Abstract: Improving Climate Data Access through integration of the DARE Platform, the ENES CDI and EUDAT B2 Services

Christian Pagé
CECI, Université de Toulouse, CNRS, Cerfacs
Toulouse, France

Wim Som de Cerff
KNMI
De Bilt, The Netherlands

Xavier Pivan
Toulouse, France

Alessandro Spinuso
KNMI
De Bilt, The Netherlands

Maarten Plieger
KNMI
De Bilt, The Netherlands

ABSTRACT

Supporting data analytics in climate research with respect to data access is a challenge due to increasing data volumes. This is especially true for researchers in other scientific domains using climate data, such as crop modelers, hydrologists, and other climate change impacts communities, who have limited bandwidth and hardware to support remote large data volumes. Several international and European initiatives have emerged and provide standalone solutions that offer potential for interoperability.In Europe, the IS-ENES (https://is.enes.org) consortium has developed a platform, that is a component of the ENES CDI (Climate Data Infrastructure), to ease access to climate data for the climate impact community (C4I: https://climate4impact.eu). It exposes data from ESGF data nodes as well as any OpenDAP server. It provides user interfaces, wizards and services for search and discovery, visualization, processing and downloading. This platform is very useful for many categories of users, and it will be presented.

One of the important aspect of the C4I platform is that it enables users to perform on-demand data analysis calculations through its backbone based on a collection of OGC WPS (Web Processing Service). These, coupled with authorisation mechanisms based on access tokens, enable the delegation of the calculations onto distributed infrastructures and the controlled management of the results. These characteristics have been further extended with provenance integration, especially to obtain the traceable calculation of climate impact indicators, in the context of the CLIPC project [3,6]. A solution based on a standard representation (W3C-PROV) and a set of lineage management and workflows tools [5] that will scale to other computational use cases, and that will be interoperable with ongoing european initiatives.

In Europe, an emerging e-infrastructure is being designed and built for several scientific domains, lead by EUDAT (https://eudat.eu) and EGI (https://egi.eu), which will form the basis of the future European Open Science Cloud (EOSC) to support scientific researchers. This e-infrastructure provides services within the EUDAT CDI (Collaborative Data Infrastructure). The ENES climate community is participating in the EUDAT CDI. Also, the EU Big Data Europe project (https://www.big-data-europe.eu) has developed the BDE Platform, which combine several technologies aimed at scientific data processing.

Within the new EU DARE H2020-funded project (http://project-dare.eu), the DARE Platform (formerly BDE Platform) will be integrated with the C4I Platform, through the use of EUDAT B2-Services. It will also integrate the ESGF Computing and Data Nodes that are part of the ENES CDI. An example Use Case of a relevant climate data analysis is to produce multi-model, multi-scenario, time-series average datasets of the Earth's surface temperatures for Western-Europe, using CMIP [7] data. Input data for this Use Case requires the processing of high-resolution datasets with

daily temperature covering the period 1950-2100, and produced by several climate models and with several Greenhouse Gas scenarios. New high-resolution climate projections that are part of the upcoming CMIP6 will generate a large amount of data (HighResMIP experiments). Each model produces 5-year files of approximately 5 Gigabytes (highly variable). A common analysis is to calculate climate indices, for example the Daytime Temperature Range (DTR) which uses the maximum and minimum daily temperature. If one would want to calculate the DTR anomaly between the future climate for 4 different Greenhouse Gas Scenarios, for all the CMIP models (36) and all ensemble members (10), this would lead to a total size of 2768 files and 14 Tb. This is way beyond the capacity of researchers working in the climate change impact community. To bring analytics closer to data storage, this input data must be processed in stages of location-dependent workflows, accessing close-to-data computational facilities for data-reduction tasks, as well as intermediate and on-demand resources for the generation of derived products. Schematics of the integration will be presented, along with some background information and motivations.

Keywords— science-gateway; climate; research; provenance; cloud; analytics; processing; infrastructure; EOSC; EGI; EUDAT; ENES; DARE; H2020

References

[1] Semeijn, Bram, 2013: Climate4impact: a new gateway for the global climate impact community. GeoQ, **6**, p. 24-25.

[2] Pagé, Christian, Wim Som de Cerff, Maarten Plieger, Chris Jack, Antonio S. Cofiño, Manuel Vega, Michael Kolax, Lars Bärring, Ole Bøssing Christensen, Marius Matreata, 2017: Report on derived products in CLIMATE4IMPACT. FP7 IS-ENES2 D11.6 Deliverable, 23 pp.

[3] A. Mihajlovski, A. Spinuso, M. Plieger, and W. Som de Cerff. Enabling datadriven provenance in NetCDF, via OGC WPS operations. climate analysis services use case. *In AGU Fall Meeting Abstracts*, 2016.

[4] R. Filgueira, A. Krause, M. Atkinson, I. Klampanos, A. Spinuso, and S. Sanchez-Exposito. dispel4py: An agile framework for data-intensive escience. In 11th IEEE International Conference on e-Science, pages 454-464, Aug 2015.

[5] S-ProvFlow, https://github.com/KNMI/s-provenance, Accessed March 2018.

[6] CLIPC project. http://www.clipc.eu. Accessed: February 2018.

[7] CMIP5 - Coupled Model Intercomparison. https://cmip.llnl.gov/cmip5. Accessed: March 2018.