EGU2020-17031
https://doi.org/10.5194/egusphere-egu2020-17031
EGU General Assembly 2020

A Python-oriented environment for climate experiments at scale in the frame of the European Open Science Cloud

Donatello Elia[1,2], Fabrizio Antonio[1], Cosimo Palazzo[1], Paola Nassisi[1], Sofiane Bendoukha[3], Regina Kwee-Hinzmann[3], Sandro Fiore[1], Tobias Weigel[3], Hannes Thiemann[3], and Giovanni Aloisio[1,2]

[1]Euro-Mediterranean Center on Climate Change (CMCC) Foundation, Lecce, Italy

[2]University of Salento, Lecce, Italy

[3]Deutsches Klimarechenzentrum (DKRZ), Hamburg, Germany

Scientific data analysis experiments and applications require software capable of handling domain-specific and data-intensive workflows. The increasing volume of scientific data is further exacerbating these data management and analytics challenges, pushing the community towards the definition of novel programming environments for dealing efficiently with complex experiments, while abstracting from the underlying computing infrastructure.

ECASLab provides a user-friendly data analytics environment to support scientists in their daily research activities, in particular in the climate change domain, by integrating analysis tools with scientific datasets (e.g., from the ESGF data archive) and computing resources (i.e., Cloud and HPC-based). It combines the features of the ENES Climate Analytics Service (ECAS) and the JupyterHub service, with a wide set of scientific libraries from the Python landscape for data manipulation, analysis and visualization. ECASLab is being set up in the frame of the European Open Science Cloud (EOSC) platform - in the EU H2020 EOSC-Hub project - by CMCC (https://ecaslab.cmcc.it/) and DKRZ (https://ecaslab.dkrz.de/), which host two major instances of the environment.

ECAS, which lies at the heart of ECASLab, enables scientists to perform data analysis experiments on large volumes of multi-dimensional data by providing a workflow-oriented, PID-supported, server-side and distributed computing approach. ECAS consists of multiple components, centered around the Ophidia High Performance Data Analytics framework, which has been integrated with data access and sharing services (e.g., EUDAT B2DROP/B2SHARE, Onedata), along with the EGI federated cloud infrastructure. The integration with JupyterHub provides a convenient interface for scientists to access the ECAS features for the development and execution of experiments, as well as for sharing results (and the experiment/workflow definition itself). ECAS parallel data analytics capabilities can be easily exploited in Jupyter Notebooks (by means of PyOphidia, the Ophidia Python bindings) together with well-known Python modules for processing and for plotting the results on charts and maps (e.g., Dask, Xarray, NumPy, Matplotlib, etc.). ECAS is also one of the compute services made available to climate scientists by the EU H2020 IS-ENES3 project.

Hence, this integrated environment represents a complete software stack for the design and run

of interactive experiments as well as complex and data-intensive workflows. One class of such large-scale workflows, efficiently implemented through the environment resources, refers to multi-model data analysis in the context of both CMIP5 and CMIP6 (i.e., precipitation trend analysis orchestrated in parallel over multiple CMIP-based datasets).