**TITLE: AtlantECO [WP2] – Traditional microscopy dataset –** **Jellyfish (Cnidaria+Ctenophora) abundance and biomass concentration data**

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

**1.- INTRODUCTION**

This dataset contains 590 371georeferenced abundance and biomass concentration records of 109 accepted scientific names of Jellyfish (i.e., Cnidaria and Ctenophora together) of various taxonomic levels. This dataset is a compilation of the following five main global and regional datasets that reported abundances of planktonic marine jellyfish:

* 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_4030000.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 Cnidaria concentrations recorded through the plankton net samples of the Tara Oceans expeditions and that were analyzed through the ZooScan imaging system (Brandão, Benedetti et al., 2021) - <https://doi.org/10.1038/s41598-021-94615-5>

**2.- METHODOLOGY USED**

The datasets listed above were first downloaded and re-formatted to the common AtlantECO WP2 data format. To homogenize the taxonomic classification of all the AtlantECO WP2 datasets containing microbiome (i.e., plankton) diversity data, the species names of each dataset were automatically compared to the list of species names accepted in the World Register of Marine Species (WoRMS), using the ‘worms’ R package version 0.2.2 (Holstein, 2018). Typos, synonyms and deprecated species names were corrected and the unique numerical identifiers of each accepted scientific name (i.e., AphiaID) were provided as well. Then, the datasets were progressively combined in as follows.

First, the NMFS-COPEPOD cnidaria records (n = 51 590) were used as a basis since it corresponds to the most widely used compilation of historic zooplankton records. 4 286 records were removed because they were associated with missing abundance values (total n records = 47 304). Second, we added the 198 746 abundance records of cnidaria and ctenophora from the other global compilation of jellyfish records that is JeDI, which is known to contain parts of the NMFS-COPEPOD datasets. A single JeDI record was first discarded as it displayed a negative abundance value, which was considered erroneous. Furthermore, we noted that many (n = 140 357; corresponding to nearly 70% of all the records) JeDI records showed missing sampling depth parameters (i.e., missing values of ‘Depth’, ‘MinDepth’ and ‘MaxDepth’ at the same time). These had to be which shortened the contribution of JeDI to the present dataset to 58 388 records (total n records = 105 686). Because we knew JeDI and NMFS-COPEPOD have duplicate records, we checked for potential duplicates at this stage through the definition of an ‘occurrence ID’ that was based on the following parameters: decimal longitude, decimal latitude, sampling date, maximum sampling depth, scientific name and measured abundance value (‘MeasurementValue’). We found 79 788 unique records (75.5% of the data) and we removed the 25 898 duplicates that were defined based on the parameters chosen to define the occurrence ID.

Then, we added the 362 019 cnidaria abundance records from the SO-CPR survey (no ctenophora records were present in this dataset). No records were removed because of missing abundance values (total n records = 441 807). Third, we added the 148 176 cnidaria records from the AusCPR survey (again, no ctenophora records were present in this dataset; n total records = 589 983). Then, we added the Tara Oceans imaging dataset available (Brandão, Benedetti et al., 2021), which added 388 records of Cnidaria abundance (total n records = 590 371).

The cleaned abundance sheet were passed through the following stages of processing to reach the final data set files.

1. The taxonomy of the individual carbon content values was checked against WoRMS. Individual carbon content values are individual carbon content measurements and Mean/Median/Min/Max/Stdev estimates (in mgC.ind-1). For Jellyfish, 25 carbon content measurements derived from the following sources were compiled: Moriarty, R., Buitenhuis, E., Le Quéré, C., & Gosselin, M.-P. (2013). Distribution of known macrozooplankton abundance and biomass in the global ocean. Earth System Science Data, 5(2), 241-257 and Kiørboe, T. (2013). Zooplankton body composition. Limnology and oceanography, 58(5), 1843-1850. doi:https://doi.org/10.4319/lo.2013.58.5.1843. These values can be found on sheet 3 of “AtlantECO-BASEv1\_dataset\_Amphipoda\_ind\_carbon\_values\_20220930.xlsx”. This is also done for the field observations in the abundance sheet.
2. The life form variable gives information of the type of population organization- and the ecological organization of the organism recorded (e.g."singular","colonial","symbiotic","free living" etc.). As data is taken from different sources, the ‘LifeForm’ entries show a range of different information and formats. All the life form entries within the abundance sheet ("AtlantECO-BASEv1\_dataset\_Jellyfish\_abundances\_22\_04\_22.csv”) and individual carbon content values are standardized by subdividing the information in more detailed variables; Life Form Phase, Stage Name, Stage Number, Life form Min Size and Sex.
3. The unit of which the field observation is measured in it standardized. All entries in #/m2 are converted to #/m3. This is done by dividing the MeasurementValue (numeric value of the measurement ie. Concentration) by the difference between maximum and minimum depth (MeasurementValue /(MaxDepth – MinDepth)).
4. To calculate biomass concentrations, estimates of individual carbon content at the most precise taxonomic resolution possible (i.e., species-level mean carbon content to convert a species-level abundance observations) were determined. We took the highest known taxonomic classification of the observation and match it to the individual carbon content values. When there was no immediate match, we worked up the taxonomic ranking until a match is found. For example, for an observation identified down to the species level, *Aegina citrea,* no associated individual carbon content values are found until the Class level (Hydrozoa). Therefore, it is matched with all individual carbon content values that have class Hydrozoa. In this instance the mean carbon mass calculated for this observation is composed of 13 different individual carbon content values. The minimum (Min) and maximum (Max) carbon mass is the lowest and highest values that we used to compute the mean (Mean) and its standard deviation (Stdev). When additional information of life form attributes is known, of either minimum size, phase, larval stage name, larval stage number and/or sex, it is considered when looking for matches. In some instances, the highest classification is to group level, which is an arbitrary name we assigned to functional groups to organize data sheets. When this occurs, all inidivual carbon content values are averaged.
   1. One exception to this were the observations identified only as “Jellyfish”. The term “Jellyfish” has been used through the past decades in a very heterogeneous way, meaning that it did not always meant to gather the same types of zooplanktonic organisms. Nowadays, this terms is mainly used to describe holoplanktonic, meroplanktonic and sometimes benthic Cnidarians (e.g., Scyphozoa, Cubozoa, Hydrozoa etc.) and Ctenophores of different size classes and life history traits. Therefore, we did not consider it as a robust functional grouping and decided not to derive any individual carbon content value from the literature for those field observations only identified as “Jellyfish”.
5. Each abundance record was converted to minimum, maximum and mean carbon biomass concentration (‘MinBiomass’, ‘MaxBiomass’ and ‘MeanBiomass’, respectively) expressed in mgC.m-3 based on estimates minimum, maximum and mean individual carbon content (‘MinBiomassConversion’, ‘MaxBiomassConversion’ and ‘MeanBiomassConversion’, respectively) expressed in mgC.ind-1.
6. Then the biomass concentrations of sampling events were summed. A sample event is defined by the Latitude, Longitude, Event Date, Depth, Year, Month and Day. This produces the total biomass concentration of all organisms of that specific event. The average biomass value was taken of organisms that had multiple replications at a sample event to avoid increase in biomass due to sampling replication.
7. The summed biomasses are rasterized by latitude, longitude and depth levels (WOA Depth Levels). As the depth levels do not show even intervals, spatial cubes of different y lengths are produced. Then to produce 1 lat x 1 long x 122 depth x 12 month dimensions the mean biomass per month over all the years is taken per 1 x 1 x 122. The minimum biomass is (MinBiomass) is therefore comes from the month with the lowest value and the maximum biomass (MaxBiomass) is the month with the highest value.

The main R packages used to implement ZooBase v2 were: ‘rlang’ version 1.0.6 (Henley and Wickham, 2022), ‘fst’ version 0.9.8 (Klik, 2022), ‘writexl’ version 1.4.0 (Ooms and McNamara, 2021), ‘reshape2’ version 1.4.4 (Wickham, 2020), ‘RNetCDF’ version 2.5-2 (Michna and Woods, 2022), ‘lubridate’ version 1.8.0, (Grolemund and Wickham, 2011), ‘progress’ version 1.2.2 (Csárdi, 2018), ‘units’ version 0.7-2 (Pebesma et al. 2016), ‘stringr’ version 1.4.1 (Wickham, 2022), ‘tidytable’ version 0.9.0 (Fairbanks et al., 2022), ‘readxl’ version 1.4.1 (Wickam and Bryan, 2022), ‘worms’ version 0.4.4 (Chamberlain and Bart, 2022, ‘taxize’ version 0.9.100 (Chamberlain et al., 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 4911m.

**Time period covered by the dataset:** From 09-07-1938 to 31-01-2021.

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

**Date of dataset creation:** 20/12/2022.

**Raw dataset repository:** AtlantECO’s GeoNode (<https://atlanteco-geonode.eu/>).

**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, MaxDepth, ScientificName, MeasurementValue.

**5.- DATA OVERVIEW**

*Diagram

Description automatically generated*

**6.- CONTRIBUTORS**

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

- Ruby Bader (ruby.bader@usys.ethz.ch), ETH Zürich, Switzerland.

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