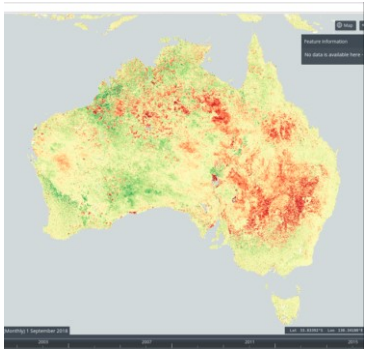
Australian eResearch & Data Skills Summit 2020

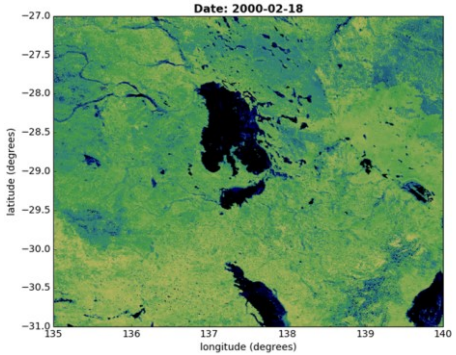
Successful Data Training Story at National Computational Infrastructure

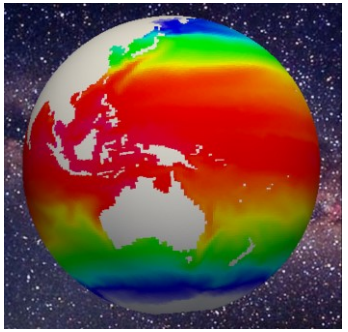
Jingbo Wang and Ben Evans

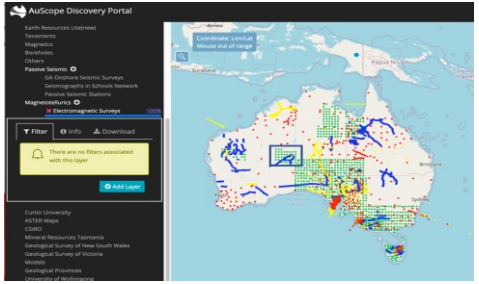
10PB+ Research managed repository of datasets:

- tuned within NCI for computational intensive methods and data analysis
- made available through data services for broader access and informatics









NCI AUSTRALIA

Welcome to the GSKY Manual

User Guide

High performance data services

nci.org.au

Setup for Pangeo Environment on Gadi

In this notebook we will go through:

- Configuring the Pangeo environment using your Gadi account
- Submitting and monitoring multi-node Pangeo jobs to Gadi
- Running Pangeo Jupyter notebooks in batch mode non-interactively

The Pangeo software ecosystem involves open source tools such as xarray, iris, dask, Jupyter and many other packages.

This notebook provides instructions on how to use the Pangeo environment to run a multi-node parallel Jupyter notebook within the queues on Gadi and shows how to interact with it from your desktop.

Configuring your account on Gadi

Step 1: Enabling Pangeo in your shell envorinment

To enable the Pangeo environment, you can use the following command within jobs, or within an interactive environment:

```
$ module load pangeo
      Loading pangeo/2020.05
        Loading requirement: intel-mkl/2019.3.199 python3/3.7.4 hdf5/1.10.5 netcdf/4.7.3
```

Note that Pangeo has its own Python installation.

Step 2: Configuring your JupyterLab password on Gadi

We will use JupyterLab to load notebooks and monitor jobs. JupyterLab is bundled within the Pangeo environment. To setup this environment, run the following two commands:

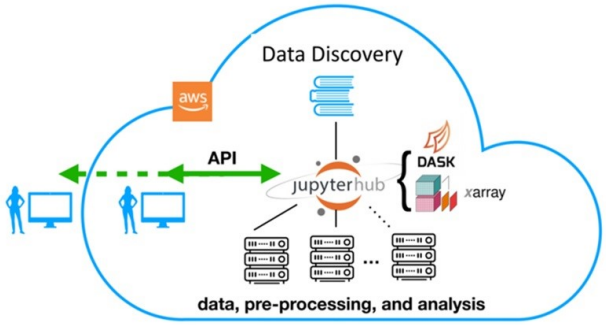
```
$ jupyter notebook --generate-config
$ jupyter notebook password
```

This is a stand-alone password that you will use later for accessing the JupyterLab server. You can use this command to reset your password at any time.

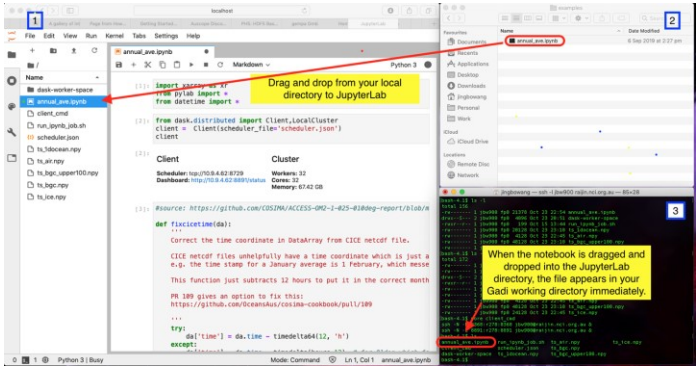
Step 3: Exiting the Pangeo environment

Python research environment

Pangeo's open-source ecosystem

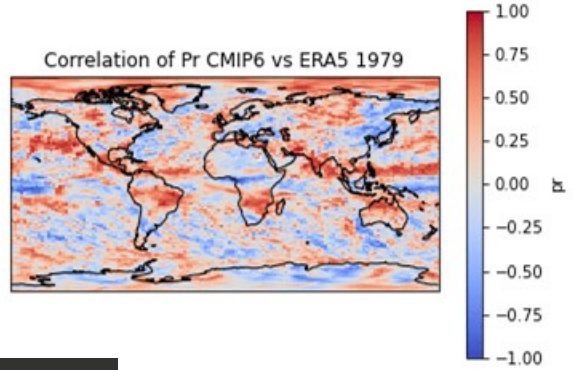


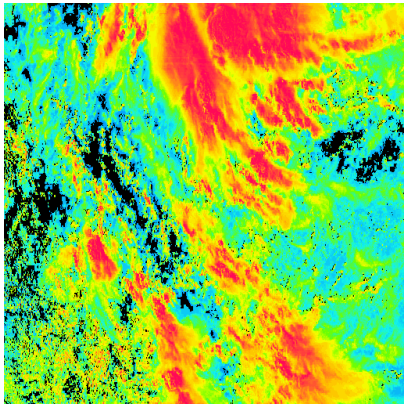
(Image courtesy of Pangeo)



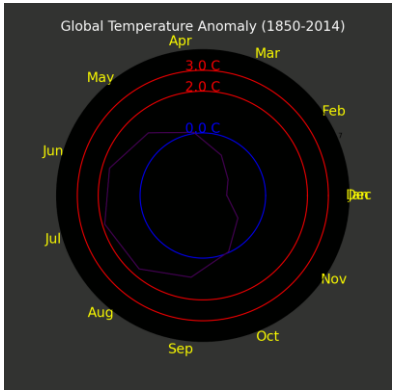
Current data training focus:

- Awareness of the datasets
- Awareness of how to use data services
- Awareness of data analysis platforms

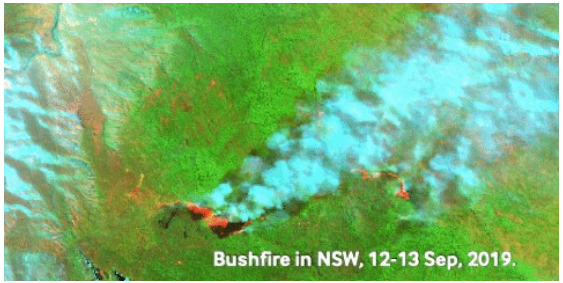




Cyclone Debbie 2017



Temp anomaly spiral

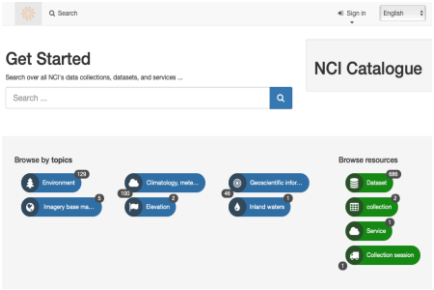


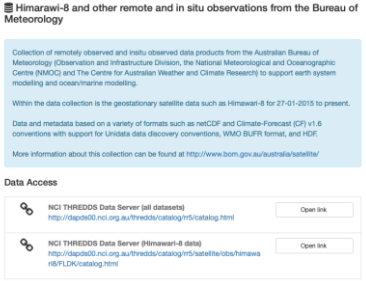
Bushfire 2019 from National Map

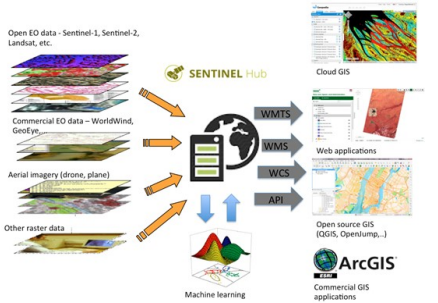
What do we teach?

Training topics includes:

- How to search the datasets
- How to access the datasets
- What is a data service
- What data service NCI offers to our users
- How to use NCI's data services
- What is data portal
- How NCI contribute various data portals
- How to process data using various libraries
- How to improve your computation performance
- How to scale up your program to work with large dataset







Current training materials in dev

Training materials	Purpose
Presentation	Share/update information
Wiki, webpages	Share/update information
Jupyter notebooks	Data analysis examples
User guide	Self learning on users' own pace

User guide and Jupyter notebooks are the focus in 2020 due to more compute capacity on Gadi and cloud-based data analysis platforms.

Training materials development

Training course curriculum is designed based on information from a few channels:

- Regular interaction with stakeholders
- Help desk tickets (label tickets and NLP analysis)
- Domain requirements (e.g., CMIP6 and ERA5 information session)
- Highlight newly released datasets and data services
- Build experience with how to programmatically use data
- Learn from software/data carpentry

This is the most time-consuming part!

Training materials development

Pages / NCI Help / NCI Data Training

NCI Data Webinar

Date: staring on the 10th of September, 2019, with a general plan for sessions fortnightly

Times: 11am for 1 hour (Canberra time)

Registration: https://anu.zoom.us/webinar/register/WN_r61cvTvvTG-c5kULHSgm0w

Zoom Meeting ID: https://anu.zoom.us/j/440830453

Scope of the webinar:

- Walk-through of various data services at NCI, including how to find datasets and access them
- Information on specific datasets, including a chance to talk to key data owners and users
- Information on NCI dataset management
- Examples of how to use the datasets using popular applications, portals, software, Virtual Research Environments
- Other relevant use-cases for specific communities

Who should attend: NCI data users, general researchers who are interested in the following domains such as climate and weather, earth observation, geosciences, hydrology, environment, astronomy, and son on.

Format: The first part will be a talk (30-40min), followed by Q&A.

Agenda:

Date	Topic	Presenter	links to the material
10 Sep 2019	Introduction to the NCI's Managed Datasets	Jingbo Wang	slides.pdf
24 Sep 2019	Introduction to NCI's Virtual Desktop Infrastructure (VDI)	Jingbo Wang	slides.pdf
8 Oct 2019	Introduction to NCI's Geospatial Data Server (GSKY) (part I)	Jingbo Wang	slides.pdf
22 Oct 2019	Introduction to NCI's Geospatial Data Server (GSKY) (part II)	Jingbo Wang	slides.pdf
5 Nov 2019	no webinar. NCI Training, 5 Nov, 2019 Australasian Leadership Computing Symposium (ALCS) 6-8 Nov, 2019 register here if you are interested.		

CMIP Community

Blog

PAGE TREE

- Overview on CMIP
- Data Access Information
- CMIP Data Publication - NCI ESGF Node
- Datasets and Available Variables
- Data Download Request
- CMIP Data User Tutorials and Training
- Scientific Validation
- CMIP Data Citation
- FAQs on CMIP
- Known Issues and Errata

Pages

CMIP Community Home

Created by Kate Snow, last modified by Clare Richards on Nov 20, 2019

Welcome to NCI's CMIP Community Page

This is the homepage for information and updates relating to the Coupled Model Intercomparison Project (CMIP) data and activities at NCI for use by the Australian climate science community.

Acknowledgements

Announcements

CMIP6 replication underway at NCI
NCI has begun replication of initial CMIP6 data into the project space oi10 - request membership through my.nci.org.au/mancini/.

On track for CMIP6 Publications
Testing of the upgrades to NCI's ESGF node for CMIP6 is complete with the production node deployment underway: esgf.nci.org.au.

20 Mar 2019 **CMIP6 Data download continuing as data becomes available**

Currently ~30TB of CMIP6 data replicated to NCI under project oi10.

20 Mar 2019 **Retraction of CMIP6 data at NCI**

To keep up to date on downloaded data affected by logged errata, watch the CMIP6 Dataset Errata page.

01 Aug 2018 **Initial release of CMIP6**

CMIP6 data is available for download and analysis under a trial period.

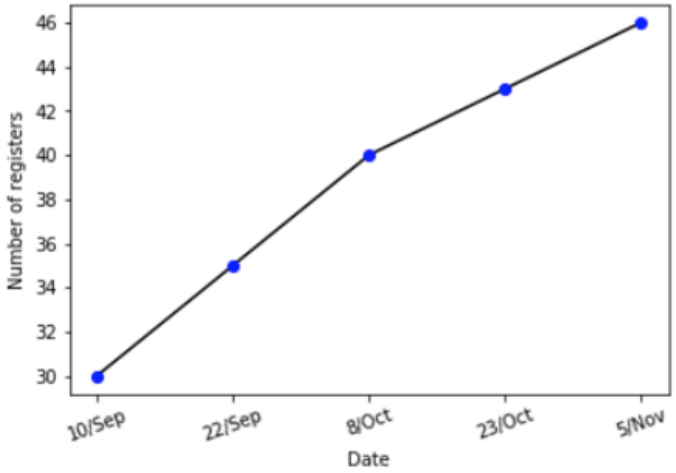
CMIP6 Status

The NCI ESGF portal to CMIP6 model output is open. Contributions from modelling centers is expected to continue throughout 2019 and 2020.

Fortnightly data webinar

1. Overview of NCI's data collection
2. Introduction to Virtual Desktop Infrastructure
3. Introduction to GSKY – part I
4. Introduction to GSKY – part II
5. Electromagnetic Geophysical Data

Number of registrations for the data webinar



Major training event in 2019

AMOS half day training course

- CMIP information update
- Xarray and dask examples

Part of NCI's Climate Science Data-enhanced Virtual Laboratory project.

Notebook examples

Data Access and Quick Preview

- ncdump
- ncview
- GDALinfo
- Python_GDAL_netCDF
- Search and Access CMIP5 data using Clef and Xarray
- Search and Access CMIP6 data using Clef and Xarray

Data Manipulation

- CDO - calculate monthly anomally and Nino Index using CMIP6 data
- CDO - compare model and observational data
- CDO - mask ocean or land
- Xarray - open and read data
- Xarray - subset and plot data (sliding and dicing)
- Xarray - calculate surface temperature anomalies in Australian
- Xarray - calculate Nino 34 time series
- Xarray - common operations, resampling, rolling and climatology
- Netcdf Subset Service (NCSS) with Python and Siphon
- Panoply CMIP5
- Panoply eReef and ANU water
- Paraview

Major training event in 2019

NCI hosts a global Hackathon (G)EOHack19 that ran a series of regional consultations in various Indigenous communities around the world to identify challenges that can be addressed using open EO data.

Teams across Alaska, Germany and Australia had 36 hours to develop a solution to these challenges. The winners, Team Triton from Australia, were announced at the GEO Week Plenary.

Major training event in 2019

One day High Performance Data Training, topics include:

- Introduction to NCI's national reference datasets
- How to set up python environment on your local computer, Virtual Desktop Infrastructure
- How to manage the I/O performance using data chunking
- How to set up Pangeo environment on NCI's supercomputer platform.

ALCS 19
AUSTRALASIAN LEADERSHIP COMPUTING SYMPOSIUM

Australasia's research supercomputing users' forum, as well as a flagship promotion of high-performance computing (HPC) and high-performance data (HPD) in Australasia.

Training delivery format - workshop

	Room A	Room B
Morning Block 9:30am - 12:30pm	Intro to HPC Getting started on Gadi – for new users	HPD High performance data training session The NCI National Research Data Collection is Australia's largest collection of research data, encompassing more than 10 PB of nationally and internationally significant datasets. They are available through NCI's National Environmental Research Data Interoperability Platform (NERDIP), which provides an integrated platform for users to access datasets managed at NCI. At the high performance data training session, we will offer introduction on NCI's research data collections and jupyter notebook examples on accessing data through various services and platforms.
Afternoon Block 2:00pm – 5:00pm	Transitioning from Raijin to Gadi Are you a Raijin user who wants to make a smooth transition to Gadi	High performance data training session The NCI National Research Data Collection is Australia's largest collection of research data, encompassing more than 10 PB of nationally and internationally significant datasets. They are available through NCI's National Environmental Research Data Interoperability Platform (NERDIP), which provides an integrated platform for users to access datasets managed at NCI. At the high performance data training session, we will offer introduction on NCI's research data collections and jupyter notebook examples on accessing data through various services and platforms.

We offer different levels of training modules:

Attendees' preference:

HPC Intro = 7
HPC advanced = 25

HPD Intro = 10
HPD advanced = 25

Major training event in 2020

One day High Performance Data Training, topics include:

- Introduction to NCI's national reference datasets
- How to set up python environment on your local computer, Virtual Desktop Infrastructure
- How to manage the I/O performance using data chunking
- How to set up Pangeo environment on NCI's supercomputer platform.

The 4th Australian Climate and Water Summer Institute (ACAWSI).

The Summer Institute was held from 3rd February – 14th February, 2020 at Fenner school, ANU.

Lesson learned

Successes

- Collaboration with data experts
- Help desk is informative on building training materials
- Webinar is popular
- Online tutorials for self learning
- Training motivates new topics!

Challenges

- Satisfying user demand
- Address user questions
- Keep material updated
- Effort to promote training events