HAL open science

# Large scale patterns of marine diatom richness: Drivers and trends in a changing ocean

Greta Busseni, Luigi Caputi, Roberta Piredda, Paul Fremont, Bruno Hay Mele, Lucia Campese, Eleonora Scalco, Colomban Vargas, Chris Bowler, Francesco d'Ovidio, et al.

## To cite this version:

Greta Busseni, Luigi Caputi, Roberta Piredda, Paul Fremont, Bruno Hay Mele, et al.. Large scale patterns of marine diatom richness: Drivers and trends in a changing ocean. Global Ecology and Biogeography, 2020, 29 (11), pp.1915-1928. 10.1111/geb.13161 . hal-02992584

**HAL Id: hal-02992584**

**https://hal.science/hal-02992584v1**

Submitted on 19 Apr 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Check for updates

Received: 24 November 2019 | Revised: 17 June 2020 | Accepted: 19 June 2020

DOI: 10.1111/geb.13161

RESEARCH PAPERS

Global Ecology and Biogeography A Journal of Macroecology WILEY

# Large scale patterns of marine diatom richness: Drivers and trends in a changing ocean

Greta Busseni[1] | Luigi Caputi[1] | Roberta Piredda[1] | Paul Fremont[1,2] | Bruno Hay Mele[1] | Lucia Campese[1] | Eleonora Scalco[1] | Colomban de Vargas[3,4] | Chris Bowler[5] | Francesco d'Ovidio[6] | Adriana Zingone[1] | Maurizio Ribera d'Alcalà[1] | Daniele Iudicone[1]

[1]Stazione Zoologica Anton Dohrn, Villa Comunale, Naples, Italy

[2]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

[3]Sorbonne Université, CNRS, Station Biologique de Roscoff, Roscoff, France

[4]Research Federation for the study of Global Ocean Systems Ecology and Evolution, Tara Oceans GOSEE, Paris, France

[5]Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Université Paris, Paris, France

[6]Sorbonne Université, CNRS, Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques (LOCEAN-IPSL), Paris, France

**Correspondence**
Daniele Iudicone, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Naples, Italy.
Email: daniele.iudicone@szn.it

**Editor:** Derek Tittensor

## Abstract

**Aim:** Plankton diversity is a pivotal element of marine ecosystem stability and functioning. A major obstacle in the assessment of diversity is the lack of consistency between patterns assessed by molecular and morphological data. This work aims to reconcile the two in a single richness measure, to investigate the environmental drivers affecting this measure, and finally to predict its spatio-temporal patterns.

**Location and time period:** This is a global scale study, based on data collected within the 2009–2013 interval during the *Tara* Oceans expedition.

**Major taxa studied:** The focus of this study is diatoms. They play an important role in several biogeochemical cycles and within marine food webs, and display high taxonomic and functional richness.

**Methods:** We integrate measures of diatom richness across the global ocean using molecular and morphological approaches, giving particular attention to 'the rare biosphere'. We then perform a machine-learning-based analysis of these reconciled patterns to extrapolate diatom richness at the global scale and to identify the main environmental processes governing it. Finally, we model the response of diatom richness to climate change.

**Results:** By filtering out 0.3% of the rarest operational taxonomic units, molecular-based richness patterns show the best possible match with the morphological approach. Temperature, phosphate, chlorophyll *a* and the Lyapunov exponent are the major explainers of these reconciled patterns. Global scale predictions provide a first approximation of the global geography of diatom richness and of the possible impacts of climate change.

**Main conclusions:** Our models suggest that diatom richness is controlled by different processes characteristic of distinct environmental scenarios: lateral mixing in highly dynamic regions, and both nutrient availability and temperature elsewhere. We present herein the effects of these processes on richness and how these same effects

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2020 The Authors. Global Ecology and Biogeography published by John Wiley & Sons Ltd

differ from other diversity indices because of the main component of richness: the rare biosphere.

**KEYWORDS**

diatoms, diversity, machine learning, metabarcoding, microscopy, richness, Tara Oceans

## 1 | INTRODUCTION

The role of plankton diversity in marine ecosystem functioning is recurrently debated in view of ongoing climate change (Beaugrand, Edwards & Legendre, 2010; Tittensor et al., 2010). To better understand the relationship between plankton diversity and its roles in environmental functioning, a more resolved mapping of plankton distributions and diversity in the contemporary ocean is needed. However, diversity assessments are not as straightforward as they may appear. Traditional methods to classify the different units of diversity rely on morphological analyses, from the classical Utermöhl method (Utermöhl, 1958) to the more detailed scanning electron microscopy and transmission electron microscopy (Tomas, 1997) methods. Molecular screening based upon metabarcode-based information is a more recent method to identify units of diversity (Leray & Knowlton, 2016; Zimmermann, Glöckner, Jahn, Enke & Gemeinholzer, 2015). The morphology- and metabarcode-based approaches are differentiated by their measuring unit, which is the species in the traditional morphological sense (e.g., Tomas, 1997) – to the extent to which species can be defined (De Queiroz, 2007) – for the former, and the operational taxonomic unit (OTU) for the latter. However, OTUs, independently of the clustering method implemented for their definition, are not always a good proxy for morphological species, since: (a) different morpho-species may be clustered in the same OTU (Ratnasingham & Hebert, 2013); (b) others may be clustered across multiple OTUs (Ratnasingham & Hebert, 2013); and (c) artefacts [e.g., sequencing errors, polymerase chain reaction (PCR) amplification artefacts, chimaeras; Brown et al., 2015] might be misunderstood as OTUs. The reconciliation of morpho-species-based and OTU-based diversity assessments is one of the current challenges in marine ecology as there could be great advantages for community composition assessments (Muller-Karger et al., 2018). In brief, reconciling the two sources of information would merge their strengths and would provide a more robust estimate of the number of taxa present in a location of interest. Once taxonomic units have been selected, the diversity within an assemblage can be quantified by several indices. Among them, richness is the optimal, unbiased descriptor of how many different species co-occur at a given site at a specific time (Magurran, 1988). This value results from the interplay of biotic interactions and physicochemical factors, providing clues to understand communities at ecological and biological levels. Richness may be controlled by bottom-up processes, as higher nutrient availability fosters higher richness (Dutkiewicz, Follows & Bragg, 2009), but also by water dynamics whenever the local strength of oceanic mixing and transport is sufficiently high (Barton, Dutkiewicz, Flierl, Bragg & Follows, 2010; Lévy, Jahn, Dutkiewicz, Follows & d'Ovidio, 2015). Moreover, richness is the key information to unravel the Hutchinson (1961) plankton paradox. The evidence that many more species than those expected considering the competitive exclusion principle can coexist in a small, homogeneous water volume is still puzzling the scientific community, even though several hypotheses have been proposed (Roy & Chattopadhyay, 2007).

Richness values are dramatically affected by the presence of rare species and, therefore, are extremely sensitive to the sampling and analytical effort (Cermeño, Teixeira, Branco, Figueiras & Marañón, 2014). Rare species, which are intentionally underrated by other metrics of diversity (e.g., Shannon *H* diversity index), represent an incredibly large quota of microbial communities (Caron & Countway, 2009) to the point that they have been named 'the rare biosphere' (Sogin et al., 2006). Even though they have a minor cumulative impact on total abundance, they are the main component of richness and the primary determinant of phylogenetic diversity. The rare biosphere has been interpreted as a vast diversity reservoir, presumably made up of ecologically redundant species that, although maybe on their way to extinction, are able to rapidly interact with the environment and thus quickly drive community structure rearrangements (Caron & Countway, 2009; Lynch & Neufeld, 2015). Nevertheless, the potential contribution of the rare biosphere to ecological resilience is often neglected. Moreover, a fundamental question in ecology and evolution is whether taxonomic and genotypic units are discriminating organisms that we would classify as diverse because of their different responses to the same ecological context, which again leads to the plankton paradox. Setting the number of players in the game in a robust way is, therefore, a prerequisite to answering this question and it can be achieved by producing a consistent, unified picture of the units, even when very rare.

Diatoms (Bacillariophyta) are one of the major phytoplanktonic groups and are among the most diverse classes of organisms on Earth (Mann & Vanormelingen, 2013). While recent studies have already shown how comparison of molecular and morphological taxonomic units can lead to greater analytical precision (Malviya et al., 2016; Zimmermann et al., 2015), we herein aim to integrate the two diatom richness estimation approaches, to assess how co-occurrence of many diatom species varies at the global scale, by adequately considering the different ways the rare biosphere is accounted for by these methods. Our analysis builds on *Tara* Oceans data, which have already demonstrated the potential of metabarcode analysis for several taxa (de Vargas et al., 2015; Le Bescot et al., 2015)

including diatoms (Malviya et al., 2016; Pierella Karlusich, Ibarbalz, & Bowler, 2020). The unified view of diatom richness patterns that we obtain by reconciling morphological and molecular markers allows us, through machine learning approaches (a) to describe diatom variability at the global scale, (b) to analyse how such variability matches with spatial patterns of environmental variables and, finally, as a proof of concept, (c) to attempt a prediction of how climate change might affect the observed richness patterns.

## 2 | MATERIALS AND METHODS

### 2.1 | Metabarcoding data

Metabarcoding data for *Tara* Oceans samples were exploited for the present study. Total nucleic acids (DNA + RNA) were extracted from all the samples, and the hypervariable V9 region of the nuclear 18S ribosomal DNA was amplified through PCR (see Alberti et al., 2017; de Vargas et al., 2015). A quality filtering based on reads 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. Within this dataset, 237,565 V9 diatom-assigned ribotypes were detected (de Vargas et al., 2015). We focus on the 20–180 μm size fraction dataset, which contains 183 net samples encompassing 125 stations sampled at the subsurface (5 m) and 58 of them at deep chlorophyll maximum depth (DCM).

### 2.2 | Bioinformatics pipeline

The initial taxonomic assignation of reads was confirmed and refined using a custom version of the Protist Ribosomal Reference database ($PR^2$; Guillou et al., 2013) containing a selection of new and curated reference sequences available for diatoms in GenBank in December 2018. Taxonomic assignment to diatoms of ribotypes was performed in two steps. We first made a local blast against the custom version of the $PR^2$ database retaining only results showing similarity >90% over a query coverage with the reference > 109 bp. Reference and environmental sequences were aligned with MAFFT v.7 (Katoh, Rozewicki & Yamada, 2017) using the experimental service for large numbers of highly similar and short sequences (https://mafft.cbrc.jp/alignment/server/large.html?aug31). Poorly aligned sequences and hard-to-align blocks within the alignment were removed. Ribotypes taxonomically annotated to diatoms were furthermore filtered through a phylogenetic approach. Phylogenetic analyses were performed using the approximately-maximum likelihood method (Yang, 1994) implemented in the FastTree2 software (Price, Dehal & Arkin, 2010). Ribotypes validated by the taxonomic and phylogenetic check (Supporting Information Data S1) were clustered into OTUs applying the Swarm approach (Mahé, Rognes, Quince, de Vargas & Dunthorn, 2014). Swarm aggregation was performed at different clustering levels (*d*) from 1 to 5, using the standard values for all the other parameters through the SWARM software (Mahé et al., 2014). Moreover, OTUs in the range 95%–99% were calculated using the vsearch distance-based greedy clustering algorithm (method = dgc) through MOTHUR (Rognes, Flouri, Nichols, Quince, & Mahé, 2016; Schloss et al., 2009).

### 2.3 | Morphology-based data

Morphology-based counting was implemented for subsamples of the same size fraction used for molecular analyses (20–180 μm), obtained from a few to more than 100 L of seawater, at surface and DCM depth from the stations in the Cape Agulhas region, the South Atlantic transect, South Pacific Ocean and the Southern Ocean (Malviya et al., 2016). Additional samples were analysed from the Atlantic, Indian and Pacific Oceans, the Mediterranean Sea and the Arctic. Up to 3 mL of each net sample was placed in an Utermöhl chamber. The whole sedimentation chamber bottom was observed and cells (> 5 μm) were identified up to the species level whenever possible using a light inverted microscope (Axiophot200, Carl Zeiss, Oberkochen, Germany) at 400 × magnification (Utermöhl, 1958). The morphology-based richness was computed as the number of different taxa identified in the samples (Supporting Information Data S2).

### 2.4 | Filtering process of molecular data

The process was repeated using ribotypes, or differentially clustered OTUs (see 'Bioinformatics pipeline' in Materials and methods): Swarm OTUs at five different clustering levels (*d*) and vsearch OTUs at five different distance thresholds (*s*). Each molecular sample was filtered at a series of ordered thresholds (from 1% to 100%). Ribotypes or OTUs were discarded from the rarest to the most abundant in the sample in order to progressively exclude an increasing amount of reads. Filtering thresholds were thus measured in terms of the total relative abundance maintained by the filtering. Pairwise Pearson correlations of the richness computed over the 11 types of differently clustered metabarcode information filtered at different thresholds were performed against the morphology-based richness. A false discovery rate *p*-value adjustment of the *p*-value was implemented. The optimal correlation between morphology-based richness and any barcode-based richness at different thresholds was selected as the most significant correlation (adjusted $p$-value < .05) with the highest correlation coefficient (*rho*). This optimal correlation drove the choice of the molecular data type other than of the filtering threshold to be applied to the whole dataset. This latter filtered information will be used for all downstream analysis in this paper. The filtering process is schematized in Supporting Information Figure S1. The cumulative threshold globally applied corresponded to a different absolute threshold for each station, computed as the maximum relative abundance of the discarded OTUs of each station.

### 2.5 | Discarded OTUs

OTU frequency is measured as the number of stations where each OTU is detected. OTUs were classified into three classes according to the filtering results: (a) OTUs always discarded, (b) OTUs always kept and (c) OTUs that are kept or discarded depending on the station. Taxonomic annotation of discarded OTUs was compared to the annotation of the pool of OTUs retained by the filtering.

### 2.6 | Environmental data

Nine environmental variables were considered to investigate the processes behind diatom richness dynamics. Variables included descriptors of hydrodynamic mixing, nutrient availability, temperature and chlorophyll *a* concentration. Local confluence and mixing were estimated through the finite-size Lyapunov exponent, a measure of the front intensification rate computed as the backward-in-time relative separation of water parcels from altimetry-derived surface currents (Lehahn, d'Ovidio & Kohen, 2018). An altimetry-based three-days advection scheme was also applied to 25-km resolution infrared sea surface temperature (SST) images to estimate the SST gradients at km scales at sampling sites (d'Ovidio, De Monte, Alvain, Dandonneau & Levy, 2010). Temperature, salinity, silicate, phosphate and nitrate availability were extracted at 5-m depth from the World Ocean Atlas 2013 (WOA13) database, while iron availability was derived using the pelagic interactions scheme for carbon and ecosystem studies (PISCES)-v2 model (Aumont, Ethé, Tagliabue, Bopp & Gehlen, 2015). Chlorophyll *a* concentration was extracted from the World Ocean Atlas 2001 (WOA01) database (Conkright et al., 2002). When no data were available at the latitude and longitude coordinates of the sample, a search was done within a 2° square around the sampling location and values found within this square were averaged.

### 2.7 | Earth system models

WOA and PISCES-v2 data were used to obtain present-day global environmental conditions. Future environmental conditions were derived by computing the mean of six Earth system models over the 2006–2015 (present-day) and 2090–2099 (end of the century) decades under the greenhouse gas emission scenario RCP8.5. The models included in the analysis are IPSL-CM5A-LR/MR, GFDL-ESM2G/M, MPI-ESM-LR/MR, CESM1-BGC, HadGEM2-ES and NorESM1. We considered the average value of the models as it tends to smooth errors and incongruences between models (Bopp et al., 2013). The delta between the end of the century and the present-day mean models was computed (Supporting Information Figure S2) and added to the WOA and PISCES v2 global scale data to obtain future environmental conditions.

### 2.8 | Machine learning modelling integration

In order to model surface diatom richness, as measured by the optimally filtered molecular information, we integrated four machine learning approaches: boosted regression tree (BRT), random forest (RF), fully connected neural network (NN) and generalized additive models (GAM). The parameterization of each model was optimized (Supporting Information Text S1). The final models were then trained on the whole dataset. Model performance was investigated by computing the mean residuals and the Pearson correlations between the observed and predicted richnesses for the whole dataset (Supporting Information Table S2). Models with a cross-validated root-mean-square deviation below 50 and mean residuals below 30 were selected as significant. Two sets of models were built and optimized: one including all the nine variables previously described to investigate the roles of the variables on the model and a second one excluding the hydrodynamic parameters (SST gradient and Lyapunov exponent) to allow the projection at global scale, as we lack global scale data for these two variables. In both cases, three out of four models were considered as significant, excluding the GAM model from both exercises (Supporting Information Table S2). The projection of the seven-variable models was performed at the global scale using both the present time and the end of the century conditions. The variable importance (Fisher, Rudin & Dominici, 2018) was computed for the nine-variable models using the *DALEX* R package (Biecek, 2018) as the difference between the loss function calculated for validation data with every variable being shuffled and the loss function calculated for the validation dataset.

Furthermore, with the *iBreakDown* R package (Gosiewska & Biecek, 2019) we computed the variables contribution to single predictions and we averaged it across models. We used the nine-variable models to evaluate *Tara* Oceans stations predictions and the seven-variable models to evaluate global scale predictions. Using the seven-variable models, the local variable contributions were computed on global scale data seven more times (one per variable), using each time the future conditions of all the parameters except one, which kept present-day condition. Finally, the difference between the future-present-day variation of richness using all the future conditions and the future-present-day variation of richness using as future conditions all the future parameters except one was computed. All data-mining and statistical analyses were performed in R (version 3.4.1; R Core Team, 2017).

## 3 | RESULTS AND DISCUSSION

### 3.1 | Integration of morphological and metabarcoding counts

Our first step was to strengthen the consistency among diatom richness estimates from molecular and morphological data. There have been previous efforts to compare metabarcoding- and

morphology-based information for diatoms (e.g., Malviya et al., 2016; Piredda et al., 2018; Zimmermann, et al., 2015). However, while previous studies limited their analysis to a descriptive comparison check, here we follow a different approach, proposing a reproducible procedure to reasonably reconcile the two measures. The *Tara* Oceans metabarcode dataset, following a stringent quality filtering, contains 174,267 ribotypes annotated as Bacillariophyta. This study focuses on diatoms from 183 sampling stations addressing the 20–180 μm size fraction, because for those net samples morphology-based data adequately cover richness, thus allowing a reliable comparison with metabarcoding data. This size fraction actually contains a wide variety of diatoms including small-sized ones, which may be trapped because of net clogging and cell aggregation (Leblanc et al., 2018; Piredda et al., 2018). Indeed, because of the wide intraspecific size variability during their life cycle and their pronounced shape anisotropies, size fractionation has a limited applicability to diatoms, which are recorded in all size fractions regardless of their nominal size (Piredda et al., 2018). Within the 20–180 μm size fraction, the clustering produced different numbers of OTUs depending on the clustering threshold (Supporting Information Table S1), resulting in up to 5,830 Swarm OTUs (clustering $d = 1$), and up to 41,381 vsearch OTUs (clustering $s = 99\%$), while morphology-based identification resulted in a list of 256 units at genus or species levels. The most represented genus was *Chaetoceros* (23% of the total OTUs), followed by *Pseudo-nitzschia* (12%) and *Proboscia* (8% of the total OTUs, Supporting Information Data S3). These results only partially match those of Malviya et al. (2016), which were, however, based on all size classes.

The mismatch between metabarcoding- and morphology-based information is due to multiple reasons. Metabarcoding has a much higher resolution and detection power (Leray & Knowlton, 2016), being capable of identifying both cryptic and rare species largely unresolved or missed by morphological methods. On the other hand, the remarkably larger number of units detected by metabarcoding could reflect intraspecific or even intraclonal diversity in some cases, although closely related species are at times clustered in the same OTUs (Piredda et al., 2018). From the quantitative viewpoint, a remarkable match of relative abundances between morphological and metabarcoding information has been observed for diatoms (Malviya et al., 2016; Piredda et al., 2016, 2018), although this is not usually the case for protists (Abad et al., 2016; Massana et al., 2015).

The rationale of the approach herein proposed was to obtain a sufficient covariance between the results of the two methods in order to have a similar spatial pattern of richness while still retaining the high resolution of the metabarcoding. In addition, this will allow comparison with patterns based on microscopic counts (e.g., Righetti, Vogt, Gruber, Psomas & Zimmermann, 2019). With these goals, the datasets obtained through different clustering algorithms were progressively filtered (see Materials and methods, Supporting Information Figure S1), and the resulting richness correlated to the morphology data (Figure 1a, Supporting Information Figure S3a, Data S4). The removal of the least abundant OTUs produced a progressive improvement of the Pearson correlation value across all the different clustering thresholds (Figure 1a), as expected since microscopy-based analyses are much less likely to detect rare species than metabarcoding. The maximum *rho* = .68 was obtained by the Swarm $d = 1$ removing only the cumulative 0.3% of the

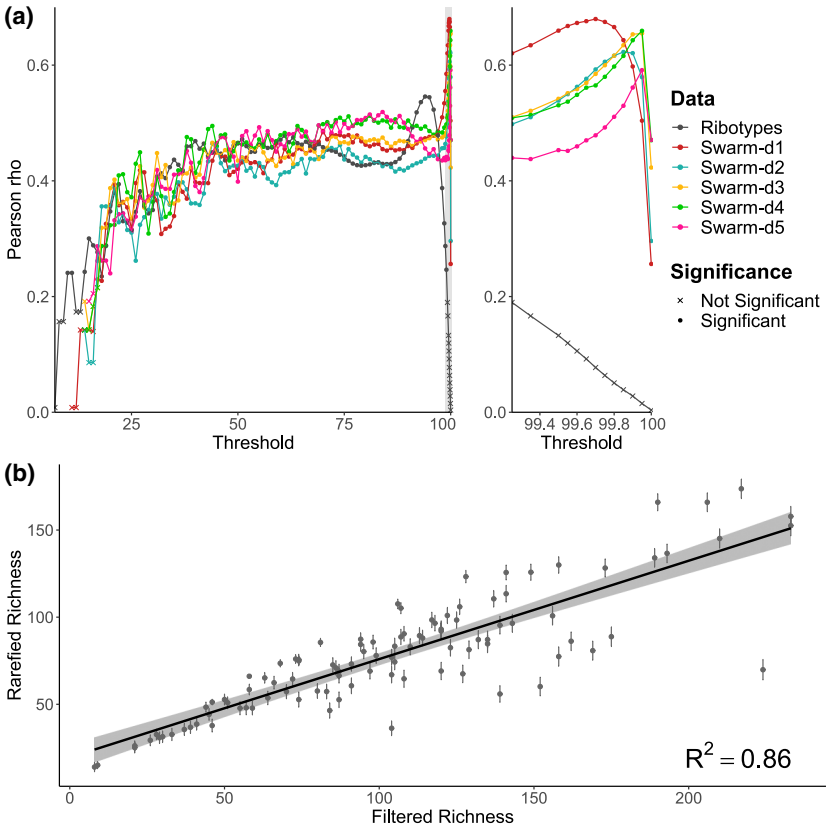


**FIGURE 1** Tuning and validation of the filtering process. (a) Pearson *rho* between diatom morphology-based richness and diatom metabarcode-based richness (according to the colour: ribotypes or Swarm clustering) progressively filtered. Correlation results are ordered along the *x* axis according to the filtering threshold applied. The '×' symbols indicate non-significant correlations (adjusted *p*-value > .05). On the right a blow-up of the grey panel in the 99%–100% threshold range. (b) Scatterplot of diatom richness as measured through the optimal filtering procedure and as obtained by the rarefaction exercise, subsampling to 3,000 reads per sample [Colour figure can be viewed at wileyonlinelibrary.com]

total abundance of OTUs. The station-specific relative abundance under which OTUs were discarded varied across stations with a median of only 0.02% (Supporting Information Figure S3b), highlighting how rare units were largely preserved by the filtering at most sites.

To further test the reliability of our filtering method, results were compared to those obtained with a standard rarefaction analysis, a classical approach to validate richness comparisons (Gotelli & Colwell, 2001). Samples were rarefied to 3,000 reads, which reduced the number of stations to 102 from the initial 183 because of the lower read abundances of 81 stations. The rarefied richness strongly correlated (*rho* = .86) with the filtering process above (Figure 1b).

## 3.2 | The identity of discarded OTUs

Under the hypothesis that filtering mainly discarded genetic variants of more abundant haplotypes, we investigated the nature of the excluded OTUs by assigning both retained and discarded OTUs to named species. Only 24% of the OTUs filtered out belonged to species that were not detected in the retained dataset, while the vast majority of them were presumably variants of more abundant haplotypes at the same station

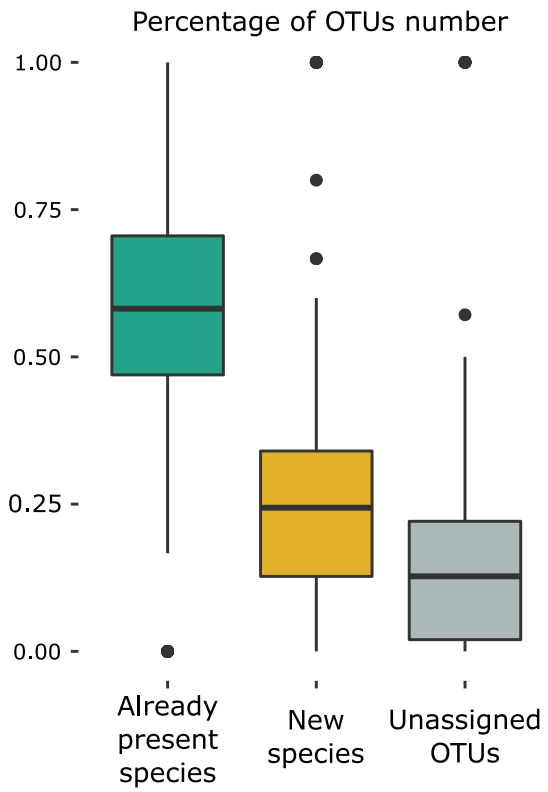


**FIGURE 2** Overview of the identity and distribution of the filtering discards. The boxplots show the discarded operational taxonomic units (OTUs) percentage over the total number of diatom OTUs discarded. The plot aggregates this information calculated in every station and then divided in three classes. The three classes are: (a) the discarded OTUs assigned to species still represented in a station (green), (b) discarded OTUs assigned to species that are not present in the retained set in a given station (yellow) and (c) non-annotated discarded OTUs (grey) [Colour figure can be viewed at wileyonlinelibrary.com]

(Figure 2). Further support for this hypothesis comes from the strong relationship between the abundance of diatoms in a sample and the number of discarded OTUs (Figure 3). This relationship suggests a higher intraspecific variability in diatom-rich stations, which can be explained by (a) a lower detectability of variants in samples with low abundance of diatom reads (Elbrecht, Vamos, Steinke & Leese, 2018), (b) high intraspecific diversity in conditions favourable to diatom growth as a consequence of cell proliferation, sexual reproduction and resting stage germination (Godhe & Rynearson, 2017; Lebret, Kritzberg, Figueroa & Rengefors, 2012), or both.

The distribution of discarded OTUs further supports their interpretation as genetic variants. The majority of OTUs (57%) were discarded by the filtering process and mostly found in only one or two samples (orange in Supporting Information Figure S4a–c). However, 29% of the OTUs discarded in one sampling site were relatively more abundant at other stations, more widely distributed across the sites (green in Supporting Information Figure S4a,b), or both. Only 14% of the OTUs were never discarded, being at times even almost ubiquitous (blue in Supporting Information Figure S4a,b). All the above suggests that the filtering procedure has removed units that would have increased the richness (see next section).

## 3.3 | The reconciled distribution patterns of diatom richness in the global ocean

Because of the above-mentioned relationship between the abundance of diatom sequences in a sample and the number of discarded OTUs, the filtering procedure applied in this study has a higher impact over stations with a large diatom population size and, hence, primary productivity. Indeed, upon filtering, the strong monotonous relationship between richness and relative abundance of reads (as a proxy of diatom productivity) was replaced by a unimodal relationship (red, Figure 3). The decline of richness in diatom-dominated stations recalls the relationships observed between phytoplankton richness and biomass (e.g., observed when excluding rare species by Vallina et al., 2014) generally explained by the dominance of a few species in the case of intense blooms (Mittelbach et al., 2001). However, we rather believe that the decline in richness observed in the cited reports may reflect poor sampling of the rare biosphere, as it is unlikely for conditions favouring the accumulation of specific diatom species to be so unfavourable to other species as to cause their disappearance. Considering all the above, we hypothesize that richness in stations with high diatom abundance is underestimated both in general and by our cautious filtering procedure. Indeed, correcting the filtering results with what we previously detected as likely fake artefacts (discarded OTUs annotated to a species not present in the corresponding retained set and found over the filtering threshold in a second station) led to a trend that confirms the general pattern but likely fixes the supposed underestimate of richness at high abundance (light red, Supporting Information Figure S5a). As this correction did not substantially impact the results (Supporting Information Figure S5b), downstream analyses have been performed on uncorrected filtered richness, so as to keep a better comparability with morphology-based datasets.

We investigated the spatial patterns of diatom richness at the global scale (Figure 4). As designed, the filtering procedure only

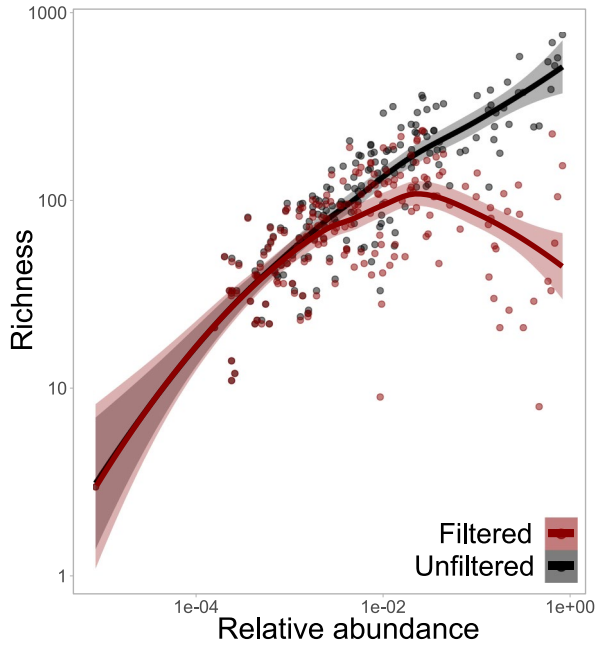


**FIGURE 3** Diatom richness–productivity relationship. The filtered and unfiltered Swarm metabarcoding richness values are related to the relative diatom abundance in the sample, expressed as the number of diatom reads in that sample over the total number of reads sequenced in the same sample. Two regression smoothing lines computed by the loess function fit the two types of data and the relative shading areas reflect the confidence intervals [Colour figure can be viewed at wileyonlinelibrary.com]

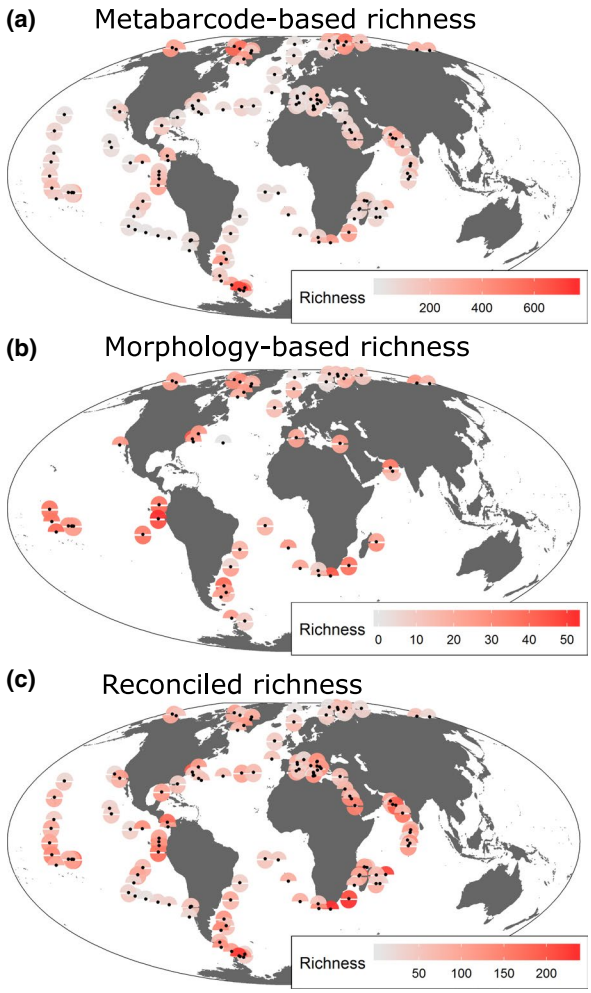


**FIGURE 4** Diatom richness distribution in the global ocean. (a) Diatom richness from *Tara* Oceans sites based on the Swarm metabarcoding in the size-class 20–180 μm. (b) Diatom richness based on morphological data (i.e., light microscopy counts) from the same size-class samples (20–180 μm). (c) Reconciled richness pattern derived by the filtering of the Swarm metabarcoding in size-class 20–180 μm optimized over the morphological observations. All maps employ the Mollweide's projection [Colour figure can be viewed at wileyonlinelibrary.com]

slightly lowered the metabarcoding resolution but allowed the richness pattern to be reconciled with that obtained from morphology-based data. Indeed, both sets of data showed higher richness in the Equatorial Pacific and the Agulhas regions (Figure 4b,c), while the peaks of richness that were molecularly identified at the poles by the unfiltered metabarcoding data (Figure 4a) are now bevelled to the medium–low values observed in the morphology-based data.

To extend the distribution of diatom richness calculated from the *Tara* Oceans stations to the global scale, we applied an integrated machine learning modelling approach, using as predictor variables physical parameters (temperature and salinity), chemical parameters (nitrate, iron, silicate, phosphate) and a proxy of the trophic status of the system (chlorophyll *a*; see Materials and methods). Both observed (Figure 4c) and modelled (Figure 5) results delineated a spatial distribution of diatom richness with maxima in the Tropical Pacific, in the North Indian Ocean and off South Africa. In particular, peaks in diatom richness are observed and modelled near to the well-known upwelling regions associated with the Benguela Current (off southern Africa) and the Humboldt Current (off Peru and Chile).

## 3.4 | Environmental and ecological drivers of diatom richness

A long-standing debate on marine plankton communities is how the environment modulates their variability. In particular, model studies (Follows, Dutkiewicz, Grant & Chisholm, 2007) have consistently proposed hydrodynamic features like fronts as drivers of plankton richness, along with traditional variables like nutrients and temperature. To further explore this issue, a second set of machine learning models was applied, including, in addition to the seven physico-chemical and trophic variables, two hydrodynamic variables: SST gradient and finite-size Lyapunov exponents, which detect fronts as kinematic boundaries (confluences) among contrasted hydrodynamic regions (d'Ovidio et al., 2010). Three models (BRT, RF and NN) out of four had overall high prediction power (Supporting Information Table S2). The scenario slightly changes among models but there is a notable coherence in defining temperature, followed by the Lyapunov index, chlorophyll *a* and phosphate, as the most important variables (Supporting Information Figure S6).

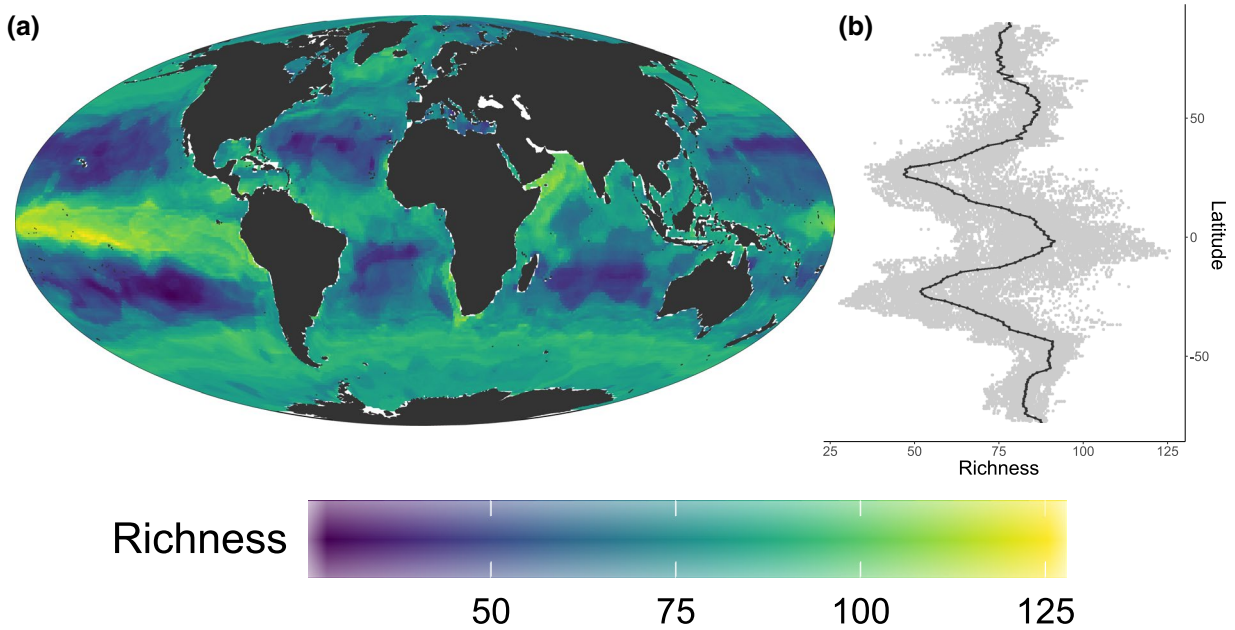


**FIGURE 5** Global prediction of diatom richness. (a) World scale map (Mollweide's projection) of predicted present-day annual mean diatom richness computed at each grid point as the average of the predicted richness of the random forest (RF) model. Present-day environmental conditions are provided by the World Ocean Atlas database with the addition of pelagic interactions scheme for carbon and ecosystem studies-v2 biogeochemical model information for iron. Panel (b) shows the latitudinal distribution of the same measure of diatom richness depicted in panel (a). The black line shows the latitudinal median computed for every degree of latitude [Colour figure can be viewed at wileyonlinelibrary.com]

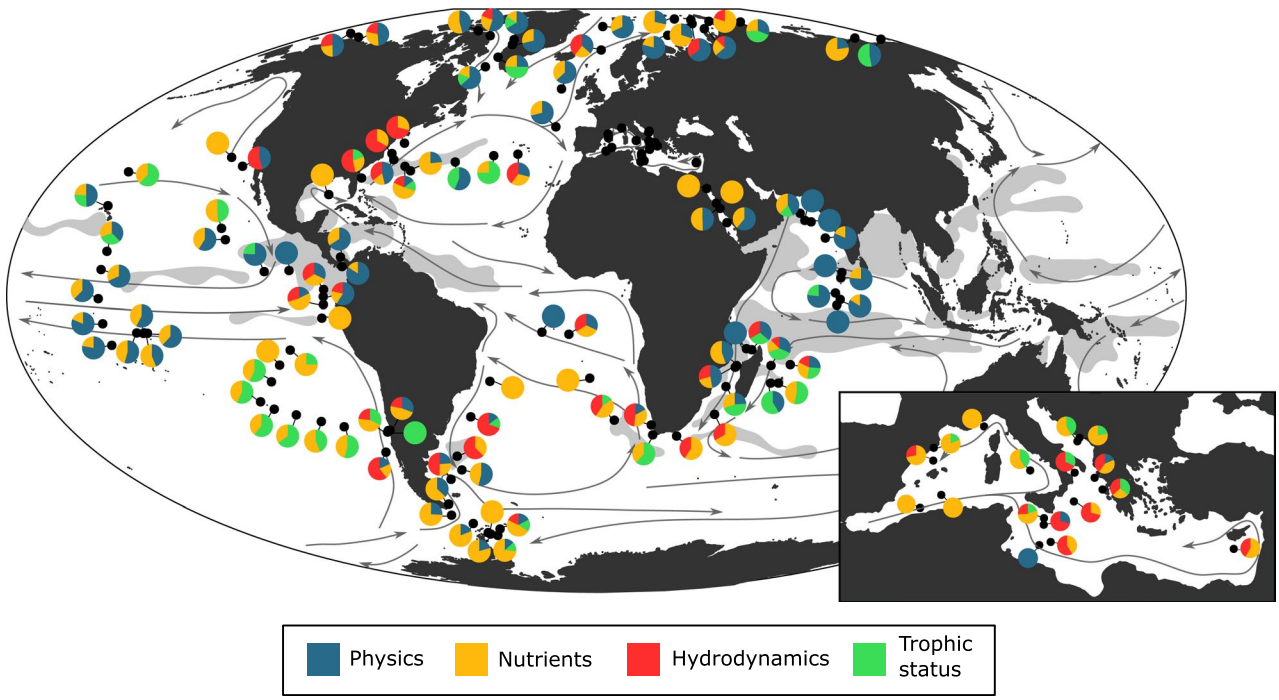


**FIGURE 6** Relative contribution of environmental variables to the prediction of diatom richness at the *Tara* Oceans stations. Only contributions higher than 20% are taken into account and results from the three employed models [boosted regression tree (BRT), neural network (NN) and random forest (RF)] are aggregated. The global scale map employs the Mollweide's projection. In the background the main oceanic currents are depicted by arrows and the areas of high lateral diffusivity (according to Abernathey & Marshall, 2013) are represented by light grey areas [Correction added on 24 August 2020 after first online publication: Figure 6 has been corrected to include the missing pie in this version.] [Colour figure can be viewed at wileyonlinelibrary.com]

Based on these results, we hypothesize diatom richness to be influenced by different processes that vary according to the specific hydrographic structure, nutrient availability and hydrodynamics of the site (Figure 6). The spatial distribution of the contributions of major variables (>20%, Figure 6) indeed shows clear geographical zonation. Temperature, the most crucial variable in the model (Supporting Information Figure S6), is dominant at the extremes of its range, mostly in the Arctic and tropical regions. More ambiguous is the case of the semi-enclosed basins such as the Mediterranean and Red Seas where salinity acquires in some parts an important role, sometimes concealed by contributions of other factors, due to the smaller scales of variability of these basins (e.g., Malanotte-Rizzoli et al., 2014). In oligotrophic oceanic regions, high chlorophyll *a* may be a consequence of equatorial upwelling, which may allow phytoplankton accumulation, locally increasing the number of *r* strategists species, sensu Margalef, as diatoms are supposed to be. Nutrients show high importance across the global ocean. Interestingly, a predominance of hydrodynamic predictors is observed in regions characterized by substantial lateral transfer due to warm currents such as the Brazil Current, the Norwegian Current, the Agulhas Current and the strong Gulf Stream.

### 3.5 | Projecting richness distribution in the future ocean

Having characterized and ranked the environmental variables affecting diatom richness distribution patterns, we analysed the impact of the climate change predicted by models. To this goal we applied to the WOA global dataset the delta values of environmental predictions between 2006–2015 (present-day) and 2090–2099 (end of the century) Earth system models (Supporting Information Figure S7; see Materials and methods). Our results suggest that climate change may lead to a general decrease of diatom richness (with a mean negative variation of 4%, down to a negative minimum of 56%), with a narrowing of hotspot regions (Figure 7a, e.g., Tropical Pacific) and a relocation of richer communities towards the poles. A remarkable example are the peaks of increased richness in the subtropical zones, which at present display very low richness (Figure 7b, Supporting Information Figure S7).

We analysed the impact of the different environmental changes through computation of the contribution of the different variables on the global scale prediction at the present time and at the end of the century (Figure 8a). Strongest positive contributions were observed for nitrate, whereas lower contributions were detected for iron and salinity, which were identified as the least influential variables in the models. Widespread contributions, both positive and negative, were observed for temperature, nitrates and silicates, highlighting a major role played by both physics and the trophic status at individual sites. Moreover, comparing the present day with the end of the century predictions, a substantial shift in temperature contribution was detected from positive to negative contributions. A similar but minor shift is observed for nitrate. This difference supports the fact that nitrate availability, together with its interaction with temperature, may dramatically affect diatom sensitivity to climate change (Thomas, Kremer, Klausmeier & Litchman, 2012). Interestingly, contribution

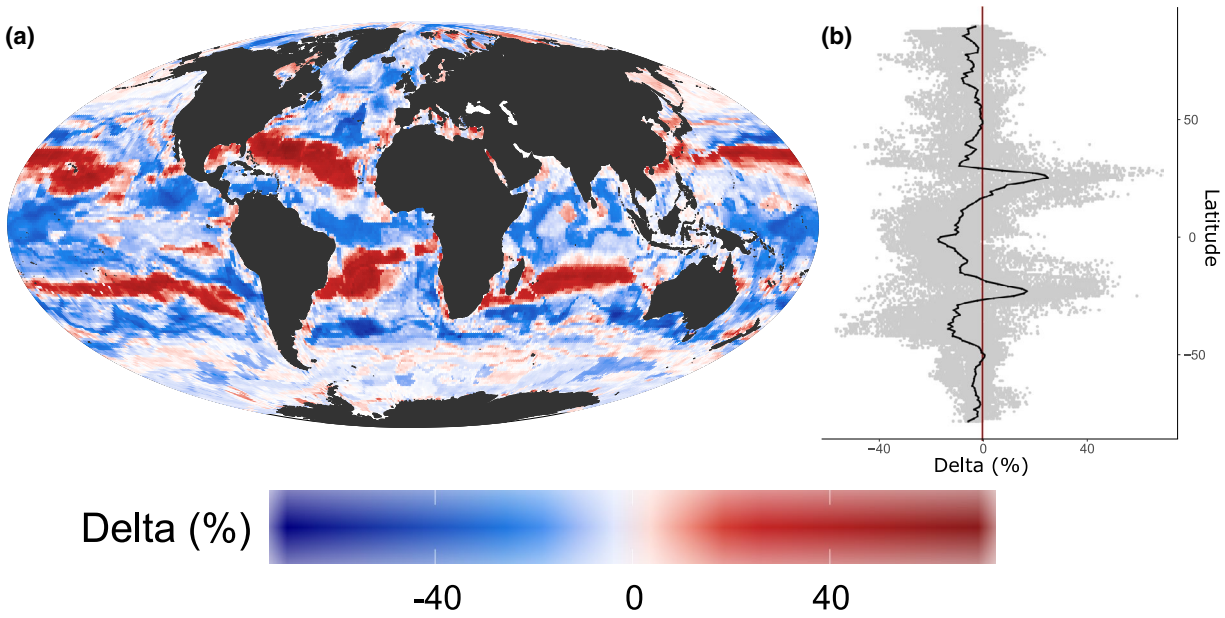


**FIGURE 7** Global future prediction of diatom richness. (a) World scale map (Mollweide's projection) of the percentage of variation of richness estimates (delta) from present-day to the end of the century, predicted using the random forest (RF) machine learning technique. The percentage of variation is computed as the difference between the two environmental conditions over the present-day conditions. (b) Latitudinal distribution of the percentage of variation of richness depicted in panel (a). The black line shows the latitudinal median computed for each degree of latitude while the red line represents the null variation, i.e., where the delta is equals to zero and there is no difference between the predictions [Colour figure can be viewed at wileyonlinelibrary.com]

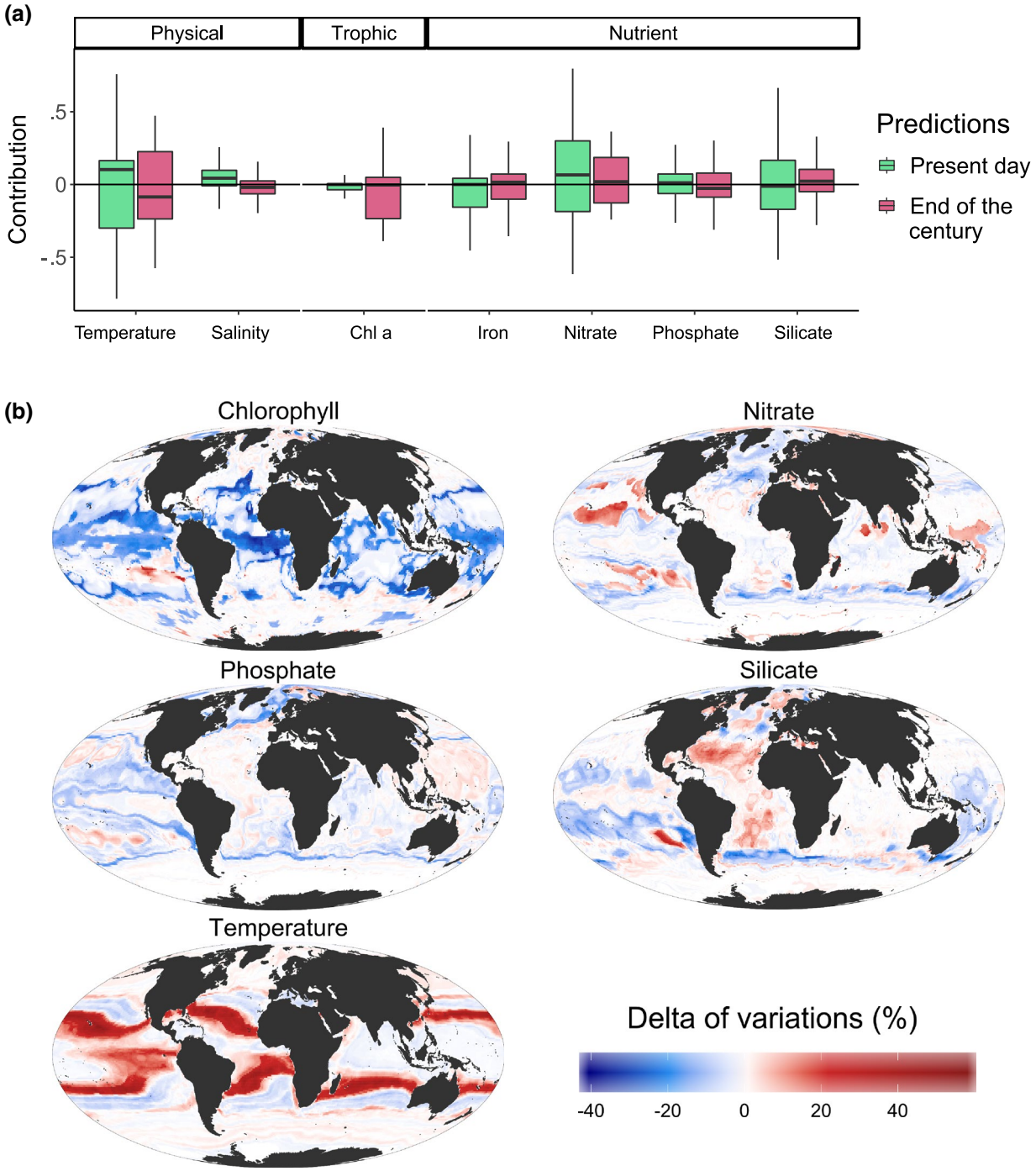


**FIGURE 8** Contribution of the environmental variables to future predictions. In panel (a) the boxplots of the mean of each variable contribution applied by the three machine learning hydrodynamic-free models [boosted regression tree (BRT), neural network (NN) and random forest (RF)] to predict diatom richness at a global scale at the present day (green) and end of the century (red), taking into account only the grid available for both times. Outliers have been excluded. In panel (b) the results of the RF predictions are compared in every sub-panel to the prediction using future conditions for all the variables except one, for which present-day conditions are employed. The delta value of variations are expressed in percentage. They corresponds to the difference between the future-present-day predictions variation (see Figure 7) and the future-present-day predictions variation using as future conditions all the future variables except one. Negative values correspond to regions where climate change variations of the variable lead to a decrease of future richness, and vice versa [Colour figure can be viewed at wileyonlinelibrary.com]

variations were detected also for chlorophyll *a* and silicates, suggesting that climate change may affect the previously documented relationship between diatom biomass and their richness.

To investigate the local impact of single environmental variables in future predictions we investigated the sensitivity of the model to future projected changes of each explanatory variable (see Materials and methods, Figure 8b). The variables locally playing a significant role in changing diatom richness in future conditions are chlorophyll *a*, with a mostly negative role and a maximum effect in the tropics, and temperature, which exhibits highly positive longitudinal bands in the subtropics, that is, in oligotrophic regions. This patterning corresponds to a specific temperature range (Supporting Information Figure S8), suggesting that higher temperatures will have a direct impact in very low richness regions (Figure 5). Diatom richness strongly deviates from predictions from metabolic theories (Supporting Information Figure S8, Righetti et al., 2019). In turn, this implies that predictions based on temperature only do not capture the complexity of the response to climate change (Thomas et al., 2012). Phosphate and nitrate show similar patterns in their impact, displaying negative contributions moving from the North Atlantic to the Arctic. Conversely, silicate shows a complementary pattern with positive impacts on most of the Atlantic Ocean and the Arctic as well.

### 3.6 | Towards a unified picture

The comparison between metabarcode- and morphology-based diatom richness estimates in the global *Tara* Oceans survey displayed over a 20:1 ratio between the former and the latter. Such imbalance highlights the need to clearly specify which kind of richness is herein being studied. Since our goal was to characterize spatial patterns of diatom richness globally, we decided to integrate both types of information assuming that morphological and molecular data could not be uncorrelated. To that aim, we selected the threshold below which the information provided by the two estimates was consistent: this allowed us to include most of the rare biosphere and to reproduce the expected unimodal relationship between diatom abundance and observed richness. This approach is not purely methodological since our filtering procedure allowed a large number of the OTUs to be kept, those above the empirically-found threshold of abundance, which are cumulatively structured in space very much like the canonical, morphology-based species. This, in turn, suggests that this reconciled OTU dataset follows the same ecological patterns of the canonical species on which our knowledge on diatom ecology is largely based, while being one order of magnitude richer.

From an evolutionary perspective, species richness may result either from neutral or adaptive processes. While the former could pass undetected by microscopy counts, the latter might lead to different morphotypes and thus increase also microscopy-based richness. We posit that microscopic counts always tend to underestimate or miss the rare biosphere. Furthermore, even without equating OTUs to morphotypes, the OTUs' distribution is a better representation of the diversity of a sample, in the sense of the number of distinct taxonomic units, than microscopic counts.

Global estimated patterns of phytoplankton richness, although not numerous, have been obtained by model simulations (e.g., Barton et al., 2010), or from direct observations (e.g., Righetti et al., 2019; or Ibarbalz et al., 2019 based on the same dataset). Our spatial patterns are based on global-scale data using different machine learning methods that consider not only the intrinsic structure of spatial variance, as in geostatistical methods, but also other variables that might affect the richness, but without assuming any driving mechanism as in biogeochemical models. Our maps mimic the global richness patterns of phytoplankton distributions obtained by biogeochemical models (e.g., Barton et al., 2010), albeit with a few discrepancies. We interpret some of them as due to inadequate coverage of *Tara* Oceans sampling across specific regions, particularly when those regions have characteristics not present elsewhere (e.g., the Antarctic region). This obviously weakens the predictive potential of statistical models. For areas with sufficient coverage, for example, the Gulf Stream, two upwelling systems and the high nutrient low chlorophyll regions in the subtropical Pacific, our extrapolations are very close to the results of the biogeochemical model. In addition, our analysis, and the derived extrapolations, allowed two remarkable features to be captured: a reduction of richness within tropical and towards polar oceans and a high richness within intertropical and temperate regions (Figure 5b). The former sets a significant difference between diatoms and both phyto- (Righetti et al., 2019; Supporting Information Figure S8) and whole plankton (Ibarbalz et al., 2019) distribution patterns. Diatoms behave similarly to the rest of phytoplankton in the equatorial regions or particular tropical hotspots regions such as the Indian Ocean and the Peru Current upwelling. While diatoms are assumed to be highly diverse in upwelling regions, the high richness in equatorial areas is remarkable and hints at the presence of a suite of strategies to cope with substrate limitations of different origins (e.g., Caputi et al., 2019, Kemp & Villareal, 2018). In other tropical regions, where other phytoplankton display high richness, possibly due to higher metabolic rates (Righetti et al., 2019), diatoms display a clear minimum; in temperate and up to the subpolar regions, where other phytoplankton start their decline in richness (Ibarbalz et al., 2019; Righetti et al., 2019), diatoms peak again. We hypothesize that this behaviour might also be explained by the desynchronized seasonal cycles and by the multiphase life strategies of many diatoms. Even assuming that we underestimated the richness in high latitude regions, as suggested by the correction based on taxonomic annotation (Supporting Information Figure S5), diatom richness in these areas seems lower but still relatively higher than other phytoplankton (Righetti et al., 2019). This matches with the general perception of diatoms optimally thriving in these regions.

At the first order of approximation, the observed global patterns can be explained by three different scenarios in terms of environmental variables or processes. In regions highly active in horizontal hydrodynamics (maxima in Lyapunov exponents) richness is likely enhanced by the confluence of fronts as conjectured by d'Ovidio et al. (2010), or supported by model studies (Lévy et al., 2015). A

second scenario is detected in oligotrophic regions, where nutrient availability (Dutkiewicz et al., 2009) strictly controls the phyto biomass, that is, the chlorophyll *a*, and the richness, herein at its minimum. Finally, the last scenario is found at the extremes of the temperature gradients, in tropical and polar regions, where temperature stands out as a major explainer of diatom richness. This evidence seems to support the metabolic theory hypothesis (Righetti et al., 2019). We believe that the identification of temperature as the main predictor is due to the similar richness observed at the extremes of its range and is not a direct indicator of the processes behind richness increase, likely driven by complex dynamics.

Intriguingly, machine learning results predict a richness increase in regions where stratification and oligotrophy should increase in a global warming scenario. This forecast seems to contradict the prevalent views on diatoms as adapted to nutrient-rich and turbulent environments but we think that this is not necessarily a contradiction. In fact, rather than in biomass our machine learning approach predicts an increase in richness, which is in line with the view that diatoms may persist in oligotrophic areas because they may quickly respond to episodic nutrient pulses, as proposed by McCarthy and Goldman (1979) and recently also shown in other *Tara* Oceans studies by Malviya et al. (2016) and Caputi et al. (2019). The latter authors also demonstrated a broad suite of responses to the same environmental perturbation, which highlights the capability of different diatom species to occupy a wide range of ecological niches. This evidence expands the potential set of conditions under which diatoms may thrive, warranting a higher richness than the one expected by the prevailing view, which is mainly based on well-known, opportunistic coastal species.

Moreover, we point out how the latitudinal variation of diatom richness herein described differs significantly from the pattern of the Shannon diversity index shown by Ibarbalz et al. (2019) for the same group. This highlights the different meanings of the richness (how many distinct units) and the Shannon index (how abundance of units is distributed) in characterizing diversity and explains why they cannot be simply compared. Their differences show that the relative weight of rare versus more abundant taxa varies latitudinally, implicitly supporting our first scenario, and fosters a more in-depth study of the rare biosphere.

## 4 | CONCLUDING REMARKS

Machine learning tools are useful to integrate extensive, multivariate datasets to set the starting point of a mechanistic interpretation of patterns that should complement the evidence provided by them.

Molecular data strongly suggest that diatoms are less exposed to a temperature-dependent latitudinal decrease in richness and, supported also by microscopic counts, 'the variation of richness with latitude deviates from the pattern of most plankton.

Overall, our analysis confirms and reinforces that even for a single phytoplankton group, the apparent paradox of many species coexisting in the same water parcel holds true (Hutchinson, 1961). There is no unique driver of such patterns, with non-equilibrium (first scenario), hump-shaped richness–productivity relationship (second scenario), and latitudinal dependence (third scenario) always concurring, albeit with different geographical patterns, to modulate richness.

## ACKNOWLEDGMENTS

We thank Dr Olivier Jaillon and Dr Lucie Zinger for their suggestions. We thank Dr Marion Gehlen for her support on analysis of the climate model outputs. G.B. and L.Cam acknowledge a fellowship funded by the Stazione Zoologica Anton Dohrn (SZN) within the SZN-Open University Ph.D. program. R.P. was supported by the Italian MIUR Flagship Project RITMARE and the European Union's Horizon 2020 Research and Innovation Program EMBRIC (GA 654008). E.S. was partially supported by a grant from the Ministero dell'Istruzione dell'Università e della Ricerca RITMARE project, and by a postdoc fellowship 'Lina Rizzo' by the Accademia Nazionale Dei Lincei (Rome, Italy). F.d'O. acknowledges support by the NASA/CNES TOSCA BIOSWOT-AdAC project.

## DATA AVAILABILITY STATEMENT

All the data exploited to perform this study are available on Dryad: https://doi.org/10.5061/dryad.wh70rxwk6.

## ORCID

*Greta Busseni* https://orcid.org/0000-0001-6307-5366
*Luigi Caputi* https://orcid.org/0000-0002-5724-1943
*Bruno Hay Mele* https://orcid.org/0000-0001-5579-183X

## REFERENCES

Abad, D., Albaina, A., Aguirre, M., Laza-Martínez, A., Uriarte, I., Iriarte, A., … Estonba, A. (2016). Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy. *Marine Biology*, *163*(7), 1–13. https://doi.org/10.1007/s00227-016-2920-0

Abernathey, R. P., & Marshall, J. (2013). Global surface eddy diffusivities derived from satellite altimetry. *Journal of Geophysical Research: Oceans*, *118*(2), 901–916. https://doi.org/10.1002/jgrc.20066

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., … Wincker, P. (2017). Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition. *Scientific Data*, *4*, 1–20. https://doi.org/10.1038/sdata.2017.93

Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., & Gehlen, M. (2015). PISCES-v2: An ocean biogeochemical model for carbon and ecosystem studies. *Geoscientific Model Development*, *8*(8), 2465–2513. https://doi.org/10.5194/gmd-8-2465-2015

Barton, A. D., Dutkiewicz, S., Flierl, G., Bragg, J., & Follows, M. J. (2010). Patterns of diversity in marine phytoplankton. *Science*, *327*(5972), 1509–1511. https://doi.org/10.1126/science.1184961

Beaugrand, G., Edwards, M., & Legendre, L. (2010). Marine biodiversity, ecosystem functioning, and carbon cycles. *Proceedings of the National Academy of Sciences USA*, *107*(22), 10120–10124. https://doi.org/10.1073/pnas.0913855107

Biecek, P. (2018). DALEX: Explainers for complex predictive models. *The Journal of Machine Learning Research*, *19*(1), 3245–3249.

Bopp, L., Resplandy, L., Orr, J. C., Doney, S. C., Dunne, J. P., Gehlen, M., … Vichi, M. (2013). Multiple stressors of ocean ecosystems in the 21st

century: Projections with CMIP5 models. *Biogeosciences*, *10*(2013), 6225–6245. https://doi.org/10.5194/bg-10-6225-2013

Brown, S. P., Veach, A. M., Rigdon-Huss, A. R., Grond, K., Lickteig, S. K., Lothamer, K., … Jumpponen, A. (2015). Scraping the bottom of the barrel: Are rare high throughput sequences artifacts? *Fungal Ecology*, *13*, 221–225. https://doi.org/10.1016/j.funeco.2014.08.006

Caputi, L., Carradec, Q., Eveillard, D., Kirilovsky, A., Pelletier, E., Pierella Karlusich, J. J., … Iudicone, D. (2019). Community-level responses to iron availability in open ocean planktonic ecosystems. *Global Biogeochemical Cycles*, *33*(3), 391–419. https://doi.org/10.1029/2018gb006022

Caron, D. A., & Countway, P. D. (2009). Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquatic Microbial Ecology*, *57*(3), 227–238. https://doi.org/10.3354/ame01352

Cermeño, P., Teixeira, I. G., Branco, M., Figueiras, F. G., & Marañón, E. (2014). Sampling the limits of species richness in marine phytoplankton communities. *Journal of Plankton Research*, *36*(4), 1135–1139. https://doi.org/10.1093/plankt/fbu033

Conkright, M. E., Locarnini, R. A., Garcia, H. E., O'Brien, T. D., Boyer, T. P., Stephens, C., & Antonov, J. I. (2002). World Ocean Atlas 2001: Objective analyses, data statistics, and figures CD-ROM documentation. *National Oceanographic Data Center Internal Report (NOAA Atlas NESDIS)*.

d'Ovidio, F., De Monte, S., Alvain, S., Dandonneau, Y., & Levy, M. (2010). Fluid dynamical niches of phytoplankton types. *Proceedings of the National Academy of Sciences USA*, *107*(43), 18366–18370. https://doi.org/10.1073/pnas.1004620107

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology*, *56*(6), 879–886. https://doi.org/10.1080/10635150701701083

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., … Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237). https://doi.org/10.1126/science.1261605

Dutkiewicz, S., Follows, M. J., & Bragg, J. G. (2009). Modeling the coupling of ocean ecology and biogeochemistry. *Global Biogeochemical Cycles*, *23*(4), 1–15. https://doi.org/10.1029/2008GB003405

Elbrecht, V., Vamos, E. E., Steinke, D., & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*, *6*, e4644. https://doi.org/10.7717/peerj.4644

Fisher, A., Rudin, C., & Dominici, F. (2018). All Models are wrong but many are useful: Variable Importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint arXiv:1801.01489*

Follows, M. J., Dutkiewicz, S., Grant, S., & Chisholm, S. W. (2007). Emergent biogeography of microbial communities in a model ocean. *Science*, *315*(5820), 1843–1846. https://doi.org/10.1126/science.1138544

Godhe, A., & Rynearson, T. (2017). The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728), 20160399. https://doi.org/10.1098/rstb.2016.0399

Gosiewska, A., & Biecek, P. (2019). iBreakDown: Uncertainty of model explanations for non-additive predictive models. *arXiv preprint arXiv:1903.11420*

Gotelli, N. J., & Colwell, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, *4*(4), 379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., … Christen, R. (2013). The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604. https://doi.org/10.1093/nar/gks1160

Hutchinson, G. E. (1961). The paradox of the plankton. *The American Naturalist*, *95*(882), 137–145. https://doi.org/10.1086/282171

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., … Zinger, L. (2019). Global trends in marine plankton diversity across kingdoms of life. *Cell*, *179*(5), 1084–1097. https://doi.org/10.1016/j.cell.2019.10.008

Katoh, K., Rozewicki, J., & Yamada, K. D. (2017). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1–7. https://doi.org/10.1093/bib/bbx108

Kemp, A. E. S., & Villareal, T. A. (2018). The case of the diatoms and the muddled mandalas: Time to recognize diatom adaptations to stratified waters. *Progress in Oceanography*, *167*(2018), 138–149. https://doi.org/10.1016/j.pocean.2018.08.002

Le Bescot, N., Mahé, F., Audic, S., Dimier, C., Garet, M.-J., Poulain, J., … Siano, R. (2015). Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. *Environmental Microbiology*, *18*(2), 609–626. https://doi.org/10.1111/1462-2920.13039

Leblanc, K., Quéguiner, B., Diaz, F., Cornet, V., Michel-Rodriguez, M., Durrieu De Madron, X., … Conan, P. (2018). Nanoplanktonic diatoms are globally overlooked but play a role in spring blooms and carbon export. *Nature Communications*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-03376-9

Lebret, K., Kritzberg, E. S., Figueroa, R., & Rengefors, K. (2012). Genetic diversity within and genetic differentiation between blooms of a microalgal species. *Environmental Microbiology*, *14*(9), 2395–2404. https://doi.org/10.1111/j.1462-2920.2012.02769.x

Lehahn, Y., d'Ovidio, F., & Koren, I. (2018). A satellite-based Lagrangian view on phytoplankton dynamics. *Annual Review of Marine Science*, *10*, 99–119. https://doi.org/10.1146/annurev-marine-121916-063204

Leray, M., & Knowlton, N. (2016). Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150331. https://doi.org/10.1098/rstb.2015.0331

Lévy, M., Jahn, O., Dutkiewicz, S., Follows, M. J., & d'Ovidio, F. (2015). The dynamical landscape of marine phytoplankton diversity. *Journal of the Royal Society Interface*, *12*(111), 20150481. https://doi.org/10.1098/rsif.2015.0481

Lynch, M. D. J., & Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology*, *13*(4), 217–229. https://doi.org/10.1038/nrmicro3400

Magurran, A. E. (1988). *Ecological diversity and its measurement*. Princeton, NJ: Princeton University Press.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ*, *2*, e593. https://doi.org/10.7717/peerj.593

Malanotte-Rizzoli, P., Artale, V., Borzelli-Eusebi, G. L., Brenner, S., Civitarese, G., Crise, A., … Triantafyllou, G. (2014). Physical forcing and physical/biochemical variability of the Mediterranean Sea: A review of unresolved issues and directions for future research. *Ocean Science Discussions*, *10*, 1205–1280. https://doi.org/10.5194/osd-10-1205-2013

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., … Bowler, C. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences USA*, *113*(11), E1516–E1525. https://doi.org/10.1073/pnas.1509523113

Mann, D. G., & Vanormelingen, P. (2013). An inordinate fondness? The number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology*, *60*(4), 414–420. https://doi.org/10.1111/jeu.12047

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., … de Vargas, C. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing.

*Environmental Microbiology*, *17*(10), 4035–4049. https://doi.org/10.1111/1462-2920.12955

McCarthy, J. J., & Goldman, J. C. (1979). Nitrogenous nutrition of marine phytoplankton in nutrient-depleted waters. *Science*, *203*(4381), 670–672.

Mittelbach, G. G., Steiner, C. F., Scheiner, S. M., Gross, K. L., Reynolds, H. L., Waide, R. B., … Gough, L. (2001). What is the observed relationship between species richness and productivity? *Ecology*, *82*(9), 2381–2396. https://doi.org/10.1890/0012-9658(2001)082[2381:WITORB]2.0.CO;2

Muller-Karger, F. E., Miloslavich, P., Bax, N. J., Simmons, S., Costello, M. J., Sousa Pinto, I., … Geller, G. (2018). Advancing marine biological observations and data requirements of the complementary Essential Ocean Variables (EOVs) and Essential Biodiversity Variables (EBVs) frameworks. *Frontiers in Marine Science*, *5*, 1–15. https://doi.org/10.3389/fmars.2018.00211

Pierella Karlusich, J. J., Ibarbalz, F. M., & Bowler, C. (2020). Phytoplankton in the *Tara* Ocean. *Annual Review of Marine Science*, *12*, 233–265.

Piredda, R., Claverie, J.-M., Decelle, J., de Vargas, C., Dunthorn, M., Edvardsen, B., … Zingone, A. (2018). Diatom diversity through HTS-metabarcoding in coastal European seas. *Scientific Reports*, *8*(1), 18059. https://doi.org/10.1038/s41598-018-36345-9

Piredda, R., Tomasino, M. P., D'Erchia, A. M., Manzari, C., Pesole, G., Montresor, M., … Zingone, A. (2016). Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. *FEMS Microbiology Ecology*, *93*(1), fiw200. https://doi.org/10.1093/femsec/fiw200

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: The barcode index number (BIN) system. *PLoS ONE*, *8*(7), e66213. https://doi.org/10.1371/journal.pone.0066213

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Righetti, D., Vogt, M., Gruber, N., Psomas, A., & Zimmermann, N. E. (2019). Global pattern of phytoplankton diversity driven by temperature and environmental variability. *Science Advances*, *5*(5), eaau6253. https://doi.org/10.1126/sciadv.aau6253

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Roy, S., & Chattopadhyay, J. (2007). Towards a resolution of 'the paradox of the plankton': A brief overview of the proposed mechanisms. *Ecological Complexity*, *4*(1–2), 26–33. https://doi.org/10.1016/j.ecocom.2007.02.016

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., … Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences USA*, *103*(32), 12115–12120. https://doi.org/10.1073/pnas.0605127103

Thomas, M., Kremer, C., Klausmeier, C. A., & Litchman, E. (2012). A global pattern of thermal adaptation in marine phytoplankton. *Science*, *338*(338), 1085–1088. https://doi.org/10.1126/science.1224836

Tittensor, D. P., Mora, C., Jetz, W., Lotze, H. K., Ricard, D., Berghe, E. V., & Worm, B. (2010). Global patterns and predictors of marine biodiversity across taxa. *Nature*, *466*(7310), 1098–1101.

Tomas, C. R. (Ed.). (1997). *Identifying marine phytoplankton*, San Diego, CA: Elsevier. https://doi.org/10.1029/98EO00066

Utermöhl, H. (1958). Zur Vervollkommnung der quantitativen Phytoplankton-Methodik. *Internationale Vereinigung für Theoretische und Angewandte Limnologie: Mitteilungen*, *9*(1), 1–38. https://doi.org/10.1080/05384680.1958.11904091

Vallina, S. M., Follows, M. J., Dutkiewicz, S., Montoya, J. M., Cermeno, P., & Loreau, M. (2014). Global relationship between phytoplankton diversity and productivity in the ocean. *Nature Communications*, *5*, 1–10. https://doi.org/10.1038/ncomms5299

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, *39*(3), 306–314. https://doi.org/10.1007/BF00160154

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., & Gemeinholzer, B. (2015). Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources*, *15*(3), 526–542. https://doi.org/10.1111/1755-0998.12336

## BIOSKETCH

The oceanography research group at Stazione Zoologica Anton Dohrn (Italy) has a broad interest in marine plankton ecology and evolution. Research focuses on the role of physical, biological and biogeochemical processes in setting the structure and functioning of plankton communities. As a constitutive member of the *Tara* Oceans Consortium (https://www.embl.de/tara/tara-oceans-science/index.html), we share the vision of aiming at a deeply multidisciplinary view of the evolution of plankton in the seascape. Therefore, deeply embedded within the Consortium activities our research approach combines advanced numerical modelling and experimental approaches in ocean physics with remote sensing, biology, genomics, ecology and chemistry providing stakeholders with the scientific background needed for meaningful environmental policies.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

**How to cite this article:** Busseni G, Caputi L, Piredda R, et al. Large scale patterns of marine diatom richness: Drivers and trends in a changing ocean. *Global Ecol Biogeogr.* 2020;29:1915–1928. https://doi.org/10.1111/geb.13161