



Using empirical orthogonal functions derived from remote-sensing reflectance for the prediction of phytoplankton pigment concentrations

A. Bracher^{1,2}, M. H. Taylor³, B. Taylor¹, T. Dinter^{1,2}, R. Röttgers⁴, and F. Steinmetz⁵

¹Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bussestraße 24, 27570 Bremerhaven, Germany

²Institute of Environmental Physics, University of Bremen, Otto-Hahn-Allee 1, 28359 Bremen, Germany

³Leibniz Center for Tropical Marine Ecology, Fahrenheitstraße 6, 28359 Bremen, Germany

⁴Helmholtz Zentrum Geesthacht Center of Materials and Coastal Research, Max-Planck-Str., 21502 Geesthacht, Germany

⁵HYGEOS, Euratechnologies, 165 Avenue de Bretagne, 59000 Lille, France

Correspondence to: A. Bracher (astrid.bracher@awi.de)

Received: 30 July 2014 – Published in Ocean Sci. Discuss.: 11 September 2014

Revised: 29 November 2014 – Accepted: 6 January 2015 – Published: 3 February 2015

Abstract. The composition and abundance of algal pigments provide information on phytoplankton community characteristics such as photoacclimation, overall biomass and taxonomic composition. In particular, pigments play a major role in photoprotection and in the light-driven part of photosynthesis. Most phytoplankton pigments can be measured by high-performance liquid chromatography (HPLC) techniques applied to filtered water samples. This method, as well as other laboratory analyses, is time consuming and therefore limits the number of samples that can be processed in a given time. In order to receive information on phytoplankton pigment composition with a higher temporal and spatial resolution, we have developed a method to assess pigment concentrations from continuous optical measurements. The method applies an empirical orthogonal function (EOF) analysis to remote-sensing reflectance data derived from ship-based hyperspectral underwater radiometry and from multispectral satellite data (using the Medium Resolution Imaging Spectrometer – MERIS – Polymer product developed by Steinmetz et al., 2011) measured in the Atlantic Ocean. Subsequently we developed multiple linear regression models with measured (collocated) pigment concentrations as the response variable and EOF loadings as predictor variables. The model results show that surface concentrations of a suite of pigments and pigment groups can be well predicted from the ship-based reflectance measurements, even when

only a multispectral resolution is chosen (i.e., eight bands, similar to those used by MERIS). Based on the MERIS reflectance data, concentrations of total and monovinyl chlorophyll *a* and the groups of photoprotective and photosynthetic carotenoids can be predicted with high quality. As a demonstration of the utility of the approach, the fitted model based on satellite reflectance data as input was applied to 1 month of MERIS Polymer data to predict the concentration of those pigment groups for the whole eastern tropical Atlantic area. Bootstrapping explorations of cross-validation error indicate that the method can produce reliable predictions with relatively small data sets (e.g., < 50 collocated values of reflectance and pigment concentration). The method allows for the derivation of time series from continuous reflectance data of various pigment groups at various regions, which can be used to study variability and change of phytoplankton composition and photophysiology.

1 Introduction

Optical measurements taken from various platforms have been successfully used to determine the total chlorophyll *a* (TChl *a*) concentration (e.g., see the summary by McClain 2009). Those measurements can be taken continuously,

thereby allowing for the estimation of TChl *a* concentration at a much higher temporal and spatial resolution than possible from chemical measurements in the laboratory, e.g., by high-performance liquid chromatography (HPLC) analysis of discrete water samples. Chl *a* is the major pigment in all phytoplankton species and is often used as an indicator of phytoplankton biomass. When pigments are measured by HPLC, TChl *a* is defined as the sum of monovinyl Chl *a* (MVChl *a*), divinyl Chl *a* (DVChl *a*) and chlorophyllide *a* (which is mainly formed as an artifact of the former two during the extraction process and therefore included in the calculation). DVChl *a* exists only in the prokaryotic genus *Prochlorococcus*, while MVChl *a* is the Chl *a* pigment for all other phytoplankton (other cyanobacteria and eukaryotes). Besides Chl *a*, there are many other pigments in phytoplankton that are either involved in light harvesting, such as chlorophyll *b* (Chl *b*), chlorophyll *c* (Chl *c*) and photosynthetic carotenoids (PSC), or in protecting Chl *a* and other sensitive pigments from photodamage, such as photoprotective carotenoids (PPC). Some pigments only occur in certain phytoplankton groups and thus are indicator pigments for their identification, e.g., peridinin in dinoflagellates (e.g., Letelier et al., 1993; Vidussi et al., 2001).

When analyzing biogeochemical fluxes in the oceans, however, it is inadequate to consider phytoplankton as a single variable (i.e., TChl *a*) because various groups have different roles in the biogeochemical processes (such as carbon fixation and export, nitrogen fixation, and silicon uptake). TChl *a* is far from being a sole function of phytoplankton biomass and varies, as other phytoplankton pigments do, with taxonomic composition and mean physiological algal assemblage state in response to several factors such as light, temperature and nutrients (Behrenfeld and Boss, 2006). Thus, knowledge of a wider array of phytoplankton pigment concentrations provides insight into phytoplankton composition, overall light absorption and physiological state. Phytoplankton absorption bears the imprints of different types of pigments and can be measured by optical measurements. However, different phytoplankton pigments may correlate in parts of their spectrum, making individual pigment detection difficult.

Several recent studies have investigated the potential of using continuous optical data to derive surface concentrations of pigments other than TChl *a*, with the advantage of being able to supply estimates over larger spatial and temporal scales than obtained with in situ water sampling. Chase et al. (2013) decomposed a large global data set of hyperspectral particulate absorption measurements into Gaussian function components and assessed the magnitude of specific Gaussian functions in relation to the absorption by specific pigments or pigment groups. The method provided robust results for obtaining concentrations of TChl *a*, TChl *b* (sum of different types of Chl *b*), TChl *c* (sum of different types of Chl *c*), PSC, PPC and phycoerythrin (PE). Organelli et al. (2013) used a multivariate approach applied to fourth-

derivative spectra of phytoplankton or particulate absorption (a_{ph} and a_p , respectively) data to retrieve TChl *a*, the total concentrations of seven diagnostic pigments and three phytoplankton size classes. However, a_p and a_{ph} are inherent optical properties (IOP) which cannot be directly determined from satellite ocean-color measurements (after successful atmospheric correction), such as the apparent optical properties (AOP). The estimation of IOP from AOP is based on a certain inversion model (e.g., the Quasi-Analytical Algorithm by Lee et al., 2002), which introduces additional uncertainty.

The water-leaving reflectance (ρ_w) is related not only to phytoplankton absorption but also to the scattering and absorption of water and other water constituents and to changes in the radiance distribution in response to environmental conditions such as observation geometry, surface waves and atmospheric conditions.

Pan et al. (2010) developed empirical algorithms based on reflectance ratios to approximate key phytoplankton pigment concentrations. The band-ratio algorithms were developed from underwater radiometric measurements collocated to pigment data taken in northeastern US coastal waters and were successful in deriving the concentration of TChl *a*, TChl *b*, TChl *c* and nine different carotenoids. However, such band-ratio algorithms require a very large database (> 400 collocations with satellite data) from a certain region to derive robust results. Pan et al. (2013) later described that the algorithm had to be adapted by modifying the pigment-specific coefficients based on a regionally specific data set.

Craig et al. (2012) developed local models to estimate TChl *a* and a_{ph} at different wavelengths from hyperspectral in situ measurements of remote-sensing reflectance, $R_{rs}(\lambda)$, in an optically complex water body. The models were based on empirical orthogonal functions (EOF) analysis of normalized $R_{rs}(\lambda)$ spectra and a subsequent linear fitting of measured TChl *a* concentration and $a_{ph}(\lambda)$ as response variables to EOF loadings as predictor variables. Taylor et al. (2013) showed that the method could be used similarly to derive PE concentrations from underwater upwelling radiance spectra, $L_u(\lambda)$, which enabled continuous profile predictions of PE concentrations.

The present study aims to use the spectral information contained in reflectance data to derive the optical signature of different pigments by an automatic and generic technique. The EOF analysis is applied to R_{rs} and to ρ_{wN} (i.e., normalized ρ_w just above surface) data measured in the field and by satellite sensors, respectively, in the Atlantic Ocean. The dominant EOF loadings were subsequently assessed as predictors in a multiple linear regression for the concentration of phytoplankton pigments and pigment groups as response variables. The prediction error of each model is evaluated by a permuted cross-validation routine, which is used to estimate the critical sample sizes necessary for reliable prediction. In addition, we demonstrate the approach's utility in estimating the large-scale distribution and photophysiology of the phytoplankton assemblage.

2 Material and methods

Two sets of optical and pigment data from the Atlantic Ocean were used in the analysis. The first model setup used a data set which included only optical measurements taken in situ (as depth profiles) and collocated surface pigment data collected during three transatlantic RV *Polarstern* cruises in 2008 and 2010. These data enabled us to study the difference in EOF methods between hyper- and multispectral resolution. In the following, we call this data set “field data set”. For a second data set, the “satellite-based data set”, we considered water reflectance measurements from the satellite sensor Medium Resolution Imaging Spectrometer (MERIS), collocated to pigment data from various researchers in the tropical Atlantic Ocean. These data enabled us to study the generic application of the method.

2.1 Field data set

Samples for the field data set were collected during three RV *Polarstern* cruises: the expeditions *ANTXXIV/4* in April/May 2008 and *ANTXXVI/4* in April/May 2010 followed a south-to-north transect through the Atlantic Ocean from Punta Arenas (Chile) to Bremerhaven (Germany); *ANTXXV/1* in November 2008 followed a north-to-south transect through the eastern Atlantic Ocean from Bremerhaven to Cape Town (South Africa) (see Fig. 1; for more details see Table S1, upper panel in the Supplement). Sampling was generally conducted at 12:00 local time and involved conductivity temperature density (CTD) casts with water samplers, below-water radiance and irradiance measurements and above-water irradiance measurements. Water samples from surface water (< 10 m) for pigment analysis and for PE analysis were filtered on GF/F filters and on $0.4 \mu\text{m}$ polycarbonate filters, respectively. Filters were immediately shock-frozen in liquid nitrogen and stored at -80°C until further analysis at the laboratories of the Alfred-Wegener-Institute Helmholtz Centre for Polar and Marine Research (AWI).

2.1.1 Pigment data

The composition of pigments that were soluble in organic solvents was analyzed by HPLC following the method by Barlow et al. (1997) and adjusted to our temperature-controlled instruments (a Waters 600 controller combined with a Waters 2998 photodiode array detector, a Water717plus auto sampler and a LC Microsorb C8 HPLC column) as detailed in Taylor et al. (2011). We determined the list of pigments shown in Table 1 of Taylor et al. (2011) and applied the method by Aiken et al. (2009) for quality control of the pigment data. HPLC data for *ANTXXV/1*, as opposed to the other two cruises, were already published in Taylor et al. (2011) and are available from PANGAEA (doi.pangaea.de/10.1594/PANGAEA.819070). The relative concentration of PE was taken from the data set published

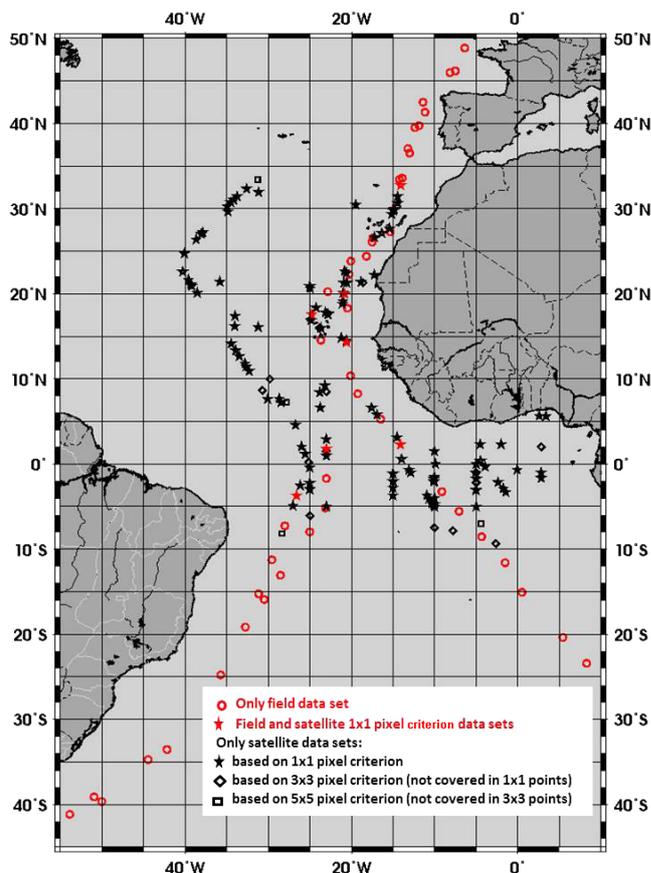


Figure 1. Position of pigment samples used in this study. Red: field data set; black: samples which are collocated to satellite-based but not to field reflectance data; circles: samples which are collocated to field but not to satellite-based reflectance data; stars, diamonds and squares: collocations to MERIS Polymer data based on the 1×1 , 3×3 and 5×5 pixel criteria, respectively.

for all three cruises in PANGAEA (doi.pangaea.de/10.1594/PANGAEA.819624) and analyzed in Taylor et al. (2013). As outlined in Taylor et al. (2013), the PE concentration is expressed as a relative value, while all other pigments concentrations are directly measured values.

2.1.2 Reflectance data field data set

For all three cruises as AOP input data, we used $R_{rs}(\lambda)$ data obtained from profiles of radiance and irradiance from 320 to 950 nm, with an optical resolution of 3.3 nm and a spectral accuracy of 0.3 nm, measured with hyperspectral radiometers (RAMSES, TriOS GmbH, Germany) at the same time and place as pigment data of Sect. 2.1.1. R_{rs} data of *ANTXXV/1* were already published in Taylor et al. (2011) and are available from PANGAEA (doi.pangaea.de/10.1594/PANGAEA.819506). For the other two cruises we applied the same technique and instrumentation as in Taylor et al. (2011) to derive the R_{rs} spectrum at each station. To test the influence of

spectral range and resolution of AOPs, the hyperspectral field $R_{rs}(\lambda)$ data were used within the range of 350 to 700 nm and 380 to 700 nm and reduced to the multispectral bands (412, 443, 490, 510, 560, 620, 665 and 681 nm) of MERIS by taking the integral over all wavebands within one band (± 10 nm around the center wavelength except when $681 \text{ nm} \pm 7.5 \text{ nm}$ was used).

2.2 Satellite-based data set

For this data set, pigment concentrations had been determined from the sea surface (< 10 m) with HPLC by several investigators within the area of 35°N – 10°S and 42°W – 3°E during the MERIS/ENVISAT mission lifetime (2002–2012; for more details on the data set see Supplement Table S1, lower panel). A large part of those data are publicly available from the SEABASS and BODC databases. The remaining pigment data are from the field data set within this area, including additional data from stations where no radiometric measurements had been taken and from four other cruises: pigment data from the RV *Maria S. Merian* cruise *MSM-18/3* were analyzed by AWI as described above in Sect. 2.1.1; data from two RV *Polarstern* cruises (*ANTXXIII/1* and *ANTXXIV/1*) were analyzed by HZG following Zapata et al. (2000); data from the *Bonus Good Hope* (BGH) cruise, conducted by the Laboratoire d’Océanographie de Villefranche, were acquired as outlined in Speich et al. (2008) and analyzed following the method by Ras et al. (2008).

AOP input data is from the MERIS Polymer level 2 $\rho_{wN}(\lambda)$ product given for the same eight wavebands as listed in Sect. 2.1.2. The Polymer algorithm (for details see Steinmetz et al., 2011) provides a powerful atmospheric correction. It is an iterative spectral matching method over the whole available sensor spectrum and uses two decoupled models. First, the water reflectance is modeled using two parameters: the Chl a concentration and the particle backscattering coefficient. Second, the reflectance of the atmosphere, including aerosols and contamination by sun glint, is simplified by using an analytical expression that can account for multiple interactions between molecular and aerosol scatterings (and glitter) without referring to a specific aerosol model. Hence, it allows for the retrieval of large amounts of MERIS observations in sun glint, thin clouds or heavy aerosol plumes; these contaminated conditions could not be treated correctly by standard atmospheric correction schemes extrapolating from the near infrared. MERIS Polymer products thus improve the spatial coverage by almost a factor of 2 and have proven successful for retrieving MERIS Ocean Colour products: Polymer was selected as the MERIS processor for atmospheric correction for the Ocean Colour Climate Change Initiative after an extensive validation and inter-comparison with other atmospheric correction algorithms in which each algorithm’s uncertainty was assessed (Müller and Krasemann, 2012). However, additional uncertainties proba-

bly result from the difference in spatial resolution between satellite (1 km by 1 km) and ship-based (20 cm by 20 cm) sampled data.

Matchups between pigment data and MERIS Polymer $\rho_{wN}(\lambda)$ and TChl a products were determined according to the MERMAID (MERIS MATCHUP In-situ Database) as 1×1 (within the MERIS pixel), 3×3 and 5×5 pixels, respectively, measured on the same day around the field observation (see Barker et al., 2008). For the 3×3 and 5×5 MERIS pixel match-ups, the mean $\rho_{wN}(\lambda)$ and TChl a concentrations from the MERIS products were calculated. Then the 1×1 , mean 3×3 and mean 5×5 MERIS $\rho_{wN}(\lambda)$ matchup data were used for deriving predicted (modeled) pigment concentrations, as outlined in Sect. 2.3. The mean MERIS Polymer TChl a data were validated with the in situ TChl a data of the satellite-based data set. The R^2 , percent bias (PB), mean percent difference (MPD) and root mean square error (RMSE) between the two collocated data sets were calculated as outlined in Werdell et al. (2013) and used to determine pigment prediction full-fit statistics (see Sect. 2.3.2).

2.3 Statistical methods to retrieve pigment concentrations from reflectance

Figure 1 presents the distribution of collocated pigment and reflectance measurements for both field and satellite-based data sets that were used separately as input for the EOF prediction analysis. The field data set covered 53 collocated reflectance and pigment data points (Fig. 1, red points). We used three setups of the field $R_{rs}(\lambda)$ spectra for the development of pigment-specific models:

1. $R_{rs}(\lambda)$ data in hyperspectral (1 nm resolved, “hyper_ R_{rs} ”) from 350 to 700 nm,
2. “hyper_ R_{rs} ” from 380 to 700 nm and
3. $R_{rs}(\lambda)$ data in MERIS band resolution (“band_ R_{rs} ”).

The three satellite-based data sets consisted of 139, 155 and 160 collocated reflectance and pigment data points from 2002 to 2012 for the 1×1 (Fig. 1, stars), 3×3 (Fig. 1, diamonds) and 5×5 (Fig. 1, squares) pixel collocation criteria, respectively, covering all months except January, March and December (details on the spatial and temporal distribution of collocations are given in the Supplement Table S1). Eighteen collocations of the field data matched the 1×1 pixel satellite-based data set (Fig. 1, red stars), but no additional field data matched the two other (3×3 and 5×5 pixel) satellite-based data sets.

The following pigments were included in the construction of all different statistical models (further explained below): MVChl a , DVChl a , chlorophyllide a , MVChl b , DVChl b , Chl $c_{1/2}$ (Chl c type 1 and 2), Chl c_3 (Chl c type 3), Hex (19-hexanoyl-fucoxanthin), But (19-butanoyl-fucoxanthin), Allo (alloxanthin), Diadino (diadinoxanthin), Diato (diatoxanthin), Fuco (fucoxanthin), Lut (lutein), Peri (peridinin),

Viola (violaxanthin) and Zea (zeaxanthin). In addition, the concentrations of different pigment groups were considered, such as the sum concentration of α - and β -carotene (Caro); MVChl *a*, DVChl *a* and chlorophyllide *a* (TChl *a*); MVChl *b* and DVChl *b* (TChl *b*); all pheopigments (TPheo); according to Hooker et al. (2005) and Roy et al. (2011), the photosynthetically active carotenoids (PSC: Fuco, But, Hex, Peri), and the photoprotective carotenoids (PPC: Allo, Diadino, Diato, Zea, Caro). PE, although measured by a different technique, was also included since it also had a major impact on the optical $L_u(\lambda)$ data of the field data set (see Taylor et al., 2013). Neo and PE were only considered in the models using the field data set because, for some pigment samples of the data set collocated to the satellite data, these pigments have not been analyzed.

Figure 2 gives an overview describing the various steps of the development and validation of our EOF method to predict various pigments and pigment groups' concentrations, which are described in detail in the following subsections.

2.3.1 Empirical orthogonal function analysis

Following Taylor et al. (2013), the spectral data were subjected to an EOF analysis, also known as a principal component analysis, in order to reduce the high dimensionality of the data and derive the dominant signals (“modes”) that best describe variance within the data set. In addition to dimension reduction of spectral data, the use of EOF modes in statistical model building also avoids problems associated with multicollinearity amongst the original predictor variables. All calculations in the following were done with the statistical computing software R (R Development Core Team, 2013).

Spectral data were contained in a data matrix \mathbf{X} with dimensions M , sample rows, by N , reflectance band columns. Spectral samples were collocated to the respective pigment data set \mathbf{Y} with dimensions M by P , pigment columns (pigments and pigment groups included are outlined above). While hyper_ R_{rs} data consisted of 350–700 nm ($N = 351$) or 380–700 nm ($N = 321$) bands, band_ R_{rs} and the satellite_ ρ_{wN} data consisted of the eight MERIS visual wavebands ($N = 8$). As in Taylor et al. (2013), spectral data sets \mathbf{X} were standardized for each sample row by first subtracting the mean spectral value (centering) followed by division by the spectral standard deviation (scaling), which focused the analysis on the spectral shape rather than the magnitude. The standardized matrix \mathbf{X} was then subjected to singular value decomposition (SVD) in order to derive EOF modes:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad x_{ij} = \sum_{k=1,N} u_{ik}\sigma_k v_{kj}, \quad (1)$$

where \mathbf{V} is a $N \times N$ matrix containing the EOFs (spectral pattern), \mathbf{U} is an $M \times N$ matrix containing the principal components (PCs), $\mathbf{\Sigma}$ is an $N \times N$ matrix containing the sin-

gular values on the diagonal and k is the EOF mode index (length N). Only EOFs $\leq \min(M, N)$ will carry information. This notation differs slightly from that presented in Taylor et al. (2013), where a covariance matrix of the data set was subjected to Eigen decomposition with subsequent projection of data onto EOFs to derive PCs. The results of both approaches are similar except that \mathbf{U} derived via SVD is unitary, and $\mathbf{\Sigma}$ contains standard deviation rather than variance. The SVD method is presented here due to its more straightforward notation: EOFs and PCs are determined in a single step whereas the alternate Eigen decomposition is a three-step calculation (Fig. 2, the upper part of the panel on the left summarizes these steps).

2.3.2 Log transformed general linear model

A general linear model was used to predict log-transformed pigment concentrations of each pigment, y_p , based on a subset of PCs, \mathbf{U} , as covariates (Fig. 2, the lower part of the panel on the left summarizes these steps). The linear model uses log-transformed pigment concentrations. Since only positive, non-zero values are permissible with this transformation, a small value was added to all concentrations ($0.00001 \text{ mg m}^{-3}$) to allow for the inclusion of samples where pigment concentrations were essentially zero or below the detection limit. A truncated subset of PCs was used as defined by the magnitude of their standard deviation. PCs with standard deviations of ≤ 0.0001 times the standard deviation of the first component were omitted. The resulting multiple regression had the form

$$\log(y_p) = a + b_1u_1 + b_2u_2 + \dots + b_nu_n, \quad (2)$$

where $\log(y_p)$ is the natural log-transformed concentration of pigment p , $u_{1,2,\dots,n}$ are the leading n PC scores from \mathbf{U} , a is the intercept and $b_{1,2,\dots,n}$ are the regression coefficients. A bidirectional stepwise routine was used to search for smaller multiple regression models based on fewer predictor terms. Best linear models were selected through minimization of the Akaike information criterion (AIC). Once the best linear model was determined, the relative importance of included terms was defined by the change in AIC (ΔAIC) following each term's removal.

Since the range of concentration varies greatly among the different pigments, we calculated mainly relative error statistics. According to Werdell et al. (2013), the coefficient of determination (R^2), the RMSE, the slope (S) and the intercept (a) of the linear regression are based on the log-scaled predicted ($\log(y_p)$) as opposed to the log-scaled observed ($\log(y_o)$) pigment concentration data, while the MPD, the PB and the median percent difference (MDPD) are based on the non-log-transformed pigment concentrations. The following

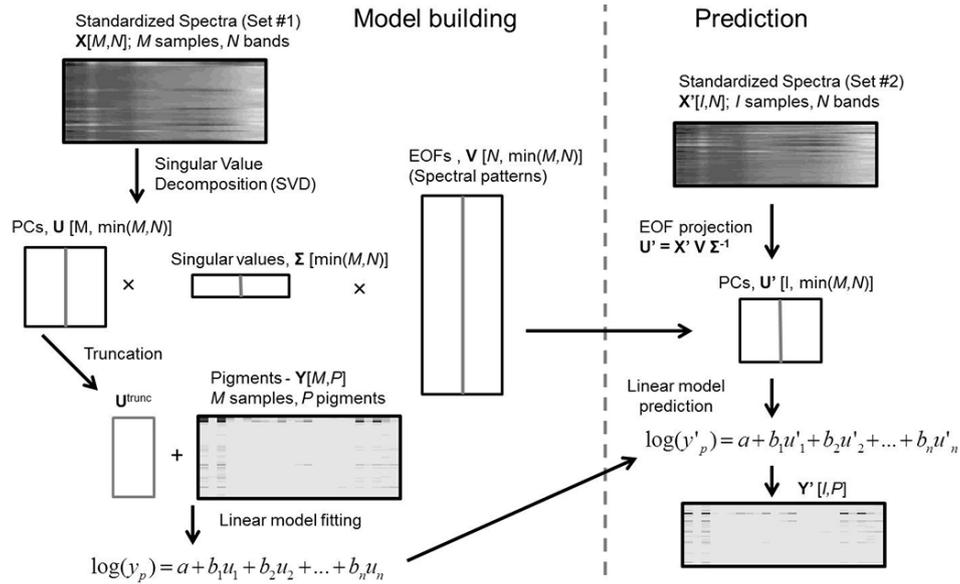


Figure 2. Schematic overview of the steps used in model building and prediction. Multiple linear regression models are fit to log-transformed pigment concentrations, y_p , as the response variable and EOFs derived from a spectral (reflectance) data set, X , as predictor variables. Model building (left) is used for “full-fit” models to all data samples (M) or to a training subset of samples for cross-validation (Sect. 2.3.3). Prediction (right) is used for the assessment of the model error on a validation subset of samples (I) for cross-validation (Sect. 2.3.3) or in the extrapolation of model predictions to a new data set of reflectance spectra, as was done for the larger area of the tropical eastern Atlantic region in this study (Sect. 2.3.4).

equations for these statistics were used:

$$\text{RMSE} = \text{root of } \frac{1}{N} \sum_{i=1}^N (\log(y_{pi}) - \log(y_{oi}))^2, \quad (3)$$

$$\text{MPD} = \frac{100}{N} \sum_{i=1}^N |(y_{pi} - y_{oi})| / (y_{oi}) \quad [\%], \quad (4)$$

$$\text{PB} = \frac{100}{N} \sum_{i=1}^N (y_{pi} - y_{oi}) / (y_{oi}) \quad [\%], \quad (5)$$

$$\text{MDPD} = \text{Median of } [|(y_{pi} - y_{oi})| / (y_{oi}) \times 100] \\ i = 1, N \quad [\%]. \quad (6)$$

2.3.3 Model prediction error

In addition to the statistics performed for each pigment linear model (Sect. 2.3.2), we performed a cross-validation of the linear model fitting in order to better test the robustness of the models’ prediction error. Data were split into two groups: the first part of the data was used for model fitting (Fig. 2, left panel), while the second part was used for prediction validation (Fig. 2, right panel). According to Craig et al. (2012), we assessed the number of observations required to achieve adequate predictions by the pigment linear models using the variable jack-knife procedure of Wu (1986). So the proportion used for data splitting for the cross-validation procedure was varied as follows, where n is the total number of samples, tp is the number of training points and vp is the number

of points used for validation:

$$tp = n \times d, \text{ with } d = 0.1, 0.15, 0.2, \dots, 0.9, \quad (7)$$

$$vp = n(1 - d). \quad (8)$$

Since the number of permutations for data splitting affects the overall computing time, the procedure was run for 500 permutations, similar to the recommendation of Craig et al. (2012). Such a high number of permutations rules out the model error being assessed based on a spatially or temporally biased data set.

Each cross-validation procedure was as follows:

1. For 500 permutations, do steps 2–8.
2. Randomly select $n \times d$ of collocated samples to include in training sets $\mathbf{X}^{\text{train}}$ and $\mathbf{Y}^{\text{train}}$ for spectra and pigment data, respectively. Remaining $n(1 - d)$ of samples are allocated to the validation sets $\mathbf{X}^{\text{valid}}$ and $\mathbf{Y}^{\text{valid}}$.
3. Standardize $\mathbf{X}^{\text{train}}$ and perform EOF following Eq. (1) to obtain $\mathbf{U}^{\text{train}}$, Σ^{train} and $\mathbf{V}^{\text{train}}$.
4. For each pigment concentration y_p^{valid} of $\mathbf{Y}^{\text{valid}}$, do steps 5–9.
5. Fit linear model to log-transformed pigment concentrations using selected $\mathbf{U}^{\text{train}}$ as in Eq. (2):

$$\log(y_p^{\text{train}}) = a + b_1 u_1^{\text{train}} + b_2 u_2^{\text{train}} + \dots + b_n u_n^{\text{train}}. \quad (9)$$

6. Perform bidirectional stepwise search for smaller linear model.
7. Standardize validation set and project $\mathbf{X}^{\text{valid}}$ onto the EOFs $\mathbf{V}^{\text{train}}$ and the inverse of singular values $\Sigma^{\text{train}^{-1}}$ to derive their PCs $\mathbf{U}^{\text{valid}}$:

$$\mathbf{U}^{\text{valid}} = \mathbf{X}^{\text{valid}} \cdot \mathbf{V}^{\text{train}} \cdot \Sigma^{\text{train}^{-1}}. \quad (10)$$

Use selected PCs of $\mathbf{U}^{\text{valid}}$ as variables in Eq. (10) in order to predict pigment concentrations for the validation data set:

$$\log(y_p^{\text{valid}}) = a + b_1 u_1^{\text{valid}} + b_2 u_2^{\text{valid}} + \dots + b_n u_n^{\text{valid}}. \quad (11)$$

8. Record pairs of observed and predicted validation pigment concentrations y_o and y_p^{valid} in a new object for all permutations for later calculation of prediction error.

For each permutation, the R^2 based on the log-scaled predicted ($\log(y_p^{\text{valid}})$) versus the log-scaled measured ($\log(y_o)$) were derived and finally, over all permutations, the mean value (R^2_{cv}) was calculated. In accordance with statistics in Sect. 2.3.2, the prediction error was described in terms of the absolute squared difference based on log-transformed pigment concentrations, $(\log(y_p^{\text{valid}}) - \log(y_o))^2$, and relative difference based on non-log-transformed pigment concentrations, $(y_p^{\text{valid}} - y_o)/y_o$. Mean and median relative difference (MPD_{cv} and MDPD_{cv}, respectively) and the root mean square absolute difference (RMSE_{cv}) over all permutation were determined as follows:

$$\text{MPD}_{\text{cv}} = \frac{100}{N} \sum_{i=1}^N \left[(y_p^{\text{valid}} - y_o) / y_o \right] \quad [\%], \quad (12)$$

$$\text{RMSE}_{\text{cv}} = \sqrt{\frac{1}{N} \sum_{i=1}^N \left[\log(y_p^{\text{valid}}) - \log(y_o) \right]^2}, \quad (13)$$

$$\text{MDPD}_{\text{cv}} = \text{median value of} \left(\frac{y_{i,p}^{\text{pred}} - y_{i,p}^{\text{valid}}}{y_{i,p}^{\text{valid}}} \times 100 \right), \quad (14)$$

$i = 1, N[\%].$

2.3.4 Pigment concentration predictions with MERIS reflectance data

In order to predict pigment concentration from MERIS $\rho_{\text{wN}}(\lambda)$ for a whole month of data in November 2008, for which we did not have corresponding pigment measurements, the following method was applied: we projected standardized MERIS $\rho_{\text{wN}}(\lambda)$ data onto the EOF loading (\mathbf{V}) to derive their principal components (\mathbf{U}), which were subsequently used for the prediction with the fitted linear model (as in Sect. 2.3.3, step 7, Eq. 11, Fig. 2, right panel), where $b_{1,2,\dots,n}$ are taken from the EOF model developed with the 1×1 MERIS Polymer $\rho_{\text{wN}}(\lambda)$ matchups (following Fig. 2, left panel).

3 Results and discussion

3.1 Characteristics of input data sets

Figure 3 shows the original and standardized spectra of the field and satellite-based data sets. Considering the conversion of $R_{\text{rs}}(\lambda)$ to $\rho_{\text{wN}}(\lambda)$ data by a factor of π , the magnitude and shape of the original and standardized spectra are similar for the band-resolved data sets, except that the standardized satellite- ρ_{wN} data set contains only one spectrum with maximum reflectance in the green at 560 nm, while the standardized field data set contains four spectra with maxima at 510 nm.

The composition and range of pigments (as detailed with maximum, minimum, mean and standard deviation in Supplement Table S2) show, for all pigments, that the collocations to the field data set contain higher maxima and minima than the collocations to the satellite-based data set (except for Fuco, for which it is equal and for Zea, for which it is inverted). For most pigments, mean values are very similar for both data sets. However, standard deviations for the field data set are 2 to 3 times higher than the mean for all pigments. In the satellite data set, the standard deviation is of a similar magnitude to the mean value. The higher concentration of total pigments in the field data set may explain the small differences in the shape of the reflectance spectra of the two (field versus satellite-based) data sets. However, DVChl *b*, MVChl *b*, TChl *b*, Allo, Diato, Lut, Neo, Peri, Viola and TPheo had values of 0 mg m^{-3} in more than 20 % of all stations in both data sets. Also, Chl *c*₃ had a concentration of 0 mg m^{-3} in one sample collocated to the field and in over 30 % of samples collocated to the satellite-based data set. Several pigments had concentrations of 0 mg m^{-3} only occasionally ($< 10 \%$) in samples collocated to the satellite-based data set (Caro, Chl *c*_{1/2}, But, Hex, Zea, DVChl *a*, Diadino and Fuco) and in the field data sets (DVChl *a*, Diadino and Fuco). All other pigments not listed here had detectable concentrations in all samples.

3.2 EOF analysis – shape of modes and relevance for predictions

Following the EOF truncation criteria outlined in Sect. 2.3.2, the decomposition of the standardized spectra resulted in nine modes (EOF-1 to EOF-9) for the hyper- R_{rs} and seven modes for the band- R_{rs} and satellite- ρ_{wN} data sets (the first four modes are presented in Fig. 4). EOF modes for the three satellite- ρ_{wN} data sets were nearly identical. For simplicity we only show (Fig. 4) and discuss the EOF modes of the 1×1 pixel collocation data set. For all data sets, the first three modes explain over 99.8 % of the variance for all three data sets, with EOF-1 explaining between 94.5 % and 96 % of the variance (Table 1).

The shapes of the first three EOF modes are very similar among all three reflectance data sets. They are nearly iden-

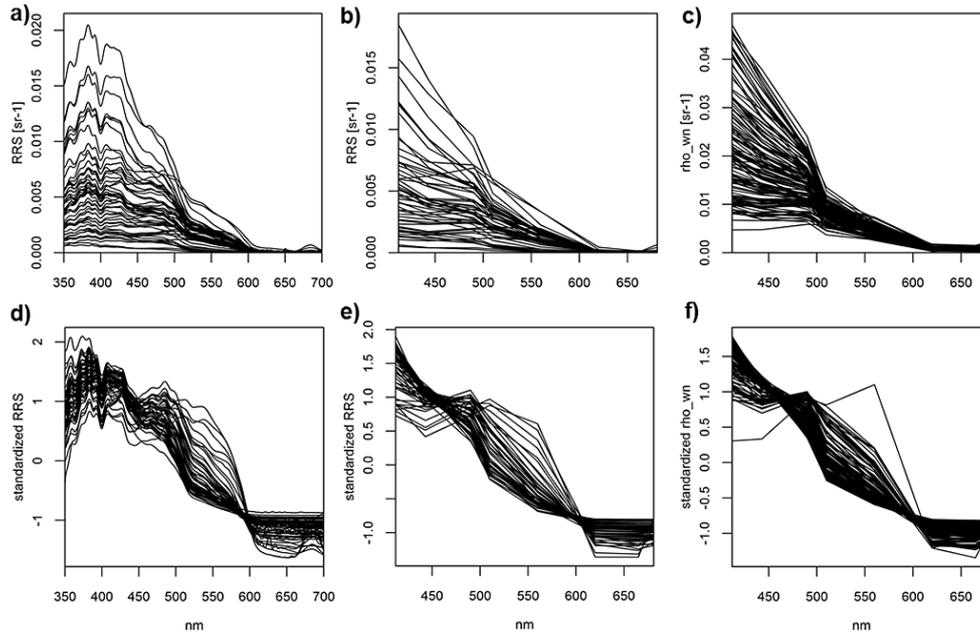


Figure 3. (a)–(c) Original (in sr^{-1}) and (d)–(f) standardized (subtracted mean and divided by standard deviation) reflectance spectra of hyper_ R_{RS} (a) and (d), band_ R_{RS} (b) and (e) and satellite_ ρ_{wN} (from MERIS Polymer) data within the 1×1 pixel collocation boxes (c) and (f).

Table 1. Percent of total variance explained (Expl. variation; upper panel) and cumulative proportion (Cum. proport.; lower panel) by the significant EOFs derived from field R_{RS} spectra in hyperspectral resolution (hyper_ R_{RS}) and multispectral resolution (band_ R_{RS}) and from satellite_ ρ_{wN} (from MERIS Polymer) using the 1×1 pixel collocation criterion.

% Expl. variation	EOF-1	EOF-2	EOF-3	EOF-4	EOF-5	EOF-6	EOF-7	EOF-8	EOF-9
hyper_ R_{RS}	95.0	4.1	0.7	0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
band_ R_{RS}	94.5	5.0	0.4	< 0.1	< 0.1	< 0.1	< 0.1		
satellite_ ρ_{wN}	95.9	3.9	0.1	< 0.1	< 0.1	< 0.1	< 0.1		
% Cum. proport.	EOF-1	EOF-2	EOF-3	EOF-4	EOF-5	EOF-6	EOF-7	EOF-8	EOF-9
hyper_ R_{RS}	95.0	99.1	99.8	99.9	99.965	99.981	99.990	99.994	99.996
band_ R_{RS}	94.5	99.5	99.9	99.95	99.990	99.999	100.0		
satellite_ ρ_{wN}	95.9	99.7	99.9	99.95	99.985	99.996	100.0		

tical for the band_ R_{RS} and the satellite_ ρ_{wN} data sets but show smoother shapes and peaks for hyper_ R_{RS} for the first two modes. Due to the limited number of wavelengths for the two multispectral data sets, EOFs show evidence of a shift in peak location, starting with EOF-3 (peak at 412 and 443 nm for EOF-3 and EOF-4, respectively), as compared to hyper_ R_{RS} (peak at 360 and 410 nm for EOF-3 and EOF-4, respectively). This is likely due to the increased spectral resolution of the hyperspectral data, which allows for more precision in identifying spectral regions of higher variance. For EOF-4, the satellite_ ρ_{wN} mode is much flatter beyond 500 nm and shows no trough between 600 and 650 nm compared to the EOF-4 for the other two data sets. Not much similarity is seen among the EOF-5 modes of the different spectra data sets; for EOF-6, the two field data sets are simi-

lar in the overall shape, but peak locations are shifted towards longer wavelengths for the satellite data set. EOF-7 and EOF-8 show very similar shapes for hyper_ R_{RS} and deviate from EOF-7 in the band data sets, while EOF-9 from hyper_ R_{RS} looks much more like the later ones.

The EOF analyses identify dominant modes of variance, which can be interpreted as imprints of changes in the optical properties of water constituents in the water column. For this study, only reflectance spectra taken in high TChl a waters with measurable mineral fraction (identified as cluster V for the ANTXXV/1 data in Taylor et al., 2011) show any resemblance to spectral shapes obtained in the case 2 waters of Lubac and Loisel (2007, e.g., class 5) and Craig et al. (2012). The remaining spectra (typical case 1 water) show characteristics not observed in those studies. This difference explains

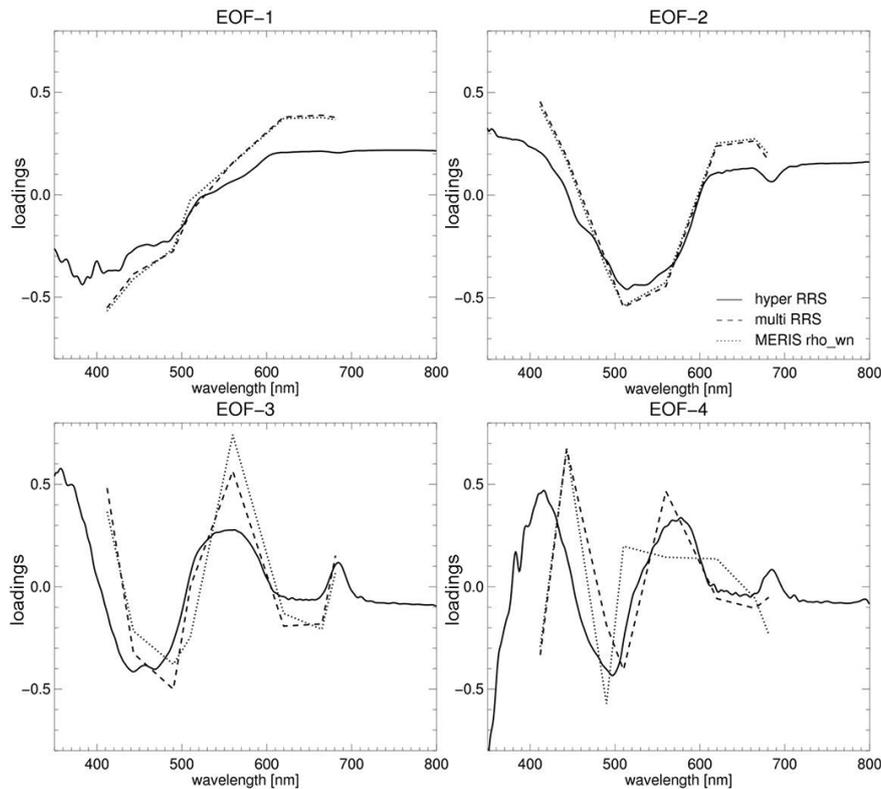


Figure 4. First four EOF modes (EOF-1 to EOF-4) derived from field R_{rs} data set in hyperspectral resolution (hyper- R_{rs} , solid lines) and in multispectral band resolution (band- R_{rs} , dashed lines) and from using satellite- ρ_{wN} (from MERIS Polymer, dotted line) data within the 1×1 pixel collocation box.

the minor variations in the shape and loading of EOFs between their and our data sets. In the following, we focus the discussion on our hyper- R_{rs} data set results with specific comparison to the study by Craig et al. (2012), which was also based on hyperspectral R_{rs} data.

Our first three EOF modes correspond to the ones derived for the hyperspectral case 2 reflectance data set of Craig et al. (2012). As pointed out in their study, EOF-1 is likely the signature of bulk oscillations in phytoplankton biomass concentration (including its effect on backscattering). However, our EOF-1 already explains much more of the variance than in Craig et al. (2012), where it only accounted for 72.4% and showed much more structure and a weaker exponential decrease from 400 to 550 nm. EOF-2 superficially resembles the overall changes in the total absorption over broad band structures. It strongly decreases from 350 to 510 nm and increases again above 570 nm, which is connected to total pigment and water absorption, respectively. There is a peak around 683 nm which can be linked to MVChl a and DVChl a fluorescence. While this peak is present in EOF-1 in the Craig et al. (2012) data set, it is not in the EOF-1 of our data set likely because of the lower TChl a concentrations.

EOF-3 of our data set as compared to the one of Craig et al. (2012) shows a much steeper decrease with wavelength

in the blue spectral range. These changes may reflect concomitant changes of absorption by chlorophyll, colored dissolved organic matter and non-algal particles expected to be co-varying and of much lower concentration in our case 1 waters. Scattering by particles other than phytoplankton was much higher in the case 2 water of Craig et al. (2012), leading to a less steep slope of this EOF mode. EOF-4 appears different in relation to the three peaks. Similar to EOF-2 and EOF-3, these differences are caused by the different composition and overall loading of water constituents of our and their sampled stations.

In summary, in contrast to more coastal waters where measurable mineral fraction can affect R_{rs} properties, the total attenuation is much more affected by total pigment concentration in our open-ocean, case 1 data set. Our data set was largely composed of samples from waters with lower TChl a concentration, ranging from 0.005 to 3.553 mg m⁻³, while in the study of Craig et al. (2012) it ranged from 0.584 to 18.02 mg m⁻³. EOFs greater than 4 were not presented in Craig et al. (2012) because they were not used to predict TChl a from R_{rs} data, as was the case for our TChl a (and MVChl a) linear model predictions (Sect. 3.3.3). Higher EOF modes probably reflect the influence of specific pigment

groups or pigments, as indicated by the results of the ΔAIC values and further discussed in Sect. 3.3.3.

3.3 Pigment prediction by linear models

3.3.1 Field data set linear models

All pigments that were detected in the full set of the field data samples were well predicted by linear models based on hyperspectral ($\text{hyper}_{R_{rs}}$) or the reduced eight-band ($\text{band}_{R_{rs}}$) resolution spectra. The correlations between predicted and observed concentrations for these pigments were highly significant ($p < 0.0001$) and cross-validation statistics reached reasonable quality with $R^2_{cv} \geq 0.5$, $\text{MDPD}_{cv} \leq 45\%$ and $\text{MPD}_{cv} \leq 60\%$ (Table 2a, upper part). For some pigments (TChl *a*, PSC, MVChl *a*, Hex, Caro) EOFs based on 380 to 700 nm produced much better linear model results using $\text{hyper}_{R_{rs}}$ data than based on 350 to 700 nm (for all statistical parameters see Supplement Table S3; models based on $\text{hyper}_{R_{rs}}$ (a) at 350 to 700 nm and (b) at 380 to 700 nm and (c) on $\text{band}_{R_{rs}}$). Lower quality for one statistical parameter for both linear models was reached for Zea (R^2_{cv} 0.35 and 0.28), But (MPD_{cv} 81 and 95 %) and for two parameters for PE (MDPD_{cv} 65 and 67 %, MPD_{cv} 139 and 156 %).

Plots of observed versus predicted values for the full data set of well-predicted pigments TChl *a*, PSC, PPC, Hex and Zea are shown in Fig. 5. For pigment groups and pigments with a high range of data (TChl *a*, PSC and Hex), covering about 3 orders of magnitude, the intercept is much lower and the regression closely aligns with the 1 : 1 reference line. The predicted versus observed regression for Zea was of lower quality ($R^2 < 0.6$) likely due to a much lower range of observed concentrations.

For all other pigments, predictions were of low quality (results not shown), demonstrating that the linear model approach does not produce robust predictions for situations where pigment were not detected (i.e., 0 mg m^{-3}) in every sample (see results for all pigment predictions in Supplement Table S3). Even pigments that were only occasionally undetected (e.g., DVChl *a*, TChl *b*, MVChl *b*) showed increased error in cross-validation prediction as revealed by MDPD_{cv} and RMSE_{cv} values far above 100 % and 1, respectively. We re-ran the predictions for specific pigments where only a few samples ($< 10\%$) had concentrations of 0 mg m^{-3} , as was the case for DVChl *a*, Fuco, Diadino and Chl *c*₃ (see Supplement Table S2). In those specific linear model runs we only included as input data the data points where the specific pigment concentrations were $> 0 \text{ mg m}^{-3}$. The resulting predictions (Table 2a, lower part; for DVChl *a* see full-fit results in Fig. 5d) from using the adjusted input data for those pigments show robust and significant cross-validation results within the same quality range as for the pigments which were detected in all data. For other pigments, where non-detection occurred more frequently ($> 20\%$ of the samples), the re-

moval of non-detection samples did not result in robust predictions (results not shown).

Cross-validation results of well-predicted pigments (Table 2a) show that, especially regarding the R^2_{cv} and RMSE_{cv} values, $\text{hyper}_{R_{rs}}$ -based linear models perform either the same (PSC), slightly better (PPC, Chl *c*_{1/2}) or much better (TChl *a*, MVChl *a*, But, Hex, Zea, Caro, PE, DVChl *a*, Chl *c*₃, Diadino, Fuco) than predictions based on eight wavelengths ($\text{band}_{R_{rs}}$ data set). In particular, RMSE_{cv} is much improved for several pigment predictions where RMSE_{cv} reaches high values ($> 0.65 \text{ mg m}^{-3}$), i.e., for PE, Fuco, But, Chl *c*₃, Diadino and Hex. The benefit was less clear when observing the statistics of MDPD_{cv} and MPD_{cv} in several pigments (MVChl *a*, Chl *c*_{1/2}, TChl *a* and PSC predictions). For these pigments the multispectral resolution appears to be sufficient for obtaining similarly robust predictions. TChl *a* (in line with MVChl *a*) and PSC dominate the overall phytoplankton pigment composition and absorption. TChl *a* concentrations have been well retrieved by band-ratio algorithms as a main phytoplankton biomass indicator (e.g., see Brewin et al., 2014). For pigments very similar in spectral range, such as But, Hex and Fuco, the hyperspectral resolution of the linear models provides much more robust pigment predictions (Table 2a). The $\text{hyper}_{R_{rs}}$ linear models also produced better predictions for DVChl *a*, Zea, Diadino and PPC, where the specific linear models included a much larger set of EOF modes (see Sect. 3.3.3) which may indicate the importance of higher-resolution spectral details not available in the $\text{band}_{R_{rs}}$ data.

3.3.2 Satellite-based data set linear models

Results for the models predicting pigment concentration from the satellite-based data set were very similar when using 1×1 , 3×3 or 5×5 collocated MERIS ρ_{wN} data (for all statistical parameters see Supplement Table S3: $\text{satellite}_{\rho_{wN}}$ models based on 1×1 (d), 3×3 (e) and 5×5 (f) collocations). Deviations were within 1 to 3 % for all statistical parameters. R^2_{cv} values were best in all cases for well-predicted pigment concentrations in the 1×1 collocations, while MPD_{cv} was best in the 3×3 collocations. Results clearly show that even models based on 5×5 pixel collocations can produce robust results. For simplicity, in the following we present and discuss the results of the 1×1 collocated reflectance data only.

In line with field data linear model results, pigment groups and pigments, which were detected in every sample (MVChl *a*, TChl *a*, PSC and PPC; the full-fit linear model results are shown in Fig. 6a–c), are well predicted with similar cross-validation statistic values using the $\text{satellite}_{\rho_{wN}}$ data set (Table 2b, upper part). Also, good predictions for some pigments (DVChl *a*, Zea, Diadino, Hex, But, Fuco and Chl *c*_{1/2}) could be obtained by re-running the linear model analysis with concentrations of 0 mg m^{-3} excluded (Table 2b, lower part). For example, the full-fit linear model

Table 2. Statistics of linear models using EOF modes based on a) field R_{rs} data in hyperspectral (hyper; normally 350–700 nm; when * then 380–700 nm) resolution and multispectral (band) resolution and (b) the satellite- ρ_{wN} (from MERIS Polymer) using the 1×1 pixel collocation criterion data set. Cross-validation results are presented with 500 permutations for data splitting into 80 % of the data used for training and 20 % for validation. Only well-retrieved pigment prediction results, with correlations being highly significant at $p < 0.0001$, are given. Abbreviations of pigments are explained in Sect. 2.3.1. Pigments listed in the upper part of each table show high-quality results using the entire data set. In the lower part of each table (listed under “ $> 0 \text{ mg m}^{-3}$ ”) models are based only on the data set of collocated R_{rs} samples where the respective pigment reached concentrations above 0 mg m^{-3} . Bold: here band-model performs better than hyper-model. Red signifies only medium quality as specified in the text.

(a)	Band/hyper	MDPDcv		MPDcv		RMSEcv		R^2_{cv}		
		<i>N</i>	band	hyper	band	hyper	band	hyper	band	hyper
	TChl <i>a</i> *	53	28	32	42	43	0.54	0.49	0.72	0.77
	PSC*	53	32	34	51	53	0.62	0.62	0.75	0.75
	PPC	53	28	27	51	49	0.57	0.56	0.52	0.52
	MVChl <i>a</i> *	53	31	34	45	44	0.54	0.50	0.77	0.79
	Chl <i>c</i> _{1/2} *	53	34	36	48	49	0.57	0.56	0.79	0.80
	But	53	44	43	95	81	0.90	0.82	0.50	0.55
	Hex*	53	37	36	57	53	0.69	0.60	0.64	0.70
	Zea	53	31	28	45	42	0.52	0.48	0.28	0.35
	Caro*	53	34	33	55	50	0.62	0.54	0.55	0.62
	PE*	53	67	65	156	139	1.28	1.16	0.65	0.69
$> 0 \text{ mg m}^{-3}$:										
	DVChl <i>a</i>	49	26	24	45	39	0.49	0.44	0.59	0.67
	Chl <i>c</i> ₃ *	52	40	37	61	58	0.86	0.69	0.71	0.73
	Fuco*	52	45	40	92	73	1.02	0.82	0.65	0.66
	Diadino	52	37	33	59	53	0.71	0.61	0.57	0.67

(b) satellite- ρ_{wN}	<i>N</i>	MDPDcv	MPDcv	RMSEcv	R^2_{cv}
TChl <i>a</i>	139	32	47	0.64	0.67
PSC	139	42	59	0.79	0.70
PPC	139	27	40	0.49	0.50
MVChl <i>a</i>	139	40	55	0.67	0.72
$> 0 \text{ mg m}^{-3}$:					
DVChl <i>a</i>	124	40	59	0.73	0.25
Chl <i>c</i> _{1/2}	127	41	89	0.84	0.64
X19_But	126	39	55	0.63	0.74
X19_Hex	132	40	60	0.68	0.68
Fuco	128	40	65	0.85	0.71
Zea	138	29	46	0.51	0.40
Diadino	132	45	91	0.84	0.57

results for DVChl *a*, Hex and Zea are shown in Fig. 6d–f. Nevertheless, some of these pigments show only medium quality for one cross-validation statistical parameter (lower R^2_{cv} for DVChl *a* and Zea, higher MPDcv for Fuco, Chl *c*_{1/2} and Diadino).

The full-fit results shown in Fig. 6 show that the models based on the satellite data show much poorer predictions (e.g., *a*, R^2 and RMSE) than the field data models for all pigment or pigment groups (except Zea) even though the satellite data models are based on more samples. This may be

caused by the lower quality of water-leaving reflectance data obtained from the satellite as opposed to direct radiometric measurements in the water column. Another explanation may be that the lower standard deviation of the pigments in the satellite-based data set leads to less precision of the EOF-based models. The latter may explain why the full-fit results for predicting Zea concentrations are very similar for the two model types.

Similar to the field data linear models, no robust predictions were obtained for all other pigments that reached

Table 3. Δ AIC for the robust pigment predictions of the pigment groups TChl *a*, PSC and PPC and the pigments MVChl *a*, Zea and DVChl *a* by the EOF models based on field R_{rs} in (a) hyperspectral resolution (hyper_ R_{rs}) and (b) multispectral resolution (band_ R_{rs}) and (c) the satellite_ ρ_{wN} (from MERIS Polymer) using the 1×1 pixel collocation criterion. The pigments listed under “no 0 mg m^{-3} ” were predicted using a reduced data set where the respective pigment reached concentrations above 0 mg m^{-3} . Bold highlights the EOF mode with the highest Δ AIC.

(a) hyper_ R_{rs}	EOF-1	EOF-2	EOF-3	EOF-4	EOF-5	EOF-6	EOF-7	EOF-8	EOF-9
TChl <i>a</i>		80	16	16					
PSC		78	9	8	4				
PPC		35	33		0.4		3	2	
MVChl <i>a</i>		88	16	11					
Zea	5	6	21		10	7	1	5	9
no 0 mg m^{-3} :									
DVChl <i>a</i>	7	10	21	21	1	20	9	4	7
(b) band_ R_{rs}	EOF-1	EOF-2	EOF-3	EOF-4	EOF-5	EOF-6	EOF-7		
TChl <i>a</i>		86	10	15	3	1			
PSC		80	5	5	6				
PPC		38	20	9		1			
MVChl <i>a</i>		90	10	9	1				
Zea	2	5	6	10	4	7	0.3		
no 0 mg m^{-3} :									
DVChl <i>a</i>	12	17		20	23				
(c) satellite_ ρ_{wN}	EOF-1	EOF-2	EOF-3	EOF-4	EOF-5	EOF-6	EOF-7		
TChl <i>a</i>	7	148				1			
PSC	5	146		0.2		16			
PPC	10	62		3		3	4		
MVChl <i>a</i>	4	151				5			
no 0 mg m^{-3} :									
Zea	38	63	17		1	11	6		
DVChl <i>a</i>	13		26	35	2				

$> 0 \text{ mg m}^{-3}$ in less than 80 % of all samples concentrations, even when only data points with specific pigment concentrations $> 0 \text{ mg m}^{-3}$ were included (results not shown).

3.3.3 EOF modes relevant for pigment predictions

Table 3 presents the results of EOF significance based on Δ AIC from their removal as model terms. For the hyper_ R_{rs} data set, the prediction linear models used EOF-2 and EOF-3 for all pigments. EOF-2 was the most relevant in the respective models for all pigment prediction except for Zea and DVChl *a*, for which EOF-3 was the most important, closely followed by several other EOF modes. For all other well-predicted pigments, EOF-3 followed EOF-2 in importance, except for Chl *c*₃ (EOF-4) and PE (EOF-1). Besides PE, only EOF-1 was included (with medium importance) for the prediction of But, DVChl *a* and Zea concentrations. Nearly all linear models using the hyper_ R_{rs} data set to predict pigment concentrations incorporated the loadings of three to five EOF

modes. In contrast, predictive models for DVChl *a*, Zea and PPC incorporated nine, eight and six EOF modes, respectively.

As discussed in Sect. 3.2, EOF-2 reflects the optical imprint of all phytoplankton pigments. The high Δ AIC value of EOF-2 for most pigments' linear models is probably caused by the increase in concentration of these specific pigments and most phytoplankton groups when TChl *a* increases. In contrast to that, cyanobacteria and especially its subgroup *Prochlorococcus*, containing the marker pigments Zea and DVChl *a*, respectively, are the most abundant phytoplankton under low TChl *a* concentrations. This has manifested in the abundance-based algorithms to retrieve picoplankton from TChl *a* data (Uitz et al., 2006; Hirata et al., 2011) and may explain why predictions of those marker pigments by our linear models show lower Δ AIC for EOF-2 and require several different EOF modes in their linear models.

As in Craig et al. (2012), EOF-2 to EOF-4 were relevant for our hyper_ R_{rs} -based TChl *a* and MVChl *a* predictions.

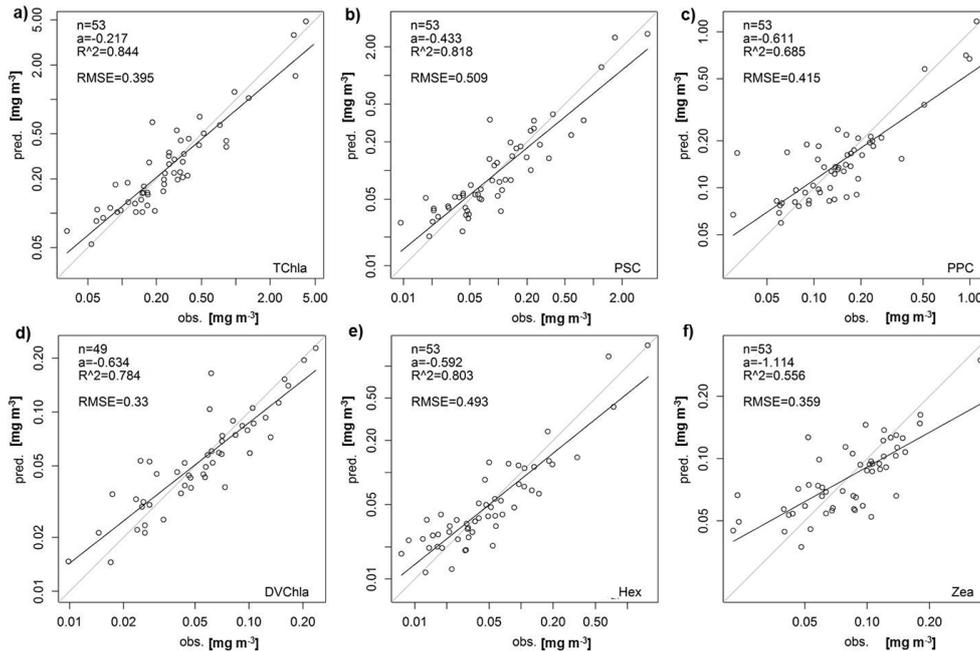


Figure 5. Examples of regressions between observed (obs.) and predicted (pred.) concentrations for pigment groups, (a) TChl *a*, (b) PSC and (c) PPC, and specific pigments, (d) DVChl *a*, (e) Hex and (f) Zea. Observed values have been measured by HPLC (obs.), while predictions are made using a linear model based on EOF modes derived from field R_{rs} data in hyperspectral resolution (hyper- R_{rs}). For DVChl *a*, the model data set was reduced by excluding collocated samples where DVChl *a* had concentrations of 0 mg m^{-3} .

EOF models developed by Taylor et al. (2013) to predict PE concentrations based on L_u data required the first four EOF modes, while our PE prediction based on R_{rs} data required the first three EOFs only. For all other pigments, the higher EOFs were also necessary for robust predictions.

Similarly to the hyper- R_{rs} linear models, the two multispectral linear models also showed EOF-2 to be the most important predictor for specific pigment models except for DVChl *a* (both models) and Zea (only band- R_{rs}).

3.4 Number of data points to construct robust models

Our presented linear models to predict specific pigment or pigment group concentration are calibrated for an ocean-color data set of a specific region with coincidental pigment measurements. Results of the variable jack-knife procedure indicate that the minimal number of training points needed to set up a robust linear model varies among pigments and pigment groups, as revealed by several statistical error measures: the ratio of R^2_{cv} to R^2 (R^2_{cv}/R^2), the ratio of MPD_{cv} to MPD (MPD_{cv}/MPD) and the ratio of RMSE to RMSE_{cv}. Examples for predicting TChl *a*, PSC, PPC and PE are shown in Fig. 7. The ratio R^2_{cv}/R^2 for PPC in all linear models (Fig. 7a, d) drops below 0.8 after a threshold of 50 training data and then decreases exponentially with diminishing data, while other pigments can maintain a high ratio with as few as 30 samples and even 15 samples in the case of the hyper- R_{rs} PE linear model. The threshold where

the slope increases in RMSE_{cv}/RMSE (Fig. 7c, f) is for all pigments and linear models probably around 20 to 30 training points. MPD_{cv}/MPD ratios below 1.4, which would indicate robust fits, are obtained for all pigments above 50 training points for the satellite- ρ_{wN} (Fig. 7e) and above 30 for the hyper- R_{rs} data sets (Fig. 7b). Generally, we observe that band- R_{rs} -based models are more sensitive to training sample size as compared to the hyper- R_{rs} -based models, especially for TChl *a* and PE. As a general recommendation, a requirement of at least 45 to 50 training data points is advised for most cases, while some pigments (e.g., TChl *a*) may be well predicted with as few as 25 training samples when using models based on hyper- R_{rs} data. Based on these results, we are confident that the models presented in Sect. 3.3 are able to provide robust predictions for both field and the satellite-based data. In the case of PE, the number of samples seems to have been too small, especially for the multispectral resolution, to provide robust PE predictions.

3.5 Comparison to other approaches deriving pigment concentration

Our hyper- R_{rs} TChl *a* linear model results ($R^2 = 0.84$, $RMSE = 0.4$, $R^2_{cv} = 0.77$, $RMSE_{cv} = 0.49$; Fig. 5 and Table 2) are comparable to results by Craig et al. (2012; $R^2 = 0.84$, $RMSE = 0.3$, $R^2_{cv} = 0.76$, $RMSE_{cv} = 0.21$). Craig et al. (2012) used measurements only from a single location, sampled about weekly over the course of 1 year, while

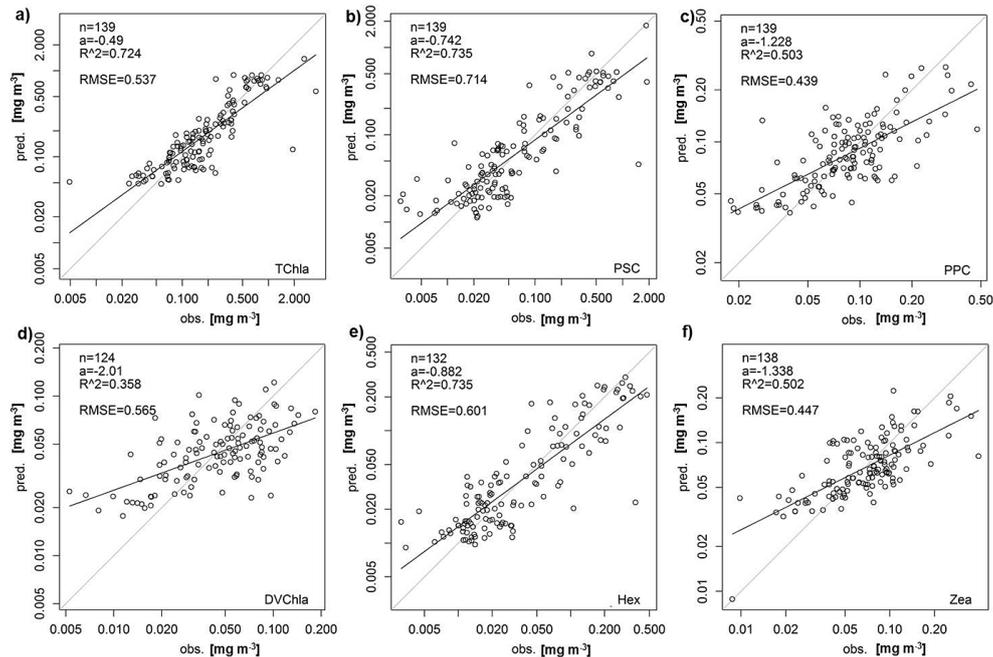


Figure 6. Examples of regressions between observed (obs.) and predicted (pred.) concentrations for pigment groups, (a) TChl *a*, (b) PSC and (c) PPC, and specific pigments, (d) DVChl *a*, (e) Hex and (f) Zea: observed values have been measured by HPLC (obs.), while predictions are made using a linear model based on EOF modes derived from satellite ρ_{wN} (from MERIS Polymer) data within the 1×1 pixel collocation box. For DVChl *a*, Hex and Zea, the model data set was reduced by excluding collocated samples where DVChl *a* had concentrations of 0 mg m^{-3} .

our field data set was from a much larger region (covering 95° in latitude and 85° in longitude) and sampled during two seasons in 2008 and 2010 only. In their study the same linear model setup was used with collocated in situ reflectance and TChl *a* data sampled at Compass Buoy station in the Bedford Basin near Halifax as input data. The number of collocations used for training to obtain robust results for TChl *a* predictions was also similar for both studies, with more than 25 recommended for our hyper- R_{RS} linear model and more than 15 in the Craig et al. (2012) study.

Chase et al. (2013) used Gaussian functions to derive different chlorophyll types, PSC and PPC concentrations from a large global data set of hyperspectral particulate absorption measurements. Their validation results showed MDPD values between predicted and observed concentrations of 30 and 36 %, 40 and 53 %, 49 % and 51 % for TChl *a*, TChl *c*, PSC and PPC, respectively. Our linear models show similar (TChl *a* 27–32 %) or even much better MDPDcv values (Chl $c_{1/2}$: 34–41 %, PSC: 32–43 %, PPC: 24–28 %). We believe that this further indicates the robustness of our approach, especially given that we use a more indirect measure of pigments, AOP (reflectance), as opposed to the IOPs used in their study.

Pan et al. (2010) developed pigment specific band-ratio algorithms with collocated in situ $R_{RS}(\lambda)$ and pigment measurements from the northeastern coast of the United States.

Those algorithms are based on deriving pigment-specific coefficients for third-order polynomial functions using the band ratio of either 490–550 nm or 490–670 nm (for SeaWiFS; for MODIS changed accordingly to MODIS bands 488 and 547 nm). Validation of results with collocated satellite (SeaWiFS and MODIS) reflectance data and pigment concentrations showed very good-quality predictions for several pigments (TChl *a*, TChl *c*, Caro, Fuco, Diadino and Zea) using SeaWiFS bands (MPD from 36 to 48 %, RMSE from 0.23 to 0.29, and R^2 from 0.65 to 0.90; similar results were also obtained using MODIS bands). This method was modified to the northern South China Sea using globally derived relationships and locally identified links between pigment concentration and sea surface temperature (Pan et al., 2013) with similar validation results as in Pan et al. (2010). Compared to our linear model results, the quality of pigment concentration prediction is similar: while our results for MPDcv and R^2_{cv} are slightly worse (42–50 % and 0.61–0.80, respectively), our results for RMSEcv ($0.48\text{--}0.61 \text{ mg m}^{-3}$, except Fuco: 0.82 mg m^{-3}) are much better.

PE is not well predicted by both our linear models based on the field data set. Still, hyper- R_{RS} linear model cross-validation measures are much better than the PE band- R_{RS} linear model. In Taylor et al. (2013), PE concentrations were predicted from the same underwater light measurements but using L_u instead of R_{RS} data and the model was based on pig-

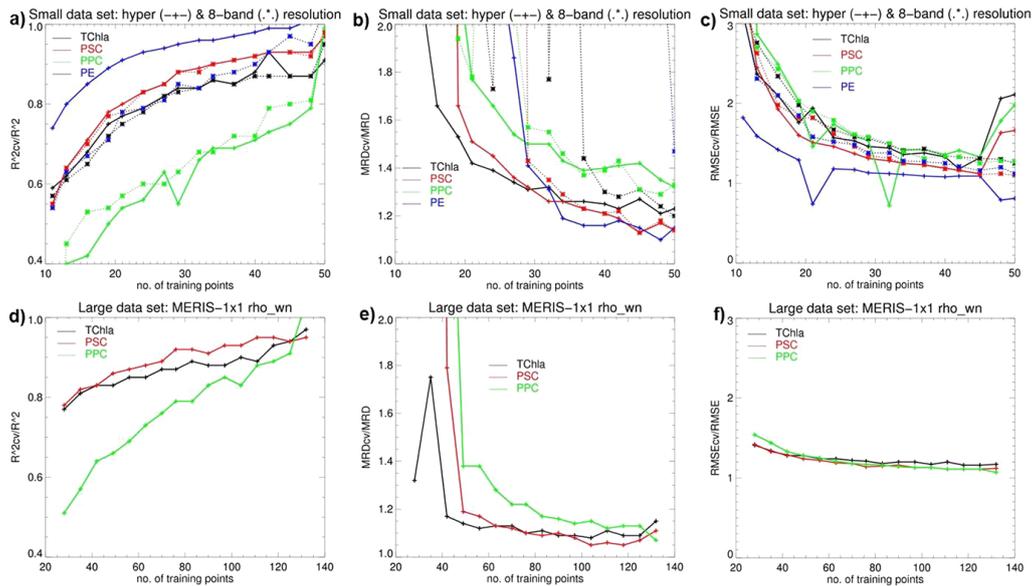


Figure 7. R^2_{cv}/R^2 (a, d), MRD_{cv}/MRD (b, e) and $RMSE_{cv}/RMSE$ (c, f) as a function of number of training points (tp) for the linear models. Shown are results for specific models for TChl *a*, PSC, PPC and PE using reflectance data from the field (a–c) in hyperspectral (hyper_ R_{RS} , solid lines) and multispectral band (band_ R_{RS} , dotted lines) resolution and from satellite MERIS Polymer within the 1×1 pixel collocation box (satellite_ ρ_{wN} , (d–f)). The number of total samples points was $n = 53$ for hyper_ R_{RS} and band_ R_{RS} and $n = 139$ for satellite_ ρ_{wN} . Cross validation is based on 500 permutations using tp for training and as number of validation points (vp): $vp = n - tp$.

ment concentrations at surface and deeper depths. No cross-validation was performed within their study. Our results for R^2_{cv} (0.69) are even better than their results for using the data from all three cruises for predictions (R^2 of 0.58). The data set of Taylor et al. (2013) was nearly 3 times larger than our field data set and a log-link generalized linear model (GLM) was used instead of a log-transformed linear model. For the latter we tested both settings for our pigment linear models. Cross-validation revealed a similar prediction error for PE using the log-link GLM instead of the log-transformed linear model, but the error increased when GLM was used for other pigment predictions.

As for TChl *a* predictions from the satellite_ ρ_{wN} linear model, validation results of the MERIS Polymer TChl *a* product collocations with in situ TChl *a* from the satellite-based data set showed marginal differences for the 1×1 , 3×3 or 5×5 pixel collocations (Table 4, upper panel). The TChl *a* Polymer product obtained 3 % higher *MPD* and similar R^2 , *RMSE* and *PB* values (of about 0.74, 0.51 and 10 % on average, respectively) to the TChl *a* linear model predictions. In the global validation by Brewin et al. (2015), the OC4V6 (Ocean-Chlorophyll-4 algorithm version 6; O’Reilly et al., 2000) was selected from amongst various TChl *a* satellite products as the best TChl *a* algorithm. This algorithm is used to produce the MERIS Polymer TChl *a* from atmospheric-corrected MERIS Polymer data. Global validation by Brewin et al. (2015), with 1039 collocations and retrievals of TChl *a* directly from in situ $\rho_{wN}(\lambda)$ data, showed an R^2 of 0.87 and a *RMSE* of 0.29 for OC4V6 based on non-log-transformed

concentrations (which compares to our *RMSE* values on log-scale shown in Table 4 of Bracher et al., 2014). We conclude that both MERIS Polymer TChl *a* products, the level 2 and linear models, show high quality within the eastern Atlantic Ocean although they are retrieved from satellite data and not in situ ρ_{wN} data.

The comparison with other methods of retrieving pigment concentrations from reflectance data shows that our method, based on a linear model using EOFs from reflectance data, gives robust results for pigment groups and pigments that are always present in the region investigated. To test our EOF methods for independent data sets using the method established by a certain testing data set, we have used the cross-validation technique. The technique allows the re-sampling of all data for 500 different subsets (i.e., run by 500 permutations) into testing and validation data sets.

The advantage of our approach is that it allows for the estimation of several pigments and pigment groups using either reflectance data measured directly in the ocean water or obtained from a satellite ocean-color sensor. For the eastern tropical Atlantic Ocean data set, these additional pigments (other than TChl *a*) include PPC, PSC, DVChl *a* and MChl *a*. Additional pigments may also be accurately predicted with this approach; however, the results suggest that the prediction error increases for pigments that are found in lower concentrations or with a high number of samples below the detection limit (i.e., referred to in statistics as “censoring”). This poor performance may be in part due to the fact that pigments found in small concentrations are likely to have a limited ef-

Table 4. TChl *a* validation statistics for MERIS Polymer TChl *a* (left panel) and TChl *a* obtained from the full-fit linear model on satellite_ρ_{wN} (from MERIS Polymer data, EOF full-fit model, right panel) with different collocation criteria (either 1 × 1 or the mean of 3 × 3 or 5 × 5 pixel values) for the MERIS Polymer data compared to the in situ (from HPLC) value.

MERIS		Polymer				EOF full-fit model			
Collocation	<i>N</i>	<i>R</i> ²	RMSE	PB	MPD	<i>R</i> ²	RMSE	PB	MPD
1 × 1 pixel	139	0.74	0.51	6	39	0.72	0.54	10	37
3 × 3 pixel	155	0.75	0.50	10	39	0.74	0.52	10	36
5 × 5 pixel	160	0.73	0.51	11	40	0.74	0.52	10	36

fect on spectral shape, but further modeling work may also need to focus on better approaches for the treatment of censored values. Generally, we can also see from the field data linear models that using a coherent in situ data set, where all pigments have been measured by the same method and instrumentation, may be better suited for the modeling approach due to the homogeneous error across the range of pigment concentrations. An advantage of our linear method to pigment-specific band algorithms is that we require a much smaller data set for establishing the prediction (about 50 as opposed to several hundreds) of collocated pigment and reflectance data.

3.6 Application of linear model to study large-scale pigment distributions

For demonstrating the application of our linear model, we used the satellite_ρ_{wN} specific pigment's full-fit models for TChl *a*, MVChl *a*, PSC and PPC and ran these specific models using November 2008 MERIS Polymer ρ_{wN} level 2 data to retrieve those pigments for an example time period on a larger spatial scale. By subtracting the MVChl *a* value from TChl *a* we also derived concentrations of DVChl *a*. Figure 8 shows the monthly averages for those various pigment groups and pigments. Also, the MERIS Polymer TChl *a* concentration for the same time and region is shown.

The distributions of TChl *a* from the EOF model prediction or from the Polymer algorithm are very similar, ranging from 0.00003 to 7.52 mg TChl *a* m⁻³. For this particular month, the total biomass of phytoplankton shows a strong phytoplankton bloom (> 2 mg m⁻³) at the Mauritanian upwelling spread in two parts, 190–24° N and 14–7° N, and high values (> 0.5 mg m⁻³) at all coastal areas of the African continent. Enhanced TChl *a* concentrations > 0.3 mg m⁻³ are also spreading into the open ocean especially at 5–20° N and 30–40° W, along the 0° latitude from Africa to South America, and south of this at 3–10° S from 3° E to about 25° W. MVChl *a* follows more or less the TChl *a* distribution, however, only reaching the magnitude indicated by the TChl *a* values at the northern bloom. The deviation between TChl *a* and MVChl *a* is obvious in the distribution of DVChl *a*, which indicates that at the northern part of the Mauritanian upwelling bloom, *Prochlorococcus* (the only

phytoplankton genus which contains DVChl *a*) seems to have contributed to this bloom by only a very minor fraction (i.e., a few percent), while elsewhere it presents a substantial background of about 30 % of all phytoplankton.

Our predicted PPC concentrations show values in the same range as TChl *a* at the oligotrophic areas and about 50 % in the enhanced TChl *a* areas and the southern part of the bloom. As for DVChl *a*, in the northern part of the bloom PPC concentrations are significantly lower and only contribute less than 10 % to the total pigment concentrations. PSC concentration in the oligotrophic and enhanced TChl *a* areas are much lower than PPC or even DVChl *a* concentrations but reflect the TChl *a* distribution more or less on the large scale. Within the northern part of the Mauritanian upwelling PSC concentrations reach values even as high as for TChl *a*, while concentrations at the bloom further south contribute to less than 10 % of the total pigment concentrations. In Taylor et al. (2011) the analysis of pigment and additional microscopic data clearly showed very high concentrations of Fuco, a main pigment of PSC, and a high dominance of diatoms within water samples at the northern bloom collected at the same time period.

From our results, we can conclude that the northern phytoplankton bloom at the Mauritanian upwelling seems to have been freshly growing with very high photosynthetic activity, while for most of the other areas a lot of the energy build-up via photosynthesis was used for photoprotection. We have no information on photodegradation since no significant prediction linear model could be developed for phaeopigments. These pigments had only been identified in less than 60 % of all samples collocated to the field and satellite-based data sets, and the results show that this pigment group was not well predicted by the linear model. Based on the biogeography of Longhurst (2006), the oligotrophic areas on our maps fall in the North Atlantic Subtropical Gyre Province East at > 25° N (the border between the two is the subtropical convergence) and the North Atlantic Tropical Gyre Province at 25° N to about 12° N. At the eastern corner towards the coast of these provinces, in the Canary Coastal Province (CNRy), concentrations of all predicted pigments and pigment groups may have been increased due to eddy-driven processes that increase the supply of nutrients. In Taylor et al. (2011), the two blooms analyzed by field samples at CNRY have been

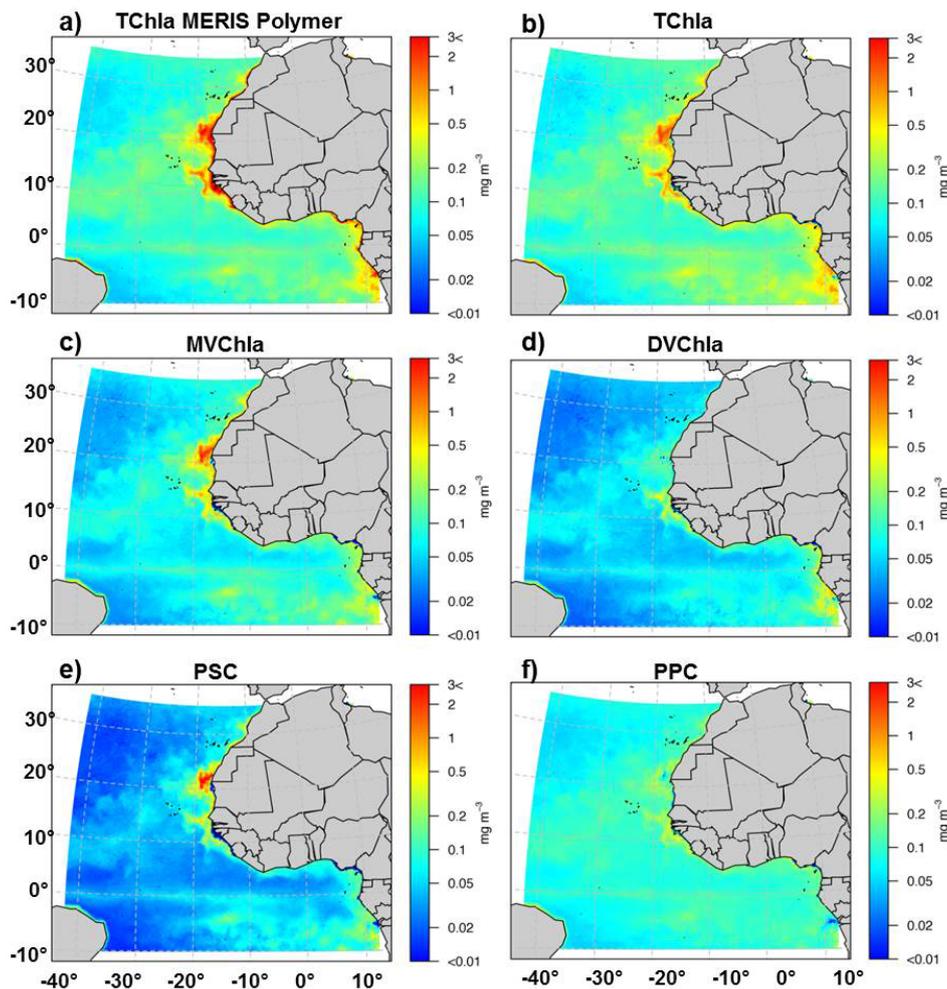


Figure 8. Monthly mean concentrations (in 0.25° grid resolution) for November 2008 of (a) TChl *a* of the MERIS Polymer algorithm (TChl *a* MERIS Polymer) and predicted (b) TChl *a*, (c) MVChl *a*, (d) DVChl *a*, (e) PSC and (f) PPC by the LM, based on the full fit of satellite ρ_{wN} data within the 1×1 pixel collocation box and the EOFs of this month's MERIS Polymer ρ_{wN} data.

identified to cluster differently due to their pigment composition. The northern “fresh” bloom with low photoinhibition, high dominance of phytoplankton and strong photosynthetic efficiency was related to a major upwelling focusing in the area south of Cape Blanc (western Sahara) off the coast of Mauritania. DVChl *a* was absent in this bloom, which is in line with our results obtained from the linear model. The southern part of the CNRY bloom was placed within the African dust veil where mineral-rich dust fertilizes the ocean. In the northern bloom stations, the spectral shape and absolute values of particulate non-phytoplankton absorption spectra, presented in Taylor et al. (2011) and coinciding with the pigment data used in our study, clearly indicated that mineral particle absorption was very high.

Comparisons of our predictions to pigment data not used for the development and validation of our EOF model show consistent results: Partensky et al. (1996) measured TChl *a* concentrations of about 1.2 mg m^{-3} in December 1992 (EU-

MELI 5 cruise) at a station within a phytoplankton bloom at $18^\circ 29' \text{ N}$ and $21^\circ 05' \text{ W}$, similar to the range of our predicted values at the southern edge of the northern bloom. Barlow et al. (2002) measured, within the area of our predictions, concentrations of TChl *a*, DVChl *a*, PSC and PPC during the AMT-3 cruise in October 1996 at 20° N and 20° W (0.4 , 0.05 , 0.175 and above 0.09 mg m^{-3} , respectively) and 30° N and 22° W (0.05 , 0.01 , 0.022 and above 0.04 mg m^{-3} , respectively), similar to our predicted concentrations for the same pigments just east of the northern bloom of the Mauritanian upwelling and at the North Atlantic Subtropical Gyre Province East, respectively.

4 Conclusions

We present robust predictions of concentrations of various pigments and pigment groups from linear models based on fitting empirical orthogonal function on a set of reflectance

data to collocated pigment concentrations. Spectral shapes of the reflectance spectra from the eastern Atlantic and of their derived EOF modes reflect typical case 1 water characteristics. In our study, it was shown that EOFs derived from both hyperspectral underwater radiometric measurements and multispectral reflectance data from field or satellite (MERIS Polymer) enable reliable predictions of the concentration of nine different pigments/pigment groups (TChl *a*, PPC, PSC, MVChl *a*, Chl *c*_{1/2}, But, Hex, Zea, Caro, PE). A limitation of all predictions was that only those pigments can be predicted that have been identified in every collocated sample; adding a small value (0.0001 mg m⁻³) to censored samples was not an appropriate solution to this problem.

The method proves for the first time to be applicable for predicting concentrations of not only TChl *a* and PE but also of other pigments and pigment groups with weaker, but spectrally unique, features on the underwater light field. Statistical resampling used for cross-validation indicates that predictions were robust ($R^2_{cv} \geq 0.5$, MDPD_{cv} $\leq 44\%$ and MPD_{cv} $\leq 60\%$) for all pigments (except for PE, Zea and But, which deviated for one of these measures) and pigment groups. Hyperspectral linear models proved to be already stable with less collocated samples for most pigment or pigment groups used for training ($n > 30$ to 40) than linear models based on multispectral reflectance data ($n > 50$). The linear models using MERIS Polymer reflectance data as input were applied to 1 month of satellite data to predict the concentrations of TChl *a*, PSC, PPC, MVChl *a* and DVChl *a* for the whole eastern tropical Atlantic. For the first time a consistent picture of several phytoplankton pigments indicating group-specific behavior and photophysiology on a larger spatial scale for this area was shown.

Our presented linear models are generic and can be applied to even a small, consistently collocated reflectance and pigment data set to enable various specific pigment predictions from continuous optical measurements. The optical data can be obtained from radiometric measurements based on various platforms (buoys, gliders, floats or satellite). On a global scale, TChl *a*, PSC and PPC are consistently accurately predicted, while other pigments may be better predicted on smaller spatial scales. Highly temporally resolved time series data, which – depending on the platform – may even provide good spatial coverage, can be used to study variability and change of overall phytoplankton and photophysiological responses to environmental variables. While we established the linear models for prediction of various pigments in typical case 1 waters, the method should be tested in the future for its applicability in case 2 waters as well.

The Supplement related to this article is available online at doi:10.5194/os-11-139-2015-supplement.

Author contributions. A. Bracher designed and ran the experiments and wrote the manuscript. M. H. Taylor developed the statistical method and wrote the R code; B. Taylor supplied all field and part of the additional pigment data, prepared all input data and designed part of the experiment; T. Dinter contributed to programming; R. Röttgers contributed to R_{rs} data and additional pigment data; and F. Steinmetz supplied MERIS Polymer data and match-ups. All co-authors assisted in writing the manuscript.

Acknowledgements. We thank AWI, Helmholtz Impulse Fond (HGF Young Investigators Group Phytooptics), Total Foundation (project PHYTOSCOPE) and the Leibniz Association (project OCEANET) for funding. We thank Erika Allhusen, Helmke Hepach, Kerstin Heymann, Laila Bentama, Anja Bernhardt and Sonja Wiegmann for work in the laboratory and on-board RVs *Polarstern* and *Merian S. Merian*. We thank the crew, principal investigators and other scientists on *ANTXXIII/1*, *ANTXXIV/1*, *ANTXXIV/4*, *ANTXXV/1*, *ANTXXVI/4* and MSM 18-3 for support on board and fruitful discussions afterwards. We thank Oliver Zielinski and his group at ICBM, University Oldenburg, previously at the University of Applied Sciences, Bremerhaven, for the use of their equipment. We thank ESA for MERIS level 1 data. We thank Amélie Talec for BGH sample collection and Hervé Claustre and Josephine Ras (all LOV) for BGH pigment sample analysis. This study contains data supplied by the Natural Environment Research Council (NERC) via BODC and NASA via SEABASS. We thank all researchers behind these large pigment data sets.

Edited by: O. Zielinski

References

- Aiken, J., Pradhan, Y., Barlow, R., Lavender, S., Poulton, A., Holligan, P., and Hardman-Mountford, N.: Phytoplankton pigments and functional types in the Atlantic Ocean: A decadal assessment, 1995–2005, *Deep-Sea Res. Pt. II*, 56, 899–917, doi:10.1016/j.dsr2.2008.09.017, 2009.
- Barker, K., Mazeran, C., Lerebourg, C., Bouvet, M., Antoine, D., Ondrusek, M., Zibordi, G., and Lavender, S.: MERMAID: MERIS MATchup in situ Database. Proceedings of the 2nd (A)ATSR and MERIS Workshop, Frascati, Italy, September 2008, 2008.
- Barlow, R. G., Cummings, D. G., and Gibb, S. W.: Improved resolution of mono- and divinyl chlorophylls a and b and zeaxanthin and lutein in phytoplankton extracts using reverse phase C-8 HPLC, *Marine Ecol. Progr. Ser.*, 161, 303–307, 1997.
- Barlow, R. G., Aiken, J., Holligan, P. M., Cummings, D. G., Martorena, S., and Hooker, S.: Phytoplankton pigment and absorption characteristics along meridional transects in the Atlantic, *Deep-Sea Res. Pt. I*, 47, 637–660, 2002.

- Behrenfeld, M. I. and Boss, E. B.: Beam attenuation and chlorophyll concentration as alternative optical indices of phytoplankton biomass. *J. Marine Res.*, 64, 431–451, 2006.
- Bracher, A., Taylor, M. H., Taylor, B., Dinter, T., Röttgers, R., and Steinmetz, F.: Using empirical orthogonal functions derived from remote sensing reflectance for the prediction of concentrations of phytoplankton pigments, *Ocean Sci. Discuss.*, 11, 2073–2117, doi:10.5194/osd-11-2073-2014, 2014.
- Brewin, R. J. W., Sathyendranath, S., Müller, D., Brockmann, C., Deschamps, P.-Y., Devred, E., Doerffer, R., Fomferra, N., Franz, B. A., Grant, M. G., Groom, S. B., Horseman, A., Hu C., Krasemann, H., Lee, Z., Maritorea, S., Mélin, F., Peters, M., Platt, T., Regner, P., Smyth, T., Steinmetz, F., Swinton, J., Werdell, J., and White III, G. N.: The Ocean Colour Climate Change Initiative: A round-robin comparison on in-water bio-optical algorithms, *Remote Sens. Environ.*, in press, 2015.
- Chase, A., Boss, E., Zaneveld, R., Bricaud, A., Claustre, H., Ras, J., Dall’Olmo, G., and Westberry, T. K.: Decomposition of in situ particulate absorption spectra, *Methods Oceanogr.*, 7, 110–124, 2013.
- Craig, S. E., Jones, C. T., Li, W. K. W., Lazin, G., Horne, E., Caverhill, C., and Cullen, J. J.: Deriving optical metrics of ecological variability from measurements of coastal ocean colour, *Remote Sens. Environ.*, 119, 72–83, 2012.
- Hirata, T., Hardman-Mountford, N. J., Brewin, R. J. W., Aiken, J., Barlow, R., Suzuki, K., Isada, T., Howell, E., Hashioka, T., Noguchi-Aita, M., and Yamanaka, Y.: Synoptic relationships between surface Chlorophyll-*a* and diagnostic pigments specific to phytoplankton functional types, *Biogeosciences*, 8, 311–327, doi:10.5194/bg-8-311-2011, 2011.
- Hooker, S. B., Van Heukelem, L., Thomas, C. S., Claustre, H., Ras, J., and Barlow, R.: The second SeaWiFS HPLC analysis round robin experiment (SeaHARRE-2). NASA TM/2005-212785, NASA Goddard Space Flight Center, Greenbelt, Maryland, 2005.
- Lee, Z. P., Carder, K. L., and Arnone, R. A.: Deriving inherent optical properties from water color: A multi-band quasi-analytical algorithm for optically deep waters, *Appl. Opt.*, 41, 5755–5772, 2002.
- Letelier, R. M., Bidigare, R. R., Hebel, D. V., Ondrusek, M., Winn, C. D., and Karl, D. M.: Temporal variability of phytoplankton community structure based on pigment analysis. *Limnol. Oceanogr.*, 38, 1420–1437, 1993.
- Longhurst, A.: *Ecological Geography of the Sea*, 2nd Edn., Elsevier Academic press, USA, 2006.
- Lubac, B. and Loisel, H.: Variability and classification of remote sensing reflectance spectra in the eastern English Channel and southern North Sea, *Remote Sens. Environ.*, 110, 45–48, 2007.
- McClain, C. R.: A Decade of Satellite Ocean Color Observations, *Ann. Rev. Marine Sci.*, 1, 19–42, doi:10.1146/annurev.marine.010908.163650, 2009.
- Müller, D. and Krasemann, H.: Product validation and algorithm selection report, part I – atmospheric correction. Tech. Rep. AO-1/6207/09/I-LG D2.5, European Space Agency, ESRIN, 2012.
- O’Reilly, J. E., Maritorea, S., Siegel, D., and O’Brien, M. C.: Ocean color chlorophyll algorithms for SeaWiFS, OC2, and OC4: Technical report, edited by: Toole, D., Mitchell, B. G., Kahru, M., Chavez, F. P., Strutton, P., Cota, G., Hooker, S. B., McClain, C. R., Carder, K. L., Muller-Karger, F., Harding, L., Magnuson, A., Phinney, D., Moore, G. F., Aiken, J., Arrigo, K. R., Letelier, R., Culver, M., Hooker, S. B., and Firestone, E. R.: SeaWiFS postlaunch calibration and validation analyses, Part 3. NASA, Goddard Space Flight Center, Greenbelt, Maryland. SeaWiFS Postlaunch Technical Report Series, 11, 9–23, 2000.
- Organelli, E., Bricaud, A., Antoine D., and Uitz, J.: Multivariate approach for the retrieval of phytoplankton size structure from measured light absorption spectra in the Mediterranean Sea (BOUSSOLE site), *Appl. Opt.*, 52, 2257–2273, 2013.
- Partensky, F., Blanchot, J., Lantoin, F., Neveux, J., and Marie, D.: Vertical structure of picophytoplankton at different trophic sites of the tropical northeastern Atlantic Ocean, *Deep-Sea Res. Pt. I*, 43, 1191–1213, 1996.
- Pan, X., Mannino, A., Russ, M. E., Hooker, S. B., and Harding, L. W.: Remote sensing of phytoplankton pigment distribution in the United States northeast coast, *Remote Sens. Environ.*, 114, 2403–2416, 2010.
- Pan, X., Wong, G. T. F., Ho T.-Y., Shiah, F.-K., and Liu H.: Remote sensing of picophytoplankton distribution in the northern South China Sea, *Remote Sens. Environ.*, 128, 162–175, 2013.
- R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, available at: <http://www.R-project.org/>, last access: 14 December 2013.
- Ras, J., Claustre, H., and Uitz, J.: Spatial variability of phytoplankton pigment distributions in the Subtropical South Pacific Ocean: comparison between in situ and predicted data, *Biogeosciences*, 5, 353–369, doi:10.5194/bg-5-353-2008, 2008.
- Roy, S., Llewellyn, C. A., and Egeland, E. S.: *Phytoplankton pigments: characterization, chemotaxonomy and applications in oceanography*, Cambridge University Press, 2011.
- Speich, S., Boyé, M., and Dehairs, F.: IPY BONUS-GoodHope cruise Report. May 2008, IPEV-INSU-LEFE-IFREMER-ANR-IRD-IUEM-UBO, 2008.
- Steinmetz F., Deschamps P.-Y., and Ramon D.: Atmospheric correction in presence of sun glint: application to MERIS, *Opt. Expr.*, 19, 9783–9800, 2011.
- Taylor, B. B., Torrecilla, E., Bernhardt, A., Taylor, M. H., Peeken, I., Röttgers, R., Piera, J., and Bracher, A.: Bio-optical provinces in the eastern Atlantic Ocean and their biogeographical relevance, *Biogeosciences*, 8, 3609–3629, doi:10.5194/bg-8-3609-2011, 2011.
- Taylor, B. B., Taylor, M., Dinter, T., and Bracher, A.: Estimation of relative phycoerythrin concentrations from hyperspectral underwater radiance measurements – a statistical approach, *J. Geophys. Res.-Ocean.*, 118, 2948–2960, doi:10.1002/jgrc.20201, 2013.
- Uitz, J., Claustre, H., Morel, A., and Hooker, S.: Vertical distribution of phytoplankton communities in open ocean: an assessment based on surface chlorophyll, *J. Geophys. Res.*, 111, C08005, doi:10.1029/2005JC003207, 2006.
- Vidussi, F., Claustre, H., Manca, B. B., Lucheta, A., and Marty, J.-C.: Phytoplankton pigment distribution in relation to upper thermocline circulation in the eastern Mediterranean Sea during winter, *J. Geophys. Res.*, 106, 19939–19956, 2001.
- Werdell, P. J., Franz, B. A., Bailey, S. W., Feldman, G. C., Boss, E., Brando, V. E., Dowell, M., Hirata, T., Lavender, S. J., Lee, Z. P., Loisel, H., Maritorea, S., Melin, F., Moore, T. S., Smyth, T. J., Antoine, D., Devred, E., d’Andon, O. H. F., and Mangin, A.:

- Generalized ocean color inversion model for retrieving marine inherent optical properties, *Appl. Opt.*, 52, 2019–2037, 2013.
- Wu, C. F.: Jackknife, bootstrap and other resampling methods in regression analysis, *Annals Stat.*, 14, 1261–1295, 1986.
- Zapata, M., Rodriguez, F., and Garido, J. L.: Separation of chlorophylls and carotenoids from marine phytoplankton: a new HPLC method using a reversed phase C8 column and pyridine-containing mobile phases, *Marine Ecol. Progr. Ser.*, 195, 29–45, 2000.