A global diatom database – abundance, biovolume and biomass in the world ocean

Leblanc[1] K., Arístegui[2] J., Armand[3] L., Assmy[4] P., Beker[5] B., Bode[6] A ., Breton[7abc] E., Cornet[1] V., Gibson[8] J., Gosselin[9] M.-P., Kopczynska[10] E., Marshall[11] H., Peloquin[12] J., Piontkovski[13] S., Poulton[14] A.J., Quéguiner[1] B., Schiebel[15] R., Shipe[16] R., Stefels[17] J., van Leeuwe[17] M.A., Varela[6] M., Widdicombe[18] C., Yallop[19], M.

[1]{Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU, IRD, MIO, UM 110, 13288, Marseille, Cedex 09, France.}
[2]{Instituto de Oceanografía y Cambio Global. Universidad de Las Palmas de Gran Canaria. 35017 Las Palmas. Spain}
[3]{Department of Biological Sciences, Macquarie University, North Ryde, NSW, 2109, Australia}
[4]{Norwegian Polar Institute, Fram Centre, Hjalmar Johansens gt. 14, 9296 Tromsø, Norway}
[5]{Laboratoire des Sciences de l'Environnement Marin (UMR CNRS 6539), Institut Universitaire Européen de la Mer (IUEM), Place Nicolas Copernic, Technopôle Brest Iroise, 29280 Plouzané, France}
[6]{Instituto Español de Oceanografía, Centro Oceanográfico de A Coruña Apdo. 130, E-15080, A Coruña, Spain}
[7]{[a]Univ Lille Nord de France, F-59000 Lille; [b]ULCO, LOG, F-62930 Wimereux; [c]CNRS, UMR 8187 LOG, F-62930 Wimereux, France.}
[8]{Institute for Marine and Antarctic Studies, University of Tasmania, Private Bag 129, Hobart, Tasmania 7001, Australia}
[9]{The Freshwater Biological Association, The Ferry Landing, Far Sawrey, Ambleside, LA22 0LP, U.K.}
[10]{Institute of Biochemistry and Biophysics, Department of Antarctic Biology, Polish Academy of Sciences, 02-141 Warszawa, Poland}
[11]{Department of Biological Sciences, Old Dominion University, Norfolk, VA, USA}
[12]{Inst.f.Biogeochemie u.Schadstoffdynamik, Universitätstrasse 16, 8092 Zürich, Switzerland}
[13]{Department of Marine Sciences, Sultan Qaboos University, Sultanate of Oman}
[14]{National Oceanography Centre, Waterfront Campus, Southampton, SO14 3ZH, U.K.}
[15]{Laboratoire des Bio-Indicateurs Actuels et Fossiles (BIAF), UPRES EA 2644, Université d'Angers,49045 Angers CEDEX 01, France}
[16]{UCLA, Los Angeles, California 90095, USA}
[17]{University of Groningen, Centre for Life Sciences Ecophysiology of Plants, PO Box 11103, 9700 CC Groningen, the Netherlands}
[18]{Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth, PL1 3DH, United Kingdom}
[19]{School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG,U.K.}

Correspondance to: K. Leblanc (karine.leblanc@univ-amu.fr)

Abstract

Phytoplankton identification and abundance data are now commonly feeding plankton distribution databases worldwide. This study is a first attempt to compile the largest possible body of data available from different databases as well as from individual published or unpublished datasets regarding diatom distribution in the world ocean. The data obtained originate from time series studies as well as spatial studies. This effort is supported by the Marine Ecosystem Model Inter-Comparison Project (MAREMIP), which aims at building consistent datasets for the main Plankton Functional Types (PFT) in order to help validate biogeochemical ocean models by using carbon (C) biomass derived from abundance data. In this study we collected over 293 000 individual geo-referenced data points with diatom abundances from bottle and net sampling. Sampling site distribution was not homogeneous, with 58% of data in the Atlantic, 20% in the Arctic, 12% in the Pacific, 8% in the Indian and 1% in the Southern Ocean. A total of 136 different genera and 607 different species were identified after spell checking and name correction. Only a small fraction of these data were also documented for biovolumes and an even smaller fraction was converted to C biomass. As it is virtually impossible to reconstruct everyone's method for biovolume calculation, which is usually not indicated in the datasets, we decided to undertake the effort to document, for every distinct species, the minimum and maximum cell dimensions, and to convert all the available abundance data into biovolumes and C biomass using a single standardized method. Statistical correction of the database was also adopted to exclude potential outliers and suspicious data points. The final database contains 90 648 data points with converted C biomass. Diatom C biomass calculated from cell sizes spans over eight orders of magnitude. The mean diatom biomass for individual locations, dates and depths is 141.19 µg C L^{-1}, while the median value is 11.16 µg C L^{-1}. Regarding biomass distribution, 19% of data are in the range 0-1 µg C L^{-1}, 29% in the range 1-10 µg C L^{-1}, 31 % in the range 10-100 µg C L^{-1}, 18% in the range 100-1 000 µg C L^{-1}, and only 3% >1 000 µg C L^{-1}. Interestingly, less than 50 species contributed to >90% of global biomass, among which centric species were dominant. Thus, placing significant efforts on cell size measurements, process studies and C quota calculations on these species should considerably improve biomass estimates in the upcoming years. A first-order estimate of the diatom biomass for the global ocean ranges from 444 to 582 Tg C, which converts to 3 to 4 Tmol Si and to an average Si biomass turnover rate of 0.15 to 0.19 d^{-1}.

Link to the dataset : http://doi.pangaea.de/10.1594/PANGAEA.777384.

1. Introduction

Marine ecosystems are characterized by large species diversity, yet the succession and distribution of the main taxa are still poorly understood. Plankton diversity is often narrowed down to the notion of functional group, which can be defined as a group of organisms operating the same biogeochemical process and driving the flux of the main biogenic elements differently from other groups. Functional groups have been further organized into Plankton Functional Types (PFT) (Le Quéré et al., 2005; Hood et al., 2006), in order to help construct biogeochemical models including diversity in a simplified way. Main PFT include diatoms, calcifying organisms, nitrogen fixers, pico-autotrophs, pico-heterotrophs and various zooplankton groups. Diatoms are a large component of marine biomass and produce ~25% of the total C fixed on Earth (Nelson et al., 1995; Field et al., 1998), producing more organic C than all rainforests combined. Another striking image to consider is that they produce one fifth of the oxygen we breathe. Therefore they have a major ecological significance and impact on the global elemental Si and C cycles (Tréguer et al., 1995; Ragueneau et al., 2000; Tréguer, 2002; Jin et al., 2006). Diatoms also have a high export/production ratio due to elevated sedimentation rates by forming aggregates and incorporation into fast sinking zooplankton faeces. Diatoms are, along with dinoflagellates, today's most diverse planktonic flora. A current estimate of all living diatoms ranges from 10 000 to 100 000 species, but a smaller fraction, from 1 400 to 1 800 species, are recognized as marine planktonic (Sournia et al., 1991). Major progress has been made in the last decades on in situ Si dynamics, thereby improving models, but the knowledge of biological factors such as species composition, cell morphology and aggregation processes still needs to be improved (Hood et al., 2006).

Satellite data now allow a closer definition of functional groups from space (Alvain et al., 2005; Uitz et al., 2006), and this effort has been most fruitful on coccolithophores (Yoder and Brown, 1994; Iglesias-Rodriguez et al. 2002) but has also been recently attempted on *Trichodesmium* (Dupouy et al., 2008) and diatoms (Sathyendranath et al., 2004). However many challenges remain with this approach, a major bias being the impossibility to capture subsurface blooms but also to assess variable cellular pigment quotas. Hence, Dynamic Green Ocean Models (DGOM) still need validating with datasets giving C biomass estimates for each PFT. Improving

the parameterization for diatoms in various biogeochemical models would thus help improve the global C budget and the subsequent fate of exported particulate matter with respect to depth estimations.

Phytoplankton identification and abundance data are now regularly added to plankton databases worldwide but need to be regrouped so that they can be useful to the biogeochemistry and modeling community. This study is the first attempt to compile the largest possible body of available data from these different databases as well as from individual datasets regarding diatom distribution in the world ocean. This study is supported by the MAREMIP program, which aims at building consistent datasets for the major PFT in order to provide validation sets for biogeochemical ocean models. This paper is part of the special issue dedicated to providing global databases (named Marine Ecosystem Data - MAREDAT) on the nine main PFT for their abundance and C biomass.

Diatom cell sizes range from a few micrometers up to 2 millimeters and their cellular biovolumes span over nine orders of magnitude. Subsequent C conversion estimates are therefore prone to large errors if cell size is not correctly assessed. The challenge posed by compiling a global database on diatom abundance, biovolume and biomass is the large intraspecific variability observed in diverse parts of the world ocean and in the same area depending on environmental conditions and life-stages.

Plankton identification and counting is sometimes rewarding, but is most often considered a tedious task, one that cannot be completed “without ruin of the body and mind” as Haeckel (1890) humorously phrased it. Systematic cell size measurements, biovolume and biomass conversion are even more challenging. An additional objective of this study is to provide a tool for taxonomists worldwide to facilitate these measurements and calculation in a standardized way during routine cell counts.

The objective of this study is to promote the construction of an extensive diatom database with standardized methods for collection, counting, data management and conversion to biomass used to assess the global importance of diatoms in marine productivity and provide field data for biogeochemical models including PFT. An extensive bibliographic search was undertaken to compile all available diatom dimensions for all reported species. This will allow a first estimation of the contribution of diatoms to the global C budgets based on field data. A quantitative and qualitative description of the main features of diatom biomass distribution is presented in the

following study. This effort has been initiated in the PANGAEA database, where individual collections are available, but should be the object of supplementary effort to systematically include cell sizes in a standardized way (see material and methods section) in future studies.

2. Methods

2.1. Data collection

Data were collected through a first round of mail enquiries addressed to an extensive list of taxonomists. A second round of enquiries was sent to the administrators of the main known databases (PANGAEA, BODC, NODC, NMSF-Copepod…) for access to their datasets. Finally, recent oceanographic cruises or research programs or time-series that were known to include taxonomic data were identified and permission for use in the present database was acquired from each owner. The entries for each data point included date of collection, sampling depth, latitude, longitude, taxonomic information, abundance with unit and if possible, sampling, preservation and counting methods. The latter information was most difficult to obtain for old datasets where the contact person could not be identified or had retired.

We collected over 293 000 individual geo-referenced data points with diatom abundances mostly from bottle sampling (Niskin, Hansen or other appropriate bottle sampling device). A very small fraction of the database included net hauls or Continuous Plankton Recorder (CPR) data, which were excluded from the present database as it is quite difficult to reconstruct quantitative cellular concentrations from them and because of their bias towards collecting larger cells. After filtering out zero abundance data, net haul data, erroneous data and after statistical treatment (see section 2.4), 91 704 data points with associated cell abundance remained, 90 648 of which were converted to C biomass. A total of 607 different taxonomic species and 136 different genera were identified after spell checking and taxonomic nomenclatural verification. The entire data treatment process is described in the flow diagram in Fig. 1.

2.2. Biomass conversion procedure

Measured cell sizes are rarely or vaguely indicated in phytoplankton databases. Clearly, more effort is needed on building accurate taxonomic databases with associated species size range for each oceanic and coastal region. In order to reconstruct each species cell size, one

option is to consider the minimum and maximum dimensions of each species and derive minimum, maximum and average biovolumes and associated C biomass. Such efforts have for instance been successfully undertaken in the Baltic Sea by the HELCOM Phytoplankton Expert Group (PEG), and resulted in a report compiling a complete list of species with their measured dimensions and biovolumes (Olenina et al., 2006). In this study, the authors put an emphasis on the « hidden dimension » of cells, as some algal dimensions are seldom visible in the microscope during routine cell counts and hence are almost never documented. This is typically the case for the pervalvar axis of many diatoms, which most often lie on their valve face after sedimentation on a glass slide. In most cases assumptions are made regarding this hidden dimension (an example for an assumption can be pervalvar axis = 1/3 of the apical axis) but this information is mostly absent from taxonomic guides, which give at best one or two of the cell dimensions. Hence, further attentiveness is required to document consistent ratios between visible and hidden dimensions for the main diatom species.

In the last decade, a couple of significant studies (Hillebrand et al., 1999; Sun and Liu, 2003) have produced detailed guides of biovolume calculations for phytoplankton species taking into account the variety and complexity of the numerous diatom shapes by assimilating them into standardized geometric models (19 different shapes were used for this study), which should help harmonize biovolume calculations considerably. As it is not possible to measure every cell's dimensions in one sample, it is usually recommended to measure all dimensions for 25 cells of each species and use the mean value of the obtained cell volume for all occurrences of the same species, although in most cases the standard error in mean biovolume calculation is <5% after the measurements of 10 cells (Sun and Liu, 2003). However, Hillebrand et al. (1999) emphasized that seasonal, inter annual, spatial and life cycle variations render it inaccurate to use average biovolume data of species throughout the year. Therefore, strict quality standards imply that biovolume should be calculated for each subset of samples, sometimes including different sampling depths of the same water body (Hillebrand et al., 1999).

2.3. Data file content

The data file consists of an excel file containing several spreadsheets. A spreadsheet named "dimension-biovolume-biomass" lists all the different name entries, with their corrected names, and associated World Register of Marine Species code (WoRMS, http://www.marinespecies.org).

In total, 1 364 different taxonomic entries were found, but were reduced to 727 different taxonomic lines after name correction. The original entry and its associated correction following WoRMS are indicated in two different columns. Up to 607 WoRMS species codes were attributed, but 24 entries were not found in the WoRMS register and labeled ‘nf1’ to ‘nf24’. Entry lines were also tagged with a “C” for centrics, “P” for pennates and “U” for unidentified diatoms (this last group was not converted to C biomass because of the large uncertainty on cell size). In most instances, taxonomic entries were not associated with cell size measurements. On other occasions, biovolume measurements were provided but lacked corresponding cell size data. Hence, it was virtually impossible to reconstruct each individual calculation method employed for estimating biovolume, when this was often not indicated in the datasets. Keeping the original published biovolumes would almost certainly have introduced a bias between different datasets. We therefore chose to exclude such data, and have documented instead, for every distinct species, the minimum, average and maximum known cell dimensions. The dimensions extracted from the literature were then used to convert all the available abundance data into biovolumes and C biomass using a single standardized method. Each species is allocated one of the 19 possible diatom shapes identified in Sun and Liu (2003) in order to derive the biovolume (V) and surface area (S) calculation formulas. The figures for the different shapes and formulas extracted from Sun and Liu (2003) are shown in another spreadsheet “diatom shapes” for a quick visual check of the diatom cell shapes. In the spreadsheet “dimension-biovolume-biomass”, the known minimum and maximum dimensions for each species are indicated. In the column “other info”, the taxonomist’s original observations regarding size are indicated, but most often refers to a unique value – the largest dimension or diameter of the cell. When indications of cell size are given, minimum and maximum dimensions columns are amended to fit the observations (indicated by a yellow color). The bibliographical references used to find dimensions for each species are indicated for each entry as a number, which refers to the “reference” spreadsheet, where full references are given. Dimensions written in black correspond to referenced measurements; dimensions written in red refer to a value deduced from illustrations or drawings when a scale bar was present, showing a ratio between two different axes of the cells. Cells labeled in pink indicate that an assumption was made on the ratio between one of the known dimensions and the hidden dimension. The assumption made is always explicitly indicated in another column - for instance for some *Coscinodiscus* species pervalvar axis =1/3 diameter. Minimum and maximum

biovolume, surface area and S/V ratios are calculated for every single entry depending on the given dimensions. The cellular biovolumes ranged from 3 μm^3 (*Thalassiosira* sp.) to 4.71 x 10^9 μm^3 (*Ethmodiscus* sp.). The total biovolume obtained was then converted to C biomass similarly to the method used in Cornet-Barthaux et al. (2007) using the equation of Eppley et al. (1970) corrected by UNESCO (1974) and Smayda (1978):

$$\log_{10}C\ (pg) = 0.76 \log [\text{cell volume } (\mu m^3)] - 0.352$$

The spreadsheet "diatom database" is the actual diatom compiled database with the complete information regarding date, location, depth, methods, and taxonomic information. Each line starts with a unique primary key indicator which enables rapid restoration back to the original data file in the event that database sorting or filter commands are used for further computations. Biovolume, surface area, and cellular C content are automatically retrieved from the previous spreadsheet based on the recognition of the original name entry. Abundance data are standardized to one unit (cells L^{-1}) and multiplied with C content per cell (pg $cell^{-1}$) to derive total C biomass (converted to $\mu g\ C\ L^{-1}$). Minimum, maximum and average data of size, biovolume and biomass are indicated in the file, however in this paper, generally averaged data estimates for biomass will be used in discussion.

2.4. Quality control

A first run through the database was done to check for all spelling errors and invalid data entries. Suspicious data, for which the abundance values or units were not clear were systematically discarded. A statistical treatment, using Chauvenet's criterion test, was then applied to the database to filter out potential outliers. Only 151 data were identified as outliers using this criterion, and they all corresponded to entry lines with "unidentified diatom species" or "diatom spp.". This is not surprising, as the biomass conversion used in this case is the average between the minimum and maximum biomass found for all diatoms, and logically leads to very spurious biomass values (usually overestimating, probably because unidentified cells are mostly of small sizes). After correcting the database by excluding these outliers, a few average biomass values remained conspicuously elevated. On investigation, they were found to correspond to "unidentified diatom species" or "diatom spp." lines. Therefore, we chose to discard the biovolume calculations for all these entry lines ("U") because the assumptions made on their

biovolume were too imprecise, nevertheless the abundance data from these locations were kept, in order to preserve the 1 056 relevant data points.

3. Results

3.1. Spatial distribution of data

The database contains 91 704 individual lines (90 648 with converted biomass). There are 9 930 unique locations, time and depth points (but with multiple species entries) and 2 971 unique location and time points (all depths combined). Regarding the spatial distribution of data, the oceanic regions best represented included the North Atlantic, the North Indian, Equatorial Atlantic, Arctic, Antarctic and North Pacific areas (Fig.2). Indonesia, the Gulf of Mexico & Caribbean, the South Pacific, South Atlantic and South Indian are less well covered. This does not mean that samples were not collected and counted, but simply that the data have not been released for public use by their owner or have remained the property of a given government. The largest number of observations was reported in the northern hemisphere (NH) between the Equator and 70°N (Fig. 3a). Table 1 shows that the distribution of biomass data, according to latitudinal bands, is clearly skewed towards the mid-northern hemisphere with 43.9% of data between 40° and 60°N.

3.2. Temporal distribution of data

Most observations were commenced in the 1970s, but a few datasets date as far back as 1933-1934 and 1954-1956 (Fig. 3b). As expected, data frequency diminishes after 2000, as newer data need to be published by the relevant PIs before being submitted to databases, a process that usually occurs a few years after the end of a research program. Data were mostly obtained during boreal spring and autumn (37% in March, April and November), while the boreal winter months were less well covered (11% in December, January and February).

3.3. Global abundance characteristics

Diatom abundances ranged from 1 to 6.95 x 10^7 cells L^{-1}. The highest abundances reported in the database, representing massive blooms (>10 millions cells L^{-1}) were found in Antarctica in the Ross Sea in December 2004 and January 2005, and at the Antarctic Davis station in January

1995. These occurrences are represented by *Chaetoceros socialis* blooms, *Thalassiosira* spp. and unidentified pennates. Abundances of up to several million cells L^{-1} were also reported in a coastal area during the Galicia program off NW Spain (again identified as *Chaetoceros socialis*). The smallest abundance values were reported for the Indian Ocean and the Mediterranean Sea. The average diatom cell abundance for each time, location and depth was 263 099 cells L^{-1} and the median value was 7 056 cells L^{-1}.

3.4. Global biomass characteristics

Diatom C biomass calculated from cell sizes span over eight orders of magnitude (Fig. 4). The mean diatom biomass for the entire database is 141.19 µg C L^{-1}, while the median value is 11.16 µg C L^{-1}. The mean diatom biomass for the NH is 141.22 µg C L^{-1} (median 12.60 µg C L^{-1}) and 141.27 µg C L^{-1} (median 4.67 µg C L^{-1}) for the Southern Hemisphere (SH). For the whole database, 19% of biomass data are in the range 0-1 µg C L^{-1}, 29% in the range 1-10 µg C L^{-1}, 31 % in the range 10-100 µg C L^{-1}, 18% in the range 100-1 000 µg C L^{-1}, and only 3% >1 000 µg C L^{-1}.

The maximum biomass in the NH (12 299 µg C L^{-1}) was reported off the coast of NW Spain (43.42°N-8.43°E) at the surface in July 1990 (Fig. 5a). The biomass maximum was associated to a bloom of *Dactyliosolen fragilissimus* and *Chaetoceros* spp. The maximum biomass in the SH (11 174 µg C L^{-1}) was observed in the Peruvian upwelling region in March 1974. Here, the surface water bloom was comprised of *Dactyliosolen fragilissimus*, *Leptocylindrus danicus* and *Guinardia delicatula*.

The biomass uncertainty was calculated as a percentage of the difference between the maximum biomass and minimum biomass normalized to the mean biomass (Fig. 5b). The biomass uncertainty comprised between 100 and 200% of the average biomass for 96% of the data, and between 0 and 100% for the remaining 4% of data. Uncertainty is strongly sensitive to cell size, and therefore diatom species that span wide size ranges provide the least precise estimates. Only the accurate determination of cell sizes for each species and for each program, location, date and depth will significantly improve this bias.

3.5. Latitudinal and depth distribution of biomass estimates

The vast majority of biomass estimates were collected in the 0-100 m layer (Fig. 6a), which is well covered in terms of vertical resolution, while deeper estimates are mostly found at fixed depths below 100 m (150, 200 m) and are more scarce.

The largest range of biomass estimates corresponds to the latitudinal bands most often sampled, between 40° and 60°N (Fig. 6b). Estimates are scant in the SH, but all latitudes are reasonably well covered. There is no clear tendency towards lower or higher biomass according to latitude, except potentially in the Arctic where the range of variation seems to be lower than elsewhere.

3.6. Seasonal distribution

There are no clear seasonal trends in the monthly distribution of biomass estimates in the NH (Fig.7a). The largest range of estimates is observed in June and the lowest in November, but wide amplitude of variation is observed almost for every month. Seasonality seems a bit more marked for the SH, with the lowest range of variations observed between June and September and the highest range between November and March (Fig. 7b). This weak display of seasonality probably originates from the fact that a mix of warm and cold waters, eutrophic and oligotrophic areas are represented in both hemispheres.

3.7. Dominant genera and species

Biomass data for all identical taxonomic entries were summed for the entire database, for either genera (Fig. 8) or for individual species (Fig. 9). Out of the 136 identified genera in the database, 32 genera represent 99 % of the total estimated biomass. A boxplot of estimated averaged biomass for all 32 genera is shown in Fig. 8. The median values for all individual genera roughly range between 0.1 and 10 µg C L^{-1}. Taking into account the 5^{th} and 95^{th} percentiles, average biomass ranges between 0.002 µg C L^{-1} and 826 µg C L^{-1}. The largest range of biomass is found for the genus *Thalassiosira* and the narrowest for *Paralia.* The percentage contribution of each genus ranked by decreasing order of importance is reported in Table 2. The dominant genus in the database is *Rhizosolenia*, representing 17.4% of the total diatom biomass, followed by *Chaetoceros* (14.5%) and *Thalassiosira* (12.6%). Unidentified pennate and centric

diatoms were included in the calculation, and if determined down to genus would inevitably change the relative order of the dominant genera, as they represent 8.2 and 6.6% of the total biomass, respectively. The other important genera are *Dactyliosolen* (7.6%) and *Guinardia* (7.3%). Centric diatoms are by far the largest contributors to total biomass (86%) and the cylindrical shape is dominant overall.

A second boxplot figure is presented in Fig. 9 with the same calculations as in the preceeding Fig. 8, but using only the taxonomic entries that were identified down to the species level and excluding all other undetermined species (e.g. *Chaeotoceros* spp.). Out of the 552 identified species (which may be reduced to a slightly smaller number after elimination of all synonyms in the database), only 43 species contribute 90 % of the total diatom biomass for identified species (47.5% of the total biomass in the database including all undifferentiated taxa). The median value for these dominant species ranges roughly from 0.1 to 10 µg C L^{-1}. When extending to the 5^{th} and 95^{th} percentiles, biomass data range from 0.002 µg C L^{-1} to 439 µg C L^{-1}. The largest range of biomass is found for *Rhizosolenia imbricata* and the narrowest for *Coscinodiscus wailesii*. The percentage contribution of each species ranked by decreasing order of importance is reported in Table 3. The predominant species, contributing up to 19% of total biomass (excluding all unidentified species data) were *Dactyliosolen fragilissimus* (13.6%), *Rhizosolenia imbricata* (10.8%) and *Guinardia striata* (8.2%). The *Rhizosolenia* species in this list (6/43) alone represent 20.8% of total biomass (identified to the species level). The seven major *Chaetoceros* species combined represent 6.1% of biomass. The most dominant *Chaetoceros* species in terms of average total biomass was found to be *Chaetoceros socialis* (2.6%) followed by *Chaetoceros compressus* (1.6%). Again the dominant species contributing to the average total biomass overall were principally represented by centric diatom species.

4. Discussion

This study is the first effort to compile robust global biomass estimates for marine diatoms. A summary boxplot diagram (Fig. 10) shows that 78% of the data (without consideration of taxa) range between 0.01 and 100 µg C L^{-1} for the average diatom biomass estimates per depth. However, there remain numerous biases in the present database that require resolution, before an accurate diatom biomass dataset can be fully realised in the future. We have identified several

major biases from this compilation and acknowledge that resolving them at this point in time is beyond the scope of this paper. These biases are:

1. If the temporal distribution seems to be well covered (Fig. 7), the spatial coverage is still inhomogeneous (Fig. 2) and vast parts of the ocean (in particular the SH) remain under sampled and/or the data remain inaccessible.

2. Blooming/productive areas are often better investigated than oceanic deserts, and when programs do occur in oligotrophic regions, researchers can often refrain from running accurate cell counts when the abundance of a group is very low. Figures 8 and 9 show that for individual genera or species the distribution of data around the median values are mostly skewed towards the higher biomasses. Such a feature indicates cell abundances have been assessed more thoroughly when cells are abundant. Similarly, large cells are more easily identified in light microscopy than smaller cells (typically <10-20 µm).

3. Most cell counts are run on fixed samples, and even if diatoms are usually not considered to be impacted by preservatives, there is some evidence that diatoms do shrink or swell with Lugol's solution, sometimes by up to 30%, depending on its final concentration in the sample (Montagnes et al., 1994; Menden-Deuer et al., 2001). However these studies were carried out on a small number of diatom species, and more work is needed to determine the accurate effect of Lugol's preservation on cell size and biovolume measurements.

4. The biovolume used to convert μm^3 into pg C $cell^{-1}$ is calculated from the frustule outer dimensions, which do not necessarily match that of the cytoplasm. The latter can be, depending on the species, considerably smaller than the frustule itself. This issue can only be resolved by culture work to determine cellular C content on the main identified species. The impact of this issue means all C biomass estimates must be considered as overestimates and a maximum value per genus or species.

5. Cells change size through their life cycle, season, depth and it is therefore inadequate to use average values for cell size, and subsequently for biovolume and carbon biomass calculations. Cell sizes should be measured systematically (for the dominant species) between subsamples and between different areas. This could not be done in the database, where minimum and maximum ranges for each species were considered, and distinction in sizes according to the geographic area could not be taken into account. According to Viličić (1985) the use of literature data from other

oceanic regions should be avoided and measuring cell dimensions for each dataset is the only way to estimate the total cell volume without major error.

6. Regarding the average cell size, Hillebrand et al. (1999) further stated that the biovolume should be calculated from the median of measured linear dimensions, not as a mean (or median) of a set of individually calculated biovolumes. Here, we were not able to calculate median dimensions for lack of data on cell size measurements, so we decided to use the average biovolume calculated from the literature minimum and maximum dimensions, but we acknowledge that this is a rough approximation.

7. In most cases, the hidden dimension of diatoms is not indicated, and cannot be obtained without further manipulation of the cells on glass slides using needles, a task that can be daunting to most people. In this study, assumptions were made on the hidden dimension using ratios between for instance the diameter and pervalvar axis for centric diatoms. Clearly, more attention needs to be given to these calculations, and this hidden dimension should be better indicated in taxonomic guides.

8. The cellular carbon content is assumed to be constant and a function of cell volume. However, it is known that depending on growth conditions (irradiance, temperature, nutrients), a degree of plasticity in the cellular C content can be achieved (Finenko et al., 2003). Applying the same conversion factor over a wide size range, as is the case for diatoms, leads to systematic errors and this formulation should also be improved (Menden-Deuer and Lessard, 2000).

These biases are well established and acknowledged in modern treatments of biovolume and biomass estimates (e.g. Cornet-Barthaux et al., 2007) yet nevertheless remain challenging. Substantial progress could be achieved by placing more efforts on the globally dominant species. This database allows the first estimate of the relative contribution of the main diatom genera and species to global biomass, and reveals that a small number of them (<50) represent between 90 and 99 % of the biomass. Improving size and biovolume determinations on these particular species, as well as according to geographical area, season and life cycle should thus substantially improve diatom biomass estimates. Guillard and Kilham (1978) published an extensive description of the diatom flora for the main biogeographical provinces, which similarly showed that only a few dozen species were dominant in each province. At a coastal site in the Gulf of Lions (North Western Mediterranean Sea), a bimonthly survey over 11 years showed that out of

the 91 diatom species that were identified, only 16 species represented 97 % of the combined cell abundances. Incidentally, 10 of these 16 species also appear in the top 50 species identified in Fig. 9. We, therefore, advocate the systematic use of regional atlases reporting full description of cell sizes and biovolume ranges for the dominant species present, which are usually much less numerous than the full extent of diatom diversity. Focusing on improving biomass estimates for the most abundant species identified here should be an achievable task within the next few years, and should considerably improve global diatom biomass estimates. This list of dominant species should of course not be considered as a static unchanging list, as climate change and environmental modifications are highly susceptible to change the order of species dominance in the ocean. However some species identified here as globally important are seldom the object of laboratory culture work and little is known of their physiology and biogeochemical characteristics.

This study, together with the other datasets compiled for the main Planktonic Functional Types, should allow a first comparison of a PFT's relative importance, as well as an estimation of the global heterotrophic to autotrophic planktonic biomass ratio. Looking at coastal and open ocean data separately should also allow for the validation or otherwise of the trophic chain pyramid models proposed by Gasol et al. (1997). By compiling simultaneous reports for most planktonic groups (phytoplankton, bacteria, mesozooplankton and heterotrophic protists) from the literature and in various environments, Gasol et al. (1997) showed that the heterotrophic:autotrophic biomass ratio was higher in open ocean/less productive systems, indicating an inverted biomass pyramid, while coastal/productive areas were characterized by a smaller contribution of heterotrophs relative to autotrophs. According to the authors, these differences reflect consumer-controlled systems in the first case, and resource-controlled systems in the latter. The different databases compiled in this special issue could be used to run such comparisons (see also Buitenhuis et al., introductory paper in this issue).

Despite the identified biases, the biovolume data compiled in this study are in the same order of magnitude as the literature data. Considering a global integration depth of 100 m as a rough estimate for the euphotic zone depth, diatom biomass data are mostly comprised between 0.01 and 10 g C m^{-2}, which is in the same order of magnitude as the total autotrophic plankton biomass (diatoms + other groups) by Gasol et al. (1997), which ranged between 0.02 and 31.8 g C m^{-2}. However, a more extensive comparison with the literature remains difficult because global

estimates derived from satellite products are most often given in chlorophyll *a* concentrations or as net primary production.

Finally, we present an attempt at a first-order estimate of the global diatom biomass (Table 4 and 5). Following the method described in Luo et al. (this issue), depth-integrated biomass values (a minimum of three depths were required for the calculation) were binned to 3° x 3° grid to partially smooth out the uneven spatial distribution of data. The total area of the five main oceans was multiplied by the geometric or arithmetic means of diatom biomass for each ocean. The geometric mean is considered preferentially for this calculation as it is the exact representation of the mean for log-normal distributed data. The dataset was furthermore sorted out between coastal (defined here as bathymetry <100 m) and open ocean data, representing 552 and 3826 different sites respectively. The binning procedure is inadequate to use on coastal data only (too little spatial coverage), hence the calculations were run on the entire dataset first (Table 4), then on open ocean data alone (Table 5), the difference reflecting the weight of coastal data. Considering either 100 or 200 m as the depth of integration yields diatom biomass values for the global ocean using all data of 488-470 Tg C (geometric mean) and 2942-3023 Tg C (arithmetic mean) respectively. These values vary slightly considering open ocean data alone (Table 5) and amount to 582-444 Tg C (geometric mean) and 3636-3433 Tg C respectively (arithmetic mean). After conversion to Si biomass using a Si:C ratio of 0.093, as the average between Si-stressed diatoms (0.056, DeLaRocha et al., 2010) and Si-replete diatoms (0.130, Brzezinski et al., 2011), the global Si budget for diatom biomass amounts to 3.6-3.8 Tmol Si for the global ocean (Table 4) and 3.4-4.5 Tmol Si for the open ocean with coastal data excluded (Table 5). By considering the global gross Si production annual estimate of 240 Tmol Si y^{-1} given by Nelson et al. (1995), this converts to a Si biomass turnover rate comprised between 0.15 and 0.19 d^{-1} (geometric mean). The arithmetic means yield a Si turnover rate of 0.02-0.03 d^{-1}, which seems to be highly underestimated for diatoms.

Next, the mean integrated BSi biomass over 0-200 m (in mmol Si m^{-2}) is presented for each basin and compared to literature data for various oceanic provinces (Table 6). Diatom biomass is usually available indirectly through particulate Si measurements in ocean studies, allowing a comparison between our dataset and actual measurements after conversion from C to Si biomass. Our estimates for open ocean data are comprised between 3.3 and 26.9 mmol Si m^{-2}, which is quite similar to the estimate given in Adjou et al. (2011) of 2 to 26 mmol Si m^{-2} for

HNLC and oligotrophic regions. However, the range of variations of integrated BSi data in various hydrological environments can be quite large and may locally be one to three orders of magnitude higher than our basin averages as evidenced in Table 5.

Unfortunately, we did not find any integrated BSi data for the Arctic Ocean to compare with our data. This region presents a 215% increase of biomass estimates when looking at open ocean data alone (9.9 mmol Si m^{-2}), compared to the entire dataset estimate (4.6 mmol Si m^{-2}), while the Atlantic, Pacific and Indian ocean all show a slight decrease (-3 to -7%) when excluding coastal data, which are generally expected to be skewed towards higher biomasses. This particular feature of the Arctic could be explained by the presence of a broad continental shelf and the impact of large riverine inputs, which could induce large differences between coastal and open ocean biomass. The Atlantic Ocean average estimate (combining data from the Baltic and Mediterranean) is the lowest of all regions (3.3-3.4 mmol m^{-2}) and compares well with literature data for the Mediterranean Sea, the Bermuda Time Series (BATS) and the North Atlantic. Much larger values were found in the Atlantic sector of the ACC (Antarctic Circumpolar Current), which is at the boundary with the Southern Ocean and reflects a very different environment. The Pacific Ocean estimate also compares well with open ocean data (HOT, ALOHA, the Central, Equatorial and Southern Pacific), but is much lower than coastal measurements obtained at Monterey Bay or the Santa Barbara basin which are highly productive coastal systems. The Southern Ocean is the region where the discrepancy between our estimates and measurements is highest, with much lower values than expected for diatoms, and a global budget close to that of the Arctic and Atlantic Oceans. This may be due to poor sampling coverage in the dataset, which is visible on Figure 5, where very few sampling sites are actually documented. The Indian Ocean shows the highest estimates (26.9-29.1 mmol Si m^{-2}) in our dataset and is probably skewed by data from the Kerguelen Plateau, which displays a massive diatom bloom every year. The only data available for BSi are found in the Subantarctic region but unfortunately no other data for the Central and Northern Indian Ocean could be found for comparison.

5. Conclusion

This study provides the first attempt to compile global abundance and biomass data for diatoms in a unique database, with uniform data treatment. Quantitative and qualitative information are provided, but much more information on species distribution, succession and relative importance between biogeographical provinces and coastal/open ocean systems can be derived from the present database, although such coverage is beyond the scope of this paper. Despite significant identified biases in biovolume calculations and C content conversions, these first estimates may be used in global biogeochemical models implementing diatoms as a model variable. First estimates for the global ocean produce a diatom biomass of 37-49 Tmol C and 3-4 Tmol Si, and an average Si biomass turnover rate of 0.15 to 0.19 d^{-1}. Spatial coverage, species identification and cell size assessments may still be improved and taxonomists are encouraged to submit future data to data repositories such as PANGAEA so that they may be used to refine future dataset aggregation projects such as this one.

We emphasize that less than 50 species represent >90% of the total biomass, and that placing more efforts to resolve the listed biases for these dominant species first (which are sometimes less well studied) should help to improve the global biomass estimates considerably. Hence the huge diversity of diatom species in the modern ocean may be reduced down, for more complete studies of size, biovolume and cellular C content assessments, to a more managable number of taxa for global modeling efforts. But we should keep in mind that climate and environmental change may alter this dominance list at any time, and that continued taxonomic identification and counting efforts of the entire plankton flora remains crucial. Another goal was to provide a usable data file for taxonomists worldwide so that they can add further diatom count data and compute their biovolume and C biomass in a similar way. This file is available in open access through the PANGAEA database center (see Appendix A), and will evolve with new data submissions.

Along with other papers of this special issue, this study also clearly highlights that taxonomic work and phytoplankton identification skills are far from obsolete and are needed more than ever if we are to achieve robust datasets of planktonic biomass.

6. APPENDIX A

6.1. Data table

A full table containing all biomass/abundance data points can be downloaded from the data archive PANGAEA, http://doi.pangaea.de/10.1594/PANGAEA.777384. See description of the file in the “Data file content” section (2.3). The excel file allowing for automatic biovolume calculation can be used as a starting tool to create regional diatom databases and is available upon demand to the first author. New data additions to this database are welcomed and will be implemented when available.

6.2. Gridded netcdf biomass product

The biomass data has been gridded onto a 360 x 180° grid, with a vertical resolution of six depth levels: 0-5m, 5-25m, 25-50m, 50-75m, 75-100m and >100m. Data has been converted to netcf format for ease of use in model calculation exercises. The netcdf file can be downloaded from PANGAEA, http://doi.pangaea.de/10.1594/PANGAEA.777384.

Table 1: Latitudinal distribution of biomass data in %.

Latitudinal band	Biomass data in %
90°S-80°S	0.0
80°S-70°S	0.8
70°S-60°S	0.6
60°S-50°S	5.3
50°S-40°S	2.2
40°S-30°S	1.3
30°S-20°S	0.8
20°S-10°S	2.8
10°S-0°	6.9
0°N-10°N	6.5
10°N-20°N	2.4
20°N-30°N	1.3
30°N-40°N	5.5
40°N-50°N	24.5
50°N-60°N	19.4
60°N-70°N	11.8
70°N-80°N	5.1
80°N-90°N	2.9

Table 2: Diatom genera in ascending order of contribution to total biomass. 32 genera amount to 99 % of global biomass. Note that unidentified pennate and centric diatoms represent a non negligible 14.8 % of the total biomass. If they were identified down to genera, the order of dominance for the most abundant groups might change.

Genera	% contribution to total	Genera	% contribution to total
Rhizosolenia	17.4	*Denticulopsis*	0.7
Chaetoceros	14.5	*Fragilariopsis*	0.7
Thalassiosira	12.6	*Paralia*	0.6
Pennate	8.2	*Pseudo-nitzschia*	0.6
Dactyliosolen	7.6	*Asterionellopsis*	0.5
Guinardia	7.3	*Pleurosigma*	0.5
Centric	6.6	*Eucampia*	0.4
Detonula	4.2	*Bacteriastrum*	0.4
Coscinodiscus	3.1	*Actinocyclus*	0.3
Leptocylindrus	3.0	*Thalassionema*	0.2
Nitzschia	2.3	*Navicula*	0.2
Skeletonema	1.8	*Amphiprora*	0.2
Lauderia	1.3	*Corethron*	0.2
Cerataulina	1.1	*Thalassiothrix*	0.2
Proboscia	1.0	*Cyclotella*	0.1
Ditylum	0.9	*Cylindrotheca*	0.1

Table 3: Diatom species (all taxa not identified down to species level were left out of the calculation) in ascending order of contribution to total biomass. 43 species amount to 90 % of global diatom biomass (identified species only).

Species	% contribution to total biomass	Species	% contribution to total biomass
Dactyliosolen fragilissimus	13.6	*Proboscia alata*	0.9
Rhizosolenia imbricata	10.8	*Chaetoceros curvisetus*	0.8
Guinardia striata	8.1	*Guinardia flaccida*	0.8
Detonula pumila	7.7	*Pseudo-nitzschia pungens*	0.7
Guinardia delicatula	4.5	*Fragilariopsis oceanica*	0.7
Leptocylindrus danicus	4.2	*Nitzschia longissima*	0.6
Skeletonema costatum	3.4	*Thalassiosira gravida*	0.6
Rhizosolenia chunii	3.0	*Eucampia zodiacus*	0.5
Chaetoceros socialis	2.6	*Proboscia inermis*	0.5
Rhizosolenia setigera	2.5	*Rhizosolenia hebetata*	0.5
Lauderia annulata	2.5	*Chaetoceros debilis*	0.5
Rhizosolenia robusta	2.4	*Chaetoceros decipiens*	0.5
Cerataulina pelagica	2.1	*Chaetoceros didymus*	0.4
Ditylum brightwellii	1.8	*Guinardia cylindrus*	0.4
Chaetoceros compressus	1.6	*Coscinodiscus wailesii*	0.4
Rhizosolenia styliformis	1.6	*Proboscia indica*	0.4
Leptocylindrus mediterraneus	1.4	*Thalassiosira rotula*	0.4
Coscinodiscus oculus-iridis	1.3	*Thalassionema nitzschioides*	0.4
Thalassiosira nordenskioeldii	1.3	*Nitzschia closterium*	0.3
Paralia sulcata	1.1	*Chaetoceros lorenzianus*	0.3
Asterionellopsis glacialis	1.0	*Detonula confervacea*	0.3
Chaetoceros affinis	0.9		

Table 4 : Global ocean budget of diatom biomass for the entire dataset expressed in Tg C, Tmol C and Tmol Si and Si biomass turnover rate estimates in d^{-1} (see discussion section for calculation details).

	All data 0-100 m		**All data 0-200 m**	
Global Ocean diatom biomass	geometric mean	arithmetic mean	geometric mean	arithmetic mean
Tg C	488	2942	470	3023
Tmol C	41	245	39	252
Tmol Si	3.8	22.8	3.6	23.4
Si biomass turnover rate (d^{-1})	0.17	0.03	0.18	0.03

Table 5 : Global open ocean budget of diatom biomass for the dataset without coastal sites (where bathymetry <100 m) expressed in Tg C, Tmol C and Tmol Si and Si biomass turnover rate estimates in d^{-1} (see discussion section for calculation details).

	Open ocean data 0-100 m		**Open ocean data 0-200 m**	
Global Open Ocean diatom biomass	geometric mean	arithmetic mean	geometric mean	arithmetic mean
Tg C	582	3626	444	3433
Tmol C	49	302	37	286
Tmol Si	4.5	28.1	3.4	26.6
Si biomass turnover rate (d^{-1})	0.15	0.02	0.19	0.02

Table 6: Mean integrated BSi (over 200 m) in mmol m^{-2} calculated from the present database are indicated by the geometric mean and arithmetic means, using a Si:C conversion factor of 0.093 (see discussion section for calculation details). A distinction was made between all available data and open ocean data alone (considering all data points below the 100 m isobath as coastal data). These results are compared to other regional data published in various studies, indicated either as min and max values or by an average ± SD. The areal surface considered for each ocean were 14.056, 76.762, 155.557, 68.556, 20,327 (in 10^{12} m^2) for the Arctic, Atlantic + Mediterranean + Baltic, Pacific, Indian and Southern Oceans respectively.

[1]Leblanc et al., 2005 ; [2]Leblanc et al., 2009 ; [3]Krause et al., 2009 ; [4]Nelson et al., 1995 ; [5]Brzezinski and Kosman,1996 ; [6]Queguiner and Brzezinski, 2002 ; [7]Shipe et al., 2006 ; [8]Peinert and Miquel, 1994; [9]Leblanc et al., 2003; [10]Leblanc et al., 2004; [11]Crombet et al., 2011 ; [12]Brzezinski et al., 2012 ; [13]Brzezinski et al., 1998 ; [14]Krause et al., 2011 ; [15]Brzezinski et al., 2003 ; [16]Brzezinski et al., 1997 ; [17]Shipe et al., 2001 ; [18]Brzezinski et al., 2005 ; [19]Brzezinski et al., 2001 ; [20]Mosseri et al., 2008 ; [21]Leblanc et al., 2002.

Oceanic region	Province	BSi (mmol m^{-2}) *(geom.mean ; arith.mean)*	References
Arctic	*All data*	*4.6 ; 12.9*	*this study*
	Open Ocean data	*9.9 ; 23.1*	*this study*
Atlantic	North Atlantic (POMME)	1.6 – 60.9	1
	North Atlantic (NABE)	17.7 – 102.2	2
	BATS	11.7 – 50.8	3
	BATS	4.0 ± 6.8	4
	Sargasso Sea	1.2 – 109.1	3, 5
	ACC	30.2 – 1231.2	6
	Amazon plume waters	2.0 – 55.9	7
Mediterranean	Western basin	1.0 – 50.0	8,9,10,11
	Eastern basin	3.9 – 6.4	11
Atlantic, Mediterranean & Baltic	*All data*	*3.4 ; 27.7*	*this study*
	Open Ocean data	*3.3 ; 28.3*	*this study*
Pacific	HOT	<10.0	12
	ALOHA	3.0	12
	Central North Pacific	1.8 – 18.4	13
	Eastern Equatorial Pacific	3.8 – 18.0	14
	Monterey Bay	16.3 – 175	15
	Monterey Bay – upwelling event	56 – 566	16
	Santa Barbara basin	6.6 – 380	17
	SOFEX unfertilized North patch (56°S)	4.9 – 13.1	18
	All data	*8.0 ; 52.4*	*this study*
	Open Ocean data	*7.1 ; 75.4*	*this study*
Southern Ocean	Pacific sector (60-66°S)	386 ± 203	19
	SOFEX unfertilized South patch (66°S)	19.1 – 89.8	18
	All data	*4.0 ; 7.8*	*this study*
	Open Ocean data	*4.4 ; 8.4*	*this study*
Indian Ocean	Kerguelen Plateau (KEOPS I)	605 - 2105	20
	Polar Front Zone	46.6 ± 18.7	21
	Subantarcic Zone	31.6 ± 10.1	21
	Subtropical Zone	19.8 ± 2.8	21
	All data	*29.1 ; 186.8*	*this study*
	Open Ocean data	*26.9 ; 178.0*	*this study*

Figures :

Fig.1: Flow diagram of the methodology used to derive diatom biomass estimates from abundance data.

Fig.2: Data distribution according to main oceanic regions (1) North Atlantic, (2) Equatorial Atlantic, (3) South Atlantic, (4) North Pacific, (5) Equatorial Pacific, (6) South Pacific, (7) North Indian, (8) South Indian, (9) Arctic, (10) Antarctic, (11) Baltic, (12) Bering Sea, (13) Gulf of Mexico & Caribbean, (14) Indonesia, (15) Mediterranean.

Fig.3: Frequency of data distribution according to latitude (a) and year (b).

Fig.4: Mean log-normalized diatom biomass ($\log_{10}$ µg C L^{-1}) for different depth layers.

Fig.5: Mean surface log-normalized diatom biomass ($\log_{10}$ µg C L^{-1}) (a) and uncertainty in cell biomass in % of the mean, due to the uncertainty of cell size [=(max biomass-min biomass)/mean biomass*100] (b).

Fig.6: Distribution of log-normalized diatom biomass ($\log_{10}$ µg C L^{-1}) as a function of depth (a) and latitude (b).

Fig.7: Seasonal distribution of log-normalized diatom biomass data ($\log_{10}$ µg C L^{-1}) for the Northern (a) and Southern (b) Hemispheres.

Fig.8: Boxplot of the main diatom genera, contributing to 99 % of the total biomass ($\log_{10}$ µg C L^{-1}) in the database. Red dots represent the 5^{th} and 95^{th} percentiles. Genus contribution to total biomass is arranged in decreasing order of abundance from top to bottom (see Table 2 for relative importance).

Fig.9: Boxplot of the main diatom species, contributing to 90 % of the total biomass ($\log_{10}$ µg C L^{-1}) in the database. Red dots represent the 5^{th} and 95^{th} percentiles. Species contribution to total biomass is arranged in decreasing order of abundance from top to bottom (see Table 3 for relative importance). All undetermined genera (example *Chaetoceros spp.)* were left out of the calculation to focus on identified species.

Fig.10: Boxplot of the minimum, mean and maximum estimates of diatom biomass ($\log_{10}$ µg C L^{-1}). Red dots represent the 5th and 95th percentiles and black circles the outliers.

Acknowledgments

We wish to acknowledge the contribution of several other contributors to this dataset: D. Harbour, M. Estrada, M. Fiala, M-J. Chrétiennot-Dinet, F. Figueiras, F. Gomez, D. Karentz, J. Ramos, T. Robert, M. Silver, G. Tarran and the Plymouth Marine Laboratory (PML) as well as many other anonymous providers of data obtained through data centers such as NODC, BODC, PANGAEA, NMSF-Copepod and the French JGOFS/PROOF programs (BIOSOPE, PROSOPE). Depth-integrated data (e.g. P. Ajani) for diatom abundances were also available, but have not been included in the present database and manuscript. We wish to thank Corinne LeQuéré for leading this excellent initiative, as well as Erik Buitenhuis, Vogt Meike, Scott Doney, Colleen O'Brien and Yawei Luo for putting together and kindly helping out with the Matlab scripts, gridded netcdf files and statistical treatment protocols. Finally, we wish to thank Stéphane Pesant for his work on integrating the present database into PANGAEA in an open access format. Submission of new data is very welcome (please contact karine.leblanc@univ-amu.fr).

References

1. Adjou, M., Tréguer, P., Dumousseaud, C., Corvaisier, R., Brzezinski, M.A., and Nelson D.M. Particulate silica and Si recycling in the surface waters of the Eastern Equatorial Pacific. Deep-Sea Res. II, 58, 449-461, 2011.
2. Alvain, S., Moulin, C., Dandonneau, Y., and Bréon, F. M.: Remote sensing of phytoplankton groups in case I waters from global SeaWifs imagery, Deep-Sea Res. I, 52, 1989-2004, 2005.
3. Brzezinski, M. A.: The Si:C:N ratios of marine diatoms : Interspecific variability and the effect of some environmental variables, J. Phycol., 21, 345-357, 1985.
4. Cornet-Barthaux, V., Armand, L., and Quéguiner, B.: Biovolume and biomass estimates of key diatoms in the southern ocean, Aquatic Microbial Ecology, 48, 295-308, 10.3354/ame048295, 2007.
5. Dupouy, C., Neveux, J., Dirberg, G., Tenório, M. M. B., Röttgers, R., and Ouillon, S.: Bio-optical properties of the marine cyanobacteria *Trichodesmium spp.*, Journal of Applied Remote Sensing 02, 023503, 2008.
6. Eppley, R. W., Reid, F. M. H., and Strickland, J. D. H.: The ecology of the plankton off La Jolla, California, in the period April through September, 1967. Iii. Estimates of phytoplankton crop, size, growth rate, and primary production, Bull. Scripps Inst. Oceanogr., 17, 33-42, 1970.
7. Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P.: Primary production of the biosphere: Integrating terrestrial and oceanic components, Science, 281, 237-240, 10.1126/science.281.5374.237, 1998.
8. Finenko, Z. Z., Hoepffner, N., Williams, R., and Piontkovski, S. A.: Phytoplankton carbon to chlorophyll a ratio: Response to light, temperature and nutrient., Marine Ecological Journal, 2, 40-64, 2003.
9. Gasol, J. M., Del Giorgio, P. A., and Duarte, C. M.: Biomass distribution in marine planktonic communities., Limnol. Oceanogr., 42, 1353-1363, 1997.
10. Guillard, R. R. L., and Kilham, P.: The ecology of marine planktonic diatoms. In : The biology of diatoms, edited by: Werner, D., University of California Press, Berkeley, 1978.

11. Haeckel, V. E.: Plankton-studien. Vergleichende Untersuchungen über die Bedeutung und Zusammensetzung der pelagischen Fauna und Flora. Edited by: Fischer, V. V. G., Jena, 105, 1890.
12. Hillebrand, H., Dürselen, C.-D., Kirtschel, D., Pollingher, U., and Zohary, T.: Biovolume calculation for pelagic and benthic microalgae, J. Phycol., 35, 403-424, 1999.
13. Hood, R. R., Laws, E. A., Armstrong, R. A., Bates, N. R., Brown, C. W., Carlson, C. A., Chai, F., Doney, S. C., Falkowski, P. G., Feely, R. A., Friedrichs, M. A. M., Landry, M. R., Keith Moore, J., Nelson, D. M., Richardson, T. L., Salihoglu, B., Schartau, M., Toole, D. A., and Wiggert, J. D.: Pelagic functional group modeling: Progress, challenges and prospects, Deep-Sea Res. II, 53, 459-512, 2006.
14. Jin, X., Gruber, N., Dunne, J. P., Sarmiento, J. L., and Armstrong, R. A.: Diagnosing the contribution of phytoplankton functional groups to the production and export of particulate organic carbon, $CaCO_3$, and opal from global nutrient and alkalinity distributions, Global Biogeochem. Cycles, 20, 10.1029/2005gb002532, 2006.
15. Le Quéré, C., Harrison, S. P., Prentice, I. C., Buitenhuis, E. T., Aumont, O., Bopp, L., Claustre, H., Cotrim Da Cunha, L., Geider, R., Giraud, X., Klaas, C., Kohfeld, K. E., Legendre, L., Manizza, M., Platt, T., Rivkin, R. B., Sathyendranath, S., Uitz, J., Watson, A. J., and Wolf-Gladrow, D.: Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models, Global Change Biology, 11, 2016-2040, 2005.
16. Menden-Deuer, S., and Lessard, E. J.: Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton., Limnol. Oceanogr., 45, 569-579, 2000.
17. Menden-Deuer, S., Lessard, E. J., and Satterberg, J.: Effect of preservation on dinoflagellate and diatom cell volume and consequences for carbon biomass predictions, Mar. Ecol. Prog. Ser., 222, 41-50, 10.3354/meps222041, 2001.
18. Montagnes, D. J. S., Berges, J. A., Harrison, P. J., and Taylor, F. J. R.: Estimating carbon, nitrogen, protein, and chlorophyll a from volume in marine phytoplankton., Limnol. Oceanogr., 39, 1044-1060, 1994.
19. Nelson, D. M., Tréguer, M. A., Brzezinski, M. A., Leynaert, A., and Quéguiner, B.: Production and dissolution of biogenic silica in the ocean : Revised global estimates, comparison with regional data and relationship to biogenic sedimentation., Global Biogeochemical Cycle, 9, 359-372, 1995.

20. Olenina, I., Hajdu, S., Edler, L., Andersson, A., Wasmund, N., Busch, S., Göbel, J., Gromisz, S., Huseby, S., Huttunen, M., Jaanus, A., Kokkonen, P., Ledaine, I., and Niemkiewicz, E.: Biovolumes and size-classes of phytoplankton in the Baltic Sea., HELCOM Balt.Sea Environ. Proc., 106, 144 pp, 2006.
21. Ragueneau, O., Tréguer, P., Leynaert, A., Anderson, R. F., Brzezinski, M. A., DeMaster, D. J., Dugdale, R. C., Dymond, J., Fisher, G., François, R., Heinze, C., Maier-Reimer, E., Martin-Jézéquel, V., Nelson, D. M., and Quéguiner, B.: A review of the Si cycle in the modern ocean : Recent progress and missing gaps in the application of biogenic opal as a paleoproductivity proxy., Global and Planetary Change, 26, 317-365, 2000.
22. Smayda, T. J.: From phytoplankton to biomass. 6. Phytoplankton manual, Monographs on oceanographic methodology, edited by: Sournia, A., UNESCO, Paris, 1978.
23. Sournia, A., Chrétiennot-Dinet, M.-J., and Ricard, M.: Marine phytoplankton: How many species in the world ocean?, J. Plankton Res., 13, 1093-1099, 10.1093/plankt/13.5.1093, 1991.
24. Sun, J., and Liu, D.: Geometric models for calculating cell biovolume and surface area for phytoplankton, J. Plankton Res., 25, 1331-1346, 10.1093/plankt/fbg096, 2003.
25. Tréguer, P., Nelson, D. M., Van Bennekom, A. J., D.J., D., Leynaert, A., and Quéguiner, B.: The silica balance in the world ocean : A reestimate., Science, 268, 375-379, 1995.
26. Tréguer, P.: Silica and the cycle of carbon in the ocean, Comptes Rendus Geosciences, 334, 3-11, 2002.
27. Uitz, J., Claustre, H., Morel, A., and Hooker, S. B.: Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll., J. Geophys. Res., 111, C08005, doi:08010.01029/02005JC003207, 2006.
28. UNESCO: A review of methods used for quantitative phytoplankton studies, UNESCO Tech Pap Mar Sci 18, UNESCO, Paris, 1974.
29. Viličić, D.: An examination of cell volume in dominant phytoplankton species of the central and southern Adriatic Sea., Int. Revue ges. Hydrobiol., 70, 829-843, 1985.