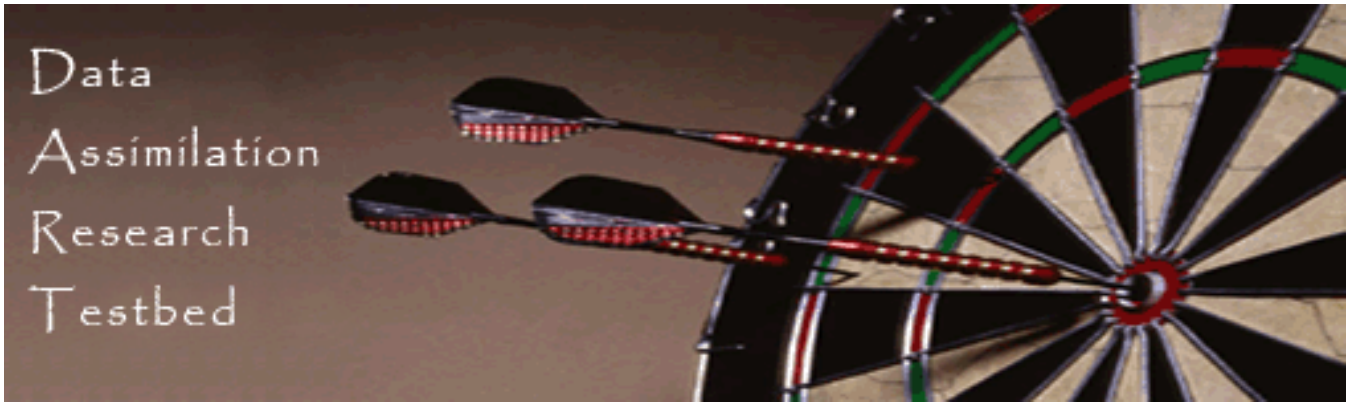
Data Assimilation with CLM & DART



Tim Hoar: *National Center for Atmospheric Research*

with a whole lot of help from:

Bill Sacks, Mariana Vertenstein, Tony Craig, Jim Edwards: *NCAR*

Andrew Fox: *National Ecological Observatory Network (NEON)*

Nancy Collins, Kevin Raeder, Jeff Anderson: *NCAR*

Yongfei Zhang: *University of Texas Austin*

Is DA different for NWP and ecosystem models?

	Data Assimilation in NWP	Data Assimilation in CLM
Main objective	Improved initial conditions Forecast improvement	Process understanding Regional quantification Forecasting
Dynamics	Physics – essentially well known from first principles	Physical, biological, chemical – Only partially known, empirical relationships
Observations	High spatial and temporal density	Very different spatial and temporal characteristics
Mathematical problem	Optimization of initial conditions	Initial value problem (e.g. pools) Boundary conditions (e.g. fluxes) Parameter optimization

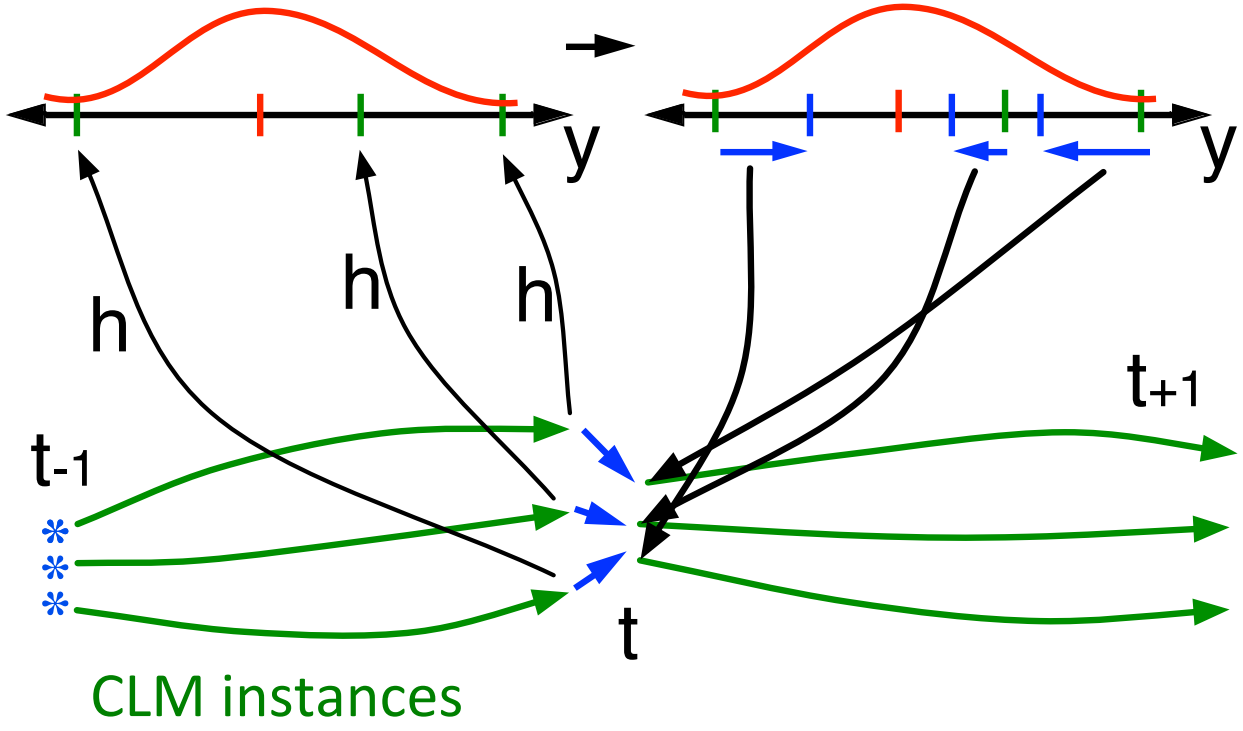
NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

NSF

Slide courtesy of NEON

A generic ensemble filter system like DART needs:

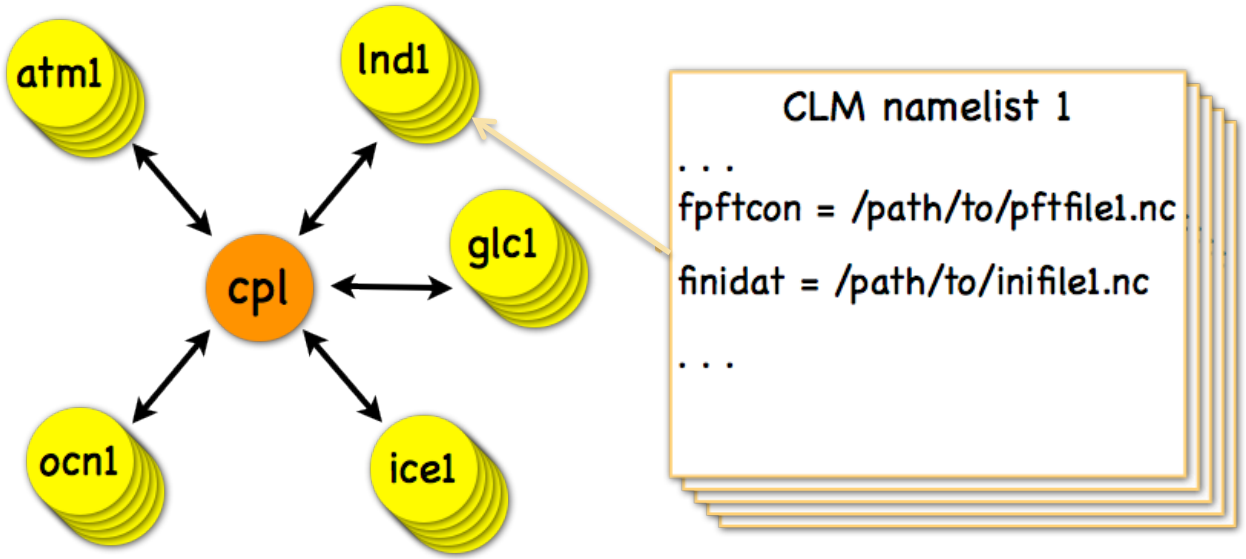
1. A way to make model forecasts.
2. A way to estimate what the observation would be – given the model state. This is the observation operator – h.



The **increments** are regressed onto as many CLM **state variables** as you like. If there is a correlation, the CLM state gets adjusted in the restart file.

Multi-instance CESM code

- A multi-instance version of CESM has been developed that more easily facilitates ensemble-based DA
- For example, multiple land models can be driven by multiple data-atmospheres in a single executable.
- This capability should be available in the next CESM release.

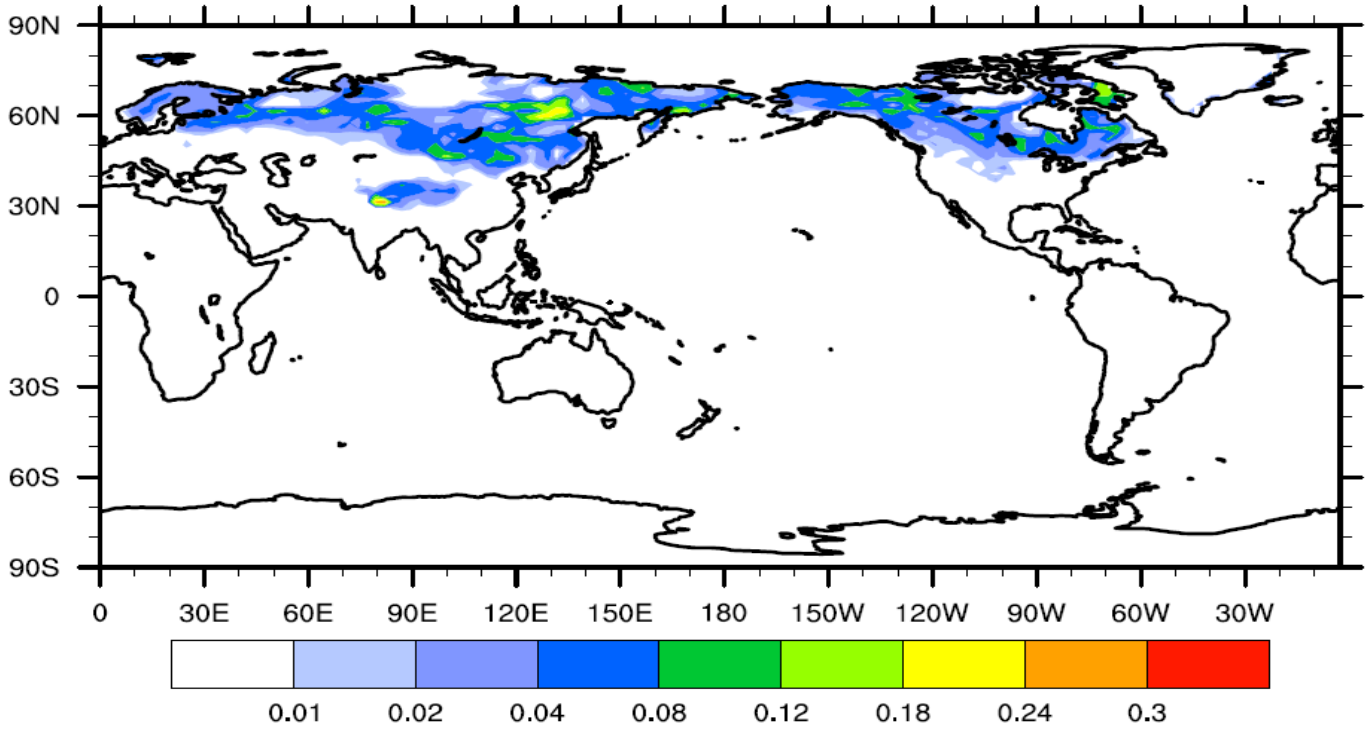


Slide courtesy of NEON

Assimilation of MODIS snow cover fraction

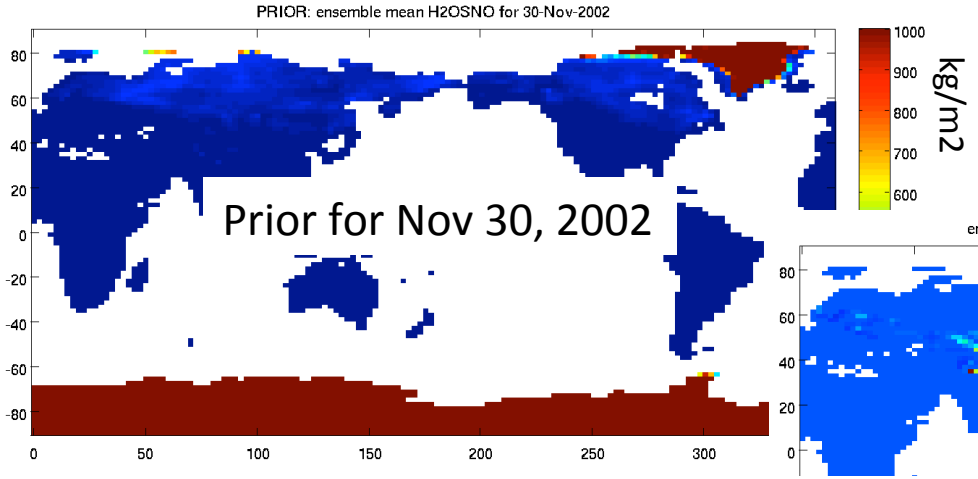
- 80 member ensemble for onset of NH winter
- Assimilate once per day
- Level 3 MODIS product – regridded to a daily 1 degree grid
- Observation error variance is 0.1 (for lack of a better value)
- Observations can impact state variables within 200km
- CLM variable to be updated is the snow water equivalent "H2OSNO"

Standard deviation of the snow cover fraction initial conditions for Oct. 2002

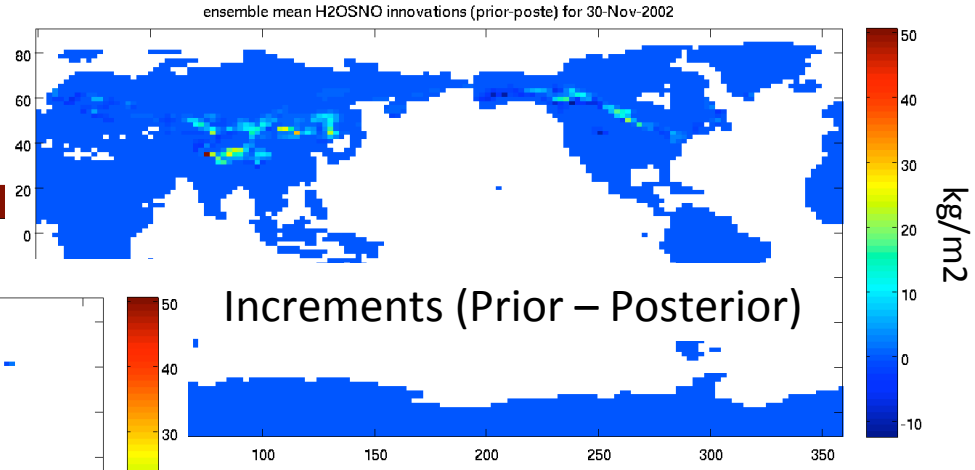


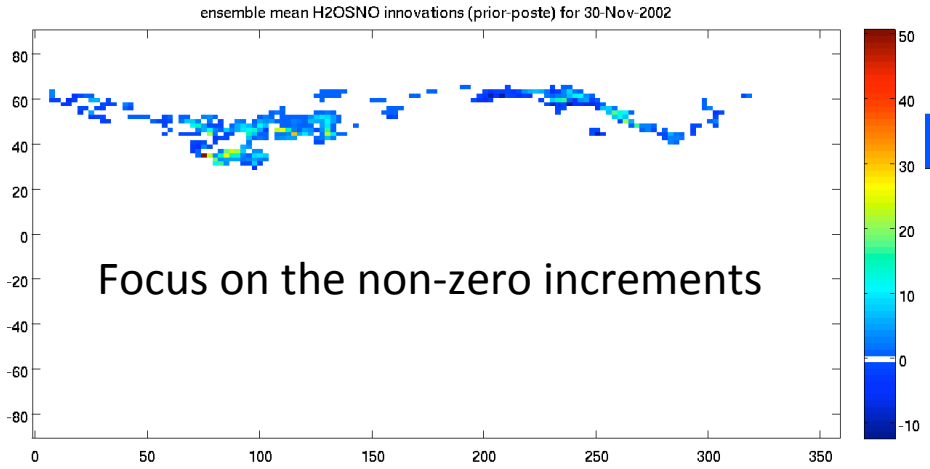
Slide courtesy of Yongfei Zhang

An early result: assimilation of MODIS *snowcover fraction* on total *snow water equivalent* in CLM.



Thanks Yongfei!





The model state is changing in reasonable places, by reasonable amounts. At this point, that's all we're looking for.

Slide courtesy of Yongfei Zhang

What can CLM-DART do *right now*:

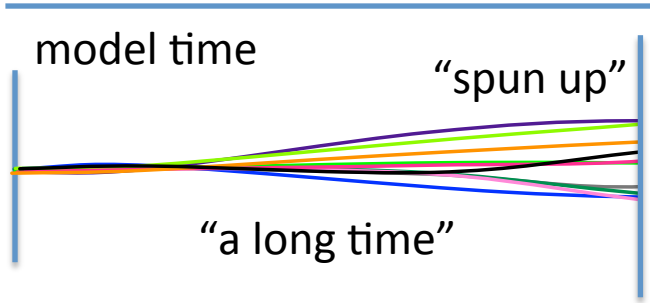
- Use the CESM multi-instance capability to run simultaneous instances of CLM.
- Force each instance with different *realistic* atmospheric conditions (say, from an offline CAM/DART assimilation).
- Assimilate observations every time CESM stops.
- Modify the CLM restart file contents to be more consistent with observations – *and not just at the observation location!*
- Use CLM history files to provide model states to compare with observations, i.e. the observation operator IS the history file (GRACE observations, NEE, …).

What can CLM-DART do *right now*:

- Use the CESM multi-instance capability to run up to 80 simultaneous instances of CLM
- Force each instance with different *realistic* atmospheric conditions (from an offline CAM/DART assimilation)
- Use the multi-instance capability to assimilate every midnight
- Modify the CLM restart file contents to be more consistent with observations – *and not just at the observation location*
- Can use CLM history files to provide model states to compare with observations, i.e. the observation operator IS the history file (GRACE observations, NEE, …)
- *Defeat any (and all?) balance checks Erik can throw at us …*
- *Blow your file quota on any machine, any time, without breaking a sweat …*

Creating the initial ensemble of CLM.

Replicate what we have N times.
Use a unique (and different!) *realistic* DATM for each.
Run them forward for "a long time".



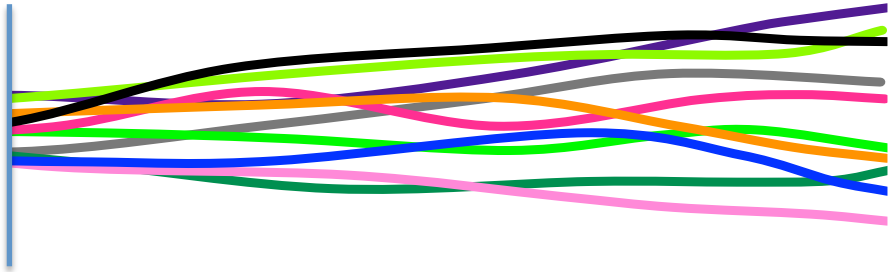
Getting a proper initial ensemble is an area of active research.

We don't know how much spread we NEED to capture the uncertainty in the system.

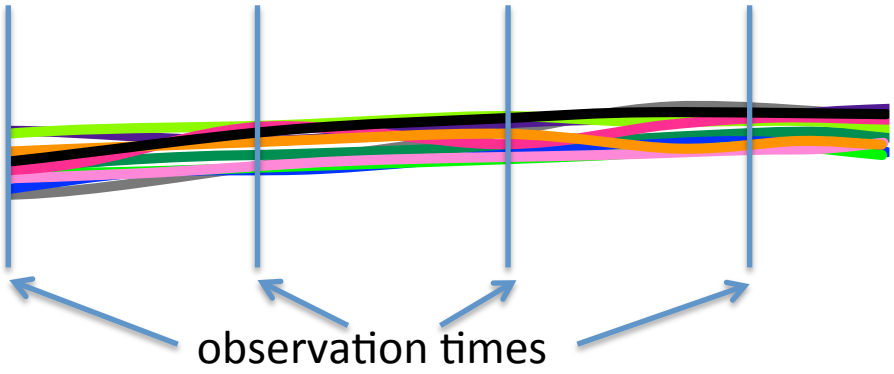
The ensemble advantage.

You can represent uncertainty.

In a free run,
the ensemble spread
frequently grows.

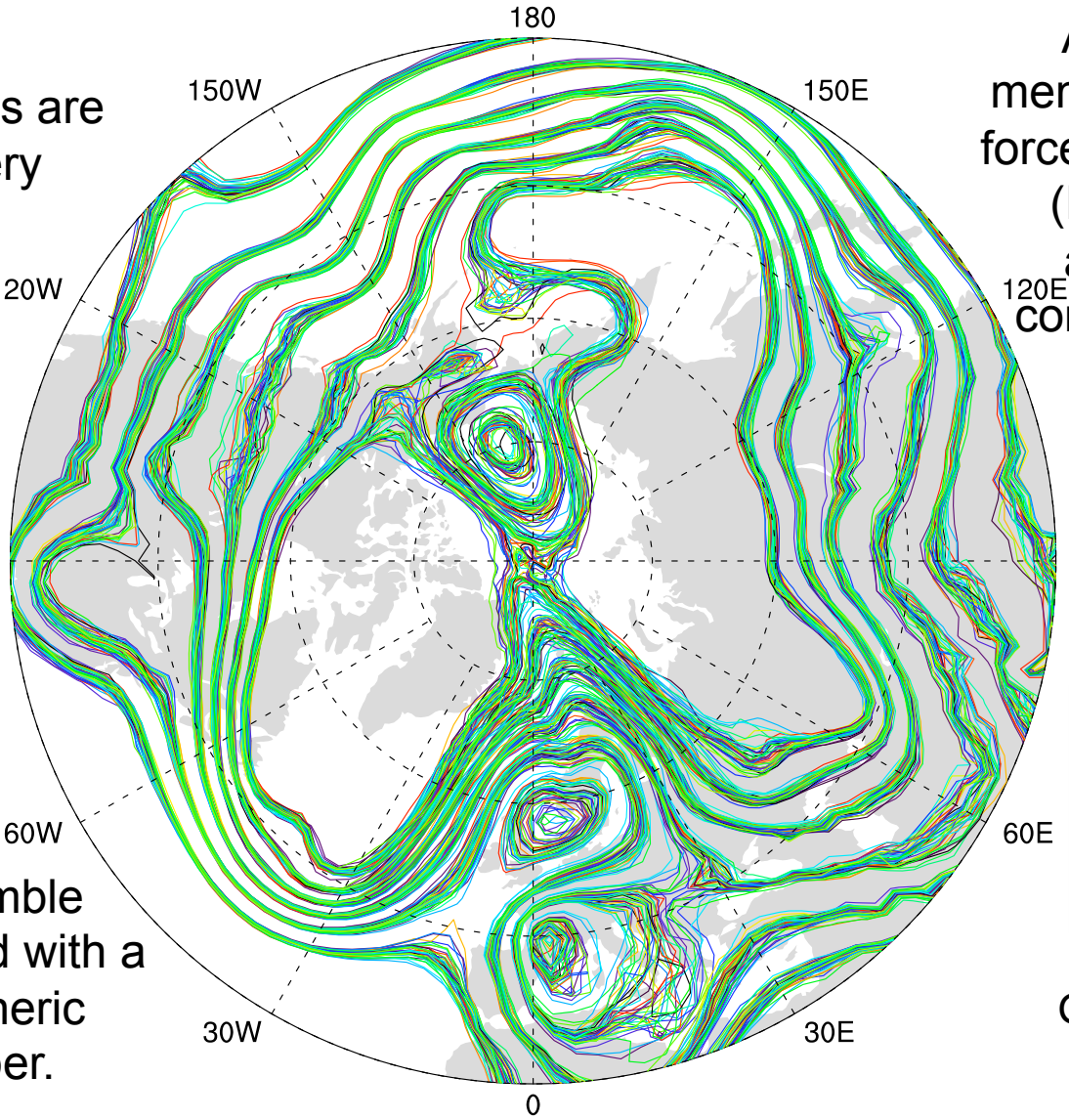


With a good assimilation:
ensemble spread ultimately
remains stable and small
enough to be informative



Atmospheric Reanalysis

O(1 million) atmospheric obs are assimilated every day.



500 hPa GPH
Feb 17 2003

Assimilation uses 80 members of 2º FV CAM forced by a single ocean (Hadley+ NCEP-OI2) and produces a very competitive reanalysis.

1998-2010
4x daily
is available.

Each CLM ensemble member is forced with a different atmospheric reanalysis member.

Generates spread in the land model.

CLM-DART coupling

- Our goal has been to “Do no harm” to CLM
- DART’s namelist allows you to choose what CLM variables get updated by the assimilation
- New routines communicate between CLM and DART
- At predetermined assimilation intervals:

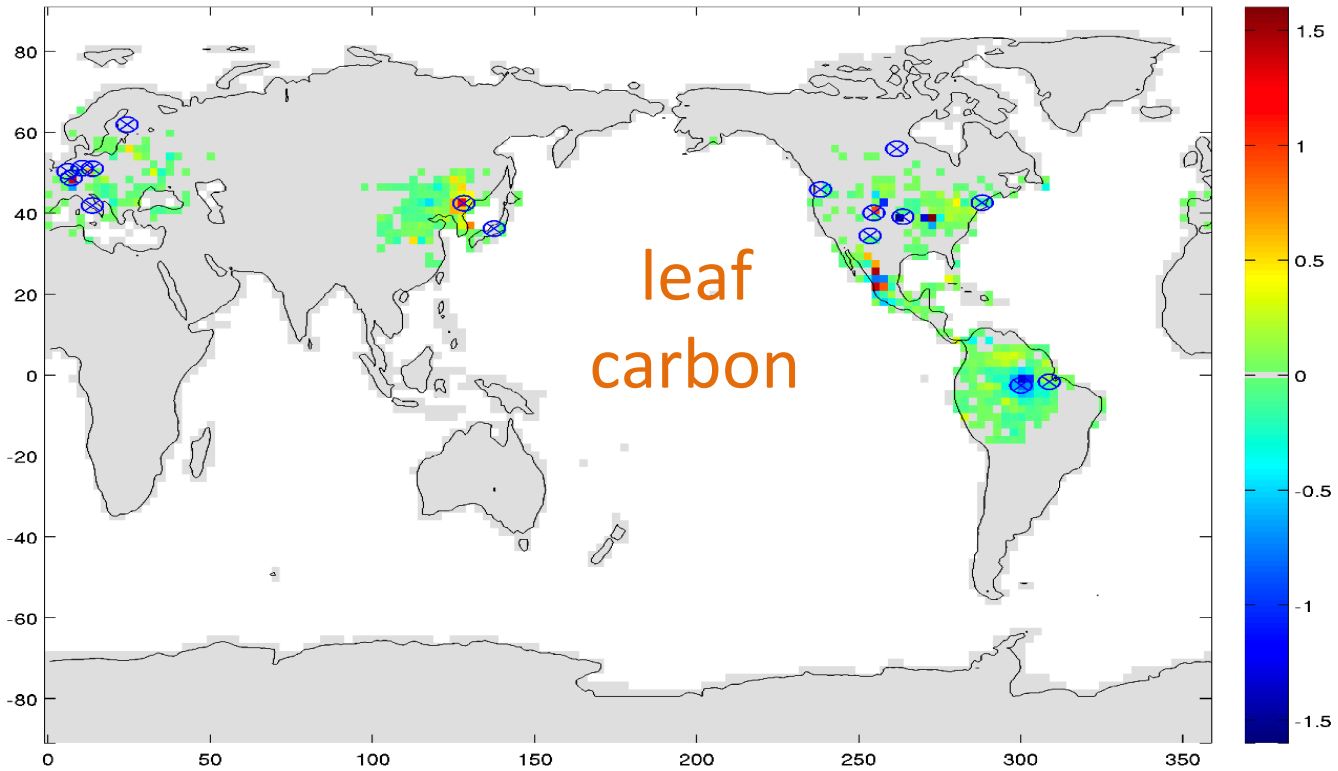
1. CESM/CLM stops and writes restart and history files
2. DART state vector extracted from CLM restart & history files
3. Increments calculated and applied to DART state vector
4. CLM restart files updated with adjusted DART state vector
5. CESM postrun script executes

Proof-of-concept using Perfect Model Experiment of *leafc* follows.

- 18 synthetic observation locations of leaf carbon
- 40 CLM instances spun up for several months

Innovation map of leafc on 4 May 2000

- **Information** from a site is **extrapolated** across space through the covariance matrix represented by the ensemble of CLM instances.
- Generally, largest updates closest to observation sites.



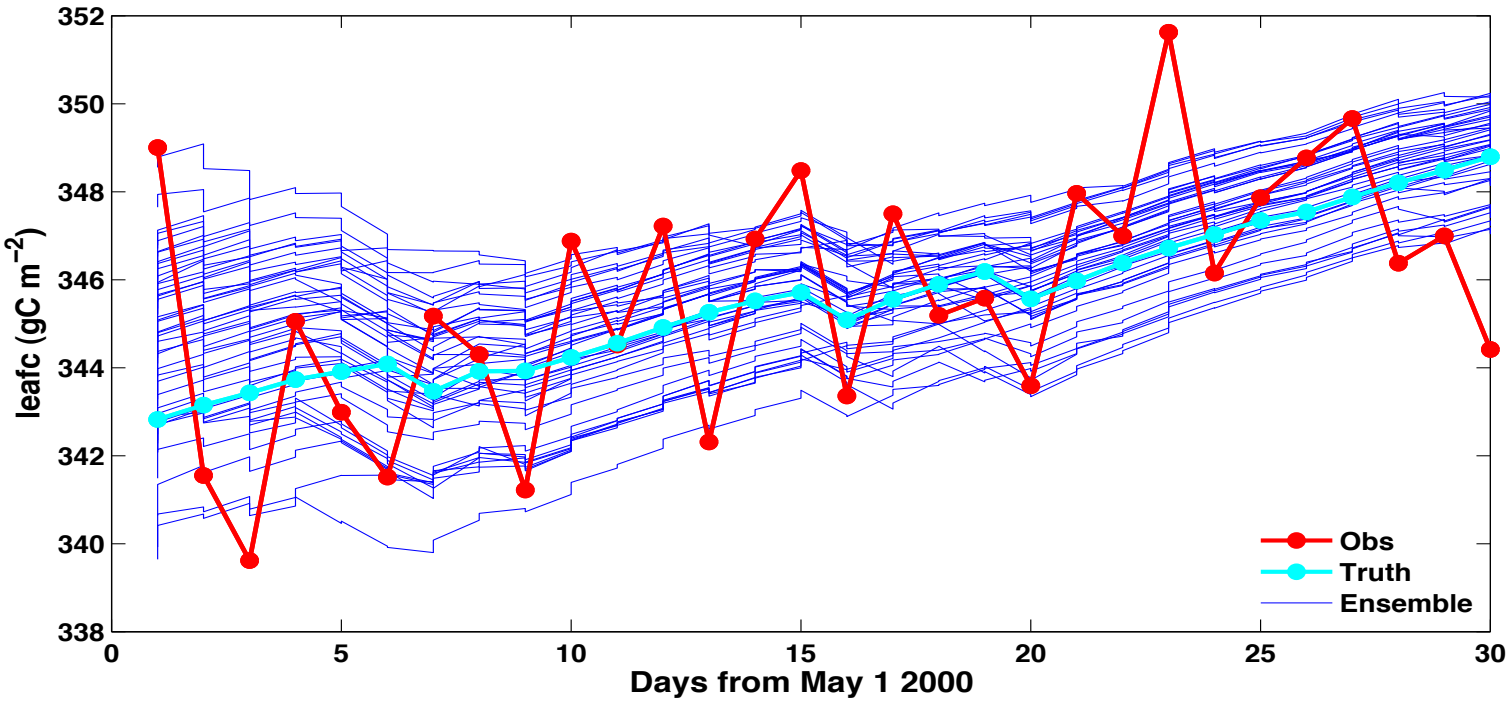
NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

NSF

Slide courtesy of NEON

Time series of “truth”, obs and 40 ens members

- 40 member ensemble of `leafc` in a single grid cell corresponding to 60.21°W, 2.61°S (Manaus, Brazil).
- Ensemble members (blue lines) show impact of assimilation.



- Andy has a CSL proposal for 420,000 core-hours on Yellowstone to continue.

EOS

EOS, TRANSACTIONS, AMERICAN GEOPHYSICAL UNION

VOLUME 93 NUMBER 23 5 JUNE 2012

IN THIS ISSUE:

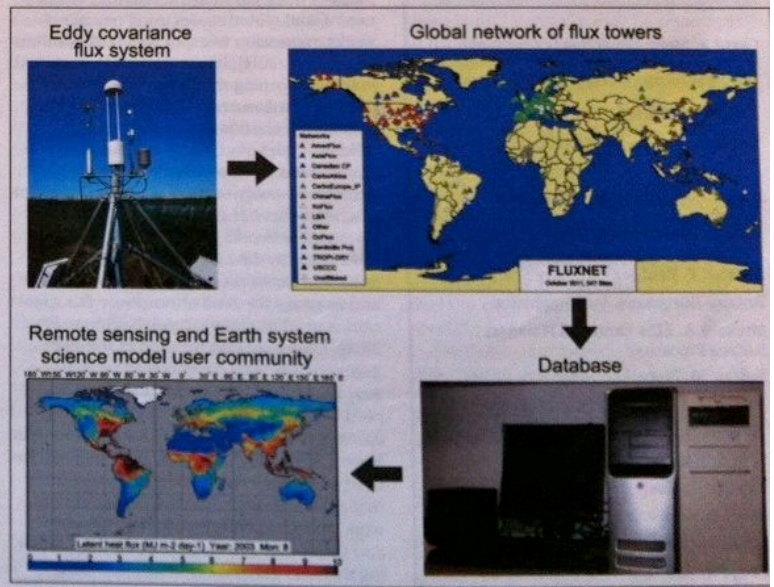
The Role of Trace Gas Flux Networks in the Biogeosciences

Vast networks of meteorological sensors ring the globe, providing continuous measurements of an array of atmospheric state variables such as temperature, humidity, rainfall, and the concentration of carbon dioxide [*New et al.*, 1999; *Tans et al.*, 1996]. These measurements provide input to weather and climate models and are key to detecting trends in climate, greenhouse gases, and air pollution. Yet to understand how and why these atmospheric state variables vary in time and space, biogeoscientists need to know where, when, and at what rates important gases are flowing between the land and the atmosphere. Tracking trace gas fluxes provides information on plant or microbial metabolism and climate-ecosystem interactions.

The existence of trace gas flux networks is a relatively new phenomenon, dating back to research in 1984. The first gas flux measurement networks were regional in scope and were designed to track pollutant gases such as sulfur dioxide, ozone, nitric acid, and nitrogen dioxide. Atmospheric observations and model simulations were used

A key attribute of the eddy covariance method is its ability to measure fluxes in situ with minimal disturbance to the environment, at a spatial scale of hundreds of meters, and on time scales spanning hours, days, and years.

For the eddy covariance method to work, trace gas sensors must be able to respond to fluctuations in atmospheric gas concentrations over as little as a tenth of a second, maintain a stable calibration, possess a high signal-to-noise ratio, and, in cases where pumps are needed to move air to the sensor, have access to a power line. The current generation of carbon dioxide and water vapor sensors easily meets these criteria, and a revolution in instrument development is producing trace gas sensors capable of measuring a broad suite of compounds at high sampling rates with high sensitivity and precision. Those measuring stable isotopes of carbon, oxygen, and carbonyl sulfide can help partition fluxes between the vegetation and the soil. Those measuring methane and nitrous oxide can assess microbial activity

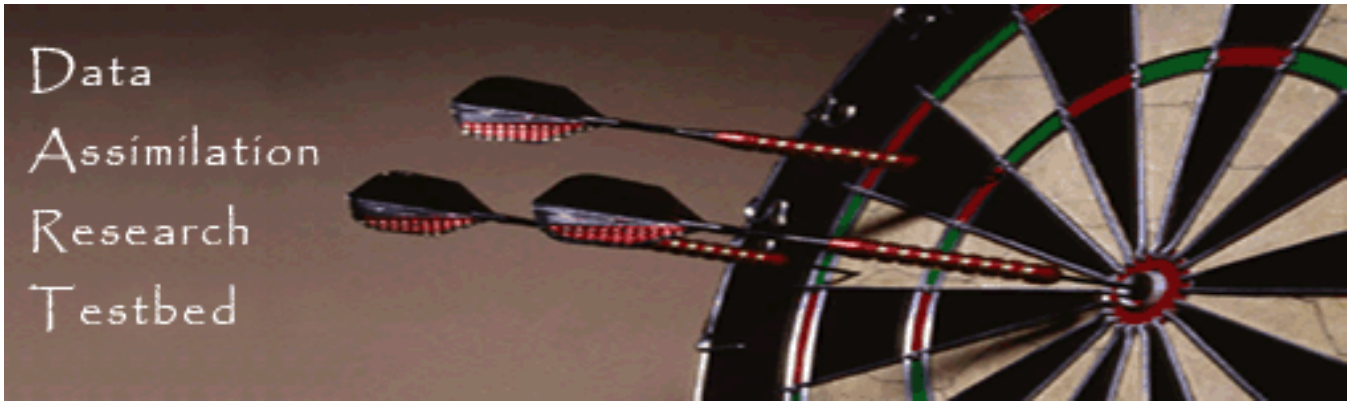


The components of a flux network: (top left) An eddy covariance flux tower, which acts as one site in (top right) a global network (FLUXNET, a component of NASA's Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)) that produces data that are merged and distributed through (bottom right) an integrative database. These data, used in conjunction with

Problems to be solved:

- Proper initial ensemble
- Creating snow with the right characteristics
- Bounded quantities
- When all ensembles have identical values the observations cannot have any effect with the current algorithms
- Forward operators – many flux observations are over timescales that are inconvenient – need soil moisture from last month and now…
- CLM has **a lot** of carbon species, hard to support all the forward operators required
- CLM's abstraction of grid cells, land units, etc., make the treatment of observations very peculiar. All land units in a grid cell share a location. Easy to have 'contradictory' observations.

For more information:



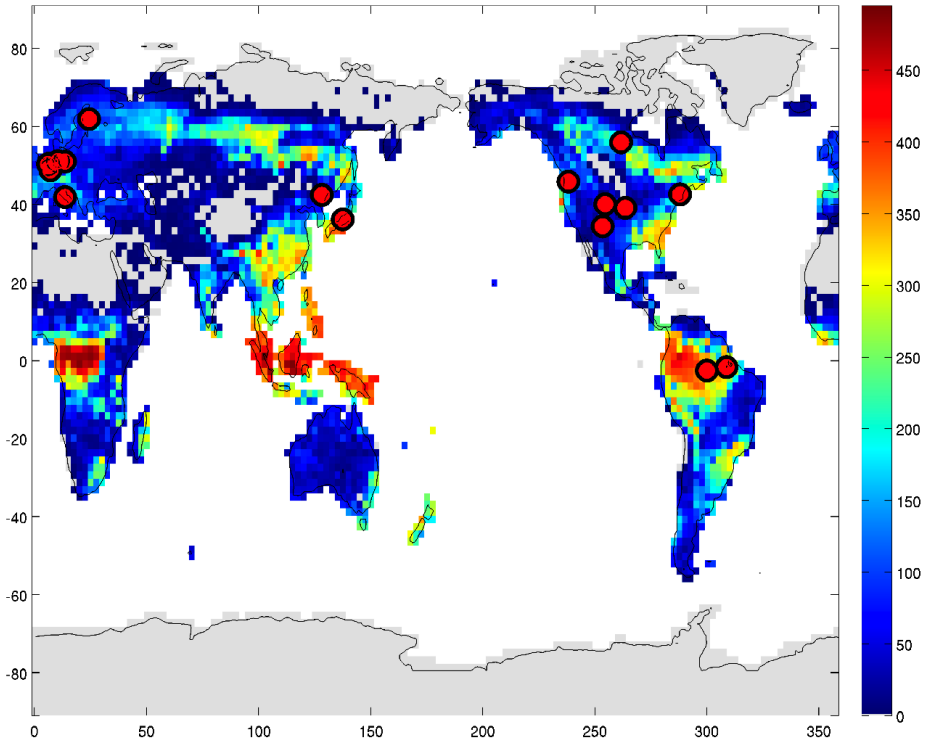
www.image.ucar.edu/DAReS/DART

dart@ucar.edu

An example of data assimilation in the CLM

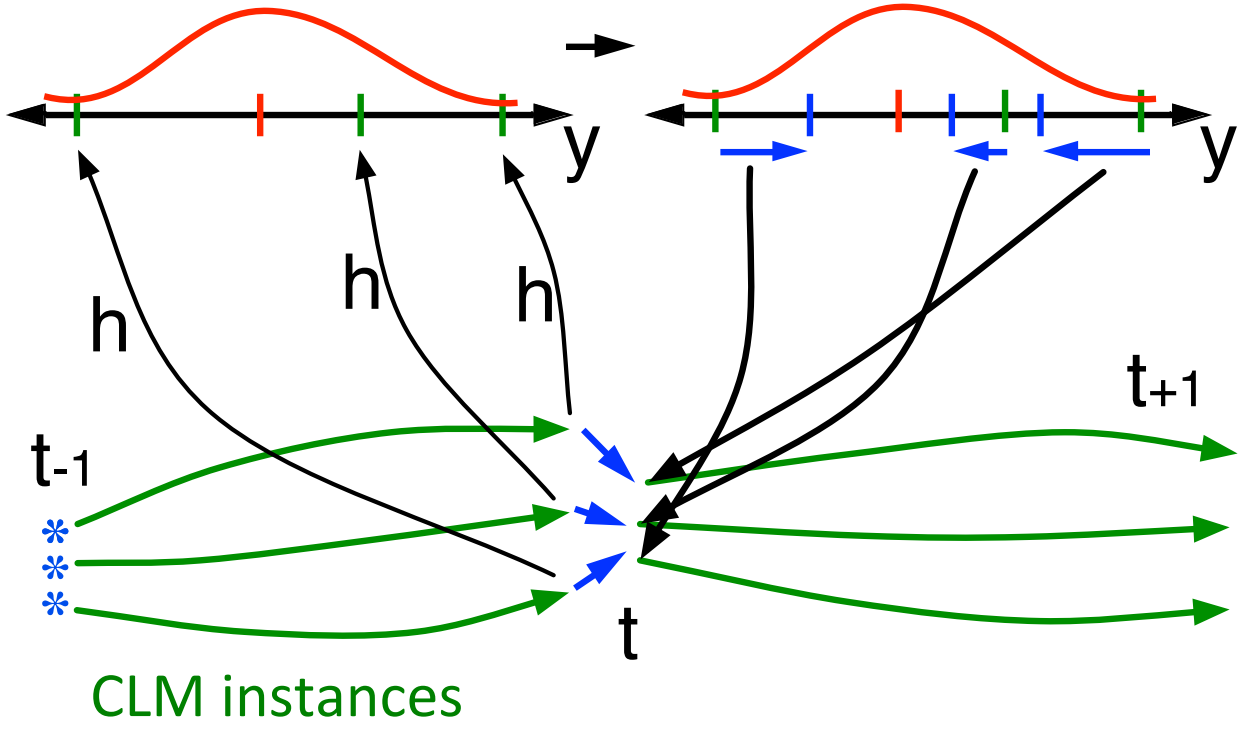
- 40 member ensemble of CLM forced with meteorology from 40 different data atmospheres in 2° grid global runs
- Leaf carbon is a key variable in CLM strongly influencing productivity, evapotranspiration and radiation dynamics
- Run 1 ensemble member forward from 1 May 2000, harvesting daily observations of `leafc` at 16 FLUXNET locations
- Run 40 ensemble members forward from 1 May 2000 for 30 days, assimilating synthetic observations

Global `leafc`, 1 May 2000



A generic ensemble filter system like DART needs:

1. A way to make model forecasts;
2. A way to compute the observation operators, h.

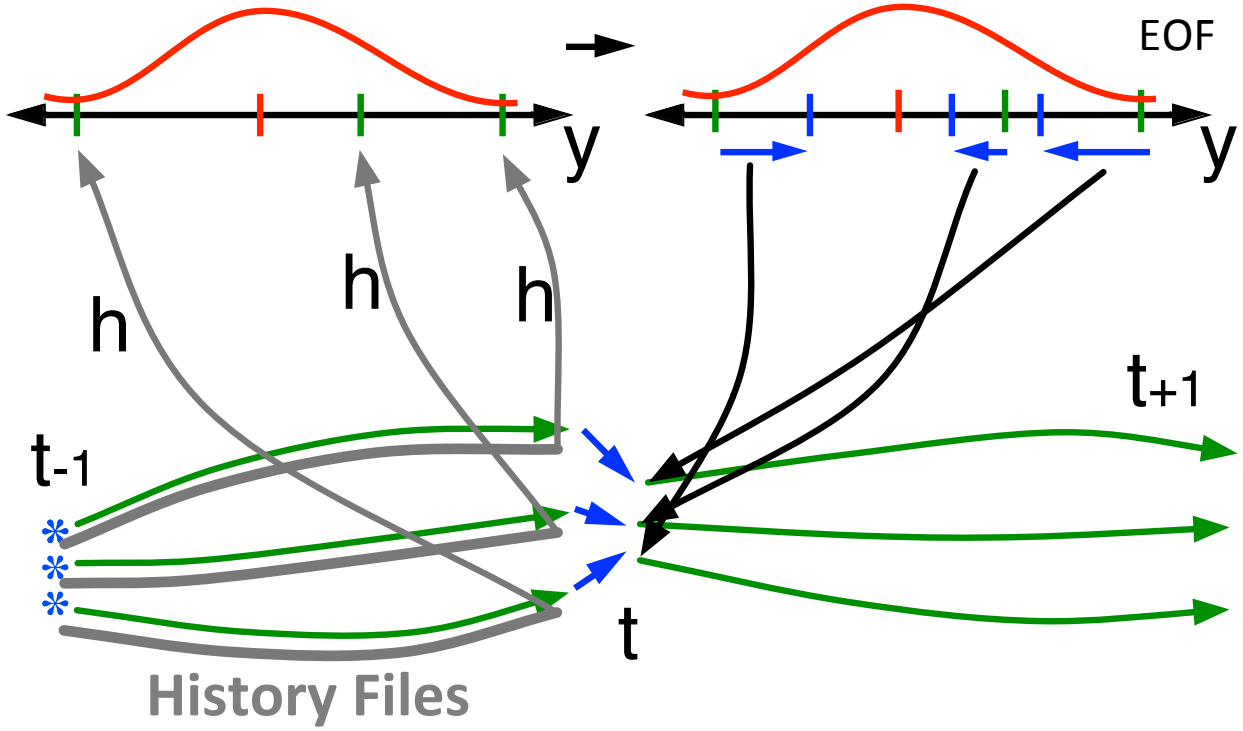


The **increments** are regressed onto as many CLM **state variables** as you like. If there is a correlation, the CLM state gets adjusted in the restart file.

History file games:

We can query a history file for the CLM state at

```
cat << EOF >! user_nl_clm
&clm_inparm
 hist_empty_htapes = .false.
 hist_fincl1 = 'NEP'
 hist_fincl2 = 'NEP'
 hist_nhtfrq = -24,1,
 hist_mfilt  = 1,48
 hist_avgflag_pertape = 'A','A'
/
EOF
```



The HARD part is: *What do we do when only SOME (or none!) of the ensembles have [snow,leaves,…] and the observations indicate otherwise?*

Corn Snow?

New Snow?

Sugar Snow?

Dry Snow?

Wet Snow?

"Champagne Powder"?

Crusty Snow?

Slushy Snow?

Old Snow?

Dirty Snow?

Early Season Snow?

Packed Snow?

Snow Density?

Snow Albedo?

The ensemble ***must*** have some uncertainty, it cannot use the same value for all. The model expert must provide guidance. It's even worse for the hundreds of carbon-based quantities!

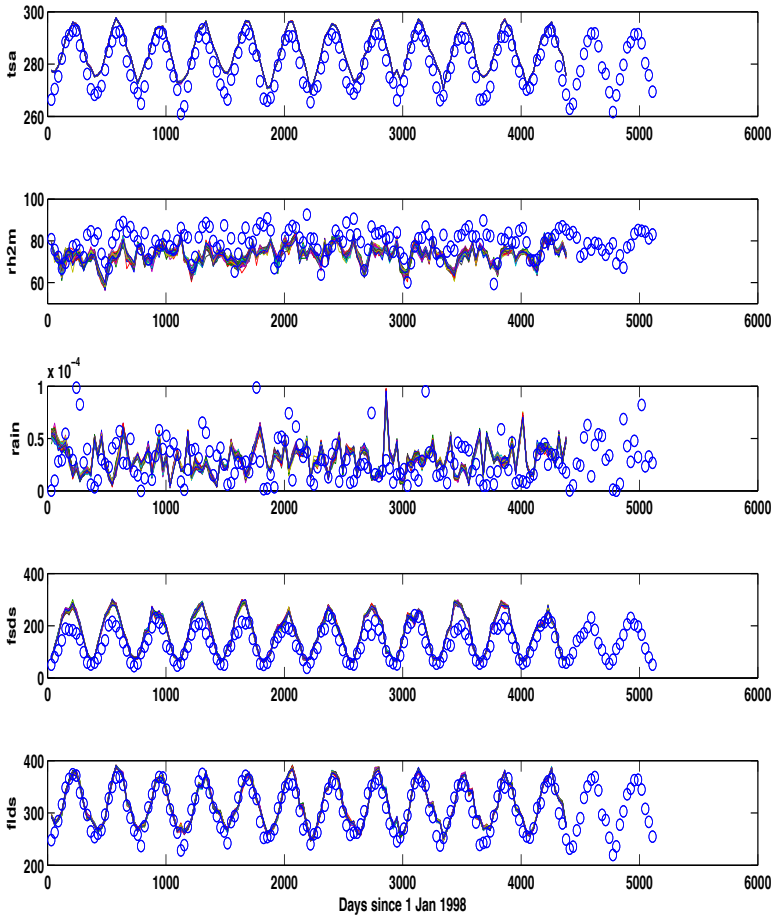
Details

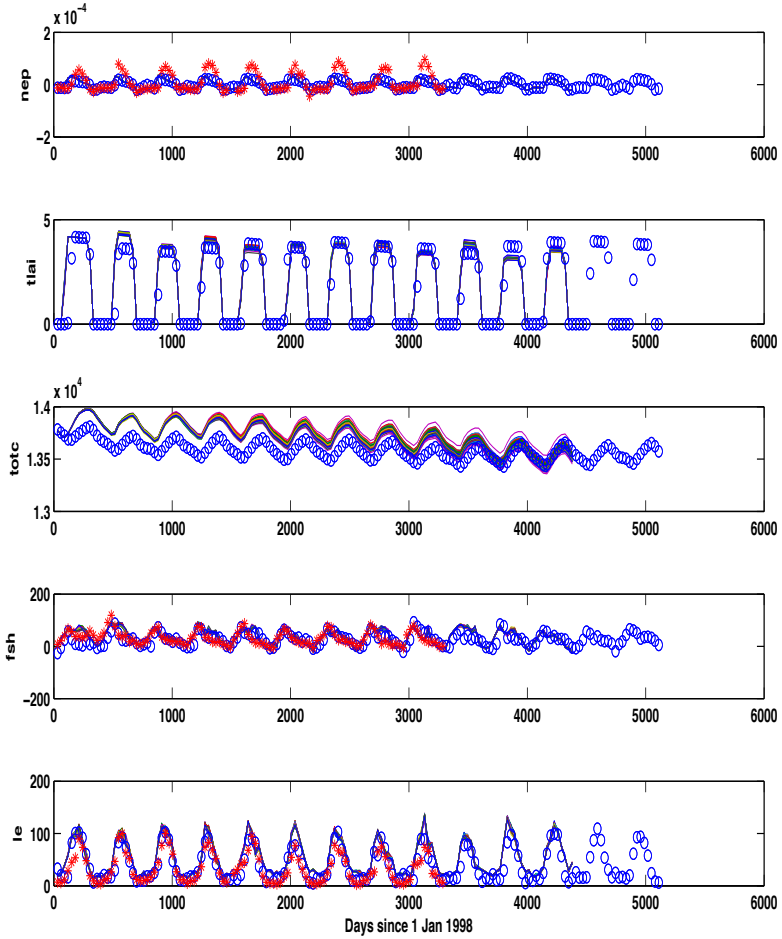
- DART allows you to choose what **CLM variables** get updated by the assimilation.

```
&clm_vars_nml
   clm_state_variables = 'frac_sno',      'KIND_SNOWCOVER_FRAC',
                         'DZSNO',         'KIND_SNOW_THICKNESS',
                         'H2OSNO',        'KIND_SNOW_WATER',
                         'T_SOISNO',      'KIND_SOIL_TEMPERATURE',
                         'leafc',         'KIND_LEAF_CARBON' /
```

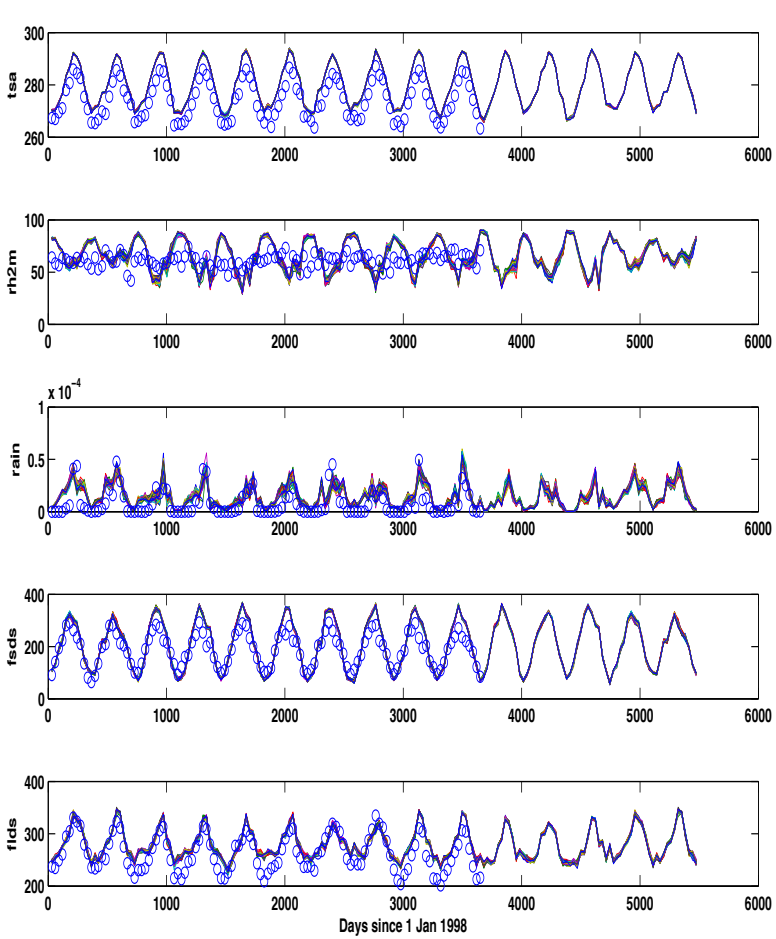
- These are read from a CLM restart file and reinserted after the assimilation.
- **Potential problem … balance/consistency?**

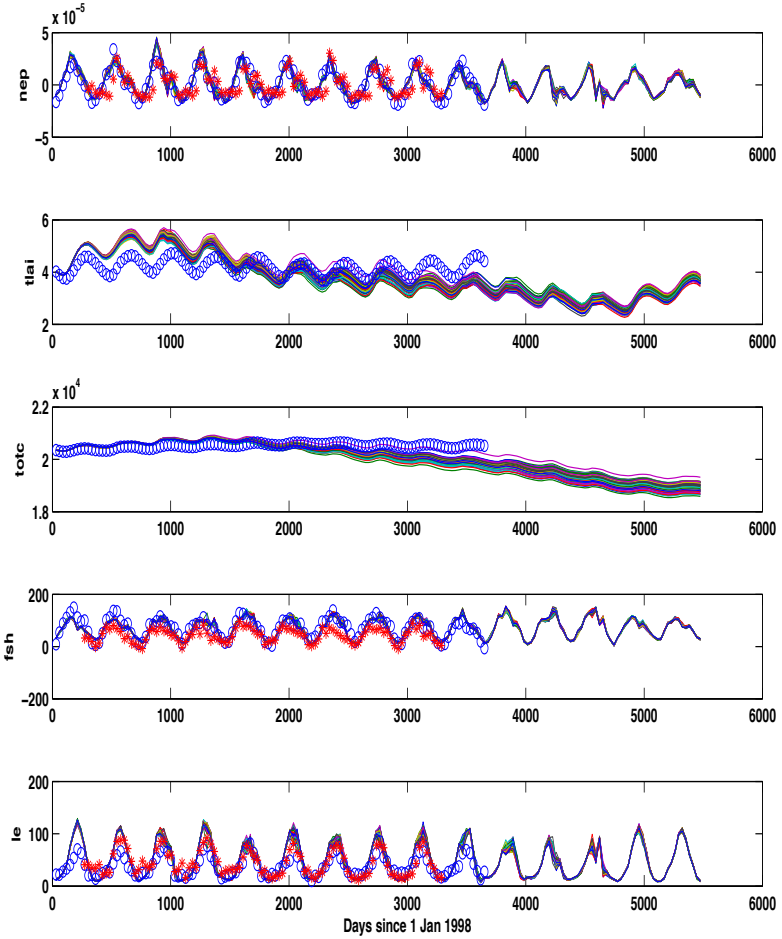
Harvard Forest – monthly averages 1998-2010ish





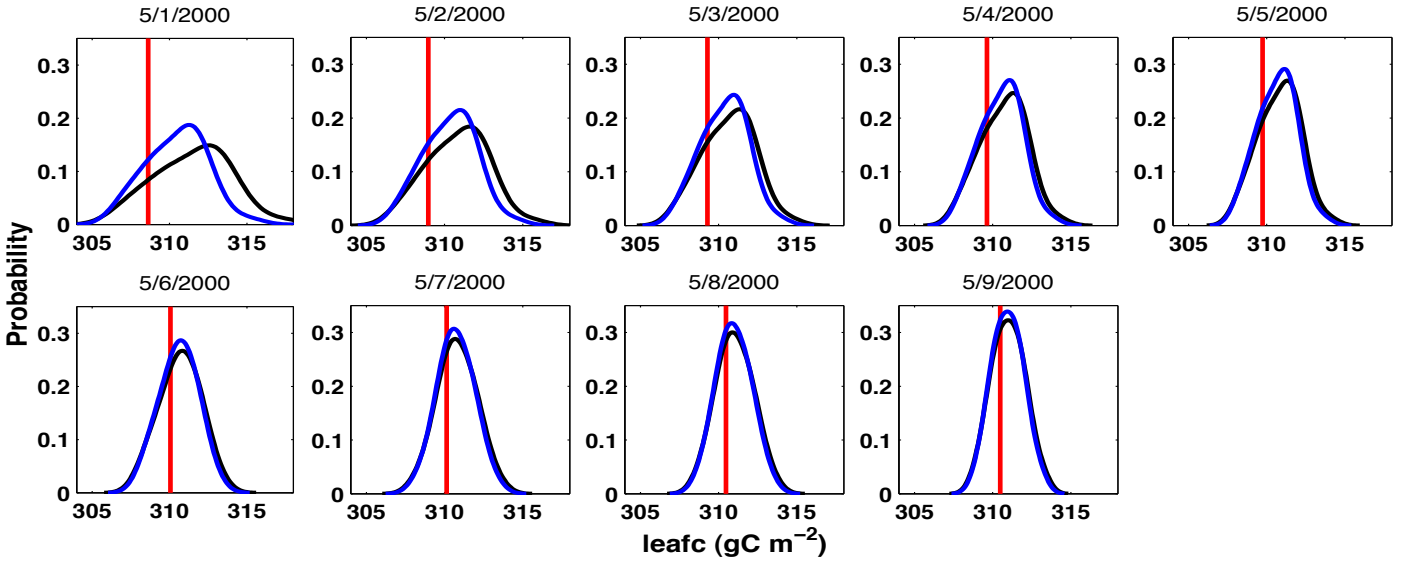
Niwot Ridge

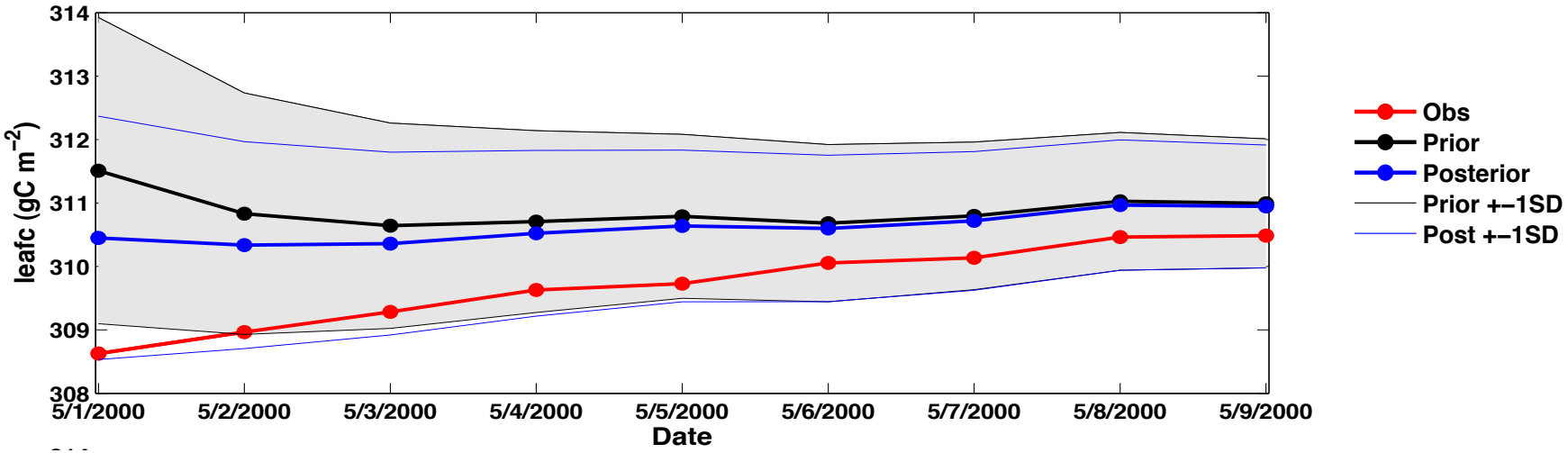

tsa
rh2m
rain
fsds
flds
Days since 1 Jan 1998


nep
tlai
totc
fsh
le
Days since 1 Jan 1998

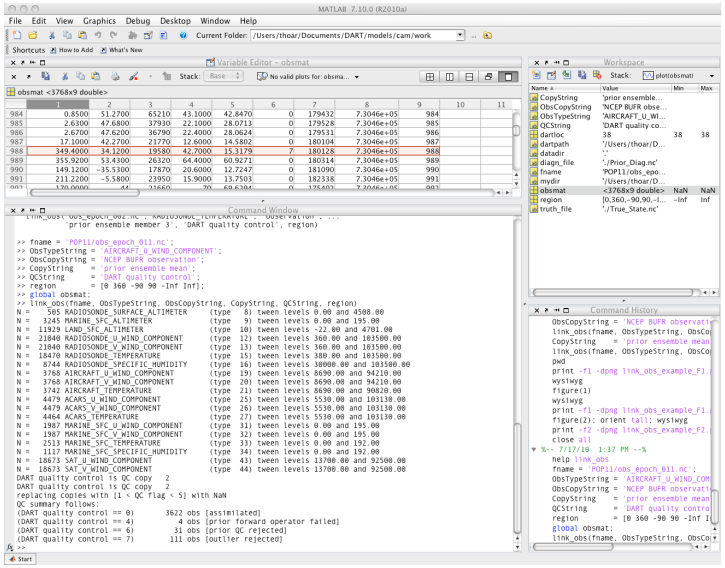
Proof-of-concept with leaf carbon

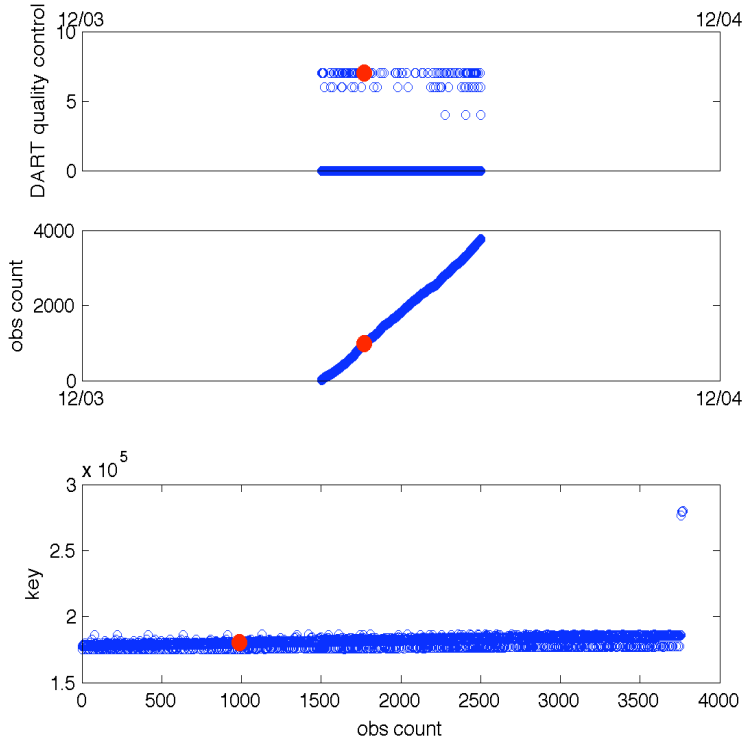
Prior and posterior probability distributions of leaf carbon in a single grid cell at 60°W, 4°S for nine days of assimilation

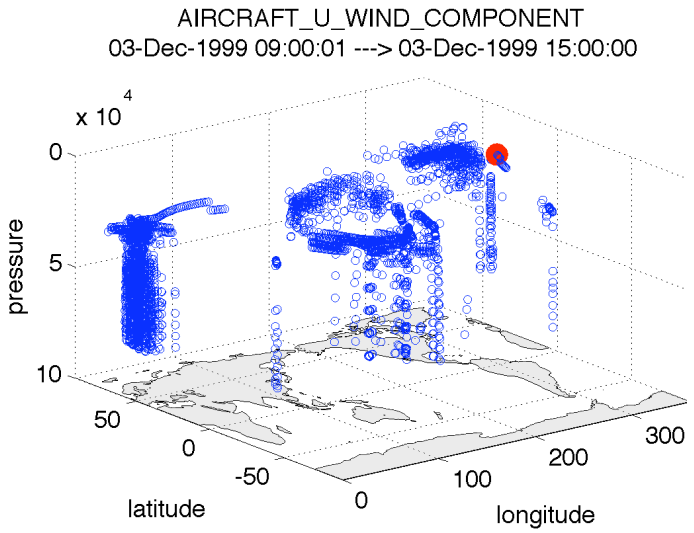


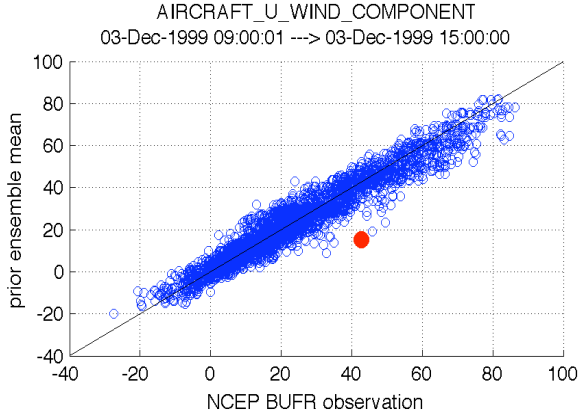


slide held in reserve

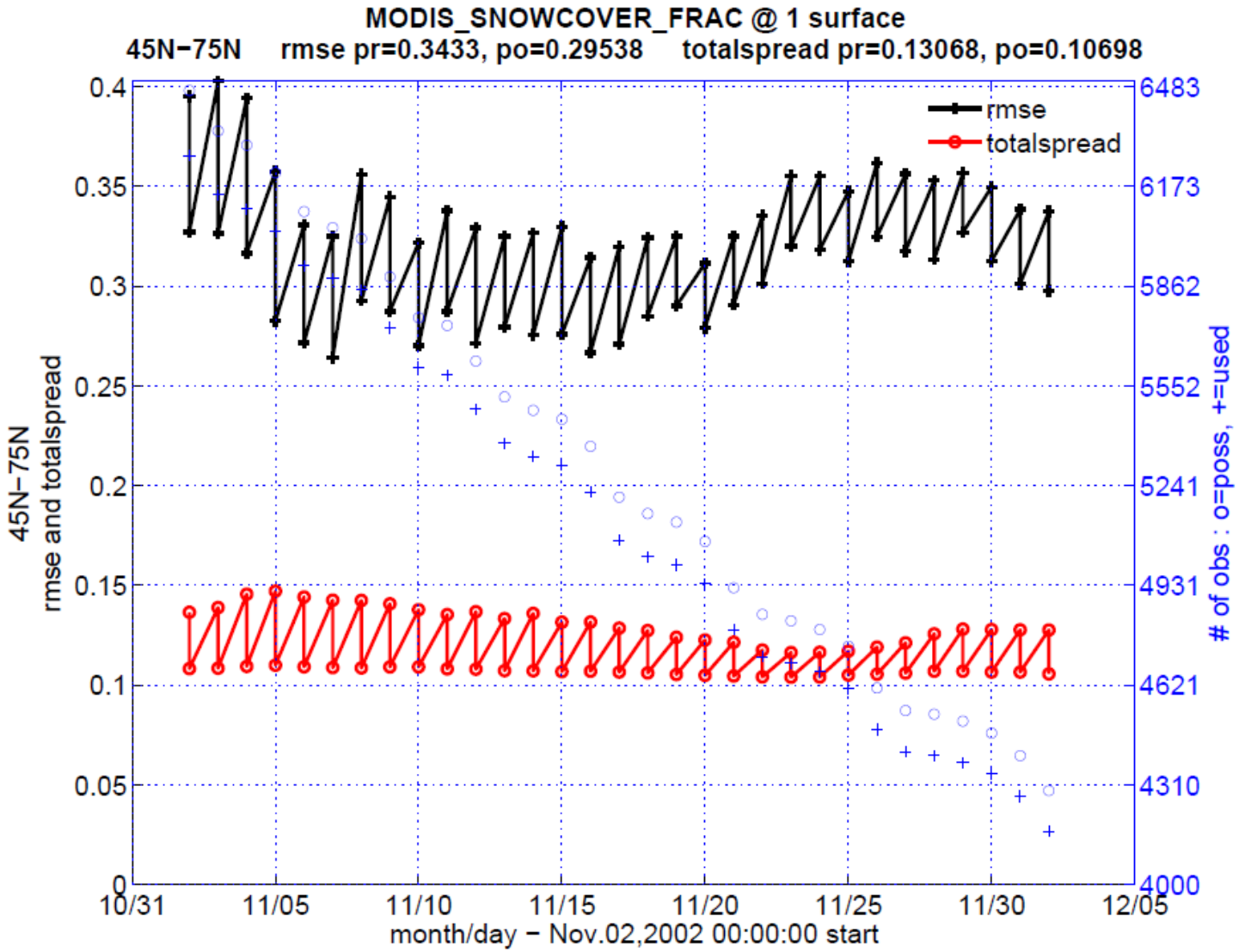








slide held in reserve



Ensemble Filter for Large Geophysical Models

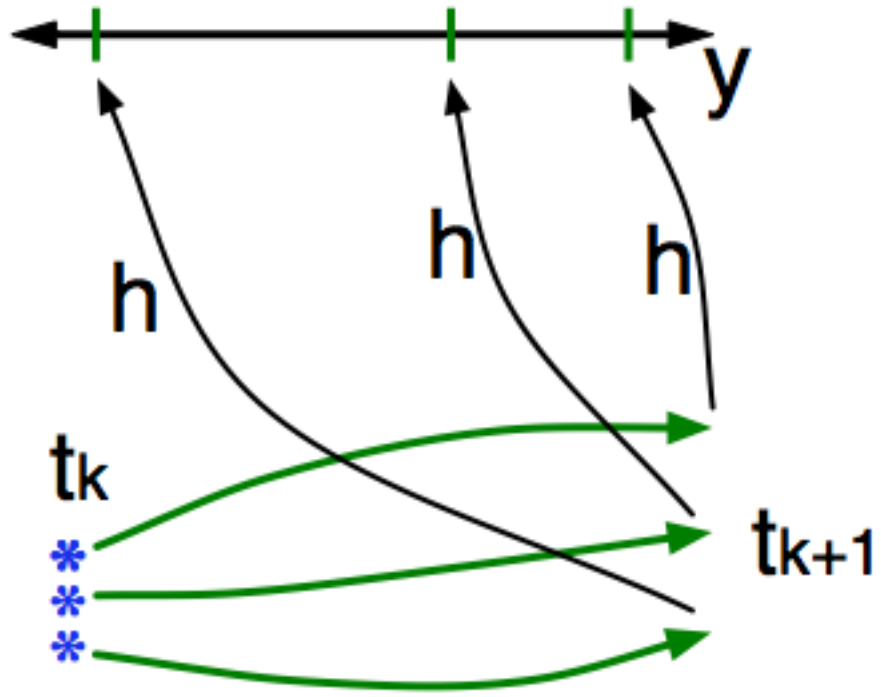
1. Use model to advance ensemble (3 members here) to time at which next observation becomes available.

Ensemble state estimate after using previous observation (analysis)

Ensemble state at time of next observation (prior)

Ensemble Filter for Large Geophysical Models

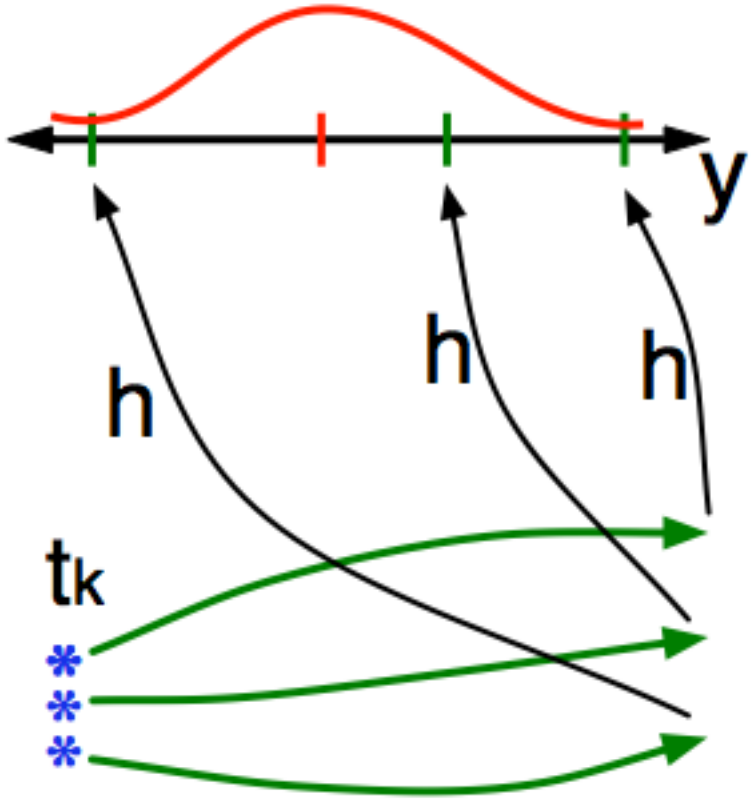
2. Get prior ensemble sample of observation, y = h(x), by applying forward operator **h** to each ensemble member.



Theory: observations from instruments with uncorrelated errors can be done sequentially.

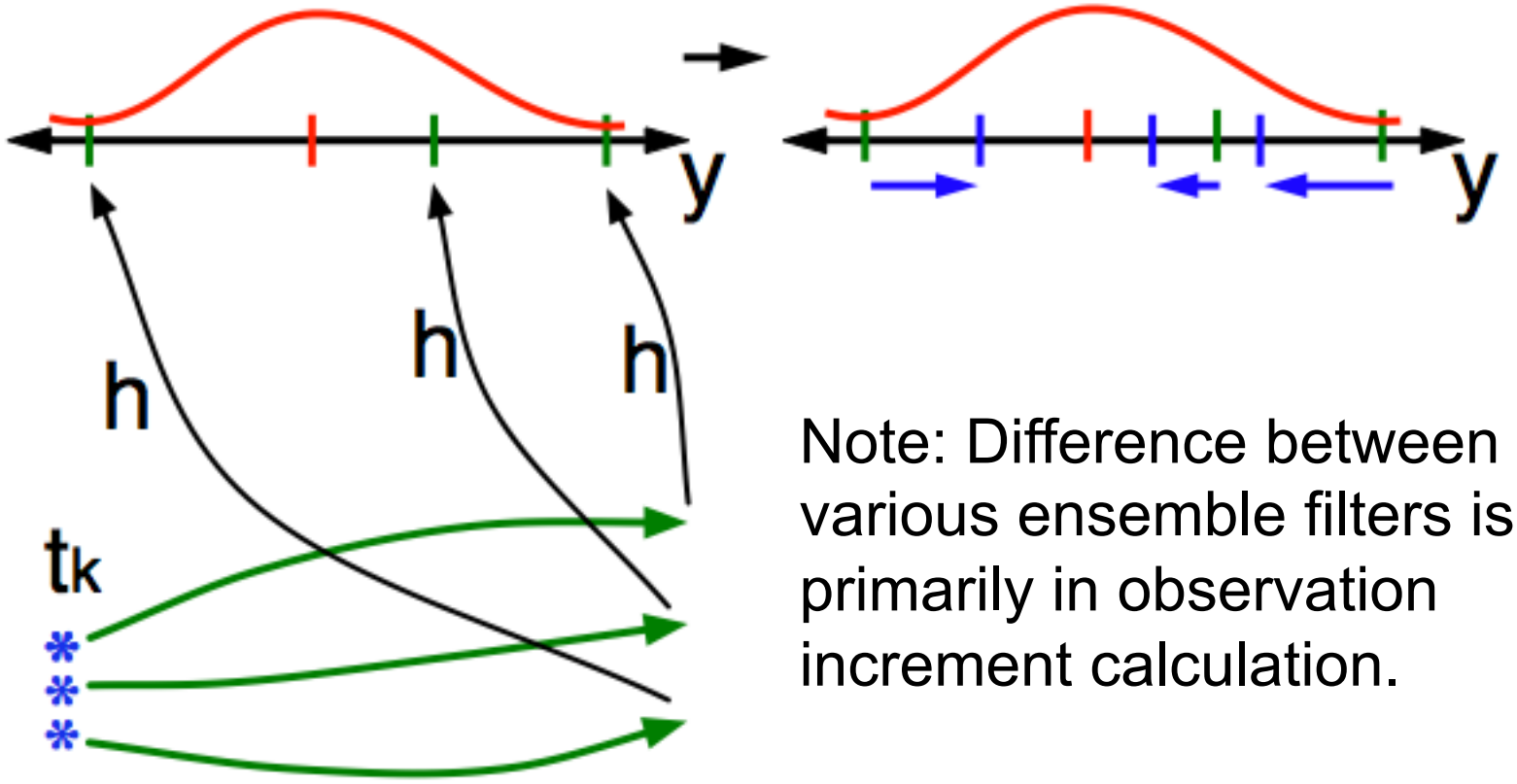
Ensemble Filter for Large Geophysical Models

3. Get observed value and observational error distribution from observing system.



Ensemble Filter for Large Geophysical Models

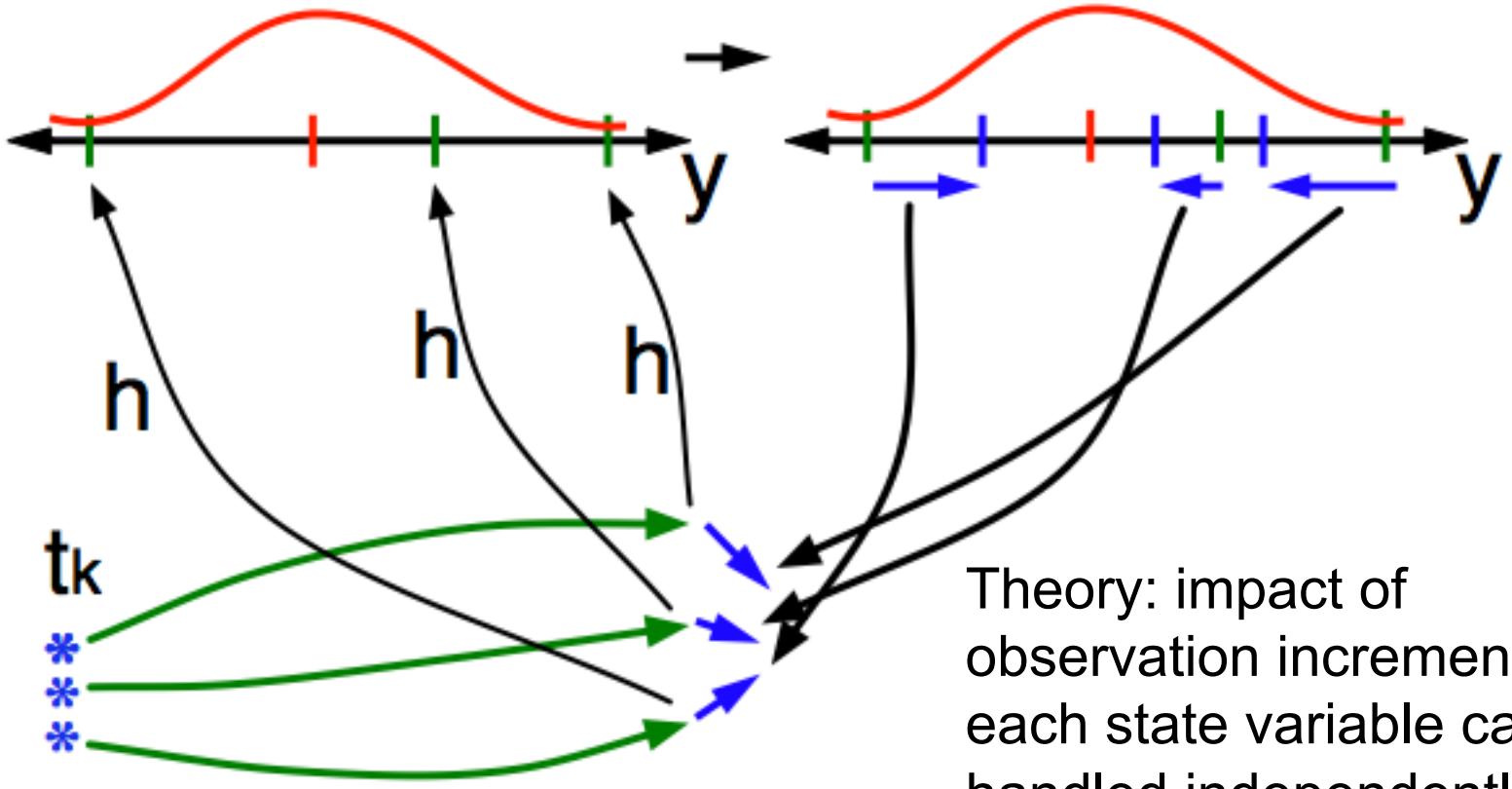
4. Find the increments for the prior observation ensemble (this is a scalar problem for uncorrelated observation errors).



Note: Difference between various ensemble filters is primarily in observation increment calculation.

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 …