Data: the elephant in the room

JASMIN:
One step on the road to dealing with the elephant

Bryan Lawrence (@bnlawrence)

National Centre for Atmospheric Science
University of Reading
Science and Technology Facilities Council
(Centre for Environmental Data Archival)

(With thanks to
Jonathan Churchill,
Matt Pritchard
& Ag Stephens
as well as folks named
as we go along!)

Outline

Intro to CEDA and motivation.

Data Deluge:

  - Examples: UPSCALE and CMIP5

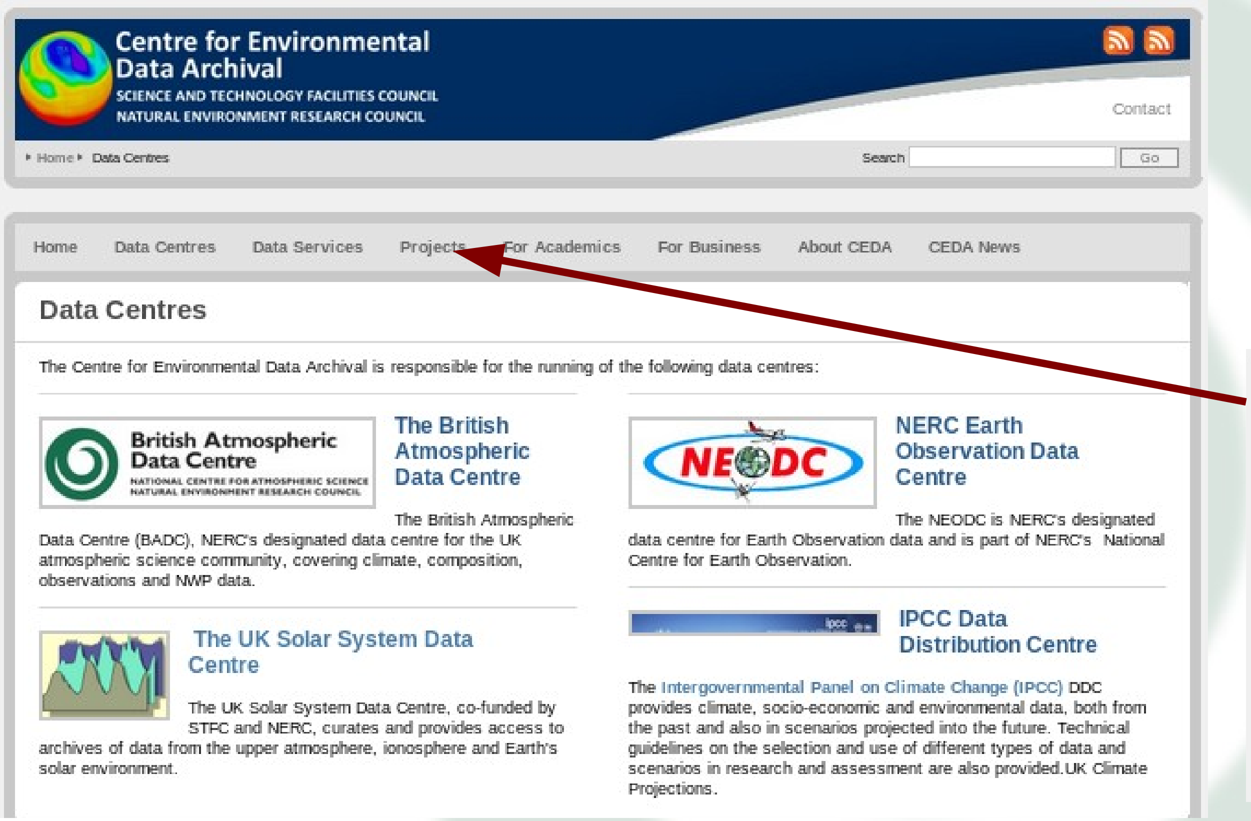
Introduction to JASMIN (&CEMS)

Some JASMIN usage

  - Archive management
  - UPSCALE
  - EO Reprocssing (with CEMS)

Just one example of why we (sometimes) need to write data out and not do post-processing in the machine!

BADC and CEDA? www.ceda.ac.uk


Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL
Contact
Home ▸ Data Centres
Search
Go
Home
Data Centres
Data Services
Projects
For Academics
For Business
About CEDA
CEDA News
British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL
NEODC

Data Centres

The Centre for Environmental Data Archival is responsible for the running of the following data centres:

The British Atmospheric Data Centre

The British Atmospheric Data Centre (BADC), NERC's designated data centre for the UK atmospheric science community, covering climate, composition, observations and NWP data.

The UK Solar System Data Centre

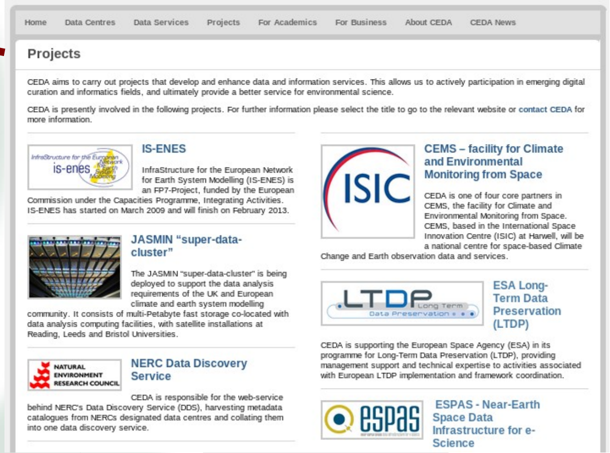
The UK Solar System Data Centre, co-funded by STFC and NERC, curates and provides access to archives of data from the upper atmosphere, ionosphere and Earth's solar environment.

NERC Earth Observation Data Centre

The NEODC is NERC's designated data centre for Earth Observation data and is part of NERC's National Centre for Earth Observation.

IPCC Data Distribution Centre

The Intergovernmental Panel on Climate Change (IPCC) DDC provides climate, socio-economic and environmental data, both from the past and also in scenarios projected into the future. Technical guidelines on the selection and use of different types of data and scenarios in research and assessment are also provided.UK Climate Projections.


Home
Data Centres
Data Services
Projects
For Academics
For Business
About CEDA
CEDA News
InfraStructure for the European Network for Earth System Modelling
is-enes
ISIC
LTDP Long Term Data Preservation
NATURAL ENVIRONMENT RESEARCH COUNCIL
espas

Projects

CEDA aims to carry out projects that develop and enhance data and information services. This allows us to actively participation in emerging digital curation and informatics fields, and ultimately provide a better service for environmental science.

CEDA is presently involved in the following projects. For further information please select the title to go to the relevant website or contact CEDA for more information.

IS-ENES

InfraStructure for the European Network for Earth System Modelling (IS-ENES) is an FP7-Project, funded by the European Commission under the Capacities Programme, Integrating Activities. IS-ENES has started on March 2009 and will finish on February 2013.

JASMIN "super-data-cluster"

The JASMIN "super-data-cluster" is being deployed to support the data analysis requirements of the UK and European climate and earth system modelling community. It consists of multi-Petabyte fast storage co-located with data analysis computing facilities, with satellite installations at Reading, Leeds and Bristol Universities.

NERC Data Discovery Service

CEDA is responsible for the web-service behind NERC's Data Discovery Service (DDS), harvesting metadata catalogues from NERCs designated data centres and collating them into one data discovery service.

CEMS – facility for Climate and Environmental Monitoring from Space

CEDA is one of four core partners in CEMS, the facility for Climate and Environmental Monitoring from Space. CEMS, based in the International Space Innovation Centre (ISIC) at Harwell, will be a national centre for space-based Climate Change and Earth observation data and services.

ESA Long-Term Data Preservation (LTDP)

CEDA is supporting the European Space Agency (ESA) in its programme for Long-Term Data Preservation (LTDP), providing management support and technical expertise to activities associated with European LTDP implementation and framework coordination.

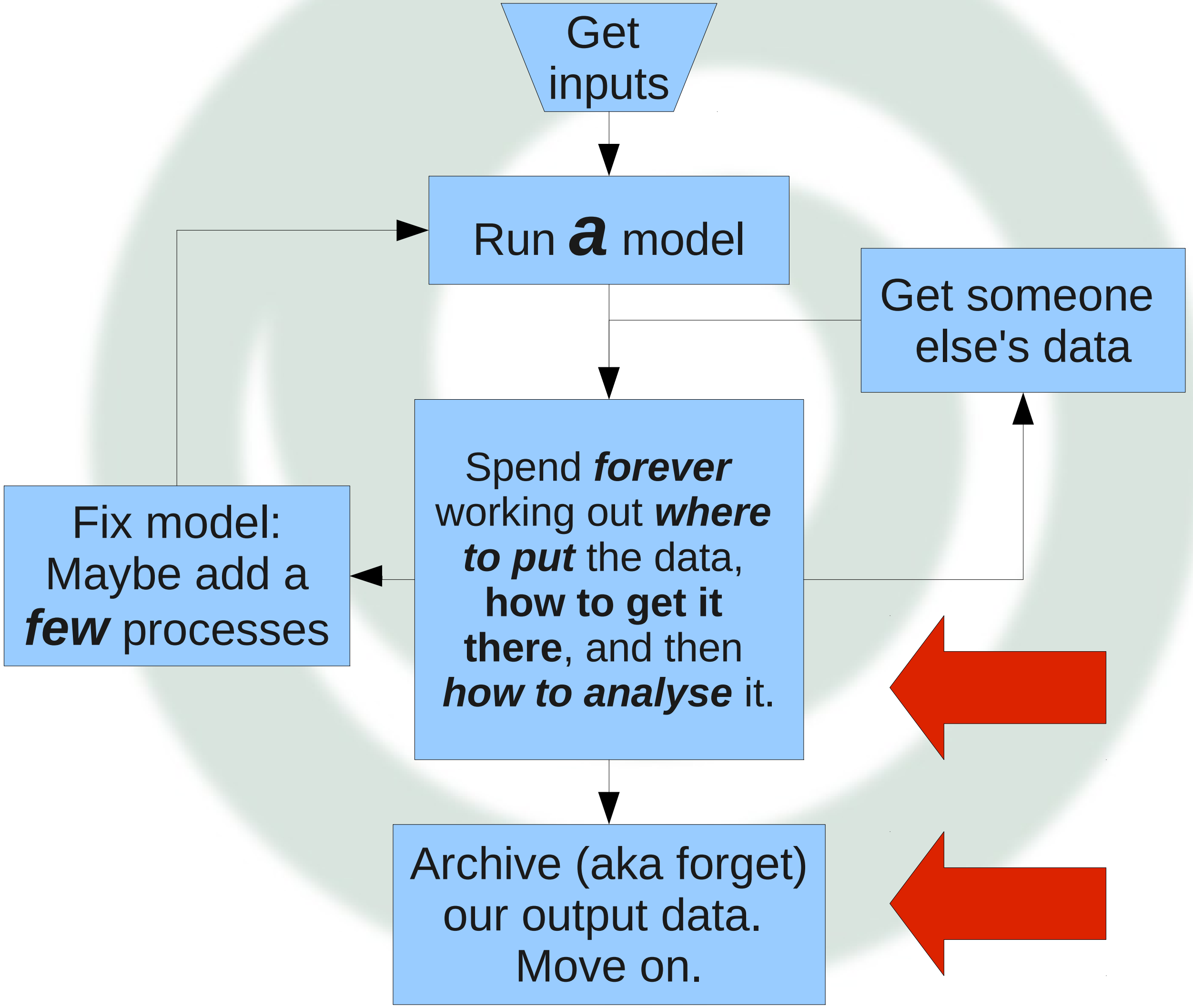
ESPAS - Near-Earth Space Data Infrastructure for e-Science

Lots more

Mission: Curation AND **facilitation!**

Motivation

This is how our community thinks about models and data:



Get inputs

Run ***a*** model

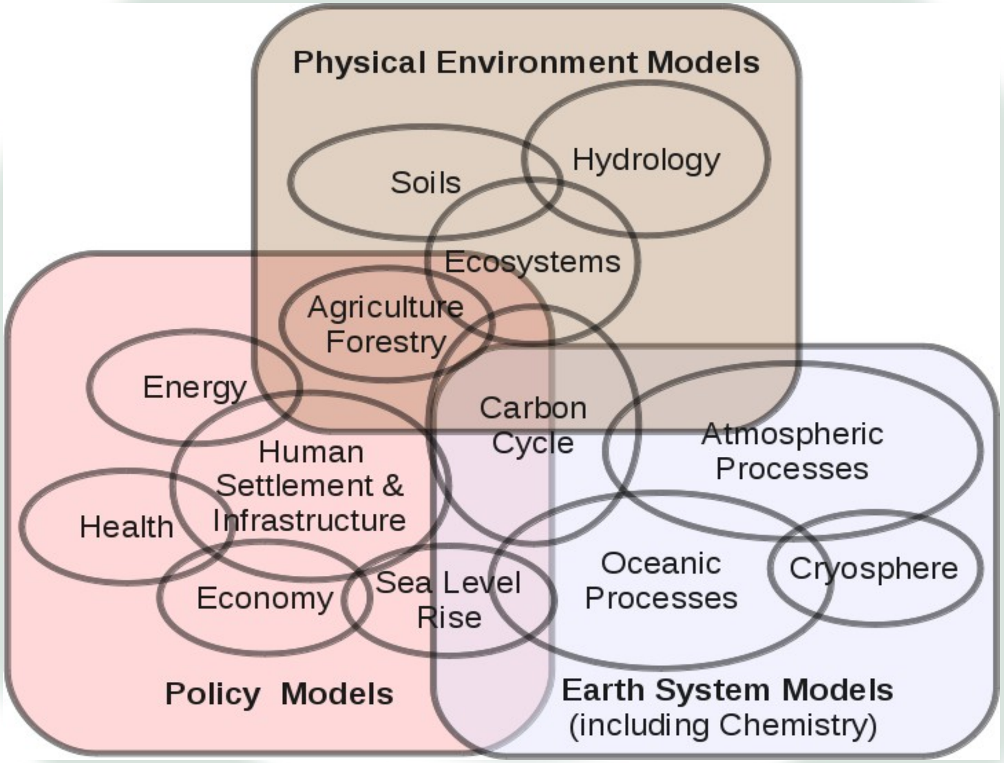
Get someone else's data

Spend ***forever*** working out ***where to put*** the data, **how to get it there**, and then ***how to analyse*** it.

Fix model: Maybe add a ***few*** processes

Archive (aka forget) our output data. Move on.

But it's not just about us! Many, many processes, so many, many communities *might* use some of that data!



We can't add all these processes into our models, we have to interact via data! Lots of it. At high resolution. Whether we like it or not.

(Figure adapted from Moss et al., 2010).

A quick (broken) calculation

Consider a grand ensemble: Let's say it's EC-earth running at 10 years/day, at 25 km grid resolution on 5000 cores.

Not too far from reality?

It would be entirely reasonable (scientifically) to consider a 50 member initial condition ensemble. Let's add in a few variants to try and capture structural (model) uncertainty. Let's say 4.

That's a million core experiment. Feasible with our models today!

But not feasible with our existing HPC, but let's pretend we had a million core machine, which EC-earth could use with “similar” core performance.

Now run that “grand ensemble” for 25 years: 2.5 days in the machine. Only 60 million core hours!

Output?

A 1.25 degree (actually T159L62) model produces roughly 9 GB of data per simulation month in a real application (Colin Jones). Let's say 10 GB to make life easy. (Conservative!)

This simulation could produce (10x5x5: 250 GB model month). The grand ensemble output has: 25x50x4x12=60K months.

So, that's 60K x 250 GB ~15 PB = 6 PB/day (=0.5 Tbit/s!)

Data Lifetime is much much longer than the time the model spends in the HPC!

(conclusion: you can't share disk in an archive in the same way as you share an HPC – although obviously you can share support for *your* disk!).

I'll be coming back to that.

Consequences?

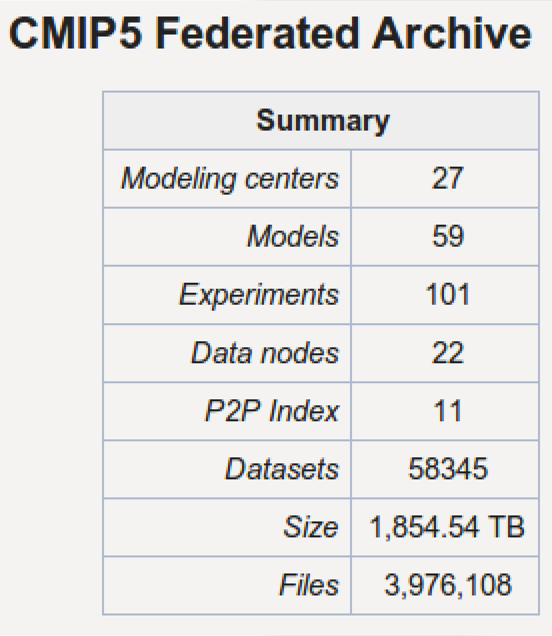
There's a lot wrong with that calculation,

But the bottom line is that a relatively modest improvement in our software, with a pretty significant improvement in hardware, and hardware availability, along with some realistic efforts to attack uncertainty, will create a data nightmare!

Sure, we can make choices about what to write out - in this future we can start to think about the FLOPS as free, and the BYTEs as the significant cost! And analysis as the place to invest, not so much performance!

So, that's the future … what about now?

Welcome to CMIP5



CMIP5 Federated Archive

Summary	
Modeling centers	27
Models	59
Experiments	101
Data nodes	22
P2P Index	11
Datasets	58345
Size	1,854.54 TB
Files	3,976,108

Compare with CMIP3 a little over six year ago:
35 TB held in one location!

BADC has just over 700 TB of that (90 TB of UKMO data + 620 TB replicated!)

(Including data currently being staged, we have around 900 TB and growing.)

BADC intends to keep this data INDEFINITELY!

Welcome to UPSCALE

(the things Pier-Luigi mentioned on the way …)

Never mind how fast the UM is … UPSCALE has produced roughly 500 TB of data in a year, on a single machine in Germany with minimal archive.

- cf CMIP5: currently 1.8 PB from o(30) centres over a couple of years ...

We can store it (in the UK), and we can move the data, just! Sustained 1.5 TB/day over a year (peaking over 10 TB/day).

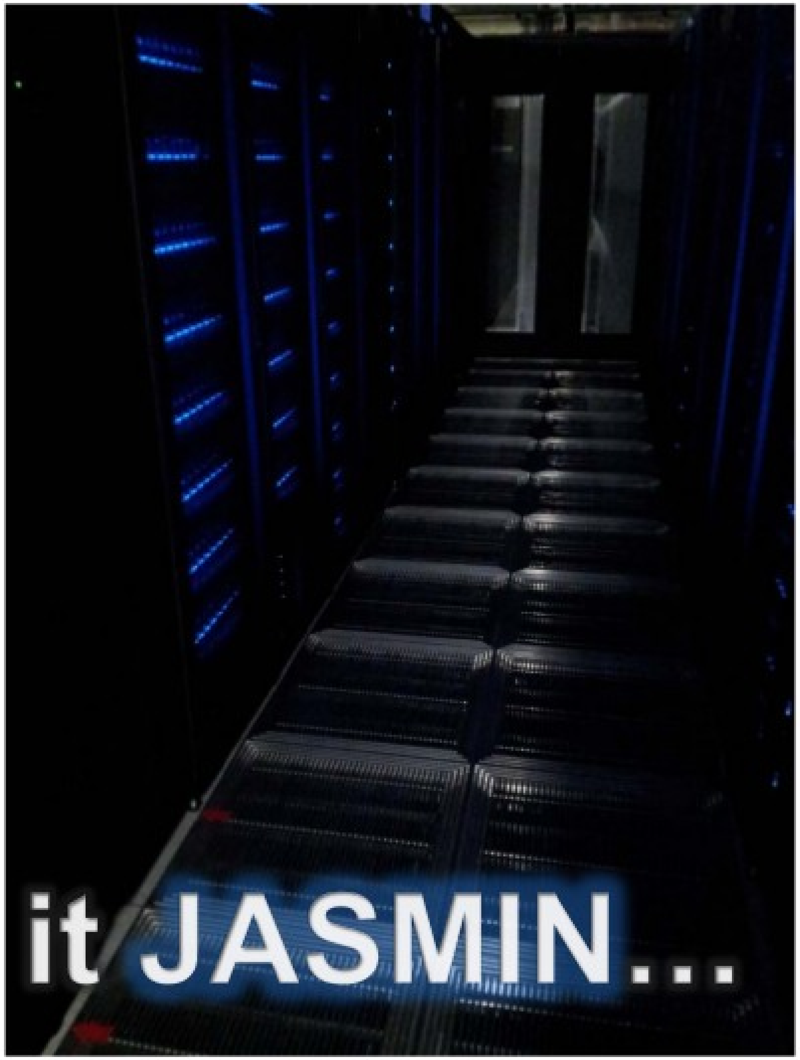
- cf with our thought experiment: 6 PB/day!

How do we analyse the data?

Hardware and Software. Well, we've done something about the former …

A new computer! A “Super Data Cluster”

http://arxiv.org/abs/1204.3553



We’ve called it JASMIN...

Life before JASMIN

CEDA went petascale a while ago, and our computer environment at the beginning of 2012 consisted of:

- of the order of 200 milliion files
- using o(1.5 PB) of NAS disk, on
- o(150) disk partitions, and split into o(600) datasets
- o(300) different computers (virtual machines), on
- o(30) hypervisors
- (And I cant find the number of tapes in the Atlas data store and in remote storage).

Things were grim!

CEDA, and particularly BADC, was grinding to a halt. We had inexplicable network problems. We spent lots of time moving data as machine lifetimes expired, and user services suffered.

(This was not a designed environment, it was organic, it grew over a decade)

(It has taken us a year to migrate our data to JASMIN, it'll probably take nearly as long to retire all the services on legacy hardware).

But why? Run anywhere, analyse in one place!

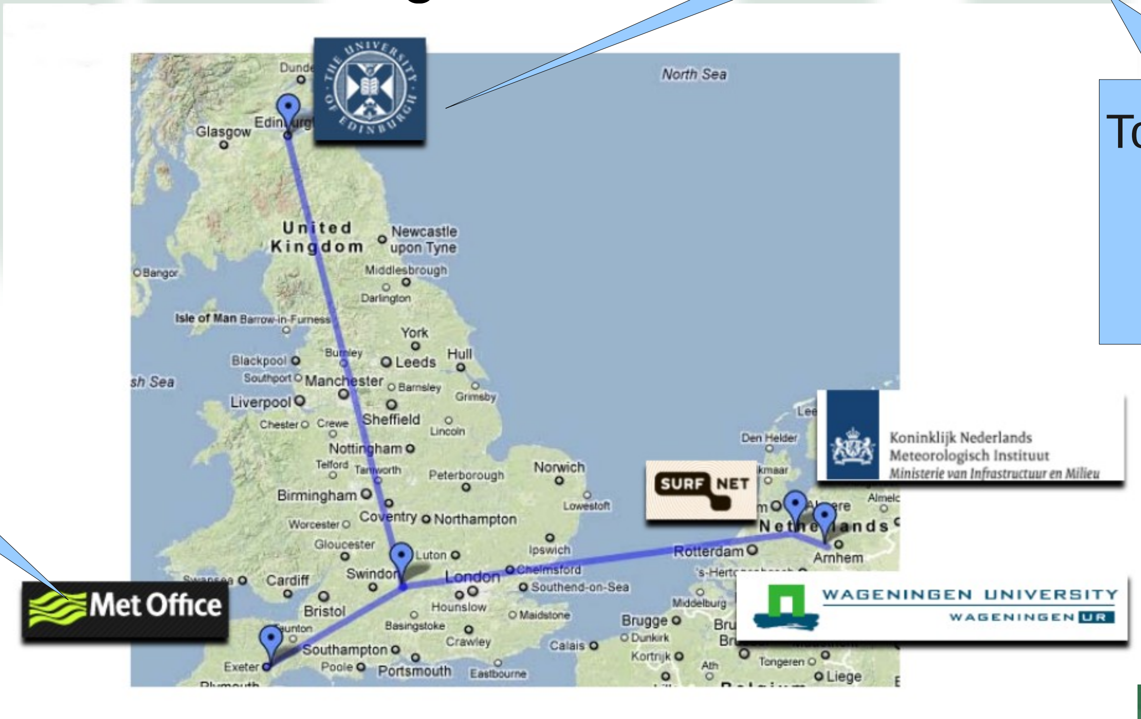
We modellers are a fickle lot.

We'll use anyone's HPC

- If we can (currently a big IF for European models)

But we want to combine our data!

- NXN data transfers are out!
- N will be bad enough!



HECToR

Cray XE6
90K cores

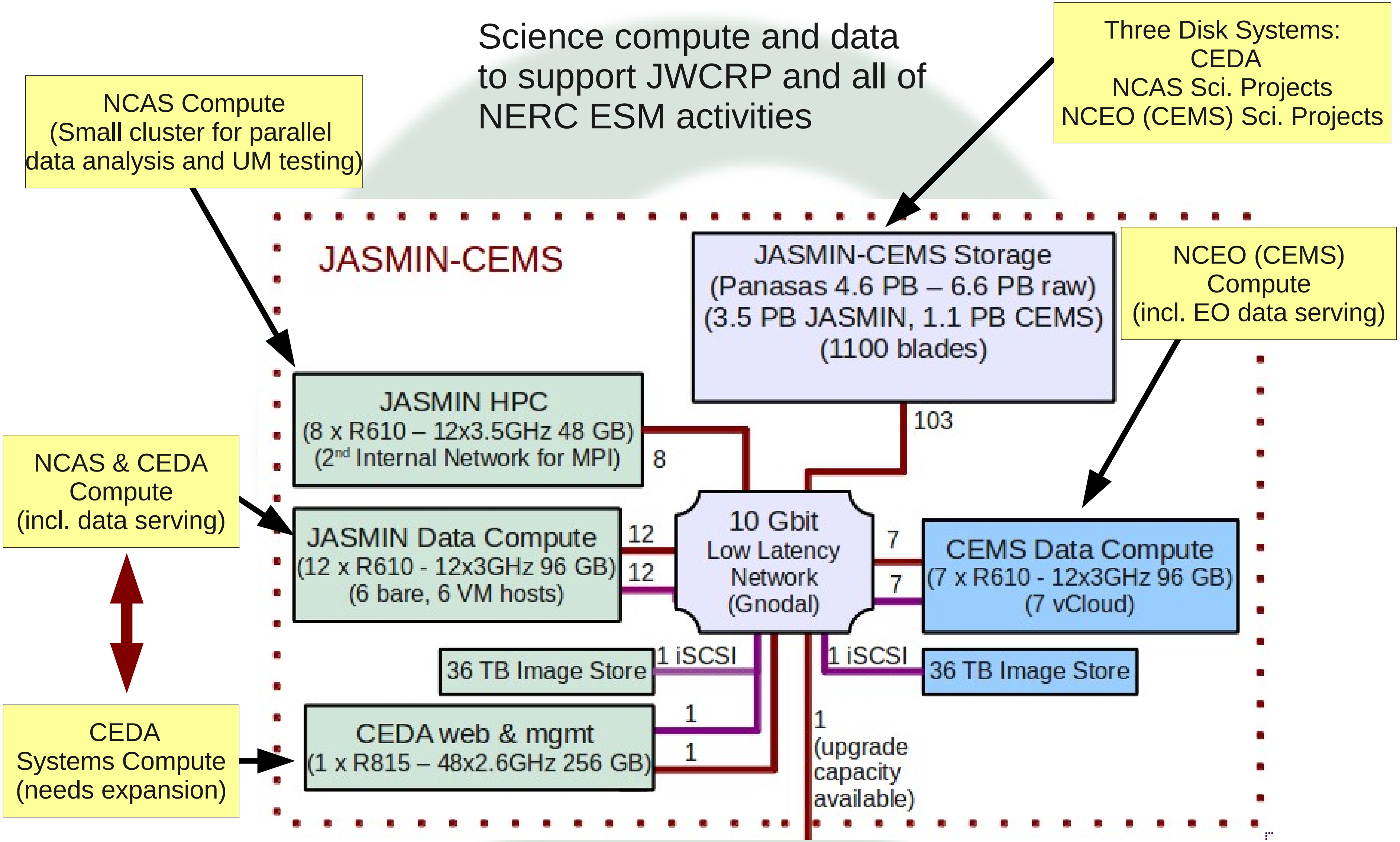
To be replaced Dec 2013 With ARCHER

MONSooN

5000 core
IBM P7

JASMIN

Science compute and data to support JWCRP and all of NERC ESM activities



NCAS Compute
(Small cluster for parallel data analysis and UM testing)

Three Disk Systems:
CEDA
NCAS Sci. Projects
NCEO (CEMS) Sci. Projects

NCAS & CEDA Compute
(incl. data serving)

NCEO (CEMS) Compute
(incl. EO data serving)

CEDA
Systems Compute
(needs expansion)

JASMIN-CEMS

JASMIN-CEMS Storage
(Panasas 4.6 PB – 6.6 PB raw)
(3.5 PB JASMIN, 1.1 PB CEMS)
(1100 blades)

JASMIN HPC
(8 x R610 – 12x3.5GHz 48 GB)
(2nd Internal Network for MPI)

JASMIN Data Compute
(12 x R610 - 12x3GHz 96 GB)
(6 bare, 6 VM hosts)

10 Gbit
Low Latency
Network
(Gnodal)

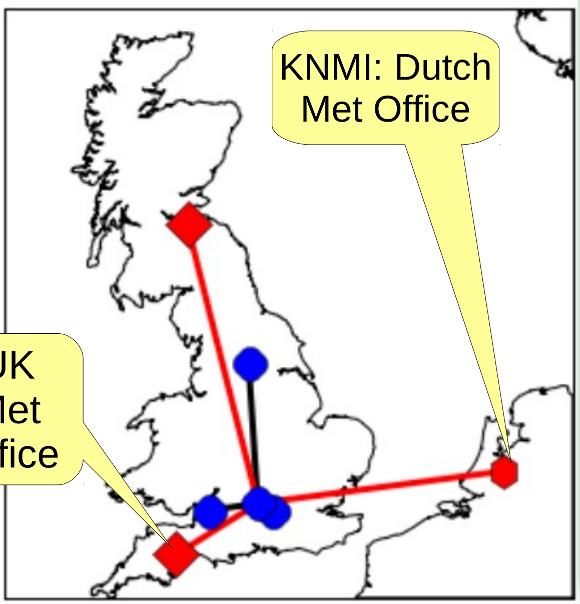
CEMS Data Compute
(7 x R610 - 12x3GHz 96 GB)
(7 vCloud)

36 TB Image Store

36 TB Image Store

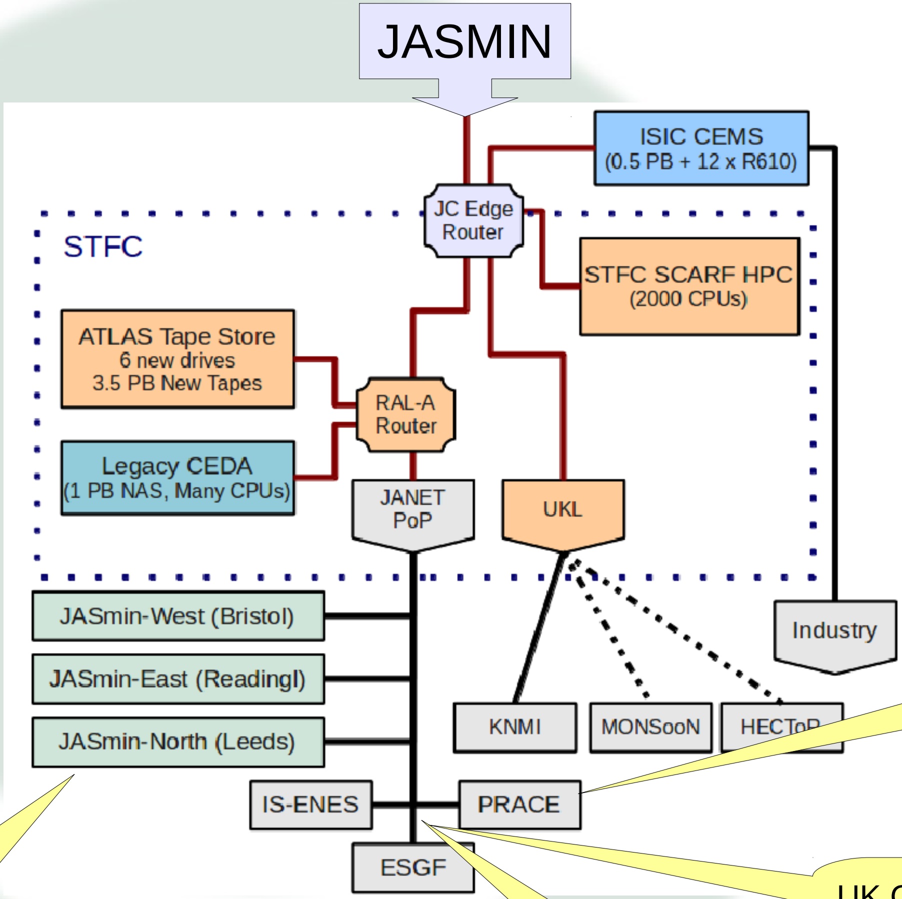
CEDA web & mgmt
(1 x R815 – 48x2.6GHz 256 GB)

Networks & “Greater JASMIN”.



KNMI: Dutch Met Office

UK Met Office



Major JASMIN components Also at LEEDS READING BRISTOL

e.g. UPSCALE

UK CMIP5 and AR5 dependent on this link

Key link to DKRZ (German Climate Computing)

4.6 PB of disk:
Why doesn't a data centre use tape?

We do! Enuff said. But that's not all we do!

At petascale, if the data of interest is on tape, we rapidly become a WORN archive: Write Once, Read Never!

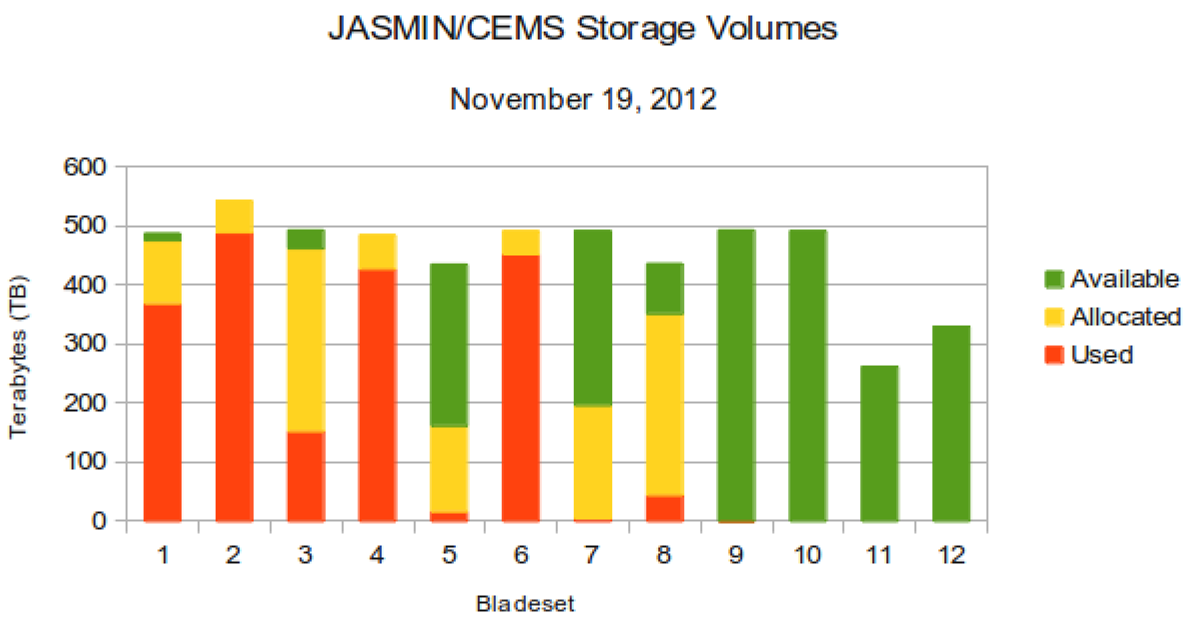
We can't put in place the same sort of procedures and policies that an operational centre has – academic scientists just don't behave like that.

We can (and will) give scientists access to tape, but for big data sets to be usable, they need to be online!

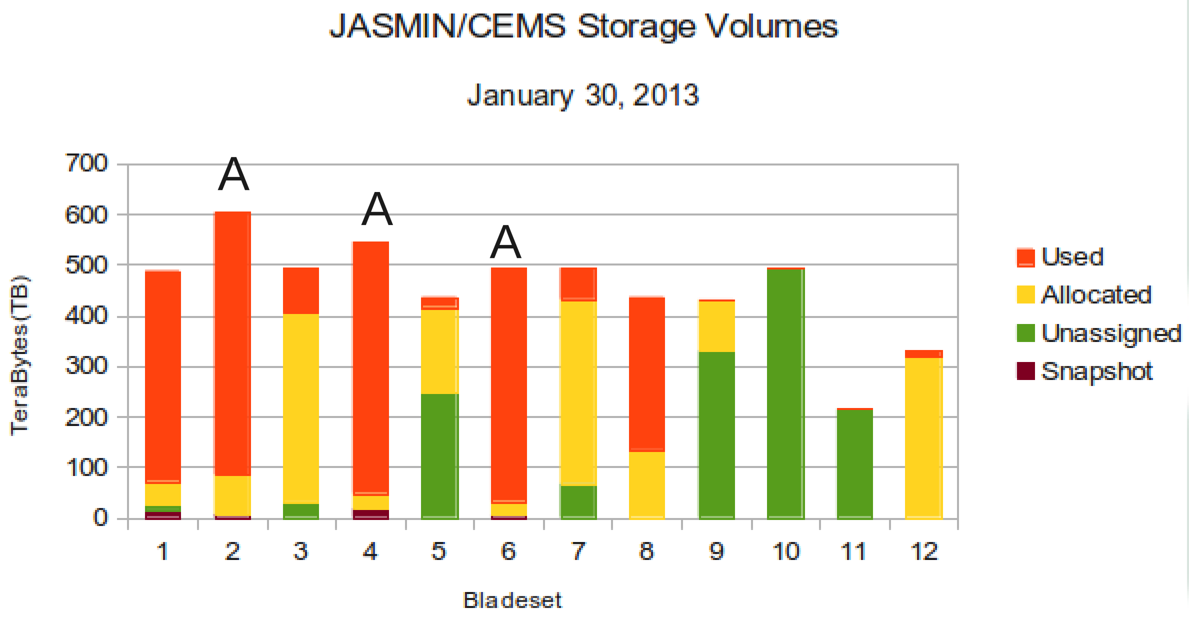
If not, where will they put it when it comes off tape? How will we managed scratch access at scale when we don't know who is going to what when? How will they do be big multi-model ensemble analyses at high resolution?

November 2012
35% Used
22% Allocated
44% Free

57%
“used”


JASMIN/CEMS Storage Volumes

November 19, 2012


JASMIN/CEMS Storage Volumes

January 30, 2013

January 2013
42% Used
29% Allocated
28% Free

71%
“used”

14% in three months!
Expect to “fill” in 2013
(UPSCALE and archive migration may have stopped, but we have WISER to come + EO ...)

On bladesets

We have compromised on our performance between I/O, space, reliability, and rebuild times.

- e.g. 1 PB given up to vertical parity.
- Small files RAID1, big files RAID5.
- Files striped within bladesets.

Most bladesets ~500 TB ~100 blades.

“Archive” on dedicated bladesets.

“Big Bruisers” like UPCALE on their own bladesets. Have no impact on other users!

So, although we have 1 Tbit/s bandwidth, most applications will only see max of 100 Gbit/s. But they wont share that (at least from the I/O perspective.)

Three examples of JASMIN/CEMS usage:

(1) Archive support

(2) Modelling

(3) Earth Observation

(1) Supporting the archive: ESGF on JASMIN

(Stephen Pascoe, CEDA)

Managing our server environment.

JASMIN has allowed us to scale ESGF nodes to the size required to run the Thredds Data Server on CMIP5 whilst keeping the flexibility of virtual machines.

ESGF index node (& UI)

- Medium VM: 2CPU, 8GB RAM

ESGF data node for replicas

- Large VM: 4CPU, 32GB RAM

Extra data node:

- Same configuration as replica node. In preparation for CORDEX

Archive Management

Checksumming an important part of believing in archive integrity.

Quality Controlling data an important part of data ingestion for long term maintenance.

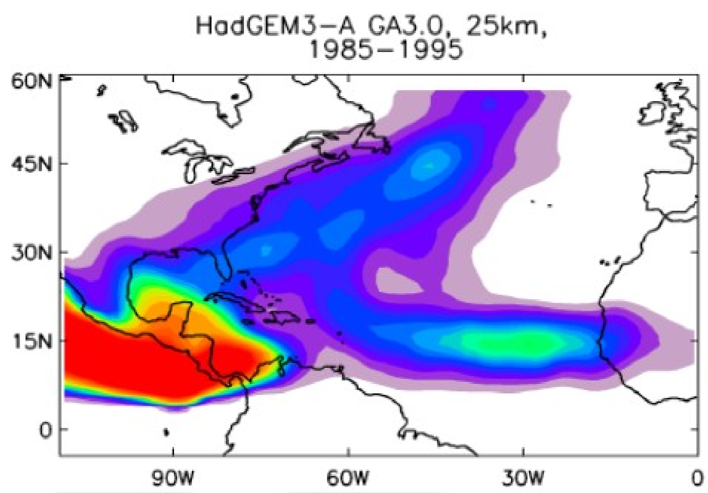
eg Checksumming 53.5TB of MPI-M CMIP5 replicas

- Split into 16 parallel jobs
- Parallel execution time: 3:56:34. Total process time: 20:28:14. 5x speed-up!

Running CMIP5 QC on Met Office and IPSL data. Running time from weeks to days!

Upscale Cyclone Tracking (Malcolm Roberts, UKMO)

The TRACK feature tracking algorithm (Hodges, 1994) takes global 6 hourly winds as input, processes these in a number of stages (about 20), to produce tropical cyclone tracks as output. These stages involve calculation of vorticity on several model levels, filtering to lower resolution grids, as well as the tracking. The code is a mixture of Fortran and C.



(a la Pier Luigi Vidale yesterday)

The input file size for each 7 month season of the 25km N512 model is a 42GB netcdf file. Due to the fairly intensive IO and processing, each of these seasons takes approximately 55 hours to process on the MONSooN post-processor (which is several times longer than it takes the model to run 7 months)!
On JASMIN, the same executable takes around 22 hours for the same period.
There's a lot to understand here, but this is probably just better I/O, we haven't begun to think about the parallelisation opportunities we have here!

ATSR Reprocessing (Steve Donegan, CEDA)

Along Track Scanning Radiometer

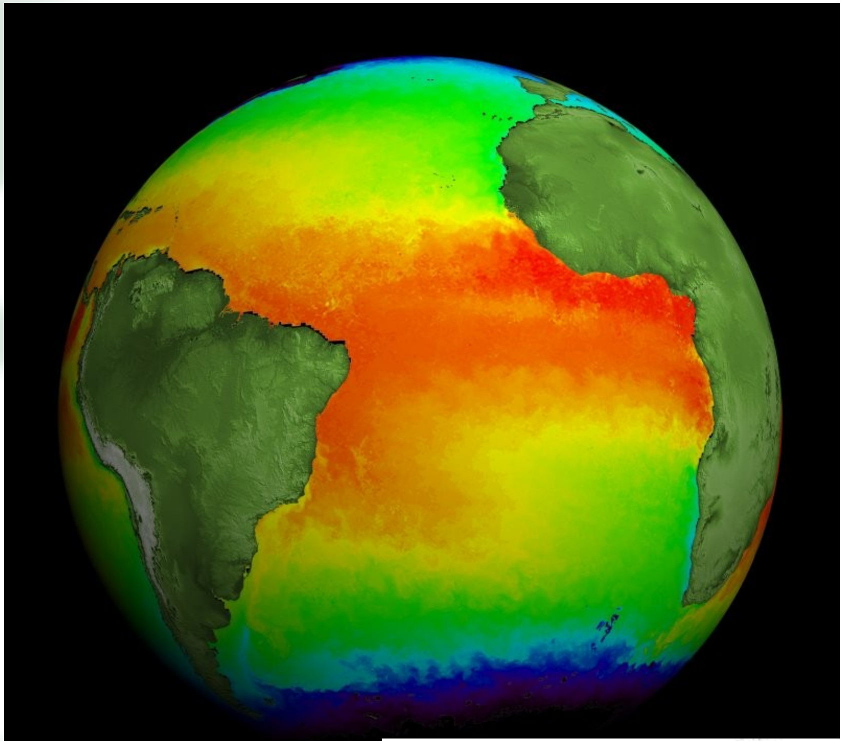
(measures sea surface temperature to 0.3K precision with pixel size ~ 1km^2 and decadal stability around 0.1K)

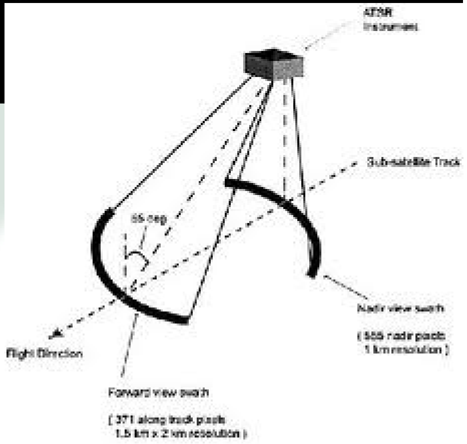
(Cunning use of on-board black bodies and etc to provide reliable calibration.)

Three Instruments:

- ATSR1: (1991-1997)
- ATSR2: (1995-2003)
- AATSR: (2002-present)

Periodic reprocessing of basic brightness temperatures (in instrumental radiance bands) to produce temperatures (and other products).





Reprocessing ATSR1 and ATSR2 at RAL using LOTUS (JASMIN-HPC)

Last reprocessing took place in 2007-2008.

Used 10 dedicated servers to process data and place product in archive

Previous reprocessing complicated by lack of sufficient contiguous storage for output and on archive.

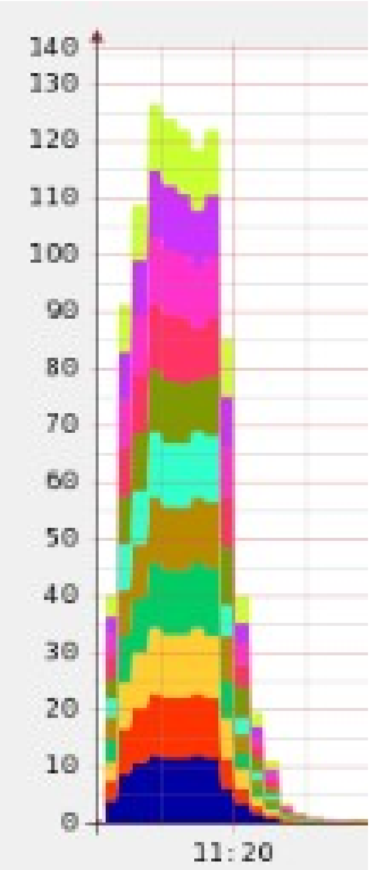
Reprocessing of 1 month of ATSR2 L1B data using original system took ~**3 days**: using JASMIN-HPC Lotus: **12 minutes**.

Entire 7 year archive of ATSR2 L1B should/could take 2-3 days (see caveats) whereas previously took 6 months

ATSR1 data ~5 year archive should take less time as fewer channels/smaller source data.

Level2 data should be even quicker (takes reprocessed L1B data as input).

Increased processing speed using new system makes it easier to deal with/rectify errors if anything spotted.



132 processors flat out (NOT I/O bound) for 12 minutes)!

Just one more example as to why we sometimes need to write out lots of data – and not always do the processing “in” the job!

Large Domain Cloud System Resolving Simulations of the Tropics

A cast of thousands;

Steve Woolnough, Doug Parker, Adrian Matthews, Mike Blackburn,
Robin Hogan, **Grenville Lister**, Chris Holloway, Barney Love, Nick Dixon, Thorwald Stein, Kevin Pearson, Guiying Yang
and
Many people in the Met Office in Atmospheric Process and Parametrization and the Joint Centre for Mesoscale Modelling, led by Paul Field

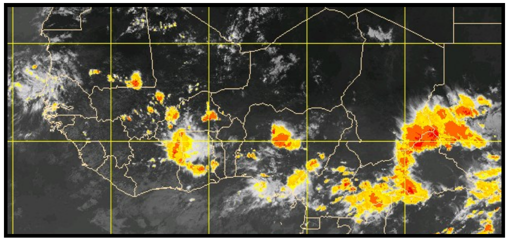
Cascade *Organised convection and scale interactions in the tropical atmosphere*

Case Studies

West Africa

African Easterly Waves

Diurnal Cycle

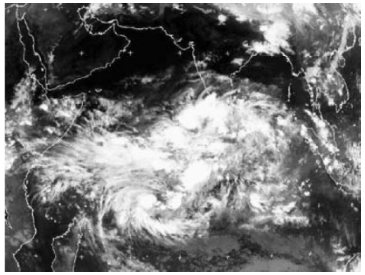


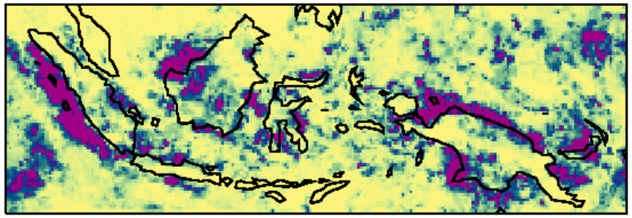
Warm Pool

MJO

Maritime Continent

Diurnal Cycle

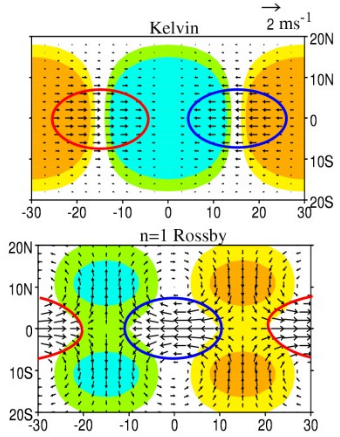




Idealised

Warm Pool Convection

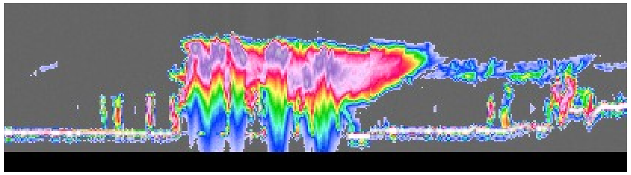
Equatorial Waves



Model Evaluation against Observations

CloudSat/CALIPSO: vertical cloud properties

GERB/SEVIRI/MTSAT: horizontal and time



Synthesis

Analysis of scale interactions

Insight into physical processes

Compare with climate / NWP resolution

Conclusions for parameterization

Cascade

Data requirements (1)

Aims again..

- Advance understanding of convective organisation and scale interactions
- Inform the development of new approaches to convective parameterization

Calculating things like spectra, terms in the budgets e.g.

$$\frac{\partial\bar{\theta}}{\partial t}+\nabla(\bar{\mathbf{v}}\bullet\bar{\theta})+\frac{\partial}{\partial p}(\bar{\omega}\bar{\theta})=Q_1+Q_R$$
$$\frac{\partial\bar{q}}{\partial t}+\nabla(\bar{\mathbf{v}}\bullet\bar{q}_v)+\frac{\partial}{\partial p}(\bar{\omega}q_v)=Q_2$$

where

$$Q_1=-\frac{\partial}{\partial p}(\overline{w'\theta'})+(c-e)$$
$$Q_2=-\frac{\partial}{\partial p}(\overline{w'q_v'})+(e-c)$$

or correlation terms in the energy budgets e.g. $\overline{Q_1'T'}$

Means we need either

- High temporal and spatial resolution data, or
- To do a lot of the analysis during the model integrations - computational expensive and ***not very flexible*** (matters in this case)
- Also need analysis machines with “large” memory

Cascade

Data requirements (2)

West Africa case studies (total ~50TB)

- 4km Africa (1d field 6.8MB, 3d field 482MB)
- 4TB/model day
- 16TB for the 40 day production run
- 1.5km Africa -1d field 32MB, 3d field 2.2GB
- 2TB/model day
- 30TB for the production run

Warm Pool Case studies (total ~125 TB)

- 4km (1d field 35MB, 3d field 2.4GB)
- 3.2TB/model day
- 64TB 2x10 day production runs
- 1.5km (1d field 236MB, 3d field 16GB)
  (hrly over sub-domain, 3hrly over full domain)
- 10TB/model day
- 40TB production run data

+ 35Tb of data from idealized simulations

This project "finished" in 2012, but analysis continues, and new projects are planned.

These data volumes were troubling, and the next generation projects will also cause Issues …

We hope JASMIN will help.

For process modelling studies, when the purpose of running at high resolution is to be able to analyse the high resolution data at high frequency, data management is as important as compute!

Conclusion

Can't begin to do justice to the data analysis issues here!

However, we know we will have to write out lots of data from these models – even if I/O is hard, our models cannot include all the things that user communities need!

Haven't even touched on one of the most important issues we have: getting scientists to exploit parallel analysis, when they're used to IDL with a bit of shell scripting!

Still, we believe a dedicated data analysis “super data cluster” will be an integral part of our support for “big data” in climate and earth observation for the foreseeable future.

It's also an integral part of our future support for “big data” curation. We could not have gone on as we were.