Response to Reviews, 12/02/2020

Summary by Righetti et al. (DR)

We thank the three reviewers for their constructive comments, which have provided the basis to increase the quality, reproducibility, and accuracy of the database and manuscript. In essence, reviewer 1 advised us to implement minor corrections regarding table 1 and 2. Reviewer 2 suggested minor changes with respect to 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 the reviewer's points in detail. Red markings indicate textual edits that have been implemented in a revised version of the manuscript.

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 them 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 phytoplankton 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 313: We add a grey bar to each panel in figure 3, indicating probability density.

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: 790'224 data points for 1498 species, with 54.9% of points being unique to this source; OBIS: 823'861 data points for 1325 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 sources in the revised version of the manuscript: Line 100ff: *"To create PhytoBase, we compiled marine phytoplankton occurrences 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 species distribution evidence, striving for a global synthesis of data. 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 have not been included in GBIF or OBIS at the time of query (closing window, March 2017). While our selection of additional sources was not exhaustive, it strived for the inclusion of quality controlled, large-scale plankton 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 also specify the subsequent section: 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 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)."*

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; <0.0001% of raw data), (ii) if they had no indication on the year or month of collection (excluding 7.23% GBIF raw data and 0.87% 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 6.55% 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 reconciled with Darwin Core, we present additional columns (e.g., columns ending in "gbif" relate to GBIF sourced data)."*
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 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, 18.57% 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, 21.64% contained a "depthAccuracy", and only marginally more records contained a MinimumDepthInMeter (25.72%) or minimumDepthInMeter (23.99%). 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; 18.57% of records with entries) while OBIS raw* *data contained the fields "depthprecision" (21.64% of data with entries), "minimumDepthInMeters" (25.72% of data with entries) and "maximumDepthInMeters" (23.99% 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 occurrenceID: 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 add a note to Table 1 on the higher-rank taxonomic nomenclature: Line 198: *"†Nomenclature (phylum, class) follows OBIS backbone taxonomy (retrieved 6 March 2017; using R packages RPostgreSQL and devtools), which extensively relies on the World Register of Marine Species (www.marinespecies.org)".*

- 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 were uniquely provided by MareDat, while additional count data on trichomes for the genus Trichodesmium may be accessed 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 flags OBIS 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 names associated with each record in separate columns specific to each data source (i.e.,*

"scientificNameOriginal_<source>"), which allows tracing back the harmonized name to its original name(s). Retaining these original names ensures that any taxonomic changes or updated methods for taxonomic harmonization can be readily implemented in the future."

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 also 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 marine database. However, we retained the following species (…)."

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., folder "download_and_prepare_data" and folder "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 original research cruises and survey programs. While small or fragile species may escape detection by the CPR program (Richardson et al., 2006), the resolution of seawater samples is heavily influenced by sampling volume and taxonomic expertise (Cermeño et al., 2014). We have shown that the average number of species detected per sampling event ranges from three to above 50 between cruises. Global spatiotemporal biases have been similarly present in data collections of heterotrophic marine taxa (Menegotto & Rangel 2018), but sampling resolution biases and divergent sampling protocols may be even more common for the phytoplankton."* In accordance with this changes, we adjust line 499: "(...)be additionally implemented to overcome data limitations. The latter statistical (...)."

Finally, we highlight the benefit of integrating molecular data, in line with the point by RE3:
Line 508ff: *"The detection of rare species and their integration into PhytoBase may become possible via molecular methods, including metagenomic approaches (Bork et al., 2015; Sogin et al., 2006). DNA (...)"*

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: *“We thus recommend careful screening of the metadata of occurrence records, retrievable via the record specific information (e.g. datasetKeys for GBIF records, resourceIDs for OBIS records), to reduce biases in biogeographic characterizations of species.”*

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 phytoplankton data across several methodological or taxonomic dimensions."*

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 marine phytoplankton taxa recorded as autotrophs (de Vargas et al., 2015; Falkowski et al., 2004), but it is clear that some of these species may be mixotrophic, particularly for the Dinophyceae (de Vargas et al., 2015)."*