An Introduction to Ensemble Data Assimilation and the Data Assimilation Research Testbed (DART)

Jeffrey Anderson, Nancy Collins, Tim Hoar, Hui Liu, Glen Romine, Kevin Raeder
NCAR Institute for Math Applied to Geophysics
What is Data Assimilation?

Observations combined with a Model forecast…

…to produce an analysis (best possible estimate).
Example: Estimating the Temperature Outside

An observation has a value (*),

![Graph showing observed temperature is 1°C]
Example: Estimating the Temperature Outside

An observation has a value (\( \ast \)),

and an error distribution (red curve) that is associated with the instrument.
Example: Estimating the Temperature Outside

Thermometer outside measures 1°C.

Instrument builder says thermometer is unbiased with +/- 0.8°C gaussian error.
Example: Estimating the Temperature Outside

Thermometer outside measures 1°C.

The red plot is $P(T \mid T_o)$, probability of temperature given that $T_o$ was observed.
Example: Estimating the Temperature Outside

We also have a prior estimate of temperature.

The green curve is $P(T | C)$; probability of temperature given all available prior information $C$. 

Example: Estimating the Temperature Outside

Prior information $C$ can include:

1. Observations of things besides $T$;
2. Model forecast made using observations at earlier times;
3. \textit{A priori} physical constraints ($T > -273.15'C$);
Combining the Prior Estimate and Observation

Bayes Theorem:

\[ P(T | T_o, C) = \frac{P(T_o | T, C)P(T | C)}{\text{Normalization}} \]

Posterior: Probability of T given observations and Prior. Also called update or analysis.

Prior

Likelihood: Probability that \( T_o \) is observed if T is true value and given prior information C.
Combining the Prior Estimate and Observation

\[ P(T \mid T_o, C) = \frac{P(T_o \mid T, C)P(T \mid C)}{\text{normalization}} \]
Combining the Prior Estimate and Observation

\[ P(T \mid T_o, C) = \frac{P(T_o \mid T, C)P(T \mid C)}{\text{normalization}} \]
Combining the Prior Estimate and Observation

\[ P(T | T_o, \mathcal{C}) = \frac{P(T_o | T, \mathcal{C}) P(T | \mathcal{C})}{\text{normalization}} \]
Combining the Prior Estimate and Observation

\[ P(T \mid T_o, C) = \frac{P(T_o \mid T, C) P(T \mid C)}{\text{normalization}} \]
Combining the Prior Estimate and Observation

$$P(T \mid T_o, C) = \frac{P(T_o \mid T, C)P(T \mid C)}{\text{normalization}}$$
Consistent Color Scheme Throughout Tutorial

Green = Prior
Red = Observation
Blue = Posterior
Black = Truth
Combining the Prior Estimate and Observation

\[ P(T | T_o, C) = \frac{P(T_o | T, C)P(T | C)}{\text{normalization}} \]

Generally no analytic solution for Posterior.
Combining the Prior Estimate and Observation

\[ P(T \mid T_o, C) = \frac{P(T_o \mid T, C)P(T \mid C)}{\text{normalization}} \]

Gaussian Prior and Likelihood -> Gaussian Posterior
Combining the Prior Estimate and Observation

For Gaussian prior and likelihood...

Prior

\[ P(T \mid C) = \text{Normal}(T_p, \sigma_p) \]

Likelihood

\[ P(T_o \mid T, C) = \text{Normal}(T_o, \sigma_o) \]

Then, Posterior

\[ P(T \mid T_o, C) = \text{Normal}(T_u, \sigma_u) \]

\[ \sigma_u = \sqrt{\left( \sigma_p^{-2} + \sigma_o^{-2} \right)^{-1}} \]

With

\[ T_u = \sigma_u^2 \left[ \sigma_p^{-2} T_p + \sigma_o^{-2} T_o \right] \]
What is Ensemble Data Assimilation?

Use an ensemble (set) of model forecasts.

Use sample statistics to get covariance between state and observations.
A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation
A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation

Fit a Gaussian to the sample.
A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation

Get the observation likelihood.
Compute the continuous posterior PDF.
A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation

Use a deterministic algorithm to ‘adjust’ the ensemble.
First, ‘shift’ the ensemble to have the exact mean of the posterior.
First, ‘shift’ the ensemble to have the exact mean of the posterior. Second, linearly contract to have the exact variance of the posterior. Sample statistics are identical to Kalman filter.
Multivariate Ensemble Kalman Filter

So far, we have an observation likelihood for single variable.

Suppose the model prior has additional variables.

Use linear regression to update additional variables.
Ensemble filters: Updating additional prior state variables

Assume that all we know is prior joint distribution.
One variable is observed.
What should happen to the unobserved variable?
Ensemble filters: Updating additional prior state variables

Assume that all we know is prior joint distribution.
One variable is observed.
Compute increments for prior ensemble members of observed variable.
Ensemble filters: Updating additional prior state variables

Assume that all we know is prior joint distribution.

One variable is observed.

Using only increments guarantees that if observation had no impact on observed variable, unobserved variable is unchanged (highly desirable).
Ensemble filters: Updating additional prior state variables

Assume that all we know is prior joint distribution.

How should the unobserved variable be impacted?

First choice: least squares.
Equivalent to linear regression.
Same as assuming binormal prior.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.
How should the unobserved variable be impacted?
First choice: least squares.
Begin by finding least squares fit.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.
Next, regress the observed variable increments onto increments for the unobserved variable.
Equivalent to first finding image of increment in joint space.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.

Next, regress the observed variable increments onto increments for the unobserved variable.

Equivalent to first finding image of increment in joint space.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.

Next, regress the observed variable increments onto increments for the unobserved variable.

Equivalent to first finding image of increment in joint space.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.

Next, regress the observed variable increments onto increments for the unobserved variable.

Equivalent to first finding image of increment in joint space.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.

Next, regress the observed variable increments onto increments for the unobserved variable.

Equivalent to first finding image of increment in joint space.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.
Regression: Equivalent to first finding image of increment in joint space.
Then projecting from joint space onto unobserved priors.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.

Regression: Equivalent to first finding image of increment in joint space.
Then projecting from joint space onto unobserved priors.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.
Regression: Equivalent to first finding image of increment in joint space.
Then projecting from joint space onto unobserved priors.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.
Regression: Equivalent to first finding image of increment in joint space.
Then projecting from joint space onto unobserved priors.
Ensemble filters: Updating additional prior state variables

Have joint prior distribution of two variables.

Regression: Equivalent to first finding image of increment in joint space.

Then projecting from joint space onto unobserved priors.
Ensemble filters: Updating additional prior state variables

Now have an updated (posterior) ensemble for the unobserved variable.
Ensemble filters: Updating additional prior state variables

Now have an updated (posterior) ensemble for the unobserved variable. Fitting Gaussians shows that mean and variance have changed.
Ensemble filters: Updating additional prior state variables

Now have an updated (posterior) ensemble for the unobserved variable.
Fitting Gaussians shows that mean and variance have changed.
Other features of the prior distribution may also have changed.
Ensemble Filter for Large Geophysical Models

1. Use model to advance ensemble (3 members here) to time at which next observation becomes available.
2. Get prior ensemble sample of observation, $y = h(x)$, by applying forward operator $h$ to each ensemble member.

Theory: observations from instruments with uncorrelated errors can be done sequentially.
3. Get observed value and observational error distribution from observing system.
Ensemble Filter for Large Geophysical Models

4. Find the increments for the prior observation ensemble (this is a scalar problem for uncorrelated observation errors).

Note: Difference between various ensemble filters is primarily in observation increment calculation.
5. Use ensemble samples of $y$ and each state variable to linearly regress observation increments onto state variable increments.

Theory: impact of observation increments on each state variable can be handled independently!
Ensemble Filter for Large Geophysical Models

6. When all ensemble members for each state variable are updated, there is a new analysis. Integrate to time of next observation …
Ensemble Filter for Large Geophysical Models

A generic ensemble filter system like DART just needs:

1. A way to make model forecasts;
Ensemble Filter for Large Geophysical Models

A generic ensemble filter system like DART just needs:
1. A way to make model forecasts;
2. A way to compute forward operators, $h$. 
DART is used at:
43 UCAR member universities
More than 100 other sites

Public domain software for ensemble Data Assimilation
  – Well-tested, portable, extensible, free!

Models
  – Toy to HUGE

Observations
  – Real, synthetic, novel

An extensive Tutorial
  – With examples, exercises, explanations

People: The DARES Team

Cornell University March 2012
DART is:

Education

Exploration

Research

Operations
DART works with many geophysical models

**Global Atmosphere models:**

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Institute</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM</td>
<td>Community Atmosphere Model</td>
<td>NCAR</td>
</tr>
<tr>
<td>CAM/CHEM</td>
<td>CAM with Chemistry</td>
<td>NCAR</td>
</tr>
<tr>
<td>WACCM</td>
<td>Whole Atmosphere Community Climate Model</td>
<td>NCAR</td>
</tr>
<tr>
<td>AM2</td>
<td>Atmosphere Model 2</td>
<td>NOAA/GFDL</td>
</tr>
<tr>
<td>NOGAPS</td>
<td>Navy Operational Global Atmospheric Prediction System</td>
<td>US Navy</td>
</tr>
<tr>
<td>ECHAM</td>
<td>European Centre Hamburg Model</td>
<td>Hamburg</td>
</tr>
<tr>
<td>Planet WRF</td>
<td>Global version of WRF</td>
<td>JPL</td>
</tr>
<tr>
<td>MPAS</td>
<td>Model for Prediction Across Scales (under development)</td>
<td>NCAR/DOE</td>
</tr>
</tbody>
</table>
DART works with many geophysical models

### Regional Atmosphere models:

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Operator</th>
</tr>
</thead>
<tbody>
<tr>
<td>WRF/ARW</td>
<td>Weather Research and Forecast Model</td>
<td>NCAR</td>
</tr>
<tr>
<td>WRF/CHEM</td>
<td>WRF with Chemistry</td>
<td>NCAR</td>
</tr>
<tr>
<td>NCOMMAS</td>
<td>Collaborative Model for Multiscale Atmospheric Simulation</td>
<td>NOAA/NSSL</td>
</tr>
<tr>
<td>COAMPS</td>
<td>Coupled Ocean/Atmosphere Mesoscale Prediction System</td>
<td>US Navy</td>
</tr>
<tr>
<td>CMAQ</td>
<td>Community Multi-scale Air Quality</td>
<td>EPA</td>
</tr>
<tr>
<td>COSMO</td>
<td>Consortium for Small-Scale Modeling</td>
<td>DWD</td>
</tr>
</tbody>
</table>
DART works with many geophysical models

**Ocean models:**

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>POP</td>
<td>Parallel Ocean Program</td>
<td>DOE/NCAR</td>
</tr>
<tr>
<td>MIT OGCM</td>
<td>Ocean General Circulation Model</td>
<td>MIT</td>
</tr>
<tr>
<td>ROMS</td>
<td>Regional Ocean Modeling System (under development)</td>
<td>Rutgers</td>
</tr>
<tr>
<td>MPAS</td>
<td>Model for Prediction Across Scales (Under development)</td>
<td>DOE/LANL</td>
</tr>
</tbody>
</table>
DART works with many geophysical models

**Upper Atmosphere/Space Weather models:**

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROSE</td>
<td></td>
<td>NCAR</td>
</tr>
<tr>
<td>TieGCM</td>
<td>Thermosphere Ionosphere</td>
<td>NCAR/HAO</td>
</tr>
<tr>
<td></td>
<td>Electrodynamic GCM</td>
<td></td>
</tr>
<tr>
<td>GITM</td>
<td>Global Ionosphere</td>
<td>Michigan</td>
</tr>
<tr>
<td></td>
<td>Thermosphere Model</td>
<td></td>
</tr>
</tbody>
</table>
DART works with many geophysical models

Land Surface models:

CLM  Community Land Model  NCAR
Focus on DART Science with CAM

Basic Capability: Ensemble Analyses and Forecasts
Works for all CAM versions since 2002

20 of 80 members

6-hour forecast
500 hPa height
18Z 14 Jan 2007

contours from 5400 to 5880 by 80
Ensemble Analyses and Forecasts

Sample collaborations:

Edmund Chang, Stony Brook
Pacific storm track/cyclogenesis

Nedjeljka Zagar, Ljubljana University
Normal mode analysis of general circulation

Maria Tsukernik, Monash/Brown
Antarctic cyclones
Ensemble Analyses and Forecasts

Sample collaborations:

Rahul Mahajan, U. Washington
Real-time ensemble forecasts
For Pacific Northwest.

Ibrahim Hoteit, KAUST Saudi Arabia
Gulf of Mexico Ocean Prediction.

Ryan Torn, SUNY Albany
Real-time Atlantic hurricane forecasts.
Diagnosing and Correcting Errors in the CAM Finite Volume core with DART

Kevin Raeder*
Jeff Anderson*
Peter Lauritzen+
Tim Hoar*

*NCAR/CISL/IMAGe/DAReS
+NCAR/ESSL/CGD/AMPS

Cornell University March 2012
Gridpoint noise detected in CAM/DART analysis

Ensemble Mean V at 266 hPa at 6 hours

CAM FV core - 80 member mean - 00Z 25 September 2006
Suspicions turned to the polar filter (DPF)

CAM FV core - 80 member mean - 00Z 25 September 2006
Continuous polar filter (alt-pft) eliminated noise.
Differences mostly in transition region of default filter.
The use of DART diagnosed a problem that had been unrecognized (or at least undocumented).

Could have an important effect on any physics in which meridional mixing is important.

The problem can be seen in ‘free runs’ - it is not a data assimilation artifact.

Without assimilation, can’t get reproducing occurrences to diagnose.
Cloud response to the 2007 Arctic sea ice loss in CAM3.5 and CAM4

Jennifer E. Kay
National Center for Atmospheric Research (NCAR)
Colorado State University (CSU)
Collaborators: Julienne Stroeve (NSIDC), Andrew Gettelman, Kevin Raeder, Jeff Anderson (NCAR), Graeme Stephens, Tristan L’Ecuyer, Chris O’Dell (CSU)
CAM4’s cloud response to sea ice loss; July 2006 to 2007

24-hour forecasts started from DART/CAM analyses identified erroneous cloud response to disappearing sea ice. Jen Kay found that low clouds were only diagnosed over open water, not ice, and the low cloud scheme should have required a well mixed boundary layer.

Short forecasts with a climate model from analyses, compared against observations, point to model improvements.
Exploring the Impact of Novel Observations:
Impact of COSMIC GPS Observations in Cam Analyses

IONOSPHERE

NEUTRAL ATMOSPHERE

Occulting GPS Satellite

Time Delay & Bending Angle
Provide Density vs. Altitude

Occulting LEO Satellite

EARTH

Cornell University March 2012
Impact of COSMIC GPS Observations in Cam Analyses

Observations 1 December 2006

GPS

ACARS and Aircraft

Radiosondes

Sat Winds

Cornell University March 2012
CAM 6-hour forecast Bias from Radiosonde Specific Humidity (Q) - December 2006
Conclusions

- GPS has significant information, especially about moisture;
- Most important where other observations are sparse;
- Ensemble assimilation can do full multivariate improvement;
- Must carefully consider planning of future obs systems.

- CAM biases can be reduced with GPS observations.
Estimating CO with MOPITT remote sensing observations in CAM/Chem
Ave Arellano, ACD
(now U. Arizona)

Ave extended DART/CAM for CAM/Chem.
Then he added MOPITT observations. Improved fit to aircraft CO obs.

Estimating CO with MOPITT remote sensing observations in CAM/Chem

Ave Arellano, ACD (now U. Arizona)
Estimating CO with MOPITT remote sensing observations in CAM/Chem

Ave Arellano, ACD
(now U. Arizona)

This system was used for real-time support for ARCTAS field campaign.
Moving towards coupled assimilation for earth system models.

Tim Hoar, Nancy Collins, Kevin Raeder, Jeffrey Anderson, NCAR Institute for Math Applied to Geophysics
Data Assimilation Research Section

Steve Yeager, Mariana Vertenstein, Gokhan Danabasoglu, Alicia Karspeck, and Joe Tribbia
NCAR/NESL/CGD/Oceanography

Cornell University March 2012
Ocean Data Assimilation Motivation

- Climate change over time scales of 1 to several decades has been identified as very important for mitigation and infrastructure planning.
- CGD needs ocean initial conditions for the IPCC decadal prediction program.
Hypothesis: Need Ensemble of Atmospheres to Force Ensemble Assimilation for Ocean

• Case 1: 23 POP members forced by a single atmosphere.
• Case 2: 48 POP members forced by 48 CAM/DART analyses.
• Generates additional ocean spread, improved analyses.
Current POP Assimilation

Obs

DART

3D state

3D restart

POP

Coupler

DATM

2D forcing from CAM assimilation

2D forcing

from CAM assimilation

NCAR

NSF

Cornell University March 2012
World Ocean Database T,S observation counts

These counts are for 1998 & 1999 and are representative.

FLOAT_SALINITY  68200
FLOAT_TEMPERATURE  395032
DRIFTER_TEMPERATURE  33963
MOORING_SALINITY  27476
MOORING_TEMPERATURE  623967
BOTTLE_SALINITY  79855
BOTTLE_TEMPERATURE  81488
CTD_SALINITY  328812
CTD_TEMPERATURE  368715
STD_SALINITY  674
STD_TEMPERATURE  677
XCTD_SALINITY  3328
XCTD_TEMPERATURE  5790
MBT_TEMPERATURE  58206
XBT_TEMPERATURE  1093330
APB_TEMPERATURE  580111

- temperature observation error standard deviation == 0.5 K.
- salinity observation error standard deviation  == 0.5 msu.
Ensemble *Spread* for Pacific 100m XBT

Spread of the “climatological” ensemble

Twice as much!

Small spread!
100m Mooring Temperature RMSE – Pacific

POP/CAM as good or better RMSE
Physical Space: 1998/1999 SST Anomaly from HadIOI-SST

Coupled Free Run

POP forced by observed atmosphere (hindcast)

23 POP 1 DATM

3 POP 48 CAM
Fully coupled assimilation will need data from all models at the same time.
March 28 Tornado Outbreak

May 4 (Greensburg, KS) Tornado Case

Cornell University March 2012
Hurricane Katrina Sensitivity Analysis
(Ryan Torn, SUNY Albany)

Contours are ensemble mean 48h forecast of deep-layer mean wind.

Color indicates change in the longitude of Katrina.
DART Includes Many Diagnostic Tools
Observation Visualization Example
DART Includes Many Algorithms to Improve Performance

- Adaptive inflation to maintain spread
- Adaptive localization to reduce computation
- Group filter to design localization
- Sampling error correction to reduce errors

- General parallel implementation
Code to implement all of the algorithms discussed is freely available from:

http://www.image.ucar.edu/DAReS/DART/