A Pelagic Size Structure database (PSSdb) to support biogeochemical modeling: 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 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), the ZooScan (Gorsky et al. 2010), and the Underwater Vision Profiler (Picheral et al. 2010). The data sources come 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, additional documentation, and updates to the data and methods (*e.g.*, as the project progresses) will be available on the PSSdb webpage (https://pssdb.net).

‡: All artifacts were removed, but observations from IFCB and UVP include living and detrital particles, whereas Zooscan only includes living particles. Note that samples with less than 95% validation of the taxonomic annotation were discarded for 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 initial dataset is composed of two products, separated by imaging device:

- **Product 1a** includes the normalized biovolume (NB) of plankton and particles within a set of pre-defined size classes, averaged by year and month, and in 1-degree longitude/latitude grid cells.

- **Product 1b** includes the results of normalized biovolume size spectra (NBSS) calculations, averaged by year and month, and in 1-degree longitude/latitude grid cells.

  The NBSS calculations are a set of ordinary least squares linear regressions applied to natural log (ln) transformed NB and biovolume size class values. Regression slopes, intercept and regression coefficient, or r^2, values, along with their standard deviation (when n > 1) are given. NBSS values are given as an average of a maximum of 16 spatial and temporal subsets (0.5° x 0.5° x 1 week) used 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 ln-transformed NBSS estimates, between the size class where the maximum NB is observed and the largest size class before three empty consecutive size classes.

Product 1a: Normalized biovolume database for a single imaging system:

The name of these files follows the format:

*Instrument*_1a_Biovolume-by-Size_v*Year-Month*.csv

Where *Instrument* refers to one of the three plankton imaging systems included in this first data release: Imaging FlowCytobot (IFCB), 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** (code): ocean or inland sea associated with the latitude/longitude coordinates: o01 = Arctic Ocean, o02 = North Atlantic, o03 = South Atlantic, o04 = Southern Ocean, o05 = Indian Ocean, o06 = South Pacific, o07 = North Pacific, o21 = Baltic Sea, o22 = Mediterranean Sea, o24 = Red Sea.
- **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.

Product 1b: NBSS least-squares linear regression results:

Least squares linear regression for each NBSS reported in (1a) was performed as:

ln(normalized_biovolume) = (slope x ln(biovolume_size_class)) + intercept

The name of these files follows the format:

*Instrument*_1b_NBSS-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), 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** (code): ocean or inland sea associated with the latitude/longitude coordinates: o01 = Arctic Ocean, o02 = North Atlantic, o03 = South Atlantic, o04 = Southern Ocean, o05 = Indian Ocean, o06 = South Pacific, o07 = North Pacific, o21 = Baltic Sea, o22 = Mediterranean Sea, o24 = Red Sea.
- **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.
- **slope_mean** (In liters per cubic micrometers): mean NBSS slopes per year and month, in 1-degree grid cells.
- **intercept_mean** In cubic micrometers per liters per cubic micrometers)**:** mean NBSS intercepts (in In) per year and month, in 1-degree grid cells.
- **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 ("n") *is greater than 1.*

- **slope_std** (In liters per cubic micrometers)**:** standard deviation of NBSS slopes per year and month, in 1-degree grid cells, only computed when n > 1.
- **intercept_std** (In cubic micrometers per liters per cubic micrometers)**:** standard deviation of NBSS intercepts (in In), per year and month, in 1-degree grid cells, only computed when n > 1.
- **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.

Data sources: Contributing projects and programs

The file *PSSdb_data-sources_v2023-05.xlsx* contains information on the data repositories and original data owners that have contributed to this first release of PSSdb.

Column descriptions:

- **Instrument:** one of the three plankton imaging systems included in this first data release: Imaging FlowCytobot (IFCB), ZooScan, and Underwater Vision Profiler (UVP).
- **Project ID:** project identifier used in Ecotaxa/ IFCB dashboards
- **Project:** text identifier for each dataset in the repository
- **Program:** name of the program associated with this project
- **Funding sources**: funding used to generate the datasets
- ***Reference DOIs*** for the data, publication, cruise, and similar
- **Data source:** repository that contained the dataset

- **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

Acknowledgements

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

References:

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