Supplement of Biogeosciences, 23, 1043–1064, 2026
https://doi.org/10.5194/bg-23-1043-2026-supplement

Biogeosciences Open Access EGU

Supplement of

Machine learning for estimating phytoplankton size structure from satellite ocean color imagery in optically complex Pacific Arctic waters

Hisatomo Waga et al.

Correspondence to: Hisatomo Waga (hwaga@alaska.edu)

Section S1

Phytoplankton pigment

At each station other than those from the PB21 cruise, between 200 and 2000 mL of surface seawater was filtered onto polycarbonate membrane or nylon mesh filters (20, 10, 5, and 2 μm pore size) and GF/F filters (0.7 μm pore size) under low vacuum pressure (<0.013 MPa), with various combinations of pore sizes used to separate water samples for different Chlorophyll *a* (Chl *a)* size fractionations (Table 2). Filters for size-fractionated measurements with fluorometric analysis were immediately soaked in *N*, *N*-dimethylformamide, and size-fractionated Chl *a* (Chl a_{size_obs}) concentrations were determined using the non-acidification technique (Welschmeyer, 1994) after a 24 h extraction in the dark at −20 °C (Suzuki and Ishimaru, 1990). During the PB21 cruise, the filter samples for Chl a_{size_obs} measurements were obtained by filtering between 500 and 1000 mL of surface seawater using the same method as with the other cruises but were promptly frozen in liquid nitrogen and then stored in a deep freezer (−80 °C).

Across all cruises, between 500 and 5000 mL of bulk, unfractionated surface seawater samples were filtered onto GF/F filters to determine the concentrations of major phytoplankton pigments. These filters were promptly frozen in liquid nitrogen and stored in a deep freezer (−80 °C) until analysis. Pigment extraction for the filters (including those of Chl a_{size_obs} samples obtained during the PB21 cruise) and the subsequent HPLC analysis were conducted at multiple labs using several HPLC systems, following the method of (Van Heukelem and Thomas, 2001): at Hokkaido University using a CLASS VP system (Shimadzu Corporation) for samples collected in 2007–2013, at Japan Agency for Marine-Earth Science and Technology (JAMSTEC) using an Agilent 1300 series (Agilent Technologies) for samples collected in 2016 and 2017, and at NASA Goddard Space Flight Center (GSFC) using an Agilent 1200 series (Agilent Technologies) for samples collected during the PB21 cruise.

Section S2

Absorption coefficient

Particles in surface seawater samples (between 500 and 5000 mL) were collected on a GF/F filter until the filter had sufficient coloration to measure the absorption coefficient of phytoplankton ($a_{ph\_obs}(\lambda)$). The absorption coefficient of particles ($a_{p\_obs}(\lambda)$) on the filter was measured in the spectral range from 300 to 850 nm at 1 nm intervals using an MPS-2400 (Shimadzu Corporation), MPS-2450 (Shimadzu Corporation) or Cary 100 (Agilent Technologies) spectrophotometer. The quantitative filter technique (QFT) was used to determine $a_{ph\_obs}(\lambda)$ for samples measured with the MPS-2400 and MPS-2450 instruments (i.e., all cruises but PB21), following the procedure described by Mitchell (Mitchell, 1990), whereas $a_{ph\_obs}(\lambda)$ for the PB21 samples was determined with GF/F filters placed inside a 15-cm integrating sphere connected to the Cary 100 (IOCCG, 2018). Following the measurement for $a_{p\_obs}(\lambda)$, the absorption coefficient of NAP ($a_{NAP\_obs}(\lambda)$) was measured after soaking the filter in 95% methanol or sodium hypochlorite, and $a_{ph\_obs}(\lambda)$ was finally obtained by subtracting a_{NAP_obs} (λ)

from $a_{p\_obs}(\lambda)$. The absorption coefficient of CDOM ($a_{CDOM\_obs}(\lambda)$) at wavelengths from 250 to 750 nm at 1 nm intervals was measured using the same spectrophotometers as for the particulate absorption measurements, with the exception of the PB21 samples, which were analyzed using a Cary 300 (Agilent Technologies) spectrophotometer with 5-cm quartz cuvettes. The summed measurements of individual constituent absorption coefficients allow estimation of the total absorption coefficient of seawater, defined as:

$$a_{\mathrm{total\_obs}}(\lambda) = a_{\mathrm{ph\_obs}}(\lambda) + a_{\mathrm{NAP\_obs}}(\lambda) + a_{\mathrm{CDOM\_obs}}(\lambda) + a_{\mathrm{w}}(\lambda), \quad \text{(S1)}$$

where $a_w(\lambda)$ is the spectral absorption coefficient of pure water (Pope and Fry, 1997).

Section S3

Remote sensing reflectance

In situ spectral radiance and irradiance measurements were acquired using a PRR-800/810 (Biospherical Instruments), C-OPS (Biospherical Instruments), or HyperPro (Satlantic) spectroradiometer. The PRR-800/810 and C-OPS measured underwater downward spectral irradiance ($E_d(\lambda, z)$) and upward spectral radiance ($L_u(\lambda, z)$) at 17 (380 to 765 nm) and 19 wavelengths (320 to 875 nm), respectively. The HyperPro was deployed as a surface buoy and acquired $E_d(\lambda, +0)$ and $L_u(\lambda, -0.24)$ between 400 and 800 nm at approximately 3 nm intervals. Remote sensing reflectance ($R_{rs\_obs}(\lambda)$) was calculated as the ratio of the water-leaving radiance ($L_w(\lambda)$) to the above-water downward spectral irradiance ($E_s(\lambda)$):

$$R_{\mathrm{rs\_obs}}(\lambda) = L_{\mathrm{w}}(\lambda)/E_{\mathrm{s}}(\lambda). \quad \text{(S2)}$$

For PRR-800/810 and C-OPS data, $L_w(\lambda)$ was estimated from $L_u(\lambda, z)$ just beneath the water surface ($L_u(\lambda, -0)$) following (Gordon et al., 1983), which was determined by extrapolating near-surface $L_u(\lambda, z)$ between 0.5–3 and 10 m deep to the surface. $L_u(\lambda, -0)$ was propagated through the water-air interface by applying the multiplicative factor 0.544 (Darecki and Stramski, 2004) to get $L_w(\lambda)$ as follows:

$$L_{\mathrm{u}}(\lambda, -0) = L_{\mathrm{u}}(\lambda, z) \times \exp(-K_{\mathrm{u}}(\lambda) \times z), \quad \text{(S3)}$$

$$L_{\mathrm{w}}(\lambda) = 0.544 \times L_{\mathrm{u}}(\lambda, -0), \quad \text{(S4)}$$

where z and $K_u(\lambda)$ are the depth and diffuse attenuation coefficients of $L_u(\lambda, z)$, respectively. For the HyperPro data, $E_s(\lambda)$ was propagated through the water surface and down to the depth of the radiometer (0.24 m) using the diffuse attenuation for downwelling light, $K_d(\lambda)$, derived from in situ measurements of total absorption and light scatter. Remote sensing reflectance was then computed as:

$$R_{\mathrm{rs\_obs}}(\lambda) = \frac{0.52 r_{\mathrm{rs\_obs}}(\lambda)}{1.0 - 1.7 r_{\mathrm{rs\_obs}}(\lambda)}, \tag{S5}$$

where $r_{rs\_obs}(\lambda)$ is the sub-surface reflectance, calculated as:

$$r_{\mathrm{rs\_obs}}(\lambda) = L_u(\lambda, -0.24) / \left[(1 - \rho_s) E_d(\lambda, +0)\, e^{-0.24\, K_d(\lambda)}\right]. \tag{S6}$$

The surface reflectance (ρ_s) was computed as a weighted sum of surface reflectance due to the direct sunlight (ρ_{sun}), computed as Fresnel reflectance with water refractive index set equal to 1.34, and sunlight reflected from clouds (ρ_{cloud}), which was assumed to be 0.05. Weights for ρ_{sun} and ρ_{cloud} were based on observed estimates of fractional cloud cover (F_c) as follows:

$$\rho_s = (1 - F_c)\rho_{sun} + F_c \rho_{\mathrm{cloud}}\,. \tag{S7}$$

Since $K_d(\lambda)$ was not measured directly, it was estimated as an average between the surface and 0.24 m depth using the Hydrolight radiative transfer model (Mobley, 1995).

$R_{rs\_obs}(\lambda)$ was resampled at ten MODIS bands in the visible range (i.e., 412, 443, 469, 488, 531, 547, 555, 645, 667, and 678 nm) from the original wavelengths of each instrument using spline interpolation (Wang et al., 2015). Note that, based on hyperspectral $R_{rs\_obs}(\lambda)$ data at wavelengths from 350 and 900 nm at 1 nm intervals measured by a HR-512i handheld spectroradiometer (Spectra Vista Corporation) during the PB21 cruise (N = 10), the spline interpolation showed median percent differences between the observed and resampled R_{rs_obs} (λ) at ten MODIS bands were 0.15%, 0.00%, and 0.01% for the PRR-800/810, C-OPS, and HyperPro data, respectively. This supports the robustness of using spline interpolation to resample $R_{rs\_obs}(\lambda)$ values at MODIS wavelengths from spectroradiometer data with slightly different observation wavelengths. Finally, a modified version of the Quasi-Analytical Algorithm (QAA; Lee et al., 2002) for the Pacific Arctic (Fujiwara et al., 2016) was used to estimate $a_{ph}(\lambda)$ ($a_{ph\_QAA}(\lambda)$) from *in situ* $R_{rs}(\lambda)$ ($R_{rs\_obs}(\lambda)$) and satellite $R_{rs}(\lambda)$ ($R_{rs\_sat}(\lambda)$). Here, $a_{ph\_QAA}(\lambda)$ estimated from $R_{rs\_obs}(\lambda)$ and $R_{rs\_sat}(\lambda)$ is denoted as $a_{ph\_QAAobs}(\lambda)$ and $a_{ph\_QAAsat}(\lambda)$, respectively. To avoid the retrieval of negative $a_{ph\_QAA}(\lambda)$, the modified version of QAA uses an optimized spectral slope of the absorption coefficient of combined CDOM and non-algal particles (S_{dg}) obtained by reconstructing the S_{dg} based on a dataset collected in the Pacific Arctic (Fujiwara et al., 2016). The $a_{ph\_QAAobs}(\lambda)$ was used to validate the performance of the modified version of the QAA by comparing it with $a_{ph\_obs}(\lambda)$.

Section S4

Pigment-based identification of phytoplankton taxonomic composition

An open-source R software package, *phytoclass* (ver 1.0.0), was used to determine the Chl *a* biomass of different phytoplankton groups from their accessory pigments (Hayward et al., 2023). The *phytoclass* package is a Chl *a* taxonomic partitioning software package similar to the widely used CHEMTAX software (Mackey et al., 1996). However, *phytoclass* has been shown to be more accurate and does not rely on initial assumptions of pigment to Chl *a* ratios for each phytoplankton group (Hayward et al., 2023).

For this study, eight target taxonomic groups (diatoms, chrysophytes, dinoflagellates, prymnesiophytes, chlorophytes, prasinophytes, cryptophytes, and cyanobacteria) and 11 marker pigments for each taxonomic group (peridinin, 19'-butanoyloxyfucoxanthin, fucoxanthin, 19'-hexanoyloxyfucoxanthin, neoxanthin, prasinoxanthin, violaxanthin, alloxanthin, lutein, zeaxanthin, and chlorophyll *b*) were selected following (Zhuang et al., 2016), as these groupings have been used previously for CHEMTAX analysis in the Chukchi Sea shelf region. Before calculating group-specific Chl *a* biomass using *phytoclass*, a hierarchical cluster analysis on the pigment concentrations normalized to Chl *a* was conducted using the Ward method (Punj and Stewart, 1983) to partition the dataset into similar pigment compositions. This clustering is recommended by Hayward et al. (Hayward et al., 2023) because pigment ratios can change at the phytoplankton genus or species level and with environmental conditions (Henriksen et al., 2002; Schlüter et al., 2000).

Figure S1

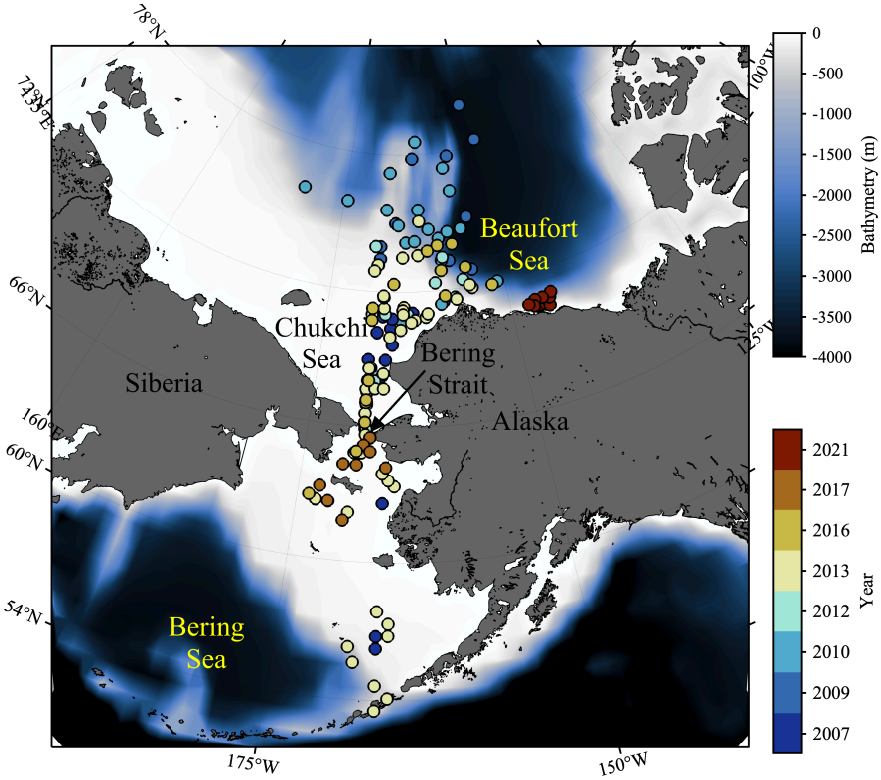

Figure S1. Sampling locations of *in situ* data used in this study. Colors of each plot indicate cruise years, whereas background color represent the bathymetry.

Figure S2

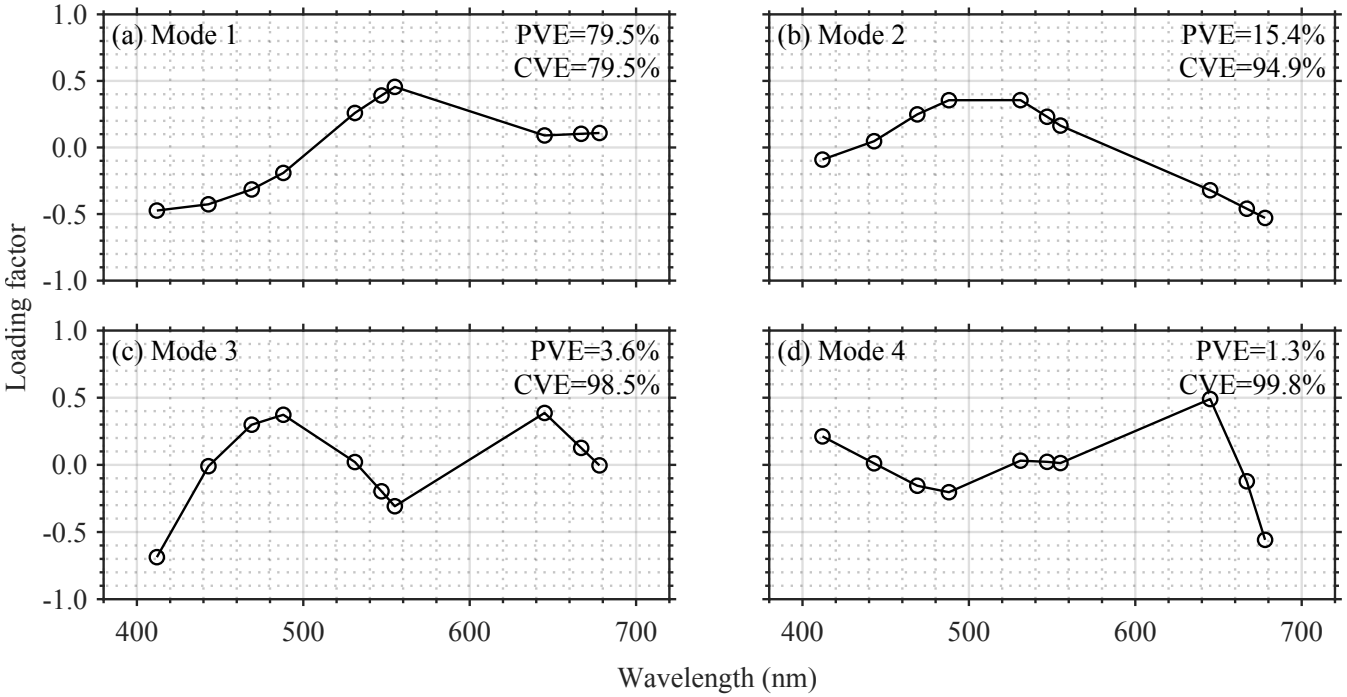

Figure S2. Loading factors derived from spectral variations of $\hat{R}_{\mathrm{rs\_obs}}(\lambda)$. The (a) first, (b) second, (c) third, and (4) fourth modes of principal component analysis (PCA). PVEs and CVEs indicate the proportion of variance and cumulative variance explained by each mode, respectively.

Figure S3

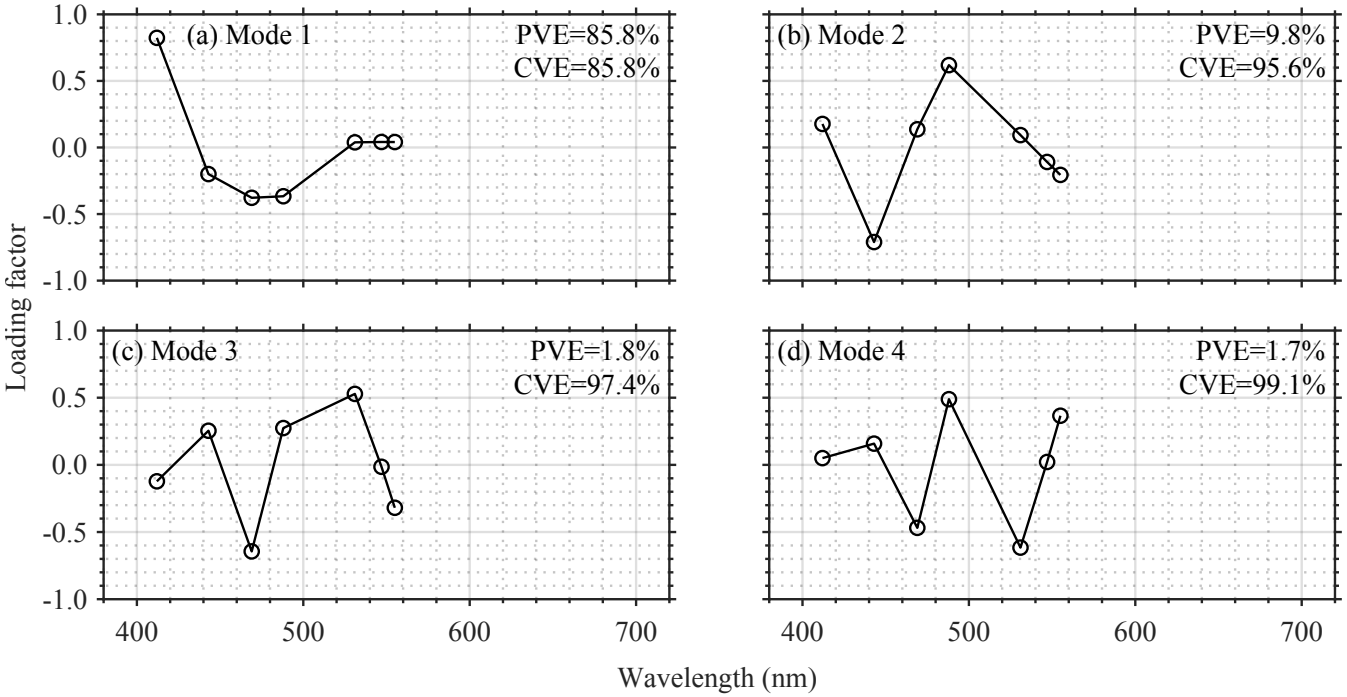

Figure S3. Loading factors derived from spectral variations of $\hat{a}_{\mathrm{ph\_obs}}(\lambda)$. The (a) first, (b) second, (c) third, and (4) fourth modes of PCA.

Figure S4

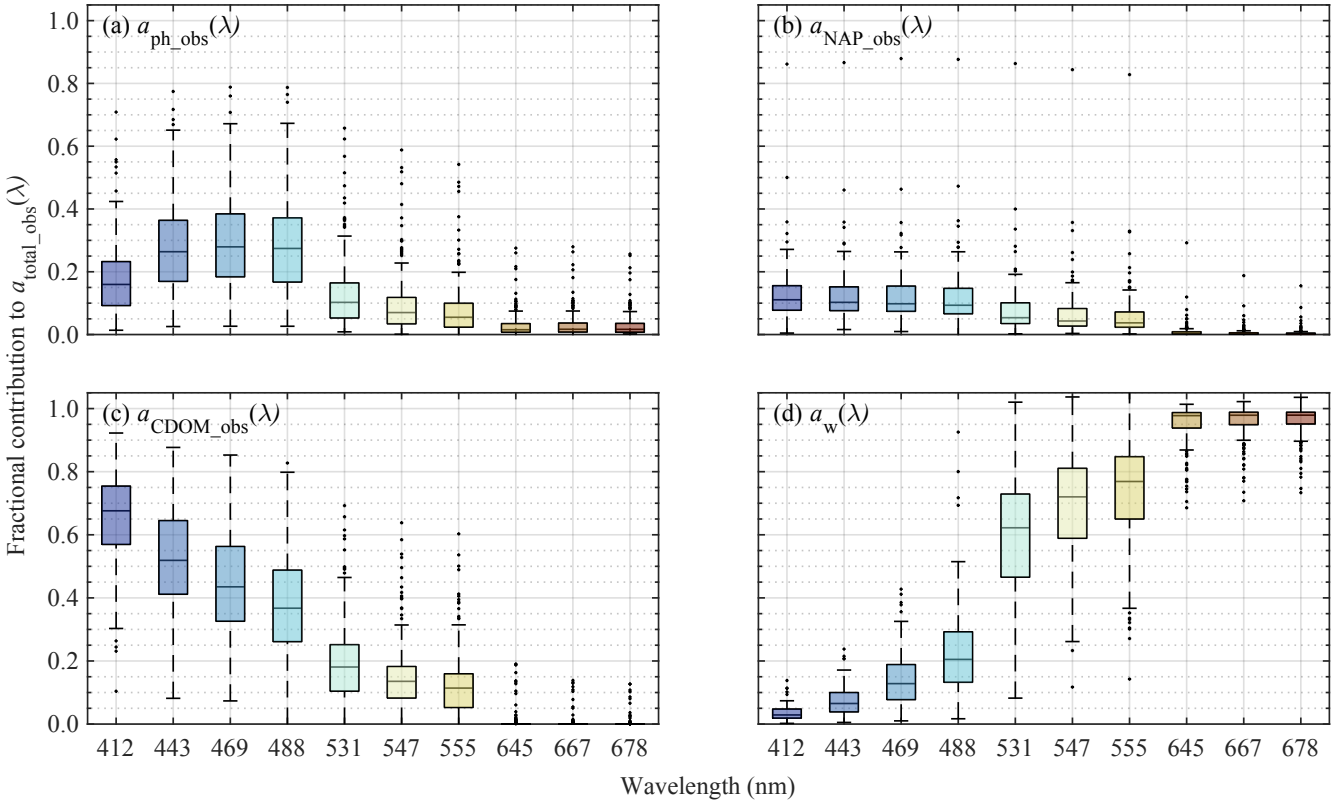

Figure S4. Light absorption by each water constituent. Fractional contributions to the total absorption coefficient ($a_{total\_obs}(\lambda)$) by (a) phytoplankton ($a_{ph\_obs}(\lambda)$), (b) non-algal particles ($a_{NAP\_obs}(\lambda)$), (c) colored dissolved organic matter ($a_{CDOM\_obs}(\lambda)$), and (d) pure water ($a_w(\lambda)$) at ten MODIS-A wavebands. Values that are more than 1.5 times the interquartile range away from the bottom or top of the box are marked as outliers.

Figure S5

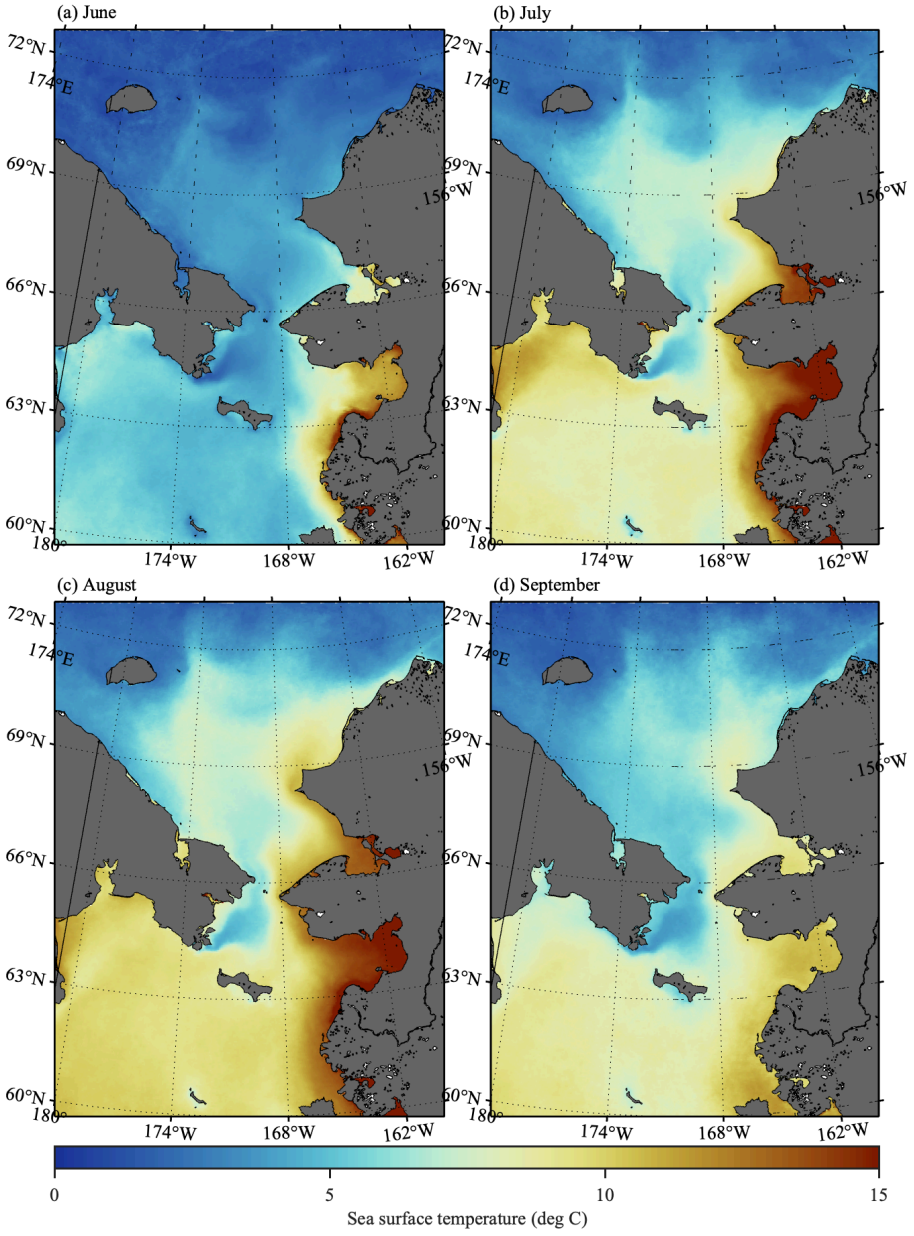

Figure S5. Monthly climatology of sea surface temperature (SST) values in (a) June, (b) July, (c) August, and (d) September in the Pacific Arctic for 2002–2022.

Table S1

Table S1. Pigment to Chl *a* ratios for each phytoplankton taxa used for the *phytoclass* analyses. Abbreviations: Peri, peridinin; But, 19'-butanoylofucoxanthin; Fuco, fucoxanthin; Hex, 19'-hexanoyloxyfucoxanthin; Neo, neoxanthin; Pras, prasinoxanthin; Viola, violaxanthin; Allo, alloxanthin; Lut, lutein; Zea, zeaxanthin; Chl b, chlorophyll b; Chryso, chrysophytes; Dino, dinoflagellates; Prym, prymnesiophytes; Chloro, chlorophytes; Pras, prasinophytes; Crypto, cryptophytes; Cyano, cyanobacteria.

		Pigment:Chl *a* ratio											
		Peri	But	Fuco	Hex	Neo	Pras	Viola	Allo	Lut	Zea	Chl *b*	Chl *a*
Cluster 1	Diatom	0	0	0.45	0	0	0	0	0	0	0	0	1
	Chryso	0	0.83	1.32	0.37	0	0	0	0	0	0	0	1
	Dino	0.77	0	0	0	0	0	0	0	0	0	0	1
	Prym	0	0	0	0.20	0	0	0	0	0	0	0	1
	Chloro	0	0	0	0	0.30	0	0.98	0	1.48	1.01	1.14	1
	Pras	0	0	0	0	0.20	0.21	0.20	0	0.20	0	1.37	1
	Crypto	0	0	0	0	0	0	0	0.20	0	0	0	1
	Cyano	0	0	0	0	0	0	0	0	0	1.49	0	1
Cluster 2	Diatom	0	0	0.36	0	0	0	0	0	0	0	0	1
	Chryso	0	1.49	0.21	0.21	0	0	0	0	0	0	0	1
	Dino	0.48	0	0	0	0	0	0	0	0	0	0	1
	Prym	0	0	0	0.35	0	0	0	0	0	0	0	1
	Chloro	0	0	0	0	1.49	0	0.21	0	0.37	0.24	1.17	1
	Pras	0	0	0	0	0.20	0.20	0.20	0	0.20	0	1.42	1
	Crypto	0	0	0	0	0	0	0	0.36	0	0	0	1
	Cyano	0	0	0	0	0	0	0	0	0	0.86	0	1
Cluster 3	Diatom	0	0	0.39	0	0	0	0	0	0	0	0	1
	Chryso	0	1.50	0.21	0.21	0	0	0	0	0	0	0	1
	Dino	1.50	0	0	0	0	0	0	0	0	0	0	1
	Prym	0	0	0	0.36	0	0	0	0	0	0	0	1
	Chloro	0	0	0	0	0.96	0	0.71	0	0	1.49	1.36	1
	Pras	0	0	0	0	0.20	0.55	0.26	0	0	0	1.49	1
	Crypto	0	0	0	0	0	0	0	0.20	0	0	0	1
	Cyano	0	0	0	0	0	0	0	0	0	0.20	0	1
Cluster 4	Diatom	0	0	0.40	0	0	0	0	0	0	0	0	1
	Chryso	0	0.25	0.20	0.20	0	0	0	0	0	0	0	1
	Dino	1.49	0	0	0	0	0	0	0	0	0	0	1
	Prym	0	0	0	0.59	0	0	0	0	0	0	0	1
	Chloro	0	0	0	0	0.20	0	1.05	0	0.20	0.20	0.20	1
	Pras	0	0	0	0	0.69	0.20	0.26	0	0.21	0	0.20	1
	Crypto	0	0	0	0	0	0	0	0.32	0	0	0	1
	Cyano	0	0	0	0	0	0	0	0	0	1.49	0	1

Table S2.

Table S2. Comparison of $a_{ph–obs}(\lambda)$ and $a_{ph\_QAAobs}(\lambda)$ values estimated from $R_{rs\_obs}(\lambda)$ using the quasi-analytical algorithm (QAA) for the Pacific Arctic. MAE stands for mean absolute error.

	Wavelength (nm)									
	412	443	469	488	531	547	555	645	667	678
MAE	0.02	0.02	0.02	0.01	0.02	0.02	0.02	0.08	0.13	0.13
Bias	1.03	1.04	1.03	1.01	0.99	0.97	0.97	1.19	1.29	1.31

Table S3

Table S3. Model parameters (β_0 and C_j in Eqs. (8) and (9)) for the CSD models optimized with the principal component analysis (PCA)-based approach. Model parameters are indicated for each predictor: i.e., $\hat{R}_{\mathrm{rs}}(\lambda)$ and $\hat{a}_{\mathrm{ph}}(\lambda)$, with the further threshold for $\hat{a}_{\mathrm{ph}}(\lambda)$.

Predictor	Threshold	Model parameters										
		β_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
$\hat{a}_{\mathrm{ph}}(\lambda)$	$a_{\mathrm{ph}}(412) > a_{\mathrm{ph}}(469)$	0.05	0.65	−0.50	0.98	−1.62	2.57	−0.27	−1.79	-	-	-
	$a_{\mathrm{ph}}(412) \leq a_{\mathrm{ph}}(469)$	−0.18	0.36	−0.41	−0.67	0.71	2.15	−0.40	−1.74	-	-	-
$\hat{R}_{\mathrm{rs}}(\lambda)$	-	−0.08	0.21	−0.06	−0.14	−0.15	−0.01	0.13	0.20	−0.47	0.02	0.27

Table S4

Table S4. Training results of the CSD models based on the diverse machine learning approaches (i.e., model type and preset) using $\hat{R}_{\mathrm{rs}}(\lambda)$ as the predictor. The four statistical metrics, including the root mean square error (RMSE), mean squared error (MSE), coefficient of determination (r^2), and mean absolute error (MAE), are given as mean ± std derived from ten repeats of five-fold cross-validation.

Rank	Model type	Preset	RMSE			MSE			R2			MAE		
1	Linear Regression	Linear	0.16	±	0.01	0.03	±	0.00	0.76	±	0.02	0.12	±	0.01
2	Linear Regression	Robust Linear	0.16	±	0.01	0.03	±	0.00	0.76	±	0.02	0.12	±	0.00
3	SVM	Linear SVM	0.17	±	0.01	0.03	±	0.00	0.74	±	0.03	0.13	±	0.00
4	Stepwise Linear Regression	Stepwise Linear	0.18	±	0.02	0.03	±	0.01	0.70	±	0.08	0.13	±	0.01
5	Efficient Linear	Effiicient Linear SVM	0.18	±	0.00	0.03	±	0.00	0.69	±	0.01	0.14	±	0.00
6	Efficient Linear	Efficient Linear Least Squares	0.20	±	0.00	0.04	±	0.00	0.63	±	0.01	0.15	±	0.00
7	Gaussian Process Regression	Exponential GPR	0.21	±	0.01	0.05	±	0.00	0.59	±	0.02	0.16	±	0.00
8	SVM	Medium Gaussian SVM	0.24	±	0.01	0.06	±	0.00	0.48	±	0.03	0.18	±	0.00
9	Gaussian Process Regression	Squared Exponential GPR	0.25	±	0.04	0.06	±	0.02	0.41	±	0.18	0.15	±	0.01
10	Gaussian Process Regression	Rational Quadratic GPR	0.25	±	0.04	0.07	±	0.02	0.41	±	0.19	0.15	±	0.01
11	Ensemble	Bagged Trees	0.26	±	0.01	0.07	±	0.00	0.40	±	0.02	0.19	±	0.00
12	Gaussian Process Regression	Matern 5/2 GPR	0.27	±	0.04	0.07	±	0.02	0.35	±	0.17	0.15	±	0.01
13	Ensemble	Boosted Trees	0.27	±	0.01	0.07	±	0.01	0.36	±	0.05	0.19	±	0.01
14	SVM	Coarse Gaussian SVM	0.27	±	0.00	0.07	±	0.00	0.33	±	0.01	0.21	±	0.00
15	SVM	Fine Gaussian SVM	0.28	±	0.00	0.08	±	0.00	0.30	±	0.02	0.20	±	0.00
16	Kernel	Least Squared Regression Kernel	0.28	±	0.01	0.08	±	0.00	0.29	±	0.04	0.20	±	0.01
17	Kernel	SVM Kernel	0.29	±	0.01	0.08	±	0.01	0.24	±	0.06	0.21	±	0.01
18	Tree	Coarse Tree	0.30	±	0.00	0.09	±	0.00	0.21	±	0.02	0.23	±	0.00
19	Tree	Medium Tree	0.30	±	0.01	0.09	±	0.01	0.19	±	0.07	0.21	±	0.01
20	Tree	Fine Tree	0.31	±	0.02	0.10	±	0.01	0.14	±	0.12	0.22	±	0.01
21	Neural Network	Narrow Neural Network	0.38	±	0.11	0.15	±	0.09	-0.37	±	0.77	0.20	±	0.02
22	Neural Network	Bi-layered Neural Network	0.40	±	0.07	0.16	±	0.06	-0.46	±	0.56	0.23	±	0.02
23	Neural Network	Tri-layered Neural Network	0.47	±	0.20	0.25	±	0.25	-1.29	±	2.26	0.25	±	0.04
24	Neural Network	Medium Neural Network	0.64	±	0.21	0.45	±	0.31	-3.09	±	2.87	0.27	±	0.03
25	SVM	Quadratic SVM	0.71	±	0.41	0.66	±	0.93	-4.93	±	8.37	0.25	±	0.07
26	Neural Network	Wide Neural Network	0.84	±	0.36	0.82	±	0.65	-6.35	±	5.88	0.32	±	0.05
27	Linear Regression	Interactions Linear	3.21	±	1.26	11.74	±	9.11	-104.16	±	80.99	0.55	±	0.13
28	SVM	Cubic SVM	24.10	±	32.60	1537.37	±	3416.27	-13771.77	±	30640.37	2.69	±	3.08

Table S5

Table S5. Training results of the CSD models based on the diverse machine learning approaches (i.e., model type and preset) using $\hat{a}_{\mathrm{ph}}(\lambda)$ as the predictor. The four statistical metrics, including the root mean square error (RMSE), mean squared error (MSE), coefficient of determination (r^2), and mean absolute error (MAE), are given as mean ± std derived from ten repeats of five-fold cross-validation.

Rank	Model type	Preset	RMSE			MSE			R2			MAE		
1	SVM	Medium Gaussian SVM	0.13	±	0.01	0.02	±	0.00	0.80	±	0.02	0.10	±	0.00
2	Gaussian Process Regression	Squared Exponential GPR	0.13	±	0.00	0.02	±	0.00	0.80	±	0.02	0.10	±	0.00
3	Gaussian Process Regression	Matern 5/2 GPR	0.13	±	0.01	0.02	±	0.00	0.80	±	0.02	0.10	±	0.00
4	Gaussian Process Regression	Rational Quadratic GPR	0.13	±	0.01	0.02	±	0.00	0.79	±	0.02	0.11	±	0.00
5	Gaussian Process Regression	Exponential GPR	0.14	±	0.00	0.02	±	0.00	0.78	±	0.01	0.11	±	0.00
6	Kernel	SVM Kernel	0.17	±	0.01	0.03	±	0.00	0.66	±	0.04	0.13	±	0.01
7	Ensemble	Bagged Trees	0.18	±	0.01	0.03	±	0.00	0.63	±	0.03	0.14	±	0.01
8	Ensemble	Boosted Trees	0.18	±	0.01	0.03	±	0.00	0.62	±	0.02	0.14	±	0.01
9	Kernel	Least Squared Regression Kernel	0.18	±	0.01	0.03	±	0.00	0.62	±	0.03	0.14	±	0.01
10	SVM	Coarse Gaussian SVM	0.18	±	0.00	0.03	±	0.00	0.62	±	0.01	0.15	±	0.00
11	Efficient Linear	Effiicient Linear SVM	0.20	±	0.00	0.04	±	0.00	0.53	±	0.01	0.17	±	0.00
12	Linear Regression	Linear	0.21	±	0.01	0.04	±	0.00	0.51	±	0.03	0.16	±	0.00
13	Tree	Fine Tree	0.21	±	0.01	0.04	±	0.00	0.50	±	0.06	0.16	±	0.01
14	Linear Regression	Robust Linear	0.21	±	0.00	0.04	±	0.00	0.50	±	0.02	0.15	±	0.00
15	Tree	Medium Tree	0.21	±	0.01	0.04	±	0.00	0.49	±	0.06	0.17	±	0.01
16	SVM	Linear SVM	0.21	±	0.01	0.04	±	0.00	0.48	±	0.03	0.15	±	0.00
17	SVM	Fine Gaussian SVM	0.23	±	0.00	0.05	±	0.00	0.39	±	0.02	0.17	±	0.00
18	Efficient Linear	Efficient Linear Least Squares	0.23	±	0.00	0.05	±	0.00	0.39	±	0.01	0.19	±	0.00
19	Tree	Coarse Tree	0.25	±	0.01	0.06	±	0.01	0.28	±	0.07	0.19	±	0.01
20	Neural Network	Narrow Neural Network	0.34	±	0.12	0.13	±	0.09	-0.48	±	1.09	0.18	±	0.03
21	Neural Network	Tri-layered Neural Network	0.36	±	0.08	0.14	±	0.07	-0.60	±	0.78	0.22	±	0.02
22	Neural Network	Wide Neural Network	0.37	±	0.10	0.14	±	0.09	-0.66	±	0.99	0.20	±	0.02
23	Neural Network	Medium Neural Network	0.37	±	0.09	0.14	±	0.08	-0.68	±	0.97	0.23	±	0.02
24	Neural Network	Bi-layered Neural Network	0.57	±	0.28	0.40	±	0.38	-3.65	±	4.49	0.26	±	0.03
25	Stepwise Linear Regression	Stepwise Linear	0.71	±	0.26	0.56	±	0.35	-5.51	±	4.06	0.21	±	0.03
26	Linear Regression	Interactions Linear	1.18	±	0.36	1.50	±	0.96	-16.46	±	11.04	0.27	±	0.04
27	SVM	Quadratic SVM	1.35	±	0.28	1.91	±	0.76	-21.18	±	8.92	0.28	±	0.03
28	SVM	Cubic SVM	5.10	±	2.66	32.39	±	34.40	-376.09	±	402.63	0.67	±	0.24

Table S6

Table S6. Model parameters (β_0 and C_j) for the CSD $\text{model}_{\text{LR}-\hat{R}_{\text{rs}}(\lambda)}$ optimized with the machine-learning (ML)-based approach. The formular is expressed as $\eta = b_0 + \sum_{j=1}^{m} a_j \hat{R}_{\text{rs}}(\lambda_j)$.

Model parameters										
b_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
1.73	−3.75	−3.13	0.00	−7.01	30.58	−94.13	55.29	−3.93	0.00	−5.13

Table S7

Table S7. Training results of the top-five and bottom-five CSD models based on the diverse machine learning approaches (i.e., model type and preset) with reduced data subset. The four statistical metrics, including the root mean square error (RMSE), mean squared error (MSE), coefficient of determination (r^2), and mean absolute error (MAE), are given as mean ± std derived from ten repeats of five-fold cross-validation.

Predictor	Rank	Model type	Preset	RMSE	MSE	r^2	MAE
$\hat{R}_{\text{rs}}(\lambda)$	1	Linear Regression	Linear	0.18 ± 0.01	0.03 ± 0.00	0.70 ± 0.04	0.14 ± 0.01
	2	Linear Regression	Robust Linear	0.19 ± 0.01	0.04 ± 0.00	0.68 ± 0.03	0.14 ± 0.00
	3	SVM	Linear SVM	0.19 ± 0.01	0.04 ± 0.00	0.67 ± 0.03	0.15 ± 0.01
	4	Efficient Linear	Effiicient Linear SVM	0.20 ± 0.01	0.04 ± 0.00	0.63 ± 0.03	0.16 ± 0.01
	5	Stepwise Linear Regression	Stepwise Linear	0.20 ± 0.04	0.04 ± 0.02	0.62 ± 0.17	0.14 ± 0.01
	⋮						
	24	Neural Network	Medium Neural Network	0.79 ± 0.27	0.68 ± 0.41	-5.18 ± 3.75	0.34 ± 0.05
	25	Neural Network	Wide Neural Network	0.81 ± 0.21	0.69 ± 0.35	-5.27 ± 3.08	0.34 ± 0.04
	26	SVM	Quadratic SVM	0.91 ± 0.44	1.00 ± 1.15	-8.04 ± 10.26	0.28 ± 0.05
	27	Linear Regression	Interactions Linear	2.11 ± 0.90	5.17 ± 4.68	-45.75 ± 42.15	0.65 ± 0.12
	28	SVM	Cubic SVM	14.58 ± 15.21	420.97 ± 806.34	-3795.55 ± 7262.99	1.99 ± 1.67
$\hat{a}_{\text{ph}}(\lambda)$	1	SVM	Medium Gaussian SVM	0.13 ± 0.01	0.02 ± 0.00	0.81 ± 0.02	0.10 ± 0.01
	2	Gaussian Process Regression	Squared Exponential GPR	0.13 ± 0.01	0.02 ± 0.00	0.81 ± 0.01	0.10 ± 0.00
	3	Gaussian Process Regression	Matern 5/2 GPR	0.13 ± 0.01	0.02 ± 0.00	0.81 ± 0.02	0.10 ± 0.00
	4	Gaussian Process Regression	Rational Quadratic GPR	0.14 ± 0.00	0.02 ± 0.00	0.78 ± 0.01	0.11 ± 0.00
	5	Gaussian Process Regression	Exponential GPR	0.14 ± 0.01	0.02 ± 0.00	0.77 ± 0.02	0.11 ± 0.01
	⋮						
	24	Neural Network	Medium Neural Network	0.61 ± 0.24	0.43 ± 0.38	-3.75 ± 4.38	0.31 ± 0.05
	25	Stepwise Linear Regression	Stepwise Linear	0.96 ± 0.40	1.05 ± 0.71	-10.63 ± 7.70	0.27 ± 0.05
	26	SVM	Quadratic SVM	1.49 ± 0.43	2.39 ± 1.44	-25.58 ± 16.35	0.35 ± 0.05
	27	SVM	Cubic SVM	1.57 ± 1.06	3.49 ± 4.72	-37.97 ± 53.38	0.35 ± 0.13
	28	Linear Regression	Interactions Linear	1.90 ± 0.39	3.74 ± 1.48	-40.26 ± 16.15	0.43 ± 0.05

References

Darecki, M. and Stramski, D.: An evaluation of MODIS and SeaWiFS bio-optical algorithms in the Baltic Sea, Remote Sens. Environ., 89, 326–350, https://doi.org/10.1016/j.rse.2003.10.012, 2004.

Gordon, H. R., Clark, D. K., Brown, J. W., Brown, O. B., Evans, R. H., and Broenkow, W. W.: Phytoplankton pigment concentrations in the Middle Atlantic Bight: comparison of ship determinations and CZCS estimates, Appl. Opt., 22, 20–36, https://doi.org/10.1364/ao.22.000020, 1983.

Hayward, A., Pinkerton, M. H., and Gutierrez-Rodriguez, A.: phytoclass: A pigment-based chemotaxonomic method to determine the biomass of phytoplankton classes, Limnol. Oceanogr. Methods, 21, 220–241, https://doi.org/10.1002/lom3.10541, 2023.

Henriksen, P., Riemann, B., Kaas, H., Sørensen, H. M., and Sørensen, H. L.: Effects of nutrient-limitation and irradiance on marine phytoplankton pigments, J. Plankton Res., 24, 835–858, https://doi.org/10.1093/plankt/24.9.835, 2002.

IOCCG: Ocean optics and biogeochemistry protocols for satellite ocean colour sensor validation; Volume 1.0. Inherent optical property measurements and protocols: Absorption coefficient, edited by: Neeley, A. R. and Mannino, A., International Ocean Colour Coordinating Group (IOCCG), Dartmouth, NS, Canada, https://doi.org/10.25607/OBP-119, 2018.

Mackey, M. D., Mackey, D. J., Higgins, H. W., and Wright, S. W.: CHEMTAX - a program for estimating class abundances from chemical markers:application to HPLC measurements of phytoplankton, Mar. Ecol. Prog. Ser., 144, 265–283, https://doi.org/10.3354/meps144265, 1996.

Mitchell, B. G.: Algorithms for determining the absorption coefficient for aquatic particulates using the quantitative filter technique, in: Ocean Optics X, 137–148, https://doi.org/10.1117/12.21440, 1990.

Mobley, C.: Hydrolight 3.0 User's Guide, Menlo Park, California, 65 pp., ADA299649,1995.

Pope, R. M. and Fry, E. S.: Absorption spectrum (380-700 nm) of pure water. II. Integrating cavity measurements, Appl. Opt., 36, 8710–8723, https://doi.org/10.1364/AO.36.008710, 1997.

Punj, G. and Stewart, D. W.: Cluster Analysis in Marketing Research: Review and Suggestions for Application, J. Mark. Res., 20, 134–148, https://doi.org/10.1177/002224378302000204, 1983.

Schlüter, M., Sauter, E. J., Schäfer, A., and Ritzrau, W.: Spatial budget of organic carbon flux to the seafloor of the northern North Atlantic (60°N–80°N), Global Biogeochem. Cycles, 14, 329–340, https://doi.org/10.1029/1999GB900043, 2000.

Suzuki, R. and Ishimaru, T.: An improved method for the determination of phytoplankton chlorophyll using N, N-dimethylformamide, J. Oceanogr. Soc. Japan, 46, 190–194, https://doi.org/10.1007/BF02125580, 1990.

Van Heukelem, L. and Thomas, C. S.: Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments, J. Chromatogr. A, 910, 31–49, https://doi.org/10.1016/S0378-4347(00)00603-4, 2001.

Welschmeyer, N. A.: Fluorometric analysis of chlorophyll a in the presence of chlorophyll b and pheopigments, Limnol. Oceanogr., 39, 1985–1992, https://doi.org/10.4319/lo.1994.39.8.1985, 1994.

Zhuang, Y., Jin, H., Li, H., Chen, J., Lin, L., Bai, Y., Ji, Z., Zhang, Y., and Gu, F.: Pacific inflow control on phytoplankton community in the Eastern Chukchi Shelf during summer, Cont. Shelf Res., 129, 23–32, https://doi.org/10.1016/j.csr.2016.09.010, 2016.