- Phytoplankton pigment assemblages in the open ocean identified using hyperspectral data
- Potential usefulness of derivative spectroscopy of absorption and remote-sensing reflectance
- Cluster-based approach to identify the most suitable spectral ranges and derivative parameters

Cluster analysis of hyperspectral optical data for discriminating phytoplankton pigment assemblages in the open ocean

E. Torrecilla[1], D. Stramski[2], R. A. Reynolds[2], E. Millán-Núñez[2,3], and J. Piera[1]

[1]*Mediterranean Marine and Environmental Research Centre, Marine Technology Unit (UTM-CSIC), Pg. Marítim Barceloneta 37, Barcelona 08003, Spain (*torrecilla@utm.csic.es*)*

[2]*Marine Physical Laboratory, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093-0238, USA*

[3]*CICESE, Ecología Marina, Carretera Ensenada-Tijuana No. 3918 Fraccionamiento Zona Playitas, Ensenada, B. C., Código Postal 22860, México*

Keywords: Ocean optics, ocean reflectance, absorption coefficient, phytoplankton pigments, cluster analysis

Abstract

Optical measurements including remote sensing provide a potential tool for the identification of dominant phytoplankton groups and for monitoring spatial and temporal changes in biodiversity in the upper ocean. We examine the application of an unsupervised hierarchical cluster analysis to phytoplankton pigment data and spectra of the absorption coefficient and remote-sensing reflectance with the aim of discriminating different phytoplankton assemblages in open ocean environments under non-bloom conditions. This technique is applied to an optical and phytoplankton pigment data set collected at several stations within the eastern Atlantic Ocean, where the surface total chlorophyll-*a* concentration (TChl*a*) ranged from 0.11 to 0.62 mg m^{-3}. Stations were selected on the basis of significant differences in the ratios of the two most dominant accessory pigments relative to TChl*a*, as derived from High Performance Liquid Chromatography (HPLC) analysis. The performance of cluster analysis applied to absorption and remote-sensing spectra is evaluated by comparisons with the cluster partitioning of the corresponding HPLC pigment data, in which the pigment-based clusters serve as a reference for identifying different phytoplankton assemblages. Two indices, cophenetic and Rand, are utilized in these comparisons to quantify the degree of similarity between pigment-based and optical-based clusters. The use of spectral derivative analysis for the optical data was also evaluated, and sensitivity tests were conducted to determine the influence of parameters used in these calculations (spectral range, smoothing filter size, band separation). The results of our analyses indicate that the second derivative calculated from hyperspectral (1 nm resolution) data of the phytoplankton absorption coefficient, $a_{ph}(\lambda)$, and remote-sensing reflectance, $R_{rs}(\lambda)$, provide better discrimination of phytoplankton pigment assemblages than traditional multispectral band-ratios or ordinary (non-differentiated) hyperspectral data of absorption and remote-sensing reflectance. The most useful spectral region for this discrimination extends generally from

wavelengths of about 425 - 435 nm to wavelengths within the 495 - 540 nm range, although in the case of phytoplankton absorption data a broader spectral region can also provide satisfactory results.

1. Introduction

In situ and remotely-sensed optical observations of ocean waters provide information regarding the concentrations of optically significant constituents in seawater, and offer the ability to observe important biological and biogeochemical variables (e.g., Chang et al., 2006). Numerous studies over the past three decades have focused on the development of bio-optical algorithms linking measurable optical properties to the primary pigment in phytoplankton, chlorophyll-*a*, a proxy for the phytoplankton biomass (e.g., Morel, 1988; Bricaud et al., 1998; O'Reilly et al., 2000; Reynolds et al., 2001). In recent years, efforts to expand the use of optical measurements for estimating other biogeochemically important ocean variables and phenomena have increased considerably. For example, optical measurements including satellite remote sensing have been used to detect harmful algal blooms (Cullen et al., 1997; Stumpf et al., 2003), surface concentrations of particulate inorganic and organic carbon (Balch et al., 2005; Stramski et al., 2008), particle size distribution (Kostadinov et al., 2009), phytoplankton community composition and size structure (Alvain et al., 2005; Uitz et al., 2006; Nair et al., 2008; Ciotti and Bricaud, 2006; Aiken et al., 2007), and phytoplankton class-specific primary production (Uitz et al., 2010).

Recent advances in measuring ocean optical properties and light fields within and leaving the ocean have included a progressive shift from using multispectral to high spectral resolution (hyperspectral) acquisition systems (Chang et al., 2004). New technologies and the miniaturization of electro-optical components have permitted the development of accurate, low-cost, and energy-efficient hyperspectral sensors suitable for deployments from a variety of platforms such as in-water vertical profiling systems, moorings, drifters, autonomous vehicles, air-borne and space-borne platforms (Perry & Rudnick, 2003; Dickey et al., 2006). The capability to obtain measurements at hundreds of narrow and closely spaced wavelength bands from the ultraviolet to near-infrared, with a resolution better than 10 nm, has become one of the

most powerful and fastest growing areas of technology in the field of ocean optics.

Hyperspectral optical data provide the opportunity for improvements in spectral shape analysis and subsequent extraction of environmental information compared with low spectral resolution optical data. Derivative spectroscopy is one powerful technique of spectral shape analysis which enhances subtle features in hyperspectral data, and has been used successfully to obtain information about optically significant water constituents. For example, Craig et al. (2006) assessed the feasibility of detection of a toxic bloom of the dinoflagellate *Karenia brevis* from the analysis of the fourth derivative of phytoplankton absorption spectra, estimated from in situ hyperspectral measurements of remote-sensing reflectance $R_{rs}(\lambda)$ (λ is light wavelength in vacuo). The advantages offered by hyperspectral measurements of $R_{rs}(\lambda)$ in combination with derivative spectroscopy for identifying algal blooms were also demonstrated by Lubac et al. (2008), who based their analysis on the position of the maxima and minima of the second derivative of the spectral $R_{rs}(\lambda)$. Louchard et al. (2002) assessed major sediment pigments of benthic substrates from derivative spectra of hyperspectral $R_{rs}(\lambda)$ measured in shallow marine environments. In general, the optical detection of specific algal blooms appears feasible because certain accessory pigments with specific absorption features are unique to individual phytoplankton taxa and can be better differentiated in hyperspectral absorption data than in multispectral data with a limited number of wavelengths.

The advantages and increasing availability of high spectral resolution measurements suggest that the effectiveness of hyperspectral optical information for assessing phytoplankton diversity should be further explored. In particular, there is a need to test whether the hyperspectral approach, which has proven useful in inland and coastal waters (e.g., Lee & Carder, 2004; Hunter et al., 2008; Lubac et al., 2008), can be also effective for the identification of different phytoplankton assemblages at large spatial scales in open ocean waters. These tests are also especially important for the common situation in which various phytoplankton groups co-exist at significant concentrations, and no single species dominates the assemblage (i.e., a non-bloom

79 condition).

80 In this study, we analyze phytoplankton pigment data in conjunction with optical data of
81 absorption coefficients and remote-sensing reflectance, which were determined along a north-to-
82 south transect in the eastern Atlantic Ocean. Our primary goal is to examine the feasibility of
83 classifying different open ocean environments under non-bloom conditions in terms of
84 phytoplankton pigment assemblages from analysis of hyperspectral absorption and remote-
85 sensing reflectance measurements. In order to address this question, an unsupervised
86 hierarchical cluster analysis is applied to the pigment data set obtained from High Performance
87 Liquid Chromatography (HPLC) analysis of seawater samples and to the optical data sets
88 including the spectra of absorption coefficients and remote-sensing reflectance and their second
89 derivative spectra. For this analysis, the pigment data and the corresponding optical data were
90 selected to represent distinct differences in major accessory pigments present in the samples. We
91 view our analysis basically as a proof-of-concept study in which our approach is to use a
92 relatively small but carefully selected set of data which exhibits significant contrasts in the
93 composition of pigments, rather than to indiscriminately use large data sets. The pigment-based
94 clusters provide a reference for partitioning the selected data set into distinct subsets, each
95 characterized by different phytoplankton pigment composition. Two indices, cophenetic and
96 Rand, are examined to quantify the degree of similarity between the pigment-based clusters and
97 optical-based clusters, and are ultimately used to illustrate the effectiveness of optical
98 classification. The degree of similarity between clusters was evaluated for calculations involving
99 different spectral ranges of optical data. Because the quality of derivative analysis also depends
100 on parameters involved in data processing and computations, especially smoothing filter size and
101 derivative band separation (Tsai & Philpot, 1998; Lee & Carder, 2002; Vaiphasa, 2006), a
102 sensitivity of cluster analysis to the choice of these parameters was performed.
103

2. Measurements and Data Analysis

The approach in this study consists of three main components: (i) collection of field data of phytoplankton pigments and ocean optical properties and selection of a subset of data characterized by distinct differences in major accessory pigments for the cluster analysis, (ii) radiative transfer modeling to compute hyperspectral remote sensing-reflectance, and (iii) cluster analysis of pigment and optical data. The methodology of each component is described below.

2.1. Field measurements

Measurements of phytoplankton pigment composition and seawater optical properties were obtained during the ANT-XXIII/1 expedition of the R/V Polarstern along a north-to-south transect in the eastern Atlantic Ocean during October and November, 2005 (Fig. 1). The investigated area spanned a wide range of different oceanic environments between the English Channel and the waters off the African coast of Namibia. Typically, one full station was conducted daily near local noon throughout the cruise. These full stations consisted of in situ measurements of seawater inherent and apparent optical properties along with laboratory analyses of water samples collected from discrete depths with the ship's CTD/rosette system. For the present study, a subset of nine stations (see Fig. 1 for station locations) was selected for cluster analysis based on the observation of distinct differences in the ratios of dominant accessory pigments to total chlorophyll-*a* (further details in sec. 3.1). The selected data from the nine stations are representative of surface waters within the top 5-10 m of the ocean, as our main interest lies in the methodology for estimating variability in phytoplankton communities from remote-sensing reflectance. A brief description of the measurements is provided in the following three subsections. More methodological details, especially for the radiometric and backscattering measurements, can be found in Stramski et al. (2008).

2.1.1. HPLC pigment analysis

Concentrations of chlorophyll-*a* and accessory pigments in phytoplankton were measured on surface water samples from each station using HPLC techniques. Two sets of replicate samples were collected and analyzed at two laboratories, the Center for Hydro-Optics and Remote Sensing (CHORS) laboratory at San Diego State University (California, USA) and the GKSS Research Centre in Geesthacht (Germany). The CHORS analysis was based on a method described in Van Heukelem and Thomas (2001), and the GKSS samples were analyzed following the method of Zapata et al. (2000). The CHORS analysis included identification and quantification of more pigments (27) than the GKKS method (23) including alternative forms of chlorophyll-*a*, and was chosen as the primary pigment data set for identifying phytoplankton assemblages using cluster analysis. Throughout the rest of this paper, as a measure of chlorophyll-*a* we use the CHORS values of the total chlorophyll-*a*, TChl*a*, which is defined as the summed contributions of concentrations of monovinyl chlorophyll-*a* (MVChl*a*), divinyl chlorophyll-*a* (DVChl*a*), chlorophyllide-*a* (Chlide), and the allomeric and epimeric forms of chlorophyll-*a*.

Following completion of our analyses, the potential for errors in the CHORS results was identified. The NASA team tasked with investigating HPLC data quality problems at CHORS recommended that overall "These data are not validated and should not be used as sole basis for a scientific result, conclusion, or hypothesis – independent corroborating evidence is required" (Hooker and Van Heukelem, 2009). Based on field data obtained in the SeaHARRE-3 intercalibration experiment, corrections for specific individual pigments (MVChla, DVChla) were developed (C. Trees, personal communication), which are used in our data set in the determination of TChl*a*. These corrections are described in Stramski et al. (2008). The corrected TChl*a* data exhibit reasonable agreement with fluorometrically-derived chlorophyll measurements, and provide realistic estimates of chlorophyll-specific phytoplankton absorption

154 coefficients within the red peak of chlorophyll-*a*. We also applied similar corrections to three
155 other pigments in the data set; monovinyl chlorophyll-*b*, *β*-carotene, and alloxanthin. We caution
156 that our corrections were developed from a limited set of pigment data and intercalibration
157 results, and should not be used indiscriminately with other data sets affected by the CHORS data
158 quality problems.

159 We also compared the corrected CHORS results with independent pigment determinations
160 done by GKSS. Some differences in the concentrations of individual pigments between the two
161 data sets were observed. For example, the sum of monovinyl chlorophyll-*a* (MVChl*a*) and
162 divinyl chlorophyll-*a* (DVChl*a*) was generally higher (on average by 20% with a standard
163 deviation of 23%, number of samples 25) for CHORS compared with GKSS. However, the
164 CHORS data yielded more reasonable estimates of chlorophyll-specific phytoplankton
165 absorption.

166 Despite such differences in the estimates of some individual pigment concentrations, with
167 regard to the present application it is important to note that both laboratories provided similar
168 characterization of samples in terms of the relative pigment composition as described by ratios of
169 various individual pigments to TChl*a*. The same dominant accessory pigments for any given
170 station and the same trends in the pigment ratios among the stations were obtained from both sets
171 of HPLC analyses. This is an essential result for our study because in the cluster analysis we
172 utilize only pigment ratios, and not individual pigment concentrations. For the nine stations
173 selected in our analysis, the cluster techniques applied independently to the CHORS and GKSS
174 sets of HPLC pigment ratios yielded a very similar partitioning of stations into clusters. We
175 therefore chose the one set of corrected pigment results from CHORS for all subsequent analyses
176 in this paper (sec. 3.1 presents CHORS-based pigment clusters). This choice is further supported
177 by comparisons of cluster analysis of phytoplankton absorption data with the CHORS and
178 GKSS-based pigment clusters. The measures of cluster similarity (see sec. 2.3) between the
179 absorption-based clusters and pigment-based clusters were found to be similar when either the

CHORS and GKSS pigment data were used.

Because of a large range of pigment compositions across different phytoplankton classes, determinations of phytoplankton composition from HPLC pigment data is not straightforward (e.g., Jeffrey et al., 1997). Whereas certain diagnostic pigments can serve as unambiguous markers for some phytoplankton classes (e.g., peridinin in dinoflagellates, alloxanthin in cryptophytes), many important pigments are shared by more than one algal taxa (e.g., fucoxanthin in diatoms, haptophytes, chrysophytes, and raphidophytes). Nevertheless, because many of the classes have distinctive suites of marker pigments, HPLC data can be useful for indicating their presence and abundance in a mixed phytoplankton population. Specifically, a useful indication of contributing phytoplankton classes can be obtained from the ratios of the concentrations of specific pigments to chlorophyll-*a* or the ratios of specific diagnostic pigments to the sum of these diagnostic pigments, because these ratios can differ between taxonomic groups (Mackey et al., 1996; Wright et al., 1996; Vidussi et al., 2001).

For each of the nine stations selected in this study, we calculated two sets of pigment ratios for subsequent use in the cluster analysis. The first set of pigment ratios consisted of ratios of the concentration of each individual pigment to the TChl*a* concentration, as obtained from HPLC measurements at the CHORS laboratory. The following 24 pigments were included in these calculations: monovinyl chlorophyll-*a*, divinyl chlorophyll-*a*, chlorophyllide-*a*, chlorophyll-*a* allomer, chlorophyll-*a* epimer, monovinyl chlorophyll-*b*, divinyl chlorophyll-*b*, chlorophyll-*c2*, chlorophyll-*c3*, α-carotene, β-carotene, alloxanthin, diadinoxanthin, diatoxanthin, fucoxanthin, 19'-hexanoyloxyfucoxanthin, 19'-butanoyloxyfucoxanthin, neoxanthin, prasinoxanthin, violaxanthin, zeaxanthin, peridinin, pheophorbide-*a*, and lutein. The CHORS pigment data set also included a few additional pigments (chlorophyll-*c1*, gyroxanthin-diester, and pheophytin-*a*) which were below the level of detection for the nine stations. These pigments were not included in our cluster analysis as they would have no effect on the results.

205 The second set of pigment ratios was based on 8 diagnostic pigments: divinyl chlorophyll-*a*
206 (DVChl*a*), divinyl chlorophyll-*b* (DVChl*b*), alloxanthin (Allo), fucoxanthin (Fuco), 19'-
207 hexanoyloxyfucoxanthin (Hex), 19'-butanoyloxyfucoxanthin (But), peridinin (Peri), and
208 zeaxanthin (Zea). This list of pigments is consistent with that proposed by Vidussi et al. (2001)
209 with the exception that we added DVChl*a* which is a diagnostic pigment for prochlorophytes.
210 For each station the 8 ratios were calculated by dividing a concentration of a given diagnostic
211 pigment to the sum of the 8 diagnostic pigment concentrations (see Uitz et al., 2006).
212
213 *2.1.2. Inherent optical properties*
214 The inherent optical properties (IOPs) of seawater have a two-fold application in our study.
215 First, the absorption, scattering, and backscattering coefficients are used to define IOP inputs to
216 radiative transfer simulations that generate hyperspectral data of remote-sensing reflectance (sec.
217 2.2). Second, the total absorption coefficient, $a(\lambda)$, and the phytoplankton absorption coefficient,
218 $a_{ph}(\lambda)$, are utilized directly in the cluster analysis (sec. 2.3 and 3.2).
219 The spectral absorption coefficients of particles, $a_p(\lambda)$, and colored dissolved organic matter
220 (CDOM), $a_{cdom}(\lambda)$ in m^{-1}, were determined at 1-nm intervals from high spectral resolution
221 measurements on freshly-collected discrete water samples with a point-source integrating cavity
222 absorption meter (PSICAM) over the range 350-750 nm (Röttgers et al., 2005, Röttgers &
223 Doerffer, 2007). As the PSICAM did not provide data below 350 nm, the $a_p(\lambda)$ values within the
224 300 to 350 nm spectral range were obtained from filter pad measurements on discrete water
225 samples collected on GF/F filters and frozen in liquid nitrogen until analysis with a dual-beam
226 spectrophotometer (Lambda 18, Perkin Elmer). The filter pad measurements were made with the
227 transmittance-reflectance (T-R) technique of Tassan and Ferrari (1995; 2002) using a correction
228 for the pathlength-amplification factor from Stramska et al. (2006). We have chosen to use the
229 PSICAM data of $a_p(\lambda)$ over the majority of the spectrum because the PSICAM technique
230 involves a direct measurement of absorption on particle suspension with minimal scattering

artifacts, which is expected to be generally superior to the filter pad measurements. The data of $a_{cdom}(\lambda)$ below 350 nm were obtained from an exponential fit to the PSICAM-measured $a_{cdom}(\lambda)$. A null point correction based on wavelengths in the far red or near-infrared was applied to all $a_p(\lambda)$ and $a_{cdom}(\lambda)$ spectra. The total spectral absorption coefficient, $a(\lambda)$, was determined as the sum of $a_p(\lambda)$, $a_{cdom}(\lambda)$, and the pure water component, $a_w(\lambda)$. The latter was obtained from Pope & Fry (1997) for the spectral range 380-727 nm and from Fry et al. (2006) for the range 300-379 nm. We note that our primary interest is in the spectral information contained at wavelengths longer than about 350 nm extending throughout the visible part of the spectrum up to 725 nm where most phytoplankton pigments exhibit significant absorption features. However, data at wavelengths shorter than 350 nm are useful for our analysis, especially in the context of derivative spectra whose discrete values at specific wavelengths were calculated in our study using data covering a bandwidth on the order of 10 - 30 nm.

The spectra of the phytoplankton absorption coefficient, $a_{ph}(\lambda)$, were determined as a difference between the absorption coefficient of particles, $a_p(\lambda)$, and the non-phytoplankton component of particulate absorption, $a_d(\lambda)$, which is commonly referred to as detrital absorption. These determinations were based on the T-R filter pad measurements, in which the $a_d(\lambda)$ spectra were measured on GF/F sample filters following treatment with sodium hypochlorite NaClO (Ferrari and Tassan, 1999). In this treatment, the particles on the sample filter were exposed to a small amount of a 2% NaClO solution for several minutes to bleach phytoplankton pigments.

The spectral beam attenuation coefficient of particles and CDOM, $c_{p,cdom}(\lambda)$ in m^{-1}, was determined at each station from in situ measurements with two single-wavelength C-Star transmissometers (488 and 660 nm; WET Labs, Inc.). Note that the C-star derived values represent the total beam attenuation, $c(\lambda)$, with pure seawater contribution, $c_w(\lambda)$, subtracted, i.e. $c_{p,cdom}(\lambda) = c(\lambda) - c_w(\lambda) = a_p(\lambda) + a_{cdom}(\lambda) + b_p(\lambda)$, where $b_p(\lambda)$ is the spectral scattering coefficient of particles. The values of $b_p(\lambda)$ at 488 and 660 nm were thus calculated from C-Star

attenuation and PSICAM absorption measurements as $b_p(\lambda) = c_{p,cdom}(\lambda) - a_p(\lambda) - a_{cdom}(\lambda)$. A power function fit was then applied to these values to produce the spectral data of $b_p(\lambda)$ over the range 300-750 nm with a 1 nm resolution. The pure seawater scattering coefficient, $b_w(\lambda)$, was calculated using the Buiteveld et al. (1994) equations with measured water temperature and salinity (see Twardowski et al., 2007 and Stramski et al., 2008 for details). The total scattering coefficient, $b(\lambda)$, was obtained as a sum $b_w(\lambda) + b_p(\lambda)$.

The spectral backscattering coefficient, $b_b(\lambda)$ in m^{-1}, was determined by combining in situ measurements with three instruments, a Hydroscat-6 and two a-βeta sensors (HOBI Labs, Inc.), to yield a total of eight spectral bands: 420, 442, 470, 510, 550, 589, 620, and 671 nm (Stramski et al. 2008). Because $b_b(\lambda)$ is generally expected to be a smooth monotonic function of wavelength, especially in the open ocean under non-bloom conditions, the experimental measurements were fitted to a power function to obtain hyperspectral resolution over the 300-750 spectral range. The spectral backscattering coefficient of particles, $b_{bp}(\lambda)$ in m^{-1}, was determined as a difference between the total and pure seawater backscattering coefficients, $b_b(\lambda) - b_{bw}(\lambda)$, in which the pure seawater component, $b_{bw}(\lambda)$, was calculated as $0.5\, b_w(\lambda)$.

From the values of $b_p(\lambda)$ and $b_{bp}(\lambda)$, we calculated the particle backscatter fraction $B_p(\lambda) = b_{bp}(\lambda) / b_p(\lambda)$. These data were then fitted to a power function, $B_p(\lambda) = B_p(\lambda_0)\,(\lambda_0 / \lambda)^m$, where λ_0 is the reference wavelength 550 nm. The backscattering fraction $B_p(\lambda_0)$ at the reference wavelength and the exponent m represent the best fit parameters of the linear regression analysis performed for the log-transformed data of $B_p(\lambda)$ vs. λ for each station. The parameters of the power function fit of $B_p(\lambda)$ were used as input to radiative transfer simulations (see sec. 2.2).

2.1.3. Remote-sensing reflectance

Values of multispectral remote-sensing reflectance, $R_{rs}(\lambda)$ in sr^{-1}, were estimated at each station at 13 wavelengths from in situ measurements of underwater vertical profiles of spectral nadir upwelling radiance, $L_u(\lambda, z)$ in W m^{-2} sr^{-1} nm^{-1}, and spectral downwelling plane irradiance,

282 $E_d(\lambda, z)$ in W m^{-2} nm^{-1}, where z is depth. These measurements were made with a freefall
283 spectroradiometer, the SeaWiFS Profiling Multichannel Radiometer (SPMR, Satlantic, Inc.).
284 The wavelengths for these measurements are 339, 380, 412, 443, 470, 490, 510, 532, 554, 589,
285 619, 666, and 683 nm. The radiometric measurements and data processing were consistent with
286 methods recommended in NASA protocols (Mueller et al., 2003).
287
288 *2.2. Modeled hyperspectral reflectance*
289 Because hyperspectral radiometric measurements were not conducted during the ANT-
290 XXIII/1 cruise and our primary interest is in the analysis of hyperspectral optical data, we
291 performed numerical simulations of radiative transfer (RT) to estimate the hyperspectral remote-
292 sensing reflectance $R_{rs}(\lambda)$ for each of the nine selected stations. The radiative transfer model
293 Hydrolight/Ecolight version 5.0 (Sequoia Scientific, Inc.) was used (Mobley 1994; 2008). An
294 important prerequisite for undertaking these RT simulations was the availability of a
295 comprehensive suite of IOPs for each station for use as input to the simulations, and also the
296 availability of the multispectral $R_{rs}(\lambda)$ derived from in situ measurements for use in validating
297 the simulated hyperspectral $R_{rs}(\lambda)$.
298 The RT calculations were carried out within the spectral region 300 nm to 725 nm with high
299 spectral resolution (1 nm). Similarly to the absorption, our main interest is in the reflectance data
300 at wavelengths longer than about 350 nm. However, in addition to the requirements associated
301 with derivative calculations, the radiative transfer simulations below 350 nm are needed to
302 account for Raman scattering contributions observed at $\lambda > 350$ nm. The ocean was assumed to
303 be infinitely deep and optically homogeneous, and the simulations included the Raman scattering
304 and fluorescence by colored dissolved organic matter within the ocean. The sea surface
305 boundary conditions were estimated from observations of wind speed and sky conditions
306 (cloudiness) at each station site, and the solar zenith angle was calculated for the corresponding

date and geographic coordinates. The inherent optical properties of the water column required as input to the simulations were derived from the IOP measurements in the surface waters and additional relevant determinations of $a(\lambda)$ and $b(\lambda)$ as described in sec. 2.1.2. The selection of the particulate scattering phase function, which is also part of IOP inputs to the RT simulations, was based on the particle backscatter fraction $B_p(\lambda)$. We used the Fournier-Forand phase functions which are parameterized in terms of $B_p(\lambda)$ and are built into the Hydrolight/Ecolight model.

Fig. 2 compares the model-simulated hyperspectral $R_{rs}(\lambda)$ with the measured multispectral $R_{rs}(\lambda)$ for two selected stations. The model results compare reasonably well with measurements, which lends confidence to the use of hyperspectral $R_{rs}(\lambda)$ in our cluster analysis. This level of consistency between the model and measurements suggests that the suite of parameters used as input to the RT simulations realistically represent the actual field conditions. The ability to define realistic inputs derives, in turn, from a comprehensive suite of IOP measurements that were carried out during the cruise.

2.3. Hierarchical cluster analysis and similarity indices between dendrograms

A hierarchical cluster analysis (HCA) was used to classify the 9 selected stations into distinct groups on the basis of several types of input data vectors (or objects), which included the HPLC pigments and optical data derived from spectral absorption coefficients and remote-sensing reflectance. For a given type of data, the input to the cluster analysis consisted of 9 numerical data vectors, each representing one of the 9 stations. For the input data representing the ratio of individual pigment concentrations to TChl*a*, an object for a given station is a data vector $\{p_1, p_2, p_3, \ldots, p_{24}\}$ where the consecutive elements p_i represent the ratio of each of the 24 individual pigment concentrations to TChl*a* concentration. Another type of pigment data vector used in the cluster analysis is of the form $\{d_1, d_2, d_3, \ldots, d_8\}$, where the consecutive elements represent a ratio

of one of the 8 diagnostic pigment concentrations to the sum of diagnostic pigment concentrations.

Several types of optical data vectors were used as input to the HCA analysis, including objects consisting of hyperspectral data of the remote-sensing reflectance, $R_{rs}(\lambda)$, the phytoplankton absorption coefficient, $a_{ph}(\lambda)$, the sum of pure water and phytoplankton absorption coefficients, $a_w(\lambda) + a_{ph}(\lambda)$, and the total absorption coefficient, $a(\lambda)$. The input characterizing the hyperspectral remote-sensing reflectance for any given station was used in the form of the following data vector $\{R_{rs}(\lambda_1)/R_{rs}(555), R_{rs}(\lambda_2)/R_{rs}(555), R_{rs}(\lambda_3)/R_{rs}(555),\ldots, R_{rs}(\lambda_n)/R_{rs}(555)\}$, where the consecutive elements represent the values of R_{rs} at successive light wavelengths normalized to R_{rs} at 555 nm over the spectral range from λ_1 to λ_n. Similar input vectors were created for the different components of spectral absorption. Because our analysis is focused on the spectral shapes, all the optical spectra used in the cluster analysis were normalized by the value of the optical variable at 555 nm at which variations in R_{rs} within the open ocean are generally small. The spectra involving the absorption coefficients were additionally normalized by TChl*a* concentration to minimize variability in absorption associated with changes in phytoplankton biomass. The rationale for selecting the data of $a(\lambda)$, $a_w(\lambda) + a_{ph}(\lambda)$, and $a_{ph}(\lambda)$ to create input data vectors for the cluster analysis stems from the fact that the variation in the spectral shape of $a(\lambda)$ is typically a major determinant of the variation in the spectral shape of $R_{rs}(\lambda)$. In turn, the variations in the spectral shape of $a_w(\lambda) + a_{ph}(\lambda)$ or $a_{ph}(\lambda)$ can be viewed as an important or dominant source of variation in the spectral shape of $a(\lambda)$ in open ocean situations.

We also created vectors from the second derivative spectra of the hyperspectral reflectance and absorption objects for input into the cluster analysis. The estimation of the second derivative spectra from these data was made with a finite divided difference algorithm, the so-called "finite approximation", which computes the changes in curvature of a given spectrum over a sampling

interval ($\Delta\lambda$) or band separation (*BS*) defined as $\Delta\lambda = \lambda_j - \lambda_i$, where $j > i$. Because the identification of spectral details in the derivative spectra depends on the selection of the band separation, we tested the sensitivity of cluster results to the choice of *BS*. The derivative technique is also sensitive to signal noise, thus smoothing was applied to the hyperspectral optical data prior to computation of derivative spectra. Specifically, a mean-filter smoothing method was used in which the extent of spectral smoothing depends on the size of the filter window (*WS*) used for averaging. We tested the sensitivity of cluster results to different values of *WS*. The sensitivity analysis over a range of *BS* and *WS* values allowed us to achieve the best compromise between the ability to resolve fine spectral details and the reduction of noise effects in the second derivative spectra. As discussed below (sections 3.2 and 3.3), the optimal values of *BS* and *WS* chosen in this study for the derivative analysis of absorption data are 9 nm. For the derivative analysis of reflectance data, these values are 27 nm. Therefore, although the derivative calculations were made using data from the spectral range 300 – 725 nm, our results from the derivative analysis for absorption will be reported between the wavelengths of λ_{min} = 309 nm (≡ 300 + 9) and λ_{max} = 716 nm (≡ 725 – 9). For the reflectance derivative, the results will be reported between λ_{min} = 327 nm (≡ 300 + 27) and λ_{max} = 698 nm (≡ 725 – 27).

With regard to the analysis of remote-sensing reflectance, the cluster analysis was also applied to the multispectral reflectance data obtained from in situ SPMR measurements at several discrete wavelengths. We examined the 3-element objects {$R_{rs}(443)/R_{rs}(554)$, $R_{rs}(490)/R_{rs}(554)$, $R_{rs}(510)/R_{rs}(554)$}, which consist of band ratios that are similar to those used in current research based on satellite ocean color observations such as the Sea-viewing Wide Field-of-View Sensor (SeaWiFS). We also examined the vectors consisting of 13 band ratios of remote-sensing reflectance with $R_{rs}(554)$ in the denominator, as determined from SPMR measurements at 13 wavebands.

The HCA method, schematically presented in Fig. 3, was applied using the above defined pigment and optical data vectors as input objects. This method utilizes an unsupervised

383 classification algorithm which creates a hierarchical cluster tree (dendrogram) that partitions a
384 given set of input data into clusters or groups of objects (Jain et al., 1999; Berkhin, 2006). Each
385 group includes objects that are similar to each other, but different from objects in other groups.
386 The cluster tree is obtained using a linkage algorithm based on initial calculations of the
387 pairwise distance between all objects included in the input data set. In this study the similarity
388 between each pair of objects was the cosine distance, *d*, calculated as one minus the cosine of the
389 angle θ between each pair of objects:

390 $$d(x_1, x_2) = 1 - \cos\theta = 1 - \left(\frac{x_1 \cdot x_2}{\|x_1\| \times \|x_2\|} \right) \quad (1)$$

391 where the objects x_1 and x_2 include the two considered input data vectors and the cosine of the
392 angle between the vectors is obtained as the ratio of the dot product of the vectors to the product
393 of norms of the vectors. Note that as the angle between the objects decreases the cosine
394 approaches 1, resulting in a smaller distance between the input data vectors and therefore higher
395 similarity. We also tested other measures of similarity between the input objects, e.g., Euclidean
396 distance using a similar approach to that proposed in Robila (2005). The cosine distance was
397 selected as the most appropriate measure for our study because it reflects mainly the differences
398 in the spectral shape of optical data rather than magnitude. The cosine distance is also
399 advantageous because it is scale invariant, i.e., insensitive to normalization of optical spectra at a
400 specific wavelength.
401 As a linkage algorithm, the shortest distance *D*, also referred to as the nearest neighbor, was
402 computed to measure the distance between two clusters of objects in the tree:

403 $$D(a,b) = \min\left[dist\left(x_{ai}, x_{bj}\right)\right] \qquad i \in (1,\ldots,n_a) \text{ and } j \in (1,\ldots,n_b) \quad (2)$$

404 where x_{ai} is the *i*th object in cluster *a* and x_{bj} is the *j*th object in cluster *b*. In the traditional
405 graphical representation of a dendrogram, the individual objects appear at one end and a single
406 cluster containing all objects at the other end (e.g., Jain et al., 1999). In our presentation of

dendrograms, pairs of objects showing a small cosine distance between them (i.e., with similar pigment composition or spectral properties) provide small linkage distance and therefore appear closer to each other in the cluster tree.

To evaluate the utility of optical data for discriminating phytoplankton pigment assemblages, we compared the dendrograms obtained for the different spectral optical data with a reference dendrogram obtained using the pigment composition data (see final step in Fig. 3). For this analysis, we utilize two objective criteria of cluster similarity, the cophenetic index (Sokal & Rolf, 1962) and the Rand index (Rand, 1971).

The cophenetic index (r_C) is a measure of how precisely two dendrograms preserve the pairwise distances between data objects. This index is computed from the cophenetic matrix (C) associated with each dendrogram. The elements of a cophenetic matrix ($c_{i,j}$) encode the distance between two objects (i, j), representing in the dendrogram the height of the link at which those two objects are first joined. This height is the distance between the two clusters that are merged by this link. The cophenetic index r_C represents the correlation between two cophenetic matrices (C_1 and C_2):

$$r_C = \frac{\sum_i \sum_j \left(c_{1_{i,j}} - \overline{c_1}\right)\left(c_{2_{i,j}} - \overline{c_2}\right)}{\sqrt{\left(\sum_i \sum_j \left(c_{1_{i,j}} - \overline{c_1}\right)^2\right)\left(\sum_i \sum_j \left(c_{2_{i,j}} - \overline{c_2}\right)^2\right)}} \quad (3)$$

where $\overline{c_1}$ and $\overline{c_2}$ are the mean values of the elements of the matrices C_1 and C_2, respectively.

The Rand index (r_I) provides a measure of the similarity between two hierarchical dendrograms in terms of the proportion of pairs of objects whose relationship is the same in both dendrograms. The r_I value of 1 means that all pairs of objects are clustered in the same way in both dendrograms. Note that this index has to be computed using all dendrograms cut horizontally at a level (i.e., at a specific linkage distance) which yields the optimal number of clusters (k). Otherwise, r_I would always provide a proportion of 100% because a complete dendrogram always decomposes the input data all the way through the lowest level (i.e., until the

branches consist only of single objects). Detecting natural groupings in the dendrogram and selecting the optimal number of clusters is performed by analyzing a diagram of the increasing linkage distances along the dendrogram. Based on the points at which the linkage distances between the objects change abruptly (which is associated with a steep increase of the within cluster variance), the optimal number of clusters *k* is determined and all objects located below the point where the hierarchical tree is cut off are assigned to a single cluster (Salvador & Chan, 2004).

3. Results and discussion

3.1. Classification of stations based on pigment composition

For the 9 stations selected in the study, the estimate of the TChl*a* concentration ranges from about 0.11 mg m^{-3} at the southernmost station 59 in the open ocean off the coast of Namibia to 0.62 mg m^{-3} at the northernmost station 1 in the English Channel (Fig. 1). The variability in pigment composition for the 9 stations is summarized in Table 1, which provides the ratios of the concentration of several dominant pigments to TChl*a*. Apart from MVChl*a*, which is a principal pigment common to all phytoplankton, the second most important pigment at different stations was either DVChl*a*, zeaxanthin (Zea), 19'-hexanoyloxyfucoxanthin (Hex), or fucoxanthin (Fuco). Table 1 also identifies the two dominant pigments (excluding MVChl*a*) which yield the highest ratio to TChl*a* at each station. The values for the ratios of the two dominant pigments to TChl*a* were used as a basis for selecting the 9 stations. As these stations represent different pigment compositions, we assigned a class label A, B, C, D, E, or F to each station.

Most stations visited during the cruise within the tropical and subtropical regions of the Atlantic were dominated by Zea and DVChl*a*, but the relative predominance of these two pigments varied between the stations. These pigments are diagnostic of picophytoplankton that

include DVChl*a*- and Zea-containing prochlorophytes and Zea-containing cyanobacteria (mainly *Synechococcus* in the open ocean waters). We selected 5 stations (6, 12, 37, 44, and 46) to represent this type of pigment assemblage. Note that 4 stations dominated by Zea and DVChl*a* with fairly similar ratios DVChl*a*/TChl*a* and Zea/Chl*a* are grouped within the same class C with a label C1, C2, C3, and C4. The station 6, where DVChl*a*/TChl*a* is significantly higher than Zea/TChl*a*, is considered as a separate class B. This station (or class B) is dominated by prochlorophytes as DVChl*a* is an unambiguous marker of this group. The class C stations also show significant role of prochlorophytes. However, this class exhibits a relatively higher contribution of Zea than class B, which is likely indicative of an increased role of cyanobacteria.

Fuco and MVChl*b* are the predominant accessory pigments at station 1 (class A). As these two pigments are not confined to one phytoplankton class, this station could have been dominated by Fuco-rich diatoms, haptophytes, and/or dinoflagellates, as well as MVChl*b*-rich prasinophytes and/or chlorophytes. The predominant accessory pigment at stations 48 (class D) and 51 (class E) is Hex, which suggests that haptophytes and/or chrysophytes are major phytoplankton groups at these locations. These stations are designated as different classes because they clearly differ in accessory pigments that follow Hex in ranking. Zea and Fuco are the second most important diagnostic pigments at stations 48 (class D) and 51 (class E), respectively. Finally, station 59 (class F) also shows a significant role of Hex-rich phytoplankton although Zea is the most important diagnostic pigment at this location, indicating potential significance of cyanobacteria and/or prochlorophytes.

Fig. 4a shows the hierarchical cluster tree obtained for the input data consisting of the ratios of concentrations of 24 individual pigments to TChl*a* at each station. The optimal number of clusters (k) is derived from a diagram of linkage distances along the dendrogram (Fig. 4b). The first steep increase in the linkage distance observed in this diagram, which is associated with an increase of the within cluster variance, suggests an optimal partitioning of the pigment data into 5 clusters. The linkage distance of 0.023 can be selected to characterize this steep increase in

variance (see dashed lines in Fig. 4a and 4b). For the dendrogram cut at a level of linkage distance of 0.023, all clusters are single object (i.e., single station) clusters, except for a multi-object cluster that includes stations C1, C2, C3, C4, and B. The results of this cluster analysis are quite consistent with the preliminary classification obtained by considering just two dominant diagnostic pigments (see Table 1). Note that stations C1, C2, C3, C4, and B are all characterized by relatively high ratios of DVChl*a* and Zea to TChl*a*. Some differences between these stations in terms of the relative roles of DVChl*a* and Zea do not, however, produce significant distances between the corresponding pigment data vectors and hence these 5 stations are grouped into a single cluster. The dendrogram also indicates that the stations classified as A, D, E, and F display significant dissimilarities between each other and when compared to the stations classified as B and C. Note that the stations 48 (class D) and 59 (class F) have the Hex and Zea as dominant diagnostic pigments, albeit in reverse ranking (see Table 1), so these stations appear closer to one another in the dendrogram (Fig. 4a).

Fig. 4c also depicts results from the cluster analysis of pigment data but for the input data vectors consisting of the ratios of concentrations of the 8 diagnostic pigments to the sum of the 8 diagnostic pigments. The partitioning of the stations obtained on the basis of these pigment ratios is qualitatively identical to the partitioning based on the ratios of 24 pigments to TChl*a*. The stations C1, C2, C3, C4, and B are again grouped within a single cluster and each of the remaining stations represent a separate cluster. The results in Figs. 4a and 4c along with the preliminary classification of stations based on two dominant diagnostic pigments (Table 1) suggest that there is a certain degree of flexibility in the selection of pigment ratios as a basis for discriminating different pigment assemblages in a consistent fashion. In the following analysis of optical data as a means for assessing differences in pigment assemblages, the pigment-based cluster partitioning obtained with the 24 pigment ratios (as shown in Fig. 4a) is used as a reference.

3.2. Classification of stations based on absorption spectra

Fig. 5 shows hyperspectral data of absorption coefficients first normalized at 555 nm and then divided by the total chlorophyll-*a* concentration for the nine stations. These spectra are referred to as the spectral chlorophyll (Chl)-specific normalized absorption coefficients. Specifically, we examine the Chl-specific normalized coefficients for the total absorption, $a_n^*(\lambda)$, the absorption of pure seawater plus phytoplankton, $a_{n,w+ph}^*(\lambda)$, and the absorption of phytoplankton alone, $a_{n,ph}^*(\lambda)$. The differences in the shape of phytoplankton absorption in the UV and blue spectral regions are generally quite large between most stations (Fig. 5a). With the addition of the pure water contribution, differences in the spectral shape of $a_{n,w+ph}^*(\lambda)$ continue to be seen but are considerably smaller (Fig. 5b). Finally, upon further addition of the contributions associated with non-phytoplankton particles and CDOM, the spectral shape of total absorption again shows larger differences between the stations at wavelengths shorter than the normalization point at 555 nm (Fig. 5c). From the visual inspection of these plots it is, however, difficult to deduce to what extent the observed differences might be consistent with the classification of stations based on pigment composition.

Fig. 6 illustrates the results from cluster analysis applied to the absorption spectra presented in Fig. 5 and the corresponding second derivative spectra over the entire spectral range from 300 nm to 725 nm. In nearly all cases (Fig. 6a-e), the absorption-based cluster trees differ significantly from the pigment-based cluster tree shown in Fig. 4a. Thus, the full hyperspectral data of $a_n^*(\lambda)$, $a_{n,w+ph}^*(\lambda)$, and $a_{n,ph}^*(\lambda)$ as well the second derivative spectra of $a_n^*(\lambda)$ and $a_{n,w+ph}^*(\lambda)$ do not provide useful information for discriminating the differences in pigment assemblages at the examined stations. The only case in which stations are classified within the dendrogram in a similar way to the pigment-based cluster tree is when the second derivative of phytoplankton absorption spectra is considered (Fig. 6f). When this absorption-based dendrogram is cut horizontally at a level of linkage distance of 0.023 that yields 5 clusters, the

same stations are grouped in separate clusters as in the pigment-based cluster tree. This result supports the potential usefulness of the second derivative of phytoplankton absorption spectra for discriminating different pigment assemblages.

In the analysis above we considered a spectral range from 300 nm to 725 nm, which is much broader than the spectral region where specific absorption imprints caused by accessory pigments occur. It is therefore useful to examine whether the cluster analysis of absorption data yields similarity with pigment-based clusters if different, narrower spectral ranges are considered. Fig. 7 illustrates the degree of similarity between the absorption-based and pigment-based cluster trees for different spectral ranges of absorption data. The degree of similarity is shown in terms of cophenetic and Rand indices. The three absorption spectra, $a_n^*(\lambda)$, $a_{n,w+ph}^*(\lambda)$, and $a_{n,ph}^*(\lambda)$, are considered in this analysis. In each graph, the distribution of values for the cophenetic or Rand index is shown as a function of the spectral range considered, with the lower limit of the spectral range, λ_{min}, displayed along the y-axis (ordinate) and the upper limit, λ_{max}, along the x-axis (abscissa). The similarity indices are thus shown for many spectral ranges represented by many combinations of λ_{min} and λ_{max}. The higher values of indices, depicted by darker areas in the graphs, correspond to better similarity between a given absorption-based cluster tree and pigment-based tree. The best degree of similarity is obtained when the indices are close to 1, indicated by the nearly black areas in the graphs. According to the distributions of cophenetic index, this is the case when the phytoplankton absorption spectrum $a_{n,ph}^*(\lambda)$ is analyzed over the spectral range approximately from λ_{min} = 425 nm to λ_{max} = 540 nm (Fig. 7e). The distribution of the Rand index indicates that the best similarity between the $a_{n,ph}^*(\lambda)$-based cluster tree and the pigment-based tree occurs within a broader spectral region, approximately between λ_{min} = 390 nm and λ_{max} = 610 nm (Fig. 7f). These optimal spectral regions generally overlap with the wavelength range where absorption characteristics of main accessory pigments appear (e.g., Bricaud et al., 2004). The remaining results in Fig. 7

(panels a, b, c, d) show generally poor similarity between the absorption data of $a_n^*(\lambda)$ or $a_{n,w+ph}^*(\lambda)$ and pigment composition, regardless of the spectral range considered.

Fig. 8 depicts similar results, but for the similarity between pigment composition and the second derivative spectra of $a_n^*(\lambda)$, $a_{n,w+ph}^*(\lambda)$, and $a_{n,ph}^*(\lambda)$. The use of derivative spectra generally improves the similarity as indicated by the presence of darker areas or the larger extent of dark areas in the distributions of the cophenetic and Rand indices. For example, compared to the results for the ordinary spectra of $a_{n,w+ph}^*(\lambda)$ in Fig. 7d, a significant increase in the Rand index is observed for the second derivative spectra of $a_{n,w+ph}^*(\lambda)$ within the spectral range from λ_{min} = 440 nm to λ_{max}= 650 nm (Fig. 8d). The improvement is even more striking for the results involving the second derivative spectra of phytoplankton absorption (Fig. 8e, f). The spectral regions where the cophenetic and Rand indices assume high values near or equal to 1 are much larger compared with the analysis of ordinary (non-differentiated) spectra of $a_{n,ph}^*(\lambda)$. The darkest areas in the distributions of the indices in Fig. 8e, f cover a broad spectral range, approximately from λ_{min} = 370 nm to λ_{max} = 716 nm. This result indicates that a high degree of similarity between the second derivative spectra of $a_{n,ph}^*(\lambda)$ and pigment composition can be obtained for many different combinations of spectral ranges of absorption data (i.e., different combinations of λ_{min} and λ_{max}).

The improvement in the similarity between the pigment-based and absorption-based cluster trees achieved as a result of utilization of second derivative spectra compared to the ordinary spectra of phytoplankton absorption, $a_{n,ph}^*(\lambda)$, is clearly illustrated by histograms of cophenetic index obtained on the basis of all possible combinations of spectral ranges examined (Fig. 9). We note that the cophenetic index does not require a priori selection of the optimal number of clusters which is somewhat subjective and, therefore, this index facilitates the comparison of results for ordinary and derivative spectra shown in Fig. 7e and Fig. 8e. Compared with the histogram obtained using the ordinary spectra (Fig. 9a), the histogram for the derivative spectra (Fig. 9b) shows a substantial shift to higher values of the cophenetic index. Specifically, there is

584 a significant increase in the frequency of occurrence of high values (> 0.9) of the cophenetic
585 index. This increased frequency is associated with a broader overall spectral region, or
586 equivalently a larger number of spectral ranges, for which the index is higher than 0.9.
587 Whereas the above results illustrate the advantages of derivative analysis, such analysis can
588 be highly sensitive to parameters chosen for the calculation of derivative spectra, specifically the
589 size of the filter window (*WS*) used in the spectral smoothing of the ordinary spectra and the
590 band separation (*BS*) used in the calculation of derivatives. To examine the sensitivity of cluster
591 analysis to the selection of these parameters, we computed the distribution of cophenetic index
592 between the pigment-based and absorption-based cluster trees using the second derivative
593 spectra of $a_{n,ph}^*(\lambda)$ within the spectral range of 420 - 515 nm as input (Fig. 10). This spectral
594 range is adequate for this sensitivity analysis because it showed very high values of cophenetic
595 and Rand indices in Fig. 7e and Fig. 8e. The distribution of cophenetic index in Fig. 10 is shown
596 with the smoothing parameter *WS* varying from 1 to 29 consecutive samples (with a step of 2
597 samples) along the y-axis and the band separation parameter *BS* varying also from 1 to 29
598 samples with a step of 2 samples along the x-axis. Note that the number of consecutive samples
599 is equivalent to the wavelength interval in nanometers because our spectral data (samples) have
600 the resolution of 1 nm. The highest values of the cophenetic index are obtained for intermediate
601 values of *WS* and *BS* around 9 nm – 10 nm. These are the optimal values for our derivative
602 analysis of absorption spectra. This result is consistent with the general expectation that if the
603 values of *WS* and *BS* are too small, the derivative spectra are sensitive to noise and exhibit false
604 spectral features, and on the other hand if the *WS* and *BS* are too large, the real significant
605 spectral features get smoothed out and essentially removed from the analysis. As the best
606 compromise, the cluster analyses presented in Fig. 8 and 9b were obtained with *WS* and *BS* of 9
607 nm.
608

609 *3.3. Classification of stations based on remote-sensing reflectance*

610 The relationship between the spectral remote-sensing reflectance, $R_{rs}(\lambda)$, of the ocean and
611 phytoplankton pigment composition is less direct and far more complicated than that for the
612 spectral phytoplankton absorption coefficient, mainly due to the presence of many optically
613 significant non-phytoplankton constituents in seawater. The investigation of $R_{rs}(\lambda)$ is of
614 particular interest, however, because information contained in this measurement provides a
615 potential means for remote-sensing applications. Fig. 11 shows the Hydrolight-simulated $R_{rs}(\lambda)$
616 spectra normalized at 555 nm for the nine stations identified as classes from A (station 1) through
617 F (station 59). In general, there are significant differences in the UV and blue spectral regions
618 between these normalized spectra, with the largest contrast between the classes A (station 1) and
619 C1 (station 12).

620 The dendrograms obtained from cluster analysis as applied to four different sets of input data
621 vectors containing information about remote-sensing reflectance (as described in sec 2.3) are
622 displayed in Fig. 12. The limited spectral information, i.e., the three reflectance band ratios used
623 commonly in satellite ocean color applications (Fig. 12a) and the 13 band ratios corresponding to
624 multispectral measurements with the SPMR instrument (Fig. 12b), provide a very dissimilar
625 classification of stations compared with the pigment-based cluster analysis (see Fig. 4a). All
626 stations, with only the exception of class A, show very little separation in the cluster-tree based
627 on multispectral reflectance data. The high spectral resolution (1 nm) normalized reflectance
628 spectra over the entire spectral range 300 – 725 nm also produce a dendrogram (Fig. 12c) that is
629 very different from the pigment-based cluster tree. Although the stations belonging to class C
630 are closer to one another compared with the multispectral-based cluster tree, they are grouped
631 together with two other stations (class E and D). In addition, station B forms a separate single-
632 object cluster in Fig. 12c, whereas it is grouped in a single multi-object cluster together with the
633 stations from class C in the pigment analysis. The only case when the cluster analysis of
634 reflectance data provides a high degree of similarity with pigment analysis (i.e., Rand index of

635 0.78) is for the second derivative of hyperspectral normalized reflectance over the entire spectral
636 range 327 - 698 nm (Fig. 12d). We note that the derivative reflectance spectra were calculated
637 with the parameters *WS* and *BS* of 27 nm (as supported by the sensitivity analysis discussed
638 below). The stations A, E, and F in Fig. 12d form single-object clusters at a significant distance
639 from the remaining stations. Similarly to pigment analysis, the stations C1, C2, C3, C4, and B
640 are grouped relatively close to one another. However, station D also belongs to that group,
641 which is not the case in the pigment-based cluster tree. This may be attributable to the fact that
642 Zea and DVChl*a*, which are the two most dominant diagnostic pigments at stations C1, C2, C3,
643 C4, and B, also play a significant role at station D where they are ranked as the second and third
644 most important diagnostic pigments (see Table 1).
645 The progression of linkage distances corresponding to the four dendrograms from Fig. 12
646 clearly illustrates the advantage of the second derivative spectra over the multispectral data or
647 non-differentiated spectra of reflectance (Fig. 13). The improved separation seen in terms of the
648 larger linkage distances between the clusters of stations obtained with derivative spectra
649 indicates that this approach enables better identification of the differences in the magnitude and
650 shape between the high resolution spectra. In contrast, in the analysis of multispectral data and
651 ordinary spectra, stations are linked at a very small distance, which indicates that these types of
652 reflectance data will be essentially useless for obtaining information about pigment assemblages
653 from cluster analysis.
654 The improvement in the similarity between the pigment-based and reflectance-based
655 classification achieved with the use of second derivative spectra as opposed to multispectral data
656 or ordinary spectra of reflectance is presented in Table 2. The results for the derivative analysis
657 of hyperspectral normalized $R_{rs}(\lambda)$ over the entire spectral range 300 - 725 nm show a significant
658 increase in both the cophenetic and Rand index when compared with multispectral and ordinary
659 spectral data. However, the best performance is obtained when the derivative analysis is

660 restricted to the spectral range from 435 to 495 nm (as supported by the sensitivity analysis
661 discussed below). In this case, the similarity indices are highest.
662 Fig. 14 shows distributions of the cophenetic and Rand index which identify the optimal
663 spectral ranges for the cluster analysis of the second derivative of remote-sensing reflectance.
664 The cophenetic index is generally close or slightly higher than 0.5 for most spectral ranges
665 examined, that is for most combinations of λ_{min} and λ_{max} (Fig. 14a). In the spectral region from
666 435 nm to 510 nm, this index is about 0.65. This can be considered as an optimal spectral range
667 for the application of derivative approach with potential for good similarity between the
668 pigment-based and reflectance-based cluster trees. This result is also supported by very high
669 value of the Rand index of 0.86 in that spectral range (Fig. 14b). The Rand index attains even
670 higher value of about 1 within somewhat narrower wavelength range from λ_{min} = 435 nm to λ_{max}
671 = 495 nm, which defines an alternative optimal spectral range. In addition, the Rand index
672 suggests good performance of the derivative-based analysis over a broader spectral region
673 including shorter wavelengths of λ_{min} from the near-UV. In general, this index is quite high for
674 λ_{min} varying between 350 nm and 450 nm, for example as high as 1 when the spectral range is
675 from λ_{min} = 365 nm to λ_{max} = 480 nm. However, because the optical roles of different diagnostic
676 pigments in the near-UV are insignificant or certainly less important than in the blue region, the
677 use of the spectral range 435 – 510 nm or 435 – 495 nm, where both the cophenetic and Rand
678 indices are relatively high, appears to be most reasonable.
679 The effects of the size of the filter window (*WS*) used in the smoothing of the ordinary
680 spectra and the band separation (*BS*) used in the calculation of derivatives on the similarity
681 between the cluster trees from the analysis of pigments and the second derivative reflectance
682 spectra is illustrated in Fig. 15. These results are shown for the derivative spectra calculated over
683 one of the optimal spectral ranges, specifically 435 - 495 nm, which showed high values for both
684 cophenetic and Rand indices. The best similarity with cophenetic index of about 0.65 and the
685 Rand index of 1 is obtained when the calculations of second derivative spectra are made with

686 relatively large values of *WS* and *BS*. For example, a very good result is obtained if both *WS* and
687 *BS* assume a value of 27 consecutive spectral samples (i.e., 27 nm as the resolution of our
688 hyperspectral reflectance data is 1 nm). We recall that this value was used to compute the results
689 pertinent to the derivative reflectance spectra presented in Figs. 12 - 14. Similarly good results
690 are obtained with a smaller *WS* (~14 nm) and a larger *BS* (~37 nm), or vice versa. In contrast, if
691 both *WS* and *BS* are small (less than about 10 nm) or large (above ~40 nm) the cophenetic and
692 Rand indices are reduced significantly. This sensitivity analysis, in agreement with similar
693 analysis for absorption spectra, supports strong dependence of the derivative-based cluster trees
694 on the selection of parameters *WS* and *BS* for the derivative calculations.
695
696 **4. Conclusions**
697 By applying the unsupervised hierarchical cluster analysis to pigment and optical data from
698 the eastern Atlantic Ocean we demonstrated the potential usefulness of hyperspectral data of
699 absorption coefficient and remote-sensing reflectance for discriminating different phytoplankton
700 pigment assemblages in the open ocean under non-bloom conditions. The most promising
701 results were obtained with the second derivative spectra of phytoplankton absorption coefficient
702 covering the spectral range as wide as 370 nm - 725 nm (or narrower spectral regions from
703 within that range), and the second derivative spectra of remote-sensing reflectance over the
704 spectral range from about 435 nm to 510 nm. Our ability to discriminate different phytoplankton
705 pigment assemblages from the derivative-based cluster trees was optimized by selecting the most
706 suitable parameters used in the spectral derivative calculations (see also Torrecilla et al., 2009).
707 In particular, for these optical data with a 1 nm resolution, we determined that the optimal values
708 for the smoothing filter window and band separation used in the calculations of second derivative
709 spectra are 9 nm for absorption and 27 nm for remote-sensing reflectance. These derivative
710 spectra are presented in Fig. 16 for the nine Atlantic stations selected for this study. The

711 assessment of similarity between these derivative spectra and phytoplankton pigment
712 composition was made using the similarity-based cluster algorithm and two indices, cophenetic
713 and Rand. In this cluster algorithm, the partitioning of data into clusters is based on the
714 determinations of the angular distance between each pair of examined input data vectors. This
715 type of similarity-based cluster technique accounts for complete spectral behavior of optical data
716 with no need to identify specific spectral features. This approach is particularly useful in cases
717 when it is difficult or impossible to define explicitly a set of unambiguous diagnostic spectral
718 features in the original optical data (Duin et al., 1997; Pekalska and Duin, 2000).
719 In addition to the second derivative spectra of phytoplankton absorption and remote-sensing
720 reflectance, we examined other absorption and reflectance data but they generally showed either
721 more limited value or no usefulness at all for discriminating phytoplankton pigment
722 assemblages. For example, the cluster analysis of the ordinary (non-differentiated) reflectance
723 spectra at 1 nm resolution or multispectral (13 wavebands) reflectance data showed very poor
724 similarity with pigment-based clusters. Similar results were obtained for the ordinary spectra of
725 the total absorption coefficient. However, the ordinary spectra of phytoplankton absorption are
726 useful, especially within the spectral range 425 nm - 540 nm.
727 We demonstrated that the quantification of similarity between the optical-based clusters and
728 pigment-based clusters with the cophenetic and Rand indices provides a valuable methodology
729 for identifying optical variables and their spectral ranges most suitable for characterizing the
730 phytoplankton pigment assemblages, and for selecting the optimal values of the parameters used
731 in the calculation of derivative spectra. Whereas the present study uses a limited data set, albeit
732 carefully selected to represent distinct differences in phytoplankton assemblages in terms of
733 dominant accessory pigments, further work is needed to evaluate or refine the proposed
734 methodology with larger data sets from various oceanic environments. Given significant interest
735 in the development of the capabilities for large-scale characterization of phytoplankton
736 biodiversity from optical measurements including remote-sensing observations, one may expect

737 further expansion of comprehensive databases consisting of simultaneously collected pigment
738 and hyperspectral optical data in the near future. We expect that this will support further work on
739 the cluster-based approach and other techniques, such as neural networks (Raitsos et al., 2008;
740 Aymerich et al., 2009), which exploit optical measurements as a source of information on
741 phytoplankton community composition.
742
743 **Acknowledgements**
744 This study was supported by the NASA Biodiversity and Ecological Forecasting Program (Grant
745 NNX09AK17G), the NASA Ocean Biology and Biogeochemistry Program (Grant
746 NNG04GO02G), and the Spanish National Research Council CSIC (project ANERIS PIF08-
747 015). Part of this study was performed during a visit of E. T. at Scripps Institution of
748 Oceanography supported also by CSIC (Program I3P). The Alfred Wegener Institute for Polar
749 and Marine Research (Bremerhaven, Germany) kindly made it possible for us to participate in
750 the cruise in the eastern Atlantic. We thank R. Röttgers for providing PSICAM data and HPLC
751 data from analysis at GKKS Research Centre (Geesthacht, Germany).
752
753 **References**

754 Aiken, J., Fishwick, J. R., Lavender, S. J., Barlow, R., Moore, G., Sessions, H., et al. (2007).
755 Validation of MERIS reflectance and chlorophyll during the BENCAL cruise October, 2002:
756 Preliminary validation of new products for phytoplankton functional types and
757 photosynthetic parameters. *International Journal of Remote Sensing*, *28*, 497-516.
758 Alvain, S., Moulin, C., Dandonneau, Y., & Bréon, F. M. (2005). Remote sensing of
759 phytoplankton groups in case 1 waters from global SeaWiFS imagery. *Deep-Sea Research I*,
760 *52*, 1989-2004.
761 Aymerich, I. F., Piera, J., Soria-Frisch, A., & Cros, Ll. (2009). A rapid technique for classifying

762 phytoplankton fluorescence spectra based on self-organizing maps. *Applied Spectroscopy, 63,*
763 716-726.
764 Balch, W. M., Gordon, H. R., Bowler, B. C., Drapeau, D. T., & Booth, E. S. (2005). Calcium
765 carbonate measurements in the surface global ocean based on Moderate-Resolution Imaging
766 Spectroradiometer data. *Journal of Geophysical Research, 110,* C07001,
767 doi:10.1029/2004JC002560.
768 Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In J. Kogan, C. Nicholas
769 and M. Teboulle (Eds.), *Grouping multidimensional data: Recent advances in clustering* (pp.
770 25-71). Berlin-Heidelberg: Springer.
771 Bricaud, A., Morel, A., Babin, M., Allali, K., & Claustre, H. (1998). Variations of light
772 absorption by suspended particles with chlorophyll-*a* concentration in oceanic (case 1)
773 waters: Analysis and implications for bio-optical models. *Journal of Geophysical Research,*
774 103, 31033-31044.
775 Bricaud, A., Claustre, H., Ras, J., & Oubelkheir, K. (2004). Natural variability of phytoplankton
776 absorption in oceanic waters: influence of the size structure of algal population. *Journal of*
777 *Geophysical Research,* 109, C11010, doi:10.1029/2004JC002419.
778 Buiteveld, H., Hakvoort, J. H. M., & Donze, M. (1994). The optical properties of pure water. In
779 J. S. Jaffe (Ed.), *Ocean Optics XII, Proc. SPIE, 2258* (pp. 174-183). Bellingham: SPIE.
780 Chang, G. C., Mahoney, K., Briggs-Whitmire, A., Kohler, D., Mobley, C., Moline, M., Lewis,
781 M., Boss, E., Kim, M., Philpot, W., & Dickey, T. (2004). The New Age of Hyperspectral
782 Oceanography. *Oceanography, 17(2),* 22-29.
783 Chang, G. C., Dickey, T., & Lewis, M. (2006). Toward a global ocean system for measurements
784 of optical properties using remote sensing and in situ observations. In J. Gower (Ed.), *Remote*
785 *Sensing of the Marine Environment: Manual of Remote Sensing* (pp. 285-326). New York:
786 John Wiley and Sons.
787 Ciotti, A. M., & Bricaud, A. (2006). Retrievals of a size parameter for phytoplankton and spectral

light absorption by colored detrital matter from water-leaving radiances at SeaWiFS channels in a continental shelf region off Brazil. *Limnology and Oceanography: Methods, 4*, 237-253.

Craig, S. E., Lohrenz, S. E., Lee, Z., Mahoney, K. L., Kirkpatrick, G. J., Schofield, O. M., & Steward, R. G. (2006). Use of hyperspectral remote sensing reflectance for detection and assessment of the harmful alga, *Karenia brevis*. *Applied Optics, 45*, 5414-5425.

Cullen, J. J., Ciotti, A. M., Davis, R. F., & Lewis, M. R. (1997). Optical detection and assessment of algal blooms. *Limnology and Oceanography*, *42*, 1223-1239.

Dickey, T., Lewis, M., & Chang, G. (2006). Optical oceanography: recent advances and future directions using global remote sensing and in situ observations. *Reviews of Geophysics*, *44*, RG1001, 1-39.

Duin, R. P. W., de Ridder, D., & Tax, D. M. J. (1997). Experiments with a featureless approach to pattern recognition. *Pattern Recognition Letters*, *18,* 1159-1166.

Ferrari, G. M., & Tassan, S. (1999), A method using chemical oxidation to remove light absorption by phytoplankton pigments, *Journal of Phycology*, *35*, 1090-1098.

Fry, E. S., Lu, Z., & Qu, X. (2006), Optical absorption of pure water throughout the visible and near ultraviolet. *Proceedings of Ocean Optics Conference XVIII*, Montreal, Canada.

Hooker, S.B., & Van Heukelem, L. (2009), *CHORS HPLC Uncertainties: Final Report.* World Wide Web page, from URL: http://oceancolor.gsfc.nasa.gov/DOCS/ . NASA Goddard Space Flight Center, Greenbelt, Maryland, 62 pp.

Hunter, P. D., Tyler, A. N., Présing, M., Kovács, A. W., & Preston, T. (2008). Spectral discrimination of phytoplankton colour groups: The effect of suspended particulate matter and sensor spectral resolution. *Remote Sensing of Environment*, 112, 1527-1544.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*, 264-323.

812 Jeffrey, S. W., Wright, S. W., & Zapata, M. (1997). Recent advances in HPLC pigment analysis
813 of phytoplankton. *Marine and Freshwater Research*, *50*, 879-896.
814 Kostadinov, T. S., Siegel, D. A., & Maritorena, S. (2009). Retrieval of the particle size
815 distribution from satellite ocean color observations. *Journal of Geophysical Research, 114*,
816 C09015, doi:10.1029/2009JC005303.
817 Lee, Z., & Carder, K. L. (2002). Effect of spectral band numbers on the retrieval of water column
818 and bottom properties from ocean color data. *Applied Optics, 41*, 2191-2201.
819 Lee, Z.P., & Carder, K.L. (2004). Absorption spectrum of phytoplankton pigments derived from
820 hyperspectral remote-sensing reflectance. *Remote Sensing Environment*, *89*, 361-368.
821 Louchard, E. M., Reid, R. P., Stephens, C. F., Davis, C. O., Leathers, R. A., Downes, T. V., &
822 Maffione, R. (2002). Derivative Analysis of Absorption Features in Hyperspectral Remote
823 Sensing Data of Carbonate Sediments. *Optics Express, 10*,1573-1584.
824 Lubac, B., Loisel, H., Guiselin, N., Astoreca, R., Felipe Artigas, L., & Mériaux, X. (2008).
825 Hyperspectral and multispectral ocean color inversions to detect Phaeocystis globosa blooms
826 in coastal waters. *Journal of Geophysical. Research, 113*, C06026,
827 doi:10.1029/2007JC004451.
828 Mackey, M. D., Mackey, D. J., Higgins, H. W., & Wright, S. W. (1996). CHEMTAX - a program
829 for estimating class abundances from chemical markers: Application to HPLC measurements
830 of phytoplankton. *Marine Ecology Progress Series, 144*, 265-283.
831 Mobley, C. D. (1994). *Light and Water: Radiative Transfer in Natural Waters.* (pp. 592). San
832 Diego: Academic Press.
833 Mobley, C. D. (2008). Hydrolight Ecolight 5.0 User's Guide. (pp. 99). Mercer Island,
834 Washington: Sequoia Scientific Inc.
835 Morel, A. (1988). Optical modeling of the upper ocean in relation to its biogenous matter content
836 (case I waters), *Journal of Geophysical Research*, *93*, 10749-10768.
837 Mueller, J. L., Fargion, G. S., & McClain, C. R. Eds. (2003). Ocean Optics Protocols for Satellite

838 Ocean Color Sensor Validation, Revision 4, Volume III: Radiometric Measurements and Data
839 Analysis Protocols, *NASA/TM-2003-211621/Rev4-Vol. III*, 78 pp., NASA Goddard Space
840 Flight Center, Greenbelt, Maryland.
841 Nair, A., Sathyendranath, S., Platt, T., Morales, J., Stuart, V., Forget, M., Devred, E., & Bouman,
842 H. (2008). Remote sensing of phytoplankton functional types. *Remote Sensing of*
843 *Environment*, *112*, 3366-3375.
844 O'Reilly, J. E., Maritorena, S., Siegel, D. A. et al. (2000). Ocean Color chlorophyll-*a* Algorithms
845 for SeaWiFS, OC2, and OC4: Version 4. SeaWiFS Postlaunch Technical Report Series,
846 Volume. 11, SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3, edited by
847 Hooker, S. B. and Firestone, E. R., Greenbelt, Maryland, NASA/TM-2000-206892, 9-27.
848 Pekalska, E., & Duin, R. P. W. (2000). Classifier for dissimilarity-based pattern recognition.
849 *Proceedings of the 15th International Conference on Pattern Recognition*, 12–16.
850 Perry, M. J., & Rudnick, D.L. (2003). Observing the oceans with autonomous and Lagrangian
851 platforms and sensors: The role of ALPS in sustained ocean observing systems.
852 *Oceanography*, *16(4)*, 31-36.
853 Pope, R. M., & Fry, E. S. (1997). Absorption spectrum (380-700nm) of pure water. II. Integrating
854 cavity measurements. *Applied Optics*, *36*, 8710-8723.
855 Raitsos, D. E., Lavender, S. J., Maravelias, C. D., Haralabous, J. A., Richardson, J., & Reid, P.
856 (2008). Identifying four phytoplankton functional types from space: An ecological approach,
857 *Limnology and Oceanography*, *53*, 605-613.
858 Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the*
859 *American Statistical Association*, *66*, 846-850.
860 Reynolds, R. A., Stramski D., & Mitchell, B. G. (2001). A chlorophyll-dependent semianalytical
861 model derived from field measurements of absorption and backscattering coefficients within
862 the Southern Ocean. *Journal of Geophysical Research*, *106*, 7125-7138.

Robila, S. A. (2005). Using spectral distances for speedup in hyperspectral image processing. *International Journal of Remote Sensing*, 26:24, 5629 – 5650.

Röttgers, R., & Doerffer, R. (2007). Measurements of optical absorption by chromophoric dissolved organic matter using a point-source integrating-cavity absorption meter. *Limnology and Oceanography: Methods*, *5*, 126–135

Röttgers, R., Schönfeld, W., P-R. Kipp, P-R., & Doerffer, R. (2005), Practical test of a point-source integrating cavity absorption meter: The performance of different collector assemblies. *Applied Optics*, *44*, 5549-5560.

Salvador, S., & Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *Proceedings of 16th IEEE International Conference on Tools with AI*, 576-584.

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon, 11(2)*, 33-40.

Stramska, M., Stramski, D., Kaczmarek, S., Allison, D. B., & Schwarz, J. (2006). Seasonal and regional differentiation of bio-optical properties within the north polar Atlantic. *Journal of Geophysical Reserach*, *111*, C08003, doi:10.1029/2005JC003293.

Stramski, D., Reynolds, R. A., Babin, M., Kaczmarek, S., Lewis, M. R., Röttgers, R., Sciandra, A., Stramska, M., Twardowski, M. S., Franz, B. A., & Claustre, H. (2008). Relationships between the surface concentration of particulate organic carbon and optical properties in the eastern South Pacific and eastern Atlantic Oceans. *Biogeosciences, 5*, 171-201.

Stumpf, R. P., Culver, M. E., Tester, P. A., Tomlinson, M., Kirkpatrick, G. J., Pederson, B. A., Truby, E., Ransibrahmanakul, V., & Soracco, M. (2003). Monitoring *Karenia brevis* blooms in the Gulf of Mexico using satellite ocean color imagery and other data. *Harmful Algae, 2*, 147-160.

Tassan, S., & Ferrari, G. M. (1995). An alternative approach to absorption measurements of aquatic particles retained on filters, *Limnology and Oceanography*, *40*, 1358-1368.

889 Tassan, S., & Ferrari, G. M. (2002). A sensitivity analysis of the 'Transmittance-Reflectance'
890 method for measuring light absorption by aquatic particles, *Journal of Plankton Research*, *24*,
891 757-774.
892 Torrecilla, E., Piera J., & Vilaseca, M. (2009). Derivative analysis of oceanographic
893 hyperspectral data. In G. Jedlovec (Ed.), *Advances in Geoscience and Remote Sensing* (pp.
894 597-619). Vienna: InTech.
895 Tsai, F., & Philpot, W. D. (1998). Derivative analysis of hyperspectral data. *Remote Sensing of*
896 *Environment*, *66*, 41-51.
897 Twardowski, M. S., Claustre, H., Freeman, S. A., Stramski, D., & Huot, Y. (2007). Optical
898 backscattering properties of the "clearest" natural waters. *Biogeosciences*, *4*, 1041-1058.
899 Uitz, J., Claustre, H., Gentili, B., & Stramski, D. (2010). Phytoplankton class-specific primary
900 production in the world's oceans: Seasonal and interannual variability from satellite
901 observations. *Global Biogeochemical Cycles*, *24*, GB3016, doi:10.1029/2009GB003680.
902 Uitz, J., Claustre, H., Morel, A., & Hooker, S. (2006). Vertical distribution of phytoplankton
903 communities in open-ocean: An assessment based on surface chlorophyll. *Journal of*
904 *Geophysical Research*, *111*, CO8005, doi:10.1029/2005JC003207.
905 Vaiphasa, C. (2006). Consideration of smoothing techniques for hyperspectral remote sensing.
906 *Journal of Photogrammetry and Remote Sensing*, *60*, 91-99.
907 Van Heukelem, L., & Thomas, C. S. (2001). Computer-assisted high-performance liquid
908 chromatography method development with applications to the isolation and analysis of
909 phytoplankton pigments. *Journal of Chromatography A*, *910*, 31-49.
910 Vidussi, F., Claustre, H., Manca, B. B., Luchetta, A., & Marty, J.-C. (2001). Phytoplankton
911 pigment distribution in relation to upper thermocline circulation in the eastern Mediterranean
912 Sea during winter. *Journal of Geophysical Research*, *106*, 19939–19956.

913 Wright, S. W., Thomas, D. P., Marchant, H. J., Higgins, H. W., Mackey, M. D., & Mackey, D. J.
914 (1996). Analysis of phytoplankton of the Australian sector of the Southern Ocean:
915 comparisons of microscopy and size-frequency data with interpretations of pigment HPLC
916 data using the CHEMTAX matrix factorisation program. *Marine Ecology Progress Series*,
917 *144*, 285.98.
918 Zapata, M., Rodriguez, F., & Garrido, J. L. (2000). Separation of chlorophylls and carotenoids
919 from marine phytoplankton: A new HPLC method using a reversed-phase C8 column and
920 pyridine-containing mobile phases. *Marine. Ecology Progress Series*, *195*, 29-45.
921

Table 1. Summary of phytoplankton pigment data for the nine stations selected in this study. The stations are sorted into different classes characterized by differing pigment assemblages based upon the ratios of the concentrations of two dominant accessory pigments to total chlorophyll-*a*, TChl*a* (see the 2nd and 3rd columns from the left). The ratios of the concentrations of six dominant pigments to TChl*a* are also displayed, with the two most dominant accessory pigments indicated within the shaded areas. Pigment abbreviations are: MVChl*a* = monovinyl chlorophyll-*a*, DVChl*a* = divinyl chlorophyll-*a*, Fuco = fucoxanthin, Hex = 19' hexanoyloxyfucoxanthin, But = 19' butanoloxyfucoxanthin, MVChl*b* = monovinyl chlorophyll-*b*, Chl*c2* = chlorophyll-*c2*, Zea = zeaxanthin, Pra = prasinoxanthin, Dia = diadinoxanthin, and α-caro = *α*-carotene.

Station ID	2 dominant pigments	Class	TChl*a* [mg/m^3]	Ratio of concentrations of dominant pigments to TChl*a*					
1	Fuco ≈ MVChl*b*	A	0.62	MVChl*a*	Fuco	MVChl*b*	Chl*c2*	Pra	Hex
				0.95	0.22	0.21	0.11	0.07	0.07
6	DVChl*a* > Zea	B	0.28	DVChl*a*	MVChl*a*	Zea	Hex	*α*-caro	Chl*c2*
				0.50	0.47	0.31	0.18	0.09	0.09
12	DVChl*a* ≈ Zea	C1	0.14	MVChl*a*	DVChl*a*	Zea	Hex	Chl*c2*	But
				0.54	0.44	0.40	0.24	0.09	0.09
37	DVChl*a* ≈ Zea	C2	0.15	Zea	DVChl*a*	MVChl*a*	Hex	Dia	*a*-caro
				0.63	0.58	0.40	0.17	0.10	0.10
44	DVChl*a* ≈ Zea	C3	0.22	DVChl*a*	Zea	MVChl*a*	Hex	*α*-caro	Chl*c2*
				0.56	0.50	0.42	0.17	0.08	0.08
46	DVChl*a* ≈ Zea	C4	0.14	Zea	DVChl*a*	MVChl*a*	Hex	Dia	MVChl*b*
				0.52	0.50	0.49	0.20	0.09	0.08
48	Hex > Zea	D	0.21	MVChl*a*	Hex	Zea	DVChl*a*	But	Dia
				0.76	0.48	0.25	0.21	0.18	0.16
51	Hex > Fuco	E	0.26	MVChl*a*	Hex	Fuco	Dia	Chl*c2*	But
				0.86	0.35	0.24	0.23	0.18	0.13
59	Zea ≈ Hex	F	0.11	MVChl*a*	Zea	Hex	Dia	DVChl*a*	But
				0.80	0.39	0.34	0.21	0.18	0.12

936 Table 2. A comparison of similarity indices between pigment-based clusters and reflectance-
937 based clusters for the different sources of reflectance data that are depicted in Fig. 12. For the
938 case of the second derivative of hyperspectral reflectance, the result of computations for two
939 different spectral regions is given.

Reflectance data	Rand index	Cophenetic index
3 band ratios based on 4 SeaWiFS bands	0.69	0.39
Multispectral (13 bands)	0.69	0.39
Hyperspectral (325 bands, range 300-725 nm)	0.69	0.39
Hyperspectral 2nd derivative (300-725 nm)	0.78	0.53
Hyperspectral 2nd derivative (435-495 nm)	1	0.65

940

941

Figure captions

Figure 1. Map depicting the location of full stations sampled along the north-to-south ANT-XXIII/1 cruise track in the eastern Atlantic during October and November, 2005. Each full station consisted of in situ optical measurements accompanied by discrete water sample analyses. Stations chosen for use in the cluster analysis are identified by filled circles and labeled with the station ID.

Figure 2. Hyperspectral (1 nm) determinations of the remote-sensing reflectance $R_{rs}(\lambda)$ obtained from radiative transfer simulations (solid line) compared with in situ multispectral measurements at 13 discrete bands (solid circles). Each panel illustrates a different station location.

Figure 3. A schematic diagram illustrating the general approach to hierarchical cluster analysis and similarity determination. The dendrogram obtained with pigment composition as the input (upper pathway) is used as the reference for comparison with results obtained utilizing various optical data as input.

Figure 4. (a) Dendrogram obtained for the nine stations using 24 pigment to total chlorophyll-*a* (TChl*a*) ratios determined from the CHORS HPLC analysis. (b) Linkage distances obtained from the cluster analysis shown in (a) as a function of distance along the dendrogram. (c) Similar to (a), but using input ratios calculated from 8 diagnostic pigments (see text for details).

Figure 5. Chlorophyll-specific normalized absorption coefficients for the nine stations

corresponding to: (a) absorption of phytoplankton, $a_{n,ph}^*(\lambda)$, (b) absorption of pure seawater plus phytoplankton, $a_{n,w+ph}^*(\lambda)$, and (c) total absorption, $a_n^*(\lambda)$.

Figure 6. Results of cluster analysis applied to absorption data from the nine stations. The left panels represent dendrograms obtained using the different absorption components of (a) $a_n^*(\lambda)$, (c) $a_{n,w+ph}^*(\lambda)$, and (e) $a_{n,ph}^*(\lambda)$, and the right panels (b, d, f) illustrate results obtained using each respective component's second derivative spectrum.

Figure 7. Similarity indices between absorption-based and pigment-based cluster trees obtained for the nine stations using different combinations of spectral range for (a) $a_n^*(\lambda)$, (b) $a_{n,w+ph}^*(\lambda)$ and (c) $a_{n,ph}^*(\lambda)$. The y-axis indicates the lower limit of the spectral range (λ_{min}) and the x-axis the upper limit of the spectral range (λ_{max}) utilized in the cluster analysis. Left and right panels depict the cophenetic and Rand indices, respectively.

Figure 8. Similar to Fig. 7, but based on cluster trees obtained using different spectral range combinations for the second derivative spectra of (a) $a_n^*(\lambda)$, (b) $a_{n,w+ph}^*(\lambda)$ and (c) $a_{n,ph}^*(\lambda)$. Optimal values for band separation and window size determined from prior analyses ($BS = WS =$ 9 nm, see text for details) were used in the calculation of derivative spectra.

Figure 9. Histograms of cophenetic indices obtained for all combinations of spectral ranges shown in Fig. 8 based on (a) the absorption of phytoplankton, $a_{n,ph}^*(\lambda)$, and (b) its second derivative spectra.

Figure 10. Cophenetic indices obtained from the cluster and similarity analysis of the second derivative spectra of $a_{n,ph}$*, in which different parameter sets for the derivative analysis are considered. The analysis was conducted using the optimal spectral region from 420 to 515 nm. The y-axis indicates the size of the filter window used in smoothing of the absorption spectra (*WS*), and the x-axis represents the band separation used in the calculation of the derivative calculation (*BS*).

Figure 11. Hydrolight-simulated $R_{rs}(\lambda)$ spectra, normalized at 555 nm, computed for the nine stations using measured IOPs as input.

Figure 12. Dendrograms resulting from cluster analysis of the nine stations calculated using four different sets of input data vectors: (a) three reflectance band ratios of $R_{rs}(\lambda)$ based on 4 SeaWiFS wavebands obtained from measurements with SPMR instrument, (b) 13 band ratios corresponding to multispectral measurements of $R_{rs}(\lambda)$ with SPMR instrument, (c) hyperspectral (1 nm) ordinary (non-differentiated) normalized $R_{rs}(\lambda)$ spectra computed from the Hydrolight simulations, and (d) second derivative of hyperspectral normalized $R_{rs}(\lambda)$ spectra obtained using optimal values for band separation and smoothing filter window (i.e., $BS = WS = 27$ nm).

Figure 13. Linkage distances as a function of distance along the dendrogram for each cluster tree depicted in Fig.12.

Figure 14. Similarity indices between reflectance-based and pigment-based cluster trees obtained for the nine stations using different combinations of spectral ranges for the second derivative of the hyperspectral normalized $R_{rs}(\lambda)$. Optimal values of *BS* = *WS* = 27 nm, determined from prior analyses, were used for the calculation of derivative spectra. The y-axis indicates the lower limit of the spectral range (λ_{min}) and the x-axis the upper limit of the spectral range (λ_{max}) utilized in the cluster analysis. Panels (a) and (b) depict the cophenetic and Rand indices, respectively.

Figure 15. (a) Cophenetic and (b) Rand indices obtained from the comparison of pigment-based cluster trees with trees computed from the second derivative of hyperspectral normalized $R_{rs}(\lambda)$, in which different choices for parameters of the derivative analysis are considered. The analysis has been carried out for the optimal spectral region from 435 and 495 nm. The y-axis indicates the size of the filter window used in the smoothing of spectra (*WS*), and the x-axis the band separation used in the derivative calculation (*BS*).

Figure 16. The second derivative spectra at each station of the (a) chlorophyll-specific normalized phytoplankton absorption coefficient, $a_{n,ph}^*(\lambda)$, and (b) normalized hyperspectral remote-sensing reflectance, $R_{rs}(\lambda)/R_{rs}(555)$. The derivative spectra are depicted for the optimal spectral ranges of 370 - 716 nm for absorption, and 435 - 510 nm for reflectance. Optimal values for band separation and smoothing filter window were used for the derivative calculations, i.e., *BS* = *WS* = 9 nm for the absorption data and *BS* = *WS* = 27 nm for the reflectance data.

Figure 1

Click here to download high resolution image

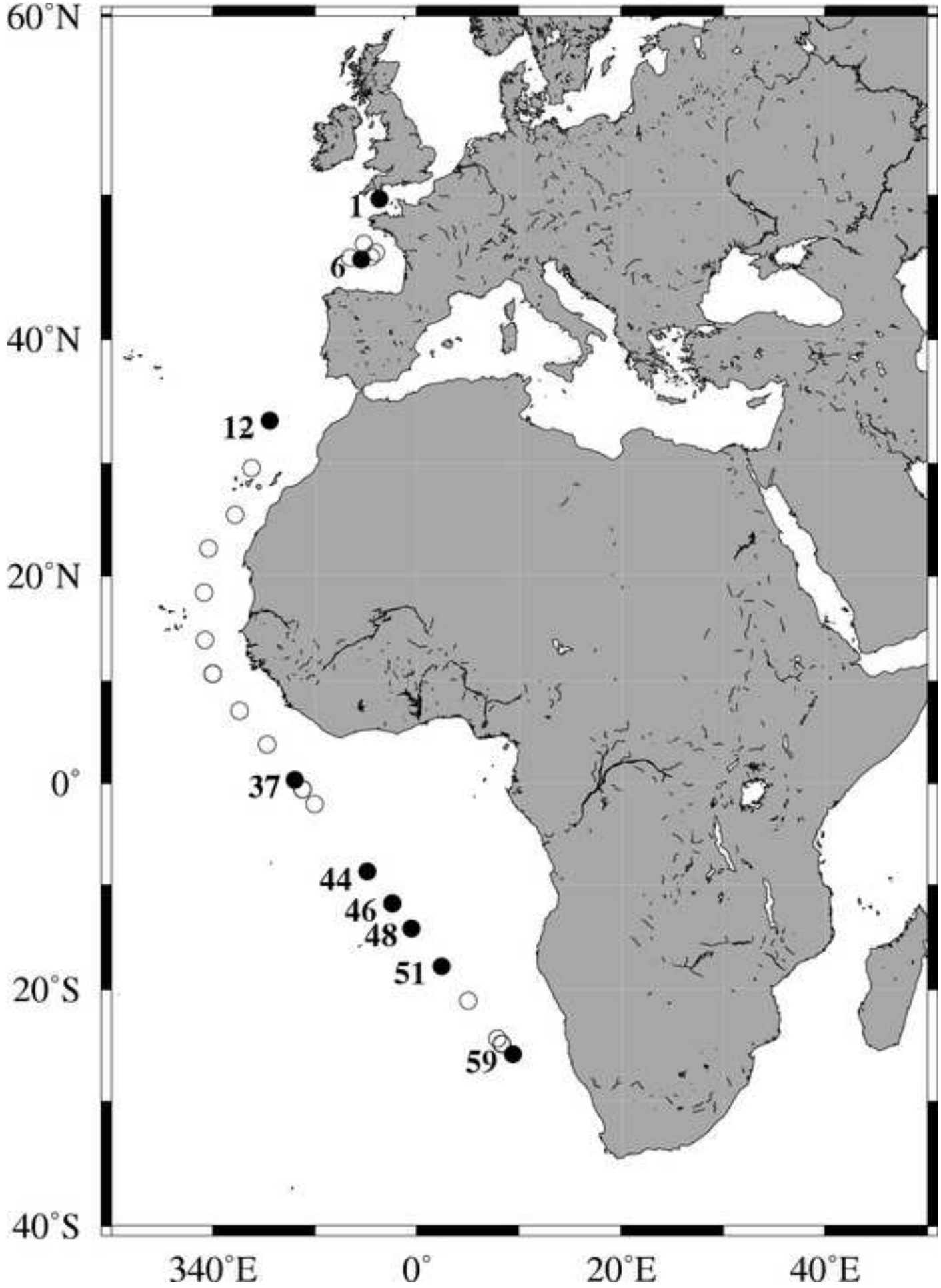

Figure 2

Click here to download high resolution image

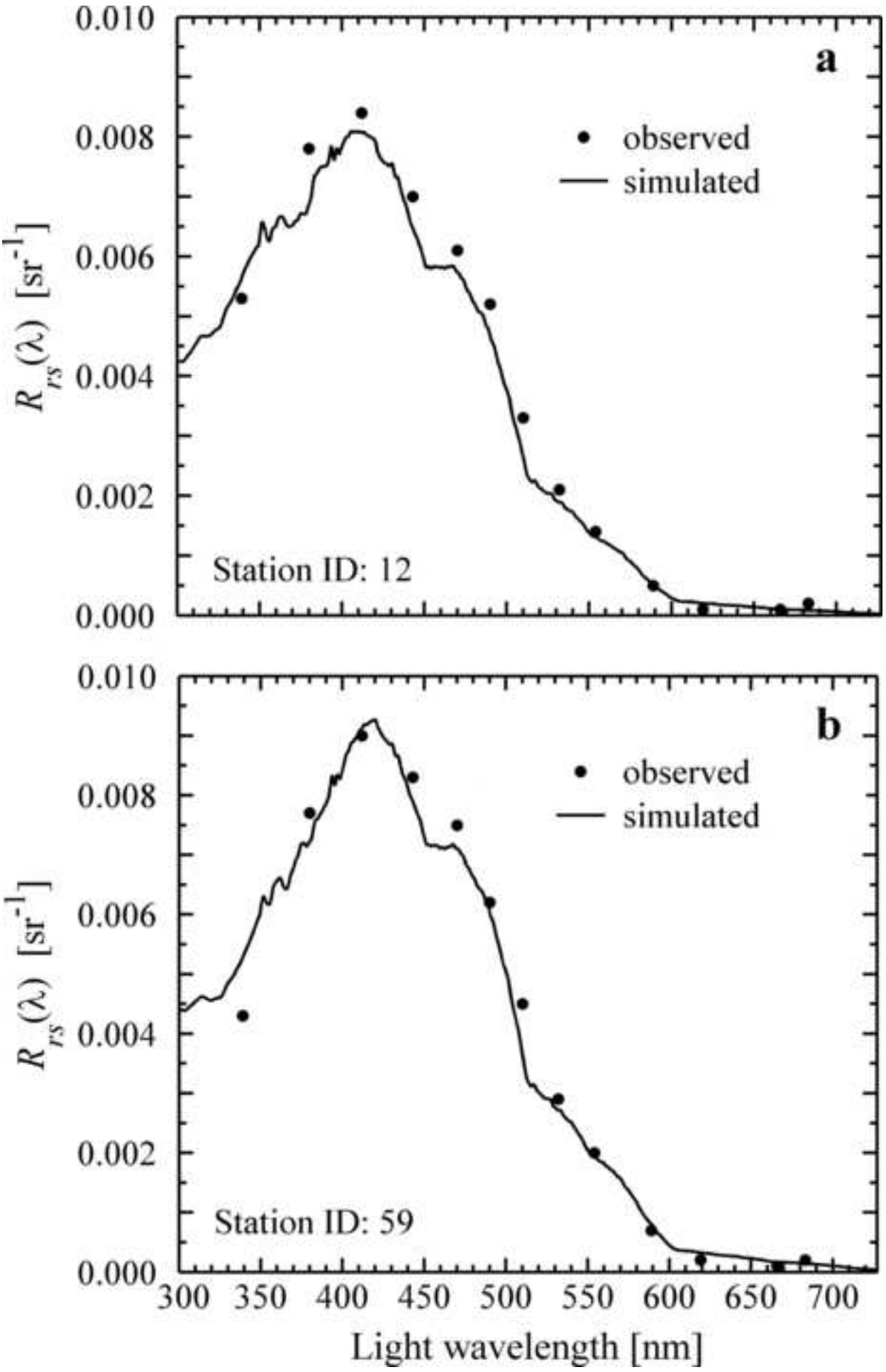

Figure 3

Click here to download high resolution image

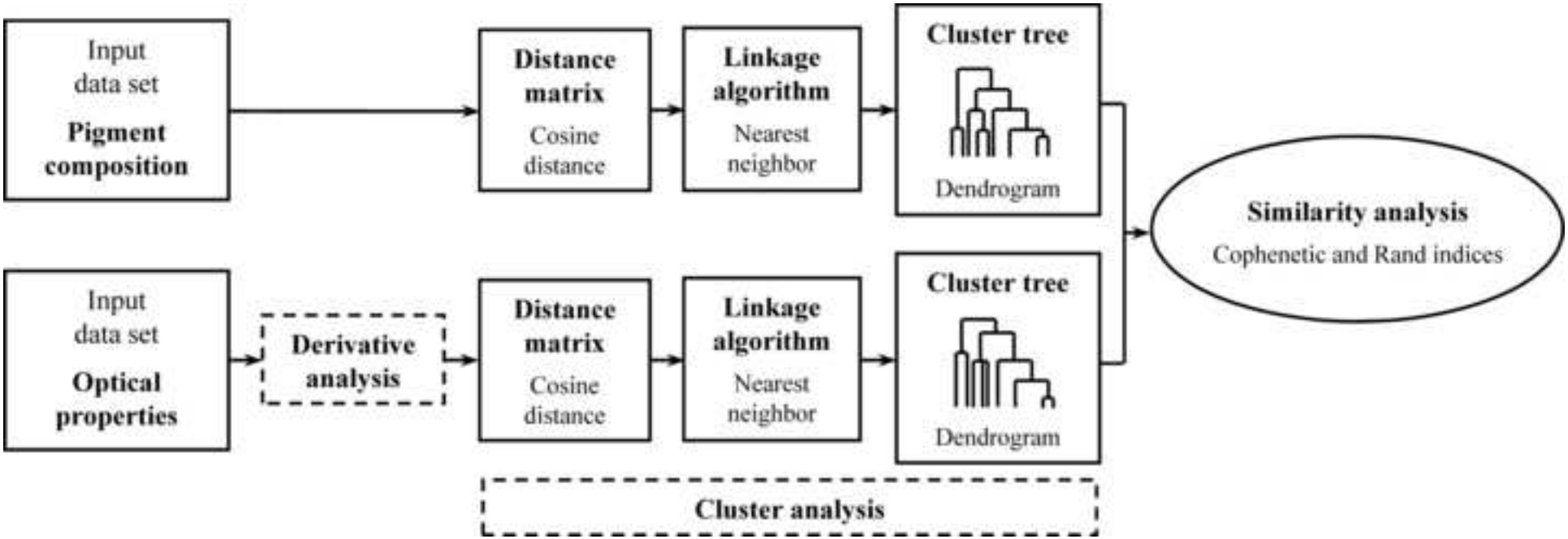

Figure 4

Click here to download high resolution image

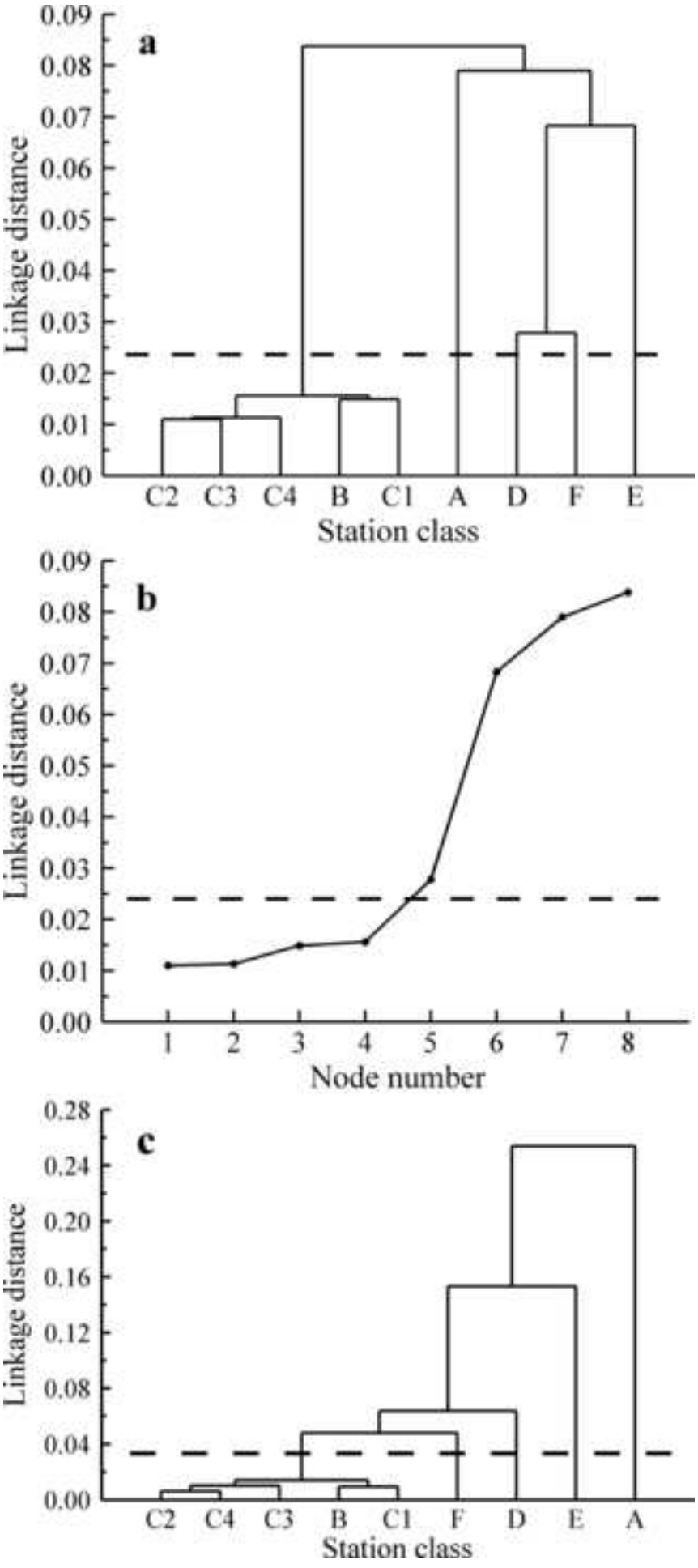

Figure 5

Click here to download high resolution image

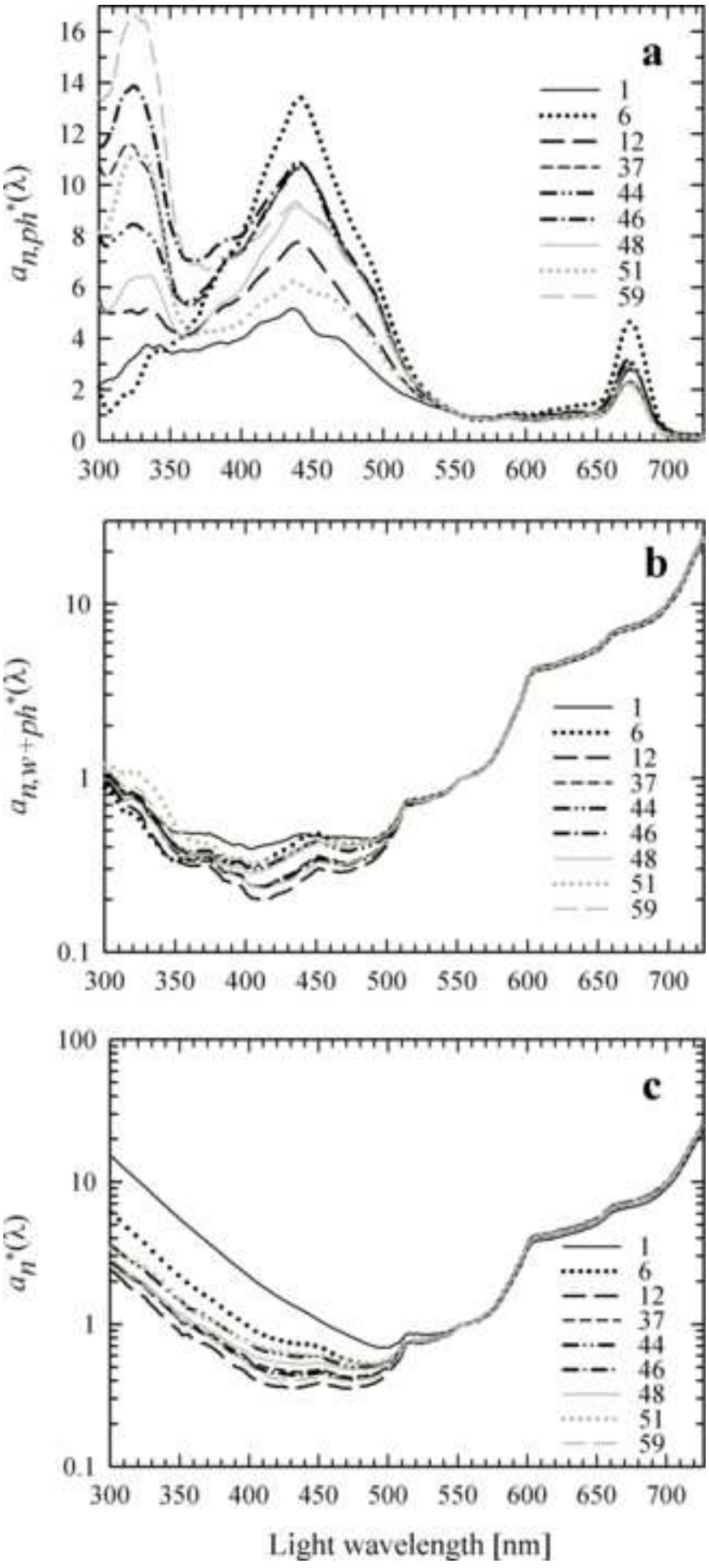

Figure 6

Click here to download high resolution image

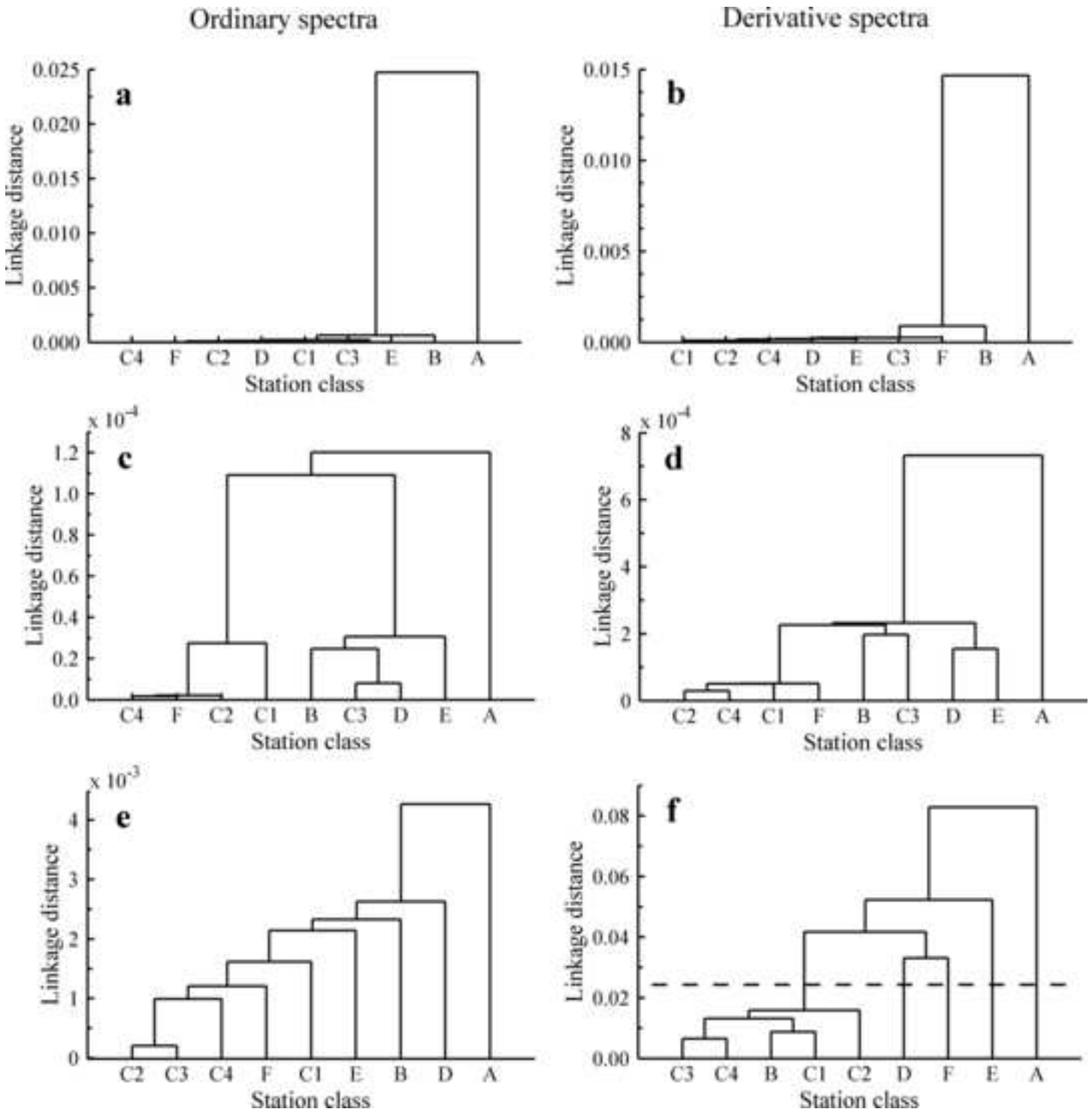

Figure 7

Click here to download high resolution image

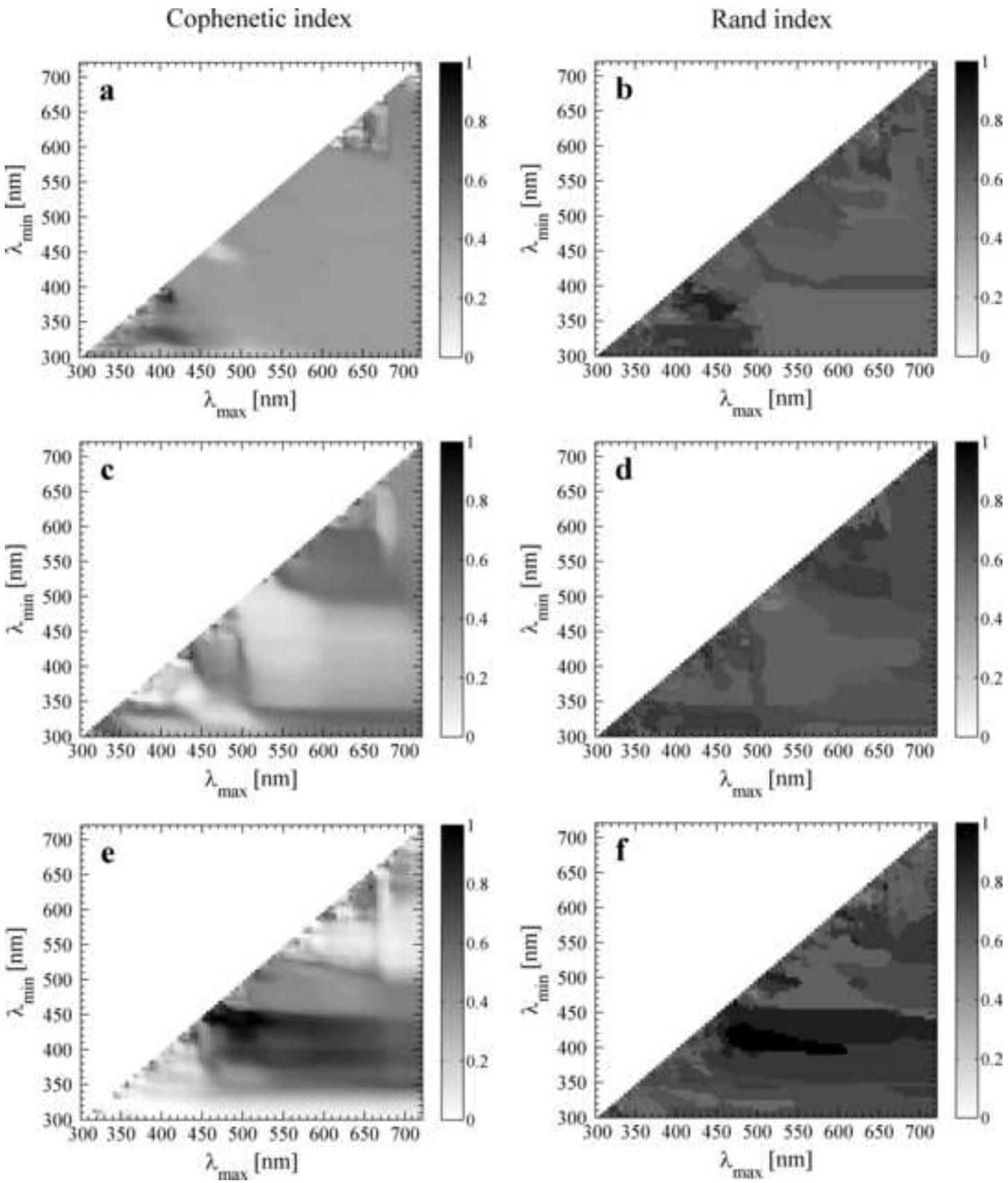

Figure 8

Click here to download high resolution image

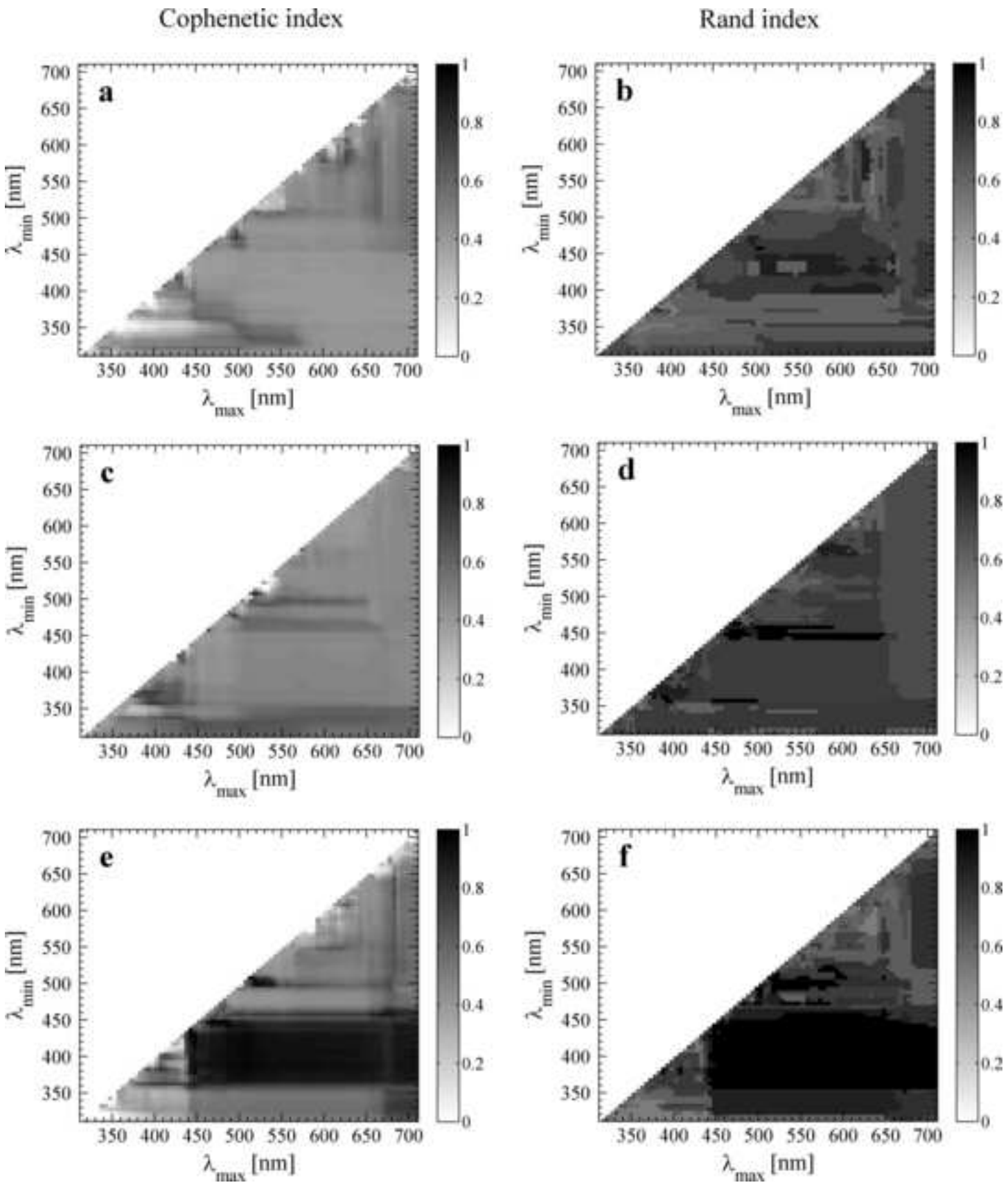

Figure 9

Click here to download high resolution image

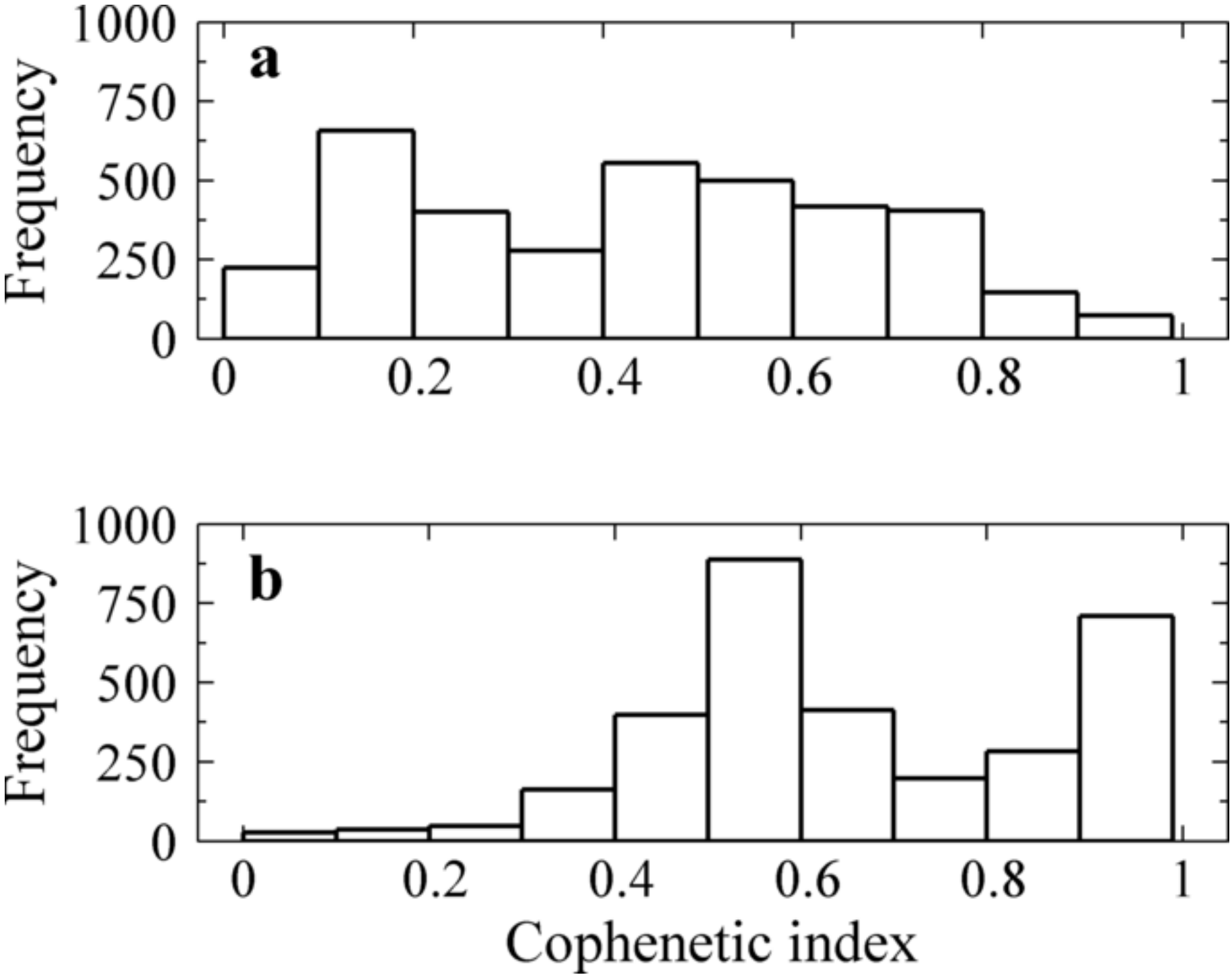

Figure 10

Click here to download high resolution image

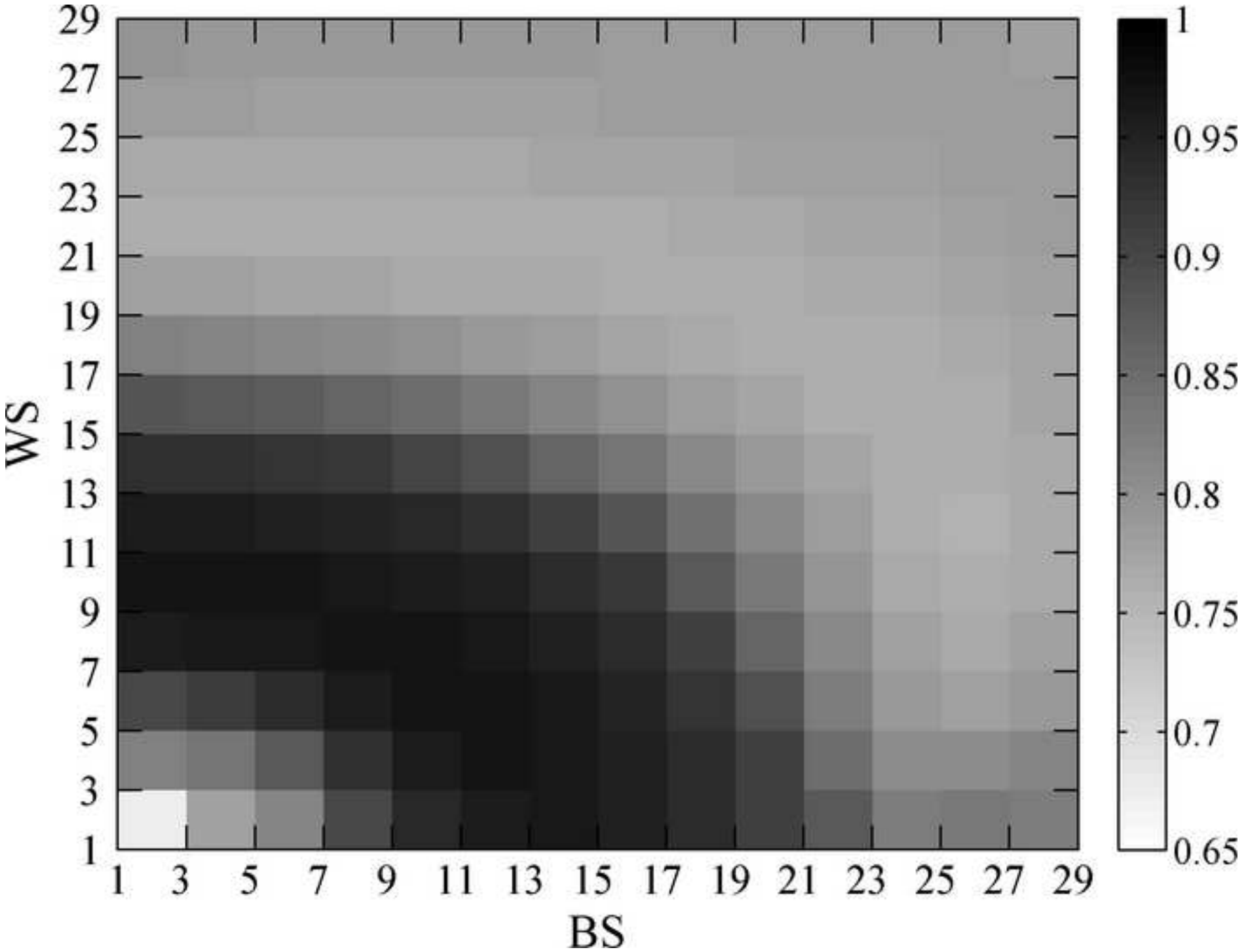

Figure 11

Click here to download high resolution image

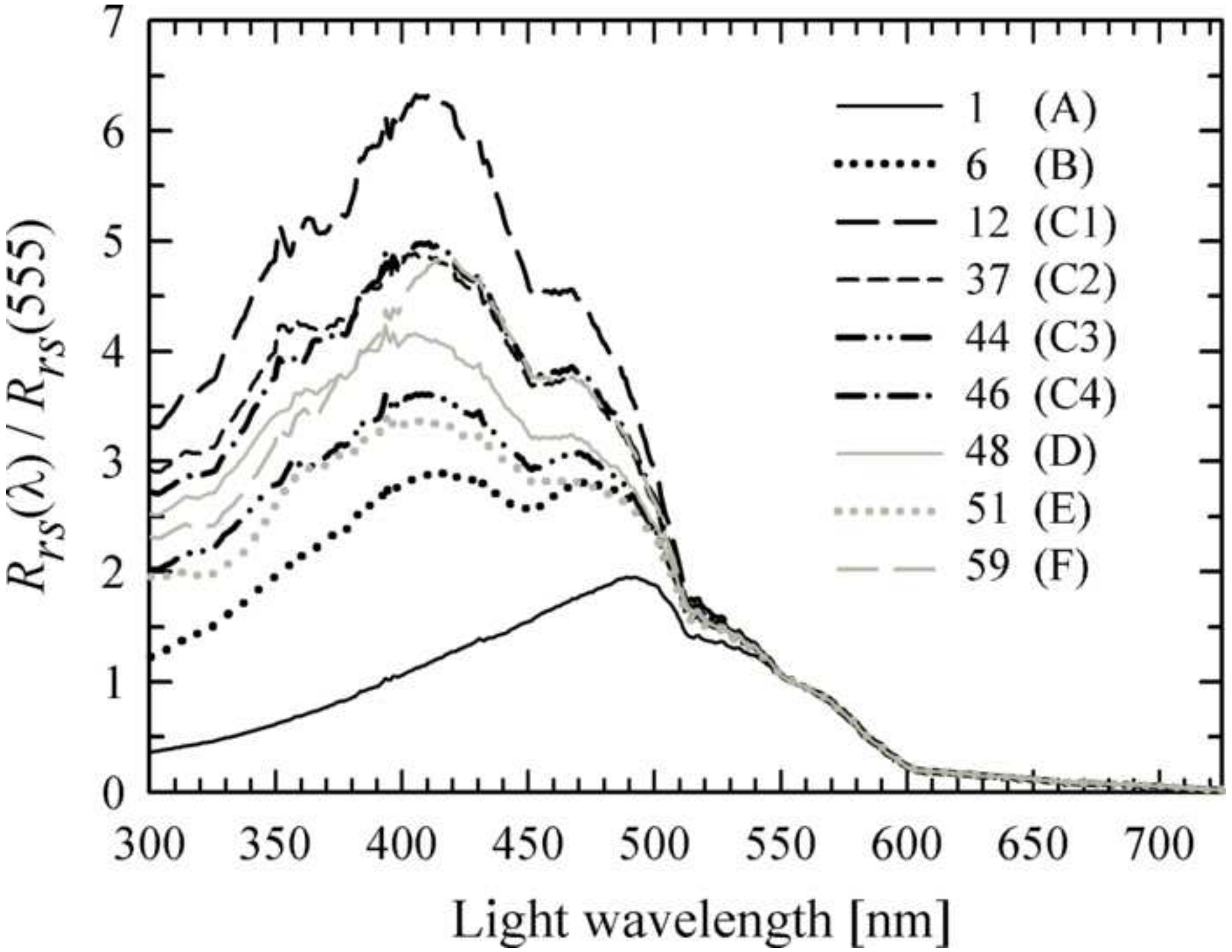

Figure 12

Click here to download high resolution image

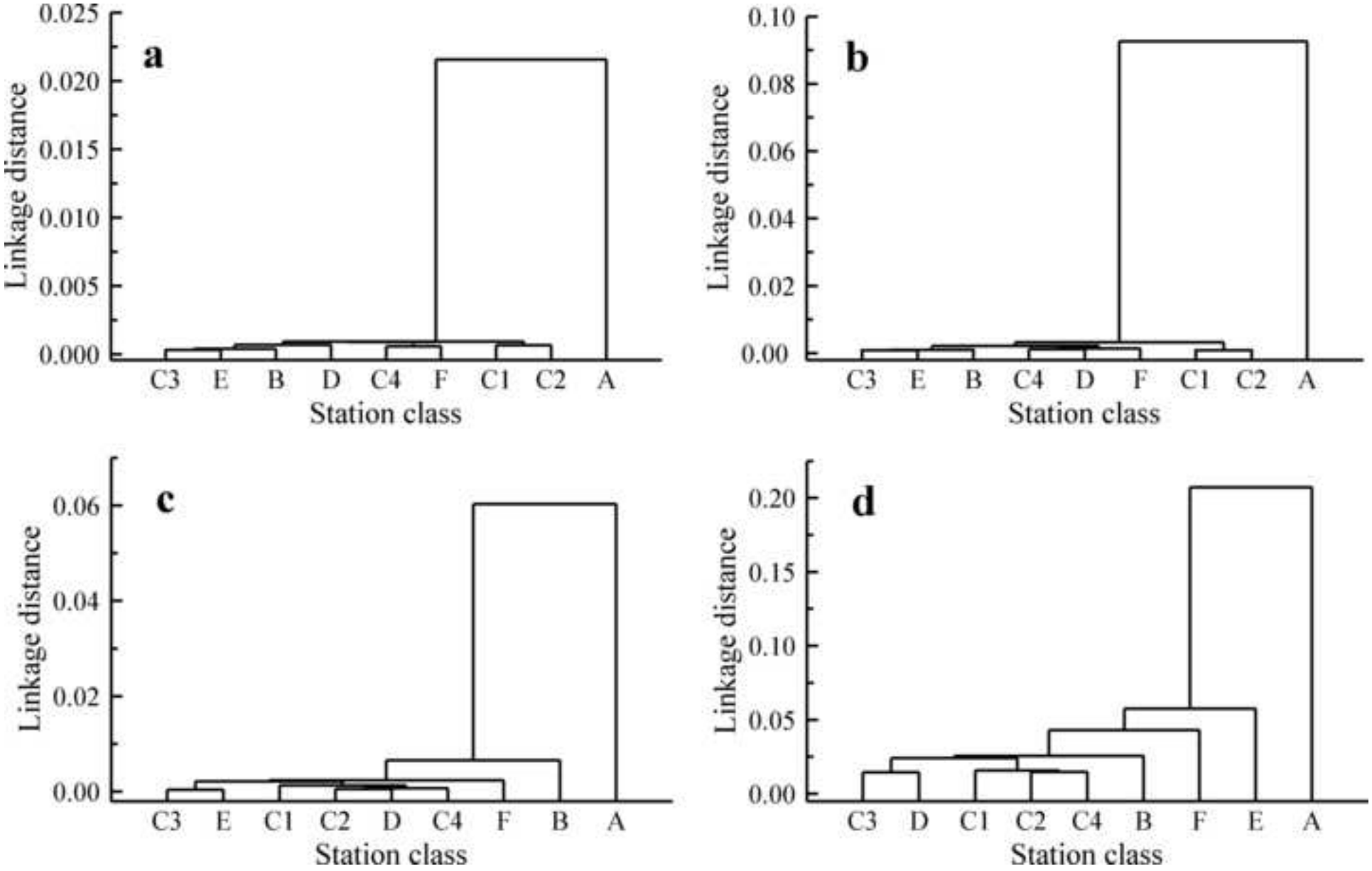

Figure 13

Click here to download high resolution image

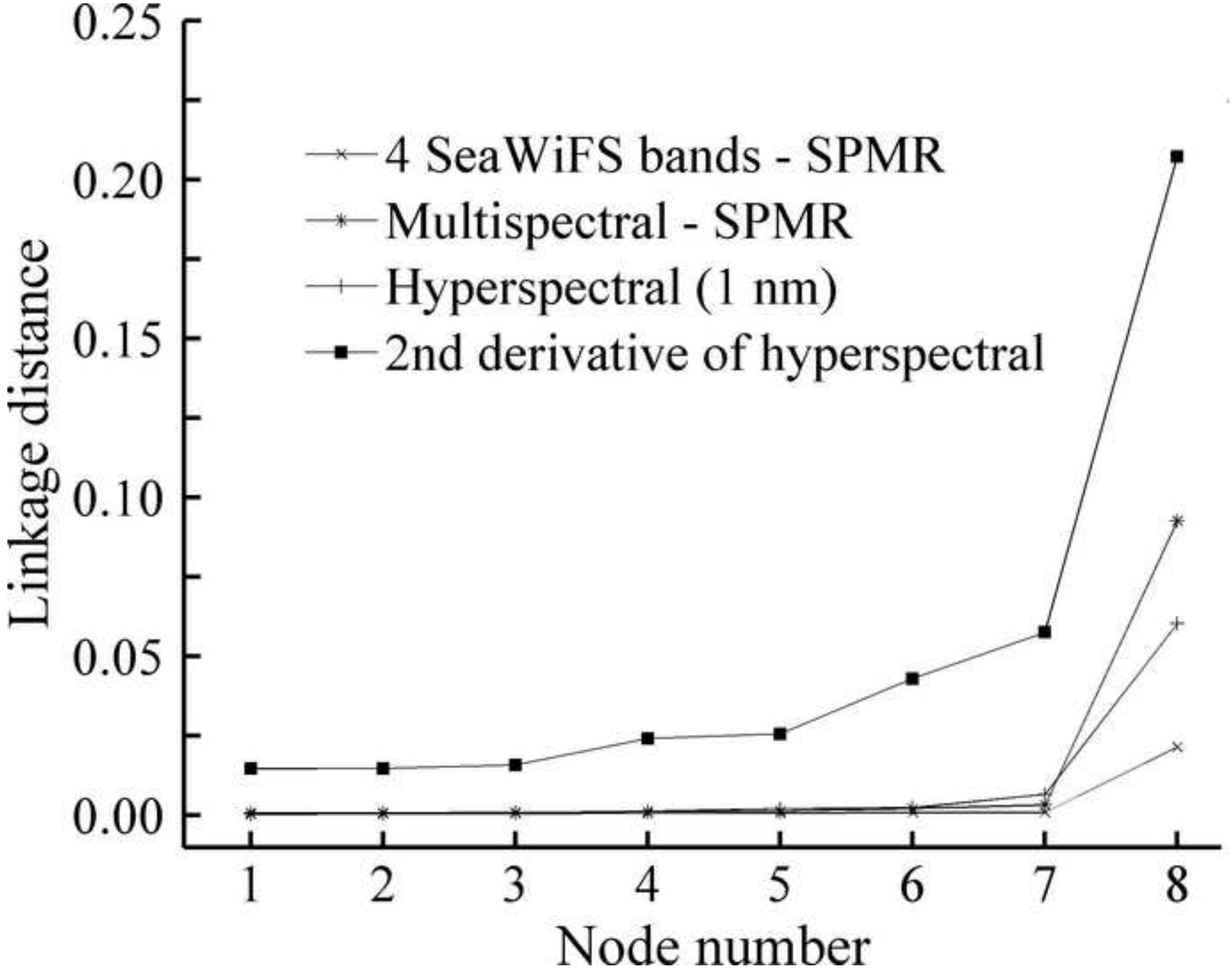

Figure 14

Click here to download high resolution image

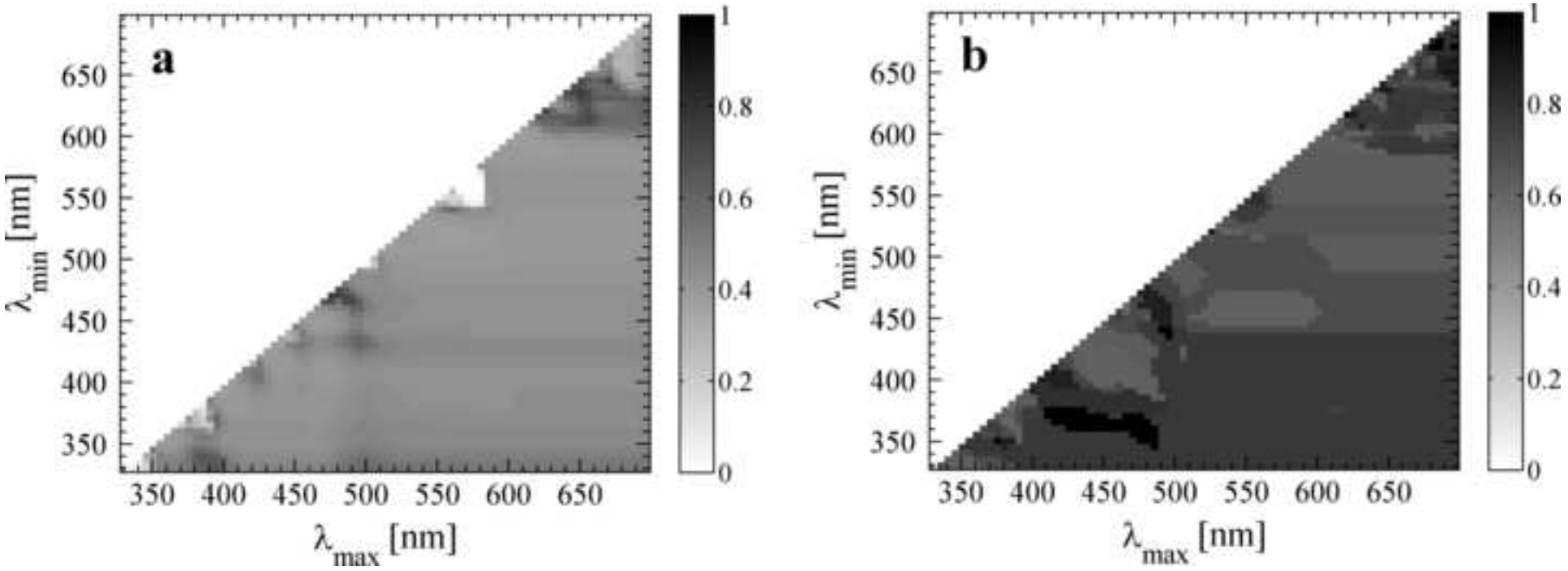

Figure 15

Click here to download high resolution image

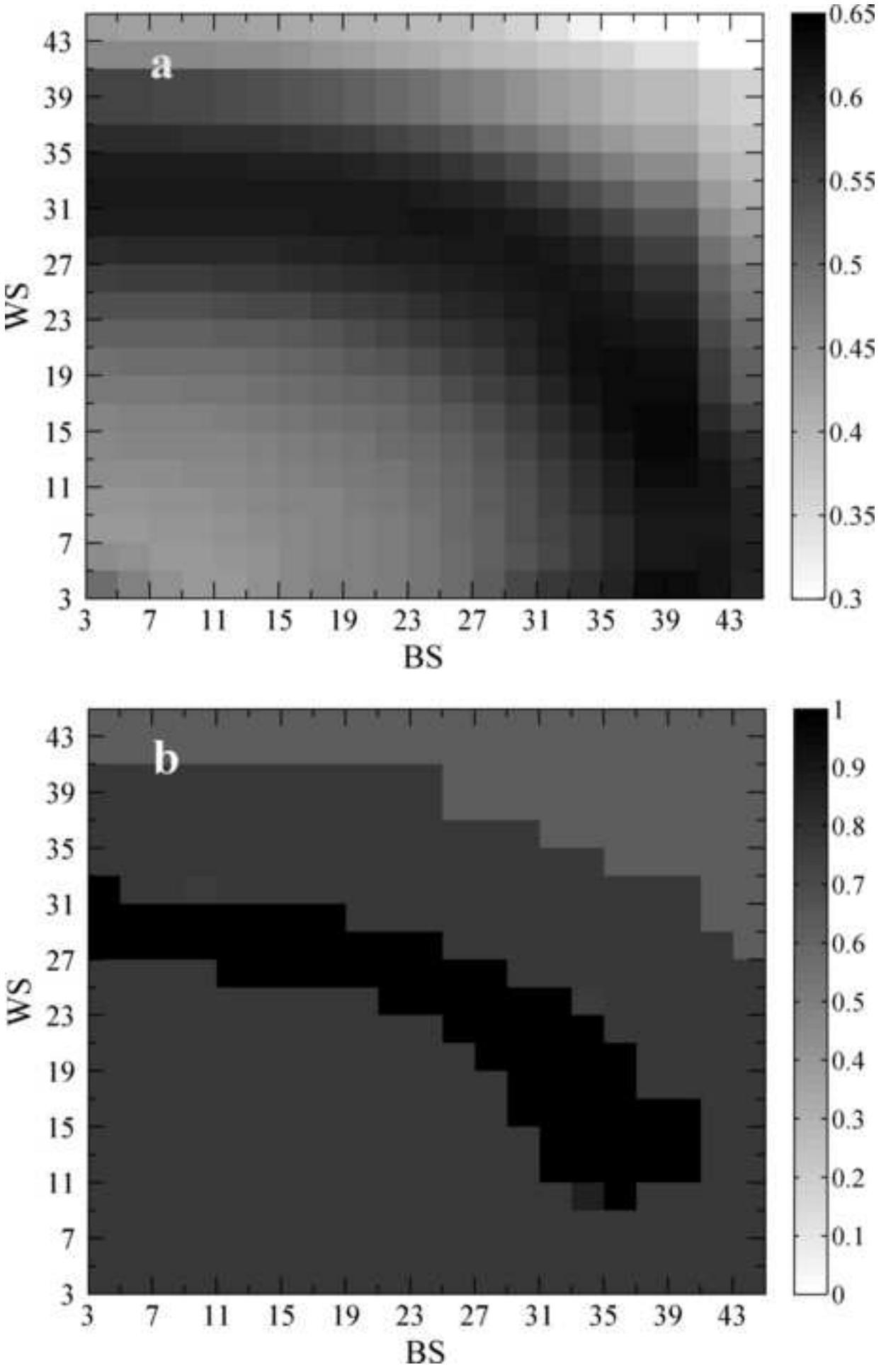

Figure 16

Click here to download high resolution image