eosc | Blue-Cloud2026

WP3
developing & testing analytical Blue Cloud workbenches for generating highly qualified data collections (EOVs)

Dominique Obaton

IFREMER

Funded by
the European Union

WP3 workbenches

What ?

TOOLS

To build sustained pipelines that will enable integration and combination of datasets from various sources for further analysis as metadata homogenisation, duplicate management, quality control

and DATA COLLECTIONS as results

To illustrate what is possible through examples

To get first user feedback from “upgraded” datasets proposed. For a continuous improved loop

Why?

→ There is a need to improve different datasets: EMODnet, Copernicus, SeaDataNet, World Ocean Database (WOD2018), ELIXIR …

- ❖ Metadata
- ❖ Duplicates
- ❖ Quality control (QC)

To ensure whatever data collection an expert / scientist /data manager uses (can be a part of different datasets), QC is good & homogenous.

→ This will increase confidence of datasets usage

- ❖ whatever the source

→ Then this will increase the number of users

WP3 workbenches

How?

CHALLENGES

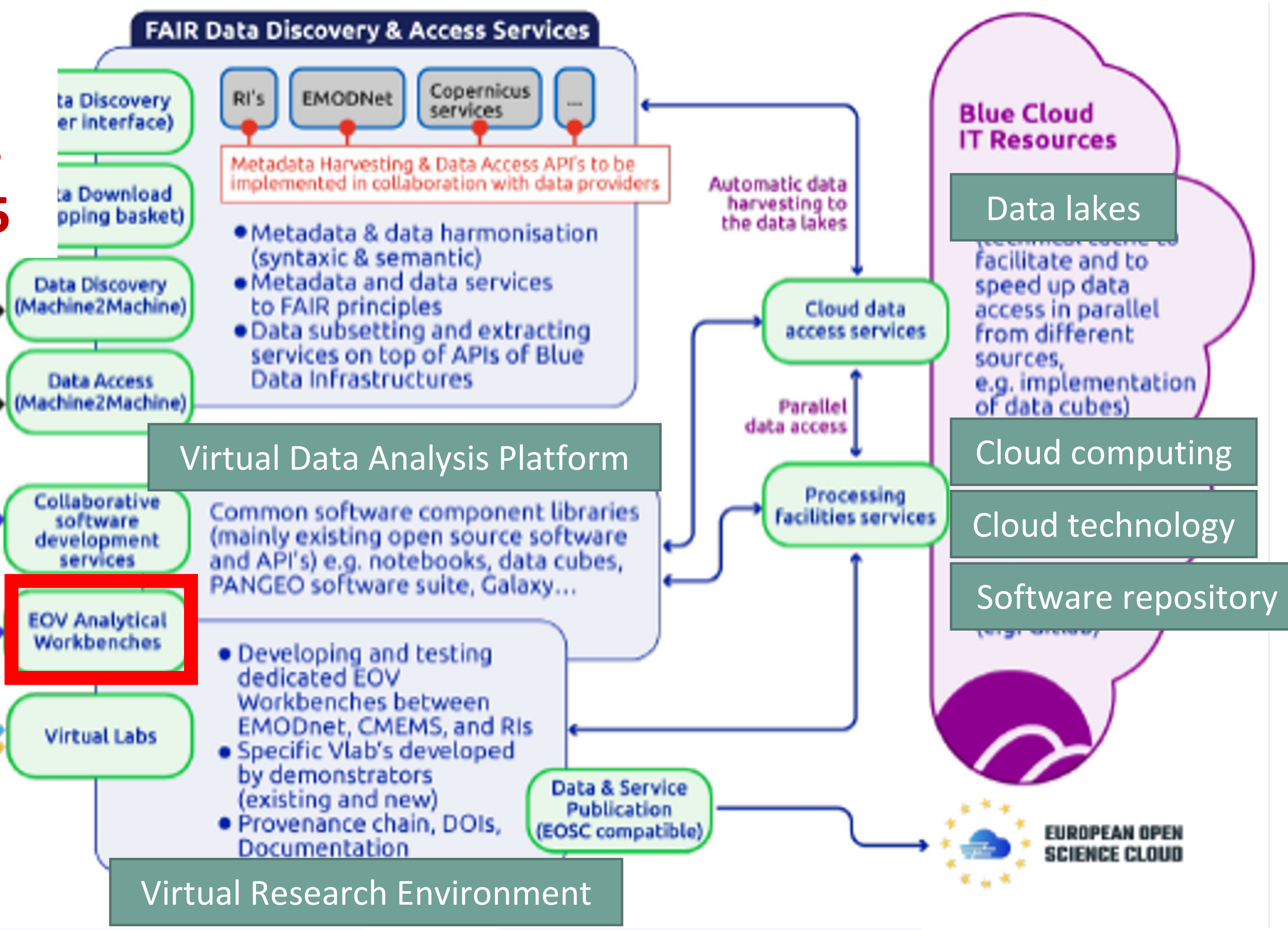
Big challenges and strong interactions with WP2 and WP5

→ Possible thanks to a Blue Cloud IT resources with

- Data lakes
- Cloud computing
- Cloud technology
- Software repository

→On a virtual environment

- Virtual data analysis platform
- Virtual research environment


FAIR Data Discovery & Access Services
ta Discovery
er interface)
ta Download
pping basket)
Data Discovery (Machine2Machine)
Data Access (Machine2Machine)
RI's
EMODNet
Copernicus services
...
Metadata Harvesting & Data Access API's to be implemented in collaboration with data providers
Metadata & data harmonisation (syntaxic & semantic)
Metadata and data services to FAIR principles
Data subsetting and extracting services on top of APIs of Blue Data Infrastructures
Virtual Data Analysis Platform
Collaborative software development services
EOV Analytical Workbenches
Virtual Labs
Common software component libraries (mainly existing open source software and API's) e.g. notebooks, data cubes, PANGEO software suite, Galaxy...
Developing and testing dedicated EOV Workbenches between EMODnet, CMEMS, and RIs
Specific Vlab's developed by demonstrators (existing and new)
Provenance chain, DOIs, Documentation
Data & Service Publication (EOSC compatible)
Virtual Research Environment
Automatic data harvesting to the data lakes
Cloud data access services
Parallel data access
Processing facilities services
Blue Cloud IT Resources
Data lakes
facilitate and to speed up data access in parallel from different sources, e.g. implementation of data cubes)
Cloud computing
Cloud technology
Software repository
EUROPEAN OPEN SCIENCE CLOUD

N.B. similar development in the FAIR-EASE project for the whole Earth's ecosystem. **To work in strong interactions**

eosc | Blue-Cloud2026

Blue Cloud 2026 challenges

What we are now able to do

Current constraints on a PC of 16GB memory

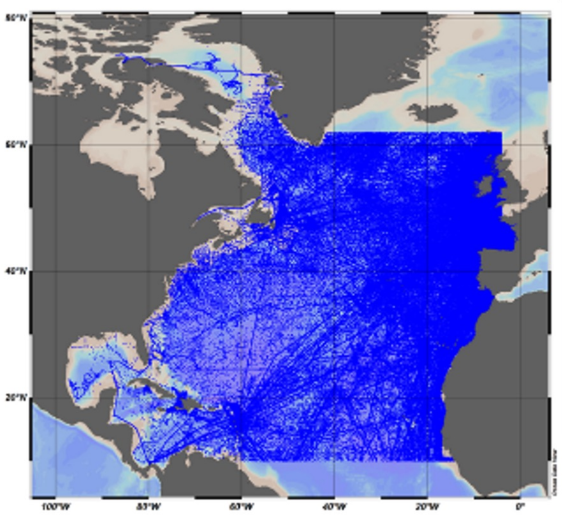
→ North Atlantic ocean

→ period: 1950-2019

→ SeaDataNet + Copernicus data collections

~ **5 GB**

- → **Limit of data volume to work on PC**
- → **Difficult to make analysis on such volume**



What we want to do (final objective)

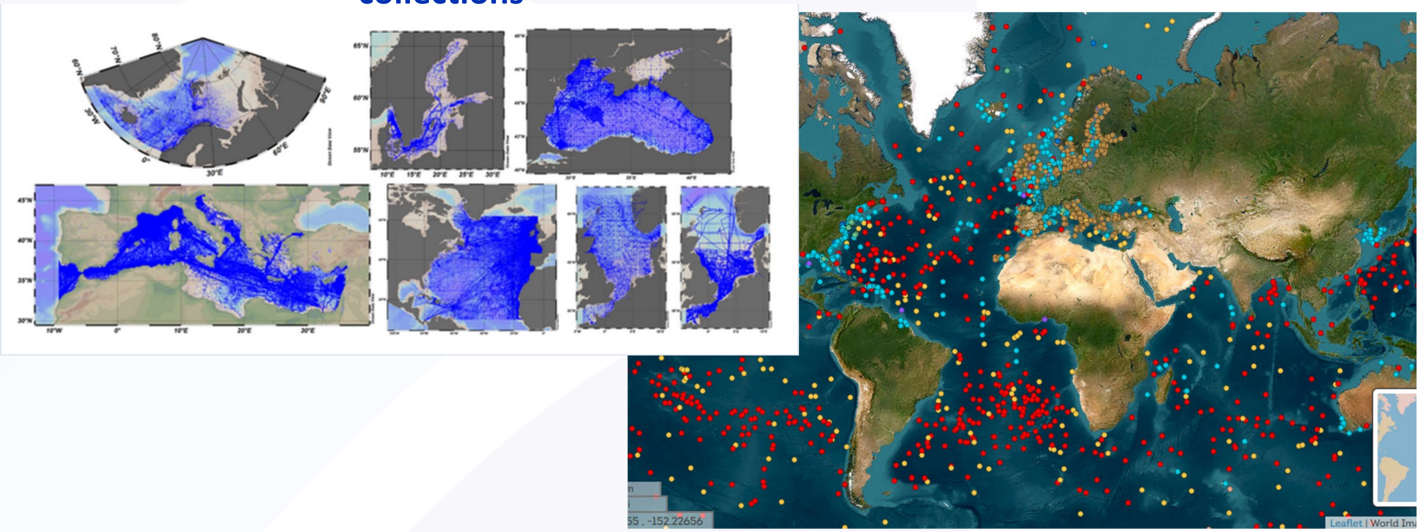
Global ocean

period: 1950-current time

SeaDataNet + Copernicus + WOD2018 datasets

= estimate **>20 GB**

- → **Possible thanks to Blue Cloud IT resources (Data lakes, processing services …)**
- → **That will offer the possibility to have the data at one place and make the analysis at the same place for large and important data collections**



WP3 workbenches

Why? → BENEFIT BLUE CLOUD 2026

→ Large data collections homogenised with a good quality control

- ❖ Not possible up to now
- ❖ Thanks to Blue Cloud IT including big data techno

→ Workbenches or pipelines developed for several variables (EOVs)

- ❖ Reusable for other / updated datasets
- ❖ Depending on specific validation wanted (e.g. open sea vs coastal)
- ❖ Portable on other e-infrastructure than Blue Cloud after the project

→ Workbenches are prototypes

- ❖ Than can be used in operational oceanography
- ❖ That will contribute to the building of the Digital Twin of the Ocean

Who?

13 partners altogether

Leader: IFREMER

Partners: AWI, CMCC, SSBE, INGV, Oceanscope, Pokapok, HCMR, OGS, SMHI, ETHZ, SU, EMBL

Through 4 tasks

For 209 months of people

During the whole project

WP3 workbenches -tasks

T3.1 coordination, definition and design

Lead: Ifremer Dominique Obaton

Partners: AWI Reiner Schlitzer*, CMCC* Massimiliano Drudi*, SSBE* Julia Vera

T3.2 EOV workbench for physics: temperature and salinity

Lead: INGV Simona Simoncelli

Partners: Ifremer Christine Coatanoan*, Oceanscope* Tanguy Szekely*, Pokapok* Jérôme Gourrion*, HCMR* Sissi Iona

T3.3 EOV workbench for eutrophication: chlorophyll, nutrients, oxygen.

Lead: OGS Alessandra Giorgetti

Partners: Ifremer Julie Gatti*, Pokapok* Virginie Racapé*, SMHI* Lindh Markus

T3.4 EOV workbench for ecosystems (genomic)

Lead: ETHZ Meike Vogt

Partners: SU Jean-Olivier Irisson*, EMBL* Stéphane Pesant

Task 3.2 Workbench for physics

T3.2 EOV workbench for physics: temperature and salinity

Lead: INGV Simona Simoncelli

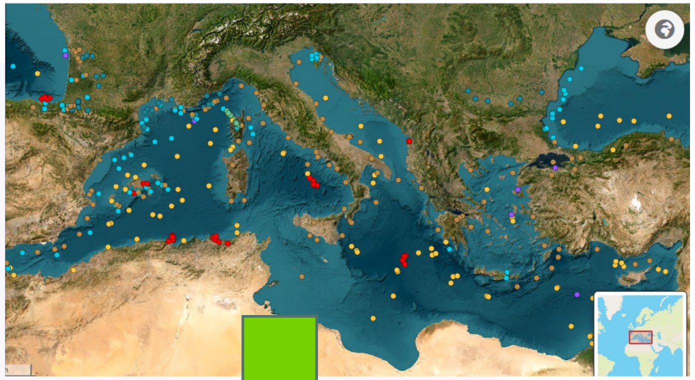
Partners: Ifremer Christine Coatanoan, *Oceanscope* Tanguy Szekely, *Pokapok* Jérôme Gourrion, *HCMR* Sissi Iona

→ Datasets: SeaDataNet + Copernicus physics + WOD2018 datasets

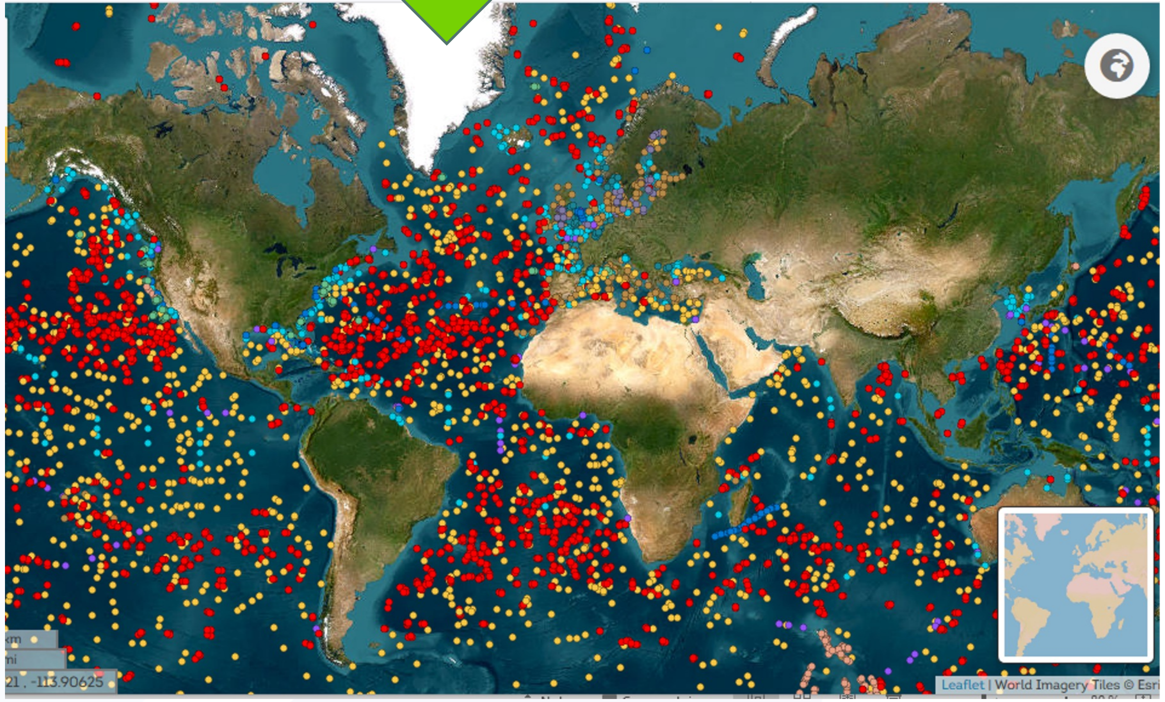
→ Variables: Temperature & salinity

→ Methods: ODV, ISAS, MinMax

→ Start with the Mediterranean Sea, go to the global ocean



Mediterranean Sea Plot: 10 days, 400 data.
All database, 1 million data



Plot: 10 days, 16500 data. All database, 30 millions data

Task 3.3 Workbench for eutrophication

T3.3 EOV workbench for eutrophication: chlorophyll, nutrients, oxygen

Lead: OGS Alessandra Giorgetti

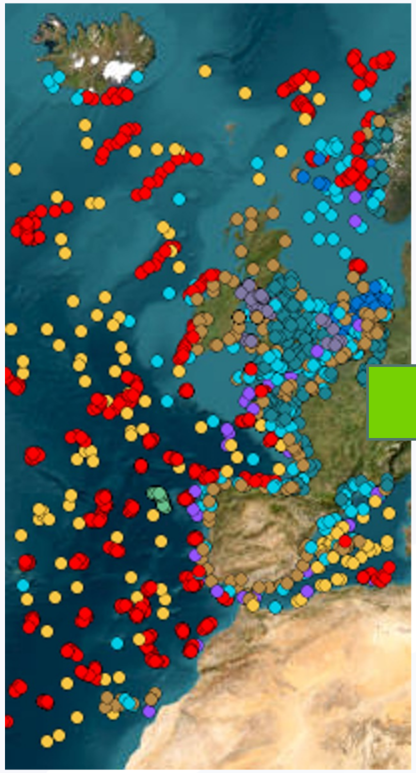
Partners: Ifremer Julie Gatti, *Pokapok* Virginie Racapé, *SMHI* Lindh Markus

→ Datasets: EMODnet chemistry + Copernicus +WOD2018 datasets

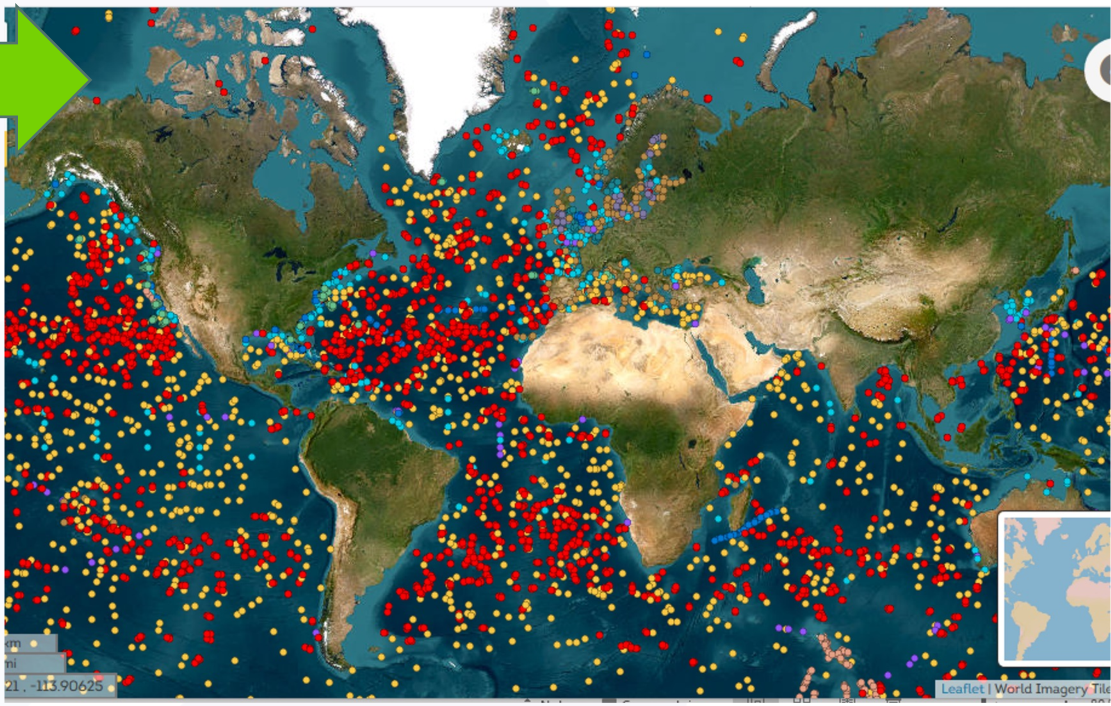
→ Variables: chlorophyll, nutrients, oxygen

→ Methods: ODV, Copernicus procedures

→ Start with the North East Atlantic, go to the global ocean



North East Atlantic
Plot: 10 days, 1790 data.
All database, 2,4 millions data



Plot: 10 days, 16500 data. All database, 30 millions data

Task 3.4 Workbench for ecosystem

T3.4 EOV workbench for ecosystems (genomic)
Lead: ETHZ Meike Vogt
Partners: SU Jean-Olivier Irisson, *EMBL* Stéphane Pesant

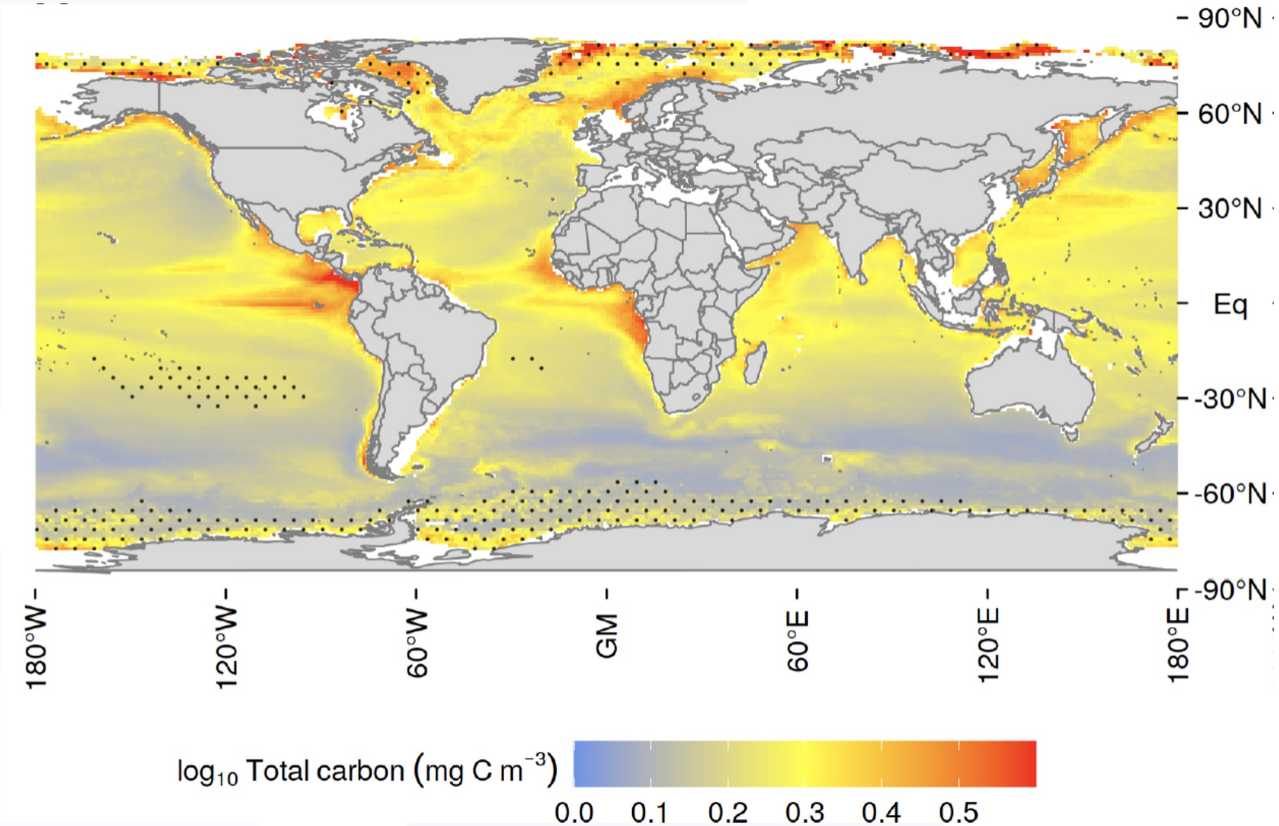
→ Datasets: large collections of plankton observations EMODnet biology/EurOBIS, ELIXIR

→ Variables: plankton diversity, biogeography, biomass and relative abundance

→ Methods: Mgnify, EcoTaxa

→ Global ocean

Aggregation and harmonization of large collections of plankton observations based on traditional counts, quantitative imaging and genomic methods



Task 3.1, within the WP

T3.1 coordination, definition and design

Lead: Ifremer Dominique Obaton

Partners: AWI Reiner Schiltzer, *CMCC* Massimiliano Drudi, *SSBE* Julia Vera

Within the workpackage

→ Coordination

→ Common points, developments between tasks, especially methods in workbenches 1 and 2

+

With other workpackages

→ Strong dependency of WP2 (data lakes, Virtual Data Analysis Platform -VDAP) and WP5 (Virtual Research Environment -VRE)

→ Highly validated datasets from workbenches + some applications/use for/within the use cases of developed in WP4

Task 3.1, outside WP

+

→ Follow and liaise with project coordination (WP1)

→ Contribution to the communication – outreach & engagement (WP6)

→ Contribute to the exploitation and sustainability (WP7) ***Big challenge***

Deliverables

Deliverable (number)	Deliverable name & description	WP number	Delivery date
D3.1	**1st release of aggregated and harmonised EOV datasets**: preliminary aggregated and harmonised EOV datasets obtained thanks to each workbench.	WP3	M20
D7.1	**Individual Exploitation Plans of Workbenches**: A compilation of the individual Exploitation Plans of each of the Workbenches set up in WP3	WP7	M24
D3.2	**Final release of open aggregated and harmonised EOV datasets**: final and open aggregated and harmonised EOV datasets stamped BC2026 obtained thanks to the workbenches.	WP3	M42
D3.3	**Workbenches and tools (notebooks, model, containers with instructions)**: Portable prototypes for operational services and data infrastructures.	WP3	M42

Workbenches task leaders presentations

Task 3.2 workbench for physics: temperature and salinity

Simona Simoncelli, INGV

First task meeting (remote) before Technical Scientific Committee: 28 February at 14:00 CET

Task 3.3 workbench for eutrophication: chlorophyll, nutrients, oxygen

Alessandra Giorgetti, OGS

First task meeting (remote) before TSC: 3rd March at 11.00 am CET

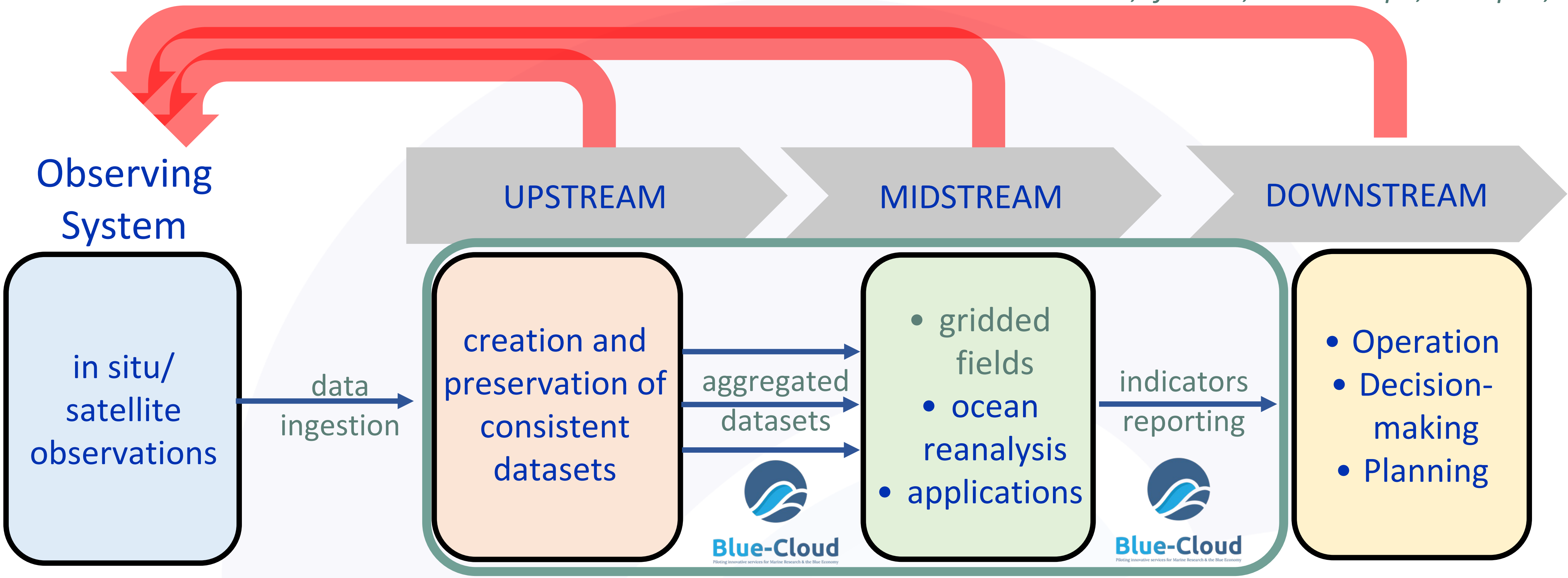
Task 3.4 workbench for ecosystems

Meike Vogt, ETHZ

First in person meetings (a) between ETHZ-EMBL: January 10-13, 2023; (b) between ETHZ-SU: February 6-9, 2023

Task 3.2 workbench for physics: temperature and salinity

***INGV**, Ifremer, Oceanscope, Pokapok, HCMR*



Observing System

UPSTREAM

MIDSTREAM

DOWNSTREAM

in situ/ satellite observations

data ingestion

creation and preservation of consistent datasets

aggregated datasets

- gridded fields
- ocean reanalysis
- applications

indicators reporting

- Operation
- Decision-making
- Planning

Blue Data Infrastructures

Copernicus Marine Service

SeaDataNet

EMODnet

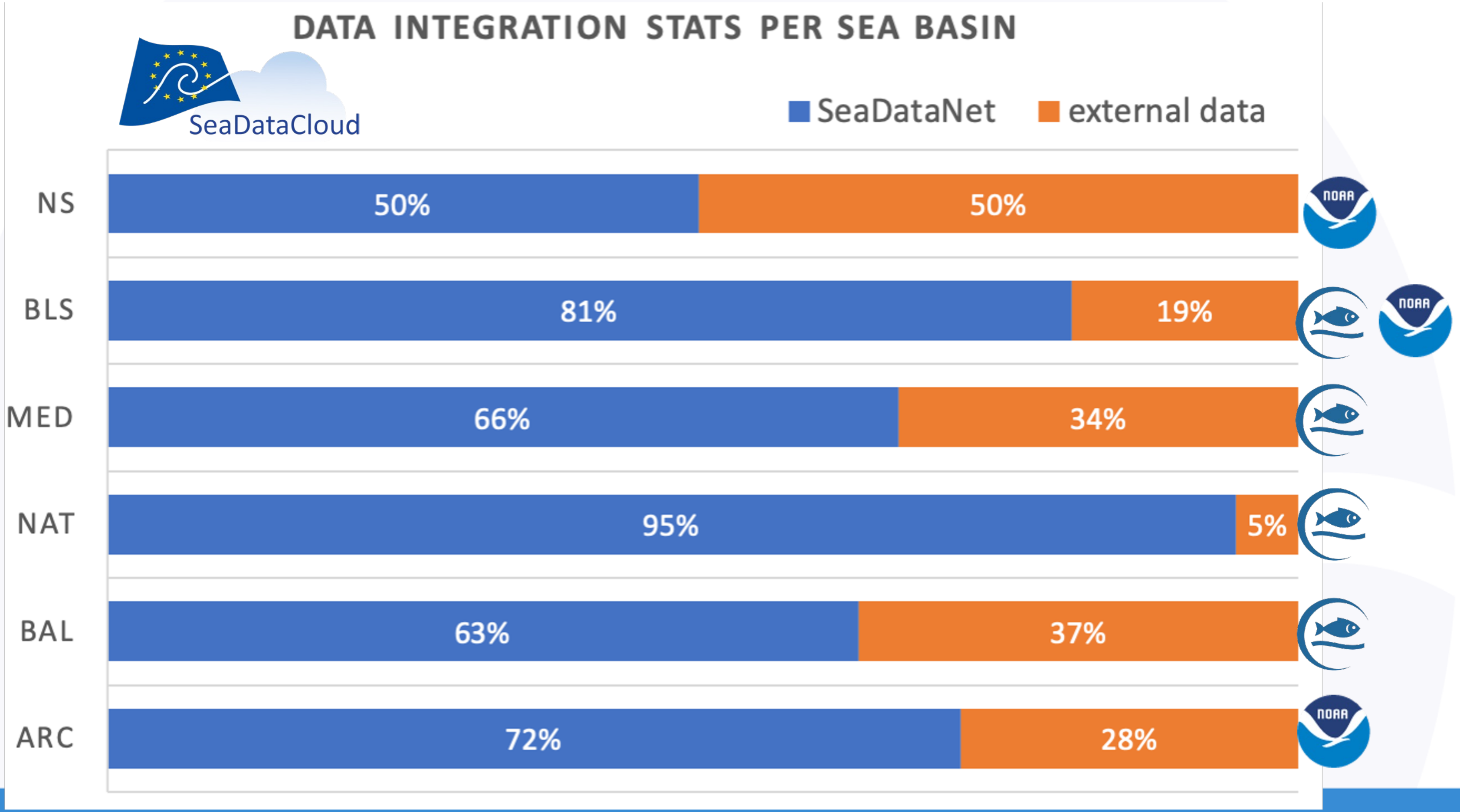
NOAA

Simoncelli et al. 2022 In *Ocean Science Data*
https://doi.org/10.1016/B978-0-12-823427-3.00001-3

Task 3.2 workbench for physics: temperature and salinity

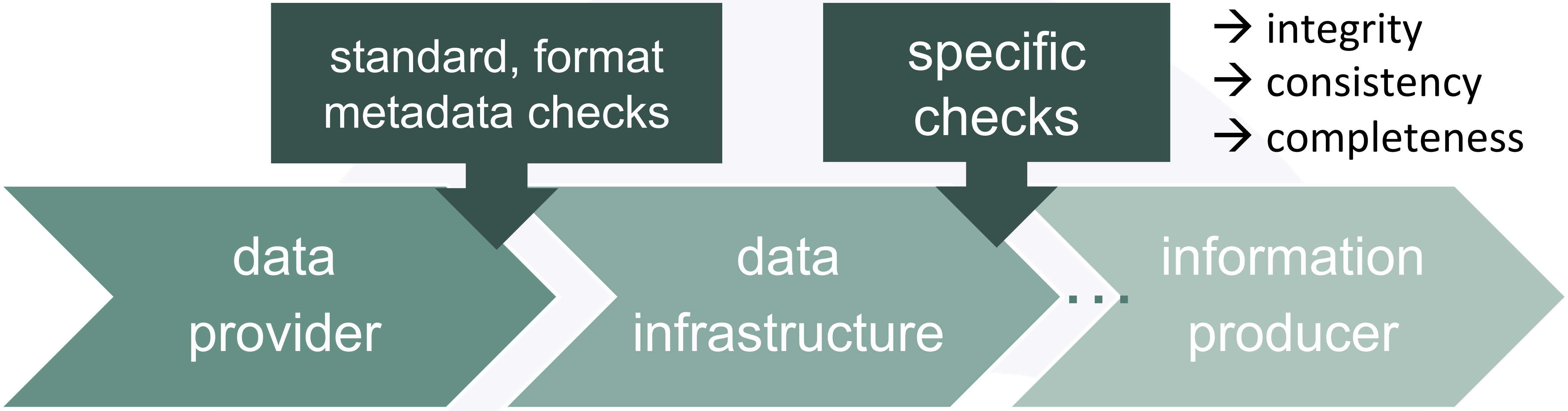
__INGV__, Ifremer, Oceanscope, Pokapok, HCMR

Integration of SeaDataNet T&S data with other data sources (Copernicus, NOAA-WOD2018) → **duplicates removal**



Task 3.2 workbench for physics: temperature and salinity

INGV, *Ifremer, Oceanscope, Pokapok, HCMR*



standard, format metadata checks

specific checks

→ integrity
→ consistency
→ completeness

data provider

data infrastructure

...

information producer

Primary QA/QC

Secondary QA/QC

Product QA/QC

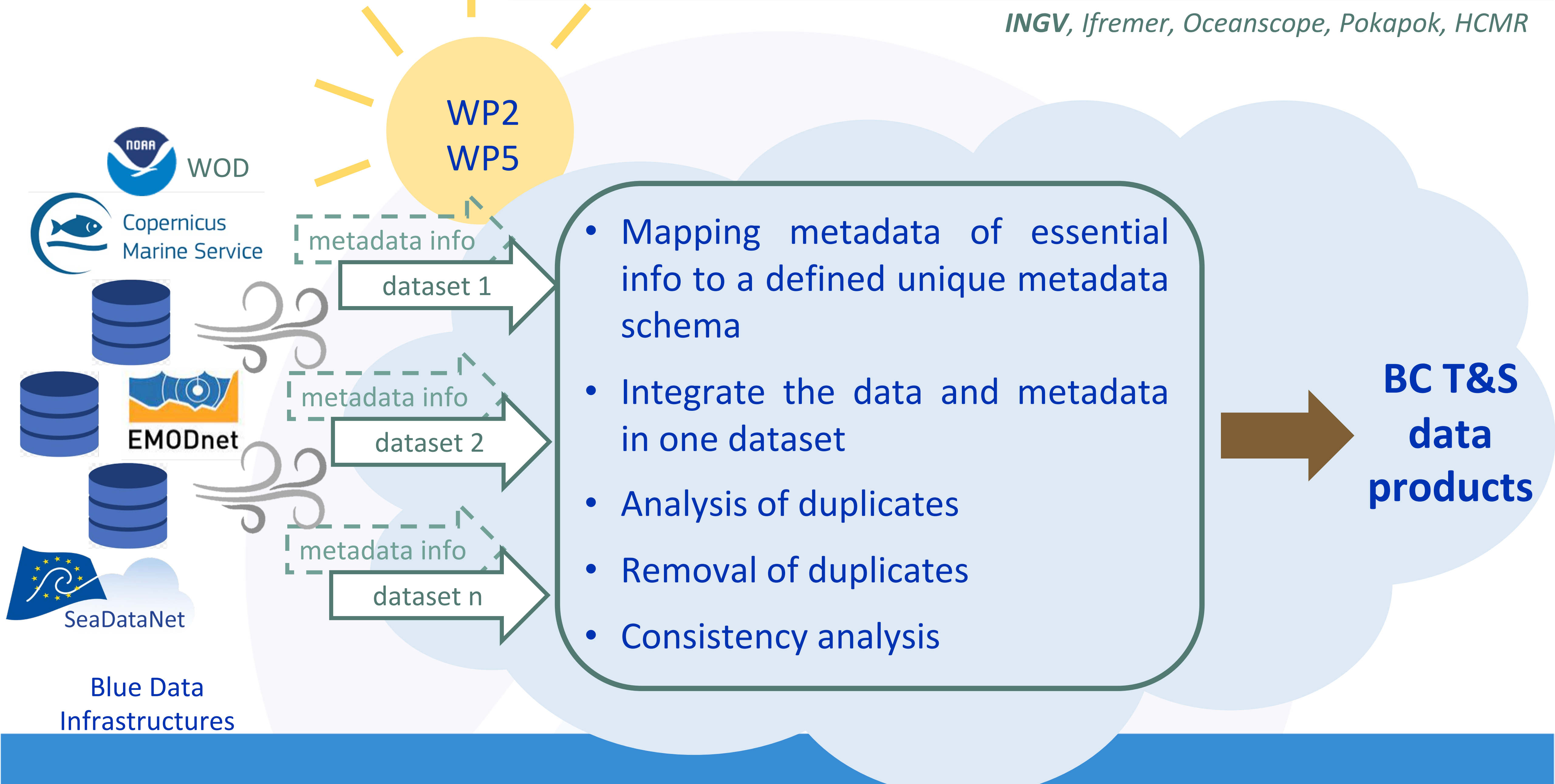
→ several actors apply QA/QC procedures along with the data lifecycle and value chain

→ provenance and quality information are key elements to preserve

Simoncelli et al. 2022 In *Ocean Science Data* https://doi.org/10.1016/B978-0-12-823427-3.00001-3

Task 3.2 phase 1: dataset integration

***INGV**, Ifremer, Oceanscope, Pokapok, HCMR*



- Mapping metadata of essential info to a defined unique metadata schema
- Integrate the data and metadata in one dataset
- Analysis of duplicates
- Removal of duplicates
- Consistency analysis

BC T&S data products

eosc | Blue-Cloud2026

Task 3.2 phase 2: workflow development

INGV, Ifremer, Oceanscope, Pokapok, HCMR

WP2
WP5

integrate data and metadata

- interactive data navigation and visualization
- interactive duplicate check tool

jupyter **ML/AI**

Duplicates

analysis and removal

Quality Control

consistency

analysis

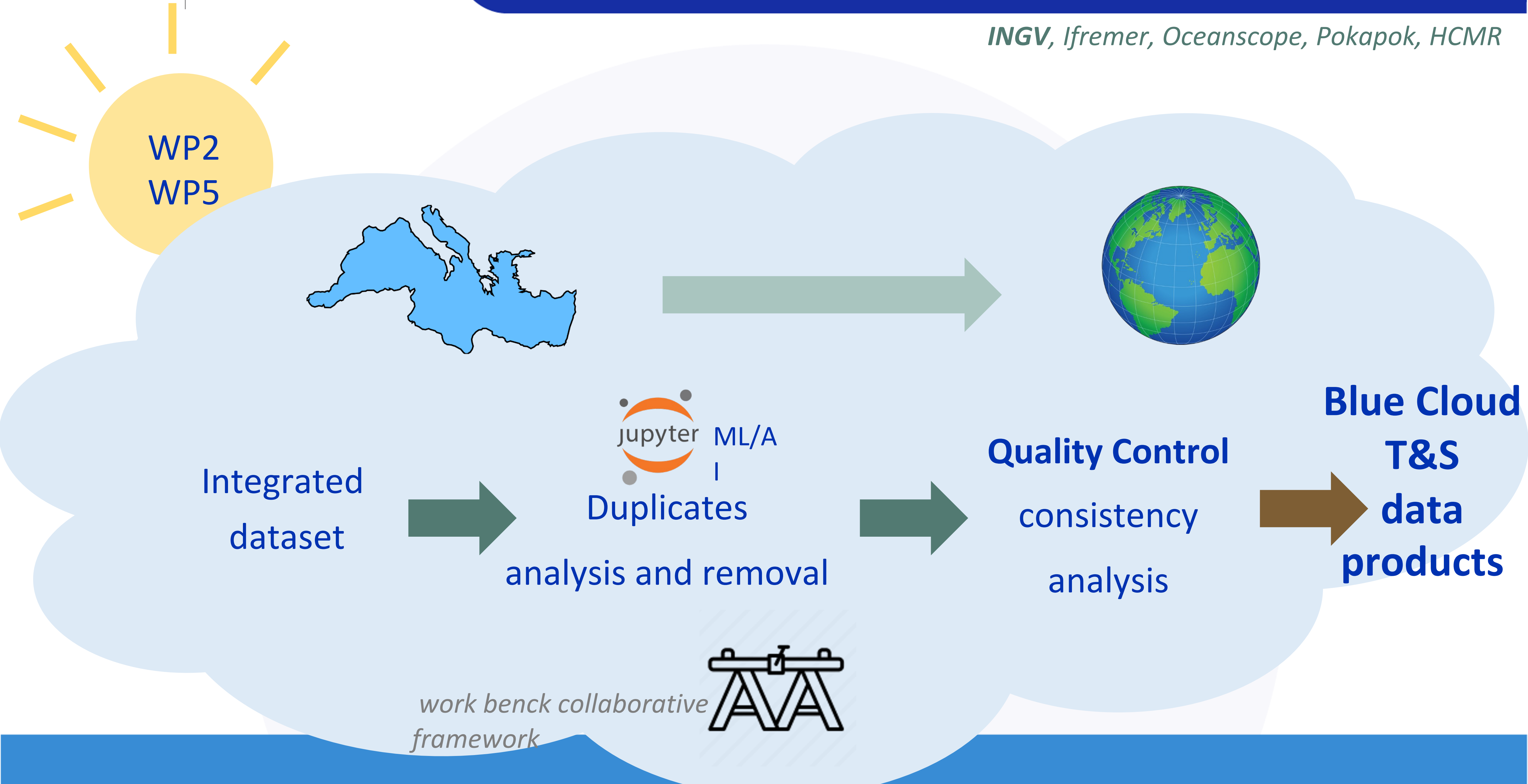
BlueCloud T&S data products

learning on collaborative framework

work benck collaborative framework

Task 3.2 phase 3: testing and optimisation

__INGV__, Ifremer, Oceanscope, Pokapok, HCMR



WP2
WP5

Integrated dataset

jupyter ML/A
I

Duplicates

analysis and removal

Quality Control

consistency

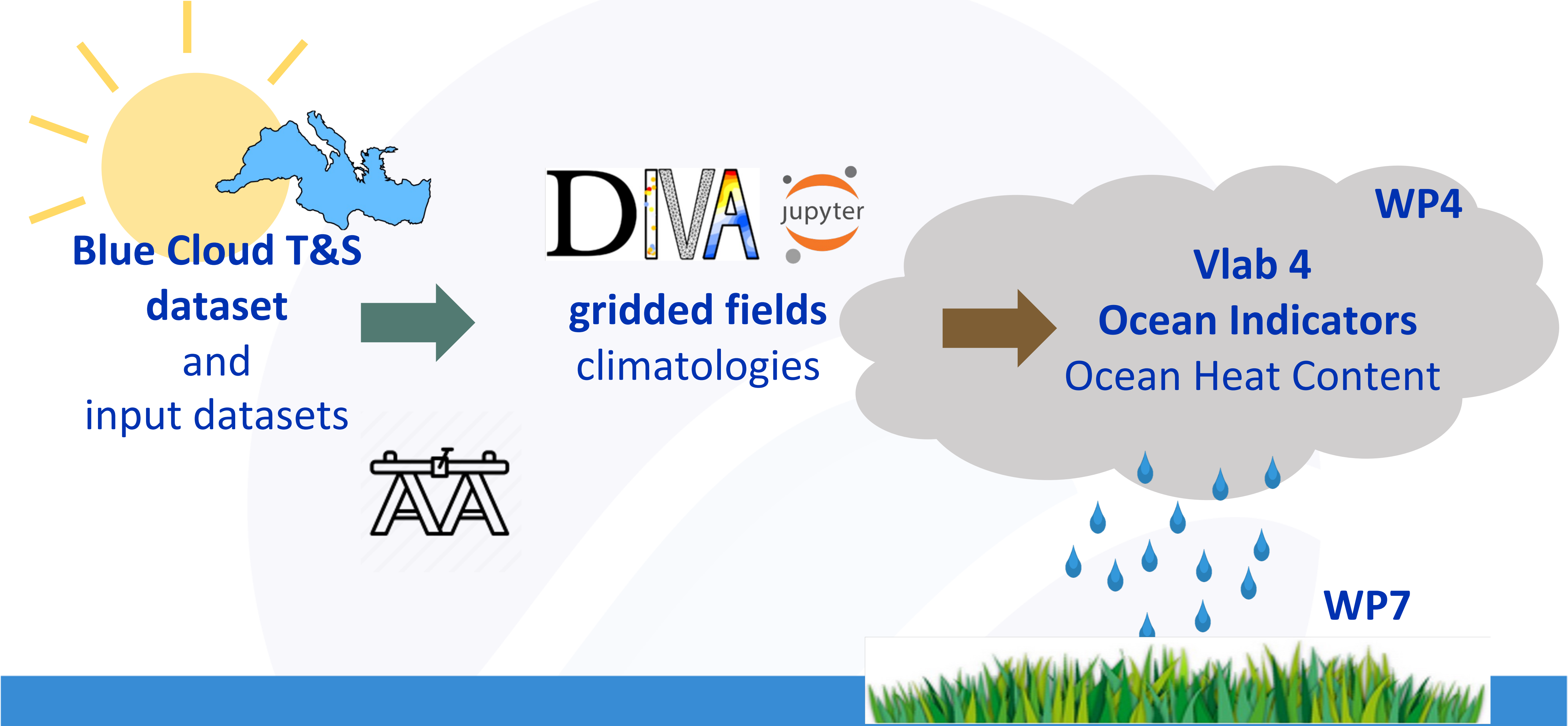
analysis

Blue Cloud T&S data products

work benck collaborative framework

Task 3.2 phase 4: verification and downstream

***INGV**, Ifremer, Oceanscope, Pokapok, HCMR*



Blue Cloud T&S dataset
and
input datasets

gridded fields
climatologies

WP4

Vlab 4
Ocean Indicators
Ocean Heat Content

WP7

Workbench 2: aims and methodology

***OGS**, Ifremer, Pokapok, SMHI*

Ambition and scope:

- Providing a toolbox, configurable by the user, to **build customizable datasets from different combinations of collections** or QC procedures and to compute various statistical parameters to assess the consistency of observed and derived quantities.

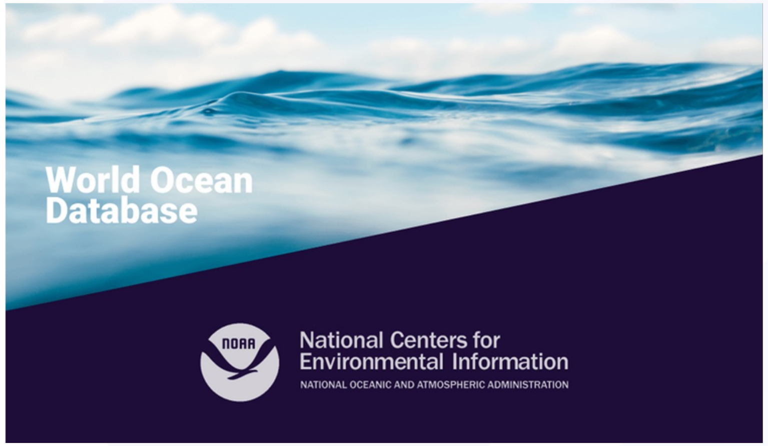
How:

- Definition and implementation of an efficient production **workflow that will merge multi-source datasets** to obtain an integrated and most complete dataset for the North East Atlantic.
- **Comparison** of EMODnet chemistry and Copernicus marine in-situ **QC methods** to provide a set of procedures depending on regions and/or user's aims.
- **Testing** the statistical parameters to i) evaluate the consistency of the initial input dataset and ii) to compare the data selection obtained after applying different QC strategies.

Data sources

***OGS**, Ifremer, Pokapok, SMHI*

- **Currently** several datasets in different data infrastructures with their own QC/QA are synchronised manually. These need to be further integrated within the Blue-Cloud 2026 data infrastructure.


World Ocean
Database
NOAA
National Centers for
Environmental Information
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Task 3.3: aim and methods

***OGS**, Ifremer, Pokapok, SMHI*

Methods:

- ODV, DIVAnd, Copernicus procedures.

Description:

- The task will aim at **generating harmonised and validated EOV data collections** for Chlorophyll, nutrients and oxygen integrating several datasets released from different EU and non-EU data infrastructures.

- The information associated with the observations, metadata, will be **mapped** into a common schema and the quality control procedures will be **exchanged**. A dedicated protocol will be jointly set out to identify and handle the potential duplicate observations.

- The Workbench will be **developed and tested for the Northeast Atlantic Sea** with the aim of further **extending** it to the global ocean and will be made available for implementation by other data infrastructures.

Technical requirements

OGS*, Ifremer, Pokapok, SMHI*

- ODV, DIVAnd, Copernicus procedures **will need further developments** to support the interaction with chosen technologies to handle the large volumes of data the WorkBenches will have to deal with (Task 3.1)

- IT functionalities and FAIRness of **BlueCloud services** (WP2 & WP5) will need to be used for dataset integration and analysis.

- **Interoperability services** of the data infrastructures, common vocabularies and brokering services will be needed to allow dataset aggregation and harmonisation.

- To be further discussed at the **task meeting planned for March 3rd, 2023 at 11.00 am CET** & in Technical and Scientific Committee (TSC) meeting in Amsterdam fixed for 28-29 March 2023

Task 3.4: Aims and methods

***ETHZ**, SU, EMBL*

Aims:

- Improve **the availability, quality and interoperability of large collections of plankton observations** based on traditional counts, quantitative imaging and genomic methods available from the EMODnet/EurOBIS and ELIXIR data infrastructures

- Develop **a generic machine learning modelling pipeline** in Blue-Cloud 2026 to generate high-quality interpolated maps of the global distribution of plankton biogeography and diversity (binary data; presence-absence), biomass (quantitative data; e.g. biomass) and community composition (relative abundance; e.g. percent reads)

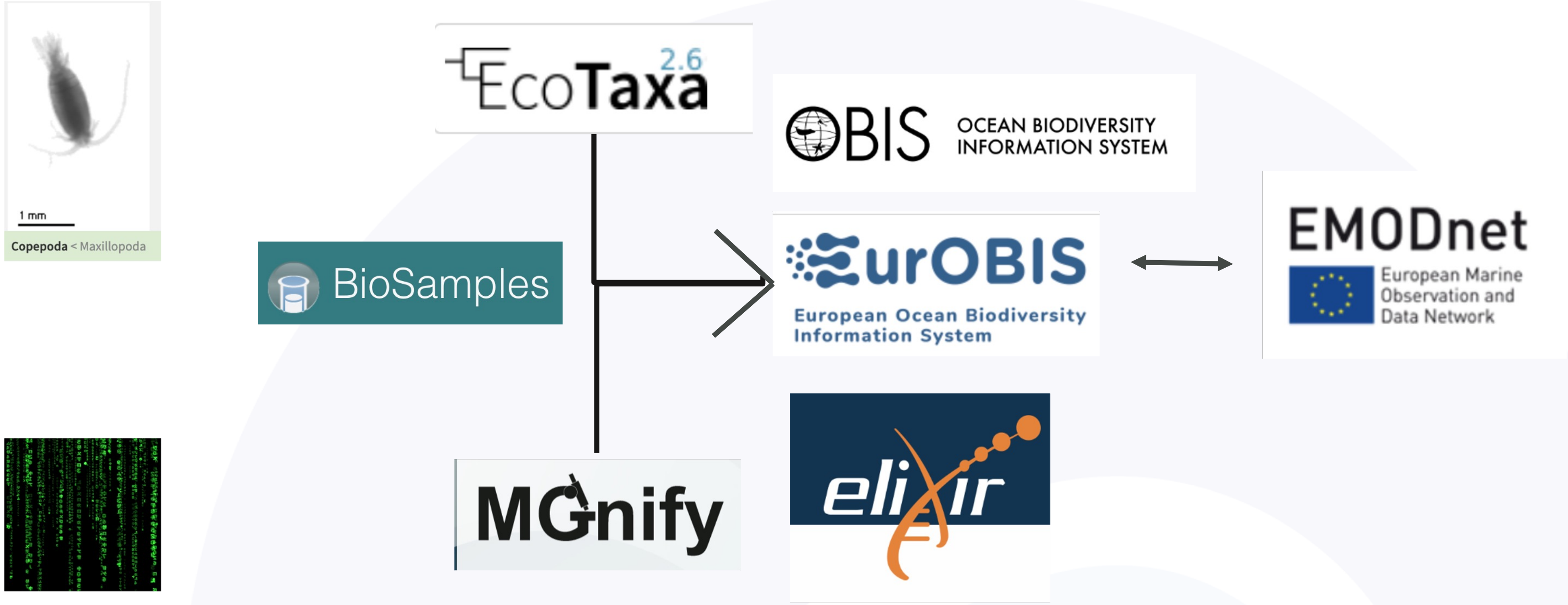
- Provide objective, multi-criteria quality assessments of output layers of added value based on based on multi-model ensemble projections

Data and Methods:

- Repositories: EurOBIS, OBIS, GBIF - plankton data from traditional methods, imaging and ‘omics
- Data resources/analysis pipelines/infrastructures: MGnify, Ecotaxa, BioSamples (EMODnet)
- Methods: Species distribution modelling (common machine learning tools)

Task 3.4: Implementation strategy - Data

ETHZ, ***SU, EMBL***

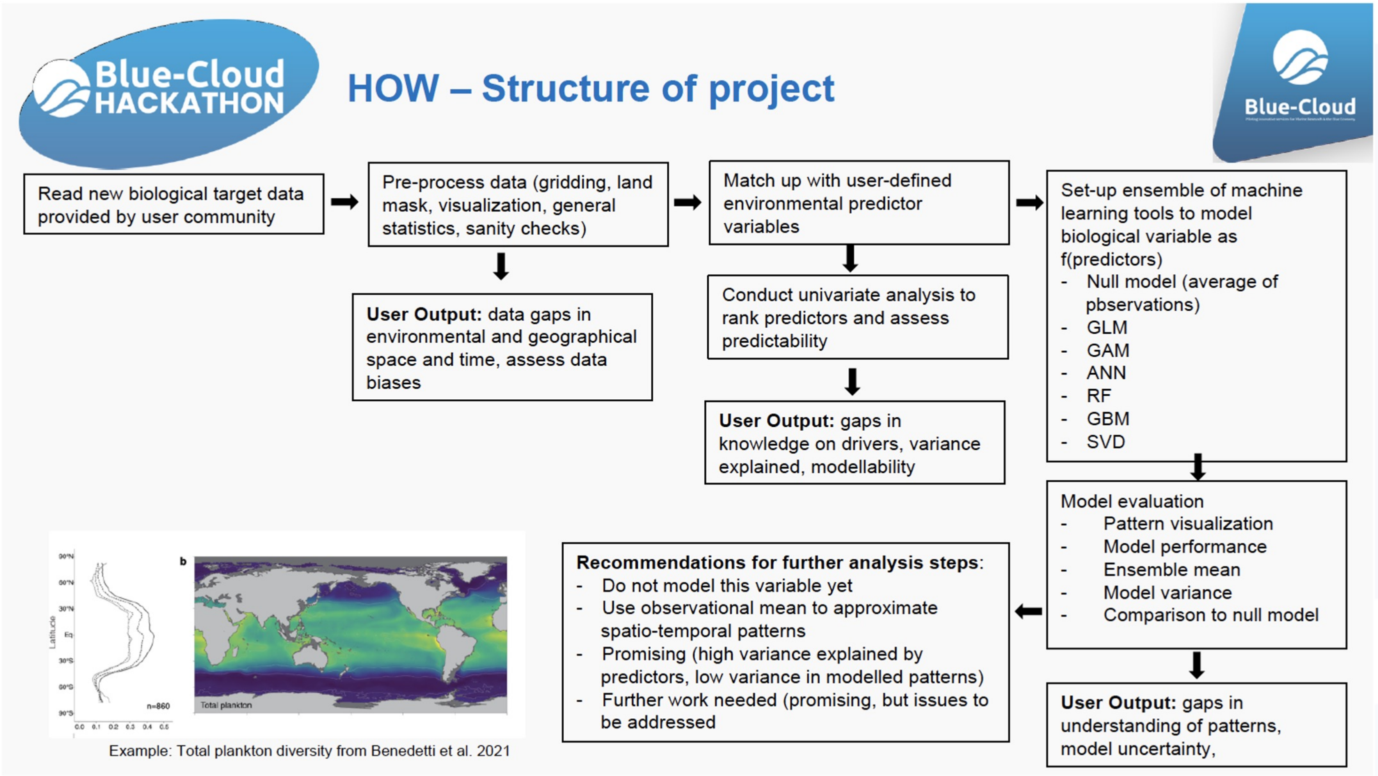

1 mm
Copepoda < Maxillopoda
EcoTaxa 2.6
BioSamples
MGnify
OBIS OCEAN BIODIVERSITY INFORMATION SYSTEM
EurOBIS European Ocean Biodiversity Information System
EMODnet European Marine Observation and Data Network
elixir

Facilitated data access and increased biological data availability:

- Improved **exchange of interoperable plankton data** between EcoTaxa/MGnify and key global data repositories (e.g. EurOBIS) and research infrastructures (e.g. EMODnet)
- **Provenance metadata curation pipeline** (BioSamples) for omics data

Task 3.4: Implementation strategy - Models

ETHZ, SU, *EMBL*



HOW – Structure of project

Read new biological target data provided by user community →

Pre-process data (gridding, land mask, visualization, general statistics, sanity checks) →

Match up with user-defined environmental predictor variables →

Set-up ensemble of machine learning tools to model biological variable as f(predictors)
- Null model (average of pbservations)
- GLM
- GAM
- ANN
- RF
- GBM
- SVD

User Output: data gaps in environmental and geographical space and time, assess data biases

Conduct univariate analysis to rank predictors and assess predictability

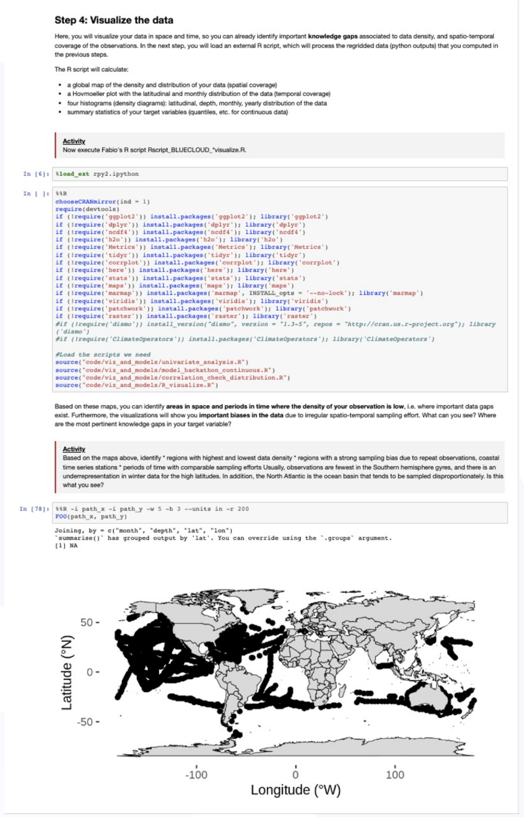
User Output: gaps in knowledge on drivers, variance explained, modellability

Model evaluation
- Pattern visualization
- Model performance
- Ensemble mean
- Model variance
- Comparison to null model

User Output: gaps in understanding of patterns, model uncertainty,

Recommendations for further analysis steps:
- Do not model this variable yet
- Use observational mean to approximate spatio-temporal patterns
- Promising (high variance explained by predictors, low variance in modelled patterns)
- Further work needed (promising, but issues to be addressed)

Example: Total plankton diversity from Benedetti et al. 2021



Step 4: Visualize the data

Here, you will visualize your data in space and time, so you can already identify important **knowledge gaps** associated to data density, and spatio-temporal coverage of the observations. In the next step, you will load an external R script, which will process the regridded data (python outputs) that you computed in the previous steps.

The R script will calculate:

- a global map of the density and distribution of your data (spatial coverage)
- a Hovmoeller plot with the latitudinal and monthly distribution of the data (temporal coverage)
- four histograms (density diagrams): latitudinal, depth, monthly, yearly distribution of the data
- summary statistics of your target variables (quantiles, etc. for continuous data)

Activity
Now execute Fabio's R script Rscript_BLUECLOUD_*visualize.R.

```
In [6]: %load_ext rpy2.ipython
```

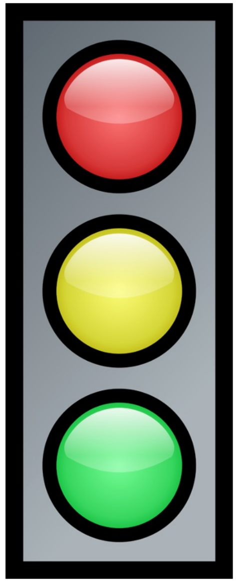
```
In [ ]: %%R
chooseCRANmirror(ind = 1)
require(devtools)
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
if (!require('ncdf4')) install.packages('ncdf4'); library('ncdf4')
if (!require('h2o')) install.packages('h2o'); library('h2o')
if (!require('Metrics')) install.packages('Metrics'); library('Metrics')
if (!require('tidyr')) install.packages('tidyr'); library('tidyr')
if (!require('corrplot')) install.packages('corrplot'); library('corrplot')
if (!require('here')) install.packages('here'); library('here')
if (!require('stats')) install.packages('stats'); library('stats')
if (!require('maps')) install.packages('maps'); library('maps')
if (!require('marmap')) install.packages('marmap', INSTALL_opts = '--no-lock'); library('marmap')
if (!require('viridis')) install.packages('viridis'); library('viridis')
if (!require('patchwork')) install.packages('patchwork'); library('patchwork')
if (!require('raster')) install.packages('raster'); library('raster')
#if (!require('dismo')) install_version("dismo", version = "1.3-5", repos = "http://cran.us.r-project.org"); library
('dismo')
#if (!require('ClimateOperators')) install.packages('ClimateOperators'); library('ClimateOperators')

#Load the scripts we need
source("code/viz_and_models/univariate_analysis.R")
source("code/viz_and_models/model_hackathon_continuous.R")
source("code/viz_and_models/correlation_check_distribution.R")
source("code/viz_and_models/R_visualize.R")
```

Based on these maps, you can identify **areas in space and periods in time where the density of your observation is low**, i.e. where important data gaps exist. Furthermore, the visualizations will show you **important biases in the data** due to irregular spatio-temporal sampling effort. What can you see? Where are the most pertinent knowledge gaps in your target variable?

Activity
Based on the maps above, identify * regions with highest and lowest data density * regions with a strong sampling bias due to repeat observations, coastal time series stations * periods of time with comparable sampling efforts Usually, observations are fewest in the Southern hemisphere gyres, and there is an underrepresentation in winter data for the high latitudes. In addition, the North Atlantic is the ocean basin that tends to be sampled disproportionately. Is this what you see?

```
In [78]: %%R -i path_x -i path_y -w 5 -h 3 --units in -r 200
FOO(path_x, path_y)
```

```
Joining, by = c("month", "depth", "lat", "lon")
`summarise()` has grouped output by 'lat'. You can override using the `.groups` argument.
[1] NA
```

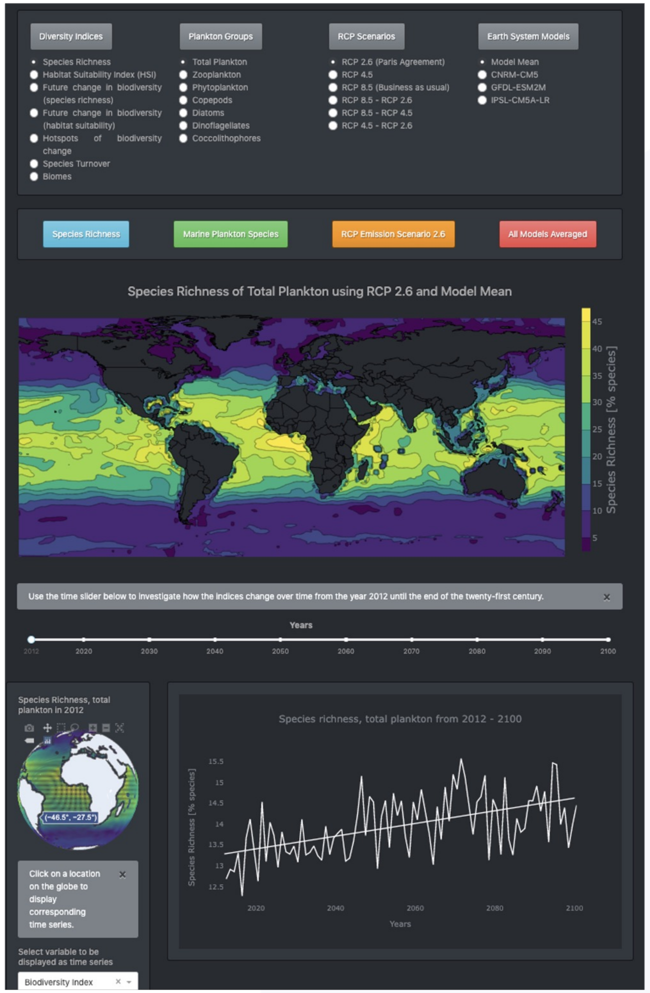


Prototype modelling pipeline developed during BlueCloud hackathon 2021 (proof-of-concept):

- **Automatized, ensemble-based modelling/machine learning pipeline** for the projections of binary (presence-absence), continuous quantitative (biomass) and proportional (percent genetic reads) data
- Rigorous quality-control of output products using **a multi-metric quality assessment**

Task 3.4: Output products - Layers of added value (EOVs)

ETHZ, SU, EMBL



Diversity Indices

- ● Species Richness
- ○ Habitat Suitability Index (HSI)
- ○ Future change in biodiversity (species richness)
- ○ Future change in biodiversity (habitat suitability)
- ○ Hotspots of biodiversity change
- ○ Species Turnover
- ○ Biomes

Plankton Groups

- ● Total Plankton
- ○ Zooplankton
- ○ Phytoplankton
- ○ Copepods
- ○ Diatoms
- ○ Dinoflagellates
- ○ Coccolithophores

RCP Scenarios

- ● RCP 2.6 (Paris Agreement)
- ○ RCP 4.5
- ○ RCP 8.5 (Business as usual)
- ○ RCP 8.5 - RCP 2.6
- ○ RCP 8.5 - RCP 4.5
- ○ RCP 4.5 - RCP 2.6

Earth System Models

- ● Model Mean
- ○ CNRM-CM5
- ○ GFDL-ESM2M
- ○ IPSL-CM5A-LR

Species Richness | Marine Plankton Species | RCP Emission Scenario 2.6 | All Models Averaged

Species Richness of Total Plankton using RCP 2.6 and Model Mean

Use the time slider below to investigate how the indices change over time from the year 2012 until the end of the twenty-first century. ×

Years

2012 2020 2030 2040 2050 2060 2070 2080 2090 2100

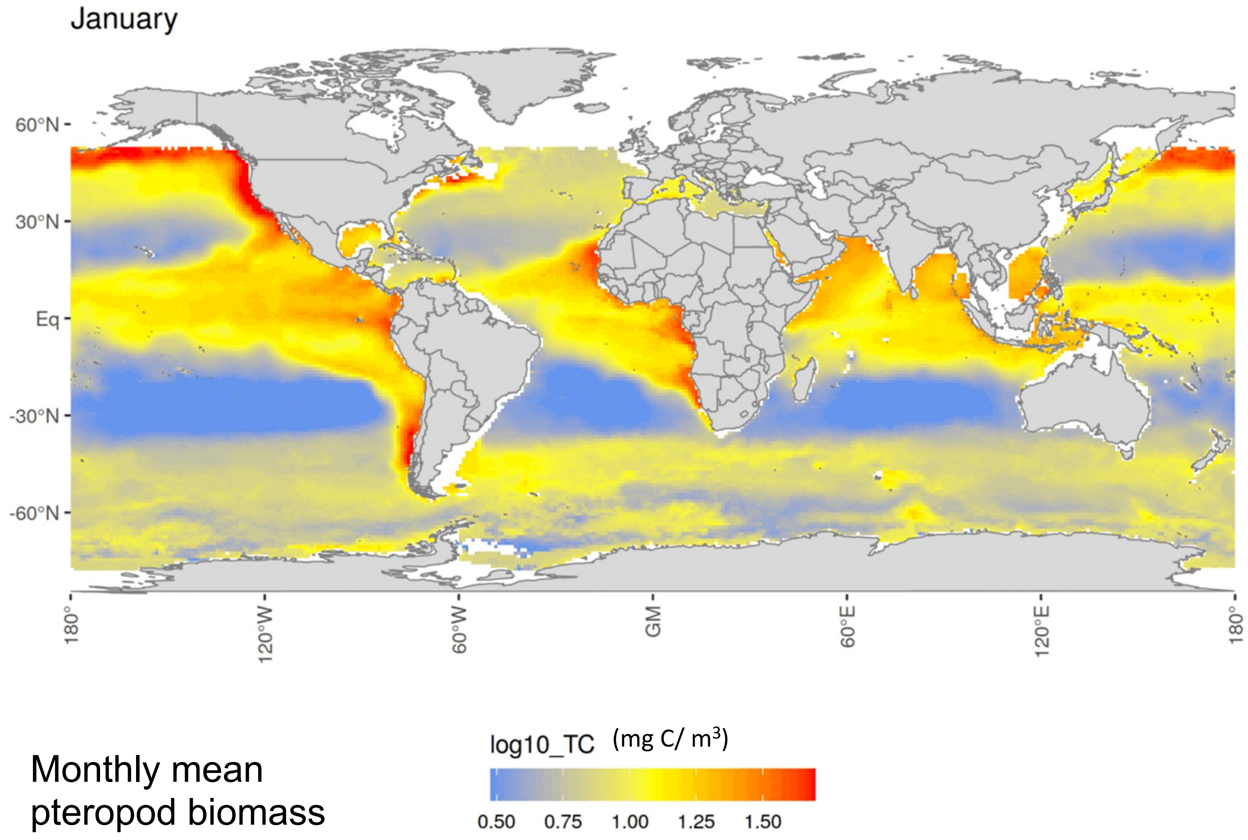
Species Richness, total plankton in 2012

Click on a location on the globe to display corresponding time series. ×

Select variable to be displayed as time series

Biodiversity Index × ▾

Species richness, total plankton from 2012 - 2100



Monthly mean pteropod biomass

Explorable layers of added value for ecosystem monitoring, conservation and policy making

Knecht et al., submitted

eosc | Blue-Cloud2026

blue-cloud.org

@bluecloudeu

blue-cloud org

Funded by
the European Union