CAPT Datasets

a case for higher-level data management

Mike Fiorino
January 14, 2005

0.0 Bottom-line Recommendations

1) process and organize model and obs data to a common standard in the same way obs were setup for amip2 as defined in:

http://www-pcmdi.llnl.gov/old/obs/pods/

2) bring the ARM continuous forcing data set to level 2

1.0 Introduction

The technical challenge of model-observation intercomparison is to form model-obs differences with minimal data transformations. Minimizing the data transformations helps preserve the closeness of the data measurement to its source.

In the CAPT project, the observational data come from the high-quality meteorological and radiation sensors of the ARM project and the model data come from climate model solutions (and associated physical processes from the model parameterizations) initialized with analyses that are close or representative of to the atmospheric conditions under which the ARM observations were taken. As the sources of the model (theory) and obs (reality) data are very different, care must be taken in reconciling the data differences – an "apples v apples" problem, for in the end, the quality of the science derived from the model-obs comparison is wholly dependent on consistent, efficient and documented data transformations.

The purpose of this note is to review the state of the data and transformations in the CAPT project against a 5-level data state model. It is recommended that both model and obs be transformed to a common level/state that follows a recognized convention for metadata and variables. Data at this common level can then be transformed, in a variety of ways, for meaningful model-obs differences. Essentially, it is argued that a higher level of data management is needed.

2.0 Data States

The model-obs comparison problem can be viewed as data transformation, i.e., to change heterogeneous data into a common or homogeneous form. We use a 5-level or 5 "state" model to characterize CAPT data. At the highest level, the model-obs data have been

analyzed to diagnostics or plots from which scientific conclusions are made. Regardless of starting point, the data must go through all five levels to produce science.

The table below describes the 5-level model for both observational and model data

5-level Data State Model

OBS	MODEL
Level 0 – Data Level - source or base data	
• calibrated data from sensors • conversion to geophysical variables	• solution in memory
Level 1 – Analysis Level - QC-analysis processed	
• QC of level-0 obs (throw out bad data) • an analysis process that puts the observations onto a constant-increment (generally) space-time grid • physical/dynamical consistency (e.g., variational analysis enforcing mass conservation)	• output of grids, MOLTS • in-model, post-processing, e.g., conversion to pressure levels, staggered to unstaggered grids
Level 2 – Intercompare Level – common metadata, following a convention	
• common variables • unit conversion • content-wise identical to level 1 • level 1->2 is lossless and reversible	• base time for time axis • variables (names, units)
Level 3 – Merge Level – Obs-Model technically identical	
• transform ob to model space • identical space-time coordinates (dimensions or axes)	• observation or forward modeling to move model to observation • typical observation model is interpolation • forward models to retrieve satellite brightness temperature for model T, q
Level 4 – Science Level – diagnostics, analysis	

• plots of differences • statistics	• final data state; the basis of scientific statements and research conclusions

The main point is that the model-obs data exist in each of these states. The issue is how the transformations are made and whether the data in each state are saved as files.

A separate doc reviews specific data sets (capt.dataset.anal.20050120.txt), but the main findings are:

1. Most ARM data are at level 1, some (e.g., TWP soundings) are below level 1 and close to being raw obs.
2. time models are all different and some not even CF compliant (e.g., 'day since' v 'days since' (correct)
3. multiple versions of same data, especially ARSCL and IOPs
4. little documentation on the source of individual data sets in /pcmdi/capt1/dat/ob/arm (I forced organization by directory)
5. inconsistent variable names
6. units of some variables not defined (e.g., ARSCL)
7. different vertical coordinates for ARSCL (z in TWP and p in SGP), both are needed since interpolating z-p for cloud fields can filter out sharp gradients
8. discrepancies between original level 1 (e.g., .txt, tarballs) and rewrite in the CF format (.nc)
9. change in some ARM continuous forcing data files (I use a mirroring utility (wget) to detect changes in source data on the ARM site and /pcmdi/capt1)
10. CAM MOLTS use descriptive dimension variable names which are not CF and impossible to write general access code
11. land output (grids in the "land" directory) in the CAM MOLTS has large time gaps

The problem with this level of disorganization is that each science-making analysis process (transforming from level 1 – level 4) is one off. It is even unclear if different investigators are using the same data for the same obs.

3.0 Conclusion and Recommendations

The CAPT obs data sets I examined are essentially unmanaged and highly inconsistent. Data transformation is a project cost that must be high given the state of these data sets. The question is not whether, but how much the project pays for reaching level 4 (science) results.

The success of the PCMDI Observational Data Set (http://www-pcmdi.llnl.gov/old/obs/pods/) for AMIP? is how it moved all disparate source obs data to level 2 (and even level 3) to greatly improved the efficiency in forming model-obs

differences. A similar development is needed for CAPT; the continuous forcing data would be a good starting point.