Ease access to climate simulations for researchers:

IS-ENES climate4impact

Christian Pagé

CECI

Université de Toulouse, CNRS, Cerfacs

Toulouse, France

christian.page@cerfacs.fr

Wim Som de Cerff

KNMI

Koninklijk Nederlands Meteorologisch Instituut

De Bilt, Netherlands

wim.som.de.cerff@knmi.nl

Maarten Plieger

KNMI

Koninklijk Nederlands Meteorologisch Instituut

De Bilt, Netherlands

maarten.plieger@knmi.nl

Alessandro Spinuso

KNMI

Koninklijk Nederlands Meteorologisch Instituut

De Bilt, Netherlands

alessandro.spinuso@knmi.nl

Xavier Pivan

France

xavier.pivan.ds@gmail.com

Index Terms—climate, conceptualisation, data-driven science, scientific workflows, provenance, metadata

I. INTRODUCTION

Proper climate data access is getting to another dimension for climate scientific researchers, as data volumes are increasing very fast. The current Coupled Model Intercomparison Experiments 6 (CMIP6 [1]) contains a very large number of experiments, climate models (with increased spatial and temporal resolutions), greenhouse gas scenarios, etc. The CMIP is a collaborative framework designed to improve knowledge of climate change. It was first organized in 1995 by the Working Group on Coupled Modelling (WGCM) of the World Climate Research Programmes (WCRP). It is developed in phases to foster the climate model improvements but also to support national and international assessments of climate change.

The current trend of much larger data volumes result in difficulties to process and analyze needed data for research and applications. This is especially true for end users and researchers that have limited computing and bandwidth resources. The whole climate data archive is expected to reach a volume of an estimated 30 Pb within the next few years and up to potentially 2000 Pb later in the decade. On-demand data processing solutions as close as possible to the data storage are emerging and are absolutely needed, thanks to newly developed standards, provenance and infrastructures.

In Europe several initiatives are taking place to support scientific on-demand data analytics at the European scale. However, they use heterogeneous systems and often incompatible authentication and semantics. But they offer the huge potential of interoperability, as for example the DARE e-science platform (http://project-dare.eu), designed for efficient and traceable development of complex experiments and domain-specific services on the Cloud. Also, the IS-ENES (https://is.enes.org) consortium has developed a platform to ease access to climate data for the climate impact community (C4I: https://climate4impact.eu). The platform is based on existing standards such as Web Processing Service (WPS). It is important to enable generic access and cross-domain interoperability, as well as providing compliance and integration with the future European Open Science Cloud (EOSC) platform (https://www.eosc-portal.eu).

II. MOTIVATIONS: CURRENT SITUATION

The IS-ENES C4I platform is being developed since 2009 ([2] and [3]), currently within the IS-ENES3 H2020 European...
project. It has evolved from a web portal presenting impact-specific national Use Cases with documentation and guidance to a platform of standardized re-usable services and building blocks. The evolution of the C4I platform has been user driven, according to user consultations throughout the years as well as suggestions and feedback from users and scientific conferences ([4]). It became obvious that the proper targeted users of the platform are scientific researchers, either being from the climate domain or other scientific domains using climate simulations.

Currently, in the climate research community, the users are just beginning to get away from a download-locally-then-analyze type of workflow. This is critical as this is no longer a possible workflow with the current data volumes needed for doing proper scientific research. There has been some national repositories set up with a subset of the most commonly used datasets, but even this approach is difficult to sustain and is not scalable.

III. MOVING TOWARD THE INTEGRATION OF HETEROGENEOUS SYSTEMS

One of the most pressing needs is for remote data processing, e.g. as near as the data storage as possible. It is not necessary that processing systems be co-located with the data servers, it can also be on intermediate systems and platforms located with a very high bandwidth and capacity compared to what users have available at their location. Several solutions are available, such as those offered by EUDAT (https://www.eudat.eu), or those of the future DARE Platform ([5]), the European Open Science Cloud (EOSC), commercial clouds (Amazon, etc.), national infrastructures, etc. An integration approach is only possible with proper standardization, sufficient metadata, lineage and provenance. This is very important when doing delegations of calculations and processing. The challenge is to enable users to transparently use remote computing resources that are available to them. But this is also a strong requirement.

IV. FUTURE STEPS

It has been shown that climate science researchers need support in getting access to data. It is getting critical with respect to the large data volume increase of the datasets. The IS-ENES C4I Platform is helping those users, but the current architecture and implementation is not scalable. Currently the C4I architecture is being completely rethought and redesigned, not only the organization of the underlying architecture, but also on how users interact with the data. One of the idea is to go away from a POSIX file-based approach to a more database-like approach (data-centric).

The current development plan and new architecture sketch of C4I will be shown: the ways it will be redesigned with respect to user interaction, as well as its new underlying architecture. Major improvements on how users interact with data, such as an improved search interface, a better scalable access to processing with transparent access to external resources, etc., will be presented. In the redesigned version, C4I users will use the front-end specific wizards (along with guidance) to trigger the execution of data processing, seamlessly and transparently. Crucial aspects related to metadata, lineage and provenance will also be discussed and exposed, especially the importance of sufficient and relevant metadata in the current context (workflow composition, dataset locations, mapping with different systems, etc.)

Further ahead, C4I will also propagate transparently the authentication and authorization of the user where needed, such as credentials to access data and/or computing resources and eventually storage as well. Security aspects (and those are critical for the success of this kind of architecture) will need to be assessed. However, it will not be discussed here.

V. ABBREVIATIONS AND ACRONYMS

- C4I: IS-ENES climate4impact Platform
- DARE: Delivering Agile Research Excellence EU-funded H2020 Project
- ENES: European Network for Earth System modelling
- EOSC: European Open Science Cloud
- EU H2020: European Commission Horizon 2020 Research and Innovation programme
- EUDAT CDI: EUDAT Collaborative Data Infrastructure
- IS-ENES: Infrastructure of the European Network for Earth System modelling
- WPS: Web Processing Service

ACKNOWLEDGEMENTS

IS-ENES Climate4impact builds on a succession of prior EU projects: IS-ENES phase 1 and 2, CLIPC, EUDAT phase 1 and 2, ESIWACE, on nationally funded projects as well as in-kind contributions. The authors also thank the developers and systems teams that have delivered and tested the platform over the years and that are not cited explicitly here. The IS-ENES3 project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824084 while the DARE project is under grant agreement No 777413.

REFERENCES

[1] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, K. E. Taylor, Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization, Geoscientific Model Development 9 (5) (2016) 1937–1958. doi:10.5194/gmd-9-1937-2016.
URL: https://www.geosci-model-dev.net/9/1937/2016/
[2] C. Pagè, M. Vega, A. Cofino, C. Jack, M. Plieger, Review CLIMATE4IMPACT services and objectives, EC FP7-IS-ENES2 Project Milestone (2016).
[3] C. Pagè, M. Plieger, W. S. de Cerff, C. Jack, A. Cofino, M. Vega, M. Kolax, L. Barring, O. B. Christensen, M. Mieatea., Report on derived products in CLIMATE4IMPACT, EC H2020-DARE Project Deliverable (2017).
[4] C. Dénandreis, C. Pagè, P. Bracconnet, L. Barring, E. Bucchignani, W. S. de Cerff, R. Hutjes, S. Joussaume, C. Mares, S. Planton, M. Plieger, Towards a dedicated impact portal to bridge the gap between the impact and climate communities : Lessons from use cases, Climatic Change 125 (3) (2014) 333–347. doi:10.1007/s10584-014-1139-7.
URL: https://doi.org/10.1007/s10584-014-1139-7
[5] C. Pagè, A. Spinuso., Requirements and Test Cases I, EC H2020-DARE Project Deliverable (2018).