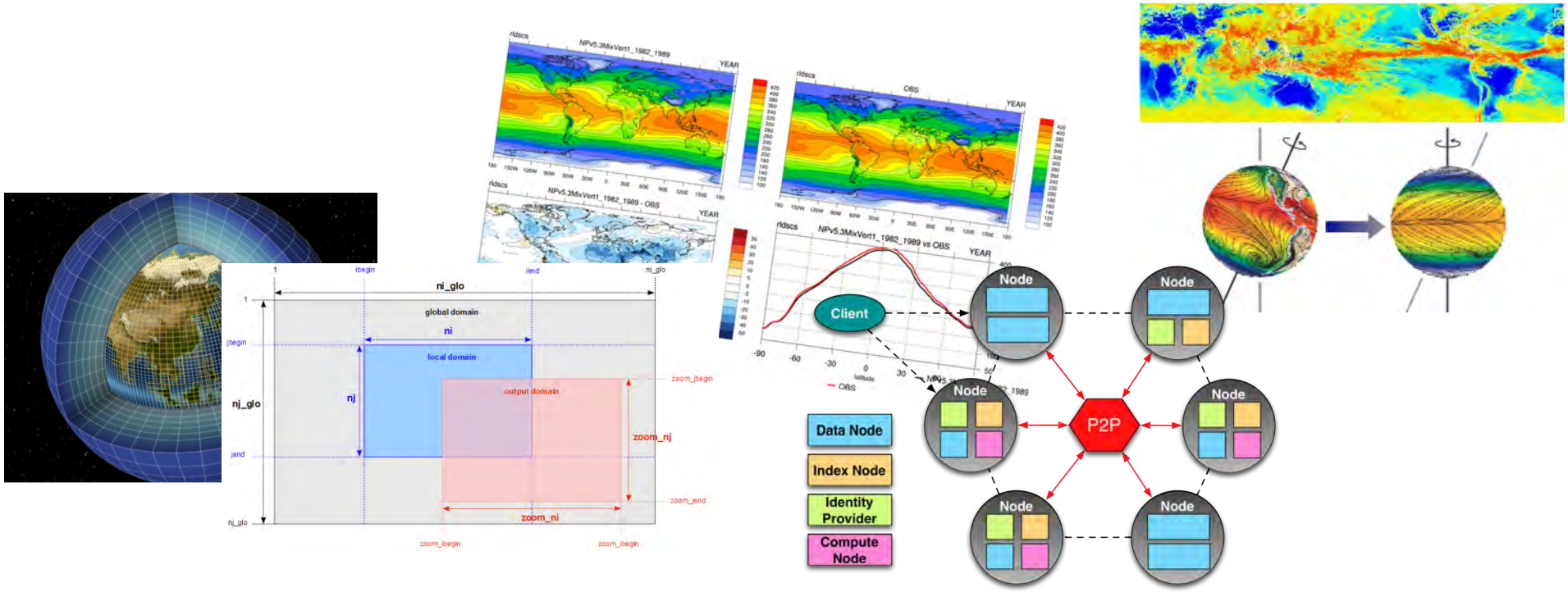
Earth System Grid Federation

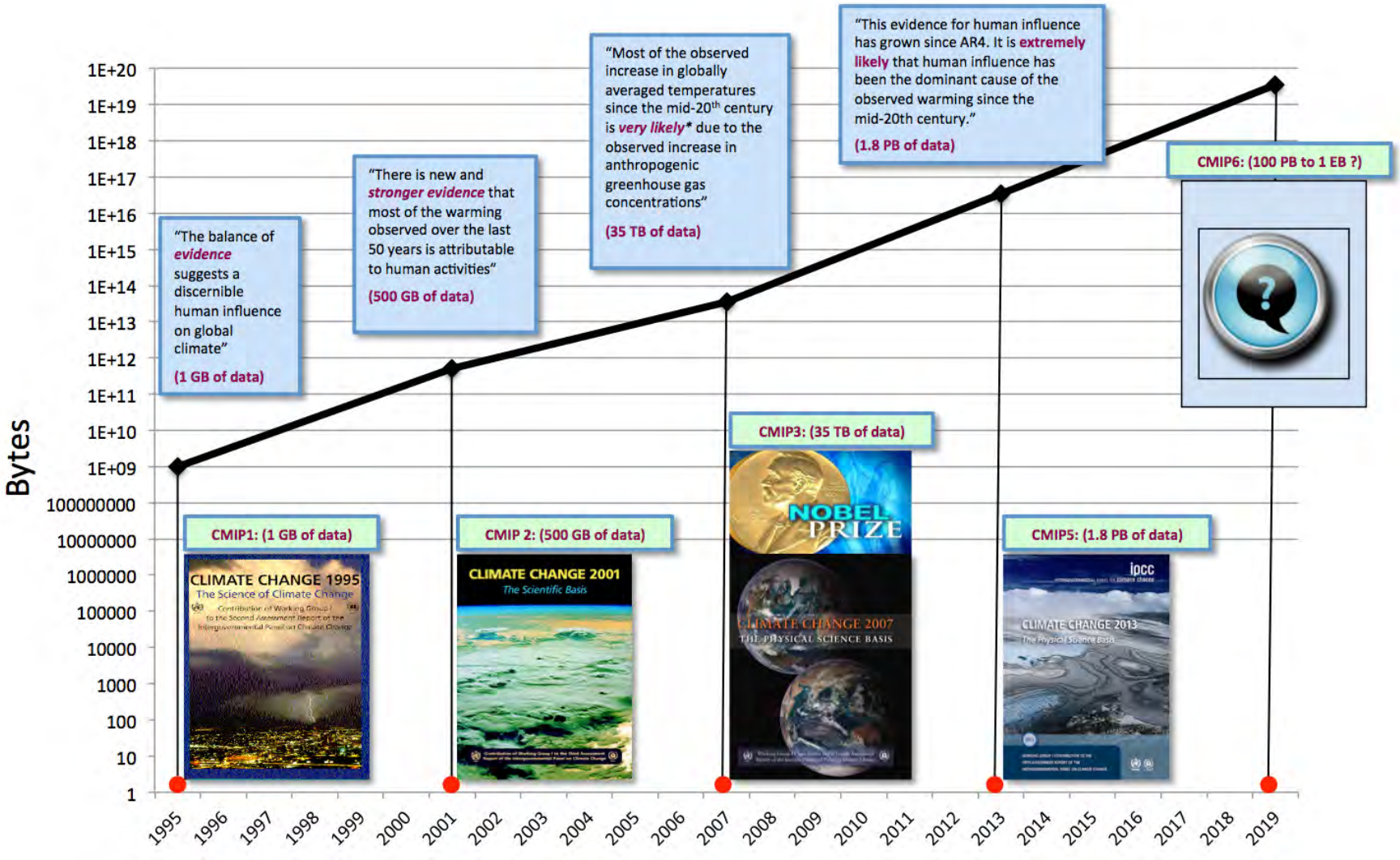
Sébastien Denvil (CNRS/IPSL). With contributions from ESGF Executive Committee and WGCM Infrastructure Panel



Climate Sciences Programs

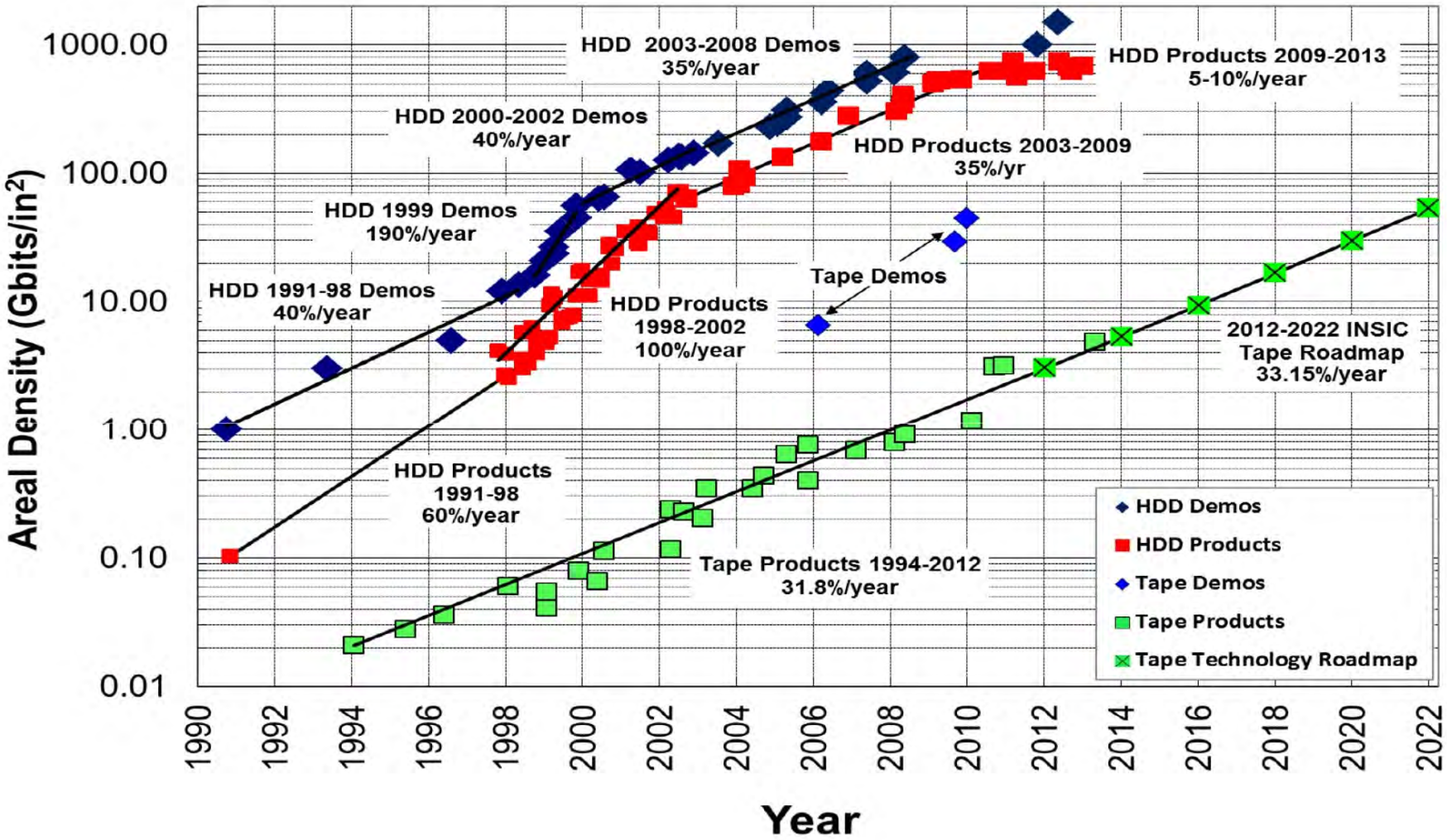
CMIP : +76 %/year

Moore's law : +42%/year



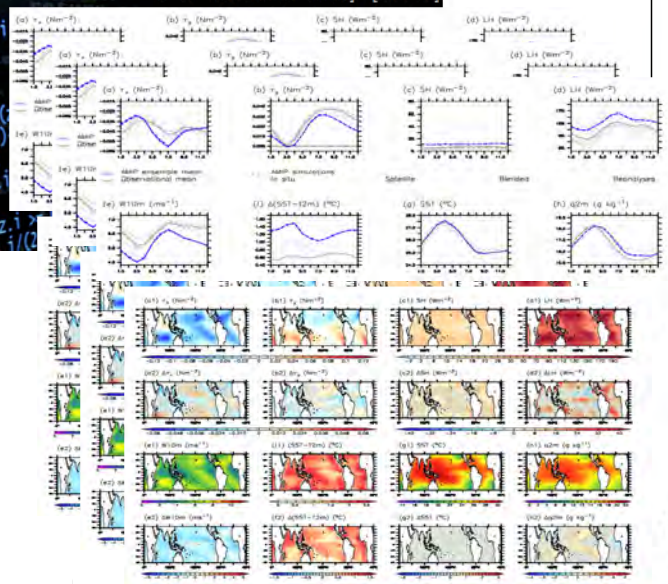
Climate Sciences Programs

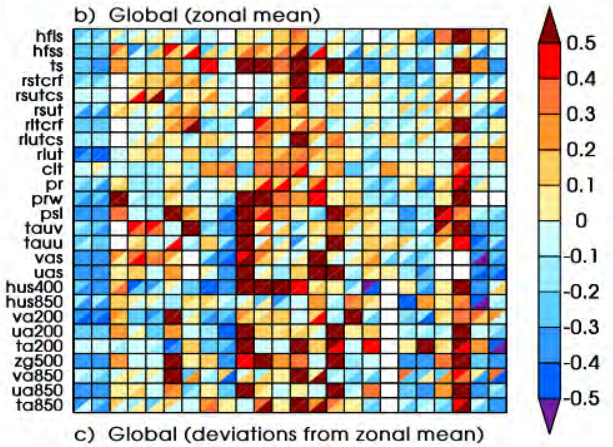
CMIP : +76 %/year
HDD : +45 %/year



Missions

Data ≠ Informations ≠ Knowledge





Global data infrastructure

CMIPs, and in general any science involving cross-model comparisons, critically depend on the global data infrastructure – the “vast machine” (Edwards 2010) – making this sort of data-sharing possible.

Data consumers

Scientists perform sequences of computations (e.g “poleward heat transport”, “length of growing season”) on datasets. Typically this is scripted in some data analysis language, and ideally it should be possible to apply the script to diverse datasets.

Data producers

Observational and model output data in the climate-ocean-weather (COW) community is initially generated in some “native” non-standard format, and any subsequent relative analyses requires considerable effort to systematise. Issues include moving and transient data sources, lossy data formats, curvilinear and other “exotic” coordinates.

Data organizers

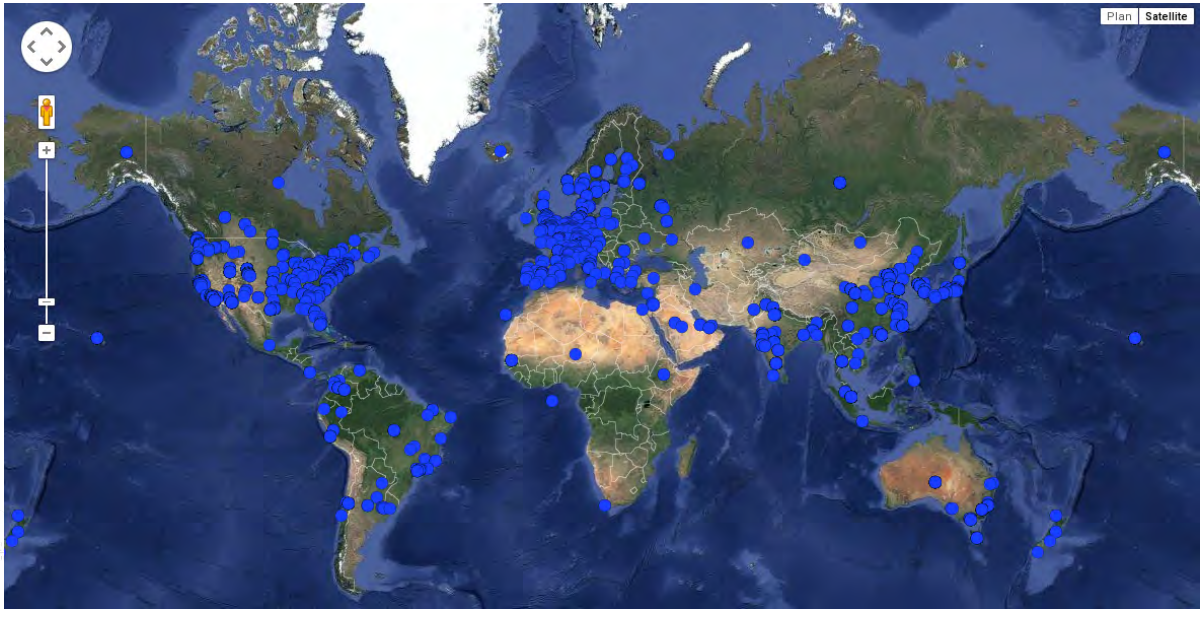
Data organizers are the community within this ecosystem that facilitates the transformation of source dependent data to a neutral and readily consumable form. They maintain the standards for describing data in a manner that permits these transformations, and develop tools to perform them.

ESGF Data Infrastructure

esgf.llnl.org

ESGF
Earth System Grid Federation

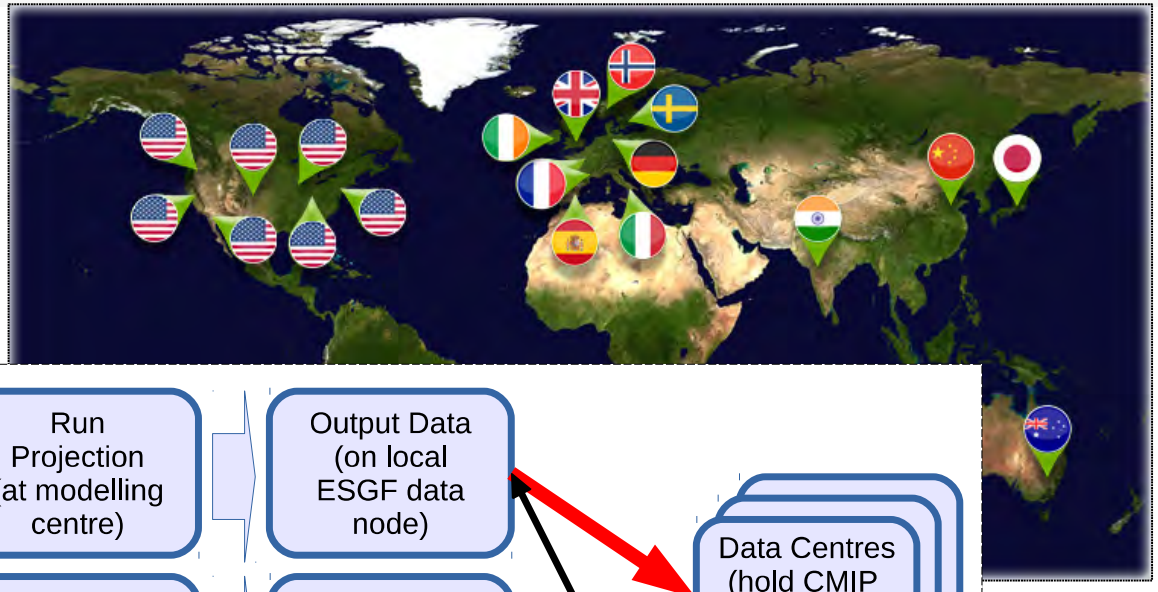
ESGF represents a multinational effort to securely access, monitor, catalog, transport, and distribute petabytes of data for climate change research experiments and observations.

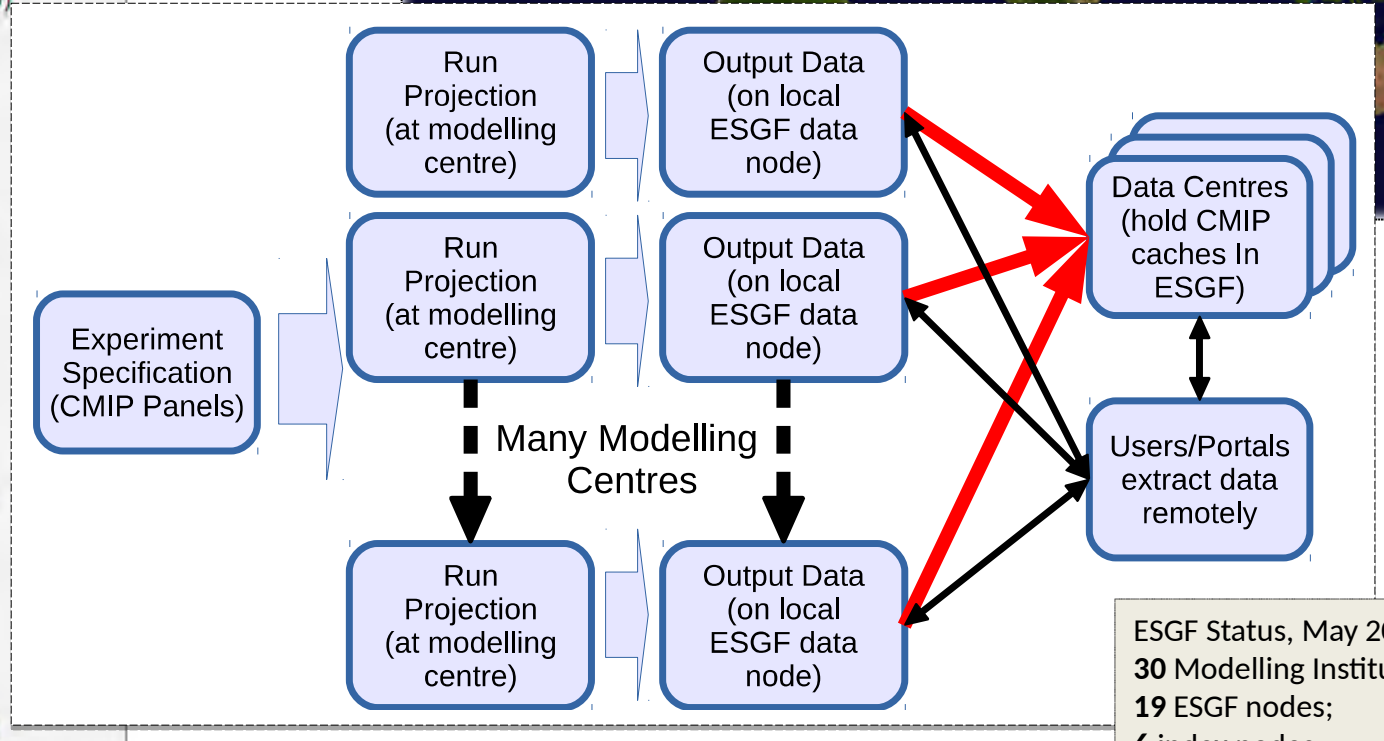


ESGF Data Infrastructure

Institute

- [ ] BCC (6891)
- [ ] BNU (512)
- [ ] CCCMA (21879)
- [x] CMCC (1189)
- [x] CNRM-CERFACS (5422)
- [ ] COLA-CFS (884)
- [ ] CSIRO-BOM (625)
- [ ] CSIRO-QCCCE (3120)
- [ ] FIO (230)
- [x] ICHEC (2966)
- [ ] INM (486)
- [ ] INPE (24)
- [x] IPSL (10637)
- [ ] LASG-CESS (2568)
- [ ] LASG-IAP (418)
- [ ] MIROC (11509)
- [x] MOHC (23452)
- [x] MPI-M (11654)
- [ ] MRI (4804)
- [ ] NASA-GISS (4781)
- [ ] NASA-GMAO (1620)
- [ ] NCAR (4922)
- [x] NCC (1014)
- [ ] NCEP (870)
- [ ] NICAM (15)
- [ ] NIMR-KMA (63)
- [ ] NOAA-GFDL (14932)
- [ ] NSF-DOE-NCAR (949)
- [x] SMHI (1840)
- [ ] UNSW (26)





ESGF Status, May 2015:
30 Modelling Institutes;
19 ESGF nodes;
6 index nodes.
140,032 datasets of which
58,174 European (**42%**)

ESGF Software Infrastructure

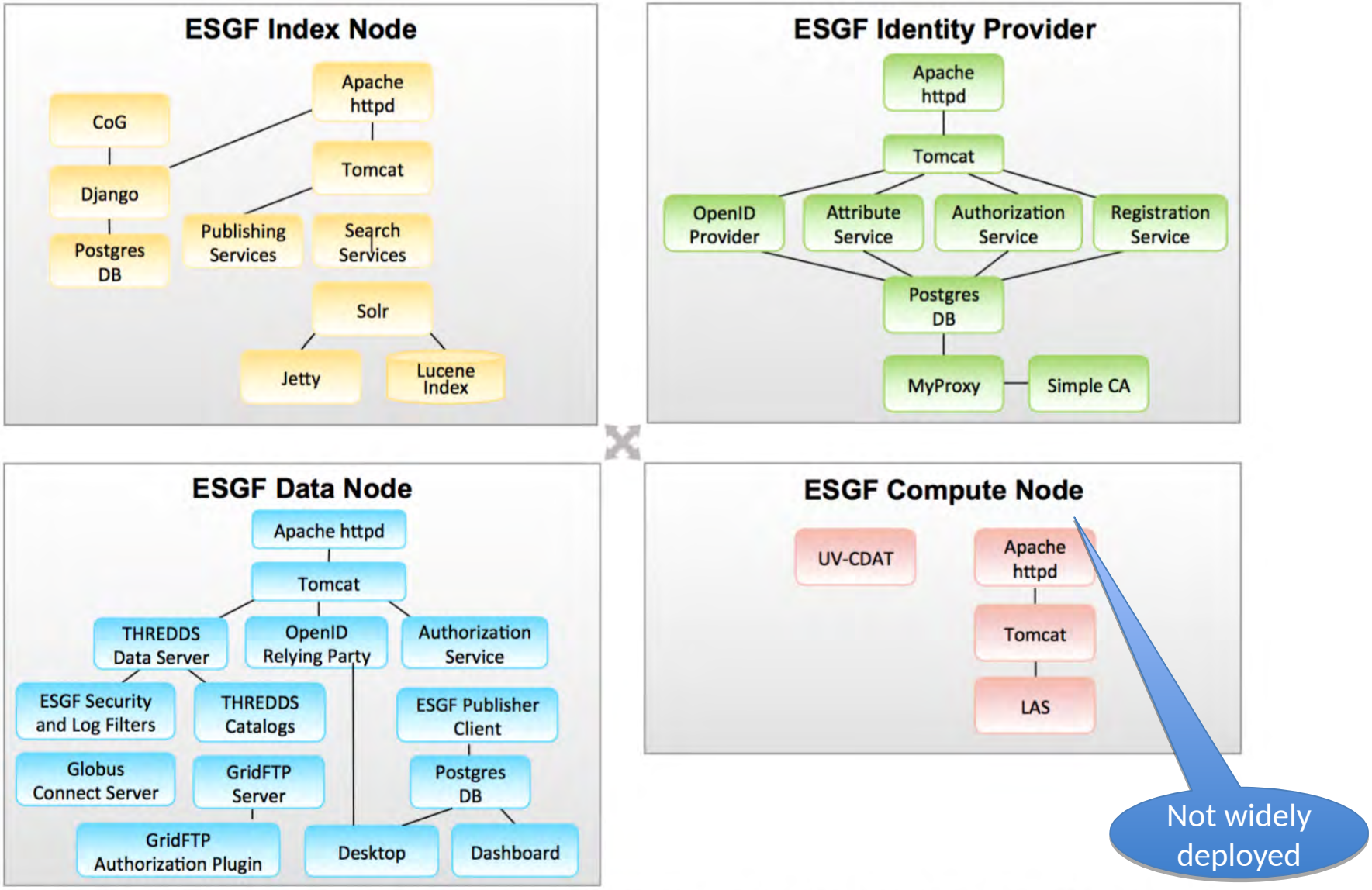
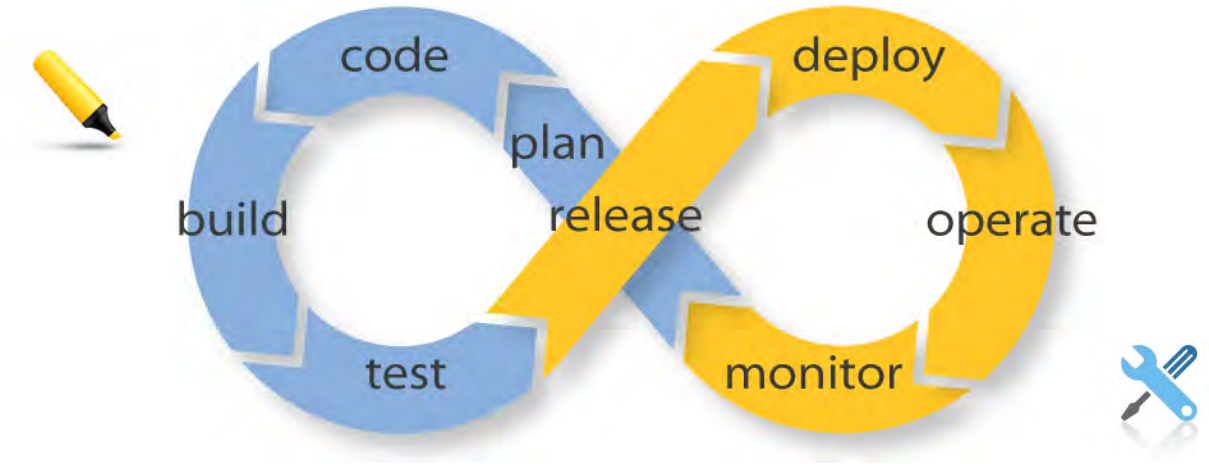


Figure 1. *Current ESGF software stack architecture at the beginning of 2016, representing Release Version 2.2.3.*

ESGF release management



Missions

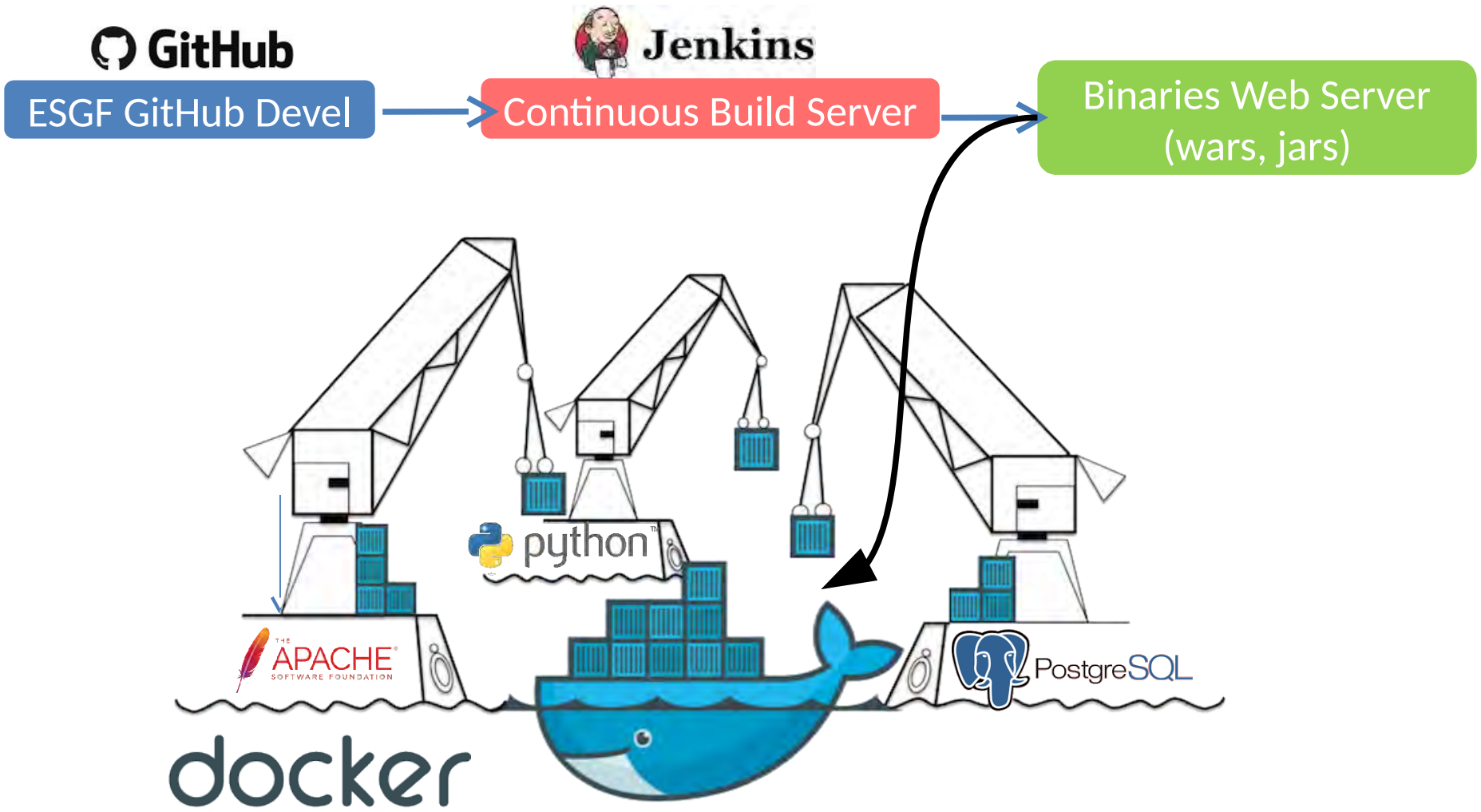
- ✓ Release management
- ✓ Build, test and validate
- ✓ Provide installation tools
- ✓ Secure deployments
- ✓ Administrators training and support

Challenges

- ➢ Automated builds and tests
- ➢ Easier installation

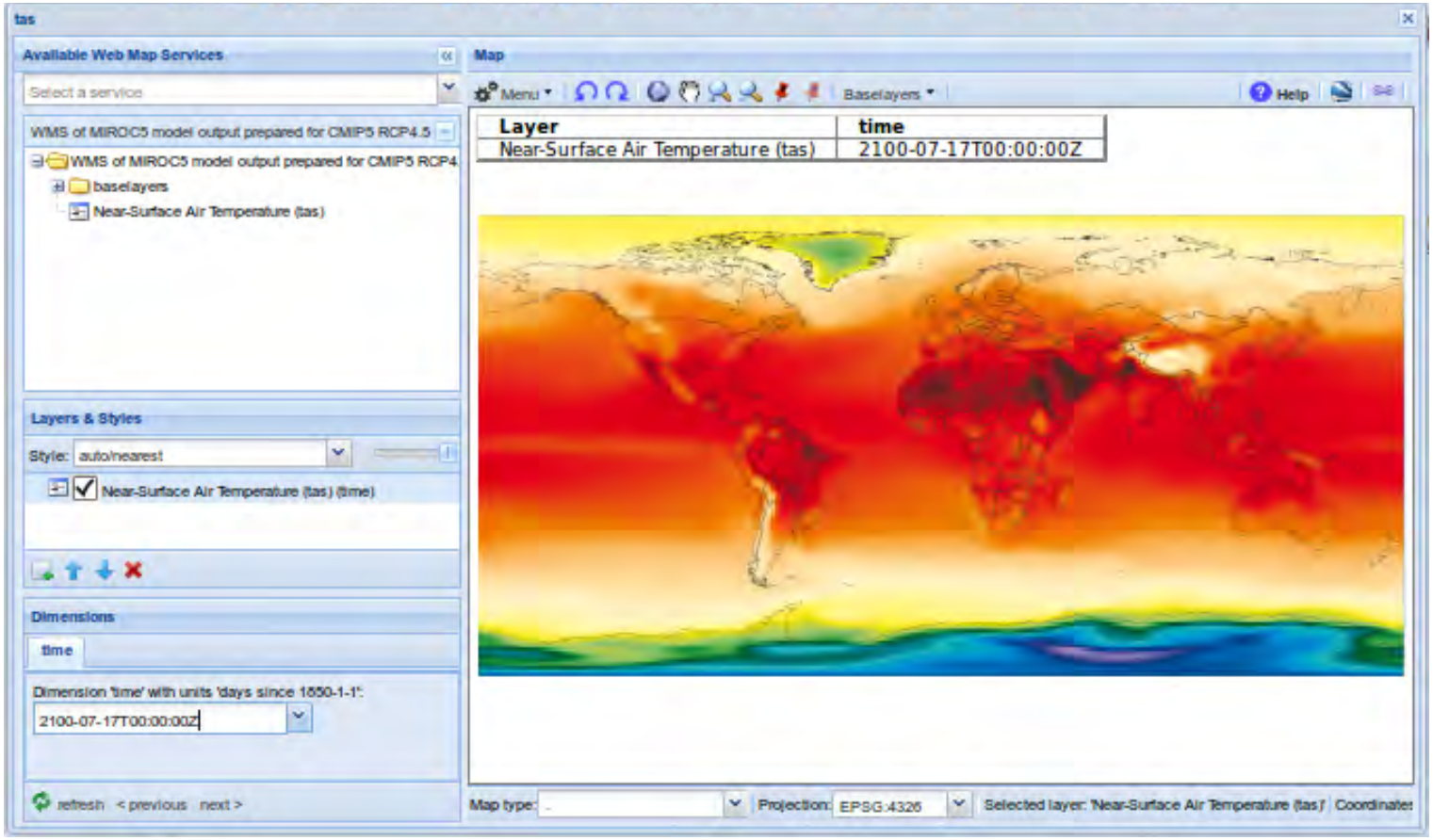
Node set up in less than one hour

Deployment and integration



ESGF Supports federated systems

KNMI:ADAGUC viewer in the climate4impact.eu portal.



Data Provider
MIROC (Japan)

Distribution
DIAS (Japan)

Identity
provider CEDA
(UK)

Authorisation
PCMDI (USA)

Visualisation
KNMI (NL)

Visualisation completely decoupled from ESGF storage: uses OpenDAP

Global data infrastructure

Dividing the work into components (i.e., data, computer, storage, and software) is easy enough, but putting together individual submissions to create a workflow for getting work done is not.

Data discovery, compute resource selection, data manipulation, derived data storage site selection, and software selection at each stage of the workflow is challenging at best.

Minimizing the time spent finding, using, and storing the data are among the more pressing concerns for users when collaborating in ESGF.

Global data infrastructure

How long does it take you on average to discover and access the date and ressources you need ?

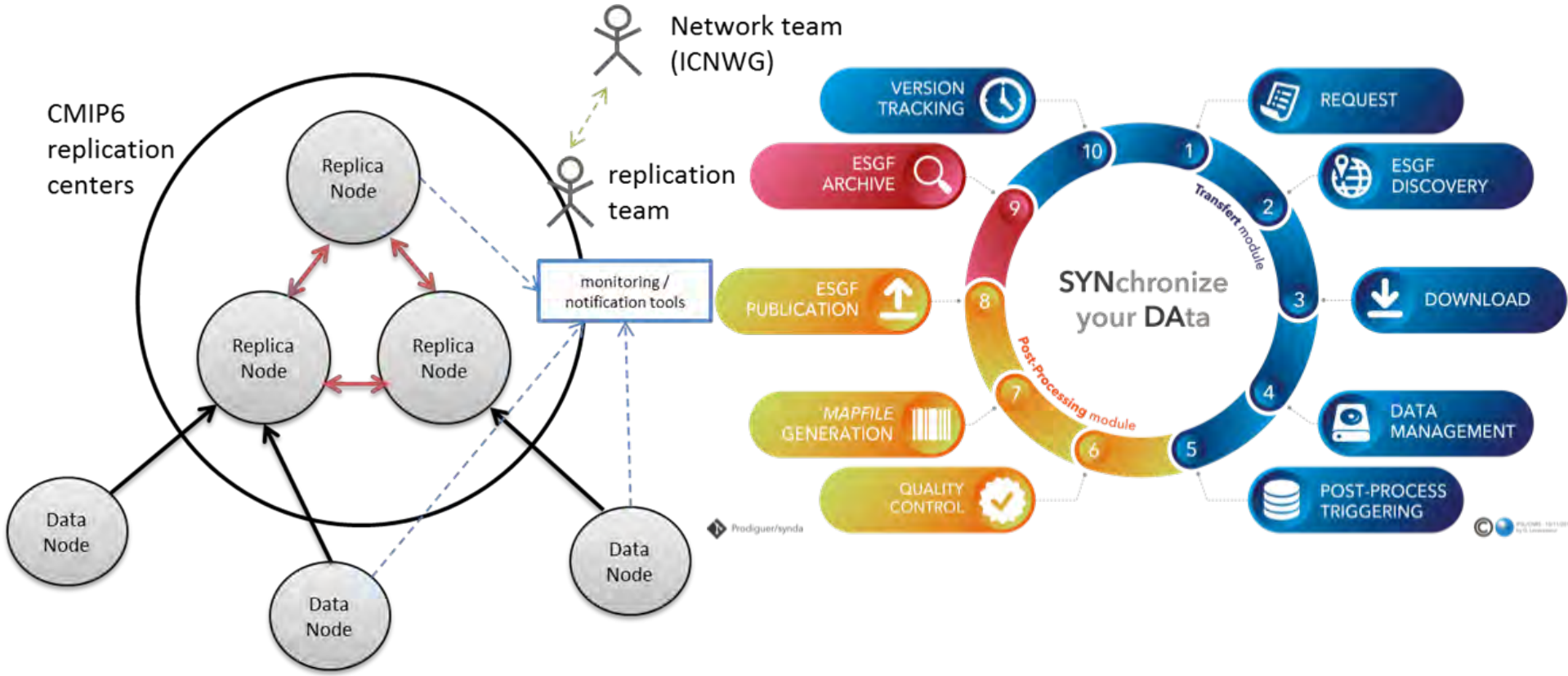
Table 8. How long does it take you on average to discover and access the data and resources you need?										
	Minutes		Hours		Days		Can't Find/Access		Total	Weighted Average
Discover	49.53%	105	36.79%	78	10.38%	22	3.30%	7	**212**	**1.67**
Access/Download	12.92%	27	42.11%	88	41.63%	87	3.35%	7	**209**	**2.35**

Which takes the longest to discover and use ?

Table 9. Which takes the longest to discover and use?										
	1 (Shortest)		2		3		4 (Longest)		Total	Weighted Average
Data	16.96%	29	22.22%	38	25.15%	43	35.67%	61	**171**	**2.80**
Computer	31.01%	49	49.37%	78	14.56%	23	5.06%	8	**158**	**1.94**
Storage	21.52%	34	36.08%	57	25.95%	41	16.46%	26	**158**	**2.37**
Software	24.53%	39	34.59%	55	20.75%	33	20.13%	32	**159**	**2.36**

Replication & versioning

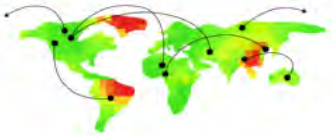
- Impact on CMIP6 data management (DM) and ESGF governance (ESGF)
- Stable processes which are supervised by a board (the CDNOT Team) are needed for CMIP6 data consistency in ESGF
- CMIP6 data replication architecture:



Network Improvement

Earth System Grid Federation ESGF

ICNWG International Climate Network Working Group



Esnet to ICNWG Site Packet Loss Testing

Better network performance needed for CMIP6

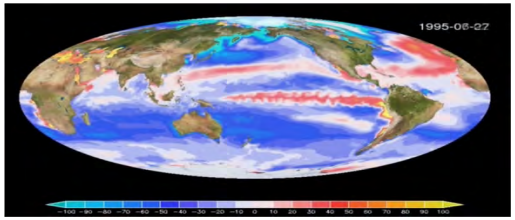
ICNWG uses perfSonar to analyze networks performance between the collaborating sites, track the health of the network connections and verify the data paths between the end sites.

CMIP6 data is estimated to be 30PB. This amount of data will require a high quality network between replication sites.

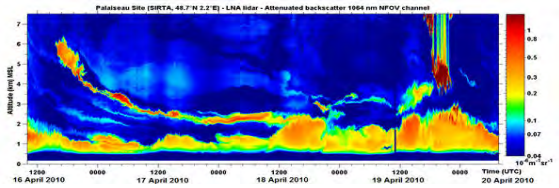
Crossing data

All data gathered together, coming from field campaign, from observational network or from numerical simulations. Data are available to the scientific community. Data are transfered to the civil society for operational applications (Climate Services, Copernicus program...).

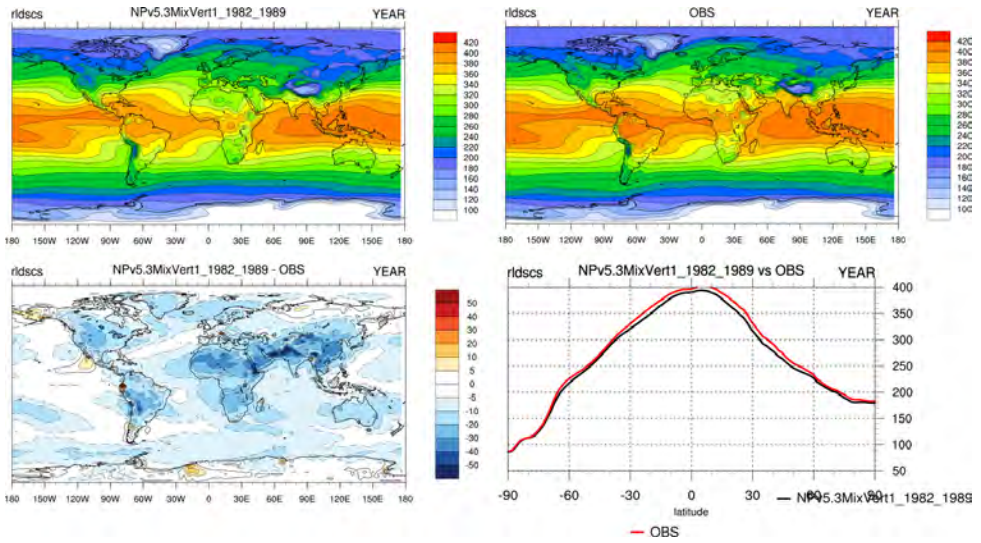
IPSL mesoscale computing and data centre hosts data and computing services relevant for climate research.



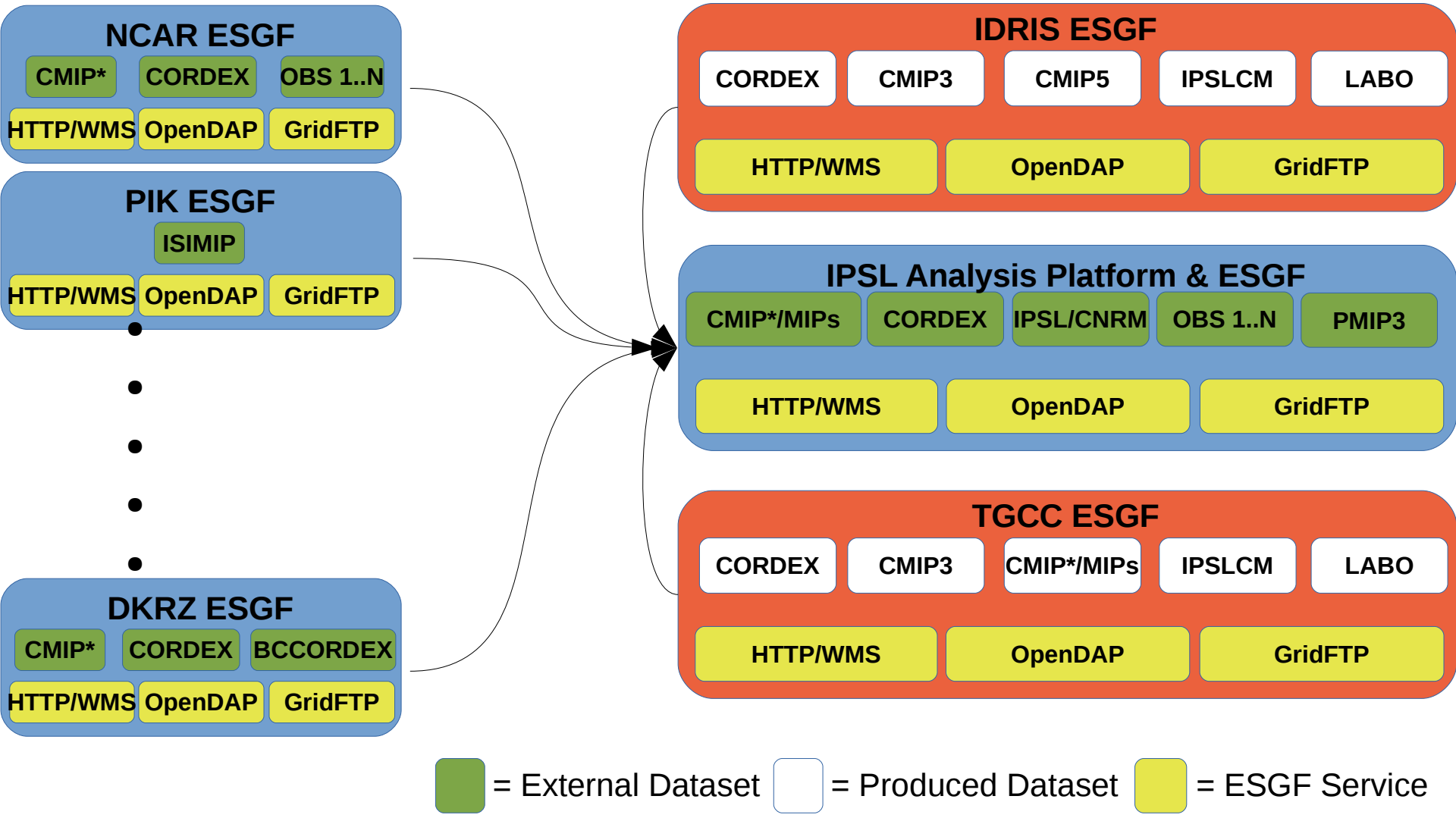
Models data



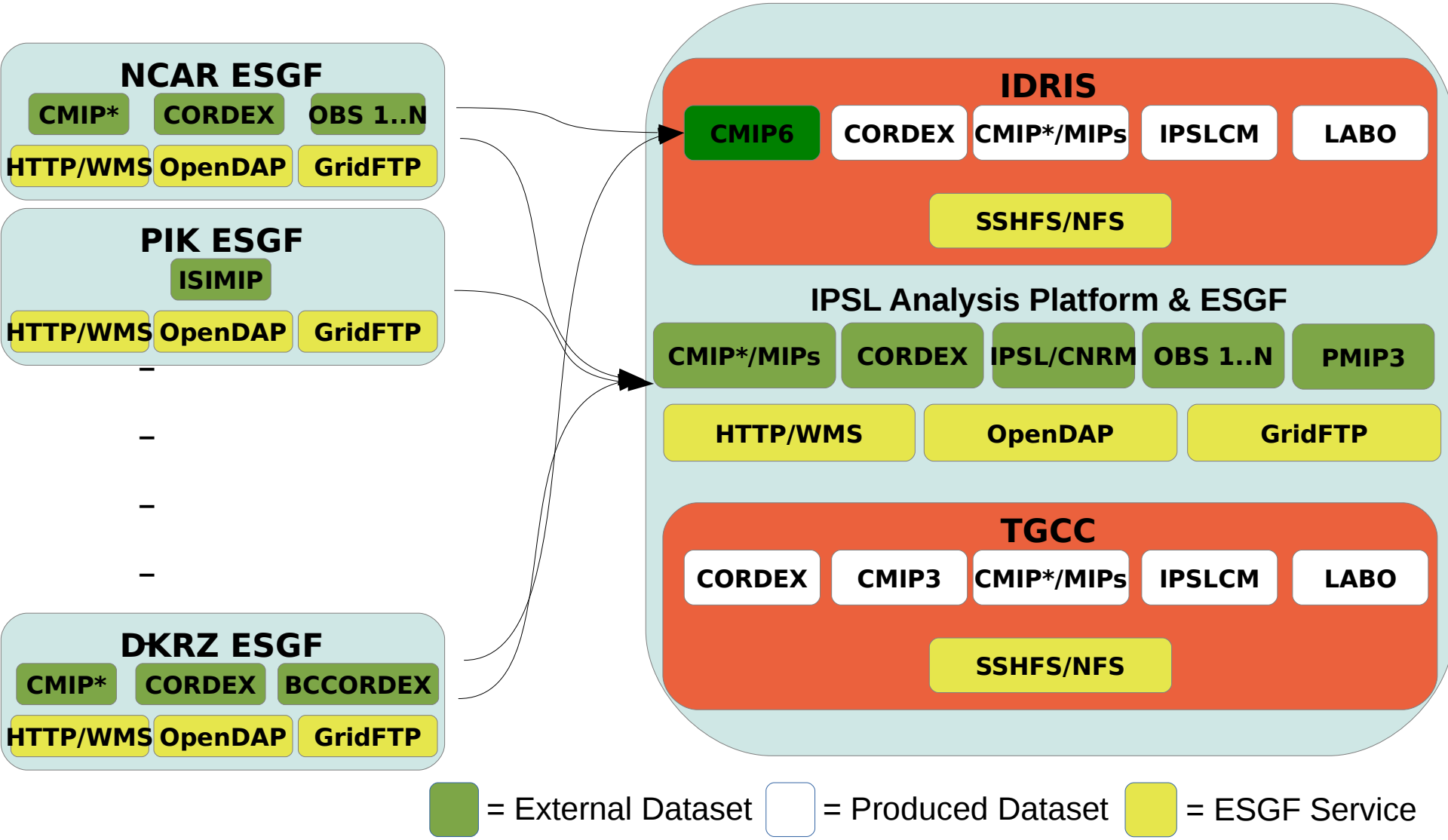
Ground observations

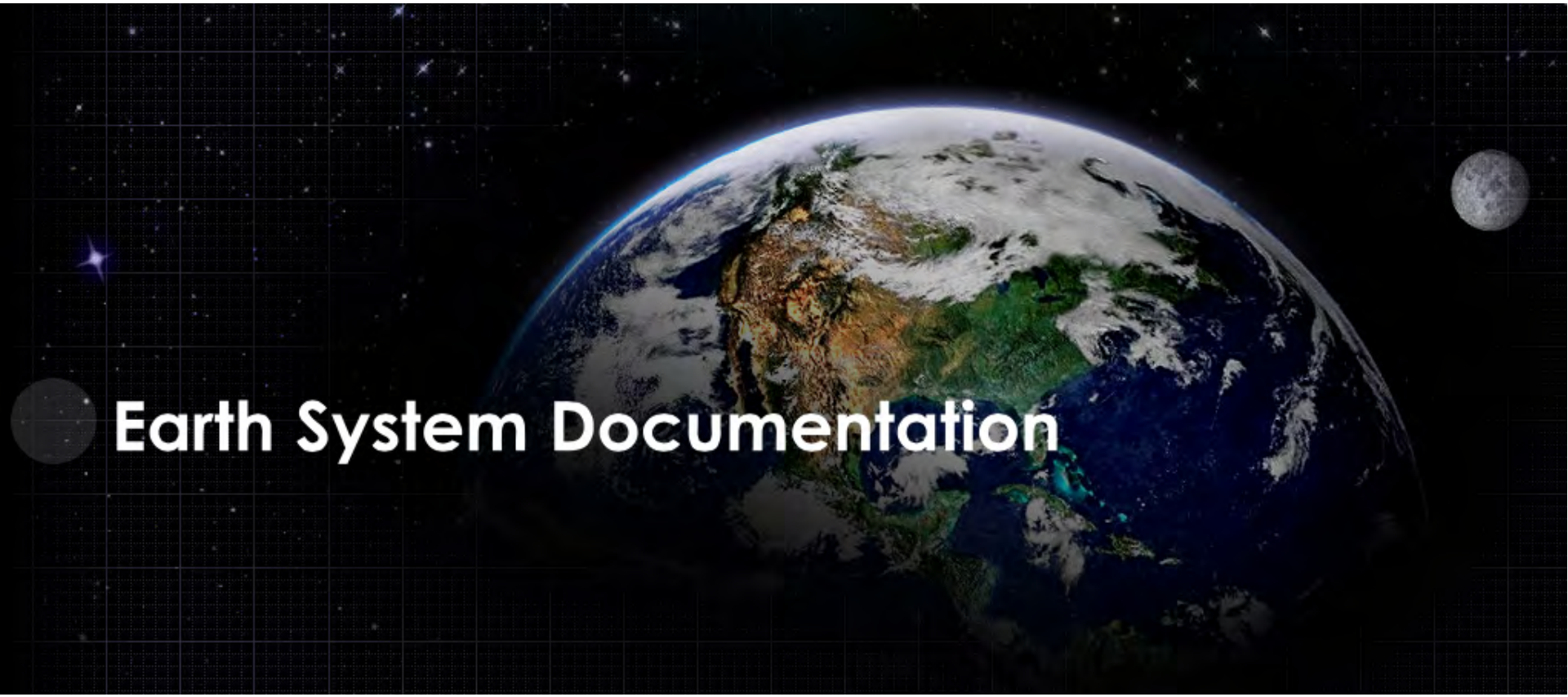


CMIP5 (2010-2016)



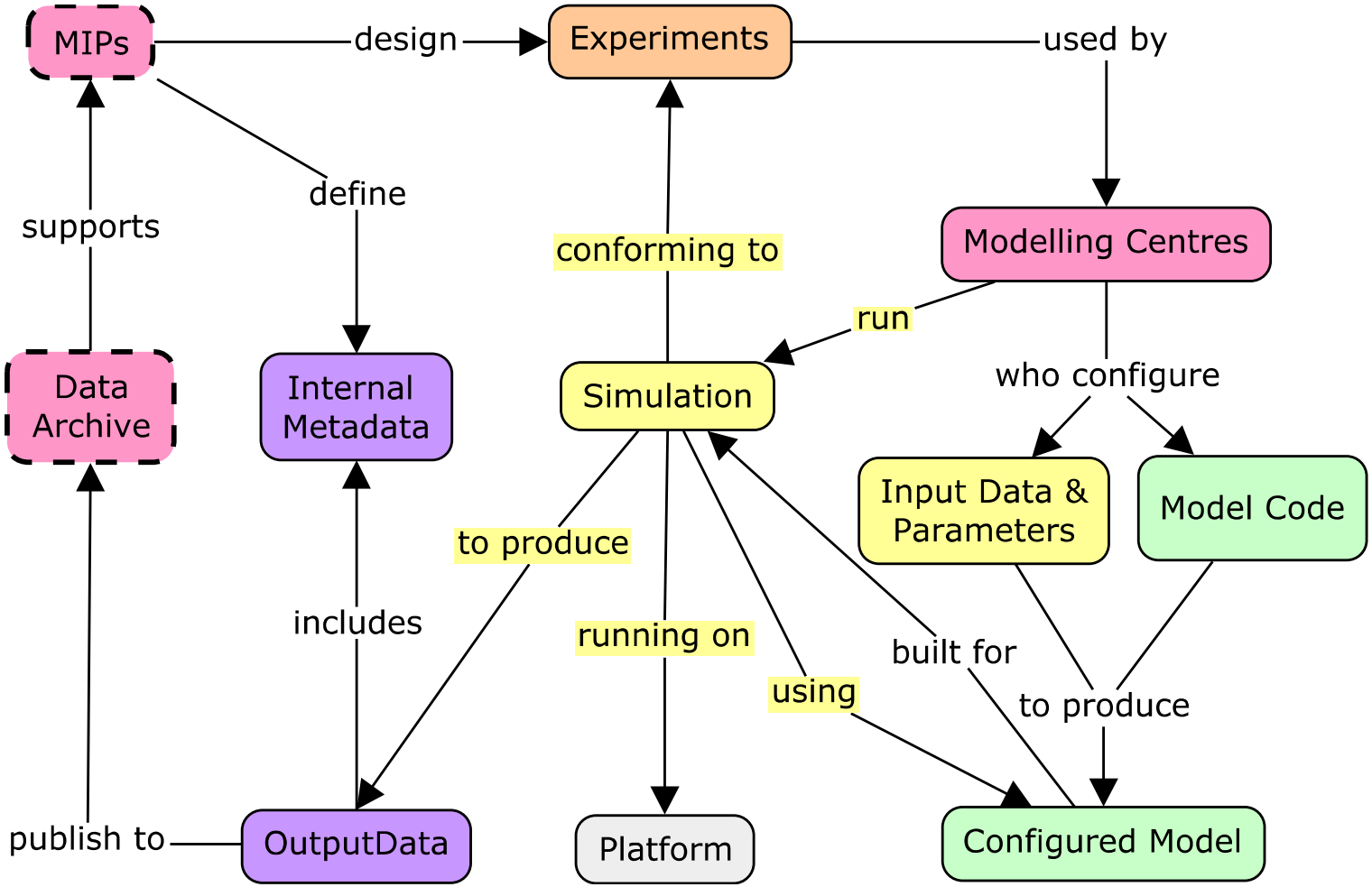
CMIP6 (2017-2023)



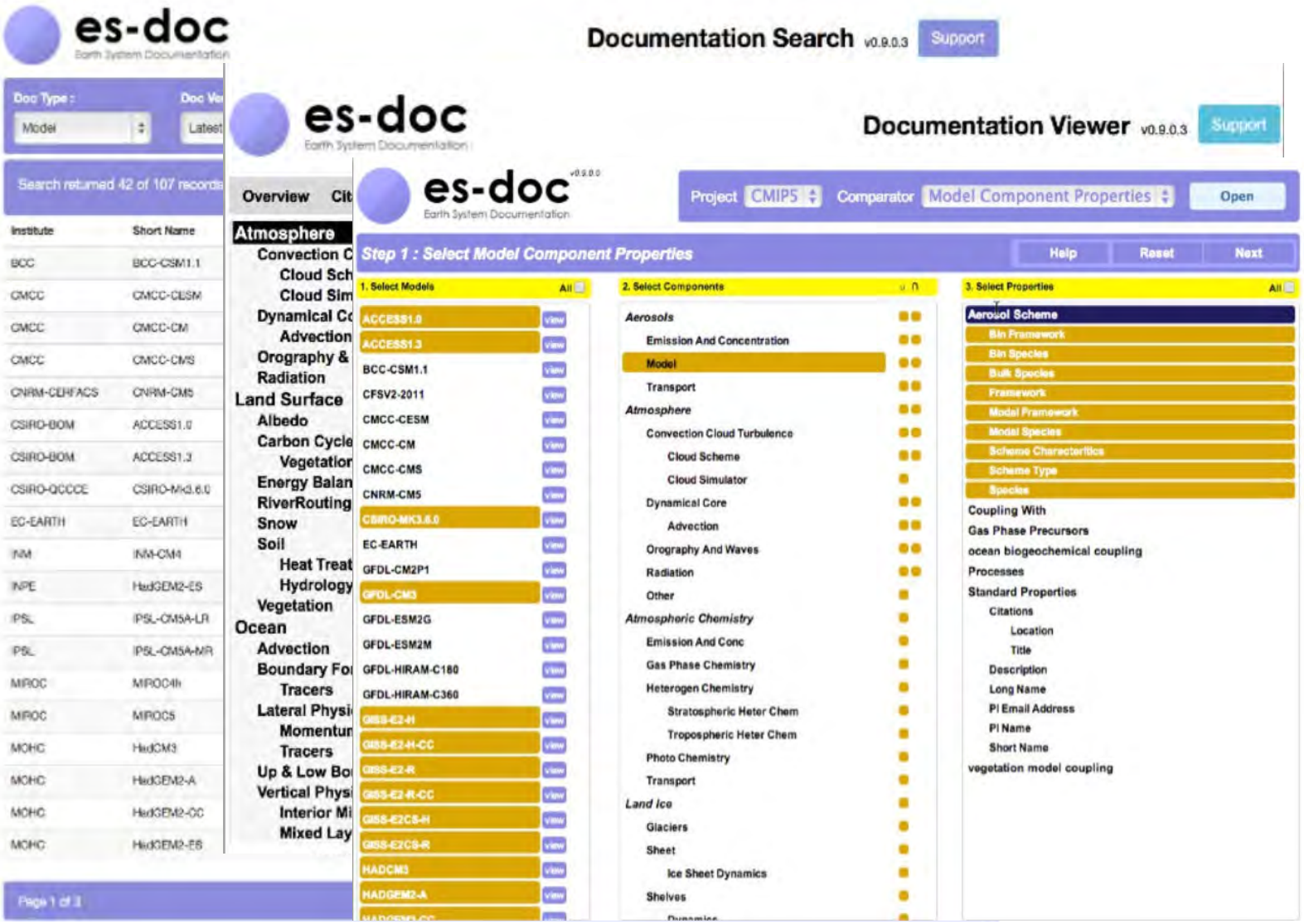


es-doc
Earth System Documentation

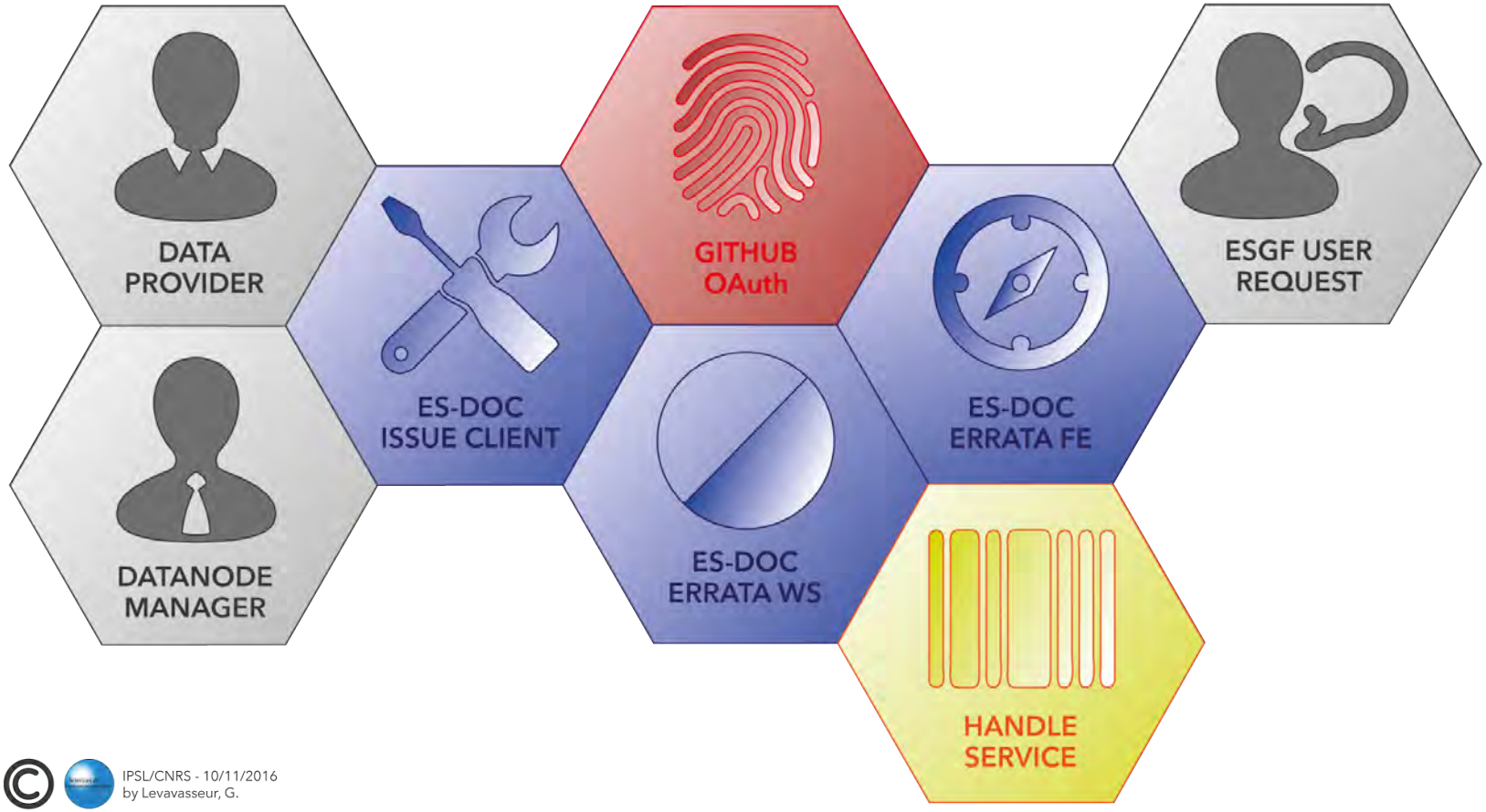
ESDOC classes and concepts



ESDOC view & search tools



ESDOC CMIP6 Errata Service



IPSLCM6 Development cost

- Total Physical Source Lines of Code (SLOC) = 761,464
- Development Effort Estimate, Person-Years = 209
- Schedule Estimate, Years = **4.08**
- Estimated Average Number of Developers (Effort/Schedule) = **51.29**
- Total Estimated Cost to Develop = **$ 28,271,263**

- SLOCCount, Copyright (C) 2001-2004 David A. Wheeler

ESGF Development cost

- Total Physical Source Lines of Code (SLOC) = 141,935
- Development Effort Estimate, Person-Years = 36
- Schedule Estimate, Years = **2.10**
- Estimated Average Number of Developers (Effort/Schedule) = **17.33**
- Total Estimated Cost to Develop = **$ 4,912,874**

- SLOCCount, Copyright (C) 2001-2004 David A. Wheeler

ESDOC Development cost

- Total Physical Source Lines of Code (SLOC) = 66,805
- Development Effort Estimate, Person-Years = 16.48
- Schedule Estimate, Years = **1.55**
- Estimated Average Number of Developers (Effort/Schedule) = **10.61**
- Total Estimated Cost to Develop = **$ 2,226,850**

- SLOCCount, Copyright (C) 2001-2004 David A. Wheeler

Aspects not covered here

List of aspects which are not covered in this presentation but which are in process in ESGF working groups and in the WIP:

- CMOR (Climate Model Output Rewriter) (Data Management)
- Control Vocabularies (DM)
- DRS (Data Reference Syntax) (DM)
- Licensing (DM)
- GUI (ESGF)
- AAI (ESGF)

ESGF: http://esgf.llnl.gov/
WIP/WGCM: https://earthsystemcog.org/projects/wip/

The WIP : Work In Progress

WIP (WGCM Infrastructure Panel) produced CMIP6 Position Papers:

Final paper:

- CDNOT (CMIP Data Node Operation Team) Terms of Reference
- CMIP6 Persistent Identifiers Implementation Plan
- CMIP6 Replication and Versioning
- CMIP6 Licensing and Access Control
- CMIP6 Data Citation and Long Term Archival
- CMIP6 Quality Assurance
- CMIP6 ESGF Publication Requirements

Working Papers:

- CMIP6_errata_system
- CMIP6_Reference_Vocabularies:lists
- CMIP6 Data Reference Vocabularies
- CMIP6 Data Request: Structure and Process
- CMIP6_global_attributes_filenames_CVs

Thank you for your attention