A Data Science environment for climate change research at CMCC

D. Elia[1,2], P. Nassisi[1], C. Palazzo[1], F. Antonio[1], B. Sbarro[1], A. D'Anca[1], S. Fiore[1], G. Aloisio[1,2]

[1] Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Lecce, Italy - [2] Università del Salento, Lecce, Italy

ENVIRONMENT GOAL & MOTIVATION

The **CMCC Data Science environment** aims to provide a *data science and learning based end-to-end scientific environment to support climate change research at scale,* seamlessly integrated into a single high performance problem solving environment, deployed at the CMCC SuperComputing Centre [1].

It will enable climate scientists to address key challenges, such as:

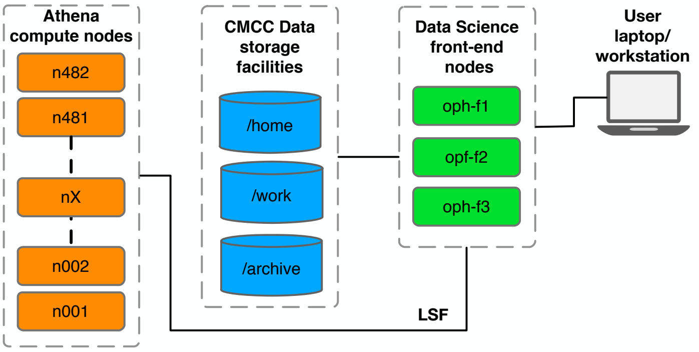
- *managing large scientific end-to-end climate experiments (aka workflows);*
- *performing interactive data exploration;*
- *analyzing massive datasets;*
- *developing user-oriented high-level data science applications* (e.g. applying ML and DL techniques) *to extract patterns and insights*;

with the ultimate goal of tackling *much larger* and *more complex* (data) science problems than are possible today in the climate change domain.

ENVIRONMENT SETUP

The first implementation of the environment includes the:

- *setup* and configuration of 3 fat-nodes at the CMCC SuperComputing Centre to host the environment software stack;
- *deployment* of a first set of Data Science technologies including, among others, *JupyterHub, Jupyter Notebooks, Python scientific libraries* and *Ophidia***;**
- *integration* with the Athena environment in terms of user management, data storage (*GPFS*) and compute resources (through *LSF*).

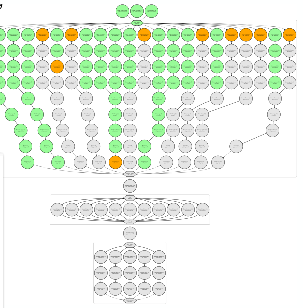


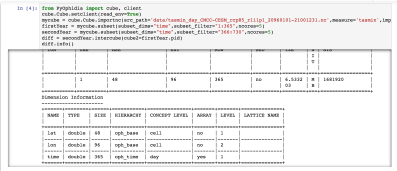
OPHIDIA HPDA FRAMEWORK

The *Ophidia High Performance Data Analytics (HPDA) framework* addresses On-Line Analytical Processing (OLAP) scientific data management by joining HPC paradigms and big data approaches. Key use cases enabled by Ophidia are analytics workflows for *climate indicators, climate diagnostics, multi-model analysis*, and *interactive data analysis*. Ophidia also provides the Python bindings (*PyOphidia*), which allows an easy integration/interaction with the wide Python-based Data Science ecosystem.

Among others, it provides the features to perform:

- *data reduction, subsetting and intercomparison;*
- *metadata and provenance management;*
- *time series analysis (with more than 100 array-based primitives);*
- *interactive and batch processing.*





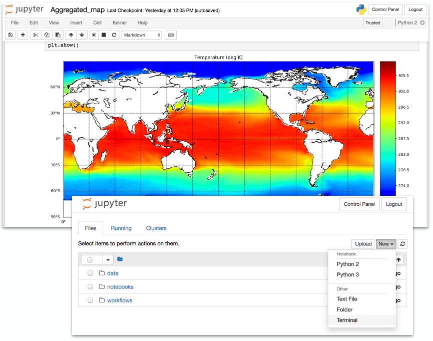
```
In [4]: from PyOphidia import cube, client
        cube.Cube.setclient(read_env=True)
        mycube = cube.Cube.importnc(src_path='data/tasmin_day_CMCC-CESM_rcp85_r1i1p1_20960101-21001231.nc',measure='tasmin',imp
        firstYear = mycube.subset(subset_dims="time",subset_filter="1:365",ncores=5)
        secondYear = mycube.subset(subset_dims="time",subset_filter="366:730",ncores=5)
        diff = secondYear.intercube(cube2=firstYear.pid)
        diff.info()
```

```
| ION     | UBE     | ASE              | ENT          | ROW          | SED      | IZE     | N | NTS             |
|         |         |                  |              |              |          |         | I |                 |
|         |         |                  |              |              |          |         | T |                 |
+=========+=========+==================+==============+==============+==========+=========+===+=================+
|         | 1       | 48               | 96           | 365          | no       | 6.5332  | M | 1681920         |
|         |         |                  |              |              |          | 03      | B |                 |
+=========+=========+==================+==============+==============+==========+=========+===+=================+
Dimension Information
---------------------
+======+========+======+===========+===============+=======+=======+=============+
| NAME | TYPE   | SIZE | HIERARCHY | CONCEPT LEVEL | ARRAY | LEVEL | LATTICE NAME |
|      |        |      |           |               |       |       |              |
+======+========+======+===========+===============+=======+=======+=============+
| lat  | double | 48   | oph_base  | cell          | no    | 1     |              |
|------|--------|------|-----------|---------------|-------|-------|--------------|
| lon  | double | 96   | oph_base  | cell          | no    | 2     |              |
|------|--------|------|-----------|---------------|-------|-------|--------------|
| time | double | 365  | oph_time  | day           | yes   | 1     |              |
+======+========+======+===========+===============+=======+=======+=============+
```

JUPYTERHUB & PYTHON ENVIRONMENT

JupyterHub is a web-based service enabling multiple users to create, execute and share *Jupyter Notebooks* (Python-based) for live-coding and visualization.

It serves as the *environment front-end*, providing access to the underlying resources and to a comprehensive eco-system of well-known *data science Python modules* for analysis and visualization.



CYLC WORKFLOW ENGINE

Cylc is a workflow engine for orchestrating complex distributed suites of interdependent cycling (repeating) tasks, as well as ordinary non-cycling workflows. It was developed with an innovative self-organizing scheduling algorithm, based on which each task knows exactly its own inputs and outputs and negotiates dependencies, so that correct scheduling emerges naturally at run time.

It has been widely adopted for weather, climate, and forecasting applications, as well as further developed and exploited in the EU H2020 *Centre of Excellence in Simulation of Weather and Climate in Europe (ESiWACE)* project.

In the context of the CMCC Data Science environment, Cylc has been selected as the *end-to-end workflow management system* to enable large-scale climate simulations at the Data Center level.

DOWNLOAD MANAGER

The download manager is a tool for searching and downloading files from the *Earth System Grid Federation (ESGF)* archive. It represents a valid alternative to the ESGF web front-end as it can be easily used as backend service: a daemon handles a dataset list to be downloaded and synchronizes a local repository to the federated archive.

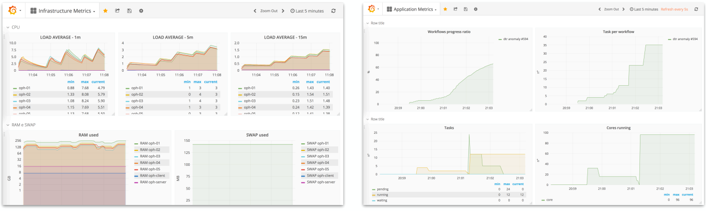
The download manager has been included in the Data Science environment to retrieve selected outputs for multiple models in the context of *CMIP5* and *CMIP6*.

MIP_Table	Model	Experiment	Ensemble	Version
3hr	ACCESS1-0	rcp45	r1i1p1	v20110323
day	ACCESS1-0	historical	r2i1p1	v20130331
day	ACCESS1-0	rcp45	r1i1p1	v20130331
…				
Amon	CMCC-CESM	rcp45	r2i1p1	v20110601
3hr	CMCC-CESM	historical	r1i1p1	v20120301
3hr	CMCC-CESM	rcp45	r1i1p1	v20110601
Amon	CMCC-CM	historical	r10i1p1	v20110601
Amon	CMCC-CM	rcp45	r2i1p1	v20110601
3hr	CMCC-CMS	rcp85	r2i1p3	v20160513
Amon	CMCC-CMS	historical	r1i1p1	v20160502
Amon	CMCC-CMS	rcp45	r1i1p1	v20120724
Amon	CMCC-CMS	rcp45	r1i1p1	v20160802
Amon	CMCC-CMS	rcp85	r1i1p1	v20160512
Amon	CMCC-CMS	rcp85	r1i1p1	v20130808
Amon	CMCC-CMS	historical	r10i1p1	v20110601
…				

MONITORING SYSTEM

The Data Science software modules run both on the Athena compute nodes, as well as on the front-end nodes. Additionally, concerning Ophidia, this can execute *single operators, massive tasks* and *workflows of multiple tasks.*

Accordingly, a *Grafana*-based system is included to monitor *resource usage* and *job submissions* by the single users at the level of the environment.



ESiWACE and ESiWACE2 projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 675191 and 823988 respectively.

[1] S. Fiore, D. Elia, C. Palazzo, A. D'Anca, I. Foster and G. Aloisio, "An HPC-enabled Data Science and Learning Environment for Climate Change Experiments at Scale", International Conference on Parallel Computing (ParCo2019), Prague, Czech Republic, 10-13 September 2019 [Accepted]