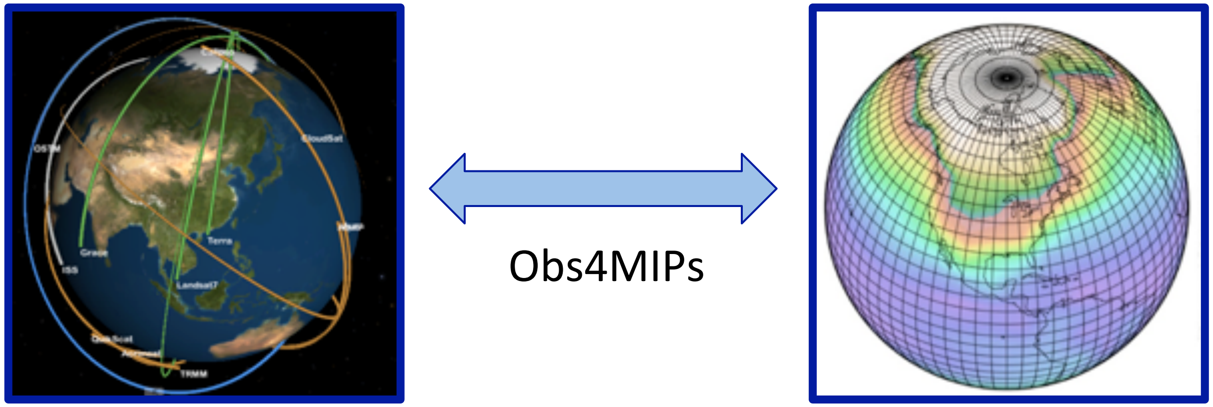
# ESGF Functionality Needed for obs4MIPs and Other Datasets



Robert Ferraro

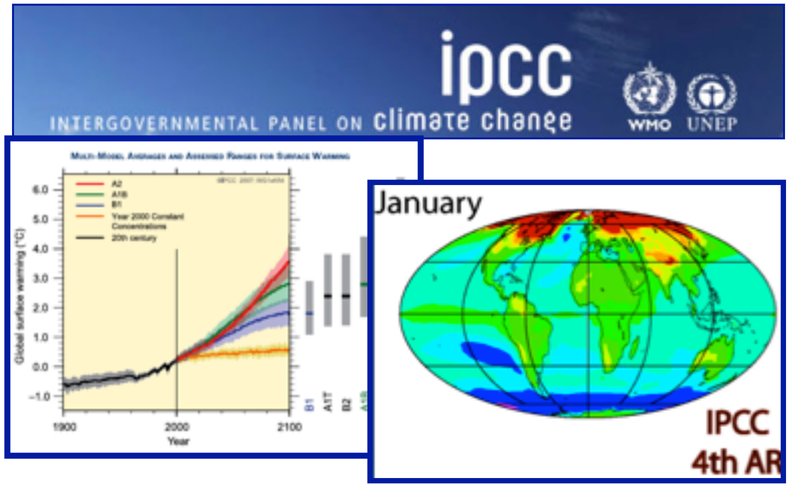
*Jet Propulsion Laboratory*

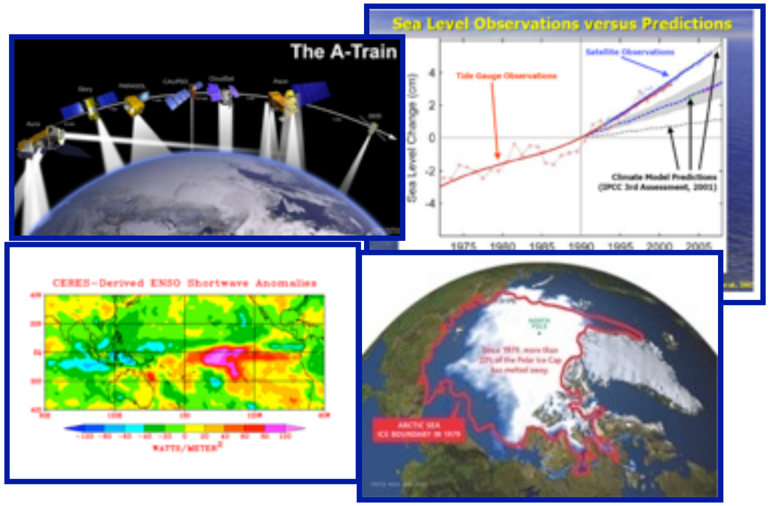
Jet Propulsion Laboratory
California Institute of Technology

PCMDI
Program for Climate Model Diagnosis and Intercomparison

# Initial Sentiments Behind obs4MIPS

*Under-Exploited Observations for Model Evaluation*




The A-Train
Sea Level Observations versus Predictions
Satellite Observations
Tide Gauge Observations
Sea Level Change (cm)
Climate Model Predictions
(IPCC 3rd Assessment, 2001)
1975
1980
1985
1990
1995
2000
2005
CERES-Derived ENSO Shortwave Anomalies
WATTS/METER²
ARCTIC SEA ICE BOUNDARY IN 1979

How to bring as much observational scrutiny as possible to the CMIP/IPCC process?

How to best utilize the wealth of satellite observations for the CMIP/IPCC process?

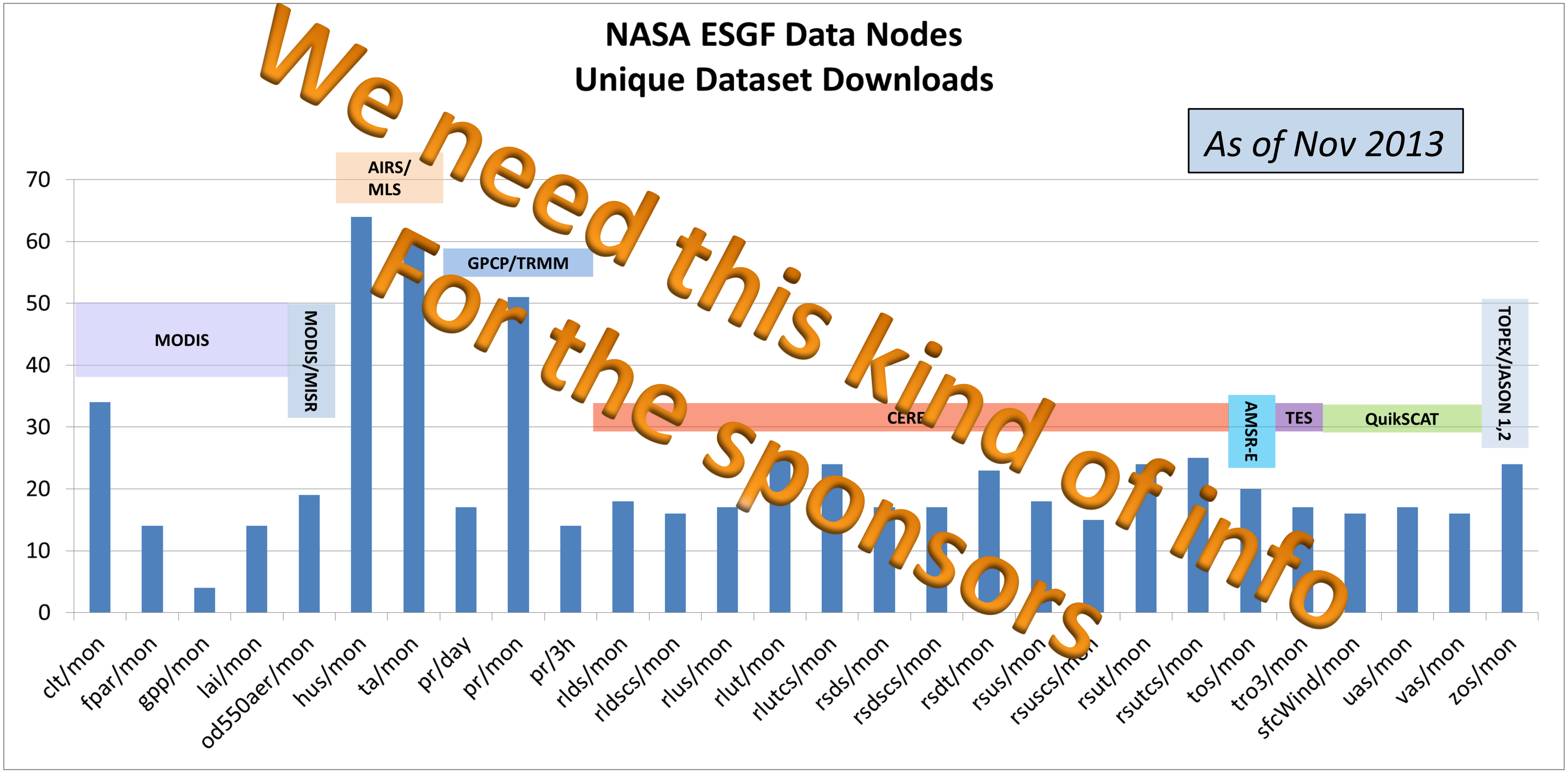
# obs4MIPs Current Holdings

- 50+ (and growing) observational datasets covering a variety of (mostly) CMIP5 output variables
  - CMIP5 output conforming (NetCDF, CF variable names)
  - Modified CMIP5 attributes list (changes that make sense for obs)
  - Modified DRS structure (trying to follow the CMIP convention, but …)
  - Starting to run into problems adhering strictly to the CMIP standards …
- 20+ (and growing) technical notes (supposed to be one for each dataset …)
  - PDF documents
  - They don't show up as individual documents using the standard search (but you can get them if you know how …)
- We have had requests to also host movies made from the files
  - Potentially valuable as quick look products for the end user
  - Doesn't fit the current ESGF data model (same problem as tech notes)

# obs4MIPs: Access Statistics

(NASA Datasets Only)

**119 unique* users from 16 countries**
**641 dataset downloads in 2012-13**



**NASA ESGF Data Nodes**
**Unique Dataset Downloads**

*As of Nov 2013*

We need this kind of info
For the sponsors

* "Unique" counts unique user ID downloads of a complete dataset, not individual files.
Repeat downloads of the same dataset by the same user were counted only once.

# Obs4MIPs – CMIP6 Planning Meeting

## Major findings relevant to ESGF development –

- Include higher frequency datasets, and higher frequency model output

  *Volume, file counts, discoverability, searchability*

- Reliable and defendable error characterization/estimation of observations; Precise definitions of data products, including biases, and precise definitions of the model output variables are required

  *More documentation, ultimately …*

- Include datasets in support of off-line simulators (aka observation proxies)

  *Maybe no impact, as CMIP also requests simulator relevant model output, but perhaps some new attributes?*

- Collocated observations are particularly valuable for diagnosing certain processes. Consideration should be given to arranging for collocated datasets, including sparser in situ data, of known value to specific evaluations.

  *In-situ doesn't fit well within the current data model (not gridded ...)*

  *New attributes*

# Obs4MIPs – CMIP6 Planning Meeting

Additional Recommendations (not consensus) that potentially impact the ESGF –

- Requiring averaging kernels for the retrieval observations.

  *Probably don't look like gridded data files*

- No gatekeeping … Just format enforcement ….

  *This is not likely to happen, but what may evolve is a rating system to be applied to each dataset*

  *Could be handled via attributes, I suppose*

  *BUT format enforcement would be very valuable ( nothing gets published without passing a format test, hopefully automated …)*

# Requirements (ok – Desirements …)
# In no particular order …

- Handle non-NetCDF files seamlessly
  - Visible in search, can be added to the data cart for automated download
- Automated Metrics Collection (standardize record format)
  - Need - who, what file, when, what node
  - Right now we get this data in inconsistent forms due to variations in the hosting DRS, file naming conventions, and multiple duplicate OpenIDs – need standards enforcement to avoid a lot of labor in preparing reports
  - AND need to be able to remove testing records (developers like to do lots of downloads during testing that skew the statistics …)
  - AND need to be able to harvest this data without asking node managers to run special scripts ….
- Expose ALL attributes in the files to Search
  - Time period, in particular, because right now finding just a particular time segment is a completely manual process
  - AND need to figure out how to add searchable attributes for non-NetCDF files …

# Requirements (Continued)

- Format Enforcement as a Condition to Publish
  - This may not appeal to all projects, but the ability to require that a file pass some set of format tests before it is published would be very useful to us
- Project Controls on what get published
  - Again, not everybody will want this
  - But right now there is no way to prevent a node from publishing a file to any project they want
  - (yes, Luca did implement something, but it depends on voluntary cooperation amongst all node managers …)
- Subsetting … Across multiple files …
  - Because there are a lot of users who just want to look at data within a region of space or time, without downloading everything
- Question – Who owns the data cart? ESGF or CoG?
  - The answer will help in assigning requirements ….