MISSIONE 4
ISTRUZIONE
RICERCA

Deliverable 8.12 – OPERATIONAL VERSION OF THE CZ VRE SERVICE AND THE ASSOCIATED USER GUIDE

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Deliverable number:	D8.12
Work package:	WP8
Intermediate Objective:	
Deliverable type:	X Document, report
•	☐ Websites, patent filings, videos, etc.
•	• ☐ Other: please specify ....
Dissemination level:	X Public
•	• ☐ Restricted
Estimated delivery (bimester):	B12
Actual delivery date:	October 2024
Author(s) (Partner-OU):	Main authors: Pasquale Bove Cnr-Igg, Francesca Caparrini Cnr-Igg, Simona Gennaro Cnr-Igg, Silvio Marta Cnr-Igg Contributing Authors: Francesca Avogadro Di Valdengo Cnr-Igg, Ilaria Baneschi Cnr-Igg, Alice Baronetti Cnr-Igg, Marco Chimenti Cnr-Igg, Gianpaolo Coro Cnr-Isti, Letizia Costanza Cnr-Igg, Andrea Dini Cnr-Igg, Marco Doveri Cnr-Igg, Mariasilvia Giamberini Cnr-Igg, Silvana Goiran Cnr-Igg, Matia Menichini Cnr-Igg, Jasmine Natalini Cnr-Igg, Pasquale Pagano Cnr-Isti, Angelica Parisi Cnr-Igg, Brunella Raco Cnr-Igg, Gianna Vivaldo Cnr-Igg, Antonello Provenzale Cnr-Igg
Reviewed by:	ITINERIS Executive Board
	•

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System - CUP B53C22002150006 (D.D. n. 130/2022)
Funded by EU - Next Generation EU
Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Summary

1. THE "CZ VRE" – WP 8 TASK 8.1

Most critical issues currently faced by our society involve key environmental challenges. Pollution, land use and climate changes, and their impacts on biodiversity and ecosystem integrity must be addressed urgently, providing knowledge and quantitative data useful to define strategies and policies to mitigate them. For studying the entire Earth System, a multi-disciplinary approach, combining field and lab measurements, data analysis and modelling tools across all the environment domains, is fundamental. To achieve this functional and integrated approach, national nodes from different research infrastructures (RIs) agreed to work for a thematic and comprehensive network in the fields of the geo and environmental sciences.

The main focus of the ITINERIS WP 8 Task 8.1 is the creation of the CZ VRE, i.e. a Virtual Research Environment (VRE) containing the information from the Critical Zone observatories (CZOs) active in Italy and abroad and managed mainly by Italian research teams.

The Critical Zone (CZ) represents the thin layer between the unweathered bedrock and the top of the vegetation canopy where "rock meets life" (Giardino & Houser, 2015). According to this definition, the CZ includes rocks, soil, water, microbiota, vegetation and fauna, the services they provide to humankind, all the processes supporting terrestrial ecosystems and the soil-vegetation-atmosphere interactions (Ashley, 1998).
As such, CZ is a fully trans-disciplinary topic which needs to be collectively addressed by a network of intercommunicating observatories providing shared models.
The CZ VRE aims at integrating data, tools for data visualization and analysis, and modelling services in a web-based (e-Science) platform that allows researchers to access and analyse data from different observation sites (CZOs), and to share data and models with other researchers and stakeholders.

1.1 The Virtual Research Environment and the D4Science e-Infrastructure

A Virtual Research Environment (VRE) is an e-Science environment built to support the principles of Open Science and data FAIRness (data which meet the principles of findability, accessibility, interoperability, and reusability). The goal of a VRE is to enhance collaboration among researchers to find appropriate solutions and address specific scientific and/or management questions (Assante et al., 2019, 2023). Thus, to simplify the concept of VRE, we could say that it is not a "simple" repository. In fact, VREs require not only the collection and harmonization of data, but also the collection and development of analysis tools, modelling solutions and graphical tools. VREs design and creation in general follow some simple steps that can be summarized as follows:

- identification of questions (and/or gaps) and stakeholders potentially interested in the issue underlying the VRE;
- collection and harmonization of validated datasets, modelling solutions, analysis and graphical tools;
- development of new data, knowledge and solutions to address specific scientific and/or management questions;
- definition and collection of the metadata to improve data FAIRness.

The VRE implementation was carried out using the D4science e-Infrastructure (https://www.d4science.org, last access: 25/09/2024). D4science is an organization, hosted by the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" of the National Research Council of Italy (ISTI, CNR), which offers a data infrastructure since the last 10 years. D4Science data infrastructure consists of a network of hardware and software resources (e.g., databases, services, machines) and is managed by a team of Information and Communication Technology (ICT) professionals and researchers having as main objective the infrastructure maintenance, updating, operation and support for users. The infrastructure relies on the *gCube* technology (https://gcube.wiki.gcube-system.org/gcube/About_gCube, last access: 26/09/2024), that is an open-source software toolkit specifically conceived for the construction and development of VREs (Assante et al., 2019). Currently, D4Science hosts 20 different gateways (https://services.d4science.org/thematic-gateways, last access: 26/09/2024) and 176 active VREs with a total of more than 24k users worldwide (Aug. 2024; Figure 1).

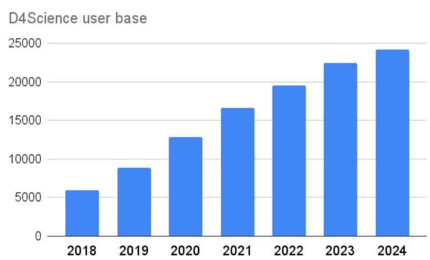

Figure 1. D4Science infrastructure supports collaborative Science and the sharing of expertise with the creation of VREs. The histogram shows the increase of D4Scierce users since 2018 (Aug. 2024).

The use of the D4Science infrastructure thus enhances the effectiveness of the Open Science approach, supporting collaborative science and the sharing of competences with the creation of VREs (Assante et al., 2019).

The CZ VRE development, powered by the D4Science infrastructure is based on the needs of the scientific communities and the specific stakeholders identified by researchers. Based

on the identified needs, the D4Science team supports researchers on the implantation and/or development of new packages, functions and tools.
The following sections will be focussed on some tips and suggestions for the use of the CZ VRE in general and on some specific brief user guides describing specific CZ VRE tools.

1.2 The design of the "CZ VRE"

The design and development of a Virtual Research Environment (VRE) dedicated to the CZ is the fundamental part of the ITINERIS WP8, Task 8.1 project (funded by EU – Next Generation EU PNRR, Mission 4 "Education and Research" – Component 2: "From research to business" – Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures").
LifeWatch ERIC, ICOS ERIC, SIOS and eLTER RI are the reference infrastructures of the Critical Zone Virtual Research Environment (CZ VRE). The main stakeholders of the CZ VRE were identified as conservationists and managers of protected areas, together with the scientific community interested in the analysis of the complex processes involved in the CZ dynamics.
In general, the main sources (in terms of data and metadata) feeding CZ VRE are (i) the IGG-CNR Critical Zone Observatories which provide measurements from automated and portable accumulation chambers and eddy covariance stations, and data from instrumentation deployed by specific projects such as IDROCZ and WinterCZ (cf. Table 1 for a list of currently available datasets); (ii) the Copernicus and MODIS services for metadata and data acquisition procedures.
The CZ VRE contains different types of products, in terms of codes in different programming languages (e.g., R and Python), open-source applications and models (e.g., Maxent, OGGM) and of web application for creating and sharing computational documents (e.g., Jupyter Notebook) (Table 2).

Environment	***Site***	***Instrument***	***Data type***	***N records***
Arctic	**Ny Ålesund**	Portable chamber	CO_2 flux, vegetation class, atmospheric pressure, soil volumetric water content, soil temperature, air temperature, air relative humidity, solar irradiance, green fractional cover	496 (248 paired)
Semi-arid	**Pianosa**	Automated chambers	CO_2 flux, atmospheric pressure, soil volumetric water content, soil temperature, air temperature, air relative humidity, wind direction, wind speed, rain gauge	29,768

Mountain	**Lauson**	Portable chamber	CO_2 flux, atmospheric pressure, soil volumetric water content, soil temperature, air temperature, air relative humidity, solar irradiance	168
Mountain	**Lavassey**	Portable chamber	CO_2 flux, atmospheric pressure, soil volumetric water content, soil temperature, air temperature, air relative humidity, solar irradiance	79
Mountain	**Forni**	Portable chamber	CO_2 flux, atmospheric pressure, soil volumetric water content, soil temperature, air temperature, air relative humidity, solar irradiance	97
Mountain	**Nivolet**	Portable chamber	CO_2 flux, atmospheric pressure, soil volumetric water content, soil temperature, air temperature, air relative humidity, solar irradiance	7,180 (3,590 paired)
Mountain	**Nivolet**	Automated chambers	CO_2 flux, H_2O flux, atmospheric pressure, soil volumetric water content, soil temperature, air temperature	3,550
Mountain	**Nivolet**	Eddy covariance	CO_2 flux, air temperature, air relative humidity, wind direction, wind speed, maximum wind speed, wind direction	2,008

Table 1. Main sources (in terms of data and metadata) for the CZ VRE are the IGG-CNR Critical Zone Observatories (CZOs). In particular, the table contains a synthesis of the variables measured by the IGG-CNR group in different environments and currently available at present in the CZ VRE.

Product	***Programming language***	***Data type (original)***	***Product type***
Software	Python	Glacier outlines	OGGM – Open Global Glacier Model (Maussion et al., 2019)
Code	R	Eddy covariance	Visualization of time series of CO_2 fluxes and environmental variables
Code	R	Glacier outlines	High-res prediction of soil temperature in proglacial areas
Code	R	Portable chambers	Non-linear modelling of CO_2 fluxes
Code	R	Portable chambers	Visualization of spatial trends in CO_2 fluxes and environmental variables

Code	R	Automated chambers	Smoothing and visualization of time series of CO_2 fluxes
Software	Data Miner		Bayesian Methods, Feed Forward Neural Network Trainer and Regressor
Software	Data Miner		Ecosystems: Habitat Representativeness Score
Software	Data Miner		Geo Processing: BIMAC
Software	Data Miner		Machine Learning Methods: Support Vector Machines Modelling
Software	Data Miner		Occurrences Obis Data Retrieval
Notebook	Python		Interactive tools for data visualization and processing

Table 2. Main types of products in terms of codes and software implemented and/or available in the CZ VRE. "Data type (original)" refers to the data the code / software was originally conceived for.

2. USER GUIDE ON THE "CZ VRE"

2.1 Access to the VRE

A specific section of the ITINERIS Project website will be devoted to the ITINERIS Hub (https://itineris.cnr.it/itineris-hub/, last access: 25/09/2024). The ITINERIS Hub will be the unique access point to all the services, facilities, tools and datasets collected and provided by the Italian RIs in the environmental domains (Figure 2). The Hub is conceived as a sort of "unique comprehensive catalogue" where researchers will also publish the VREs metadata and users will be able to access the ITINERIS VREs Gateway.

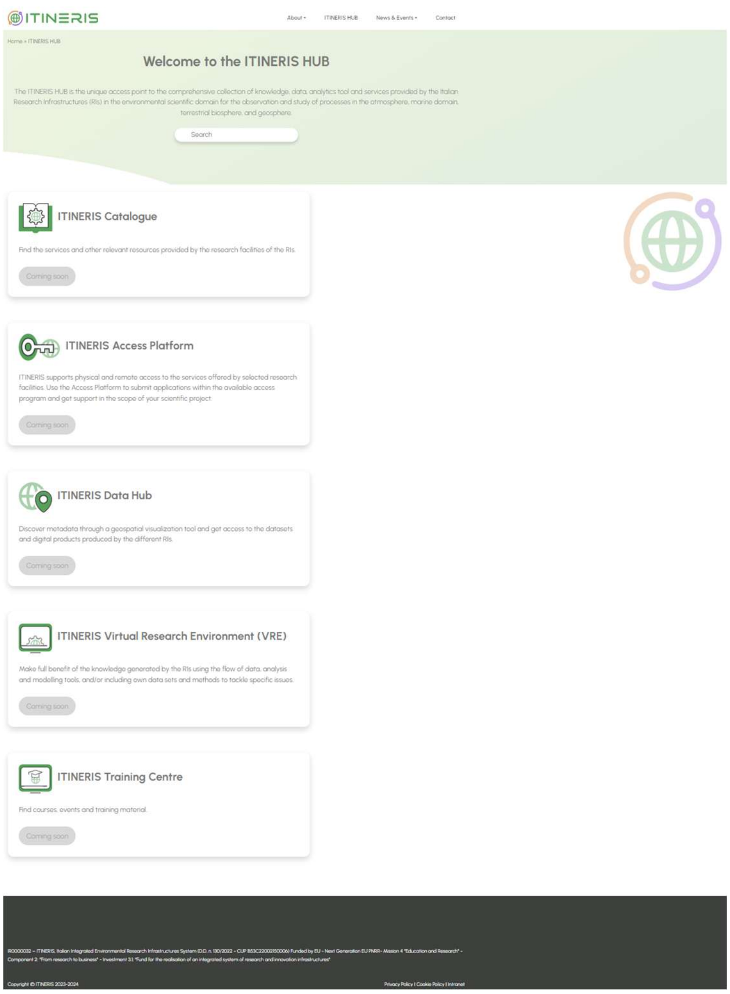

Figure 2. View of the screen of the ITINERIS Hub (https://itineris.cnr.it/itineris-hub/, last access: 25/09/2024).

The D4Science infrastructure allowed to create a unique Gateway to access multiple VREs: the ITINERIS VREs Gateway (https://itineris.d4science.org/, last access: 25/09/2024; Figure3a). D4Science is equipped with an Identity and Access Management (IAM) system.

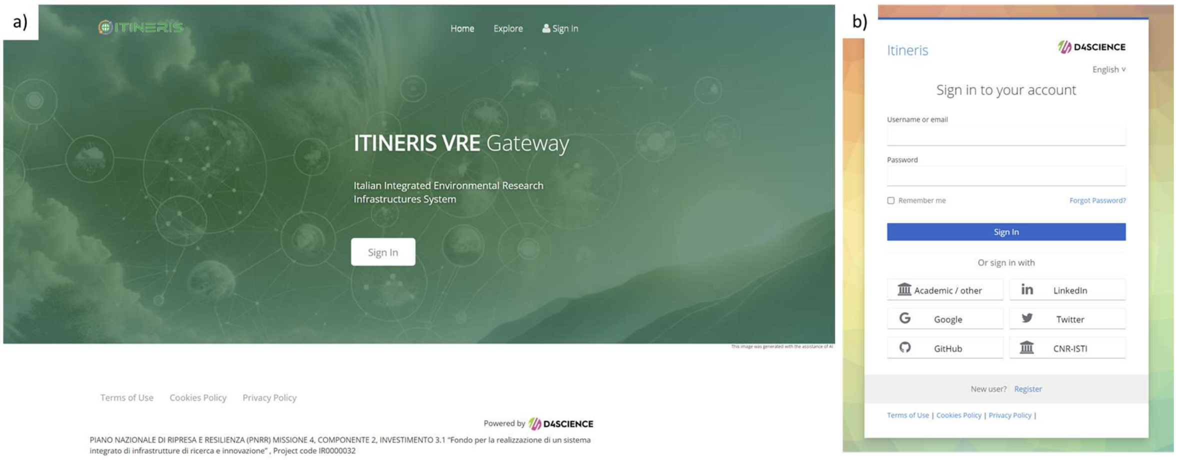

Figure 3. a) View of the screen of the ITINERIS VREs Gateway. b) View of the screen of the sign-in page where the user can insert the username and the password or select one of the different methods available for login. For the creation of a new D4Science account, the users can select the "Register" button and proceed.

After the login with the D4Science's username and password or selecting another method for the login (by using existing accounts, e.g., ORCID, GitHub, Google, LinkedIn; Figure3b), the VREs will be available and accessible. In particular, multiple teams of the ITINERIS WP8, with a high level of trans-disciplinarity, are working on the development of several thematic VREs [e.g., Critical Zone (CZ VRE); Aquatic Biomass services (BIOMASS VRE); Essential Variables (EV VRE); Aerosol-biosphere (AERO VRE); Carbon Cycle services (CARBON VRE); Indicators and Impacts of Climate Change (CLIMA VRE); Downstream Effects of Environmental Change (DOWNSTREAM VRE); Isotope Database (ISOTOPE VRE)].

Once the users have identified the VRE they are interested in, they can click the button "Request Access" (Figure 4) and confirm the request after stating a motivation for accessing to the VRE (this represents a mandatory field).

Each VRE has moderators and managers who can review the access requests. In general, in a brief time (typically within a few hours) the user receives a notification via email about the outcome of the request.

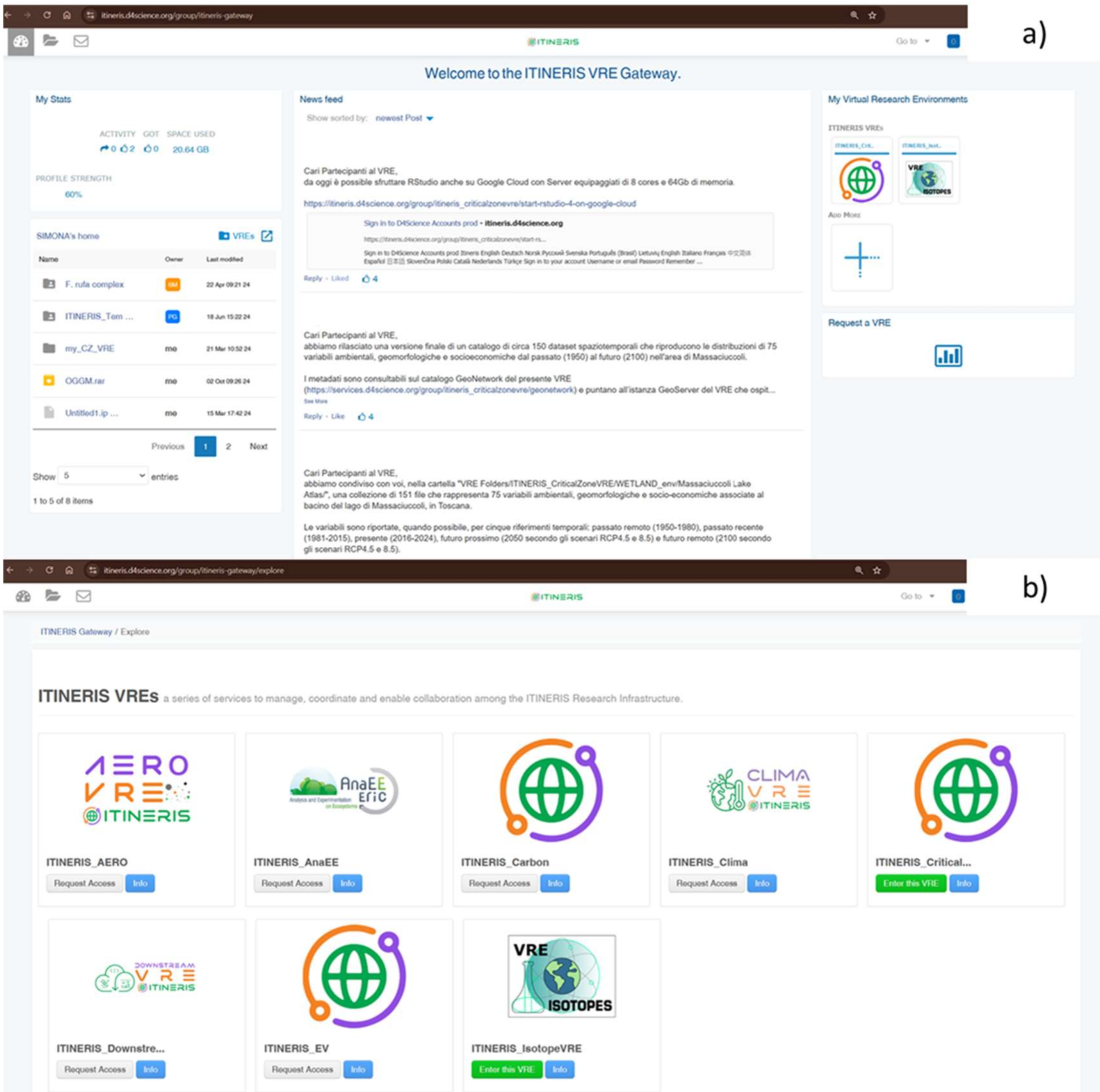

Figure 4. a) Panel on the top shows how the users can see their gateway after the login. On the left the panel "My Virtual Research Environments" allows to add one or more VRE(s) by clicking on the button "+" ("Add More"). b) By clicking on "Add More", the users can start to see the VREs available. Once the users have identified the VRE they are interested in, they can click the grey button "Request Access" and confirm the request after stating a motivation for accessing to the VRE (this represents a mandatory field).

2.2 Primary services of the CZ VRE: the workspace and the social network board

The primary services offered by the CZ VRE are a **workspace** with high computing capabilities, a shareable folder system, and a **social networking board system**. In particular, these basic elements are described below.

The VRE is equipped with an area for data collection with Collaborative Storage Framework and a collaborative workspace that promotes teamwork and offers the opportunity to share digital objects. In particular, the "**workspace**" is an exceptionally reliable, distributed, online file system that allows one to upload files, create private folders, share folders with specific colleagues, publish folders, share data, and directly publish data in a catalogue. The VRE users will find (named with the name of the user: "USER's workspace") a space named "**VRE Folders**" in their workspace, identified by a white star on the icon of the folder. Files and digital objects added or created in the VRE folders are automatically shared with the other VRE's members. Users also can make use of a restricted area; folders and/or files in this area are private, unless the user shares them. To share folders and/or files, users must use the share option (Figure 5) and select the user(s) to share with and the adequate sharing policy (Figure 6). The selectable policies ("Permission", Figure 5) are three: *Write Any* (all users can modify all files), *Read Only* (users other than the owner can only read files), and *Write Own* (all users can create files but only modify files they own). These three different options are also identified by different icons (Figure 6).

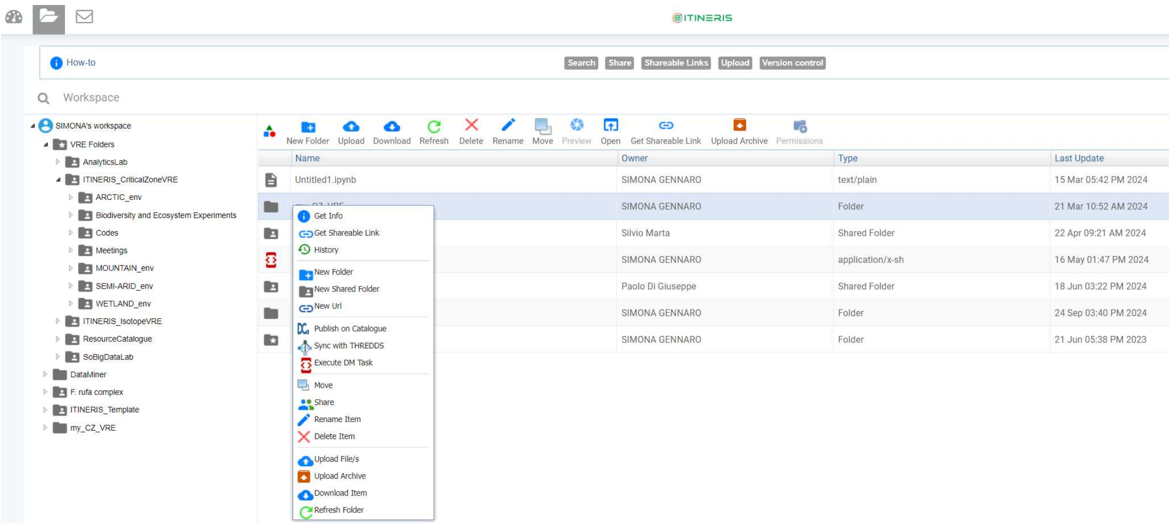

Figure 5. View of the workspace and of the drop-down menu available by right-clicking on a file or a folder in the private space.

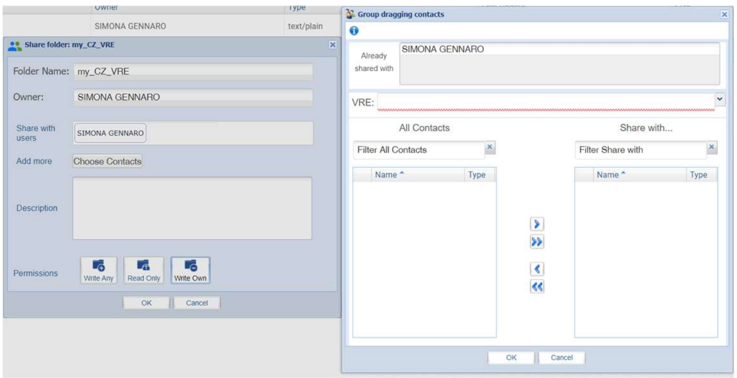

Figure 6. View of the "Share folder" window. At the bottom of this window, the user can select permissions. By clicking on "Choose Contacts" in the "Share folder" window, the user can access the window named "Group dragging contacts", where the users can manage the selection of the contacts.

The **social networking board system** integrated into the VRE allows the exchange of messages between users. The board (also called "feed"; available by clicking on "Communication", see Figure 6) for open discussions allows to share updates, links and files from the VRE folders. It is also possible to use the "@" symbol to directly mention other participants, or a tag ("#") to identify topics and keywords. Members of the VRE can also reply to a specific post or add a "Like" as reaction. VRE's members can set specific options to receive notification (via email and/or on the ITINERIS VRE Gateway) for the social network board, the workspace, as well as for the catalogue.

2.3 The Spatial Data Services and the Metadata Catalogue of the CZ VRE

About data publishing, the Publishing Framework facilitates the dissemination of research outcomes by means of the Metadata Catalogue (provided in the framework of the Spatial Data Services; Figure 7), which helps organising and making research results available to stakeholders and the broader scientific community.

All data, products and services available in the ITINERIS CZ VRE (e.g. data, codes, software and/or other types of digital research outputs) should be FAIR - *Findable*, *Accessible*, *Interoperable* and *Reusable* (Di Muri et al., 2024). The ITINERIS catalogue, currently under development, is designed to promote openness principles and interoperability among the various ITINERIS VREs.

The metadata, in general, will facilitate discovery, access, (re-)use, and preservation of data, codes, software and/or diverse types of digital research outputs. Interoperability in particular is ensured by the use of existing vocabularies (thesauri) to describe data and facilities. In particular, for example, for the meteo-climatic variables, the CF Standard Name Table (https://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html, last access: 26/09/2024) will be used; while the dataset collected by the eddy covariance facilities used in the CZOs will follow the ICOS (Integrated Carbon Observatory System; https://www.icos-cp.eu/) standards and the FLUXNET vocabulary (https://fluxnet.org/data/aboutdata/data-variables/ and https://fluxnet.org/data/fluxnet2015-dataset/fullset-data-product/, last access: 25/09/2024).

VRE members can participate in the VRE and can contribute to creating and editing metadata.

The **Metadata Catalogue** is powered by GeoNetwork and is available at https://itineris.d4science.org/group/itineris_criticalzonevre/geonetwork (last access: 30/09/2024). **GeoNetwork** is an open-source application devoted to managing and searching spatially referenced resources (https://www.geonetwork.it/, last access: 30/09/2024).

All metadata will be compliant with the standards, i.e. the **INSPIRE** (Infrastructure for Spatial Information in Europe) Directive, and the **ISO19139** XML scheme implementation. Moreover, metadata will be compliant with the Italian guidelines as from **RNDT 2.0**

(*Repertorio Nazionale dei Dati Territoriali 2.0*; https://geodati.gov.it/geoportale/manuale-rndt, last access: 30/09/2024). Considering these metadata standards, the RNDT 2.0 is necessarily compliant with INSPIRE, but the opposite may not be true (so it is better to stick to RNDT 2.0 in order to have metadata "automatically" compliant with the INSPIRE Directive). A set of information that should be included in the metadata record to be compliant with the standards EU INSPIRE Directive and ISO19139 is reported at the link https://docs.geonetwork-opensource.org/4.2/annexes/standards/iso19139/ (last access: 30/09/2024). At this link, a description of the fields that should be compiled can be found.
VRE's members can access the GeoNetwork spatial data catalogue by clicking on it in the Spatial Data Services menu (Figure 7a). The metadata catalogue allows to use three specific tools (Figure 7c): "search", "makeYourMap" and "contribute", described in the following.

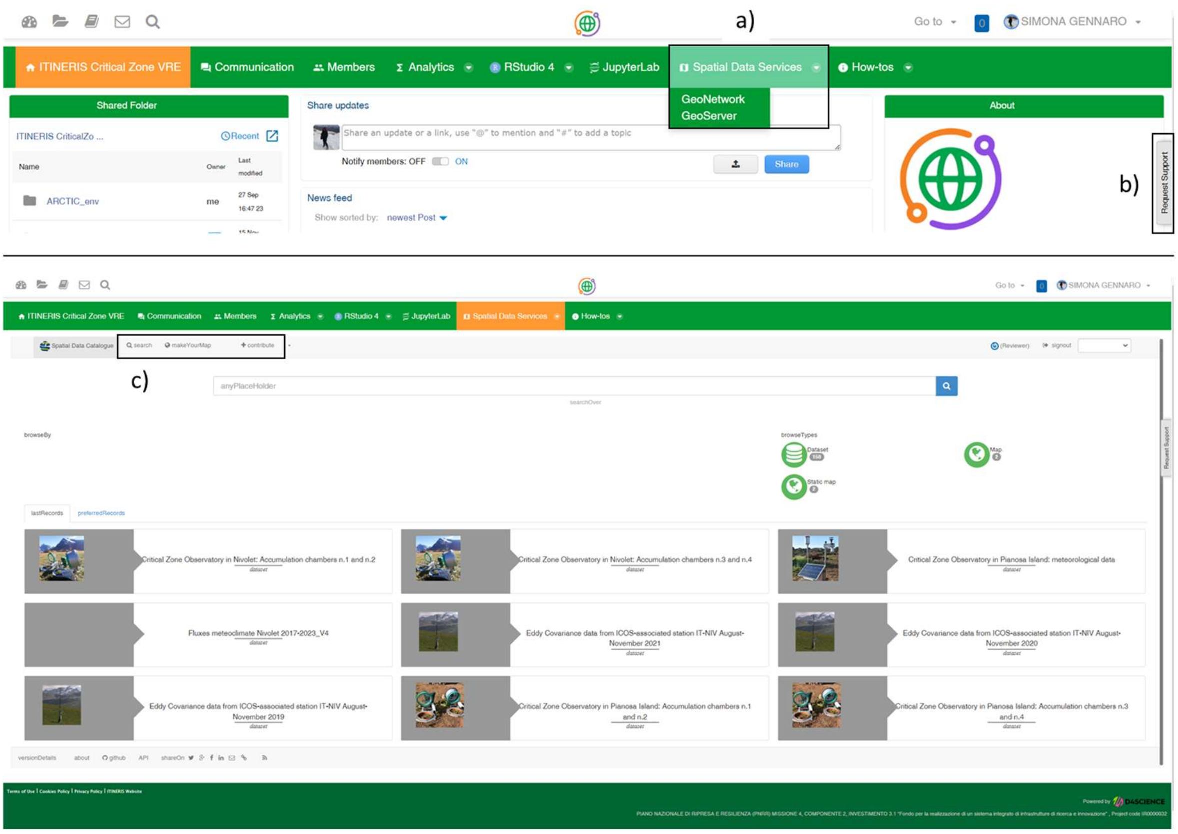

Figure 7. a) The panel on the top shows the tools available in the CZ VRE and, in particular, the Spatial Data Services menu highlighted by a). Letter b) indicates the "Request Support" button. The panel on the bottom shows the general view of the GeoNetwork Spatial Data Catalogue (https://itineris.d4science.org/group/itineris_criticalzonevre/geonetwork, last access: 30/09/2024). Here, letter c) evidence some tools available in the GeoNetwork Spatial Data Catalogue (i.e., "search", "makeYourMap" and "contribute").

The "search" button in GeoNetwork allows to search digital objects in the data catalogue by keywords and/or applying filters for making more specific queries (Figure 8).

The VRE user can also simply visualize the geospatial data of the CZ VRE by using the “makeYourMap” tool (Figure 9) and selecting a specific layer (by using the button “view” – with a “world” icon – visible on the “card” of a specific record, see Figure 8).

Finally, the “contribute” button allows to contribute to the metadata catalogue by adding new metadata records, importing records, and editing or removing existing metadata (Figure 10 and 11). VRE members may also decide to apply some edits as batch editing.

Operationally, the user can click on “contribute” > “add record” to create a new record (Figure 11). From the metadata template list, the users should select a template choosing a group from the dropdown menu available and click “createMetadata” and then they must complete the fields provided by default in the template and can create an image to illustrate the record in the search results of the catalogue.

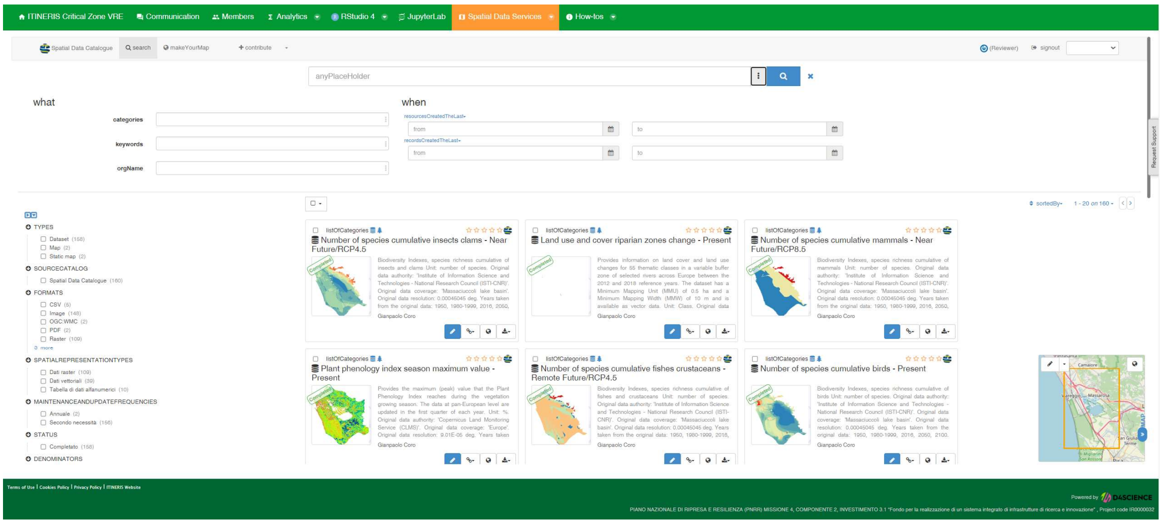

Figure 8. View of the “searching” tool of the GeoNetwork Spatial Data Catalogue.

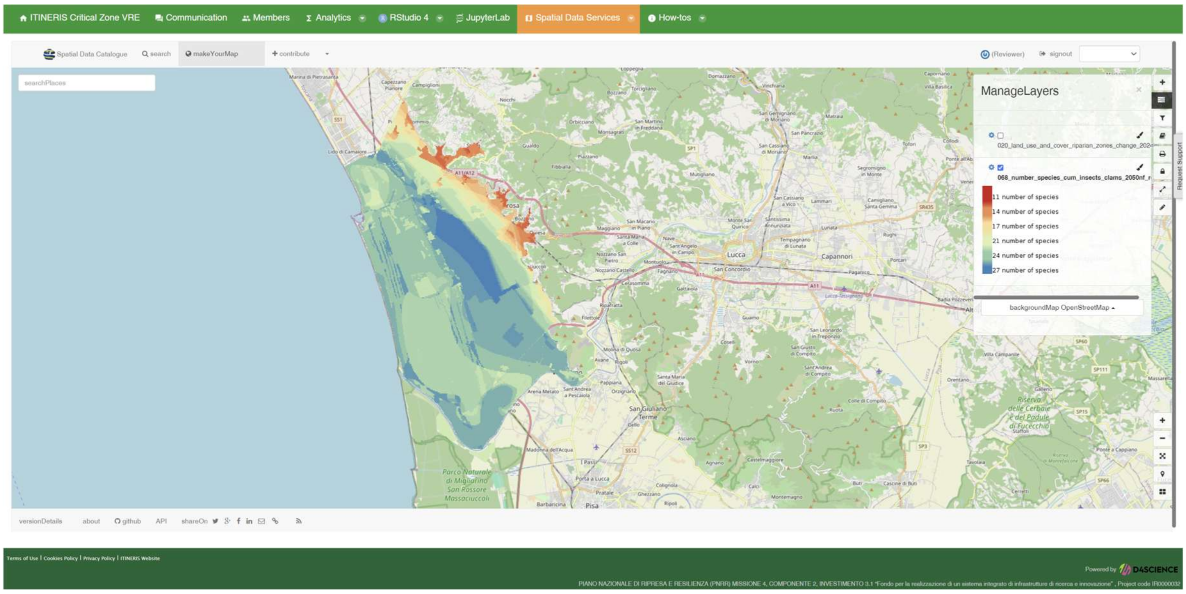

Figure 9. Viewer of the geospatial data with an example of the biodiversity indices produced in the context of the ITINERIS PNRR Italian project - project code No. IR0000032 - ESFRI Environment: Data harmonization process - Gian Luca Vannini - PhD Thesis UniPi - Supervisors Prof. G. Brunori, Dott. G. Coro - Supported by D4Science infrastructure.

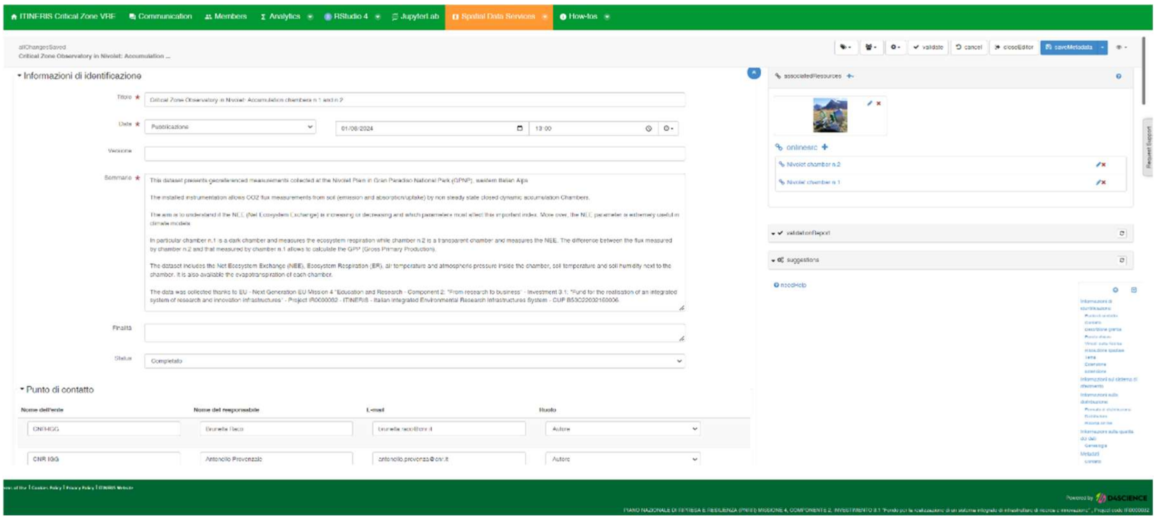

Figure 10. View of the editing metadata record panel.

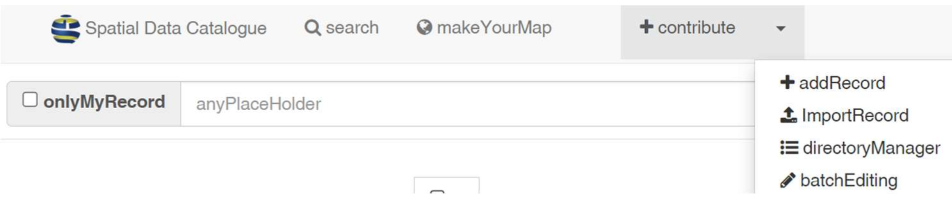

Figure 11. How to create a new metadata record operationally.

Metadata records can also be created and added in the GeoNetwork Spatial Data Catalogue by using the **GeoServer** instance implemented in the CZ VRE

(https://itineris.d4science.org/group/itineris_criticalzonevre/geoserver, last access: 30/09/2024). This is an open-source Java software server which allows users to publish and share geospatial data using open standards, thus allowing the implementation of some of the most used Open Geospatial Consortium (OGC) protocols (e.g., Web Map Service - WMS, Web Feature Service - WFS). The GeoServer instance is not open to all the VRE members, and it is accessible through a specific login. Enabled VRE members can manage and load geospatial data via the GeoServer interface by adding and configuring specific workspaces (only if necessary) to group similar data. Specific documentation on the GeoServer software interface is available at https://docs.geoserver.org/latest/en/user/ (last access: 30/09/2024).
In general, data associated with the metadata collected in the dedicated catalogue available at https://itineris.d4science.org/group/itineris_criticalzonevre/geonetwork, can be collected in this repository based on the GeoServer software and/or in the VRE's workspace.

2.4 The Analytics Engine Framework of the CZ VRE

The **Analytics Engine Framework** of the CZ VRE is equipped with instruments for data analytics. The CZ VRE allows the analysis of datasets, as well as the possibility to create and share codes, models and research products. In particular, the VRE offers the opportunity of using and implementing a system of integrated analytical methods (DataMiner) using integrated web-based interactive development environments (RStudio or JupyterLab) and integrated importer software on a virtual machine connected to the workspace folder system.
The CZ VRE integrates two versions of RStudio, a Standard (4 Cores / 8G RAM) and a Large (8 Cores / 32G RAM) one on a RStudio Server (2023.03.0, Build 386, © 2009-2023 Posit Software, PBC). About the JupyterLab, the CZ VRE integrates Python 3.10.11 and preinstalled modules *pandas*, *gdal*, *scipy*, *netCDF4*, *matplotlib*, *h5netcdf* and other for data driven research workflows (Default Server - 4 Cores / 16G RAM).
The CZ VRE is also equipped with the Software Algorithm Importer (known as SAI or Methods Importer) that is, a web interface tool to import new algorithms and scrips.

2.5 The DataMiner and Operators

The DataMiner allows one to use pre-filled operators over data of the workspace.
DataMiner operators are the implementations of the scripts and algorithms once they are published by the DataMiner and made available. A schematic view of the thematic organization of DataMiner operators is shown in the Figure 12.

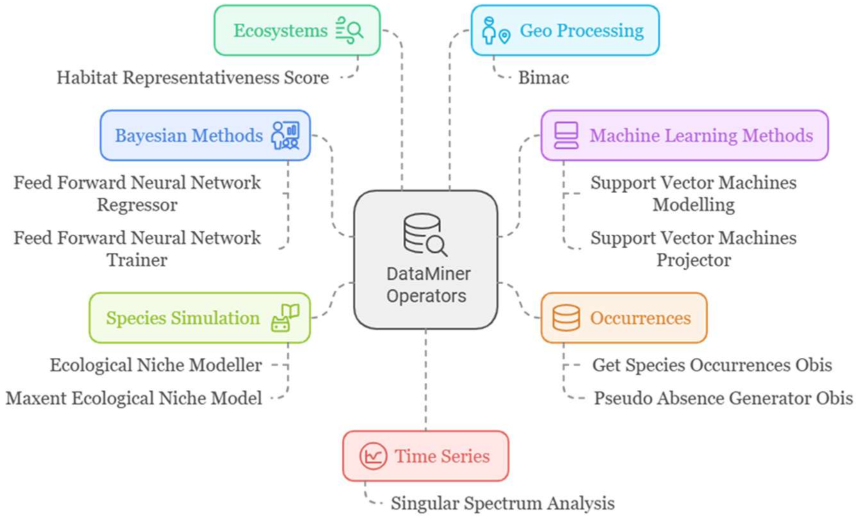

Figure 12. Thematic organization of DataMiner operators.

Users can publish new ones by submitting their scripts via the Software Algorithm Importer. To use DataMiner in a shared and collaborative way we use the idea of “experiments”. Each run of an operator defines an experiment (Figure 13).

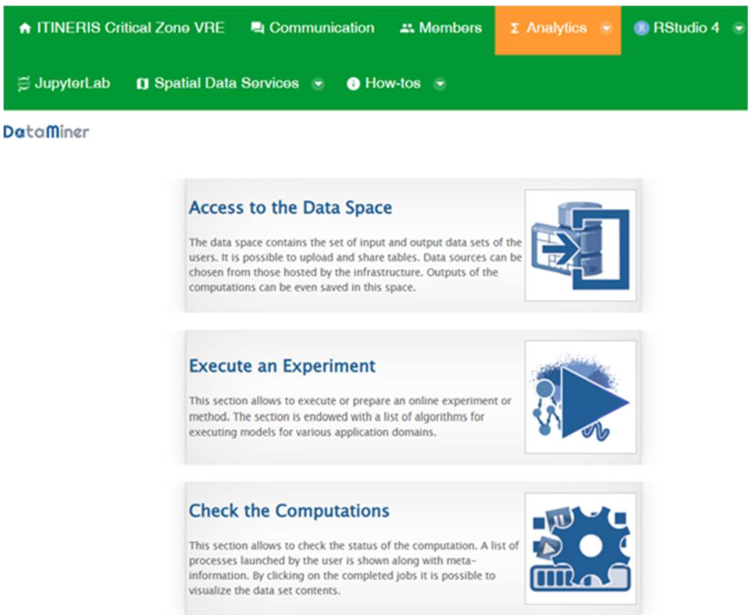

Figure 13. Panel showing the tools available in the DataMiner implemented in the CZ VRE.

An experiment is determined by its inputs, outputs, computation logs and its provenance. Each of these components is shared among the users with whom the experiment is shared.

It is possible to run an experiment and keep track of its input and output data by accessing the Data Space as shown in the Figure 14.

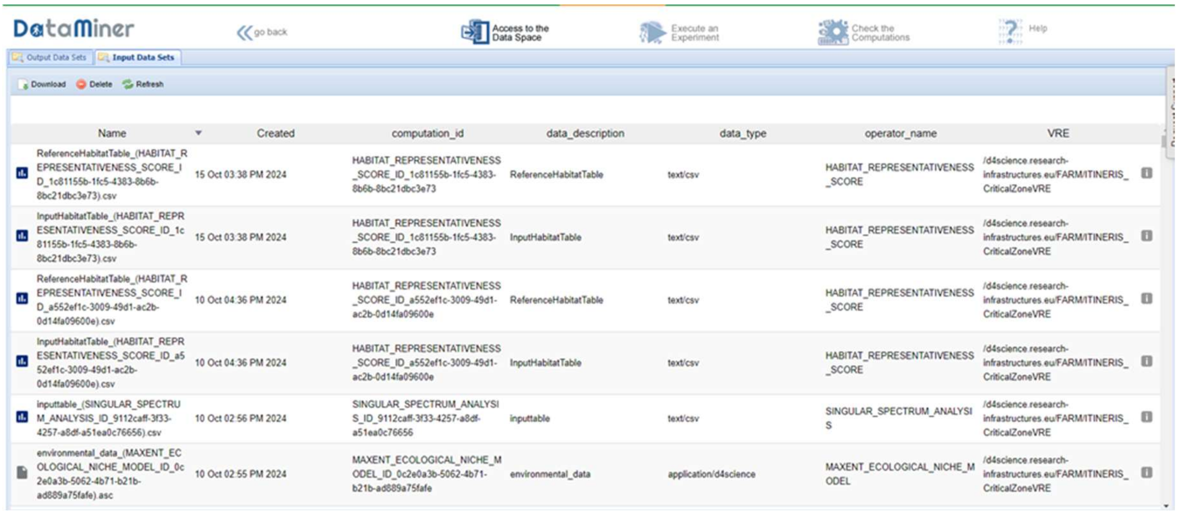

Figure 14. Each experiment produces outputs that can be consulted and shared.

Moreover, as shown in Figure 15, users can view the computation logs and the related metadata (which are not automated added to the metadata catalogue).

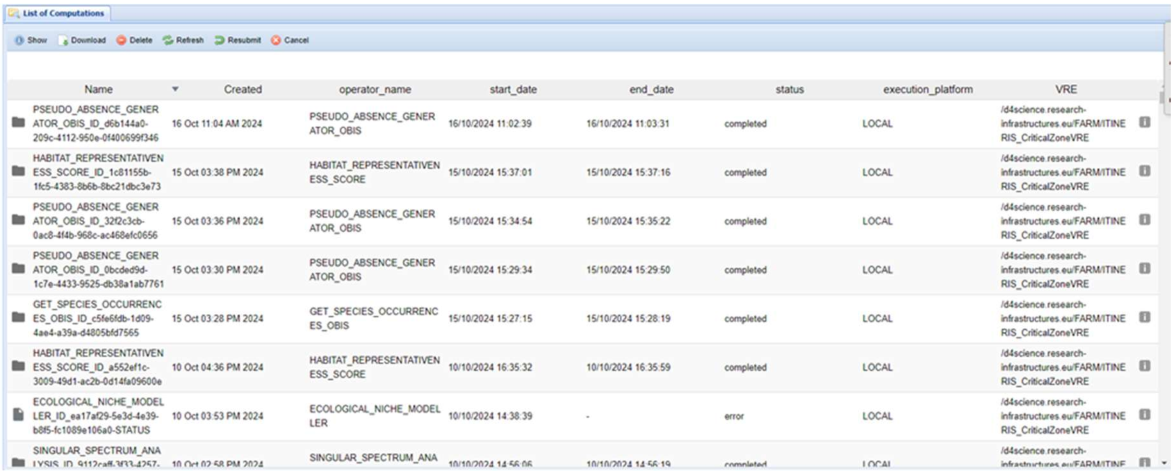

Figure 15. Each experiment produces computational logs that can be viewed and shared.

The “Execute an Experiment” main feature of the DataMiner allows access to the list of operators to run individual experiments, each with its graphical interface. Currently, several operators (Figure 16) are already available on the platform, while others can be added later via the Software Algorithm Importer to meet the research needs of users.

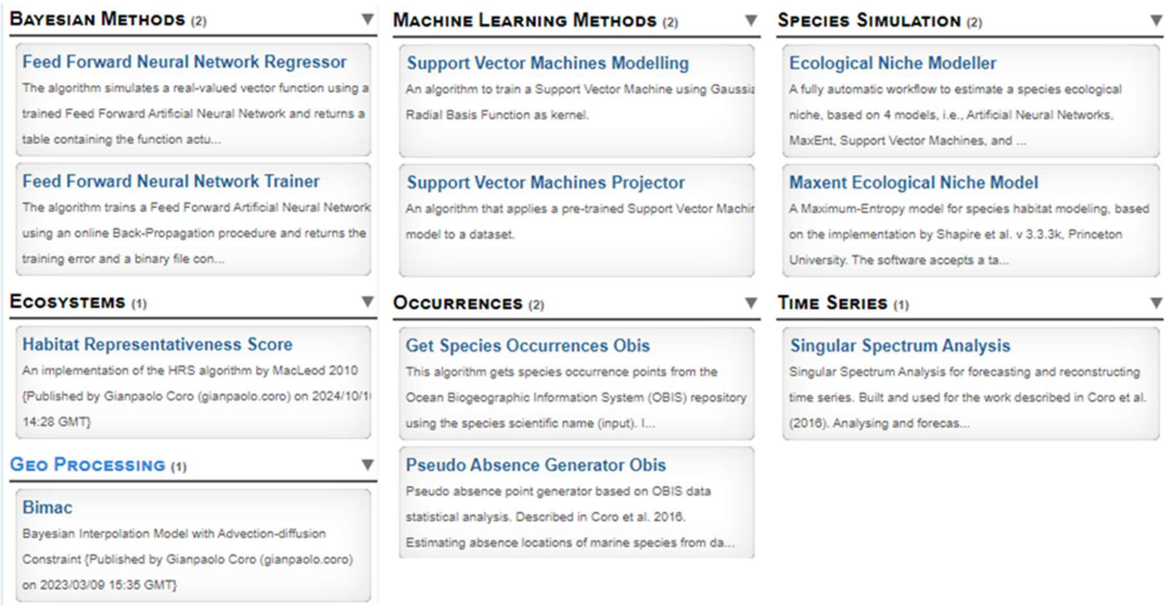

Figure 16. DataMiner operators list.

Operators use the data stored in the workspace by accessing it directly or via Shareable Link provided by the workspace system itself (Figure 17).

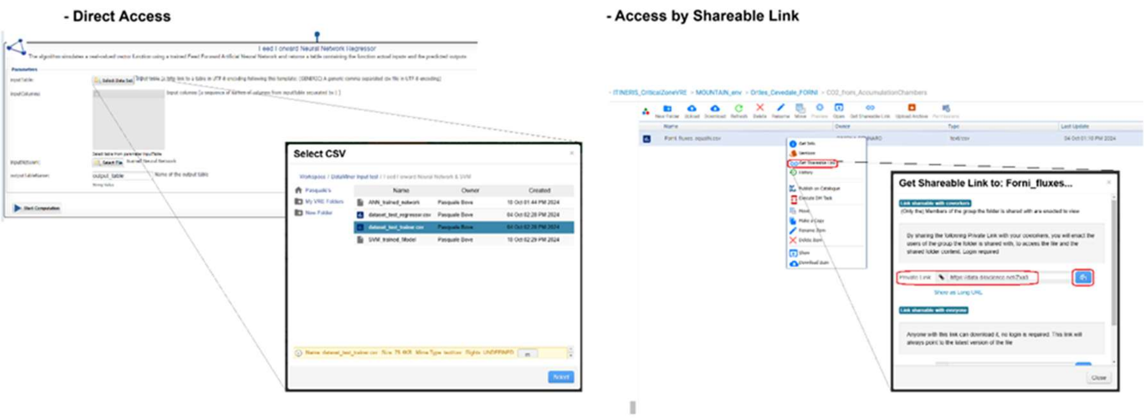

Figure 17. Data access mode in the workspace.

The current operators in the DataMiner cover a variety of tools that we have used so far in our experiments. Mainly these tools have been used for ecosystem-type assessments related to the biodiversity of the critical zone or to reconstruct environmental datasets.

Bayesian Methods: Feed Forward Neural Network Trainer

The algorithm trains a Feed Forward Artificial Neural Network (Figure 18) using an online Back-Propagation procedure and returns the training error and a binary file containing the trained network.

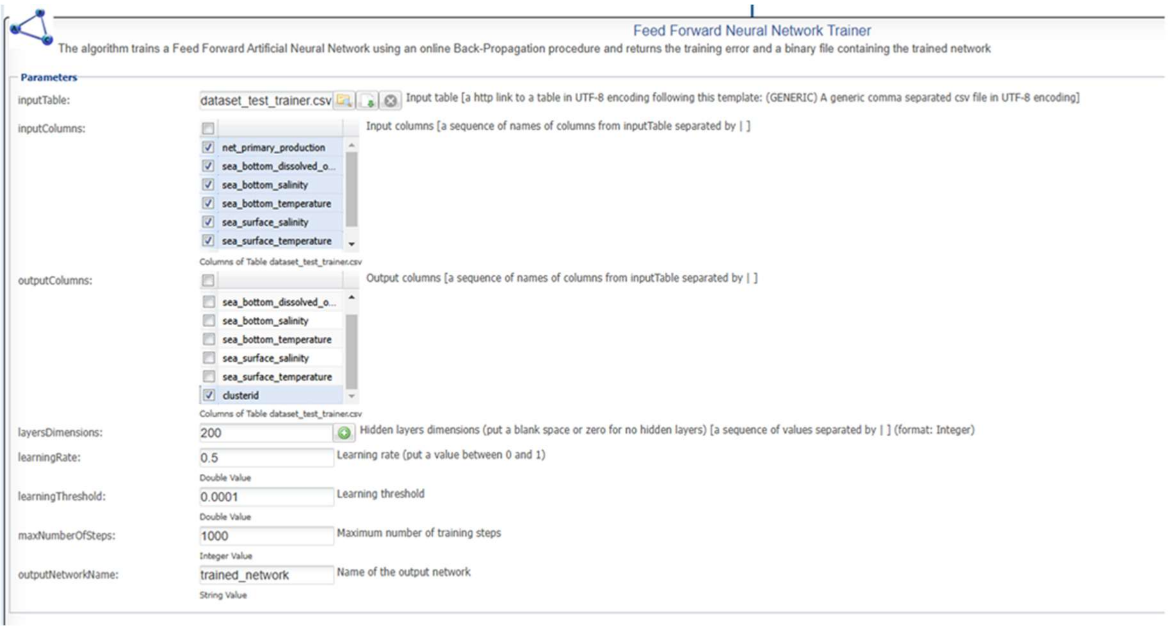

Figure 18. Feed Forward Neural Network Trainer features selection.

The trainer takes as input from the workspace a table in .csv format containing the features on which to train the network plus a column with the target values of the function. It then allows you to select from the table with an interactive menu the features, the column with the target values, and parameters to obtain better accuracy of the model.

The output result provides a trained model that can be used in the regressor operator to obtain forecasts on other data sets homogeneous to the training set.

Bayesian Methods: Feed Forward Neural Network Regressor

The algorithm simulates a real-valued vector function using a trained Feed Forward Artificial Neural Network (Figure 19) and a data set of variables to returns a table containing the predicted outputs.

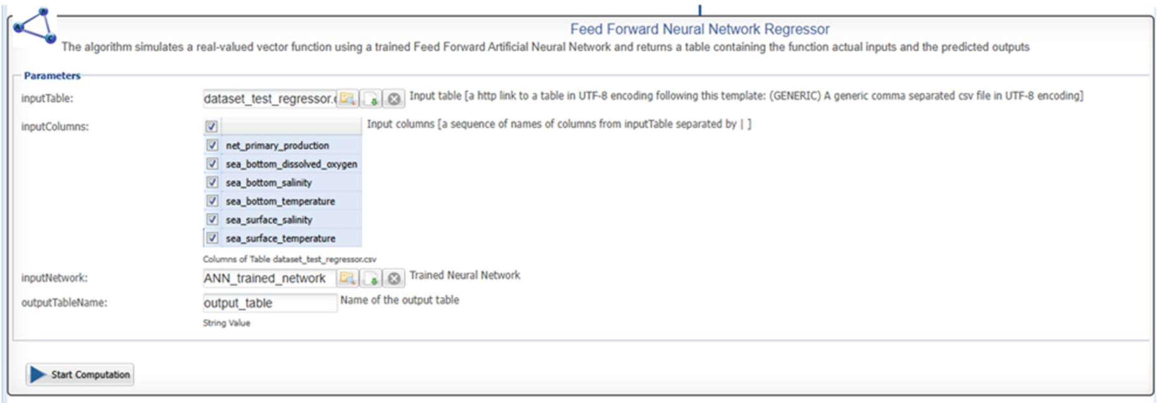

Figure 19. Feed Forward Neural Network Regressor features selection.

The Regressor takes as input a dataset of features homogeneous to those on which the model was trained, the model trained by the trainer and allows to select the features to be used for the forecast. The result is a table with the forecasts of the neural network on the input values.

Ecosystems: Habitat Representativeness Score

An implementation of the HRS algorithm presented in MacLeod et al. (2010). The habitat representativeness score (HRS) can be calculated to provide an objective assessment of whether a specific survey coverage will collect (or has collected) data that are representative of all available habitat variable combinations in an area. The operator takes two habitat tables as input. CSV files with one row for each feature vector, one column for each feature, and no headers. The output provides an assessment of the HRS value of the second habitat relative to the first as well as metadata associated with the computation.

Geo Processing: BIMAC

The Bayesian Interpolation Model with Advection-diffusion Constraint (BIMAC, Figure 20), introduced in (Coro, 2024), is an Open Science oriented, open-source, scalable and efficient workflow for 2D marine environmental parameters. It combines a fast, efficient interpolation method with a Bayesian hierarchical model embedding the stationary advection–diffusion equation as a constraint.

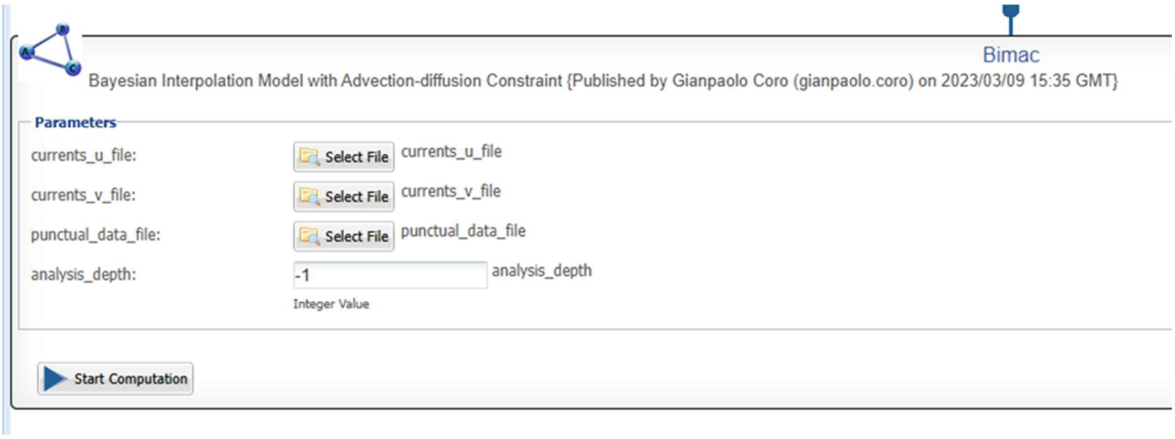

Figure 20. BIMAC Operator interface window.

The operator takes as input two optional files to describe the vectors of the marine currents in the area of interest, a .csv file with the detected point values and as output a raster file in .asc format in which the variable under examination is interpolated starting from the inputs.

Machine Learning Methods: Support Vector Machines Modelling

An algorithm to train a Support Vector Machine using Gaussian Radial Basis Function as kernel. The operator has input files similar to those of the neural network trainer, in addition to specific parameterization values of the support vector machines. Again, the output result is a trained model that can later be used by the projection operator.

Machine Learning Methods: Support Vector Machines Projector

An algorithm that applies a pre-trained Support Vector Machine model to a dataset (Figure 21). The Support Vector Machines Projector operator takes as input a previously trained SVM model and a table with features homogeneous to the training ones, allows one to select which of these to use and then produce a forecast according to the model.

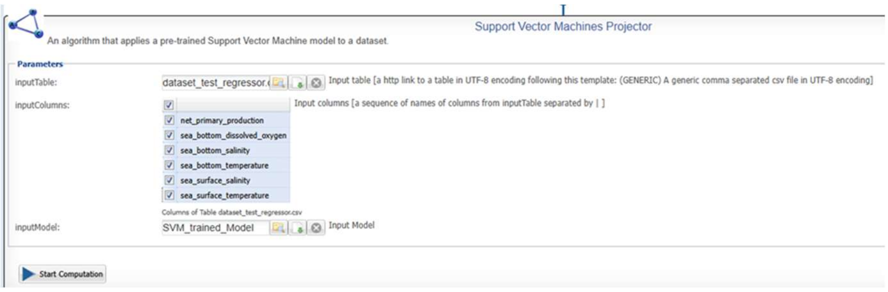

Figure 21. Support Vector Machines Projector interface window.

Occurrences: Get Species Occurrences Obis

This algorithm gets species occurrence points from the Ocean Biogeographic Information System (OBIS) repository using the species' scientific name (input). It returns a CSV file using Darwin Core-compliant field names specified in (Berghe et al., 2015)

Occurrences: Pseudo Absence Generator Obis

Pseudo absence point generator based on OBIS data statistical analysis. Described in (Coro et al., 2016). Estimating absence locations of marine species from data of scientific surveys in OBIS (Figure 22).

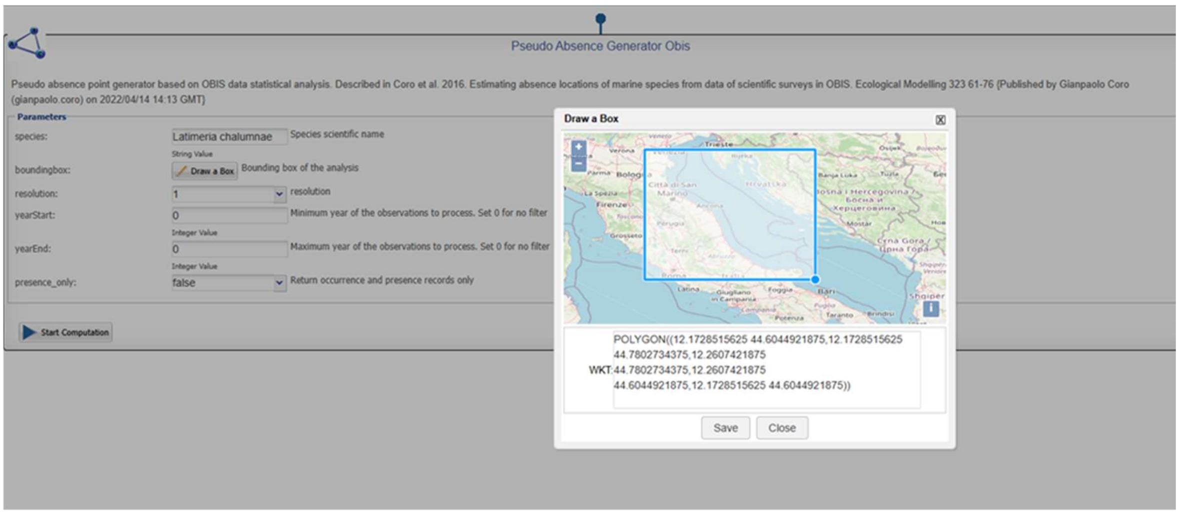

Figure 22. Pseudo absence point generator bounding box selection.

It is possible to define a bounding box and a time window in which to query the OBIS database to obtain both the points of presence of the species and to estimate those of pseudoabsence.

Species Simulation: Ecological Niche Modeller

A fully automatic workflow to estimate a species ecological niche, based on 4 models, i.e., Artificial Neural Networks, Maxent, Support Vector Machines, and AquaMaps. The workflow also produces an ensemble of these models as described in (Coro et al., 2024). This type of operator is often used to estimate the presence of a species in an area and, overall, to assess the richness of biodiversity present or expected in the future.

Species Simulation: Maxent Ecological Niche Model

A Maximum-Entropy model for species habitat modeling, based on the implementation by Shapiro et al. (1995) v 3.3.3k, Princeton University. The software accepts a table with species, longitude, latitude columns and a set of environmental variables as ASC files. Moreover it requires the specification of the species prevalence value.

Time Series: Singular Spectrum Analysis

Singular Spectrum Analysis for forecasting and reconstructing data time series. This technique is often used to produce future forecasts or reconstruct missing data series using historical datasets, as described in Coro et al. (2016).

2.6 The JupyterLab tool

The CZ VRE is equipped with JupyterLab (Figure 23), a web-based interactive environment for the creation and management of workflows of data, codes, and documents. It provides tools such as Jupyter Notebook, that can be used to create documents containing text, codes, plots, widgets etc. Jupyter Notebooks therefore allow to combine documentation, scripts and outputs in the same environment and are widely used by researchers and data scientists, for example, for educational purposes or debugging. Although originally developed for Python, more than 40 scripting languages are supported including R and Julia.

In the ITINERIS CZ VRE, researchers can create notebooks to help end-users to interactively explore and visualize data, but also incorporate scripts and processing tools.

The interactive layout of notebooks is useful for users to acquaint themselves with scripting languages, to customize or create their own tools.

As a demo application, a notebook for the visualization of CO_2 accumulation chambers data from the CZO in Pianosa Island is made available (Figure 24). The notebook contains a set of simple Python scripts to read the csv files that contain the measured variables (dataset's metadata have been published in the metadata catalogue). The user can select the time range of interest and the variable to plot.

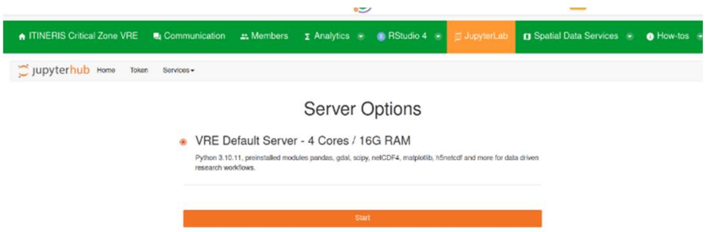

Figure 23. View of the JupyterLab launcher page. Users can access to this page by clicking on the "JupyterLab" button (on the top) and can see the specific information on the VRE default server.

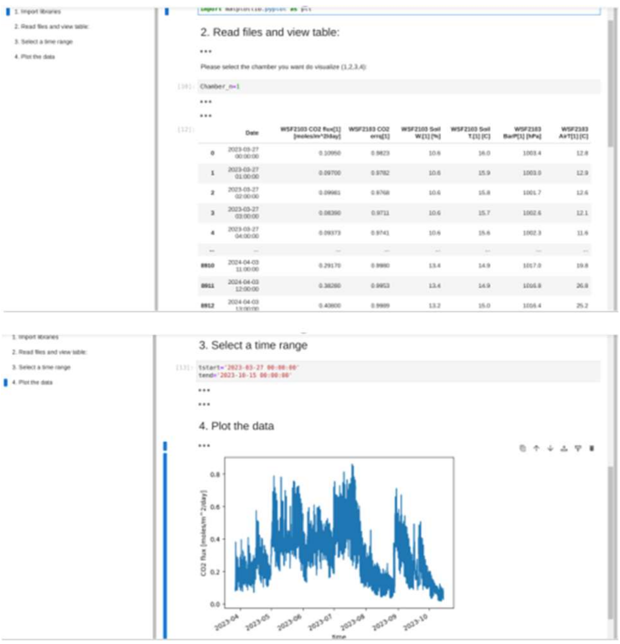

Figure 24. View of the Jupyter Notebook that allows to visualize the CO_2 accumulation chambers data from the CZO in Pianosa Island by selecting the time range of interest.

2.7 Demonstrators

2.7.1 Integrating the Open Global Glacier Model (OGGM)

A detailed pipeline is made available to allow users to set the Open Global Glacier Model (OGGM; Maussion et al., 2019) up on the VRE and to run some example codes. The initial setup code guides the user through the installation, configuration and testing of OGGM directly from the VRE Terminal (Figure 25a). This will allow to the users to avoid installing this complex model on their machine and, thus, running the analyses on the ready-for-use CZ VRE. After the installation and the setup of the model on the VRE, users can directly run the example codes imported from OGGM website. The examples codes are a series of Python tutorials (from https://tutorials.oggm.org/stable/book/10minutes.html, last access: 06/10/2024), which allow the user to run the first simple analyses with the open-source glacier model.
Moreover, codes that allow to the users to modify their own glaciological database in order to be compliant with the RGI format (v. 6.0; Randolph Glacier Inventory, 2017) – used in general by the model – is provided (Figure 25b).

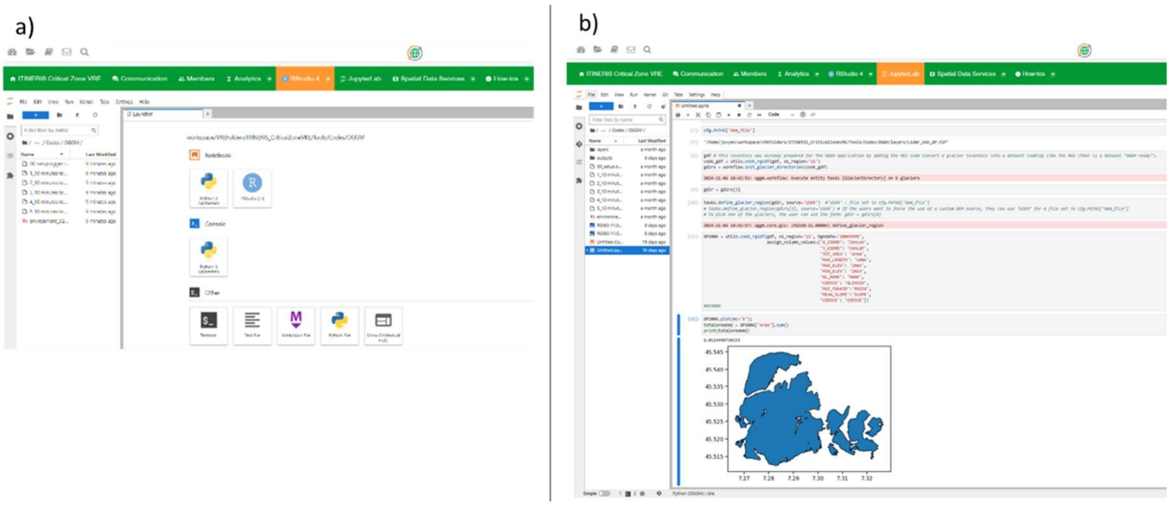

Figure 25. a) View of the Launcher panel. On the left side of the screen, users can navigate through the personal and collaborative workspaces. The figure shows the codes made available on the CZ VRE. b) View of a part of the code that allow users to modify their own glaciological database in order to be compliant with the RGI format The example code was performed by using glacier outlines which are part of the glaciological database referred to 2006 (Salvatore et al., 2015) built according to the WGMS guidelines (Paul et al., 2009, 2010).

2.7.2 High resolution soil temperature in glacier forelands

A step-by-step guide is made available to set up the R interface for Google Earth Engine (*rgee* R package – user registration required to access the Earth Engine services), and to produce maps of near-surface soil temperature in glacier forelands following the method proposed by Marta et al. (2023). Apart for the main code, a series of functions are provided to implement the functions from the *microclima* R package, and to composite ERA-Land images. Most of the layers needed to run the code are automatically retrieved from the GEE archives. Additional layers needed are made available in the workspace directory (/ITINERIS_CriticalZoneVRE/Tools/Codes/Marta et al 2023/layers). Specifically, they refer to a) an example shapefile of glacier outline (Valnontey, Cogne – 2006 glacier outlines, Salvatore et al., 2015); b) a global map of permafrost probability of occurrence (Gruber, 2012); c) a high-resolution digital terrain model (2008 LiDAR data of the Valle d'Aosta Region, available at https://geoportale.regione.vda.it/download/dtm/). Users can select their own period of interest (in term of both months and years) and use their own digital terrain model (in planar coordinate reference system) and glacier outline map to produce estimates for different periods and/or areas. As stated above, the user registration to the Google Earth Engine services is mandatory to run the code.

2.7.3 Non-linear modelling of CO_2 fluxes from alpine ecosystems

A R routine was developed to fit non-linear models for carbon dioxide (CO_2) fluxes measured using the flux chamber method in high-elevation or high-latitude ecosystems such as Arctic tundra and alpine grasslands (Lenzi et al., 2023; Magnani & Lenzi, 2024). The original codes were developed in MATLAB and allowed fitting exponential (for ecosystem respiration, ER) and Michaelis-Menten (for gross primary production, GPP) equations to flux measurements. In the framework of the ITINERIS PNRR Project, the MATLAB routines were translated and optimized for R in order to be compliant with FAIR and Open Science principles. The code integrated in the ITINERIS CZ VRE allows reproducing the model selection (AIC-based) and fitting process as proposed by Lenzi et al. (2023), and to produce the "Measured *vs* modelled" plot reported by the authors. Users can either run the code *as is* and inspect results or feed the model selection with a different set of environmental variables from the test dataset, or, finally, analyse their own dataset.

2.7.4 Visualizing CO_2 fluxes and environmental variables from glacier forelands data

A simple R routine was developed to summarize data, calculate gross primary production (GPP) and visualize trends in CO_2 fluxes measured using the accumulation chamber method along glacier forelands (Figure 26). Measurements were taken along three glacier forelands (Forni glacier – Ortles-Cevedale, and Lauson and Lavassey glaciers – Gran Paradiso), mostly at sites with known age (2 to approx. 180 years BP). By implementing a space-for-time substitution approach, patterns along the foreland can be interpreted as temporal trends, based either on age or on a development stage. Along with the trends in ecosystem respiration (ER) and net ecosystem exchange (NEE), trends in environmental variables known to influence CO_2 fluxes (solar irradiance, air temperature and relative humidity, soil temperature and volumetric water content) are also summarized and displayed. Users can simply inspect trends, try themselves to look for relationships between fluxes and environmental variables or adapt the code to summarize and visualize their own dataset.

Figure 26. Trends in CO_2 fluxes acquired by using the accumulation chamber method along glacier forelands. Violin diagram shows the ecosystem respiration (ER, yellow) and the gross primary production (GPP, green) at different development stages (increasing time from glacier melting).

2.8 Smoothing time series of CO_2 fluxes from automated chambers data

This R routine was developed to calculate gross primary production (GPP), fit a simple locally-weighted polynomial regression (i.e., LOWESS) and visualize trends in CO_2 fluxes measured using automated chambers. Automated chambers measure ecosystem respiration and net ecosystem exchange continuously, at a user-defined frequency, along with a series of environmental variables. Each pair of transparent (for NEE) and dark (for ER) chambers measures fluxes with a time lag, so that NEE and ER data need to be paired to calculate GPP. LOWESS regression requires the specification of the neighbourhood of interest (i.e., α), that is, the set of points influencing (via a tricubic weighting based on distance) the outcome for the ith point. Example data associated with the code refer to one year of hourly measurement of fluxes, with a 15-minute lag between ER and NEE; to represent different timescales (and levels of smoothing) we used α values of about 1/4, 1/12 and 1/52 to represent smoothing at the seasonal, monthly and weekly scales, respectively (Figure 27). Users can inspect trends from different automated chambers from the same place (e.g., Pianosa Island), adapt the code to compare trends between semi-arid and alpine ecosystems (e.g., Pianosa vs Nivolet datasets) or run the code on their own dataset.

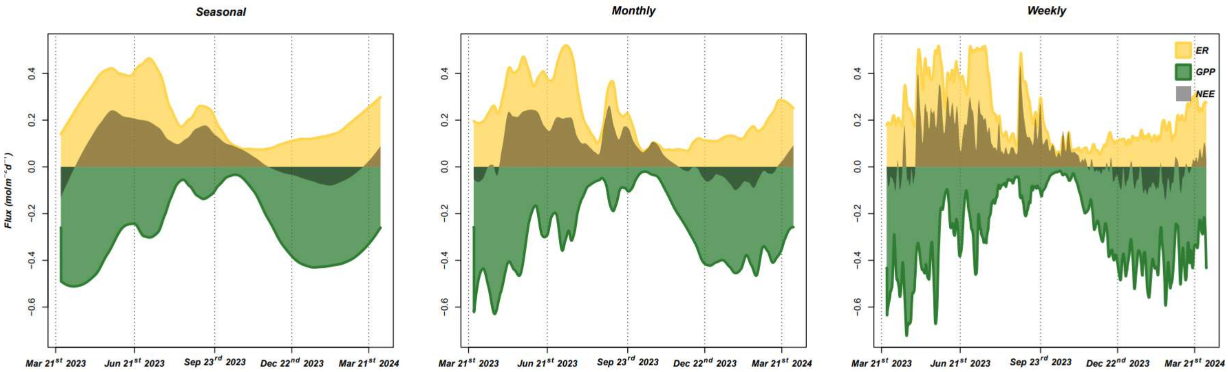

Figure 27. Annual trends in ecosystem respiration (ER, yellow), gross primary production (GPP, green) and net ecosystem exchange (NEE, brownish / greenish). Decreasing α (seasonal to monthly to weekly) produce trends with decreasing smoothing, allowing to better visualize short-term fluctuations in fluxes.

2.8.1 Visualizing CO_2 fluxes and environmental variables from Eddy Covariance data

A simple routine to visualize plots of eddy covariance dataset is proposed in the CZ VRE. The routine was developed in R language in order to be compliant with FAIR and Open Science principles and by using the RStudio Server directly integrated in the ITINERIS CZ VRE as RStudio, v4 Large (8 Cores / 32G RAM).

In particular, the code allows to the user to visualize CO_2 turbulent fluxes (μmol/m^2/s), wind speed (m/s) and other environmental variables (e.g., air temperature (°C) and RH (%)), saving a unique file with four different plots (Figure 28). Furthermore, simple graphs of the temporal evolution of CO_2 fluxes and wind speed were also performed (Figure 29). Considering the huge amounts of data collected by the eddy covariance towers, an example code to plot a subset of the whole dataset is provided (Figure 29).

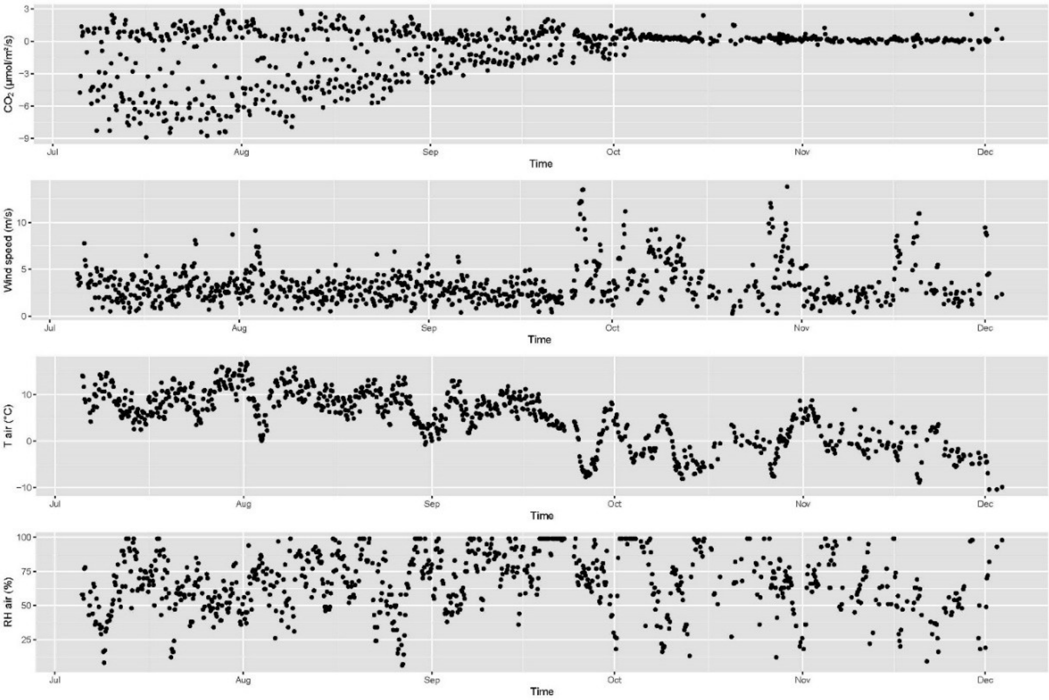

Figure 28. Temporal evolution of CO_2 turbulent flux, wind speed and environmental variables (air temperature and relative humidity) collected in 2020 from July to December by the eddy covariance tower installed at the Nivolet CZO. Dataset: Vivaldo et al., 2022.

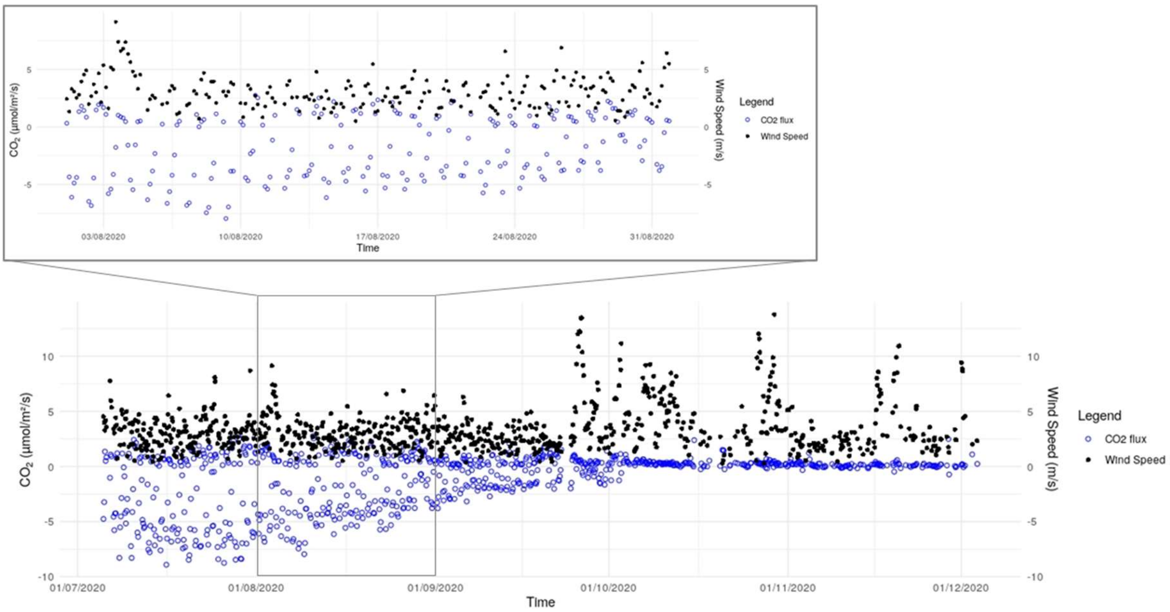

Figure 29. Temporal evolution of CO_2 turbulent flux and wind speed collected in 2020 from July to December by the eddy covariance tower installed at the Nivolet CZO and plot of a subset of the entire dataset. Dataset: Vivaldo et al., 2022.

3. REFERENCES

Ashley, G.M., 1998. Where are we headed? “Soft” rock research into the new millennium. Geological Society of America Abstract/Program, vol. 30. p. A-148.

Assante M., Candela L., Castelli D., et al., 2019. Enacting open science by D4Science. Futur. Gener. Comput. Syst., 101: 555-563. https://doi.org/10.1016/j.future.2019.05.063

Assante M., Candela L., Castelli D., Cirillo R., Coro G., Dell'Amico A., Frosini L., Lelii L., Lettere M., Mangiacrapa F., Pagano P., Panichi G., Piccioli T., Sinibaldi F., 2023. Virtual research environments co-creation: The D4Science experience. Concurrency Computat. Pract. Exper., 35(18):e6925. https://doi.org/10.1002/cpe.6925

Berghe, E.V., Coro, G., Bailly, N., Fiorellato, F., Aldemita, C., Ellenbroek, A. and Pagano, P., 2015. Retrieving taxa names from large biodiversity data collections using a flexible matching workflow. *Ecological Informatics*, *28*, pp.29-41. https://doi.org/10.1016/j.ecoinf.2015.05.004

Coro, G., 2024. An Open Science oriented Bayesian interpolation model for marine parameter observations. Environmental Modelling & Software, 172, p.105901. https://doi.org/10.1016/j.envsoft.2023.105901

Coro, G., Magliozzi, C., Berghe, E.V., Bailly, N., Ellenbroek, A. and Pagano, P., 2016. Estimating absence locations of marine species from data of scientific surveys in OBIS. Ecological Modelling, 323, pp.61-76. https://doi.org/10.1016/j.ecolmodel.2015.12.008

Coro, G., Sana, L., Bove, P., 2024. An open science automatic workflow for multi-model species distribution estimation. International Journal of Data Science and Analytics, pp.1-20. https://doi.org/10.1007/s41060-024-00517-w

Di Muri, C., Pulieri, M., Raho, D., et al., 2024. Assessing semantic interoperability in environmental sciences: variety of approaches and semantic artefacts. Sci Data 11, 1055. https://doi.org/10.1038/s41597-024-03669-3

Giardino, J., Houser, C., 2015. Principles and Dynamics of the Critical Zone. 19. Elsevier, Amsterdam, ISBN: 978-0-444-63369-9, 674 pp.

Gruber, S., 2012. Derivation and analysis of a high-resolution estimate of global permafrost zonation. *The Cryosphere*, *6*(1), 221-233. https://doi.org/10.5194/tc-6-221-2012

Lenzi, S., Magnani, M., Baneschi, I., Giamberini, M.S, Raco, B., Vivaldo, G., & Provenzale, A., 2023. Spatial and temporal variability of carbon dioxide fluxes in the Alpine Critical Zone: The case of the Nivolet Plain, Gran Paradiso National Park, Italy. *Plos one*, *18*(5), e0286268. https://doi.org/10.1371/journal.pone.0286268

MacLeod, C.D., 2010. Habitat representativeness score (HRS): a novel concept for objectively assessing the suitability of survey coverage for modelling the distribution of marine species. Journal of the Marine Biological Association of the United Kingdom, 90(7), pp.1269-1277. https://doi.org/10.1017/S0025315410000408

Magnani, M. & Lenzi, S., 2024. Codes for building and comparing multi regression models of carbon dioxide fluxes in the Alpine grasslands at Nivolet Plain, Gran Paradiso National Park (IGG-CNR-CZO@NIVOLET) (2.0). Zenodo. https://doi.org/10.5281/zenodo.12622506

Marta, S., Zimmer, A., Caccianiga, M., et al., 2023. Heterogeneous changes of soil microclimate in high mountains and glacier forelands. Nat Commun **14**, 5306. https://doi.org/10.1038/s41467-023-41063-6

Maussion, F., Butenko, A., Champollion, N., Dusch, M., Eis, J., Fourteau, K., Gregor, P., Jarosch, A. H., Landmann, J., Oesterle, F., Recinos, B., Rothenpieler, T., Vlug, A., Wild, C. T., Marzeion, B., 2019. The Open Global Glacier Model (OGGM) v1.1, Geosci. Model Dev., 12, 909-931. https://doi.org/10.5194/gmd-12-909-2019

Paul, F., Barry, R.G., Cogley, J.G., Frey, H., Haeberli, W., Ohmura, A., Ommanney, C.S.L., Raup, B., Rivera, A. & Zemp, M., 2009. Recommendations for the compilation of glacier inventory data from digital sources. Annals of Glaciology, 50 (53), 119-126. https://doi.org/10.3189/172756410790595778

Paul, F., Barry, R.G., Cogley, J.G., Frey, H., Haeberli, W., Ohmura, A., Ommanney, C.S.L., Raup, B., Rivera, A. & Zemp, M., 2010. Guidelines for the compilation of glacier inventory data from digital sources. World Glacier Monitoring Service (WGMS), Zurich, Switzerland, pp. 1-23. Link: http://globglacier.ch/docs/guidelines_inventory.pdf

RGI Consortium, 2017. Randolph Glacier Inventory - A Dataset of Global Glacier Outlines, Version 6.0. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center, https://doi.org/10.7265/4m1f-gd79

Salvatore, M. C., Zanoner, T., Baroni, C., Carton, A., Banchieri, F. A., Viani, C., Giardino, M., Perotti, L., 2015. The state of Italian glaciers: A snapshot of the 2006-2007 hydrological period. *Geografia Fisica e Dinamica Quaternaria*, *38*(2), 175-198. https://doi.org/10.4461/GFDQ.2015.38.16

Shapiro, J.L. and Prügel-Bennett, A., 1995, April. Maximum entropy analysis of genetic algorithm operators. In AISB Workshop on Evolutionary Computing (pp. 14-24). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-60469-3_21

Vivaldo, G., Baneschi, I., Boiani, V., Catania, M., Giamberini, M.S., Lenzi, S., Magnani, M., Raco, B., Parisi, A., Pennisi, M., Provenzale, A., 2022. Eddy Covariance data from ICOS-associated station IT-NIV at Nivolet Plain, Gran Paradiso National Park, 2020 (IGG-CNR-CZO@NIVOLET) (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6428619