The ensemble Kalman filter: The Movie

Douglas Nychka
Geophysical Statistics Project
National Center for Atmospheric Research

- The data/update cycle
- prior, likelihood, posterior
- ozone surface data
- linear regression for the update
- EKF: the movie

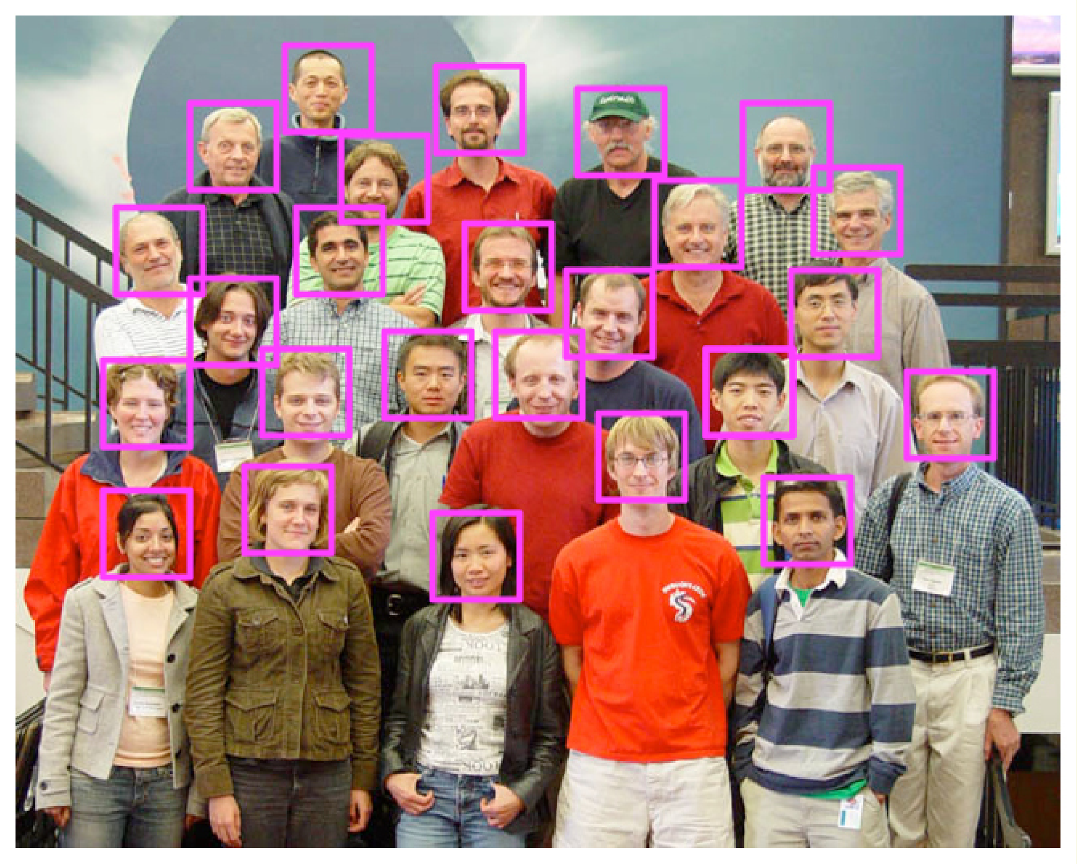
Supported by the National Science Foundation DMS

Why ensembles are a good idea.

An ensemble of states

Who wants the mean?

The ensemble mean=waste of time?



The main idea

An ensemble is a *sample* **useful for approximating the continuous distribution including covariances among variables.**

In fact it allows us to assimilate observations sequentially using a simple algorithm (*The Machine***) – even when the observations are all taken at the same time.**

The movie will be Groundhog Day ...

We will assimilate a vector of observations collected on the same day sequentially.

The basic data assimilation cycle

The problem is to estimate the state of a system g_t (or a field) at different times.

- Forecast for the state at time t
- New data, y comes in at time t
- Update g_t in light of the data
- Forecast ahead to time $t+1$ using updated state.
- *Cycle repeats.*

I am only going to talk about the update part

Cycle between the new data and update steps.

How a statistician describes the update

PRIOR: We have a probability distribution describing the uncertainty of the forecasted state.

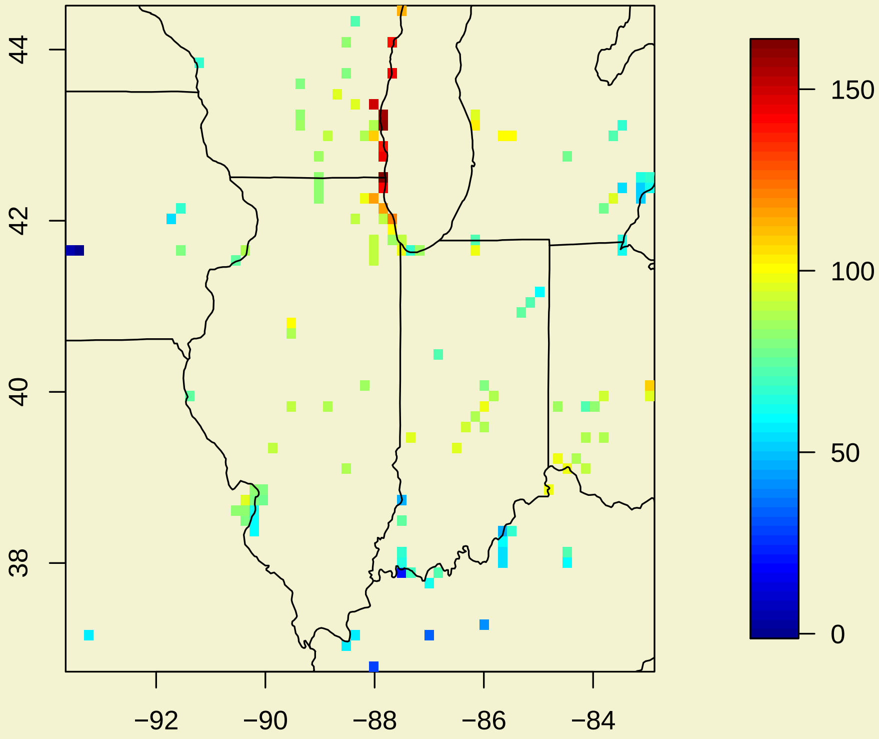
LIKELIHOOD: We know how the data is related to the true state of the system.

POSTERIOR: We want to know the probability distribution of the of the state *given* the data.

POSTERIOR is found using Bayes formula ....

Observed surface ozone, June 19, 1987

Goal: Estimate the surface



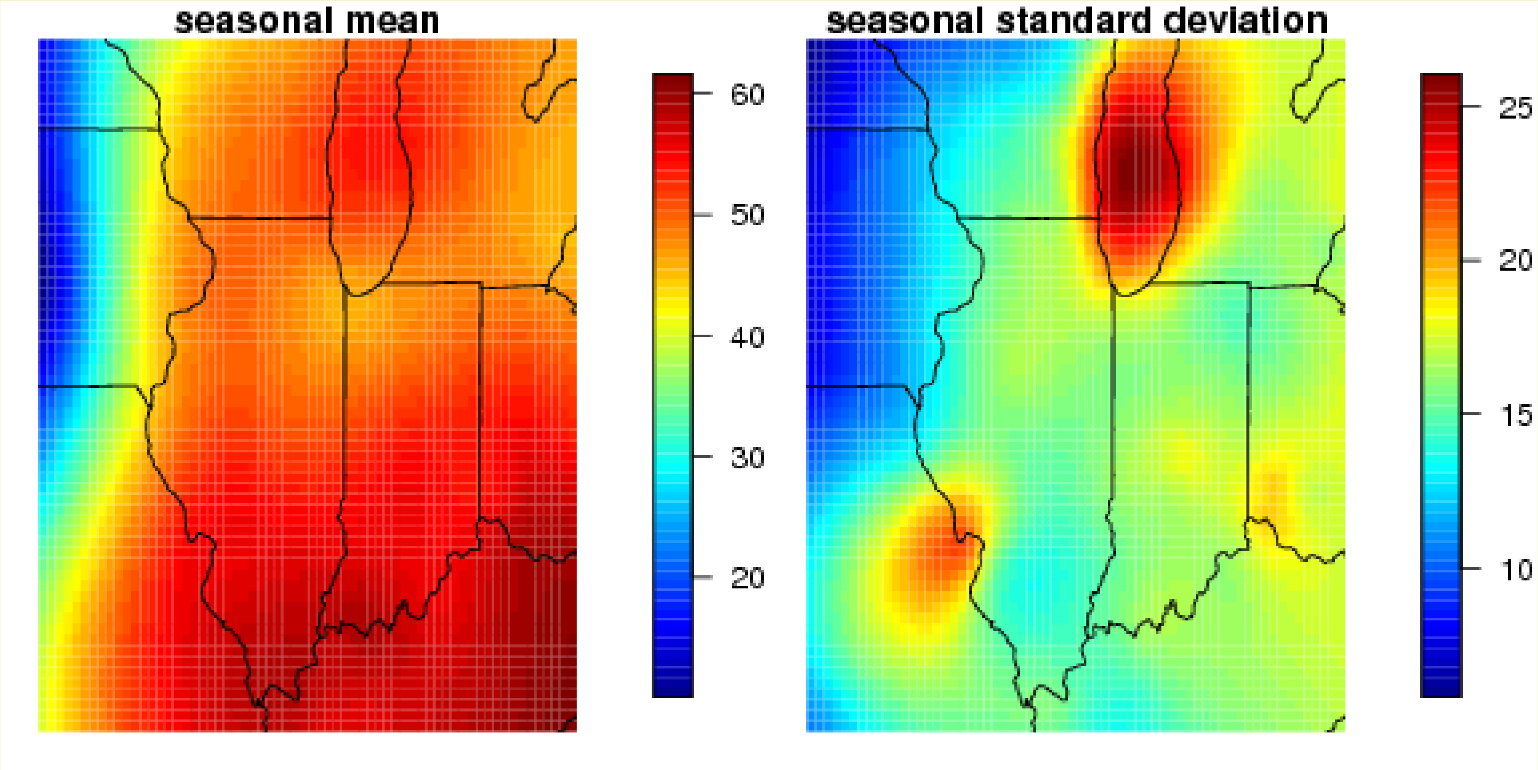
The statistical ingredients for the PRIOR

For 1987 summer ozone season we have a PRIOR for the ozone field over this region – essentially a summer ”climatology”.

Based on some data analysis, it is (roughly) multivariate normal with a mean around 60PPB a variance from 10 to 25 PPB and a correlation range of about 300 miles.

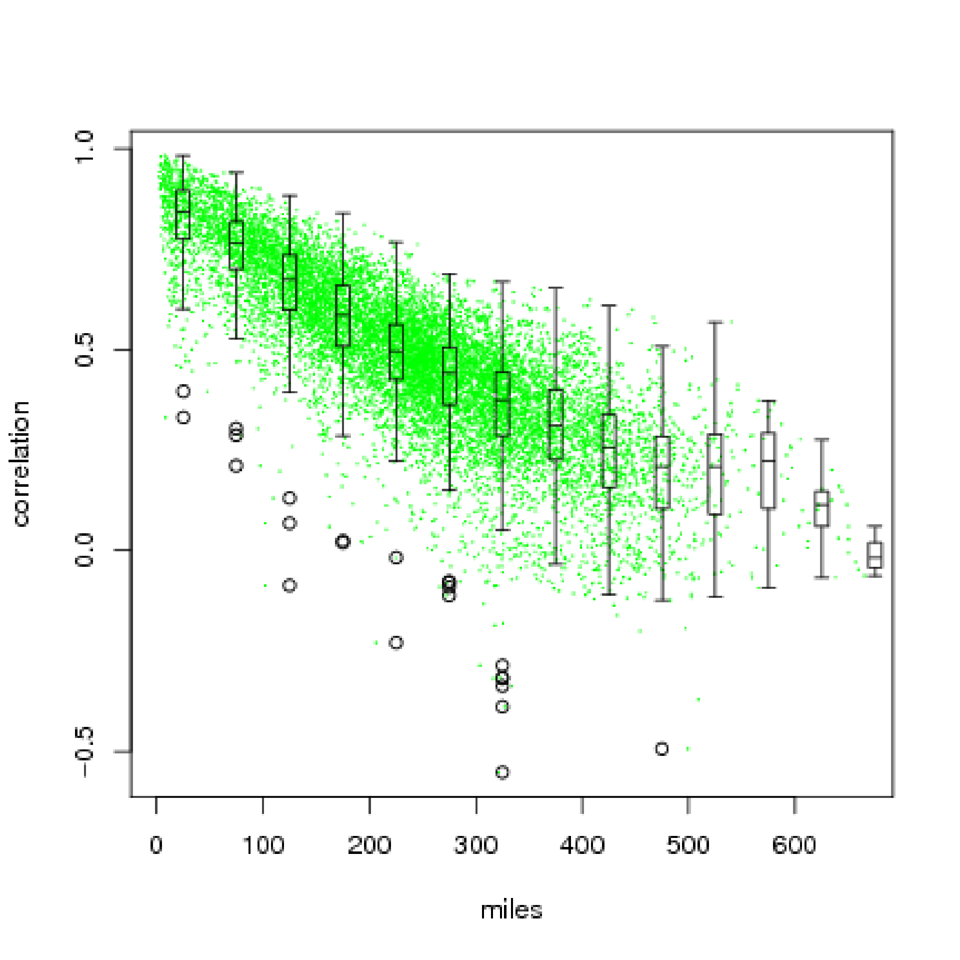
This is only our guess and will be modified as we assimilate observations.

Ozone prior from 79 days, summer 1987.



PRIOR Correlation scale

Sample correlations of ozone stations by distance of separation.



The Likelihood

Observations at irregular station locations

$$Y_k = g(\boldsymbol{x}_k) + e_k$$

Daily ozone measured with a small amount of error that is normally and independent among stations.

For this case Y is 133 observations.

Posterior

Some facts

- Posterior is multivariate normal with a mean and variance found by the Kalman Filter (or 3DVAR with the right background covariance).
- Because the observation errors are independent the observations can be assimilated sequentially and in any order.
- The ensemble filter will reproduce the Kalman Filter results as the ensemble size increases.

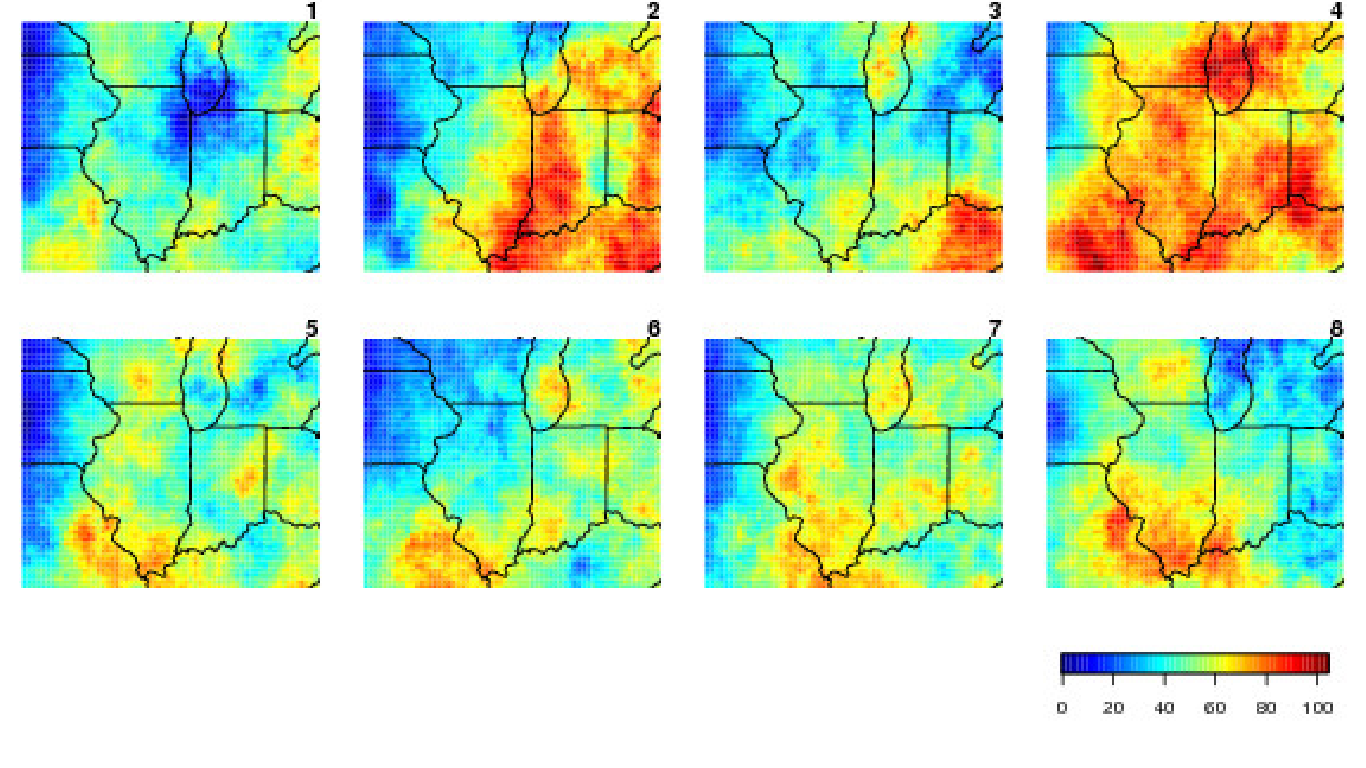
At this point in most coherent talks one writes down the Kalman filer equations ...

PINK FLOYD
WELCOME TO THE MACHINE

A different explanation: the Machine

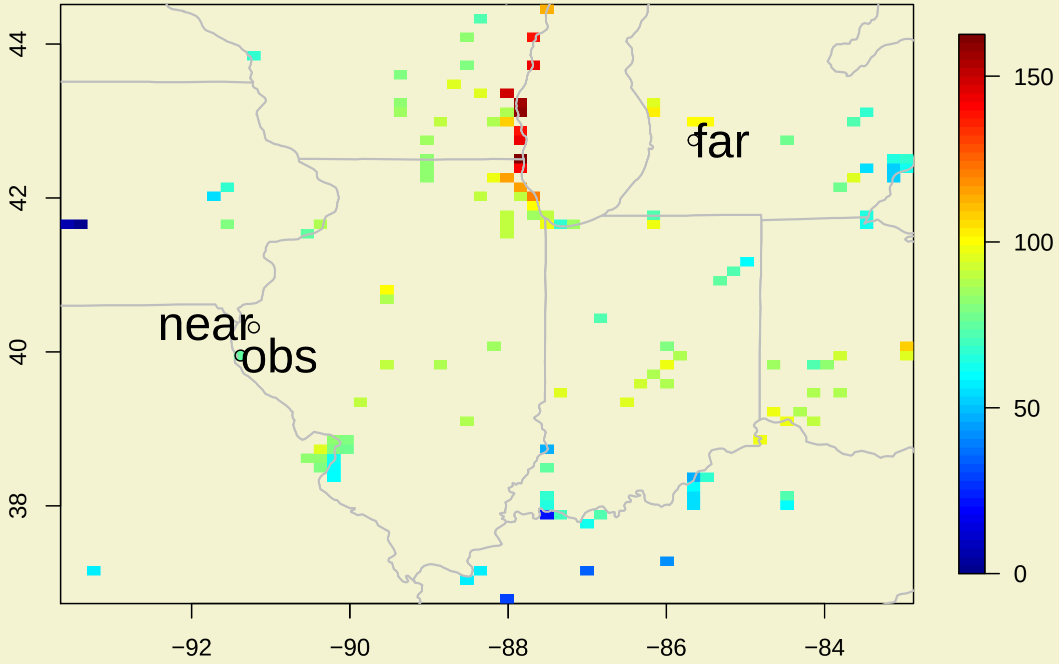
Generate a 100 member ensemble from the prior. These are random fields consistent with the ozone summer 1987 ”climatology”.

The first 8 members of initial ensemble fields



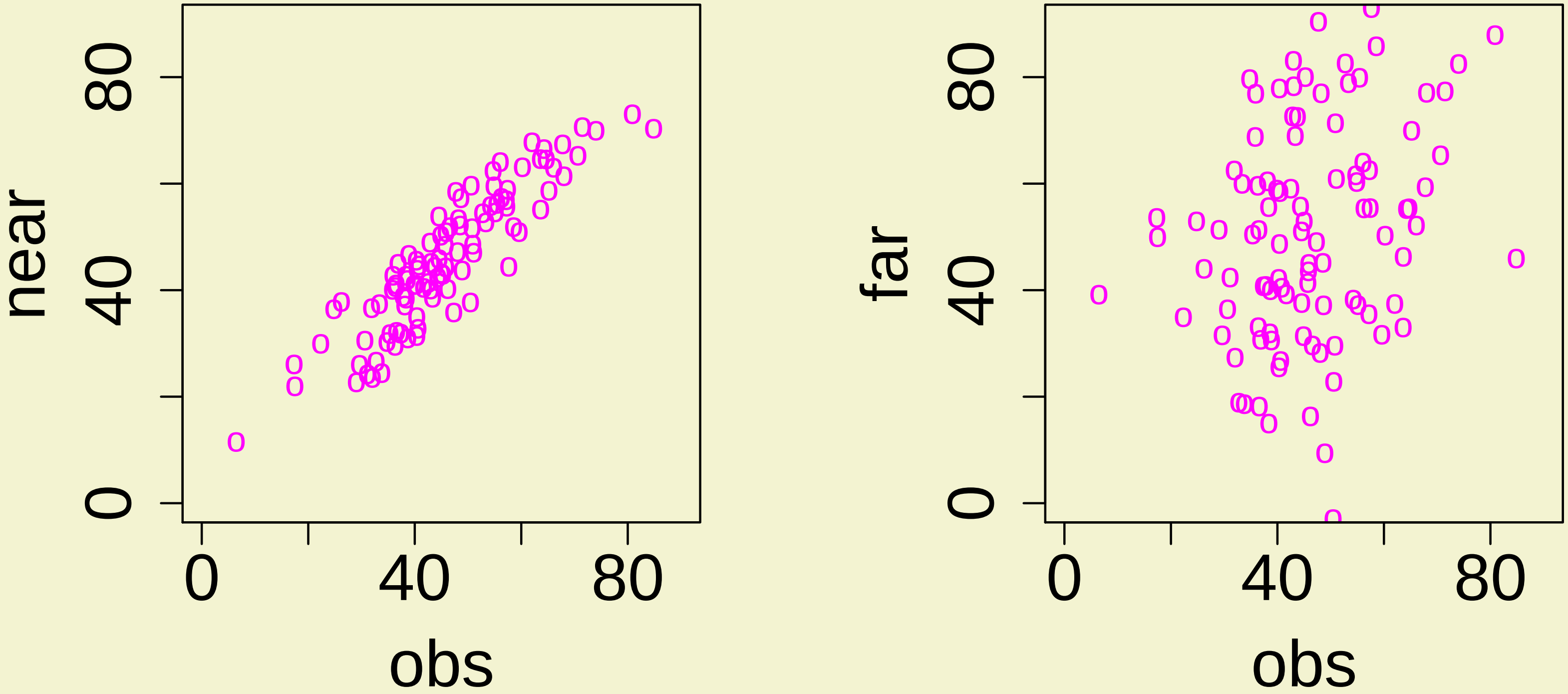
Updating the first observation

Observation has value 75 PPB

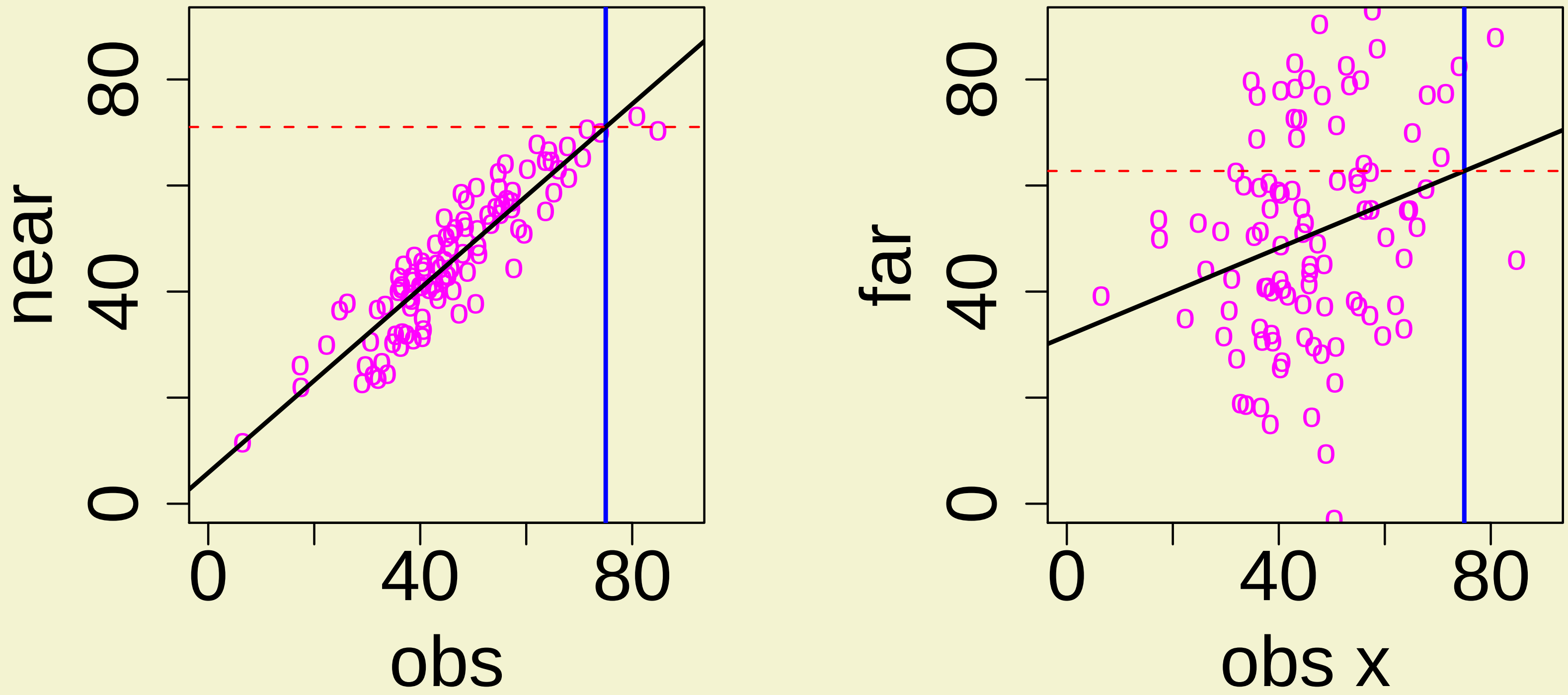


Consider updates at near and far points.

Scatterplots of the ensemble members

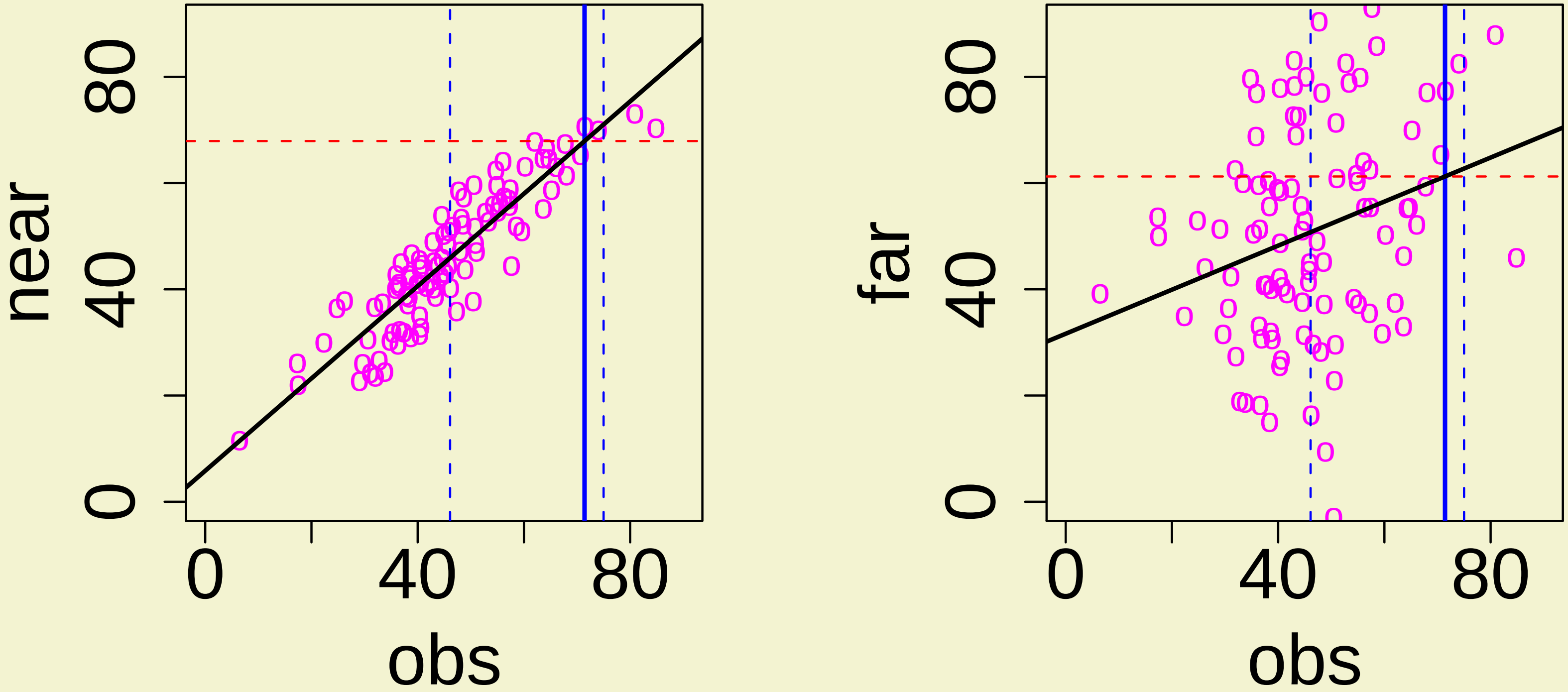


With no measurement error



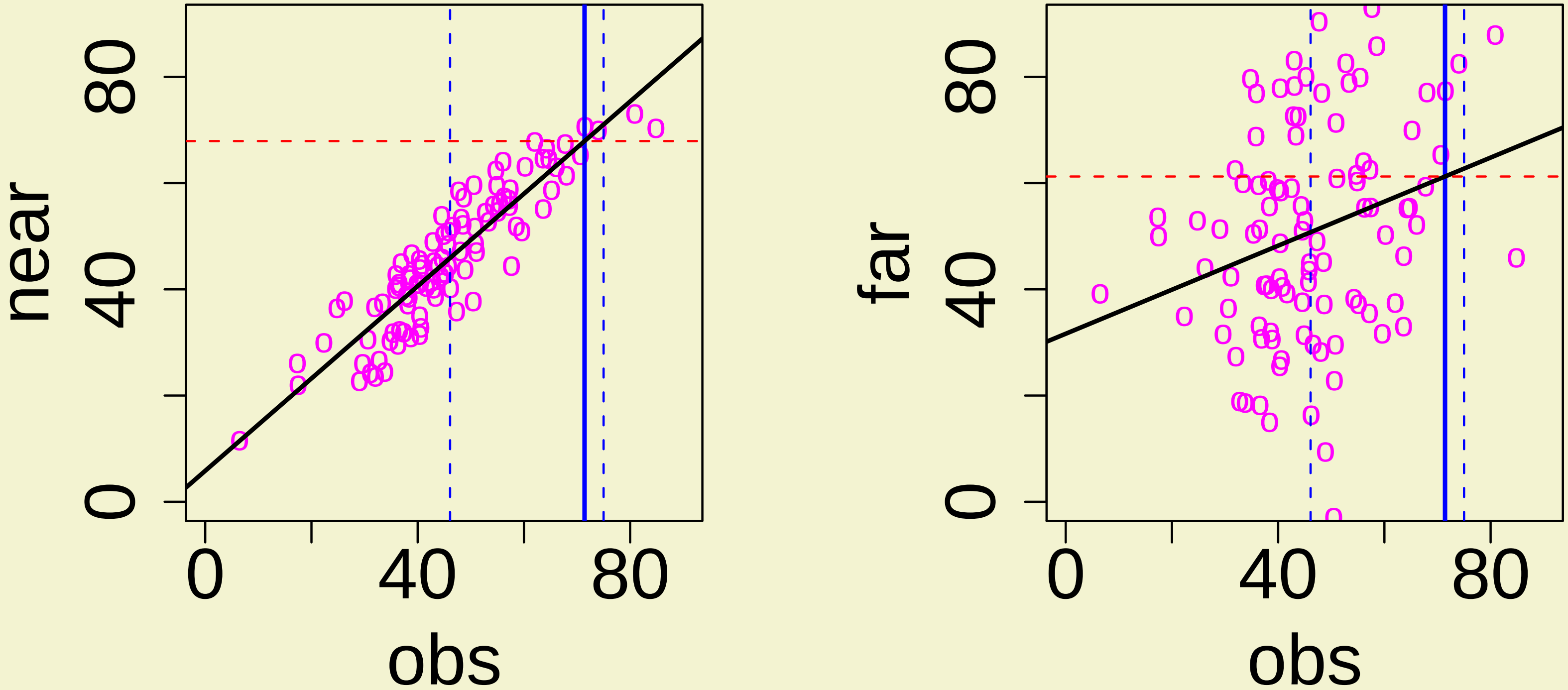
These are least squares lines.

Adjusting for measurement error



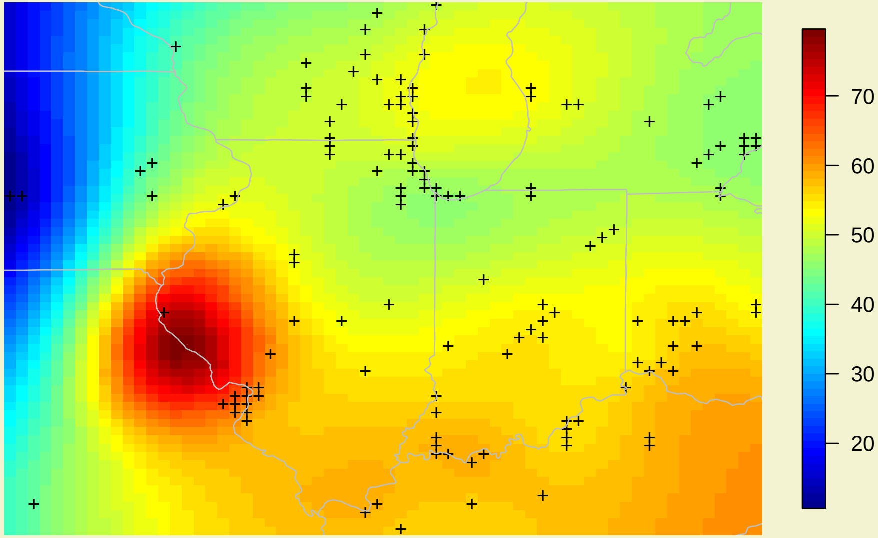
Y= 75 has some error, so adjust for this by shrinking toward the ensemble mean. The Kalman filter tells you how to do this.

The Machine = (up and over)(shrink to mean)[data]



The estimated mean ozone surface

Apply the machine to estimate all grid points.



There is something wrong here?

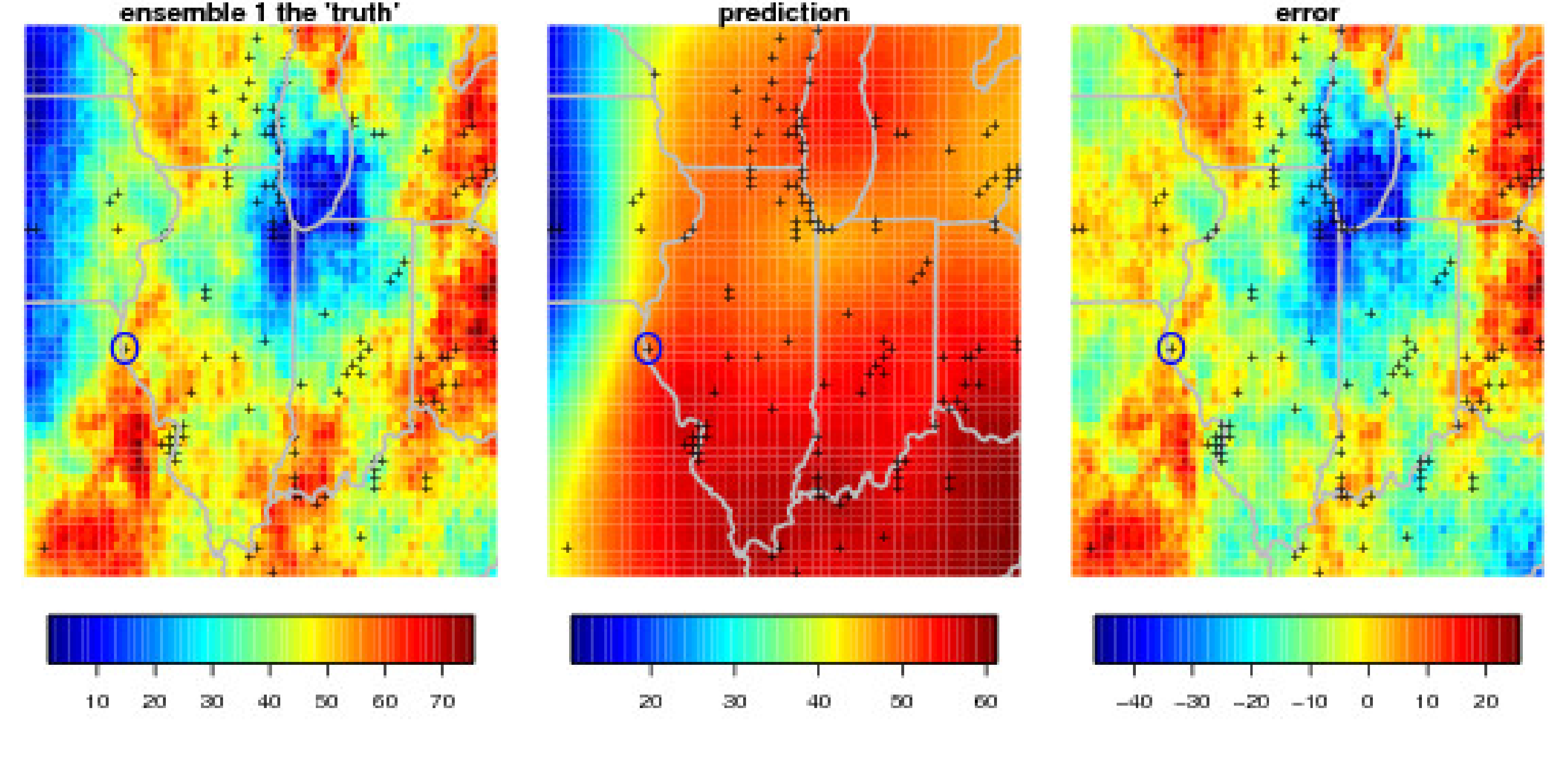
Need to include the uncertainty

Simulating an error field

- Choose an ensemble member (from the prior) and call this ”truth”.
- Generate a pseudo observation at the observation location by adding noise to the ensemble value.
- Based on the pseudo data predict the field using The Machine.
- (prediction - truth) is a draw from the error distribution.

NOTE: this is completely unrelated to the actual data!

Simulating the error field with pictures



This is a likely error field from assimilating the first observation.

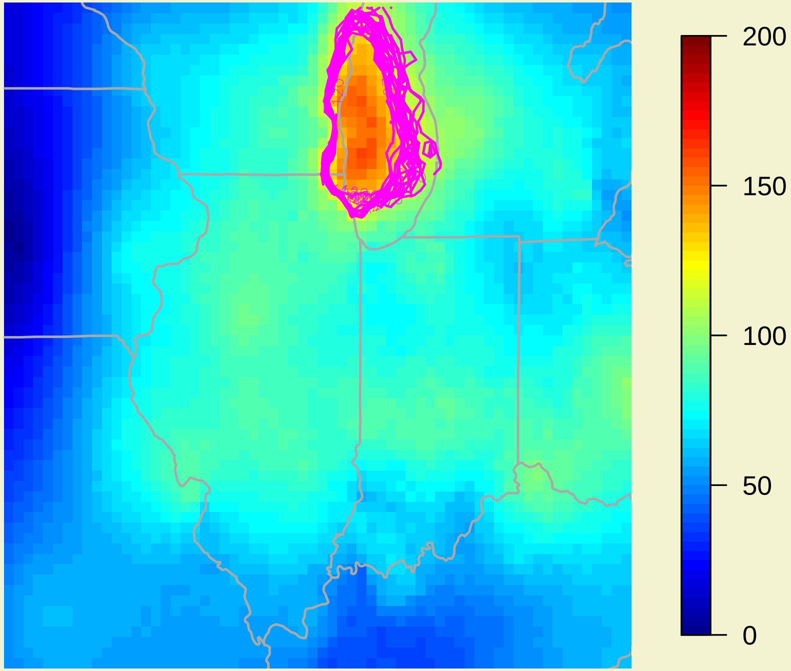
Putting it all together

- For an observation use The Machine to estimate the mean $\hat{g}$
- Use each ensemble member and The Machine to simulate an error field, u_k
- $\hat{g} + u_k$ are the new ensemble members
- Repeat with next observation.

The ensemble members will tend to agree in data rich areas.

The Movie

An inference: Ensemble contours exceeding 120PPB



Doing the math

Prior is $N(\mu, \Sigma)$ **and** $Y = H\boldsymbol{g} + e$ **with** $e \sim N(0, R)$

THE MACHINE

$$\hat{\boldsymbol{g}} = \mu + K(Y - H\mu)$$

$$\hat{\boldsymbol{g}} = \mu + \Sigma H^t (H\Sigma H^t)^{-1} (H\Sigma H^t)(H\Sigma H^t + R)^{-1}(Y - H\mu)$$

least squares **shrink to** μ

Draw an error

$z \sim N(\mu, \Sigma)$ and psuedo data $Y^* = Hz + e^*$

$u = \textit{THE MACHINE } (Y^*) - z$

$\hat{g} + u$ is a draw from the posterior.

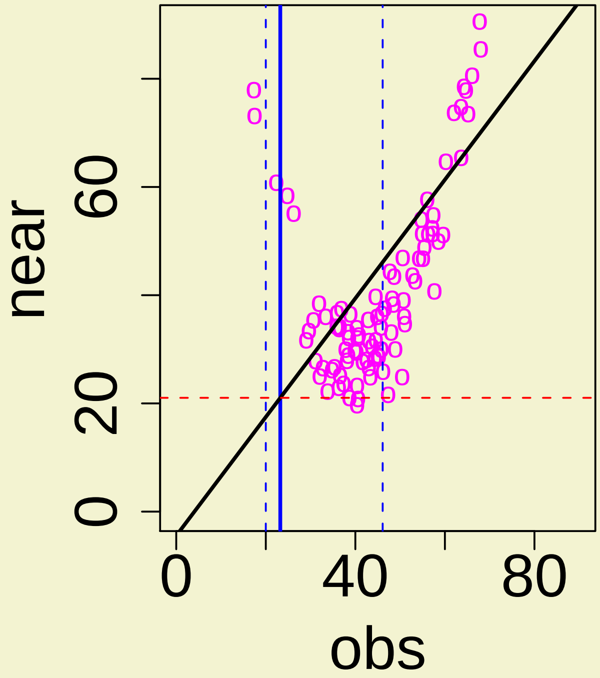
When these are combined into a single step it is the perturbed obs method.

The ensemble approximation

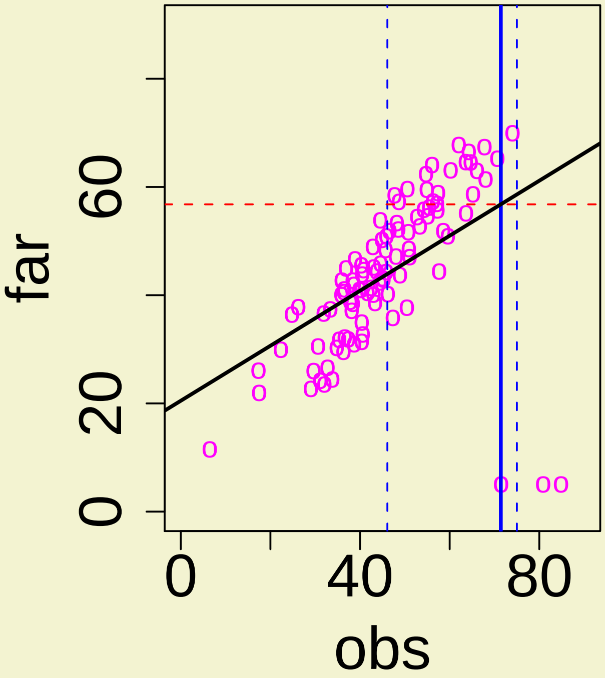
- Replace the prior mean by the ensemble sample mean.
- Replace the prior covariance by the ensemble sample covariance.
- Use the ensemble members as draws from the prior.
- Adjust and downweight for the variability of the ensemble statistics.

Some Research

Nonlinear relationships



or outliers



Need a new MACHINE!

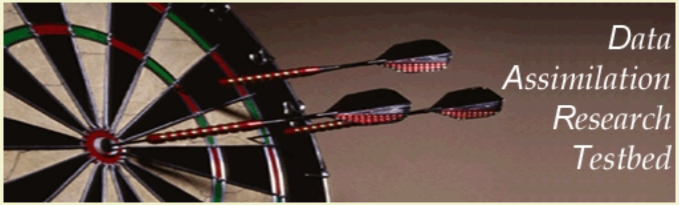
Summary

Updating an ensemble with a single observation is a simple operation related to linear regression of the observed state on the unknown ones.

The variability in the ensemble can be simulated using the same operation.

Sequentially updating can represent a complex surface and a useful measure of uncertainty.

Try this in your home or office



DART *Data Assimilation Research Testbed*