Supporting Information for "The impact of zooplankton calcifiers on the marine carbon cycle"

[1]Nielja S. Knecht, [1]Fabio Benedetti, [1]Urs Hofmann Elizondo, [2,3]Nina Bednaršek, [4,5,6]Sonia Chaabane, [7]Catharina de Weerd, [7,8]Katja T. C. A. Peijnenburg, [6]Ralf Schiebel, [1]Meike Vogt

[1]Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, Zurich, Switzerland

[2]National Institute of Biology, Marine Biological Station, Piran, Slovenia

[3]Cooperative Institute for Marine Resources Studies, Oregon State University, Oregon, USA

[4]Aix-Marseille Université, CNRS, IRD, INRAE, CEREGE, Aix-en-Provence, France

[5]French Foundation for Research on Biodiversity (FRB-CESAB), Paris, France

[6]Department of Climate Geochemistry, Max-Planck-Institute for Chemistry, Mainz, Germany

[7]Plankton Diversity and Evolution, Naturalis Biodiversity Center, Leiden, The Netherlands

[8]Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands

Contents of this file

1. Figures S1 to S26
2. Tables S1 to S8

Corresponding author: N. S. Knecht, Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, Zurich, Switzerland (nknecht@ethz.ch)

Introduction

The supporting information includes additional figures and tables relating to the original observation data, the abundance-to-biomass conversions and the modelling process. The outputs of various sensitivity analyses are also shown as described and referenced in the main text.

Figures

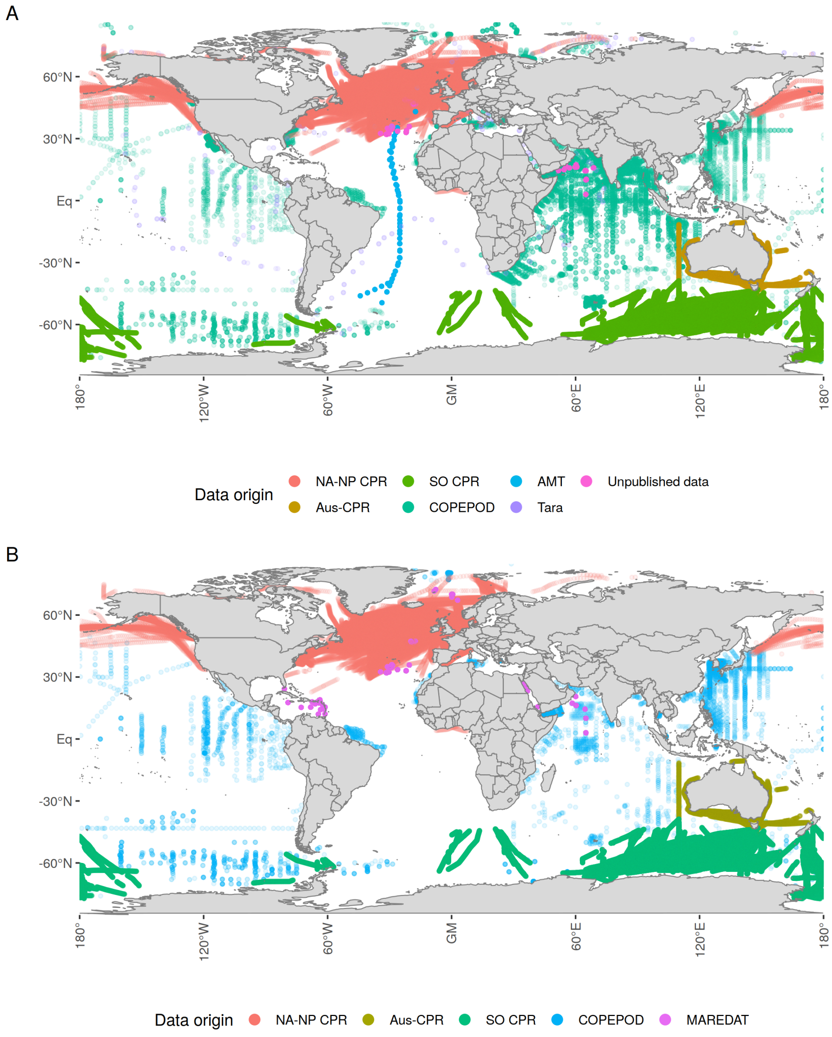


Figure S1. Pteropoda (**A**) and planktic foraminifers (**B**) data sources. CPR refers to the Continuous Plankton Recorder (NA-NP: North Atlantic and North Pacific, Aus: Australia, SO: Southern Ocean), COPEPOD to the Coastal and Oceanic Plankton Ecology, Production and Observation Database, AMT to the Atlantic Meridional Transect and MAREDAT to the MARine Ecosystem DATabase. See section 2.1.1 for more details.

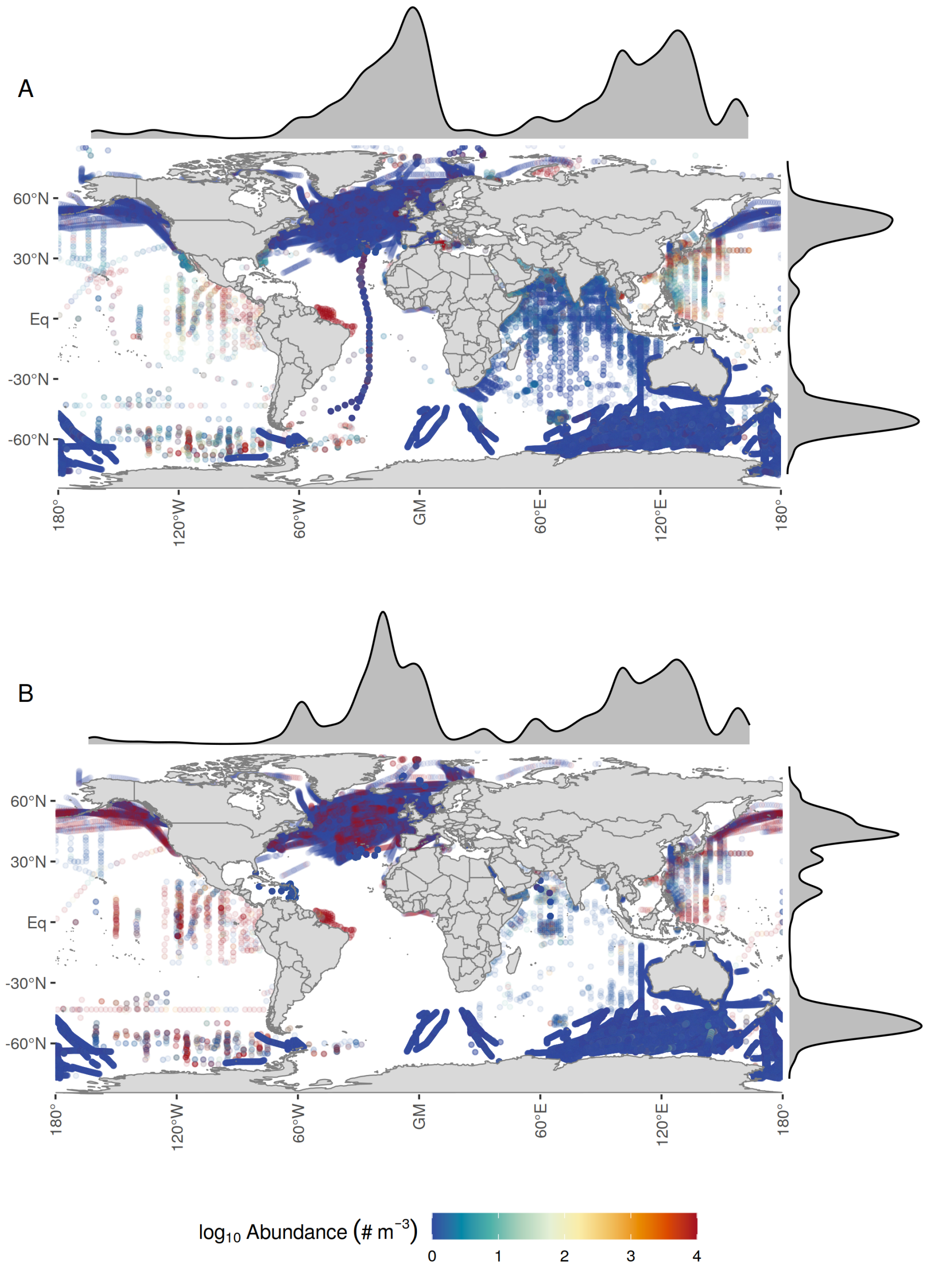


Figure S2. Pteropoda (**A**) and planktic foraminifers (**B**) abundance observation data from the full quality controlled AtlantECO dataset. The marginal plots show the density of observations and highlight the dominant role of the North Atlantic and North Pacific Continuous Plankton Recorder (NA-NP CPR) survey, the Southern Ocean CPR (SO-CPR) survey as well as a spatially confined, highly resolved dataset in the North Atlantic.

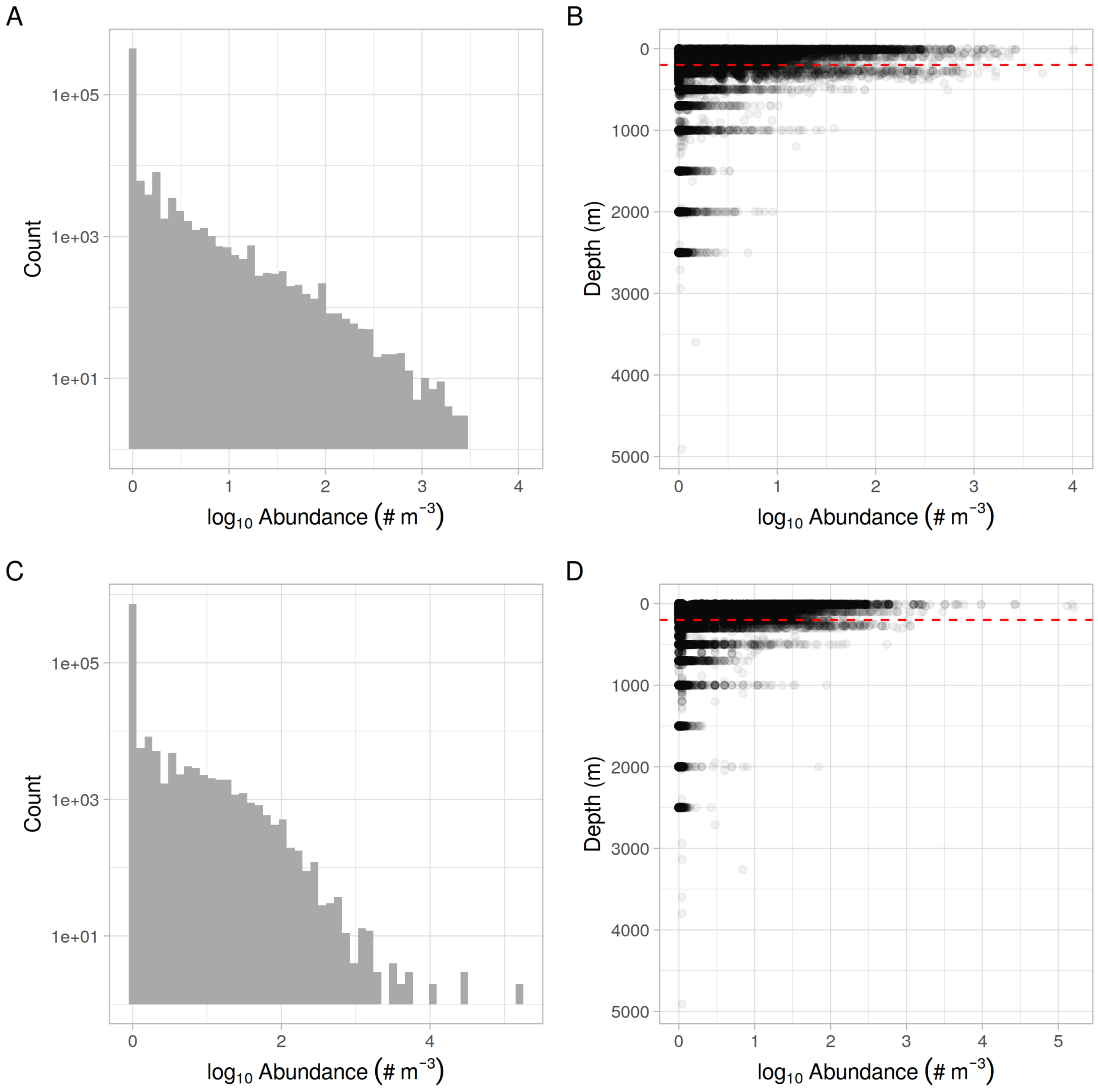


Figure S3. **A, C**: histogram of abundance observations for pteropods (**A**) and planktic foraminifers (**C**). The prevalence of zero abundances is evident. **B, D**: depth distribution of the sampling data for pteropods (**B**) and foraminifers (**D**). The dashed red line indicates the cut-off of 200 m. All data above this depth were used for the modelling.

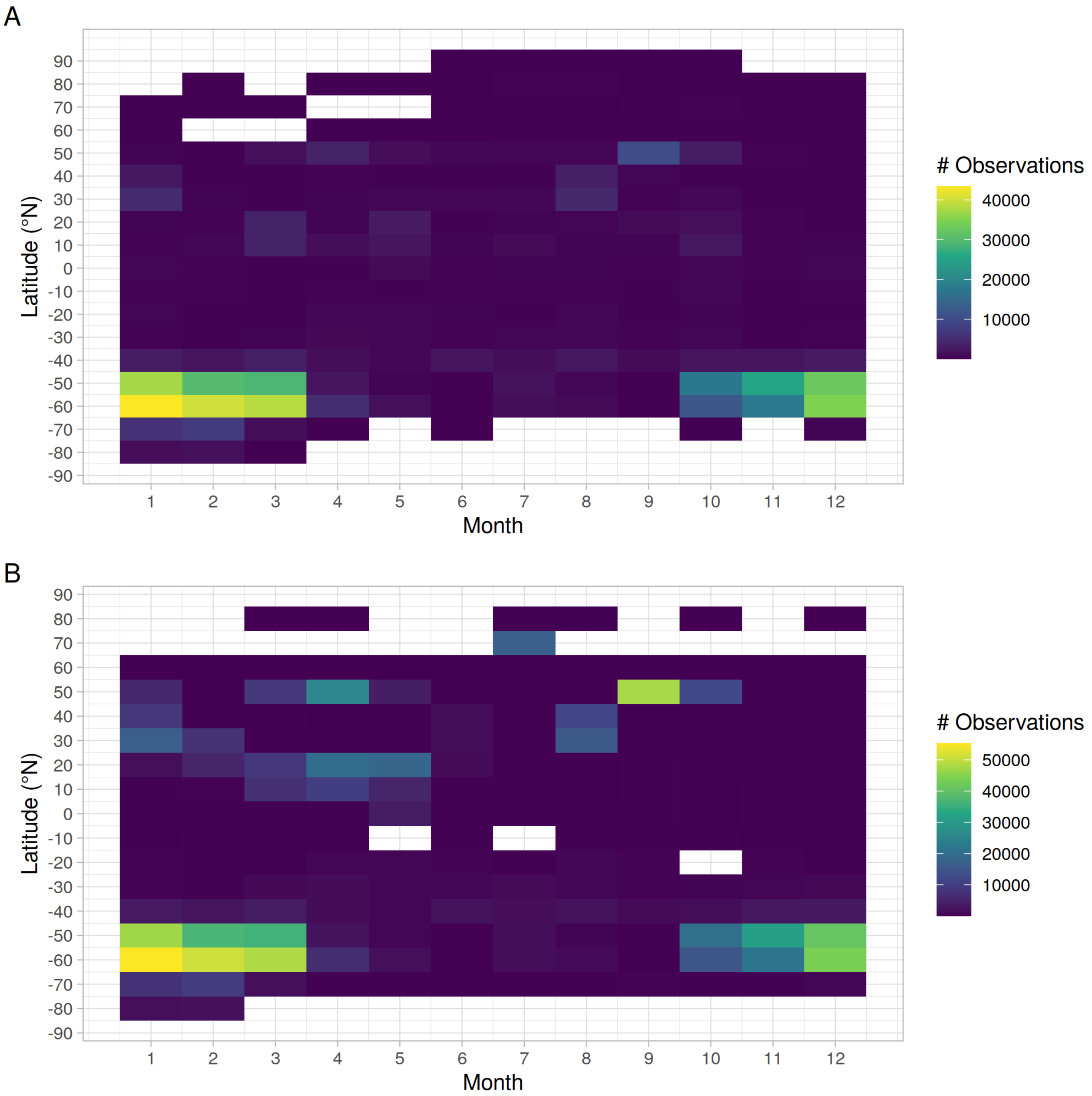


Figure S4. Hovmoeller diagrams showing the density of pteropod (**A**) and planktic foraminifer (**B**) sampling points as a function of month and latitude. The dominance of the Southern Ocean Continuous Plankton Recorder (SO-CPR) during the summer of the Southern Hemisphere as well as increased sampling effort in the Northern Hemispheric summer can be seen for both groups.

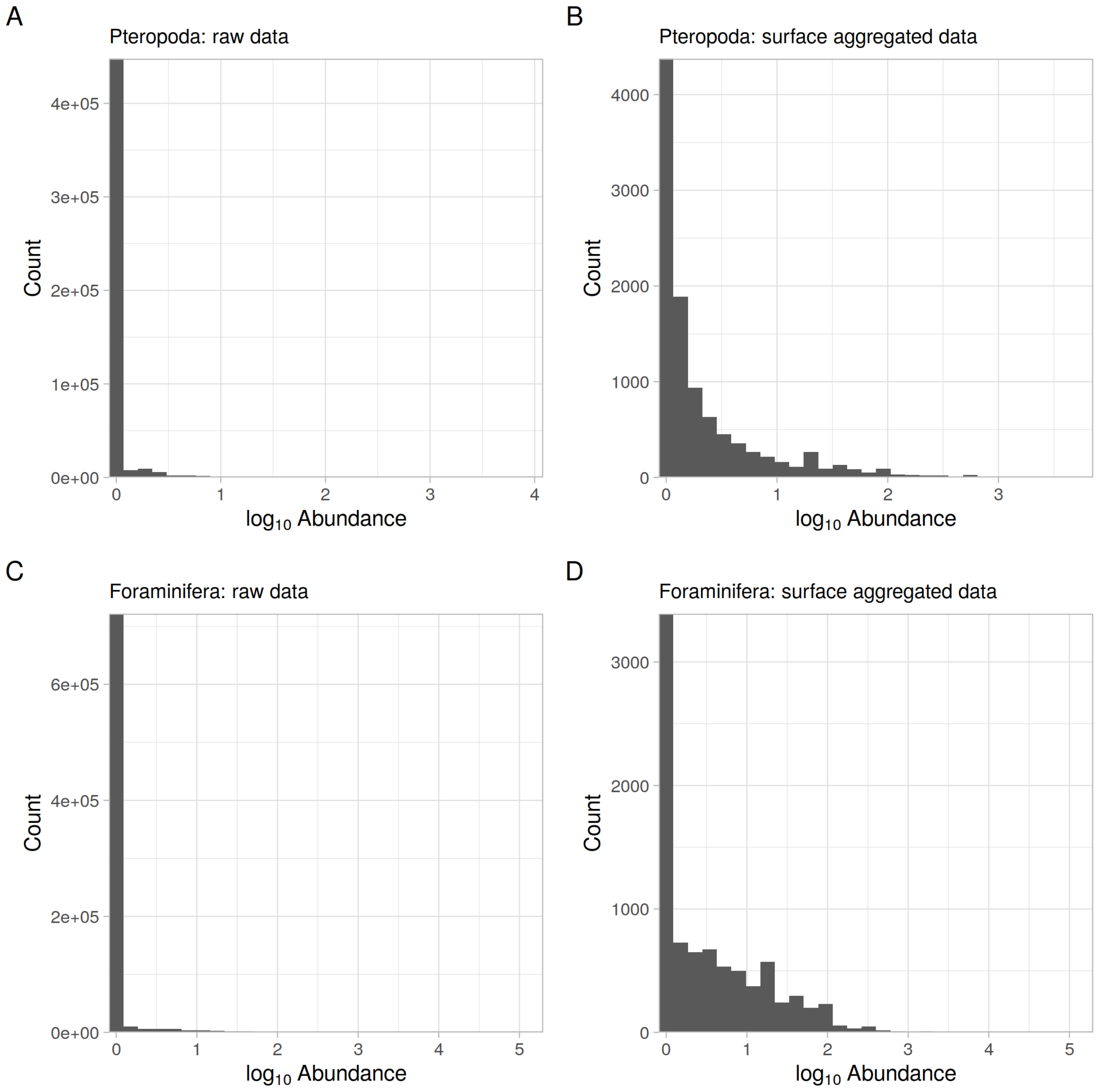


Figure S5. Effect of surface data aggregation on abundance data distribution, note the different axes limits. **A** and **C** show the distribution of raw observation data for pteropods and planktic foraminifers, respectively. Plots **B** and **D** show the histograms after the surface ocean aggregation. There is a notable reduction in points with zero abundance and the histograms are less skewed.

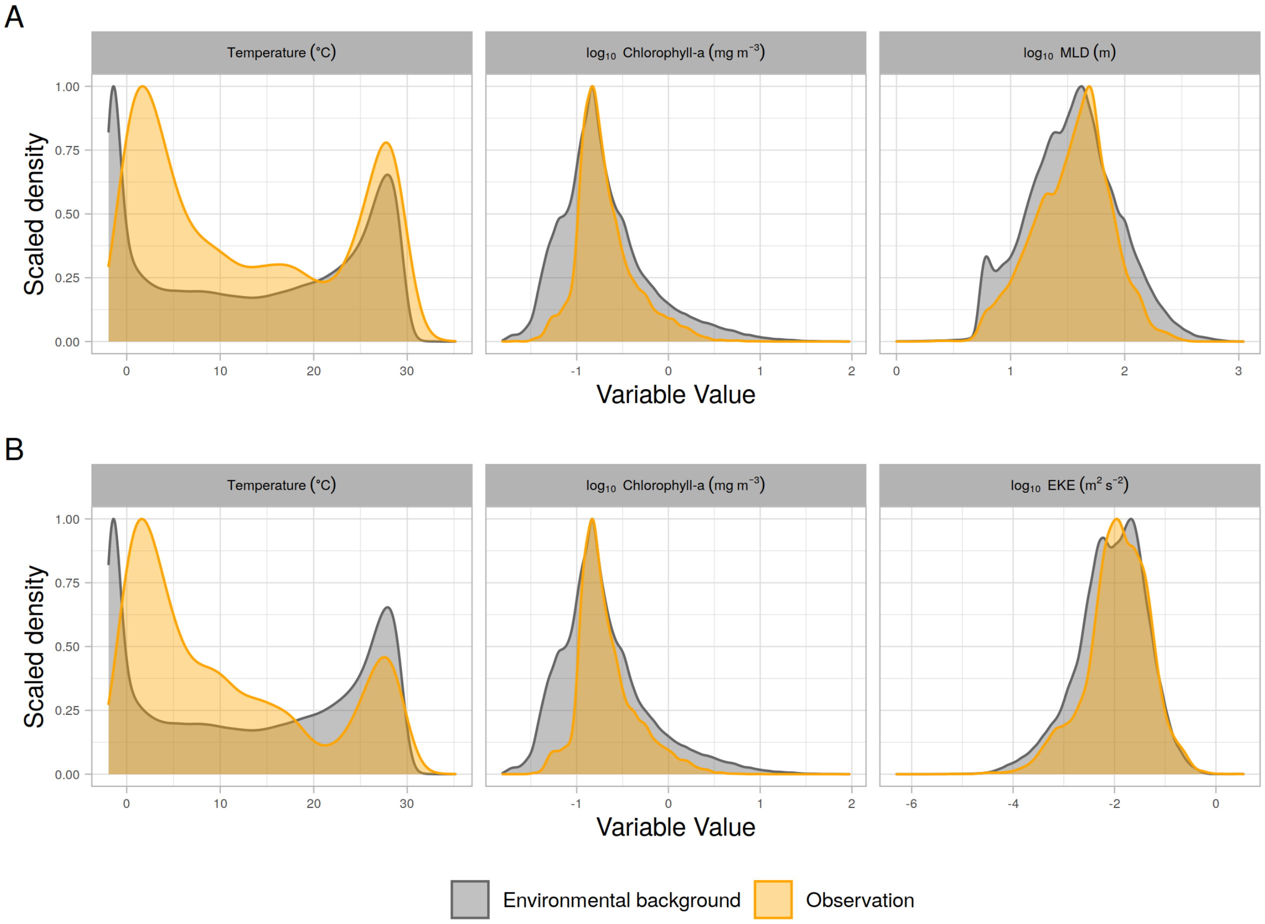


Figure S6. Coverage of the total global environmental background space by the observation data for **A** pteropods and **B** planktic foraminifers. Grey shading indicates the environmental background data and orange shading the environmental conditions at the spatio-temporal location of the sampling points after the surface ocean aggregation. The density curves are scaled to reach a maximum value of 1.

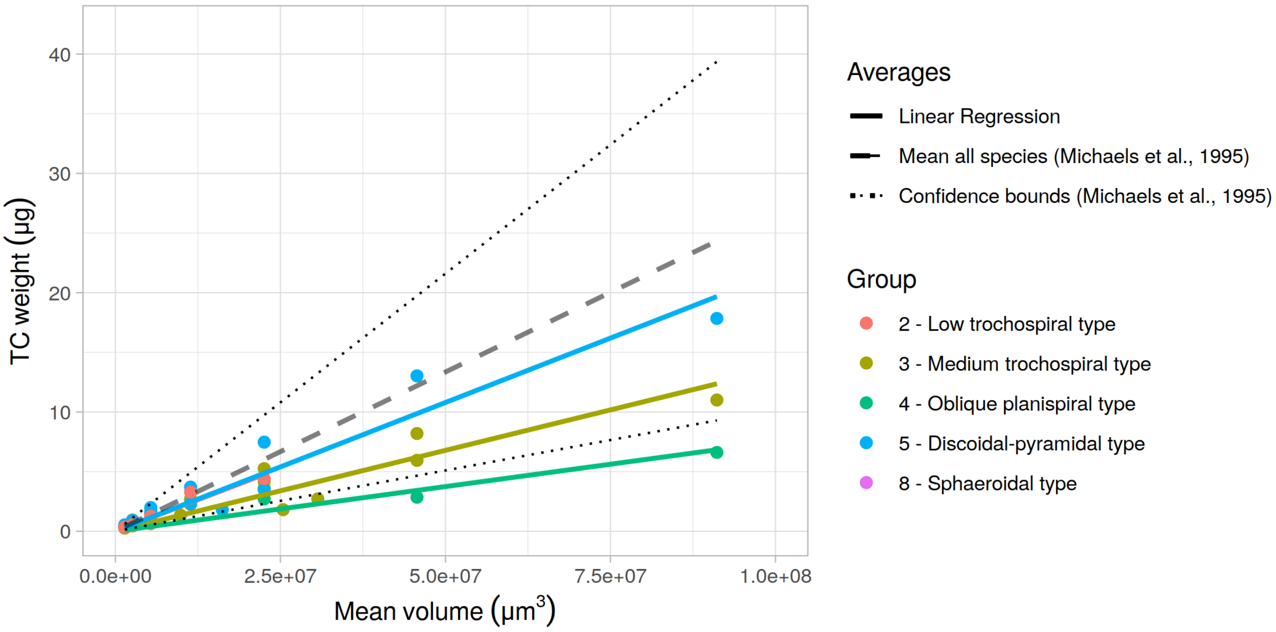


Figure S7. Linear fits of foraminifer total carbon (TC) weight as a function of mean volume based on sampling data from Schiebel and Hemleben (2000) and Takahashi and Bé (1984). The colors indicate the different shape groups. The dashed line denotes the mean value as calculated per Michaels et al. (1995) and the dotted lines the corresponding confidence interval.

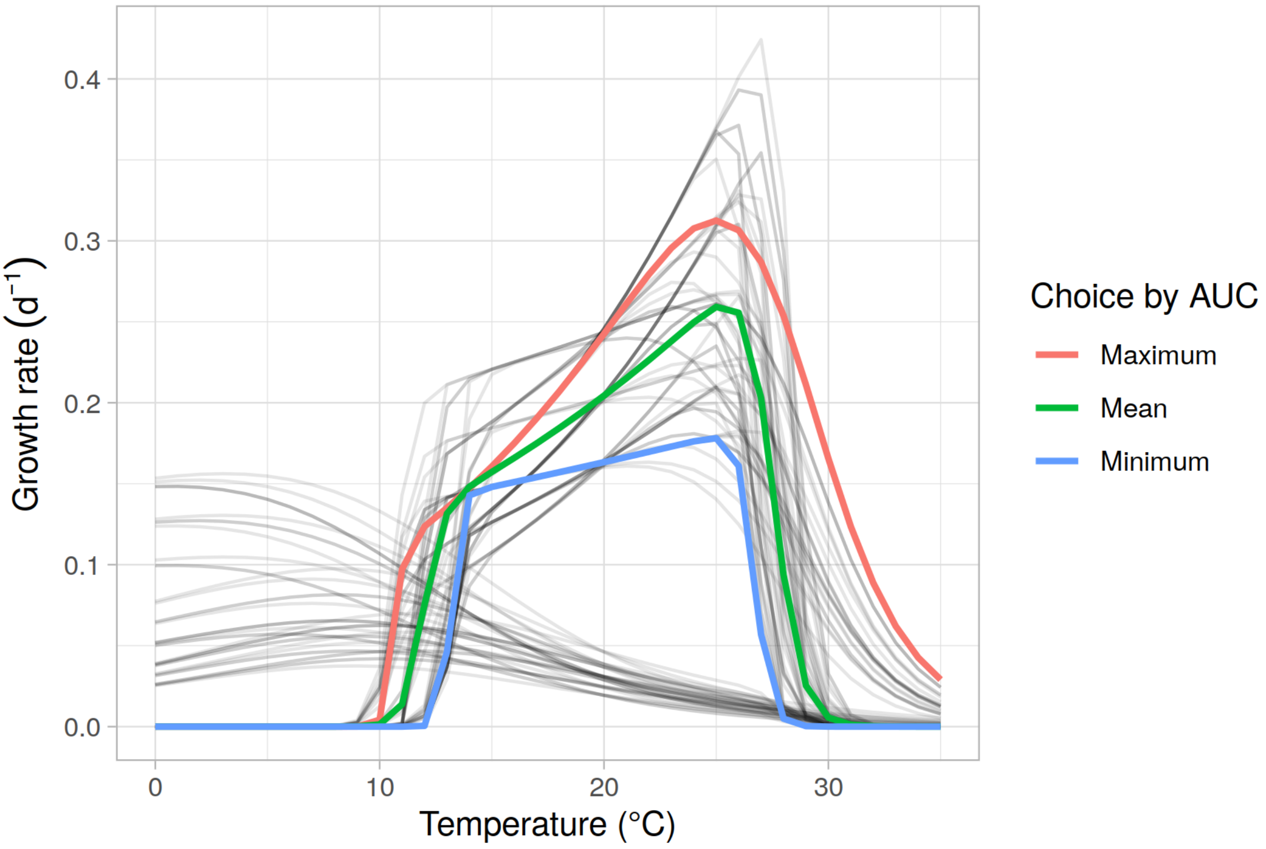


Figure S8. Foraminifer daily growth rates as derived from Lombard et al. (2009). Black lines indicate all possible curves from the range of parameter values given. Colored lines indicate the final choices for the modelling. Minimum and maximum curves were chosen based on the minimal (maximal) area under the curve (AUC) between 0°C and 30°C while retaining ecologically sensible shapes. This means the curves with a growth rate maximum between 0°C and 10°C were not chosen despite their lower AUC as they are deemed non-representative of the entire foraminifera phylum.

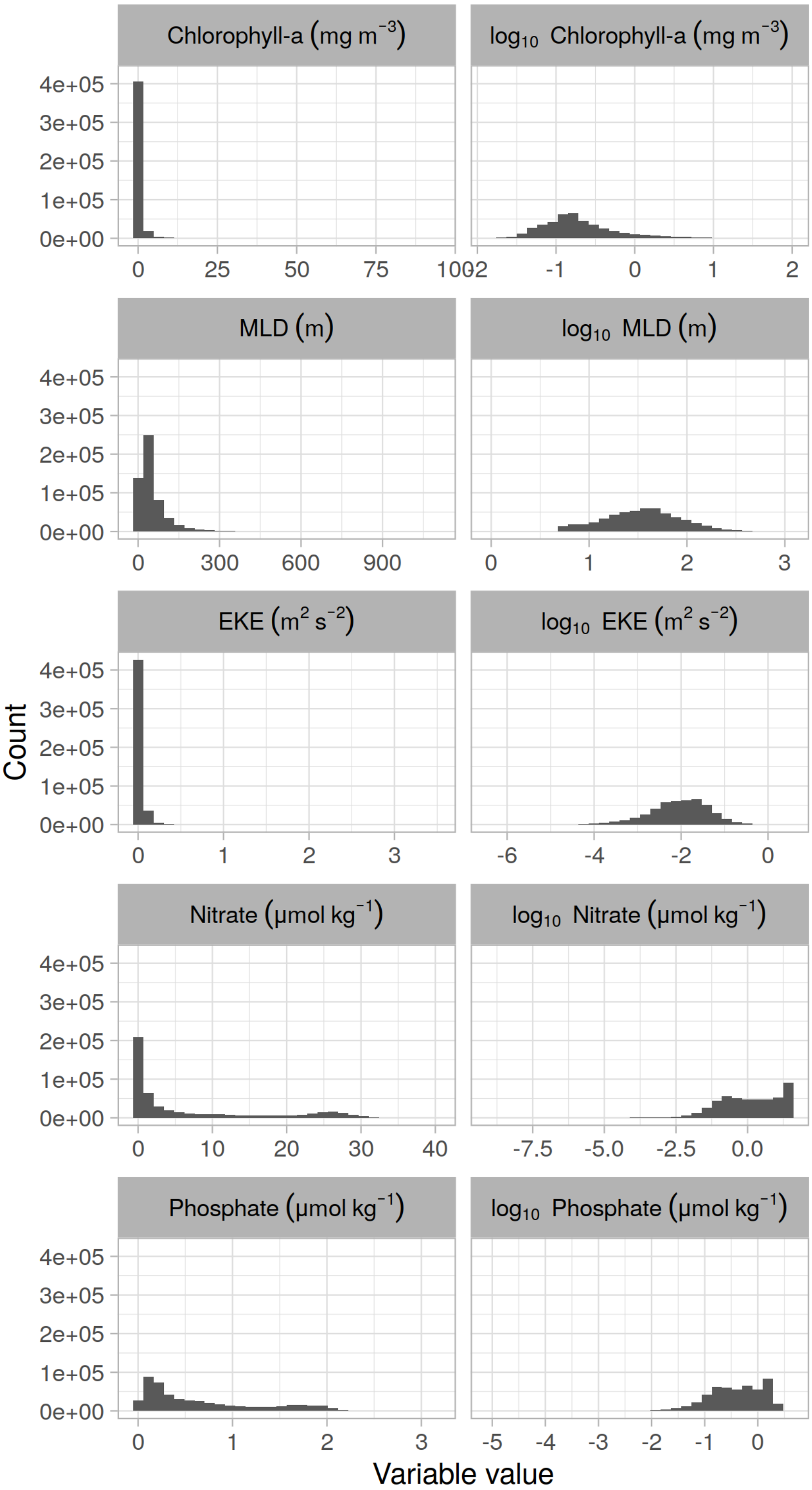


Figure S9. Histograms depicting the global distribution of values for the environmental predictors that were later log-transformed. The left column shows the histograms for the original values and the right column those for the log-transformed ones. One can see that the transformation causes all variables to be more normally distributed than originally.

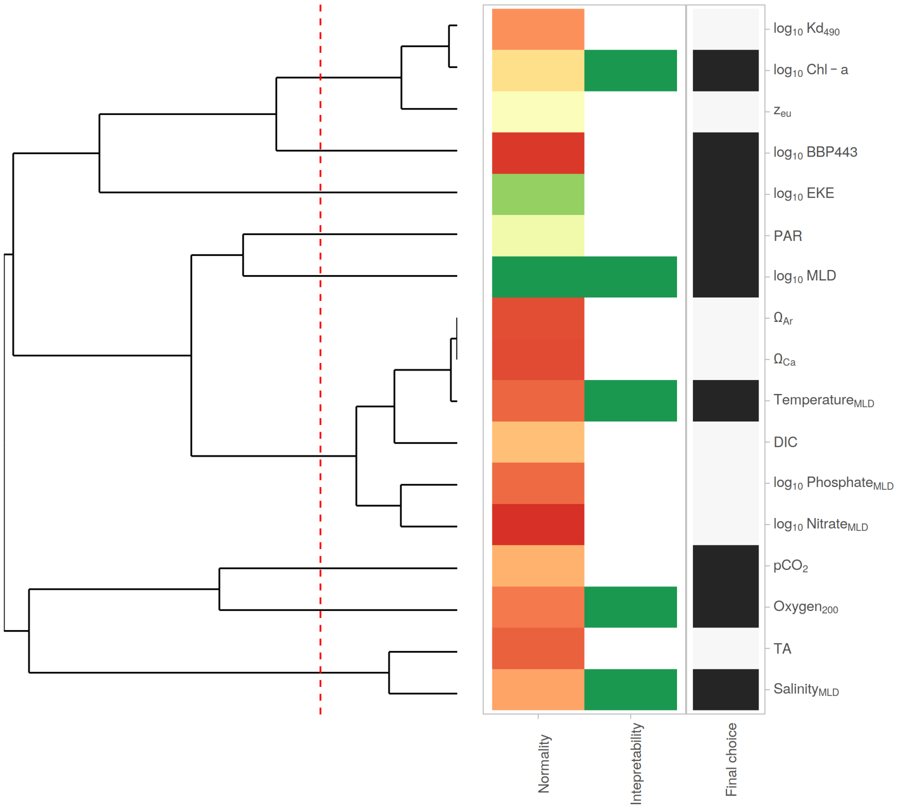


Figure S10. Plot depicting the steps taken to select the final set of environmental predictors for the pteropod species distribution models (SDMs). The dendrogram on the left shows the correlation structure of the environmental predictors as assessed at the grid points where observation data are present. The red dashed line indicates a correlation level of $|r| = 0.7$, i.e. all clusters right of this line are correlated to a higher degree. From each cluster, only one environmental predictor can be chosen and the red-green tile plot in the middle shows an evaluation of the two selection criteria, with green indicating a positive choice and red a negative one. 1) More normally distributed predictors are preferred. The normality column in the tile plot is a measure of the normality of the distribution of each environmental predictor. The values shown are the log-transformed and subsequently normalized p-values of the Shapiro-Wilk test. 2) Predictors with clearer known relevance for zooplankton abundances and hence simpler interpretability are preferred. These choices were made manually, with green shading indicating the most easily interpretable predictor. Finally, the last, black-and-white column highlights the final chosen predictors which were in the next step assessed for their predictive power.

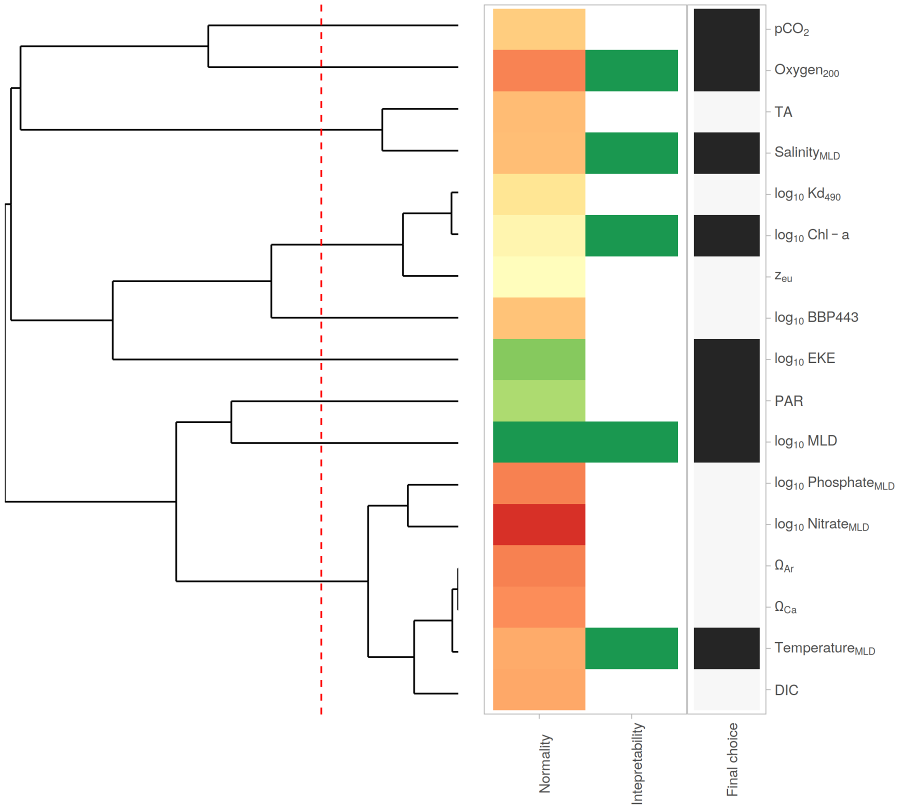
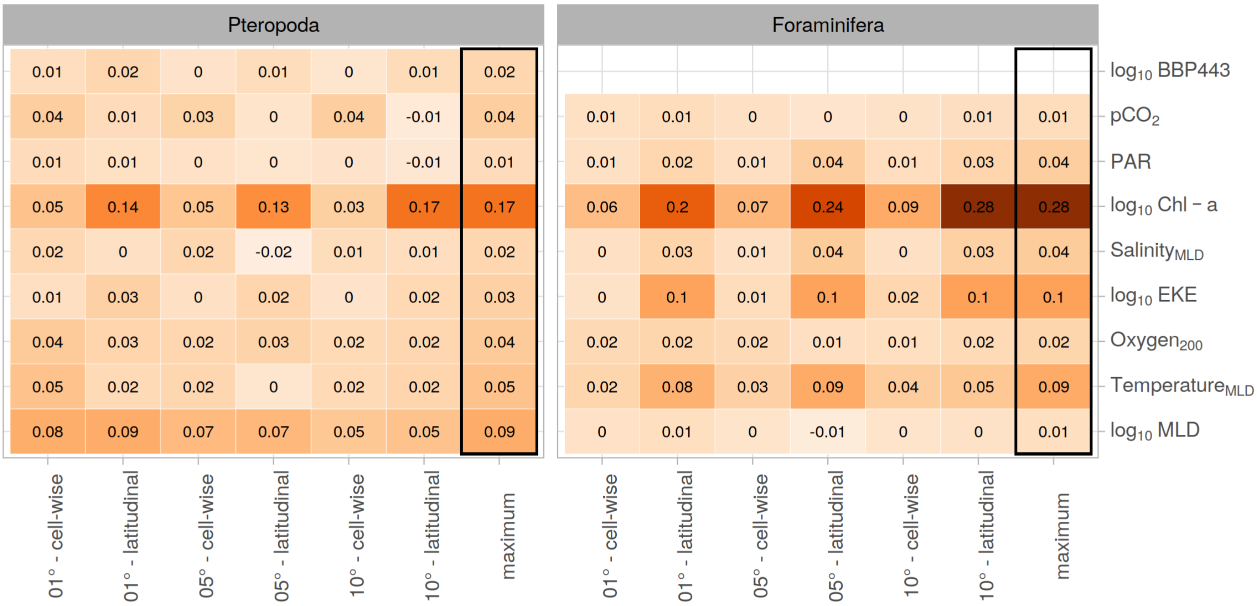


Figure S11. Plot depicting the steps taken to select the final set of environmental predictors for the foraminifer species distribution models (SDMs). See figure S10 for an extensive explanation of the plot structure.



Pteropoda

	01° - cell-wise	01° - latitudinal	05° - cell-wise	05° - latitudinal	10° - cell-wise	10° - latitudinal	maximum
$\log_{10}$ BBP443	0.01	0.02	0	0.01	0	0.01	0.02
pCO_2	0.04	0.01	0.03	0	0.04	-0.01	0.04
PAR	0.01	0.01	0	0	0	-0.01	0.01
$\log_{10}$ Chl − a	0.05	0.14	0.05	0.13	0.03	0.17	0.17
Salinity_{MLD}	0.02	0	0.02	-0.02	0.01	0.01	0.02
$\log_{10}$ EKE	0.01	0.03	0	0.02	0	0.02	0.03
Oxygen_{200}	0.04	0.03	0.02	0.03	0.02	0.02	0.04
Temperature_{MLD}	0.05	0.02	0.02	0	0.02	0.02	0.05
$\log_{10}$ MLD	0.08	0.09	0.07	0.07	0.05	0.05	0.09

Foraminifera

	01° - cell-wise	01° - latitudinal	05° - cell-wise	05° - latitudinal	10° - cell-wise	10° - latitudinal	maximum
$\log_{10}$ BBP443							
pCO_2	0.01	0.01	0	0	0	0.01	0.01
PAR	0.01	0.02	0.01	0.04	0.01	0.03	0.04
$\log_{10}$ Chl − a	0.06	0.2	0.07	0.24	0.09	0.28	0.28
Salinity_{MLD}	0	0.03	0.01	0.04	0	0.03	0.04
$\log_{10}$ EKE	0	0.1	0.01	0.1	0.02	0.1	0.1
Oxygen_{200}	0.02	0.02	0.02	0.01	0.01	0.02	0.02
Temperature_{MLD}	0.02	0.08	0.03	0.09	0.04	0.05	0.09
$\log_{10}$ MLD	0	0.01	0	-0.01	0	0	0.01

Figure S12. Variance explained by the different environmental predictors as assessed by three univariate models (GLM, GLM with quadratic terms and GAM) across the grid-wise and latitudinal aggregation levels for pteropods and foraminifers. The last column of both plots shows the maximum deviance explained across any of the assessed spatial aggregation levels. These are the values used for deciding which predictors to include in the species distribution models. The subscript MLD refers to variables that were averaged over the mixed layer depth. The value of oxygen was taken at 200 m depth.

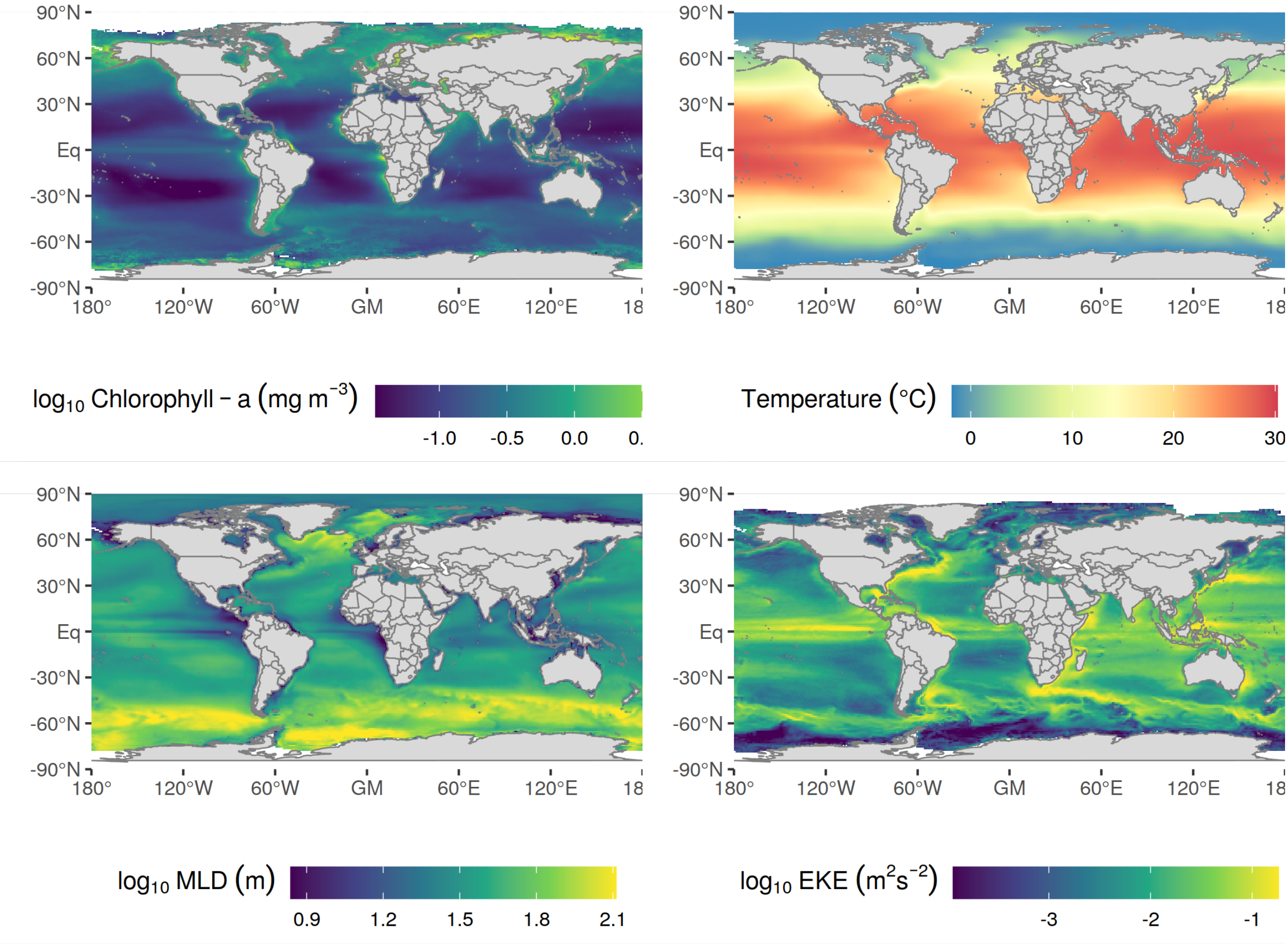


Figure S13. Annually averaged distribution of the four environmental predictors used in the modelling process.

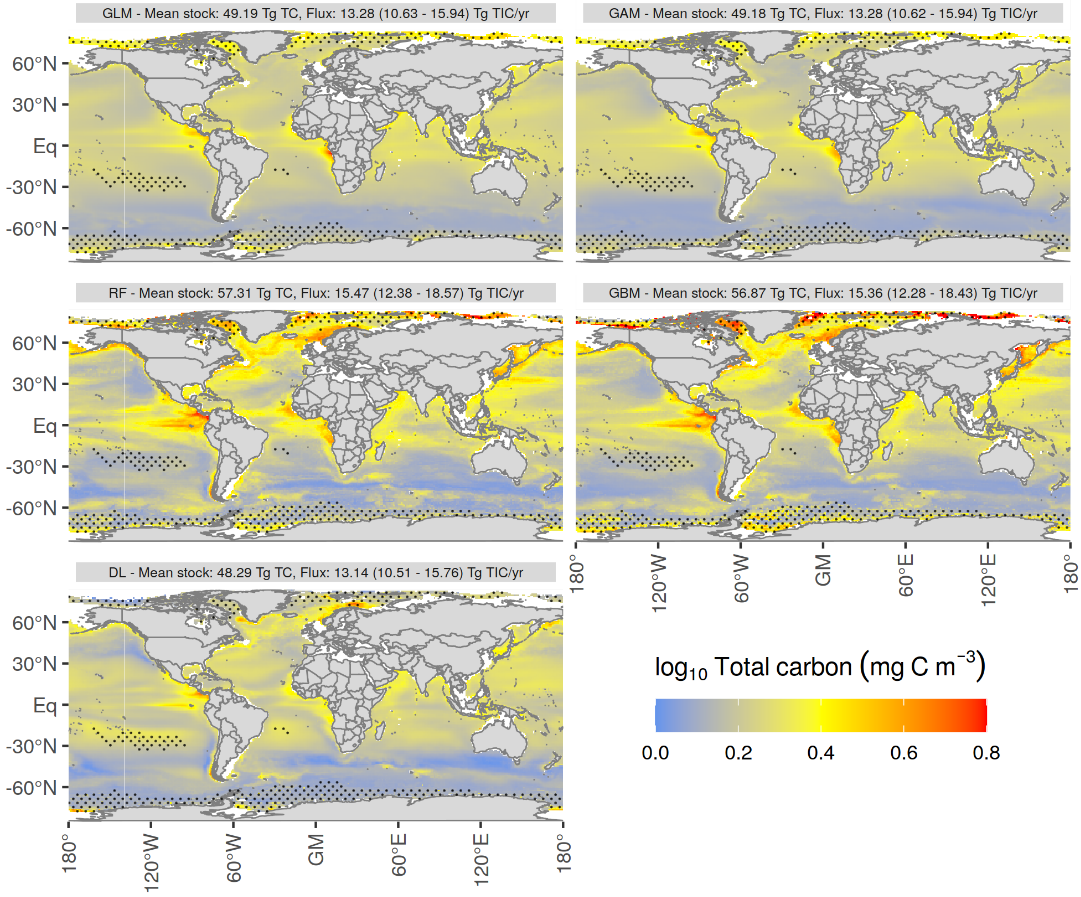

GLM - Mean stock: 49.19 Tg TC, Flux: 13.28 (10.63 - 15.94) Tg TIC/yr
GAM - Mean stock: 49.18 Tg TC, Flux: 13.28 (10.62 - 15.94) Tg TIC/yr
RF - Mean stock: 57.31 Tg TC, Flux: 15.47 (12.38 - 18.57) Tg TIC/yr
GBM - Mean stock: 56.87 Tg TC, Flux: 15.36 (12.28 - 18.43) Tg TIC/yr
DL - Mean stock: 48.29 Tg TC, Flux: 13.14 (10.51 - 15.76) Tg TIC/yr
60°N
30°N
Eq
-30°N
-60°N
180°
120°W
60°W
GM
60°E
120°E
180°
log10 Total carbon (mg C m-3)
0.0
0.2
0.4
0.6
0.8

Figure S14. Mean annual pteropod total carbon (TC) biomass predictions as calculated by the five different models. Values are shown as $\log_{10}(TC + 1)$. Stippled areas indicate grid points where the environmental conditions were outside the training dataset for more than six months of the year as calculated with the Multivariate Environmental Similarity Surfaces (MESS) analysis. The headers denote the mean TC biomass stock and the annual global total inorganic carbon (TIC) flux with the range of uncertainty resulting from different choices of the TIC-TC conversion factor and the growth rate formulation.

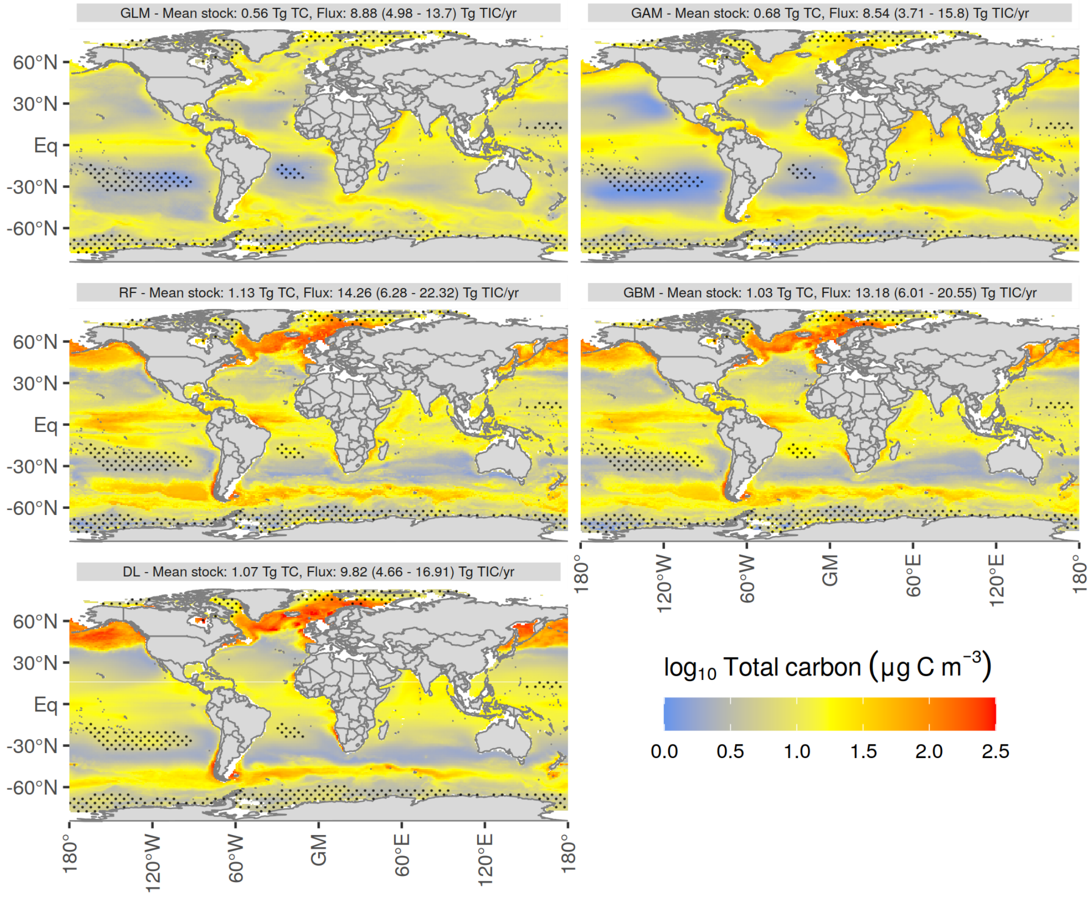

GLM - Mean stock: 0.56 Tg TC, Flux: 8.88 (4.98 - 13.7) Tg TIC/yr
GAM - Mean stock: 0.68 Tg TC, Flux: 8.54 (3.71 - 15.8) Tg TIC/yr
RF - Mean stock: 1.13 Tg TC, Flux: 14.26 (6.28 - 22.32) Tg TIC/yr
GBM - Mean stock: 1.03 Tg TC, Flux: 13.18 (6.01 - 20.55) Tg TIC/yr
DL - Mean stock: 1.07 Tg TC, Flux: 9.82 (4.66 - 16.91) Tg TIC/yr
60°N
30°N
Eq
-30°N
-60°N
180°
120°W
60°W
GM
60°E
120°E
180°
log10 Total carbon (µg C m-3)
0.0
0.5
1.0
1.5
2.0
2.5

Figure S15. Mean annual foraminifer total carbon (TC) biomass predictions as calculated by the five different models. Values are shown as $\log_{10}(TC + 1)$. Stippled areas indicate grid points where the environmental conditions were outside the training dataset for more than six months of the year as calculated with the Multivariate Environmental Similarity Surfaces (MESS) analysis. The headers denote the mean TC biomass stock and the annual global total inorganic carbon (TIC) flux with the range of uncertainty resulting from different choices of the TIC-TC conversion factor and the growth rate formulation.

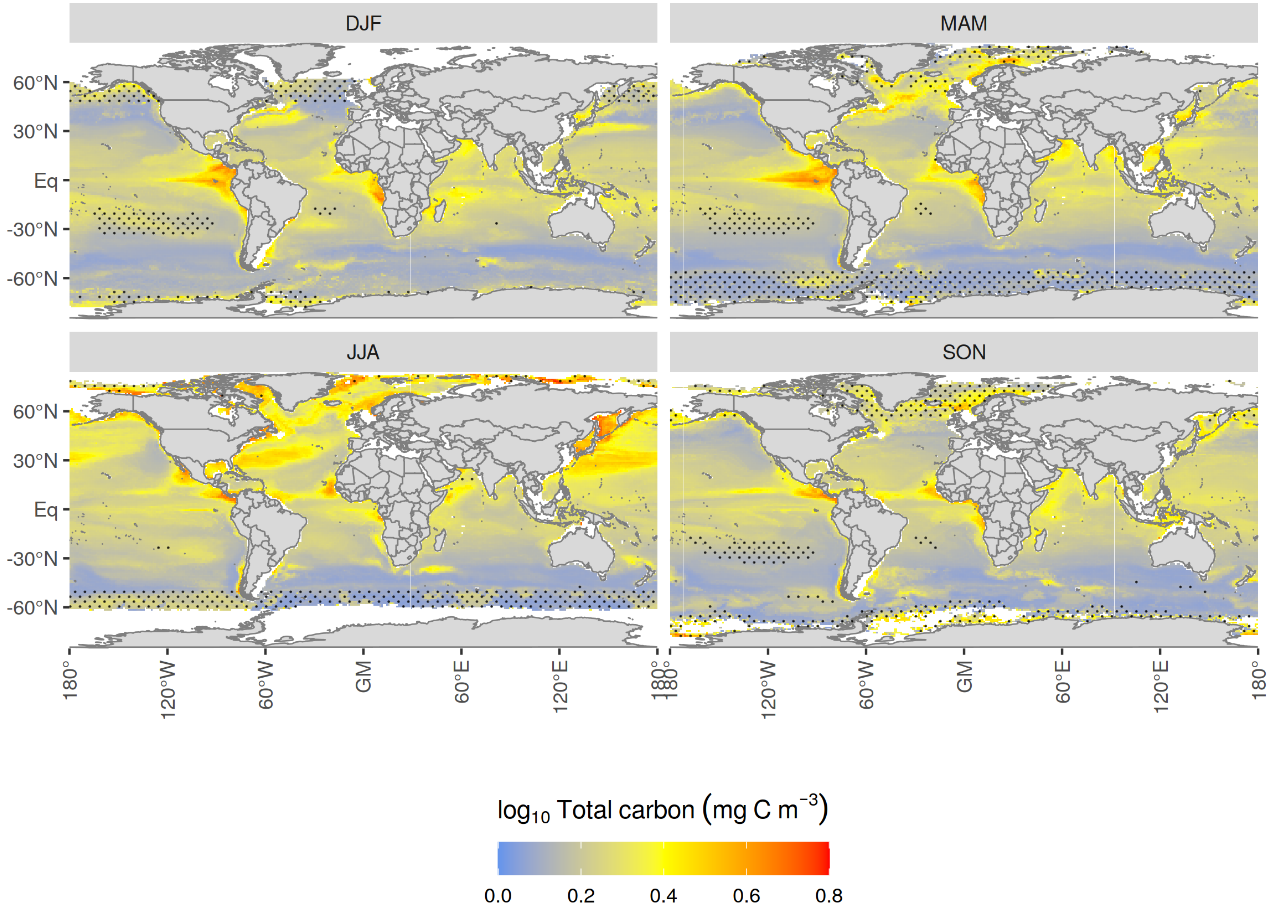


Figure S16. Seasonal mean pteropod total carbon (TC) biomass predictions as mean over the five models (DJF = December - February, MAM = March - May, JJA = June - August, SON = September - November). Values are shown as $\log_{10}(TC + 1)$. Stippled areas indicate grid points where the environmental conditions were outside the training dataset for more than one month of the respective season as calculated with the Multivariate Environmental Similarity Surfaces (MESS) analysis.

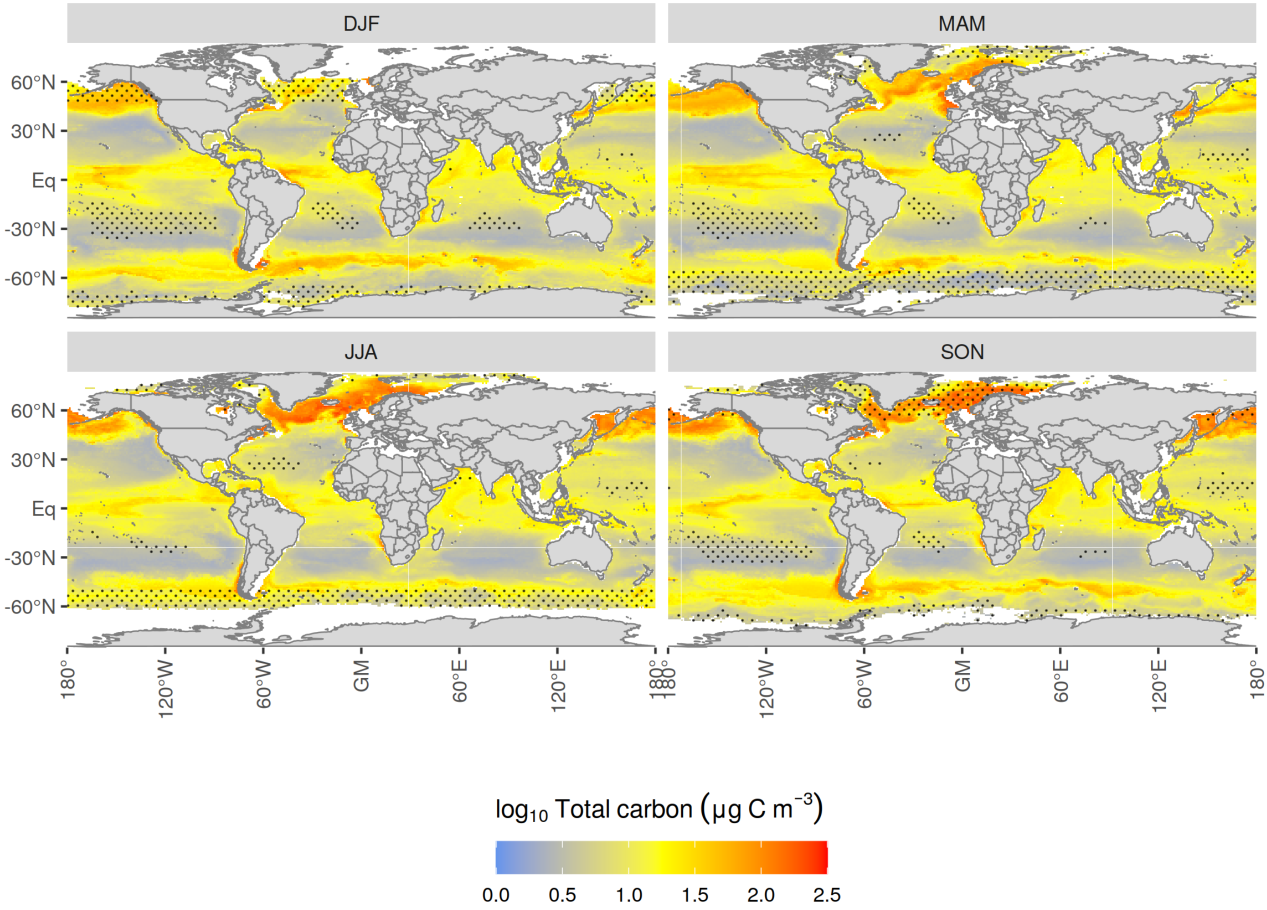


Figure S17. Seasonal mean foraminifer total carbon (TC) biomass predictions as mean over the five models (DJF = December - February, MAM = March - May, JJA = June - August, SON = September - November). Values are shown as $\log_{10}(TC + 1)$. Stippled areas indicate grid points where the environmental conditions were outside the training dataset for more than 1 months of the respective season as calculated with the Multivariate Environmental Similarity Surfaces (MESS) analysis.

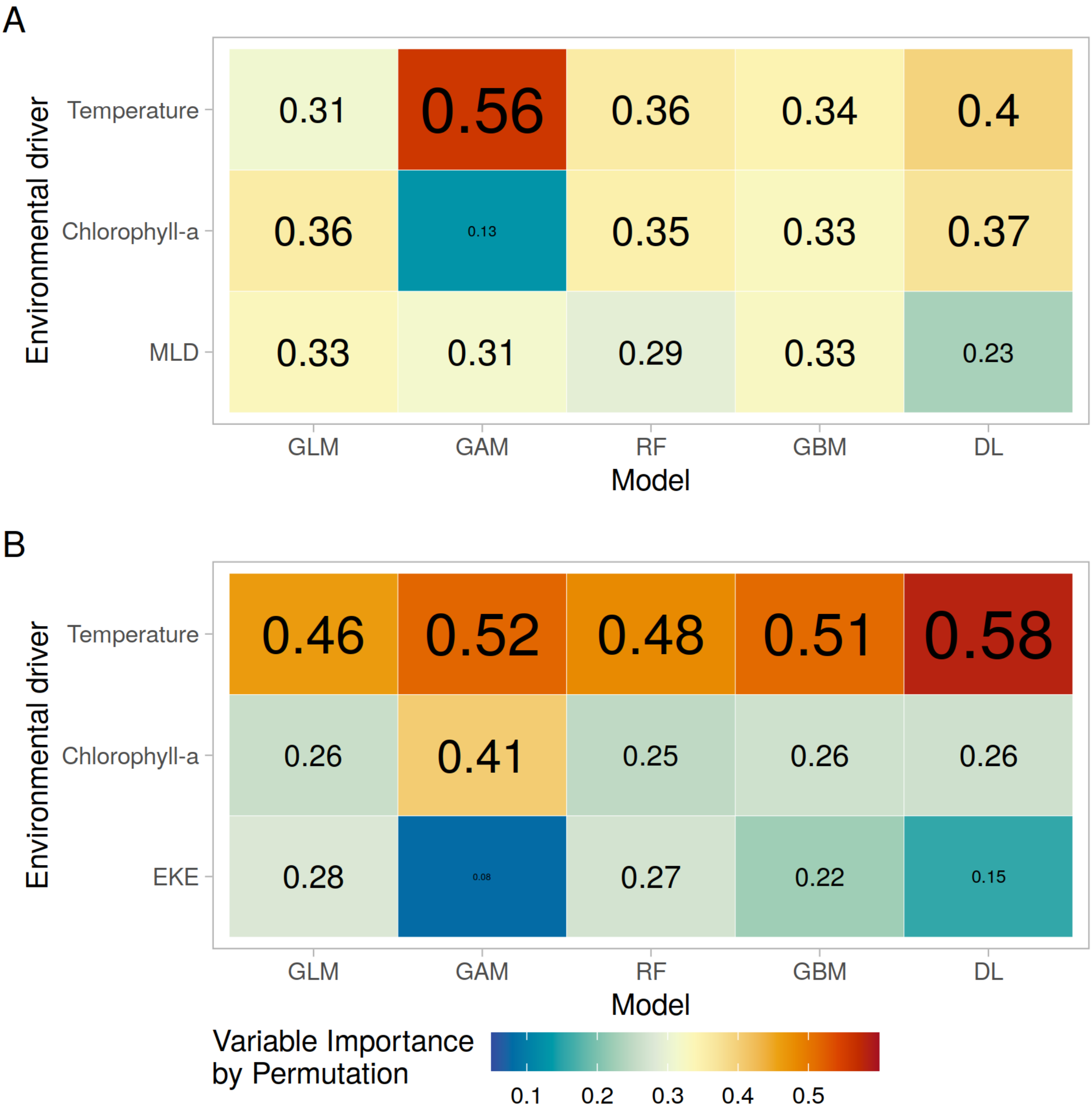


Figure S18. Normalized pteropod (**A**) and foraminifer (**B**) predictor variable importance as calculated with a permutation analysis across the five species distribution models (SDMs). A high value indicates that a change in this variable has a large effect on the predicted biomass values. All importance values are normalized to sum to one for each model.

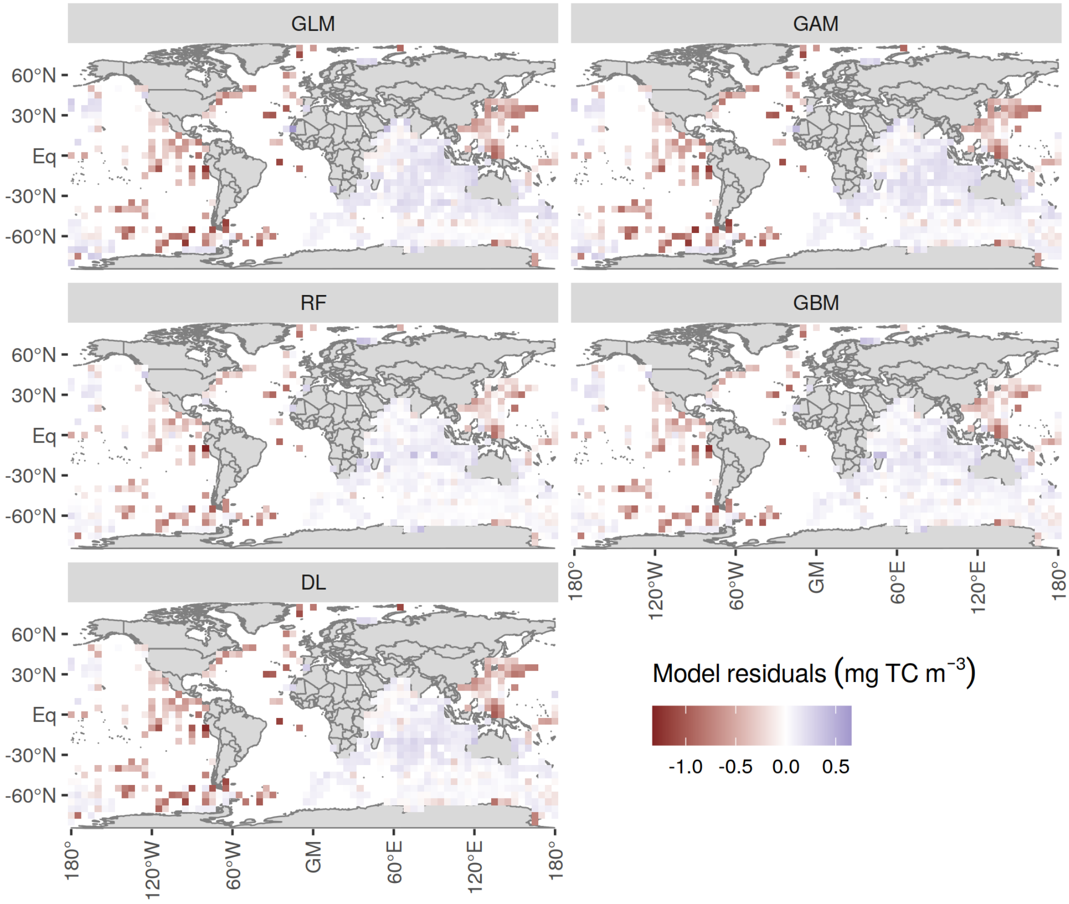


Figure S19. Pteropod total carbon (TC) biomass prediction residuals, averaged over all months and 5°grid bins. Negative residuals, i.e. an underestimation of the true values can be seen in the tropical ocean as well as the North Atlantic and the South-eastern Pacific. In contrast, an overestimation of the true values occurs mostly in the Indian Ocean and to a small extent in the Southern Ocean between 0°E and 150°E. These patterns correspond to the biomass predictions in that regions of high productivity are generally still underestimated, as the bloom dynamics here cause very high biomass concentrations. Areas of lower productivity are generally slightly overestimated.

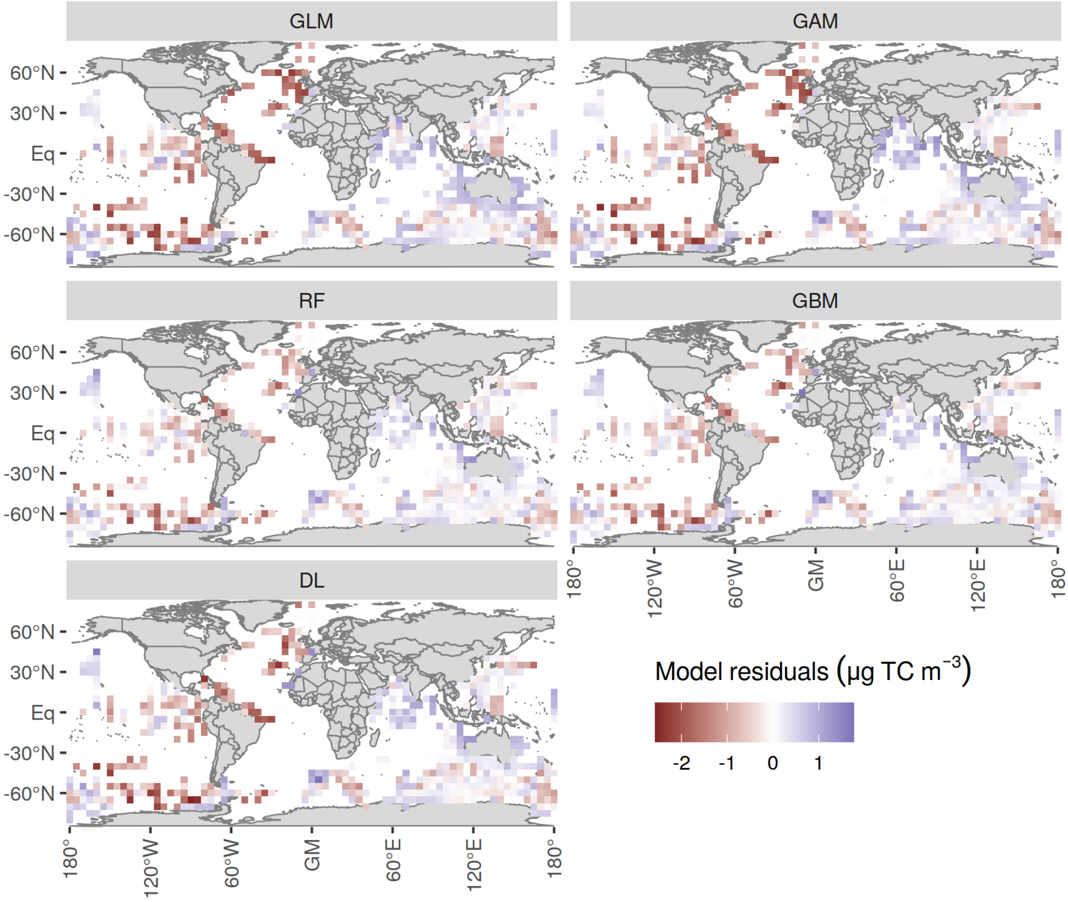


Figure S20. Foraminifer total carbon (TC) biomass prediction residuals, averaged over all months and 5° grid bins. The Random Forest model (RF) performs overall best, with lowest residual values everywhere, followed by the Boosted Regression Tree (GBM). The Generalized Linear Model (GLM) and Generalized Additive Model (GAM) strongly underestimate biomass concentrations in the highly productive regions of the North Atlantic, the equatorial region and the Southern Ocean between 180°W and 60°W. This trend is seen to a lesser extent in the Neural Network (DL) as well. In the GLM, GAM and DL, a slight overestimation of the true biomass values can be seen in the Indian Ocean and around Australia.

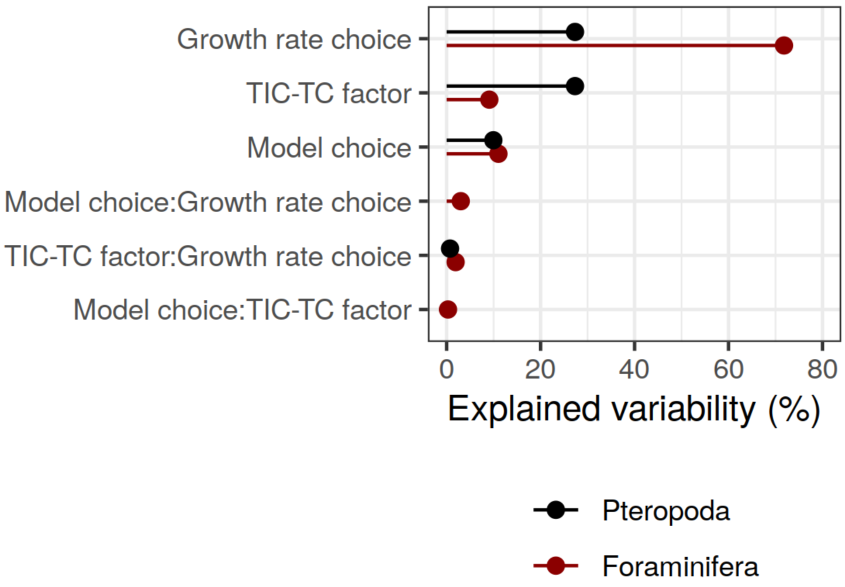


Figure S21. Percentage of variance in mean annual pteropod and foraminifer total inorganic carbon (TIC) export fluxes explained by different model setup choices as assessed with a multivariate Analysis of Variance (mANOVA).

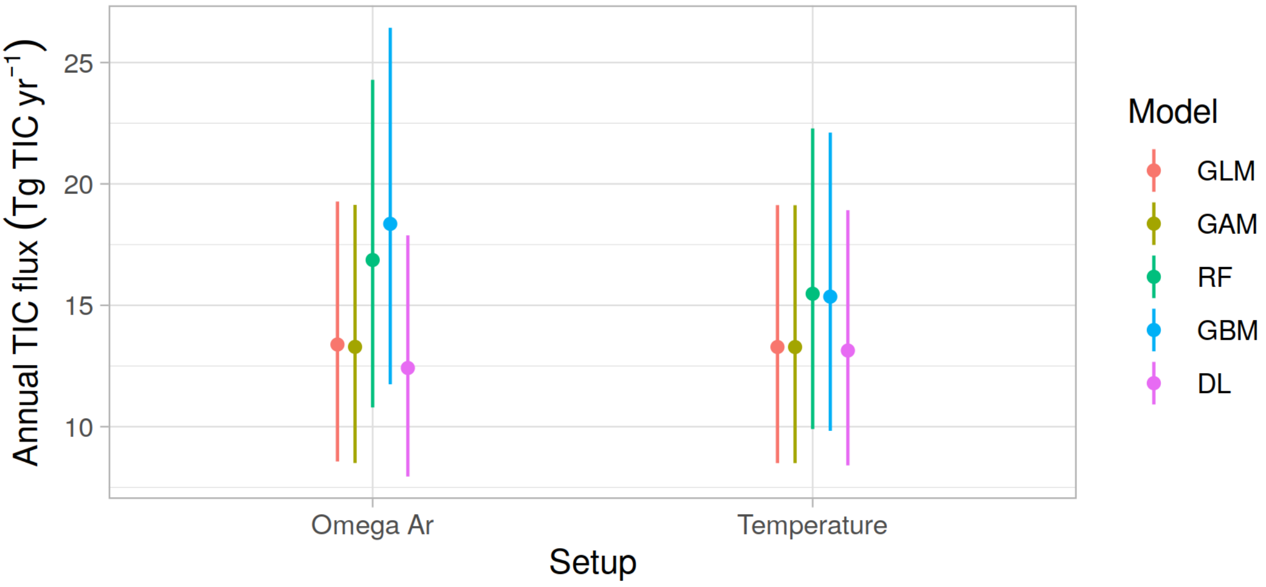


Figure S22. Global annual total inorganic carbon (TIC) fluxes for pteropods as calculated on the main predictor set including temperature averaged over the mixed layer and when replacing temperature by the aragonite saturation state (Ω_{Ar}) per SDM type. The range of values shown depicts the uncertainty range based on the TIC-TC conversion factor and the growth rate parametrization. For both plankton types, the difference in global annual TIC fluxes between the two setups is not statistically significant.

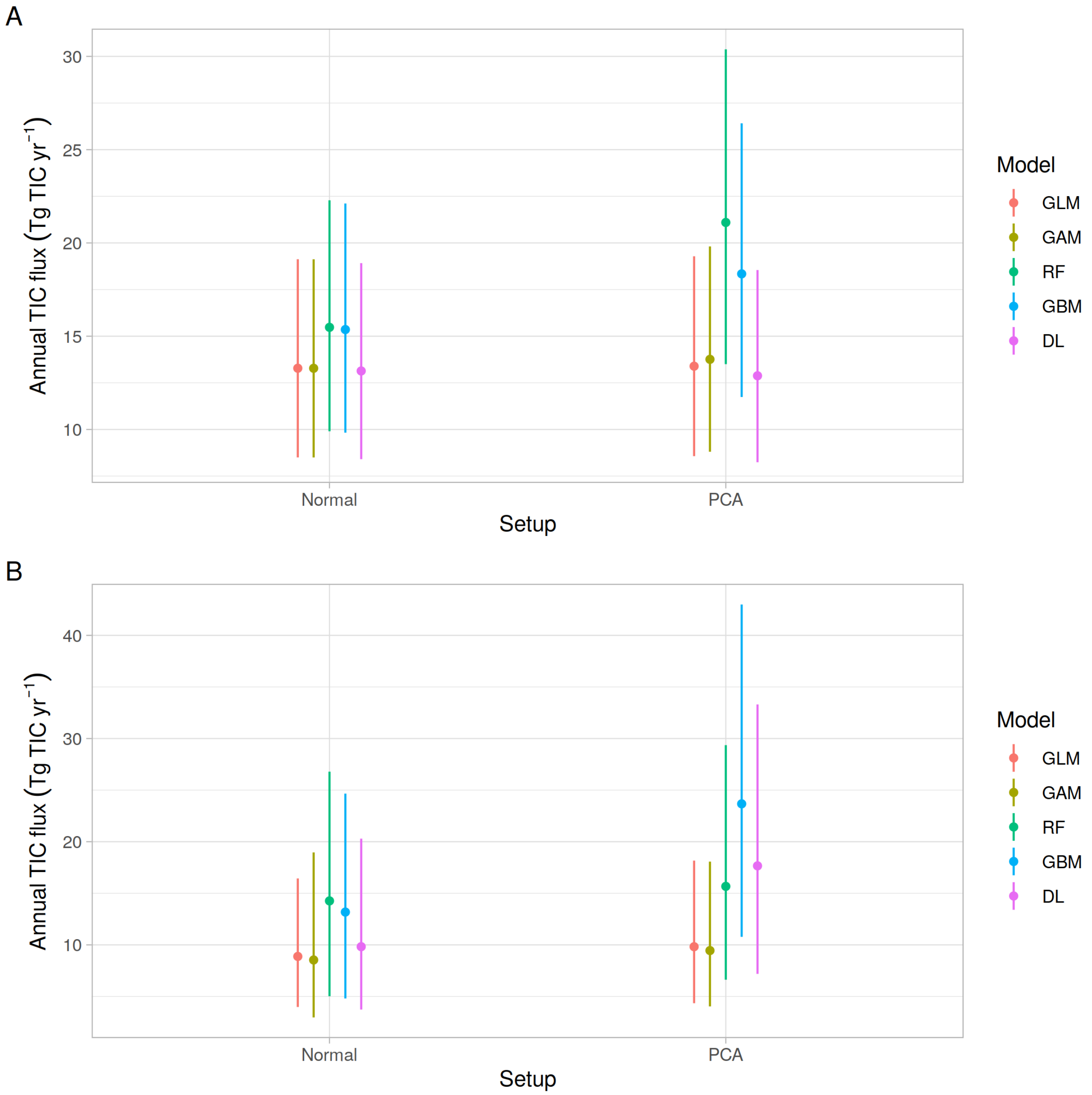


Figure S23. Global annual total inorganic carbon (TIC) fluxes for **A** pteropods and **B** foraminifers as calculated on the main predictor set and on a Principle Component Analysis (PCA) transformation of all environmental variables shown in table **??** per SDM type. The range of values shown depicts the uncertainty range based on the TIC-TC conversion factor and the growth rate parametrization. For both plankton types, the difference in global annual TIC fluxes between the regular setup and the PCA-setup is not statistically significant.

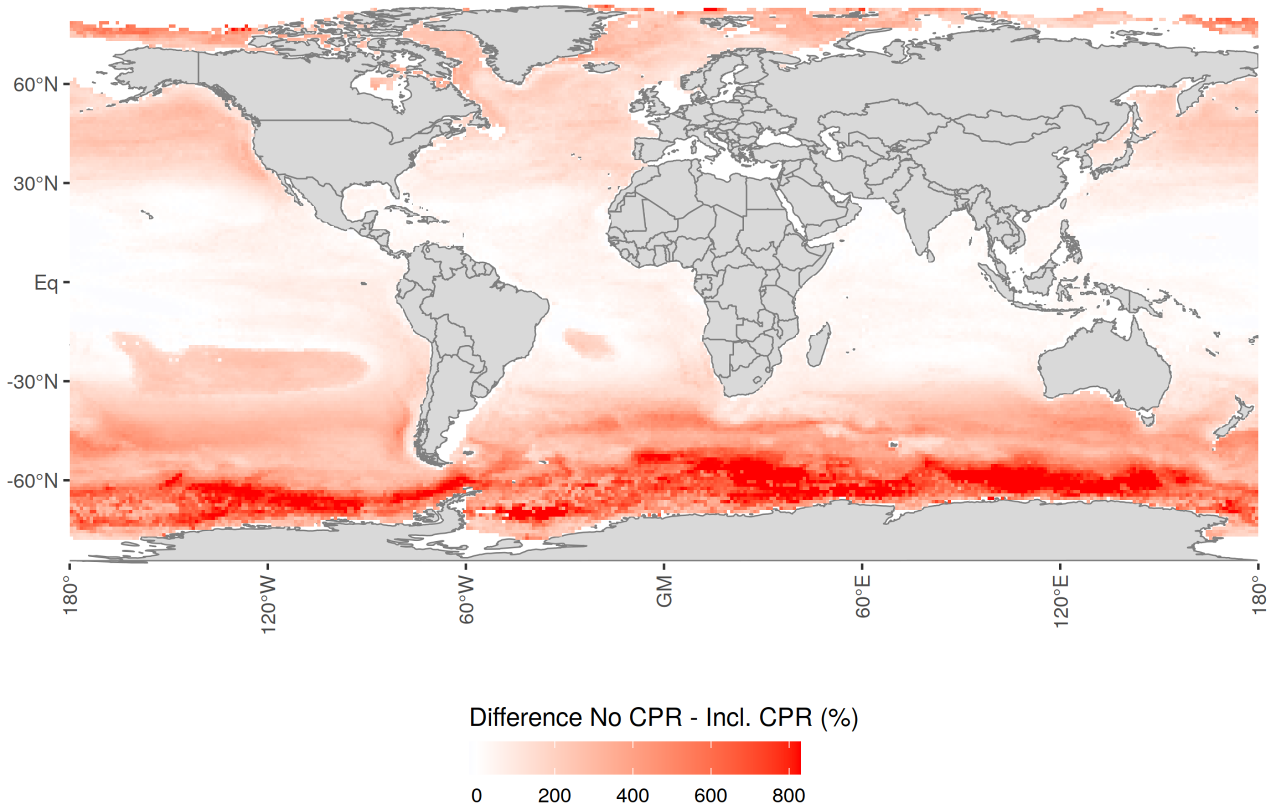


Figure S24. Relative change in pteropod biomass concentrations to baseline model when removing all CPR data.

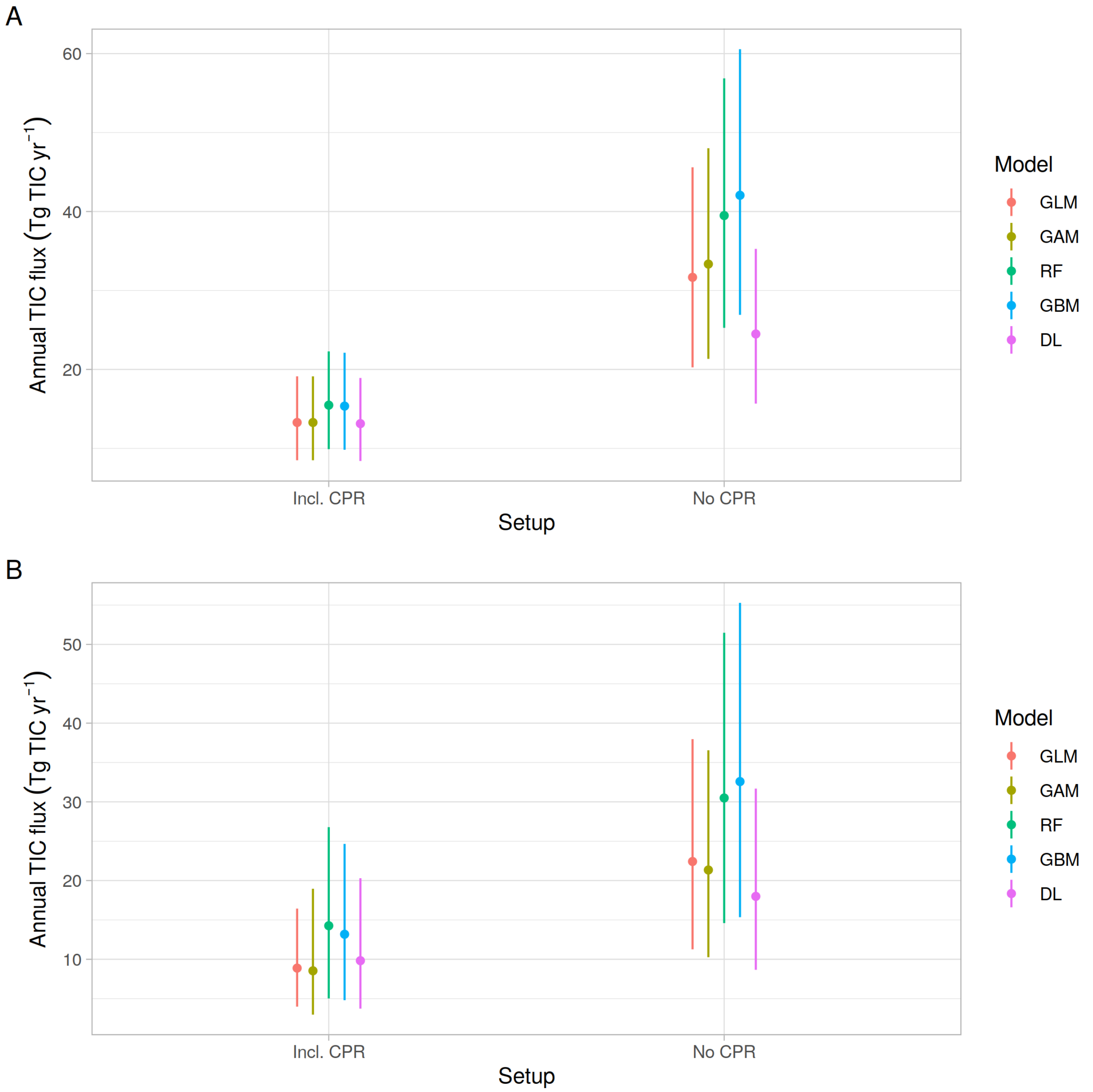


Figure S25. Global annual total inorganic carbon (TIC) fluxes for **A** pteropods and **B** foraminifers per SDM as calculated on the full dataset and only on non-CPR data, respectively. The range of values shown depicts the uncertainty range based on the TIC-TC conversion factor and the growth rate parametrization. For both plankton types, emitting all CPR values leads to a statistically significant increase in global annual TIC fluxes.

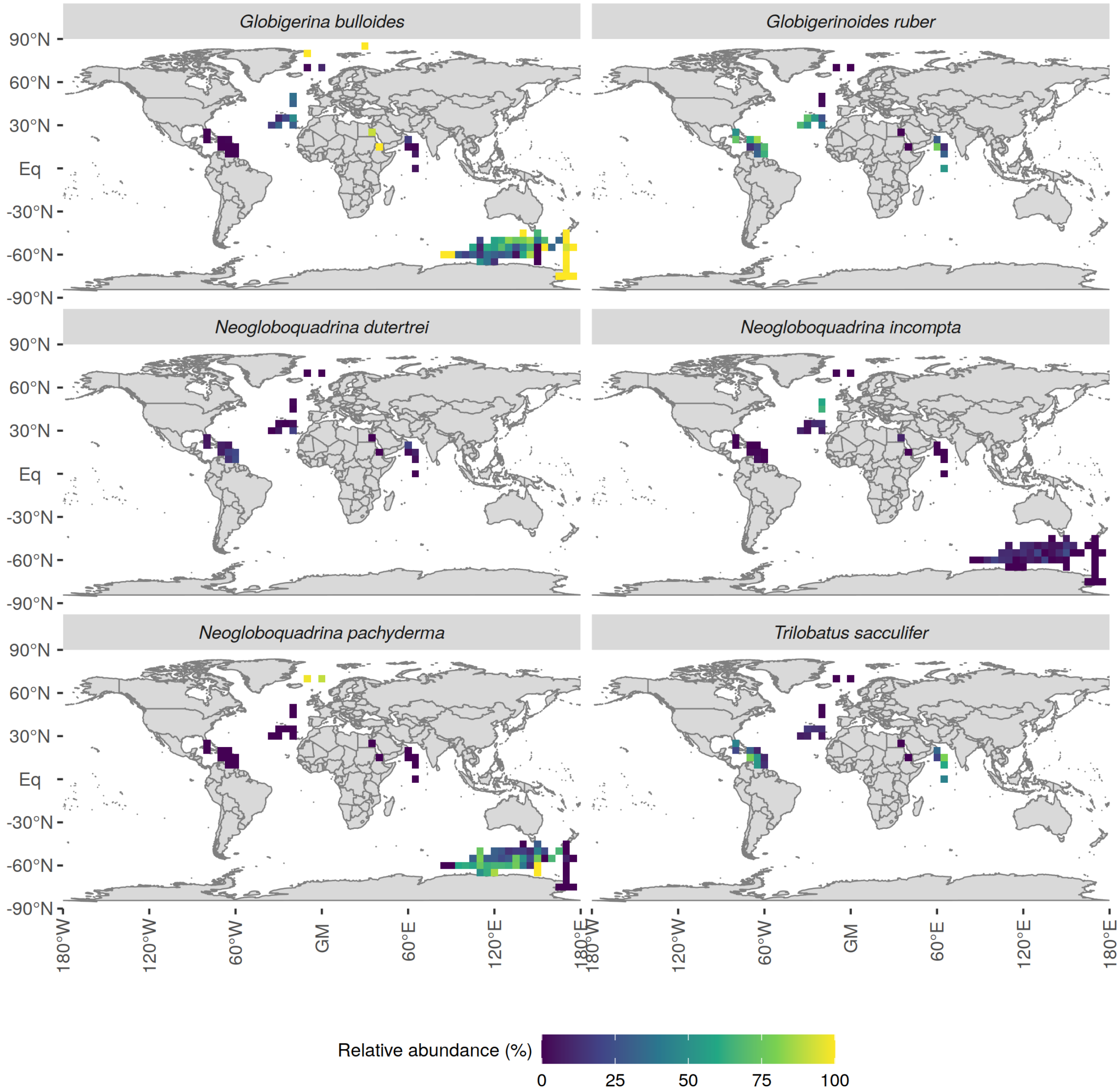


Figure S26. Relative species abundance in % of six common foraminifer species. The species-specific abundances were calculated by summing all unique counts of one species from a single tow and subsequently computing 5 × 5° gridded annual means. The relative abundance values were then calculated as a species-specific fraction of the sum over the six species' abundances. The patterns agree reasonably well with those found in Kretschmer et al. (2018) and Lombard et al. (2011) with the exception of edge cases in the Antarctic.

Tables

Table S1. Pteropod biomass conversion equations to compute wet weight (WW) or dry weight (DW) in mg based on an organisms length L or diameter D (collection adapted from Bednaršek et al. (2012)). Equations are from [1] Bednaršek et al. (2012), [2] Little and Copley (2003) and [3] Davis and Wiebe (1985).

Species	Group	Source	Equation
Limacina helicina	Round/cylindrical/globular	[1]	$DW = 0.137 * D^{1.5005}$
Limacina spp.	Round/cylindrical/globular	[1]	$WW = 10^{(2.533*\log_{10}(L)-3.89095)} * 10^5$
Clione spp.	Barell/oval-shaped (naked)	[2]	$WW = \pi * L^{(3*3/25)}$
Hyalocylis spp.	Cone/needle/tube/bottle-shaped	[2]	$WW = \pi * L^{(3*3/25)}$
Styliola spp.	Cone/needle/tube/bottle-shaped	[2]	$WW = 10^{(2.533*\log_{10}(L)-3.89095)} * 10^5$
Spongiobranchaea spp.	Barell/oval-shaped (naked)	[2]	$WW = 10^{(2.533*\log_{10}(L)-3.89095)} * 10^5$
Pneumodermopsis spp.	Barell/oval-shaped (naked)	[2]	$WW = 10^{(2.533*\log_{10}(L)-3.89095)} * 10^5$
Paedocline spp.	Barell/oval-shaped (naked)	[2]	$WW = 10^{(2.533*\log_{10}(L)-3.89095)} * 10^5$
Cavolinia spp.	Triangular/pyramidal	[2]	$WW = 0.2152 * L^{2.293}$
Clio spp.	Triangular/pyramidal	[2]	$WW = 0.2152 * L^{2.293}$
Creseis spp.	Cone/needle/tube/bottle-shaped	[2]	$WW = \pi * L^{(3*3/25)}$
Cuvierina spp.	Cone/needle/tube/bottle-shaped	[2]	$WW = \pi * L^{(3*3/25)}$
Diacria spp.	Triangular/pyramidal	[2]	$WW = 0.2152 * L^{2.293}$
Euthecosomata	Shelled	[3]	$WW = 0.2152 * L^{2.293}$
Gymnosomata	Naked	[3]	$WW = 10^{(2.533*\log_{10}(L)-3.89095)} * 10^3$
Pteropoda	Shelled	[3]	$WW = 0.2152 * L^{2.293}$

Table S2: Pteropod average length values (mm; from Bednaršek et al. (2012)) for different taxa as used in the analysis. The third column indicates the number of data points corresponding to this taxon in the full quality controlled dataset, and the fourth one indicates the number of non-zero abundances. Where no length value was available in Bednaršek et al. (2012), the fifth column indicates the choices taken. Note that for Pseudothecosomata without a given length value, the average value for the entire pteropod taxon was used.

Taxon	Length (mm)	# Obs	#Obs (non-zero)	Comment for length value
Cavolinia gibbosa	6.2	62	2	Family value used
Cavolinia globulosa	6			
Cavolinia inflexa	7.7	247	50	Mean of subspecies
Cavolinia inflexa imitans	8			
Cavolinia inflexa inflexa	7			
Cavolinia inflexa labiata	8			
Cavolinia longirostris angulosa	3.9			
Cavolinia longirostris longirostris	6.2			
Cavolinia longirostris strangulata	4			
Cavolinia uncinata	6.3	62	3	Mean of subspecies
Cavolinia uncinata pulosatupsilla	6.1			
Cavolinia uncinata uncinata	6.5			
Cavolinia spp.	6.2	23849		
Clio convexa	8	3292	217	
Clio cuspidata	20	62	10	
Clio piatkowskii	13.5			

Table S2 continued from previous page

Taxon	Length (mm)	# Obs	#Obs (non-zero)	Comment for length value
Clio pyramidata	20	56077	645	
Clio pyramidata antarctica	17	31	3	
Clio pyramidata lanceolata	20			
Clio pyramidata martensi	17			
Clio pyramidata spp.	18.5			
Clio recurva	16.5	31	1	Family value used
Clio spp.	16.5	52136		
Clione limacina antarctica	40	51717		66
Clione limacina meridionalis	20			
Clione limacina larvae	0.3			
Clione limacina spp.	12	1589		
Clione spp.	14.57	51739		
Corolla	8.9	31	3	Pteropod value used
Creseis acicula acicula	33			
Creseis acicula clava	6			
Creseis acicula spp.	19.5	524		
Creseis clava	11.5	31	14	Family value used
Creseis conica	11.5	62	17	Family value used
Creseis spp.	11.5	11211		
Creseis virgula conica	7			
Creseis virgula constricta	3.5			
Creseis virgula spp.	5.5	557		
Creseis virgula virgula	6			

Table S2 continued from previous page

Taxon	Length (mm)	# Obs	#Obs (non-zero)	Comment for length value
Cuvierina atlantica	8.1	31	3	Thecosomata value used
Cuvierina columnella columnella	10			
Cuvierina spp.	8.1	62		Thecosomata value used
Desmopterus papilio	8.9	1	1	Pteropod value used
Diacavolinia spp.	8.1	62		Thecosomata value used
Diacria costata	2.3			
Diacria danae	1.7	31	14	
Diacria major	10.7	31	1	
Diacria quadridentata	3			
Diacria rampali	9.5			
Diacria trispinosa	8	277	56	Mean of subspecies
Diacria trispinosa trispinosa	8			
Diacria spp.	5.9	3708		
Gleba spp.	8.1	31		Thecosomata value used
Heliconoides inflatus	8.1	4755	2970	Thecosomata value used
Hyalocylis	8	162	9	Mean of subspecies
Hyalocylis striata	8	217	9	
Hydromylidae	12	7056	1	Gymnosomata value used
Limacina bulimoides	2	3732	466	
Limacina helicina antarctica	5	31	6	
Limacina helicina antarctica rangii	2			
Limacina helicina helicina	6	31	1	
Limacina helicina pacifica	5			

Table S2 continued from previous page

Taxon	Length (mm)	# Obs	#Obs (non-zero)	Comment for length value
Limacina helicina spp.	4.22	1538		
Limacina inflata	1.3	104	104	
Limacina lesueuri	0.8	1073		
Limacina rangii	2.98	62	7	Family value used
Limacina retroversa	2.5	9070	1422	
Limacina retroversa australis	2.5	62	3	Species value used
Limacina spp.	2.98	62618		
Limacina trochiformis	1	3741	1389	
Paedoclione doliiformis	1.5	3	3	
Peracle bispinosa	8.9	31	3	Pteropod value used
Peracle diversa	8.9	31	10	Pteropod value used
Peracle reticulata	8.9	524	133	Pteropod value used
Peracle valdiviae	8.9	31	5	Pteropod value used
Peracle spp.	8.9	4193		Pteropod value used
Pneumodermopsis	6.5	5	5	
Pneumodermopsis canephora	12			
Pneumodermopsis ciliata	15	1		
Pneumodermopsis macrochira	2			
Pneumodermopsis paucidens	5			
Pneumodermopsis polycotyla	5			
Pneumodermopsis pulex	8			
Pneumodermopsis simplex	5			
Pneumodermopsis spoeli	3			

Table S2 continued from previous page

Taxon	Length (mm)	# Obs	#Obs (non-zero)	Comment for length value
Pneumodermopsis teschi	9.1			
Spongiobranchaea australis	22	58773	103	
Spongiobranchaea australis larvae	10			
Spongiobranchaea spp.	15			
Styliola	13	8	8	Mean of subspecies
Styliola subula	13	66	29	
Telodiacria danae	8.1	62	11	Thecosomata value used
Telodiacria quadridentata	8.1	337	5	Thecosomata value used
Thielea helicoides	8.1	3184	119	Thecosomata value used
Euthecosomata	8.1	340250	43596	
Gymnosomata	12	2331	741	
Pteropoda	8.9	79613	14713	
Total		**841239**	**66978**	

Table S3. Foraminifer shape groups as defined for the following analysis. The images are exemplary for each shape type. Sources refer to the images.

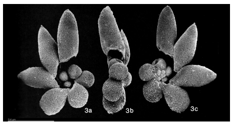
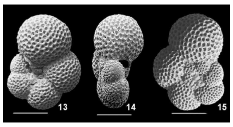
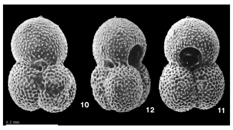
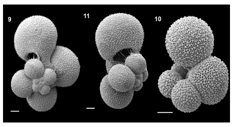
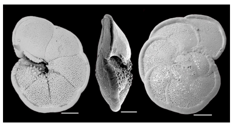
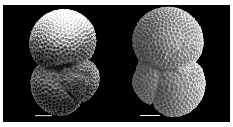
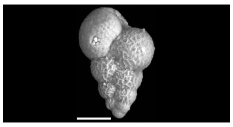
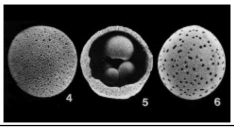
Group	Species	Example	Source
1 Digitate type	*Beella digitata* *Globigerinella adamsi* *Hastigerinella digitata*		Saito, Thompson, and Breger (1976)
2 Low trochospiral type	*Berggrenia pumilia* *Dentigloborotalia anfracta* *Tenuitella fleisheri* *Tenuitella iota* *Tenuitella parkerae* *Globigerinita humilis* *Turborotalita quinqueloba* *Orcadia riedeli* *Globigerinita minuta* *Globigerinoides tenellus* *Globorotaloides hexagonus* *Neogloboquadrina dutertrei* *Neogloboquadrina incompta* *Neogloboquadrina pachyderma*		Coxall and Spezzaferri (2018)
3 Medium trochospiral type	*Candeina nitida* *Globigerina bulloides* *Globigerina falconensis* *Globigerinita glutinata* *Globigerinoides ruber* *Globoquadrina conglomerata* *Globoturborotalita rubescens* *Spaeroidinella dehiscens* *Trilobatus sacculifer*		Loeblich and Tappan (1994)
4 Oblique planispiral type	*Hastigerina pelagica* *Globigerinella calida* *Globigerinella siphonifera*		Weiner, Weinkauf, Kurasawa, Darling, and Kucera (2015)
5 Discoidal-pyramidal type	*Globorotalia scitula* *Globorotalia theyeri* *Globorotalia crassaformis* *Globorotalia hirsuta* *Globorotalia menardii* *Globorotalia tumida* *Globorotalia ungulata* *Globorotalia truncatulinoides* *Globorotalia inflata*		Lam and Leckie (2020)
6 Subsphaeroidal type	*Pulleniatina obliquiloculata* *Globigerinoides conglobatus* *Sphaeroidinella dehiscens* *Trilobatus sacculifer*		Lam and Leckie (2020)
7 Elongate type	*Globigerinita uvula* *Streptochilus globigerus*		Miranda-Martínez, Carreño, and McDougall (2017)
8 Sphaeroidal type	*Orbulina universa*		Srinivasan and Kennett (1983)

Table S4: Average length values (µm) for the different foraminifer taxa. Values for individual species were collected from the images in Schiebel and Hemleben (2017). The third column indicates the number of data points per taxon present in the full quality controlled dataset, while the fourth column shows the number of non-zero abundance observations. For higher taxonomic levels than the species level, the fifth column indicates the choices taken for the length calculation.

Taxon	Length (µm)	# Obs	#Obs (non-zero)	Comment for length value
Beella digitata	300	5650	4	
Berggrenia pumilio	100	5650	3	
Candeina nitida	250	5650	4	
Dentigloborotalia anfracta	100	5650	80	
Globigerina bulloides	250	57372	3445	
Globigerina falconensis	250	5650	141	
Globigerina spp.	250	11	11	Mean of species used
Globigerinella adamsi	400	5650	29	
Globigerinella calida	300	5650	194	
Globigerinella siphonifera	300	5650	1018	
Globigerinita glutinata	250	5650	1986	
Globigerinita minuta	100	5650	87	
Globigerinita uvula	150	57367	117	
Globigerinoides conglobatus	300	5650	34	
Globigerinoides ruber	250	11300	1971	
Globigerinoides tenellus	150	11300	498	

Table S4 continued from previous page

Taxon	Length (µm)	# Obs	#Obs (non-zero)	Comment for length value
Globoquadrina conglomerata	300	5650	11	
Globorotalia theyeri	300	5650	150	
Globorotalia crassaformis	250	5650	74	
Globorotalia hirsuta	250	5650	539	
Globorotalia inflata	250	57367	1212	
Globorotalia menardii	400	5650	273	
Globorotalia scitula	150	5650	990	
Globorotalia truncatulinoides	300	5650	757	
Globorotalia tumida	300	5650	31	
Globorotalia ungulata	300	5650	33	
Globorotalia spp.	278	65829	128	Mean of species used
Globorotaloides hexagonus	250	5650	185	
Globoturborotalita rubescens	150	5650	264	
Hastigerina pelagica	500	5650	169	
Hastigerinella digitata	500	5650	12	
Neogloboquadrina dutertrei	250	5650	630	
Neogloboquadrina incompta	200	57367	2001	
Neogloboquadrina pachyderma	200	57367	1235	
Orbulina universa	400	5650	242	
Orcadia riedeli	150	51717	2	
Pulleniatina obliquiloculata	250	5650	46	
Sphaeroidinella dehiscens	300	5650	23	
Tenuitella fleisheri	100	5650	15	

Table S4 continued from previous page

Taxon	Length (µm)	# Obs	#Obs (non-zero)	Comment for length value
Tenuitella iota	100	5650	54	
Tenuitella parkerae	100	5650	159	
Trilobatus sacculifer	300	11300	929	
Turborotalita humilis	125	5650	181	
Turborotalita quinqueloba	150	57367	1258	
Planktic foraminifers	242	344819	80782	Mean of all species used
Total		**1021283**	**102007**	

Table S5. Total carbon (TC) biomass conversion factors (BCF) for foraminifers. These factors are derived from length and test weight measurements from Schiebel and Hemleben (2000) and Takahashi and Bé (1984). The conversion factor for foraminifers in total is derived from Michaels et al. (1995). All conversion factors are converted to total carbon (TC) biomass, using equations (4) to (6) in the main document.

Taxon	Biomass conversion factor (µg TC µm^{-3})
Species	
Globigerina bulloides	$1.1645 * 10^{-7}$
Globigerina falconensis	$1.9051 * 10^{-7}$
Globigerinella siphonifera	$0.7496 * 10^{-7}$
Globigerinita glutinata	$1.9304 * 10^{-7}$
Globorotalia hirsuta	$2.1544 * 10^{-7}$
Globorotalia scitula	$1.7367 * 10^{-7}$
Neogloboquadrina incompta	$2.1566 * 10^{-7}$
Turborotalita quinqueloba	$1.3571 * 10^{-7}$
Shape groups	
2 - Low trochospiral type	$1.7568 * 10^{-7}$
3 - Medium trochospiral type	$1.6667 * 10^{-7}$
4 - Oblique planispiral type	$0.7496 * 10^{-7}$
5 - Discoidal-pyramidal type	$1.9456 * 10^{-7}$
Foraminifers	$1.2109 * 10^{-7}$

Table S6. Hyperparameter options for the Random Forest (RF) model, the untuned parameter value and the final parameter choices for pteropods and foraminifers as determined via a grid search by assessing all hyperparameter options for those that would minimize the root mean squared error (RMSE). n_{tree} denotes the number of bootstrap samples created from the original dataset, using a fraction of r_{sample} of the entire data for each bootstrap. m_{try} refers to the number of predictors evaluated at each node for their ability to discriminate the data most clearly. min_{rows} describes the minimum number of observations in each terminal node and max_{depth} the maximum size of the tree. For an extensive description of the hyperparameters and their effects, refer to Boehmke and Greenwell (2019c).

Hyperparameter	Parameter values tested	Untuned parameter	Final value pteropods	Final value foraminifers
n_{tree}	30, 130, 230, 330, 430, 530, 630, 730, 830, 930	50	830	330
m_{try}	1, 2, 3	1	1	2
min_{rows}	1, 3, 5, 10	1	3	2
max_{depth}	10, 20, 30	20	30	10
r_{sample}	0.55, 0.632, 0.70, 0.80	0.632	0.80	0.632

Table S7. Hyperparameter options for the Gradient Boosting Machine (GBM) model, the untuned parameter value, and the final parameter choices for pteropods and foraminifers as determined via a grid search by assessing all hyperparameter options for those that would minimize the root mean squared error (RMSE). max_{depth} describes the maximum size of each individual tree and min_{rows} denotes the minimum number of observations in each terminal node. The model's learning rate is determined by r_{learn}. Each of the individual trees that together make up the GBM is trained on a a random fraction r_{sample} of the data, using a fraction $r_{samplecolumns}$ of the predictors. For an extensive description of the hyperparameters and their effects, refer to Boehmke and Greenwell (2019b).

Hyperparameter	Parameter values tested	Untuned parameter	Final parameter pteropods	Final parameter foraminifers
max_{depth}	1, 3, 5	6	5	5
min_{rows}	1, 5, 10 1		1	1
r_{learn}	0.01, 0.05, 0.1	0.3	0.01	0.01
r_{sample}	0.5, 0.75, 1	1	0.75	0.5
$r_{samplecolumns}$	$\frac{1}{3}$, $\frac{2}{3}$, 1	1	1	1

Table S8. Hyperparameter options for the Deep Learning (DL) model, the untuned parameter value, and the final parameter choices for pteropods and foraminifers as determined via a grid search by assessing all hyperparameter options for those that would minimize the root mean squared error (RMSE). The activation function describes the non-linear transformation applied at each neuron. The hidden layer structure determines the number of layers and the number of neurons per layer, e.g. (10, 10) denotes a network with two hidden layers of ten neurons each. λ_{L_1} and λ_{L_2} are weight parameters used for penalizing complexity. To avoid overfitting, L_1 (Lasso regression) or L_2 (Ridge regression) can be employed to add a penalty term based on the network weights. The strength of this penalizing factor is determined by the respective parameter λ. For an extensive description of all hyperparameters, refer to Boehmke and Greenwell (2019a).

Hyperparameter	Parameter values tested	Untuned parameter	Final parameter pteropods	Final parameter foraminifers
activation function	Rectifier, Rectifier with dropout, Tanh, Maxout, Maxout with dropout	Rectifier	Tanh	Tanh
hidden layer structure	(5, 5), (10, 10), (15, 15), (20, 20), (50, 50, 50)	(5) (20, 20)	(15, 15)	
λ_{L_1}	$0, 1 * 10^{-3}, 1 * 10^{-5}$	0	0	$1 * 10^{-3}$
λ_{L_2}	$0, 1 * 10^{-3}, 1 * 10^{-5}$	0	$1 * 10^{-3}$	$1 * 10^{-5}$

References

Bednaršek, N., Mozina, J., Vogt, M., O'Brien, C., & Tarling, G. A. (2012). The global distribution of pteropods and their contribution to carbonate and carbon biomass in the modern ocean. *Earth System Science Data*, *4*(1), 167–186. doi: 10.5194/essd-4-167-2012

Boehmke, B., & Greenwell, B. (2019a). Deep Learning. In *Hands-on machine learning with r* (1st ed., chap. 12). New York: Chapman and Hall/CRC. doi: https://doi.org/10.1201/9780367816377

Boehmke, B., & Greenwell, B. (2019b). Gradient Boosting. In *Hands-on machine learning with r* (1st ed., chap. 11). New York: Chapman and Hall/CRC. doi: https://doi.org/10.1201/9780367816377

Boehmke, B., & Greenwell, B. (2019c). Random Forests. In *Hands-on machine learning with r* (1st ed., chap. 11). New York: Chapman and Hall/CRC. doi: https://doi.org/10.1201/9780367816377

Coxall, H. K., & Spezzaferri, S. (2018). Taxonomy, biostratigraphy and phylogeny of Oligocene Catapsydrax, Globorotaloides and Protentelloides. *Cushman Found Foraminifer Res Spec Publ*, *46*, 79–124.

Davis, C. S., & Wiebe, P. H. (1985). Macrozooplankton biomass in a warm-core Gulf Stream ring: time series changes in size structure, taxonomic composition, and vertical distribution. *Journal of Geophysical Research*, *90*(C5), 8871–8884. doi: 10.1029/JC090iC05p08871

Kretschmer, K., Jonkers, L., Kucera, M., & Schulz, M. (2018). Modeling seasonal and vertical habitats of planktonic foraminifera on a global scale. *Biogeosciences*, *15*(14), 4405–4429. doi: 10.5194/bg-15-4405-2018

Lam, A. R., & Leckie, R. M. (2020, 7). Subtropical to temperate late Neogene to Quaternary planktic foraminiferal biostratigraphy across the Kuroshio Current Extension, Shatsky Rise, northwest Pacific Ocean. *PLOS ONE*, *15*(7), e0234351. Retrieved from `https://doi.org/10.1371/journal.pone.0234351`

Little, W. S., & Copley, N. J. (2003). *WHOI silhouette DIGITIZER version 1.0 user's guide* (Tech. Rep. No. July). Woods Hole: Woods Hole Oceanographic Institute. doi: 10.1575/1912/62

Loeblich, A. R., & Tappan, H. (1994). *Foraminifera of the Sahul Shelf and Timor Sea* (Vol. 31). Cushman Foundation for Foraminiferal Research.

Lombard, F., Labeyrie, L., Michel, E., Bopp, L., Cortijo, E., Retailleau, S., ... Jorissen, F. (2011). Modelling planktic foraminifer growth and distribution using an ecophysiological multi-species approach. *Biogeosciences*, *8*(4), 853–873. doi: 10.5194/bg-8-853-2011

Michaels, A. F., Caron, D. A., Swanberg, N. R., & Howse, F. A. (1995). Primary productivity by symbiont-bearing planktonic sarcodines (Acantharia, Radiolaria, Foraminifera) in surface waters near Bermuda. *Journal of Plankton Research*, *17*(1), 103–129. doi: 10.1093/plankt/17.1.103

Miranda-Martínez, A. Y., Carreño, A. L., & McDougall, K. (2017). The Neogene genus Streptochilus (Brönnimann and Resig, 1971) from the Gulf of California. *Marine Micropaleontology*, *132*, 35–52.

Saito, T., Thompson, P. R., & Breger, D. (1976). Skeletal ultramicrostructure of some elongate-chambered planktonic foraminifera and related species. In Y. Takayanagi & T. Saito (Eds.), *Progress in micropaleontology: selected papers in honor of prof. kiyoshi asano, special publication* (pp. 278–304). New York: ARRAY(0x55981508b818). Retrieved from `https://oceanrep.geomar.de/id/eprint/33439/`

Schiebel, R., & Hemleben, C. (2000). Interannual variability of planktic foraminiferal populations and test flux in the eastern North Atlantic Ocean (JGOFS). *Deep-Sea Research Part II: Topical Studies in Oceanography*, *47*(9-11), 1809–1852. doi: 10.1016/S0967-0645(00)00008-4

Srinivasan, M. S., & Kennett, J. P. (1983). The oligocene-miocene boundary in the South Pacific. *Geological Society of America Bulletin*, *94*(6), 798–812.

Takahashi, K., & Bé, A. W. (1984). Planktonic foraminifera: factors controlling sinking speeds. *Deep Sea Research Part A, Oceanographic Research Papers*, *31*(12), 1477–1500. doi: 10.1016/0198-0149(84)90083-9

Weiner, A. K. M., Weinkauf, M. F. G., Kurasawa, A., Darling, K. F., & Kucera, M. (2015). Genetic and morphometric evidence for parallel evolution of the Globigerinella calida morphotype. *Marine Micropaleontology*, *114*, 19–35.