A Climate Analytics Hub for multi-model analysis

IS-ENES3/ESGF Virtual Workshop on Compute and Analytics

2nd December, 2019

D. Elia[1,2], P. Nassisi[1], C. Palazzo[1], F. Antonio[1], S. Fiore[1], G. Aloisio[1,2]

[1] *Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Lecce, Italy*
[2] *University of Salento, Lecce, Italy*

IS-ENES3 receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Multi-model climate analysis challenges & issues

Multi-model data analysis requires

- ✔ access to data produced by large-scale simulations for multiple climate models
- ✔ running workflows with hundreds of data analytics operators

Several **key challenges** and practical **issues** related to large-scale climate analysis

- ✔ Input data from multiple models needed
- ✔ **Data download** is a **big barrier** for climate scientists
- ✔ **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

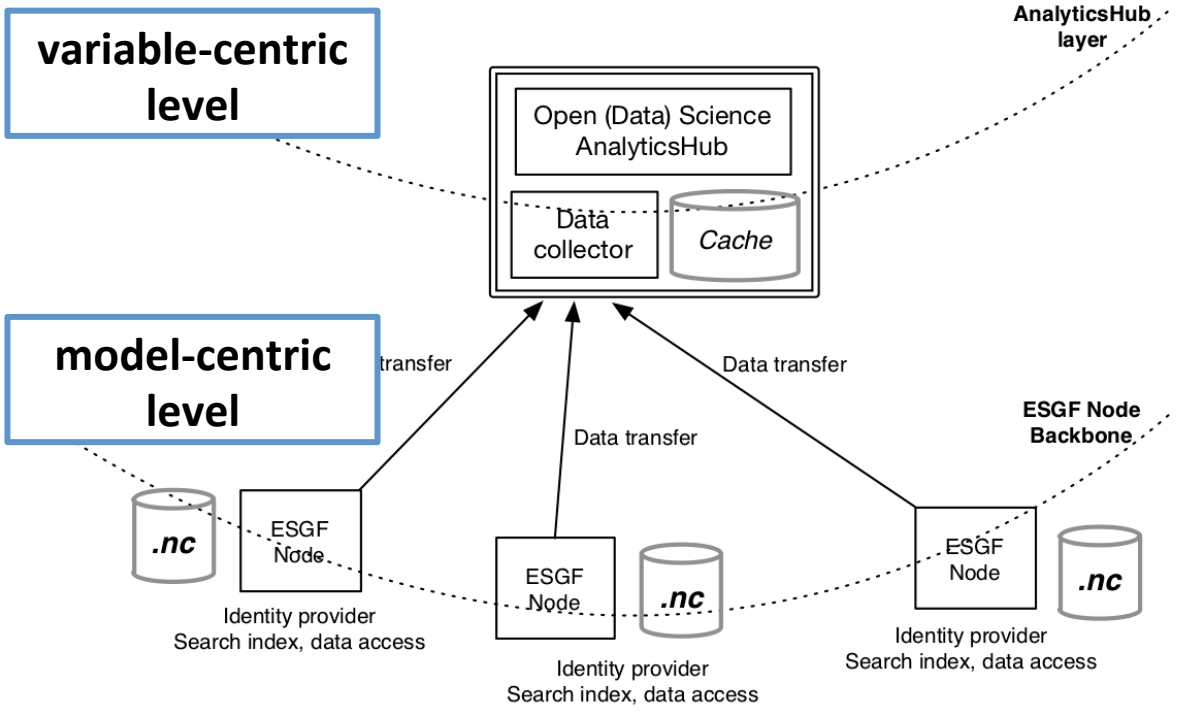
Climate Analytics-Hub

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

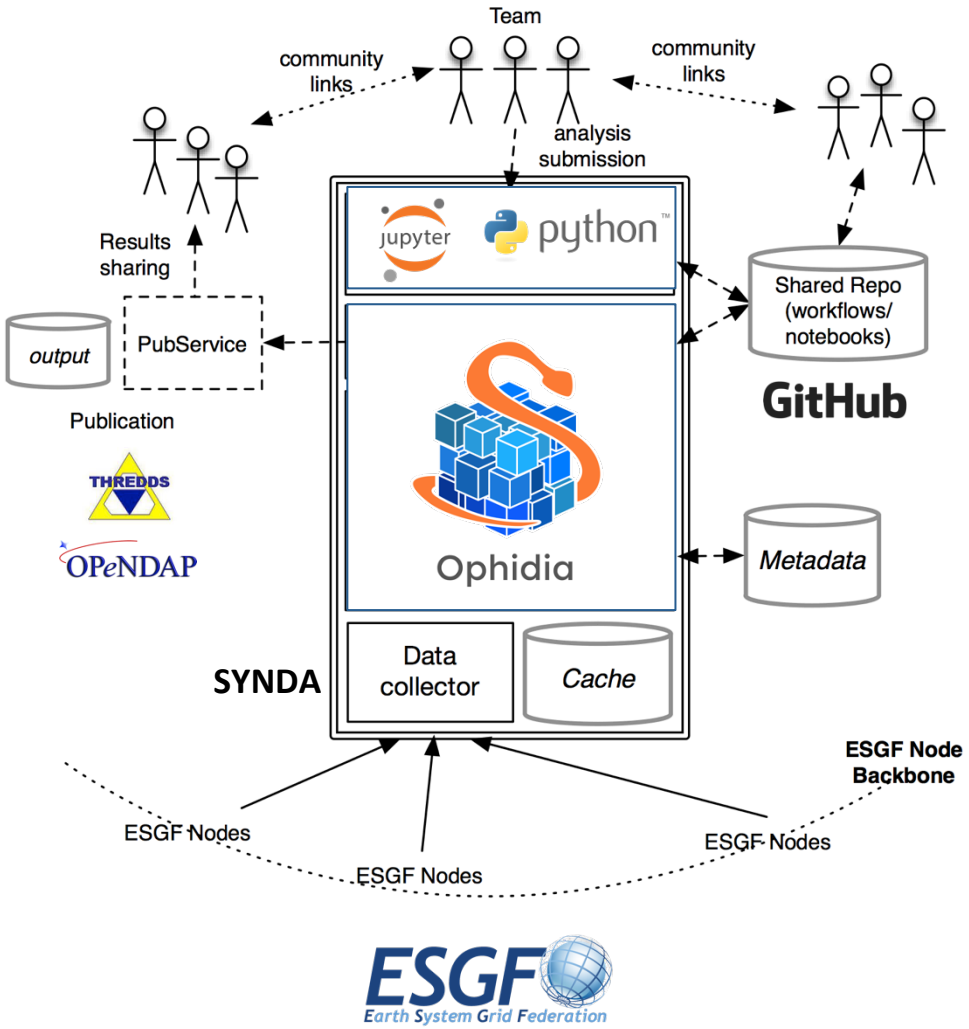
The Analytics-Hub provides Open Data Science oriented computing and analytics capabilities.

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



The CMCC Analytics Hub

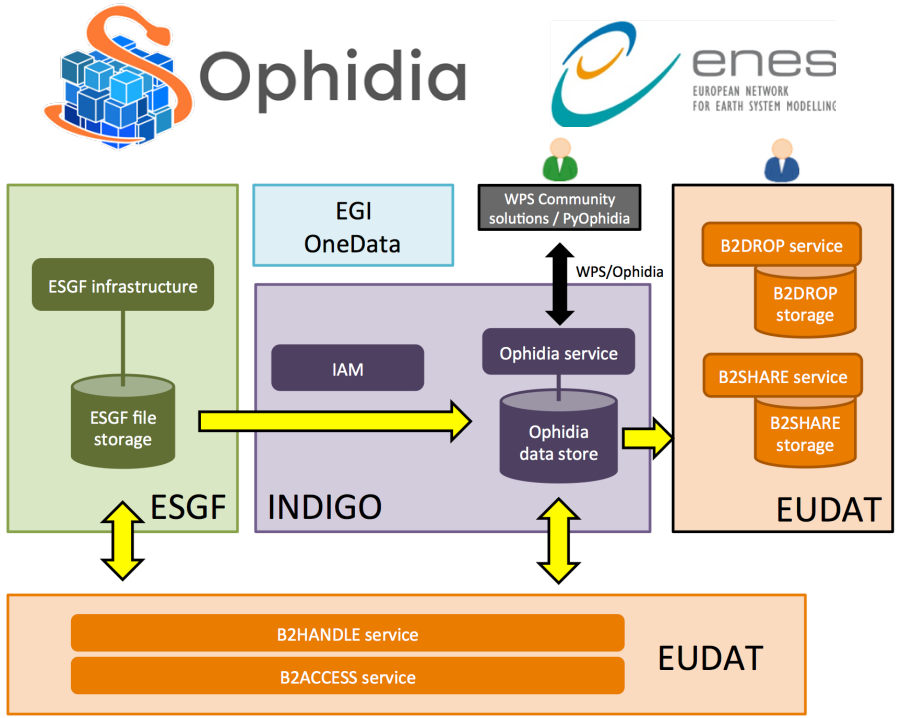


The **Analytics-Hub** consists of several components:

- ✔ A graphical **virtual environment** for (open) data science (e.g. JupyterHub)
- ✔ A wide set of high-level (Python) **scientific libraries** for analysis and plotting
- ✔ **Data Analytics and Machine Learning (HPDA) frameworks** for data science
- ✔ A user-oriented **monitoring system** to track application execution
- ✔ The **data collector** (Synda) and the local storage to gather relevant datasets from ESGF

ENES Climate Analytics Service (ECAS)

- ✔ The Analytics-Hub is a paradigm joining data and computing able to provide a **multi-model environment** for CMIP-based analytics experiments in ESGF
- ✔ The **ENES Climate Analytics Service (ECAS)**, proposed by CMCC & DKRZ in EOSC-hub supports climate data analysis
- ✔ It is one of the **EOSC-Hub Thematic Services**
- ✔ ECAS builds on top of the **Ophidia big data analytics framework** with components from INDIGO-DataCloud, EUDAT and EGI



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.

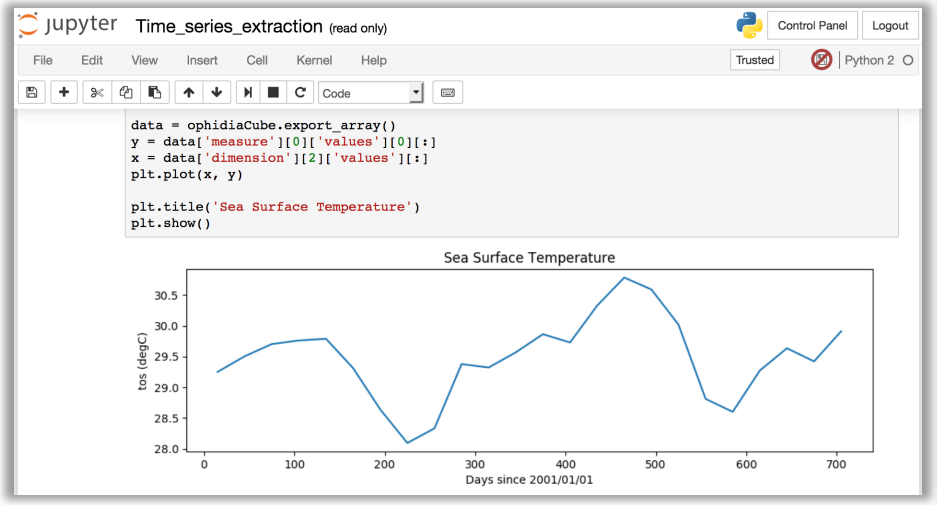
EOSC-hub cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici DKRZ DEUTSCHES KLIMARECHENZENTRUM

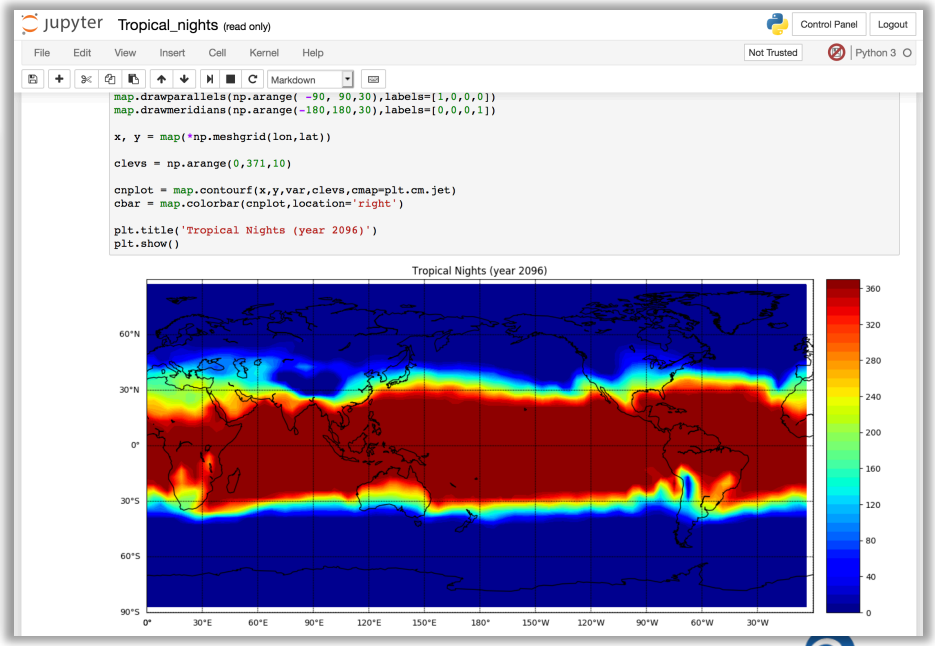
EOSC-hub receives funding from the EU's Horizon 2020 research and innovation programme under grant agreement No. 777536.

ECASLab: Python environment for Data Science

ECASLab provides a ready-to-use environment based on JupyterHub, Ophidia and other services from EUDAT and EGI, bundled with a wide set of well-known Python scientific and data management modules, some examples:

- NumPy, SciPy, Pandas
- NetCDF, PyOphidia, Scikit-learn
- Matplotlib, basemap, Cartopy





Programmatic access to ECAS through the PyOphidia class

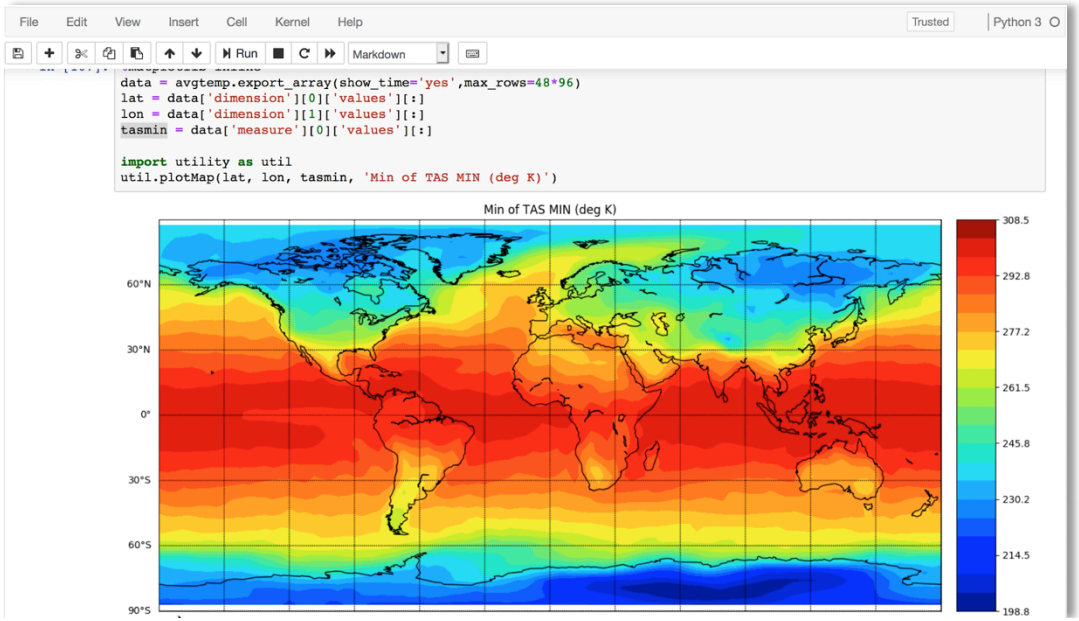
- ✔ **PyOphidia** provides a Python interface to submit commands to ECAS and to retrieve/deserialize the results (e.g. in Jupyter Notebooks)
- ✔ Two modules implemented:
  - ✔ **Client**: supports the submissions of Ophidia commands and workflows, as well as the management of session from Python code (similar to the Ophidia Terminal)
  - ✔ **Cube class**: provides the datacube type abstraction and the methods to manipulate, process and get information on cubes objects

```
from PyOphidia import cube, client
cube.Cube.setclient(read_env=True)

mycube =
cube.Cube.importnc(src_path='/public/data/ecas_training
/file.nc', measure='tos', imp_dim='time',
import_metadata='yes', ncores=5)
mycube2 = mycube.reduce(operation='max',ncores=5)
mycube3 = mycube2.rollup(ncores=5)
data = mycube3.export_array()

mycube3.exportnc2(output_path='/home/test',
export_metadata='yes')
```

https://github.com/OphidiaBigData/PyOphidia
https://pypi.org/project/PyOphidia/
https://anaconda.org/conda-forge/pyophidia

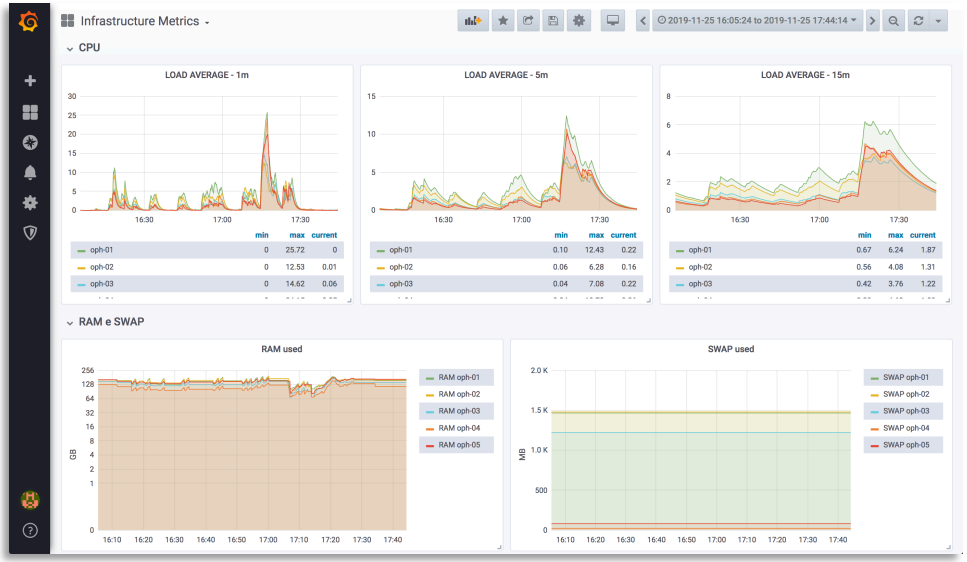


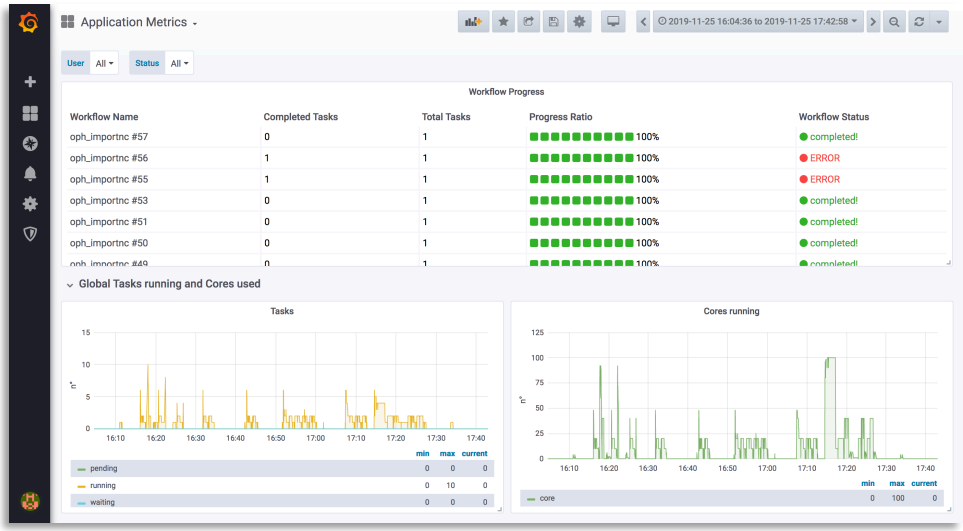
Real-time monitoring of applications

A **Grafana**-based system is used for real-time monitoring of the analytics-hub environment and the applications being executed.

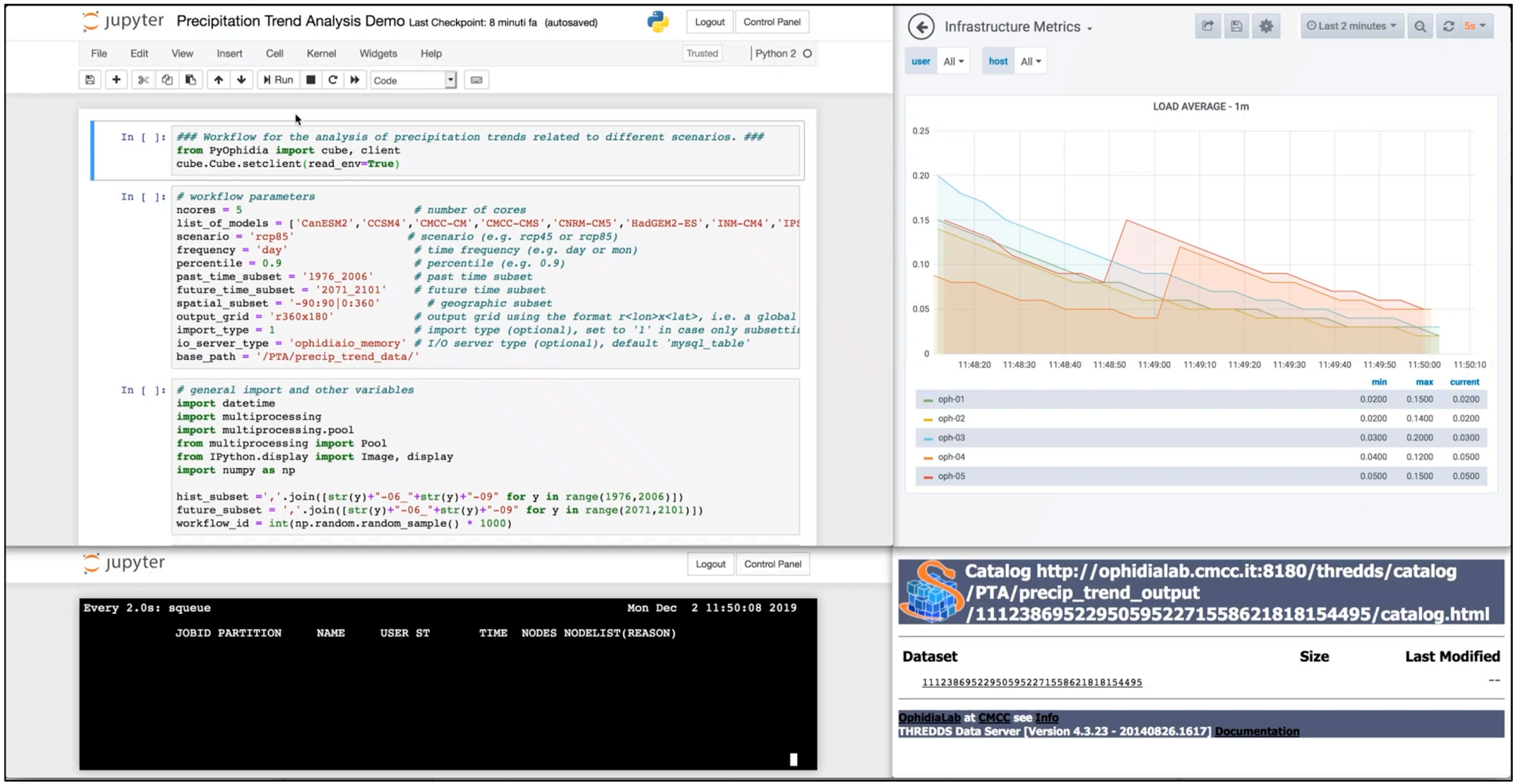
Custom dashboards have been designed to track:

- ✔ **Infrastructure** metrics:
  - resource usage on the environment nodes (CPU, memory, disk)
- ✔ **Application** metrics:
  - status of each operators and number of workflows/operator running





Multi-model experiment demo



Conclusions & next steps

Recap:

- ✔ The Analytics-Hub
  - defines a **variable-centric** environment for data analysis
  - **reduces time-to-solution** by removing the download barrier and providing server-side and high performance data analytics
  - provides a **user-friendly** environment for development, testing and execution of data analytics experiments

Next steps include:

- ✔ Implementation of multi-model experiments with CMIP6 data
- ✔ Deployment of the Analytics-Hub on CMCC new supercomputer
- ✔ Integration of additional data analytics tools to target a wider set of climate applications

Thanks

Useful links:

- ✔ CMCC ECASLab instance: https://ecaslab.cmcc.it/
- ✔ Ophidia Website: http://ophidia.cmcc.it
- ✔ Ophidia Doc: http://ophidia.cmcc.it/documentation
- ✔ PyOphidia repository: https://github.com/OphidiaBigData/PyOphidia
- ✔ Contact us at: ophidia-info@cmcc.it