Scalable Computing Challenges in Ensemble Data Assimilation

CAS2K13
Nancy Collins
NCAR - IMAGe/DARes
12 Sept 2013
Overview

• What is Data Assimilation?

• What is DART?

• Current Work on Highly Scalable Systems
Overview of Data Assimilation

Prediction Model
Overview of Data Assimilation

Prediction Model

Observing System
Overview of Data Assimilation

Prediction Model

Observing System

Forecasts

Observations

Data Assimilation

\[ t_k \rightarrow y \]

\[ \text{Forecast} \]

\[ \text{Observation} \]
Overview of Data Assimilation

- Prediction Model
- Observing System

Forecasts → Data Assimilation → Observations

Analysis → Data Assimilation → Diagnostics

12 Sept 2013
CAS2K13 - Annecy
Overview of Data Assimilation

- Prediction Model
- Observing System
- Data Assimilation
- Analysis
- Diagnostics

Initial Conditions
Forecasts
Observations
Overview of Data Assimilation

Typical Numerical Weather Prediction Process

Prediction Model → Data Assimilation → Analysis

Observing System → Data Assimilation → Diagnostics
Overview of Data Assimilation

Prediction Model

Observing System

Data Assimilation

Analysis

Diagnostics

Identify Systematic Errors
Overview of Data Assimilation

Prediction Model

Observing System

Data Assimilation

Analysis

Diagnostics

Estimate model Parameters
Overview of Data Assimilation

Prediction Model

Observing System

Data Assimilation

Analysis

Diagnostics

Estimate Errors, Design Systems
Data Assimilation Research Testbed (DART)

DART is a community ensemble assimilation facility.
Data Assimilation Types

• Variational Systems
  – Used by operational NWP forecasting centers

• Ensemble Systems
  – Make many forecasts
  – Easier to develop a DA system, especially for large models
  – Feasible for individual researchers, small groups
  – Produces uncertainty information
Data Assimilation Research Testbed

• DART software is used for:
  – Building Ensemble 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:

48 UCAR member universities
More than 100 other sites
DART Models

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

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

• Economic, Epidemiological, Ecosystem, etc
Lorenz Models

Lorenz 63 (butterfly)

Lorenz 96
Lorenz 96 Free Run

State Variable

-10
0
10
time 001
truth

5 10 15 20 25 30 35 40
Lorenz 96 Ensembles

![Graph showing state variable over time with 'truth' and 'ensemble' labels.]
Lorenz 96 with DA
DART Models

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

• Full 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, GPS (T, Q)

- **Ocean Obs**
  - Temperature, Salinity, Sea Surface Temp/Height

- **Land Obs**
  - Snow cover, CO Fluxes from Towers

- **Novel Obs Types**
  - Gravity/Length of Day, Leaf Area Index, COSMOS Neutron Soil Moisture
Examples of Observation Density by Obs Type

Observations 1 December 2006

GPS

ACARS and Aircraft

Radiosondes

Sat Winds

12 Sept 2013

CAS2K13 - Annecy
Atmospheric Reanalysis

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

O(1 million) atmospheric obs assimilated every day.

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

500 hPa GPH
Feb 17 2003

CONTOUR FROM 5200 TO 5700 BY 100
Observation Visualization Tools

![Image of observation visualization](image.png)
Parallel Computation Issues

• Model algorithms are usually grid based
  – Subregions of the model grid are distributed to different processors for parallel computation
  – Best distribution puts nearest neighbors on same processors and communicates across boundaries

• DART parallelizes differently than most apps
  – 3 distinct data decompositions for parallelism
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)
Ensemble Filter For Large Geophysical Models

1. Use model to advance ensemble (3 members here) to time at which next observation becomes available.
Ensemble Filter For Large Geophysical Models

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.
2. Get prior ensemble sample of observation, $y = h(x)$, by applying forward operator $h$ to each ensemble member.

Ensemble Filter For Large Geophysical Models

Now: Entire state available for forward ops
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.
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

5. Use ensemble samples of $y$ and each state variable to linearly regress observation increments onto state variable increments.

Now: Distribution ensures good load balancing
6. When all ensemble members for each state variable are updated, there is a new analysis. Integrate to time of next observation ...
6. When all ensemble members for each state variable are updated, there is a new analysis. Integrate to time of next observation ...
DART Evolution Challenges

• DART runs well on $O(10 - 1000)$ processors
• New architectures $O(100,000)$ processors
• Highly scalable systems require less global communication, more asynchronicity
  – Less memory per node, more nodes, lower power
  – Harder to program Geophysical applications
Addressing Shrinking Memory Sizes

• Redesigning forward operator algorithms to avoid the need for entire state of one ensemble member in single task memory

• Requires additional communication for some types of forward operators

• Keeping spatial locality lowers communication overhead but presents load balancing issues
Ensemble Filter For Large Geophysical Models

2. Get prior ensemble sample of observation, \( y = h(x) \), by applying forward operator \( h \) to each ensemble member.

Old: Entire state available for forward ops
2. Get prior ensemble sample of observation, $y = h(x)$, by applying forward operator $h$ to each ensemble member.
2. Get prior ensemble sample of observation, $y = h(x)$, by applying the forward operator $h$ to each ensemble member.

Ensemble Filter For Large Geophysical Models

New: But some communication may be needed
Avoiding Global Communication

• Current implementation transposes data for load balancing during state adjustment phase
• Global operations prohibitively expensive on O(100,000) processor counts
• Avoiding transposes avoids global operation but again raises more load balancing issues
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 can be handled independently!
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.

New: Less communication; load balance issues
5. Use ensemble samples of $y$ and each state variable to linearly regress observation increments onto state variable increments.

New: Potentially very big load balance issues
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

5. Use ensemble samples of $y$ and each state variable to linearly regress observation increments onto state variable increments.
DART Evolution for MPP Systems

• Allow single ensemble state to span multiple tasks
  – Decompose across a small number of nodes
  – Data movement confined to subsets of nodes

• Support distributed forward operator computations
  – Spatially local decomposition minimizes communication
  – One-sided MPI-2 communication avoids barriers

• Avoid global communication at state adjustment phase
  – Smarter decomposition for load balancing
  – Parallel adjustments of disjoint observation sets
DART Evolution (cont)

• Maintain reasonable interfaces that enable user-extensible sections of the code
  – Support for modification by domain scientists
  – Clear and understandable process for adding new models and new observation operators
  – Encapsulate MPI code at a level where user does not have to understand the details

• Identical results to serial implementation??
Thank you!

nancy@ucar.edu

www.image.ucar.edu/DARes
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.
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