DARE

Delivering Agile Research Excellence on European E-Infrastructures

3rd ENES Workshop on Workflows funded by ESiWACE 13-14 Sep 2018
Brussels, Belgium

Interfacing e-infrastructures to cope with large data volumes for end-users of climate data

Christian Pagé

Research Engineer / Climate Research Domain

Toulouse, France

CERFACS

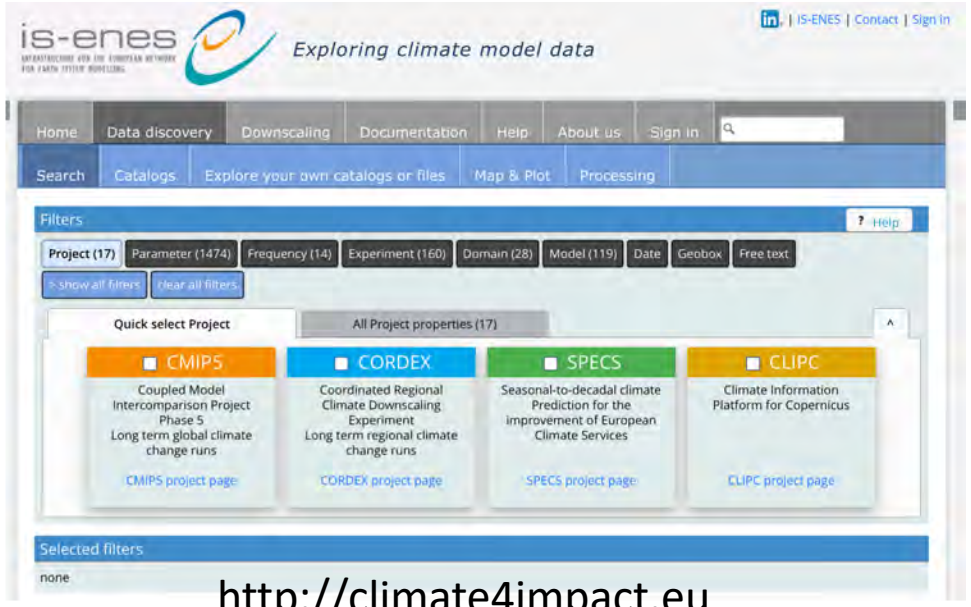
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

EUDAT Collaborative Data Infrastructure

DEMOKRITOS INGV THE UNIVERSITY of EDINBURGH CERFACS UNIVERSITY OF LIVERPOOL Koninklijk Nederlands Meteorologisch Instituut Ministerie van Infrastructuur en Milieu grnet Fraunhofer ATHENA

Motivations: Societal

- Provide climate projections data to climate change impact researchers, facilitators, practitioners
  - Ease data access with better intuitive interfaces
  - Provide more common data formats
  - Generate tailored products from data processing **workflows**

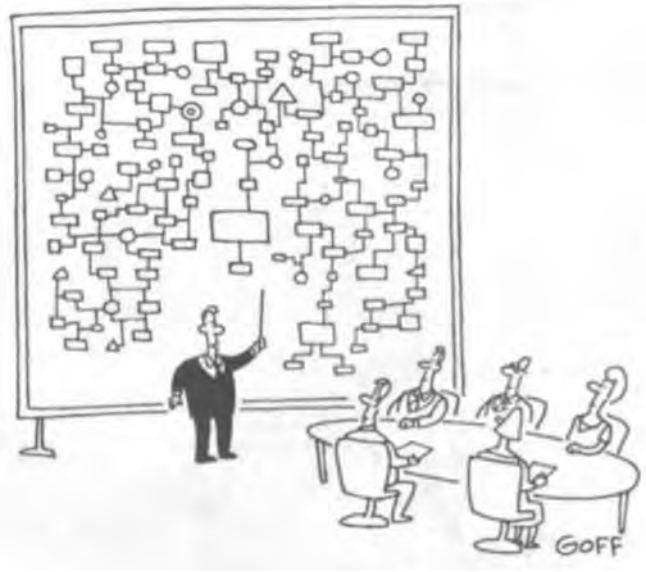


http://climate4impact.eu

Motivations: Scientific

- 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

- Achieve a more robust and flexible **Data Life Cycle**
  - More robust experiments setup
    - Explore several experiment configurations to answer scientific questions
  - **Reproducible** experiments

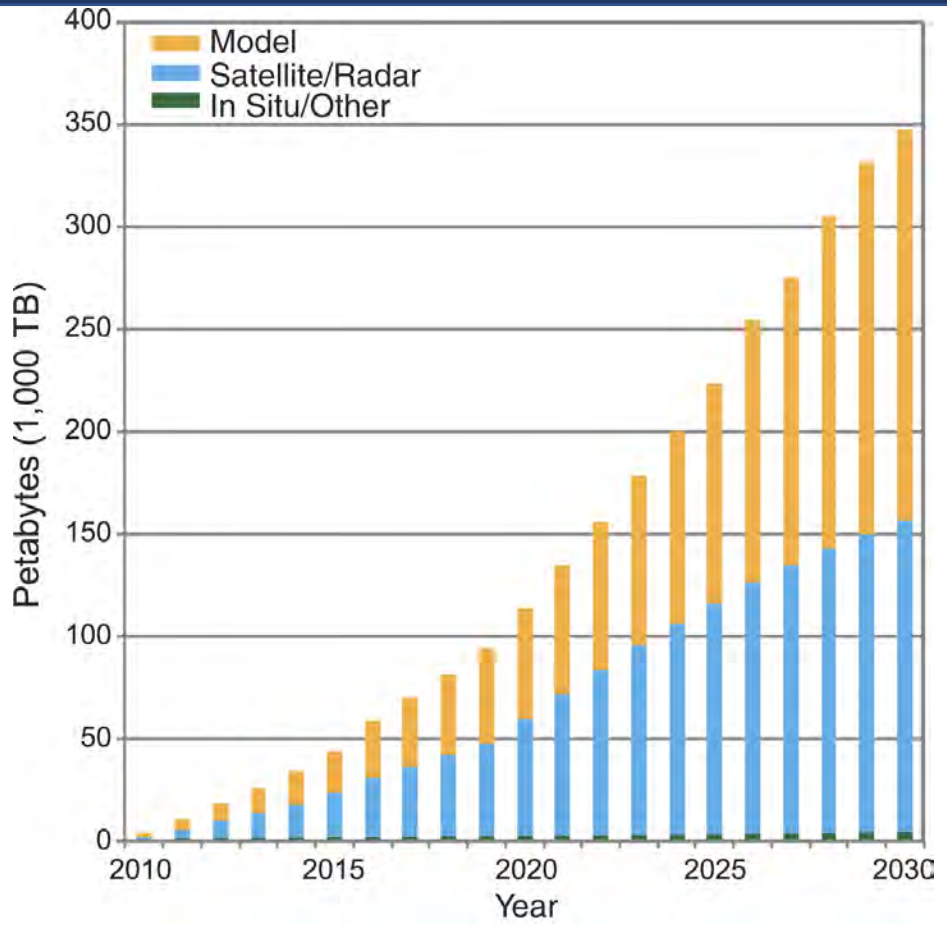


Motivations: Scientific, Technical, Societal

Technical

- Process large data volumes, ideally near(er) the data storage
  - Data Analytics
  - Data Life Cycle
- Streamline the data processing workflow
- Proper metadata description of the data objects
- Properly track provenance information
- Interconnect e-infrastructures and research infrastructures services, interfaces & platforms
  - EUDAT <=> ESGF

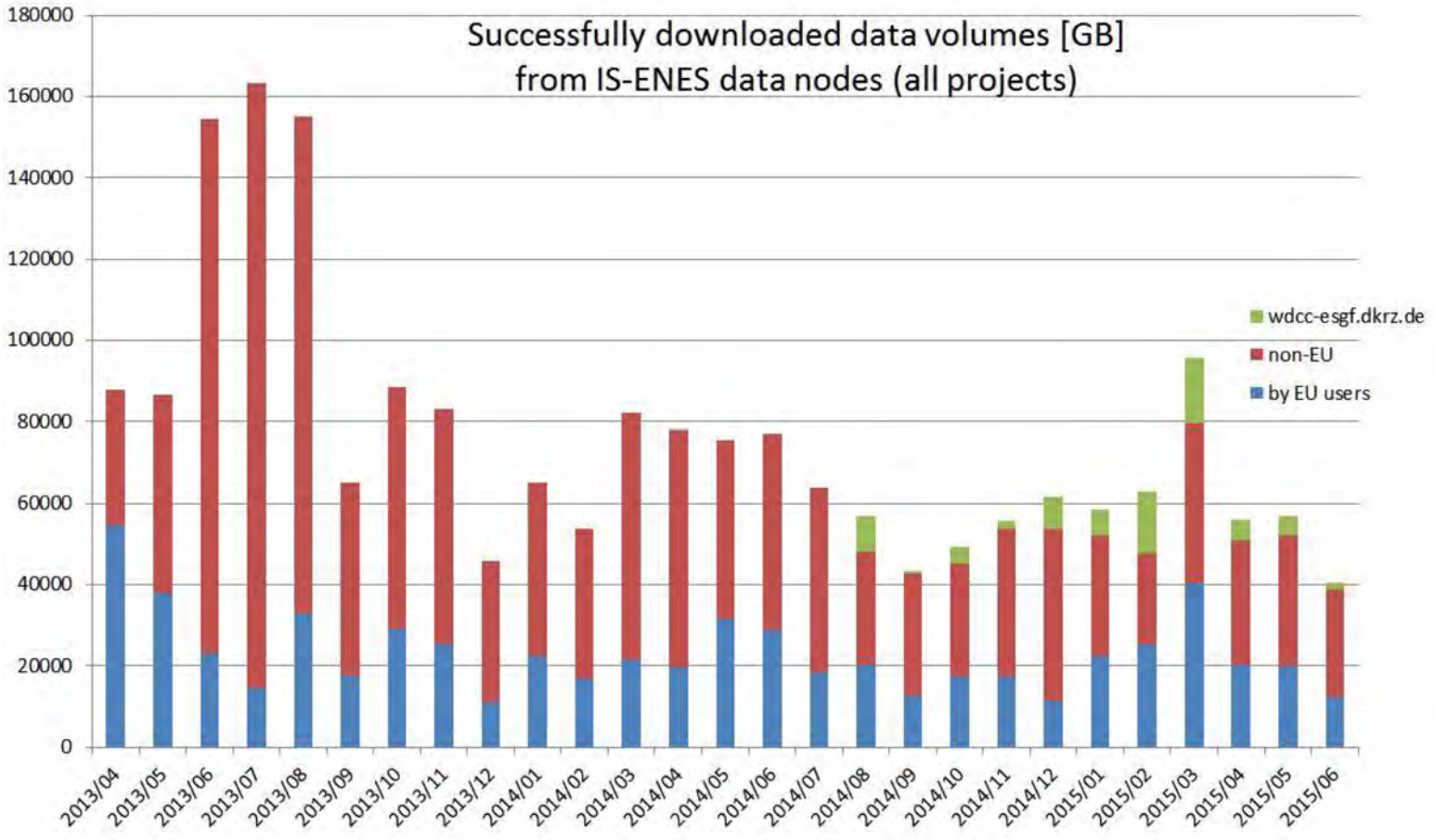
Current situation



Projected increase in global climate data for climate models, remotely sensed data, and in situ instrumental/proxy data. From Overpeck et al. Science, 2011

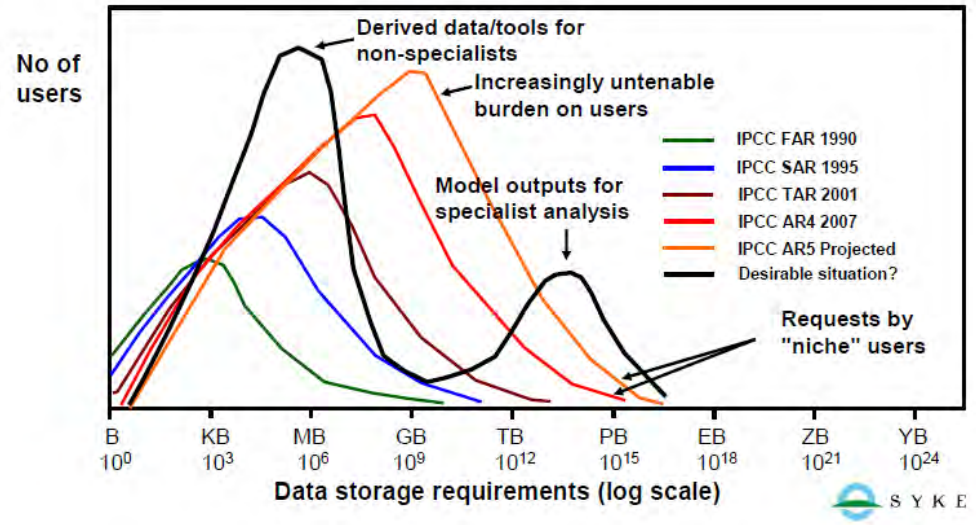
Current situation

Downloaded data volumes – European ESGF data nodes

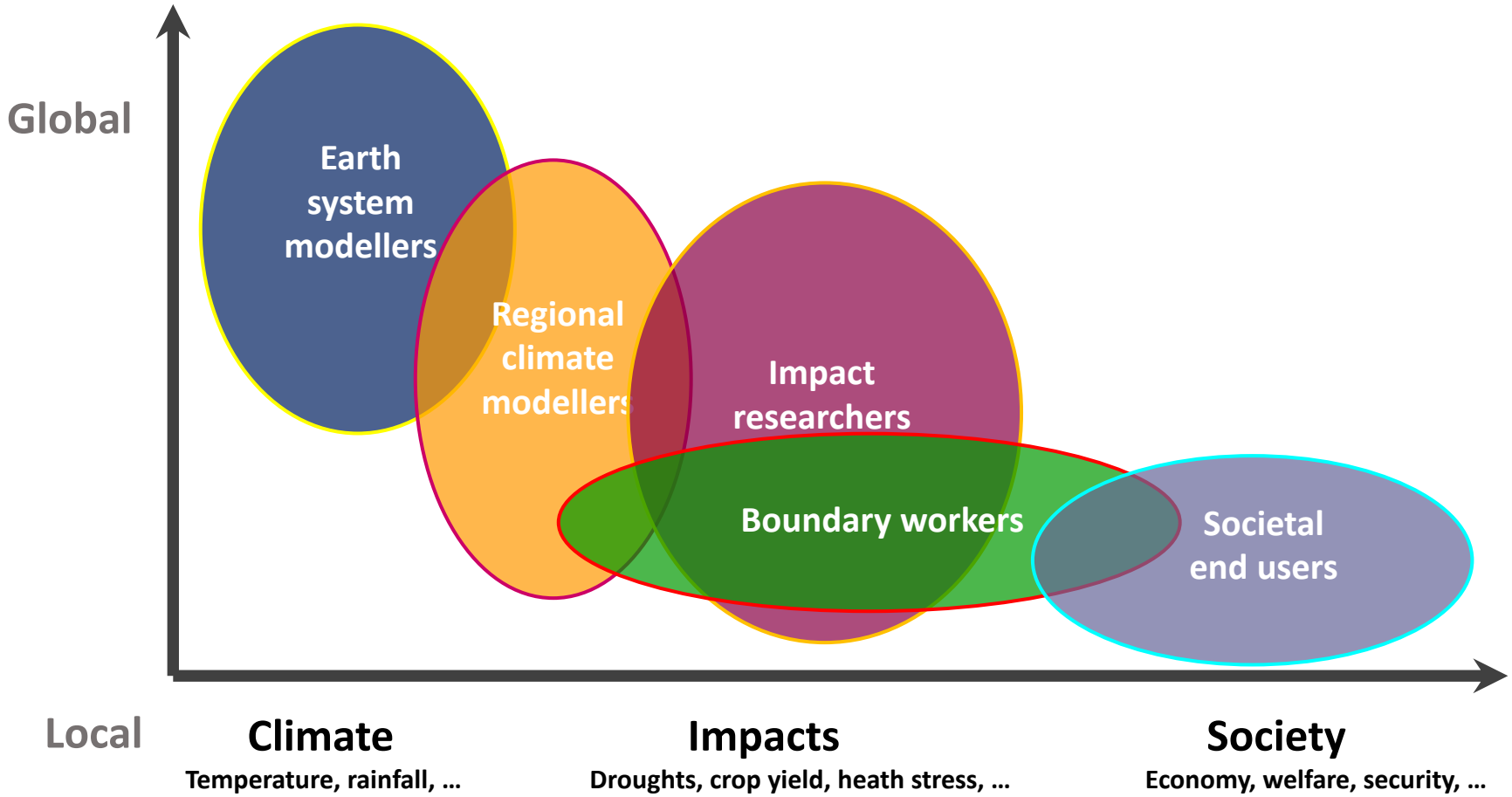


Climate Data Users: Current situation

- **Data available for scientific analysis**: a very large trend
  - Limitations in data access means limitations in data analytics and scientific results
- **Download locally then Analyze**: a workflow that cannot be sustained
  - Climate researchers
  - Impact researchers



Climate Data Users: Current situation

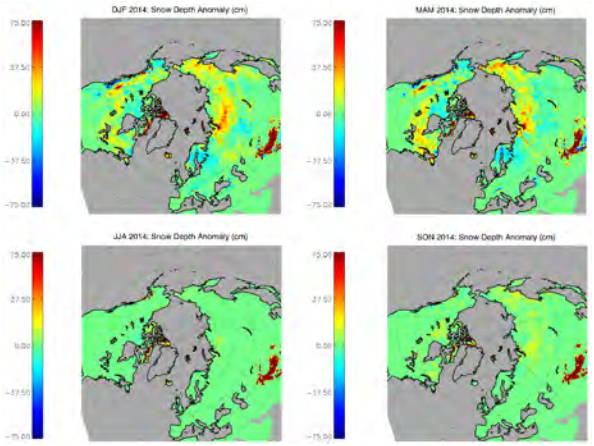


Lars Bärring, SMHI Rossby Centre, Circle-2 Conference on European Climate Change Adaptation Research and Practice, Lisbon, 10-12 March 2014

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**



- 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

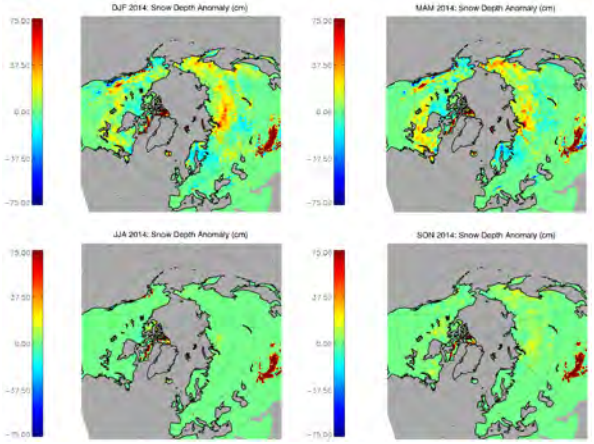
- ~11 200 files of ~50 Mb each per field

TOTAL: ~560 Gb

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**



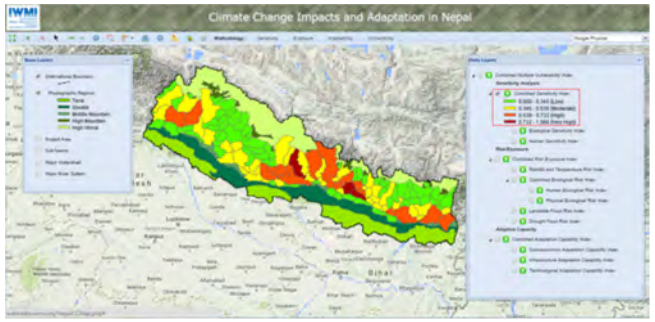
Needs and questions

- 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.**



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

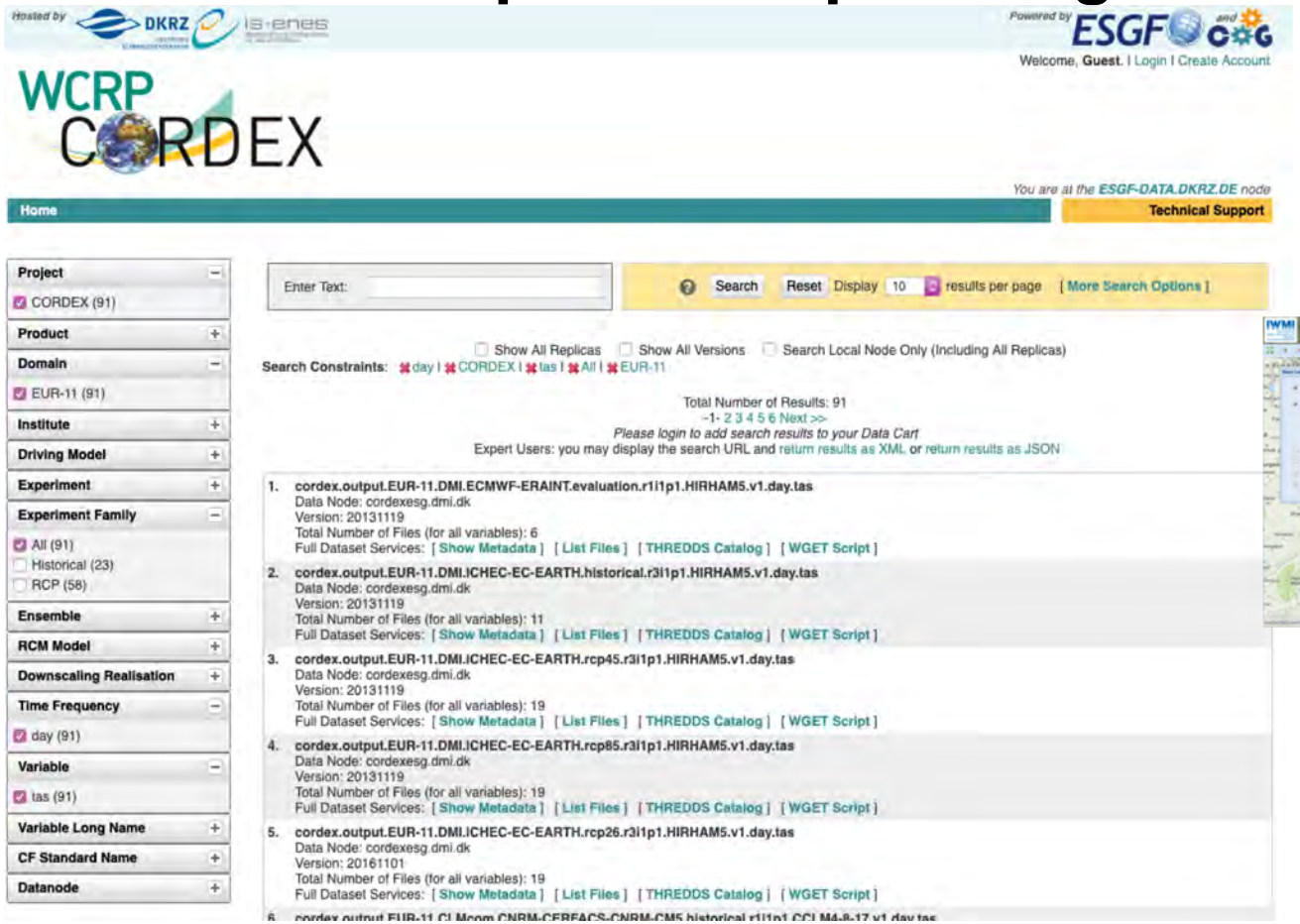
- Historical + All RCPs:
- Combination of models an ensemble members
- EUR-11 Euro-Cordex Grid

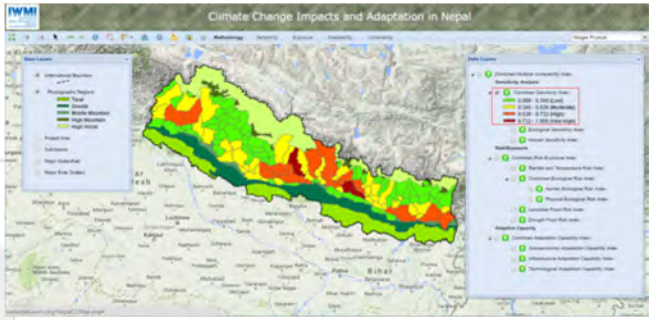
- 1378 files of ~600 Mb each per field

TOTAL: ~5 Tb

Climate Data Users: Current situation

Practical Example: An Impact Engineer

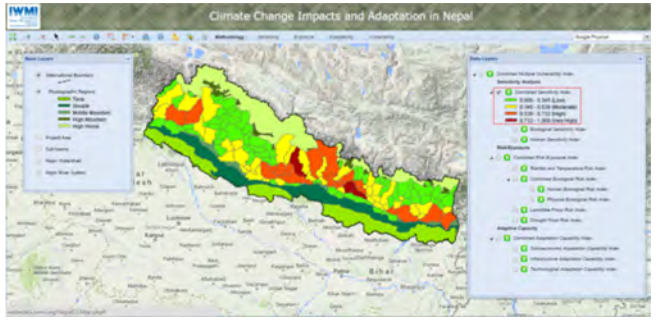




Climate Data Users: Current situation

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.**



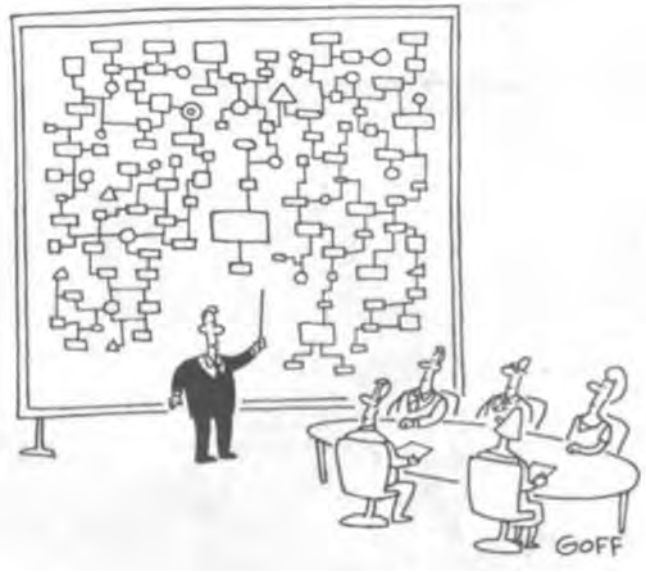
Needs and questions

- 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



CERFACS

DARE

Climate Data Distribution: ESGF RI

ESGF Data Nodes 2015

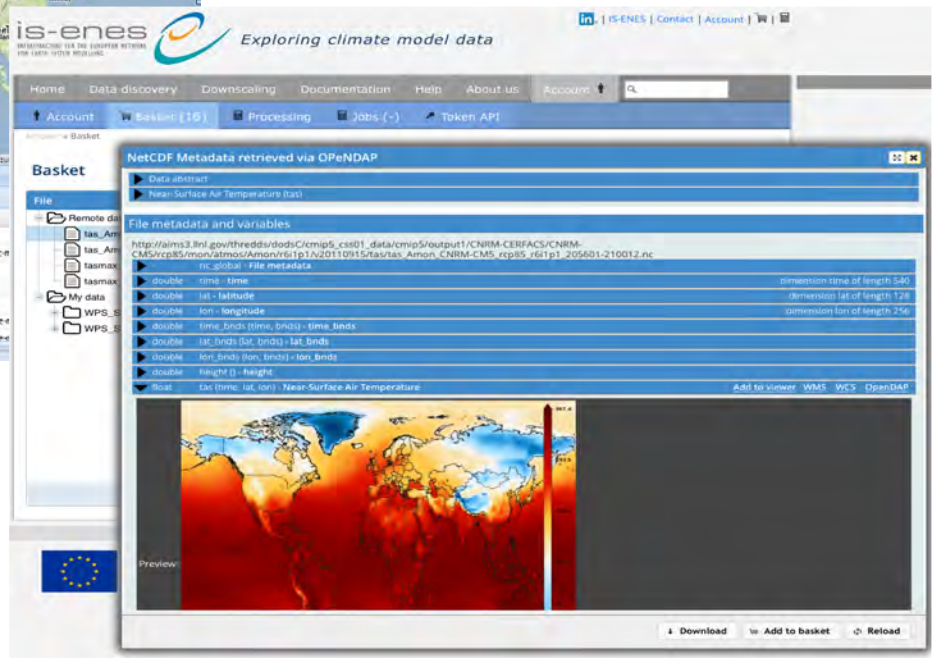
- 40 worldwide
- 18 in Europe
  (coordinated in IS-ENES)

	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**

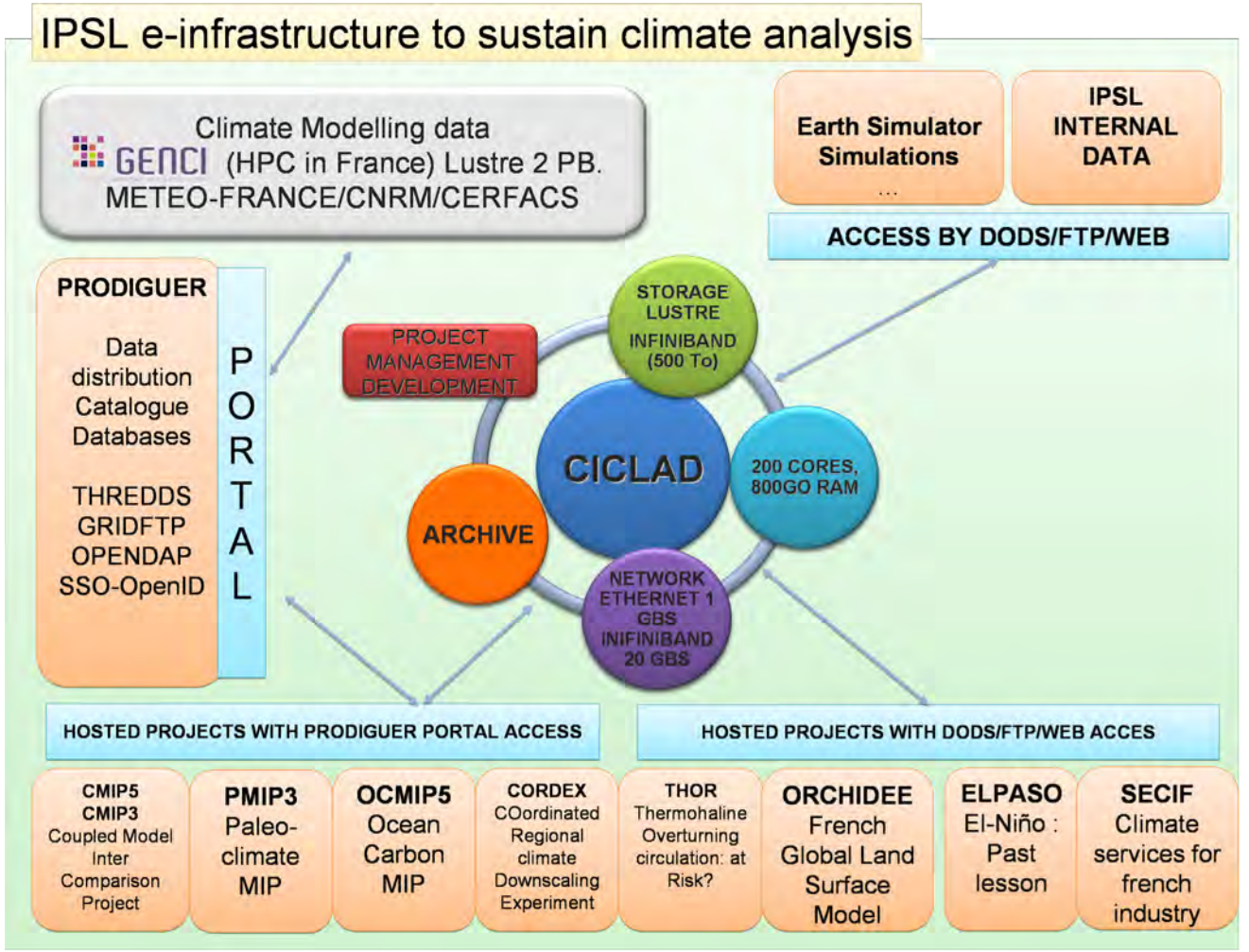
Courtesy from S. Denvil, IPSL

IS-ENES CDI C4I

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



Available Solutions: CICLAD



Courtesy from S. Denvil, IPSL

Available Solutions: JASMIN Analysis Platform

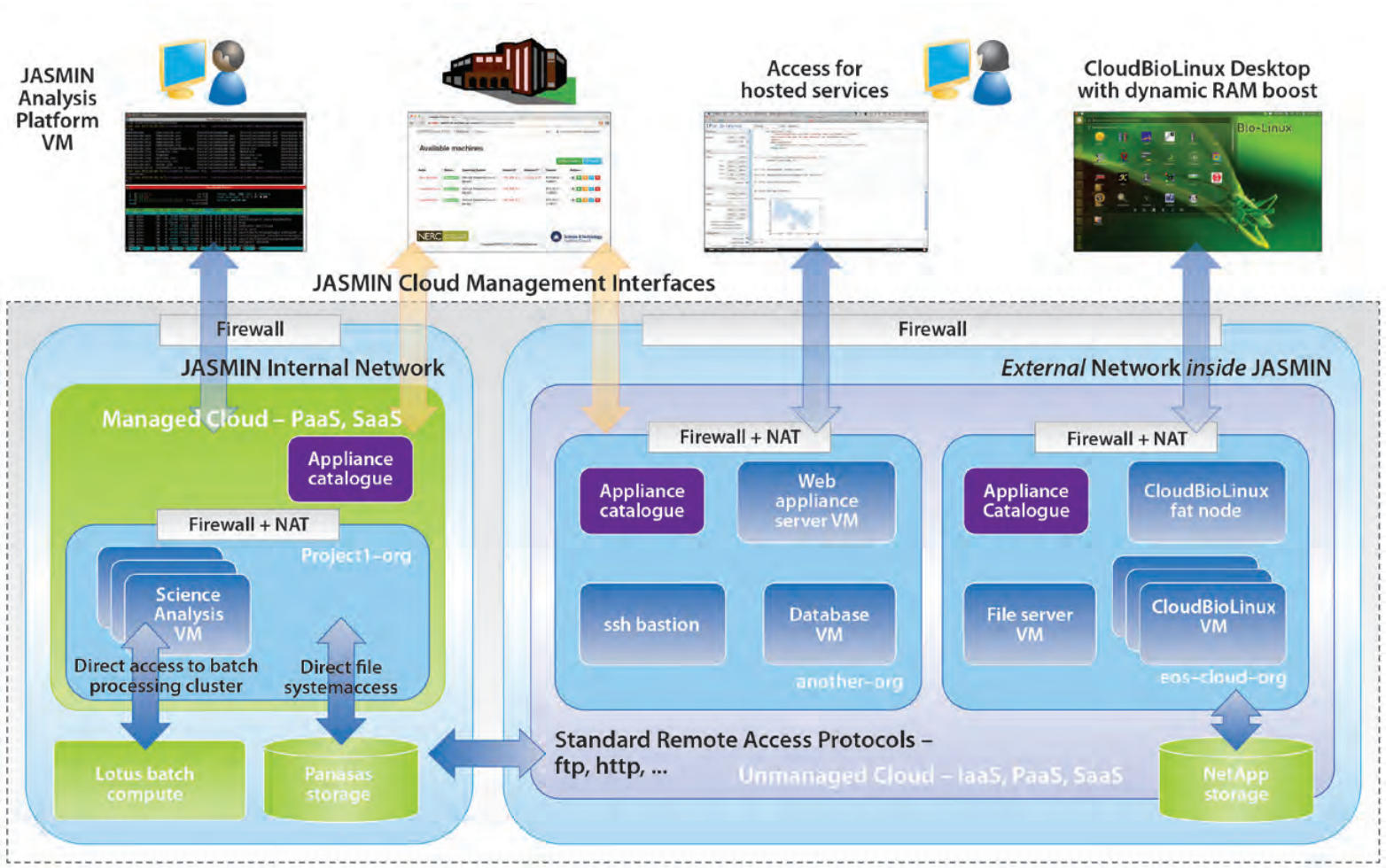
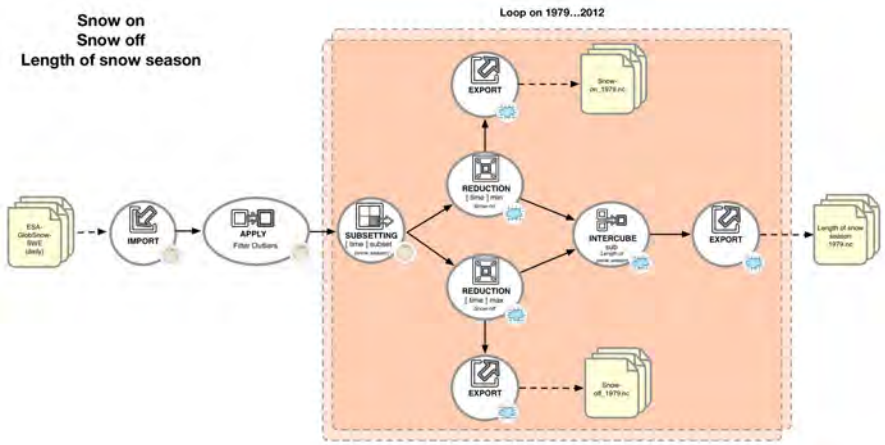


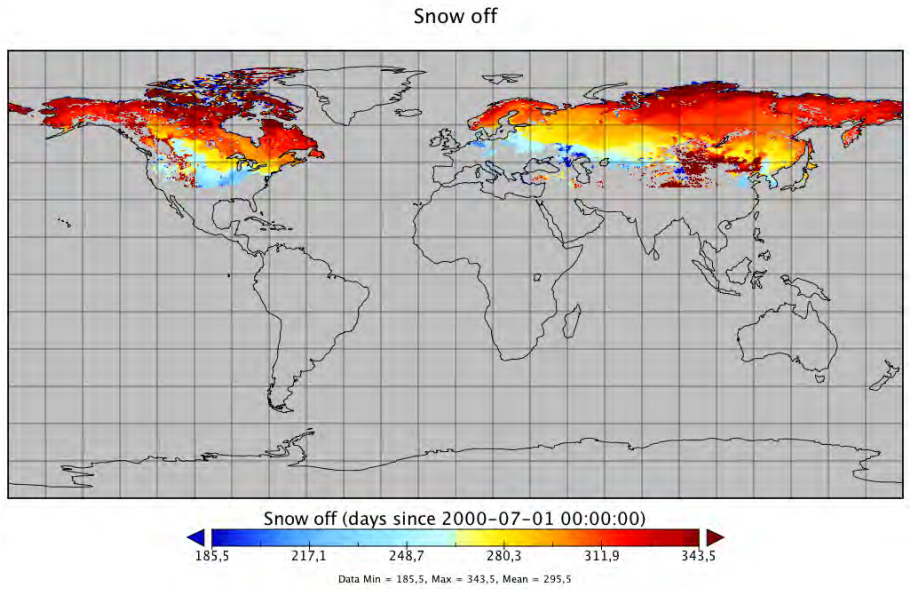
Fig. 13. CEDA's JASMIN analysis platform. *JASMIN integrates cloud architecture, container technologies, and virtual machines to improve flexibility and performance and track maintenance.*

Available Solutions: Ophidia/ECAS

Snow on/off – length of snow season

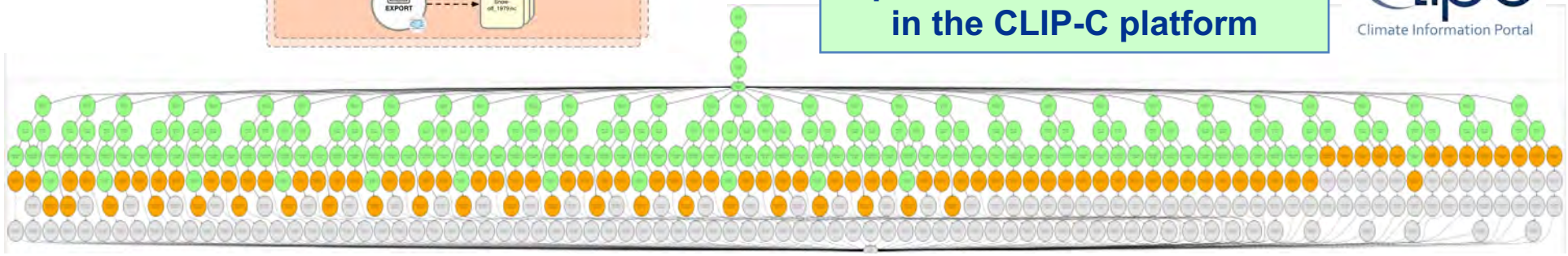
Dataset time range: 1979-2012
***6341** nc files**, 50 GB** of input data*
***99** NetCDF output files (**6MB** each)*
***21** tasks in the experiment description*
***599 tasks** performed at runtime!*





Indicators produced using Ophidia are available online in the CLIP-C platform

Climate Information Portal

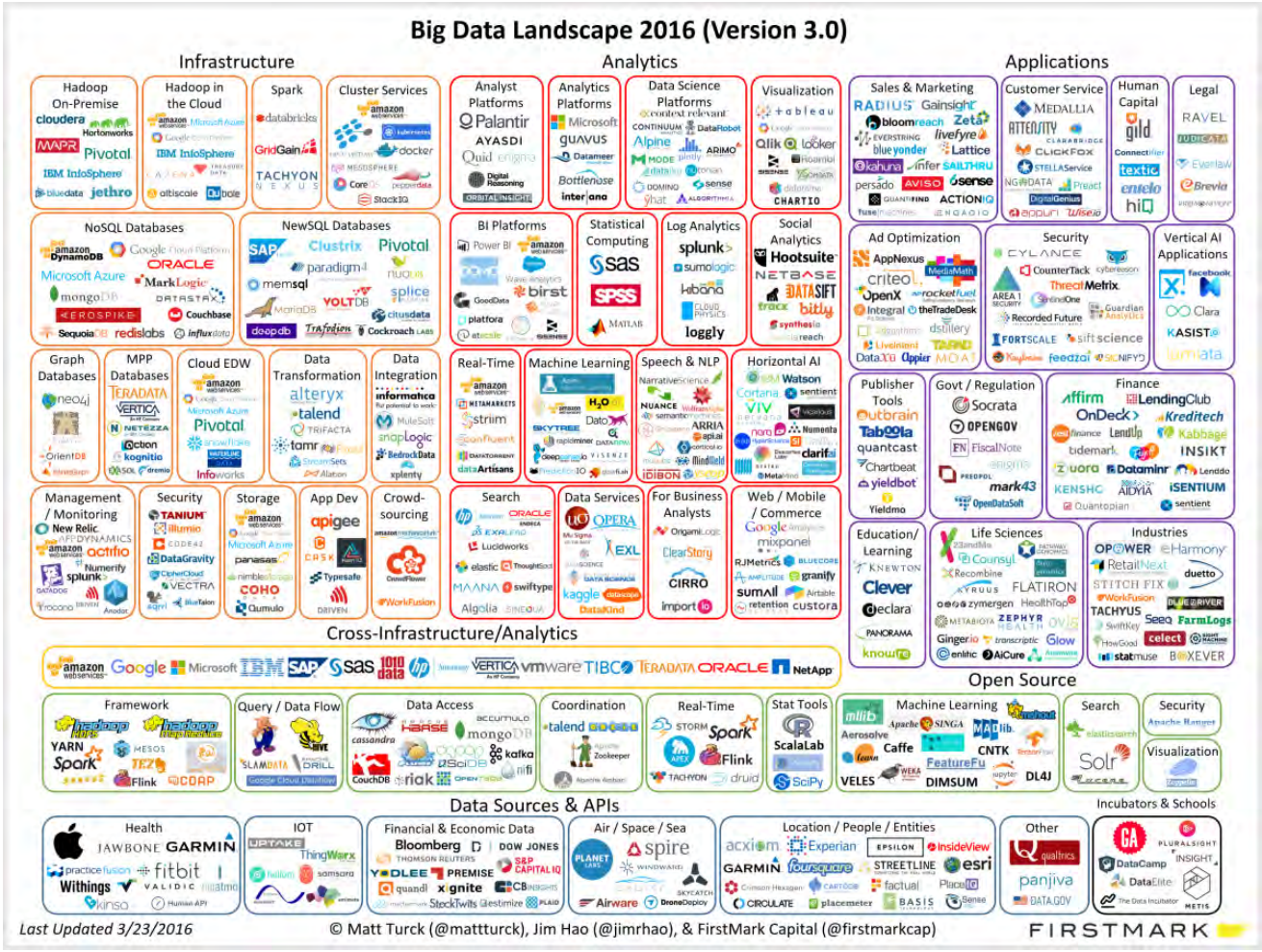


Courtesy from S. Fiore, CMCC

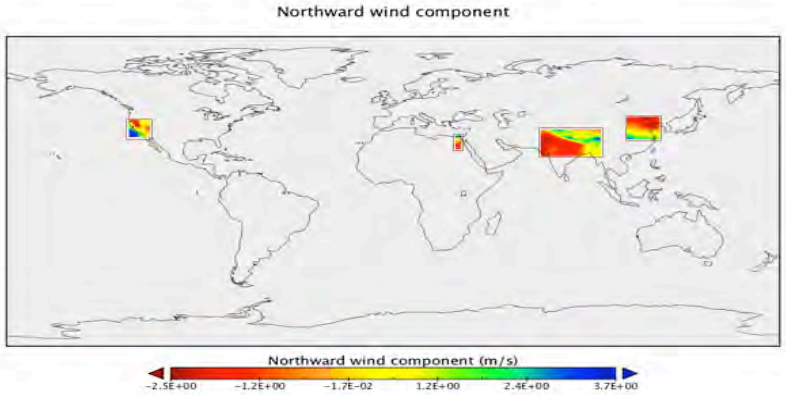
***Youtube video**: https://www.youtube.com/watch?v=RtZ6Ekh-fl0*

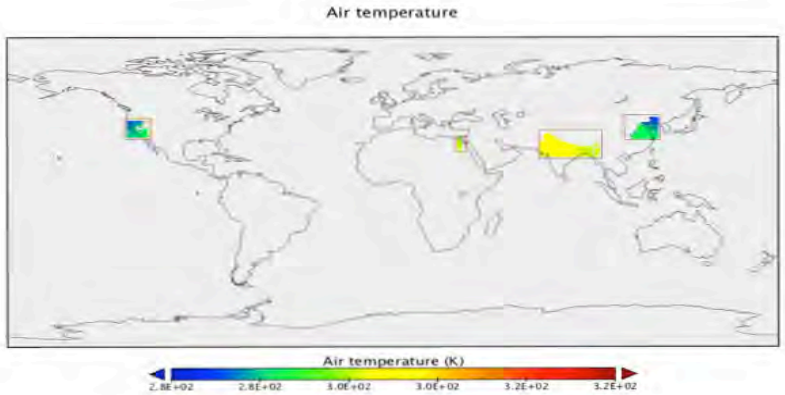
Big Data?

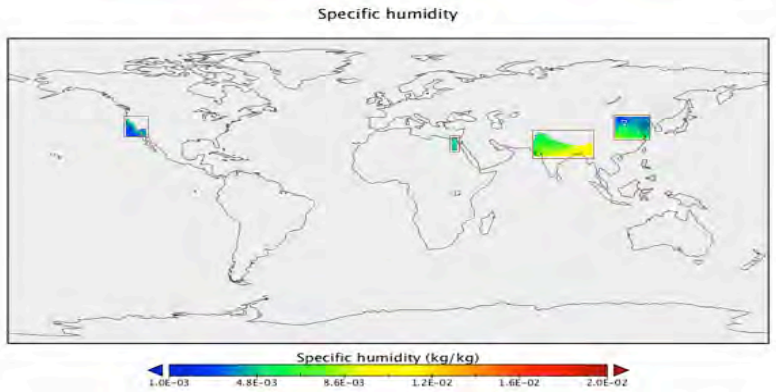
What about Big Data Technologies and Analytics??



Big Data: Hadoop and Climate Data @NASA







Wei, et al.

- ~8.4 TB transferred from archive to local workstation (weeks)
- Clipping, averaging performed by Fortran program on local workstation (days)

MERRA/AS

- Clipping, averaging performed by MERRA/AS (~28 hrs)
- Only ~35 GB final product transferred to local workstation (minutes)

- Significant time savings in data wrangling,
- rapid screening over monthly means files takes minutes, and
- there's a possibility of folding Dr. Wei's modeling algorithm back into the CDS API …

Applying Apache Hadoop to NASA's Big Climate Data: Glenn Tamkin, John Schnase, Dan Duffy, Hoot Thompson, Denis Nadeau, Scott Sinno, Savannah Strong,

CERFACS

DARE

Big Data: Hadoop and Climate Data @NASA

The original MapReduce application utilized standard Hadoop Sequence Files. Later they were modified to support three different formats called Sequence, Map, and Bloom.

Dramatic performance increases were observed with the addition of the Bloom filter (~30-80%).

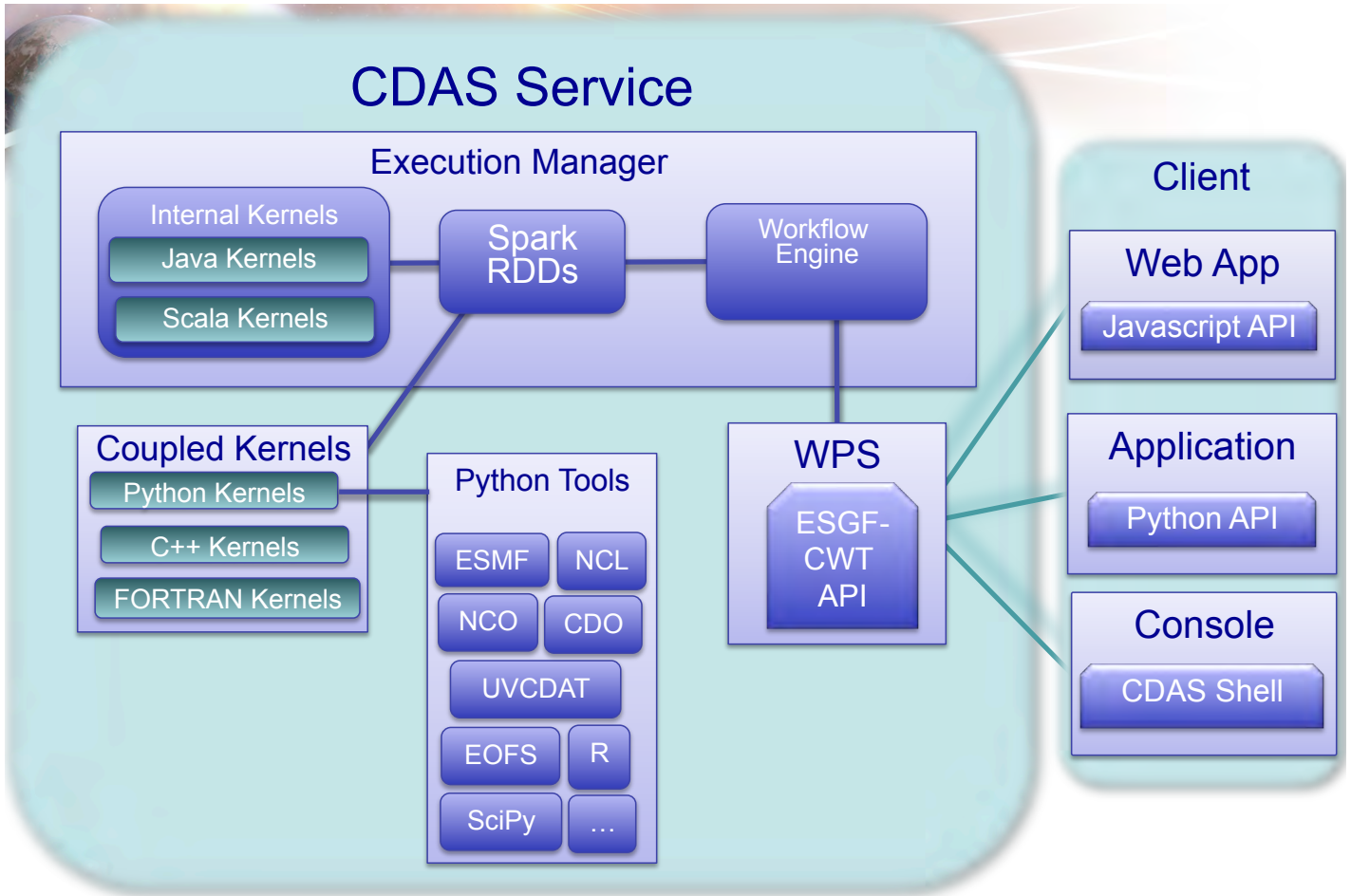
Job Description	Host	Sequence (sec)	Map (sec)	Bloom (sec)	Percent Increase
Read a single parameter (“T”) from a single sequenced monthly means file	Standalone VM	6.1	1.2	1.1	+81.9%
Single MR job across 4 months of data seeking “T” (period = 2)	Standalone VM	204	67	36	+82.3%
Generate sequence file from a single MM file	Standalone VM	39	41	51	-30.7%
Single MR job across 4 months of data seeking “T” (period = 2)	Cluster	31	46	22	+29.0%
Single MR job across 12 months of data seeking “T” (period = 3)	Cluster	49	59	36	+26.5%

Applying Apache Hadoop to NASA's Big Climate Data: Glenn Tamkin, John Schnase, Dan Duffy, Hoot Thompson, Denis Nadeau, Scott Sinno, Savannah Strong,

CERFACS

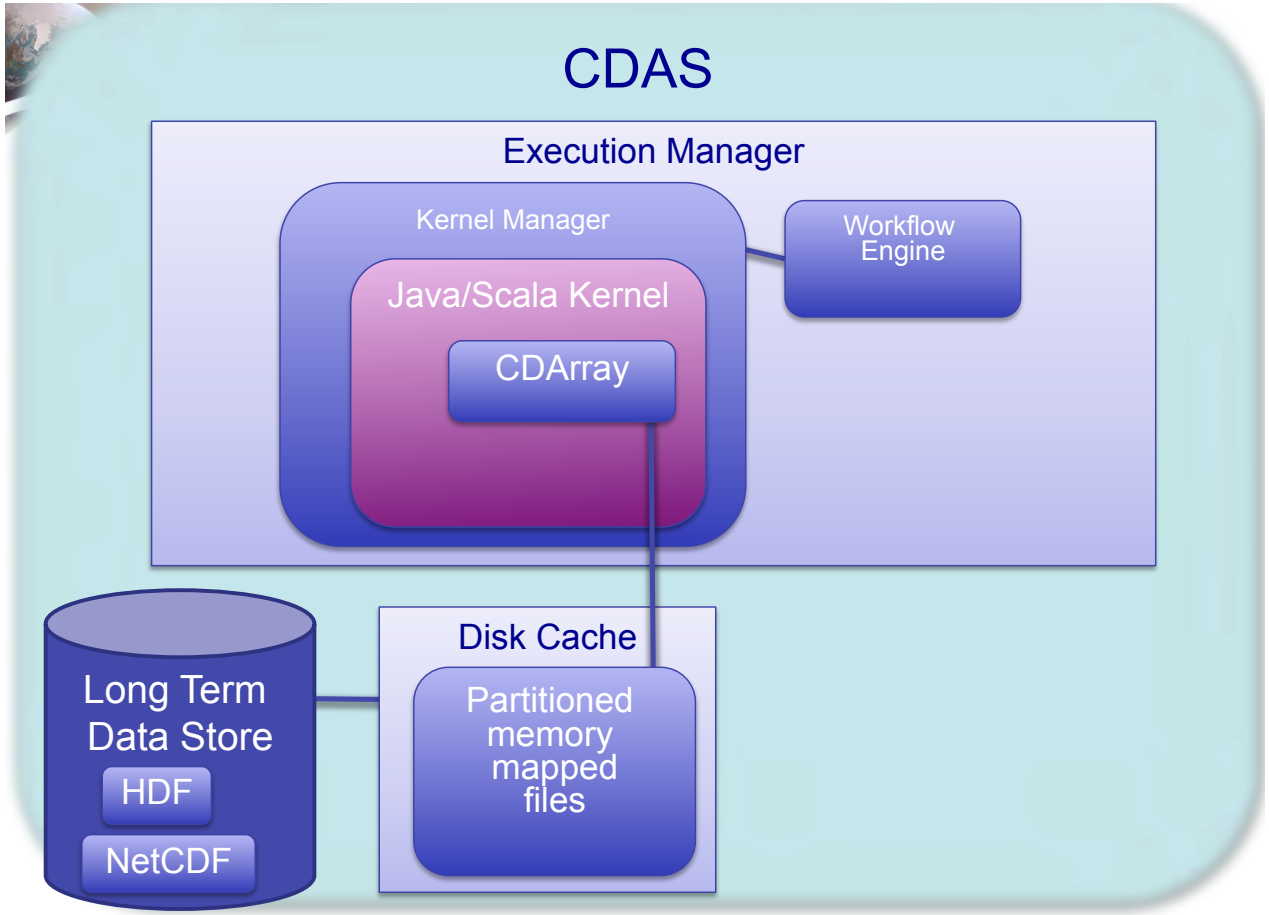
DARE

Big Data: Spark & Hadoop / CDAS @NASA



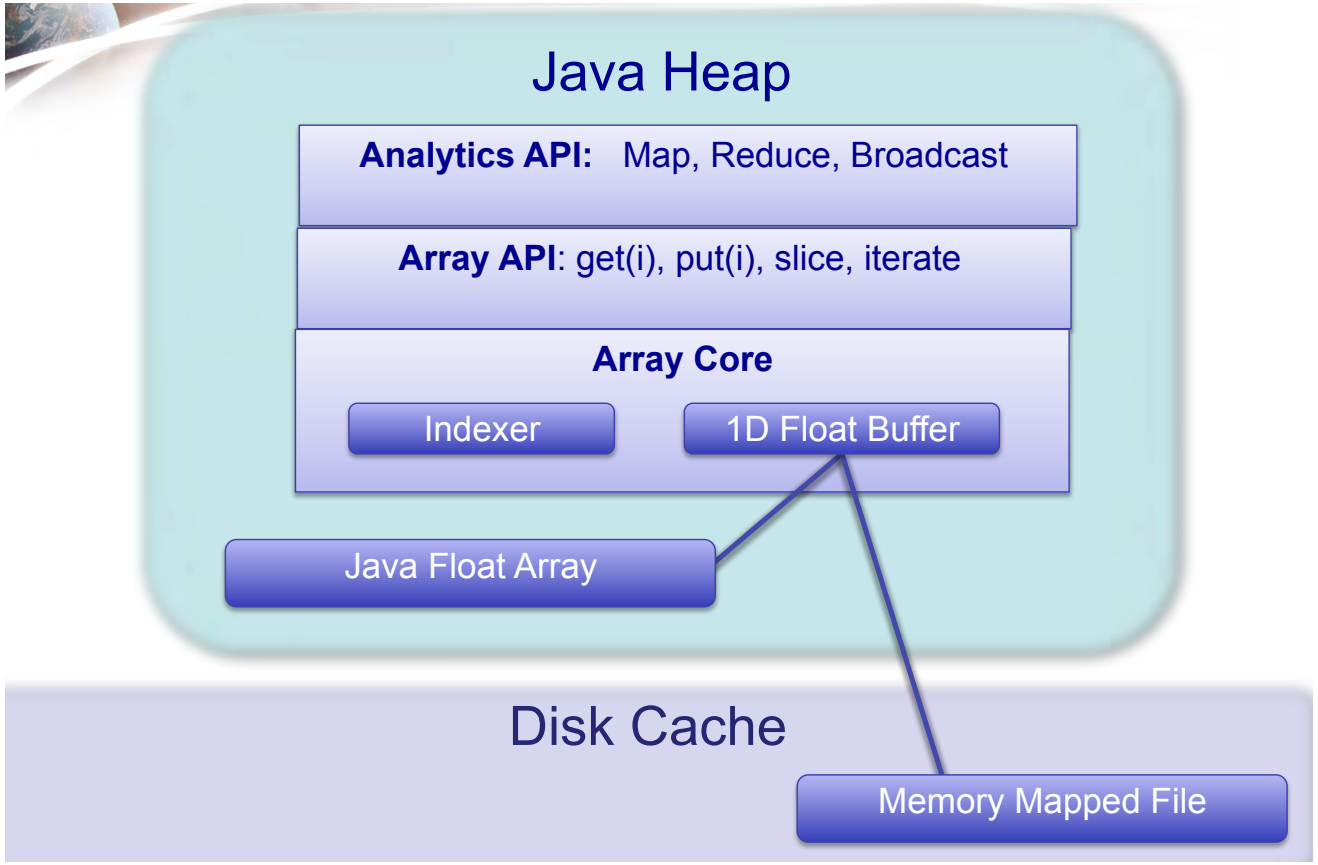
Climate Data Services Framework (CDAS). Thomas Maxwell and Dan Duffy. NASA.

Big Data: Spark & Hadoop / CDAS @NASA



Climate Data Services Framework (CDAS). Thomas Maxwell and Dan Duffy. NASA.

Big Data: Spark & Hadoop / CDAS @NASA

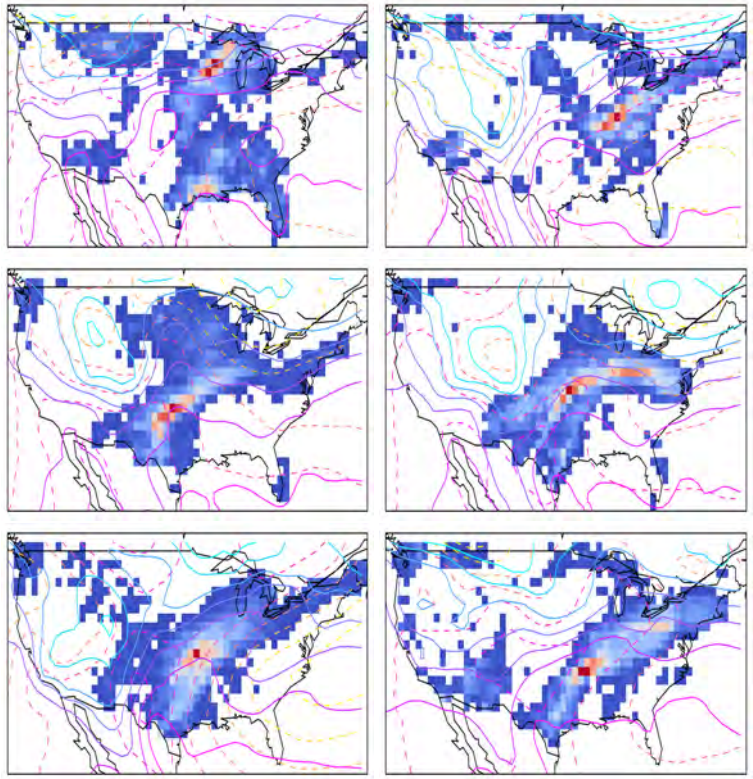


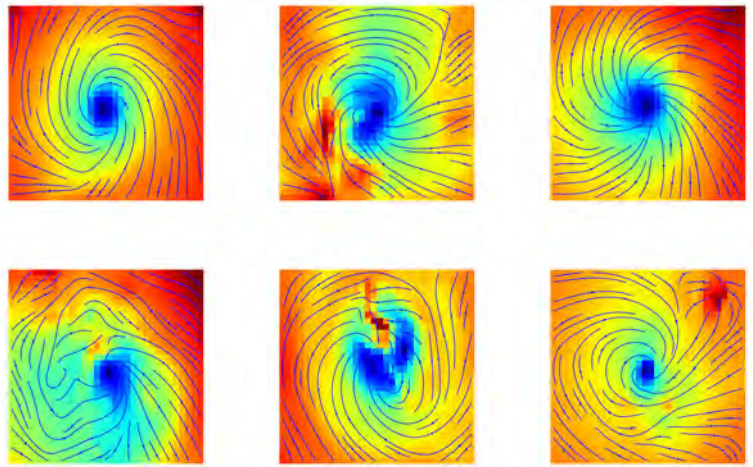
Climate Data Services Framework (CDAS). Thomas Maxwell and Dan Duffy. NASA.

CERFACS

DARE

Big Data Analytics on Climate Data





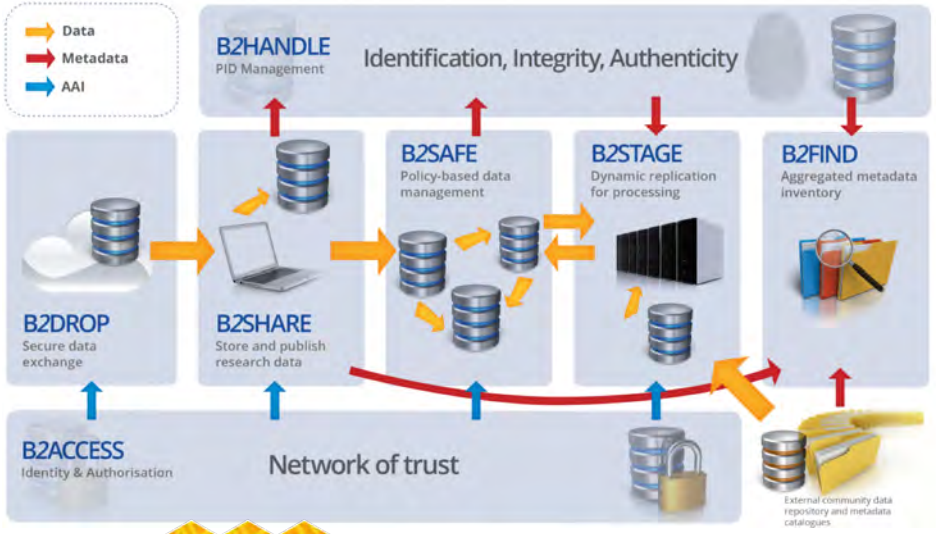
Weather fronts (left) and Tropical Cyclones (right) as detected by a convolutional neural network.

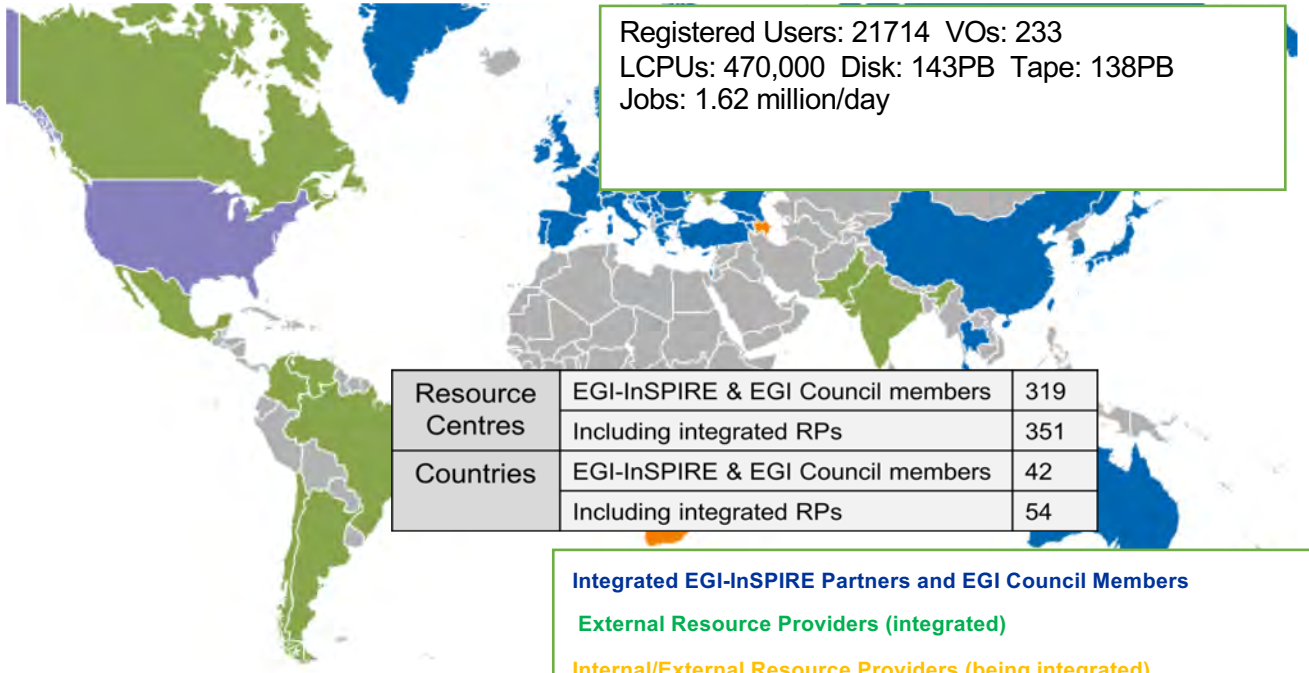
Liu et al. KDD 2016 August 13-17, San Francisco, CA, USA

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)





- Computing Power (FedCloud) resources

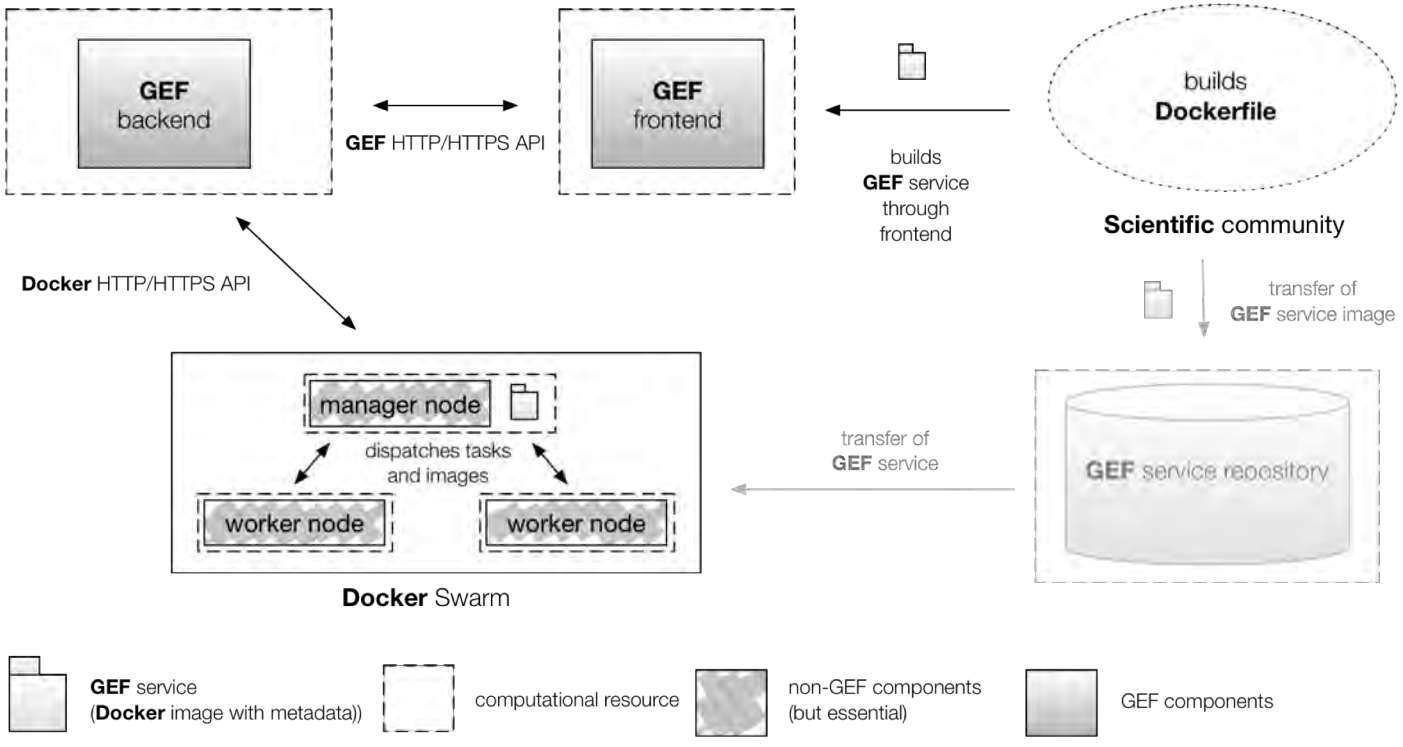
ESGF Compute Nodes

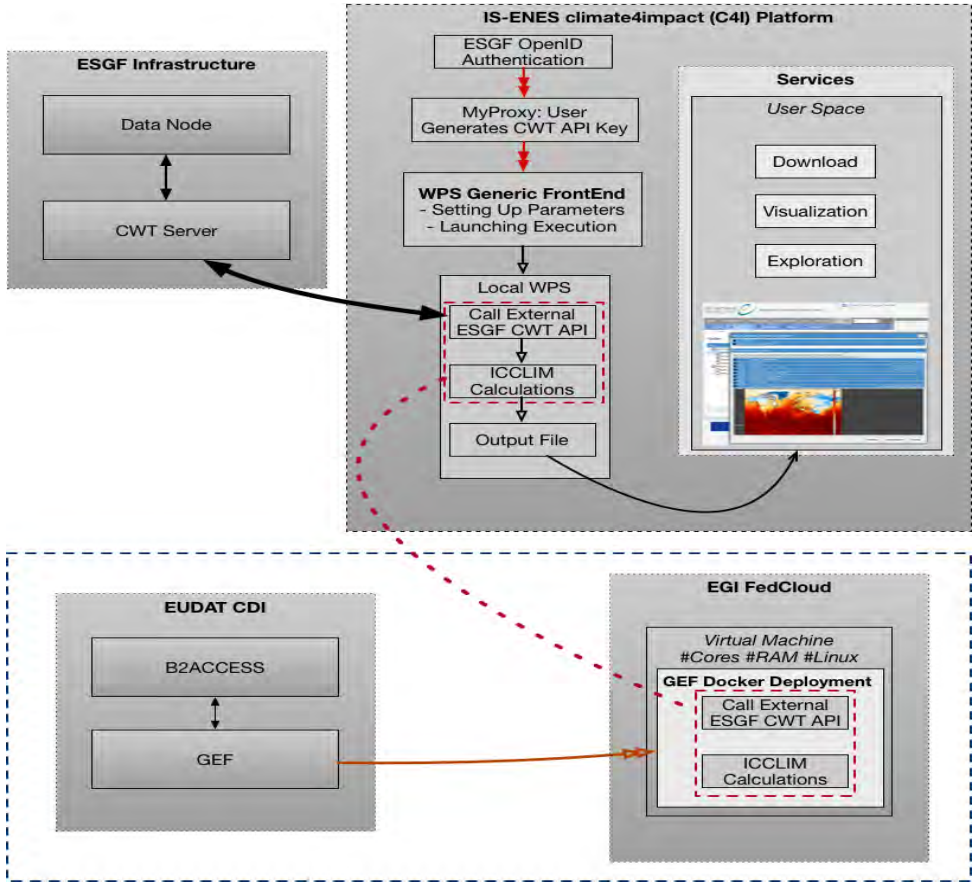
ESGF Future Computing Nodes: API

- **Goal**: perform data analysis near the data storage
  - Better data access
  - Move away from the download/analyze workflow

European Landscape & Components
EUDAT GEF & EGI

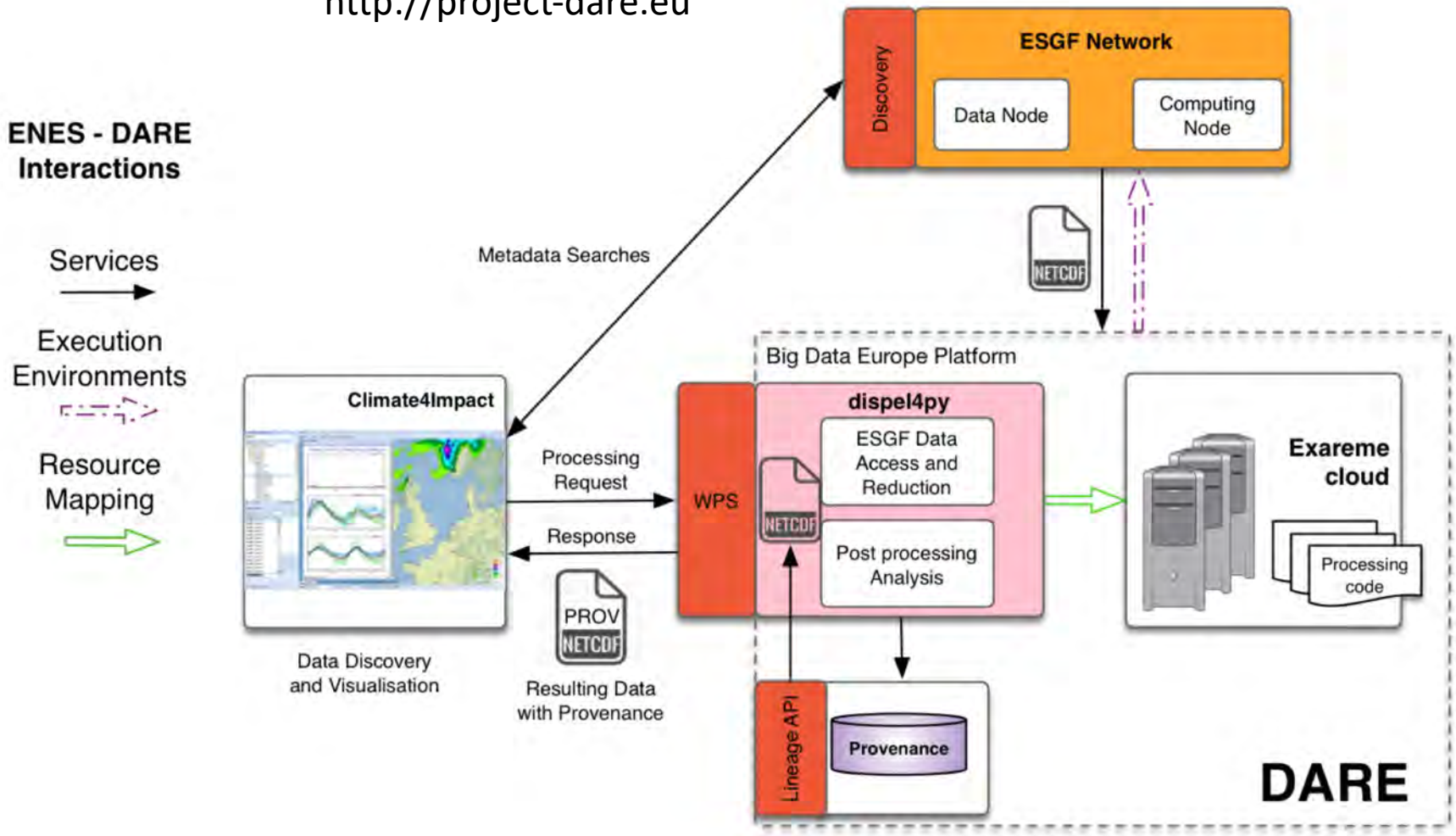
https://github.com/EUDAT-GEF/GEF





DARE IS-ENES Climate Use Case Draft Architecture

http://project-dare.eu



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

Questions & Comments! ☺

http://project-dare.eu

christian.page@cerfacs.fr