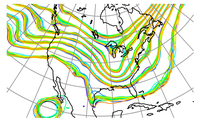


Ensemble Data Assimilation for Very Large Atmosphere, Ocean and Coupled Models with the Data Assimilation Research Testbed

K. Raeder, A. Karspeck, G. Romine, N. Collins, J. Hendricks, T. Hoar, J. Anderson, H. Kershaw, S. Karol, C. Schwartz, R. Sobash

National Center for Atmospheric Research
Boulder, Colorado, USA
dart@ucar.edu

1. DART 2.0 Highlights

- ★ Able to handle much larger model states. This is needed for higher-resolution and/or strongly-coupled DA with multiple components. Distributing the model state across all tasks during the entire filter run means no single task must store the entire state at any time.
- ★ One-sided MPI communication allows tasks to request remote data items from other tasks without interrupting their execution or arranging which data items will be needed in advance.
- ★ Computing the forward operators for all ensemble members at the same time leads to code that vectorizes better.
- ★ Native netCDF support eliminates a conversion step that translated between netCDF model files and a DART binary format file. This also reduces the high-water mark for disk requirements.
- ★ Ensemble data can be read and distributed across all tasks on a variable-by-variable basis, reducing the maximum memory requirements.
- ★ Diagnostic state space files are now written in parallel with state-space restart files, resulting in faster I/O and lower memory requirements.
- ★ Support for externally-computed forward observation operators.
- ★ Support for per-observation-type localization radii.

2. DART is ...

The Data Assimilation Research Testbed (DART) is an open source community software facility for ensemble data assimilation developed at the National Center for Atmospheric Research (NCAR). DART works with a wide variety of climate and weather models and observations and has been free and publically available for more than 10 years. Building an interface between DART and a new model does not require an adjoint and generally requires no modifications to the model code. DART works with dozens of models of varying complexity, including (but not limited to):

- weather models, e.g. WRF, COAMPS, COSMO, MPAS Atmosphere,
- components of climate models, e.g. CAM, POP, CLM, WACCM, MPAS Ocean, MITgcm-Ocean, GCCOM, ROMS, JULES, FESOM, CICE5,
- atmospheric chemistry models, e.g. CAM-CHEM, WRF-CHEM,
- ionosphere/thermosphere models, e.g. TIEGCM, GITM,
- low-order and simple research models

DART assimilates a wide variety of observation types including:

- temperature, winds, moisture from NCEP, MADIS, and SSEC,
- total precipitable water, radar observations, radio occultation observations from GPS satellites,
- ocean temperature and salinity from the World Ocean Database,
- land observations such as snow cover fraction, ground water depth, tower fluxes, cosmic ray neutron intensity, and microwave brightness temperature observations.

DART provides both state-of-the-art ensemble data assimilation capabilities and an interactive educational platform to researchers and students.

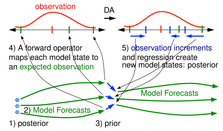


Figure 1: *Schematic for a toy ensemble size of 3.*

J. Anderson, T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Arellano, 2009: *The Data Assimilation Research Testbed: A Community Data Assimilation Facility, BAMS* **90** No. 9 pp. 1283–1296

http://www.image.ucar.edu/DAReS/DART has information about how to download DART, the DART educational materials, and how to contact us.

3. Acknowledgments

The National Center for Atmospheric Research is sponsored by the National Science Foundation. Some computational resources were provided by the Computational and Information Systems Laboratory at NCAR.

4. Motivation for DART 2.0

4.1 Data Layout

One of the challenges for assimilation with truly massive geophysical models is the data layout. Here is a simple example using 4 ensemble members and 4 processing elements (PEs). Each PE is a separate color, a representative observation location is represented by the red star, and some hypothetical area expected to be influenced by the observation is outlined in red.

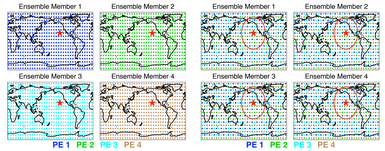


Figure 2: *Two possibilities for data layouts. Left: 'Whole State'. Each ensemble member exists exclusively on a single processing element (PE). Right: 'Distributed'. Each ensemble member is randomly distributed over multiple PEs. Both layouts have advantages and disadvantages and an expensive data transpose is required to go between layouts.*

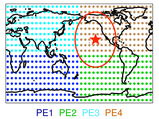
Both data layouts were used in the first version of DART: one for forward operators and one for the assimilation phase.

4.1.1 Characteristics of Whole State Layout

- If each PE has a complete state, forward operators require no communication.
- Forward operators for all ensemble members are computed in parallel.
- **However**, it is not memory scalable and cannot effectively exploit PE counts greater than the ensemble size.
- **Additionally**, newer architectures have large numbers of PEs with less memory per PE.

4.1.2 Characteristics of the Distributed Layout

- Each processing element (PE) stores all ensemble copies of a subset of the model state.
- All ensemble members for a state variable are on one PE.
- Can compute state mean, variance without communication.
- Forward operators probably require communication.
- All increments are computed in parallel.
- One PE broadcasts observation increments.
- Models frequently use a 'block' style layout as shown here. **However**, load balancing is an issue. PE4 has lots of work, PE1 has none. 'Randomly' distributing the ensemble states is likely to result in a well load-balanced system at the expense of increased communication to compute the forward operators.



Anderson, J. and N. Collins, 2007: *Scalable Implementations of Ensemble Filter Algorithms for Data Assimilation. J. Atmos. Oceanic Technol.*, **24**, pp. 1452-1463, doi: 10.1175/JTECH2049.1

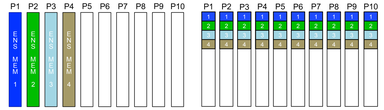


Figure 3: *Two data layouts. Left: whole state vectors are collected on each PE, up to the number of ensemble members. Right: parts of each of the ensemble of state vectors are distributed across all PEs. The more uniform use of memory as in the right panel is a key to efficient computation of forward operators in DART 2.0.*

5. MPI2 one-sided communication

The previous implementation of DART required an all-to-all data transpose of the model state. DART 2.0 replaces local array access with off-core memory retrieval using passive target MPI2 one-sided communication. This removes a previous hard limit on the maximum problem size. The use of one-sided communication allows any task to request data from another task on demand without needing to sychronize.

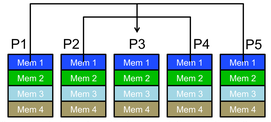


Figure 4: *P3 needs data from other PEs for computations. With MPI2 one-sided communication the other processes are not interrupted. By using more PEs each PE has fewer observations to manage. The forward operator for all ensemble members is computed at once so the code vectorizes better.*

5.0.1 Impact to Code

Most user-visible changes look like the following excerpt from the lorenz_96/model_mod.f90. Note that the new operation *vectorizes*.

Old: (the ensemble state is in array 'x')

```
obs_val = (1.0 - lctnfrac) * x(lower_index) + lctnfrac * x(upper_index)
```

New:

```
x_lower(:) = get_state(lower_index, state_handle)
x_upper(:) = get_state(upper_index, state_handle)
expected_obs(:) = (1.0 - lctnfrac) * x_lower(:) + lctnfrac * x_upper(:)
```

6. Performance of DART 2.0

A series of tests were performed on NCAR's Yellowstone with the Weather Research and Forecasting model (WRF) with 50 ensemble members and 54000 observations. Each ensemble member has a state of 184 million elements, which is approximately 74 GB. Yellowstone has Intel Sandy Bridge processors, a Mellanox InfiniBand, 25 GB of user-usable memory per node, and 8 dual-socket cores for 16 CPUs per node.

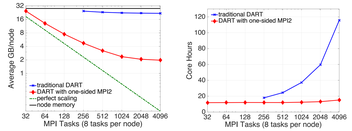


Figure 5: *Left: strong memory scaling for DART 2.0 for a large WRF model. Right: Job core-hour cost almost flat as number of nodes increases. Adding more cores results in a corresponding decrease in wall-clock time.*

6.1 Native netCDF support for I/O

DART 2.0 directly reads and writes netCDF model restart files. Since netCDF is such a common format, this eliminates the need for most models to convert model restarts to a DART format. This reduces the complexity of the cycling scripting and reduces the high-water amount of disk storage required. Models that do not use netCDF will still need conversion routines. Diagnostic files have always been in netCDF and will continue to be so. Ancillary files for things like inflation values are also now in netCDF rather than the old DART restart file format.

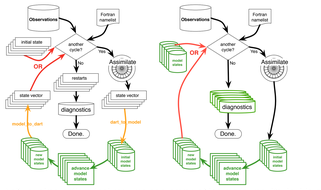


Figure 6: *Left: existing pattern that requires conversion of the model input/output to an intermediate DART format. Right: new pattern for models that use netCDF files for I/O. There is no need for specific DART restart files.*

7. High Resolution POP2

7.1 Background

The 0.1° POP2 model is designed to resolve ocean eddies, which are considerably smaller than the typical resolution of 1.0°.

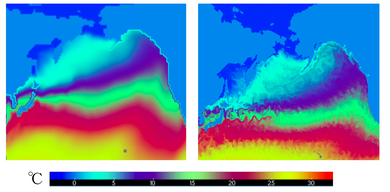


Figure 7: *Comparison of 1° (left) and 0.1° (right) POP output reveals the necessity of high resolution for resolving ocean eddies. These pictures show just the North Pacific portion of the global ocean state.*

7.2 Memory and Input/Output

The higher resolution results in the state vector increasing to more than 17 Gb. Combined with static data and the rest of the program, this does not fit in the memory of a yellowstone node. Each ensemble member must be distributed across multiple nodes. One-sided MPI2 communication makes this possible without creating excessive communication demands. Note the improved memory scaling in the left side of Figure 5. Currently the static memory is not distributed, so there are redundant copies on all the processors. This accounts for much of the flattening of the "DART with one-sided MPI2" curve in Figure 5.
In the previous DART, the entire ensemble plus additional products, such as the ensemble mean and spread, were written to a single NetCDF file. Gathering that data to a single processor is impossible for a model like the 0.1° POP, so DART2 writes out the ensemble members and other DART diagnostic output in parallel to multiple NetCDF files. Even with this change, the I/O for the 0.1° requires 70% of the assimilation time, while for the 1.0° it takes only 10%. These numbers vary with the number of processors and observations used.

8. Real Time WRF-ARW

8.1 WRF+DART Forecast System

NCAR has maintained a real-time, continuously cycled, ensemble data assimilation system since mid-March of 2015. This system currently includes an 80-member ensemble analysis that is updated every 6 hours with conventional observations on a mesoscale grid (Figure 8). The WRF model is used to advance the analysis state each cycle, and has been configured following testing of physics suites to find the combination with minimal systematic model bias as identified in prior analysis statistics. The assimilations are performed with DART configured as an ensemble adjustment Kalman filter. This analysis system demonstrates the reliability and performance of the next-generation DART2 (RMA branch) toolkit .

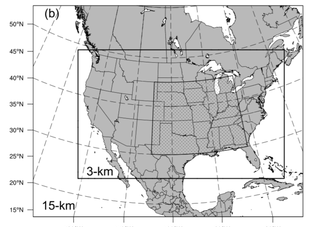


Figure 8: *The WRF+DART assimilation system uses the outer, 15-km grid, which covers areas beyond the conterminous US (CONUS). The ensemble forecasts, which start from the 15-km analyses, use the 3-km grid over the smaller, CONUS region.*

8.2 Ensemble Forecasts

The mesoscale analyses are used once daily (at 00 UTC) to initialize a 10-member convection-permitting (CP; 3-km horizontal grid spacing) ensemble forecast over CONUS. Forecasts are integrated for 48h with results posted to the web (http://ensemble.ucar.edu) where several novel approaches are demonstrated to convey probabilistic forecast information. For this project we also use DART for point based observation verification of the CP ensemble forecasts to investigate model error characteristics across a range of flow regimes. Collectively, this project demonstrates an ensemble analysis and forecast system design with a singular model core and physics suite that follows a consistent approach for forecast verification.
Evaluation of the NCAR ensemble system performance has motivated several additional studies including understanding model error characteristics and predictability of weather extreme events. The first example shows analysis increments averaged over a summer month to illustrate systematic model bias in surface temperature and moisture, particularly from the Plains and across the southern states (Fig. 9). Here, model surface conditions are too warm and dry, with assimilation of surface observations leading to cooling and moistening. Next, Fig. 10 shows a time series of threshold exceedance events for a severe storm surrogate (updraft helicity), relative to the observed number of severe storms. During the warm season, a surrogate threshold value of 75 gives about the right frequency of events, but a lower threshold is more appropriate during the cool season.

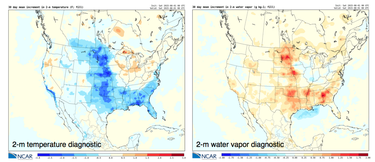


Figure 9: *30-day, mean increment of 2-m temperature (left) and 2-m water vapor produced by the DART assimilations.*

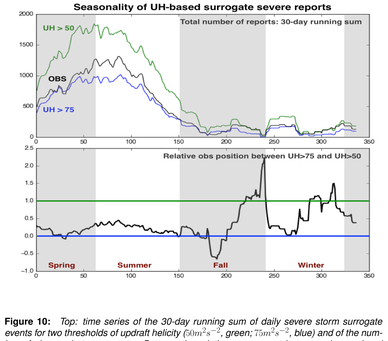


Figure 10: *Top: time series of the 30-day running sum of daily severe storm surrogate events for two thresholds of updraft helicity ($5 m^2 s^{-2}$, green; $75 m^2 s^{-2}$, blue) and of the number of observed storm reports. Bottom: the relative event count between observed and surrogates at the two thresholds: (OBS - UH75)/(UH50 - UH75).*

9. Coupled Assimilation in CESM

9.1 "Weakly Coupled" Configuration

CESMDART is a prototype global coupled ensemble data assimilation system. In-situ ocean and atmosphere data from 1970-1981 (Figure 11) are assimilated in a "weakly coupled" framework using a 30 member ensemble adjustment Kalman filter (Figure 12). The model is run at nominal 1 degree resolution, in a standard CESM "workhorse" configuration.

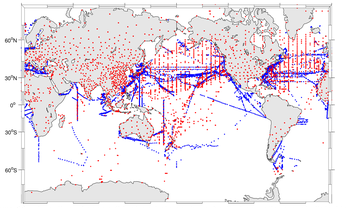


Figure 11: *Location of ocean (blue) and atmosphere (red) observations assimilated in the month of Jan 1975.*

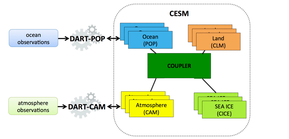


Figure 12: *The CESMDART weakly coupled framework assimilates ocean observations only into the ocean model, and similarly for the atmosphere. No land or sea ice observations are currently assimilated, but all 4 components are affected indirectly by all observations, through the interactions of the components through the CESM coupler during the forecasts.*

9.2 Evaluation of Analyses

Results are promising, indicating that the CESM can be constrained to a historical representation of the climate system.

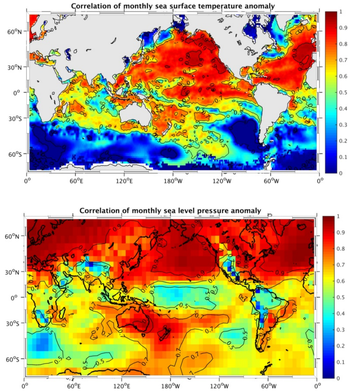


Figure 13: *Anomaly correlation between CESMDART and HADISST (top) and HADSLP (bottom). Agreement is high in regions where observations were available in the 1970s. SLP is completely independent data (no sea level pressure was assimilated). CESMDART does not use any SST products, but draws from the same raw in-situ sea surface temperature data sources sources as HADISST.*