microARC modelling methods

Aidan Hunter, Ben Ward

June 1, 2021

1 Introduction

2 Material and Methods

2.1 Data

2.1.1 Observational data

Nutrient, organic matter and plankton size spectra were measured from water samples collected by CTD casts from onboard the RV Polarstern research vessel at the LTER (Long Term Ecological Research) HAUSGARTEN sites in Fram Strait during three separate cruises (PS99, PS107 and PS114) in 2016–2018 (table 1). Only the (PS114) 2018 data were used because the full complement of required measurements (nutrient, organic matter/chlorophyll *a*, size spectra, and forcing data) was unavailable for earlier years. Our model simulates total dissolved nitrogen without distinguishing between compounds. We therefore extracted the nitrite-plus-nitrate concentration (NO_2 + NO_3) column from the nutrient data table (Torres-Valdés et al., 2019), as well as the sample event (time and place) and depth covariates. Let these dissolved nitrogen data be denoted as $\widehat{N}_{d,s,g}$, where d, s and g index depth, sample event, and some grouping variable. The $\widehat{\ }$ notation represents observations. Measured concentrations of PON and POC (particulate organic carbon and nitrogen) and chlorophyll *a* were extracted from the organic matter data table (von Jackowski et al., 2020). Depth and sample event covariates were also extracted; these differed from those associated with the nutrient data as different data types were not all gathered at the same sample events nor necessarily from the same depths (fig. 4). As the nutrient and organic matter data follow a structurally identical sample design comprising single concentration measurements at each depth and sample event, together we refer them as the scalar data. The scalar data may be written in general form $\widehat{Y}_{d,s,g}$, where $Y \in \{N, \mathrm{PON}, \mathrm{POC}, \mathrm{Chl}\ a\}$ arbitrarily represents any data type. All of the scalar data used to inform our model are displayed in fig. 1.

Planktonic size spectra were derived through meticulous microscopy then kernel density estimation methods. A subset of the CTD water samples, collected from various locations and across a small range of depths, were used to derive unique size spectra corresponding to different sampling events and depths. From 50 mL aliquots of each water sample, plankton cells were identified, measured, and counted under varying degrees of magnification — the smallest cells were only identified by genera. Kernel density methods were then applied to these count-data to generate smooth size spectra. See Lampe et al. (2021) for a full description of the size spectra derivation.

The size spectra data tables contain measured distributions of autotrophic and heterotrophic cell concentration with respect to ESD (equivalent spherical diameter). Each distribution is represented by two vectors: the independent variable, ESD (μm), a sequence of cell sizes with successive elements separated by 0.005 on the $\log_{10}$ scale; and the dependent variable of cell concentration density (cells $\mathrm{m}^{-3} \log_{10}(\mathrm{ESD}/1\mu\mathrm{m}))^{-1}$, evaluated for each element in the ESD sequence. As measured concentration densities were negligible for ESD $> 200\mu$m and were relatively uncertain for ESD $< 1\mu$m we truncated all size spectra to the interval ESD $\in [1, 200]\,\mu$m. We then used the sphere volume equation to convert cell concentration densities into bio-volume densities. The data were then aggregated into two groups by designating each sample event as Atlantic or Arctic depending upon the origin of the water samples (section 2.1.2). We then derived single size spectra for both regional groupings by averaging bio-volume densities across sampled depths and sample events (times and locations). This produced four separate size spectra: autotroph and heterotroph averaged bio-volume density from all water samples of Atlantic or Arctic origin (fig. 2).

The size spectra data (fig. 2) are continuous and therefore not directly comparable with our model outputs. We derived *binned* size data by choosing n contiguous size class intervals at much coarser resolution than the raw data, then integrating the bio-volume density vectors piecewise across each size interval to generate vectors where each of the n elements stores the total bio-volume attributed to a single size class interval (fig. 3).

Table 1: Cruise dates and availability of each data type — dissolved inorganic nitrogen, organic matter, size spectra and physical model forcing data.

Cruise	Dates	DIN	OM	Size	Forcing
PS99	25/6 – 10/7/2016	✓	✓	✓	✗
PS107	25/7 – 15/8/2017	✓	✗	✓	✓
PS114	16/7 – 27/7/2018	✓	✓	✓	✓

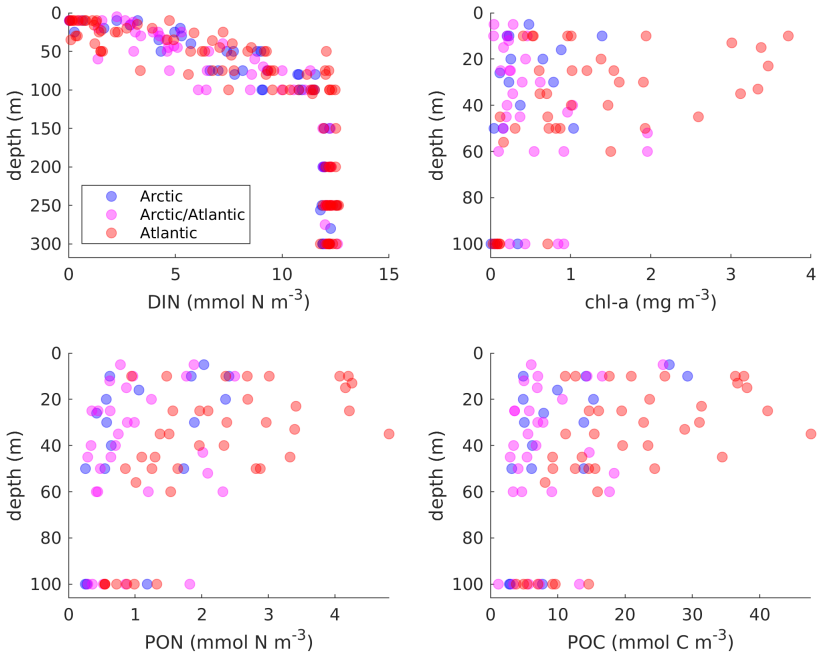


Figure 1: Nutrient and organic matter data collected during the 2018 RV Polarstern cruise – all sample events and depths used for model optimisation. Mechanistic relationships with depth are apparent; the within-depth variability results from multiple sampling events. Colours indicate the water origin of each sample, determined by the closest physical model trajectories.

2.1.2 Model-simulated forcing data

Our NPZD model represents a 1D vertical water column that is horizontally advected by ocean currents. Advection is specified by oceanographic model output. SINMOD is currently being used for this purpose, however, we intend to employ NEMO-MEDUSU simulations for the final analyses. The oceanographic model outputs comprise several thousand spatial trajectories that represent individual "particles" moving with ocean currents. Each trajectory is associated with time series of depth-discrete water temperatures and diffusivities, and surface irradiance and ice cover, that are used to drive the NPZD model. Running the NPZD model along these trajectories allows us to explicitly model plankton dynamics across depth while the entire water column is transported horizontally through space. This *Lagrangian* representation of the fluid dynamics is more intuitive than an *Eulerian* approach, and reduces the required computation.

The particle trajectories that we used in the model were extracted from SINMOD outputs as follows. Circles of radius 25 km drawn around each in-

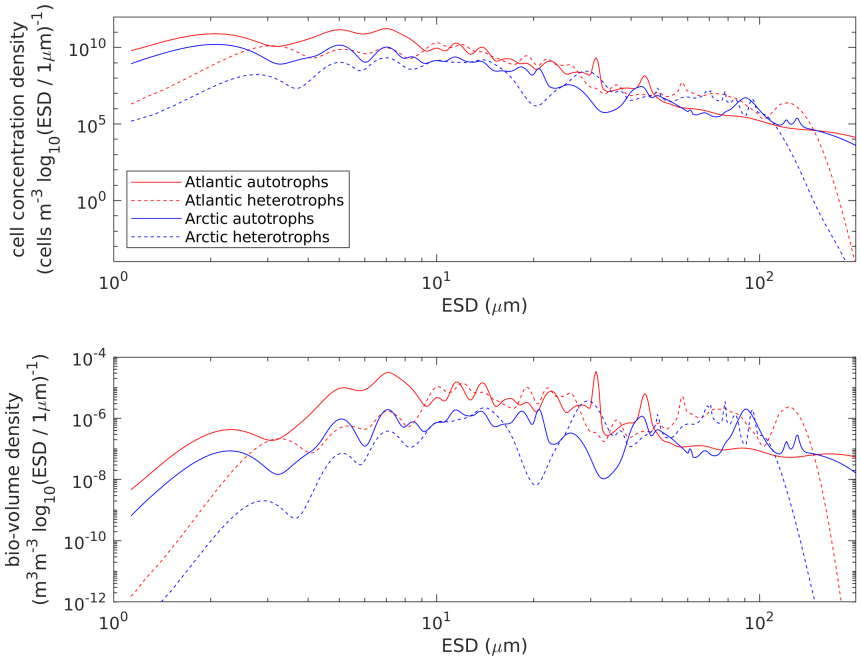


Figure 2: Plankton size spectra data derived using samples from the 2018 Polarstern cruise. Lines display observed spectra, averaged over depths and sample events, for autotrophic and heterotrophic plankton and grouped by samples from water originating from the Atlantic and the Arctic.

situ sampling site were combined to form a polygon. All particle trajectories passing through the polygon at the time of sampling were stored. This yielded several thousand trajectories that we could potentially use as forcing data.

The number of trajectories had to be reduced in order to limit model run-time. As many of the trajectories represented particles initiated close to each other, data associated with numerous different trajectories were similar. Thus, we could sensibly omit many trajectories that did not provide much information due to essentially duplicating data from nearby particles. We decided to extract and use ten unique trajectories for each in-situ sampling event. These were filtered out from the set of all trajectories as follows.

A catchment area, of radius 25 km, was centred at each sampling event. Any particle outside the catchment area at the time of sampling was omitted from consideration for that event. The remaining particles were considered close enough to the sampling event to potentially be included. These were filtered using a clustering algorithm to extract the ten most dissimilar trajectories. This process was conducted for each sampling event to produce a set

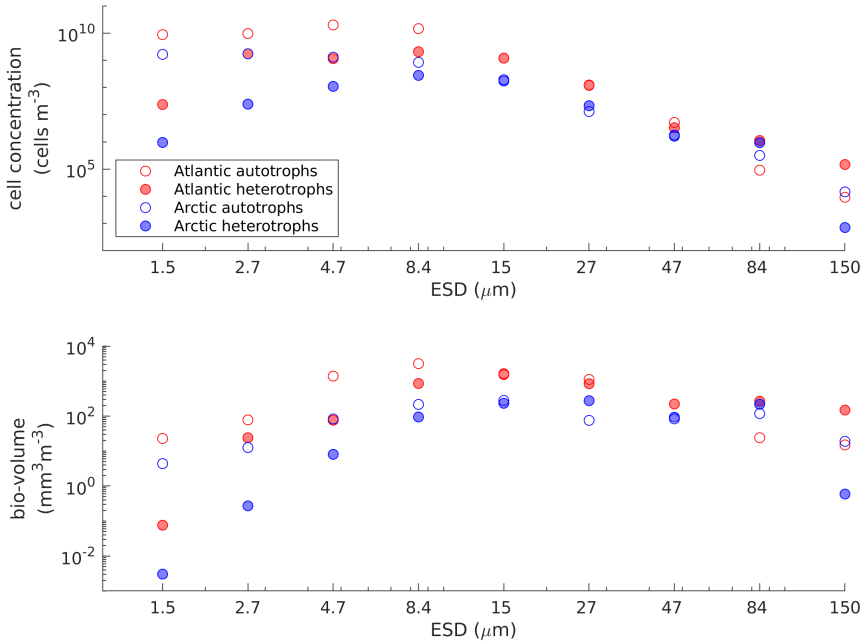


Figure 3: Observed plankton density-at-size. Points associated to each size group display total plankton within cell size intervals that are equally sized and evenly spaced on a logarithmic scale.

of 215 (for 2018) trajectories used as forcing data. The number of trajectories within this set is not divisible by ten due to several duplicate trajectories that were used for multiple sampling events. Such duplicates reduced the total number of required trajectories, further reducing computational burden. These may also be considered as the most "useful" trajectories because they pass through multiple sampling events, thus, it may be possible to observe temporal shifts in plankton community structure along these trajectories *and* fit such shifts to in-situ data (we have not attempted this, but mention it as a possibility).

2.2 Model

2.2.1 Flux equation system

Nutrient fluxes are described by a differential equation system (eqs. (1) to (3)) closely based on Ward et al. (2012) and Ward and Follows (2016). The state variables are concentrations of inorganic nitrogen, N, plankton, $B_{i,j}$, and organic matter, $M_{i,k}$. Concentrations have units of mmol element m^{-3} or mg chlorophyll a m^{-3}. Subscripts i, j, and k respectively

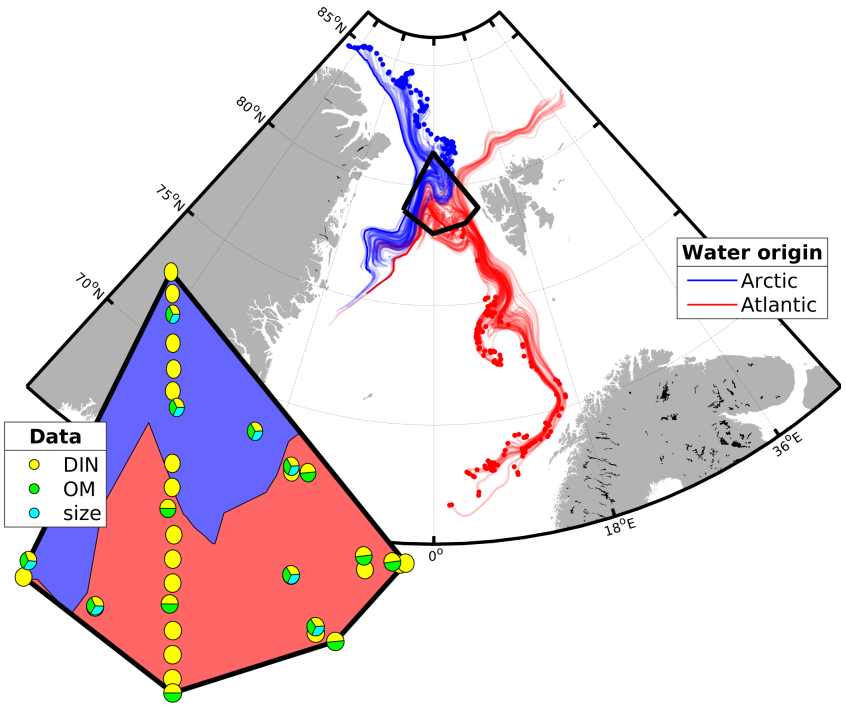


Figure 4: The Fram Strait study area. Ship-board sample sites are surrounded by The black polygon surrounds ship-board sample sites. The map inset magnifies this polygon to display positions of each 2018 sample and the data types measured at each site. Particle trajectories from the physical model are displayed as lines representing horizontal transport from 1$^{\text{st}}$ Jan. until 31$^{\text{st}}$ Oct. [check this date]. Points at one end of each line indicate initial locations of trajectories, which all lay within the polygon at the time of sampling. Trajectories originating from the Arctic and Atlantic are coloured blue and red to represent cold and warm water masses, which is also indicated in the magnified map inset.

index nutrients, plankton cell sizes, and type of organic matter. Nutrients modelled as planktonic and organic matter variables are: $i \in \{\mathrm{C}, \mathrm{N}\,\mathrm{Chl}\}$ for autotrophs; and $i \in \{\mathrm{C}, \mathrm{N}\}$ for heterotrophs and organic matter. Autotrophic and heterotrophic plankton are combined into a single matrix, $\boldsymbol{B}$, where the autotrophs are stacked atop the heterotroph variables in the j

size dimension. Thus, there are n_p autotroph and n_z heterotroph size classes indexed by $j \in \{\boldsymbol{j_p}, \boldsymbol{j_z}\} = \{1, \ldots, n_p, n_p + 1, \ldots, n_p + n_z\}$. Dissolved and particulate organic matter are indexed by $k \in \{\mathrm{DOM}, \mathrm{POM}\}$. All terms in eqs. (1) to (3) are defined in table 2, and model parameters are defined in tables 3 and 4.

$$\frac{\partial N}{\partial t} = \frac{\partial}{\partial z} K \frac{\partial N}{\partial z} - \sum_j V_{\mathrm{N},j}\, B_{\mathrm{C},j} + \sum_k r_{\mathrm{N},k}\, M_{\mathrm{N},k} \tag{1}$$

$$\frac{\partial B_{i,j}}{\partial t} = \frac{\partial}{\partial z} K \frac{\partial B_{i,j}}{\partial z} - w_{Pj} \frac{\partial B_{i,j}}{\partial z} + V_{i,j}\, B_{\mathrm{C},j} - \sum_{j_z} G_{i,j_z,j}\, B_{\mathrm{C},j_z} + \sum_j \lambda_{i,j_z}\, G_{,ij_z,j}\, B_{\mathrm{C},j_z} - m_j\, B_{i,j} \tag{2}$$

$$\frac{\partial M_{i,k}}{\partial t} = \frac{\partial}{\partial z} K \frac{\partial M_{i,k}}{\partial z} - w_k \frac{\partial M_{i,k}}{\partial z} - r_{i,k}\, M_{i,k} + S^{\mathrm{M}}_{i,k} \tag{3}$$

Table 2: Definition of all terms in flux equations (eqs. (1) to (3)).

Notation	Description
K	Vertical diffusivity from physical model
$V_{i,j}$	Nutrient uptake rate
$r_{i,k}$	Remineralisation rates of DOM and POM
$G_{i,j_z,j}$	Grazing rate of heterotrophs j_z on plankton j
λ_{i,j_z}	Assimilation efficiency of heterotrophs
m_j	Background mortality rate
w_{Pj}	Sinking rate of plankton
w_k	Sinking rate of organic matter
$S^{\mathrm{M}}_{i,k}$	Sources of organic matter
t	Time
z	Depth

2.2.2 Cell quotas & limiting terms

Plankton biomass is tracked as the carbon concentration, $B_{\mathrm{C},j}$. Concentrations of other nutrients vary, relative to carbon, within limits preventing excessive accumulation or depletion of any one nutrient. Cellular quotas of nitrogen and chlorophyll are modelled as ratios with carbon.

$$Q_{i,j} = \frac{B_{i,j}}{B_{\mathrm{C},j}} \tag{4}$$

Nitrogen quotas may vary between limiting values $Q^{\min}_{N,j}$ and $Q^{\max}_{N,j}$. Positive correlation between $Q^{\min}_{N,j}$ and $Q^{\max}_{N,j}$ hinders numerical optimisation of these

parameters. To reduce correlation we define $\widetilde{Q}^{\max}_{N,j} = Q^{\max}_{N,j}/(Q^{\max}_{N,j} - Q^{\min}_{N,j})$, and optimise parameters $Q^{\min}_{N,j}$ and $\widetilde{Q}^{\max}_{N,j}$. Maximum nitrogen quotas are then calculated as

$$Q^{\max}_{N,j} = \frac{Q^{\min}_{N,j}}{1 - 1/\widetilde{Q}^{\max}_{N,j}} \tag{5}$$

Nutrient limitation of production is a linear function of nitrogen quota.

$$\gamma_{\mathrm{N},j} = \frac{Q_{\mathrm{N},j} - Q^{\min}_{\mathrm{N},j}}{Q^{\max}_{\mathrm{N},j} - Q^{\min}_{\mathrm{N},j}} \tag{6}$$

Nutrient uptake rates are down-regulated as cell quotas increase, and are zeroed when quotas are full. The nutrient uptake regulation terms are defined as

$$Q^{\mathrm{stat}}_{\mathrm{N},j} = 1 - \left(\frac{Q_{\mathrm{N},j} - Q^{\min}_{\mathrm{N},j}}{Q^{\max}_{\mathrm{N},j} - Q^{\min}_{\mathrm{N},j}} \right)^{1/h} \tag{7}$$

which are modified from Ward and Follows (2016) by reflecting the $Q^{\mathrm{stat}}_{\mathrm{N},j}$ $(\gamma_{\mathrm{N},j})$ curves across the $Q^{\mathrm{stat}}_{\mathrm{N},j} = \gamma_{\mathrm{N},j}$ line. This modification was made solely for numerical stability, and the resulting uptake regulation curve retains a similar shape to those used by Ward and Follows (2016).

Temperature, T, influences nutrient uptake, photosynthesis, and grazing rates, which are adjusted by a temperature regulation term

$$\gamma_T = e^{A(T - T^{\mathrm{ref}})} \tag{8}$$

where T^{ref} is a fixed reference temperature and A is temperature sensitivity.

2.2.3 Nutrient uptake

Nutrient uptake is modelled with Michaelis-Menton functions, modified by quota- and temperature-limitation terms.

$$V_{\mathrm{N},j} = \frac{v^{\max}_j \, \alpha_j \, N}{\alpha_j \, N + v^{\max}_j} \, Q^{\mathrm{stat}}_{\mathrm{N},j} \, \gamma_T \tag{9}$$

All uptake rates of heterotrophs are set to zero, $V_{i,j_z} = 0$.

2.2.4 Photosynthesis

Carbon-specific light-saturated photosynthetic rate is modelled as a size-dependent maximum rate restricted by temperature and nitrogen quota limitation terms.

$$P^{\mathrm{sat}}_j = P^{\max}_j \, \gamma_T \, \gamma_{\mathrm{N},j} \tag{10}$$

Photosynthetic rate is defined as a Poisson function of irradiance, I, and chlorophyll quota.

$$P_j = P_j^{\text{sat}} \left(1 - \exp\left(\frac{-\alpha_p \, Q_{\text{Chl},j} \, I}{P_j^{\text{sat}}}\right)\right) \tag{11}$$

Carbon-specific production rate is then defined as the difference between photosynthetic rate and the cost of biosynthesis

$$V_{\text{C},j} = P_j - \xi \, V_{\text{N},j} \tag{12}$$

where cost of biosynthesis is a linear function of nitrogen uptake.

Chlorophyll production is coupled to nitrogen uptake

$$V_{\text{Chl},j} = \rho_j \, V_{\text{N},j} \tag{13}$$

and is progressively down-regulated from maximum, θ, as the photosynthetic rate, p_j, decreases below the theoretical maximum-efficiency rate, $\alpha_p \, Q_{\text{Chl,j}} \, I$, at high irradiances.

$$\rho_j = \theta \frac{P_j}{\alpha_p \, Q_{\text{Chl},j} \, I} \tag{14}$$

2.2.5 Predation

Grazing rates of prey carbon are the product of predator's maximum grazing rates and prey saturation, switching, and refuge terms.

$$G_{\text{C},j_z,j} = \underbrace{\gamma_T \, G_{j_z}^{\max}}_{\text{max. rate}} \underbrace{\frac{F_{C,j_z}}{k_G + F_{C,j_z}}}_{\text{saturation}} \underbrace{\Phi_{j_z,j}}_{\text{switching}} \underbrace{\left(1 - e^{\Lambda \, F_{C,j_z}}\right)}_{\text{refuge}} \tag{15}$$

Prey saturation is modelled with Michaelis-Menton functions of total prey carbon available to each predator

$$F_{C,j_z} = \sum_j \phi_{j_z,j} \, B_{C,j} \tag{16}$$

where $\phi_{j_z,j}$ is the availability of each prey class, j, to each predator class, j_z. Prey availability is modelled as a function of predator-to-prey diameter ratios, $\delta_{j_z,j}$

$$\phi_{j_z,j} = \exp\left[-\left(\ln\left(\frac{\delta_{j_z,j}}{\delta_{\text{opt}}}\right)\right)^2 / \left(2\sigma^2\right)\right] \tag{17}$$

where δ_{opt} and σ are the optimum ratio maximising prey availability, and variability around the optimum. As this function is a log-normal probability density without the scaling terms, it has maximum value of 1 and a log-normal shape.

The prey switching term regulates predation losses by targeting grazing on the most abundant prey classes.

$$\Phi_{j_z,j} = \frac{(\phi_{j_z,j}\, B_{C,j})^2}{\sum_j (\phi_{j_z,j}\, B_{C,j})^2} \tag{18}$$

This prevents overgrazing any single prey class when other available prey are more abundant.

The prey refuge term prevents overgrazing by reducing predator grazing rates when total available prey, F_{C,j_z}, is low.

Carbon-specific grazing rates of nitrogen and chlorophyll are calculated as the product of carbon grazing rate and cell quotas.

$$G_{i,j_z,j} = Q_{i,j}\, G_{C,j_z,j} \tag{19}$$

Assimilation efficiency of consumed prey is given by

$$\lambda_{C,j_z} = \lambda^{\max}\, \gamma_{N,j} \tag{20}$$

$$\lambda_{N,j_z} = \lambda^{\max}\, Q^{\text{stat}}_{N,j_z} \tag{21}$$

which down-regulates prey assimilation from the maximum, $\lambda^{\max}$, when predator cell quotas approach their limits.

2.2.6 Background mortality

The background mortality, $m_j\, B_{i,j}$, is linear with respect to plankton abundance, and the mortality rate, m_j, is size dependent. This size dependency is a modification of the background mortality rates from Ward et al. (2012) and Ward and Follows (2016), who used scalar values. We assume that background mortality rate of large cells is less than or equal to that of smaller cells, $m_j \leq m_{j-1}$. Background mortality rate is modelled using a power function of volume with constraints $m_a > 0$ and $m_b \leq 0$

$$m_j = m_{\min} + (m_a - m_{\min})\, \text{Vol}_j^{m_b} \tag{22}$$

where a minimum permissible mortality, $m_{\min}$, prevents $m_j \to 0$ as volume increases.

2.2.7 Organic matter

Organic matter is generated from mortality and messy feeding. As cells die, all their unassimilated nutrient is transferred into organic matter and allocated into DOM and POM categories. The proportions, $\beta_{j,k}$, allocated to DOM and POM are modelled as volume-dependent using a three-parameter

double-logistic function

$$\beta_{j,\mathrm{DOM}} = \frac{b_1}{1+e^{(x-b_3)}} + \frac{b_1\, b_2}{1+e^{(b_3-x)}} \tag{23}$$

$$\beta_{j,\mathrm{POM}} = 1 - \beta_{j\mathrm{DOM}} \implies \sum_k \beta_{j,k} = 1 \tag{24}$$

where $x = \log_{10}(\mathrm{Vol})$, and where $0 < b_1 < 1$ and $0 < b_1\, b_2 < 1$ are necessary constraints. By constraining $0 < b_2 < 1$, we ensure that $\beta_{j,\mathrm{DOM}}$ decreases monotonically with cell volume to enforce the assumption that, upon expiration, small cells produce proportionally more DOM than relatively large cells.

Sources of organic matter are the sum of mortality and messy feeding terms

$$S^{\mathrm{M}}_{i,k} = \underbrace{\sum_j \beta_{j,k}\, m_j\, B_{i,j}}_{\text{mortality}} + \underbrace{\sum_{j_z} B_{C,j_z} \sum_j \beta_{j,k}\, (1-\lambda_{i,j_z})\, G_{i,j_z,j}}_{\text{messy feeding}} \tag{25}$$

where $(1-\lambda_{i,j_z})\, G_{i,j_z,j}$ are the predator carbon-specific rates of organic matter production through messy feeding.

2.3 Parameter optimisation

2.3.1 Data standardisation

Model parameter values were numerically optimised to minimise discrepancies between model outputs and data. The scalar data, $\widehat{Y}_{d,s,g}$, were standardised with respect to the depth and sampling event covariates using LMMs (linear mixed models). Standardising the data eliminated the possibility of parameter estimates being biased by variability attributable to depth and sample event or to differences between the typical magnitude of different data types (fig. 1).

The nitrate, PON, POC, and chlorophyll a data groups each follow the same sampling design. They contain measurements from unique sampling events, where each sample is a set of single measurements taken at a range of depths. Let $Y_{d,s,g}$ denote these data, where $d \in \{1, \ldots, d_n\}$, $s \in \{1 \ldots, s_n\}$, and $g \in \{\mathrm{N}, \mathrm{PON}, \mathrm{POC}, \mathrm{Chl}\ a\}$ index depth, sampling event, and data group. As they vary with depth, z, and sampling event, measurements from each data group, Y_g, are not independent. Optimising parameters using the raw data would therefore bias results because the depth- and sample-dependencies create unevenly weighted data points. We accounted for depth- and sample-dependent variability in the data by using linear mixed models to generate standardised data sets, $\widetilde{Y}_g$. Using $\widetilde{Y}_g$ as the fitting data limits potential parameter estimation bias due to the depth and sample covariates

Table 3: Size-independent parameters. Numerically optimised parameter values are displayed above their bounds. Values displayed without bounds are fixed parameters.

Parameter	Notation	Value	Units
Rate-limiting parameters			
Reference temperature	T^{ref}	20	°C
Temperature sensitivity	A	0.05	dimensionless
Uptake regulation curvature	h	0.1	dimensionless
Photosynthesis			
Initial slope of P-I curve	α_p	0.24 $[0, 0.5]$	mmol C (mg Chl a)$^{-1}$
Cost of biosynthesis[(a)]	ξ	2.33	mmol C (mmol N)$^{-1}$
Max. Chl a-to-nitrogen ratio[(a)]	θ	4.2	mg Chl a (mmol N)$^{-1}$
Grazing			
Optimum predator:prey length ratio[(b)]	δ_{opt}	10	dimensionless
Geometric SD of prey availability	σ	2	dimensionless
Total prey half saturation	k_G	8.39 $[0.5, 10]$	mmol C m^{-3}
Prey refuge parameter	Λ	-1	dimensionless
Max. assimilation efficiency[(a)]	$\lambda^{\max}$	0.7	dimensionless
Organic matter			
DOM sinking speed	w_{DOM}	0	m d^{-1}
POM sinking speed	w_{POM}	0.51 $[0.5, 10]$	m d^{-1}
DOM remineralisation rates	$r_{N,\text{DOM}}$	0.052 $[0.005, 0.06]$	d^{-1}
	$r_{C,\text{DOM}}$	$r_{N,\text{DOM}}$	d^{-1}
POM remineralisation rates	$r_{N,\text{POM}}$	0.078 $[0.01, 0.12,]$	d^{-1}
	$r_{C,\text{POM}}$	$r_{N,\text{POM}}$	d^{-1}

[(a)] Geider et al. (1998); [(b)] Kiørboe (2008)

(fig. 5). One of two linear mixed models were fitted separately to each data group

$$\begin{aligned} \ln(Y_{d,s,g}) &= (a_g + a_{s,g}) + (b_g + b_{s,g})\, z_{d,s,g} + \epsilon_{d,s,g} \qquad (26) \\ &= \mu_{d,s,g} + \epsilon_{d,s,g} \\ \text{or} \quad Y_{d,s,g} &= (a_g + a_{s,g}) + (b_g + b_{s,g})\, \ln(z_{d,s,g}) + \epsilon_{d,s,g} \qquad (27) \\ &= \mu_{d,s,g} + \epsilon_{d,s,g} \end{aligned}$$

where a_g and b_g are the "fixed" effects of depth upon measured values, $a_{s,g}$ and $b_{s,g}$ are the "random" effects associated with sampling event, $\epsilon_{d,s,g} \sim \mathcal{N}(0, \sigma_g)$ are normally distributed residual errors, and σ_g are the residual error standard deviations. The standardised data, $\widetilde{Y}_{d,s,g}$, are equivalent to the standard-normal distributed residual errors. These are calculated from

Table 4: Size-dependent parameters ($x = a\,\mathrm{Vol}^b$). Numerically optimised parameter values are displayed above their bounds. Values displayed without bounds are fixed parameters.

Parameter	Notation	a	b	Units
Nutrient quotas				
Carbon quota[a]	Q_C	1.7×10^{-11}	0.88	mmol C cell^{-1}
Min. nitrogen:carbon quota[a]	$Q_N^{\min}$	0.15 $[0.07, 0.23]$	-0.10 $[-0.11, 0.04]$	mmol N (mmol C)$^{-1}$
Max. N:C quota transform[a,b]	$\widetilde{Q}_N^{\max}$	0.98 $[0.16, 1.00]$	-0.0021 $[-0.19, -0.001]$	dimensionless
Nutrient uptake				
Max. uptake rate[a]	$v^{\max}$	0.093 $[0.067, 0.167]$	0.166 $[0.01, 0.18]$	mmol N (mmol C)$^{-1}$
Nutrient affinity[c]	α	0.88 $[0.28, 1.60]$	-0.30 $[-0.36, -0.15]$	m^3 (mmol C)$^{-1}$
Photosynthesis				
Max. photosynthetic rate	$P^{\max}$	1.37 $[0.5, 5]$	-0.098 $[-0.5, -0.01]$	d^{-1}
Grazing				
Max. prey capture rate	$G^{\max}$	5.4 $[5, 35]$	-0.19 $[-0.5, -0.01]$	d^{-1}
Mortality				
Background mortality rate[d]	m	0.05	-0.013 $[-1, -0.01]$	d^{-1}
Sinking				
Plankton sinking rate	w_P	3.3×10^{-5} $[0, 3.4 \times 10^{-5}]$	0.27 $[0, 0.67]$	m d^{-1}

[a] Marañón et al. (2013); [b] Transformed $Q_N^{\max}$ is defined in section 2.2.2;
[c] Litchman et al. (2007); [d] m is defined by non-standard power function (section 2.2.6)

the raw data and the fitted linear mixed models parameters.

$$\widetilde{Y}_{d,s,g} = \frac{1}{\sigma_g}(\ln(Y_{d,s,g}) - \mu_{d,s,g}) \tag{28}$$

$$\text{or} \quad \widetilde{Y}_{d,s,g} = \frac{1}{\sigma_g}(Y_{d,s,g} - \mu_{d,s,g}) \tag{29}$$

The choice of model (eqs. (26) and (28) or eqs. (27) and (29)) depended on which one linearised the data from each group. The PON, POC, and chlorophyll a concentrations tended to decrease with sample depth, and the concentration-depth relationships were approximately linearised by log-transforms of the concentrations. In contrast, as nitrate concentration tended to increase with depth, a log-transform of the independent variable, depth, approximately linearised the concentration-depth relationship. Equations (26) and (28) were therefore used to standardise the PON, POC, and chlorophyll a measurements, while nitrate concentrations were standardised with eqs. (27) and (29). Each standardised data set is distributed as approximately standard normal, $\widetilde{Y}_g \sim \mathcal{N}(0, 1)$. Thus, if the corresponding model outputs are transformed identically to the data, so that standardised values are used within the cost function, then no data group, sampling event, or

depth should bias the fitting process, i.e., all data points have approximately equal weight.

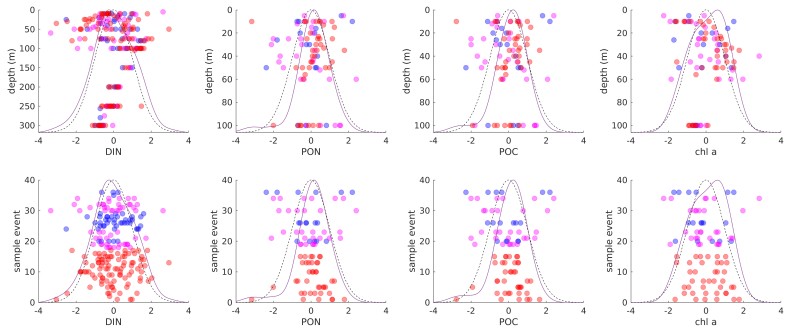


Figure 5: Data after standardisation using linear mixed effects models

2.3.2 Cost function

Hellinger distance method

Model fit to data was optimised by numerically minimising a cost function. The cost function returns a scalar, the "cost", representing discrepancy between data, $\boldsymbol{Y}$, and the modelled approximations to those data, $\widehat{\boldsymbol{Y}}$ — hatted variables ($\widehat{\ }$) denote modelled outputs. As $\widehat{\boldsymbol{Y}}$ depends upon the choice of model parameters, numerical algorithms can be used to search for parameter values that minimise the cost to produce close fits to data. The abundance-at-size data and nutrient data differ in structure and sampling design, thus required different cost function terms.

For data group $g \in \{\mathrm{N}, \mathrm{PON}, \mathrm{POC}, \mathrm{Chl}\ a\}$ there is an observation vector $\boldsymbol{Y}_g$ of length n_g. Let these observation vectors be the standardised data, $\boldsymbol{Y}_g \equiv \widetilde{\boldsymbol{Y}}_g$, (eqs. (28) and (29)) where the tilde ($\widetilde{\ }$) notation is dropped in this section. Note that since the standardised data are independent of depth and sample event the d and s indices (section 2.3.1) are omitted and the data are denoted as vectors. For each set of particle trajectories, $q \in \{1, \ldots, n_q\}$, the model returns vectors, $\widehat{\boldsymbol{Y}}_{g,q}$, containing the modelled equivalents to observed $\boldsymbol{Y}_g$. Since $\boldsymbol{Y}_g \sim \mathcal{N}(0,1)$, close fits to data would produce $\widehat{\boldsymbol{Y}}_{g,q}$ values that also have approximate standard normal distributions. As the standardised data, $\boldsymbol{Y}_g$, and the abundance-at-size data are both easily represented as distributions, we used Hellinger distances as cost function terms. Hellinger distance, $H(\boldsymbol{p}, \boldsymbol{q})$, is a bounded scalar metric representing the misfit between probability densities $\boldsymbol{p}$ and $\boldsymbol{q}$.

$$H(\boldsymbol{p}, \boldsymbol{q}) \equiv \left(1 - \sum_i \left(p_i\, q_i\right)^{\frac{1}{2}}\right)^{\frac{1}{2}} \tag{30}$$

Hellinger distance is maximised, $H(\boldsymbol{p},\boldsymbol{q})=1$, for disparate $\boldsymbol{p}$ and $\boldsymbol{q}$ ($p_i\,q_i = 0\;\forall i$), and is minimised, $H(\boldsymbol{p},\boldsymbol{q})=0$, when $\boldsymbol{p}=\boldsymbol{q}$.

For each $\boldsymbol{Y}_g$ and $\widehat{\boldsymbol{Y}}_{g,q}$ we calculated a probability density vector, $\boldsymbol{\mathcal{P}}_g$ or $\widehat{\boldsymbol{\mathcal{P}}}_{g,q}$ as follows. Cumulative distribution vectors, $\boldsymbol{c}^r$, were generated from a composition of two functions

$$f(\boldsymbol{x}) = \text{sort}(\boldsymbol{x})\ \text{(ascending)} \tag{31}$$

$$g(x_i) = \frac{\sum_{j=1}^{i} x_j}{\sum_j x_j} \tag{32}$$

so that $\boldsymbol{c}_g^r = g(f(\boldsymbol{Y}_g))$ and $\hat{\boldsymbol{c}}_{g,q}^r = g(f(\widehat{\boldsymbol{Y}}_{g,q}))$ are observed and modelled cumulative distributions. These are denoted with an r superscript to indicate that these are "raw" distributions. To compare distributions $\boldsymbol{c}_g^r$ and $\hat{\boldsymbol{c}}_{g,q}^r$ we first need to interpolate them over a uniform grid to generate comparable vectors, $\boldsymbol{c}_g$ and $\hat{\boldsymbol{c}}_{g,q}$, that contain probability values at matching grid locations. We define interpolation grids of n_g nodes spanning the interval $[\min(\boldsymbol{Y}_g, \widehat{\boldsymbol{Y}}_{g,q}), \max(\boldsymbol{Y}_g, \widehat{\boldsymbol{Y}}_{g,q})]$, then linearly interpolate the raw $\boldsymbol{c}_g^r = g(f(\boldsymbol{Y}_g))$ and $\hat{\boldsymbol{c}}_{g,q}^r = g(f(\widehat{\boldsymbol{Y}}_{g,q}))$ values to produce comparable cumulative distribution vectors, $\boldsymbol{c}_g$ and $\hat{\boldsymbol{c}}_{g,q}$. Probability density vectors were calculated as

$$\mathcal{P}_{g(1)} = c_{g(1)};\ \widehat{\mathcal{P}}_{g,q(1)} = \hat{c}_{g,q(1)} \tag{33}$$

$$\mathcal{P}_{g(i)} = c_{g(i)} - c_{g(i-1)};\ \widehat{\mathcal{P}}_{g,q(i)} = \hat{c}_{g,q(i)} - \hat{c}_{g,q(i-1)},\ i>1. \tag{34}$$

Hellinger distances were calculated as $H_{g,q} = H(\boldsymbol{\mathcal{P}}_g, \widehat{\boldsymbol{\mathcal{P}}}_{g,q})$, then averaged over particle trajectories, $H_g = \frac{1}{n_q}\sum_q H_{g,q}$. We averaged over the four groups to produce a single cost term describing combined model misfit to all nutrient data, $H^{\text{nutrient}} = \frac{1}{4}\sum_{g=1}^{4} H_g$.

The plankton size data are vectors, $\boldsymbol{Y}_g$, of measured biovolume for n_g size classes, where g indexes four groups: autotrophs and heterotrophs from waters of Atlantic or Arctic origin. These data and their modelled approximations, $\widehat{\boldsymbol{Y}}_{g,q}$, are decomposed into scalars of total biovolume, $Y_g^{\text{tot}} = \sum_j Y_{g(j)}$, and simplices of relative biovolume-at-size, $\boldsymbol{Y}_g^s = \boldsymbol{Y}_g / Y_g^{\text{tot}}$. The cost terms for relative biovolume-at-size were calculated as Hellinger distances, $H_{g,q} = H(\boldsymbol{Y}_g^s, \widehat{\boldsymbol{Y}}_{g,q}^s)$, between the simplices $\boldsymbol{Y}_g^s$ and $\widehat{\boldsymbol{Y}}_{g,q}^s$, which are equivalent to probability mass distributions. Averaging over sets of particle trajectories produces cost terms describing average model misfit to each group of relative abundance data, $H_g = \frac{1}{n_q}\sum_q H_{g,q}$. Then averaging over trophic group and water origin produces a single cost term describing overall model misfit to relative biovolume-at-size, $H^{\text{size}} = \frac{1}{4}\sum_g H_g$.

The total biovolume components, Y_g^{tot}, of decomposed size data are single scalar values. As these data are not distributions, Hellinger distances cannot be used to gauge model misfit. Instead, we used a metric, J, with

similar properties.

$$J(x,y) \equiv \frac{1-\exp(-a\,u)}{1+\exp(-a\,u)} \tag{35}$$
$$u = |\ln(x/y)|$$
$$a = \ln(3)/\ln(2)$$

Like Hellinger distances, this metric is symmetric and bounded in the $[0,1]$ interval, with $x = y \implies J(x,y) = 0$, and $x \gg y$ or $x \ll y \implies J(x,y) = 1$. Thus, H and J metrics are easily combined in a single cost function. Metric J has a shape parameter, a, chosen such that $x/y = 1/2$ or $x/y = 2 \implies J(x,y) = 1/2$, thus J lies at the centre of it range when modelled values are $1/2$ or 2 times the observed values. The cost terms for total biovolume were calculated as $J_{g,q} = J\left(\widehat{Y}^{\text{tot}}_{g,q}, Y^{\text{tot}}_{g}\right)$. The average cost of model misfit to total biovolume for each group — autotrophs and heterotrophs in Atlantic and Arctic waters — is found by averaging over particle trajectories, $J_g = \frac{1}{n_q}\sum_q J_{g,q}$. Then a single cost component representing overall model fit to total biovolume is calculated by averaging over groups, $J^{\text{size}} = \frac{1}{4}\sum_g J_g$.

The cost function returns a single scalar value, $0 \le \mathcal{C} \le 1$, representing overall model misfit.

$$\mathcal{C} = \frac{1}{2}\left(H^{\text{nutrient}} + \frac{1}{2}\left(H^{\text{size}} + J^{\text{size}}\right)\right) \tag{36}$$

The cost function construction gives equal weight to nutrient and size data: the averaged cost of all nutrient data is weighted equally to the averaged cost of all biovolume-at-size data. Within the size components of cost, the total and the relative biovolumes-at-size are also ascribed equal weightings.

Likelihood-based method

Model fit to data was optimised by numerically minimising a cost function. The cost function returns values representing the discrepancy between the data $\widetilde{Y}^{\text{obs}}$, and the modelled approximations to those data, $\widetilde{Y}$. As $\widetilde{Y}$ depends upon the choice of model parameters, numerical algorithms can be used to seek out parameter values that minimise the cost to produce close fits to data. We define a "synthetic" likelihood function (Wood, 2010) that includes a separate term for each data type. The standardised "scalar" data are described using Gaussian distributions, and the size spectra data are described using Dirichlet and lognormal distributions.

The likelihood function

$$\begin{aligned}\mathcal{L} = &\left(\prod \mathcal{N}\left(\tilde{N}^{\text{obs}}|\tilde{N},\sigma^2_N\right)\right)^{n_N^{-1}} \cdot \left(\prod \mathcal{N}\left(\tilde{M}^{\text{obs}}_{C,\text{POM}}|\tilde{M}_{C,\text{POM}},\sigma^2_{M_C}\right)\right)^{n_{M_C}^{-1}} \cdots \\ &\left(\prod \mathcal{N}\left(\tilde{M}^{\text{obs}}_{N,\text{POM}}|\tilde{M}_{N,\text{POM}},\sigma^2_{M_N}\right)\right)^{n_{M_N}^{-1}} \cdot \left(\prod \mathcal{N}\left(\tilde{B}^{\text{obs}}_{\text{Chl},j_p}|\tilde{B}_{\text{Chl},j_p},\sigma^2_{\text{Chl}}\right)\right)^{n_{\text{Chl}}^{-1}} \cdots \\ &\text{Dir}(Y^{\text{obs}}_P|Y_P,c_P)\,.\,\text{Dir}(Y^{\text{obs}}_Z|Y_Z,c_Z)\,.\,\text{LN}\left(\tilde{Y}^{\text{obs}}_P|\tilde{Y}_P,\sigma^2_P\right)\,.\,\text{LN}\left(\tilde{Y}^{\text{obs}}_Z|\tilde{Y}_Z,\sigma^2_Z\right)\end{aligned} \tag{37}$$

is the product of four Gaussian likelihoods for each scalar data type, two separate Dirichlet distributions for the relative abundances of sizes classes of autotrophs and heterotrophs, and two lognormal distributions for the total abundances derived from the size spectra data.

Each of the distribution variability parameters were derived using the variabilities from running the model over all trajectories.

The cost function to numerically minimise is the negative log-likelihood.

$$\mathcal{C} = -\ln(\mathcal{L}) \tag{38}$$

2.3.3 Optimising algorithm

The cost function was minimised using a genetic algorithm. The *ga* MatLab function was used with the default settings.

3 Results

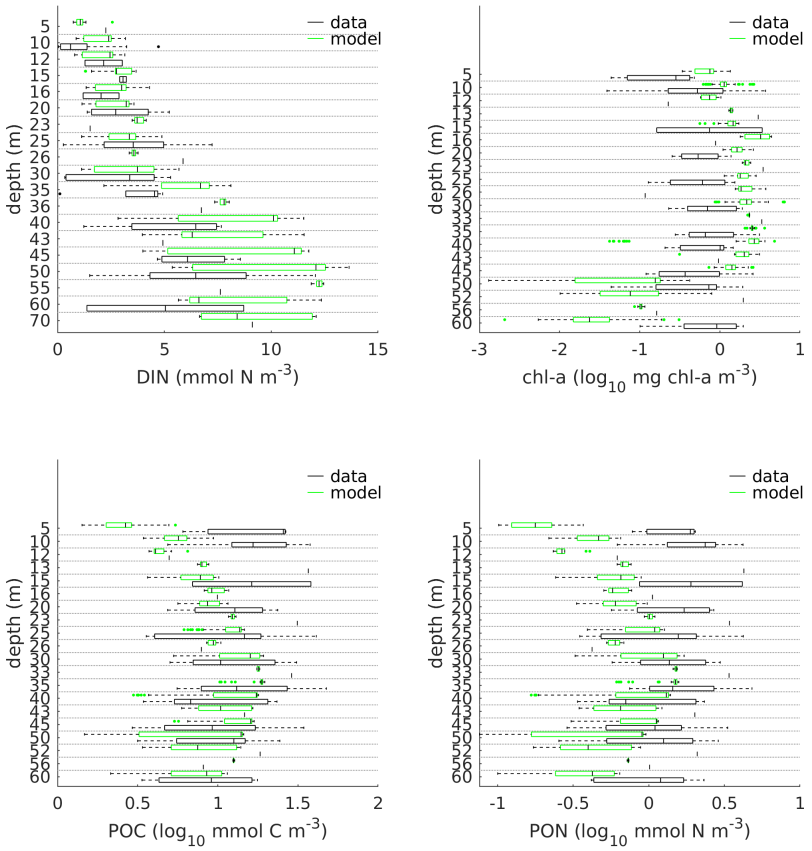


Figure 6: Model fit to depth-discrete, scalar observations.

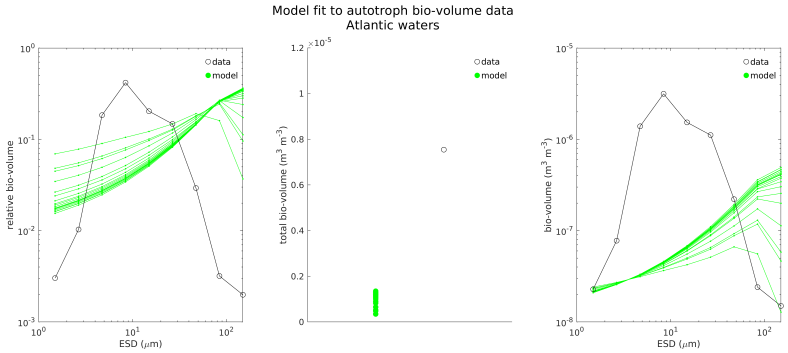


Figure 7: Model fit to Atlantic autotroph bio-volume data. Relative abundance (left); total abundance (centre); abundance-at-size (right).

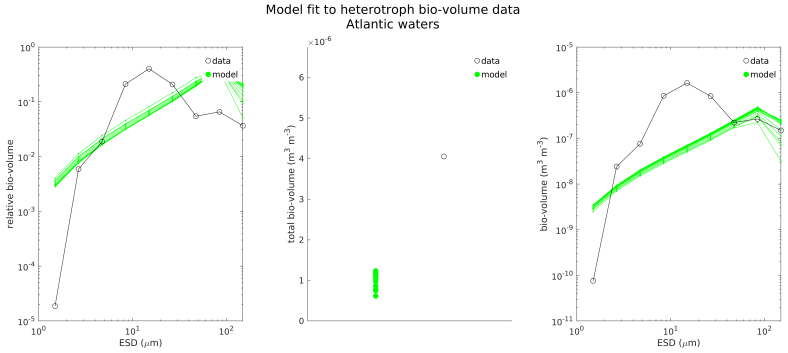
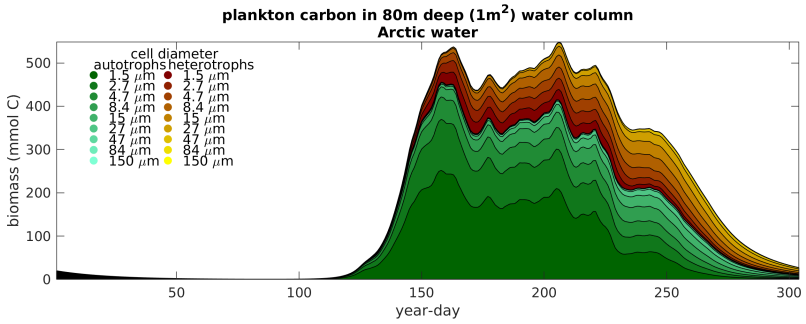


Figure 8: Model fit to Atlantic heterotroph bio-volume data. Relative abundance (left); total abundance (centre); abundance-at-size (right).



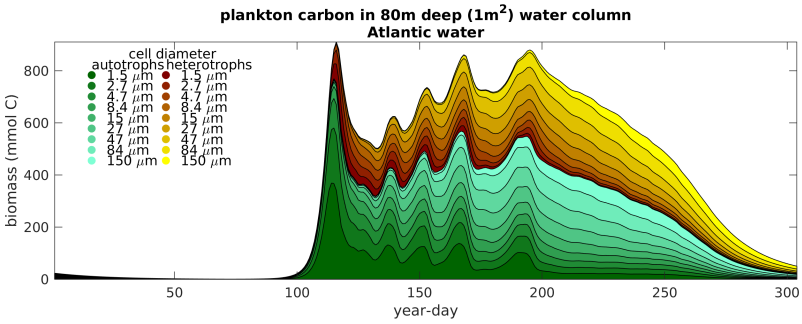


Figure 9: Time series of plankton biomass, averaged over trajectories originating from Arctic and from Atlantic waters.

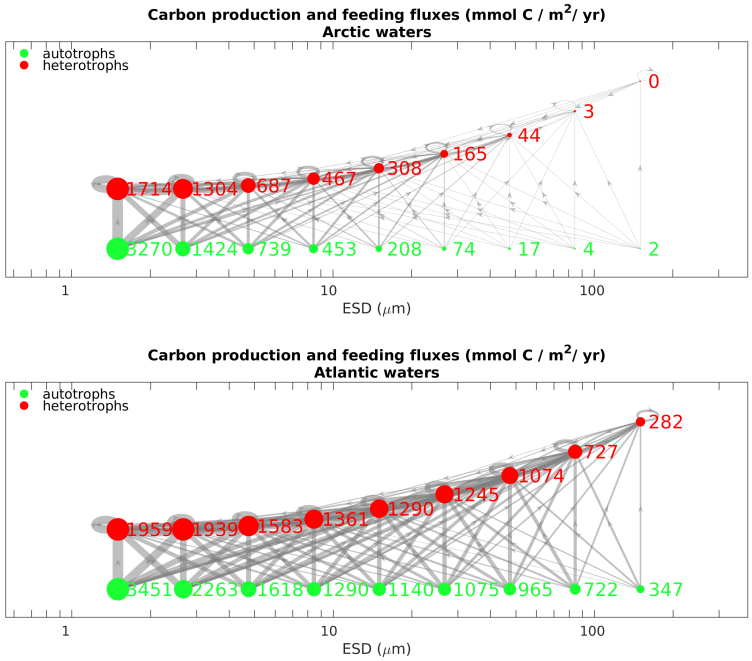


Figure 10: Total annual carbon production and feeding fluxes, averaged over trajectories originating from Arctic and from Atlantic waters.

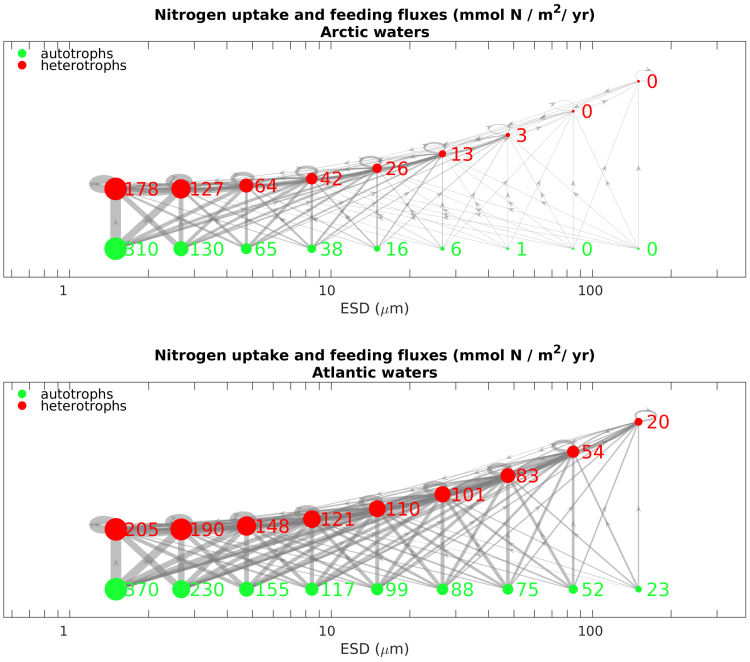


Figure 11: Total annual nitrogen production and feeding fluxes, averaged over trajectories originating from Arctic and from Atlantic waters.

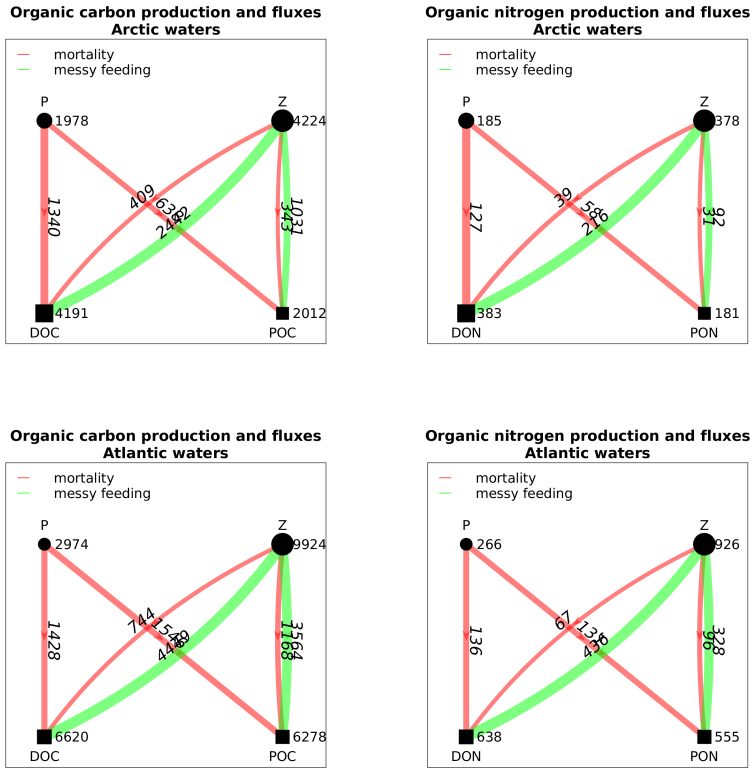


Figure 12: Total annual organic matter production and fluxes (mmol {C, N} / m^2 / yr) from autotrophs (P) and heterotrophs (Z).

4 Discussion

References

Geider, R. J., MacIntyre, H. L. and Kana, T. M. (1998). A dynamic regulatory model of photoacclimation to light, nutrients and temperature, *Limnology and Oceanography* **43**: 679–694.

Kiørboe, T. (2008). *A Mechanistic Approach to Plankton Ecology*, Princeton University Press.

Lampe, V., Nöthig, E.-M. and Schartau, M. (2021). Spatio-temporal variations in community size structure of arctic protist plankton in the fram strait, *Frontiers in Marine Science* **7**: 1239.

Litchman, E., Klausmeier, C. A., Schofield, O. M. and Falkowski, P. G.

(2007). The role of functional traits and trade-offs in structuring phytoplankton communities: scaling from cellular to ecosystem level, *Ecology Letters* **10**(12): 1170–1181.

Marañón, E., Cermeño, P., López-Sandoval, D. C., Rodríguez-Ramos, T., Sobrino, C., Huete-Ortega, M., Blanco, J. M. and Rodríguez, J. (2013). Unimodal size scaling of phytoplankton growth and the size dependence of nutrient uptake and use, *Ecology Letters* **16**: 371–379.

Torres-Valdés, S., Morische, A. and Wischnewski, L. (2019). Nutrient measurements from polarstern cruise ps114 (lter hausgarten).

von Jackowski, A., Grosse, J., Nöthig, E.-M. and Engel, A. (2020). Organic matter and bacteria measurements of polarstern cruise ps114 and maria s. merian cruise msm77.

Ward, B. A., Dutkiewicz, S., Jahn, O. and Follows, M. J. (2012). A size-structured food-web model for the global ocean, *Limnology and Oceanography* **57**(6): 1877–1891.

Ward, B. A. and Follows, M. J. (2016). Marine mixotrophy increases trophic transfer efficiency, mean organism size, and vertical carbon flux, *Proceedings of the National Academy of Sciences* **113**(11): 2958–2963.

Wood, S. (2010). Statistical inference for noisy nonlinear ecological dynamic systems, *Nature Letters* **466**(26): 1102–1104.