Leveraging and Easing End Users' Climate Data Access by Interfacing Infrastructures

Christian Page[1], Wim Som de Cerff[2], Alessandro Spinuso[2], Maarten Plieger[2], Iraklis Klampanos[3], Malcolm Atkinson[4], and Vangelis Karkaletsis[5]

[1]CERFACS European Centre for Research and Advanced Training in Scientific Computation
[2]Royal Netherlands Meteorological Institute
[3]National Centre for Scientific Research "Demokritos"
[4]University of Edinburgh
[5]National Center for Scientific Research Demokritos

November 21, 2022

Abstract

End Users of Climate data have nowadays to struggle with accessing the data they need for their research because of the rapid increase in data volumes. The whole climate data archive is expected to reach a volume of 30 Pb in 2018 and up to 2000 Pb in 2022 (estimated). On-demand data processing solutions as close as possible to the data storage are emerging, 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. 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 (ISO and OGC), such as WPS (Web Processing Service). DARE will integrate services from the EUDAT CDI, enabling generic access and cross-domain interoperability, as well as providing compliance and integration with the future EOSC platform. The DARE platform will use containerization technologies, so that it can be easily deployed on heterogeneous architectures. A scientific pilot has been designed within the DARE project for the ENES community (climate domain). The objectives are to enable delegation of on-demand computational-intensive calculations to the DARE platform, from the IS-ENES C4I interface, seamlessly. The DARE architecture and the solutions being implemented will be presented, along with the generic and agile approach taken to implement the pilot.

Leveraging and Easing End Users' Climate Data Access by Interfacing Infrastructures

Christian Pagé (Cerfacs, France)

christian.page@cerfacs.fr http://linkedin.com/in/pagechristian https://www.researchgate.net/profile/Christian_Page http://cerfacs.fr/~page

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

Wim Som de Cerff & Maarten Plieger & Alessandro Spinuso (KNMI, Netherlands)

Iraklis Angelos Klampanos & Vangelis Karkaletsis (NCSR Demokritos, Greece) Malcolm Atkinson (University of Edinburgh, UK)

IN33E-0889

Abstract 340841

I New Challenges for Science

Alias	City	Node Type	Software Version	Software Release
	Huntsville	Compute Idp Index Data	v0.0.0-devel	Flatbush
	Paris	Compute Idp Index Data	v1.5.0-brower_park-release-2-gfad439c-master	brower_park
a.gov	Sylmar	Compute Idp Index Data	v1.6.0-3-g39599da-devel	brower_park
	Livermore	Compute Idp Index Data	v1.6.1-bushwick_myrtle-release-devel	bushwick_myrtle
	Livermore	Compute Idp Index Data	v1.5.0-brower_park-release-master	brower_park

FIG 1: Earth System Grid Federation (ESGF) of Servers

- The Earth System Grid Federation is a Peer2Peer International Network of Servers
- Currently, the *Download & Analyze* workflow is no longer sustainable
- European Data Nodes are managed by IS-ENES (Infrastructure for the European Network of Earth System Modelling)

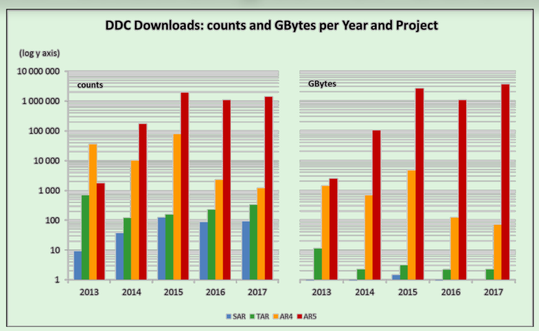

FIG 2: DKRZ ESGF Data Node Data Volume (GB) Downloaded for SAR, TAR, AR4 and AR5

- Increase in Data Production is significant
- Moving data will be soon much more expensive than computing time
- A change of paradigm

	CMIP5	CMIP6	CMIP7
Year	2012	2017	2022
Power factor	1	30	1000
Npp	200	357	647
Resolution [km]	100	56	31
Number of mesh points [millions]	3,2	18,1	108,4
Ensemble size	120	214	388
Number of variables	800	1068	1439
Interval of 3-dimensional output (hours)	6	4	3
Years simulated	90000	120170	161898
Storage density	0,00002	0,00002	0,00002
Distributed Archive Size (Pb)	**3,19**	**86,05**	**2260,20**

FIG 3: Climate Model Intercomparison Projects (CMIP) Archive Size (PB)

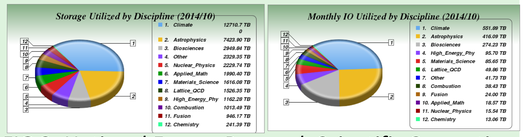

FIG 3: National Energy Research Scientific Computing Center (NERSC) Storage and I/O by Discipline

- Climate has large needs for Storage and I/O
- Large and heterogeneous communities of users

II C4I IS-ENES Tailored UI

- https://climate4impact.eu
- Developed and managed by IS-ENES
- Platform for researchers to explore climate data and perform analysis
- Not only UI, but also Standard Services (WPS, WCS, ...)
- Connects to ESGF web services
- Tailored for end-users
- Supports on-demand data processing and statistical downscaling
- Now containerized version
  - docker & docker-compose
- Compatible with CMIP6, CORDEX, etc.

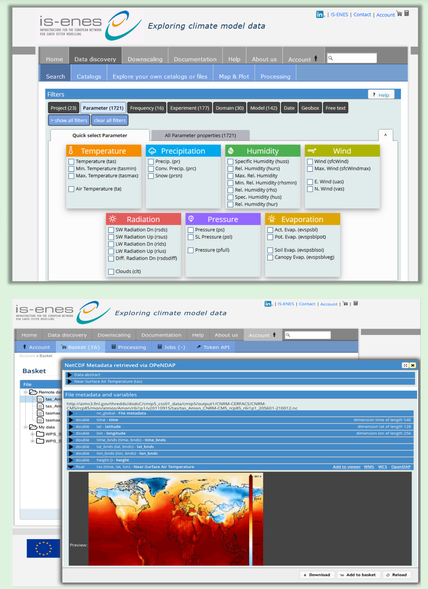

FIG 4: C4I Faceted Search and Interface

III DARE Platform

- dispel4py: High-level streaming dataflow specification API/library. Automated collection of lineage.
- S-ProvFlow: Reproducibility as a Service. Based on W3C-PROV.
- Exareme: Large-scale dataflow processing on the cloud.
- Semagrow: Semantics and linked-data.
- BigDataEurope: Big Data Analytics

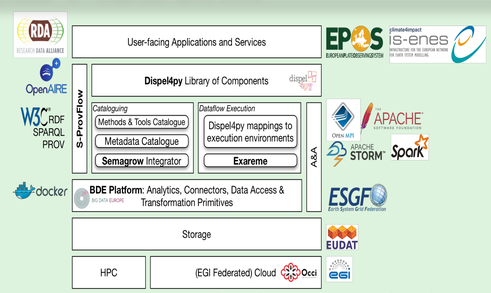

FIG 5: DARE provisional architecture stack indicating related platforms, infrastructures and technologies

IV Mapping Communities' Needs

- User Stories Approach
  - Bottom-Up
- Translating a User Story to
  - Feature
  - Capability
- Leads to specification of Components

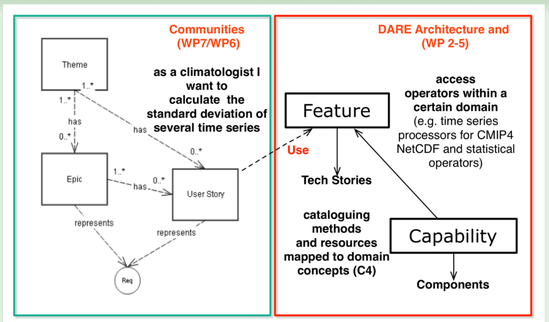

FIG 6: Mapping User Stories to Technical Architecture

V IS-ENES Climate Domain Pilot

- Supports a very large number of Use Cases
- Integration of several components
- icclim Climate Indices backend
- Interfaces to external sources
- Major DARE contributions
  - Provenance/Lineage
  - Seamless integration to high-performance data analytics platform

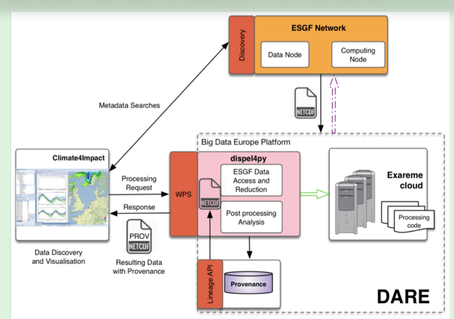

FIG 7: IS-ENES Climate Domain Pilot Schematics

VI Tracking Provenance & Lineage

- Deployed using docker-compose
- Data captured in a document-oriented database based on JSON representation (MongoDB)
- S-ProvFlow collection of methods exposed through OpenApi2.0 API abstraction layer
- Visualization Tool: Provenance Relationships

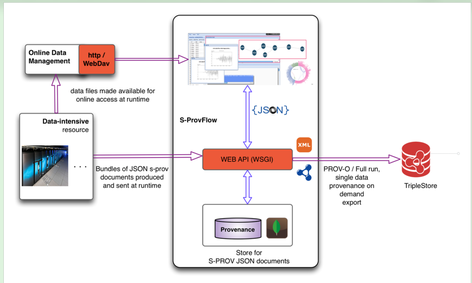

FIG 8: Schematic architecture exploiting the S-ProvFlow system for acquisition, visualisation, data access and provenance export services

VII Future Work

- Finalize the DARE Architecture (Dec 2018)
- Implement dispel4py Processing Elements for both Domain Pilots
- Develop the OGC WPS interface to dispel4py
- Evaluate where Exareme can be useful within some Domain Pilots Use Cases
- Write interfaces from dispel4py to/from ESGF
  - Both for the data nodes as well as for the computing nodes
- Interface to external services (see Fig. 5, e.g. EUDAT B2 Services)
- Develop the Provenance Model and Semantics for Climate Data and Processing
- Perform first execution tests on the DARE testbed
- Collaborate with H2020/IS-ENES3 to:
  - Design the C4I UI to deal with complete workflows

Get the Poster here!!

Some References

H2020/DARE Project: http://project-dare.eu

IS-ENES C4I: https://climate4impact.eu

EUDAT CDI: https://www.eudat.eu/eudat-collaborative-data-infrastructure-cdi

icclim: https://github.com/cerfacs-globc/icclim

The DARE project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 777413

https://tinyurl.com/AGU-DARE