**TITLE: AtlantECO [WP2] – Traditional microscopy dataset –** **Euphausiacea abundance and biomass concentration data**

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

**1.- INTRODUCTION**

This dataset contains **1 071 450** georeferenced abundance and biomass concentration records of **39** accepted scientific names of Euphausiacea (i.e., krill) of various taxonomic levels. This dataset is a compilation of the following four main global and regional datasets that reported abundances of planktonic marine euphausiids:

* 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_4218000.html>
* 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 KRILLBASE dataset (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**

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 dataset was used as a basis since it corresponds to the most widely used compilation of historic zooplankton records. 2 791 records were removed because they were associated with missing abundance values (total n records = 34 659). Older CPR records present in NMFS-COPEPOD were kept as the available data from the North Atlantic and North Pacific CPR survey (Richardson et al., 2006) did not contain krill records. Second, we added the 4 447 662 euphausiids abundance records from the SO-CPR survey. 1 189 491 records were removed because they were associated with missing abundance values (total n records = 3 292 830). Third, we added the 451 584 euphausiids abundance records from the AusCPR survey. A single record was removed because it displayed a missing abundance values (total n records = 3 744 413). Finally, we added the KRILLBASE dataset, which added 11 593 records of euphausiids abundance (2 950 records were discarded because of missing abundance data; total n records = 3 756 006).

**WARNING**: KRILLBASE’s standardized abundance estimates were originally provided in ind.m-2. To be comparable with the abundance measurements from the other datasets, they can be easily converted to ind.m-3 based on sampling depth parameters (‘MinDepth’ and ‘MaxDepth’; see Atkinson et al., 2017).

Finally, we checked for potential duplicates again 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, measured abundance value (‘MeasurementValue’) and the measurement’s unit (ind.m-3 or ind.m-2 for the KRILLBASE data). No more than 5 records were removed prior to occurrence ID definition as they displayed missing maximum sampling depth (‘MaxDepth’) values. We found 1 071 450 unique records (71.5% of the data) and we thus removed the 2 684 551 duplicate records that were identified based on the parameters chosen to define the occurrence ID.

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 values was checked against WoRMS. Individual carbon values are individual carbon content measurements and Mean/Median/Min/Max/Stdev estimates (in mgC.ind-1). For Euphausiacea, 51 carbon content measurements derived fromMoriarty, 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, Kiørboe, T. (2013). Zooplankton body composition. Limnology and oceanography, 58(5), 18431850. doi:https://doi.org/10.4319/lo.2013.58.5.1843 and Virtue, P., Meyer, B., Freier, U., Nichols, P. D., Jia, Z., King, R., . . . Kawaguchi, S. (2016). Condition of larval (furcilia VI) and one year old juvenile Euphausia superba during the winter–spring transition in East Antarctica. Deep Sea Research Part II: Topical Studies in Oceanography, 131, 182-188. doi:https://doi.org/10.1016/j.dsr2.2016.02.001. These values can be found on sheet 3 of “AtlantECO-BASEv1\_dataset\_Euphausiacea\_ind\_carbon\_values\_20220930.xlsx”. This is also done for the field observations in the abundance sheet.
2. The life form variable gives information about 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 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, i.e., organisms 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, *Euphausia crystallorophiasm,* no associated individual carbon content values are found until the genus level (*Euphausia*). Therefore, it is matched with all individual carbon content values that have genus *Euphausia*. In this instance the mean carbon mass calculated for this observation is composed of 36 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 individual carbon content values are averaged.
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 have 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) 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), and ‘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 3884m.

**Time period covered by the dataset:** From 03-03-1926 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, MeasurementUnit.

**5.- DATA OVERVIEW**

****

**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.