ENSEMBLE-BASED DATA ASSIMILATION FOR THE COMMUNITY LAND MODEL

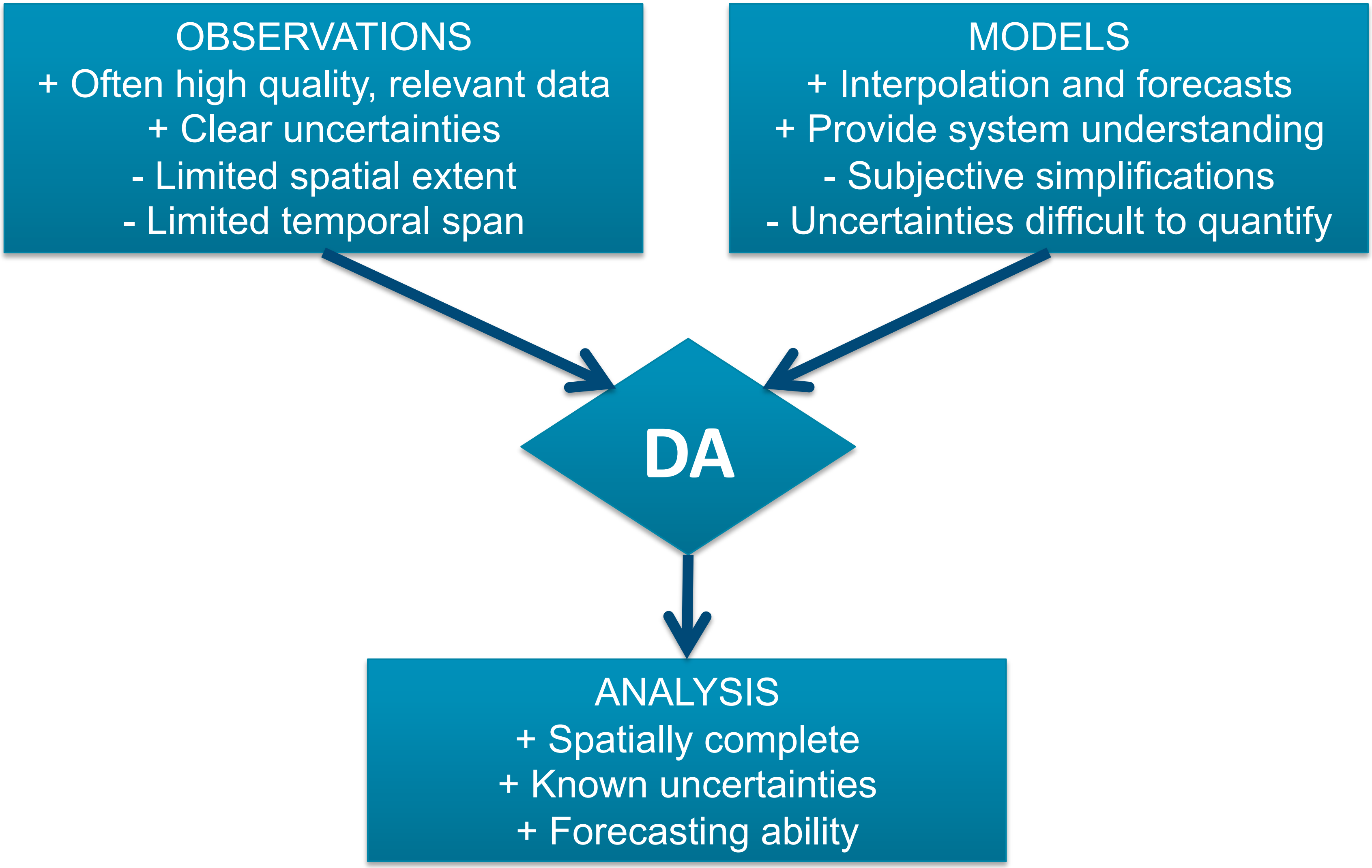
A. M. Fox[1], T. Hoar[2], Y. Zhang[3], D. Schimel[1], W. J. Sacks[2], T. Craig[2], J. Anderson[2]

1. National Ecological Observatory Network; 2. NCAR; 3. University of Texas, Austin

What is Data Assimilation?

- Systematic combination of data and models
- Taking into account the uncertainties in both
- Process model provides an analytical framework for data interpretation, synthesis, extrapolation
- If done well:
  - Modeled state becomes more consistent with observations
  - Make forecasts more accurate
  - Make reanalysis better represent the state of the system

DA for improving BGC models



OBSERVATIONS
+ Often high quality, relevant data
+ Clear uncertainties
- Limited spatial extent
- Limited temporal span

MODELS
+ Interpolation and forecasts
+ Provide system understanding
- Subjective simplifications
- Uncertainties difficult to quantify

DA

ANALYSIS
+ Spatially complete
+ Known uncertainties
+ Forecasting ability

Is DA different for NWP and BGC models?

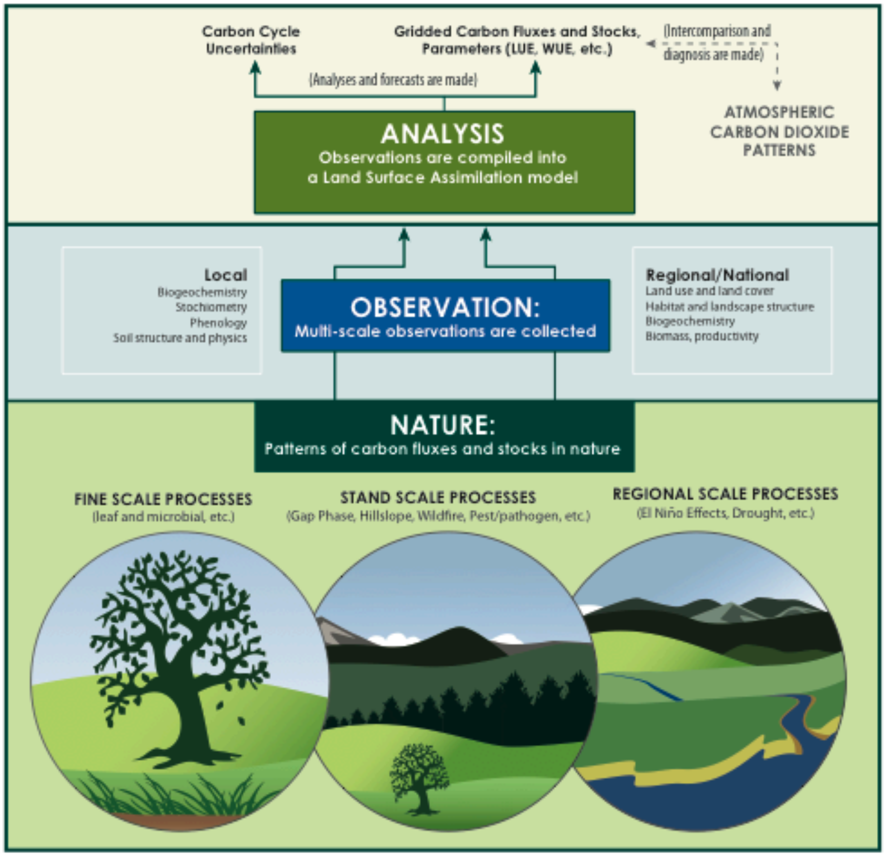
	Data Assimilation in NWP	Data Assimilation in CLM
Main objective	Forecast improvement	Process understanding Regional quantification Forecasting
Dynamics	Physics – essentially well known from first principles	Physical, biological, chemical – Only partially known, empirical relationships
Main drivers, boundary conditions	Well known	Mostly identified Poorly quantified
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

DA is able to improve CLM

- Providing estimates of state variables, initial conditions, and parameters
- Quantifying uncertainties with respect to modeled states of an ecosystem, initial conditions and parameters
- Helping to select between alternative model structures
- Providing a quantitative basis to evaluate sampling strategies for future experiments and observations that will enable improvements to BOTH models and forecasts

How is DA useful for NEON?

- The goal of NEON is to enable understanding and forecasting of the impacts of climate change, land use change and invasive species on continental-scale ecology



Carbon Cycle Uncertainties

Gridded Carbon Fluxes and Stocks, Parameters (LUE, WUE, etc.)

(Intercomparison and diagnosis are made)

(Analyses and forecasts are made)

ATMOSPHERIC CARBON DIOXIDE PATTERNS

ANALYSIS

Observations are compiled into a Land Surface Assimilation model

Local

Biogeochemistry
Stochiometry
Phenology
Soil structure and physics

OBSERVATION:

Multi-scale observations are collected

Regional/National

Land use and land cover
Habitat and landscape structure
Biogeochemistry
Biomass, productivity

NATURE:

Patterns of carbon fluxes and stocks in nature

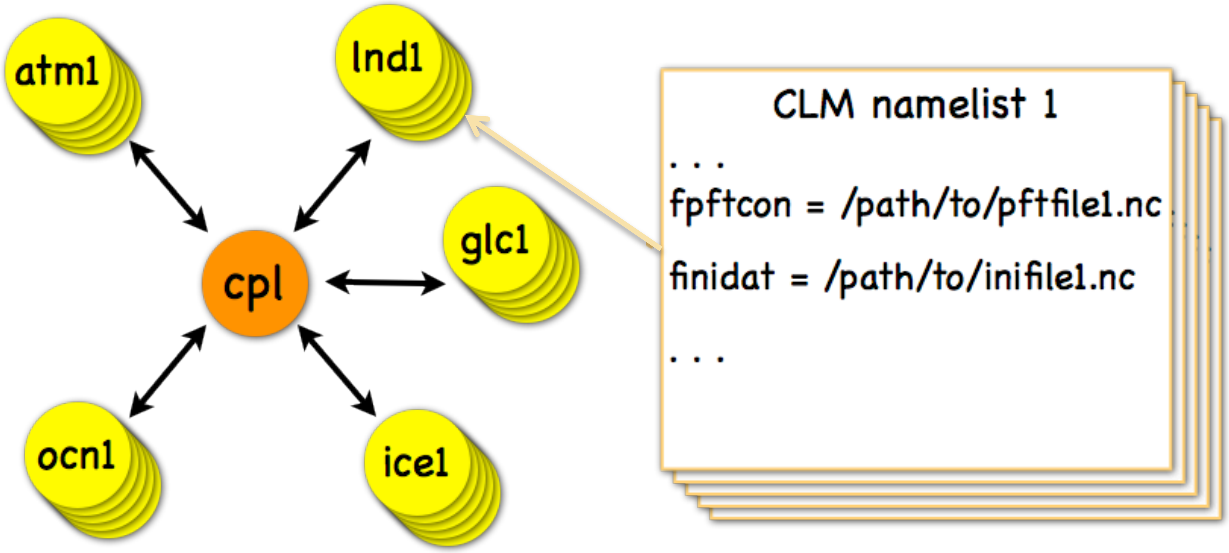
FINE SCALE PROCESSES (leaf and microbial, etc.)

STAND SCALE PROCESSES (Gap Phase, Hillslope, Wildfire, Pest/pathogen, etc.)

REGIONAL SCALE PROCESSES (El Niño Effects, Drought, etc.)

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.

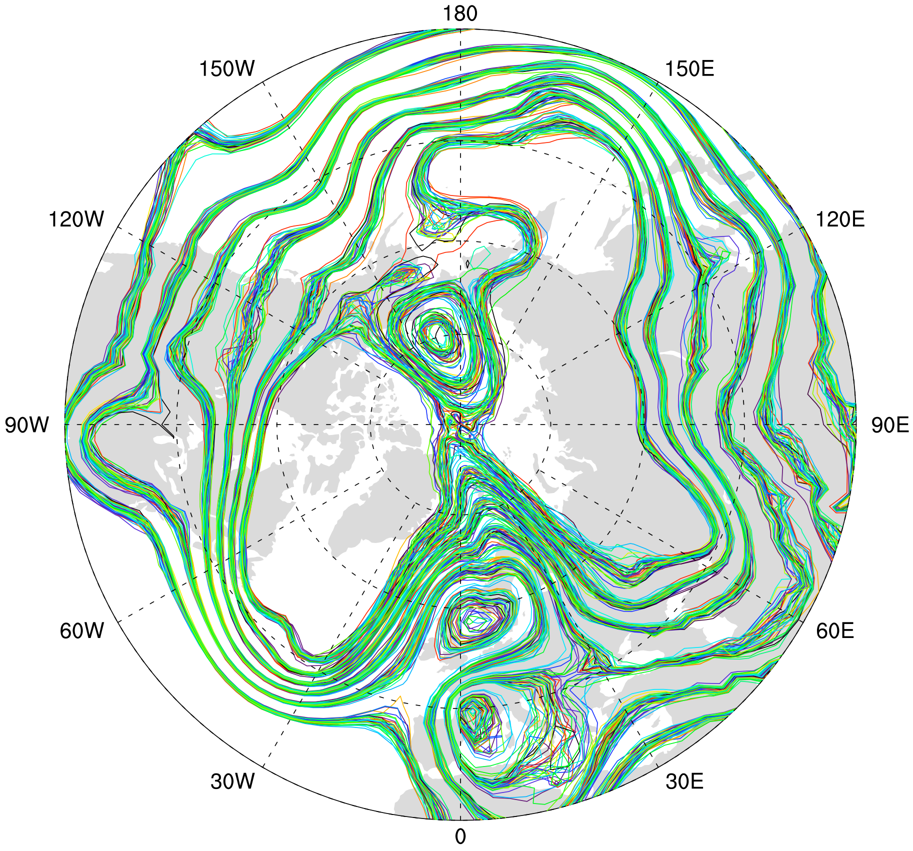


CLM namelist 1

```
. . .
fpftcon = /path/to/pftfile1.nc
finidat = /path/to/inifile1.nc
. . .
```

Multi-instances of data atmospheres

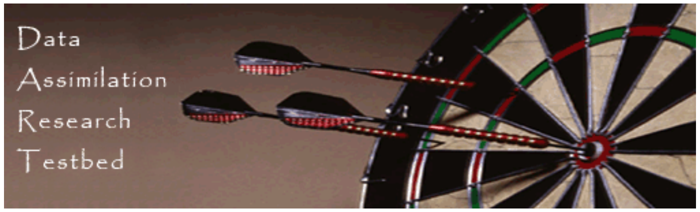
- Assimilation uses 80 members of 2^o FV CAM forced by a single ocean
- O (1 million) atmospheric obs are assimilated every day
- 1998 – 2010, 6 hrly reanalysis available
- Each CLM ensemble member is forced with a different atmospheric reanalysis member
- Generates spread in the land model

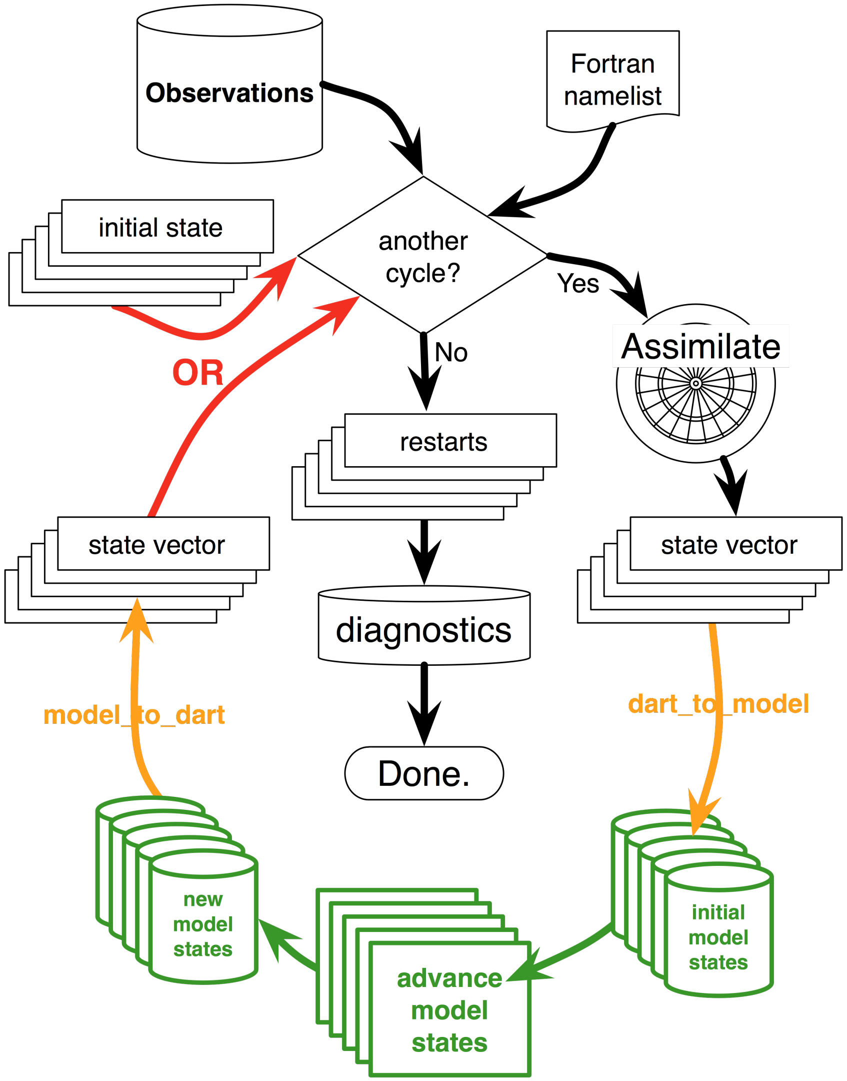


500 hPa GPH
Feb 17 2003

Data Assimilation Research Testbed (DART)

- DART is a community facility for ensemble DA
- Uses a variety of flavors of filters
  - Ensemble Adjustment Kalman Filter
- Many enhancements to basic filtering algorithms
  - Adaptive inflation
  - Localization
- Extensive documentation, tutorials and diagnostic tools





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

```
&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 variables are:

  i. read from a CLM restart file

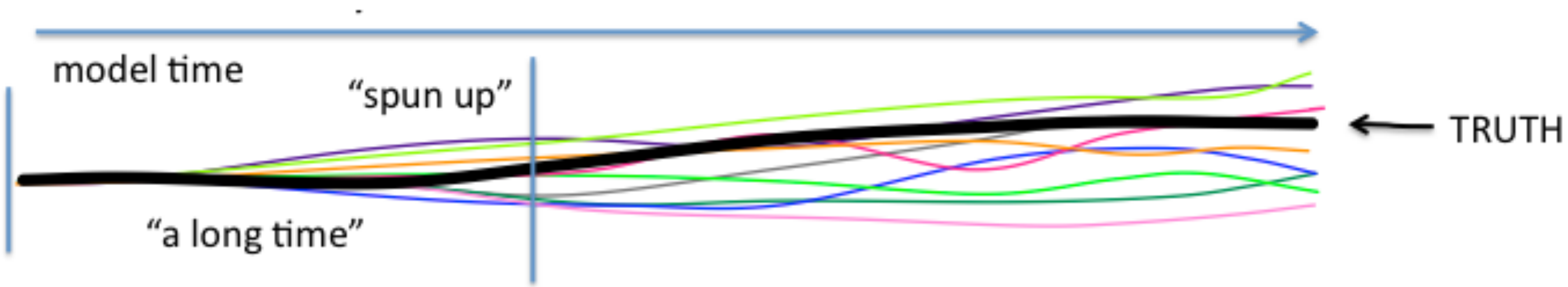
  ii. Converted into a .ics file for `filter`

  iii. Increments calculated by DART (EAKF)

  iv. Updated values are converted back and inserted into restart file

Perfect model experiment design 1.

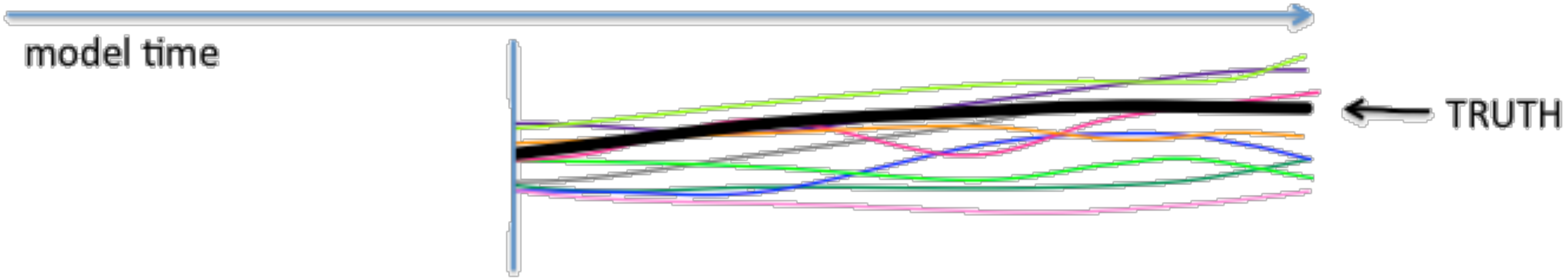
- Testing the success of the DART-CLM implementation with perfect model experiments



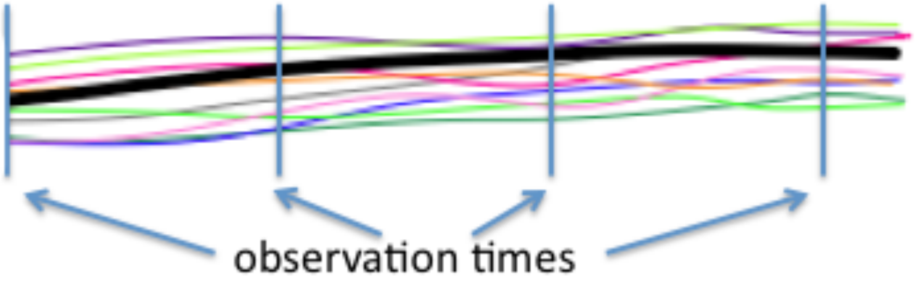
- Each line represents the evolution of individual instances of CLM
- Pick one and declare if the truth
- Harvest synthetic "observations" from this truth at predefined time intervals and locations, adding a prescribed noise/ uncertainty

Perfect model experiment design 2.

- Without assimilation:
  - frequently ensemble spread with grow



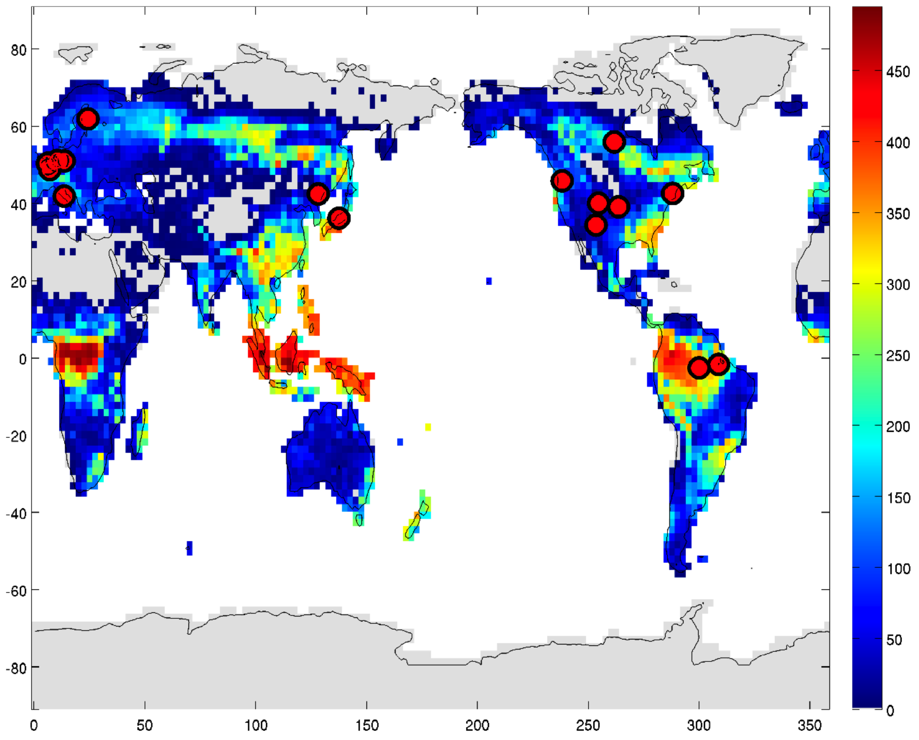
- With assimilation:
  - Ensemble spread remains stable, is small enough to be informative, but does not collapse away from truth



Perfect model experiment methodology

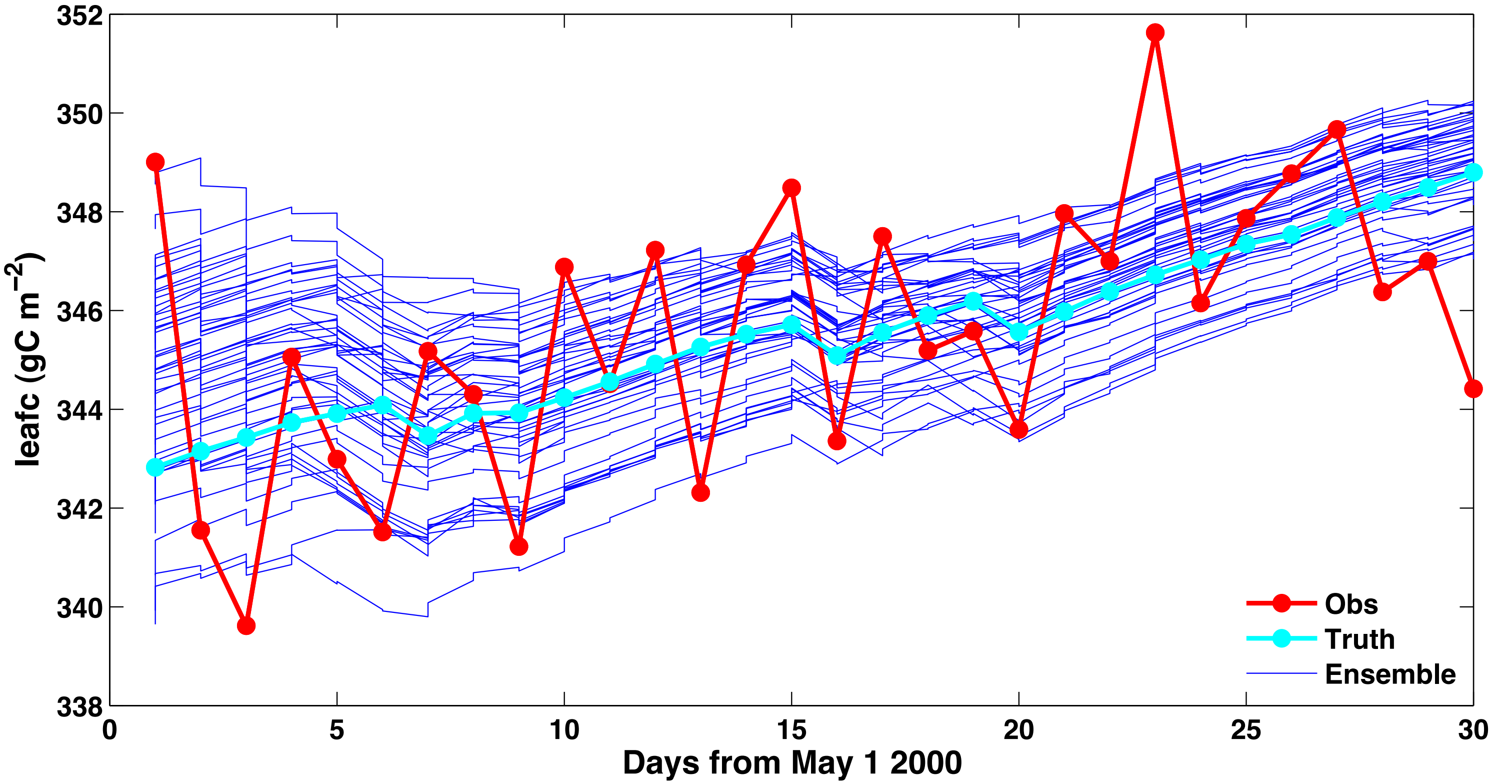
- Use 40 DATMS (CPLHIST) with 40 instances of CLM
- 2^{o} ICN compset, global run
- Start from 1 Jan 2000 spun up state
- Run for 4 months to generate spread
- Run 1 ensemble member forward from 1 May 2000, harvesting daily observations of `leafc` at 16 locations
- Run 40 ensemble members forward from 1 May 2000 for 30 days, assimilating synthetic observations

Global `leafc`, 1 May 2000



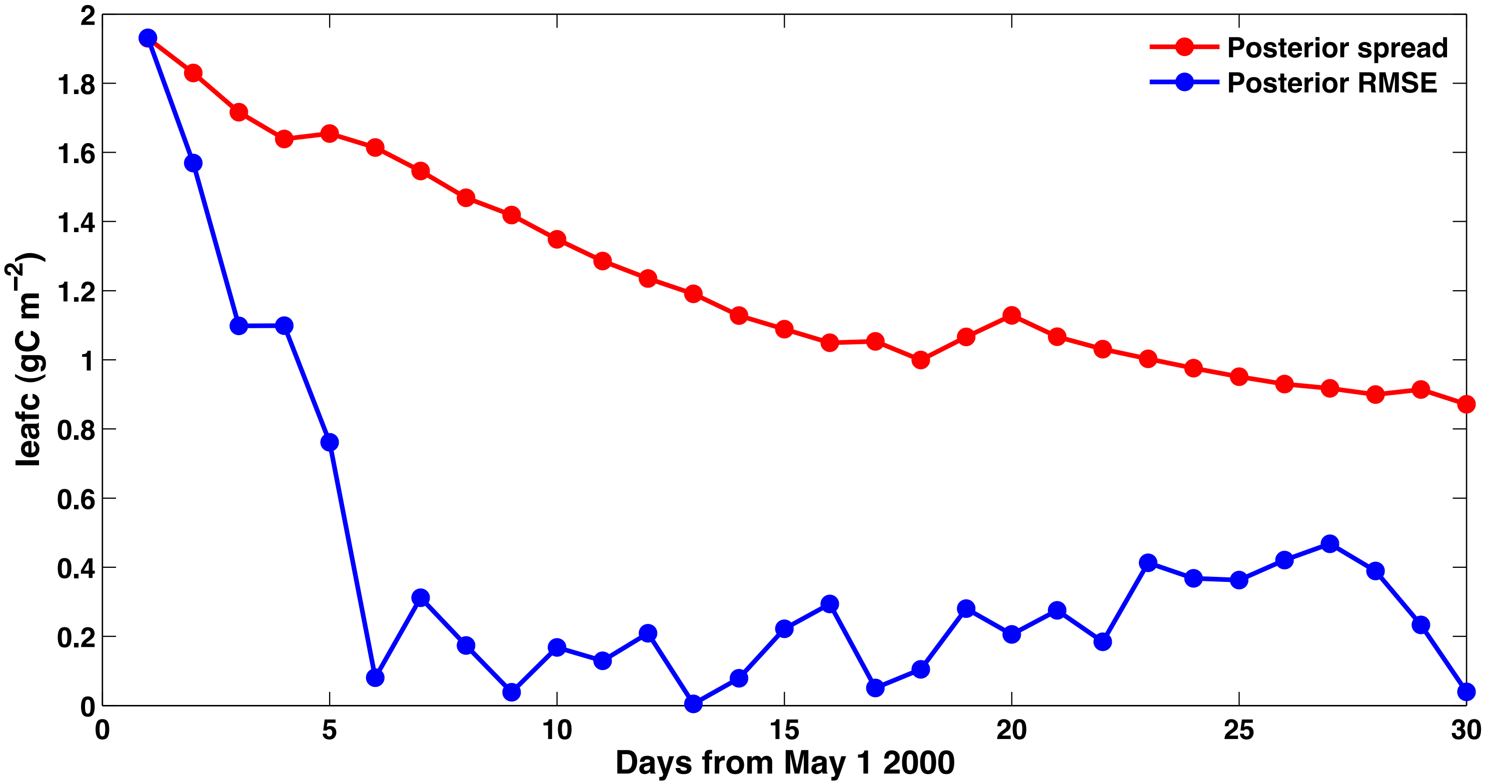
Time series of "truth", obs and 40 ens members

- 40 member ensemble of `leafc` in a single grid cell corresponding to 60.21°W, 2.61°S (Manaus, Brazil)
- "Sawtooth" pattern describes increments



Time series of RMS error of ens mean and spread

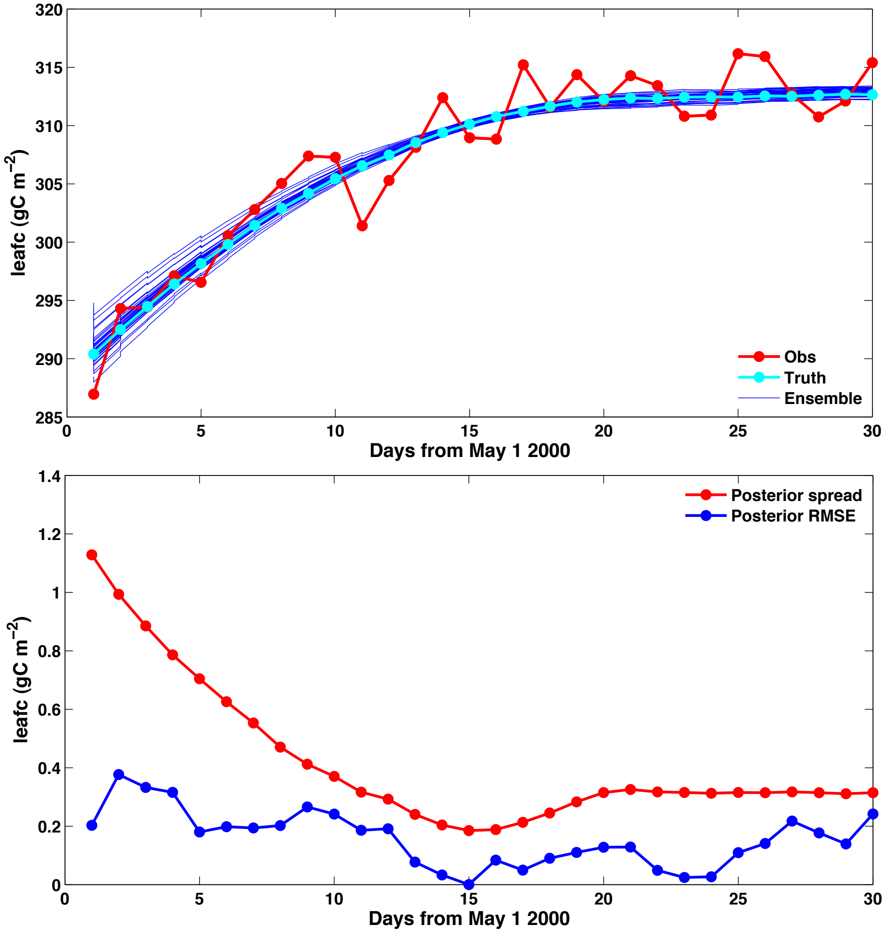
- Assimilation is consistent with truth
- Reduction in RMSE between ensemble mean and truth
- Reduction in ensemble spread, but not too much



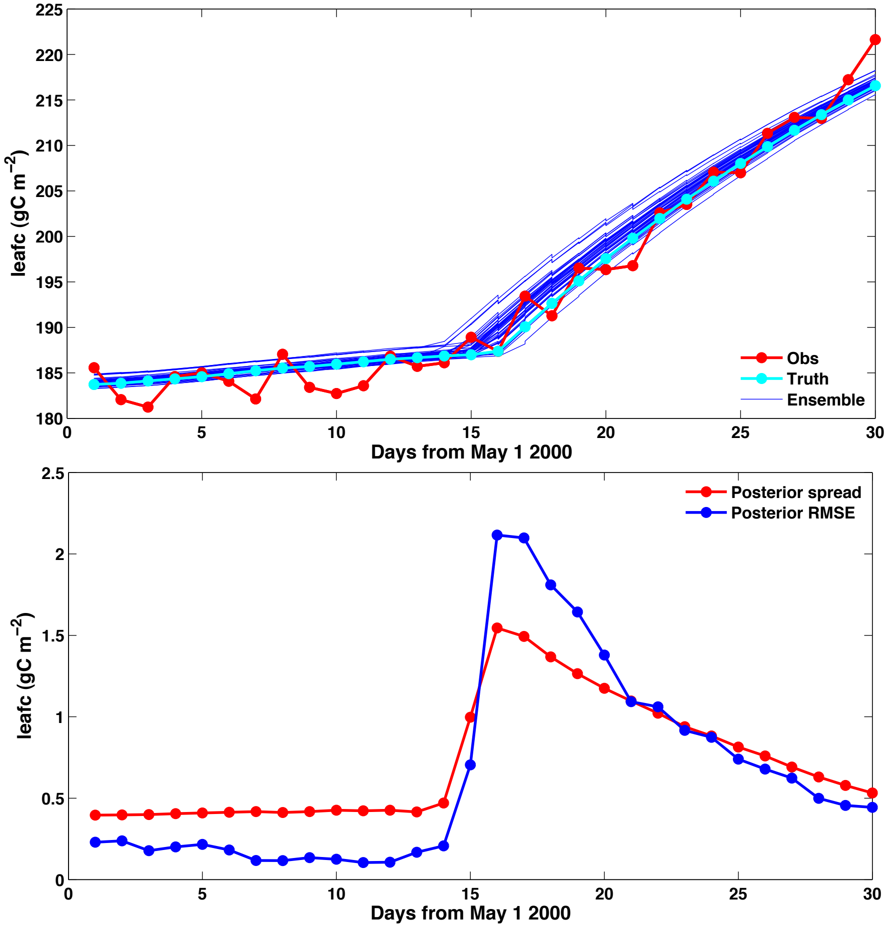
Some other examples

- See variation in confidence about the state

137.42°W, 36.15°N (Takayama, Japan)

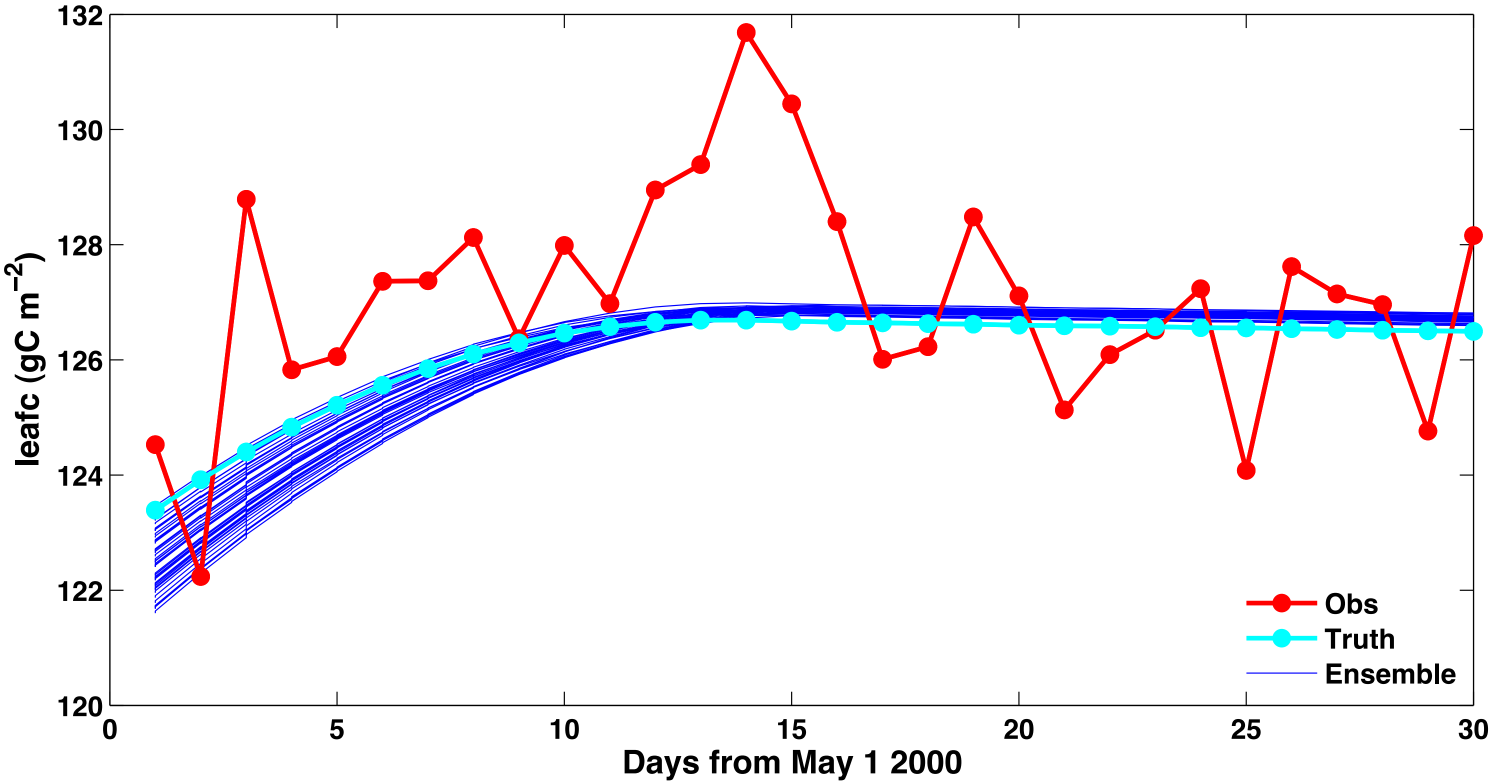


72.17°W, 42.54°N (Harvard Forest, USA)



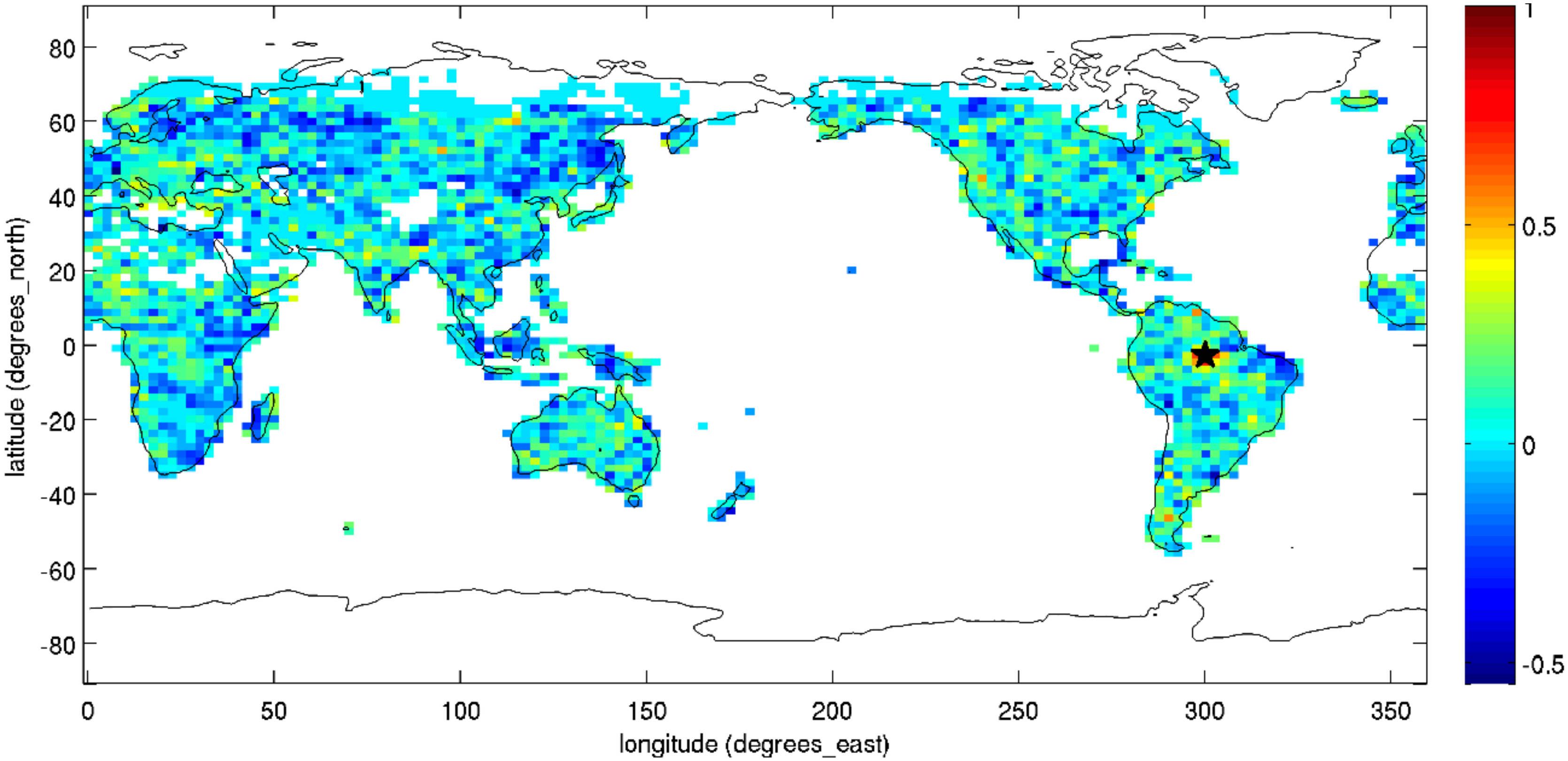
But they don’t always look so good

- Time series of truth, observations and ensemble members for grid cell at 7.07°E, 48.67°N (Hesse, France)
- Collapse of ensemble, overly confident, and wrong



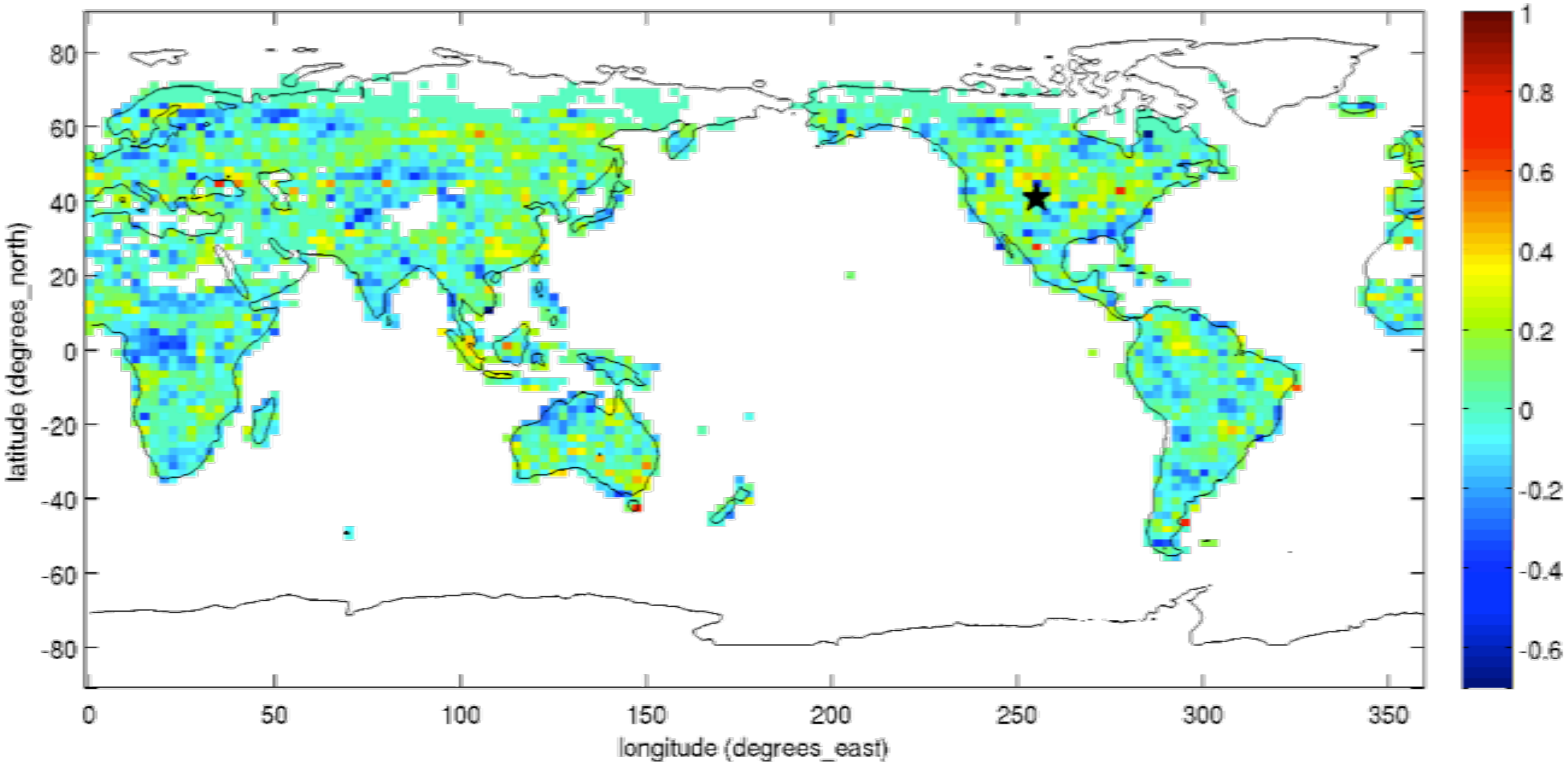
Patterns of spatial correlation

- We also have spatial correlations in ensemble members
- 4 May 2000 between 60.21°W, 2.61°S (Manaus, Brazil) and everywhere else using `leafc` state variable



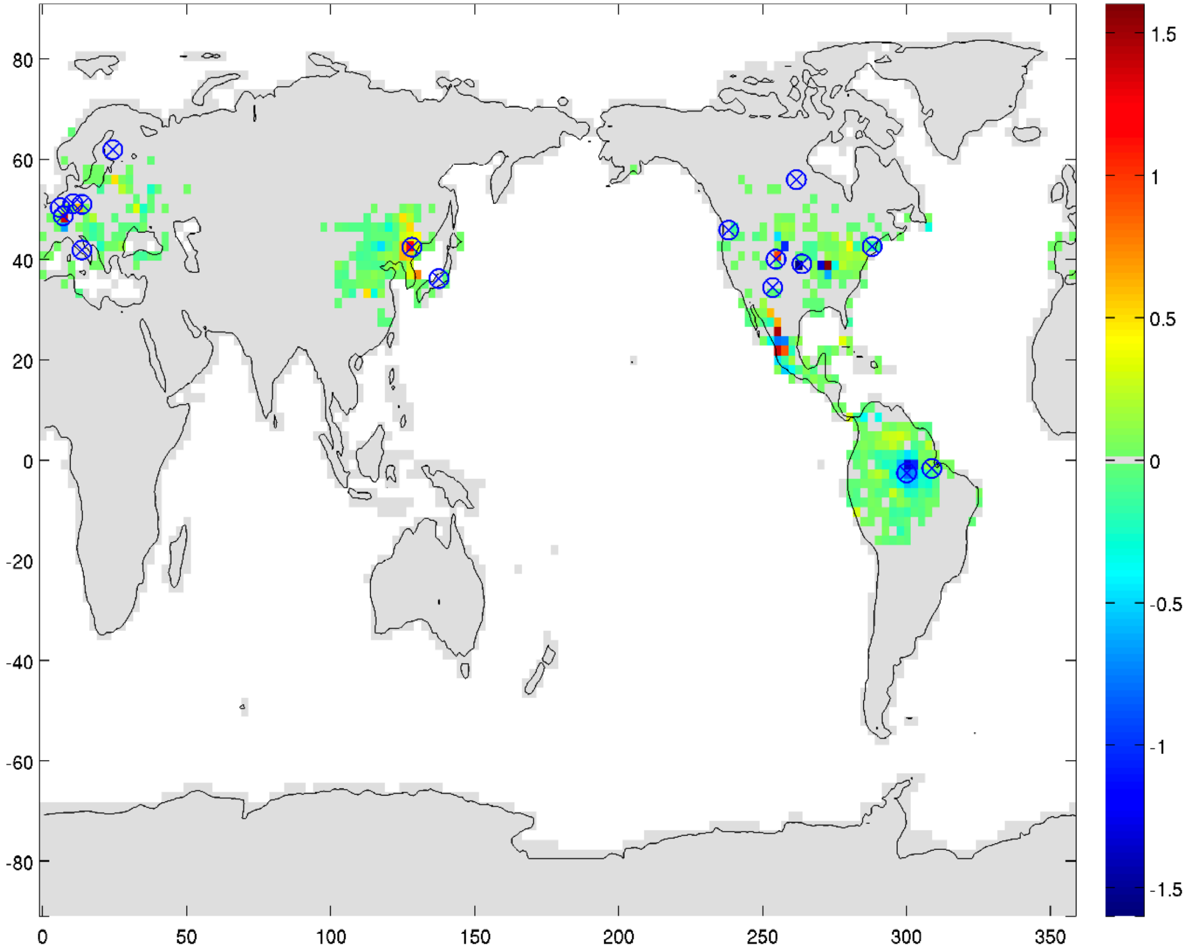
See some unexpected high correlations

- Distant model grid cells maybe correlated with observation site
- 4 May 2000 between 105.55°W, 40.03°N (Niwot Ridge, USA) and everywhere else with `leafc`



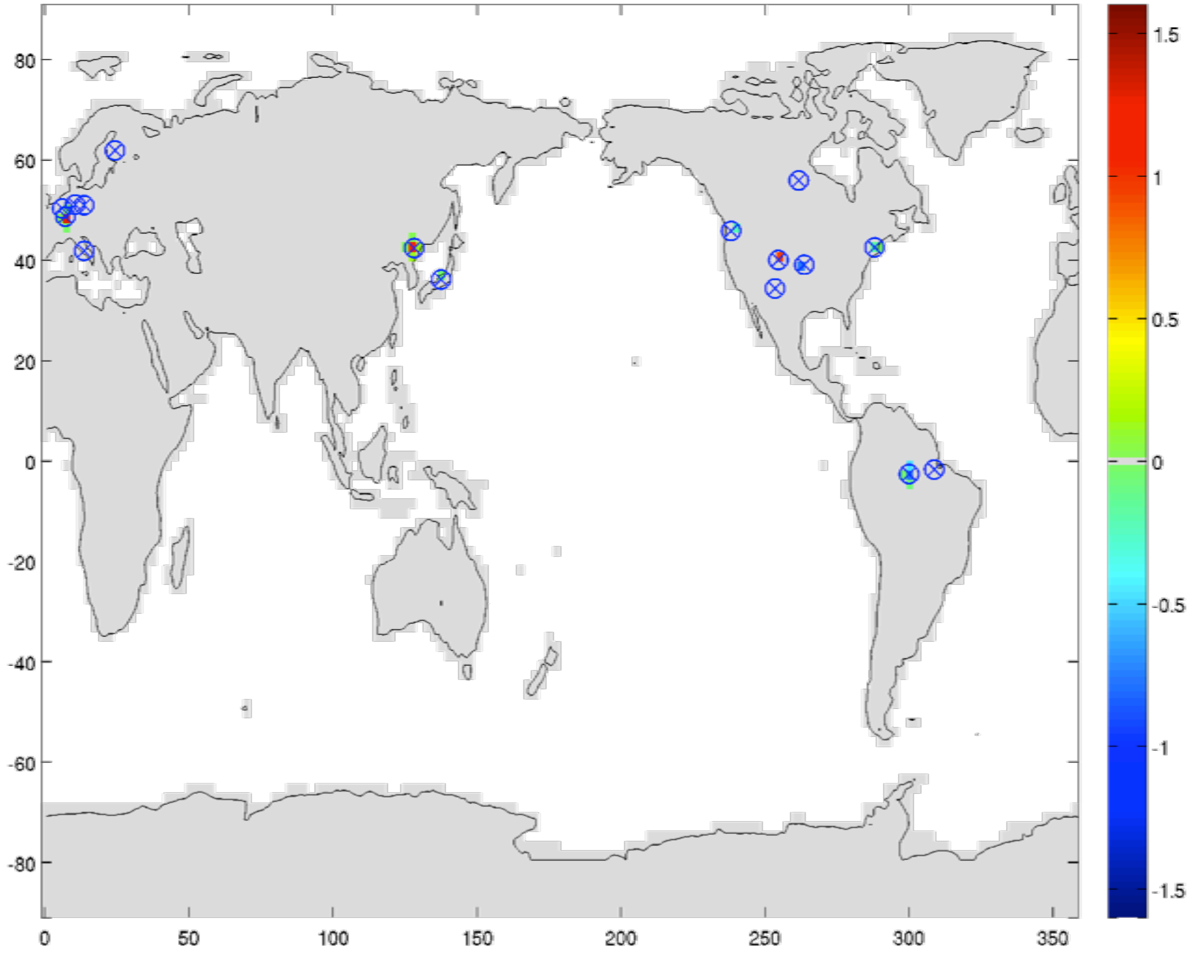
Innovation map of `leafc` on 4 May 2000

- Large areas of the globe are being affect by observations
- **Cutoff at 0.3 radians** (~2000km)



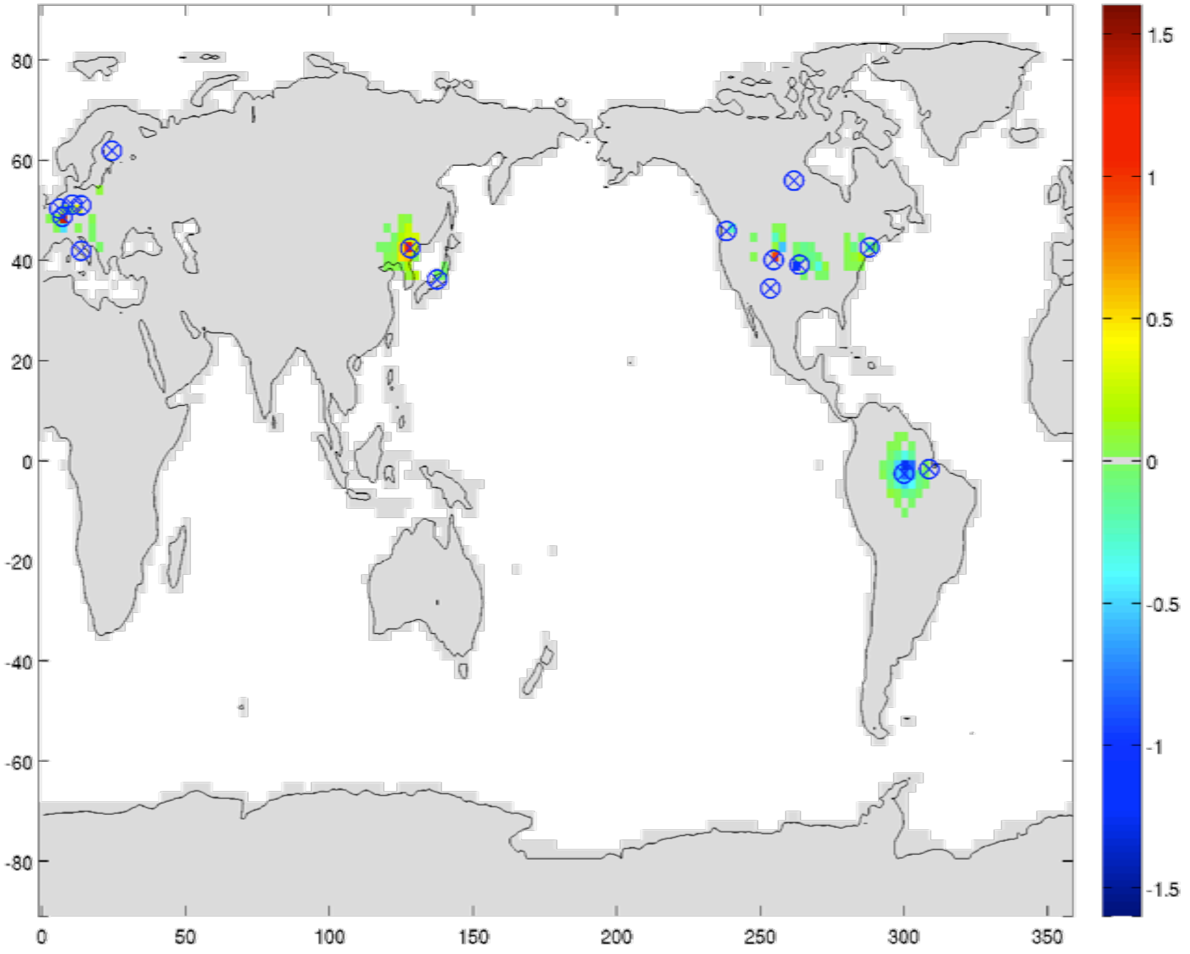
Localization limits increments to adjacent cells

- Innovation map of `leafc` on 4 May 2000
- **Cutoff 0.03 radians** (~200km)



A cutoff value in between

- Innovation map of `leafc` on 4 May 2000
- **Cutoff 0.1 radians** (~200km)



Many big questions remain

- How to create initial ensemble spread – how large should it be?
- How to maintain ensemble spread – is climate forcing variability the best approach?
- What do we do about carbon/water balance – its lost at the moment and balance checks are removed?
- Are there useful patterns of spatial correlation/covariance in CLM variables – does it make sense to use them in a model with no "physics" connection model grid cells?
- What are the most appropriate observations to use – and can we develop appropriate observation operators to link them with CLM state?
- How can we best use an ensemble DA approach for parameter estimation – we can augment DART state vector with CLM parameters, but which ones?

If you have any answers…

- Or if you'd like to get in involved in data assimilation with CLM using DART
- Or if there's a capability you particularly want
- Please don't hesitate to speak with myself or DART folks

The National Ecological Observatory Network is a project sponsored by the National Science Foundation and managed under cooperative agreement by NEON Inc.