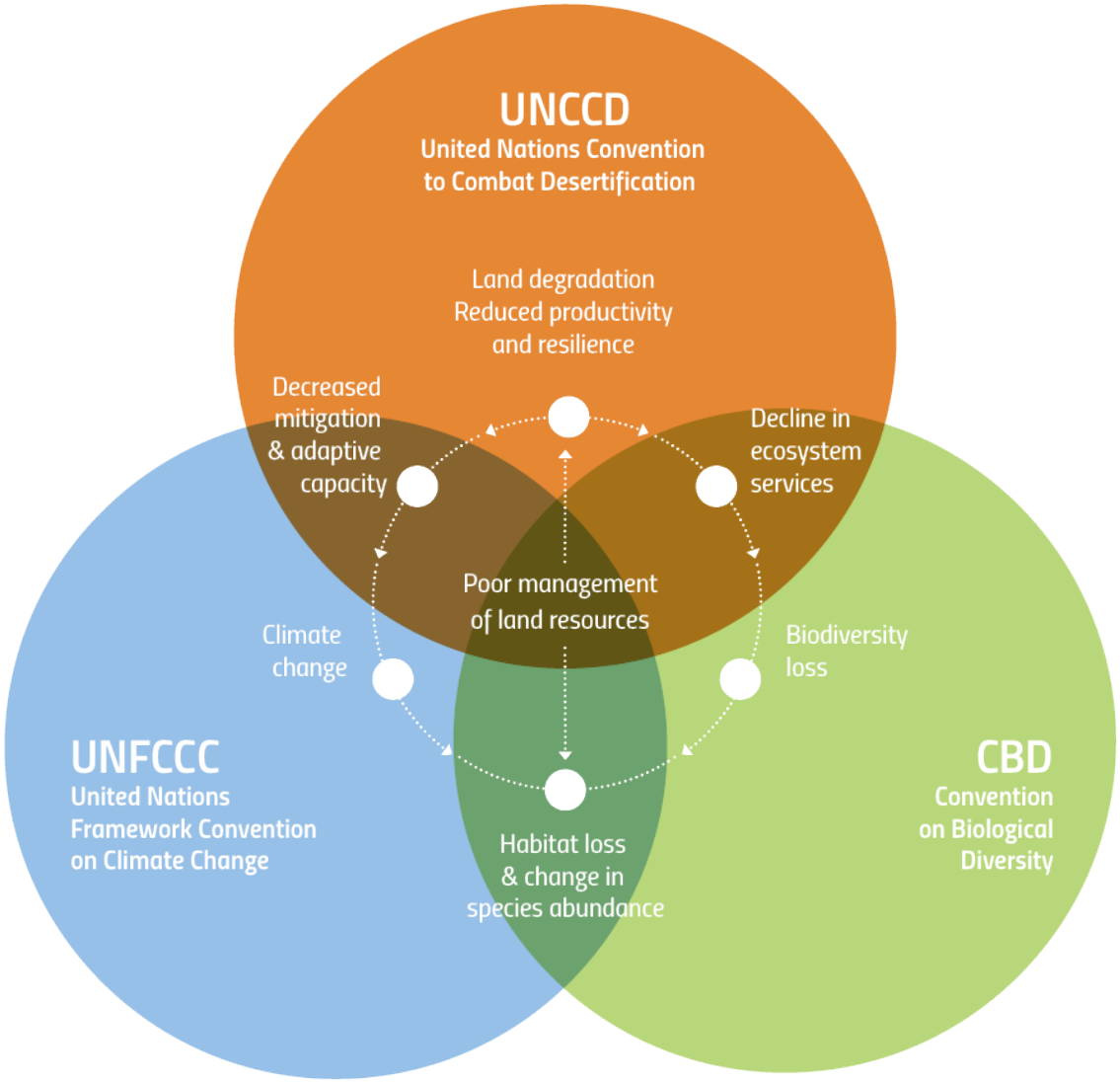
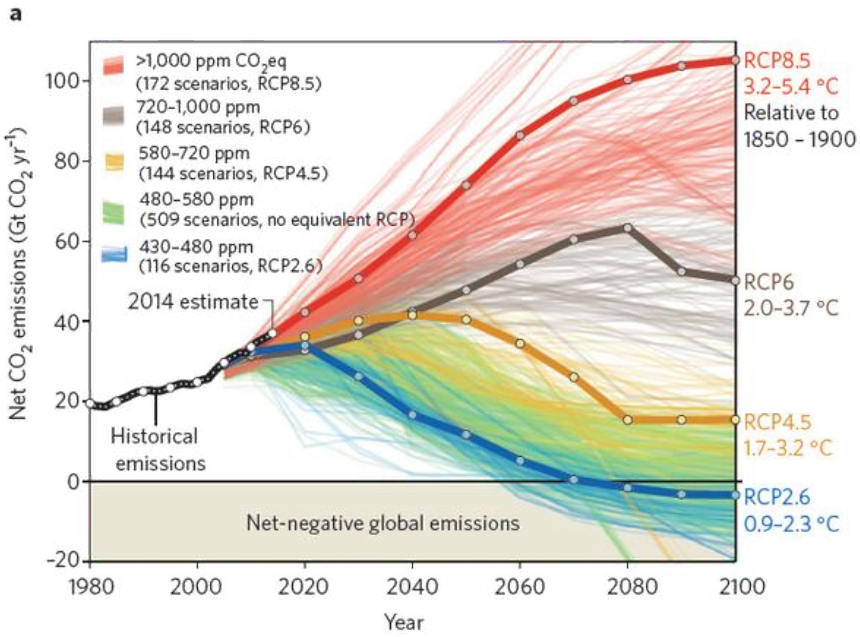
# Climate Data Science Outlook and Trends

The CODATA-RDA Research Data Science Advanced Workshops - Climate Data Sciences
Trieste 21 August 2018

Graziano Giuliani - ICTP ESP

# The Mission

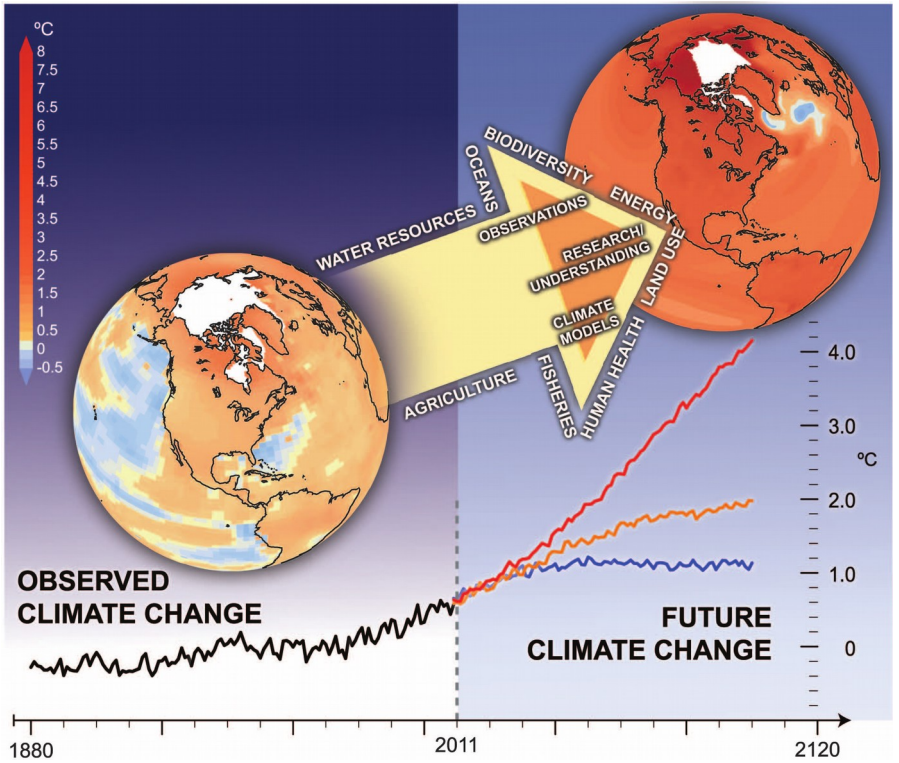



ICTP

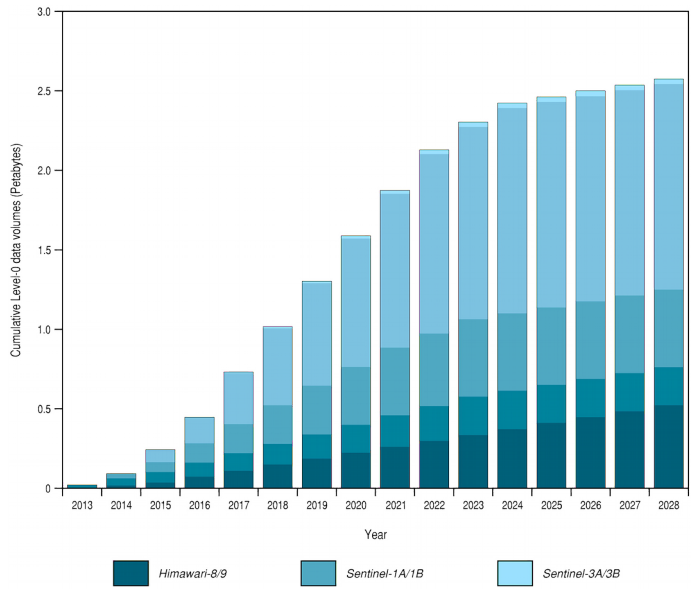
# Climate Research

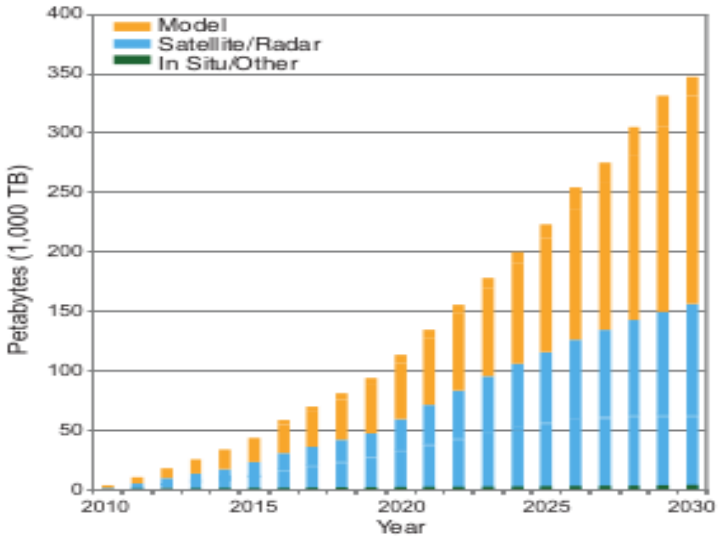
- Documenting the past, historical and paleo earth climates
- Observing actual system status
- Modeling future status
- Ecosystem, health and economic drivers and impacts




Overpeck, Science 2011

# Climate Data Numbers

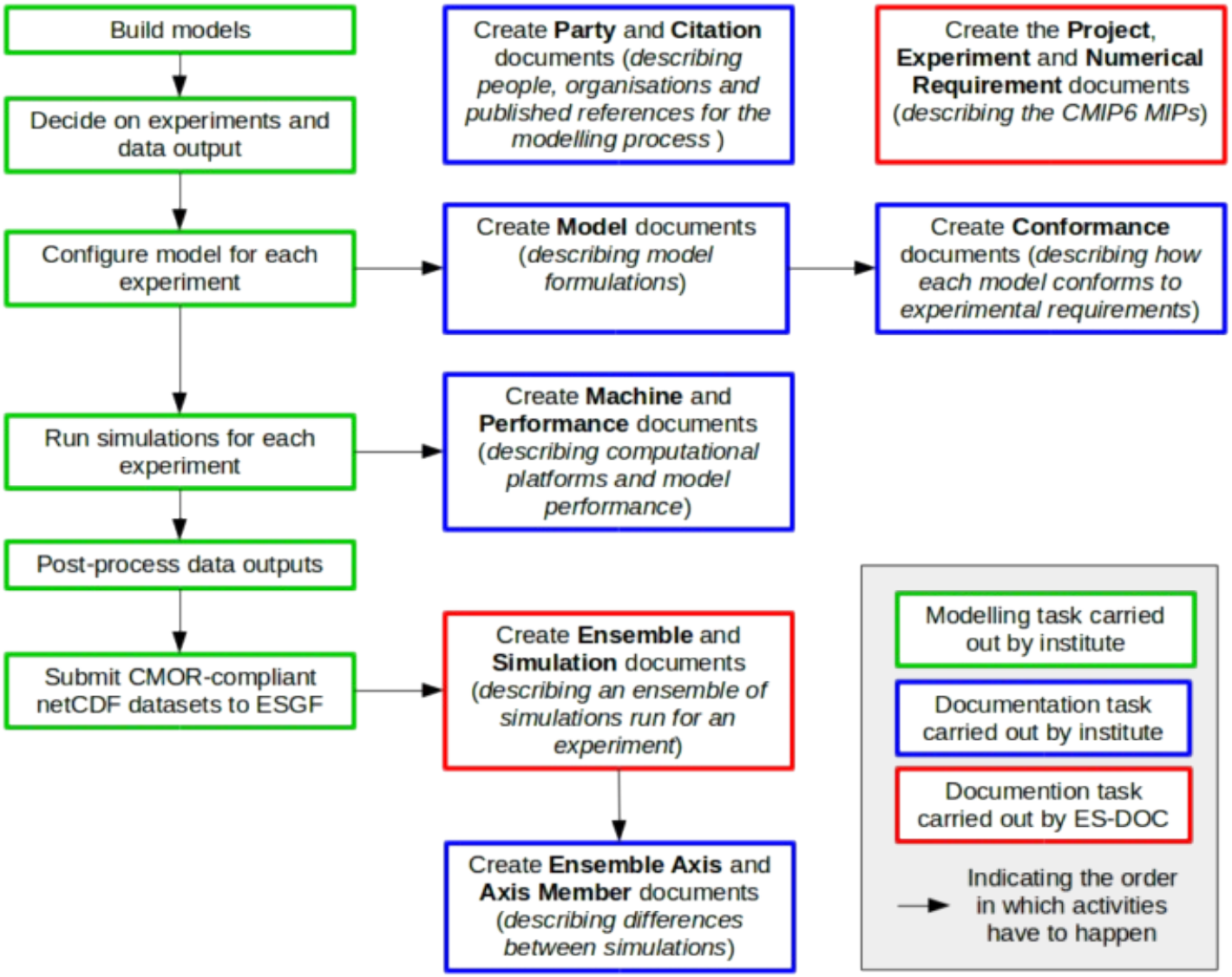



- By the end of 2017, the operational Sentinel-1, 2 and 3 satellites alone will continuously collect a volume of 20 Terabytes per day.

- The CMIP data archives have grown from the 50TB of the the CMIP3 project to the 2.5PB of the CMIP5 project. The same trend is expected for CMIP6 to reach ~100PB of disk storage space.

# IPCC GCM Model



# Acronyms

- The World Climate Research Program : WCRP
- Working Group on Climate Modeling : WGCM
- WGCM Infrastructure Panel : WIP
- The Program for Climate Model Diagnostics and Intercomparison : PCMDI
- Coupled Model Intercomparison Projects : CMIP

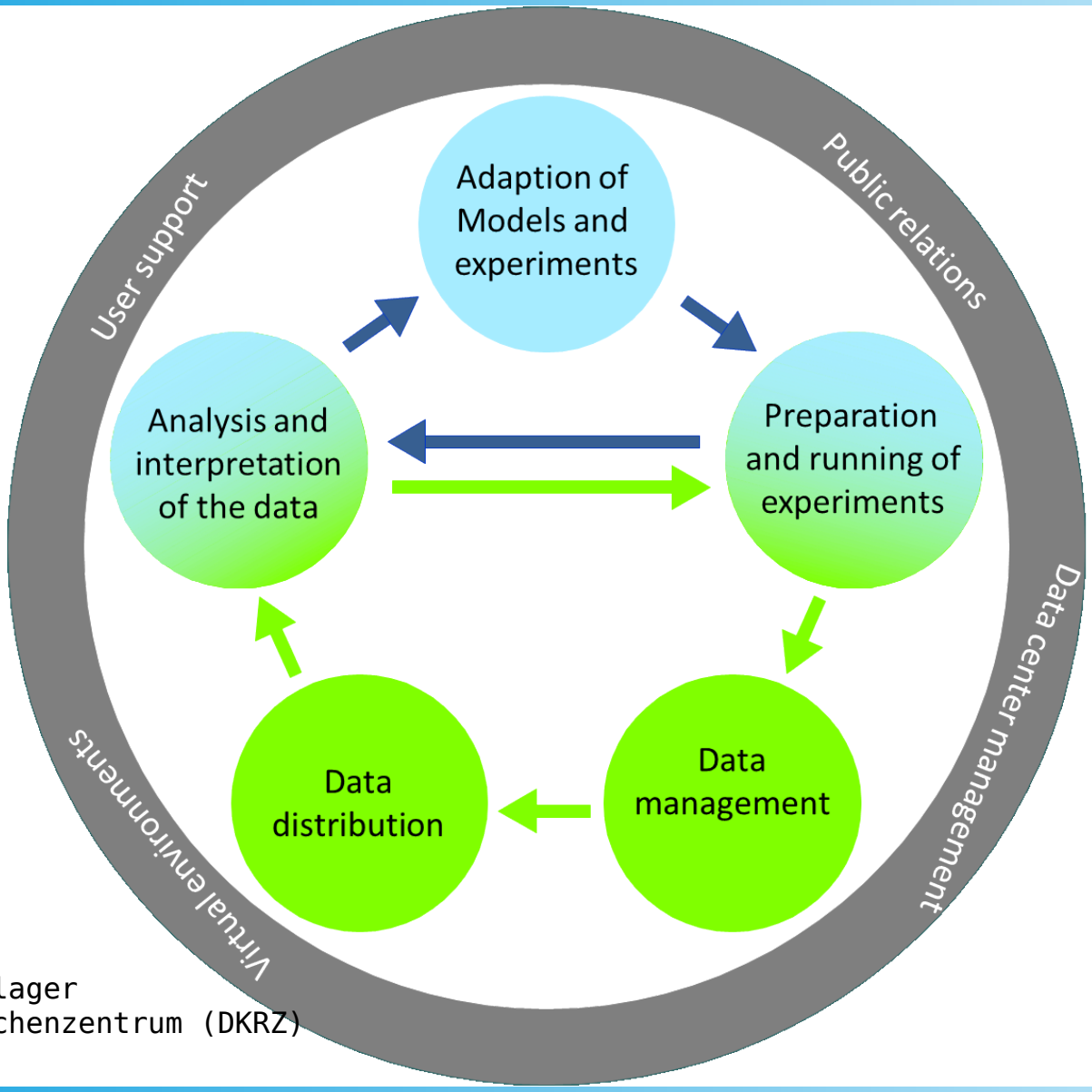
Requirements for a global data infrastructure in support of CMIP6

Geosci. Model Dev. Discuss. https://doi.org/10.5194/gmd-2018-52

# WIP CMIP6 guidelines

- The global computational and data infrastructure needs to be formally examined as an integrated element.
- Focus shifting to Impact Studies
- Scientific reproducibility and durability and provenance of data
- Systematic and routine evaluation of Earth System Models (ESMs)
- Mechanisms to identify costs and benefits in developing new models, performing CMIP simulations, and disseminating the model output
- Experimental specifications as machine-readable experiment design on all of the controlled vocabularies
- Review the management of information about users to simplify communications with them

# Climate Center Service Structure



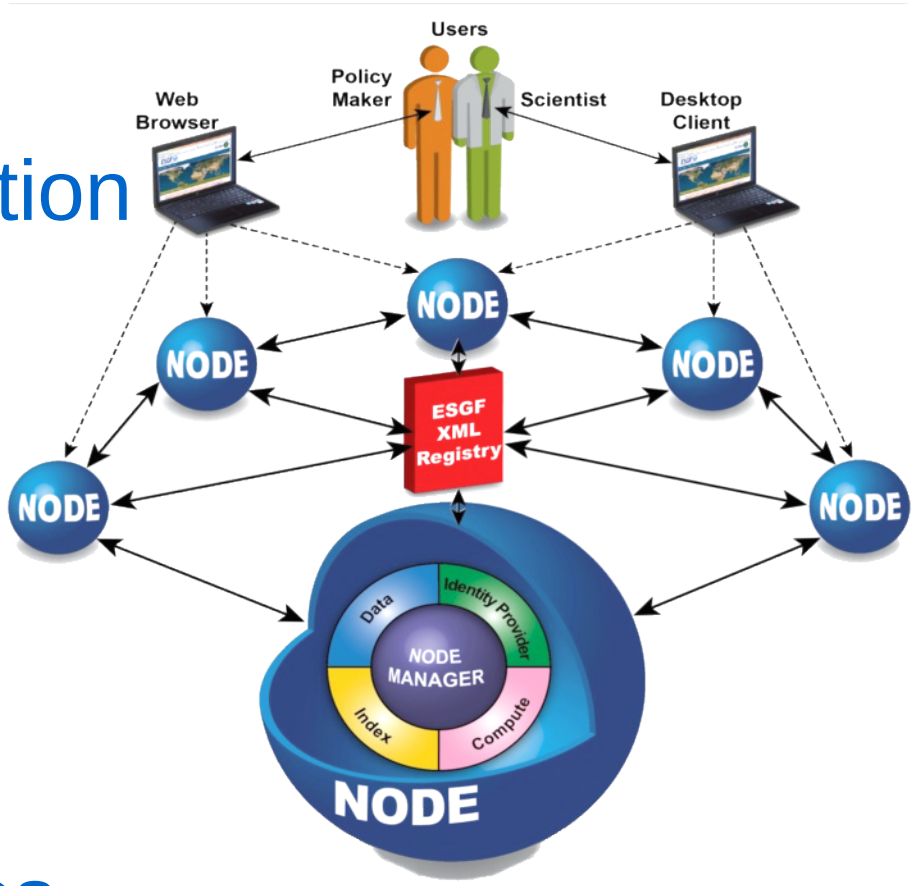

Michael Lautenschlager
Deutsches Klimarechenzentrum (DKRZ)

# IPCC Model data repositories

- Earth System Grid Federation

- National Sites
- Impact Portals

- Climate Service Companies
  PAY SERVICES



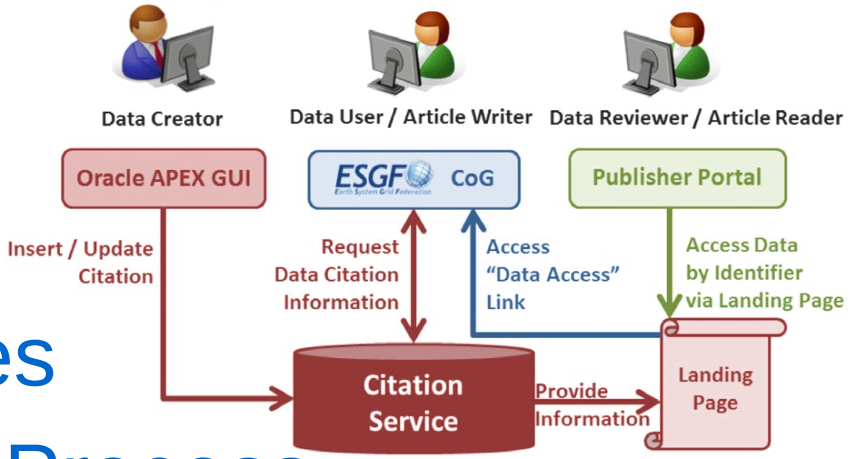

ESGF.LLNL.GOV

ICTP

# Data Management

https://www.earthsystemcog.org/projects/wip/position_papers

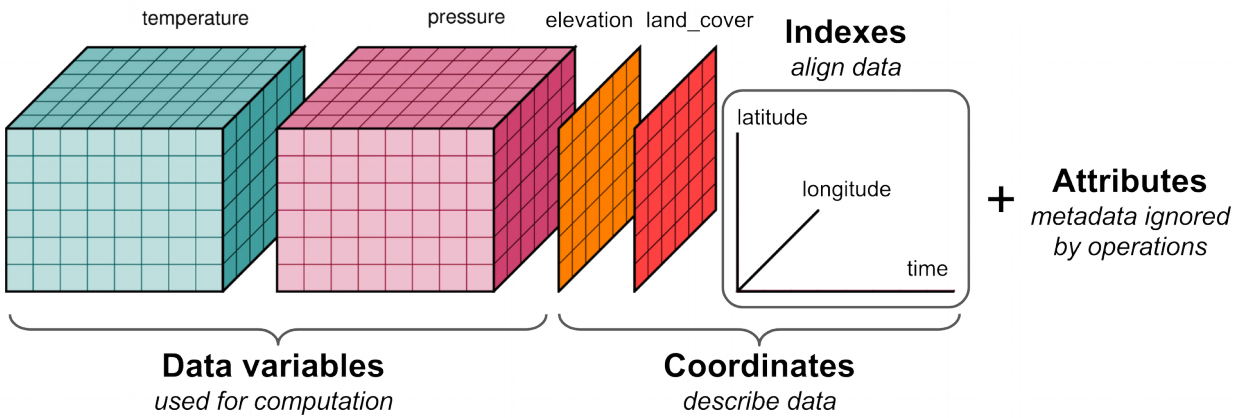
- Replication and Versioning
- Use of Persistent Identifiers
- Data Reference Vocabularies
- Data Request Structure and Process
- Data Quality Assurance
- Data Citation and Long-term Archiving
- File Names and Global Attributes
- Licensing and Access Control
- Errata service




Stockhause, M and Lautenschlager, M 2017 CMIP6 Data Citation of Evolving Data. Data Science Journal, DOI: https://doi.org/10.5334/dsj-2017-030

# Data Format

- NetCDF format




- Climate Model Output Rewriter CMOR3
  - Each file contains a single primary output variable (along with coordinate/grid variables, attributes and other metadata) from a single model and a single simulation
  - Variable number of time slices (samples) can be stored in a single file
  - Metadata written are defined MIP-specific tables of information
  - Unit of measure checking through UDUNITS library

# Data Analysis Workflow

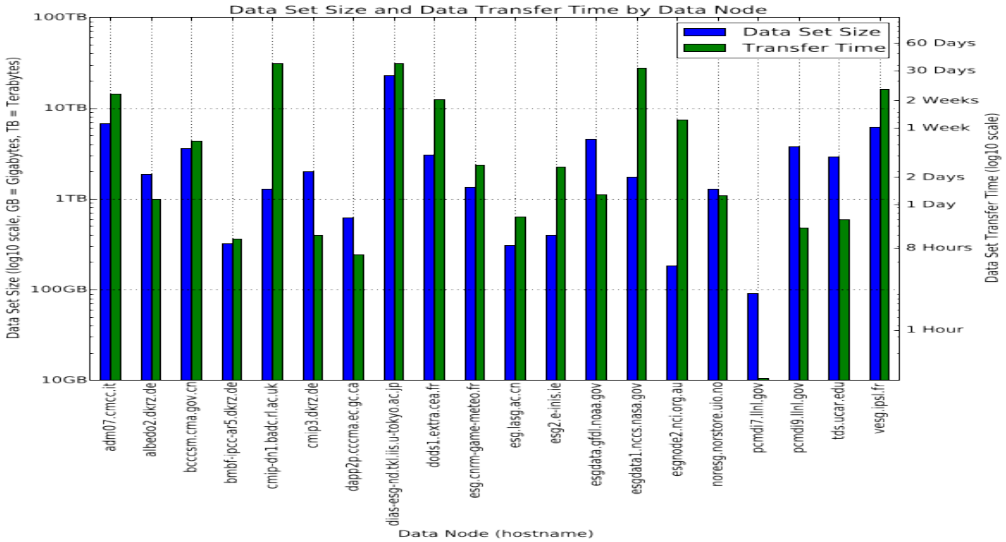
- Data Collection (STAGING)
- Data Pre-Processing (ADAPTATION)
- Scientific work (PROCESSING)
- Result check (VERIFICATION)

Publication and peer review process

# Timings

An Assessment of Data Transfer Performance for Large-Scale Climate Data Analysis and Recommendations for the Data Infrastructure for CMIP6 - Dart, Wehner, Prabhat

- STAGING – 3 months
- ADAPTATION – 3 weeks
- PROCESSING – 2 days
- VERIFICATION – 10 minutes




STAGING (Data Transfer)

is the bottleneck for data analysis

# Data Analytic Storage Systems

- Traditional :
  - Move data from Storage to Compute
  - Computation
  - Move results to Storage

- Emerging :
  - DASS
    - Move Analytics to storage/compute nodes
    - Results kept on storage/compute nodes
  - ESGF OGC WPS Interfaces
    - Climate analytics through Web Processing Services

hadoop

ICTP

# Reinvent the wheel?

- A data cube (or datacube) is a multi-dimensional ("n-D") array of values. Typically, the term datacube is applied in contexts where these arrays are massively larger than the hosting computer's main memory; examples include multi-Terabyte/Petabyte data warehouses and time series of image data.
- Google Earth Engine combines a multi-petabyte catalog of satellite imagery and geospatial datasets with planetary-scale analysis capabilities and makes it available for scientists, researchers, and developers to detect changes, map trends, and quantify differences on the Earth's surface.

# WIP remark

"In the future, datasets and software with provenance information will be first-class entities of scientific publication, alongside the traditional peer-reviewed article […] Data analytics at large scale is increasingly moving toward machine learning and other directly data-driven methods of analysis, which will also be dependent on data with provenance tracking."

# Hands-on Lab

- Python tools for CMIP5 data processing and plotting

Open a terminal and type:

```
cd /scratch/$USER

wget http://clima-dods.ictp.it/Workshops/CODATA_2018/codata_2018_climate_data.ipynb
```