Towards an Open (Data) Science Analytics-Hub for Reproducible Multi-Model Climate Analysis at Scale

Sandro Fiore*, Donatello Elia*†, Cosimo Palazzo*, Alessandro D'Anca*,
Fabrizio Antonio*, Dean N. Williams‡, Ian Foster§ and Giovanni Aloisio*†
*Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Lecce, Italy
†University of Salento, Lecce, Italy
‡Lawrence Livermore National Laboratory, Livermore, USA
§University of Chicago & Argonne National Laboratory, Chicago, USA

***Abstract*—Open Science is key to future scientific research and promotes a deep transformation in the whole scientific research process encouraging the adoption of transparent and collaborative scientific approaches aimed at knowledge sharing. Open Science is increasingly gaining attention in the current and future research agenda worldwide. To effectively address Open Science goals, besides Open Access to results and data, it is also paramount to provide tools or environments to support the whole research process, in particular the design, execution and sharing of transparent and reproducible experiments, including data provenance (or lineage) tracking. This work introduces the *Climate Analytics-Hub*, a new component on top of the Earth System Grid Federation (ESGF), which joins big data approaches and parallel computing paradigms to provide an Open Science environment for reproducible multi-model climate change data analytics experiments at scale. An operational implementation has been set up at the SuperComputing Centre of the Euro-Mediterranean Center on Climate Change, with the main goal of becoming a reference Open Science hub in the climate community regarding the multi-model analysis based on the Coupled Model Intercomparison Project (CMIP).**

***Index Terms*—Open Science, provenance, analytics-hub, reproducibility, data analytics**

I. Introduction

Open science is becoming increasingly crucial for scientific research and can have a significant impact on the whole research cycle. It leverages new ways to perform research and share the results through open digital technologies and collaborative tools [1]. There is no clear definition of Open Science; it can actually be considered as an umbrella term covering a broad range of aspects related to scientific knowledge sharing and research collaboration, embracing other terms such as Open Access, Open Data, Open Source software and Open reproducible research [2] [3]. In [4] review, the following definition of Open Science is proposed: *"Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks"*. From this review it emerges, hence, that the key aspect of Open Science is transparent, accessible, shared and collaborative developed knowledge. Transparency, openness and reproducibility are also mentioned as key factors for an Open (Science) research culture [5].

In the European landscape, Open Science is considered strategic for future research programmes. In 2015, the EU Commission actually set Open Science, Open Innovation and Open to the world as three main goals for future research and innovation in the EU [6]. From this perspective, *research*, *data*, and *dissemination* represent three key dimensions for Open Science in Europe. Several initiatives and projects have therefore been funded by the EU commission to promote open science and innovation. A very important initiative in this direction is OpenAIRE (Open Access Infrastructure for Research in Europe), that has been supported since 2006 by a series of EU projects to ease the adoption of Open Access in Europe, by providing open access to the research outputs funded by the EU [7]. Another example is the FOSTER portal, which has been supported by the FP7 FOSTER (Facilitate Open Science Training for European Research) and H2020 FOSTER Plus (Fostering the practical implementation of Open Science in Horizon 2020 and beyond) EU projects and provides training resources to aid researchers and other stakeholders in the development of Open Science practices [8].

Currently, one of the most important initiatives carried out by the EU is the European Open Science Cloud (EOSC), which *"aims to create a trusted environment for hosting and processing research data to support EU science in its global leading role"* [9].

One of the key aspects in Open Science is the FAIR Reproducibility principle [10] [11]. Several efforts have been made towards addressing computational reproducibility, as seen in literature [12] [13] [14].

This work introduces the *Climate Analytics-Hub*, a new component built on top of the Earth System Grid Federation (ESGF), which joins big data approaches and parallel computing paradigms with the aim of providing an Open Science-ready environment for reproducible multi-model climate analytics experiments at scale based on the Coupled Model Intercomparison Project (CMIP).

The rest of this paper is organized as follows: Section II

describes multi-model climate data analytics, along with the key concepts and main challenges, in the context of the CMIP experiments and the ESGF federation, whereas Section III introduces the architecture of the Climate Analytics-Hub together with the main requirements it addresses. Section IV describes the internal design of the Climate Analytics-Hub, its infrastructural view and implementation details as well as Open Science aspects related to analytics workflows and applications, such as, in particular, reproducibility. Then, Section V describes the implementation of multi-model climate data analysis, emphasizing the analytics workflow runtime execution and the available provenance support. Finally, Section VI draws the main conclusions and hints at future work.

II. Multi-model climate data analytics in the CMIP context

This section describes multi-model climate data analytics in the CMIP context, introducing the CMIP experiment and the ESGF infrastructure, as well as presenting the key concepts, main challenges and issues of these analyses.

A. The CMIP experiments and Earth System Grid Federation

The increased models resolution in the development of comprehensive Earth System Models is rapidly leading to a very large climate simulations output that poses significant scientific data management challenges in terms of data sharing, processing, analysis, visualization, preservation, curation, and archiving [15] [16] [17].

In this domain, large-scale global experiments for climate model intercomparison (CMIP* [18]) have led to the development of the Earth System Grid Federation (ESGF [19]). It is a federated data infrastructure that involves a large set of data providers/modelling centres around the globe and includes the European contribution through the IS-ENES project (by the European Network for Earth System Modelling (ENES) community). The Coupled Model Intercomparison Project (CMIP) has been established by the Working Group on Coupled Modelling [20] (WGCM) under the World Climate Research Programme (WCRP).

From an infrastructural standpoint, ESGF provides production-level support for search & discovery, browsing and access to climate simulation data and observational data products. It should be noted that:

- ESGF has been serving the Coupled Model Intercomparison Project Phase 5 (CMIP5) experiment, providing access to about 2PB of data produced around the globe by 26 institutes (groups) and 60 models.
- ESGF is supporting the CMIP6 experiments, which are expected to publish around 20PB of data (a 10X factor with respect to CMIP5).

It is also important to point out that, today, ESGF primarily provides a large-scale, federated data sharing infrastructure. Nevertheless, several efforts are currently being made to include analytics and computing capabilities in production as future plan for 2019 onward. In such a context, CMIP-based *multi-model analyses* are clearly one of the most relevant exercises that can be run by scientists on top of the ESGF data archive.

B. Multi-model climate data analysis: key concepts

Multi-model data analysis requires access to data produced by large-scale inter-comparison experiments (e.g. CMIP) and made available through the ESGF federated data archive, as well as running *workflows* with tens/hundreds of data analytics operators. Examples of multi-model analysis are, among others: anomaly analysis, trend analysis and climate change signal analysis.

In the context of the H2020 INDIGO-Datacloud project [21], the Precipitation Trend Analysis (PTA) was selected as a pilot case [22] [23] since it is scientifically relevant and also general enough to validate the infrastructural aspects that also apply to other classes of data analysis (e.g. outlier analysis). Fig. 1 shows the workflow designed for the PTA in the CMIP5 context.

The proposed analysis consists of two main stages:

- the first part includes a number of identical sub-workflows, each associated with a specific climate model involved in the CMIP experiment and independent of the others; a future climate scenario must also be defined as input for this step;
- the second part considers a final workflow to perform statistical analysis on the set of output provided at the end of each sub-workflow at the first stage.

In Fig. 1, the sub-workflows are shown within cyan rectangles. The tasks related to historical data process are in green rectangles, whereas the tasks that process data resulting from the model are in red rectangles. It should be noted that the time domain related to historical data is fixed; for instance, the 1976-2005 range is adopted for the experiment. The time domain related to models shall have the same duration (e.g. 30 years) though it clearly refers to a future time range, like 2071-2100.

Each sub-workflow performs the following tasks in the first phase of the experiment: (i) discovery of the two input datasets (historical and future scenario data), (ii) spatio/temporal subsetting based on the user's input, (iii) evaluation of the precipitation trend for both datasets, (iv) trends comparison over the considered domain, and (v) 2D map generation (output).

In the second phase of the experiment, the multi-model statistical analysis includes the following four steps: (i) data gathering from the first phase (NetCDF files [24]), (ii) data regridding, (iii) statistical analysis, and (iv) final 2D maps related to the inferred statistical indicators. The final data or maps can then be published or shared with the whole experiment flow definition.

C. Multi-model climate analysis: challenges and issues

To fully understand some key challenges and very practical issues related to *multi-model climate analysis*, it is important to analyse the entire user's scientific workflow behind it. To perform multi-model climate analysis, the end-users must:

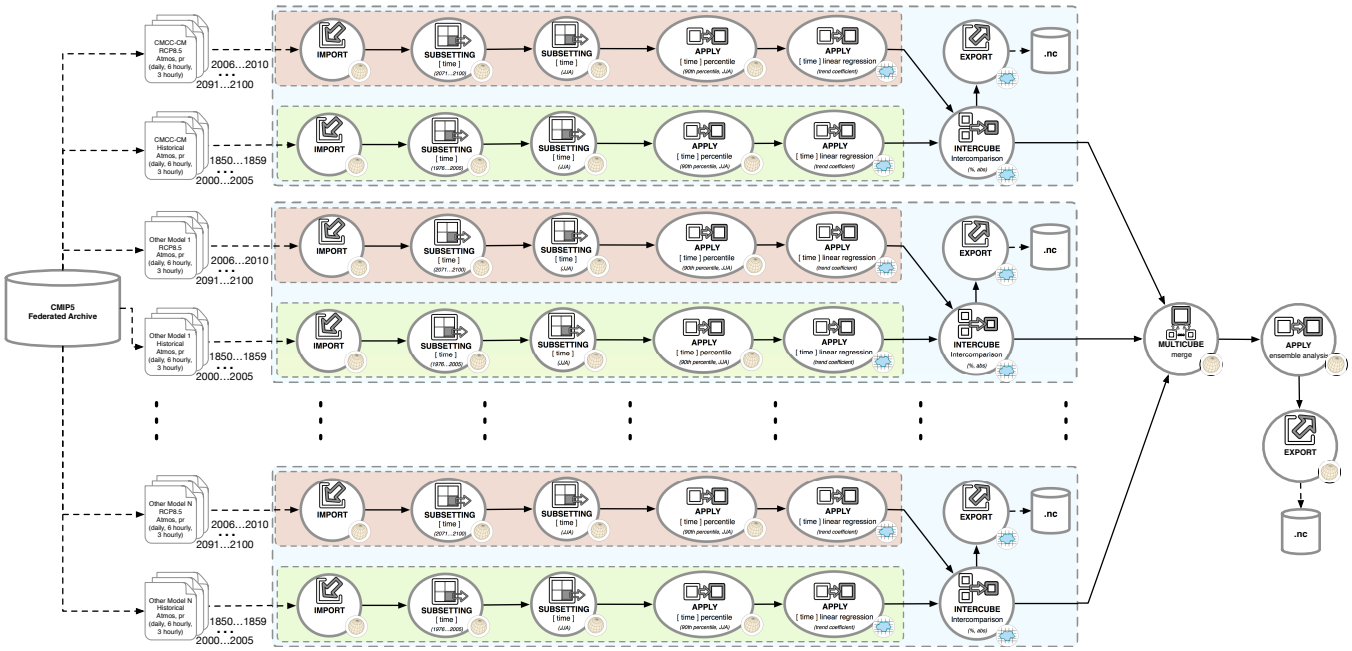

Fig. 1. Definition of the Precipitation Trend Analysis workflows.

1) *download* all the needed input datasets from the distributed ESGF data nodes to their local machines (local could mean the scientist's workstation or the user account on a HPC facility). Such a preparatory step represents a strong barrier for climate scientists, as the data download can take a significant amount of time (depending on the amount of data required by the analysis). Moreover, downloads can suffer from network instability, dropped connections, etc. which make the entire process even more painful.
2) *prepare* a set of batch scripts that can properly process all the collected data. To this end, analyzing large datasets involves running multiple data operators, from a set of domain-oriented command line interface (CLI) tools (mostly *sequential*). This is usually done via scripts on the client side and requires climate scientists to take care of, implement and replicate workflow-like control logic aspects in their scripts, along with the expected application-level part. At this level, re-usability of scripts has never (or very poorly) been addressed.
3) *install and update* all the required data analysis tools/libraries on their local machines. To this end, the proper setup of the ICT environment (which requires system management and technical skills) is key to run the analysis, as the user generally leverages a wide set of tools and the compatibility at ecosystem level (e.g. libraries), mainly related to software versions, can raise several issues.
4) *run* the analysis taking into account the available computational and storage resources. This could lead to user-specific solutions about how to split the analysis, exploit parallelism, use the available resources, etc. In this regard, the large volume of data and the strong I/O requirements pose additional challenges related to performance as well as data handling.

In such a context, the *reproducibility* of the multi-model analyses has never been fully addressed from an Open Science perspective. Indeed, it can be easily argued that the client-side nature of the workflow is a major barrier towards the implementation of an *Open Science driven climate analytics environment*. The next section provides a detailed description of the approach inspired by Open Science principles (e.g. reproducibility) and useful to address the mentioned challenges and issues.

III. Climate Analytics-Hub: architectural view and key requirements

This section presents the architectural view of the Climate Analytics-Hub in the large as well as its role with respect to the legacy ESGF infrastructure, as well as its key requirements to address the multi-model analytics challenges described in the previous section.

A. Architectural view in the large

The proposed architecture (Fig. 2) implements a *Climate Analytics-Hub* (hereafter *Analytics-Hub*) level on top of the existing ESGF data nodes backbone to allow the execution of multi-model climate analyses on a single location. The Analytics-Hub is responsible for providing Open Science oriented computing and analytics capabilities on top of a data collection layer which both (i) pre-stages and caches the data relevant to the analyses from the different ESGF data nodes and (ii) keeps the local copy of data synchronised with the remote copy available in the ESGF infrastructure.

Of course, a centralized storage location, like in the Analytics-Hub, cannot represent a scalable solution for the whole CMIP data archive (approximately 20PB expected for CMIP6), but it can be considered as a suitable approach for the analysis of one or more selected variables (depending on storage availability). As a consequence, multiple, distributed Analytics-Hubs could serve the entire community by addressing the full spectrum of variables. Such scenario provides a centralised, *variable-centric* and Analytics-Hub-based infrastructural paradigm for multi-model climate analysis, on top of the distributed, *model-centric* and data nodes-based paradigm available through the ESGF infrastructure, mostly serving data access needs.

In previous work [22] [23], a distributed solution based on a two-level workflow approach was proposed. That was the first step towards the Analytics-Hub concept, which was not mature enough at that time. The design was mainly driven by the data distribution requirement inherently coming from the legacy of the ESGF infrastructure as well as by the need to avoid large-scale data movement simply through the adoption of server-side analytics solutions. While the solution proved to be effective with regard to the time-to-solution dimension of the multi-model climate analysis, it was noted that it could not be the proper solution in production environments, since they suffer from network instability, sites unavailability, services downtime, and non-uniform service release deployment across sites. Such elements were key to move towards a more centralised, single-level workflow, Analytics-Hub concept.

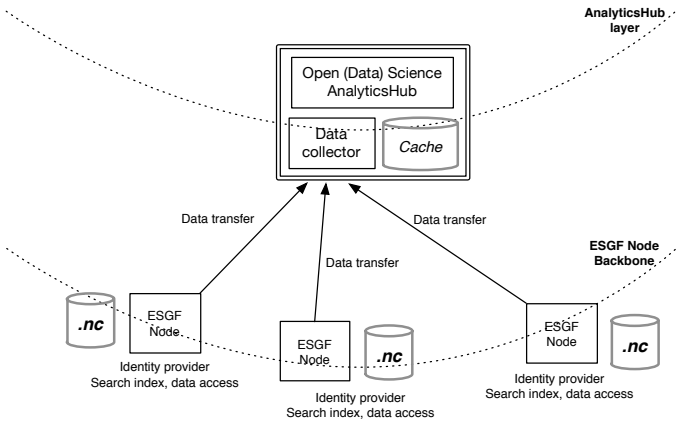

Fig. 2. Analytics-Hub architecture in the large.

B. Analytics-Hub requirements

To tackle and address the large-scale multi-model climate data analysis challenges and issues described in Section II-C, the envisioned Analytics-Hub component has to fulfil some key requirements, such as: server-side analytics, parallel/big data approaches, workflow analytics support, data consistency, metadata management, provenance and reproducibility, social and cultural implications, and, finally, Open (Data) Science-ready environments.

a) Server-side analytics: As described in Section II-C, the workflow for multi-model climate analysis is still based on a server-side data management (data access) and client-side (desktop-based) data analysis. This workflow is not feasible at large-scale, since the ever-larger scientific datasets that are going to be produced by experiments/simulations (e.g. CMIP6):

- (i) make data download no longer a viable option for users to collect all the data;
- (ii) cannot be properly handled with the available client-side data management tools due to the critical *volume* dimension of the analysis.

Using a server-side paradigm, data (input, output, and intermediate products), provenance and even sessions can be managed on the remote side and only the final results of the analysis (typically megabytes or even kilobytes) can be downloaded by the end-users. Such an approach reduces (i) the downloaded data, (ii) the makespan for the analysis task, and (iii) the complexity related to the analysis software to be installed on the end-users machines, thus fully addressing several issues mentioned in Section II-C. Additionally, the server-side paradigm can straightforwardly enable Open Science principles, leading, for instance, to a better re-use of data (e.g. intermediate/final products), improved analyses (e.g. server-side jobs) and user's sessions, etc. Still, the provenance management can represent the proper foundation to fully support reproducibility. Finally, storing all the information on the server-side, knowledge-driven features (e.g. based on data mining algorithms) can be added to the analytics system with the aim of suggesting, recommending and predicting.

b) Big data and HPC-based analytics: Big data and HPC approaches (e.g. High Performance Data Analytics - HPDA) can represent the proper answer to deal with the big data nature of the multi-model analysis. Presently, the big data and HPC convergence is an open, challenging and vibrant research topic under discussion by the HPC scientific community (e.g. Big Data and Extreme-scale Computing initiative [25]). With respect to the user's workflow described in Section II-C, HPDA frameworks allow the implementation of a new approach, based on a server-side analysis paradigm and data-intensive facilities close to the data storage. Performance is a key challenge addressed by HPDA solutions.

c) Data consistency: Data consistency arises when a data replication scenario comes into play. The Analytics-Hub downloads the data relevant to the multi-model climate analysis from ESGF and caches it into a local storage. However, since new versions of a dataset can be published into the ESGF data archive, it is of paramount importance that the Analytics-Hub cache should not get into an inconsistency status. To address that, the local cache must reflect the new status of the ESGF federated repository, by downloading new datasets versions as soon as they are published and made available in ESGF.

d) Workflow-enabled analytics: To manage large-scale multi-model climate analysis, end-users need to deal with tens/hundreds of analytics operators. Workflow support is then

key to both (i) mapping a climate analysis onto a Direct Acyclic Graph and (ii) properly managing its run time execution (dependencies, failures, etc.). From an Open Science perspective, FAIR principles [10] can be applied to workflows; indeed, workflow documents can be shared among scientists (re-usability), described using standards/recommendations (interoperability), as well as published on well-structured (findability) and public (openness and accessibility) repositories (e.g. GitHub, MyExperiment [26]).

e) Metadata management, provenance and reproducibility : Metadata is a key point for scientific data management systems in general and server-side analytics systems in particular, due to the potential scale of data, experiments and users they target. Metadata can be scientific dataset attributes, provenance information, storage mapping information, persistent identifiers (e.g. DOIs), etc. Besides the well-known data discovery, metadata is also key to addressing analysis experiments *reproducibility*, thus strongly contributing to the adoption of Open Science principles.

f) Social/cultural implications: The proposed Analytics-Hub approach can help develop new community-oriented tools towards much more open, multi-level and collaborative scientific forms/approaches. From a social perspective, scientists should actually move from isolated ways of conducting their research towards new and more collaborative approaches/environments for multi-model climate analysis, to differently cope with the way scientists interact with each other both inside (for research purposes) and outside (for dissemination and scholarly communication purposes) the scientific community. In this respect, the Analytics-Hub aims to support a social and cultural shift moving from a *single-user* to a (distributed) *team-driven* analysis approach, where multiple users can share thoughts, exchange ideas and collaborate on the same analysis experiment by working on several aspects and branches of the full analysis workflow.

g) Open (Data) Science-ready environment: Open (Data) Science requires systems capable of fostering collaboration through scientists and sharing research results. In such a context, Jupyter Notebooks represents a very valuable and easy-to-use tool to share and replicate the code and results of scientific experiments, jointly with the explanatory comments in human-readable form [27].

IV. Analytics-Hub: architectural design, infrastructural view and implementation details

This section provides a detailed description of the Analytics-Hub by presenting its internal architectural design, the infrastructural view, some implementation details as well as Open Science aspects related to analytics workflows and applications such as, in particular, reproducibility.

A. Architectural design

Based on the above listed requirements, the internal design of the Analytics-Hub consists of several components:

- (i) an interface/GUI providing an Open (data) Science-ready environment where scientists can run their own Data Science applications, perform interactive and exploratory data analysis, run analytics workflows, perform data visualization, manage collaborative sessions, share analysis experiments, etc.;
- (ii) a workflow-enabled, secure, and interoperable Analytics-Hub front-end able to address the user's requests both in terms of single tasks and workflows;
- (iii) an analytics framework back-end able to perform data analysis at scale and support metadata management at different levels: datasets (e.g. data attributes), infrastructure (e.g. data partitioning & mapping onto the storage system, computational and data resources, software ecosystem), and processing (e.g. provenance, logging and bookkeeping);
- (iv) a data collector and its local storage to gather the relevant datasets from ESGF and keep them in sync with the remote repositories.

As shown in Fig. 3, the proposed Open (Data) Science environment has relevant social implications; actually, it also includes (i) *publication services* to enable shareability and reusability of results (open data) across the community, and (ii) *open development platforms* to host, review, manage and share code (e.g. workflows, applications) by means of an open and community-oriented approach.

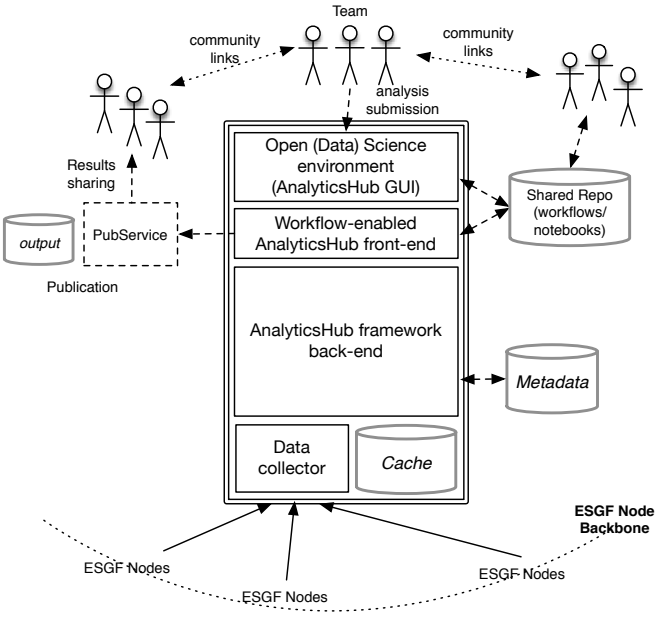

Fig. 3. Analytics-Hub architecture in the small.

B. Infrastructural view

From an infrastructural standpoint, the proposed Analytics-Hub integrates several open source software solutions. More specifically (i) the Open (Data) Science environment is implemented through JupyterHub [28]; (ii) the Analytics-Hub front-end and back-end are based on the Ophidia HPDA framework [29] [30]; and (iii) the data collector is based on Synda [31], which allows the download of datasets and the (one-way)

synchronization of local data repositories with data hosted on the ESGF data infrastructure.

In the proposed ecosystem, the deployed *publication services* are mainly OPeNDAP/THREDDS and Apache HTTP services, providing open access (e.g. based on creative commons licenses). The *open development platform* selected to share workflows and the applications code in the system is GitHub; by its nature, it tracks workflow evolution provenance.

C. Implementation details

The Ophidia HPDA framework is the main component of the Analytics-Hub. It is a complete open source solution [32], released as open source software (under GPLv3 license) and used to perform scientific data analytics by means of HPC paradigms and in-memory based big data approaches [33]. The platform has been successfully used in several scientific experiments (e.g. climate change and astronomy) as well as smart cities applications [34]. It supports access, management, analysis and mining of n-dimensional array-based data structures, leveraging the datacube abstraction. Relevant to this paper is the collaborative session support provided by the Ophidia server front-end, which enables a team-oriented analytics session management. Sessions are server-side managed and they can be paused/resumed; they also support group-based authorization to manage multiple roles in a team of scientists participating in the same experiment.

The Ophidia workflow management system [35] is a core component of the Ophidia platform. It allows coordinating and orchestrating the execution of scientific experiments composed of multiple data analytics, processing and visualization operators (e.g. operational processing/analysis chains). In terms of execution, the Ophidia HPDA framework supports different types of tasks: (i) single tasks of one operator; (ii) HTC tasks (parameter sweep tasks), where a single operator is executed multiple times on different input, according to user-defined filters; and (iii) complex workflows (DAG) composed of multiple, single or HTC tasks, jointly with flow control and management tasks (i.e., iterations, conditionals). To simplify the interface, all the three different types of tasks are actually managed as workflows; they are coded in JavaScript Object Notation (JSON) in compliance with a request schema [36]. The schema specifies how to describe tasks and dependencies (both data and flow dependencies), input and output data, metadata information and flow management operations; it is used to validate workflow instances. Analysis experiments can be designed according to this schema and can easily be shared with other users (e.g. through GitHub), fostering experiment reuse and inherently providing a means of experiment reproducibility. In fact, given the JSON workflow and the input data, it is possible to rerun the experiment through Ophidia and reproduce the experiment outcome. Additionally, the JSON schema allows creating easy-to-process, interoperable, machine-readable documents.

Besides the workflow management support, Ophidia also provides the Python bindings, called PyOphidia, which allow a programmable integration of Ophidia operators and workflows into more articulated and shareable Data Science applications. Hence, PyOphidia can be used together with other Python modules for the creation, execution and sharing of end-to-end data analytics workflows within Python-based Jupyter Notebooks.

D. Analytics-Hub workflows & applications reproducibility

From an Open Science principles perspective, the *Ophidia workflow document* enables *workflow replicability*. Moreover, due to the open source nature of the framework, *Ophidia workflows* are also *extendable* and *modifiable*. Still, the more detailed *Ophidia analytics document* enables *analysis reproducibility*. It extends the Ophidia workflow document (whose specific version is uniquely identified by its associated commit in GitHub) with additional information on (i) the computing environment (e.g. platform, compilers, libraries, etc.), (ii) the analytics ecosystem (e.g. Ophidia release, NetCDF library version, Python modules and related software dependencies, etc.) and (iii) the input data (e.g. through DOIs). Indeed, the first two points mentioned above capture system-level provenance information, which is key to enabling portability, as a pre-condition for reproducibility. Reproducibility in turn, fosters and addresses *re-usability*, one of the FAIR guiding data principles.

The information needed to reproduce an experiment can be obtained from its *provenance*, that is the description of the different stages data has undergone during the analysis process, from its origin to the final outcome. As mentioned in Section III-B, (tracking) provenance is a strong requirement for the Analytics-Hub. Besides the static *prospective provenance* tracked by the workflow document, Ophidia also supports the more dynamic *retrospective provenance*, which means it tracks at run time the provenance of each datacube imported or produced within the framework. In this respect, each new datacube is linked to the set of input datacubes (the multi-dimensional datasets) it has been generated from, together with the applied operator; to identify a datacube, a unique persistent identifier (PID) is automatically generated by the framework and attached to it.

However, as the information about the compute environment and the analytics ecosystem is not captured by the Analytics-Hub yet, the *reproducibility* can only be addressed through the more complete *Ophidia analytics document*, which may require human intervention/input to fully describe any missing provenance information (e.g. computing ecosystem and platform-level information). This shows the multifaceted nature of provenance and the existence of several classes of *provenance information* that must be taken into account. Such variety of information enables spotting issues that may not only be related to the application itself, but also to the surrounding software ecosystem [37]. Whereas the complete Ophidia analytics document is aimed at enabling reproducibility, its machine-readability (JSON format) represents a pre-condition for the *reproducible executability* of the analysis. Such concept is well-connected, from a technological perspective, to virtualized/cloud environments and automated deploy-

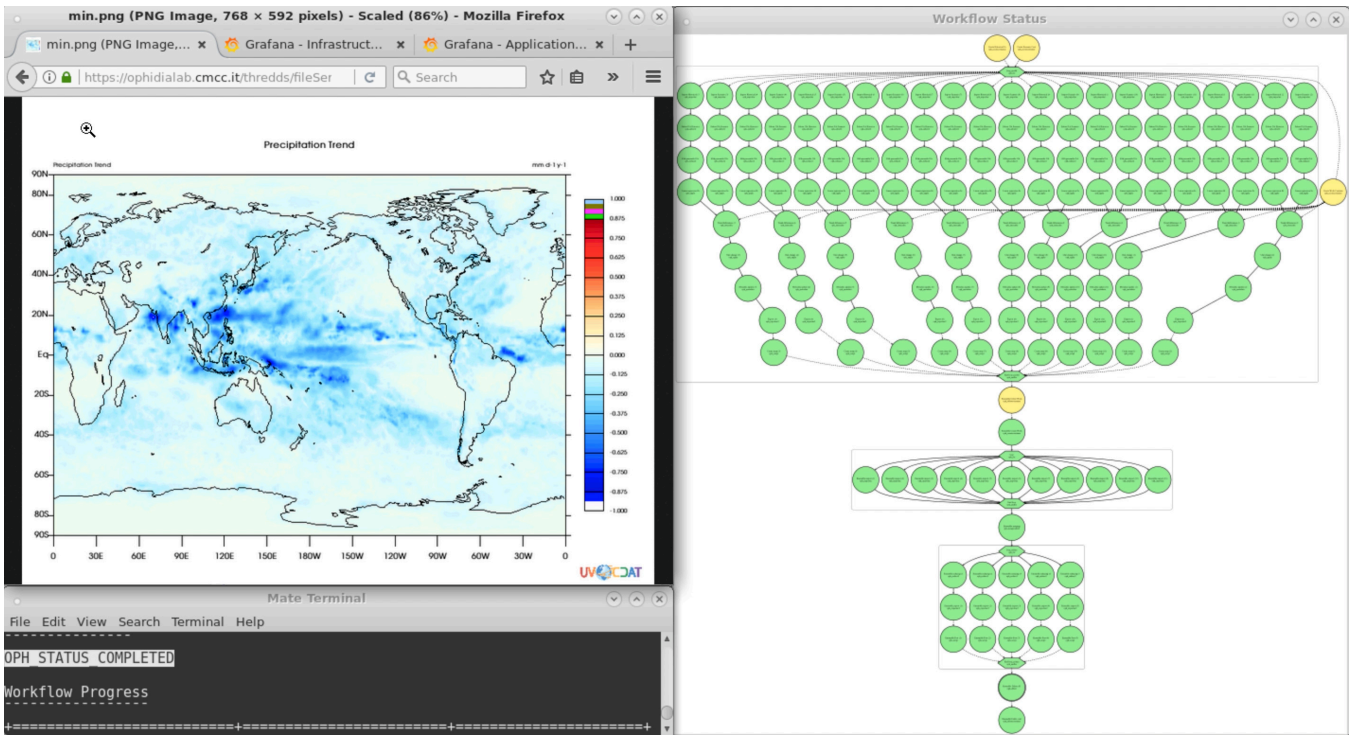

Fig. 4. A snapshot of single-site multi-model analysis workflow (runtime execution and output of the workflow experiment).

ment. Due to page-limit constraints, this topic will be further discussed in a future work. Moreover, the analytics document is stored and managed in versioned repositories (i.e. GitHub) and its evolution is easily tracked through GitHub commits, thus enabling the *analytics document evolution provenance*, which is beyond reproducibility and leads directly to the *citability* of the analysis for scientific publications.

It is worth mentioning that the retrospective provenance support implemented in the Analytics-Hub also applies to Python data analysis *applications*. Similarly to the workflow approach, the provided provenance system is able to track the complete analytics operators' flow throughout the full application execution, by storing all the relevant information in the provenance database (provDB). From a technical standpoint, the provDB is a knowledge graph for analytics experiments, implemented as a graph database running on top of Neo4j graphDB engine. Presently, the provDB is mainly explored for retrospective provenance through the native, available query support. More specifically, the current support allows end-users to explore, navigate, reason, make inference, and, if needed, manually change the workflow document or the application code. Future work on this specific topic concerns the development of graph mining algorithms with the ultimate goal of addressing *AI-enabled reproducibility* scenarios.

V. Multi-model climate analysis implementation

A real implementation of the PTA test case described in Section II-B has been implemented as demonstrator on top of the proposed *Analytics-Hub* as an Ophidia workflow; it has been tested on the Analytics-Hub infrastructure set up at the CMCC SuperComputing Centre, which aims to become a reference Open Science hub in the climate community regarding the CMIP-based multi-model analysis applied to some key variables (e.g. precipitation).

The Analytics-Hub paradigm creates new, refined and open variable-centric data stores, it eases and democratizes the analysis process overcoming key barriers related to data download & preparation, and promotes Open Science principles; in particular, re-usability, openness and sharing of data, workflows and source code, fostering new opportunities for open research and collaborations.

The PTA multi-model workflow [38] has been executed on 11 models from the CMIP5 experiment (for a total of 181 tasks, as can be seen in Fig. 4, which shows the runtime of the experiment). Additionally, the PTA workflow has been generalized to support the implementation of a variety of indicators in a multi-model fashion, thus addressing multi-model climate data analyses more in general. The solution includes a multi-model framework, where single-model indicator workflows can be plugged in the overall workflow, as a black-box, through a specific API.

From a *provenance* standpoint, the *oph_cubeio* operator can be used in the Ophidia framework to retrieve the whole data lineage related to a particular PID (i.e. associated with a datacube) from the provDB. Fig. 5 shows a graphical representation of provenance (created from the Ophidia CLI) for a datacube produced by the PTA workflow. In particular, it refers to one of the identical single-model blocks executed

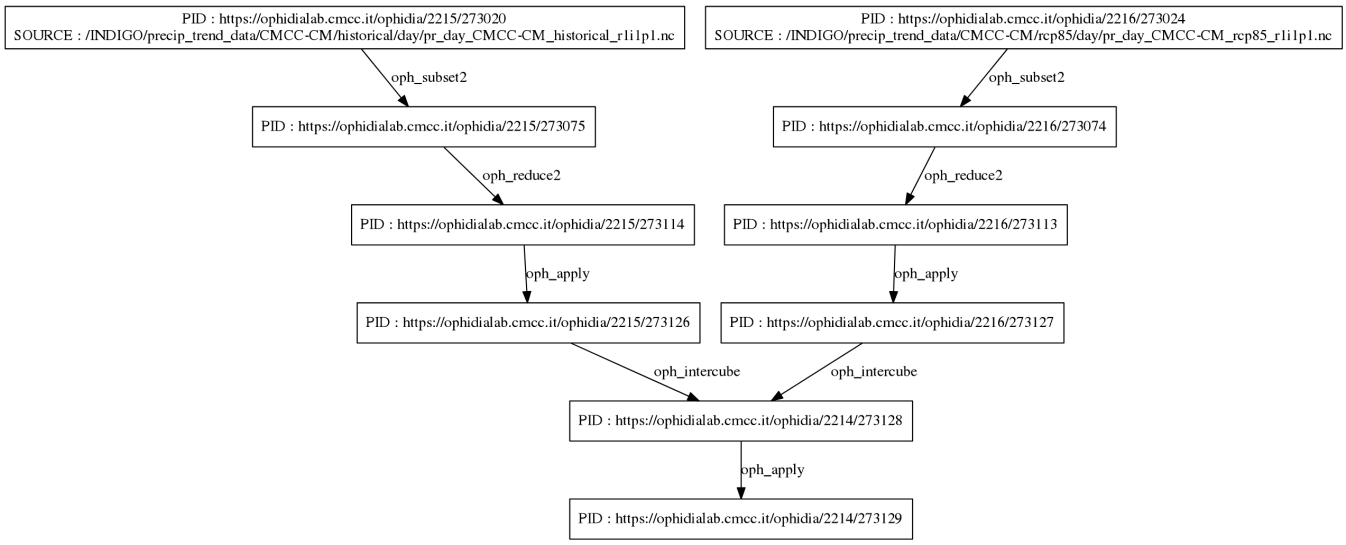

Fig. 5. Ophidia data provenance diagram related to the first stage (single-model block) of the PTA workflow.

during the first part of the workflow on each of the 11 models. The first two nodes are related to the import of the input dataset of the model, whereas the node at the bottom represents the last datacube produced during this stage; the edges among the nodes are labelled with the operator executed to run the analytics.

From a reproducibility point of view, the *Ophidia analytics document* of the performed PTA includes information about the Analytics-Hub platform running at the CMCC SuperComputing Centre, the Analytics-Hub software stack, the Ophidia workflow document related to the PTA analysis, as well as all the involved input data from the CMIP5 federated data archive.

VI. Conclusions and future work

This paper presents the *Climate Analytics-Hub*, a new component built on top of the Earth System Grid Federation, which joins big data approaches and parallel computing paradigms to provide an Open Science environment for reproducible multi-model climate change data analytics experiments at scale. The paper highlights the rationale behind the *Analytics-Hub* as well as its role on top of ESGF. Additionally, it delves into architectural aspects and infrastructural details to provide an in-depth view of this component. The adoption of Open Science principles (in particular reproducibility, but also openness and reusability) with respect to the Analytics-Hub workflows and applications is also extensively presented. A real multi-model analytics use case related to the study of the precipitation trend analysis in the CMIP5 is also thoroughly discussed in terms of experiment design, implementation details and provenance aspects.

Future work is mainly aimed at Open Science principles and in particular *AI-enabled reproducibility* with the support of graph mining applied to the provDB, to enable proactive knowledge-based approaches (e.g. recommendation systems) for advanced data provenance exploitation scenarios. Moreover, a larger-scale Analytics-Hub setup is planned at the CMCC SuperComputing Centre, to support reproducible multi-model analytics experiments in the CMIP6 context.

Acknowledgments

This work was partially supported by the Italian Ministry of Education, Universities and Research (MIUR) under the GEMINA contract, and partially by the EU H2020 Excellence in SImulation of Weather and Climate in Europe (ESiWACE) project (Grant Agreement 675191) and the EU H2020 Integrating and managing services for the European Open Science Cloud (EOSC-Hub) Project (Grant Agreement 777536). Moreover, the authors would like to acknowledge Antonio Aloisio for his editing and proofreading work on this paper.

References

[1] *Open innovation, open science, open to the world - A Vision for Europe*. Directorate-General for Research and Innovation (European Commission), 05 2016. [Online]. Available: https://publications.europa.eu/en/publication-detail/-/publication/3213b335-1cbc-11e6-ba9a-01aa75ed71a1/language-en

[2] B. Fecher and S. Friesike, *Open Science: One Term, Five Schools of Thought*. Cham: Springer International Publishing, 2014, pp. 17–47. [Online]. Available: https://doi.org/10.1007/978-3-319-00026-8_2

[3] N. Pontika, P. Knoth, M. Cancellieri, and S. Pearce, "Fostering open science to research using a taxonomy and an elearning portal," in *iKnow: 15th International Conference on Knowledge Technologies and Data Driven Business*, 2015. [Online]. Available: http://oro.open.ac.uk/44719/

[4] R. Vicente-Saez and C. Martinez-Fuentes, "Open science now: A systematic literature review for an integrated definition," *Journal of Business Research*, vol. 88, pp. 428 – 436, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0148296317305441

[5] B. A. Nosek, G. Alter *et al.*, "Promoting an open research culture," *Science*, vol. 348, no. 6242, pp. 1422–1425, 2015. [Online]. Available: http://science.sciencemag.org/content/348/6242/1422

[6] Goals of research and innovation policy. European Commission. [Online]. Available: https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy_en

[7] About OpenAIRE, OpenAIRE portal. [Online]. Available: https://www.openaire.eu/mission-and-vision

[8] FOSTER portal. [Online]. Available: https://www.fosteropenscience.eu/about

[9] European Open Science Cloud (EOSC). European Commission. [Online]. Available: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[10] M. D. Wilkinson, M. Dumontier *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, 2016.

[11] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson, "Cloudy, increasingly fair; revisiting the fair data guiding principles for the european open science cloud," *Information Services & Use*, vol. 37, no. 1, pp. 49–56, 2017.

[12] J. Freire, P. Bonnet, and D. Shasha, "Computational reproducibility: State-of-the-art, challenges, and database research opportunities," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '12. New York, NY, USA: ACM, 2012, pp. 593–596. [Online]. Available: http://doi.acm.org/10.1145/2213836.2213908

[13] Z. Yuan, D. H. T. That, S. Kothari, G. Fils, and T. Malik, "Utilizing provenance in reusable research objects," *Informatics*, vol. 5, no. 1, p. 14, 2018. [Online]. Available: https://doi.org/10.3390/informatics5010014

[14] F. S. Chirigati, D. E. Shasha, and J. Freire, "Reprozip: Using provenance to support computational reproducibility," in *5th Workshop on the Theory and Practice of Provenance, TaPP'13, Lombard, IL, USA, April 2-3, 2013*, 2013. [Online]. Available: https://www.usenix.org/conference/tapp13/technical-sessions/presentation/chirigati

[15] J. Dongarra, P. Beckman *et al.*, "The international exascale software project roadmap," *Int. J. High Perform. Comput. Appl.*, vol. 25, no. 1, pp. 3–60, Feb. 2011. [Online]. Available: http://dx.doi.org/10.1177/1094342010391989

[16] (2012, October) PRACE - the scientific case for high performance computing in europe 2012-2020. PRACE. [Online]. Available: http://www.prace-ri.eu/IMG/pdf/prace_-_the_scientific_case_-_full_text_-.pdf

[17] G. Aloisio and S. Fiore, "Towards exascale distributed data management," *Int. J. High Perform. Comput. Appl.*, vol. 23, no. 4, pp. 398–400, Nov. 2009. [Online]. Available: http://dx.doi.org/10.1177/1094342009347702

[18] WCRP Coupled Model Intercomparison Project (CMIP). [Online]. Available: https://www.wcrp-climate.org/wgcm-cmip

[19] L. Cinquini, D. Crichton *et al.*, "The earth system grid federation: An open infrastructure for access to distributed geospatial data," *Future Generation Computer Systems*, vol. 36, pp. 400 – 417, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X13001477

[20] Working Group on Coupled Modelling. [Online]. Available: https://www.wcrp-climate.org/wgcm-overview

[21] D. Salomoni, I. Campos *et al.*, "Indigo-datacloud: a platform to facilitate seamless access to e-infrastructures," *Journal of Grid Computing*, vol. 16, no. 3, pp. 381–408, Sep 2018. [Online]. Available: https://doi.org/10.1007/s10723-018-9453-3

[22] S. Fiore, M. Płóciennik *et al.*, "Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the earth system grid federation eco-system," in *2016 IEEE International Conference on Big Data (Big Data)*, Dec 2016, pp. 2911–2918.

[23] M. Płóciennik, S. Fiore, G. Donvito, M. Owsiak, M. Fargetta, R. Barbera, R. Bruno, E. Giorgio, D. N. Williams, and G. Aloisio, "Two-level dynamic workflow orchestration in the indigo datacloud for large-scale, climate change data analytics experiments," *Procedia Computer Science*, vol. 80, pp. 722 – 733, 2016, international Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050916308341

[24] R. Rew and G. Davis, "Netcdf: an interface for scientific data access," *IEEE Computer Graphics and Applications*, vol. 10, no. 4, pp. 76–82, July 1990.

[25] M. Asch, T. Moore *et al.*, "Big data and extreme-scale computing: Pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry," *The International Journal of High Performance Computing Applications*, vol. 32, no. 4, pp. 435–479, 2018. [Online]. Available: https://doi.org/10.1177/1094342018778123

[26] myExperiment. [Online]. Available: https://www.myexperiment.org/home

[27] T. Kluyver, B. Ragan-Kelley *et al.*, "Jupyter notebooks - a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Scmidt, Eds. IOS Press, 2016, pp. 87–90. [Online]. Available: https://eprints.soton.ac.uk/403913/

[28] Jupyterhub. [Online]. Available: http://jupyter.org/hub

[29] S. Fiore, A. D'Anca, C. Palazzo, I. Foster, D. Williams, and G. Aloisio, "Ophidia: Toward big data analytics for escience," *Procedia Computer Science*, vol. 18, pp. 2376 – 2385, 2013, 2013 International Conference on Computational Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050913005528

[30] S. Fiore, C. Palazzo, A. D'Anca, I. T. Foster, D. N. Williams, and G. Aloisio, "A big data analytics framework for scientific data management," in *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*. IEEE, 2013, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/BigData.2013.6691720

[31] (2016, February) Data access and synchronization with synda. [Online]. Available: https://portal.enes.org/data/data-metadata-service/search-and-download/synda

[32] Ophidia HPDA framework project - github. [Online]. Available: https://github.com/OphidiaBigData/

[33] D. Elia, S. Fiore, A. D'Anca, C. Palazzo, I. T. Foster, D. N. Williams, and G. Aloisio, "An in-memory based framework for scientific data analytics," in *Proceedings of the ACM International Conference on Computing Frontiers, CF'16, Como, Italy, May 16-19, 2016*, 2016, pp. 424–429. [Online]. Available: http://doi.acm.org/10.1145/2903150.2911719

[34] S. Fiore, C. Palazzo *et al.*, "Big data analytics on large-scale scientific datasets in the indigo-datacloud project," in *Proceedings of the Computing Frontiers Conference*, ser. CF'17. New York, NY, USA: ACM, 2017, pp. 343–348. [Online]. Available: http://doi.acm.org/10.1145/3075564.3078884

[35] C. Palazzo, A. Mariello, S. Fiore, A. D'Anca, D. Elia, D. N. Williams, and G. Aloisio, "A workflow-enabled big data analytics software stack for escience," in *2015 International Conference on High Performance Computing Simulation (HPCS)*, July 2015, pp. 545–552.

[36] Ophidia json request schema. [Online]. Available: http://ophidia.cmcc.it/documentation/users/appendix/json_request.html

[37] J. Freire and F. S. Chirigati, "Provenance and the different flavors of reproducibility," *IEEE Data Eng. Bull.*, vol. 41, no. 1, pp. 15–26, 2018. [Online]. Available: http://sites.computer.org/debull/A18mar/p15.pdf

[38] PTA workflow document. [Online]. Available: https://github.com/OphidiaBigData/ophidia-workflow-catalogue/tree/master/indigo/precip_trend_analisys