ESGF within NCI
NCI part of Government’s broad National Research Infrastructure Review, which includes nearly all large scale national infrastructure

National “mood” has largely moved from century scale to seasonal-to-decadal

CMIP activity seen as important national role for ACCESS/Climate research but pressure for more practical outcomes. e.g. food security, agriculture, water mgt, environmental, urban impacts.

CMIP part of a number of significant data resources that are important to make interoperable with HPC/HPDA.

Value of NCI acknowledged but still large number of questions about funding futures. e.g. who pays for storage

NCI has a cloud but should we have more in commercial clouds and have users pay and is it beneficial to us?
NCI Setting

- NCI has top 100 supercomputer, a high performance OpenStack cloud, large Lustre site-wide storage for datasets, and deep tape archive.

- NCI other key environmental, earth system and solid earth simulation, data and data analysis. CMIP is currently ~10% of the total volume of interest. i.e. 10% of data under ESGF type management.

- Has developed processes for managing all types of simulation and observation data inc point and line data, licensed and some auth.

- Datasets include ([http://datacatalogue.nci.org.au.au](http://datacatalogue.nci.org.au.au))
  
  - 1. Climate/ESS Model Assets and Data Products
  - 2. Earth and Marine Observations and Data Products
  - 3. Geoscience Collections
  - 4. Terrestrial Ecosystems Collections
  - 5. Water Management Collections
• Major model to be published from Australian will be the ACCESS-CM2 model and possibly an ESM model.
• Additional MIPs and CORDEX data to be published, but expect smaller and perhaps more obs4MIP
• Expect replicated data to be ~5 Pbytes
• Master node of the Australian datasets (including CMIP5 era)
• full replica of the international index
• ongoing replication of the “key variables”, and republishing
• Expect more systematic assignment of DOIs
• Data replication with major ESGF nodes via gridftp end-points
Computational and climate data services

- Most serious users will access via direct access to NCI filesystems using either supercomputer, VDI access or virtual labs
- Data will be indexed via our standard (non-ESGF) data catalogues – including geospatial index as well as catalogue systems (geonetwork + data discovery portal)
- Publishing data via non-authenticated OPeNDAP and OGC services
  - Expect more via authenticated methods in future
- Climate expect more data portal services (e.g., climate4Impact and other international examples available locally)
- Expect more WPS service (e.g. birdhouse, zoo, bespoke)
Data Analysis and visualisation

- In-situ and data service approach to data analysis
  - Need to move away from “data download/shopping cart”
  - Non-bespoke authentication/auth approach to access data
- ESGF/UV-CDAT tools available within a broader VDI service that allows interactive login.
- Broad range of data analysis, workflow and visualisation requirements that goes across very broad science domains
- Data available in-situ and via services for other workflows and publication
- Data workflows with hooks for reproducibility, PROV standard capture and publishing
Operational matters

- Puppet-delivered deployments in VMs using full repeatable processes for redeploy. RPMs preferred.
- Old-style bash script approach and partial upgrade to hack solutions are not sustainable.
- Docker not supported (as yet) because of security concerns.
Absolutely top priority issues for ESGF data for now

1. Robust publishing and well-known procedures for making data available at nodes and how available internationally
2. Full documentation about datasets available.
3. Advertising international dataset availability and changes
4. Automated and fast data replication and updates for node-subscribed datasets/variables onto the local node and to remote
   1. Must be tested and work with data publication and data mgt.
5. Bullet-proof software and data update processes (i.e., test environment first)
6. Ensure that all other compute and viz software developments are available across the nodes. (I see as a major gap at the moment).
• NWP and Forecasts
  UM, APS3 (Global, Regional, City), ACCESS-TC
• Coupled Seasonal and Decadal Climate
  ACCESS-GC2/3 (GloSea5)
• Data Assimilation
  3D-VAR, 4D-VAR (Atmosphere), EnKF (Ocean)
• Ocean Forecasting and Research
  OceanMaps, BlueLink, MOM5, CICE/SIS, WW3, ROMS
• Fully-Coupled Earth System Models

• Water availability and usage over time
• Catchment zone
• Vegetation changes
• Data fusion with point-clouds and local or other measurements
• Statistical techniques on key variables
NCI GSKY: Himawari-8 Data processing on the fly through standard OGC API
Data Quality Strategy (DQS): What does it involve?

1. Underlying HPD file format
2. Close collaboration with data custodians and managers
   - Planning, designing, or reassessing the data collections
3. Quality control through compliance with recognised community standards
4. Data assurance through demonstrated functionality across common platforms, tools, and services

DATA SERVING: THREDDS, OpenDAP, WMS, etc.
DATA USAGE: Matlab, R, Python, GDAL, etc.

Digital Object Identifiers (DOI) minting, Making metadata/data available and discoverable online

FILE (GRANULE)-LEVEL
- Climate and Forecasts (CF) Convention
- Attribute Convention Dataset Discovery (ACDD)
- Additional discipline specific standards

COLLECTION & DATASET-LEVEL
- Data Management Plans (ISO 19115, ANZLIC, etc.)

Self-describing file formats (e.g., NetCDF, HDF)
<table>
<thead>
<tr>
<th>Program/Service</th>
<th>Test</th>
<th>File 1</th>
<th>File 2</th>
<th>File 3</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NetCDF Utilities</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>netcdf (v.4.3.3.1)</td>
<td>Read netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>NCO (v.4.5.2)</td>
<td>Read netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>GDAL Utilities (v.1.9.3)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gdalinfo-1</td>
<td>Read netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>gdalinfo-2</td>
<td>Read netCDF CRS information.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Data Viewers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>netcdf-viewer (v.2.1.1)</td>
<td>Visually inspect netCDF contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Parasol (v.4.6.1)</td>
<td>Read and plot netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>THREDDS Data Server (v.4.4)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ONESAP (access and submitting)</td>
<td>Read/submit netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>NetCDF Subset Service (NCSS)</td>
<td>Request subset of netCDF contents using spatial/temporal query.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Gridne Viewer</td>
<td>View netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>WMS GetMap (v.1.1.1)</td>
<td>Request netCDF file using WMS.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>WCS GetCoverage (v.1.0.0)</td>
<td>Request netCDF file using WCS.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Python (2.7.x)</strong></td>
<td>netCDF4-python (v.1.2.2) Read/inspect netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>GDAL-py (1.11.1)</td>
<td>Read/inspect netCDF file contents.</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>h5py (v.2.5.6)</td>
<td>Read/inspect netCDF file contents.</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><strong>MATLAB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2012b</td>
<td>Read/inspect netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>R2012b</td>
<td>Read/inspect netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>R2015a</td>
<td>Read/inspect netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>R2018a</td>
<td>Read/inspect netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>R (v.3.1.8)</td>
<td>nco (v.1.15) Read/inspect netCDF file contents.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>QGIS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v2.2.0 Valmiera</td>
<td>Add data from netCDF as raster layer</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Add data as WMS layer (served by THREDDS)</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><strong>Visualisation Tools</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ParaView** (v.5.8.1)</td>
<td>Read/visualise netCDF file.</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>
• Quality Assurance against performance metrics

• We need to scale data so that you can analyse in real-time and in-situ.

• Need to combine/overlay/slice-dice all manner of data at high precision from vast reference with highly specific data.

• We need faster, automated systems for real world activities, decision making capability using smart new algorithms and programmatic techniques:

  - real data feeds, cross-referencing longitudinal data, geospatial or other key "metadata" queries.
PROMS v3 uses an extension to the PROV ontology as its data model.

- Entities
- Activities
- Agent


The Provenance Service captures information at each step within the end-to-end workflow, and stores it within the Provenance Repository.