A Highly Biased View of the 'ART' of Data Assimilation

Jeffrey L. Anderson
NCAR Data Assimilation Initiative
8 December, 2003

I. Overview and some methods
    Big problems require clever simplification

II. Challenges
    A. Model bias
    B. Balances and attractors
    C. Assimilation and discrete distributions

III. Opportunities
    Field is maturing; theory and methods that are easy to apply
    Software engineering advances make it easier to get started
    Efforts like Data Assimilation Research Testbed (DART) underway

The Data Assimilation Problem

Given:

1. A physical system (atmosphere, ocean...)

---

2. Observations of the physical system

   Usually sparse and irregular in time and space
   Instruments have error of which we have a (poor) estimate
   Observations may be of ‘non-state’ quantities
   Many observations may have very low information content

---

3. A model of the physical system

   Usually thought of as approximating time evolution
   Could also be just a model of balance (attractor) relations
   Truncated representation of ‘continuous’ physical system
   Often quasi-regular discretization in space and/or time
   Generally characterized by ‘large’ systematic errors
   May be ergodic with some sort of ‘attractor’

We want to increase our information about all three pieces:

---

1. Get an improved estimate of state of physical system

Includes time evolution and ‘balances’
Initial conditions for forecasts
High quality analyses (re-analyses)

---

2. Get better estimates of observing system error characteristics

Estimate value of existing observations
Design observing systems that provide increased information

---

3. Improve model of physical system

Evaluate model systematic errors
Select appropriate values for model parameters
Evaluate relative characteristics of different models

Examples:

1. Numerical Weather Prediction

Model: Global troposphere / stratosphere O(1 degree by 50 levels)

Observations: radiosondes twice daily, surface observations, satellite winds, aircraft reports, satellite radiances, etc.

2. Tropical Upper Ocean State Estimation (ENSO prediction)

Model: Global (or Pacific Basin) Ocean O(1 degree by 50 levels)

Observations: Surface winds (possibly from atmospheric assimilation), TAO buoys, XBTs, satellite sea surface altimetry

3. Mesoscale simulation and prediction

Model: Regional mesoscale model (WRF), O(1km resolution)

Observations: Radial velocity from Doppler radar returns

4. Global Carbon Sources and Sinks

Nonlinear Filtering (A Bayesian Perspective)

Dynamical system governed by (stochastic) DE:

$$dx_t = f(x_t, t) + G(x_t, t)d\beta_t, \qquad t \geq 0 \tag{1}$$

Observations at discrete times:

$$y_k = h(x_k, t_k) + v_k; \;\; k = 1, 2, \ldots; \qquad t_{k+1} > t_k \geq t_0 \tag{2}$$

Observational error is white in time and Gaussian (nice, not essential)

$$v_k \rightarrow N(0, R_k) \tag{3}$$

Complete history of observations is:

$$Y_\tau = \{y_l; \;\; t_l \leq \tau\} \tag{4}$$

Goal: Find probability distribution for state at time t:

$$p(x, t | Y_t) \tag{5}$$

Nonlinear Filtering (cont.)

State between observation times obtained from DE

Need to update state given new observation:

$$p\left(x, t_k \middle| Y_{t_k}\right) = p\left(x, t_k \middle| y_k, Y_{t_{k-1}}\right) \tag{6}$$

Apply Bayes' rule:

$$p\left(x, t_k \middle| Y_{t_k}\right) = \frac{p\left(y_k \middle| x_k, Y_{t_{k-1}}\right) p\left(x, t_k \middle| Y_{t_{k-1}}\right)}{p\left(y_k \middle| Y_{t_{k-1}}\right)} \tag{7}$$

Noise is white in time (3) so:

$$p\left(y_k \middle| x_k, Y_{t_{k-1}}\right) = p(y_k \middle| x_k) \tag{8}$$

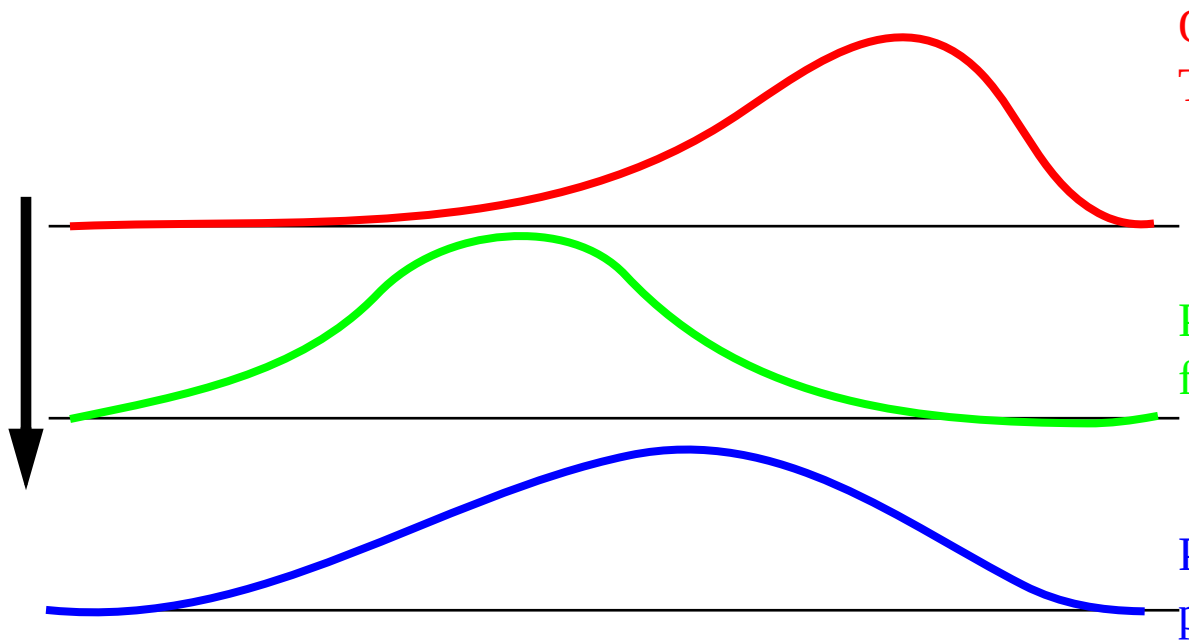
Also have:

$$p\left(y_k \middle| Y_{t_{k-1}}\right) = \int p(y_k \middle| x) p\left(x, t_k \middle| Y_{t_{k-1}}\right) dx \tag{9}$$

Nonlinear Filtering (cont.)

Probability after new observation:

$$p\left(x, t_k \middle| Y_{t_k}\right) = \frac{p(y_k|x)\, p\left(x, t_k \middle| Y_{t_{k-1}}\right)}{\int p(y_k|\xi)\, p\left(\xi, t_k \middle| Y_{t_{k-1}}\right) d\xi} \qquad (10)$$



Observational Error Term

Prior State Estimate, from Model

Posterior (normalized product)

General methods for solving the filter equations are known:

1. Advancing state estimate in time

2. Taking product of two distributions

But, these methods are far too expensive for problems of interest

1. Huge model state spaces (10 is big!), NWP models at O(10 million)

2. Need truncated representations of probabilistic state to avoid exponential solution time and storage

The ART of Data Assimilation:

Find heuristic simplifications that make approximate solution affordable

1. Localization (spatial or other truncated basis)

2. Linearization of models, represent time evolution as linear
   (around a control non-linear trajectory)

3. Represent distributions as Gaussian (or sum of Gaussians)

4. Monte Carlo methods

5. Application of simple balance relations

6. Many others...

Kalman Filter

Simplifications:

1. Linearization of model around non-linear control trajectory

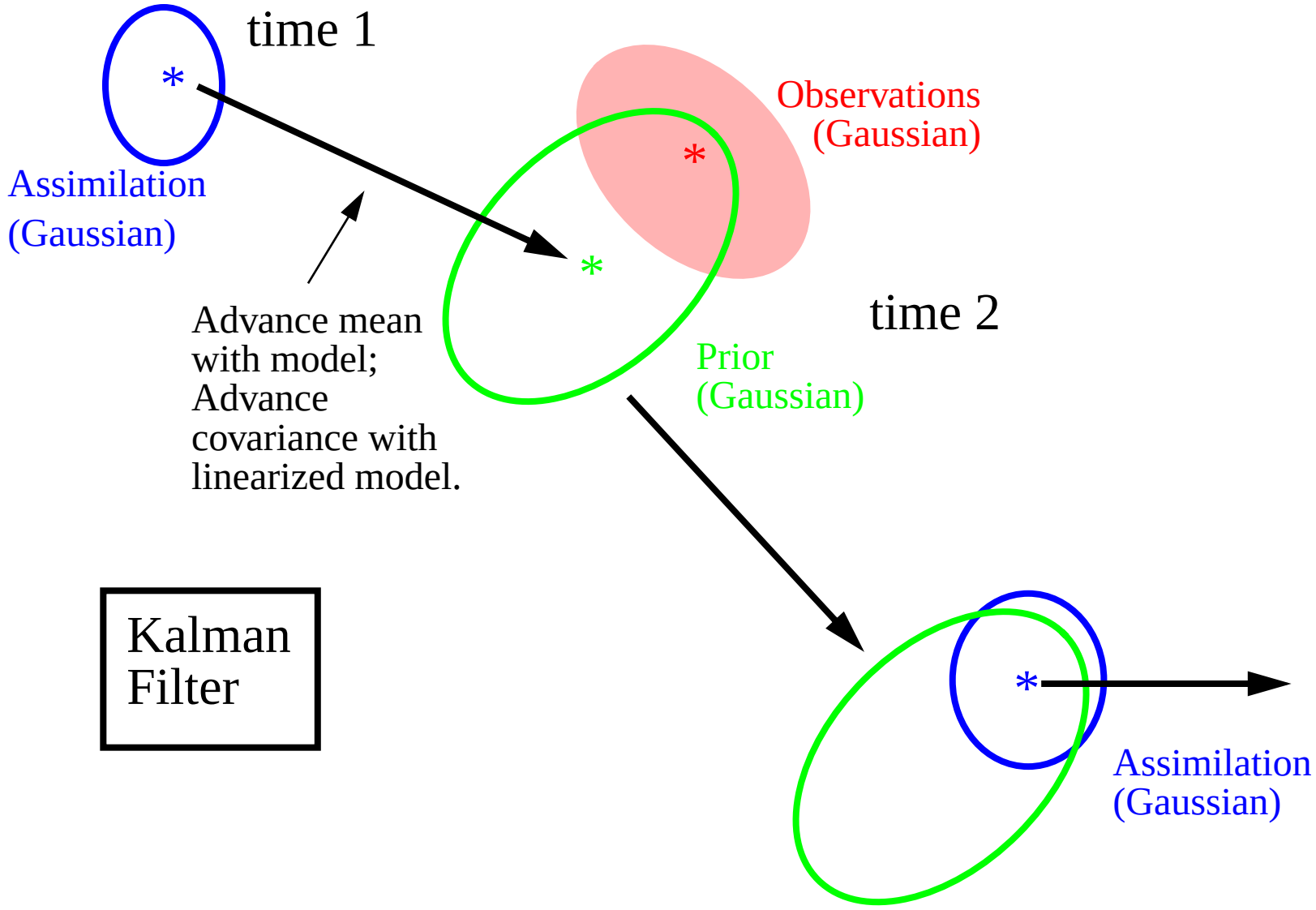
2. Error distributions assumed Gaussian

Fundamental Problem:

Still too expensive for large models

Advancing covariance in linearized model is at least:

O(model_size * model_size)


time 1
Assimilation
(Gaussian)
Observations
(Gaussian)
Advance mean
with model;
Advance
covariance with
linearized model.
time 2
Prior
(Gaussian)
Kalman
Filter
Assimilation
(Gaussian)

Reduced Space Kalman Filters:

Additional simplification:

Assume that covariance projects only on small subspace of model state

Evolving covariance in linearized model projected on subspace may be cheap

Subspace selection:

1. Dynamical: use simplified model based on some sort of scaling

2. Statistical: use long record of model (or physical system) to find reduced basis in which most variance occurs (EOF most common to date)

Problems:

1. Dynamics constrained to subspace may provide inaccurate covariance evolution

2. Observations may not ‘project strongly’ on subspace

3. Errors orthogonal to subspace unconstrained, model bias in these directions can quickly prove fatal

Ensemble Kalman Filters:

Simplifications:

1. Monte Carlo approximation to probability distributions

2. Localization in space, avoids degeneracy from samples smaller than state space and reduces sampling noise

3. Gaussian representation of probability distributions generally used for computing update

Problems:

1. Selecting initial samples for ensembles (Monte Carlo samples)

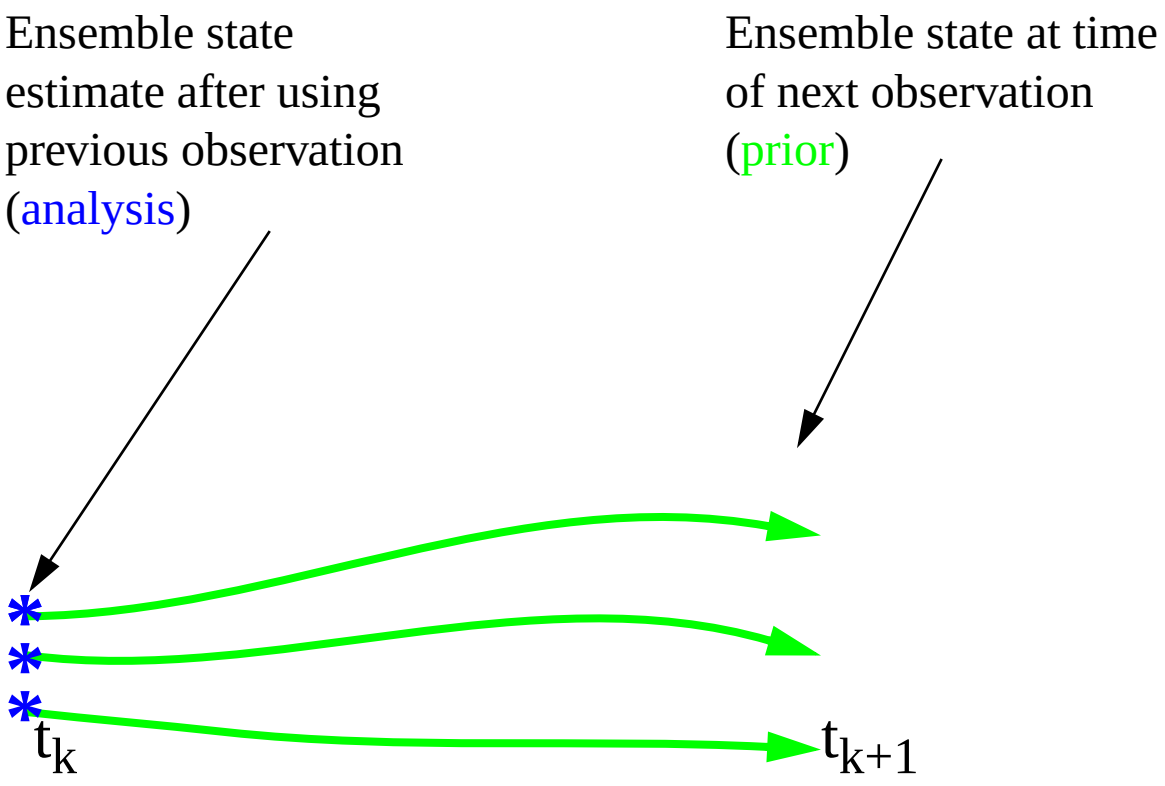
2. Determining degree of spatial localization; sampling error

3. Maintaining appropriate model ‘balances’ in ensemble members

BUT, UNPRECEDENTED EASE OF INITIAL APPLICATION

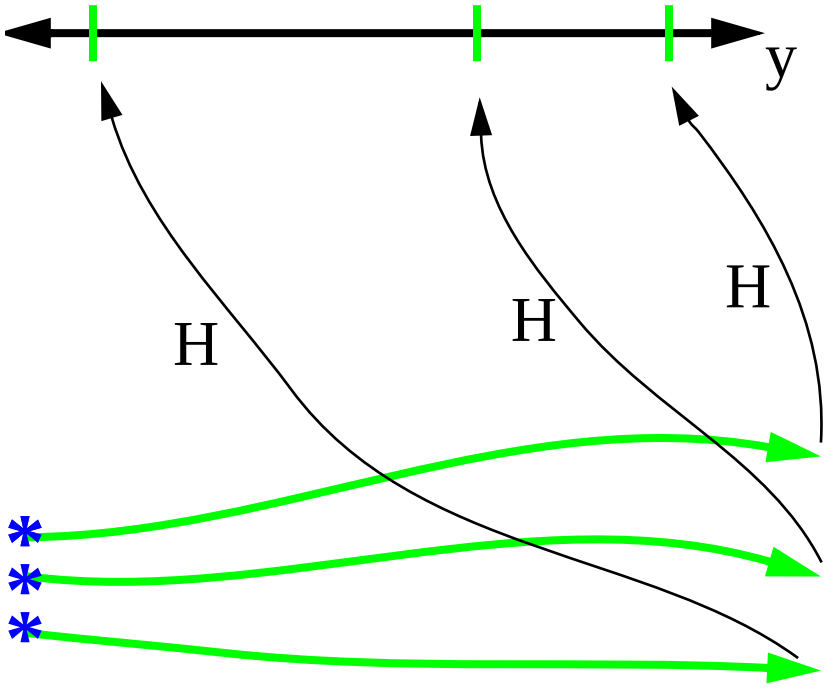
How an Ensemble Filter Works

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



How an Ensemble Filter Works

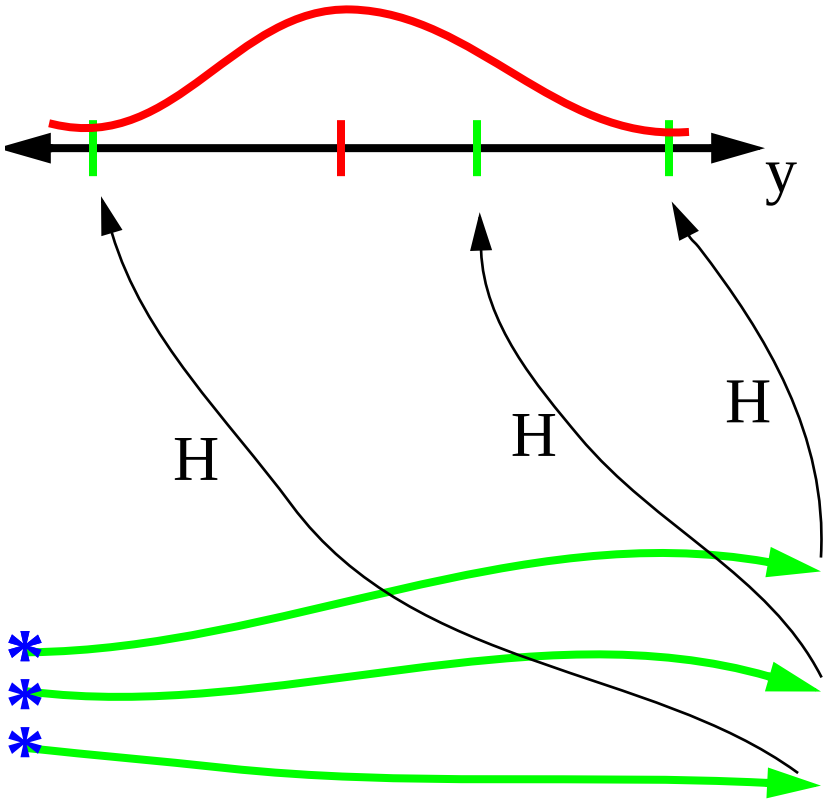
2. Get prior ensemble sample of observation, y=H(x), by applying forward operator H to each ensemble member



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

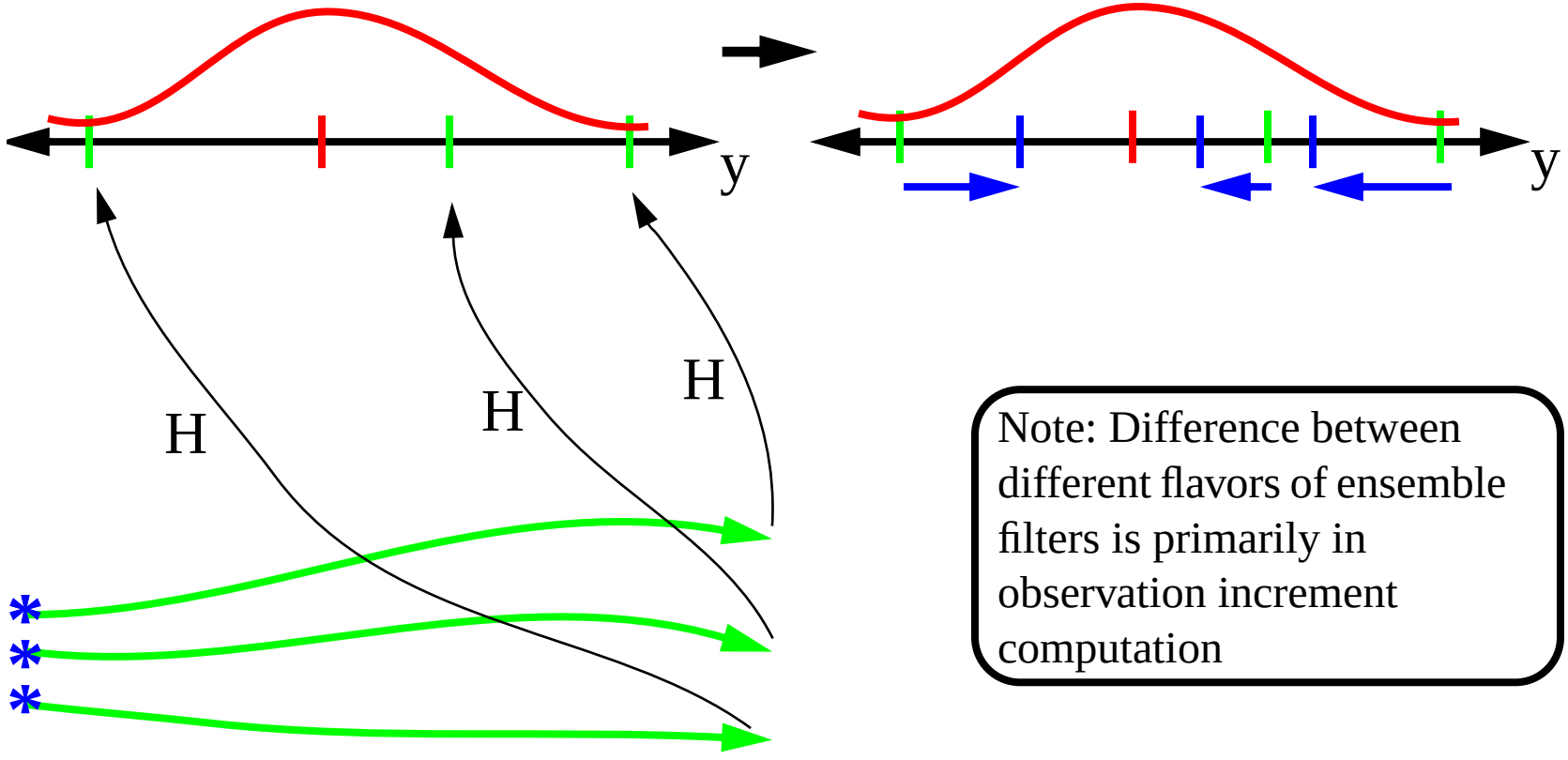
How an Ensemble Filter Works

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



How an Ensemble Filter Works

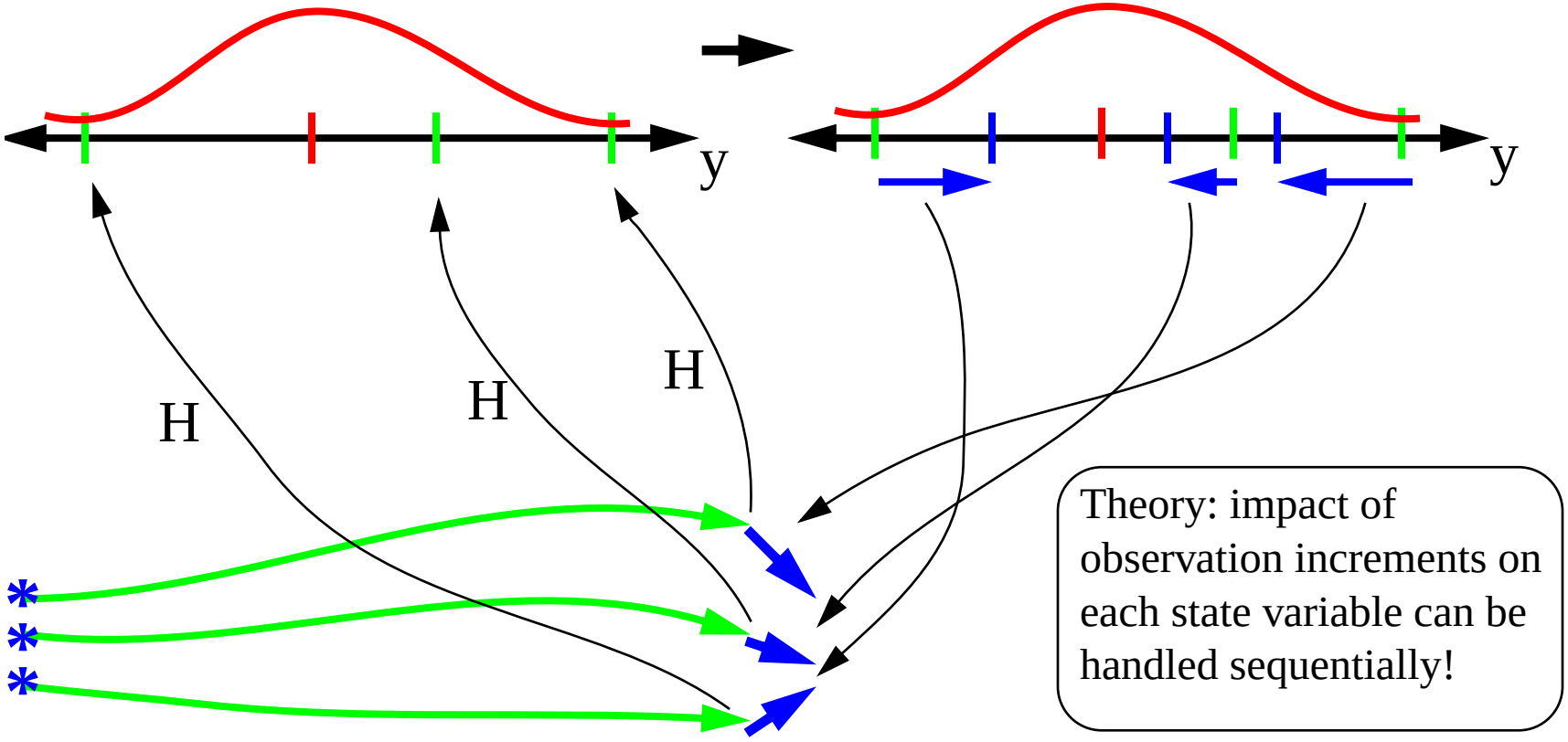
4. Find increment for each prior observation ensemble
(this is a scalar problem for uncorrelated observation errors)



Note: Difference between different flavors of ensemble filters is primarily in observation increment computation

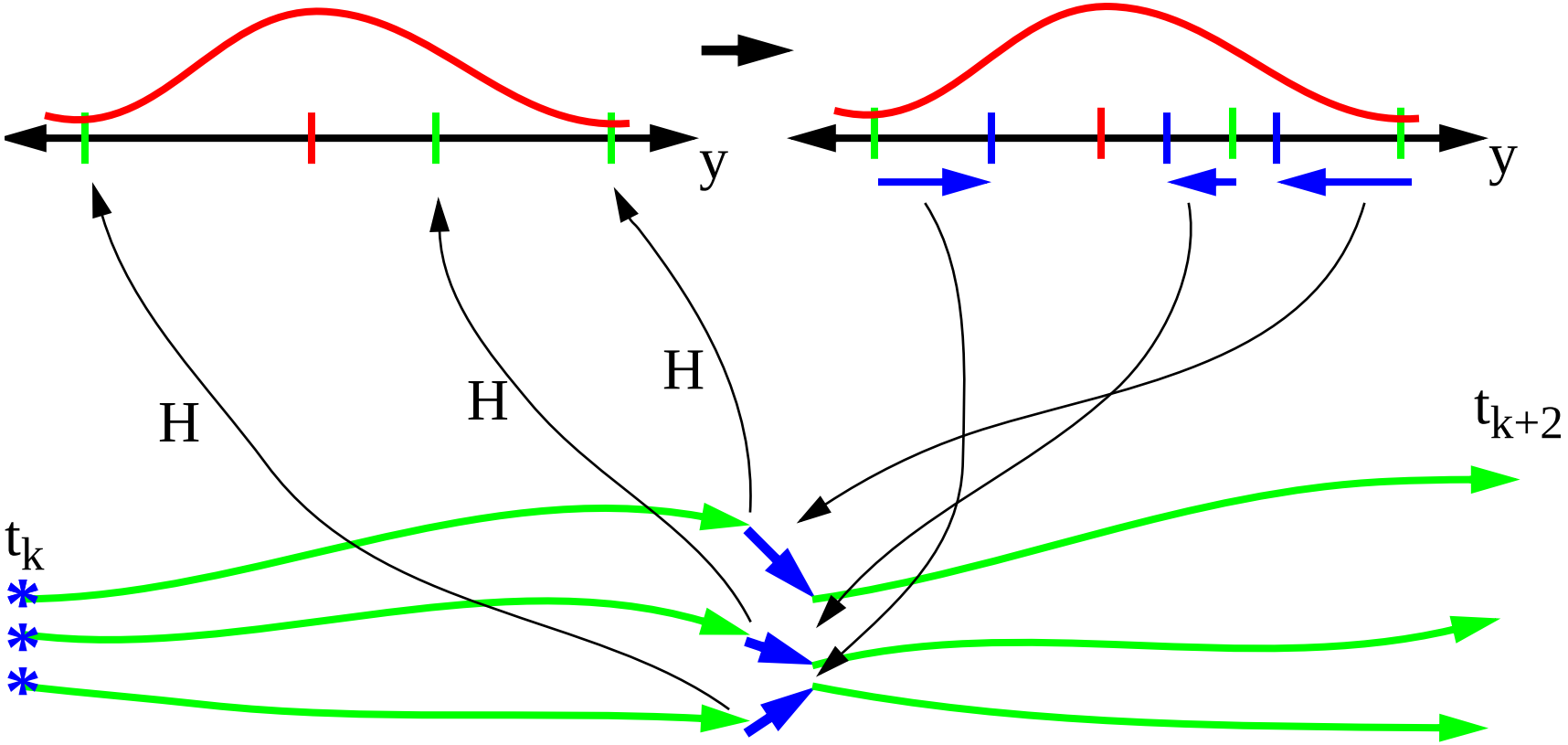
How an Ensemble Filter Works

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



How an Ensemble Filter Works

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



Details of Step 4: Finding Increments for Observation Variable Ensemble, y

Scalar Problem: Wide variety of options available and affordable. Examples:

1. Perturbed Observation Ensemble Kalman Filter (EnKF); stochastic

2. Ensemble Adjustment Kalman Filter (EAKF); deterministic

Key to Kalman Filters: Product of Gaussians is Gaussian

Prior ensemble sample mean $\bar{y}^p$ and variance Σ^p

Observation y^o with observational error variance Σ^o

Posterior Variance is:

$$\Sigma^u = \left[\left(\Sigma^p\right)^{-1} + \left(\Sigma^o\right)^{-1} \right]^{-1} \qquad (11)$$

and mean is:

$$\bar{y}^u = \Sigma^u \left[\bar{y}^p / \Sigma^p + y^o / \Sigma^o \right] \qquad (12)$$

Details of Step 4: Perturbed Observation Ensemble Kalman Filter (EnKF)

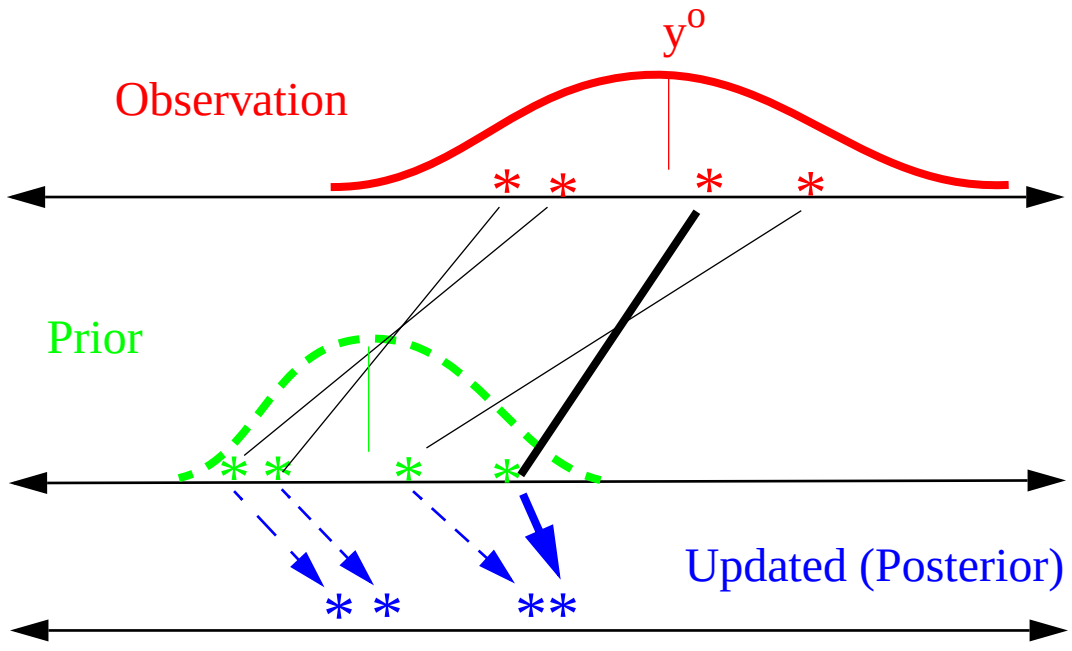
1. Apply (11) once to compute updated covariance Σ^u

2. Create N-member sample of observation dist. by adding samples of obs. error to y^o

3. Apply (12) N times to compute updated ensemble members, $\bar{y}^u_i$

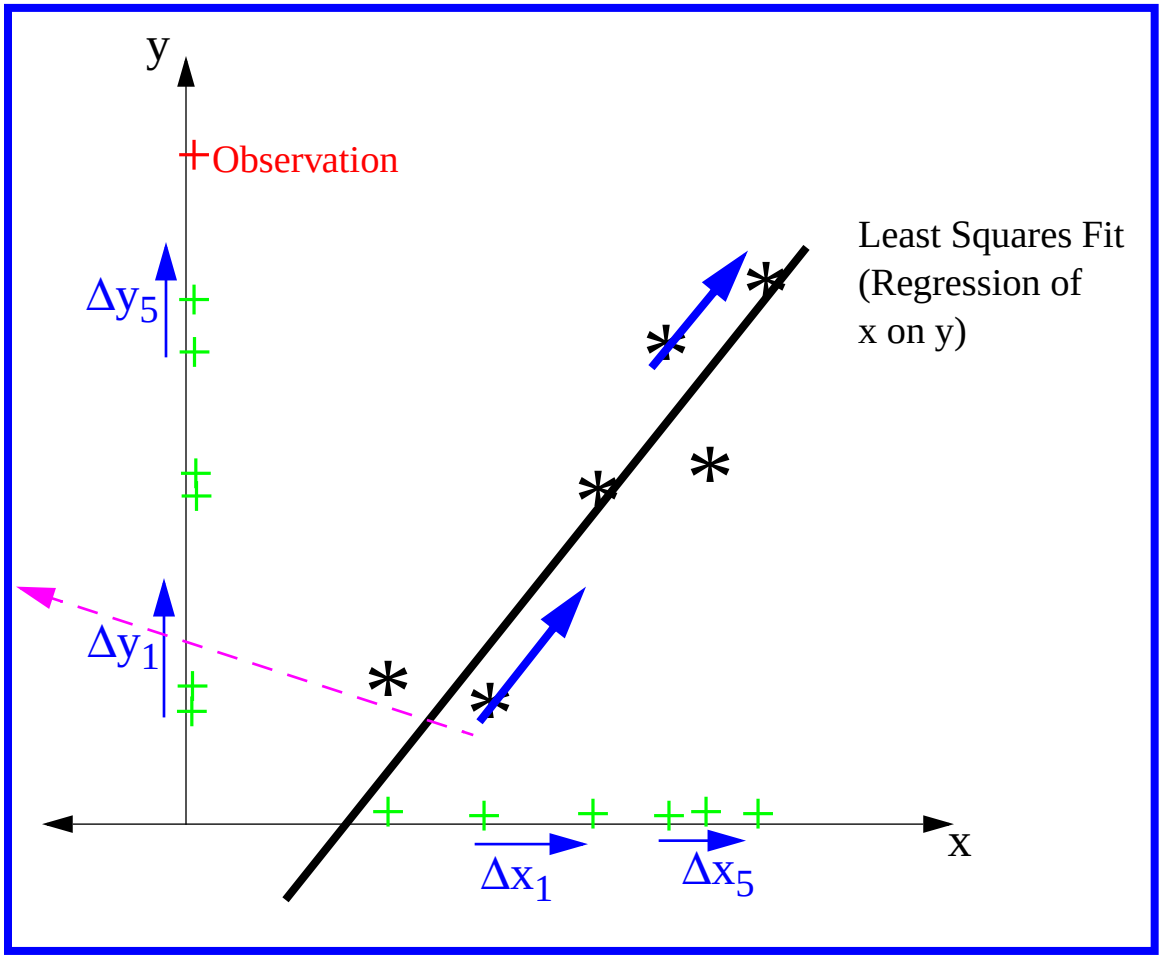
   Replace $\bar{y}^p$ with ith prior ensemble member, y^p_i

   Replace y^o with ith value from random sample, y^o_i



Details of Step 5: Compute state var. increments from obs. variable increments

Regression using joint sample statistics from ensembles: can be done sequentially!



Regression begins with least squares fit to sample, *

Increments for state variable, x, multiplied by |correl(x, y)|

Large sample size needed to filter ‘noise’

Trade-offs with ‘local’ linearization:
Precision vs. accuracy

A Host of Challenges Remain

Problem 1. Sampling error impacts estimates of increments

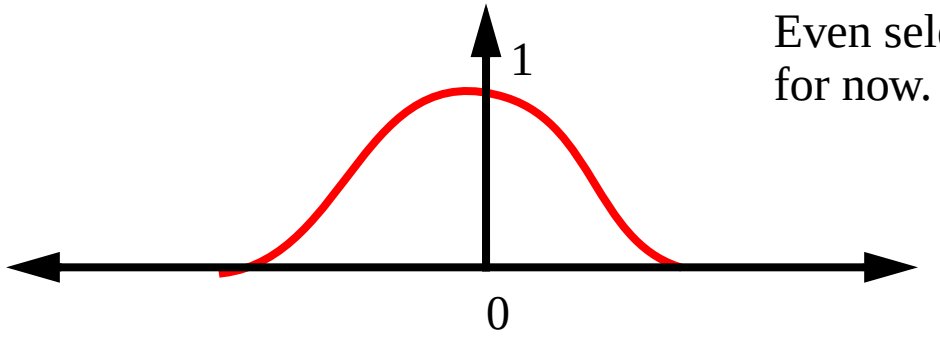
Key: estimates of regression coefficients have errors

Many obs. with small (or zero) expected correlations => error build-up

Solution: Reduce impact of observations as function of ensemble size, sample correlation, and expected distribution of correlation

But...need this prior estimate (may be mostly unknown?)

For now, use distance dependent envelope to reduce impact of remote observations



Even selecting envelope is tricky for now.

Problem 2. Initial conditions for ensembles

Key: Bayesian, assumes initial ensembles are magically available

Solution: For ergodic models spin-up by running ensemble a very long time from arbitrary initial perturbations, slowly ‘turn on’ observations

But... this may be impossible for some models (WRF regional applications)

Given prior knowledge of expected correlations (see problem 1) should be able to generate appropriate ensemble ICs

Still a topic for ongoing research

Problem 3. Assimilation of variables with discrete distributions

Key: ensemble prior may indicate zero probability of an event that is occurring

i.e. All ensemble members say no rain but rain is observed

Directly related to existence of discrete convective cells

Solutions: Apply methods for accounting for model error

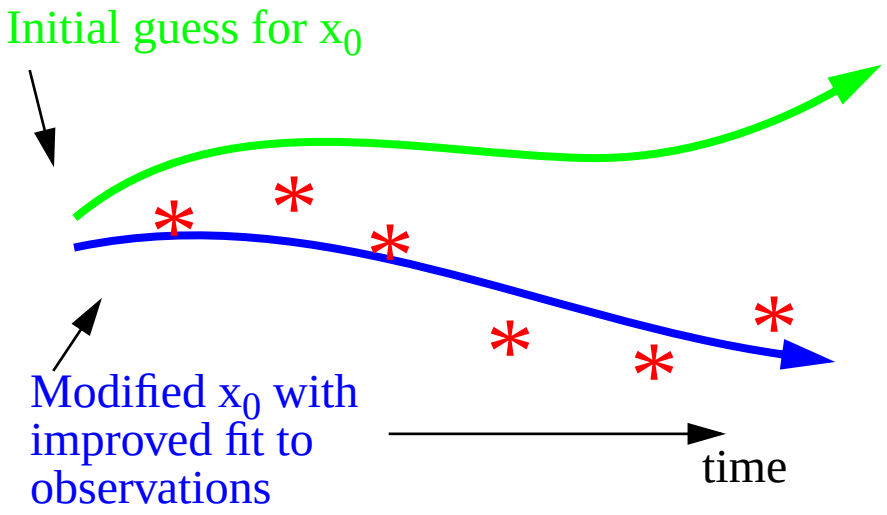
Redefine state variables to avoid discrete probability densities

Research on this problem is in its infancy

4D-Variational (4D-Var): State of the Art for Global Weather Prediction

Find model trajectory through time minimizing measure of departure from observations

Applied over some finite period of observations



For optimization, need gradient of norm with respect to initial state, X_0

Key: integrating adjoint of linear tangent model linearized around forward non-linear trajectory backward in time allows computation of gradient with single integration pair

This makes 4D-Var feasible as long as period is short and number of iterations needed for optimization is small

Additional problems:

1. Model 'balance' constraints may not be satisfied for finite optimization periods
2. Still hard to generate adjoints for complicated models
3. May need to relax constraints to deal with model BIAS

Exciting Opportunities Abound in Data Assimilation

Field is maturing, basic theory well-understood

Increasingly powerful heuristic methods being developed

Some new methods (like filter) are very simple to implement (naively)

Software engineering advances make it easier to access models and data

NCAR/NOAA are building a prototype facility for exploring DA

1. The challenges are opportunities!

2. Plethora of models and observations that have not been touched!

3. Improved assimilation application to existing high profile problems!

   Example: Getting more from existing data, surface pressure observations

   Example: Quality control of observations: using good data, rejecting bad

Data Assimilation Opportunities (cont.)

4. Using data to improve models!

   Example: Application in simple low-order model

5. Stochastic (ensemble) prediction

6. Evaluating and designing observing systems!

   (Observing / Assimilation System Simulation Experiments)

   What is information content of existing observations?

   What is value of additional proposed observations?

   Use of targeted (on demand) observations

   Potential for extremely high impact (if you can stand the heat)

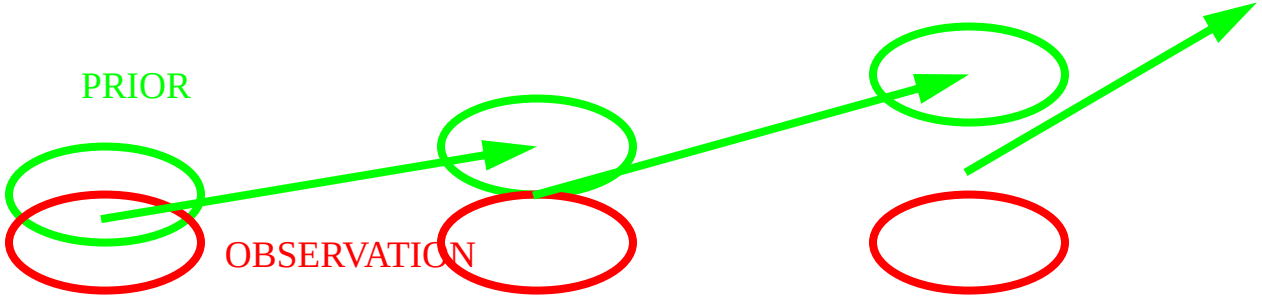
Challenge #1: Model Bias (Systematic Error)

Filter equations assume prior estimate (and observations) are unbiased

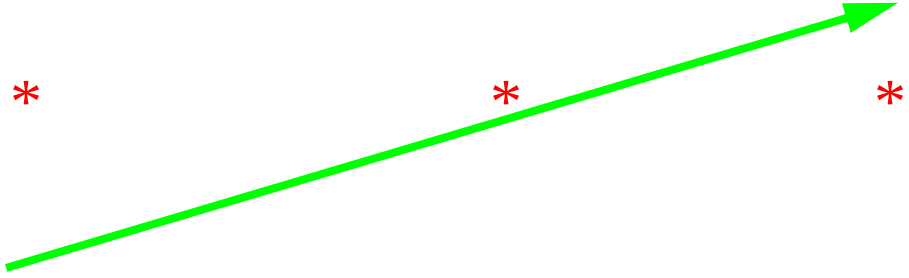
Questionable for Observations, ridiculous for Models

Biased prior estimate will cause observations to be given too little weight

Repeated applications lead to progressively less weight, estimate can diverge



Implications are obvious for 4D-Var, too



Dealing with model bias is mostly an open question:

1. Can reduce confidence in model prior estimates by some constant factor

2. Explicitly model the model bias as an extended state vector and assimilate coefficients of this bias model

Model: $dx/dt = F(x)$

Model plus bias model: $dx/dt = F(x) + \varepsilon(t)$; $d\varepsilon/dt = 0$

where ε is a vector of the same length as x

Very tricky: if we knew much about modeling the bias, we could remove it

Challenge #2: Balances and Attractors

Many models of interest have balances, both obvious (say geostrophic) and subtle

The structure of the model ‘attractors’ may be extremely complex

In some cases, off-attractor perturbations may lead to ‘large’ transient response

Example: High frequency gravity waves in some Primitive Equation models

The behavior of these transients can lead to model bias

In this sense, even perfect model experiments can have large model bias

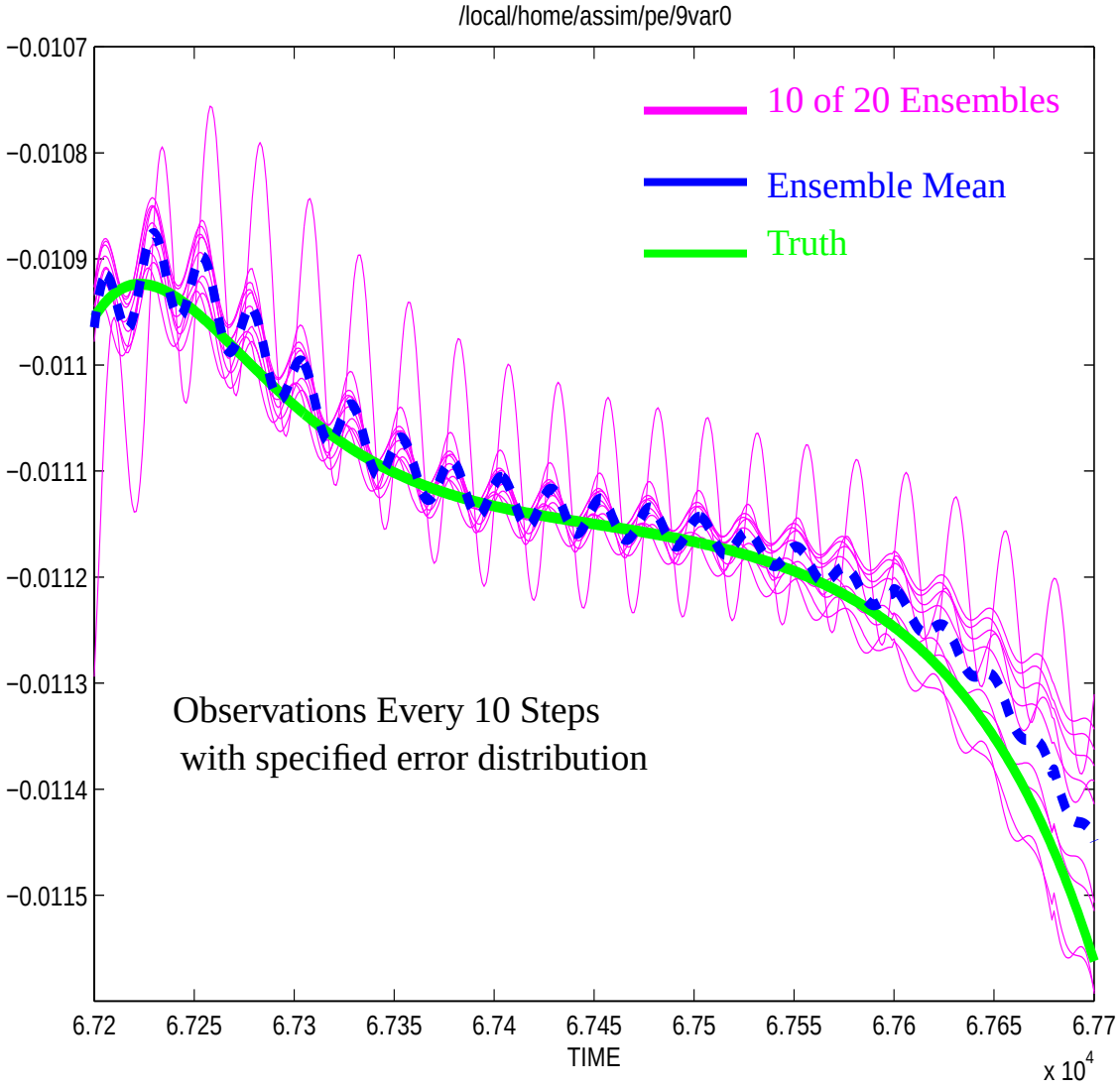
Understanding how to minimize this behavior or limit its impact is a fun problem

The continuous system may also have balances, obvious and subtle

Unclear how differences between model and continuous ‘attractors’ impacts assimilation

Example of model balances: Lorenz 9-Variable Model

Time series of Ensemble Filter Assimilation for variable X_1



Equilibrated model has small high frequency variability

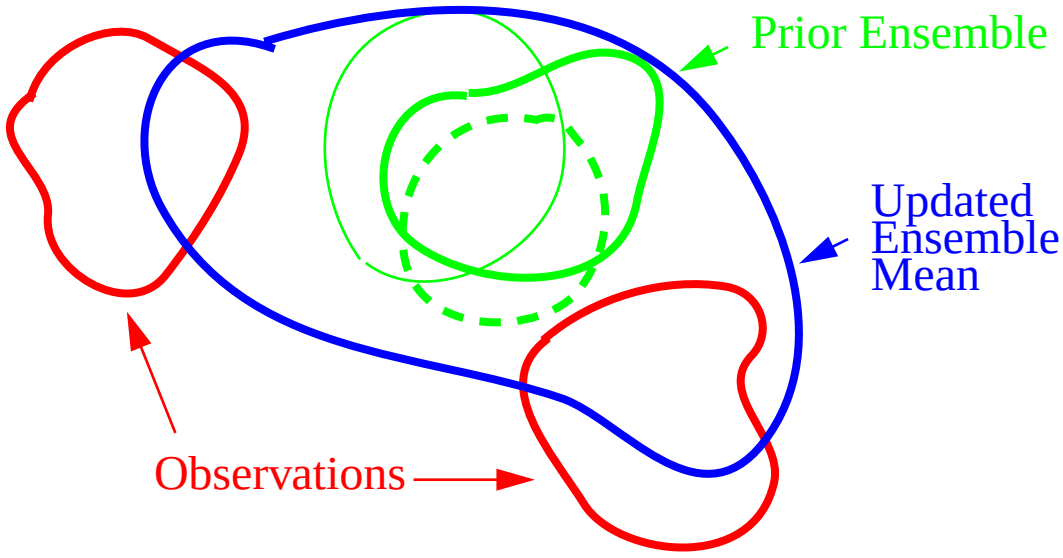
Perturbing off attractor leads to transient, high amplitude, high frequency waves

Assimilation can't avoid some off attractor error

Ensemble high frequency waves tend to be in phase

Result is apparent model bias (all ensemble members on one side of truth)

Challenge #3: Assimilation of Discrete Distributions



Example: assimilation of convective elements

Prior is 'certain' that there are no convective cells outside the green areas

Observations indicate discrete areas outside the green

This is indicative of highly non-linear problem

Ensemble techniques, at best, tend to smear out prior discrete structures

4D-Var is likely to have non-global local minima

But, we think we know what we want to do

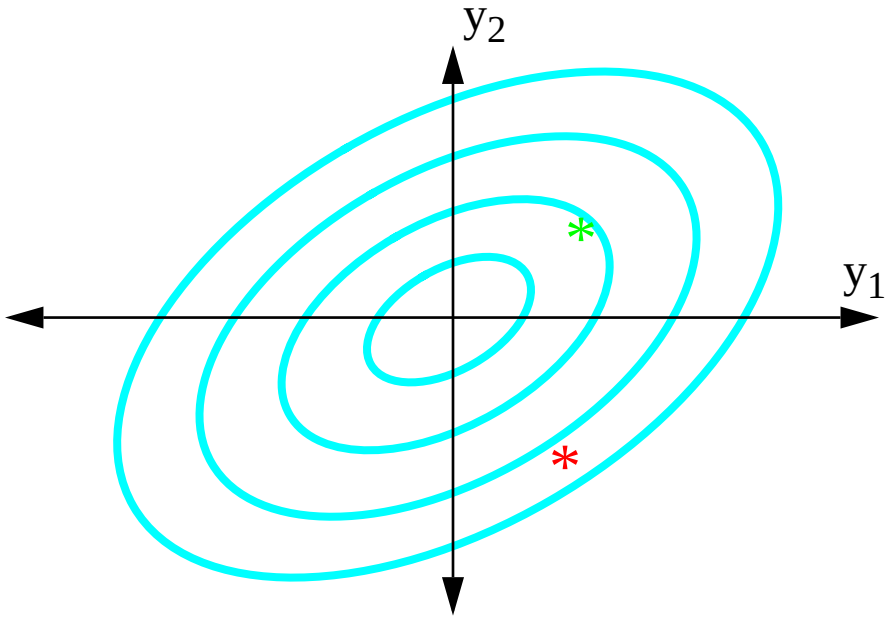
- Keep information from prior on larger scale 'background'
- Introduce cells where observed

Requires new norms or ways to deal with model bias as function of scale

Quality Control of Observations

Methods to exclude erroneous observations:

1. Discard impossible values (negative R.H.)
2. Discard values greatly outside climatological range
3. Discard values that are more than α prior ensemble sample standard deviations away from prior ensemble mean
4. 'Buddy' checks for pairs of observations: just apply chi-square test using prior ensemble covariance and label pair as inconsistent if threshold value exceeded



When both prior and observations can be inconsistent with our prior expectations, detecting and excluding these errors can be VERY DIFFICULT

Using Data Assimilation to Constrain Model Parameters

Example from another low-order model: Lorenz-96 Model

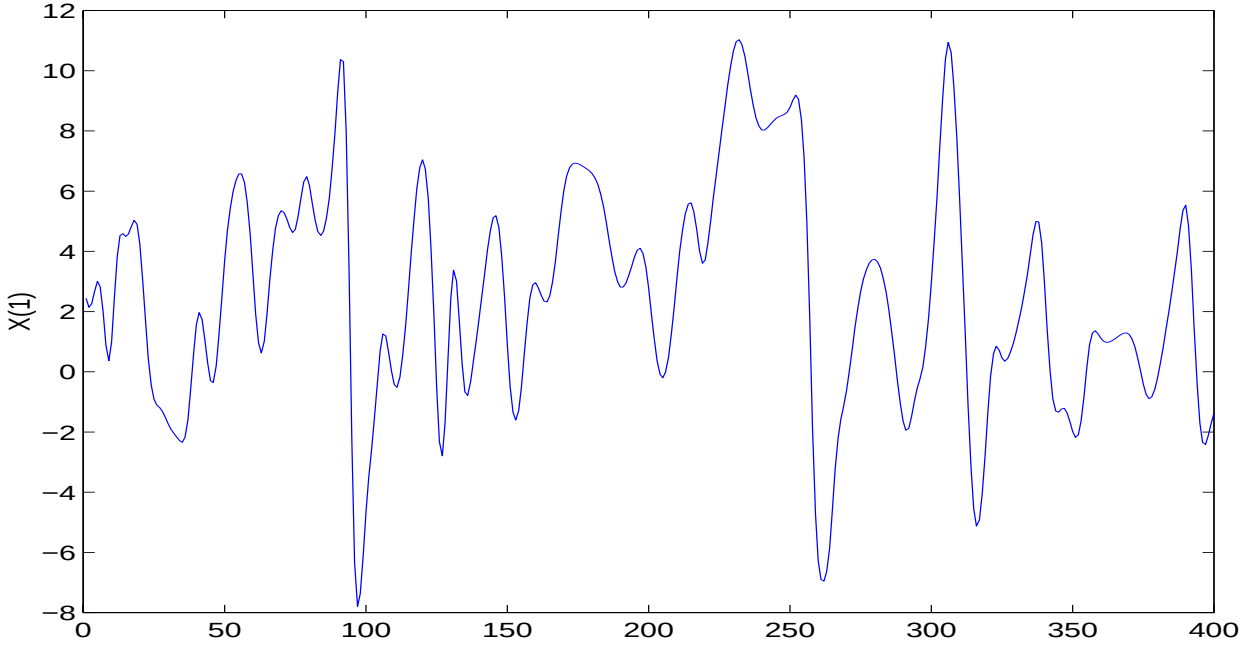
Variable size low-order dynamical system

N variables, $X_1, X_2, ..., X_N$

$dX_i / dt = (X_{i+1} - X_{i-2})X_{i-1} - X_i + F$

$i = 1, ..., N$ with cyclic indices

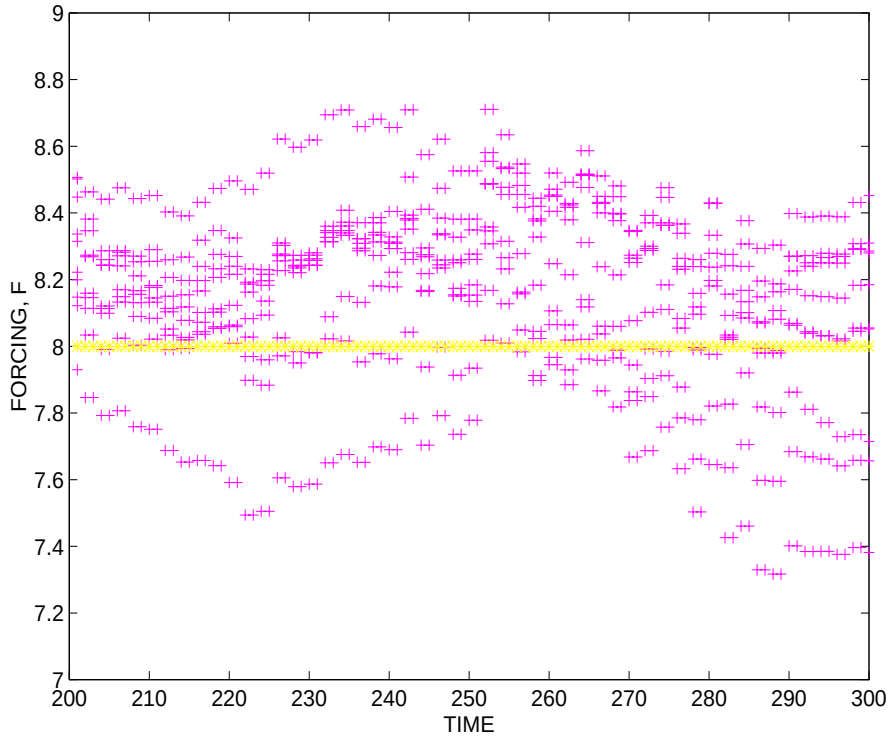
Use N = 40, F = 8.0, 4th-order Runge-Kutta with dt=0.0



Lorenz-96 Free Forcing Model Filter

20 Member Ensemble (10 Plotted) Obs Every 2 Steps of State Variables only

Truth in Yellow (8.0) Ensemble



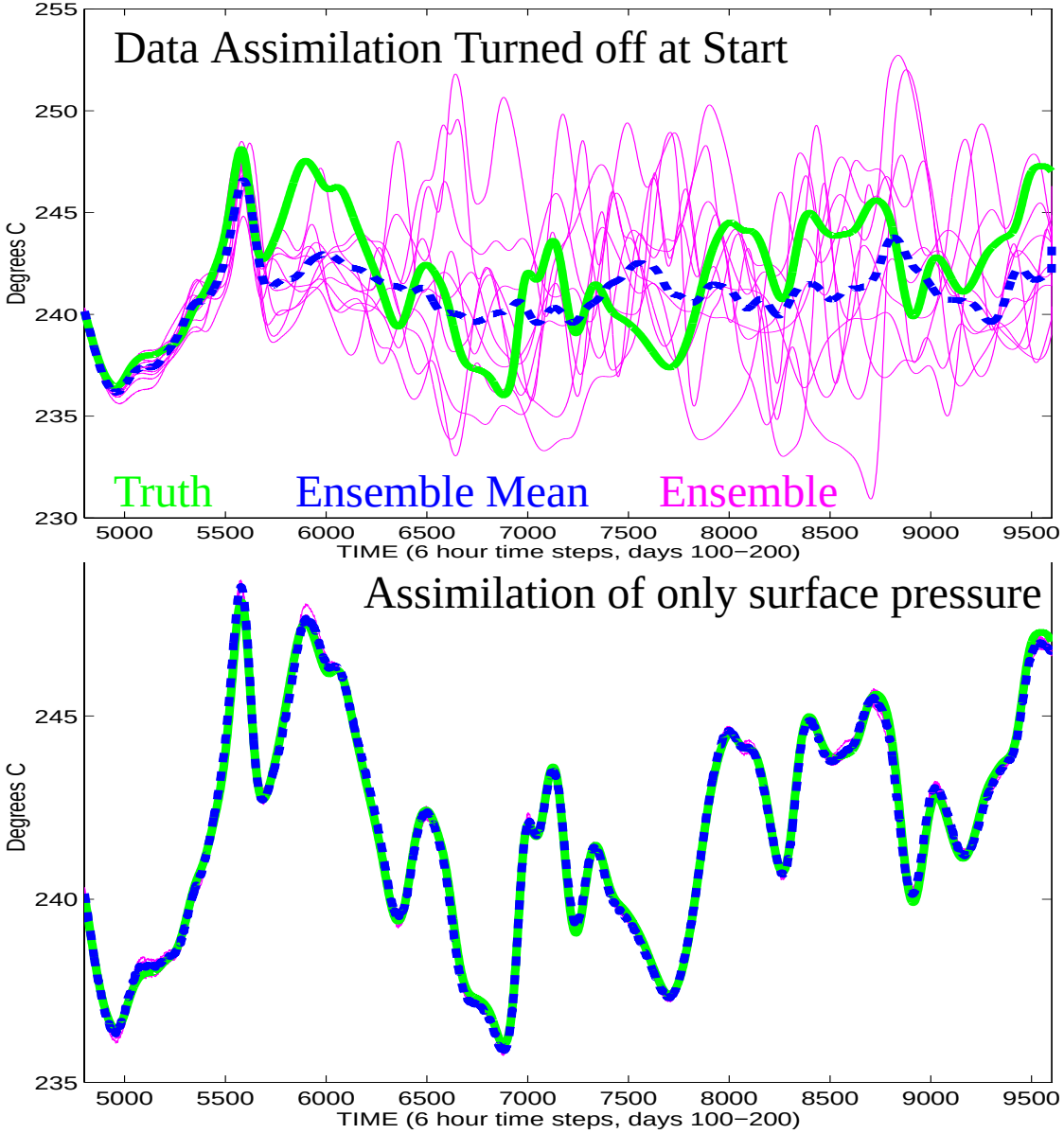
Can treat model parameters as free parameters

Here, the forcing F is assimilated along with the state

This is potential mechanism for dynamic adjustment of unknown parameters and for dealing with unknown model systematic error

Many models include a number of poorly know free parameters
May be able to improve models by using data to constrain these
Observation system parameters can also be constrained by data (obs. error for instance)

Evaluating and Designing Observing Systems: Information Content of Observations



Example: What is information content of surface pressure observations in an atmospheric GCM?

Observing System Simulation Experiment (OSSE) with an Ensemble Filter

Observations generated from a model run (truth in green)

Same model assimilates

Surface pressure is able to closely constrain entire atmosphere

Figure shows mid-latitude, mid-troposphere temperature

Advances and Opportunities in Data Assimilation for many Geophysical Studies

Advances:

Field is maturing; theory and methods that are easy to apply
Software engineering advances make it easier to get started
Efforts like Data Assimilation Research Testbed (DART) underway

Opportunities:

A. A plethora of untouched models and observations
B. Improved assimilation methods for existing problems
C. Improved use of existing observations; quality control
D. Using data to improve models
E. Evaluating value of existing observations
F. Evaluating future observing systems
G. Adaptive observations