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

Open Science in Big Data 2018 Workshop
Seattle, December 10th, 2018

S. Fiore[1], D. Elia[12], C. Palazzo[1], A. D'Anca[1], F. Antonio[1], D. N. Williams[3], I. Foster[4], G. Aloisio[12]
[1] Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Lecce, Italy
[2] University of Salento, Lecce, Italy
[3] Lawrence Livermore National Laboratory, Livermore, USA
[4] University of Chicago & Argonne National Laboratory, Chicago, USA

The ESiWACE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675191 http://www.esiwace.eu

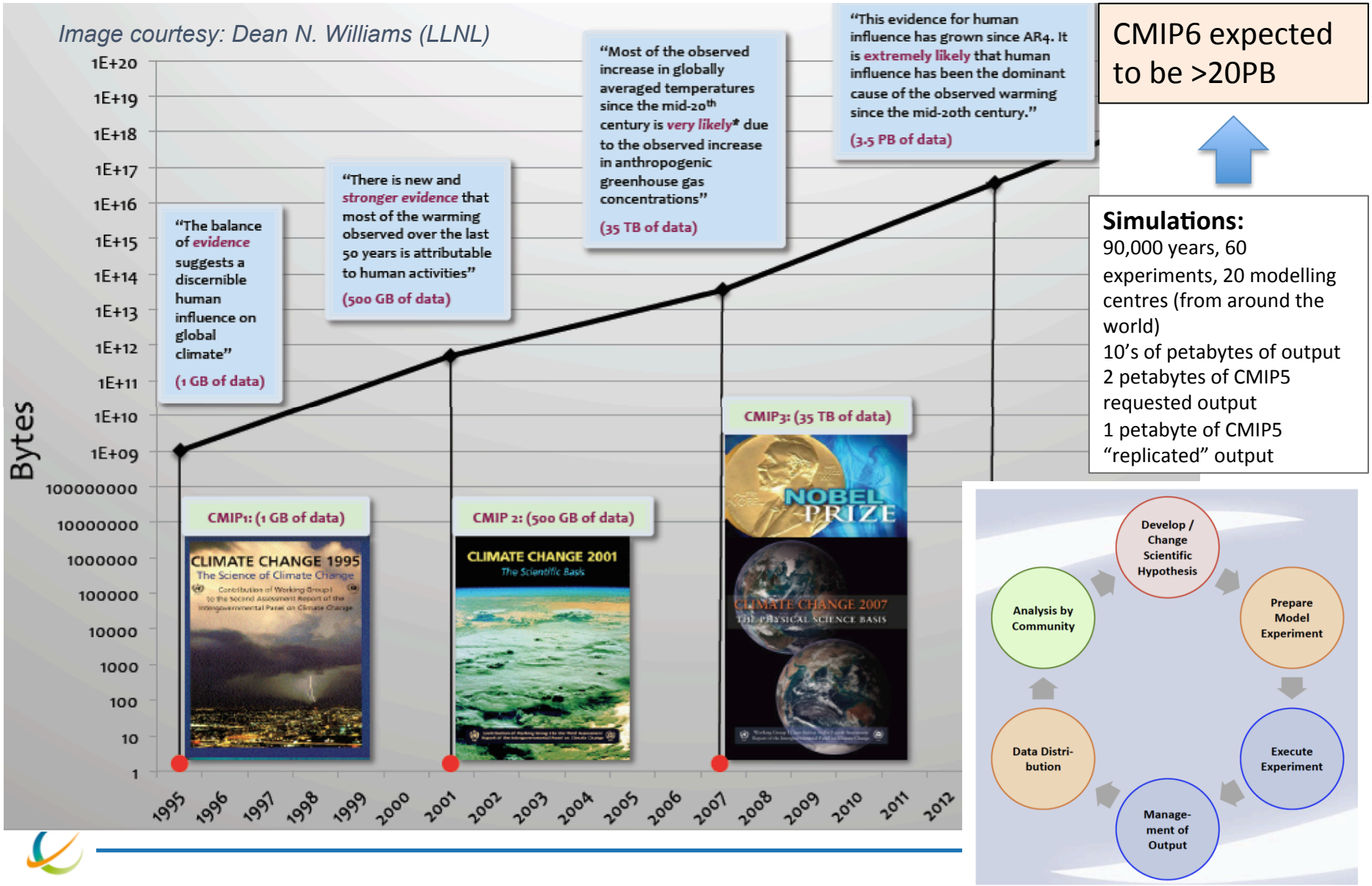
Outline

- Context: CMIP/ESGF
- Multi-model climate analysis
  - challenges and issues
- Climate Analytics-Hub
  - Architectural view
  - Key requirements
  - Architectural design and infrastructural view
  - Ophidia big data analytics framework
- Precipitation Trend Analysis case study
  - Ophidia analytics workflow
- Analytics-Hub workflows and applications reproducibility
- Conclusions and future work

Multi-model climate analysis: context and case study

- The proposed work is mainly related to the climate change community
- It is directly connected to the Coupled Model Intercomparison Project (CMIP) and the Earth System Grid Federation (ESGF) infrastructure
- CMIP is a collaborative framework designed to improve knowledge of climate change by comparing multiple model simulations with the observations and with each other

A (big) data perspective of the CMIP experiments

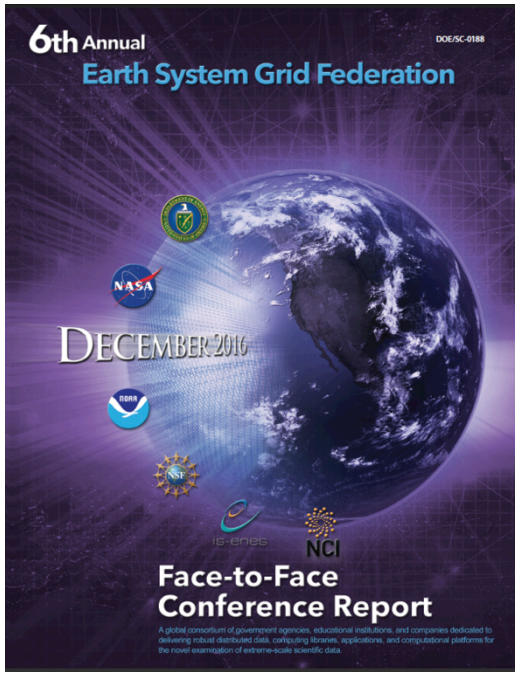


CMIP6 expected to be >20PB

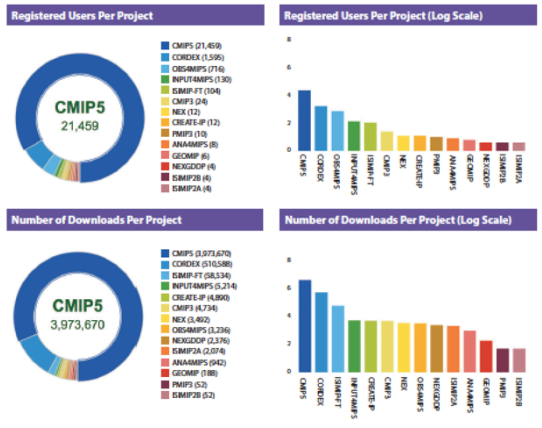
Simulations:
90,000 years, 60 experiments, 20 modelling centres (from around the world)
10's of petabytes of output
2 petabytes of CMIP5 requested output
1 petabyte of CMIP5 "replicated" output

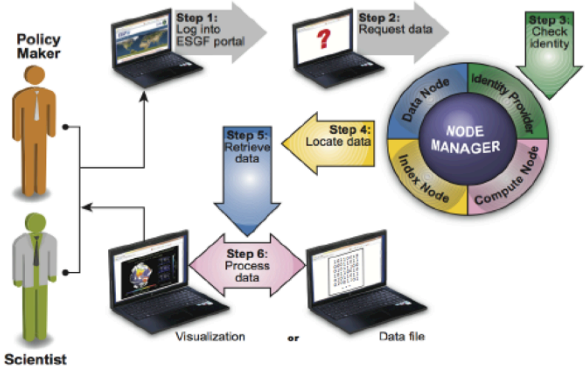
Getting access to CMIP data: ESGF

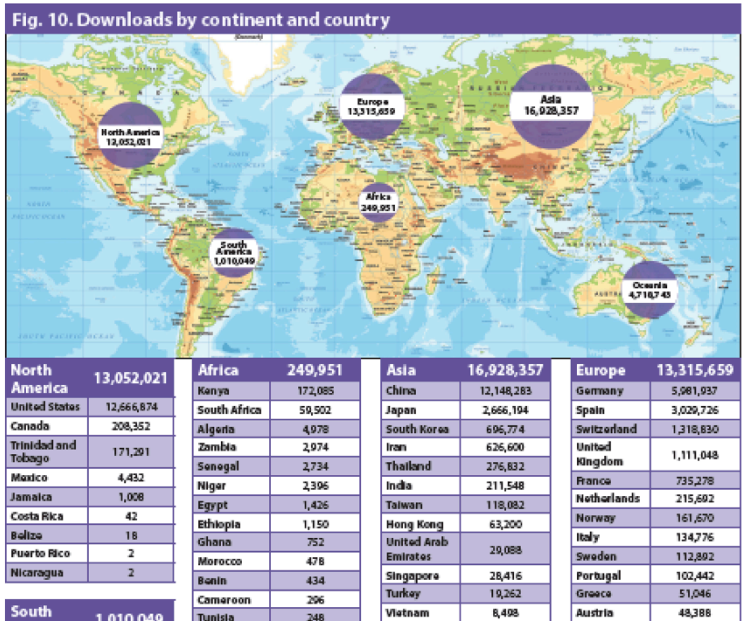
ESGF[1] is a coordinated multiagency, international collaboration of institutions that continually develop, deploy, and maintain software needed to facilitate and empower the study of climate.



6th Annual ESGF F2F Conference
December 5–9, 2016, Washington, D.C.
Convened by DOE, NASA, NOAA, NSF
IS-ENES, NCI







North America	13,052,021
United States	12,666,874
Canada	208,352
Trinidad and Tobago	171,291
Mexico	4,432
Jamaica	1,008
Costa Rica	42
Belize	18
Puerto Rico	2
Nicaragua	2

Africa	249,951
Kenya	172,085
South Africa	59,502
Algeria	4,978
Zambia	2,974
Senegal	2,734
Niger	2,396
Egypt	1,426
Ethiopia	1,150
Ghana	752
Morocco	478
Benin	434
Cameroon	296
Tunisia	248

Asia	16,928,357
China	12,148,283
Japan	2,666,194
South Korea	696,774
Iran	626,600
Thailand	276,832
India	211,548
Taiwan	118,082
Hong Kong	63,200
United Arab Emirates	29,088
Singapore	28,416
Turkey	19,262
Vietnam	8,498

Europe	13,315,659
Germany	5,981,937
Spain	3,029,726
Switzerland	1,318,830
United Kingdom	1,111,048
France	735,278
Netherlands	215,692
Norway	161,670
Italy	134,776
Sweden	112,892
Portugal	102,442
Greece	51,046
Austria	48,388

J. Chen, A. Choudhary, et al. "Synergistic Challenges in Data- Intensive Science and Exascale Computing," DOE ASCAC Data Subcommittee Report, Department of Energy Office of Science, March, 2013.

Multi-model climate analysis challenges & issues

- Multi-model data analysis experiments allow a large-scale data analytics in order to compare multiple climate models simulations

- Several **key challenges** and practical **issues** related to the multi-model climate analysis
  - Input data from multiple models needed
  - **Data download is a big barrier** for climate scientists
    - download can take a significant amount of time
    - network instability, dropped connections, etc.
  - **Data analysis** mainly performed using **client-side** & **sequential approaches**
  - Installation and update of **data analysis tools and libraries** needed
  - **Strong requirements** in terms of **computational and storage resources**

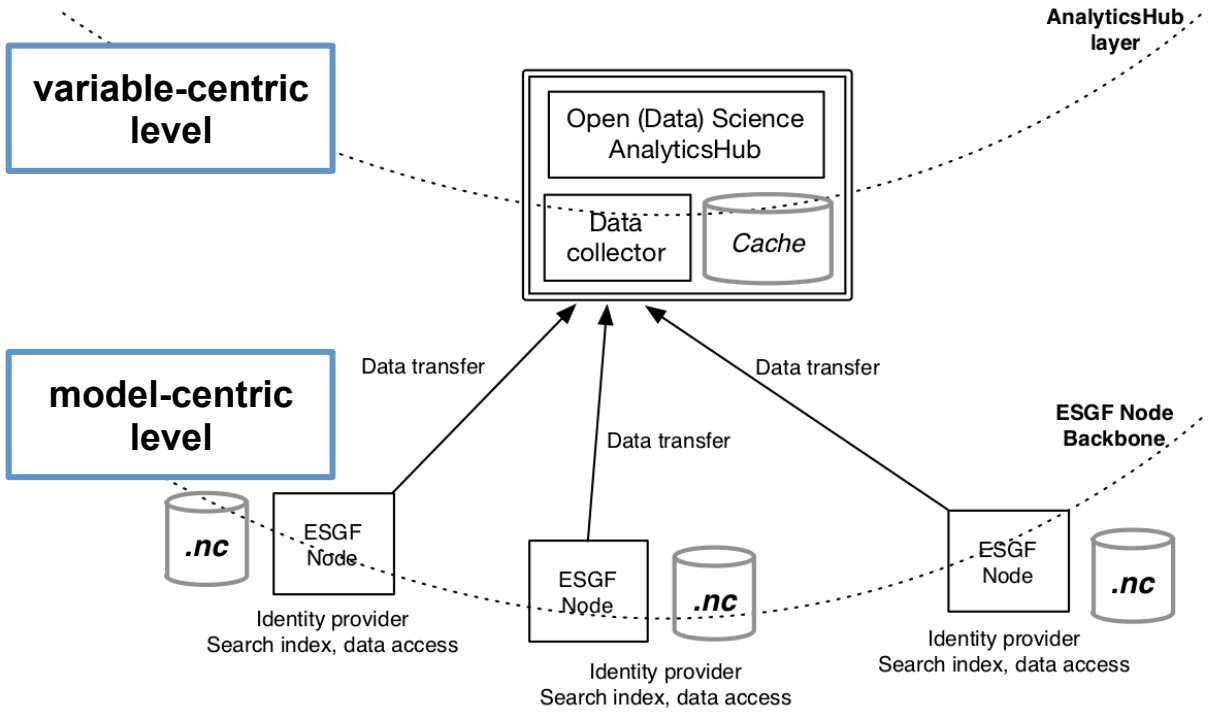
Climate Analytics-Hub (in the large view)

The proposed solution implements the **Climate Analytics-Hub level** on top of the ESGF data nodes to allow the execution of multi-model climate analyses on a single location.

The **Analytics-Hub** provides Open (Data) Science oriented computing and analytics capabilities.

The **data collector** layer

- pre-stages and caches data relevant to the analyses from the different ESGF data nodes
- synchronizes the local copy of the data with the ESGF remote repositories



Analytics-Hub requirements

- **Server-side analytics**
  - Downloaded data, task makespan and software complexity reduced
- **Big data and HPC approaches**
- **Data consistency**
  - Synchronization between Analytics-Hub cache and ESGF repository
- **Workflow-enabled analytics**
  - FAIR principles can be applied to workflows
- **Social and cultural implications**
  - From a single-user to a distributed team-driven analysis approach
- **Open (Data) Science-ready environments**
  - Code and research results sharing, collaboration through scientists


Analytics-Hub architecture and SW infrastructure

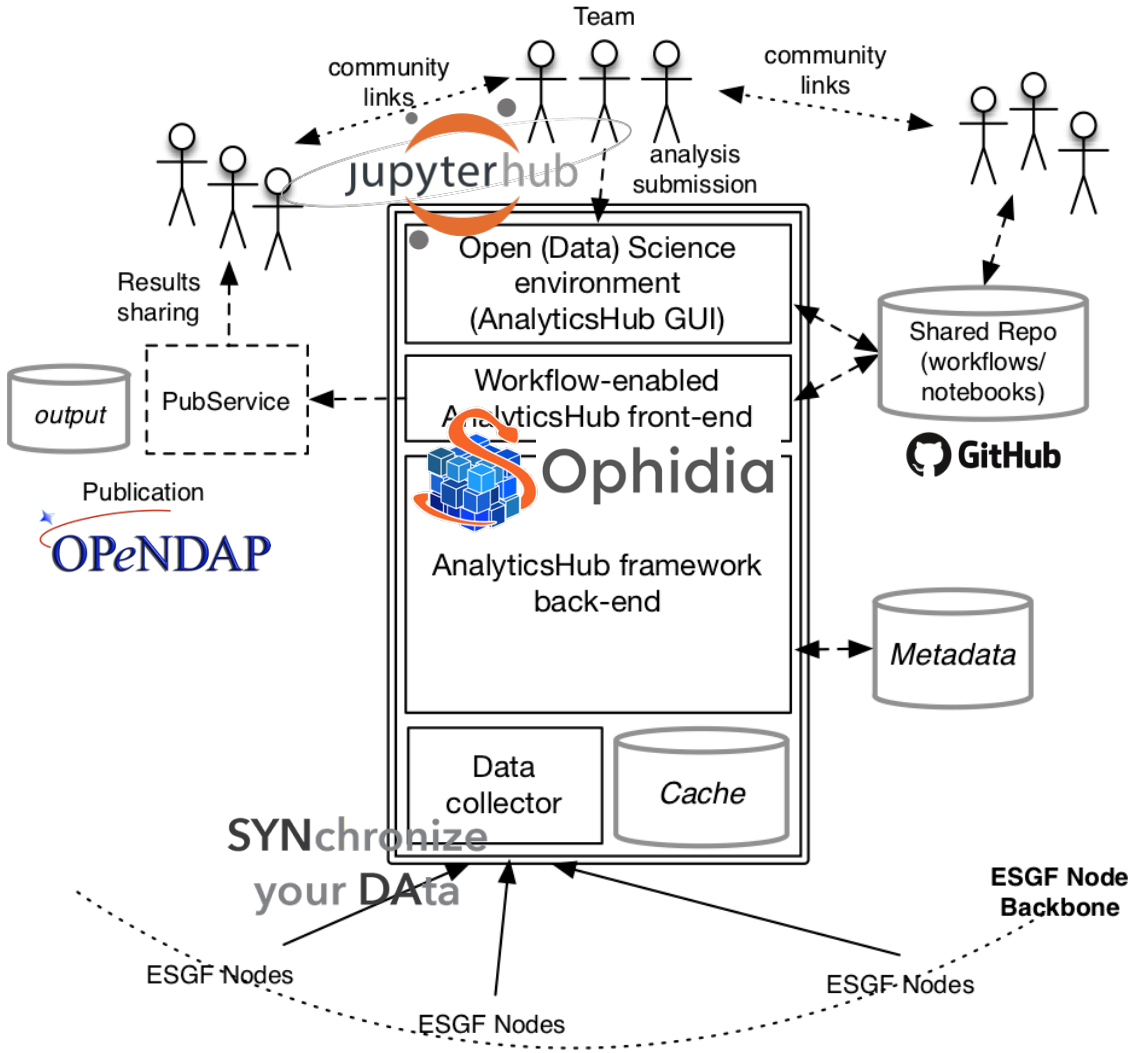
The **Analytics-Hub** consists of several components:

i. a **GUI** providing an Open (Data) Science-ready environment

ii. a **workflow-enabled, secure and interoperable front-end** to address user's requests

iii. an **analytics framework back-end** to perform data analysis at scale

iv. a **data collector** and its **local storage** to gather relevant datasets from ESGF



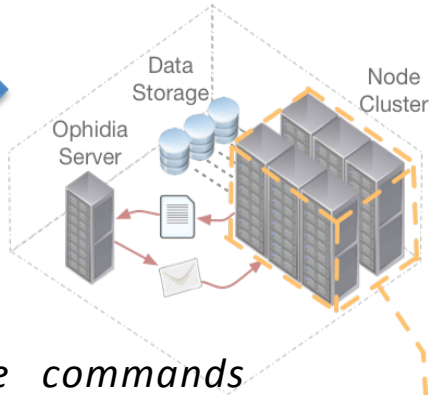
Ophidia: a scientific big data analytics framework

Ophidia (http://ophidia.cmcc.it) is a CMCC Foundation research project addressing fast and big data challenges for eScience

It provides:

- support for declarative, parallel, server-side data analysis exploiting parallel computing techniques and database approaches
- end-to-end mechanisms to support complex experiments and large processing workflows on scientific datacubes

Server-side paradigm and the datacube abstraction

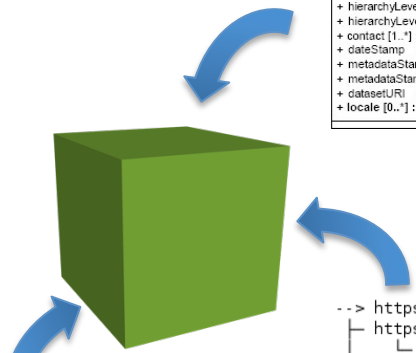


Oph_Term: *a terminal-like commands interpreter serving as client for the Ophidia framework*

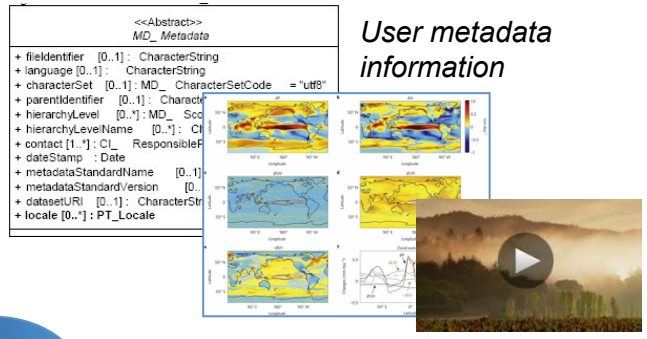
Ophidia framework: *declarative, parallel server-side processing*

Through the ***oph_term*** *the user can send commands to the Ophidia framework to manipulate datasets*

Three interaction modes:
Operators, Workflows, Python Apps



System metadata of the datacube (size, distribution, etc.)



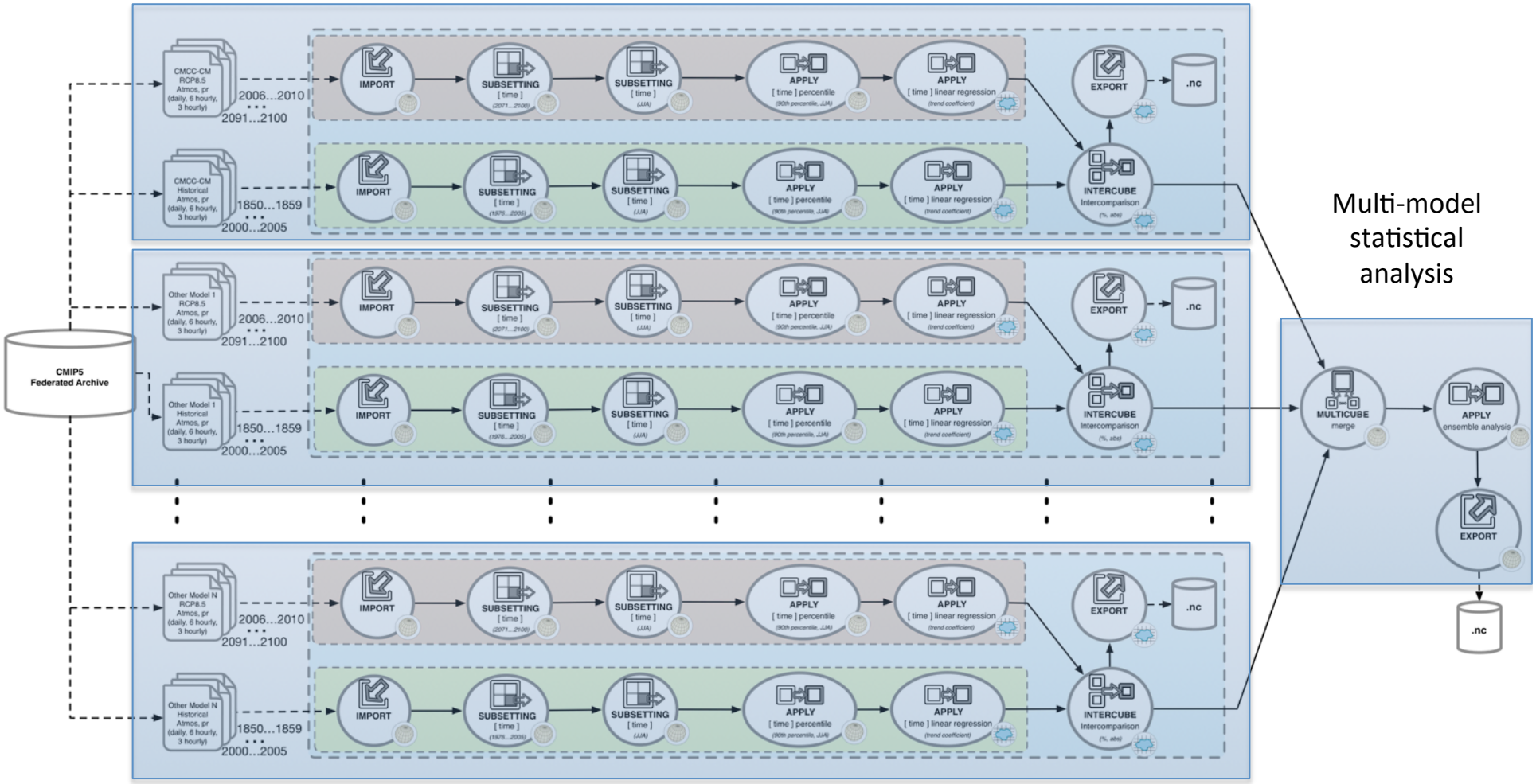
User metadata information

Metadata provenance

```
--> https://ophidia.cmcc.it:8443/162/169 (ROOT)
  ├─ https://ophidia.cmcc.it:8443/162/170 (oph_reduce)
  │    └─ https://ophidia.cmcc.it:8443/162/171 (oph_merge)
  │         ├─ https://ophidia.cmcc.it:8443/162/172 (oph_aggregate2)
  │         └─ https://ophidia.cmcc.it:8443/162/173 (oph_rollup)
  │              ├─ https://ophidia.cmcc.it:8443/162/174 (oph_reduce)
  │              └─ https://ophidia.cmcc.it:8443/162/175 (oph_reduce)
  ├─ https://ophidia.cmcc.it:8443/162/176 (oph_aggregate)
  └─ https://ophidia.cmcc.it:8443/162/177 (oph_aggregate)
```

Multi-model precipitation trend analysis (I) experiment design

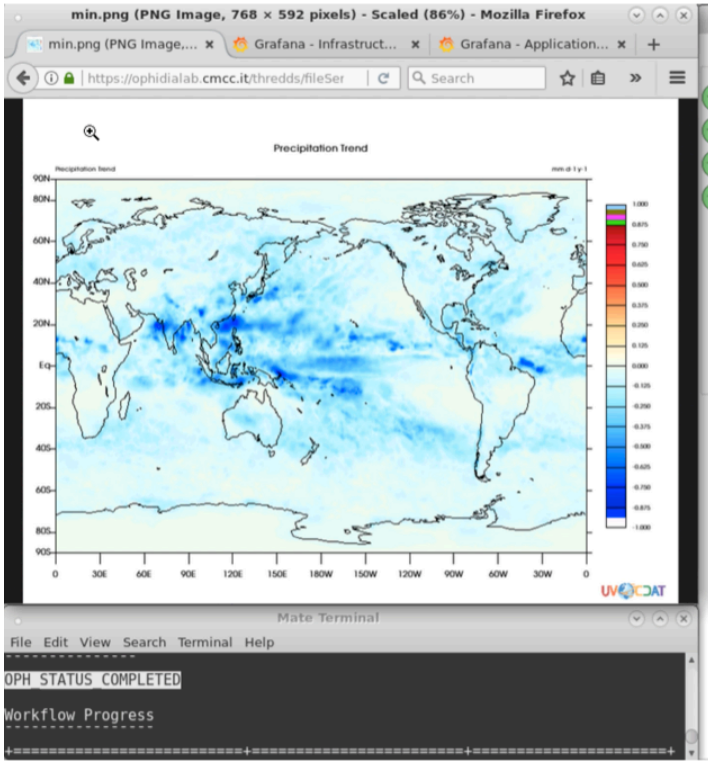
Single model precipitation trend analysis

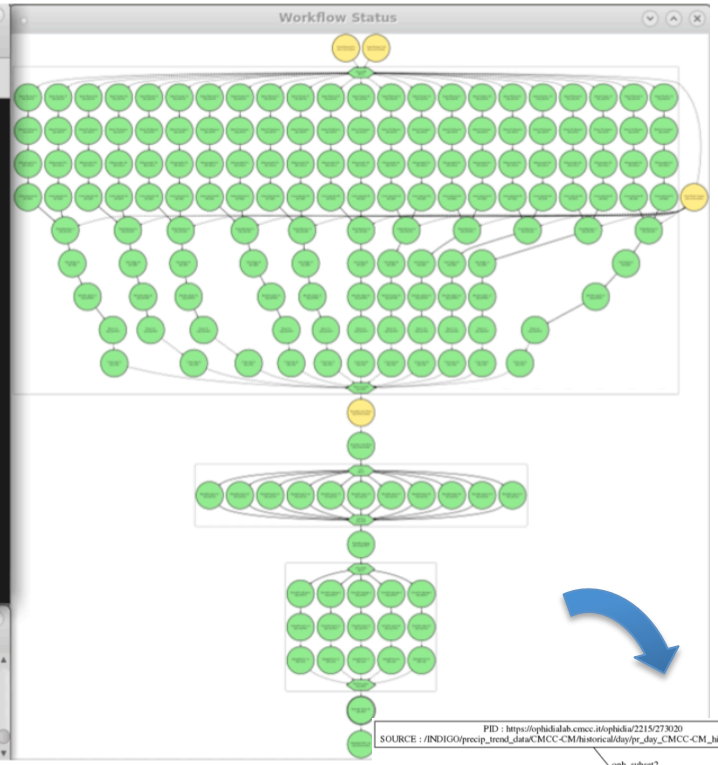


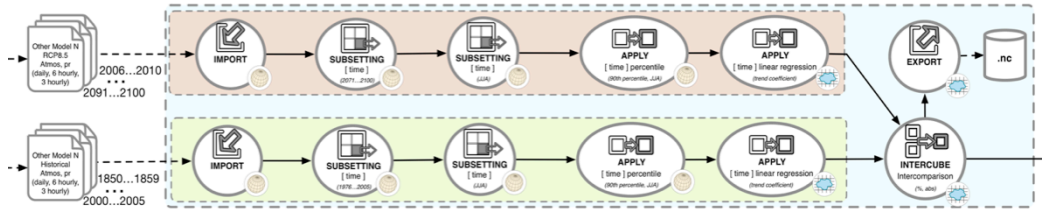
Multi-model precipitation trend analysis (II) experiment implementation

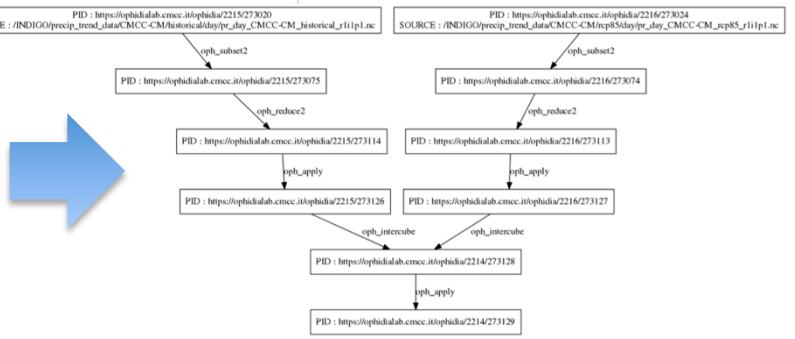
PTA test case implemented as an Ophidia analytics workflow

- 11 models from CMIP5 experiment, 181 tasks





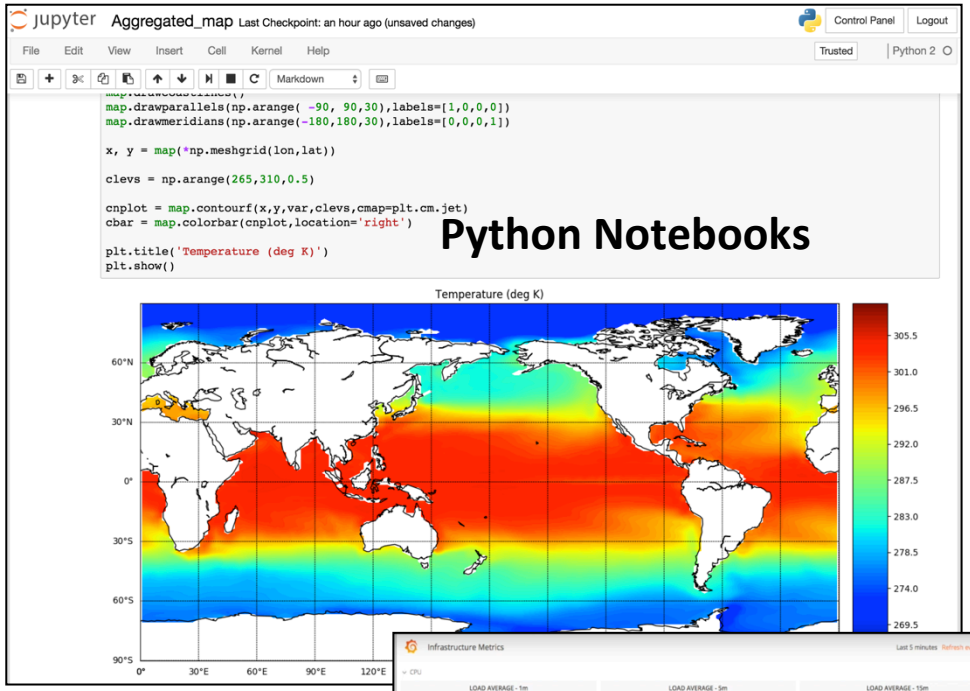




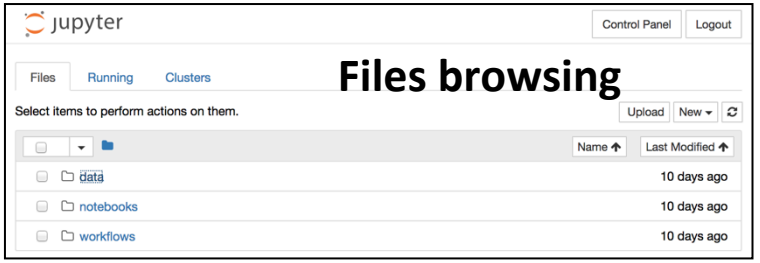
Provenance: Ophidia *oph_cubeio* operator used to retrieve the whole data lineage related to a particular datacube

YouTube video: https://youtu.be/fJgFf3UDDF4

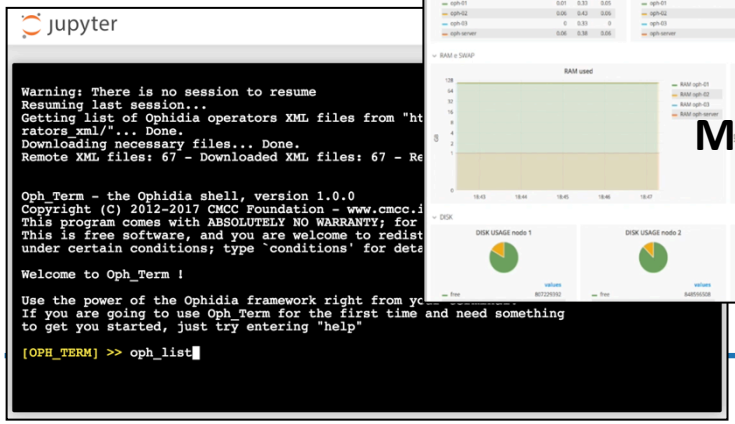
Multi-model precipitation trend analysis (III) experiment ecosystem



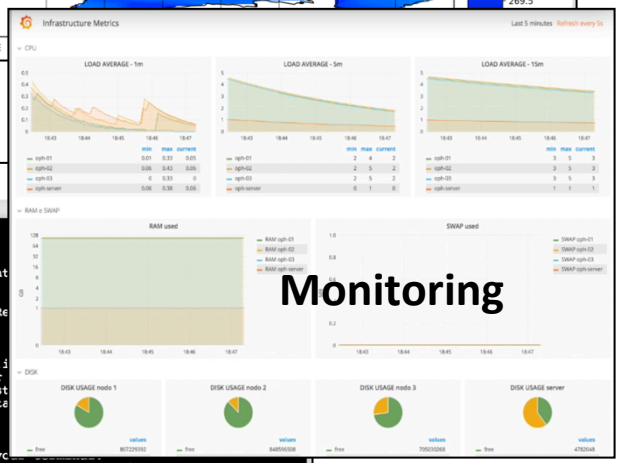
Python Notebooks



Files browsing



ECAS Terminal



Monitoring

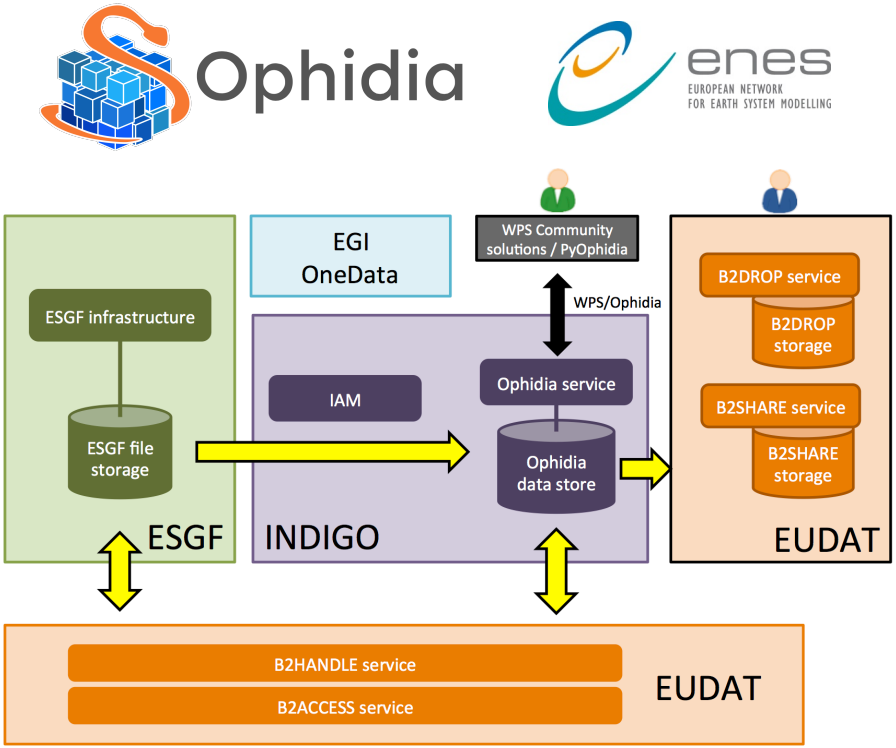
QuickStart

Open Science principles adoption
Analytics-Hub workflows & applications reproducibility

- The **Ophidia analytics document** extends the Ophidia workflow document
  - computing environment (platform, compilers, libraries, …)
  - analytics eco-system (Ophidia release, NetCDF library, Python modules, …)
  - input data (through DOIs, …)
- System-level provenance → portability → **reproducibility** → re-usability
- Ophidia supports two types of provenance
  - static prospective provenance: tracked by workflow document
  - dynamic retrospective provenance: to track at run time the provenance of each datacube
    - also for Python data analysis applications
- The Ophidia analytics document also allows
  - **reproducible executability**: by its machine-readability (JSON format)
  - **analytics document evolution provenance** → citability


EOSC Landscape

- In the **EOSC landscape**, CMCC is involved into the ECAS Thematic Service activity jointly with DKRZ
- **ECAS** is a concrete implementation of a community service supporting multiple kinds of climate data analyses
- The proposed **Analytics-hub** is mainly a novel, variable-centric, reference architecture designed on top of the **ESGF data node backbone**, specifically tailored on and addressing multi-model analysis in the **CMIP context**.
- Future work could relate to bringing the Analytics-Hub into EOSC, for instance by providing a specific ECAS release, implementing the Analytics-Hub concept/architecture.



The European Commission launched the European Open ScienceCloud Initiative to capitalise on the data revolution. EOSC will provide European science, industry and public authorities with world-class digital infrastructure that bring state of the art computing and data storage capacity to the fingertips of any scientists and engineer in the EU.

EOSC-hub receives funding from the EU's Horizon 2020 research and innovation programme under grant agreement No. 777536.

Conclusions

The **Climate Analytics-Hub paradigm**

- *creates new, refined and open **variable-centric data stores***
- ***makes the analysis process easier** by overcoming key barriers related to data download and preparation*
- *promotes **Open (Data) Science principles**: re-usability, openness and sharing of data, workflows and source code, fostering new opportunities for open research and collaborations*

Future work

- Climate Analytics-Hub set up at the CMCC Supercomputing Center on a **larger HPC facility**
- Support to **reproducible** multi-model analytics experiments in the CMIP6 context
- New **provenance support** available in the next release

Thanks!

Questions?