Revised Manuscript Materials essd-2019-159

PhytoBase: A global synthesis of open ocean phytoplankton occurrences

Damiano Righetti*, Meike Vogt, Niklaus E. Zimmermann, Michael D. Guiry, Nicolas Gruber

*Corresponding author. Email: damiano.righetti@env.ethz.ch

This PDF file includes:
Response to Reviews
Marked copy of revised manuscript (track changes)
Revised manuscript (clean version)

In addition: all quality figures and the word document can be accessed via
https://polybox.ethz.ch/index.php/s/e78FVPd6Gn8NJeS

Response to Reviews, 24/02/2020

Summary by Righetti et al. (DR)

We thank the three reviewers for their constructive comments, which provided a valuable basis for increasing the quality, reproducibility, and accuracy of the database and ms. In essence, reviewer 1 advised us to implement minor specifications in three data items. Reviewer 2 suggested minor changes with respect to the data structuring and methodology, with a main focus on facilitating future updates of our database and its curation over time. Reviewer 3 suggested a set of general discussion points and minor specifications.

We address each of these points in detail. Red markings indicate textual edits that have been implemented in the revised version of the manuscript. All statistics and figures in the manuscript have been thoroughly updated.

Reviewer 1:

This paper presents PhytoBase, a global dataset that is essentially a compilation of the existing GBIF and OBIS phytoplankton species occurrence datasets, and a few other smaller datasets. The synthesis and harmonization of these databases results in a substantial increase in phytoplankton occurrence records and yields the largest global database of phytoplankton occurrences. The PhytoBase dataset of spatiotemporal observations of species occurrences may contribute to studies that determine and forecast species distributions and studies aimed at understanding the drivers behind the distribution patterns. The limitations of the database are the spatially highly uneven data density, and, more importantly, strong biases due to differences in sampling methods (e.g. sampling volume, taxonomic resolution etc.). These limitations, appropriately addressed in the paper, prevent the use of PhytoBase for direct analyses of species diversity patterns and biogeography studies, and severely limit the accuracy of data analyses. The authors thus correctly advise that statistical techniques be used to overcome the various biases present in PhytoBase.

I recommend publication of this database in ESSD. I only have a few very minor comments: i) What is the difference between Columns 1-2 and 3-4 in Tables 2 and 3? ii) Can you add a colorbar for the frequency distribution in Figure 3?

Specific responses by Righetti et al (DR) to Reviewer 1 (RE1):

DR: We thank RE1 for the careful check of our data items and change these as follows:
i) Line 227ff and 266ff: Using intersected lines, we now highlight that the first two columns summarize the total records, while the third and fourth column summarize the subset of records with a depth-statement. This distinction is important, as plankton compositions often shift with depth and analyses may thus focus on records with a depth that can be associated to the well-mixed upper water column (mixed layer depth).
ii) Line 313ff: We have added grey bars to each panel in Fig. 3 and specified the caption.

Reviewer 2:

Review summary

The authors present a compilation, named PhytoBase, of five data sources on phytoplankton occurrence records targeting open ocean, including two main data sources: the Global Biodiversity Information Facility (GBIF; www.gbif.org), and the Ocean Biogeographic Information System (OBIS; www.obis.org), complemented by three other sources: the Marine Ecosystem Data initiative (MareDat; Buitenhuis et al. 2013), a marine micro-phytoplankton dataset (Sal et al., 2013), and with a subset of the data collected during the TARA Oceans cruise (Villar et al., 2015). To my knowledge, this compilation leads to the largest dataset on open ocean phytoplankton. A huge effort of data harmonization is recognized, on several aspects of both data structure and taxonomy but also on data qualification (cleaning) required to ensure data quality. This database opens perspectives for phytoplankton research on niche modelling, species distributions, especially within the context of global changes. This database, if updated and maintained in time, will be a valuable bibliographic source for future phytoplankton studies.

I would suggest the acceptance of the paper with 'Minor changes', but I give some recommendations that if addressed will contribute to strengthen and improve the quality of the present paper but also PhytoBase, in particular on data structuring and processes to maintain the valuable product. In my vision, the potential of PhytoBase lies in the compilation of existing data sources but essentially lies in its reproducibility, sustainability and maintenance in time instead of one single snapshot, even more because PhytoBase relies on existing data sources that are maintained and grow in time. That is the reason why

some comments insist on sustainability and maintenance aspects, with emphasis on the need of reproducible processes.

Interpretation of the aspects raised by Reviewer 2 (RE2):

DR: We thank RE2 for the thorough analysis and constructive comments, which greatly improved the quality of our manuscript. We share every interest to facilitate future updates of PhytoBase. To ensure this "dynamic component", we increase transparency and clarity on our methods, in particular with regard to synthesizing original data and columns across sources (textual edits, see lines indicated below), and we now publish the 21 relevant R-scripts used to do download, clean, and synthesize PhytoBase on gitlab (https://gitlab.ethz.ch/phytobase/supplementary). In addition, we now publish the "synonymy table" on gitlab, which lists the original 3303 species names (or generic names) in the raw data together with the harmonized species names (or generic names).
Line 540ff: *"PhytoBase is publicly available through PANGAEA, doi:10.1594/PANGAEA.904397 (Righetti et al., 2019a). Associated R scripts and the synonymy table used to harmonize species' names are available through https://gitlab.ethz.ch/phytobase/supplementary.*

Detailed comments

Abstract No comments

1. Introduction No comments

2. Compilation of occurrences

2.1.Data origin
- Line 100-104: The authors should argue further why they have chosen these three complementary data sources in particular: the MareDat, data from Sal et al. and TARA data collection subset. Did the authors proceed to some extensive bibliographic work to search for potentially valuable datasets and were the sources used the only available open datasets? If yes, this deserves further statements on this bibliographic work, and eventual criteria (if applicable) to choose the data sources.

Specific responses by Righetti et al (DR):

DR: We clarify our initial choice of data sources: The primary focus was set on retrieving data from GBIF (www.gbif.org) and OBIS (www.obis.org); firstly, because GBIF and OBIS promised the largest gain of data-points, as a function of time and effort spent (GBIF: 790'103 data points for 1492 species, with 54.9% of points being unique to this source; OBIS: 823'836 data points for 1320 species, with 56.3% of points being unique). Second, we focused on GBIF and OBIS, because a framework including these two growing archives, will ensure an efficient gathering of phytoplankton data in the future, in line with the mission statement of GBIF ("GBIF (…) is aimed at providing anyone, anywhere, open access to data about all types of life on Earth"; https://www.gbif.org/what-is-gbif , accessed 27.02.2020) and OBIS ("Vision: To be the most comprehensive gateway to the world's ocean biodiversity and biogeographic data (…)"; https://www.obis.org/about/, accessed 27.02.2020). Due to their strive for completeness, we expect OBIS and GBIF to remain leading archives for sharing biological data between multiple datasets and sources, and will serve themselves as key attractors for future datasets from various sources, including datasets from TARA Oceans, the MALASPINA expedition, and other marine diversity efforts. In this context, it will be the key task of individual institutions and cruises to inject their data into these two archives, rather than spreading data across multiple repositories, and to reconcile taxonomy with reference standards. Our work demonstrates how data can be efficiently inter-compared and merged between major plankton data archives.

Our choice of the three additional sources was, indeed, not exhaustive. It included a large dataset that was acquired, quality-controlled and published by our group, the MAREDAT data set, which we are highly familiar with (e.g., O'Brien et al., 2016; Brun et al., 2015; MAREDAT: 101'969 records, among which 94.7% were new to PhytoBase). We also strived to include data from the global TARA Oceans cruise, yet at the time of data download (closing window, March 2017) not all data from TARA Oceans were publicly available, and we thus limited the inclusion to the quality-controlled dataset of Villar et al. (2015). Last but not least, we added the global dataset from the AMT data series by Sal et al. (2013), which is unique in aspects of taxonomic standardization and consistency in sampling methodology. The inclusion of other, smaller datasets was beyond the scope of this project.

We thoroughly specify the selection of data in the revised version of the manuscript:
Line 100ff: *"To create PhytoBase, we compiled marine phytoplankton occurrences (i.e.,*

presences or abundances) from five sources, including the two largest open-access species occurrence archives: the Global Biodiversity Information Facility (GBIF; www.gbif.org), and the Ocean Biogeographic Information System (OBIS; www.obis.org). These two archives represent leading efforts to globally gather species distribution evidence. We augmented the data with records from the Marine Ecosystem Data initiative (MareDat; Buitenhuis et al. 2013), records from a micro-phytoplankton dataset (Sal et al., 2013), and records from the global TARA Oceans cruise (Villar et al., 2015), which were not included in GBIF or OBIS at the time of data query (closing window, March 2017). While our selection of additional data was not exhaustive, it strived for the inclusion of quality controlled large-scale phytoplankton datasets. Specifically, MareDat represents a previous global effort in gathering marine plankton data for ecological analyses (e.g., Brun et al., 2015; O'Brien et al., 2016), while Sal et al. (2013) and Villar et al. (2015) are unique in aspects of taxonomic standardization and consistency in methodology."

DR: To avoid redundancies and increase clarity, we specify the subsequent sections:
Lines 132ff : *"(…). Occurrence data from the TARA Ocean cruise included the Bacillariophyceae and Dinoflagellata (Villar et al., 2015; their Tables W8 and W9). Occurrence data from MareDat included five phytoplankton papers (Buitenhuis et al., 2012; Leblanc et al., 2012; Luo et al., 2012; O'Brien et al., 2013; Vogt et al., 2012). Additional data processed by the TARA Oceans or Malaspina expedition (Duarte, 2015) may provide valuable context for a future data synthesis, yet here we have focused on publicly available sources until March 2017. The raw sources that underpin GBIF and OBIS, and MAREDAT, represent decades to centuries of efforts spent in collecting phytoplankton data, including a substantial amount of data from the CPR program (Richardson et al., 2006). In addition, a large fraction of data from the AMT program (cruises 1 to 6) are represented in Sal et al. (2013)."*
Line 483ff: "*The harmonization of different archives striving to collect global species evidence, therefore substantially expanded the empirical basis of phytoplankton records."*

Lines 120-121: R packages RPostgreSQL and devtools should be properly cited and referenced
DR: We agree and cite the packages. In addition we reference the package 'robis'.
Line 124 ff: *"The data from OBIS were first retrieved on 5 December 2015 using the R package robis (Provoost and Bosch, 2015) and the OBIS taxonomic backbone, accessed on 4 December 2015 via the R packages RPostgreSQL (Conway et al., 2015) and devtools (Wickham, H. and*

Chang, 2015). Data were updated for the taxa selected on 6 March 2017 (using the OBIS taxonomic backbone, accessed on 6 March 2017 via the same R packages).”

Line 575 ff:

”Provoost, P. and Bosch, S.: robis: R client for the OBIS API. R package version 0.1.5. https://cran.r-project.org/package=robis, 2015.

Conway, J., Eddelbuettel, D., Nishiyama, T., Prayaga, S. K., Tiffin, N.: RPostgreSQL: R interface to the PostgreSQL database system. R package version 0.4.1. https://cran.r-project.org/package=RPostgreSQL, 2015.

Wickham, H. and Chang, W.: Devtools: Tools to make developing R packages easier. R package version 1.12.0. https://cran.r-project.org/package=devtools, 2015.”

2.2.Data selection

2.2.1. Data accessed through GBIF and OBIS

- Lines 146-149: please provide percentage of records excluded with filters on year and missing date.

DR: We revisited our statistics and now present this information in the main text:

Line 149ff: *“To filter out raw data of presumably inferior quality, records from OBIS and GBIF were removed: (i) if their year of collection indicated >2017 or <1800 (excluding 110 records; <0.001% of raw data), (ii) if they had no indication on the year or month of collection (excluding 7.2% GBIF raw data and 0.9% OBIS raw data) or (iii) if they had geographic coordinates outside the range -180 to 180 for longitude and/or outside -90 to 90 for latitude. The latter criterion did not lead to any data exclusion, as (…)”*

Line 154ff has now been adjusted and specified accordingly: *“Records with negative recording depths (0% of GBIF and 6.6% of OBIS raw data) were flagged and changed to positive, assuming that their original sign was mistaken.”*

Line 171ff has now been adjusted accordingly: *“(…) we flagged rather than excluded data with reported recording before year 1800 (564 records; values 6, 10 or 11) and unrealistic day entries (58 340 records; values -9 or -1).”*

2.3.Concatenation of source datasets

- Line 188, In table 1: The authors do not mention in the main document that the two main data sources GBIF and OBIS extensively rely on the Darwin Core standard (https://dwc.tdwg.org/terms/). This explains why most of column names are the same. In

addition, the authors should precise that an attempt was done to match Darwin Core standard in the final column names, as working to comply with a standard is in general a best practice and an added value for the work. Due to the fact the two main sources are aligned on Darwin Core, not mentioning Darwin Core might be seen as regression. To understand that the Darwin Core standard has been exploited by the authors, we have to refer to the CSV table (http://hs.pangaea.de/Projects/PHYTOBASE/Column_definition_for_phytoplankton_harmonized_database.zip) available under section 'further details' of PhytoBase PANGEA record available at https://doi.pangaea.de/10.1594/PANGAEA.904397

DR: We thank the reviewer for this point. We now explain our naming convention at the first instance in the main text, which aligns with Darwin Core (dwc) standard wherever possible. We now provide an overview on the full column structure contained in PhytoBase in Table1 and highlight the column names that are in line with dwc. Upon contacting the GBIF secretariat, we received an additional expert opinion on the possibility for alignment of our original column names in PhytoBase with dwc.
Line 187ff: "*Columns match Darwin Core standard (https://dwc.tdwg.org) where original data structure could be reconciled with this standard, following GBIF and OBIS that widely rely on Darwin Core. Where critical metadata could not be assigned to Darwin Core, we use additional columns (e.g., columns ending in "gbif" present metadata from GBIF).*"
We highlight the column names in line with dwc by a "*", adding a note to Table 1:
Line 196: "**Column names following Darwin Core standard (https://dwc.tdwg.org).*"
We adjust the table's header, line 192: *"Table 1: Harmonization of original column names (data-fields) between data sources and final column name structure in PhytoBase"*
We shorten the main text: Line 147: "(…) assuming that the latter was based on observation ~~(see Table 1 for an overview of the metadata retained).~~

- Lines 188, In table 1 : The table intends to harmonize the column names, and tries to use Darwin Core when possible. This is achieved with the following fields: scientificName, basisOfRecord, decimalLongitude, decimalLatitude, taxonRank, individualCount, year, month, day, but not for other fields. Indeed, for fields that are common to at least two data sources, such as Darwin Core column names for GBIF and OBIS, the table results in a some kind of de-harmonization and de-standardization of column names, such as for:
o institutionCode (Darwin Core term), that is split in two separate columns specific to

GBIF/OBIS, ie institutionCode_gbif / institutionCode_obis. It would have been preferable to keep the standard column name, and act at content level to keep source provenance, for example adding a prefix or URN such as urn:gbif:<institutioncode> or urn:obis:<institutioncode> as content of a single institutionCode

o cellsPerLitre : This is a non standard term. It is recommended to keep aligning on the Darwin Core standard by relying on columns relative to measurements or facts, in particular to use standard terms measurementType ("number of cells"), measurementValue (value of cell number), and measurementUnit (number of cells per litre)

o Depth: This is an non standard term and it deserves a reflexion whether the use of standard terms minimumDepthInMeters and maximumDepthInMeters could be relevant.

DR: We now present the full column structure of PhytoBase in Table 1. The column names align with our revised naming convention (see above, revised line 189ff). We mark all column names in Table 1 that are in line with dwc standard.

DR: InstitutionCode: Entries on the "InstitutionCode" of records stemming from both OBIS and GBIF have been identical. We hence could perfectly merge the columns institutionCode_gbif and institutionCode_obis to a single column named "InstitutionCode", in line with dwc.

DR: cellsPerLitre: In line with RE2, and upon contacting the GBIF secretariat, we now split this column into two dwc terms: "organismQuantity" (here, we present the values) and "organismQuantityType" (i.e., "number_of_cells_per_L").

DR: Depth: We carefully examined the benefit of including the minimum– and maximum depth statement. However, "MinimumDepthInMeters" and "MaximumDepthInMeters" were not available for original GBIF records. By contrast, 18.6% of GBIF raw records contained a statement on "DepthAccuracy". This is because GBIF sticks to the term "depth" (as differing from dwc) and the single matching term "depthAccuracy". Similarly, among the OBIS records, 21.6% contained a "depthAccuracy", and only marginally more records contained a MinimumDepthInMeter (25.7%) or minimumDepthInMeter (24.0%). To enhance compatibility between the two major source archives in PhytoBase, we hence stick to the term "depth" together with "depthAccuracy", in line with GBIF data conventions. We now elaborate this point in the main text.

Line 190 ff: *"With regard to sampling depth, GBIF raw data contained the field*

"depthAccuracy" (i.e., a non Darwin Core term; 18.6% of records with entries) while OBIS raw data contained the fields "depthprecision" (21.6% of data with entries), "minimumDepthInMeters" (25.7% of data with entries) and "maximumDepthInMeters" (24.0% of data with entries), i.e., two Darwin Core terms. To enhance compatibility between GBIF and OBIS, we therefore used the column "depth", together with "depthAccuracy", and we integrated "depthprecision" into the latter column."

DR: We note that depth accuracy statements have not been present in the raw data of Maredat, Villar et al. (2015) or Sal et al. (2013). This is mainly because discrete samples at specific depths have been analyzed for phytoplankton abundance and taxonomic identity.

RE2: In similar way, it is recommended that authors check in depth about existence of Darwin core terms that match the other column names: originDatabase, datasetKey, collectionCode, resname, resourceID, cruiseOrStationId, cruise, sampleId; while avoiding the use of source-specific column names, as illustrated above with the *institutionCode*.

DR: We agree, and check the remaining column names for compatibility with Darwin Core.

DR: "originDatabase_maredat" refers uniquely to MareDat (original name: "Origin Database" or "Database", depending on the MareDat paper). This column presents acronyms of original databases to which records belonged inside MareDat. In line with our naming convention provided in lines 187ff of the revised ms (i.e., Darwin Core where possible, specific columns for other relevant metadata where needed) we stick to the current term.

DR: "datasetKey" is a non dwc term, inherent to GBIF terminology: A closely related dwc term would be "datasetID". We thus tested whether we can merge "datasetKey" (inherent to GBIF data) and "resourceID" (inherent to OBIS data) into the single column named "datasetID", without creating ambiguity to which original source (GBIF, OBIS, MAREDAT, Villar or Sal) merged entries in "resourceID" would belong. We find that for 26.1% of data in PhytoBase, this merger would lead to two entries for "resourceID" – one leading back to OBIS, one back to GBIF. This is because a substantial part of the records have origin in both GBIF and OBIS. To keep column entries slim and retain important metadata, traceable to OBIS and GBIF, we decide to stick to the current columns, in line with our naming convention. In addition, we find that there are many more datasetKeys (GBIF) than resourceIDs (OBIS). Hence, retaining the detail of resolution seems advantageous.

DR: In line with our naming convention, we retain "collectionCode_obis", "resname_obis",

"resourceID_obis", "cruiseOrStationID_maredat", "cruise_sal", and "sampleID_sal" as separate columns. These columns contain metadata at different levels of detail, reflecting data structure in underlying source archives. This original resolution is important for future data users, as it allows associating the records to different cruises or protocols, and thus potentially different methodologies used in phytoplankton collection.

DR: We have removed column "ID", which is not conform with dwc. However, we now add a note to Table 1 guide the reader/user with the potential creation of an occurrence ID:
Line 195: *"Each record in PhytoBase is uniquely identifiable by the occurrence ID: scientificName, decimalLongitude, decimalLatitude, year, month, day, depth"*

- Lines 188, In table 1, row about cellsPerLitre: The authors should also revise the corresponding table row as it seems information has been wrongly copied-pasted ("taxonRank")
DR: Indeed. "taxonRank" has been deleted from the erroneous places in Table 1.

- Line 210. The authors make use of a column "group" to add either the Phylum or Class. It is recommended to keep using Darwin Core standard terms phylum and class as separate columns.
DR: We have added the columns "phylum" and "class" (dwc standard terms) to PhytoBase, and remove "group". We thouroughly checked higher order taxonomy and adjusted the ms accordingly. We add a note to Table 1 on the higher order taxonomic hierarchy:
Line 198: *"†Higher order taxonomy (phylum, class) follows OBIS (taxonomic backbone; retrieved 6 March 2017), which relies on the World Register of Marine Species (www.marinespecies.org)".*
DR: We update Table 4 in the MS accordingly (Line 351ff), We change *Dinoflagellatae* to *Dinophyceae* throughout the ms, and we adjust/clarify the names of key taxa:
Lines 15ff: *"(…) spanning the principal groups of the diatoms, dinoflagellates, and haptophytes, as well as several other groups."*
Line 112ff: *"More specifically, within the Ochrophyta, we considered the classes Bacillariophyceae (diatoms), Chrysophyceae, Dictyochophyceae, Pelagophyceae and Raphidophyceae. Within the Myzozoa, we considered the class Dinophyceae (dinoflagellates)."*
Line 478ff: *"This new database contains 1 360 621 records (1 280 103 records at the level of species), including 1716 species of seven phyla."*
Line 380: *"‡Including one species of the syster class Pelagophyceae."*
Lines 478ff: *"This new database contains 1 360 621 records (1 280 103 records at the level of species), including 1716 species of seven phyla."*
Lines 547ff: "*In PhytoBase, we compiled more than 1.36 million marine phytoplankton records*

that span 1704 species ~~and ten major groups,~~ including the key taxa Bacillariophyceae, Dinophyceae, Haptophyta, Cyanobacteria and others."

- Line 212: The authors make use of a column "sourceArchive" to refer to the data source from which the record comes from. It is recommended to look carefully at Darwin Core standard to find the appropriate standard term to use for referencing the data source.

DR: We agree that standard terms are preferable. Our column "sourceArchive" is unique to the PhytoBase compilation, indicating from what original, large archive (GBIF, OBIS, MAREDAT, Villar, or Sal) each record stems. The associated column "yearOfDataAccess" presents the year, in which data were downloaded from archives. We find no suitable matchup terms in dwc system for these purposes, and stick to the current terms.

- Beyond the harmonization of column names highlighted in Table 1, since I believe it is the core of paper describing the set-up of PhytoBase, it would be highly valuable to include in the main document the final data structure retained in the PhytoBase (as set in table http://hs.pangaea.de/Projects/PHYTOBASE/Column_definition_for_phytoplankton_harmonized_database.zip), including column definitions, and for the extra columns added by authors, to proceed with an in-depth check about existence of Darwin core terms to use instead of adhoc column names, as recommended with column names enumerated above. In fact, the high potential of PhytoBase and perspective to exploit it will be fostered by such Darwin Core standard compliance. By relying on Darwin Core, this will offer perspectives to facilitate growing of source global information systems such as GBIF or OBIS with datasets not yet available through it, while benefiting from data already harmonized and standardized through PhytoBase.

DR: We agree with RE2 that a comprehensive presentation of column names is desirable. We adjust Table 1 accordingly. We now also elucidate the content of many columns in the footnotes of Table 1. Yet, given space constraints, we describe each column and their content more thoroughly in the Excel sheet, which is presenting all columns (accompanying PhytoBase on Pangaea). Moreover, Table 1 has been annotated to indicate dwc terms. See also our discussion, to what degree we make columns compatible with dwc in our response to the RE2's general comment on "2.3. Concatenation of source datasets".

DR: We checked the compatibility of added columns with dwc:
Regarding "sourceArchive" and "yearOfDataAccess" we stick to the original terms, in accordance with our response to line 212 (RE2, see above). We now explain why we include

the two columns in the main text.

Line 208ff: *"To indicate the source from which records were obtained (GBIF, OBIS, MAREDAT, VILLAR or SAL) and the year of data access, we added the columns "sourceArchive" and "yearOfDataAccess".*

DR: Regarding "colonialFormCellsPerLitre": We now integrate the column "colonialFormCellsPerLitre" into the columns "organismQuantity" and "organismQuantityType", using "number_of_colonial_ form_cells_per_L" as the entry for the latter. To maintain source attribution we highlight that quantifications for "colonial type cells" stem from MAREDAT

Line 166: *"Across all sources, data on colonial cells were uniquely provided by MareDat, (…)."*

DR: Regarding "totalColonialorSingleCells_or_trichomes_l": We remove this column, as it cannot be reconciled with dwc, while adding only very minor additional data to PhytoBase. To compensate for this exclusion, we refer to the additional data in the text.

Line 166: *"Across all sources, data on colonial cells could be uniquely accessed via MareDat (and additional count data on trichomes of genus Trichodesmium are available from Luo et al., 2012)."*

DR: Regarding "recordWithinMLD_clim" and "depthOriginal". Both columns cannot be reconciled with dwc. We remove the first column (presenting climatological reference data from de Boyer and Montegut, 2004) and leave it now up to the data user to define the mixed-layer depth (if required to select data). The second column ("depthOriginal") can be reconstructed via the column "depth" and a new column "flag" (below). We hence delete it.

DR: Regarding "unrealisticDayOrYear" and "basisPresumablySedimentary": We replace these columns by a quality flag column, termed "flag". We explain the purpose of this column to the reader in the main text.

Line 210ff: *"Last, we added a quality flag column, termed "flag". This column denotes records with originally negative collection depth entries (N) (sect. 2.2.1), unrealistic day (D) or year (Y) entries (sect. 2.2.2), and/or records collected from sediment samples or traps (S), rather than seawater samples (sect. 2.3.2).*

Line 273 ff: *We flagged phytoplankton records from OBIS and GBIF in the database associated with surface sediment traps or sediment cores (denoted by an "S", in the flag column) (…)".*

DR: Accordingly, we correct all column names, and their explanation in the excel sheet that accompanies PhytoBase on Pangea.

DR: Owing to the changes in column name structure, in line with the inputs by RE2, the following sentences or sub-clauses have been deleted from the manuscript:
Line 164ff: ~~The column "unrealisticDayOrYear" in PhytoBase indicates day or year entries, originally associated with MareDat. Data selected from MareDat were merged to a single dataset, containing the columns: "scientificName", "longitude", "latitude", "year", "month", "day", "group", "Origin Database", "Cruise or station ID", "basis", "depth", and "rank".~~

Line 203ff: ~~We added the column "group" to the database, denoting to which phylum or class records belong: i.e., *Cyanobacteria*, *Bacillariophyceae*, *Chlorophyta*, *Chrysophyceae*, *Cryptophyta*, *Dinoflagellata*, *Euglenophyta*, *Haptophyta*, *Raphidophyceae* or picoeukaryotes, and the column "sourceArchive", indicating the source from which records were obtained (GBIF, OBIS, MAREDAT, VILLAR or SAL).~~

Line 251 ff: ~~Furthermore, we added the column "yearOfDataAccess", indicating the year of data download (2015, 2017 or both) and the column "containedWithinMLD_clim", which distinguishes records stemming from waters deeper than the oceanic mixed-layer (monthly climatology, de Boyer Montégut 2004) (11.5% of records) from those inside the mixed-layer.~~

Line 265 ff: *"(…) this does not exclude the possibility that occurrence records of extant species in the GBIF and OBIS source datasets originated partially from sediment traps or sediment core samples~~, rather than from seawater samples~~."*

2.3.1. Extant species selection and taxonomic harmonization
- Lines 223-227: The authors refer to a screening process performed by Algaebase founder and director, as personal communication. This screening led to exclude a relatively significant number of taxa and associated data. Hence, such process seems to appear as key harmonization task for PhytoBase. In my opinion, such process should be further described in the actual PhytoBase and paper materials & methods. In addition, there is no statement that make understand whether the screaning process was done manually or through a semi-automated procedure. If it is a manual process, this may be seen as a limitation referring to reproducibility, sustainability and maintenance of PhytoBase, even more

because it has not been operated by PhytoBase creators/maintainers. It is then strongly recommended to describe further such screening process within the main document (or through an appendix), and, if done manually, to suggest how this could be replaced or at least complemented by a semi-automated and reproducible process , thus leading to the possibility for future users to get an updated PhytoBase in time.

DR: First, we provide the necessary basis that any updated (or different) method can be implemented to standardize or harmonize the species names in PhytoBase:
Line 197: *"¶We retain all original scientificName(s) and synonyms used in individual sources as additional columns with the format "scientificNameOriginal_<source>"*
Line 257ff: "*In particular, we retained the original taxonomic name(s) associated with each record in separate columns of the type "scientificNameOriginal_<source>", which allows tracing back the harmonized name to its original name(s). Retaining original names ensures that future taxonomic changes or updated methods can be readily implemented."*

DR: Second, we agree with RE2 that the harmonization procedure should be further specified, which has now been implemented as follows.
Line 223ff: *"(ii) We extracted all scientific names (mostly at species level, including all synonyms and spelling variants) associated with at least one depth-referenced record from the raw database (Table 2). This resulted in 3302 names, which were validated in August 2017 against the 150 000+ specific and infraspecific names in Algaebase (www.algaebase.org), and matched using a relational database of current names and synonyms; orthography was made as compatible as possible with the International Code of Nomenclature (Turland et al., 2018), particularly in relation to the gender of specific epithets. ~~Each name was verified by M. Guiry, the founder and director at Algaebase (M. Guiry, pers. comm.) in August 2017~~. This ~~expert~~ screening led to the exclusion of 459 names (…).*
(iii) We excluded species (and their data) classified as "fossil only" or "fossil", based on Algaebase (accessed August 2017) or the World Register of Marine Species (WoRMS; www.marinespecies.org, accessed August 2017). We further excluded species belonging to genera with fossil types denoted by Algaebase, under the condition that these species lacked habitat information on ~~both~~ Algaebase ~~and WoRMS~~, assuming that the latter species have been collected based on sedimentary or fossilized materials. Species uniquely classified as "freshwater" on ~~both~~ Algaebase ~~and WoRMS~~ were discarded, as these were beyond the scope of our open ocean database. However, we retained species classified as (…)."

DR: We add Turland et al. (2018) to the references.

Line 727 ff: "*Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T. W., McNeill, J., Monro, A. M., Prado, J., Price, M. J. & Smith, G. F., editors. International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Regnum Vegetabile, Vol. 159. pp. [i]-xxxviii, 1-253. Glashütten: Koeltz Botanical Books, 2018. doi:10.12705/Code.2018.*"

DR: We now also include M. D. Guiry as co-author on the revised manuscript.

Line 3: "*Damiano Righetti[1], Meike Vogt[1], Niklaus E. Zimmermann[2], Michael D. Guiry[3], Nicolas Gruber[1]*" *[3]AlgaeBase, Ryan Institute, NUI, Galway, University Road, Galway H91 TK33, Ireland*

2.3.2. Data merger and synthesis

- Line 270: The rgbif R package should be properly cited and referenced. In addition, please note that there is a typo with the package name ('rgibf' instead of rgbif).

DR: Excellent catch. rgbif has now been spellchecked and cited.

Line 275: *"(…) using the function datasets in the R package rgbif (Chamberlain, 2015)(…)"*

Line 609ff: *Chamberlain, S.: rgbif: Interface to the Global Biodiversity Information Facility API. R package version 0.9.7. https://cran.r-project.org/package=rgbif, 2015.*

3. Results

3.1. Data

- This section is very welcome and acknowledged.

3.1.1. Spatiotemporal coverage

- Line 283: It is recommended to add the EPSG code of the World Geodetic System (WGS84). In addition, I recommend to include this as standard Darwin Core column in PhytoBase using the term geodeticDatum.

DR: We now mention the EPSG code in the first instance in the MS:

Line 152ff: "*However, the latter criterion was fulfilled by all records, as these were standardized to -180 to 180 degrees longitude (rather than 0 to 360 longitude East) and -90 to 90 degrees latitude (WGS84).*"

WGS84 had also been included in the Excel sheet (for columns: decimalLatitude, and

decimalLongitude), which accompanies PhytoBase on Pangaea. We consider this information redundant with an additional column added to PhytoBase and prefer to keep the number of columns in the database to the minimum possible, since this increases the usability of the data set, and facilitates treatment of data in analysis software packages.

5. Data availability

- In principle, it is highly recommended, based on principles of open and reproducible science and sustainability, that authors make available already the R scripts together with the PhytoBase on PANGAEA, and avoid provision on demand through emails to the authors.

DR: We agree with this point. We now provide all 21 R scripts used to do download, clean, and synthesize PhytoBase (and to match data columns with Darwin core terms) through gitlab: https://gitlab.ethz.ch/phytobase/supplementary. Due to the large amount of scripts required to perform each successive step of the database assembly, we gather the scripts into two folders, i.e., "download_and_prepare_data" and "merge_and_harmonize_data".

References:

Brun, P., Vogt, M., Payne, M. R., Gruber, N., O'Brien, C. J., Buitenhuis, E. T., Le Quéré, C., Leblanc, K. and Luo, Y.-W.: Ecological niches of open ocean phytoplankton taxa, Limnol. Oceanogr., 60(3), 1020–1038, doi:10.1002/lno.10074, 2015.

O'Brien, C. J., Vogt, M. and Gruber, N.: Global coccolithophore diversity: Drivers and future change, Prog. Oceanogr., 140, 27–42, doi:10.1016/j.pocean.2015.10.003, 2016.

Reviewer 3:

The MS entitled "PHYTOBASE: A global synthesis of open ocean phytoplankton occurrences" by Righetti et al. represents an interesting effort of combining major existing marine phytoplankton diversity information gathered by microscopy observation, discrimination, identification and, for some of them cells and colony counts, all over ocean systems around the Globe. The authors take into account not only abundance (quantitative) but also presence (qualitative) information in the same database, as well as different sampling methodologies which have an impact on the results obtained, considering bigger or smaller organisms (according to mesh/silk size discrimination and/or microscopy limitations), delicate or robust species (which will not be disrupted by mesh collection), rare or abundant species (depending on the volume of sample analysed). The description of the data as well as the combination methodology, quality control, flagging and taxonomic

relevance/correction of the datasets before and after merging them, are clear. The authors make it possible to address a more complete picture by providing a direct and easier access to current knowledge of phytoplankton distribution all over the oceanic realm, identifying properly the uneven distribution od sampling effort and, consequently, of biodiversity assessment or phytoplankton in large areas mainly identified in the Southern Hemisphere. Moreover, they made also an assessment of which are the taxa well known in comparison which the taxa relatively poorly known, mainly concerning small phytoplankton. Finally, they clearly demonstrate the new possibilities in developing ecological models and predictions on the distribution of phytoplankton taxa in open ocean systems.

I therefore recommend this MS to be published in Earth System Science Data after some small technical corrections (see below).

Some general considerations:

One issue to be reminded is that one cannot state for sure, even considering areas which have been well sampled for decades, that some species are not present in a precise area, mostly because, in the corresponding existing databases, studies combining different sampling approaches and, to some extent, also different approaches for considering either morphology, molecular or functional diversity, are scarce.

It remains important then to make this new database as informative as possible, not only concerning the correct nomenclature to be used (and a big effort for make old and new names was also carried out by the present work) but also by considering biases due to different sampling strategies (either nets or tows, Niskin bottles, continuous pumping at a considered depth). One recommendation would be to maintain taxonomic and phylogenetical research as a complement of routine monitoring efforts, providing more accurate consideration of rare species by considering higher sample volumes, concentration by different manners and, the most important, taxonomist expertise which, combined to molecular phylogeny, will certainly make it possible to extract more information from metabarcoding and metagenomic approaches. Moreover, it is also important to consider also new automated approaches which would make it possible to extend the sampling effort on different platforms, addressing most of the time a most limited taxonomical resolution but recalling on functional diversity which, to some extent, would complete taxonomical information included in a marine phytoplankton global database.

Interpretation of the aspects raised by Reviewer 3 (RE3):

We thank for the comments raised by RE3. Indeed, we share the view that omission of rare species is a limitation in our work [e.g., Line 350ff: *"However these estimates only represent the fraction of species detectable via light microscopy, and other methods underlying our database, preferentially omitting very rare or small species (Cermeño et al., 2014; Ser-Giacomi et al., 2018; Sogin et al., 2006)*].

DR: We have strengthened the point that several diversity dimensions and methodological approaches combined would amplify the benefit of PhytoBase.

Line 135ff: *"Additional data processed by the TARA Oceans or Malaspina expedition (Duarte, 2015) may provide valuable context for a future synthesis, and may eventually combine molecular with traditional approaches, yet here we have focused on (…)."*

DR: We also strengthen the discussion about potential species omission:

Line 483ff: *"Second, sampling priorities with respect to taxonomic groups, size classes or species resolution differ widely between research cruises and programs. While small or fragile species may escape detection by the CPR program (Richardson et al., 2006), the resolution of seawater samples is influenced by sampling volume and taxonomic expertise (Cermeño et al., 2014). Our results show that (…)."*

Finally, we highlight the benefit of integrating molecular data, in line with the point by RE3:

Line 512ff: *"The detection of rare species and their integration into PhytoBase may become possible via molecular methods (Bork et al., 2015; Sogin et al., 2006). DNA sequencing has become an alternative approach to (…)."*

Some details:

Page 3 line 74: ". . .onto a 270 µm silk roll. . ." as it is important to remind the particular sampling conditions of CPR.

DR: We agree and include the detail in mesh size.

Line 74ff: *"(…) in which plankton are sampled by filtering seawater onto a silk roll (270 µm mesh size) within a recorder device that is towed behind research and commercial ships (Richardson et al., 2006)."*

Line 427 ff: *"The mesh size of the silk employed in CPR of 270 µm under-samples small phytoplankton species (<10 µm)."*

Page 6 line 170; what about other essential metadata as "collection device" and "analytical tool" (type of microscope) and "volume analysed"? Would this information be available/included/easy to access?

DR: In line with the need to retrieve metadata (depending on the purpose of analysis) we retained datasetKeys, resourceIDs and cruiseIDs that link back to specific source archives in PhytoBase as separate columns. Unfortunately, essential metadata on the specific sample collection method are, more often than not, not automatically included in the data retrieved from archives such as GBIF and OBIS. Essentially, we would need to check every dataset key (GBIF) or resourceID (obis), which potentially links metadata with individual datasets in these archives. We consider the inclusion of this information for all taxa considered beyond the scope of this work. Yet, we now refer more explicitly to the option to retrieve metadata:

Line 205: *"§§ datasetKey_gbif and resourceID_obis are keys to access metadata of original datasets in GBIF and OBIS via API, including information on sampling methods."*

Line 494ff: *"Thus, without careful screening and checking of the data (via e.g. datasetKeys for GBIF records, resourceIDs for OBIS records), the characterization of biogeographies at the species level might be highly biased."*

Page 16: Figure 5 caption: ". . .temperate seas. . .of Southern Hemisphere (E), cold seas . . .of Southern Hemisphere (F). . ."

DR: The caption has been corrected.

Page 18 lines 419-420: what about other biases of CPR collection as fragile unarmored species, small but also big as ciliates? An extra comment on this issue will be welcomed, as these surveys are one of the most sustained and complete surveys of plankton in some targeted areas.

DR: We agree with RE3 that the CPR data contain methodological limitations, with influence the database collected, meaning that fragile or unarmored species, as also rare species, will be underrepresented in the present study. We added additional explanation and discussion with regard to this – and other – sources of bias in our manuscript. Please see our adjustments above, in response to the first (general) comment of RE3.

Page 20 Figure 8 caption: References García et al. 2013; Locarinio et al., 2013 and de Boyer Montegut, 2004 are missing from the reference list.

DR: The references have been included.

Page 22 line 500: To what extent DNA sequencing have really become an alternative to microscopy for characterizing phytoplankton biogeography instead of a complementary and, to some extent supplementary to morphological microscopic identification?
DR: In our view, this is not a question that can be conclusively addressed. We are in close collaboration with e.g. members of the TARA consortium, and believe that in the future, data collection will tend towards the collection and analysis of environmental (meta)genomic samples, with a move away from traditional microscopy. We believe that classical morphological identification is essential to validate metagenomic information, especially with regard to abundance, biomass or dominance of species. We believe that a merger of traditional and metagenomic data in terms of presence/absence data will be possible, but further efforts need to be made, as come 30% of all oceanic metagenomic data is currently taxonomically unassigned (de Vargas et al., 2015). However, metagenomic data may give us better information eventually on rare and morpholoigically indistinguishable taxa, such as e.g. the vast diversity of picophytoplankton (some of which are included in PhytoBase via MareDat) or haptophytes that cannot be identified using traditional methods.

DR: Our view that metagenomic data and traditional data have become *complementary* approaches to characterize phytoplankton biogeography is reflected in the following edit:
Line 516ff: *"However, we expect that an integration of detailed genetic data with traditional sampling data may soon become possible, allowing to combine several methodological or taxonomic dimensions in databases."*

Page 23 line 535: to what extent have you only considered photosynthetical microbial organisms only, especially in some major taxa where both heterotrophs and pigmented cells (mixotrophs or autotrophs) occur? Thanks for precising this in the Materials and Methods section.
DR: It is currently not known how much heterotrophy is involved in algae in general, but it is well known that mixotrophy is an issue for the dinoflagellates. We modify the Materials and Methods section to include information with regard to this aspect:
Line 114ff: "This selection of phyla or classes strived to include all autotrophic marine phytoplankton taxa (de Vargas et al., 2015; Falkowski et al., 2004), but it is clear that some of the species may be mixotrophic, particularly for the Dinophyceae (Jeong et al, 2010)."

PHYTOBASE: A global synthesis of open ocean phytoplankton occurrences

Damiano Righetti[1], Meike Vogt[1], Niklaus E. Zimmermann[2], Michael D. Guiry[3], Nicolas Gruber[1]

[1]Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, Universitätstrasse 16, 8092 Zürich, Switzerland
[2]Dynamic Macroecology, Landscape Dynamics, Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland
[3]AlgaeBase, Ryan Institute, NUI, Galway, University Road, Galway H91 TK33, Ireland

Correspondence to: Damiano Righetti (damiano.righetti@env.ethz.ch)

Abstract. Marine phytoplankton are responsible for half of the global net primary production and perform multiple other ecological functions and services of the global ocean. These photosynthetic organisms comprise more than 4300 marine species, but their biogeographic patterns and the resulting species diversity are poorly known, mostly owing to severe data limitations. Here, we compile, synthesize, and harmonize marine phytoplankton occurrence records from the two largest biological occurrence archives (Ocean Biogeographic Information System; OBIS, and Global Biodiversity Information Facility; GBIF) and three independent recent data collections. We bring together over 1.36 million phytoplankton occurrence records (1.28 million at the level of species) for a total of 1704 species, spanning the principal groups of the diatoms, dinoflagellates, and haptophytes, as well as several other groups. This data compilation increases the amount of marine phytoplankton records available through the single largest contributing archive (OBIS) by 65%. Data span all ocean basins, latitudes and most seasons. Analyzing the oceanic inventory of sampled phytoplankton species richness at the broadest spatial scales possible, using a resampling procedure, we find that richness tends to saturate in the pantropics at ~93% of all species in our database, at ~64% in temperate waters, and at ~35% in the cold Northern Hemisphere, while the Southern Hemisphere remains underexplored. We provide metadata on the cruise, research institution, depth and date for each data record, and we include phytoplankton cell counts for 193 763 entries. We strongly recommend consideration of spatiotemporal biases in sampling intensity and varying taxonomic sampling scopes between research cruises or institutions when analyzing the occurrence data spatially. Including such information into statistical analysis tools, such as species distribution models may serve to project the diversity, niches, and distribution of species in the contemporary and future ocean, opening the door for quantitative insights into macroecological phytoplankton patterns. PhytoBase can be downloaded from PANGAEA, doi:10.1594/PANGAEA.904397 (Righetti et al., 2019a).

1 Introduction

Phytoplankton are photosynthetic members of the plankton, responsible for about half of the global net primary production (Field et al., 1998). While more than 4300 phytoplankton species have been described (Sournia et al., 1991), spanning at least six major clades (Falkowski et al., 2004), there are likely many more species living in the ocean, perhaps more than

10 000 (de Vargas et al., 2015). Some of these species (e.g. *Emiliania huxleyi*, *Gephyrocapsa oceanica*) are abundant and occur throughout the ocean (Iglesias-Rodríguez et al., 2002), but a majority of plankton species form low abundance populations (Ser-Giacomi et al., 2018) and remain essentially uncharted; i.e., the quantitative description of where they live, and where not, is rather poor. This biogeographic knowledge gap stems from a lack of systematic global surveys, as have been undertaken for inorganic carbon (WOCE/JGOFS/GOSHIP; Wallace 2001) or for trace metals (GEOTRACES; Mawji et al. 2015). Owing to logistic and financial challenges associated with internationally coordinated surveys, our knowledge of phytoplankton biogeography is, with a few exceptions (Bork et al., 2015; McQuatters-Gollop et al., 2015), mostly based on spatially very limited surveys or basin scale studies (e.g., Endo et al., 2018; Honjo and Okada, 1974). Marine phytoplankton occurrence data are unevenly distributed, incomplete in remote areas, and orders of magnitude higher in more easily accessed areas, especially near coasts (Buitenhuis et al., 2013). Additional factors that have impeded progress in developing a good biogeographic understanding of the phytoplankton are difficulties in species identification, linked to their microscopic body size. This is well reflected in the current geographic knowledge on phytoplankton species richness from direct observations (e.g. Rodríguez-Ramos et al., 2015), which is much more limited compared to that of other marine taxa, such as zooplankton (e.g., Rombouts et al., 2010), fishes (e.g, Jones and Cheung, 2015), sharks (e.g., Worm et al., 2005) or krill (e.g., Tittensor et al., 2010), even though many of these taxa also suffer from deficiencies in sampling efforts (Menegotto and Rangel, 2018).

Initial efforts to overcome the data sparseness and patchiness for phytoplankton by the MareDat project (Buitenhuis et al., 2012; Leblanc et al., 2012; Luo et al., 2012; O'Brien et al., 2013; Vogt et al., 2012) resulted in the compilation and synthesis of 119 phytoplankton species from 17 240 sampling events. While representing a large step forward, the coverage remained relatively limited, largely owing to MareDat's focus on abundance data, motivated by the need to use the data for model evaluation and other quantitative assessments (Buitenhuis et al., 2013). But during these efforts, it became clear that there are at least an order of magnitude more data in archives around the world if one relaxed the abundance criterion and considered all observations that included presences. The potential for the use of presences to constrain e.g., phytoplankton community structure and richness, is large, as demonstrated by Righetti et al. (2019b), who recently produced the first global map of phytoplankton species richness. This application was also made possible thanks to the rapid developments in data mining and statistical analysis tools, such as species distribution models (SDMs) (Guisan and Zimmermann, 2000) that permit scientists to account for some of the limitations stemming from spatiotemporal sampling biases underlying species' occurrence data (Breiner et al., 2015; Phillips et al., 2009).

A key enabler for the compilation and synthesis of phytoplankton occurrences (presence or abundance records) is the existence of two digital biological data archives, i.e., the Global Biodiversity Information Facility (GBIF; www.gbif.org), and the Ocean Biogeographic Information System (OBIS; www.obis.org). GBIF is the world's largest archive for species occurrence records, while OBIS is the largest occurrence database on marine taxa. Both archives have gathered a large number of phytoplankton occurrence records and make them freely available to the global community. In addition to MareDat (Buitenhuis et al., 2013), marine surveys such as those conducted with the Continuous Plankton Recorder (CPR)

(McQuatters-Gollop et al., 2015), the Atlantic Meridional Transect (AMT) (Aiken et al., 2000; Sal et al., 2013) and other programs provide relevant phytoplankton occurrence records, including data on species' abundance. A global synthesis of species occurrence records, including those from GBIF and OBIS has been attempted for upper trophic marine organisms, gathering 3.44 million records across nine taxa from zooplankton to sharks (Menegotto & Rangel 2018). But so far, no effort has been undertaken to bring the various sources together for the lowest trophic marine organisms and merge them into a single harmonized database. This study aims to address this gap and to create PhytoBase, the world's largest open ocean phytoplankton occurrence database, which may substantially reduce the global limitations associated with undersampling.

The majority of the existing occurrence data of phytoplankton species have been collected via seawater samples of ~5–25 mL (Lund et al., 1958; Utermöhl, 1958), followed by microscopic specimen identification. Another key source of occurrence data is the continuous plankton recorder (CPR) program, in which plankton are sampled by filtering seawater onto a silk roll (270 µm mesh size) within a recorder device that is towed behind research and commercial ships (Richardson et al., 2006). The plankton are then picked from the screens and identified by microscopy. DNA sequencing has become an alternative method to record and monitor marine phytoplankton at large scales (e.g. de Vargas *et al.* 2015; Sunagawa *et al.* 2015). However, within the recent global TARA Oceans cruise, ca. $^{1}/_{3}$ of DNA sequences of plankton from seawater samples could not yet be assigned to any taxon (de Vargas et al., 2015). For the most species-rich phytoplankton group (*Bacillariophyceae*), 58% of DNA sequences from seawater could be assigned to genus level in the same cruise (Malviya et al., 2016), but the majority of species have lacked reference DNA sequences needed for their identification. Additional factors have hampered the study of global phytoplankton biogeography: Some surveys lack resolution in terms of the species recorded (Richardson et al., 2006; Villar et al., 2015) and abundance information in terms of cells or biomass of species is often not available in the archived records (e.g. from GBIF). Second, the taxonomic identification and chronic undersampling of the species present in local communities via seawater samples (Cermeño et al., 2014) pose challenges, which can be resolved only by trained experts or larger sampling volumes. In addition, the rapidly evolving taxonomy (e.g. Jordan 2004) has led to varying use of nomenclature. These limitations need to be assessed and possibly overcome in a data synthesis effort.

Here, we compile 1 360 621 phytoplankton occurrence records (94.1% resolved to the level of species; n = 1704 species) and demonstrate that combining data from OBIS and GBIF increases the number of occurrence records by 52.7% relative to the data solely obtained from OBIS. This gain increases to 65.2% when adding occurrence data from marine surveys, including MareDat (Buitenhuis et al., 2013), AMT cruises (Sal et al., 2013), and initial TARA Oceans results (Villar et al., 2015). With respect to species abundance information, we retain cell count records whenever available from all sources, resulting in 193 763 quantitative entries. We harmonize and update the taxonomy between the sources, focusing on extant species and open ocean records. The resulting PhytoBase dataset allows for studying global patterns in the biogeography, diversity, and composition of phytoplankton species. Using statistical SDMs, the data may serve as a starting point to examine species' niche differences across all major phytoplankton taxa and their potentially shifting distributions under climate change. The dataset can be accessed through PANGAEA, doi:10.1594/PANGAEA.904397 (Righetti et al., 2019a).

2 Compilation of occurrences

2.1 Data origin

To create PhytoBase, we compiled marine phytoplankton occurrences (i.e., presences and abundances larger than zero) from five sources, including the two largest open access species occurrence archives: the Global Biodiversity Information Facility (GBIF; www.gbif.org), and the Ocean Biogeographic Information System (OBIS; www.obis.org). These two archives represent leading efforts to globally gather species distribution evidence. We augmented the data with records from the Marine Ecosystem Data initiative (MareDat; Buitenhuis *et al.* 2013), records from a micro-phytoplankton dataset (Sal et al., 2013), and records from the global TARA Oceans cruise (Villar et al., 2015), which were not included in GBIF or OBIS at the time of data query (closing window, March 2017). While our selection of additional data was not exhaustive, it strived for the inclusion of quality controlled large-scale phytoplankton datasets. Specifically, MareDat represents a previous global effort in gathering marine plankton data for ecological analyses (e.g., Brun et al., 2015; O'Brien et al., 2016), while Sal et al. (2013) and Villar et al. (2015) are unique in aspects of taxonomic standardization and consistency in methodology.

We retrieved occurrences at the level "species" or below (e.g., "subspecies", "variety" and "form", as indicated by the taxonRank field in GBIF and OBIS sourced data) for seven phyla: *Cyanobacteria*, *Chlorophyta* (excluding macroalgae), *Cryptophyta*, *Myzozoa*, *Haptophyta*, *Ochrophyta*, and *Euglenozoa.* More specifically, within the *Ochrophyta*, we considered the classes *Bacillariophyceae* (diatoms), *Chrysophyceae*, *Dictyochophyceae*, *Pelagophyceae* and *Raphidophyceae.* Within the *Myzozoa*, we considered the class *Dinophyceae* (dinoflagellates). Within the *Euglenozoa*, we considered the class *Euglenoidea*. This selection of phyla or classes strived to include all autotrophic marine phytoplankton taxa (de Vargas et al., 2015; Falkowski et al., 2004), but it is clear that some of the species may be mixotrophic, particularly for the *Dinophyceae* (Jeong et al., 2010). At the genus level, we additionally retrieved occurrences for *Prochlorococcus* and *Synechococcus* from all sources, as the latter two genera are often highly abundant (Flombaum et al., 2013), but rarely determined to the species level. We also retrieved occurrence records for the functionally relevant genera *Phaeocystis*, *Richelia*, *Trichodesmium*, and the "picoeukaryote" group from MareDat. For simplicity, we treat genera as "species" in statistics herein.

For the selected taxa, occurrence data from GBIF and OBIS were first downloaded in December 2015 and updated in February 2017. Specifically, the initial retrieval of the GBIF data occurred on 7 December 2015 (using the taxonomic backbone from https://doi.org/10.15468/39omei, accessed on 14 July 2015), and the data were updated on 27 February 2017 (using an updated taxonomic backbone, accessed via http://rs.gbif.org/datasets/backbone, released 27 February 2017). The data from OBIS were first retrieved on 5 December 2015 using the R package *robis* (Provoost and Bosch, 2015) and the OBIS taxonomic backbone, accessed on 4 December 2015 via the R packages *RPostgreSQL* (Conway et al., 2015) and *devtools* (Wickham and Chang, 2015). Data were updated for the taxa selected on 6 March 2017 (using the OBIS taxonomic backbone, accessed on 6 March 2017 via the same R packages). The update in 2017 expanded the occurrences retrieved from GBIF substantially, with over 20 000 additional phytoplankton records stemming from an Australian CPR program alone

(AusCPR, https://doi.org/10.1016/j.pocean.2005.09.011, accessed via www.gbif.org on 6 March 2017). We retained any GBIF sourced data that were retrieved in 2015, but deleted from GBIF before March 2017 (such as CPR data, with dataset key 83986ffa-f762-11e1-a439-00145eb45e9a). Occurrence data from the TARA Ocean cruise included the *Bacillariophyceae* and *Dinophyceae* (Villar et al., 2015; their Tables W8 and W9). Occurrence data from MareDat included five phytoplankton papers (Buitenhuis et al., 2012; Leblanc et al., 2012; Luo et al., 2012; O'Brien et al., 2013; Vogt et al., 2012). Additional data processed by the TARA Oceans or Malaspina expedition (Duarte, 2015) may provide valuable context for a future synthesis, and may eventually combine molecular with traditional approaches, yet here we have focused on publicly available sources until March 2017. The raw sources that underpin GBIF and OBIS, and MareDat, represent decades to centuries of efforts spent in collecting phytoplankton data, including a substantial amount of data from the CPR program (Richardson et al., 2006) and a large fraction of data from the AMT program (cruises 1 to 6) (Sal et al., 2013).

2.2 Data selection

We excluded occurrences from waters less than 200 m deep (Amante and Eakins, 2009), from enclosed seas (Baltic Sea, Black Sea or Caspian Sea), and from seas with a surface salinity below 20, using the globally gridded (spatial 1° x 1°) monthly climatological data of Zweng et al. (2013). This salinity-bathymetry threshold served to select data from open oceans, excluding environmentally more complex, and often more fertile, near-shore waters.

2.2.1 Data accessed through GBIF and OBIS

We included GBIF occurrence records on the basis of "human observation", "observation", "literature", "living specimen", "material sample", "machine observation", "observation", and "unknown", assuming that the latter was based on observation. With respect to OBIS data, we included data records on the basis of "O" and "D", whereby "O" refers to observation and "D" to literature-based records. To filter out raw data of presumably inferior quality, records from OBIS and GBIF were removed: (i) if their year of collection indicated >2017 or <1800 (excluding 110 records; <0.001% of raw data), (ii) if they had no indication on the year or month of collection (excluding 7.2% GBIF raw data and 0.9% OBIS raw data) or (iii) if they had geographic coordinates outside the range -180 to 180 for longitude and/or outside -90 to 90 for latitude. However, the latter criterion was fulfilled by all records, as these were standardized to -180 to 180 degrees longitude (rather than 0 to 360 longitude East) and -90 to 90 degrees latitude (WGS84). Records with negative recording depths (0% of GBIF and 6.6% of OBIS raw data) were flagged and changed to positive, assuming that their original sign was mistaken.

2.2.2 Data accessed through MAREDAT

We included occurrence records at the species level for the *Bacillariophyceae* (Leblanc et al., 2012) and *Haptophyta* (O'Brien et al., 2013) and species presence records on *Bacillariophyceae* host cells from Luo et al. (2012). Harmonization of *Haptophyta* species names from MareDat (O'Brien et al., 2013) was guided by a synonymy table provided by O'Brien (*pers. comm.*) (Table A1). Harmonization of *Bacillariophyceae* species names in MareDat was in progress at the time of first data

access (24 August 2015) and completed (Table A2). In addition, we selected all genus and species level records available for *Trichodesmium*, *Richelia* (Luo et al., 2012), *Phaeocystis* (Vogt et al., 2012), *Synechococcus* (using the data-field "SynmL") and *Prochlorococcus* (using the data-field "PromL") (Buitenhuis et al., 2012). We included genus level records from the latter taxa, as they represent functionally important phytoplankton groups (Le Quéré, 2005), and as information on the presence and abundance of their cells or colonial cells often only existed at genus level (Buitenhuis et al., 2012; Luo et al., 2012; Vogt et al., 2012). Across all sources, data on colonial cells were uniquely provided by MareDat, while additional count data on trichomes for the genus *Trichodesmium* may be accessed from Luo et al. (2012). In addition, we retained records on the "picoeukaryote" group, which were not determined to species or genus level (Buitenhuis et al., 2012). For all taxa, we retained records with reported abundances (i.e., cell counts) larger than zero, while excluding records with zero entries or missing data (NA), as our database focuses on presence-only or abundance records. Given that data of the MareDat have been scrutinized previously, we flagged rather than excluded data with reported recording before year 1800 (564 records; values 6, 10 or 11) and unrealistic day entries (58 340 records; values -9 or -1).

2.2.3 Data accessed through Villar et al. (2015)

We compiled presence records of species of *Bacillariophyceae* and *Dinophyceae* from the tables W8 and W9 of Villar et al. (2015). We excluded species names containing "cf" (e.g *Bacteriastrum cf. delicatulum*), as such nomenclature is typically used to refer to closely related species of an observed species. We retained all species (n = 3), which contained "group" in their names (e.g. *Pseudo-nitzschia delicatissima group*). *Tripos lineatus/pentagonus complex* was considered as *Tripos lineatus*. The cleaning of spelling variants of original names from Villar et al. (2015) is presented in Table A3.

2.2.4 Data accessed through Sal et al. (2013)

We considered occurrence records of the *Bacillariophyceae*, *Dictyochophyceae, Dinophyceae*, *Haptophyta* and *Peridinea* and at species level or below, using the species name in the final database. These data included 5891 records from 314 species and 543 samples. The dataset of Sal et al. (2013) represents a highly complementary source of phytoplankton occurrence records, i.e., it had no duplicated records with any of the other sources. This data collection contains in situ samples subjected to consistent methodology performed by the same taxonomist.

2.3 Concatenation of source datasets

Column names or data-fields were adjusted and harmonized to establish compatibility in the dimensions of the different source datasets (Table 1). Columns match Darwin Core standard (https://dwc.tdwg.org) where original data structure could be reconciled with this standard, following GBIF and OBIS that widely rely on Darwin Core. Where critical metadata could not be assigned to Darwin Core, we use additional columns (e.g., columns ending in "gbif" present metadata from GBIF). With regard to sampling depth, GBIF raw data contained the field "depthAccuracy" (18.6% of data with entries) while OBIS raw data contained the fields "depthprecision" (21.64% of data with entries), "minimumDepthInMeters" (Darwin Core term;

Table 1: Harmonization of column names (data-fields) between data sources and final column name structure in PhytoBase

Original column names							Final column names
GBIF (2015)	GBIF (2017)	OBIS (2015)	OBIS (2017)	MareDat	Villar et al	Sal et al	(sources merged)
species	species	species	species	species	species	species	scientificName*,¶
decimalLongitude	longitude	longitude	longitude	Longitude	Longitude	Lon	decimalLongitude*
decimalLatitude	latitude	latitude	latitude	Latitude	Latitude	Lat	decimalLatitude*
year	year	yearcollected	year	Year	Date	Date	year*
month	month	monthcollected	month	Month	Date	Date	month*
day	day	daycollected	day	Day	Date	Date	day*
depth	depth	depth	depth	Depth	Depth	Depth	depth
-	depthAccuracy	depthprecision	depthprecision	-	-	-	depthAccuracy
taxonRank	taxonRank	-	-	rank	-	-	taxonRank*,†
-	occurrencestatus	-	occurrencestatus	-	-	-	occurrenceStatus*
phylum	phylum	phylum	phylum	-	-	-	phylum*,‡
class	class	class	class	-	-	-	class*,‡
basisOfRecord	basisOfRecord	basisofrecord	basisOfRecord	-	-	-	basisOfRecord*
-	institutionCode	institutioncode	institutionCode	-	-	-	institutionCode*,§
-	-	-	-	-	-	-	sourceArchive
datasetKey	datasetKey	-	-	-	-	-	datasetKey_gbif‖,§§
publishingOrgKey	-	-	-	-	-	-	publishingOrgKey_gbif§
-	-	collectioncode	collectionCode	-	-	-	collectionCode_obis‖
-	-	-	resname	-	-	-	resname_obis‖
-	-	resource_id	resource_id	-	-	-	resourceID_obis‖,§§
-	-	-	-	Origin Database	-	-	originDatabase_maredat§
-	-	-	-	CruiseorStationID	-	-	cruiseOrStationID_ maredat‖
-	-	-	-	-	Station	-	taraStation_villar‖
-	-	-	-	-	-	Cruise	cruise_sal‖
-	-	-	-	-		SampleID	sampleID_sal
-	-	-	-	-	Mixed Layer Depth (m)	MLD	MLD_villar_sal
-	-	-	-	cellsL^{-1},cellsmL^{-1}	-	organism-quantity	organismQuantity*<and> organismQuantityType*
-	individualCount	-	observedindivi-	-	-	-	individualCount*,+
-	-	-	dualcount]	-	-	-	yearOfDataAccess
-	-	-	-	-	-	-	flag

GBIF data were downloaded in 2015 (www.gbif.org; retrieved 7 December 2015) and 2017 (retrieved 27 February 2017)
OBIS data were downloaded in 2015 (www.iobis.org; retrieved 5 December 2015) and 2017 (retrieved 6 March 2017)
Each occurrence record in PhytoBase is uniquely identifiable by the occurrence ID: scientificName, decimalLongitude, decimalLatitude, year, month, day and depth
* Column names following Darwin Core standard (https://dwc.tdwg.org).
¶ We retain all original scientificName(s) and synonyms used in individual sources as additional columns with the format "scientificNameOriginal_<source>"
† The "TaxonRank" field indicates the level of taxonomic resolution (species or genus) of the observation record. Records of subspecies, varieties, and forms were generally extracted from original sources, but considered at the species level (using the genus and specific epithet).
‡ Higher order taxonomy (phylum, class) follows OBIS (taxonomic backbone; retrieved 6 March 2017), which relies on the World Register of Marine Species (www.marinespecies.org).
§ These fields indicate the organization or institution by which original records were collected.
‖ These fields are indicators of different research cruises or resources, to which original records belonged.
+ "individualCount" and "observedindividualcount" had equivalent entries for records that overlapped between GBIF and OBIS, and were merged into one column.
§§ datasetKey_gbif and resourceID_obis are keys to access metadata of original datasets in GBIF and OBIS via API, including information on sampling methods.

25.7% of data with entries) and "maximumDepthInMeters" (Darwin Core term; 24.0% of data with entries). To retain depth precision information from both GBIF and OBIS, we integrated "depthprecision" into the column "depthAccuracy", presented together with a column on "depth" of sampling. To indicate the source from which records were obtained (GBIF, OBIS, MareDat, Villar or Sal) and the year of data access, we added the columns "sourceArchive" and "yearOfDataAccess". Last, we added a quality flag column, termed "flag". This column flags records with originally negative collection depth (N) changed to positive (sect. 2.2.1), unrealistic day (D) or year (Y) entries (sect. 2.2.2), and/or records collected from sediment samples or traps (S) rather than seawater samples (sect. 2.3.2). We concatenated the sources into a raw database, which contained 1.51 million depth-referenced occurrence records, 3300 phytoplankton species (including five genera) and 247 385 sampling events (Table 2). Sampling events are thereby (and herein) defined as unique combinations of decimalLongitude, decimalLatitude, depth, and time (year, month, day) in the occurrence data.

2.3.1 Extant species selection and taxonomic harmonization

We strived for a selection of occurrence data of extant phytoplankton species and a taxonomic harmonization of their multiple spelling variants (merging synonyms, while clearing misspellings or unaccepted names). This procedure included three steps:

(i) We discarded all species (and their data) that did not have any depth-referenced record. This choice was made on the basis that these species may have been predominantly recorded via fossil materials or have been associated with large uncertainty with respect to their sampling depth, which would infringe the scope of our database.

(ii) We extracted all scientific names (mostly at species level, including all synonyms and spelling variants) associated with at least one depth-referenced record from the raw database (Table 2). This resulted in 3300 names, which were validated in August 2017 against the 150 000+ specific and infraspecific names in AlgaeBase (www.algaebase.org), and matched using a relational database of current names and synonyms; orthography was made as compatible as possible

Table 2: Summary statistics of the raw database by source

Source	Number of observations (%unique to source)		Number of species* (%unique to source)		Number of observations (%unique to source)		Number of species* (%unique to source)	
	full data				data with depth-reference			
GBIF	970 927	(65.6)	3 977	(60.4)	908 995	(64.2)	2676	(51.5)
OBIS	853 981	(60.5)	2 305	(25.2)	823 968	(60.1)	1812	(25.4)
MareDat	102 621	(94.6)	123	(1.1)	102 467	(94.7)	123	(1.5)
Villar et al.	202	(100.0)	87	(0.0)	202	(100.0)	87	(0.0)
Sal et al.	5891	(100.0)	314	(0.0)	5867	(100.0)	313	(0.1)
Total	**1 594 649**		**4741**		**1 511 351**		**3300**	

Numbers of observations (with % of observations unique to the source in parentheses) and the numbers of species (with % of species unique to the source in parentheses) presented for each data source. 27 537 observation records of Picoeukaryotes (not identified to species or genus level) are included among the total records and stem from MareDat (all of which contained a depth-reference).

*Including synonyms or spelling variants.

with the International Code of Nomenclature (Turland et al., 2018), particularly in relation to the gender of specific epithets. This screening led to the exclusion of 459 names (and their data), which could not be traced back to any taxonomically accepted name at the time of query, and to the creation of a "synonymy table" in which each original name (including its potentially multiple synonyms and spelling errors) was matched to a corrected or accepted name.

(iii) We excluded species (and their data) classified as "fossil only" or "fossil", based on AlgaeBase (www.algaebase.org, accessed August 2017) or the World Register of Marine Species (WoRMS; www.marinespecies.org, accessed August 2017). We also excluded species belonging to genera with fossil types denoted by AlgaeBase, under the condition that these species lacked habitat information on AlgaeBase, assuming that the latter species have been collected based on sedimentary or fossilized materials. Species uniquely classified as "freshwater" on AlgaeBase were discarded, as these were beyond the scope of our open ocean database. However, we retained species classified as "freshwater", which had at least 24 open ocean (sect 2.2) records and thus were assumed to thrive also in marine habitats: *Aulacoseira granulata, Chaetoceros wighamii, Diatoma rhombica, Dinobryon balticum, Gymnodinium wulffii, Tripos candelabrum, Tripos euarcuatus*. These cleaning steps led to a remaining set of 2032 original species names, synonyms or spelling variants, corresponding to 1709 taxonomically harmonized species (including five genera not resolved to species level).

2.3.2 Data merger and synthesis

We removed duplicate records, considering the columns "scientificName", "decimalLongitude", "decimalLatitude", "year", "month", "day", and "depth". Removing duplicates meant that any relevant metadata of the duplicated (and hence removed) records were added to the metadata of the record retained, either in an existing or additional column (e.g., information on the original dataset-keys, two which the merged records belonged). We assigned the corrected and/or harmonized taxonomic species name to each original species name in the database on the basis of the synonymy table. We removed duplicates with respect to exact combinations of the harmonized "scientificName", and "decimalLongitude", "decimalLatitude", "year", "month", "day", "depth". This resulted in the harmonized database containing 1 360 621 occurrence records (of which 95.8% had a depth-reference), 1709 species (including five genera), and 242 074 sampling events (Table 3). We retained meta-information on the dataset ID, cruise number, and further attributes when removing duplicates. In particular, we retained the original taxonomic name(s) associated with each record in separate columns of type "scientificNameOriginal_<source>", which allows tracing back the harmonized name to its original name(s). Retaining the original names ensures that future taxonomic changes or updated methods for taxonomic harmonization can be readily implemented. Besides the presences, the final database includes 193 777 count records of individuals or cells, spanning 1126 species. Among these, 105 242 records included a volume basis (spanning 335 species), with a predominant origin from MareDat (n = 99 498) and Sal et al. (2013) (n = 5744). Last, we flagged sedimentary records, indicated by the column "flag". Although we excluded probably many records based on fossil materials during cleaning step (i), this does not exclude the possibility that occurrence records of extant species in the GBIF and OBIS source datasets originated partially from sediment traps or sediment core samples.

Table 3: Summary statistics of the harmonized database by source

Source	Number of observations (%unique to source)		Number of species* (%unique to source)		Number of observations (%unique to source)		Number of species* (%unique to source)	
	full data				data with depth-reference			
GBIF	790 103	(54.9)	1492	(31.5)	751 227	(53.7)	1444	(31.3)
OBIS	823 836	(56.3)	1320	(21.6)	796 907	(56.0)	1283	(22.0)
MareDat	101 969	(94.7)	120	(2.7)	101 816	(94.8)	121	(2.7)
Villar et al.	202	(100.0)	87	(0.0)	202	(100.0)	87	(0.0)
Sal et al.	5744	(100.0)	291	(0.0)	5721	(100.0)	290	(0.0)
Total	**1 360 765**		**1709**		**1 303 721**		**1709**	

Numbers of observations (with % of observations unique to the source in parentheses) and numbers of species (with % of species unique to the source in parentheses) presented for each data source.

*Including 1711 species names and the genera *Phaeocystis*, *Trichodesmium*, *Richelia*, *Prochlorococcus* and *Synechococcus*. 27 537 observation records of Picoeukaryotes (not identified to species or genus level) are included among the total records and stem from MareDat (all of which contained a depth-reference).

Marine sediments can conserve phytoplankton cells that are exported to depth. We flagged phytoplankton records from OBIS and GBIF in the database associated with surface sediment traps or sediment cores (using an "S" in the flag column) by checking the metadata of each individual source dataset of GBIF (using the GBIF datasetKey) and OBIS (using the OBIS resourceID), using the function *datasets* in the R package *rgbif* (Chamberlain, 2015) and the online portal of OBIS (http://iobis.org/explore/#/dataset, accessed 24 October 2018). This check resulted in the flagging of 2.7% of records. We did not attempt to clean or remove sediment type records in MareDat, assuming that information on sampling depth, associated with records of MareDat led to the exclusion of sedimentary records previously. Data from Sal et al. (2013) and Villar et al. (2015) were uniquely based on seawater samples.

3 Results

3.1 Data

3.1.1 Spatiotemporal coverage

Phytoplankton occurrence records contained in PhytoBase cover all ocean basins, latitudes, longitudes, and months (Fig. 1). However, data density is globally highly uneven (Fig 1B, C; histograms) with 44.7% of all records falling into the North Atlantic alone, while only 1.4% of records originate from the South Atlantic, and large parts of the South Pacific basin are devoid of records (Fig. 1A). Analyzing the data by latitude (Fig. 1B) and longitude (Fig. 1C) reveals that sampling has been particularly thin at high latitudes (>70°N and S) during wintertime. Occurrences cover a total of 18 863 monthly cells of 1° latitude × 1° longitude (using the World Geodetic System of 1984 as the reference coordinate system; WGS 84), which corresponds to 3.9% of all monthly (n = 12 months) 1° cells of the open ocean (definition; sect. 2.2). Without monthly distinction, records cover 6098 spatial 1° cells, which is a fraction of 15.5% of all 1° cells of the open ocean.

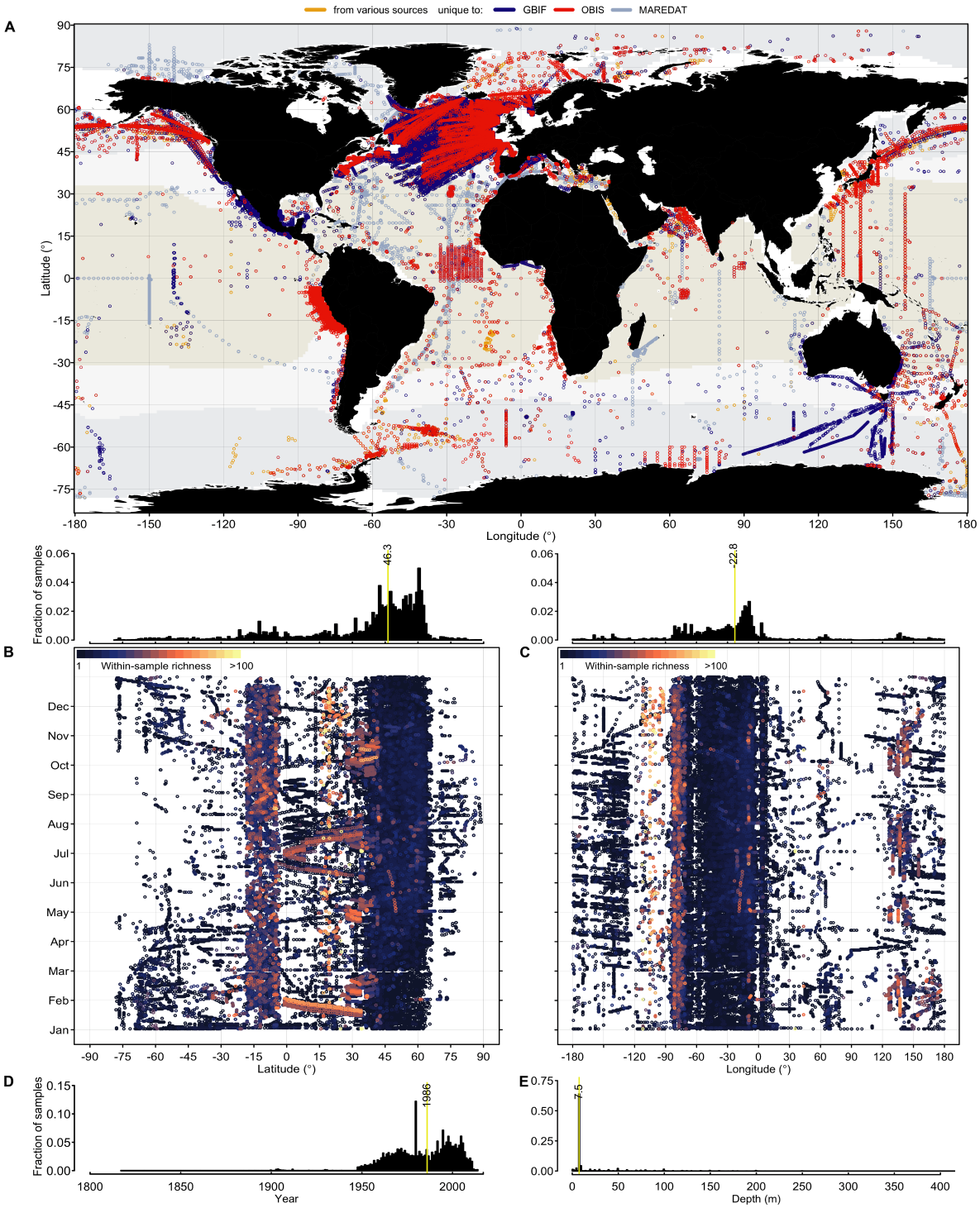

Figure 1: Global distribution of phytoplankton occurrence records of PhytoBase. (**A**) Circles show the position of in situ occurrence records (n = 1 360 765, including 1 280 103 records at the level of species), with the color indicating the source of the data. Map shading indicates the extent of tropical (T >20°C; yellow), temperate (10°C≤ T≤ 20°C; snow-white), and cold (T <10°C; light-blue) seas, based on

the annual mean sea surface temperature (Locarini et al., 2013). (**B-C**) Records plotted as a function of month and latitude (B) or longitude (C). Colors of dots show the number of species detected in each sample (defined as any exact combination of time, location, and depth, in the final dataset). Histograms above panels (B-C) show the frequency of these samples by latitude (B) or longitude (C). (**D-E**) Histograms of sample frequency by year (D), and by depth (E). Vertical yellow lines show the median.

Record quantities are not evenly balanced between major phytoplankton taxa, and global sampling schemes differ between these taxa (Fig. 2). CPR observations are highly condensed in the North Atlantic (and to a lesser extent south of Australia) for the *Bacillariophyceae* and *Dinophyceae* (Fig. 2A, B), but this aggregation is less clear for the *Haptophyta* (Fig. 2C), whose species have typically much smaller cells (often <10 µm) than species of the former two taxa. These three principal phytoplankton taxa have been well surveyed along the north-south AMT cruises, but they lack data in large areas of the South Pacific. Among the less species-rich taxonomic groups, including the *Cyanobacteria* (Fig. 2D) and *Chlorophyta*, global occurrence data coverage has been sparser (Fig. 2D, E). Since all of the principal taxa (Fig. 2) are globally abundant and widespread, the distribution of data indicates sampling efforts rather than a lack of phytoplankton.

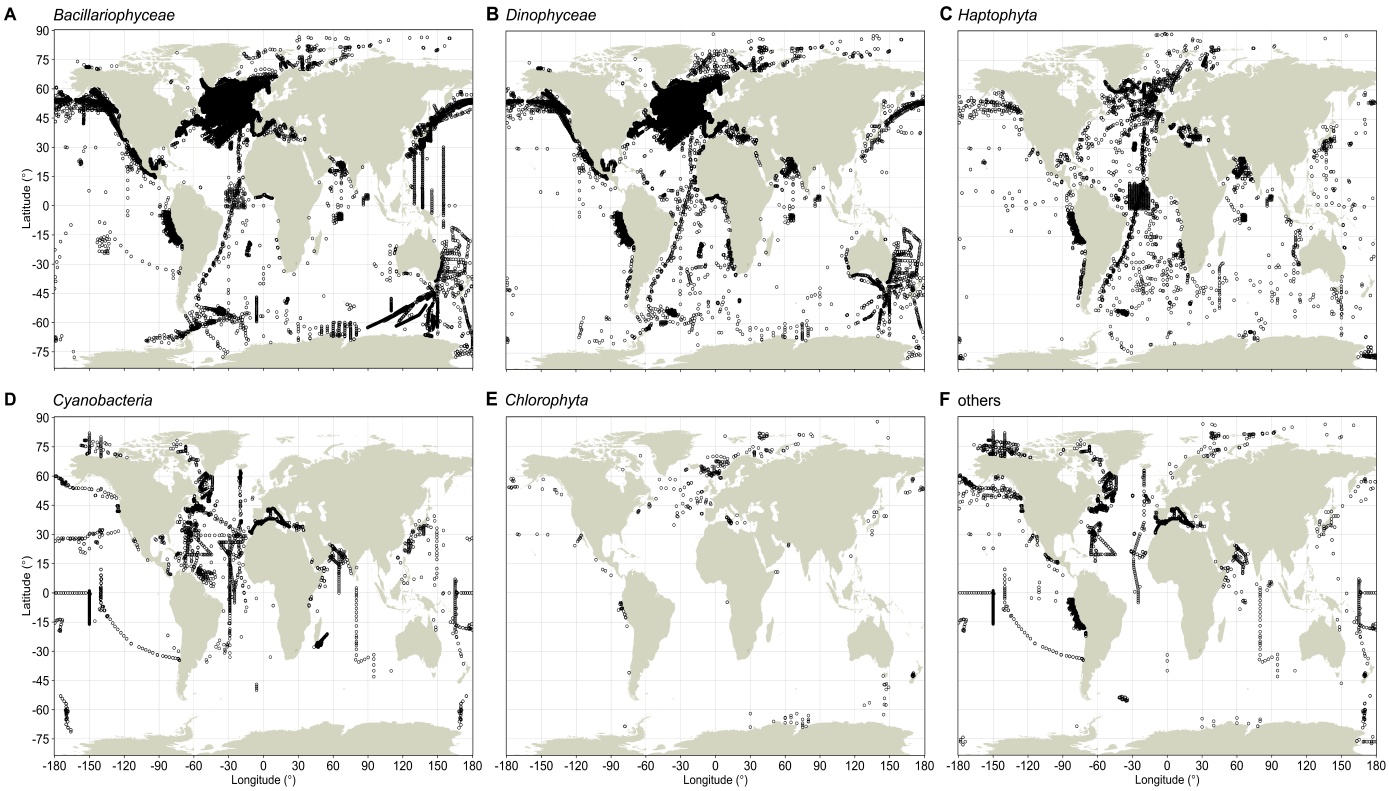

Figure 2: Global distribution of phytoplankton occurrence records in PhytoBase for individual taxa. Black circles show the distribution of in situ records for the five largest phyla or classes in the database that constitute 97.6% of all records (**A-E**) and for the remaining taxa (**F**). Records may overlap at any particular location.

3.1.2 Environmental coverage

The phytoplankton occurrences cover the entire temperature range and a broad part of nitrate and mixed layer conditions found in the global surface ocean (Fig. 3A, B). To visualize the environmental data coverage, figure 3 matches the occurrence records of PhytoBase with climatological sea surface data on nitrate (Garcia et al., 2013), temperature (Locarini

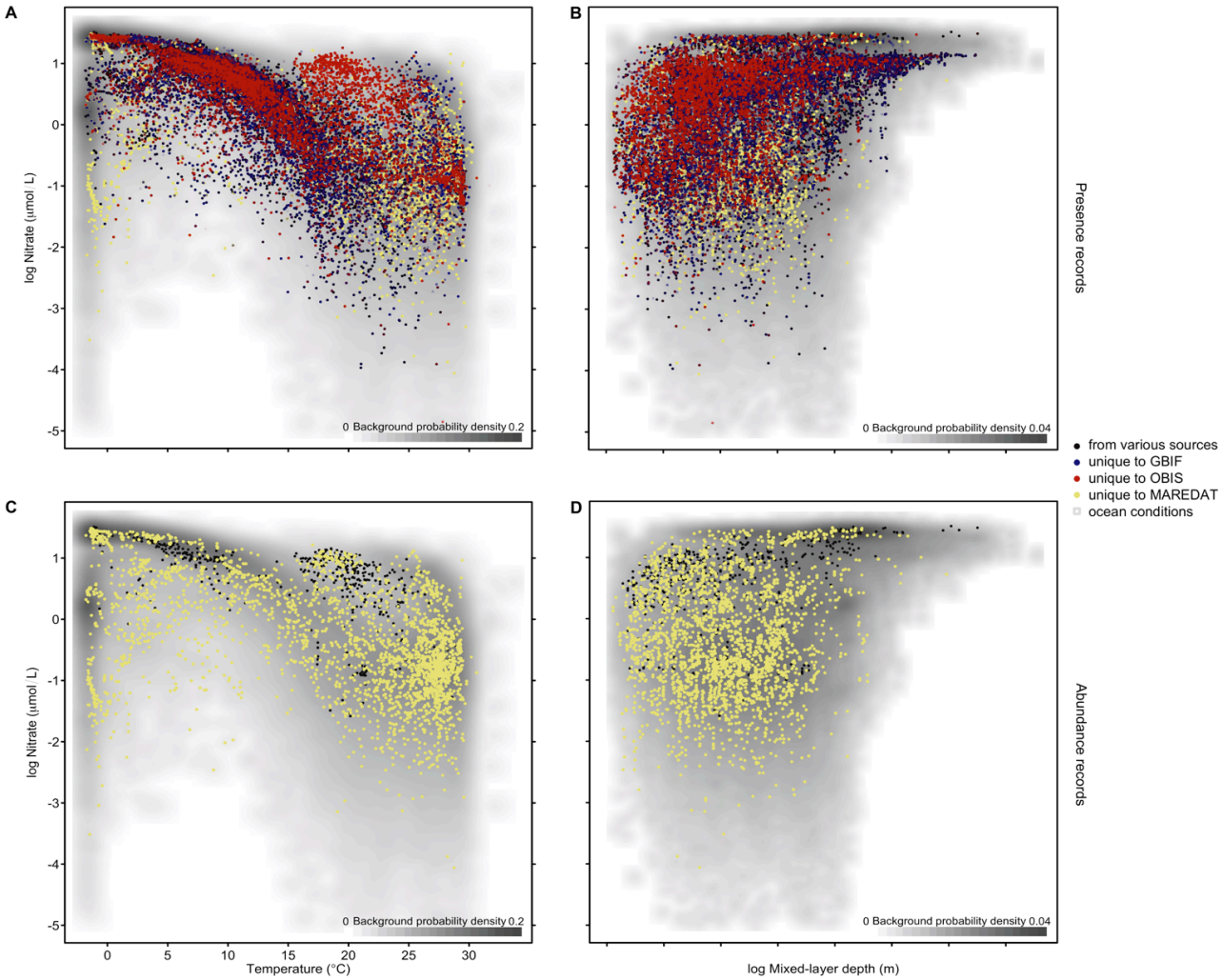

Figure 3: Phytoplankton records in environmental parameter space. (**A-B**) Dots display in situ records (n = 1 360 621) as a function of sea temperature and nitrate concentration (A), and as a function of mixed-layer depth (MLD) and nitrate concentration (B). The scale is logarithmic for MLD and nitrate. Shading indicates the frequency of environmental conditions appearing in the open ocean at surface, with darker grey shade indicating higher frequency (bivariate Gaussian kernel density estimate). The colors of the dots denote the source of data, indicating complementarity or overlap of the environmental gradients sampled between sources. (**C-D**) Show the subset of records that contain information on species' cell counts per liter (n = 105 242), stemming largely from MareDat.

et al., 2013) and mixed-layer depth (de Boyer Montégut, 2004) at monthly 1° × 1° resolution. Records are concentrated in areas with intermediate conditions, which are relatively more frequent at the global scale (gray shade; Fig. 3A, B). Data on cell counts (7.7% of total) show a similar coverage as the full data (Fig. 3A, B), but are much thinner (Fig. 3C, D).

3.1.3 Taxonomic coverage

We assessed what fraction of the known marine phytoplankton species (Falkowski et al., 2004; Jordan, 2004; de Vargas et al., 2015) is represented in PhytoBase. The records compiled include all major taxa of marine phytoplankton known (n = 16 classes), including the *Bacillariophyceae*, *Dinophyceae*, and *Haptophyta*. Records span roughly half of the known marine species of the *Haptophyta* (Jordan, 2004) and a similar fraction of the known marine species of *Bacillariophyceae* and *Dinophyceae* (Table 4). By contrast, species of the less species rich taxa tend to be more strongly underrepresented and account for a relatively small fraction (<3%) of all species in PhytoBase.

Record quantities are unevenly distributed between individual species (Fig. 4). Half of the species contain at least 30 presence records, but multiple species contribute one or two records (Fig. 4A). The species with less than 30 records account for as little as 0.54% of all species records in PhytoBase. Similarly, half of all genera contain at least 110 records each, while genera with less than 110 records each contribute as little as 8.2% to the total of records. A similar data distribution applies to the subset of species (n = 330), for which cell count entries (with volume reference) are available (Fig. 4B). Half of these species contribute at least 16 records, and among genera with cell counts, half contribute at least 76 records.

3.1.4 Completeness of species richness inventories at large spatial scales

We analyzed the ocean inventory of phytoplankton species richness in the database for three different regimes of ocean temperature by means of species accumulation curves (SACs) (Thompson and Withers, 2003) (Fig. 5). These curves present the cumulative species richness detected as a function of sampling effort (or survey area) and they are expected to increase asymptotically before they saturate above a certain threshold of sampling effort (i.e., when the system has been exhaustively sampled). Using the number of sampling events (i.e., unique combinations of time, depth, location in our database) as a surrogate for sampling effort (x-axis), we find that the richness detected (y-axis) and the completeness of species richness detection (degree of saturation), differ notably between regimes. In the Southern temperate– (Fig. 5E) and cold seas (Fig. 5F), species richness has been incompletely sampled with respect to all taxa (black lines) or key taxa (colored lines). By contrast, SACs in the Northern Hemisphere start to saturate at ~40 000 samples, suggesting that the sampling has recorded a majority of the species. Specifically, SACs suggest that species richness will saturate at around ~1500 species in the tropical regime (>20°C), at ~1100 species in northern mid latitudes (≥10°C, ≤ 20°C), and at ~600 species in the cold Northern seas (<10°C). This corresponds to 93%, 64% and 35% of the ~1700 species collected in PhytoBase, respectively. However, these estimates only represent the fraction of species detectable via light microscopy and other methods underlying our database, preferentially omitting very rare or small species (Cermeño et al., 2014; Ser-Giacomi etal., 2018; Sogin et al., 2006). Thus, the richness will likely increase (at low rates) with additional sampling efforts. Theoretical models have suggested that

Table 4: Statistics on the number of records and species contained in the database for key taxa

Taxon	Range (mean) of known marine species	Sources contributing to database	Records in database	Number of species or taxa in database (%)	% of marine species known
Bacillariophyceae (Cl.)	1800[†]-5000[§] (3400)	GBIF, OBIS, MareDat, Villar et al., Sal et al.	699 111	705 (41.2)	14-39
Dinophyceae (Cl.)	1780[†]-1800[§] (1790)	GBIF, OBIS, Villar et al., Sal et al.	527 293	778 (45.5)	43-44
Haptophyta (Ph.)	300[†,‖]-480[§] (360)	GBIF, OBIS, Sal et al., MareDat	47 183	166 (9.7)	34-55
Chlorophyta (Ph.)	100[§]-128[†] (114)	GBIF, OBIS	1304	22 (1.3)	17-22
Chrysophyceae (Cl.)	130[†]-800[§] (465)	GBIF, OBIS, Sal et al.	288	6 (0.4)	1-5
Cryptophyta (Ph.)	78[†]-100[§] (89)	GBIF, OBIS	2312	11 (0.6)	4-5
Cyanobacteria (Ph.)	150[§]	GBIF, OBIS, MareDat	53 060	7 (0.4)	5
Dictyochophyceae (Cl.)	200[†]	GBIF, Sal et al.	1824	8 (0.5)[‡]	4
Euglenoidea (Cl.)	30[§]-36[†] (33)	GBIF, OBIS	701	3 (0.2)	8-10
Raphidophyceae (Cl.)	4[†]-10[§] (7)	GBIF, OBIS	8	3 (0.2)	30-75
Picoeukaryotes	-	MareDat	27 537	1	-
Total	**4530[†,¶]-16 940[§]** (10 735)	**5**	**1 360 621**	**1710**	**10-38**

Cl., class. Ph, phylum.

The table summarizes the occurrence records for the ten major taxa in PhytoBase and describes to what degree the species in each taxon represent the total number of marine species known (for which exact numbers are still debated; we therefore provide upper and lower bounds, and mean values in parentheses).

[§]Falkowski et al. (2004). This estimate includes both coastal and open ocean taxa, while PhytoBase focuses primarily on data from the open ocean.

[†]de Vargas et al. (2015)

[‖]Jordan et al. (2004)

[‡]Including one species of the syster class *Pelagophyceae*.

[¶]The estimate by de Vargas et al. (2015) excluded prokaryotes. A number of 150 prokaryotes (Falkowski et al., 2004) were added to obtain the mean.

communities with many rare species lead to SACs with "low shoulders" meaning that SACs have a long upward slope to the asymptote (Thompson and Withers, 2003), consistent with our SACs (Fig. 5).

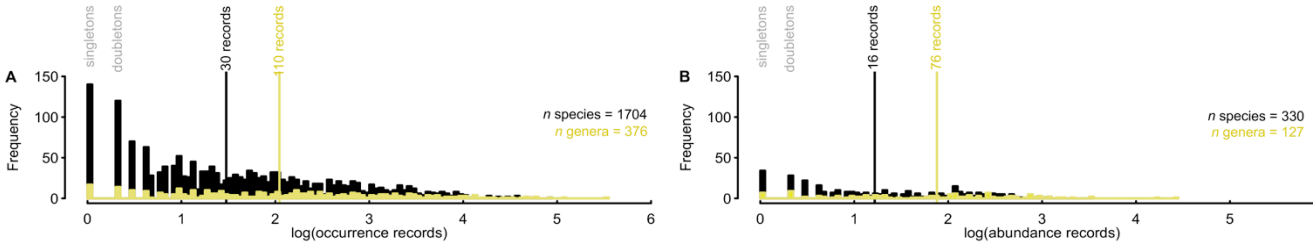

Figure 4: Distribution of occurrence records between species or genera. Histograms show the frequency of species (black) and genera (yellow) with a certain amount of (**A**) presence or (**B**) abundance records, separately. Vertical lines (black, yellow) indicate the median value. *X*-axes are logarithmic to the base ten.

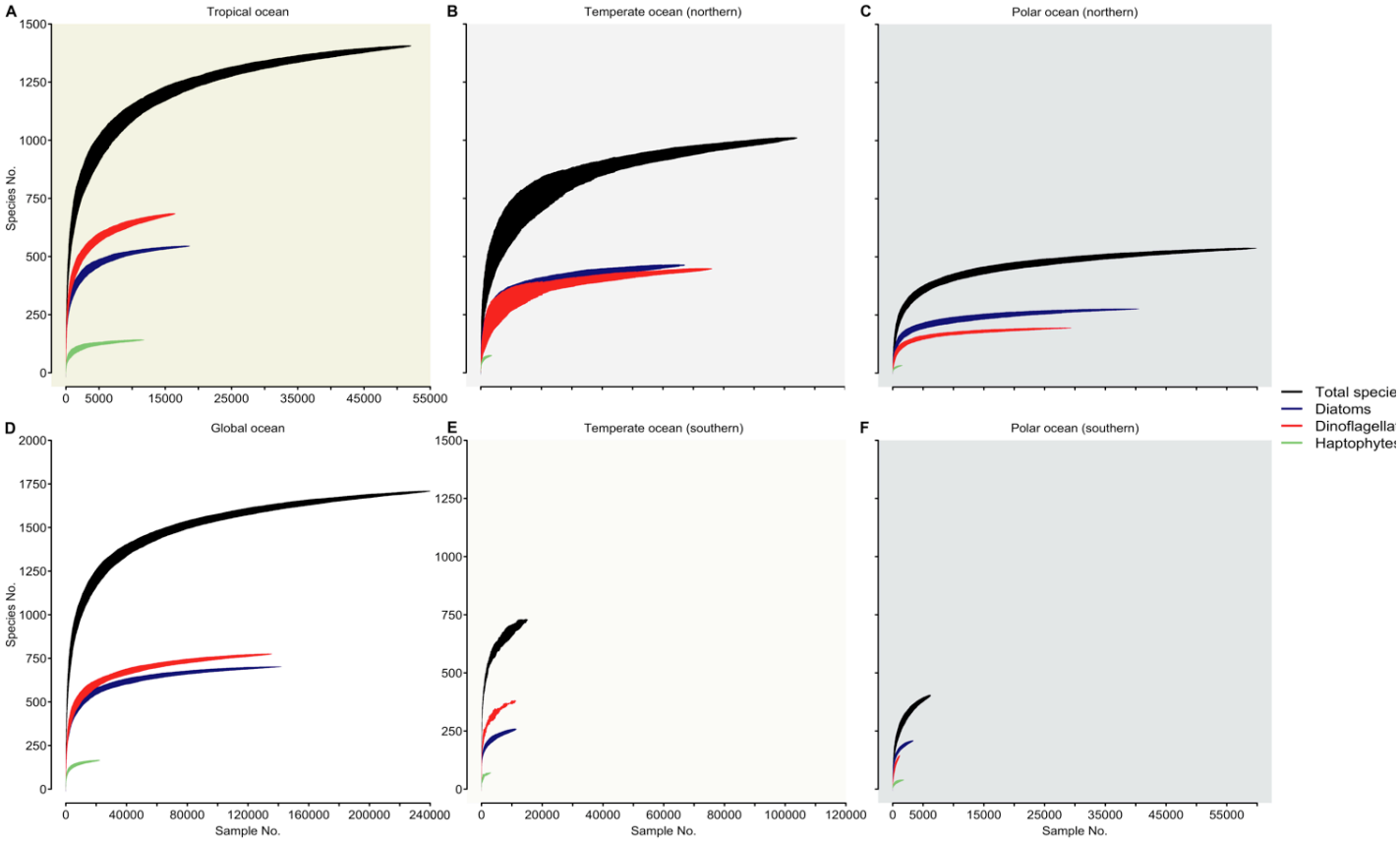

Figure 5: Accumulation of species richness as a function of sampling effort by region. Curves show the cumulative species richness as a function of samples (i.e., unique combinations of space, time and depth in the database, drawn at random) drawn at random from the database, using 100 runs (shadings around the curves indicate ± 1 S.D). Shown are species accumulation curves for all species (black) and three major taxa (colours) for (**A**) the tropics, defined as regions with a sea surface temperature (T) >20°C. (**B**) Temperate seas (10°C≤ T≤ 20°C) of the Northern Hemisphere. (**C**) Cold seas (T< 10°C) of the Northern Hemisphere. (**D**) Global ocean. (**E**) Temperate seas (10°C≤ T≤ 20°C) of the Southern Hemisphere. (**F**) Cold seas (T< 10°C) of the Southern Hemisphere. Background colors refer to figure 1A.

3.1.5 Species richness documented within 1° cells

To explore how completely species richness has been sampled at much smaller spatial scales, we binned data at 1° × 1° resolution, and analyzed the number of species in the pooled data per cell as a function of sampling effort. Hotspots in directly observed phytoplankton richness at the 1° cell level emerge in near-shore waters of Peru, around California, south-east of Australia, in the North Atlantic, along AMT cruises, and along research transects south of Japan (Fig. 6A). The species richness detected per 1° cell is positively correlated with sampling effort, using the number of samples collected per cell as a surrogate of sampling effort (Spearman's ρ = 0.47, P < 0.001). In particular, richness of *Bacillariophyceae* (ρ = 0.88, P < 0.001) and of *Dinophyceae* (ρ = 0.92, P < 0.001), is positively correlated with effort, while this is less the case for *Haptophyta* (ρ = 0.27; P < 0.001). Analyzing species richness as a function of "sampling events" for different thermal

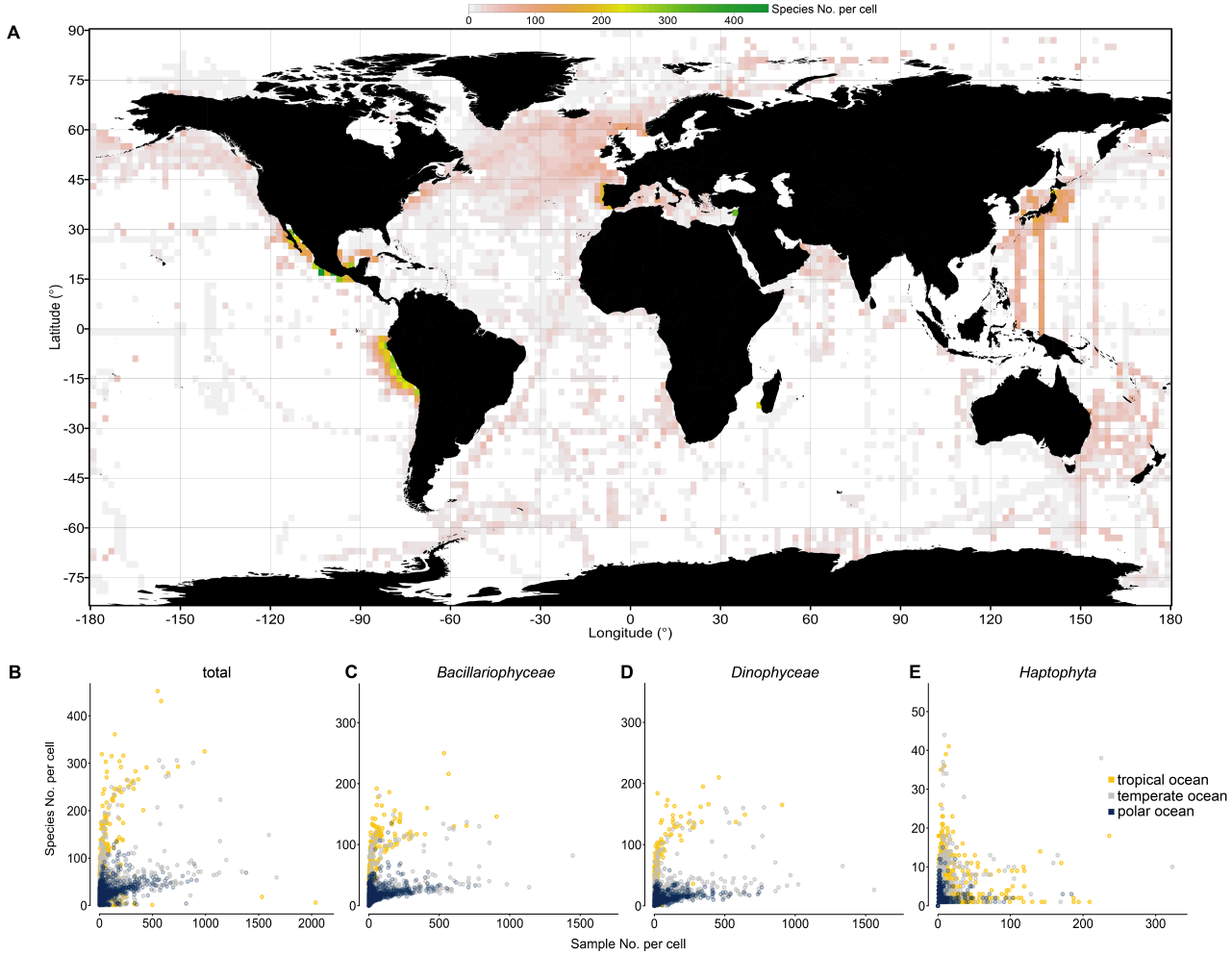

Figure 6: Species richness observed within 1° cells. (**A**) Global map visualizing the species richness detected within each 1° latitude x 1° longitude cell of the ocean. (The means of four 1° cells are depicted at 2°-resolution). (**B-E**) The number of species detected within each 1°-cell is plotted as a function of sampling effort (i.e., number of sampling events, defined as unique combinations of position, time and depth in the database), with colours indicating data originating from different regions: tropical (T >20°C; yellow), temperate (10°C≤ T≤ 20°C; snow-white), and polar 1° cells (T< 10°C; light-blue), as defined by the annual mean temperature at sea surface (Locarini et al., 2013; see shading of map in figure 1). The richness-effort relationship is shown for all taxa (B), and major taxa separately (C-E).

regimes separately reveals that tropical areas (yellow dots; Fig. 6B-E) yield higher cumulative per cell richness at moderate to high sampling effort (> 50 samples), than temperate (grey dots) and polar areas (blue dots). Although data are thin and scattered, species richness in cold areas tends to saturate at ~70 species per cell (Fig. 6B; blue dots) at an effort of ~500 samples collected per cell. In contrast, species richness of the tropical areas tends to reach ~290 species per cell at the same effort (~500 samples). This suggests that tropical phytoplankton richness at the cell level is about four times higher than that of cold northern areas, but richness may further increase with additional sampling effort. Analyzing the data of the major

taxa separately suggests that ~200 species of *Bacillariophyceae* and *Dinophyceae* can be collected at high sampling effort (~500 samples) per 1° cell, yet data are sparse for the *Haptophyta*, which generally lack 1° cells with more than 100 samples available (Fig. 6E).

The analysis of detected species richness per 1° cells suggests that approximately one third to one fifth of all species inventoried in the entire tropical or polar regime (see Fig. 5) might be detected within a single well-sampled 1° cell of the same regime (above ~500 samples) (Fig. 6B). This result is in coarse agreement with the result obtained at the large spatial scale (Fig. 5), where the cumulative richness in the tropical regime was close to three times that of the northern cold regime.

3.1.6 Comparative spatial and taxonomic analysis of source datasets

We considered the sources obtained from within the GBIF archive as an exemplary case for a more detailed analysis of original source dataset coverage, as GBIF provided relatively detailed information on its sources via dataset keys. CPR is the single largest source dataset obtained from GBIF, which covers the North Atlantic and North Pacific (Fig. 7A-D; brown dots), and parts of the ocean south of Australia (Fig. 7A-D; blue dots). CPR records obtained via GBIF contribute 33.9% to all records in PhytoBase. CPR data show relatively low species numbers captured on average per "sample" (Fig. 7I), with samples being defined as exact combinations of geographic position, depth, and time in the data records. This may be owing to the continuous collection of species or incomplete reporting of taxa. The mesh size of the silk employed in CPR of 270 µm undersamples small phytoplankton species (<10 µm). Yet, small species nevertheless get regularly captured in CPR, as they get attached to the screens (Richardson et al., 2006). Within the 16 largest source datasets obtained via GBIF, the average number of species collected per sample is below four for the CPR program and increases to more than 50 for other datasets (Fig. 7I). These 16 test datasets (excluding datasets containing sedimentary records) highlight that the taxonomic resolution strongly differs between samples of individual cruises or survey programs. By latitude, different surveys or cruises thus contribute to PhytoBase to a varying degree (Fig. 7E-H). Systematic differences in the species detected per sample and the varying contribution of sources to the database along latitude (Fig. 7E-H) are important considerations when, for example, analyzing species richness directly.

Analysing the 16 largest source datasets from GBIF in environmental parameter space (Fig. 8) reveals that different domains of the global sea surface temperatures, nitrate levels or mixed-layer depths have been sampled (Fig. 8). Datasets originating from the tropics and subtropics (mean temperature of sampling of 20°C or higher; Fig. 8A) tend to be associated with higher taxonomic detail (~25 species detected per sample on average; Fig. 7I), compared to datasets collected in colder areas. Yet, this likely also reflects an overall higher number of species occurring in tropical areas (Figs. 5A) than in extratropical ones.

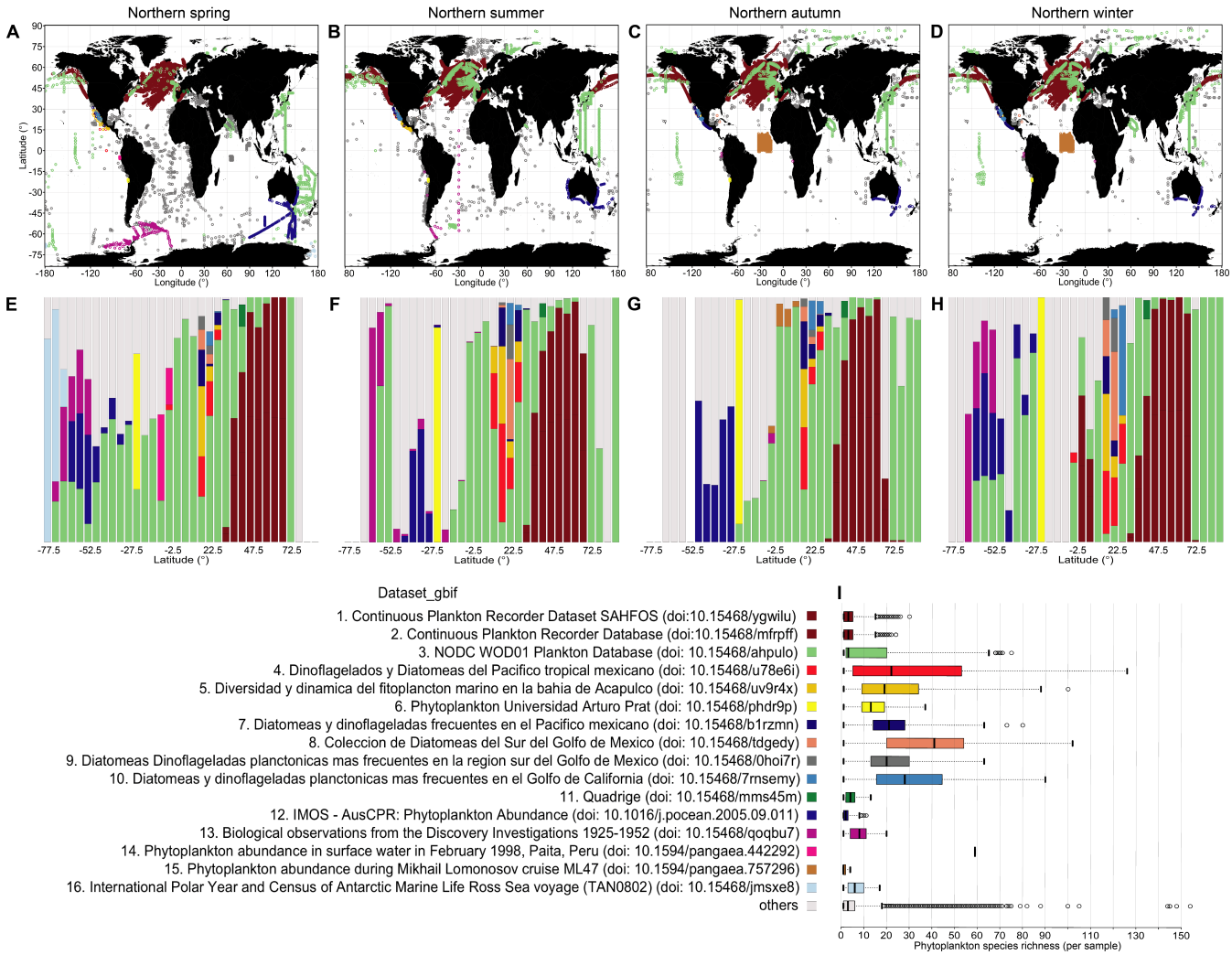

Figure 7: Spatial extent of the 16 largest datasets from GBIF and average per-sample richness. (**A-D**) Maps display the spatial distribution of the 16 largest contributing datasets to the GBIF-sourced data in PhytoBase, for each season separately. The datasets presented comprise 54.8% of all records and 94.0% of GBIF-sourced records. GBIF data is shown as an exemplary case, as it contributes a variety of source datasets defined by dataset keys (datasetKey_gbif). Panels (**E-H**) show the importance of contributing datasets, by latitude. The width of coloured sub-bars reflects the amount of occurrences from each dataset, in 5° latitude bands. Panels (E–H) correspond to the data shown in (A–D). (I) Boxplots highlight the average species richness (thick vertical lines) detected per sample in each dataset, and the first and third quartiles for richness distribution around the mean (boxes). Whiskers denote 2.5 times the inter-quartile range. Note that the same analysis may be performed for OBIS-sourced data using the field "resourceID_obis".

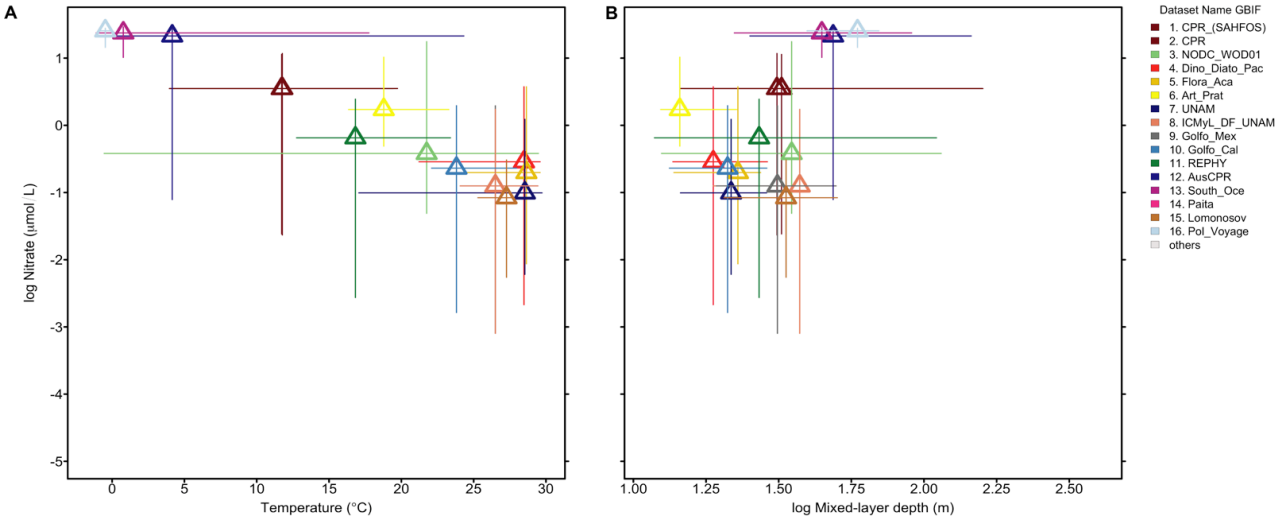

Figure 8: Environmental range of the 16 largest datasets from GBIF. (**A-B**) The range of 16 datasets contained within GBIF-sourced data, and the range of the dataset from Sal et al. (2015) are represented by thin lines in parameter space: (A) temperature *vs.* logarithmic nitrate concentration in the surface ocean, and (B) logarithmic mixed-layer depth vs. logarithmic nitrate (using climatological environmental data from Garcia et al., 2013; Locarini et al., 2013; de Boyer Montégut, 2004; matched with records at monthly climatological 1°-resolution). Lines span the minimum to maximum environmental condition associated with the records of each dataset separately. Triangles display the mean environmental condition of the records per dataset.

3.1.7 Sensitivity of data to taxonomic harmonization and coordinate rounding

While GBIF-derived data contributed roughly 14% more records to the raw database than OBIS (Table 2), this relative contribution changed after the harmonization of species names and their synonyms. GBIF finally contributed 790 103 records, and OBIS 823 836 records to the harmonized PhytoBase. Hence, the exclusion of non-marine, fossil or doubtful species and the taxonomic harmonization step, were overall more stringent for GBIF-sourced than OBIS-sourced data.

We tested to what degree the number of unique records in the harmonized database changed when rounding decimal positions in the raw data from each of the five data sources, prior to their merger. We find that the total number of unique records in PhytoBase declines continuously from 1.36 million to 1.07 million, when rounding the coordinates of records in the raw data to the 6th, 5th, 4th, 3rd, and 2nd decimal place. This result may be explained by the fact that large parts of the data came from CPR. Records collected by CPR are progressively binned into coarser sampling units when rounding their decimal positions. The harmonized database (without coordinate rounding) gained 65.2% occurrence records, relative to its largest individual source archive. This gain was similar in magnitude for the non-harmonized raw database and increased to ca. 73% when rounding coordinates to varying decimals. This shows that the different sources contribute a substantial fraction of unique records to PhytoBase, irrespective of the coordinate rounding to varying decimals.

4 Discussion

4.1 Data coverage, uncertainties, and recommendations

Spatiotemporal data on species occurrence are an essential basis to assess and forecast species' distributions and to understand the drivers behind these patterns. Following recent calls to gather species occurrences into global databases (Edwards, 2000; Meyer et al., 2015), we merged occurrence data of marine phytoplankton from three data sources and from the two largest open access biological data archives into PhytoBase. This new database contains 1 360 621 records (1 280 103 records at the level of species), including 1716 species of seven phyla. Our effort addresses a gap in marine species occurrence data, as previous studies of marine taxa (Tittensor et al. 2010; Chaudhary et al. 2016; Menegotto & Rangel 2018) had no easy access to data sufficiently complete for global analyses of phytoplankton. The synthesis and harmonization of GBIF data with OBIS and other sources results in a substantial gain of phytoplankton occurrence records (> 60% additional records), relative to phytoplankton records residing in either of the two archives. The harmonization of different archives striving to gather global species evidence, therefore substantially expanded the empirical basis of phytoplankton records.

PhytoBase presents, to our knowledge, the currently largest global database of marine phytoplankton species occurrences. However, two main limitations remain: First, the global data density is spatiotemporally highly uneven and gaps persist across large swaths of the ocean, e.g., in the South Pacific and the central Indian. Second, sampling priorities with respect to taxonomic groups, size classes or species resolution differ widely between research cruises and programs. While small or fragile species may escape detection by the CPR program (Richardson et al., 2006), the resolution of traditional samples is influenced by sampling volume and taxonomic expertise (Cermeño et al., 2014). Our results show that the average number of species detected per sample varies from three to above 50 between different cruises or programs. A global spatial bias in collection density of marine species has been similarly found for heterotrophic taxa (Menegotto & Rangel 2018), but sampling biases and divergent sampling protocols may be even more common for phytoplankton.

Owing to these limitations, we recommend that direct analyses we recommend that direct analyses of the database be undertaken and interpreted with caution. For example, our data analysis has shown that direct species richness estimates are sensitive to the number of sampling events. In addition, many species have low occurrence numbers in the database, making any inference about their ecological niche or geographic distribution very uncertain. Thus, without careful screening and checking of the data (via e.g. datasetKeys for GBIF records, resourceIDs for OBIS records), the characterization of biogeographies at the species level might be highly biased.

Statistical techniques such as rarefaction (Rodríguez-Ramos et al., 2015), randomized resampling (Chaudhary et al., 2017), analysis of sampling gaps (Woolley et al. 2016; Menegotto & Rangel 2018), and species distribution modeling (Zimmermann and Guisan, 2000) may be implemented to overcome these limitations. The latter statistical technique may be particularly promising, as species distribution models can be set up to account for variation in presence data sampling (Phillips et al., 2009) and data scarceness (Breiner et al., 2015). Based on observed associations between species'

occurrences and environmental factors (Guisan and Thuiller, 2005), these models estimate the species' ecological niche, which is projected into geographic space, assuming that the species' niche and its geographic habitat are directly interrelated (Colwell and Rangel, 2009). Another advantage of species distribution models is that they can circumvent geographic sampling gaps through a spatial projection of the niche, as long as environmental conditions relevant to describe the niche of the species have been sufficiently well sampled and the species fills its ecological niche. This is the approach used by Righetti et al. (2019b), building on a large fraction of the PhytoBase (77.6% of the records, accessed in 2015 and falling into the monthly climatological mixed-layer; de Boyer Montégut, 2004), to analyze global richness patterns of phytoplankton.

DNA sequencing has become an alternative approach to characterize phytoplankton biogeography (de Vargas et al., 2015). These data have two advantages over traditional taxonomic data: First, the sensitivity of metagenomic methods to detect rare taxa is relatively much higher. The detection of rare species and their integration into PhytoBase may hence become possible via molecular methods (Bork et al., 2015; Sogin et al., 2006). Second, metagenomic data have been collected in a methodologically consistent way in recent global surveys, such as the TARA Oceans cruise (de Vargas et al., 2015). But there are also drawbacks associated with DNA based methods. A large disadvantage of current metagenomic data is the lack of catalogued reference gene sequences for most species. As a result, the majority of the metagenomic sequences can only be determined to the level of genus (Malviya et al., 2016). However, we expect that an integration of detailed genetic data with traditional sampling data may soon become possible, allowing to combine several methodological or taxonomic dimensions. At any point in the future, changing taxonomic nomenclature can be implemented in PhytoBase, as we retained the original name variants and synonyms from the raw data sources together with the harmonized name for each record.

4.2 Data use

Our data compilation and synthesis product PhytoBase was designed to support primarily the analysis of the distribution, diversity, and abundance of phytoplankton species and related biotic or abiotic drivers in macroecological studies. But PhytoBase is far from limited to this set of applications, and may include the analysis of ecological niche differences between species or clades, linkages between species' ecological niches and phylogenetic or functional relatedness, current or future spatial projections of species' niches, tests on whether presence-absence patterns of multiple species can predict community trait indices, studies on how well species' traits predict spatial patterns of species, or joint analyses of species' distribution and trait data to project trait biogeographies. The database may also be used to validate the increasingly complex marine ecosystem models included in regional to global climate models.

The accuracy of data analyses may be limited by sampling biases underlying PhytoBase, including the spatiotemporal variation in sampling efforts and varying taxonomic detail between data sources or research cruises. The latter limitation might be alleviated by considering different methodologies associated with varying cruises or collecting organisations in spatial analyses. Where possible, we thus retained the information on the original dataset ID or dataset key along with each occurrence record in the database. Moreover, statistical analysis tools may be used to address spatiotemporal variation in

global sampling efforts. New data from undersampled areas such as the South Pacific will likely lead to new species discoveries and may greatly improve the global observational basis of phytoplankton occurrence data in the future. Data inclusion from recent cruises, which are still under evaluation, appears as a natural next step. These data may come from the Malaspina expedition (Duarte, 2015), TARA Oceans (Bork et al., 2015) and Southern Ocean transects (Balch et al., 2016).

5 Data availability

PhytoBase is publicly available through PANGAEA, doi:10.1594/PANGAEA.904397 (Righetti et al., 2019a). Associated R scripts and the synonymy table used to harmonize species' names are available through https://gitlab.ethz.ch/phytobase/supplementary.

6 Conclusions

In PhytoBase, we compiled more than 1.36 million marine phytoplankton records that span 1704 species including the key taxa *Bacillariophyceae*, *Dinophyceae*, *Haptophyta*, *Cyanobacteria* and others. The database addresses photosynthetic microbial organisms, which play crucial roles in global biogeochemical cycles and marine ecology. We have provided an analysis of the current status of marine phytoplankton occurrence records accessible through public archives, their spatial and methodological limitations, and the completeness of species richness information for different ocean regions. PhytoBase may stimulate studies on the biogeography, diversity, and composition of phytoplankton and serve to calibrate ecological or mechanistic models. We recommend accounting carefully for data structure and metadata, depending on the purpose of analysis.

7 Appendices

Table A1: Harmonization of 113 taxon names in the MareDat dataset of O'Brien et al. (2013). Only the 113 names that changed during harmonization are shown, out of a total of 197 names.

Group	Original name	Harmonized name
Haptophyta	*_P. pouchetii*	*Phaeocystis pouchetii*
	*_P. pouchetii_*	*Phaeocystis pouchetii*
	_Phaeocystis pouchetii	*Phaeocystis pouchetii*
	_Phaeocystis pouchetii (Subcomponent: bladders)	*Phaeocystis pouchetii*
	_Phaeocystis spp.	*Phaeocystis*
	Phaeocystis spp.	*Phaeocystis*
	_Phaeocystis spp. (Subgroup: motile)	*Phaeocystis*
	_Phaeocystis spp. (Subgroup: non-motile)	*Phaeocystis*
	ACANTHOICA QUATTROSPINA	*Acanthoica quattrospina*
	Acanthoica acanthos	*Anacanthoica acanthos*

Acanthoica sp. cf. quattraspina	*Acanthoica quattrospina*
Algirosphaera oryza	*Algirosphaera robusta*
Algirosphaera robsta	*Algirosphaera robusta*
Anoplosolenia	*Anoplosolenia brasiliensis*
Anoplosolenia braziliensis	*Anoplosolenia brasiliensis*
Anoplosolenia sp. cf. brasiliensis	*Anoplosolenia brasiliensis*
Anthosphaera robusta	*Algirosphaera robusta*
CALCIDISCUS leptoporus	*Calcidiscus leptoporus*
Calcidiscus leptopora	*Calcidiscus leptoporus*
Calcidiscus leptoporus (inc. Coccolithus pelagicus)	*Calcidiscus leptoporus*
Calcidiscus leptoporus (small + intermediate)	*Calcidiscus leptoporus*
Calcidiscus leptoporus intermediate	*Calcidiscus leptoporus*
Calciosolenia MURRAYI	*Calciosolenia murrayi*
Calciosolenia brasiliensis	*Anoplosolenia brasiliensis*
Calciosolenia granii v closterium	*Anoplosolenia brasiliensis*
Calciosolenia granii v cylindrothecaf	*Calciosolenia murrayi*
Calciosolenia granii v cylindrothecaforma	*Calciosolenia murrayi*
Calciosolenia granii var closterium	*Anoplosolenia brasiliensis*
Calciosolenia granii var cylindrothecaeiformis	*Calciosolenia murrayi*
Calciosolenia murray	*Calciosolenia murrayi*
Calciosolenia siniosa	*Calciosolenia murrayi*
Calciosolenia sinuosa	*Calciosolenia murrayi*
Calciosolenia sp. cf. murrayi	*Calciosolenia murrayi*
Caneosphaera molischii	*Syracosphaera molischii*
Caneosphaera molischii and similar	*Syracosphaera molischii*
Coccolithus fragilis	*Oolithotus fragilis*
Coccolithus huxley	*Emiliania huxleyi*
Coccolithus huxleyi	*Emiliania huxleyi*
Coccolithus leptoporus	*Calcidiscus leptoporus*
Coccolithus sibogae	*Umbilicosphaera sibogae*
Crenalithus sessilis	*Reticulofenestra sessilis*
Crystallolithus cf rigidus	*Calcidiscus leptoporus*
Cyclococcolithus fragilis	*Oolithotus fragilis*
Discophaera tubifer	*Discosphaera tubifera*
Discosphaera thomsoni	*Discosphaera tubifera*
Discosphaera tubifer	*Discosphaera tubifera*
Discosphaera tubifer (inc. Papposphaera.lepida)	*Discosphaera tubifera*
Discosphaera tubifera	*Discosphaera tubifera*
Emiliana huxleyi	*Emiliania huxleyi*
Emiliania huxleyi A1	*Emiliania huxleyi*
Emiliania huxleyi A2	*Emiliania huxleyi*
Emiliania huxleyi A3	*Emiliania huxleyi*

	Emiliania huxleyi C	*Emiliania huxleyi*
	Emiliania huxleyi Indet.	*Emiliania huxleyi*
	Emiliania huxleyi var. Huxleyi	*Emiliania huxleyi*
	Florisphaera profunda var. profunda	*Florisphaera profunda*
	Halopappus adriaticus	*Michaelsarsia adriaticus*
	Helicosphaera carteri var. Carteri	*Helicosphaera carteri*
	Michelsarsia elegans	*Michaelsarsia elegans*
	Oolithotus fragilis var. Fragilis	*Oolithotus fragilis*
	Oolithus spp. cf fragilis	*Oolithotus fragilis*
	Ophiaster hydroideuss	*Ophiaster hydroideus*
	Ophiaster spp. cf. Hydroides	*Ophiaster hydroideus*
	P. antarctica	*Phaeocystis antarctica*
	*P. antarctica_*	*Phaeocystis antarctica*
	PHAEOCYSTIS	*Phaeocystis*
	*PHAEOCYSTIS_*	*Phaeocystis*
	PHAEOCYSTIS POUCHETII	*Phaeocystis pouchetii*
	*PHAEOCYSTIS POUCHETII_*	*Phaeocystis pouchetii*
	PHAEOCYSTIS sp.	*Phaeocystis*
	PHAEOCYSTIS sp._	*Phaeocystis*
	Palusphaera sp.	*Rhabdosphaera longistylis*
	Palusphaera vandeli	*Rhabdosphaera longistylis*
	*Phaeocystis antarctica_*	*Phaeocystis antarctica*
	Phaeocystis cf. pouchetii	*Phaeocystis pouchetii*
	*Phaeocystis cf. pouchetii_*	*Phaeocystis pouchetii*
	*Phaeocystis globosa_*	*Phaeocystis globosa*
	Phaeocystis motile	*Phaeocystis*
	*Phaeocystis motile_*	*Phaeocystis*
	Phaeocystis sp.	*Phaeocystis*
	Phaeocystis sp._	*Phaeocystis*
	Phaeocystis spp.	*Phaeocystis*
	Pontosphaera huxleyi	*Emiliania huxleyi*
	Rhabdosphaera sp. cf. claviger (inc. var. stylifera)	*Rhabdosphaera clavigera*
	Rhabdosphaera claviger	*Rhabdosphaera clavigera*
	Rhabdosphaera clavigera var. Clavigera	*Rhabdosphaera clavigera*
	Rhabdosphaera clavigera var. Stylifera	*Rhabdosphaera clavigera*
	Rhabdosphaera stylifera	*Rhabdosphaera clavigera*
	Rhabdosphaera tubifer	*Discosphaera tubifera*
	Rhabdosphaera tubulosa	*Discosphaera tubifera*
	Syrachosphaera pulchra	*Syracosphaera pulchra*
	Syracosphaera brasiliensis	*Anoplosolenia brasiliensis*
	Syracosphaera cf. Pulchra	*Syracosphaera pulchra*
	Syracosphaera confuse	*Ophiaster hydroideus*

	Syracosphaera corii	*Michaelsarsia adriaticus*
	Syracosphaera cornifera	*Helladosphaera cornifera*
	Syracosphaera corri	*Michaelsarsia adriaticus*
	Syracosphaera mediterranea	*Coronosphaera mediterranea*
	Syracosphaera molischii s.l.	*Syracosphaera molischii*
	Syracosphaera oblonga	*Calyptrosphaera oblonga*
	Syracosphaera quadricornu	*Algirosphaera robusta*
	Syracosphaera sp. cf. prolongata (inc. S.pirus)	*Syracosphaera prolongata*
	Syracosphaera tuberculata	*Coronosphaera mediterranea*
	Umbellosphaera hulburtiana	*Umbilicosphaera hulburtiana*
	Umbellosphaera sibogae	*Umbilicosphaera sibogae*
	Umbellosphaera spp. cf. irregularis + tenuis	*Umbellosphaera irregularis*
	Umbilicosphaera mirabilis	*Umbilicosphaera sibogae*
	Umbilicosphaera sibogae (Weber-van-Bosse) Gaarder	*Umbilicosphaera sibogae*
	Umbilicosphaera sibogae sibogae	*Umbilicosphaera sibogae*
	Umbilicosphaera sibogae var. Sibogae	*Umbilicosphaera sibogae*
	Umbilicosphaera spp. (U.sibogae)	*Umbilicosphaera sibogae*
	Umbillicosphaera sibogae	*Umbilicosphaera sibogae*

Note. An empty space in the original taxon name is indicated by "_".

560 **Table A2:** Harmonization of 156 taxon names in the MareDat dataset of Leblanc et al. (2012). Only the 156 names that changed during harmonization are shown, out of a total of 248 names.

Group	Original name	Harmonized name
Bacillariophyceae	*Actinocyclus coscinodiscoides*	*Roperia tesselata*
	Actinocyclus tessellatus	*Roperia tesselata*
	Asterionella frauenfeldii	*Thalassionema frauenfeldii*
	Asterionella glacialis	*Asterionellopsis glacialis*
	Asterionella mediterranea subsp pacifica	*Lioloma pacificum*
	Asterionellopsis japonica	*Asterionellopsis glacialis*
	Bacteriastrum varians	*Bacteriastrum furcatum*
	Cerataulina bergonii	*Cerataulina pelagica*
	Cerataulus bergonii	*Cerataulina pelagica*
	Ceratoneis closterium	*Cylindrotheca closterium*
	Ceratoneis longissima	*Nitzschia longissima*
	Chaetoceros angulatus	*Chaetoceros affinis*
	Chaetoceros atlanticus f. bulosus	*Chaetoceros bulbosus*
	Chaetoceros audax	*Chaetoceros atlanticus*
	Chaetoceros borealis f. concavicornis	*Chaetoceros concavicornis*
	Chaetoceros cellulosus	*Chaetoceros lorenzianus*
	Chaetoceros chilensis	*Chaetoceros peruvianus*
	Chaetoceros contortus	*Chaetoceros compressus*
	Chaetoceros convexicornis	*Chaetoceros peruvianus*

Chaetoceros dichaeta	*Chaetoceros distans*
Chaetoceros dispar	*Chaetoceros atlanticus*
Chaetoceros grunowii	*Chaetoceros decipiens*
Chaetoceros jahnischianus	*Chaetoceros distans*
Chaetoceros javanis	*Chaetoceros affinis*
Chaetoceros peruvio-atlanticus	*Chaetoceros peruvianus*
Chaetoceros polygonus	*Chaetoceros atlanticus*
Chaetoceros radians	*Chaetoceros socialis*
Chaetoceros radiculus	*Chaetoceros bulbosus*
Chaetoceros ralfsii	*Chaetoceros affinis*
Chaetoceros remotus	*Chaetoceros distans*
Chaetoceros schimperianus	*Chaetoceros bulbosus*
Chaetoceros schuttii	*Chaetoceros affinis*
Chaetocros vermiculatus	*Chaetoceros debilis*
Corethron criophilum	*Corethron pennatum*
Corethron hystrix	*Corethron pennatum*
Corethron valdivae	*Corethron pennatum*
Coscinodiscus anguste-lineatus	*Thalassiosira anguste-lineata*
Coscinodiscus gravidus	*Thalassiosira gravida*
Coscinodiscus pelagicus	*Thalassiosira gravida*
Coscinodiscus polychordus	*Thalassiosira anguste-lineata*
Coscinodiscus rotulus	*Thalassiosira gravida*
Coscinodiscus sol	*Planktoniella sol*
Coscinodiscus sublineatus	*Thalassiosira anguste-lineata*
Coscinosira polychordata	*Thalassiosira anguste-lineata*
Dactyliosolen mediterraneus	*Leptocylindrus mediterraneus*
Dactyliosolen meleagris	*Leptocylindrus mediterraneus*
Detonula delicatula	*Detonula pumila*
Diatoma rhombica	*Fragilariopsis rhombica*
Dicladia bulbosa	*Chaetoceros bulbosus*
Dithylim inaequale	*Ditylum brightwellii*
Dithylum trigonum	*Ditylum brightwellii*
Eucampia balaustium	*Eucampia antarctica*
Eucampia Britannica	*Eucampia zodiacus*
Eucampia nodosa	*Eucampia zodiacus*
Eucampia striata	*Guinardia striata*
Eupodiscus tesselatus	*Roperia tesselata*
Fragilaria arctica	*Fragilariopsis oceanica*
Fragilaria kerguelensis	*Fragilariopsis kerguelensis*
Fragilaria obliquecostata	*Fragilariopsis obliquecostata*
Fragilaria rhombica	*Fragilariopsis rhombica*
Fragilariopsis antarctica	*Fragilariopsis oceanica*

Fragilariopsis sublinearis	*Fragilariopsis obliquecostata*
Fragilaris sublinearis	*Fragilariopsis obliquecostata*
Fragillariopsis antarctica	*Fragilariopsis kerguelensis*
Gallionella sulcata	*Paralia sulcata*
Guinardia baltica	*Guinardia flaccida*
Hemiaulus delicatulus	*Hemiaulus hauckii*
Henseniella baltica	*Guinardia flaccida*
Homeocladia closterium	*Cylindrotheca closterium*
Homeocladia delicatissima	*Pseudo-nitzschia delicatissima*
Lauderia borealis	*Lauderia annulata*
Lauderia pumila	*Detonula pumila*
Lauderia schroederi	*Detonula pumila*
Leptocylindrus belgicus	*Leptocylindrus minimus*
Melosira costata	*Skeletonema costatum*
Melosira marina	*Paralia sulcata*
Melosira sulcata	*Paralia sulcata*
Moerellia cornuta	*Eucampia cornuta*
Navicula mebranacea	*Meuniera membranacea*
Navicula planamembranacea	*Ephemera planamembranacea*
Navicula pseudomembranacea	*Meuniera membranacea*
Nitzschia actydrophila	*Pseudo-nitzschia delicatissima*
Nitzschia angulate	*Fragilariopsis rhombica*
Nitzschia Antarctica	*Fragilariopsis rhombica*
Nitzschia birostrata	*Nitzschia longissima*
Nitzschia closterium	*Cylindrotheca closterium*
Nitzschia curvirostris	*Cylindrotheca closterium*
Nitzschia delicatissima	*Pseudo-nitzschia delicatissima*
Nitzschia grunowii	*Fragilariopsis oceanica*
Nitzschia heimii	*Pseudo-nitzschia heimii*
Nitzschia kergelensis	*Fragilariopsis kerguelensis*
Nitzschia obliquecostata	*Fragilariopsis obliquecostata*
Nitzschia pungens	*Pseudo-nitzschia pungens*
Nitzschia seriata	*Pseudo-nitzschia seriata*
Nitzschiella longissima	*Nitzschia longissima*
Nitzschiella tenuirostris	*Cylindrotheca closterium*
Orthoseira angulate	*Thalassiosira angulata*
Orthoseira marina	*Paralia sulcata*
Orthosira marina	*Paralia sulcata*
Paralia marina	*Paralia sulcata*
Planktoniella wolterecki	*Planktoniella sol*
Podosira subtilis	*Thalassiosira subtilis*
Proboscia alata f. alata	*Proboscia alata*

Proboscia alata f. gracillima	*Proboscia alata*
Proboscia gracillima	*Proboscia alata*
Pyxilla baltica	*Rhizosolenia setigera*
Rhizosolenia alata	*Proboscia alata*
Rhizosolenia alata f. indica	*Proboscia indica*
Rhizosolenia alata var. indica	*Proboscia indica*
Rhizosolenia amputata	*Rhizosolenia bergonii*
Rhizosolenia antarctica	*Guinardia cylindrus*
Rhizosolenia calcar	*Pseudosolenia calcar-avis*
Rhizosolenia calcar avis	*Pseudosolenia calcar-avis*
Rhizosolenia calcar-avis	*Pseudosolenia calcar-avis*
Rhizosolenia cylindrus	*Guinardia cylindrus*
Rhizosolenia delicatula	*Guinardia delicatula*
Rhizosolenia flaccida	*Guinardia flaccida*
Rhizosolenia fragilima	*Dactyliosolen fragilissimus*
Rhizosolenia fragilissima	*Dactyliosolen fragilissimus*
Rhizosolenia genuine	*Proboscia alata*
Rhizosolenia gracillima	*Proboscia alata*
Rhizosolenia hebetata f hiemalis	*Rhizosolenia hebetata*
Rhizosolenia hebetata f. hebetata	*Rhizosolenia hebetata*
Rhizosolenia hebetata f. semispina	*Rhizosolenia hebetata*
Rhizosolenia hensenii	*Rhizosolenia setigera*
Rhizosolenia indica	*Proboscia indica*
Rhizosolenia japonica	*Rhizosolenia setigera*
Rhizosolenia murrayana	*Rhizosolenia chunii*
Rhizosolenia semispina	*Rhizosolenia hebetata*
Rhizosolenia stolterfothii	*Guinardia striata*
Rhizosolenia strubsolei	*Rhizosolenia imbricata*
Rhizosolenia styliformis var. longispina	*Rhizosolenia styliformis*
Rhizosolenia styliformis var. polydactyla	*Rhizosolenia styliformis*
Rhizosolenia styliformis var. semispina	*Rhizosolenia hebetata*
Schroederella delicatula	*Detonula pumila*
Spingeria bacillaris	*Thalassionema bacillare*
Stauroneis membranacea	*Meuniera membranacea*
Stauropsis membranacea	*Meuniera membranacea*
Synedra nitzschioides	*Thalassionema nitzschioides*
Synedra thalassiothrix	*Thalassiothrix longissima*
Terebraria kerguelensis	*Fragilariopsis kerguelensis*
Thalassionema elegans	*Thalassionema bacillare*
Thalassiosira condensata	*Detonula pumila*
Thalassiosira decipiens	*Thalassiosira angulate*
Thalassiosira polychorda	*Thalassiosira anguste-lineata*

	Thalassiosira rotula	*Thalassiosira gravida*
	Thalassiosira tcherniai	*Thalassiosira gravida*
	Thalassiothrix curvata	*Thalassionema nitzschioides*
	Thalassiothrix delicatula	*Lioloma delicatulum*
	Thalassiothrix frauenfeldii	*Thalassionema frauenfeldii*
	Thalassiothrix fraunfeldii	*Thalassionema nitzschioides*
	Thalassiothrix mediterranea var. pacifica	*Lioloma pacificum*
	Trachysphenia australis v kerguelensis	*Fragilariopsis kerguelensis*
	Triceratium brightwellii	*Ditylum brightwellii*
	Zygoceros pelagica	*Cerataulina pelagica*
	Zygoceros pelagicum	*Cerataulina pelagica*

Table A3: Harmonization of the total of 109 species names in the data from Villar et al. (2015). Only the 109 names that changed during harmonization are shown, out of a total of 201 names.

Group	Original name	Harmonized name
Bacillariophyceae	*Asteromphalus cf. flabellatus*	*Asteromphalus*
	Asteromphalus spp.	*Asteromphalus*
	Bacteriastrum cf. delicatulum	*Bacteriastrum*
	Bacteriastrum cf. elongatum	*Bacteriastrum*
	Bacteriastrum cf. furcatum	*Bacteriastrum*
	Bacteriastrum cf. hyalinum	*Bacteriastrum*
	Bacteriastrum spp.	*Bacteriastrum*
	Biddulphia spp.	*Biddulphia*
	Chaetoceros atlanticus var. neapolitanus	*Chaetoceros atlanticus*
	Chaetoceros bulbosum	*Chaetoceros bulbosus*
	Chaetoceros cf. atlanticus	*Chaetoceros*
	Chaetoceros cf. coarctatus	*Chaetoceros*
	Chaetoceros cf. compressus	*Chaetoceros*
	Chaetoceros cf. danicus	*Chaetoceros*
	Chaetoceros cf. densus	*Chaetoceros*
	Chaetoceros cf. dichaeta	*Chaetoceros*
	Chaetoceros cf. laciniosus	*Chaetoceros*
	Chaetoceros cf. lorenzianus	*Chaetoceros*
	Chaetoceros spp.	*Chaetoceros*
	Climacodium cf. fravenfeldianum	*Climacodium*
	Climacodium spp.	*Climacodium*
	Corethron cf. pennatum	*Corethron*
	Corethron spp.	*Corethron*
	Coscinodiscus spp.	*Coscinodiscus*
	Cylindrotheca spp.	*Cylindrotheca*
	Ditylum spp.	*Ditylum*

	Eucampia antartica	*Eucampia antarctica*
	Eucampia spp.	*Eucampia*
	Eucampia zodiacus f. cylindrocornis	*Eucampia zodiacus*
	Fragilariopsis spp.	*Fragilariopsis*
	Haslea wawrickae	*Haslea wawrikae*
	Hemiaulus spp.	*Hemiaulus*
	Hemidiscus cf. cuneiformis	*Hemidiscus*
	Lauderia spp.	*Lauderia*
	Leptocylindrus cf. danicus	*Leptocylindrus*
	Leptocylindrus cf. minimus	*Leptocylindrus*
	Lithodesmium spp.	*Lithodesmium*
	Nitzschia spp.	*Nitzschia*
	Odontella spp.	*Odontella*
	Pseudo-nitzschia cf. fraudulenta	*Pseudo-nitzschia*
	Pseudo-nitzschia cf. subcurvata	*Pseudo-nitzschia*
	Pseudo-nitzschia delicatissima group	*Pseudo-nitzschia delicatissima*
	Pseudo-nitzschia pseudodelicatissima group	*Pseudo-nitzschia pseudodelicatissima*
	Pseudo-nitzschia seriata group	*Pseudo-nitzschia seriata*
	Pseudo-nitzschia spp.	*Pseudo-nitzschia*
	Rhizosolenia cf. acuminata	*Rhizosolenia*
	Rhizosolenia cf. bergonii	*Rhizosolenia*
	Rhizosolenia cf. curvata	*Rhizosolenia*
	Rhizosolenia cf. decipiens	*Rhizosolenia*
	Rhizosolenia cf. hebetata	*Rhizosolenia*
	Rhizosolenia cf. imbricata	*Rhizosolenia*
	Rhizosolenia spp.	*Rhizosolenia*
	Skeletonema spp.	*Skeletonema*
	Thalassionema spp.	*Thalassionema*
	Thalassiosira spp.	*Thalassiosira*
Dinophyceae	*Amphidinium spp.*	*Amphidinium*
	Archaeperidinium cf. minutum	*Archaeperidinium*
	Blepharocysta spp.	*Blepharocysta*
	Ceratocorys cf. gourreti	*Ceratocorys*
	Ceratocorys spp.	*Ceratocorys*
	Dinophysis cf. acuminata	*Dinophysis*
	Dinophysis cf. ovum	*Dinophysis*
	Dinophysis cf. uracantha	*Dinophysis*
	Dinophysis spp.	*Dinophysis*
	Diplopsalis group	*Diplopsalis*
	Gonyaulax cf. apiculata	*Gonyaulax*
	Gonyaulax cf. elegans	*Gonyaulax*
	Gonyaulax cf. fragilis	*Gonyaulax*

	Gonyaulax cf. hyalina	*Gonyaulax*
	Gonyaulax cf. pacifica	*Gonyaulax*
	Gonyaulax cf. polygramma	*Gonyaulax*
	Gonyaulax cf. scrippsae	*Gonyaulax*
	Gonyaulax cf. sphaeroidea	*Gonyaulax*
	Gonyaulax cf. spinifera	*Gonyaulax*
	Gonyaulax cf. striata	*Gonyaulax*
	Gonyaulax spp.	*Gonyaulax*
	Gymnodinium spp.	*Gymnodinium*
	Gyrodinium spp.	*Gyrodinium*
	Histioneis cf. megalocopa	*Histioneis*
	Histioneis cf. striata	*Histioneis*
	Oxytoxum cf. laticeps	*Oxytoxum*
	Oxytoxum spp.	*Oxytoxum*
	Paleophalacroma unicinctum	*Palaeophalacroma unicinctum*
	Phalacroma cf. rotundatum	*Phalacroma*
	Prorocentrum cf. balticum	*Prorocentrum*
	Prorocentrum cf. concavum	*Prorocentrum*
	Prorocentrum cf. nux	*Prorocentrum*
	Protoceratium spinolosum	*Protoceratium spinulosum*
	Protoperidinium cf. bipes	*Protoperidinium*
	Protoperidinium cf. breve	*Protoperidinium*
	Protoperidinium cf. crassipes	*Protoperidinium*
	Protoperidinium cf. diabolum	*Protoperidinium*
	Protoperidinium cf. divergens	*Protoperidinium*
	Protoperidinium cf. globulus	*Protoperidinium*
	Protoperidinium cf. grainii	*Protoperidinium*
	Protoperidinium cf. leonis	*Protoperidinium*
	Protoperidinium cf. monovelum	*Protoperidinium*
	Protoperidinium cf. nudum	*Protoperidinium*
	Protoperidinium cf. ovatum	*Protoperidinium*
	Protoperidinium cf. ovum	*Protoperidinium*
	Protoperidinium cf. pyriforme	*Protoperidinium*
	Protoperidinium cf. quarnerense	*Protoperidinium*
	Protoperidinium cf. steinii	*Protoperidinium*
	Protoperidinium cf. variegatum	*Protoperidinium*
	Protoperidinuim spp.	*Protoperidinium*
	Schuettiella cf. mitra	*Schuettiella*
	Tripos arietinum	*Tripos arietinus*
	Tripos lineatus/pentagonus complex	*Tripos lineatus*
	Tripos massiliense	*Tripos massiliensis*

Note. Data of genera (using the harmonized names) were excluded from the database.

8 Author contributions

MV, NG, NEZ and DR conceived the study. MG performed the taxonomic expert screening. MV compiled substantial parts of the MareDat database. DR compiled the data, developed the code, and led the writing, with inputs by all authors.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

We thank all biologists, taxonomists and cruise organizers for their essential efforts in collecting and sharing occurrence data of marine phytoplankton, synthesized in this database. We thank for expertise on taxonomic nomenclature provided by M. Estrada. We thank M. Döring and P. Provoost for their support with retrieval of phytoplankton occurrence data from GBIF (www.gbif.org) and OBIS (www.obis.org). Funding for this effort came from ETH Zürich under grant ETH-52 13-2.

References

Aman, A. A. and Bman, B. B.: The test article, J. Sci. Res., 12, 135–147, doi:10.1234/56789, 2015.

Aiken, J., Rees, N., Hooker, S., Holligan, P., Bale, A., Robins, D., Moore, G., Harris, R. and Pilgrim, D.: The Atlantic Meridional Transect: overview and synthesis of data, Prog. Oceanogr., 45(3–4), 257–312, doi:10.1016/S0079-6611(00)00005-7, 2000.

Amante, C. and Eakins, B. W.: ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis, (NOAA Tech. Memo. NESDIS NGDC-24, Natl. Geophys. Data Center, NOAA, 2009)., doi:10.7289/V5C8276M, 2009.

Balch, W. M., Bates, N. R., Lam, P. J., Twining, B. S., Rosengard, S. Z., Bowler, B. C., Drapeau, D. T., Garley, R., Lubelczyk, L. C., Mitchell, C. and Rauschenberg, S.: Factors regulating the Great Calcite Belt in the Southern Ocean and its biogeochemical significance, Global Biogeochem. Cycles, 30(8), 1124–1144, doi:10.1002/2016GB005414, 2016.

Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E. and Wincker, P.: Tara Oceans studies plankton at planetary scale, Science, 348(6237), 873–873, doi:10.1126/science.aac5605, 2015.

de Boyer Montégut, C.: Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology, J. Geophys. Res., 109(C12), C12003, doi:10.1029/2004JC002378, 2004.

Breiner, F. T., Guisan, A., Bergamini, A. and Nobis, M. P.: Overcoming limitations of modelling rare species by using ensembles of small models, Methods Ecol. Evol., 6(10), 1210–1218, doi:10.1111/2041-210X.12403, 2015.

Brun, P., Vogt, M., Payne, M. R., Gruber, N., O'Brien, C. J., Buitenhuis, E. T., Le Quéré, C., Leblanc, K. and Luo, Y.-W.: Ecological niches of open ocean phytoplankton taxa, Limnol. Oceanogr., 60(3), 1020–1038, doi:10.1002/lno.10074, 2015.

Buitenhuis, E. T., Li, W. K. W., Vaulot, D., Lomas, M. W., Landry, M. R., Partensky, F., Karl, D. M., Ulloa, O., Campbell, L., Jacquet, S., Lantoine, F., Chavez, F., Macias, D., Gosselin, M. and McManus, G. B.: Picophytoplankton biomass distribution in the global ocean, Earth Syst. Sci. Data, 4(1), 37–46, doi:10.5194/essd-4-37-2012, 2012.

Buitenhuis, E. T., Vogt, M., Moriarty, R., Bednaršek, N., Doney, S. C., Leblanc, K., Le Quéré, C., Luo, Y.-W., O'Brien, C., O'Brien, T., Peloquin, J., Schiebel, R. and Swan, C.: MAREDAT: towards a world atlas of MARine Ecosystem DATa, Earth Syst. Sci. Data, 5(2), 227–239, doi:10.5194/essd-5-227-2013, 2013.

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

Chaudhary, C., Saeedi, H. and Costello, M. J.: Bimodality of Latitudinal Gradients in Marine Species Richness, Trends Ecol. Evol., 31(9), 670–676, doi:10.1016/j.tree.2016.06.001, 2016.

Chaudhary, C., Saeedi, H. and Costello, M. J.: Marine Species Richness Is Bimodal with Latitude: A Reply to Fernandez and Marques, Trends Ecol. Evol., 32(4), 234–237, doi:10.1016/j.tree.2017.02.007, 2017.

Colwell, R. K. and Rangel, T. F.: Hutchinson's duality: The once and future niche, Proc. Natl. Acad. Sci., 106(Supplement_2), 19651–19658, doi:10.1073/pnas.0901650106, 2009.

Conway, J., Eddelbuettel, D., Nishiyama, T., Prayaga, S. K., Tiffin, N.: RPostgreSQL: R interface to the PostgreSQL database system. R package version 0.4.1. https://cran.r-project.org/package=RPostgreSQL, 2015.

Chamberlain, S.: rgbif: Interface to the Global Biodiversity Information Facility API. R package version 0.9.7. https://cran.r-project.org/package=rgbif, 2015.

Duarte, C. M.: Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition, Limnol. Oceanogr. Bull., 24(1), 11–14, doi:10.1002/lob.10008, 2015.

Edwards, J. L.: Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop, Science, 289(5488), 2312–2314, doi:10.1126/science.289.5488.2312, 2000.

Endo, H., Ogata, H. and Suzuki, K.: Contrasting biogeography and diversity patterns between diatoms and haptophytes in the central Pacific Ocean, Sci. Rep., 8(1), 10916, doi:10.1038/s41598-018-29039-9, 2018.

Falkowski, P. G., Katz M. E., Knoll, A. H., Quigg, A., Raven, J. A., Schofield, O. and Taylor, F. J. R.: The evolution of modern eukaryotic phytoplankton, Science, 305(5682), 354–360, doi:10.1126/science.1095964, 2004.

Field, C. B., Behrenfeld, M. J., Tanderson, J. T. and Falkowski, P.: Primary production of the biosphere: Integrating

terrestrial and oceanic components, Science, 281(5374), 237–240, doi:10.1126/science.281.5374.237, 1998.

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincon, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K. W., Lomas, M. W., Veneziano, D., Vera, C. S., Vrugt, J. A. and Martiny, A. C.: Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus, Proc. Natl. Acad. Sci., 110(24), 9824–9829, doi:10.1073/pnas.1307701110, 2013.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., Reagan, J. R. and Johnson, D. R.: World Ocean Atlas 2013, Vol. 4 Dissolved Inorg. Nutr. (phosphate, nitrate, Silic. S. Levitus, Ed.; A. Mishonov, Tech. Ed., 25, 2013.

Guisan, A. and Thuiller, W.: Predicting species distribution: Offering more than simple habitat models, Ecol. Lett., 8(9), 993–1009, doi:10.1111/j.1461-0248.2005.00792.x, 2005.

Guisan, A. and Zimmermann, N. E.: Predictive habitat distribution models in ecology, Ecol. Modell., 135(2–3), 147–186, doi:10.1016/S0304-3800(00)00354-9, 2000.

Honjo, S. and Okada, H.: Community structure of soccolithophores in the photic layer of the mid-pacific, Micropaleontology, 20(2), 209, doi:10.2307/1485061, 1974.

Iglesias-Rodríguez, M. D., Brown, C. W., Doney, S. C., Kleypas, J., Kolber, D., Kolber, Z., Hayes, P. K. and Falkowski, P. G.: Representing key phytoplankton functional groups in ocean carbon cycle models: Coccolithophorids, Global Biogeochem. Cycles, 16(4), 47-1-47–20, doi:10.1029/2001GB001454, 2002.

Jeong, H. J., Yoo, Y. Du, Kim, J. S., Seong, K. A., Kang, N. S. and Kim, T. H.: Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs, Ocean Sci. J., 45(2), 65–91, doi:10.1007/s12601-010-0007-2, 2010.

Jones, M. C. and Cheung, W. W. L.: Multi-model ensemble projections of climate change effects on global marine biodiversity, ICES J. Mar. Sci., 72(3), 741–752, doi:10.1093/icesjms/fsu172, 2015.

Jordan, R. W.: A revised classification scheme for living haptophytes, Micropaleontology, 50(Suppl_1), 55–79, doi:10.2113/50.Suppl_1.55, 2004.

Leblanc, K., Arístegui, J., Armand, L., Assmy, P., Beker, B., Bode, A., Breton, E., Cornet, V., Gibson, J., Gosselin, M.-P., Kopczynska, E., Marshall, H., Peloquin, J., Piontkovski, S., Poulton, A. J., Quéguiner, B., Schiebel, R., Shipe, R., Stefels, J., van Leeuwe, M. A., Varela, M., Widdicombe, C. and Yallop, M.: A global diatom database – abundance, biovolume and biomass in the world ocean, Earth Syst. Sci. Data, 4(1), 149–165, doi:10.5194/essd-4-149-2012, 2012.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., Garcia, H. E., Baranova, O. K., Zweng, M. M., Paver, C. R., Reagan, J. R., Johnson, D. R., Hamilton, M. and Seidov, D.: World Ocean Atlas 2013, Vol. 2 Temp. S. Levitus, Ed., A.

Lund, J. W. G., Kipling, C. and Le Cren, E. D.: The inverted microscope method of estimating algal numbers and the statistical basis of estimations by counting, Hydrobiologia, 11(2), 143–170, doi:10.1007/BF00007865, 1958.

Luo, Y.-W., Doney, S. C., Anderson, L. A., Benavides, M., Berman-Frank, I., Bode, A., Bonnet, S., Boström, K. H., Böttjer, D., Capone, D. G., Carpenter, E. J., Chen, Y. L., Church, M. J., Dore, J. E., Falcón, L. I., Fernández, A., Foster, R. A., Furuya, K., Gómez, F., Gundersen, K., Hynes, A. M., Karl, D. M., Kitajima, S., Langlois, R. J., LaRoche, J., Letelier, R. M., Marañón, E., McGillicuddy, D. J., Moisander, P. H., Moore, C. M., Mouriño-Carballido, B., Mulholland, M. R., Needoba, J. A., Orcutt, K. M., Poulton, A. J., Rahav, E., Raimbault, P., Rees, A. P., Riemann, L., Shiozaki, T., Subramaniam, A., Tyrrell, T., Turk-Kubo, K. A., Varela, M., Villareal, T. A., Webb, E. A., White, A. E., Wu, J. and Zehr, J. P.: Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates, Earth Syst. Sci. Data, 4(1), 47–73, doi:10.5194/essd-4-47-2012, 2012.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L., Zingone, A. and Bowler, C.: Insights into global diatom distribution and diversity in the world's ocean, Proc. Natl. Acad. Sci., 113(11), E1516–E1525, doi:10.1073/pnas.1509523113, 2016.

Mawji, E., Schlitzer, R., Dodas, E. M., Abadie, C., Abouchami, W., Anderson, R. F., Baars, O., Bakker, K., Baskaran, M., Bates, N. R., Bluhm, K., Bowie, A., Bown, J., Boye, M., Boyle, E. A., Branellec, P., Bruland, K. W., Brzezinski, M. A., Bucciarelli, E., Buesseler, K., Butler, E., Cai, P., Cardinal, D., Casciotti, K., Chaves, J., Cheng, H., Chever, F., Church, T. M., Colman, A. S., Conway, T. M., Croot, P. L., Cutter, G. A., de Baar, H. J. W., de Souza, G. F., Dehairs, F., Deng, F., Dieu, H. T., Dulaquais, G., Echegoyen-Sanz, Y., Lawrence Edwards, R., Fahrbach, E., Fitzsimmons, J., Fleisher, M., Frank, M., Friedrich, J., Fripiat, F., Galer, S. J. G., Gamo, T., Solsona, E. G., Gerringa, L. J. A., Godoy, J. M., Gonzalez, S., Grossteffan, E., Hatta, M., Hayes, C. T., Heller, M. I., Henderson, G., Huang, K., Jeandel, C., Jenkins, W. J., John, S., Kenna, T. C., Klunder, M., Kretschmer, S., Kumamoto, Y., Laan, P., Labatut, M., Lacan, F., Lam, P. J., Lannuzel, D., le Moigne, F., Lechtenfeld, O. J., Lohan, M. C., Lu, Y., Masqué, P., McClain, C. R., Measures, C., Middag, R., Moffett, J., Navidad, A., Nishioka, J., Noble, A., Obata, H., Ohnemus, D. C., Owens, S., Planchon, F., Pradoux, C., Puigcorbé, V., Quay, P., Radic, A., Rehkämper, M., Remenyi, T., Rijkenberg, M. J. A., Rintoul, S., Robinson, L. F., Roeske, T., Rosenberg, M., van der Loeff, M. R., Ryabenko, E., et al.: The GEOTRACES intermediate data product 2014, Mar. Chem., 177, 1–8, doi:10.1016/j.marchem.2015.04.005, 2015.

McQuatters-Gollop, A., Edwards, M., Helaouët, P., Johns, D. G., Owens, N. J. P., Raitsos, D. E., Schroeder, D., Skinner, J. and Stern, R. F.: The Continuous Plankton Recorder survey: How can long-term phytoplankton datasets contribute to the assessment of Good Environmental Status?, Estuar. Coast. Shelf Sci., 162, 88–97, doi:10.1016/j.ecss.2015.05.010, 2015.

Menegotto, A. and Rangel, T. F.: Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness, Nat. Commun., 9(1), 4713, doi:10.1038/s41467-018-07217-7, 2018.

Meyer, C., Kreft, H., Guralnick, R. and Jetz, W.: Global priorities for an effective information basis of biodiversity distributions, Nat. Commun., 6(1), 8221, doi:10.1038/ncomms9221, 2015.

O'Brien, C. J., Peloquin, J. A., Vogt, M., Heinle, M., Gruber, N., Ajani, P., Andruleit, H., Arístegui, J., Beaufort, L., Estrada, M., Karentz, D., Kopczyńska, E., Lee, R., Poulton, a. J., Pritchard, T. and Widdicombe, C.: Global marine plankton functional type biomass distributions: coccolithophores, Earth Syst. Sci. Data, 5(2), 259–276, doi:10.5194/essd-5-259-2013, 2013.

O'Brien, C. J., Vogt, M. and Gruber, N.: Global coccolithophore diversity: Drivers and future change, Prog. Oceanogr., 140, 27–42, doi:10.1016/j.pocean.2015.10.003, 2016.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. and Ferrier, S.: Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data, Ecol. Appl., 19(1), 181–197, doi:10.1890/07-2153.1, 2009.

Provoost, P. and Bosch, S.: robis: R client for the OBIS API. R package version 0.1.5. https://cran.r-project.org/package=robis, 2015.

Le Quéré, C.: Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models, Glob. Change Biol., 11(11), 2016–2040, doi:10.1111/j.1365-2486.2005.01004.x, 2005.

Richardson, A. J., Walne, A. W., John, A. W. G., Jonas, T. D., Lindley, J. A., Sims, D. W., Stevens, D. and Witt, M.: Using continuous plankton recorder data, Prog. Oceanogr., 68(1), 27–74, doi:10.1016/j.pocean.2005.09.011, 2006.

Righetti, D., Vogt, M., Zimmermann, N. E. and Gruber, N.: PHYTOBASE: A global synthesis of open ocean phytoplankton occurrences, doi:10.1594/PANGAEA.904397, 2019a.

Righetti, D., Vogt, M., Gruber, N., Psomas, A. and Zimmermann, N. E.: Global pattern of phytoplankton diversity driven by temperature and environmental variability, Sci. Adv., 5(5), 10, doi:10.1126/sciadv.aau6253, 2019b.

Rodríguez-Ramos, T., Marañón, E. and Cermeño, P.: Marine nano- and microphytoplankton diversity: redrawing global patterns from sampling-standardized data, Glob. Ecol. Biogeogr., 24(5), 527–538, doi:10.1111/geb.12274, 2015.

Rombouts, I., Beaugrand, G., Ibañez, F., Gasparini, S., Chiba, S. and Legendre, L.: A multivariate approach to large-scale variation in marine planktonic copepod diversity and its environmental correlates, Limnol. Oceanogr., 55(5), 2219–2229, doi:10.4319/lo.2010.55.5.2219, 2010.

Sal, S., López-Urrutia, Á., Irigoien, X., Harbour, D. S. and Harris, R. P.: Marine microplankton diversity database, Ecology, 94(7), 1658, doi:10.1890/13-0236.1, 2013.

Ser-Giacomi, E., Zinger, L., Malviya, S., De Vargas, C., Karsenti, E., Bowler, C. and De Monte, S.: Ubiquitous abundance distribution of non-dominant plankton across the global ocean, Nat. Ecol. Evol., 2(8), 1243–1249, doi:10.1038/s41559-018-

0587-2, 2018.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M. and Herndl, G. J.: Microbial diversity in the deep sea and the underexplored "rare biosphere," Proc. Natl. Acad. Sci., 103(32), 12115–12120, doi:10.1073/pnas.0605127103, 2006.

Sournia, A., Chrdtiennot-Dinet, M.-J. and Ricard, M.: Marine phytoplankton: how many species in the world ocean?, J. Plankton Res., 13(5), 1093–1099, doi:10.1093/plankt/13.5.1093, 1991.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., D'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., Bork, P., Boss, E., Bowler, C., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E. G., Sardet, C., Sieracki, M. and Velayoudon, D.: Structure and function of the global ocean microbiome, Science, 348(6237), 1261359–1261359, doi:10.1126/science.1261359, 2015.

Thompson, G. G. and Withers, P. C.: Effect of species richness and relative abundance on the shape of the species accumulation curve, Austral Ecol., 28(4), 355–360, doi:10.1046/j.1442-9993.2003.01294.x, 2003.

Tittensor, D. P., Mora, C., Jetz, W., Lotze, H. K., Ricard, D., Berghe, E. Vanden and Worm, B.: Global patterns and predictors of marine biodiversity across taxa, Nature, 466(7310), 1098–1101, doi:10.1038/nature09329, 2010.

Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T. W., McNeill, J., Monro, A. M., Prado, J., Price, M. J. & Smith, G. F., editors. International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Regnum Vegetabile, Vol. 159. pp. [i]-xxxviii, 1-253. Glashütten: Koeltz Botanical Books, 2018. doi:10.12705/Code.2018.

Utermöhl, H.: Zur Vervollkommnung der quantitativen Phytoplankton-Methodik, SIL Commun. 1953-1996, 9(1), 1–38, doi:10.1080/05384680.1958.11904091, 1958.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Luke, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., Boss, E., Follows, M., Karp-Boss, L.,

Krzic, U., Reynaud, E. G., Sardet, C., Sullivan, M. B. and Velayoudon, D.: Eukaryotic plankton diversity in the sunlit ocean, Science, 348(6237), 1261605–1261605, doi:10.1126/science.1261605, 2015.

Villar, E., Farrant, G. K., Follows, M., Garczarek, L., Speich, S., Audic, S., Bittner, L., Blanke, B., Brum, J. R., Brunet, C., Casotti, R., Chase, A., Dolan, J. R., D'Ortenzio, F., Gattuso, J.-P., Grima, N., Guidi, L., Hill, C. N., Jahn, O., Jamet, J.-L., Le Goff, H., Lepoivre, C., Malviya, S., Pelletier, E., Romagnan, J.-B., Roux, S., Santini, S., Scalco, E., Schwenck, S. M., Tanaka, A., Testor, P., Vannier, T., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S. G., Bork, P., Boss, E., de Vargas, C., Gorsky, G., Ogata, H., Pesant, S., Sullivan, M. B., Sunagawa, S., Wincker, P., Karsenti, E., Bowler, C., Not, F., Hingamp, P. and Iudicone, D.: Environmental characteristics of Agulhas rings affect interocean plankton transport, Science, 348(6237), 1261447–1261447, doi:10.1126/science.1261447, 2015.

Vogt, M., O'Brien, C., Peloquin, J., Schoemann, V., Breton, E., Estrada, M., Gibson, J., Karentz, D., Van Leeuwe, M. A., Stefels, J., Widdicombe, C. and Peperzak, L.: Global marine plankton functional type biomass distributions: Phaeocystis spp., Earth Syst. Sci. Data, 4(1), 107–120, doi:10.5194/essd-4-107-2012, 2012.

Wallace, D. W. R.: Chapter 6.3 Storage and transport of excess CO2 in the oceans:The JGOFS/WOCE global CO2 survey, in Eos, Transactions American Geophysical Union, vol. 82, pp. 489–521., 2001.

Wickham, H. and Chang, W.: Devtools: Tools to make developing R packages easier. R package version 1.12.0. https://cran.r-project.org/package=devtools, 2015.

Woolley, S. N. C., Tittensor, D. P., Dunstan, P. K., Guillera-Arroita, G., Lahoz-Monfort, J. J., Wintle, B. A., Worm, B. and O'Hara, T. D.: Deep-sea diversity patterns are shaped by energy availability, Nature, 533(7603), 393–396, doi:10.1038/nature17937, 2016.

Worm, B., Sandow, M., Oschlies, A., Lotze, H. K. and Myers, R. a: Global patterns of predator diversity in the open oceans., Science, 309(5739), 1365–9, doi:10.1126/science.1113399, 2005.

Zimmermann, N. E. and Guisan, A.: Predictive habitat distribution models in ecology, Ecol. Modell., 135(2–3), 147–186, doi:10.1016/S0304-3800(00)00354-9, 2000.

Zweng, M. M., Reagan, J. R., Antonov, J. I., Locarini, R. A., Mishonov, A. V., Boyer, T. P., Garcia, H. E., Baranova, O. K., Johnson, D. R., Seidov, D. and Biddle, M. M.: World Ocean Atlas 2013, Volume 2: Salinity., S. Levitus, A. Mishonov, Eds. (NOAA Atlas NESDIS 74, 2013), 39 pp., 2013.

PHYTOBASE: A global synthesis of open ocean phytoplankton occurrences

Damiano Righetti[1], Meike Vogt[1], Niklaus E. Zimmermann[2], Michael D. Guiry[3], Nicolas Gruber[1]

[1]Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, Universitätstrasse 16, 8092 Zürich, Switzerland
[2]Dynamic Macroecology, Landscape Dynamics, Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland
[3]AlgaeBase, Ryan Institute, NUI, Galway, University Road, Galway H91 TK33, Ireland

Correspondence to: Damiano Righetti (damiano.righetti@env.ethz.ch)

Abstract. Marine phytoplankton are responsible for half of the global net primary production and perform multiple other ecological functions and services of the global ocean. These photosynthetic organisms comprise more than 4300 marine species, but their biogeographic patterns and the resulting species diversity are poorly known, mostly owing to severe data limitations. Here, we compile, synthesize, and harmonize marine phytoplankton occurrence records from the two largest biological occurrence archives (Ocean Biogeographic Information System; OBIS, and Global Biodiversity Information Facility; GBIF) and three independent recent data collections. We bring together over 1.36 million phytoplankton occurrence records (1.28 million at the level of species) for a total of 1704 species, spanning the principal groups of the diatoms, dinoflagellates, and haptophytes, as well as several other groups. This data compilation increases the amount of marine phytoplankton records available through the single largest contributing archive (OBIS) by 65%. Data span all ocean basins, latitudes and most seasons. Analyzing the oceanic inventory of sampled phytoplankton species richness at the broadest spatial scales possible, using a resampling procedure, we find that richness tends to saturate in the pantropics at ~93% of all species in our database, at ~64% in temperate waters, and at ~35% in the cold Northern Hemisphere, while the Southern Hemisphere remains underexplored. We provide metadata on the cruise, research institution, depth and date for each data record, and we include phytoplankton cell counts for 193 763 records. We strongly recommend consideration of spatiotemporal biases in sampling intensity and varying taxonomic sampling scopes between research cruises or institutions when analyzing the occurrence data spatially. Including such information into predictive tools, such as statistical species distribution models may serve to project the diversity, niches, and distribution of species in the contemporary and future ocean, opening the door for quantitative macroecological analyses of phytoplankton. PhytoBase can be downloaded from PANGAEA, doi:10.1594/PANGAEA.904397 (Righetti et al., 2019a).

1 Introduction

Phytoplankton are photosynthetic members of the plankton, responsible for about half of the global net primary production (Field et al., 1998). While more than 4300 phytoplankton species have been described (Sournia et al., 1991), spanning at least six major clades (Falkowski et al., 2004), there are likely many more species living in the ocean, perhaps more than

10 000 (de Vargas et al., 2015). Some of these species (e.g. *Emiliania huxleyi*, *Gephyrocapsa oceanica*) are abundant and occur throughout the ocean (Iglesias-Rodríguez et al., 2002), but a majority of plankton species form low abundance populations (Ser-Giacomi et al., 2018) and remain essentially uncharted; i.e., the quantitative description of where they live, and where not, is rather poor. This biogeographic knowledge gap stems from a lack of systematic global surveys, as have been undertaken for inorganic carbon (WOCE/JGOFS/GOSHIP; Wallace 2001) or for trace metals (GEOTRACES; Mawji et al. 2015). Owing to logistic and financial challenges associated with internationally coordinated surveys, our knowledge of phytoplankton biogeography is, with a few exceptions (Bork et al., 2015; McQuatters-Gollop et al., 2015), mostly based on spatially very limited surveys or basin scale studies (e.g., Endo et al., 2018; Honjo and Okada, 1974). Marine phytoplankton occurrence data are unevenly distributed, incomplete in remote areas, and orders of magnitude higher in more easily accessed areas, especially near coasts (Buitenhuis et al., 2013). Additional factors that have impeded progress in developing a good biogeographic understanding of the phytoplankton are difficulties in species identification, linked to their microscopic body size. This is well reflected in the current geographic knowledge on phytoplankton species richness from direct observations (e.g. Rodríguez-Ramos et al., 2015), which is much more limited compared to that of other marine taxa, such as zooplankton (e.g., Rombouts et al., 2010), fishes (e.g, Jones and Cheung, 2015), sharks (e.g., Worm et al., 2005) or krill (e.g., Tittensor et al., 2010), even though many of these taxa also suffer from deficiencies in sampling efforts (Menegotto and Rangel, 2018).

Initial efforts to overcome the data sparseness and patchiness for phytoplankton by the MareDat project (Buitenhuis et al., 2012; Leblanc et al., 2012; Luo et al., 2012; O'Brien et al., 2013; Vogt et al., 2012) resulted in the compilation and synthesis of 119 phytoplankton species from 17 240 sampling events. While representing a large step forward, the coverage remained relatively limited, largely owing to MareDat's focus on abundance data, motivated by the need to use the data for model evaluation and other quantitative assessments (Buitenhuis et al., 2013). But during these efforts, it became clear that there are at least an order of magnitude more data in archives around the world if one relaxed the abundance criterion and considered all observations that included presences. The potential for the use of presences to constrain e.g., phytoplankton community structure and richness, is large, as demonstrated by Righetti et al. (2019b), who recently produced the first global map of phytoplankton species richness. This application was also made possible thanks to the rapid developments in data mining and statistical analysis tools, such as species distribution models (SDMs) (Guisan and Zimmermann, 2000) that permit scientists to account for some of the limitations stemming from spatiotemporal sampling biases underlying species' occurrence data (Breiner et al., 2015; Phillips et al., 2009).

A key enabler for the compilation and synthesis of phytoplankton occurrences (presence or abundance records) is the existence of two digital biological data archives, i.e., the Global Biodiversity Information Facility (GBIF; www.gbif.org), and the Ocean Biogeographic Information System (OBIS; www.obis.org). GBIF is the world's largest archive for species occurrence records, while OBIS is the largest occurrence database on marine taxa. Both archives have gathered a large number of phytoplankton occurrence records and make them freely available to the global community. In addition to MareDat (Buitenhuis et al., 2013), marine surveys such as those conducted with the Continuous Plankton Recorder (CPR)

(McQuatters-Gollop et al., 2015), the Atlantic Meridional Transect (AMT) (Aiken et al., 2000; Sal et al., 2013) and other programs provide relevant phytoplankton occurrence records, including data on species' abundance. A global synthesis of species occurrence records, including those from GBIF and OBIS has been attempted for upper trophic marine organisms, gathering 3.44 million records across nine taxa from zooplankton to sharks (Menegotto & Rangel 2018). But so far, no effort has been undertaken to bring the various sources together for the lowest trophic marine organisms and merge them into a single harmonized database. This study aims to address this gap and to create PhytoBase, the world's largest open ocean phytoplankton occurrence database, which may substantially reduce the global limitations associated with undersampling.

The majority of the existing occurrence data of phytoplankton species have been collected via seawater samples of ~5–25 mL (Lund et al., 1958; Utermöhl, 1958), followed by microscopic specimen identification. Another key source of occurrence data is the continuous plankton recorder (CPR) program, in which plankton are sampled by filtering seawater onto a silk roll (270 μm mesh size) within a recorder device that is towed behind research and commercial ships (Richardson et al., 2006). The plankton are then picked from the screens and identified by microscopy. DNA sequencing has become an alternative method to record and monitor marine phytoplankton at large scales (e.g. de Vargas *et al.* 2015; Sunagawa *et al.* 2015). However, within the recent global TARA Oceans cruise, ca. $^1/_3$ of DNA sequences of plankton from seawater samples could not yet be assigned to any taxon (de Vargas et al., 2015). For the most species-rich phytoplankton group (*Bacillariophyceae*), 58% of DNA sequences from seawater could be assigned to genus level in the same cruise (Malviya et al., 2016), but the majority of species have lacked reference DNA sequences needed for their identification. Additional factors have hampered the study of global phytoplankton biogeography: Some surveys lack resolution in terms of the species recorded (Richardson et al., 2006; Villar et al., 2015) and abundance information in terms of cells or biomass of species is often not available in the archived records (e.g. from GBIF). Second, the taxonomic identification and chronic undersampling of the species present in local communities via seawater samples (Cermeño et al., 2014) pose challenges, which can be resolved only by trained experts or larger sampling volumes. In addition, the rapidly evolving taxonomy (e.g. Jordan 2004) has led to varying use of nomenclature. These limitations need to be assessed and possibly overcome in a data synthesis effort.

Here, we compile 1 360 621 phytoplankton occurrence records (94.1% resolved to the level of species; n = 1704 species) and demonstrate that combining data from OBIS and GBIF increases the number of occurrence records by 52.7% relative to the data solely obtained from OBIS. This gain increases to 65.2% when adding occurrence data from marine surveys, including MareDat (Buitenhuis et al., 2013), AMT cruises (Sal et al., 2013), and initial TARA Oceans results (Villar et al., 2015). With respect to species abundance information, we retain cell count records whenever available from all sources, resulting in 193 763 quantitative entries. We harmonize and update the taxonomy between the sources, focusing on extant species and open ocean records. The resulting PhytoBase dataset allows for studying global patterns in the biogeography, diversity, and composition of phytoplankton species. Using statistical SDMs, the data may serve as a starting point to examine species' niche differences across all major phytoplankton taxa and their potentially shifting distributions under climate change. The dataset can be accessed through PANGAEA, doi:10.1594/PANGAEA.904397 (Righetti et al., 2019a).

2 Compilation of occurrences

2.1 Data origin

To create PhytoBase, we compiled marine phytoplankton occurrences (i.e., presences and abundances larger than zero) from five sources, including the two largest open access species occurrence archives: the Global Biodiversity Information Facility (GBIF; www.gbif.org), and the Ocean Biogeographic Information System (OBIS; www.obis.org). These two archives represent leading efforts to gather global species distribution evidence. We augmented the data with records from the Marine Ecosystem Data initiative (MareDat; Buitenhuis *et al.* 2013), records from a micro-phytoplankton dataset (Sal et al., 2013), and records from the global TARA Oceans cruise (Villar et al., 2015), which were not included in GBIF or OBIS at the time of data query (closing window, March 2017). While our selection of additional data was not exhaustive, it strived for the inclusion of quality controlled large-scale phytoplankton datasets. Specifically, MareDat represents a previous global effort in gathering marine plankton data for ecological analyses (e.g., Brun et al., 2015; O'Brien et al., 2016), while Sal et al. (2013) and Villar et al. (2015) are unique in aspects of taxonomic standardization and consistency in methodology.

We retrieved occurrence records at the level "species" or below (e.g., "subspecies", "variety" and "form", as indicated by the taxonRank field in GBIF and OBIS downloads) for seven phyla: *Cyanobacteria*, *Chlorophyta* (excluding macroalgae), *Cryptophyta*, *Myzozoa*, *Haptophyta*, *Ochrophyta*, and *Euglenozoa.* More specifically, within the *Ochrophyta*, we considered the classes *Bacillariophyceae* (diatoms), *Chrysophyceae*, *Dictyochophyceae*, *Pelagophyceae* and *Raphidophyceae.* Within the *Myzozoa*, we considered the class *Dinophyceae* (dinoflagellates). Within the *Euglenozoa*, we considered the class *Euglenoidea*. This selection of phyla or classes strived to include all autotrophic marine phytoplankton taxa (de Vargas et al., 2015; Falkowski et al., 2004), but it is clear that some of the species may be mixotrophic, particularly for the *Dinophyceae* (Jeong et al., 2010). At genus level, we additionally retrieved occurrences for *Prochlorococcus* and *Synechococcus* from all sources, as the latter two genera are often highly abundant (Flombaum et al., 2013), but rarely determined to the species level. Last, we considered records for the functionally relevant genera *Phaeocystis*, *Richelia*, *Trichodesmium*, and non-specified picoeukaryotes from MareDat. For simplicity, we treat genera as "species" in statistics herein.

For the taxa selected, occurrence data from GBIF and OBIS were first downloaded in December 2015 and updated in February 2017. Specifically, the initial retrieval of the GBIF data occurred on 7 December 2015 (using the taxonomic backbone from https://doi.org/10.15468/39omei, accessed on 14 July 2015), and the data were updated on 27 February 2017 (using an updated taxonomic backbone, accessed via http://rs.gbif.org/datasets/backbone, released 27 February 2017). The data from OBIS were first retrieved on 5 December 2015 using the R package *robis* (Provoost and Bosch, 2015) and the OBIS taxonomic backbone, accessed on 4 December 2015 via the R packages *RPostgreSQL* (Conway et al., 2015) and *devtools* (Wickham and Chang, 2015). Data were updated for the taxa selected on 6 March 2017 (using the OBIS taxonomic backbone, accessed on 6 March 2017 via the same R packages). The update in 2017 expanded the occurrences retrieved from GBIF substantially, with over 20 000 additional phytoplankton records stemming from an Australian CPR program alone

(AusCPR, https://doi.org/10.1016/j.pocean.2005.09.011, accessed via www.gbif.org on 6 March 2017). We retained any GBIF sourced data that were retrieved in 2015, but deleted from GBIF before March 2017 (such as CPR data, with dataset key 83986ffa-f762-11e1-a439-00145eb45e9a). Occurrence data from the TARA Ocean cruise included the *Bacillariophyceae* and *Dinophyceae* (Villar et al., 2015; their Tables W8 and W9). Occurrence data from MareDat included five phytoplankton papers (Buitenhuis et al., 2012; Leblanc et al., 2012; Luo et al., 2012; O'Brien et al., 2013; Vogt et al., 2012). Additional data processed by the TARA Oceans or Malaspina expedition (Duarte, 2015) may provide valuable context for a future synthesis, and may eventually combine molecular with traditional approaches, yet here we have focused on publicly available sources until March 2017. These sources reflect decades to centuries of efforts spent in collecting phytoplankton data, including a substantial amount of data from the CPR program (Richardson et al., 2006) and a large fraction of data from the AMT program (cruises 1 to 6) (Sal et al., 2013).

2.2 Data selection

We excluded occurrences from waters less than 200 m deep (Amante and Eakins, 2009), from enclosed seas (Baltic Sea, Black Sea or Caspian Sea), and from seas with a surface salinity below 20, using the globally gridded (spatial 1° x 1°) monthly climatological data of Zweng et al. (2013). This salinity-bathymetry threshold served to select data from open oceans, excluding environmentally more complex, and often more fertile, near-shore waters.

2.2.1 Data accessed through GBIF and OBIS

We included GBIF occurrence records on the basis of "human observation", "observation", "literature", "living specimen", "material sample", "machine observation", "observation", and "unknown", assuming that the latter was based on observation. With respect to OBIS data, we included data records on the basis of "O" and "D", whereby "O" refers to observation and "D" to literature-based records. To filter out raw data of presumably inferior quality, records from OBIS and GBIF were removed: (i) if their year of collection indicated >2017 or <1800 (excluding 110 records; <0.001% of raw data), (ii) if they had no indication on the year or month of collection (excluding 7.2% GBIF raw data and 0.9% OBIS raw data) or (iii) if they had geographic coordinates outside the range -180 to 180 for longitude and/or outside -90 to 90 for latitude. However, the latter criterion was fulfilled by all records, as these were standardized to -180 to 180 degrees longitude (rather than 0 to 360 longitude East) and -90 to 90 degrees latitude (WGS84). Records with negative recording depths (0% of GBIF and 6.6% of OBIS raw data) were flagged and changed to positive, assuming that their original sign was mistaken.

2.2.2 Data accessed through MAREDAT

We included occurrence records at the species level for the *Bacillariophyceae* (Leblanc et al., 2012) and *Haptophyta* (O'Brien et al., 2013) and species presence records on *Bacillariophyceae* host cells from Luo et al. (2012). Harmonization of *Haptophyta* species names from MareDat (O'Brien et al., 2013) was guided by a synonymy table provided by O'Brien (*pers. comm.*) (Table A1). Harmonization of *Bacillariophyceae* species names in MareDat was in progress at the time of first data

access (24 August 2015) and completed (Table A2). In addition, we retained all genus and species level records available for *Trichodesmium*, *Richelia* (Luo et al., 2012), *Phaeocystis* (Vogt et al., 2012), *Synechococcus* (using the data-field "SynmL") and *Prochlorococcus* (using the data-field "PromL") (Buitenhuis et al., 2012). We included genus level records from the latter taxa, as they represent functionally important phytoplankton groups (Le Quéré, 2005), and as information on the presence and abundance of their cells or colonial cells often only existed at genus level (Buitenhuis et al., 2012; Luo et al., 2012; Vogt et al., 2012). Across all sources, data on colonial cells could be uniquely accessed via MareDat (and additional count data on trichomes of genus *Trichodesmium* are available from Luo et al., 2012). We also retained records of the "picoeukaryote" group, which were not determined to species or genus level (Buitenhuis et al., 2012). For all taxa, we retained records with reported abundances (i.e., cell counts) larger than zero, while excluding records with zero entries or missing data entries, as our database focuses on presence-only or abundance records. Given that data of the MareDat have been scrutinized previously, we flagged rather than excluded data with reported recording before year 1800 (n = 564; values 6, 10 or 11) and unrealistic day entries (n = 58 340; values -9 or -1).

2.2.3 Data accessed through Villar et al. (2015)

We compiled presence records of species of *Bacillariophyceae* and *Dinophyceae* from the tables W8 and W9 of Villar et al. (2015). We excluded species names containing "cf" (e.g *Bacteriastrum cf. delicatulum*), as such nomenclature is typically used to refer to closely related species of an observed species. We retained all species (n = 3), which contained "group" in their names (e.g. *Pseudo-nitzschia delicatissima group*). *Tripos lineatus/pentagonus complex* was considered as *Tripos lineatus*. The cleaning of spelling variants of original names from Villar et al. (2015) is presented in Table A3.

2.2.4 Data accessed through Sal et al. (2013)

We considered occurrence records of the *Bacillariophyceae*, *Dictyochophyceae, Dinophyceae*, *Haptophyta* and *Peridinea* and at species level or below, using the species name in the final database. These data included 5891 records from 314 species and 543 samples. The dataset of Sal et al. (2013) represents a highly complementary source of phytoplankton occurrence records, i.e., it had no duplicated records with any of the other data sources considered. This data collection consists of in situ samples subjected to consistent methodology performed by the same taxonomist.

2.3 Concatenation of source datasets

Column names or data-fields were adjusted and harmonized to establish compatibility in the dimensions of the different source datasets (Table 1). Columns match Darwin Core standard (https://dwc.tdwg.org) where original data structure could be reconciled with this standard, following GBIF and OBIS that widely rely on Darwin Core. Where critical metadata could not be assigned to Darwin Core, we use additional columns (e.g., columns ending in "gbif" present metadata from GBIF). With regard to sampling depth, GBIF raw data contained the field "depthAccuracy" (18.6% of data with entries) while OBIS raw data contained the fields "depthprecision" (21.64% of data with entries), "minimumDepthInMeters" (Darwin Core term;

Table 1: Harmonization of column names (data-fields) between data sources and final column name structure in PhytoBase

Original column names							Final column names
GBIF (2015)	GBIF (2017)	OBIS (2015)	OBIS (2017)	MareDat	Villar et al	Sal et al	(sources merged)
species	species	species	species	species	species	species	scientificName*,¶
decimalLongitude	longitude	longitude	longitude	Longitude	Longitude	Lon	decimalLongitude*
decimalLatitude	latitude	latitude	latitude	Latitude	Latitude	Lat	decimalLatitude*
year	year	yearcollected	year	Year	Date	Date	year*
month	month	monthcollected	month	Month	Date	Date	month*
day	day	daycollected	day	Day	Date	Date	day*
depth	depth	depth	depth	Depth	Depth	Depth	depth
-	depthAccuracy	depthprecision	depthprecision	-	-	-	depthAccuracy
taxonRank	taxonRank	-	-	rank	-	-	taxonRank*,†
-	occurrencestatus	-	occurrencestatus	-	-	-	occurrenceStatus*
phylum	phylum	phylum	phylum	-	-	-	phylum*,‡
class	class	class	class	-	-	-	class*,‡
basisOfRecord	basisOfRecord	basisofrecord	basisOfRecord	-	-	-	basisOfRecord*
-	institutionCode	institutioncode	institutionCode	-	-	-	institutionCode*,§
-	-	-	-	-	-	-	sourceArchive
datasetKey	datasetKey	-	-	-	-	-	datasetKey_gbif‖,§§
publishingOrgKey	-	-	-	-	-	-	publishingOrgKey_gbif§
-	-	collectioncode	collectionCode	-	-	-	collectionCode_obis‖
-	-	-	resname	-	-	-	resname_obis‖
-	-	resource_id	resource_id	-	-	-	resourceID_obis‖,§§
-	-	-	-	Origin Database	-	-	originDatabase_maredat§
-	-	-	-	CruiseorStationID	-	-	cruiseOrStationID_ maredat‖
-	-	-	-	-	Station	-	taraStation_villar‖
-	-	-	-	-	-	Cruise	cruise_sal‖
-	-	-	-	-		SampleID	sampleID_sal
-	-	-	-	-	Mixed Layer Depth (m)	MLD	MLD_villar_sal
-	-	-	-	cellsL^{-1},cellsmL^{-1}	-	organism-quantity	organismQuantity*<and> organismQuantityType*
-	individualCount	-	observedindivi-	-	-	-	individualCount*,+
-	-	-	dualcount]	-	-	-	yearOfDataAccess
-	-	-	-	-	-	-	flag

GBIF data were downloaded in 2015 (www.gbif.org; retrieved 7 December 2015) and 2017 (retrieved 27 February 2017)

OBIS data were downloaded in 2015 (www.iobis.org; retrieved 5 December 2015) and 2017 (retrieved 6 March 2017)

Each occurrence record in PhytoBase is uniquely identifiable by the occurrence ID: scientificName, decimalLongitude, decimalLatitude, year, month, day and depth

* Column names following Darwin Core standard (https://dwc.tdwg.org).

¶ We retain all original scientificName(s) and synonyms used in individual sources as additional columns with the format "scientificNameOriginal_<source>"

† The "TaxonRank" field indicates the level of taxonomic resolution (species or genus) of the observation record. Records of subspecies, varieties, and forms were generally extracted from original sources, but considered at the species level (using the genus and specific epithet).

‡ Higher order taxonomy (phylum, class) follows OBIS (taxonomic backbone; retrieved 6 March 2017), which relies on the World Register of Marine Species (www.marinespecies.org).

§ These fields indicate the organization or institution by which original records were collected.

‖ These fields are indicators of different research cruises or resources, to which original records belonged.

+ "individualCount" and "observedindividualcount" had equivalent entries for records that overlapped between GBIF and OBIS, and were merged into one column.

§§ datasetKey_gbif and resourceID_obis are keys to access metadata of original datasets in GBIF and OBIS via API, including information on sampling methods.

25.7% of data with entries) and “maximumDepthInMeters” (Darwin Core term; 24.0% of data with entries). To enhance compatibility between GBIF and OBIS, we therefore used the column "depth", together with "depthAccuracy", and we integrated "depthprecision" into the latter column. To indicate the source from which records were obtained (GBIF, OBIS, MareDat, Villar or Sal) and the year of data access, we added the columns “sourceArchive” and “yearOfDataAccess”. Last, we added a quality flag column, termed “flag”. This column denotes records with originally negative collection depth entries (N) changed to positive (sect. 2.2.1), unrealistic day (D) or year (Y) entries (sect. 2.2.2), and/or records collected from sediment samples or traps (S) rather than seawater samples (sect. 2.3.2). We concatenated the sources into a raw database, which contained 1.51 million depth-referenced occurrence records, 3300 phytoplankton species (including five genera) and 247 385 sampling events (Table 2). Sampling events are thereby (and herein) defined as unique combinations of decimalLongitude, decimalLatitude, depth, and time (year, month, day) in the data.

2.3.1 Extant species selection and taxonomic harmonization

We strived for a selection of occurrence data of extant phytoplankton species and a taxonomic harmonization of their multiple spelling variants (merging synonyms, while clearing misspellings or unaccepted names). This procedure included three steps:

(i) We discarded all species (and their data) that did not have any depth-referenced record. This choice was made on the basis that these species may have been predominantly recorded via fossil materials or have been associated with large uncertainty with respect to their sampling depth, which would infringe the scope of our database.

(ii) We extracted all scientific names (mostly at species level, including all synonyms and spelling variants) associated with at least one depth-referenced record from the raw database (Table 2). This resulted in 3300 names, which were validated in August 2017 against the 150 000+ specific and infraspecific names in AlgaeBase (www.algaebase.org), and matched using a relational database of current names and synonyms; orthography was made as compatible as possible

Table 2: Summary statistics of the raw database by source

Source	Number of observations (%unique to source)		Number of species* (%unique to source)		Number of observations (%unique to source)		Number of species* (%unique to source)	
	full data				data with depth-reference			
GBIF	970 927	(65.6)	3 977	(60.4)	908 995	(64.2)	2676	(51.5)
OBIS	853 981	(60.5)	2 305	(25.2)	823 968	(60.1)	1812	(25.4)
MareDat	102 621	(94.6)	123	(1.1)	102 467	(94.7)	123	(1.5)
Villar et al.	202	(100.0)	87	(0.0)	202	(100.0)	87	(0.0)
Sal et al.	5891	(100.0)	314	(0.0)	5867	(100.0)	313	(0.1)
Total	**1 594 649**		**4741**		**1 511 351**		**3300**	

Numbers of observations (with % of observations unique to the source in parentheses) and the numbers of species (with % of species unique to the source in parentheses) presented for each data source. 27 537 observation records of Picoeukaryotes (not identified to species or genus level) are included among the total records and stem from MareDat (all of which contained a depth-reference).

* Including synonyms or spelling variants.

with the International Code of Nomenclature (Turland et al., 2018), particularly in relation to the gender of specific epithets. This screening led to the exclusion of 459 names (and their data), which could not be traced back to any taxonomically accepted name at the time of query, and to the creation of a "synonymy table" in which each original name (including its potentially multiple synonyms and spelling errors) was matched to a corrected or accepted name.

(iii) We excluded species (and their data) classified as "fossil only" or "fossil" on AlgaeBase (www.algaebase.org, accessed August 2017) or the World Register of Marine Species (WoRMS; www.marinespecies.org, accessed August 2017). We further excluded species belonging to genera with fossil types denoted by AlgaeBase, under the condition that these species lacked habitat information on AlgaeBase, assuming that the latter species have been collected based on sedimentary or fossilized materials. Species uniquely classified as "freshwater" on AlgaeBase were discarded, as these were beyond the scope of our open ocean database. However, we retained species classified as "freshwater", which had at least 24 open ocean (sect 2.2) records and thus were assumed to thrive also in marine habitats: *Aulacoseira granulata, Chaetoceros wighamii, Diatoma rhombica, Dinobryon balticum, Gymnodinium wulffii, Tripos candelabrum, Tripos euarcuatus*. These cleaning steps led to a remaining set of 2032 original species names, synonyms or spelling variants, corresponding to 1709 taxonomically harmonized species (including five genera not resolved to species level).

2.3.2 Data merger and synthesis

We removed duplicate records, considering the columns "scientificName", "decimalLongitude", "decimalLatitude", "year", "month", "day", and "depth". Removing duplicates meant that any relevant metadata of the duplicated (and hence removed) records were added to the metadata of the record retained, either in an existing or additional column (e.g., information on the original dataset-keys, two which the merged records belonged). We assigned the corrected and/or harmonized taxonomic species name to each original species name in the database on the basis of the synonymy table. We removed duplicates with respect to exact combinations of the harmonized "scientificName", and "decimalLongitude", "decimalLatitude", "year", "month", "day", "depth". This resulted in the harmonized database containing 1 360 621 occurrence records (of which 95.8% had a depth-reference), 1709 species (including five genera), and 242 074 sampling events (Table 3). We retained meta-information on the dataset ID, cruise number, and further attributes when removing duplicates. In particular, we retained the original taxonomic name(s) associated with each record in separate columns of the type "scientificNameOriginal_<source>", which allows tracing back the harmonized name to its original name(s). Retaining original names ensures that future taxonomic changes or updated methods can be readily implemented. Besides the presences, the final database includes 193 777 count records of individuals or cells, spanning 1126 species. Among these, 105 242 records included a volume basis (spanning 335 species), with a predominant origin from MareDat (n = 99 498) and Sal et al. (2013) (n = 5744). Last, we flagged sedimentary records, indicated by the column "flag". Although we excluded probably many records based on fossil materials during cleaning step (i), this does not exclude the possibility that occurrence records of extant species in the GBIF and OBIS source datasets originated partially from sediment traps or sediment core samples.

Table 3: Summary statistics of the harmonized database by source

Source	Number of observations (%unique to source)		Number of species* (%unique to source)		Number of observations (%unique to source)		Number of species* (%unique to source)	
	full data				data with depth-reference			
GBIF	790 103	(54.9)	1492	(31.5)	751 227	(53.7)	1444	(31.3)
OBIS	823 836	(56.3)	1320	(21.6)	796 907	(56.0)	1283	(22.0)
MareDat	101 969	(94.7)	120	(2.7)	101 816	(94.8)	121	(2.7)
Villar et al.	202	(100.0)	87	(0.0)	202	(100.0)	87	(0.0)
Sal et al.	5744	(100.0)	291	(0.0)	5721	(100.0)	290	(0.0)
Total	**1 360 765**		**1709**		**1 303 721**		**1709**	

Numbers of observations (with % of observations unique to the source in parentheses) and numbers of species (with % of species unique to the source in parentheses) presented for each data source.

*Including 1711 species names and the genera *Phaeocystis*, *Trichodesmium*, *Richelia*, *Prochlorococcus* and *Synechococcus*. 27 537 observation records of Picoeukaryotes (not identified to species or genus level) are included among the total records and stem from MareDat (all of which contained a depth-reference).

Marine sediments can conserve phytoplankton cells that are exported to depth. We flagged phytoplankton records from OBIS and GBIF in the database associated with surface sediment traps or sediment cores (using an "S" in the flag column) by checking the metadata of each individual source dataset of GBIF (using the GBIF datasetKey) and OBIS (using the OBIS resourceID), using the function *datasets* in the R package *rgbif* (Chamberlain, 2015) and the online portal of OBIS (http://iobis.org/explore/#/dataset, accessed 24 October 2018). This check resulted in the flagging of 2.7% of records. We did not attempt to clean or remove sediment type records in MareDat, assuming that information on sampling depth, associated with records of MareDat led to thorough exclusion of sedimentary records previously. Data from Sal et al. (2013) and Villar et al. (2015) were uniquely based on seawater samples.

3 Results

3.1 Data

3.1.1 Spatiotemporal coverage

Phytoplankton occurrence records contained in PhytoBase cover all ocean basins, latitudes, longitudes, and months (Fig. 1). However, data density is globally highly uneven (Fig 1B, C; histograms) with 44.7% of all records falling into the North Atlantic alone, while only 1.4% of records originate from the South Atlantic, and large parts of the South Pacific basin are devoid of records (Fig. 1A). Analyzing the data by latitude (Fig. 1B) and longitude (Fig. 1C) reveals that sampling has been particularly thin at high latitudes (>70°N and S) during wintertime. Occurrences cover a total of 18 863 monthly cells of 1° latitude × 1° longitude (using the World Geodetic System of 1984 as the reference coordinate system; WGS 84), which corresponds to 3.9% of all monthly (n = 12 months) 1° cells of the open ocean (sect. 2.2). Without monthly distinction, records cover 6098 spatial 1° cells, which is a fraction of 15.5% of all 1° cells of the open ocean.

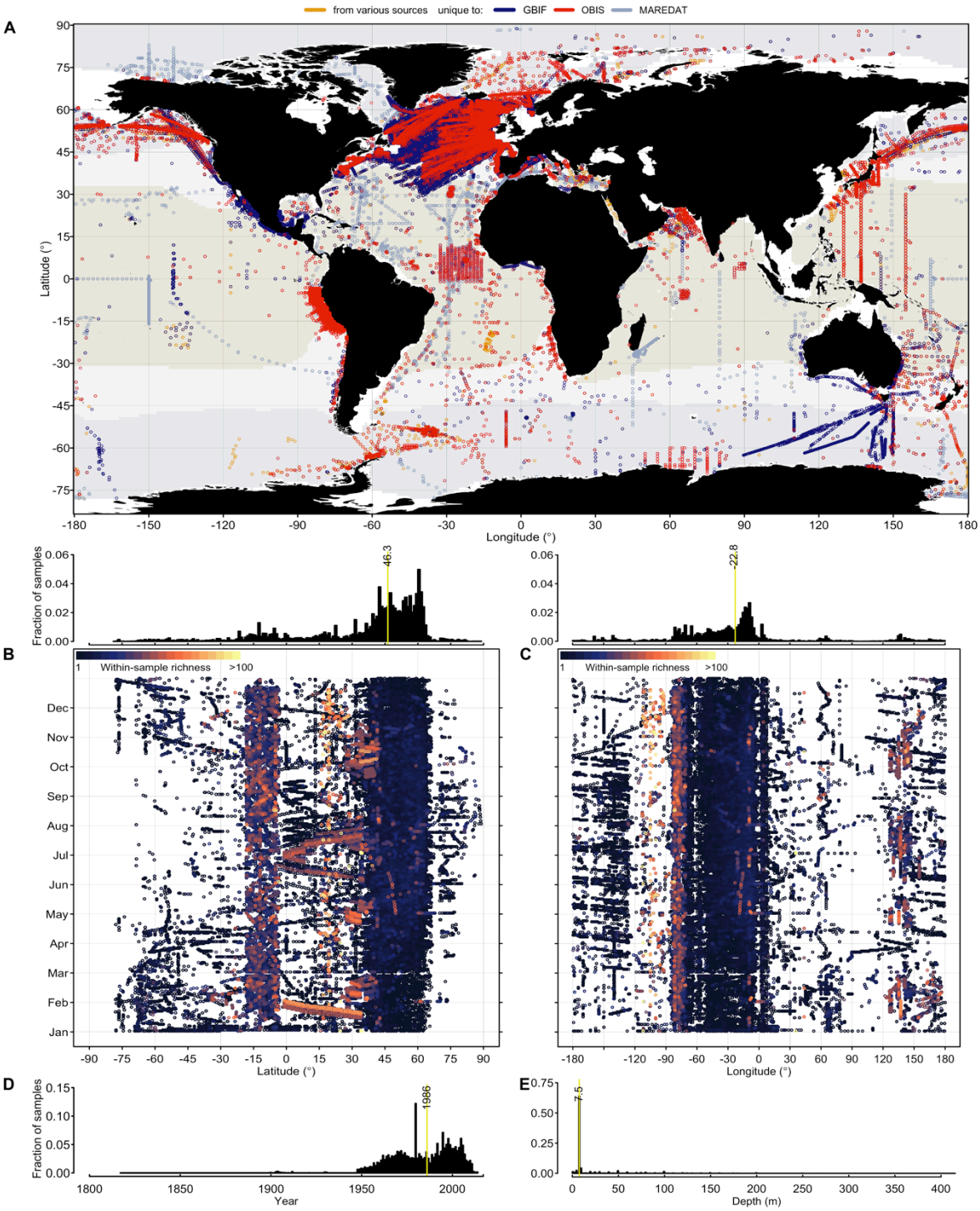

Figure 1: Global distribution of phytoplankton occurrence records of PhytoBase. (**A**) Circles show the position of in situ occurrence records (n = 1 360 765, including 1 280 103 records at the level of species), with the color indicating the source of the data. Color shading indicates the extent of tropical (T >20°C; yellow), temperate (10°C≤ T≤ 20°C; snow-white), and cold (T <10°C; light-blue) seas, based on

the annual mean sea surface temperature (Locarini et al., 2013). (**B-C**) Sampling locations (dots) are plotted as a function of the month of sampling, and (B) latitude, or (C) longitude. Colors display the species number detected in individual samples (each sample is defined as an exact combinations of time, location, and depth, in the dataset). Histograms above panels (B) and (C) show the frequency of samples by latitude (B) or longitude (C). (**D-E**) Histograms of sample frequency by year (D), and by depth (E). Vertical yellow lines show the median.

Record quantities are not evenly distributed between major taxa, and global sampling schemes differ between these taxa (Fig. 2). CPR observations are highly condensed in the North Atlantic (and to a lesser extent south of Australia) for the *Bacillariophyceae* and *Dinophyceae* (Fig. 2A, B), but this aggregation is less clear for the *Haptophyta* (Fig. 2C), whose species have typically much smaller cells (often <10 µm) than the species of the former two groups. These three principal phytoplankton taxa have been well surveyed along the north-south AMT cruises, but they lack data in large areas of the South Pacific. Among the less species-rich taxonomic groups, including the *Cyanobacteria* (Fig. 2D) and *Chlorophyta*, global occurrence data coverage has been sparser (Fig. 2D, E). Since all of the principal phytoplankton taxa are globally abundant and widespread, the absence of records likely reflects a lack of sampling efforts rather than a lack of phytoplankton.

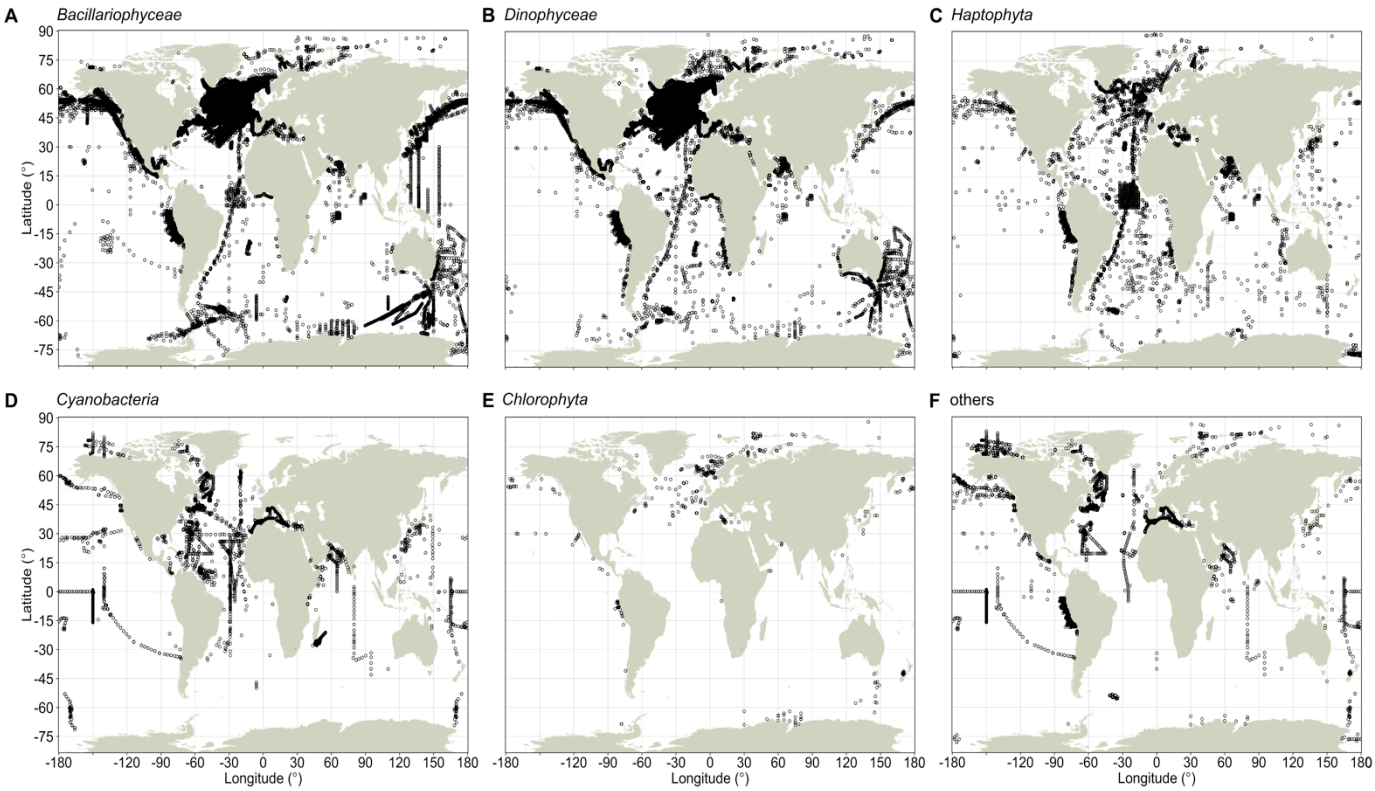

Figure 2: Global distribution of phytoplankton occurrence records in PhytoBase for individual taxa. Black circles show the distribution of in situ records for the five largest phyla or classes in the database that constitute 97.6% of all records (**A-E**) and for the remaining taxa (**F**). Records may overlap at any particular location.

3.1.2 Environmental coverage

The phytoplankton occurrences compiled cover the entire temperature range and a broad part of nitrate and mixed layer conditions found in the global ocean (Fig. 3A, B). To visualize the environmental data coverage, figure 3 matches occurrence records of PhytoBase with climatological sea surface data on nitrate (Garcia et al., 2013), temperature (Locarini

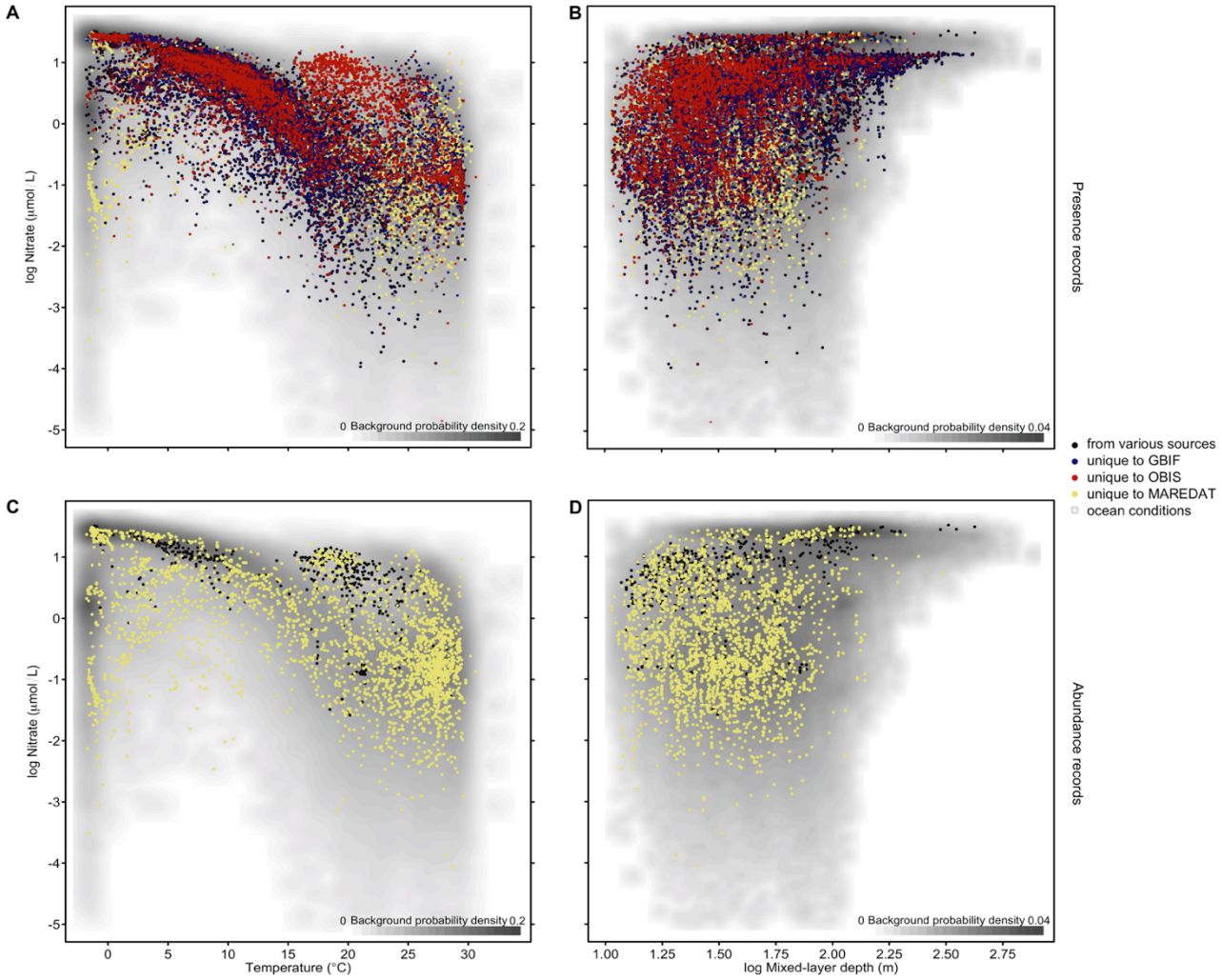

Figure 3: Phytoplankton records in environmental parameter space. (**A-B**) Dots display in situ records (n = 1 360 621) as a function of sea temperature and nitrate concentration (A), and as a function of mixed-layer depth (MLD) and nitrate concentration (B). The scale is logarithmic for MLD and nitrate. Shading indicates the frequency of environmental conditions appearing in the open ocean at surface, with darker grey shade indicating higher frequency of occurrence (bivariate Gaussian kernel density estimate). The colors of the dots denote the source of data, indicating complementarity or overlap of the environmental gradients sampled between sources. (**C-D**) Show the subset of records that contain information on species' cell counts per liter (n = 105 242), stemming largely from MareDat.

et al., 2013) and mixed-layer depth (de Boyer Montégut, 2004) at monthly 1° × 1° resolution. Records are concentrated in areas with intermediate conditions, which are relatively more frequent at the global scale (gray shade; Fig. 3A, B). Data on cell counts (7.7% of total) show a similar coverage as the full data (Fig. 3A, B), but are much thinner (Fig. 3C, D).

3.1.3 Taxonomic coverage

We assessed what fraction of the known marine phytoplankton species (Falkowski et al., 2004; Jordan, 2004; de Vargas et al., 2015) is represented in PhytoBase. The records include all major marine taxa of phytoplankton known (n = 16 classes), including the *Bacillariophyceae*, *Dinophyceae*, and *Haptophyta*. Records span roughly half of the known marine species of the *Haptophyta* (Jordan, 2004) and a similar fraction of the known marine species of *Bacillariophyceae* and *Dinophyceae* (Table 4). By contrast, species of the less species rich taxa tend to be more strongly underrepresented and account for a relatively small fraction (<3%) of all species in PhytoBase.

Record quantities are unevenly distributed between individual species (Fig. 4). Half of the species contain at least 30 presence records, but multiple species contribute one or two records (Fig. 4A). The species with less than 30 records account for as little as 0.54% of all species records in PhytoBase. Similarly, half of all genera contain at least 110 records each, while genera with less than 110 records each contribute as little as 8.2% to the total of records. A similar data distribution applies to the subset of species (n = 330), for which cell count records (with a volume basis) are available (Fig. 4B). Half of these species contribute at least 16 records, and among the genera containing cell counts, half contribute at least 76 records.

3.1.4 Completeness of species richness inventories at large spatial scales

We analyzed the ocean inventory of phytoplankton species richness in the database for three different regimes of ocean temperature by means of species accumulation curves (SACs) (Thompson and Withers, 2003) (Fig. 5). These curves present the cumulative species richness detected as a function of sampling effort (or survey area) and are expected to increase asymptotically before they saturate above a certain threshold of sampling effort (i.e., when the system has been exhaustively sampled). Using the number of sampling events (i.e., unique combinations of time, depth, location in our database) as a surrogate for sampling effort (x-axis), we find that the richness detected (y-axis) and the completeness of species richness detection (degree of saturation), differ notably between regimes. In the Southern temperate– (Fig. 5E) and cold seas (Fig. 5F), species richness has been incompletely sampled with respect to all taxa (black lines) or key taxa (colored lines). By contrast, SACs in the Northern Hemisphere start to saturate at ~40 000 samples, suggesting that the sampling has recorded a majority of the species. Specifically, SACs suggest that species richness will saturate at around ~1500 species in the tropical regime (>20°C), at ~1100 species in northern mid latitudes (≥10°C, ≤ 20°C), and at ~600 species in the cold Northern seas (<10°C). This corresponds to 93%, 64% and 35% of all ~1700 species collected in PhytoBase. However, these estimates only represent the fraction of species detectable via light microscopy and other methods underlying our database, preferentially omitting very rare or small species (Cermeño et al., 2014; Ser-Giacomi etal., 2018; Sogin et al., 2006). Thus, the richness will likely increase (at low rates) with additional sampling effort. Theoretical models have suggested that

Table 4: Statistics on the number of records and species contained in the database for key taxa

Taxon	Range (mean) of known marine species	Sources contributing to database	Records in database	Number of species or taxa in database (%)	% of marine species known
Bacillariophyceae (Cl.)	1800[†]-5000[§] (3400)	GBIF, OBIS, MareDat, Villar et al., Sal et al.	699 111	705 (41.2)	14-39
Dinophyceae (Cl.)	1780[†]-1800[§] (1790)	GBIF, OBIS, Villar et al., Sal et al.	527 293	778 (45.5)	43-44
Haptophyta (Ph.)	300[†,‖]-480[§] (360)	GBIF, OBIS, Sal et al., MareDat	47 183	166 (9.7)	34-55
Chlorophyta (Ph.)	100[§]-128[†] (114)	GBIF, OBIS	1304	22 (1.3)	17-22
Chrysophyceae (Cl.)	130[†]-800[§] (465)	GBIF, OBIS, Sal et al.	288	6 (0.4)	1-5
Cryptophyta (Ph.)	78[†]-100[§] (89)	GBIF, OBIS	2312	11 (0.6)	4-5
Cyanobacteria (Ph.)	150[§]	GBIF, OBIS, MareDat	53 060	7 (0.4)	5
Dictyochophyceae (Cl.)	200[†]	GBIF, Sal et al.	1824	8 (0.5)[‡]	4
Euglenoidea (Cl.)	30[§]-36[†] (33)	GBIF, OBIS	701	3 (0.2)	8-10
Raphidophyceae (Cl.)	4[†]-10[§] (7)	GBIF, OBIS	8	3 (0.2)	30-75
Picoeukaryotes	-	MareDat	27 537	1	-
Total	**4530[†,¶]-16 940[§]** (10 735)	**5**	**1 360 621**	**1710**	**10-38**

Cl., class. Ph, phylum.

The table summarizes the occurrence records for the ten major taxa in PhytoBase and describes to what degree the species in each taxon represent the total number of marine species known (for which exact numbers are still debated; we therefore provide upper and lower bounds, and mean values in parentheses).

[§]Falkowski et al. (2004). This estimate includes both coastal and open ocean taxa, while PhytoBase focuses primarily on data from the open ocean.

[†]de Vargas et al. (2015)

[‖]Jordan et al. (2004)

[‡]Including one species of the syster class *Pelagophyceae.*

[¶]The estimate by de Vargas et al. (2015) excluded prokaryotes. A number of 150 prokaryotes (Falkowski et al., 2004) were added to obtain the mean.

communities with many rare species lead to SACs with "low shoulders" meaning that SACs have a long upward slope to the asymptote (Thompson and Withers, 2003), consistent with our SACs (Fig. 5).

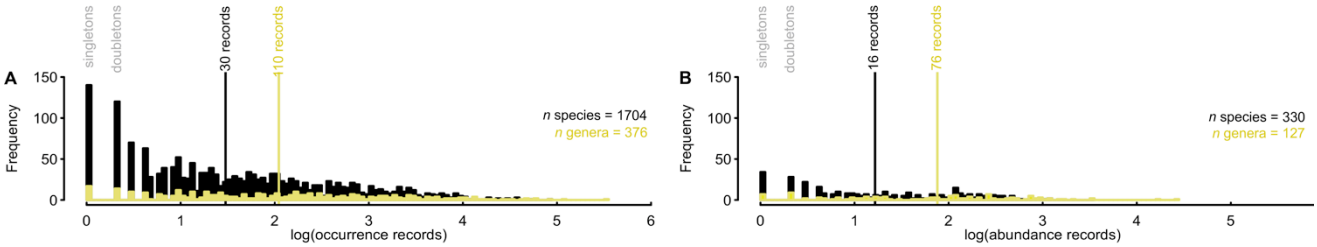

Figure 4: Distribution of occurrence records between species or genera. Histograms show the frequency of species (black) and genera (yellow) with a certain amount of (**A**) presence or (**B**) abundance records, separately. Vertical lines (black, yellow) indicate the median value. *X*-axes are logarithmic to the base ten.

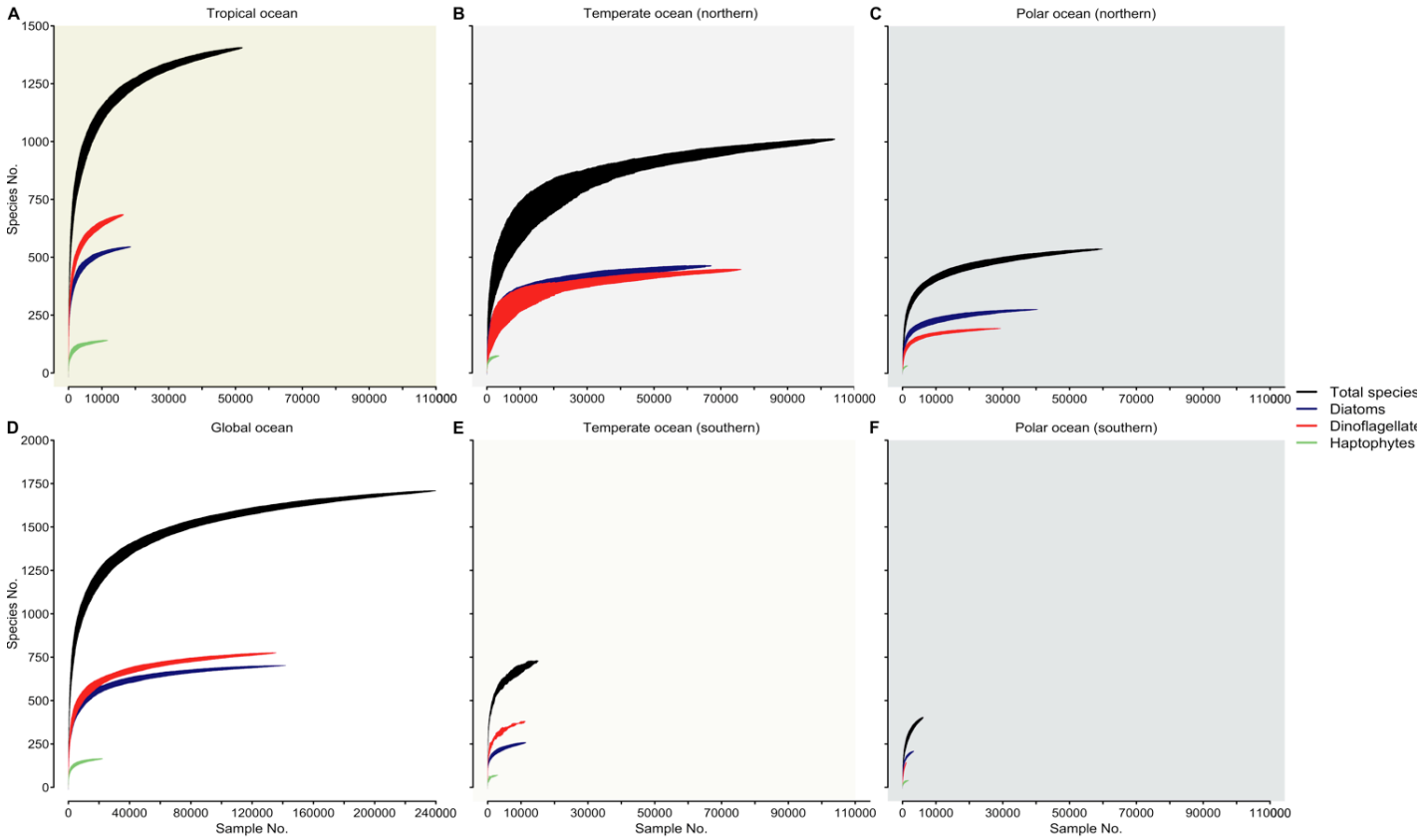

Figure 5: Accumulation of species richness as a function of sampling effort by region. Curves show the cumulative species richness as a function of samples (i.e., unique combinations of space, time and depth in the database, drawn at random) drawn at random from the database, using 100 runs (shadings around the curves indicate ± 1 S.D). Shown are species accumulation curves for all species (black) and three major taxa (colours) for (**A**) the tropics, defined as regions with a sea surface temperature (T) >20°C. (**B**) Temperate seas (10°C≤ T≤ 20°C) of the Northern Hemisphere. (**C**) Cold seas (T< 10°C) of the Northern Hemisphere. (**D**) Global ocean. (**E**) Temperate seas (10°C≤ T≤ 20°C) of the Southern Hemisphere. (**F**) Cold seas (T< 10°C) of the Southern Hemisphere. Background colors refer to figure 1A.

3.1.5 Species richness documented within 1° cells

To explore how completely species richness has been sampled at much smaller spatial scales, we binned data at 1° × 1° resolution, and analyzed the number of species in the pooled data per cell as a function of sampling effort. Hotspots in directly observed phytoplankton richness at the 1° cell level emerge in near-shore waters of Peru, around California, south-east of Australia, in the North Atlantic, along AMT cruises, and along research transects south of Japan (Fig. 6A). The species richness detected per 1° cell is positively correlated with sampling effort, using the number of samples collected per cell as a surrogate of sampling effort (Spearman's $\rho = 0.47$, $P < 0.001$). In particular, the richness of *Bacillariophyceae* ($\rho = 0.88$, $P < 0.001$) and of *Dinophyceae* ($\rho = 0.92$, $P < 0.001$) is positively correlated with effort, while this is less the case for the *Haptophyta* ($\rho = 0.27$; $P < 0.001$). Analyzing species richness as a function of "sampling events" for different thermal

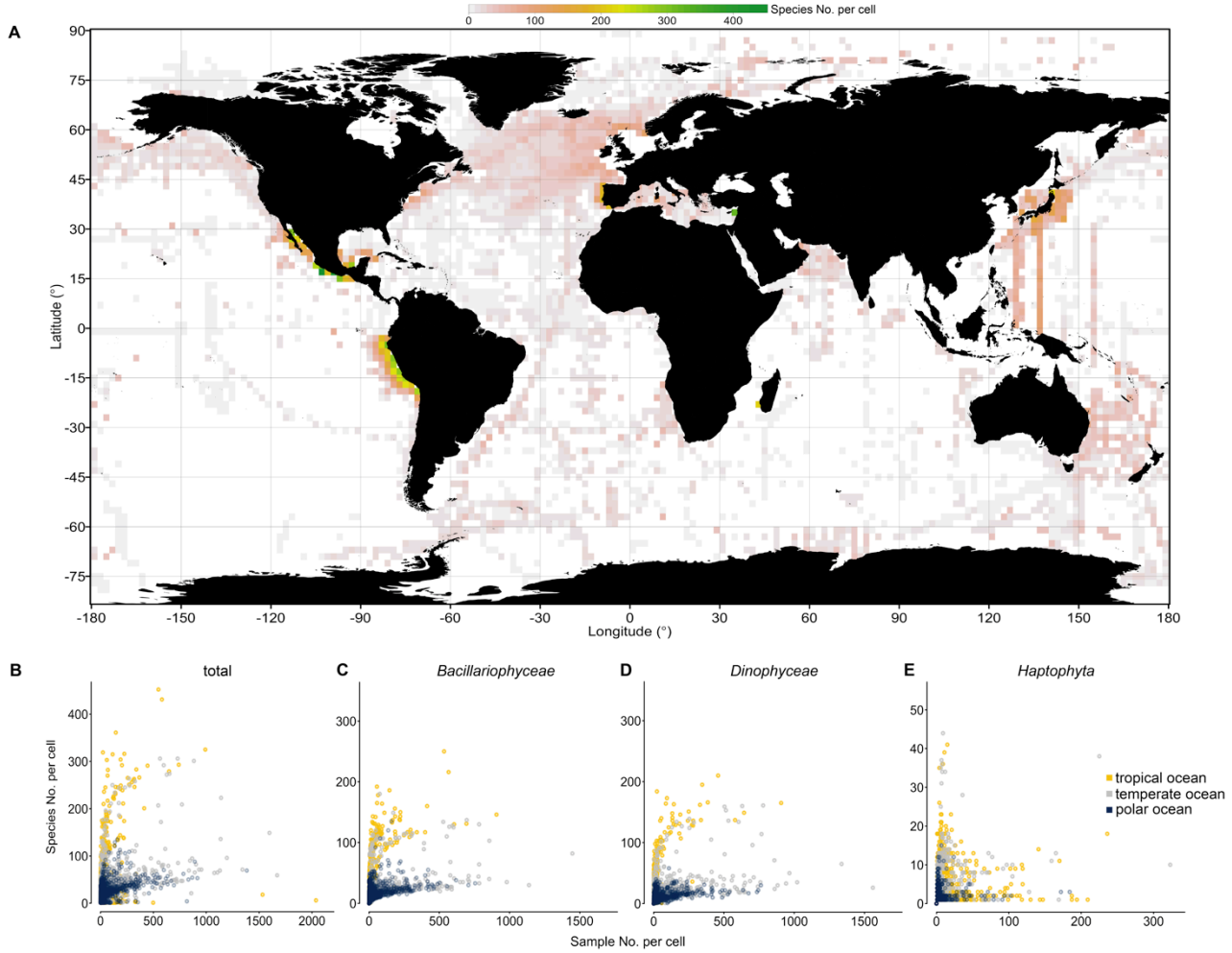

Figure 6: Species richness observed within 1° cells. (**A**) Global map visualizing the species richness detected within each 1° latitude x 1° longitude cell of the ocean. (The means of four 1° cells are depicted at 2°-resolution). (**B-E**) The number of species detected within each 1° cell is plotted as a function of sampling effort (i.e., number of sampling events, defined as unique combinations of position, time and depth in the database), with colours indicating data originating from different regions: tropical (T >20°C; yellow), temperate (10°C≤ T≤ 20°C; snow-white), and polar 1° cells (T< 10°C; light-blue), as defined by the annual mean temperature at sea surface (Locarini et al., 2013; see shading of map in figure 1). The richness-effort relationship is shown for all taxa (B), and major taxa separately (C-E).

regimes separately reveals that tropical areas (yellow dots; Fig. 6B-E) yield higher cumulative per cell richness at moderate to high sampling effort (> 50 samples), than temperate (grey dots) and polar areas (blue dots). Although data are thin and scattered, species richness in cold areas tends to saturate at ~70 species per cell (Fig. 6B; blue dots) at an effort of ~500 samples collected per cell. In contrast, species richness of the tropical areas tends to reach ~290 species per cell at the same effort (~500 samples). This suggests that tropical phytoplankton richness at the cell level is about four times higher than that

of cold northern areas, but richness may further increase with additional sampling effort. Analyzing the data of the major taxa separately suggests that roughly 200 species of *Bacillariophyceae* and *Dinophyceae* can be collected per cell at high sampling effort (~500 samples), yet data are sparse for *Haptophyta*, which broadly lack 1° cells with more than 100 samples collected (Fig. 6E).

The analysis of species richness detected per 1° cells suggests that approximately $^1/_3$ to $^1/_5$ of all species inventoried in the tropical or polar regime (see Fig. 5) through our database can be detected within a single 1°-cell of these regimes at high sampling effort (~500 samples) (Fig. 6B). This result is in coarse agreement with the result obtained at the large spatial scale (Fig. 5A-C), where the richness detected in the tropical regime was close to three times that of the (northern) cold regime.

3.1.6 Comparative spatial and taxonomic analysis of source datasets

We used the sources obtained from within the GBIF archive as an exemplary case for a more detailed examination of original source dataset coverage, as GBIF provides relatively detailed information on its sources via dataset keys. CPR is the single largest source dataset obtained from GBIF, which covers the North Atlantic and North Pacific (Fig. 7A-D; brown dots), and parts of the ocean south of Australia (Fig. 7A-D; blue dots). CPR records obtained via GBIF contribute 33.9% to all records in PhytoBase. CPR data show relatively low species numbers captured on average per "sample" (Fig. 7I), with samples being defined as exact combinations of geographic position, depth, and time in the data records. This may be owing to the continuous collection of species or incomplete reporting of taxa. The mesh size of the silk employed in CPR of 270 µm undersamples small phytoplankton species (<10 µm). Yet, small species nevertheless get regularly captured in CPR, as they get attached to the screens (Richardson et al., 2006). Within the 16 largest source datasets obtained via GBIF, the average number of species collected per sample is below four for the CPR program and increases to more than 50 for other datasets (Fig. 7I). These 16 test datasets (excluding datasets containing sedimentary records) highlight that the taxonomic resolution strongly differs between samples of individual cruises or survey programs. By latitude, different surveys or cruises thus contribute to PhytoBase to a varying degree (Fig. 7E-H). Systematic differences in the species detected per sample and the varying contribution of sources to the database along latitude (Fig. 7E-H) are important considerations when, for example, analyzing species richness directly.

Analysing the 16 largest source datasets from GBIF in environmental parameter space (Fig. 8) reveals the association of individual datasets with subdomains of the global temperatures, nitrate levels or mixed-layer depths (Fig. 8). GBIF datasets collected in the tropics and subtropics (mean temperature of sampling ≥20°C; Fig. 8A) tend to be associated with higher taxonomic detail (~25 species detected per sample on average; Fig. 7I), compared to datasets collected in colder areas. Yet, this likely also reflects an overall higher number of species occurring in tropical areas (Figs. 5A) than extratropical ones.

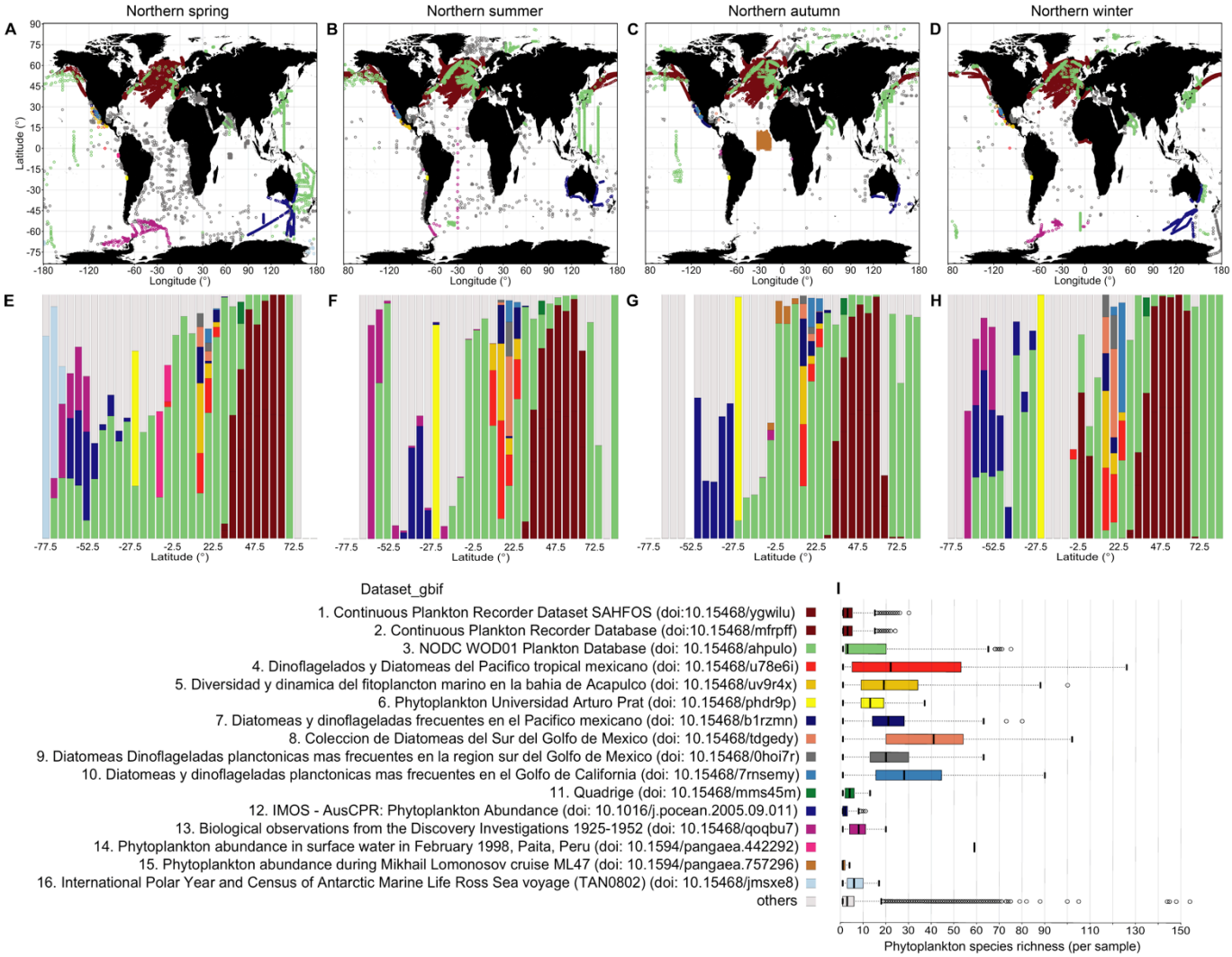

Figure 7: Spatial extent of the 16 largest datasets from GBIF and average per-sample richness. (**A-D**) Maps display the spatial distribution of the 16 largest contributing datasets to the GBIF-sourced data in PhytoBase, showing each season separately. The datasets presented comprise 54.3% of all records and 93.5% of GBIF-sourced records. GBIF data is shown as an exemplary case, as GBIF contributes a variety of source datasets referenced by dataset keys (datasetKey_gbif). Panels (**E-H**) show the relative contribution of datasets to the occurrence data per 5° latitude. Coloured sub-bars represent specific datasets (see **I**) and their widths the amount of data contributed. Panels (E–H) present the data shown in (A–D). (**I**) Analysis of within-sample species richness. Boxes show the mean species richness (thick vertical lines) detected per sample of specific datasets, and the first and third quartiles around the mean (boxes). Whiskers denote 2.5 times the inter quartile range. Note that the same analysis may be performed for OBIS data using the field “resourceID_obis”.

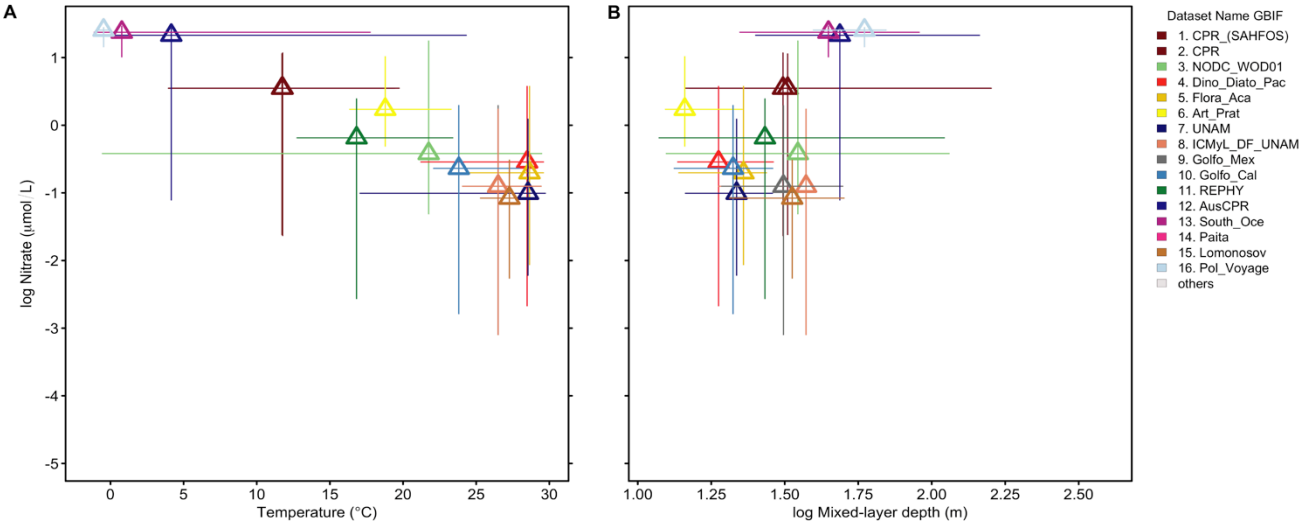

Figure 8: Environmental range of the 16 largest datasets from GBIF. (**A-B**) The range of 16 datasets contained within GBIF-sourced data, and the range of the dataset from Sal et al. (2015) are represented by thin lines in parameter space: (A) temperature *vs.* logarithmic nitrate concentration in the surface ocean, and (B) logarithmic mixed-layer depth vs. logarithmic nitrate (using climatological environmental data from Garcia et al., 2013; Locarini et al., 2013; de Boyer Montégut, 2004; matched with records at monthly climatological 1°-resolution). Lines span the minimum to maximum environmental condition associated with the records of each dataset separately. Triangles display the mean environmental condition of the records per dataset.

3.1.7 Sensitivity of data to taxonomic harmonization and coordinate rounding

Compared to OBIS, GBIF contributed roughly 14% additional records to the raw database (Table 2), yet this relative contribution changed after the harmonization step of species names. GBIF finally contributed 790 103 records, and OBIS 823 836 records to the harmonized PhytoBase. Hence, the exclusion of non-marine, fossil or doubtful species and the taxonomic harmonization step were more stringent for GBIF sourced records than OBIS sourced records.

We tested to what degree the number of unique records in the harmonized database changed when decimal positions in the coordinates of each of the five data sources were rounded prior to their merger. We find that the total number of unique records in PhytoBase declines continuously from 1.36 million to 1.07 million, when rounding the coordinates of records in the data to the 6th, 5th, 4th, 3rd, and 2nd decimal place. This result may be explained by the fact that large parts of the data come from CPR. The records of CPR start to be binned into coarser sampling units when rounding their decimal positions. The harmonized database (without coordinate rounding) contained 65.2% additional records compared to its largest contributing source. This gain was similarly high in the non-harmonized raw database and increased to ca. 73% when rounding coordinates to varying decimals. This shows that different sources contributed highly complementary records to PhytoBase, regardless of a coordinate rounding test to varying decimals.

4 Discussion

4.1 Data coverage, uncertainties, and recommendations

Spatiotemporal data on species occurrence are an essential basis to assess and forecast species' distributions and to understand the drivers behind these patterns. Following recent calls to gather species occurrences into global databases (Edwards, 2000; Meyer et al., 2015), we merged occurrence data of marine phytoplankton from three data sources and from the two largest open access biological data archives into PhytoBase. This new database contains 1 360 621 records (1 280 103 records at the level of species), describing 1716 species of seven phyla. Our effort addresses a gap in marine species occurrence data, as previous studies of marine taxa (Tittensor et al. 2010; Chaudhary et al. 2016; Menegotto & Rangel 2018) had no easy access to data sufficiently complete for global analyses of phytoplankton. The synthesis and harmonization of GBIF data with OBIS and other sources results in a substantial gain of phytoplankton occurrence data (> 60% additional data), relative to phytoplankton data residing in either of the two archives. The harmonization of different archives, which collect global species distribution evidence, therefore substantially expanded the empirical basis of phytoplankton records.

PhytoBase presents, to our knowledge, the currently largest global database of marine phytoplankton species occurrences. However, two main limitations remain: First, the global data density is spatially highly uneven and gaps persist across large swaths of the ocean, e.g., in the South Pacific and the central Indian. Second, sampling priorities with respect to taxonomic groups, size classes or species resolution differ widely between research cruises and programs. While small or fragile species may escape detection by the CPR program (Richardson et al., 2006), the resolution of traditional samples is influenced by sampling volume and taxonomic expertise (Cermeño et al., 2014). Our results show that the average number of species detected per sample varies from three to above 50 between different cruises or programs. A global spatial bias in collection density of marine species has been similarly found for heterotrophic taxa (Menegotto & Rangel 2018), but sampling biases and divergent sampling protocols may be even more common for phytoplankton.

Owing to these limitations, we recommend that direct analyses we recommend that direct analyses of the database be undertaken and interpreted with caution. For example, our data analysis has shown that direct species richness estimates are sensitive to the number of sampling events. In addition, many species have low occurrence numbers in the database, making any inference about their ecological niche or geographic distribution very uncertain. Thus, without careful screening and checking of the data (via e.g. datasetKeys for GBIF records, resourceIDs for OBIS records), the characterization of biogeographies at the species level might be highly biased.

Statistical techniques such as rarefaction (Rodríguez-Ramos et al., 2015), randomized resampling (Chaudhary et al., 2017), analysis of sampling gaps (Woolley et al. 2016; Menegotto & Rangel 2018), and species distribution modeling (Zimmermann and Guisan, 2000) may be implemented to overcome these limitations. The latter statistical technique may be particularly promising, as species distribution models can be set up to account for variation in presence data sampling (Phillips et al., 2009) and data scarceness (Breiner et al., 2015). Based on observed associations between species'

occurrences and environmental factors (Guisan and Thuiller, 2005), these models estimate the species' ecological niche, which is projected into geographic space, assuming that the species' niche and its geographic habitat are directly interrelated (Colwell and Rangel, 2009). Another advantage of species distribution models is that they can circumvent geographic sampling gaps through a spatial projection of the niche, as long as environmental conditions relevant to describe the niche of the species have been sufficiently well sampled and the species fills its ecological niche. This is the approach used by Righetti et al. (2019b), building on a large fraction of the PhytoBase (77.6% of the records, falling into the monthly climatological mixed-layer; de Boyer Montégut, 2004), to analyze global richness patterns of phytoplankton.

The detection of rare species and their integration into PhytoBase may become possible via molecular methods (Bork et al., 2015; Sogin et al., 2006). DNA sequencing has become an alternative approach to characterize phytoplankton biogeography (de Vargas et al., 2015). These data have two advantages over traditional taxonomic data: First, the sensitivity of metagenomic methods to detect rare taxa is relatively much higher. Second, metagenomic data have been collected in a methodologically consistent way in recent global surveys, such as the TARA Oceans cruise (de Vargas et al., 2015). But there are also drawbacks associated with DNA based methods. A large disadvantage of current metagenomic data is the lack of catalogued reference gene sequences for most species. As a result, the majority of the metagenomic sequences can only be determined to the level of genus (Malviya et al., 2016). However, we expect that an integration of detailed genetic data with traditional sampling data may soon become possible, allowing to combine several methodological or taxonomic dimensions in databases. At any point in the future, changing taxonomic nomenclature may be readily implemented, as we retained the original names and synonyms from raw data sources together with the harmonized name for each record in Phytobase.

4.2 Data use

Our data compilation and synthesis product PhytoBase has been designed to support primarily the analysis of the distribution, diversity, and abundance of phytoplankton species and related biotic or abiotic drivers in macroecological studies. But PhytoBase is far from limited to this set of applications, and may include the analysis of ecological niche differences between species or clades, linkages between species' ecological niches and phylogenetic or functional relatedness, current or future spatial projections of species' niches and composition, tests on whether presence-absence patterns of multiple species can predict community trait indices or joint analyses of species' distribution and trait data to project trait biogeographies. The database may also be used to validate the increasingly complex marine ecosystem models included in regional to global climate models.

The accuracy of data analyses may be limited by sampling biases underlying PhytoBase, including the spatiotemporal variation in sampling efforts and varying taxonomic detail between data sources. The latter limitation might be alleviated by considering different methodologies associated with varying cruises or collecting organisations in analyses. Where possible, we thus retained the information on the original dataset ID or dataset key along with each occurrence record in the database. Moreover, statistical analysis tools may be used to address spatiotemporal variation in global sampling efforts. Given the

critical undersampling of the Southern Hemisphere, data from areas such as the South Pacific will likely lead to new species discoveries and may greatly improve the global observational basis of phytoplankton occurrences in the future. Data inclusion from recent cruises, which are still under evaluation, appears as a natural next step. These data may come from the Malaspina expedition (Duarte, 2015), TARA Oceans (Bork et al., 2015) and Southern Ocean transects (Balch et al., 2016).

5 Data availability

PhytoBase is publicly available through PANGAEA, doi:10.1594/PANGAEA.904397 (Righetti et al., 2019a). Associated R scripts and the synonymy table used to harmonize species' names are available through https://gitlab.ethz.ch/phytobase/supplementary.

6 Conclusions

In PhytoBase, we compiled more than 1.36 million marine phytoplankton records that span 1704 species including the key taxa *Bacillariophyceae*, *Dinophyceae, Haptophyta*, *Cyanobacteria* and others. The database addresses photosynthetic microbial organisms, which play crucial roles in global biogeochemical cycles and marine ecology. We have provided an analysis of the current status of marine phytoplankton occurrence records accessible through public archives, their spatial and methodological limitations, and the completeness of species richness information for different ocean regions. PhytoBase may stimulate studies on the biogeography, diversity, and composition of phytoplankton and serve to calibrate ecological or mechanistic models. We recommend accounting carefully for data structure and metadata, depending on the purpose of analysis.

7 Appendices

Table A1: Harmonization of 113 taxon names in the MareDat dataset of O'Brien et al. (2013). Only the 113 names that changed during harmonization are shown, out of a total of 197 names.

Group	Original name	Harmonized name
Haptophyta	*_P. pouchetii*	*Phaeocystis pouchetii*
	*_P. pouchetii_*	*Phaeocystis pouchetii*
	_Phaeocystis pouchetii	*Phaeocystis pouchetii*
	_Phaeocystis pouchetii (Subcomponent: bladders)	*Phaeocystis pouchetii*
	_Phaeocystis spp.	*Phaeocystis*
	Phaeocystis spp.	*Phaeocystis*
	_Phaeocystis spp. (Subgroup: motile)	*Phaeocystis*
	_Phaeocystis spp. (Subgroup: non-motile)	*Phaeocystis*
	ACANTHOICA QUATTROSPINA	*Acanthoica quattrospina*
	Acanthoica acanthos	*Anacanthoica acanthos*

Acanthoica sp. cf. quattraspina	*Acanthoica quattrospina*
Algirosphaera oryza	*Algirosphaera robusta*
Algirosphaera robsta	*Algirosphaera robusta*
Anoplosolenia	*Anoplosolenia brasiliensis*
Anoplosolenia braziliensis	*Anoplosolenia brasiliensis*
Anoplosolenia sp. cf. brasiliensis	*Anoplosolenia brasiliensis*
Anthosphaera robusta	*Algirosphaera robusta*
CALCIDISCUS leptoporus	*Calcidiscus leptoporus*
Calcidiscus leptopora	*Calcidiscus leptoporus*
Calcidiscus leptoporus (inc. Coccolithus pelagicus)	*Calcidiscus leptoporus*
Calcidiscus leptoporus (small + intermediate)	*Calcidiscus leptoporus*
Calcidiscus leptoporus intermediate	*Calcidiscus leptoporus*
Calciosolenia MURRAYI	*Calciosolenia murrayi*
Calciosolenia brasiliensis	*Anoplosolenia brasiliensis*
Calciosolenia granii v closterium	*Anoplosolenia brasiliensis*
Calciosolenia granii v cylindrothecaf	*Calciosolenia murrayi*
Calciosolenia granii v cylindrothecaforma	*Calciosolenia murrayi*
Calciosolenia granii var closterium	*Anoplosolenia brasiliensis*
Calciosolenia granii var cylindrothecaeiformis	*Calciosolenia murrayi*
Calciosolenia murray	*Calciosolenia murrayi*
Calciosolenia siniosa	*Calciosolenia murrayi*
Calciosolenia sinuosa	*Calciosolenia murrayi*
Calciosolenia sp. cf. murrayi	*Calciosolenia murrayi*
Caneosphaera molischii	*Syracosphaera molischii*
Caneosphaera molischii and similar	*Syracosphaera molischii*
Coccolithus fragilis	*Oolithotus fragilis*
Coccolithus huxley	*Emiliania huxleyi*
Coccolithus huxleyi	*Emiliania huxleyi*
Coccolithus leptoporus	*Calcidiscus leptoporus*
Coccolithus sibogae	*Umbilicosphaera sibogae*
Crenalithus sessilis	*Reticulofenestra sessilis*
Crystallolithus cf rigidus	*Calcidiscus leptoporus*
Cyclococcolithus fragilis	*Oolithotus fragilis*
Discophaera tubifer	*Discosphaera tubifera*
Discosphaera thomsoni	*Discosphaera tubifera*
Discosphaera tubifer	*Discosphaera tubifera*
Discosphaera tubifer (inc. Papposphaera.lepida)	*Discosphaera tubifera*
Discosphaera tubifera	*Discosphaera tubifera*
Emiliana huxleyi	*Emiliania huxleyi*
Emiliania huxleyi A1	*Emiliania huxleyi*
Emiliania huxleyi A2	*Emiliania huxleyi*
Emiliania huxleyi A3	*Emiliania huxleyi*

	Emiliania huxleyi C	*Emiliania huxleyi*
	Emiliania huxleyi Indet.	*Emiliania huxleyi*
	Emiliania huxleyi var. Huxleyi	*Emiliania huxleyi*
	Florisphaera profunda var. profunda	*Florisphaera profunda*
	Halopappus adriaticus	*Michaelsarsia adriaticus*
	Helicosphaera carteri var. Carteri	*Helicosphaera carteri*
	Michelsarsia elegans	*Michaelsarsia elegans*
	Oolithotus fragilis var. Fragilis	*Oolithotus fragilis*
	Oolithus spp. cf fragilis	*Oolithotus fragilis*
	Ophiaster hydroideuss	*Ophiaster hydroideus*
	Ophiaster spp. cf. Hydroides	*Ophiaster hydroideus*
	P. antarctica	*Phaeocystis antarctica*
	*P. antarctica_*	*Phaeocystis antarctica*
	PHAEOCYSTIS	*Phaeocystis*
	*PHAEOCYSTIS_*	*Phaeocystis*
	PHAEOCYSTIS POUCHETII	*Phaeocystis pouchetii*
	*PHAEOCYSTIS POUCHETII_*	*Phaeocystis pouchetii*
	PHAEOCYSTIS sp.	*Phaeocystis*
	PHAEOCYSTIS sp._	*Phaeocystis*
	Palusphaera sp.	*Rhabdosphaera longistylis*
	Palusphaera vandeli	*Rhabdosphaera longistylis*
	*Phaeocystis antarctica_*	*Phaeocystis antarctica*
	Phaeocystis cf. pouchetii	*Phaeocystis pouchetii*
	*Phaeocystis cf. pouchetii_*	*Phaeocystis pouchetii*
	*Phaeocystis globosa_*	*Phaeocystis globosa*
	Phaeocystis motile	*Phaeocystis*
	*Phaeocystis motile_*	*Phaeocystis*
	Phaeocystis sp.	*Phaeocystis*
	Phaeocystis sp._	*Phaeocystis*
	Phaeocystis spp.	*Phaeocystis*
	Pontosphaera huxleyi	*Emiliania huxleyi*
	Rhabdosphaera sp. cf. claviger (inc. var. stylifera)	*Rhabdosphaera clavigera*
	Rhabdosphaera claviger	*Rhabdosphaera clavigera*
	Rhabdosphaera clavigera var. Clavigera	*Rhabdosphaera clavigera*
	Rhabdosphaera clavigera var. Stylifera	*Rhabdosphaera clavigera*
	Rhabdosphaera stylifera	*Rhabdosphaera clavigera*
	Rhabdosphaera tubifer	*Discosphaera tubifera*
	Rhabdosphaera tubulosa	*Discosphaera tubifera*
	Syrachosphaera pulchra	*Syracosphaera pulchra*
	Syracosphaera brasiliensis	*Anoplosolenia brasiliensis*
	Syracosphaera cf. Pulchra	*Syracosphaera pulchra*
	Syracosphaera confuse	*Ophiaster hydroideus*

	Syracosphaera corii	*Michaelsarsia adriaticus*
	Syracosphaera cornifera	*Helladosphaera cornifera*
	Syracosphaera corri	*Michaelsarsia adriaticus*
	Syracosphaera mediterranea	*Coronosphaera mediterranea*
	Syracosphaera molischii s.l.	*Syracosphaera molischii*
	Syracosphaera oblonga	*Calyptrosphaera oblonga*
	Syracosphaera quadricornu	*Algirosphaera robusta*
	Syracosphaera sp. cf. prolongata (inc. S.pirus)	*Syracosphaera prolongata*
	Syracosphaera tuberculata	*Coronosphaera mediterranea*
	Umbellosphaera hulburtiana	*Umbilicosphaera hulburtiana*
	Umbellosphaera sibogae	*Umbilicosphaera sibogae*
	Umbellosphaera spp. cf. irregularis + tenuis	*Umbellosphaera irregularis*
	Umbilicosphaera mirabilis	*Umbilicosphaera sibogae*
	Umbilicosphaera sibogae (Weber-van-Bosse) Gaarder	*Umbilicosphaera sibogae*
	Umbilicosphaera sibogae sibogae	*Umbilicosphaera sibogae*
	Umbilicosphaera sibogae var. Sibogae	*Umbilicosphaera sibogae*
	Umbilicosphaera spp. (U.sibogae)	*Umbilicosphaera sibogae*
	Umbillicosphaera sibogae	*Umbilicosphaera sibogae*

Note. An empty space in the original taxon name is indicated by "_".

Table A2: Harmonization of 156 taxon names in the MareDat dataset of Leblanc et al. (2012). Only the 156 names that changed during harmonization are shown, out of a total of 248 names.

Group	Original name	Harmonized name
Bacillariophyceae	*Actinocyclus coscinodiscoides*	*Roperia tesselata*
	Actinocyclus tessellatus	*Roperia tesselata*
	Asterionella frauenfeldii	*Thalassionema frauenfeldii*
	Asterionella glacialis	*Asterionellopsis glacialis*
	Asterionella mediterranea subsp pacifica	*Lioloma pacificum*
	Asterionellopsis japonica	*Asterionellopsis glacialis*
	Bacteriastrum varians	*Bacteriastrum furcatum*
	Cerataulina bergonii	*Cerataulina pelagica*
	Cerataulus bergonii	*Cerataulina pelagica*
	Ceratoneis closterium	*Cylindrotheca closterium*
	Ceratoneis longissima	*Nitzschia longissima*
	Chaetoceros angulatus	*Chaetoceros affinis*
	Chaetoceros atlanticus f. bulosus	*Chaetoceros bulbosus*
	Chaetoceros audax	*Chaetoceros atlanticus*
	Chaetoceros borealis f. concavicornis	*Chaetoceros concavicornis*
	Chaetoceros cellulosus	*Chaetoceros lorenzianus*
	Chaetoceros chilensis	*Chaetoceros peruvianus*
	Chaetoceros contortus	*Chaetoceros compressus*
	Chaetoceros convexicornis	*Chaetoceros peruvianus*

Chaetoceros dichaeta	*Chaetoceros distans*
Chaetoceros dispar	*Chaetoceros atlanticus*
Chaetoceros grunowii	*Chaetoceros decipiens*
Chaetoceros jahnischianus	*Chaetoceros distans*
Chaetoceros javanis	*Chaetoceros affinis*
Chaetoceros peruvio-atlanticus	*Chaetoceros peruvianus*
Chaetoceros polygonus	*Chaetoceros atlanticus*
Chaetoceros radians	*Chaetoceros socialis*
Chaetoceros radiculus	*Chaetoceros bulbosus*
Chaetoceros ralfsii	*Chaetoceros affinis*
Chaetoceros remotus	*Chaetoceros distans*
Chaetoceros schimperianus	*Chaetoceros bulbosus*
Chaetoceros schuttii	*Chaetoceros affinis*
Chaetocros vermiculatus	*Chaetoceros debilis*
Corethron criophilum	*Corethron pennatum*
Corethron hystrix	*Corethron pennatum*
Corethron valdivae	*Corethron pennatum*
Coscinodiscus anguste-lineatus	*Thalassiosira anguste-lineata*
Coscinodiscus gravidus	*Thalassiosira gravida*
Coscinodiscus pelagicus	*Thalassiosira gravida*
Coscinodiscus polychordus	*Thalassiosira anguste-lineata*
Coscinodiscus rotulus	*Thalassiosira gravida*
Coscinodiscus sol	*Planktoniella sol*
Coscinodiscus sublineatus	*Thalassiosira anguste-lineata*
Coscinosira polychordata	*Thalassiosira anguste-lineata*
Dactyliosolen mediterraneus	*Leptocylindrus mediterraneus*
Dactyliosolen meleagris	*Leptocylindrus mediterraneus*
Detonula delicatula	*Detonula pumila*
Diatoma rhombica	*Fragilariopsis rhombica*
Dicladia bulbosa	*Chaetoceros bulbosus*
Dithylim inaequale	*Ditylum brightwellii*
Dithylum trigonum	*Ditylum brightwellii*
Eucampia balaustium	*Eucampia antarctica*
Eucampia Britannica	*Eucampia zodiacus*
Eucampia nodosa	*Eucampia zodiacus*
Eucampia striata	*Guinardia striata*
Eupodiscus tesselatus	*Roperia tesselata*
Fragilaria arctica	*Fragilariopsis oceanica*
Fragilaria kerguelensis	*Fragilariopsis kerguelensis*
Fragilaria obliquecostata	*Fragilariopsis obliquecostata*
Fragilaria rhombica	*Fragilariopsis rhombica*
Fragilariopsis antarctica	*Fragilariopsis oceanica*

Fragilariopsis sublinearis	*Fragilariopsis obliquecostata*
Fragilaris sublinearis	*Fragilariopsis obliquecostata*
Fragillariopsis antarctica	*Fragilariopsis kerguelensis*
Gallionella sulcata	*Paralia sulcata*
Guinardia baltica	*Guinardia flaccida*
Hemiaulus delicatulus	*Hemiaulus hauckii*
Henseniella baltica	*Guinardia flaccida*
Homeocladia closterium	*Cylindrotheca closterium*
Homeocladia delicatissima	*Pseudo-nitzschia delicatissima*
Lauderia borealis	*Lauderia annulata*
Lauderia pumila	*Detonula pumila*
Lauderia schroederi	*Detonula pumila*
Leptocylindrus belgicus	*Leptocylindrus minimus*
Melosira costata	*Skeletonema costatum*
Melosira marina	*Paralia sulcata*
Melosira sulcata	*Paralia sulcata*
Moerellia cornuta	*Eucampia cornuta*
Navicula mebranacea	*Meuniera membranacea*
Navicula planamembranacea	*Ephemera planamembranacea*
Navicula pseudomembranacea	*Meuniera membranacea*
Nitzschia actydrophila	*Pseudo-nitzschia delicatissima*
Nitzschia angulate	*Fragilariopsis rhombica*
Nitzschia Antarctica	*Fragilariopsis rhombica*
Nitzschia birostrata	*Nitzschia longissima*
Nitzschia closterium	*Cylindrotheca closterium*
Nitzschia curvirostris	*Cylindrotheca closterium*
Nitzschia delicatissima	*Pseudo-nitzschia delicatissima*
Nitzschia grunowii	*Fragilariopsis oceanica*
Nitzschia heimii	*Pseudo-nitzschia heimii*
Nitzschia kergelensis	*Fragilariopsis kerguelensis*
Nitzschia obliquecostata	*Fragilariopsis obliquecostata*
Nitzschia pungens	*Pseudo-nitzschia pungens*
Nitzschia seriata	*Pseudo-nitzschia seriata*
Nitzschiella longissima	*Nitzschia longissima*
Nitzschiella tenuirostris	*Cylindrotheca closterium*
Orthoseira angulate	*Thalassiosira angulata*
Orthoseira marina	*Paralia sulcata*
Orthosira marina	*Paralia sulcata*
Paralia marina	*Paralia sulcata*
Planktoniella wolterecki	*Planktoniella sol*
Podosira subtilis	*Thalassiosira subtilis*
Proboscia alata f. alata	*Proboscia alata*

Proboscia alata f. gracillima	*Proboscia alata*
Proboscia gracillima	*Proboscia alata*
Pyxilla baltica	*Rhizosolenia setigera*
Rhizosolenia alata	*Proboscia alata*
Rhizosolenia alata f. indica	*Proboscia indica*
Rhizosolenia alata var. indica	*Proboscia indica*
Rhizosolenia amputata	*Rhizosolenia bergonii*
Rhizosolenia antarctica	*Guinardia cylindrus*
Rhizosolenia calcar	*Pseudosolenia calcar-avis*
Rhizosolenia calcar avis	*Pseudosolenia calcar-avis*
Rhizosolenia calcar-avis	*Pseudosolenia calcar-avis*
Rhizosolenia cylindrus	*Guinardia cylindrus*
Rhizosolenia delicatula	*Guinardia delicatula*
Rhizosolenia flaccida	*Guinardia flaccida*
Rhizosolenia fragilima	*Dactyliosolen fragilissimus*
Rhizosolenia fragilissima	*Dactyliosolen fragilissimus*
Rhizosolenia genuine	*Proboscia alata*
Rhizosolenia gracillima	*Proboscia alata*
Rhizosolenia hebetata f hiemalis	*Rhizosolenia hebetata*
Rhizosolenia hebetata f. hebetata	*Rhizosolenia hebetata*
Rhizosolenia hebetata f. semispina	*Rhizosolenia hebetata*
Rhizosolenia hensenii	*Rhizosolenia setigera*
Rhizosolenia indica	*Proboscia indica*
Rhizosolenia japonica	*Rhizosolenia setigera*
Rhizosolenia murrayana	*Rhizosolenia chunii*
Rhizosolenia semispina	*Rhizosolenia hebetata*
Rhizosolenia stolterfothii	*Guinardia striata*
Rhizosolenia strubsolei	*Rhizosolenia imbricata*
Rhizosolenia styliformis var. longispina	*Rhizosolenia styliformis*
Rhizosolenia styliformis var. polydactyla	*Rhizosolenia styliformis*
Rhizosolenia styliformis var. semispina	*Rhizosolenia hebetata*
Schroederella delicatula	*Detonula pumila*
Spingeria bacillaris	*Thalassionema bacillare*
Stauroneis membranacea	*Meuniera membranacea*
Stauropsis membranacea	*Meuniera membranacea*
Synedra nitzschioides	*Thalassionema nitzschioides*
Synedra thalassiothrix	*Thalassiothrix longissima*
Terebraria kerguelensis	*Fragilariopsis kerguelensis*
Thalassionema elegans	*Thalassionema bacillare*
Thalassiosira condensata	*Detonula pumila*
Thalassiosira decipiens	*Thalassiosira angulate*
Thalassiosira polychorda	*Thalassiosira anguste-lineata*

	Thalassiosira rotula	*Thalassiosira gravida*
	Thalassiosira tcherniai	*Thalassiosira gravida*
	Thalassiothrix curvata	*Thalassionema nitzschioides*
	Thalassiothrix delicatula	*Lioloma delicatulum*
	Thalassiothrix frauenfeldii	*Thalassionema frauenfeldii*
	Thalassiothrix fraunfeldii	*Thalassionema nitzschioides*
	Thalassiothrix mediterranea var. pacifica	*Lioloma pacificum*
	Trachysphenia australis v kerguelensis	*Fragilariopsis kerguelensis*
	Triceratium brightwellii	*Ditylum brightwellii*
	Zygoceros pelagica	*Cerataulina pelagica*
	Zygoceros pelagicum	*Cerataulina pelagica*

560

Table A3: Harmonization of the total of 109 species names in the data from Villar et al. (2015). Only the 109 names that changed during harmonization are shown, out of a total of 201 names.

Group	**Original name**	**Harmonized name**
Bacillariophyceae	*Asteromphalus cf. flabellatus*	*Asteromphalus*
	Asteromphalus spp.	*Asteromphalus*
	Bacteriastrum cf. delicatulum	*Bacteriastrum*
	Bacteriastrum cf. elongatum	*Bacteriastrum*
	Bacteriastrum cf. furcatum	*Bacteriastrum*
	Bacteriastrum cf. hyalinum	*Bacteriastrum*
	Bacteriastrum spp.	*Bacteriastrum*
	Biddulphia spp.	*Biddulphia*
	Chaetoceros atlanticus var. neapolitanus	*Chaetoceros atlanticus*
	Chaetoceros bulbosum	*Chaetoceros bulbosus*
	Chaetoceros cf. atlanticus	*Chaetoceros*
	Chaetoceros cf. coarctatus	*Chaetoceros*
	Chaetoceros cf. compressus	*Chaetoceros*
	Chaetoceros cf. danicus	*Chaetoceros*
	Chaetoceros cf. densus	*Chaetoceros*
	Chaetoceros cf. dichaeta	*Chaetoceros*
	Chaetoceros cf. laciniosus	*Chaetoceros*
	Chaetoceros cf. lorenzianus	*Chaetoceros*
	Chaetoceros spp.	*Chaetoceros*
	Climacodium cf. fravenfeldianum	*Climacodium*
	Climacodium spp.	*Climacodium*
	Corethron cf. pennatum	*Corethron*
	Corethron spp.	*Corethron*
	Coscinodiscus spp.	*Coscinodiscus*
	Cylindrotheca spp.	*Cylindrotheca*
	Ditylum spp.	*Ditylum*

	Eucampia antartica	*Eucampia antarctica*
	Eucampia spp.	*Eucampia*
	Eucampia zodiacus f. cylindrocornis	*Eucampia zodiacus*
	Fragilariopsis spp.	*Fragilariopsis*
	Haslea wawrickae	*Haslea wawrikae*
	Hemiaulus spp.	*Hemiaulus*
	Hemidiscus cf. cuneiformis	*Hemidiscus*
	Lauderia spp.	*Lauderia*
	Leptocylindrus cf. danicus	*Leptocylindrus*
	Leptocylindrus cf. minimus	*Leptocylindrus*
	Lithodesmium spp.	*Lithodesmium*
	Nitzschia spp.	*Nitzschia*
	Odontella spp.	*Odontella*
	Pseudo-nitzschia cf. fraudulenta	*Pseudo-nitzschia*
	Pseudo-nitzschia cf. subcurvata	*Pseudo-nitzschia*
	Pseudo-nitzschia delicatissima group	*Pseudo-nitzschia delicatissima*
	Pseudo-nitzschia pseudodelicatissima group	*Pseudo-nitzschia pseudodelicatissima*
	Pseudo-nitzschia seriata group	*Pseudo-nitzschia seriata*
	Pseudo-nitzschia spp.	*Pseudo-nitzschia*
	Rhizosolenia cf. acuminata	*Rhizosolenia*
	Rhizosolenia cf. bergonii	*Rhizosolenia*
	Rhizosolenia cf. curvata	*Rhizosolenia*
	Rhizosolenia cf. decipiens	*Rhizosolenia*
	Rhizosolenia cf. hebetata	*Rhizosolenia*
	Rhizosolenia cf. imbricata	*Rhizosolenia*
	Rhizosolenia spp.	*Rhizosolenia*
	Skeletonema spp.	*Skeletonema*
	Thalassionema spp.	*Thalassionema*
	Thalassiosira spp.	*Thalassiosira*
Dinophyceae	*Amphidinium spp.*	*Amphidinium*
	Archaeperidinium cf. minutum	*Archaeperidinium*
	Blepharocysta spp.	*Blepharocysta*
	Ceratocorys cf. gourreti	*Ceratocorys*
	Ceratocorys spp.	*Ceratocorys*
	Dinophysis cf. acuminata	*Dinophysis*
	Dinophysis cf. ovum	*Dinophysis*
	Dinophysis cf. uracantha	*Dinophysis*
	Dinophysis spp.	*Dinophysis*
	Diplopsalis group	*Diplopsalis*
	Gonyaulax cf. apiculata	*Gonyaulax*
	Gonyaulax cf. elegans	*Gonyaulax*
	Gonyaulax cf. fragilis	*Gonyaulax*

	Gonyaulax cf. hyalina	*Gonyaulax*
	Gonyaulax cf. pacifica	*Gonyaulax*
	Gonyaulax cf. polygramma	*Gonyaulax*
	Gonyaulax cf. scrippsae	*Gonyaulax*
	Gonyaulax cf. sphaeroidea	*Gonyaulax*
	Gonyaulax cf. spinifera	*Gonyaulax*
	Gonyaulax cf. striata	*Gonyaulax*
	Gonyaulax spp.	*Gonyaulax*
	Gymnodinium spp.	*Gymnodinium*
	Gyrodinium spp.	*Gyrodinium*
	Histioneis cf. megalocopa	*Histioneis*
	Histioneis cf. striata	*Histioneis*
	Oxytoxum cf. laticeps	*Oxytoxum*
	Oxytoxum spp.	*Oxytoxum*
	Paleophalacroma unicinctum	*Palaeophalacroma unicinctum*
	Phalacroma cf. rotundatum	*Phalacroma*
	Prorocentrum cf. balticum	*Prorocentrum*
	Prorocentrum cf. concavum	*Prorocentrum*
	Prorocentrum cf. nux	*Prorocentrum*
	Protoceratium spinolosum	*Protoceratium spinulosum*
	Protoperidinium cf. bipes	*Protoperidinium*
	Protoperidinium cf. breve	*Protoperidinium*
	Protoperidinium cf. crassipes	*Protoperidinium*
	Protoperidinium cf. diabolum	*Protoperidinium*
	Protoperidinium cf. divergens	*Protoperidinium*
	Protoperidinium cf. globulus	*Protoperidinium*
	Protoperidinium cf. grainii	*Protoperidinium*
	Protoperidinium cf. leonis	*Protoperidinium*
	Protoperidinium cf. monovelum	*Protoperidinium*
	Protoperidinium cf. nudum	*Protoperidinium*
	Protoperidinium cf. ovatum	*Protoperidinium*
	Protoperidinium cf. ovum	*Protoperidinium*
	Protoperidinium cf. pyriforme	*Protoperidinium*
	Protoperidinium cf. quarnerense	*Protoperidinium*
	Protoperidinium cf. steinii	*Protoperidinium*
	Protoperidinium cf. variegatum	*Protoperidinium*
	Protoperidinuim spp.	*Protoperidinium*
	Schuettiella cf. mitra	*Schuettiella*
	Tripos arietinum	*Tripos arietinus*
	Tripos lineatus/pentagonus complex	*Tripos lineatus*
	Tripos massiliense	*Tripos massiliensis*

Note. Data of genera (using the harmonized names) were excluded from the database.

565

8 Author contributions

MV, NG, NEZ and DR conceived the study. MG performed the taxonomic expert screening. MV compiled substantial parts of the MareDat database. DR compiled the data, developed the code, and led the writing, with inputs by all authors.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

We thank all biologists, taxonomists and cruise organizers for their essential efforts in collecting and sharing occurrence data of marine phytoplankton, synthesized in this database. We thank for expertise on taxonomic nomenclature provided by M. Estrada. We thank M. Döring and P. Provoost for their support with retrieval of phytoplankton occurrence data from GBIF (www.gbif.org) and OBIS (www.obis.org). Funding for this effort came from ETH Zürich under grant ETH-52 13-2.

References

Aman, A. A. and Bman, B. B.: The test article, J. Sci. Res., 12, 135–147, doi:10.1234/56789, 2015.

Aiken, J., Rees, N., Hooker, S., Holligan, P., Bale, A., Robins, D., Moore, G., Harris, R. and Pilgrim, D.: The Atlantic Meridional Transect: overview and synthesis of data, Prog. Oceanogr., 45(3–4), 257–312, doi:10.1016/S0079-6611(00)00005-7, 2000.

Amante, C. and Eakins, B. W.: ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis, (NOAA Tech. Memo. NESDIS NGDC-24, Natl. Geophys. Data Center, NOAA, 2009)., doi:10.7289/V5C8276M, 2009.

Balch, W. M., Bates, N. R., Lam, P. J., Twining, B. S., Rosengard, S. Z., Bowler, B. C., Drapeau, D. T., Garley, R., Lubelczyk, L. C., Mitchell, C. and Rauschenberg, S.: Factors regulating the Great Calcite Belt in the Southern Ocean and its biogeochemical significance, Global Biogeochem. Cycles, 30(8), 1124–1144, doi:10.1002/2016GB005414, 2016.

Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E. and Wincker, P.: Tara Oceans studies plankton at planetary scale, Science, 348(6237), 873–873, doi:10.1126/science.aac5605, 2015.

de Boyer Montégut, C.: Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology, J. Geophys. Res., 109(C12), C12003, doi:10.1029/2004JC002378, 2004.

Breiner, F. T., Guisan, A., Bergamini, A. and Nobis, M. P.: Overcoming limitations of modelling rare species by using ensembles of small models, Methods Ecol. Evol., 6(10), 1210–1218, doi:10.1111/2041-210X.12403, 2015.

Brun, P., Vogt, M., Payne, M. R., Gruber, N., O'Brien, C. J., Buitenhuis, E. T., Le Quéré, C., Leblanc, K. and Luo, Y.-W.: Ecological niches of open ocean phytoplankton taxa, Limnol. Oceanogr., 60(3), 1020–1038, doi:10.1002/lno.10074, 2015.

Buitenhuis, E. T., Li, W. K. W., Vaulot, D., Lomas, M. W., Landry, M. R., Partensky, F., Karl, D. M., Ulloa, O., Campbell, L., Jacquet, S., Lantoine, F., Chavez, F., Macias, D., Gosselin, M. and McManus, G. B.: Picophytoplankton biomass distribution in the global ocean, Earth Syst. Sci. Data, 4(1), 37–46, doi:10.5194/essd-4-37-2012, 2012.

Buitenhuis, E. T., Vogt, M., Moriarty, R., Bednaršek, N., Doney, S. C., Leblanc, K., Le Quéré, C., Luo, Y.-W., O'Brien, C., O'Brien, T., Peloquin, J., Schiebel, R. and Swan, C.: MAREDAT: towards a world atlas of MARine Ecosystem DATa, Earth Syst. Sci. Data, 5(2), 227–239, doi:10.5194/essd-5-227-2013, 2013.

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

Chaudhary, C., Saeedi, H. and Costello, M. J.: Bimodality of Latitudinal Gradients in Marine Species Richness, Trends Ecol. Evol., 31(9), 670–676, doi:10.1016/j.tree.2016.06.001, 2016.

Chaudhary, C., Saeedi, H. and Costello, M. J.: Marine Species Richness Is Bimodal with Latitude: A Reply to Fernandez and Marques, Trends Ecol. Evol., 32(4), 234–237, doi:10.1016/j.tree.2017.02.007, 2017.

Colwell, R. K. and Rangel, T. F.: Hutchinson's duality: The once and future niche, Proc. Natl. Acad. Sci., 106(Supplement_2), 19651–19658, doi:10.1073/pnas.0901650106, 2009.

Conway, J., Eddelbuettel, D., Nishiyama, T., Prayaga, S. K., Tiffin, N.: RPostgreSQL: R interface to the PostgreSQL database system. R package version 0.4.1. https://cran.r-project.org/package=RPostgreSQL, 2015.

Chamberlain, S.: rgbif: Interface to the Global Biodiversity Information Facility API. R package version 0.9.7. https://cran.r-project.org/package=rgbif, 2015.

Duarte, C. M.: Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition, Limnol. Oceanogr. Bull., 24(1), 11–14, doi:10.1002/lob.10008, 2015.

Edwards, J. L.: Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop, Science, 289(5488), 2312–2314, doi:10.1126/science.289.5488.2312, 2000.

Endo, H., Ogata, H. and Suzuki, K.: Contrasting biogeography and diversity patterns between diatoms and haptophytes in the central Pacific Ocean, Sci. Rep., 8(1), 10916, doi:10.1038/s41598-018-29039-9, 2018.

Falkowski, P. G., Katz M. E., Knoll, A. H., Quigg, A., Raven, J. A., Schofield, O. and Taylor, F. J. R.: The evolution of modern eukaryotic phytoplankton, Science, 305(5682), 354–360, doi:10.1126/science.1095964, 2004.

Field, C. B., Behrenfeld, M. J., Tanderson, J. T. and Falkowski, P.: Primary production of the biosphere: Integrating

terrestrial and oceanic components, Science, 281(5374), 237–240, doi:10.1126/science.281.5374.237, 1998.

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincon, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K. W., Lomas, M. W., Veneziano, D., Vera, C. S., Vrugt, J. A. and Martiny, A. C.: Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus, Proc. Natl. Acad. Sci., 110(24), 9824–9829, doi:10.1073/pnas.1307701110, 2013.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., Reagan, J. R. and Johnson, D. R.: World Ocean Atlas 2013, Vol. 4 Dissolved Inorg. Nutr. (phosphate, nitrate, Silic. S. Levitus, Ed.; A. Mishonov, Tech. Ed., 25, 2013.

Guisan, A. and Thuiller, W.: Predicting species distribution: Offering more than simple habitat models, Ecol. Lett., 8(9), 993–1009, doi:10.1111/j.1461-0248.2005.00792.x, 2005.

Guisan, A. and Zimmermann, N. E.: Predictive habitat distribution models in ecology, Ecol. Modell., 135(2–3), 147–186, doi:10.1016/S0304-3800(00)00354-9, 2000.

Honjo, S. and Okada, H.: Community structure of soccolithophores in the photic layer of the mid-pacific, Micropaleontology, 20(2), 209, doi:10.2307/1485061, 1974.

Iglesias-Rodríguez, M. D., Brown, C. W., Doney, S. C., Kleypas, J., Kolber, D., Kolber, Z., Hayes, P. K. and Falkowski, P. G.: Representing key phytoplankton functional groups in ocean carbon cycle models: Coccolithophorids, Global Biogeochem. Cycles, 16(4), 47-1-47–20, doi:10.1029/2001GB001454, 2002.

Jeong, H. J., Yoo, Y. Du, Kim, J. S., Seong, K. A., Kang, N. S. and Kim, T. H.: Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs, Ocean Sci. J., 45(2), 65–91, doi:10.1007/s12601-010-0007-2, 2010.

Jones, M. C. and Cheung, W. W. L.: Multi-model ensemble projections of climate change effects on global marine biodiversity, ICES J. Mar. Sci., 72(3), 741–752, doi:10.1093/icesjms/fsu172, 2015.

Jordan, R. W.: A revised classification scheme for living haptophytes, Micropaleontology, 50(Suppl_1), 55–79, doi:10.2113/50.Suppl_1.55, 2004.

Leblanc, K., Arístegui, J., Armand, L., Assmy, P., Beker, B., Bode, A., Breton, E., Cornet, V., Gibson, J., Gosselin, M.-P., Kopczynska, E., Marshall, H., Peloquin, J., Piontkovski, S., Poulton, A. J., Quéguiner, B., Schiebel, R., Shipe, R., Stefels, J., van Leeuwe, M. A., Varela, M., Widdicombe, C. and Yallop, M.: A global diatom database – abundance, biovolume and biomass in the world ocean, Earth Syst. Sci. Data, 4(1), 149–165, doi:10.5194/essd-4-149-2012, 2012.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., Garcia, H. E., Baranova, O. K., Zweng, M. M., Paver, C. R., Reagan, J. R., Johnson, D. R., Hamilton, M. and Seidov, D.: World Ocean Atlas 2013, Vol. 2 Temp. S. Levitus, Ed., A.

Mishonov Tech. Ed.; NOAA Atlas NESDIS 73, 40, 2013.

Lund, J. W. G., Kipling, C. and Le Cren, E. D.: The inverted microscope method of estimating algal numbers and the statistical basis of estimations by counting, Hydrobiologia, 11(2), 143–170, doi:10.1007/BF00007865, 1958.

Luo, Y.-W., Doney, S. C., Anderson, L. A., Benavides, M., Berman-Frank, I., Bode, A., Bonnet, S., Boström, K. H., Böttjer, D., Capone, D. G., Carpenter, E. J., Chen, Y. L., Church, M. J., Dore, J. E., Falcón, L. I., Fernández, A., Foster, R. A., Furuya, K., Gómez, F., Gundersen, K., Hynes, A. M., Karl, D. M., Kitajima, S., Langlois, R. J., LaRoche, J., Letelier, R. M., Marañón, E., McGillicuddy, D. J., Moisander, P. H., Moore, C. M., Mouriño-Carballido, B., Mulholland, M. R., Needoba, J. A., Orcutt, K. M., Poulton, A. J., Rahav, E., Raimbault, P., Rees, A. P., Riemann, L., Shiozaki, T., Subramaniam, A., Tyrrell, T., Turk-Kubo, K. A., Varela, M., Villareal, T. A., Webb, E. A., White, A. E., Wu, J. and Zehr, J. P.: Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates, Earth Syst. Sci. Data, 4(1), 47–73, doi:10.5194/essd-4-47-2012, 2012.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L., Zingone, A. and Bowler, C.: Insights into global diatom distribution and diversity in the world's ocean, Proc. Natl. Acad. Sci., 113(11), E1516–E1525, doi:10.1073/pnas.1509523113, 2016.

Mawji, E., Schlitzer, R., Dodas, E. M., Abadie, C., Abouchami, W., Anderson, R. F., Baars, O., Bakker, K., Baskaran, M., Bates, N. R., Bluhm, K., Bowie, A., Bown, J., Boye, M., Boyle, E. A., Branellec, P., Bruland, K. W., Brzezinski, M. A., Bucciarelli, E., Buesseler, K., Butler, E., Cai, P., Cardinal, D., Casciotti, K., Chaves, J., Cheng, H., Chever, F., Church, T. M., Colman, A. S., Conway, T. M., Croot, P. L., Cutter, G. A., de Baar, H. J. W., de Souza, G. F., Dehairs, F., Deng, F., Dieu, H. T., Dulaquais, G., Echegoyen-Sanz, Y., Lawrence Edwards, R., Fahrbach, E., Fitzsimmons, J., Fleisher, M., Frank, M., Friedrich, J., Fripiat, F., Galer, S. J. G., Gamo, T., Solsona, E. G., Gerringa, L. J. A., Godoy, J. M., Gonzalez, S., Grossteffan, E., Hatta, M., Hayes, C. T., Heller, M. I., Henderson, G., Huang, K., Jeandel, C., Jenkins, W. J., John, S., Kenna, T. C., Klunder, M., Kretschmer, S., Kumamoto, Y., Laan, P., Labatut, M., Lacan, F., Lam, P. J., Lannuzel, D., le Moigne, F., Lechtenfeld, O. J., Lohan, M. C., Lu, Y., Masqué, P., McClain, C. R., Measures, C., Middag, R., Moffett, J., Navidad, A., Nishioka, J., Noble, A., Obata, H., Ohnemus, D. C., Owens, S., Planchon, F., Pradoux, C., Puigcorbé, V., Quay, P., Radic, A., Rehkämper, M., Remenyi, T., Rijkenberg, M. J. A., Rintoul, S., Robinson, L. F., Roeske, T., Rosenberg, M., van der Loeff, M. R., Ryabenko, E., et al.: The GEOTRACES intermediate data product 2014, Mar. Chem., 177, 1–8, doi:10.1016/j.marchem.2015.04.005, 2015.

McQuatters-Gollop, A., Edwards, M., Helaouët, P., Johns, D. G., Owens, N. J. P., Raitsos, D. E., Schroeder, D., Skinner, J. and Stern, R. F.: The Continuous Plankton Recorder survey: How can long-term phytoplankton datasets contribute to the assessment of Good Environmental Status?, Estuar. Coast. Shelf Sci., 162, 88–97, doi:10.1016/j.ecss.2015.05.010, 2015.

Menegotto, A. and Rangel, T. F.: Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness, Nat. Commun., 9(1), 4713, doi:10.1038/s41467-018-07217-7, 2018.

Meyer, C., Kreft, H., Guralnick, R. and Jetz, W.: Global priorities for an effective information basis of biodiversity distributions, Nat. Commun., 6(1), 8221, doi:10.1038/ncomms9221, 2015.

O'Brien, C. J., Peloquin, J. A., Vogt, M., Heinle, M., Gruber, N., Ajani, P., Andruleit, H., Arístegui, J., Beaufort, L., Estrada, M., Karentz, D., Kopczyńska, E., Lee, R., Poulton, a. J., Pritchard, T. and Widdicombe, C.: Global marine plankton functional type biomass distributions: coccolithophores, Earth Syst. Sci. Data, 5(2), 259–276, doi:10.5194/essd-5-259-2013, 2013.

O'Brien, C. J., Vogt, M. and Gruber, N.: Global coccolithophore diversity: Drivers and future change, Prog. Oceanogr., 140, 27–42, doi:10.1016/j.pocean.2015.10.003, 2016.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. and Ferrier, S.: Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data, Ecol. Appl., 19(1), 181–197, doi:10.1890/07-2153.1, 2009.

Provoost, P. and Bosch, S.: robis: R client for the OBIS API. R package version 0.1.5. https://cran.r-project.org/package=robis, 2015.

Le Quéré, C.: Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models, Glob. Change Biol., 11(11), 2016–2040, doi:10.1111/j.1365-2486.2005.01004.x, 2005.

Richardson, A. J., Walne, A. W., John, A. W. G., Jonas, T. D., Lindley, J. A., Sims, D. W., Stevens, D. and Witt, M.: Using continuous plankton recorder data, Prog. Oceanogr., 68(1), 27–74, doi:10.1016/j.pocean.2005.09.011, 2006.

Righetti, D., Vogt, M., Zimmermann, N. E. and Gruber, N.: PHYTOBASE: A global synthesis of open ocean phytoplankton occurrences, doi:10.1594/PANGAEA.904397, 2019a.

Righetti, D., Vogt, M., Gruber, N., Psomas, A. and Zimmermann, N. E.: Global pattern of phytoplankton diversity driven by temperature and environmental variability, Sci. Adv., 5(5), 10, doi:10.1126/sciadv.aau6253, 2019b.

Rodríguez-Ramos, T., Marañón, E. and Cermeño, P.: Marine nano- and microphytoplankton diversity: redrawing global patterns from sampling-standardized data, Glob. Ecol. Biogeogr., 24(5), 527–538, doi:10.1111/geb.12274, 2015.

Rombouts, I., Beaugrand, G., Ibañez, F., Gasparini, S., Chiba, S. and Legendre, L.: A multivariate approach to large-scale variation in marine planktonic copepod diversity and its environmental correlates, Limnol. Oceanogr., 55(5), 2219–2229, doi:10.4319/lo.2010.55.5.2219, 2010.

Sal, S., López-Urrutia, Á., Irigoien, X., Harbour, D. S. and Harris, R. P.: Marine microplankton diversity database, Ecology, 94(7), 1658, doi:10.1890/13-0236.1, 2013.

Ser-Giacomi, E., Zinger, L., Malviya, S., De Vargas, C., Karsenti, E., Bowler, C. and De Monte, S.: Ubiquitous abundance distribution of non-dominant plankton across the global ocean, Nat. Ecol. Evol., 2(8), 1243–1249, doi:10.1038/s41559-018-

0587-2, 2018.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M. and Herndl, G. J.: Microbial diversity in the deep sea and the underexplored "rare biosphere," Proc. Natl. Acad. Sci., 103(32), 12115–12120, doi:10.1073/pnas.0605127103, 2006.

Sournia, A., Chrdtiennot-Dinet, M.-J. and Ricard, M.: Marine phytoplankton: how many species in the world ocean?, J. Plankton Res., 13(5), 1093–1099, doi:10.1093/plankt/13.5.1093, 1991.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., D'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., Bork, P., Boss, E., Bowler, C., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E. G., Sardet, C., Sieracki, M. and Velayoudon, D.: Structure and function of the global ocean microbiome, Science, 348(6237), 1261359–1261359, doi:10.1126/science.1261359, 2015.

Thompson, G. G. and Withers, P. C.: Effect of species richness and relative abundance on the shape of the species accumulation curve, Austral Ecol., 28(4), 355–360, doi:10.1046/j.1442-9993.2003.01294.x, 2003.

Tittensor, D. P., Mora, C., Jetz, W., Lotze, H. K., Ricard, D., Berghe, E. Vanden and Worm, B.: Global patterns and predictors of marine biodiversity across taxa, Nature, 466(7310), 1098–1101, doi:10.1038/nature09329, 2010.

Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T. W., McNeill, J., Monro, A. M., Prado, J., Price, M. J. & Smith, G. F., editors. International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Regnum Vegetabile, Vol. 159. pp. [i]-xxxviii, 1-253. Glashütten: Koeltz Botanical Books, 2018. doi:10.12705/Code.2018.

Utermöhl, H.: Zur Vervollkommnung der quantitativen Phytoplankton-Methodik, SIL Commun. 1953-1996, 9(1), 1–38, doi:10.1080/05384680.1958.11904091, 1958.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Luke, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., Boss, E., Follows, M., Karp-Boss, L.,

Krzic, U., Reynaud, E. G., Sardet, C., Sullivan, M. B. and Velayoudon, D.: Eukaryotic plankton diversity in the sunlit ocean, Science, 348(6237), 1261605–1261605, doi:10.1126/science.1261605, 2015.

Villar, E., Farrant, G. K., Follows, M., Garczarek, L., Speich, S., Audic, S., Bittner, L., Blanke, B., Brum, J. R., Brunet, C., Casotti, R., Chase, A., Dolan, J. R., D'Ortenzio, F., Gattuso, J.-P., Grima, N., Guidi, L., Hill, C. N., Jahn, O., Jamet, J.-L., Le Goff, H., Lepoivre, C., Malviya, S., Pelletier, E., Romagnan, J.-B., Roux, S., Santini, S., Scalco, E., Schwenck, S. M., Tanaka, A., Testor, P., Vannier, T., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S. G., Bork, P., Boss, E., de Vargas, C., Gorsky, G., Ogata, H., Pesant, S., Sullivan, M. B., Sunagawa, S., Wincker, P., Karsenti, E., Bowler, C., Not, F., Hingamp, P. and Iudicone, D.: Environmental characteristics of Agulhas rings affect interocean plankton transport, Science, 348(6237), 1261447–1261447, doi:10.1126/science.1261447, 2015.

Vogt, M., O'Brien, C., Peloquin, J., Schoemann, V., Breton, E., Estrada, M., Gibson, J., Karentz, D., Van Leeuwe, M. A., Stefels, J., Widdicombe, C. and Peperzak, L.: Global marine plankton functional type biomass distributions: Phaeocystis spp., Earth Syst. Sci. Data, 4(1), 107–120, doi:10.5194/essd-4-107-2012, 2012.

Wallace, D. W. R.: Chapter 6.3 Storage and transport of excess CO2 in the oceans:The JGOFS/WOCE global CO2 survey, in Eos, Transactions American Geophysical Union, vol. 82, pp. 489–521., 2001.

Wickham, H. and Chang, W.: Devtools: Tools to make developing R packages easier. R package version 1.12.0. https://cran.r-project.org/package=devtools, 2015.

Woolley, S. N. C., Tittensor, D. P., Dunstan, P. K., Guillera-Arroita, G., Lahoz-Monfort, J. J., Wintle, B. A., Worm, B. and O'Hara, T. D.: Deep-sea diversity patterns are shaped by energy availability, Nature, 533(7603), 393–396, doi:10.1038/nature17937, 2016.

Worm, B., Sandow, M., Oschlies, A., Lotze, H. K. and Myers, R. a: Global patterns of predator diversity in the open oceans., Science, 309(5739), 1365–9, doi:10.1126/science.1113399, 2005.

Zimmermann, N. E. and Guisan, A.: Predictive habitat distribution models in ecology, Ecol. Modell., 135(2–3), 147–186, doi:10.1016/S0304-3800(00)00354-9, 2000.

Zweng, M. M., Reagan, J. R., Antonov, J. I., Locarini, R. A., Mishonov, A. V., Boyer, T. P., Garcia, H. E., Baranova, O. K., Johnson, D. R., Seidov, D. and Biddle, M. M.: World Ocean Atlas 2013, Volume 2: Salinity., S. Levitus, A. Mishonov, Eds. (NOAA Atlas NESDIS 74, 2013), 39 pp., 2013.