| Project Title | **Blue-Cloud 2026: A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters** |
|---|---|
| Project Acronym | Blue-Cloud 2026 |
| Project Number | 101094227 |
| Type of project | RIA – Research and Innovation Action |
| Topics | HORIZON-INFRA-2022-EOSC-01 |
| Starting date of Project | 01 January 2023 |
| Duration of the project | 42 months |
| Website | www.blue-cloud.org |

# D4.1 – VLabs Technical Requirements

| Work Package | WP4 \| Blue-Cloud Virtual Labs for demonstrating cross-domain web-based open science |
|---|---|
| Task | T4.1 \| WP4 coordination |
| Lead author | Patricia Cabrera (VLIZ) |
| Contributors | Alexander Barth (ULiège), Charles Tropin (ULiège), Francesco Palermo (CMCC), Joao Vitorino (IH), Julien Barde (IRD), Bastien Grasset (IRD), Steven Pint (VLIZ) |
| Peer reviewers | Massimiliano Assante (CNR), Dominique Obaton (IFREMER), Sara Pittonet (Trust-IT) |
| Version | V1.0 |
| Due Date | 30/04/2023 |
| Submission Date | 28/04/2023 |

**Dissemination Level**

| | |
|---|---|
| X | PU: Public |
| | CO: Confidential, only for members of the consortium (including the Commission) |
| | EU-RES. Classified Information: RESTREINT UE (Commission Decision 2005/444/EC) |
| | EU-CON. Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC) |
| | EU-SEC. Classified Information: SECRET UE (Commission Decision 2005/444/EC) |

Funded by
the European Union

## Version History

| Revision | Date | Editors | Comments |
|---|---|---|---|
| 0.1 | 5/03/2023 | Patricia Cabrera (VLIZ) | Table of Content |
| 0.2 | 10/03/2023 | Patricia Cabrera (VLIZ) | Introduction, WP4 Summary, Blue-Cloud Virtual Research Environment |
| 0.3 | 1/04/2023 | Alexander Barth (ULiège), Charles Tropin (ULiège), Francesco Palermo (CMCC), Joao Vitorino (IH), Julien Barde (IRD), Bastien Grasset (IRD), Steven Pint (VLIZ) | Blue-Cloud Virtual Labs for demonstrating cross-domain web-based open science |
| 0.4 | 14/04/2023 | Massimiliano Assante (CNR) | First reviewer |
| 0.5 | 20/04/2023 | Dominique Obaton (IFREMER) | Second reviewer |
| 1.0 | 27/04/2023 | Patricia Cabrera (VLIZ) | Final version |

## Glossary of terms

| Item | Description |
|---|---|
| API | Application Programming Interface |
| CMEMS | Copernicus Marine Environment Monitoring Service |
| CNR | Italian National Research Council |
| DD&AS | Data Discovery & Access Service |
| D4Science Infrastructure | Data Infrastructure promoting Open Science (managed by CNR) |
| ECMWF | European Centre for Medium-Range Weather Forecasts |
| EMODnet | European Marine Observation and Data Network |
| EOSC | European Open Science Cloud |
| EOVs | Essential Ocean Variables |
| GRSF | Global Record of Stocks and Fisheries |
| RFMO | Regional Fisheries Management Organisations |
| TBD | To Be Defined |
| VLab | Virtual Laboratory |
| VRE | Virtual Research Environment |

## Keywords

EOSC; Virtual Labs; Big Data; Virtual Research Environment; Data infrastructures

## Disclaimer

The Blue-Cloud 2026 project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101094227. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

## EXECUTIVE SUMMARY

Blue-Cloud 2026 builds upon the pilot Blue-Cloud project which established a pilot cyber platform, providing researchers access to multi-disciplinary datasets from observations, analytical services, and computing facilities essential for blue science. The Blue-Cloud Open Science Platform developed a collaborative environment where different services are available. A **data discovery and access service** was developed as an overarching service to facilitate smart sharing of multi-disciplinary datasets with human and machine users. The Blue-Cloud **Virtual Research Environment (VRE)** orchestrates the computing and analytical services in specific integrated and managed applications exploiting federated Blue-Cloud data resources as well as external data resources**.** The Blue-Cloud innovation potential was explored and unlocked by **five dedicated demonstrators** as Virtual Labs (VLabs) co-designed with top-level marine researchers to demonstrate the power of the Blue-Cloud Open Science platform.

In Blue-Cloud 2026, three of these VLabs are expanded and two new ones are created, with the objective to further evolve the Blue-Cloud Virtual Research Environment (VRE). The VLabs will also bring new data types, such as currents or carbon data, and will also be able to incorporate the Data Discovery and access (DD&AS) Service to their workflows.

This deliverable describes and summarises the technical requirements concerning the input data, workflows and outputs towards the implementation of the five VLabs in the VRE. These technical requirements are assessed by CNR-ISTI, partner responsible for the Blue-Cloud VRE (WP5), to ensure they meet the Platform requirements and further develop the services provided by Blue-Cloud.

This document is structured in 5 sections. Sections 1 and 2 give an introduction to the document and a summary of the WP. Section 3 is dedicated to the Blue-Cloud VRE, on top of which the VLabs will be developed. Section 4 introduces the five VLabs and their objectives along with the workflows, data sources and technical requirements. Section 5 summarises the main findings and next steps.

## Table of Contents

# 1. Introduction

This deliverable gathers the technical requirements from each of the five Virtual Laboratories (VLabs) to be implemented in the Blue-Cloud Virtual Research Environment (VRE). These requirements will be used by WP5 'Blue-Cloud VRE platform evolution and integration with EOSC resources and services', that is responsible for the creation of the VLabs and it might involve expanding and/or upgrading current functionalities. First a summary of the activities of WP4 is presented, followed by a short description about the services provided by the Blue-Cloud VRE and then, a description of each of the VLab along with their technical requirements.

During the implementation phase of the VLabs, close collaboration with WP5 is expected, with the objective that VLabs adapt their technologies and tools to the Blue-Cloud VRE. Almost all the steps of the workflows are expected to be developed/integrated within the Blue-Cloud VRE, except if a very large computation capacity is needed, while some pre-processing steps can also be performed externally (High Performance Computing, HPC). Similarly, VLabs that use public datasets available from the Blue Data Infrastructures (BDIs) should make use of the Blue-Cloud Data Discovery and Access Service (DD&AS) (WP2), especially for very large datasets that cannot be uploaded locally to the VRE.

This deliverable will also serve as input for a series of workshops delivered by WP5 to introduce and explain the VRE functionalities to VLab developers and for the next deliverable *D4.2 "VLabs implementation guidelines"* due to month six of the project.

# 2. WP4 summary

WP4 *"Blue-Cloud Virtual Labs for demonstrating cross-domain web-based open science*" designs, manages and runs the creation of five VLabs in the Blue-Cloud Virtual VRE.

- VLab #1 Coastal oceans observations along Europe (led by IH)
- VLab #2 Coastal currents from observations (led by ULiège)
- VLab #3 Carbon-Plankton dynamics (led by VLIZ)
- VLab #4 Marine Environmental Indicators (led by CMCC)
- VLab #5 Global Fisheries Atlas (led by IRD)

The VLabs cover a set of marine multidisciplinary data from the BDIs and other marine Research Infrastructures (RIs), like JERICO. These data may be from very different types and origins and are usually not gathered together (coastal and open sea data, plankton, hydrodynamics and fisheries data). The VLabs innovative services, in the form of data products and/or analytical tools, demonstrate the

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

added-value of web-based open science as promoted by European Ocean Science Cloud (EOSC). VLabs 1 and 2 are newly developed in Blue-Cloud 2026, while VLabs 3 to 5 were first developed in the previous Blue-Cloud project and will be expanded during this new phase.

The development of the VLabs includes several tasks and phases, from data collection and processing to the development of analytical services, and to the publication of the results in the Blue-Cloud Data and Service Catalogue and EOSC Catalogues. This work is divided in four phases:

- **Phase 1 [M1-M3]: Description of the workflows. I**dentification of datasets, tools, and technical requirements to implement/upgrade the VLabs in the VRE.
- **Phase 2 [M4-M24]: Implementation.** Beta versions of VLabs developed & implemented in VRE.
- **Phase 3 [M25-M36]: Testing & fine tuning.** Beta version fully matured and integrated in the VRE, making use of VRE services (DD&AS, and other services developed under WP2-3-5).
- **Phase 4 [M37-M42]:** Dissemination and user support: In collaboration with WP6-7, VLabs are presented at different events and exploited by users (e.g. hackathon). Feedback is used for service updates.

# 3. Blue-Cloud Virtual Research Environment

The Blue-Cloud VRE is based upon the existing **D4Science** (Assante et al. 2019a) **e-infrastructure** as developed and managed by CNR-ISTI. , where the Blue-Cloud VLabs are hosted. It exploits cloud-based hardware resources (hardware layer) organised as a dynamic pool of virtual machines, supporting computation and storage. The Blue-Cloud VRE serves a federation of computing platforms and analytical services, offering a (1) **Collaborative Framework** to share a workspace with other researchers and interact with the community, (2) the **Analytics Computing Framework**, that allows the execution of analytics tasks in a collaborative environment, and the (3) **Blue-Cloud VRE Data adn ServiceCatalogue,** where the generated products are published and pushed to the EOSC (Assante et al. 2020).

## 3.1 Blue-Cloud Collaborative Framework

The Collaborative Framework is realised through a combination of software components (services and libraries) powered by the gCube System (Assante et al. 2019b). Three main subsystems characterise the Collaborative Framework:

- **Shared workspace System** provides a remote (Cloud) folder-based file system, supporting sharing of folders and different item types (ranging from binary files to information objects representing, for instance, tabular data, workflows, distribution maps and statistical algorithms).
- **User Management System** provides functionality to manage personal profiles and users in the VLabs. Users are assigned a specific role and this defines their privileges.
- **Social Networking System** comprises services conceptually close to the common ones promoted by social networks – e.g. posting news, commenting on posted news, likes, private messages and notifications.

## 3.2 Blue-Cloud Analytics Computing Framework

The Analytics Computing Framework includes a set of services for performing data processing and mining. Algorithms are executed in parallel, using the same e-infrastructure nodes as working nodes. Services performing Data Analytics operations are deployed according to a distributed architecture, in order to balance the load of those procedures requiring local resources. The core services are the following:

- **JupyterHub** enables users to develop and execute *Jupyter notebooks*. This environment is integrated with the rest of VLab facilities, e.g. it is possible to use files from the workspace, to store new files within the workspace and provides access to the Workspace enabling sharing of resources much more easily. It is preconfigured with libraries and packages to ease the execution of common Data Analytics tasks.
- **RStudio** allows users to perform online statistical analyses with the R software for statistical computing, and it is possible to use files from the workspace and to store new files within the workspace. The user is responsible to implement proper R scripts as described in the RStudio documentation. The VRE platform makes accessible the RStudio Application by ensuring its operation and orchestration in a cluster. The cluster is composed of multiple hosts, each of which is assigned in exclusive mode to a user for an entire online session. At the end of the session, all the content stored in that host may be removed. All the data, scripts, and other resources that the user needs to persist have to be stored into the Workspace that is accessible through the RStudio Application.
- The **RShiny Application** is delivered by the user to the Blue-Cloud VRE, as a complete Shiny Application on the Docker Registry. The user is responsible to build the standard Docker image as described in the ShinyProxy documentation by writing the Docker file, building the Docker image and providing the configuration for ShinyProxy. D4Science.org makes accessible the Shiny Application on the infrastructure by ensuring its operation and orchestration in a cluster. The

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

cluster is composed of multiple Docker hosts which run in parallel mode and act as managers and workers.

- **Analytics Engine**: permits the execution of an array of analytics *Methods* by transparently relying on distributed computing infrastructure. Executions can run either on shared multi-core virtual machines or on dedicated virtual machines. New software can be integrated by using the dedicated Software Importer (SAI). *Methods* integration Tutorial: https://data.d4science.net/Kuh.

## 3.3 Blue-Cloud Data and Service Catalogue

The Blue-Cloud Data and Service Catalogue enables users to publish, search and browse data, products, and resources of interest from the Blue-Cloud community. It features datasets and products resulting from the Blue-Cloud VLabs and the methods used to generate them. Every item is accompanied by rich metadata to enhance FAIRness:

- title and creator(s);
- accessibility properties and intellectual properties, e.g. licences;
- technical properties, e.g. size and format;
- legal and ethical attributes, e.g. whether containing personal data;
- Input datasets and methods used.

In particular, the Catalogue is exploited to publish the Blue-Cloud Services and exchange them with the EOSC Marketplace Services Catalogue. This allows EOSC users to discover and access the VLabs and the derived data products. Additionally, the Catalogue ensures direct publication on items into the Blue-Clod Zenodo community: https://zenodo.org/communities/bluecloud/.

# 4. Blue-Cloud Virtual Labs for demonstrating cross-domain web-based open science

## 4.1. VLab #1 Coastal oceans observations along Europe

### 4.1.1. Scientific background and VLab objective

Many types of ocean data are available from a variety of sources, but they are not all coordinated or interoperable. Integration of diverse datasets needs to be addressed and demonstrated to facilitate the creation of an effective knowledge base for application and policy decisions.

This VLab implements an environment that is specifically designed to explore the added value of exploitation and integration of observations collected in European coastal ocean areas. It brings together observations collected by partners of the Joint European Research Infrastructure for Coastal Observatories (JERICO-RI) with other available data and complementary information and implements anlaytical tools as well as interactive state-of-the-art visualisations and advanced processing and post-processing facilities to get unprecedented insights on key processes affecting European coastal ocean environments.

Three Thematic Services (TS) will be implemented in this VLab. **TS#1** will address **"transboundary processes and connectivity along the European margins",** focusing on processes such as biological connectivity, contaminants spread, and along margin impacts of river outflows. **TS#2 "Extreme Events"** will focus on coastal impacts of major storms, and **TS#3 "Ocean glider"** will show the added value of repeated glider sections. Implementation will benefit from JERICO-Coastal Ocean Resource Environment (JERICO-CORE) developments, where its functionalities (API, web services, client libraries, tools, ...) will be expanded.

### 4.1.2. Workflow

- **Data identification and collection**: Different data sources (Figure 1a) are used as inputs for the derivation of exploitation/integration products. The majority of these data sources are public and already available in standard formats. VLab#1 will implement mechanisms to allow users to input their own data and to use it in the derivations of the added-value exploitation/integration products. Each TS is based on a specific time period. TS#1 will focus on the period from January to May 2020, during which a particularly comprehensive set of observations is available to explore transboundary processes and connectivity. TS#2 will focus on two periods, the first one

from 28 February to 10 March 2018, during which strom Emma affected the European coasts and the second one from 10 to 20 October 2018, corresponding to the final stages of evolution of hurricane Leslie to the impact with the Portuguese coast.

- **Data pre-processing**: If required, data sets are processed to standard formats or to meet specific requirements of the TS they feed. In the thematic services TS#1 and TS#2, this stage includes, for example, the preparation of data sets eventually provided by the users, which need to be processed to meet the standard formats used for the subsequent analysis. For the specific case of TS#1, this also includes the low-pass filtering procedure required to proceed, in the analysis phases, with the analysis of the subinertial variability observed in coastal ocean areas. In TS#3 this stage will process glider raw data to OG1.0 format datasets.
- **Data exploitation and integration**: Different tools are implemented to develop different levels of exploitations of the individual datasets and of integrated analysis. The thematic services TS#1 and TS#2 will implement tools for data exploitation (including basic statistics, modal structure of variability, water mass analysis and transport derivation), tools for data integration (including consistency analysis, cross-correlation analysis, bottom impacts mapping and coastal ocean response analysis). TS#3 will implement specific tools for derivation of advanced products from glider data (such as water mass analysis).
- **Results outputs and visualisations**: Results from the exploitations and integration tools are provided using different formats, including NetCDF and ASCII files with different resolutions in x,y,z and time. Advanced visualisations are implemented and used to improve the integration analysis and to allow (for some of the products) user interaction.

*Table 1a – VLab 1 Data sources*

| Dataset/Variable | TS# | Data Infrastructure | Data Access | Link to dataset |
|---|---|---|---|---|
| **IN SITU OBSERVATIONS** | | | | |
| Surface Currents measured by coastal HF radars *(JERICO partners)* | TS#1<br>TS#2 | CMEMS *(INSITU_GLO_PHY_UV_DISCRETE_MY_013_044)*<br>or<br>EMODnet Physics | FTP | https://doi.org/10.17882/86236 |
| Current profiles measured by offshore buoys *(JERICO partners)* | TS#1 | Instituto Hidrografico Puertos del Estado | HTTP | TBD |

| Dataset/Variable | TS# | Data Infrastructure | Data Access | Link to dataset |
|---|---|---|---|---|
| Sea Level Height measured by coastal tide gauge stations *(JERICO partners)* | TS#1 TS#2 | CMEMS (*INSITU_GLO_PHY_SSH_DISCRETE_MY_013_053)* or EMODnet Physics | FTP | https://doi.org/10.17882/93670 |
| Near Surface Temperature measured by offshore buoys *(JERICO partners)* | TS#1 TS#2 | Instituto Hidrografico Puertos del Estado | HTTP | TBD |
| Water column T,S measured by underwater gliders *(JERICO partners)* | TS#1 | PLOCAN | HTTP | http://data.plocan.eu/thredds/catalog/glider/catalog.html |
| SOCIB Glider - Canales Endurance Line. | TS#3 | SOCIB Data Repository | REST API | https://doi.org/10.25704/JD07-SV9 |
| Water Column Temperature measured by multiparametric buoys *(JERICO-RI partners)* | TS#1 TS#2 | Instituto Hidrografico Puertos del Estado | HTTP | TBD |
| Wave parameters measured by offshore buoys *(JERICO-RI partners)* | TS#2 | CMEMS (*INSITU_GLO_WAV_DISCRETE_MY_013_045)* EMODnet Physics | FTP | https://doi.org/10.17882/70345 |
| River outflow from River Gauge Stations | TS#1 | EMODnet Physics | TBD | TBD |
| Seabed Substrate Coastal Type | TS#1 TS#2 | EMODnet Geology | TBD | TBD |
| Coastal Migration | TS#2 | EMODnet Geology | TBD | TBD |

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

| Dataset/Variable | TS# | Data Infrastructure | Data Access | Link to dataset |
|---|---|---|---|---|
| Monthly Vessel Density Maps (Tankers) | TS#1 | EMODnet Human Activities | TBD | TBD |
| **REMOTE SENSING OBSERVATIONS** | | | | |
| Sea Surface Temperature measured by satellite | TS#1<br>TS#2 | CMEMS<br>*(SST_ATL_SST_L4_NRT_OBSERVATIONS_010_025)* | HTTP | https://doi.org/10.48670/moi-00152 |
| Sea Surface Height measured by satellite | TS#1 | CMEMS<br>*(SEALEVEL_EUR_PHY_L4_NRT_OBSERVATIONS_008_060)* | HTTP | https://doi.org/10.48670/moi-00142 |
| Surface Geostrophic Current derived from satellite | TS#1 | CMEMS<br>*(SEALEVEL_EUR_PHY_L4_NRT_OBSERVATIONS_008_060)* | HTTP | https://doi.org/10.48670/moi-00142 |
| Chlorophyll-a concentration | TS#1 | CMEMS<br>*(OCEANCOLOUR_ATL_BGC_L4_NRT_009_116 ; OCEANCOLOUR_ATL_BGC_L4_MY_009_118)* | HTTP | https://doi.org/10.48670/moi-00288<br><br>https://doi.org/10.48670/moi-00289 |
| **MODEL DATA** | | | | |
| 3D analysis of T,S, Current, sea surface height from NEMO model | TS#1<br>TS#2 | CMEMS<br>*(IBI_ANALYSISFORECAST_PHY_005_001)* | HTTP | https://doi.org/10.48670/moi-00027 |
| 3D analysis of wind, air temperature, humidity | TS#1<br>TS#2 | C3S<br>*(ERA5 HOURLY DATA ON SINGLE LEVELS)* | HTTP | https://doi.org/10.24381/cds.adbb2d47 |
| Habitat Suitability Maps for macro algae species | TS#1 | EMODnet Biology | DD&AS | TBD |
| **BATHYMETRY** | | | | |
| DTM for domain selected | TS#1<br>TS#2<br>TS#3 | EMODnet Bathymetry | DD&AS | TBD |

Funded by the European Union

Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

*Table 1b – VLab 1 Technical requirements*

| Computing Capacity | Storage | VRE service used & Programming language | Output Data Format, Visualisation tools |
|---|---|---|---|
| TBD | Order 1TB | RStudio, JupyterHub, Analytics Engine, methods importer, containerised applications (Docker); Python, Fortran, Java, Julia, R | NetCDF, ASCII - different resolutions x,y,z(time), binary images (e.g. png, jpeg)<br>RShiny, Viewers based on web application (to plot TS#3 advanced data products), Specific tools to be implemented in Python or RShiny for advanced visualisation of data products from TS#1 and TS#2 |

## 4.2. VLab #2 Coastal currents from observations

### 4.2.1. Scientific background and VLab objective

The VLab aims to provide a new service to generate integrated ocean surface current maps from High Frequency (HF) radar, drifter data and geostrophic currents from altimetry data using the DIVAnd (data interpolating variational analysis in n dimensions) method. The merging and analysis of these datasets will be performed using various constraints, in particular the presence of the coastline, constraints on horizontal divergence as well as a momentum balance (between acceleration, Coriolis force and surface pressure gradient) as published in Barth et al. (2021) which is an outcome of the SeaDataCloud and EMODnet Physics projects. It is expected that different data sources will have a different accuracy which will be taken into account. The VLab will provide an easy way to use the analysis techniques without installation and with a set of data sources already preconfigured.

The main output of this VLab is a service in the form of easily customizable Jupyter notebooks that allow users to generate surface currents maps for a user-chosen coastal region (when data is available and in particular the availability of HF radar data which extents depending on the configuration about 50 km - 200 km offshore). The user would also be able to make Lagrangian simulations based on these currents maps to visualise the movements of artificial drifters released at a user-chosen location (assuming suitable data coverage).

In addition, visualisation of the surface currents will be realised using the JavaScript library leaflet-velocity, giving to a broad range of interested users an intuitive understanding of the ocean circulation.

As an application of the surface currents, the derived product will be used to initialise the oil spill forecasting model (MEDSLIK-II) described in Liubartseva et al. (2016) to simulate the dispersion of an oil spill accident. MEDSLIK-II is currently developed, among others, by CMCC and University of Bologna, and with an embracing operational chain it is already able to run multi-model simulations and predictions of the transport and weathering of oil spills. It supports several forecasting systems with different resolutions, with the ability to reach up to 50 m in coastal areas.

### 4.2.2. Workflow

The workflow (Figure 1) will be composed of the following steps:

- **Data identification and collection:** the data preprocessing should be part of the reproducible workflow. Therefore we will write scripts to download the data from Table 2a favouring where possible stable URLs and APIs. The provenance of the input data will be clearly documented.

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

- **Data preprocessing:** merging and harmonising the format of the input data files and saving the consolidated data collection as a NetCDF file.
- **Data analysis:** the surface current data (analysis dataset) will be analysed by DIVAnd using different dynamical constraints.
- **Modelling:** Drive the MEDSLIK-II model using the analysed currents.
- **Data publication:** Publish the result from the example domain in the Blue-Cloud Data and Service Catalogue.

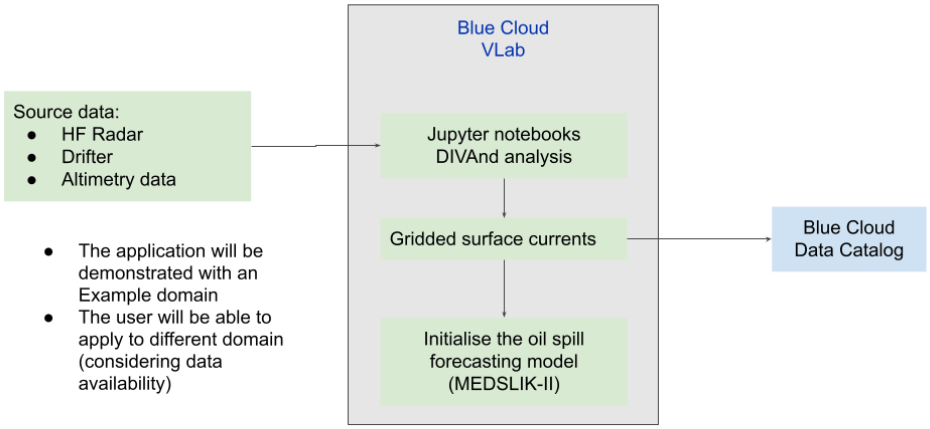


Figure 1: *VLab 2 Workflow diagram showing data sets and processes*

*Table 2a – VLab 2 Data sources*

| Dataset/Variable | Data Infrastructure | Data Access | Link to dataset |
|---|---|---|---|
| Copernicus Marine In Situ - Global Ocean-Delayed Mode in situ Observations of surface (drifters, HFR) and sub-surface (vessel-mounted ADCPs) water velocity | CMEMS *INSITU_GLO_PHY_UV_DISCRETE_MY_013_044* | FTP | https://doi.org/10.17882/86236 |
| European Seas Along Track L 3 Sea Surface Heights Reprocessed 1993 Ongoing | CMEMS SEALEVEL_EUR_PHY_L3_MY_008_061 | FTP | https://doi.org/10.48670/moi-00139 |

| Dataset/Variable | Data Infrastructure | Data Access | Link to dataset |
|---|---|---|---|
| Bathymetry | EMODnet Bathymetry | HTTP | https://emodnet.ec.europa.eu/en/bathymetry#bathymetry-products |
| Global bathymetry | GEBCO | HTTP | https://www.gebco.net |
| GSHHG coastline | NOAA | HTTP | https://www.ngdc.noaa.gov/mgg/shorelines/data/gshhg/latest/ |
| Coastline | Open Street Map | HTTP | https://osmdata.openstreetmap.de/data/coastlines.html |
| Wind speed at 10m height | ECMWF | FTP | |

*Table 2b – VLab 2 Technical requirements*

| Computing Capacity & storage | Programming language | VRE service used | Output Data Format and resolution |
|---|---|---|---|
| 10 GB | Julia, Python, Fortran and nco tools for Medslik | JupyterHub, DataMiner for Medslik | NetCDF (lon/lat/time), the resolution will be a user-chosen parameter, but typically 4 km spatial resolution at hourly time resolution (for an area when HF radar data is present) |

## 4.3. VLab #3 Carbon-Plankton dynamics

### 4.3.1 Scientific background and VLab objective

Marine phytoplankton is at the base of the marine food web and regulates functions in coastal ecosystems. Changes observed in the marine plankton community are expected to have a knock-on effect throughout the food web and the biogeochemical dynamics of marine ecosystems (and to an extent physical). Therefore, understanding how primary production changes through time and space is of key importance to better quantify the effects of human activities and their impact on the ocean.

The objective of this VLab is to offer a workflow to analyse which factors drive the phytoplankton dynamics and how these factors change in space and time using the Nutrient-Phytoplankton-Zooplankton-Detritus (NPZD) model (Soetaert and Herman, 2009). The workflow is focused on a marine system and follows a similar methodology as described in Everaert et al. (2015). The model will be converted to run in carbon units, allowing one to integrate carbon information (carbonate system concentrations and fluxes from ICOS Belgian and Italian stations BE-SOOP-Simon Stevin & BE-FOS-Thornton Buoy, IT-FOS-PALOMA, …) to infer carbon sequestration as part of the detritus state variable. Additionally the model will be applied with data from the Northern Adriatic Sea to improve the reproducibility of the scripts.

### 4.3.2 Workflow

- **Data identification and collection:** Several data sources (table 3a) are used as input data to feed the model. All data is publicly available, either via de Blue-Cloud DD&AS, or via APIs available from the different sources.
- **Data preprocessing**: Integration of data sources: Carbonate system concentrations and fluxes combined with plankton and environmental variables. Processing of these data according to the NPZD model. When data is not available at the time resolution needed generalised additive models will be used.
- **Model calibration:** Conversion of the model to run in Carbon units, calibration and validation against field observations, i.e. calculating the Root Mean Square Error.
- **Model simulation:** Phytoplankton biomass dynamics will be simulated and the relative contributions of the environmental parameters estimated.
- **Quantification of Carbon sequestration:** Combining NPZD model with a Carbon Nitrogen Regulated ecosystem model.
- **Output visualisation:** Creating graphs with the results.
- **Model application to Northern Adriatic Sea:** Application and calibration of the model with input data from the Northern Adriatic Sea.

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

*Table 3a – VLab 3 Data sources (Belgian part of the North Sea in black and Northern Adriatic Sea in blue)*

| Dataset/Variable | Data Infrastructure | Data Access | Link to dataset |
|---|---|---|---|
| Phytoplankton abundances (Chla) | LifeWatch<br>EMODnet Chemistry | API<br>HTTP | https://rshiny.lifewatch.be/station-data/<br>https://cdi-chemistry.seadatanet.org/search |
| Zooplankton abundances | LifeWatch<br>EurOBIS | API<br>HTTP | https://rshiny.lifewatch.be/zooscan-data/<br>https://www.eurobis.org//imis?module=institute&insid=4325 |
| Carbon | ICOS- Carbon Portal, SOCAT<br>ICOS- Carbon Portal, SOCAT | HTTP | https://data.icos-cp.eu/portal/<br>https://www.socat.info/index.php/data-access/ |
| Carbon fluxes | OCADS<br>OCADS | HTTP | https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/SPCO2_1982_present_ETH_SOM_FFN.html |
| Sea-surface Temperature (SST) | Meetnet Vlaamse Banken (SST can also be a product from CMEMS or EMODNet or ICOS CP (i.e. ICOS BE Stations) | API<br>HTPP | https://rshiny.lifewatch.be/mvb-data/ |
| Nutrient concentrations | LifeWatch, ICOS<br>EMODnet Chemistry | API<br>HTTP | https://rshiny.lifewatch.be/station-data/<br>https://cdi-chemistry.seadatanet.org/search |

*Table 3b – VLab 3 Technical requirements*

| Computing Capacity & storage | Programming language | VRE service used | Output Data Format, Visualisation tools |
|---|---|---|---|
| 10-20 fast CPUs & up to 100 GB | R | RStudio/R Markdown | CSV files, PNG graphs in R, RShiny |

## 4.4. VLab #4 Marine Environmental Indicators

### 4.4.1 Scientific background and VLab objective

The Marine Environmental Indicators (MEI) VLab allows users to monitor and assess the environmental status of marine areas and support the decision-making process for ocean management. Multiple data sources are exploited in a unique data analysis service, which will allow the online computation of indicators through *Jupyter Notebooks* or a customised web application that exploits the *VRE Analytics Engine* services for computing the indicator's outputs. The produced indicator outputs can contribute to initiatives such as the Copernicus Marine Ocean State Report or Med-CORDEX.

Functionalities developed during the pilot phase of Blue-Cloud will be improved, including new algorithms for producing new environmental indicators and new data sources (physics, biogeochemistry, biology, chemical data). The visualisation functionalities of data and the overall user experience will also be improved. During the first two years, new algorithms will be developed, to then be implemented as Jupyter notebooks and development of the user interactive interface for the data analytics and display. During the last two years, the focus will be to improve the interoperability with other services, dissemination and exploitation of the service (webinars, hackathon).

### 4.4.2 Workflow

The general workflow is going to be organised in the following steps. It can change in some parts by staring at each algorithm's needs. This workflow can be followed by *Jupyter Notebooks* development but it is mainly specific for methods implemented on the *VRE Analytics Engine* .

- **Data identification**: Selection of input data with different temporal resolutions and origin (e.g. model data, in-situ data from BDIs or other resources) to feed the indicators algorithms.
- **Definition of parameters and data preprocessing**: potential integration of data sources, selection of sub-region domain, specification of the temporal range and levels of depth in case of 3D dataset using the parameters provided by the users (or services). Setting of algorithm's specific parameters.
- **Model simulation**: submission of the algorithm with the parameters defined and the input data selected in the previous steps.
- **Output visualisation**: the outputs will be available in the form of time series or maps in NetCDF or other formats (e.g. images).
- **Output publication**: the outputs produced by the algorithm can be published in a catalogue service (Blue-Cloud Data and Service Catalogue) to be afterwards processed and reused by other services.

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

*Table 4a – VLab 4 Data sources*

| Dataset/Variable | Data Infrastructure | Data Access | Link to dataset |
|---|---|---|---|
| Global Ocean CORA In-situ Observations *INSITU_GLO_PHY_TS_DISCRETE_MY_013_001 and INSITU_GLO_PHY_TS_OA_MY_013_052* | CMEMS-MDS Wekeo | HTTP/API/ FTP | https://www.seanoe.org/data/00351/46219/ https://data.marine.copernicus.eu/product/INSITU_GLO_PHY_TS_DISCRETE_MY_013_001/ https://data.marine.copernicus.eu/product/INSITU_GLO_PHY_TS_OA_MY_013_052 |
| Mediterranean Sea - Temperature and salinity Historical Data | SeaDataNet | HTTP | https://sextant.ifremer.fr/record/2a2aa0c5-4054-4a62-a18b-3835b304fe64/ |
| Various | WOD | HTTP | https://www.ncei.noaa.gov/access/world-ocean-database-select/dbsearch.html |
| CS3 ERA5 / 10m neutral wind speed | Wekeo | HTTP/API | https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview |
| Mediterranean Sea Physics Reanalysis *MEDSEA_MULTIYEAR_PHY_006_004* | CMEMS-MDS Wekeo | HTTP/API | https://data.marine.copernicus.eu/product/MEDSEA_MULTIYEAR_PHY_006_004/description?view=-&product_id=-&option=- |
| Global Ocean Physics Reanalysis *GLOBAL_MULTIYEAR_PHY_001_030* | CMEMS-MDS Wekeo | HTTP/API | https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_001_030/description |
| Chemistry | EMODnet | TBD | TBD |

*Table 4b – VLab 4 Technical requirements*

| Computing Capacity | Storage | VRE service used & Programming language | Output Data Format, Visualisation tools |
|---|---|---|---|
| 32 CPUs, 32GB RAM | 100GB | Jupyter Notebooks, Data Miner, Storage Hub, Blue-Cloud Data and Service Catalogue, Docker Python, Julia, Diva | Output data format: NetCDF, PNG Visualisation services: WMS service, OpenDap service |

## 4.5 VLab #5 Global Fisheries Atlas

### 4.5.1 Scientific background and VLab objective

The scientific background of this VLab revolves around the study and management of tuna and other fish populations in the world's oceans. Tuna are important commercially and ecologically, but they are also highly migratory and can cross international boundaries, making their management a complex and global issue. Fisheries are an important source of food and livelihoods for many people, but overfishing and unsustainable fishing practices have led to declines in many fish populations around the world. The Fisheries Atlas is an online platform that provides access to various data sets related to global fisheries, including catch data, fishing effort, and stock assessments. The GRSF (Global Record of Stocks and Fisheries) catalogue is a database of information on the status of global fish stocks, including tuna.

The objective of this VLab is to provide a hands-on learning experience to explore the use of the Global Tuna Atlas and GRSF catalogue for studying and managing fish populations. The VLab will allow users to navigate and analyse the data provided by these tools, including creating maps and visualisations of tuna distribution and migration patterns. Through this VLab, users will also gain an understanding of the challenges and complexities of managing highly migratory and economically important species like tuna. Overall, this VLab will provide a comprehensive and interactive learning experience that will enhance global understanding of tuna biology and management.

The automated task sequence that processes the data through a specific path from initiation to completion for Global Tuna Atlas and Global Record of Stocks and Fisheries are detailed in the sections below.

### 4.5.2 Global Tuna Atlas Workflow

The following workflow is dependent on the data provided, thus it could change during the time of the project as work is done in collaboration with the providers to update input data in the Global Tuna Atlas Workflow.

- **Data collection:** The first step in creating the Global Tuna Atlas is to collect data from tuna regional fisheries management organisations (tRFMOs), and scientific research institutions. This data includes information on tuna catch and effort, as well as other factors such as conversion factors from catches in number to catches in tons.
- **Data standardisation:** Collected data is standardised to ensure that it is consistent and can be compared across different sources. This involves reconciling differences in data formats, units of measurement, mapping of codelist, and other factors that can affect the accuracy of the data.

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

- **Data processing:** Once standardised, data can be treated following different choices to reduce inconsistency of data or to remove duplicates. Some choices are made on the species kept in the dataset to ensure that adequate data is retained to preserve important information, while discarding inconsistent and not desired data.
- **Mapping and visualisation:** The results of the data analysis are then visualised using maps and other graphical representations as well as shiny apps and explanation papers. This allows users of the Global Tuna Atlas VLab to explore the data and gain insights into the state of the global tuna fishery.
- **Records publication:** During this step the contents of the GRSF knowledge base are published in the Blue-Cloud Data and Service Catalogue offered by the VRE platform. At this stage, a Digital Object Identifier (DOI) is assigned to the outputs hosted on Zenodo, facilitating their identification and referencing. Additionally, both the data and metadata are made available in a standardised format to ensure accessibility and interoperability with other systems.

### 4.5.3 Global Record of Stocks and Fisheries Workflow

- **Data harvesting:** The first step of the workflow is to fetch the data from the original data sources. Currently GRSF collects data from four different data sources; FIRMS[1], RAM[2], FishSource[3], and FAO SDG 14.4.1 Questionnaire[4]. This process does not affect the data from the remote database sources. The data are collected using various methods and processes. The harvested data do not have a common structure or format. FIRMS data are in XML format, RAM data are provided as tabular data (i.e. in MS Excel), FishSource data are in JSON format, and FAO SDG 14.4.1 data in RDF format.
- **Data transformation**: After fetching the data it is important to transform them so that they have a similar structure and semantics. At this stage, data is transformed from XML, JSON and MS Access to RDF format. Specifically, data are transformed into instances of the MarineTLO[5] ontology with respect to the identified GRSF requirements. Information harvested from the database sources will be mapped to the agreed GRSF standards, when not already compliant.
- **Data normalisation:** The normalisation step applies a set of normalisation filters to the transformed data so that they are compliant with the GRSF standards. These filters may alter or add information to assist the instance matching functionalities of the merging process. Examples of normalisation filters include the addition of the corresponding FAO 3Alpha code that is

[1] FAO FIRMS (https://firms.fao.org/firms/en)
[2] RAM Legacy Stock Assessment Database (https://www.ramlegacy.org/)
[3] Sustainable Fisheries Partnership FishSource (https://www.fishsource.org/)
[4] FAO Sustainable Development Goals Indicator 14.4.1 (https://www.fao.org/sustainable-development-goals/indicators/14.4.1/en)
[5] https://projects.ics.forth.gr/isl/MarineTLO/

associated with the scientific name of a species, the addition of ISO3 codes to a country or the specification of a water area standard regarding a specific code.

- **Data cleaning:** The data cleaning step includes all the necessary modifications of the source data in order to correct observed errors. The errors are being corrected either by the application of automatic filters (such as the genus capitalization for the scientific names of species) or by the notification of the sources to refine their data and re - harvest the altered content. More examples of data cleaning include the spelling correction in the scientific names, gear codes or water area codes.
- **Records merging:** This step ensures that the contents that have been added in the GRSF knowledge base are properly connected based on a set of criteria. This is achieved by linking records that have the same values on particular fields (specifically time-independent values) for producing a new single GRSF record. For example, if there are stock records having the same species and water area, we can merge them into a single stock. During this process, we also use external knowledge to detect similarities among different names and terminologies used in the database sources (i.e. species names).
- **Records dissection:** A process applied to aggregated source fishery records so that they will construct concrete GRSF fishery records compliant with the GRSF standards. The process is applied on particular fields of the aggregated record (i.e. species, fishing gears, and flag states) so that the constructed GRSF record is uniquely described and suitable for traceability purposes. Considering that the source fishery record example contains two different species, the dissection process produces two distinct GRSF fishery records.
- **Records publication**: During this step the contents of the GRSF knowledge base are published in the Blue-Cloud Data and Service Catalogue offered by the VRE platform. This way GRSF contents are exposed from the corresponding facilities of the dedicated VLab, allowing the experts to inspect the contents of the GRSF and curate them appropriately. During this step, Universally Unique Identifiers (UUID) and human readable semantic identifiers are generated and associated with each GRSF record. The former are generated based on RFC 4122 and are used to uniquely identify records. The latter are generated using various GRSF fields and populated with standard codes and allow the identification and interpretation of records by humans.

### 4.5.4 Merging of the workflows

In the Blue-Cloud 2026 project, the objective is the merging of the two above workflows into one. The first steps of each workflow have separated inputs and outputs. The merging will occur in the Geonetwork catalogue which will be supplied by both workflows. Directly from the database outputs for the Global Tuna Atlas workflow and via CKAN for the Global Record of Stocks and Fisheries Workflow.

Figure 2 summarises the steps explained above:

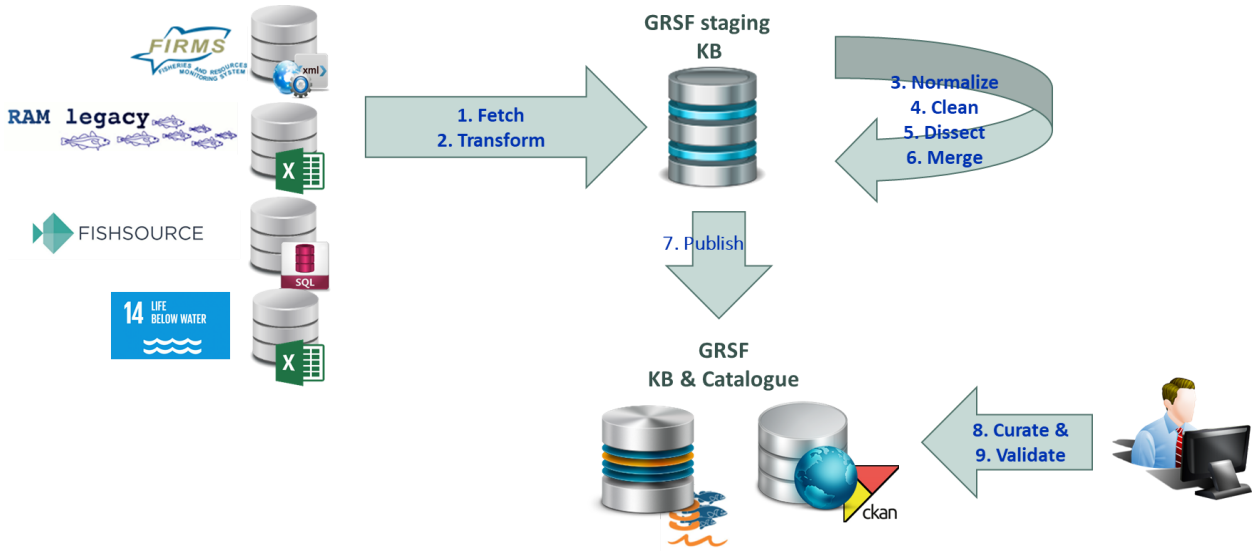


Figure 2: *VLab 5- GRSF construction workflow*

*Table 5a – VLab 5 Data sources*

| Dataset/Variable | Data Infrastructure | Data Access | Link to dataset |
|---|---|---|---|
| Georeferenced catches (1 deg/ 5 deg/All) | Tuna RFMOs | HTTP | https://www.fao.org/fishery/geoserver/tunaatlas/ https://dx.doi.org/10.5281/zenodo.5746041 https://dx.doi.org/10.5281/zenodo.5745986 https://dx.doi.org/10.5281/zenodo.5747174 https://doi.org/10.5281/zenodo.5745958 |
| Georeferenced efforts (1 deg/ 5 deg/ All) | Tuna RFMOs | TBD | TBD |
| Georeferenced catches harmonised units of measures (1 deg/ 5 deg/ All) | Tuna RFMOs | TBD | TBD |

| Dataset/Variable | Data Infrastructure | Data Access | Link to dataset |
|---|---|---|---|
| GRSF | RAM legacy database, FIRMS, FishSource | TBD | TBD |

*Table 5b – VLab 5 Technical requirements*

| Computing Capacity | Storage | VRE service used & Programming language | Output Data Format, Visualisation tools |
|---|---|---|---|
| 32GB RAM | < 100 GB | GeoNetwork<br>GeoServer<br>Postgres / Postgis<br>OpenFairViewer<br>Shinyproxy<br>R (RStudio)<br>SQL<br>Java | SQL<br>OGC XML metadata<br>OGC access protocols : WMS / WFS<br>RDF / SPARQL |

## 5. Conclusions and next steps

This document highlights the technical requirements as requested by the VLab developers. These may change or evolve as the work progresses and will contribute to the development of the Blue-Cloud VRE.

The VLabs focus on specific thematic fields, and as such their requirements are not the same. For example, some VLabs are oriented in bringing different types of data together, providing a workflow to harmonise, combine and make this data available to users. Other VLabs develop their own algorithms and provide indicators to end users. Because of these differences, some VLabs require the use of notebooks or API's, while in other cases workflows are implemented using the analytical services from the VRE, such as the Analytics Engine. Most of the VLabs use data and data products as inputs from several Blue Data Infrastructures, such as CMEMS, several EMODnet lots and SeaDataNet. Some of the methodologies make use of Artificial Intelligence methods such as machine learning and/or deep learning. Almost all VLabs request data visualisation tools to display the outputs (e.g. RShiny, time series, graphs, etc.)

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

The further development of the VLabs will take place in a close cooperation between WP2 "FAIR compliant Discovery and Access services for marine domains & beyond", WP3 *"Developing and testing analytical Blue Cloud WorkBenches for generating highly qualified data collections (Essential Ocean Variables, EOV)"* and WP5 **"***Blue-Cloud VRE platform evolution and integration with EOSC resources and service***s".** WP2 will provide access to the data in multiple Blue Data Infrastructures that are planned to be used in each VLab. WP3 is expecting to use some of the data products derived in WP4, while WP4 also will benefit from the outputs from WP3. This has to be well coordinated since this work will be developed in parallel. Finally, WP5 will further optimise, expand and/or upgrade its functionalities according to the needs requested by WP4.

As a next step, after the VLabs will be created, the deliverable D4.2 *'VLabs Implementation guidelines'* will describe the implementation recommendations of the VLabs in the VRE and how to explore the existing services of the VRE in collaboration and exchanges between WP4 and WP5, which will be supported by a workshop that is planned in the month of June 2023.

Funded by the European Union

**Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227**

## References

M. Assante et al. (2019a) Enacting open science by D4Science. Future Gener. Comput. Syst. 101: 555-563 10.1016/j.future.2019.05.063

M. Assante et al. (2019b) The gCube system: Delivering Virtual Research Environments as-a-Service. Future Gener. Comput. Syst. 95: 445-453 10.1016/j.future.2018.10.035

Assante M., Candela, L., and Pagano P. (2020). D4.2 Blue Cloud VRE Common Facilities - Release 1 (Version 1). Zenodo. https://doi.org/10.5281/zenodo.6335691

Barth, A., Troupin, C., Reyes, E., Alvera-Azcárate, A., Beckers J.-M. and Tintoré J. (2021): Variational interpolation of high-frequency radar surface currents using DIVAnd. Ocean Dynamics, 71, 293–308. https://doi.org/10.1007/s10236-020-01432-x

Everaert, G., De Laender, F., Goethals, P.L.M, Janssen, C.R (2015). Relative contribution of persistent organic pollutants to marine phytoplankton biomass dynamics in the North Sea and the Kattegat. *Chemosphere* 134, 76-83. http://dx.doi.org/10.1016/j.chemosphere.2015.03.084

Liubartseva, S., Coppini, G., Pinardi, N., De Dominicis, M., Lecci, R., Turrisi, G., Cretì, S., Martinelli, S., Agostini, P., Marra, P., and Palermo, F. (2016). Decision support system for emergency management of oil spill accidents in the Mediterranean Sea. *Natural Hazards and Earth System Sciences*, *16*(8), 2009-2020. https://doi.org/10.5194/nhess-16-2009-2016

Soetaert, K., Herman, P.M.J. (2009). A practical guide to ecological modelling. Using R as a Simulation Platform. Springer-Verlag, New York, US, p. 54 - 58.