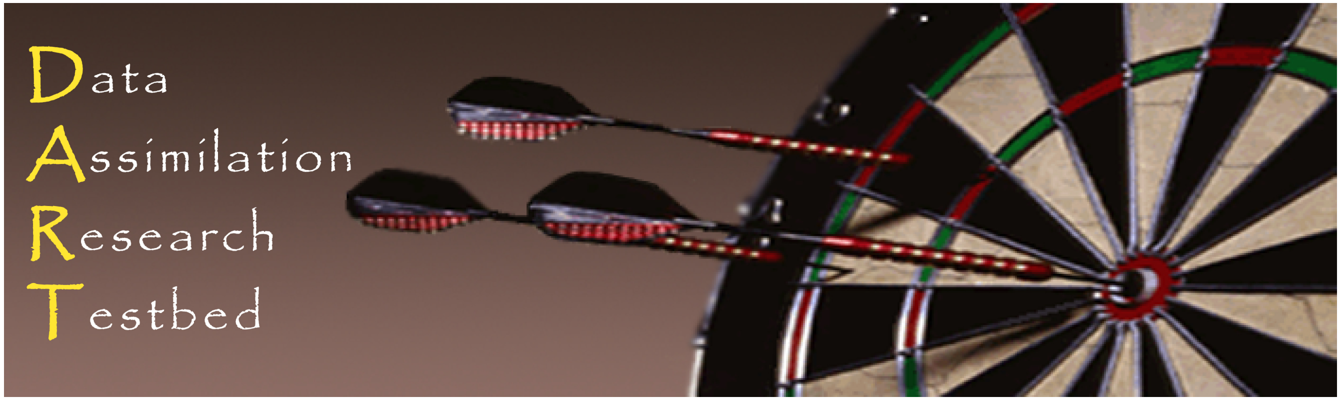
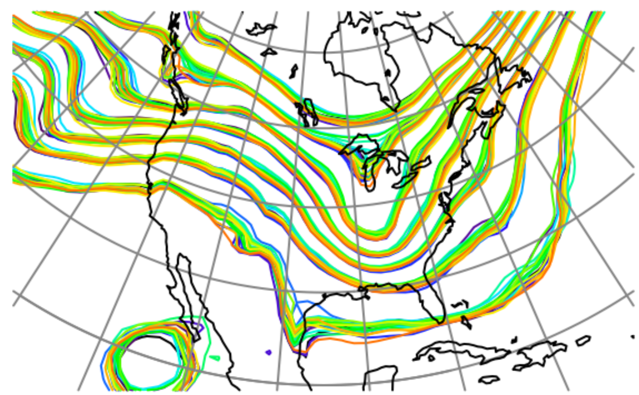


The keys to ensemble data assimilation.

Tim Hoar, Data Assimilation Research Section, NCAR



The National Center for Atmospheric Research is sponsored by the National Science Foundation.
Any opinions, findings and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

NCAR | UCAR National Center for Atmospheric Research

Outline

1. My pet peeve.
2. A brief overview of ensemble assimilation.
3. Why localization and inflation are necessary.
4. Diagnosing what went right.
5. Diagnosing what went wrong.
6. Common mistakes.
7. Some things to think about.
8. Where to learn more.

Motivation

"I spent the last N years developing a method and compared it to an E*KF that I knocked out in a day and -WOW- my method beat the E*KF! It's a ***MIRACLE!*** "

I am simply tired of all the inappropriate comparisons.
I really don't care who wins, just be fair.

At the very least : don't compare this:

Your fully-tested, optimized final product.

To this:

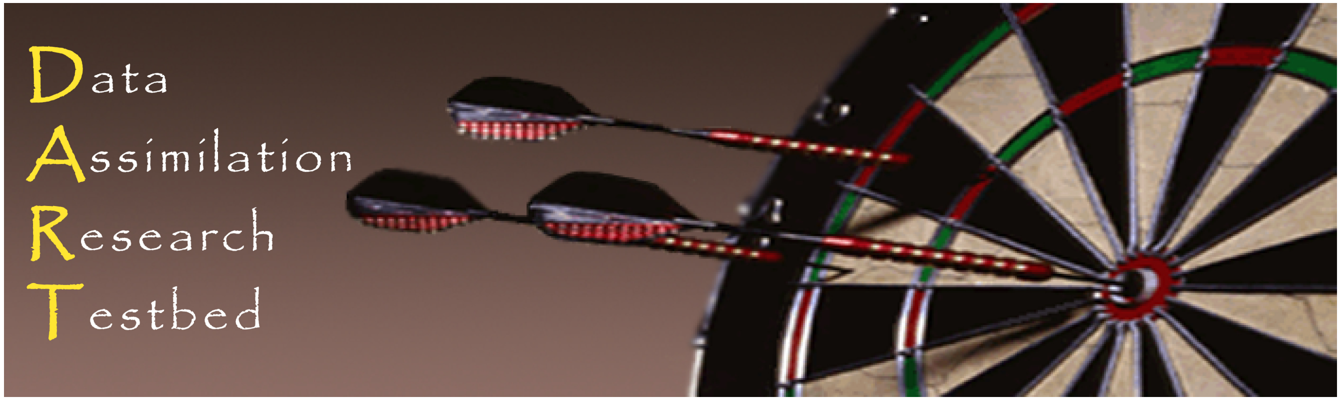
Something full of unrealized potential.

Or even more disheartening:

Don’t compare this to this.

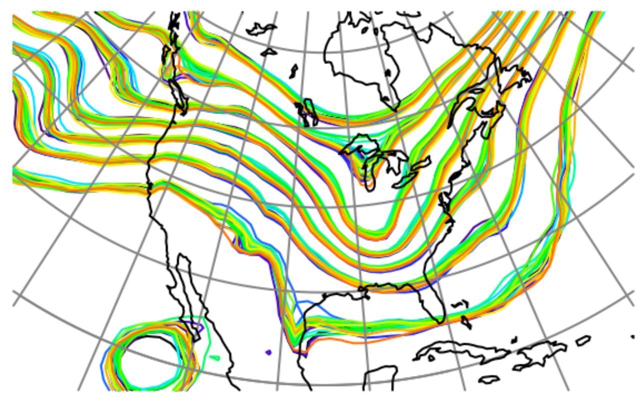
It is possible to sabotage (even unintentionally) a method to produce poor results.

Sadly, it happens!



successful
The keys to ensemble data assimilation.

Tim Hoar, Data Assimilation Research Section, NCAR

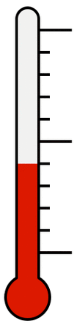


The National Center for Atmospheric Research is sponsored by the National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

NCAR | National Center for
UCAR | Atmospheric Research

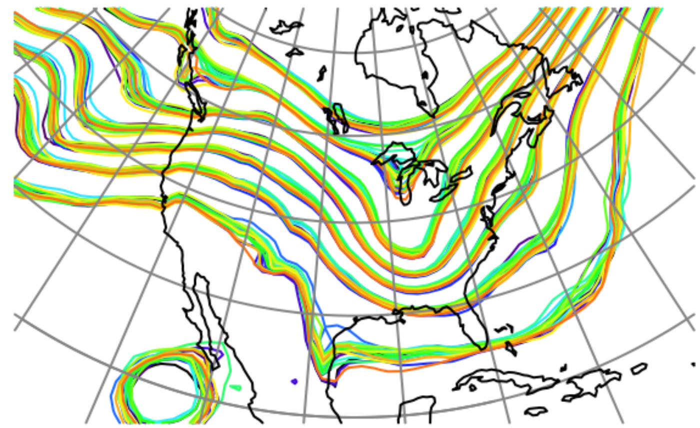
What is Data Assimilation?

Observations combined with a Model forecast…



+

=



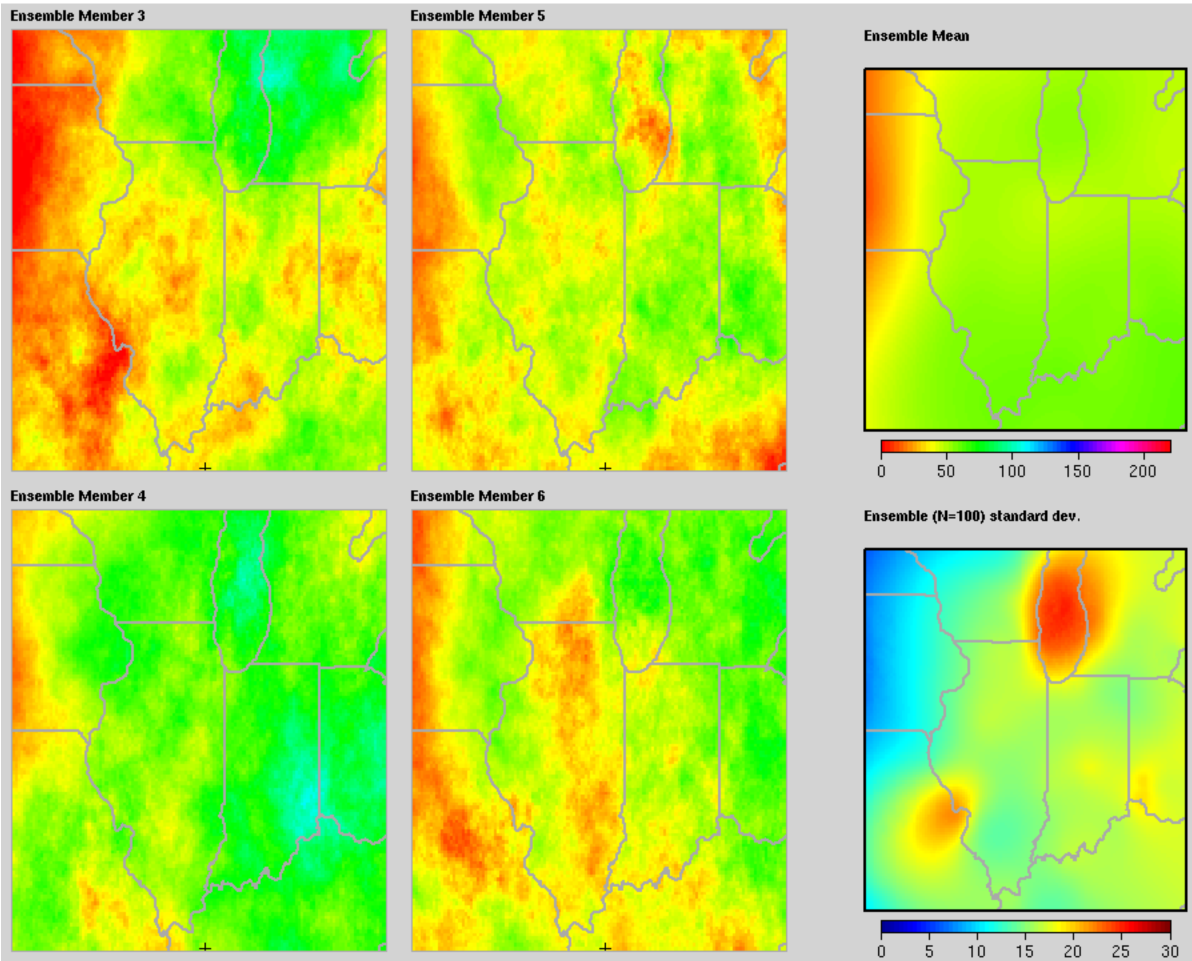
… to produce an analysis.

Overview article of the Data Assimilation Research Testbed (DART):

Anderson, Jeffrey, T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, A. Arellano, 2009: The Data Assimilation Research Testbed: A Community Facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296. doi:10.1175/2009BAMS2618.1

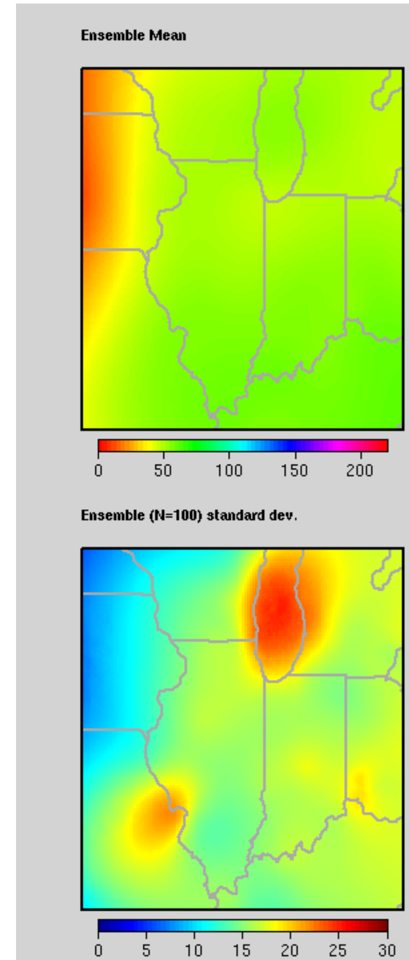
Ozone fields example

4 estimates of Ozone – all equally likely.



Mean of 100 estimates.

Variability of 100 estimates.



Schematic of an Ensemble Filter for Geophysical Data Assimilation

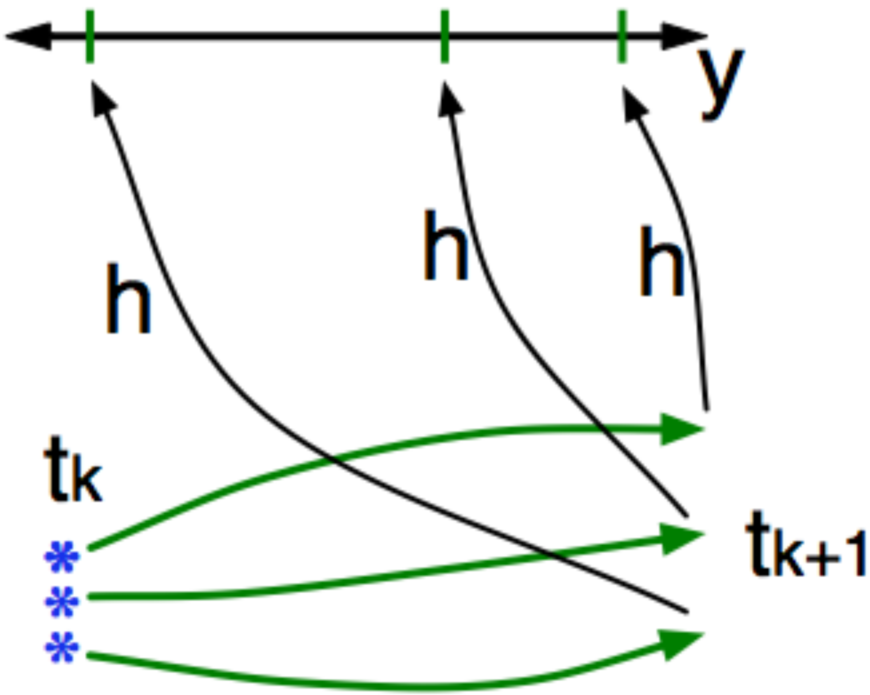
1. Use model to advance ensemble (3 members here) to time at which next observation becomes available.

Ensemble state estimate after using previous observation (analysis)

Ensemble state at time of next observation (prior)

Schematic of an Ensemble Filter for Geophysical Data Assimilation

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

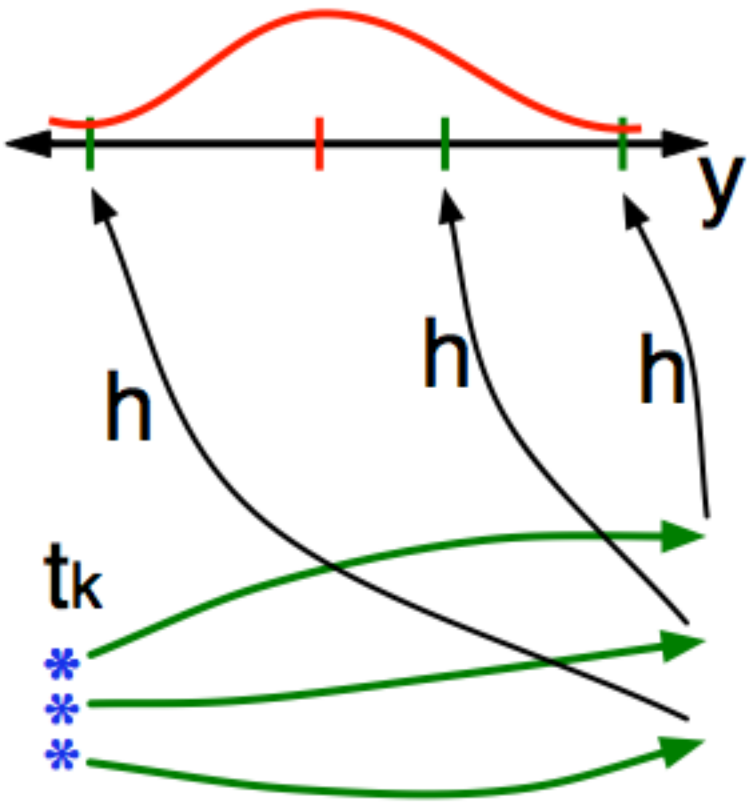


Theory: observations from instruments with uncorrelated errors can be done sequentially.

Houtekamer, P.L. and H.L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123-137

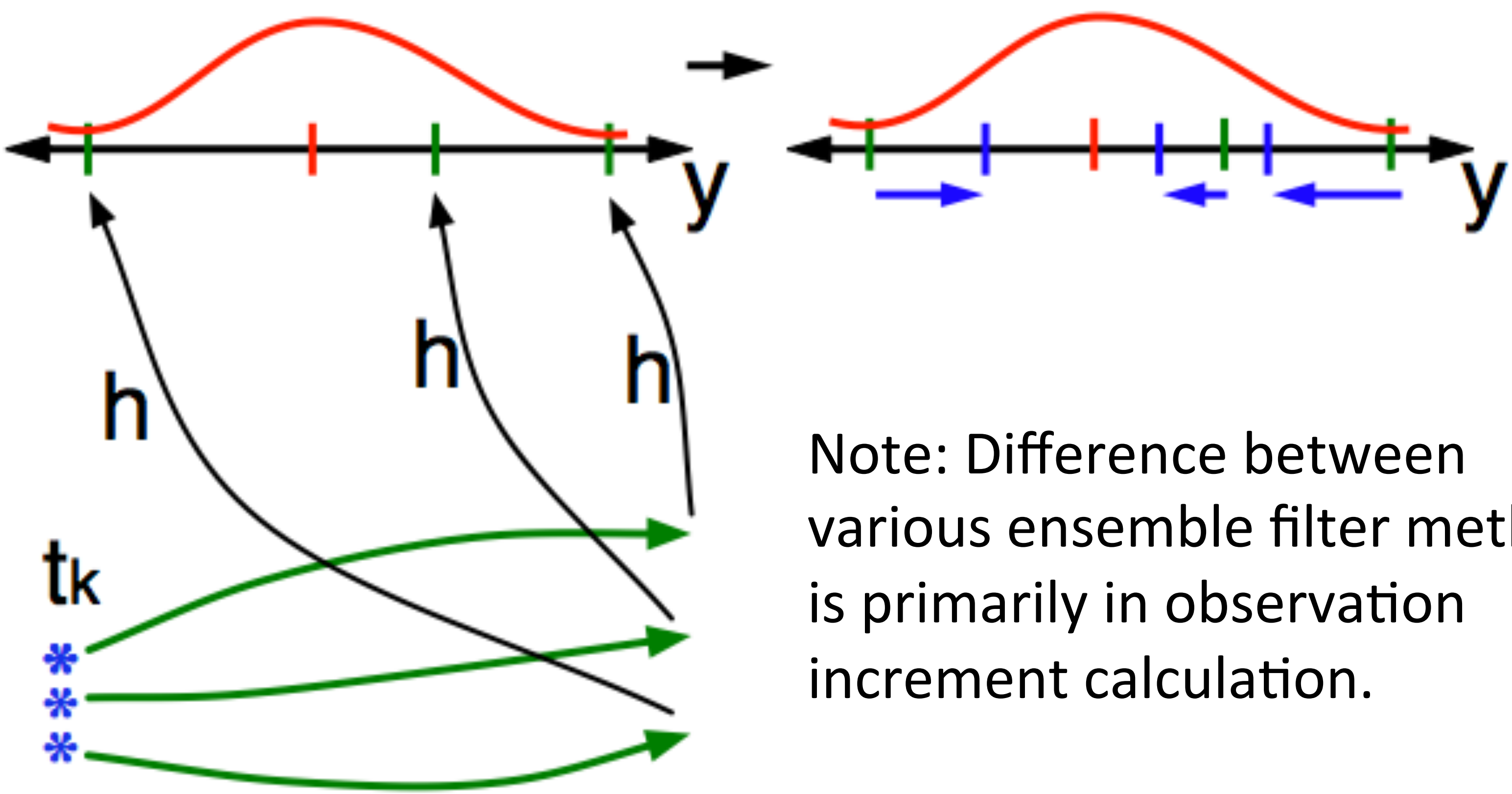
Schematic of an Ensemble Filter for Geophysical Data Assimilation

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



Schematic of an Ensemble Filter for Geophysical Data Assimilation

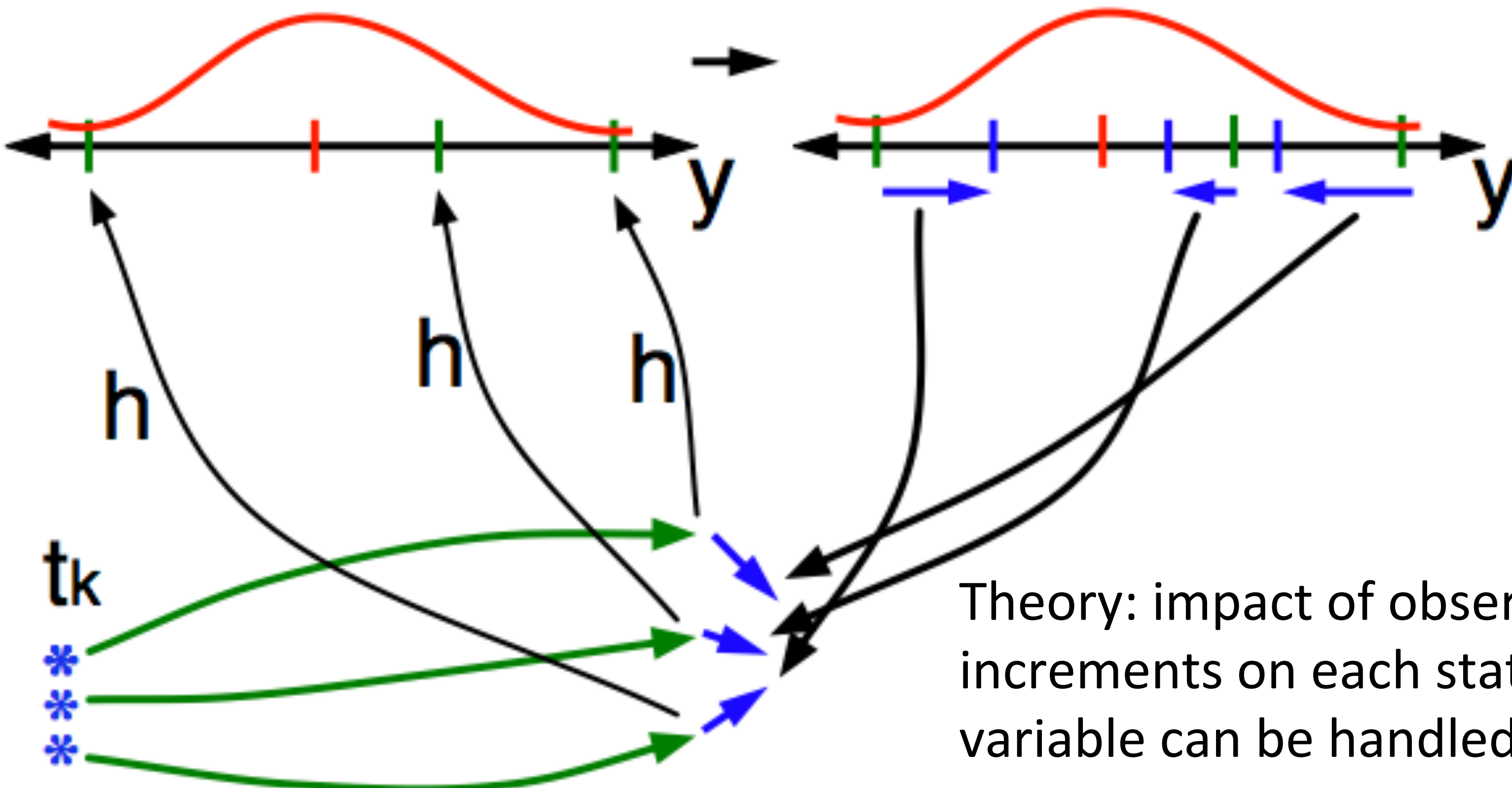
4. Find the increments for the prior observation ensemble (this is a scalar problem for uncorrelated observation errors).



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

Schematic of an Ensemble Filter for Geophysical Data Assimilation

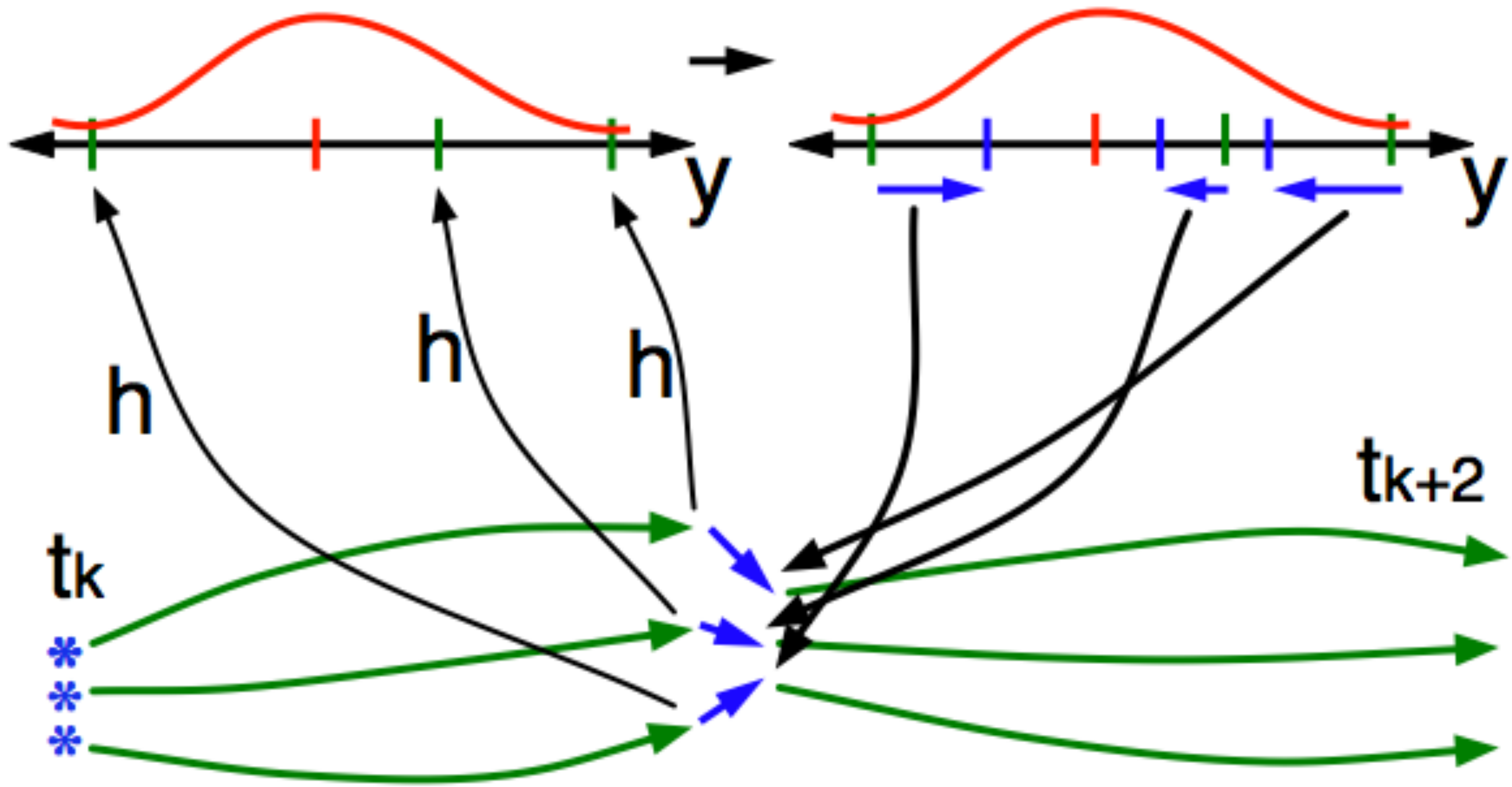
5. Use ensemble samples of y and each state variable to linearly regress observation increments onto state variable increments.



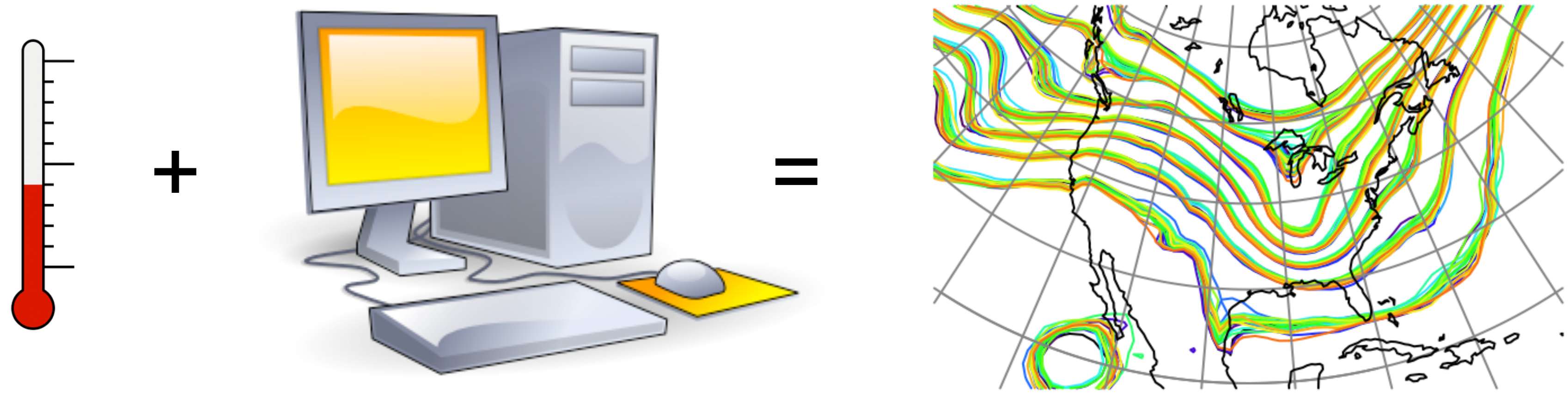
Theory: impact of observation increments on each state variable can be handled independently!

Schematic of an Ensemble Filter for Geophysical Data Assimilation

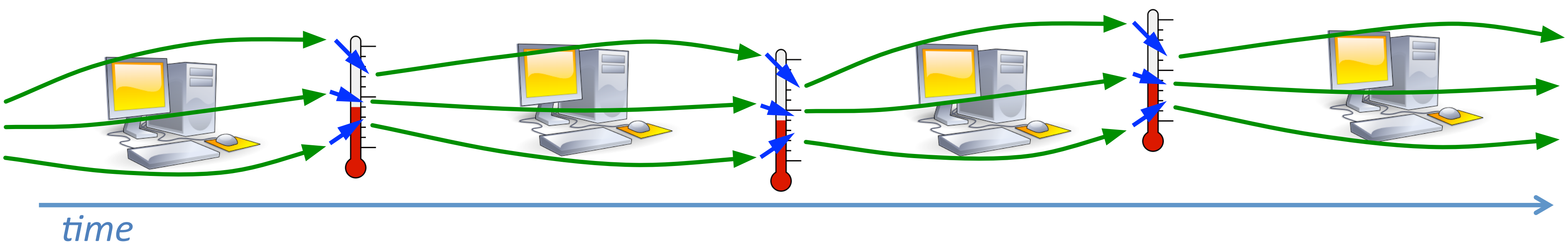
6. When all ensemble members for each state variable are updated, there is a new analysis. Integrate to time of next observation …



Once is not enough!



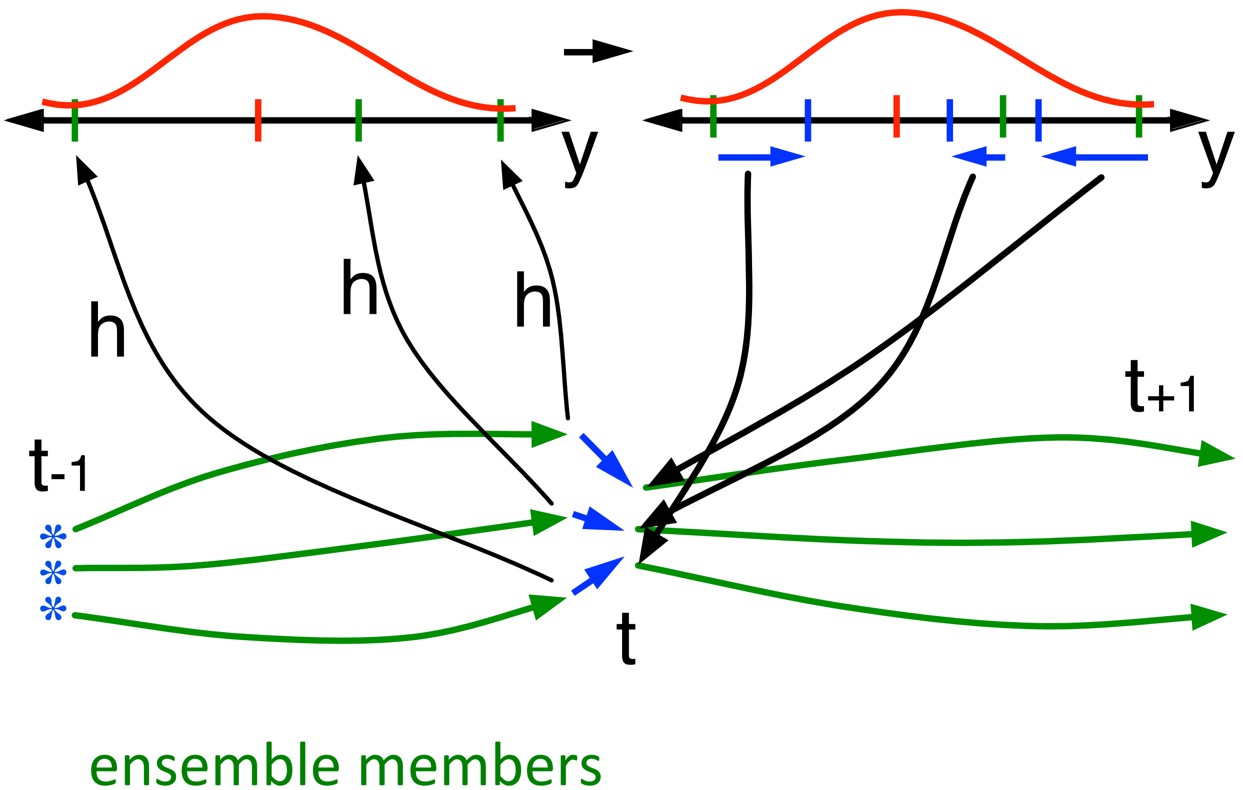
We want to assimilate **over and over** to steadily make the model states more consistent with the observations.



I used to know what *'coupled'* data assimilation meant.
I don't anymore. Ditto for '*hybrid*' methods.

A generic ensemble filter system like DART needs:

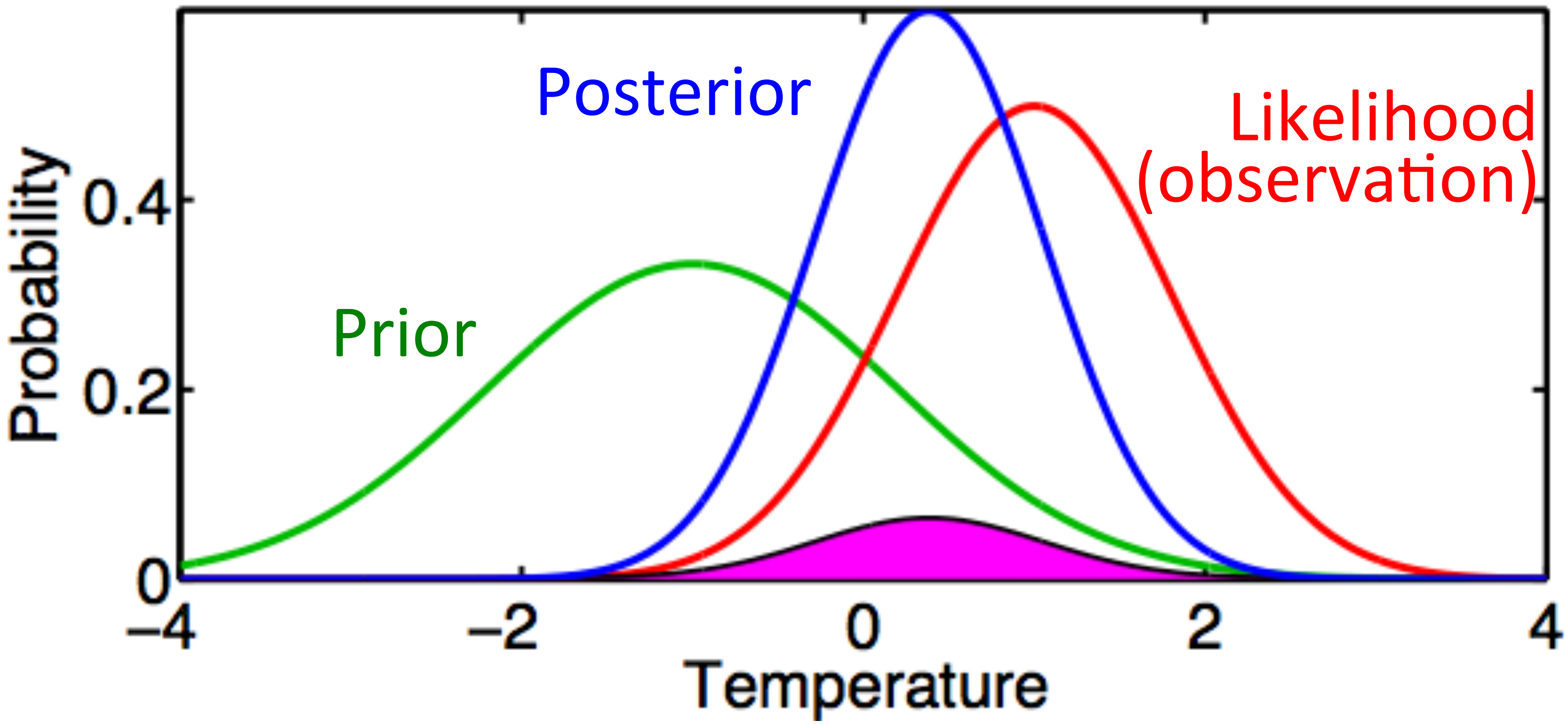
1. A way to make model forecasts.
2. A way to estimate what the observation would be – given the model state. This is the forward observation operator – h.



The **increments** are regressed onto as many **state variables** as you like. If there is a correlation, the state gets adjusted. The new states are used as new initial conditions.

Combining the Prior Estimate and Observation

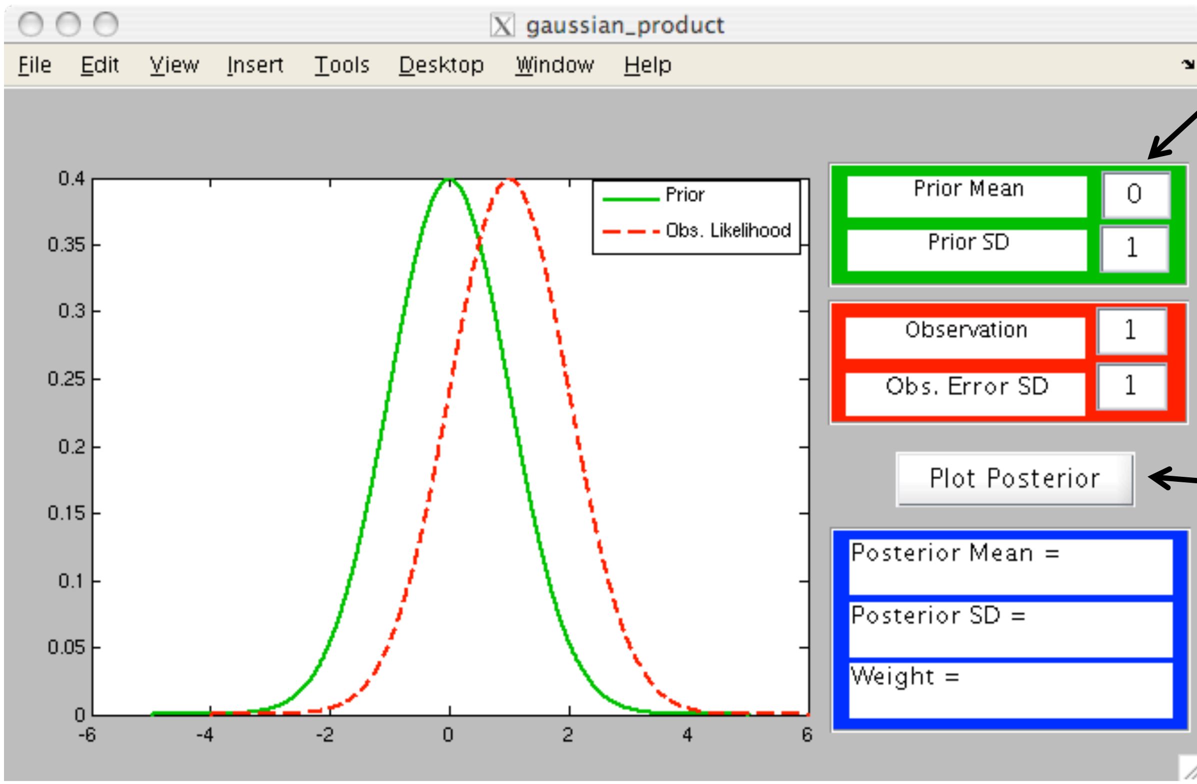
$$P\left(T \mid T_0, C\right) = \frac{P(T_0 \mid T, C)P(T \mid C)}{normalization}$$



The example here shows gaussians, not required …

Matlab Hands-on: gaussian_product

Purpose: Explore the gaussian posterior that results from taking the product of a gaussian prior and a gaussian likelihood.



1) Set Prior Mean and Standard Deviation.

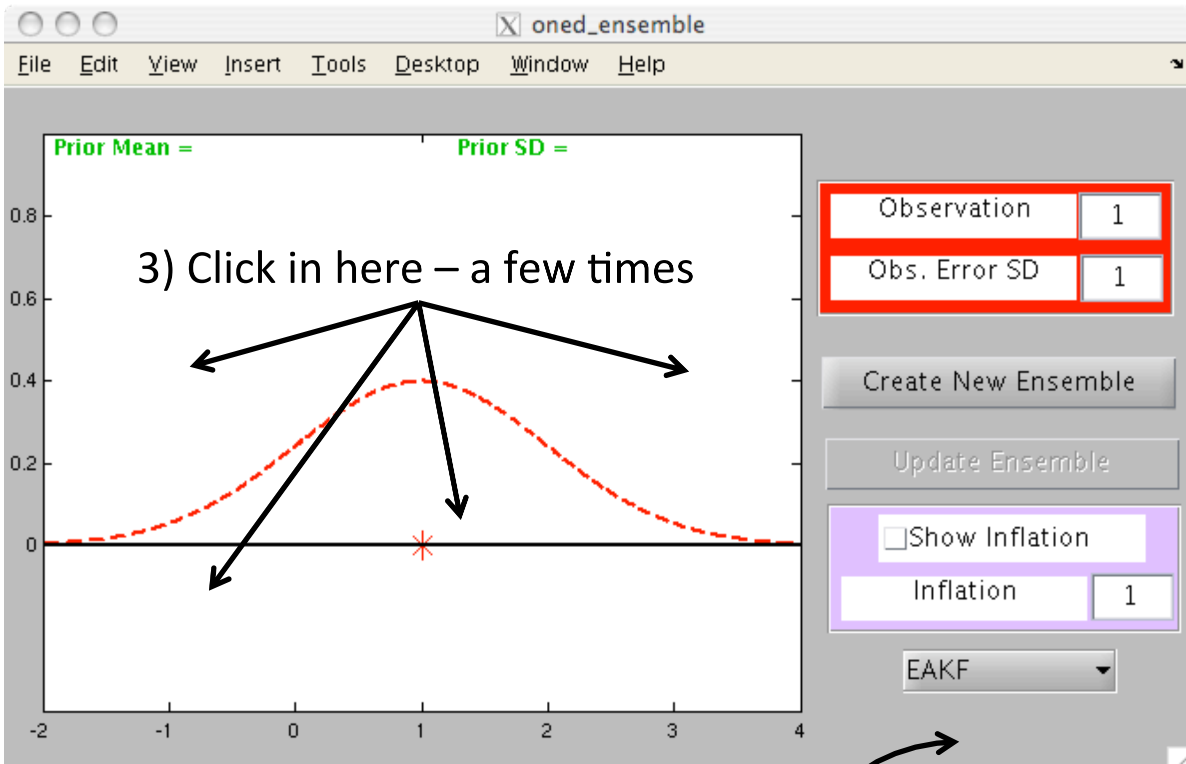
2) Set Observation Mean and Observation Error Standard Deviation.

3) Select Plot Posterior to Update the items in blue.

Matlab Hands-On: oned_ensemble

Matlab GUI **oned_ensemble** demonstrates how the increments are calculated.

Purpose: Explore how ensemble filters update a prior ensemble.



1) change these if you want to.

2) Click on Create New Ensemble

5) Click on Update Ensemble

4) Click outside the axis on the gray (anywhere) to finish defining the ensemble.

Ignore the Inflation and EAKF menus for now.

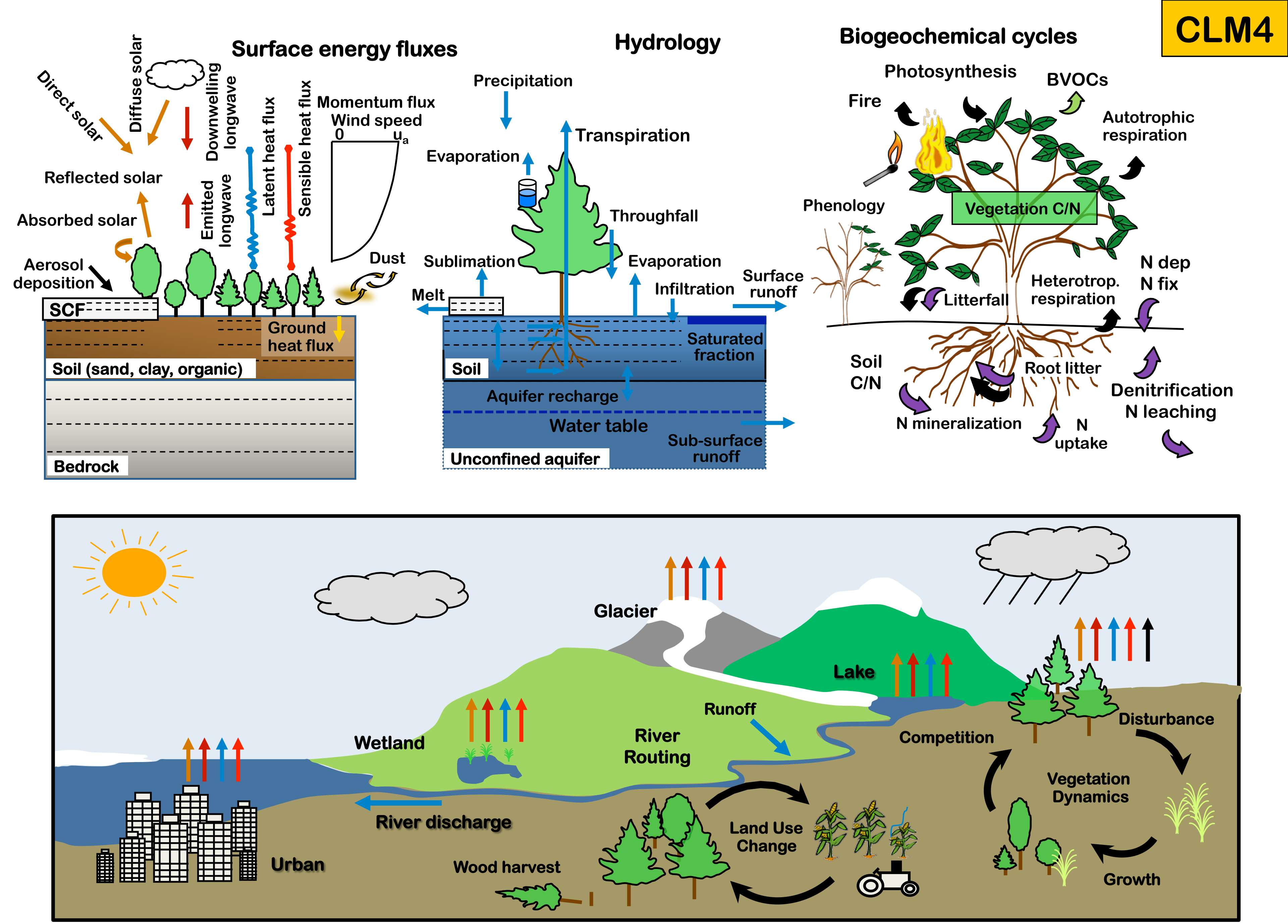
OK – so now we have increments …

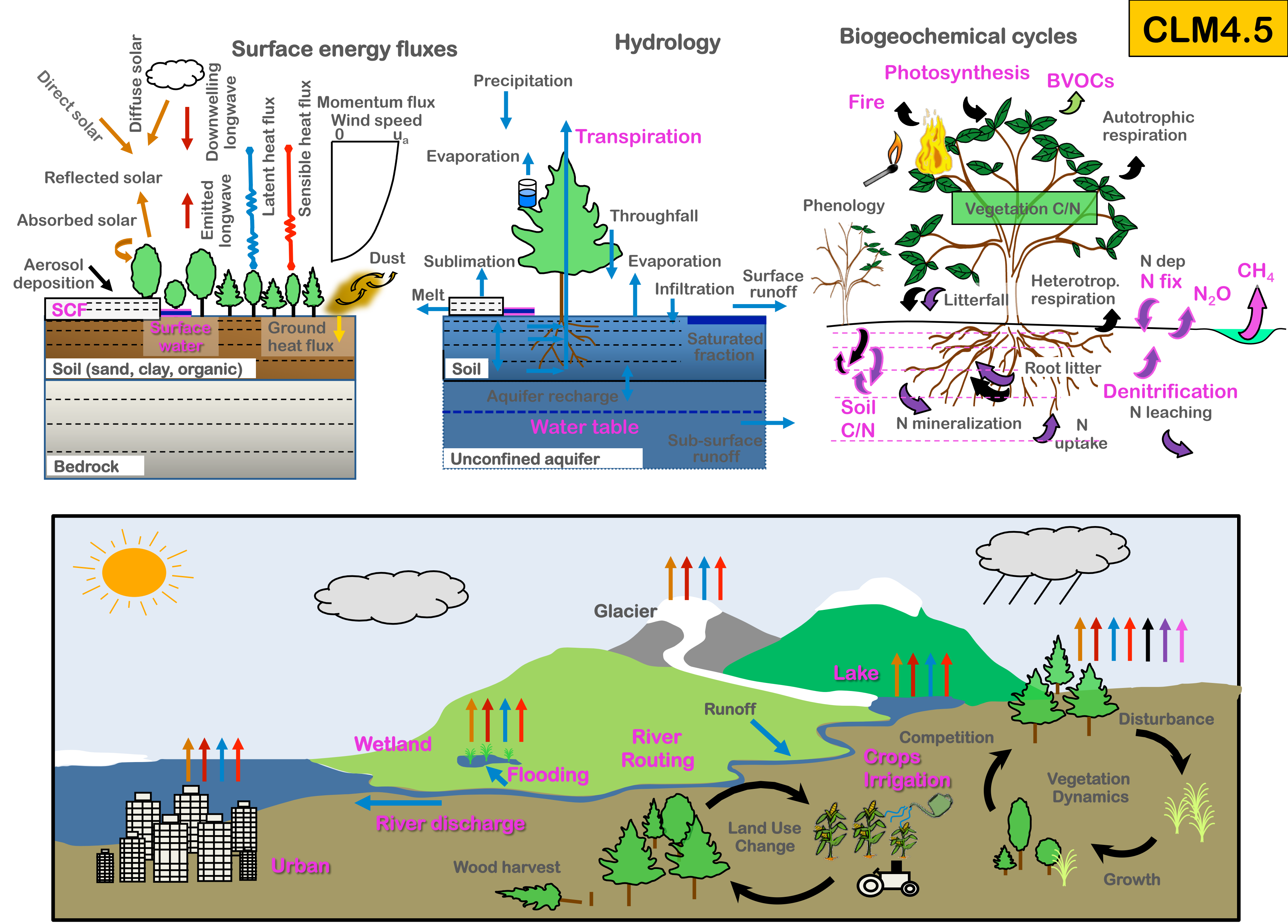
We need to know how to use the increments.
"We regress them onto the model state."

Time for a quick tour of
DART/DART_LAB/DART_LAB.html

1. Concepts in 1D
2. What can the increments impact?
3. What ***should*** the increments impact?

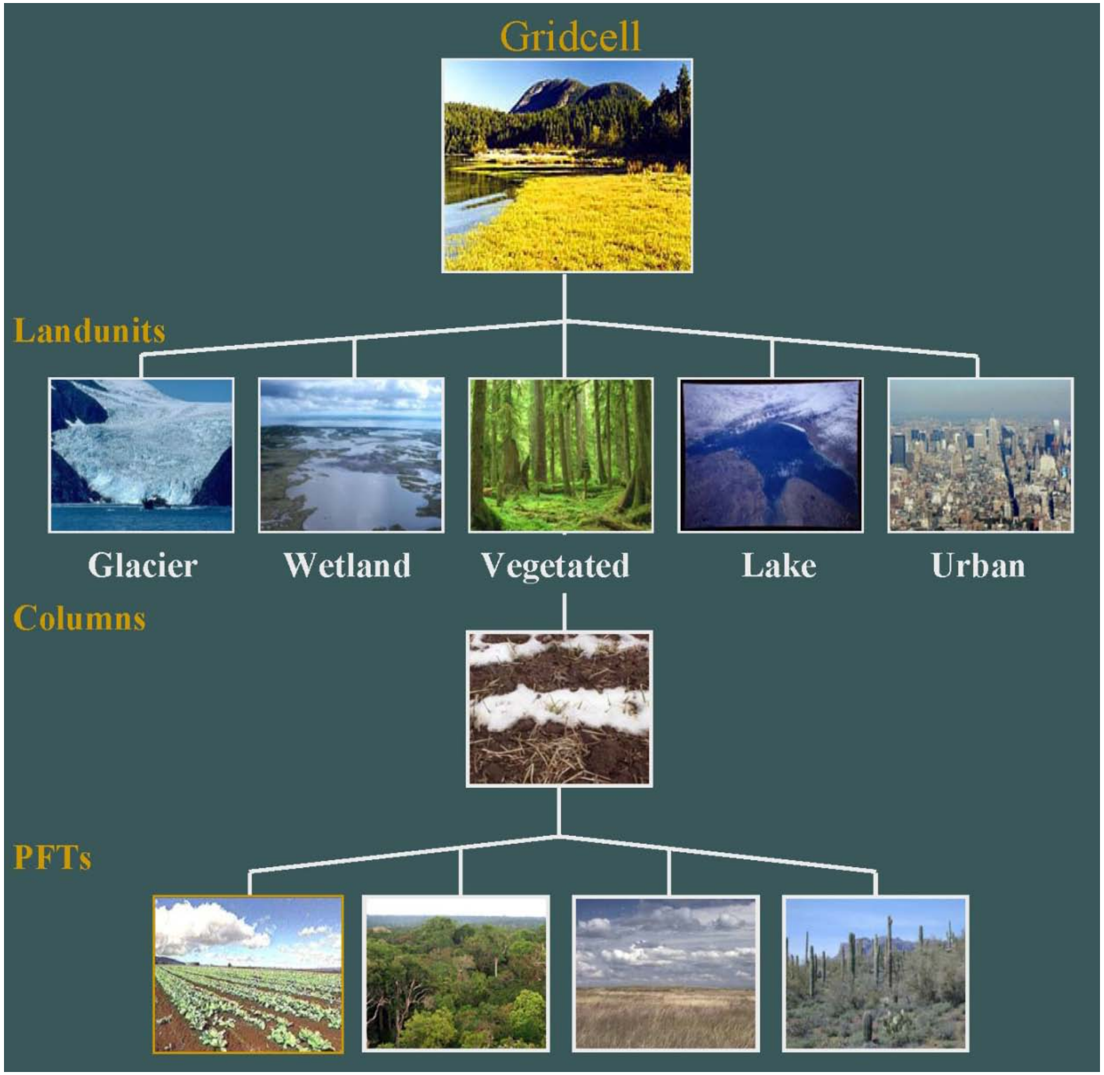
The next slide shows some of the processes in the Community Land Model.
There are more than 200 variables at each gridpoint. What do you do?


CLM4
Surface energy fluxes
Direct solar
Diffuse solar
Downwelling longwave
Reflected solar
Absorbed solar
Emitted longwave
Latent heat flux
Sensible heat flux
Momentum flux
Wind speed
0
u_a
Aerosol deposition
Dust
SCF
Ground heat flux
Soil (sand, clay, organic)
Bedrock
Hydrology
Precipitation
Transpiration
Evaporation
Throughfall
Sublimation
Evaporation
Melt
Infiltration
Surface runoff
Saturated fraction
Soil
Aquifer recharge
Water table
Sub-surface runoff
Unconfined aquifer
Biogeochemical cycles
Photosynthesis
BVOCs
Fire
Autotrophic respiration
Phenology
Vegetation C/N
Litterfall
Heterotrop. respiration
N dep
N fix
Soil C/N
Root litter
Denitrification
N leaching
N mineralization
N uptake
Glacier
Lake
Wetland
Runoff
River Routing
Competition
Disturbance
Vegetation Dynamics
Growth
River discharge
Land Use Change
Urban
Wood harvest



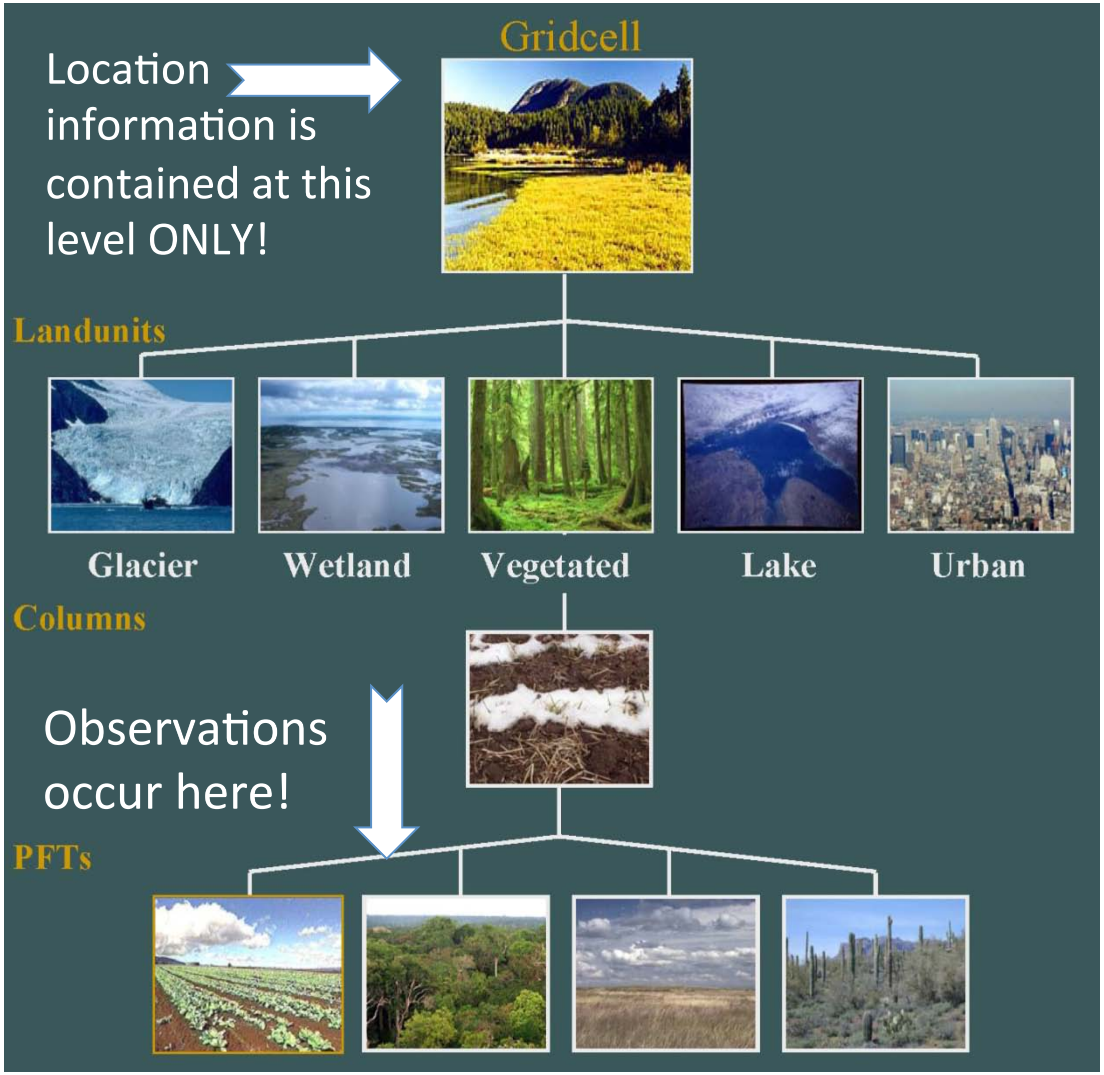
CLM4.5
Surface energy fluxes
Direct solar
Diffuse solar
Downwelling longwave
Latent heat flux
Sensible heat flux
Momentum flux
Wind speed
0
u_a
Reflected solar
Absorbed solar
Emitted longwave
Aerosol deposition
Dust
SCF
Surface water
Ground heat flux
Soil (sand, clay, organic)
Bedrock
Hydrology
Precipitation
Transpiration
Evaporation
Throughfall
Sublimation
Evaporation
Melt
Infiltration
Surface runoff
Saturated fraction
Soil
Aquifer recharge
Water table
Sub-surface runoff
Unconfined aquifer
Biogeochemical cycles
Photosynthesis
BVOCs
Fire
Autotrophic respiration
Phenology
Vegetation C/N
Litterfall
Heterotrop. respiration
N dep
N fix
N_2O
CH_4
Root litter
Soil C/N
N mineralization
N uptake
Denitrification
N leaching
Glacier
Lake
Runoff
Wetland
River Routing
Flooding
Competition
Crops
Irrigation
Disturbance
Vegetation Dynamics
River discharge
Land Use Change
Urban
Wood harvest
Growth

As if it weren't complicated enough …



Models that abstract the gridcell into a "nested gridcell hiearchy of of multiple landunits, snow/soil columns, and Plant Function Types" are particularly troublesome when trying to convert the model state to the expected observation value.

As if it weren't complicated enough …

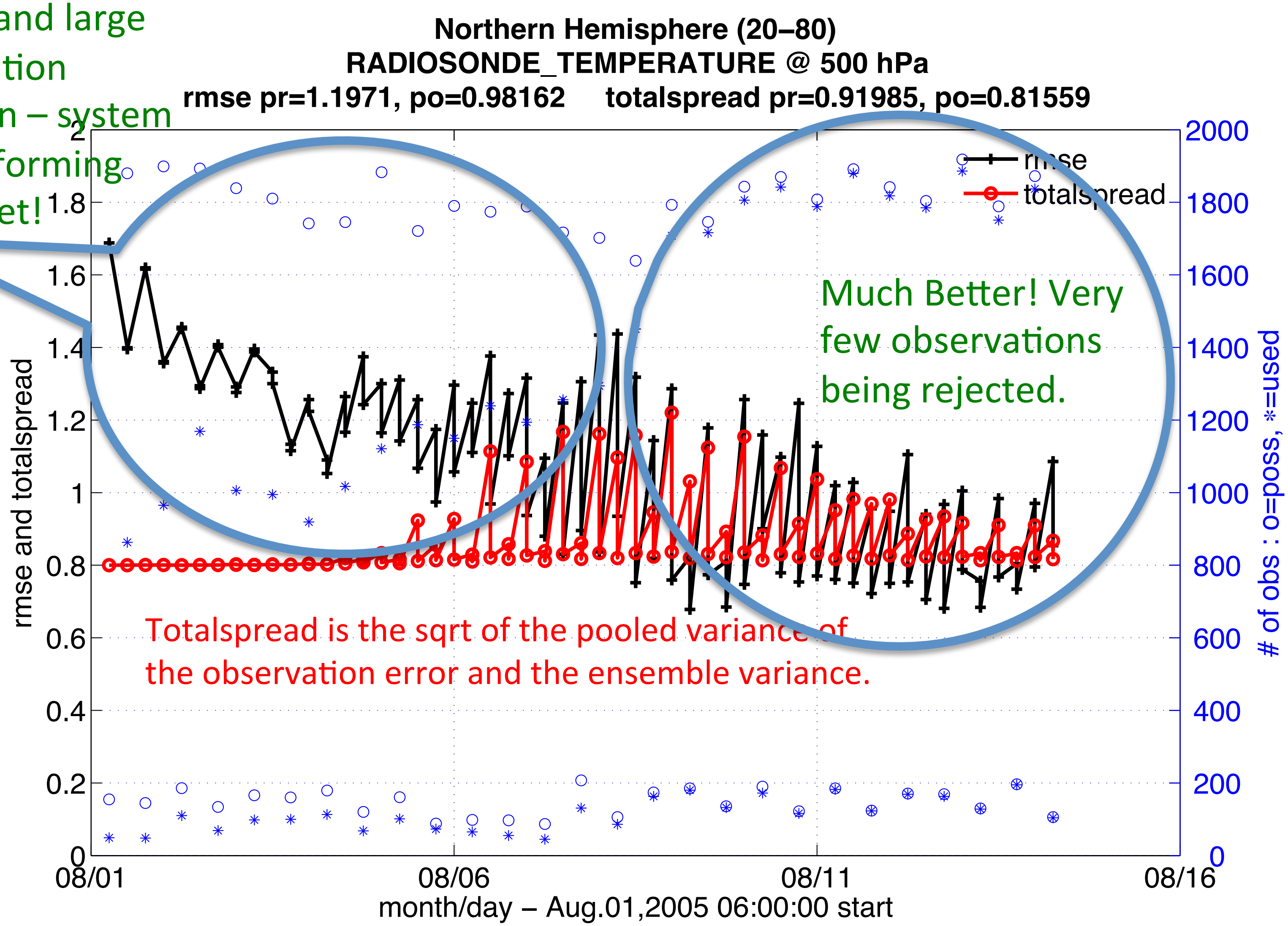


Models that abstract the gridcell into a "nested gridcell hiearchy of of multiple landunits, snow/soil columns, and Plant Function Types" are particularly troublesome when trying to convert the model state to the expected observation value.

Given a soil temperature observation at a specific lat/lon, which PFT did it come from? **No way to know!** *Unless obs have more metadata!*

Performance and Rejection

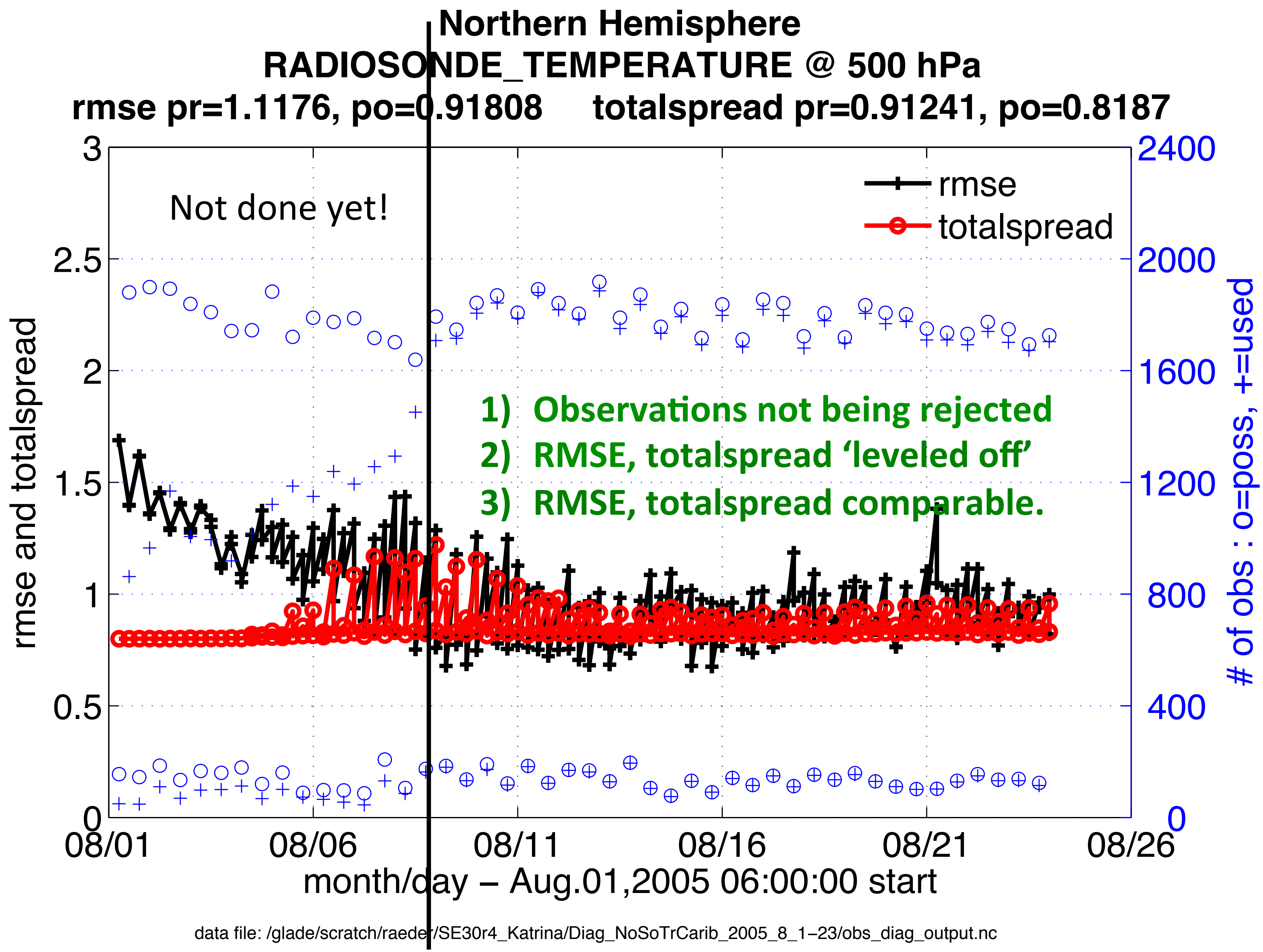
Initially tiny spread and large observation rejection – system not performing well – yet!



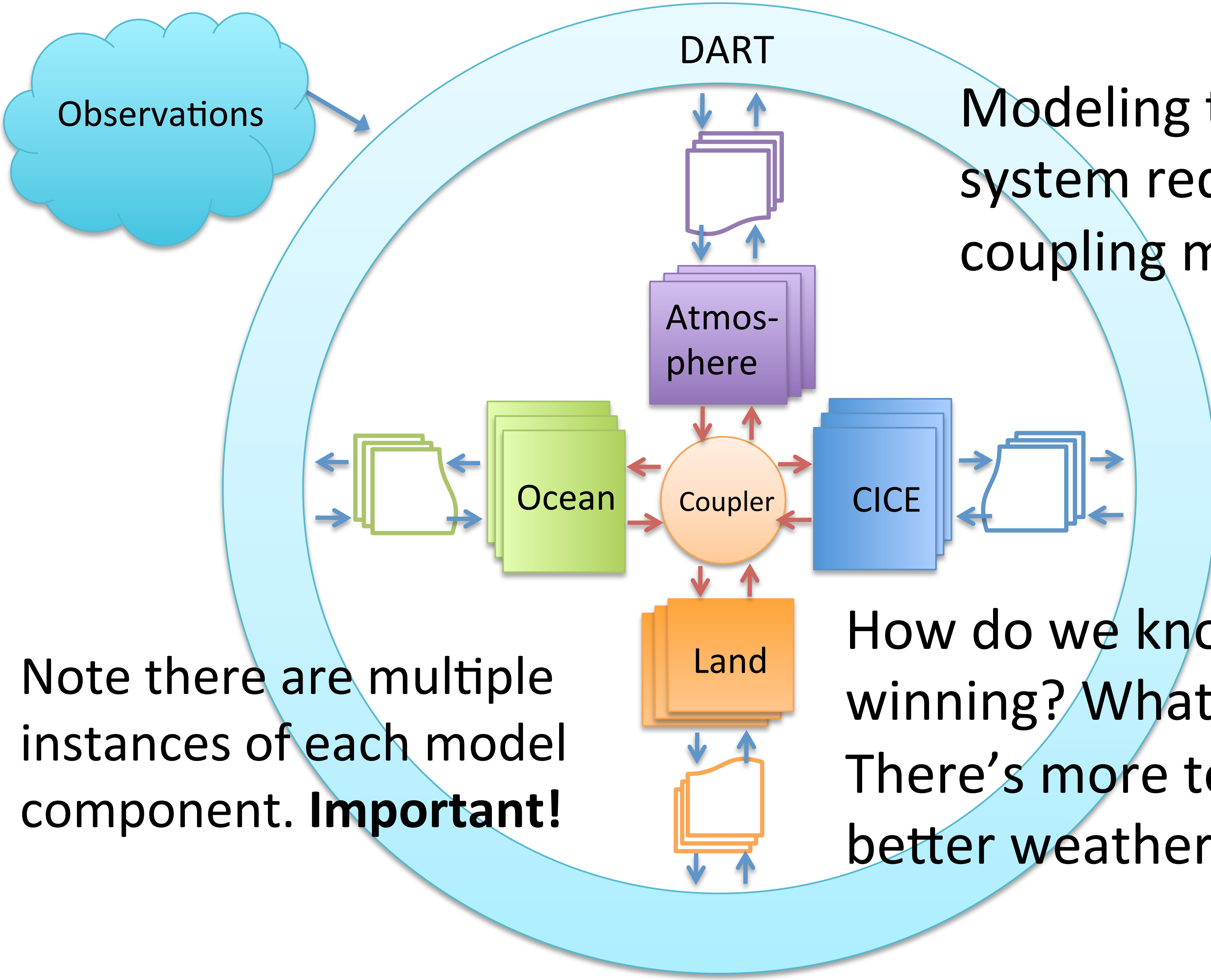
Much Better! Very few observations being rejected.

Totalspread is the sqrt of the pooled variance of the observation error and the ensemble variance.

A good-looking experiment.



Sometimes the models are *PRETTY COMPLEX*

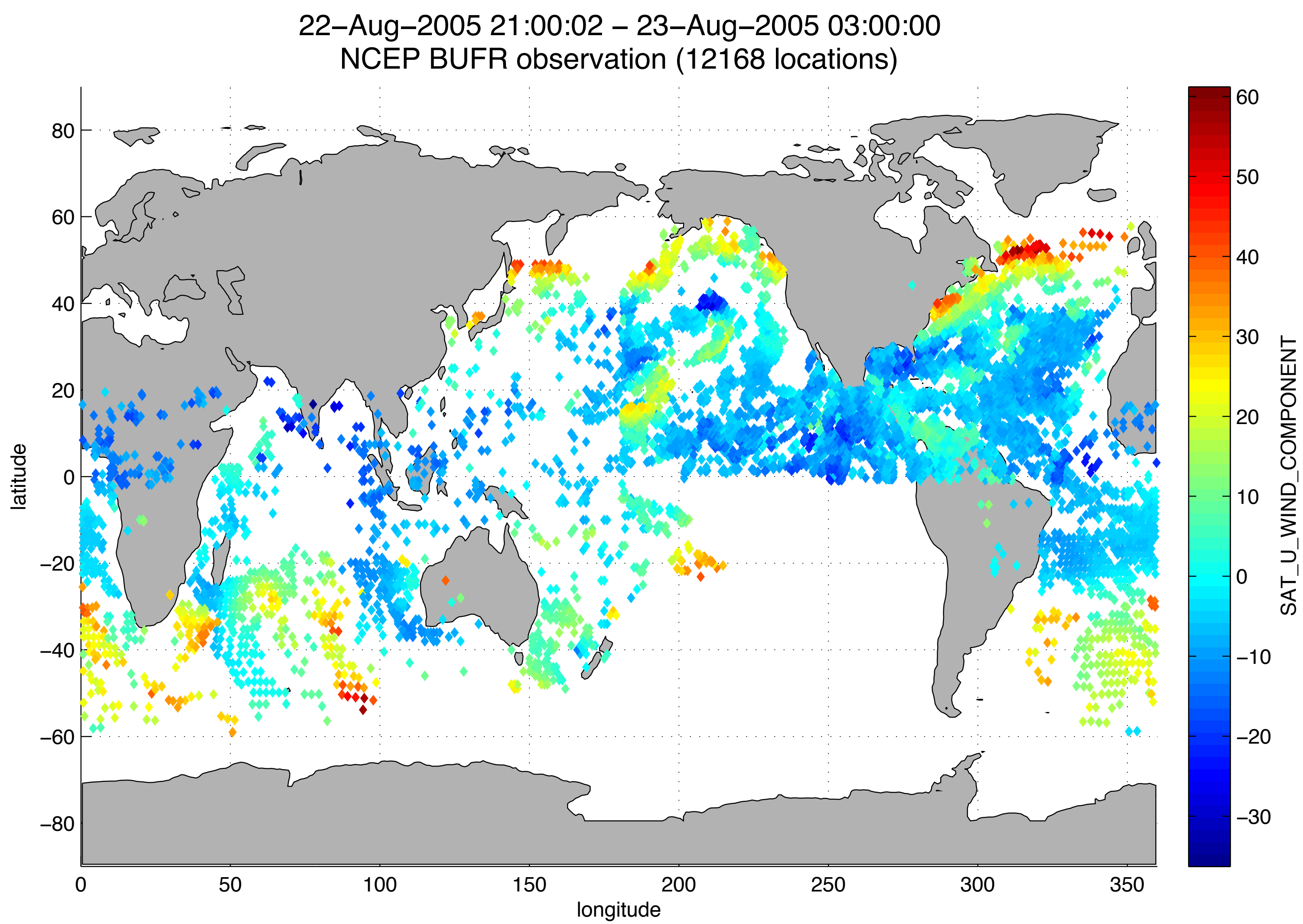


Modeling the Earth system requires coupling models.

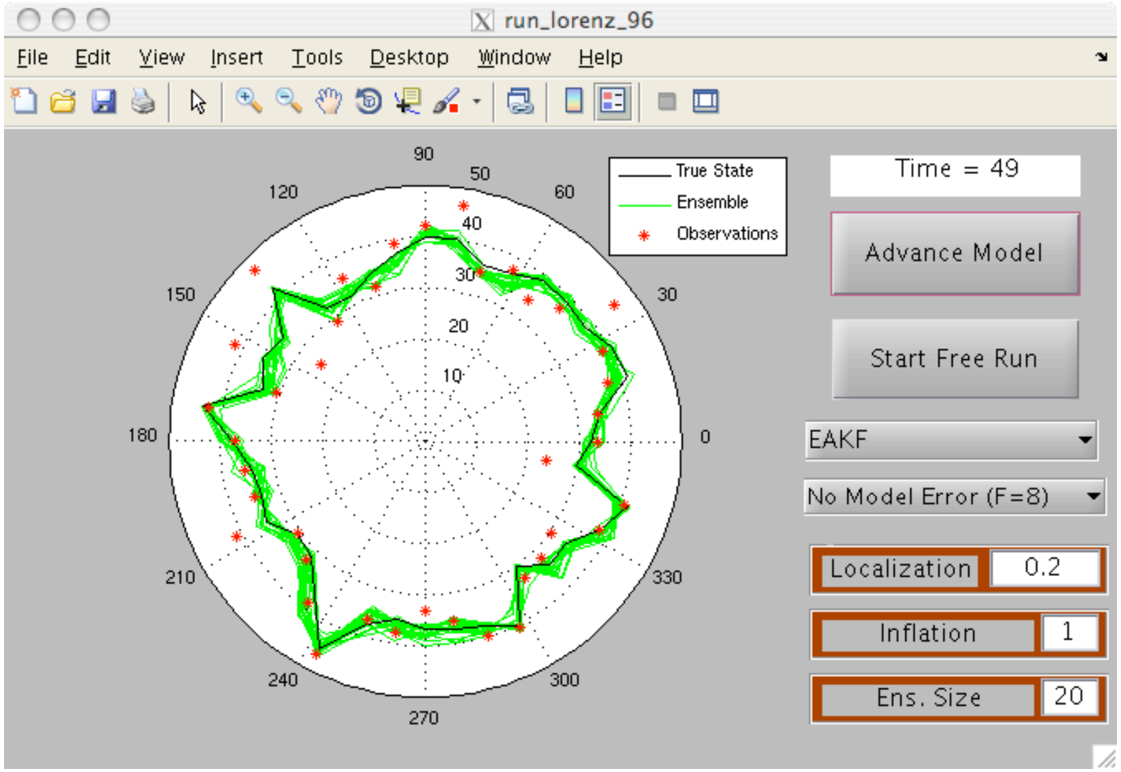
Note there are multiple instances of each model component. **Important!**

How do we know if we are winning? What is success? There's more to it than a better weather forecast!

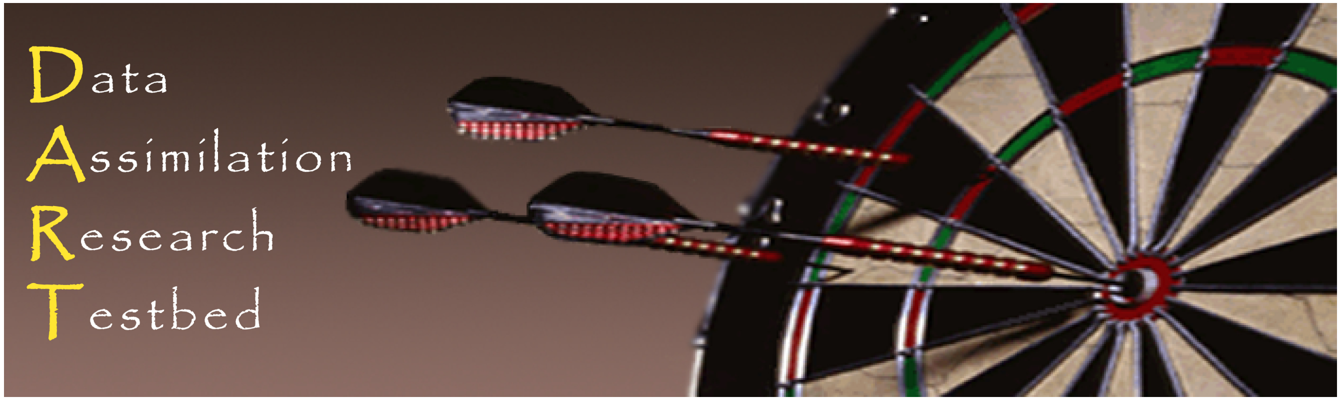
The argument for localization …



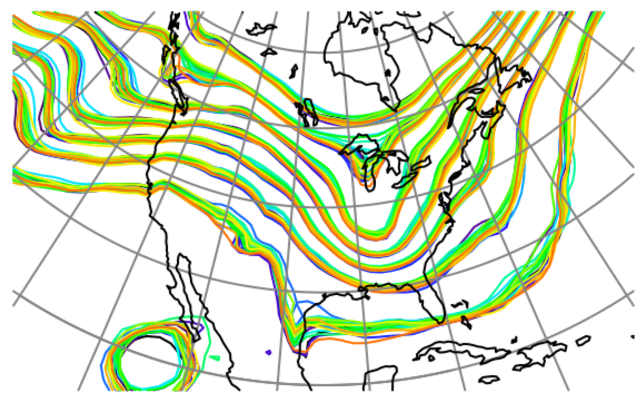
Localization & Sampling Error



Tim – don't forget to run Matlab GUI **run_lorenz_96**



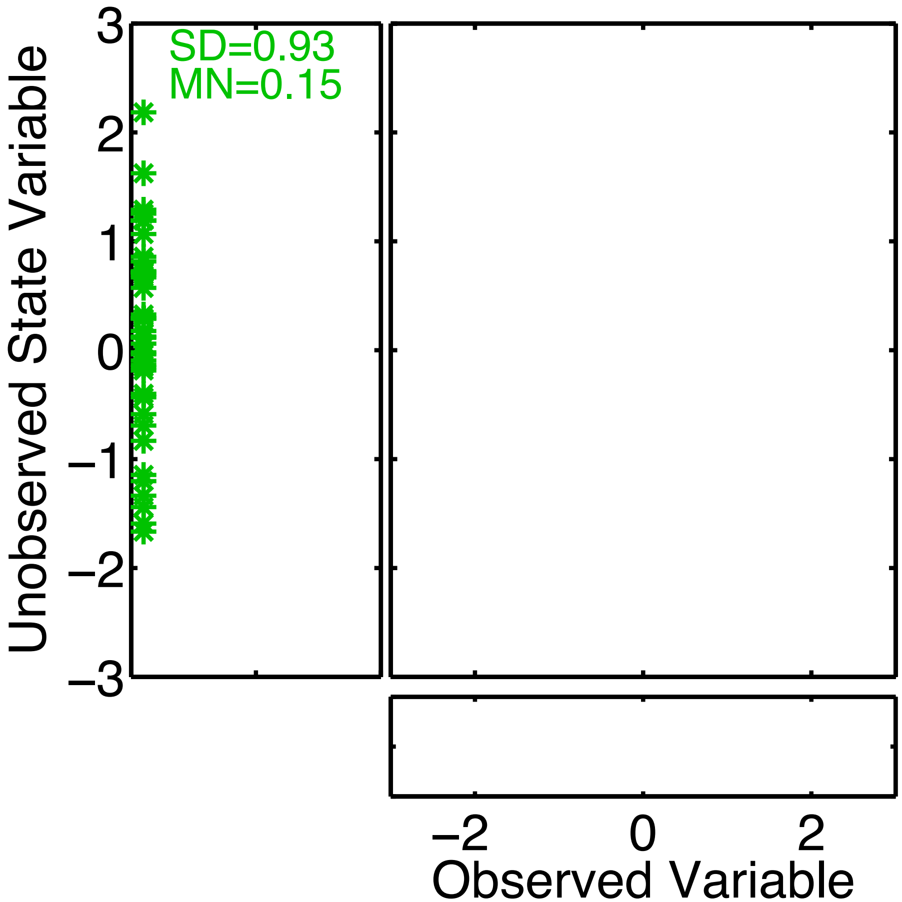
DART_LAB Tutorial Section 3: Sampling error and localization.



The National Center for Atmospheric Research is sponsored by the National Science Foundation.
Any opinions, findings and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

NCAR | National Center for
UCAR | Atmospheric Research

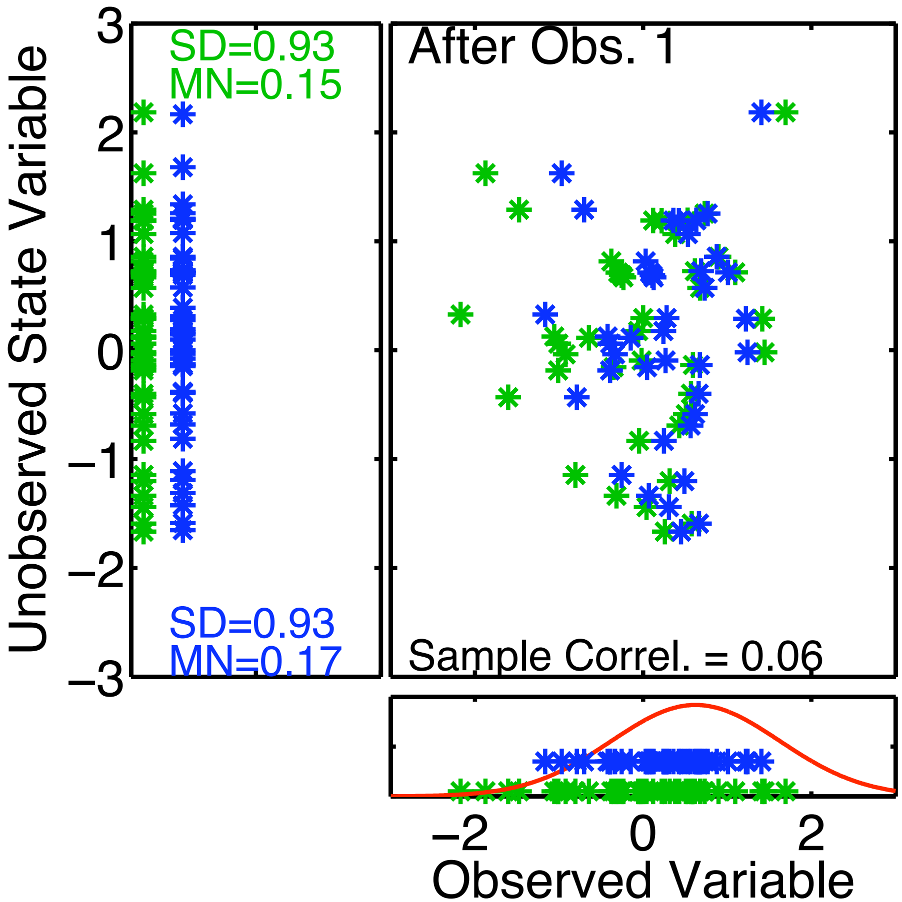
Regression Sampling Error



Suppose unobserved state variable is known to be unrelated to observed variables.

Unobserved variable should remain unchanged.

Regression Sampling Error

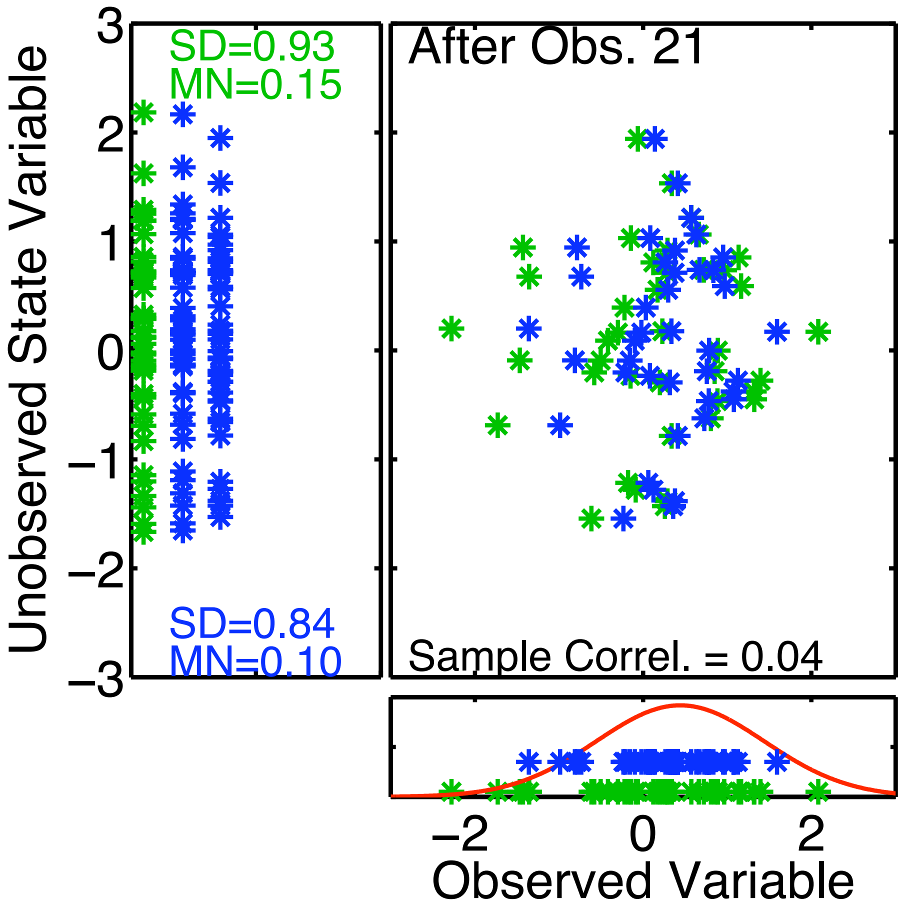


Suppose unobserved state variable is known to be unrelated to observed variables.

Finite samples from joint distribution have non-zero correlation, expected |corr| = 0.19 for 20 samples.

After one observation, unobserved variable mean and standard deviation change.

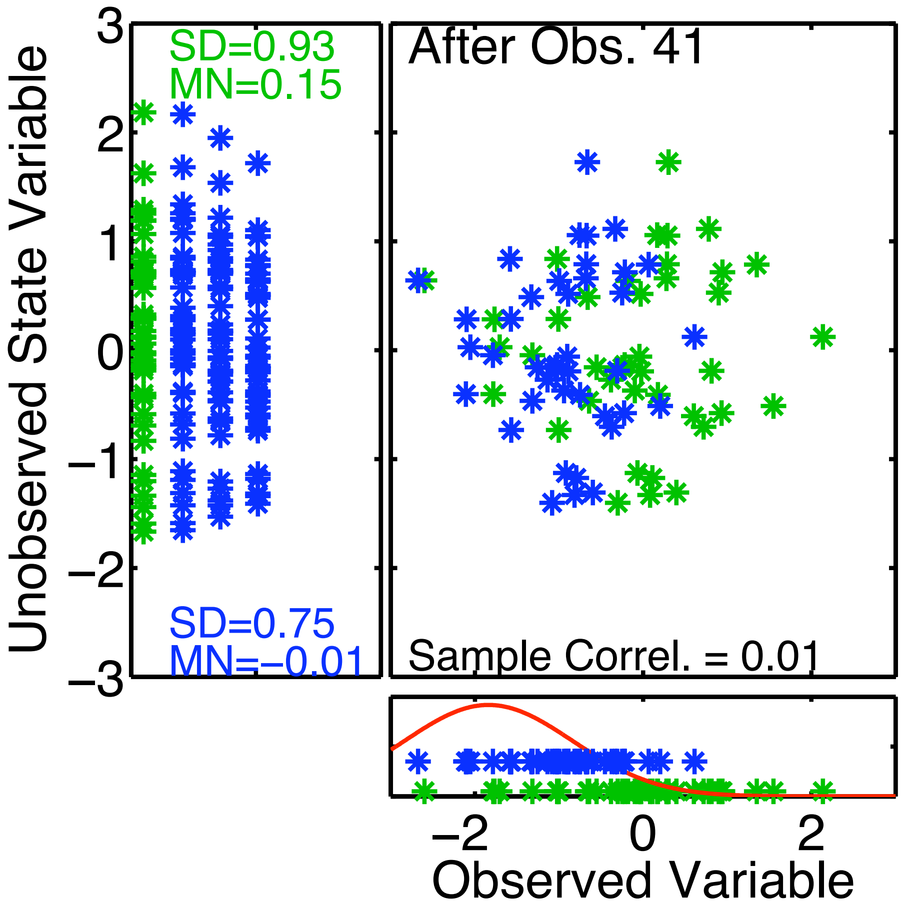
Regression Sampling Error



Suppose unobserved state variable is known to be unrelated to observed variables.

Unobserved mean follows a random walk as more observations are used.

Regression Sampling Error

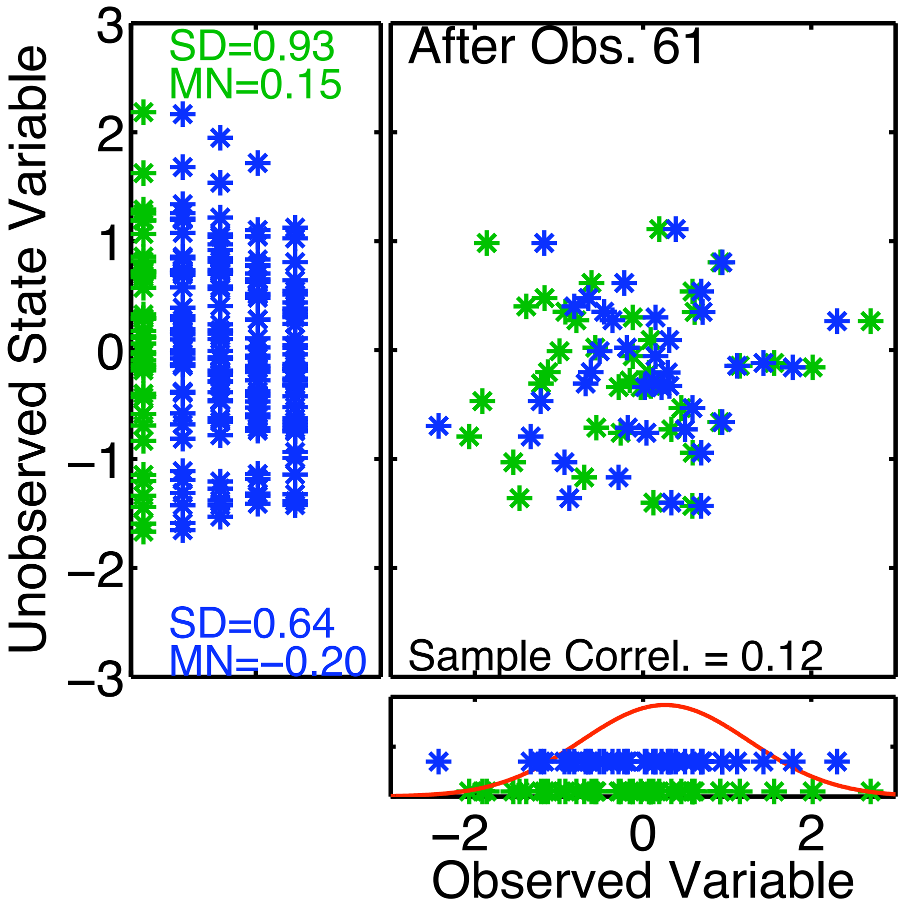


Suppose unobserved state variable is known to be unrelated to observed variables.

Unobserved mean follows a random walk as more observations are used.

Unobserved standard deviation consistently decreases.

Regression Sampling Error

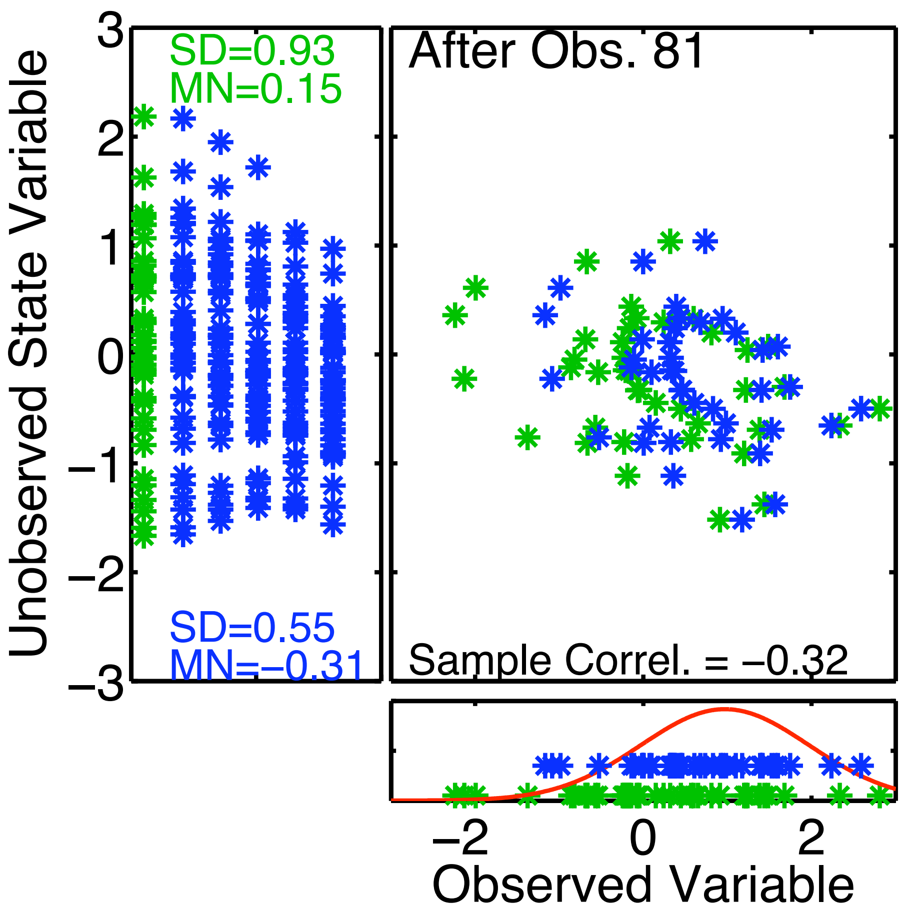


Suppose unobserved state variable is known to be unrelated to observed variables.

Unobserved mean follows a random walk as more observations are used.

Unobserved standard deviation consistently decreases.

Regression Sampling Error

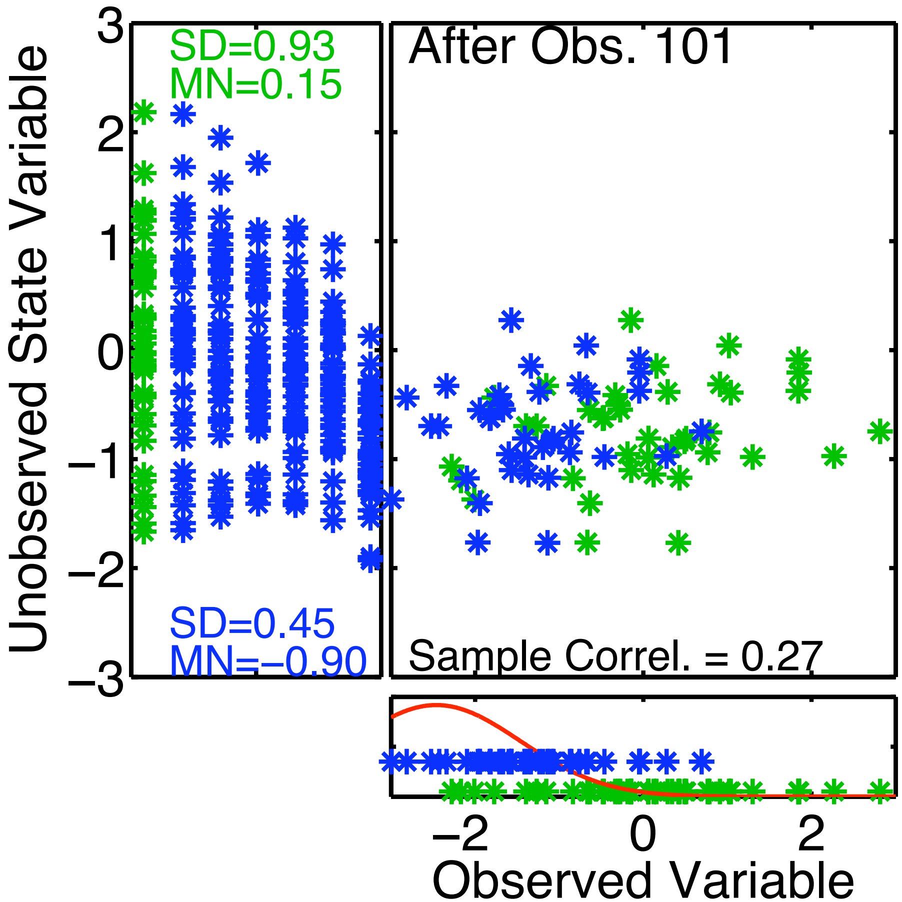


Suppose unobserved state variable is known to be unrelated to observed variables.

Unobserved mean follows a random walk as more observations are used.

Unobserved standard deviation consistently decreases.

Regression Sampling Error

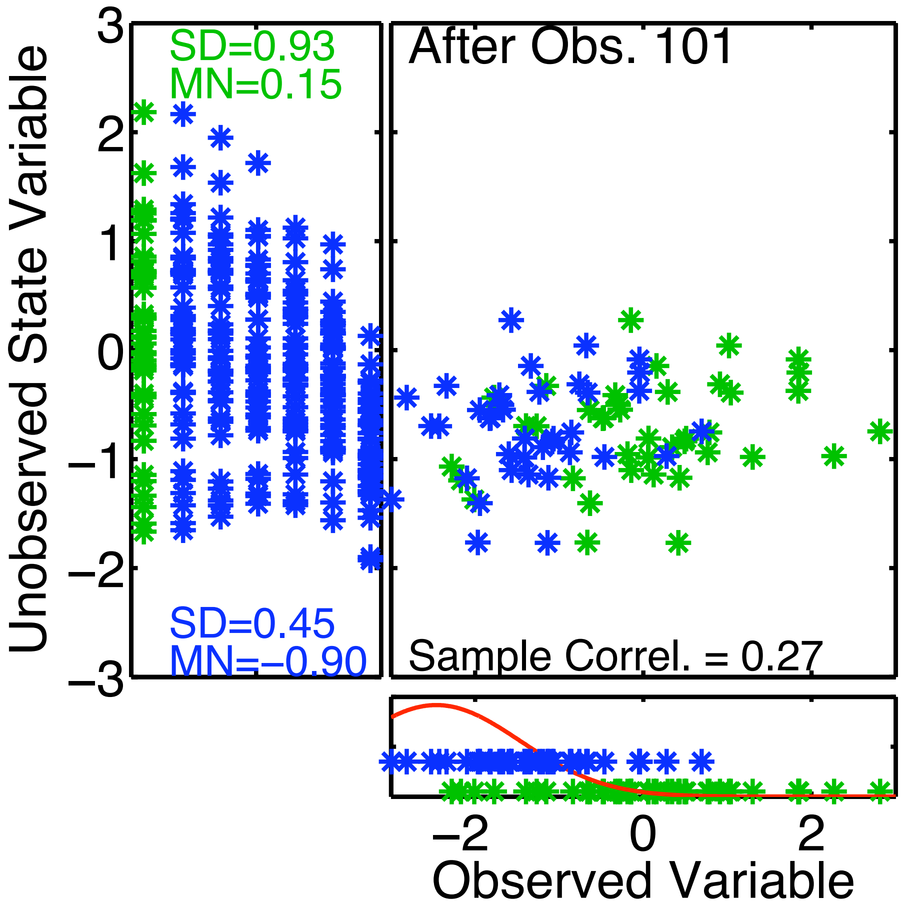


Suppose unobserved state variable is known to be unrelated to observed variables.

Unobserved mean follows a random walk as more observations are used.

Unobserved standard deviation consistently decreases.

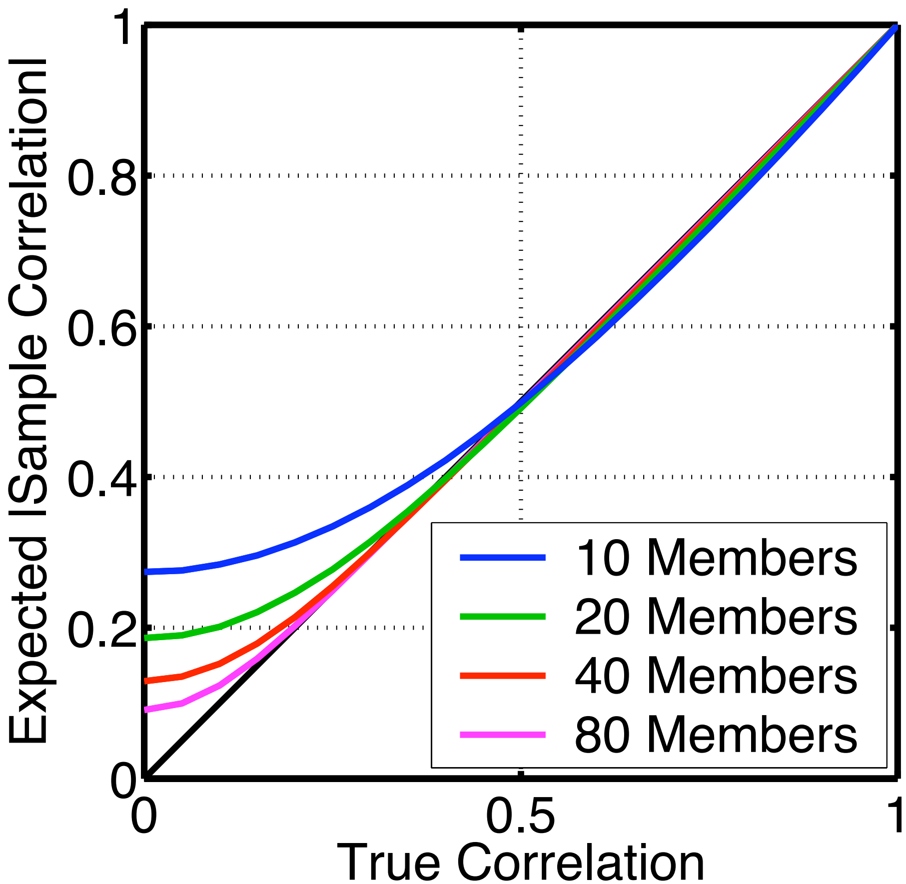
Regression Sampling Error



Suppose unobserved state variable is known to be unrelated to observed variables.

- Estimates of unobserved are too confident.
- Give less weight to subsequent meaningful observations.
- Meaningful observations can end up being ignored.

Regression Sampling Error

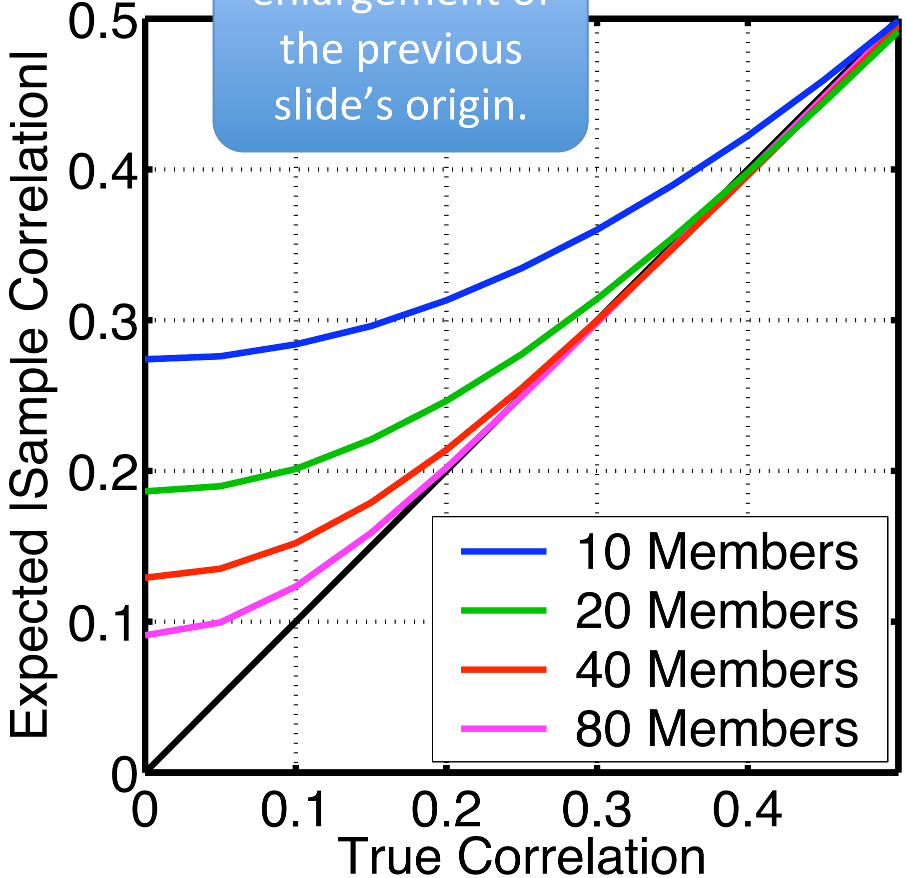


Absolute value of expected sample correlation vs. true correlation.

Errors decrease for large ensembles and for correlations with absolute value close to 1.

Regression Sampling Error

This is just an enlargement of the previous slide's origin.



For small true correlations, sampling errors are undesirably large even for 80 members!

So - the primary tool to fight this is ***localization***. Don't let observations that are known to be unrelated to model variables impact those model variables. Lots of strategies here. Physical distance, chemical properties, geographic separation (e.g. watersheds) … added benefit: **computational efficiency!**

Data Assimilation Research Testbed

So … how do we assess performance?

1. We are trying to achieve an ensemble that is indistinguishable from the physical realization of the modeled system. (we want our ensemble of models to generate synthetic observations that have the same PDF as the real observation)

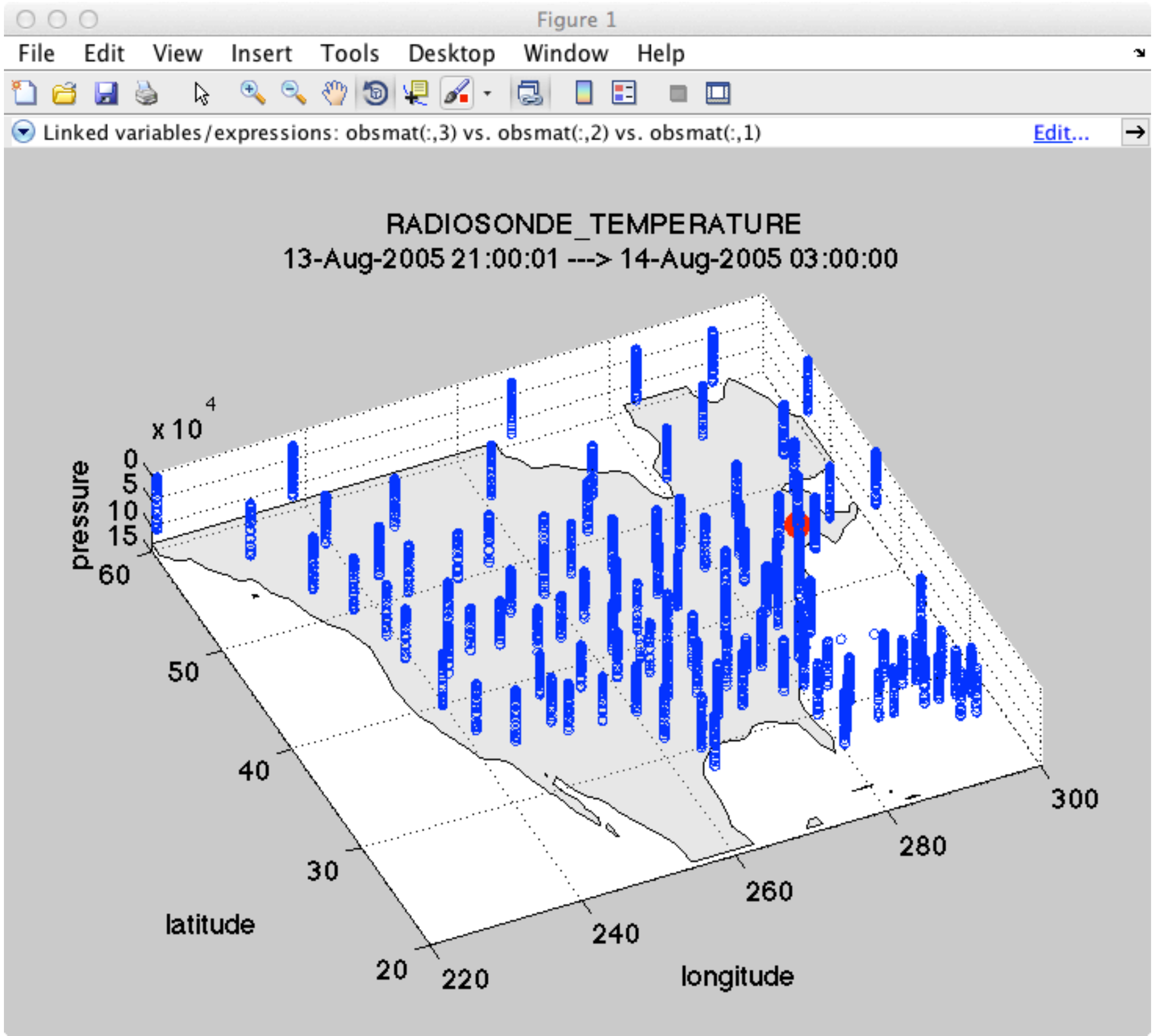
2. We want the ensemble to be as informative as possible and still capture our uncertainty in the system.

3. It is trivial to develop a method to have a terrific **posterior** RMSE compared to observations. 'Direct replacement'. This was done in the early days of atmospheric DA and it was shown to have **really poor** forecast properties.

4. It is also possible to get a great RMSE by rejecting all the observations that disagree with your ensemble. This is called 'filter divergence' and is the #1 undesirable property of ensemble methods.

Rank histograms can assess #1 and #2.
Observation-space diagnostics of the ***PRIOR*** can assess #3 and #4.

Rejection … where and why?



Tim – don't forget to run Matlab GUI **link_obs**

Recap:

So that's how to assess whether or not the assimilation was effective:

1. Are the observations getting rejected?
2. Is the ensemble collapsing?
3. Is the RMSE more-or-less steady?
4. Do the rank histograms look reasonable?

More things to think about:

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

1. Replicate an equilibrated state N times.
2. Use a unique (and different) *realistic* forcing for each to induce separate model trajectories.
3. Run them forward for "a long time".

Does this work for your model?

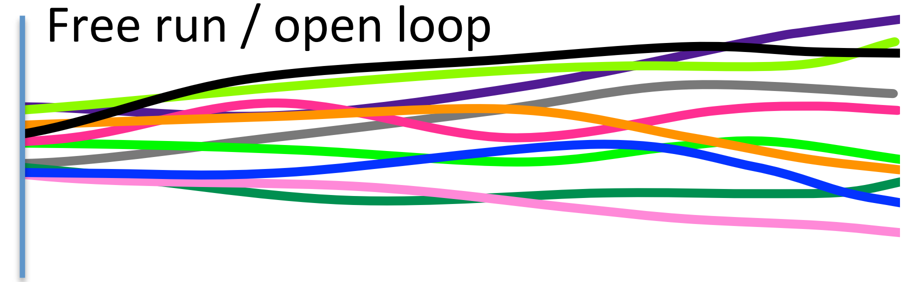
DART has tools we are using to explore how much spread we NEED to capture the uncertainty in the system.

The ensemble advantage.

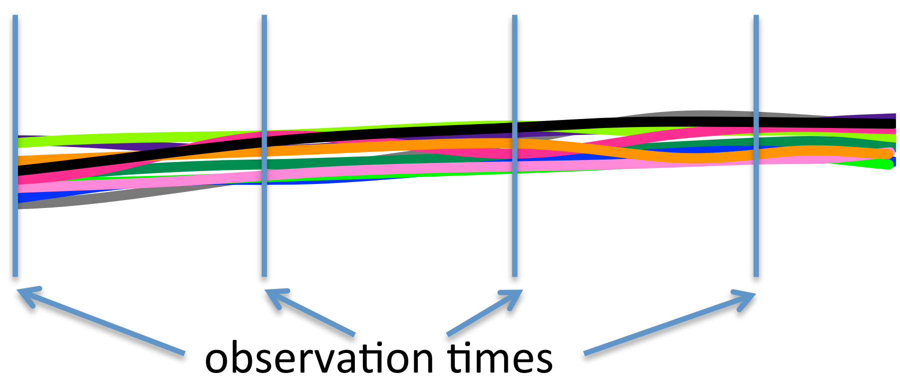
You can represent uncertainty.

time

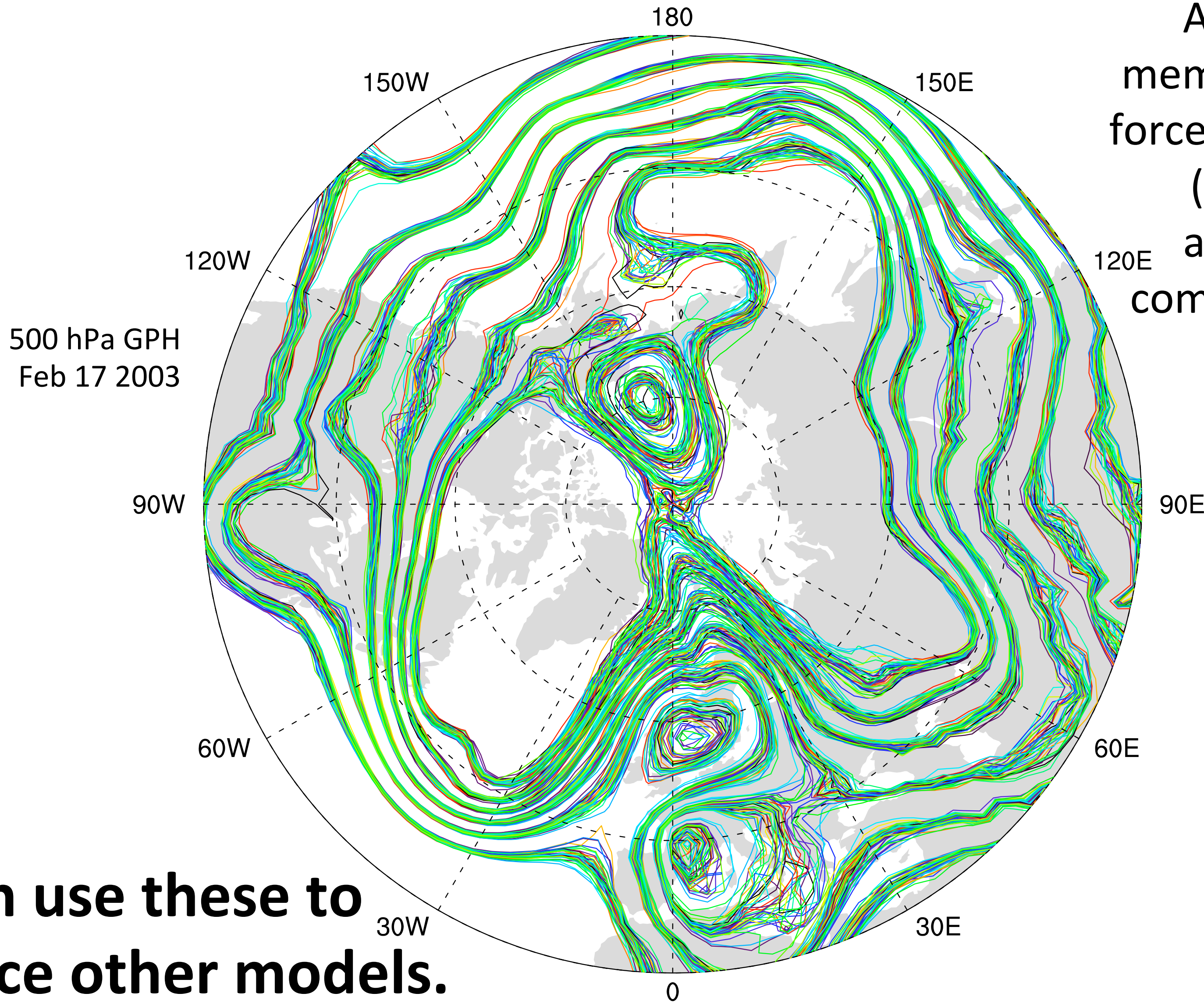
The ensemble spread frequently grows in a free run of a dispersive model.



A good assimilation reduces the ensemble spread and is still representative and informative.



Atmospheric Ensemble Reanalysis



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

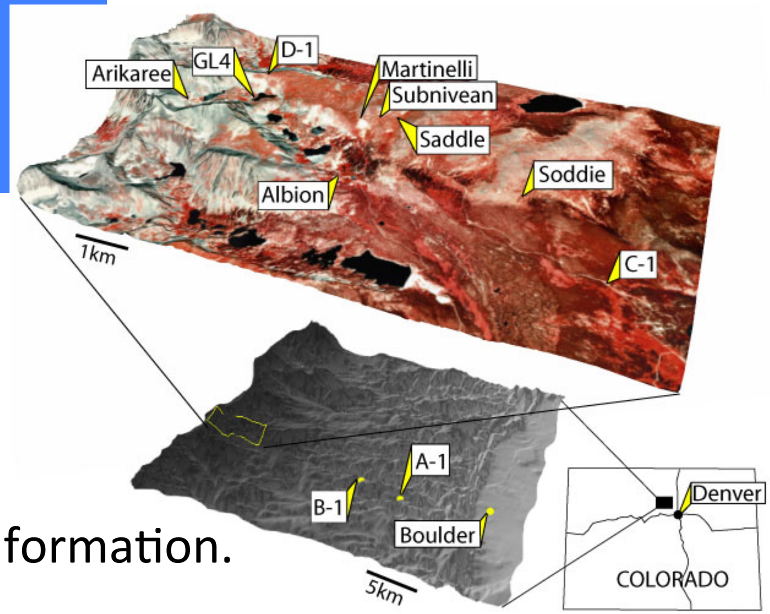
O(1 million) atmospheric obs are assimilated every day.

1998-2010+ 4x daily is available.

Can use these to force other models.

A land model experiment at a single site.

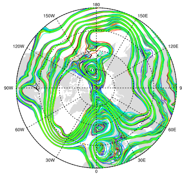
In collaboration with Andy Fox (NEON): An experiment at Niwot Ridge

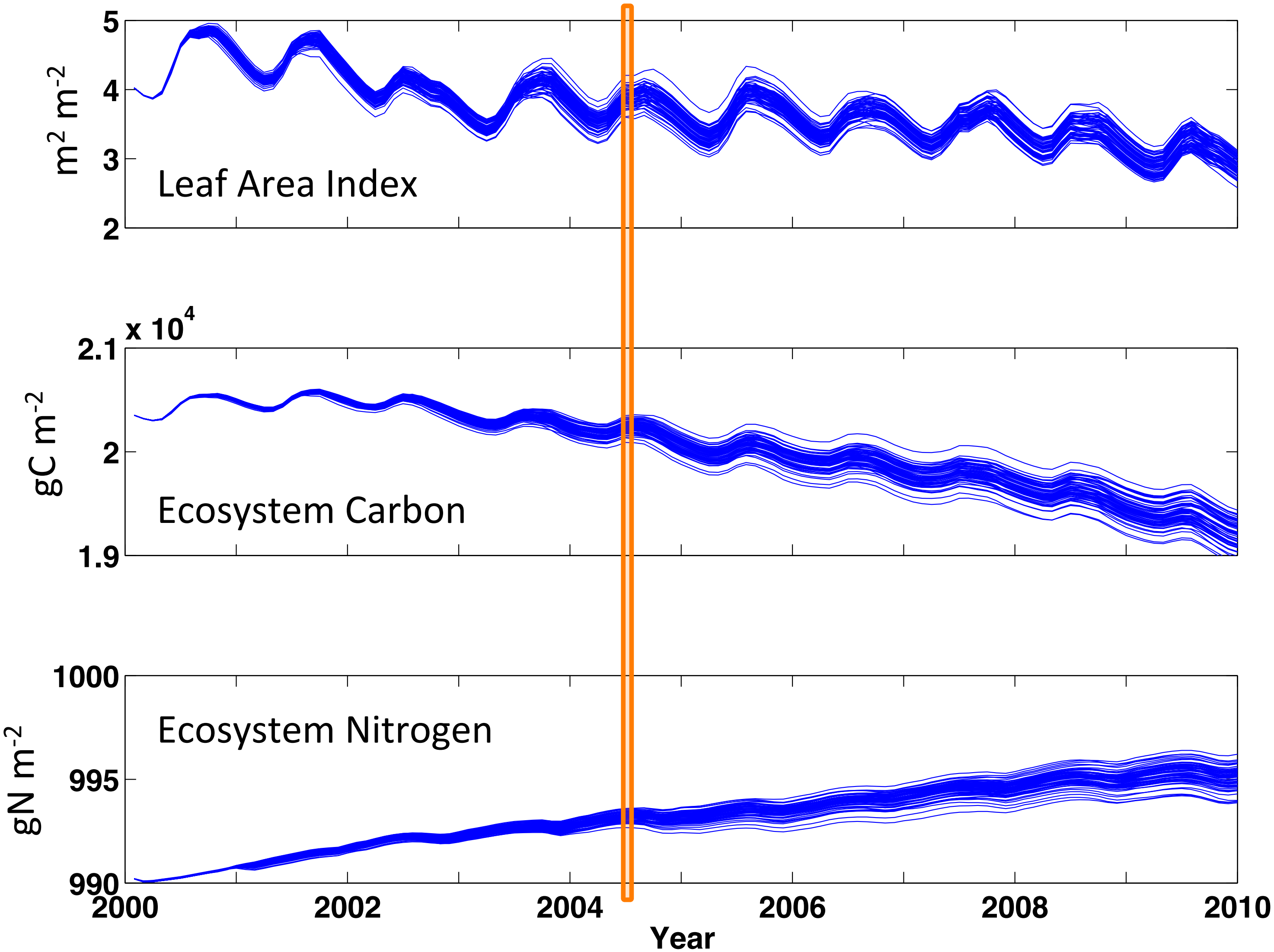


- 9.7 km east of the Continental Divide
- C-1 is located in a Subalpine Forest
- (40º 02' 09'' N; 105º 32' 09'' W; 3021 m)
- One column of Community Land Model (CLM)
  - Spun up for 1500 years with site-specific information.
- 64 ensemble members
- Forcing from the DART/CAM reanalysis,
- Assimilating tower fluxes of latent heat (LE), sensible heat (H), and net ecosystem production (NEP).
- Impacts CLM variables: LEAFC, LIVEROOTC, LIVESTEMC, DEADSTEMC, LITR1C, LITR2C, SOIL1C, SOIL2C, SOILLIQ … all of these are *unobserved*.

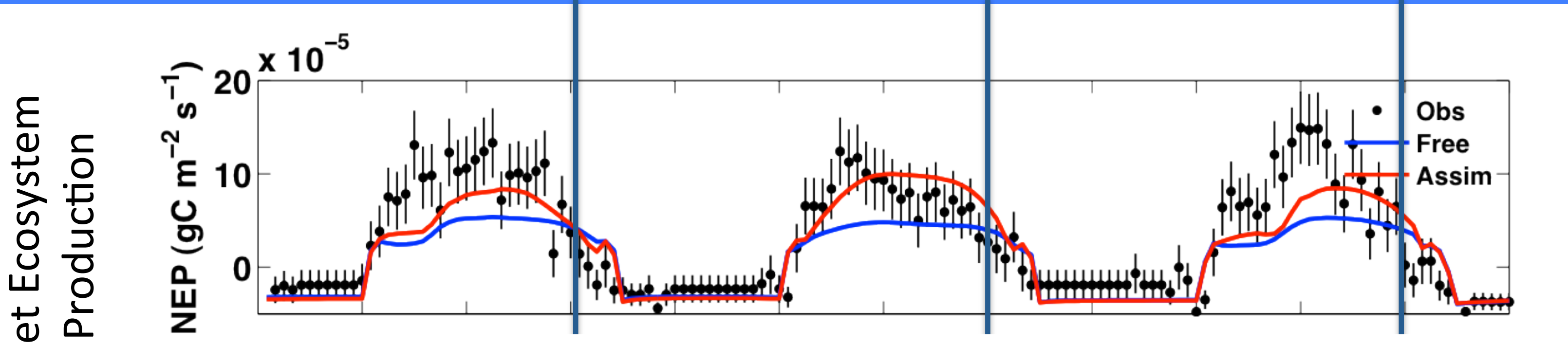
This is the sort of information that needs to be disclosed!

Free Runs of CLM driven by 64 CAM reanalyses

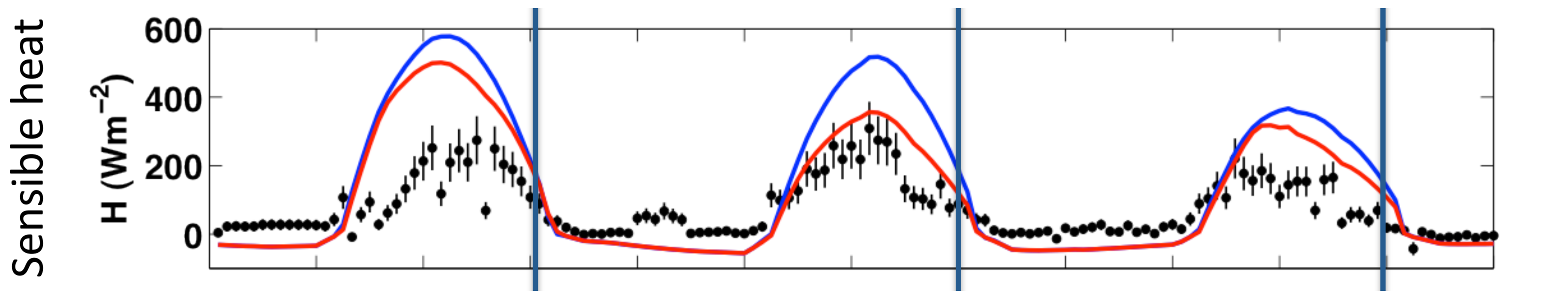




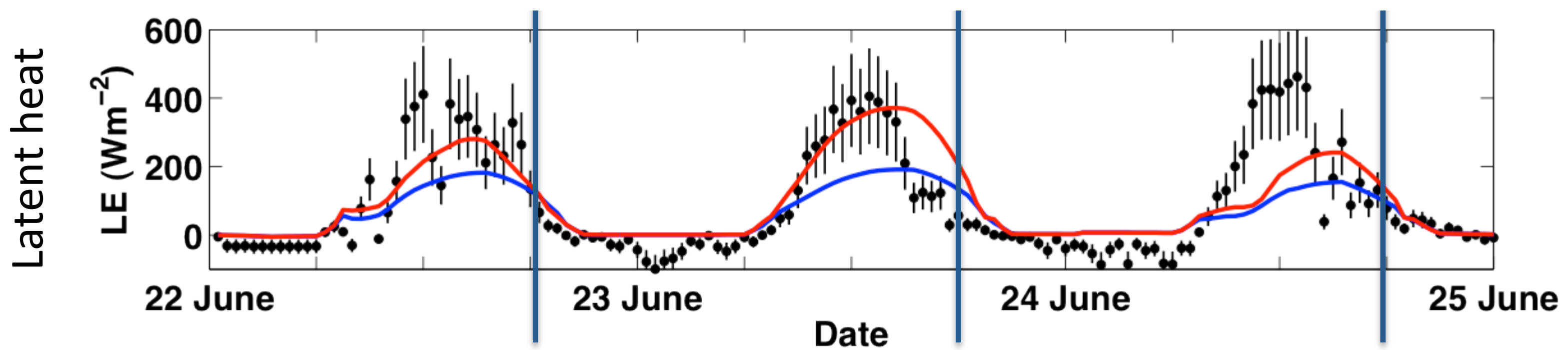
In collaboration with Andy Fox (NEON): Focus on the ensemble means (for clarity)



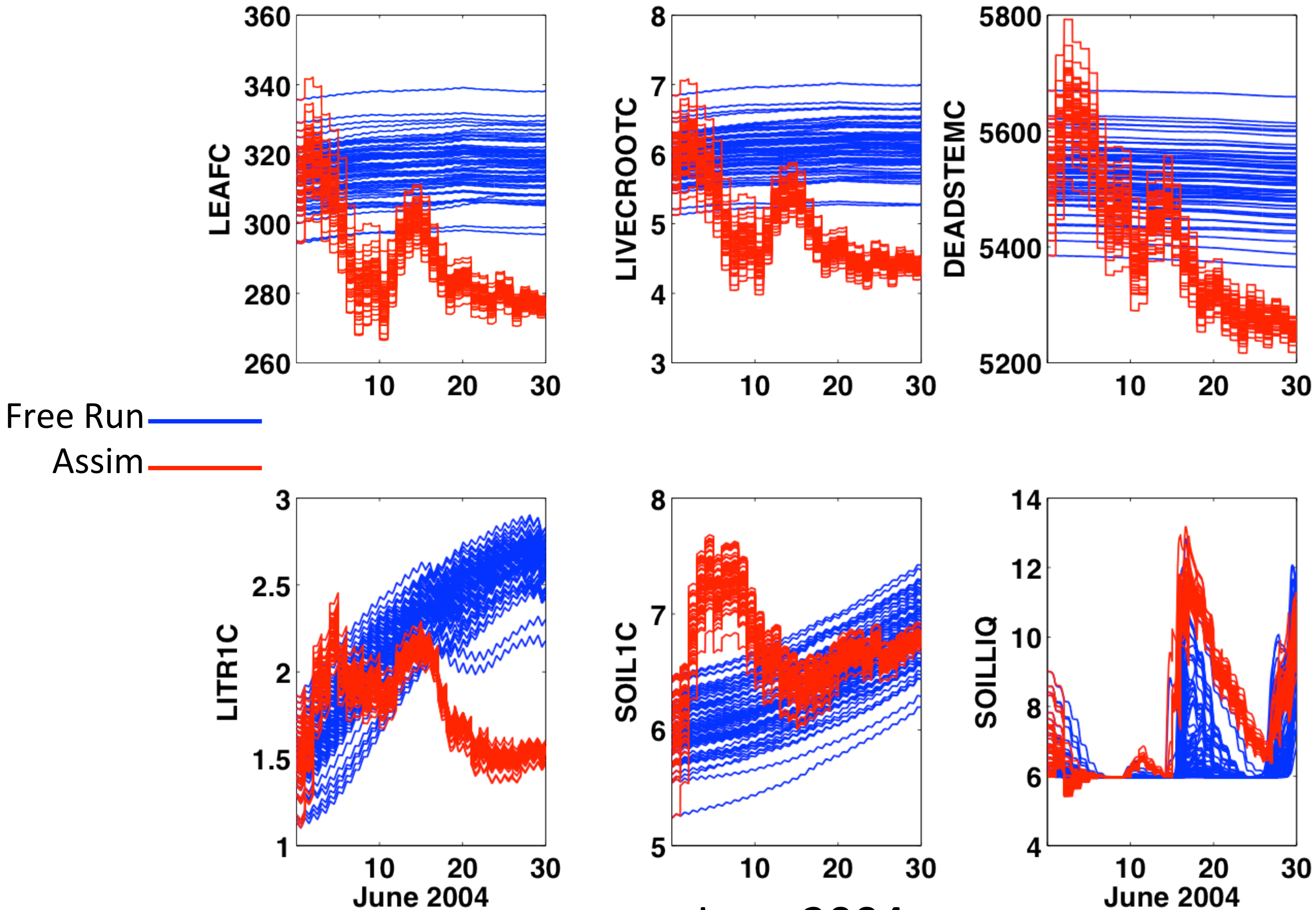
FYI: We are assimilating flux observations ...



The model states are being updated at about 8PM local time.

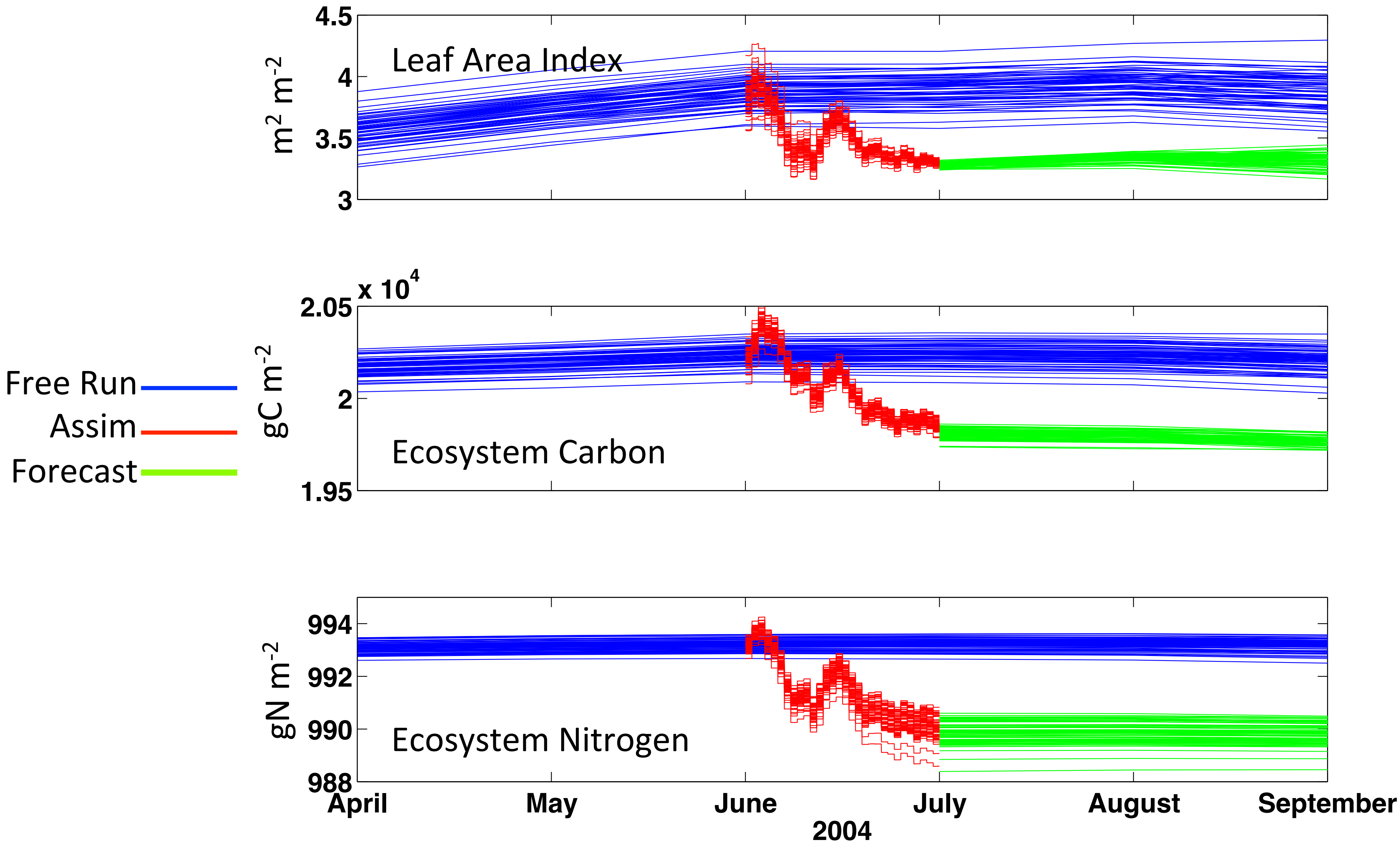


These are all unobserved variables.

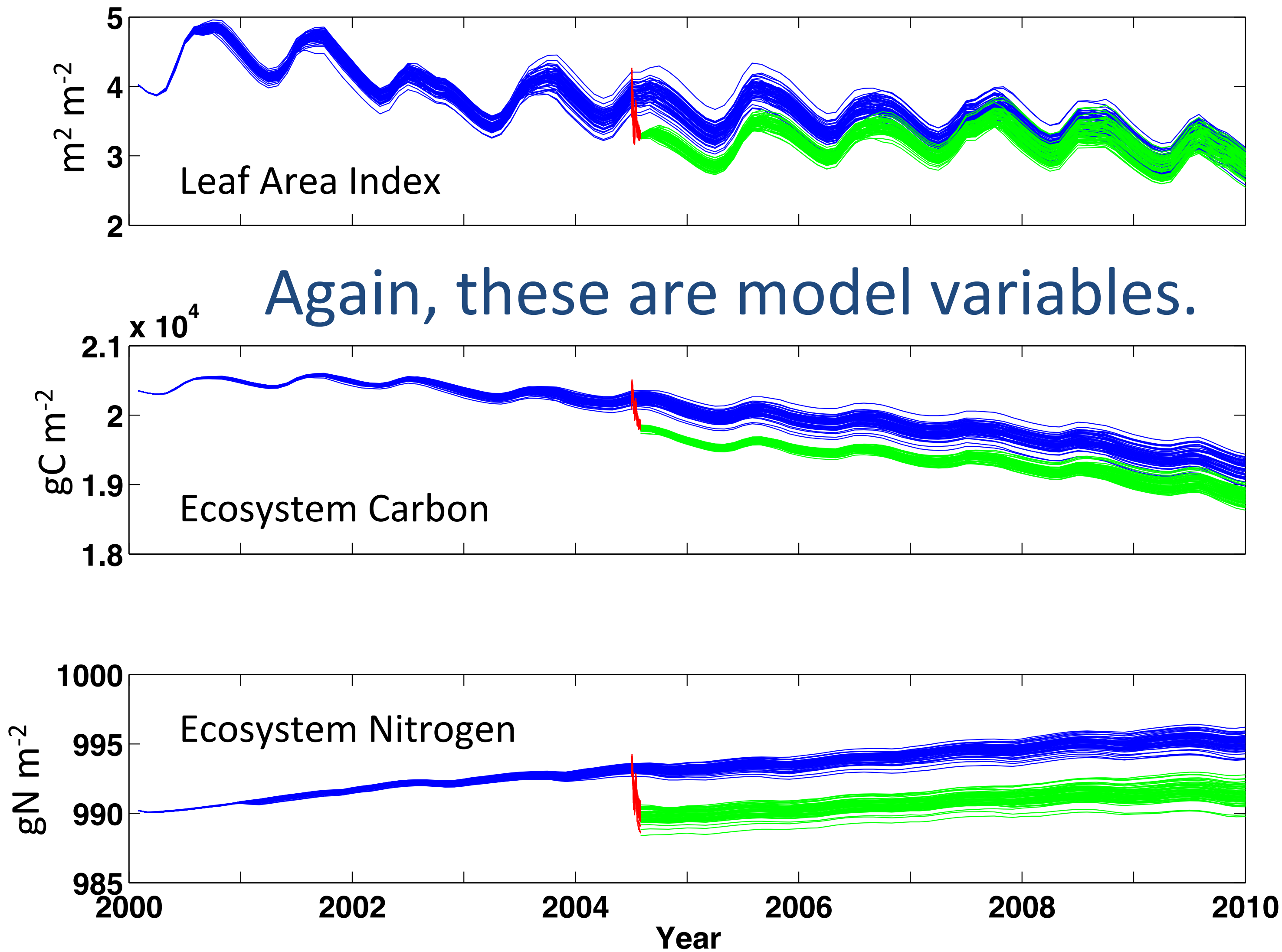


June 2004

Effect on short-term forecast on unobserved variables.



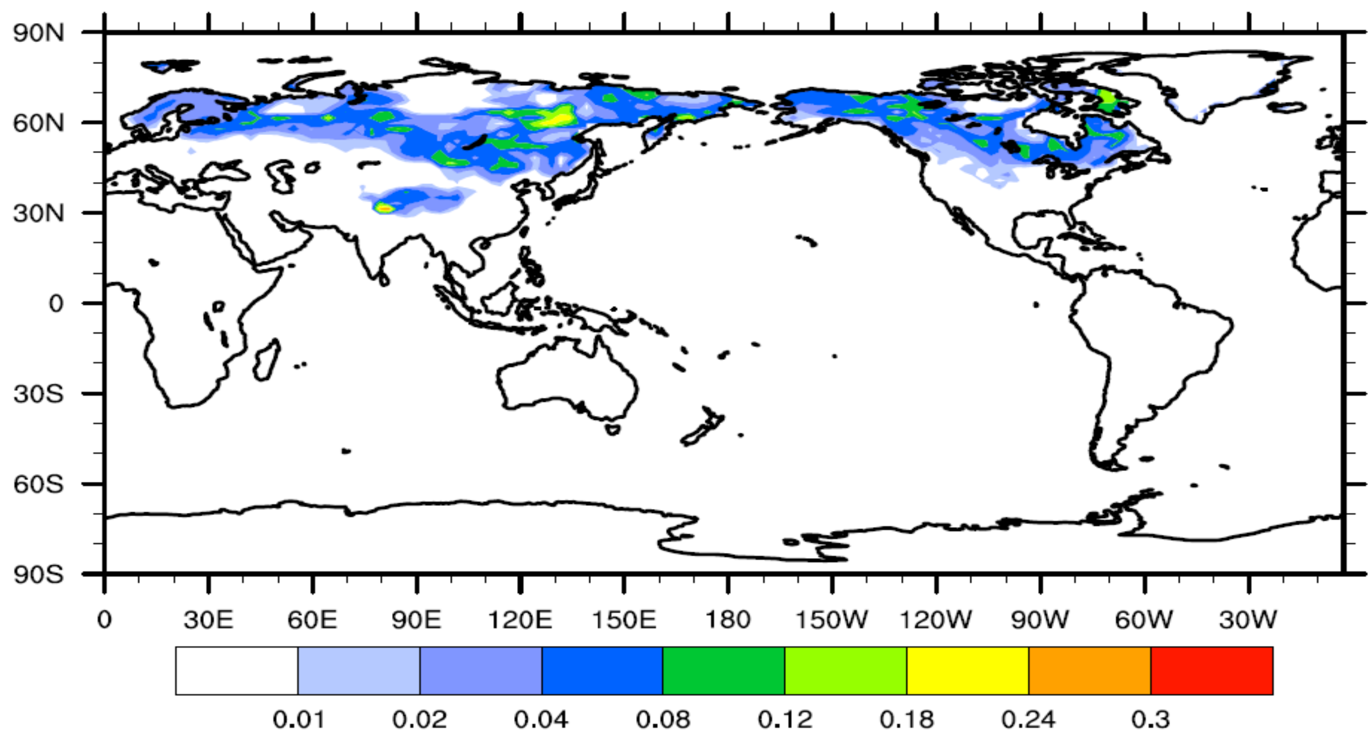
Effect on longer-term forecast



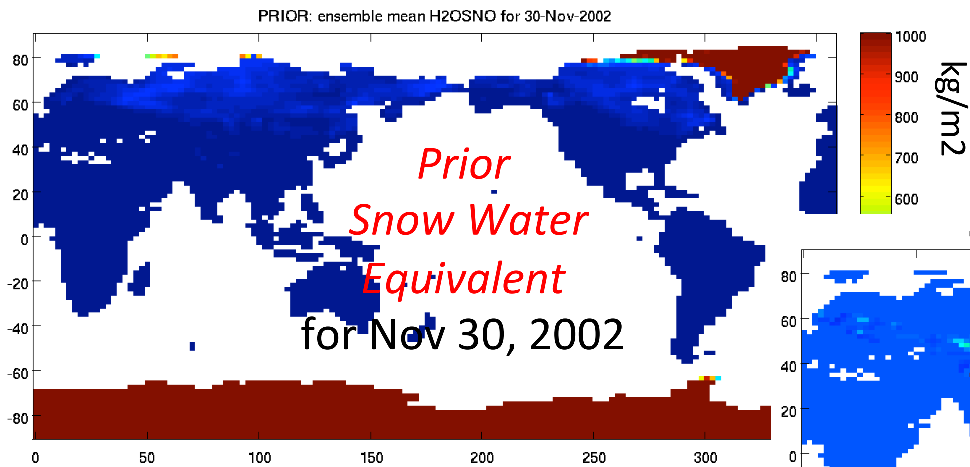
Assimilation of MODIS snow cover fraction

- 80 member ensemble for onset of NH winter, assimilate once per day
- Level 3 MODIS product – regridded to a daily 1 degree grid
- Observations can impact state variables within 200km
- CLM variable to be updated is the snow water equivalent "H2OSNO"
- **Analogous to precipitation** …

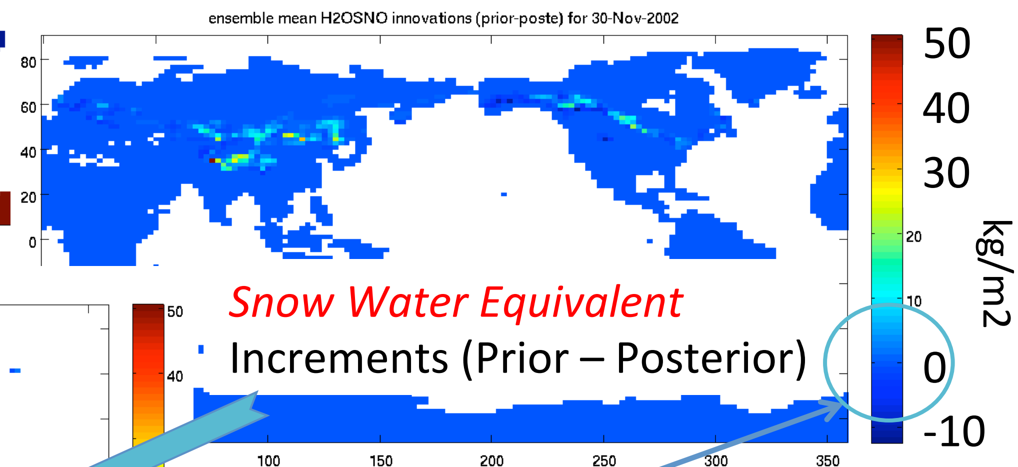
Standard deviation of the CLM snow cover fraction initial conditions for Oct. 2002

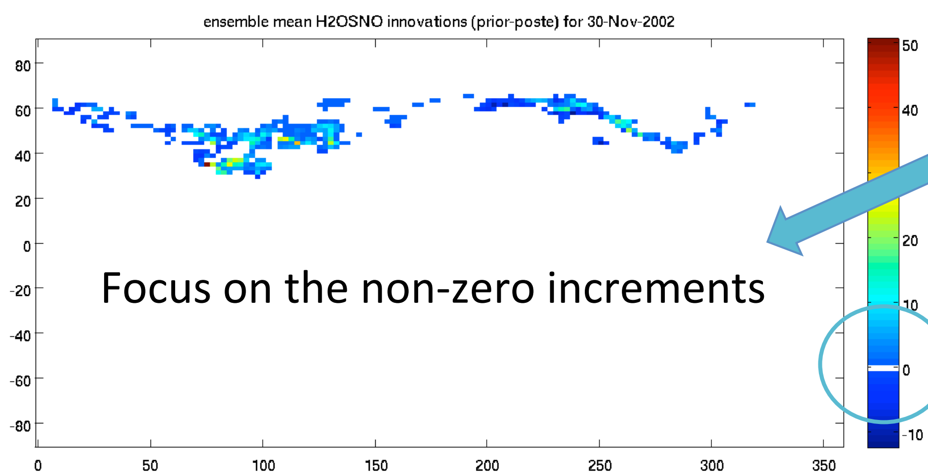


An early result: assimilation of MODIS *snow cover fraction* on total *snow water equivalent* in CLM.



Thanks Yongfei!





The model state is changing in reasonable places, by reasonable amounts.

The HARD part:

What do we do when SOME (or none!) of the ensembles have [snow,leaves,precipitation, …] 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?

Packed Snow?

Early Season Snow?

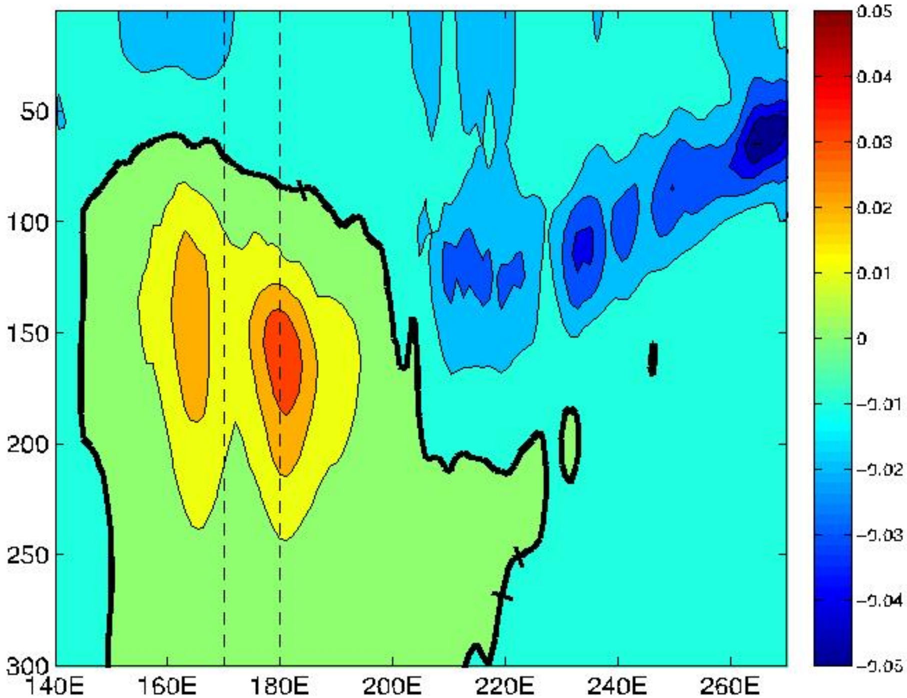
Snow Density?

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!

Ocean Considerations

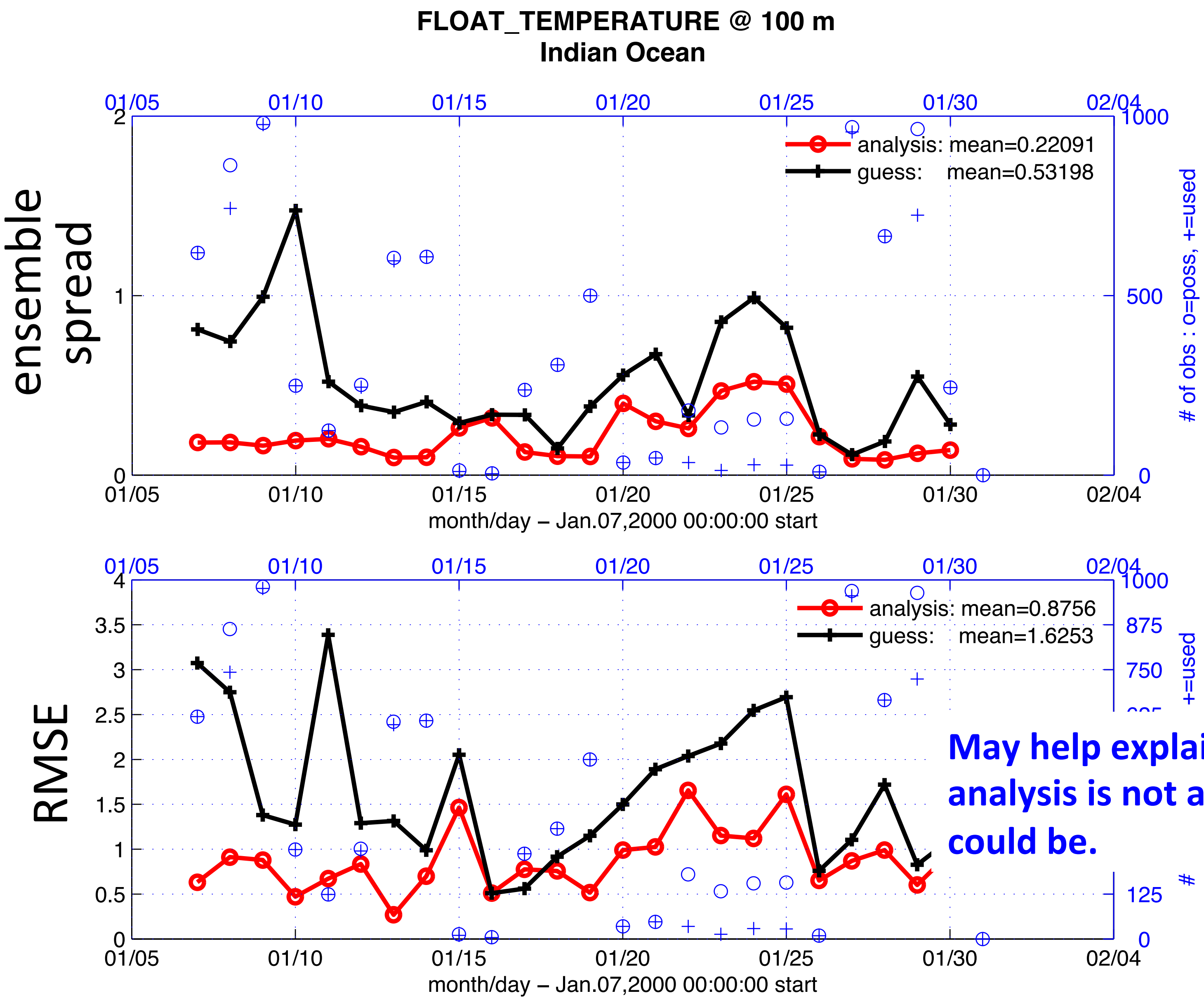
Alicia R. Karspeck, Steve Yeager, Gokhan Danabasoglu, Tim Hoar, Nancy Collins, Kevin Raeder, Jeffrey Anderson, and Joseph Tribbia, 2013: An Ensemble Adjustment Kalman Filter for the CCSM4 Ocean Component. *J. Climate*, **26**, 7392–7413. doi: http://dx.doi.org/10.1175/JCLI-D-12-00402.1



The 2005 average temperature increment for the POP-DART ocean data assimilation in the equatorial Pacific (2.5S to 2.5N) for 1-day assimilation cycles. This represents a tilting and sharpening of the equatorial thermocline.

Regional models have to consider Boundary conditions.
Buoyancy effects from observations not in 'profiles'
Model states that cannot be numerically supported – sharp boundary currents!

Diagnostics – Ocean Example



# of observations available AND # actually used!

May help explain why analysis is not as good as it could be.

Key Questions for Ensemble DA:

- What parts of the model 'state' do we update?
- What is a proper initial ensemble?
- Is an ensemble of boundary conditions necessary?
- Localization considerations
- How many ensemble members are needed to mitigate regression error?
- What is the proper observation error specification? It is not just instrument error but also mismatch in representativeness.
- Can models tolerate new assimilated states? Silently fail? Violently fail?
- Snow (vegetation) … depths, layers, characteristics, content.
- Forward observation operators
  - Many observations are over timescales or are quantities that are inconvenient
- Bounded quantities? When all ensembles have identical values the observations cannot have any effect with the current algorithms.

Climate Modeler's Commandments

by John Kutzbach (Univ. of Wisconsin).

1. Thou shalt not worship the climate model.
2. Thou shalt not worship the climate model, but thou shalt honor the climate modeler, that it might be well with thee.
3. Thou shalt use the model that is most appropriate for [illegible]ion at hand.
4. Thou shalt not change more than one thing at [illegible]st.
5. In making sensitivity experiments, thou [illegible] model hard enough to make it notice you.
6. Thou shalt not covet fine-scale res[illegible]oarse-scale model.
7. Thou shalt follow the rules f[illegible]ce testing and remember the model's inherent variability.
8. Thou shalt know the model [illegible]es and remember that model biases may lead to biased sensitivity estimates.
9. Thou shalt run the same experiment with different models and compare the results.
10. **Thou shalt worship good observations of the spatial and temporal behavior of the earth system. Good models follow such observations. One golden observation is worth a thousand simulations.**

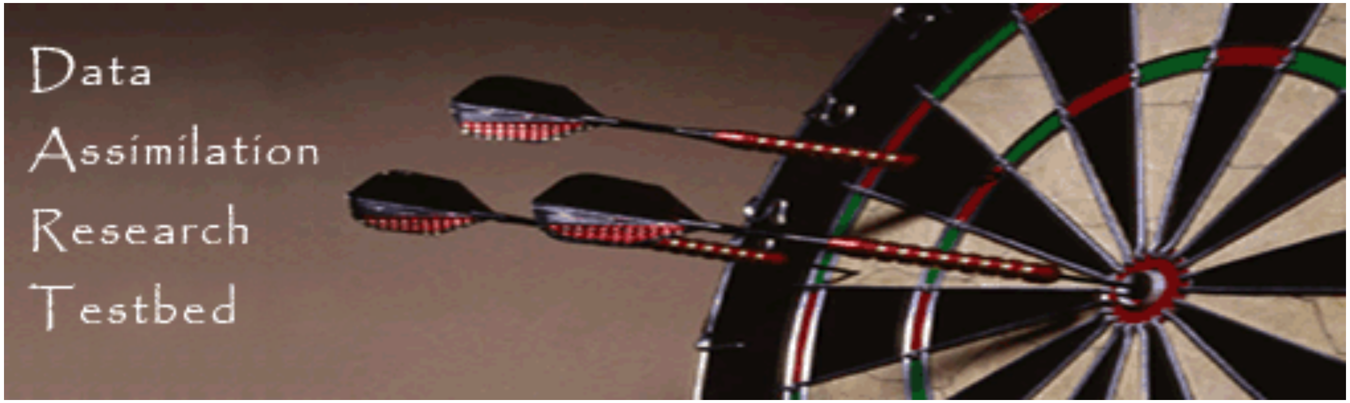
Amen Brother!

For more information:

CAM GCOM CAM-Chem ROMS WRF

GITM WRF-Hydro WACCM

CLM



POP

AM2

BGRID

COAMPS

www.image.ucar.edu/DAReS/DART

NOAH

dart@ucar.edu

MPAS_ATM

MITgcm_ocean

SQG NAAPS MPAS_OCN TIEGCM COAMPS_nest

NCOMMAS PE2LYR PBL_1d

CABLE WRF-Chem

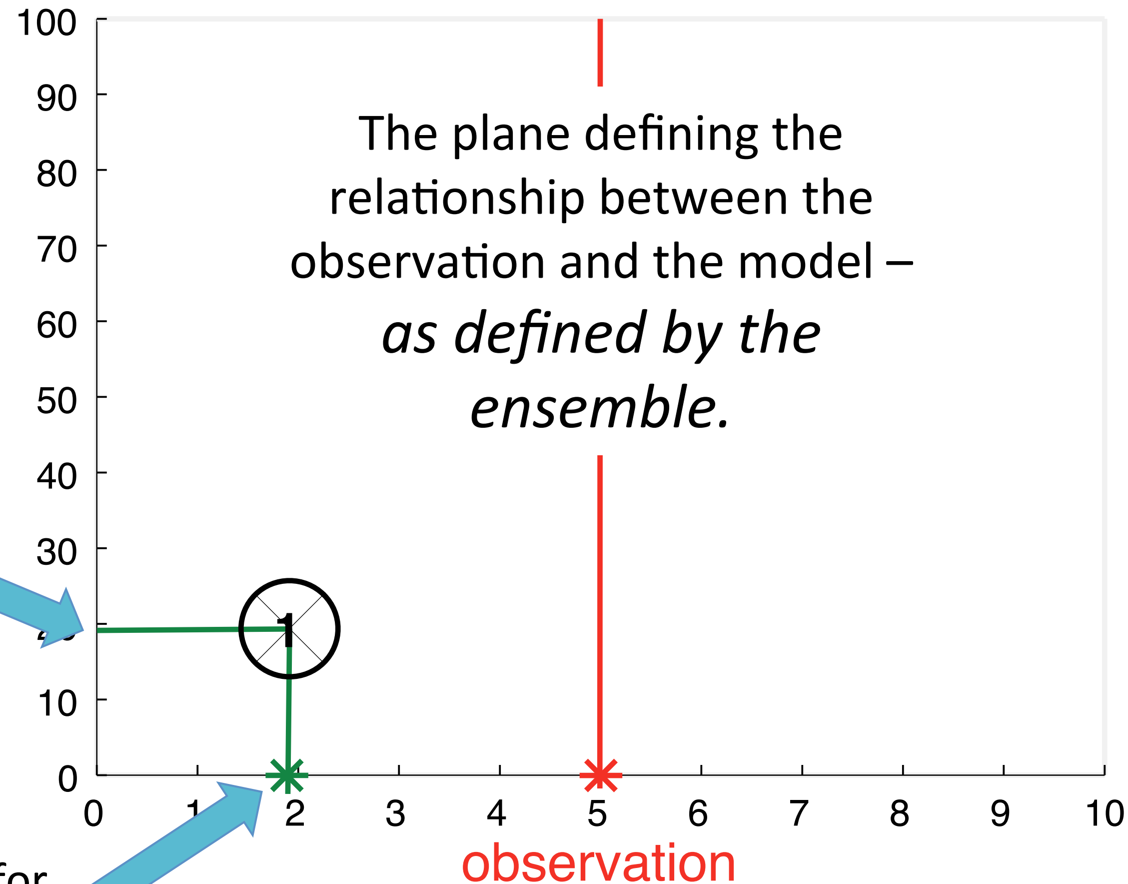
San Diego is very nice, but …

Everything after here held in reserve.

Looking at it another way:

Some unobserved state variable. e.g. live root carbon, dead root carbon, canopy water …

Directly from ensemble member 1

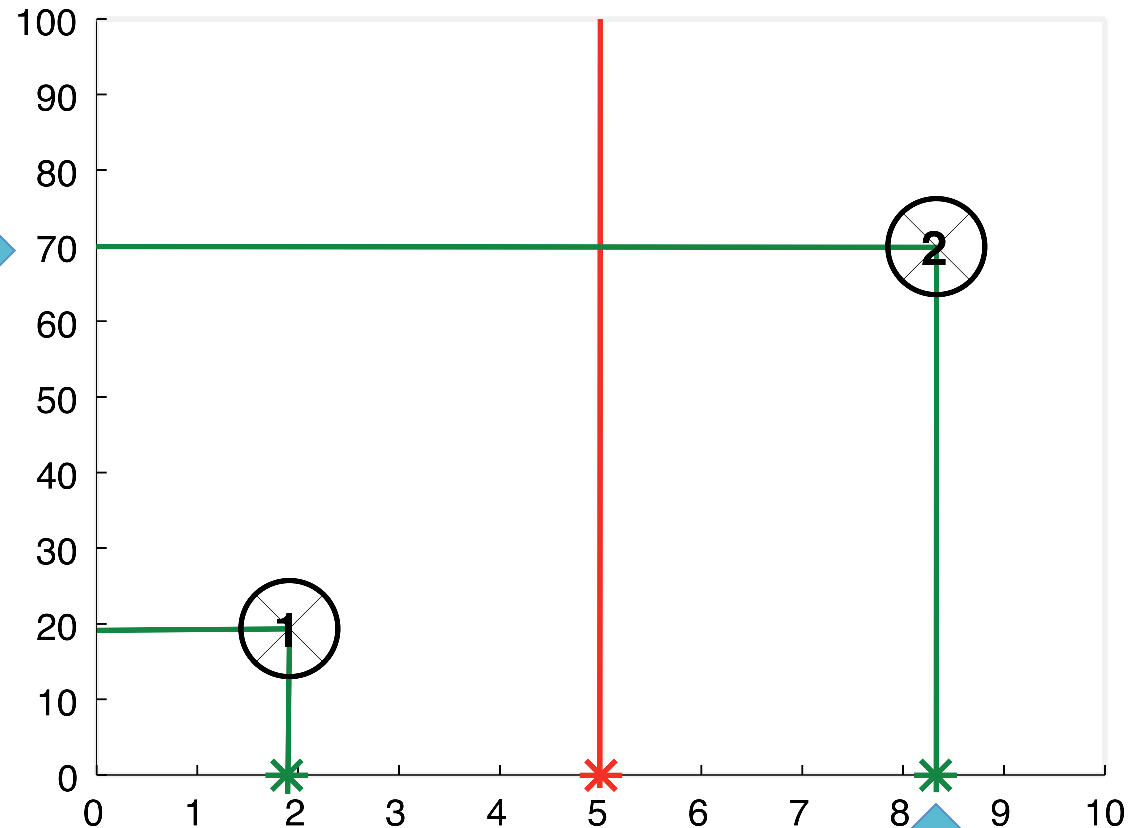


Result of the forward observation operator for ensemble member 1

Could be Soil Temperature

Looking at it another way:

Directly from ensemble member 2

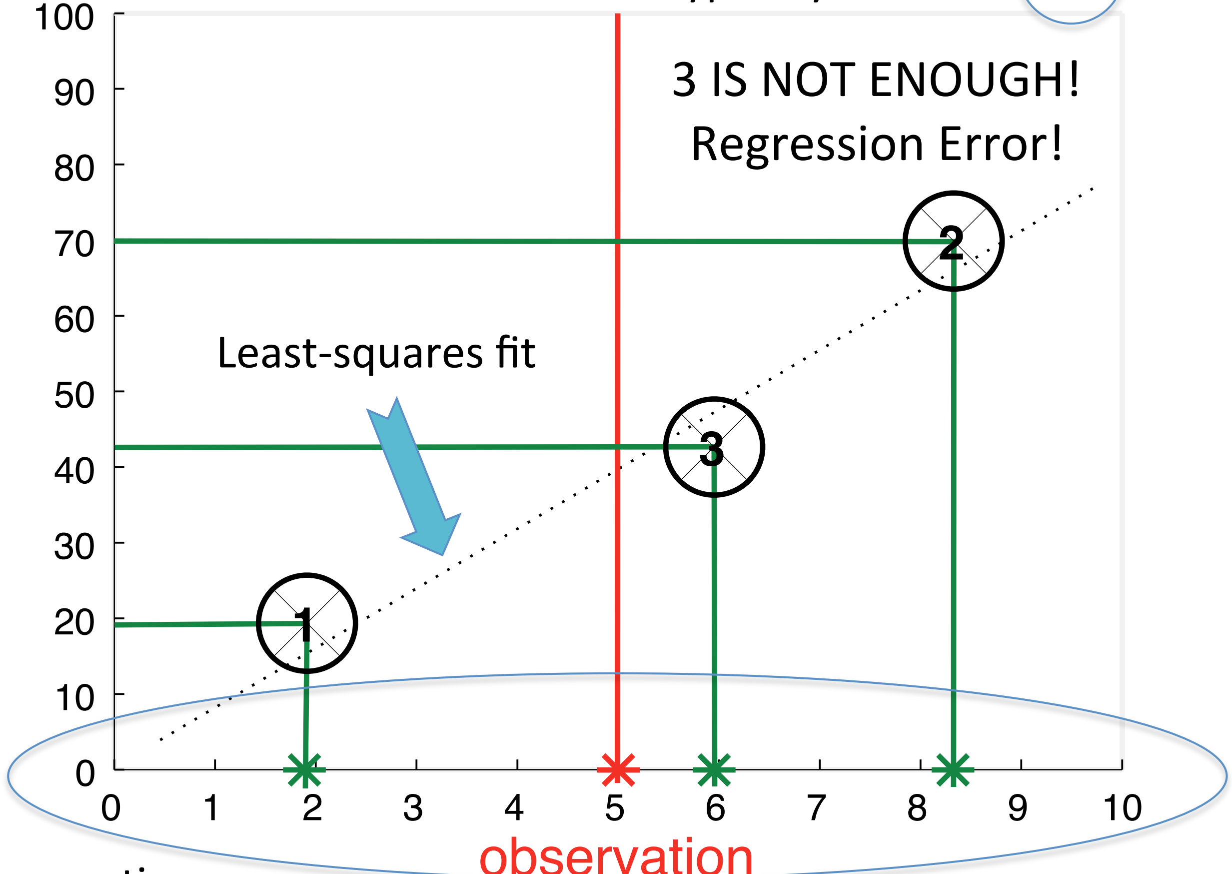


observation

"observation" from ensemble member 2

Looking at it another way:

In our assimilations, we typically use order 80.



Now, we can calculate out observation increments any way we want.

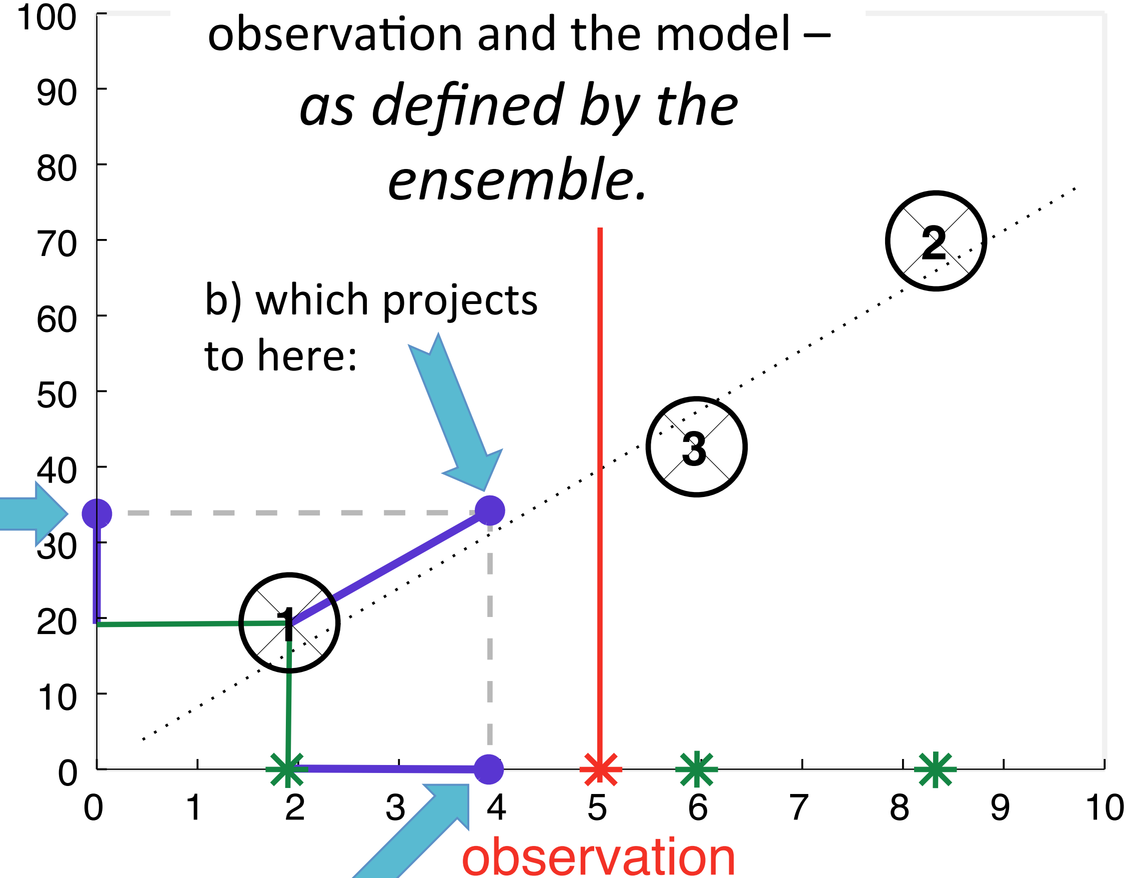
Looking at it another way:

Anderson, J.L., 2003: A local least squares framework for ensemble filtering. *Mon. Wea. Rev.*, **131**, 634-642

The plane defining the relationship between the observation and the model – *as defined by the ensemble.*

b) which projects to here:

c) Which means the unobserved Posterior should be:

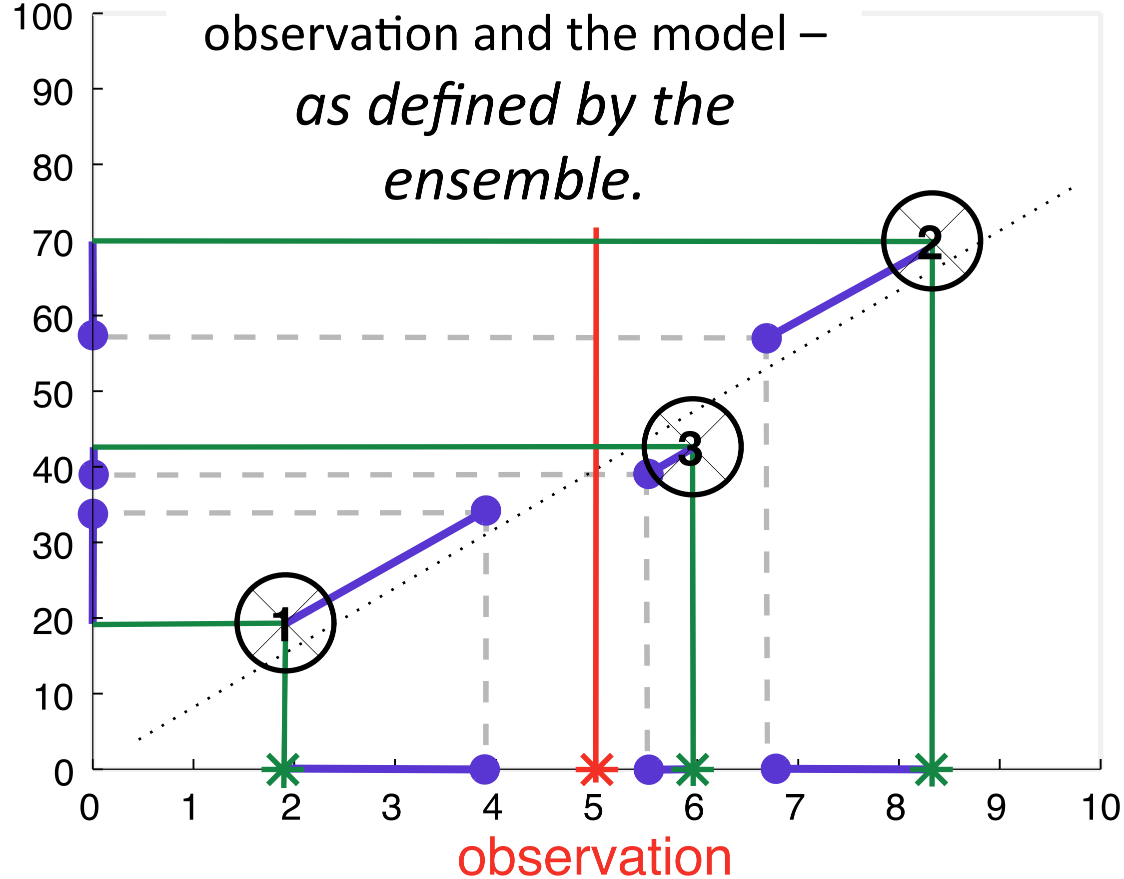


a) The “observation” Posterior for member 1

Looking at it another way:

Any part of the model:
snow cover fraction,
root carbon,
canopy water …
Could even be a model parameter!

The plane defining the relationship between the observation and the model – *as defined by the ensemble.*

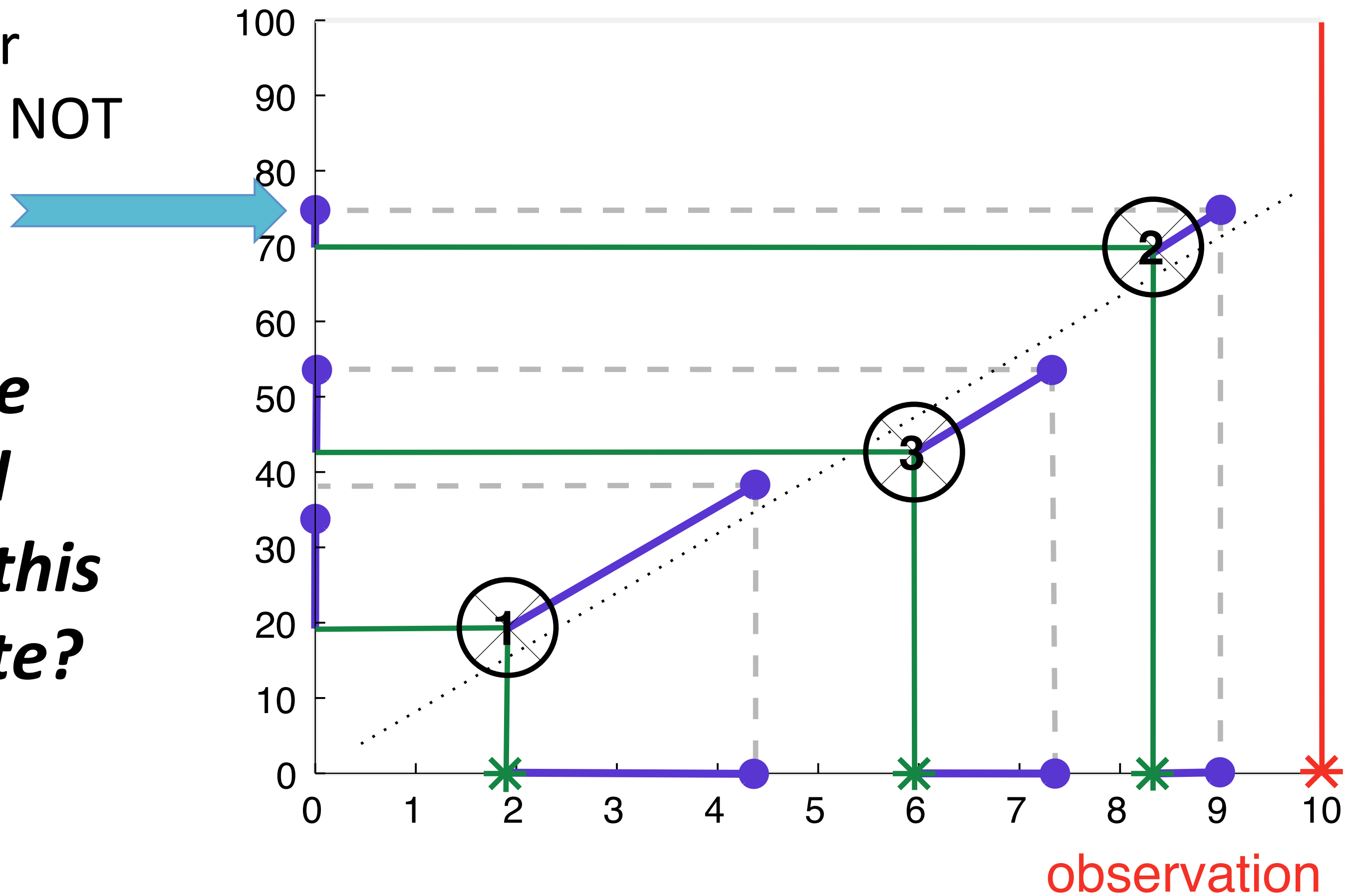


Could be Soil Temperature

Potential Problem

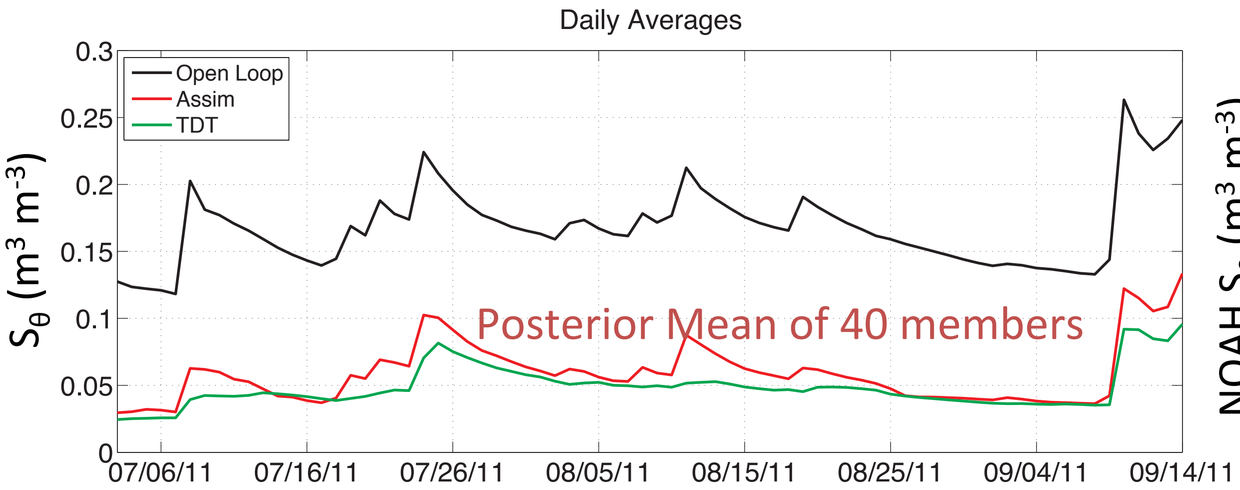
This posterior MAY or MAY NOT be realistic!

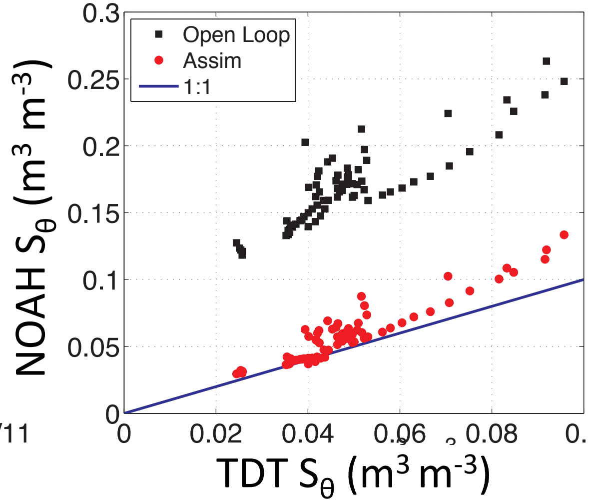
Can the model tolerate this new state?



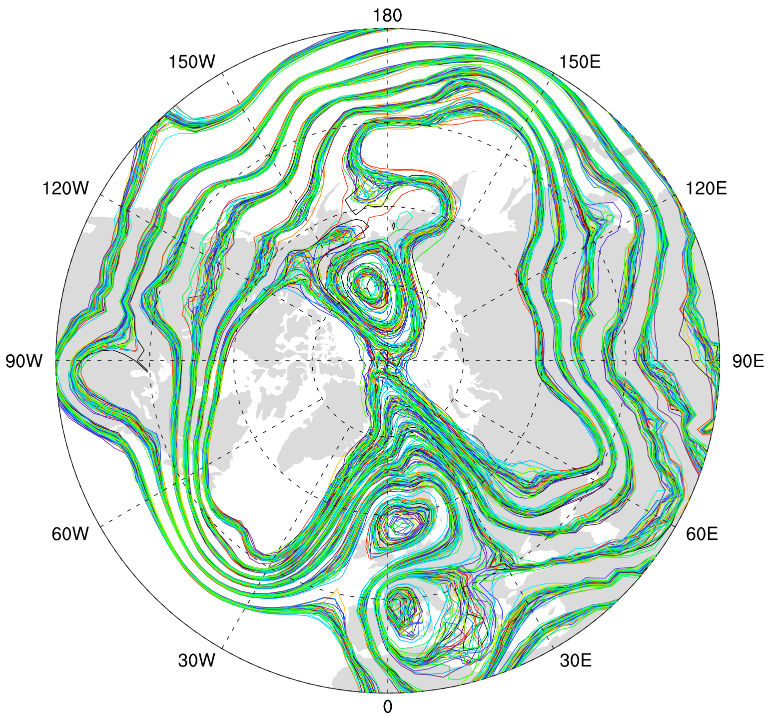
If the observation is "too far" away, it is rejected.
What is "too far"?

NOAH-DART: Integrated Soil Moisture





Pros and Cons

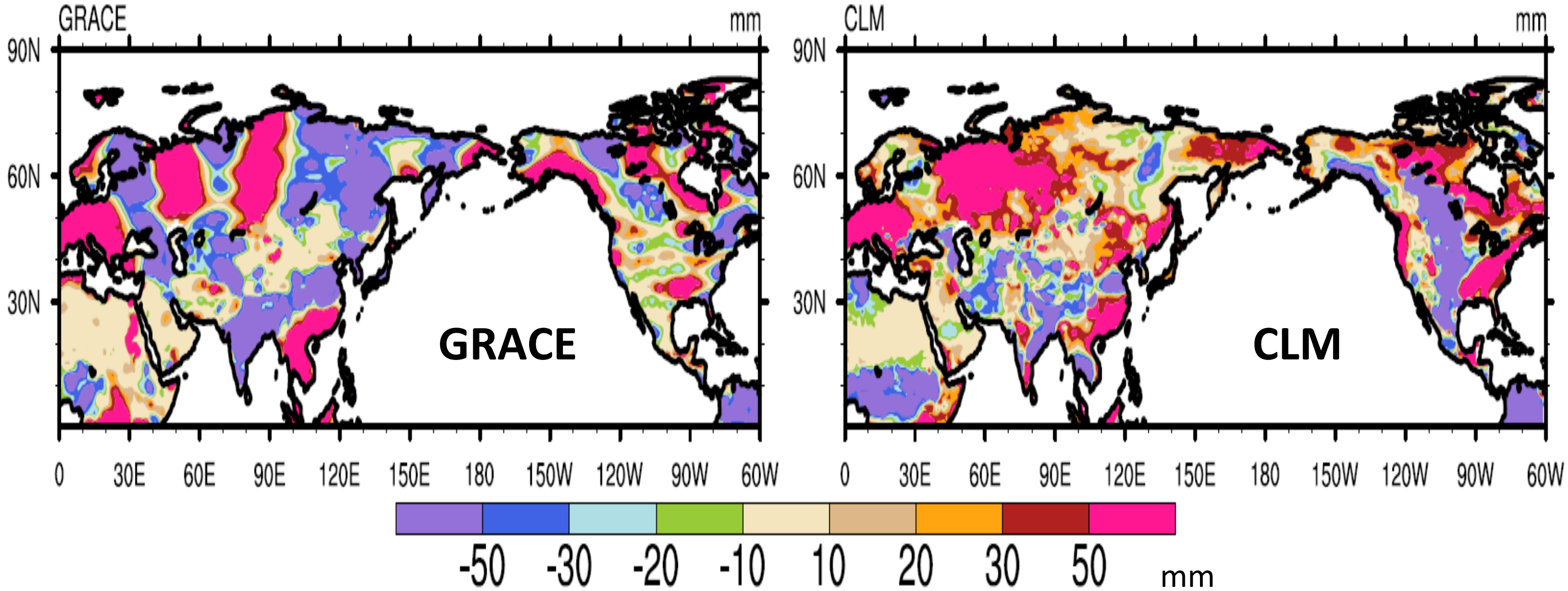


- **80 realizations/members**
- **Model states are self-consistent**
- **Model states consistent with obs**
- **Available every 6 hours for 12+ years**

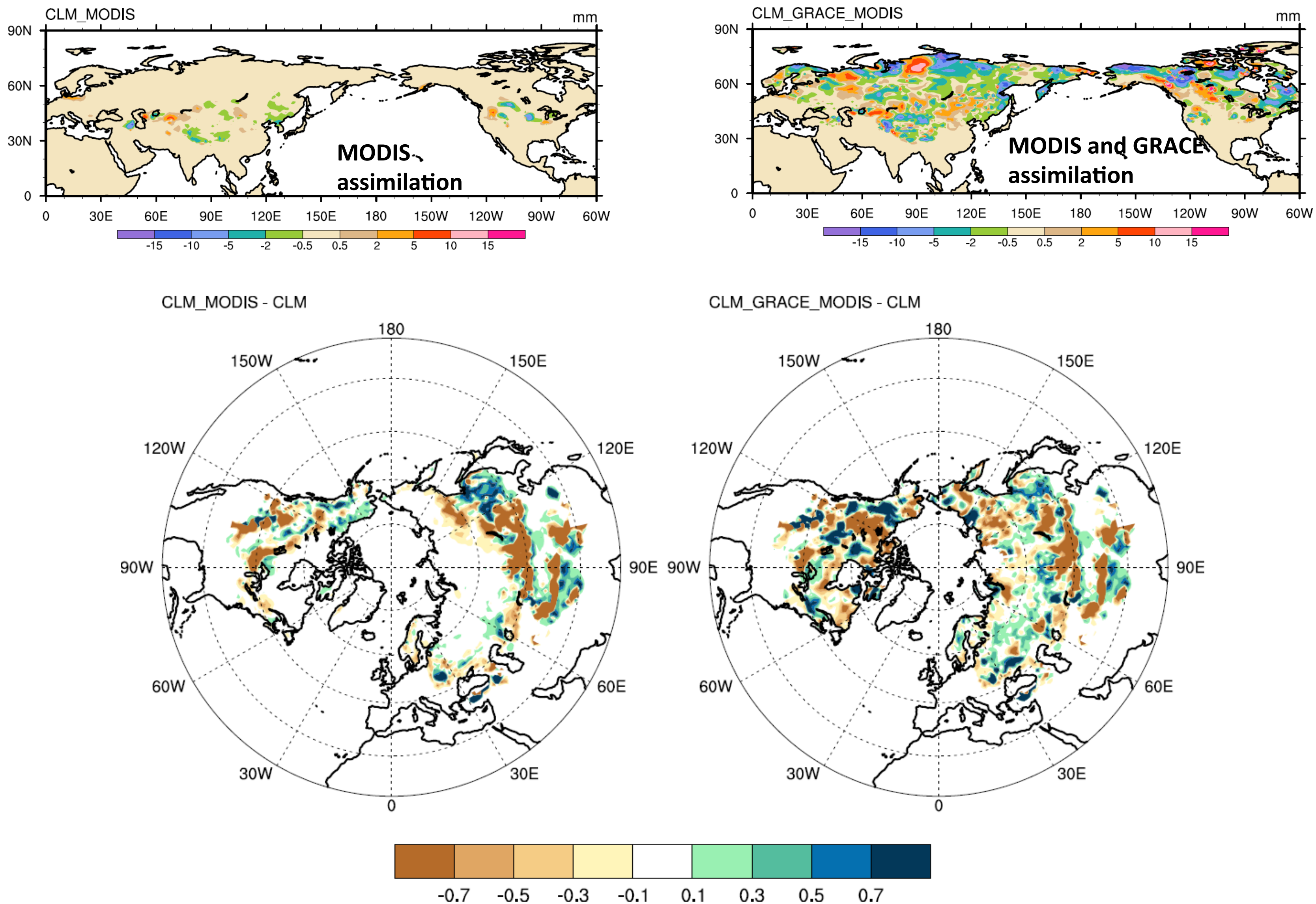
- Relatively low spatial resolution has implications for regional applications.
- Suboptimal precipitation characteristics.
- Available every 6 hours
    - higher frequency available if needed.
- Only have 12 years … enough?

I'm not going to prove it here, but I believe having an **ensemble** of forcing data is **crucial** to land/ocean data assimilation.

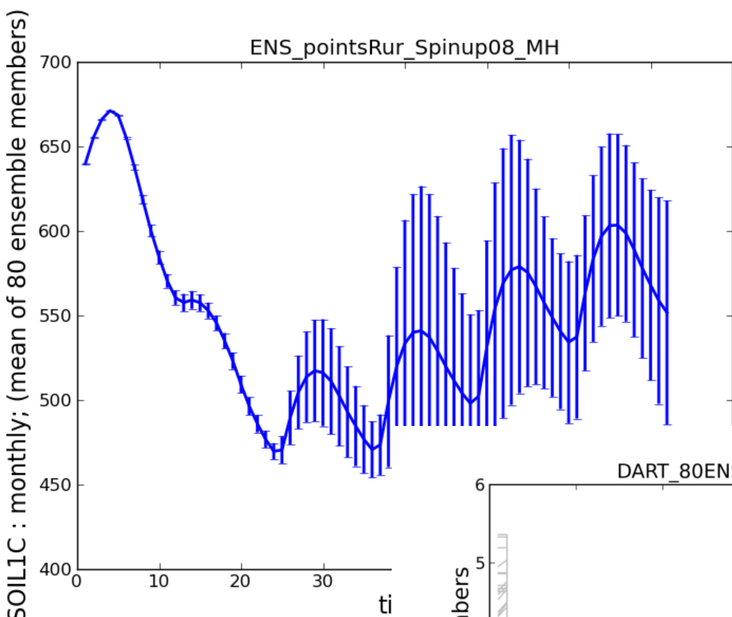
Total Water Storage change
Jan 2003

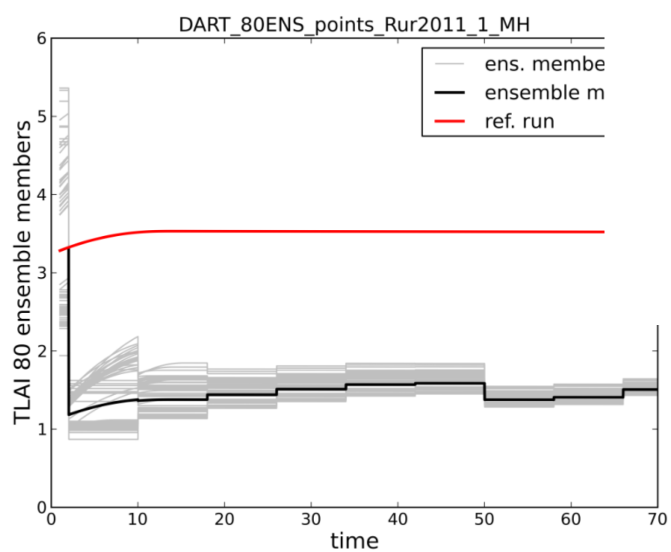


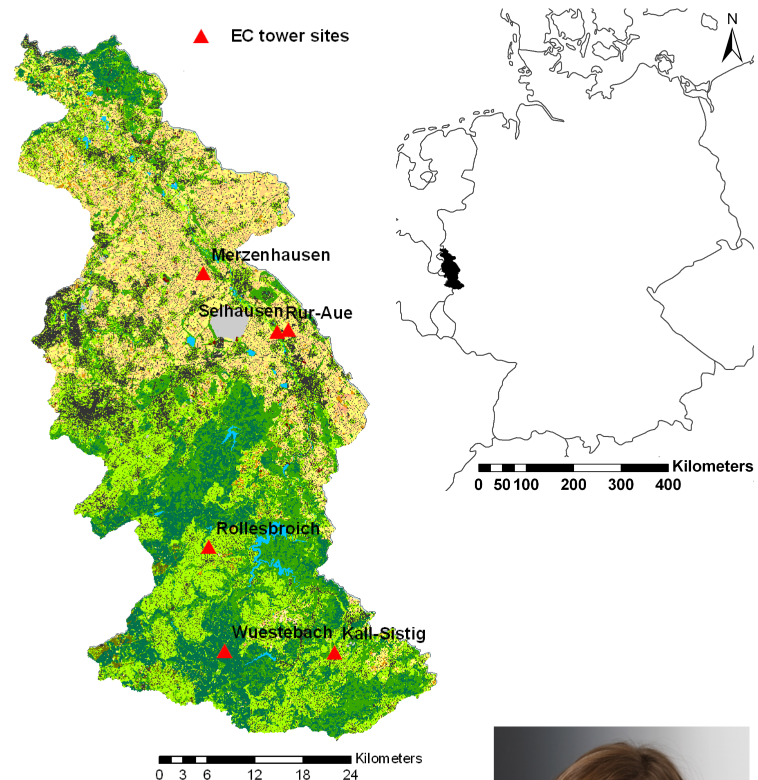
Assimilation Results



- Assimilation of eddy covariance fluxes & MODIS LAI data and CLM upscale NEE from plot to catchment scale



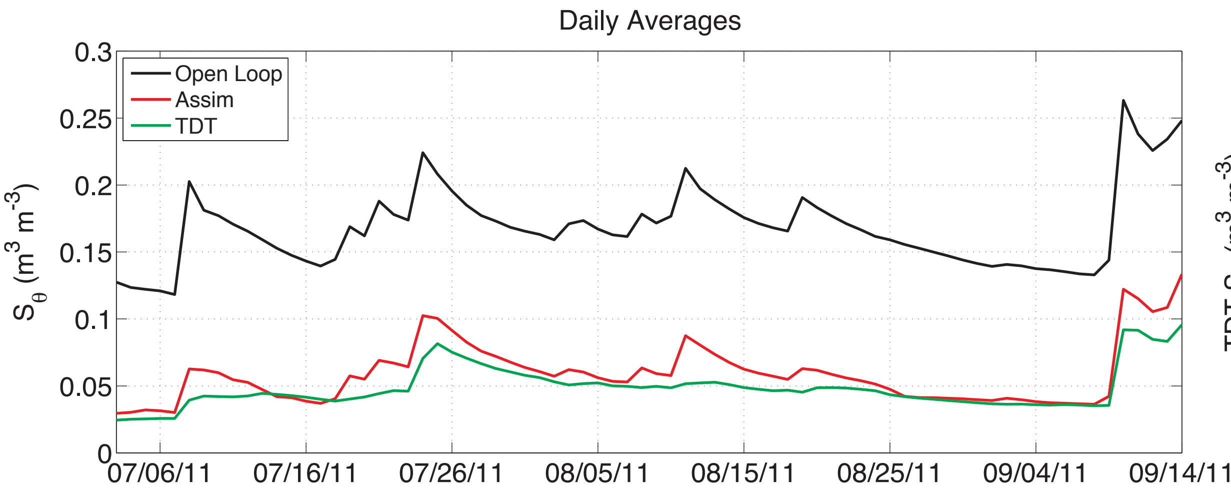


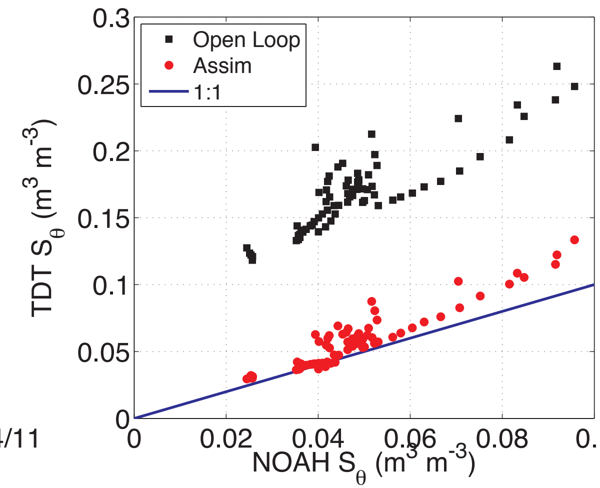


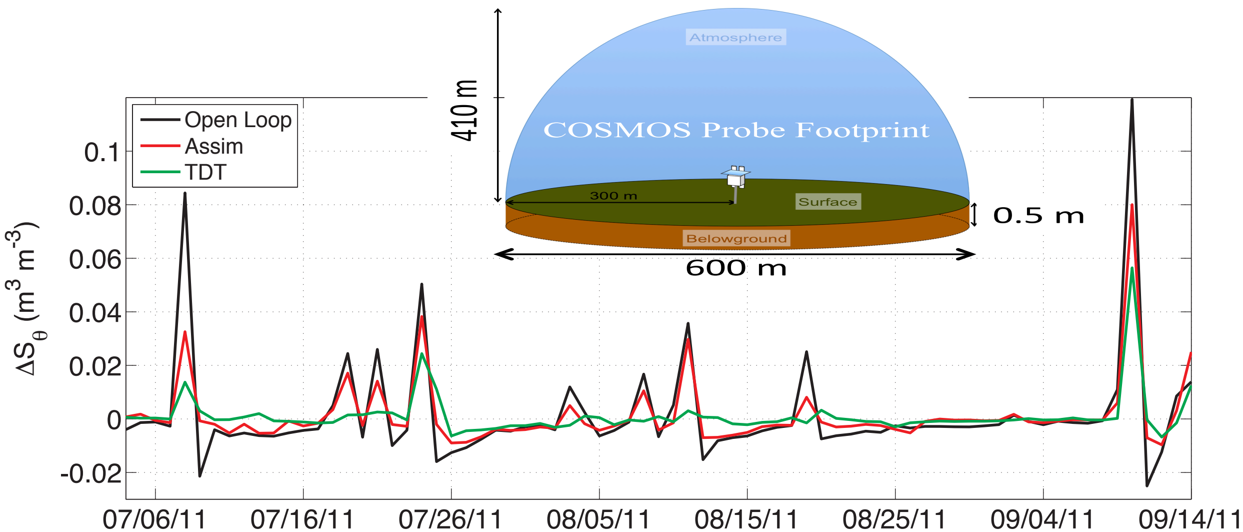
Hanna Post visited Gordon Bonan, Andy Fox and me for 3 months earlier this year.

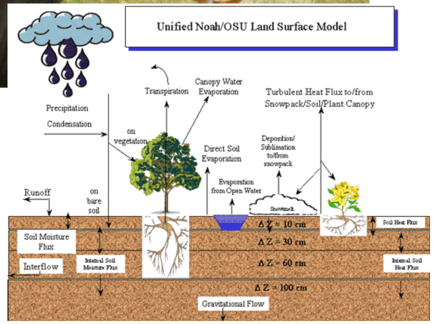
Hanna Post, IBG-3: Agrosphere

NOAH-DART: Integrated Soil Moisture









Future Work: AKA “What I didn’t talk about.”

- ✓ Improved observation metadata / peculiar land model hierarchies …
- ✓ Snow … destroying is easy, making ‘brand new’ snow is hard …
- ✓ Forcing files/data for the resolutions desired …
- ✓ Forward observation operators in support of the instruments …
- ✓ Supporting non-local localizations (eg. watersheds) …
- ✓ The initial ensemble & spread …
- ✓ Identifying model variables that *NEED* to be updated …

And a whole lot more …