



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The Climate Data eXchange: Bringing NASA's Observational Data to the IPCC Community

Amy Braverman¹ Dan Crichton¹ Chris Mattmann¹
Rob Raskin¹ Mike Gunson¹ Dean Williams²

Jet Propulsion Laboratory,
California Institute of Technology
Mail Stop 306-463, 4800 Oak Grove Drive
Pasadena, CA 91109

Program for Climate Model Diagnosis and Intercomparison,
Lawrence Livermore National Laboratory
Mail Code L-103, 7000 East Avenue
Livermore, CA 94550

October 7, 2009



One role for observations in climate studies and IPCC AR5

NASA's Observational Data

Why is the delivery of NASA observations different from that of model output?

SOA Paradigms

Challenges

Summary



Role of observations in climate science:

1. process understanding

- ▶ exploratory data analysis
- ▶ hypothesis formulation

2. parameterization and model development

- ▶ statistical description of sub-grid-scale processes
- ▶ hypothesis testing

3. model evaluation (IPCC)

- ▶ comparison of model output against observations
- ▶ weighting multi-model ensemble members ("scoring")



The reliability of projections could be improved if the models were weighted according to some measure of skill. . . Since there is no verification for a climate forecast on timescales of decades to centuries, the skill or performance of the models needs to be defined, for example, by comparing simulated patterns of present day climate to observations.

–Co-Chairs of IPCC Working Group I and Working Group II

(From “Scoping the IPCC 5th Assessment Report, Proposal for an IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections”)



Remote sensing data are:

1. Massive

- ▶ provide detailed information about processes through multivariate distributions on multiple spatial and temporal scales

2. Heterogeneous

- ▶ have variety of organizational structures, retrieval methods, sampling characteristics, and meaning (not like model output!)

3. Distributed

- ▶ are stored all over the country and the world

4. Virtual

- ▶ need to be assembled "on-the-fly"

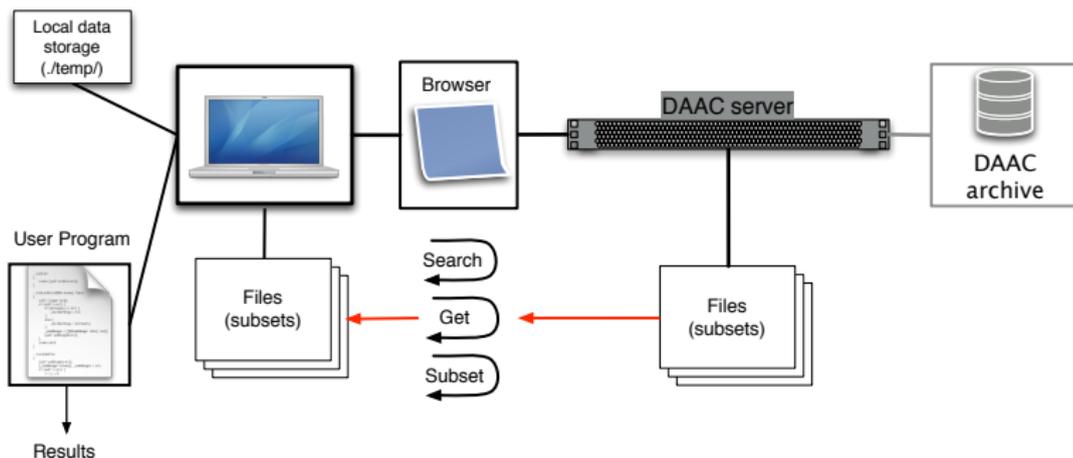


Why are observations different?

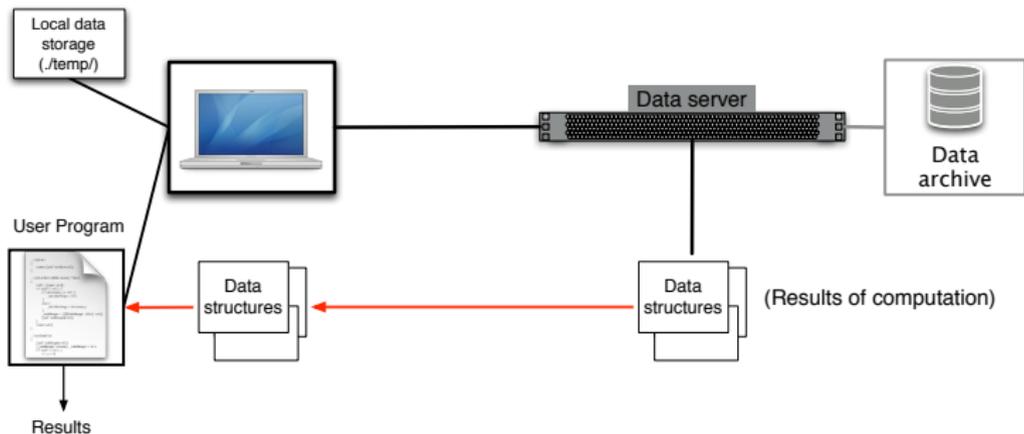
- ▶ How to access data?
- ▶ How to rectify data so they are comparable to model output?
- ▶ How to compare?

Service-oriented architecture allows for remote computation. Create building-block services to push as much computation as possible to data site, only move results.

Requires a paradigm that does not attempt to separate analysis from access.



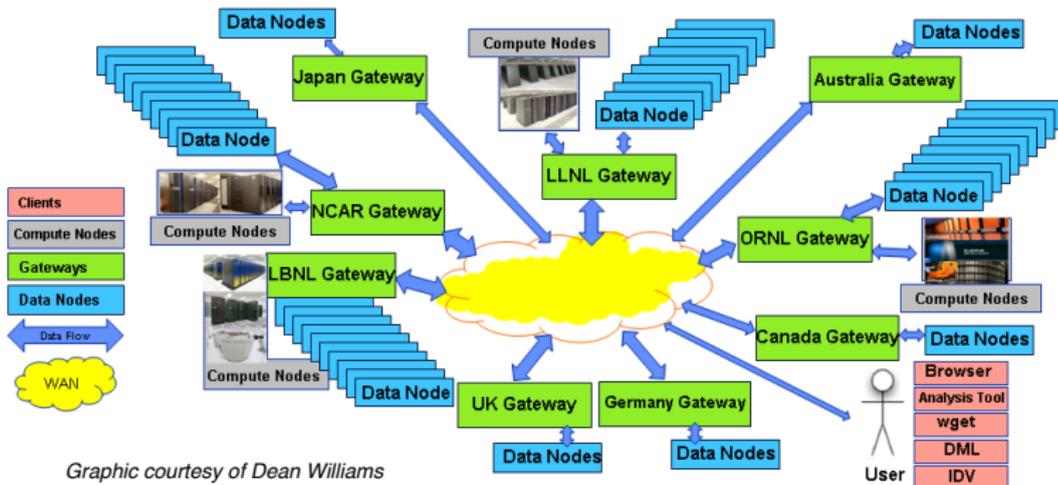
- ▶ User program must encode all functionality beyond gross-level access.
- ▶ Requires knowledge of specific instrument characteristics including retrieval methods, format, measurement errors and biases, etc.
- ▶ Difficulties multiply with more than one data source.



- ▶ Push as much computation as possible to locations where data reside; minimize movement of data.
- ▶ *How to “choreograph” data analysis to take advantage of this?*
- ▶ How does the network topology defined by the system architecture constrain data analysis?
- ▶ How do data analysis objectives constrain the architecture?

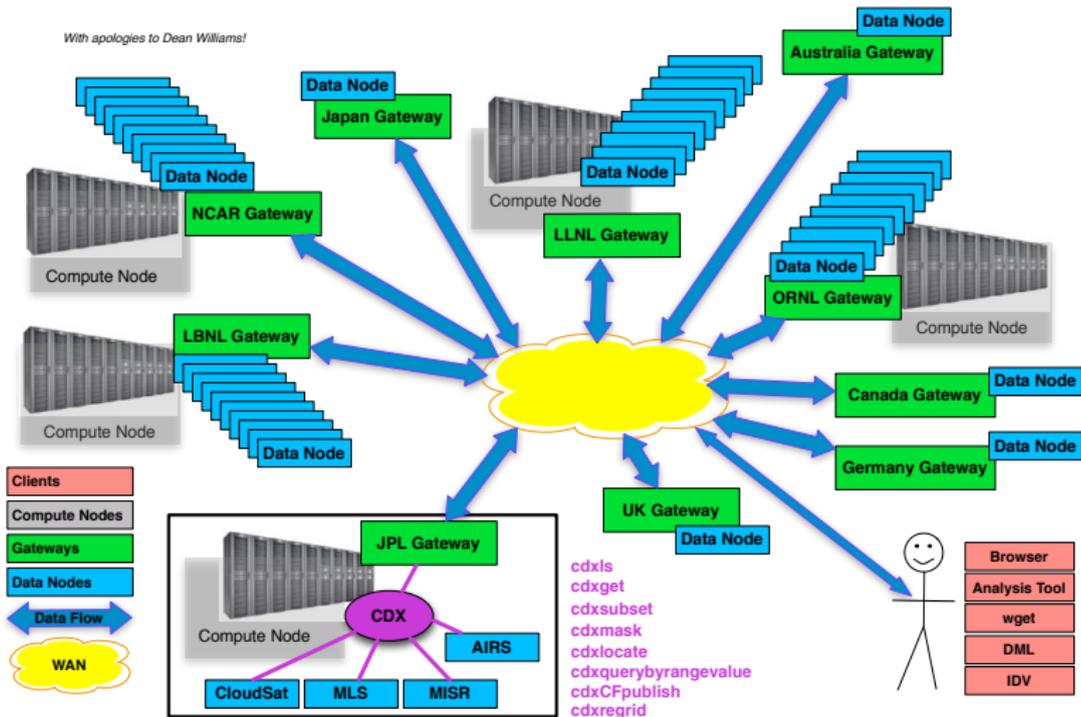


- ▶ Independent gateways federating metadata, users.
- ▶ Individual data nodes responsible for publishing services.
- ▶ Designed for model output data sets.





The Next-generation ESG with Observations





- ▶ What architectural design produces the most efficient system topology for the types of data movement that will be required given scientific objectives?
- ▶ How do we design computational methods that exploit the system topology and its distributed nature? Need algorithms for distributed versions of statistics of interest.
- ▶ How to design the hierarchy of web services (building-blocks) for distributed computation to provide maximum flexibility and utility to scientists?
 - ▶ Bottom of hierarchy is simple access (list, get, subset, find, etc.)
 - ▶ Next level is simple statistics (means, variances, correlations, etc.) with flexible arguments and filtering.
 - ▶ More complex functions (e.g., create a time series, regrid, interpolate, match-up, etc.)
- ▶ Data analysis choreography: how to assemble the building blocks most efficiently given a set of analysis goals? How to optimize data movement?



- ▶ CDX is a service-oriented architecture built on top of Object Oriented dataTechnology (OODT) to facilitate access to and analysis of massive, distributed remote sensing and model output data sets.
- ▶ CDX is a distributed science analysis environment, **not** a data distribution system.
- ▶ Significant efficiencies are achieved by embedding data access functions of appropriate granularity within data analysis functions. Instrument team knowledge of data encoded directly.
- ▶ In this paradigm, data sets are not static entities. They are virtual, possibly streaming data structures flowing across the internet, manipulated and combined on-the-fly as necessary for specific analyses.
- ▶ Flexibility is key: users must be able to revert to lower-level data access functions (e.g., getting whole files) if they don't like the way higher-level functions work.

Copyright 2009, California Institute of Technology. Government sponsorship acknowledged.