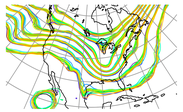

Extending the CAM6+DART Reanalysis Past 2020, for use in CLM, POP, and CICE Data Assimilation, CAM Model Development, and Machine Learning

K. Raeder[1], **J. Anderson**[1], **M. Gharamti**[1], **B. Johnson**[1]

1 National Center for Atmospheric Research; CISL/DAReS, Boulder, CO;

dart@ucar.edu

CISL Computational & Information Systems Lab

Data Assimilation Research Testbed

1. Overview

The CAM6+DART Reanalysis is being extended from the original span of 2011-2019 through 2021. The publicly available data products can be used for

+ identifying biases in CAM6-FV,
+ forcing (ensemble) hindcasts using CESM surface models,
+ ensemble data assimilation experiments using CLM, POP, CICE,...
+ machine learning research applied to the atmosphere,
+ starting (ensemble) CAM6 and coupled forecasts from reanalyses with no external models biases.
+ comparing observation sets commonly used for model validation,
+ investigating the response of CLM plant growth variables,
+ uncertainty quantification in CAM6 constrained by observations. to a realistic range of atmospheric forcing.

Several examples are described here and several more in other CESM Workshop presentations:

+ Brett Raczka: "Applying the Data Assimilation Research Testbed (DART) towards improved CLM simulations of Earth System Carbon: Water and Energy Cycling" LMWG Tues. 12:00
+ Dan Amrhein: poster about POP+DART results and efforts with MOM6 and JEDI forward operators. Tues. 3:20-4:30.

2. CAM Reanalysis

The Data Assimilation Research Testbed (DART) is a community facility for ensemble data assimilation developed and maintained at NCAR. Data assimilation combines short model hindcasts with observations to produce an ensemble description of the state of the Earth system. This description is a balance between the observations and the model formulations, which takes into account the uncertainties in each. This process of confronting the model with observations facilitates model evaluation and improvement.

This 80-member DART reanalysis using CESM2.1's version of the Community Atmospheric Model (CAM6-FV) at 1° resolution will span 2011-2021. The compset = HIST CAM60 CLM50%BGC-CROP CICE%PRES DOCN%DOM MOSART SGLC SWAV.

The data ocean is daily, $1/4^\circ$ AVHRR. Every 6 hours we assimilate up to a million observations; most conventional atmospheric observations, temperature from AIRS, and atmospheric refractivity ($\approx$density) from GPS satellites (Figure 1).

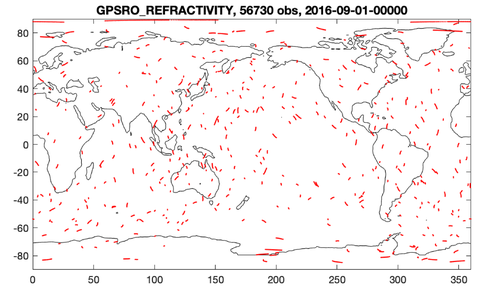

Figure 1: *The horizontal locations of the refractivity observations from a variety of Global Positioning System receivers (e.g. COSMIC1) in a 6 hour span centered on the date shown. This time has 56,730 of these observations.*

The observations directly change the CAM+DART "state vector" {PS, T, US, VS, Q, CLDLIQ, CLDICE} and indirectly change all of CAM's other variables and the active component models; CLM, CICE, MOSART. CAM's "other variables" include the surface fluxes which are passed to the other components (e.g. Figure 4).

3. Evaluation of the Reanalysis

An essential part of the Reanalysis was an evaluation of the performance by comparing CAM's estimates of the observations with the actual observations after every month of assimilation. Our evaluation used selected, broad regions, included statistics such as RMSE and bias of the reanalysis with respect to each of the observation types, and included vertical profiles of all of the statistics.

It includes "total spread", which is a combination of the uncertainties of the model (ensemble spread) and the observations (observation error estimates). This is useful as a measuring stick to judge whether the difference between the reanalysis and observations is meaningful. For example, if the RMSE is statistically indistinguishable from the total spread, then the model and observations are consistent with each other and the assimilation is performing as hoped. The spread in the hindcast or assimilation ensembles varies with location, time, and model variable, and is an objective estimate of the uncertainty. The monthly diagnostic pictures are included in the datasets. They can be used to judge the suitability of the reanalysis for a chosen application. DART has software for making similar evaluations, based on users' needs.

Examples are shown in Figures 2 and 3 for the month when observations from airplanes disappeared due to COVID-19. Note the larger RMSE, totalspread, and bias in the last 10 days, even for independent observations such as AIRS satellite retrievals. The loss of these observations degraded the assimilation in a measurable way.

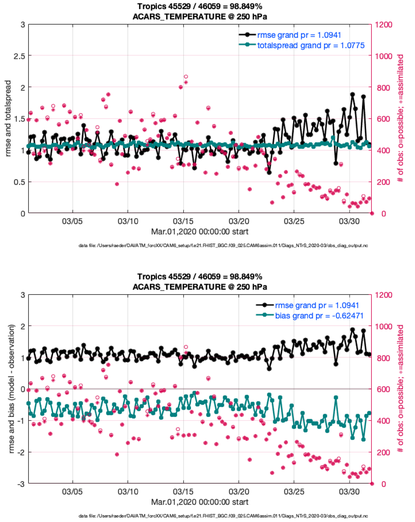

Figure 2: *Root mean square error (black), total spread (green, top), and bias (green, bottom) of the reanalysis observation estimates relative to the actual observations in the Tropics ($|lat| < 20$) for March 2020. The red circles mark the number of available observations (right axis) and the red * mark the number used. The title also lists the total observations in the month ("available/used = %"), the observation type, and atmospheric layer.*

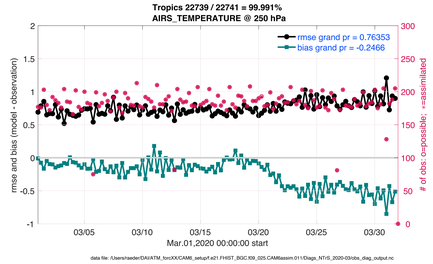

Figure 3: *Similar to Figure 2, but for the bias relative to AIRS temperature observations*

4. Data Products Useful for Surface Component Models

The observation space diagnostics (above) and the datasets described below are freely available in NCAR's Research Data Archive ds345.0 (compressed) and in /glade/collections/rda/data/ds345.0/cpl unzipped (uncompressed).

The most relevant dataset for surface model hindcasts is the ensembles of atmospheric forcing. They have CAM-FV's 1° horizontal resolution and hourly, 3-hourly, and 6-hourly frequencies. The ensemble represents the uncertainty in the knowledge of the atmosphere; each member is equally valid. Data assimilation using the surface components requires the variability in the atmospheric forcing, which maintains the ensemble spread in the assimilation process. (See Figure 4).

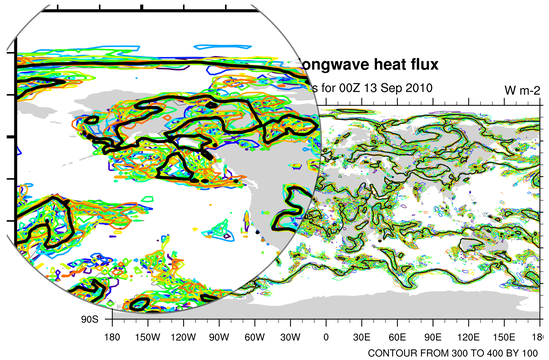

Figure 4: *The variability of the downward longwave heat flux at the surface. The ensemble mean (black curve) and 20 of the 80 members (rainbow colors) are shown here.*

These forcings are in the form of CESM's "data atmosphere" DATM coupler history files, which can be easily used with all of the other CESM components; CICE, CISM, CLM(CTSM), POP(MOM), ..., They are well suited for ensemble forecasts, sensitivity studies, ensemble data assimilation, and model parameter tuning efforts. They are packaged into year-long files, one member per file, for convenient use. The forcing files contain the variables which are written when histaux {a2x3hr, a2x24hr, a2x1hri, a2x1hr, r2x} = .true., which were the settings used in the J1850G CISM spinup case, plus river forcing of other components. J compsets have all components active, except for the atmosphere.

A third dataset is an 80 member ensemble of restart files for CICE, CAM, CLM, and MOSART, which is available weekly. They are consistent with CAM's representation of the actual atmospheric states. These can be used as initial conditions for hindcasts of historical environmental events. One example is the "kiloCAM" experiments being run on KAUST's Shaheen computer by M. Gharamti and B. Johnson. The 80 member CAM6+DART restart file sets are the basis for defining initial ensembles of 250, 500, and 1000 members of 1° CAM6, to investigate the benefits of large ensembles in global scale data assimilation.

5. Using the DATM forcing Files

Figure 5 shows some of the data motion in an assimilation which uses DATM files to force POP. Currently CLM and CICE can be substituted for POP in that figure because there are interfaces to DART for those components. If the DART branch of the figure is left off, then it represents an ensemble hindcast with no assimilation. Then any component (except CAM) can replace POP. Finally, the ensemble size can be 1, which is just a traditional hindcast forced by a single data atmosphere.

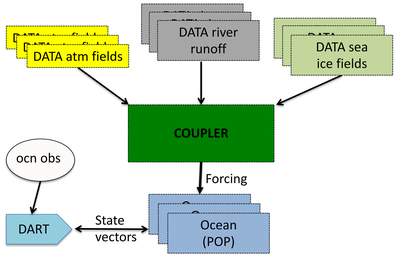

Figure 5: *CESM's coupler, run in multidriver mode, ingests an ensemble of DATM forcing files and passes the data to the active components which need it.*

High resolution (0.1°) POP2+DART will be used for early testing of NCAR's next supercomputer "Derecho" through the Advanced Scientific Discovery project "Enhancing Earth System Predictability by Diagnosing Model Error in Mode Water Regions" Ben Johnson (CISL), Moha Gharamti (CISL), Anna-Lena Deppenmeier (CGD), Ian Grooms (CU-Boulder). The 80 member POP2 ensemble will be forced by the CAM6+DART Reanalysis forcing files.

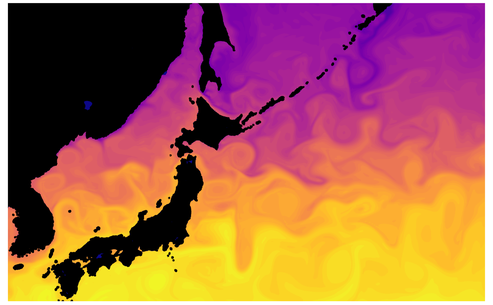

Figure 6: *High resolution POP2 interfaced to DART can reproduce the Kuroshio current (B. Johnson, NCAR;DAReS)*

6. Machine Learning

The observation files used in Section *Evaluation of the Reanalysis* contain the original observations, the Reanalysis ensemble estimates of them, and quality control labels, plus other metadata, for each observation. There are O(10^{10}) observations spanning the most recent decade at high frequency (actual observations times; not restricted to assimilation times). We're not aware of any similar set of observations freely available to researchers in atmospheric science.

7. Plans

- Finish the extension of the Reanalysis through 2021.
- Collaborate with people in the CESM surface model communities to use the data for:
  - recreating recent historical conditions in CESM,
  - model improvement through identification of biases,
  - investigating model response to variability in the forcing.
- Evaluate the feasibility and usefulness of developing an interface between CISM and DART.
- Exploit the 10^{10} labeled observations for machine learning research

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977. We would like to acknowledge high-performance computing support from Cheyenne (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.