly productive area with typical phytoplankton species. Group 2 data, which came from several different campaigns (Fig. 5) are mainly associated with mesotrophic waters (Fig. 14), whereas groups 4, 5, and 6 waters are oligotrophic. The mean absorption spectrum of group 2 being very different in amplitude and shape from those of groups 4, 5, and 6, we may argue that the phytoplankton composition is quite different from that of groups 4, 5, and 6.

Therefore the SOM + HAC groups correspond to a water type and to a mean pigmentary composition.

Appendix B: Linear Regression

To estimate the quality of a linear or log-linear relationship, two parameters, R^2 and s , are usually calculated. Let us consider a data set of n samples $(x_i, y_i^{\text{observed}})$. The determination of the parameter R^2 , which is a measure of “goodness of fit,” represents the part of the variation in y explained by x . It is given by

$$R^2 = \frac{\sum (y_i^{\text{estimated}} - \bar{y})^2}{\sum (y_i^{\text{observed}} - \bar{y})^2}, \quad (\text{B1})$$

where \bar{y} is the mean of the y_i^{observed} . The parameter s , which is referred to as the root mean square (rms), is an estimate of the error on y . It is given by

$$s = \left(\frac{\sum (y_i^{\text{observed}} - y_i^{\text{estimated}})^2}{n - 2} \right)^{1/2}. \quad (\text{B2})$$

This work is a contribution to the European project NAOC (EVG1-CT-2000-00034) and to several projects which have been funded by the PROOF national programme (EUMELI, EPOPE, FRONTAL, PROSOPE, POMME). The authors thank the chief scientists of the cruises during which the *in situ* data were collected, and all the colleagues who participated in the sample collection, HPLC measurements, or absorption measurements (H. Claustre, K. Oubelkheir, K. Allali, C. Cailliau, J. C. Marty, N. Sadoudi, D. Tailliez, F. Vidussi). We also thank the two anonymous reviewers and R. Griffiths for stimulating discussions.

References

1. M. Saraceno, C. Provost, and M. Lebbah, “Biophysical regions identification using an artificial neuronal network: a case study in the South Western Atlantic,” *Adv. Space Res.* **37**, 793–805 (2006).

2. Y. Liu and R. H. Weisberg, “Patterns of ocean current variability on the West Florida Shelf using the self-organizing map,” *J. Geophys. Res.* **110**, C06003, doi:10.1029/2004JC002786 (2005).
3. Y. Liu, R. H. Weisberg, and R. He, “Sea surface temperature patterns on the West Florida Shelf using growing hierarchical self-organizing maps,” *J. Atmos. Oceanic Technol.* **23**, 325–338 (2006).
4. A. Chazottes, A. Bricaud, M. Crépon, and S. Thiria, “Statistical analysis of a database of absorption spectra of phytoplankton and pigment concentrations using self-organizing maps,” *Appl. Opt.* **45**, 8102–8115 (2006).
5. T. Kohonen, *Self-Organizing Maps* (Springer Verlag, 1984).
6. A. Bricaud, M. Babin, A. Morel, and H. Claustre, “Variability in the chlorophyll-specific absorption coefficients of natural phytoplankton: analysis and parameterization,” *J. Geophys. Res.* **100**, 13331–13332 (1995).
7. A. Bricaud, A. Morel, M. Babin, K. Allali, and H. Claustre, “Variations of light absorption by suspended particles with chlorophyll a concentration in oceanic (case 1) waters: analysis and implications for bio-optical models,” *J. Geophys. Res.* **103**, 31033–31044 (1998).
8. F. Vidussi, H. Claustre, J. Bustillos-Guzman, C. Cailliau, and J. C. Marty, “Rapid HPLC method for determination of phytoplankton chemotaxonomic pigments: separation of chlorophyll a from divinyl-chlorophyll-a, and zeaxanthin from lutein,” *J. Plankton Res.* **18**, 2377–2382 (1996).
9. A. Morel and S. Maritorena, “Bio-optical properties of oceanic waters: a reappraisal,” *J. Geophys. Res.* **106**, 7763–7780 (2001).
10. A. Niang, L. Gross, S. Thiria, F. Badran, and C. Moulin, “Automatic neural classification of ocean colour reflectance spectra at the top of the atmosphere with introduction of expert knowledge,” *Remote Sens. Environ.* **86**, 257–271 (2003).
11. Y. Liu, R. H. Weisberg, and C. N. K. Mooers, “Performance evaluation of the self-organizing map for feature extraction,” *J. Geophys. Res.* **111**, C05018, doi:10.1029/2005JC003117 (2006).
12. G. Dreyfus, *Neural Networks: Methodology and Applications* (Springer-Verlag, 2005).
13. L. Mémery, G. Reverdin, S. Paillet, and A. Oschlies, “Introduction to the POMME special section: thermocline ventilation and biogeochemical tracer distribution in the northeast Atlantic Ocean and impact of mesoscale dynamics,” *J. Geophys. Res.* **110**, C07S01, doi: 10.1029/2005JC002976 (2005).
14. A. Bricaud, H. Claustre, J. Ras, and K. Oubelkheir, “Natural variability of phytoplanktonic absorption in oceanic waters: influence of the size structure of algal populations,” *J. Geophys. Res.* **109**, C11010, doi:10.1029/2004JC002419 (2004).
15. M. Levy, Y. Lehahn, J.-M. Andre, L. Mémery, H. Loisel and E. Heifetz, “Production regimes in the Northeast Atlantic: a study based on Sea-viewing Wide Field-of-view Sensor (SeaWiFS) chlorophyll and ocean general circulation model mixed layer depth,” *J. Geophys. Res.* **110**, C07S10, doi:10.1029/2004JC002771 (2005).
16. A. Paci, G. Caniaux, M. Gavart, H. Giordani, M. Lévy, L. Prieur, and G. Reverdin, “A high-resolution simulation of the ocean during the POMME experiment: simulation results and comparison with observations,” *J. Geophys. Res.* **110**, C07S09, doi:10.1029/2004JC002712 (2005).
17. S. Alvain, C. Moulin, Y. Dandonneau, H. Loisel, and F.-M. Bréon, “A species-dependent bio-optical model of case I waters for global ocean color processing,” *Deep-Sea Res.* **153**, 917–925 (2006).