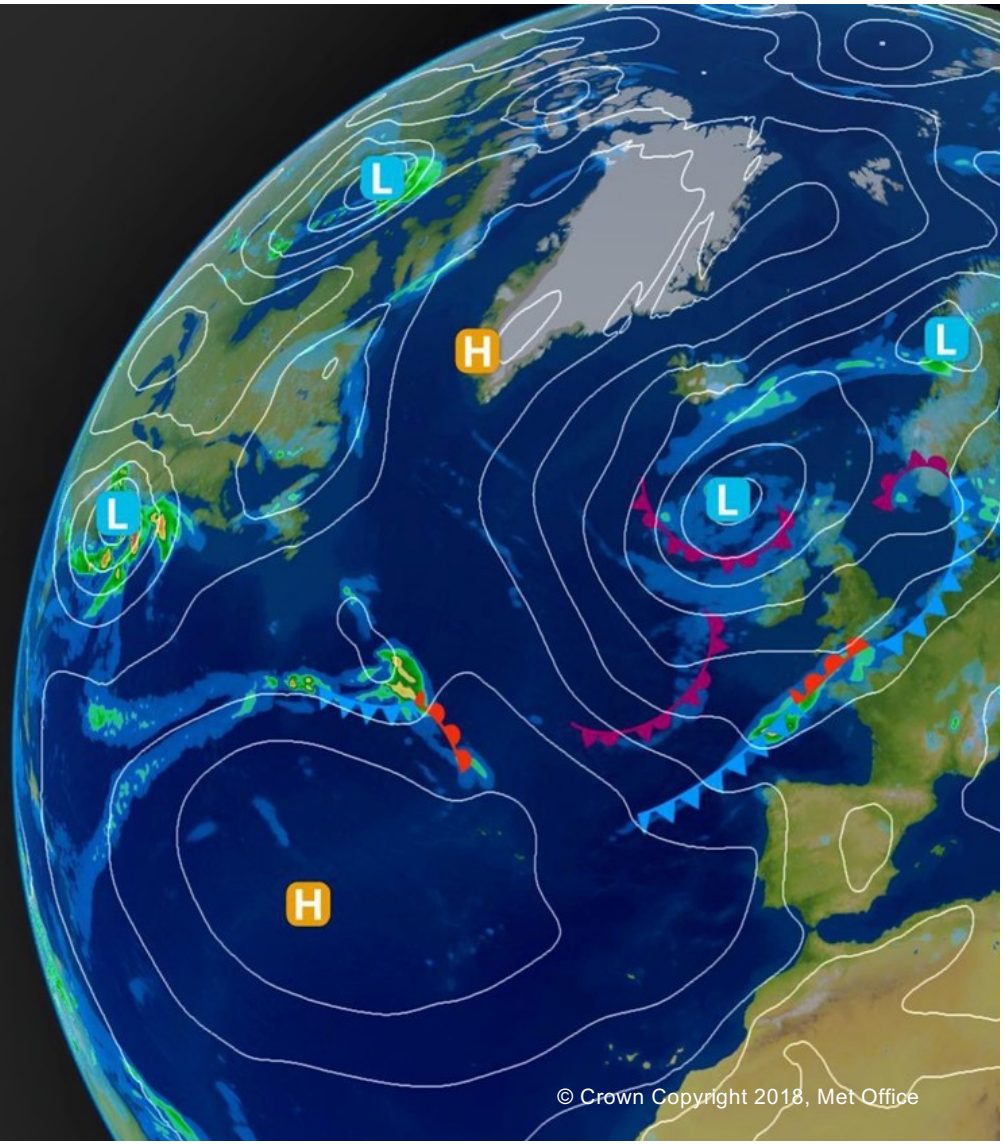
Met Office
Hadley Centre

Analysis facilities and Cloud

Matthew Mizielinski
Paul Durack
and the WIP

WGCM-24
9th December 2021 - Virtual



Met Office
Hadley Centre

Contributors to CMIP6

ESGF nodes

Analysis facilities

https://pcmdi.llnl.gov/CMIP6/

Met Office
Hadley Centre

NCAR: GLADE & Casper

- 100 node Casper cluster including high-memory, GPU and high throughput nodes
- Access to local archive of CMIP, and other project data on GLADE storage facility
- Access available to NSF funded researchers and US university staff and students

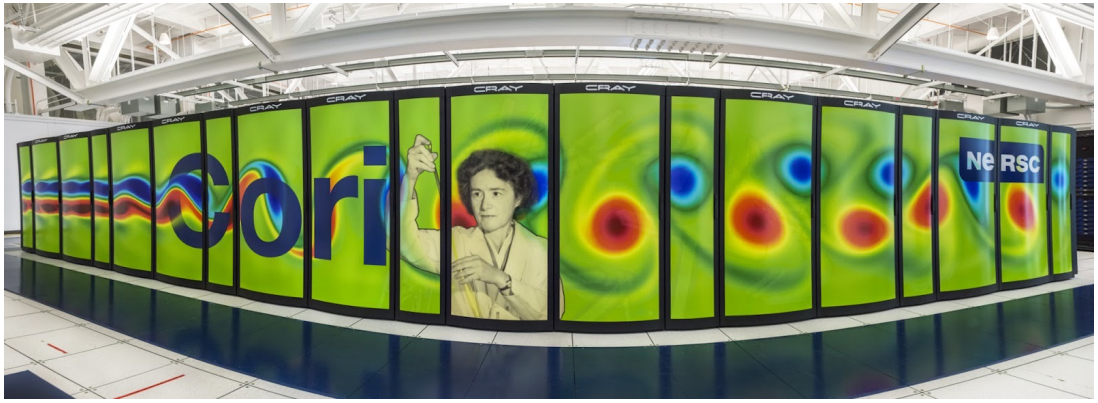
NCAR UCAR CISL Computational & Information Systems Lab
NCAR is sponsored by National Science Foundation

https://www2.cisl.ucar.edu/resources/cmip-analysis-platform

Met Office
Hadley Centre

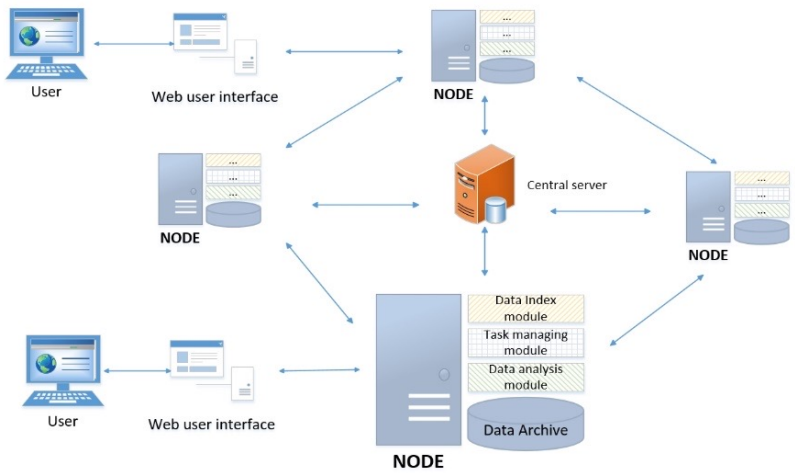
DOE: NERSC Cori

- Cray XC40 HPC with large number of conventional and Xeon Phi nodes
- Large HPSS archives and various types of local storage, including flash
- Open to DoE Office of Science projects, significant use by climate science



Typical climate model analysis facilities in Asia (1)

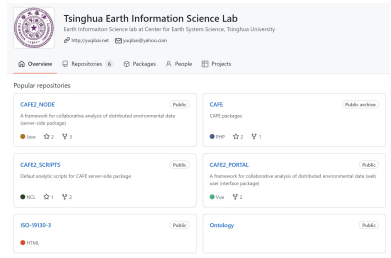
China: A Collaborative Analysis Framework for distributed gridded Environmental data (CAFÉ)



Xu, Hao, Sha Li, Yuqi Bai, Wenhao Dong, Wenyu Huang, Shiming Xu, Yanluan Lin et al. "A collaborative analysis framework for distributed gridded environmental data." Environmental Modelling & Software 111 (2019): 324-339.

- **A new "ZERO Download" mode**
  - Multiple data nodes establish a federation.
  - Users could submit data analysis tasks and then get the analysis result.
  - Data analysis are performed where the data reside.
  - Users do not have to download the model data.

- **A GitHub project**

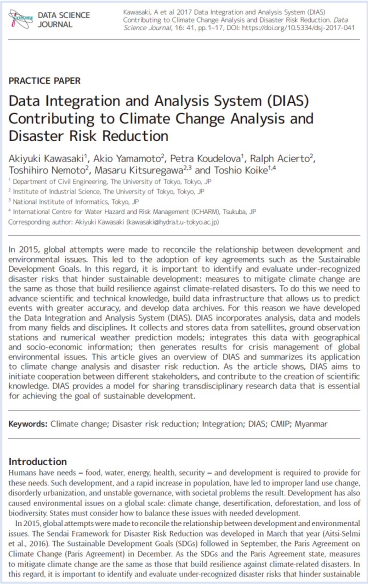


- **A Patented Technology**

Courtesy Yuqi Bai

Typical climate model analysis facilities in Asia (2)

Japan: Data Integration and Analysis System (DIAS)



DATA SCIENCE JOURNAL

Kawasaki, A et al 2017 Data Integration and Analysis System (DIAS) Contributing to Climate Change Analysis and Disaster Risk Reduction. *Data Science Journal*, 16: 41, pp. 1–17, DOI: https://doi.org/10.5334/dsj-2017-041

PRACTICE PAPER

Data Integration and Analysis System (DIAS) Contributing to Climate Change Analysis and Disaster Risk Reduction

Akiyuki Kawasaki[1], Akio Yamamoto[2], Petra Koudelova[1], Ralph Acierto[2], Toshihiro Nemoto[2], Masaru Kitsuregawa[2,3] and Toshio Koike[1,4]

[1] Department of Civil Engineering, The University of Tokyo, Tokyo, JP
[2] Institute of Industrial Science, The University of Tokyo, Tokyo, JP
[3] National Institute of Informatics, Tokyo, JP
[4] International Centre for Water Hazard and Risk Management (ICHARM), Tsukuba, JP
Corresponding author: Akiyuki Kawasaki (kawasak@hydra.t.u-tokyo.ac.jp)

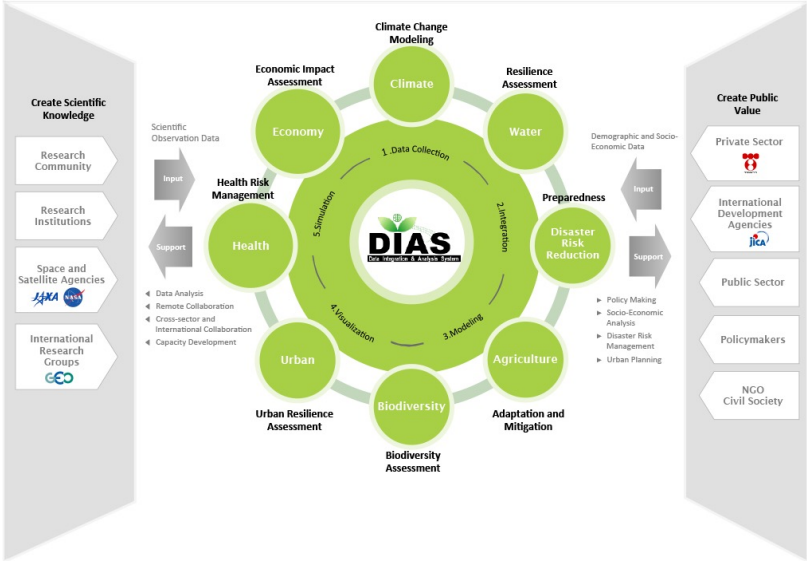
In 2015, global attempts were made to reconcile the relationship between development and environmental issues. This led to the adoption of key agreements such as the Sustainable Development Goals. In this regard, it is important to identify and evaluate under-recognized disaster risks that hinder sustainable development: measures to mitigate climate change are the same as those that build resilience against climate-related disasters. To do this we need to advance scientific and technical knowledge, build data infrastructure that allows us to predict events with greater accuracy, and develop data archives. For this reason we have developed the Data Integration and Analysis System (DIAS). DIAS incorporates analysis, data and models from many fields and disciplines. It collects and stores data from satellites, ground observation stations and numerical weather prediction models; integrates this data with geographical and socio-economic information; then generates results for crisis management of global environmental issues. This article gives an overview of DIAS and summarizes its application to climate change analysis and disaster risk reduction. As the article shows, DIAS aims to initiate cooperation between different stakeholders, and contribute to the creation of scientific knowledge. DIAS provides a model for sharing transdisciplinary research data that is essential for achieving the goal of sustainable development.

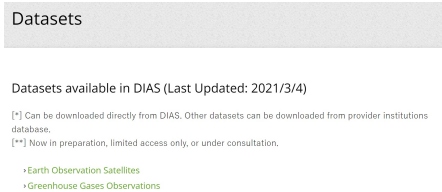
Keywords: Climate change; Disaster risk reduction; Integration; DIAS; CMIP; Myanmar

Introduction

Humans have needs – food, water, energy, health, security – and development is required to provide for these needs. Such development, and a rapid increase in population, have led to improper land use change, disorderly urbanization, and unstable governance, with societal problems the result. Development has also caused environmental issues on a global scale: climate change, desertification, deforestation, and loss of biodiversity. States must consider how to balance these issues with needed development.

In 2015, global attempts were made to reconcile the relationship between development and environmental issues. The Sendai Framework for Disaster Risk Reduction was developed in March that year (Aitsi-Selmi et al., 2016). The Sustainable Development Goals (SDGs) followed in September, the Paris Agreement on Climate Change (Paris Agreement) in December. As the SDGs and the Paris Agreement state, measures to mitigate climate change are the same as those that build resilience against climate-related disasters. In this regard, it is important to identify and evaluate under-recognized disaster risks that hinder sustainable





Datasets

Datasets available in DIAS (Last Updated: 2021/3/4)

[*] Can be downloaded directly from DIAS. Other datasets can be downloaded from provider institutions database.
[**] Now in preparation, limited access only, or under consultation.

- Earth Observation Satellites
- Greenhouse Gases Observations
- Terrestrial Ecosystems / Carbon Flux Observations
- Weather Observations
- Watershed Observations
- Ocean Observations
- Reanalysis
- Prediction
- Downscaled Data
- Natural Disasters
- Land Use
- Health Hazard

Kawasaki, Akiyuki, Akio Yamamoto, Petra Koudelova, Ralph Acierto, Toshihiro Nemoto, Masaru Kitsuregawa, and Toshio Koike. "Data integration and analysis system (DIAS) contributing to climate change analysis and disaster risk reduction." Data Science Journal 16 (2017).

Courtesy Yuqi Bai

Typical climate model analysis facilities in Asia (3)

Australia: NCI High Performance Computing and High Performance Data Platform

The NCI High Performance Computing and High Performance Data Platform to Support the Analysis of Petascale Environmental Data Collections

Ben Evans[1], Lesley Wyborn[1], Tim Pugh[2], Chris Allen[1], Joseph Antony[1], Kashif Gohar[1], David Porter[1], Jon Smillie[1], Claire Trenham[1], Jingbo Wang[1], Alex Ip[3], and Gavin Bell[4]

[1] National Computational Infrastructure (NCI), Australian National University, Canberra, Australia
[2] Bureau of Meteorology, Melbourne, Australia
[3] Geoscience Australia, Canberra, Australia
[4] The 6th Column Project, Berlin, Germany
{Ben.Evans,Lesley.Wyborn,Chris.Allen,Joseph.Antony, Kashif Gohar,David.Porter,Jon.Smillie, Claire.Trenham,Jingbo.Wang}@anu.edu.au, T.Pugh@bom.gov.au, Alex.Ip@ga.gov.au, gavin@6thcolumn.org

Abstract. The National Computational Infrastructure (NCI) at the Australian National University (ANU) has co-located a priority set of over 10 PetaBytes (PBytes) of national data collections within a HPC research facility. The facility provides an integrated high-performance computational and storage platform, or a High Performance Data (HPD) platform, to serve and analyse the massive amounts of data across the spectrum of environmental collections – in particular from the climate, environmental and geoscientific domains. The data is managed in concert with the government agencies, major academic research communities and collaborating overseas organisations. By co-locating the vast data collections with high performance computing environments and harmonising these large valuable data assets, new opportunities have arisen for Data-Intensive interdisciplinary science at scales and resolutions not hitherto possible.

Keywords: high performance computing, high performance data, cloud Computing, data-intensive science, scalable data services, data cube, virtual laboratories.

1 Introduction

The National Computational Infrastructure (NCI) at the Australian National University (ANU) has organised a priority set of large volume national environmental data assets on a High Performance Data (HPD) Node within a High Performance

R. Denzer et al. (Eds.): ISESS 2015, IFIP AICT 448, pp. 569-577, 2015.

Australia's preeminent high-performance data, storage and computing facility.

The National Computational Infrastructure (NCI) is Australia's leading high-performance data, storage and computing organisation, providing expert services to benefit all domains of science, government and industry.

NCI brings the Australian Government and the Australian research sector together through a broad collaboration involving the largest national science agencies, universities, industry and the Australian Research Council.

NCI empowers government agencies, universities, and industry across multiple domains of research. Our integrated hardware, services and expertise drive high-impact research and groundbreaking outcomes for Australia.

Evans, Ben, Lesley Wyborn, Tim Pugh, Chris Allen, Joseph Antony, Kashif Gohar, David Porter et al.
"The NCI high performance computing and high performance data platform to support the analysis of petascale environmental data collections."
In International Symposium on Environmental Software Systems, pp. 569-577. Springer, Cham, 2015.

Courtesy Yuqi Bai

European Service Offering: climate model data analysis (e.g. CMIP6)

Tiered of downloading Terabytes out of Petabytes of climate model data ?

Large European climate data centers offer the possibility to directly exploit locally available large climate data pools (e.g. CMIP6 data)

The offering from DKRZ, IPSL-CNRS, UKRI-CEDA, CMCC:

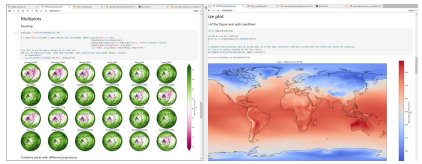
- **Access to large European climate model data pools**
  (multi-PByte data collections including CMIP6, CORDEX, ..)
- **Access to associated HPC compute ressources**
- **Access to interactive analysis environments**
  (including jupyter-hub installations at DKRZ, CMCC and STFC)
  - support for e.g. pangeo sw stack (xarray, dask), cdo, ESMValTool and user tailored environments..

Two types of service:

- **"Jump start service":**
  + minimal application procedure
  - limited compute resources
- **Analysis platform service:**
  - short project proposal required
  + guaranteed resource allocation

Interested? Further information:

- **Climate Analytics service (ECAS):** https://portal.enes.org/data/data-metadata-service/climate-analytics-service
- **Analysis platforms application:** https://portal.enes.org/data/data-metadata-service/analysis-platforms
- **Demos, use-cases, example jupyter notebooks:** https://github.com/IS-ENES-Data/Climate-data-analysis-service



Next deadline: 20.12.2021

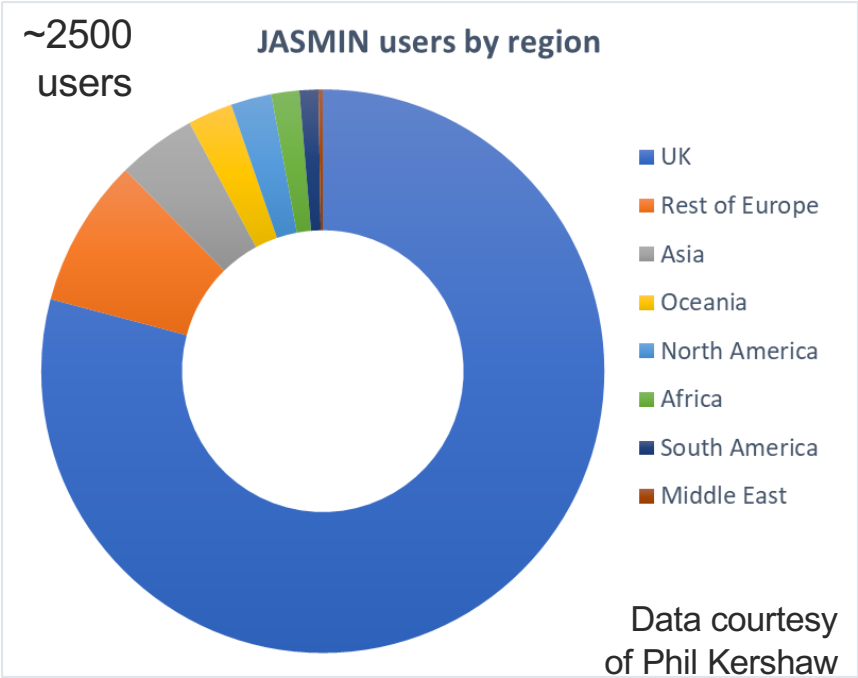
The IS-ENES3 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Stephan Kindermann

Met Office
Hadley Centre

Access to analysis facilities

- Most facilities are restricted access
  - Local users
  - Funded collaborations
- Gap in provision for Africa and South America

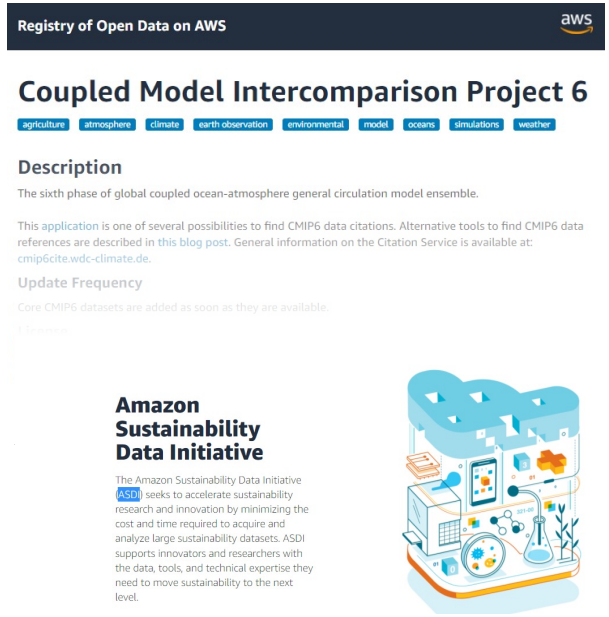


Met Office Hadley Centre

Pangeo and commercial cloud

PANGEO

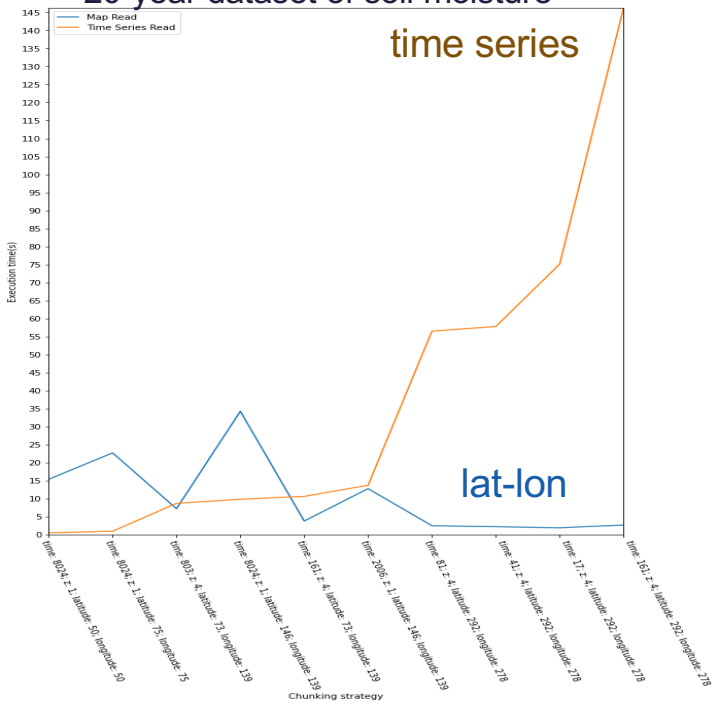
- AWS providing hosting for a subset of CMIP6 data in Zarr format along with compute through ASDI
  - Some free access for users
  - Many other datasets available
- Pangeo providing and managing data
  - Collaborating with GFDL and CEDA
- Change in data storage, and compression, has implications for analysis and archiving options



Object Store: Different storage strategies showed radically different performance

Adapted from slide by Phil Kershaw

20-year dataset of soil moisture



- Experiment with different storage chunking arrangements
- Way in which data is written and stored has significant impact on performance when used
- Rewriting with alternate chunking is fast, needs planning

Science and Technology Facilities Council
Natural Environment Research Council

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Met Office
Hadley Centre

Summary

- Host of national and regional facilities available for local science communities
  - Some outreach for specific projects
  - Existing collaborations
- Need to advertise these facilities and encourage their use
  - Common tool sets appearing based around core set used by Pangeo
- Commercial cloud spreading general access to other regions
  - Cost and funding
  - Access is not universal