**TITLE: AtlantECO [WP2] – Traditional microscopy dataset –** **Thaliacea (Salpida+Doliolida+Pyromosomatida) abundance and biomass concentration data**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**1.- INTRODUCTION**

This dataset contains **447 920** georeferenced abundance and biomass concentration records of **23** accepted scientific names of Thaliacea (i.e., Salpida, Doliolida and Pyrosomatida together) of various taxonomic levels. This dataset is a compilation of the following five main global and regional datasets that reported abundances of marine planktonic thaliaceans:

* The Coastal & Oceanic Plankton Ecology, Production & Observation Database (NMFS-COPEPOD, O’Brien, 2014) from the National Oceanic and Atmospheric Administration - <https://www.st.nmfs.noaa.gov/copepod/atlas/html/taxatlas_4350000.html>
* The Jellyfish Database Initiative (JeDI; Lucas et al., 2014) - doi:10.1111/geb.12169
* The Southern Ocean CPR (SO-CPR) survey (Hosie, 2021) - doi:10.26179/ksds-s610
* The Australian CPR (AusCPR) survey (AusCPR) - <https://catalogue-imos.aodn.org.au/geonetwork/srv/eng/catalog.search#/metadata/c1344e70-480e-0993-e044-00144f7bc0f4>
* The salps observations (mostly *Salpa thompsoni*) of KRILLBASE (Atkinson et al., 2017) - [www.earth-syst-sci-data.net/9/193/2017/](http://www.earth-syst-sci-data.net/9/193/2017/)

**2.- METHODOLOGY USED**

We compiled an exhaustive dataset of in situ pelagic tunicates (i.e., Thaliaceans) concentrations from large scale plankton monitoring programs and previous plankton data compilations to derive monthly field of pelagic tunicates biomass (in mgC.m-3) that can be used as a standard data set to evaluate the FFGM biomass estimated by PISCES-FFGM. First, five main data sources were retrieved: NOAA’s Coastal and Oceanic Plankton Ecology, Production, and Observation Database (COPEPOD; O’Brien et al. 2014), the Jellyfish Database Initiative (JeDI; Lucas et al. 2014), KRILLBASE (Atkinson et al. 2017), the Australian Continuous Plankton Recorder (CPR) survey (AusCPR; IMOS 2021) and the Southern Ocean CPR survey (SO-CPR; Hosie et al. 2021). This compilation gathered planetary scale plankton concentration measurements collected through a broad variety of sampling devices over the last 150 years, with taxonomic identification of varying precision and scientific names, some of which changed through time. Therefore, we curated the scientific names and the taxonomic classification of each observation to harmonize names across all data sets and to correct deprecated names and synonyms based on the backbone classification of the World Register of Marine Species (WoRMS; Horton et al. 2022) using the ‘worms’ R package version 0.2.2 (Holstein et al. 2018). Then, only those observations corresponding to an organism belonging to the Class Thaliacea were kept. Observations without a precise sampling date and and at least one sampling depth indicator (usually maximum sampling depth, in meters) were discarded. All data sets provided concentrations in ind.m-3 except KRILLBASE which provided Salp (mostly *Salpa* *thompsoni*) densities in ind.m-2 which we converted to ind.m-3 based on the maximum sampling depth of the corresponding net tows. In KRILLBASE, 5 186 observations of Thaliaceans with missing density values were discarded (35.6% of the original 14 543 observations). In COPEPOD, concentrations are standardized as if they were all taken from a plankton net equipped with a 330 µm mesh (Moriarty et al., 2013). 862 point observations with missing concentration values were discarded (3.5% of the original 24 316 observations). We examined the composition of the original data sources compiled within JeDI and COPEPOD by assessing the recorded institution codes as well as their corresponding spatio-temporal distributions to evaluate the observations overlapping between these two previous data syntheses. We logically observed a very high overlap between COPEPOD and JeDI as the former data set was the main data contributor to the latter. Therefore, overlapping records were identified based on their sampling metadata, scientific names, concentration values, the recorded institution codes and recorded data sources, and they were removed from JeDI. This removed 14 198 (74.1%) of the JeDI’s original Thaliaceans observations (total n records = 491 529).

Finally, we checked for potential further duplicates again through the definition of an ‘occurrence ID’ that was based on the following set of metadata: ‘decimalLatitude’, ‘decimalLongitude’, ‘Day’, ‘Month’, ‘Year’, ‘Depth’, ‘MinDepth’, ‘MaxDepth’, ‘ScientificName’, ‘MeasurementValue’ and ‘MeshSize’. We found 447 920 unique zooplankton occurrences (corresponding to 91.1% of the data) and 43 607 were duplicates based on the parameters chosen to define the occurrence ID.

Most of the records showed a fairly precise taxonomic resolution as 39% of the data was species- resolved 0.19% genus-resolved and 38% family-resolved. Therefore, we were able to perform taxon-specific conversions from individual concentrations to biomass concentrations (in mgC.m-3) for each point observation. We used the taxon-specific carbon weights (mgC.ind-1) summarized by Lucas et al. (2014) which were based on the group-specific length–mass or mass–mass linear and logistic regression equations of Lucas et al. (2011). Not all the observations had a precise counterpart in the carbon weights compilation of Lucas et al. (2014) because they were not identified at the species or the genus level (e.g., Class-level, Order-level or Family-level observations). In these cases, we computed the median carbon weight of those taxa reported in Lucas et al. (2014) and which composed the higher level taxonomic group (i.e., the carbon weight of Salpidae corresponded to the average carbon weight of all Salpidae species), and used this average carbon weight to convert the individual concentrations to carbon concentrations.

More extensive details about the implementation of the dataset and the way the abundances were converted to biomass concentrations are available in Clerc et al. (2023) - <https://doi.org/10.5194/egusphere-2022-1282>.

On top of the libraries already mentioned above, the main R packages used to implement this dataset were: ‘tidyverse’ (Wickham et al., 2019), ‘reshape2’ version 1.4.4 (Wickham, 2007), ‘marmap’ version 1.0.6 (Pante & Simon-Bouhet, 2013), ‘lubridate’ version 1.8.0 (Grolemund, 2011) and ‘raster’ version 3.5-15 (Hijmans, 2022).

**3.- DATASET DESCRIPTION**

**Data type:** Abundances converted to biomass concentrations.

**Latitude/Longitude format:** WGS 84 (-180°E/+180°E).

**Geographic area covered by the dataset:** Global Ocean.

**Depth range covered by the dataset:** From 0m to 3600m.

**Time period covered by the dataset:** From 11-02-1926 to 31-01-2021.

**Dataset format:** .csv file withsemicolon-delimited columns.

**Date of dataset creation:** 22/09/2022.

**Raw dataset repository:** Zenodo.

**4.- MAIN VARIABLE DESCRIPTION**

MeasurementTypeID: Has not been defined within AtlantECO

MeasurementValue: Organisms concentration (i.e., abundance) in ind.m-3

MeasurementID: Has not been defined within AtlantECO

occurrenceID: combination of ‘decimalLatitude’, ‘decimalLongitude’, ‘Day’, ‘Month’, ‘Year’, ‘Depth’, ‘MinDepth’, ‘MaxDepth’, ‘ScientificName’, ‘MeasurementValue’ and ‘MeshSize’.

**5.- DATA OVERVIEW**

**

**6.- CONTRIBUTORS**

- Corentin Clerc (cclerc@lmd.ipsl.fr), ENS-LMD, France.

- Fabio Benedetti (fabio.benedetti@usys.ethz.ch), ETH Zürich, Switzerland.

- Olivier Aumont (olivier.aumont@ird.fr), IRD-LOCEAN, France.

- Laurent Bopp (bopp@lmd.ens.fr), LSCE, France.

- Meike Vogt (meike.vogt@usys.ethz.ch), ETH Zürich, Switzerland.