Climate Data Access: Re-thinking our Data Analysis Workflows

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 & Ernst de Vreede (KNMI, Netherlands) Niels Drost (Netherlands eScience Center)

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

Download Poster!

I New Challenges for Science

	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 1: Climate Model Intercomparison Projects (CMIP) Archive Size (PB)

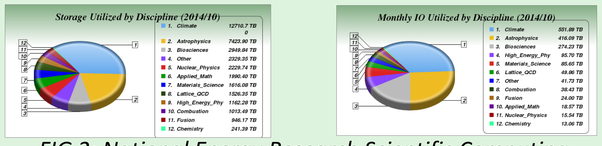


FIG 2: National Energy Research Scientific Computing Center (NERSC) Storage and I/O by Discipline

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

II Common Users' Needs

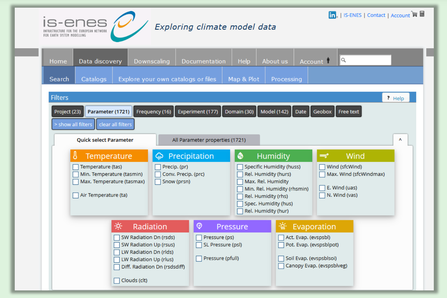
- Guidance/tools for data and scenarios subsetting: selecting a subset of representative climate scenarios
- Lower significantly the total data size to download
- Calculate as much as possible remotely
- Reformat/Repackage the data into easier formats
- Have full Provenance and Lineage information
- Proper Metadata description, especially for derived data
- Variety of Access Interfaces for adoption: GUI, OGC, REST APIs, Jupyter Notebooks, ...

Take Home Messages

1. **Climate Datasets are getting too large for comfort**
2. **Current Data Analysis Workflow is no longer possible: processing delegation is needed**
3. **Heterogeneous Processing Backends are available: Clouds, EUDAT, DARE, EOSC, ESGF CWT, etc.**
4. **It is possible to hide underlying complexity**
5. **Provenance & Lineage is essential**
6. **Precise (Metadata-)Standards are mandatory**

III climate4impact 1.0

- https://climate4impact.eu
- Developed and managed by IS-ENES since 2010
- Platform for researchers to explore climate data and perform analysis
- Not only UI, but also Standard Services (WPS, WCS, ...)
- Tailored for end-users
- Supports on-demand data processing and statistical downscaling
- Now containerized version
  - docker
  - docker-compose



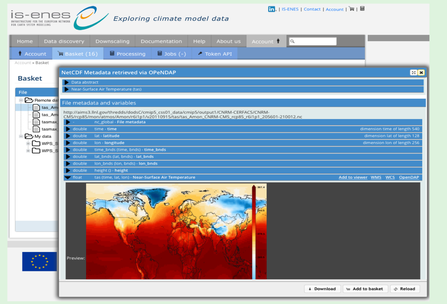


FIG 3: C4I Faceted Search and Interface

IV Needs for Evolution

- Modernize the front-end: Drupal ➤ ReactJS.
- Separate the front-end and the processing back-end.
- Ease the installation process using docker and docker-compose
- Improve the ergonomics and the responsiveness of the Search Interface.
- Improve the users' experience by having a more streamlined workflow.
  - Hide complexity from the users
  - Move from a file to a more data-oriented approach
- Incorporate more complete Provenance & Lineage
- Have users control and build their own workflows, and share them with the community

V Connecting External Resources

- Connect to external computing resources, such as:
  - Clouds (Private and Public)
  - e-infrastructures:
    - EUDAT CDI
    - European Science Cloud (EOSC)
    - DARE Platform
    - ESGF Computing Nodes (CWT)

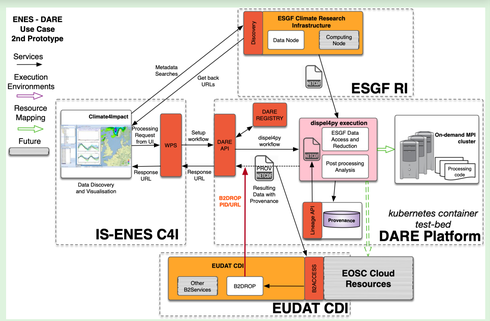


FIG 4: Prototype Integration with DARE Platform, EUDAT CDI and EOSC

VI Upcoming Work: C4I 2.0

- Evaluate the possibility of using a micro-services approach
  - Python/Flask-based as much as possible
  - Reuse old java code from version 1.0 if needed
- Refactor whole documentation and guidance using S3 Bucket for content storage
- Implement a Vocabulary Service
- Restructure and optimize icclim backend processing open-source software
- Evaluate possible C4I/WPS Proxy
- Services with external APIs
  - MyCollection (Basket)
  - OGC-WPS using Birdhouse Framework
  - ...
- Support for Climate Infrastructure (ESGF) Computing Nodes for Pre-Processing Data

EUDAT CDI: https://www.eudat.eu/eudat-cdi climate4impact: https://climate4impact.eu/ ESGF CWT: https://github.com/ESGF/esgf-compute-api DARE Platform: http://project-dare.eu

The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°824084, and the DARE project under N°777413

http://bit.ly/climate-workflows