eosc | Blue-Cloud2026

CEPHALOPOD: a standardized and flexible framework for marine habitat modelling across data types and sources

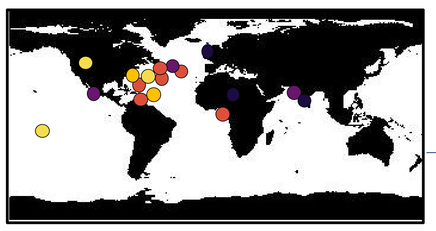
Alexandre Schickele[1]*, Corentin Clerc[1]*, Fabio Benedetti[1], Jean-Olivier Irisson[2], Matthias Münnich[1], Meike Vogt[1]
[1]ETH Zürich, Institute of Biogeochemistry and Pollutant, Zurich, Switzerland
[2]Sorbonne Université, CNRS, LOV, Villefranche-sur-Mer, France
*equally contributed to this poster

Abstract

Here, we introduce the Cloud-based Ensemble Pipeline for Habitat modelling Across Large-scale Ocean Plankton Observation Datasets (CEPHALOPOD). Through two example of applications, we illustrate the framework ability to extrapolate plankton distributions in space and time based on traditional net observations or omics datasets. CEPHALOPOD addresses to a broad spectrum of stakeholders, including researchers, policymakers, conservationists, and educational institutions.

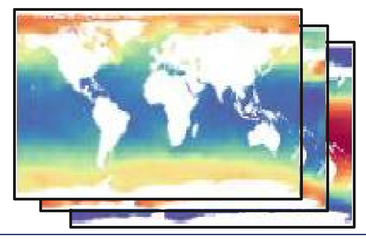
Accessible repositories (AtlantECO, OBIS Ocean Biodiversity Information System, GBIF Global Biodiversity Information Facility, MGnify, EMODnet European Marine Observation and Data Network)

Observations



Variable selection

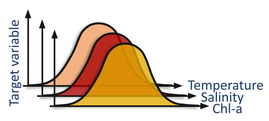
Oxygen and nutrients | Physics | Carbonate chemistry | Biology



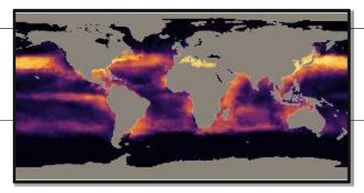
Model fitting

7 algorithms

- Generalised Linear Models, Generalised Additive Models, Random Forest, Multi Layers Perceptron, Boosted Regression Trees, Support Vector Machine
- **Multivariate Boosted Tree Regressor (proportion only)**

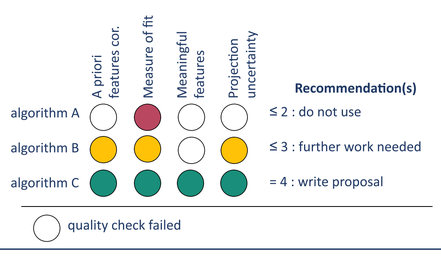


Model evaluation



Projection

Recommendations

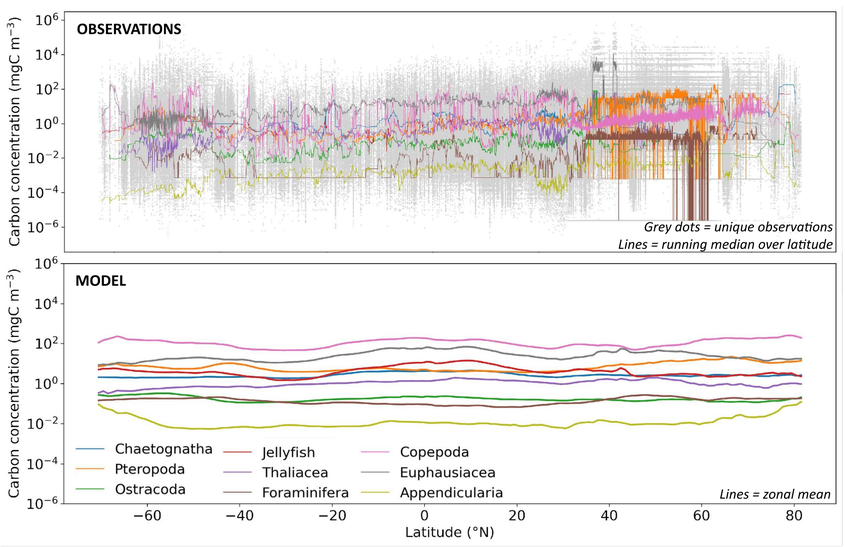


STEP(S)

1. **AVAILABLE BIOLOGICAL DATA** — Occurrence OR Biomass OR Metagenomic reads OR Custom
   - Binarisation
   - Shannon index
2. **BIOLOGICAL INPUT** TARGET — Binary OR Continuous OR Proportions
3. **ENVIRONMENT** FEATURES
   - Regrid to lowest common resolution
   - Average target value per cell x time
4. **PSEUDO ABSENCES**
   - Biased by presence density — Binary
5. **QUERY CHECK**
   - Target outliers
   - Correlated features
   - ***A priori* features correlation with target**
   - Multivariate Environmental Similarity Surfaces
6. **HYPERPARAMETER**
   - Hyper-parameter definition & tuning range
7. **FOLDS**
   - Training and testing folds
   - Evaluation fold
8. **CALIBRATION**
   - Hyper-parameter tuning
   - Range of algorithms
9. **EVALUATION**
   - **Measure of fit**
   - **Meaningful features**
10. **PROJECTION**
    - Full model
    - Bootstrap resampling
    - **Projection uncertainty**

    ENSEMBLE
    - Algorithm disagreement
11. **OUTPUTS**
    - Global maps of target variable
    - Feature selection & *a priori* importance
    - Feature importance
    - Algorithm-level & ensemble projections
    - Algorithm-level & ensemble uncertainty
    - Partial feature dependency
    - Summary .pdf

Example 1 : Evaluate zooplankton global biomass and distribution



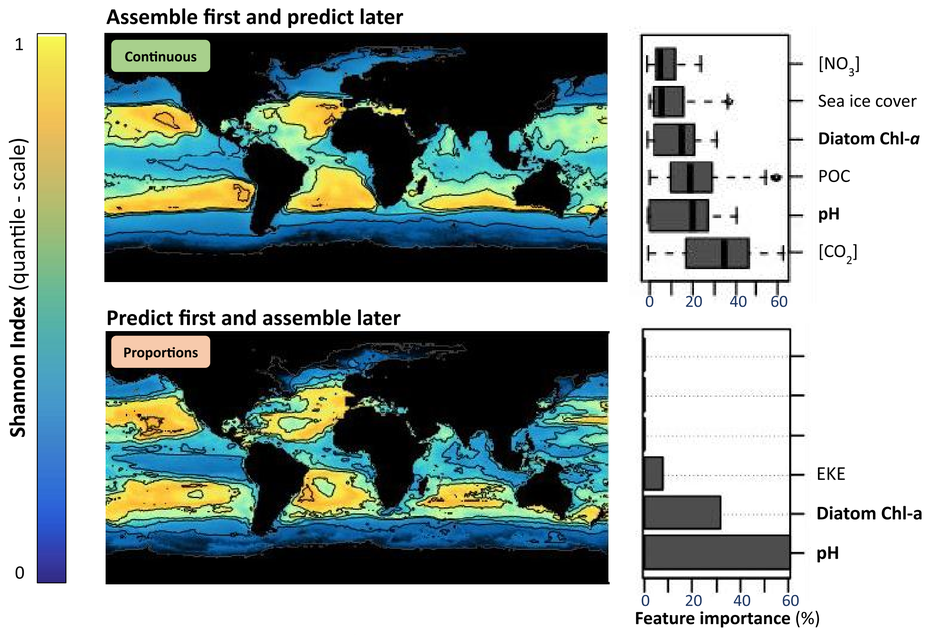
Pipeline input: biomass retrieved from traditional net zooplankton observations*

Pipeline ouput: extrapolated spatio temporal PFT-resolved plankton distribution

Applications: **Earth System Model evaluation** (PFT-specific), global **plankton ecology** (trophic pyramid, size spectrum, traits biogeography)

*Compilation of zooplankton observation from various methods (AtlanTECO-BASE). Zooplankton abundances (including observations from MAREDAT; Buitenhuis et al., 2013), the CPR surveys, BODC, COPEPOD, JeDI, KRILLBASE, Malaspina, Tara, among other efforts). 15,294,171 zooplankton abundance observations distributed in 10 plankton-functional types and 1,262 accepted taxonomic names. Conversion to biomasses based on a taxonomically resolved abundance to biomass conversion (unpublished).

Example 2 : Shannon diversity index for coccolithophores



Pipeline input : coccolitophores MAGs*

Comparison between **predict first** (model each species independently and then compute diversity) and **assemble first** (compute diversity within sample and then model diversity distribution)

Similar map, different environmental predictors: **diversity is an inherent property** of the system ?

Intercomparability of different data types within the same pipeline lead to new ecological insight about biodiversity and underlying processes

*MAG-level (i.e., Metagenome Assembled Genome; considered as species) relative genome reads from the MATOU catalogue (Marine Atlas of Tara Ocean Unigenes) among the 0.5 – 8 µm size filter, corresponding to the *Prymnesiophyceae* taxa.

Corentin

Alexandre

Funded by the European Union

@BlueCloudEU | company/blue-cloud-org | BlueCloudorg | blue-cloud.org