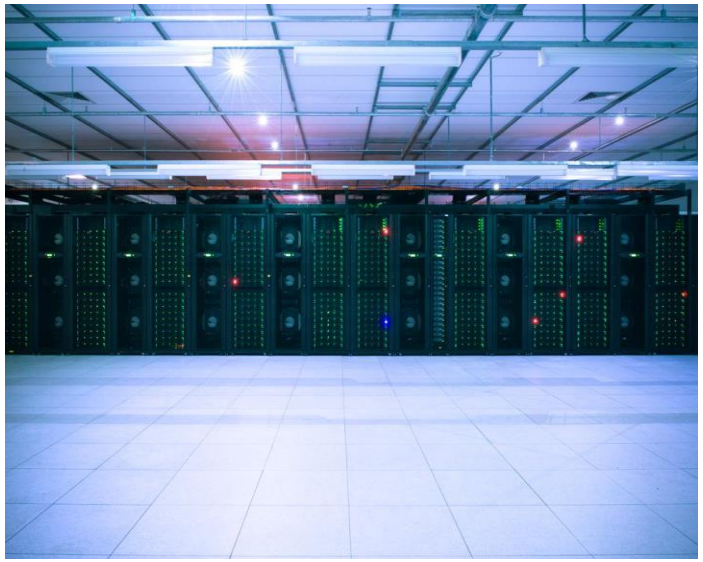
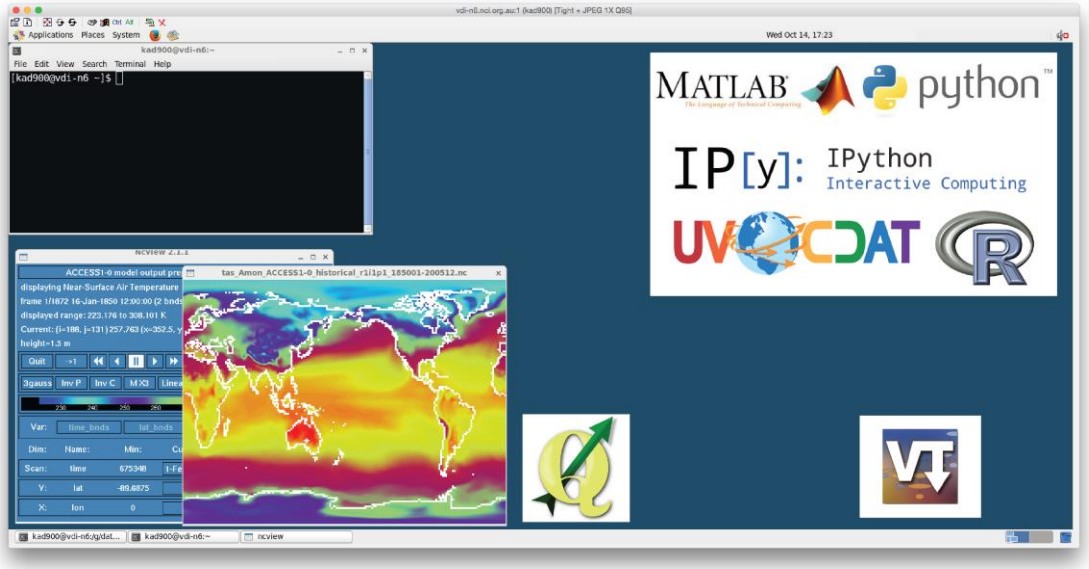
The Australian Climate Science Data-enhanced Virtual Laboratory

***Clare Richards**, Ben Evans, Kate Snow, Chris Allen, Matt Nethery,*
Paola Petrelli, Claire Trenham, Aurel Moise, Sean Pringle,
Claire Carouge, Scott Wales, Louise Wilson, Tim Erwin.

The National Research Environment for Climate Science

- The history
- The Climate Science DeVL project
- Preparing for the future





Clare.Richards@anu.edu.au

About the National Computational Infrastructure

HPC and cloud infrastructure

- Computationally- and data-intensive science
- Big domains and long simulations

15+PB of reference data collections:

- **Climate and Weather**
- Environmental
- Earth Observation
- Geophysical
- Optical Astronomy

- Other data:

  Genomics and Social Sciences

BoM | GA | CSIRO | ANU | Other National | Inter-national

Clare.Richards@anu.edu.au

Understanding the needs for Climate research

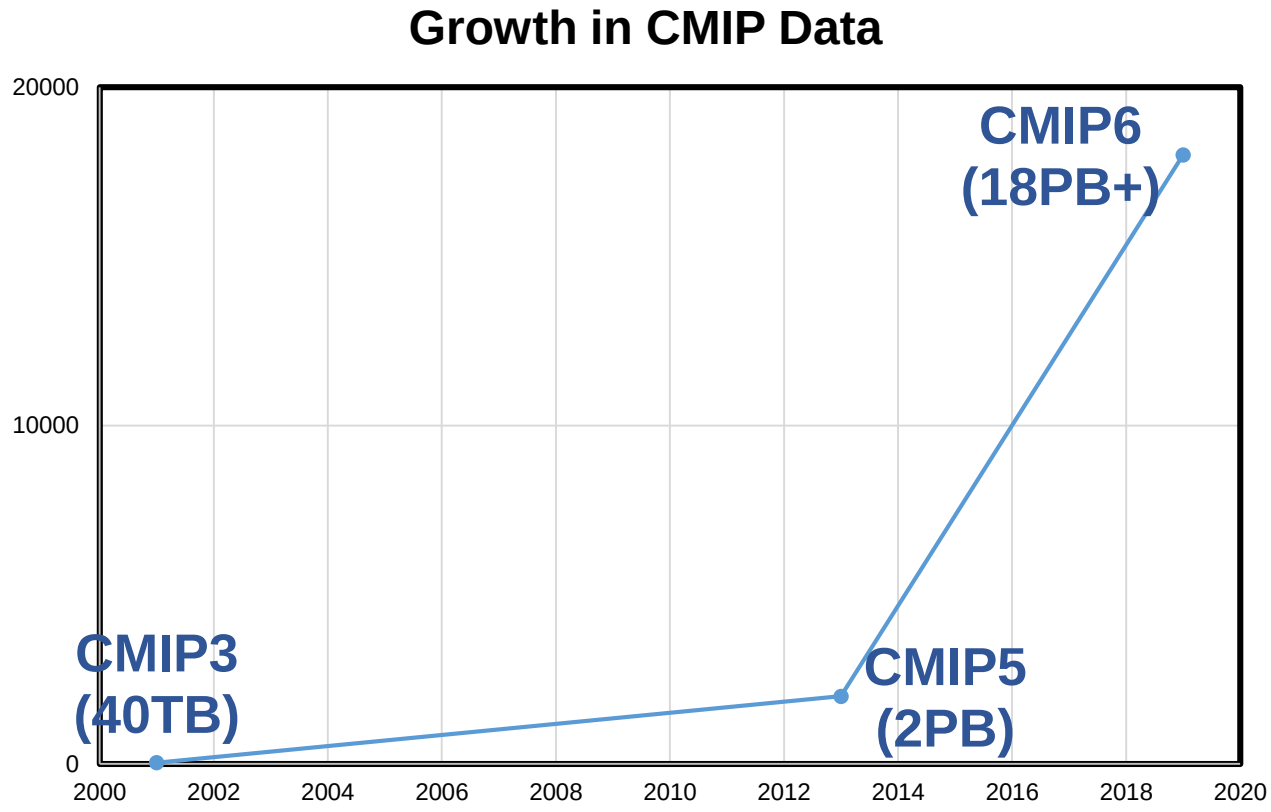
One of the most **computationally demanding** research domains in the environmental sciences.

Volume and complexity of data is growing

- International collaboration involves sharing a lot of data.
- Coupled Model Intercomparison Project (CMIP)

Need a research platform for model development and intensive data analysis

- Must adapt as user needs change.



Balaji V, et al (2018) *Requirements for a global data infrastructure in support of CMIP6*

Clare.Richards@anu.edu.au

Digital environments for Climate and Weather Science

Climate and Weather Science Virtual Laboratory (2013-2016)

- Integrated compute and data analysis platforms

Research Data Storage and Services (2011-2017)

- Repository for National Reference Data Collections established
- Easier access to quality data
- Co-located with Petascale HPC and Cloud

Collaborative user-focused development

- Users with range of skills, experience.

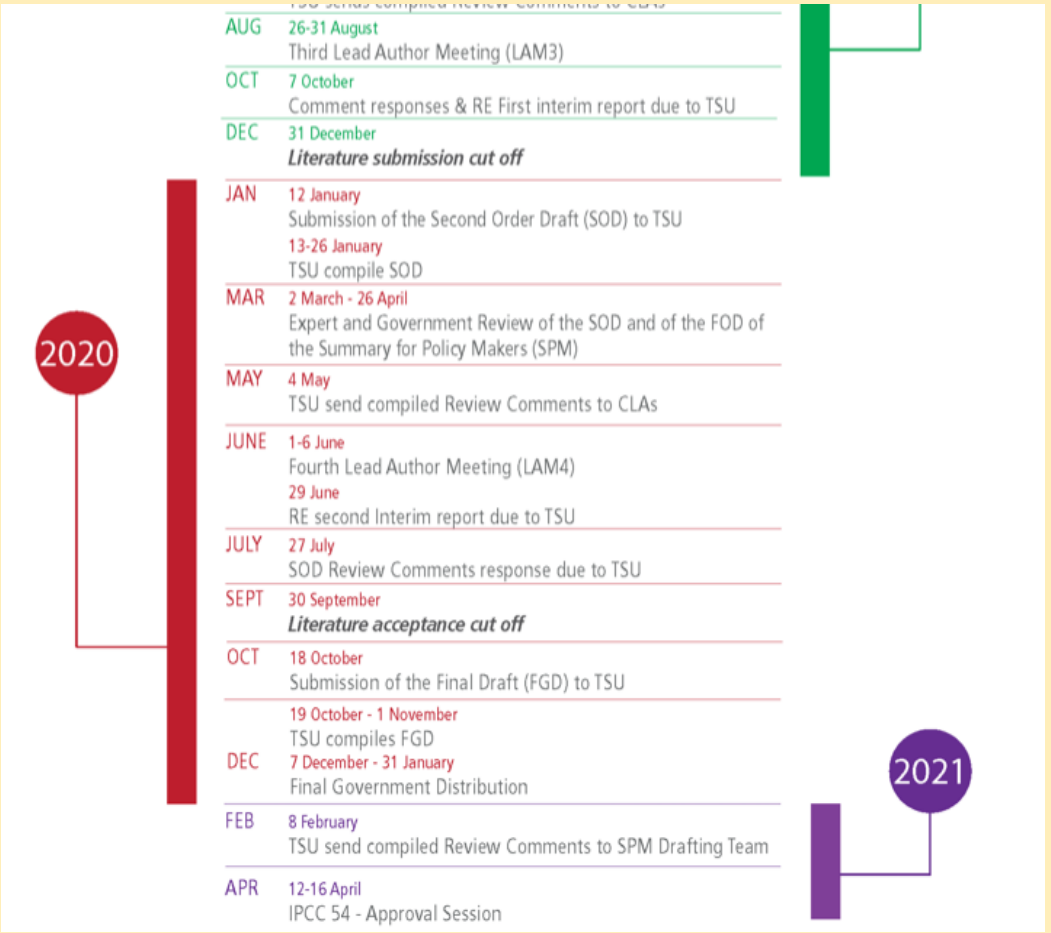
NCI AUSTRALIA · CSIRO · Australian Government Bureau of Meteorology · ARC CENTRE OF EXCELLENCE FOR CLIMATE EXTREMES · Australian Research Data Commons

Australian climate research 2018-2020+

- Coupled Model Intercomparison Project Phase 6 (CMIP6)
  - World Climate Research Programme (WCRP)
  - Internationally coordinated research activity
  - Basis of assessments by the IPCC (Intergovernmental Panel on Climate Change)
  - Informs global climate agreements
  - Used by Australian research, government, business, agriculture and industry

- Australian Research effort for CMIP6
  - Model Development (ACCESS-CM2, ACCESS-ESM1.5)
  - Data management, access and analysis

IPCC Timeline



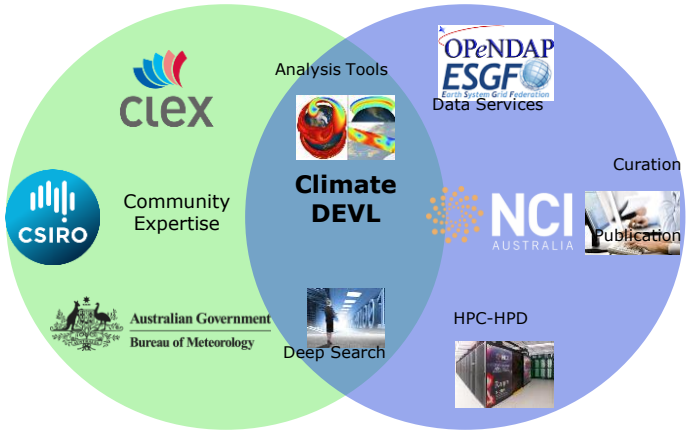
Source: https://wg1.ipcc.ch/AR6/AR6.html

Clare.Richards@anu.edu.au

Enhancing climate research capabilities in Australia

Climate Science Data-enhanced Virtual Laboratory (DeVL) project (2018-19)

- Reliable access to Climate and Weather data collections (local and internationally)
  - CMIP6 will be biggest climate dataset ever!
  - Earth System Grid Federation (ESGF) software upgrades
- Updated data analysis environments at NCI
  - Improved tools
- Improved FAIRness of data
  - Scalable data search
- User support and training



NCI AUSTRALIA

CSIRO

Australian Government
Bureau of Meteorology

ARC CENTRE OF EXCELLENCE FOR CLIMATE EXTREMES

ARDC
Australian Research Data Commons

Clare.Richards@anu.edu.au

ESGF Deployment and Data Replication

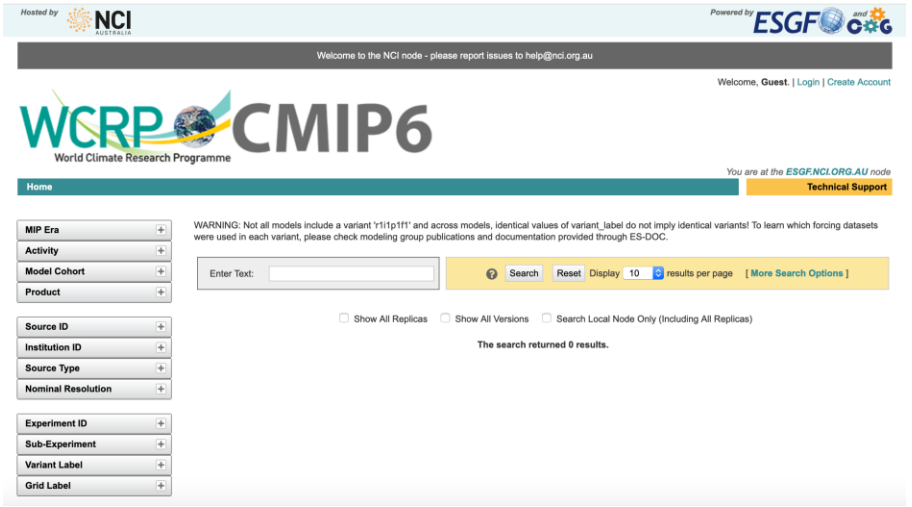
ESGF Deployment and CMIP data replication and management

a. International coordination - NCI Tier-1 site
b. New ESGF node deployed (https://esgf.nci.org.au)
    i. Components of stack are separate

a. Replication and management of international data
    i. Synda software deployed on DTN
    ii. Fully evaluated by redownload of CMIP5 (~615TB)
    iii. CMIP6 download approx. 42TB but expecting 3PB+
    iv. Majority of download mid-late 2019

a. Publishing Australian data to the ESGF
    i. Upwards of 400TB (mid year)

For more information on ESGF:
https://esgf.llnl.gov/



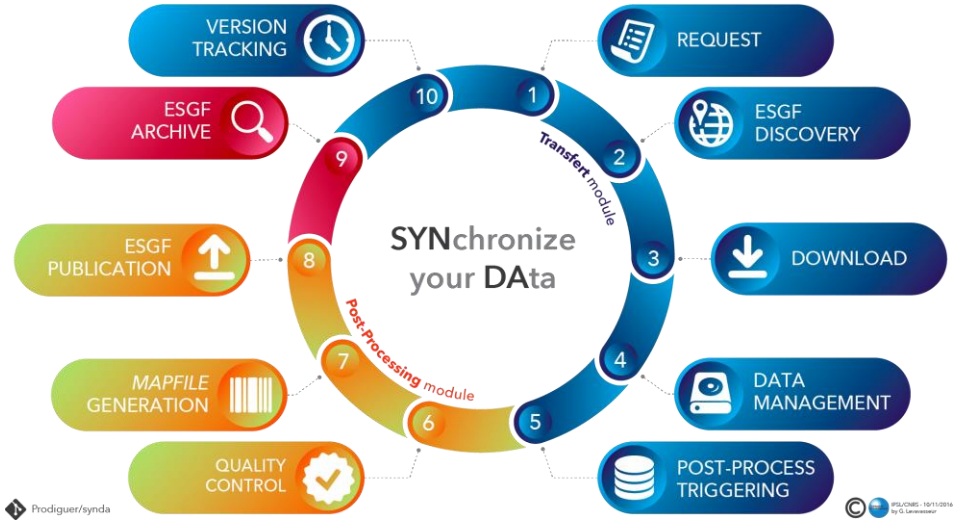


Image Source: http://prodiguer.github.io/synda/sdp/replication.html

Clare.Richards@anu.edu.au

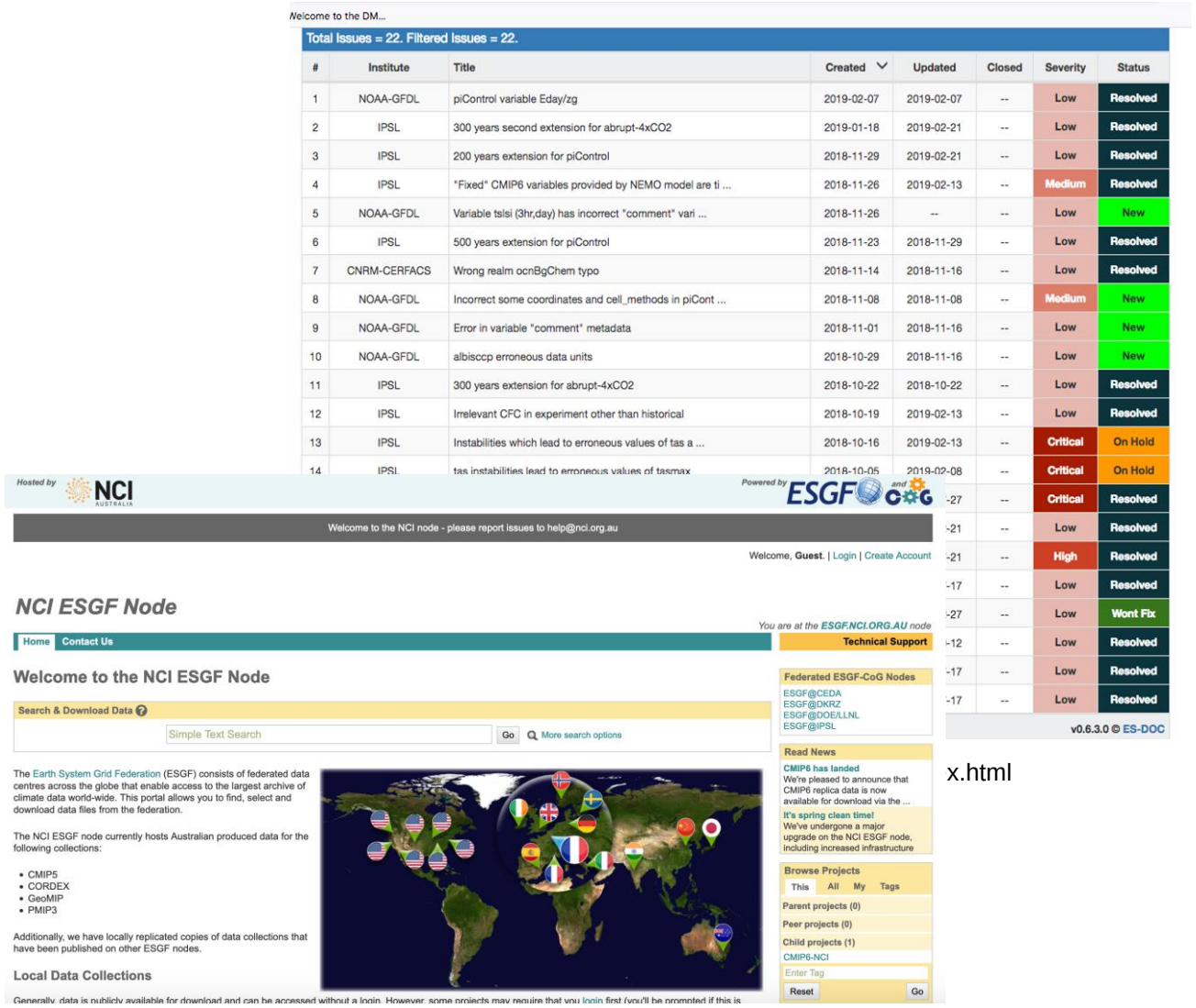
Internationally Federated Data Management

- Errata Service coordination and updates

- Data Management Planning
  - CMIP6 expected at 20PB, approx. 6PB likely requirement for Australian climate community.
  - Finite storage available.

- Data Management approach
  - Increased focus on quality and reliability

- Data Publication Services
  - ESGF
  - THREDDS
  - Modelling centre support

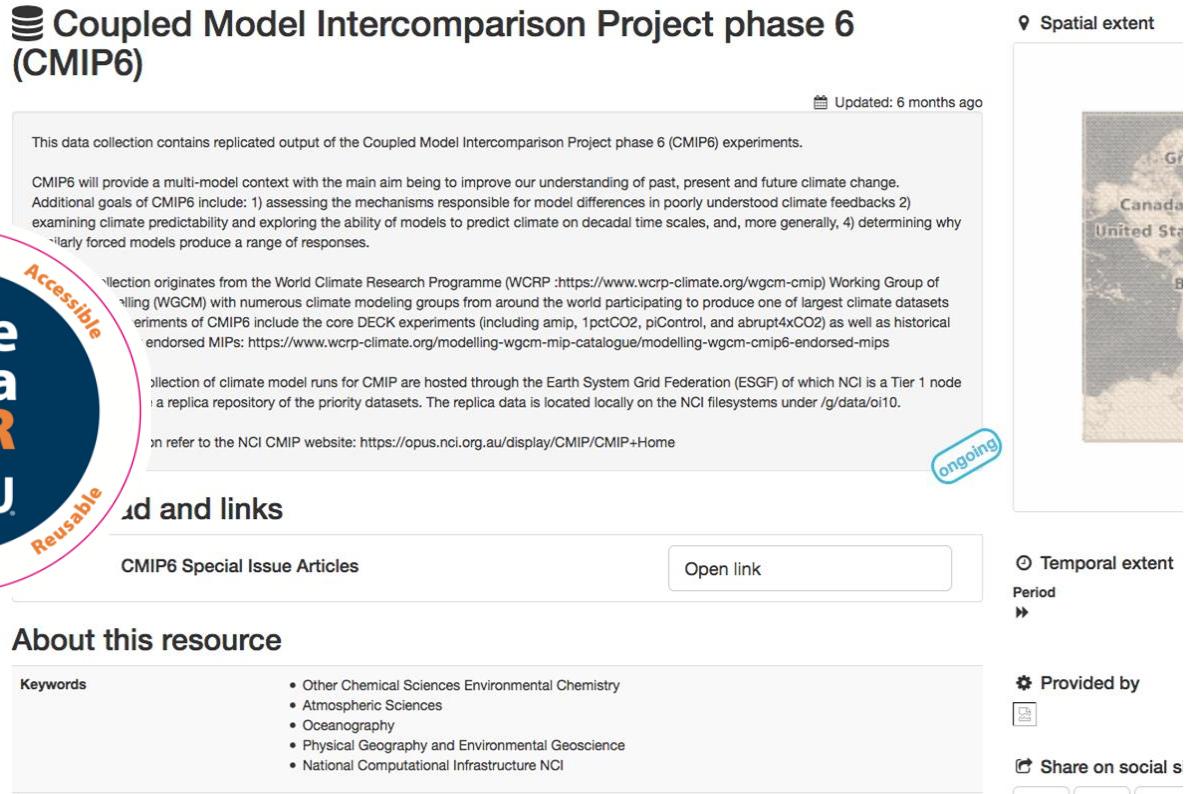


x.html

Clare.Richards@anu.edu.au

Improved Data Discovery

- CMIP datasets registered in NCI data catalogue (geonetwork.nci.org.au)
- ARDC's Research Data Australia (RDA) data discovery portal harvests these catalogue entries
- Will be harvested into Google Dataset Search



Improved Data Search and Availability

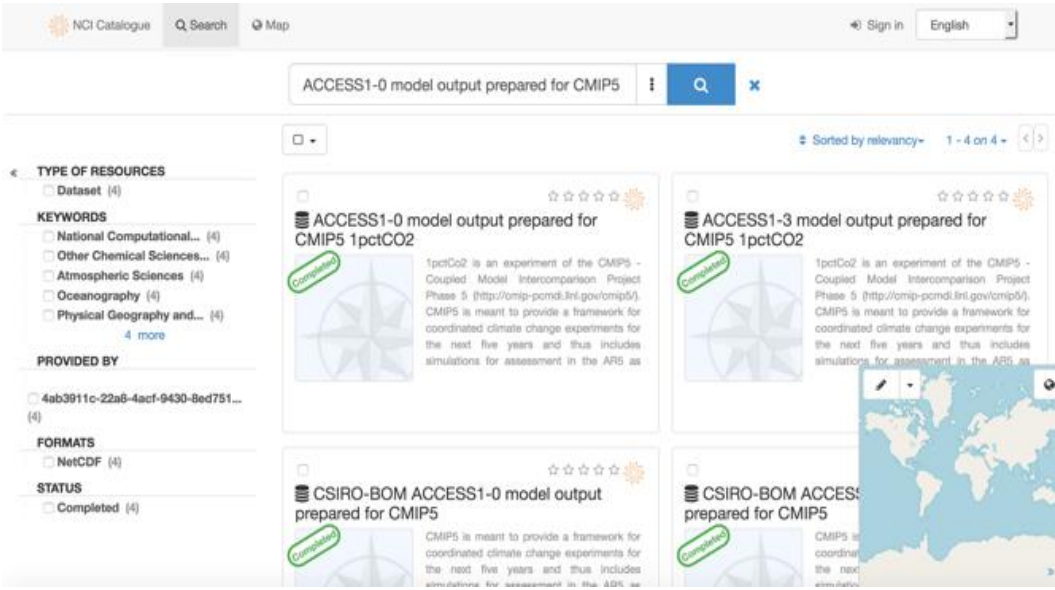
- Data index and search
  - All official CMIP data at NCI indexed by the NCI Metadata Attribute Service (MAS)
  - MAS integrated with the CleF API* for improved user search of large datasets
  - NCI ESGF node extended for overseas ESGF index

- Improvements to NCI infrastructure and services
  - Notification service for new data versions
  - Synda automated replication as new data becomes available
  - Automatic republishing to NCI ESGF node

```
[pxp581@vdi-n24 clef]$ clef --help
Usage: clef [OPTIONS] COMMAND [ARGS]...

Options:
  --remote   returns only ESGF search results
  --local    returns only local files matching ESGF search
  --missing  returns only missing files matching ESGF search
  --request  send NCI request to download missing files matching ESGF search
  --help     Show this message and exit.

Commands:
  cmip5  Search local database for files matching the...
  cmip6  Search local database for files matching the...
```



*See Claire Trenham's talk … on now!

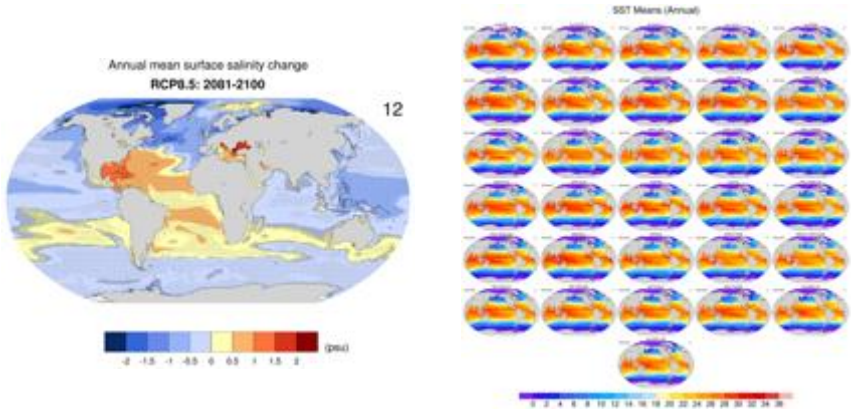
Clare.Richards@anu.edu.au

Best-practice tools and common workflows

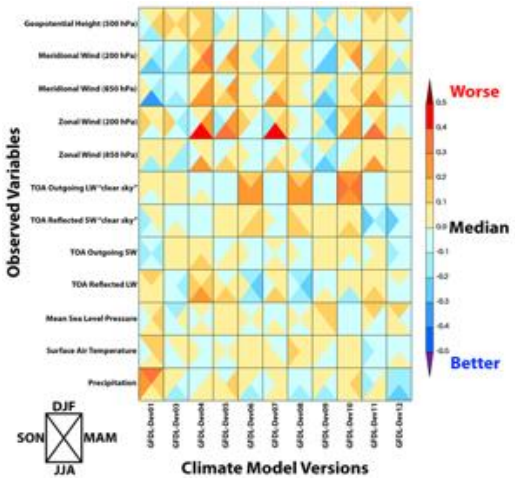
New analysis tools and updates to the existing tools

a. Review of the CSIRO CMIP pipeline*
b. Testing of international model evaluation tools for the CMIP6 data analysis
   - Earth System Model eValuation Tool (ESMValTool) – diagnostics and performance metrics
   - PCMDI Metrics Packages (PMP) – model evaluation and statistical error measures

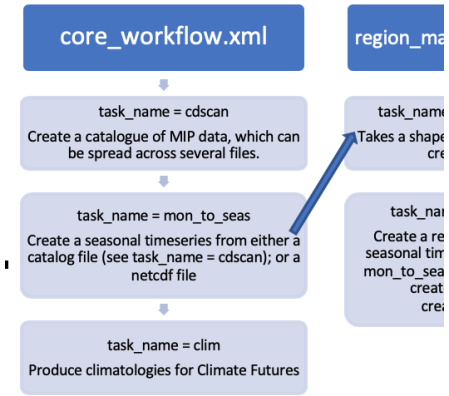
ESMValTool



PMP



CMIP Pipeline



*See Claire Trenham's talk … on now!

Clare.Richards@anu.edu.au

Data Analysis Tools and Environments

- Prototype of JupyterHub
  - Notebook based workflow
  - Host collection of core notebooks for climate analysis
  - Simpler training platform
  - Central administration of server environment
  - Investigation into Pangeo

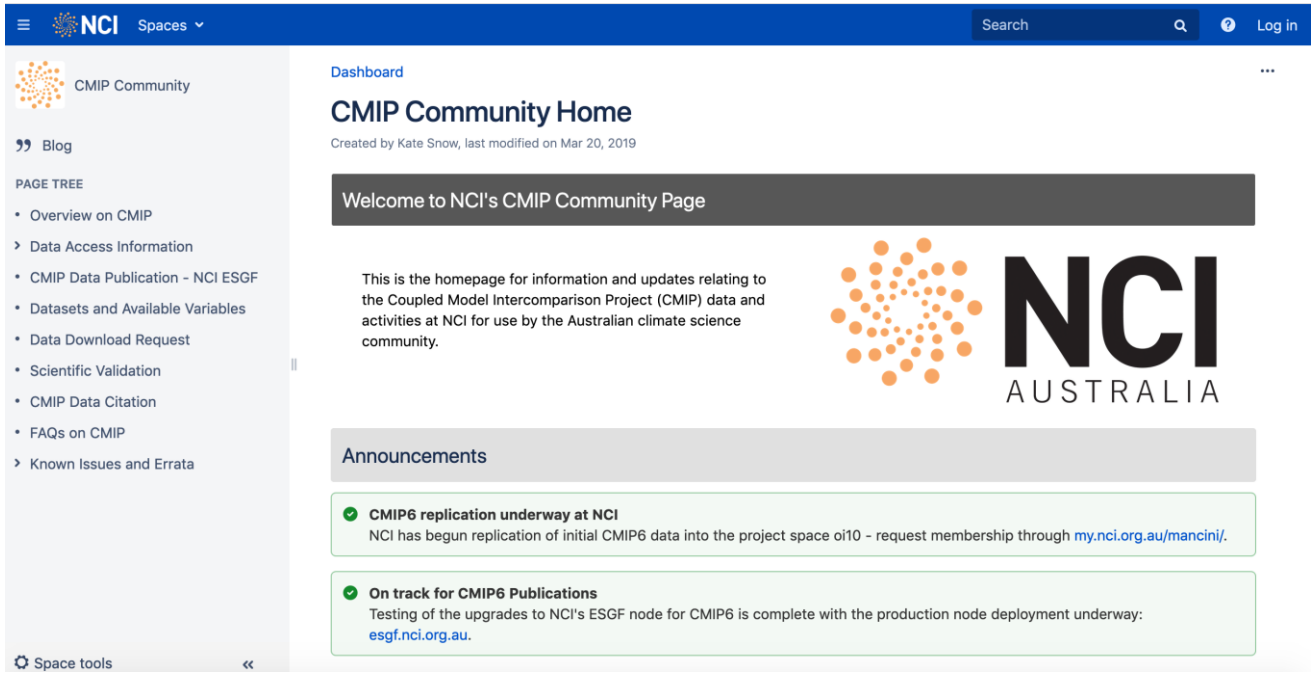
- Suite of R tools and workflows
  - Most climate support focused on Python so looking to understand R user requirements.

Clare.Richards@anu.edu.au

User Support and Training

User support - Documentation and Outreach/Training

a. User documentation and updated CMIP Community website at NCI.
b. New training material for the CMIP6 data and tools – to be available online.
c. Contact help@nci.org.au to discuss training opportunities and user support.



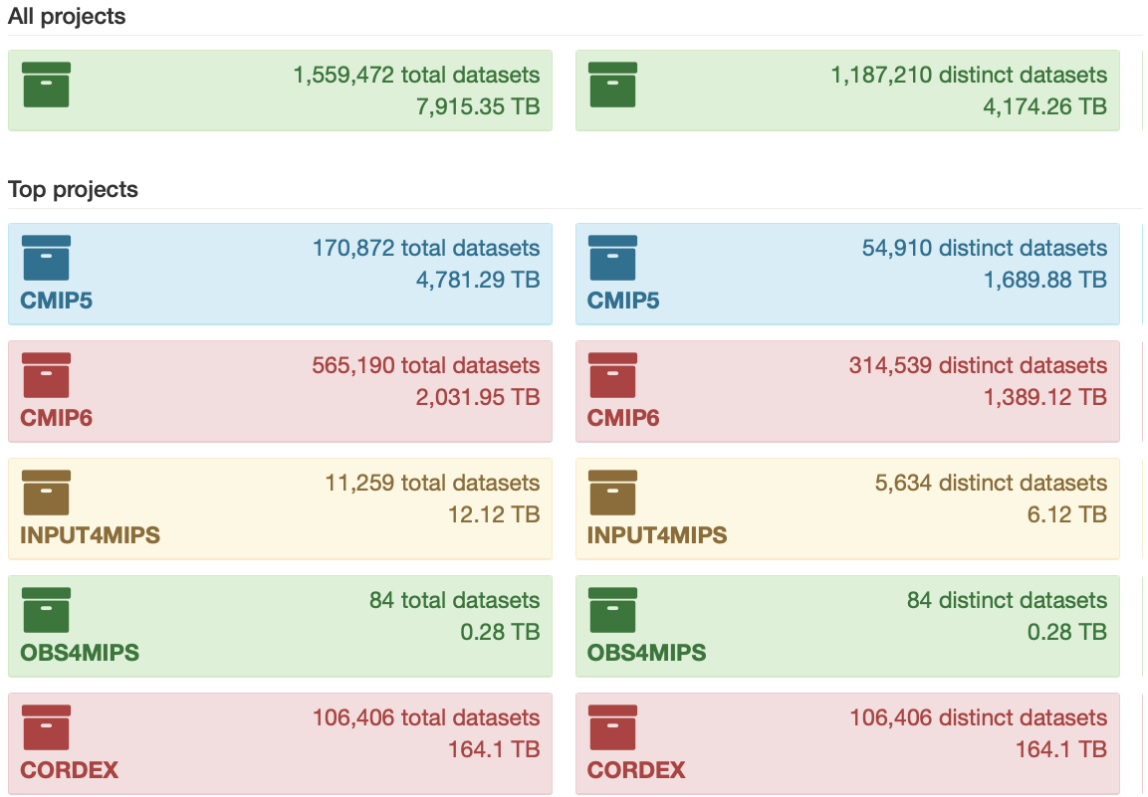
https://opus.nci.org.au/display/CMIP/CMIP+Community+Home

Case studies and sustainability

- Improved data usage and tracking
  - Improved tracking (CleF records all data searches, requests, project membership, ...) and reporting,
  - ESGF dashboard API and usage

- Development of case studies and user feedback.
  - Helps define priorities and future work requirements.
  - Demonstrates impacts.

- Establishment of ongoing sustainability requirements beyond June 2019.
  - Expected needs for CMIP6 in coming years
  - Maintaining infrastructure and services
  - Support and training



http://esgf-ui.cmcc.it:8080/esgf-dashboard-ui/pages/index.html

Clare.Richards@anu.edu.au

Preparing for the future

Making science efficient and reliable so researchers can focus on their work

- Supporting the Australian Climate Research Community for CMIP6:
  - Data Management, replication and publishing
  - Software and Tools for data-intensive analysis
  - Improved data access
  - User support and training
  - Integration with HPC – model development and data analysis

- Some challenges along the way:
  - Download of massive datasets
  - Storage availability

- Project ends mid-2019 but new requirements are emerging:
  - New data (ERA5 &?) and research objectives
  - Increasing requirements for tools and resources
  - Establishing sustainable ongoing support

Acknowledgements

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

Australian Government
Australian Research Council

Australian National University

Australian Government
Bureau of Meteorology

CSIRO

Australian Government
Geoscience Australia