Deep Earth System Data Laboratory (DeepESDL)

Anca Anghelea [1], Ewelina Dobrowolska[2], Gunnar Brandt[4], Martin Reinhardt[3], Miguel Mahecha[3], Tejas Morbagal Harish[4], Stephan Meissl[5]

[1] Department of Climate, Sustainability and Science, European Space Agency, 00044 Frascati, Italy – anca.anghelea@esa.int
[2] Serco Italia, 00044 Frascati, Italy, – Ewelina.Dobrowolska – ewelina.dobrowolska@serco.com
[3]Remote Sensing Centre for Earth System Research, University of Leipzig, Germany – martin.reinhardt.3@uni-leipzig.de
[4] Brockmann Consult, 21029 Hamburg, Germany – gunnar.brandt@brockmann-consult.de
[5]EOX IT Services GmbH, 1090 Wien, Austria – stephan.meissl@eox.at

Keywords: Earth System Science, Multivariate analysis, Earth Observation, Open Source, Open Science, Artificial Intelligence

Abstract

The Deep Earth System Data Lab (DeepESDL) provides an AI-ready, collaborative environment for researchers aiming to study the Earth's complex dynamics using various datasets and empirical approaches. Recently opened to Early Adopters, it builds on projects like CAB-LAB and ESDL, utilizing well-established Python and Julia technology stacks. DeepESDL offers programmatic access to extensive analysis-ready data cubes and computational resources, enabling researchers to focus on analysis without extensive preparations. Scientists can use persistently available data cubes or generate user-tailored cubes from own data or publicly available datasets. The goal is to streamline data processing through empirical or AI methods within high-dimensional Earth Observation workflows. DeepESDL addresses the complete research cycle, from discovery of earth data to powerful analyses, collaborative scientific research, advanced data visualisation and publication of results, promoting FAIR and Open Science. Apart from serving as a research environment, DeepESDL showcases scientific use cases and supports educational purposes through capacity building, academic programs, and Open Science initiatives. This paper presents an overview of DeepESDL.

1. Introduction

The Deep Earth System Data Lab (or DeepESDL, https://earthsystemdatalab.net) provides an AI-ready, collaborative environment enabling researchers to understand the complex dynamics of the Earth System using numerous datasets and multi-variate, empirical approaches.

The solution (which has recently opened to selected Early Adopters with a baseline set of features and support) builds on the work done in the frame of projects such as CAB-LAB and ESDL which established the technical foundations and created measurable value for the scientific community (Linscheid 2020), (Krich 2020), (Flach 2018), (Mahecha 2020), (Estupinan-Suarez 2024).

DeepESDL relies primarily on the well-established and complete technology stacks in Python (and Julia), which have become quasi standards in the research communities far beyond the users of ESDL and ensure their availability in well-tested and maintained ready-to-use environments.

The core of the DeepESDL is represented by the provision of programmatic access to a large offer of analysis-ready data organised in tailored cubes combined with adequate computational resources and capabilities to allow researchers to focus on analysis and processing without costly preparations, which is often directly at the expense of scientific work.

Scientists can make use of the persistently available datacubes, readily provided by the project (such as the Earth System Data Cube) or can use the tools made available by DeepESDL, such as open-source "recipes" (DeepESDL 2024) for the generation of user-tailored datacubes (DeepESDL 2024a).

Ultimately, the aim is to support the user experience towards an efficient processing of data (through empirical methods or AI approaches) within a scientific workflow that requires high-dimensional Earth Observation (EO) or derived data products.

With increasing number of available data sets, their discoverability and documentation become more and more important, particularly for those approaches that involve multiple variables, for which the researcher may not always be a thematic expert. To address this, DeepESDL is thus providing an informative catalogue to find all available data and to find all required metainformation describing them. This includes not only standard information e.g., regarding spatial and temporal coverage, versioning, but also on the specific the transformation methods applied during data cube generation which may, for instance reduce numerical accuracy, or any other known limitation relevant to the researcher.

The system design has openness, collaboration, and dissemination as key guiding principles. As science teams need proper tooling support to efficiently work together in this virtual environment, one of the key elements of the architecture is represented by the DeepESDL Hub, providing teams of scientific users with the means for collaboration and exchange of versioned results, source codes, models, execution parameters, and other artifacts and outcomes of their activities in a simple, safe, and reliable way.

Apart from providing a technological solution, DeepESDL looks to provide a collection of scientific use cases demonstrating the various platform capabilities, and providing resources and examples for researchers in Earth Sciences. Additionally, DeepESDL is serving educational purposes, being actively used in capacity building for Earth Observation, academic programmes as well as supporting competitions and Open Science intiaitives.

This paper gives an overview of DeepESDL and how it contributes to the scientific objectives of an increasing number of scientific users.

2. DeepESDL Architecture

DeepESDL is a platform providing analysis-ready data cube services in a powerful, virtual laboratory to the Earth Science research community. The DeepESDL suite of services facilitate data exploitation, collaborating on data and source code, and publication of results. DeepESDL supports Machine Learning and artificial intelligence approaches, including the preparation of AI-ready datasets, providing a programming environment with relevant libraries and packages, and the resources to execute processing pipelines.

2.1. DeepESDL Applications

DeepESDL provides a suite of applications and services designed to facilitate the use and analysis of Earth system data. These tools and services are intended to enhance accessibility, visualization, and processing capabilities for both individual researchers and projects. DeepESDL includes several applications, as detailed below.

T**he xcube Catalogue** (in development) is a tool for browsing available data sets. It aims to provide a user-friendly entry point for accessing data, similar to ESA’s EuroDataCube Collections.

Visualization Tools with various levels of integration and complementary capabilities, include adapted versions of the xcube Viewer (xcube viewer 2024), Lexcube (Lexcube 2024) and a 4D Viewer, which are designed to enhance the visual analysis of data.

The Jupyter Notebook Service allows users to run Jupyter Notebooks tailored to specific Machine Learning (ML) environments related to their projects. This service facilitates the use of interactive and reproducible research tools.

Additionally, DeepESDL provides access to a number of services, including:

1. **xcube Server**: This service allows users to browse, access, and publish gridded data cubes. It plays a crucial role in managing large-scale geospatial data.
2. **geoDB**: A database instance that supports the browsing, accessing, and publishing of vector datasets, providing essential tools for spatial data management.
3. **Workspace**: This is a collaborative environment where projects can store and share various types of data, including ML workflows. This facilitates teamwork and data sharing.
4. **Access to Data Cubes**: Projects have access to both shared and project-specific data cubes stored in an object storage system. This ensures data availability and ease of access.

DeepESDL common resources include a cluster of worker nodes and a common xcube ARDC (Analysis-Ready Data Cubes) Service, which can access various data services and convert them into ARDC. The data cubes in DeepESDL maintain a consistent structure and format, making them uniform and easier to manage.

Data access is possible in two main manners within DeepESDL. xcube data stores provide on-the-fly access to datasets via the data stores framework. Other datasets are made anaylsis-ready and persisted in object storage for fastest access.

2.2. Support for Machine Learning on Earth data

Furthermore, the DeepESDL supports the implementation and execution of Machine Learning workflows on Analysis Ready Data Cubes in a reproducible and FAIR way, allowing sharing and versioning of all ML artifacts like code, data, models, execution parameters, metrics, and results as well as tracking each step in the ML workflows (supported by integration with Open-Source tools like TensorBoard or Mlflow) for an experiment so that others can reproduce them and contribute.

To support efficient machine learning, the ARDC Generator uses a configurable ML sampling module for training and validation. Worker nodes provide GPU acceleration for intensive ML training tasks. Core libraries like Keras/Tensorflow and PyTorch are available, along with advanced tools for model evaluation such as TensorBoard and MLflow.

2.3. Published Project Data

DeepESDL supports the publication of project data through a common xcube Server and geoDB, serving the public catalogue and visualization tools. A public REST API is available for programmatic access to these services. Additionally, a Jupyter Book (JB) or Notebook Viewer (NBviewer) service will allow selected Project-JNBs to be published as story-telling books, aiding in the dissemination of scientific information.

2.4. System Architecture

The numerous services and applications in DeepESDL are containerized and executed in a common cloud environment. Kubernetes (K8s), a popular container orchestrator, is used for orchestration, monitoring, maintenance, and scaling. This ensures that DeepESDL is deployable on any cloud environment capable of running a K8s service, including AWS, Google, Microsoft Azure, and DIAS instances. The initial deployment is on AWS in the EU-Central-1 region, located in Frankfurt, Germany, utilizing the managed K8s service EKS for cost efficiency and reliability.

2.5. Technology Foundation:

To meet its ambitious goals, DeepESDL relies heavily on existing technologies. Many system components have been developed in previous ESA projects and are reused, adapted, extended, and branded for DeepESDL. Examples include:

- **DeepESDL Hub**, based on EOxHub used in the ESA EuroDataCube project.
- **xcube tools**: The xcube cube generator tools, xcube Server, and xcube Viewer are derived from components in the xcube Toolkit.
- **Cube 4D Viewer**, an adapted version of Earthwave’s 4D Viewer.

This strategic reuse of established technologies ensures robustness, cost-efficiency, and reliability in DeepESDL’s services. The complete documentation for the DeepESDL is available at (DeepESDL 2024b).

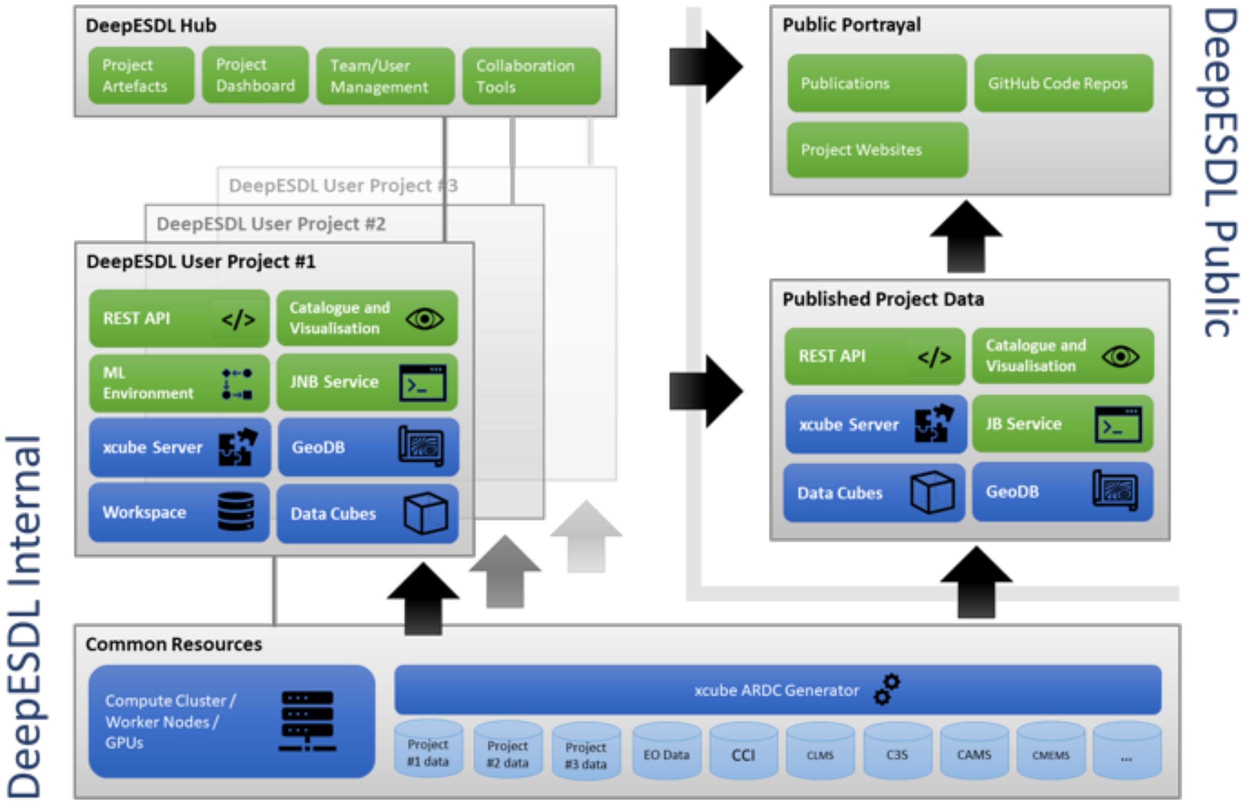

Figure 1. High-level concepts and the architecture of the DeepESDL project. It comprises the internal DeepESDL Hub and user projects and the publicly visible parts. Both are served by a common infrastructure and common resources.

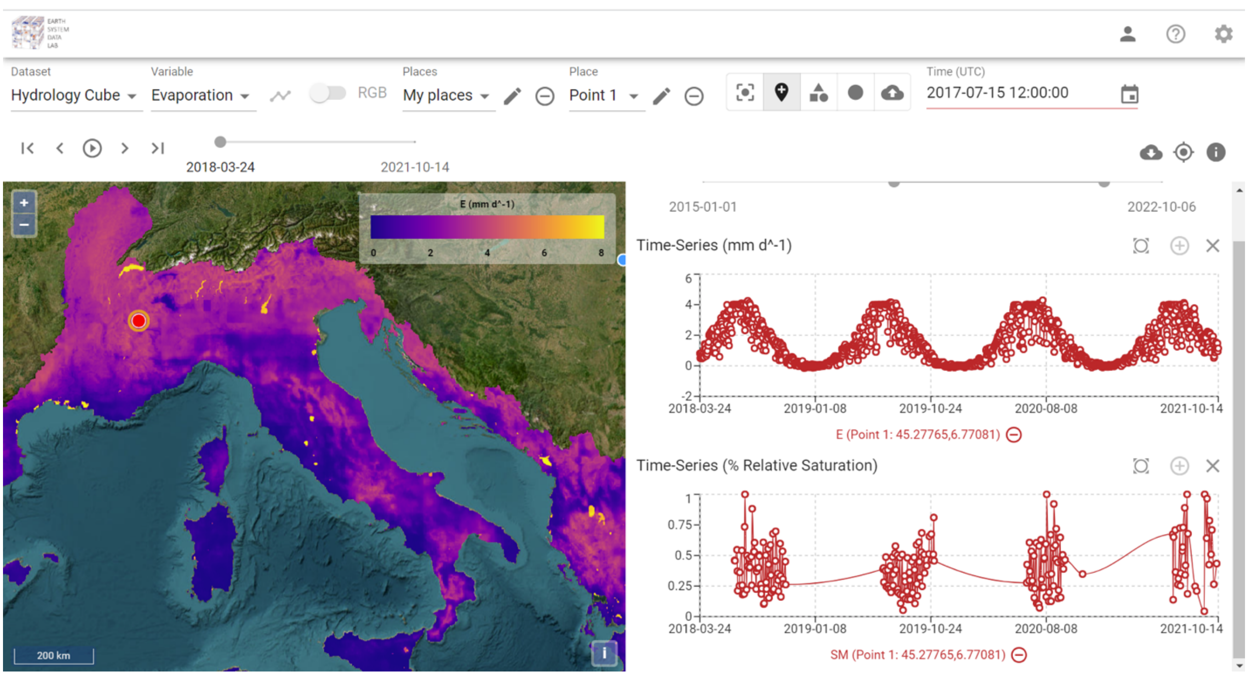

Figure 2. DeepESDL xcube viewer [7] is a simple, single-page web application to be used with the xcube server. The viewer contains public datasets only, but will later also provide user/team cubes when logged in. The login will be the same as for the DeepESDL JupyterLab.

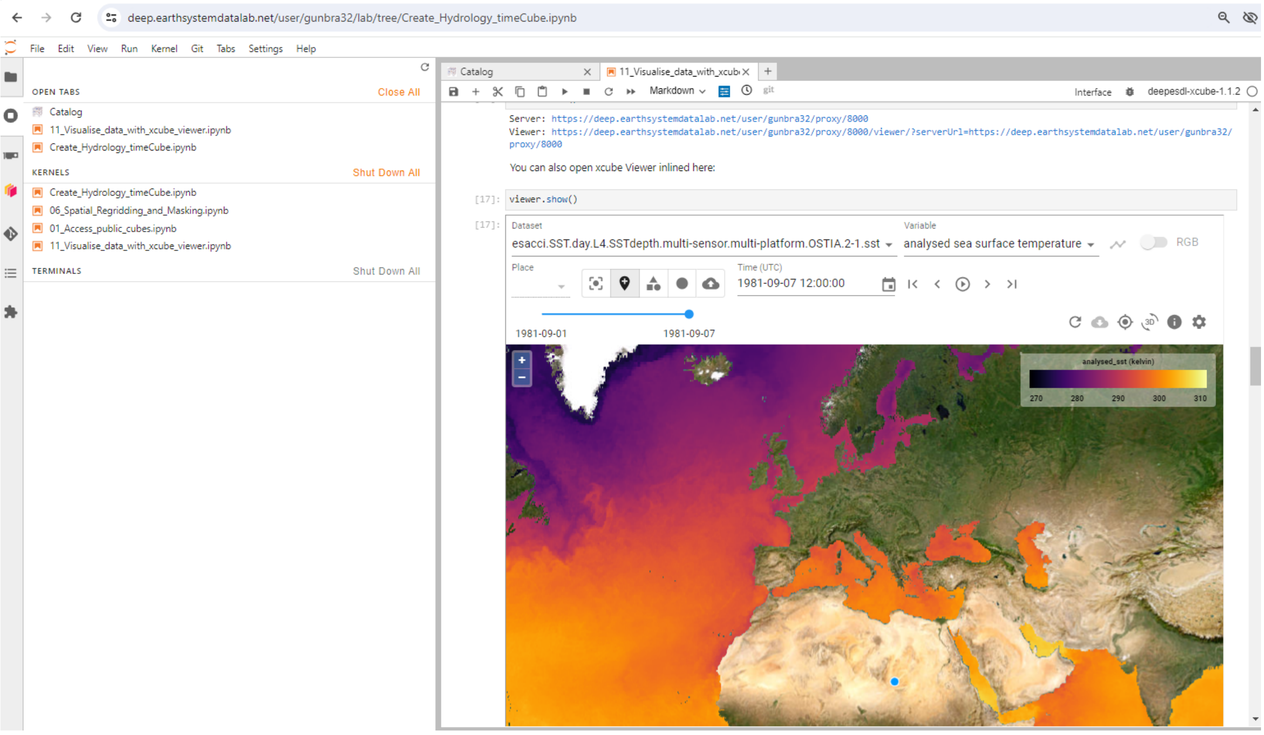

Figure 3. Besides the public instance of the xcube viewer, which comprises published versions of data cubes, the application is also available as a versatile and useful tool for exploring their intermediate results in the development phase of their work. To this end, xcube viewer is available as an on-demand self-service within the DeepESDL Hub. Users may simply start an xcube server in a Notebook using the resources of their Jupyter session, add local or remote data sets, and start a viewer comprising the datasets, which have been added to the server. This entire process is very intuitive and lean and as simple as a common plot command. It provides the user, however, with a fully functional viewer instance, which allows for interactive inspection of several data cubes. The following screenshot show the viewer inside a Juypter Notebook cell.

Finally, dissemination is essential for the Open Science spirit of the DeepESDL. In the predecessor ESDL project, the visualisation of the ESDC was primary done through a web-based xcube Viewer and using the multiple plotting capabilities of Jupyter Notebooks. DeepESDL goes beyond that, enhancing the xcube Viewer into a visualisation and data analytics toolbox and adding a special viewer app that allows for spectacular and immersive 4D visualisations and animations. Both xcube Viewer and 4D Viewer use the same RESTful data service API provided by xcube Server. The latter also provides an OGC WMTS so other OGC-compliant applications, such as QGIS3, will be able to visualise DeepESDL’s ARDCs.

To foster collaboration, additional features such as publishing individual Jupyter Notebooks as storytelling documents or even books using Jupyter Books or the Executable Book Project are being explored, together with concepts such as storytelling and DeepESDL User Project Dashboards which may also link to the viewers and Notebooks.

3. Scientific Visualisation

DeepESDL offers three main applications for scientific data visualisation, each catering to a different need and addressing users with varying levels of expertise, from advanced users that have an intimate knowledge of the data and require close interactions and manipulations, to higher level users for which intuitive and potentially immersive visualisations are more important.

Each application has specific features to disseminate data cubes generated in the platform.

The xcube viewer is the default viewer with image and time-series functionalities that works out-of-the box with all data cubes complying to the xcube dataset convention. It is available as a public viewer with a fixed URL (Figure 2) as well as in the DeepESDL Hub, where it may be launched on demand by users (Figure 3).

Secondly, the 4D viewer offers unique capabilities of rendering three-dimensional data onto a fourth dimension, e.g., on a terrain. It offers sophisticated features to create scenes by defining the rendering of surfaces and the characteristics of light sources. xcube server may be used as a back end providing image tiles to the 4D viewer, thus connecting all data cube generated within DeepESDL to the application. The 4D Viewer, initially created as part of the polar cluster digital twin Earth precursor project Digital Twin Antarctica, is a visualisation tool which provides a truly 4D immersive experience for the user. It has been developed using the Unity game engine and combines functionalities of standard geospatial analysis platforms within a 4D framework. It allows the construction of 3D worlds, the projection of datasets as layers within that world, and finally the evolution of these datasets through time. A web build of the 4D viewer, set up to provide data cubes accessible on the public DeepESDL S3 bucket, is available at https://esdl-4d-viewer.web.app/.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-2024-13-2024 | © Author(s) 2024. CC BY 4.0 License.

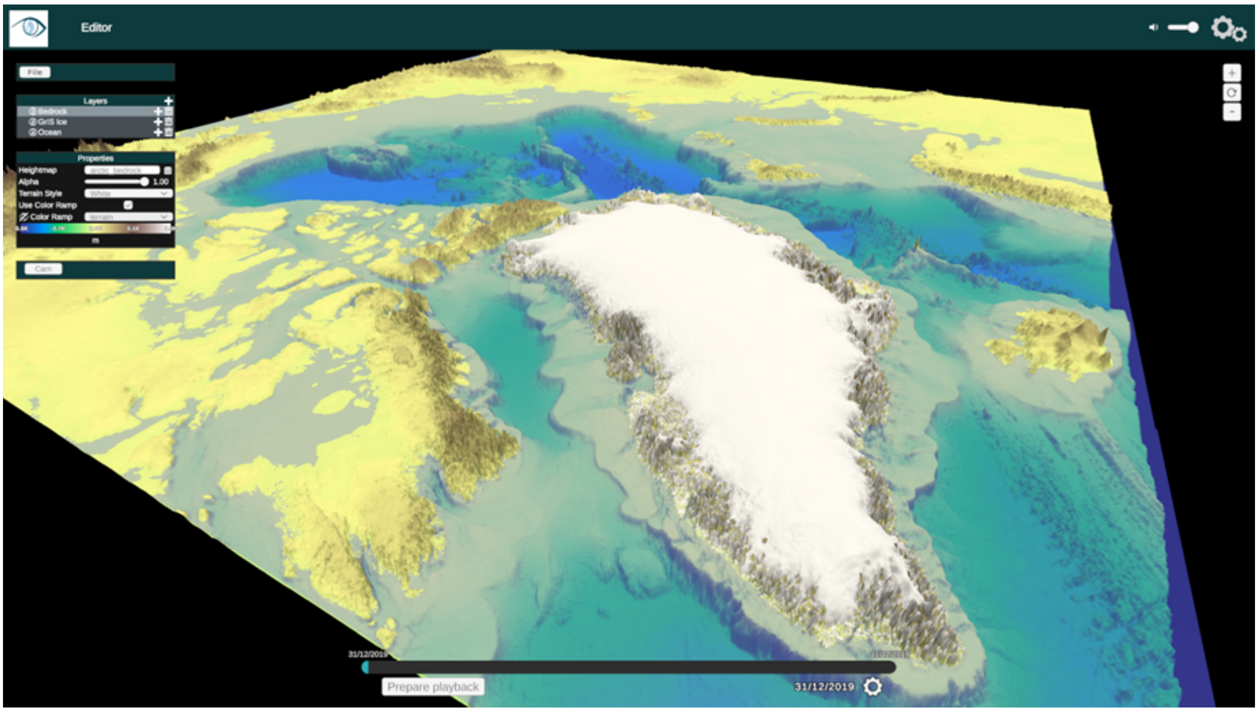

Figure 4. 4D Viewer examples showing: base terrain over the Arctic circle, including a DEM for the Greenland ice sheet

Finally, Lexcube [10] offers sophisticated 3D visualisation capabilities and scrolling through data cubes across any dimension. It also comes with a Jupyter Hub integration allowing for users to launch an instance with their own data. Lexcube is a library for interactively visualizing three-dimensional floating-point data as 3D cubes in Jupyter notebooks. It supports several data formats including numpy.ndarray (with exactly 3 dimensions), xarray.DataArray (with exactly 3 dimensions, rectangularly gridded). Possible data sources are any gridded Zarr or NetCDF data set (local or remote, e.g., accessed with S3), Copernicus Data Storage, e.g., ERA5 data, Google Earth Engine (using xee). The library is available at (Lexcube 2024a).

4. Scientific User Community and Community

Up to now, DeepESDL has been in a hybrid mode comprising substantial development and deployment activities but also continuous operations for a growing number of users. Essential to the adoption is the provision of the DeepESDL as part of ESA’s Network of Resources portfolio (ESA NoR 2024). The Network of Resources (NoR) is an initiative by the European Space Agency that facilitates the use of cloud environments for Earth Observation (EO) data, supporting research, development, and pre-commercial users in transitioning from a data download paradigm to a “bring the user to the data” paradigm.

Through this mechanism, ESA supports an equitable access to science and computational resources, enabling worldwide researchers to access fit-for-purpose platforms and conduct meaningful studies, potentially at large scale. Since its availability on the NoR portfolio, a growing number of scientific users have been onboarded to the DeepESDL. Notably, due to the collaborative features and the design “fit for teams”, DeepESDL is being adopted by a growing number of activities, including, e.g., DeepExtremes, EO4Health, or the Hydrology4D projects.

Finally, DeepESDL is supporting a growing number of educational and capacity building activities, including the new “Earth System Science Challenges” (ESA Science Hub 2024) organised at the ESRIN Science Hub, in which students typically at PhD or Master level are attempting to solve Earth data science problems applied to concrete scientific questions and use cases.

5. Conclusions

This paper made a summary of the key features and capabilities of DeepESDL – AI-ready, collaborative environment for researchers aiming to study the Earth's complex dynamics using various datasets and empirical approaches, which addresses the full research cycle while promoting a FAIR and Open approach to science.

References

Linscheid N., Estupinan-Suarez L. M., Brenning A., Carvalhais N., Cremer F., Gans F., Rammig A., Reichstein M., Sierra C. A., Mahecha M. D., (2020): Towards a global understanding of vegetation–climate dynamics at multiple timescales. *Biogeosciences*, 17, 945–962, https://doi.org/10.5194/bg-17-945-2020

Krich C., Runge J., Miralles, D. G., Migliavacca M., Perez-Priego O., El-Madany T., Carrara A., Mahecha M. D. (2020): Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach. *Biogeosciences*, 17, 1033–1061, https://doi.org/10.5194/bg-17-1033-2020

Flach M., Sippel S., Gans F., Bastos A., Brenning A., Reichstein M., Mahecha M. D. (2018): Contrasting biosphere responses to hydrometeorological extremes: revisiting the 2010 western Russian heatwave. *Biogeosciences*, 15, 6067–6085, 2018, https://doi.org/10.5194/bg-15-6067-2018

Mahecha M. D., Gans F., Brandt G., Christiansen R., Cornell S. E., Fomferra N., Kraemer G., Peters J., Bodesheim P., Camps-Valls G., Donges J. F., Dorigo W., Estupinan-Suarez L. M., Gutierrez-Velez V. H., Gutwin M., Jung M., Londoño M. C., Miralles D. G., Papastefanou P., Reichstein M. (2020): Earth system data cubes unravel global multivariate dynamics. *Earth System Dynamics*, 11, 201–234, https://doi.org/10.5194/esd-11-201-2020

Estupinan-Suarez, L. M., Mahecha, M. D., Brenning, A., Kraemer, G., Poveda, G., Reichstein, M., & Sierra, C. A. (2024). Spatial patterns of vegetation activity related to ENSO in northern South America. *Journal of Geophysical Research: Biogeosciences*, 129(1), https://doi.org/10.1029/2022JG007344

DeepESDL 2024, Datasets available in DeepESDL https://deepesdl.readthedocs.io/en/latest/datasets/datasets/ (01 July 2024)

DeepESDL 2024a, DeepESDL Cube Generation Recipes https://github.com/deepesdl/cube-gen (01 July 2024)

xcube viewer 2024, xcube viewer application documentation https://xcube.readthedocs.io/en/latest/viewer.html# (01 July 2024)

DeepESDL 2024b, DeepESDL Documentation https://deepesdl.readthedocs.io/ (01 July 2024)

ESA NoR 2024, Network of Resources https://nor-discover.org/ (01 July 2024)

Lexcube 2024, Lexcube viewer www.lexcube.org (01 July 2024)

Lexcube 2024a, Lexcube Project https://github.com/msoechting/lexcube (01 July 2024)

ESA Science Hub 2024, Earth System Science Hub Challenge https://sciencehub.esa.int/2024/05/09/3rd-earth-system-science-challenge/ (01 July 2024)