Analyzing Big Data in the Atmospheric Sciences: Community Intercomparison Suite

Nick Schutgens
Zak Kipling
Philip Stier
Philip Kershaw
Stephen Pascoe

UNIVERSITY OF OXFORD

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

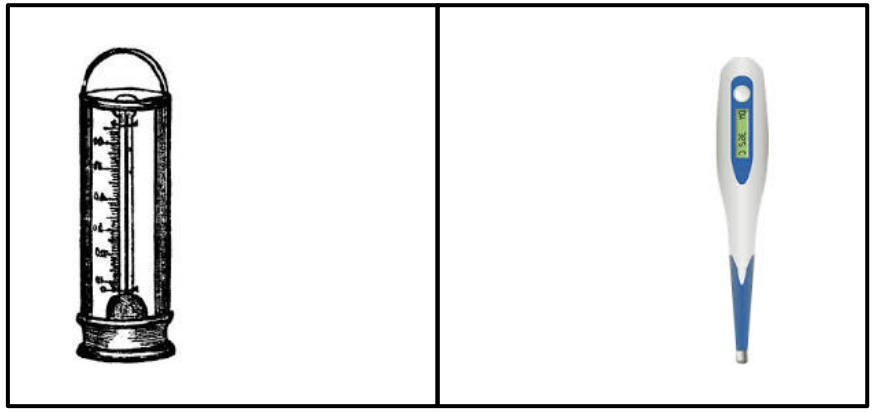
Issues with Big Data

- Well, it's… big
- But it's also diverse:
  - Formats
  - Spatio-temporal sampling
  - Grid structure
  - Content (spatio-temporal extent / variables)
- It also allows many choices:
  - (Statistical) analyses
  - Plot types

User sensory overload

Another issue with data

- You may want to compare it to other data…

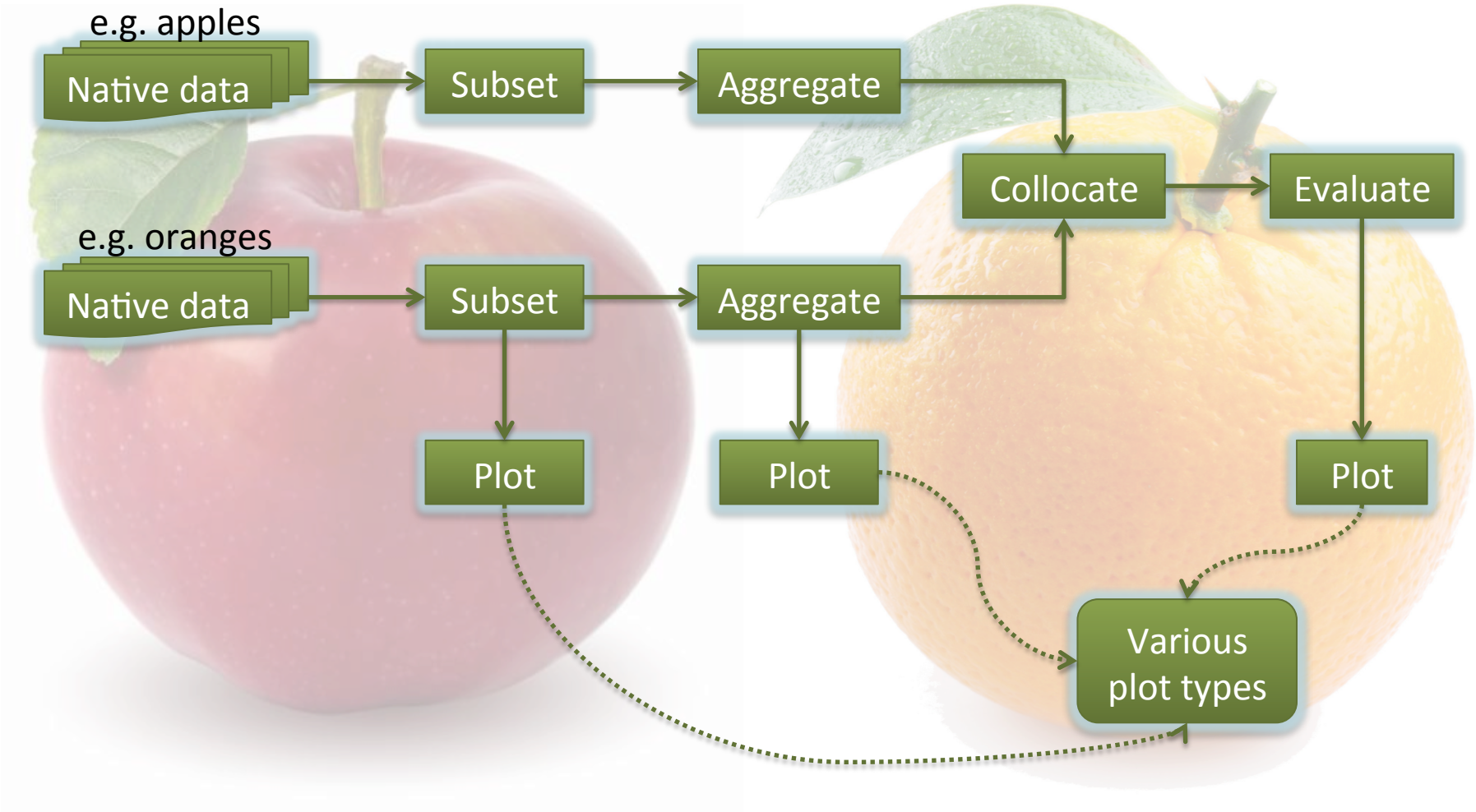


- Spatio-temporal collocation of datasets is an important part of scientific work (or should be!): subsetting, aggregation & interpolation

Community Intercomparison Suite

- *Generic* tool for *visualising and collocating native* datasets used in (atmospheric) sciences
- *Simple syntax* with *many options*
- Flexible approach through plug-ins by users
- Open Python source code with base code developed by software company

A typical CIS work-flow

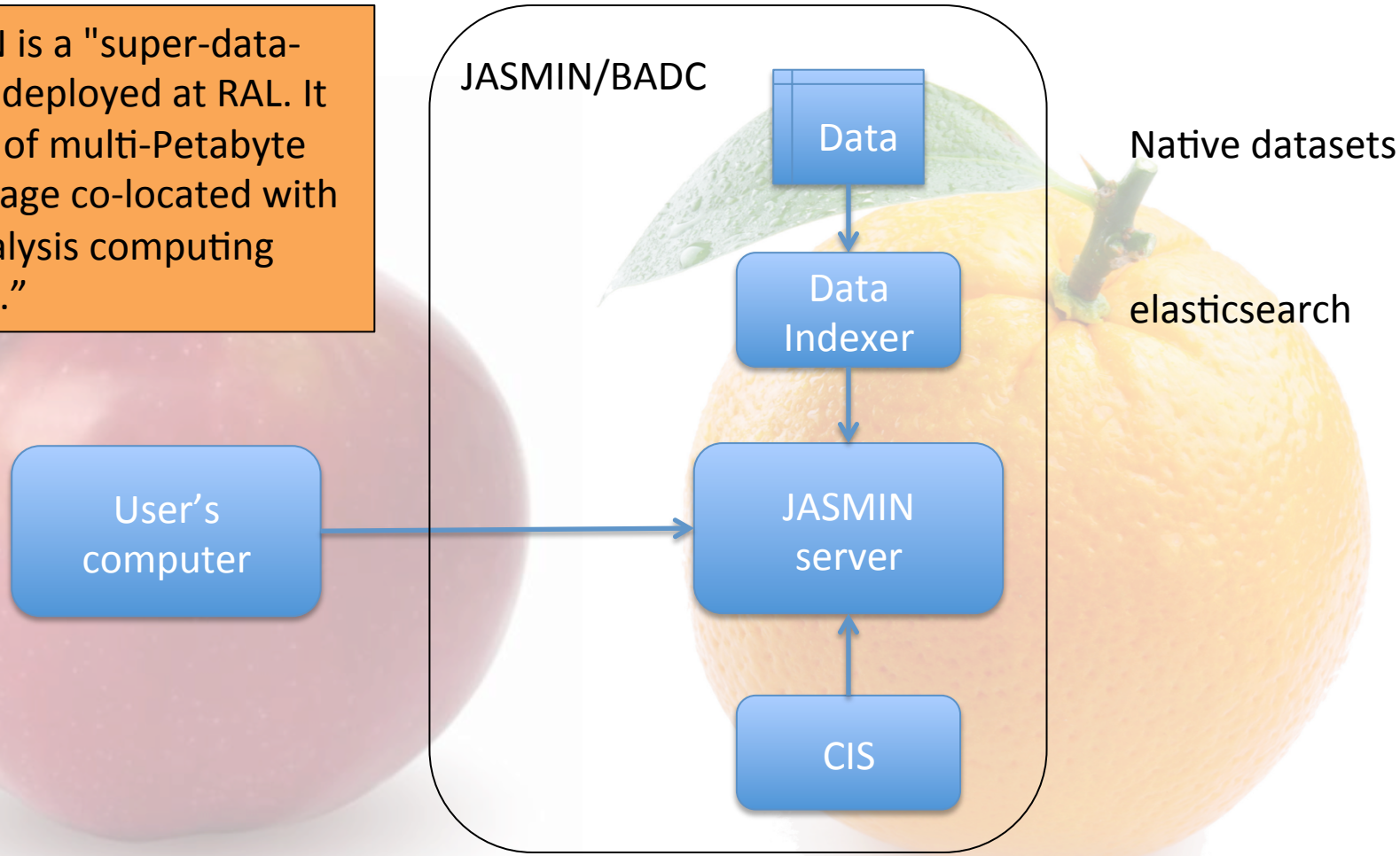


Basic CIS syntax

- cis info <files>
- cis plot <variable>:<files>
- cis subset <variable>:<files> <domain>
- cis aggregate <variable>:<files> <domain & bins>
- cis col <variable>:<files> <other files>
- cis eval <variable>:<files> <other variable>:<other files> <arithmetic operation>
- cis stats <variable>:<files> <other variable>:<other files>

CIS on JASMIN

"JASMIN is a "super-data-cluster" deployed at RAL. It consists of multi-Petabyte fast storage co-located with data analysis computing facilities."



But CIS should work on any system with Python and IRIS

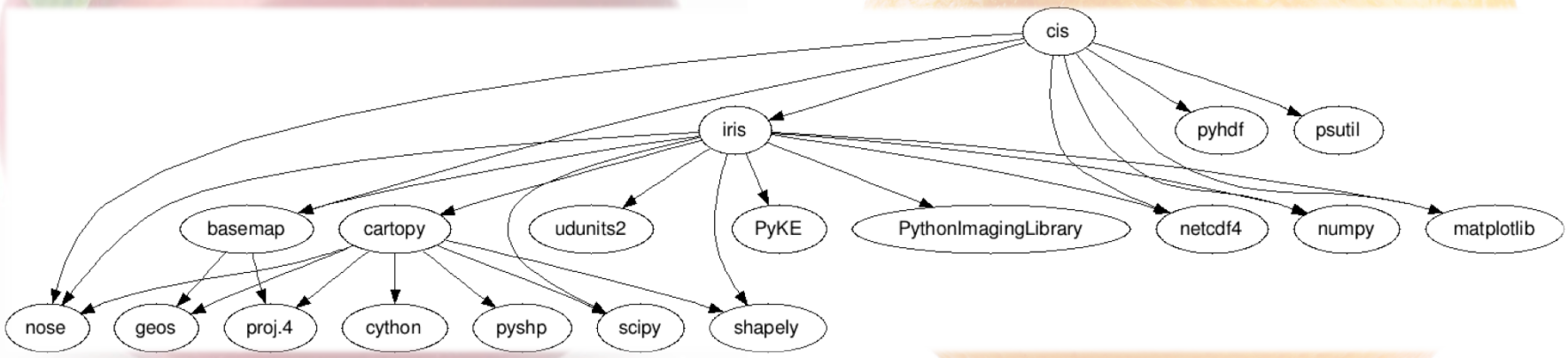
CIS objects & dependencies & API's

Gridded object

- Regular lat,lon,z (or p), time
- Used for model data and aggregated CIS products
- Relies on IRIS

Ungridded object

- Irregular lat,lon, z (or p) or time
- Used for (most) observations
- Sets CIS apart from NCO, CDO



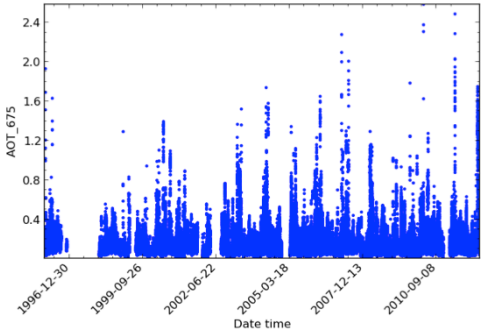
CIS API's documented & available for third party use (e.g. Data Indexer)

Flexibility of CIS

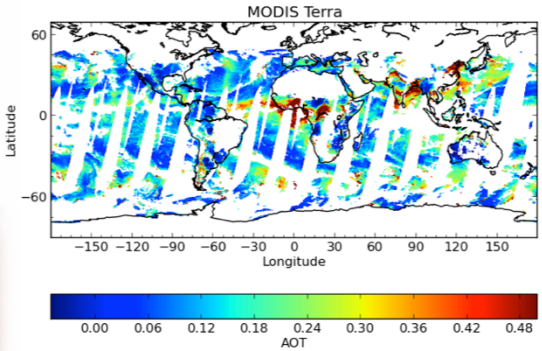
- Generic NetCDF & HDF readers
- Data read-in plug-ins
  - Product specific, currently based on filename
  - Establish logical relationships among data
  - Correct for information lacunae
  - CIS comes with plug-ins for many aerosol/cloud related datasets
- Default behaviour, easily overridden
- Collocation plug-ins
  - Define arbitrary averaging kernels & statistics

Plotting original data

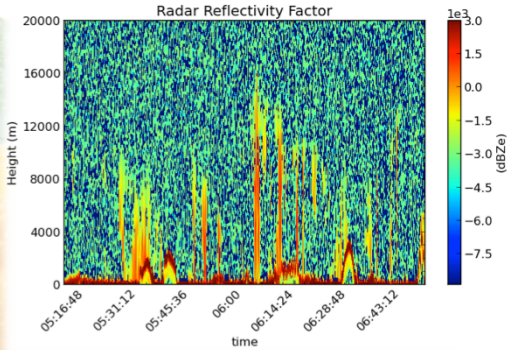
AERONET station data

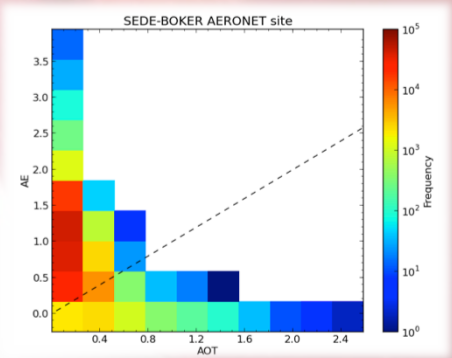


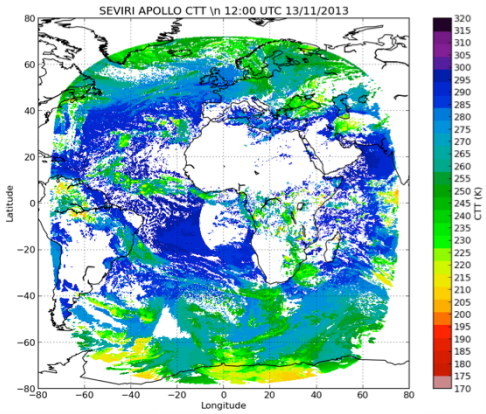
Satellite imager data

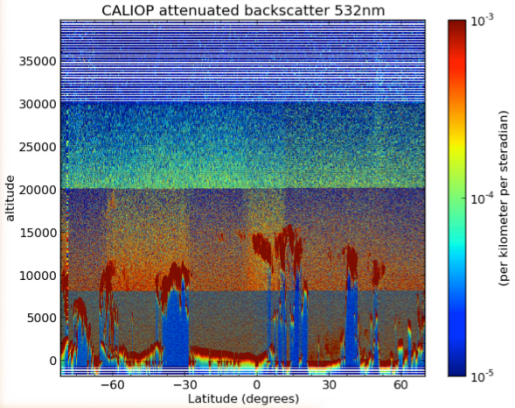


Satellite active profiler data









1

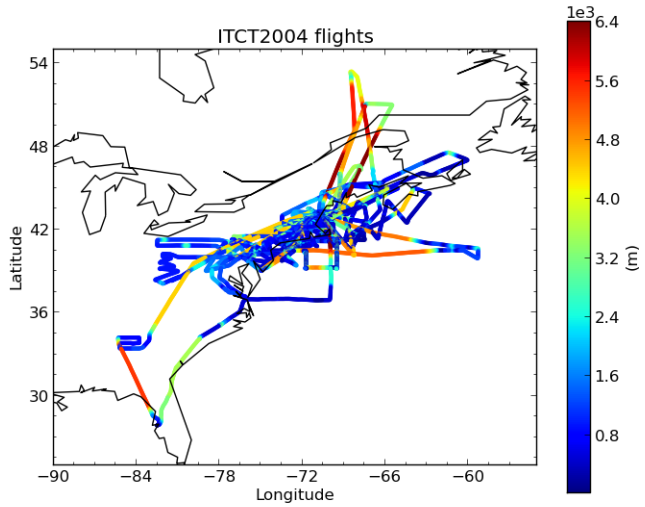
Number of required commands to create plot

```
cis plot <variable>:<files> <options>
```

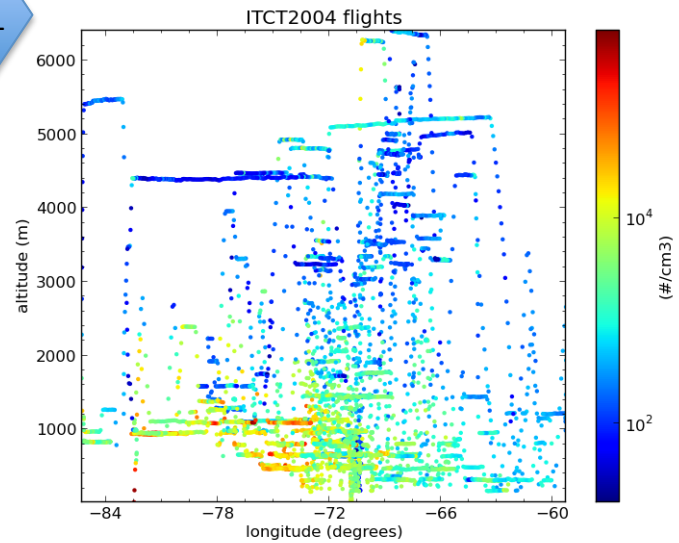
Flight Campaign data

Collaboration: project GASSP homogenizes many Flight Campaign datasets (PI: K. Carslaw, Leeds)

1



1



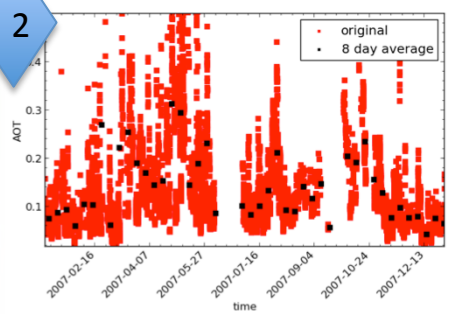
18 files, each a single flight over a period of a month

```
cis plot <variable>:<files> --xaxis=<variable> --yaxis=<variable>
```

Aggregating data

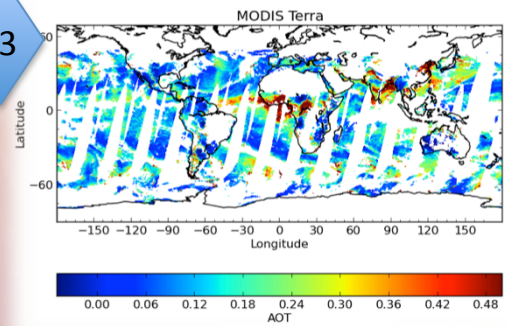
Aggregation averages data and puts it on a regular grid. For each new gridbox, standard deviation as well as total nr of points of the original data is also available.

2

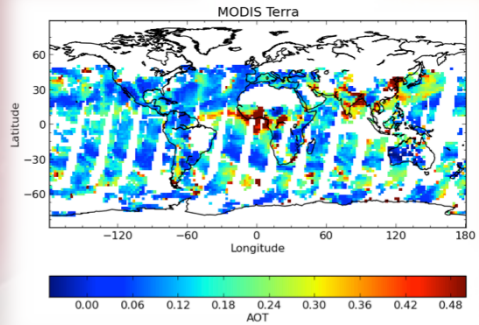


3

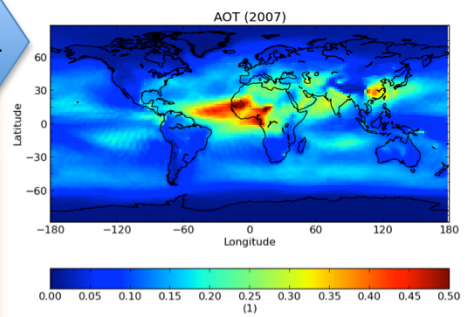
134 files 200 MB

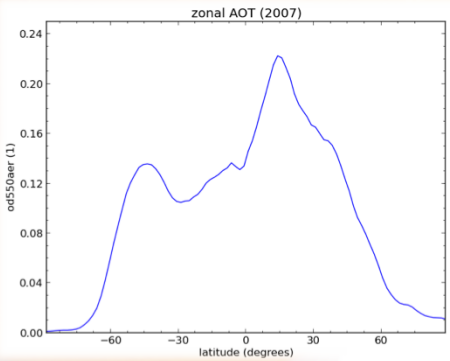


1 file 1.6 MB



4





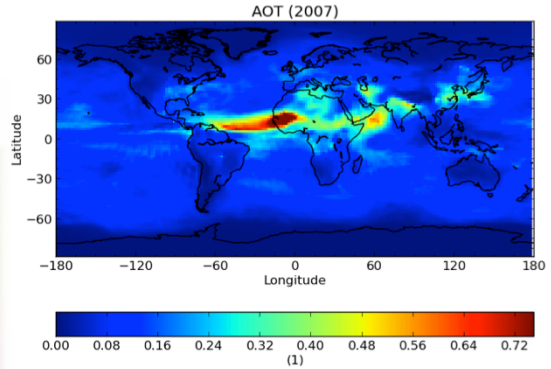
Purpose:

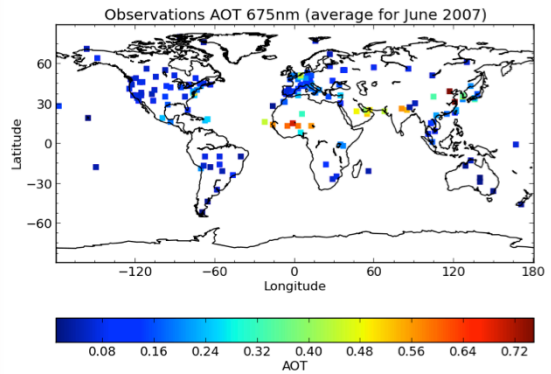
- Regularize grid
- Reduce data size
- Smooth field
- Analysis

```
cis aggregate <variables>:<files> <domain & bins> -o <new file>
```

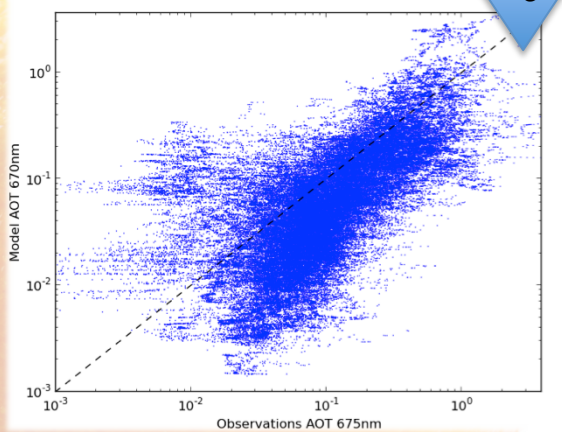
Spatio-temporal Collocation

Model gives global output every 3 hours for a full month





Observations are day-time site measurements, every 15 min for a full month

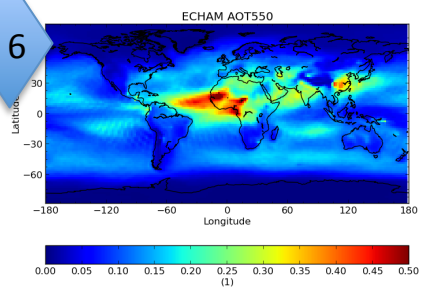


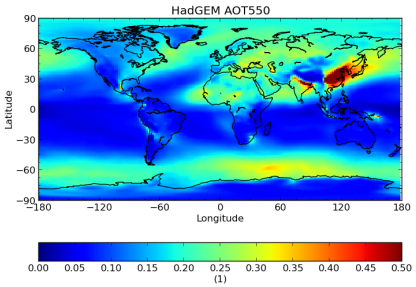
6

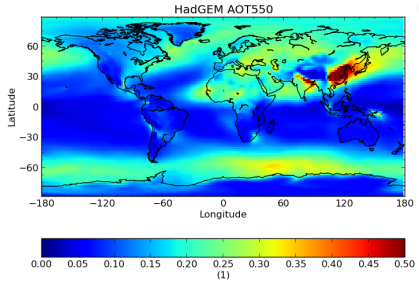
```
cis col <variable>:<file> <file>:<method> -o <new file>
```

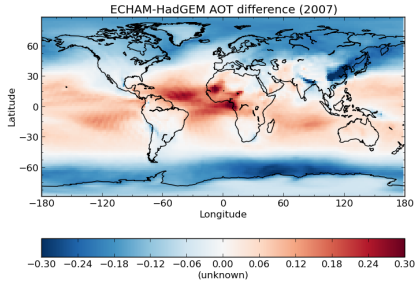
Arithmetic & Masking

6

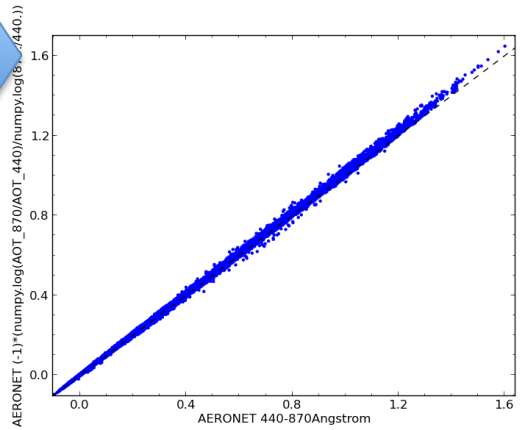




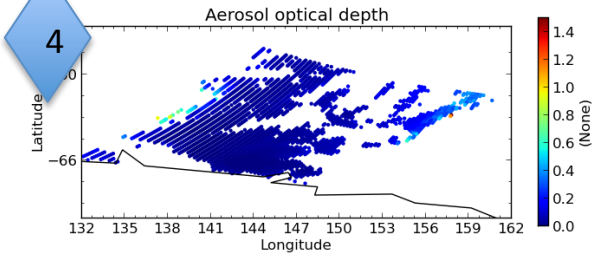


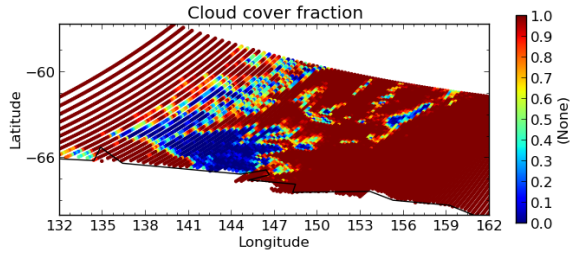


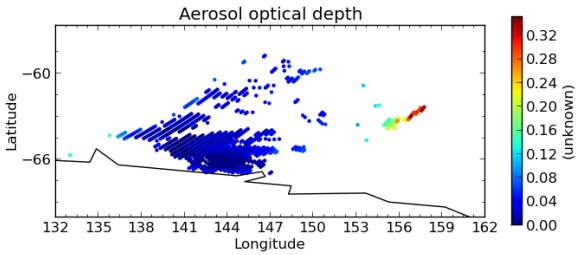
2



4

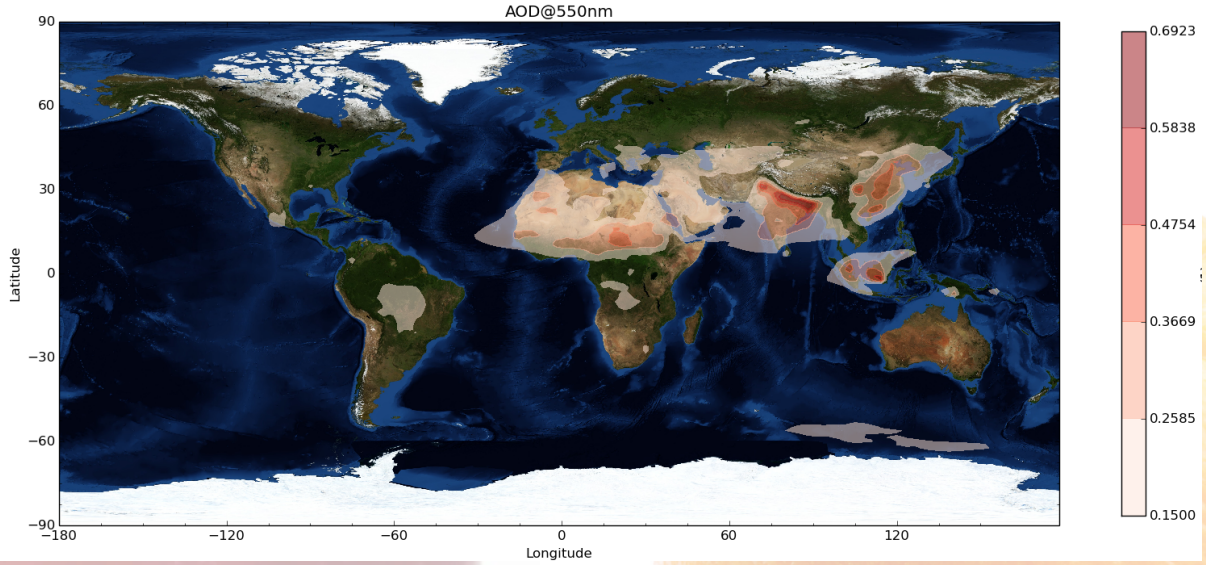







```
cis eval <variable>:<file> <variable>:<file> <expression> -o <new file>
```

Community Intercomparison Suite



- First release imminent
- We are thinking of a one-day course
  - Support from devs
- We have a website and soon a better one:
  - http://proj.badc.rl.ac.uk/cedaservices/wiki/JASMIN/CommunityIntercomparisonSuite