







Radiometric approach for the detection of picophytoplankton assemblages across oceanic fronts

PRISCILA KIENTECA LANGE,^{1,2,3,*}  P. JEREMY WERDELL,¹
ZACHARY K. ERICKSON,^{1,4}  GIORGIO DALL'OLMO,⁵  ROBERT J.
W. BREWIN,⁶ MIKHAIL V. ZUBKOV,⁷ GLEN A. TARRAN,⁵  HEATHER
A. BOUMAN,⁸ WAYNE H. SLADE,⁹ SUSANNE E. CRAIG,^{1,2} NICOLE J.
POULTON,¹⁰ ASTRID BRACHER,^{11,12} MICHAEL W. LOMAS,¹⁰ AND
IVONA CETINIĆ^{1,2}

¹*Ocean Ecology Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 21077, USA*

²*Universities Space Research Association, 7178 Columbia Gateway Drive, Columbia, MD 21046, USA*

³*Blue Marble Space Institute of Science, Seattle, WA 98154, USA*

⁴*NASA Postdoctoral Program Fellow, Ocean Ecology Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 21077, USA*

⁵*National Centre for Earth Observation, Plymouth Marine Laboratory, Plymouth PL1 3DH, UK*

⁶*College of Life and Environmental Sciences, University of Exeter, Penryn, Cornwall TR10 9EZ, UK*

⁷*Scottish Association for Marine Science, Scottish Marine Institute, Dunbeg, Oban, Argyll, PA37 1QA Scotland, UK*

⁸*Department of Earth Sciences, University of Oxford, South Parks Rd, Oxford OX1 3AN, UK*

⁹*Sequoia Scientific, Inc., 2700 Richards Road, Suite 107, Bellevue, WA 98005, USA*

¹⁰*Bigelow Laboratory for Ocean Sciences, 60 Bigelow Drive, East Boothbay, ME 04544, USA*

¹¹*Climate Sciences, Alfred-Wegener-Institute Helmholtz Center for Polar and Marine Research, D-27570 Bremerhaven, Germany*

¹²*Institute of Environmental Physics, Department of Physics and Electrical Engineering, University of Bremen, D-28359 Bremen, Germany*

*priscila@bmsis.org

Abstract: Cell abundances of *Prochlorococcus*, *Synechococcus*, and autotrophic picoeukaryotes were estimated in surface waters using principal component analysis (PCA) of hyperspectral and multispectral remote-sensing reflectance data. This involved the development of models that employed multilinear correlations between cell abundances across the Atlantic Ocean and a combination of PCA scores and sea surface temperatures. The models retrieve high *Prochlorococcus* abundances in the Equatorial Convergence Zone and show their numerical dominance in oceanic gyres, with decreases in *Prochlorococcus* abundances towards temperate waters where *Synechococcus* flourishes, and an emergence of picoeukaryotes in temperate waters. Fine-scale *in-situ* sampling across ocean fronts provided a large dynamic range of measurements for the training dataset, which resulted in the successful detection of fine-scale *Synechococcus* patches. Satellite implementation of the models showed good performance ($R^2 > 0.50$) when validated against *in-situ* data from six Atlantic Meridional Transect cruises. The improved relative performance of the hyperspectral models highlights the importance of future high spectral resolution satellite instruments, such as the NASA PACE mission's Ocean Color Instrument, to extend our spatiotemporal knowledge about ecologically relevant phytoplankton assemblages.

Published by The Optical Society under the terms of the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

1. Introduction

Observing spatiotemporal changes in the composition of phytoplankton assemblages over broad areas of the ocean increases our understanding of the response of these critical photoautotrophs to environmental and climatic processes. The smallest phytoplankton cells, most often categorized as picophytoplankton ($< 2 \mu\text{m}$ [1]) or ultraphytoplankton ($< 3 \mu\text{m}$ [2]), are the most abundant primary producers in the global ocean. Despite their individually low biomass relative to other primary producers [3,4], picophytoplankton are dominant in $\sim 50\%$ of the world's surface oceans, where the reduced availability of inorganic nutrients limits the growth of larger phytoplankton cells [5–7]. Composed of the cyanobacteria *Prochlorococcus* ($\sim 0.8 \mu\text{m}$) and *Synechococcus* ($\sim 1 \mu\text{m}$), as well as a polyphyletic group of picoeukaryotes, picophytoplankton are responsible for 50 to 90% of all primary production in open ocean ecosystems [8,9]. They therefore play a substantial role in the maintenance of the marine food web and contribute up to 30% of the total carbon export to the deep ocean [10–12].

Given the important ecological and biogeochemical roles of picophytoplankton, the oceanographic community invests substantially in improving our scientific understanding of their spatiotemporal patterns. Ship-based *in-situ* measurements of phytoplankton composition have revealed important paradigms in their diversity [13–18]. In the Atlantic Ocean, for example, *Prochlorococcus* inhabits warmer and mostly oligotrophic waters surrounded by spatially adjacent fronts of sub-mesoscale *Synechococcus* patches [8,13,18]. These fronts often reside at boundaries where phytoplankton communities start to transition to higher concentrations of larger eukaryotic cells, such as picoeukaryotes and nanoeukaryotic flagellates [8,19] (Fig. 1). Hence, identification of *Prochlorococcus* and *Synechococcus* distributions may conceptually be used to identify trophic boundaries in oceanic ecosystems [20], in addition to providing insight into productivity, food web regimes, and carbon export.

Ocean color satellite instruments provide a tool for capturing and retrospectively analyzing phytoplankton spatiotemporal patterns on synoptic and long-term scales that are unattainable by conventional *in-situ* methods [21–23]. These instruments measure visible and near-infrared radiances at discrete wavelengths at the top-of-the atmosphere. Atmospheric correction algorithms are applied to remove contributions of the atmosphere and surface reflection from the total signal, leaving estimates of spectral remote-sensing reflectances ($R_{rs}(\lambda)$; sr^{-1}), the light exiting the water column normalized to the downwelling surface irradiance [24]. Bio-optical algorithms are subsequently applied to the $R_{rs}(\lambda)$ to produce estimates of near-surface concentrations of the photosynthetic pigment chlorophyll-a (*Chl*; mg m^{-3}) and other metrics of phytoplankton community composition [25–27]. Other existing bio-optical algorithms provide abundances or biomass of different phytoplankton using unique empirical relationships between cell abundance and $R_{rs}(\lambda)$, as well as additional satellite observables such as sea surface temperature (*SST*; $^{\circ}\text{C}$) and photosynthetically active radiation (*PAR*; $\mu\text{E m}^{-2} \text{s}^{-1}$) [9,28–30].

To date, the majority of bio-optical algorithms that explore phytoplankton community composition exploit the capabilities of multispectral ocean color satellites, using only a few wavelengths of an $R_{rs}(\lambda)$ spectrum [21,23,31]. More recent approaches consider increased spectral resolution, following the development of commercial off-the-shelf instrumentation allowing the hyperspectral *in-situ* measurement of $R_{rs}(\lambda)$ and the expectation that hyperspectral ocean color satellite instruments will be launched in the foreseeable future [32]. Given the higher information content of hyperspectral radiometry, sophisticated statistical methods have been successfully applied to assess its variability and correlation with phytoplankton attributes of interest [18,33–39]. The forthcoming NASA Plankton, Aerosol, Cloud, ocean Ecosystem (*PACE*) mission is expected to increase the interest and demand for hyperspectral methods for global phytoplankton community composition assessment [40].

In this paper, we present empirical algorithms based on principal component regressions that provide estimates of surface abundances of *Prochlorococcus*, *Synechococcus*, and autotrophic

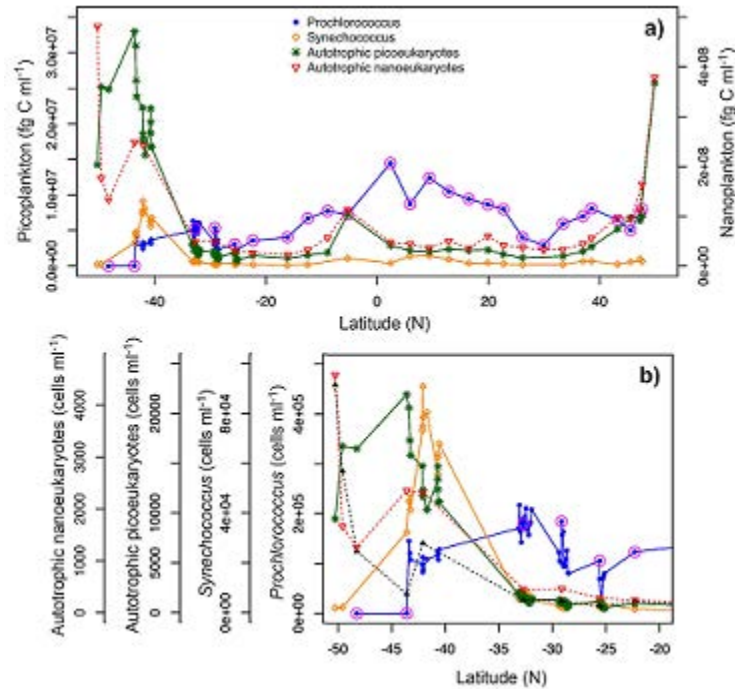


Fig. 1. **a)** Carbon concentration estimated from flow-cytometric cell counts across the Atlantic Meridional Transect, and **b)** cell abundance (scaled to group-specific maximum cell abundance) of *Prochlorococcus* (blue), *Synechococcus* (orange), autotrophic picoeukaryotes (green) and autotrophic nanoeukaryotes (red) in surface waters of the frontal system between the South Atlantic Gyre and temperate waters of the South Atlantic (subset of the southern portion of the transect in (a)). Data collected during AMT24 (2014). Red circles in *Prochlorococcus* indicate samples that were taken from CTD casts. The remaining samples (across the *Synechococcus* front) were taken from the ship's underway system.

picoeukaryotes, derived from *in-situ* datasets of measured cell abundances and hyperspectral $R_{rs}(\lambda)$. First, we explore the viability of principal component techniques for the identification of some of the smallest phytoplankton community members using hyperspectral and multispectral $R_{rs}(\lambda)$. This exploration includes an assessment of performance enhancement using both $R_{rs}(\lambda)$ and remotely sensed *SST* as an additional predictor. Second, we evaluate the relative performance of multi- and hyperspectral implementations of these algorithms. These comparisons quantify improvements in *Prochlorococcus* and *Synechococcus* retrievals when additional spectral information is used. Knowledge of such performance differences provides a metric of relative uncertainty to be considered when evaluating results from heritage multispectral satellite instruments in comparison with forthcoming hyperspectral satellite instruments such as NASA's PACE mission [40].

2. Material and methods

2.1. Algorithm training *in-situ* dataset

Radiometric, hydrographic, and phytoplankton abundance *in-situ* data for algorithm training were collected during the Atlantic Meridional Transect 24 (AMT24) oceanographic expedition, which took place between the United Kingdom and the Falkland Islands during boreal autumn

(September 30th to November 1st, 2014) onboard the RRS James Clark Ross. AMT24 covered most biogeochemical provinces of the Atlantic Ocean (Fig. 2), capturing several marine ecosystems inclusive of ocean gyres, the highly productive Equatorial Convergence Zone, and the high-latitude boundaries of the ocean gyres [8,41].

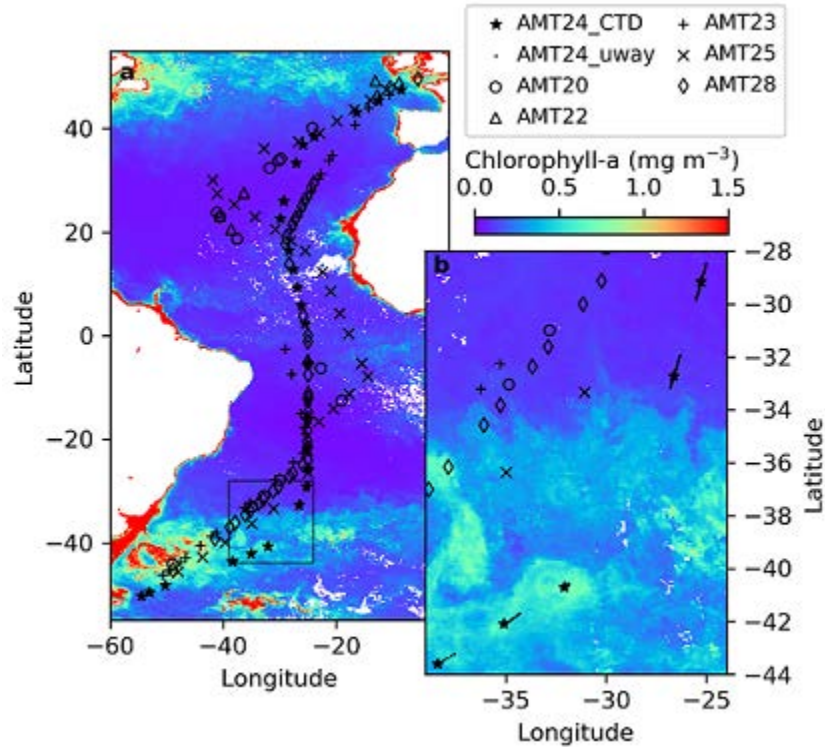


Fig. 2. **a)** CTD stations for the training dataset (Atlantic Meridional Transect 24 - AMT24), and the validation datasets (AMTs 20, 22, 23, 25 and 28), with the monthly composite of chlorophyll (MODerate resolution Imaging Spectroradiometer onboard Aqua - Aqua-MODIS) in October/2014, and **b)** magnification of frontal region between the South Atlantic Gyre and temperate waters, highlighting the frequent underway samples (dots).

The sampling strategy to generate an appropriate dataset to develop a predictive algorithm targeted to a phytoplankton group must be designed according to the spatial scales of variability for this group. As such, consideration of previous knowledge about the biology and ecology of this phytoplankton group is useful. With that in mind, we considered two different approaches to collect discrete samples for the analysis of picophytoplankton community structure. First, daily surface (< 10 m depth) samples were collected at 13:00 (local time) using a Niskin bottle deployed as part of the CTD rosette (Fig. 2). Second, additional surface samples were collected every 30 minutes from the underway system of the ship (Fig. 2) while crossing the front between the South Atlantic Gyre and temperate waters (latitude from 25°S to 45°S). This is the region where we expect a transition from *Prochlorococcus* dominance into the sub-mesoscale *Synechococcus* patches. Water temperature was measured using a CTD (Sea-Bird Electronics SBE 9/11) installed on the rosette profiler or using the hull-mounted shipboard CTD unit (SBE 3P). More details on the underway sampling can be found in Brewin et al. [42].

Above-water radiometric data were collected in continuous underway mode using three Sea-Bird Electronics HyperSAS radiometer systems (measuring total upwelling radiance $L_r(\lambda)$,

sky radiance $L_{sky}(\lambda)$, and planar downwelling irradiance $E_d(\lambda)$ as described by Brewin et al. [43]. The radiometers have nominal spectral resolution of 10 nm and spectral sampling of 3.3 nm. The procedure to process radiometric data followed protocols described in the same reference, with the following modifications: 1) raw radiometric data were converted to physical quantities using calibration coefficients computed as the average between the pre- and post-cruise calibrations; 2) corrections for dark counts, interpolated in time and over a common wavelength range, were done as in Brewin et al. [43]; 3) continuous measurements of pitch and roll were used to compute tilt angles and all radiometric measurements corresponding to tilt angles $\geq 5^\circ$, or with solar zenith angles $\geq 80^\circ$ and $\leq 10^\circ$ were discarded; 4) the relative azimuth angle ($\Delta\phi$) between sensor (ϕ) and sun (ϕ_0) was computed as $\Delta\phi = \phi - \phi_0$ and all radiometric measurements with $\Delta\phi \geq 170^\circ$ and $\Delta\phi \leq 50^\circ$ were discarded; and, 5) an existing technique based on the assumed absence of upward radiance in the near infrared in open-ocean waters [44] was adapted to minimize sun glint. For the latter, we divided the continuous underway dataset into 1-minute intervals and for each interval we only retained the data corresponding to the $L_t(\lambda)$ spectrum that had the minimum $L_t(\lambda)$ in the near-infrared spectral region as determined by the average of values in the 750-800 nm range. Water-leaving radiance ($L_w(\lambda)$) was computed by subtracting the influence of sky and sunlight specularly reflected by the sea surface using the following equation:

$$L_w = L_t - \rho_{sky}L_{sky} - L_{NIR}, \quad (1)$$

where ρ_{sky} and L_{NIR} are scalar coefficients that we obtained by minimizing the following cost function:

$$C = \sum_{\lambda=750}^{\lambda=800} |L_t(\lambda) - \rho_{sky}L_{sky}(\lambda) - L_{NIR}|. \quad (2)$$

In practice, this minimization routine ensures that the derived $L_w(\lambda)$ is approximately zero and spectrally flat between 750 and 800 nm. Finally, remote-sensing reflectances were computed by dividing $L_w(\lambda)$ by $E_d(\lambda)$.

Once processed, $R_{rs}(\lambda)$ from 414 to 660 nm were interpolated (2 nm resolution), then quality-controlled by removing: 1) measurements collected earlier than 09:00 local time or later than 17:00 local time; 2) spectra that showed negative values in the visible range (400-700 nm); and, 3) spectra with second derivative values higher than $2 \times 10^{-4} \text{ sr}^{-1} \text{ nm}^{-1}$ or lower than $-2 \times 10^{-4} \text{ sr}^{-1} \text{ nm}^{-1}$ in the spectral region from 610 to 660 nm, as a means of noise removal. Coincidence between in-situ $R_{rs}(\lambda)$ measurements and discrete sampling locations was determined by time (date, hour, and minute of sampling). Prior to the numerical analysis, each $R_{rs}(\lambda)$ spectrum was standardized ($R_{rs}'(\lambda)$) [33,35] following:

$$R_{rs}'(\lambda = i) = \frac{R_{rs}(\lambda = i) - \text{mean}[R_{rs}]_{414}^{660}}{\text{sd}[R_{rs}]_{414}^{660}}, \quad (3)$$

where $R_{rs}(\lambda=i)$ is the R_{rs} at the i^{th} wavelength, and $\text{mean}[R_{rs}]_{414}^{660}$ and $\text{sd}[R_{rs}]_{414}^{660}$ are the average and standard deviations of $R_{rs}(\lambda)$ of values between 414 and 660 nm in one $R_{rs}(\lambda)$ spectrum. This standardization of the $R_{rs}(\lambda)$ curves highlights spectral features of $R_{rs}(\lambda)$ and minimizes variance due to amplitude. Within open ocean (case 1) waters, the variability in the shapes of spectral features are mostly governed by phytoplankton absorption properties (i.e., pigments and packaging) [45], which provide the most useful spectral characteristics to differentiate between taxonomic groups. Features caused by changes in the spectral slope of backscattering and absorption by colored dissolved organic matter (CDOM) are still reflected in the shape of standardized $R_{rs}'(\lambda)$ spectra. Less spectrally distinct changes in $R_{rs}(\lambda)$ result from backscattering effects driven by particle morphological characteristics and refractive indices, and from processing errors in underway measured $R_{rs}(\lambda)$ such as sea-surface correction and cloud effects. The measured $R_{rs}'(\lambda)$ spectra from the AMT24 dataset are shown in Fig. 3.

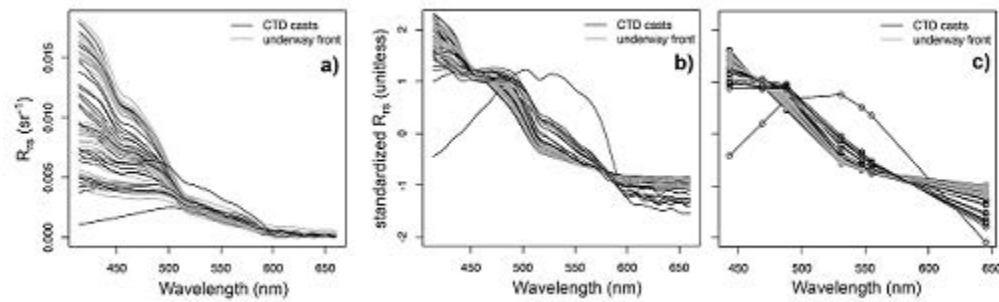


Fig. 3. Remote-sensing reflectances ($R_{rs}(\lambda)$) measured at discrete sampling locations across the Atlantic Ocean during AMT24: **a)** original hyperspectral measurements; **b)** standardized hyperspectral measurements; **c)** standardized multiband measurements at the central wavelengths of seven Aqua-MODIS bands: 443, 469, 488, 531, 547, 555, and 645 nm.

Picophytoplankton cell concentrations (cells ml^{-1}) were analyzed in 1.6 ml seawater samples preserved with paraformaldehyde using a FACSCalibur (Becton Dickinson) flow cytometer. Yellow-green 0.5 and 1.0 μm reference beads (Fluoresbrite Microparticles, Polysciences, Warrington, PA, USA) were used as an internal standard for both fluorescence and flow rates [46]. For *Prochlorococcus* and *Synechococcus*, samples were stained with a 1% commercial stock solution of SYBR Green 1 (Molecular Probes, Inc.) in Milli-Q water, then mixed with 300 mol m^{-3} tripotassium citrate (24.5 mol m^{-3} final concentration) [47]. This method allows the distinction of different populations of microbes based on their DNA content and right-angle light scatter (*RALS*), regardless of their intracellular *Chl* content (red fluorescence) [46]. Autotrophic eukaryotes were quantified based on their red fluorescence and *RALS*, using the method described in Olson et al. [48]. The AMT24 picoplankton dataset is freely available [49].

2.2. Validation in-situ datasets

Radiometric, hydrographic, and phytoplankton abundance *in-situ* data for algorithm validation were collected during several oceanographic expeditions. First, cross-validation (see section 2.4) was performed using the same AMT24 dataset that was used for training the model. Then, a satellite implementation was tested using flow cytometric counts from five additional AMT cruises (AMT20, 22, 23, 25, and 28) [50–53] and coincident $R_{rs}(\lambda)$ and *SST* satellite retrievals (see details in section 2.3), provided by the British Oceanographic Data Centre (BODC) [52]. Flow cytometric quantification of *Prochlorococcus*, *Synechococcus* and autotrophic picoeukaryotes was conducted using the method described in Olson et al. [48], except on AMTs 23 and 25 where *Prochlorococcus* was quantified following Zubkov et al. [47]. The collection and processing of flow cytometric data on these cruises followed the methods described in Lange et al. [28]. The five AMT cruises surveyed similar locations and occurred in similar seasons (late September to early November) spanning 2010 to 2018 (detailed information on cruise tracks and dates are described in the Atlantic Meridional Transect website [54]).

2.3. Satellite data

MODerate resolution Imaging Spectroradiometer onboard Aqua (Aqua-MODIS) data were acquired from the NASA Ocean Biology Processing Group [55]. This included Level-3, 4-km global maps of $R_{rs}(\lambda)$ and *SST* spanning the following periods: daily and 8-day composites from September 30th to November 1st, 2014 (the duration of AMT24); and 8-day composites spanning October 12th to November 25th 2010 (AMT20), October 10th to November 24th 2012 (AMT22),

October 3rd to November 4th 2013 (AMT23), September 11th to November 4th 2015 (AMT25) and September 23rd to October 30th 2018 (AMT28). Data from 8-day satellite composites were considered to match *in-situ* sampling locations when the date of the *in-situ* collection fell within the 8-day window of the composite and its location was located inside a valid 4-km satellite pixel. The October 2014 monthly cell abundance composites were created by averaging products that used 8-day composites from October 2014 as input.

Although the temporal interval between in-situ and satellite data may be long (for instance 3-4 days) when using 8-day satellite composites, the abundance of picophytoplankton cells are not expected to change abruptly over time in stratified environments where they are most abundant (i.e. ocean gyres and Equatorial divergence zone). Phytoplankton community structure in these regions gradually changes over the seasons, with a much less dynamic behavior than temperate waters and shelf seas. Thus, these operationally-viable retrievals from 8-day satellite composites show their distribution patterns in enough detail and an acceptable associated uncertainty led by temporal mismatch. Data processing and quality assurance followed the OBPG reprocessing configuration 2018.0 [55]. Available visible Aqua-MODIS $R_{rs}(\lambda)$ from the OBPG were used at 443, 469, 488, 531, 547, 555, and 645 nm wavelengths. Satellite $R_{rs}(\lambda)$ spectra were standardized according to Eq. (3) before being utilized for model implementation.

2.4. Model development

Following Craig et al. [33] and Bracher et al. [35], we used principal component regression to derive empirical relationships for the prediction of the abundances of *Prochlorococcus*, *Synechococcus* and autotrophic picoeukaryotic cells from scores of a principal component analysis (PCA) of *in-situ* $R_{rs}'(\lambda)$ from AMT24. We also considered the SST measured in the AMT24 stations as an additional predictor to improve the performance of the PCA score-based empirical models. The decomposition of standardized $R_{rs}(\lambda)$ spectra via PCA was performed in R using the function *prcomp* (package stats [56]), using: 1) hyperspectral $R_{rs}'(\lambda)$ spanning 414-660 nm with 2 nm intervals, hereafter referred to as PCAh, and 2) $R_{rs}(\lambda)$ measurements at the seven Aqua-MODIS wavelengths (443, 469, 488, 531, 547, 555, 645 nm) available in the HyperSAS measurement range, hereafter referred to as PCAm. The matrix \mathbf{X} with the $R_{rs}'(\lambda)$ spectra was decomposed into principal components (PC) via:

$$\mathbf{X}(n, w) = \mathbf{U}(n, p) \sum(p) \mathbf{V}(w, p)^T, \quad (4)$$

where the matrix \mathbf{V} of loadings (also known as eigenvectors) shows the spectral contributions to each PC (or mode), the vector \sum contains the singular values (square-root of scores), and the matrix \mathbf{U} of scores (or eigenvalues) consists of the projection of samples at each PC driven by the variability of $R_{rs}'(\lambda)$ in distinct sections of the spectrum [35]. The values n , w , and p in parentheses indicate dimensions of the matrices and correspond to the number of observations, number of wavelengths, and number of PCs, respectively, where the number of PCs is equal to the smallest number between n and w . Derived PCs with a standard deviation lower than 0.1% of the standard deviation of the first PC were discarded, resulting in 20 PCs from PCAh and 5 PCs from PCAm. Additional PCs were discarded based on their significance as a predicting variable in the empirical model (p-values > 0.05), resulting in 14 PCs for PCAh and 3 PCs for PCAm.

The PC scores were used as predictors in multilinear regression analyses targeting the abundances of *Prochlorococcus* (*Pro*) (Eq. (5)), *Synechococcus* (*Syn*) (Eq. (6)), and autotrophic picoeukaryotes (*Apeuk*) (Eq. (7)). The initial empirical models were developed using SST and all PC scores as predictors. Irrelevant predictors (highest p-value in the regression model) were then systematically discarded using backward stepwise selection. As each predictor was discarded, the new model (without the discarded predictor) was compared with the previous model (including that predictor) using the Akaike Information Criteria (AIC), and the model with the lower AIC value was selected. This process was interrupted when the model that included a target predictor

showed lower AIC than the model where it was removed. Then, the other variables were removed one by one, and the AIC was re-calculated to assure the best selection of variables, including those with low p-values in the regression. In the final regressions, *SST* was used as an additional predictor for *Prochlorococcus* and picoeukaryotes, composing the following formulations:

$$y_{pro} = a + b_0 \log_{10}(SST) + b_1 u_1 + b_2 u_2 + \dots + b_p u_p \quad (5)$$

$$\log_{10}(y_{syn}) = a + b_1 u_1 + b_2 u_2 + \dots + b_p u_p, \quad \text{and} \quad (6)$$

$$\log_{10}(y_{Apeuk}) = a + b_0 \log_{10}(SST) + b_1 u_1 + b_2 u_2 + \dots + b_p u_p \quad (7)$$

where y is the concentration of cells (cells ml^{-1}), $u_{1,2,\dots,p}$ is the score of a $R_{rs}(\lambda)$ spectrum in the p^{th} PC from the matrix \mathbf{U} , a is the intercept, and $b_{0,1,2,\dots,p}$ are the regression coefficients. The explanatory variable *SST* and the response variables (cell abundances of *Synechococcus* and autotrophic picoeukaryotes) were log-transformed for the multilinear regression analysis to achieve a normal distribution. In contrast, cell abundances of *Prochlorococcus* demonstrated normal distribution, thus log-transformation was not required and, when implemented for testing, significantly reduced the performance of the empirical model. The workflow of calculations is displayed in Fig. 4.

2.5. Model uncertainty assessment

To assess the robustness of the empirical models, cell abundance estimates were compared with the *in-situ* observations using the approach proposed by Seegers et al. [57], which includes two statistical metrics for uncertainty: average bias (Eq. (8)) and mean absolute error (*MAE*, Eq. (9)), assuming the normal frequency distribution of the variables. Here, we also calculate the adjusted coefficient of determination (R^2 , Eq. (10)). These metrics were calculated as follows:

$$\text{bias} = 10 \left(\frac{1}{n} \sum_{i=1}^n \log_{10}(X_i^P) - \log_{10}(X_i^O) \right), \quad (8)$$

$$\text{MAE} = 10 \left(\frac{1}{n} \sum_{i=1}^n | \log_{10}(X_i^P) - \log_{10}(X_i^O) | \right), \quad \text{and} \quad (9)$$

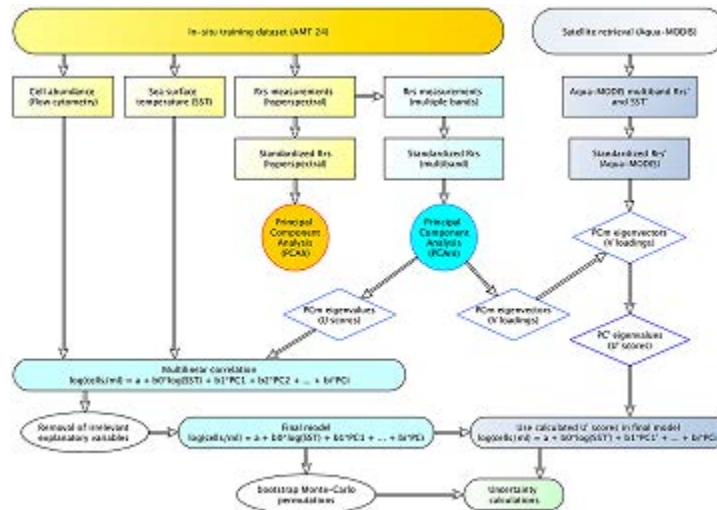


Fig. 4. Workflow of calculations performed in the predictive models: Model design (yellow and blue) and model application to Aqua-MODIS data (grey).

$$R^2 = 1 - (1 - R^2) \left[\frac{n - 1}{n - (k + 1)} \right] \quad (10)$$

where n is the number of observations, X^P is the predicted variable, X^O is the observed variable, and k is the number of independent variables in the equation. For consistency across all phytoplankton assemblages, all metrics were calculated in logarithmic space, and reported values therefore can be assessed as relative or percentage uncertainties (i.e., Eqs. (3) and (4) from Seegers et al. [57]). Uncertainties were calculated using the following dataset arrangements:

- 1) *Full-fit in-situ predictions*: Models trained with the AMT24 dataset were used to compute cell abundances from *in-situ* $R_{rs}(\lambda)$ measurements from AMT24 and predictions were compared to *in-situ* observations of cell abundances from AMT24, which were also used for developing the models (Tables 1 and 2);
- 2) *Cross-validation based on in-situ predictions*: Models trained with randomly sub-sampled training datasets (80% of the original AMT24 dataset) were used to compute cell abundances using the remaining 20% of the dataset, and these predictions were compared with observations from this 20% sub-dataset (bootstrap method). This process was repeated (2000 Monte-Carlo permutations) and the average performance metrics were computed (Tables 1 and 2);
- 3) *Satellite predictions using full-fit multispectral in-situ models*: Models trained with the AMT24 dataset were used to compute cell abundances from Aqua-MODIS $R_{rs}(\lambda)$ and SST retrievals (daily and 8-day composites) matching the time and location of sampling of AMT24, and predictions were compared to *in-situ* observations of cell abundances which were used to develop the prediction models (Table 3); and,
- 4) *Validation of satellite predictions with independent datasets*: Models trained with the AMT24 dataset were used to compute cell abundances from Aqua-MODIS $R_{rs}(\lambda)$ and SST retrievals (8-day composites) matching the time and location of sampling of five AMT cruises (AMTs 20, 22, 23, 25 and 28), and predictions were compared to *in-situ* observations of cell abundances (Table 3).

Arrangements 1 and 2 assess model performance and robustness against the selection of input data, respectively. Arrangement 2 (cross-validation) allows an assessment of whether or not the

Table 1. Arrangement 1 uncertainty calculations for cell abundance (cells ml⁻¹) model estimates of Prochlorococcus, Synechococcus and autotrophic picoeukaryotes during AMT24, with or without SST. Bias and MAE were calculated in log₁₀ normal space, thus are expressed in relative values corresponding to the percentage deviation from 1 (i.e., 1.09 = +9%, 0.93 = -7%). R² was calculated in log₁₀ normal space for Synechococcus and picoeukaryotes, but with untransformed data for Prochlorococcus because Prochlorococcus abundances naturally show a normal distribution. The best performing hyperspectral and multispectral models using either PCs + SST or PCs only are indicated in bold, with corresponding results shown in Fig. 6.

Spectral resolution	Predicted variable	AMT24 (PCs only)				AMT24 (PCs + SST)			
		<i>n</i>	<i>bias</i>	<i>MAE</i>	<i>R</i> ²	<i>n</i>	<i>bias</i>	<i>MAE</i>	<i>R</i> ²
Hyperspectral	<i>Prochlorococcus</i> ^a	73	1.13	1.49	0.42	73	1.08	1.31	0.82
	<i>Synechococcus</i>	73	~ 1	1.27	0.92	73	~ 1	1.27	0.92
	Picoeukaryotes	78	~ 1	1.21	0.95	78	~ 1	1.21	0.95
Multispectral	<i>Prochlorococcus</i> ^a	73	1.09	1.46	0.50	73	1.09	1.33	0.76
	<i>Synechococcus</i>	73	~ 1	1.45	0.81	73	~ 1	1.45	0.81
	Picoeukaryotes ^a	78	1	1.27	0.90	78	1	1.24	0.92

^aModels chosen to use sea surface temperature (SST) as an additional predictor.

full-fit model is overtrained (i.e., not generalizable to datasets other than its training dataset). If the full-fit and the cross-validation performance metrics show similar results, the model is robust (i.e., not overtrained). Arrangements 3 and 4 are used to assess the performance of the model in terms of application to satellite data to assess its uncertainty by validation with independent datasets. All statistical analyses were performed using the R packages *stats* [56], *MASS* [58], and *devtools* [59].

Table 2. Arrangement 1 (full-fit) versus 2 (cross-validation) uncertainty calculations for cell abundance model estimates of *Prochlorococcus*, *Synechococcus* and autotrophic picoeukaryotes during AMT24. Bias and MAE were calculated in \log_{10} space, thus are expressed in relative values corresponding to the percentage deviation from 1 (i.e., 1.09 = +9%, 0.93 = -7%). R^2 was calculated in \log_{10} space for *Synechococcus* and picoeukaryotes, but with untransformed data for *Prochlorococcus* because *Prochlorococcus* abundances naturally show a normal distribution.

Spectral resolution	Predicted variable	All AMT24 (Arrangement 1)				Re-sampled AMT24 (Arrangement 2)			
		<i>n</i>	<i>bias</i>	<i>MAE</i>	R^2	<i>n</i>	<i>bias</i>	<i>MAE</i>	R^2
Hyperspectral	<i>Prochlorococcus</i> ^a	73	1.08	1.31	0.82	16	1.08	1.35	0.78
	<i>Synechococcus</i>	73	~ 1	1.27	0.92	16	~ 1	1.36	0.85
	Picoeukaryotes	78	~ 1	1.21	0.95	16	~ 1	1.26	0.92
Multispectral	<i>Prochlorococcus</i> ^a	73	1.09	1.33	0.76	16	1.11	1.38	0.74
	<i>Synechococcus</i>	73	~ 1	1.45	0.81	16	1.01	1.50	0.74
	Picoeukaryotes ^a	78	1	1.24	0.92	16	0.94	1.39	0.76

^aModels using sea surface temperature (*SST*) as an additional predictor.

Table 3. Arrangements 3 and 4 uncertainty calculations for cell abundance model estimates (cells ml^{-1}) of *Prochlorococcus*, *Synechococcus* and autotrophic picoeukaryotes using Aqua-MODIS $R_{rs}(\lambda)$ 8-day-composite retrievals for time and location of AMT24 sampling sites and those of AMTs 20, 22, 23, 25 and 28. Bias and MAE were calculated in \log_{10} normal space, thus are expressed in relative values corresponding to the percentage deviation from 1 (i.e., 1.09 = +9%, 0.93 = -7%). R^2 was calculated in \log_{10} normal space for *Synechococcus* and picoeukaryotes, but with untransformed data for *Prochlorococcus* because *Prochlorococcus* abundances naturally show a normal distribution.

Spectral resolution	Predicted variable	AMT24 Aqua-MODIS (8-day composites)				AMTs 20,22-25, 28 Aqua-MODIS (8-day)			
		<i>n</i>	<i>bias</i>	<i>MAE</i>	R^2	<i>n</i>	<i>bias</i>	<i>MAE</i>	R^2
Multispectral	<i>Prochlorococcus</i> ^a	60	1.09	1.37	0.58	113	1.75	2.26	0.54
	<i>Synechococcus</i>	68	0.62	2.04	0.50	120	0.93	2.20	0.40
	Picoeukaryotes ^a	65	0.91	1.28	0.92	120	1.05	1.53	0.60

^aModels using sea surface temperature (*SST*) as an additional predictor.

3. Results

3.1. Selection of explanatory variables

The backward selection of explanatory variables resulted in 14 PCs for PCAh and 3 PCs for PCAm. The loadings of the first 6 PCs for the PCAh and PCAm datasets are shown in Fig. 5. The spectral distribution of PC loadings is akin to results from prior similar approaches [33–35], indicating spectral features related to the optical properties of the seawater constituents. The spectral variability of the first PC is driven mainly by the particulate backscattering of the in-water constituents and the absorption of water molecules, and explained more than 96% of the data covariance for both multi- and hyperspectral $R_{rs}'(\lambda)$ datasets. The second PC highlights spectral features related to the absorption by *Chl* at the ocean surface, explaining ~3.5% of the dataset

covariance; and the third PC is driven by the spectral variation of $R_{rs}'(\lambda)$ due to the absorption of accessory pigments and explained $\sim 0.16\%$ of the dataset covariance [33–35]. These first three PCs were similar between hyperspectral and multispectral models, indicating that the most significant $R_{rs}'(\lambda)$ features were captured by multispectral data (Fig. 5).

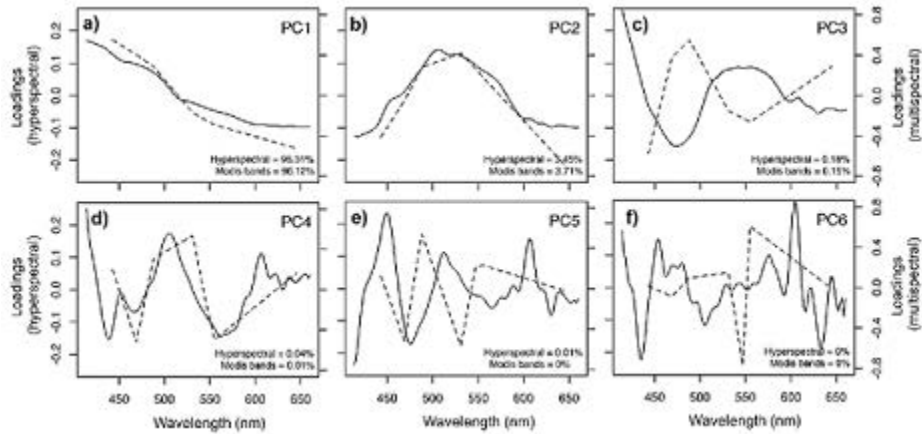


Fig. 5. Spectral distribution of loadings of the first six principal components. Solid lines (primary Y axis) show loadings of the PCA using hyperspectral $R_{rs}'(\lambda)$ (PCAh), whereas dashed lines (secondary Y axis) show those of the PCA using $R_{rs}'(\lambda)$ at the Aqua-MODIS bands (PCAm). Relative (percentage) explanation of the variability of the data by each PC is shown on the bottom right of each plot.

In the multispectral models, PCs 1 and 2 were strong predictors for all targeted picophytoplankton taxa, whereas PC3 was utilized to predict the abundance of *Synechococcus* and picoeukaryotes. For the hyperspectral models, the prediction of *Prochlorococcus* utilized PCs 1 and 2 combined with two other PCs, *Synechococcus* was associated with PCs 1, 2 and 3 in association with seven other PCs, and picoeukaryotes were predicted using PCs 2 and 3 with five additional PCs. In addition to using the PCs' scores as predictors, *SST* was included as a predicting variable and the improvement of the models was evaluated.

3.2. Model performance assessed with the training dataset

3.2.1. *SST* as an additional predictor

Regardless of differences in performance metrics, both hyperspectral and multispectral models are capable of detecting the changes in cell concentrations along the AMT 24 transect. However, hyperspectral based models were superior to multispectral ones regardless of the targeted picophytoplankton group (Table 1). In addition, for Arrangement 1 (section 2.4), the inclusion of *SST* as a predictor considerably improved the performance of both the multispectral and hyperspectral models to predict *Prochlorococcus* when compared to models that only used PCs as predictors (Table 1). For the hyperspectral approach, the *MAE* decreased from 1.49 (49%) to 1.31 (31%) and R^2 increased from 0.42 to 0.82 when *SST* was added to the predictive model of *Prochlorococcus* (Table 1, Fig. 6). Likewise, for the multispectral approach, the *MAE* decreased from 1.46 to 1.33 and R^2 increased from 0.50 to 0.76 (Table 1, Fig. 6). The inclusion of *SST* as a predictor did not improve either the multispectral or hyperspectral models to predict *Synechococcus*, with biases and *MAEs* remaining unchanged (Table 1). For autotrophic picoeukaryotes, uncertainty metrics remained effectively unchanged when considering *SST* for the hyperspectral approach and, in the multispectral model, the *MAE* and R^2 showed an improvement when adding *SST* (1.27 to 1.24 and 0.90 to 0.92, respectively) (Table 1). We opted to use *SST*

as an additional predictor in the models to estimate the abundance of *Prochlorococcus* using both multi- and hyperspectral $R_{rs}'(\lambda)$, and in the model to predict the abundance of autotrophic picoeukaryotes using multispectral $R_{rs}'(\lambda)$.

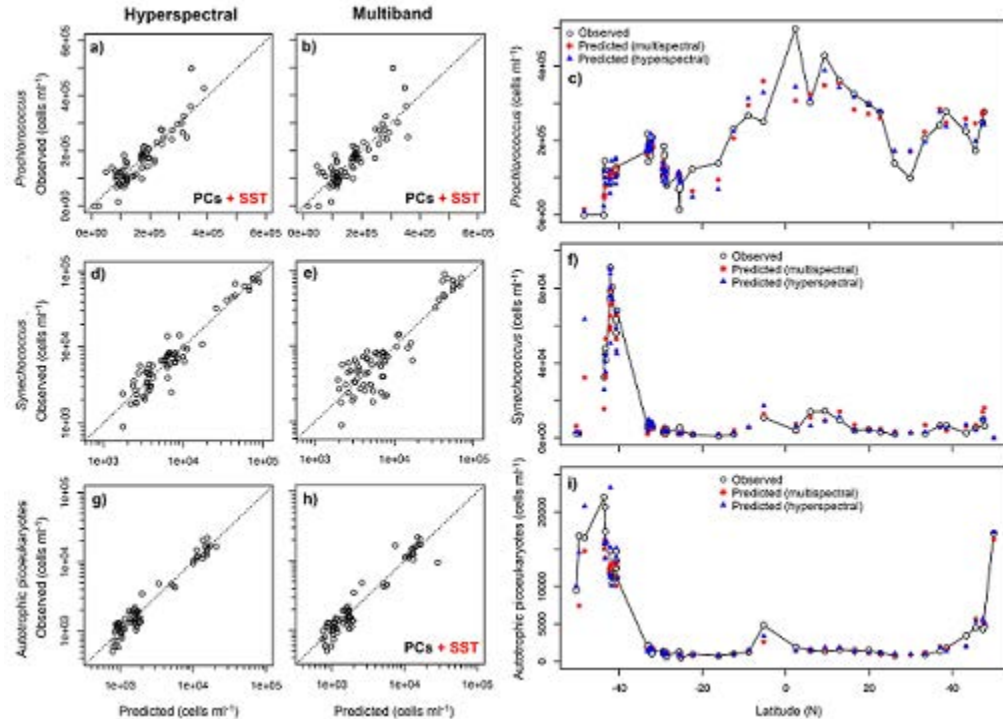


Fig. 6. Performance of developed models (Arrangement 2) for **a-c** *Prochlorococcus* (first row), **d-f** *Synechococcus* (second row) and **g-i** autotrophic picoeukaryotes (third row) cell abundance, using PCAm and PCAh approaches. In panels **d-e** and **g-h**, abundances of *Synechococcus* and picoeukaryotes are plotted in log10 scale as this transformation was implemented for model development. *SST* was used as an additional predictor for both *Prochlorococcus* models and the multispectral picoeukaryotes model since it was found to improve performance (Table 1). Regardless of difference in performance metrics, both hyperspectral and multispectral *in-situ* models are capable of detecting the changes in cell concentrations along the AMT 24 transect (**c,f,i**).

3.2.2. Multispectral versus hyperspectral cross-validation

Model performance improved when using hyperspectral $R_{rs}'(\lambda)$ compared to consideration of only Aqua-MODIS bands (see Fig. 6, Tables 1 and 2). For *Synechococcus* abundance estimation, biases were negligible (Table 2) while multispectral *MAEs* exceeded hyperspectral *MAEs* in both Arrangements 1 (full-fit) and 2 (cross-validation) (1.45 vs. 1.27 and 1.50 vs. 1.36, respectively). For the prediction of *Prochlorococcus* and picoeukaryote abundances, the hyperspectral biases and *MAEs* were also reduced relative to their multispectral counterparts for both Arrangements 1 and 2 (Table 2). Finally, the R^2 for predicting *Prochlorococcus*, *Synechococcus*, and autotrophic picoeukaryote abundances increased by 6% on average when using hyperspectral approach compared to the multispectral approach. Nevertheless, and despite underperforming relative to the hyperspectral approach, patterns in the latitudinal variability in the abundance of these groups

were still reasonably captured by the multispectral approach, using *SST* when applicable, across the full dynamic range of cell concentrations for each phytoplankton group (see Fig. 6).

3.3. Model implementation using satellite data (Aqua-MODIS)

3.3.1. Satellite retrievals from AMT cruises

Assessment of our multispectral model using 8-day Aqua-MODIS $R_{rs}'(\lambda)$ and *SST* imagery (September 30th to October 7th, 2014) as input yielded reasonable retrievals of cell concentrations when compared to *in-situ* samples collected during the AMT24 cruise (Table 3). The *MAE* of 1.37 for *Prochlorococcus*, 2.04 for *Synechococcus* and 1.28 for picoeukaryotes was higher than the one encountered for *in-situ* $R_{rs}'(\lambda)$ data (Table 1), indicating a degradation in performance when moving to the satellite $R_{rs}'(\lambda)$. The *bias* in *Prochlorococcus* prediction remained around 1.09 (9%) when using Aqua-MODIS $R_{rs}(\lambda)$, similar to that using *in-situ* $R_{rs}'(\lambda)$ measurements. However, increases in the *bias* of *Synechococcus* (0.62 (−38%) from Aqua-MODIS and ~ 1 (~ 0%) from *in-situ* $R_{rs}'(\lambda)$) and picoeukaryote retrievals (0.91 (−9%) from Aqua-MODIS and 1 (0%) from *in-situ* $R_{rs}(\lambda)$) were more evident. These underestimations of cell abundances when using satellite data are likely associated with the “patchy” nature of their spatial distribution, further augmented by mismatch between *in-situ*/satellite sampling times and areas (1.6 ml discrete sample vs. 8 days/4 km composites) (Fig. 7).

The temporal portability of multispectral models was assessed using cell abundance predictions computed from Aqua-MODIS data retrieved from sampling time/locations of AMTs 20, 22, 23, 25 and 28. *Prochlorococcus* abundance was overestimated in these 5 AMT cruises, as indicated by the increase in *MAE* (2.26) and in *bias* (1.75) when compared to satellite retrievals from AMT24 (*MAE* = 1.37, *bias* = 1.09), especially in the North Atlantic (Fig. 8). The *Synechococcus* model predictions also showed a higher *MAE* (2.20) compared to AMT24 (2.04), whereas picoeukaryotes *MAE* increased from 1.28 on AMT24 to 1.53 for the other five AMTs with a *bias* decreasing slightly from 0.91 (− 9%) to 1.05 (5%) (Table 3).

3.3.2. Implementation using satellite imagery

The spatial distribution of these picophytoplankton groups captured by our models is shown in Fig. 9. Satellite predictions show highest abundances of *Prochlorococcus* at the Equatorial Convergence Zone and lowest abundances in the ocean gyres (despite still being higher than other phytoplankton), with an increase towards the high-latitude edges of both North and South Atlantic subtropical gyres. Despite the low abundance, *Prochlorococcus* numerically dominated the picophytoplankton in the gyres. *Synechococcus* showed highest abundances at the high-latitude edges of the ocean gyres. Autotrophic picoeukaryotes were most abundant in higher latitudes (> 45° N and S) showing similar patterns to the distribution of *Chl* (see Fig. 2), with the constraint of *Chl* concentrations being lower than 1 mg m^{−3}, since chlorophyll concentrations never reached values higher than this in the present dataset. Satellite visualization of model outputs allowed us to detect the picophytoplankton community zonation at the high-latitude gyre edges (i.e. *Prochlorococcus*-*Synechococcus*-picoeukaryotes from the inner gyres towards higher latitudes), as observed in *in-situ* measurements (see Fig. 1), demonstrating the potential use of our approach for the evaluation of ecosystem and biogeochemical models.

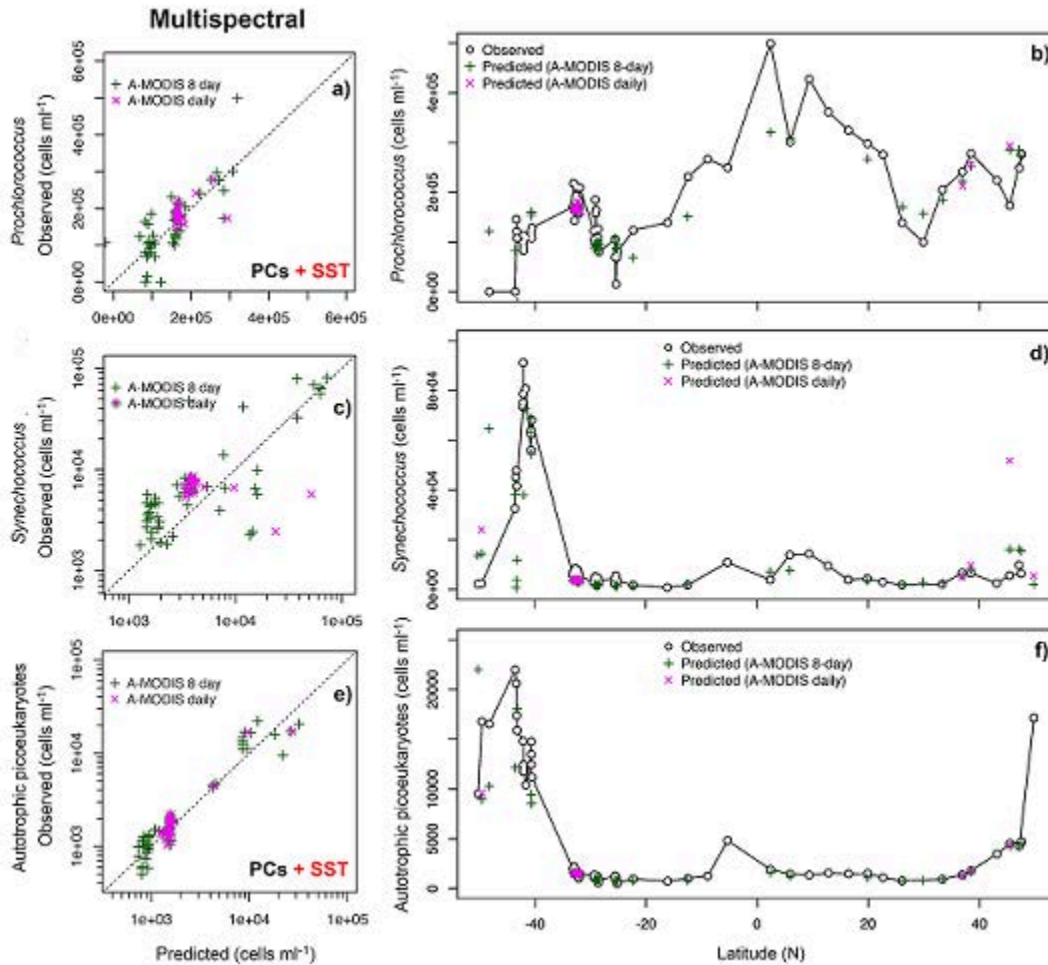


Fig. 7. Performance of developed models (PCA_m approach) for **a,b**) *Prochlorococcus* (first row), **c,d**) *Synechococcus* (second row) and **e,f**) autotrophic picoeukaryotes (third row) cell abundance, implemented using Aqua-MODIS retrievals for the cruise AMT24. In panels **c** and **e**, abundances of *Synechococcus* and picoeukaryotes are plotted in \log_{10} scale because this transformation was implemented for model development. SST was used as an additional predictor for models to predict *Prochlorococcus* and autotrophic picoeukaryotes.

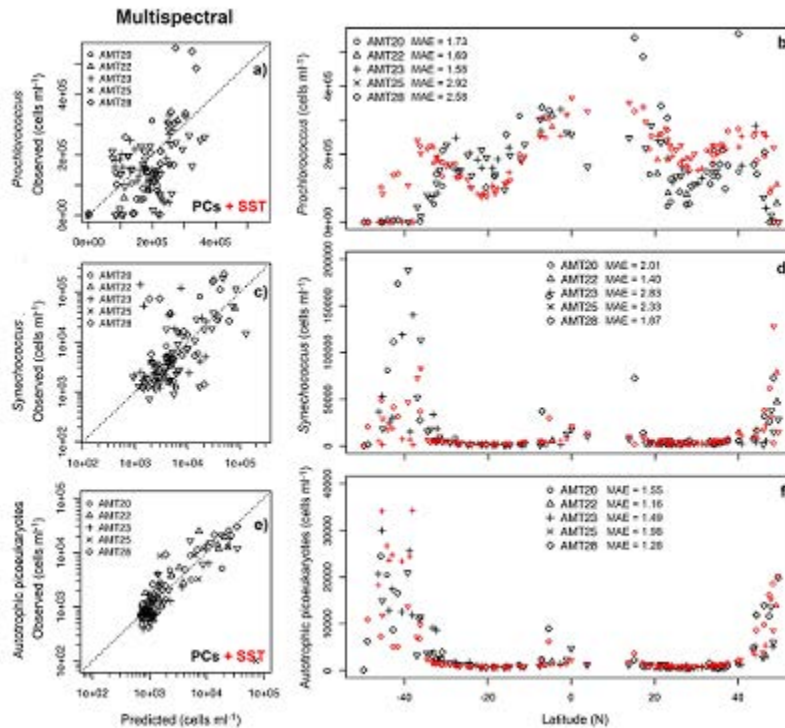


Fig. 8. Performance of developed models (PCA_m approach) for **a,b** *Prochlorococcus* (first row), **c,d** *Synechococcus* (second row) and **e,f** autotrophic picoeukaryotes (third row) cell abundance, implemented using 8-day Aqua-MODIS retrievals for AMTs 20, 22, 23, 25 and 28. In panels **b**, **d**, and **f**, black symbols indicate *in-situ* observations, while red markers indicate values retrieved using species-specific model from Aqua-MODIS, and specific MAE of modelled values from each cruise is shown. In panels **c** and **e**, abundances of *Synechococcus* and picoeukaryotes are shown in log scale, as this transformation was used in model development. *SST* was used as an additional predictor for models to predict *Prochlorococcus* and autotrophic picoeukaryotes.

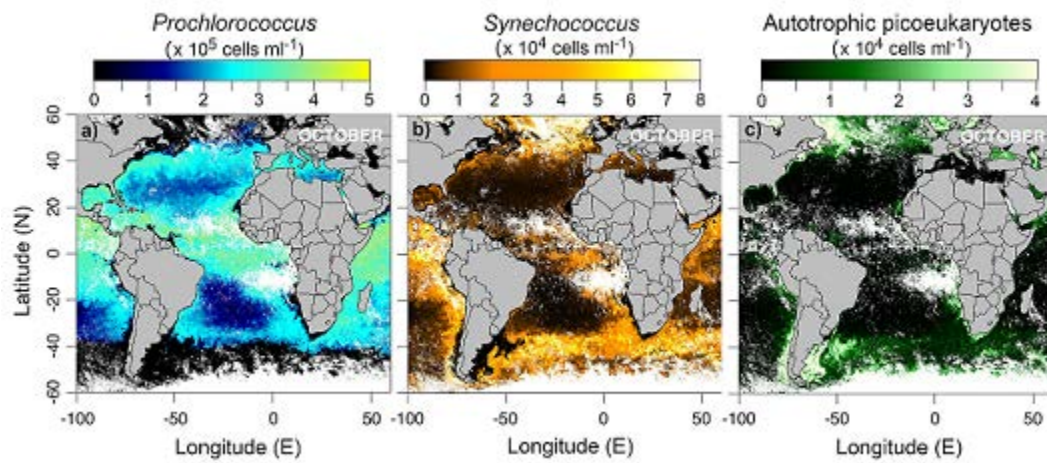


Fig. 9. Aqua-MODIS monthly composites (October 2014) showing cell abundances (cells ml^{-1}) of **a)** *Prochlorococcus*, **b)** *Synechococcus* and **c)** autotrophic picoeukaryotes at the sea surface.

4. Discussion

Principal component regression analysis provides a powerful tool to retrieve optically-significant marine variables from hyperspectral radiometry by exploring spectral variations in $R_{rs}(\lambda)$ [33,35]. With regard to assessing phytoplankton community composition, this method has been implemented most frequently in areas of high phytoplankton biomass, where changes in phytoplankton composition and biomass provide significant changes in phytoplankton absorption that are reflected in spectral variations in $R_{rs}(\lambda)$ [33,35,60]. The highest picophytoplankton abundances occur in the stable oligotrophic ocean, where the spectral signature of water is influenced not only by the present cells but also by other seawater constituents that co-vary with their abundances such as the absorption of colored dissolved organic matter and backscattering of heterotrophic bacteria, both of which alter the magnitude and shape of $R_{rs}(\lambda)$. Considering this, our analysis captures the associations between changes in ocean color and the abundance of the smallest phytoplankton, namely *Prochlorococcus*, *Synechococcus*, and autotrophic picoeukaryotes. In the PCA made with $R_{rs}(\lambda)$ spectra from the Atlantic Ocean (AMT24) and concurrent cell counts, the first three principal components displayed spectral features directly or indirectly correlated with the abundance of these taxa. For example, PC1 shows $R_{rs}(\lambda)$ features likely attributed to the backscatter slope and the spectral shape of the absorption of water molecules, having similar shape to the first PC of PCAs from hyperspectral $R_{rs}(\lambda)$ spectra of meso- and eutrophic waters [33–35]. This first PC was highly correlated with highest *Prochlorococcus* abundances and lowest abundances of larger phytoplankton cells, meaning *Prochlorococcus* is most abundant in waters where the shape of the $R_{rs}(\lambda)$ spectrum is most similar to that of PC1, thus having lower influence of the absorption of *Chl*, accessory pigments and other in-water constituents (i.e., oligotrophic waters). PCs 2 to 4 were associated with the presence of accessory pigments and higher *Chl* absorption, present in *Synechococcus* and autotrophic picoeukaryotic cells.

Increasing the spectral resolution of $R_{rs}(\lambda)$ substantially improved the prediction of all targeted groups (see Tables 1 and 2). Hyperspectral $R_{rs}(\lambda)$ provides greater information content on oceanic constituents contributing to variability in the optical signal, in particular taxon-specific light-absorbing photosynthetic pigments. Pigment-specific light absorption imposes spectral features of several nanometers in distance, leading to variations in the spectral shape of $R_{rs}(\lambda)$ signal [23,38]. As a consequence, our hyperspectral approach resulted in a higher number of usable principal components (predictors) than our multispectral approach, ultimately increasing the performance of the hyperspectral predictive models, in particular for *Synechococcus*. This result agrees with several previous comparisons of hyperspectral and multispectral algorithms that demonstrate how increasing the spectral resolution of the $R_{rs}(\lambda)$ signal improves predictive models for some phytoplankton taxa [33–35,38, 61].

Remotely-sensed physical ocean properties such as *SST* can be useful to further constrain empirical models that predict algal abundance. *SST* can be used as a powerful predictor for the accumulation of cells when direct or indirect relationships between *SST* and certain ecological conditions that favor the target taxon are well known, as previously demonstrated for the prediction of blooms of the harmful dinoflagellate *Alexandrium fundyense* in the Bay of Fundy [61] and blooms of the diatom *Pseudo-nitzschia* in Chesapeake Bay [62], and in other predicting models for the biomass of specific phytoplankton groups [30,63,64]. In our study, the inclusion of *SST* was relevant for predicting the abundances of *Prochlorococcus* and picoeukaryotes, as ecological niches of both taxa are extremely constrained by temperature [65–67]. *Prochlorococcus* is most abundant in environments with high water column stability [68–70], which is usually associated with high *SST* [71], whereas picoeukaryotes grow next to the transition between oligo- and mesotrophic waters [41,72], where *SST* is typically slightly lower than at the center of the gyres [73]. The inclusion of *SST* as a predictor was especially useful for improving multispectral models.

The performance of empirical models such as the ones presented here are highly dependent on the training datasets. For example, inclusion of a dataset collected at higher spatial frequency across the frontal region in the South Atlantic allowed for a larger dynamic range in the training dataset, yielding better retrievals for phytoplankton taxa that occur in high-abundance patches such as *Synechococcus* at the frontal system of the South Atlantic gyre southern boundary [8,29]. When we retrained the multispectral model using only samples collected on CTD casts (sparse sampling strategy), *Synechococcus* cell abundances were underestimated in these patches as sparse sampling missed small pockets of high *Synechococcus* abundances, thus not capturing the full range of *Synechococcus* cell concentrations. The increased number of samples across the *Synechococcus* patch reduced the retrieval bias from 0.71 (−29%) to ~ 1 ($\sim 0\%$) when using multispectral $R_{rs}(\lambda)$ and from 0.84 (−16%) to ~ 1 ($\sim 0\%$) when using hyperspectral $R_{rs}(\lambda)$, whereas MAE was reduced from 1.67 (67%) to 1.45 (45%) in the multispectral model and from 1.37 (37%) to 1.27 (27%) using the hyperspectral approach. As an empirical model is only good at predicting cell abundances within the cell number range of its training dataset, this result highlights the importance of understanding the scales of cell abundances and its spatial distribution patterns for the targeted phytoplankton taxon when assembling data to train empirical models. Proper design of *in-situ* sampling plans must cover the full dynamic range of cell abundances of that particular taxon. Similarly, vertical sampling needs consideration in such analyses given that *in situ* sampling does not always represent the spectrally-dependent depth range considered in the satellite retrieval. We considered the top 10 m of the water column in these analyses, which does not consider the full euphotic zone in our areas of interest, but does encompass a reasonable fraction of the optically weighted signal observed over the first e-folding depth [74].

Satellite implementation of the empirical models to monthly composites of Aqua-MODIS $R_{rs}(\lambda)$ and SST provided a qualitative view of the spatial and temporal distributions (see Fig. 9) of targeted taxa, if only to provide a visual case study to assess the portability of our model. For *Prochlorococcus*, our model predicts highest surface abundances at the edges of the ocean gyres and Equatorial Convergence, showing similar distribution to that of *in-situ* observations from AMT cruises (see Fig. 1) [8,14,52]. This *Prochlorococcus* distribution pattern agrees with predictions of other ocean color-based models, such as Alvain et al. [75], El-Hourany et al. [76], and Xi et al. [31], and the model of Lange et al. [28] which combines ocean color information with environmental variables. In turn, a model based solely on environmental variables (SST, photosynthetically-active radiation - PAR) – i.e. Flombaum et al. [29] – estimates highest *Prochlorococcus* abundances in western boundary currents such as the Gulf Stream and the Brazil Current because, in this model, SST is the most important driver of the distribution of *Prochlorococcus*. SST is a powerful predictor of *Prochlorococcus* [65–67], possibly due to its causal relationship with water column stratification [71] which favors the growth of this cyanobacterium [70]. Stratification induces oligotrophy, avoiding the growth of microbial assemblages that include herbivores of *Prochlorococcus* [77]. The direct relationship between *Prochlorococcus* and water column stability, rather than temperature, would justify the high *Prochlorococcus* abundances found in the Mediterranean Sea [78], and its absence in polar regions where stratification is seasonal or episodic. While SST may fail to predict the presence of *Prochlorococcus* in regions where salinity is important in driving stratification, ocean color variables such as $R_{rs}(\lambda)$ provide direct observation of the surface water components. $R_{rs}(\lambda)$ and spectral phytoplankton absorption coefficients ($a_{ph}(\lambda)$) provide refined information on the presence of optically-relevant phytoplankton, which are abundant in the absence of *Prochlorococcus*. In other words, *Prochlorococcus* is most abundant where the optical influence of phytoplankton on the $R_{rs}(\lambda)$ spectrum is minimal. However, concurrent high abundances of *Prochlorococcus* and other phytoplankton groups (such as diatoms, nano- and picoeukaryotes) occur in areas where nutrient input is high despite high stratification levels (i.e., high SST), such as the Equatorial

Convergence Zone [8,79]. This explains the best performance of our *Prochlorococcus* model when using ocean color information and *SST* as predictors.

Regarding *Synechococcus* estimates, our ocean color-based model finds highest abundances at the high-latitude edges of the ocean gyres, especially the South Atlantic gyre, surrounding possible blooms of larger phytoplankton cells such as coccolithophorids, similar to predictions based on *SST* and *PAR* [29]. Highest abundances of autotrophic picoeukaryotes were found at the higher latitude edges of the ocean gyres ($> 45^\circ$ N and S), mimicking patterns seen in the *Chl* distribution. However, picoeukaryotic populations slightly decrease where *Chl* concentrations reach values of $\sim 1 \text{ mg m}^{-3}$. Such spatial and temporal patterns highlight the importance of these picophytoplankton taxa as proxies for certain ecosystems or trophic conditions. For example, high abundances of *Prochlorococcus* delineate the extension of the ocean gyres, and *Synechococcus* becomes abundant in a narrow band at the transition between oligotrophic (i.e. South Atlantic gyre) and mesotrophic waters (i.e. temperate waters of higher latitudes where pico- and nanophytoplankton bloom), as also observed in several studies [8,14,18,41]. It is important to note that our model estimates cell abundances, which are highly correlated with group-specific carbon biomass but not always with pigment concentrations because of photophysiological adaptations of picophytoplankton cells to the different environmental conditions found across oceanic fronts [8,41,80–83].

In a similar way, we hypothesize that the inclusion of datasets from other parts of the ocean outside the Atlantic would improve the global model and allow for basin-specific tuning. Such models could allow for a segregated assessment of the photophysiological and optical characteristics of basin-specific ecotypes of the picocyanobacteria and picoeukaryotic flora, ultimately improving the performance of these empirical models. Furthermore, the ability of models to retrieve abundances of *Synechococcus* and autotrophic picoeukaryotes could be improved by including datasets from coastal and/or high *Chl* areas ($> 1 \text{ mg m}^{-3}$), allowing for a merged approach (similar to NASA's current operational *Chl* algorithm). In these waters, the contribution of *CDOM* and carotenoids in large phytoplankton to the spectral variability of $R_{rs}(\lambda)$ is higher, diminishing the relative influence of picophytoplankton cells. However, the spectral characteristics of these two groups are different in complex waters: *Synechococcus* ecotypes display different concentrations of accessory pigments to adapt to different optical niches [84–87], although they all contain phycobiliprotein complexes which are rather unique and likely to be detected by the PCA; and the taxonomic composition of autotrophic picoeukaryote communities is highly variable according to nutrient availability, temperature and stratification [41,88]. This could also deteriorate finding robust models for the specific groups. Xi et al. [31] used a large global matchup dataset for setting up similar Empirical Orthogonal Function (EOF) models with pigments (measured using HPLC) and satellite $R_{rs}(\lambda)$ data. While eukaryotic phytoplankton groups were very well predicted globally, the prediction skill of *Prochlorococcus* and *Synechococcus* was rather poor.

Observed changes in model performance between ocean basins or different Atlantic cruises may be expected and could stem from multiple sources. First, the occurrence of distinct ecotypes of *Prochlorococcus* and *Synechococcus* and combinations of picoeukaryotic taxa in each ocean basin, and their associated optical properties (due to the physiological acclimation and/or evolutionary adaptation) might have made our model specific to the Atlantic Ocean during the AMT sampling season(s) only. Second, the relationships between group-specific cell abundances and the $R_{rs}(\lambda)$ signature can be influenced by the structure of the ecosystem itself – that is, the presence of other phytoplankton cells (e.g., diatoms in the Equatorial Convergence Zone), or other optically-active water constituents (e.g., *CDOM* and non-algal particles). Differences in ecosystem structure, specifically in the top-down control and other loss pathways for these phytoplankton populations, could also potentially influence model predictions. In addition, flow cytometric cell counts enable a precise determination of the abundance of picophytoplankton

groups, which can be converted to carbon biomass [81,82], and do not depend on models and their associated uncertainties to attribute group-specific biomass from marker pigments. However, the use of marker pigments as proxies for phytoplankton taxa is most directly linked to the observed change in the $R_{rs}(\lambda)$ spectrum, and also provide estimates of the contribution of larger phytoplankton to the total phytoplankton biomass and its influence in the $R_{rs}(\lambda)$ spectrum, which can be useful for analysis interpretation. Lastly, while methods used to collect $R_{rs}(\lambda)$ for this study followed similar community-approved procedures, approaches used to quantify the cell abundances on different oceanographic expeditions differ, potentially adding to differing validation performances when comparing outputs of the model with alternate datasets where different flow cytometric procedures were adopted (i.e., Olson et al. [48] versus Zubkov et al. [47] for quantifying *Prochlorococcus* and *Synechococcus*).

Since the goal of the model is to detect the large-scale spatial variability in open ocean waters, where picophytoplankton cells are most abundant, the model has not been tested in shelf seas and coastal waters. We expect that the models will need to be retuned for such waters because the presence of suspended sediments and *CDOM* will change the spectral distribution of the eigenvectors of each principal component.

5. Summary and conclusions

Cell abundances of *Prochlorococcus*, *Synechococcus* and autotrophic picoeukaryotes were estimated in surface waters of the Atlantic Ocean using empirical models based on a combination of *SST* and the scores of an $R_{rs}(\lambda)$ principal component analysis, which captured the association between changes in ocean color and the abundance of these picophytoplankton groups. These models were implemented using satellite data (Aqua-MODIS), which allowed us to estimate cell abundances on a basin scale. Although these phytoplankton types occur in high abundances in oligotrophic oceans, the spectral signature of waters inhabited by these cells is highly influenced by their optical attributes and other water constituents that co-vary with their abundance, such as the absorption of *CDOM* and backscattering of heterotrophic bacteria, which modify the magnitude and shape of the $R_{rs}(\lambda)$ spectrum, being expressed in different PCs of the PCA.

The extension of the predictive models to a basin scale is feasible because of the broad swath of the reference AMT *in-situ* dataset, which covers a large range of marine environments, including the North and South Atlantic gyres where picoplankton are dominant, and the Equatorial Convergence Zone where pico-sized cells are abundant but share the environment with larger phytoplankton. Along the AMT transect, model estimates successfully demonstrate the expected distributions of *Prochlorococcus* in gyres, with higher cell concentrations at the Equatorial Convergence and near the gyre edges. The model shows the emergence of autotrophic picoeukaryotes where *Chl* concentrations increase, and latitudinal changes in the abundance of *Synechococcus* showing high-abundance patches in areas of trophic transition such as between the ocean gyres and mesotrophic waters of higher latitudes.

Our model successfully predicts the abundance of *Prochlorococcus*, *Synechococcus* and autotrophic picoeukaryotic cells in the surface oceans using remote-sensing reflectance and sea surface temperature. The models using hyperspectral $R_{rs}(\lambda)$ substantially improved the prediction of *Prochlorococcus* when compared to the multispectral model. The sampling strategy to generate an appropriate dataset to develop a predictive algorithm targeted to a phytoplankton group must be designed according to the scale of spatial variability of this group; for example, in the case of *Synechococcus* accurate algorithm retrievals necessitate fine spatial sampling to detect the full abundance range including elevated cell concentrations along transition zones between oligotrophic and mesotrophic waters. Thus, consideration of previous knowledge about the biology and ecology of the target phytoplankton group is required.

Funding

National Aeronautics and Space Administration (NNX13AC42G); Natural Environment Research Council (NE/R015953/1).

Acknowledgments

The authors are thankful for all the scientists that contributed to collection of this dataset, and captains and crews of all the research vessels that supported these and our other sea-going adventures. We thank the NASA Ocean Biology Processing Group (OBPG) for providing the satellite imagery and operational support. We also thank the British Oceanographic Data Centre (BODC) for providing *in-situ* data from the AMT cruises. This study contributes to the international IMBeR project and is contribution number 353 of the AMT programme. Finally, we are thankful to Erdem Karakoylu for valuable advice on the use of statistical methods.

Disclosures

The authors declare no conflicts of interest.

References

1. J. M. Sieburth, V. Smetacek, and J. Lenz, "Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions," *Limnol. Oceanogr.* **23**(6), 1256–1263 (1978).
2. W. K. W. Li, "Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting," *Limnol. Oceanogr.* **39**(1), 169–175 (1994).
3. W. K. W. Li, "Macroecological patterns of phytoplankton in the northwestern North Atlantic Ocean," *Nature* **419**(6903), 154–157 (2002).
4. T. S. Kostadinov, S. Milutinovi, I. Marinov, and A. Cabré, "Carbon-based phytoplankton size classes retrieved via ocean color estimates of the particle size distribution," *Ocean Sci.* **12**(2), 561–575 (2016).
5. P. Raimbault, N. Garcia, and F. Cerutti, "Distribution of inorganic and organic nutrients in the South Pacific Ocean - evidence for long-term accumulation of organic matter in nitrogen-depleted waters," *Biogeosciences* **5**(2), 281–298 (2008).
6. E. Marañón, "Cell size as a key determinant of phytoplankton metabolism and community structure," *Ann. Rev. Mar. Sci.* **7**(1), 241–264 (2015).
7. A. Landolfi, H. Dietze, and G. Volpe, "Longitudinal variability of organic nutrients in the North Atlantic subtropical gyre," *Deep Sea Res., Part I* **111**, 50–60 (2016).
8. M. V. Zubkov, M. A. Sleight, G. A. Tarran, P. H. Burkill, and R. J. G. Leakey, "Picoplanktonic community structure on an Atlantic transect from 50 degrees N to 50 degrees S," *Deep Sea Res., Part I* **45**(8), 1339–1355 (1998).
9. R. J. W. Brewin, G. H. Tilstone, T. Jackson, T. Cain, P. I. Miller, P. K. Lange, A. Misra, and R. L. Airs, "Modelling size-fractionated primary production in the Atlantic Ocean from remote sensing," *Prog. Oceanogr.* (2017).
10. M. W. Lomas and S. B. Moran, "Evidence for aggregation and export of cyanobacteria and nano-eukaryotes from the Sargasso Sea euphotic zone," *Biogeosciences* **8**(1), 203–216 (2011).
11. Z. I. Johnson and Y. Lin, "*Prochlorococcus*: Approved for export," *Proc. Natl. Acad. Sci.* **106**(26), 10400–10401 (2009).
12. T. L. Richardson, "Mechanisms and pathways of small-phytoplankton export from the surface ocean," *Ann. Rev. Mar. Sci.* (2019).
13. M. V. Zubkov, M. A. Sleight, P. H. Burkill, and R. J. G. Leakey, "Picoplankton community structure on the Atlantic Meridional Transect: a comparison between seasons," *Prog. Oceanogr.* **45**(3-4), 369–386 (2000).
14. J. L. Heywood, M. V. Zubkov, G. A. Tarran, B. M. Fuchs, and P. M. Holligan, "Prokaryoplankton standing stocks in oligotrophic gyre and equatorial provinces of the Atlantic Ocean: Evaluation of inter-annual variability," *Deep Sea Res., Part II* **53**(14-16), 1530–1547 (2006).
15. M. J. W. Veldhuis, K. R. Timmermans, P. Croot, and B. van der Wagt, "Picophytoplankton; a comparative study of their biochemical composition and photosynthetic properties," *J. Sea Res.* **53**(1-2), 7–24 (2005).
16. D. Vaulot, W. Eikrem, M. Viprey, and H. Moreau, "The diversity of small eukaryotic phytoplankton ($\leq 3 \mu\text{m}$) in marine ecosystems," *FEMS Microbiol. Rev.* **32**(5), 795–820 (2008).
17. F. Partensky, W. R. Hess, and D. Vaulot, "*Prochlorococcus*, a marine photosynthetic prokaryote of global significance," *Microbiol. Mol. Biol. Rev.* **63**(1), 106–127 (1999).
18. A. Bracher, H. Xi, T. Dinter, A. Mangin, V. Strass, W.-J. von Appen, and S. Wiegmann, "High resolution water column phytoplankton composition across the Atlantic Ocean from ship-towed vertical undulating radiometry," *Front. Mar. Sci.* **7**, 235 (2020).

19. W. M. Balch, B. C. Bowler, D. T. Drapeau, L. C. Lubelczyk, E. Lyczkowski, C. Mitchell, and A. Wyeth, "Coccolithophore distributions of the North and South Atlantic Ocean," *Deep Sea Res., Part I* **151**(June), 103066 (2019).
20. B. Mouriño-Carballido, E. Hojas, P. Cermeño, P. Chouciño, B. Fernández-Castro, M. Latasa, E. Marañón, X. A. G. Morán, and M. Vidal, "Nutrient supply controls picoplankton community structure during three contrasting seasons in the northwestern Mediterranean Sea," *Mar. Ecol.: Prog. Ser.* **543**, 1–19 (2016).
21. C. B. Mouw, N. J. Hardman-Mountford, S. Alvain, A. Bracher, R. J. W. Brewin, A. Bricaud, A. M. Ciotti, E. Devred, A. Fujiwara, T. Hirata, T. Hirawake, T. S. Kostadinov, S. Roy, and J. Uitz, "A consumer's guide to satellite remote sensing of multiple phytoplankton groups in the global ocean," *Front. Mar. Sci.* **4**(FEB), 1 (2017).
22. A. Bracher, H. A. Bouman, R. J. W. Brewin, A. Bricaud, V. Brotas, A. M. Ciotti, L. Clementson, E. Devred, A. Di Cicco, S. Dutkiewicz, N. J. Hardman-Mountford, A. E. Hickman, M. Hieronimi, T. Hirata, S. N. Losa, C. B. Mouw, E. Organelli, D. E. Raitos, J. Uitz, M. Vogt, and A. Wolanin, "Obtaining phytoplankton diversity from ocean color: A scientific roadmap for future development," *Front. Mar. Sci.* **4**(MAR), 1–15 (2017).
23. IOCCG, "Phytoplankton functional types from Space.," Reports Monogr. Int. Ocean. Coord. Gr. (2014).
24. C. D. Mobley, J. Werdell, B. Franz, Z. Ahmad, and S. Bailey, "Atmospheric Correction for Satellite Ocean Color Radiometry," *NASA Tech. Memo.* (2016).
25. A. Bracher, M. Vountas, T. Dinter, J. P. Burrows, R. Röttgers, and I. Peeken, "Quantitative observation of cyanobacteria and diatoms from space using PhytoDOAS on CIAMACHY data," *Biogeosciences* **6**(5), 751–764 (2009).
26. R. J. W. Brewin, S. Sathyendranath, G. Tilstone, P. K. Lange, and T. Platt, "A multicomponent model of phytoplankton size structure," *J. Geophys. Res.: Oceans* **119**(6), 3478–3496 (2014).
27. J. Uitz, H. Claustre, B. Gentili, and D. Stramski, "Phytoplankton class-specific primary production in the world's oceans: Seasonal and interannual variability from satellite observations," *Global Biogeochem. Cycles* **24**(3), 1 (2010).
28. P. K. Lange, R. Brewin, G. Dall'Olmo, G. Tarran, S. Sathyendranath, M. Zubkov, and H. Bouman, "Scratching Beneath the Surface: A Model to Predict the Vertical Distribution of *Prochlorococcus* Using Remote Sensing," *Remote Sens.* **10**(6), 847 (2018).
29. P. Flombaum, J. L. Gallegos, R. A. Gordillo, J. Rincon, L. L. Zabala, N. Jiao, D. M. Karl, W. K. W. Li, M. W. Lomas, D. Veneziano, C. S. Vera, J. A. Vrugt, and A. C. Martiny, "Present and future global distributions of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*," *Proc. Natl. Acad. Sci.* **110**(24), 9824–9829 (2013).
30. T. S. Moore and C. W. Brown, "Incorporating environmental data in abundance-based algorithms for deriving phytoplankton size classes in the Atlantic Ocean," *Remote Sens. Environ.* (2020).
31. H. Xi, S. N. Losa, A. Mangin, M. A. Soppa, P. Garnesson, J. Demaria, Y. Liu, O. H. F. d'Andon, and A. Bracher, "Global retrieval of phytoplankton functional types based on empirical orthogonal functions using CMEMS GlobColour merged products and further extension to OLCI data," *Remote Sens. Environ.* (2020).
32. National Research Council, *Assessing the Requirements for Sustained Ocean Color Research and Operations* (National Academies Press, 2011).
33. S. E. Craig, C. T. Jones, W. K. W. Li, G. Lazin, E. Horne, C. Caverhill, and J. J. Cullen, "Deriving optical metrics of coastal phytoplankton biomass from ocean colour," *Remote Sens. Environ.* **119**, 72–83 (2012).
34. M. Soja-Woźniak, S. E. Craig, S. Kratzer, B. Wojtasiewicz, M. Darecki, and C. T. Jones, "A novel statistical approach for ocean colour estimation of inherent optical properties and cyanobacteria abundance in optically complex waters," *Remote Sens.* **9**(4), 343 (2017).
35. A. Bracher, M. H. Taylor, B. Taylor, T. Dinter, R. Röttgers, and F. Steinmetz, "Using empirical orthogonal functions derived from remote-sensing reflectance for the prediction of phytoplankton pigment concentrations," *Ocean Sci.* **11**(1), 139–158 (2015).
36. S. L. Palacios, R. M. Kudela, L. S. Guild, K. H. Negrey, J. Torres-Perez, and J. Broughton, "Remote sensing of phytoplankton functional types in the coastal ocean from the HypSPIRI Preparatory Flight Campaign," *Remote Sens. Environ.* **167**, 269–280 (2015).
37. H. Xi, M. Hieronimi, R. Röttgers, H. Krasemann, and Z. Qiu, "Hyperspectral differentiation of phytoplankton taxonomic groups: A comparison between using remote sensing reflectance and absorption spectra," *Remote Sens.* **7**(11), 14781–14805 (2015).
38. A. P. Chase, E. Boss, I. Cetinić, and W. Slade, "Estimation of phytoplankton accessory pigments from hyperspectral reflectance spectra: toward a global algorithm," *J. Geophys. Res.: Oceans* **122**(12), 9725–9743 (2017).
39. J. Uitz, D. Stramski, R. A. Reynolds, and J. Dubranna, "Assessing phytoplankton community composition from hyperspectral measurements of phytoplankton absorption coefficient and remote-sensing reflectance in open-ocean environments," *Remote Sens. Environ.* **171**, 58–74 (2015).
40. P. J. Werdell, M. J. Behrenfeld, P. S. Bontempi, E. Boss, B. Cairns, G. T. Davis, B. A. Franz, U. B. Gliese, E. T. Gorman, O. Hasekamp, K. D. Knobelspiesse, A. Mannino, J. V. Martins, C. R. McClain, G. Meister, and L. A. Remer, "The Plankton, Aerosol, Cloud, Ocean Ecosystem Mission: Status, Science, Advances," *Bull. Am. Meteorol. Soc.* **100**(9), 1775–1794 (2019).
41. G. A. Tarran, J. L. Heywood, and M. V. Zubkov, "Latitudinal changes in the standing stocks of nano- and picoeukaryotic phytoplankton in the Atlantic Ocean," *Deep Sea Res., Part II* **53**(14-16), 1516–1529 (2006).

42. R. J. W. Brewin, G. Dall'Olmo, S. Pardo, V. van Dongen-Vogels, and E. S. Boss, "Underway spectrophotometry along the Atlantic Meridional Transect reveals high performance in satellite chlorophyll retrievals," *Remote Sens. Environ.* **183**, 82–97 (2016).
43. R. J. W. Brewin, G. Dall'Olmo, S. Pardo, V. van Dongen-Vogels, and E. S. Boss, "Underway spectrophotometry along the Atlantic Meridional Transect reveals high performance in satellite chlorophyll retrievals," *Remote Sens. Environ.* (2016).
44. S. B. Hooker, G. Lazin, G. Zibordi, and S. Mclean, "An evaluation of above- and in-water methods for determining water-leaving radiances," *J. Atmos. Ocean. Technol.* (2002).
45. IOCCG, *IOCCG Report Number 01: Minimum Requirements for an Operational, Ocean-Colour Sensor for the Open Ocean* (1998).
46. M. V. Zubkov and P. H. Burkill, "Syringe pumped high speed flow cytometry of oceanic phytoplankton," *Cytometry, Part A* **69A**(9), 1010–1019 (2006).
47. M. V. Zubkov, P. H. Burkill, and J. N. Topping, "Flow cytometric enumeration of DNA-stained oceanic planktonic protists," *J. Plankton Res.* **29**(1), 79–86 (2006).
48. R. J. Olson, E. R. Zettler, and M. D. DuRand, "Phytoplankton analysis using flow cytometry," in *Handbook of Methods in Aquatic Microbial Ecology*, P. F. Kemp, E. B. Sherr, and J. J. Cole, eds. (Lewis Publishers, 1993), pp. 175–186.
49. G. A. Tarran and M. V. Zubkov, "Abundance of microbial phytoplankton through the water column during the AMT24 (JR20140922/JR303) cruise in September–November 2014," Br. Oceanogr. Data Centre, Natl. Oceanogr. Centre, NERC, UK (2020).
50. G. A. Tarran and M. V. Zubkov, "Abundance of microbial phytoplankton through the water column during the AMT22 (JC079) cruise in October–November 2012," Br. Oceanogr. Data Centre, Natl. Oceanogr. Centre, NERC, UK (2020).
51. G. A. Tarran and M. V. Zubkov, "Abundance of microbial phytoplankton through the water column during the AMT23 (JR20131005/JR300) cruise in October–November 2013," Br. Oceanogr. Data Centre, Natl. Oceanogr. Centre, NERC, UK (2020).
52. G. A. Tarran, P. K. Lange, and M. V. Zubkov, "Abundance of microbial phytoplankton through the water column during the AMT25 (JR15001) cruise in September–November 2015," Br. Oceanogr. Data Centre, Natl. Oceanogr. Centre, NERC, UK. (2020).
53. G. A. Tarran and A. May, "Abundance of microbial bacteria and phytoplankton through the water column during the AMT28 (JR18001) cruise in September–October 2018," Br. Oceanogr. Data Centre, Natl. Oceanogr. Centre, NERC, UK (2020).
54. British Oceanographic Data Centre, Atlantic Meridional Transect. "<http://amt-uk.org/Cruises/>," Accessed on 2019/10/30. (2020).
55. NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group, "Moderate Resolution Imaging Spectroradiometer onboard Aqua (Aqua-MODIS) Ocean Color Data, NASA OB. DAAC, Greenbelt, MD, USA.," Accessed on 2019/10/30. (2018).
56. R Core Team, *R: A Language and Environment for Statistical Computing* (2017).
57. B. N. Seegers, R. P. Stumpf, B. A. Schaeffer, K. A. Loftin, and P. J. Werdell, "Performance metrics for the assessment of satellite data products: an ocean color case study," *Opt. Express* (2018).
58. W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Fourth edn (Springer, New York, 2002), 53(March).
59. H. Wickham and W. Chang, "devtools: Tools to make developing R packages easier," U.S. patent R package version 1.12.0 (2016).
60. H. Xi, M. Hieronymi, H. Krasemann, and R. Röttgers, "Phytoplankton group identification using simulated and in situ hyperspectral remote sensing reflectance," *Front. Mar. Sci.* **4**(AUG), 1–13 (2017).
61. E. Devred, J. Martin, S. Sathyendranath, V. Stuart, E. Horne, T. Platt, M. H. Forget, and P. Smith, "Development of a conceptual warning system for toxic levels of *Alexandrium fundyense* in the Bay of Fundy based on remote sensing data," *Remote Sens. Environ.* **211**(October 2017), 413–424 (2018).
62. C. R. Anderson, M. R. P. Sapiano, M. B. K. Prasad, W. Long, P. J. Tango, C. W. Brown, and R. Murtugudde, "Predicting potentially toxigenic *Pseudo-nitzschia* blooms in the Chesapeake Bay," *J. Mar. Syst.* **83**(3–4), 127 (2010).
63. B. A. Ward, "Temperature-correlated changes in phytoplankton community structure are restricted to polar waters," *PLoS One* (2015).
64. R. J. W. Brewin, S. Ciavatta, S. Sathyendranath, T. Jackson, G. Tilstone, K. Curran, R. L. Airs, D. Cummings, V. Brotas, E. Organelli, G. Dall'Olmo, and D. E. Raitsos, "Uncertainty in ocean-color estimates of chlorophyll for phytoplankton groups," *Front. Mar. Sci.* **4**, 104 (2017).
65. E. R. Zinser, Z. I. Johnson, A. Coe, E. Karaca, D. Veneziano, and S. W. Chisholm, "Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean," *Limnol. Oceanogr.* **52**(5), 2205–2220 (2007).
66. Z. I. Johnson, E. R. Zinser, A. Coe, N. P. McNulty, E. M. Woodward, and S. W. Chisholm, "Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients," *Science* **311**(5768), 1737–1740 (2006).
67. A. A. Larkin, S. K. Blinebry, C. Howes, Y. Lin, S. E. Loftus, C. A. Schmaus, E. R. Zinser, and Z. I. Johnson, "Niche partitioning and biogeography of high light adapted *Prochlorococcus* across taxonomic ranks in the North Pacific," *ISME J.* **10**(7), 1555–1567 (2016).

68. H. A. Bouman, T. Platt, S. Sathyendranath, W. K. W. Li, V. Stuart, C. Fuentes-Yaco, H. Maass, E. P. W. Horne, O. Ulloa, V. Lutz, and M. Kyewalyanga, "Temperature as indicator of optical properties and community structure of marine phytoplankton: implications for remote sensing," *Mar. Ecol.: Prog. Ser.* **258**, 19–30 (2003).
69. H. A. Bouman, O. Ulloa, D. J. Scanlan, K. Zwirgmaier, W. K. W. Li, T. Platt, V. Stuart, R. Barlow, O. Leth, L. Clementson, V. Lutz, M. Fukasawa, S. Watanabe, and S. Sathyendranath, "Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes," *Science* **312**(5775), 918–921 (2006).
70. H. A. Bouman, O. Ulloa, R. Barlow, W. K. W. Li, T. Platt, K. Zwirgmaier, D. J. Scanlan, and S. Sathyendranath, "Water-column stratification governs the community structure of subtropical marine picophytoplankton," *Environ. Microbiol. Rep.* **3**(4), 473–482 (2011).
71. A. G. O'Carroll, E. M. Armstrong, H. Beggs, M. Bouali, K. S. Casey, G. K. Corlett, P. Dash, C. Donlon, C. L. Gentemann, J. L. Høyer, A. Ignatov, K. Kabobah, M. Kachi, Y. Kurihara, I. Karagali, E. Maturi, C. J. Merchant, S. Marullo, P. Minnett, M. Pennybacker, B. Ramakrishnan, R. A. A. J. Ramsankaran, R. Santoleri, S. Sunder, S. S. Picart, J. Vázquez-Cuervo, and W. Wimmer, "Observational needs of sea surface temperature," *Front. Mar. Sci.* (2019).
72. A. R. Kirkham, C. Lepère, L. E. Jardillier, F. Not, H. Bouman, A. Mead, and D. J. Scanlan, "A global perspective on marine photosynthetic picoeukaryote community structure," *ISME J.* **7**(5), 922–936 (2013).
73. D. Demory, A. C. Baudoux, A. Monier, N. Simon, C. Six, P. Ge, F. Rigaut-Jalabert, D. Marie, A. Sciandra, O. Bernard, and S. Rabouille, "Picoeukaryotes of the *Micromonas* genus: sentinels of a warming ocean," *ISME J.* (2019).
74. H. R. Gordon and A. Y. Morel, *Remote Assessment of Ocean Color for Interpretation of Satellite Visible Imagery* (Springer-Verlag New York, 1983).
75. S. Alvain, C. Moulin, Y. Dandonneau, and H. Loisel, "Seasonal distribution and succession of dominant phytoplankton groups in the global ocean: A satellite view," *Global Biogeochem. Cycles* **22**(3), 1 (2008).
76. R. El Hourany, M. Abboud-Abi Saab, G. Faour, O. Aumont, M. Crépon, and S. Thiria, "Estimation of secondary phytoplankton pigments from satellite observations using Self-Organizing Maps (SOMs)," *J. Geophys. Res.: Oceans* **124**(2), 1357–1378 (2019).
77. S. Dutkiewicz and C. L. Follett, "The collapse of *Prochlorococcus* populations in the transition between the subtropical and subpolar gyres," in *Ocean Sciences Meeting* (2020).
78. M. L. Pedrotti, L. Mousseau, S. Marro, O. Passafiume, M. Gossaert, and J. P. Labat, "Variability of ultraplankton composition and distribution in an oligotrophic coastal ecosystem of the NW Mediterranean Sea derived from a two-year survey at the single cell level," *PLoS One* (2017).
79. H. Liu, H. A. Nolla, and L. Campbell, "*Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean," *Aquat. Microb. Ecol.* **12**(1), 39–47 (1997).
80. F. Baltar and J. Arístegui, "Fronts at the surface ocean can shape distinct regions of microbial activity and community assemblages down to the bathypelagic zone: the Azores Front as a case study," *Front. Mar. Sci.* **4**, 252 (2017).
81. J. R. Christian and D. M. Karl, "Microbial community structure at the US-Joint Global Ocean Flux Study Station ALOHA: inverse methods for estimating biochemical indicator ratios," *J. Geophys. Res.* (1994).
82. S. E. Baer, M. W. Lomas, K. X. Terpis, C. Mouginit, and A. C. Martiny, "Stoichiometry of *Prochlorococcus*, *Synechococcus*, and small eukaryotic populations in the western North Atlantic Ocean," *Environ. Microbiol.* (2017).
83. F. Martini, S. Neuer, D. Hamill, J. Robidart, and M. W. Lomas, "Clade and strain specific contributions of *Synechococcus* and *Prochlorococcus* to carbon export in the Sargasso Sea," *Limnol. Oceanogr.* (2018).
84. D. J. Scanlan, M. Ostrowski, S. Mazard, A. Dufresne, L. Garczarek, W. R. Hess, A. F. Post, M. Hagemann, I. Paulsen, and F. Partensky, "Ecological genomics of marine picocyanobacteria," *Microbiol. Mol. Biol. Rev.* **73**(2), 249–299 (2009).
85. A. Shukla, A. Biswas, N. Blot, F. Partensky, J. A. Karty, L. A. Hammad, L. Garczarek, A. Gutu, W. M. Schlachter, and D. M. Kehoe, "Phycocyanin-specific bilin lyase-isomerase controls blue-green chromatic acclimation in marine *Synechococcus*," *Proc. Natl. Acad. Sci. U. S. A.* **109**(49), 20136–20141 (2012).
86. R. J. Olson, S. W. Chisholm, and E. R. Zettler, "Pigments, size, and distribution of *Synechococcus* in the North Atlantic and Pacific Oceans," *Limnology and Oceanography* **35**(1), 45 (1990).
87. B. Palenik, "Chromatic adaptation in marine *Synechococcus* strains," *Appl. Environ. Microbiol.* **67**(2), 991–994 (2001).
88. A. R. Kirkham, L. E. Jardillier, R. Holland, M. V. Zubkov, and D. J. Scanlan, "Analysis of photosynthetic picoeukaryote community structure along an extended Ellett Line transect in the northern North Atlantic reveals a dominance of novel prymnesiophyte and prasinophyte phylotypes," *Deep Sea Res., Part 1* **58**(7), 733–744 (2011).