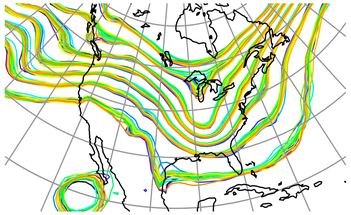


A General Purpose Data Assimilation Facility: DART

J.Anderson, T.Hoar, N.Collins, K.Raeder, H.Liu

National Center for Atmospheric Research,
Institute for Mathematics Applied to Geosciences
Data Assimilation Research Section
dart@ucar.edu

Institute for Mathematics Applied to Geosciences

1. Ensemble Data Assimilation

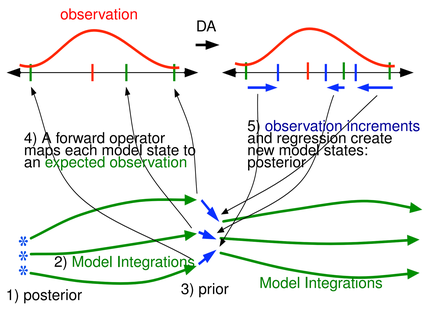
1.1 What is Data Assimilation?

Data Assimilation (DA) combines observations of a physical system with predictions from a numerical forecast model. DA can be used for many purposes, including:

- constructing initial conditions for forecasts,
- evaluating errors in the model and observations,
- finding appropriate values for model parameters,
- designing better observational systems.

The Data Assimilation Research Testbed (DART) is a community software facility that can be used for all the above purposes. DART provides a variety of ensemble filtering (EF) algorithms.

1.2 Sequential Ensemble Filtering



1.3 Geophysical applications require extensions

The basic EF algorithm does not work well when applied to large geophysical problems. Model error, sampling error from using affordable ensemble sizes, and violation of linear and Gaussian assumptions all lead to overconfidence in the ensemble priors. This can result in poor performance or filter divergence. DART has several self-tuning algorithms to address these problems that work for a wide variety of models and observations without the need for user expertise. Some of these are described in sections 3 and 4.

2. What's in DART?

DART makes it easy to learn and apply EF data assimilation.

- Has an extensive tutorial and instruction set.
- Incorporating new models and new observation types requires only minimal coding of a small set of interface routines.
- Scales linearly to hundreds of processors. Parallel performance is independent of the forecast model. Even single-threaded models can be run in parallel.

- Includes many flavors of ensemble filters:
  1. EAKF; Ensemble Adjustment Kalman Filter,
  2. EnKF; Ensemble Kalman Filter,
  3. Kernal filter,
  4. particle filter,
  5. a fixed-lag ensemble Kalman smoother.

- Provides additional algorithms for improved performance:
  1. prior and posterior inflation,
  2. automatic adaptive inflation,
  3. horizontal, vertical, multivariate localization,
  4. hierarchical filter for adaptive localization,
  5. dynamic adjustment of localization cutoff radius,
  6. *a priori* sampling error correction.

- Output is in portable netCDF files and one custom-format observation file. Matlab© scripts are provided to investigate:
  1. rank histograms,
  2. bias, error, and spread as a function of height or time,
  3. ensemble trajectories,
  4. innovations,
  5. 3D plots of observation densities and rejection attributes.

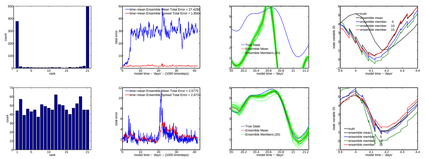


Figure 1: *Examples of some diagnostic plots which can be generated for any DART experiment, any model. These are 'perfect model' experiment results with the Lorenz 96 model. The top row of plots is from an experiment that exhibited filter divergence. The bottom row of plots used covariance inflation.*

- It is freely available via a web download using subversion, which provides for easy code upgrade paths and bug fixes.

Compliant Models and Observation Types

The distributed code includes a variety of low-order models and the following geophysical model interfaces:

1. CAM; Community Atmosphere Model (spectral, FV cores),
2. WRF; Weather Research and Forecasting system,
3. MITgcm Ocean; general circulation model,
4. ROSE; Middle atmosphere dynamics and chemistry,
5. GFDL; grid point GCM dynamical core,
6. Two-layer primitive equation model (NOAA/CDC),
7. PBL_1d; Single column (WRF) model.

Observation types that have been used include:

1. upper air: radiosondes, ACARS, satellite drift winds,
2. surface: winds(10m), T and Q(2m), P_{surf},
3. scatterometer winds,
4. Doppler radial velocity and reflectivity,
5. GPS radio occultation, refractivity,
6. ground-based GPS.

3. Hierarchical Filter for Adaptive Localization

Sampling error from using small ensembles leads to spuriously large correlations among weakly-related observations and state variables. This results in systematic underestimation of posterior variance and can lead to filter divergence. Localizing the impact of an observation to nearby state variables has been the traditional solution. This requires expert knowledge and trial-and-error to get appropriate localizations. DART provides an algorithm to automatically compute localizations using a small group of ensembles.

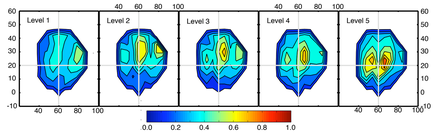


Figure 2: *Adaptive localization for a surface pressure (PS) observation on the zonal wind (V) at 5 model levels as determined from 4 groups of 20 members. The location of the PS observation is indicated by the crosshairs. Note the asymmetry! The model is the GFDL dynamical core.*

4. Adaptive Inflation

Model error and violation of linear and Gaussian assumptions are additional sources of insufficient variance in the ensemble priors. This can be ameliorated by 'inflation': where the ensemble spread is increased while maintaining the mean and sample correlations among all prior variables. Traditionally, all variables at all locations have been 'inflated' by a constant value, chosen by the user to optimize performance in some region or timespan. This tuning takes time and computer resources and can never be optimal for the entire domain. Often, a value of inflation that works well in one region will lead to uncontrolled growth of variance in another. DART has an adaptive inflation algorithm that uses the set of observations affecting a state variable to determine the best value of inflation for that variable.

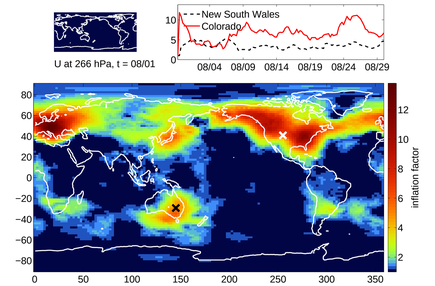


Figure 3: *Illustration of the damped adaptive spatial inflation and its evolution over time at a pair of locations. CAM T85 U winds at level 15 (≈ 266 hPa) at the end of one month of assimilating observations every 6 hours. The field started off with a uniform value of 1.0.*

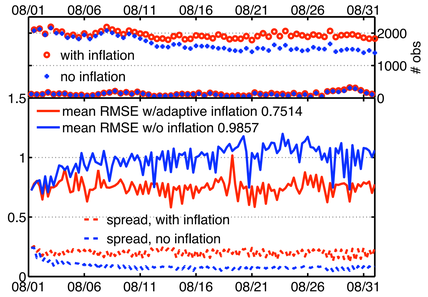


Figure 4: *Six hour forecast ensemble mean RMS error and spread of 500 hPa radiosonde temperature observations for CAM T85 assimilations with no inflation and with damped adaptive inflation. The assimilation with adaptive inflation has reduced RMS error and more consistent spread. The upper panel also shows that fewer observations are being rejected by the assimilation using inflation.*

5. New Observations

The impact of new observation types on predictions of high-impact weather like tropical storms can be assessed using DA. Figure 5 shows forecasts of typhoon Shanshan (2006) initialized from two assimilations that differ only in the use of the global positioning system (GPS) radio occultation measurements from the Constellation Observing System for Meteorology Ionosphere and Climate (COSMIC) satellites. The WRF model in a regional configuration was used for both the assimilations and the forecasts.

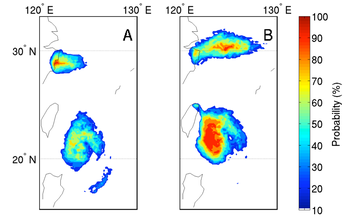


Figure 5: *Forecasts of the probability that rainfall will exceed 60mm during the period of 12Z 14 September to 12Z 15 September 2007. Forecasts were initiated 36 hours before. The assimilations producing the forecasts in panel B made use of COSMIC GPS radio occultation observations while those in panel A did not.*

The 32-member assimilations use a 45-km grid; the first 16 ensemble analyses are interpolated to a 15-km grid to perform 72-hour forecasts. The probability at each grid point is computed by dividing the number of forecast members that predicted excessive precipitation by 16, the total number of forecasts. Ensemble forecasts starting from analyses using COSMIC observations have larger probabilities of excessive precipitation and are more consistent with heavy rainfall generated by Shanshan.

6. Parallel Scaling

Scaling runs have been done with a state-of-the-art global atmospheric climate model on architectures ranging from a commodity Intel-based 32 CPU Linux cluster to Bluefire - an IBM Power 575 with 4064 POWER6^{TM} processors running at 4.7GHz. The following results are from Bluefire, and are representative.

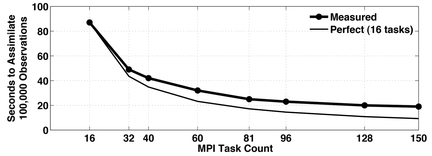


Figure 6: *Wall clock time required to assimilate 100,000 observations as a function of number of MPI tasks. The experiment used 80 ensemble members, each with 2,166,624 state variables. The 'perfect scaling' line is based on the time for 16 MPI tasks.*

The data is from the CAM FV core on a 1.9° x 2.5° grid with 26 vertical levels. In this case, the computational burden of assimilation is about equal to the burden of advancing the model 6 hours. The model advances scale independent of the assimilation; DART can advance ensemble members simultaneously or sequentially.

7. Try this at home!

Our web site is: **http://www.image.ucar.edu/DAReS/DART**
There you will find information about how to download the latest version of DART from our subversion server, information on a full DART tutorial (included with the distribution), and contact information for the DART development group.

References

[1] Anderson, J., 2008 *Spatially and temporally varying adaptive covariance inflation for ensemble filters. Tellus A,*, doi:10.1111/j.1600-0870.2008.00361.x

[2] Anderson, J., 2007 *An adaptive covariance inflation error correction algorithm for ensemble filters. Tellus A,*, **59 (2)**, 210-224, doi: 10.1111/j.1600-0870.2006.00216.x

[3] Anderson, J., 2007 *Exploring the need for localization in ensemble data assimilation using an heirarchical ensemble filter. Physica D*, **230**, 99-111, doi:10.1016/j.physd.2006.02.011

[4] Anderson, J., Collins, N., 2007, *Scalable Implementations of Ensemble Filter Algorithms for Data Assimilation. Journal of Atmospheric and Oceanic Technology*, **24** 1452-1463, doi: 10.1175/JTECH2049.1

[5] Anderson, J., 2003 *A local least squares framework for ensemble filtering. Monthly Weather Review*, **131**, 634-642, doi:10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2