BRITISH ATMOSPHERIC DATA CENTRE

Presenting a multi-terabyte dataset via the web

Ag Stephens

BADC Data Scientist

11 November 2003

http://badc.nerc.ac.uk

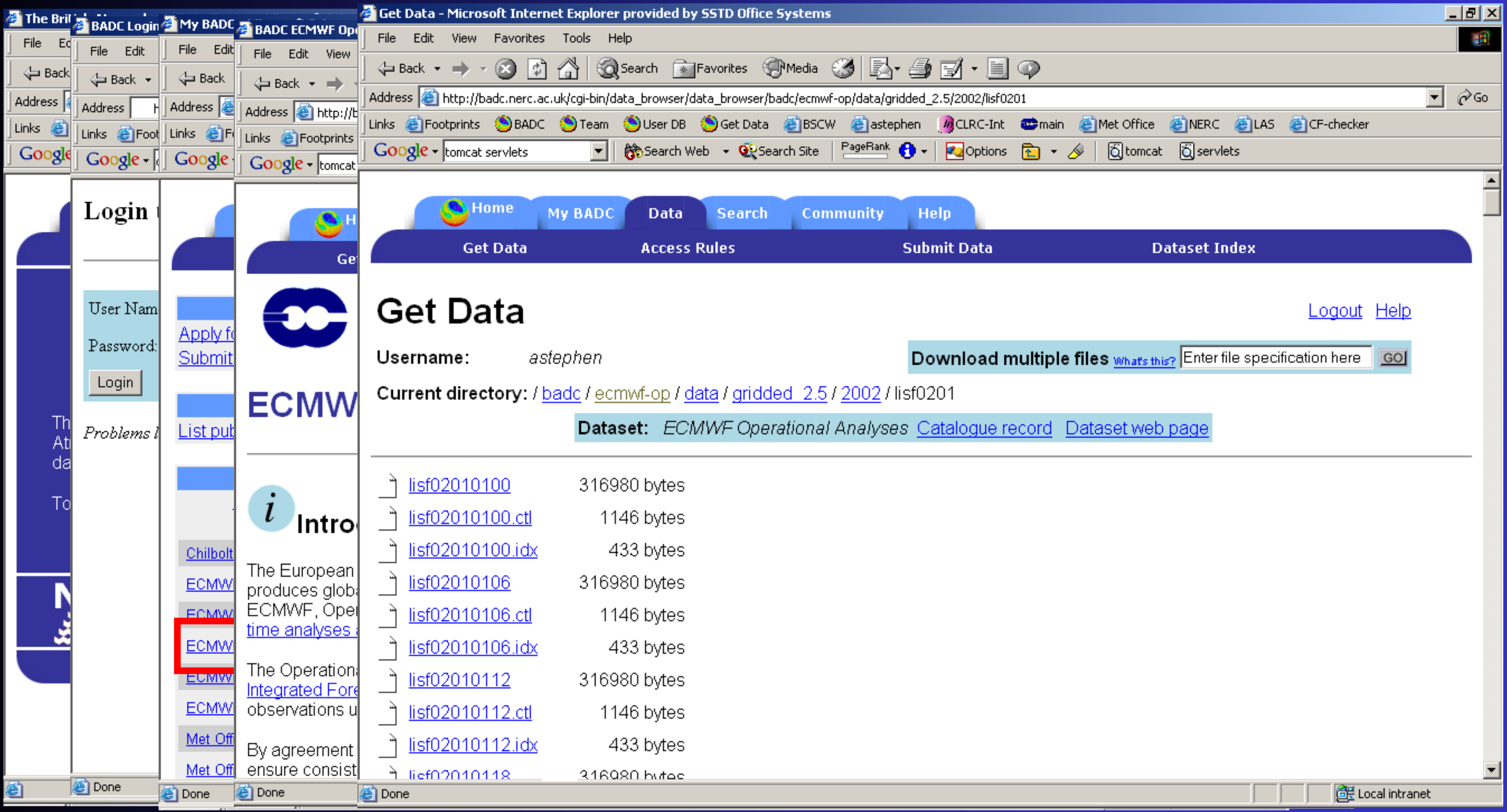
Presentation outline

- **An introduction to the BADC.**
- **The project stages for delivering a Live Access Server:**
  1. Project design.
  2. Tools to convert the data.
  3. Designing a caching architecture.
  4. Aggregation of data files.
  5. Setting up the Live Access Server.
- **Further demonstration and conclusions.**

What is the BADC?

- The NERC-designated data centre for atmospheric science.
- Over 20 TB of data.
- Serving around 5,000 users.
- Agreement with Met Office and ECMWF to distribute data.

How people use the BADC


Get Data - Microsoft Internet Explorer provided by SSTD Office Systems
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://badc.nerc.ac.uk/cgi-bin/data_browser/data_browser/badc/ecmwf-op/data/gridded_2.5/2002/lisf0201
Go
Links Footprints BADC Team User DB Get Data BSCW astephen CLRC-Int main Met Office NERC LAS CF-checker
Google tomcat servlets Search Web Search Site PageRank Options tomcat servlets
Home My BADC Data Search Community Help
Get Data Access Rules Submit Data Dataset Index
Logout Help
Download multiple files What's this?
Enter file specification here
GO
Done
Local intranet
The Bri
BADC Login
My BADC
BADC ECMWF Op
Login
User Nam
Password:
Login
Problems l
Apply fo
Submit
List pub
Chilbolt
ECMW
ECMW
ECMW
ECMW
ECMW
Met Off
Met Off
ECMW
Introd
The European
produces globa
ECMWF, Oper
time analyses
The Operation
Integrated Fore
observations u
By agreement
ensure consist
Th
At
da
To

Get Data

Username: *astephen*

Current directory: / badc / ecmwf-op / data / gridded_2.5 / 2002 / lisf0201

Dataset: *ECMWF Operational Analyses* Catalogue record Dataset web page

lisf02010100	316980 bytes
lisf02010100.ctl	1146 bytes
lisf02010100.idx	433 bytes
lisf02010106	316980 bytes
lisf02010106.ctl	1146 bytes
lisf02010106.idx	433 bytes
lisf02010112	316980 bytes
lisf02010112.ctl	1146 bytes
lisf02010112.idx	433 bytes
lisf02010118	316980 bytes

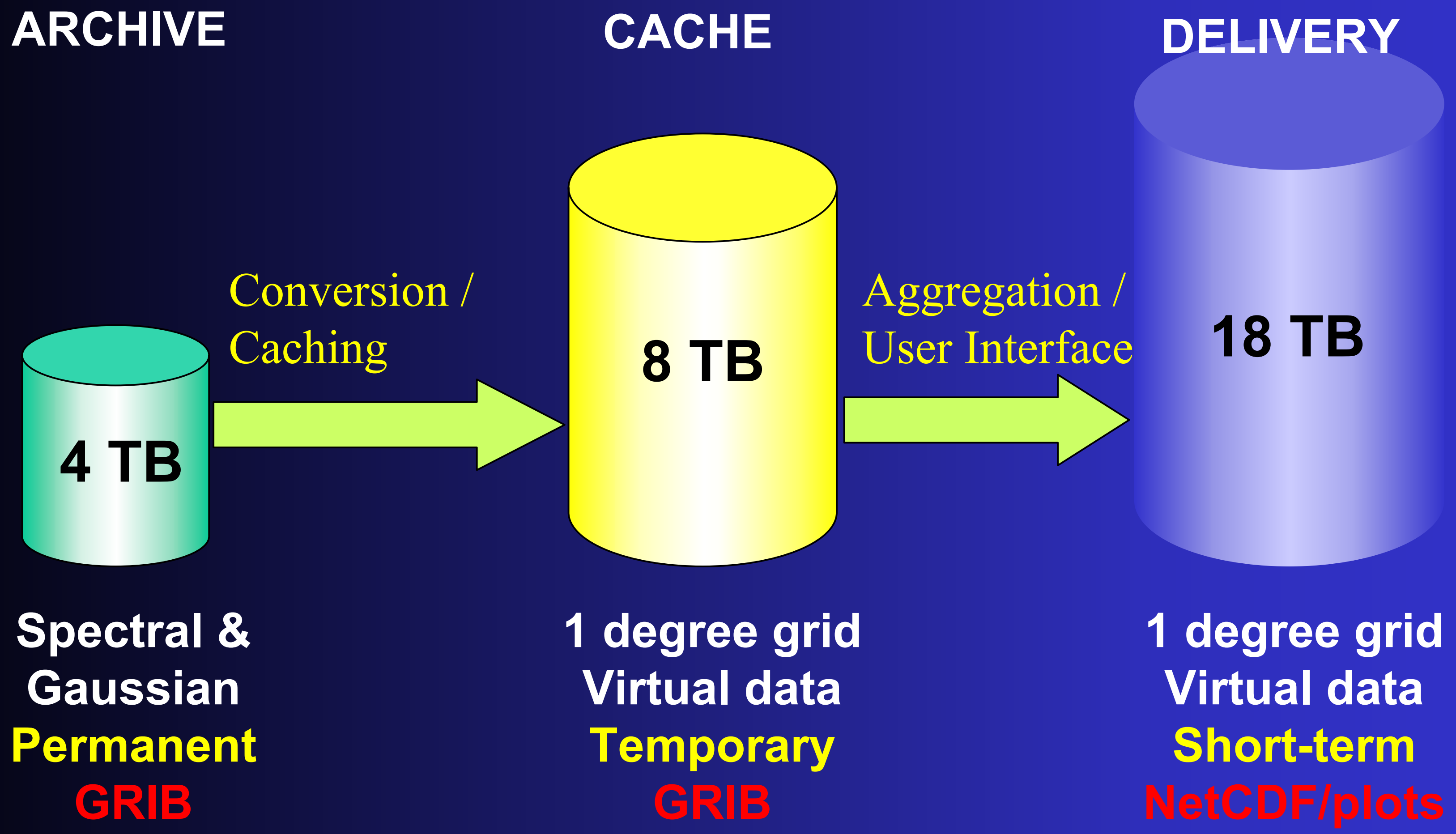
NCAS NATURAL ENVIRONMENT RESEARCH COUNCIL

http://badc.nerc.ac.uk

CLRC

NATURAL ENVIRONMENT RESEARCH COUNCIL

LAS Project Stage 1: Project Design



ARCHIVE

4 TB

Spectral &
Gaussian
Permanent
GRIB

Conversion /
Caching

CACHE

8 TB

1 degree grid
Virtual data
Temporary
GRIB

Aggregation /
User Interface

DELIVERY

18 TB

1 degree grid
Virtual data
Short-term
NetCDF/plots

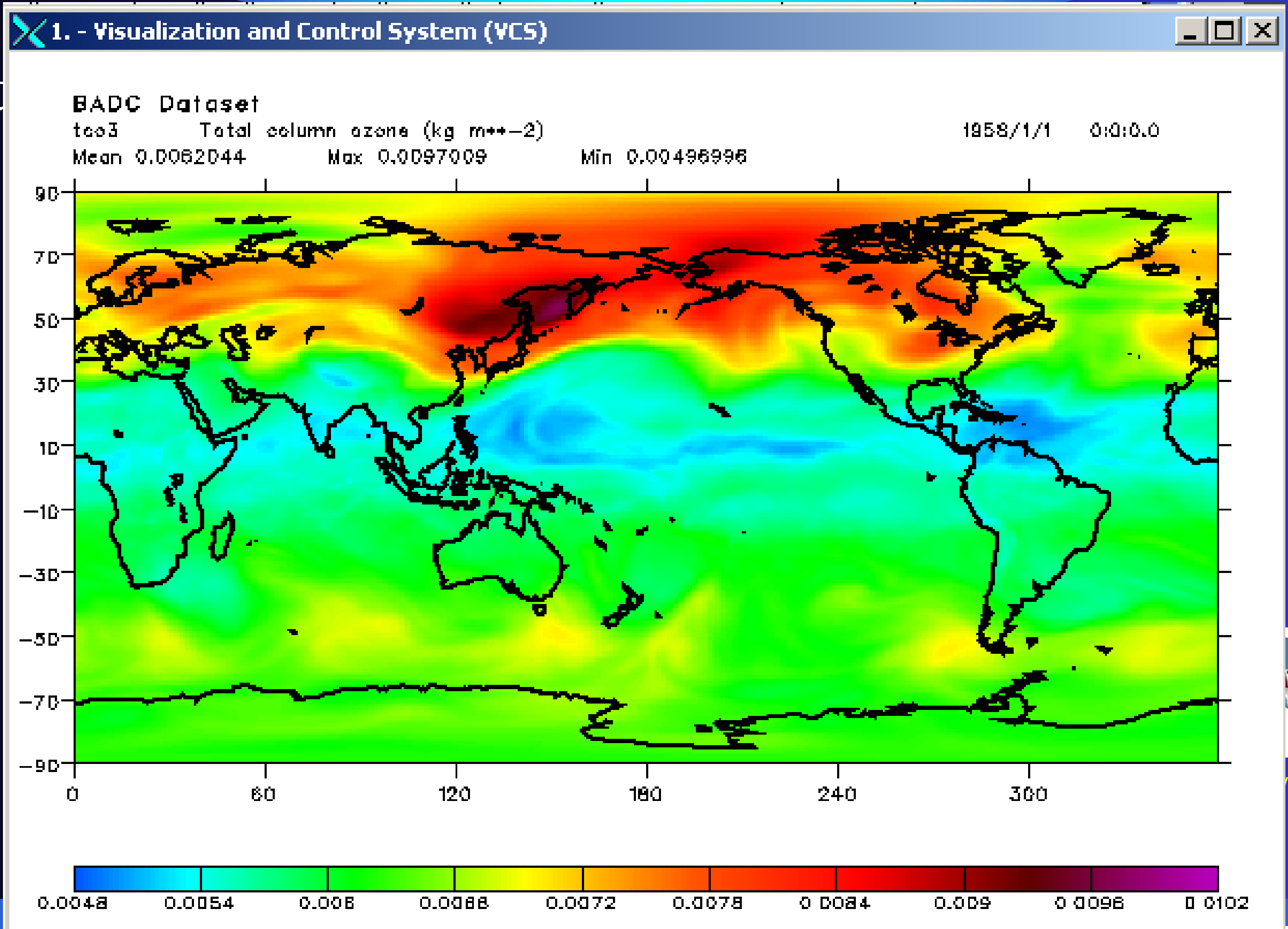
LAS Project Stage 2: Conversion Tools

Int

1.

2.

3.



1. - Visualization and Control System (VCS)

BADC Dataset

tco3 Total column ozone (kg m**-2) 1958/1/1 0:0:0.0

Mean 0.0062044 Max 0.0097009 Min 0.00496996

4.

5.

6.

7.

8.

/cdat

LAS Project Stage 3: Caching

- Cache copy of directory structure.
- Cache algorithms written in Python.
- Control data volumes.
- Analyse and process request sizes.
- Cache of about 1 TB initially.

LAS Project Stage 4: File Aggregation

Climate Data Markup Language (CDML) files are created by the cdscan utility.

CDML contains the following sections:

<dataset> - general information at the dataset level.

<axis> - axis dimension information.

<variable> - relating to individual variables.

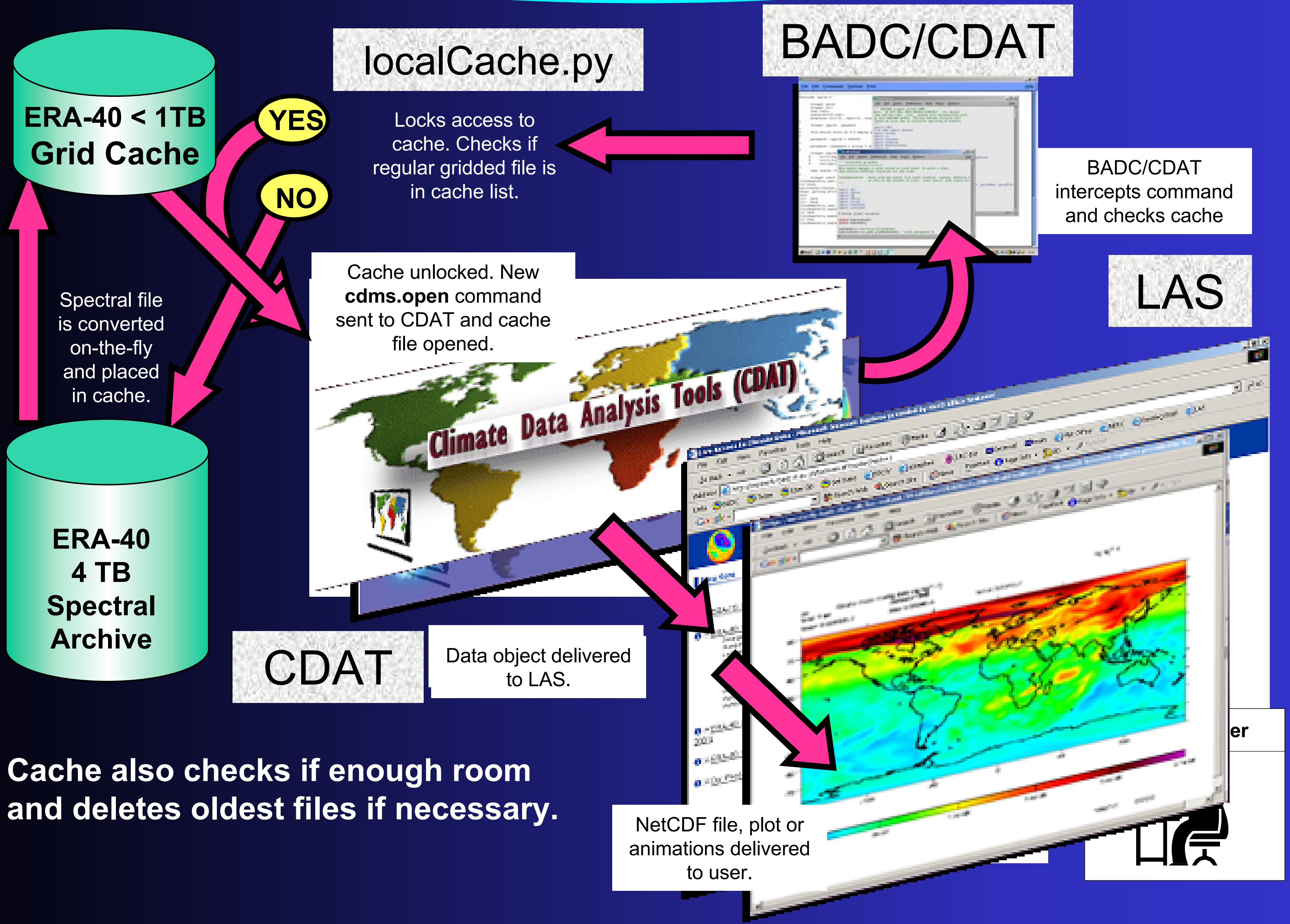
3,000,000 files from one 21KB XML file!

LAS Project Stage 5: Live Access Server

Work required to configure LAS:

1. Configuring Apache webserver (RedHat Linux).
2. Configuring Tomcat Java Servlet Engine.
3. Interfacing to MySQL database.
4. Ingesting CDML files into LAS.
5. Security layer (pending).

How it all fits together



BADC/CDAT intercepts command and checks cache

Locks access to cache. Checks if regular gridded file is in cache list.

Spectral file is converted on-the-fly and placed in cache.

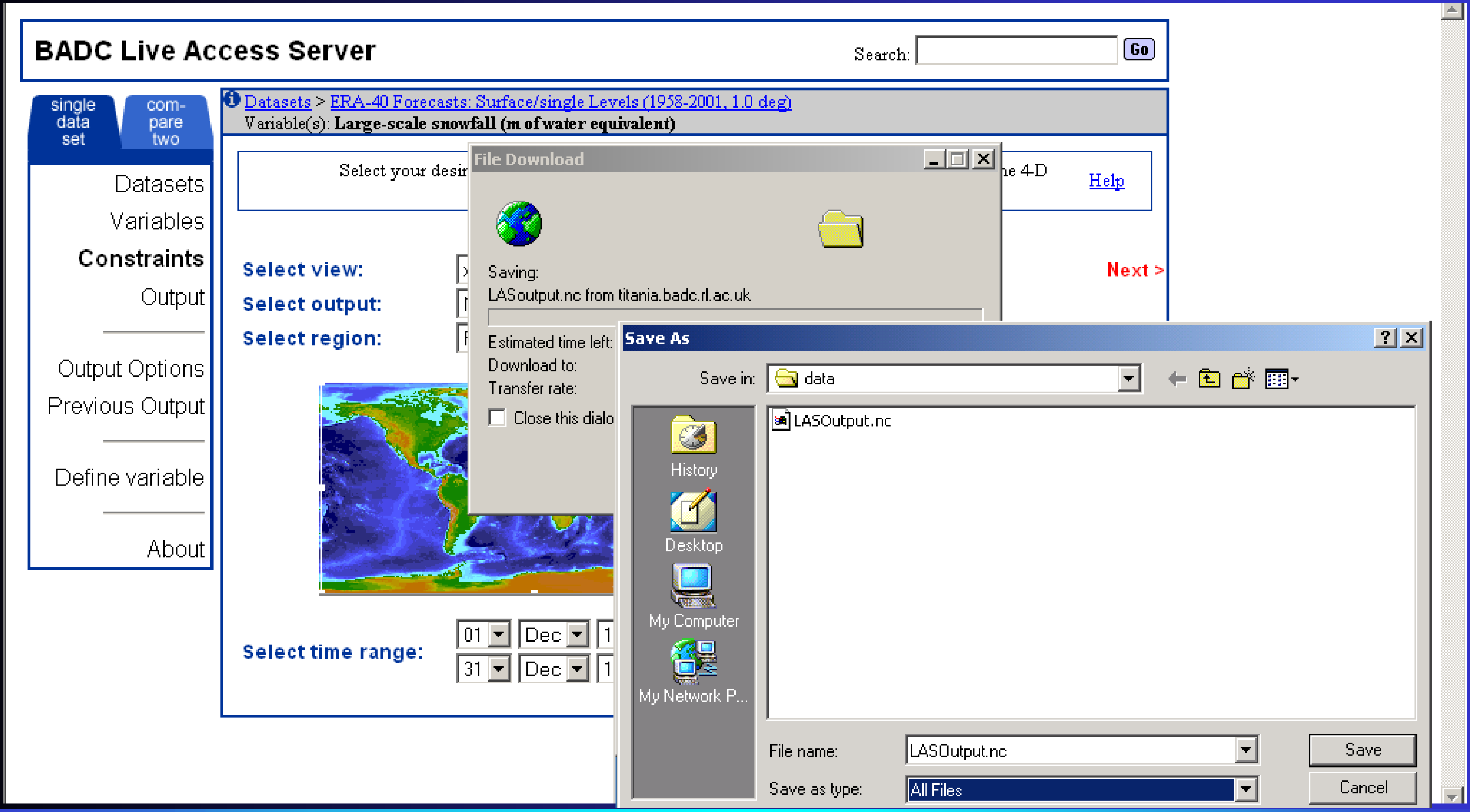
Cache unlocked. New **cdms.open** command sent to CDAT and cache file opened.

Data object delivered to LAS.

NetCDF file, plot or animations delivered to user.

Cache also checks if enough room and deletes oldest files if necessary.

BADC LAS Demo 1: 1 month to NetCDF



BADC Live Access Server

Search: Go

single data set | com-pare two

Datasets

Variables

Constraints

Output

Output Options

Previous Output

Define variable

About

Datasets > ERA-40 Forecasts: Surface/single Levels (1958-2001, 1.0 deg)

Variable(s): **Large-scale snowfall (m of water equivalent)**

Select your desir … he 4-D Help

Select view:

Select output:

Select region:

Select time range: 01 Dec 1 … 31 Dec 1 …

Next >

File Download

Saving:

LASoutput.nc from titania.badc.rl.ac.uk

Estimated time left:

Download to:

Transfer rate:

□ Close this dialo

Save As

Save in: data

History

Desktop

My Computer

My Network P...

LASOutput.nc

File name: LASOutput.nc

Save as type: All Files

Save

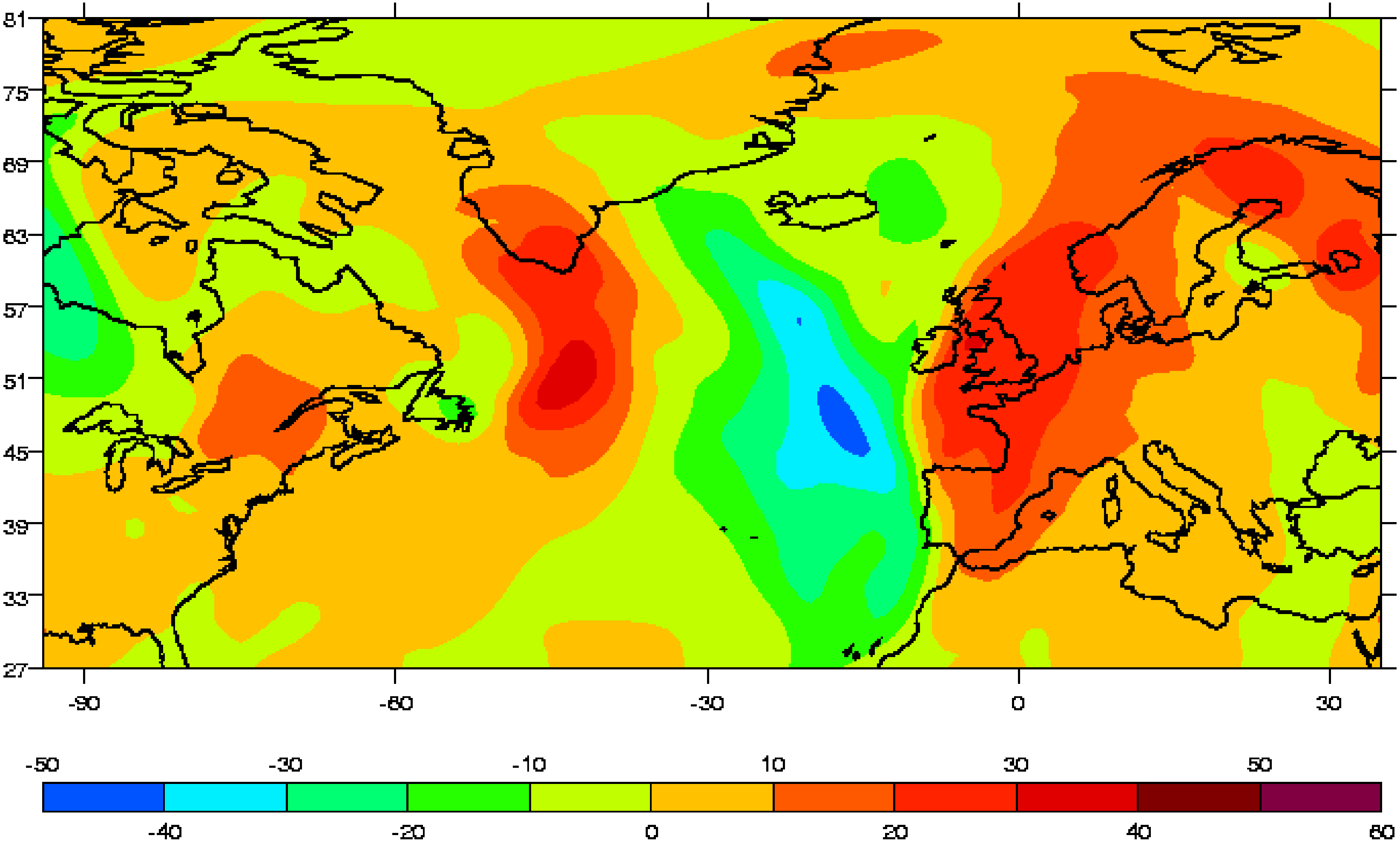
Cancel

http://badc.nerc.ac.uk

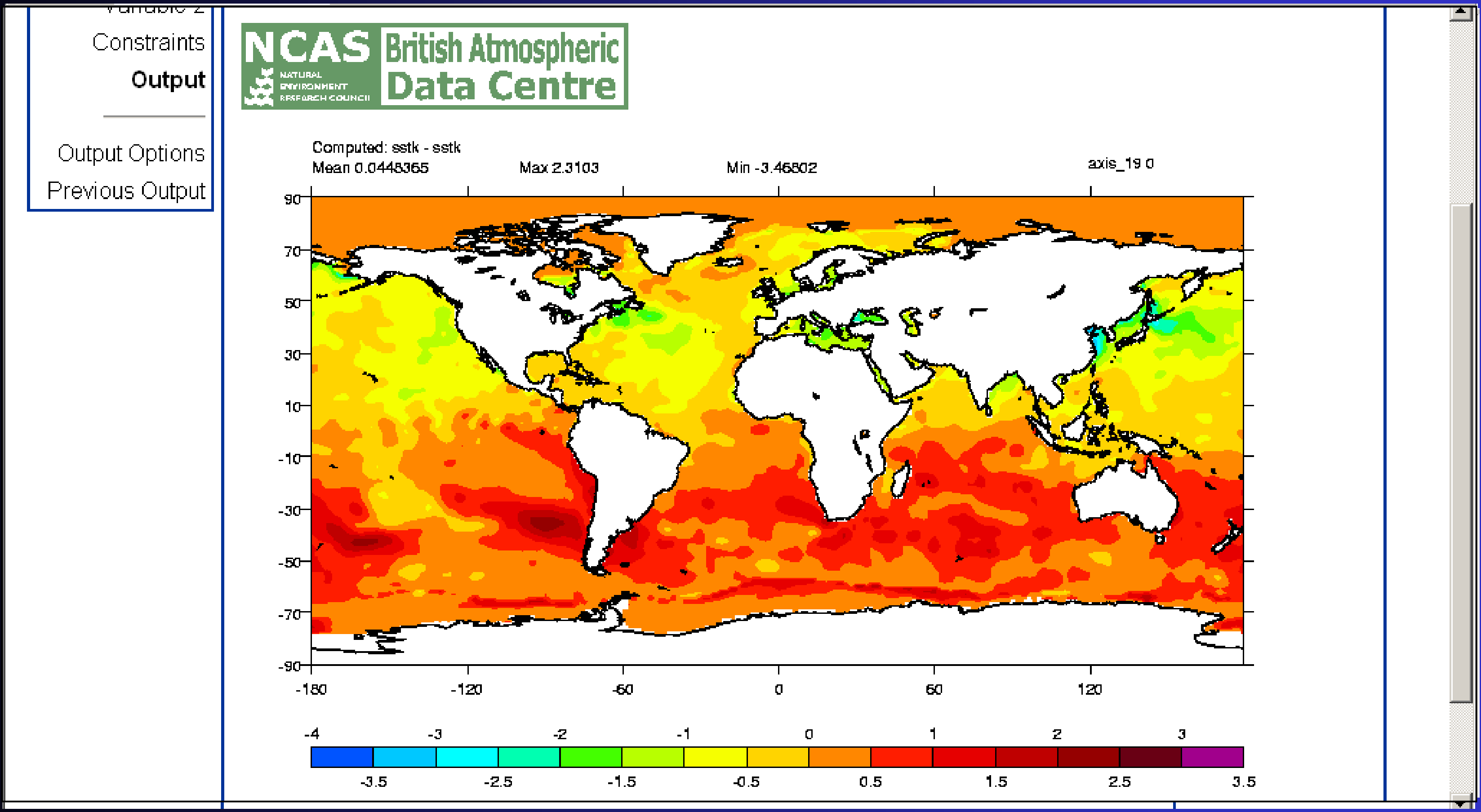
ERA-40 Re-analysis Data

v V-velocity (ms**-1) 1987/10/10 0:0:0.0

Mean 1.12876 Max 38.7089 Min -42.7208



BADC LAS Demo 3: Departure plot

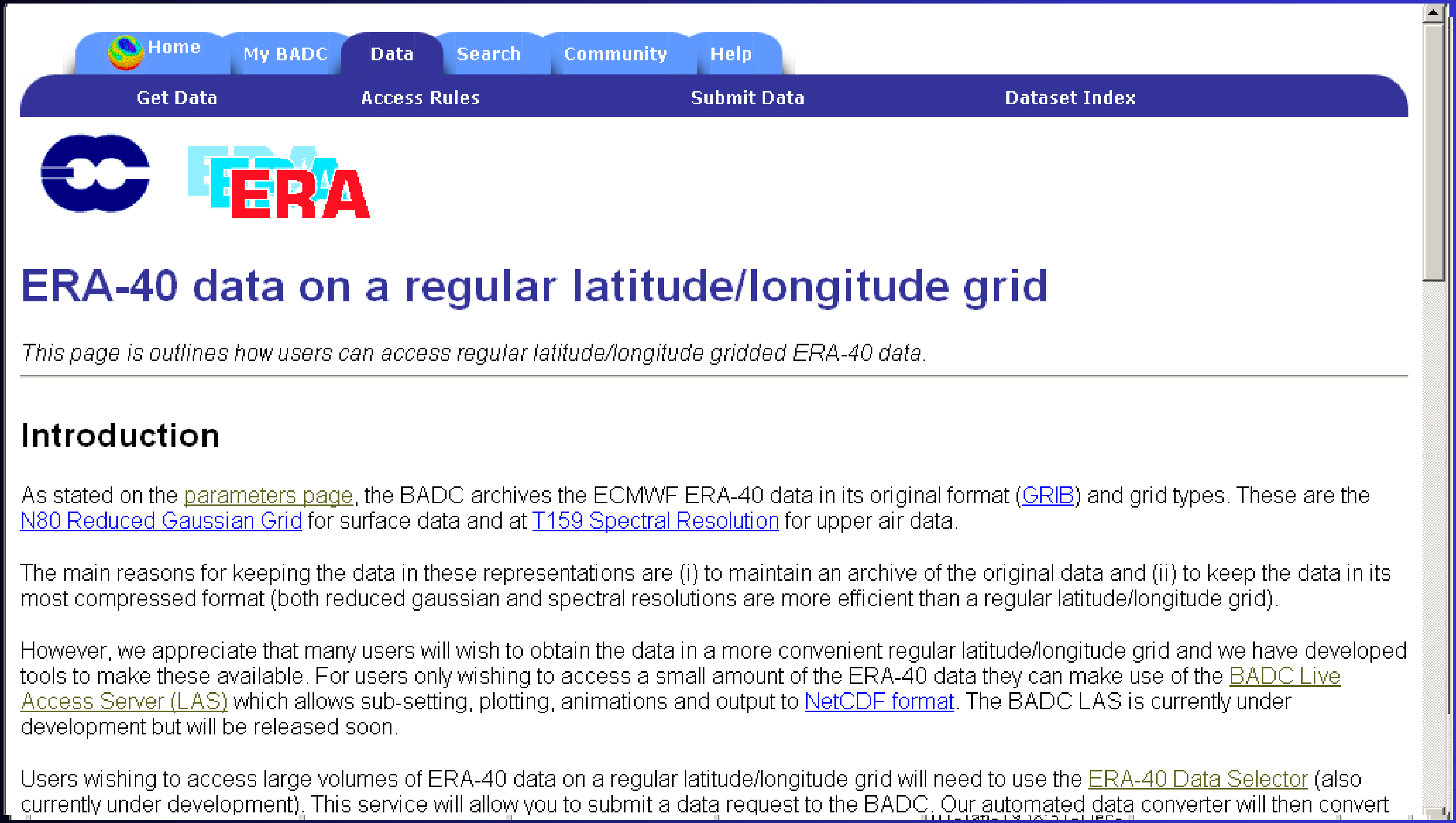


Computed: sstk - sstk

Mean 0.0448365 Max 2.3103 Min -3.46802 axis_19 0

http://badc.nerc.ac.uk

BADC LAS Demo: Metadata



ERA-40 data on a regular latitude/longitude grid

This page is outlines how users can access regular latitude/longitude gridded ERA-40 data.

Introduction

As stated on the parameters page, the BADC archives the ECMWF ERA-40 data in its original format (GRIB) and grid types. These are the N80 Reduced Gaussian Grid for surface data and at T159 Spectral Resolution for upper air data.

The main reasons for keeping the data in these representations are (i) to maintain an archive of the original data and (ii) to keep the data in its most compressed format (both reduced gaussian and spectral resolutions are more efficient than a regular latitude/longitude grid).

However, we appreciate that many users will wish to obtain the data in a more convenient regular latitude/longitude grid and we have developed tools to make these available. For users only wishing to access a small amount of the ERA-40 data they can make use of the BADC Live Access Server (LAS) which allows sub-setting, plotting, animations and output to NetCDF format. The BADC LAS is currently under development but will be released soon.

Users wishing to access large volumes of ERA-40 data on a regular latitude/longitude grid will need to use the ERA-40 Data Selector (also currently under development). This service will allow you to submit a data request to the BADC. Our automated data converter will then convert

What have we learnt?

Advantages of our approach:

- Multiple TB via one interface – some virtual!
- Users saved from files and formats.
- New options for sub-setting and plotting.
- Automatic monitoring of data usage.
- Caching system available for other purposes.
- Knowledge of CDAT and LAS for other projects.

What have we learnt?

Disadvantages:

- No automatic response to massive requests.
- Limits to configurations of plots and animations.
- Caching database is slow.
- Only one dataset presented so far.

To the future...

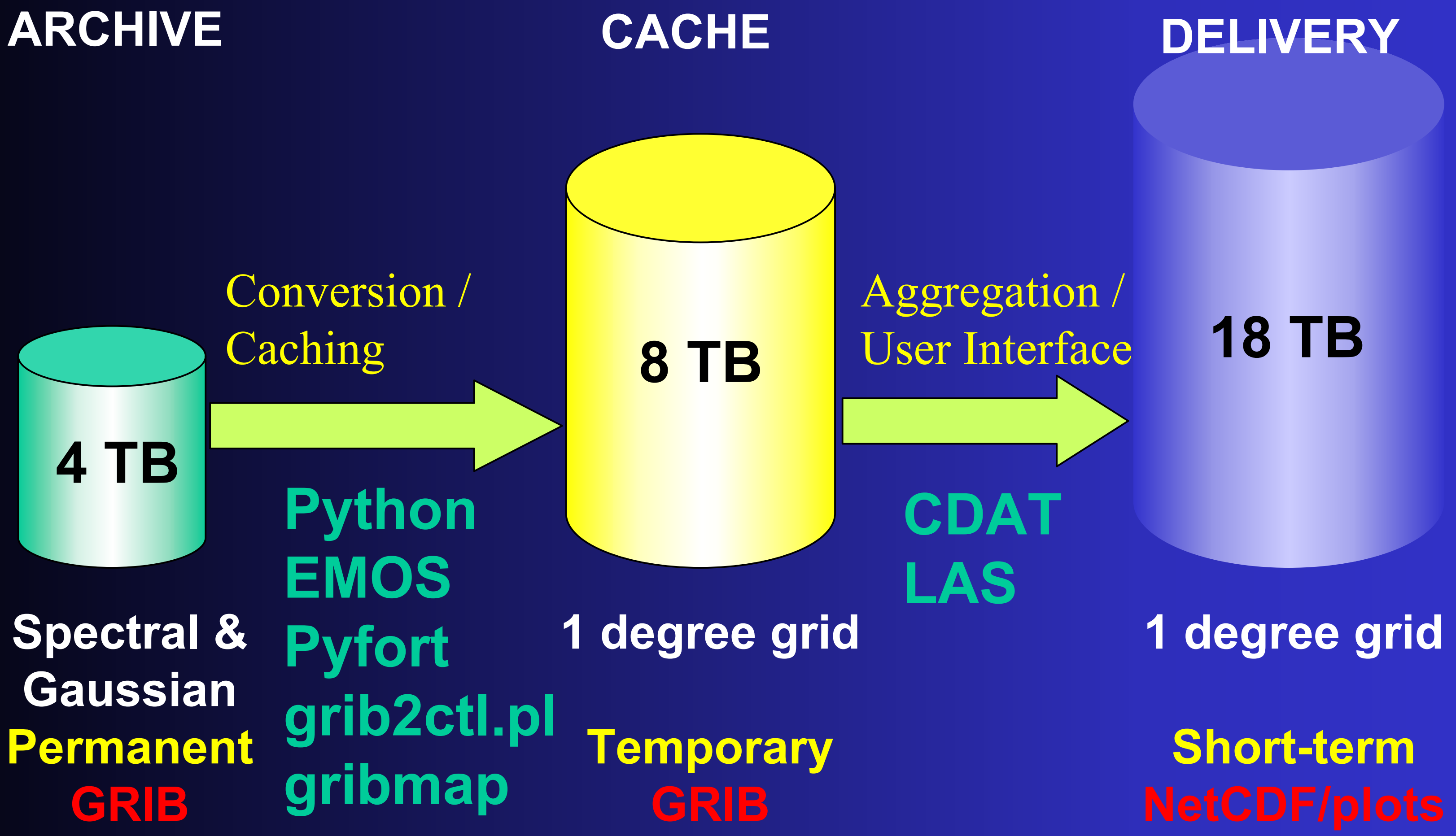
We plan to:

- Implement parallel LASes (ECMWF, UM, COAPEC).
- Implement a time algorithm to keep users informed.
- Generate user-defined LASes on-the-fly.
- Allow comparison of different datasets.
- Re-think the caching database interaction for speed.
- Look to parallelise the background file conversions.

Useful Links

BADC:	http://badc.nerc.ac.uk
CDAT:	http://esg.llnl.gov/cdat
CDML:	http://esg.llnl.gov/cdat/cdms_html/cdms-6.htm
LAS:	http://ferret.pmel.noaa.gov/Ferret/LAS/ferret_LAS.html
Pyfort:	http://pyfortran.sourceforge.net

LAS Project Overview



ARCHIVE

4 TB

Spectral & Gaussian

Permanent

GRIB

Conversion / Caching

Python
EMOS
Pyfort
grib2ctl.pl
gribmap

CACHE

8 TB

1 degree grid

Temporary

GRIB

Aggregation / User Interface

CDAT
LAS

DELIVERY

18 TB

1 degree grid

Short-term

NetCDF/plots

http://badc.nerc.ac.uk