ECOLOGY

Deviation from neutral species abundance distributions unveils geographical differences in the structure of diatom communities

Emanuele Pigani[1,2]†, Bruno Hay Mele[1]†, Lucia Campese[1], Enrico Ser-Giacomi[3], Maurizio Ribera[1], Daniele Iudicone[1]*, Samir Suweis[2,4]*

In recent years, the application of metagenomics techniques has advanced our understanding of plankton communities and their global distribution. Despite this progress, the relationship between the abundance distribution of diatom species and varying marine environmental conditions remains poorly understood. This study, leveraging data from the *Tara* Oceans expedition, tests the hypothesis that diatoms in sampled stations display a consistent species abundance distribution structure, as though they were sampled from a single ocean-wide metacommunity. Using a neutral sampling theory, we thus develop a framework to estimate the structure and diversity of diatom communities at each sampling station given the shape of the species abundance distribution of the metacommunity and the information of a reference station. Our analysis reveals a substantial temperature gradient in the discrepancies between predicted and observed biodiversity across the sampled stations. These findings challenge the hypothesis of a single neutral metacommunity, indicating that environmental differences substantially influence both the composition and structure of diatom communities.

INTRODUCTION

Recent advances in high-throughput molecular techniques have greatly advanced our understanding of microbial communities (*1*), revealing them to be highly rich and harboring a considerable population of rare, low-abundance species, known as the "rare biosphere" (*2*, *3*). Furthermore, these technological improvements have also enabled the examination of large-scale biogeographical patterns of microscopic organisms, a topic of long-standing debate (*4*, *5*).

As a relevant example, these advancements have substantially influenced the study of pelagic marine environments, where ecosystems are dominated—in terms of biomass and abundance—by plankton, an ensemble of microscopic drifting organisms with notable phylogenetic (*6*) and trophic (*7*) diversity. In these ecosystems, ocean transport processes such as advection and mixing play crucial roles in determining the spatial distribution of these organisms (*8*). Although advancements have been made, there remains a lack of comprehensive understanding of the dispersion and geographic structure of planktonic microbial communities. This underscores the need for a conceptual framework elucidating the "seascape"—a term used to denote the interplay of biological, chemical, and physical elements in the ocean, and its influence on plankton ecology (*8*). In particular, a noteworthy challenge lies in determining whether plankton samples represent a mixture of species from the same community (*9*, *10*)—a concept known as metacommunity in terrestrial ecology (*11*)—or distinct communities shaped by specific environments (*8*, *12*, *13*). Equally relevant is the question of whether plankton samples represent a mixture of species that display the same community structure everywhere, or if the structure presents distinct patterns in time and space, beyond just composition.

In community ecology, species abundance distribution (SAD)—the distribution of individuals within a species in a given community—has been a cornerstone of research (*11*, *14*), which can potentially shed light on the study of plankton microbial communities and their structure. Studying SADs not only allows for a characterization of ecological communities but also provides critical insights that enable species number estimations at larger scales beyond direct measurement, accomplished by inferring SAD distribution forms (*15*–*19*). Notably, the functional form of SADs has shown consistency across diverse ecosystems. This suggests that fundamental ecological mechanisms, such as birth/death processes and migration/speciation events, might underpin such ubiquitous patterns (*20*–*23*). Yet, the interpretation of SADs in ecological terms and the determination of their specific drivers prove challenging, as multiple mechanisms could potentially lead to the same SAD (*14*, *24*, *25*).

In the specific context of pelagic communities, the extent to which the structure of SADs is connected with oceanic environmental conditions and water mixing remains uncertain. Ser-Giacomi *et al.* (*26*) proposed and tested a density and dispersal-dependent neutral demographic model for plankton communities using the extensive *Tara* Oceans dataset (*27*). Their study focused on "Operational Taxonomic Units" (OTUs) from 152 sampling stations, representing four different organism size classes, at two depths. Their findings revealed that plankton SADs follow a power-law distribution with exponents exhibiting minimal variation across locations (*26*), suggesting an absence of a geographical signature when excluding the most abundant species in the analysis. Later studies found smaller power-law exponents in lacustrine compared to marine environments, attributing this to differences in dispersion physics, specifically chaotic advection (*28*). These mechanisms, while difficult to characterize, could explain also the lack of immediate geography in the marine environment.

[1]Stazione Zoologica Anton Dohrn, 80135 Napoli, Italy. [2]Dipartimento di Fisica e Astronomia "Galileo Galilei", Università di Padova, 35131 Padova, Italy. [3]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Istituto Nazionale di Fisica Nucleare, INFN, Sezione di Padova, 35131 Padova, Italy.
*Corresponding author. Email: iudicone@szn.it (D.I.); samir.suweis@unipd.it (S.S.)
†These authors contributed equally to this work.

While the aforementioned studies have focused on protist communities as a whole, encompassing various trophic levels and life strategies, Busseni *et al.* (*29*) took a different approach. They concentrated on diatoms in their examination of the *Tara* Oceans dataset, investigating the relationship between global diatom richness—which displays different latitudinal variation when compared to the one of the whole plankton (*30*)—and seascape properties. Their particular emphasis on diatoms was due to their primary role in plankton communities regarding diversity and biomass (*31*), important contributions to carbon and silica biogeochemical cycles (*32*), and homogeneous ecological position within the community despite varied life strategies (*33*), which is suggested by the lack of observed phagotrophic capability (*34*). They identified nutrient availability, chlorophyll *a*, temperature, and lateral mixing as nonlinear modulators of richness. However, their study did not analyze SADs.

In the present work, we fill this gap by analyzing the properties of diatom SADs and exploring their relationship with marine environmental conditions. In particular, we test the hypothesis of the existence of an ocean-wide neutral diatom metacommunity, from which we can determine the structural properties of the local diatom communities and their diversity. Given that oceanic plankton communities are persistently influenced by currents and display a biogeographical equilibrium with these currents (*8*), it is crucial to investigate whether the SADs of different plankton groups share a common structure despite community and environmental variations.

For this purpose, we leverage an OTU-based community table (see Materials and Methods) from 181 *Tara* Oceans sampling stations as proxies for different diatom species, with 123 sampled at the subsurface (SRF) and 58 at the deep chlorophyll maximum (DCM) depth. Then, following a tradition in theoretical ecology and statistical physics (*22*), we propose a theoretical framework that serves as a null model to test our hypothesis. Under the assumption that density-dependent birth and death events (*23*) occur for each species at each location and independently of species identity, we examine whether variations in observed local SADs structural properties can be exclusively attributed to differences in sample size or if there are unaccounted variations, potentially indicating the influence of external drivers or distinct biological traits.

RESULTS

A thorough examination of diatom community structures, spanning 181 *Tara* Oceans sampling stations, consistently reveals a similar pattern in diatom SADs, even amidst regional variations and differences in total read counts. This observation is vividly displayed in Fig. 1A, where each line, color-coded by the station's average temperature, represents a distinct station's empirical SAD. In particular, stations with lower temperatures exhibit higher abundances, consistent with the more favorable conditions for diatoms in those regions at the time of sampling (*35*). At the same time, stations characterized by different temperatures align closely with power-law distributions with nearly equivalent slopes. A maximum likelihood fit applied to these distributions confirms this behavior: The distribution exponents, although capable of variation, tend to cluster in a narrow range between −1.5 and −2 and do not display a clear geographical pattern (see fig. S1). This is visually consistent with a reference power-law distribution with exponent −1.5 denoted by a dashed line in Fig. 1A and supports similar findings from studies of whole plankton communities (*26*).

The stability in the slopes among different stations may suggest that they all share the same structure, and it is the sampling process itself that considerably influences the observed variability in the total abundance among SADs. To further explore this relationship, a synthetic undersampling experiment is performed.

Station #173—(79.0°N, 79.4°E)—at 5-m depth (SRF), distinguished by its rich biodiversity, is chosen (arbitrarily) as a reference station. The experiment involves a multinomial sampling of OTU reads from this station, where the number of sampled OTUs is aligned with the total counts of other stations. This process results in synthetic samples, with total OTU abundances corresponding to the empirical measurements but with a different composition.

As demonstrated in Fig. 1B, the SADs of these synthetic samples closely mirror the observed patterns, apparently affirming the efficacy of this undersampling experiment. We also find a strong correlation $r^2 = 0.7$ ($P < 0.01$) between the richness of empirical and synthetic communities, as shown in Fig. 1B. Additional results showing similar trends are detailed in fig. S2.

Neutral sampling theory

The similarity observed in SADs seems consistent with the scenario where each local diatom community represents a sample from an underlying species reservoir or, in other words, a metacommunity (*36*, *37*). To systematically and analytically investigate this hypothesis, we exploit a scaling methodology (*17*). This allows us to predict the composition of diatom communities at targeted sampling stations using the attributes of a reference station.

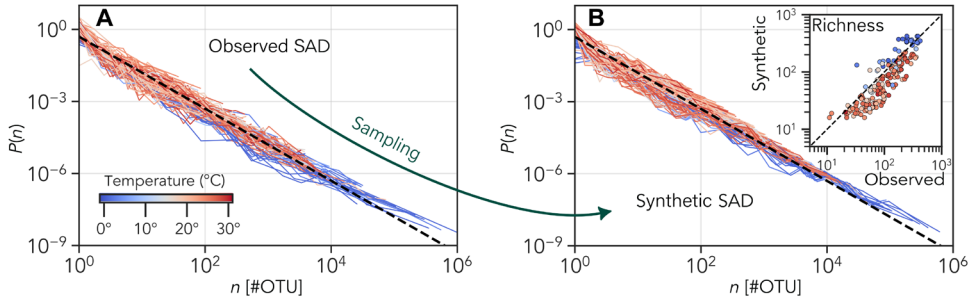

Fig. 1. Comparison of empirical and synthetic SADs. (**A**) OTU-based diatom SADs for 181 *Tara* Oceans stations. Each line is color-coded on the basis of the average temperature of its respective station. Despite higher abundances observed in lower temperature stations, the SADs maintain a consistent shape, closely approximating a power-law distribution ($\lambda \sim 1.5$, represented by the dashed line). (**B**) Synthetic SADs, generated from multinomial sampling based on the observed SAD of station #173 SRF. These synthetic SADs closely mimic the empirical ones, as evidenced by the correlation ($r^2 = 0.7$) between the synthetic and empirical richness (inset, each point is colored according to its station's temperature).

To begin with, we consider a stochastic neutral model for the dynamics of the species abundances within the hypothesized ocean metacommunity. We assume that each species undergoes birth and death events at species-independent rates given by $b_n = bn + \chi$ and $d_n = dn + \mu$, where n is the species population, b and d represent the linear per-capita birth and death rates, while χ and μ are their nonlinear corrections, thus allowing for density-dependent rates. The parameter χ represents a migration/speciation term that is different from zero even when $n = 0$, while μ accounts for a lower death probability of species with very high abundance (e.g., blooms) than individuals belonging to rare species (*26*).

Such ecological stochastic dynamics can be described in terms of a master equation (*38*), whose steady state provides the expected SAD $P(n) \equiv P(n|\alpha, \beta, r)$, which reads

$$P(n) = \theta(\alpha, \beta, r)\frac{\Gamma(\beta+1)\Gamma(n+\alpha)}{\Gamma(\alpha)\Gamma(n+\beta+1)}e^{-rn} \quad (1)$$

where we have introduced the parameters $\alpha = \chi/b$, $\beta = \mu/d$, and $r = \log b/d$, while $\Gamma(\cdot)$ is the Gamma function and $\theta(\alpha, \beta, r)$ is a normalization factor that ensures $\sum_{n\geq 1} P(n) = 1$. Asymptotically, $P(n)$ is a power-law distribution with an exponent $-\lambda = -1 + \alpha - \beta$ and an exponential cutoff that depends on r, which vanishes in the limit $b = d$ (*26*).

To investigate the existence of an ocean-wide neutral metacommunity, we thus suppose that $P(n)$ describes the SAD of the entire diatom metacommunity and that by sampling it, we can deduce the SADs and richness of local communities, as outlined in Fig. 2. Consider a metacommunity inhabiting an ocean volume V, comprising S species with abundances distributed according to Eq. 1. In a homogeneously mixed system, a local community can be seen as a sample of volume $V_p < V$ from the metacommunity, thus yielding a local SAD $\phi(n|p) \equiv \phi(n|\alpha, \beta, r, p)$ given by

$$\phi(n\,|\,p) = \begin{cases} p^n e^{-n(r-1)}\dfrac{\Gamma(n+\alpha)}{\Gamma(\alpha+1)}\dfrac{{}_2\widetilde{F}_1[n+1, n+\alpha; n+\beta+1; e^{-r}(1-p)]}{{}_2\widetilde{F}_1(1, \alpha+1; \beta+2; e^{-r})}, & \text{for } n \geq 1 \\ (1-p)\dfrac{{}_2F_1[1, \alpha+1; \beta+2; e^{-r}(1-p)]}{{}_2F_1(1, \alpha+1; \beta+2; e^{-r})}, & \text{for } n = 0 \end{cases} \quad (2)$$

as derived in Material and Methods. Here, $p = V_p/V$ represents the sampling ratio (*22*), and ${}_2\widetilde{F}_1(\cdot)$ is the regularized hypergeometric function. If species are uniformly distributed with population density

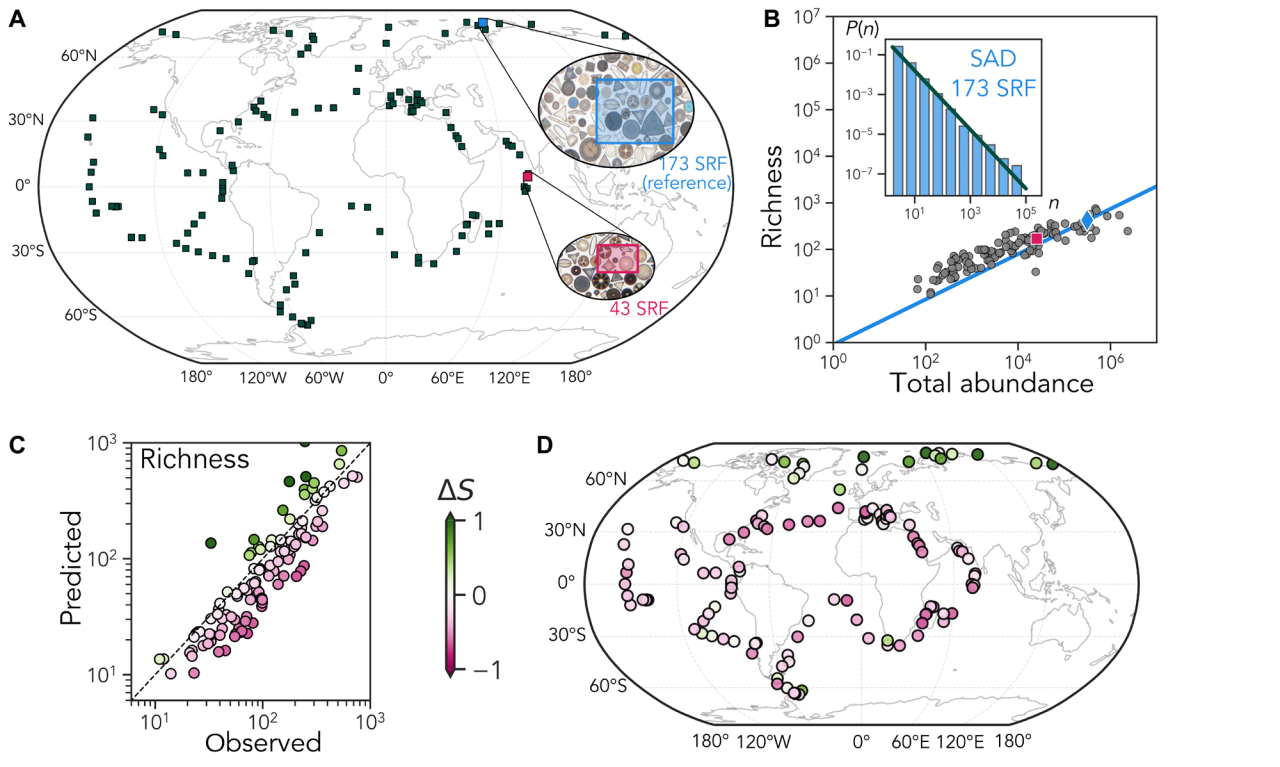

Fig. 2. Illustration and results of the neutral sampling approach. (**A**) Geographic distribution of the SRF *Tara* Oceans stations considered. Station #173 SRF (79.0°N, 79.4°E) is one of the most diverse stations, while tropical stations such as station 43 SRF (4.7°N, 73.5°E) typically have lower abundance and sampling effectiveness. (**B**) Inset: The empirical SAD for reference station #173 (blue histogram) is fitted with Eq. 1 (green curve). The fitted parameters are $\alpha = -0.2$, $\beta = 0.28$, and $r = 0.0$, indicating that the curve asymptotically approaches a power law with exponent $\lambda = -1.48$. (B) The power-law approximation of SAD allows us to downscale the richness of station #173 (blue diamond) to predict the richness of the other stations as a function of their abundances through Eq. 5 (blue line). This prediction is compared with the empirical data (red square for station 43, gray dots for all the other SRF stations). (**C**) The neutral sampling framework is validated by comparing the predicted and the observed richness. The Pearson's correlation is $r = 0.78$ with $P < 0.01$. Each point is colored on the basis of the relative error of the prediction ΔS, which is generally positive for the most biodiverse communities. (**D**) Geographic distribution of ΔS. In polar regions, $\Delta S > 0$ indicates an overestimation of richness, while underestimation occurs in tropical regions. In all panels, only SRF stations are considered.

ρ, then the total population in the sampled station is $N_p = \rho V_p$, while the total abundance in the metacommunity is $N = \rho V$, leading to $p = N_p/N$. This is typically valid when the use of ecosystem resources is saturated (*11*).

We note that $\phi(0|p)$ estimates the likelihood of the absence of a species in the local sample. Therefore, the probability of observing S_p species in V_p when the metacommunity richness is S follows a binomial distribution with an average given by

$$\langle S_p \rangle = S\left\{1 - (1-p)\frac{{}_2F_1\left[1, \alpha+1; \beta+2; e^{-r}(1-p)\right]}{{}_2F_1(1, \alpha+1; \beta+2; e^{-r})}\right\} \quad (3)$$

and variance $S[1-\phi(0|p)]\phi(0|p)$ (see Materials and Methods), aligning with several theoretical models in ecology [see, e.g., (*39*, *40*)].

The challenge of applying Eq. 3 to estimate sample diversity lies in the inability to infer p from the available data. For instance, considering the union of all 181 samples as a single meta-community is impractical due to the compositional nature of the data (*41*): Each metabarcoding sample possesses a unique library size—the total number of sequencing reads—which precludes direct comparison or amalgamation of different samples.

Nevertheless, we have devised a strategy to evaluate the single neutral metacommunity hypothesis without direct access to its parameters. This approach requires two assumptions. First, we assume the power-law asymptotic approximation of Eq. 1 as the distribution of the metacommunity SAD. If this holds, then also the local SADs should display power-law tails. We find that the majority of the stations are best fitted with Eq. 1 for $r = 0$ (see table S1) and with exponent $1 \leq \lambda < 2$. Second, we presume that the population sizes of local communities are notably smaller than the one of the metacommunity, implying $p \ll 1$. This assumption is well-founded given the nature of the data (*27*). Under these conditions, Eq. 3 simplifies to

$$\langle S_p \rangle = \underbrace{-\frac{\Gamma(-\lambda+1)}{\zeta(\lambda)}\frac{S}{(N)^{\lambda-1}}}_{K}(N_p)^{\lambda-1} \quad (4)$$

where $\zeta(\cdot)$ is the Riemann ζ-function and we introduce the metacommunity biodiversity index K, which depends solely on metacommunity properties (see Materials and Methods). Equation 4 shows the existence of a simple scaling law between the richness and the total abundance of the samples which is regulated by λ.

Moreover, by comparing Eq. 4 across two local stations and using one as a reference, we can infer the richness of the other station as

$$\langle S_p \rangle = S_{\text{ref}}\left(\frac{N_p}{N_{\text{ref}}}\right)^{\lambda-1} \quad (5)$$

Notably, this sampling relation does not involve the metacommunity index K but does depend on λ, which is nevertheless unknown beforehand. However, due to the scale invariance property of power-law distributions, a subsample from a power law also follows a power law with the same exponent. Therefore, we can estimate λ by analyzing the SAD fits of the diatom local communities, and then exploit Eq. 5 to infer the biodiversity of each local station, given the reference one. In our analysis, we use the exponent fitted in the reference community. Nevertheless, using either the average ($\lambda = 1.44$) or the mode ($\lambda = 1.51$) of the fitted λ exponents from all the local stations yields similar outcomes, as shown in fig. S3.

This neutral sampling procedure is sketched in Fig. 2 (A and B). Consistently with the synthetic sampling experiments, we chose a posteriori station #173—(79.0°N, 79.4°E)—at 5-m depth (SRF) as the reference station (Fig. 2B). Not only its community is one of the most diverse and abundant but also it is the best one in predicting the properties of the other local stations. Yet, similar results can be obtained for other choices of reference stations (see Supplementary Materials and figs. S4 and S5). Notably, the fitted parameters for the reference station are $\alpha = -0.2$, $\beta = 0.28$ (implying $\lambda = 1.48$), and $r = 0$ (Fig. 2B, inset), thus supporting the hypothesis of a power-law SAD with exponent $\lambda \sim 1.5$ [cf. (*26*)]. Consequently, Eq. 5 can be used to downscale or upscale the richness as a function of the total abundance based on the properties of the reference station (indicated by the blue diamond in Fig. 2B). This procedure yields a twofold result. On one hand, the scaling of richness with the total population of the samples qualitatively aligns with our prediction, showing a correlation coefficient $r = 0.78$. On the other hand, we observe consistent discrepancies between the data and the model. Specifically, the richness for less abundant stations is underestimated, while for the most abundant stations, it is overestimated. In the Supplementary Materials (fig. S3), we also infer the richness of the DCM stations from reference station #173 SRF, although with less accuracy.

Deviations from neutral SADs and diatom biogeography

The above results are even more evident from Fig. 2C. It shows how, from the fitted parameter λ and the properties of the reference station, we can estimate the richness of the other stations through Eq. 5. The predicted values for the richness of the SRF stations are in good agreement with the empirical ones ($r = 0.78$ with $P < 0.01$). At the same time, we find systematic deviations between the diversity estimates obtained through the neutral SAD sampling framework and the empirical ones. In particular, we analyze the distribution of relative differences between the predicted biodiversity (S_p) and the observed one (S_{loc}) in each SRF station, namely

$$\Delta S = \frac{S_p - S_{\text{loc}}}{S_{\text{loc}}} \quad (6)$$

As shown in Fig. 2D, it is clear that the estimator's error is not randomly distributed, i.e., ΔS depends on the station latitude. In other words, the discrepancies between the theory and the data, as measured by ΔS, reveal a biogeographical pattern. The Arctic stations, known for hosting a very rich and abundant diatom population (*29*), display the highest positive deviations, meaning that the estimator is overestimating richness in this region. The opposite occurs for the mid-latitudinal Atlantic stations, sampled in late winter, where the measured richness is larger than the one produced by the estimator. Last, deviations in both signs are observed in the highly oligotrophic waters of the South Pacific, while the entire tropical Pacific and Indian oceans have slightly negative anomalies.

The systematic and geographically related discrepancies of ΔS indicate that the null hypothesis of a single neutral metacommunity for oceanic diatoms is not supported by the data. This is also confirmed by the analysis of the inferred metacommunity biodiversity indexes K. The existence of a single ocean-wide metacommunity would imply

the same K for all stations. However, as shown in fig. S7, we find different values of K, for both SRF and DCM stations.

To gain a deeper understanding of the anomalies in richness prediction, we examined the correlation of ΔS with other macroecological and environmental descriptors. The most substantial correlations are illustrated in Fig. 3A, while a more detailed analysis can be found in figs. S8 and S9. Notably, ΔS exhibits a stronger negative correlation with temperature compared to richness, aligning with the observed geographical pattern for the deviation of predicted richness. Moreover, ΔS has a robust positive correlation with latitude ($r = 0.48$) and solar light ($r = 0.35$), and a negative correlation with salinity ($r = -0.43$). In all these instances, these correlations are stronger than those observed with richness. Conversely, when considering the correlation with nutrients, richness shows a moderate association with levels of NO_3 ($r = 0.34$), PO_4 ($r = 0.39$), and $Si(OH)_4$ ($r = 0.36$), while ΔS display a significant correlation only with PO_4 ($r = 0.3$).

We also investigate the relationship between ΔS and community structure properties. We find that the strongest correlation ($r = -0.68$) is with evenness, a measure that quantifies community heterogeneity. Evenness is defined as

$$E = -\frac{\sum_{i=1}^{S} x_i \log(x_i)}{\log(S)} \tag{7}$$

where x_i is the relative abundance of species i within the community, i.e., $x_i = n_i/N$ and $N = \sum_{j=1}^{S} n_j$ is the total abundance of the community. Therefore, as reported in Fig. 3B, an increase in evenness is associated with a decrease in ΔS, suggesting that ΔS captures relevant traits of the community structure with negative or positive deviations associated, among others, with lower or higher presence of dominant species characterizing the assemblage.

One might expect that the geographical distribution of evenness (reflecting its correlation with ΔS) would result in a spatial pattern for the exponents (λ) governing the SAD's power-law behavior. However, this is not the case (see fig. S11), and we can understand why. Following our previous assumptions, we derive an analytical prediction of evenness for a pure power-law SAD defined within a finite concentration range $[x_m; x_M]$ (see Eq. 25). By considering the empirical minimum (x_m) and maximum (x_M) of each station and the exponent λ obtained from the fits, we can predict the evenness of each sample, as shown in Fig. 4A. As also evident from Fig. 4B, the dependence of evenness on λ is highly non-monotonic, with maximum values for either small ($\lambda \lesssim 1$) or large ($\lambda \gtrsim 2$) and minimum for intermediate exponents ($\lambda \sim 1.5$). The reason behind this non-monotonic behavior is that for large exponents, the power-law distribution is very concentrated at x_m, i.e., the distribution is homogeneous and thus the evenness is close to 1.

Last, in the Supplementary Materials, we compare the fits for diatom (see fig. S10) and for protists (20- to 180-μm size class) communities, finding a low correlation ($r = 0.25$ with $P < 0.01$; see fig. S9). This result indicates that dispersal, which a priori would affect the diatoms as all other protists, is not the sole mechanism shaping the differences in the SAD exponents among stations. We should have found similar λ for protists and diatoms communities in this case. Moreover, a Mantel test (*42*) reveals a statistically significant relationship between ΔS and temperature T ($P < 10^{-3}$). This relationship suggests that the observed variations in SAD structure—specifically the differences between high and low evenness—may be influenced by environmental factors. In other words, these correlations demonstrate biogeographical patterns in the distribution of marine communities across the ocean, which may be shaped by a range of abiotic and biotic factors, which will require further studies.

DISCUSSION

Scaling laws are ubiquitous in nature, even though the underlying mechanisms behind them are not unique (*43*, *44*). SADs of organisms in different ecosystems fall within this category, even though their formulation may differ between ecosystems or groups. Ser-Giacomi *et al.* (*26*) showed that the SADs of eukaryotic plankton in the ocean follow a power law, provided that few, very abundant species are removed from the assemblage. They attributed such power-law behavior to a neutral model with density-dependent rates governing the exponents of the asymptotic power law. Although such exponents were fairly constant in space, the values of the composing parameters were not, possibly reflecting dispersal effects, as also proposed by Villa *et al.* (*28*).

To investigate the factors ruling the assemblage of diatoms, we have thus proposed a null neutral model, able to quantify the effects of demographic stochasticity and sampling effort in shaping diatom SAD. By establishing a theoretical framework that accounts for

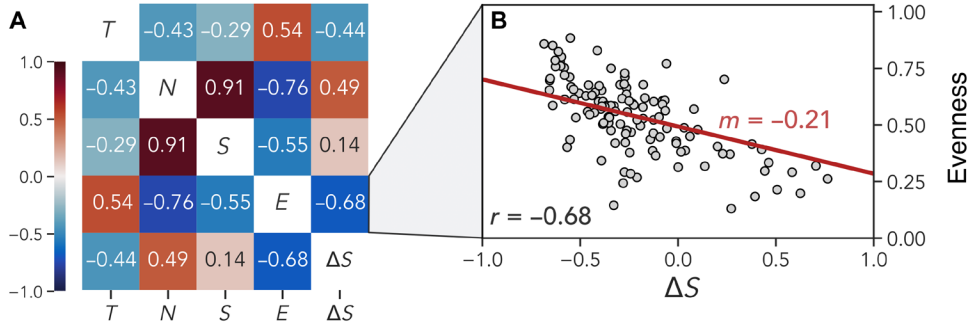

Fig. 3. Correlation analysis. (**A**) Spearman's correlation matrix for temperature (T), total abundance (N), richness (S), evenness (E), and the relative deviation between the predicted and observed richness (ΔS) for the SRF stations. All correlations are statistically significant ($P \leq 0.01$), with the exception of the correlation between S and ΔS ($P = 0.04$). The largest correlation observed between ΔS and evenness. (**B**) Scatter plot representing each SRF station, with evenness on the y axis and ΔS on the x axis. The slope of the fitted linear relation is $m = -0.21$, indicating a negative correlation between them.

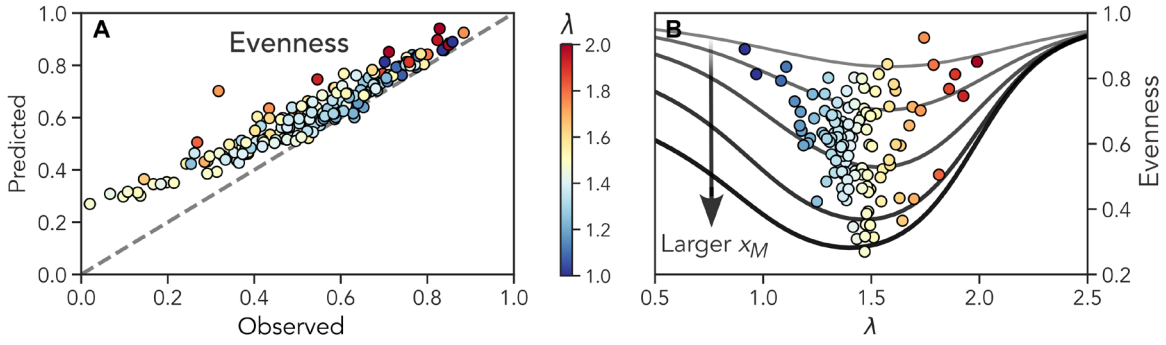

Fig. 4. Evenness relationship with power-law distribution. (**A**) Comparative analysis of predicted and empirical evenness. The empirical evenness is calculated from the data of SRF stations using Eq. 7, while the predicted evenness is derived from Eq. 25, assuming that each SAD follows a power-law distribution with exponent λ within the range of empirical minimum (x_m) and maximum (x_M) frequencies. The correlation between predicted and observed evenness is strong ($r^2 \simeq 0.9$), particularly for $E \geq 0.4$. However, the expected values generally overestimate the evenness. The gray dashed line represents the bisector, and each point is color-coded according to the value of the fitted λ. Notably, the stations with the highest evenness typically exhibit the largest (red circles) or the smallest (blue circles) λ. This relationship is further illustrated in (**B**), where the predicted (solid lines) and empirical (color-coded circles) evenness as a function of λ display a non-monotonic dependence. Solid lines represent Eq. 25 for $x_m = 10^{-6}$ and progressively larger values of x_M (from top to bottom: 0.001, 0.01, 0.1, 0.5, and 0.99).

undersampling from a neutral metacommunity, we can qualitatively predict the SRF average diversity of diatom communities at various sampling stations using the characteristics of a reference station.

The less accurate prediction of the richness of DCM stations from #173 SRF suggests that their distribution is substantially different from the SRF ones. This inaccuracy does reflect the different environmental contexts of DCMs, where diatoms are not abundant. DCMs display a wide range of environmental conditions and, in the case of seasonal DCMs, histories [see, e.g., (*45*)]. In addition, if we attempt to use a DCM reference station to predict the richness of the other DCM stations, then the correlation is still significant, but there is a systematic underestimation of the richness (see figs. S5 and S6).

We have also provided quantitative evidence with the hypothesis that a single ocean-wide neutral metacommunity, where the differences among local sampled stations are attributed solely to different sampling efforts, does not hold, at least for this specific plankton group. In particular, the analysis of the deviations between the predicted and the observed richness reveals the signature of biogeographical patterns, with ΔS being substantially correlated with temperature (*46*), as shown in fig. S8. This suggests that the environment plays a key role in shaping community structures and cannot be neglected when modeling diatom communities. It is worth stressing that temperature, along with nutrients and light—the environmental parameters we previously explored—should not be considered as direct drivers of the observed difference, but rather as proxies of different dynamics of the seascape.

Bio-oceanographic processes thus underlie the observed differences in diatom community structures among regions highlighted by ΔS and the negative monotonic trend between ΔS and the evenness E (see Fig. 3 and fig. S9). During the time of sampling, the subarctic region experienced a phytoplankton seasonal bloom (*35*), in which diatoms played a key role. This was due to high nutrient concentrations and moderate stratification, even at the low in situ temperatures (*12*). In contrast, the permanently stratified and low-silicon, nutrient-depleted tropical region hampers the accumulation of phytoplankton biomass. This particularly affects many medium-sized eukaryotic fractions, favoring the bacterial component and, more broadly, the picoplanktonic fraction, with diatoms playing a lesser role in these regions.

In addition, we have found a negative correlation between local diversity and evenness. This result, in stark contrast to what is typically observed in terrestrial ecosystems, can be attributed to the fact that uneven distributions are characterized by a few dominant species with very high abundance, such as blooming species (*47*).

We have also analyzed which parameters in the neutral, dispersal-modulated (through the constant term in the birth rate, representing immigration) model could affect the deviation between the predicted and the observed richness. Since, in most of the stations, we underestimate the observed diversity, it would be possible to reduce such deviations by assuming a higher α and/or a lower β in the reference station (see fig. S12). This, in turn, would imply a smaller density-dependent b or a larger density-independent χ for the birth rate and/or vice versa (larger d, smaller μ) for the terms related to death. In other words, the neutral model fails to properly describe the presence of boom-and-bust species (*48*). Such species can accumulate for a limited amount of time at a much higher speed than the other species, thus reducing the evenness. Moreover, when $\Delta S < 0$, the stations are even more dominated by a few diatom species. Not all diatoms display the boom-and-bust strategy, especially in environmental contexts, e.g., subtropical gyres, where the system's carrying capacity for diatom is quite small due to low silicon (Si) and iron (Fe) concentrations. However, the observed structural differences exist among whole diatom assemblages without filtering rare or abundant species, and they are also present among low abundance stations, where the very abundant species are absent.

In summary, we suggest that the community assembly differences, indicated by our estimators ΔS and K, are not only the result of different sampling efforts but also the stem from the complex interplay of life strategies [see, e.g., (*49*)], abiotic processes [see, e.g., (*50*)], and biotic interactions [see, e.g., (*51*)] that operate differently in space, time, and across species and also have varying impacts at the intra-group level. Therefore, a key area for future research will be integrating these factors with the impact of the seascape in ecological modeling. The link between ΔS, derived within our proposed framework, and the above processes deserves further investigation and may reveal

overlooked interplays between the seascape and plankton dynamics to embed in ecological modeling.

MATERIALS AND METHODS

Data

The dataset used for the analysis is the OTU-based community table presented in (*29*). Specifically, we analyzed metabarcoding data collected at 123 *Tara* Oceans epipelagic stations for the size fraction of 20 to 180 μm. This comprised a total number of 181 samples, with 123 stations sampled at the subsurface at a depth of 5 m (SRF) and 58 of them at DCM depth. Total nucleic acids (DNA + RNA) were extracted from all samples, and the hypervariable V9 region of the nuclear 18*S* ribosomal DNA was amplified through polymerase chain reaction (PCR) to obtain a metabarcoding survey of plankton (*7*). Quality filtering based on read quality checks and a minimum number of occurrences of three copies in at least two different samples was implemented to reduce PCR and sequencing errors. OTUs were subsequently formed by applying the Swarm approach (*52*) and resulted in a final number of 5308 validated surface diatom OTUs, with the information of the total number of reads in each station. Additional information about the OTU-based community table can be found in (*29*).

Ecological framework

We use a community assembly model to describe the empirical abundance distributions of the OTUs, which serve as proxies for species, in the different samples. Let us consider an ecosystem composed of N individuals belonging to S different species. Following (*38*), we assume that each species undergoes birth and death at rates that depend only on the species abundance. In particular, if n is the population of a species, then the birth and death master equation governing the dynamics of the species abundance is

$$\dot{P}_n(t) = b_{n-1}P_{n-1}(t) + d_{n+1}P_{n+1}(t) - (b_n + d_n)P_n(t) \tag{8}$$

where the (species-independent) rates are

$$\begin{cases} b_n = bn + \chi, \\ d_n = dn + \mu \end{cases} \tag{9}$$

Here, b and d represent the linear per-capita birth and death rates, while χ and μ are their nonlinear corrections, whose effect is especially important for species of low abundance. The steady-state solution of the master equation describing the ecological stochastic dynamics of such a modeled community provides the expected SAD $P(n)$, which reads (*26*, *38*)

$$P(n) = \theta(\alpha, \beta, r)\frac{\Gamma(\beta+1)\Gamma(n+\alpha)}{\Gamma(\alpha)\Gamma(n+\beta+1)}e^{-rn} \text{ when } n \geq 1 \tag{10}$$

where the parameters α, β, and r are functions of the birth and death rates, namely

$$\begin{cases} \alpha = \chi/b, \\ \beta = \mu/d, \\ r = -\log b/d \end{cases} \tag{11}$$

whereas the normalization factor $\theta(\alpha, \beta, r)$ takes into account the fact that each species at the global scale consists of at least one individual

$$\theta(\alpha, \beta, r) = \left[\sum_{n=1}^{\infty}\frac{\Gamma(\beta+1)\Gamma(n+\alpha)}{\Gamma(\alpha)\Gamma(n+\beta+1)}e^{-rn}\right]^{-1} = \frac{e^r\Gamma(1+\beta)}{\alpha_2F_1(1, 1+\alpha, 2+\beta, e^{-r})} \tag{12}$$

where ${}_2F_1(\cdot)$ indicates the hypergeometric function. Similarly to (*53*), we would like to determine the local expected SAD when we look at local subsample $p = V_p/V$, i.e., considering a subvolume V_p of the whole volume V, under the hypothesis of random sampling. As already mentioned in the main text, if species are uniformly distributed with density ρ, then the sampling effectiveness corresponds to the sampled fraction of individuals $p = N_p/N$. Let us consider a species of n individuals among the whole population. Under the random sampling hypothesis, the conditional probability for the species to have k individuals at the subsample p is given by a binomial distribution, that is

$$\mathscr{P}_{\text{binom}}(k|n, p) = \binom{n}{k}p^k(1-p)^{n-k} \tag{13}$$

Hence, the probability $\phi(k|p)$ that a species has an abundance $k \geq 1$ at a subscale p is the marginalization of the binomial over all the possible total abundances n of the species at the global scale, i.e., $\phi(k|p) = \sum_{n\geq k}\mathscr{P}_{\text{binom}}(k|n, p)P(n)$. This leads to

$$\phi(k|p) = \frac{p^k e^{r-kr}\Gamma(k+\alpha)\,{}_2\widetilde{F}_1\left[k+1, k+\alpha; k+\beta+1; e^{-r}(1-p)\right]}{\Gamma(\alpha+1)\,{}_2\widetilde{F}_1(1, \alpha+1; \beta+2; e^{-r})} \tag{14}$$

where ${}_2\widetilde{F}_1(\cdot)$ is the regularized hypergeometric function. Analogously, we can calculate the probability $\phi(k=0|p)$ that the species is not present at the local scale, with the only difference that now n should be strictly larger than 0. The explicit calculation leads to

$$\phi(0|p) = (1-p)\frac{{}_2F_1\left[1, \alpha+1; \beta+2; e^{-r}(1-p)\right]}{{}_2F_1(1, \alpha+1; \beta+2; e^{-r})} \tag{15}$$

Consequently, $1 - \phi(0|p)$ represents the probability for a species to be found in the local sample, and thus the probability of sample S_p species in the subvolume p with total richness S is given by the binomial distribution

$$p(S_p|S, p)) = \binom{S}{S_p}[1-\phi(0|p)]S_p\phi(0|p)^{S-S_p} \tag{16}$$

which has average $S[1 - \phi(0|p)]$ and variance $S[1 - \phi(0|p)]\phi(0|p)$. Therefore, the expected number of species $\langle S_p\rangle$ present with sampling effectiveness p is

$$\langle S_p\rangle = S\left\{1 - (1-p)\frac{{}_2F_1[1, \alpha+1; \beta+2; e^{-r}(1-p)]}{{}_2F_1(1, \alpha+1; \beta+2; e^{-r})}\right\} \tag{17}$$

Neutral sampling theory from a power-law SAD

Equation 1 can be asymptotically approximated by $P(n) \propto n^{-\lambda}e^{-rn}$ (*26*) with $\lambda = 1 - \alpha + \beta$. When $r = 0$, this distribution is therefore well approximated by a power-law SAD

$$P(n) = \frac{n^{-\lambda}}{\zeta(\lambda)} \tag{18}$$

where $\zeta(\cdot)$ is the Riemann ζ-function. In this approximation, the probability is not present at the local scale given by Eq. 15 becomes

$$\phi(0|p) = \frac{\mathrm{Li}_\lambda(1-p)}{\zeta(\lambda)} \tag{19}$$

where Li is the polylogarithm function $\mathrm{Li}_n(z) = \sum_{k=1}^{\infty} \frac{z^k}{k^n}$. If we assume $p \ll 1$, then we can expand this relation and get the scaling relation for the richness

$$\langle S_p \rangle = S \begin{cases} -\dfrac{\Gamma(-\lambda+1)}{\zeta(\lambda)} p^{\lambda-1} + o(p^\lambda - 1) & \text{if } 1 < \lambda < 2, \\ \dfrac{6}{\pi^2} p[1 - \log(p)] + o[p\log(p)] & \text{if } \lambda = 2, \\ \dfrac{\zeta(\lambda-1)}{\zeta(\lambda)} p + o(p) & \text{if } \lambda > 2 \end{cases} \tag{20}$$

The richness scaling depends on λ when $1 < \lambda \leq 2$ and is linear otherwise. Here, we are interested in the first regime, as λ is typically close to 1.5. In this case, we can explicitly write the scaling relation from the metacommunity properties as

$$\langle S_p \rangle = \underbrace{-\frac{\Gamma(-\lambda+1)}{\zeta(\lambda)} \frac{S}{(N)^{\lambda-1}}}_{K} (N_p)^{\lambda-1} \tag{21}$$

where K is the metacommunity biodiversity index, which is a function of the metacommunity properties. We can infer K from the sampling data, provided that we know λ, as

$$K = \frac{S_{\mathrm{loc}}}{N_{\mathrm{loc}}^{\lambda-1}} \tag{22}$$

where S_{loc} and N_{loc} are the observed local richness and total abundance, respectively.

Model fitting

We fit Eq. 1 to all samples. We optimize the parameters of Eq. 1 for each sample using the generalized simulated annealing algorithm implemented in the GenSA R library (*54*). Generalized simulated annealing is a robust and efficient method that avoids resting on local minima thanks to a Metropolis-based acceptance model. We manually tuned the lower and upper bounds for parameters.

The resulting parameters and P values are reported in table S1, and α and β are also displayed in the inset of fig. S7. A total of 143 out of 181 stations displayed $r = 0$, while the remaining 40 stations have a mean of the order of 10^{-3}, which is two orders of magnitudes higher than those reported by Ser-Giacomi *et al.* (*26*). Thus, the fits do not display exponential decay for large abundances. A Kolmogorov-Smirnov test was conducted on the fits, resulting in ~20% acceptance rate (see table S1).

We choose the most abundant stations as good candidates for being the reference station. The relative deviations ΔS between the predicted and the observed richness for the reference station candidates are reported in fig. S2. Being the one that minimizes the overall ΔS, we choose station #173 (79.0°N, 79.4°E) at 5-m depth (SRF) as the reference one.

Expected value of the evenness for a power-law SAD

Motivated by the empirical fits, which show that the parameter r is close to 0 in the majority of the stations, we derive an analytical formula for the expected evenness for a SAD that follows a power law with exponent $-\lambda$, with domain between two relative frequencies x_m and x_M, namely

$$P(x) = \frac{(\lambda-1)(x_M x_m)^\lambda}{x_M^\lambda x_m - x_M x_m^\lambda} x^{-\lambda} \tag{23}$$

By definition (see Eq. 7), the Evenness is the Shannon diversity index normalized by its maximum possible value, i.e., when every species is equally likely (and thus equal to the average). Since for the power-law distribution, the average is

$$\bar{x} = \frac{(\lambda-1)(x_m^2 x_M^\lambda - x_M^2 x_m^\lambda)}{(\lambda-2)(x_m x_M^\lambda - x_M x_m^\lambda)} \tag{24}$$

we can calculate the expected evenness E_{pl} by first averaging the evenness over $P(x)$ and then replacing S with $1/\bar{x}$. This leads to

$$E_{\mathrm{pl}}(\lambda \,|\, x_m, x_M) = \frac{x_m^2 x_M^\lambda[(\lambda-2)\log(x_M)+1] - x_M^2 x_m^\lambda[(\lambda-2)\log(x_M)+1]}{(\lambda-2)(x_m^2 x_M^\lambda - x_M^2 x_m^\lambda)\log\left[\frac{(\lambda-1)(x_m^2 x_M^\lambda - x_M^2 x_m^\lambda)}{(\lambda-2)(x_m x_M^\lambda - x_M x_m^\lambda)}\right]} \tag{25}$$

Supplementary Materials

This PDF file includes:

Figs. S1 to S12

Table S1

REFERENCES AND NOTES

1. T. Van Rossum, P. Ferretti, O. M. Maistrenko, P. Bork, Diversity within species: Interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020).
2. C. Pedrós-Alió, The rare bacterial biosphere. *Ann. Rev. Mar. Sci.* **4**, 449–466 (2012).
3. M. D. J. Lynch, J. D. Neufeld, Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* **13**, 217–229 (2015).
4. J. B. H. Martiny, B. J. Bohannan, J. H. Brown, R. K. Colwell, J. A. Fuhrman, J. L. Green, M. C. Horner-Devine, M. Kane, J. A. Krumins, C. R. Kuske, P. J. Morin, S. Naeem, L. Ovreås, A.-L. Reysenbach, V. H. Smith, J. T. Staley, Microbial biogeography: Putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).
5. C. A. Hanson, J. A. Fuhrman, M. C. Horner-Devine, J. B. H. Martiny, Beyond biogeographic patterns: Processes shaping the microbial landscape. *Nat. Rev. Microbiol.* **10**, 497–506 (2012).
6. T. Cavalier-Smith, Kingdom protozoa and its 18 phyla. *Microbiol. Mol. Biol. Rev.* **57**, 953–994 (1993).
7. C. De Vargas, S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. L. Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horák, O. Jaillon, G. Lima-Mendez, J. Lukeš, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Coordinators, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not,

H. Ogata, S. Pesant, J. Raes, M. E. Sieracki, S. Speich, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, E. Karsenti, Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
8. D. J. Richter, R. Watteaux, T. Vannier, J. Leconte, P. Frémont, G. Reygondeau, N. Maillet, N. Henry, G. Benoit, O. D. Silva, T. O. Delmont, A. Fernàndez-Guerra, S. Suweis, R. Narci, C. Berney, D. Eveillard, F. Gavory, L. Guidi, K. Labadie, E. Mahieu, J. Poulain, S. Romac, S. Roux, C. Dimier, S. Kandels, M. Picheral, S. Searson, T. O. Coordinators, S. Pesant, J.-M. Aury, J. R. Brum, C. Lemaitre, E. Pelletier, P. Bork, S. Sunagawa, F. Lombard, L. Karp-Boss, C. Bowler, M. B. Sullivan, E. Karsenti, M. Mariadassou, I. Probert, P. Peterlongo, P. Wincker, C. de Vargas, M. R. d'Alcalà, D. Iudicone, O. Jaillon, Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *eLife* **11**, e78129 (2022).
9. L. G. M. Baas Becking, *Geobiologie of Inleiding tot de Milieukunde* (WP Van Stockum & Zoon, 1934). accessible as Baas Becking's: Geobiology, Or Introduction to Environmental Science, First Edition, D. E. Canfield, Ed. (John Wiley & Sons Ltd, 2016).
10. M. A. O'Malley, The nineteenth century roots of 'everything is everywhere'. *Nat. Rev. Microbiol.* **5**, 647–651 (2007).
11. S. Hubbel, *The Unified Neutral Theory of Biodiversity and Biogeography*. (Princeton Univ. Press, 2001).
12. A. R. Longhurst, *Ecological Geography of the Sea* (Elsevier, 2010).
13. G. Sommeria-Klein, R. Watteaux, F. M. Ibarbalz, J. J. P. Karlusich, D. Iudicone, C. Bowler, H. Morlon, Global drivers of eukaryotic plankton biogeography in the sunlit ocean. *Science* **374**, 594–599 (2021).
14. B. J. McGill, R. S. Etienne, J. S. Gray, D. Alonso, M. J. Anderson, H. K. Benecha, M. Dornelas, B. J. Enquist, J. L. Green, F. He, A. H. Hurlbert, A. E. Magurran, P. A. Marquet, B. A. Maurer, A. Ostling, C. U. Soykan, K. I. Ugland, E. P. White, Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.* **10**, 995–1015 (2007).
15. J. W. F. Slik, V. Arroyo-Rodríguez, S.-I. Aiba, P. Alvarez-Loayza, L. F. Alves, P. Ashton, P. Balvanera, M. L. Bastian, P. J. Bellingham, E. van den Berg, L. Bernacci, P. da Conceição Bispo, L. Blanc, K. Böhning-Gaese, P. Boeckx, F. Bongers, B. Boyle, M. Bradford, F. Q. Brearley, M. B.-N. Hockemba, S. Bunyavejchewin, D. C. L. Matos, M. Castillo-Santiago, E. L. M. Catharino, S.-L. Chai, Y. Chen, R. K. Colwell, R. L. Chazdon, C. Clark, D. B. Clark, D. A. Clark, H. Culmsee, K. Damas, H. S. Dattaraja, G. Dauby, P. Davidar, S. J. De Walt, J.-L. Doucet, A. Duque, G. Durigan, K. A. O. Eichhorn, P. V. Eisenlohr, E. Eler, C. Ewango, N. Farwig, K. J. Feeley, L. Ferreira, R. Field, A. T. de Oliveira Filho, C. Fletcher, O. Forshed, G. Franco, G. Fredriksson, T. Gillespie, J.-F. Gillet, G. Amarnath, D. M. Griffith, J. Grogan, N. Gunatilleke, D. Harris, R. Harrison, A. Hector, J. Homeier, N. Imai, A. Itoh, P. A. Jansen, C. A. Joly, B. H. J. de Jong, K. Kartawinata, E. Kearsley, D. L. Kelly, D. Kenfack, M. Kessler, K. Kitayama, R. Kooyman, E. Larney, Y. Laumonier, S. Laurance, W. F. Laurance, M. J. Lawes, I. L. do Amaral, S. G. Letcher, J. Lindsell, X. Lu, A. Mansor, A. Marjokorpi, E. H. Martin, H. Meilby, F. P. L. Melo, D. J. Metcalfe, V. P. Medjibe, J. P. Metzger, J. Millet, D. Mohandass, J. C. Montero, M. de Morisson Valeriano, B. Mugerwa, H. Nagamasu, R. Nilus, S. Ochoa-Gaona, Onrizal, N. Page, P. Parolin, M. Parren, N. Parthasarathy, E. Paudel, A. Permana, M. T. F. Piedade, N. C. A. Pitman, L. Poorter, A. D. Poulsen, J. Poulsen, J. Powers, R. C. Prasad, J.-P. Puyravaud, J.-C. Razafimahaimodison, J. Reitsma, J. R. D. Santos, W. R. Spironello, H. Romero-Saltos, F. Rovero, A. H. Rozak, K. Ruokolainen, E. Rutishauser, F. Saiter, P. Saner, B. A. Santos, F. Santos, S. K. Sarker, M. Satdichanh, C. B. Schmitt, J. Schöngart, M. Schulze, M. S. Suganuma, D. Sheil, E. D. S. Pinheiro, P. Sist, T. Stevart, R. Sukumar, I.-F. Sun, T. Sunderland, H. S. Suresh, E. Suzuki, M. Tabarelli, J. Tang, N. Targhetta, I. Theilade, D. W. Thomas, P. Tchouto, J. Hurtado, R. Valencia, J. L. C. H. van Valkenburg, T. Van Do, R. Vasquez, H. Verbeeck, V. Adekunle, S. A. Vieira, C. O. Webb, T. Whitfeld, S. A. Wich, J. Williams, F. Wittmann, H. Wöll, X. Yang, C. Y. A. Yao, S. L. Yap, T. Yoneda, R. A. Zahawi, R. Zakaria, R. Zang, R. L. de Assis, B. G. Luize, E. M. Venticinque, An estimate of the number of tropical tree species. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7472–7477 (2015).
16. H. Ter Steege, N. C. A. Pitman, D. Sabatier, C. Baraloto, R. P. Salomão, J. E. Guevara, O. L. Phillips, C. V. Castilho, W. E. Magnusson, J.-F. Molino, A. Monteagudo, P. N. Vargas, J. C. Montero, T. R. Feldpausch, E. N. H. Coronado, T. J. Killeen, B. Mostacedo, R. Vasquez, R. L. Assis, J. Terborgh, F. Wittmann, A. Andrade, W. F. Laurance, S. G. W. Laurance, B. S. Marimon, B.-H. Marimon Jr., I. C. G. Vieira, I. L. Amaral, R. Brienen, H. Castellanos, D. C. López, J. F. Duivenvoorden, H. F. Mogollón, F. D. de Almeida Matos, N. Dávila, R. García-Villacorta, P. R. S. Diaz, F. Costa, T. Emilio, C. Levis, J. Schietti, P. Souza, A. Alonso, F. Dallmeier, A. J. D. Montoya, M. T. F. Piedade, A. Araujo-Murakami, L. Arroyo, R. Gribel, P. V. A. Fine, C. A. Peres, M. Toledo, G. A. Aymard C, T. R. Baker, C. Cerón, J. Engel, T. W. Henkel, P. Maas, P. Petronelli, J. Stropp, C. E. Zartman, D. Daly, D. Neill, M. Silveira, M. R. Paredes, J. Chave, D. de Andrade Lima Filho, P. M. Jørgensen, A. Fuentes, J. Schöngart, F. C. Valverde, A. D. Fiore, E. M. Jimenez, M. C. P. Mora, J. F. Phillips, G. Rivas, T. R. van Andel, P. von Hildebrand, B. Hoffman, E. L. Zent, Y. Malhi, A. Prieto, A. Rudas, A. R. Ruschell, N. Silva, V. Vos, S. Zent, A. A. Oliveira, A. C. Schutz, T. Gonzales, M. T. Nascimento, H. Ramirez-Angulo, R. Sierra, M. Tirado, M. N. U. Medina, G. van der Heijden, C. I. A. Vela, E. V. Torre, C. Vriesendorp, O. Wang, K. R. Young, C. Baider, H. Balslev, C. Ferreira, I. Mesones, A. Torres-Lezama, L. E. U. Giraldo, R. Zagt, M. N. Alexiades, L. Hernandez, I. Huamantupa-Chuquimaco, W. Milliken, W. P. Cuenca, D. Pauletto, E. V. Sandoval, L. V. Gamarra, K. G. Dexter, K. Feeley, G. Lopez-Gonzalez, M. R. Silman, Hyperdominance in the amazonian tree flora. *Science* **342**, 1243092 (2013).
17. A. Tovo, S. Suweis, M. Formentin, M. Favretti, I. Volkov, J. R. Banavar, S. Azaele, A. Maritan, Upscaling species richness and abundances in tropical forests. *Sci. Adv.* **3**, e1701438 (2017).
18. A. Tovo, S. Stivanello, A. Maritan, S. Suweis, S. Favaro, M. Formentin, Upscaling human activity data: A statistical ecology approach. *PLOS ONE* **16**, e0253461 (2021).
19. J. Grilli, Macroecological laws describe variation and diversity in microbial communities. *Nat. Commun.* **11**, 4743 (2020).
20. R. V. Solé, D. Alonso, A. McKane, Scaling in a network model of a multispecies ecosystem. *Phys. A Stat. Mech. Appl.* **286**, 337–344 (2000).
21. S. Azaele, S. Pigolotti, J. R. Banavar, A. Maritan, Dynamical evolution of ecosystems. *Nature* **444**, 926–928 (2006).
22. S. Azaele, S. Suweis, J. Grilli, I. Volkov, J. R. Banavar, A. Maritan, Statistical mechanics of ecological systems: Neutral theory and beyond. *Rev. Mod. Phys.* **88**, 035003 (2016).
23. I. Volkov, J. R. Banavar, F. He, S. P. Hubbell, A. Maritan, Density dependence explains tree species abundance and diversity in tropical forests. *Nature* **438**, 658–661 (2005).
24. S. Pueyo, F. He, T. Zillio, The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecol. Lett.* **10**, 1017–1028 (2007).
25. A. E. Magurran, Species abundance distributions: Pattern or process? *Funct. Ecol.* **19**, 177–181 (2005).
26. E. Ser-Giacomi, L. Zinger, S. Malviya, C. De Vargas, E. Karsenti, C. Bowler, S. De Monte, Ubiquitous abundance distribution of non-dominant plankton across the global ocean. *Nat. Ecol. Evol.* **2**, 1243–1249 (2018).
27. S. Pesant, F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot, G. Gorsky, D. Iudicone, E. Karsenti, S. Speich, R. Troublé, C. Dimier, S. Searson, Tara Oceans Consortium Coordinators, Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data.* **2**, 150023 (2015).
28. P. V. Martín, A. Buček, T. Bourguignon, S. Pigolotti, Ocean currents promote rare species diversity in protists. *Sci. adv.* **6**, eaaz9037 (2020).
29. G. Busseni, L. Caputi, R. Piredda, P. Fremont, B. Hay Mele, L. Campese, E. Scalco, C. de Vargas, C. Bowler, F. d'Ovidio, A. Zingone, M. R. D'Alcalà, D. Iudicone, Large scale patterns of marine diatom richness: Drivers and trends in a changing ocean. *Glob. Ecol. Biogeogr.* **29**, 1915–1928 (2020).
30. F. M. Ibarbalz, N. Henry, M. C. Brandão, S. Martini, G. Busseni, H. Byrne, L. P. Coelho, H. Endo, J. M. Gasol, A. C. Gregory, F. Mahé, J. Rigonato, M. Royo-Llonch, G. Salazar, I. Sanz-Sáez, E. Scalco, D. Soviadan, A. A. Zayed, A. Zingone, K. Labadie, J. Ferland, C. Marec, S. Kandels, M. Picheral, C. Dimier, J. Poulain, S. Pisarev, M. Carmichael, S. Pesant, M. Babin, E. Boss, D. Iudicone, O. Jaillon, S. G. Acinas, H. Ogata, E. Pelletier, L. Stemmann, M. B. Sullivan, S. Sunagawa, L. Bopp, C. de Vargas, L. Karp-Boss, P. Wincker, F. Lombard, C. Bowler, L. Zinger, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, F. Lombard, C. Bowler, L. Zinger, Global trends in marine plankton diversity across kingdoms of life. *Cell* **179**, 1084–1097.e21 (2019).
31. E. T. Buitenhuis, M. Vogt, R. Moriarty, N. Bednaršek, S. C. Doney, K. Leblanc, C. le Quéré, Y. W. Luo, C. O'Brien, T. O'Brien, J. Peloquin, R. Schiebel, C. Swan, Maredat: Towards a world atlas of marine ecosystem data. *Earth Syst. Sci. Data* **5**, 227–239 (2013).
32. O. Ragueneau, S. Schultes, K. Bidle, P. Claquin, B. Moriceau, Si and C interactions in the world ocean: Importance of ecological processes and implications for the role of diatoms in the biological pump. *Global Biogeochem. Cycles* **20**, (2006).
33. J. Seckbach, P. Kociolek, *The Diatom World*, vol. 19 (Springer Science & Business Media, 2011).
34. D. K. Stoecker, P. J. Hansen, D. A. Caron, A. Mitra, Mixotrophy in the marine plankton. *Ann. Rev. Mar. Sci.* **9**, 311–335 (2017).
35. S. Malviya, E. Scalco, S. Audic, F. Vincent, A. Veluchamy, J. Poulain, P. Wincker, D. Iudicone, C. De Vargas, L. Bittner, A. Zingone, C. Bowler, Insights into global diatom distribution and diversity in the world's ocean. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E1516–E1525 (2016).
36. R. A. Fisher, A. S. Corbet, C. B. Williams, The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42–58 (1943).
37. I. Volkov, J. R. Banavar, S. P. Hubbell, A. Maritan, Neutral theory and relative species abundance in ecology. *Nature* **424**, 1035–1037 (2003).
38. F. He, Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. *Funct. Ecol.* **19**, 187–193 (2005).
39. P. A. Marquet, M. Tejo, R. Rebolledo, *What Is the Species Richness Distribution* (Princeton Univ. Press, 2020).
40. N. Leibovich, J. Rothschild, S. Goyal, A. Zilman, Phenomenology and dynamics of competitive ecosystems beyond the niche-neutral regimes. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2204394119 (2022).

41. G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, J. J. Egozcue, Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
42. N. Mantel, The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220 (1967).
43. S. Thurner, R. Hanel, P. Klimek, *Introduction to the Theory of Complex Systems* (Oxford Univ. Press, 2018).
44. G. West, Scale: The universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies (reprint edition, Sl: Penguin Books, 2018).
45. J. J. Cullen, Subsurface chlorophyll maximum layers: Enduring enigma or mystery solved? *Ann. Rev. Mar. Sci.* **7**, 207–239 (2015).
46. C. I. Abreu, M. Dal Bello, C. Bunse, J. Pinhassi, J. Gore, Warmer temperatures favor slower-growing bacteria in natural marine communities. bioRxiv 2022.07.13.499956 (2022).
47. J. Carstensen, R. Klais, J. E. Cloern, Phytoplankton blooms in estuarine and coastal waters: Seasonal patterns and key species. *Estuar. Coast. Shelf Sci.* **162**, 98–109 (2015).
48. P. Assmy, V. Smetacek, M. Montresor, C. Klaas, J. Henjes, V. H. Strass, J. M. Arrieta, U. Bathmann, G. M. Berg, E. Breitbarth, B. Cisewski, L. Friedrichs, N. Fuchs, G. J. Herndl, S. Jansen, S. Krägefsky, M. Latasa, I. Peeken, R. Röttgers, R. Scharek, S. E. Schüller, S. Steigenberger, A. Webb, D. Wolf-Gladrow, Thick-shelled, grazer-protected diatoms decouple ocean carbon and silicon cycles in the iron-limited antarctic circumpolar current. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20633–20638 (2013).
49. P. von Dassow, M. Montresor, Unveiling the mysteries of phytoplankton life cycles: Patterns and opportunities behind complexity. *J. Plankton Res.* **33**, 3–12 (2011).
50. T. Wyatt, I. R. Jenkinson, The north atlantic turbine: Views of production processes from a mainly north atlantic perspective. *Fish. Oceanogr.* **2**, 231–243 (1993).
51. G. Lima-Mendez, K. Faust, N. Henry, J. Decelle, S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-Espinosa, S. Roux, F. Vincent, L. Bittner, Y. Darzi, J. Wang, S. Audic, L. Berline, G. Bontempi, A. M. Cabello, L. Coppola, F. M. Cornejo-Castillo, F. d'Ovidio, L. de Meester, I. Ferrera, M. J. Garet-Delmas, L. Guidi, E. Lara, S. Pesant, M. Royo-Llonch, G. Salazar, P. Sánchez, M. Sebastian, C. Souffreau, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, Tara Oceans coordinators, G. Gorsky, F. Not, H. Ogata, S. Speich, L. Stemmann, J. Weissenbach, P. Wincker, S. G. Acinas, S. Sunagawa, P. Bork, M. B. Sullivan, E. Karsenti, C. Bowler, C. de Vargas, J. Raes, Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
52. F. Mahé, T. Rognes, C. Quince, C. de Vargas, M. Dunthorn, Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014).
53. R. S. Etienne, D. Alonso, A dispersal-limited sampling theory for species and alleles. *Ecol. Lett.* **8**, 1147–1156 (2005).
54. Y. Xiang, S. Gubian, B. Suomela, J. Hoeng, Generalized simulated annealing for global optimization: The GenSA package. *R Journal* **5**, 13–28 (2013).

Acknowledgments: E.P. acknowledges P. Padmanabha and S. Azaele for insightful discussions and valuable suggestions. **Funding:** S.S. acknowledges INFN for the Lincoln grant and UNIPD DFA BIRD2021 grant. E.S.-G. is very grateful for support from the Simons Foundation: the Simons Collaboration on Computational Biogeochemical modeling of Marine Ecosystems (CBIOMES #549931). E.P. and L.C. acknowledge a fellowship funded by the Stazione Zoologica Anton Dohrn (SZN) within the SZN-Open University Ph.D. program. E.P. and D.I. acknowledge the AtlantECO project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 862923. **Author contributions:** S.S. and D.I. conceived the research. S.S., M.R., and D.I. designed the analyses. E.P., L.C., and B.H.M. analyzed the data. E.P. and B.H.M. did the computational and statistical analyses. S.S., E.S.-G., and E.P. performed the analytical calculations. All the authors contributed to other aspects of the paper and the writing of the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper, in the Supplementary Materials, and/or at https://doi.org/10.5281/zenodo.8217321. In addition, all data and codes are available at https://zenodo.org/badge/latestdoi/674742723.

Submitted 7 February 2023
Accepted 5 February 2024
Published 8 March 2024
10.1126/sciadv.adh0477