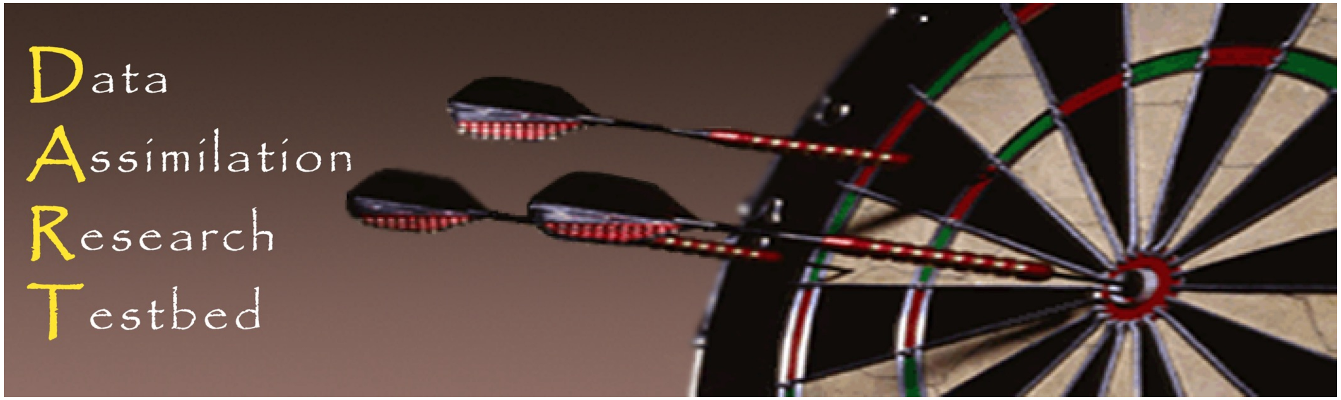
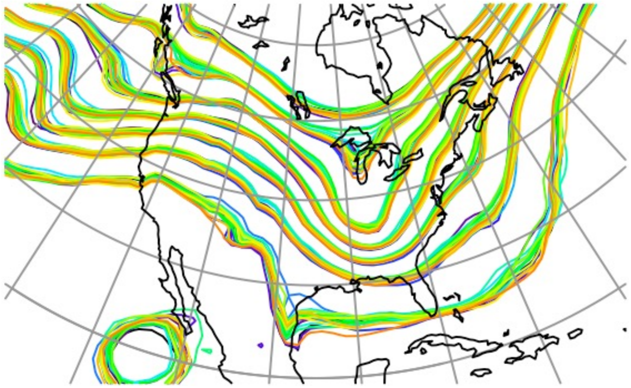


The Data Assimilation Research Testbed: An Intro to Powerful Nonlinear and Non-Gaussian Data Assimilation Tools

Jeff Anderson, NCAR/DAReS



The National Center for Atmospheric Research is sponsored by the National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

NCAR | National Center for
UCAR | Atmospheric Research

Building a Forecast System

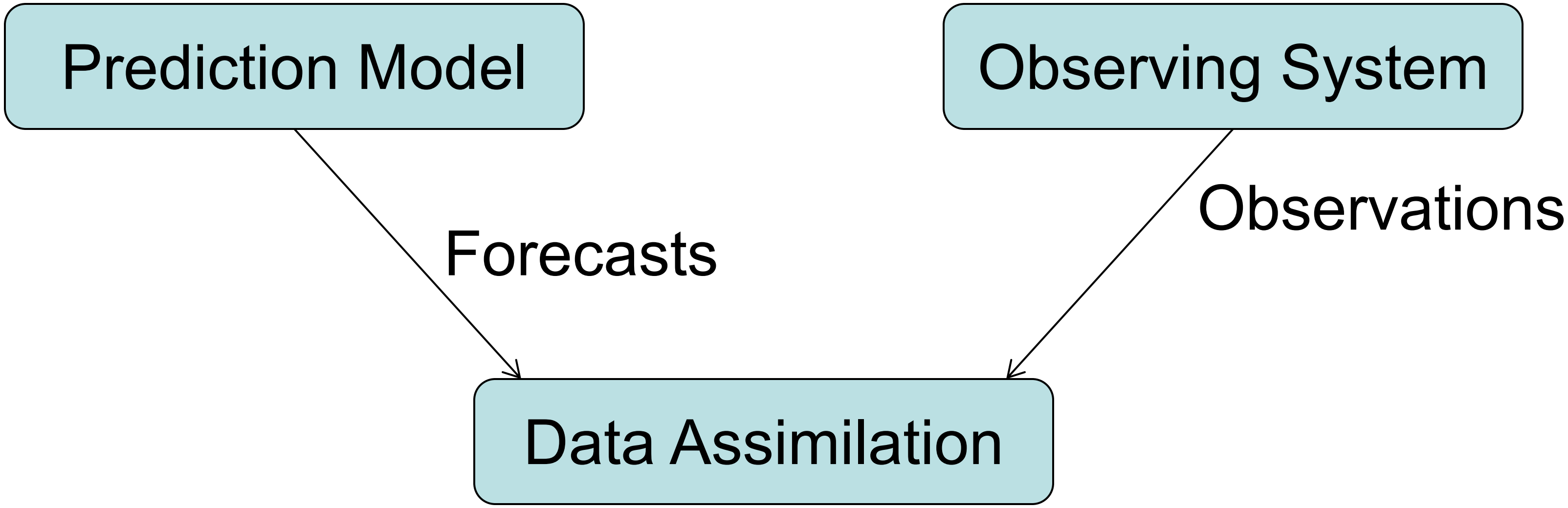
Prediction Model

Building a Forecast System

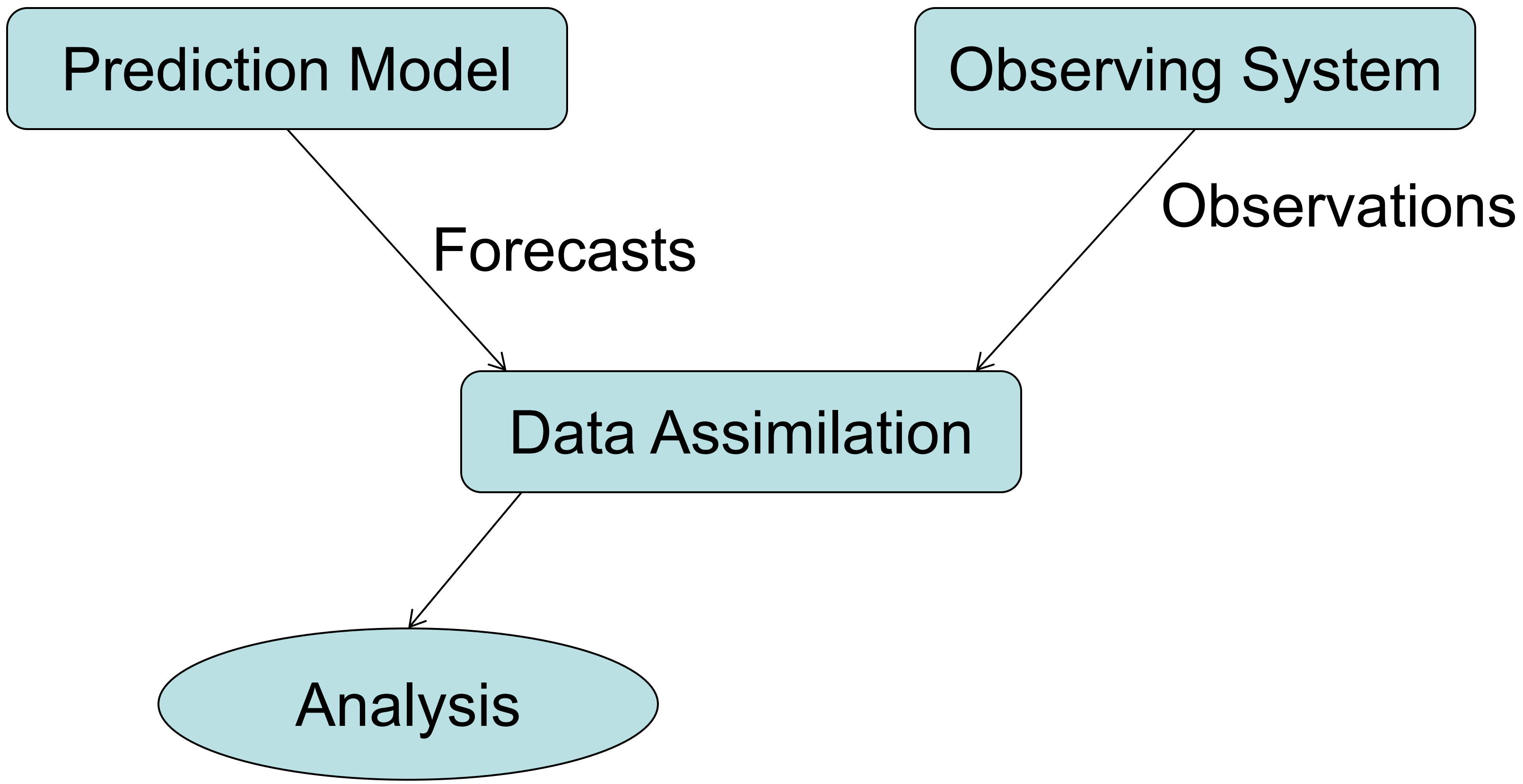
Prediction Model

Observing System

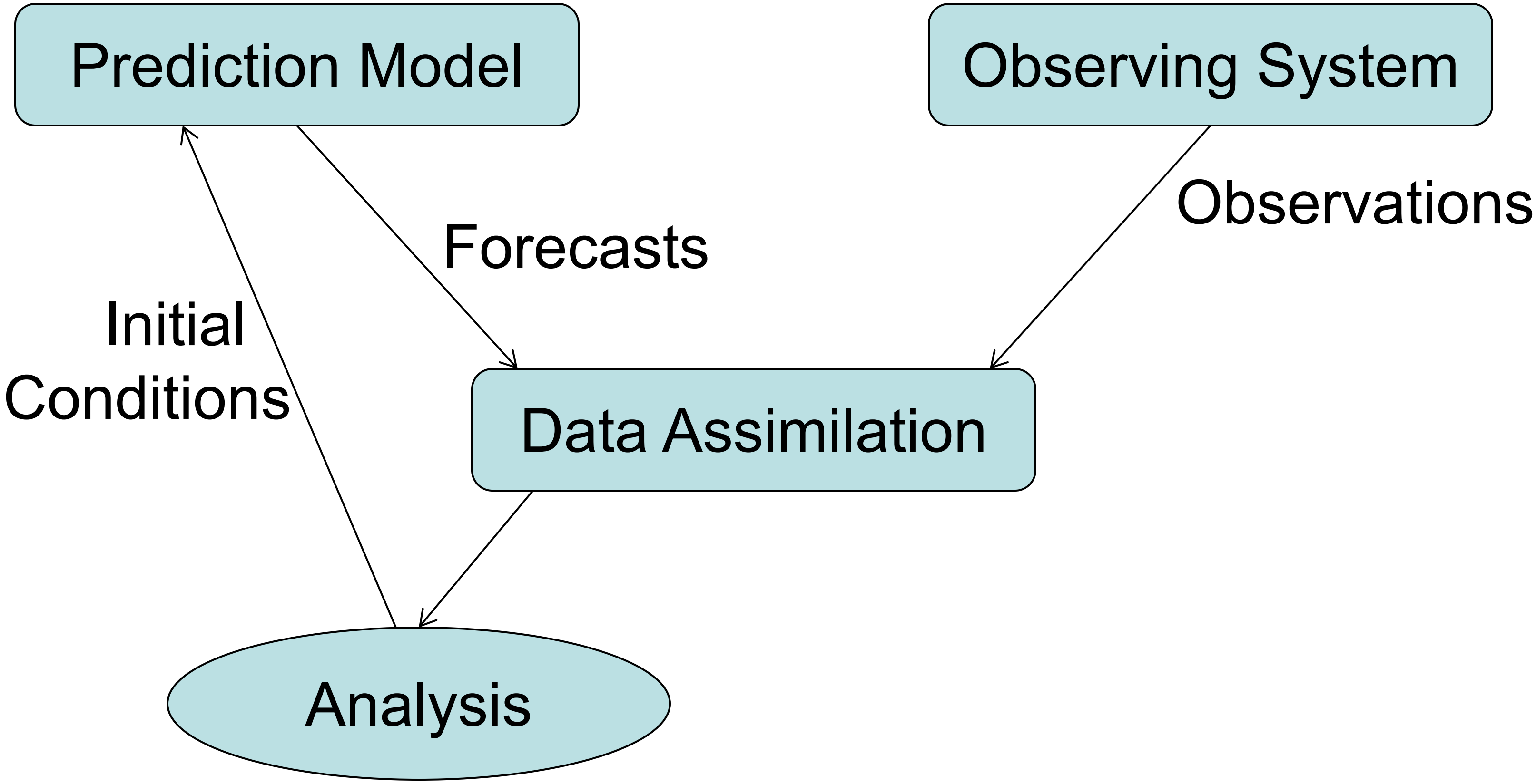
Building a Forecast System



Building a Forecast System

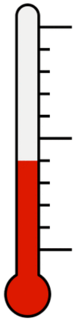


Building a Forecast System

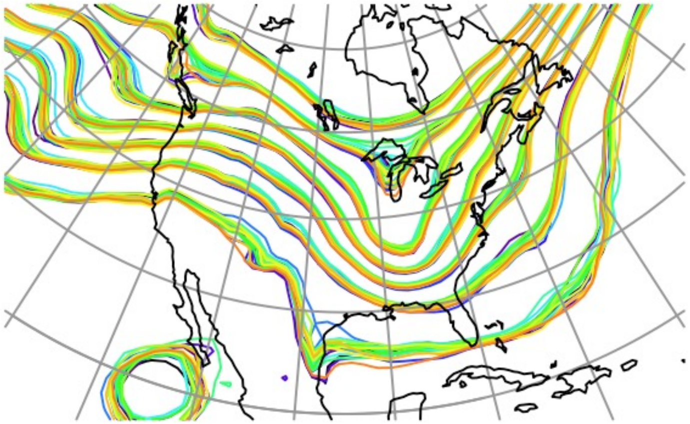


What is Data Assimilation?

Observations combined with a Model forecast…



+



…to produce an analysis
(best possible estimate).

Data Assimilation: What can it do?

- Analyses and forecasts of state variables.
- Smoothing estimates of state variables.
- Estimate model parameters.
- Estimate model errors.
- Estimate observing system errors.
- Quantitatively design observing systems.
- Estimate external forcing.
- Estimate anything correlated with model/observations.

A General Description of the Forecast Problem

A system governed by (stochastic) Difference Equation:

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

Observations at discrete times:

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

Observational error 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:

$$p(x, t \mid Y_t) \quad \text{Analysis} \qquad\qquad p(x, t^+ \mid Y_t) \quad \text{Forecast} \tag{5}$$

A General Description of the Forecast Problem

State between observation times obtained from Difference Equation. Need to update state given new observations:

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

Apply Bayes' rule:

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

Noise is white in time (3), so:

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

Integrate numerator to get normalizing denominator:

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

A General Description of the Forecast Problem

Probability after new observation:

Prior (forecast)

Likelihood

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

Posterior (analysis).

Denominator just normalization.

Methods for Solving the Forecast Problem: Kalman Filter

Assumes:

linear model

Gaussian noise

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

Gaussian state

linear forward operator,

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

Gaussian observation error

Product of Two Gaussians

Product of d-dimensional normals with means μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 is normal.

$$N(\mu_1, \Sigma_1) N(\mu_2, \Sigma_2) = cN(\mu, \Sigma)$$

Product of Two Gaussians

Product of d-dimensional normals with means μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 is normal.

$$N(\mu_1, \Sigma_1) N(\mu_2, \Sigma_2) = c N(\mu, \Sigma)$$

Covariance: $$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$$

Mean: $$\mu = \Sigma(\Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2)$$

Product of Two Gaussians

Product of d-dimensional normals with means μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 is normal.

$$N(\mu_1, \Sigma_1) N(\mu_2, \Sigma_2) = cN(\mu, \Sigma)$$

Covariance: $$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$$

Mean: $$\mu = \Sigma(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2)$$

Weight: $$c = \frac{1}{(2\Pi)^{d/2} \left|\Sigma_1 + \Sigma_2\right|^{1/2}} \exp\left\{-\frac{1}{2}\left[(\mu_2 - \mu_1)^T (\Sigma_1 + \Sigma_2)^{-1} (\mu_2 - \mu_1)\right]\right\}$$

We'll ignore the weight since we immediately normalize products to be PDFs.

The Kalman Filter

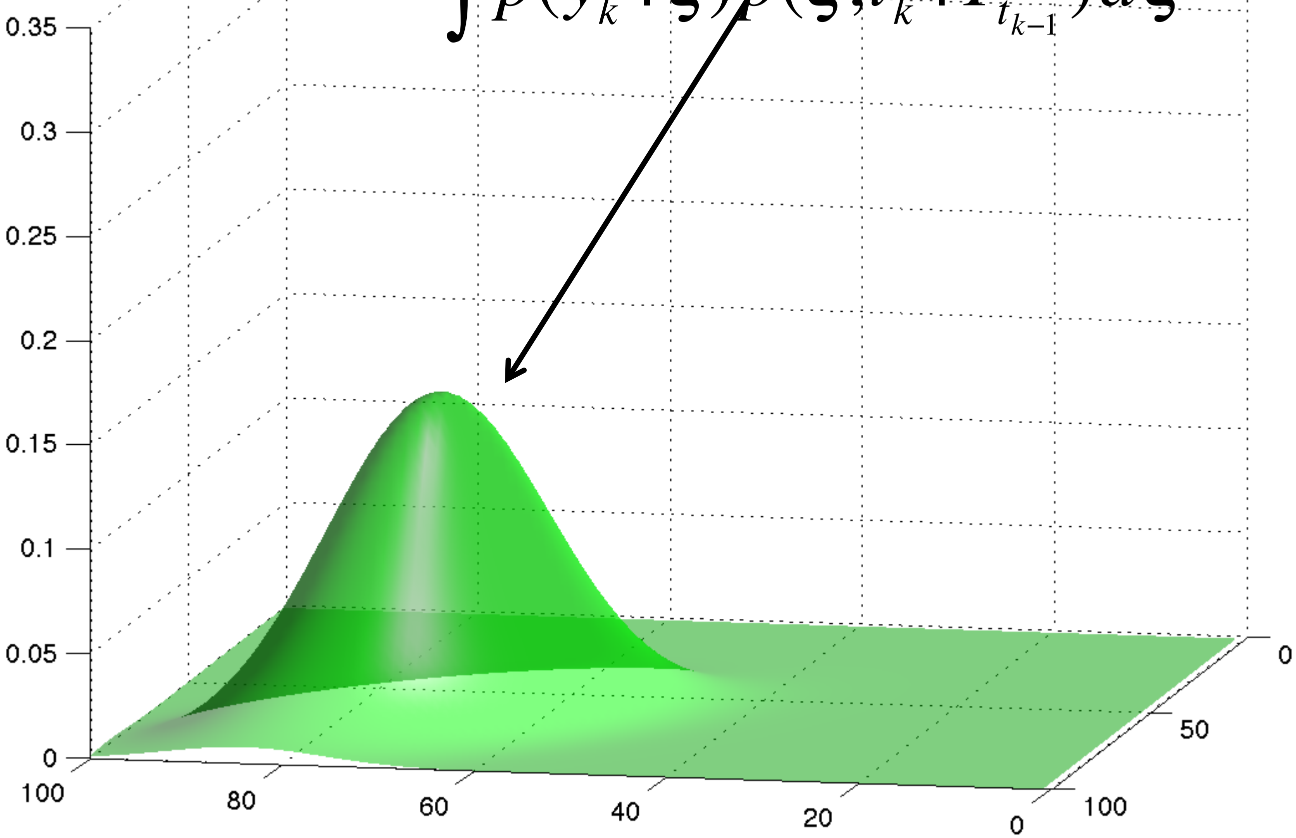
$$p\left(x,t_k \mid Y_{t_k}\right) = \frac{p\left(y_k \mid x\right) p\left(x,t_k \mid Y_{t_{k-1}}\right)}{\int p(y_k \mid \xi) p(\xi,t_k \mid Y_{t_{k-1}}) d\xi} \qquad (10)$$

Numerator is just product of two Gaussians.

Denominator just normalizes posterior to be a PDF.

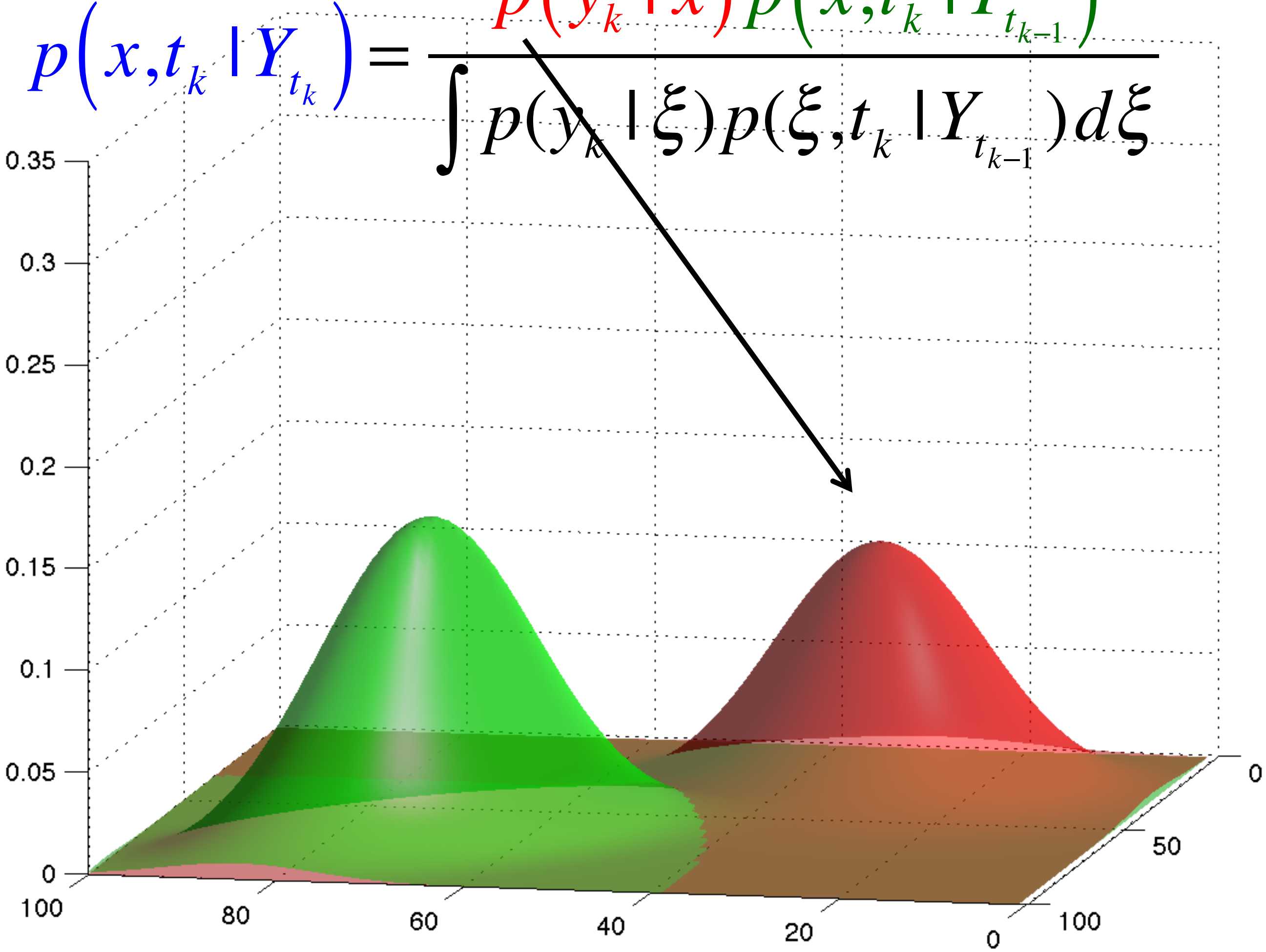
The Kalman Filter

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



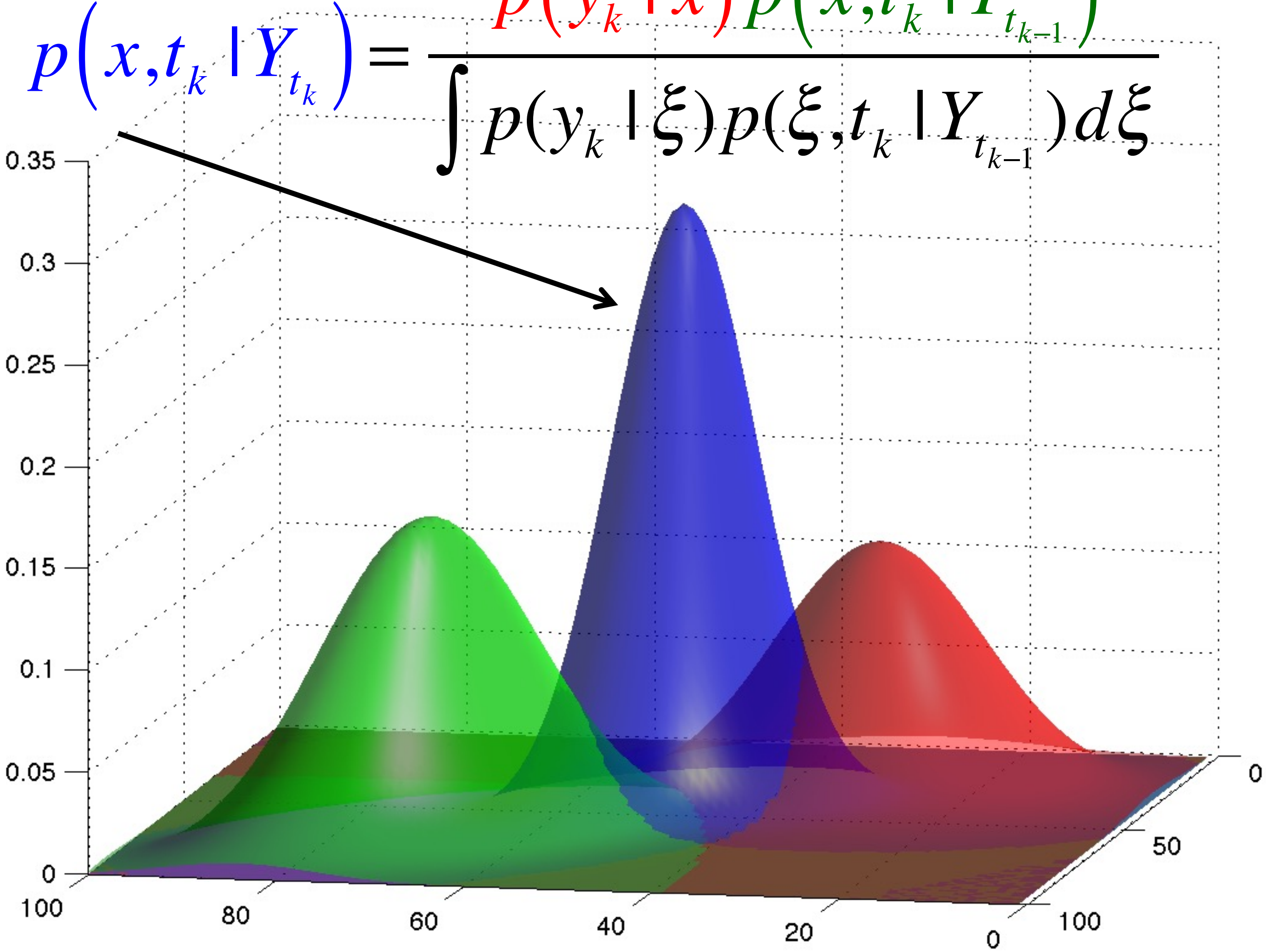
The Kalman Filter

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



The Kalman Filter

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



Kalman Filter: Cost Challenges

Product of d-dimensional normals with means μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 is normal.

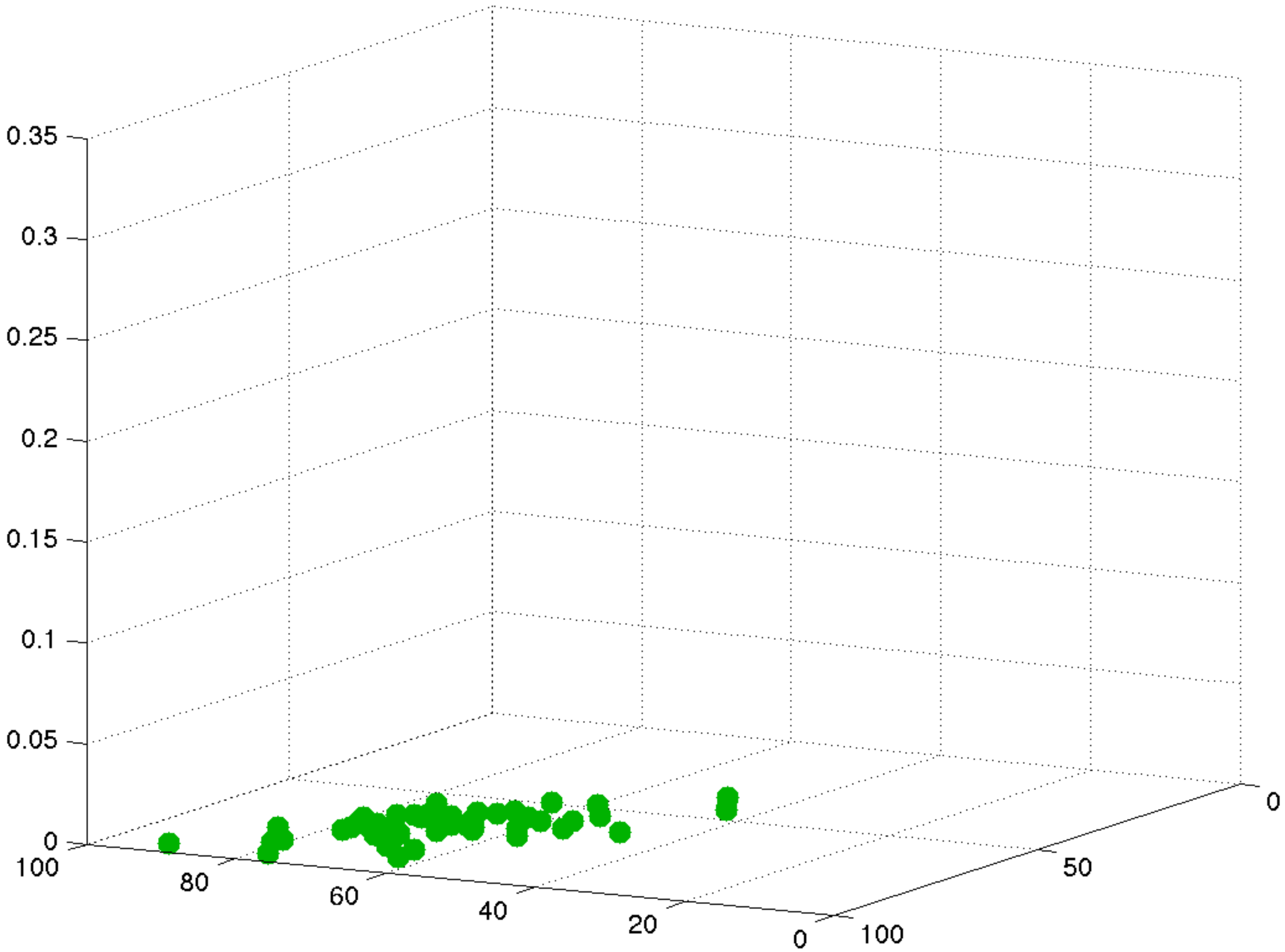
$$N(\mu_1, \Sigma_1) N(\mu_2, \Sigma_2) = cN(\mu, \Sigma)$$

Covariance: $$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$$

Mean: $$u = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} (\Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2)$$

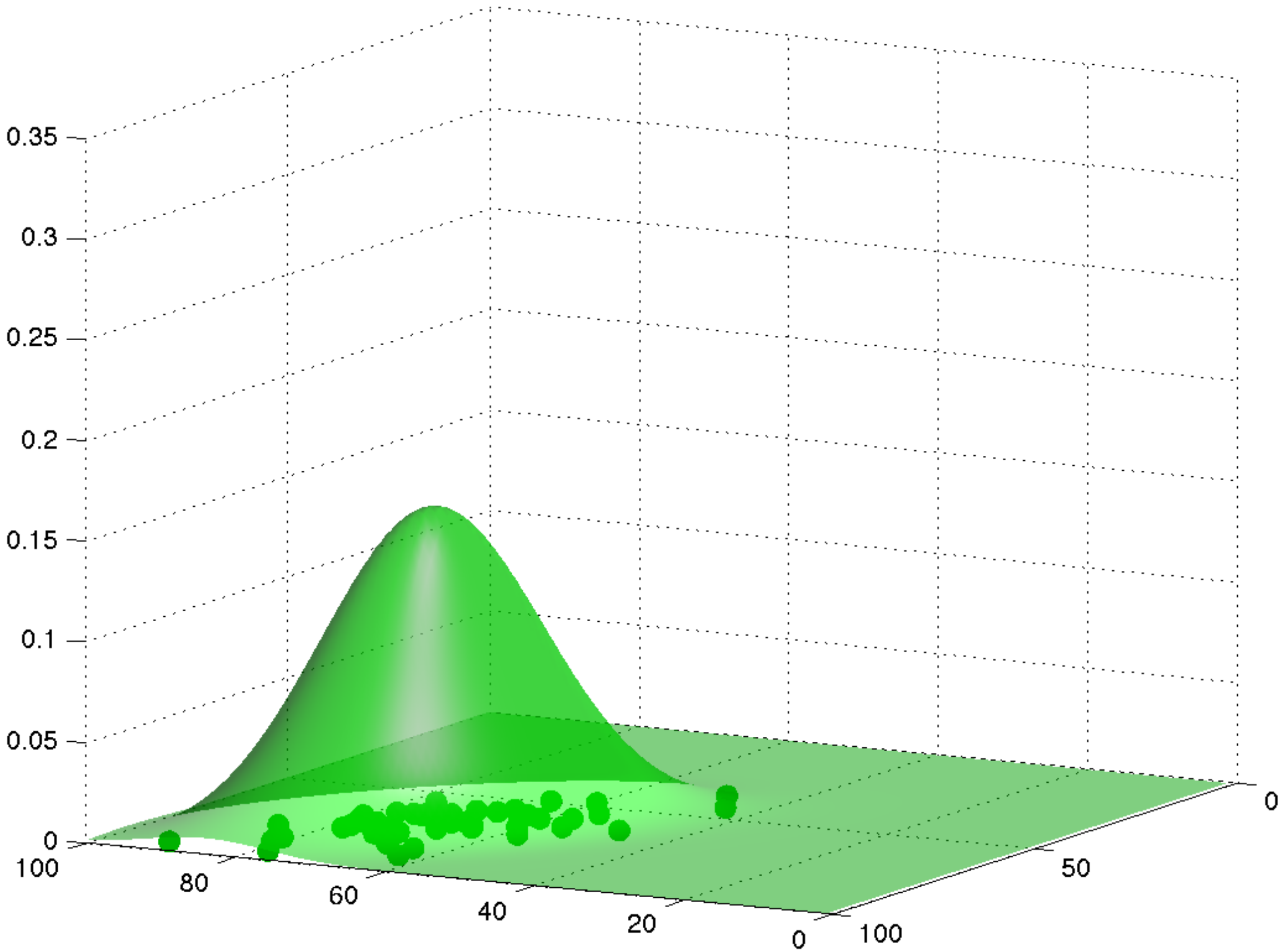
Must store and invert covariance matrices.
Too big to store for large problems.
Too costly to invert, > $O(n^2)$.

1. Start with ensemble of forecasts.



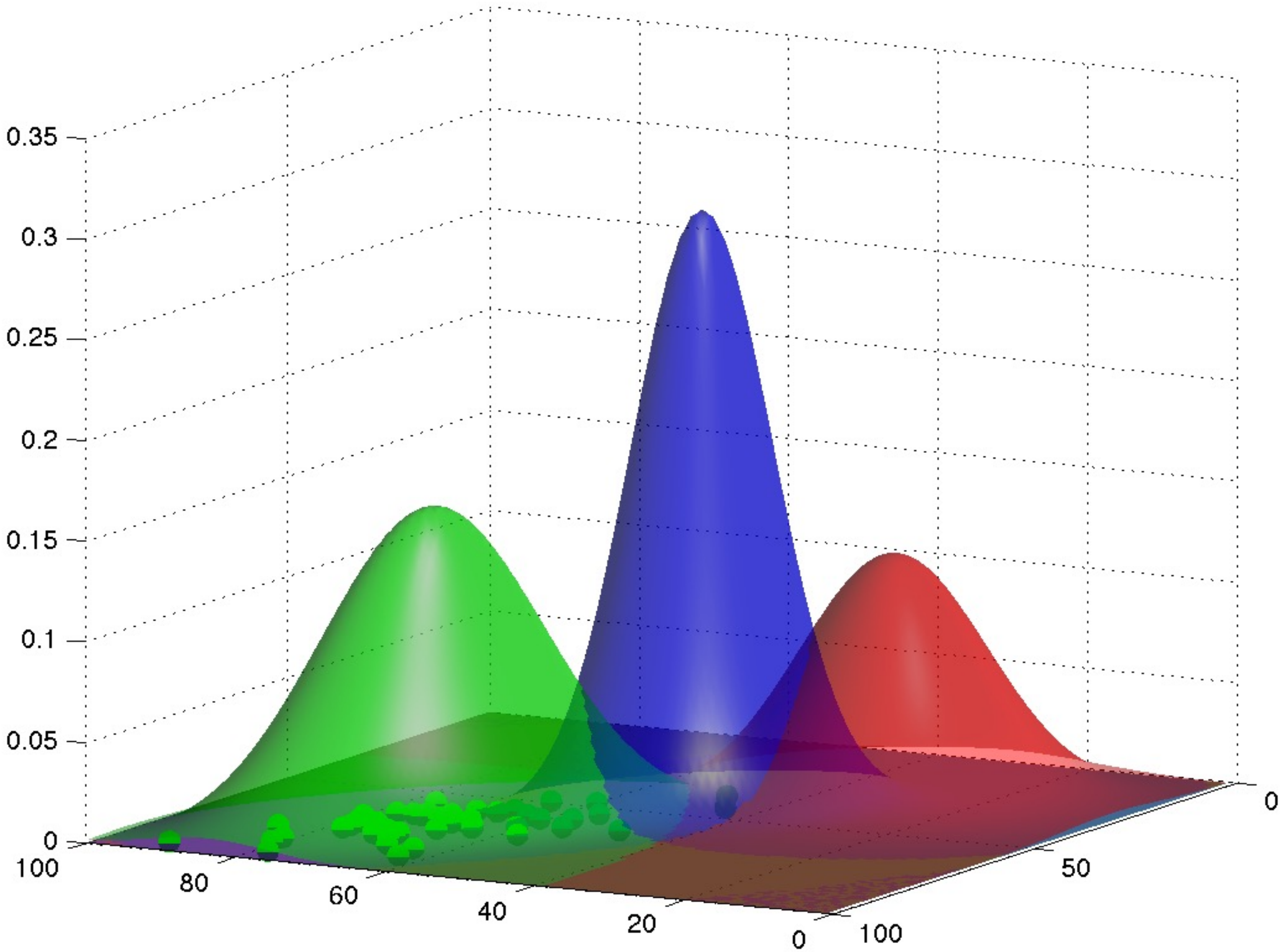
The Ensemble Kalman Filter

2. Fit a normal to ensemble.



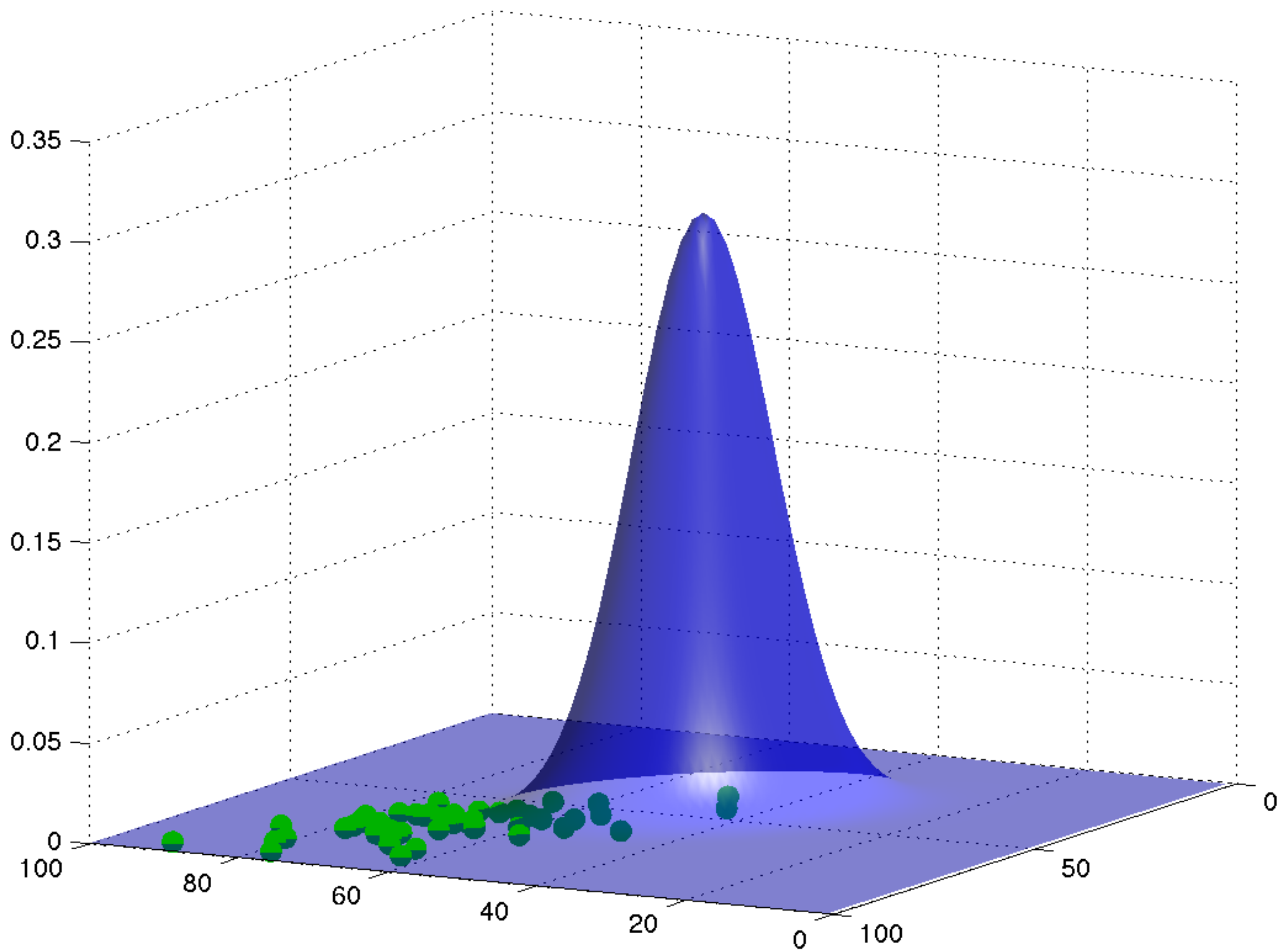
The Ensemble Kalman Filter

3. Do standard Kalman filter.

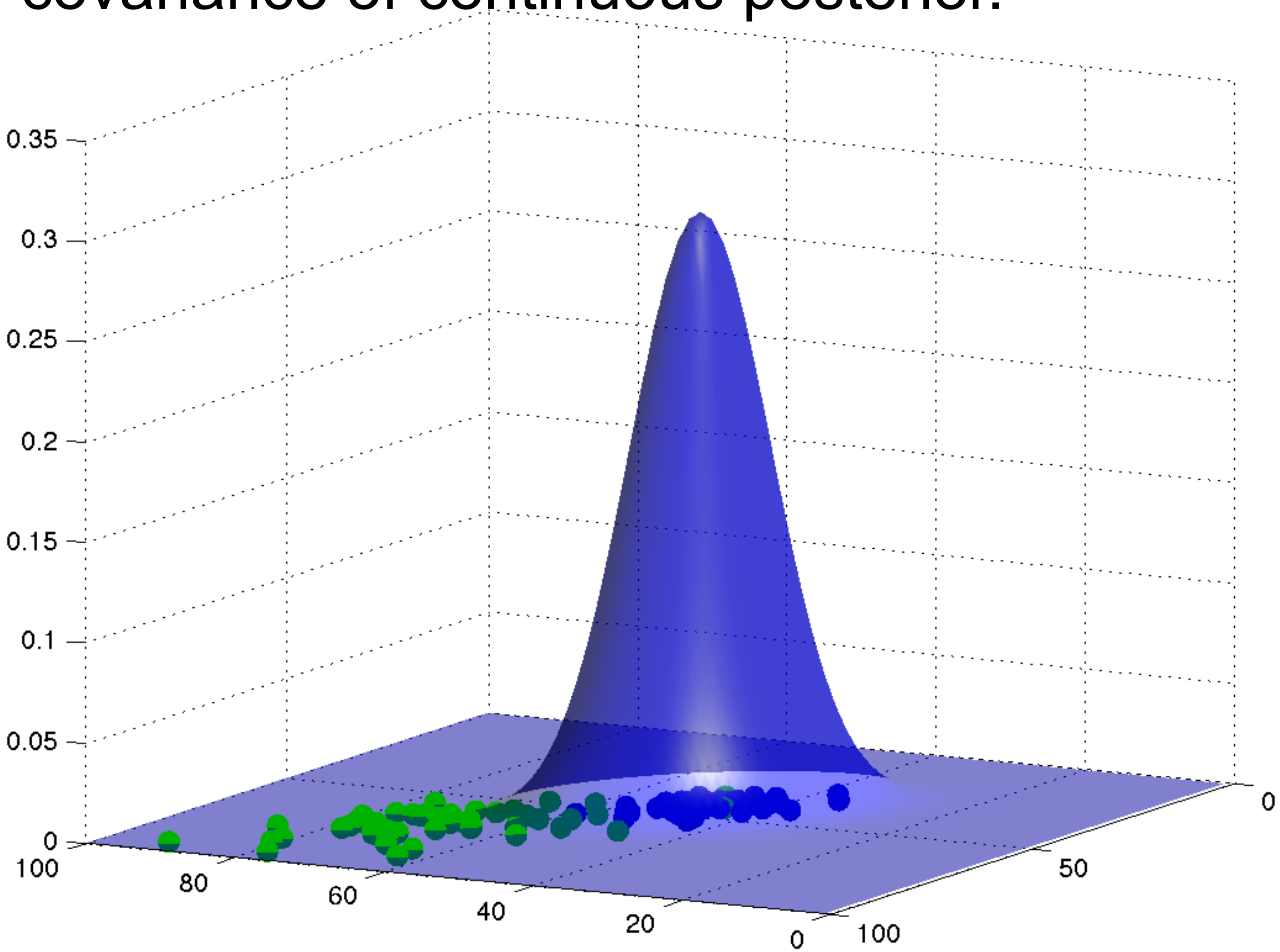


The Ensemble Kalman Filter

Have continuous posterior; need an ensemble.



4. Can create an ensemble with exact sample mean and covariance of continuous posterior.



Schematic of a Sequential Ensemble Filter

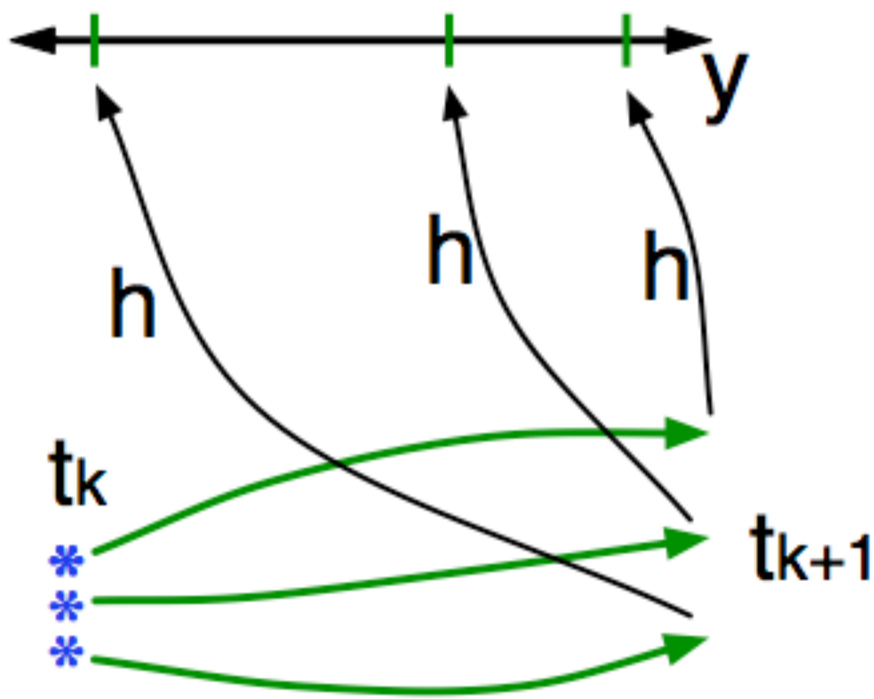
1. Use model to advance ensemble (3 members here) to time at which next observation becomes available.

Ensemble state estimate after using previous observation (analysis)

Ensemble state at time of next observation (prior)

Schematic of a Sequential Ensemble Filter

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

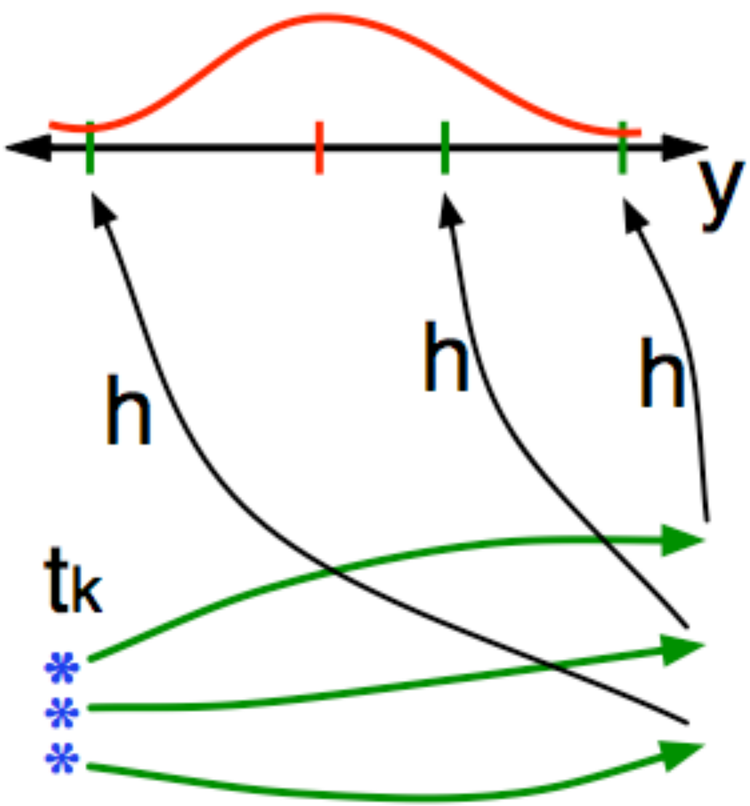


Theory: observations from instruments with uncorrelated errors can be done sequentially.

Can think about single observation without (too much) loss of generality.

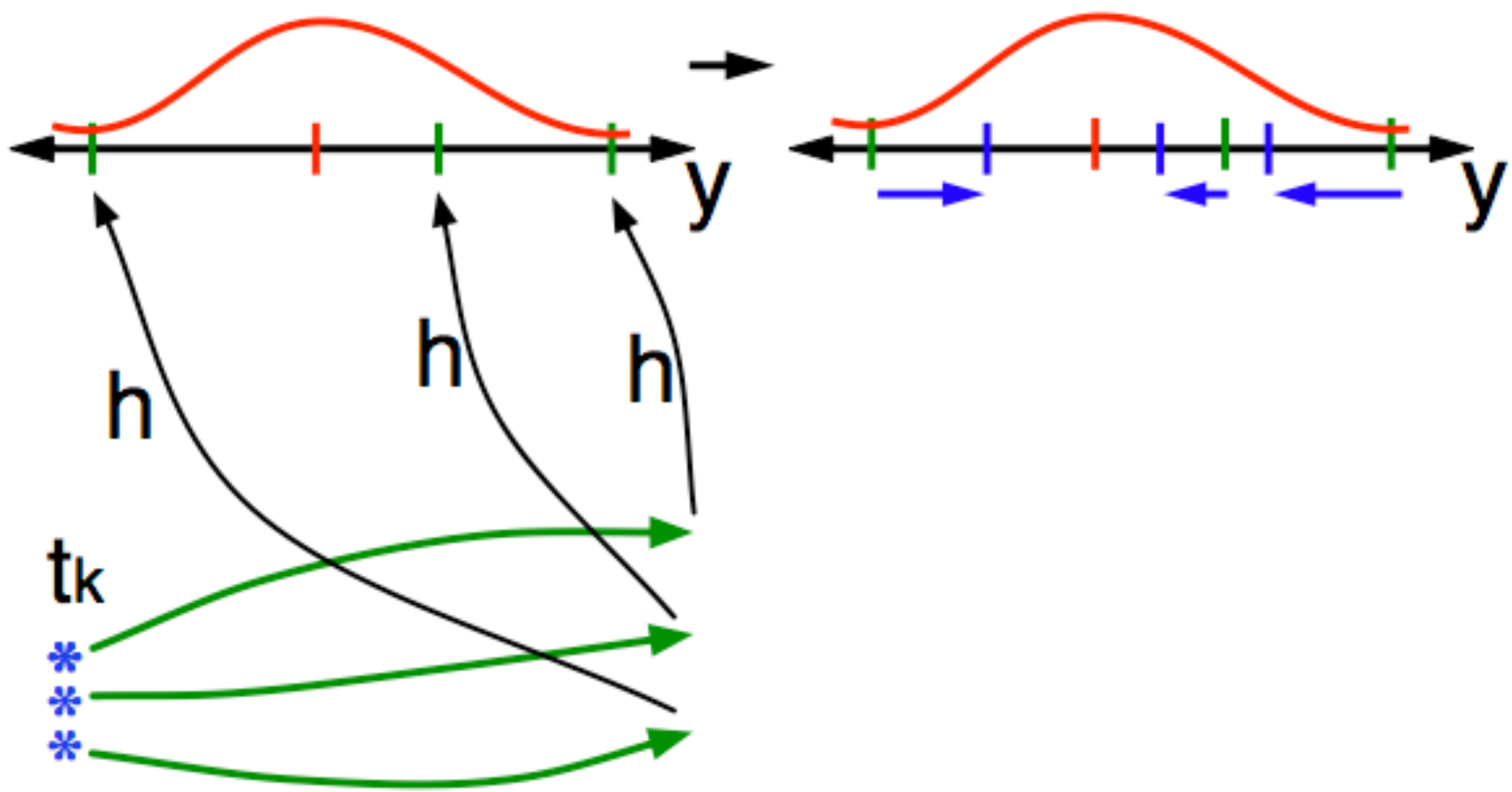
Schematic of a Sequential Ensemble Filter

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

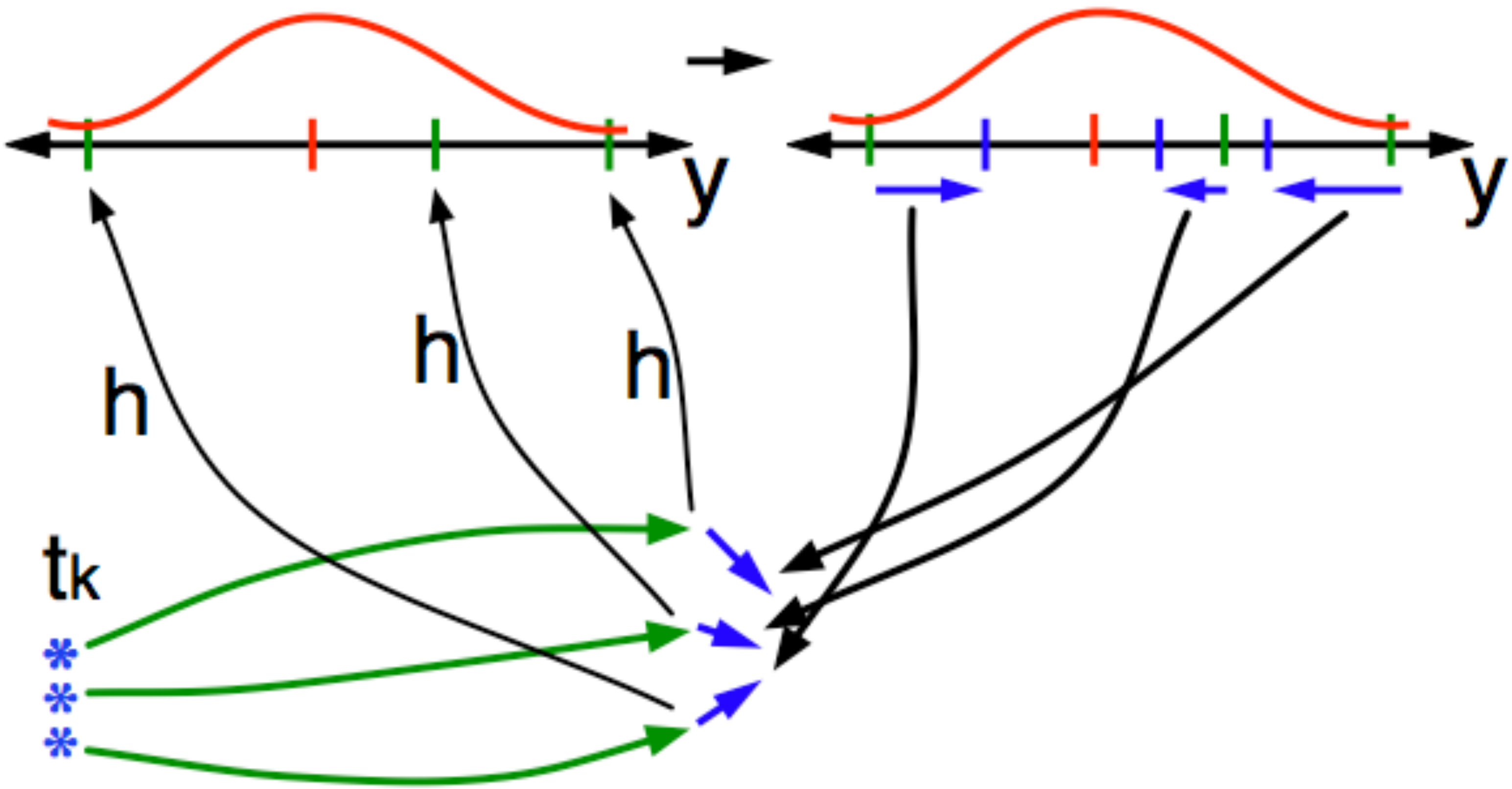


Schematic of a Sequential Ensemble Filter

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

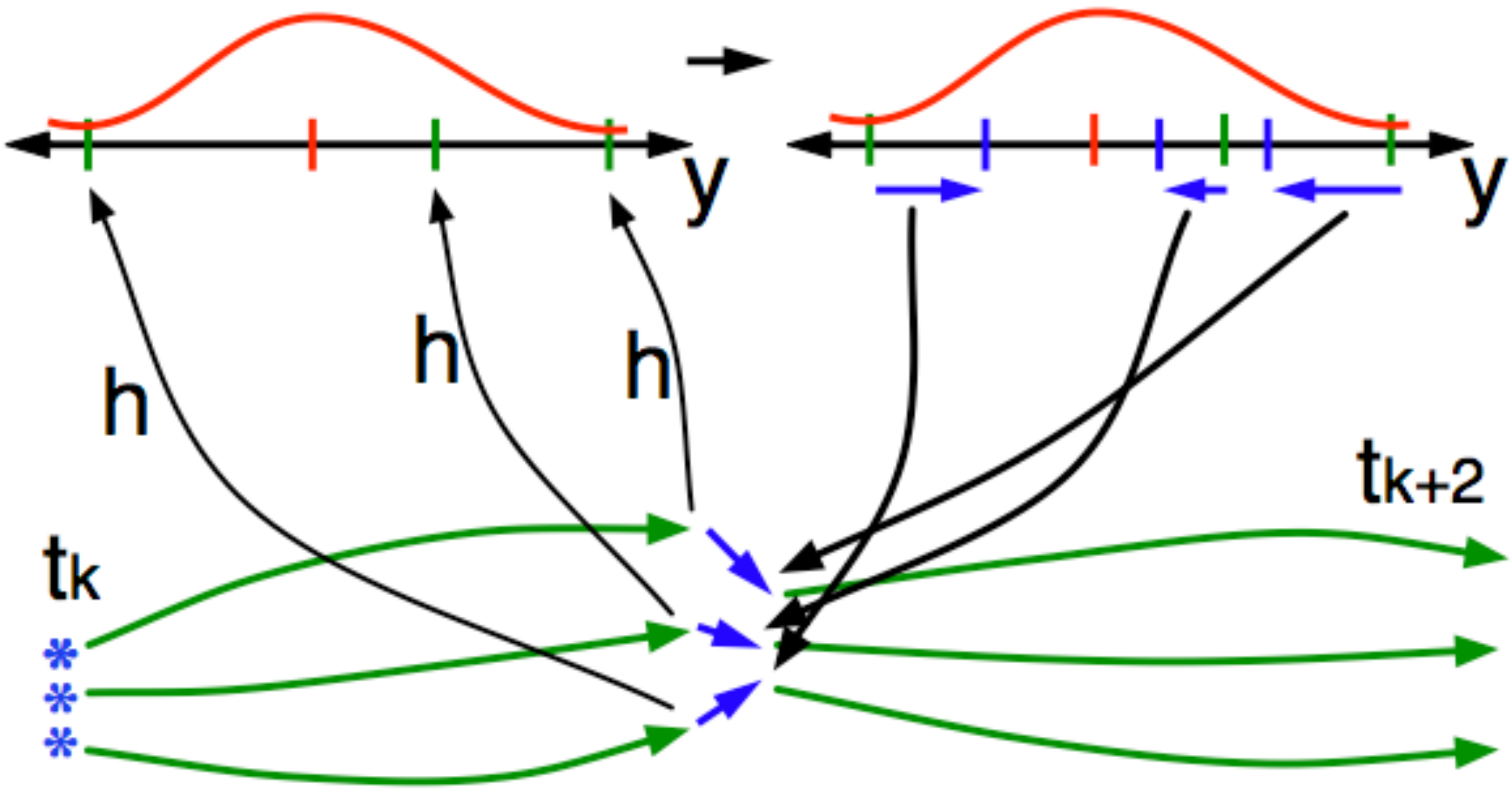


5. Use ensemble samples of y and each state variable to linearly regress observation increments onto state variable increments.

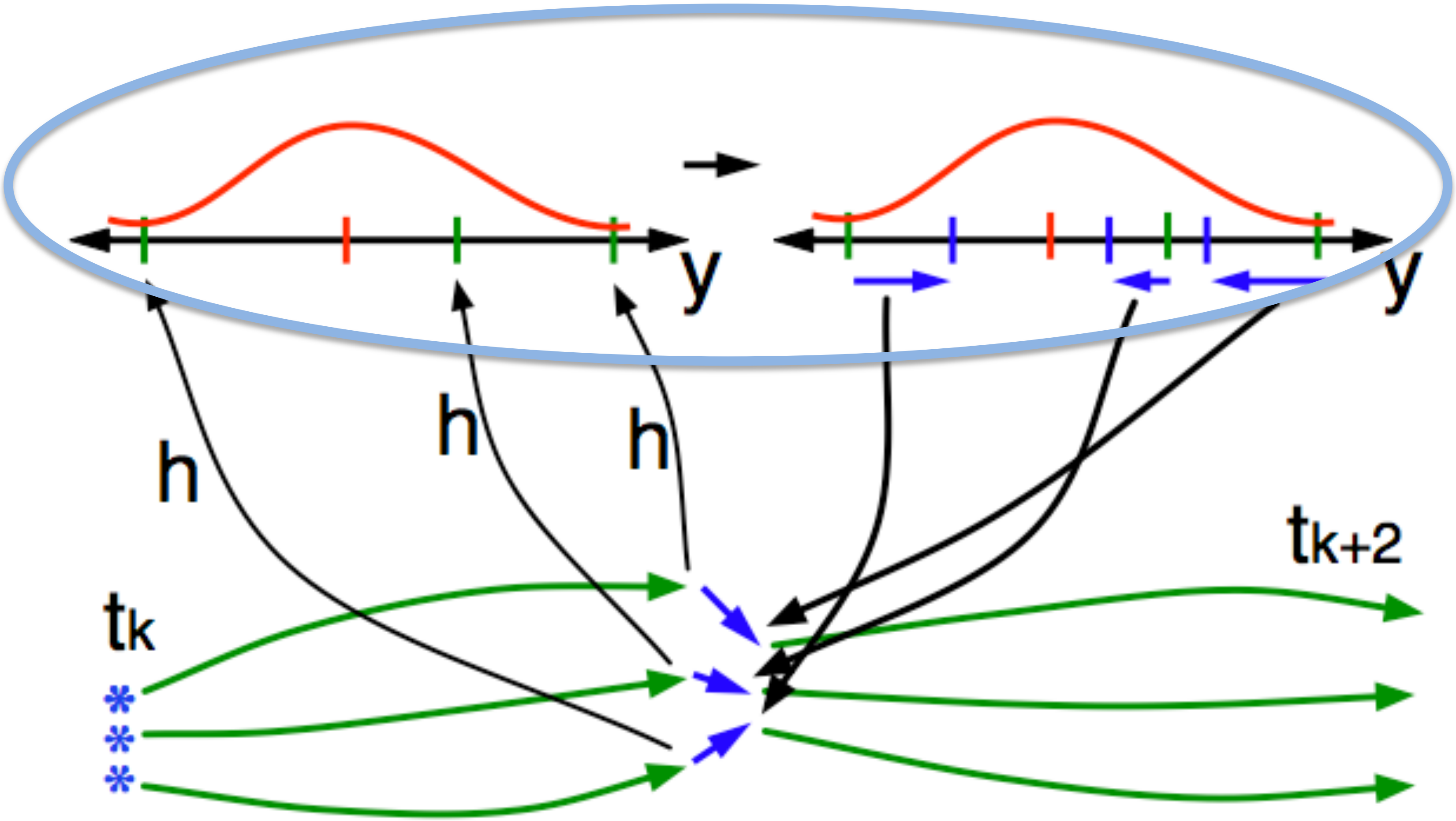


Schematic of a Sequential Ensemble Filter

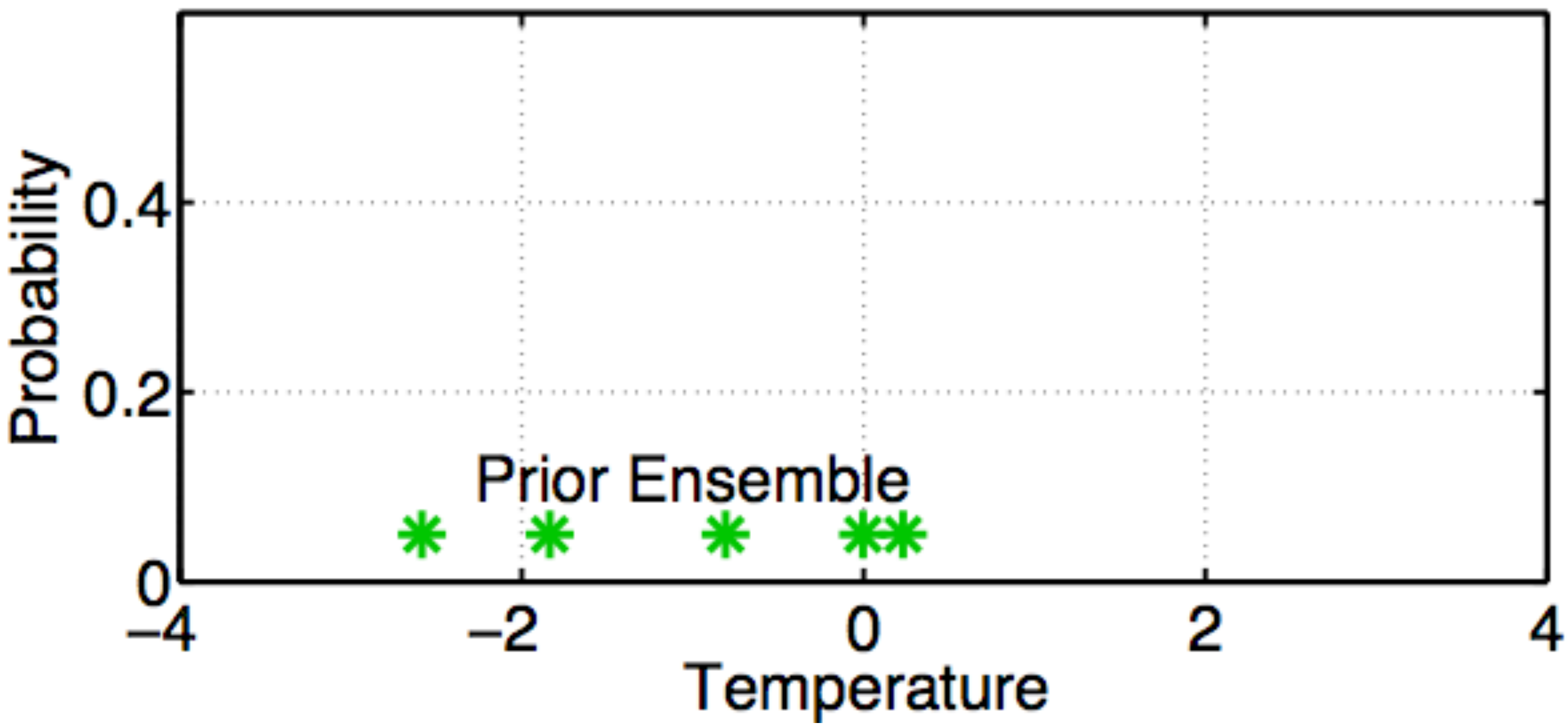
6. When all ensemble members for each state variable are updated, integrate to time of next observation …



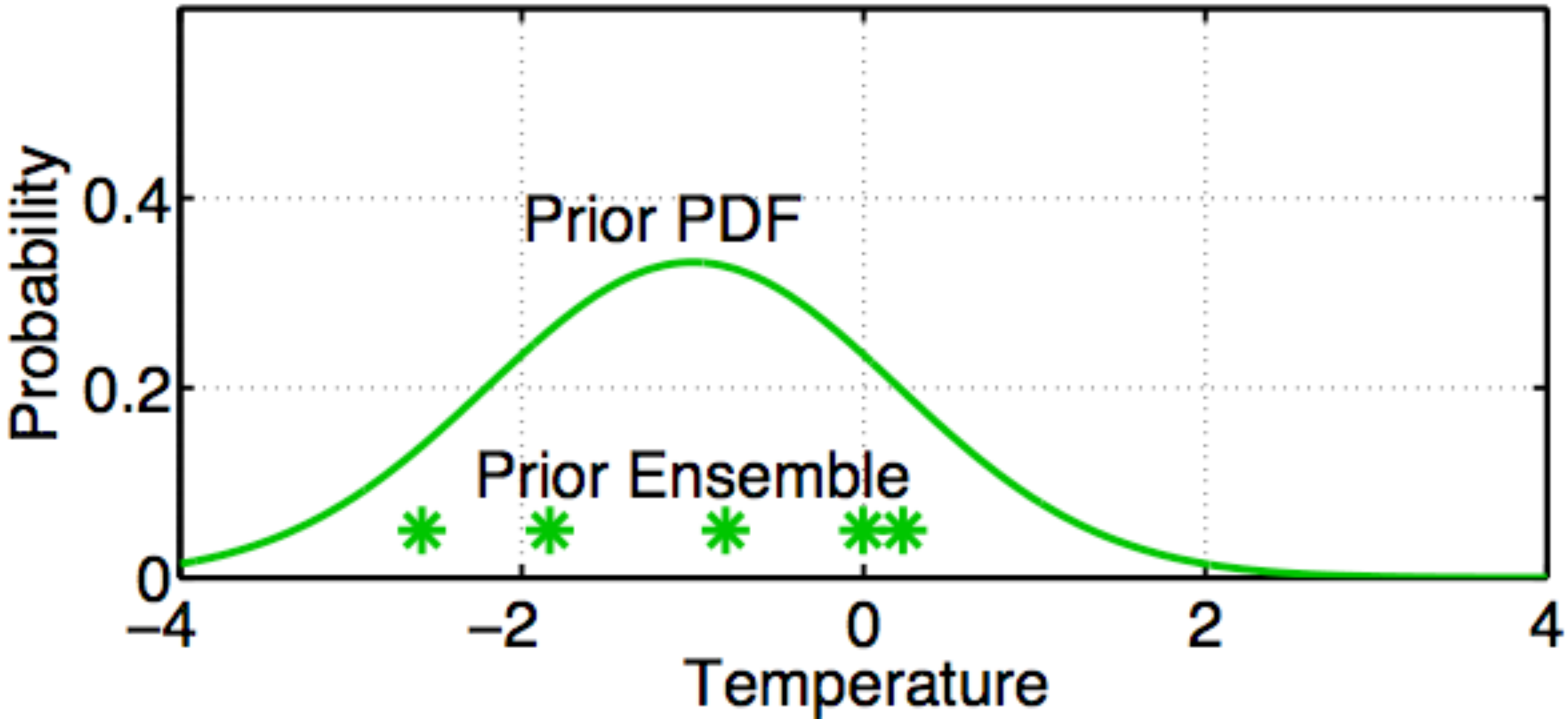
Ensemble Kalman Filter Step 1: Observation Increments



Ensemble Kalman Filter Step 1: Observation Increments

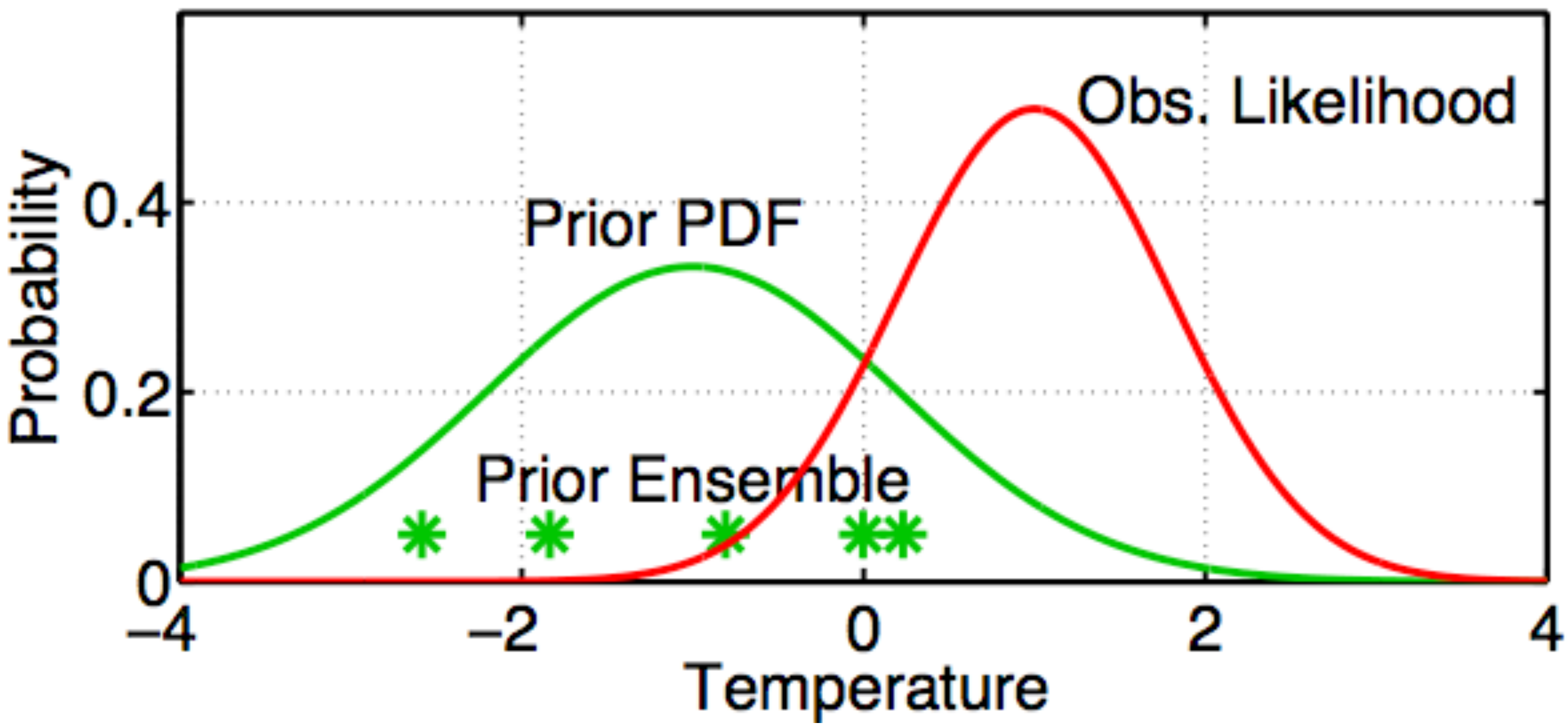


Ensemble Kalman Filter Step 1: Observation Increments



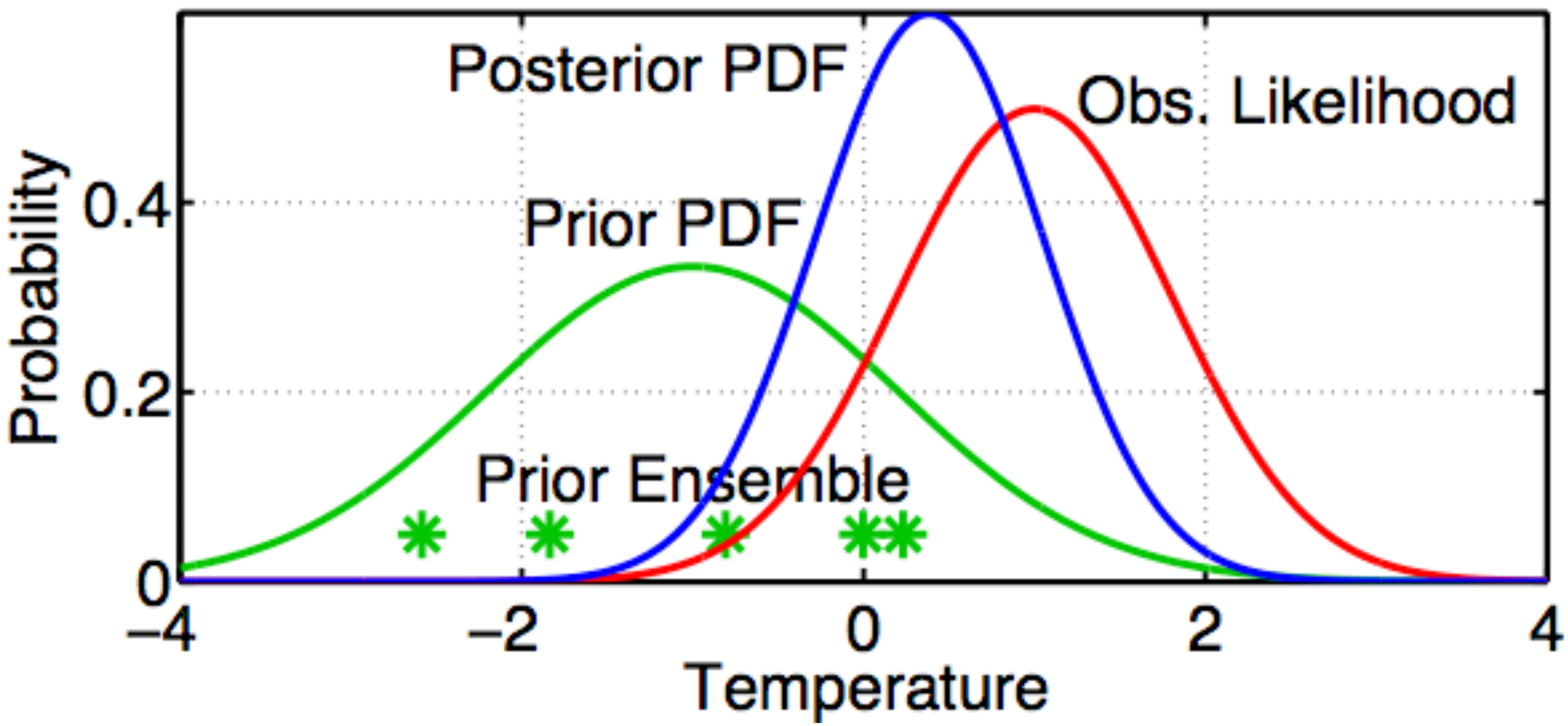
Fit a Gaussian to the sample.

Ensemble Kalman Filter Step 1: Observation Increments



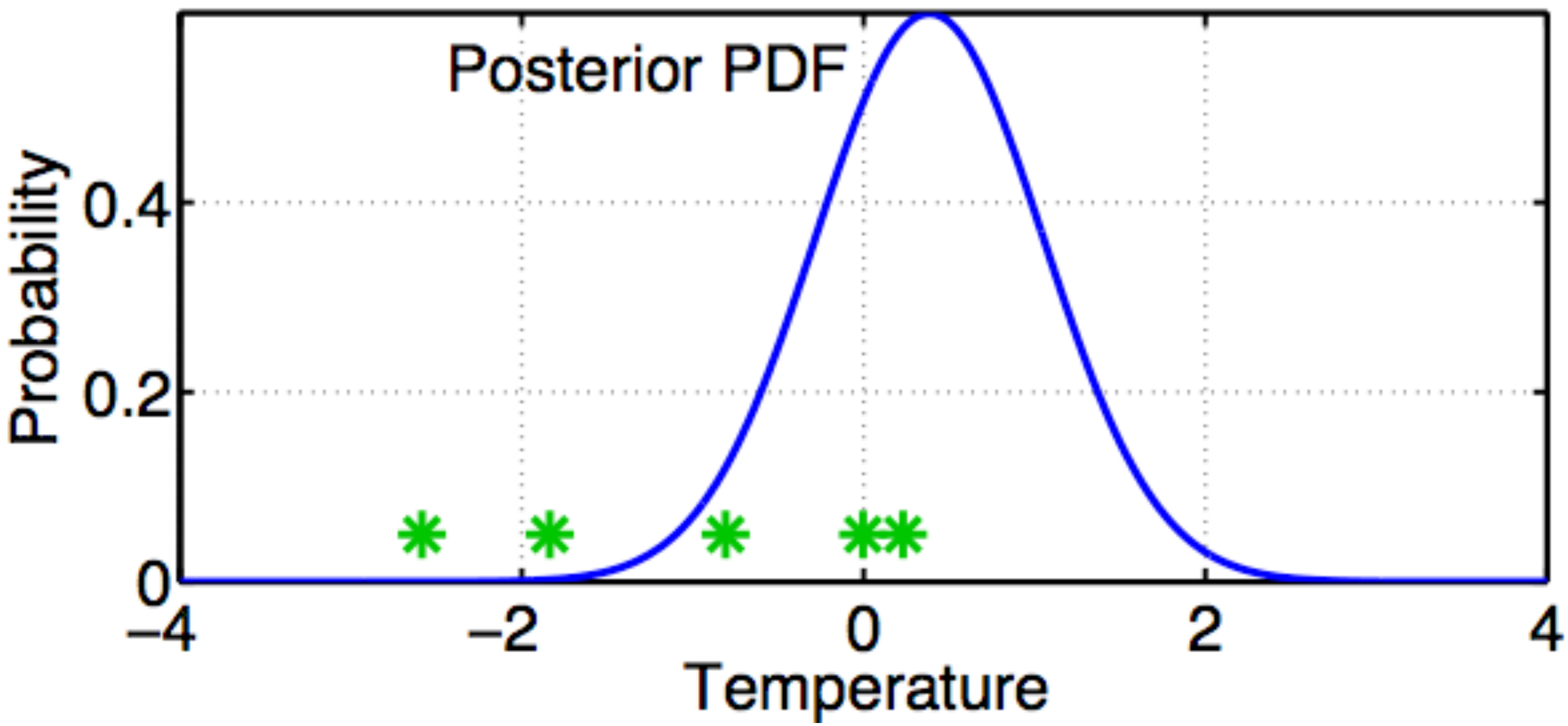
Get the observation likelihood.

Ensemble Kalman Filter Step 1: Observation Increments



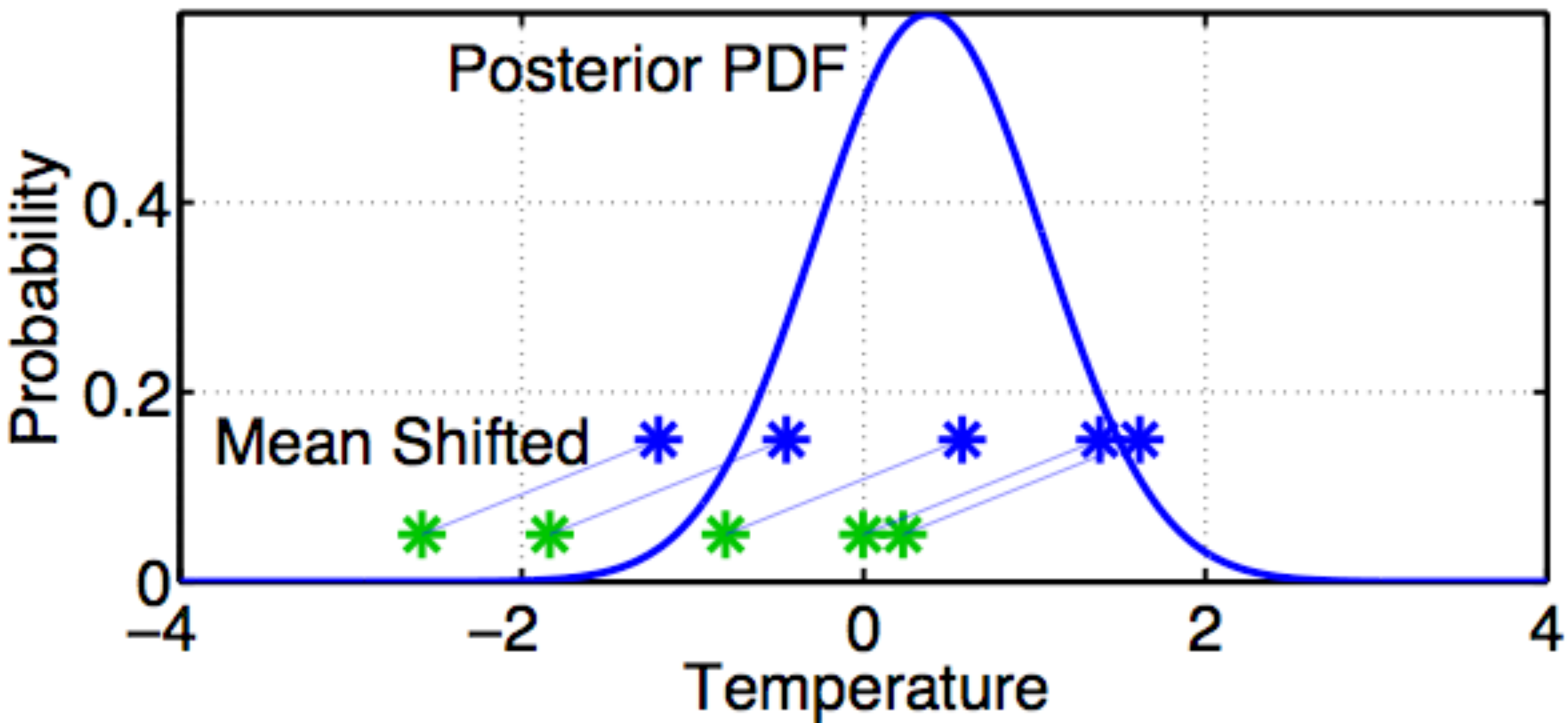
Compute the continuous posterior PDF.

Ensemble Kalman Filter Step 1: Observation Increments



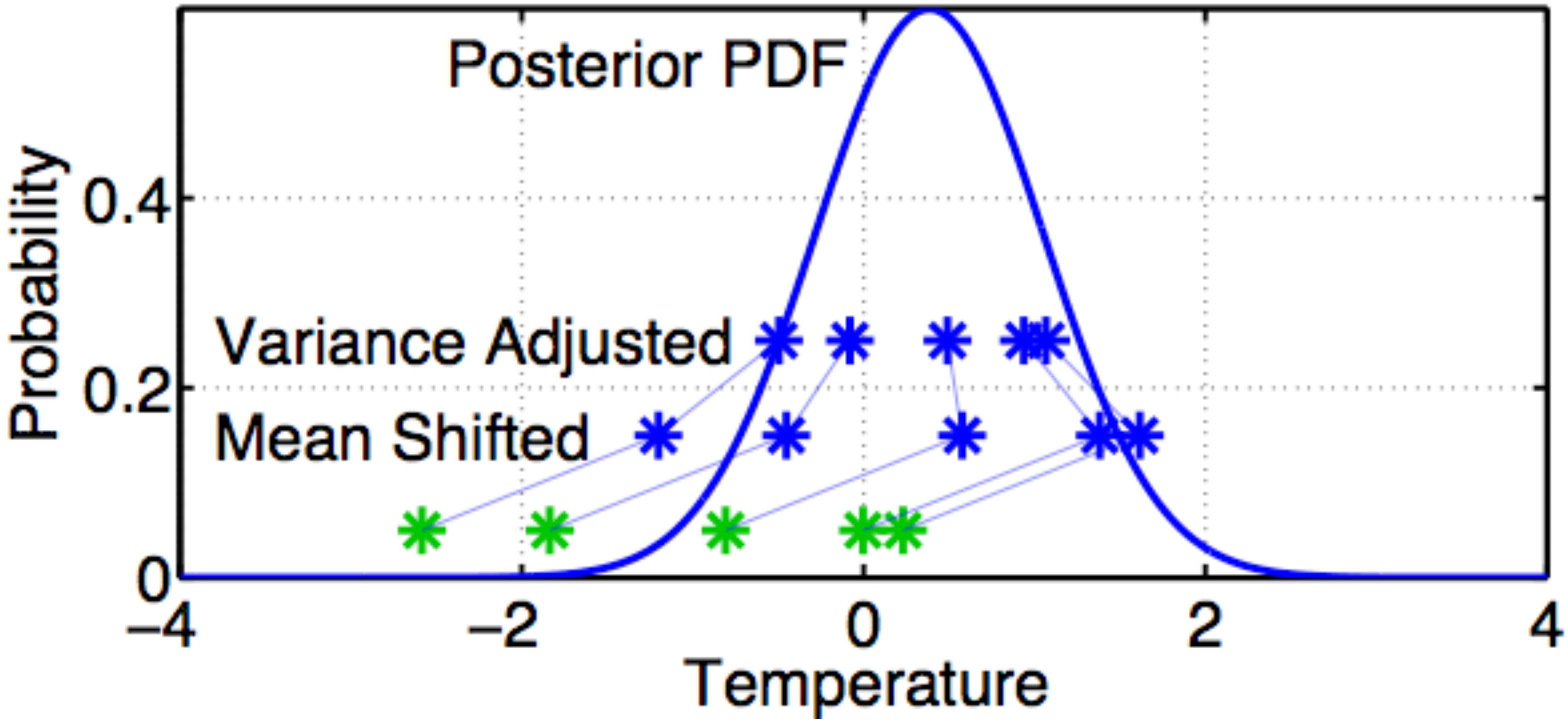
Use a deterministic algorithm to ‘adjust’ the ensemble.

Ensemble Kalman Filter Step 1: Observation Increments



First, ‘shift’ the ensemble to have the exact mean of the posterior.

Ensemble Kalman Filter Step 1: Observation Increments



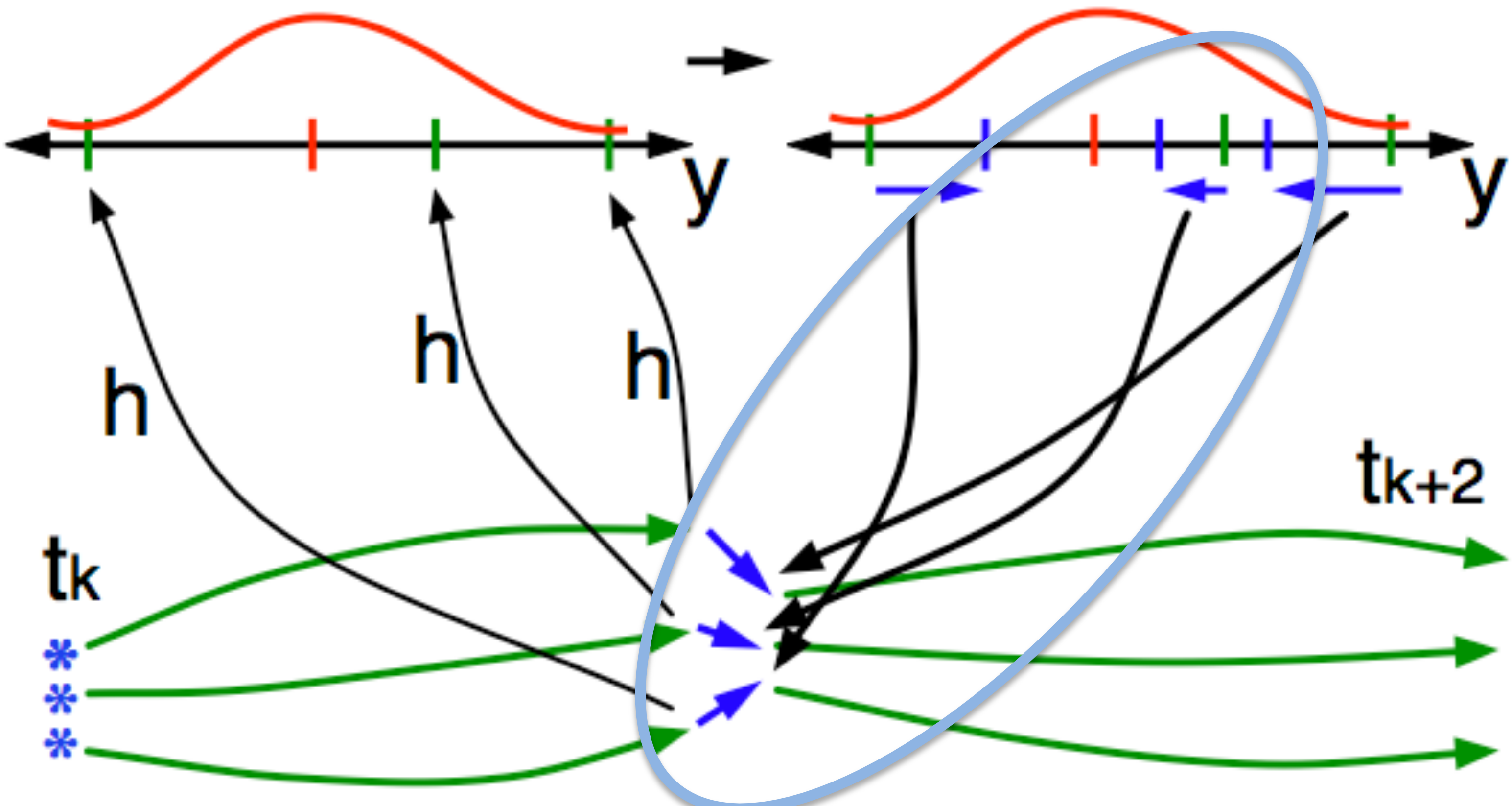
First, ‘shift’ the ensemble to have the exact mean of the posterior.

Second, linearly contract to have the exact variance of the posterior.

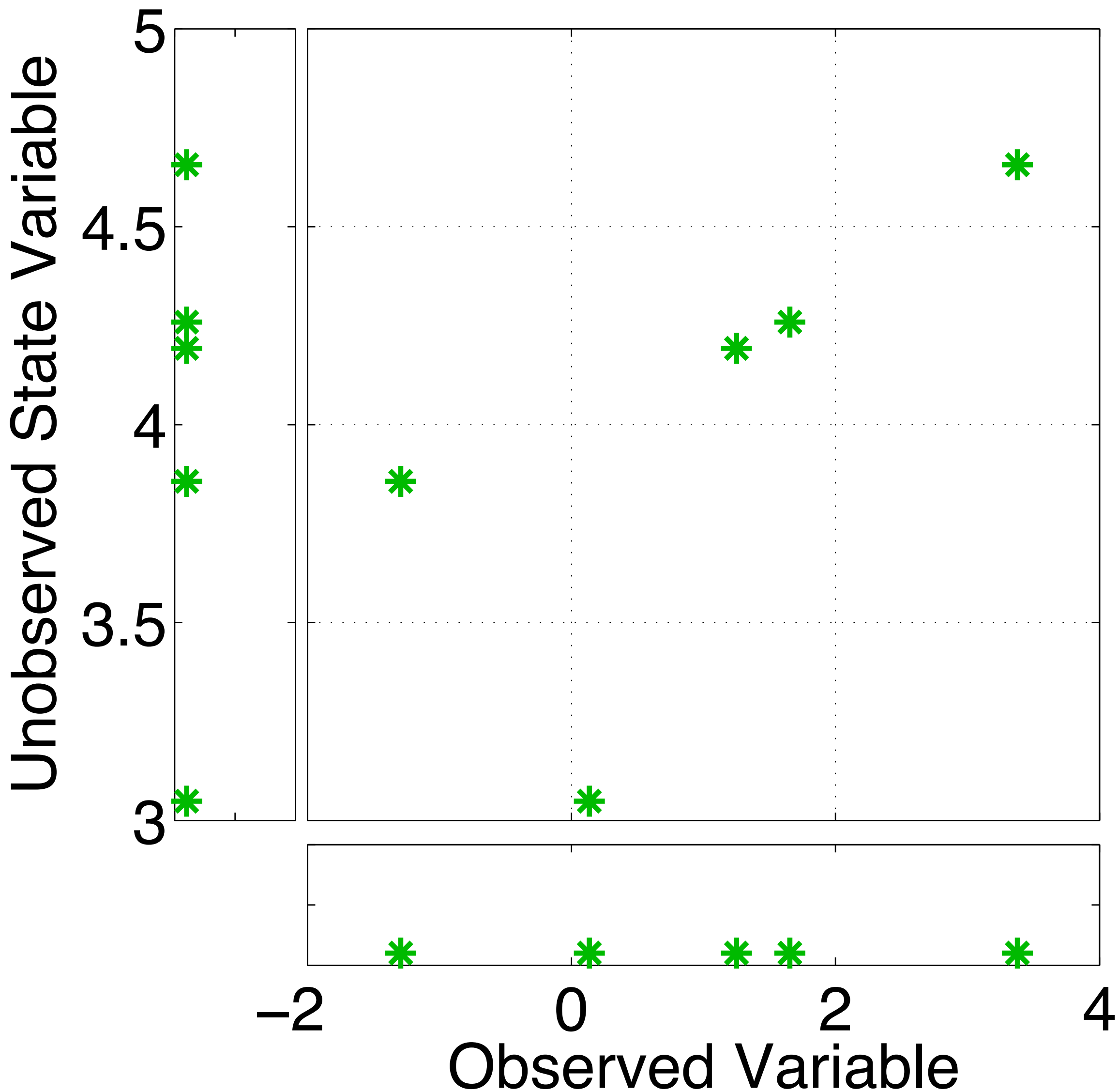
Sample statistics are identical to Kalman filter.

Ensemble Kalman Filter Step 2: Update Other Variables

Linear regression of observation increments onto each state variable independently (used for parallelism in DART).



Ensemble Kalman Filter Step 2: Update Other Variables

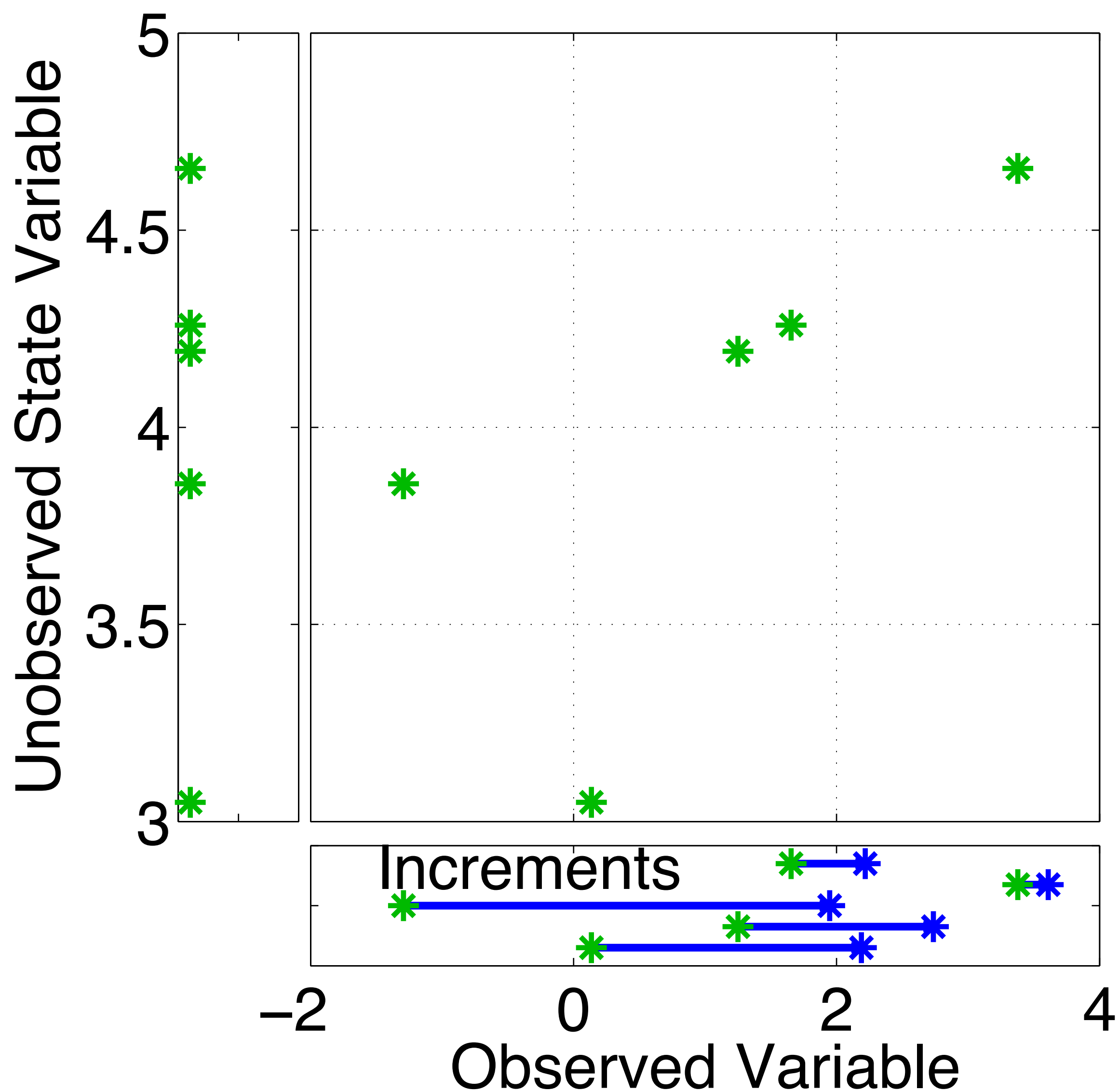


Assume that all we know is the prior joint distribution.

One variable is observed.

What should happen to the unobserved variable?

Ensemble Kalman Filter Step 2: Update Other Variables



Assume that all we know is the prior joint distribution.

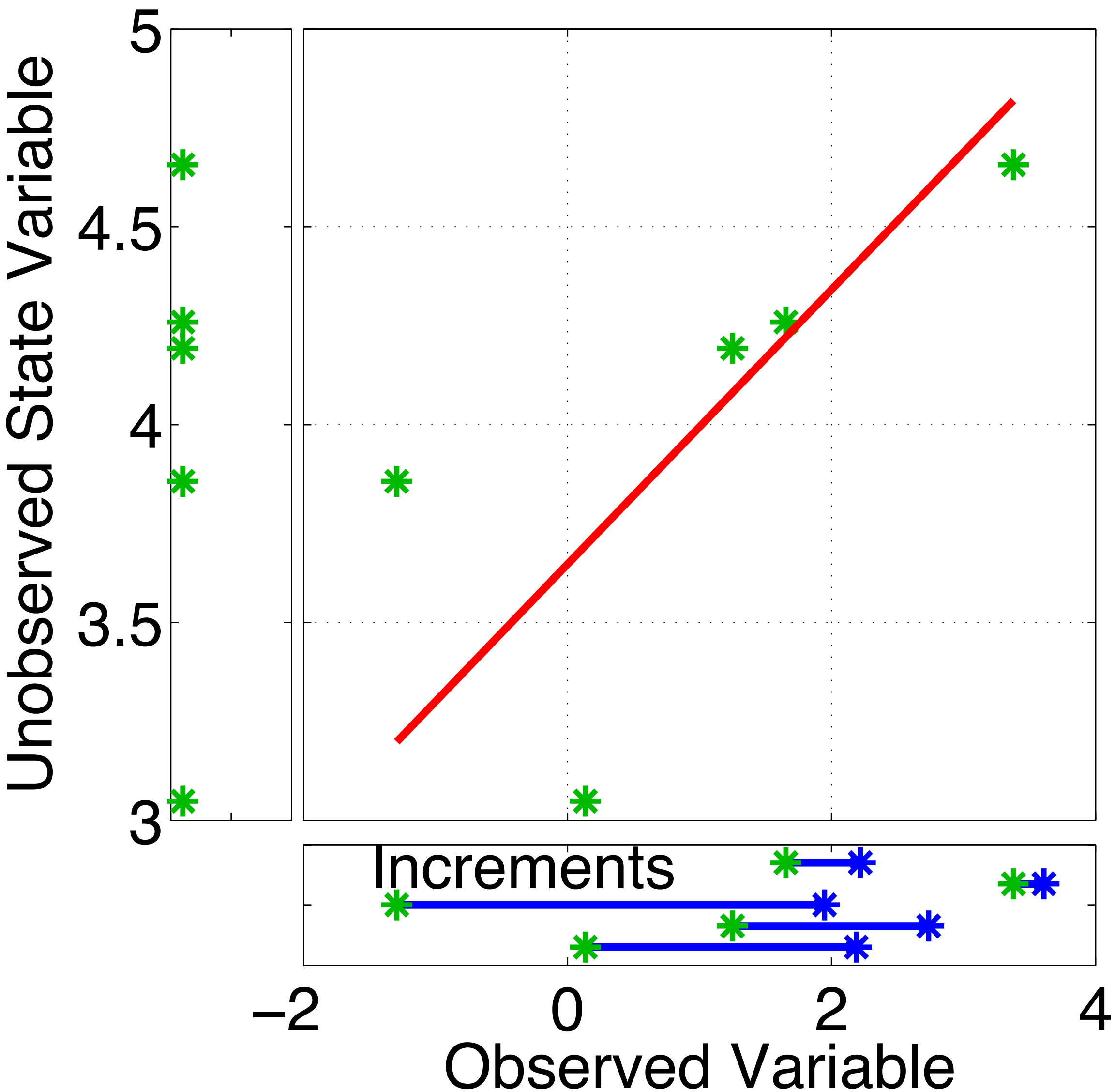
How should the unobserved variable be impacted?

1st choice: least squares

Equivalent to linear regression.

Same as assuming binormal prior.

Ensemble Kalman Filter Step 2: Update Other Variables



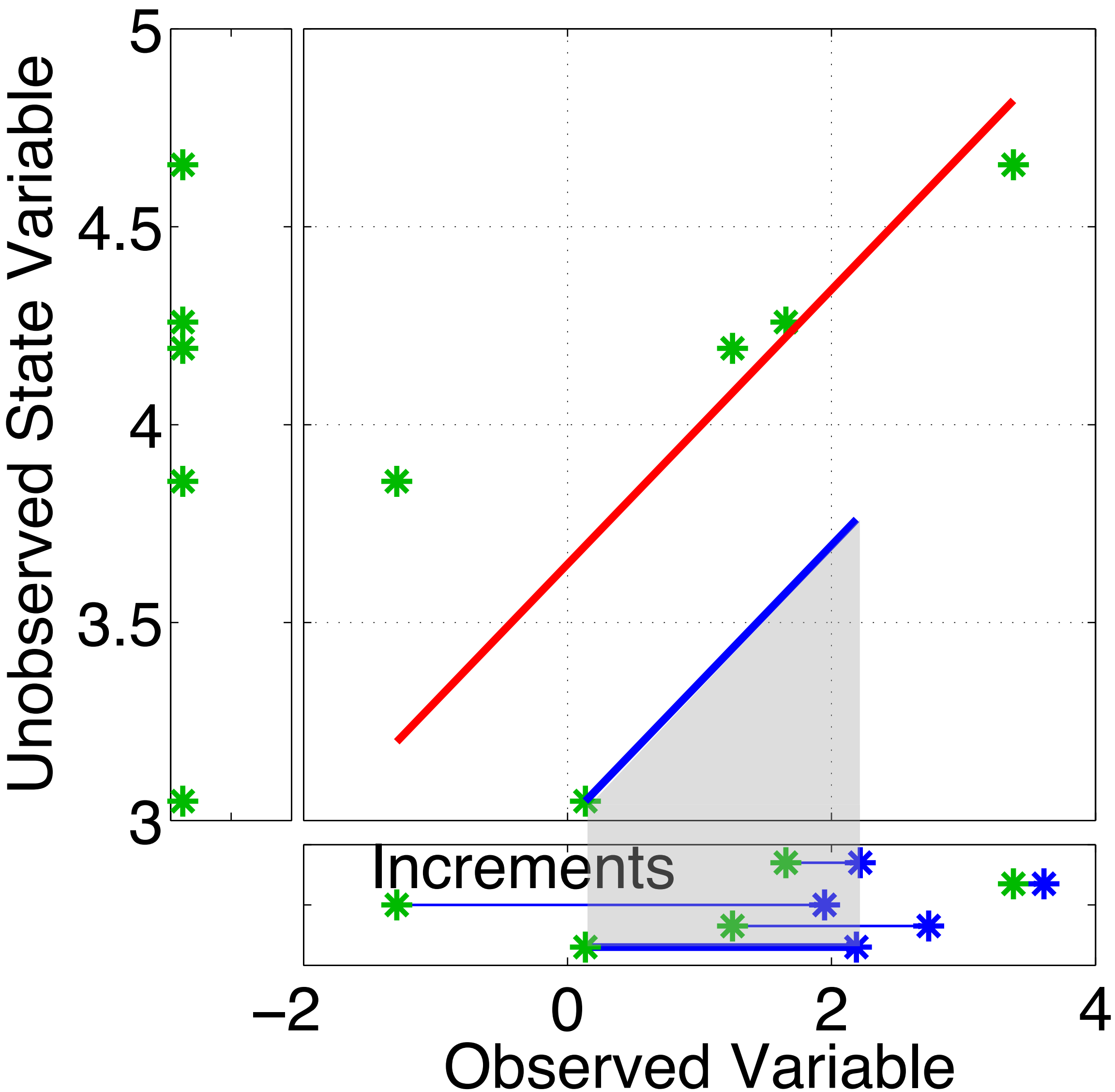
Have joint prior distribution of two variables.

How should the unobserved variable be impacted?

1st choice: least squares

Begin by finding least squares fit.

Ensemble Kalman Filter Step 2: Update Other Variables

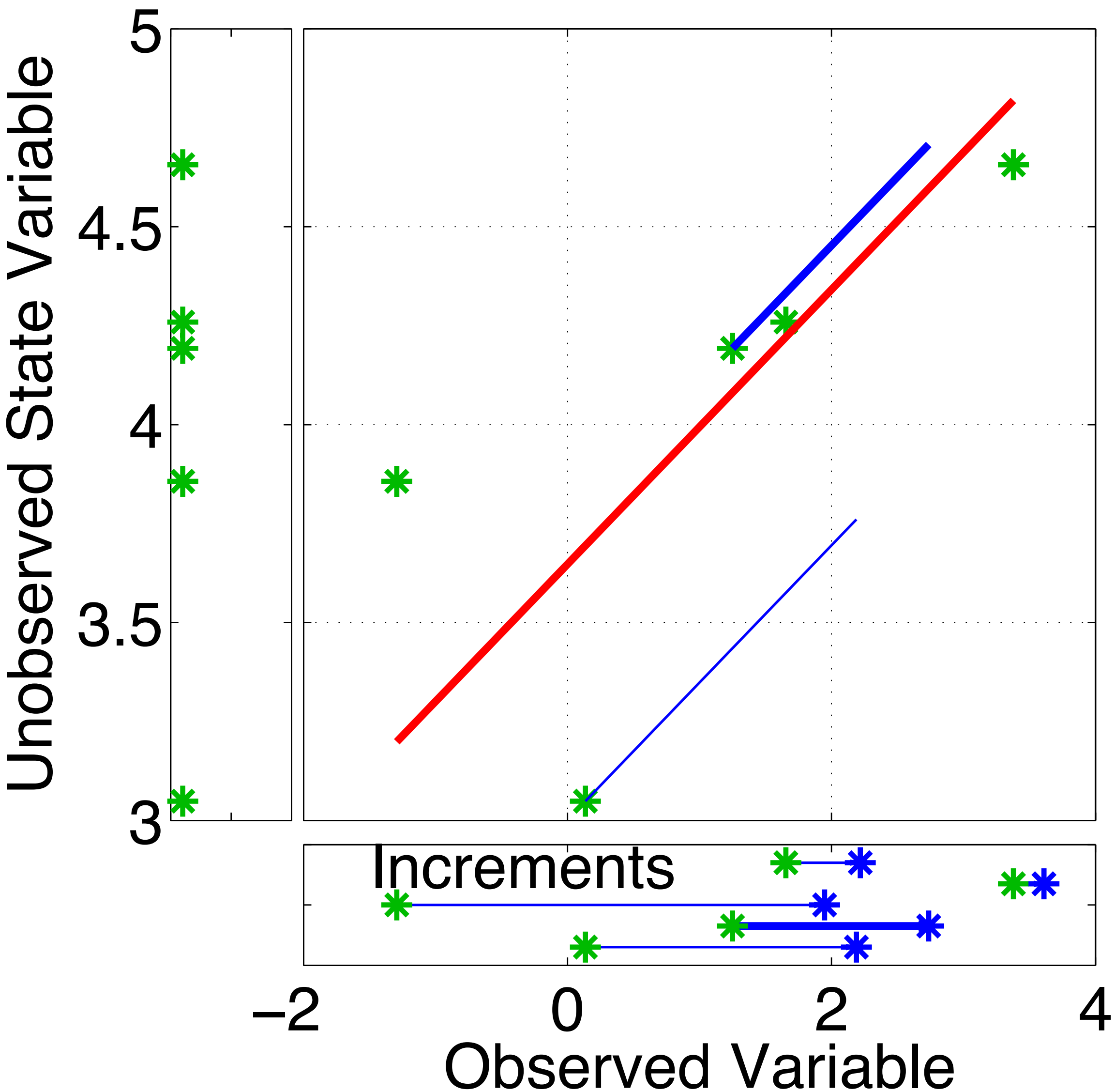


Have joint prior distribution of two variables.

Next, regress the observed variable increments onto increments for the unobserved variable.

Equivalent to first finding image of increment in joint space.

Ensemble Kalman Filter Step 2: Update Other Variables

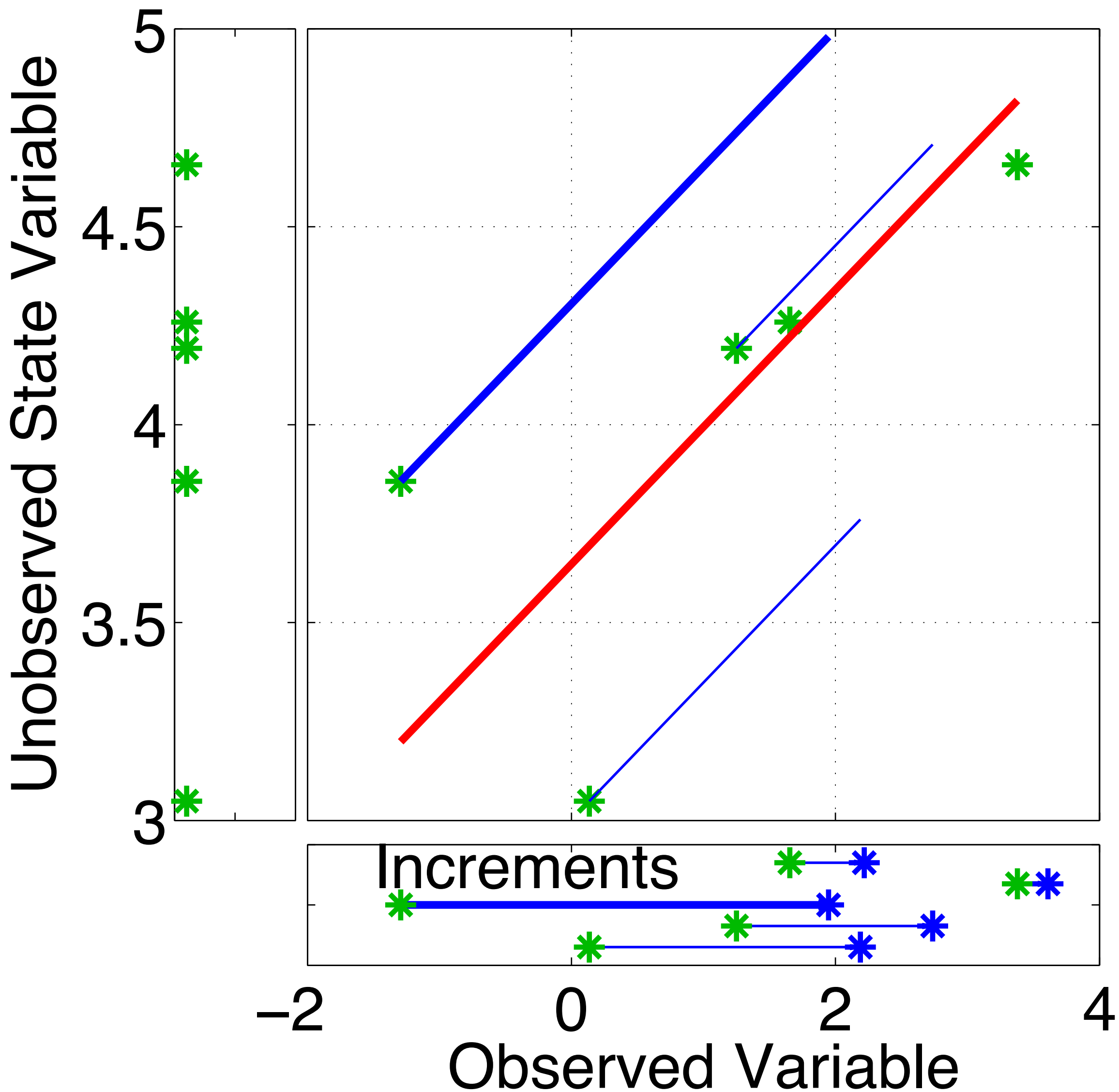


Have joint prior distribution of two variables.

Next, regress the observed variable increments onto increments for the unobserved variable.

Equivalent to first finding image of increment in joint space.

Ensemble Kalman Filter Step 2: Update Other Variables

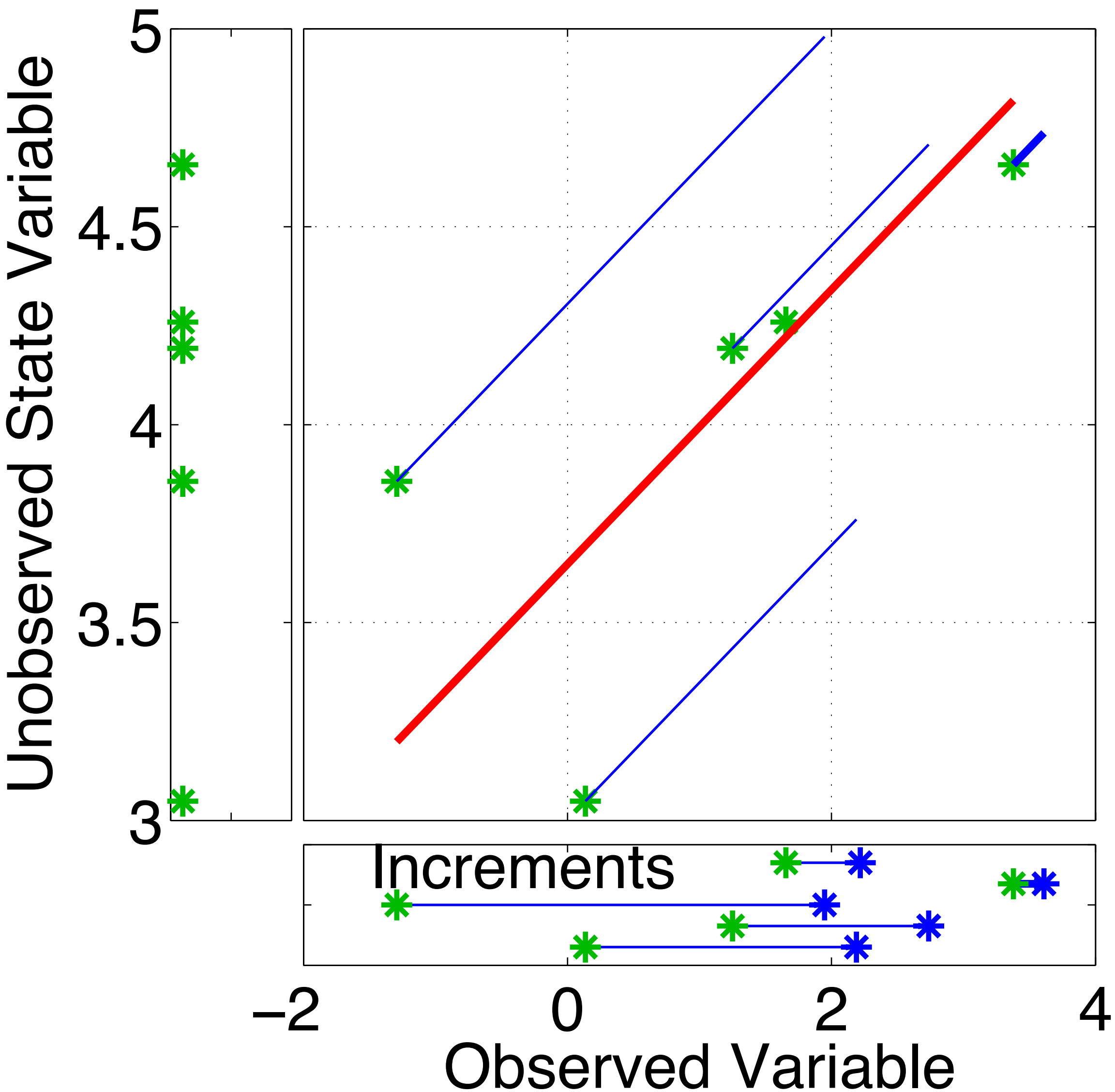


Have joint prior distribution of two variables.

Next, regress the observed variable increments onto increments for the unobserved variable.

Equivalent to first finding image of increment in joint space.

Ensemble Kalman Filter Step 2: Update Other Variables

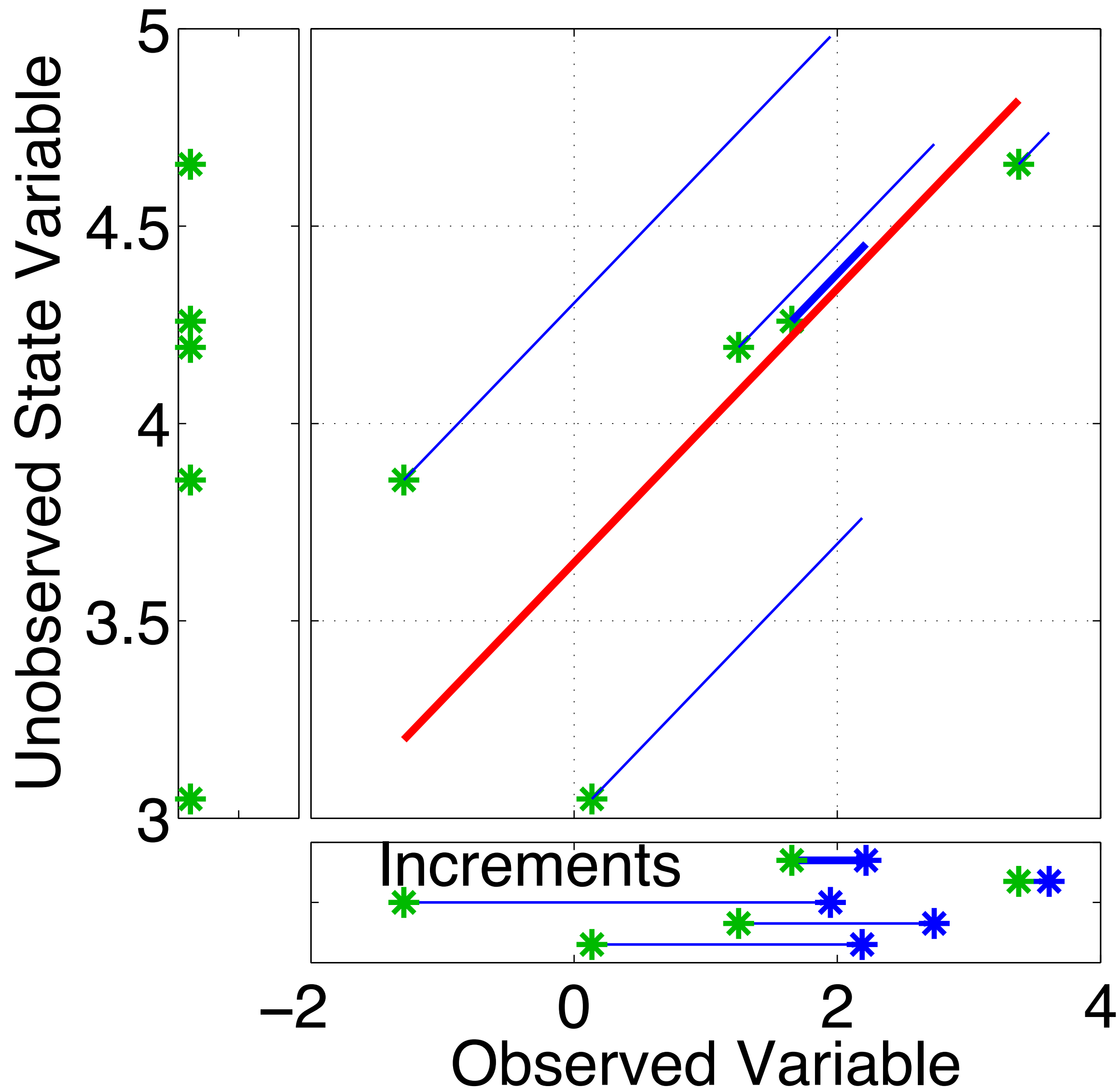


Have joint prior distribution of two variables.

Next, regress the observed variable increments onto increments for the unobserved variable.

Equivalent to first finding image of increment in joint space.

Ensemble Kalman Filter Step 2: Update Other Variables

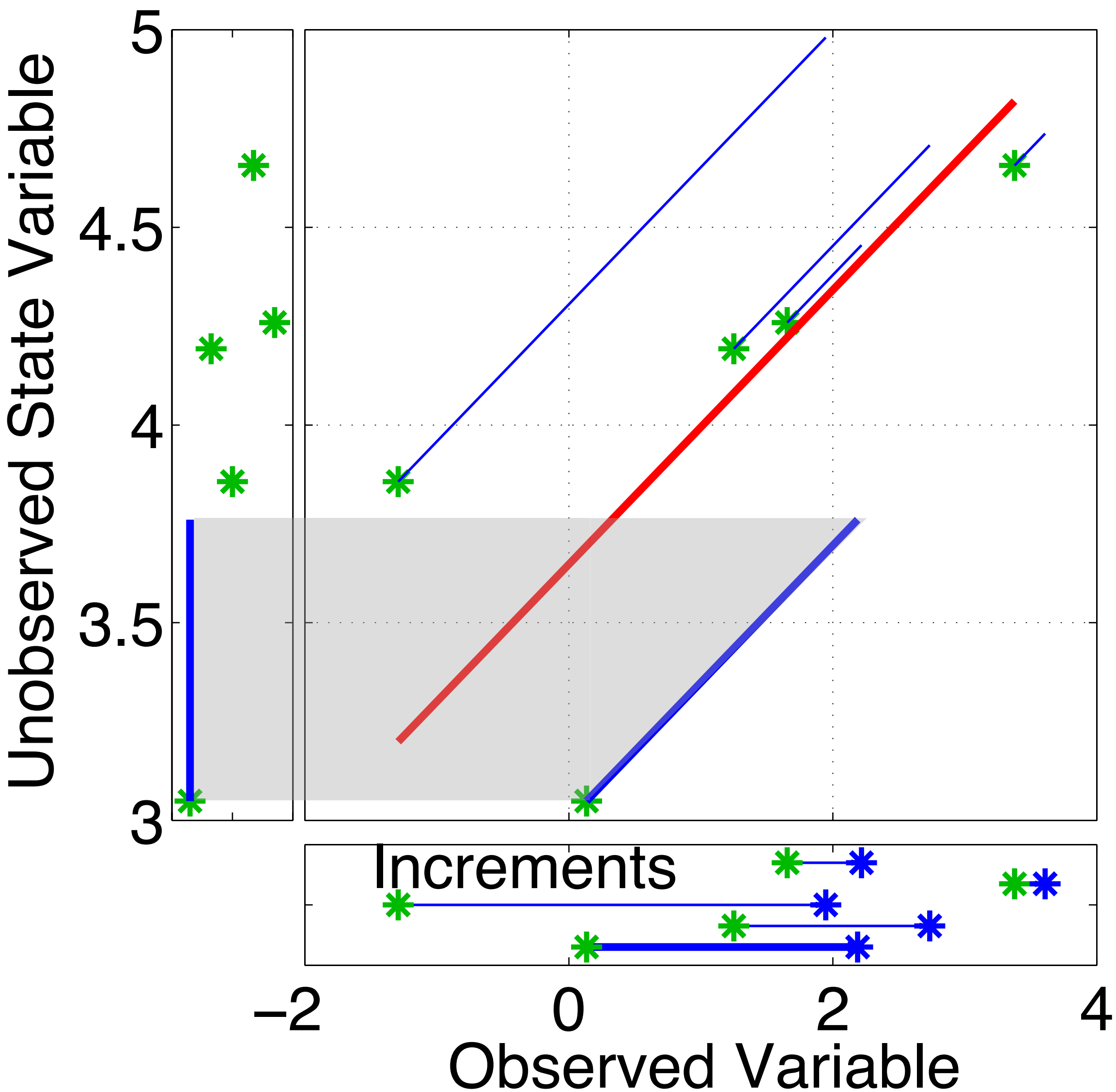


Have joint prior distribution of two variables.

Next, regress the observed variable increments onto increments for the unobserved variable.

Equivalent to first finding image of increment in joint space.

Ensemble Kalman Filter Step 2: Update Other Variables

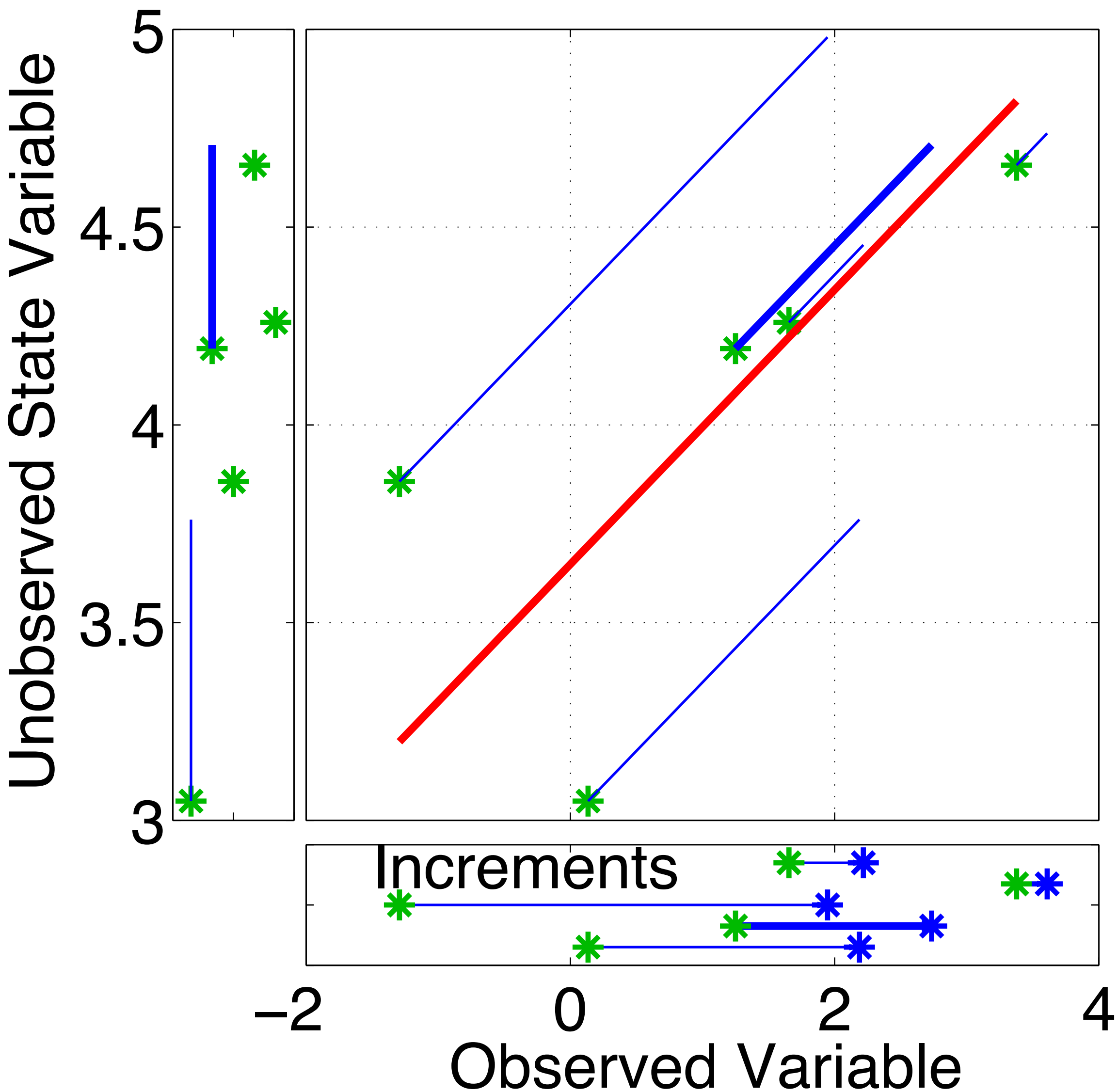


Have joint prior distribution of two variables.

Regression: Equivalent to first finding image of increment in joint space.

Then projecting from joint space onto unobserved priors.

Ensemble Kalman Filter Step 2: Update Other Variables

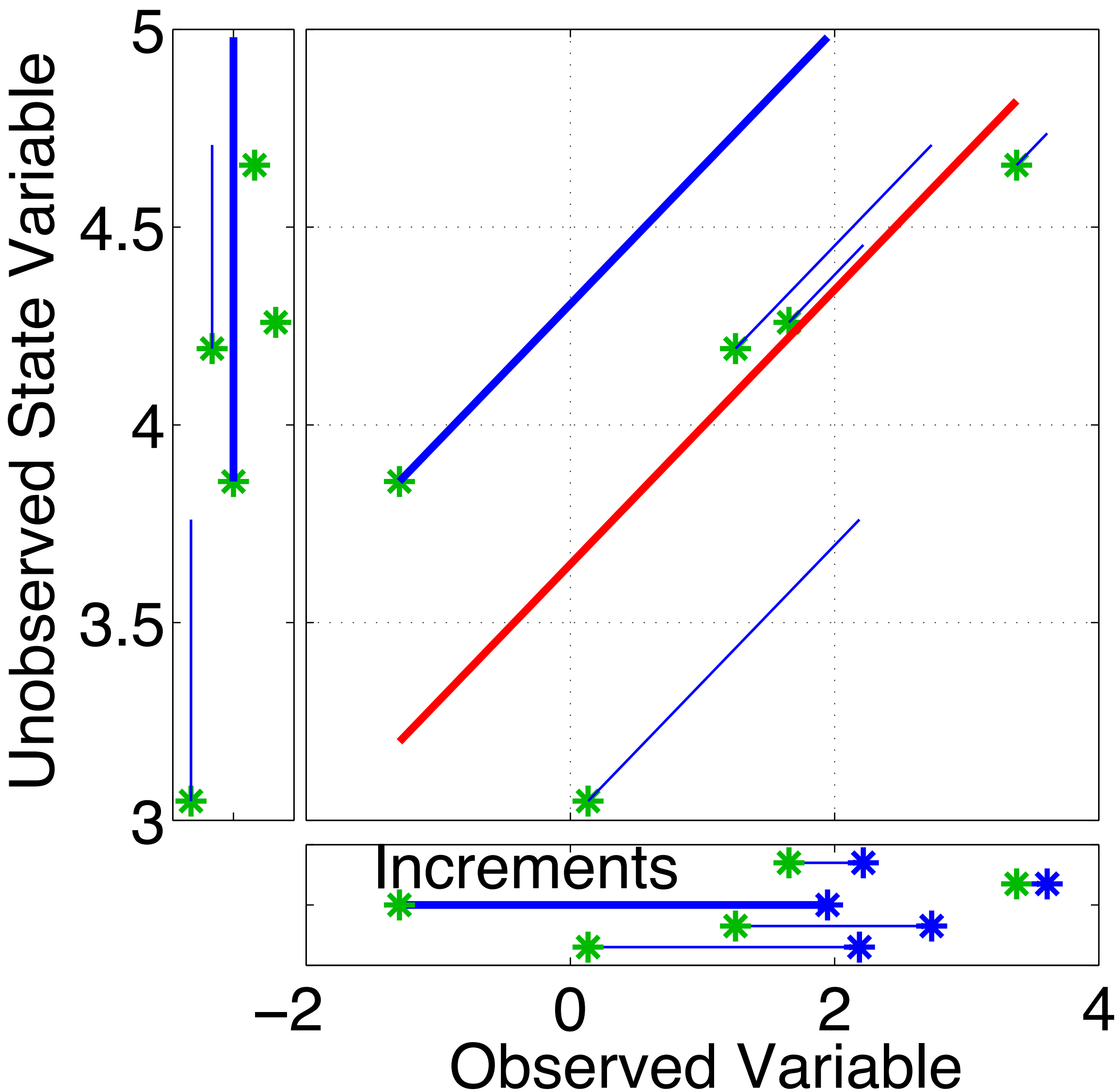


Have joint prior distribution of two variables.

Regression: Equivalent to first finding image of increment in joint space.

Then projecting from joint space onto unobserved priors.

Ensemble Kalman Filter Step 2: Update Other Variables

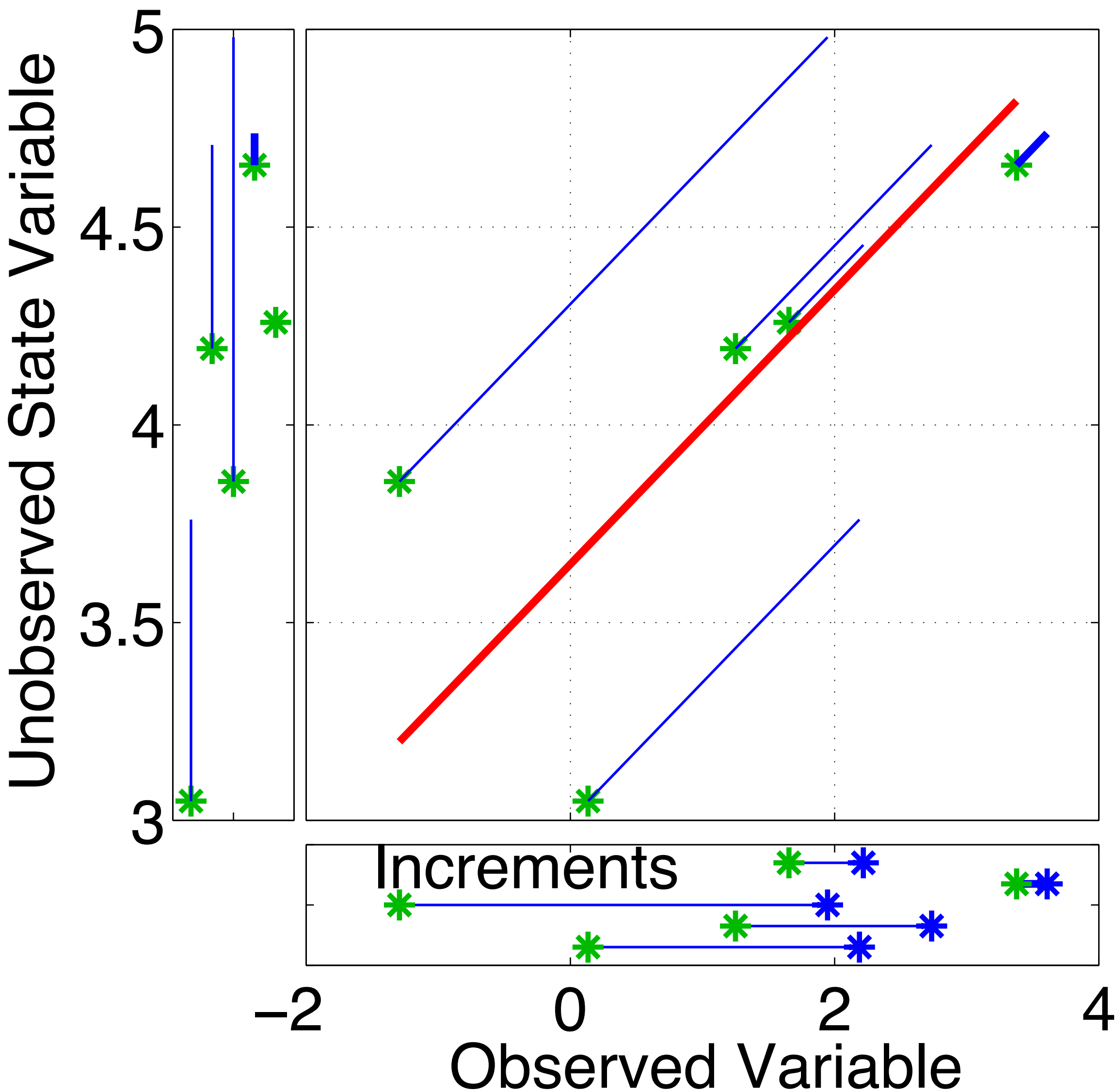


Have joint prior distribution of two variables.

Regression: Equivalent to first finding image of increment in joint space.

Then projecting from joint space onto unobserved priors.

Ensemble Kalman Filter Step 2: Update Other Variables

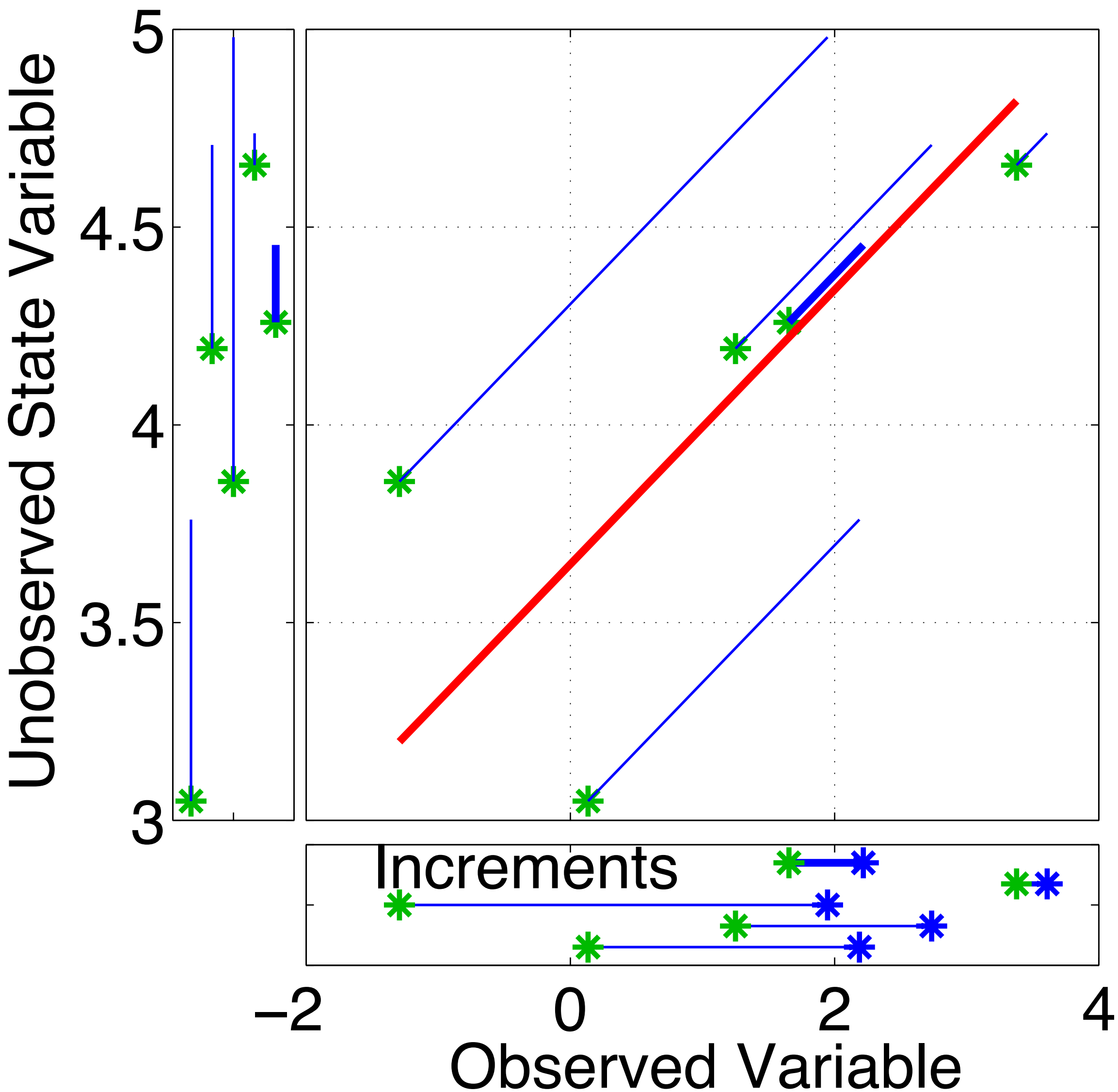


Have joint prior distribution of two variables.

Regression: Equivalent to first finding image of increment in joint space.

Then projecting from joint space onto unobserved priors.

Ensemble Kalman Filter Step 2: Update Other Variables

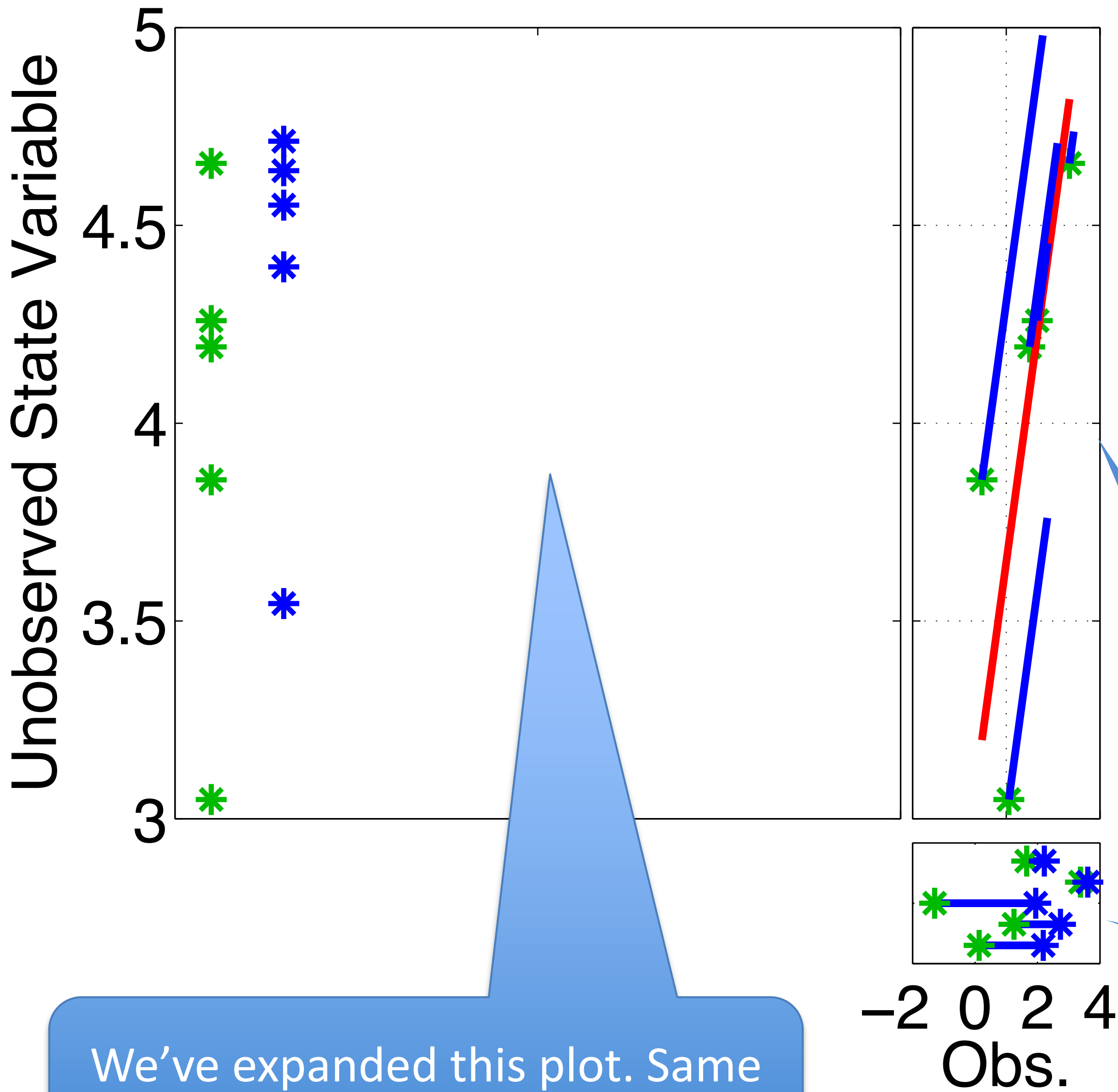


Have joint prior distribution of two variables.

Regression: Equivalent to first finding image of increment in joint space.

Then projecting from joint space onto unobserved priors.

Ensemble Kalman Filter Step 2: Update Other Variables

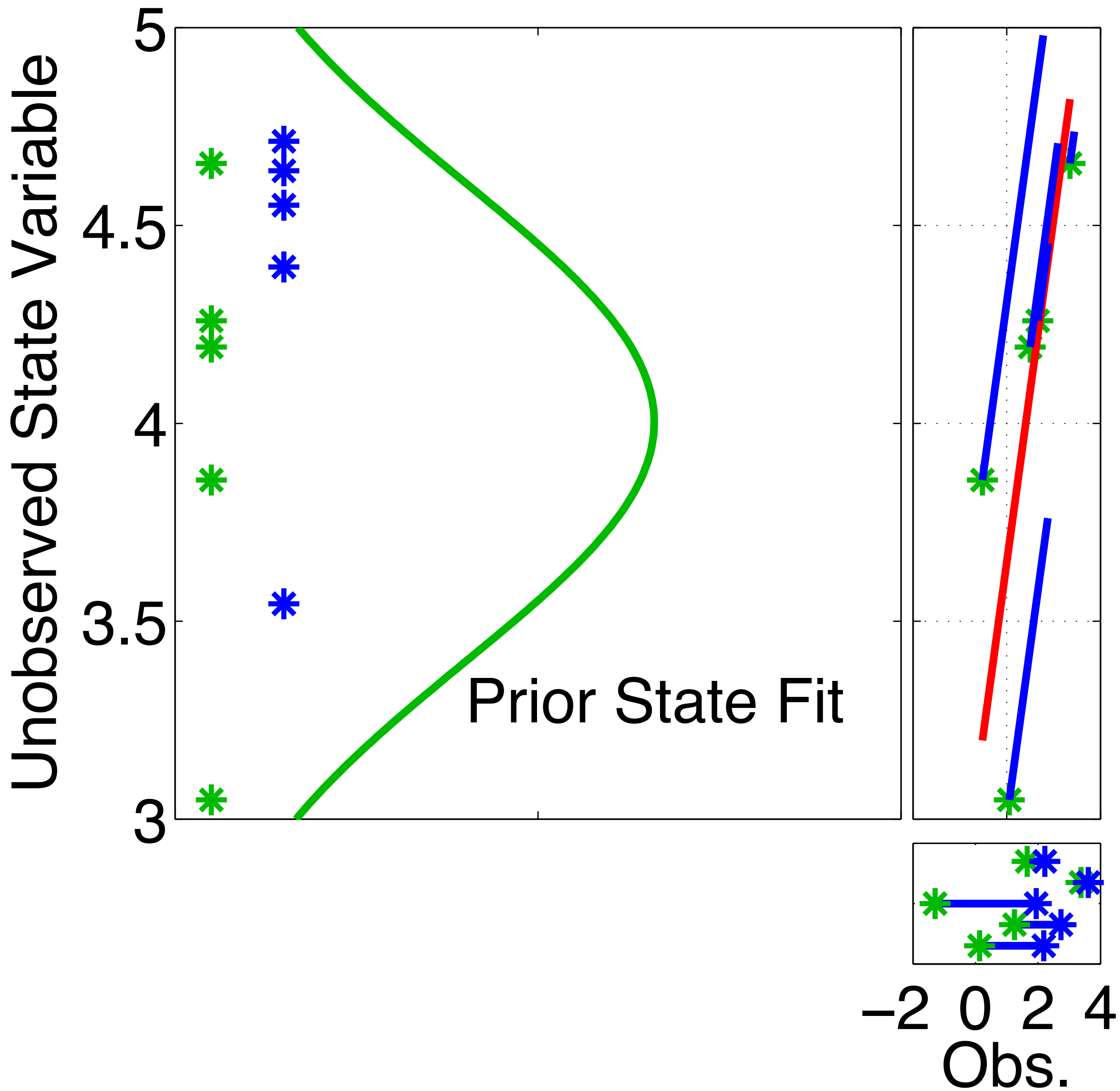


Now have an updated (posterior) ensemble for the unobserved variable.

We've expanded this plot. Same information as previous slides.

Compressed these two.

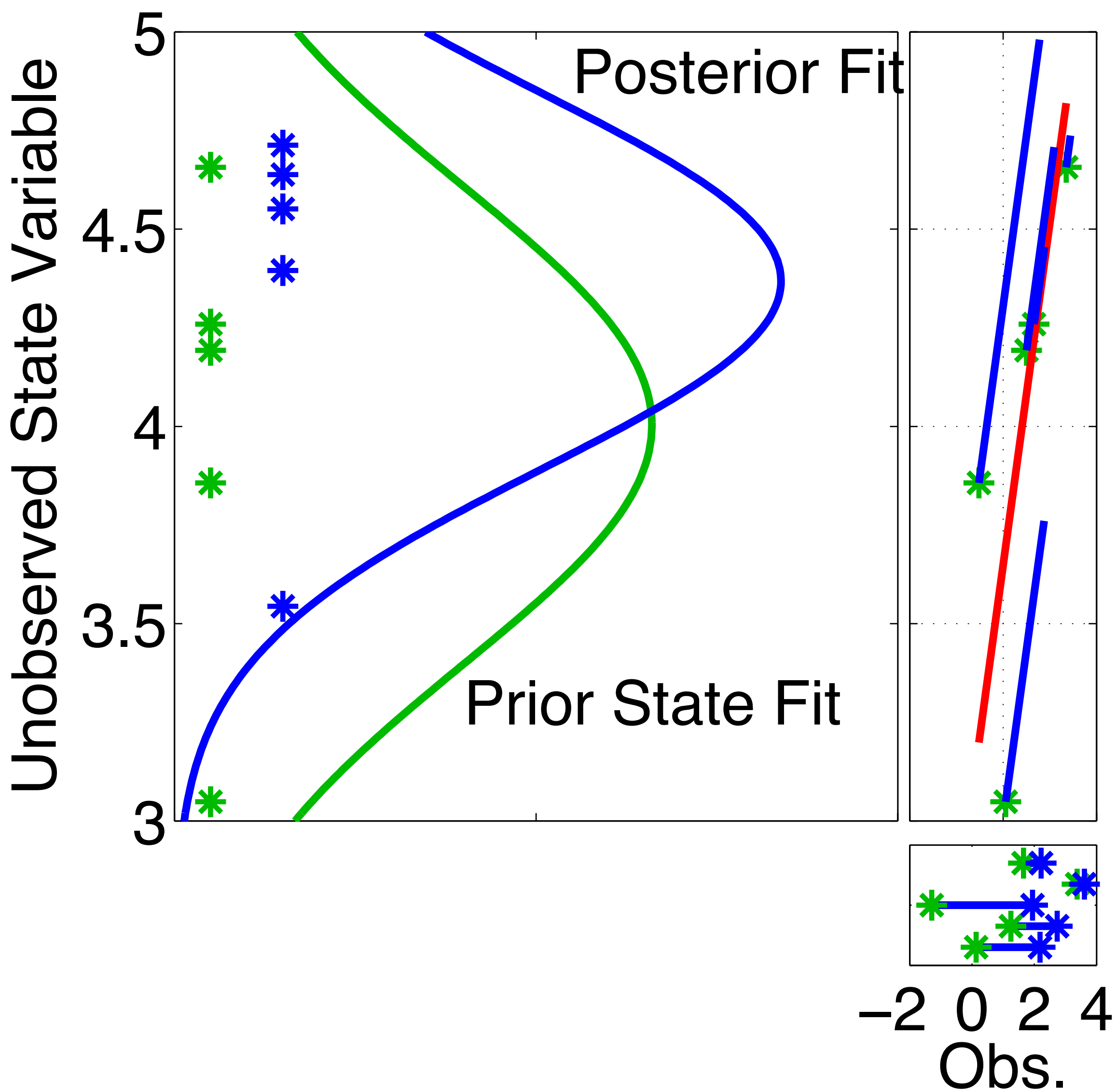
Ensemble Kalman Filter Step 2: Update Other Variables



Now have an updated (posterior) ensemble for the unobserved variable.

Fitting Gaussians shows that mean and variance have changed.

Ensemble Kalman Filter Step 2: Update Other Variables

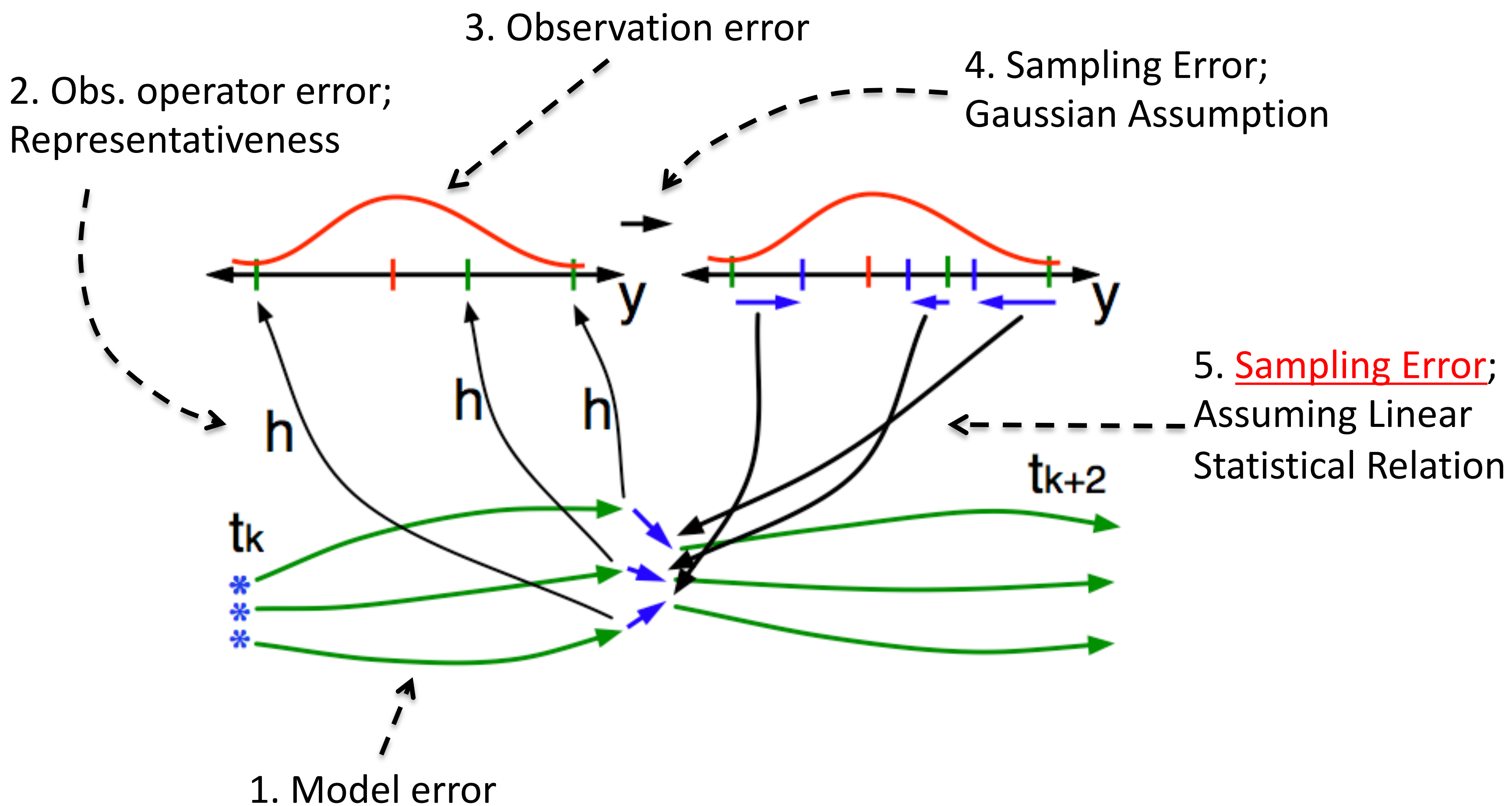


Now have an updated (posterior) ensemble for the unobserved variable.

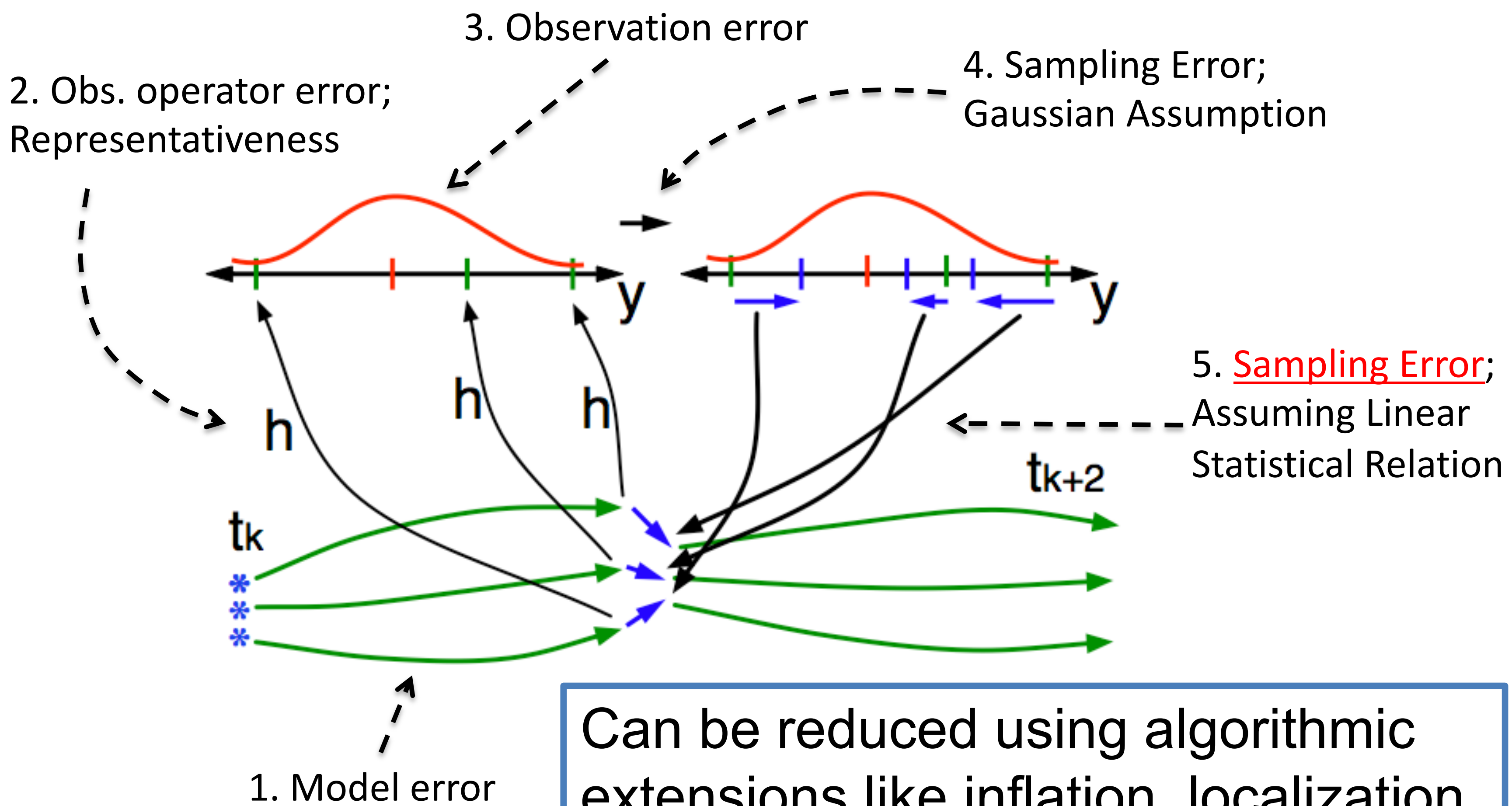
Fitting Gaussians shows that mean and variance have changed.

Other features of the prior distribution may also have changed.

Some Error Sources in Ensemble Filters



Some Error Sources in Ensemble Filters



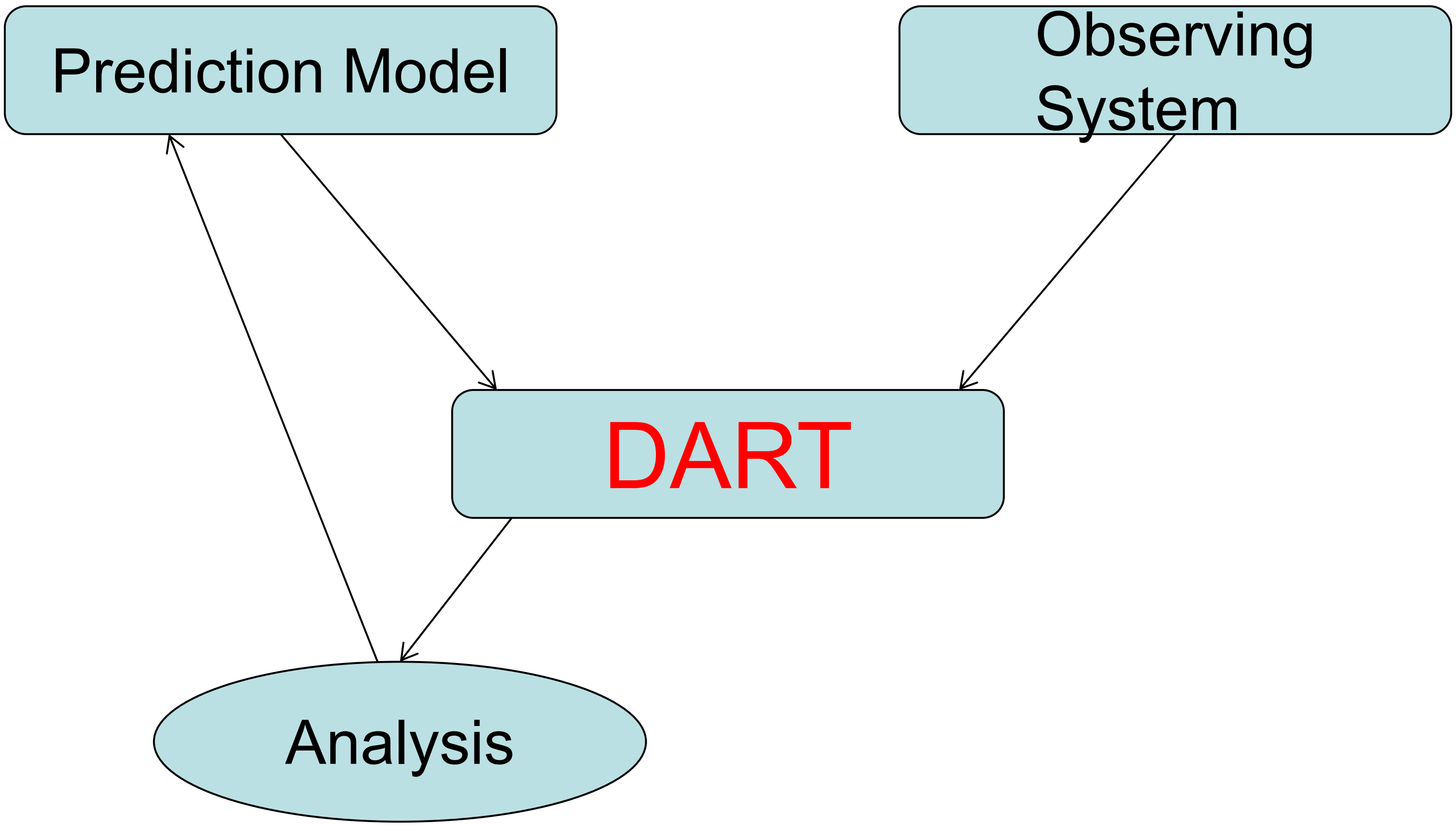
Can be reduced using algorithmic extensions like inflation, localization, sampling error correction, …

Ensemble Kalman Filter: Conclusions

- Basic Ensemble Kalman Filter is trivial.
- Good ensemble filters require inflation, localization, …
- ‘Automated’ inflation, localization algorithms exist.
- Parallel implementations for 100,000 cores for large models.
- Calibration and validation essential.
- Hybrids with variational methods.
- Other enhancements in the pipeline.

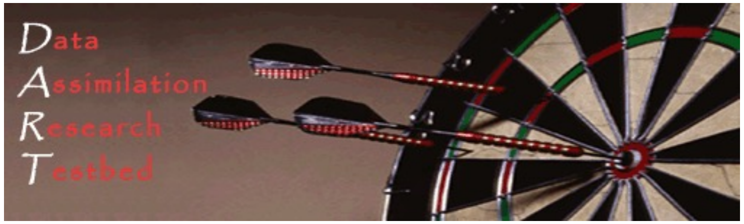
The Data Assimilation Research Testbed (DART)

DART provides data assimilation 'glue' to build ensemble forecast systems for the atmosphere, ocean, land, …



Data Assimilation Research Testbed (DART)

- A state-of-the-art Data Assimilation System for Geoscience
  - Flexible, portable, well-tested, extensible, free!
  - Works with many models.
  - Works with any observations: Real, synthetic, novel.
- A Data Assimilation Research System
  - Theory based, widely applicable general techniques.
  - Localization, Sampling Error Correction, Adaptive Inflation, …
- Professional software engineering
  - Carefully constructed and verified.
  - Excellent performance.
  - Comprehensive documentation, examples, tutorials.
- People: The DAReS Team



DART is used at:

48 UCAR member universities,
More than 100 other sites,
(More than 1500 registered users).

Colorado State University

中央氣象局 Central Weather Bureau

UNIVERSITY OF WASHINGTON

CIRES

GFDL Geophysical Fluid Dynamics Laboratory Princeton, NJ

THE UNIVERSITY OF UTAH

NOAA

PENNSTATE 1855

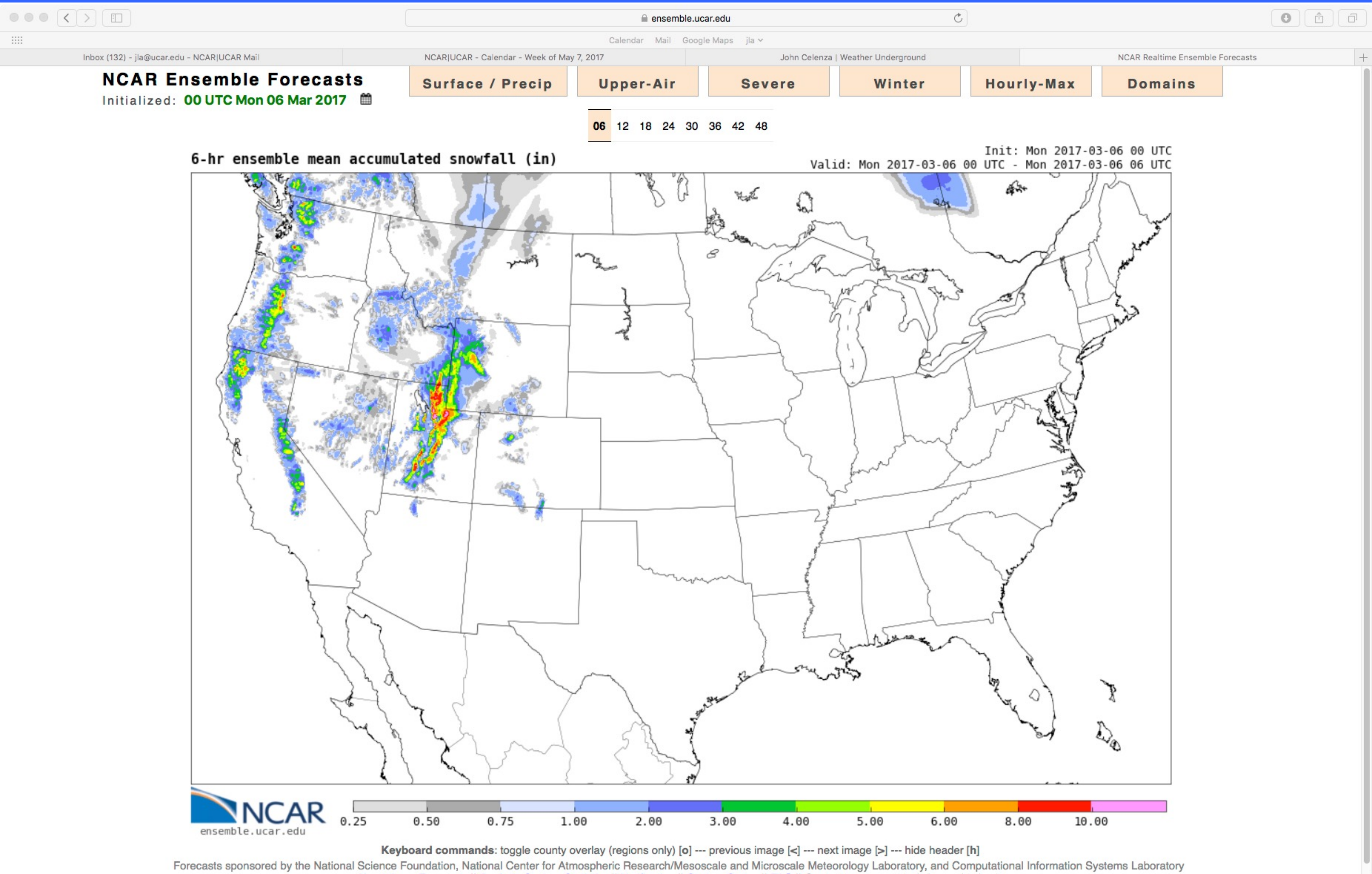
CALIFORNIA INSTITUTE OF TECHNOLOGY 1891

JPL
Jet Propulsion Laboratory
California Institute of Technology

DART Accelerates Forecast System Development

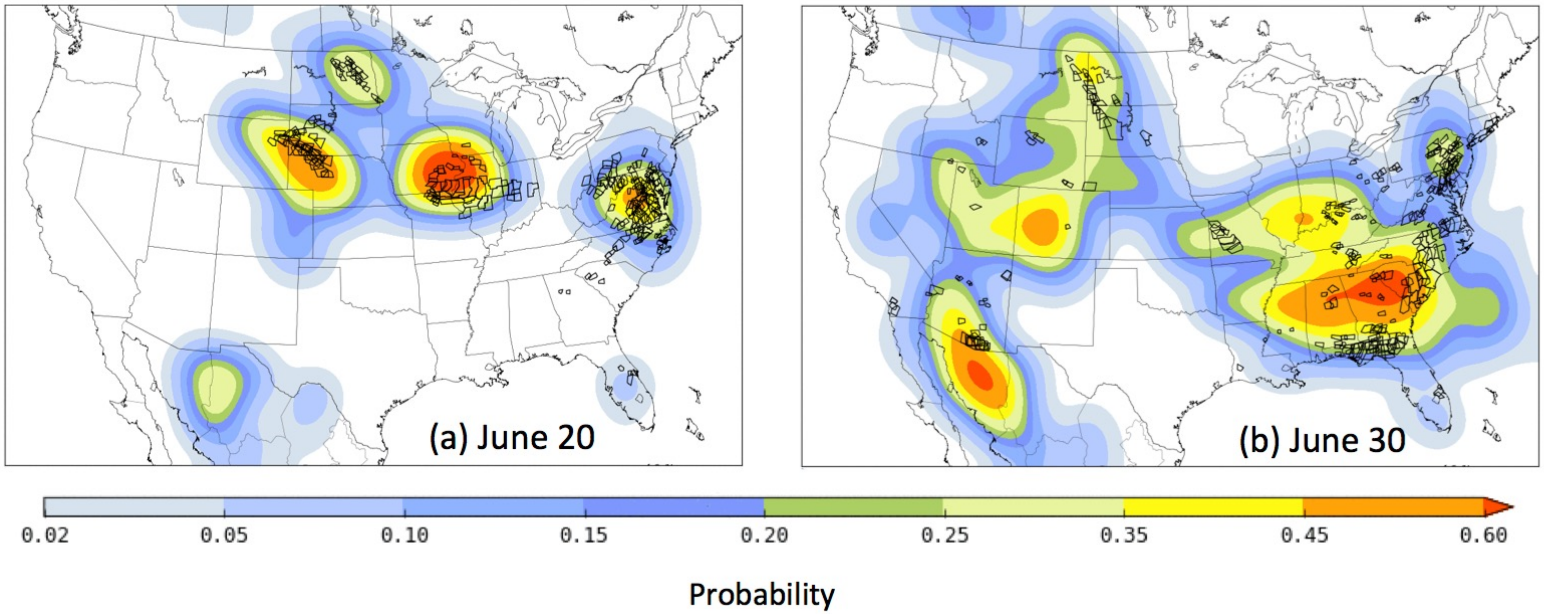
- Works with nearly all NCAR community models (dozens of other models, too).
- New models can be added in weeks.
- Adding new observations is even easier.
- Modular: models, observations and assimilation tools easily combined.
- Enables DA use by prediction scientists.
  - Doesn't require assimilation expertise.
- Fast & efficient software: laptops to supers.

Example: NCAR Real-time ensemble prediction system



Example: NCAR Real-time ensemble prediction system

Severe weather forecast for *two* days compared to NWS warnings



- WRF, 10 member ensemble, GFS for boundary conditions
- Continuous operation from April 2015 to December 2017
- 48 hour forecasts at 3km resolution
- First continuously cycling ensemble system for CONUS

DART Applications with CESM Earth System Models

DART interfaces exist for many components of NCAR's Community Earth System Model:

- Lower atmosphere: CAM-FV, CAM-SE, MPAS
- Upper atmosphere, ionosphere: WACCM, WACCMX
- Atmospheric Chemistry: CAM/Chem
- Ocean: POP
- Land surface / biosphere: CLM
- Sea Ice: CICE
- Weakly coupled DA combinations of the above

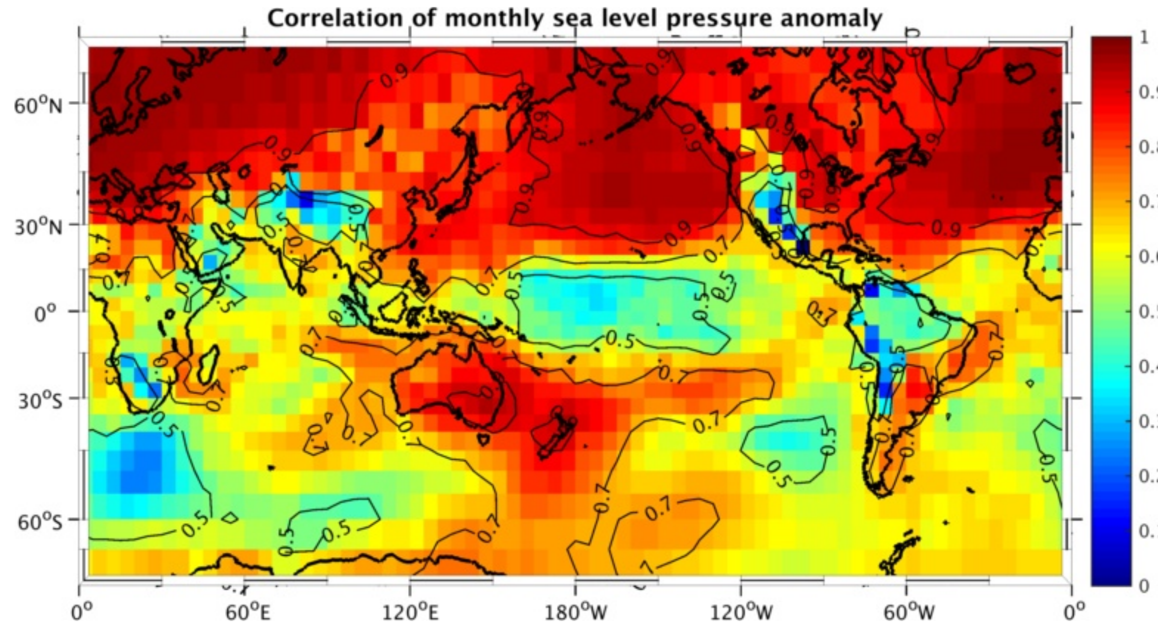
Multiple Component POP/CAM Coupled DA

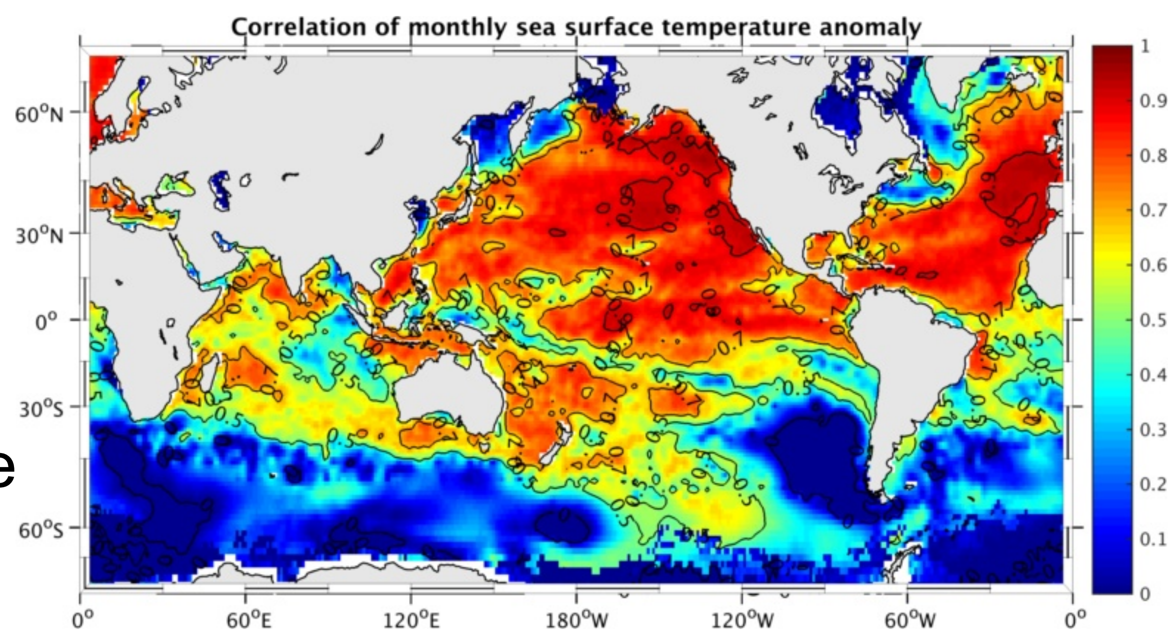
Comparisons to HADISST and HADSLP.

Correlation high where observations existed.

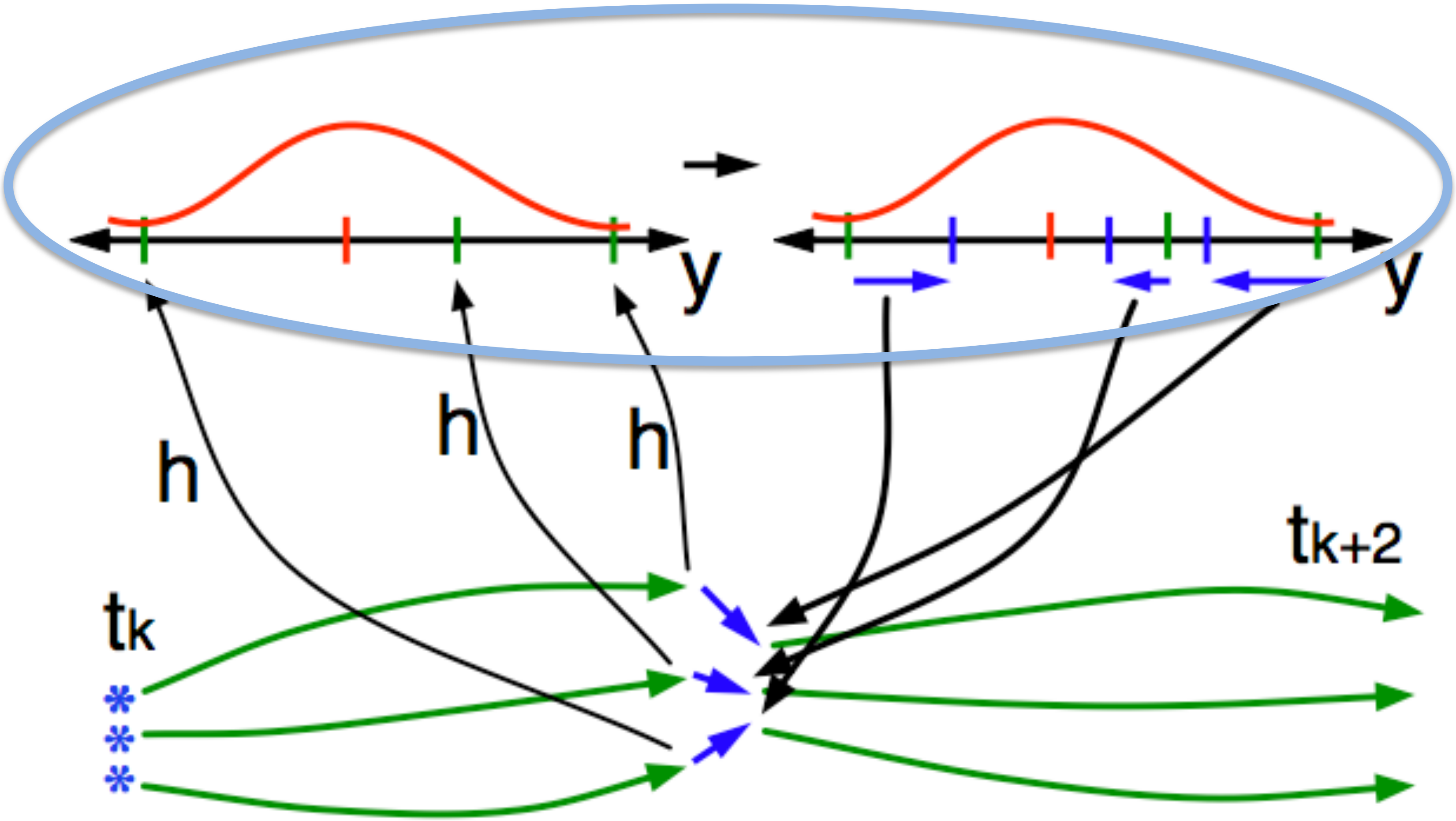
DART did not assimilate SST products or observations.

Produces competitive reanalysis.



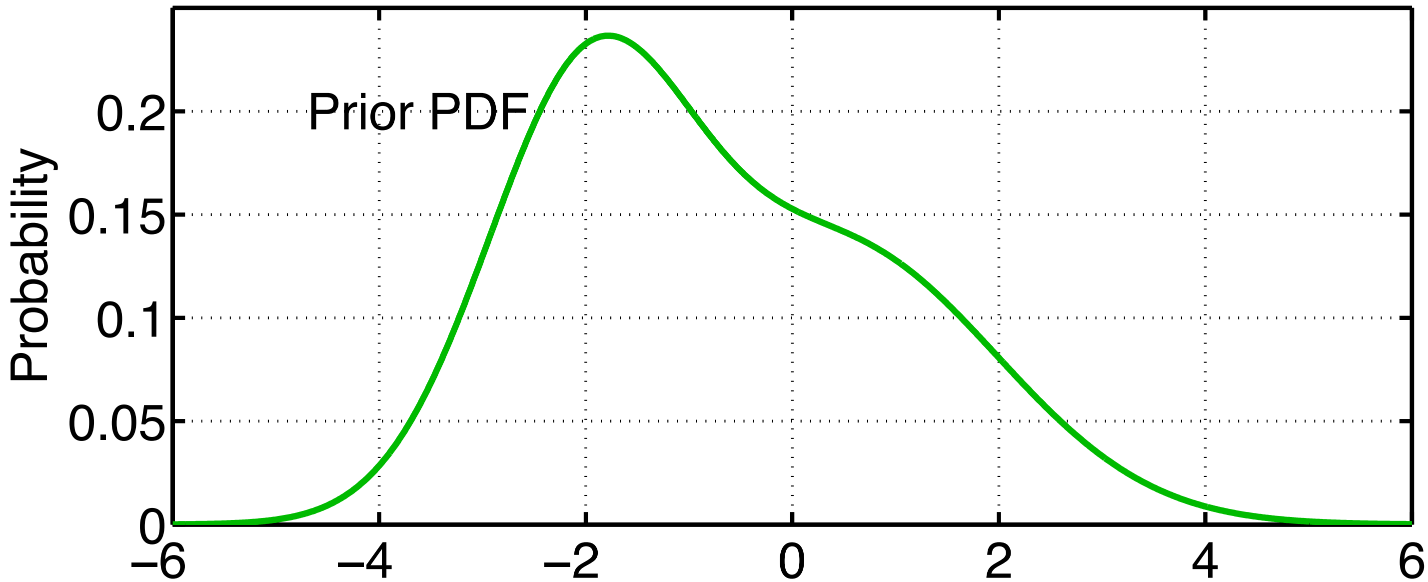


Novel Algorithm 1: General Method for Observation Increments



Bayes' Rule: Combining Forecast with Observation

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{Normalization}$$



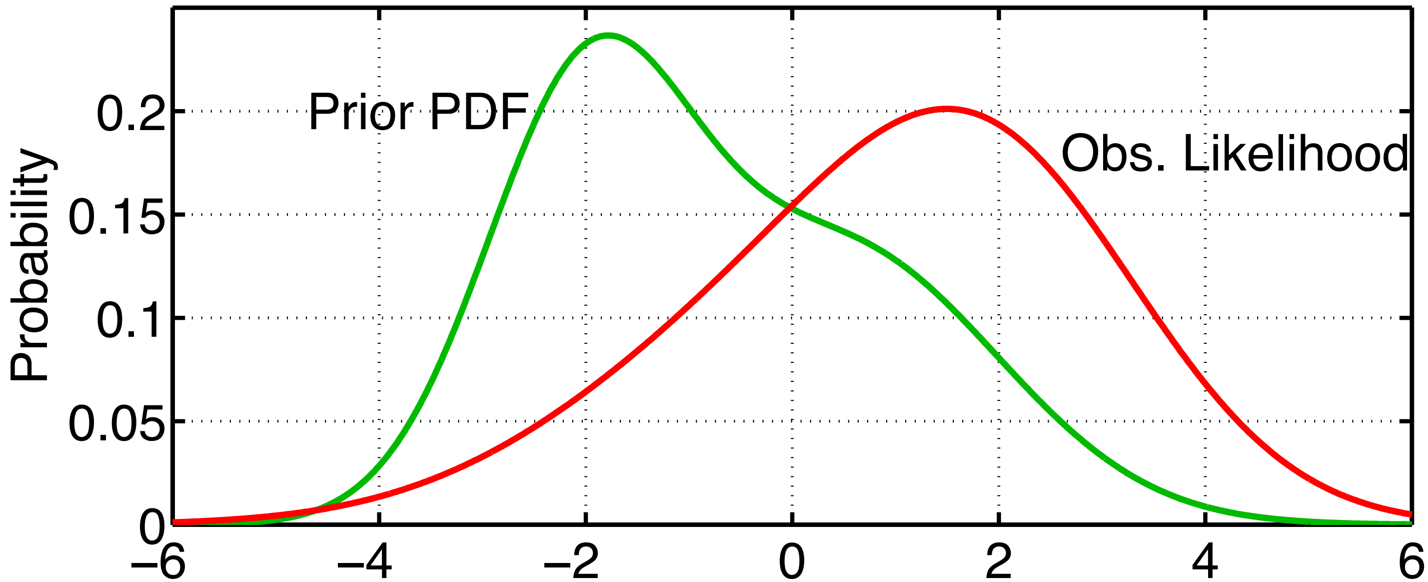
A : Forecast estimate from model (prior).

$p(B/A)$: Observation likelihood.

$p(A/B)$: Analysis (posterior) estimate combines A and B.

Bayes' Rule: Combining Forecast with Observation

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{Normalization}$$



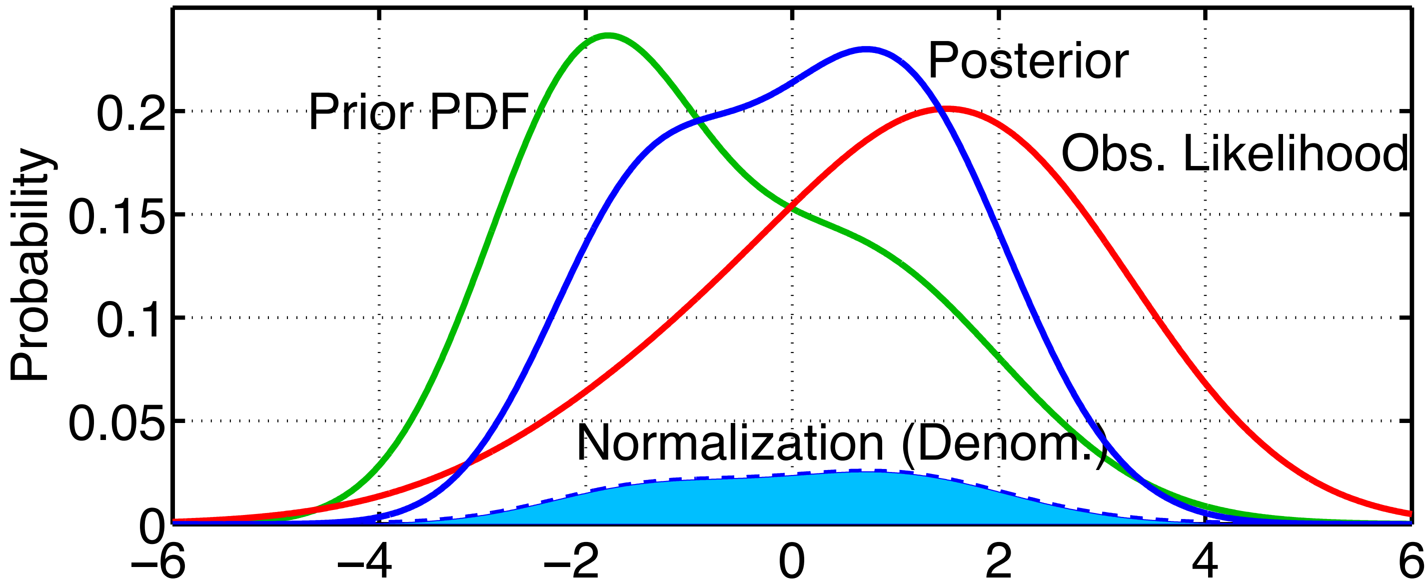
A : Forecast estimate from model (prior).

$p(B/A)$: Observation likelihood.

$p(A/B)$: Analysis (posterior) estimate combines A and B.

Bayes' Rule: Combining Forecast with Observation

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{Normalization}$$

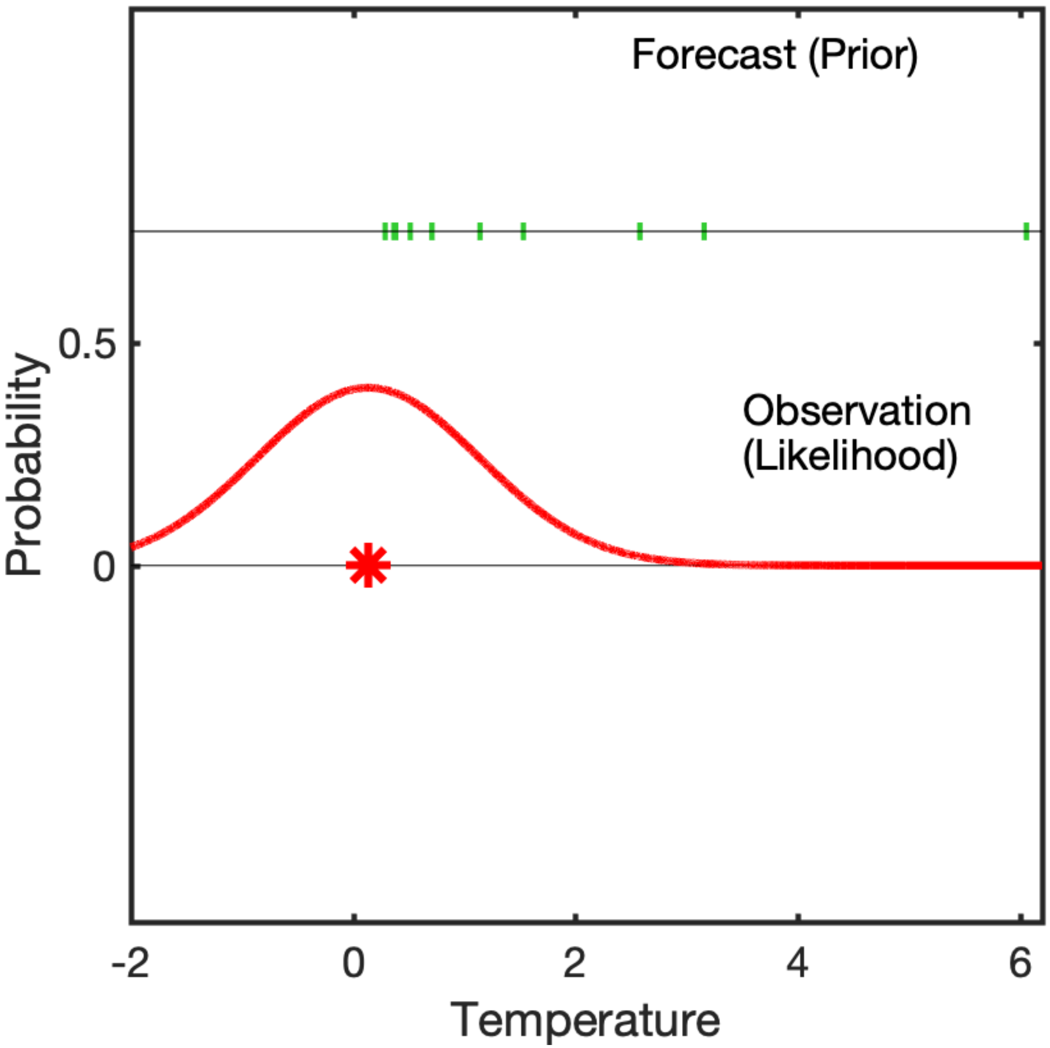


A : Forecast estimate from model (prior).

$p(B/A)$: Observation likelihood.

$p(A/B)$: Analysis (posterior) estimate combines A and B.

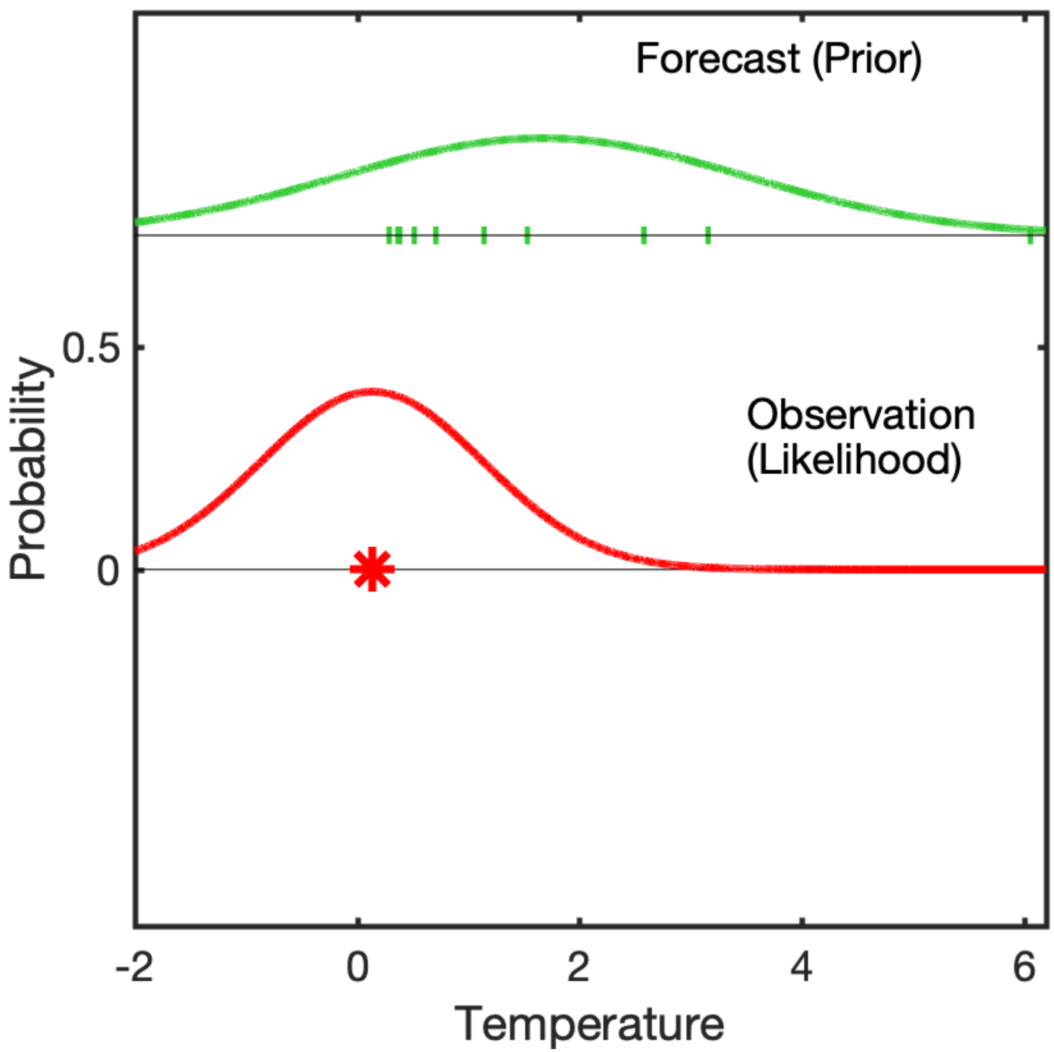
Should I Worry About Ice Going Down the Hill?



Have 10 forecasts of NCAR temperature.

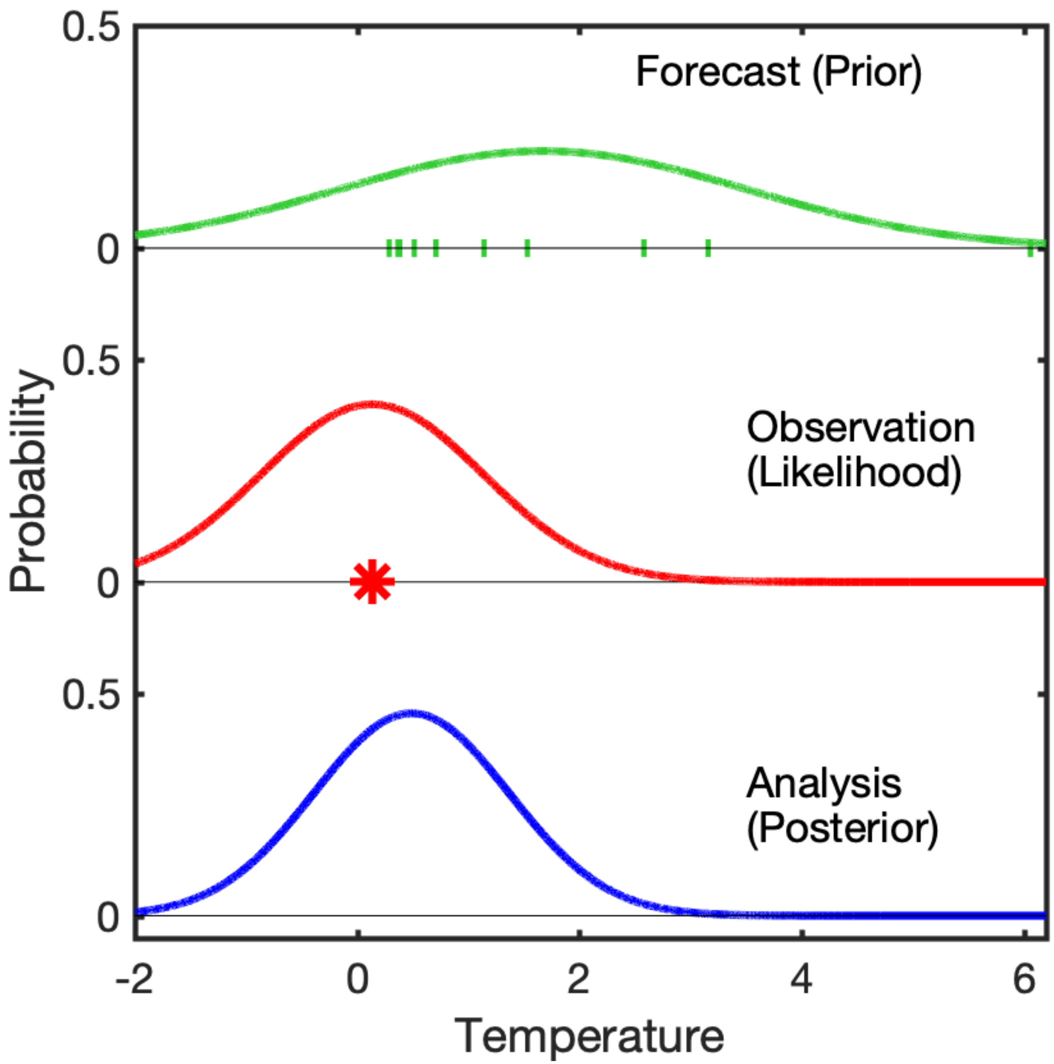
Use Bayes to combine with uncertain NCAR temperature observation.

Should I Worry About Ice Going Down the Hill?



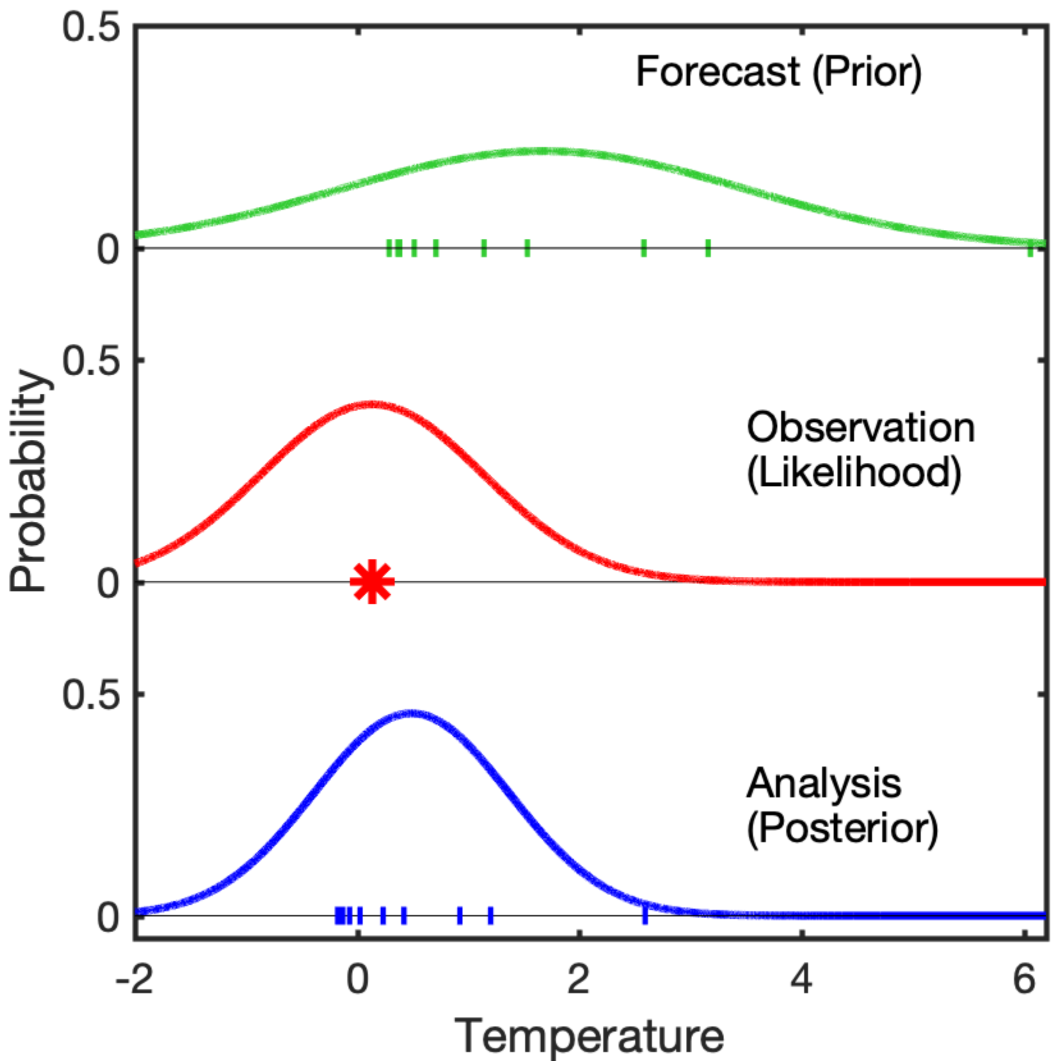
Original DART: Fit a normal to the forecast ensemble.

Should I Worry About Ice Going Down the Hill?



Bayes product gives continuous normal posterior.

Should I Worry About Ice Going Down the Hill?

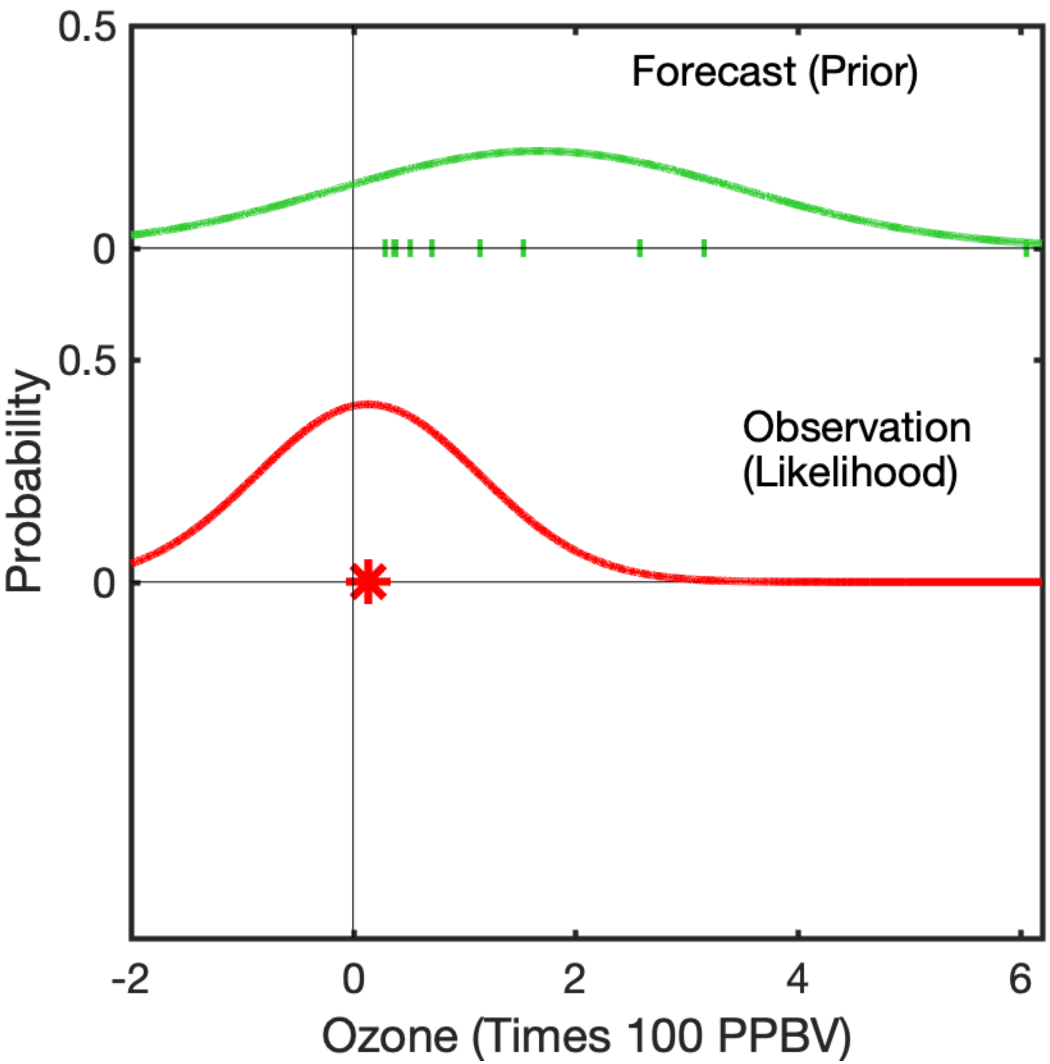


Get a posterior ensemble.

Until now, we only knew how to do this for normal distributions.

Normal may work okay for applications like NWP.

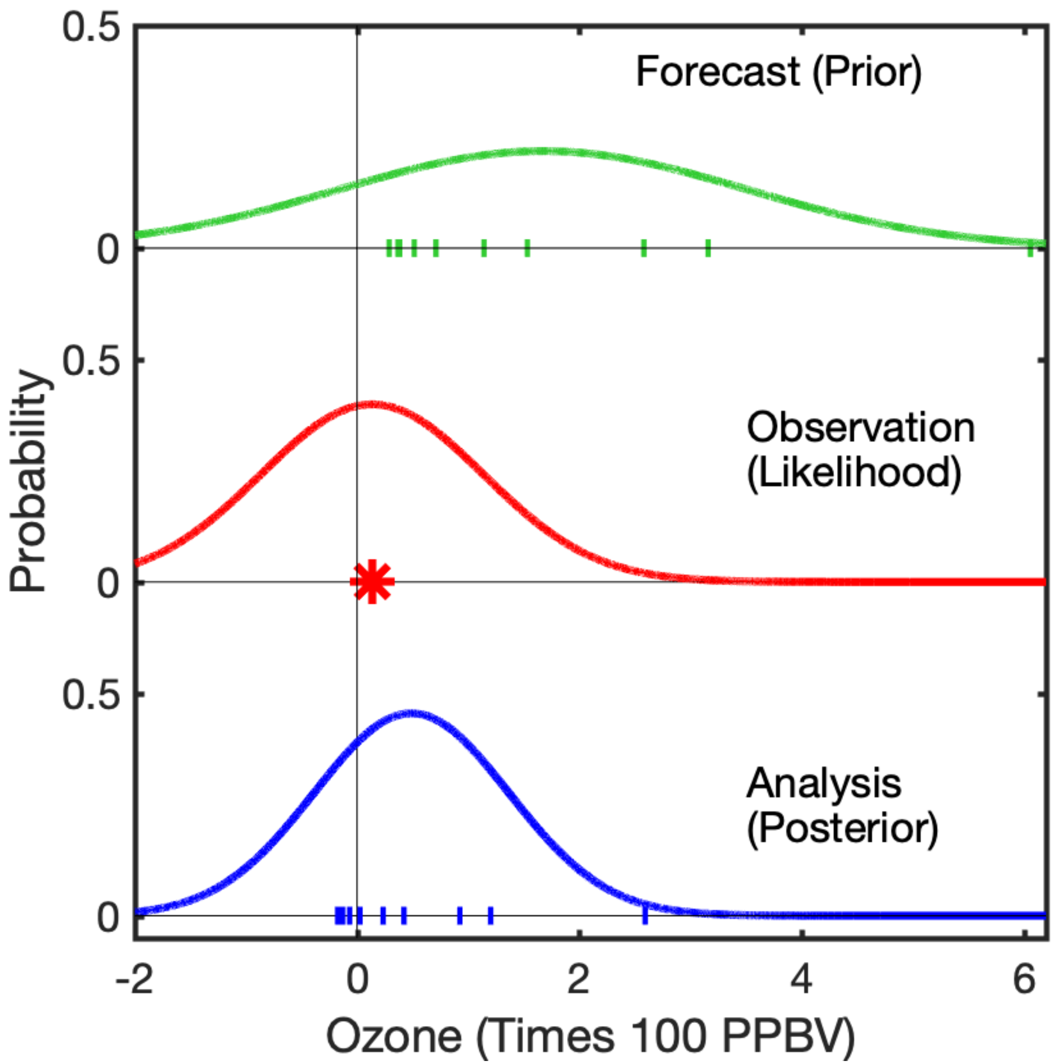
Should I Worry About Air Quality Going Down the Hill?



Forecast model knows ozone must be positive.

Fitting a normal leads to probability of negative.

Should I Worry About Air Quality Going Down the Hill?

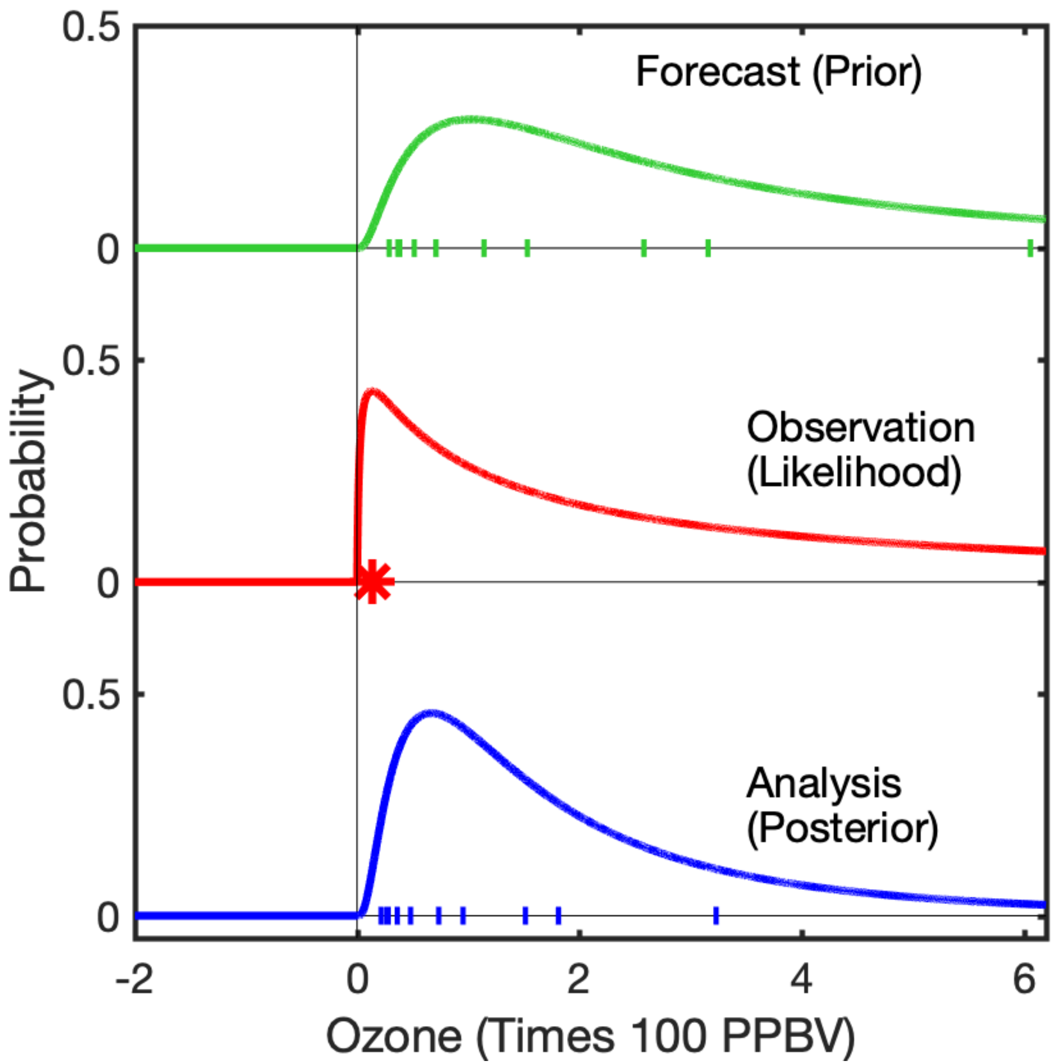


Doing the DA can lead to negative ensemble members.

What does that mean? Not sure, but nothing good.

Putting these back into model to make new forecasts is a problem, too.

Should I Worry About Air Quality Going Down the Hill?

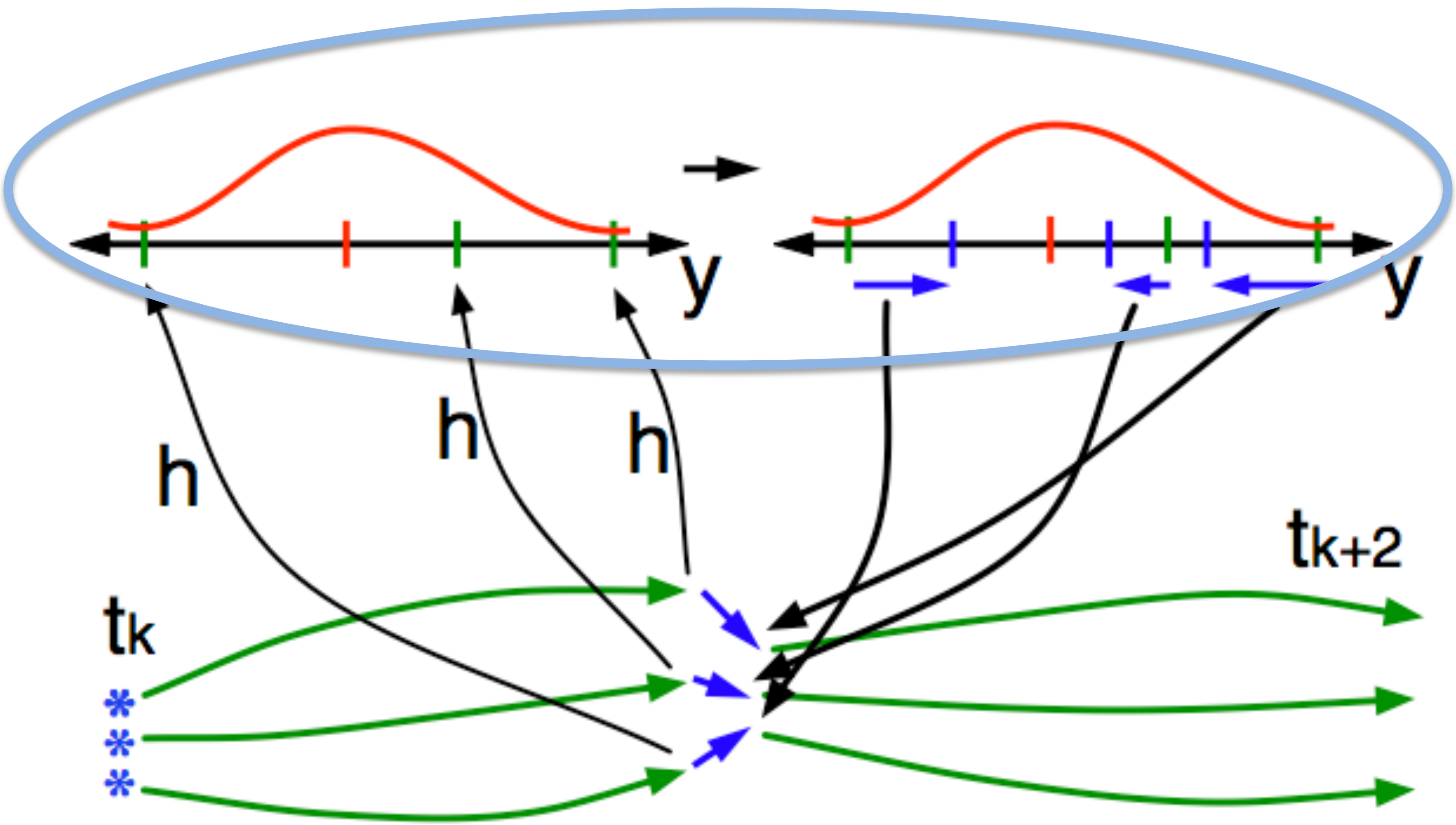


Now can do any distribution using quantile conserving ensemble algorithms.

Example: Gamma for bounded quantity like ozone.

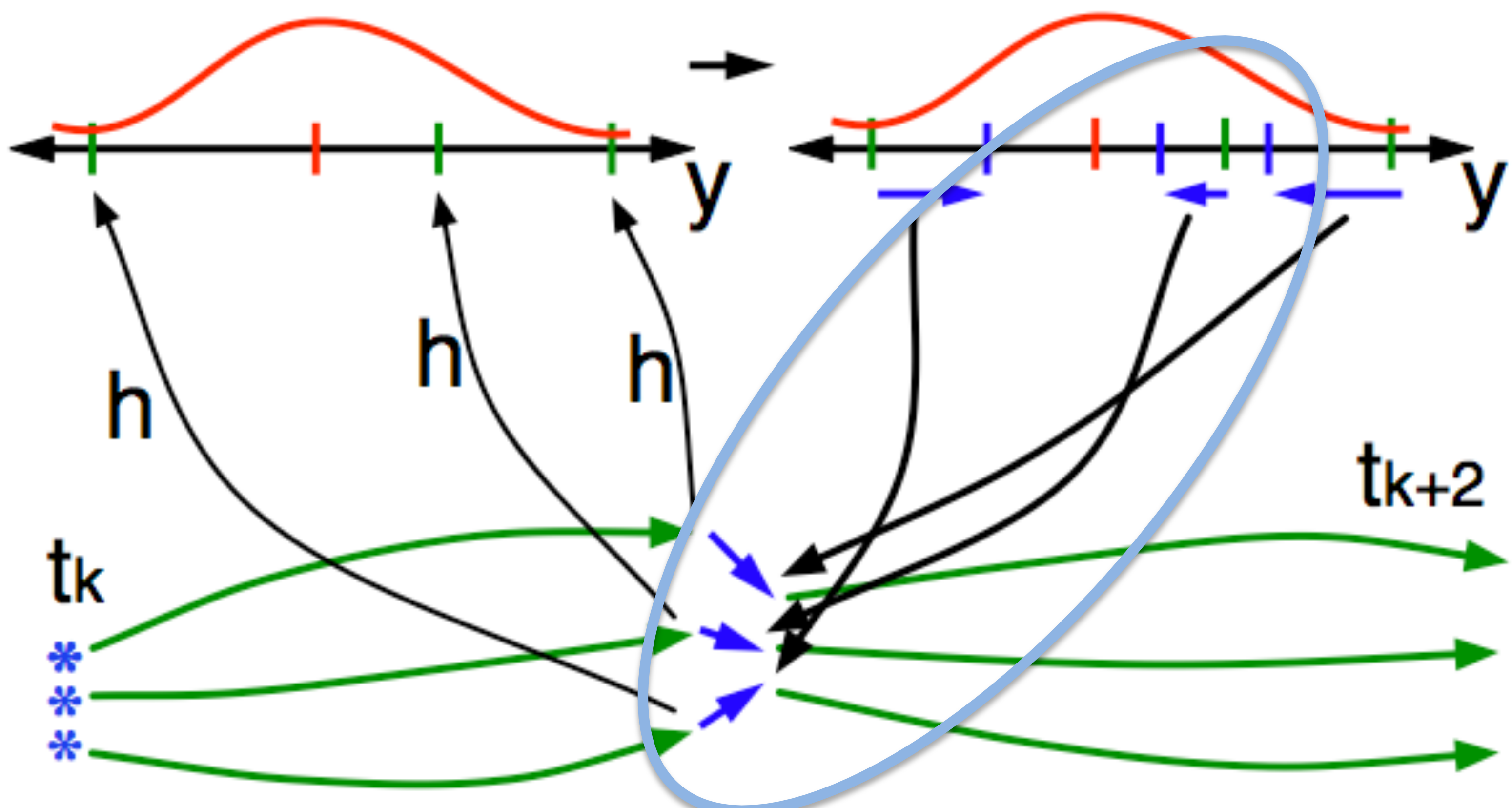
Posterior ensemble no longer crazy.

DART now provides nearly general solutions for this step
(Anderson, 2022, MWR *150*, 1061-1074).



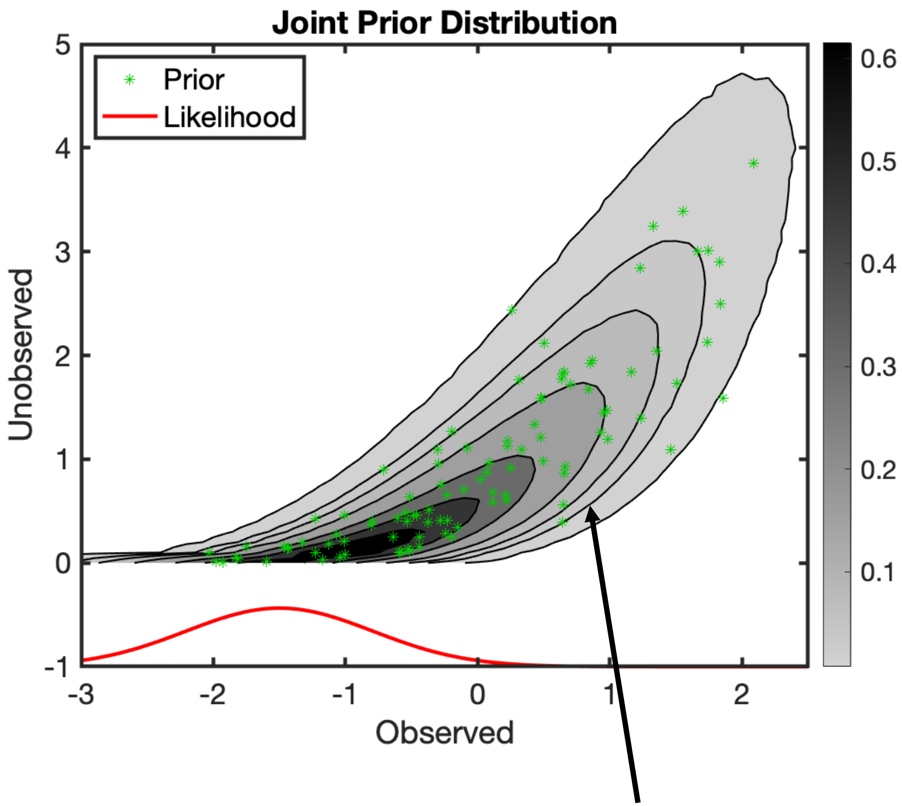
Novel Algorithm 2: Nonlinear constraint-preserving regression

Linear regression can cause inconsistent updates for other variables.



Problems with Linear Regression of Increments

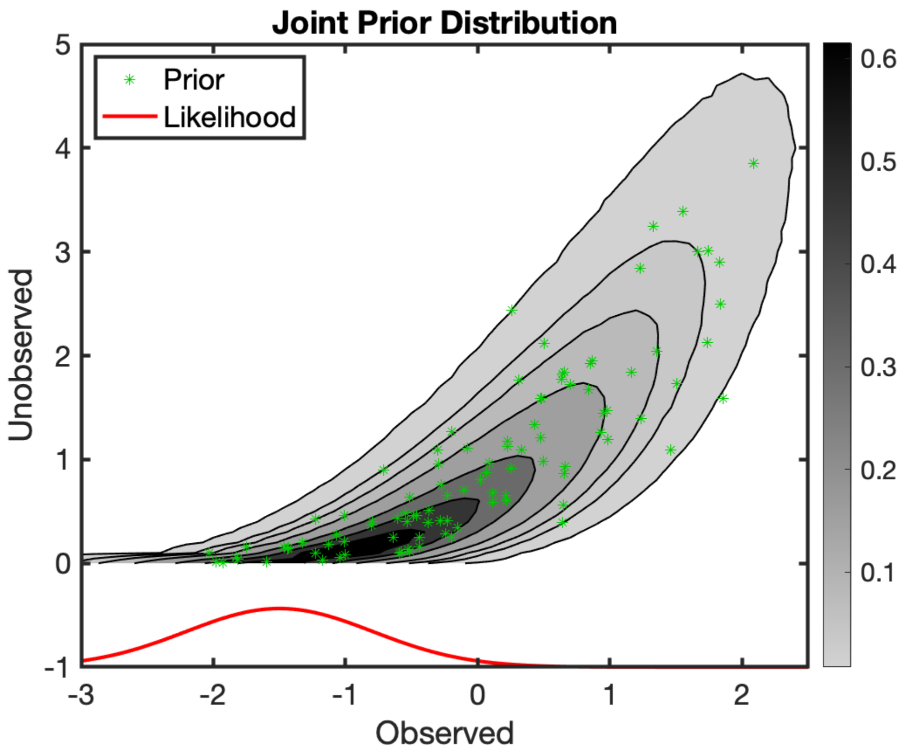
Prior for normal-gamma distribution with 100 member ensemble.



Contours are correct distribution.

Problems with Linear Regression of Increments

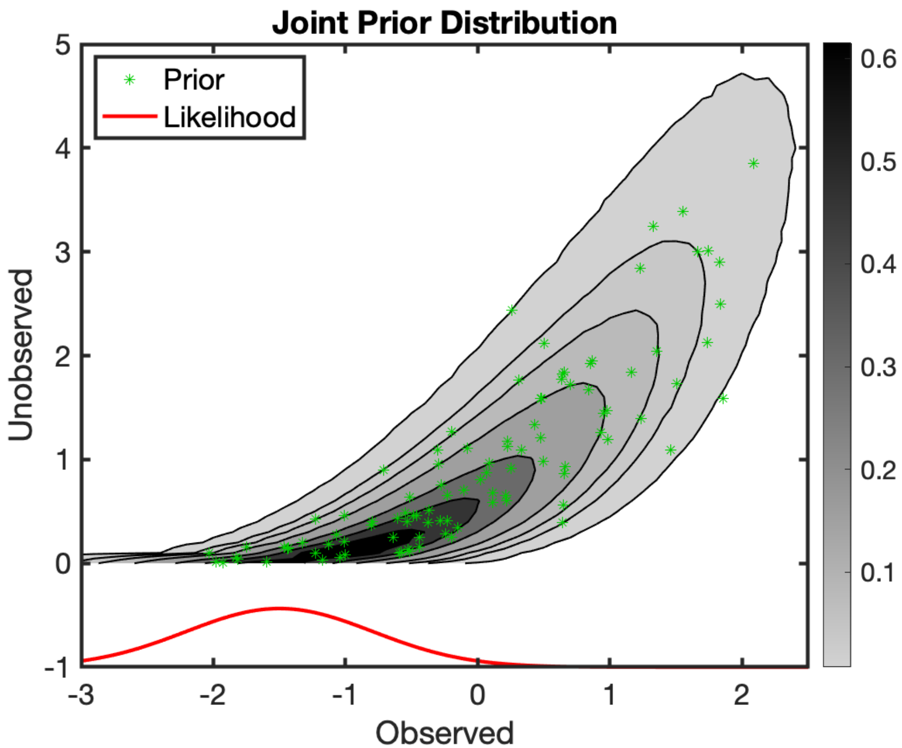
Prior for normal-gamma distribution with 100 member ensemble.



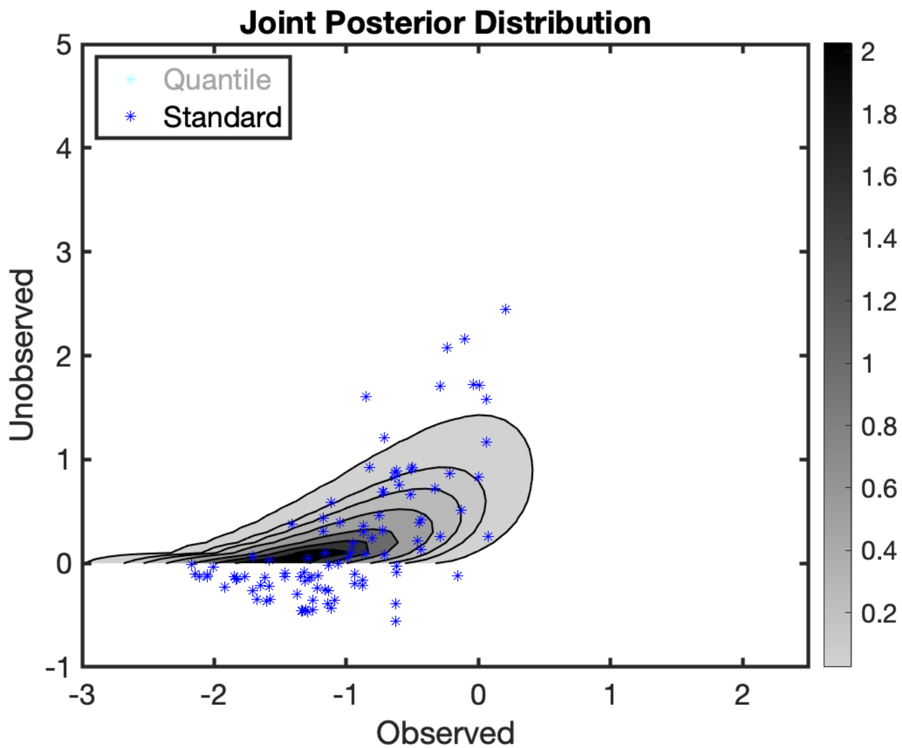
Example: Use observation of temperature to improve estimate of ozone.

Problems with Linear Regression of Increments

Prior for normal-gamma distribution with 100 member ensemble.



Posterior ensemble has problems.

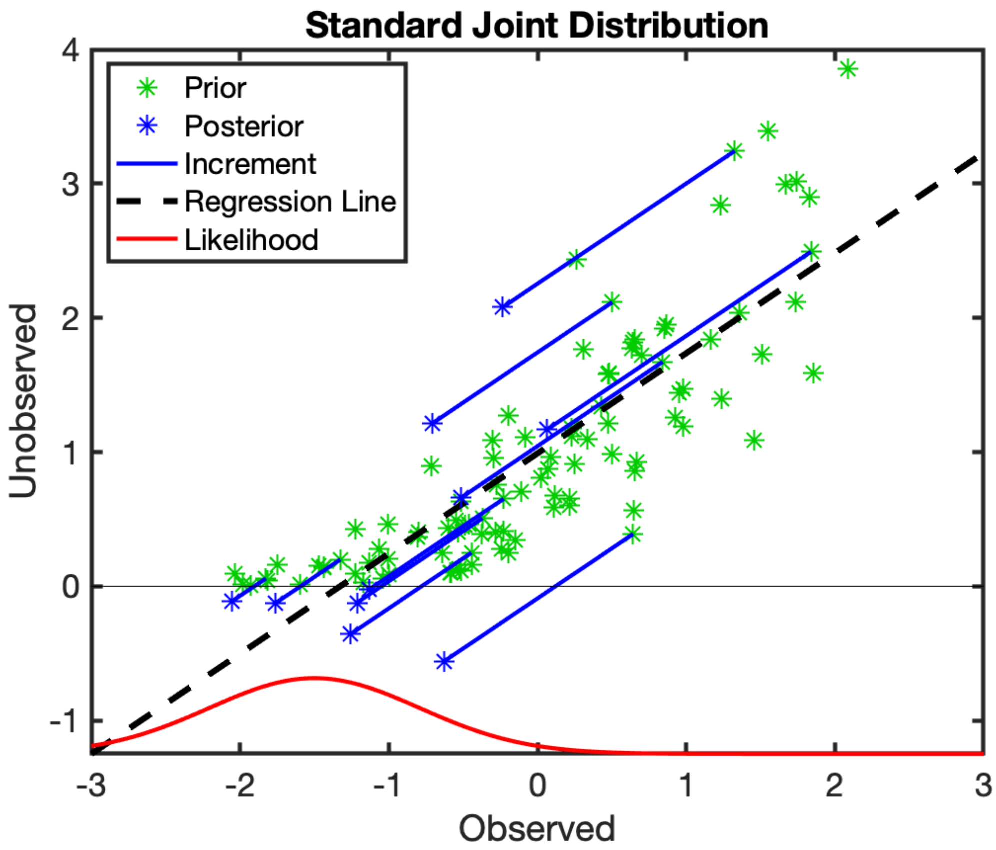


Problems with Linear Regression of Increments

Example regression increment vectors:

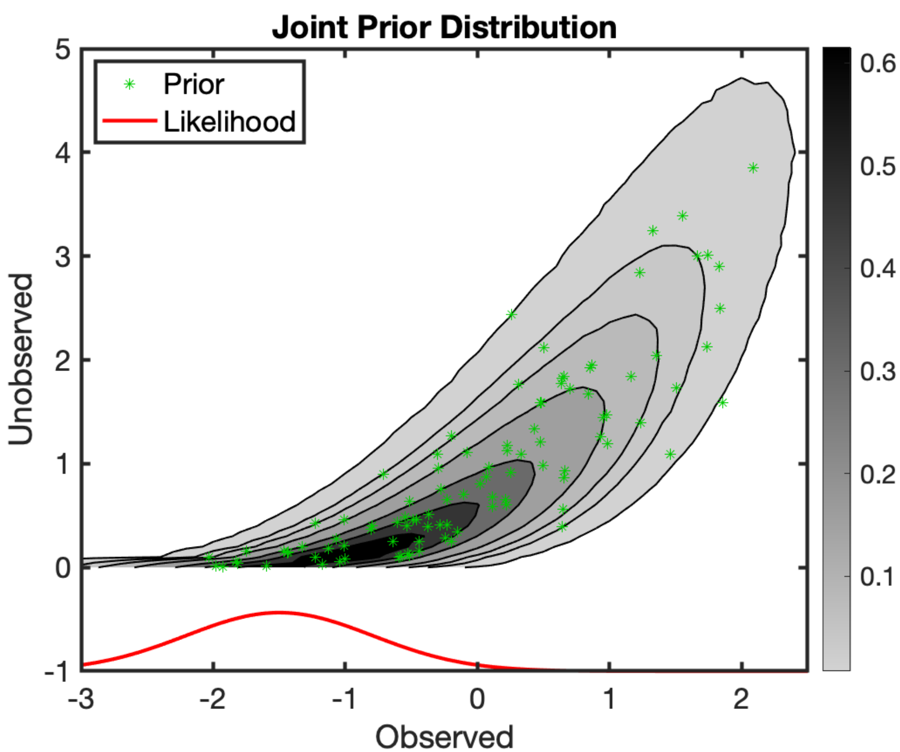
Don't respect bounds,

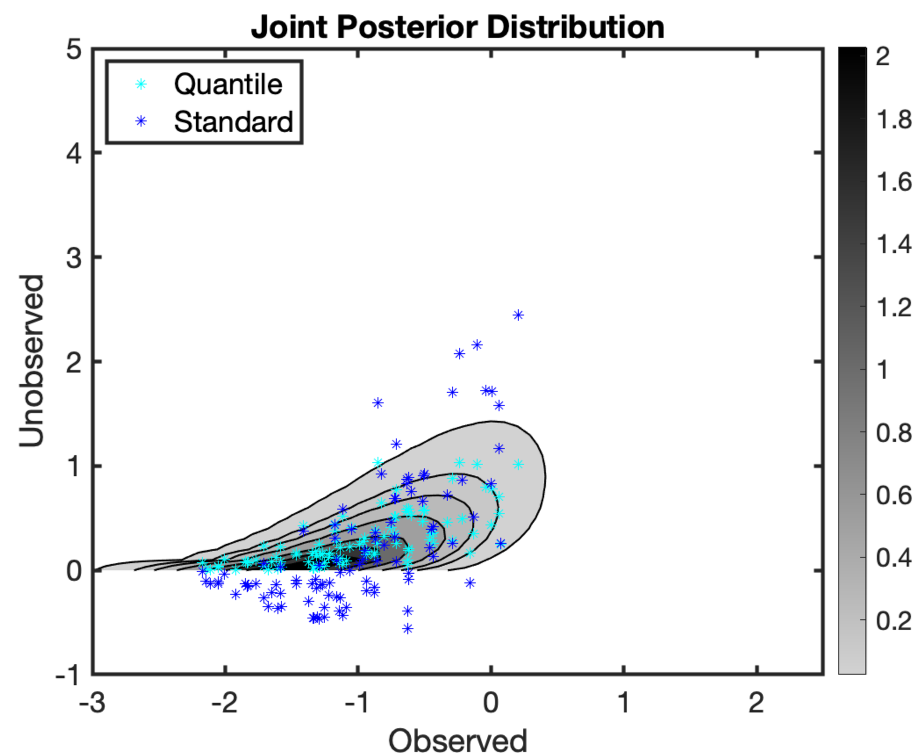
Struggle with nonlinearity.



Solution: Regression of Obs. Increments in Transformed Quantile Space

Doesn't violate prior PDF constraints like bounds.
Also deals with curvature (nonlinearity).





Crucially important for things like chemical tracers, streamflow, sea ice concentration, snow cover,…

Algorithm

y_n^p , y_n^a, x_n^p, *n=1, …N* are prior and posterior (analysis) ensembles of observed variable y and unobserved variable x

F_x^p and F_y^p are continuous CDFs appropriate for x and y

$\Phi(z)$ is the CDF of the standard normal, $\Phi^{-1}(p)$ is the probit function

$\tilde{x}_n^p = \Phi^{-1}\left[F_x^p\left(x_n^p\right)\right]$, $\tilde{y}_n^p = \Phi^{-1}\left[F_y^p\left(y_n^p\right)\right]$ and $\tilde{y}_n^a = \Phi^{-1}\left[F_y^p\left(y_n^a\right)\right]$ are probit space

$\Delta\tilde{y}_n = \tilde{y}_n^a - \tilde{y}_n^p$ is probit space observation increment

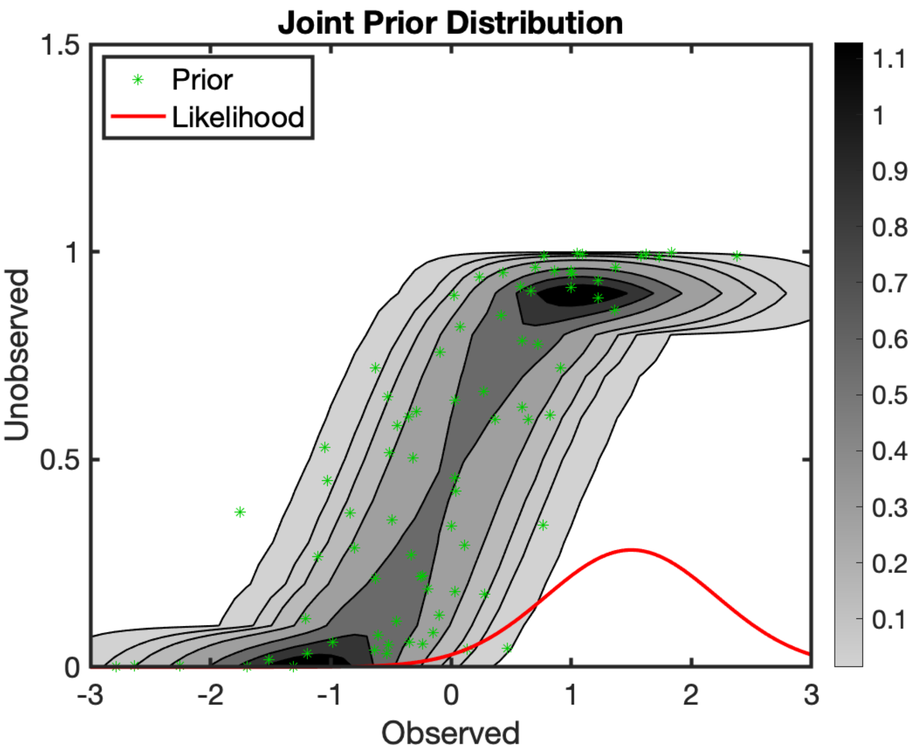
$\Delta\tilde{x}_n = \frac{\tilde{\sigma}_{x,y}}{\tilde{\sigma}_{y,y}}\Delta\tilde{y}_n$ regress increments in probit space (eq. 5 Anderson 2003)

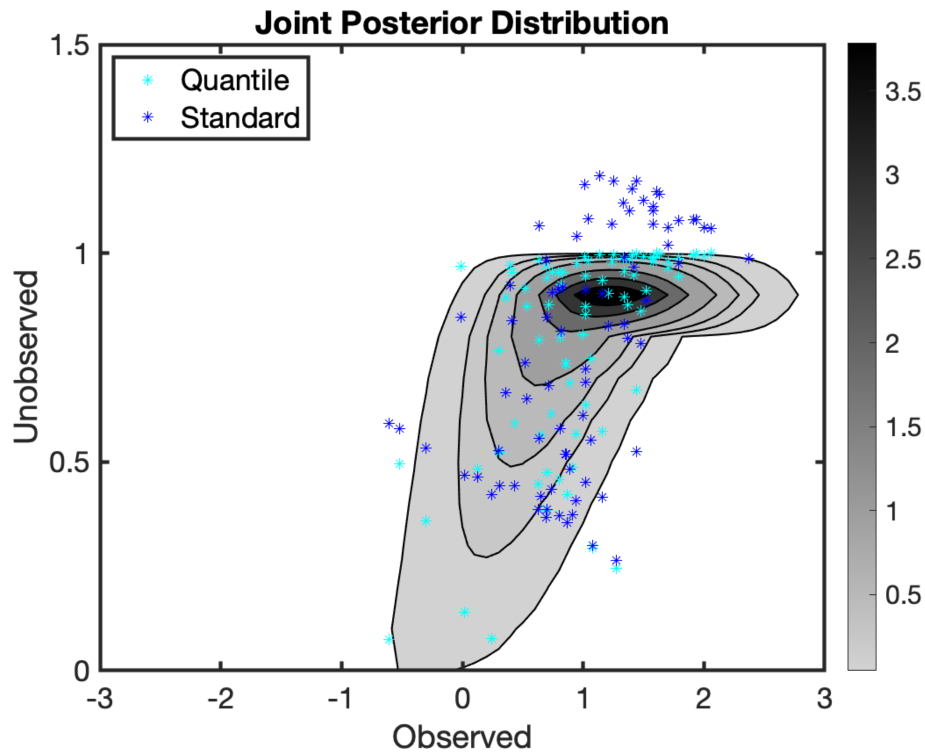
$\tilde{x}_n^a = \tilde{x}_n^p + \Delta\tilde{x}_n$ is posterior ensemble in probit space

$x_n^a = \left(F_x^p\right)^{-1}\left[\Phi\left(\tilde{x}_n^a\right)\right]$ is posterior ensemble

Example 2: Normal observed, beta unobserved

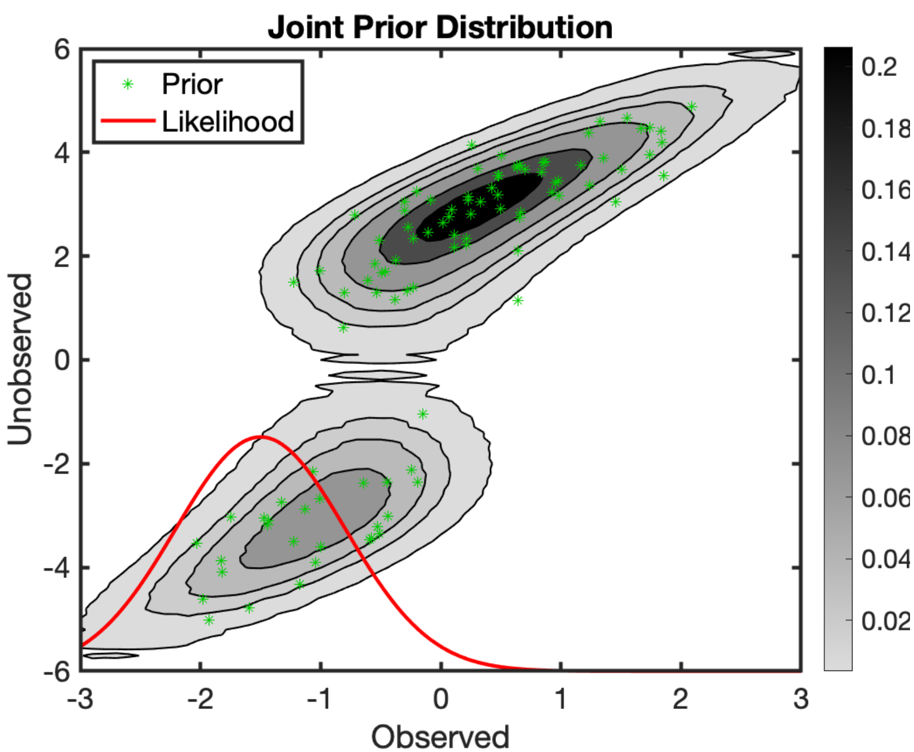
Application: Sea ice fraction, bounded between 0 and 1.

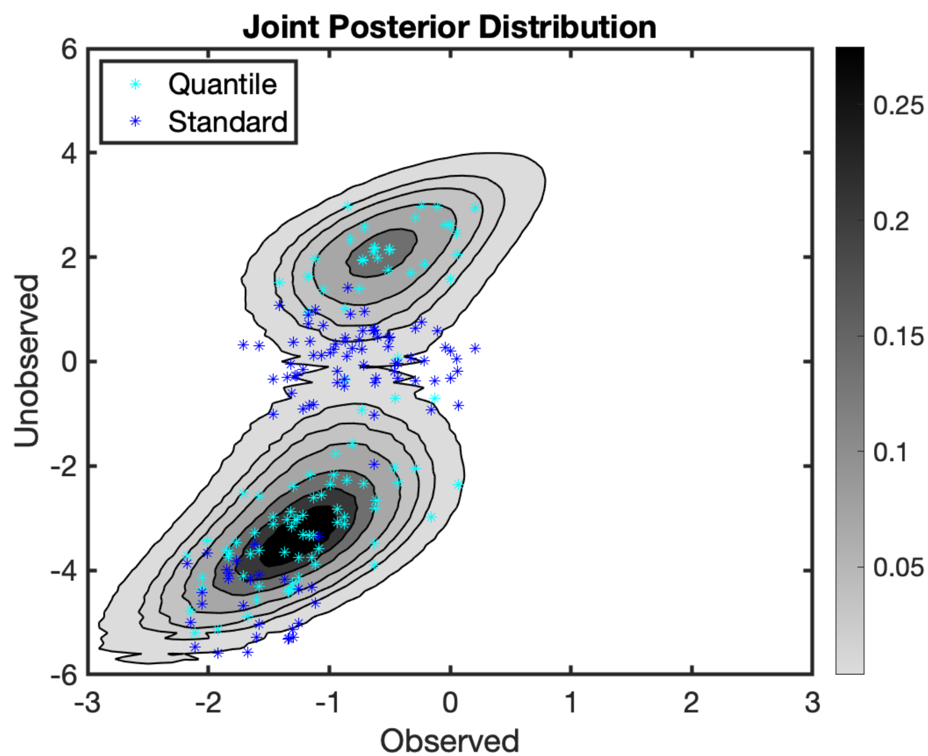




Example 3: Normal observed, binormal unobserved

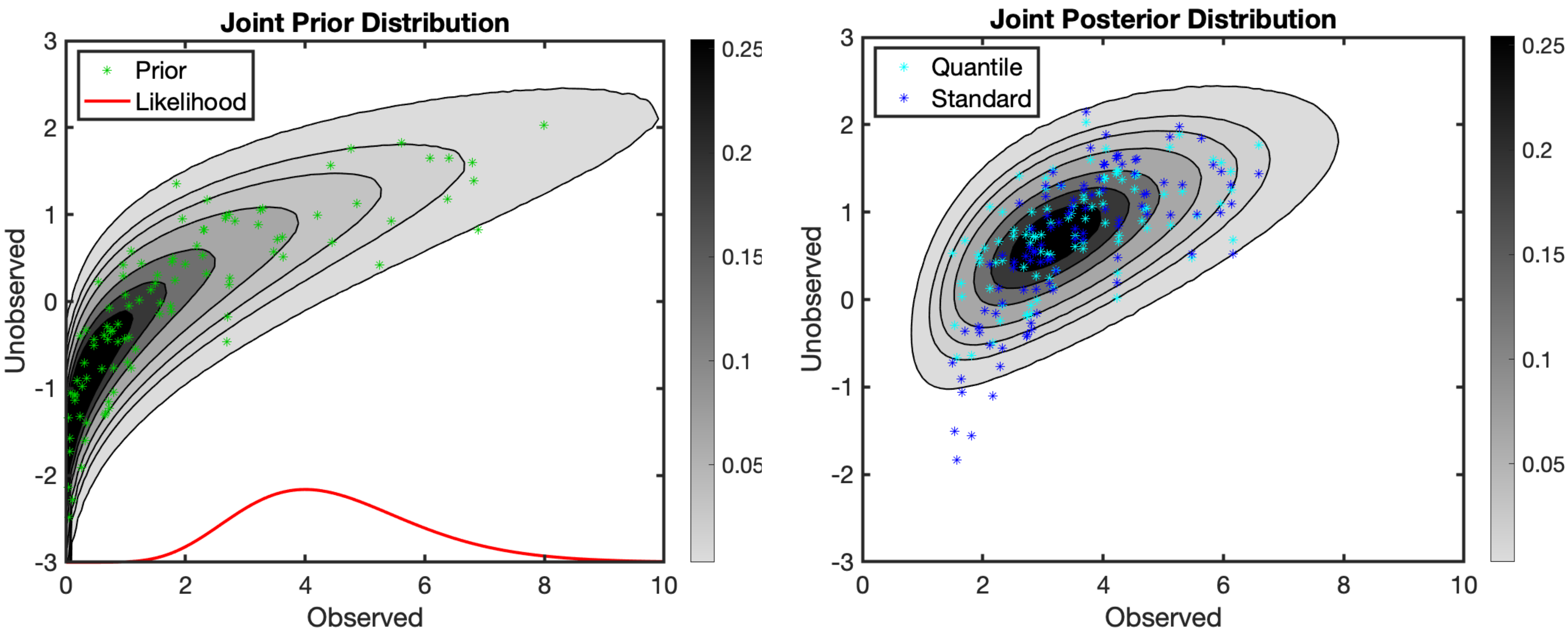
Application: Convection initiation. Either convection is occurring or it is not, partially convecting is not possible.





Example 4: Gamma observed, normal unobserved

Application: Impact of tracer observations on free atmosphere variables.



DART is Uniquely Able to Use These New Methods

- Works with DART's sequential ensemble algorithms.

- Compatible with existing DART parallel implementation.

- Works with all (dozens) of DART supported models.

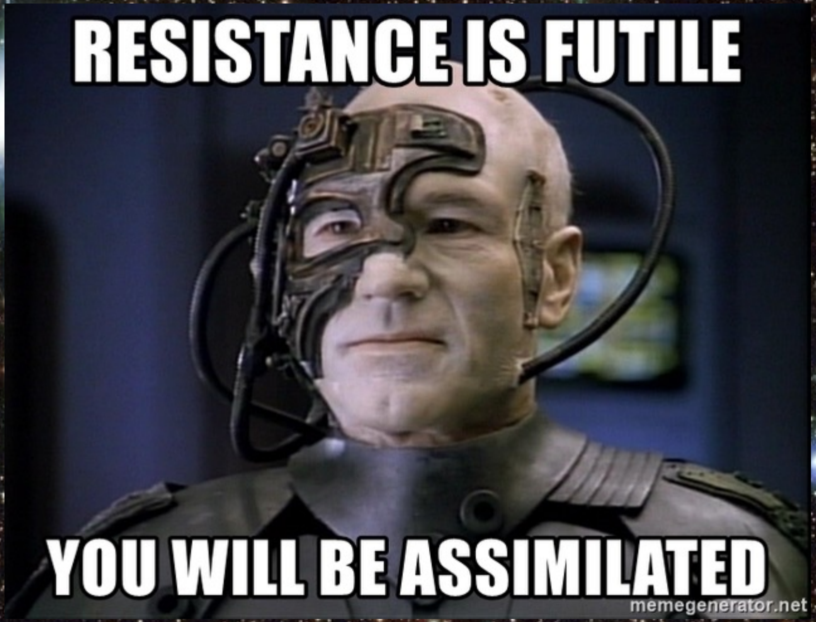
Important for Many High-Impact DA Science Collaborations

- Estimating and Predicting Bounded Quantities:
  - Atmospheric chemistry,
  - Streamflow and flooding,
  - Ocean biogeochemistry,
  - Sea ice (ASP Postdoc Chris Riedel already pushing forward),
  - Snow and land ice,
  - Land surface and biosphere,
  - Source and sink estimation,
    CO2, pollutants,
    Accidental/intentional releases,
  - Model parameter estimation.

- Non-Gaussian distributions:
  - Convection,
  - Radiation remote sensing.

DART: The Next Generation

Breakthroughs in ensemble DA algorithms being implemented in DART provide powerful and unique nonlinear and non-Gaussian capabilities for Earth system applications.



Major improvements for:
Tracers / bounded quantities;
Remote sensing observations;
Model parameter estimation.

DAReS is looking forward to an exciting and busy future accelerating science progress with these powerful new methods.

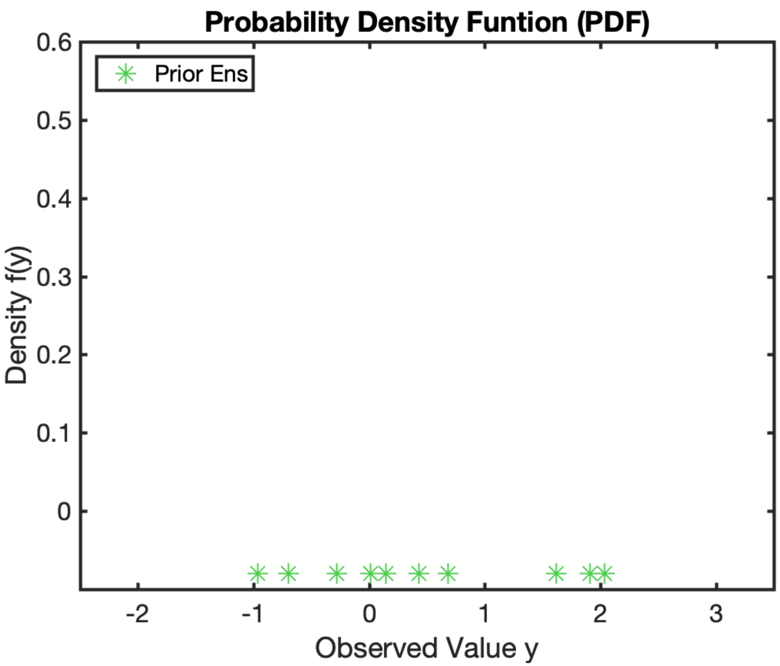
New website!

https://dart.ucar.edu

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., Arellano, A., 2009: *The Data Assimilation Research Testbed: A community facility.* BAMS, **90**, 1283—1296, doi: 10.1175/2009BAMS2618.1

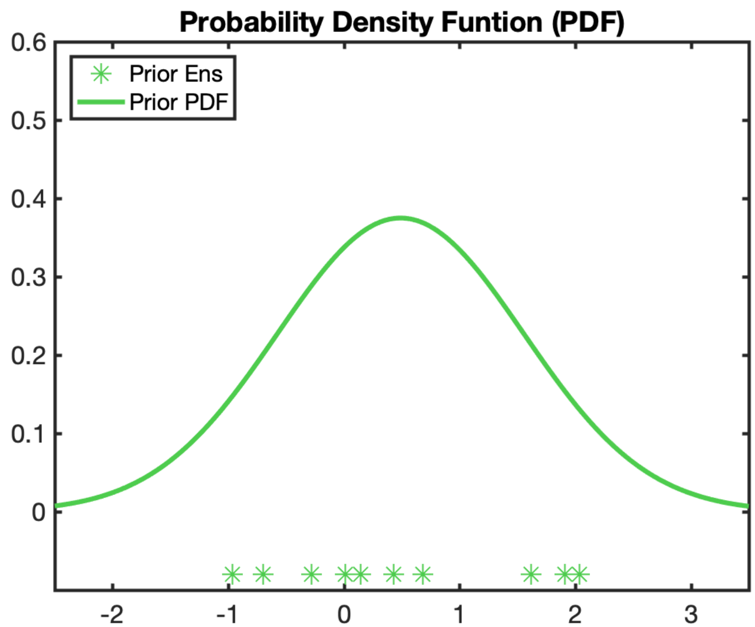
Application 1: Bayesian filtering for an observed variable

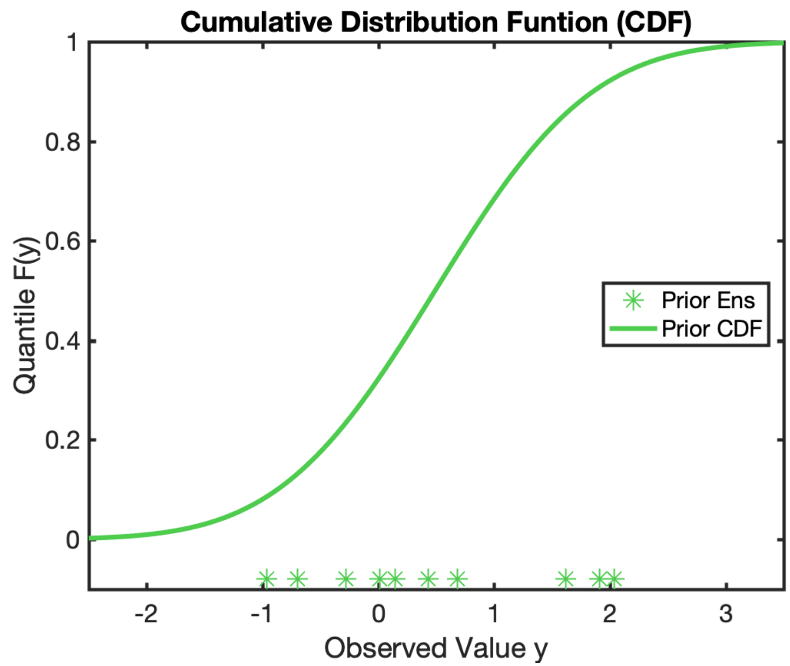
Given a prior ensemble estimate of an observed quantity, y



Application 1: Bayesian filtering for an observed variable

Fit a continuous PDF from an appropriate distribution family and find the corresponding CDF

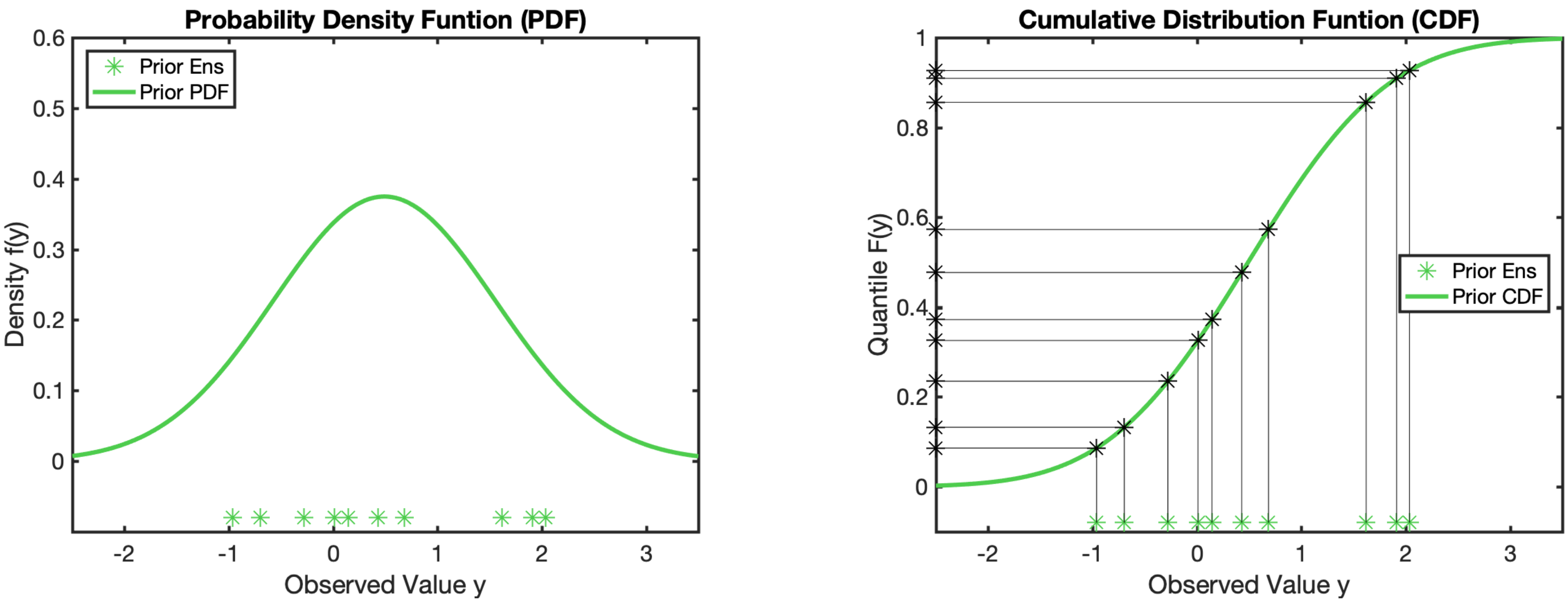




This example uses a normal PDF

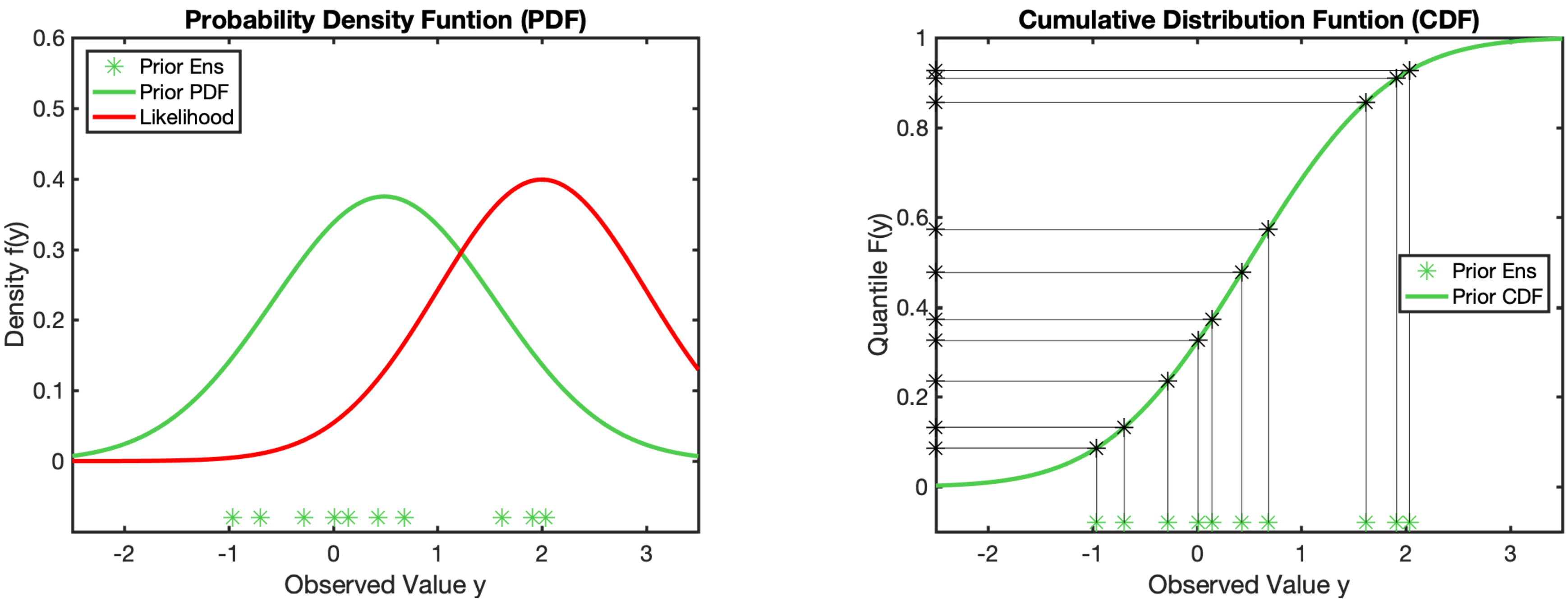
Application 1: Bayesian filtering for an observed variable

Compute the quantile of ensemble members;
just the value of CDF evaluated for each member.



This example uses a normal PDF

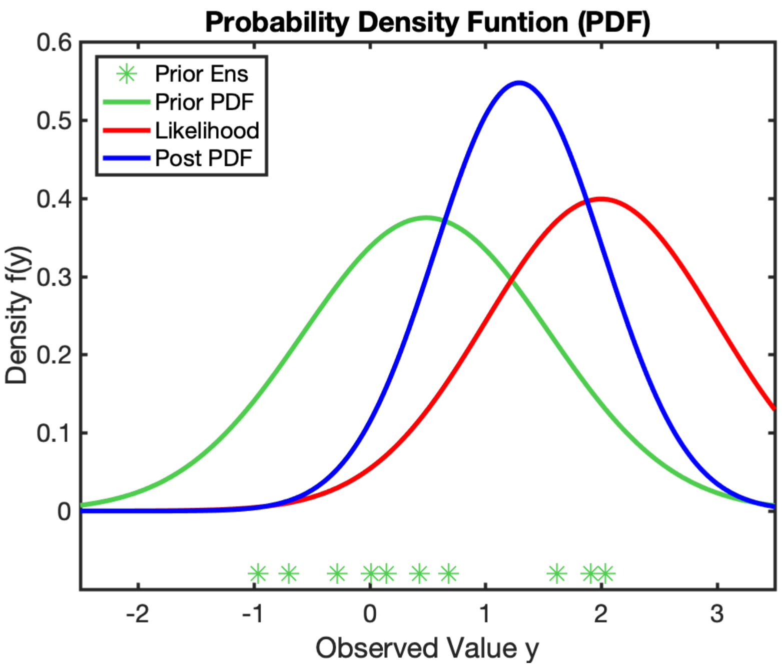
Continuous likelihood for this observation.

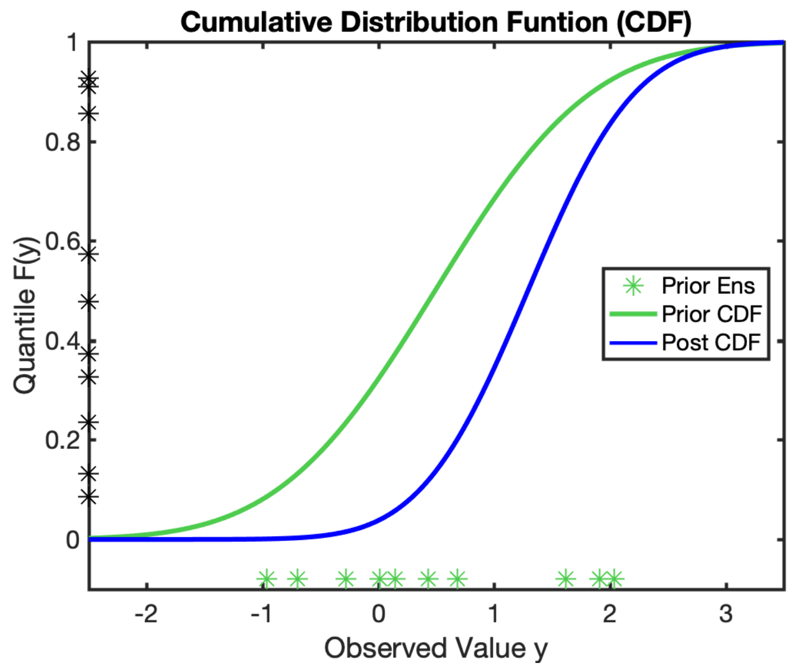


This example uses a normal PDF

Application 1: Bayesian filtering for an observed variable

Bayes tells us that the continuous posterior PDF is the product of the continuous likelihood and prior.

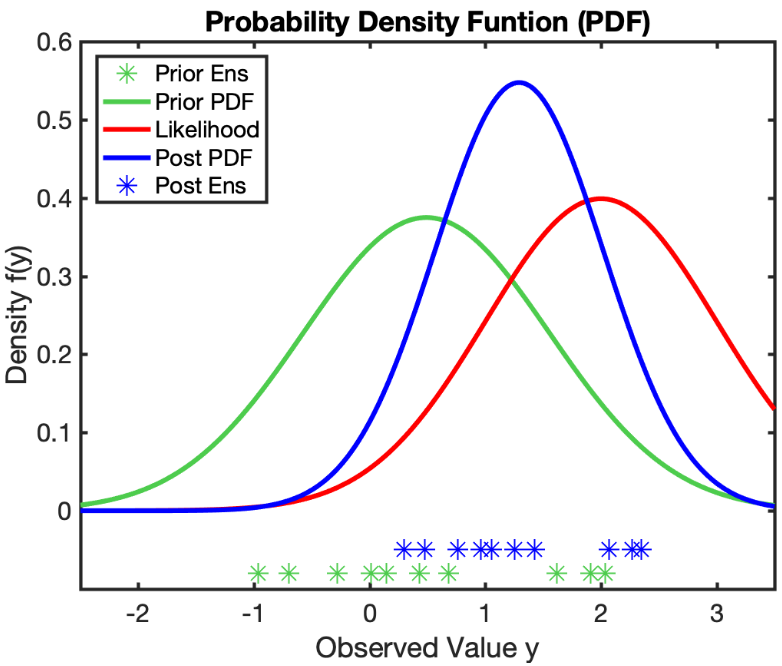


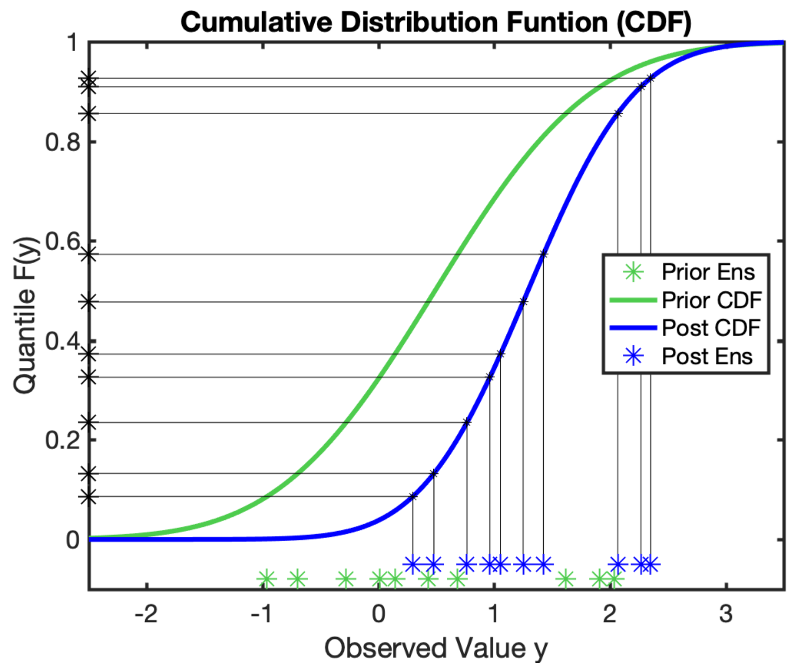


Normal times normal is normal.

Application 1: Bayesian filtering for an observed variable

Posterior ensemble members have same quantiles as prior.
This is quantile function, inverse of posterior CDF.

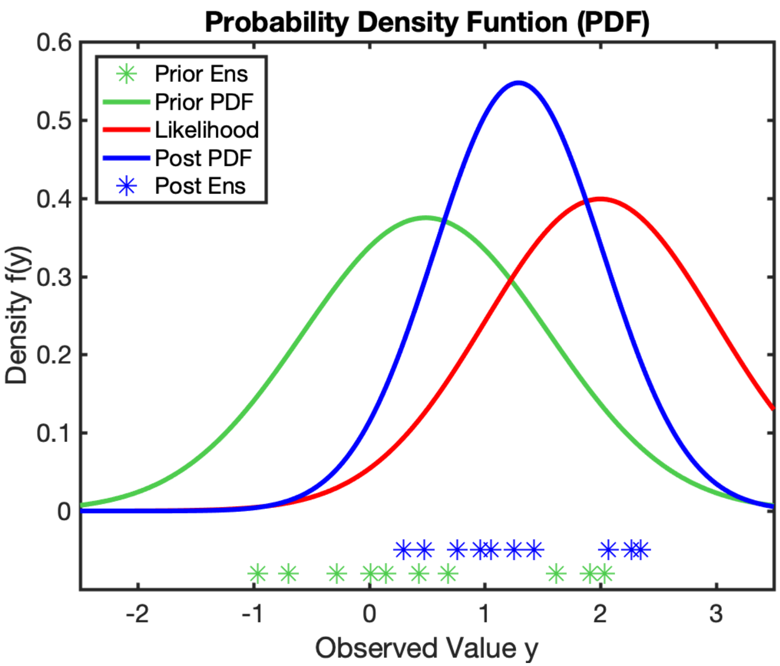


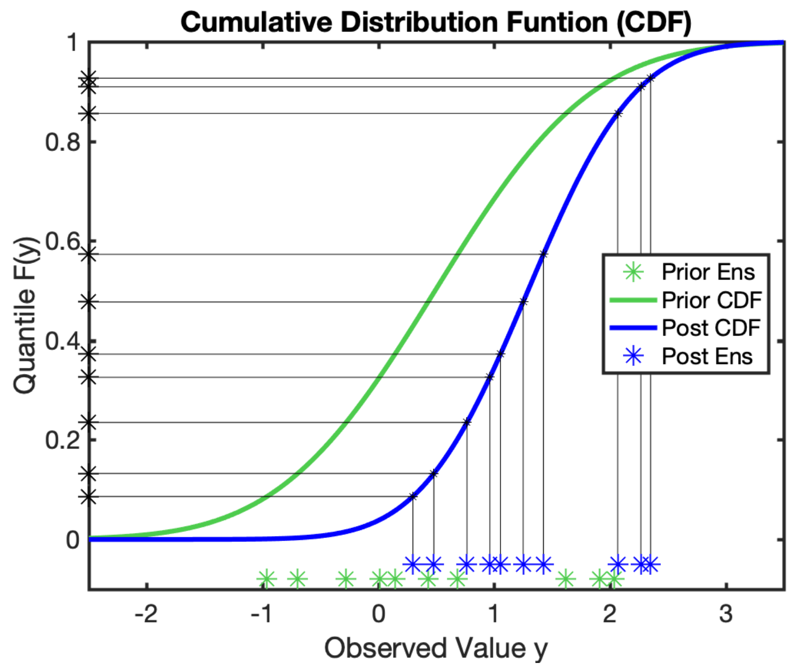


This example uses a normal PDF

Application 1: Bayesian filtering for an observed variable

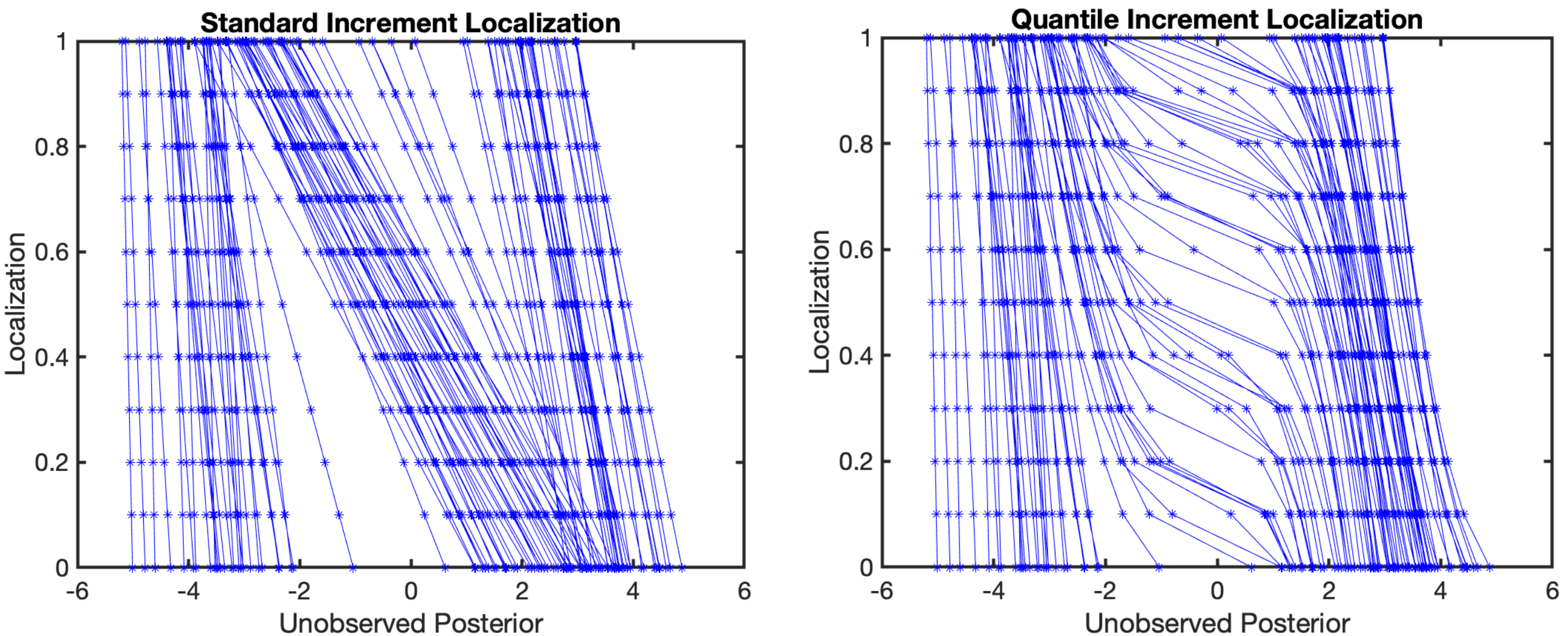
For normal prior and likelihood, this is identical to existing deterministic Ensemble Adjustment Kalman Filter (EAKF)





Localization of Quantile (or Probit) Increments: Normal-binormal example

Standard increment localization may ignore prior constraints.
Quantile increment localization 'knows' prior was binormal.



What about the normal-normal case?

$\tilde{y}_n^p = \Phi^{-1}\left[F_y^p\left(y_n^p\right)\right]$ and $\tilde{y}_n^a = \Phi^{-1}\left[F_y^p\left(y_n^a\right)\right]$

F_y^p is for $Normal\left[\mu_y^p, \left(\sigma_y^p\right)^2\right]$

This means that $y_n^p = \sigma_y^p \tilde{y}_n^p + \mu_y^p, \quad y_n^a = \sigma_y^p \tilde{y}_n^a + \mu_y^p$

Similarly $x_n^p = \sigma_x^p \tilde{x}_n^p + \mu_x^p$

Differences are $\Delta y_n = \sigma_y^p \Delta\tilde{y}_n$ and $\Delta x_n = \sigma_x^p \Delta\tilde{x}_n$

Covariance $\sigma_{x,y} = \sigma_x^p \sigma_y^p \tilde{\sigma}_{x,y}$

Variance $\sigma_{y,y} = \sigma_y^p \sigma_y^p \tilde{\sigma}_{y,y}$

The standard increment regression is $\Delta x_n = \frac{\sigma_{x,y}}{\sigma_{y,y}} \Delta y_n$

Substituting for all terms $\sigma_x^p \Delta\tilde{x}_n = \frac{\sigma_x^p \sigma_y^p \tilde{\sigma}_{x,y}}{\sigma_y^p \sigma_y^p \tilde{\sigma}_{y,y}} \sigma_y^y \Delta\tilde{y}_n$

Cancels to give $\Delta\tilde{x}_n = \frac{\tilde{\sigma}_{x,y}}{\tilde{\sigma}_{y,y}} \Delta\tilde{y}_n$

What about the normal-normal case?

Computing increments in regular space is equivalent to computing increments in probit space.

For normal-normal, just do what we have always done (and for any normal in a bivariate pair???).

Recall that the QCEFF normal filter in observation space is equivalent to our traditional EAKF in observation space.

Similarly, the method here is identical to the EAKF for unobserved updates.

The EAKF is equivalent to the Kalman Filter for normal/Gaussian cases.

The QCEFF normal combined with regression here is an ensemble generalization of the EAKF and the Kalman filter.

Caveat: quantile increment localization is NOT the same as standard increment localization even in the normal-normal case.

Computational Cost: An Efficient Workflow

Inverting quantile functions can be expensive.
As described, have to do that for each observation/unobs pair.

More efficient workflows are possible; stay in probit space:

1. Compute all forward operators for a window.
2. Compute prior probit for all joint state ensembles.
3. Loop through observations (sequential_obs loop):
    a. Invert prior probit for observed variable,
    b. Compute posterior for observed,
    c. Do probit conversion for posterior obs,
    d. Do regression of obs probit increments for each joint unobs,
4. Back to regular space for all state only at the end of all obs

Number of quantile/probit inversions down to one per extended state.