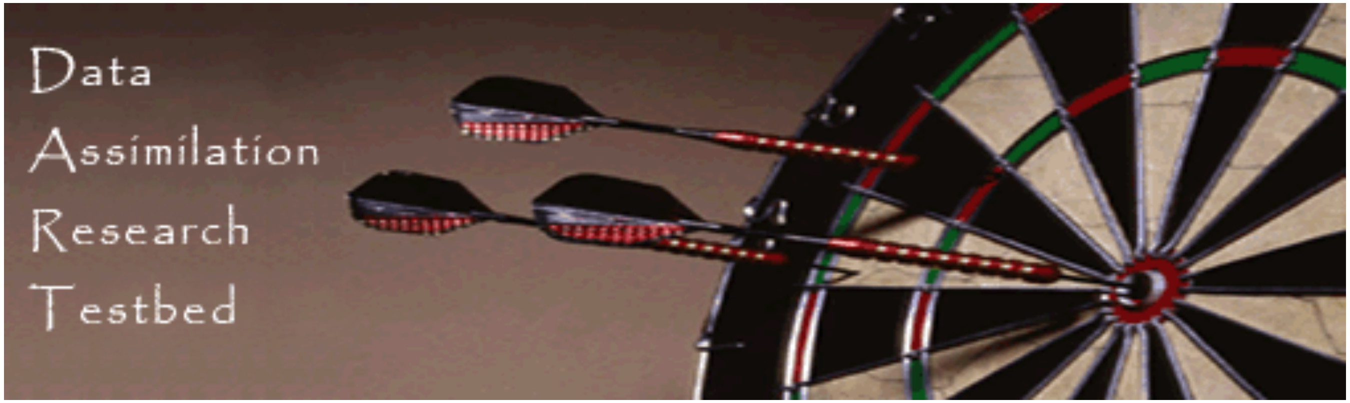
Data Assimilation for CLM: a comprehensive overview

in 12 minutes!



Tim Hoar: *NCAR* with a lot of help from:

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

Yongfei Zhang: *University of Texas Austin*

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

NSF

Motivation

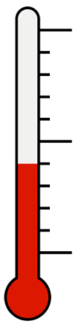
1. The ecological state of the planet is the result of an unknowable disturbance history.
2. Model spinup cannot be counted on to accurately re-create that disturbance history.

Data assimilation can put the model state more in line with the current state. This allows us to:

- Quantify ecological states
  - to establish a baseline
  - as a preface for ecological forecasting
- Better understand our models
- Improve our understanding of the underlying processes.

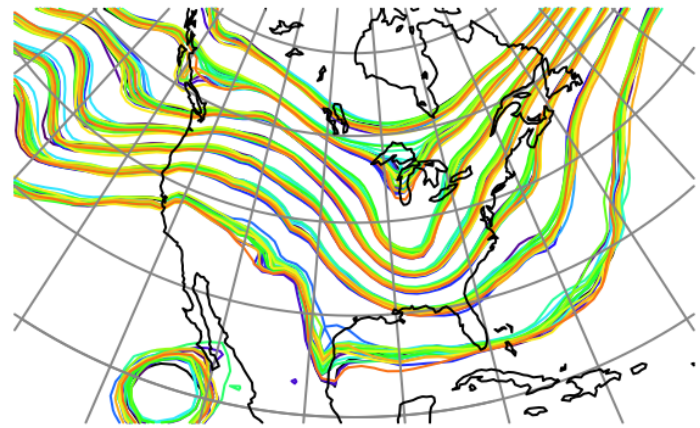
What is Data Assimilation?

Observations combined with a Model forecast…



+

=



… to produce an analysis.

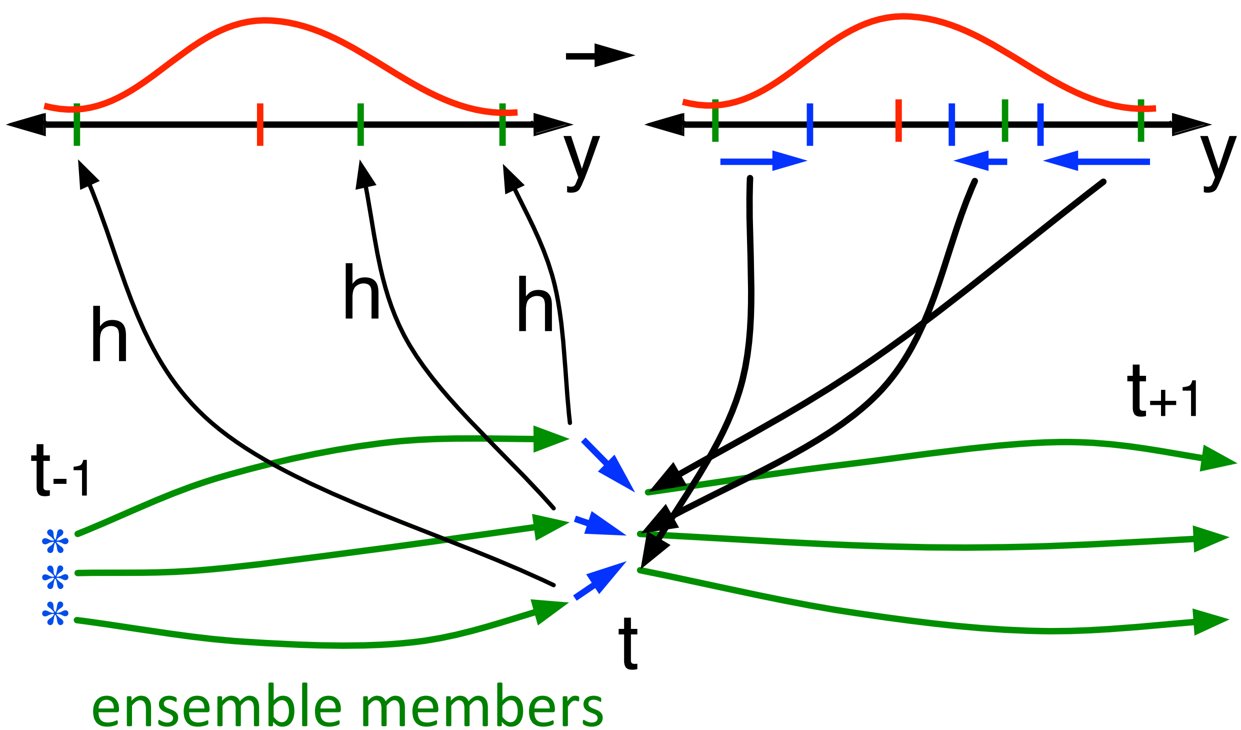
Overview article of the Data Assimilation Research Testbed (DART):

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

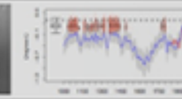
Data Assimilation Research Testbed

A generic ensemble DA 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 in the restart file.

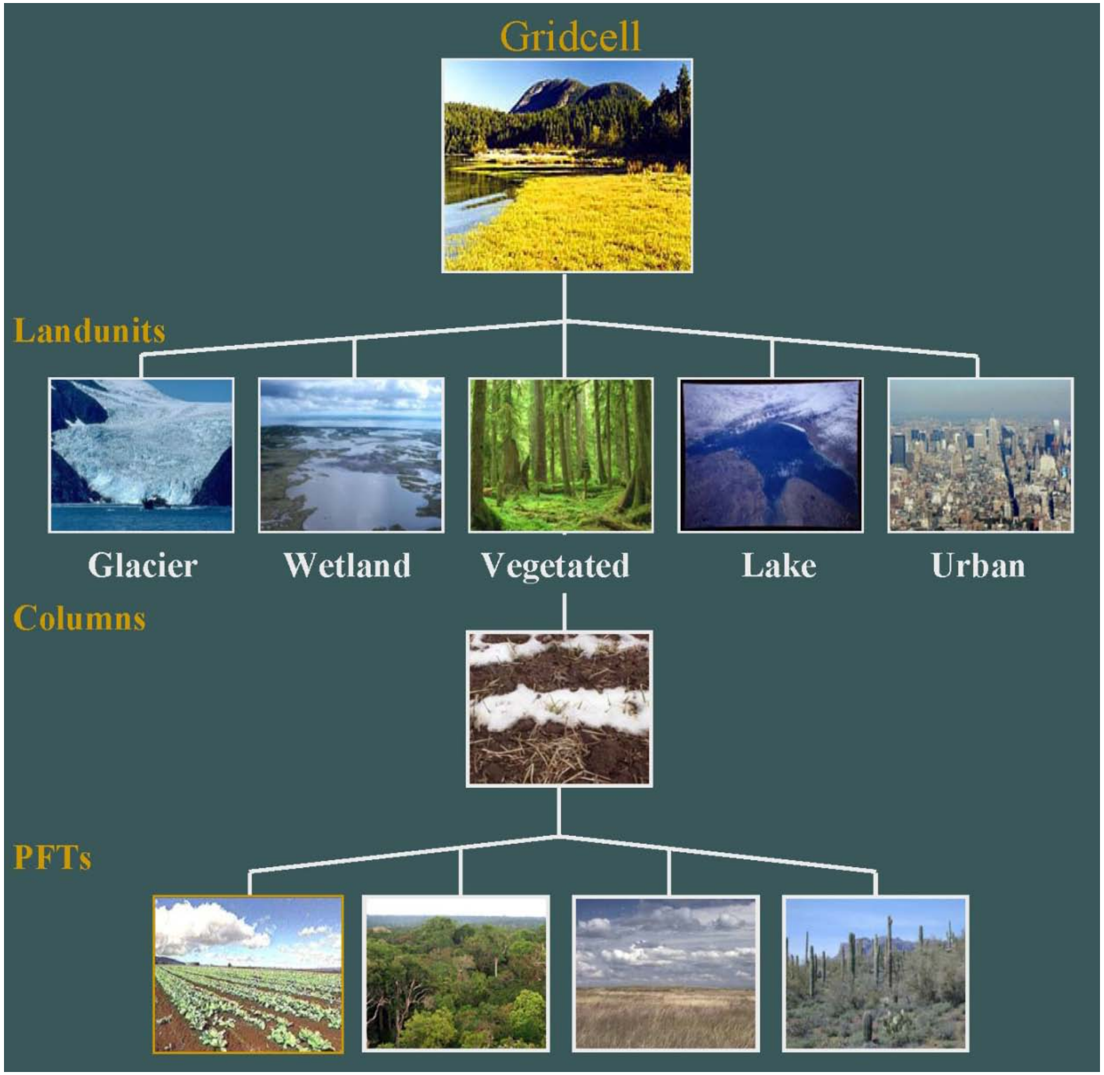


Keys to ensemble land DA:

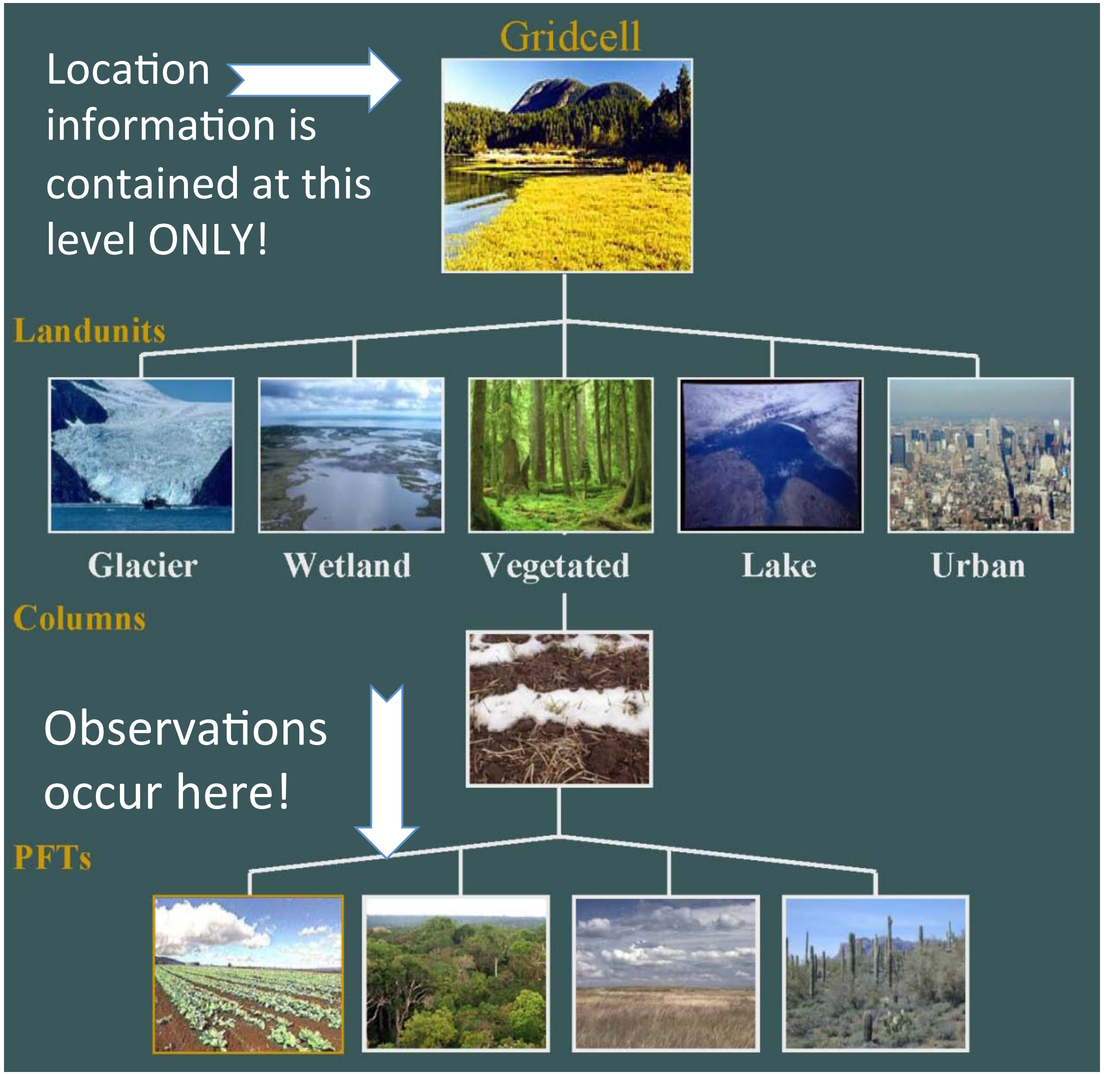
1. What parts of the model 'state' do we update?
   1. The stock CLM restart files have ***hundreds*** of variables in them. *Knowing which ones to update is up to the researcher!*
2. What is a "proper" initial ensemble?
   1. How many model instances do we need?
   2. How do we get them?
   3. Does it maintain realistic uncertainty? Is it still informative?
3. We have imperfect knowledge of the "forcing" fields.
   1. Will the inference change with slightly different forcing?
   2. Does the forcing overwhelm the sparse observations?
4. Can models tolerate new assimilated states?
   1. Model variables not necessarily 'in balance' or consistent anymore. *What happens in a coupled framework?*
   2. Silently fail?

Keys to ensemble land DA (cont'd):

5. What happens when CLM and the observations are in violent disagreement? *Can only be answered by the researcher!*
    1. Snow vs. bare ground
    2. Senescence, etc.
6. Assimilation affects bounded quantities.
    1. Soils dry beyond their physical limits, for example.
7. Need forward observation operators.
    1. How do we estimate the observation value given the CLM state? Ally Touré [NASA] here now to do this for AMSR-E brightness temperatures.
8. Observation metadata is very important for accurate forward observation operators. *This is the next thing on my to-do list.*
    1. Location information alone is insufficient. Land cover type needed.



CLM abstracts the gridcell into a "nested gridcell hiearchy of of multiple landunits, snow/soil columns, and Plant Function Types". This is particularly troublesome when trying to convert the model state to the expected observation value ***because***:



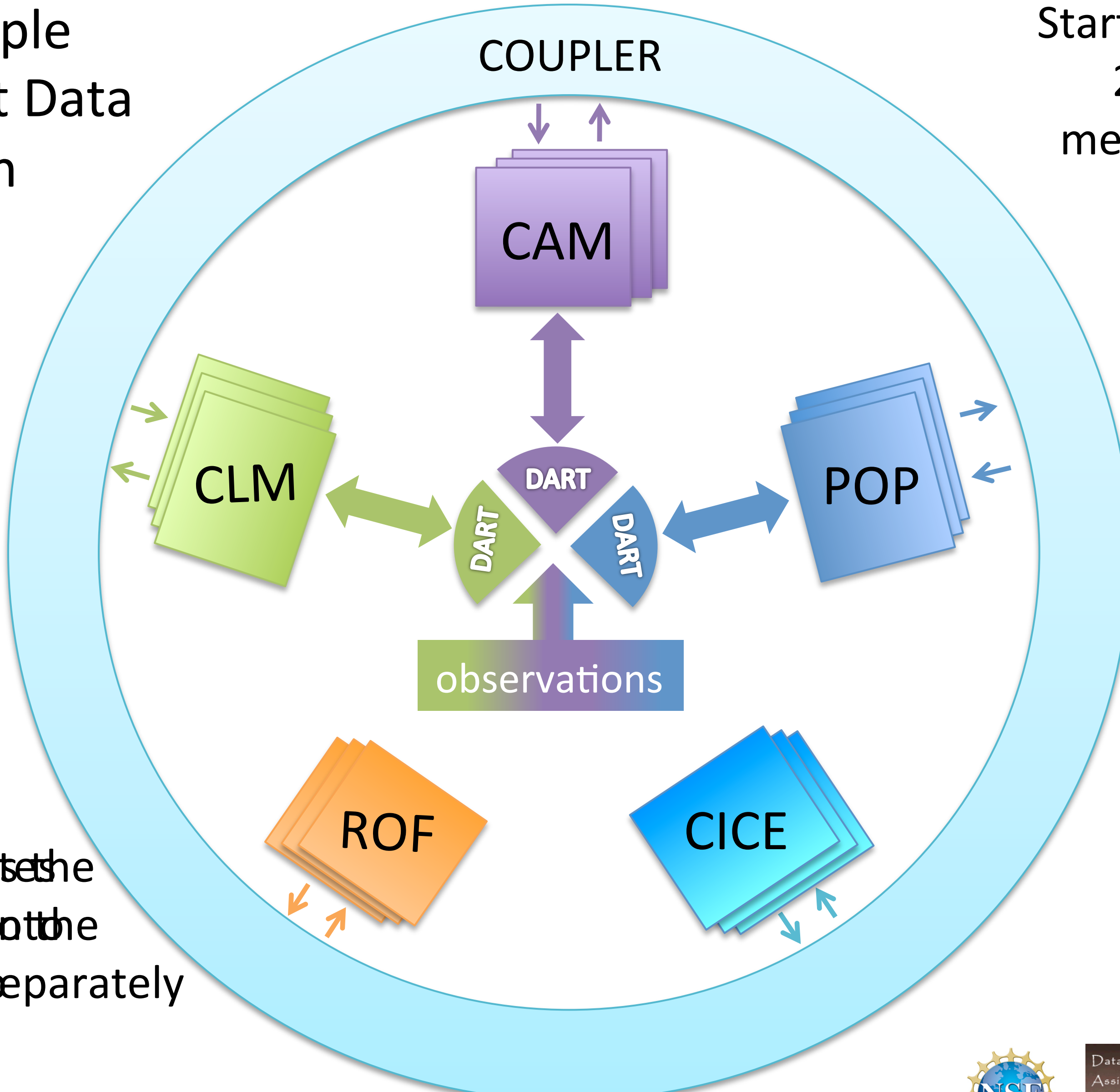
CLM abstracts the gridcell into a “nested gridcell hiearchy of of multiple landunits, snow/ soil columns, and Plant Function Types”. This is particularly troublesome when trying to convert the model state to the expected observation value ***because***:

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!*

Data Assimilation Research Testbed

DART Multiple Component Data Assimilation

Started with CCSM4 20th Century 30-member ensemble for all model components



Important!
There are *multiple* instances of each model component.

DART assimilates the observations into the components separately

B compset CESM1_1_1

Data Assimilation Research Testbed

Check out Yongfei's poster!

Assimilation of the MODIS Snow Cover Fraction data through DART/CLM4

Y. Zhang[1], T. J. Hoar[2], Z.-L. Yang[1], J. L. Anderson[2], A. Toure[3,4], M. Rodell[4]

1. Jackson School of Geosciences, University of Texas at Austin, Austin, TX, United States.
2. The National Center for Atmospheric Research, Boulder, CO, United States.
3. Universities Space Research Association (USRA), Columbia, MD, United States.
4. NASA Goddard Space Flight Center, Greenbelt, MD, United States

(yongfei@utexas.edu)

Introduction

✓ Snow plays a unique role in global water and energy cycles. The special physical properties (high albedo, low thermal conductivity, and phase change ability) largely modulate energy and water exchanges between the atmosphere and the land surface. As a common snowpack measurement, snow water equivalent (SWE) is the amount of water contained within the snowpack, which is important for water resources management and hydrological forecasts in regions where streamflow depends on snowmelt. However, high-quality large-scale SWE datasets are generally not available.

✓ Some recent studies have demonstrated the value of satellite-retrieved SWE data on local or regional scales. This study will develop and refine, on global scales, a multi-sensor data assimilation system, through which observations of MODIS SCF and GRACE terrestrial water storage (TWS) change as well as other high-quality satellite data can be assimilated.

The DART/CLM4 Data Assimilation System

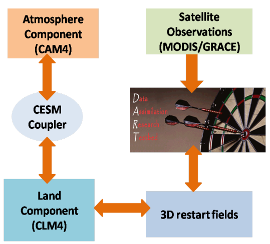


Fig. 1 Schematic of the data assimilation system.

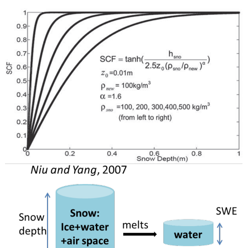


Fig. 2 (upper) The visualization of SCF scheme in CLM4. (bottom) The concept of SWE.

✓The Data Assimilation Research Testbed (DART) (http://www.image.ucar.edu/DAReS/DART/) is a comprehensive data assimilation software environment that can help modelers and observational scientists easily explore a variety of data assimilation methods and observations with different numerical models. This study represents the first effort of linking DART and a land surface model.

✓The Community Land Model version 4 (CLM4), one of the state-of-art land surface models, simulates a snowpack with multi-layers (1-5 layers) depending on its thickness, and accounting for internal physical processes such as water-heat transport, thawing-freezing, liquid water retention, and densification. The snow cover fraction is a function of snow density following Niu and Yang (2007).

Meteorological Forcings from DART/CAM4

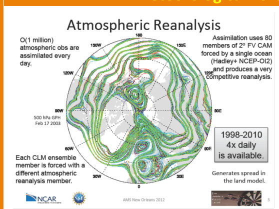


Fig. 3 Geopotential heights (500 hPa) for half (40) of the ensemble members typically used in DART/CAM assimilations for 1200 UTC 17 Feb 2003 [*Hoar et al.*, 2012].

✓ A freely available ensemble of reanalysis data created by DART and the Community Atmospheric Model (CAM4) is used to drive the CLM ensemble members.

✓ The CAM4 reanalysis is similar to the NCEP reanalysis, except the former is an ensemble and a product of the coupled DART and CAM4.

✓ The CAM-produced ensemble reanalysis forcing fields are physically and mutually consistent for a given member and exhibit ensemble spread spatio-temporally.

✓The reanalysis may inherit some of the systematic biases that are found in the CAM model.

✓ **Compare precipitation to Global Precipitation Climatology Project (GPCP)**

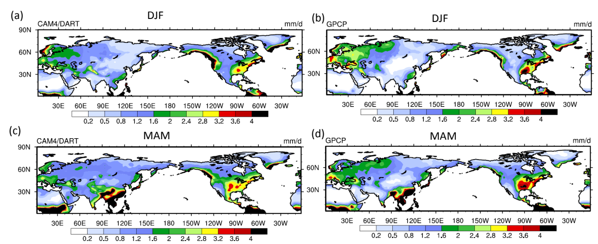


Fig. 4 10-year (1999-2008) DJF mean of precipitation for (a) CAM4/DART and (b) Global Precipitation Climatology Project (GPCP), and MAM mean of precipitation for (c) CAM4/DART and (d)GPCP.

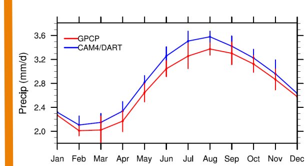


Fig. 5 10-year (1999-2008) mean seasonal cycle of precipitation for GPCP and CAM4/DART. The error bars represent 2 standard deviations.

✓ Compared to GPCP, CAM4/DART produces more precipitation over Canada, the western America and the central Siberia, and less precipitation over the eastern America, and the western Eurasia.

✓ The seasonal cycle of CAM4/DART precipitation has comparable amplitude to GPCP.

✓ While CAM4 tends to have cold bias and excessive precipitation in the Arctic region (de *Boer et al.*, 2011), GPCP is found to underestimate precipitation in some regions (*Adler et al.*, 2003)

Satellite Observations

✓ MODIS/Terra daily snow cover (MOD10C2; 0.05° resolution; northern hemisphere) Retrieved using NDSI [*Salomonson et al.*, 2004]

$$\mathrm{NDSI} = \frac{\text{band 4} - \text{band 6}}{\text{band 4} + \text{band 6}}$$

✓ Pre-processed to 0.9 ° x1.25 ° "Level 4" data following *Rodell and Houser* [2004].
Pixels with lower than 20% confidence index (percentage of clear sky over certain grids) will be discarded.

Results

✓ **Localization**: Localization: a technique to reduce sampling error by limiting the influence of observations to nearby grid cells.

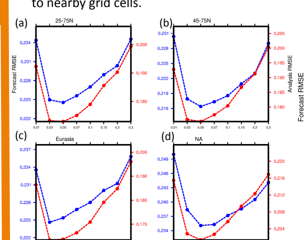


Fig. 6 Variations of forecast RMSE (blue dots on the left Y axis) and analysis RMSE (red dots on the right Y axis) of eight experiments with localization distances (radians on the X axis) for (a) 25–75N, (b) 45-75N, (c) Eurasia, and (d) NA.

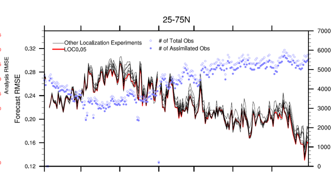


Fig. 7 Evolution of daily forecast (prior) RMSE of SCF in the latitudinal bands spanning from 25°to 75°N. Blue circles show the number of observations available and blue discs show the number of observations that are actually assimilated at each time. Red line represents RMSE of the experiment that uses a localization distance of 0.05 radians and black ensemble lines show RMSE of experiments with other seven localization distances (0.01, 0.03, 0.07, 0.1, 0.15, 0.2, and 0.3 radians).

✓ **Ensemble spread**

✓ **Innovation**

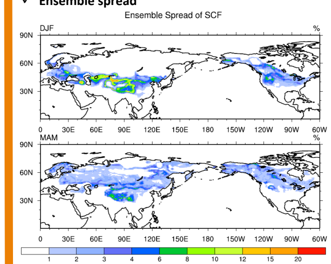


Fig. 8 Ensemble spread of SCF for (a) DJF and (b) MAM in 2002—2003. Ensemble spread is calculated as the standard deviation of SCF among 40 ensemble members. SCF values are averaged for two seasons before calculating the ensemble spread.

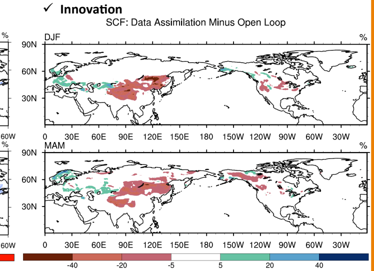


Fig. 9 The difference of SCF between the data assimilation case and the open loop case averaged for (a) DJF and (b) MAM.

✓ **Time series of SCF**

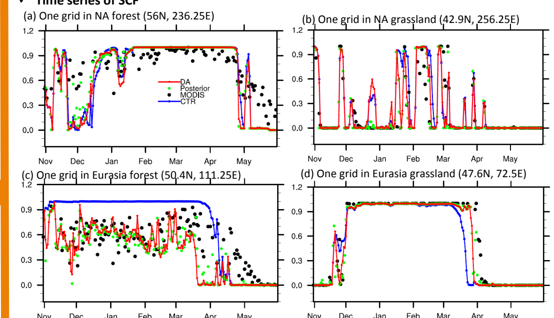


Fig. 10 Time series of SCF for four grids: (a) one grid in NA forest, (b) one grid in NA grassland, (c) one grid in Eurasia forest, and (d) one grid in Eurasia grassland. Blue line represents the CLM4 open loop simulation. Black dots represent MODIS observed SCF. Green dots represent the posterior SCF. Red line represents the DART/CLM4 result that assimilates SCF but updates SWE.

Conclusions

✓A localization distance of 0.05 radians stands out among a series of localization distances, producing the smallest RMSE.

✓In winter, SCF ensemble spread is mainly located in lower-middle latitude regions. In spring, the spatial pattern of SCF ensemble spread extends northward, indicating that the uncertainty of modeled snow in high latitude regions increases as snow starts melting.

✓Snow data assimilation shows little change on SCF at higher-middle and high latitudes in winter due to the fact that SCF in CLM4 reaches the unity too fast compared to MODIS data.

✓The effectiveness of data assimilation on model states varies with vegetation types, with mixed performance over forest regions and consistently good performance over grassland areas.

References

Adler, R. F., G. J. Huffman, A. Chang, R. Ferraro, P.-P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, A. Cruber, J. Susskind, P. Arkin,and E. Nelkin (2003), The version-2 GPCP monthly precipitation analysis(1979-present), *J. Hydrometeoro.*, *4*, 1147-1167.

de Boer, G., W. Chapman, J. Kay, B. Medeiros, M. Shupe, S. Vavrus, and J. Walsh (2011), A Characterization of the Present-Day Arctic Atmosphere in CCSM4, *J. Climate.*, *25*, 2676–2695.

Hoar. T., Data assimilation with CLM & DART, presented at the 17th CESM Workshop in Breckenridge, CO, USA.

Niu, G.-Y., and Z.-L. Yang (2007), An observation-based formulation of snow cover fraction and its evaluation over large North American river basins, *J. Geophys. Res.*, 112, D21101, doi:10.1029/2007JD008674.

Salomonson, V. V. and I. Appel (2004), Estimating fractional snow cover from MODIS using the normalized difference snow index.

Acknowledgements

This work is supported by NASA Grant NNX09AJ48G and the NCAR Advanced Study Program.

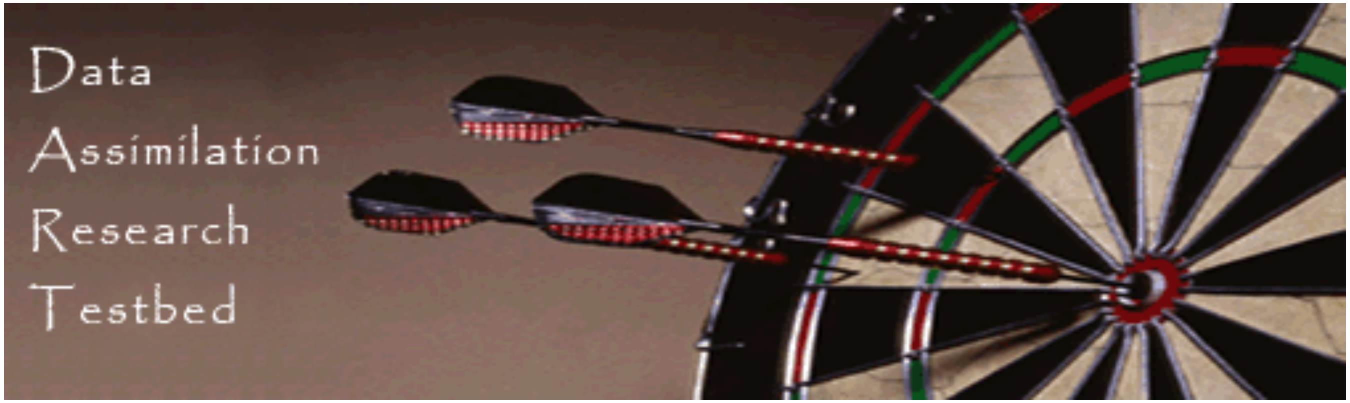
For more information:

CAM

WRF

GITM

CLM



POP

AM2

BGRID

COAMPS

www.image.ucar.edu/DAReS/DART

NOAH

dart@ucar.edu

MPAS_ATM

MITgcm_ocean

COAMPS_nest

SQG

NAAPS

MPAS_OCN

TIEGCM

PBL_1d

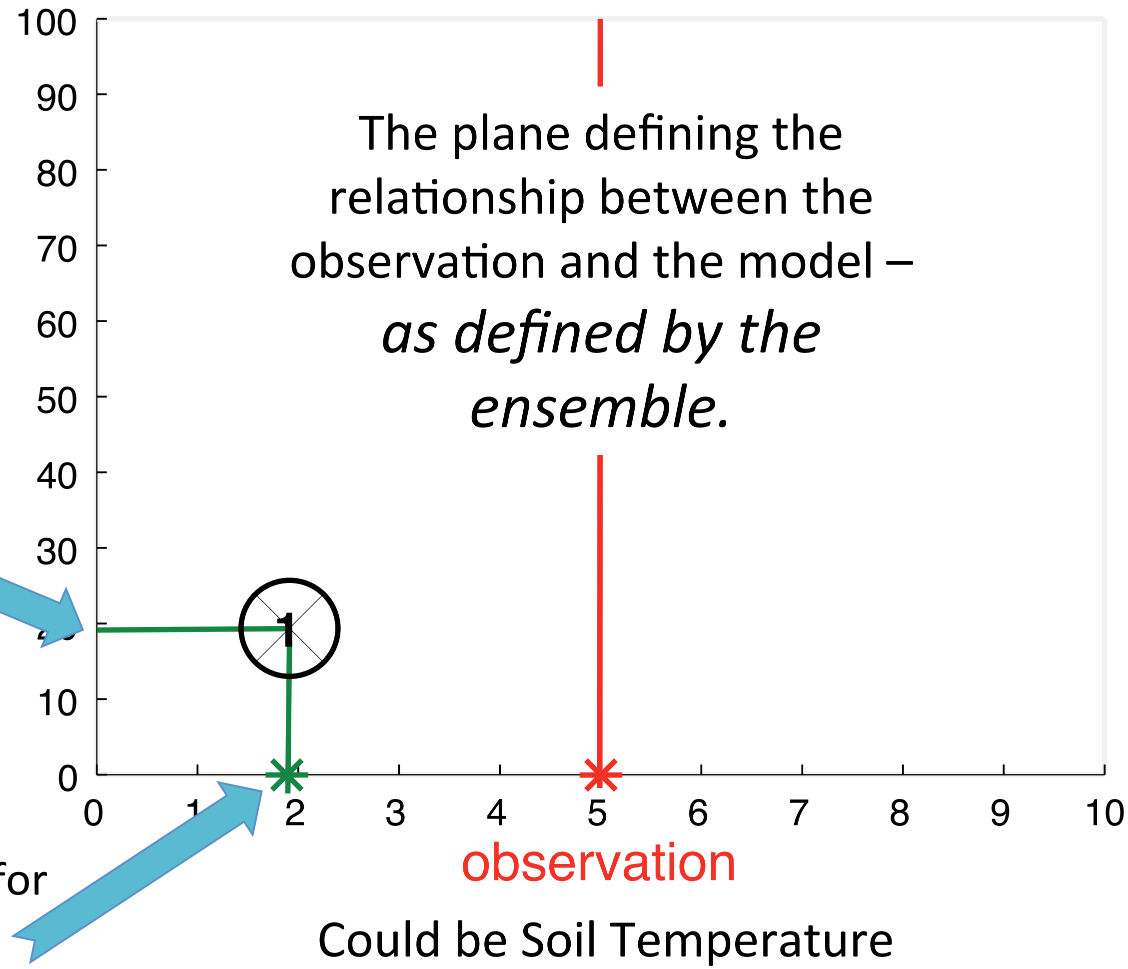
PE2LYR

NCOMMAS

Slides held in reserve

Looking at it another way:

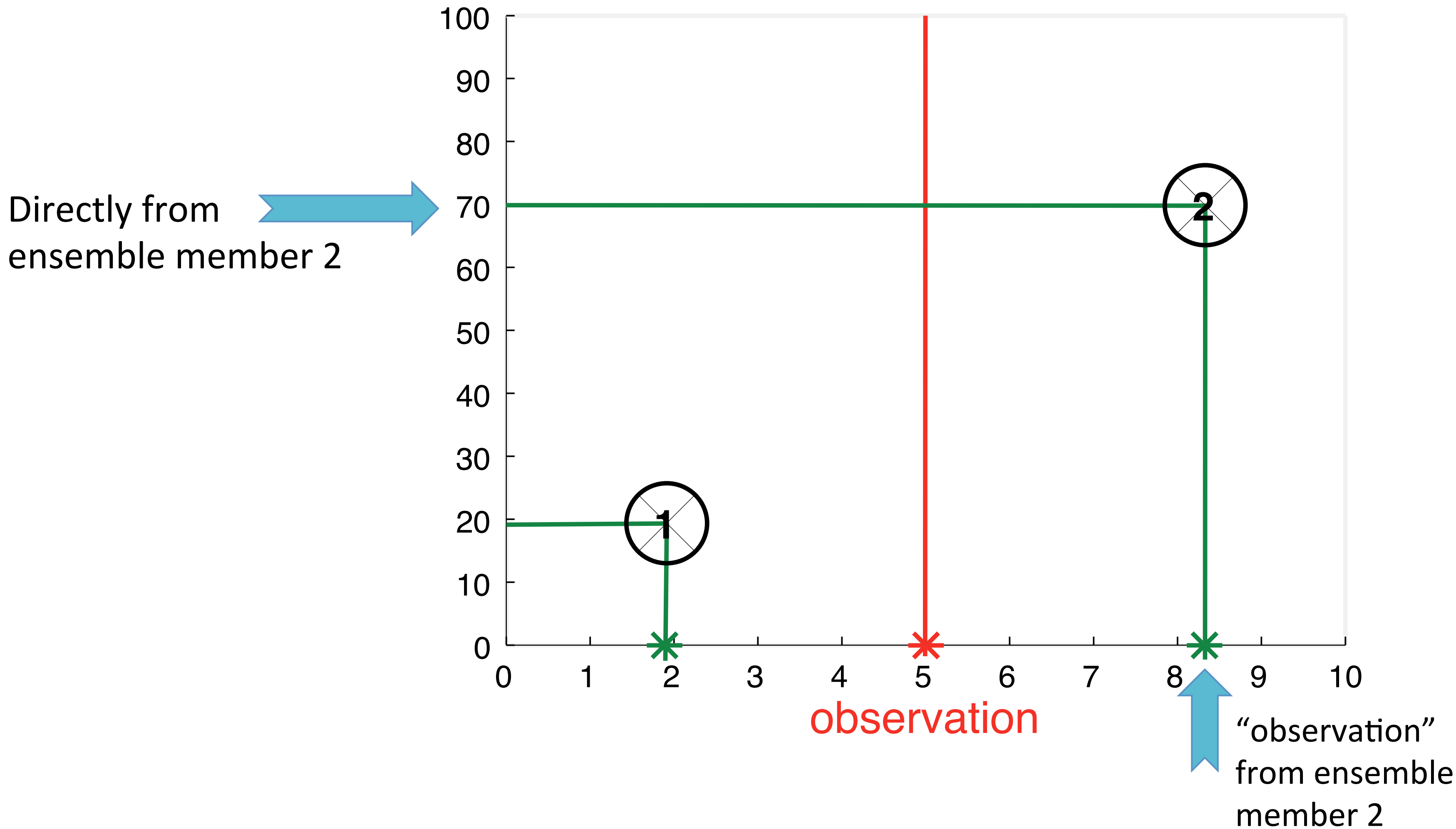
Some unobserved state variable. e.g. live root carbon, dead root carbon, canopy water …



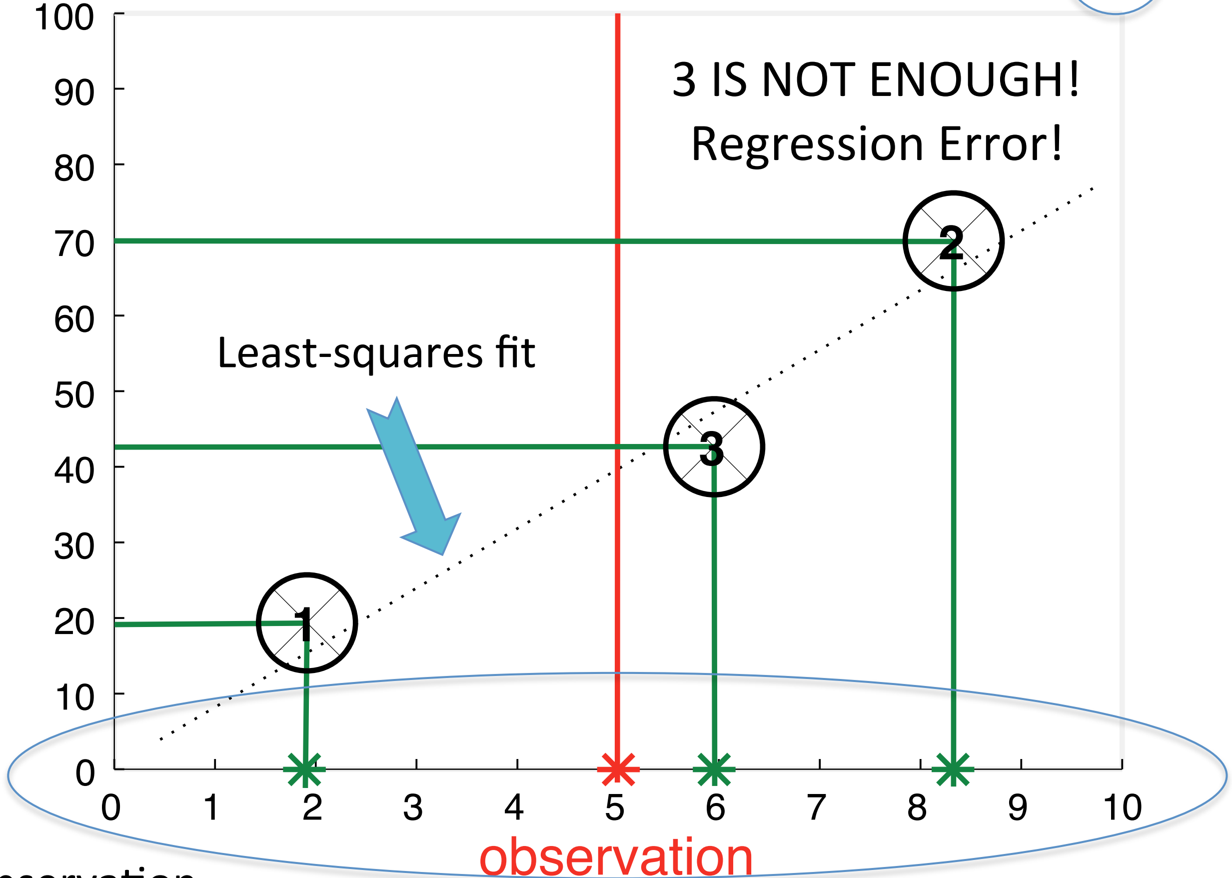
Directly from ensemble member 1

Result of the forward observation operator for ensemble member 1

Data Assimilation Research Testbed


Directly from ensemble member 2
100
90
80
70
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10
1
2
observation
"observation" from ensemble member 2

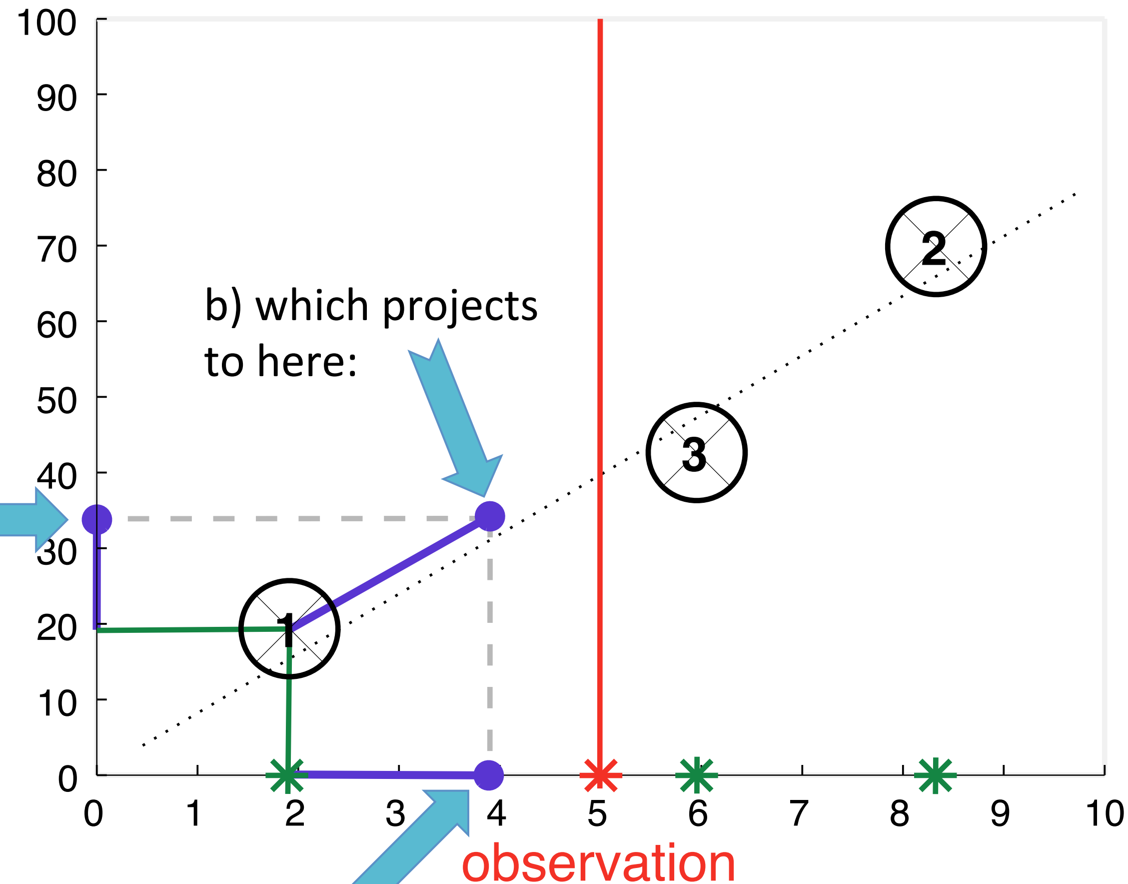
In our global atmospheric assimilations, we use 80.



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

c) Which means the unobserved Posterior should be:

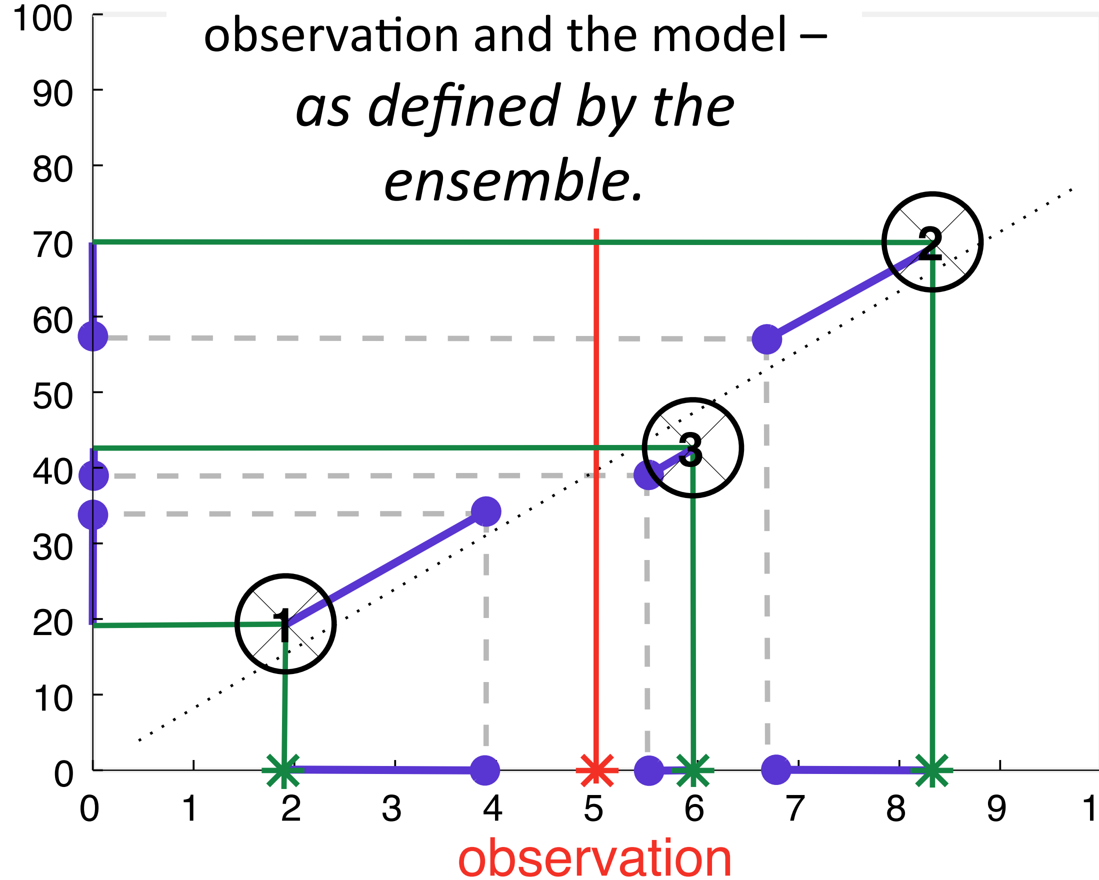
b) which projects to here:



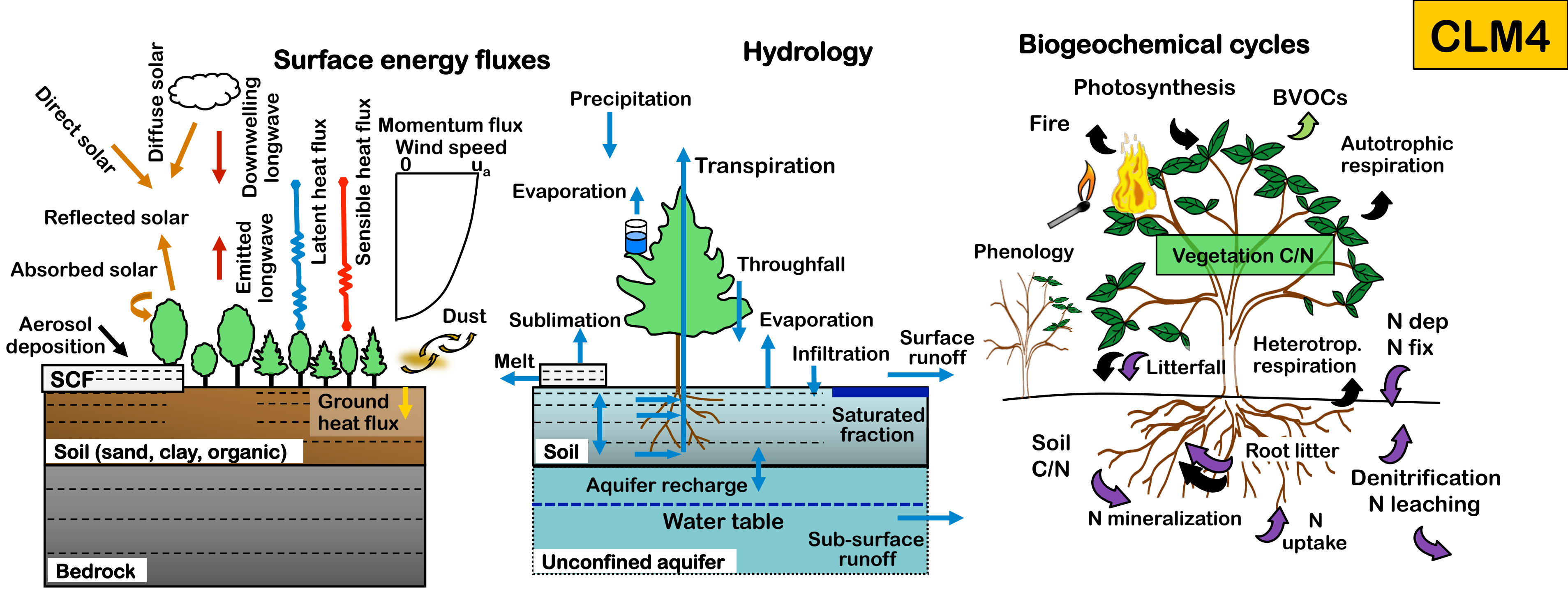
a) The "observation" Posterior for member 1

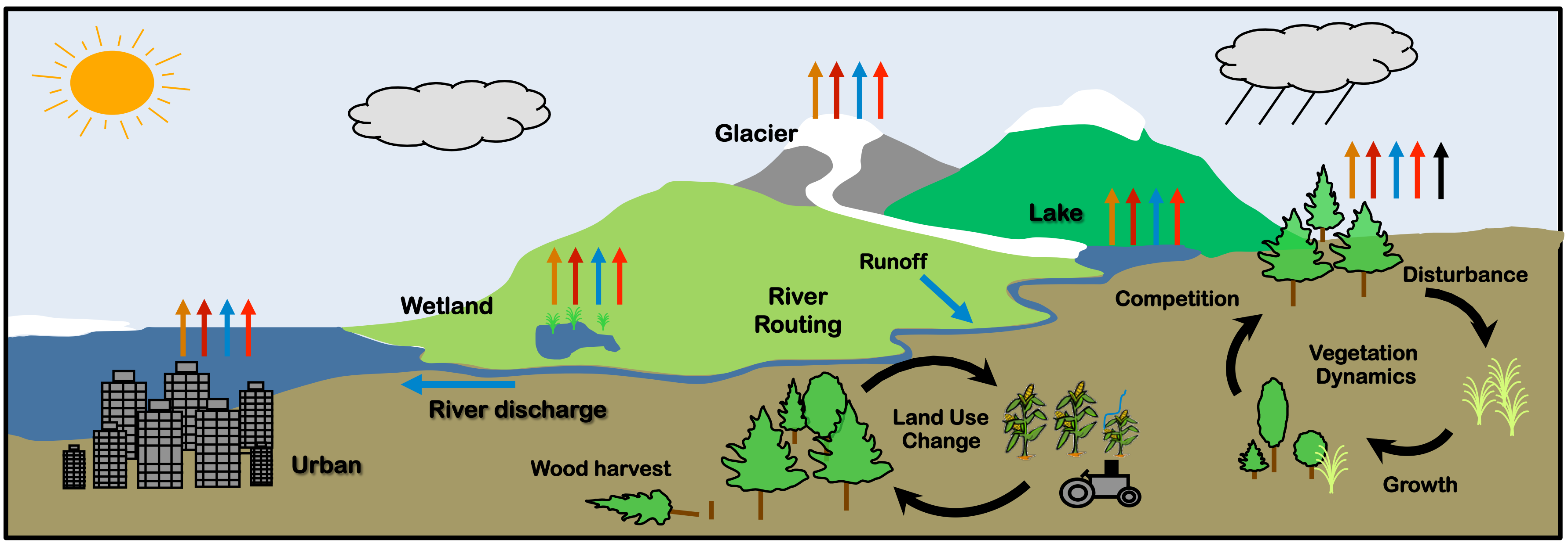
The plane defining the relationship between the observation and the model – *as defined by the ensemble.*

Some unobserved state variable like: live root carbon, dead root carbon, canopy water …

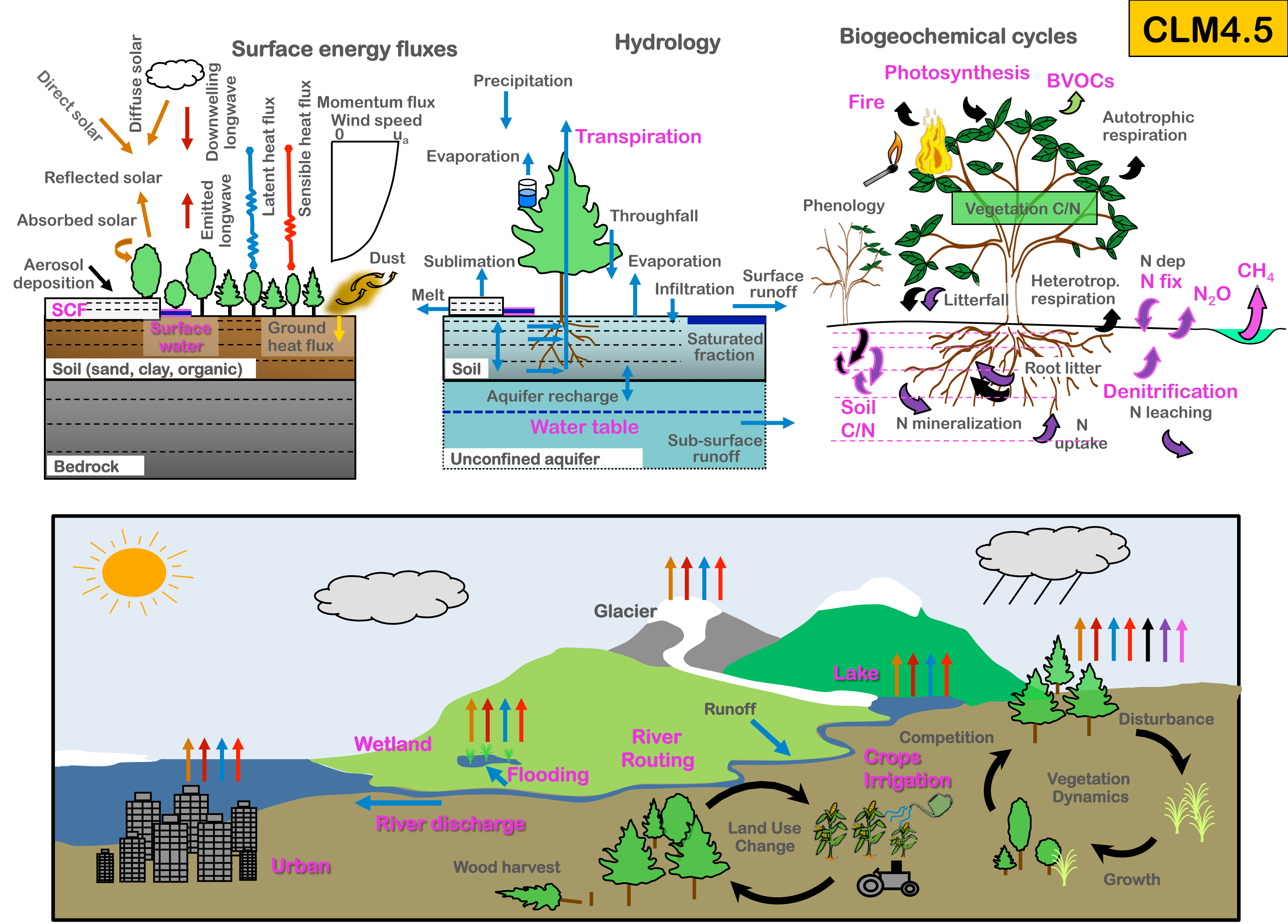


Could be Soil Temperature





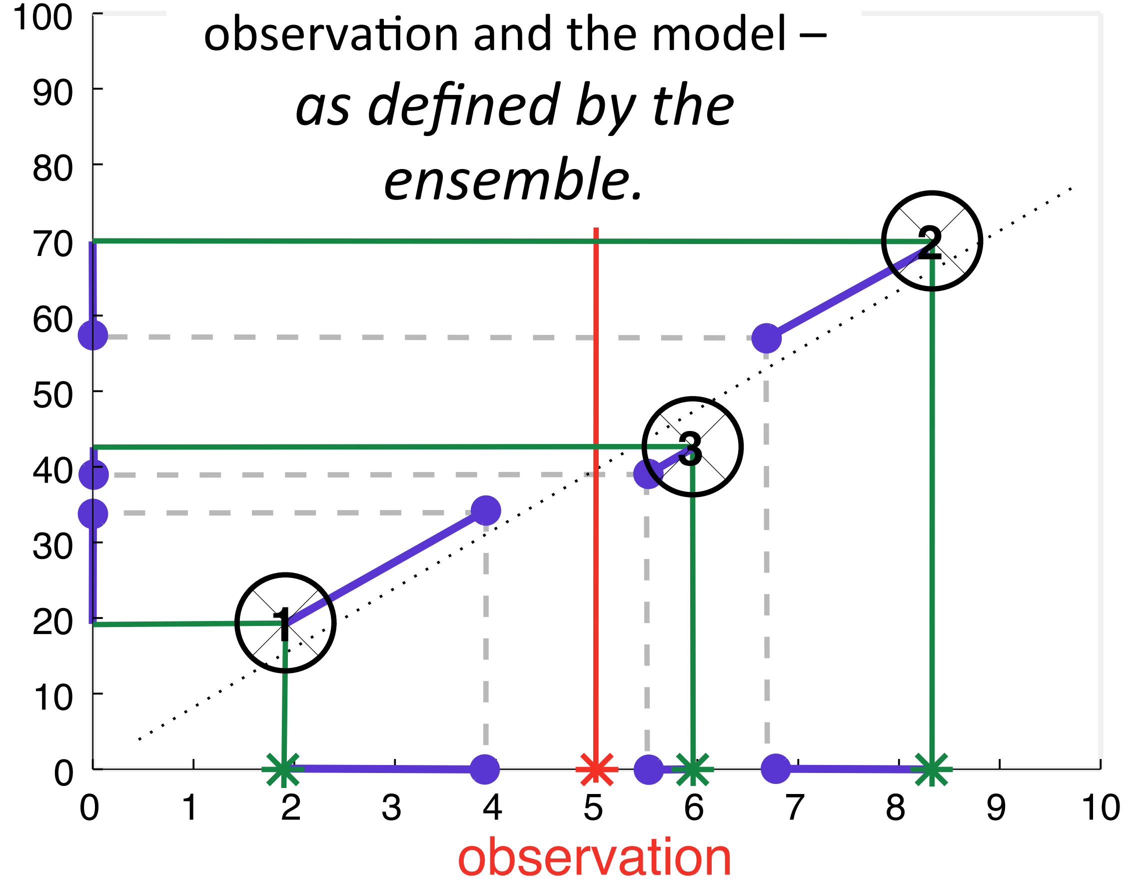
I got these from Dave Lawrence. I don't know if he made them or not – but Thanks to whomever did!



I got these from Dave Lawrence. I don't know if he made them or not – but Thanks to whomever did!

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

Some unobserved state variable like: live root carbon, dead root carbon, canopy water …

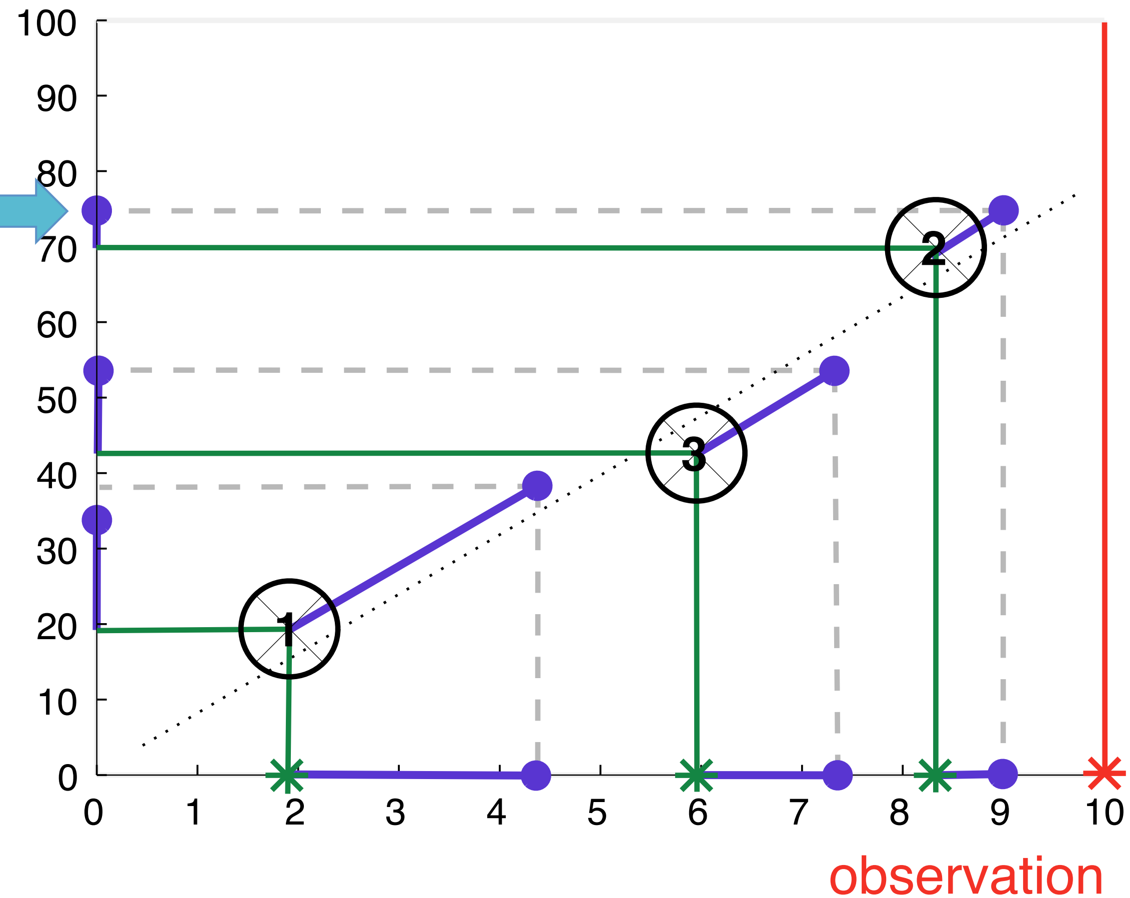


Could be Soil Temperature

REPEATED FOR REFERENCE

Potential Problem

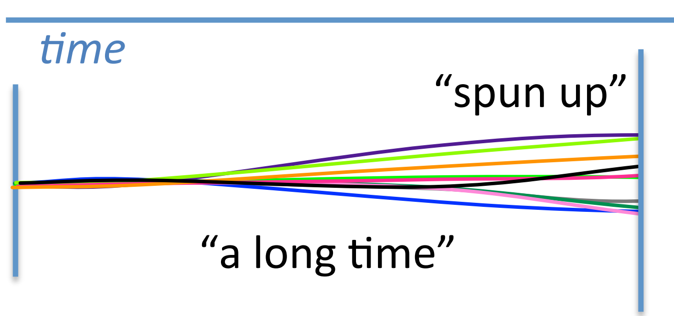
This posterior MAY or MAY NOT be realistic!



Can the model tolerate this new state?

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

Creating the initial ensemble of ...



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".

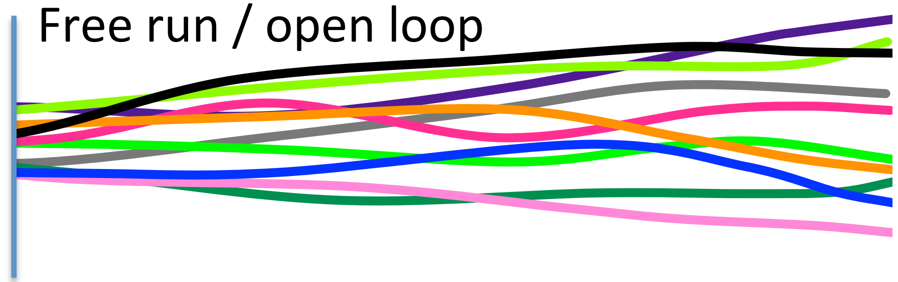
DART has tools we are using to explore how much spread we NEED to capture the uncertainty in the system.

The ensemble advantage.

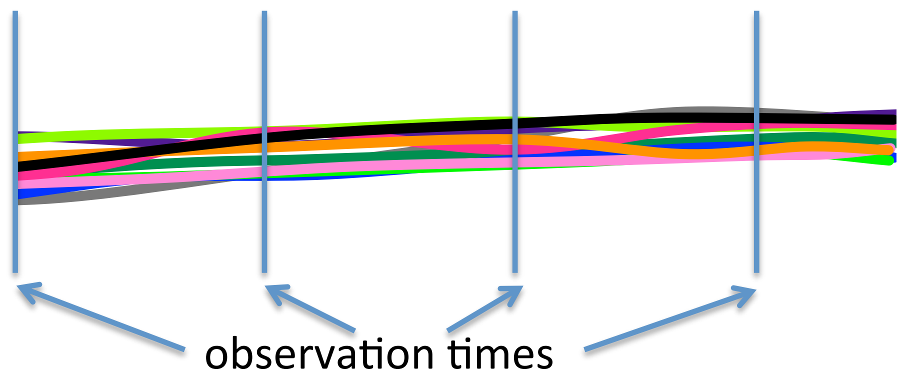
You can represent uncertainty.

time

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

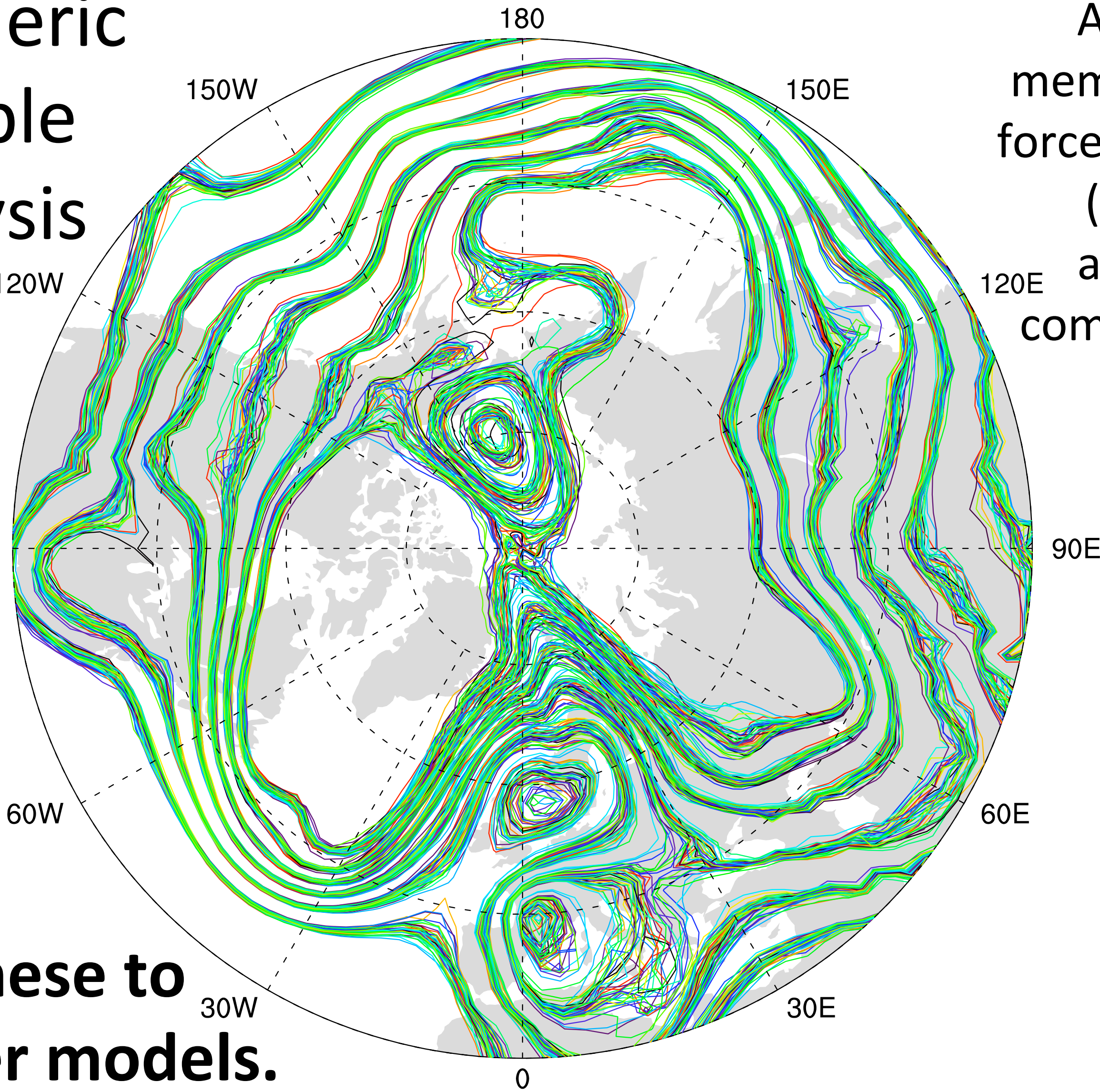


A good assimilation reduces the ensemble spread and is still representative and informative.



Atmospheric Ensemble Reanalysis

500 hPa GPH
Feb 17 2003



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

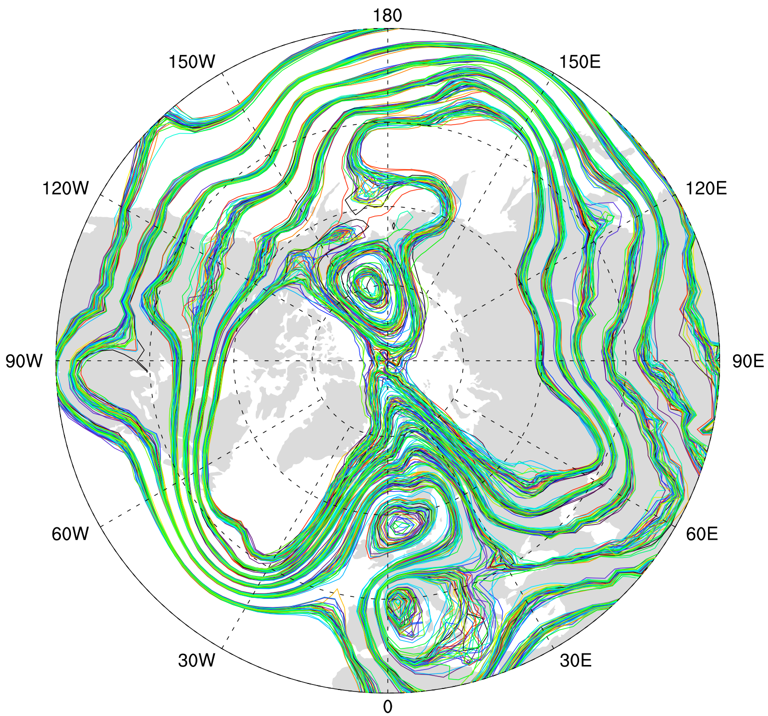
O(1 million) atmospheric obs are assimilated every day.

1998-2010+ 4x daily is available.

Can use these to force other models.

Data Assimilation Research Testbed

Pros and Cons

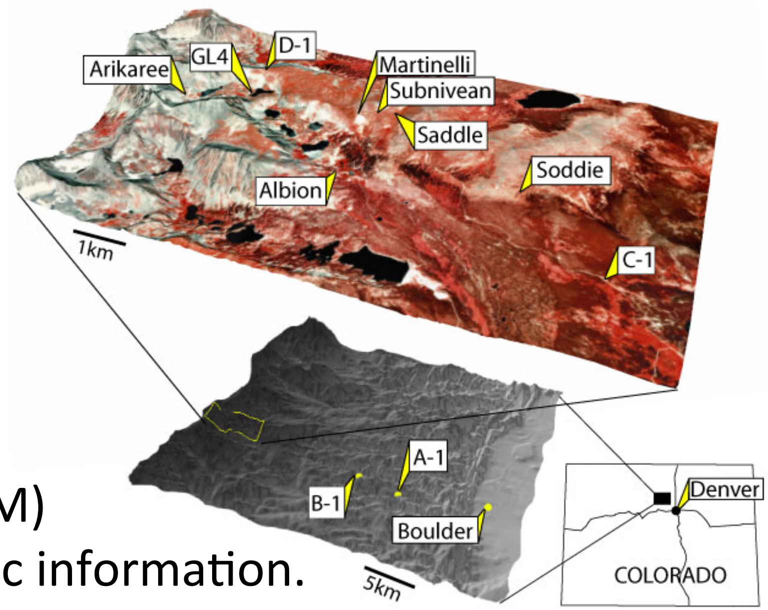


- **80 realizations/members**
- **Model states are self-consistent**
- **Model states consistent with obs**
- **Available every 6 hours for 12+ years**

- 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 data assimilation.

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*.

Assimilation of the MODIS Snow Cover Fraction Dataset through the Coupled Data Assimilation Research Testbed (DART) and the Community Land Model (CLM4)

Yongfei Zhang, Zong-Liang Yang
The University of Texas at Austin

Tim Hoar, Jeffrey Anderson
The National Center for Atmospheric Research

Ally Toure, Matthew Rodell
The National Aeronautics and Space Administration

The HARD part is: *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?

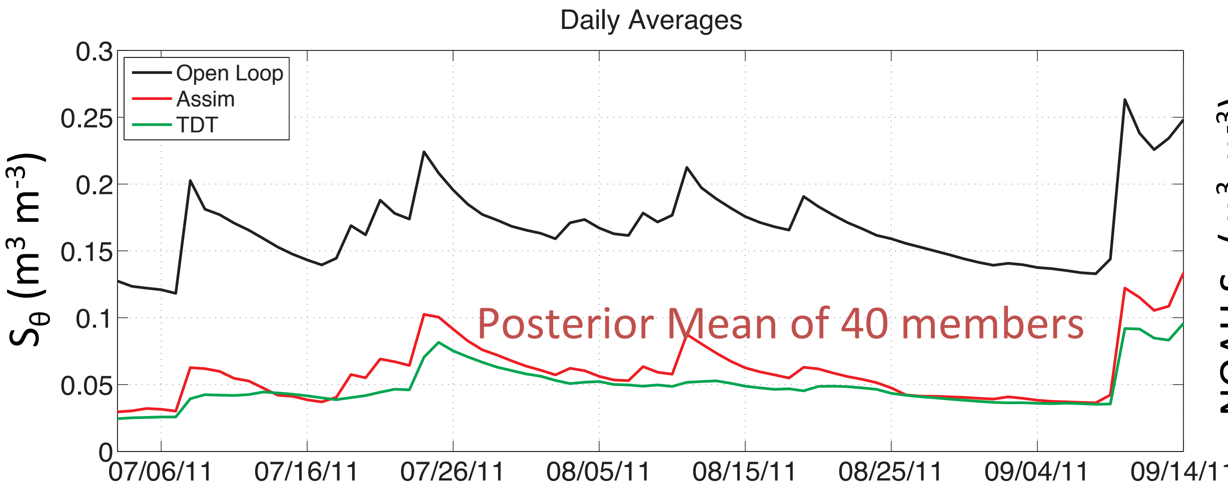
Early Season Snow?

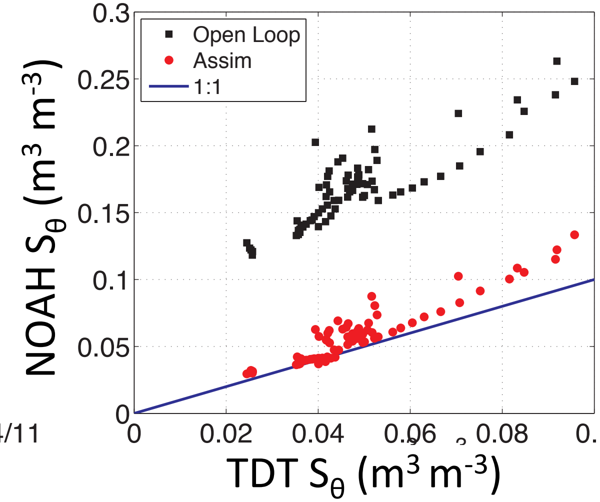
Snow Density?

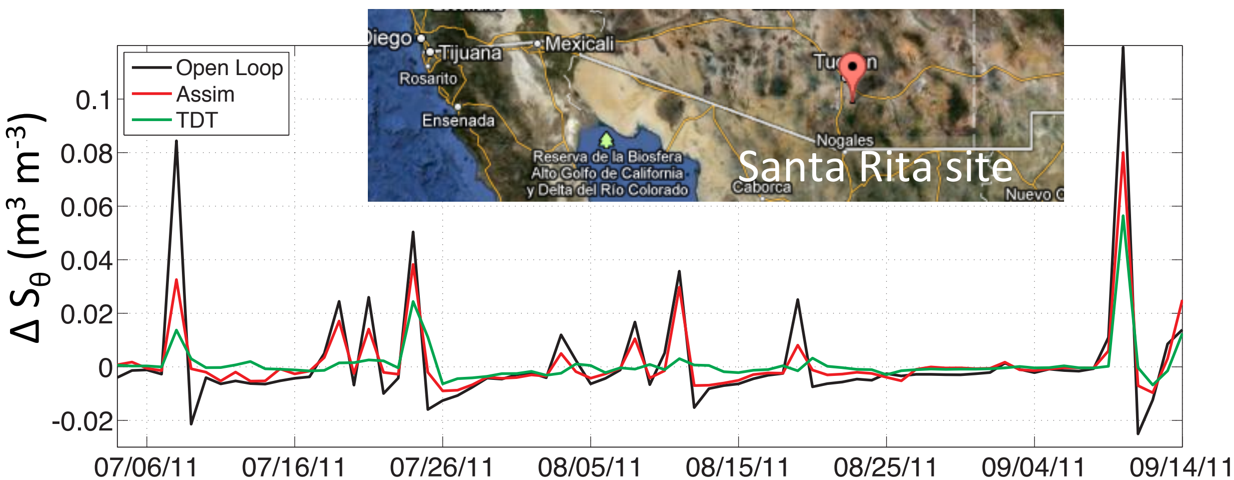
Snow Albedo?

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!

NOAH-DART: Integrated Soil Moisture







Data Assimilation Research Testbed