An Introduction to the Data Assimilation Research Testbed and Generic Sampling Theory

Report to DAI Scientific Review Panel
24 September, 2004

1. Requirements for a generic assimilation software facility

2. Tactical approach to an effective, cross-cutting initiative

3. Overview of DART

4. Key problems:
   Sampling error
   Efficient parallel filters

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

Fundamental Problems for a Data Assimilation Initiative

1. Cross-cutting initiatives can be too diffuse (everything to everyone)

2. Data Assimilation is notoriously personnel intensive

3. How to have viable collaborations with many modeling/obs. groups?

Fundamental Problems for a Data Assimilation Initiative

1. Cross-cutting initiatives can be too diffuse (everything to everyone)

2. Data Assimilation is notoriously personnel intensive

3. How to have viable collaborations with many modeling/obs. groups?

Our Solution:

Develop a flexible, powerful, generic, data assimilation facility

Ensemble filter algorithms make this possible for now

Towards a General, Flexible Assimilation Facility

Goals:

1. Assimilation that works with variety of models and obs. types
2. Coding for system must be easy to implement (weeks max.)

   (This appears to rule out variational methods at present)
3. Must allow complicated forward operators
4. GOOD assimilation results for novice users
   EXCELLENT results with added expertise/development
5. GOOD performance on variety of platforms with little effort
   EXCELLENT performance with added expertise/development

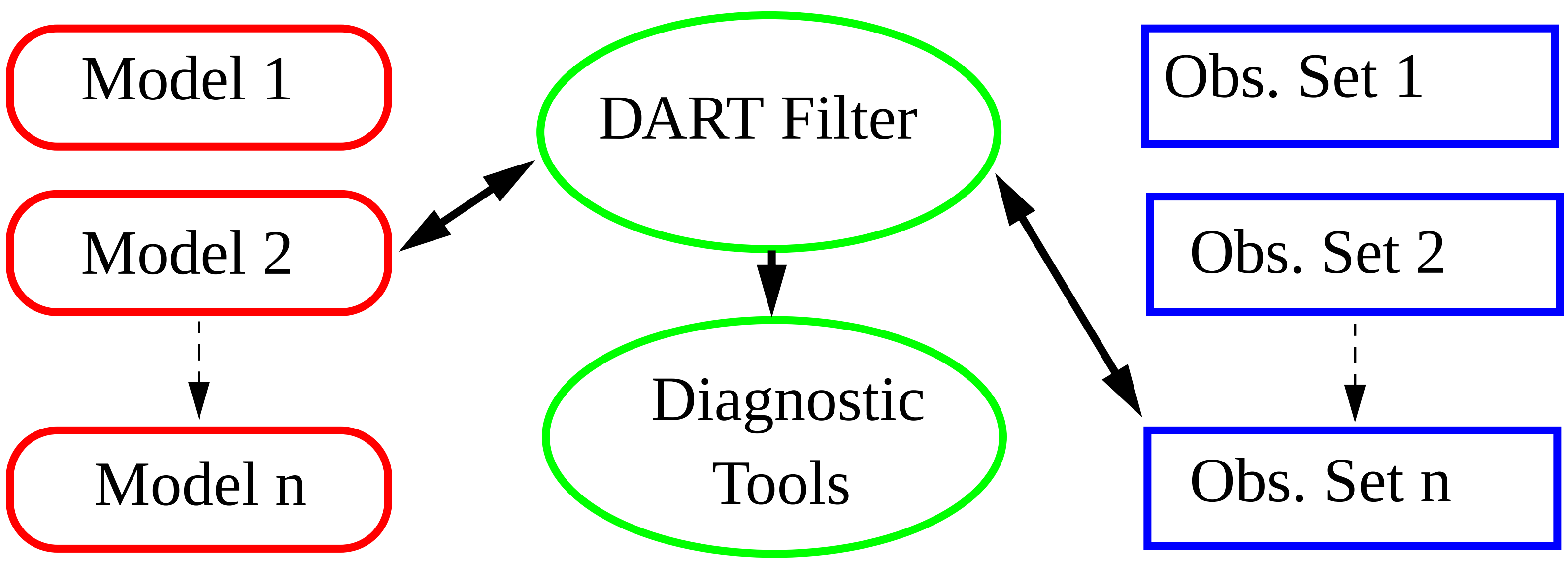
The Data Assimilation Research Testbed (DART)

A. Combine assimilation algorithms, models, and observation sets

B. Diagnostic tools for assimilation experimentation

C. Compliant models and observation sets (real and synthetic)

D. A high-quality, generic ensemble filtering algorithm



DART compliant models (largest set ever with assim system?)

1. Many low-order models (Lorenz63, L84, L96, L2004, ...)
2. Global 2-level PE model (from NOAA/CDC)
3. CGD’s CAM 2.0 & 3.0 (global spectral model)
4. GFDL FMS B-grid GCM (global grid point model)
5. MIT GCM (from Jim Hansen)
6. WRF model
7. NCEP GFS (assisted by NOAA/CDC)
8. GFDL MOM3/4 ocean model
9. ACD’s ROSE model (upper atmosphere with chemistry)

This allows for a hierarchical approach to filter development

DART compliant Forward Operators and Datasets

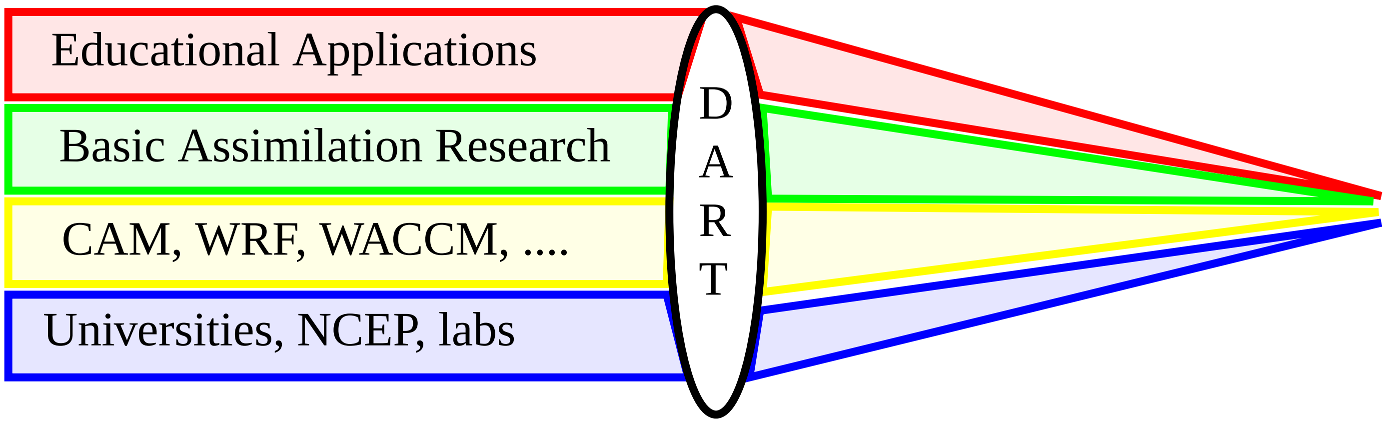
Many linear and non-linear forward operators for low-order models
U, V, T, Ps, Q, for realistic models
Radar reflectivity, GPS refractivity for realistic models

Can ingest observations from reanalysis or operational BUFR files

Can create synthetic (perfect model) observations for any of these

Tactical Vision: Use DART to Focus Efforts

Tactic used to coordinate modeling/dynamics groups
(CCSM, MM5, WRF)



KEY: DART must not be too model/data specific

Needs to be nearly generic and relatively efficient

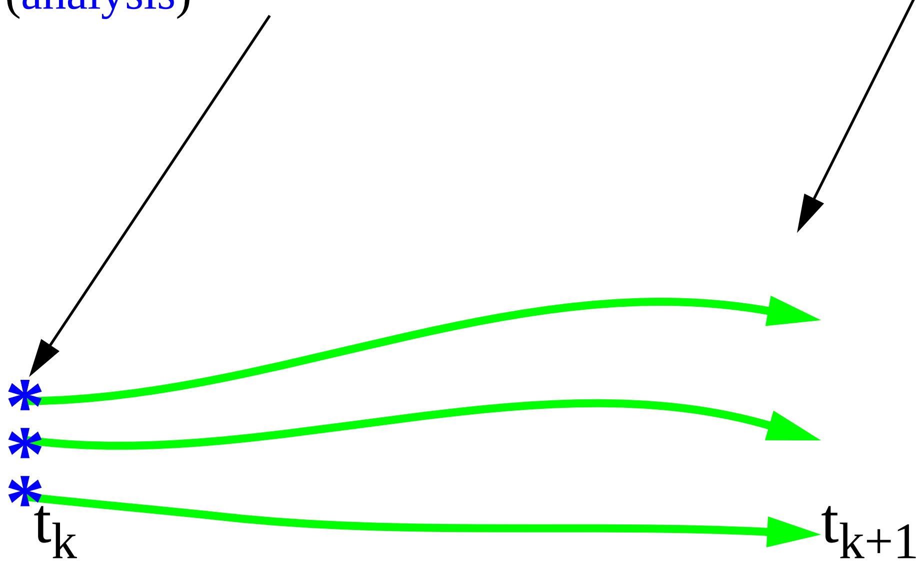
Basic research on ensemble filters in DAI has been addressing this

How an Ensemble Filter Works

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

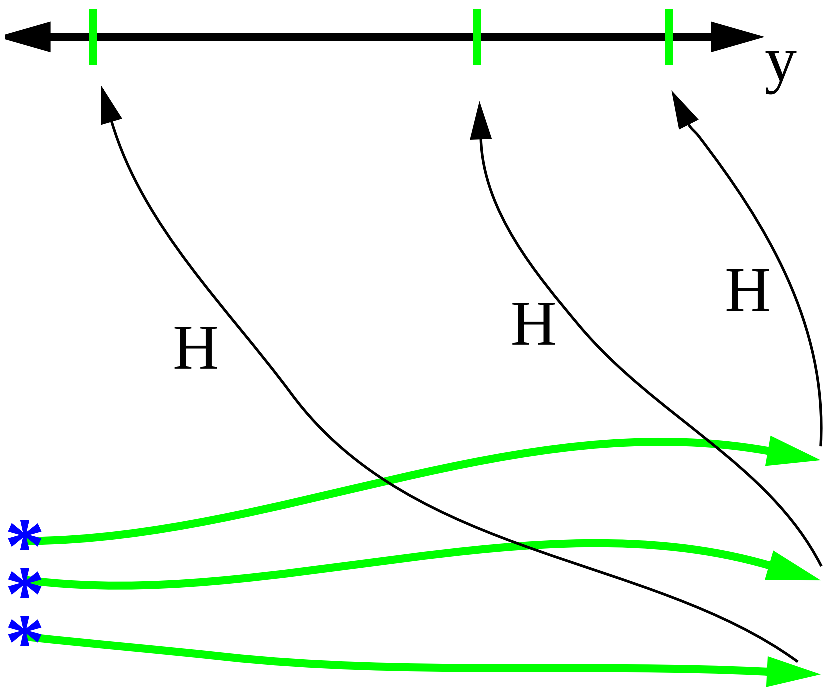
Ensemble state estimate after using previous observation (analysis)

Ensemble state at time of next observation (prior)



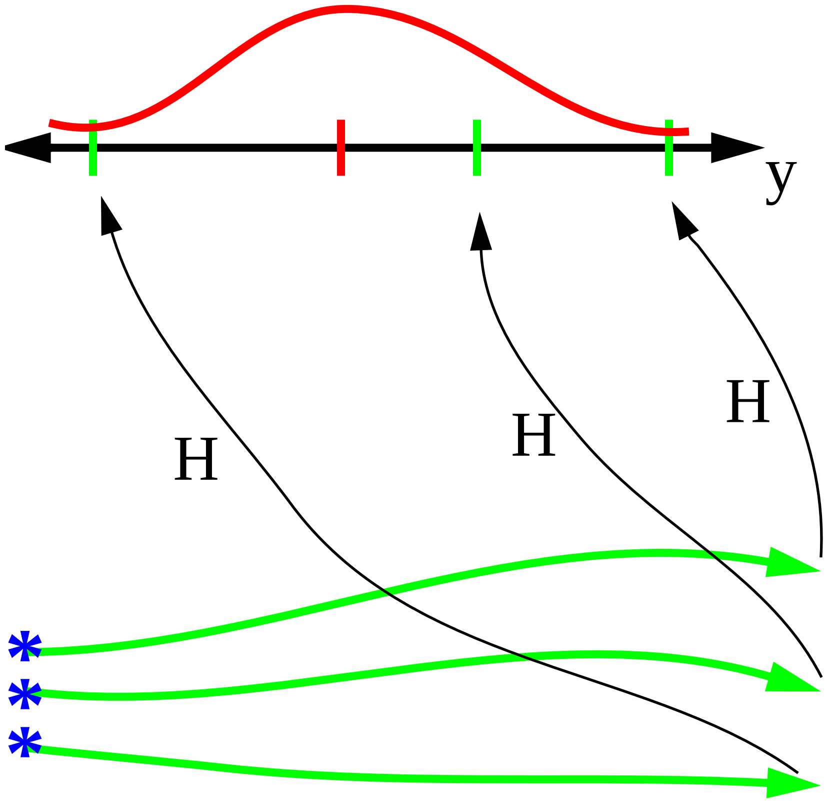
How an Ensemble Filter Works

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



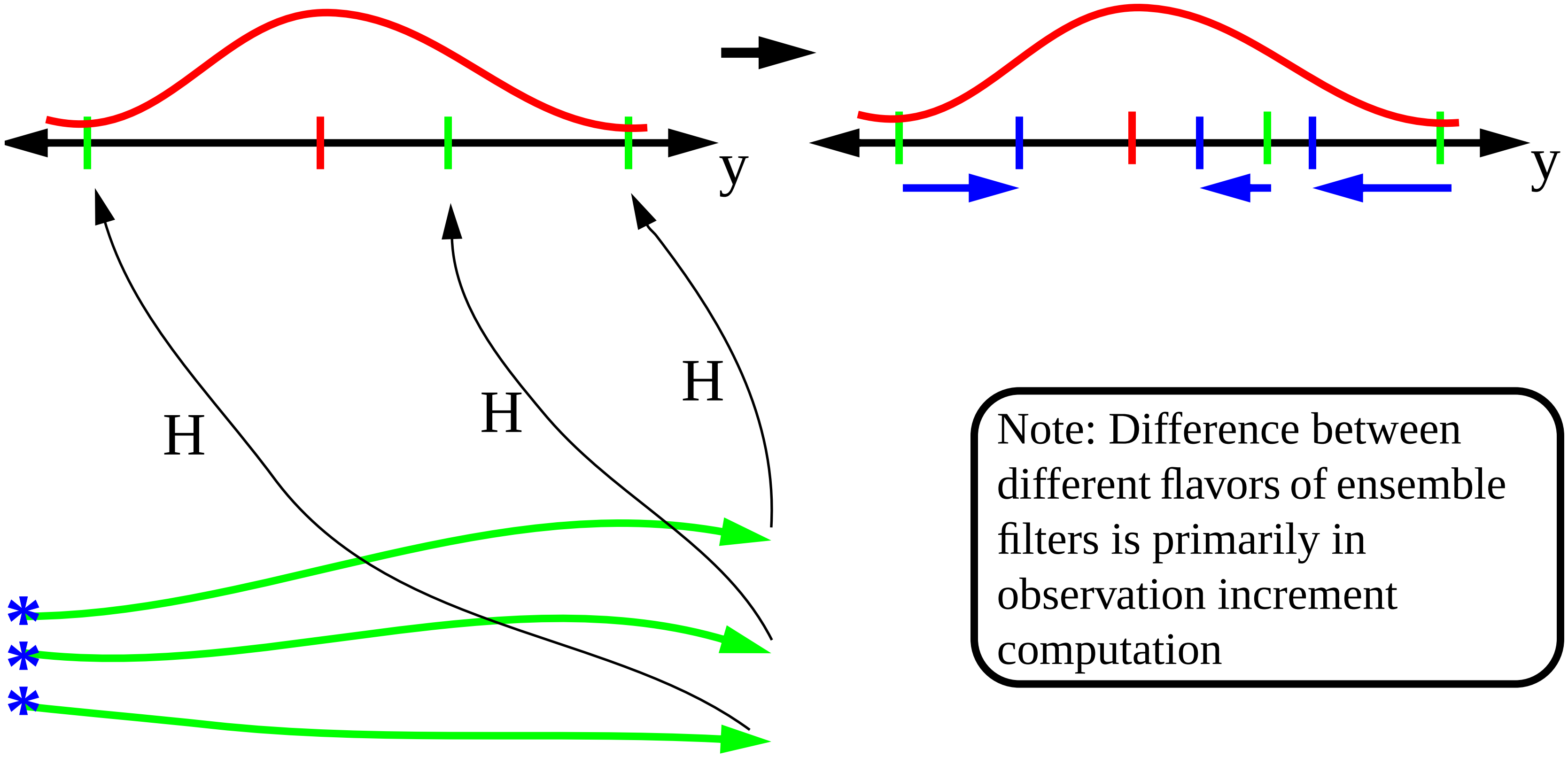
Theory: observations from instruments with uncorrelated errors can be done sequentially.

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



How an Ensemble Filter Works

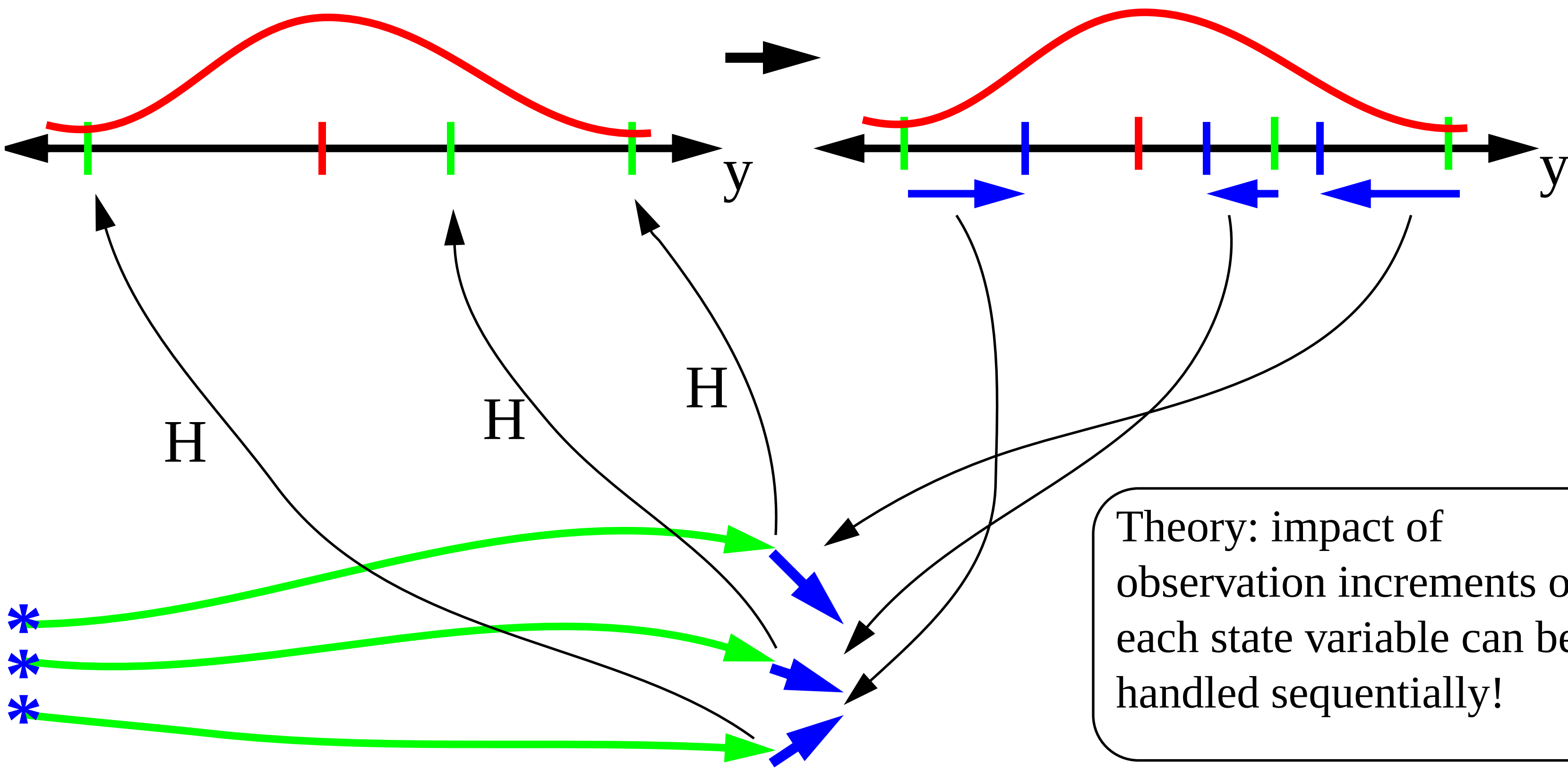
4. Find increment for each prior observation ensemble
(this is a scalar problem for uncorrelated observation errors)



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

How an Ensemble Filter Works

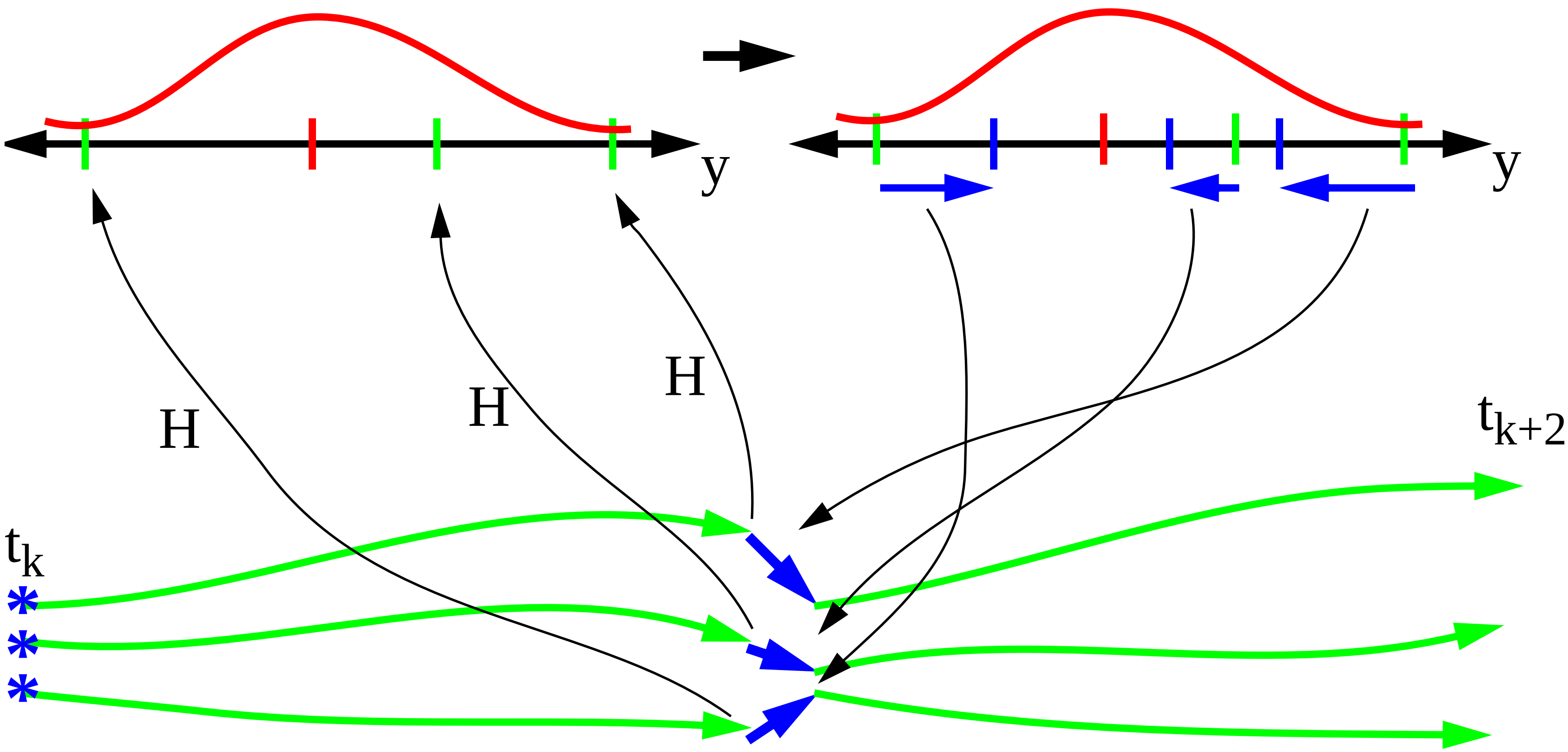
5. Use ensemble samples of y and each state variable to linearly regress observation increments onto state variable increments



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

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. Ensemble Adjustment Kalman Filter (EAKF); deterministic

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

---

Key to Kalman Filters: Product of Gaussians is Gaussian

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

Observation y^o with observational error variance Σ^o

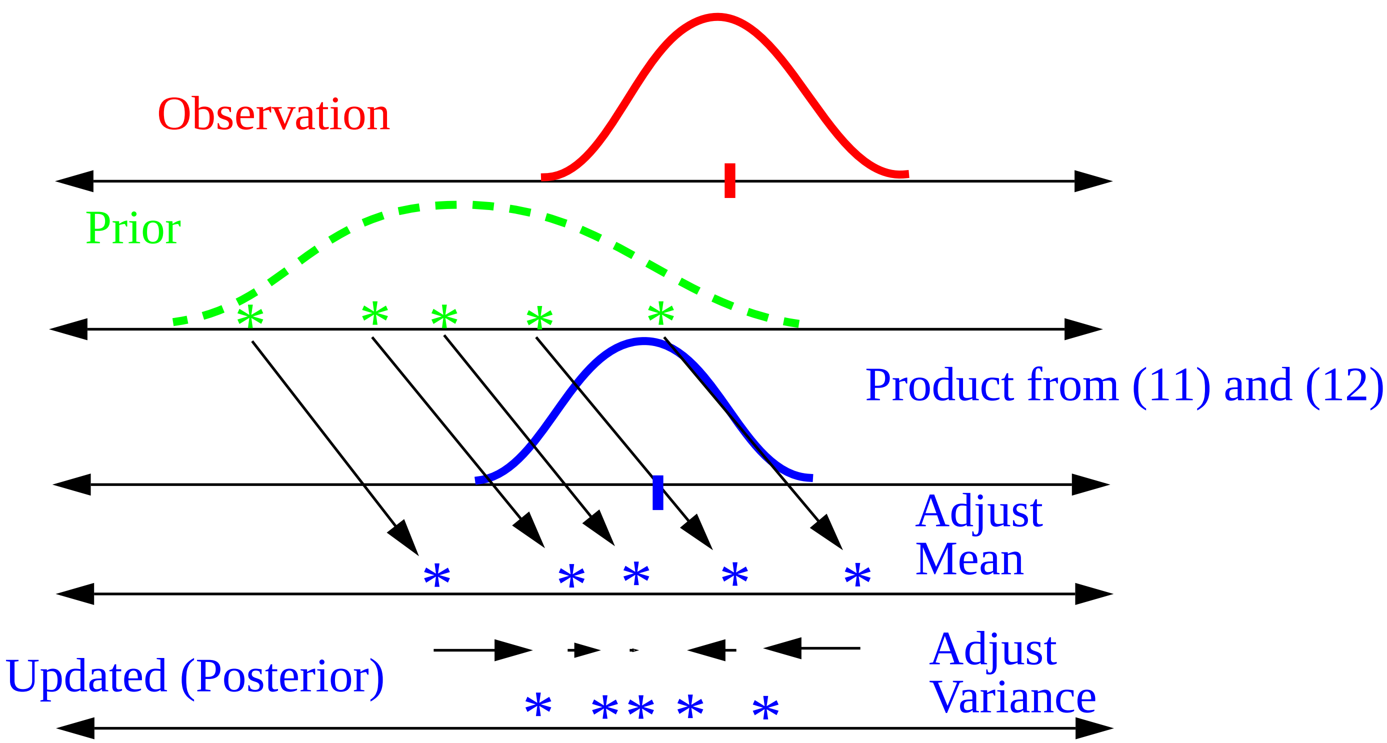
Posterior Variance is: $$\Sigma^u = \left[(\Sigma^p)^{-1} + (\Sigma^o)^{-1}\right]^{-1} \quad (11)$$

and mean is: $$\overline{y^u} = \Sigma^u\left[\overline{y^p}/\Sigma^p + y^o/\Sigma^o\right] \quad (12)$$

Details of Step 4: Ensemble Adjustment Kalman Filter

1. Compute prior sample variance and mean, Σ^p and $\bar{y}^p$
2. Apply (11) to compute updated variance, Σ^u, (12) to compute updated mean, $\bar{y}^u$
4. Adjust prior ensemble of y so that mean and variance are exactly $\bar{y}^u$ and Σ^u

$$y_i^u = \left(y_i^p - \bar{y}^p\right)\sqrt{\Sigma^u / \Sigma^p} + \bar{y}^u, \quad i = 1, \ldots N$$



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

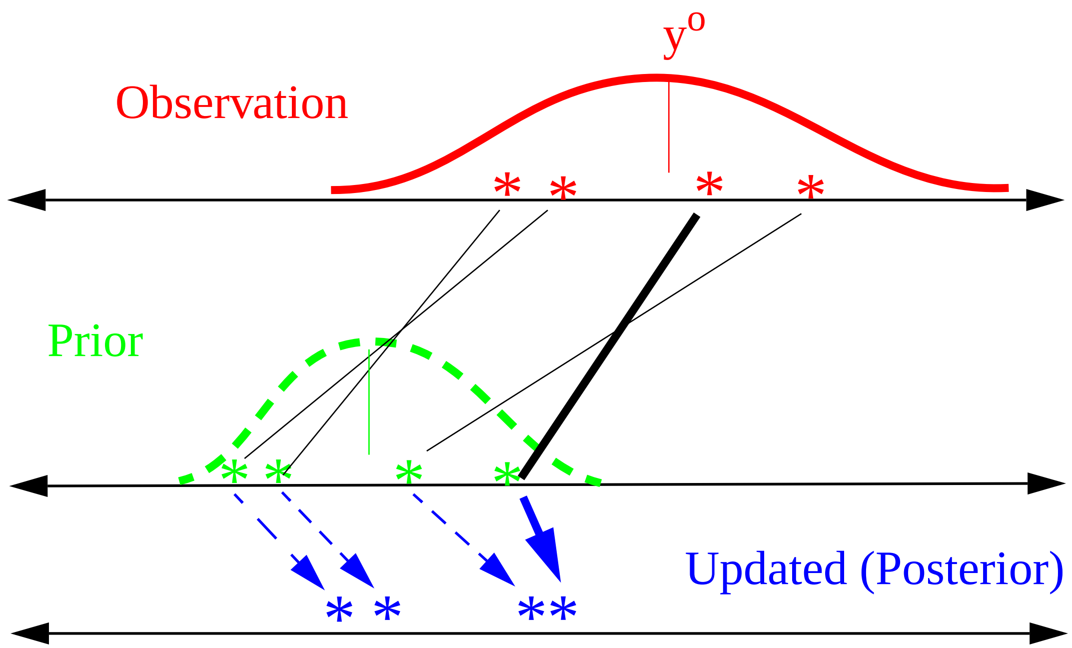
1. Apply (11) once to compute updated variance Σ^u

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, $\overline{y}^u{}_i$

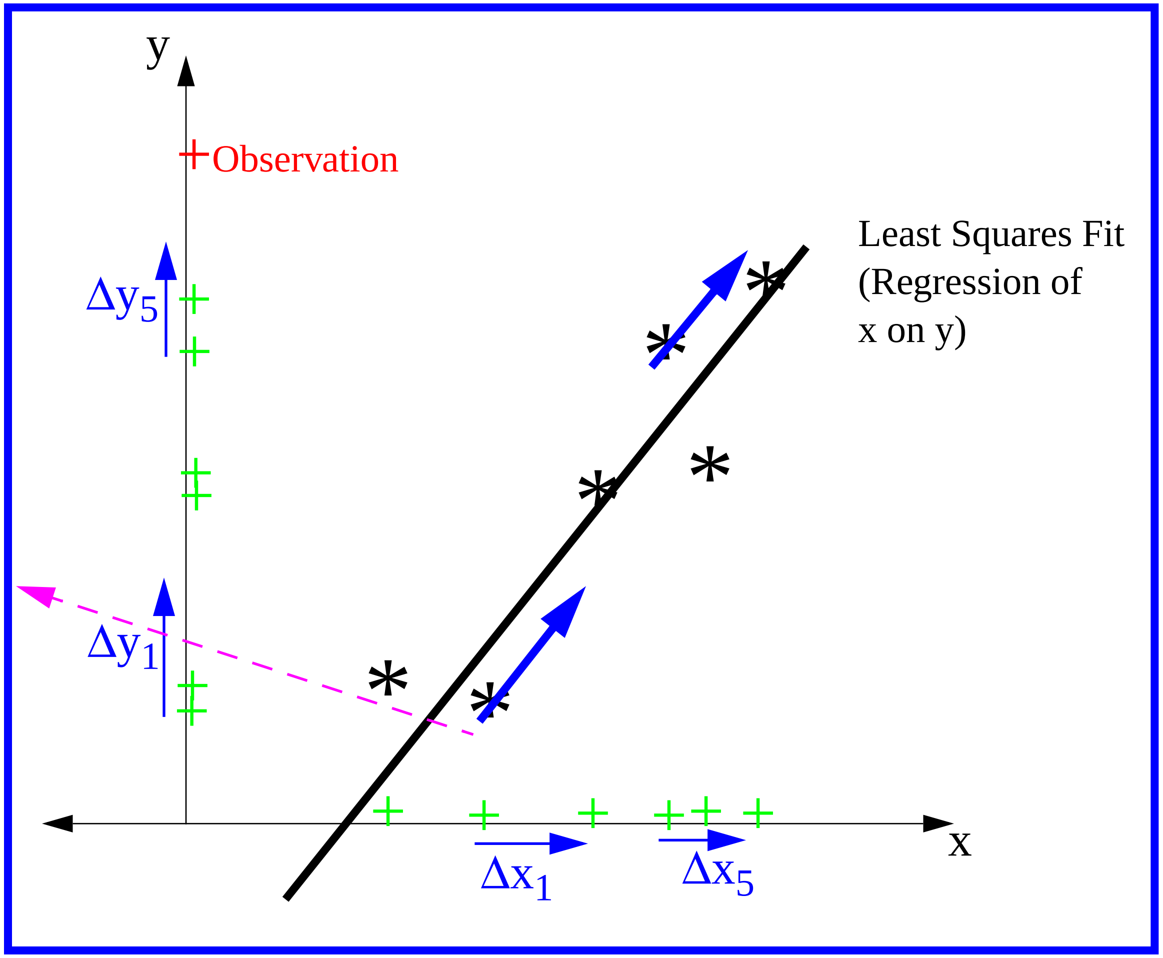
   Replace $\overline{y}^p$ with ith prior ensemble member, $y^p{}_i$

   Replace y^o with ith value from random sample, $y^o{}_i$



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

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



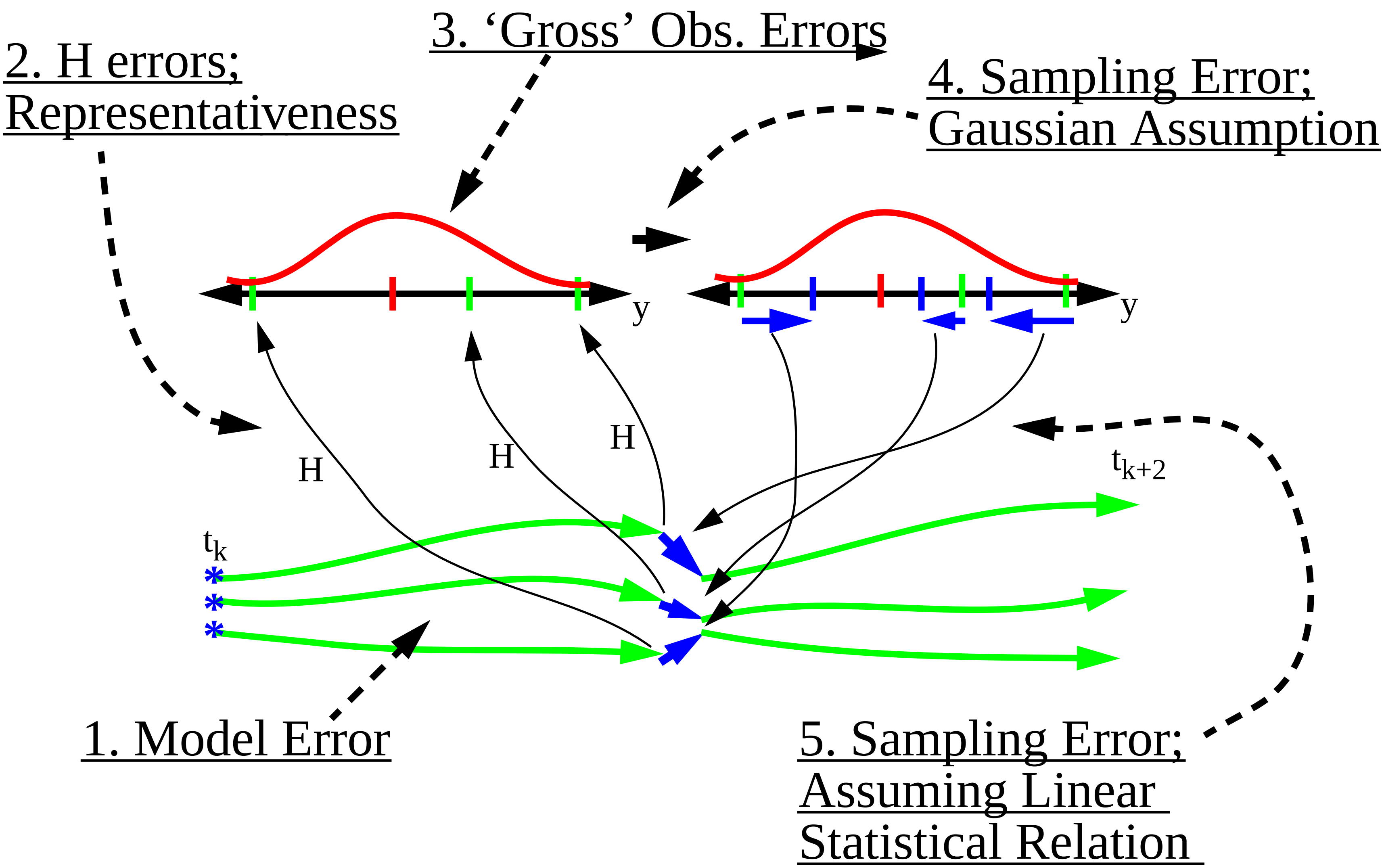
Regression begins with least squares fit to sample, *

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

Some Error Sources in Ensemble Filters



Dealing with Ensemble Filter Error Sources

Serious problem: almost all errors lead to too much certainty
(too little spread in ensemble)

Result is poor performance at best, failure at worst

Algorithms to correct for overconfidence available

1. Hierarchical filters deal with sampling error (localization)
   (see Anderson, 2004b on publications list)

2. Use inconsistency with obs for adaptive spread correction

3. A priori corrections for sampling error from small correlations

Parallel filtering algorithm developed and in DART

Scalable, bitwise reproducing (if required) algorithm

Currently implemented for Linux clusters

An MPI version will be implemented

Allows tolerably good performance on many platforms

Potential for excellent performance with effort (load balancing, ...)

(See Anderson 2004a on publication list)

DAI has many internal NCAR collaborations already in existence

Here are some you'll hear about:

1. Assimilation in Community Atmospheric Model (CAM)	Kevin Raeder
2. Assimilation in WRF	Chris Snyder
3. Use of GPS occultation observations	Hui Liu
4. Assimilation of Doppler radar observations	Alain Caya
5. Advanced statistics for ensemble assimilation	Doug Nychka
6. Middle atmosphere assimilation and observations	Tomoko Matsuo
7. Boundary layer assimilation	Josh Hacker
8. Targeting and observing system design	Shree Khare

Some NCAR DA activities don't (yet) collaborate actively with DAI
An example of these:

1. Grid-scale CO_2 flux estimation with 4D-Var	David Baker