

An Architecture and Analysis Environment for Model to Observational Data Intercomparisons

Chris Mattmann¹, Amy Braverman¹, Dan Crichton¹, and Dean Williams²

¹Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109 USA
{mattmann,ajb,crichton}@jpl.nasa.gov

²Program for Climate Model Diagnosis and Intercomparison
Lawrence Livermore National Laboratory
Livermore, CA 94550, USA
williams13@llnl.gov

December 2, 2009

The Jet Propulsion Laboratory (JPL) has within the last year initiated an effort to increase the use of its observational data in the improvement and analysis of climate model outputs. This effort, known as the Climate Data eXchange (CDX), is a multi-institutional collaboration involving representatives from JPL and from the Program for Climate Model Diagnosis and Intercomparisons (PCMDI) at Lawrence Livermore National Laboratory (LLNL).

Our early focus in the context of CDX has been on NASA Level 2 observational data products. These products vary in a number of ways incl.: (1) *format* - many of the products are stored in the Hierarchical Data Format (HDF), others in netCDF, with variation even between software versions that generated these output files within the same format; (2) *geographic distribution* - most observational data products are co-located with their scientific discipline expertise, to increase the yield of promising scientific results and to cut down on the effort for a science user to make progress; (3) *data access mechanism* - some data products are available from sophisticated web service interfaces, e.g., OPeNDAP – others are not, requiring a user to fill on an online web ordering “cart”, and have an email notification indicating availability at a later date; and (4) *size* - depending on the frequency of the instrument’s orbit, and the characteristics of the mission including the way that the instrument “sees” the Earth, the sheer volume of the Level 2 data can widely vary, ranging

from megabytes (MB) per product, to gigabytes (GB). These four dimensions are just a sampling of the characteristics of Level 2 observational data.

The goal of CDX is to deliver an open source software toolkit that allows science users to alleviate as much of the complexity of dealing with Level 2 observational data as possible, and to facilitate its comparison to model outputs. In this fashion, there are two fundamental subsystems within CDX: (1) a *Client Toolkit* – an easy to install software package, providing a set of CDX-enabled familiar commands (ls, get, subset, find, mask, locate) for science users to effectively integrate into their Python, IDL and Matlab environments; and (2) a *CDX Gateway* web service package that is deployed at each data “node” that is made part of the CDX system. Gateway services implement the server portion of the functionality needed by the client toolkit commands.

To date, we have successfully leveraged our early prototypes in subsetting, accessing and using NCAR CCSM model output data from PCMDI in time series generations involving AIRS Level 2 and 3 observational data sets. CDX has increased our scientists’ efficiency by ten-fold (decreasing data access times from nearly 9 hours to 5 minutes in the generation of a time series comparison of a month’s worth of AIRS data to its NCAR CCSM model output counterpart). Our future work on CDX includes adding even more observational datasets (currently AIRS, MLS, CloudSAT, MISR and LLNL model outputs are supported) and creating more value-added tools for scientists to perform sophisticated distributed data analyses.