DARE Delivering Agile Research Excellence on European E-Infrastructures

GA No 777413

EGU General Assembly 2020, Sharing Geoscience Online, 4-8 May, 2020

Integrating e-infrastructures for remote climate data processing

Christian Pagé

Research Engineer / Climate Research Domain

CERFACS CECI, Toulouse, France

Alessandro Spinuso, Maarten Plieger, Wim Som de Cerff, *KNMI, Netherlands*

Iraklis Klampanos, Malcolm Atkinson, Vangelis Karkaletsis, *Demokritos, Greece*

EGU2020-4658
https://doi.org/10.5194/egusphere-egu2020-4658

DEMOKRITOS INGV THE UNIVERSITY of EDINBURGH CERFACS KIT Karlsruher Institut für Technologie Koninklijk Nederlands Meteorologisch Instituut grnet Fraunhofer ATHENA

New Challenges for Science

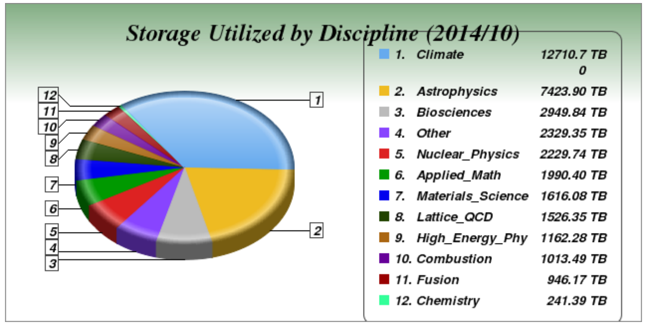
Climate Researchers and Climate Change Community End Users

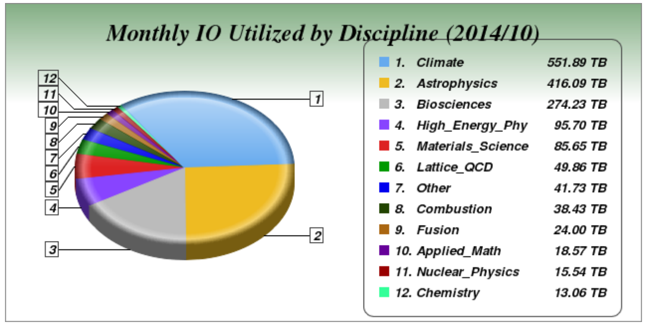
- Large needs for Storage and I/O
- Heterogeneous communities of users

Common Needs

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

	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**



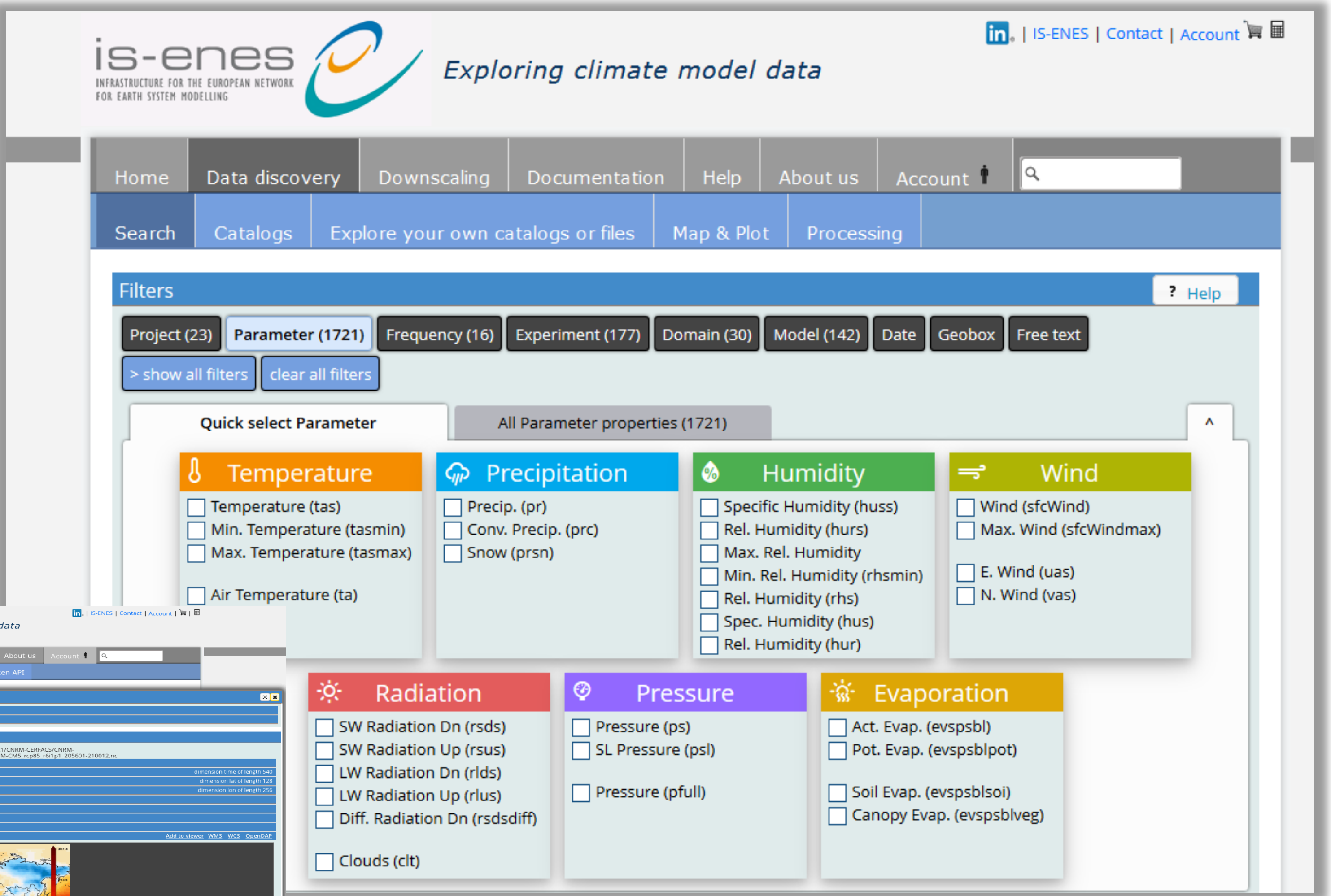


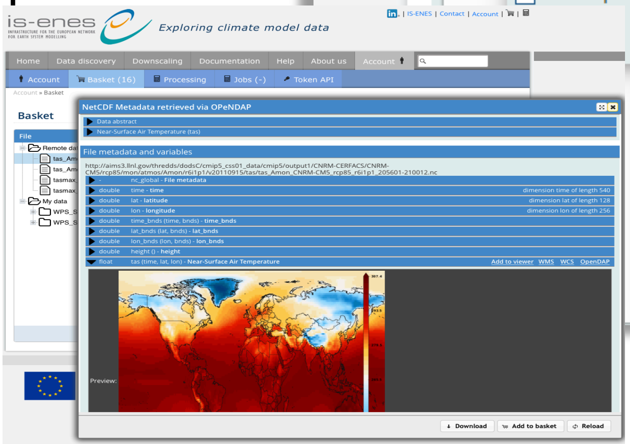
National Energy Research Scientific Computing Center (NERSC) Storage and I/O by Discipline

IS-ENES climate4impact (C4I)

https://climate4impact.eu

- Developed and managed by IS-ENES since 2010
- Not only UI, but also Services (WPS, WCS,..)
- Tailored for end-users
- Supports on-demand data processing
- Now containerized version
  - docker
  - docker-compose

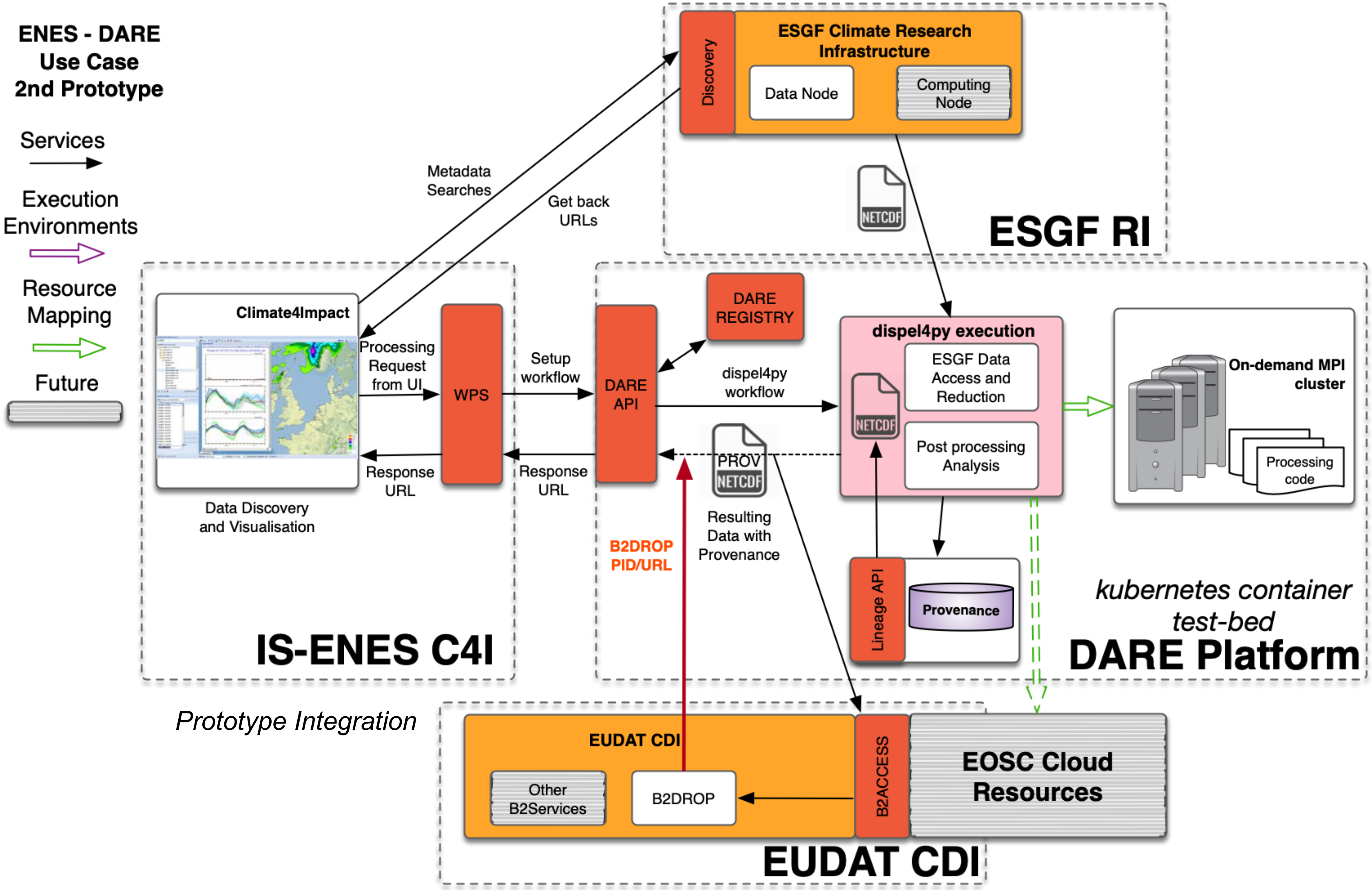




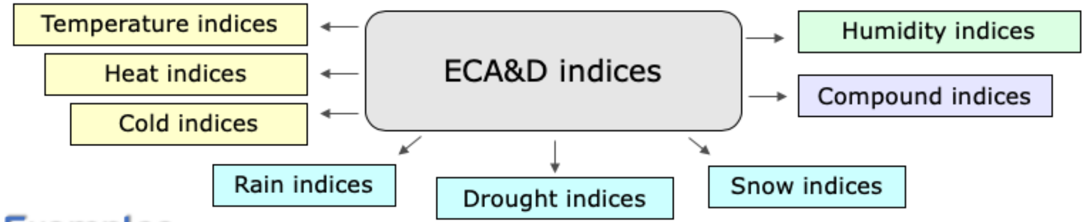
C4I Faceted Search and Interface

DARE: Execution Platform and Enabling Connections to External Resources

- Connect to external computing/storage resources:
  - Clouds (AWS, etc.)
  - e-infrastructures:
    - EUDAT CDI
    - European Science Cloud (EOSC)
    - DARE Platform
    - ESGF Computing Nodes (CWT)



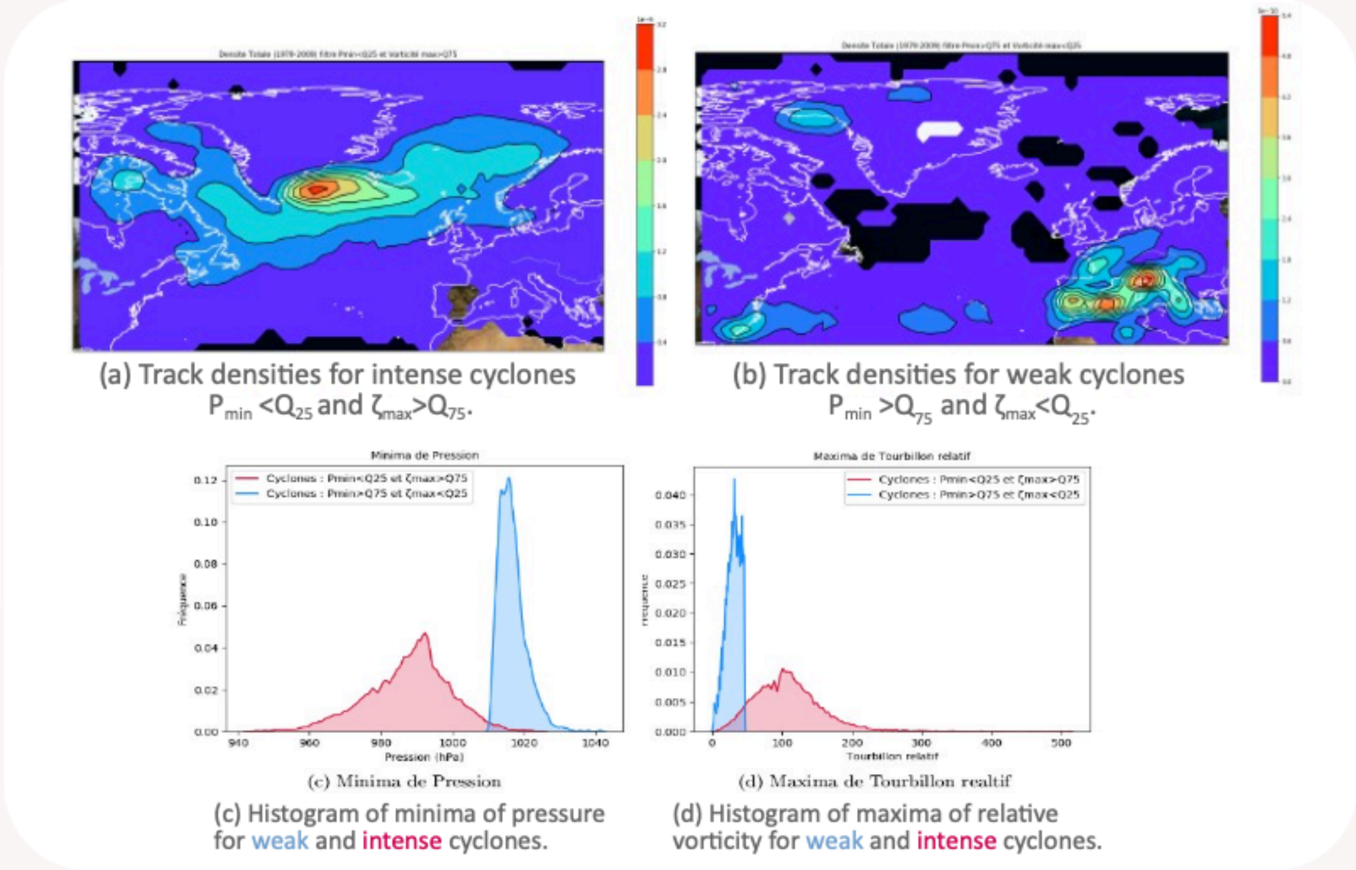
Leveraging Existing Complex Tools



Examples

- Intra-period extreme temperature range [° C] - **ETR**
- Warm days (days with mean temperature > 90th percentile of daily mean temperature) - **TG90p**
- Summer days (days with max temperature > 25 ° C) - **SU**

Calculating climate indices and climate indicators
https://github.com/cerfacs-globc/icclim



Extra-Tropical Cyclone Tracking in Climate Simulations
https://github.com/cerfacs-globc/cyclone_tracking

Discussions material

Take Home 🏡 Messages ☑

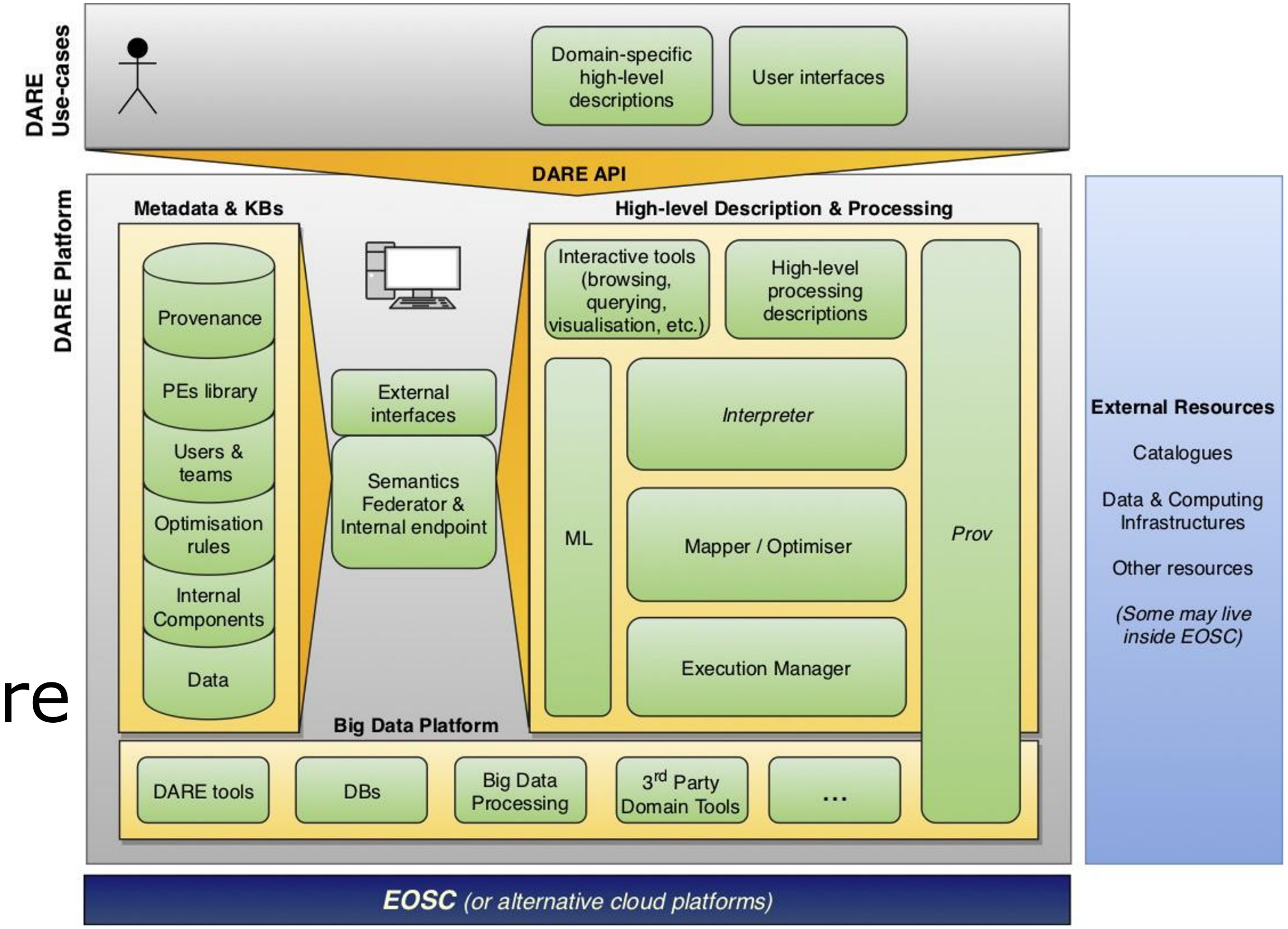
1. Climate Datasets are getting too large for comfort
2. Current Data Analysis Workflow is no longer possible: processing delegation is needed
3. Heterogeneous Processing Backends are available: Clouds, Clusters, EUDAT, EOSC, ESGF CWT, etc.
4. It is possible to hide underlying complexity
5. Provenance & Lineage is essential
6. DARE approach together with IS-ENES, EUDAT, ESGF CWT, can help Scientific Researchers

- DARE Platform
- Cyclone Tracking
- IS-ENES climate4impact
- EUDAT CDI & EOSC

DARE Platform

DARE Platform

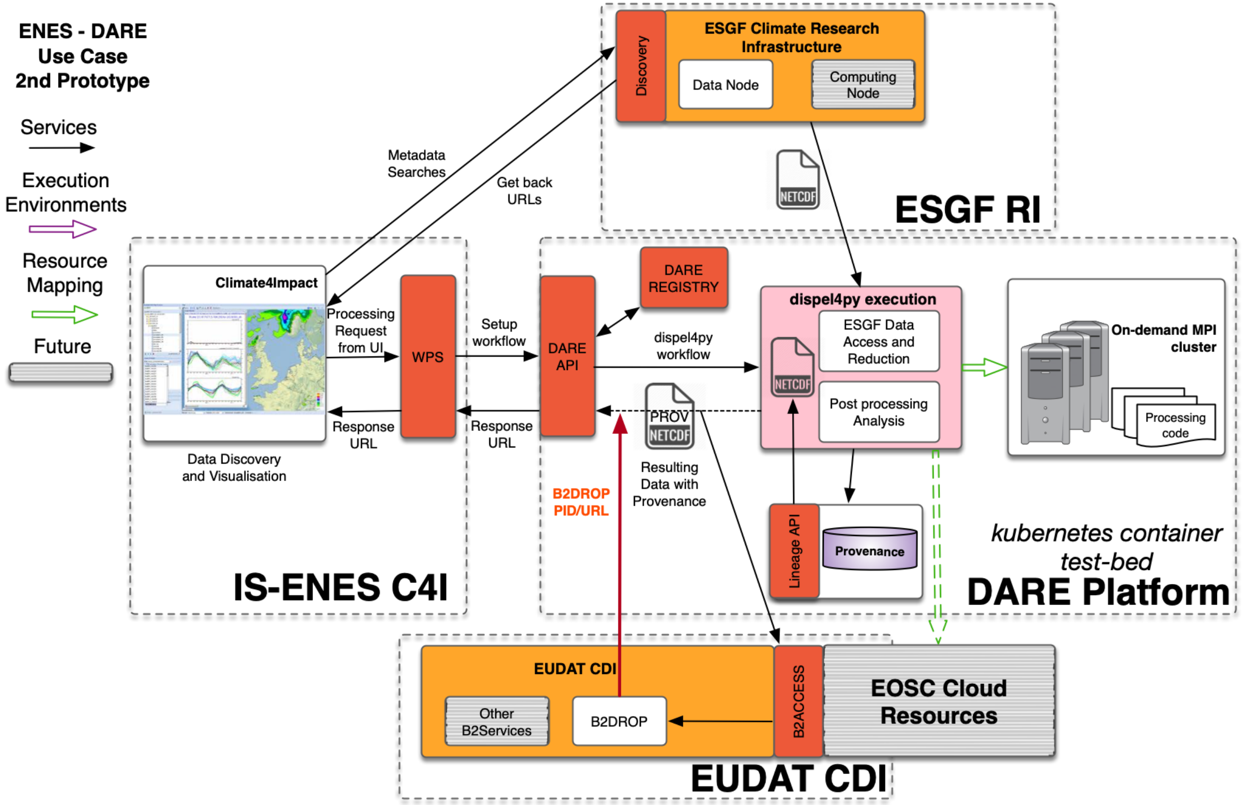
- Composition of services using containers
- Across service communication using exposed REST APIs
- Scalable and flexible due to kubernetes orchestration
- Effortless cloud infrastructure deployment
- Software isolation



DARE Platform

Simplified View of Prototype Architecture

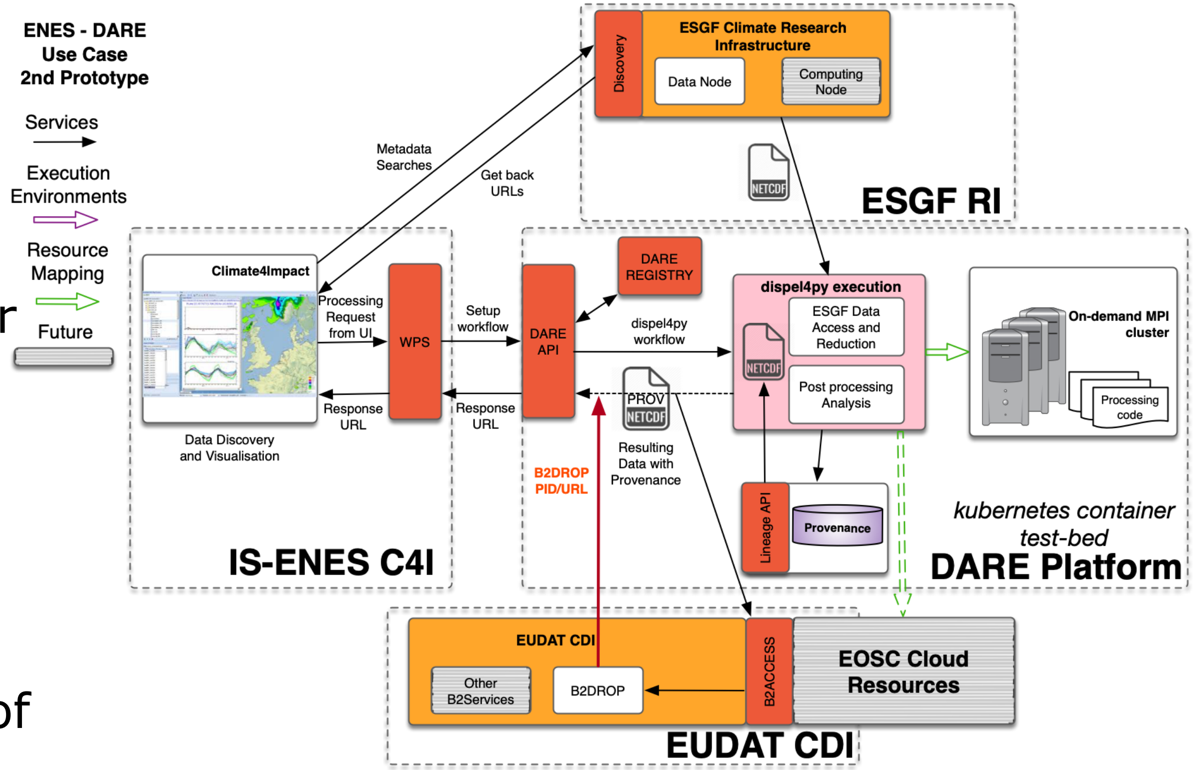
- Interfaces
  - C4I WPS / DARE API
  - DARE API / EUDAT B2DROP
  - C4I DARE Workflow / ESGF Data Nodes
  - C4I DARE Workflow / ESGF Computing Nodes
- Processing Backend
  - icclim software
- Provenance & Lineage
  - Automated by the DARE Platform
  - Community-specific information added



DARE Platform

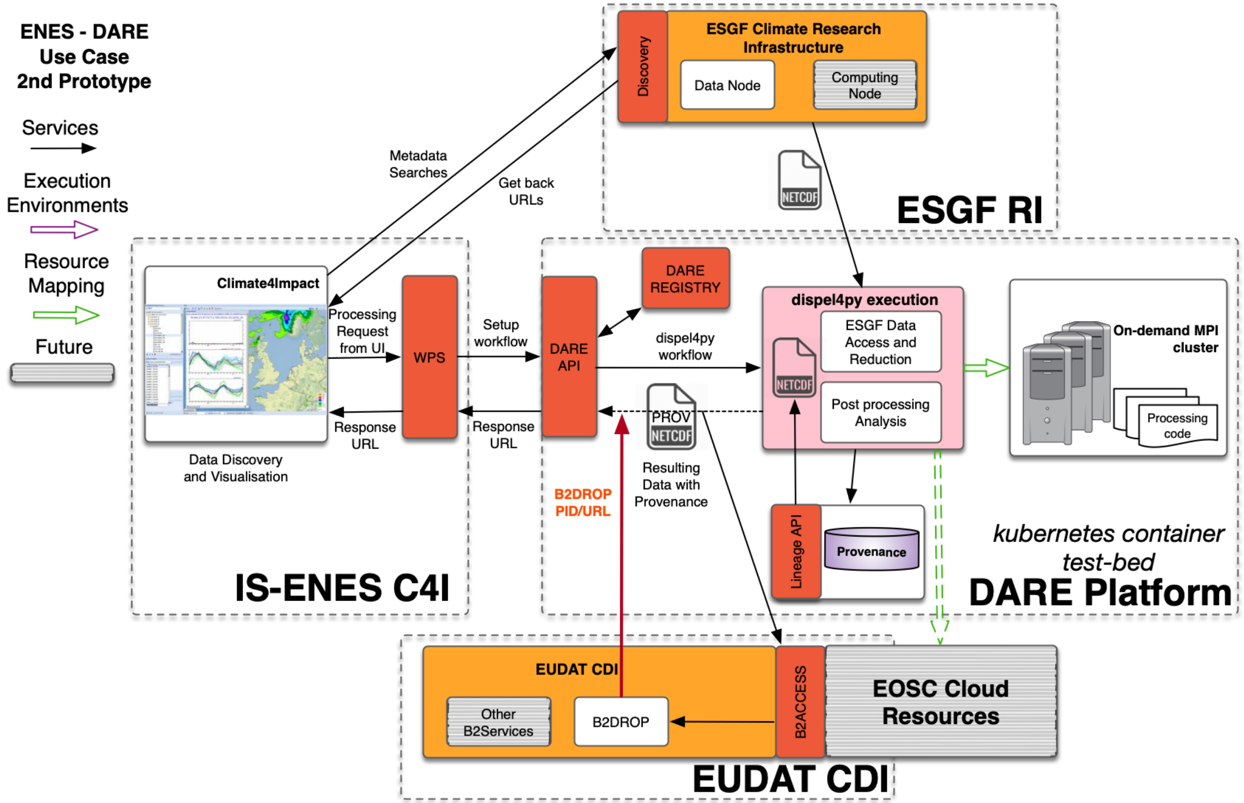
Automation/External Services to Interface and Integrate

- **EUDAT-B2DROP**
  - Store user output
  - Auth B2ACCESS
- **ESGF CWT**
  - Delegation of processing before data transfer
  - Auth API Key
- **ESGF Data Nodes**
  - Retrieve Climate Datasets
- **C4I**
  - Use ESGF Search results to get input URLs of datasets
  - Storage of output and transfer back to C4I
  - Control whole workflow



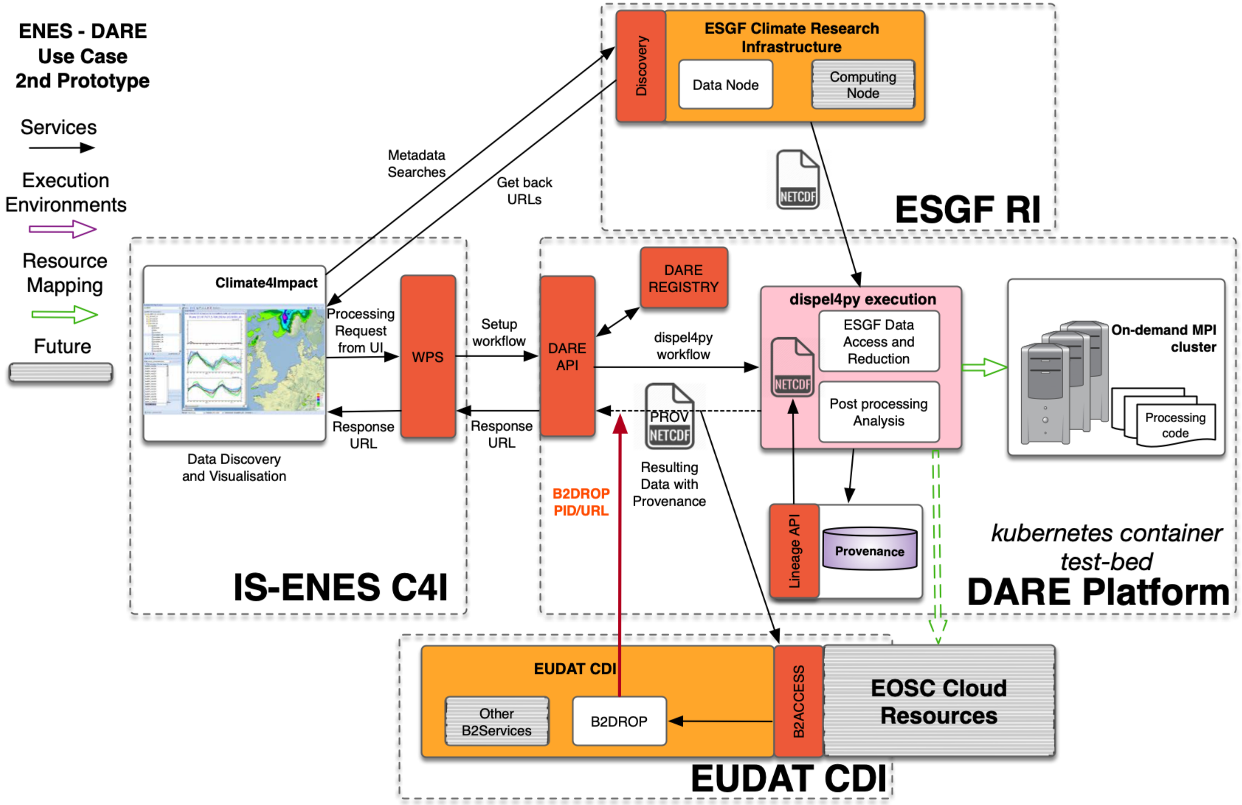
DARE Platform

- Efficient Interfacing e-Infrastructures for Researchers is Challenging
  - Technologies: fast-evolving but many are getting deprecated and obsolete
  - Authentication and Security
  - Scalability in Federated Environments
- DARE Platform
  - Hides complexity and heterogeneity
  - Provides automated Provenance & Lineage
  - Provides EUDAT & EOSC Compatibility



DARE Platform

- IS-ENES C4I
  - User-Driven Interface Development
  - Focus on Data
  - Provides API
  - Ease Access to large climate datasets
- EUDAT CDI
  - Solid Collection of Services for Scientific Data
  - Compatible with EOSC
  - Integrates community-specific aspects



Cyclone Tracking

Haran Émilie[1], Laloue Alice[1], Pourquié Clément[1] supervised by Pagé Christian[2], Sanchez Émilia[2]
[1]IENM2, École Nationale de la Météorologie, Toulouse, France
[2]Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS), Toulouse, France

Abstract

This project is focused on the study of the evolution of cyclone trajectories in a future climate. A tracking code was applied on historical climate simulations and climate projections from the PRIMAVERA project. Tools to analyze the results had to be developed : maps of cyclone track densities, maps of variability of track densities, histograms.

Introduction

Extra-tropical cyclones are low pressure systems in mid-latitudes associated to local maxima of vorticity often leading to strong winds, extreme precipitation and high waves. These cyclones affect livelihoods and infrastructure.
It is therefore important, in climate, especially in the context of climate change to be able to detect them, to follow their trajectories and their evolution, and identify areas most prone to the passage of intense cyclones. This project is part of the research carried out by the Climate Modeling and Global Change team at the CERFACS and of the European PRIMAVERA project [1]. A tool for monitoring the trajectories of extra-tropical (and tropical) cyclones was developed by M. R. Sinclair [4] and improved by the University of Quebec in Montreal (UQAM) and the Meteorological Service of Canada [2]. It was then adapted by C. Pagé at CERFACS to deal with the outputs of climate models and simulations in NetCDF format.

Methods

ERA-Interim
atmospheric reanalysis

1979-2009 *reference*
NetCDF data

PRIMAVERA ARPEGE
CNRM-CM6-1 and
CNRM-CM6-1-HR
historical simulations
(past climate)
1979-2009
NetCDF data

PRIMAVERA ARPEGE
CNRM-CM6-1 and
CNRM-CM6-1-HR
climate projections
(future climate)
2015-2045
NetCDF data

NetCFD data extraction and processing

Outputs of the tracking algorithm

A cyclone is identified as a maximum of relative vorticity at 500hPa associated to a low pressure at sea level within a radius of 150km.

Cyclone tracks are composed of at least 2 points.

```
1 48.326 322.031 0 956.29 12 01 01 1950
1 51.128 322.031 0 967.78 12 02 01 1950
1 53.929 320.625 0 980.20 12 03 01 1950
1 56.731 320.625 0 990.30 12 04 01 1950
2 65.136 350.156 240 1004.48 12 05 01 1950
2 65.136 12.656 0 991.28 12 06 01 1950
```

Sinclair, M. R. et al., 2004 [4]

Calculating cyclone tracks

Tools
Track densities and Variability

A matrix of density in which each coefficient corresponds to a spatial grid point is created.
A coefficient is implemented by one each time a cyclone track crosses the grid point it is associated with.

To calculate the variability, the standard error of several matrix of density is calculated.

Analysis of the outputs of the tracking algorithm with the tools developed

Results and Discussion

1. Track densities of extra-tropical cyclones in the ERA-Interim reanalysis

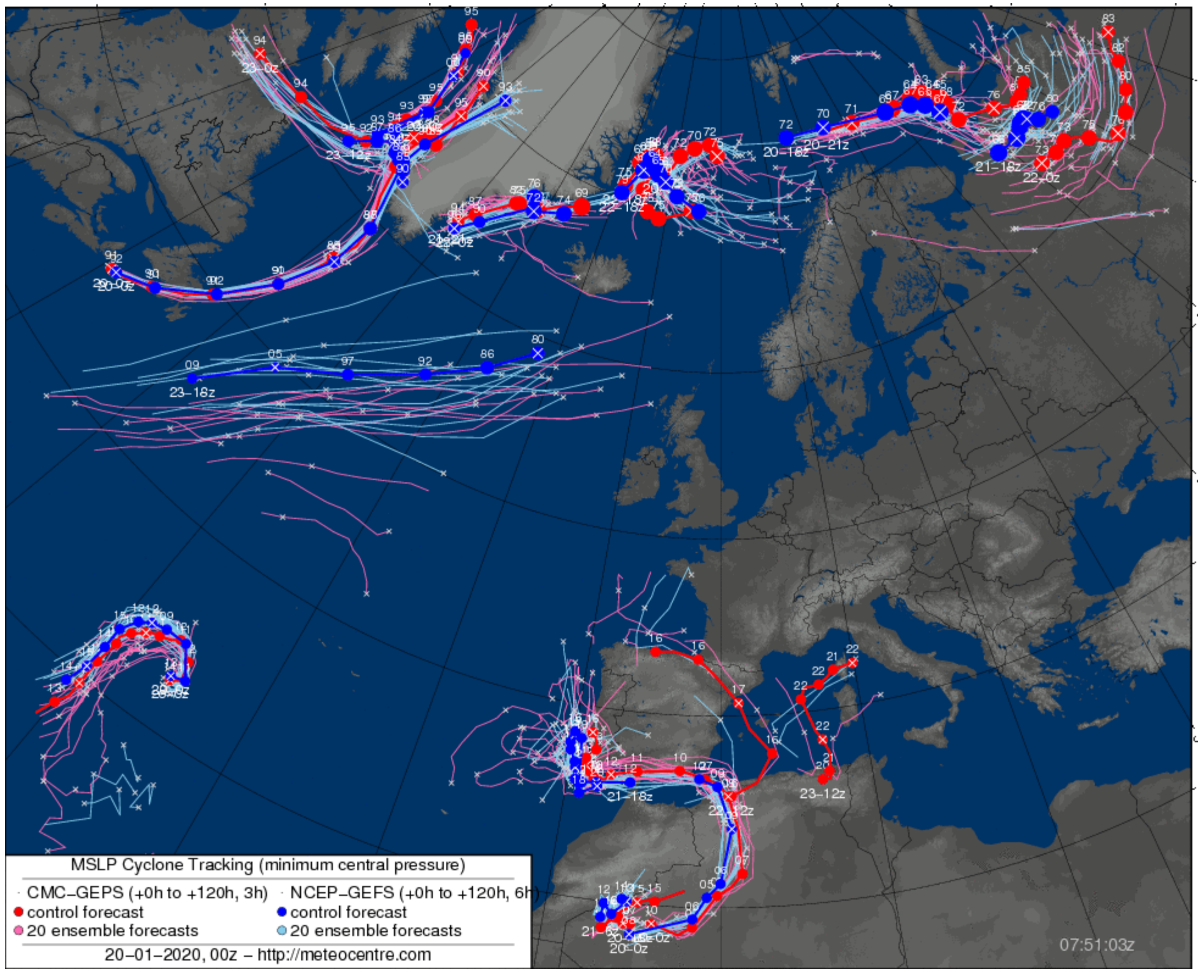
The tracking algorithm was first applied on data from the ERA-Interim reanalysis [3] - on which the new

3. Track densities of extra-tropical cyclones in the climate projections (2015 – 2045)

ENES Use Case: Storm Tracking

- The new ENES Use Case is about tracking storms in climate scenarios
- Initially it is for extra-tropical cyclones but the same implementation can be used for tropical cyclones

The Storm Tracking



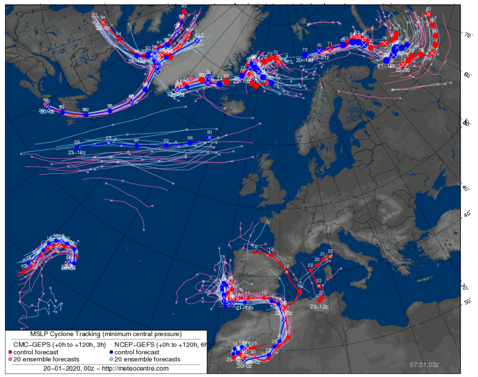
Storm Tracking Use Case

- Based on the Sinclair Methodology (1997)
- Group Effort

Canadian Meteorological Centre

Université du Québec à Montréal

High impact Lab Québec region Environment Canada



- Rabah Aider
- Jean-François Caron
- Louis-Philippe Caron
- Corina Costea
- Ronald Frenette
- Stéphane Gagnon
- Philippe Gachon
- Rares Gheti
- Anne-Marie Leduc
- Philippe Martin,
- Milka Radojevik
- Christian Saad
- Mark R Sinclair
- Katja Winger
- Ayrton Zadra
- Christian Pagé

Storm Tracking Use Case

Rational - Unified code

Tropical center diagnostic simple tracking method

Pressure/Vorticity centers elaborated tracking method

} Several versions of the code

Tropical center and transition diagnostic

AND

Pressure/Vorticity centers

Both using same elaborated tracking method

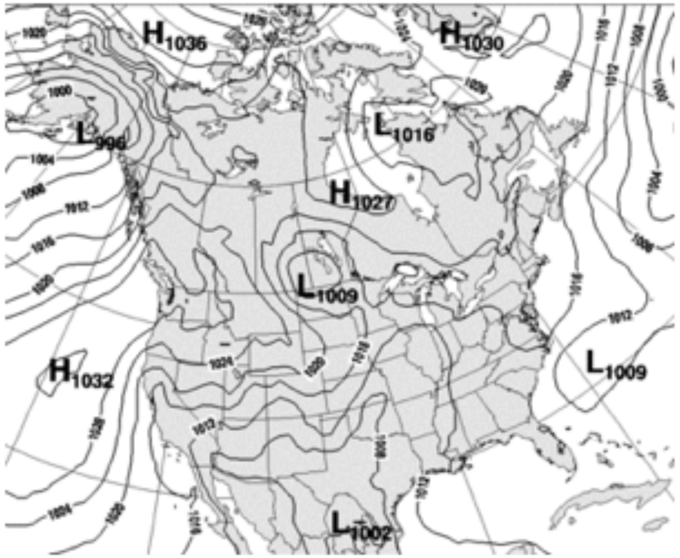
} One program, one code. Maintenance, implementation easier

Storm Tracking Use Case

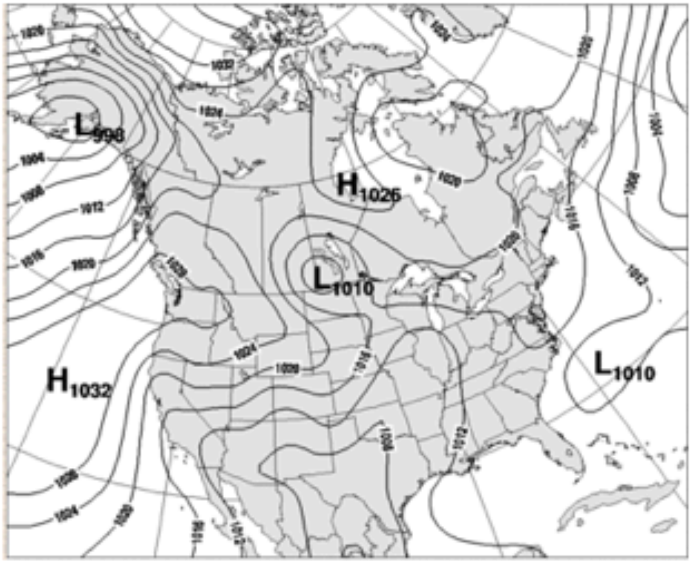
Step 1. Preprocessing the data

For pressure center tracking

Raw NWP output



300km Cressman filter to remove small scale details



Storm Tracking Use Case

Step 2a. Identifying minima

- Uses cubic splines between grid points for interpolation for a more precise location.

- Surface pressure field unreliable over higher terrain:
    - Filter with vorticy threshold.
    - Threshold varies with terrain height

Storm Tracking Use Case

Step 2b. Identifying tropical centers

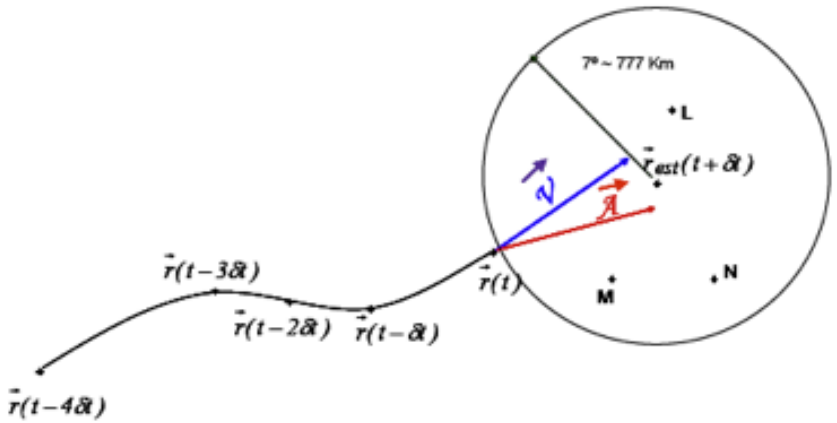
- Pressure minimum
- A 300km Cressman filter is applied to the 850 hPa vorticity field
  - A vorticity max higher the 2.5x10-5 s-1 is present within 150km of pressure center.
- 850-250 hPa thickness higher then 925 dam within 150km.
  - Indicating the presence of a deep and vertical warm center
- A 10m wind speed higher then 22 kts within 225km
- A baroclinicity in the low levels
  - Measured in terms of the asymmetry of the 900-600 hPa thickness*.

Use the Safir-Simpson surface wind thresholds to classify
•Tropical Depression : V10m < 34 kts
•Tropical Storm : 34 < V10m < 64 kts
•Hurricane : V10m > 64 kts

*Sinclair, M. R., 2004: *Extratropical Transition of Southwest Pacific Tropical Cyclones. Part II: Midlatitude Circulation Characterisitcs, Mon. Wea. Rev., 132, p. 2149.*

Storm Tracking Use Case

Step 3. Tracking the centers



- r(t*) is the position of the first point of the trajectory and r(t) is the position of the current point.
- wm is a weight function depending on the number of analysis per day (2,4,8,..)

$$\vec{r}_{est}(t+\delta t) = \vec{r}(t) + w_m \cdot \vec{A} + (1 - w_m) \cdot \vec{V} \cdot \delta t$$

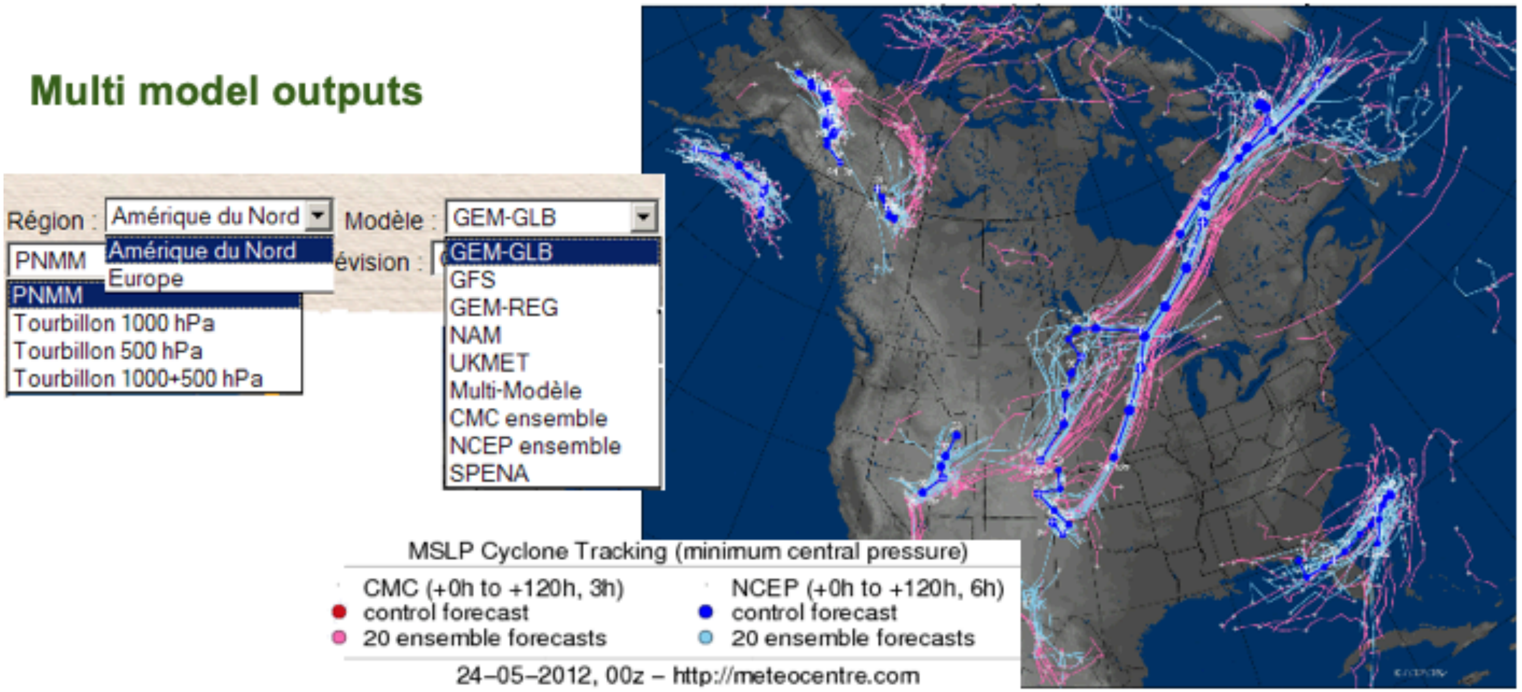
Mark R Sinclair's method

$$\vec{A} = \frac{\vec{r}(t) - \vec{r}(t - 4\delta t)}{4}$$

Storm Tracking Use Case

Step 4. Outputs

Multi model outputs

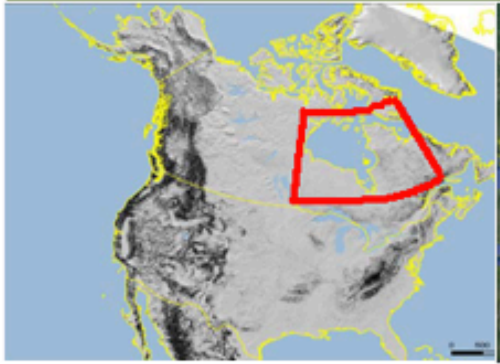


http://meteocentre.com/tracking/index_e.html

From: Jean - François Caron – UK Met

Storm Tracking Use Case

Application - Storm impacts study



Université du Québec à Montréal

CENTRE ESCER

Environment Canada Meteorological Service of Canada

Environnement Canada Service météorologique du Canada

OURANOS

Transports Québec

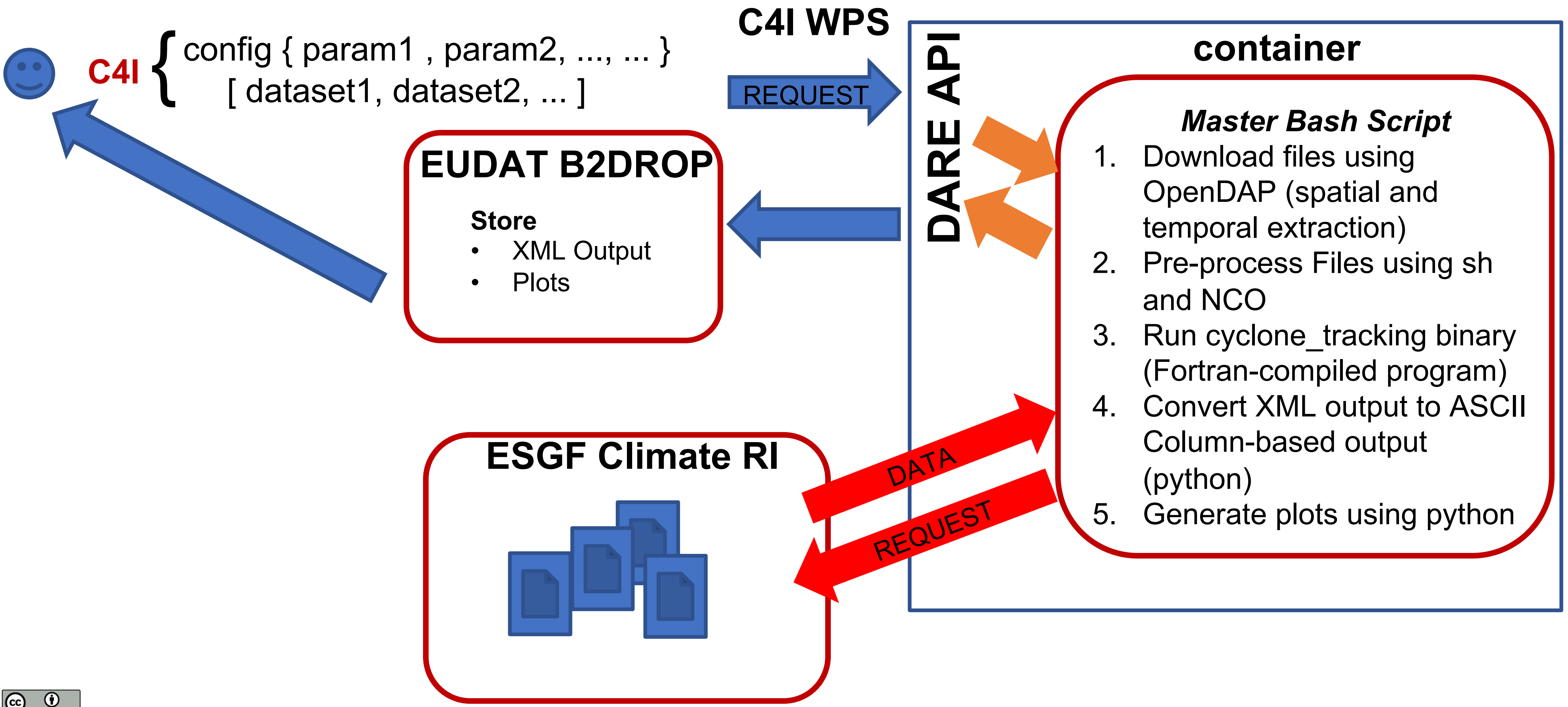
- To improve our knowledge of weather systems affecting urban and surrounding zone
  - Hudson Bay (MTQ project) and southern Québec
  - track density, storm duration, mean circulation, mean vorticity and wind 1000 hPa

- To analyze the links between storms and hazardous events (extremes, high impact)
  - 2m temperatures, precipitations and 10m wind

- To understand the impact of these extremes on population health and infrastructures.

- To predict future changes in the storm climatology and theirs effects on surface extremes
  - cerca 2050

Storm Tracking Use Case Workflow

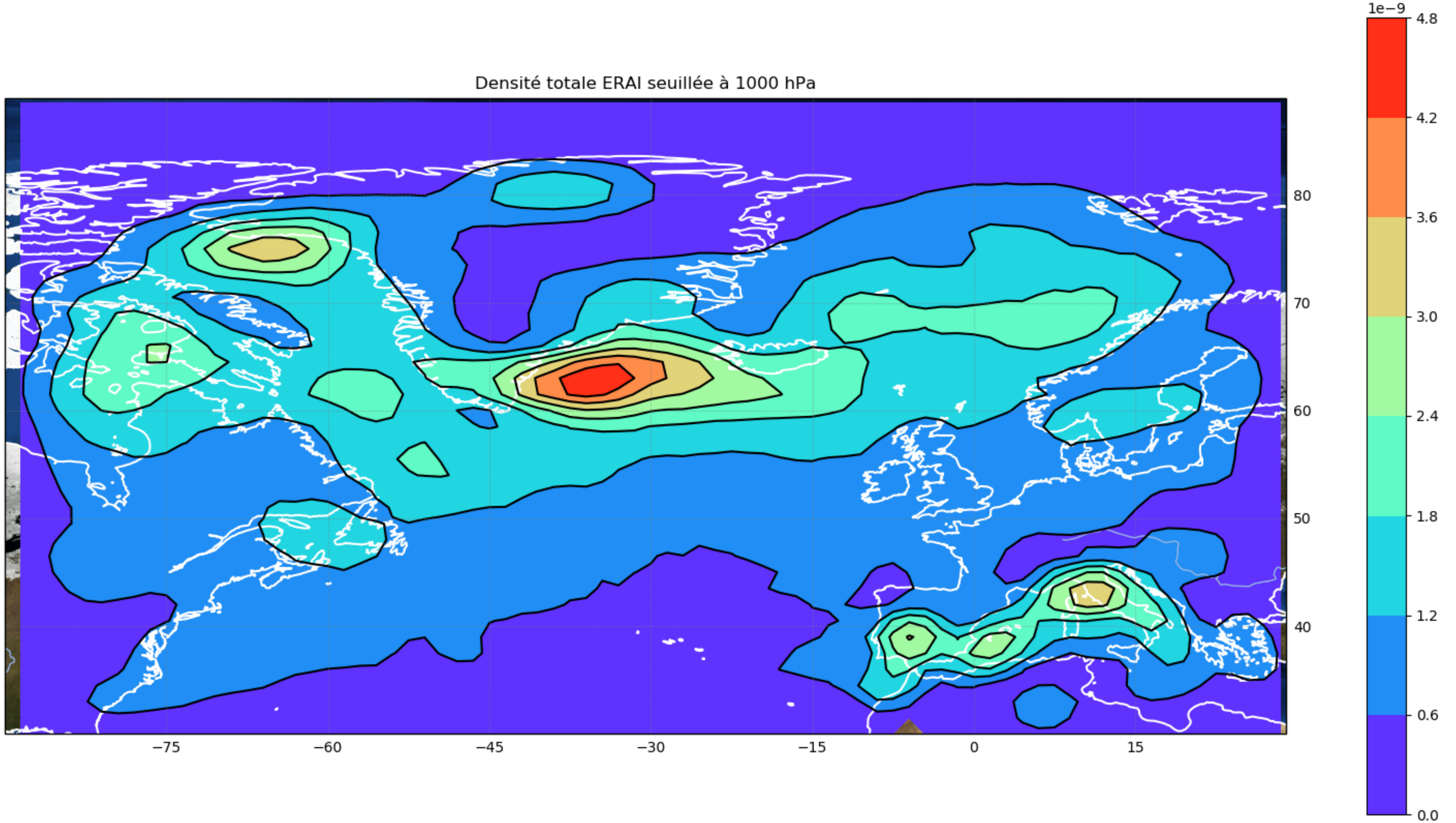


Storm Tracking Use Case Workflow Current status

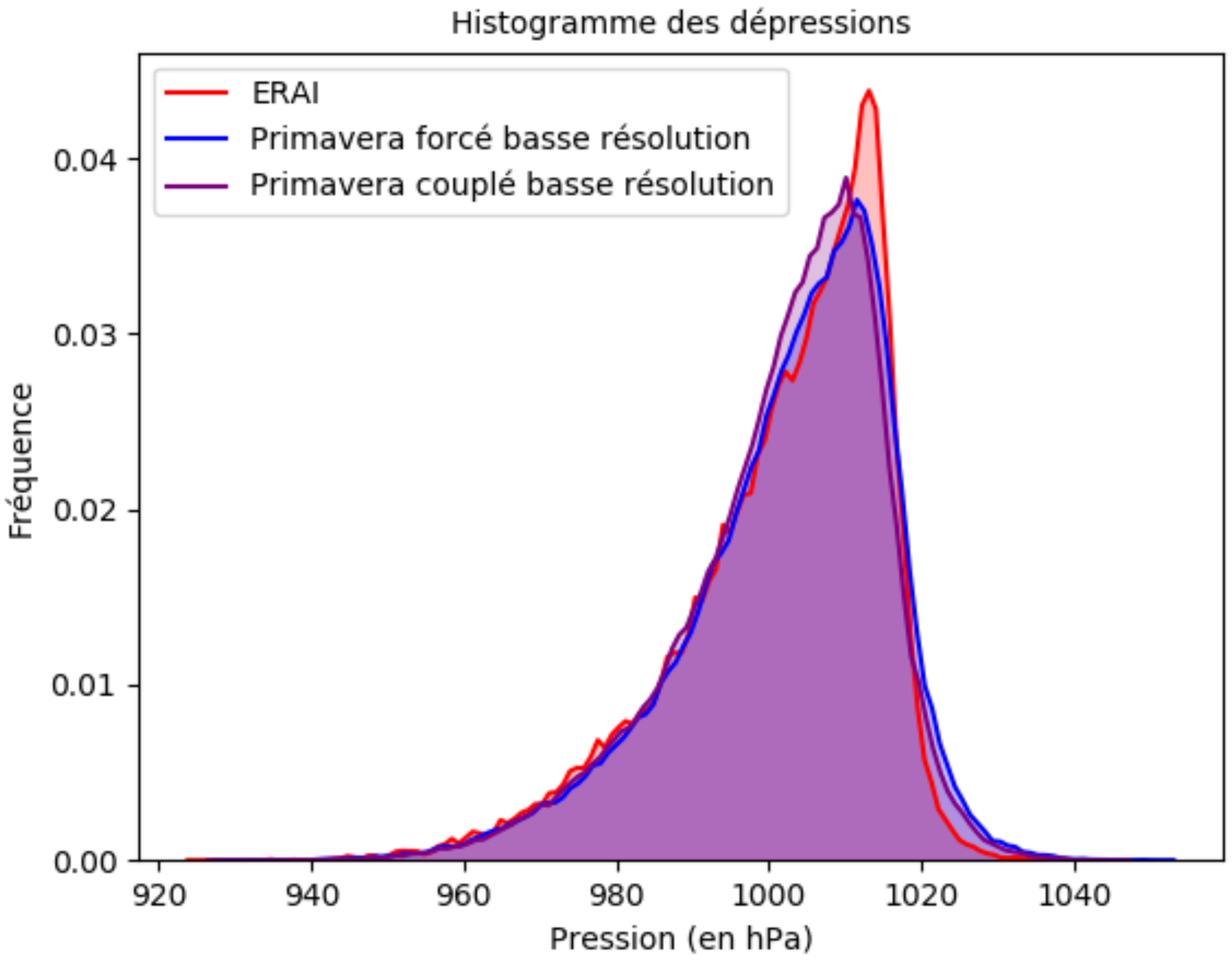
tracking_master.sh

1. Input to the script: JSON configuration file and Input data files URLs
2. Create directory structure
3. ***get_files.sh*** : Download global sample climate scenario files using wget and hard-coded B2DROP URLs
4. Loop over each time period, as climate scenario files are chunked into several files over a long time period (e.g. 5-year, 10-year, etc.)
    5. ***extract_data.sh*** : using NCO NetCDF tools
        1. Extract spatial area and time period
        2. Extract specific vertical levels for 3D variables (Geopotential@1000; Winds@500; ...)
        3. Rename coordinates and variables to standard names
        4. Remove extra vertical coordinates having 1 element
        5. Concatenate all single variables into a single file, except land-sea mask and orography
    6. ***make_tracks.abs*** : execute algorithm (compiled fortran) using appropriate command-line arguments :
        1. 3 input files: main file, land-sea mask, orography
        2. Output prefix filename (for XML output and NetCDF output)
        3. JSON Configuration file
        4. Variable name to use to retrieve grid parameters, dimensions and time steps
    7. ***tracking_xml2ascii.py*** : Convert XML output to CSV
    8. Create warmstart.txt file using storm centers of last day in time period
6. Calculate storm density. Create several plot types using python.

Storm Tracking Use Case Workflow



Storm Tracking Use Case Workflow



IS-ENES climate4impact

DARE

Climate Data Distribution

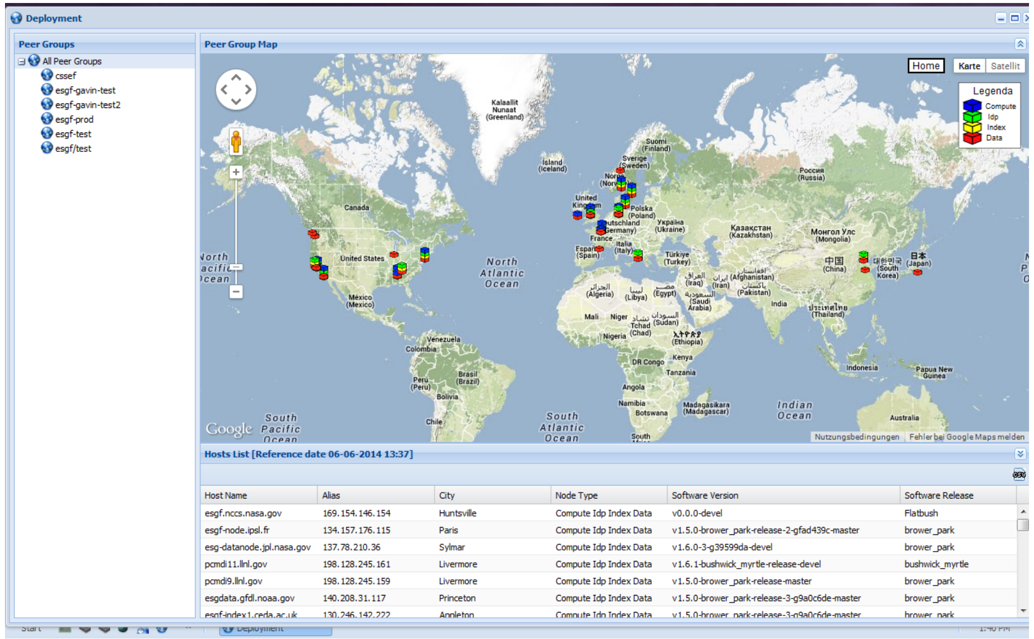
ESGF Data Nodes 2019

- 31 worldwide
- 18 in Europe (17 institutions) (coordinated by IS-ENES)

is-enes INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

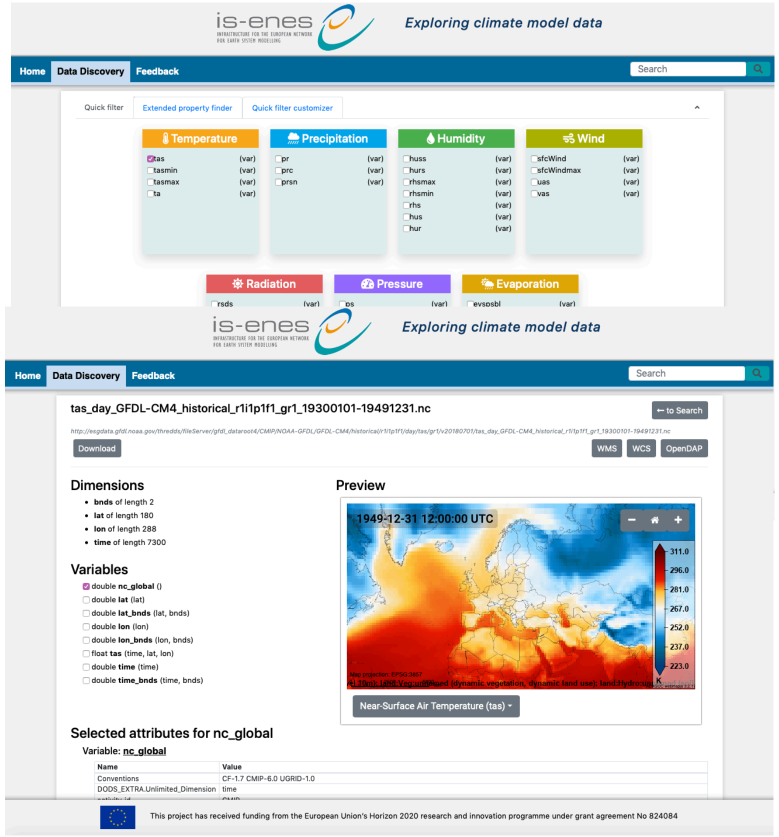
e-infrastructure

Climate Information Portal



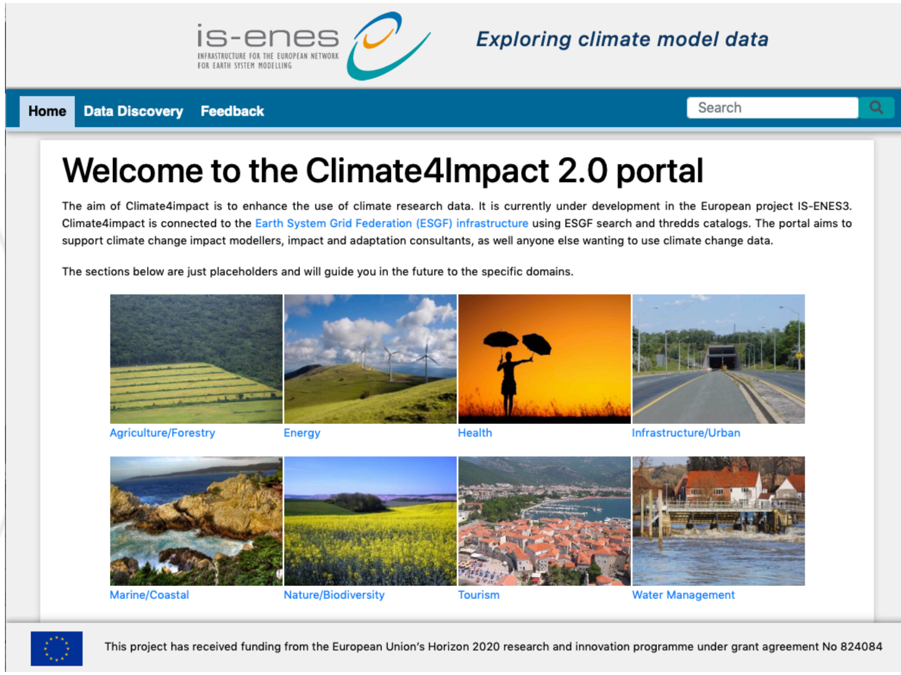
IS-ENES CDI climate4impact

- Tailored for end-users
- Supports on-demand data processing



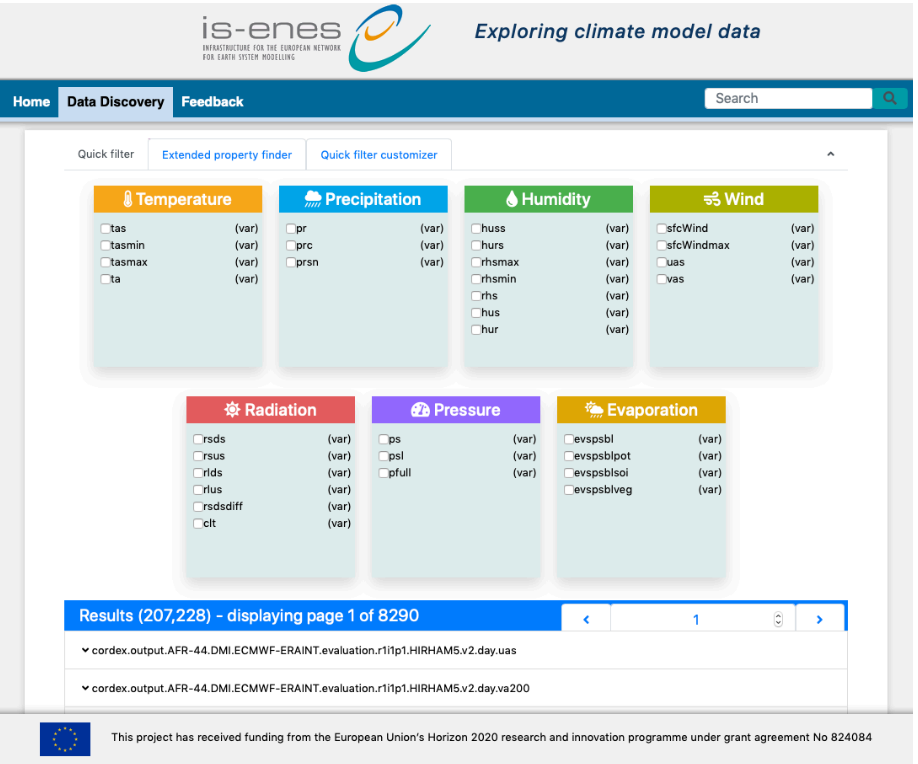
What is the climate4impact portal?

- Platform for researchers to explore climate data and perform analysis
- Connects to ESGF web services
  - Search, Catalog Support, Security
  - Several projects and experiments
- Visualization via ADAGUC Software
  - Visualization system using Web Map Services
  - Web Coverage Services for data transformation
- Analysis using (Py)WPS to perform calculations
  - icclim open-source software for on-demand climate indices calculation
  - Data sub-selection
  - Personal store for processing results
- In production
  - Deployed in the cloud
  - Is one of the official CMIP6 dissemination portals

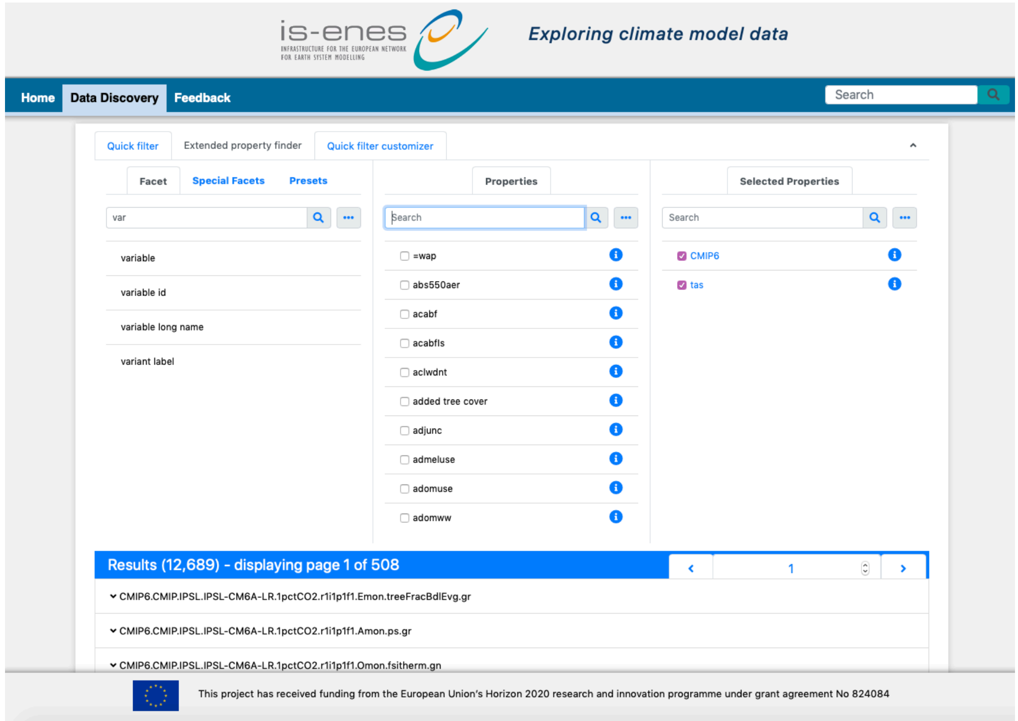


Research data lifecycle

Web based faceted search

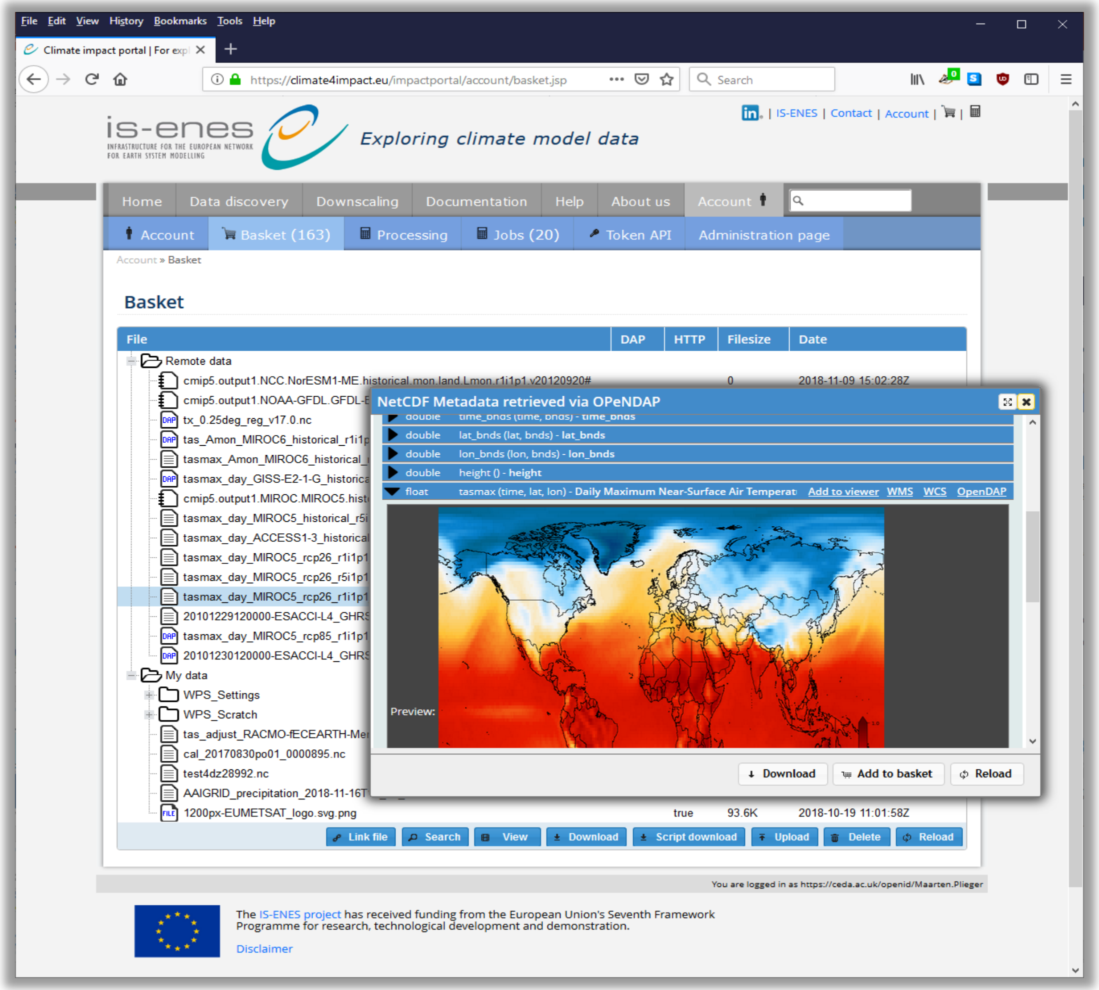


- Drill down search results
- Tooltips for acronyms
- Quick select menus, configurable
- ES-DOC integration
- Preview of data
- Save Search Parameters
- Export search list to CSV



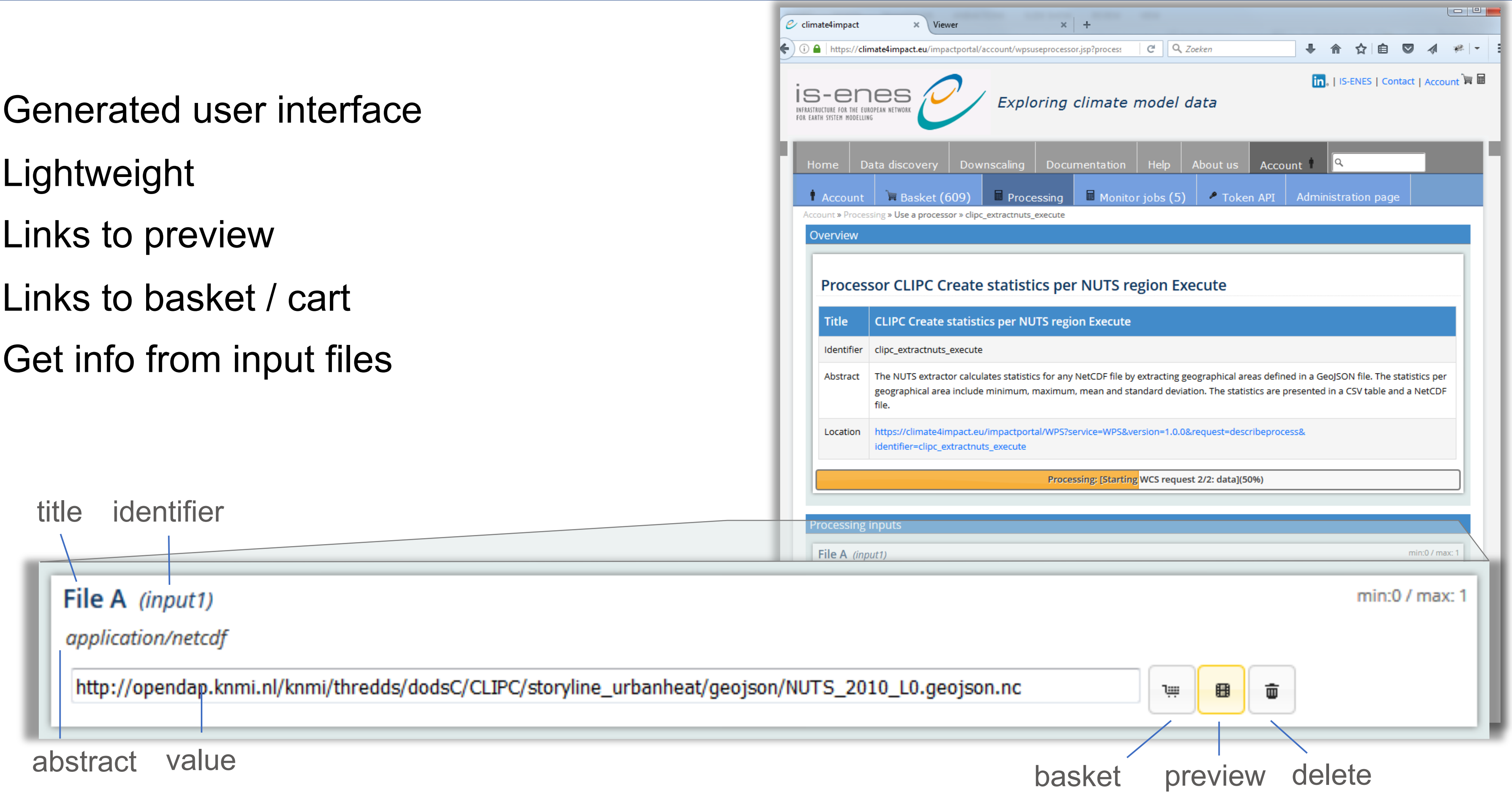
Personal User Space

- By default the basket contains:
  - “Remote data” for links
  - “My data” for your own data
- Script based download allows to select and download multiple files
- The basket allows for uploading your own files
  - Can be used in processing or visualization
  - NetCDF, CSV, GeoJSON, PNG
- Share your data located in your basket with others

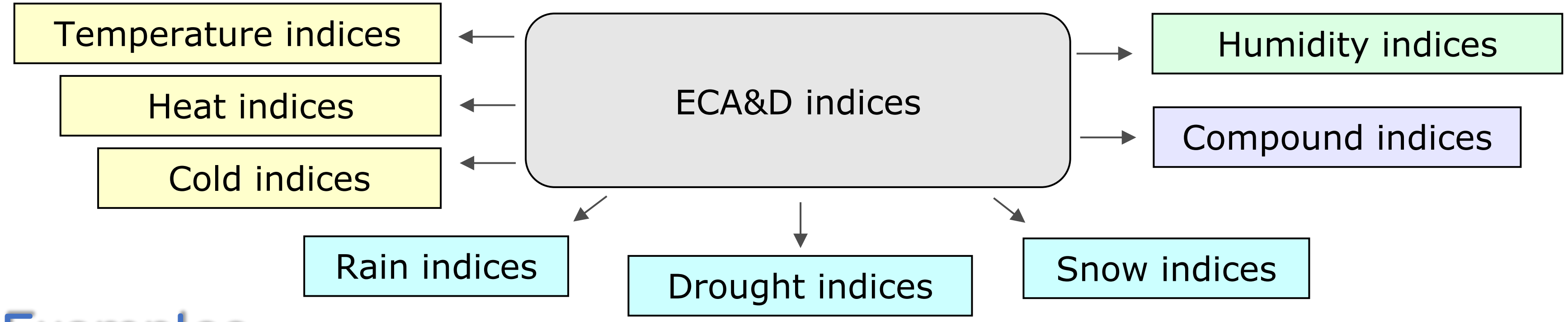


Web processing interface for data analysis

- Generated user interface
- Lightweight
- Links to preview
- Links to basket / cart
- Get info from input files



Climate Indices using icclim



Examples

- Intra-period extreme temperature range [° C] - **ETR**
- Warm days (days with mean temperature > 90th percentile of daily mean temperature) - **TG90p**
- Summer days (days with max temperature > 25 ° C) - **SU**

- Python code developed at CERFACS, started in September 2013
  - Generic and modular approach, can be reused in other environments
  - C functions called for optimization
- I/O interface is structured for optimal performance, with wrapper functions
- Some percentile-based indices (TG10p, TX10p, TN90p, etc) using bootstrap method

icclim source code and documentation is available via **https://github.com/cerfacs-globc/icclim**

An xarray/dask fork has been done and is now at an alpha stage.

Example: Calculating summer days (SU) 1/3

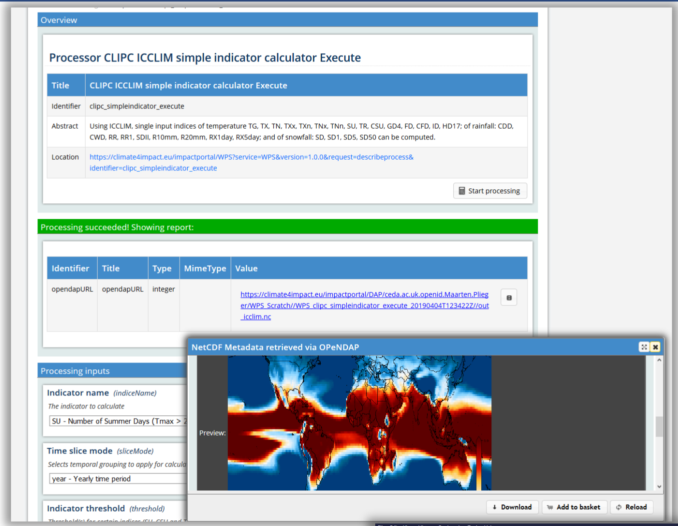
- Calculate number of days where maximum temperature is above 25 degrees per European country, based on experiment RCP 2.6 and climate model MIROC5

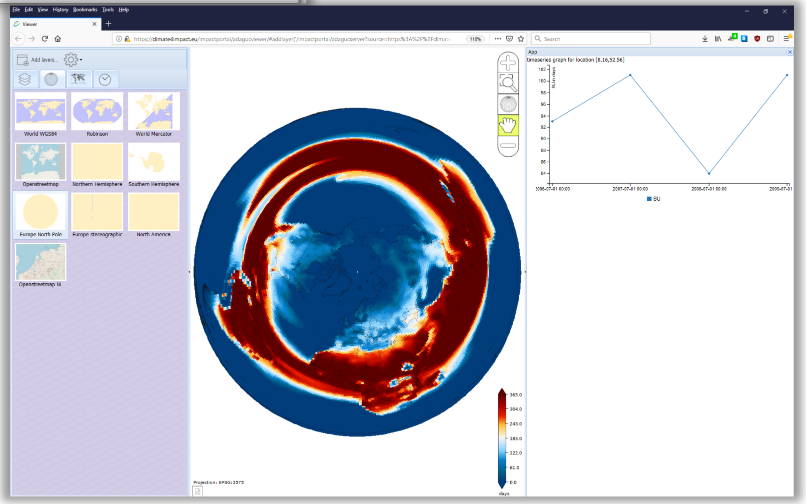
- Sign in
- Go to Search and select:
  1. Project: CMIP5
  2. Parameter: tasmax
  3. Time frequency: daily
  4. Experiment: rcp26
  5. Model: MIROC5,
  6. Ensemble: r1i1p1
  7. Select the latest version

- Select a file from the dataset and add it to your basket

Example: Calculating summer days (SU) 2/3

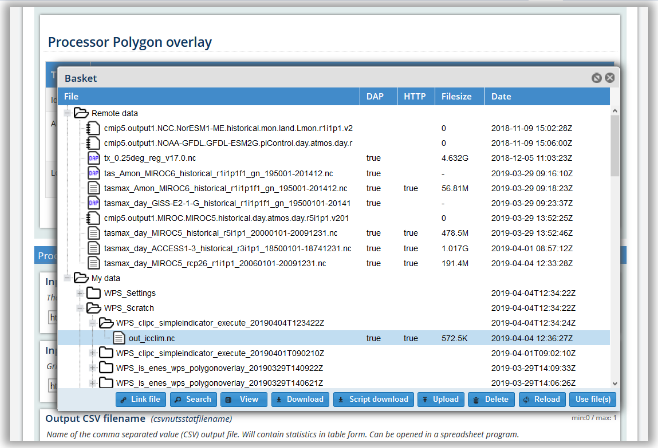
- Go to Processing and select: icclim simple indicator calculations
- Select SU, Summer days. Leave the threshold to 25 degrees Celsius
- Select the file from your basket and click “Start processing”
- Visualize the output



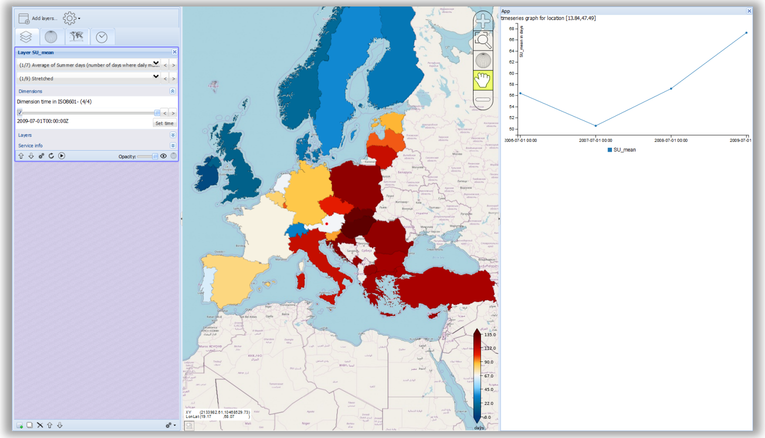


Example: Calculating summer days (SU) 3/3

- Go to Processing and select Polygon overlay
- For “Input File B - Gridded data”, choose the latest result with SU from your basket. This is the most recent folder under WPS_Scratch
- As variable select “SU”, as time range select “*”
- Click “Start processing”



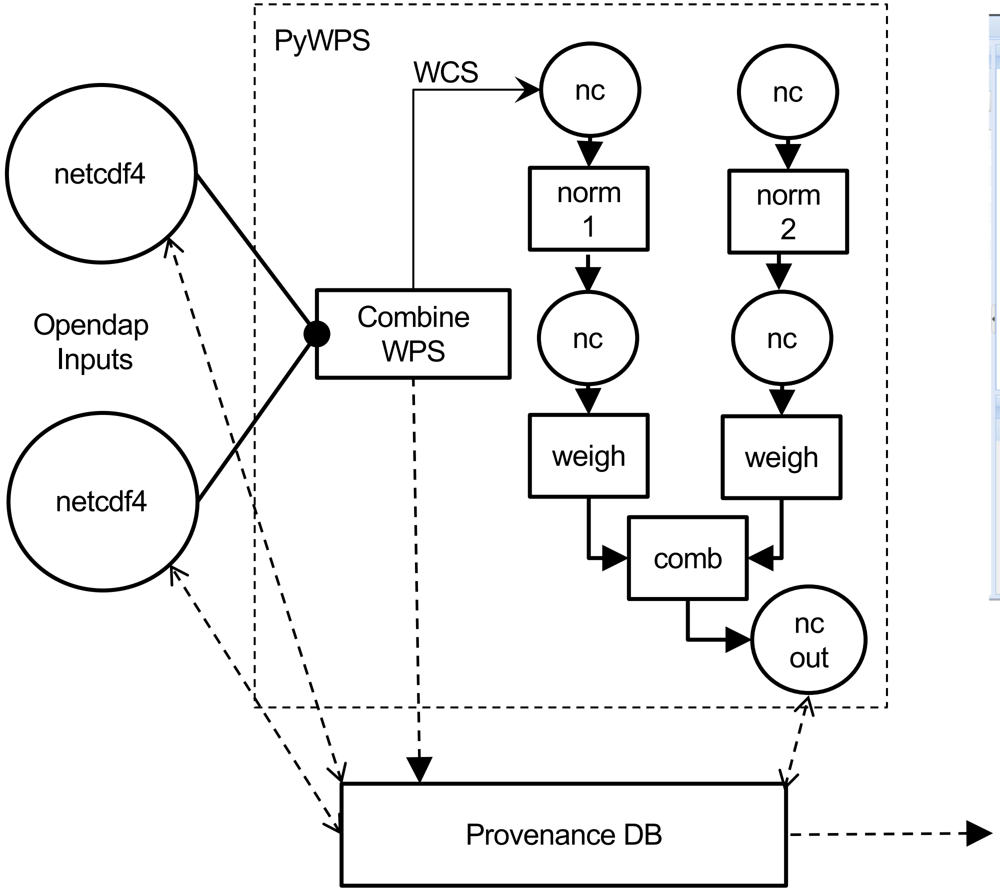
Results: Summer days per European country for MIROC5 / RCP26 !



Workflows, Provenance, Tracability and Reproducability

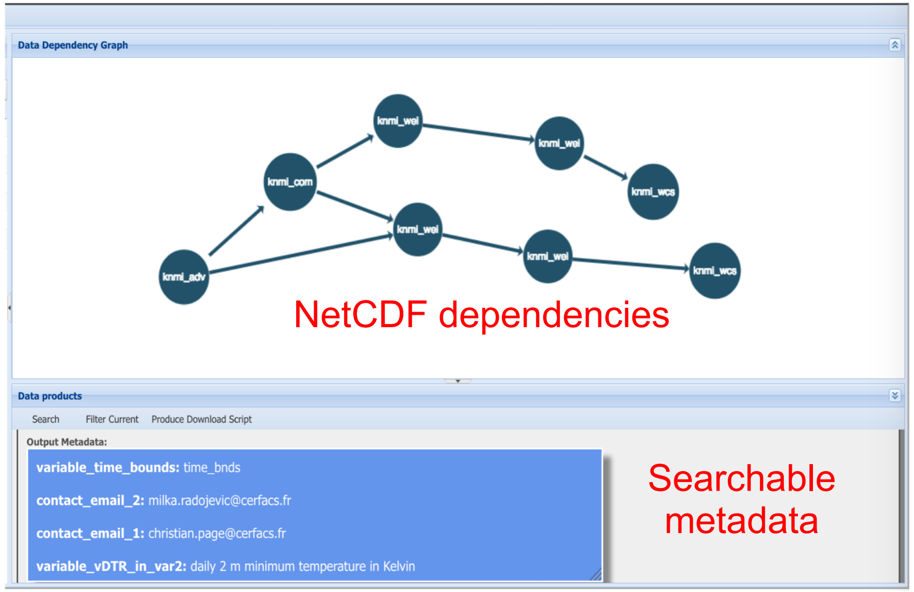
Combine WPS, five steps involved

- Provenance module: WPS_PROV
- Provenance metadata is stored in NetCDF
- W3C PROV-DM standard



Visual analytics techniques on provenance

- Highlighting data-reuse, even for cached data
- User interactions
- Exploitation of resources



What can be improved?

- **Currently C4I handles ESGF data on file level**
  - X Fragmentation of files is a barrier for many users and hurts user experience
  - → Hide file structure, work with datasets and search patterns
  - → Especially important, because now C4I is one of the official data distribution endpoints

- **Currently the Processing services are on the same machine as the portal**
  - X Currently not scalable and processing load effects the portal
  - → Make use of distributed Web Processing Services using delegation

- **Currently the frontend uses old technologies (JSP, Jquery, ExtJS)**
  - X Difficult to maintain, and it is difficult to re-use results from other work
  - → Migrate to ReactJS (Based on work done in the project C3S-Magic)
  - → Good moment to re-design the user interface in collaboration with users.

- **Currently the viewer is running in a separate tab**
  - By using ReactJS, it is easier to make use of an embedded viewer (adaguc-webmapjs)

- **Currently provenance tracking is limited to a few processes**
  - Enhance usage of W3C PROV-DM standard and WPS_PROV toolkit

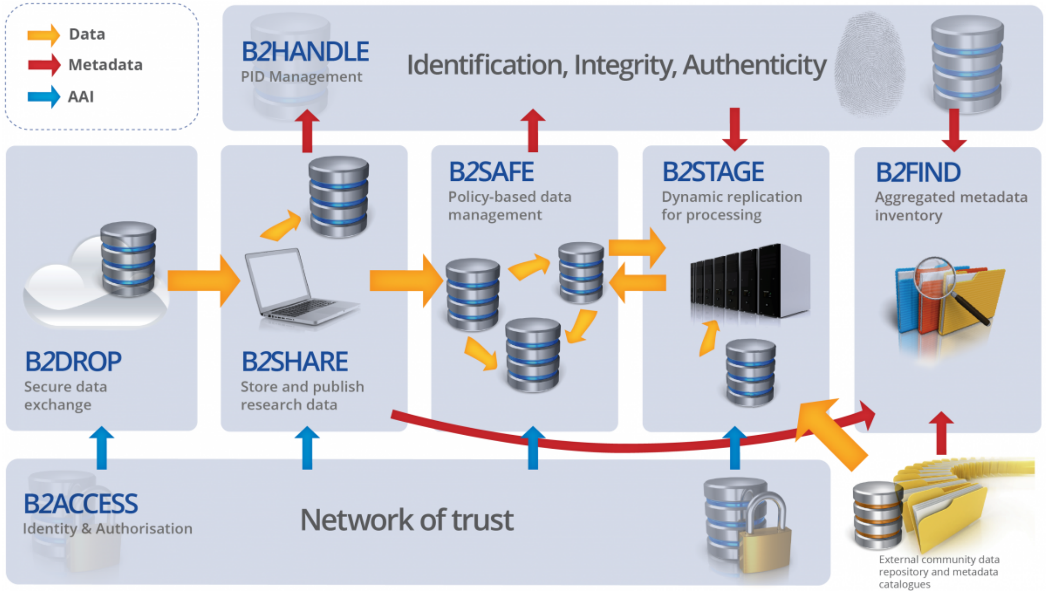
- We are looking for users who are willing to help to improve the platform!

EUDAT CDI & EOSC

European Landscape & Components EUDAT & EOSC

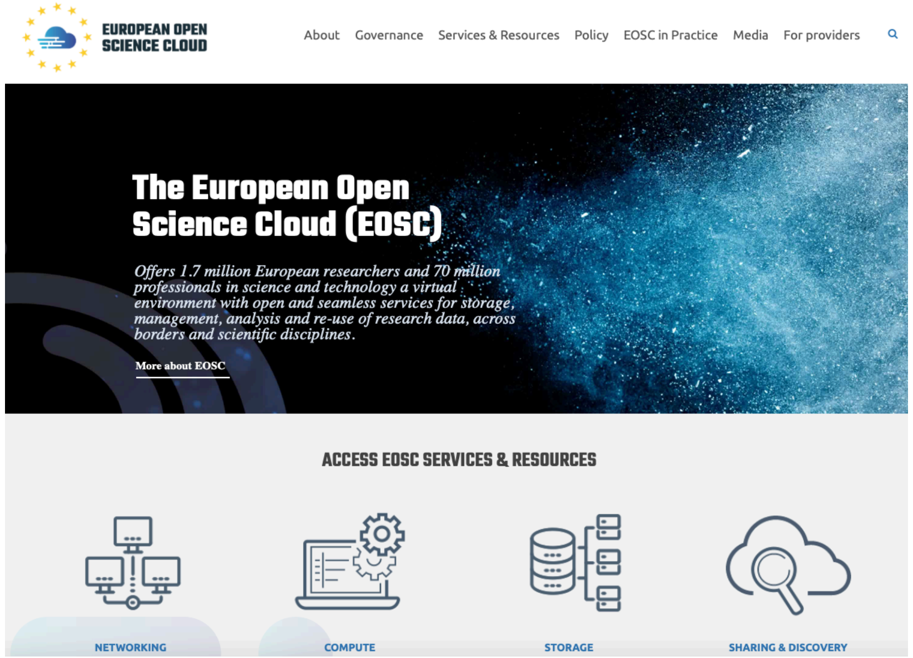
EUDAT CDI B2 Service Suite

- Integrated B2 Services
- B2ACCESS: Common AAI
- B2DROP: Secure Data Exchange
- Interface between EUDAT B2 Services and Communities infrastructures, such as Climate



European Open Science Cloud (EOSC)

- Marketplace of Services
  - Compute
  - Storage
  - Sharing, etc.



EUDAT
CDI Architecture

Mark van de Sanden
EUDAT CDI Technical Coordinator

www.eudat.eu

EUDAT receives funding from the European Union's Horizon 2020 programme - DG CONNECT e-Infrastructures. Contract No. 654065

History of the EUDAT CDI

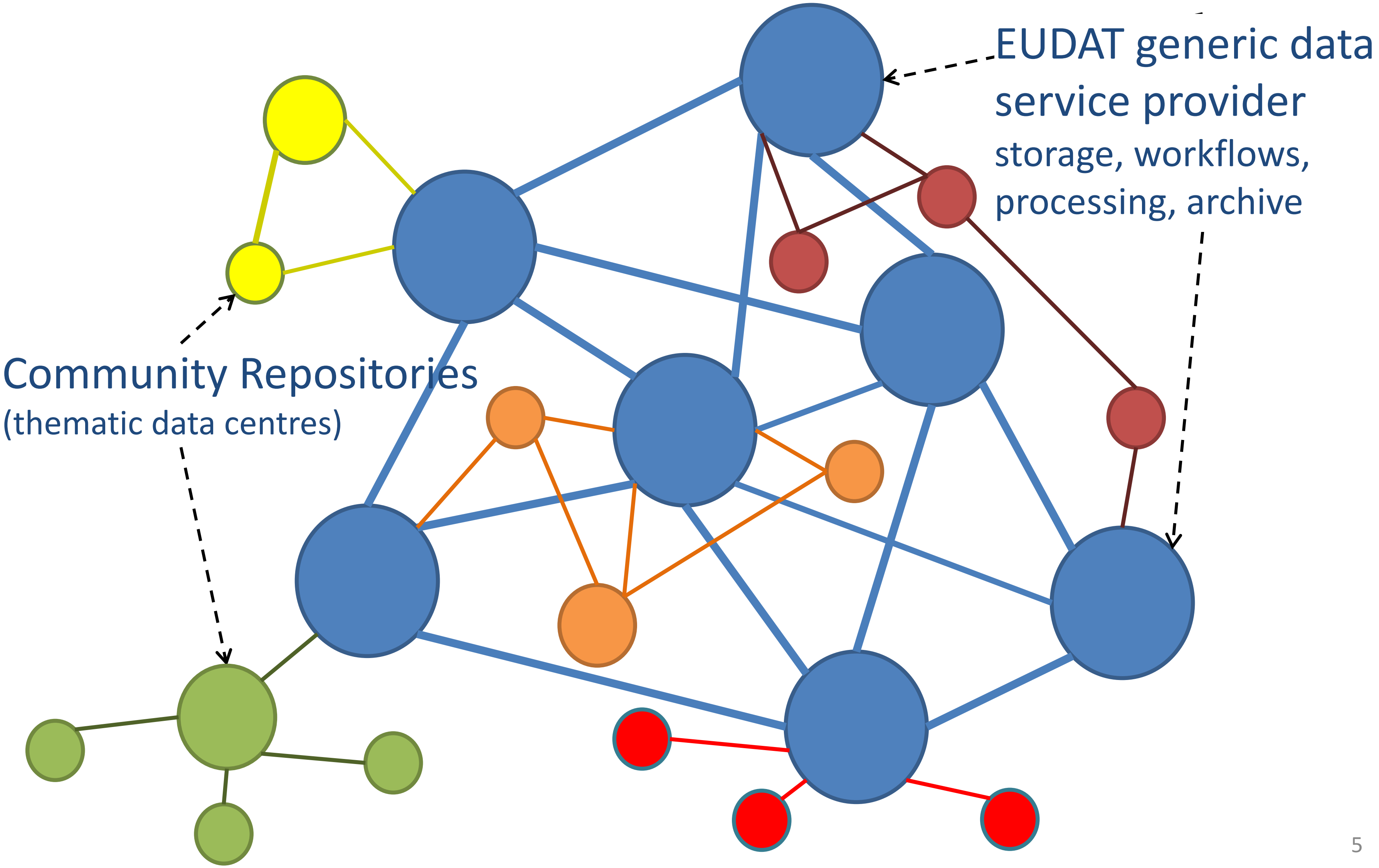
- Common Services for heterogeneous communities
  - Science data rates are exploding and will likely become continue to do so
  - Building bespoke services for new communities is not cost effective
- Initial Set of Services developed as result of community needs
  - Beyond the original ‘core’ communities
  - New services and specific community issues highlighted

If there are hundreds of Research Infrastructures, how many different data management systems can be sustained?

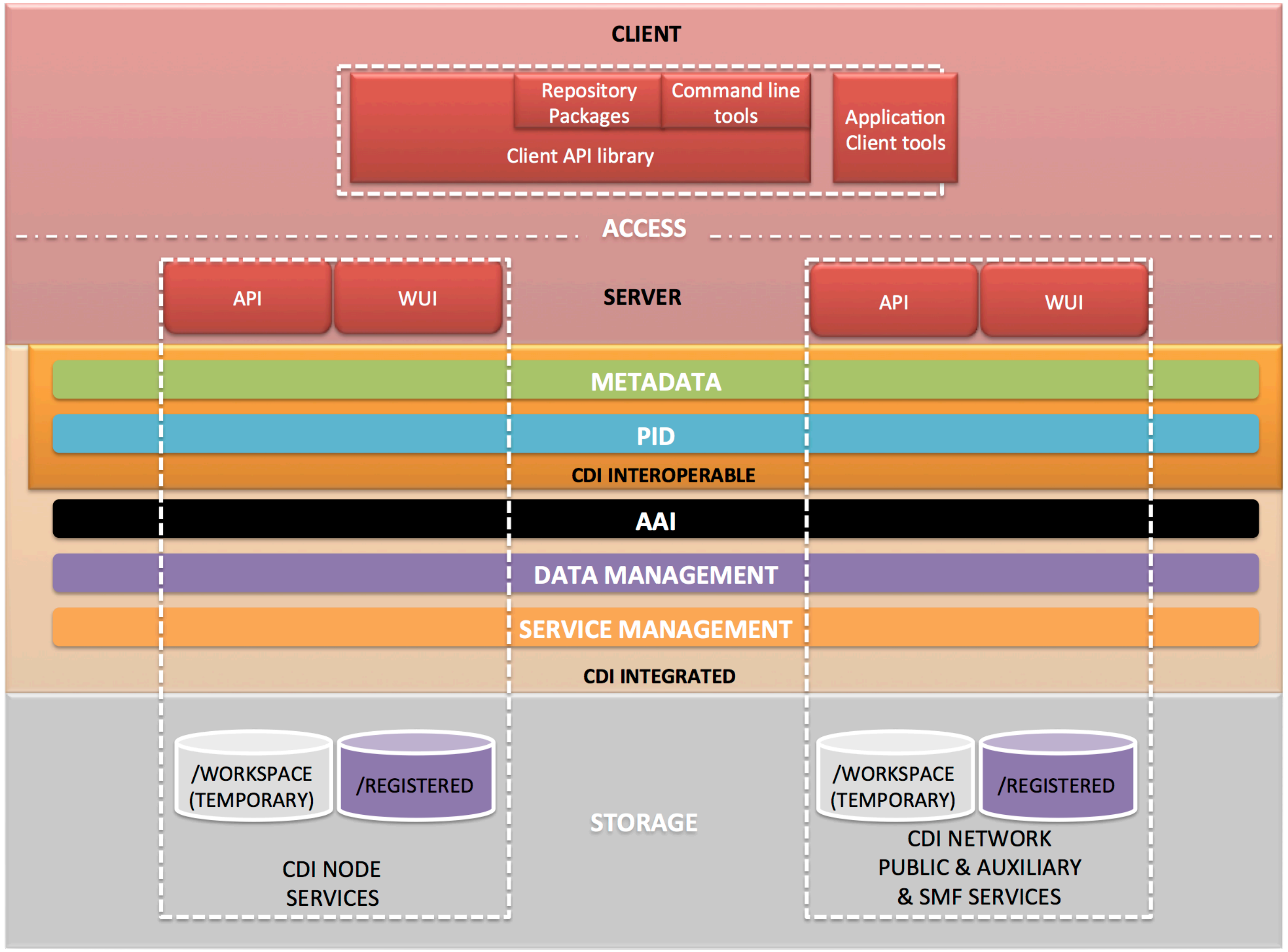
EUDAT CDI

- EUDAT CDI is partner agreement
- Signed September 2016
- Total of 22 partners
- Commitment to sustain the EUDAT CDI pan-European Infrastructure for 10 years
- Partnership for Generic and Thematic service providers (level 1 and 2)
- Membership fee to become CDI partner
- No direct funding for service development
- Plan and organise participating in EC and community projects

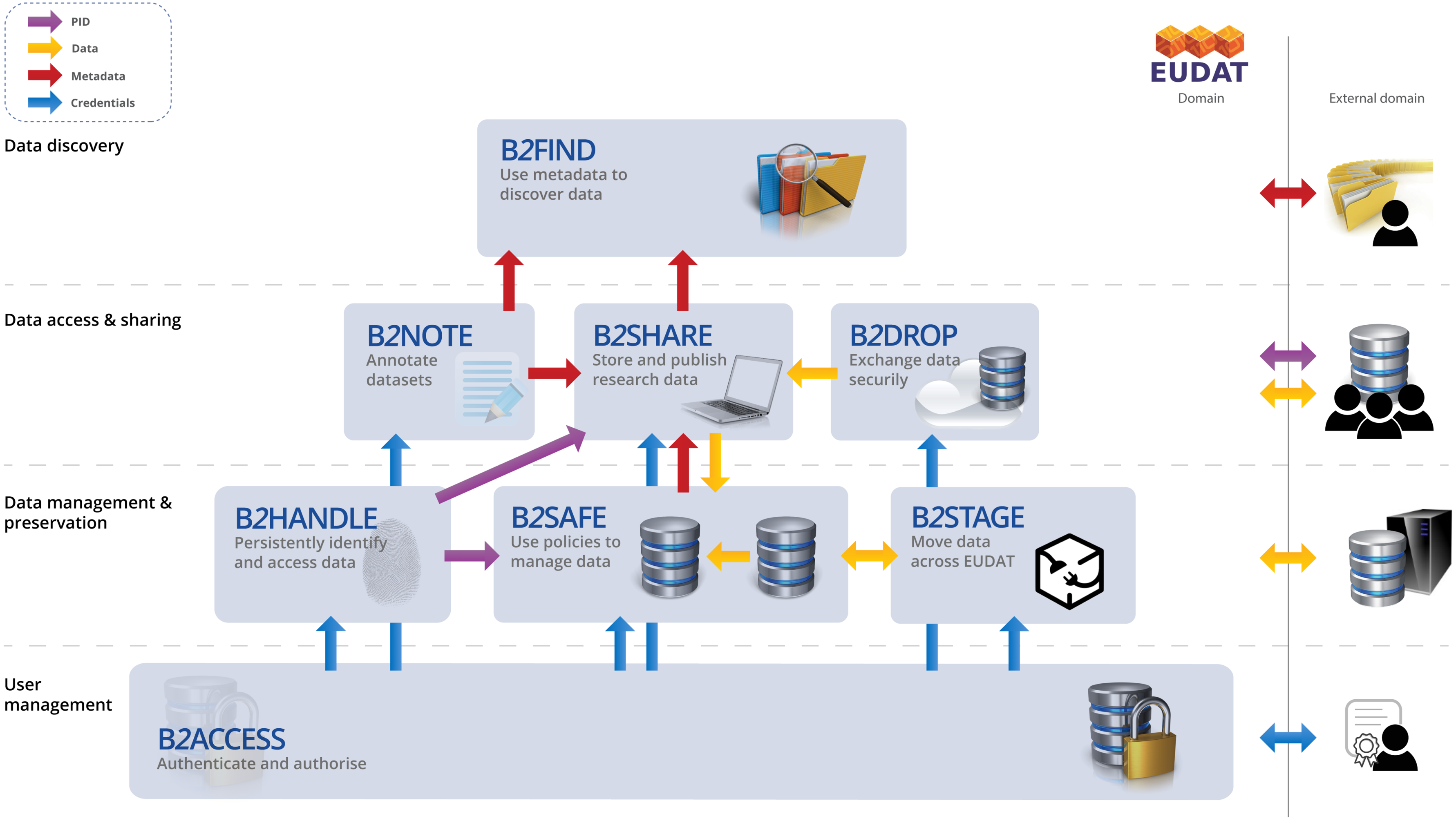
Collaborative Data Infrastructure (CDI)



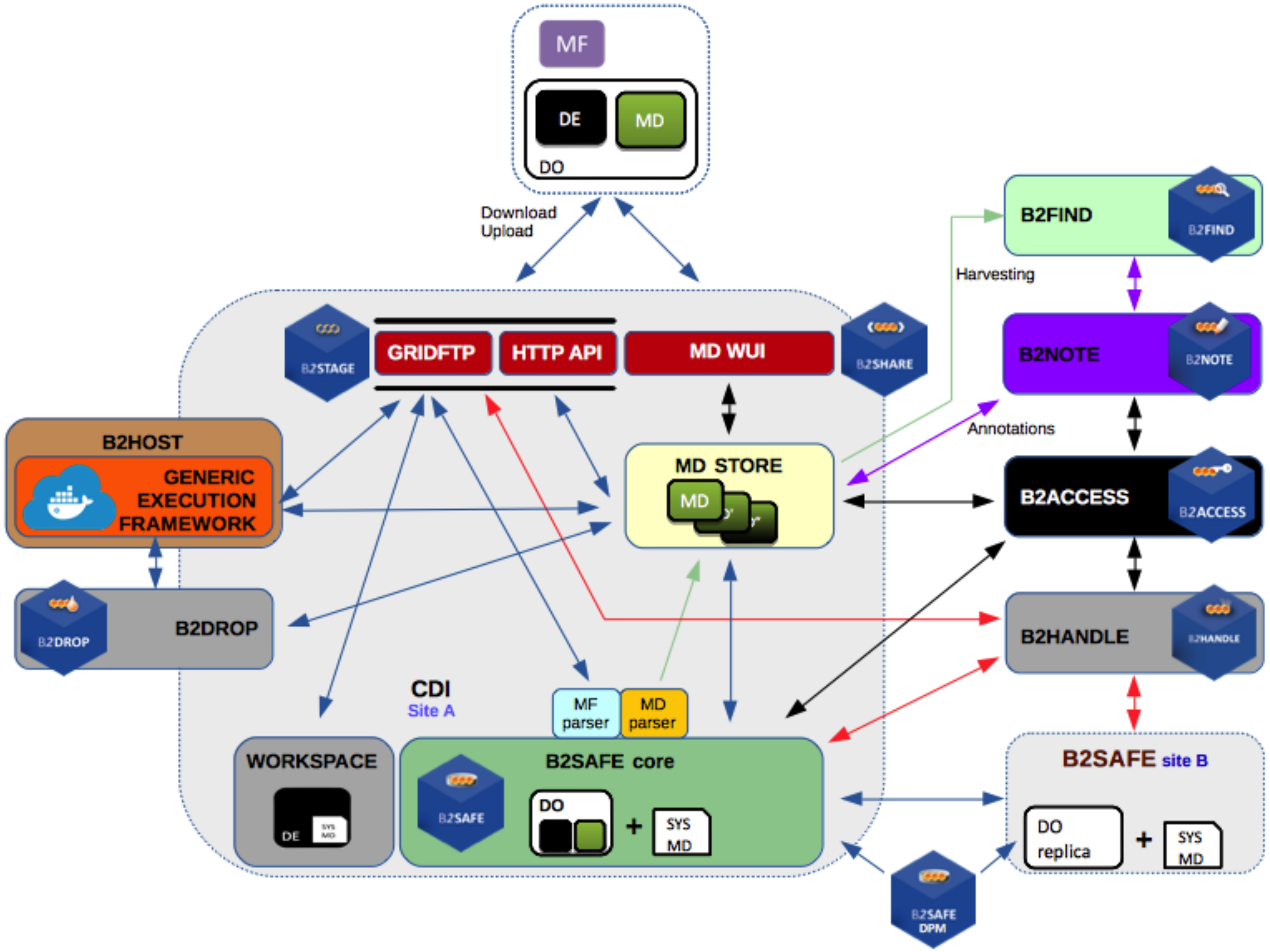
CDI Architecture



Service Diagram



Evolution of the CDI



DARE

DARE