Key link between iron and the size structure of three main mesoplanktonic groups (Crustaceans, Rhizarians, and colonial N_2-fixers) in the Global Ocean.

Authors

Mathilde Dugenne,[1]* Marco Corrales-Ugalde,[2] Jessica Y. Luo,[3] Lars Stemmann,[1] Jean-Olivier Irisson,[1] Fabien Lombard,[1] Todd O'Brien,[4] Charles Stock,[3] PSSdb data contributors consortium[5], Rainer Kiko,[1,6]*

Affiliations

[1]: Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France
[2]: Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA
[3]: NOAA/OAR Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA
[4]: NOAA Fisheries - Office of Science and Technology, Silver Spring, Maryland, USA
[5]: Full list of authors consortium in Supplementary material
[6]: GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

*** mdugenne@hotmail.com, rkiko@geomar.de**

Abstract

Size is commonly used as a master trait to characterize pelagic organisms as it affects a range of processes and impact marine biogeochemical cycles and services. Yet, a holistic understanding of what environmental factors shape size structure is lacking for most mesoplankton. As part of the Pelagic Size Structure database, we explore the linkages between environmental factors and global compilations of Rhizarian, colonial N_2-fixer, and Crustacean size spectra measured from Underwater Vision Profilers or benchtop scanners. We found that iron, alongside temperature, plays a disproportionate role in shaping their spectral biogeography. Our results highlight the effect of dust on N_2-fixers and Rhizarians while total iron, comprising organic and mineral compounds, explained most of the variance in Crustaceans size structure. Using machine learning models, we predicted their size structure at the global scale with relatively high R^2 of 0.93, 0.84, and 0.66. We hope our predictions can support further assessment of their role in biogeochemical processes under present and future forcings.

Introduction

For decades, ecologists and modelers have summarized planktonic community structure using size spectrum (*1–5*), representing the universal decline of pelagic organism abundances in increasing size classes (*1*), as they relate to ecosystem functioning and services (*6*). Key ecosystem processes linked to organism sizes include CO_2 fixation by marine phytoplankton (*7*) and its respiration by primary producers and consumers (*8*), predation (*9*), as well as organic matter packaging (*10*, *11*) and passive sinking (*12*). These eventually affect essential services of marine ecosystems: carbon export, related to long-term sequestration of atmospheric CO_2 into the Ocean's interior, and fish production. The first service -carbon export efficiency- has been shown to vary in response to plankton community structure, with higher efficiency linked to increased relative proportion of microphytoplankton (20-200 µm) or mesozooplankton (200-20,000 µm) in both data compilations (*13*) and size-structured models (*14*). The second - production of commercial fisheries stocks - has been generally linked to the transmission of energy across plankton

food webs by size-selective feeding, referred to as trophic efficiency (*15*, *16*). Since size-dependent processes underpin all links of marine food webs, changes affecting plankton communities can be observed through the lens of the size spectrum framework via three parameters (*17*): the spectral intercept (which will remain the same if the change affects all plankton groups equally), the slope (which will vary with differences in transfer efficiency), and the linearity (which may decrease under perturbations that affect only certain taxa). The size-structure framework is often used in conjunction with other traits (e.g. feeding strategy, migration, buoyancy) to represent plankton communities along a few functional axes, instead of numerous taxa that cannot all be monitored or cultivated (*18–25*). To do so, ecologists proposed to use new imaging devices that provide high-throughput proxies of functional complexity across the five orders of magnitude in plankton sizes (*24*, *26*).

These traits are deemed "functional", as they are tied to a few mechanistic principles influencing ecosystem functioning by affecting organisms' fitness via resource acquisition, growth, reproduction, or survival (*27*). The underlying changes in plankton fitness along environmental gradients eventually pertain to the self-organization of communities according to abiotic and biotic pressures exerted on specific traits. In marine ecosystems, numerous plankton communities were shown to self-assemble on a global scale according to their functional traits. For example, families of copepods (*28*), and other broad groups of zooplankton (*29*), typically present distinct biogeography primarily driven by their diet (e.g. carnivore versus omnivore versus filter-feeders). Mixotrophs inhabit diverse ecosystems depending on the mode of symbiosis (e.g. "phytoplankton" that ingest prey or "zooplankton" that steal chloroplasts or harbor photosynthetic symbionts) and their ability to form colonies (*30*). Likewise, phytoplankton with the ability to fix atmospheric N_2 into bioavailable ammonia (N_2-fixers) have been found in a range of different habitats (from coastal to open ocean and from subtropical to polar ecosystems) favoring specific forms (free-living, colonial, symbiotic) or symbiosis (*31*). All these communities vary in size with latitudinal gradients (*32–34*), in accordance with the set of empirical rules and theoretical principles linking environmental gradients to plankton size structure reviewed by Barton et al. (*5*). In this review, the authors stressed that phytoplankton are thought to respond primarily to light and nutrient availability (nitrogen for non N_2-fixers and iron/phosphorus for N_2-fixers) affecting spectral slopes and intercepts (*35*), while zooplankton, like most metazoans, have been shown to grow bigger under cold temperature. This relationship, known as the Temperature-Size Rule (*36*), has first puzzled ecologists since low temperature considerably slows down plankton growth, and is now supported by alternative hypotheses linked to organism physiology (*37*, *38*). In addition to these known bottom-up controls, plankton size biogeography is also determined by top-down pressures, which are harder to link to large-scale size variability without observing ecological succession through extended temporal surveys located across the Globe.

Overall, Barton et al. (*5*) identified seasonal and latitudinal variations of temperature and oxygen levels, resource availability and prey size structure, and ocean circulation as all major structuring elements of the pelagic realm. Yet, their effects are often hard to disentangle as most factors co-vary in marine ecosystems and some relationships appear contradicting. For example, Brun et al. (*33*) demonstrated that responses linking copepod average body size and environmental descriptors could be counter-intuitive, as body size decreased under high temperature, low oxygen levels, in accordance with the Temperature-Size Rule, but also with productivity and phytoplankton cell size. The

authors suggest that additional functional traits, namely alternative diet and feeding behaviors, should help explain this contradiction. Similarly, Barton et al. (*5*) hypothesized that opportunist mixotrophs have a greater advantage in oligotrophic ecosystems, where sources of inorganic and organic nutrients are scarce, but also in more productive systems that undergo large seasonal fluctuations in resource availability, like temperate and polar ecosystems. The resulting partition between oligotrophic and productive ecosystems has been evidenced in different classes of mixotrophic Rhizarians, monitored by non-intrusive imaging approaches due to their fragility (*39*). Barton et al. (*5*) also pointed to the large uncertainty regarding future projections of marine N_2-fixation, as greater temperature should increase the growth of large N_2-fixers like *Trichodesmium*, but also limit the supply of their essential nutrients which could benefit taxa or morphotypes with a larger surface:volume ratio. Thus, the global driver(s) of plankton size biogeography remain largely enigmatic, in part due to the paucity of direct size structure compilations and related environmental gradients (*5*, *24*). Current trait biogeographies rely on the correspondence between certain taxa and specific functional traits. Or, in the case of *Trichodesmium*, colonies typically organize in two morphotypes, spherical Puffs and fusiform Tuffs, whose shape and surface:volume ratio vary greatly. Environmental cues and mechanisms involved in this difference have remained enigmatic (*40*) since most species are able to form both morphotypes (*41*). Similarly, the first global meta-analysis of copepod trait biogeography (*33*) has shown that environmental factors like sea surface temperature, net primary production, or chlorophyll-a concentration, only explained half of the dataset's variance in space or time, leaving important sources of variability unexplained that the authors linked to further clustering within the group of Crustaceans.

Here, we present a global compilation of direct size estimates in the 200-20,000 µm size range, defined as mesoplankton by Sieburth et al. (*42*), associated with three groups that dominate community abundances in the upper ocean (*43*): Colonial N_2-fixers, Rhizarians, and Crustaceans. We depict the spectral biogeography of Colonial N_2-fixers and Rhizarians using a range of Underwater Vision Profilers (UVP), since these organisms can only be monitored by non-intrusive imaging systems, and net collected samples imaged with benchtop scanners for Crustaceans. We explore the relationships between their size distributions and environmental descriptors like temperature, resource availability, including inorganic nutrients and prey, or mesoscale circulation, and provide direct evidence of niche partitioning within these planktonic groups. We used the linkages between size structure and environmental gradients in conjunction with machine learning models (i.e. boosted regression trees) to generate global predictions of group-specific size structure. With these predictions, our overarching goal was to fulfill the four objectives of plankton trait biogeography studies defined by Barton et al. (*5*): (1) provide a mechanistic understanding of the factors structuring plankton communities, (2) understand the cause of spatio-temporal variability in community functions like carbon export or N_2-fixation, (3) validate or help refining plankton models, with the goal to (4) constrain the future projections of ecosystem processes under global change. Like Barton et al. (*5*), we recognize that mesoplankton size biogeography is ultimately determined by the interplay between both bottom-up and top-down (e.g. size-selective predation by fish or other nekton) controls, though the latter are out of the scope of this paper as this would require an extensive characterization of larger predators (through space and time) that are poorly detected with the imaging systems we used. Besides identifying potential drivers of trait distributions in the global oceans, our results provide insights into how mesoplankton affect marine ecosystems biogeochemistry and services (N_2-fixation, carbon fluxes and export) in response to environmental stressors. With the predictions of mesoplankton spectral

biogeography and associated compilations of pelagic size spectra, expressed as a function of cell abundances or carbon biomass (data not shown), our hope is to inform the development of biogeochemical and other size-structured models used to foresee the changes in ecosystem functioning and services under future climate change. Indeed, the complexity of plankton functional types represented in the 14 biogeochemical models compared as part of the latest Coupled Model Intercomparison Project, has significantly increased with the support of dataset compilations such as MAREDAT (https://doi.org/10.5194/essd-5-227-2013) and COPEPOD (https://www.st.nmfs.noaa.gov/copepod/). However, these models are still evaluated against bulk biological properties (e.g., surface chlorophyll concentration, net primary productivity, and bulk mesozooplankton biomass), despite increasing evidence that plankton size structure, rather than bulk biomass and productivity estimates, is amongst the best predictor of ecosystem services like fish production or carbon export.

Results and Discussion

1. Size spectrum of N_2-fixing *Trichodesmium*, Rhizarians and Crustaceans and evidence for their niche partitioning in the Global Ocean

Since 2004/2008, benchtop scanners and UVP have been regularly used to image mesoplankton in all major ocean basins, with the highest concentration of samples recorded within the Subtropics and less emphasis given to the Southern or Indian Ocean (Figure 1A). Here, we benefited from a recent collaborative and international effort, the Pelagic Size Structure database (PSSdb), to compile millions of size estimates with their corresponding taxonomic annotations generated during these deployments (Figure 1B). The standardization of all annotations revealed that colonial N_2-fixers, Rhizarians and Crustaceans, whose vignettes are illustrated in Figure 1C, are globally important members of the mesoplankton communities. The first two groups represent 45% and 15% of the total image diversity in UVP datasets, while the latter correspond to 81% of the overall scanner images (Figure 1B). This result is consistent with a recent study highlighting the main communities of mesoplankton in the epipelagic layer based on UVP profiles solely (*43*). In this study, we used UVP profiles to characterize the occurrences of colonial N_2-fixers and Rhizarians which are both strongly undersampled when collected with plankton nets due to mesh exclusion or destruction (*39*). Though Crustaceans are also relatively well imaged *in situ* with the UVP, known underestimation of their stocks in comparable size ranges (*44*) possibly due to active avoidance (*45*), led us to use the scanner dataset exclusively to quantify the spectral biogeography of this group. In addition, the higher resolution of the scanners also allows to classify images at higher taxonomic ranks (Figure 1B), mostly to the family level that could be associated to specific trophic groups (e.g. carnivore or omnivore) based on the compilation of Benedetti et al. (*46*).

The global compilation of mesoplankton size structure presents similarities and differences between the three main groups in the given size range (Figure 2). All groups show a log-linear decline in abundance with increasing size classes, with slopes (Figure S1) relatively close to the canonical value of -1 highlighted in the first synthesis of plankton size distributions published by Sheldon et al. (*1*). Their total abundances, obtained by integrating the size spectra, mostly range between ~0.1 and ~1000 particles m^{-3}, representing a significant spatio-temporal variability across the PSSdb datasets. Colonial N_2-fixers present the highest percentage of empty records, with 64% of the bins without occurrences, while empty bins represented only one third and one tenth of the dataset for Rhizarians and Crustaceans respectively (Figure 2B). Modeled abundances, derived from the predicted

NBSS parameters (slope, intercept, size range), showed very contrasting biogeography across the three main mesoplankton groups, with a good agreement between observations and predictions (Figure 2C) corresponding to the different habitats discussed in Panaïotis et al. (*43*). Likewise, we found that colonial N_2-fixers were restricted to the subtropical latitudes, that Rhizarians were mostly present in the subtropics although they were also relatively abundant at higher latitudes and in the Californian upwelling system, and that Crustaceans were generally ubiquitous, with the highest abundances observed in productive ecosystems (e.g. poles, upwelling regions and coasts). Caution should be taken when using or interpreting predictions in the Southern Ocean, but our results suggest that Rhizarians could also be numerically important in this undersampled region. Since these differences could result from specific adaptations of mesoplankton subgroups, also referred to as niche partitioning, we investigated the linkages between environmental factors and mesoplankton size distribution by grouping images at different taxonomic ranks.

Statistics describing the goodness of fit of the machine learning model applied to increasing partitions within the three main mesoplankton groups are summarized in Table 1. In general, all models presented relatively high R^2 (>0.6), except for rare groups of Phaeodaria (R^2=0.26) and Foraminifera (R^2=0.25). The latter was not surprising since metabarcode sequenced as part of the Tara Oceans global expedition showed that Foraminifera did not present a clear relationship with environmental factors (*30*). Grouping images based on increasing taxonomic ranks resulted in overall improved model performances for all groups, particularly after accounting for model complexity and the number of observations. The decrease in model information criteria (BIC and AIC estimates), also indicative of increased model likelihood, were more marginal for the colonial N_2-fixers due to the dominance of a specific class (i.e., Oscillatoriales) and morphotype (i.e., Tuff) in the UVP images. Their dominance has been observed across seasonal trends recorded at the Long-term Oligotrophic Habitat Assessment long-term station (Station ALOHA) in the North Pacific Subtropical Gyre as well as in the Tropical Atlantic (*47–49*). In comparison, both Rhizarians and Crustaceans presented larger differences in model performance metrics in increasing partitions of the datasets (e.g. increasing taxonomic ranks) despite lower available observations, supporting strong evidence for their niche partitioning in the Global Ocean.

2. Temperature and iron are amongst the best environmental descriptors of mesoplankton spectral biogeography.

The feature importance metric of the model fits was used to estimate which environmental factors tested (Figure S2) drove the differences in size biogeography of the main mesoplankton groups in the epipelagic layer. The top and last five environmental descriptors ranked in terms of their gradual importance in the model variance explained are presented in Figure 3A, along with their associated correlation coefficients with NBSS integral. Variables at intermediate ranks are not shown since they had relatively low scores and also lower correlation coefficients with NBSS integral (Table S1).

Sea surface temperature (SST) ranked first for colonial N_2-fixers and Rhizarians, and fourth for Crustaceans, thus representing one of the main factors explaining the global patterns in mesoplankton size distribution. This finding is consistent with the global analysis of UVP (*43*) and scanned nets collected during the Tara Oceans global expedition (*32*), as well as a meta-analysis of phytoplankton (*50*), mixotrophic Rhizarians (e.g. Collodaria, Acantharia and Foraminifera) (*30*), and copepod (*33*) biogeography. In all these

studies, temperature gradients explain up to 50% of the total variance in plankton biogeography. Here, the importance of SST in model fits reflects its high correlation with total counts of the different mesoplankton groups (Figure 3A, Table S1), with significant coefficients ranging between 0.4 and 0.5 in absolute value, except for Foraminifera which were more globally ubiquitous (*30*). While both colonial N_2-fixers and Rhizarians were positively correlated with SST, with their occurrence highest in the subtropics (Figure 2C), Crustaceans exhibited a negative correlation, consistent with their higher abundances in polar and upwelling regions (Figure 2C). Overall, Crustaceans spectral biogeography was very similar to previous model predictions of their total biomass in the epipelagic layer (*44*), so we looked at specific linkages between NBSS parameters (intercept and slope) and the best environmental predictors to further tease out the effect on changes in abundance or size on the distributions of total biomass and NBSS integral (Figures 4-6A). This analysis showed that although there were differences in the relationships between NBSS intercepts and SST between Crustaceans and colonial N_2-fixers/Rhizarians, all coefficients were consistent with respect to spectral slopes. All groups appeared larger under cold temperature (r(Colonial N_2-fixer slopes, SST)=-0.23, r(Rhizarian slopes, SST)=-0.29, r(Crustacean slopes, SST)=-0.48, p-values<0.001), in agreement with the general Temperature-Size Rule (*36*).

The first hypothesis explaining the emergence of larger metazoans in cold environments was linked to increased oxygen availability, since spatial gradients in surface temperature are opposed to that of dissolved oxygen and apparent oxygen utilisation in marine ecosystems (Figure S2). Oxygen exerts a strong metabolic constraint on Crustaceans, with concentrations between 15-50 μM of dissolved oxygen typically defined as critical to maintain aerobic reactions and survival (*51*, *52*). It is thus expected that increased oxygen availability poleward may support larger Crustaceans at cold temperatures. Here, predictions of mean crustaceans size (Figure 6C) follow this expectation, with larger organisms found in polar, temperate and upwelling ecosystems. Further, all Crustacean groups showed a positive correlation between spectral slopes and dissolved oxygen concentration, indicating a flattening of their spectra under high oxygen concentrations (Table S1). Brandão et al. (*32*) found that the trend of larger copepods at higher latitudes were mainly due to a shift in copepod families, with large omnivore-herbivore Calanaide dominating towards the pole and small carnivorous copepods preferentially found in the subtropics. Our results also follow this pattern, as indicated with our predictions of carnivorous and omnivorous copepod NBSS obtained after grouping size estimates at the family level assigned to a specific diet using published compilation (Figure 6C).

In contrast to that of Crustaceans, the average size of Rhizarians and N_2-fixers did not vary as much across ecosystems. While they showed increased sizes in specific regions – at the transition between gyres and Equatorial currents and in the Arctic Ocean for Rhizarians or in the Western Subtropical South Pacific Gyre for colonial N_2-fixers – systematic trends were not evident (Figure 4-5C). The areas where Rhizarians increased in size were co-dominated by Phaeodaria, which are exclusively heterotrophs, suggesting that the link between temperature and increased size for some Rhizarians may also be mediated by oxygen availability. However, Phaeodaria NBSS slopes did not show a significant correlation with oxygen concentration (Table S1). Since endosymbiotic Rhizarians and colonial N_2-fixers do not require high oxygen concentrations to fuel their metabolism (they rather support the production or directly produce oxygen by photosynthesis), it is unlikely that oxygen levels are responsible for the Temperature-Size Rule in these groups. In

addition, oxygen is known to inhibit nitrogenase, so higher oxygen concentration could be detrimental for colonial N_2-fixers (*53*). The other hypothesis for the Temperature-Size Rule invokes differences in anabolic (i.e growth) versus catabolic (i.e respiration) rates with respect to changes in temperature. These are typically quantified over a 10°C temperature change, and referred to as temperature coefficients (Q_{10}). According to this hypothesis, organisms can stay smaller under high temperature only if catabolic rates increase more rapidly with temperature compared to anabolic reactions. In other words, it applies to organisms that present higher Q_{10} coefficients for respiration than for growth (*36*, *54*). Elevated respiratory Q_{10}, indicative of high respiration at high temperature, were shown to support the anoxic conditions that favor N_2-fixation within colonies of *Trichodesmium* (*55–57*). Similar differences in anabolic and catabolic Q_{10} measured from enclosed experiments with Foraminifera (*58*) or Collodaria (*59*) hint that similar regulatory effects on metabolic reactions may come into play for symbiotic Rhizarians to explain their flatter size spectra at low temperature.

More surprisingly, we found that different iron proxies were systematically in the top ranked descriptors explaining the variability in mesoplankton spectral biogeography (Figure 3A). Other biogeographic studies have not included any iron proxy, which could explain why the importance of iron has been overlooked so far and why they were only able to explain data variance up to 50% (*33*). Our model included several proxies linked to different, but not mutually exclusive, pools of iron to reflect the diversity in natural iron compounds found in epipelagic marine ecosystems (*60*). Labile compounds were mostly represented by aerosols, either from desert sources (best approximated by wet dust deposition satellite products) or from other sources like volcanic and wildfire ashes or anthropogenic activities (accounted for in the total aerosol optical thickness satellite proxy). The peak of N_2-fixing *Trichodesmium* abundance observed in Figure 2C, and its enrichment in iron, has been largely attributed to Saharan dust deposition in this ecosystem (*61–63*), in agreement with the positive correlation coefficients between the satellite product and our observations ($r=0.46$, p-value<0.001). Furthermore, we found a significant relationship between wet dust deposition and our observations of Rhizarians size spectra, as dust was the second most important variable in predicting their spatio-temporal patterns (Figure 3A). To our knowledge, such a link has not been demonstrated before. A second proxy, representing the total iron concentration accounting for both mineral and organically-bound Fe ligands, from the biogeochemical model PISCES, was also among the top-5 scores in explaining the variance in Rhizarian spectral biogeography. Global patterns in monthly average observations of in situ dissolved iron concentration measured as part of the GEOTRACES program also matched that of two Rhizarian groups (i.e Phaeodaria and Foraminifera), with very high correlation coefficients (>0.77, Table S1). This gives us further confidence in the linkages found between the total iron model product and mesoplankton spectral biogeography.

To identify which iron proxy was best correlated with our main mesoplankton groups, we computed the temporal correlation coefficients between NBSS predictions and each iron proxy in individual 1°x1° grid cells (Figure S3). The product with the maximum coefficient was considered the main iron proxy for any given mesoplankton group (Figure 3B) and the total number of grid cells with a specific assignment (dust, aerosol, total iron) were later compared in broad ecosystems categories (e.g Polar, Temperate-Subpolar, Subtropical, Equatorial-Upwellings) to assess their seasonal trends in relation with our NBSS predictions (Figure 3C). This analysis revealed that the peak of modeled abundance of Rhizarians (Figure 2C), and more specifically Phaeodaria, in the Southern Ocean and upwelling

ecosystems was likely linked to increased concentrations of total iron (Figure 3B-C). Their presence at high latitudes is consistent with the first quantitative assessment of class-specific concentrations measured in 877 stations distributed worldwide, albeit in Biard et al. (*39*) the overall abundance of Phaeodaria appeared far more reduced in the austral ecosystem compared to Subtropical latitudes. Likewise, the seasonal trends between NBSS predictions and the main iron proxy support the role of dust and aerosol in shaping the size distribution of N_2-fixers in the Atlantic and Pacific Subtropical Gyres respectively (Figure 3B-C). In the Atlantic, where the main iron pool is provided by Saharan dust deposition, seasonal trends of N_2-fixer NBSS strongly followed the climatology of the dust product, marked by a peak of iron supply early in the year that progressively decreases to a minimum observed in June during the dry season. Similarly, N_2-fixer abundances were low until June and increased later in the year. Seasonal trends were relatively similar in the Pacific Ocean, although mostly linked to total aerosol, with a main peak in abundance observed in winter (Oct-Dec in the Northern hemisphere and Jun-Aug in the Southern hemisphere, Figure 3C), consistent with a negative correlation between N_2-fixer abundances and light availability (Table S1). In the North Pacific Subtropical Gyre (NPSG), N_2-fixation normally undergoes both phosphorus or iron limitation linked to climatic and seasonal anomalies in the Pacific Decadal Oscillations and dust supply from Asia respectively (102). We thus hypothesize that phosphorus limitation could be responsible for the weaker correlation between N_2-fixers and aerosol optical depth in the NPSG during the summer months (103). Correlations with aerosols were highest in the South Pacific Subtropical Gyre (SPSG) (Figure 3B), where anthropogenic dust is supplemented by ashes produced by volcanic activities or wildfires (*64*, *65*). Phytoplankton blooms, followed by satellite chlorophyll-a concentration, have been regularly linked to direct ashes deposition from Southern tropical islands (Papua New Guinea, Vanuatu and more recently Tonga) volcanic eruptions (*66*), Australian wildfires (*67*), and recently to shallow hydrothermal emissions located across the Tonga-Kermadec volcanic arc (*68*). The latter have been implicated in the iron fertilization of surface waters leading to elevated concentrations of N_2-fixers (*69*), more specifically *Trichodesmium*, and a shift in the mesoplankton community (*70*). Zooscan analyses highlighted the presence of gelatinous zooplankton like Chaetognaths or Salps in the vicinity of the volcanic arc, in an ecosystem otherwise dominated by copepods. Unfortunately, nets severely damage the fragile Rhizarians, so it was not possible to quantify the impact of iron supply by volcanic emissions, and more generally by aerosols, on this taxonomic group. The climatology in Rhizarians NBSS predictions in association with aerosol proxy in the SPSG do suggest that volcanic ashes could promote the growth of symbiotic Rhizarians like Acantharia and Collodaria, which present significant positive correlations with aerosol (Figure 3A, Table S1).

Compared to N_2-fixers and Rhizarians, Crustaceans were almost exclusively linked to total iron concentration in all ecosystems (Figure 3B). This is consistent with the relative importance of this product in explaining the observed variance in their spectral biogeography, surpassing temperature or prey availability (Figure 3A). Our model included proxies of phytoplankton particle size distribution slope (*71*), indicative of the prevalence of small picophytoplankton (0.2-2 μm) relative to larger nano-(2-20 μm) and microphytoplankton (20-50 μm), in addition to satellite chlorophyll-a concentration (Figure S2). Despite these proxies for phytoplankton productivity and community structure, total iron concentration was still the top-ranked environmental factor explaining the variance in Crustacean model fits, especially that of copepods (r(Copepoda, model iron)=0.44, (Copepoda, GEOTRACES iron)=0.5, p-value<0.001). As a result, predictions of their total abundances, obtained by integrating the size spectra, closely followed the seasonal trends in

total iron concentrations in ecosystems where both observations and predictions were maximum (Figure 2C, 3C). In the North Pacific Temperate latitudes, Crustacean abundances increased in the spring, following the peak in microphytoplankton size fraction observed by satellite. Likewise, the pool of total iron presented a peak of concentration between Feb-Apr, consistent with the increased Ekman pumping and deep mixing events occurring in the wintertime. Copepods, which dominate over other Crustacean groups across all ecosystems (Figure 6C), were strongly negatively correlated to absolute mixed layer depth globally (r(Copepod)=-0.4, p-value<0.001), in agreement with the Dilution-Recoupling hypothesis formulated by Behrenfeld et al. (*72*) to explain the spring bloom in temperate ecosystems. In comparison, the wind-driven Ekman pumping located in the Eastern Equatorial Pacific is more sustained throughout the year (*73*), although Trade winds generally intensify between June-August, resulting in a small increase in total iron and overall weaker seasonal trends caused by upwelling (Figure 3C). Yet, both the fraction in microphytoplankton observed by satellite and our predictions of Crustacean abundances seemed to be linked to the overall availability of total iron in this ecosystem, with a peak of concentration recorded in September. This analysis hence supports the long-standing recognition that dissolved iron exert a strong control on phytoplankton dynamics (*74*) and associated grazers, particularly in "High Nutrient Low Chlorophyll" (HNLC) regions located in the Equatorial Pacific and California Current upwelling systems, as well as in the Southern Ocean, subarctic Pacific, and subpolar North Atlantic. In the latter ecosystems, our knowledge on seasonal trends is limited by the sea ice coverage, the lack of sunlight and higher cloud coverage in the wintertime, as highlighted in important temporal and spatial gaps in both satellite and model products (Figure S2), resulting in only partial predictions of Crustaceans size distribution at least in the Southern Ocean (Figure 3C). Nonetheless, we found that total iron concentration and Crustacean abundances were higher during the austral summer than during the winter, which is unexpected given that phytoplankton blooms initiated in the summer are thought to considerably drawdown surface iron concentrations (*75*). The authors suggested that in the Southern Ocean, "higher dissolved iron concentrations associated with high chlorophyll-a could be reconciled by assuming high rates of Fe recycling associated with greater biomass". Iron fertilization experiments have confirmed the important role of copepods in driving iron recycling, partly sustaining phytoplankton blooms in this ecosystem (*76*). Similarly in the Arctic Ocean, we found that total iron concentration matched the seasonal trends in Crustaceans NBSS integral very closely, with both concentrations remaining high even after the spring bloom, occurring between Apr-May (Figure 3C). In this ecosystem, exogenous inputs of iron by river discharges peak in June and could thus explain elevated iron concentrations (*77*). However, an iron fertilization experiment occurring during this period also pointed to the rapid utilization of iron by copepods (*78*), suggesting that the strong correlation between Crustaceans and total iron concentration wasn't necessarily biased by exogenous inputs. Further, a recent comparison between a compilation of total dissolved iron and gridded products (*79*) similar to our analysis have shown that iron was strongly correlated to apparent oxygen utilisation (AOU), which appears positively correlated to Copepods biogeography in our analysis (Table S1). These empirical evidence, together with our global analysis linking iron pool to mesozooplankton spectral biogeography, suggest that Crustaceans could exert a strong control on iron recycling by predation and remineralization, further alleviating iron limitation of phytoplankton prey in productive ecosystems.

3. Additional resources, accounting for trophic strategies or other nutrients, and mesoscale circulation further partition the biogeography of broad mesoplankton groups.

As discussed above, significant correlation coefficients between environmental factors and one, often dominant, taxon underscore the importance of temperature and iron controls as main drivers of mesoplankton spectral biogeography (Table S1). However, many environmental factors with less importance in explaining the variance of broad mesoplankton size distributions appeared significantly correlated to rarer taxa (Table S1), and sometimes only to one of the NBSS parameters (intercept, slope, size range) used to train our habitat models (Figures 4-6A). Thus, in this section we explore the linkages between environmental factors and specific groups of colonial N_2-fixers, Rhizarians, or Crustaceans, or specific NBSS parameters (intercept, slope), and highlight some of their adaptations that help explain the niche partitioning within broad mesoplankton phyla.

3.1 Environmental adaptations of colonial N_2-fixers morphotypes

In the epipelagic layer, large-scale NBSS patterns in colonial N_2-fixers generally matched that of the two *Trichodesmium* morphotypes, Puff and Tuff. This was indicated by similar relative proportions in NBSS integrals (Figure 4C) and by the marginal increase of model performance metrics (AIC and BIC) when grouping annotations on the phylum and morphotype levels (Table 1). Nonetheless, some areas of the globe displayed elevated proportions of one particular morphotype, the Puff, near the Kuroshio current and near the Tonga-Kermadec volcanic arc (Figure 4C bottom panel). The increased proportion of Puff in these locations correspond to grid cells where predicted mean size of colonial N_2-fixers significantly increased (Figure 4C top panel), and where the main iron proxy determined by maximum correlation coefficients with monthly NBSS predictions was aerosol (Figure 3C). Using the same correlation coefficients analyses on our observations rather than on the predicted NBSS, we found that Puff were particularly responsive to aerosol in the Western Pacific Ocean, and that increased availability of anthropogenic dust or ashes had a positive effect on both the intercept and the slope of Puff NBSS (Figure 4A-B). Positive effects on the slope are indicative of flatter NBSS, and thus of an increase of the proportion of larger colonies compared to small ones. Increased proportions of Puff were also recorded at Station ALOHA (*47*) in the same months we observed elevated aerosol supply using satellite products (Figure 3C). *Trichodesmium* colonies, more specifically Puff, have been shown to increase in abundance in response to input of mineral iron, including from dust (*80*) and volcanic ashes or plumes (*70*), thanks to active motion of the colony and its microbiome (*81*). Although most studies relating dust supply to iron uptake focused on Puff, we found that Tuff were equally, if not more favored by Saharan dust deposition (r(Tuff)=0.37 and r(Puff)=0.33). Compared to other aerosols, Saharan dust did not appear to have a positive impact on the size of the colonies directly, and rather impacted spectral slopes through its effect on colony abundance (Figure S1A), as indicated by the consistent negative correlations between NBSS intercepts and slopes (Figure 4B). After tracking both morphotypes in an enclosed environment placed under stimulus, Pfreundt et al. (*82*) found that Puff likely form by merging Tuff together. This suggests that our observations could result from the initial stimuli of Tuff by Saharan dust, which increased in abundance and interacted with other Tuff colonies in spatial proximity to progressively form Puff. This mechanism would support both the similar proportions of *Trichodesmium* morphotypes predicted in the Subtropical Atlantic (Figure 4C), in addition to their similar response to the dust satellite products, as opposed to the preferential response of Puff to aerosol products (Figure 4A).

In addition, Tuff size distributions appeared linked to two proxies of mesoscale circulations: the sea level anomalies (SLA), indicative of isopycnals uplifting (negative anomalies or cyclones) and downwelling (positive anomalies or anticyclones), and finite size Lyapunov exponent (FSLE), which increases (in absolute values) at fronts due to the rapid divergence of water parcels in close proximity. While there was no significant correlation between NBSS intercepts and SLA for either morphotype, Tuff colonies appeared relatively larger under negative SLA. In the NPSG, negative SLA anomalies have been linked to the surface depletion of phosphates and their upwelled injection at the base of the chlorophyll maximum (*83*), thus larger colonies could result from an increased phosphate acquisition through their known diel vertical migration (*84*). The cyclonic side of mesoscale fronts have also been associated with blooms of *Trichodesmium* through physical accumulation (*85*). The physical accumulation of buoyant plankton was evidenced in a number of fronts indicated by high absolute FSLE, and is considered a major driver of the surface (0-20m) blooms of colonies of *Trichodesmium* and other diazotrophs (*48*). In our global compilation, we found that Tuff colonies get preferentially affected by both phosphate limitation, as indicated by the negative correlation between Tuff NBSS intercepts and phosphate concentrations, and frontal structures (Figure 4B). FSLE appears to positively impact the size of Tuff colonies, by promoting larger colonies that can effectively accumulate at the surface of fronts by forming long rafts (A long raft is effectively counted as one particle using the UVP, hence the apparent decorrelation between NBSS intercepts and FSLE) visible from space.

3.2 Environmental adaptations of Rhizarian classes

Compared to N_2-fixers morphotypes, our investigation of model predictions within the different classes of Rhizarians (Acantharia, Collodaria, Phaeodaria, Foraminifera) yielded very distinct spatial patterns indicative of important niche partitioning and environmental adaptations within this phylum (Figure 5). Indeed, this group showed the largest differences in model performance metrics, AIC and BIC estimates, when grouping size estimates and associated taxonomic annotations at the class level (Table 1). Collodaria and Acantharia were most abundant, both in absolute (data not shown) and relative concentrations, within subtropical latitudes, while Phaeodaria and Foraminifera increased in proportions in relatively colder ecosystems (Figure 5C). As discussed above, these partitions seem primarily driven by their adaptations to specific iron pools, indicated by the important positive correlations between satellite or biogeochemical proxies and class-specific NBSS intercepts (Figure 5A). The peak in Foraminifera relative proportions seems to coincide with areas of larger predicted size for the group of Rhizarians (Figure 5C, top panel), which is consistent with the size of the vignettes illustrating each Rhizarian class presented in Figure 1. The positive correlation between total iron concentration and their spectral slopes suggest that larger Foraminifera are likely found in response to increased iron availability, possibly mediated by increased phytoplankton productivity (*86*). Hence, both in the case of colonial N_2-fixers and Rhizarians, the effect of temperature gradients on organisms' spectral slopes following the Temperature-Size Rule was masked by the role of iron sources (aerosol and total iron respectively) in flattening the slopes of the largest subgroup (e.g. Puff for colonial N_2-fixers and Foraminifera for Rhizarians).

Foraminifera typically capture phytoplankton prey by extending their pseudopods to increase their encounter rates, or alternatively pair with algal symbionts (dinoflagellates, pelagophytes, or chlorophytes) in oligotrophic ecosystems to acquire photosynthesis products in exchange of host protection and nutritive by-products (*87*). This diversity in

resource acquisition strategy might explain the relatively weak correlation coefficients in our compilation, besides total iron and phosphate concentrations in accordance with other net-based studies (*86*). Like Foraminifera, Phaeodaria appeared directly affected by phosphate concentrations, as well as prey availability and size structure (Figure 5A-B). Phaeodaria are exclusively heterotrophic and tend to accumulate below the epipelagic layer (*44*) and in upwelling regions (Figure 5C), where they significantly attenuate the carbon fluxes by feeding on detrital particles sinking to the ocean's floor, and also contribute to the sinking fluxes as a result of their silica ballast (*88*). In our compilation, the spectral intercepts of Phaeodaria were positively correlated with both the fraction of microphytoplankton and silicate concentration (Figure 5A), suggesting that the source of detrital particles they feed upon was mainly silicified diatoms. Silicate concentrations also affected the other silicified Rhizarians, the Collodaria, albeit the positive correlation was observed with the spectral slopes rather than intercepts. Thus, in the presence of high concentration of silicates, Collodaria tend to be significantly larger, but also less abundant. This finding is consistent with the high Si-requirement of Collodaria, making them poor competitors against smaller, less silicified, diatoms (*89*). Like Collodaria, Acantharia mostly pair with prymnesiophytes (of the genus *Phaeocystis*) to survive the lack of essential nutrients in oligotrophic ecosystems (*90*) and use excess photosynthesis products to fuel their respiration (*91*). As such, they are mostly found in the sunlit surface waters, in conjunction with increased light availability, as indicated by the positive correlation coefficients linking Collodaria and Acantharia NBSS intercepts to daily PAR (Figure 5A). Future efforts should consider other Ocean Color products, like the relative abundance of different phytoplankton groups that could be linked to specific symbiont of these Rhizarians based on their pigment signatures, to further understand and predict the spectral biogeography of Rhizarians in relation to their environment.

3.3 Environmental adaptations of Crustaceans and copepods functional groups

Most of the patterns in Crustaceans' spectral biogeography and its linkages with the environmental factors tested were driven by a single taxon (Figure 3A). Indeed, copepods largely dominated over other Crustacean classes (e.g. Ostracoda and Malacostraca) with regard to predicted absolute and relative abundances, but also with respect to predicted size (Figure 6C). Still, we found important differences in the relative importance of environmental factors driving NBSS parameters (intercept and slope) of the different classes of Crustaceans and different families of copepods assigned to a specific diet (omnivore or carnivore). Amongst these factors, the two variables linked to mesoscale circulation played a key role in determining the size structure of Copepoda, Malacostraca, and Ostracoda (Table S1). Both Malacostraca and Ostracoda were negatively correlated to FSLE, indicating that frontal filaments may be hot-spots for these classes of Crustaceans. In addition, SLA ranked third in explaining the variability of Crustaceans spectral biogeography (Figure 3A), and negative anomalies, indicative of a strong isopycnals uplifting in the epipelagic layer and associated entrainment of deep inorganic nutrients, were significantly correlated to copepods' NBSS intercepts ($r=-0.21$, p-value<0.001). Although this finding was expected, given the long-standing evidence that nutrients brought into the epipelagic actively fuel new production in cyclonic eddies (*92*), this result is notable as it indicates that cyclonic eddies directly benefit higher trophic levels (grazers) on a global scale. Interestingly, we found that both carnivorous ($r=-0.25$, p-value<0.005) and omnivorous ($r=-0.28$, p-value<0.005) copepods were favored in cyclonic eddies, despite their opposite biogeography (Figure 6C). The spatial distribution of copepod trophic groups presented in Figure 6c agree with the recent compilation of copepod species occurrence

(61), as we found a similar dominance of carnivorous copepods, that feed on microzooplankton (e.g. zooplankton larves, ciliates and heterotrophic or mixotrophic dinoflagellates), in oligotrophic ecosystems and a prevalence of omnivorous copepods towards the poles (Figure 6C). Overall, copepods also responded positively to increased fractions of nano- and microphytoplankton (r=0.28, p-value<0.001), associated with elevated concentrations of iron rather than important macronutrients like nitrates (r=0.07, p-value>0.05, Table S1) or phosphates (r=-0.06, p-value>0.05, Table S1).

Elevated nitrate concentrations appeared rather correlated with the two main parameters describing the NBSS estimates of Malacostraca (r(intercept)=0.1 and r(slope)=0.03, p-value<0.001). Together with the concentration of chlorophyll-a, these factors likely drove the larger proportions of this class in upwelling and subpolar ecosystems like the Antarctic where they serve as a keystone group (*93*). The importance of nitrate concentrations in driving patterns in NBSS parameters compared to the satellite proxy of phytoplankton size distribution slopes (r=-0.07, p-value>0.05) is noteworthy and could suggest that prey quality, often estimated by C:N stoichiometry, could prevail over palatability (e.g. size) to explain Malacostraca food selectivity. Caution should be taken when using the phytoplankton size distribution and derived fractions measured by satellite since microphytoplankton only encompass the 20-50 μm size fraction, but feeding experiments conducted in an upwelling system did result in increased productivity of the euphausiid *Euphausia pacifica* when fed on N-rich marine snow particles (*94*). Like copepods with iron, our results could also indicate that Malacostraca play a key role in recycling nitrogen. Indeed, two models linking these mesozooplankton groups to prey stoichiometry (Malacostraca prey being either N-rich marine snow or copepods) have shown that preferential excretion of iron or fixed nitrogen could explain nutrient variability in ecosystems dominated by copepods or krill respectively (*95*, *96*). Malacostraca abundances also increased with silicate concentrations, consistent with their known preference for diatom flocs (*94*).

The lack of response of the class Copepoda to silicate concentrations was puzzling given that diatoms make up most of their gut contents in productive ecosystems (*97*), however we did find a strong linkage between copepods and silicate concentrations after grouping omnivorous families together (r(intercept)=0.32, p-value<0.001). With the exception of silicate concentrations, most factors tested presented similar correlations with either omnivorous or carnivorous copepods, including the fraction microphytoplankton (Figure 6A). Carnivorous copepods hence appeared to benefit from non-siliceous microphytoplankton (r(microphytoplankton)=0.28, p-value<0.001) like mixotrophic dinoflagellates (*98*) that dominate the autotrophic biomass in subtropical gyres (*99*). Compared to other Crustacean groups, observations of Ostracoda were marked by weaker correlation coefficients between spectral parameters and the environmental factors tested (Figure 6A), except for surface temperature (r=0.41, p-value<0.001) and nitrate concentrations (r=-0.23, p-value<0.01). Accordingly, our predictions for this relatively understudied class of opportunistic Crustaceans (*100*) yielded elevated proportions in subtropical gyres, but also at transition zones separating oligotrophic from temperate ecosystems, or subpolar ecosystems from polar ecosystems. The absence of significant correlations with any proxy linked to phytoplankton community and productivity is also consistent with their known versatility with respect to food sources (*101*).

Materials and Methods

1. Global compilation of group-specific size spectrum

To understand and predict the global size distribution of Colonial N_2-fixers, Rhizarians, and Crustaceans in the epipelagic layer, we compiled direct size estimates measured by a range of Underwater Vision Profilers of successive generations (*102*, *103*) and benchtop scanners (e.g. ZooScan, for nets deployed in all major basins within 0 and 200 m. This compilation stems from a collaborative, international effort, termed the Pelagic Size Structure database (*105*), that aims at providing standardized and homogenized size and count information collected by a variety of commercially available plankton imaging instruments. We used the Python workflow (publicly available at https://github.com/jessluo/PSSdb) developed for the first PSSdb release, described in more details in Dugenne et al. (*105*), with little modifications to produce our globally consistent estimates of group-specific Normalized Biovolume Size Spectra (hereafter refer to as NBSS).

Briefly, the general pipeline consists in five automated steps that follow image acquisition and near-real time segmentation (i.e. the process of separating individual particle from background pixels to extract their morphologic features): (1) export of raw size, count and annotation information with the associated metadata from the collaborative platform for image annotation EcoTaxa (https://ecotaxa.obs-vlfr.fr/), (2) standardization of the export tables following a unique format with standard labels and units, (3) quality control of individual datasets that account for particle count uncertainty, sufficient manual validation of the taxonomic classifications (UVP profiles of scanned nets should have 95% validated annotations to pass our quality control), differences in size calibration factors, missing information or incorrect GPS location of each sample, (4) aggregation of samples collected in spatial and temporal proximity to provide accurate measures of particle size distribution at the current resolution of biogeochemical models and other global data compilations (i.e. 1°x1° latitude/longitude and year/month), and (5) the computation of NBSS and derived parameters (slope, intercept, R^2), using pre-defined size classes matching the recent compilation of UVP5 particle size distribution by Kiko et al. (*106*).

The main addition to the Dugenne et al. (*105*) workflow consists of standardizing all taxonomic annotations assigned to individual particles by automated classifiers and later validated by taxonomic experts using the World Register of Marine Species (https://www.marinespecies.org/) as suggested in Neeley et al. (*107*). We wrote a custom script based on existing web scraping functions to automatically search for the taxonomic annotations in WORMS, allowing various formats (e.g. hierarchical or nominal category) and ranks according to the preset classes of automated classifiers, the expertise of human annotators or the research question. The taxonomic look-up table automatically generated by this script is available at https://github.com/jessluo/PSSdb/blob/main/ancillary/plankton_annotated_taxonomy.xlsx. Following this additional standardization, individual annotations were grouped in increasing taxonomic ranks (e.g. phylum, class or subclass, family to assign copepod's trophic strategy based on published compilation (*46*)) and according to their morphotypes when possible (e.g. solitary or colonial for Rhizarians, Puff or Tuff for *Trichodesmium)* to

estimate group-specific NBSS based on biovolume. Biovolume estimates were computed using equivalent circular diameter (ECD) derived from particles area following Eq. 1:

$$\text{Biovolume } (\mu m^3) = 1/6 \times \pi \times ECD(\mu m)^3 = 1/6 \times \pi \times 2\sqrt{\frac{\text{area } (\mu m^2)}{\pi}} \quad (1)$$

The sum of all biovolume estimates within the discrete size classes defined in Eq. 2 is then normalized to the volume analyzed, corresponding to the total volume cumulated over depth (0-200 m) and within each spatio-temporal bin, and to the width of each size class to produce NBSS according to Eq. 3.

$$ECD_i(\mu m) = 2^{i \times 1/3} \text{ with } i=0,1,\ldots,78 \text{ for size classes ranging between 1-53,000,000 } \mu m \quad (2)$$

$$NBSS_i(\mu m^3\ m^{-3}\ \mu m^{-3}) = \Sigma_i \text{Biovolume/volume analyzed/biovolume class width} \quad (3)$$

We then selected the maximum and the last normalized biovolume estimate before observing one empty size class respectively as the lower and upper limits of each size spectrum. This step is necessary to select the effective size range of individual mesoplankton group and to ensure that estimates are not biased by the camera resolution or the sampling gear (e.g. net mesh), or by the limited volume sampled respectively. After selection, unbiased NBSS estimates and associated midpoint of individual biovolume size classes are log_{10}-transformed to estimate spectral slopes, intercepts, and size ranges for each 1°x1° latitude/longitude and year month by linear regression (Figure S1). Intercept estimates were scaled to the minimum observed size of individual mesoplankton groups since the abundance of 1 μm^3 particles, the initial intercept for a spectrum starting at log(1 μm^3)=0, is not representative of the actual abundance of mesoplankton groups. Some of the individual spatio-temporal bins did not contain enough observations to construct a reliable NBSS and fit a log-linear regression as group-specific PSSdb products necessarily included fewer images than the first release. In addition, we selected only datasets that spanned at least 80% of the epipelagic layer (e.g. 160 m) to follow the methodology established by Drago et al. (*44*). As a result, the spatio-temporal coverage in our study is slightly diminished compared to the bulk datasets, constructed with all particles imaged except for methodological artefacts. Overall, group-specific NBSS were computed using 3,068 UVP profiles and 2,411 scans sampled between 2008-2021 and 2004-2022 respectively.

2. Linkages between mesoplankton size structure and environmental covariates

In the absence of broadly concurrent and quality controlled *in situ* hydrographic measurements, we linked mesoplankton NBSS to environmental gradients using global sets of objective analysis, satellite proxies, and biogeochemical models (all accessed and downloaded in Sept. 2023). We used the latest World Ocean Atlas objective analysis (WOA, available at https://www.ncei.noaa.gov/access/world-ocean-atlas-2018) to assign each grid cell (defined by 1°x1° latitude/longitude and month) its associated mean mixed layer depth (m), surface temperature (°C), and surface concentration in phosphate (µM), silicate (µM), and nitrate (µM). In addition, we used WOA surface apparent oxygen utilisation (µM) to represent patterns in dissolved oxygen concentration resulting from

water mass ventilation and age or remineralization rather than the temperature-dependent O_2 solubility.

Given that iron is also considered an essential micronutrient for phytoplankton growth (*74*) but *in situ* measurements are still too scarce to construct a reliable distribution of dissolved iron concentrations in the Global Ocean with enough temporal coverage (*108*), we relied on diverse proxies, satellite measurements distributed by NASA or biogeochemical models provided by Copernicus, of both mineral and total iron compounds respectively. In the surface ocean, most labile compounds present are thought to originate from atmospheric aerosols, delivering mineral dust, ashes formed by volcanic activities and wildfires, or anthropogenic compounds all relatively rich in iron (*109*). These can be approximated from space, as total aerosol optical thickness (unitless variable termed Aerosol optical thickness at 869 nm) and wet dust deposition (variable termed dust wet deposition bin 001 or duwt001, in kg m^{-2} s^{-1}), with a 9km, monthly resolution using specific satellite algorithms whose products are distributed for every year and month by NASA at https://oceandata.sci.gsfc.nasa.gov/directdataaccess/Level-3%20Mapped/Aqua-MODIS and https://goldsmr4.gesdisc.eosdis.nasa.gov/opendap/MERRA2_MONTHLY/M2TMNXADG.5.12.4/ respectively. Alternatively, organic iron-complexing ligands can also represent important sources of iron, especially for organisms synthesizing siderophore transporters, that are poorly characterized by current analytical methods due to the diversity of compounds and reactivity within this pool (*60*). To account for all compounds in the complex pool of dissolved iron, we used surface predictions of total dissolved iron concentration (µM) from the biogeochemical model PISCES (*110*), distributed for any given year month within the period 1993-2022 by Copernicus (product cmems_mod_glo_bgc_my_0.25_P1M-m available at https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_BGC_001_029/) with a resolution of 1/4° latitude/longitude. This model has shown a strong agreement with benchmark sections from the GEOTRACES program compared to other model predictions and also include all known source and sink processes (riverine, sediment, hydrothermal, or dust inputs, remineralization, scavenging) affecting iron pools in marine ecosystems (*108*).

We downloaded monthly average surface photosynthetically available radiation (PAR, in E m^{-2} d^{-1}) and chlorophyll-a concentration (CHL, in mg m^{-3}) for all sampled years (2004-2022) from the same url as the aerosol NASA satellite products. Besides chlorophyll-a concentration, we also included recent estimates of the particle size distribution slope (PSD slope, in particles L^{-1} μm^{-2}) and fractions of pico- (0.2-2 µm), nano- (2-20 µm), and micro- (20-50 µm) phytoplankton retrieved from backscatter coefficients measured by remote sensing at a resolution 4km for every month year to characterize the phytoplankton community representing the main food source for mesozooplankton groups (*71*). Lastly, proxies for mesoscale structure, like sea level anomaly (SLA, in m), that indicate cyclonic and anticyclonic eddies under negative and positive anomaly, and finite size Lyapunov exponent (FSLE, in d^{-1}), that measures the frontal divergence of mesoscale eddies or other currents in close proximity, were accessed on AVISO at https://www.aviso.altimetry.fr/en/data/products with a resolution of 1/25° latitude/longitude. Since mesoscale features are operationally defined as coherent vortex generally trackable for a month or longer with a radius scale of ~100 km (*111*), we consider that the final 1°x1° (~111km) and year month resolution of our datasets were also representative of mesoscale circulation, even at the higher end member of its operational resolution.

All variables were automatically downloaded and averaged to match the resolution of our mesoplankton NBSS estimates (if the resolution was greater than 1°x1° latitude/longitude and year month) using custom scripts available at https://github.com/mdugenne/PSSdb_Learning. Resulting spatial patterns are presented in Figure S2. The environmental datasets were then merged to our group-specific products according to the bin location (1°x1° latitude/longitude) and the sampling time (year month for yearly resolved variables or month for WOA climatologies). Locations with one or more missing environmental variable(s) (e.g. satellite products at high latitude during winter) were filtered out to compute the Spearman correlation coefficients between environmental factors and spectral parameters (slope, intercept, size range) and train the machine learning model used to predict the size distribution of mesoplankton groups globally.

3. Global predictions of mesoplankton size structure using machine learning models

We used our compilations of group-specific NBSS and associated environmental factors in combination with machine learning habitat models to estimate the global size distribution of the three major mesoplankton groups (Colonial N_2-fixers, Rhizarians, Crustaceans) in the epipelagic layer. This modelling effort builds on the recent publication of Drago et al. (*44*), who characterized and predicted the total biomass estimates of several mesozooplankton groups (e.g. Copepods, Rhizarians, not including colonial N_2-fixers) using UVP profiles globally distributed. Since elevated biomass may result from increased abundance and/or size, we expanded their work to help disentangle the effect of environmental factors on mesoplankton abundance (with the maximum abundance representing the spectral intercept) and size (whose continuum is determined by spectral slopes and effective range). Habitat models cover diverse statistical models, including the relatively recent boosted regression trees (also referred as xgboost, (*112*)), that link environmental factors to species or group distribution with the objective to predict their occurrence in unsampled locations. Like other machine learning models, xgboost learns simple or complex rules (here a hierarchy of yes/no decisions delimited by the maximum tree depth) linking the response variable (NBSS parameters) to explanatory variables (environmental factors described in Material and Methods section 2) using a subset of the dataset (here 80% random split), named the training set. The learning process is an iterative process based on the progressive evaluation of a loss function, defined as the root mean square error between observations and predictions.

Their performance is then assessed using the remaining subset (here 20% random split), termed the test set. The algorithm performance evaluation is based on different metrics like the coefficient of determination (R^2), determined by the sum of squared errors between test predictions and test set, or the log-likelihood of the model, that represents the goodness of fit between model predictions and observations according to the data distribution. We used two parameters linked to the model log-likelihood (noted l), the Akaike (AIC) and Bayesian (BIC) Information Criterion defined in Eq. 4, to compare model predictions amongst increasing taxonomic ranks. While we expect that grouping observations based on high taxonomic rank like class or families should be more appropriate to capture the variability in mesoplankton NBSS on a global scale as a result of diverging environmental adaptations (*33*), we also considered that the decrease of available observations for these ranks or the differences in model complexity could also result in

overall lower performances of the model. Since AIC and BIC take into account model likelihood, complexity, and the number of available observations, we used these estimates as a mean to provide objective comparisons between model performances. The likelihood of taxon-specific models were averaged, with a weight corresponding to the proportion of observations, across all subgroups of colonial N_2-fixers, Rhizarians and Crustaceans when increased taxonomic ranks were compared to the overall model performed at the phylum level, following a similar approach as mixture models (*113*).

$$AIC = 2k - 2l \text{ and } BIC = k\ln(n) - 2l, \quad (4)$$

with k the number of model parameters, n the number of observations, and *l* the model log-likelihood directly returned by the xgboost package.

Unlike more traditional decision tree models that predict a set of discrete values termed the leaves, xgboost models may apply to continuous variables by producing numerous leaves associated with a specific score, that are ultimately summed up across all trees to create a smooth gradient in the predicted response variable (i.e the "boosting" process). Instead of discontinuous response, the resulting predictions will thus present a smoother data distribution similar to that of spectral parameters (Figure S1). Since these models are based on direct representation of the linkages between explanatory and response variables, estimates of feature importance, measured as the average gain in model performance (log-likelihood) across all splits after the variable was used, are relatively straightforward. Such models have additional advantages as they can deal with many dependent variables (for example, dust and aerosol proxies for iron deposition, or temperature and PAR) and predict multivariate outputs. In our study, we simultaneously predicted spectral slopes, intercepts and size range while accounting for their correlation (Figure S1). The reconstructed NBSS, obtained by assuming a log-linear decline with the modeled parameters, can then be integrated to predict the total abundances of mesoplankton groups at any given space and time. Prior to model training, both the spectral intercept and size range were $\log_{10}$-transformed to ensure that predictions will be positive after reverse transformation. To minimize the effect of irrelevant explanatory variables and avoid model overfitting, several parameters were tuned, including score shrinkage (eta parameter tested between 0.1 and 0.5) to homogenize the gradient boosting process in case of overfitting, the maximum number of splits in the decision trees (max_depth parameter tested between 3 and 10) that can lead to oversimplification/complexification of the model, or the number of trees (n_estimators parameter tested between 20 and 100). The best combination of model parameters was determined objectively using the grid search function implemented in the xgboost package.

Not all mesoplankton groups were abundant enough or significantly correlated with the environmental factors tested to analyze and predict their size structure in the epipelagic layer using a machine learning model (*44*). Overall, we could accurately predict the spectral biogeography of one family of colonial N_2-fixers (Oscillatoriales), four classes of Rhizarians (Acantharia, Collodaria, Phaeodaria, Foraminifera) and three classes of Crustaceans (Copepoda, Malacostraca, Ostracoda). The family of colonial N_2-fixers was represented exclusively by *Trichodesmium* in this study although we note the presence of a few Nostocales in the Baltic Sea that were too rare for reliable predictions of their size spectrum (data not shown). Some groups were further distinguished based on their morphology (e.g. *Trichodesmium* Puff or Tuff), or into functional groups (e.g. carnivorous

or omnivorous copepods) based on published correspondence between copepod's families and trophic groups (*46*). The most abundant class of Rhizaria, Collodaria, were almost exclusively found as solitary forms in the epipelagic, so we could not explore possible niche partitioning amongst colonial and solitary Rhizarians. Predictions were generated with the optimal set of parameters in every 1°x1° grid cell for the full temporal coverage of the UVP and scanners datasets. Temporal correlations with the multiple iron proxies (total iron, dust or aerosol) were estimated in each spatial bin to determine which proxy had the strongest link with NBSS integral predictions, that result from predicted slopes, intercepts and size ranges (Figure S3). The main iron proxy was then defined based on the maximum correlation across the 3 possible correlation coefficients in broad ecosystem categories (e.g. Polar, Temperate-Subpolar, Subtropical Gyres, Equatorial-Upwellings) corresponding to adjacent Longhurst provinces (https://www.marineregions.org/sources.php#longhurst), and in diverse ocean basins (https://www.marineregions.org/sources.php#goas). For a full description of the xgboost python package used in this study, see documentation provided at https://xgboost.readthedocs.io/en/stable/python. The code to train, check the model performance and feature importance, and predict mesoplankton size distribution based on xgboost functions is available at https://github.com/mdugenne/PSSdb_Learning.

References

1. R. W. Sheldon, A. Prakash, W. H. Sutcliffe Jr., The size distribution of particles in the Ocean. *Limnology and Oceanography* **17**, 327–340 (1972).

2. T. Platt, K. Denman, Organisation in the pelagic ecosystem. *Helgolander Wiss. Meeresunters* **30**, 575–581 (1977).

3. K. H. Andersen, T. Berge, R. J. Gonçalves, M. Hartvig, J. Heuschele, S. Hylander, N. S. Jacobsen, C. Lindemann, E. A. Martens, A. B. Neuheimer, K. Olsson, A. Palacz, A. E. F. Prowe, J. Sainmont, S. J. Traving, A. W. Visser, N. Wadhwa, T. Kiørboe, Characteristic Sizes of Life in the Oceans, from Bacteria to Whales. *Annual Review of Marine Science* **8**, 217–241 (2016).

4. K. H. Andersen, A. W. Visser, From cell size and first principles to structure and function of unicellular plankton communities. *Progress in Oceanography* **213**, 102995 (2023).

5. A. D. Barton, A. J. Pershing, E. Litchman, N. R. Record, K. F. Edwards, Z. V. Finkel, T. Kiørboe, B. A. Ward, The biogeography of marine plankton traits. *Ecology Letters* **16**, 522–534 (2013).

6. I. A. Hatton, R. F. Heneghan, Y. M. Bar-On, E. D. Galbraith, The global ocean size spectrum from bacteria to whales. *Science Advances* **7**, eabh3732 (2021).

7. E. Maranón, P. Cermenó, J. Rodríguez, M. V. Zubkov, R. P. Harris, Scaling of phytoplankton photosynthesis and cell size in the ocean. *Limnology and Oceanography* **52**, 2190–2198 (2007).

8. T. Kiørboe, A. G. Hirst, Shifts in Mass Scaling of Respiration, Feeding, and Growth Rates across Life-Form Transitions in Marine Pelagic Organisms. *The American Naturalist* **183**, E118–E130 (2014).

9. B. Hansen, P. K. Bjornsen, P. J. Hansen, The size ratio between planktonic predators and their prey. *Limnology and Oceanography* **39**, 395–403 (1994).

10. K. Stamieszkin, A. J. Pershing, N. R. Record, C. H. Pilskaln, H. G. Dam, L. R. Feinberg, Size as the master trait in modeled copepod fecal pellet carbon flux. *Limnology and Oceanography* **60**, 2090–2107 (2015).

11. S. Uye, K. Kaname, Relations between fecal pellet volume and body size for major zooplankters of the Inland Sea of Japan. *J Oceanogr* **50**, 43–49 (1994).

12. J. A. Chindia, C. C. Figueredo, Phytoplankton settling depends on cell morphological traits, but what is the best predictor? *Hydrobiologia* **813**, 51–61 (2018).

13. S. Henson, F. Le Moigne, S. Giering, Drivers of Carbon Export Efficiency in the Global Ocean. *Global Biogeochemical Cycles* **33**, 891–903 (2019).

14. C. Serra-Pompei, B. A. Ward, J. Pinti, A. W. Visser, T. Kiørboe, K. H. Andersen, Linking Plankton Size Spectra and Community Composition to Carbon Export and Its Efficiency. *Global Biogeochemical Cycles* **36**, e2021GB007275 (2022).

15. R. W. Sheldon, W. H. Sutcliffe Jr., M. A. Paranjape, Structure of Pelagic Food Chain and Relationship Between Plankton and Fish Production. *J. Fish. Res. Bd. Can.* **34**, 2344–2353 (1977).

16. R. R. Rykaczewski, D. M. Checkley, Influence of ocean winds on the pelagic ecosystem in upwelling regions. *Proceedings of the National Academy of Sciences* **105**, 1965–1970 (2008).

17. R. F. Heneghan, I. A. Hatton, E. D. Galbraith, Climate change impacts on marine ecosystems through the lens of the size spectrum. *Emerging Topics in Life Sciences* **3**, 233–243 (2019).

18. M. Beck, C. Cailleton, L. Guidi, C. Desnos, L. Jalabert, A. Elineau, L. Stemmann, S.-D. Ayata, J.-O. Irisson, Morphological diversity increases with decreasing resources along a zooplankton time series. *Proceedings of the Royal Society B: Biological Sciences* **290**, 20232109 (2023).

19. L. Vilgrain, F. Maps, M. Picheral, M. Babin, C. Aubry, J.-O. Irisson, S.-D. Ayata, Trait-based approach using in situ copepod images reveals contrasting ecological patterns across an Arctic ice melt zone. *Limnology and Oceanography* **66**, 1155–1167 (2021).

20. T. Kiørboe, A. Visser, K. H. Andersen, A trait-based approach to ocean ecology. *ICES Journal of Marine Science* **75**, 1849–1863 (2018).

21. E. Litchman, M. D. Ohman, T. Kiørboe, Trait-based approaches to zooplankton communities. *Journal of Plankton Research* **35**, 473–484 (2013).

22. M. D. Ohman, J.-B. Romagnan, Nonlinear effects of body size and optical attenuation on Diel Vertical Migration by zooplankton. *Limnology and Oceanography* **61**, 765–770 (2016).

23. E. Litchman, C. A. Klausmeier, O. M. Schofield, P. G. Falkowski, The role of functional traits and trade-offs in structuring phytoplankton communities: scaling from cellular to ecosystem level. *Ecology Letters* **10**, 1170–1181 (2007).

24. S. Martini, F. Larras, A. Boyé, E. Faure, N. Aberle, P. Archambault, L. Bacouillard, B. E. Beisner, L. Bittner, E. Castella, M. Danger, O. Gauthier, L. Karp-Boss, F. Lombard, F. Maps, L. Stemmann, E. Thiébaut, P. Usseglio-Polatera, M. Vogt, M. Laviale, S.-D. Ayata, Functional trait-based approaches as a common framework for aquatic ecologists. *Limnology and Oceanography* **66**, 965–994 (2021).

25. E. C. Orenstein, S.-D. Ayata, F. Maps, É. C. Becker, F. Benedetti, T. Biard, T. de Garidel-Thoron, J. S. Ellen, F. Ferrario, S. L. C. Giering, T. Guy-Haim, L. Hoebeke, M. H. Iversen, T. Kiørboe, J.-F. Lalonde, A. Lana, M. Laviale, F. Lombard, T. Lorimer, S. Martini, A. Meyer, K. O. Möller, B. Niehoff, M. D. Ohman, C. Pradalier, J.-B. Romagnan, S.-M. Schröder, V. Sonnet, H. M. Sosik, L. S. Stemmann, M. Stock, T. Terbiyik-Kurt, N. Valcárcel-Pérez, L. Vilgrain, G. Wacquet, A. M. Waite, J.-O. Irisson, Machine learning techniques to characterize functional traits of plankton from image data. *Limnology and Oceanography* **67**, 1647–1669 (2022).

26. F. Lombard, E. Boss, A. M. Waite, M. Vogt, J. Uitz, L. Stemmann, H. M. Sosik, J. Schulz, J.-B. Romagnan, M. Picheral, J. Pearlman, M. D. Ohman, B. Niehoff, K. O. Möller, P. Miloslavich, A. Lara-Lpez, R. Kudela, R. M. Lopes, R. Kiko, L. Karp-Boss, J. S. Jaffe, M. H. Iversen, J.-O. Irisson, K. Fennel, H. Hauss, L. Guidi, G. Gorsky, S. L. C. Giering, P. Gaube, S. Gallager, G. Dubelaar, R. K. Cowen, F. Carlotti, C. Briseño-Avena, L. Berline, K. Benoit-Bird, N. Bax, S. Batten, S. D. Ayata, L. F. Artigas, W. Appeltans, Globally consistent quantitative observations of planktonic ecosystems. *Frontiers in Marine Science* **6** (2019).

27. C. Violle, M.-L. Navas, D. Vile, E. Kazakou, C. Fortunel, I. Hummel, E. Garnier, Let the concept of trait be functional! *Oikos* **116**, 882–892 (2007).

28. F. Benedetti, M. Vogt, D. Righetti, F. Guilhaumon, S.-D. Ayata, Do functional groups of planktonic copepods differ in their ecological niches? *Journal of Biogeography* **45**, 604–616 (2018).

29. J. D. Everett, R. F. Heneghan, J. L. Blanchard, I. M. Suthers, Pakhomov, Evgeny, Self-organisation of zooplankton communities produces similar food chain lengths throughout the ocean (2022). https://doi.org/10.21203/rs.3.rs-1186379/v1.

30. E. Faure, F. Not, A.-S. Benoiston, K. Labadie, L. Bittner, S.-D. Ayata, Mixotrophic protists display contrasted biogeographies in the global ocean. *ISME J* **13**, 1072–1083 (2019).

31. J. P. Zehr, D. G. Capone, Changing perspectives in marine nitrogen fixation. *Science* **368**, eaay9514 (2020).

32. M. C. Brandão, F. Benedetti, S. Martini, Y. D. Soviadan, J.-O. Irisson, J.-B. Romagnan, A. Elineau, C. Desnos, L. Jalabert, A. S. Freire, M. Picheral, L. Guidi, G. Gorsky, C. Bowler, L. Karp-Boss, N. Henry, C. de Vargas, M. B. Sullivan, L. Stemmann, F. Lombard, Macroscale patterns of oceanic zooplankton composition and size structure. *Sci Rep* **11**, 15714 (2021).

33. P. Brun, M. R. Payne, T. Kiørboe, Trait biogeography of marine copepods – an analysis across scales. *Ecology Letters* **19**, 1403–1413 (2016).

34. J. J. Pierella Karlusich, E. Pelletier, F. Lombard, M. Carsique, E. Dvorak, S. Colin, M. Picheral, F. M. Cornejo-Castillo, S. G. Acinas, R. Pepperkok, E. Karsenti, C. de Vargas, P. Wincker, C. Bowler, R. A. Foster, Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. *Nat Commun* **12**, 4160 (2021).

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

36. D. Atkinson, R. M. Sibly, Why are organisms usually bigger in colder environments? Making sense of a life history puzzle. *Trends in Ecology & Evolution* **12**, 235–239 (1997).

37. R. W. Eppley, Temperature and phytoplankton growth in the sea}. *Fishery Bulletin* **70**, 1063--1085 (1972).

38. A. G. Hirst, A. J. Bunker, Growth of marine planktonic copepods: Global rates and patterns in relation to chlorophyll a, temperature, and body weight. *Limnology and Oceanography* **48**, 1988–2010 (2003).

39. T. Biard, L. Stemmann, M. Picheral, N. Mayot, P. Vandromme, H. Hauss, G. Gorsky, L. Guidi, R. Kiko, F. Not, In situ imaging reveals the biomass of giant protists in the global ocean. *Nature* **532**, 504–507 (2016).

40. M. Eichner, K. Inomura, J. J. P. Karlusich, Y. Shaked, Better together? Lessons on sociality from Trichodesmium. *Trends in Microbiology* **31**, 1072–1084 (2023).

41. A. M. Hynes, "Diversity of the marine cyanobacterium Trichodesmium : characterization of the Woods Hole culture collection and quantification of field populations," thesis, Massachusetts Institute of Technology and Woods Hole Oceanographic Institution (2009).

42. J. McN. Sieburth, V. Smetacek, J. Lenz, Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions 1. *Limnology and Oceanography* **23**, 1256–1263 (1978).

43. T. Panaïotis, M. Babin, T. Biard, F. Carlotti, L. Coppola, L. Guidi, H. Hauss, L. Karp-Boss, R. Kiko, F. Lombard, A. M. P. McDonnell, M. Picheral, A. Rogge, A. M. Waite, L. Stemmann, J.-O. Irisson, Three major mesoplanktonic communities resolved by in situ imaging in the upper 500 m of the global ocean. *Global Ecology and Biogeography* **n/a**.

44. L. Drago, T. Panaïotis, J.-O. Irisson, M. Babin, T. Biard, F. Carlotti, L. Coppola, L. Guidi, H. Hauss, L. Karp-Boss, F. Lombard, A. M. P. McDonnell, M. Picheral, A. Rogge, A. M. Waite, L. Stemmann, R. Kiko, Global

Distribution of Zooplankton Biomass Estimated by In Situ Imaging and Machine Learning. *Frontiers in Marine Science* **9** (2022).

45. M. Geoffroy, T. Langbehn, P. Priou, Ø. Varpe, G. Johnsen, A. Le Bris, J. A. D. Fisher, M. Daase, D. McKee, J. Cohen, J. Berge, Pelagic organisms avoid white, blue, and red artificial light from scientific instruments. *Sci Rep* **11**, 14941 (2021).

46. F. Benedetti, J. Wydler, M. Vogt, Copepod functional traits and groups show divergent biogeographies in the global ocean. *Journal of Biogeography* **50**, 8–22 (2023).

47. A. E. White, K. S. Watkins-Brandt, M. J. Church, Temporal Variability of Trichodesmium spp. and Diatom-Diazotroph Assemblages in the North Pacific Subtropical Gyre. *Frontiers in Marine Science* **5** (2018).

48. E. M. Olson, D. J. McGillicuddy, G. R. Flierl, C. S. Davis, S. T. Dyhrman, J. B. Waterbury, Mesoscale eddies and Trichodesmium spp. distributions in the southwestern North Atlantic. *Journal of Geophysical Research: Oceans* **120**, 4129–4150 (2015).

49. E. M. Olson, D. J. McGillicuddy, S. T. Dyhrman, J. B. Waterbury, C. S. Davis, A. R. Solow, The depth-distribution of nitrogen fixation by Trichodesmium spp. colonies in the tropical–subtropical North Atlantic. *Deep Sea Research Part I: Oceanographic Research Papers* **104**, 72–91 (2015).

50. P. Brun, M. Vogt, M. R. Payne, N. Gruber, C. J. O'Brien, E. T. Buitenhuis, C. Le Quéré, K. Leblanc, Y.-W. Luo, Ecological niches of open ocean phytoplankton taxa. *Limnology and Oceanography* **60**, 1020–1038 (2015).

51. B. A. Seibel, J. L. Schneider, S. Kaartvedt, K. F. Wishner, K. L. Daly, Hypoxia Tolerance and Metabolic Suppression in Oxygen Minimum Zone Euphausiids: Implications for Ocean Deoxygenation and Biogeochemical Cycles. *Integrative and Comparative Biology* **56**, 510–523 (2016).

52. R. Kiko, H. Hauss, F. Buchholz, F. Melzner, Ammonium excretion and oxygen respiration of tropical copepods and euphausiids exposed to oxygen minimum zone conditions. *Biogeosciences* **13**, 2241–2255 (2016).

53. J. R. Gallon, The oxygen sensitivity of nitrogenase: a problem for biochemists and micro-organisms. *Trends in Biochemical Sciences* **6**, 19–23 (1981).

54. A. Regaudie-de-Gioux, C. M. Duarte, Temperature dependence of planktonic metabolism in the ocean. *Global Biogeochemical Cycles* **26** (2012).

55. M. Staal, F. J. R. Meysman, L. J. Stal, Temperature excludes N2-fixing heterocystous cyanobacteria in the tropical oceans. *Nature* **425**, 504–507 (2003).

56. L. J. Stal, Is the distribution of nitrogen-fixing cyanobacteria in the oceans related to temperature? *Environmental Microbiology* **11**, 1632–1645 (2009).

57. K. Inomura, S. T. Wilson, C. Deutsch, Mechanistic Model for the Coexistence of Nitrogen Fixation and Photosynthesis in Marine Trichodesmium. *mSystems* **4**, e00210-19 (2019).

58. F. Lombard, J. Erez, E. Michel, L. Labeyrie, Temperature effect on respiration and photosynthesis of the symbiont-bearing planktonic foraminifera Globigerinoides ruber, Orbulina universa, and Globigerinella siphonifera. *Limnology and Oceanography* **54**, 210–218 (2009).

59. E. Villar, V. Dani, E. Bigeard, T. Linhart, M. Mendez-Sandin, C. Bachy, C. Six, F. Lombard, C. Sabourault, F. Not, Symbiont Chloroplasts Remain Active During Bleaching-Like Response Induced by Thermal Stress in Collozoum pelagicum (Collodaria, Retaria). *Frontiers in Marine Science* **5** (2018).

60. M. Gledhill, K. Buck, The Organic Complexation of Iron in the Marine Environment: A Review. *Frontiers in Microbiology* **3** (2012).

61. A. Tovar-Sanchez, S. A. Sañudo-Wilhelmy, A. B. Kustka, S. Agustí, J. Dachs, D. A. Hutchins, D. G. Capone, C. M. Duarte, Effects of Dust Deposition and River Discharges on Trace Metal Composition of Trichodesmium

spp. in the Tropical and Subtropical North Atlantic Ocean. *Limnology and Oceanography* **51**, 1755–1761 (2006).

62. A. Fernández, B. Mouriño-Carballido, A. Bode, M. Varela, E. Marañón, Latitudinal distribution of *Trichodesmium* spp. and N_2 fixation in the Atlantic Ocean. *Biogeosciences* **7**, 3167–3176 (2010).

63. R. J. Langlois, M. M. Mills, C. Ridame, P. Croot, J. LaRoche, Diazotrophic bacteria respond to Saharan dust additions. *Marine Ecology Progress Series* **470**, 1–14 (2012).

64. P. Pinedo-Gonzalez, N. J. Hawco, R. M. Bundy, E. V. Armbrust, M. Follows, B. B. Cael, A. E. White, S. Ferron, D. M. Karl, S. G. John, Anthropogenic Asian aerosols provide Fe to the North Pacific Ocean | PNAS. https://www.pnas.org/doi/full/10.1073/pnas.2010315117.

65. J. Weis, C. Schallenberg, Z. Chase, A. R. Bowie, B. Wojtasiewicz, M. M. G. Perron, M. D. Mallet, P. G. Strutton, Southern Ocean Phytoplankton Stimulated by Wildfire Emissions and Sustained by Iron Recycling. *Geophysical Research Letters* **49**, e2021GL097538 (2022).

66. J.-E. Yoon, D. King, J. Longman, S. J. Cronin, Differential response of chlorophyll-a concentrations to explosive volcanism in the western South Pacific. *Frontiers in Marine Science* **10** (2023).

67. M. Li, F. Shen, X. Sun, 2019–2020 Australian bushfire air particulate pollution and impact on the South Pacific Ocean. *Sci Rep* **11**, 12288 (2021).

68. C. Guieu, S. Bonnet, A. Petrenko, C. Menkes, V. Chavagnac, K. Desboeufs, C. Maes, T. Moutin, Iron from a submarine source impacts the productive layer of the Western Tropical South Pacific (WTSP). *Sci Rep* **8**, 9075 (2018).

69. S. Bonnet, C. Guieu, V. Taillandier, C. Boulart, P. Bouruet-Aubertot, F. Gazeau, C. Scalabrin, M. Bressac, A. N. Knapp, Y. Cuypers, D. González-Santana, H. J. Forrer, J.-M. Grisoni, O. Grosso, J. Habasque, M. Jardin-Camps, N. Leblond, F. A. C. Le Moigne, A. Lebourges-Dhaussy, C. Lory, S. Nunige, E. Pulido-Villena, A. L. Rizzo, G. Sarthou, C. Tilliette, Natural iron fertilization by shallow hydrothermal sources fuels diazotroph blooms in the ocean. *Science* **380**, 812–817 (2023).

70. Z. Mériguet, M. Vilain, A. Baudena, C. Tilliette, J. Habasque, A. Lebourges-Dhaussy, N. Bhairy, C. Guieu, S. Bonnet, F. Lombard, Plankton community structure in response to hydrothermal iron inputs along the Tonga-Kermadec arc. *Front. Mar. Sci.* **10**, 1232923 (2023).

71. T. S. Kostadinov, L. Robertson Lain, C. E. Kong, X. Zhang, S. Maritorena, S. Bernard, H. Loisel, D. S. F. Jorge, E. Kochetkova, S. Roy, B. Jonsson, V. Martinez-Vicente, S. Sathyendranath, Ocean color algorithm for the retrieval of the particle size distribution and carbon-based phytoplankton size classes using a two-component coated-sphere backscattering model. *Ocean Science* **19**, 703–727 (2023).

72. M. J. Behrenfeld, Abandoning Sverdrup's Critical Depth Hypothesis on phytoplankton blooms. *Ecology* **91**, 977–989 (2010).

73. L.-C. Wang, Y. Jia-Yuh, Dynamics of Upwelling Annual Cycle in the Equatorial Pacific Ocean. *Sustainability* **11**, 5038 (2019).

74. C. M. Moore, M. M. Mills, K. R. Arrigo, I. Berman-Frank, L. Bopp, P. W. Boyd, E. D. Galbraith, R. J. Geider, C. Guieu, S. L. Jaccard, T. D. Jickells, J. La Roche, T. M. Lenton, N. M. Mahowald, E. Marañón, I. Marinov, J. K. Moore, T. Nakatsuka, A. Oschlies, M. A. Saito, T. F. Thingstad, A. Tsuda, O. Ulloa, Processes and patterns of oceanic nutrient limitation. *Nature Geosci* **6**, 701–710 (2013).

75. A. Tagliabue, T. Mtshali, O. Aumont, A. R. Bowie, M. B. Klunder, A. N. Roychoudhury, S. Swart, A global compilation of dissolved iron measurements: focus on distributions and processes in the Southern Ocean. *Biogeosciences* **9**, 2333–2349 (2012).

76. L. M. Laglera, A. Tovar-Sánchez, M. H. Iversen, H. E. González, H. Naik, G. Mangesh, P. Assmy, C. Klaas, M. G. Mazzocchi, M. Montresor, S. W. A. Naqvi, V. Smetacek, D. A. Wolf-Gladrow, Iron partitioning during

LOHAFEX: Copepod grazing as a major driver for iron recycling in the Southern Ocean. *Marine Chemistry* **196**, 148–161 (2017).

77. S. Winkelbauer, M. Mayer, V. Seitner, E. Zsoter, H. Zuo, L. Haimberger, Diagnostic evaluation of river discharge into the Arctic Ocean and its impact on oceanic volume transports. *Hydrology and Earth System Sciences* **26**, 279–304 (2022).

78. A. Tsuda, S. Takeda, H. Saito, J. Nishioka, I. Kudo, Y. Nojiri, K. Suzuki, M. Uematsu, M. L. Wells, D. Tsumune, T. Yoshimura, T. Aono, T. Aramaki, W. P. Cochlan, M. Hayakawa, K. Imai, T. Isada, Y. Iwamoto, W. K. Johnson, S. Kameyama, S. Kato, H. Kiyosawa, Y. Kondo, M. Levasseur, R. J. Machida, I. Nagao, F. Nakagawa, T. Nakanishi, S. Nakatsuka, A. Narita, Y. Noiri, H. Obata, H. Ogawa, K. Oguma, T. Ono, T. Sakuragi, M. Sasakawa, M. Sato, A. Shimamoto, H. Takata, C. G. Trick, Y. W. Watanabe, C. S. Wong, N. Yoshie, Evidence for the grazing hypothesis: Grazing reduces phytoplankton responses of the HNLC ecosystem to iron enrichment in the western subarctic pacific (SEEDS II). *J Oceanogr* **63**, 983–994 (2007).

79. Y. Huang, A. Tagliabue, N. Cassar, Data-Driven Modeling of Dissolved Iron in the Global Ocean. *Frontiers in Marine Science* **9** (2022).

80. S. Basu, M. Gledhill, D. de Beer, S. G. Prabhu Matondkar, Y. Shaked, Colonies of marine cyanobacteria Trichodesmium interact with associated bacteria to acquire iron from dust. *Commun Biol* **2**, 1–8 (2019).

81. S. Basu, Y. Shaked, Mineral iron utilization by natural and cultured Trichodesmium and associated bacteria. *Limnology and Oceanography* **63**, 2307–2320 (2018).

82. U. Pfreundt, J. Słomka, G. Schneider, A. Sengupta, F. Carrara, V. Fernandez, M. Ackermann, R. Stocker, Controlled motility in the cyanobacterium Trichodesmium regulates aggregate architecture. *Science* **380**, 830–835 (2023).

83. B. Barone, A. Coenen, S. Beckett, D. McGillicuddy, J. Weitz, D. Karl, The ecological and biogeochemical state of the North Pacifi c Subtropical Gyre is linked to sea surface height. *Journal of Marine Research* **77** (2019).

84. T. A. Villareal, E. J. Carpenter, Buoyancy Regulation and the Potential for Vertical Migration in the Oceanic Cyanobacterium Trichodesmium. *Microbial Ecology* **45**, 1–10 (2003).

85. M. Dugenne, M. R. Gradoville, M. J. Church, S. T. Wilson, U. Sheyn, M. J. Harke, K. M. Björkman, N. J. Hawco, A. M. Hynes, F. Ribalet, D. M. Karl, E. F. DeLong, S. T. Dyhrman, E. V. Armbrust, S. John, J. M. Eppley, K. Harding, B. Stewart, A. M. Cabello, K. A. Turk-Kubo, M. Caffin, A. E. White, J. P. Zehr, Nitrogen Fixation in Mesoscale Eddies of the North Pacific Subtropical Gyre: Patterns and Mechanisms. *Global Biogeochemical Cycles* **37**, e2022GB007386 (2023).

86. A. J. Arnold, W. C. Parker, "Biogeography of planktonic Foraminifera" in *Modern Foraminifera*, B. K. Sen Gupta, Ed. (Springer Netherlands, Dordrecht, 2003; https://doi.org/10.1007/0-306-48104-9_7), pp. 103–122.

87. B. K. Sen Gupta, *Modern Foraminifera* (Springer Netherlands, Dordrecht, 2003; http://link.springer.com/10.1007/0-306-48104-9).

88. M. R. Stukel, T. Biard, J. Krause, M. D. Ohman, Large Phaeodaria in the twilight zone: Their role in the carbon cycle. *Limnology and Oceanography* **63**, 2579–2594 (2018).

89. N. Llopis Monferrer, D. Boltovskoy, P. Tréguer, M. M. Sandin, F. Not, A. Leynaert, Estimating Biogenic Silica Production of Rhizaria in the Global Ocean. *Global Biogeochemical Cycles* **34**, e2019GB006286 (2020).

90. J. Decelle, S. Colin, R. A. Foster, "Photosymbiosis in Marine Planktonic Protists" in *Marine Protists: Diversity and Dynamics*, S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, F. Not, Eds. (Springer Japan, Tokyo, 2015; https://doi.org/10.1007/978-4-431-55130-0_19), pp. 465–500.

91. J. Mansour, "Mixotrophy of photosymbiotic Radiolaria," thesis, Sorbonne université (2021).

92. D. J. McGillicuddy, Mechanisms of Physical-Biological-Biogeochemical Interaction at the Oceanic Mesoscale. *Annual Review of Marine Science* **8**, 125–159 (2016).

93. T. Antezana, Euphausia mucronata: A keystone herbivore and prey of the Humboldt Current System. *Deep Sea Research Part II: Topical Studies in Oceanography* **57**, 652–662 (2010).

94. L. Dilling, J. Wilson, D. Steinberg, A. Alldredge, Feeding by the euphausiid Euphausia pacifica and the copepod Calanus pacificus on marine snow. *Marine Ecology Progress Series* **170**, 189–201 (1998).

95. J. Priddle, M. J. Whitehouse, P. Ward, R. S. Shreeve, A. S. Brierley, A. Atkinson, J. L. Watkins, M. A. Brandon, G. C. Cripps, Biogeochemistry of a Southern Ocean plankton ecosystem: Using natural variability in community composition to study the role of metazooplankton in carbon and nitrogen cycles - Priddle - 2003 - Journal of Geophysical Research: Oceans - Wiley Online Library. *Journal of Geophysical Research* **108**, 8082 (2003).

96. C. Richon, O. Aumont, A. Tagliabue, Prey Stoichiometry Drives Iron Recycling by Zooplankton in the Global Ocean. *Frontiers in Marine Science* **7** (2020).

97. A. F. Pasternak, S. B. Schnack-Schiel, "Feeding patterns of dominant Antarctic copepods: an interplay of diapause, selectivity, and availability of food" in *Copepoda: Developments in Ecology, Biology and Systematics: Proceedings of the Seventh International Conference on Copepoda, Held in Curitiba, Brazil, 25–31 July 1999*, R. M. Lopes, J. W. Reid, C. E. F. Rocha, Eds. (Springer Netherlands, Dordrecht, 2001; https://doi.org/10.1007/0-306-47537-5_3)*Developments in Hydrobiology*, pp. 25–36.

98. E. Saiz, A. Calbet, Copepod feeding in the ocean: scaling patterns, composition of their diet and the bias of estimates due to microzooplankton grazing during incubations. *Hydrobiologia* **666**, 181–196 (2011).

99. A. L. Pasulka, M. R. Landry, D. A. A. Taniguchi, A. G. Taylor, M. J. Church, Temporal dynamics of phytoplankton and heterotrophic protists at station ALOHA. *Deep Sea Research Part II: Topical Studies in Oceanography* **93**, 44–57 (2013).

100. F. Mesquita-Joanes, A. J. Smith, F. A. Viehberg, "Chapter 2 - The Ecology of Ostracoda Across Levels of Biological Organisation from Individual to Ecosystem: A Review of Recent Developments and Future Potential" in *Developments in Quaternary Sciences*, D. J. Horne, J. A. Holmes, J. Rodriguez-Lazaro, F. A. Viehberg, Eds. (Elsevier, 2012; https://www.sciencedirect.com/science/article/pii/B9780444536365000020)vol. 17 of *Ostracoda as Proxies for Quaternary Climate Change*, pp. 15–35.

101. J. Vannier, K. Abe, K. Ikuta, Feeding in myodocopid ostracods: functional morphology and laboratory observations from videos | Marine Biology. *Marine biology* **132**, 391–408 (1998).

102. M. Picheral, L. Guidi, L. Stemmann, D. M. Karl, G. Iddaoud, G. Gorsky, The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnology and Oceanography: Methods* **8**, 462–473 (2010).

103. M. Picheral, C. Catalano, D. Brousseau, H. Claustre, L. Coppola, E. Leymarie, J. Coindat, F. Dias, S. Fevre, L. Guidi, J. O. Irisson, L. Legendre, F. Lombard, L. Mortier, C. Penkerch, A. Rogge, C. Schmechtig, S. Thibault, T. Tixier, A. Waite, L. Stemmann, The Underwater Vision Profiler 6: an imaging sensor of particle size spectra and plankton, for autonomous and cabled platforms. *Limnology and Oceanography: Methods* **20**, 115–129 (2022).

104. G. Gorsky, M. D. Ohman, M. Picheral, S. Gasparini, L. Stemmann, J.-B. Romagnan, A. Cawood, S. Pesant, C. García-Comas, F. Prejger, Digital zooplankton image analysis using the ZooScan integrated system. *Journal of Plankton Research* **32**, 285–303 (2010).

105. M. Dugenne, M. Corrales-Ugalde, J. Luo, R. Kiko, T. O'Brien, J.-O. Irisson, F. Lombard, L. Stemmann, C. Stock, C. R. Anderson, M. Babin, N. Bhairy, S. Bonnet, F. Carlotti, A. Cornils, E. T. Crockford, P. Daniel, C. Desnos, L. Drago, A. Elineau, A. Fischer, N. Grandrémy, P.-L. Grondin, L. Guidi, C. Guieu, H. Hauss, K. Hayashi, J. A. Huggett, L. Jalabert, L. Karp-Boss, K. M. Kenitz, R. M. Kudela, M. Lescot, C. Marec, A. McDonnell, Z. Mériguet, B. Niehoff, M. Noyon, T. Panaïotis, E. Peacock, M. Picheral, E. Riquier, C. Roesler, J.-B. Romagnan, H. M. Sosik, G. Spencer, J. Taucher, C. Tilliette, M. Vilain, First release of the Pelagic Size Structure database: Global datasets of marine size spectra obtained from plankton imaging devices. *Earth System Science Data Discussions*, 1–41 (2023).

106. R. Kiko, M. Picheral, D. Antoine, M. Babin, L. Berline, T. Biard, E. Boss, P. Brandt, F. Carlotti, S. Christiansen, L. Coppola, L. de la Cruz, E. Diamond-Riquier, X. Durrieu de Madron, A. Elineau, G. Gorsky, L. Guidi, H.

Hauss, J.-O. Irisson, L. Karp-Boss, J. Karstensen, D. Kim, R. M. Lekanoff, F. Lombard, R. M. Lopes, C. Marec, A. M. P. McDonnell, D. Niemeyer, M. Noyon, S. H. O'Daly, M. D. Ohman, J. L. Pretty, A. Rogge, S. Searson, M. Shibata, Y. Tanaka, T. Tanhua, J. Taucher, E. Trudnowska, J. S. Turner, A. Waite, L. Stemmann, A global marine particle size distribution dataset obtained with the Underwater Vision Profiler 5. *Earth System Science Data* **14**, 4315–4337 (2022).

107. A. Neeley, S. E. Beaulieu, C. Proctor, I. Cetinić, J. Futrelle, I. Soto Ramos, H. M. Sosik, E. Devred, L. Karp-Boss, M. Picheral, N. Poulton, C. S. Roesler, A. Shepherd, Standards and practices for reporting plankton and other particle observations from images. (2021).

108. A. Tagliabue, O. Aumont, R. DeAth, J. P. Dunne, S. Dutkiewicz, E. Galbraith, K. Misumi, J. K. Moore, A. Ridgwell, E. Sherman, C. Stock, M. Vichi, C. Völker, A. Yool, How well do global ocean biogeochemistry models simulate dissolved iron distributions? *Global Biogeochemical Cycles* **30**, 149–174 (2016).

109. A. Tagliabue, A. R. Bowie, P. W. Boyd, K. N. Buck, K. S. Johnson, M. A. Saito, The integral role of iron in ocean biogeochemistry. *Nature* **543**, 51–59 (2017).

110. O. Aumont, C. Ethé, A. Tagliabue, L. Bopp, M. Gehlen, PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. *Geoscientific Model Development* **8**, 2465–2513 (2015).

111. D. B. Chelton, M. G. Schlax, R. M. Samelson, Global observations of nonlinear mesoscale eddies. *Progress in Oceanography* **91**, 167–216 (2011).

112. J. Elith, J. R. Leathwick, T. Hastie, A working guide to boosted regression trees. *Journal of Animal Ecology* **77**, 802–813 (2008).

113. R. J. Steele, A. E. Raftery, Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models | University of Washington Department of Statistics. *Frontiers of Statistical Decision Making and Bayesian Analysis*, 113--130 (2010).

Acknowledgments

This work was mainly funded by NOAA (Award NA21OAR4310254 to JL, RK, LS, FL, J-OI, TO'B and CS) for the project "Developing PSSdb: a Pelagic Size Structure database to support biogeochemical modelling". MD, RK, and LS received further support from the European Union project TRIATLAS (European Union Horizon 2020 program, grant agreement 817578). RK additionally acknowledges support via a Make Our Planet Great Again grant from the French National Research Agency (ANR) within the Programme d'Investissements d'Avenir #ANR-19-MPGA-0012 and from the Heisenberg Programme of the German Science Foundation #KI1387/5-1.

Funding: NOAA NA21OAR4310254 (JL, RK, LS, FL, J-OI, TO'B and CS)
European Union Horizon 2020 program grant 817578 (MD, RK, and LS)
French National Research Agency (ANR), Programme d'Investissements d'Avenir grant #ANR-19-MPGA-0012 (RK)
Heisenberg Programme of the German Science Foundation #KI1387/5-1 (RK)

Author contributions:

Conceptualization: MD, MC-U, JYL, RK, TO'B, J-OI, FL, LS, CS
Methodology: MD, MC-U, JYL, RK
Data acquisition: RK, J-OI, FL, LS, AC, LG, CG, HH, AMcD, MN, MP, and J-BR
Data collection, acquisition, or curation: NB, SB, FC, CD, LD, AE, NG, JAH, LJ, ML, ZM, BN, TP, ER, JT, CT, MV

Visualization: MD
Supervision: JYL, RK
Writing—original draft: MD
Writing—review & editing: all authors

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: The Pelagic Size Structure database workflow has been implemented in Python and is freely available at https://github. com/jessluo/PSSdb. The code to train, check the model performance and feature importance, and predict mesoplankton size distribution based on xgboost functions is available at https://github.com/mdugenne/PSSdb_Learning. All datasets and models used in this study are available at https://zenodo.org/communities/pssdb/ and https://github.com/mdugenne/PSSdb_Learning.

PSSdb data contributors consortium: Marc Picheral (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Lionel Guidi (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Andrew M. P. McDonnell (Oceanography Department, University of Alaska Fairbanks, Fairbanks, AK, United States), Laetitia Drago (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Zoé Mériguet (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Marion Vilain (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Chloé Tilliette (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Nagib Bhairy (Mediterranean Institute of Oceanography, Marseille, France); Sophie Bonnet (Mediterranean Institute of Oceanography, Marseille, France), Thelma Panaïotis (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Margaux Noyon (Department of Oceanography and Institute for Coastal and Marine Research, Nelson Mandela University, Gqeberha, South Africa), Cécile Guieu (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Astrid Cornils (Polar Biological Oceanography, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany), Jean-Baptiste Romagnan (DECOD, IFREMER, INRAE, Institut Agro, Nantes, France), Nina Grandrémy (DECOD, IFREMER, INRAE, Institut Agro, Nantes, France), François Carlotti (Mediterranean Institute of Oceanography, Marseille, France), Magali Lescot (Mediterranean Institute of Oceanography, Marseille, France), Emilie Riquier (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefarnche, Villefranche-sur-mer, France), Laetitia Jalabert (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Amanda Elineau (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Corinne Desnos (Sorbonne Université,CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-mer, France), Helena Hauss (GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany), Jan Taucher (GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany), Barbara Niehoff (Polar Biological Oceanography, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany), Jenny Huggett (Oceans and Coastal Research, Department of Forestry, Fisheries and the Environment, Cape Town, South Africa)

Figures and Tables

Table 1. Fit of mesoplankton size spectra predictions. Statistical metrics to evaluate the goodness of fit of mesoplankton size spectra predictions in increasing taxonomic ranks. Optimal models are indicated by low AIC/BIC and high R^2

	Colonial N_2-fixers			Rhizarians		Crustaceans		
Rank	Phylum	Family	FG	Phylum	Class	Phylum	Class	FG
Taxa	Cyanobacteria	Oscillatoriales	Puff Tuff	Radiozoa	Acantharia Collodaria Phaeodaria Foraminifera	Crustacea	Copepoda Malacostraca Ostracoda	Carnivore, Omnivore
BIC	145.92	145.4	143.3	150.5	132.7	145.6	135.9	134.3
AIC	144.1	144.1	142.1	149.1	133.4	152.2	134.7	135.1
R^2	0.93	0.93	0.93/0.95	0.61	0.54/0.66/0.26/0.25	0.69	0.84/0.62/0.43	0.68/0.62

Functional groups (FG): Morphotypes for colonial N_2-fixers or trophic group for copepods according to the compilation of Benedetti et al. (*46*)

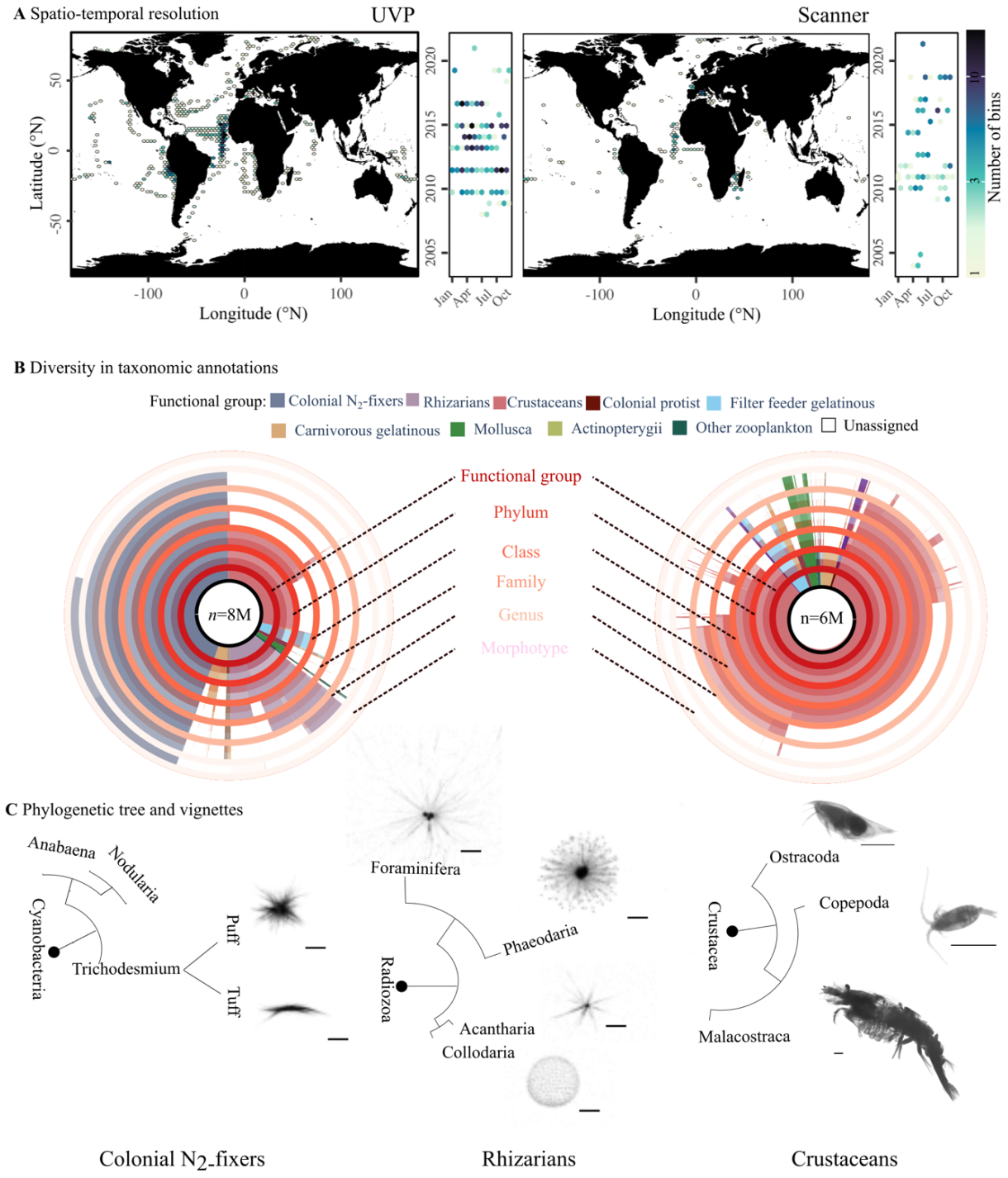


Fig. 1. Resolution and diversity in taxonomic annotations of the UVP (left) and scanner (right) imaging datasets. **(A)** Spatio-temporal resolution of the imaging datasets produced in the Pelagic Size Structure database. The colorbar represents the number of spatial (1°x1° latitude/longitude) or temporal (year month) bins in these datasets. **(B)** Diversity of taxonomic annotations (n) separated by taxonomic ranks (red to pink rings) and colored by plankton functional groups. Blank portions correspond to organisms unassigned at the considered ranks. We grouped all annotations at different levels (phylum, class, family or morphotype) to predict the biogeography of the pelagic size structure on a global scale for the three main mesoplanktonic groups (colonial N_2-fixers, Rhizarians, and Crustaceans). **(C)** Phylogenetic tree and examples of vignette illustrating the genetic and morphological diversity within colonial N_2-fixers (left), Rhizarians (middle), and Crustaceans (right). All scale bars represent 1 mm.

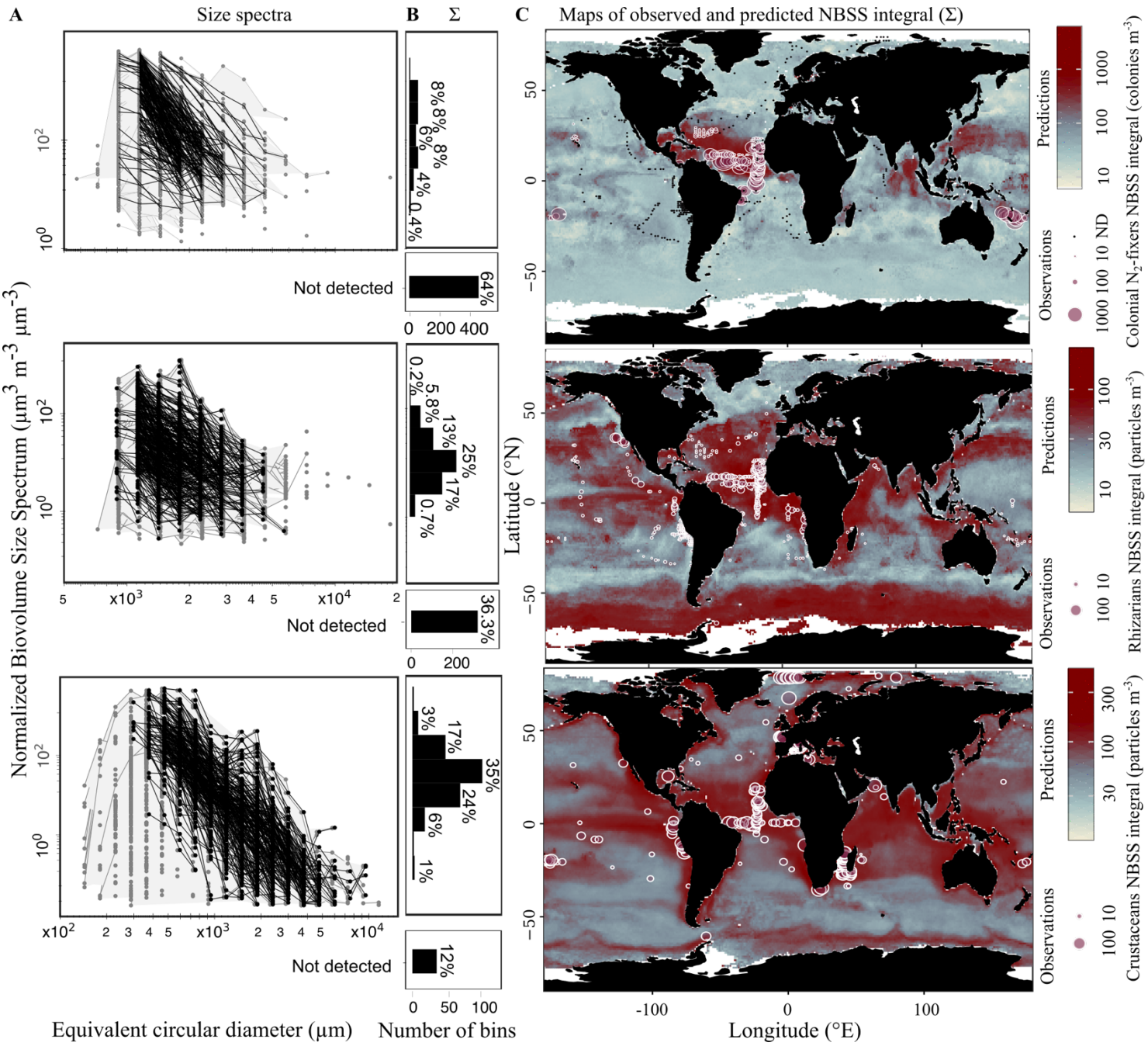


Fig. 2. Observed size spectra (left) and maps of observations and predictions (right) of colonial N_2-fixers (top panels), Rhizarians (middle panels), and Crustaceans (bottom panels) total abundance. (A) Group-specific size spectra derived from UVP (for colonial N_2-fixers and Rhizarians) and scanner (for Crustaceans) images. Gray dots mark size classes that were filtered out due to methodological bias (mesh of the nets, particles too rare to be counted accurately) and black dots indicate all final data points used to infer spectral parameters and predict NBSS. Note that x-axis limits differ depending on the instrument resolution and detection limits. **(B)** Total abundance of colonial N_2-fixers, Rhizarians, and Crustaceans in individual grid cells (1°x1° longitude/latitude and year month) distributed globally. **(C)** Maps of observed (open circles) and predicted (colored tiles) average abundance, obtained by integrating NBSS estimates using predicted parameters (slope, intercept, size range). ND: Not detected.

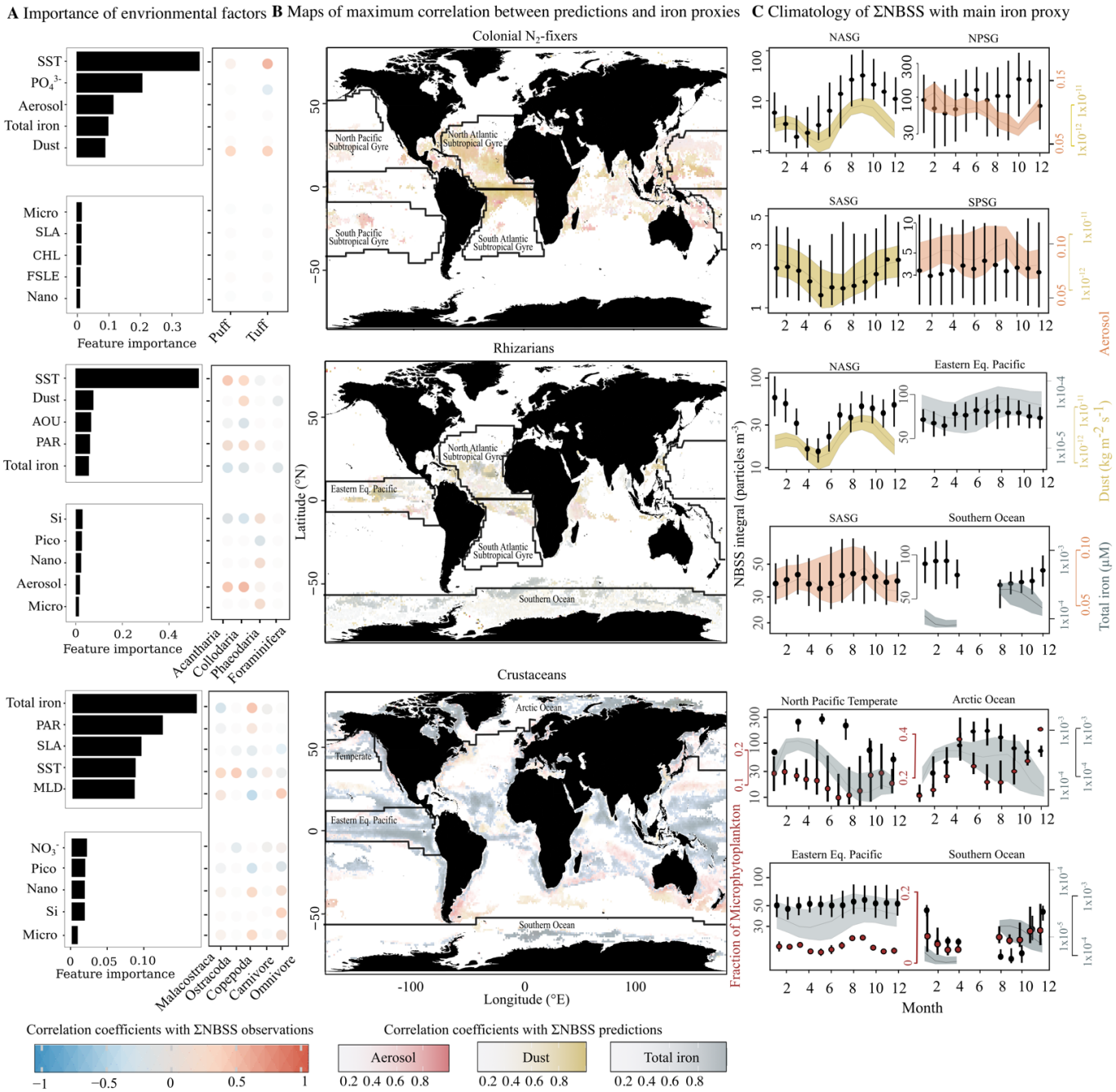


Fig. 3. Importance of environmental factors in predicting the spectral biogeography of colonial N_2-fixers (top panels), Rhizarians (middle panels), and Crustaceans (bottom panels). (A) Top and last 5 environmental factors and their respective importance in predicting mesoplankton groups spectral biogeography based on model scores (left panels) and Spearman correlation coefficients with observed abundances (right panels). Significant correlation coefficients are highlighted using a transparency gradient, with non-significant p-values (Table S1) given the highest transparency. **(B)** Maps of maximum correlation coefficients between predicted NBSS integral and multiple iron proxies used to determine the main proxy in specific ecosystems. The main iron proxy was attributed according to the variable counting the most grid cells in specific ecosystems and ocean regions. Ecosystems whose climatology is presented in panel c are outlined in black. **(C)** Climatology of mesoplankton groups (black dots) and the main iron proxy (ribbons) in ecosystems with significant correlations. Seasonal trends of microphytoplankton (red dots) are superimposed on the climatology of Crustaceans NBSS integral to link their dynamics to that of phytoplankton prey.

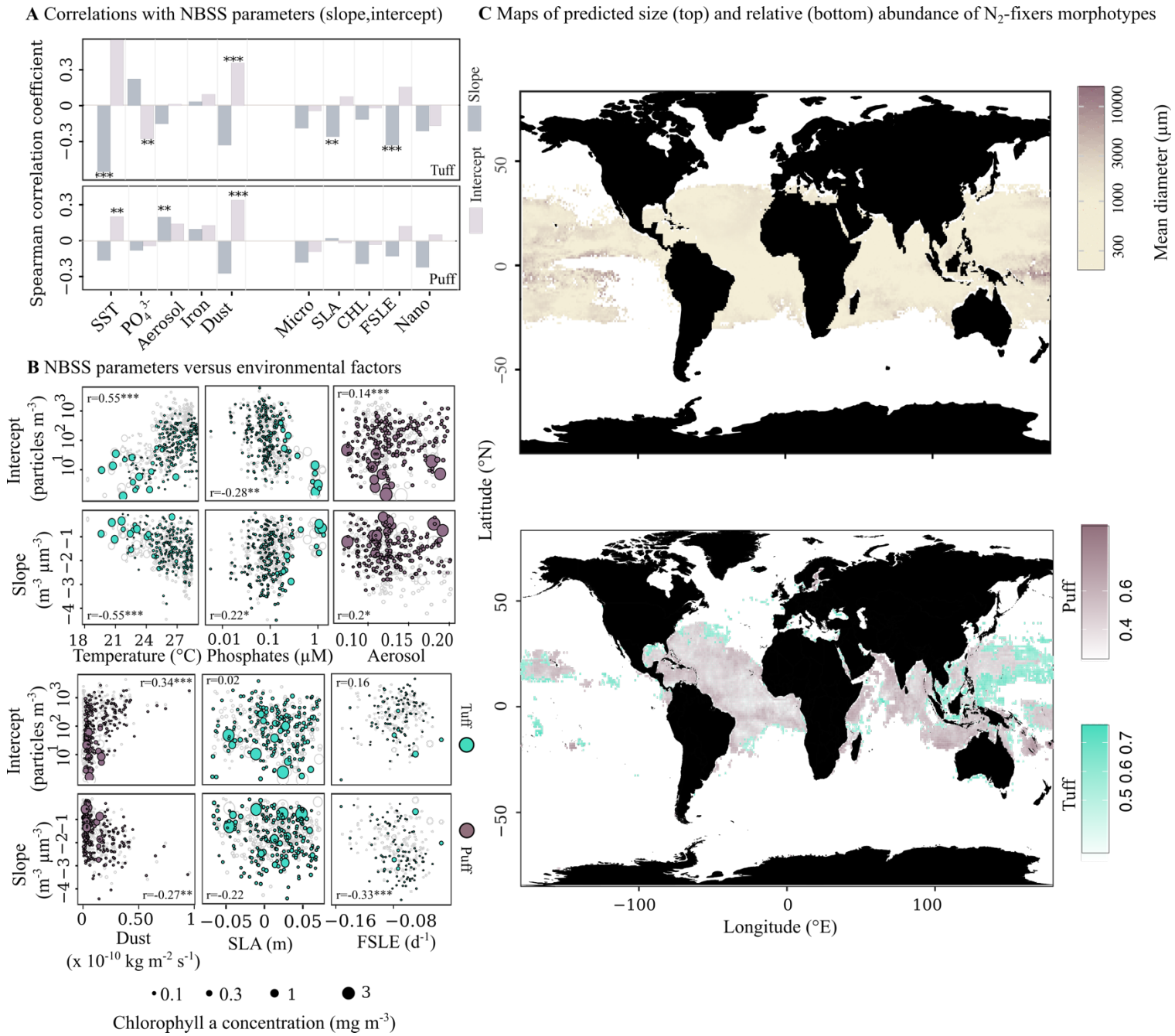


Fig. 4. Disentangling the effects of environmental factors on observations and predictions of colonial N_2-fixers size distribution. **(A)** Spearman correlation coefficients between top and last ranked environmental factors and Tuff (top panel) or Puff (bottom panel) NBSS parameters (slope in gray and intercept in purple). Symbols indicate the level of significance at 0.005 (***), 0.01 (**), 0.05 (*) and are only drawn for one parameter to simplify the panel (all significance levels are indicated in panel n). **(B)** Effects of significant environmental factors on Tuff (green dots) or Puff (brown dots) NBSS parameters (slope and intercept). Datapoints vary in size according to the mean chlorophyll-a concentration in that same bin. **(C)** Maps of colonial N_2-fixers predicted size (top) and relative (bottom) abundance of morphotypes (Tuff in green and Puff in brown). SST: sea surface temperature, PO_4^{3-}: phosphate concentration, Micro: fraction of microphytoplankton, SLA: sea level anomaly, CHL: chlorophyll-a concentration, FSLE: finite size Lyapunov exponent, Nano: Fraction of nanophytoplankton.

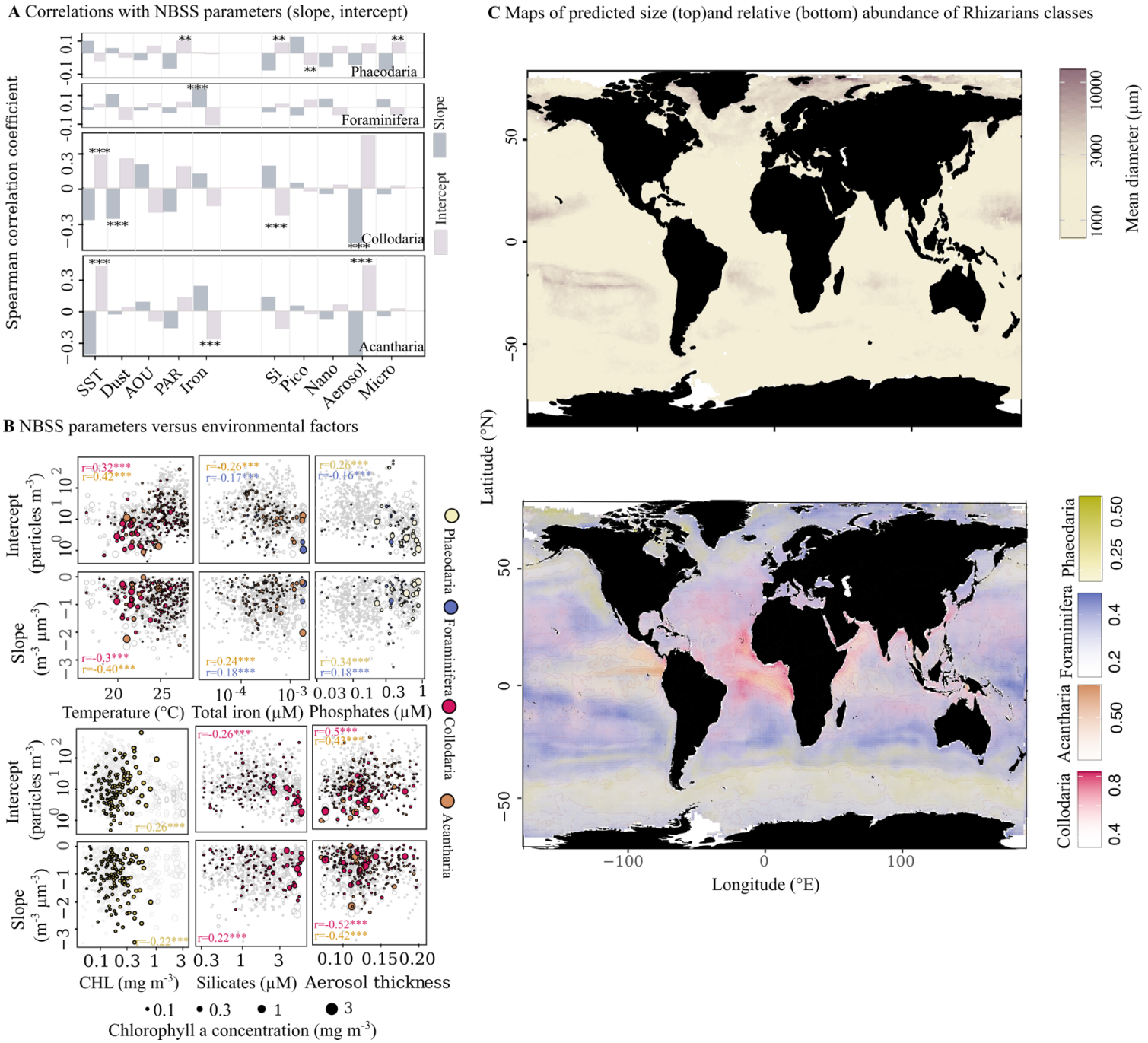


Fig. 5. Disentangling the effects of environmental factors on observations and predictions of Rhizarians size distribution. **(A)** Spearman correlation coefficients between top and last ranked environmental factors and Phaeodaria (1st panel), Foraminifera (2nd panel), Collodaria (3rd panel) or Acantharia (4th panel) NBSS parameters (slope in gray and intercept in purple). Symbols indicate the level of significance at 0.005 (***), 0.01 (**), 0.05 (*) and are only drawn for one parameter to simplify the panel (all significance levels are indicated in panel b). **(B)** Effects of significant environmental factors on Phaeodaria (beige dots), Foraminifera (blue dots), Collodaria (pink dots) or Acantharia (orange dots) NBSS parameters (slope and intercept). Datapoints vary in size according to the mean chlorophyll-a concentration in that same bin. **(C)** Maps of Rhizarians predicted size (top) and relative (bottom) abundance of classes (Phaeodaria in beige, Foraminifera in blue, Collodaria in pink, and Acantharia in orange). SST: sea surface temperature, AOU: apparent oxygen utilisation, PAR: photosynthetically available radiance, Si: Silicate concentration, Pico: fraction of picophytoplankton, Nano: Fraction of nanophytoplankton, Micro: fraction of microphytoplankton.

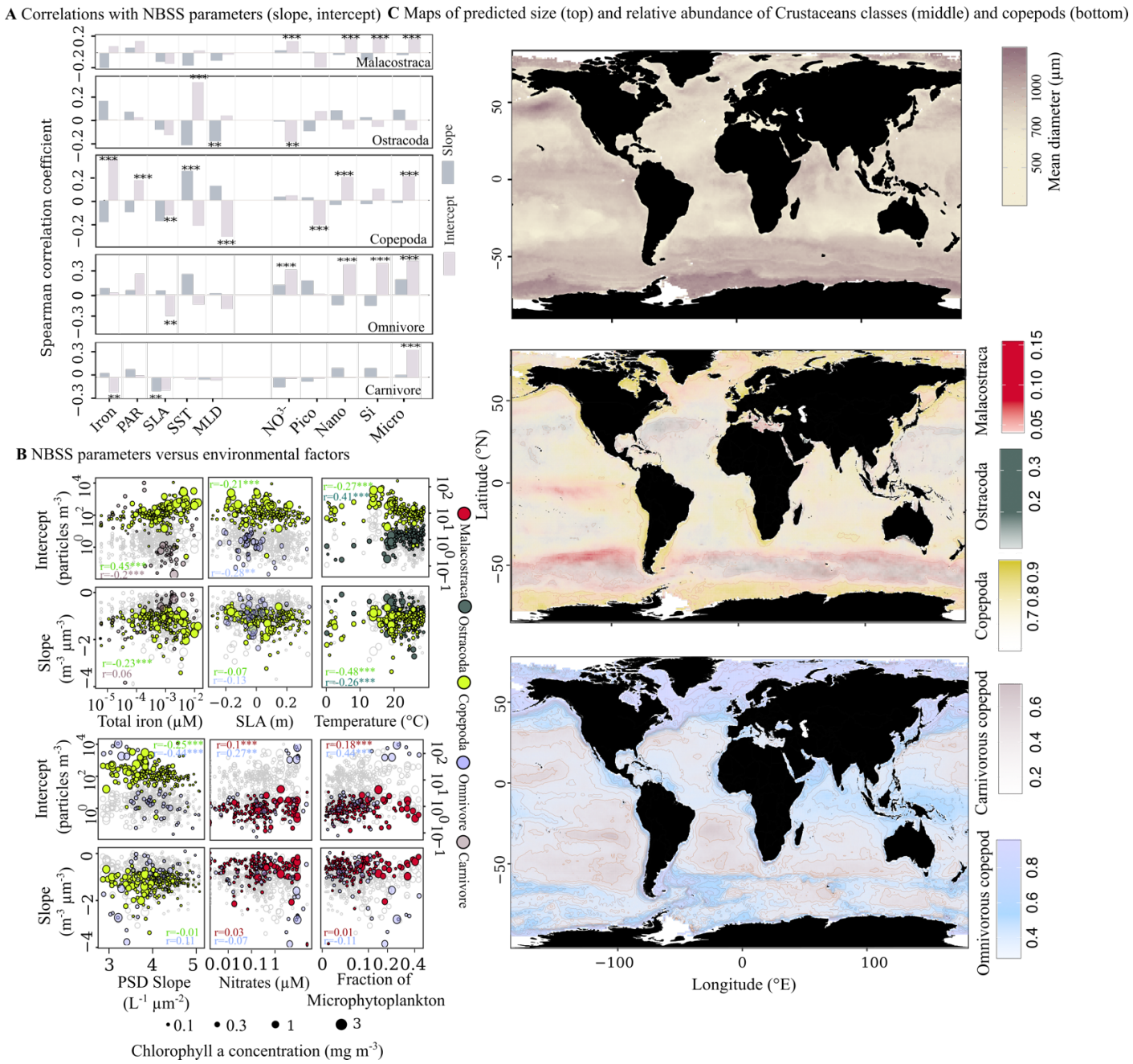


Fig. 6. Disentangling the effects of environmental factors on observations and predictions of Crustaceans size distribution. (A) Spearman correlation coefficients between top and last ranked environmental factors and Malacostraca (1st panel), Ostracoda (2nd panel), Copepoda (3rd panel), omnivorous copepods (4th panel) or carnivorous copepods (5th panel) NBSS parameters (slope in gray and intercept in purple). Symbols indicate the level of significance at 0.005 (***), 0.01 (**), 0.05 (*) and are only drawn for one parameter to simplify the panel (all significance levels are indicated in panel b). **(B)** Effects of significant environmental factors on Malacostraca (red dots, left axis), Ostracoda (green dots, left axis), Copepoda (yellow dots, left axis), omnivorous copepods (blue dots, right axis) or carnivorous copepods (brown dots, right axis) NBSS parameters (slope and intercept). Datapoints vary in size according to the mean chlorophyll-a concentration in that same bin. **(C)** Maps of Crustaceans predicted size (top) and relative abundance of classes (Malacostraca in red, Ostracoda in green, Copepoda in yellow) and copepod trophic strategy (omnivore in blue and carnivore in brown) based on a compilation by Benedetti et al. (85). PAR: photosynthetically available radiance, SLA: sea level anomaly, SST: sea surface temperature, MLD: mixed layer depth, NO_3^-: Nitrate concentration, Pico: fraction of picophytoplankton, Nano: Fraction of nanophytoplankton, Si: Silicate concentration, Micro: fraction of microphytoplankton

Supplementary Materials

Supplementary Materials included:

- **Supplementary Fig. S1. Distribution (diagonal) and correlation within spectral parameters (intercept, slope, and size range) of three mesoplankton groups. (A)** Colonial N_2-fixers. **(B)** Rhizarians. **(C)** Crustaceans. Datapoints vary in size according to the average chlorophyll-a concentration observed in satellite products in the same spatio-temporal bin. The distribution of spectral parameters is shown with violin plots on the diagonal.
- **Supplementary Fig. S2. Environmental factors used to predict the spectral biogeography of the main mesoplankton groups.** Set of average (for the month of January) environmental factors included as explanatory variables to predict the size distribution of colonial N_2-fixers, Rhizarians and Crustaceans globally. Variables include surface temperature and concentrations of nitrates, phosphates, silicates (not shown given its strong correlation with nitrates), and apparent oxygen utilization, along with mixed layer depth (MLD) from World Ocean Atlas. The surface concentration of diverse iron pools were approximated by satellite (wet dust deposition and aerosol optical thickness) or biogeochemical model (total iron) products. Total iron concentrations measured *in situ* as part of the GEOTRACES program (colored dots, downloaded at https://geotraces.webodv.awi.de/service/) are superimposed on top of the model PISCES continuous predictions (distributed by Copernicus), to validate the product used in our study. Phytoplankton concentration and composition was approximated by chlorophyll a concentration (CHL), particle size distribution slope (PSD slope), and fractions of pico-, nano- and microphytoplankton (not shown since their relative proportion affect directly PSD slopes) satellite products distributed by NASA, alongside daily photosynthetically available Radiance (PAR). Mesoscale activities were indicated by two AVISO products: Sea level anomaly (SLA), indicative of cyclonic or anticyclonic eddies under negative or positive anomaly, and finite size Lyapunov exponent (FSLE), marking the divergence rate of water parcels in close proximity.
- **Supplementary Fig. S3. Correlations between iron proxies and predictions of NBSS integral in 1°x1° grid cells. (A)** Identification of main iron proxy (top panel) based on the maximum correlation coefficients between iron proxies (dust, aerosol, or total iron) and predictions of colonial N_2-fixers NBSS integral in 1°x1° grid cells. **(B)** Identification of main iron proxy (top panel) based on the maximum correlation coefficients between iron proxies (dust, aerosol, or total iron) and predictions of Rhizarians NBSS integral in 1°x1° grid cells. **(C)** Identification of main iron proxy (top panel) based on the maximum correlation coefficients between iron proxies (dust, aerosol, or total iron) and predictions of Crustaceans NBSS integral in 1°x1° grid cells. **(D)** Mask of ecosystems (Equatorial-Upwellings: orange, Polar: green, Subtropical: blue, Temperate-Subpolar: pink) used to plot seasonal trends of NBSS predictions in association with the main iron proxy. The main proxy is identified in each ecosystem / ocean basins by calculating the number of bins associated to each iron proxy. The proxy with the maximum number of bins in a given ecosystem/ocean basin is considered the main iron pool.

- **Supplementary Table S1. Spearman correlation coefficients between environmental factors and the integral of mesoplankton size spectra calculated at the finest taxonomic resolution.** Environmental factors are ranked according to their feature importance in model fits. Symbols indicate the level of significance at 0.005 (***), 0.01 (**), 0.05 (*). Bold variables highlight the main descriptors of group-specific NBSS biogeography detailed in the main text.

Supplementary Materials for

Key link between iron and the size structure of three main mesoplanktonic groups (Crustaceans, Rhizarians, and colonial N_2-fixers) in the Global Ocean

Mathilde Dugenne *et al.*

*Corresponding authors: Mathilde Dugenne (mdugenne@hotmail.com), Rainer Kiko (rkiko@geomar.de)

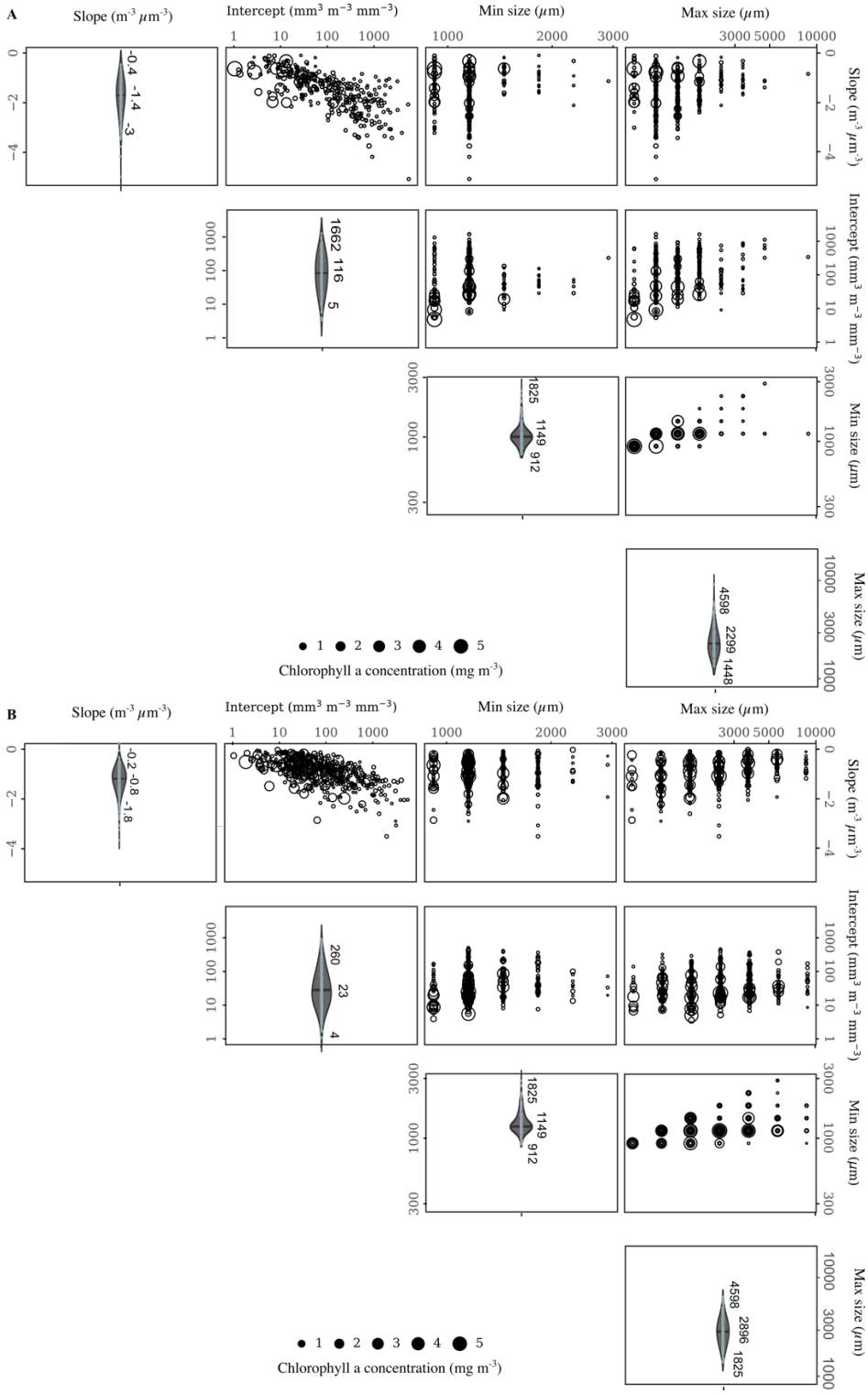
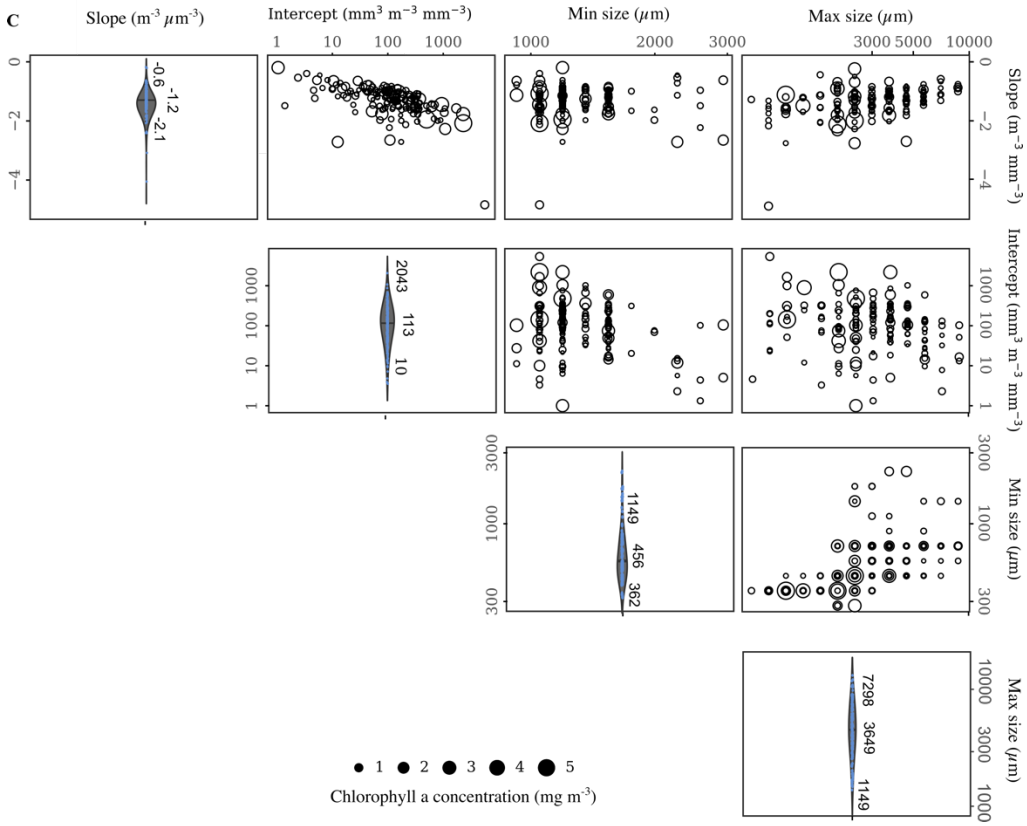
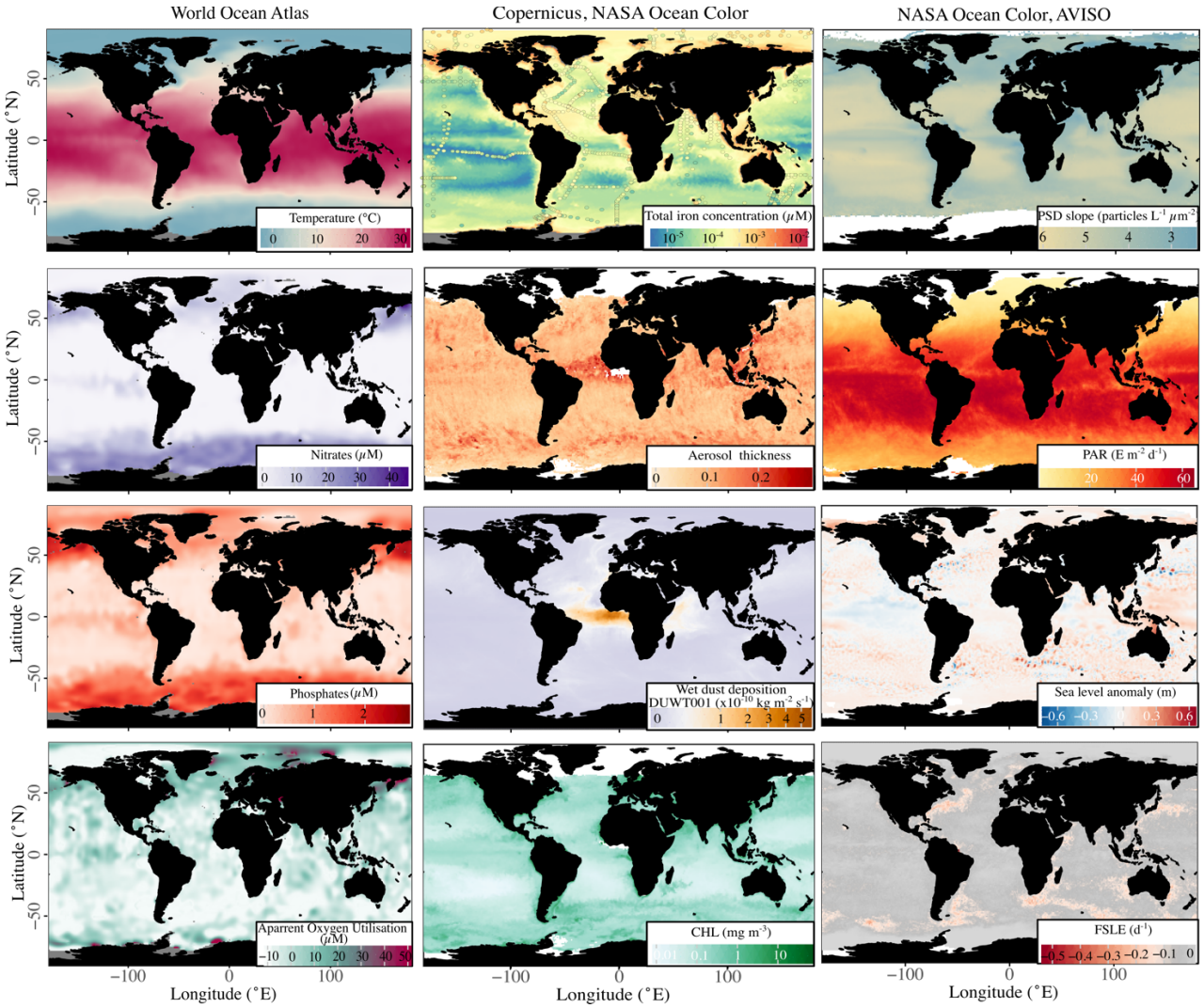


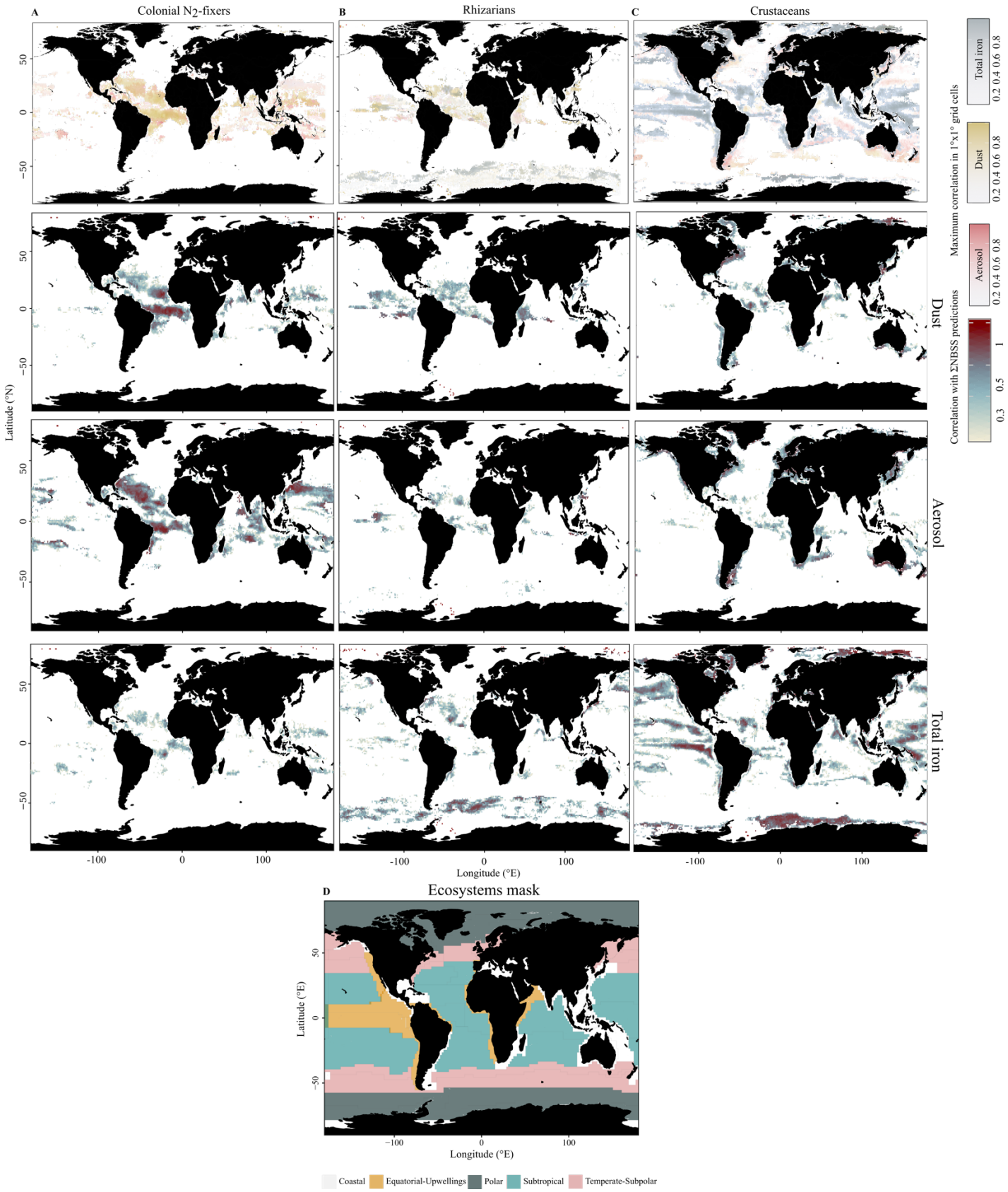
Figure continues on page 4



Supplementary Fig. S1. Distribution (diagonal) and correlation within spectral parameters (intercept, slope, and size range) of three mesoplankton groups. **(A)** Colonial N_2-fixers. **(B)** Rhizarians. **(C)** Crustaceans. Datapoints vary in size according to the average chlorophyll-a concentration observed in satellite products in the same spatio-temporal bin. The distribution of spectral parameters is shown with violin plots on the diagonal.



Supplementary Fig. S2. Environmental factors used to predict the spectral biogeography of the main mesoplankton groups. Set of average (for the month of January) environmental factors included as explanatory variables to predict the size distribution of colonial N_2-fixers, Rhizarians and Crustaceans globally. Variables include surface temperature and concentrations of nitrates, phosphates, silicates (not shown given its strong correlation with nitrates), and apparent oxygen utilization, along with mixed layer depth (MLD) from World Ocean Atlas. The surface concentration of diverse iron pools were approximated by satellite (wet dust deposition and aerosol optical thickness) or biogeochemical model (total iron) products. Total iron concentrations measured in situ as part of the GEOTRACES program (colored dots, downloaded at https://geotraces.webodv.awi.de/service/) are superimposed on top of the model PISCES continuous predictions (distributed by Copernicus), to validate the product used in our study. Phytoplankton concentration and composition was approximated by chlorophyll a concentration (CHL), particle size distribution slope (PSD slope), and fractions of pico-, nano- and microphytoplankton (not shown since their relative proportion affect directly PSD slopes) satellite products distributed by NASA, alongside daily photosynthetically available Radiance (PAR). Mesoscale activities were indicated by two AVISO products: Sea level anomaly (SLA), indicative of cyclonic or anticyclonic eddies under negative or positive anomaly, and finite size Lyapunov exponent (FSLE), marking the divergence rate of water parcels in close proximity.



Caption on page 7

Supplementary Fig. S3. Correlations between iron proxies and predictions of NBSS integral in 1°x1° grid cells. **(A)** Identification of main iron proxy (top panel) based on the maximum correlation coefficients between iron proxies (dust, aerosol, or total iron) and predictions of colonial N_2-fixers NBSS integral in 1°x1° grid cells. **(B)** Identification of main iron proxy (top panel) based on the maximum correlation coefficients between iron proxies (dust, aerosol, or total iron) and predictions of Rhizarians NBSS integral in 1°x1° grid cells. **(C)** Identification of main iron proxy (top panel) based on the maximum correlation coefficients between iron proxies (dust, aerosol, or total iron) and predictions of Crustaceans NBSS integral in 1°x1° grid cells. **(D)** Mask of ecosystems (Equatorial-Upwellings: orange, Polar: green, Subtropical: blue, Temperate-Subpolar: pink) used to plot seasonal trends of NBSS predictions in association with the main iron proxy. The main proxy is identified in each ecosystem / ocean basins by calculating the number of bins associated to each iron proxy. The proxy with the maximum number of bins in a given ecosystem/ocean basin is considered the main iron pool.

Supplementary Table S1. Spearman correlation coefficients between environmental factors and the integral of mesoplankton size spectra calculated at the finest taxonomic resolution. Environmental factors are ranked according to their feature importance in model fits. Symbols indicate the level of significance at 0.005 (***), 0.01 (**), 0.05 (*). Bold variables highlight the main descriptors of group-specific NBSS biogeography detailed in the main text.

Rank	Environmental factor	Spearman correlation coefficients for NBSS integral (+ rank for partition model fit or NBSS slope for the extra variables)	
		Colonial N_2-fixers	
		Puff	Tuff
1	**Temperature (°C)**	0.19*(1)	0.53***(2)
2	**Phosphates (μM)**	-0.05(3)	-0.27**(3)
3	**Aerosol (unitless)**	0.46***(5)	0(5)
4	**Total iron (μM)**	0.15(4)	0.11(4)
5	**Wet dust deposition (kg m^{-2} s^{-1})**	0.33***(2)	0.37***(1)
6	PAR (E m^{-2} d^{-1})	-0.25**(10)	-0.24*(6)
7	MLD (m)	0.21*(8)	0.04(12)
8	AOU (μM)	0.05(16)	0.08(13)
9	Silicates (μM)	0.15(7)	0.33***(7)
10	PSD slope (L^{-1} μm^{-2})	0.12(15)	0.16(16)
11	Nitrates (μM)	-0.32*(6)	-0.22*(8)
12	Picophytoplankton (fraction)	0.12(14)	0.16(17)
13	**Microphytoplankton (fraction)**	-0.15(12)	-0.19(15)
14	**SLA (m)**	-0.01(11)	0.05(10)
15	**CHL (mg m^{-3})**	-0.06(9)	-0.04(9)
16	**FSLE (d^{-1})**	0.11(13)	0.13(11)
17	**Nanophytoplankton (fraction)**	-0.15(17)	-0.21***(14)
Extra	GEOTRACES† dissolved iron (μM)	-0.09(-0.11)	-0.21***(-0.12)
	WOA surface dissolved oxygen (μM)	-0.04(0.12)	-0.33***(0.29)

Rank	Environmental factor	Rhizarians			
		Acantharia	Collodaria	Phaeodaria	Foraminifera
1	**Temperature (°C)**	0.42***(2)	0.31***(2)	-0.11*(12)	0.03(9)
2	**Wet dust deposition (kg m^{-2} s^{-1})**	0.02(8)	0.28***(4)	-0.05(6)	-0.13**(3)
3	**AOU (μM)**	-0.11*(11)	-0.2***(8)	0.06(14)	0.03(16)
4	**PAR (E m^{-2} d^{-1})**	0.15***(13)	0.2***(12)	0.15**(17)	0.05(5)
5	**Total iron (μM)**	-0.25***(6)	-0.18***(11)	-0.01(3)	-0.18***(1)
6	FSLE (d^{-1})	0.004(16)	-0.05(9)	-0.05(5)	-0.02(12)
7	MLD (m)	-0.2***(7)	0.06(14)	-0.26***(8)	0.004(2)
8	SLA (m)	0.09*(10)	0.11*(5)	0.22***(4)	0.06(4)
9	Phosphates (μM)	-0.13**(15)	-0.22***(3)	0.33***(1)	0.18***(6)
10	PSD slope (L^{-1} μm^{-2})	-0.01(17)	-0.01(17)	-0.19***(15)	0.04(11)
11	Nitrates (μM)	-0.14**(9)	-0.11*(13)	0.19***(7)	0.11**(8)
12	CHL (mg m^{-3})	0.01(12)	0(7)	0.22***(13)	-0.03(7)
13	**Silicates (μM)**	-0.16***(14)	-0.26***(10)	0.16***(11)	0.04(10)
14	**Picophytoplankton (fraction)**	-0.007(3)	-0.01(15)	-0.19***(10)	0.04(13)
15	**Nanophytoplankton (fraction)**	0.04(4)	0.01(16)	0.18***(16)	-0.03(14)
16	**Aerosol (unitless)**	0.43***(5)	0.51***(1)	0.11*(9)	0.21**(17)
17	**Microphytoplankton (fraction)**	0.007(1)	0.01(6)	0.22***(2)	-0.04(15)
Extra	GEOTRACES† dissolved iron (μM)	0.04 (-0.03)	-0.13 (0.3)	0.87***(0.16)	0.77***(0.3)
	WOA surface dissolved oxygen (μM)	-0.08 (0.31)	-0.28***(0.19)	-0.25***(-0.07)	-0.05(0.05)

Table continues on page 9

Rank	Environmental factor	Mesoplankton group				
				Crustaceans		
		Malacostraca	Ostracoda	Copepoda	Carnivore	Omnivore
1	**Total iron (μM)**	-0.26***(2)	-0.02(8)	0.44***(1)	-0.05(4)	-0.04(13)
2	**PAR (E m^{-2} d^{-1})**	-0.14*(4)	0.01(7)	0.19**(9)	0.04(1)	-0.15(10)
3	**SLA (m)**	-0.08(7)	-0.16*(4)	-0.2**(3)	-0.02(16)	-0.33**(12)
4	**Temperature (°C)**	0.29***(3)	0.39***(1)	-0.25***(5)	0.18*(12)	-0.1(1)
5	**MLD (m)**	0.19**(12)	0.02(14)	-0.4***(4)	0.04(10)	0.4(5)
6	PSD slope (L^{-1} μm^{-2})	-0.13(16)	0.04(16)	-0.29***(14)	0.11(11)	-0.18***(4)
7	CHL (mg m^{-3})	0.18*(17)	-0.0(9)	0.37***(8)	-0.1(15)	0.29***(2)
8	AOU (μM)	0.12(15)	-0.08(5)	0.16***(10)	0.06(6)	0.14(11)
9	Aerosol (unitless)	0.17*(11)	0.2*(2)	-0.05(13)	0.14*(2)	-0.27**(7)
10	Phosphates (μM)	-0.06(1)	-0.22**(13)	-0.06(12)	0.07(3)	0.4***(14)
11	FSLE (d^{-1})	-0.24**(10)	-0.16*(3)	-0.04(11)	-0.03(9)	0.17(8)
12	Wet dust deposition (kg m^{-2} s^{-1})	-0.04(5)	0.04(11)	0.2**(15)	-0.15*(14)	-0.61***(9)
13	**Nitrates (μM)**	0.08(9)	-0.23**(6)	0.07(2)	-0.04(5)	0.27**(15)
14	**Picophytoplankton (fraction)**	-0.14*(14)	0.04(17)	-0.29***(6)	0.11(13)	-0.43***(3)
15	**Nanophytoplankton (fraction)**	0.14*(13)	-0.04(15)	0.28***(16)	-0.12(17)	0.43***(6)
16	**Silicates (μM)**	0.13(6)	-0.03(12)	0.11(7)	0.02(7)	0.36***(16)
17	**Microphytoplankton (fraction)**	0.15*(8)	-0.03(10)	0.28***(17)	-0.12(8)	0.44***(17)
Extra	GEOTRACES† dissolved iron (μM)	-0.48***(0.19)	0.03**(0.20)	0.5***(-0.43)	0.05 (-0.04)	0.15*(-0.33)
	WOA surface dissolved oxygen (μM)	-0.33***(0.13)	-0.27***(0.22)	0.19***(0.43)	-0.17*(0.18)	0.17*(0.28)

PAR : Photosynthetic Available Radiation, MLD : Mixed Layer Depth, AOU: Apparent Oxygen Utilisation, PSD slope: Particle Size Distribution slope, SLA: Sea Level Anomaly, CHL: surface chlorophyll-a, FSLE: Finite Size Lyapunov Exponent. † Based on GEOTRACES (v. 2021 downloaded at https://geotraces.webodv.awi.de/service/IDP2021_v2%3EGEOTRACES_IDP2021_Seawater_Discrete_Sample_Data_v2) surface dissolved iron concentration measured from Niskin bottles averaged by Longhurst regions and month.