


## Optical types of inland and coastal waters

Evangelos Spyrakos <sup>1\*</sup>, Ruth O'Donnell,<sup>2</sup> Peter D. Hunter,<sup>1</sup> Claire Miller,<sup>2</sup> Marian Scott,<sup>2</sup> Stefan G. H. Simis,<sup>3</sup> Claire Neil,<sup>1</sup> Claudio C. F. Barbosa,<sup>4</sup> Caren E. Binding,<sup>5</sup> Shane Bradt,<sup>6</sup> Mariano Bresciani,<sup>7</sup> Giorgio Dall'Olmo,<sup>3</sup> Claudia Giardino,<sup>7</sup> Anatoly A. Gitelson,<sup>8</sup> Tiit Kutser,<sup>9</sup> Lin Li,<sup>10</sup> Bunkei Matsushita,<sup>11</sup> Victor Martinez-Vicente,<sup>3</sup> Mark W. Matthews,<sup>12</sup> Igor Ogashawara,<sup>10</sup> Antonio Ruiz-Verdú,<sup>13</sup> John F. Schalles,<sup>14</sup> Emma Tebbs,<sup>15</sup> Yunlin Zhang,<sup>16</sup> Andrew N. Tyler<sup>1</sup>

<sup>1</sup>Biological and Environmental Sciences, School of Natural Sciences, University of Stirling, Stirling, United Kingdom

<sup>2</sup>School of Mathematics and Statistics, University of Glasgow, Glasgow, United Kingdom

<sup>3</sup>Plymouth Marine Laboratory, Plymouth, United Kingdom

<sup>4</sup>Image Processing Division, Nacional Institute for Space Research-INPE, Sao Jose dos Campos, Sao Paulo, Brazil

<sup>5</sup>Water Science and Technology Directorate, Environment and Climate Change Canada, Burlington, Ontario, Canada

<sup>6</sup>Department of Biological Sciences, University of New Hampshire, Durham, New Hampshire

<sup>7</sup>Institute for Electromagnetic Sensing of the Environment, CNR-IREA, Milano, Italy

<sup>8</sup>Department of Civil and Environmental Engineering, Israel Institute of Technology, Technion City, Haifa, Israel

<sup>9</sup>Estonian Marine Institute, University of Tartu, Tallinn, Estonia

<sup>10</sup>Planetary and Environmental Remote Sensing Lab, Indiana University-Purdue University at Indianapolis, Indiana

<sup>11</sup>Faculty of Life & Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki, Japan

<sup>12</sup>CyanoLakes (Pty) Ltd, Cape Town, South Africa

<sup>13</sup>Image Processing Laboratory (IPL), Universitat de València Catedrático José Beltrán, Paterna, València, Spain

<sup>14</sup>Department of Biology, Creighton University, Omaha, Nebraska

<sup>15</sup>Department of Geography, King's College London, London, United Kingdom

<sup>16</sup>Taihu Lake Laboratory Ecosystem Research Station, State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, PR China

### Abstract

Inland and coastal waterbodies are critical components of the global biosphere. Timely monitoring is necessary to enhance our understanding of their functions, the drivers impacting on these functions and to deliver more effective management. The ability to observe waterbodies from space has led to Earth observation (EO) becoming established as an important source of information on water quality and ecosystem condition. However, progress toward a globally valid EO approach is still largely hampered by inconsistencies over temporally and spatially variable in-water optical conditions. In this study, a comprehensive dataset from more than 250 aquatic systems, representing a wide range of conditions, was analyzed in order to develop a typology of optical water types (OWTs) for inland and coastal waters. We introduce a novel approach for clustering in situ hyperspectral water reflectance measurements ( $n = 4045$ ) from multiple sources based on a functional data analysis. The resulting classification algorithm identified 13 spectrally distinct clusters of measurements in inland waters, and a further nine clusters from the marine environment. The distinction and characterization of OWTs was supported by the availability of a wide range of coincident data on biogeochemical and inherent optical properties from inland waters. Phylogenetic trees based on the shapes of cluster means were constructed to identify similarities among the derived clusters with respect to spectral diversity. This typification provides a valuable framework for a globally applicable EO scheme and the design of future EO missions.

\*Correspondence: [evangelos.spyrakos@stir.ac.uk](mailto:evangelos.spyrakos@stir.ac.uk)

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

The global importance of aquatic systems is incontestable since they play a fundamental role in biogeochemical cycling, the maintenance of biodiversity, and human well-being and prosperity (Galloway et al. 2004; World Resources Institute 2005; Cole et al. 2007; Borges et al. 2015; Le Quére et al. 2015) and as such are fundamental to the delivery of

the UN Sustainable Development Goals. Nevertheless, several aspects of their role in these processes remain unclear (Raymond et al. 2013), while their resilience to changing environmental conditions and anthropogenic disturbance is still poorly understood (Fabry et al. 2008; Petrescu et al. 2015). Globally valid approaches for the study of these processes based only on field data is typically hindered by their high variability in both temporal and spatial scales (Dickey 2003; Peters et al. 2007). Furthermore, the sheer number of waterbodies and their geographic remoteness hampers their systematic study (Karl 1999; Verpoorter et al. 2014).

Satellite remote sensing offers a means to quantify physical and biogeochemical processes in aquatic systems at large scales, providing valuable insights into mechanisms associated with biogeochemical cycles, the climate system and its changes (Yang et al. 2013; Guo et al. 2015; Hestir et al. 2015). The rapidly increasing rate of data collection from Earth observation (EO) missions suitable for observing waterbodies (e.g., European Space Agency [ESA] Envisat and Sentinel, National Aeronautics and Space Administration [NASA] Landsat and Aqua missions) offers long-term archives of our aquatic environments while advances in optical sensors support new and more detailed characterization of the Earth surface. Of particular interest is the remote sensing signal in the visible and infrared part of the spectrum since it comprises information on key color-forming substances such as phytoplankton pigments, suspended minerals, and dissolved compounds. Nonetheless, the wide range of possible combinations and composition of these substances found within and between aquatic systems challenges the applicability of EO techniques (Bukata 1995; Morel and Maritorena 2001; Mélin and Vantrepotte 2015). Numerous approaches have been developed for the retrieval of biogeochemical properties from remote sensing data (reviews in Acker et al. 2005; Matthews 2011; Odermatt et al. 2012; Blondeau-Patissier et al. 2014; Tyler et al. 2016) but quantifying the associated uncertainties when these are applied over different conditions has hitherto proved difficult.

Water optical typologies has been suggested as a mechanism to delineate water masses on the basis of their optical properties (Jerlov 1977; Prieur and Sathyendranath 1981; Baker and Smith 1982) and thereby schematize the application of EO methods (Arnone et al. 2004). As a result, a range of parameters linked to the observed variability in water color has been encompassed in classification schemes. These include water column parameters such as Secchi disk depth ( $Z_{SD}$ , see Table 1 for a list of symbols and acronyms) (e.g., Arnone 1985), inherent optical properties (IOPs; mainly absorption: e.g., Babin et al. 2003; Shi et al. 2014) as well as radiometric quantities measured below or above the water surface (e.g., Le et al. 2011; Moore et al. 2014).

Traditionally, the partitioning of water properties into optical types has been driven by the failure of retrieval algorithms, often developed for oceanic waters, to provide

accurate data in coastal and inland systems. In this context, Morel and Prieur (1977) distinguished two water types, depending on the predominance of phytoplankton and autochthonous production of dissolved and particulate detrital material (Case-1), or the input of external particulate and dissolved material into the system causing an uncoupling of phytoplankton with bulk optical properties (Case-2). More recent studies have moved toward the differentiation of water types in optically complex environments using in situ and/or satellite-derived reflectance data. Most of these studies have considered the range of optical classes in marine systems (English Channel and North Sea: Lubac and Loisel 2007; Tilstone et al. 2012; Vantrepotte et al. 2012, Iberian coastal waters: Spirakos et al. 2011; Adriatic Sea: Mélin et al. 2011, Yellow Sea: Ye et al. 2016; Northwest Atlantic shelf: Moore et al. 2001, global ocean: Moore et al. 2009, 2014, global coastal waters: Mélin and Vantrepotte 2015) with only a few studies focussed on inland systems (lakes and reservoirs in China: Le et al. 2011; Shen et al. 2015; Estonian and Finnish lakes: Reinart et al. 2003). Overall, these classification schemes can substantially improve the remote sensing products associated with individual optical water types (OWTs), and have demonstrated the need for a better understanding of the underlying variability especially in nearshore and inland waterbodies (Moore et al. 2014). In parallel, optical water typologies based on remote sensing data have found further applications in ecological studies (Martin Traykovski and Sosik 2003), the detection of blooms (comprehensive list in Blondeau-Patissier et al. 2014) and in the more detailed study of the relationships between absorption parameters and water constituents especially when these can be determined in large datasets from different aquatic systems (Torrecilla et al. 2011).

Several hierarchical, partitional, and hybrid (Jain et al. 1999) clustering techniques have been implemented for the classification of remote sensing reflectance ( $R_{rs}$ ) into groups based upon differences in magnitude and shape. Consequently, techniques including agglomerative hierarchical (Shi et al. 2014),  $k$ -means clustering (Palacios et al. 2012), fuzzy clustering (González Vilas et al. 2011; Moore et al. 2014), and artificial neural networks (Canziani et al. 2008) have been used to uncover clusters present in these datasets using different degrees of implicit or explicit knowledge. While these approaches provide useful insights into the differentiation of water masses based on their optical properties, they have often lacked a comprehensive analysis of the physical basis to the definition of the clusters in terms of their variability in IOPs and biogeochemical significance. Moreover, few studies have considered the relations between OWTs found in coastal and inland aquatic systems. In spite of the progress made in the development of these methodologies, a solid foundation for dealing with high data dimensionality, uncertainty due to the use of different sensors, and variability in the relevant spectral features is still lacking.

**Table 1.** Symbols and acronyms.

Symbols/acronyms	Description	Units
EO	Earth observation	—
OWTs	Optical water types	—
$\lambda$	Wavelength	nm
<i>Datasets</i>		
LIMNADES	Lake Bio-optical Measurements and Matchup Data for Remote Sensing	—
SeaBASS	SeaWiFS Bio-optical Archive and Storage System	—
I	Inland waters only	—
C	Coastal waters only	—
N	All waters	—
<i>Biogeochemical parameters</i>		
Chl <i>a</i>	(Concentration of) Chl <i>a</i>	mg m <sup>-3</sup>
PC	(Concentration of) Phycocyanin	mg m <sup>-3</sup>
TSM	(Concentration of) Total suspended matter	mg L <sup>-1</sup>
ISM	(Concentration of) Inorganic suspended matter	mg L <sup>-1</sup>
CDOM	Colored dissolved organic matter	m <sup>-1</sup>
<b>IOP</b> <b>Inherent optical properties</b>		
$a_{\text{CDOM}}(\lambda)$	Absorption coefficient at wavelength $\lambda$ of CDOM	m <sup>-1</sup>
$a_{\text{NAP}}(\lambda)$	Absorption coefficient at wavelength $\lambda$ of “non-algal” particles (NAP)	m <sup>-1</sup>
$a_{\text{ph}}(\lambda)$	Absorption coefficient at wavelength $\lambda$ of phytoplankton	m <sup>-1</sup>
$S_{\text{CDOM}}$	Slope coefficient of $a_{\text{CDOM}}$ model: $a_{\text{CDOM}}(\lambda) = a_{\text{CDOM}}(\lambda_r) e^{-S_{\text{CDOM}}(\lambda - \lambda_r)} + K$ at reference wavelength (400 nm), where $K$ is a parameter used to offset baseline shifts unrelated to the absorption of CDOM	nm <sup>-1</sup>
$S_{\text{NAP}}$	Slope coefficient of $a_{\text{NAP}}$ model: $a_{\text{NAP}}(\lambda) = a_{\text{NAP}}(\lambda_r) e^{-S_{\text{NAP}}(\lambda - \lambda_r)} + K$ at reference wavelength (400 nm), where $K$ is a parameter used to offset baseline shifts unrelated to the absorption of NAP	nm <sup>-1</sup>
<b>sIOP</b> <b>Specific inherent optical properties</b>		
$a_{\text{NAP}}^*(\lambda)$	Absorption coefficient at wavelength $\lambda$ of NAP normalized to TSM concentration	g m <sup>-2</sup>
$a_{\text{ph}}^*(\lambda)$	Absorption coefficient at wavelength $\lambda$ of phytoplankton normalized to Chl <i>a</i> concentration	m <sup>2</sup> mg <sup>-1</sup>
<b>AOP</b> <b>Apparent optical properties</b>		
$R_{\text{rs}}(\lambda)$	Remote-sensing reflectance $R_{\text{rs}}$	sr <sup>-1</sup>
$Z_{\text{SD}}$	Secchi disk depth	m

The aim of the present study is to extend our knowledge of the optical diversity of aquatic systems, and in particular inland waters. To this end, a large database of observations from a range of different systems and a wide range of water conditions is used to: (1) obtain distinct OWTs; (2) develop a methodological approach for capturing key features found in the spectra based on functional data analysis; and (3) assess similarities and differences between inland waters and coastal marine systems. It is expected that the optical diversity of inland waterbodies exceeds that of marine systems, reflecting the wide diversity in morphology and surrounding land use of inland waters. Nevertheless, we expect that within and between regions recurrent OWTs can be detected, such as systems dominated by phytoplankton or by high light absorption due to dissolved matter. We subsequently investigate the extent to which OWTs can be approximated by a limited set of wavebands available from current and future remote sensors (“Implications for implementation to satellite imagery” section).

## Datasets

A large dataset (hereafter denoted Dataset-N) of 4035 in situ hyperspectral  $R_{\text{rs}}$  spectra from inland and coastal marine waters was used in the clustering analysis. The dataset consisted of data from more than 250 inland lakes, reservoirs and large rivers (Dataset-I, inland waters) and data from 14 campaigns in marine waters (Dataset-C, coastal waters). For this study, data were sourced from the in situ bio-optical data repositories LIMNADES (Lake Bio-optical Measurements and Matchup Data for Remote Sensing: <http://www.limnades.org>) and SeaBaSS (SeaWiFS Bio-optical Archive and Storage System: <http://seabass.gsfc.nasa.gov>).

## Inland aquatic systems

The LIMNADES data (Dataset-I) used here were compiled from 16 individual datasets of bio-optical and biogeochemical measurements from a variety of natural and artificial inland aquatic systems including mainly lakes and reservoirs but also rivers and floodplains. Table 2 summarizes these

**Table 2.** Description of the datasets from inland (I) water systems used in this work.

Dataset	Principal institute	Inland system(s)	References
I-A	CAS	Lake Taihu, China	Zhang et al. (2007, 2010)
I-B	CEDEX	56 reservoirs and 2 lakes in Spain	Ruiz-Verdú et al. (2005, 2008), Simis et al. (2007)
I-C	CNR	Five lakes in the Mediterranean and subalpine eco-regions of Italy	Bresciani et al. (2011), Giardino et al. (2005, 2014 <sup>a,b</sup> , 2015), Guanter et al. (2010), Manzo et al. (2015)
I-D	CU	43 sites in U.S. inland waters	Gitelson et al. (2007), Schalles (2006), Schalles and Hladik (2012)
I-E	EC	Erie; Ontario; Winnipeg (Canada and U.S.)	Binding et al. (2008, 2010, 2011, 2013)
I-F	INPE	Lago Grande de Curuai (Brazil)	Barbosa (2007)
I-G	IU	Three drinking water reservoirs in central Indiana (U.S.)	Li et al. (2013, 2015)
I-H	NIOO-KNAW	Lakes Loosdrechtse, Plassen and IJsselmeer, Netherlands	Guanter et al. (2010), Ruiz-Verdú et al. (2008), Simis et al. (2005, 2007)
I-I	UCT	Three South African reservoirs	Matthews (2014), Matthews and Bernard (2013)
I-J	UL	Lake Bogoria, Kenya	Tebbs et al. (2013)
I-K	UNH	62 lakes in New England, U.S. and Great Salt Lake, U.S.	Bradt (2012), Moore et al. (2014)
I-L	UNL-A	Several lakes and reservoirs in eastern Nebraska and northwest Iowa, U.S.	Dall'Olmo et al. (2003, 2005), Dall'Olmo and Gitelson (2005), Gitelson et al. (2008)
I-M	UNL-B	Fremont State Lakes (U.S.) and Lake Kinneret (Israel)	Gitelson et al. (2009), Gurlin et al. (2011), Yacobi et al. (2011)
I-N	USTIR	Lake Balaton and Four neighboring aquatic systems in Hungary Five lakes in the UK	Riddick et al. (2015)
I-O	UT	Lake Peipsi, Estonia	Kutser et al. (2012, 2013)
I-P	UTSU	Five lakes in Japan and China	Yang et al. (2013), Jaelani et al. (2013), Matsushita et al. (2015)

CAS, Chinese Academy of Sciences; CEDEX, Centro de Estudios Hidrográficos; CNR, Consiglio Nazionale delle Ricerche; CU, Creighton University; EC, Environment Canada; INPE, Instituto Nacional de Pesquisas Espaciais; IU, Indiana University; NIOO-KNAW, Netherlands Institute of Ecology; UCT, University of Cape Town; UL, University of Leicester; UNH, University of New Hemisphere; UNL, University of Nebraska-Lincoln; USTIR, University of Stirling; UT, University of Tartu; UTSU, University of Tsukuba.

datasets, providing references to detailed information including sources and spatial coverage. A total of 3025  $R_{rs}(\lambda)$  spectra across a wide range of system characteristics, conditions, and geographical conditions were used in the clustering analysis. Cluster analysis was initially performed only on Dataset-I  $R_{rs}(\lambda)$  to facilitate the determination of distinct OWTs solely in inland water systems. Paired measurements of IOPs and biogeochemical parameters were then used to support the characterization of the resulting clusters.

### Coastal systems

The well-documented SeaBaSS dataset (Dataset-C, Table 3) (Werdell and Bailey 2002; Werdell et al. 2003) was used for comparison to inland optical clusters as defined by the classification analysis. Data extracted from SeaBaSS were restricted to hyperspectral  $R_{rs}(\lambda)$  ( $n = 1010$ ) spectra from mainly coastal and but also some open ocean environments originally measured above-water. Dataset-C included few spectra ( $n = 68$ ) from open ocean environments, but due to the dominance of data from coastal waters, it is considered here to represent coastal environments. Only a limited

number of these datasets also included coincident measurements of IOPs and water quality parameters. As a result, IOPs and water constituents from the marine environment were not considered in this study. Nevertheless, clustering algorithms were applied to  $R_{rs}(\lambda)$  spectra from both inland and coastal systems in order to broaden the application of the classification scheme and study commonalities in spectral patterns across inland and coastal waters.

### Definition of reflectance

Clustering analysis was based on hyperspectral  $R_{rs}$ , with a minimum resolution of 1 nm and spectral range of 400–800 nm.  $R_{rs}(\lambda)$  (in  $\text{sr}^{-1}$ ) is defined here as the upwelling radiance emerging from the water column divided by the downwelling irradiance reaching the water surface. For those cases when in situ measurements were carried out just below surface,  $R_{rs}(0-)$  was converted to  $R_{rs}(0+)$  using the air–sea interface transfer coefficients of Eq. 1 (Lee et al. 1999):

$$R_{rs}(0+) = 0.52 R_{rs}(0-)/[1 - 1.7 R_{rs}(0-)] \quad (1)$$

**Table 3.** Description of the datasets from coastal (C) systems used in this work. Datasets were downloaded from SeaBaSS (SeaWiFS Bio-optical Archive and Storage System: <http://seabass.gsfc.nasa.gov>) (Werdell and Bailey 2002; Werdell et al. 2003) on 4<sup>th</sup> of March 2016.

Dataset	Principal institute responsible for sample and data collection and analysis/experiment	Marine system(s)
C-A	NOAA_CCMA/CALIFORNIA_2002	Coast of California
C-B	NOAA_CCMA/GLORIA	North Carolina
C-C	NOAA_CCMA/NC	North Carolina
C-D	NRL/ADRIATIC	Adriatic Sea
C-E	NRL/CHESAPEAKE	Chesapeake Bay
C-F	NRL/COJET	Mississippi Sound
C-G	NRL/HORN_ISLAND	Horn Island, Mississippi Sound
C-H	NRL/HYPOXIA	Mississippi Sound
C-I	NRL/LAUDERDALE	South-eastern coast, Florida
C-J	NRL/MONTEREY	Monterey Bay
C-K	NRL/SEED	Gulf of Mexico
C-L	NRL/SO_GASEX	Southern Ocean
C-M	UCSB/BBOP	Bermuda
C-N	USF/CARIACO	CARIACO station off the continental shelf of Venezuela (south-eastern Caribbean Sea)

NOAA\_CCMA, National Oceanic and Atmospheric Administration-Center for Coastal Monitoring and Assessment; NRL, United States Naval Research Laboratory; UCSB, University California Santa Barbara; USF, University South Florida.

## Methods

### Functional data analysis

Clustering was employed in order to identify statistically robust groups of spectra, which can be used to assist the definition of distinct OWTs found in aquatic systems. In the clustering process, the approach used for preprocessing of the data can play a crucial role in determining the influence of spectral features on the clusters obtained. In previous studies, classification of radiometric quantities have mainly considered unscaled data (e.g., Moore et al. 2001; Mélin et al. 2011); however, spectra scaling has been suggested by multiple authors in order to moderate the effect of variation in amplitude attributed to changes in the concentrations of optically active constituents (Mobley 1994; Schalles 2006; Ficek et al. 2012). In the analysis presented here, the  $R_{rs}(\lambda)$  were standardized prior to clustering in order to reduce the effect of the mean spectral reflectance on the separation of clusters. It is further thought that uncertainties in  $R_{rs}(\lambda)$  are more likely to have an effect on the amplitude of the spectra rather than their shape (Craig et al. 2006). The standardization used in this study entailed division by the area between each spectra and a zero baseline, calculated using numerical integration. This standardization approach was chosen because it preserves the shape of the  $R_{rs}$  across the different parts of the spectrum (Vantrepotte et al. 2012).

Subsequently, a functional data analysis approach was used to cluster the spectra. This approach approximates each  $R_{rs}(\lambda)$  using a smooth function which is estimated via a linear combination of  $B$ -spline basis functions (full details are

provided in Ramsay 2006). Rather than treating the reflectance values measured at each wavelength as single, correlated observations they are viewed as realizations of an unobservable continuous variable. Viewing the  $R_{rs}$  spectra in this way and clustering the smooth curves allows features within the groups (i.e., commonalities in shape and mean level) to be captured, which may be neglected if clustering was applied only to a single summary value (Tarpey and Kinateder 2003; Tarpey 2007). An attractive feature of the smoothing methods used within functional data analysis is that the underlying functions can be estimated such that excessive local variability which is not of interest is removed. In addition to reducing the noise in the data, by treating the basis coefficients which estimate each curve as the quantities to be clustered, we can justify the assumption of independence amongst variables. This is a fundamental assumption of clustering which is often overlooked (Fraleay and Raftery 1998) and can be violated in hyperspectral data due to the presence of strong autocorrelation between observations at neighboring wavelengths.

The number of basis functions used to estimate each smooth  $R_{rs}$  function controls the degree of flexibility, with more basis functions resulting in more flexibility. Adaptive smoothing can also be applied via the use of a non-constant basis to enable more flexibility in regions where there is greatest variability amongst each  $R_{rs}(\lambda)$ . In general, far fewer basis functions are used to represent each smooth function than there are original measurements, leading to a large reduction in dimensionality.

### **k-means clustering**

The *k*-means approach (MacQueen 1967; Lloyd 1982) was used to generate spectrally distinct water classes from the  $R_{rs}(\lambda)$  datasets. The *k*-means algorithm is a partitional approach (Jain 2010), well known for its efficiency in the classification of large datasets (Huang 1998). For functional data, each individual and the cluster centers can be defined in terms of the sets of basis coefficients which define the curves. Multiple starting points (50) were specified for the cluster centers in order to ensure the partition identified is not sensitive to the initial selection.

As with all clustering approaches, for *k*-means, there is a choice for the appropriate number of clusters. In this case, we used a gap statistic (Tibshirani et al. 2001), which selects the statistically optimal number of clusters by comparing the change in within-cluster dispersion between the observed data and a null reference distribution that is generated using the observed data. The reference distribution assumes there is no cluster structure in the data. Fuzzy *c*-means (FCM) clustering was also explored. While the estimated membership function in FCM may be attractive, the drawback of this approach is the required specification of an additional parameter, namely a weighting exponent which determines the degree of fuzziness in the clusters. *k*-means is a special case of the FCM with the weighting exponent fixed to 1, resulting in all data points being assigned to one and only one cluster.

As a measure of proximity to cluster mean, the L2 norm distance was calculated between each individual and each cluster mean. These curve-mean distances were scaled between 0 and 1 and were used to quantify how close the curve was to each cluster mean.

In this study, we use the term:

- “Cluster” to refer to the end-member resulting from cluster analysis, i.e., set of distinct spectra as these were separated by *k*-means algorithm,
- “Group” to refer to spectra with high within similarity of the second derivatives of cluster means based on the L2 norm distances and
- “Type” for the representative spectrum (here, the mean spectrum is used) and in-water optically active compounds for a cluster.

## **Results**

### **Spectral variability, rescaling and adaptive smoothing**

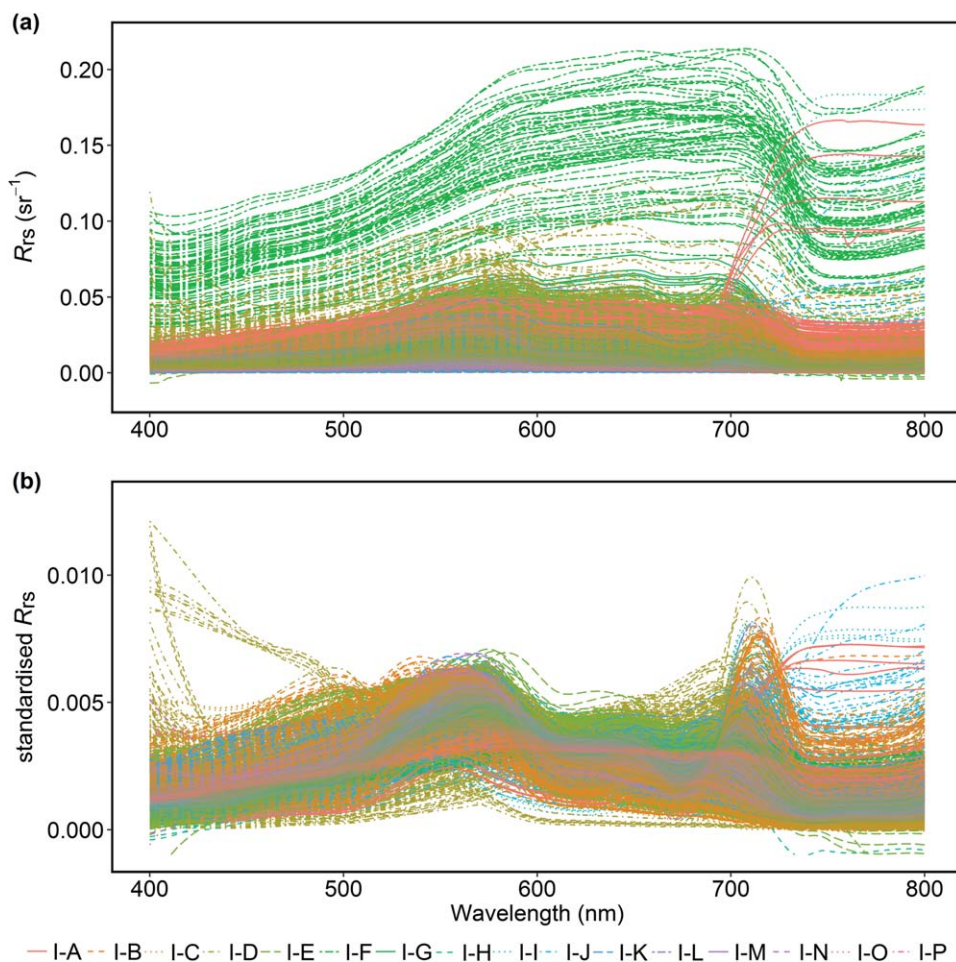
Figures 1a, 2a show the in situ  $R_{rs}(\lambda)$  spectra from Dataset-I and Dataset-C on their original scale (dataset details are provided in Tables 2, 3). Spectra from inland waters generally had higher mean reflectance than those from coastal waters but both sets demonstrated considerable variation in magnitude, even when they exhibited similar shapes. The reflectance peak in the green part of the

spectra (500–600 nm) ranged from 0.0003  $\text{sr}^{-1}$  to 0.2031  $\text{sr}^{-1}$  in Dataset-I and from 0.001  $\text{sr}^{-1}$  to 0.051  $\text{sr}^{-1}$  in Dataset-C. In the near-infrared (NIR) spectral region (680–720 nm), maximum values of the  $R_{rs}$  peak were 0.2137  $\text{sr}^{-1}$  and 0.0359  $\text{sr}^{-1}$ , respectively, for the inland and coastal data. Spectral features appearing around  $R_{rs}$  (760) could be indicative of an abnormal signal, pertaining to flaws in the measurement and processing protocols (e.g., suboptimal sensor calibration, incompatible viewing angles, or lack of synchronicity in the measurement). As shown in Figs. 1b, 2b, standardized in situ  $R_{rs}(\lambda)$  spectra are accompanied by lower variability in the overall magnitude of reflectance. Coefficient of variation varied from 98% to 236% in Dataset-I with a local maximum at 675 nm and an overall minimum at 550 nm. Similarly, for Dataset-C, this varied from 98% at 550 nm to 236% at 675 nm.

The resulting design matrix for the *B*-spline basis used in this study was based on 25 cubic basis functions (Fig. 3). This number provided the best achievable fit to the data and captured all key features of the standardized  $R_{rs}(\lambda)$ . Figure 3 also illustrates the unequally spaced basis of *B*-spline functions. The *B*-spline representation used approximately one knot every 30 nm between 400 nm and 500 nm, one knot every 15 nm between 500 nm and 750 nm, with the 750–800 nm part of the spectrum being covered by a single interval. Although the large basis function used between 750 nm and 800 nm ignores a large part of the variability in this range, it helps resolve issues of instrument noise and poor instrument calibrations that often affect this part of the spectrum (Fargion and Mueller 2002). The sparsity of the basis here will prevent features which are not of interest, or are subject to a high degree of uncertainty having a disproportional influence on the definition of clusters. Unusual spectra revealed by functional boxplots (not shown) were considered to correspond to “extreme” cases (13 spectra) or erroneous measurements (27 spectra) where successive peaks were shown. The former cases referred to very clear waters (cluster I13), while the latter cases were removed from the dataset.

### **Clustering of reflectance spectra**

The *k*-means algorithm was applied to the basis coefficients which defined the smooth  $R_{rs}(\lambda)$  spectra for three datasets: inland (Dataset-I), coastal (Dataset-C), and all waters (Dataset-N). For *k*-means clustering, the statistically optimal number of clusters determined by using the gap statistic (with 500 reference distributions) was 12 for Dataset-I and 9 for Dataset-C. An additional group of curves (I13) that were identified as being unusual by the functional boxplots was added to the 12 inland clusters identified using the *k*-means approach. All pairs of cluster means were found to be significantly different using a permutation *t*-test (Ramsay 2007), suggesting unique structural groups.



**Fig. 1.** In situ hyperspectral remote sensing reflectance ( $R_{rs}$ ) spectra of datasets (Table 1) collected at inland aquatic systems (a) on their original scale ( $\text{sr}^{-1}$ ) (b) standardized.

The optimal number of clusters for Dataset-N was identified using an approach based on analysis of similarity between a fixed number of clusters due to the large number of spectra positioned on the boundaries between several clusters. The number of clusters was initially set to 21 based on the assumption that this would represent the upper bound (as the sum of clusters resolved in dataset-I [ $n=12$ ] and Dataset-C [ $n=9$ ] separately). In order to identify clusters that could subsequently be merged, the difference between clusters was explored in terms of the L2 norm distance between the mean curves for each cluster.

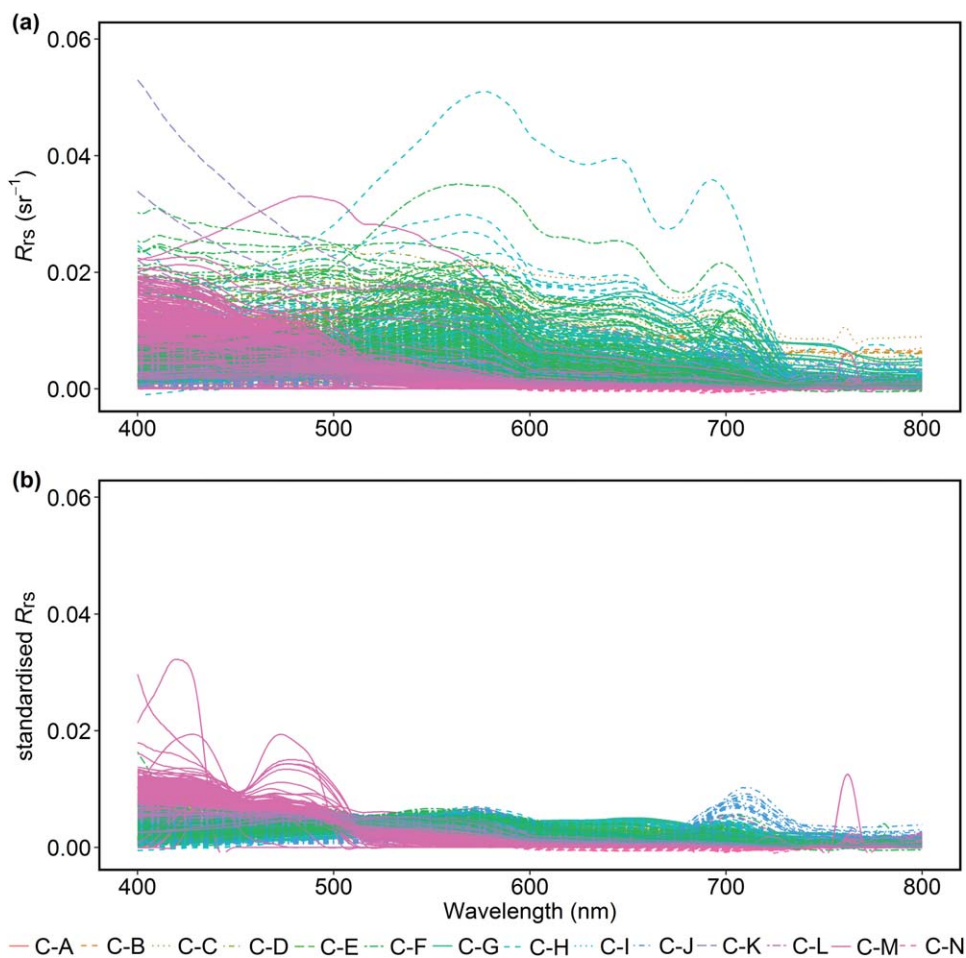
Differences between the shapes of the cluster mean curves, following a second derivative transformation, were also considered. For Dataset-I, the means of the 13 distinct standardized and non-standardized  $R_{rs}(\lambda)$  spectral clusters, as identified by the  $k$ -means algorithm, are presented in Fig. 4. The largest numbers of spectra were assigned to clusters I2 (15.3%) and I6 (14.3%). Clusters I1 and I13 collectively contained 1.1% of the data. We noted that clusters were not

strongly driven by waterbody or season but were distributed across space and time.

Figure 5 presents the mean in situ  $R_{rs}(\lambda)$  spectra before and after standardization, for the nine groups obtained by applying the  $k$ -means algorithm to the functional data from Dataset-C. The spectra were nearly equally partitioned (12.1–15.3%) between clusters C1, C3, C4, C6, C7, and C8. Conversely 22 (2.2%), 46 (4.9%), and 83 (8.4%) spectra were grouped in clusters C2, C5, and C9. Figure 6 shows  $R_{rs}$  and standard deviation for each cluster identified in inland waters. The  $k$ -means classification of all data combined (Dataset-N) resulted in the 21 sets of reflectance spectra shown in Fig. 7.

### Bio-optical properties of inland water clusters

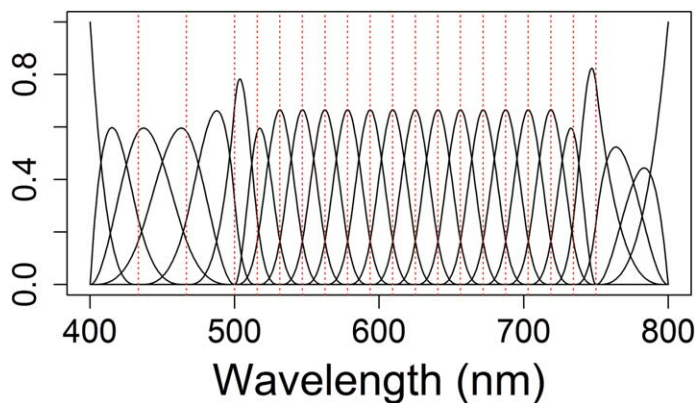
Figures 8, 9 summarize water constituents and the optical properties corresponding to each cluster of Dataset-I (I-clusters). Several parameters measured coincident to the reflectance measurements are considered here. The water constituent concentrations that were most commonly



**Fig. 2.** In situ hyperspectral remote sensing reflectance spectra ( $R_{rs}$ ) of datasets (Table 3) collected at coastal systems **(a)** on their original scale **(b)** standardized.

measured in parallel with the radiometric measurements were chlorophyll *a* (Chl *a*) ( $n = 2835$ ), total suspended matter (TSM) ( $n = 1836$ ), and absorption of colored dissolved organic matter (CDOM) at 442 nm ( $a_{CDOM(442)}$ ) ( $n = 1720$ ), while 622  $R_{rs}(\lambda)$  measurements were also accompanied by absorption coefficients of phytoplankton pigments  $a_{ph}(\lambda)$  and non-algal particles (NAP)  $a_{NAP}$ . Despite the high variability of these in-water parameters and the often complex relationships between apparent optical properties and the particulate and dissolved material found in inland waters, there are some notable differences among the groups of in situ water properties for each partition of  $R_{rs}(\lambda)$ . As expected, the optical properties and concentrations of optically active substances underpin the clustering of  $R_{rs}(\lambda)$ .

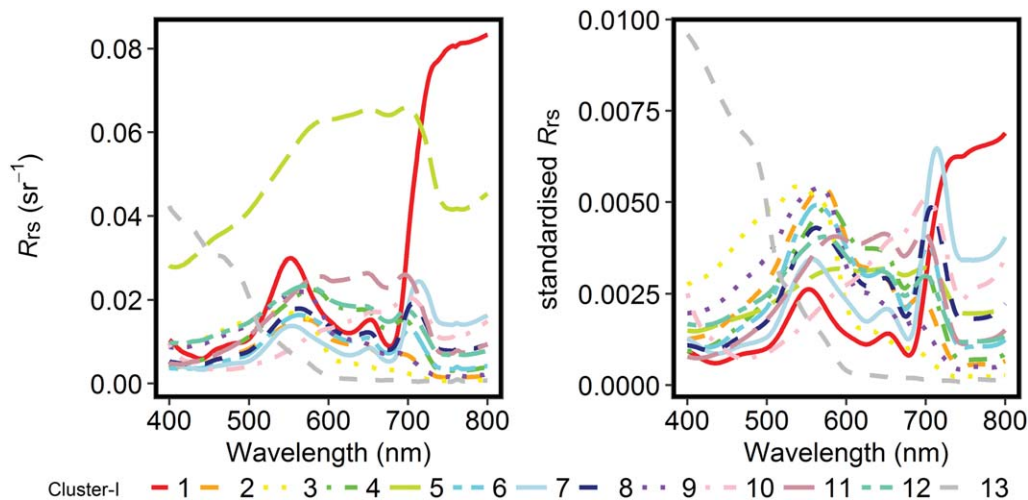
The 13 I-clusters exhibited marked differences in terms of their water constituent concentrations and IOPs. For example, clusters I1, I7, and I8 exhibited very high concentrations (mean values well above  $100 \text{ mg m}^{-3}$ ) of Chl *a* and the accessory pigment phycocyanin (PC) (mean values greater than  $200 \text{ mg m}^{-3}$ ). In contrast, Chl *a* was remarkably low in



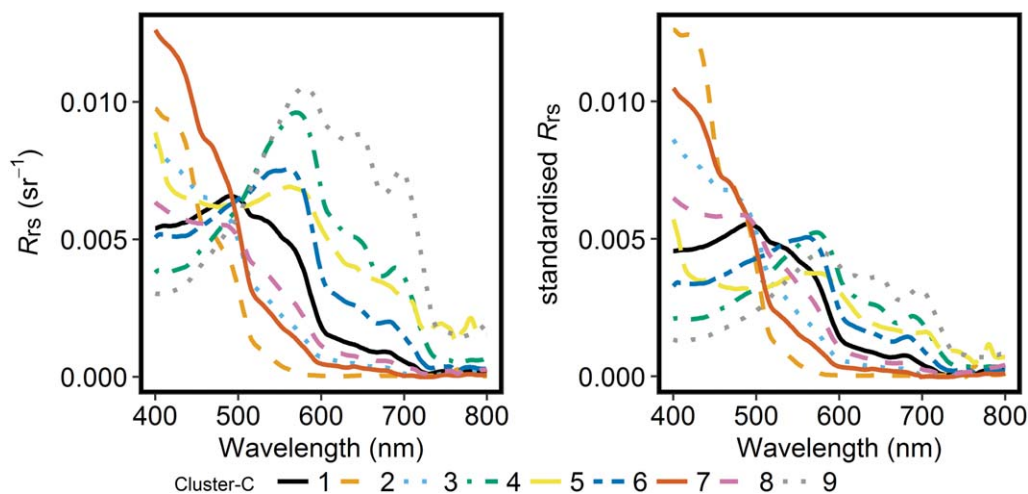
**Fig. 3.** Unequally spaced basis of B-spline functions (25) used in the study to fit the smooth curve.

clusters I3 ( $1.60 \pm 1.02 \text{ mg m}^{-3}$ ,  $n = 214$ ) and I13 ( $0.27 \pm 0.57 \text{ mg m}^{-3}$ ,  $n = 8$ ). These clusters also showed higher values of Secchi disk depth (I3:  $6.17 \pm 2.52 \text{ m}$ ,  $n = 173$ ; I13:





**Fig. 4.** Mean remote sensing reflectance spectra ( $R_{rs}$ ) for each distinct cluster obtained in inland waters as were identified by  $k$ -means algorithm applied on the functional data. Left: on their original scale. Right: standardized.

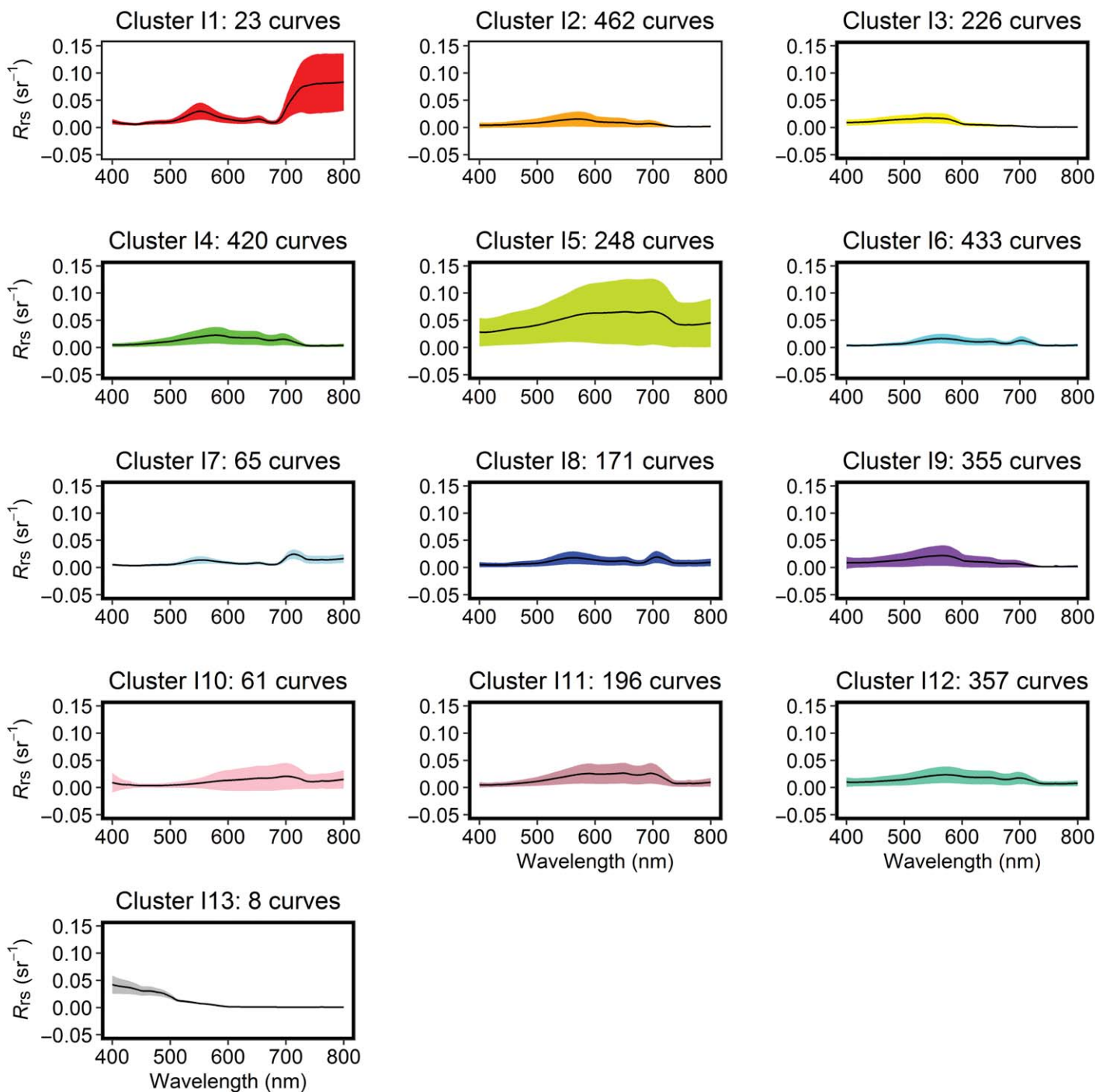


**Fig. 5.** Mean remote sensing reflectance spectra ( $R_{rs}$ ) for each distinct cluster obtained in marine waters as were identified by  $k$ -means algorithm applied on the functional data. Left: on their original scale. Right: standardized.

$18.45 \pm 4.17$  m,  $n = 2$ ) and the lowest mean concentration of TSM (I3:  $1.57 \pm 1.64$  mg L<sup>-1</sup>,  $n = 87$ ; I13:  $1.00 \pm 0.88$  mg L<sup>-1</sup>,  $n = 8$ ). We noted that the highest mean inorganic suspended matter (ISM) concentration ( $94.41 \pm 64.45$  mg L<sup>-1</sup>,  $n = 200$ ) was found in the samples grouped in cluster I5. In addition, cluster I5 was characterized by the highest  $a_{NAP}(442)$  mean ( $5.76 \pm 2.90$  m<sup>-1</sup>,  $n = 112$ ), while clusters I10 and I1 had higher  $a_{CDOM}(442)$  ( $9.00 \pm 7.35$  m<sup>-1</sup>,  $n = 50$ ) and  $a_{ph}(442)$  ( $106.49 \pm 10.28$  m<sup>-1</sup>,  $n = 11$ ), respectively. Clusters with the highest  $a_{ph}(442)$  and  $a_{NAP}(442)$  values were principally found among the groups with their lowest mass-specific absorption coefficients ( $a_{ph}(442)$ :[Chl  $a$ ] or  $a^*_{ph}(442)$  and  $a_{NAP}(442)$ :[TSM] or  $a^*_{NAP}(442)$ ) and, corresponding to a higher degree of “pigment packaging” (e.g., Bricaud et al. 1995) or cell shading and a more minerogenic NAP. In cases

where clusters had similar mean concentrations of one or more biogeochemical parameters, we generally observed differences in other variables which facilitated their distinctive characterization. For example, cluster I4 showed comparable to I5 Chl  $a$  but contrasting ISM concentrations.

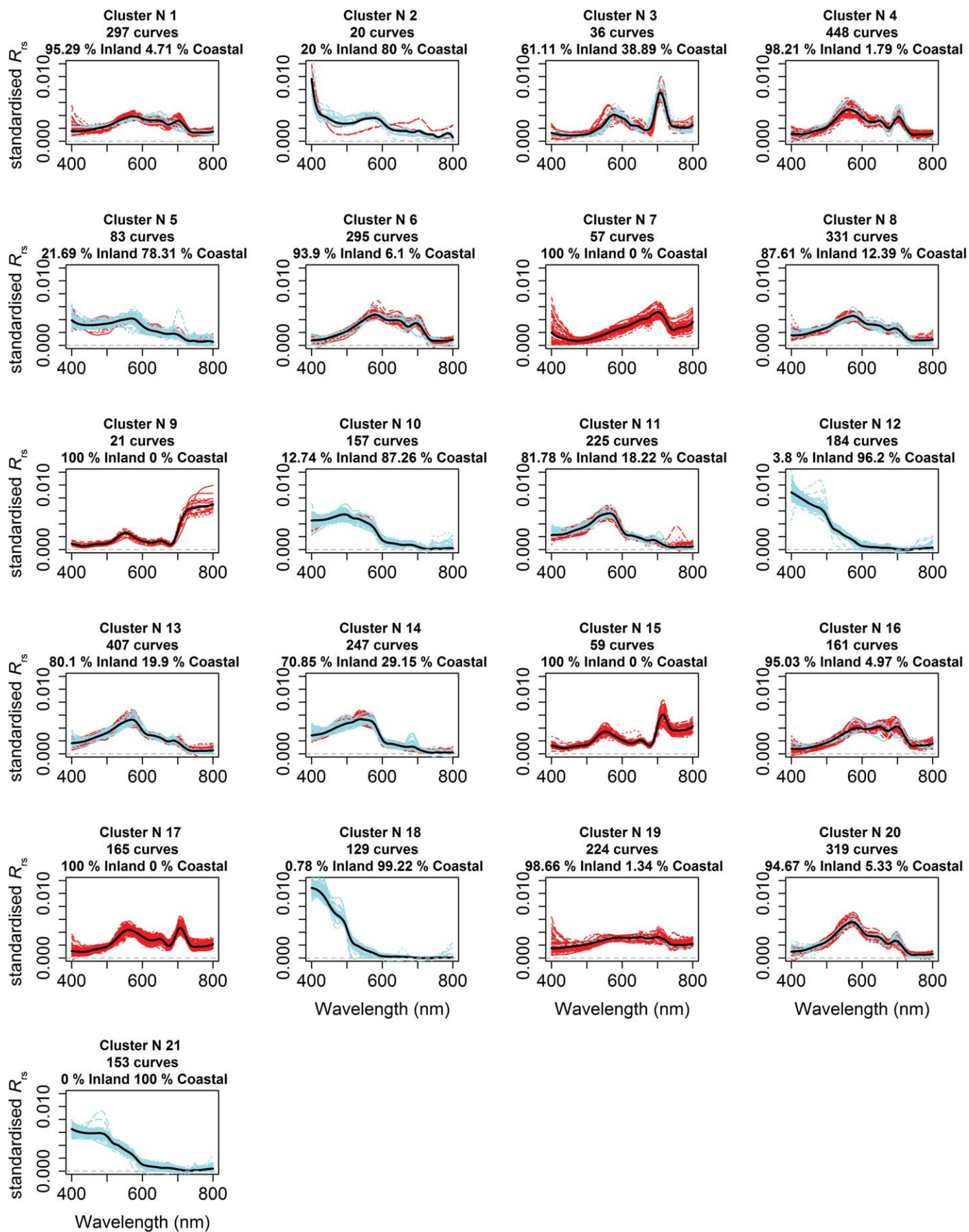
Figure 10a–c illustrate absorption spectra of CDOM and specific absorption of phytoplankton and NAP for each cluster identified by the classification analysis. In the analysis, we considered a spectral range from 400 nm to 700 nm, which corresponds to the range available for most data points. Both  $a_{CDOM}$  and its spectral slope ( $S_{CDOM}$ ) varied between the different clusters. Cluster I3 showed the lowest  $S_{CDOM}$  ( $0.0114 \pm 0.0068$  nm<sup>-1</sup>,  $n = 6$ ). Higher  $S_{CDOM}$  values were observed in clusters I9 ( $0.0173 \pm 0.0050$  nm<sup>-1</sup>,  $n = 39$ ), I2 ( $0.0161 \pm 0.0037$  nm<sup>-1</sup>,  $n = 57$ ), and I11



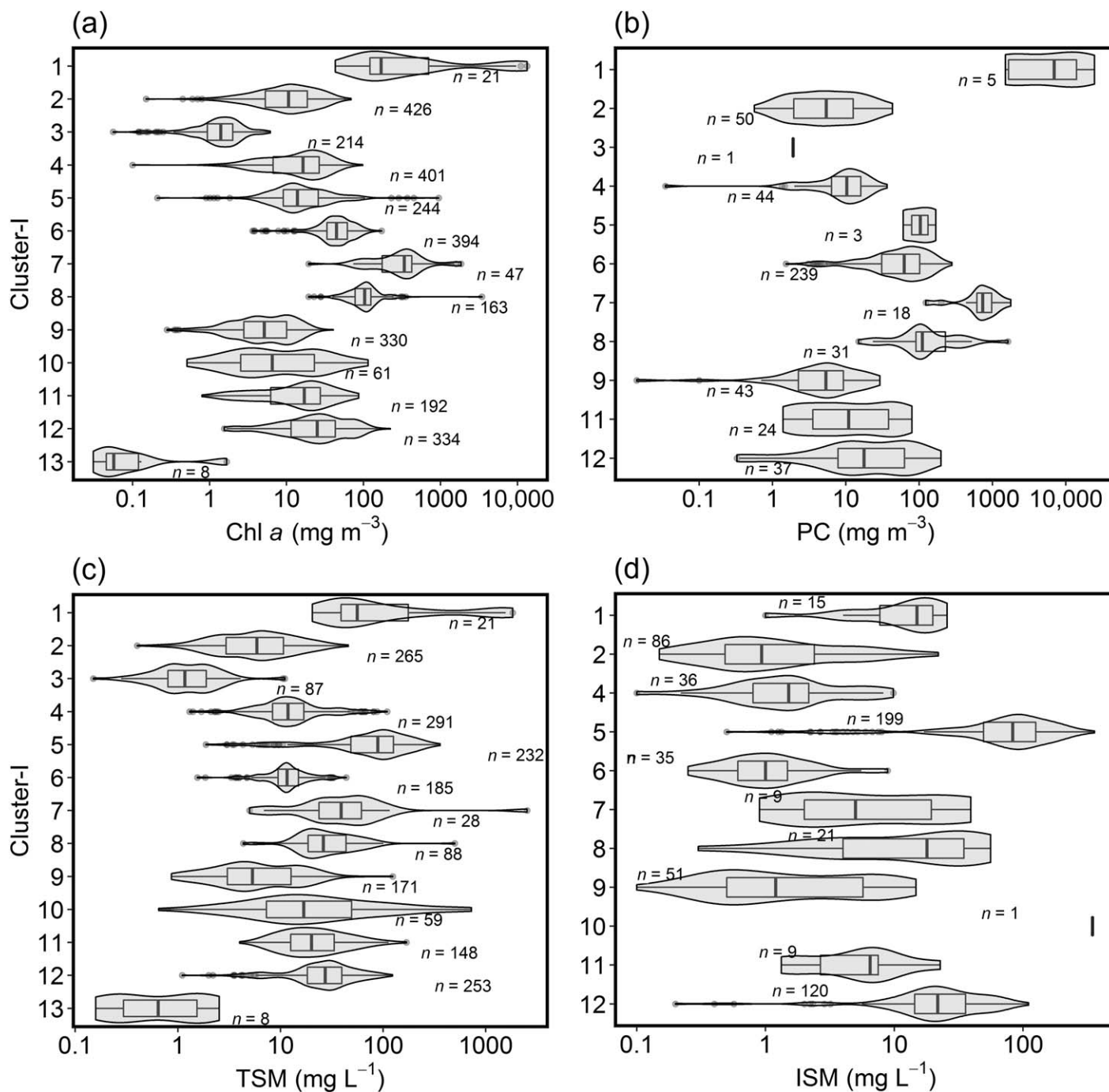
**Fig. 6.** Mean (solid black line) remote sensing reflectance ( $R_{rs}$ ) and standard deviation (shaded area) obtained in inland waters as were identified by  $k$ -means algorithm.

( $0.0150 \pm 0.0017 \text{ nm}^{-1}$ ,  $n = 30$ ).  $S_{CDOM}$  showed relatively low variability within the remaining clusters with mean values in these ranging from  $0.0139 \text{ nm}^{-1}$  to  $0.0147 \text{ nm}^{-1}$ . Figure 10b shows high variability of mean  $a_{ph}^*(\lambda)$  in both magnitude and spectral shape among the clusters where both  $a_{ph}(\lambda)$  and Chl  $a$  were measured. The differences in spectral amplitude were

mainly observed in the blue and red regions of the spectra; cluster I3 exhibited the lowest blue to red peak ratio while that ratio was higher in clusters I5, I11, and I12. These clusters were also characterized by the lowest mean value ( $0.0080 \pm 0.0017 \text{ nm}^{-1}$ ,  $n = 45$ ) of slope for NAP absorption,  $S_{NAP}$ .



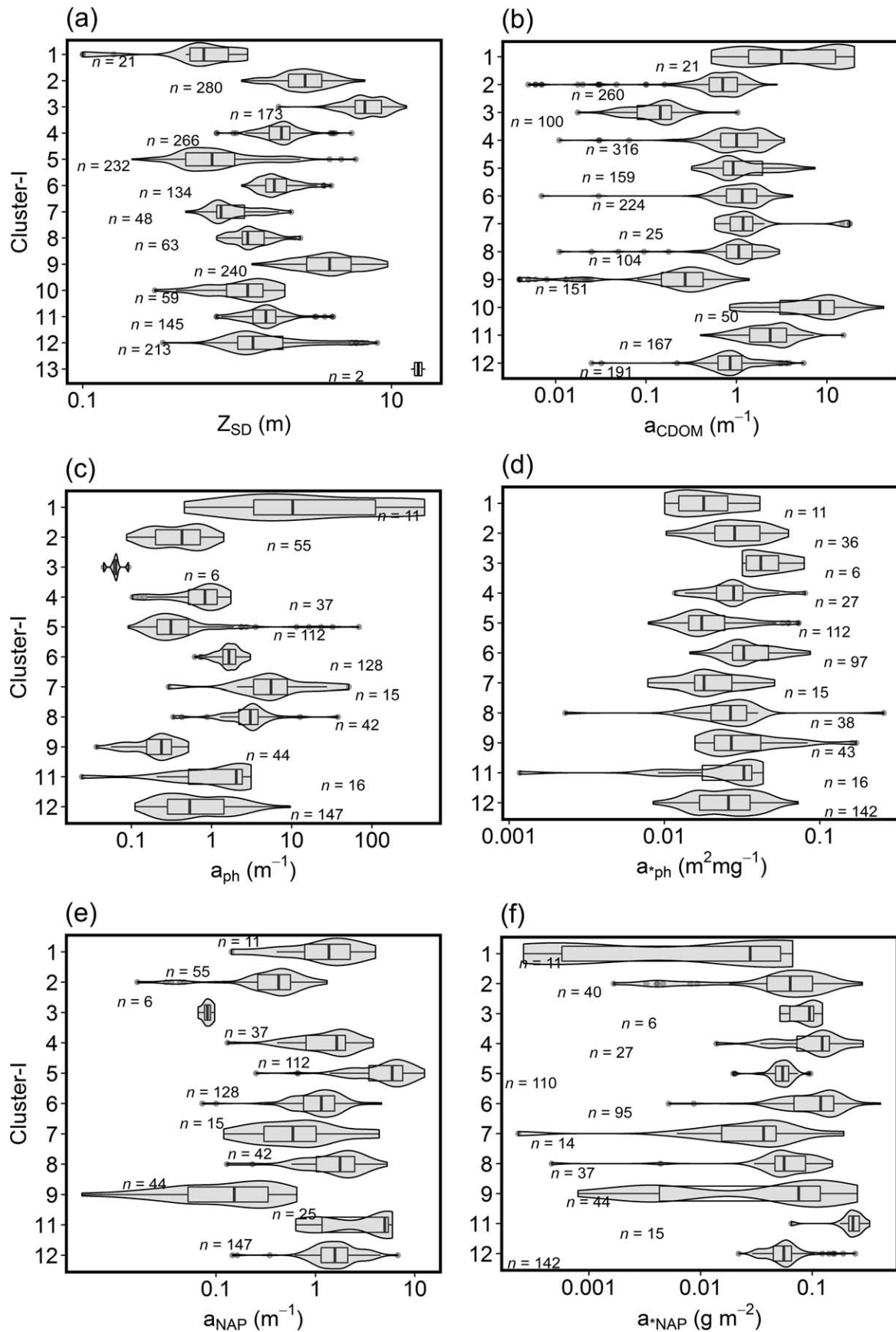
**Fig. 7.** Mean remote sensing reflectance spectra ( $R_{rs}$ ) colored by dataset for each distinct cluster obtained in natural waters identified by  $k$ -means algorithm applied on the functional data. Spectra are shown on standardized reflectance scale. Percentage of curves of each type in each group is shown in plot titles.



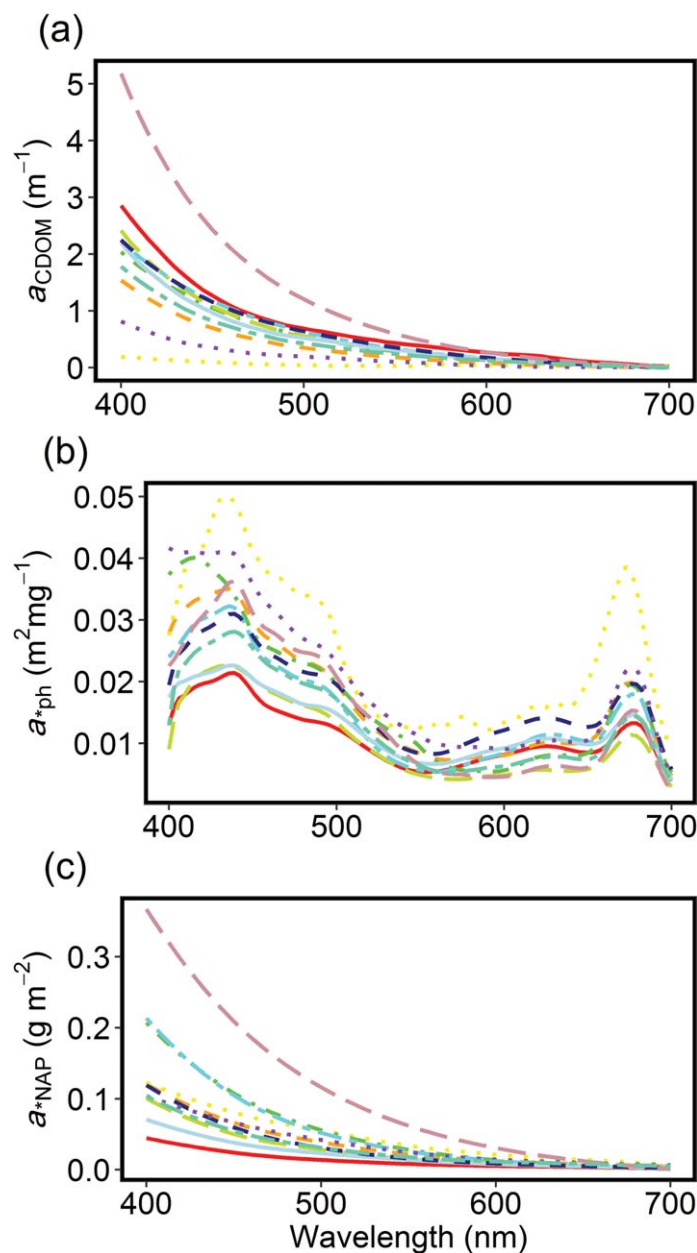
**Fig. 8.** Boxplots with probability density of (a) Chl *a*, (b) PC, (c) TSM, and (d) ISM for each optical cluster in Dataset-I. The sample median is indicated by a vertical line within the box while dots represent data beyond the bounds of the error bars and *n* the number of observation.

Figure 11 summarizes the relative contribution of optically active substances  $a_{CDOM}$ ,  $a_{ph}$ , and  $a_{NAP}$  to total absorption (minus pure water absorption) at 442 nm for each optical cluster. Phytoplankton absorption was consistently the dominant absorption component of samples grouped in clusters I1 and I7 and regularly the weakest component in I4 and I5. Spectra belonging to cluster I5

were predominantly characterized by strong relative influence of  $a_{NAP}$ .  $a_{CDOM}$  was the dominant light absorbing coefficient at 442 nm for clusters I2 and I3. Data points grouped in clusters I8 and I6 were mainly found toward the upper half of the ternary plot, whereas samples collected from clusters I11 and I12 mostly appeared at the lower half of the plot.



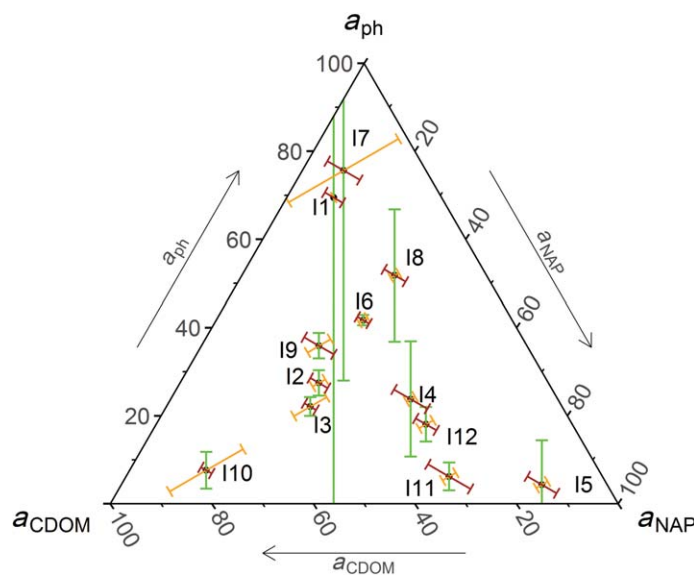
**Fig. 9.** Boxplots with probability density of (a) Secchi disk depth,  $Z_{SD}$  (b) absorption coefficient of CDOM,  $a_{CDOM}(442)$  (c) absorption coefficient of phytoplankton,  $a_{ph}(442)$  (d) absorption coefficient of phytoplankton normalized to Chl *a* concentration,  $a^*_{ph}(442)$  (e) absorption coefficient of “non-algal” particles,  $a_{NAP}(442)$  and (f) absorption coefficient of NAP normalized to TSM concentration,  $a^*_{NAP}(442)$ , for each optical cluster in Dataset-I. The sample median is indicated by a vertical line within the box while dots represent data beyond the bounds of the error bars and *n* the number of observation.



**Fig. 10.** Mean spectra of (a) absorption coefficient of CDOM,  $a_{CDOM}(\lambda)$  (b) absorption coefficient of phytoplankton normalized to Chl *a* concentration,  $a_{ph}^*(\lambda)$  and (c) absorption coefficient of NAP normalized to TSM concentration,  $a_{NAP}^*(\lambda)$ , of each optical cluster in Dataset-I. Line colors and types as shown in Fig. 4.

**Relationships among optical clusters in inland and coastal waters**

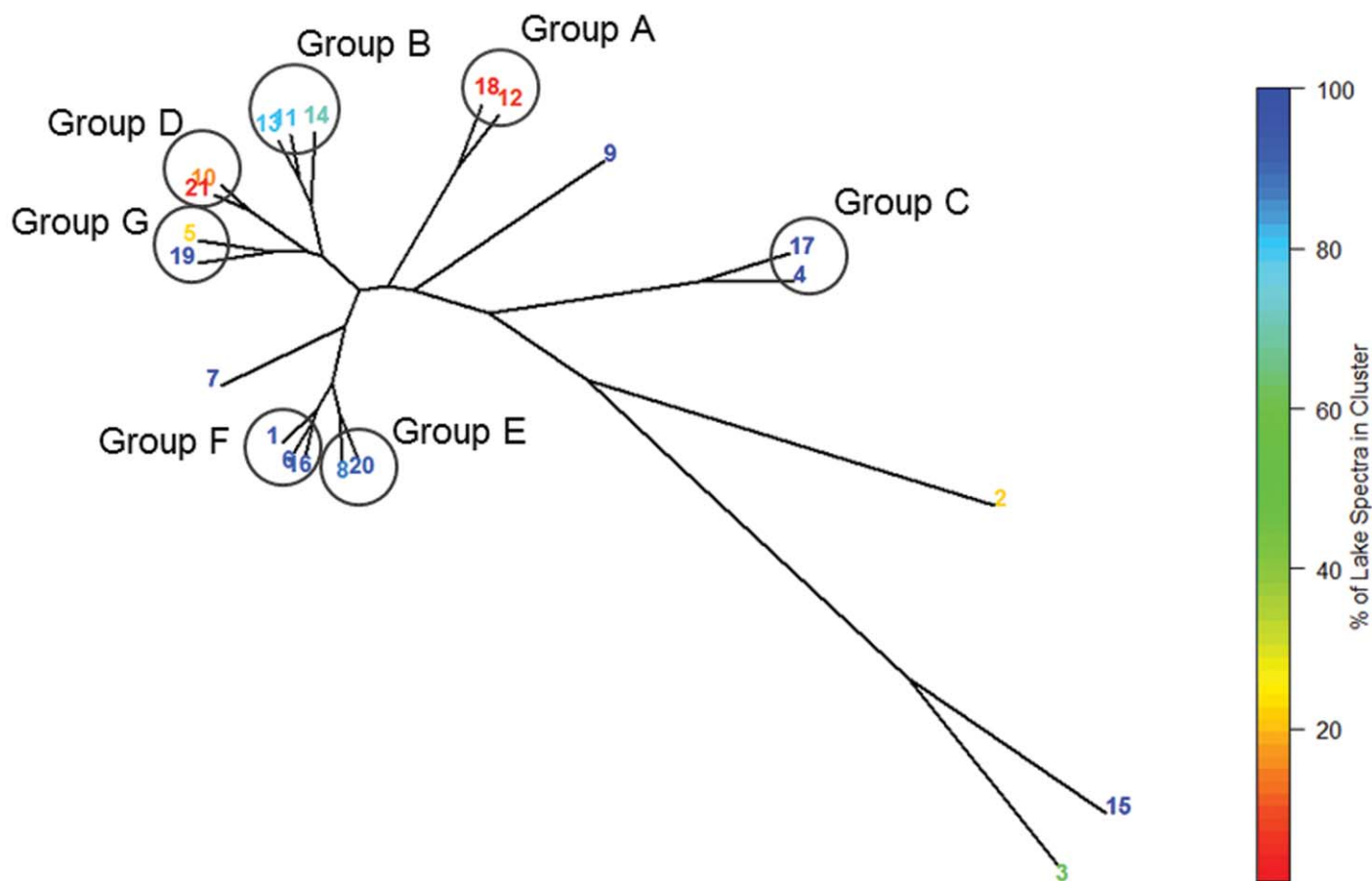
$R_{rs}(\lambda)$  spectra from coastal systems were predominant in clusters N2, N5, N10, N12, N18, and N20 (all with relative contributions above 79.3%), while the remaining clusters were largely composed by spectra from inland waters. Sixteen clusters contained spectra derived from both inland and coastal systems (Fig. 7). A phylogenetic tree was constructed



**Fig. 11.** Ternary plot representing the mean relative contribution of optically active substances (absorption coefficient of CDOM [ $a_{CDOM}$ ], absorption coefficient of phytoplankton [ $a_{ph}$ ] and absorption coefficient of “non-algal” particles [ $a_{NAP}$ ]) to total absorption at 442 nm for each optical cluster in dataset-I. Error bars indicate standard deviation of the mean.

to explore relationships among the 21 cluster means (Fig. 12). This tree represents the similarity of the second derivatives of cluster means based on the L2 norm distances. Clusters N2 ( $n = 20$ ), N3 ( $n = 36$ ), N7 ( $n = 57$ ), N9 ( $n = 21$ ), and N15 ( $n = 59$ ) can be seen to be most distinct from other clusters with N3, N15, and N2 showing most difference from all other clusters in terms of their second derivatives. Two of these clusters (N2 and N9) also contained the lowest number of  $R_{rs}(\lambda)$  spectra indicating they may be composed of uncommon spectral properties. Cluster N2 displayed spectral features in the blue and red region of the spectrum that suggests residual glint contribution in the measured signal. This group of measurements was therefore excluded from further analysis.

Using the phylogenetic tree, seven major Groups with high within group similarity were identified (Group A: N18, N12; B: N11, N13, N14; C: N4, N17; D: N10, N21; E: N8, N20; F: N1, N6, N16; G: N5, N19). Group A ( $n = 310$ ) mainly included reflectance spectra from Dataset-C with relatively high  $R_{rs}(\lambda)$  in the blue. Groups E-F both had three  $R_{rs}(\lambda)$  peaks between 500 nm and 750 nm and were mainly found in Dataset-I. The reflectance peak around 700 nm in Group F appeared associated with particulate scattering and occurred at longer wavelengths than in Group E, where cluster N5 suggests the presence of Chl *a* fluorescence at around 685 nm and cluster N19 suggests highly turbid water with a minor contribution of phytoplankton absorption. Groups B, D, and G were assembled closely (Fig. 13). These three Groups contained data from both Dataset-I and Dataset-C. Relatively



**Fig. 12.** Phylogenetic tree representing the similarity of the second derivatives of all cluster means based on L2 norm distances. The colors of the labels represent the percentage of spectra originating from inland waters in each group.

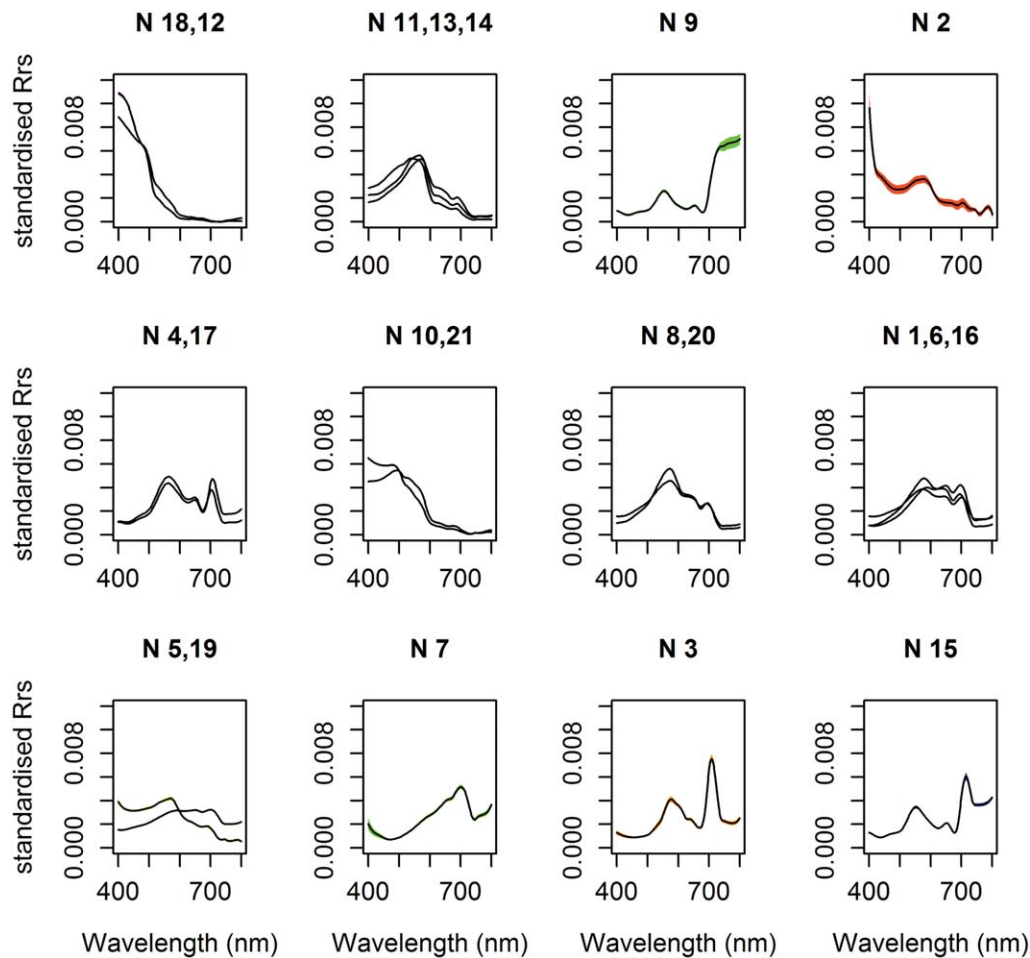
clear waters (no prominent peak near 700 nm) and a strong influence of  $a_{\text{CDOM}}$  in the blue characterize the clusters in Group B. Clusters N10 and N21 (Group D) shared a sharp  $R_{\text{rs}}(\lambda)$  decrease near 600 nm and high blue-to-green  $R_{\text{rs}}(\lambda)$  ratio suggesting clear waters, but with a lower blue-to-green ratio compared to the clusters of Group A. Last, Clusters N5 and N19 showed high similarities of the second derivatives; neither cluster shows clearly defined features beyond the attenuation of light by  $a_{\text{CDOM}}$  in the blue and absorption by water in the red to NIR domain. Interestingly, N5 contained primarily data collected in coastal systems (79.27%) whereas N19 was composed of spectra found in inland waters (98.21%).

## Discussion

### Methodological considerations

$R_{\text{rs}}(\lambda)$  holds valuable information on the concentration and composition of in-water constituents (Gordon et al. 1988; Gordon and Franz 2008) and is now readily available from multispectral ocean color satellite sensors. We present a novel approach for classification of in situ hyperspectral  $R_{\text{rs}}(\lambda)$  to help optimize the interpretation of proximal or

remotely sensed  $R_{\text{rs}}(\lambda)$  in terms of biogeochemically-relevant quantities. While  $k$ -means is a classical statistical method, its application in a functional setting is not routine, particularly when the irregularly spaced  $B$ -spline basis coefficients have been selected so the clusters are based on the areas of the spectra, which are of most interest. The robustness of this approach is potentially dependent on the smooth functions which are used to estimate the underlying smooth processes from which the observed data have arisen. The 25 cubic basis functions, with different resolution along the wavelength, employed here provided an excellent fit to the data (Fig. 3), capturing all key features of the spectra while removing local variability. This approach also proves to be an efficient way to reduce dimensionality and noise of the spectra while preserving distinctive features. FCM clustering was also explored and the adjusted Rand index (ARI) (Hubert and Arabie 1985) used for the comparison with the  $k$ -means approach suggested strong agreement (ARI greater than 0.76) between the two clustering methods. While both shape and amplitude of  $R_{\text{rs}}(\lambda)$  contain information about optically active constituents, we standardized the in situ spectra (Vantrepotte et al. 2012) in order to reduce the influence of



**Fig. 13.** Cluster merging based on the similarity between the second derivative of cluster means in terms of L2 norm distance.

spectral amplitude on clustering. This is considered a suitable approach when considering  $R_{rs}(\lambda)$  from such diverse origins. Focussing on the shape rather than amplitude of the spectra implies primary sensitivity to spectral variation in absorption coefficients in the clustering (Loisel and Morel 2001). However, since the absorption by water itself also displays a spectral dependence, the attenuation depth of the recorded signal, and therefore the light path and intensity of light scattering which primarily affects  $R_{rs}(\lambda)$  amplitude, does bear influence on the clustering results. General observations of the obtained clusters (Figs. 4, 5) show amplitude variability in dominant parts of spectra but with distinctive spectral features. However, data standardization prior to clustering has also been suggested to reduce spurious effects of unequal variances and clustering of non-standardized  $R_{rs}(\lambda)$  is still most common in the literature (Moore et al. 2014; Shen et al. 2015). Nevertheless, when Mahalanobis distance is used in the cluster analysis, data preprocessing is considered redundant unless rounding errors in the covariance matrix have not been restrained (Besset 2001; Eyob 2009).

### Optical water typology

While studies of optical water typology (e.g., Jerlov 1977; Morel and Prieur 1977; Moore et al. 2009) provided useful insights on the distinctive optical types found in aquatic systems, they were challenged by the representativeness of optical conditions and/or limited understanding of factors driving the observed variability among the different optical clusters. Our ability to identify representative clusters from 4035  $R_{rs}(\lambda)$  spectra collected in 250 inland water and several coastal systems benefits from operating over a wide range of in situ biogeochemical parameters. In this study, we were able to resolve clusters of  $R_{rs}(\lambda)$  spectra representing statistically distinct optical clusters found in inland or coastal waters and, in some cases, in both environments. The number of clusters was identified following a purely data driven approach where the number chosen was selected by the gap statistic as statistically optimal. The OWTs suggested here are considered as typical OWT found in the datasets and emerge as representations of optical conditions that are a glimpse of a natural continuum system in aquatic systems.



**Table 4.** Dominant characteristics of OWTs in inland waters.

OWT	Dominant characteristics
OWT1	Hypereutrophic waters with scum of cyanobacterial bloom and vegetation-like $R_{rs}$
OWT2	Common case waters with diverse reflectance shape and marginal dominance of pigments and CDOM over inorganic suspended particles
OWT3	Clear waters
OWT4	Turbid waters with high organic content
OWT5	Sediment-laden waters
OWT6	Balanced effects of optically active constituents at shorter wavelength
OWT7	Highly productive waters with high cyanobacteria abundance and elevated reflectance at red/near-infrared spectral region
OWT8	Productive waters with cyanobacteria presence and with $R_{rs}$ peak close to 700 nm
OWT9	Optically neighboring to OWT2 waters but with higher $R_{rs}$ at shorter wavelengths
OWT10	CDOM-rich waters
OWT11	Waters high in CDOM with cyanobacteria presence and high absorption efficiency by NAP
OWT12	Turbid, moderately productive waters with cyanobacteria presence
OWT13	Very clear blue waters

Moreover, as an extension to previous research, we provide a detailed physical interpretation of the derived clusters facilitated by extensive data on the IOPs and concentrations of color-forming biogeochemical constituents. IOP data allowed a more detailed characterization of the optical clusters and provided reference subsets. However, we recognize that different instruments, methods, and protocols have been utilized for the measurement of optical and biogeochemical parameters. Consequently, some of the variability observed in the  $R_{rs}(\lambda)$  spectra will have arisen from different instrumentations and data collection and processing methodologies. In practice, biogeochemical data covering such a wide range of ecosystem scales are scarce, and measurement protocols have often been locally refined, modified, and optimized. It may be expected that the continued contribution of in situ observations to community databases such as LIMNADES and SeaBASS will lead to a gradual convergence of methodologies and a reduction in the associated uncertainties on in situ radiometric measurements.

#### Inland waters OWTs

The classification of inland waters  $R_{rs}(\lambda)$  revealed 13 different optical clusters (Figs. 4, 6). The categorization of these clusters to OWTs was subsequently based on in-water information on absorption coefficients (i.e., Figs. 10, 11) and biogeochemical properties (i.e., Figs. 8, 9). Table 4 provides a brief description of each OWT. PC and ratio of PC to Chl  $a$  (Simis et al. 2005) indicated the presence and relative abundance of cyanobacteria in an OWT. This is of particular interest for the monitoring of cyanobacteria blooms.

OWT1 represents waters with extremely high concentrations of Chl  $a$ , PC, and high  $R_{rs}(\lambda)$  in the red to near-infrared region of the spectrum indicating high abundance of cyanobacteria near or at the water surface. High PC concentrations ( $6953.3 \pm 9778.9 \text{ mg m}^{-3}$ ) and ratios of PC to Chl  $a$  above 1 are also indicative of high abundance of cyanobacteria in

this OWT. It is not uncommon to find extremely high concentrations of pigments and vegetation-like  $R_{rs}(\lambda)$  spectra due to shallow light penetration (and therefore limited water absorption) in inland and coastal waters (Kutser et al. 2012). For all spectra pooled into OWT1, we observed an  $R_{rs}(\lambda)$  peak close to 655 nm. This has been suggested to be a combined effect of high Chl  $a$  and PC absorption either side of the peak (Kudela et al. 2015) and could also be associated with sun-induced autofluorescence of phycobilipigments.  $a_{\text{NAP}}$  while high, is largely masked by phytoplankton absorption, suggesting dominance of living material over detritus and mineral particles, and masking of  $a_{\text{CDOM}}$  influences on the spectrum due to a short light path, similar to the masking of the absorption by water.

OWT2 was the most common case in our dataset, showing diversity in reflectance shape with peaks at regions (565 nm, 645 nm, and 695 nm) where particles scatter light (Gitelson et al. 2000; Doxaran et al. 2009) and where peaks where bounded by pigment absorption maxima (Kirk 1994). In terms of the absorption budget at the blue wavelengths, OWT2 is located close to the center of the ternary plot which indicates that  $a_{\text{CDOM}}$  and  $a_{\text{NAP}}$  over  $a_{\text{ph}}$  were contributing almost equally to non-water absorption, while the high  $S_{\text{CDOM}}$  (400–700) suggests the dissolved fraction was dominated by terrestrial humic acids (Yacobi et al. 2003; Zhang et al. 2005; Fichot and Benner 2012).

OWT3 denotes clear waters characterized by high transparency and relatively low concentrations of water constituents that do not co-vary. Remote sensing applications could be challenging in these waters due to the lack of diagnostic features while still providing the optical complexity that invalidates the use of blue-green ratio ocean chlorophyll algorithms. Specific absorption of phytoplankton and NAP in this OWT was generally high and in line with values recorded in coastal areas (e.g., Tilstone et al. 2012).

OWT4 represents turbid waters with moderate concentrations of Chl *a*, PC, CDOM, and dominance of  $a_{\text{NAP}}$  combined with high  $a_{\text{ph}}$  variability at the shorter wavelengths. Specific absorption of NAP of OWT4 was substantially high. Using the available data and reported information of the sites categorized in this OWT (Dall'Olmo and Gitelson 2006; Matthews and Bernard 2013), it can be deduced that the increased  $a_{\text{NAP}}^*(442)$  is related to high organic content of TSM (Ferrari and Dowell 1998; Babin et al. 2003).

OWT5 shows the brightly reflective nature of sediment-laden waters with high reflectance across a wide range of the spectrum. Similar reflectance spectra are described in highly turbid aquatic systems (Dekker 1993; Ruddick et al. 2006; Schalles 2006). Sites belonging to this optical type were mainly shallow floodplain (e.g., Amazon) and lowland lakes (e.g., Taihu) or rivers (e.g., Missouri). The ISM contribution to TSM in these waters is high (generally above 70% and on several occasions up to 100%), while  $a_{\text{NAP}}(442)$  is noticeably high and NAP is often the dominant component of light absorption. The dominance of particles of mineral origin is likely to be related to the observed low  $a_{\text{NAP}}(442) : \text{ISM}$  (mean =  $0.0736 \text{ m}^2 \text{ g}^{-1}$ ,  $N = 109$ ) values (Mikkelsen 2002).

OWT6 includes waters with balanced effects of optically active constituents to the absorption budget. This OWT pooled samples with relatively high concentrations of Chl *a* and PC and equal contributions of CDOM, phytoplankton, and NAP to absorption at blue wavelengths. Relatively high values of PC ( $62.5 \pm 51.21 \text{ mg m}^{-3}$ ) and PC to Chl *a* ratio ( $1.4 \pm 0.9$ ) reveal a significant presence of cyanobacteria in this OWT.

OWT7 delineates waters with particularly high values of Chl *a* concentrations and cyanobacteria abundances (PC : Chl *a*:  $1.9 \pm 0.8$  and PC:  $733.4 \pm 394.1 \text{ mg m}^{-3}$ ) and high  $R_{\text{rs}}(\lambda)$  at red/near-infrared spectral region (albeit lower than OWT1). In contrast to OWT1, OWT7 exhibits a pronounced reflectance peak around 700 nm.  $a_{\text{ph}}$  dominated the absorption budget at 442 nm while  $a_{\text{CDOM}}$  was high but very variable.

OWT8 is characterized by elevated concentrations of water constituents and especially of Chl *a* and accessory pigment PC (cyanobacteria presence) is also the main characteristic of OWT8. Nevertheless, Chl *a* and PC levels are lower when compared to OWT1 and OWT7, resulting in differences in  $R_{\text{rs}}(\lambda)$  amplitude and shape particularly in the red and near-infrared parts of the spectrum. In this context,  $R_{\text{rs}}(\lambda)$  appears lower at this spectral region while the reflectance peak is closer to 700 nm.

OWT9 shows similar spectra to those of OWT2 with an absence of a well-defined peak in the red to near-infrared region and increased non-standardized and standardized  $R_{\text{rs}}(\lambda)$  between 500 nm and 600 nm. Reflectance at shorter wavelengths was generally higher in OWT9. Optically active compounds in these waters were at similar concentrations to those observed in OWT2.

OWT10 differed from any of the other optical categories in having considerable lower reflectance from 400 nm to 600 nm with no discrete peaks and troughs in this part of the spectrum. However, a  $R_{\text{rs}}(\lambda)$  peak is noticeable near 700 nm. OWT10 grouped data collected from rivers and lakes with markedly higher concentrations of CDOM, which has a strong absorption effect at the shorter wavelengths < 500 nm (Kirk 1994; Del Vecchio and Blough 2004). Similar spectra have been previously reported in CDOM-rich environments (e.g., Kallio et al. 2001). Strömbeck and Pierson (2001) have shown that CDOM, at high concentrations, can significantly absorb light even in the red region.

OWT11 appears typical for inland waters with presence of cyanobacteria, high  $a_{\text{NAP}}^*(442)$  and high concentrations of CDOM. Reflectance spectra of this OWT appear with clearly observable but flattened peaks between 550 nm and 700 nm and with high red to blue ratios. The green maximum is suppressed and shifted to longer wavelengths due to strong CDOM absorption.

OWT12 represents turbid, moderately productive waters with cyanobacteria presence.  $R_{\text{rs}}(\lambda)$  spectral shapes resemble those of OWT11 but with a shorter wavelength of the green maximum while values are higher in the blue and lower from 580 nm to 720 nm.

Finally, OWT13 shows typical clear blue waters with high reflectance at shorter wavelengths and low reflectance values in the red region of the spectra, similar to clear oceanic waters (e.g., Cannizzaro and Carder 2006). This OWT was poorly represented in Dataset-I. In general, there is a scarcity of observations below  $3 \text{ mg m}^{-3}$  of Chl *a* (14%) or below  $3 \text{ mg L}^{-1}$  of TSM (7%) in Dataset-I which reflects the recent focus of research toward eutrophic lakes and reservoirs with harmful algal blooms.

### Relationships among optical clusters in inland and coastal waters

Synthesis and analysis of datasets coming from both inland and coastal waters provided a glimpse of the optical proximity between systems with a diverse range of properties. The results highlight common as well as unique spectral characteristics found in these waters, supporting a move toward an integrated optical classification framework for inland and coastal systems. This could be of great help especially in studies of multiple-component dynamic aquatic systems (Tyler et al. 2016) and global climatic trends. Classification of all available data led to 21 clusters of reflectance spectra (Fig. 7), many of which contained data from inland and coastal systems that importantly demonstrates a continuum of OWTs that extends across system boundaries. Previous related research (Moore et al. 2001, 2009, 2014; Reinart et al. 2003; Lubac and Loisel 2007; Le et al. 2011; Mélin et al. 2011; Spyrakos et al. 2011; Vantrepotte et al. 2012; Tilstone et al. 2012; Mélin and Vantrepotte 2015; Shen

et al. 2015; Ye et al. 2016) has suggested a substantially smaller number of optical clusters but these studies were primarily conducted at regional scales where sample sizes and the global representativeness of waterbodies considered might have limited the resolution of OWTs. Sun et al. (2012, 2014) suggested a different approach for optical classification of aquatic systems based on the normalized trough depth at 675 nm and data from turbid and productive waterbodies. This approach could be extremely useful especially for the retrieval of Chl *a* but its applicability to other environments included here (e.g., clear waters, high in  $a_{CDOM}$  waters) needs to be proven.

Many of the clusters described in these previous studies are represented in Figs. 4–7. Moreover, here we have considered waters with extreme scattering and/or absorbing properties, which have typically been omitted from previous optical classification schemes as outliers. In some cases, surface waters with extreme optical properties were found to form discretely identifiable optical clusters (e.g., cluster I3 and I10). The current analyses and results show a greater number of clusters in inland than in coastal and open-sea systems. This is, at least in part, explained by the larger size and geographical and seasonal coverage of the inland water dataset. However, given the diversity in inland waters, it is not unreasonable to suggest that these system could also comprise a larger portion of the optical diversity of natural waters. Despite these differences, the cluster analysis performed here has shown that some optical clusters are common to both inland and coastal waters. The phylogenetic tree of Fig. 12 represents the similarity of the second derivatives of all cluster means based on L2 norm distances and identified seven major groups. In parallel, it provided useful information regarding the parts of the spectra responsible for the observed similarities/dissimilarities between the clusters. These principally concern  $R_{rs}(\lambda)$  peak shifts, changes in the ratio of blue to green or red and features associated to accessory phytoplankton pigments. Such information should be considered when designing future EO missions.

### **Implications for implementation to satellite imagery**

The scope of this work was to identify distinct optical clusters and suggest OWTs for natural waters based on in situ data. Clusters were defined based on hyperspectral  $R_{rs}(\lambda)$  but these can be resampled to any sensor spectral resolution to assess the capability of differentiating clusters from EO data. In order to broadly evaluate the consistency of clustering results with respect to available EO satellite sensor wavebands, we performed a preliminary analysis to test the applicability of the approach. This included a comparison between the output of a spectral matching approach applied to multispectral sensor data simulated from in situ  $R_{rs}(\lambda)$  and the above mentioned clusters identified in the in situ datasets. Consistency was expressed as agreement between the

dominant cluster identified by spectral matching to the bands of the medium resolution imaging spectrometer (MERIS) and the *k*-means output where hyperspectral data were used. Values of 1 indicate perfect agreement while zero indicates no agreement between identified clusters. MERIS was chosen as the optimal sensor for this investigation due to its long catalogue of ocean color images (2002–2012) with a spatial resolution of 300 m, making it useful for coastal and inland water applications. However, similar results may be attained with alternative sensors such as ocean land colour instrument (OLCI) on Sentinel-3 and to some extent moderate resolution imaging spectroradiometer (MODIS). Different strategies are available to accomplish cluster assignment of satellite-derived spectra, but we followed the approach described in Moore et al. (2014) and Mélin and Vantrepotte (2015) that has already been implemented in the ESA Ocean Color-CCI project.

Spectra were standardized by dividing by the spectrum integral, which in every case led to substantial improvement in the value of cluster memberships. For the 13 clusters identified in Dataset-I, the cluster membership agreement was 0.85. However, in some cases, the differences in class membership between the top and second ranking cluster were negligible. When we considered shared top ranking for differences less than 0.001 in the membership between top and second ranking clusters, a perfect agreement was achieved. When considering spectra from coastal environments, membership agreement was lower (0.65). Agreement was improved with the removal of spectral bands between 700 nm and 800 nm prior to operation of the spectral matching routine. While encouraging, further refinement of the method is necessary to justify the classification scheme when all data are considered (21 clusters). Given that the application of this scheme on satellite imagery is sensitive to the performance of atmospheric correction methods, the selection of spectral bands must be exercised with caution. We anticipate that residual errors from incomplete atmospheric correction unrepresented in the OWT spectra and partition inefficiencies can result in spectra with zero or very low membership values. These spectra could be used to provide a better understanding of the representativeness of OWT and the limitations of atmospheric correction models and clustering methods.

### **Concluding remarks**

With increased interest in monitoring aquatic systems across wide temporal and spatial scales using remote sensing data, reliable OWT classification approaches are essential to deal with the optically diverse nature of aquatic systems, and to optimize the selection of atmospheric correction and water constituent algorithms. Through the use of a comprehensive dataset and the development of an elegant but robust approach for the classification of the in situ

hyperspectral measurements, we expect to better understand the variability of OWTs across inland and coastal waters and provide a framework to support global change research in coming years. Our methods and results can be used to identify OWT-specific technological and modeling requirements for remote sensing applications and highlight gaps in knowledge and data needs. In this regard, we note the rarity of particulate scattering and backscattering data and of standard protocols for radiometric measurements and data processing. Application of this approach to satellite imagery will require careful consideration of these confounding factors as well as the influence of uncertainties associated with atmospheric correction on the reflectance signal. Public access to cluster spectral means and covariance matrices are provided through the web page <http://www.globolakes.ac.uk/>.

## References

- Acker, J., S. Ouillon, R. Gould, and R. Arnone. 2005. Measuring marine suspended sediment concentrations from space: History and potential. Proc. 8th Int. Conf. Remote Sensing for Marine and Coastal Environments, Halifax.
- Arnone, R. A. 1985. Coastal Secchi depth atlas, p. 1–30. Mississippi: NORDA Rep. **83**.
- Arnone, R. A., A. M. Wood, and R. W. Gould, Jr. 2004. The evolution of optical water mass classification. *Oceanography* **17**: 14–15. doi:10.5670/oceanog.2004.42
- Babin, M., D. Stramski, G. M. Ferrari, H. Claustre, A. Bricaud, G. Obolensky, and N. Hoepffner. 2003. Variations in the light absorption coefficients of phytoplankton, nonalgal particles, and dissolved organic matter in coastal waters around Europe. *J. Geophys. Res.* **108**: 3211. doi:10.1029/2001JC000882
- Baker, K. S., and R. C. Smith. 1982. Bio-optical classification and model of natural waters. *Limnol. Oceanogr.* **27**: 500–509. doi:10.4319/lo.1982.27.3.0500
- Barbosa, E. 2007. Sensoriamento remoto da dinâmica da circulação da água do sistema planície de curuai/rio Amazonas. Ph.D. thesis. Instituto Nacional de Pesquisas Espaciais (INPE).
- Bessey, D. H. 2001. Object-oriented implementation of numerical methods: An introduction with Java and Smalltalk. Morgan Kaufmann.
- Binding, C. E., J. H. Jerome, R. P. Bukata, and W. G. Booty. 2008. Spectral absorption properties of dissolved and particulate matter in Lake Erie. *Remote Sens. Environ.* **112**: 1702–1711. doi:10.1016/j.rse.2007.08.017
- Binding, C. E., J. H. Jerome, R. P. Bukata, and W. G. Booty. 2010. Suspended particulate matter in Lake Erie derived from MODIS aquatic colour imagery. *Int. J. Remote Sens.* **31**: 5239–5255. doi:10.1080/01431160903302973
- Binding, C. E., T. A. Greenberg, J. H. Jerome, R. P. Bukata, and G. Letourneau. 2011. An assessment of MERIS algal products during an intense bloom in Lake of the Woods. *J. Plankton Res.* **33**: 793–806. doi:10.1093/plankt/fbq133
- Binding, C. E., T. A. Greenberg, and R. P. Bukata. 2013. The MERIS Maximum Chlorophyll Index; its merits and limitations for inland water algal bloom monitoring. *J. Great Lakes Res.* **39**: 100–107. doi:10.1016/j.jglr.2013.04.005
- Blondeau-Patissier, D., J. F. R. Gower, A. G. Dekker, S. R. Phinn, and V. E. Brandt. 2014. A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. *Prog. Oceanogr.* **123**: 123–144. doi:10.1016/j.pocean.2013.12.008
- Borges, A. V., G. Abril, F. Darchambeau, C. R. Teodoru, J. Deborde, L. O. Vidal, T. Lambert, and S. Bouillon. 2015. Divergent biophysical controls of aquatic CO<sub>2</sub> and CH<sub>4</sub> in the world's two largest rivers. *Sci. Rep.* **5**: 15614. doi:10.1038/srep15614
- Bradt, S. R. 2012. Development of bio-optical algorithms to estimate chlorophyll in the Great Salt Lake and New England lakes using in situ hyperspectral measurements. Ph.D. thesis. The Univ. of New Hampshire.
- Bresciani, M., D. Stroppiana, D. Odermatt, G. Morabito, and C. Giardino. 2011. Assessing remotely sensed chlorophyll-a for the implementation of the Water Framework Directive in European perialpine lakes. *Sci. Total Environ.* **409**: 3083–3091. doi:10.1016/j.scitotenv.2011.05.001
- Bricaud, A., M. Babin, A. Morel, and H. Claustre. 1995. Variability in the chlorophyll-specific absorption coefficients of natural phytoplankton: Analysis and parameterization. *J. Geophys. Res.* **100**: 13321–13332. doi:10.1029/95JC00463
- Bukata, R. P. 1995. Optical properties and remote sensing of inland and coastal waters. CRC Press.
- Cannizzaro, J. P., and K. L. Carder. 2006. Estimating chlorophyll a concentrations from remote-sensing reflectance in optically shallow waters. *Remote Sens. Environ.* **101**: 13–24. doi:10.1016/j.rse.2005.12.002
- Canziani, G., R. Ferrati, C. Marinelli, and F. Dukatz. 2008. Artificial neural networks and remote sensing in the analysis of the highly variable Pampean shallow lakes. *Math. Biosci. Eng.* **5**: 691–711. doi:10.3934/mbe.2008.5.691
- Cole, J. J., and others. 2007. Plumbing the global carbon cycle: Integrating inland waters into the terrestrial carbon budget. *Ecosystems* **10**: 172–185. doi:10.1007/s10021-006-9013-8
- Craig, S. E., S. E. Lohrenz, Z. Lee, K. L. Mahoney, G. J. Kirkpatrick, O. M. Schofield, and R. G. Steward. 2006. Use of hyperspectral remote sensing reflectance for detection and assessment of the harmful alga, *Karenia brevis*. *Appl. Opt.* **45**: 5414–5425. doi:10.1364/AO.45.005414
- Dall'Olmo, G., A. A. Gitelson, and D. C. Rundquist. 2003. Towards a unified approach for remote estimation of chlorophyll-a in both terrestrial vegetation and turbid productive waters. *Geophys. Res. Lett.* **30**: 1938. doi:10.1029/2003GL018065

- Dall'Olmo, G., and A. A. Gitelson. 2005. Effect of bio-optical parameter variability on the remote estimation of chlorophyll-a concentration in turbid productive waters: Experimental results. *Appl. Opt.* **44**: 412–422. doi:10.1364/AO.44.000412
- Dall'Olmo, G., A. A. Gitelson, D. C. Rundquist, B. Leavitt, T. Barrow, and J. C. Holz. 2005. Assessing the potential of SeaWiFS and MODIS for estimating chlorophyll concentration in turbid productive waters using red and near-infrared bands. *Remote Sens. Environ.* **96**: 176–187. doi:10.1016/j.rse.2005.02.007
- Dall'Olmo, G., and A. A. Gitelson. 2006. Absorption properties of dissolved and particulate matter in turbid productive inland lakes. *Proc. 8th Int. Conf. Remote Sensing for Marine and Coastal Environments*, Halifax.
- Dekker, A. G. 1993. Detection of optical water quality parameters for eutrophic waters by high resolution remote sensing. Ph.D. Thesis, Earth and Life Sciences, Amsterdam, The Netherlands. Proefschrift Vrije Universiteit (Free University).
- Del Vecchio, R., and N. V. Blough. 2004. On the origin of the optical properties of humic substances. *Environ. Sci. Technol.* **38**: 3885–3891. doi:10.1021/es049912h
- Dickey, T. D. 2003. Emerging ocean observations for interdisciplinary data assimilation systems. *J. Mar. Syst.* **40–41**: 5–48. doi:10.1016/S0924-7963(03)00011-3
- Doxaran, D., K. Ruddick, D. McKee, B. Gentili, D. Tailliez, M. Chami, and M. Babin. 2009. Spectral variations of light scattering by marine particles in coastal waters, from visible to near infrared. *Limnol. Oceanogr.* **54**: 1257. doi:10.4319/lo.2009.54.4.1257
- Eyob, E. 2009. Social implications of data mining and information privacy: Interdisciplinary frameworks and solutions. IGI Global.
- Fabry, V. J., B. A. Seibel, R. A. Feely, and J. C. Orr. 2008. Impacts of ocean acidification on marine fauna and ecosystem processes. *ICES J. Mar. Sci.* **65**: 414–432. doi:10.1093/icesjms/fsn048
- Fargion, G. S., and J. L. Mueller. 2002. Ocean optics protocols for satellite ocean color sensor validation, revision 3. National Aeronautics and Space Administration, Goddard Space Flight Center.
- Ferrari, G., and M. Dowell. 1998. CDOM absorption characteristics with relation to fluorescence and salinity in coastal areas of the southern Baltic Sea. *Estuar. Coast. Shelf Sci.* **47**: 91–105. doi:10.1006/ecss.1997.0309
- Ficek, D., J. Meler, T. Zapadka, B. Woźniak, and J. Dera. 2012. Inherent optical properties and remote sensing reflectance of Pomeranian lakes (Poland). *Oceanologia* **54**: 611–630. doi:10.5697/oc.54-4.611
- Fichot, C. G., and R. Benner. 2012. The spectral slope coefficient of chromophoric dissolved organic matter (S<sub>275–295</sub>) as a tracer of terrigenous dissolved organic carbon in river-influenced ocean margins. *Limnol. Oceanogr.* **57**: 1453–1466. doi:10.4319/lo.2012.57.5.1453
- Fraley, C., and A. E. Raftery. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **41**: 578–588. doi:10.1093/comjnl/41.8.578
- Galloway, J. N., and others. 2004. Nitrogen cycles: Past, present, and future. *Biogeochemistry* **70**: 153–226. doi:10.1007/s10533-004-0370-0
- Giardino, C., G. Candiani, and E. Zilioli. 2005. Detecting chlorophyll-a in Lake Garda using TOA MERIS radiances. *Photogramm. Eng. Remote Sensing* **71**: 1045–1051. doi:10.14358/PERS.71.9.1045
- Giardino, C., M. Bresciani, I. Cazzaniga, K. Schenk, P. Rieger, F. Braga, E. Matta, and V. E. Brando. 2014a. Evaluation of multi-resolution satellite sensors for assessing water quality and bottom depth of Lake Garda. *Sensors* **14**: 24116–24131. doi:10.3390/s141224116
- Giardino, C., M. Bresciani, D. Stroppiana, A. Oggioni, and G. Morabito. 2014b. Optical remote sensing of lakes: An overview on Lake Maggiore. *J. Limnol.* **73**: 201–214. doi:10.4081/jlimnol.2014.817
- Giardino, C., M. Bresciani, E. Valentini, L. Gasperini, R. Bolpagni, and V. E. Brando. 2015. Airborne hyperspectral data to assess suspended particulate matter and aquatic vegetation in a shallow and turbid lake. *Remote Sens. Environ.* **157**: 48–57. doi:10.1016/j.rse.2014.04.034
- Gitelson, A. A., Y. Z. Yacobi, J. F. Schalles, D. C. Rundquist, L. Han, R. Stark, and D. Etzion. 2000. Remote estimation of phytoplankton density in productive waters. *Arch. Hydrobiol. Spec. Issues. Advanc. Limnol.* **55**: 121–136.
- Gitelson, A. A., J. F. Schalles, and C. M. Hladik. 2007. Remote chlorophyll-a retrieval in turbid, productive estuaries: Chesapeake Bay case study. *Remote Sens. Environ.* **109**: 464–472. doi:10.1016/j.rse.2007.01.016
- Gitelson, A. A., G. Dall'Olmo, W. Moses, D. C. Rundquist, T. Barrow, T. R. Fisher, D. Gurlin, and J. Holz. 2008. A simple semi-analytical model for remote estimation of chlorophyll-a in turbid waters: Validation. *Remote Sens. Environ.* **112**: 3582–3593. doi:10.1016/j.rse.2008.04.015
- Gitelson, A. A., D. Gurlin, W. J. Moses, and T. Barrow. 2009. A bio-optical algorithm for the remote estimation of the chlorophyll-a concentration in case 2 waters. *Environ. Res. Lett.* **4**: 12–25. doi:10.1088/1748-9326/4/4/045003
- González Vilas, L., E. Spirakos, and J. M. Torres Palenzuela. 2011. Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain). *Remote Sens. Environ.* **115**: 524–535. doi:10.1016/j.rse.2010.09.021
- Gordon, H. R., O. B. Brown, R. H. Evans, J. W. Brown, R. C. Smith, K. S. Baker, and D. K. Clark. 1988. A semianalytic radiance model of ocean color. *J. Geophys. Res.* **93**: 10909–10924, doi:10.1029/JD093iD09p10909
- Gordon, H. R., and B. A. Franz. 2008. Remote sensing of ocean color: Assessment of the water-leaving radiance bidirectional effects on the atmospheric diffuse transmittance

- for SeaWiFS and MODIS intercomparisons. *Remote Sens. Environ.* **112**: 2677–2685. doi:10.1016/j.rse.2007.12.010
- Guanter, L., and others. 2010. Atmospheric correction of ENVISAT/MERIS data over inland waters: Validation for European lakes. *Remote Sens. Environ.* **114**: 467–480. doi:10.1016/j.rse.2009.10.004
- Guo, H., L. Zhang, and L. Zhu. 2015. Earth observation big data for climate change research. *Advances in Climate Change Research* **6**: 108–117. doi:10.1016/j.accre.2015.09.007
- Gurlin, D., A. A. Gitelson, and W. J. Moses. 2011. Remote estimation of chl-a concentration in turbid productive waters - return to a simple two-band NIR-red model? *Remote Sens. Environ.* **115**: 3479–3490. doi:10.1016/j.rse.2011.08.011
- Hestir, E. L., V. E. Brando, M. Bresciani, C. Giardino, E. Matta, P. Villa, and A. G. Dekker. 2015. Measuring freshwater aquatic ecosystems: The need for a hyperspectral global mapping satellite mission. *Remote Sens. Environ.* **167**: 181–195. doi:10.1016/j.rse.2015.05.023
- Huang, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.* **2**: 283–304. doi:10.1023/A:1009769707641
- Hubert, L., and P. Arabie. 1985. Comparing partitions. *J. Classification* **2**: 193–218. doi:10.1007/BF01908075
- Jaelani, L. M., B. Matsushita, W. Yang, and T. Fukushima. 2013. Evaluation of four MERIS atmospheric correction algorithms in Lake Kasumigaura, Japan. *Int. J. Remote Sens.* **3**: 8967–8985. doi:10.1080/01431161.2013.860660
- Jain, A. K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recog. Lett.* **31**: 651–666. doi:10.1016/j.patrec.2009.09.011
- Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. Data clustering: A review. *Acm Comput. Surv.* **31**: 264–323. doi:10.1145/331499.331504
- Jerlov, N. G. 1977. Classification of sea water in terms of quanta irradiance. *ICES J. Mar. Sci.* **37**: 281–287. doi:10.1093/icesjms/37.3.281
- Kallio, K., T. Kutser, T. Hannonen, S. Koponen, J. Pulliainen, J. Vepsäläinen, and T. Pyhälähti. 2001. Retrieval of water quality from airborne imaging spectrometry of various lake types in different seasons. *Sci. Total Environ.* **268**: 59–77. doi:10.1016/S0048-9697(00)00685-9
- Karl, M. D. 1999. A sea of change: Biogeochemical variability in the North Pacific Subtropical Gyre. *Ecosystems* **2**: 181–214. doi:10.1007/s100219900068
- Kirk, J. T. O. 1994. *Light and photosynthesis in aquatic ecosystems*. Cambridge Univ. Press.
- Kudela, R. M., S. L. Palacios, D. C. Austerberry, E. K. Accorsi, L. S. Guild, and J. Torres-Perez. 2015. Application of hyperspectral remote sensing to cyanobacterial blooms in inland waters. *Remote Sens. Environ.* **167**: 196–205. doi:10.1016/j.rse.2015.01.025
- Kutser, T., B. Paavel, C. Verpoorter, T. Kauer, and E. Vahtmäe. 2012. Remote sensing of water quality in optically complex lakes. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 165–169. doi:10.5194/isprsarchives-XXXIX-B8-165-2012
- Kutser, T., E. Vahtmäe, B. Paavel, and T. Kauer. 2013. Removing glint effects from field radiometry data measured in optically complex coastal and inland waters. *Remote Sens. Environ.* **133**: 85–89. doi:10.1016/j.rse.2013.02.011
- Le, C., Y. Li, Y. Zha, D. Sun, C. Huang, and H. Zhang. 2011. Remote estimation of chlorophyll a in optically complex waters based on optical classification. *Remote Sens. Environ.* **115**: 725–737. doi:10.1016/j.rse.2010.10.014
- Le Quéré, C. R., and others. 2015. Global carbon budget 2014. *Earth Syst. Sci. Data* **7**: 47–85. doi:10.5194/essd-7-47-2015
- Lee, Z., K. L. Carder, C. D. Mobley, R. G. Steward, and J. S. Patch. 1999. Hyperspectral remote sensing for shallow waters. 2. Deriving bottom depths and water properties by optimization. *Appl. Opt.* **38**: 3831–3843. doi:10.1364/AO.38.003831
- Li, L., L. Li, K. Song, Y. Li, L. P. Tedesco, K. Shi, and Z. Li. 2013. An inversion model for deriving inherent optical properties of inland waters: Establishment, validation and application. *Remote Sens. Environ.* **135**: 150–166. doi:10.1016/j.rse.2013.03.031
- Li, L., L. Li, and K. Song. 2015. Remote sensing of freshwater cyanobacteria: An extended IOP Inversion Model of Inland Waters (IIMIWI) for partitioning absorption coefficient and estimating phycocyanin. *Remote Sens. Environ.* **157**: 9–23. doi:10.1016/j.rse.2014.06.009
- Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**: 129–137. doi:10.1109/TIT.1982.1056489
- Loisel, H., and A. Morel. 2001. Non-isotropy of the upward radiance field in typical coastal (case 2) waters. *Int. J. Remote Sens.* **22**: 275–295. doi:10.1080/014311601449934
- Lubac, B., and H. Loisel. 2007. Variability and classification of remote sensing reflectance spectra in the eastern English Channel and southern North Sea. *Remote Sens. Environ.* **110**: 45–58. doi:10.1016/j.rse.2007.02.012
- MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**: 281–297. University of California Press, Berkeley, Calif. <https://projecteuclid.org/euclid.bsm/1200512992>.
- Manzo, C., M. Bresciani, C. Giardino, F. Braga, and C. Bassani. 2015. Sensitivity analysis of a bio-optical model for Italian lakes focused on Landsat-8, Sentinel-2 and Sentinel-3. *Eur. J. Remote Sens.* **48**: 17–32. doi:10.5721/EuJRS20154802
- Martin Traykovski, L. V., and H. M. Sosik. 2003. Feature-based classification of optical water types in the Northwest Atlantic based on satellite ocean color data.

- J. Geophys. Res. C Oceans **108**: 19-1-19-18. doi:10.1029/2001JC001172
- Matsushita, B., W. Yang, G. Yu, Y. Oyama, K. Yoshimura, and T. Fukushima. 2015. A hybrid algorithm for estimating the chlorophyll-a concentration across different trophic states in Asian inland waters. *ISPRS J. Photogramm. Remote Sens.* **102**: 28–37. doi:10.1016/j.isprsjprs.2014.12.022
- Matthews, M. W. 2011. A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters. *Int. J. Remote Sens.* **32**: 6855–6899. doi:10.1080/01431161.2010.512947
- Matthews, M. W. 2014. Eutrophication and cyanobacterial blooms in South African inland waters: 10 years of MERIS observations. *Remote Sens. Environ.* **155**: 161–177. doi:10.1016/j.rse.2014.08.010
- Matthews, M. W., and S. Bernard. 2013. Characterizing the absorption properties for remote sensing of three small optically-diverse South African reservoirs. *Remote Sens.* **5**: 4370–4404. doi:10.3390/rs5094370
- Mélin, F., V. Vantrepotte, M. Clerici, D. D'Alimonte, G. Zibordi, J. Berthon, and E. Canuti. 2011. Multi-sensor satellite time series of optical properties and chlorophyll-a concentration in the Adriatic Sea. *Prog. Oceanogr.* **91**: 229–244. doi:10.1016/j.pocean.2010.12.001
- Mélin, F., and V. Vantrepotte. 2015. How optically diverse is the coastal ocean? *Remote Sens. Environ.* **160**: 235–251. doi:10.1016/j.rse.2015.01.023
- Mikkelsen, O. A. 2002. Variation in the projected surface area of suspended particles: Implications for remote sensing assessment of TSM. *Remote Sens. Environ.* **79**: 23–29. doi:10.1016/S0034-4257(01)00235-8
- Mobley, C. D. 1994. *Light and water: Radiative transfer in natural waters.* Academic Press.
- Moore, T. S., J. W. Campbell, and H. Feng. 2001. A fuzzy logic classification scheme for selecting and blending satellite ocean color algorithms. *IEEE Trans. Geosci. Remote Sens.* **39**: 1764–1776. doi:10.1109/36.942555
- Moore, T. S., J. W. Campbell, and M. D. Dowell. 2009. A class-based approach to characterizing and mapping the uncertainty of the MODIS ocean chlorophyll product. *Remote Sens. Environ.* **113**: 2424–2430. doi:10.1016/j.rse.2009.07.016
- Moore, T. S., M. D. Dowell, S. Bradt, and A. Ruiz Verdu. 2014. An optical water type framework for selecting and blending retrievals from bio-optical algorithms in lakes and coastal waters. *Remote Sens. Environ.* **143**: 97–111. doi:10.1016/j.rse.2013.11.021
- Morel, A., and L. Prieur. 1977. Analysis of variations in ocean color. *Limnol. Oceanogr.* **22**: 709–722. doi:10.4319/lo.1977.22.4.0709
- Morel, A., and S. Maritorena. 2001. Bio-optical properties of oceanic waters: A reappraisal. *J. Geophys. Res.* **106**: 7163–7180. doi:10.1029/2000JC000319
- Odermatt, D., A. Gitelson, V. Brando, and M. Schaepman. 2012. Review of constituent retrieval in optically deep and complex waters from satellite imagery. *Remote Sens. Environ.* **118**: 116–126. doi:10.1016/j.rse.2011.11.013
- Palacios, S. L., T. D. Peterson, and R. M. Kudela. 2012. Optical characterization of water masses within the Columbia River plume. *J. Geophys. Res.* **117**: C11020. doi:10.1029/2012JC008005
- Peters, D. P. C., B. T. Bestelmeyer, and M. G. Turner. 2007. Cross-scale interactions and changing pattern-process relationships: Consequences for system dynamics. *Ecosystems* **10**: 790–796. doi:10.1007/s10021-007-9055-6
- Petrescu, A. M. R., and others. 2015. The uncertain climate footprint of wetlands under human pressure. *Proc. Natl. Acad. Sci. USA.* **112**: 4594–4599. doi:10.1073/pnas.1416267112
- Prieur, L., and S. Sathyendranath. 1981. An optical classification of coastal and oceanic waters based on the specific spectral absorption curves of phytoplankton pigments, dissolved organic matter, and other particulate materials. *Limnol. Oceanogr.* **26**: 671–689. doi:10.4319/lo.1981.26.4.0671
- Ramsay, J. O. 2006. *Functional data analysis.* Wiley Online Library.
- Raymond, P. A., and others. 2013. Global carbon dioxide emissions from inland waters. *Nature* **503**: 355–359. doi:10.1038/nature12760
- Reinart, A., A. Herlevi, H. Arst, and L. Sipelgas. 2003. Preliminary optical classification of lakes and coastal waters in Estonia and south Finland. *J. Sea Res.* **49**: 357–366. doi:10.1016/S1385-1101(03)00019-4
- Riddick, C., and others. 2015. Spatial variability of absorption coefficients over a biogeochemical gradient in a large and optically complex shallow lake. *J. Geophys. Res. Oceans* **120**: 7040–7066. doi:10.1002/2015JC011202
- Ruddick, K. G., V. De Cauwer, Y.-J. Park, and G. Moore. 2006. Seaborne measurements of near infrared water-leaving reflectance: The similarity spectrum for turbid waters. *Limnol. Oceanogr.* **51**: 1167–1179. doi:10.4319/lo.2006.51.2.1167
- Ruiz-Verdú, A., J. Domínguez-Gómez, and R. Peña-Martínez. 2005. Use of CHRIS for monitoring water quality in Rosarito reservoir. *ESA Special Publication, ESA Special Publication*, **593**: pp. 26.
- Ruiz-Verdú, A., S. G. H. Simis, C. de Hoyos, H. J. Gons, and R. Peña-Martínez. 2008. An evaluation of algorithms for the remote sensing of cyanobacterial biomass. *Remote Sens. Environ.* **112**: 3996–4008. doi:10.1016/j.rse.2007.11.019
- Schalles, J. F. 2006. Optical remote sensing techniques to estimate phytoplankton chlorophyll *a* concentrations in coastal waters with varying suspended matter and CDOM concentrations, p. 27–79. *In* L. Richardson and E. Ledrew [eds.], *Remote sensing of aquatic coastal ecosystem processes: Science and management applications.* Springer.

- Schalles, J. F., and C. M. Hladik. 2012. Mapping phytoplankton chlorophyll in turbid, Case 2 estuarine and coastal waters. *Isr. J. Plant Sci.* **60**: 169–191. doi:10.1560/JJPS.60.1-2.169
- Shen, Q., J. Li, F. Zhang, X. Sun, J. Li, W. Li, and B. Zhang. 2015. Classification of several optically complex waters in China using in situ remote sensing reflectance. *Remote Sens.* **7**: 429–440. doi:10.3390/rs71114731
- Shi, K., Y. Li, Y. Zhang, L. Li, H. Lv, and K. Song. 2014. Classification of inland waters based on bio-optical properties. *J. Sel. Topics Appl. Earth Observ. Remote Sens.* **7**: 543–561. doi:10.1109/JSTARS.2013.2290744
- Simis, S., S. Peters, and H. Gons. 2005. Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water. *Limnol. Oceanogr.* **50**: 237–245. doi:10.4319/lo.2005.50.1.0237
- Simis, S. G. H., A. Ruiz-Verdú, J. A. Domínguez-Gómez, R. Peña-Martínez, S. W. M. Peters, and H. J. Gons. 2007. Influence of phytoplankton pigment composition on remote sensing of cyanobacterial biomass. *Remote Sens. Environ.* **106**: 414–427. doi:10.1016/j.rse.2006.09.008
- Spirakos, E., L. González Vilas, J. M. Torres Palenzuela, and E. D. Barton. 2011. Remote sensing chlorophyll a of optically complex waters (rias Baixas, NW Spain): Application of a regionally specific chlorophyll a algorithm for MERIS full resolution data during an upwelling cycle. *Remote Sens. Environ.* **115**: 2471–2485. doi:10.1016/j.rse.2011.05.008
- Strömbeck, N., and D. C. Pierson. 2001. The effects of variability in the inherent optical properties on estimations of chlorophyll a by remote sensing in Swedish freshwaters. *Sci. Total Environ.* **268**: 123–137. doi:10.1016/S0048-9697(00)00681-1
- Sun, D., Y. Li, Q. Wang, C. Le, H. Lv, C. Huang, and S. Gong. 2012. Specific inherent optical quantities of complex turbid inland waters, from the perspective of water classification. *Photochem. Photobiol. Sci.* **11**: 1299–1312. doi:10.1039/c2pp25061f
- Sun, D., C. Hu, Z. Qiu, J. P. Cannizzaro, and B. B. Barnes. 2014. Influence of a red band-based water classification approach on chlorophyll algorithms for optically complex estuaries. *Remote Sens. Environ.* **155**: 289–302. doi:10.1016/j.rse.2014.08.035
- Tarpey, T. 2007. Linear transformations and the k-means clustering algorithm: Applications to clustering curves. *Am. Stat.* **61**: 34–40. doi:10.1198/000313007X171016
- Tarpey, T., and J. K. Kinader. 2003. Clustering functional data. *J. Classif.* **20**: 093–114. doi:10.1007/s00357-003-0007-3
- Tebbs, E. J., J. J. Remedios, and D. M. Harper. 2013. Remote sensing of chlorophyll-a as a measure of cyanobacterial biomass in Lake Bogoria, a hypertrophic, saline-alkaline, flamingo lake, using Landsat ETM+. *Remote Sens. Environ.* **135**: 92–106. doi:10.1016/j.rse.2013.03.024
- Tibshirani, R., G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic'. *J. R. Stat. Soc. Series B Stat. Methodol.* **63**: 411–423. doi:10.1111/1467-9868.00293
- Tilstone, G. H., and others. 2012. Variability in specific-absorption properties and their use in a semi-analytical ocean colour algorithm for MERIS in North Sea and Western English Channel Coastal Waters. *Remote Sens. Environ.* **118**: 320–338. doi:10.1016/j.rse.2011.11.019
- Torrecilla, E., D. Stramski, R. A. Reynolds, E. Millán-Núñez, and J. Piera. 2011. Cluster analysis of hyperspectral optical data for discriminating phytoplankton pigment assemblages in the open ocean. *Remote Sens. Environ.* **115**: 2578–2593. doi:10.1016/j.rse.2011.05.014
- Tyler, A. N., P. D. Hunter, E. Spirakos, S. Groom, A. M. Constantinescu, and J. Kitchen. 2016. Developments in Earth observation for the assessment and monitoring of inland, transitional, coastal and shelf-sea waters. *Sci. Total Environ.* **572**: 1307–1321. doi:10.1016/j.scitotenv.2016.01.020
- Vantrepotte, V., H. Loisel, D. Dessailly, and X. Mériaux. 2012. Optical classification of contrasted coastal waters. *Remote Sens. Environ.* **123**: 306–323. doi:10.1016/j.rse.2012.03.004
- Verpoorter, C., T. Kutser, D. A. Seekell, and L. J. Tranvik. 2014. A global inventory of lakes based on high-resolution satellite imagery. *Geophys. Res. Lett.* **41**: 6396–6402. doi:10.1002/2014GL060641
- Werdell, P. J., and S. W. Bailey. 2002. The SeaWiFS Bio-optical Archive and Storage System (SeaBASS): Current architecture and implementation, p. 45. *In* G. S. Fargion and C. R. McClain [eds.], NASA Tech. Memo. 2002-211617. NASA Goddard Space Flight Center.
- Werdell, P. J., S. Bailey, G. Fargion, C. Pietras, K. Knobelspiesse, G. Feldman, and C. McClain. 2003. Unique data repository facilitates ocean color satellite validation. *EOS Trans. AGU* **84**: 377. doi:10.1029/2003EO380001
- World Resources Institute. 2005. Ecosystems and human well-being: Synthesis.
- Yacobi, Y. Z., J. J. Alberts, M. Takacs, and M. McElvaine. 2003. Absorption spectroscopy of colored dissolved organic carbon in Georgia (USA) rivers: The impact of molecular size distribution. *J. Limnol.* **62**: 41–46. doi:10.4081/jlimnol.2003.41
- Yacobi, Y. Z., W. J. Moses, S. Kaganovsky, B. Sulimani, B. C. Leavitt, and A. A. Gitelson. 2011. NIR-red reflectance-based algorithms for chlorophyll-a estimation in mesotrophic inland and coastal waters: Lake Kinneret case study. *Water research* **45**: 2428–2436. doi:10.1016/j.watres.2011.02.002
- Yang, J., and others. 2013. The role of satellite remote sensing in climate change studies. *Nat. Clim. Chang.* **3**: 875–883. doi:10.1038/nclimate1908



- Ye, H., J., and others. 2016. Spectral classification of the Yellow Sea and implications for coastal ocean color remote sensing. *Remote Sen.* **8**: 321. doi:[10.3390/rs8040321](https://doi.org/10.3390/rs8040321)
- Zhang, Y., B. Zhang, X. Wang, J. Li, S. Feng, Q. Zhao, M. Liu, and B. Qin. 2007. A study of absorption characteristics of chromophoric dissolved organic matter and particles in lake Taihu, China. *Hydrobiologia* **592**: 105–120. doi:[10.1007/s10750-007-0724-4](https://doi.org/10.1007/s10750-007-0724-4)
- Zhang, Y., L. Feng, J. Li, L. Luo, Y. Yin, M. Liu, and Y. Li. 2010. Seasonal and spatial variation and remote sensing of phytoplankton absorption in lake taihu, a large eutrophic and shallow lake in China. *J. Plankton Res.* **32**: 1023–1037. doi:[10.1093/plankt/fbq039](https://doi.org/10.1093/plankt/fbq039)
- Zhang, Y. L., B. Q. Qin, W. M. Chen, and G. W. Zhu. 2005. A preliminary study of chromophoric dissolved organic matter (CDOM) in Lake Taihu, a shallow subtropical lake in China. *Acta Hydrochim. Hydrobiol.* **33**: 315–323. doi:[10.1002/aheh.200400585](https://doi.org/10.1002/aheh.200400585)

#### Acknowledgments

Many thanks to NASA and SeaBASS for providing data through the web-page <http://seabass.gsfc.nasa.gov/>. We would like to thank the

reviewers and the associate Editor for their insightful comments on the manuscript. E. S., P. D. H., R. O'D., C. M., M. S., V. M-V., S. S., A. T. gratefully acknowledge funding from the UK NERC-funded GloboLakes project (REF NE/J024279/1) as well as collaborative support from the ESA-funded DIVERSITY II project led by Brockmann Consult and NERC Field Spectroscopy Facility. All authors gratefully acknowledge funding received toward the collection of the data and all people assisted in producing and managing the datasets.

**Author Contribution Statement:** ES, PDH, RO'D, CM, MS, SS, AT conceived and designed the analysis; ES assembled, analyzed, and interpreted data and wrote the manuscript; RO'D and ES wrote code and performed the classification analysis; RO'D worked on the functional analysis; PH, AT, SS, CM, MS supervised the work and commented on the manuscript at all stages; PDH, CB, SB, MB, GD'O, CG, AG, TK, LL, BM, VM-V, MM, AR-V, JS, SS, ET, YZ, AT provided in situ data and edited the manuscript.

#### Conflict of Interest

None declared.

*Submitted 27 March 2017*

*Revised 19 July 2017*

*Accepted 07 August 2017*

*Associate editor: David Antoine*