DARE Delivering Agile Research Excellence on European E-Infrastructures

DARE as a platform to support Climate Data Analytics using Cloud Infrastructures

Christian Pagé

Research Engineer / Climate Research Domain

CERFACS Toulouse, France

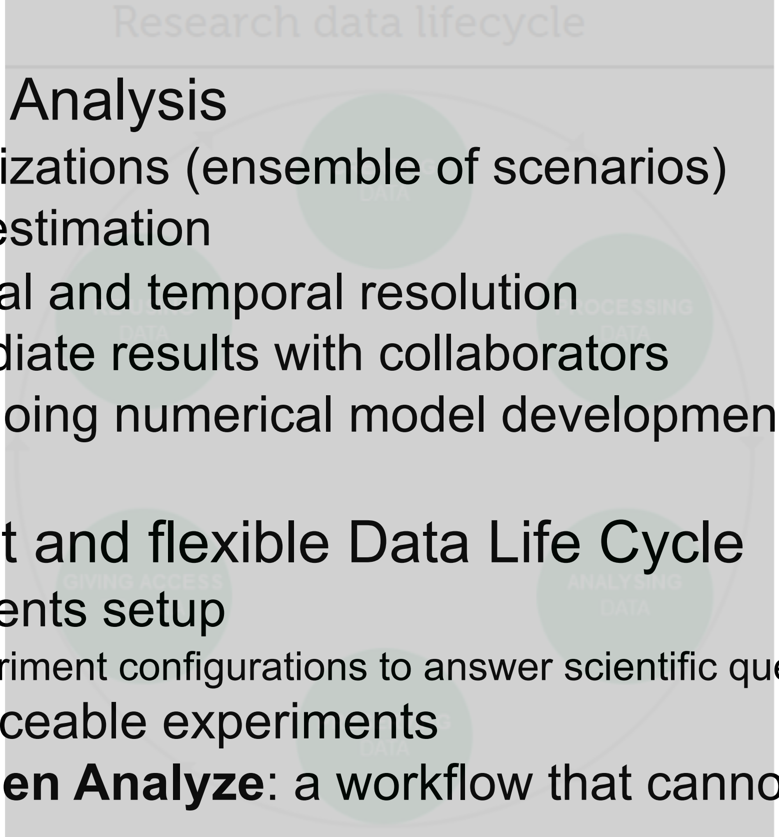
Iraklis A. Klampanos, *NCSR "Demokritos", Greece*

Malcolm Atkinson, *The University of Edinburgh, UK*

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

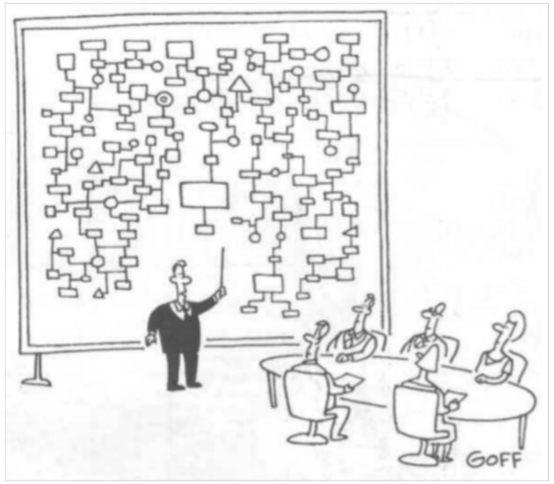
Digital Infrastructures for Research 2018, 9-11 October 2018
Lisbon, Portugal

DARE IS-ENES Climate Use Case
Motivations: Scientific, Technical, Societal



- Perform efficient Data Analysis
  - Large number of realizations (ensemble of scenarios)
  - Uncertainties range estimation
  - Process Higher spatial and temporal resolution
  - Easily share intermediate results with collaborators
  - Comparisons when doing numerical model developments

- Achieve a more robust and flexible Data Life Cycle
  - More robust experiments setup
    - Explore several experiment configurations to answer scientific questions
  - Reproducible and traceable experiments
  - **Download locally then Analyze**: a workflow that cannot be sustained



DARE Mapping Community Needs

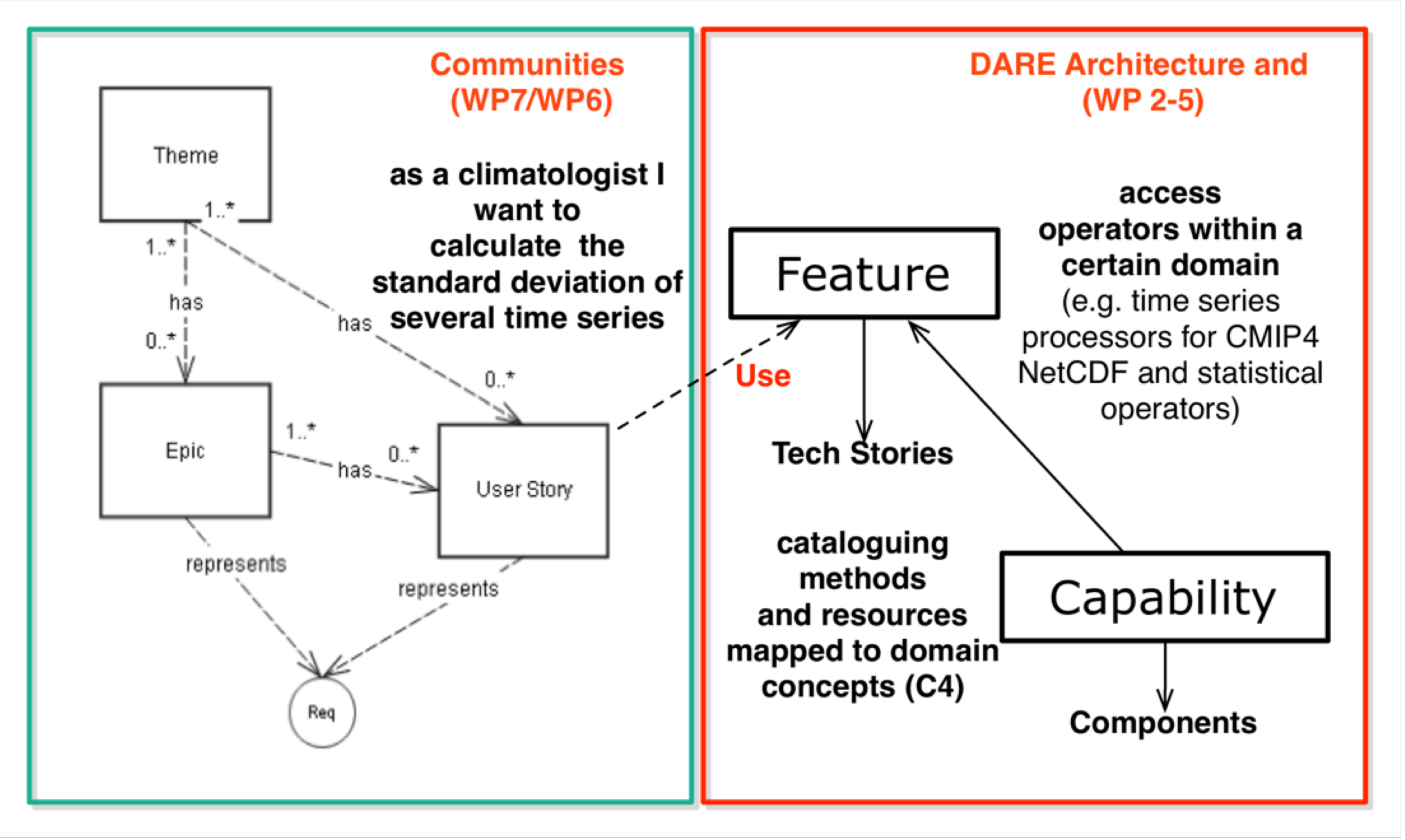
Mapping from Community User Stories to DARE Features and Capabilities

- The user-story requires to access the right implementation of a component (Feature), which may be implemented through resolving services (Tech Story).

- Cataloguing is a capability of DARE

Requirements Analysis for each Pilot

- Identify existing architecture, standards and tools components
- Identify communities' needs



Communities (WP7/WP6)

as a climatologist I want to calculate the standard deviation of several time series

DARE Architecture and (WP 2-5)

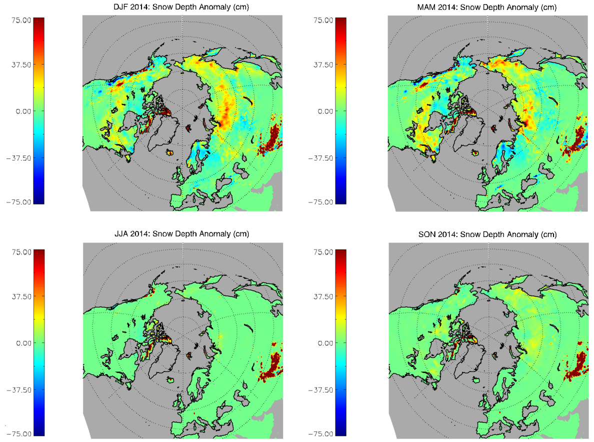
access operators within a certain domain (e.g. time series processors for CMIP4 NetCDF and statistical operators)

cataloguing methods and resources mapped to domain concepts (C4)

Climate Data Users: Current situation

Practical Example: A Climate Research PhD Student

- **I want to study how the feedback of the snow cover in Northern Europe and Russia on the weather circulation patterns and temperature extremes over Western Europe is impacted in the future climate**


DJF 2014: Snow Depth Anomaly (cm)
MAM 2014: Snow Depth Anomaly (cm)
JJA 2014: Snow Depth Anomaly (cm)
SON 2014: Snow Depth Anomaly (cm)
75.00
37.50
0.00
−37.50
−75.00

- Surface Temperature (+max/min), Pressure, Humidity, Snow Cover, Precipitation (Solid&Liquid): 8 surface fields

- Historical + All RCPs
- Combination of models an ensemble members
- EUR-44 Euro-Cordex Grid

- ~11 200 files of ~50 Mb each per field

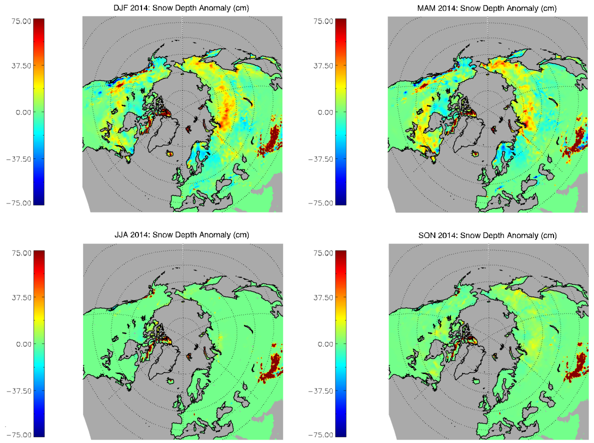
TOTAL: ~560 Gb

CERFACS DARE

Climate Data Users: Current situation

Practical Example: A Climate Research PhD Student

- **I want to study how the feedback of the snow cover in Northern Europe and Russia on the weather circulation patterns and temperature extremes over Western Europe is impacted in the future climate**


DJF 2014: Snow Depth Anomaly (cm)
MAM 2014: Snow Depth Anomaly (cm)
JJA 2014: Snow Depth Anomaly (cm)
SON 2014: Snow Depth Anomaly (cm)
75.00
37.50
0.00
−37.50
−75.00

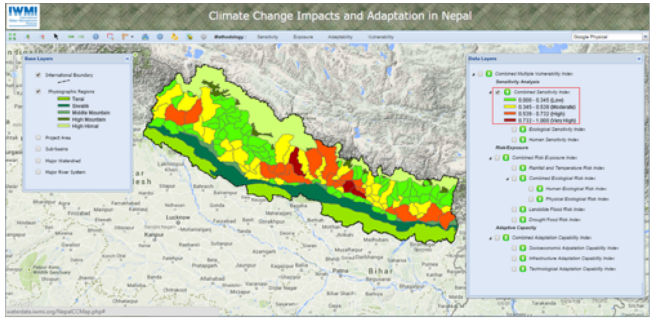
Needs and questions

- I need to calculate several statistics for analyses
- I need derived quantities (climate indices, indicators)
- I want to assess if higher resolution data is needed or other datasets
- I want to do some Quality Check
- ...

Climate Data Users: Current situation

Practical Example: An Impact Engineer

- **My region needs to assess the impact of climate change on how we perform water management. I work with GIS Software to overlay several informational data layers.**


IWMI
Climate Change Impacts and Adaptation in Nepal
Methodology
Sensitivity
Exposure
Adaptability
Vulnerability
Google Physical
Base Layers
International Boundary
Physiographic Regions
Terai
Siwalik
Middle Mountain
High Mountain
High Himal
Project Area
Sub-basins
Major Watershed
Major River System
Data Layers
Combined Multiple Vulnerability Index
Sensitivity Analysis
Combined Sensitivity Index
0.000 - 0.345 (Low)
0.345 - 0.539 (Moderate)
0.539 - 0.732 (High)
0.732 - 1.000 (Very High)
Ecological Sensitivity Index
Human Sensitivity Index
Risk/Exposure
Combined Risk Exposure Index
Rainfall and Temperature Risk Index
Combined Ecological Risk Index
Human Ecological Risk Index
Physical Ecological Risk Index
Landslide Flood Risk Index
Drought Food Risk Index
Adaptive Capacity
Combined Adaptation Capability Index
Socioeconomic Adaptation Capability Index
Infrastructure Adaptation Capability Index
Technological Adaptation Capability Index
Lucknow
Kanpur
Patna
Bihar
waterdata.iwmi.org/NepalCCMap.php#

- Surface Temperature (+max/min), Precipitation, Winds : 6 surface fields

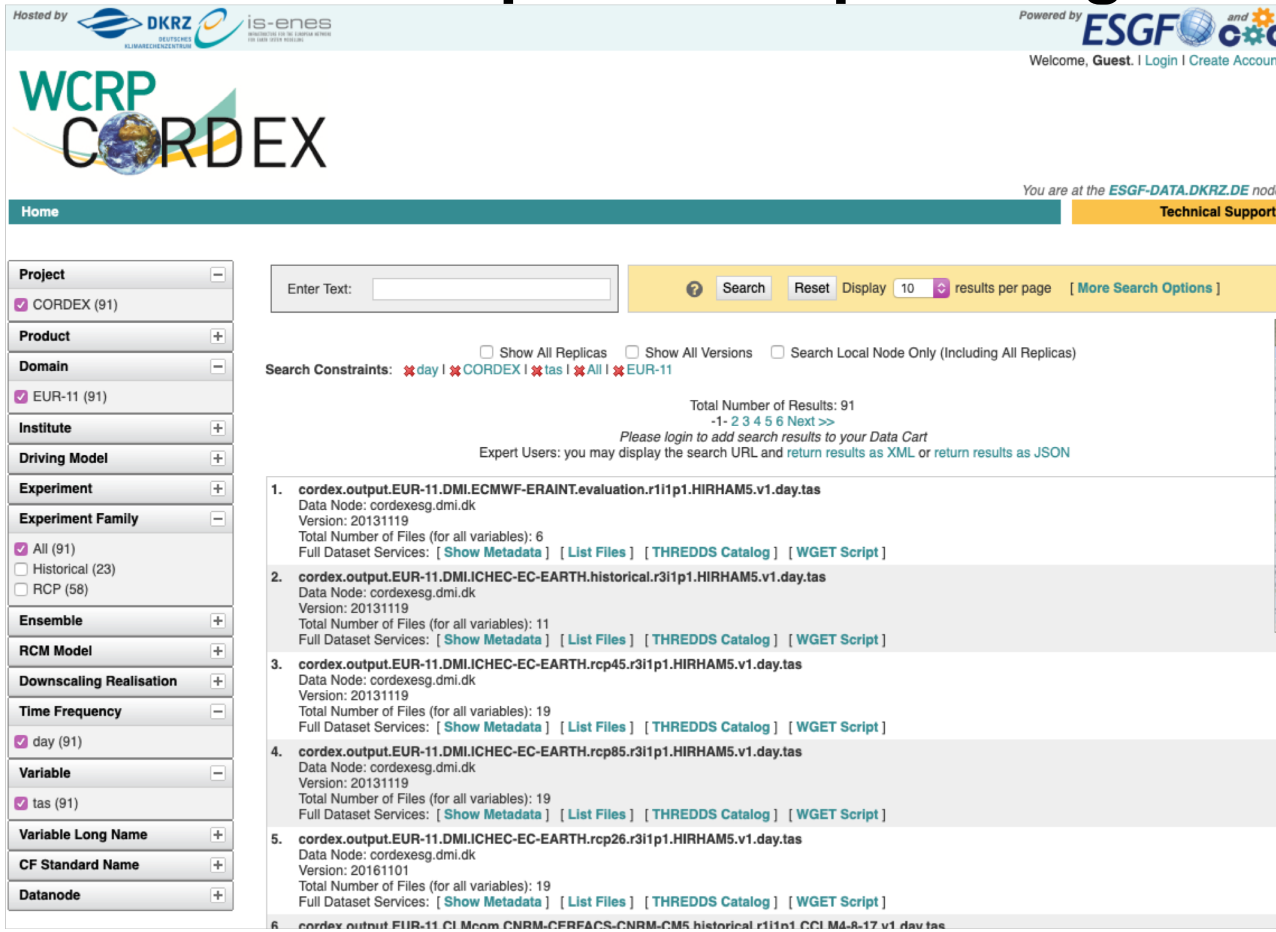
- Historical + All RCPs:
- Combination of models an ensemble members
- EUR-11 Euro-Cordex Grid

- 1378 files of ~600 Mb each per field

TOTAL: ~5 Tb

Climate Data Users: Current situation

Practical Example: An Impact Engineer



Hosted by DKRZ is-enes

Powered by ESGF and CoG

Welcome, Guest. | Login | Create Account

WCRP CORDEX

You are at the ESGF-DATA.DKRZ.DE node

Home | Technical Support

- Project ⊟
  - ☑ CORDEX (91)
- Product ⊞
- Domain ⊟
  - ☑ EUR-11 (91)
- Institute ⊞
- Driving Model ⊞
- Experiment ⊞
- Experiment Family ⊟
  - ☑ All (91)
  - □ Historical (23)
  - □ RCP (58)
- Ensemble ⊞
- RCM Model ⊞
- Downscaling Realisation ⊞
- Time Frequency ⊟
  - ☑ day (91)
- Variable ⊟
  - ☑ tas (91)
- Variable Long Name ⊞
- CF Standard Name ⊞
- Datanode ⊞

Enter Text: [ ]

? Search | Reset | Display 10 results per page [More Search Options]

□ Show All Replicas □ Show All Versions □ Search Local Node Only (Including All Replicas)

Search Constraints: ✖day | ✖CORDEX | ✖tas | ✖All | ✖EUR-11

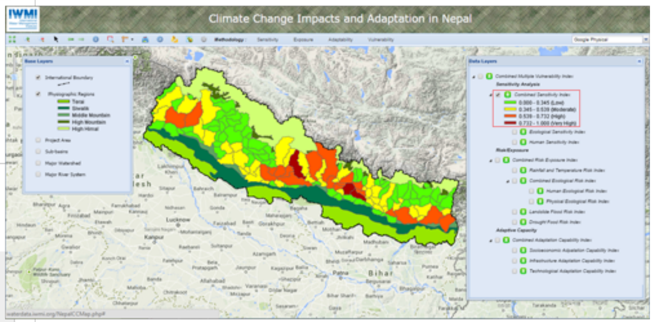
Total Number of Results: 91

-1- 2 3 4 5 6 Next >>

Please login to add search results to your Data Cart

Expert Users: you may display the search URL and return results as XML or return results as JSON

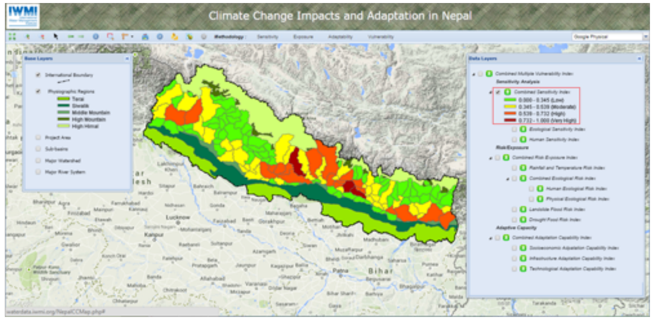
1. cordex.output.EUR-11.DMI.ECMWF-ERAINT.evaluation.r1i1p1.HIRHAM5.v1.day.tas
   Data Node: cordexesg.dmi.dk
   Version: 20131119
   Total Number of Files (for all variables): 6
   Full Dataset Services: [Show Metadata] [List Files] [THREDDS Catalog] [WGET Script]
2. cordex.output.EUR-11.DMI.ICHEC-EC-EARTH.historical.r3i1p1.HIRHAM5.v1.day.tas
   Data Node: cordexesg.dmi.dk
   Version: 20131119
   Total Number of Files (for all variables): 11
   Full Dataset Services: [Show Metadata] [List Files] [THREDDS Catalog] [WGET Script]
3. cordex.output.EUR-11.DMI.ICHEC-EC-EARTH.rcp45.r3i1p1.HIRHAM5.v1.day.tas
   Data Node: cordexesg.dmi.dk
   Version: 20131119
   Total Number of Files (for all variables): 19
   Full Dataset Services: [Show Metadata] [List Files] [THREDDS Catalog] [WGET Script]
4. cordex.output.EUR-11.DMI.ICHEC-EC-EARTH.rcp85.r3i1p1.HIRHAM5.v1.day.tas
   Data Node: cordexesg.dmi.dk
   Version: 20131119
   Total Number of Files (for all variables): 19
   Full Dataset Services: [Show Metadata] [List Files] [THREDDS Catalog] [WGET Script]
5. cordex.output.EUR-11.DMI.ICHEC-EC-EARTH.rcp26.r3i1p1.HIRHAM5.v1.day.tas
   Data Node: cordexesg.dmi.dk
   Version: 20161101
   Total Number of Files (for all variables): 19
   Full Dataset Services: [Show Metadata] [List Files] [THREDDS Catalog] [WGET Script]
6. cordex.output.EUR-11.CLMcom.CNRM-CERFACS-CNRM-CM5.historical.r1i1p1.CCLM4-8-17.v1.day.tas



Climate Data Users: Current situation

Practical Example: An Impact Engineer

- **My region needs to assess the impact of climate change on how we perform water management. I work with GIS Software to overlay several informational data layers.**


IWMI
Climate Change Impacts and Adaptation in Nepal
Methodology
Sensitivity
Exposure
Adaptability
Vulnerability
Base Layers
International Boundary
Physiographic Regions
Terai
Siwalik
Middle Mountain
High Mountain
High Himal
Project Area
Sub-basins
Major Watershed
Major River System
Data Layers
Combined Multiple Vulnerability Index
Sensitivity Analysis
Combined Sensitivity Index
0.000 - 0.345 (Low)
0.345 - 0.539 (Moderate)
0.539 - 0.732 (High)
0.732 - 1.000 (Very High)
Ecological Sensitivity Index
Human Sensitivity Index
Risk Exposure
Combined Risk Exposure Index
Rainfall and Temperature Risk Index
Combined Ecological Risk Index
Human Ecological Risk Index
Physical Ecological Risk Index
Landslide Flood Risk Index
Drought Food Risk Index
Adaptive Capacity
Combined Adaptation Capability Index
Socioeconomic Adaptation Capability Index
Infrastructure Adaptation Capability Index
Technological Adaptation Capability Index
Lucknow
Kanpur
Patna
Bihar
waterdata.iwmi.org/NepalCCMap.php#

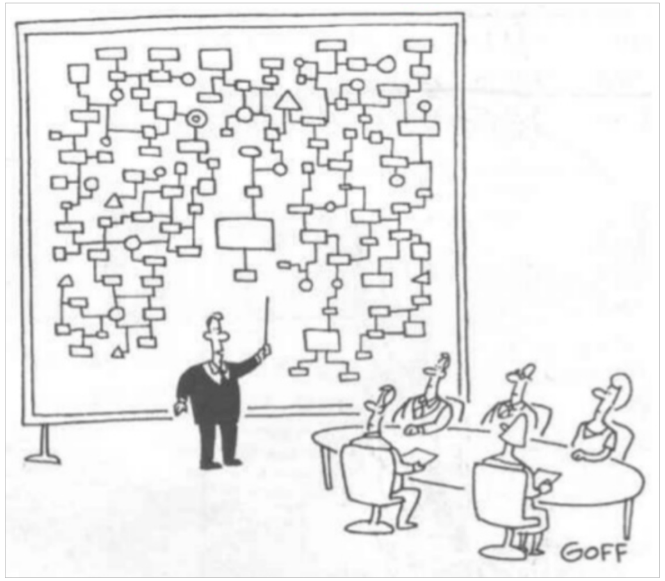
Needs and questions

- How to reduce the dataset to a representative subset?
  - My client cannot cope with too many realizations
- I need to do the calculations remotely and download the results
- I cannot use NetCDF, I need to import the data into my GIS software

Climate Data Users: Current situation

Many common needs

- Guidance and tools for data and scenarios subsetting: selecting a subset of representative scenarios
- Lower significantly the total data size to download
  - Calculate as much as possible remotely
- Reformat/Repackage the data into easier formats and organization/homogenization (implies smaller datasize)
- Full Provenance and Lineage information
- Proper Metadata description, especially for derived data
- Variety of Access Interfaces for adoption: OGC, REST, Jupyter, APIs



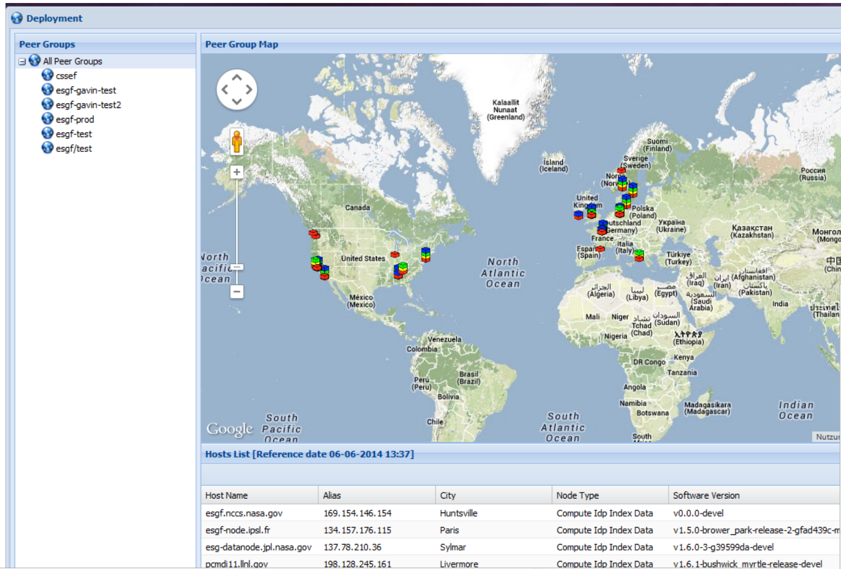
Climate Data Distribution: ESGF RI

Institut Pierre Simon Laplace – Sciences de l'environnement
DKRZ
National Centre for Atmospheric Science – NATURAL ENVIRONMENT RESEARCH COUNCIL
Science & Technology Facilities Council
DMI Vejr, klima og hav
SMHI
CMCC Centro euro-Mediterraneo sui Cambiamenti Climatici
UC Universidad de Cantabria
Royal Netherlands Meteorological Institute – Ministry of Transport, Public Works and Water Management
METEO FRANCE – Toujours un temps d'avance

ESGF Data Nodes 2015

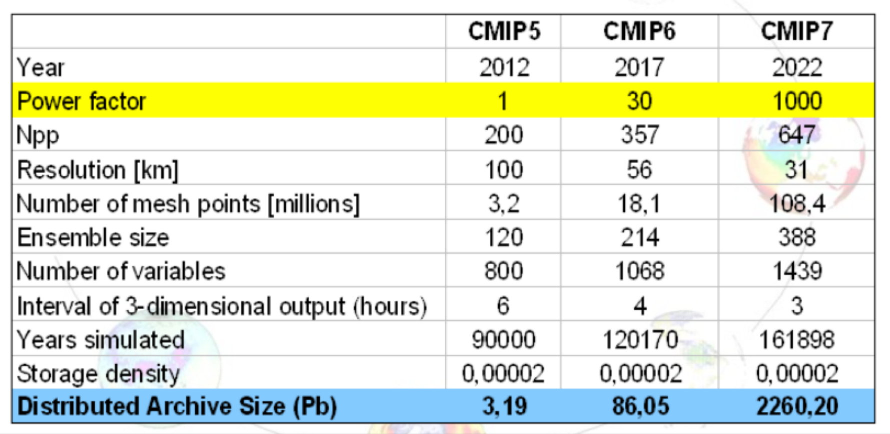
- 40 worldwide
- 18 in Europe (coordinated in IS-ENES)

is-enes – INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING
e-infrastructure
Clipc – Climate Information Portal



Hosts List [Reference date 06-06-2014 13:37]

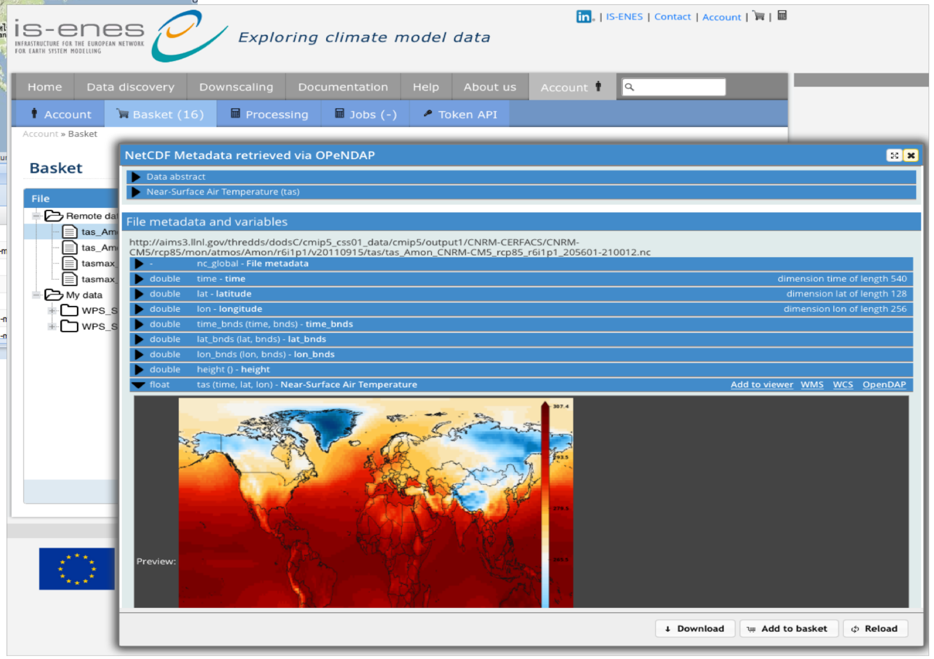
Host Name	Alias	City	Node Type	Software Version
esgf.nccs.nasa.gov	169.154.146.154	Huntsville	Compute Idp Index Data	v0.0.0-devel
esgf-node.ipsl.fr	134.157.176.115	Paris	Compute Idp Index Data	v1.5.0-brower_park-release-2-gfad439c-m
esg-datanode.jpl.nasa.gov	137.78.210.36	Sylmar	Compute Idp Index Data	v1.6.0-3-g39599da-devel
pcmdi11.llnl.gov	198.128.245.161	Livermore	Compute Idp Index Data	v1.6.1-bushwick_myrtle-release-devel



	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

IS-ENES CDI C4I

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



NetCDF Metadata retrieved via OPeNDAP

- Data abstract
- Near-Surface Air Temperature (tas)

File metadata and variables

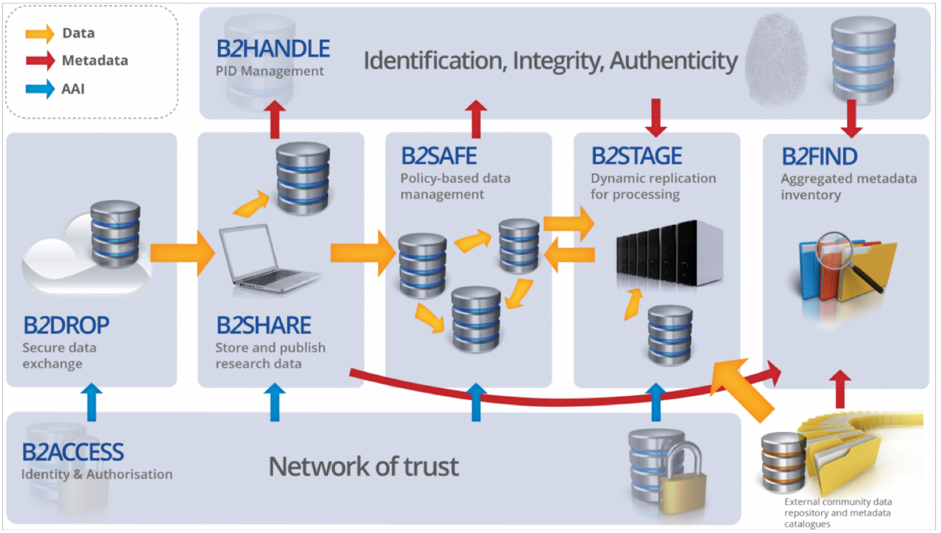
http://aims3.llnl.gov/thredds/dodsC/cmip5_css01_data/cmip5/output1/CNRM-CERFACS/CNRM-CM5/rcp85/mon/atmos/Amon/r6i1p1/v20110915/tas/tas_Amon_CNRM-CM5_rcp85_r6i1p1_205601-210012.nc

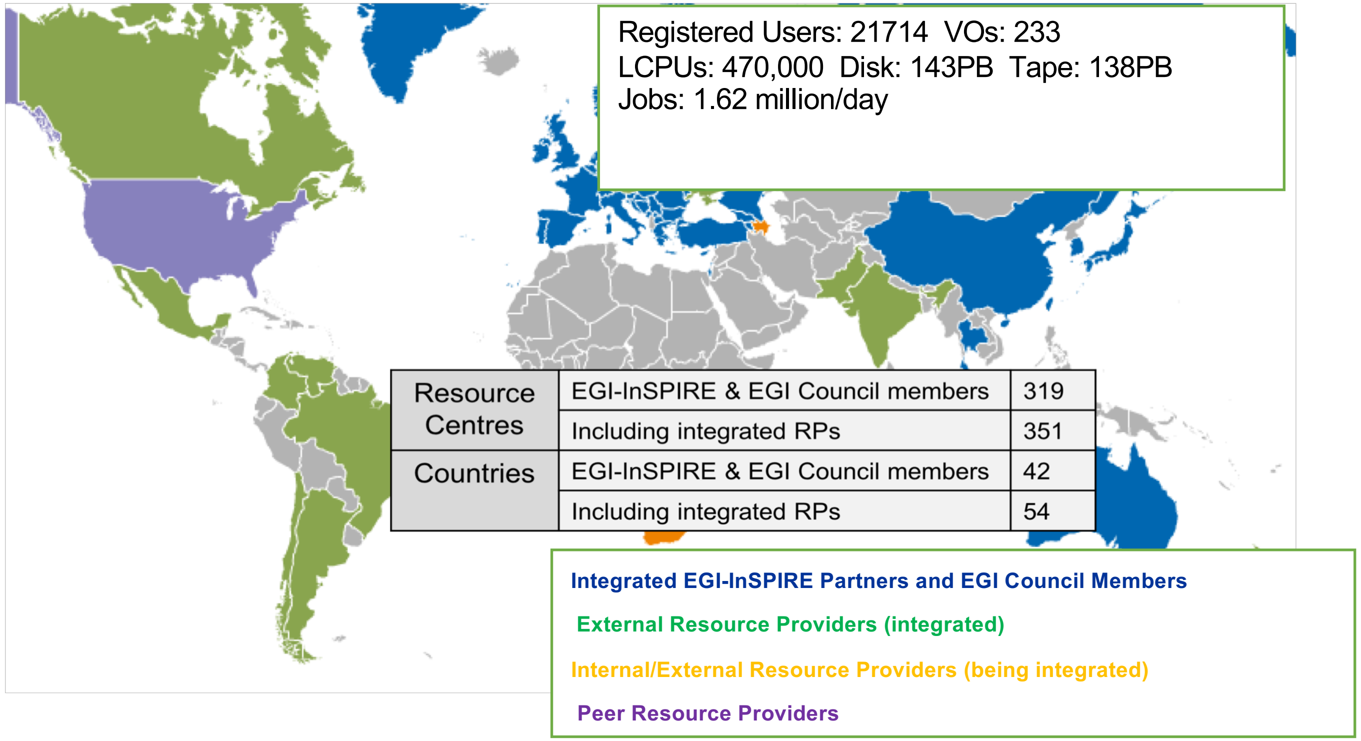
-	nc_global - **File metadata**	
double	time - **time**	dimension time of length 540
double	lat - **latitude**	dimension lat of length 128
double	lon - **longitude**	dimension lon of length 256
double	time_bnds (time, bnds) - **time_bnds**	
double	lat_bnds (lat, bnds) - **lat_bnds**	
double	lon_bnds (lon, bnds) - **lon_bnds**	
double	height () - **height**	
float	tas (time, lat, lon) - **Near-Surface Air Temperature**	Add to viewer WMS WCS OpenDAP

European Landscape & Components EUDAT & EGI

EUDAT CDI B2 Service Suite

- Integrated B2 Services
- B2ACCESS: Common AAI
- Interface between EUDAT B2 Services and Communities infrastructures, such as Climate
- Prototype Workflow Service: GEF (Generic Execution Framework)


Data
Metadata
AAI
B2HANDLE
PID Management
Identification, Integrity, Authenticity
B2SAFE
Policy-based data management
B2STAGE
Dynamic replication for processing
B2FIND
Aggregated metadata inventory
B2DROP
Secure data exchange
B2SHARE
Store and publish research data
B2ACCESS
Identity & Authorisation
Network of trust
External community data repository and metadata catalogues



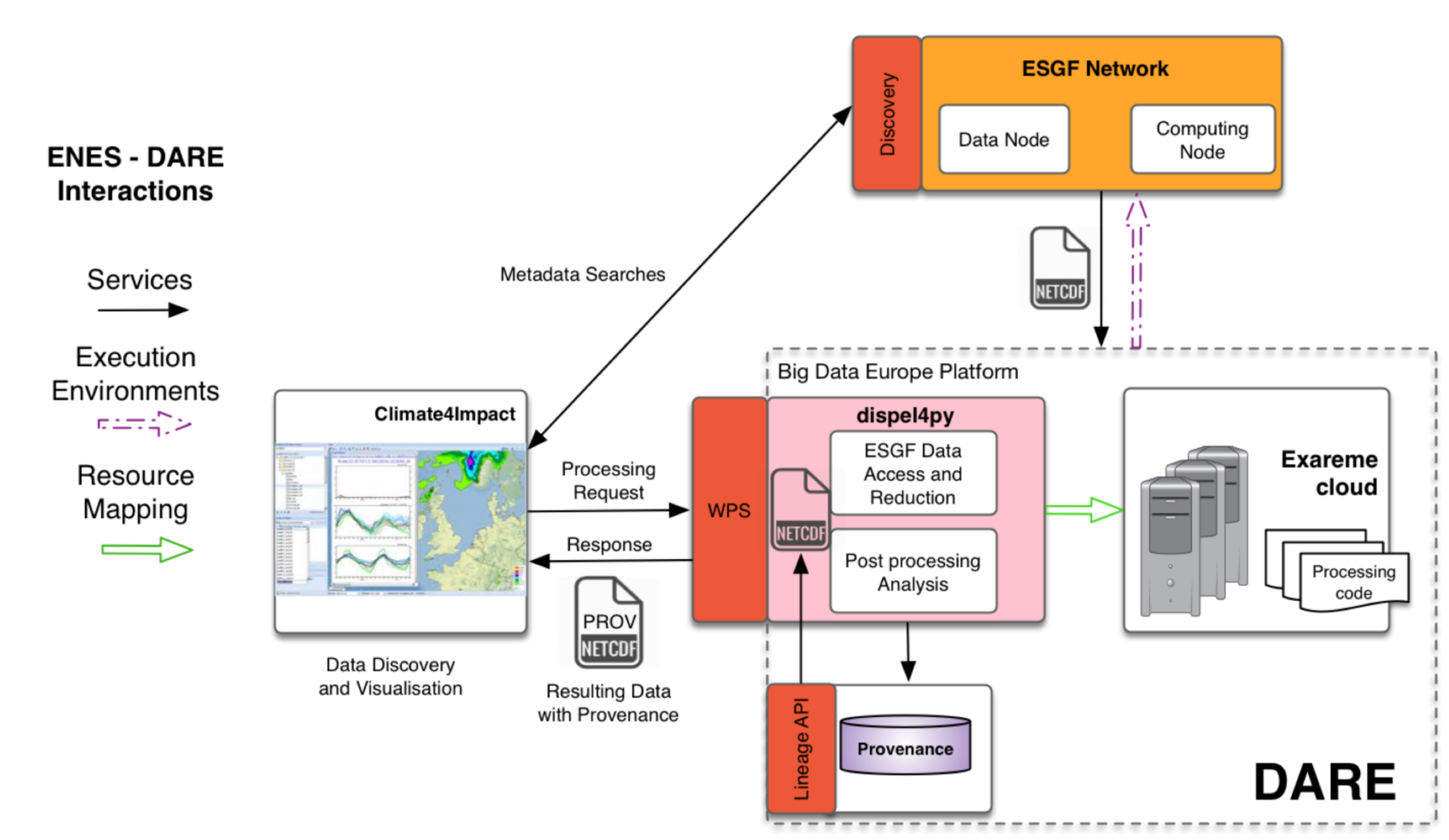
Registered Users: 21714 VOs: 233
LCPUs: 470,000 Disk: 143PB Tape: 138PB
Jobs: 1.62 million/day

Resource Centres	EGI-InSPIRE & EGI Council members	319
	Including integrated RPs	351
Countries	EGI-InSPIRE & EGI Council members	42
	Including integrated RPs	54

Integrated EGI-InSPIRE Partners and EGI Council Members
External Resource Providers (integrated)
Internal/External Resource Providers (being integrated)
Peer Resource Providers

- Computing Power (FedCloud) resources

DARE IS-ENES Climate Use Case Draft Architecture



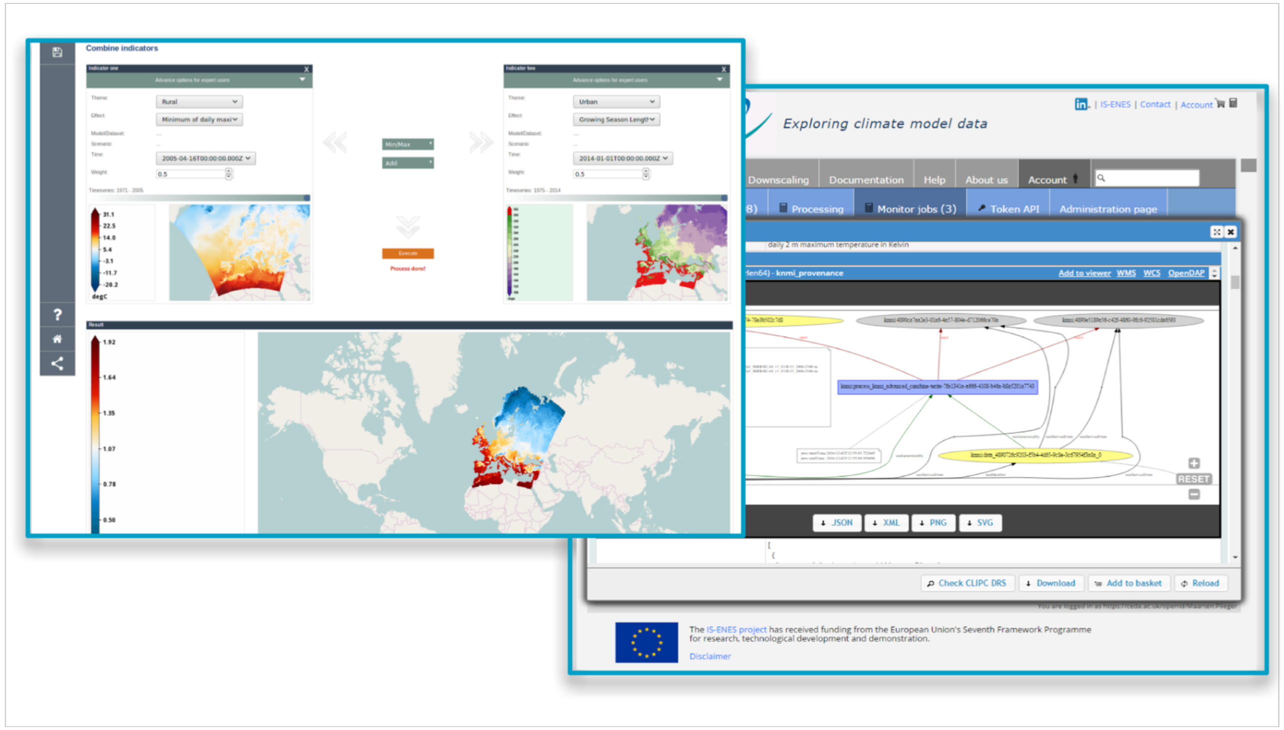
Monitoring and Exploration of WPS workflows via Provenance

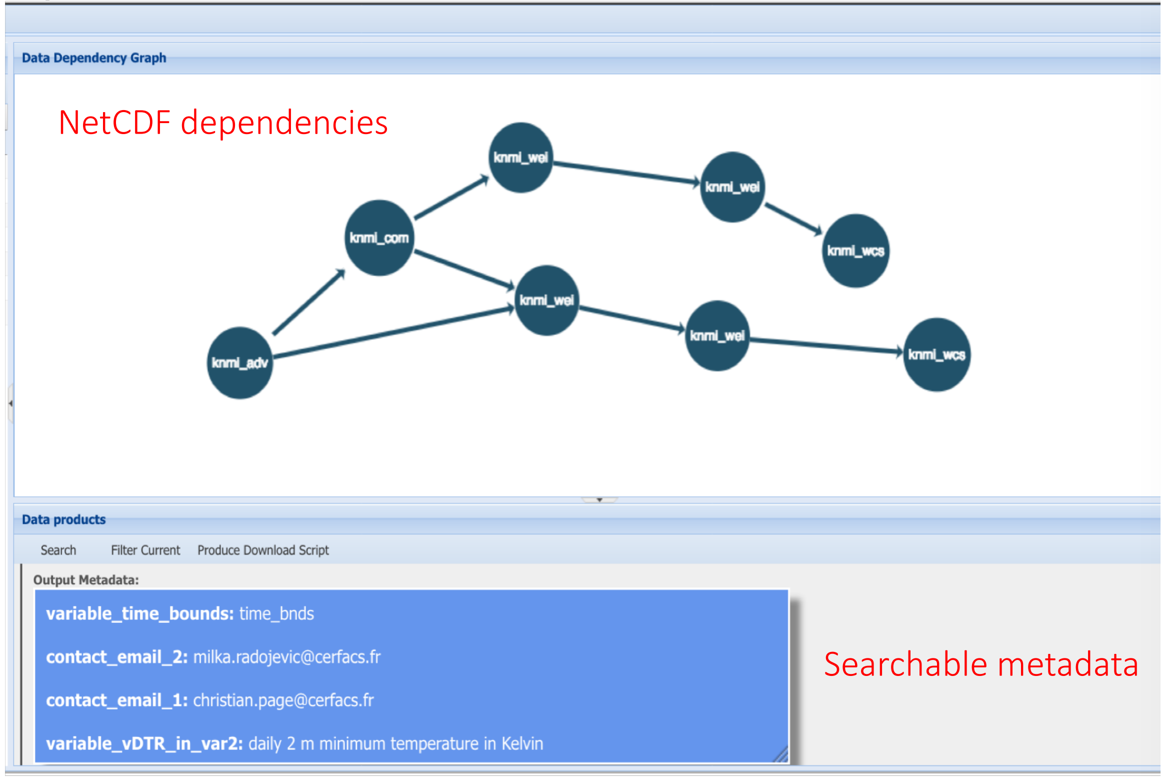
Standards

- 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





Output Metadata:

variable_time_bounds: time_bnds

contact_email_2: milka.radojevic@cerfacs.fr

contact_email_1: christian.page@cerfacs.fr

variable_vDTR_in_var2: daily 2 m minimum temperature in Kelvin

DARE IS-ENES Climate Use Case Significant benefits

- Enable the on-demand **delegation of IS-ENES C4I** Platform **Data Analytics** and **Processing** to the **DARE** infrastructure (cloud-ready).
  - Typically, data reduction **on the order of 70-90%** can be achieved, depending on the users' analyses
- Streamline and **ease** the whole **data lifecycle**, with proper data **provenance** and **lineage**

- The DARE Platform will also:
  - **Be interoperable** with **EUDAT CDI** by using its standards and B2 Services
  - **Interface** with **EOSC** and **Copernicus C3S-DIAS**

Questions & Comments! ☺

http://project-dare.eu

christian.page@cerfacs.fr

Questions & Comments! ☺

http://project-dare.eu

christian.page@cerfacs.fr

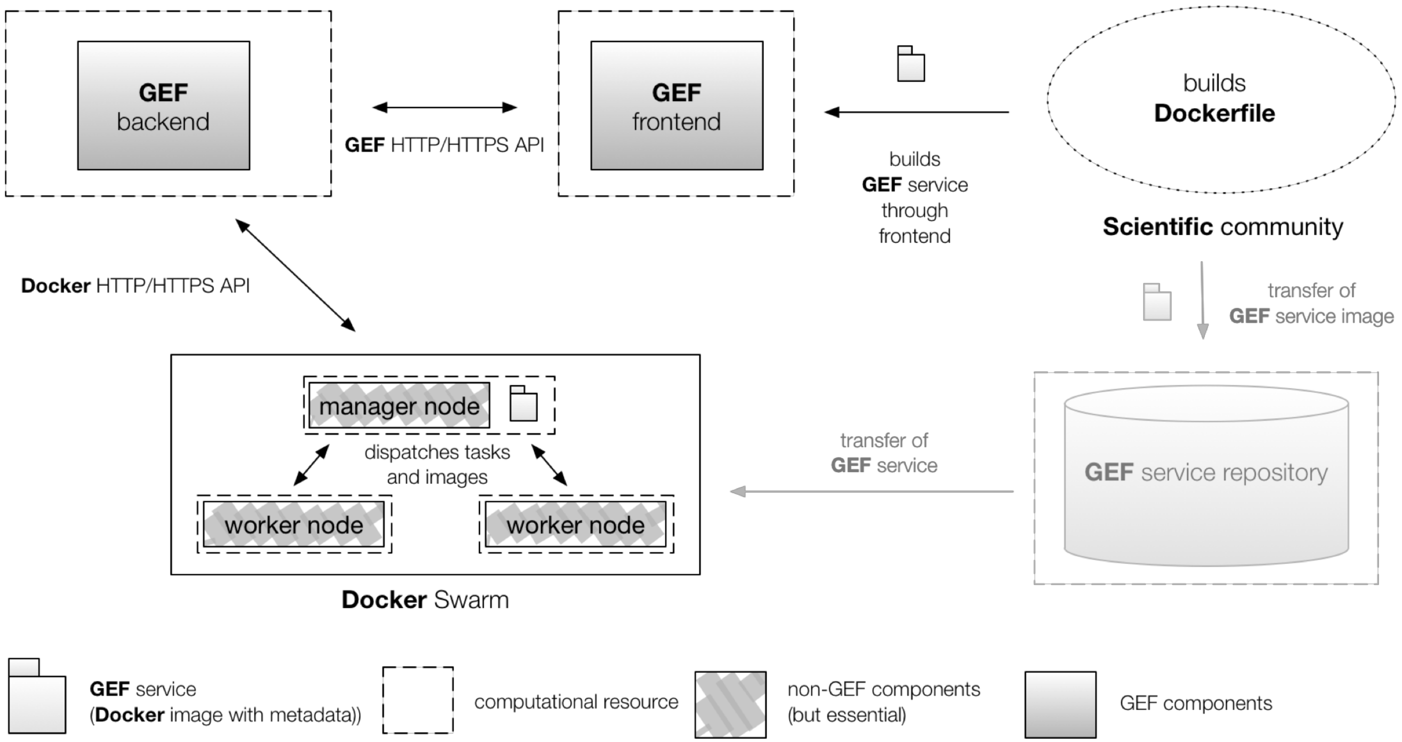
Open Questions

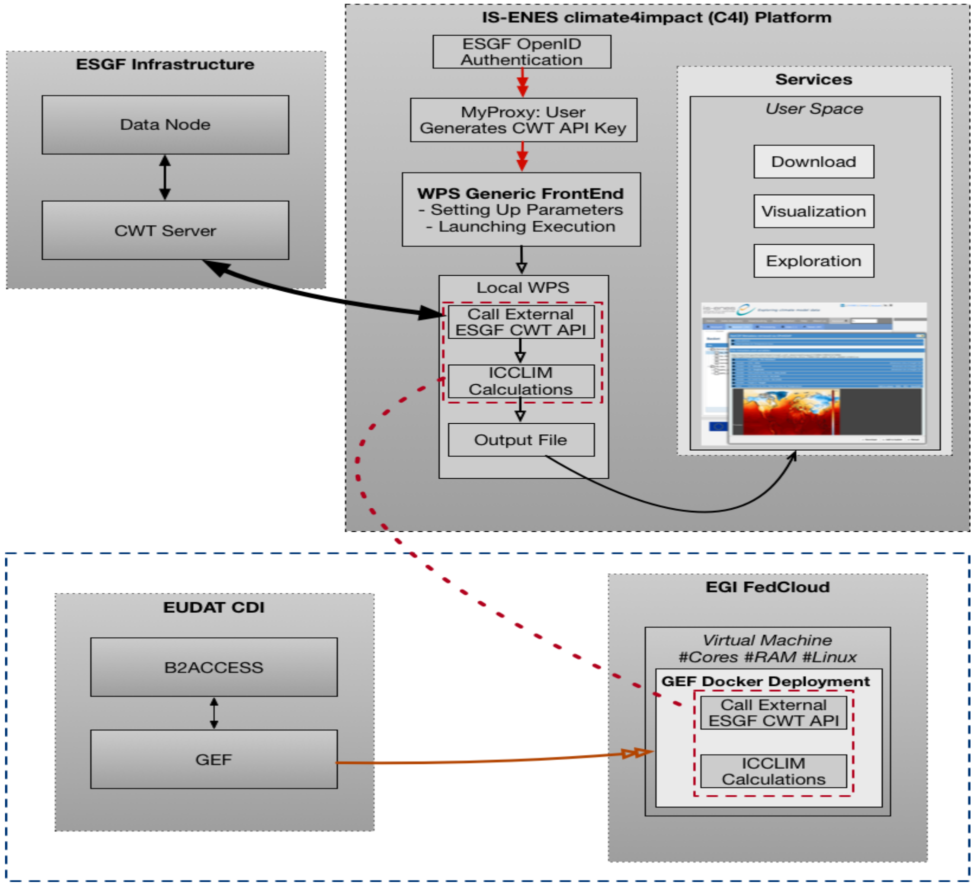
- Several European platforms will be available: C3S-DIAS, EOSC, ESGF Data/Computing Nodes, IS-ENES CDI & ECAS, EUDAT CDI, EGI, DARE, National Platforms, MAIDK
  - How do we ensure that we do not have duplicate efforts (too much)?
  - Which kind of users do they each address? How users will know which one to use? The ones they can access? With what kind of resources limitations?
  - How do we "educate" different kind of users for wide adoption and usage of those platforms?
  - How can they be interoperable? APIs, AAIs, ...
  - How to ensure that they make available promptly new datasets
  - Will they be scalable enough?

Open Questions

- How do we deal with non-mature services, changing APIs?
- On-demand remote data processing and data sharing is really needed
- Containerized solutions: distributed processing, orchestration, AAIs...
- What about Data Locality (Distributed Input Data)?
- Metadata Aspects and Reproducibility for the DLC: metadata mappings, full provenance and lineage information, PIDs

European Landscape & Components
EUDAT GEF & EGI





WP7-JRA2 IS-ENES/C4I Pilot
Generic Use Case

Objective: Generate a multi-model multi-scenario time series average of the surface temperature using CMIP5 data

Scientific Workflow

- Spatially average over Western Europe (continents only)
- Time Period 1950-2100
- RCP 8.5 GES scenario
- All Global Climate Models available
- All members available
- Calculate the average time series
- Calculate the standard deviation
- Extract separate time series of every simulation
- Plot all those time series on a single graph