A Pelagic Size Structure database (PSSdb) to support biogeochemical modeling

third update to the first release

Mathilde Dugenne*[1], **Marco Corrales-Ugalde***[2], Todd O'Brien[3], Fabien Lombard[1], Jean-Olivier Irisson[1], Lars Stemmann[1], Charles Stock[4], Rainer Kiko[5] and Jessica Y. Luo[4].

[1] Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche, 06230 Villefranche-sur-mer, France
[2] Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA.
[3] NOAA Fisheries - Office of Science & Technology - Marine Ecosystems Division, Silver Spring, Maryland, USA
[4] NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA.
[5] Department Ocean Ecosystems Biology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

* Contributed equally

General description:

This dataset represents the third update to the first release of the Pelagic Size Structure database (PSSdb, https://pssdb.net) scientific project, investigating the global particle size distributions measured from multiple pelagic[‡] imaging systems. These devices include the Imaging Flow Cytobot (Olson and Sosik 2007), benchtop scanners like the ZooScan (Gorsky et al. 2010), and the Underwater Vision Profiler (Picheral et al. 2010). The data sources originate from Ecotaxa (https://ecotaxa.obs-vlfr.fr/), Ecopart (https://ecopart.obs-vlfr.fr/), and Imaging FlowCytobot dashboards (https://ifcb.caloos.org/dashboard and https://ifcb-data.whoi.edu/dashboard). Links to the PSSdb code and documentation are available on the PSSdb webpage.

[‡]: All artifacts were removed, but observations from IFCB and UVP include living and detrital particles, whereas benchtop scanner datasets only includes living particles. Note that samples with less than 95% validation of the taxonomic annotation were discarded for scanner Zooscan and UVP projects, but not for IFCB projects, whose annotations are predictions-only. Thus, IFCB observations may include artifacts that were not classified as such, or exclude plankton that were classified as artifacts.

This updated version includes the following changes:

- Duplicate data entries and NaN values have been removed.
- Data products now include Normalized Biomass Size Spectra (NBSS), and Particle Size Distribution (PSD), two widely used methods to represent plankton and particles size distribution in marine ecology and biogeochemistry (Jonasz and Fournier 1996, San Martin et al. 2006).
- Linear regressions are now performed with $\log_{10}$ transformations of the normalized biovolume/abundance and the size classes.
- Inclusion of UVP6 and other benchtop plankton Scanner datasets from net tows, which expand the temporal and spatial coverage of the data products.
- Unbiased portion of the size spectra is selected by a new thresholding method that accounts for both uncertainties on particle sizes, limited by the camera resolution, and particle count (Schartau et al. 2010), so that only size classes with less than 20% uncertainty are retained, in addition to gaps in the size spectra.

Added on March, 2024:

- An error in the thresholding function (scripts/funcs_NBS.py) was corrected.
- A quality control function was implemented (scripts/funcs_NBS.py) to flag size spectra calculations in products 1a and 1b.

Modified on April, 2024:

- An error in the size classes defined in the ecopart_size_bins.tsv used by the size binning function (scripts/funcs_NBS.py) was corrected, the size ratio between consecutive size bins is the same across all the size ranges now.

The dataset is composed of two products, specific to each imaging device:

- Product (1a) includes the size distribution , computed from normalized biovolume, for NBSS, and normalized abundance, for PSD, of plankton and particles within a set of pre-defined size classes (expressed in both biovolume and equivalent circular diameter), averaged by year and month, and in 1-degree longitude/latitude grid cells.

- Product (1b) includes the results of NBSS and PSD regression fit parameters, slopes, intercept, and coefficient of determination (R^2), averaged by year and month, and in 1-degree longitude/latitude grid cells. The regression parameters are defined using ordinary least squares linear regressions applied to a $\log_{10}$ transformed normalized biovolume/normalized abundance and biovolume/ diameter size class values.

Size spectra parameters were averaged over a maximum of 16 spatial and temporal subsets (0.5°x0.5°x1 week) to avoid over-representation of repeated sampling events (e.g., time-series datasets) within a grid cell. Linear regressions were performed on the linear portion of the $\log_{10}$-transformed NBSS and PSD estimates, between the size classes with a size measurement or particle count uncertainty greater than 20% (Schartau et al. 2010) , and where the maximum NB/PSD is observed and the largest size class before three empty consecutive size classes.

Products 1a: Size distribution data for a single imaging system:

The name of these files follows the format:

*Instrument*_1a_Size-distribution_v*Year-Month*.csv

Where *Instrument* refers to one of the three types of plankton imaging systems included in this first data release: Imaging FlowCytobot (IFCB), Scanner (including ZooScan), and Underwater Vision Profiler (UVP). *vYear-Month* is the date when the dataset was generated (*e.g.*, “v2023-05”)..

Column descriptions:

- **year:** sampling year
- **month:** sampling month.
- **latitude** (decimal degrees)**:** latitude of the center point for each 1-degree cell
- **longitude** (decimal degrees)**:** longitude of the center point for each 1-degree cell
- **ocean** : label assigned using the Flanders Marine Institute Global Oceans and Seas (v1) datasets (https://doi.org/10.14284/542), available online at https://marineregions.org/.
- **min_depth** (meters)**:** minimum sampled depth.
- **max_depth** (meters): maximum sampled depth.
- **“n”:** number of spatial and temporal subsets used to generate a monthly average of NB per size class and per 1-degree cell. Note: in the case where a week spans two months, a simple rounded average of the sampled months was computed for a unique month identifier. E.g., if a week encompassed the last 5 days of one month and the first 2 days of the next, but the sampled data were predominantly from the second month, then resultant averaged month value would be the second month.
- **biovolume_size_class** (cubic micrometers)**:** midpoint of the size class in which the observations were categorized. Size classes were defined as spherical projections of the equivalent spherical diameter (ESD) size classes in Ecopart, and were expanded to cover all sizes of marine particles sampled by plankton imaging systems.
- **normalized_biovolume_mean** (cubic micrometers per liter per cubic micrometers)**:** normalized Biovolume for a size class and for each spatial and temporal subset (0.5°x0.5°x1 week) is calculated as:
  NB = Sum of all biovolume in a size class / (cumulative volume sampled x width of the size bin class).
  The resulting NB are averaged per year and month, in 1-degree grid cells.
- **normalized_biovolume_std** (cubic micrometers per liter per cubic micrometers)**:** standard deviation of normalized_biovolume_mean, per year and month, in 1-degree grid cells. This is only computed and present when n >1.
- **equivalent_circular_diameter_mean (micrometers):** geometric mean of the minimum and maximum values of the size class interval in which the observations were categorized. These size classes were obtained by calculating the diameter of the sphere used to calculate the ‘biovolume_size_class’ variable.

- n**ormalized_abundance_mean:** (# particles per Liter per micrometer): normalized abundance for an equivalent circular diameter size class and for each spatial and temporal subset (0.5°x0.5°x1 week) is calculated as:
  NB = Sum of all particle counts in a size class / (cumulative volume sampled x width of the equivalent circular diameter bin class).
- **normalized_abundance_std** (# particles per Liter per micrometer): standard deviation of normalize_abundance_mean, per year and month, in 1-degree grid cells. This is only computed and present when n >1.

Quality Control columns (added March 2024)

- **QC_3std_dev** (1 if data is flagged, 0 otherwise): Flag identifying spectra whose slopes are outside the mean ± 3 standard deviations.
- **QC_min_n_size_bins** (1 if data is flagged, 0 otherwise): Flag identifying spectra which only include 4 or less non-empty size classes.z
- **QC_R2** (1 if data is flagged, 0 otherwise): Flag identifying spectra whose linear fits yield a coefficient of determination below 0.8.

Products 1b: Size spectra least-squares linear regression results:

Least squares linear regression for NBSS and PSD variables reported in (1a) as:
NBSS: $\log_{10}$(normalized_biovolume) = (slope x $\log_{10}$(biovolume_size_class)) + intercept
PSD: $\log_{10}$n(normalized_abundance) = (slope x $\log_{10}$ln(equivalent_circular_diameter)) + intercept

The name of these files follows the format:

*Instrument*_1b_Size-spectra-fit_v*Year-Month*.csv

Where *Instrument* refers to one of the three plankton imaging systems included in this first data release: Imaging FlowCytobot (IFCB), Scanner (including ZooScan), and Underwater Vision Profiler (UVP). *Year-Month* is the date when the dataset was generated.

Column descriptions:

- **year:** sampling year
- **month:** sampling month.
- **latitude** (decimal degrees)**:** latitude of the center point for each 1-degree cell
- **longitude** (decimal degrees) **:** longitude of the center point for each 1-degree cell
- **ocean** : label assigned using the Flanders Marine Institute Global Oceans and Seas (v1) datasets (https://doi.org/10.14284/542), available online at https://marineregions.org/.
- **min_depth** (meters)**:** minimum sampled depth.
- **max_depth** (meters): maximum sampled depth

- **“n”:** number of spatial and temporal subsets used to generate a monthly average per 1-degree cell. Subsets are defined as weekly, 0.5 degree cells; the maximum value is 16. All size spectra calculated for the 0.5-degree cells and for each week that were used to get monthly averages for each 1-degree cell. See note from above regarding weeks that span two months.
- **NBSS_slope_mean** (log_{10} liters per cubic micrometer): mean NBSS slopes per year and month, in 1-degree grid cells.
- **NBSS_intercept_mean (**log_{10} cubic micrometers per liter per cubic micrometer)**:** mean NBSS intercepts (in ln) per year and month, in 1-degree grid cells.
- **NBSS_r2_mean:** mean determination coefficient of the NBSS calculation, per year and month, in 1-degree grid cells.

NOTE: The new three “std” columns will only be present when sample_size *is greater than 1.*

- **NBSS_slope_std** (log_{10} liter per cubic micrometer)**:** standard deviation of NBSS slopes per year and month, in 1-degree grid cells, only computed when n > 1.
- **NBSS_intercept_std** (log_{10} cubic micrometer per liter per cubic micrometer)**:** standard deviation of NBSS intercepts (in log_{10}), per year and month, in 1-degree grid cells, only computed when n > 1.
- **NBSS_r2_std:** standard deviation of the determination coefficient of the NBSS calculation, per year and month, in 1-degree cells, only computed when n > 1.

- **PSD_slope_mean** (log_{10} liter per micrometer squared): mean PSD slopes per year and month, in 1-degree grid cells.
- **PSD_intercept_mean (**log_{10} # particles per liter per micrometer)**:** mean NBSS intercepts (in log_{10}) per year and month, in 1-degree grid cells.
- **PSD_r2_mean:** mean determination coefficient of the NBSS calculation, per year and month, in 1-degree grid cells.

NOTE: The new three “std” columns will only be present when sample_size *is greater than 1.*

- **PSD_slope_std** (log_{10} liter per micrometer)**:** standard deviation of NBSS slopes per year and month, in 1-degree grid cells, only computed when n > 1.
- **PSD_intercept_std** (log_{10} # particles per liter per micrometer)**:** standard deviation of NBSS intercepts (in log_{10}), per year and month, in 1-degree grid cells, only computed when n > 1.
- **PSD_r2_std:** standard deviation of the determination coefficient of the NBSS calculation, per year and month, in 1-degree cells, only computed when n > 1.

Quality Control columns (added March 2024)

- **QC_3std_dev** (1 if data is flagged, 0 otherwise): Flag identifying spectra whose slopes are outside the mean ± 3 standard deviations.
- **QC_min_n_size_bins** (1 if data is flagged, 0 otherwise): Flag identifying spectra which only include 4 or less non-empty size classes.
- **QC_R2** (1 if data is flagged, 0 otherwise): Flag identifying spectra whose linear fits yield a coefficient of determination below 0.8.

Data sources:

This file contains information on the data repositories and original data owners that have contributed to this first release of PSSdb.

Column descriptions:

- **Data source:** repository that contained the dataset
- **Project:** identifier for each dataset in the repository
- **Instrument:** one of the three plankton imaging systems included in this first data release: Imaging FlowCytobot (IFCB), ZooScan, and Underwater Vision Profiler (UVP).
- **Data owner :** Name of the original data owner
- **Owner email**: email of the data owner
- **Additional co-authors:** other researchers involved in generating the datasets
- **Funding sources**: funding used to generate the datasets

Acknowledgements

This work was supported by NOAA-CPO GC21-407a award NA21OAR4310254.

References:

Jonasz, M. and Fournier, G. 1996. Approximation of the size distribution of marine particles by a sum of log-normal functions. *Limnol. Oceanogr.*, 41, 744-754

Olson R. J. and Sosik H. M. 2007. A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. *Limnol. Oceanogr.Methods* 5, 2007, 195–203

Gorsky G., Ohman M. D., Picheral M., Gasparini S., Stemmann L., Romagnan B., Cawood A., Pesant S., Garcia-Comas C., Prejger F. 2010. Digital zooplankton image analysis using the ZooScan integrated system. *J. Plankton Res. 32*, 2010, 285–303

Picheral M., Guidi L., Stemmann L., Karl D. M., Iddaoud G., Gorsky G. 2010. The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnol. Oceanogr.: Methods* 8, 2010, 462–473

San-Martin E., Harris R.P., Irigoien X. 2006. Latitudinal variation in plankton size spectra in the Atlantic Ocean. *Deep Sea Res. II.* 53, 2006, 1560-1572

Schartau M., Landry M.R., Armstrong R.A. 2010. Density estimation of plankton size spectra: a reanalysis of IronEX II data. *J. Plankton Res.* 32, 2010, 1167-1184