Scalable Computing Challenges in Ensemble Data Assimilation

FRCRC Symposium
Nancy Collins
NCAR - IMAGe/DAReS
14 Aug 2013
Overview

• What is Data Assimilation?

• What is DART?

• Current Work on Highly Scalable Systems
What is Data Assimilation?

Mathematical techniques for combining observations of a system with a predictive model of the system to give a better forecast of a future state of the system.
DA and Lorenz Models

• Simpler sets of equations that capture some characteristic of the actual atmosphere or other large chaotic systems
• Can be used to prototype new techniques in Data Assimilation before trying to apply them to a large weather or climate model
• “Lorenz 96” has 40 variables and might represent the air passing around the earth along a latitude circle
Lorenz 96 Free Run
Lorenz 96 Ensembles
Data Assimilation Types

• Variational Systems
  – Used by large operational weather forecasting centers
  – Requires an ‘adjoint’ for any new equations in the model

• Ensemble Systems
  – Uses statistical techniques to adjust the model values
  – Easier for small groups or individual model users

• Hybrids
  – People experimenting with small ensembles inside a variational system
Ensemble Filter For Large Geophysical Models

1. Use model to advance *ensemble* (3 members here) to time at which next observation becomes available.

Ensemble state estimate, $x(t_k)$, after using previous observation (*analysis*)

Ensemble state at time of next observation (*prior*)
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.
Ensemble Filter For Large Geophysical Models

3. Get observed value and observational error distribution from observing system.
Ensemble Filter For Large Geophysical Models

4. Compute 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.
Ensemble Filter For Large Geophysical Models

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 ...

![Diagram of ensemble filter process]
DART: Data Assimilation Research Testbed

• DART software is used for:
  – Building Data Assimilation systems
  – A Teaching tool
  – A DA Research tool

• Users can run it:
  – Out of the box
  – Add their own new models
  – Add their own new observation types
  – Change the assimilation algorithms
DART is used at:

43 UCAR member universities
More than 100 other sites

- Public domain software for 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

14 Aug 2013
FRCRC Symposium
DART Models

• 1D, 2D+
  – 6 Lorenz models, simple chaotic models (e.g. Ikeda, Null, 9var, SQG, PE2LYR, Bgrid_solo)

• Geophysical Models
  – Coupled Climate, Weather, Ocean, Land
     (e.g. CESM, WRF, POP, MITgcm, COAMPS, GITM, MPAS, TIEgcm, Rose, NOAH, NOGAPS)

• Economic, Epidemiological, Ecosystem, etc
Example Dart Observation Types

- Atmospheric Obs
  - Radiosondes (balloons) Temperature, Winds
  - Aircraft, Satellite Winds, Surface Obs
- Ocean Obs
  - Temperature, Salinity, Sea Surface Temp/Height
- Land Obs
  - Snow cover, CO Fluxes from Towers
- Novel Obs Types
  - GPS Radio Occultation (temperature, moisture)
  - Gravity/Length of Day, Leaf Area Index
Examples of Observation Density by Obs Type

Observations 1 December 2006

GPS

ACARS and Aircraft

Radiosondes

Sat Winds

14 Aug 2013

FRCRC Symposium
Atmospheric Reanalysis

O(1 million) atmospheric obs assimilated every day.

Assimilation uses 80 members of 2° FV CAM forced by a single ocean.

Used in turn to force an ensemble of ocean models where each ocean ensemble member is matched with a different atmosphere state.
Current Research Efforts

• DART runs well on O(10 – 1000) processors
• Highly scalable systems require less communication, more asynchronicity
  – Less memory per node, more nodes, lower power
  – Harder to program Geophysical applications
• DART parallelizes differently than most apps
  – 3 distinct data decompositions for parallelism
Data Decompositions

• Model Data Decomposition
  – Every model has a different data layout
  – DART uses files to exchange data with model

• Computing expected obs values (‘forward ops’)
  – Need multiple state items from a single ensemble, only parallelizes well up to N ensembles (100s)
  – Area of active development

• State adjustment (the actual assimilation)
  – Need state items from all ensemble members
  – DART parallelizes well since N obs is $O(10^K – 10^M)$
Parallelism and Communication

• Model algorithms are usually grid based
  – Best distribution puts nearest neighbors on same tasks and communicates across boundaries

• DART algorithms are pointwise
  – Great for avoiding support in DART of all possible model grids
  – Better concurrency and load balancing when neighboring points are assigned to different tasks
Typical Grid Layout

PE 1  PE 2  PE 3  PE 4
Typical Grid Layout
Typical Grid Layout
Current Work

• One-sided MPI communication
  – During the ‘forward operator’ computation
  – More concurrency because now $O(\text{number of obs})$ not $O(\text{number of ensembles})$
  – Fewer sync points, read-only data so no locking
  – Bring only the necessary data to where it needs to be used
  – Never have to fit entire state for a single ensemble member into single task memory
Current Work (cont)

• Do scatter/gather during I/O
  – DART uses files as intermediaries between it and the model – isolates us from model data decomposition
  – Read and write with parallel libraries includes scatter/gather capabilities
  – Perhaps a step towards in-memory data exchanges with parallel models (need general solution/portability)
Current Work (cont)

• Looking at places to replicate computation to save communication

• Still must address ease-of-use issues and maintain user-extensible code
  – Hide MPI code at a level where user does not have to understand all the details
  – Must be able to document and explain how to add new models and new observation operators
Thank you!

nancy@ucar.edu
www.image.ucar.edu/DAReS
Observation Visualization Tools
Other Data Assimilation Benefits

• Get a better prediction of future state of the system
  – Numerical Weather Prediction
• Uncover model deficiencies or errors
  – Biases or errors or deficient equations
• Evaluate the accuracy or information content of an observation type
  – Amount of error or impact of one type on the results
• Design new observing systems
  – Evaluate effectiveness of possible frequencies or density of new observations
Ensemble Kalman Filter (EnKF) Data Assimilation

• Run many copies of the same model with slightly different input data
• Have each model copy predict what the observation value should be
• If nearby model values are low and the predicted observation is low, increase them
• If nearby model values are high and the predicted observation is low, lower them
• You don’t have to know what equations the model is solving; you do this all with statistics and correlations between model variables and obs
User-Extensible Software Challenges

- User-extensible code means you have to make it so users can extend your code
  - Documentation of internal interfaces
  - Examples of use
  - No side effects between user-adaptable functions
  - No algorithms so exotic they make the code incomprehensible to a scientific user
Simpler is Harder than Complex

• It’s hard to write simple code
  – It’s easier to tack on more code rather than refactor existing code down to the core ideas
• Orthogonal concepts need to be kept orthogonal
  – No side effects, no linkages between unrelated concepts
• User perspective can be hard for software engineers
  – Error messages should give user guidance on to how to fix the problem
  – Parameter names and values must make sense to the user (not the coder)
DART is:

• Education
• Exploration
• Research
• Operations
**World Ocean Database**

These counts are for 1998 & 1999 and are representative.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLOAT_SALINITY</td>
<td>68200</td>
</tr>
<tr>
<td>FLOAT_TEMPERATURE</td>
<td>395032</td>
</tr>
<tr>
<td>DRIFTER_TEMPERATURE</td>
<td>33963</td>
</tr>
<tr>
<td>MOORING_SALINITY</td>
<td>27476</td>
</tr>
<tr>
<td>MOORING_TEMPERATURE</td>
<td>623967</td>
</tr>
<tr>
<td>BOTTLE_SALINITY</td>
<td>79855</td>
</tr>
<tr>
<td>BOTTLE_TEMPERATURE</td>
<td>81488</td>
</tr>
<tr>
<td>CTD_SALINITY</td>
<td>328812</td>
</tr>
<tr>
<td>CTD_TEMPERATURE</td>
<td>368715</td>
</tr>
<tr>
<td>STD_SALINITY</td>
<td>674</td>
</tr>
<tr>
<td>STD_TEMPERATURE</td>
<td>677</td>
</tr>
<tr>
<td>XCTD_SALINITY</td>
<td>3328</td>
</tr>
<tr>
<td>XCTD_TEMPERATURE</td>
<td>5790</td>
</tr>
<tr>
<td>MBT_TEMPERATURE</td>
<td>58206</td>
</tr>
<tr>
<td>XBT_TEMPERATURE</td>
<td>1093330</td>
</tr>
<tr>
<td>APB_TEMPERATURE</td>
<td>580111</td>
</tr>
</tbody>
</table>

**temperature and salinity observations**