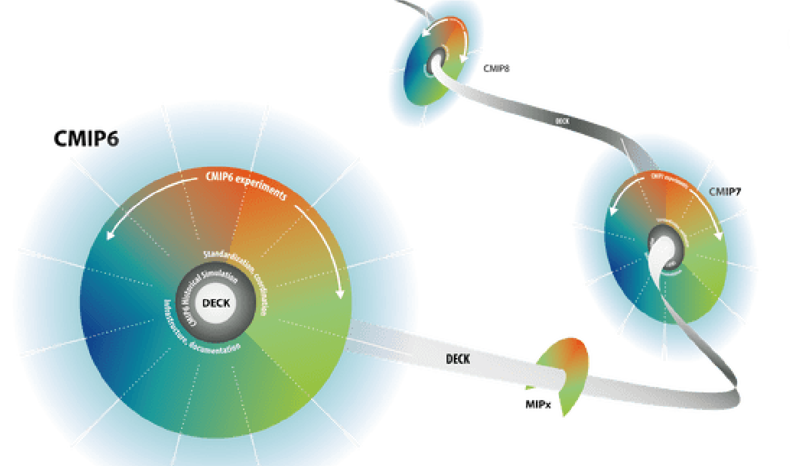

CMIP6
CMIP6 experiments
Standardization, coordination
CMIP6 Historical Simulation
DECK
Infrastructure, documentation
CMIP8
DECK
CMIP7
DECK
MIPx

Eyring et al., Geosci. Model Dev., 9, 1937-1958, 2016

Preparing for CMIP6

How to deal with multi-petabyte climate data collections

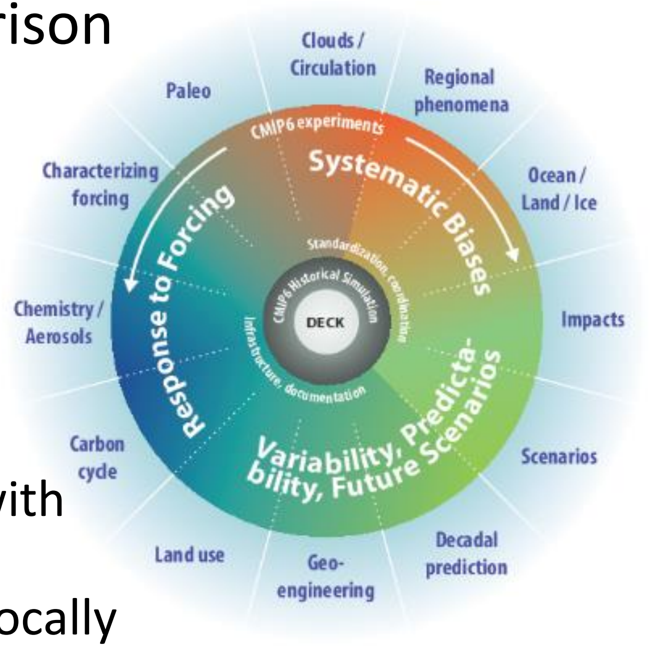
Claire Trenham | Tim Erwin, Aurel Moise, Paola Petrelli, Kate Snow, Louise Wilson, Vanessa Hernaman, Clare Richards, Craig Heady

7 May 2019

CSIRO CLIMATE SCIENCE CENTRE
www.csiro.au

What is CMIP6?

- The IPCC WCRP Coupled Model Intercomparison Project phase 6
- Largest collection of climate data to date
  - Expected total volume ~30PB
  - Currently being produced and published
- Australian climate community needs:
  - Local replica of commonly used datasets (~5PB) in Australia for local researchers to work effectively with this data
  - Be able to find what data is available globally and locally
  - Tools to work with these very large datasets



https://www.wcrp-climate.org/news/science-highlights/866-cmip-overview-paper

CMIP history in Australia

- CMIP3 (circa 2008)
  - ~30TB
  - Replicated ad hoc
  - Disks physically transported to Australia
- CMIP5 (circa 2012)
  - ~1PB of replicated + Australian model data
  - Replication a community effort by many volunteers across partner organisations downloading to NCI during Vayu/DCC/Raijin transition
  - Working through a google doc of requested variables!
  - Serious difficulties with data versioning
- CMIP6 approaching: We must not repeat the issues of the past!

What is the Climate DEVL?

- The Climate Data Enhanced Virtual Laboratory
  - An ARDC-funded collaboration between NCI, CSIRO, BoM, and CLeX to provide data access, tools and services for CMIP6.
- CMIP data is published through the Earth System Grid Federation
- NCI has
  - automated data replication for selected CMIP5/6 variables (synda tool);
  - deployed updated ESGF node and republished Australian CMIP5 models;
  - a mechanism to republish replicated data from overseas
  - developed a metadata database of CMIP data for search by CleF tool (MAS).
- For more info see Clare Richards’ talk (same time as this one!)

Information about CMIP6

- Website describing CMIP data and how to access it at NCI https://opus.nci.org.au/display/CMIP/CMIP+Community+Home
- Includes links to tables of available and requested data which will be synch'ed to NCI as it is published internationally
- Impacts for users
  - NCI manage ESGF data downloads now so **users should never download data themselves!**
  - Data is more searchable/findable than it was previously
  - Data is better documented (ESGF: version IDs, ES-DOC, errata, DOIs)

Tools for CMIP6 data in Australia

- **CleF (Climate Finder)** tool developed by CLeX & NCI
- **CMIP data processing pipeline** (CSIRO & BoM)
  - pre-process and analyse CMIP data against reanalysis/obs data for publications
  - being updated to support CMIP6 and the CleF database
- **ACCESS post-processing pipeline** (CSIRO)
  - To process the output of ACCESS CMIP6 model runs ready for publication
- Externally developed tools
  - **ESMValTool** for climate model diagnostics - under redevelopment
  - **PCMDI Metrics Package**
  - **CliMAF** and others

CleF: the Climate data Finder

- Search for data stored locally at NCI or available on the ESGF
- Developed by CleX Computational Modelling Systems team with NCI Data Services collaboration
- Can use CleF to find local and/or remote data (incl. version)

```
module use /g/data/hh5/public/modules
module load conda/analysis3-unstable
clef cmip6 -v tas -t Amon -e historical
```

- Use `clef --request` to generate download request for missing data
  - On VDI submits a request directly to help queue, on Raijin login nodes generates a request file to submit to help@nci.org.au

CMIP processing pipeline

- “The pipeline” was written in 2013-2016 by CSIRO
  - Combine CMIP5 processing and data analysis tasks into workflows for execution on NCI HPC
  - Python-based pipeline tool (using python2, CDO, NCO, R), available as a module
    - Relies on “patterns” to locate input data, which doesn’t work in new storage regime with data split across projects; outdated python modules
- Update in 2019
  - support CleF integration to specify input data locations - solves issue with new storage structure
  - review to determine which workflows do/don’t work in preparation for CMIP6
  - identify what needs updating to python3 and newer libraries
  - update documentation to reflect current state

Other analysis tools

- International community is developing tools akin to the pipeline
- ESMValTool Earth System Model eValuation Tool
  - Diagnostics and model performance metrics tool
  - Very promising but v1 problematic and v2 not yet out of development phase
  - https://www.esmvaltool.org/
- PCMDI Metrics Package
  - Aimed primarily at modelling centres investigating model performance
  - https://github.com/PCMDI/pcmdi_metrics/wiki
- Other things: many, including
  - CliMAF https://climaf.readthedocs.io/en/master/
  - Pangeo https://pangeo.io/ (not really an analysis tool)

Local infrastructure NCI

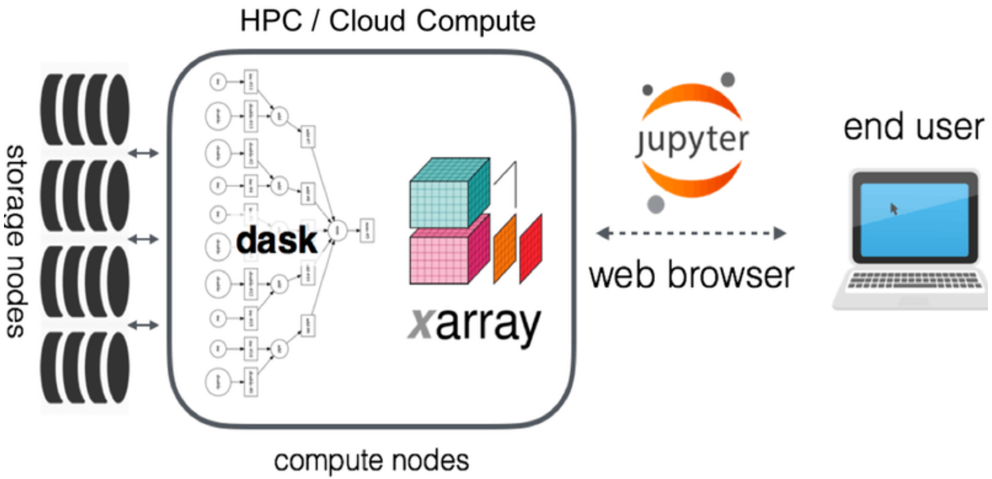
- NCI provides vital infrastructure for the Australian climate research community to work collaboratively
  - Data storage of high priority data for use by multiple researchers
    - reduce unnecessary replication of data
  - Access to HPC: run climate models; execute parallelised data processing tasks
  - Cloud-based Virtual Desktop Infrastructure for data exploration and analysis
  - Capacity for at-scale jupyter/xarray/dask workflows
- NCI ESGF node publishes Australian model output for CMIP
  - Interface to global ESGF community
  - Republish replicated data locally

Where to next? - hardware

- Data storage: we will likely need more than we currently have at NCI!
  - ACCESS-ESM1.5 and ACCESS-CM2 development and output; downscaling models using CMIP data; related models…
  - CMIP6 bulk replication (as well as other data from ESGF)
  - Observations and reanalysis data for model evaluation and comparison
  - Post-processed analysis data for publication (journals, reports, websites, brochures, advice)
- HPC: Available capacity?
- Cloud: Pangeo?

Where to next? - software

- Future of CMIP pipeline
  - update to python3?
- Contribute to ESMValTool
  - e.g. diagnostics for the Southern Hemisphere
- What can we do with Pangeo?
- International collaboration!
- ACCESS development



When will the data be available?

- Preliminary data started arriving in August 2018
- As of May 2019, have data available from 18 models at NCI
    - 40TB CMIP6 downloaded so far (and 630TB from CMIP5)
    - Mostly from DECK (historical, piControl, 1pctCO2 etc)
    - 7 models in ScenarioMIP (SSP projections)
    - Tracking page for NCI downloads http://atlantis.nci.org.au/~kxs900/cmip_tables/index_CMIP6.html
        - But note we can’t track download of data that hasn’t been published yet!
- We are conscious of IPCC AR6 publication timelines

Thank you

Climate Science Centre
Claire Trenham
Experimental Scientist

Claire.Trenham@csiro.au

How to find CMIP data?

- **Problem:** Data is now stored across multiple projects at NCI
  - need to join each project whose data you will use.
- **Solution 1 (preferred):** use CleF to locate data (CMIP5 or CMIP6)
  - Data paths on NCI's /g/data filesystems for locally stored data are returned, use my.nci.org.au to join projects to access the data.
- **Solution 2:** ua6 symlinks (CMIP5 only)
  - /g/data/ua6/DRSv3 is a symlink tree pointing to both rr3 (Australian) and al33 (replicated) data
- **Solution 3:** Search ESGF https://esgf.nci.org.au/projects/esgf_nci/ for globally published data (note different interfaces for CMIP5/6)
- NCI *may* provide symlink trees between related data collections

What data will be available?

- Data priorities set based on CMIP5 data replication and community survey
- All ACCESS-CM2 and ACCESS-ESM1.5 output
- Replicated data:
  - Focus on DECK and ScenarioMIP
  - Smaller variables (e.g. monthly frequency, or surface variables) will be higher priority than high frequency and 4D variables which take a long time to download and require a lot of space which is currently at a premium at NCI.
- We are volume limited, if we had more storage we could replicate more data in the medium-long term.
- Access to data via Pangeo CMIP6 intake??

How do I get access to the data?

- Join appropriate projects and agree to ESGF Terms of Use

NCI project	Data
rr3	Australian CMIP5 era data (incl CORDEX1)
al33	Replicated CMIP5 data (replaces ua6/unofficial-ESG-replica)
ua6	“unofficial” CMIP5 data – to be decommissioned
oi10	CMIP6 replicated data
TBA	Australian CMIP6 data
qv56	ESGF obs & reanalysis (e.g. Input4MIPs, Ana4MIPs)
cb20	CMIP3 data

- Join projects via https://my.nci.org.au
- Ask for help: cws-help@nci.org.au