IS-ENES3 Milestone M5.4

Compute service roadmap

Reporting period: 01/01/2022 to 31/03/2023

Authors: P. Nassisi (CMCC), A. Nuzzo (CMCC), F. Antonio (CMCC), G. Levavasseur (IPSL), S. Kindermann (DKRZ), M. Juckes (UKRI), P. Kershaw (UKRI), S. Fiore (UNITN), A. Spinuso (KNMI), C. Pagé (CERFACS)

Reviewers: M. Lautenschlager (DKRZ), F. Adloff (DKRZ)

Release date: 30/03/2022

ABSTRACT

This document outlines the long-term strategic roadmap of the ENES Climate Data Infrastructure compute services, starting from the requirements collected so far and taking other international efforts into account with referece to climate data analysis. It describes the status of the IS-ENES3 compute services, together with some sustainability aspects, and the broader landscape of the ESGF, EGI and Copernicus initiatives, allowing the definition of some pillars for the future of climate simulations analysis.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Table of contents

1. Introduction and goals 3
2. The ENES Climate Data Infrastructure and the compute services 4
3. Ongoing initiatives and related medium/long-term plans 8
   3.1 ESGF 8
   3.2 ENES Data Space and EGI-ACE 8
   3.3 Copernicus 10
4. Final considerations and future plans 13

1. Introduction and goals

The goal of this document is to outline a long-term strategy for the IS-ENES compute services, leveraging the work done during the last three years and the requirements collected and widely described in the project documents.

Being the compute services a relevant component of the ENES Climate Data Infrastructure (ENES CDI), which strongly relies on a close collaboration between well-established large climate compute and data centres in Europe, such a roadmap needs to carefully consider the individual institutional viewpoints as well as the ongoing and future European data infrastructure efforts and requirements of climate model data analysts.

The requirements collected so far result from an accurate analysis of the users' needs conducted within the activities of WP5, WP7 and WP10 and concluded with the production of the documents D5.1 "Compute service requirements and state of the art approaches" (M12) and M10.1 "Technical requirements on the software stack" (M14), which led to the definition of the short/medium term (IS-ENES3 timeframe) roadmap of the overall ENES CDI architecture (D10.1 Architectural document of the ENES CDI software stack (M18)).

Besides that, other notable inputs coming from project meetings, workshops, European strategies and guidelines, have contributed to the definition of new requirements and the refinement of the initial ones.

The document is structured in two main sections with some final considerations. More specifically, Section 2 describes the status of the ENES CDI and its compute services, with some references to the next steps and sustainability aspects of the analytics layer. Section 3 broadens the perspective towards other European initiatives like ESGF, EGI and Copernicus. Section 4 outlines some of the pillars that should motivate future computing services.

2. The ENES Climate Data Infrastructure and the compute services

The ENES Climate Data Infrastructure consists of a collection of services, software, and metadata specifications, which are functional to the analysis and exploitation of CMIP and CORDEX simulations with the aim of efficiently supporting decisions in the climate science domain. D10.1 presents the general architecture of the envisaged infrastructure, with technical expectations for the mid- to long-term implementation. A key component of the ENES CDI is represented by the compute layer, which fully addresses computing needs and moves towards a sustainable and integrated data analytics and processing model of CMIP6 and CORDEX data.

The ENES CDI compute services currently rely on the design carried out during the RP1 and consider the requirements coming from WP5/NA4.

Beyond the different implementations of the core analytics services developed at each site (whose development/choice was mainly driven by institutional and national requirements), the main direction of such activity in IS-ENES was to move towards a common data analytics and processing design, while preserving, at the same time, institutional needs with proper customizations.

The component diagram in Figure 1 shows the compute service architecture design that will be delivered by the end of IS-ENES3. Such a diagram highlights the strong link of each institutional implementation with the core data distribution services (ESGF data nodes) to foster the “data near processing capabilities” paradigm. It also details security aspects (grey boxes) to ensure legitimate use of the computing resources at the host facility by external users.

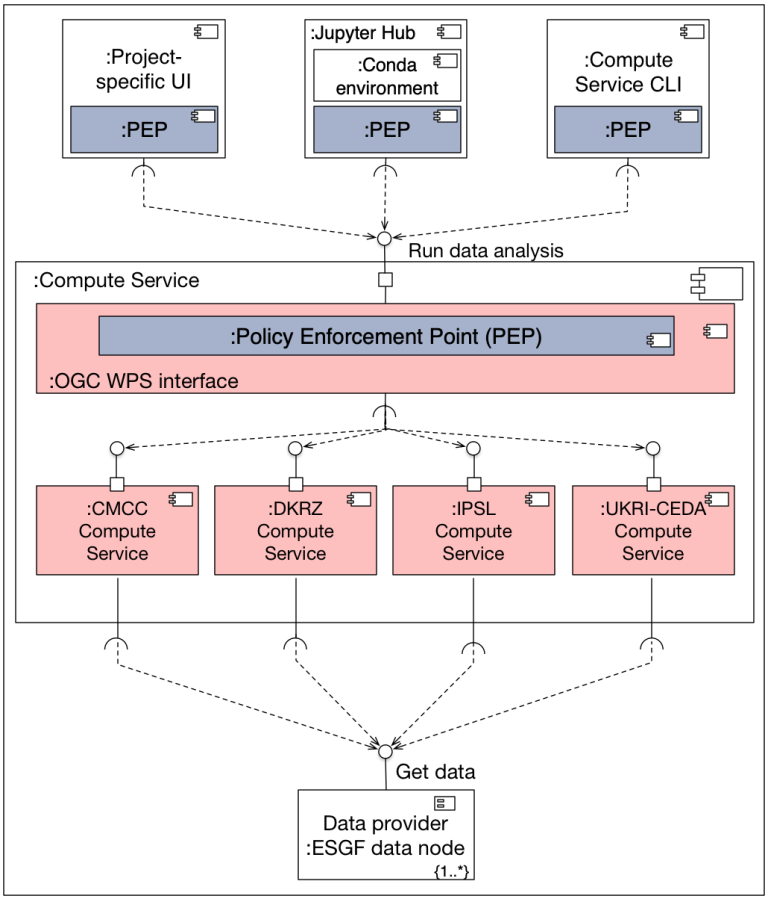

Fig. 1: ENES CDI compute layer, which provides unified access to institutional implementations of the service through an OGC WPS[1] interface and the Identity Management and Access control layer (PEP)

The IS-ENES3 compute service infrastructure is composed of four distinct facilities, hosted at DKRZ, CNRS-IPSL, UKRI and CMCC, strictly connected with institutional and federated data pools (ESGF catalogues).

The four different service implementations share the following three common aspects:

- an interoperable and flexible server front-end based on an Open Geospatial Consortium Web Processing Service (OGC-WPS) compliant interface;
- a programmatic client interface with a Python binding;
- a security infrastructure based on the work and roadmap defined with the ESGF Identity, Entitlement and Access Management (IdEA) working group activity.

[1] https://www.ogc.org/standards/wps

The access interface, based on the standard OGC-WPS, ensures the interoperability with several different clients and applications as project-specific UIs, Conda environments and Command Line Interfaces (CLI).

A Python environment based on JupyterHub is currently available on the four different deployments of the compute services. Moreover, a first implementation of a WPS layer for Copernicus has been deployed at DKRZ, allowing users to perform some basic analysis on the large CMIP data pools available. Specifically, a subsetting service called Roocs[2] performs time and area subsettings and could be extended to also provide averaging and regridding operations. During the last year of the IS-ENES3 project, the WPS will be deployed by more data providers of the ENES-CDI, being one of the common components in the compute service architecture.

The Compute Service layer also addresses the specific users' needs of another ENES CDI component, the Climate4Impact (C4I) portal[3], through which users can perform some data reduction operations on the different data pools before executing further local elaborations. All these operations produce provenance information that will be managed to foster traceability of the computations occurring across the different sites of the ENES-CDI.

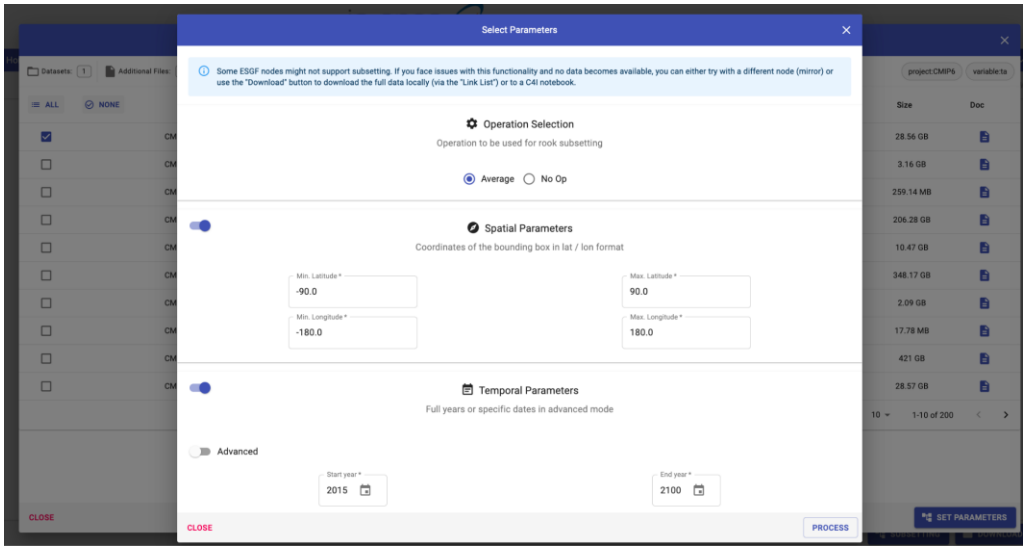

Figure 2: Interface in C4I for the interactive configuration of data reduction workflows accessing remote WPS.

[2] https://roocs.github.io/

[3] https://climate4impact.eu/impactportal/general/index.jsp

Additional ready-to-use compute environments and associated notebook kernels have also been made available; for example, pre-established ESMValTool kernels for the JupyterHub installation at DKRZ allow further analysis on model data, as well as notebooks using icclim to calculate some climate indices in a multi-scenarios approach with output example plots. Moreover, a compute service based on JupyterHub and the Ophidia framework has been developed at CMCC, exposing its capabilities through the already mentioned WPS interface and providing a variable-centric data catalogue to support additional needs of the scientific community in terms of indicators/indices calculation and multi-model analyses.

Moreover, the IS-ENES3 project has launched a new service to support server-side analysis of climate projections through a Transnational Access activity. The analysis platforms at DKRZ and UKRI allow users to log on to powerful computational resources where they can install and run their own analysis tools and access the CMIP6 and CORDEX data archives directly. Access to these resources is restricted to teams that present proposals and get awarded resources. However, initial experience of offering this service has shown that there are barriers to its exploitation, due to the high effort that teams put to overcome technical difficulties in the use of these complex resources. At both DKRZ and UKRI, the analysis platforms are heavily exploited by the national science communities. Use by the national communities is backed by a range of training and communication activities which are not fully developed in the ENES CDI service. There may also be a reluctance on the part of users to commit to a service which has a potentially limited lifespan.

Despite the difficulties in launching this service, there will clearly be an increasing demand for this type of service as data movement costs increase and the cost of holding multiple copies of high-volume data becomes unsustainable.

The sustainability aspects of the ENES CDI compute services were integrated in the sustainability discussions and roadmap as part of WP2. An initial (~4 year) sustainability perspective on WPS service development and operations is provided by COPERNICUS related funding, yet a scalable production ready deployment, as an integral part of the ENES CDI, needs to be further discussed as part of the sustainability process already started.

3. Ongoing initiatives and related medium/long-term plans

3.1 ESGF

Starting in 2019, a major activity for ESGF has been the Future Architecture initiative, aiming to re-architect and re-engineer the software system. Following a meeting of the technical representatives of the institutions involved in the collaboration, a report[4] was compiled to set out a series of proposals and a roadmap of scheduled changes. Major changes include re-engineering the system to use container technologies to allow for a more modular scalable architecture that is easier to deploy and maintain, and in addition, the adoption of a more centralised architecture making use of cloud for hosting search and identity services. These two major areas of work have been divided into two respective phases. CEDA has led the initial phase, re-engineering to use containers, which is now largely complete. This new container-based system was adopted by US ESGF partner GFDL as part of the first operational ESGF node deployment on public cloud (on Amazon Web Services). This new containerised version has also been deployed at CEDA and the deployment will proceed with other European partners over the coming months.

The second phase of the work includes the development of new search and identity services. A test instance of ESGF new Identity Provider has been adopted as part of integration pilot with the Climate4Impact Portal. In addition, the Web Processing Services (WPS) developed within the scope of the activities for C3S are being used in this pilot. This is reported earlier in this document in Section 2. It is worth noting for ESGF that these services are being proposed as candidates in a wider piece of work to integrate compute services with ESGF partners outside Europe. A Compute Working Team for ESGF has been re-established and will co-ordinate collaboration activities towards the development of common compute capabilities for the federated infrastructure.

3.2 ENES Data Space and EGI-ACE

The scientific discovery process has been deeply influenced by the data deluge started at the beginning of this century. This has caused a profound transformation in several scientific domains which are now moving towards much more collaborative processes.

In the climate domain, the ENES Data Space aims to provide a novel paradigm towards an open, scalable, and cloud-enabled data science environment for climate data analysis. It represents a data and compute ecosystem deployed on top of the EGI federated cloud infrastructure, specifically designed to address analytics needs of the ENES community. The service, developed in the context of the EGI-ACE (EGI Advanced Computing for EOSC) project, provides ready-to-use compute resources and data collections, as well as a rich ecosystem of open-source Python modules and

[4] https://doi.org/10.5281/zenodo.3928222

community-based tools (e.g., CDO, Ophidia, Xarray, Cartopy, etc.), all made available through the user-friendly Jupyter interface to enable programmatic access and the development of data science applications.

In particular, the ENES Data Space provides access to a multi-terabyte set of specific variable-centric collections from large community experiments to support researchers in climate model data analysis. The data pool of the ENES Data Space consists of a mirrored subset of CMIP datasets from the ESGF federated data archive, collected by using the Synda community tool in order to provide the most up to date datasets into a single location. Results and output products as well as workflow definitions (in the form of Jupyter Notebooks) can be easily shared among users through data sharing services, which are also being integrated in the infrastructure, such as EGI DataHub.

The service was opened in the second part of 2021 and is now accessible in the European Open Science Cloud (EOSC) through the EOSC Portal Marketplace[5].

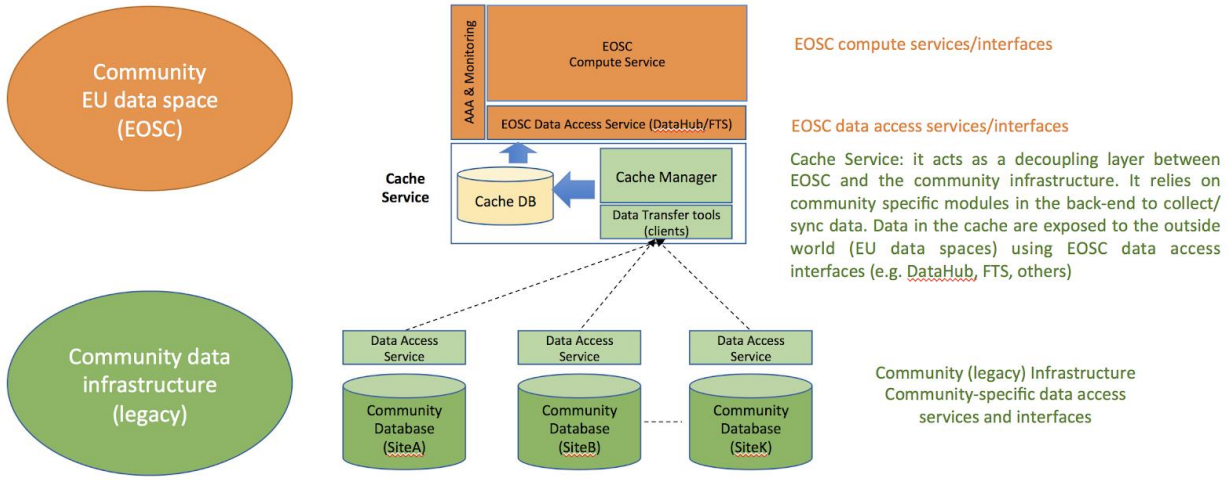

Figure 3: Architecture envisioned in the ENES Data Space for EOSC

More specifically, the ENES Data Space delivers a single entry point to an open and cloud-enabled data science environment for climate data analysis on top of the EOSC Compute Platform implemented in the project. It is built on top of the ENES Climate Analytics Service[6] (ECAS), which is one of the EOSC-Hub Thematic Services to deliver compute and analytics capabilities to the end users. Presently, 2TB CMIP6 data collections have been made available from the ESGF federated data archive, but the data pool will be further ramped up over the course of the project. The JupyterLab front-end provides data access, management, and visualization, and enables a wide

[5] https://marketplace.eosc-portal.eu/services/enes-data-space

[6] https://marketplace.eosc-portal.eu/services/enes-climate-analytics-service

range of data analysis experiments, such as trends, anomaly, climate change signal and extreme events analysis. Single and multi-model experiments are also supported via both interactive (exploratory) and batch data analysis to address different end-users' needs and requirements. Moreover, the ENES data space is intended to address both data-intensive and data-driven compute scenarios, thus covering a wide spectrum of analytics needs from the community. From an open (data) science perspective, FAIR principles will be pursued; in particular, openness and sharing of analytics applications (e.g., Jupyter Notebooks) will be fostered to increase re-use among users.

The expected contribution in the EGI-ACE project represents the first-phase implementation of a climate model data space. It leverages the compute service activity developed in IS-ENES to deliver a high-level and EOSC-enabled analytics environment for climate scientists. Future work will further evolve the environment in terms of robustness, scalability and richness of the software offering, through the integration of additional community tools and services as well as new data science Python packages.

3.3 Copernicus

The Copernicus Climate Data Store (CDS) interacts with remote data and compute services, including those that will provide regional climate projections. The first Copernicus Climate Change Service (C3S) procurements in the first phase of the Copernicus programme (Cop1) awarded several contracts towards:

1. the identification and quality-control of global and regional climate projections and predictions from the fifth and sixth phases of the Coupled Model Intercomparison Project (CMIP5/6) and the Coordinated Regional Climate Downscaling Experiment (CORDEX) led by the WCRP;
2. the creation of a dedicated infrastructure building on a set of custom Earth System Grid Federation (ESGF) nodes to publish and make information available through the cloud-based Climate Data Store (CDS), thus providing users with a single point of access to the quality-assured data;
3. providing basic compute services to reduce the amount of data relating to the global and regional climate projections served by the CDS.

A dedicated load-balanced and highly available infrastructure has been deployed for the CDS in the dedicated ESGF subsystem, as shown in figure 4.

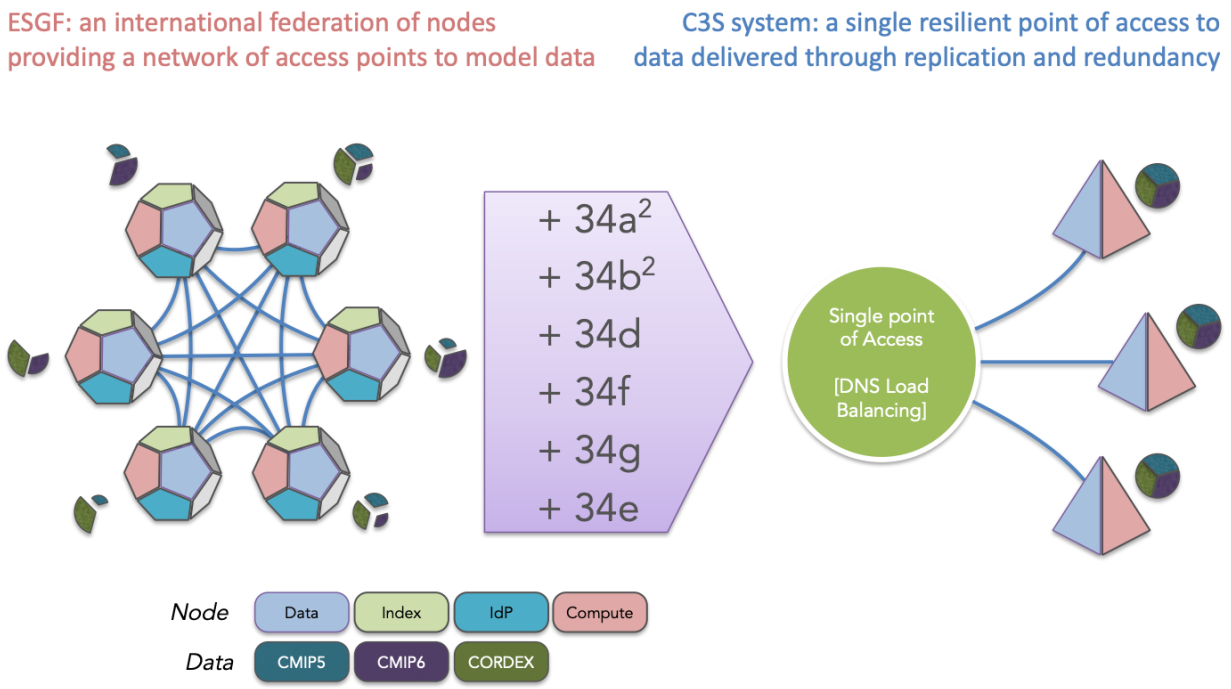

Figure 4: Dedicated load-balanced infrastructure deployed for the Climate Data Store from an ESGF-subsystem.

For the second C3S procurement, ENES partners responded to the C3S2_380 invitation to tender and proposed maintaining all infrastructure and data (global and regional climate projections and predictions) as supplied to C3S under the previous contracts. The objective of the present contract is to maintain the requisite infrastructure, based on an operational level agreement, with all the core services needed to continue the uninterrupted supply of global and regional climate projections data from previous projects to the CDS through the dedicated ESGF subsystem until the end of 2025.

The system was designed alongside the Climate Data Store itself to address the C3S willingness of a download service for climate projections data with key compute capabilities to ease server-side geographical operations. This infrastructure relies on three THREDDS instances, located in Paris (CNRS-IPSL), Didcot (STFC-CEDA) and Hamburg (DKRZ), aiming to ensure an optimal and operational service. A load-balancer is set on top of the three THREDDS instances, which allows automatic routing of the CDS requests for download based on each instance's workload. The highly available and scalable cloud Domain Name System (DNS) web service Route 53 from Amazon Web Services (AWS)[7] was selected among the existing solutions to manage load balancing. This

[7] https://aws.amazon.com/route53/

service benefits from its relatively easy set-up and configuration, along with low operational costs, and the ability to meet the required service uptime of the CDS.

In the meantime, STFC-CEDA and DKRZ have been at the forefront of delivering compute capabilities based on the OGC WPS standard specifically for C3S users. This WPS solution provides an alternative service for accessing data and is part of a new framework that lays the foundation for enhanced WPS processes that can support C3S requirements. The architecture of this new framework is called Roocs, as mentioned in Section 2.

Moreover, with regard to compute capabilities, CERFACS and KNMI are also active in the C3S_311_lot3 (C3Surf2), integrating the *icclim*[8] python package for climate indices and indicators calculation into the CDS Toolbox. This tool is developed in WP10 of IS-ENES3, while the implementation of the metadata standard related to climate indices outputs is using current work carried out on *clix-meta*[9], coordinated at the international level and developed in the context of WP3.

[8] https://github.com/cerfacs-globc/icclim

[9] https://github.com/clix-meta/clix-meta

4. Final considerations and future plans

The IS-ENES3 project has contributed to the enhancement of the ENES Climate Data Infrastructure to efficiently support decisions in the climate science domain through the exploitation of CMIP and CORDEX simulations.

As mentioned before in this document, several approaches and solutions have been developed by the different institutions participating in the IS-ENES3 project and involved in the compute service activities.

Currently, there is a clear trend towards some common aspects, with the aim of harmonising the processing services of the ENES CDI, but also preserving some peculiarities of each institutional service to better address different use cases and sectoral needs in the compute and analytics area.

The compute services should continue to be generic and interoperable as much as possible to overcome integration problems between different e-infrastructures and, at the same time, to address the needs of users from different domains and with various expertises. To this end, tailored interfaces will target specific users' groups, hiding the complexity of the compute back-end.

Due to the increasing impact of data deluge in recent years, a key aspect will be the ability of the analytics layer to scale up by accommodating larger data streams and computations, thus pursuing the near-data processing paradigm. Moreover, machine learning and artificial intelligence algorithms will help with the analysis of rapidly increasing volumes of Earth system data.

This layer should be flexible enough to be deployed on different infrastructures and this could be achieved by investing more in new containerization and container orchestration approaches. Additionally, cloud infrastructures should be exploited to provide more reliable and flexible services, also optimising the costs through efficient resource usages.

The compute layer will be strongly committed to open science principles, with the aim of democratising the analysis of climate data through the availability of tools and services, and hence foster collaboration in the climate domain.

This is in line with the European vision to empower citizens by allowing them to make better decisions based on insights coming from non-personal data. A clear focus is also on FAIR principles, which will enhance the possibilities for researchers to find, share and reuse publications, data, and software leading to innovations, higher research productivity and improved reproducibility in science. Therefore, the analytics layer will supply clear provenance information, along with transparency about initial conditions and dependencies, reproducibility in processing

and discovery services, and it will also guarantee persistence of both data and services involved in data exploitation (data and metadata standards, documentation, persistent identifier and so on).

Finally, since the ENES CDI is strongly connected to other e-infrastructures and European initiatives, future compute services will also have to take the emerging needs of climate and impact communities into account, by means of emerging technologies, besides leveraging ongoing and future European data infrastructure efforts (EOSC, Copernicus, GAIA-X, Digital Twin etc.).