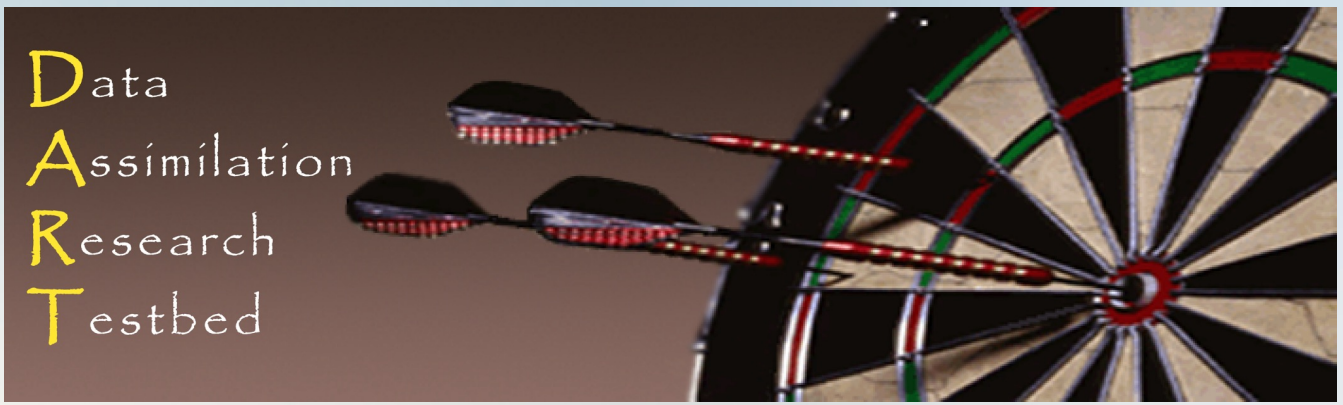


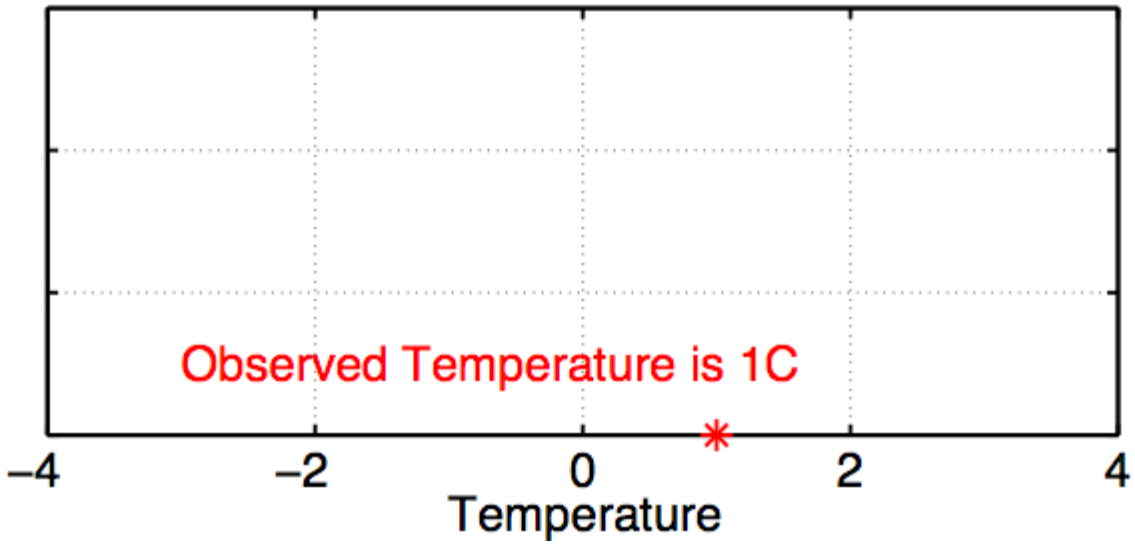
DART_LAB Tutorial Section 1: Ensemble Data Assimilation Concepts in 1D

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Example: Estimating the Temperature Outside

An observation has a value T_O(*), what the instrument measured.

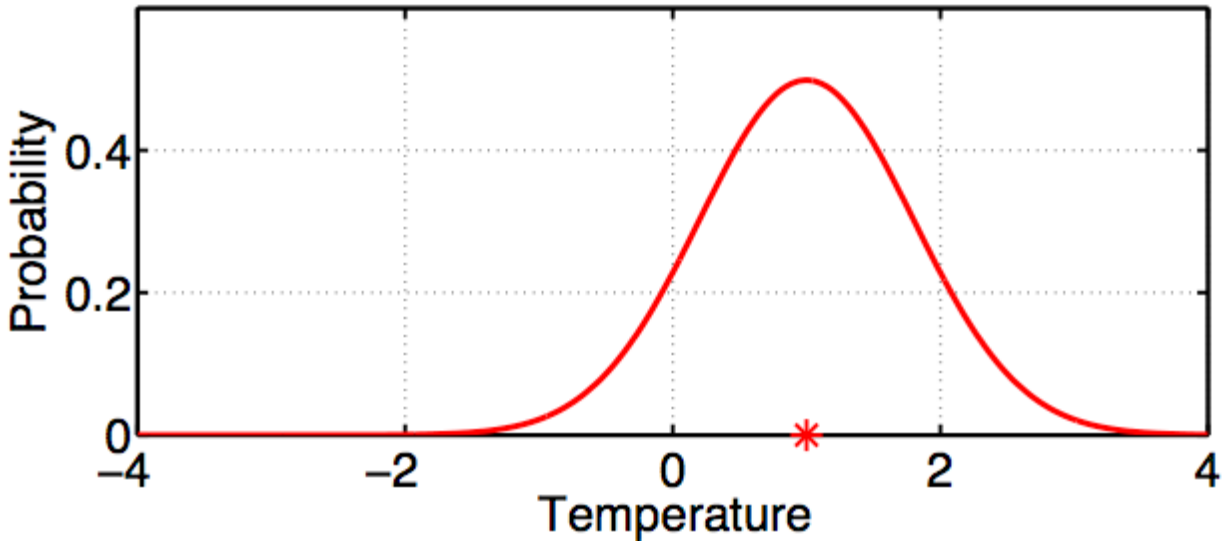
Without additional information this is meaningless.



Example: Estimating the Temperature Outside

An observation has a value T_O(*), what the instrument measured.

Without additional information this is meaningless.

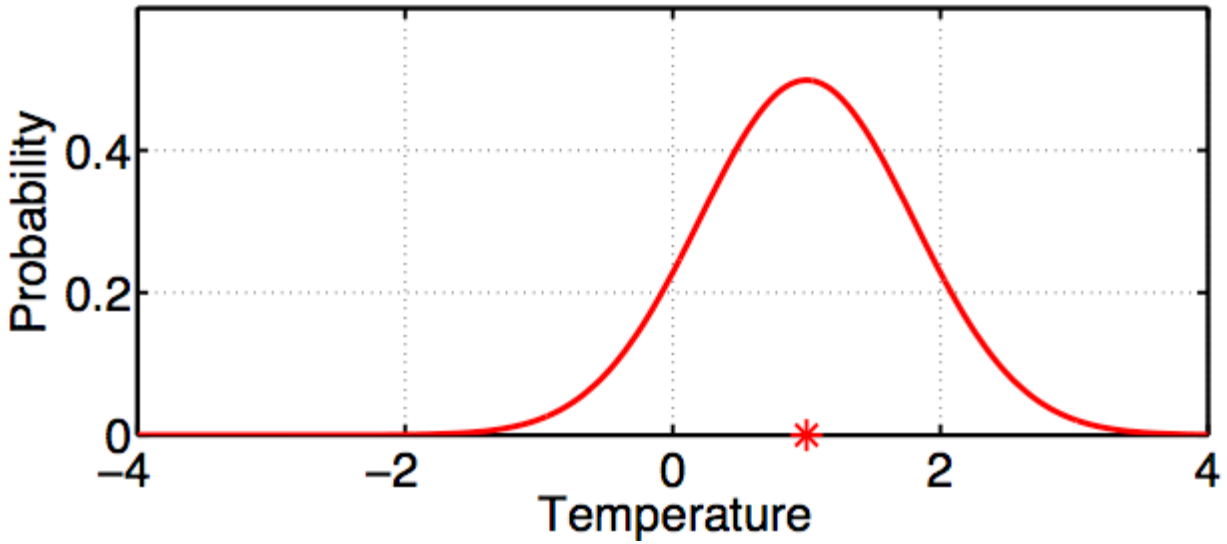


The additional information we need is a likelihood function.

Example: Estimating the Temperature Outside

An observation has a value T_O(*), what the instrument measured.

Without additional information this is meaningless.



The additional information we need is a likelihood function.

$$L(T) = P(T_O | T_{True} = T)$$

This is the relative probability that the actual temperature is T given that the instrument observed T_O.

Bayes' Theorem

Simplest form (and some notation):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Read as: "Probability of A given B is equal to the probability of B given A times the probability of A normalized by the probability of B.

Easily derived from basic definition of conditional probability:

$$P(A|B) = \frac{P(A,B)}{P(B)} \text{ and } P(B|A) = \frac{P(A,B)}{P(A)}$$

Where $P(A,B)$ represents the probability of A and B.

Bayes' Theorem

Simplest form (and some notation):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Read as: "Probability of A given B is equal to the probability of B given A times the probability of A normalized by the probability of B.

Note that statisticians might be concerned by using this discrete formulation for the continuous random variables we'll be discussing, but it all works out…

Bayes' Theorem

Simplest form (and some notation):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

We have the case where A is the temperature outside, T, and B is the observed value.

$P(T|T_O) = \frac{P(T_O|T)P(T)}{P(T_O)}$ where $P(T_O|T)$ is a more concise way of writing the likelihood.

Example: Estimating the Temperature Outside

Instrument builders know about the observation error associated with a measurement, say the thermometer is unbiased with +/- 0.8^{o} C Gaussian error.

Example: Estimating the Temperature Outside

Instrument builders know about the observation error associated with a measurement.

Define the observation error as $E = T_o - T_{True}$

Then the observation error distribution is $P(E)$.

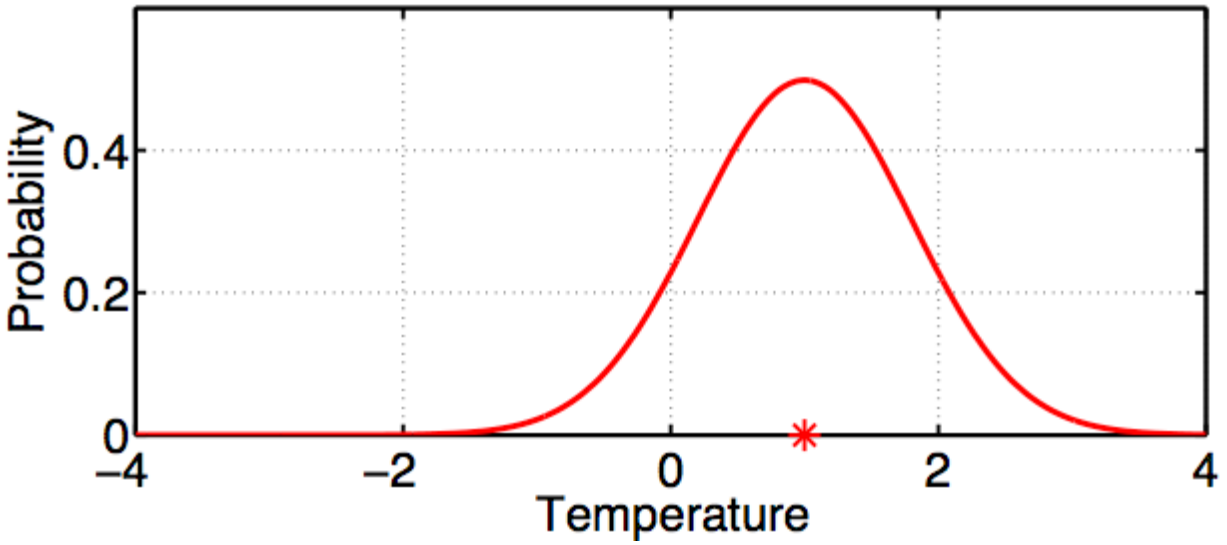
It is common for observation error to be approximated by a normal distribution with zero mean,

$$P(E) = Normal(0, \sigma_O^2)$$

Standard deviation σ_O.

Example: Estimating the Temperature Outside

The likelihood function and observation function are not the same thing.



If the error distribution is $Normal(0, \sigma_O^2)$ and the observed value is T_O, then the likelihood is $Normal(T_O, \sigma_O^2)$.

Example: Estimating the Temperature Outside

The likelihood function and observation function are not the same thing.

Be careful when the error isn't a simple normal as the relation between the observational error distribution and likelihood is more complex. (Watch out for gammas/inverse gammas for instance).

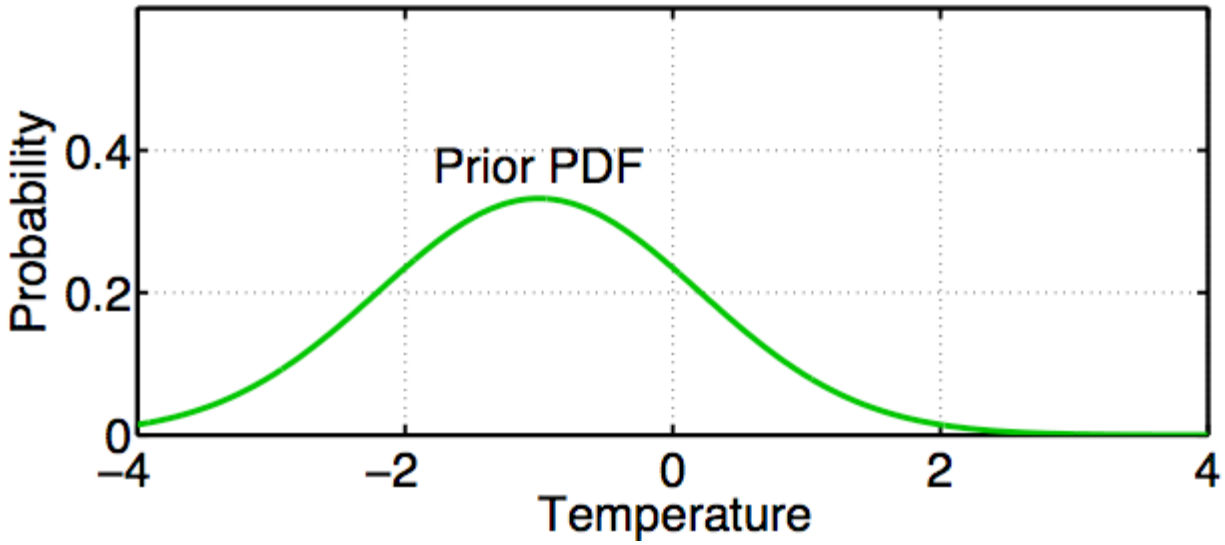
Example: Estimating the Temperature Outside

Suppose we have a second observation of the temperature. We are interested in the probability distribution of the temperature given both observations, $T_{O,1}$ and $T_{O,2}$

We will refer to our estimate using the first observation as the prior, and we want to include information from the second observation.

Example: Estimating the Temperature Outside

Prior estimate of temperature from first observation.



The green curve is $P(T \mid T_{O,1})$.

In our example so far, this is the likelihood of the first observation.

An Extended Form of Bayes

Aside: Derivation of generalized Bayes:

$$P(A, B) = \quad P(A|B) \quad P(B) \quad = P(B|A)P(A) \quad \text{(a1)}$$

An Extended Form of Bayes

Aside: Derivation of generalized Bayes:

$$P(A,B) = P(A|B)\ P(B) = P(B|A)P(A) \quad \text{(a1)}$$

$$P(A,B,C) = P\big(A,(B,C)\big) = P(A|B,C)P(B,C) \quad \text{(a2)}$$

$$P(A,B,C) = P\big(B,(A,C)\big) = P(B|A,C)P(A,C) \quad \text{(a3)}$$

Extend (a1) to get (a2) and (a3).

An Extended Form of Bayes

Aside: Derivation of generalized Bayes:

$$P(A,B) = \quad P(A|B) \quad P(B) \quad = P(B|A)P(A) \qquad \text{(a1)}$$

$$P(A,B,C) = \quad P\big(A,(B,C)\big) = P(A|B,C)P(B,C) \qquad \text{(a2)}$$

$$P(A,B,C) = \quad P\big(B,(A,C)\big) = P(B|A,C)P(A,C) \qquad \text{(a3)}$$

Solve from (a2), (a3). $$P(A|B,C) = \frac{P(B|A,C)P(A,C)}{P(B,C)} \qquad \text{(a4)}$$

Ratio from (a1). $$\frac{P(A,C)}{P(B,C)} = \frac{P(A|C)P(C)}{P(B|C)P(C)} = \frac{P(A|C)}{P(B|C)} \qquad \text{(a5)}$$

Substitute (a5) in (a4). $$P(A|B,C) = \frac{P(B|A,C)P(A|C)}{P(B|C)}$$

An Extended Form of Bayes

Aside: Derivation of generalized Bayes:

$$P(A|B,C) = \frac{P(B|A,C)P(A|C)}{P(B|C)}$$

For our case with two temperature observations:

$$P(T|T_{O,1},T_{O,2}) = \frac{P(T_{O,2}|T,T_{O,1})P(T|T_{O,1})}{P(T_{O,1}|T_{O,2})}$$

An Extended Form of Bayes

$$P(T|T_{O,1},T_{O,2}) = \frac{P(T_{O,2}|T,T_{O,1})P(T|T_{O,1})}{P(T_{O,1}|T_{O,2})}$$

We will assume that the random errors associated with the two observations are independent. This assumption will be retained throughout almost everything we do.

$$P(T|T_{O,1},T_{O,2}) = \frac{P(T_{O,2}|T)P(T|T_{O,1})}{P(T_{O,1}|T_{O,2})}$$

Combining the Prior Estimate and Observation

Bayes Theorem:

Likelihood: Probability that $T_{O,2}$ is observed if T is true value.

Prior

$$P(T|T_{O,1},T_{O,2}) = \frac{P(T_{O,2}|T)P(T|T_{O,1})}{P(T_{O,1}|T_{O,2})}$$

Posterior: Probability of T given observation and prior. Also called **update** or **analysis**.

Combining the Prior Estimate and Observation

Bayes Theorem:

Likelihood: Probability that $T_{O,2}$ is observed if T is true value.

Prior

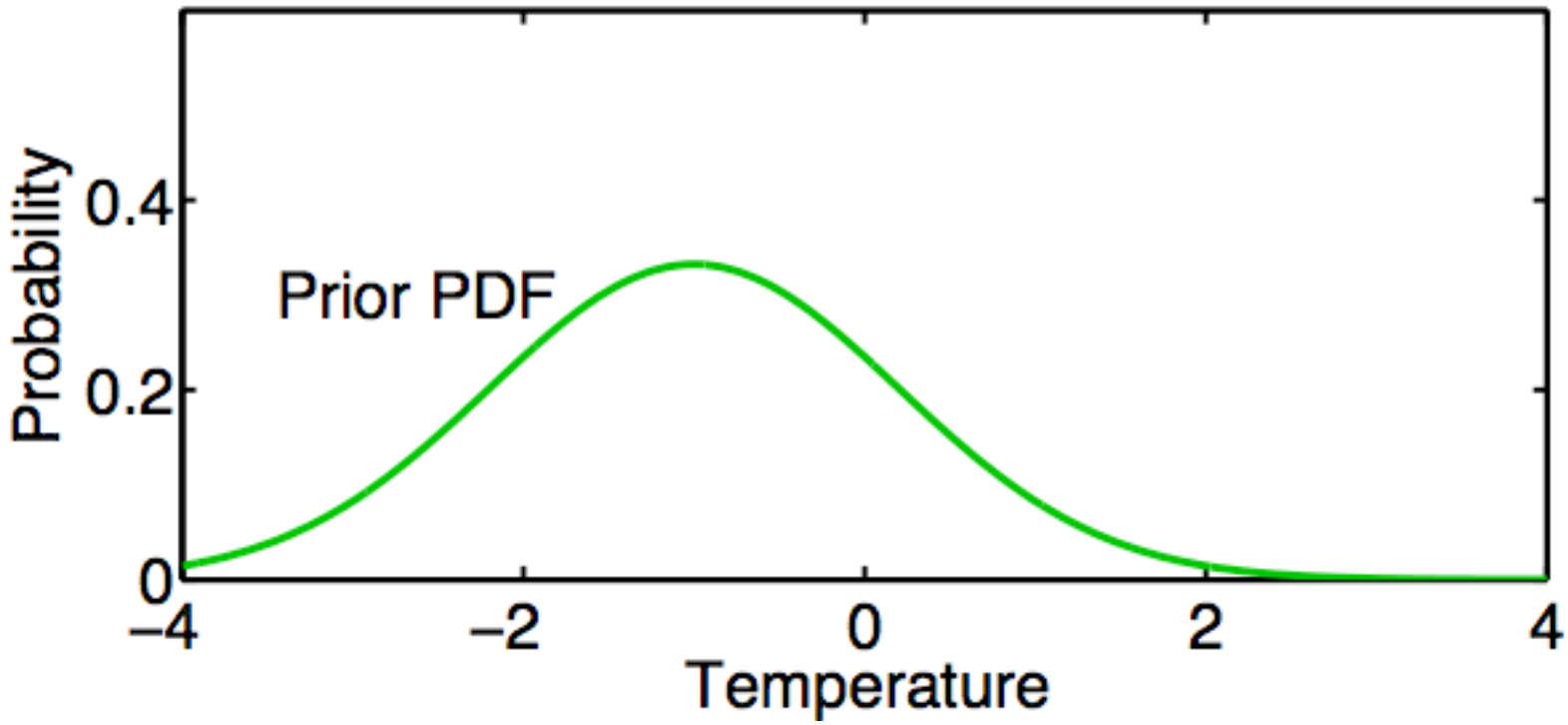
$$P(T|T_{O,1},T_{O,2}) = \frac{P(T_{O,2}|T)P(T|T_{O,1})}{normalization}$$

Posterior: Probability of T given observation and prior. Also called **update** or **analysis**.

Denominator is a normalization so that posterior is a probability distribution (PDF).

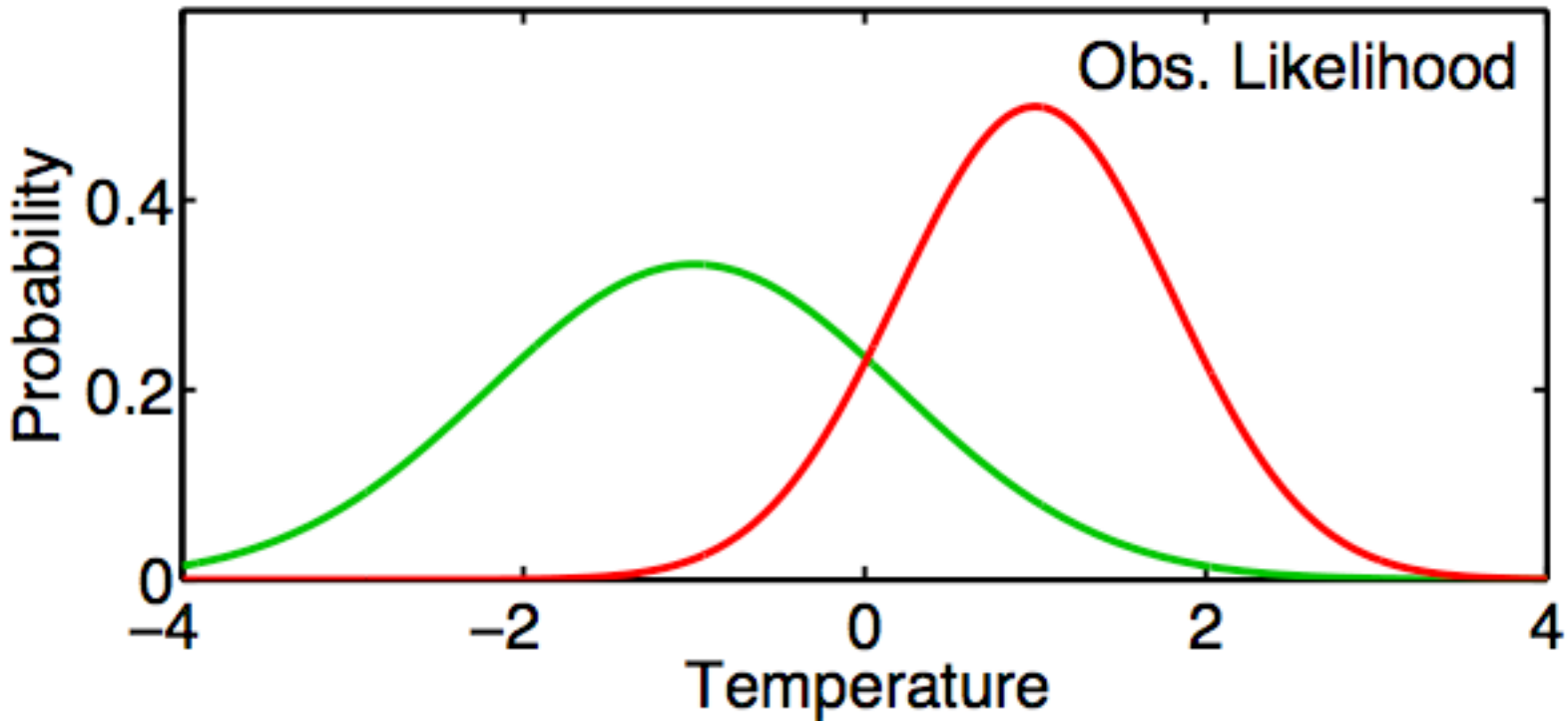
Combining the Prior Estimate and Observation

$$P(T|T_{O,1}, T_{O,2}) = \frac{P(T_{O,2}|T\ )P(T|T_{O,1})}{Normalization}$$



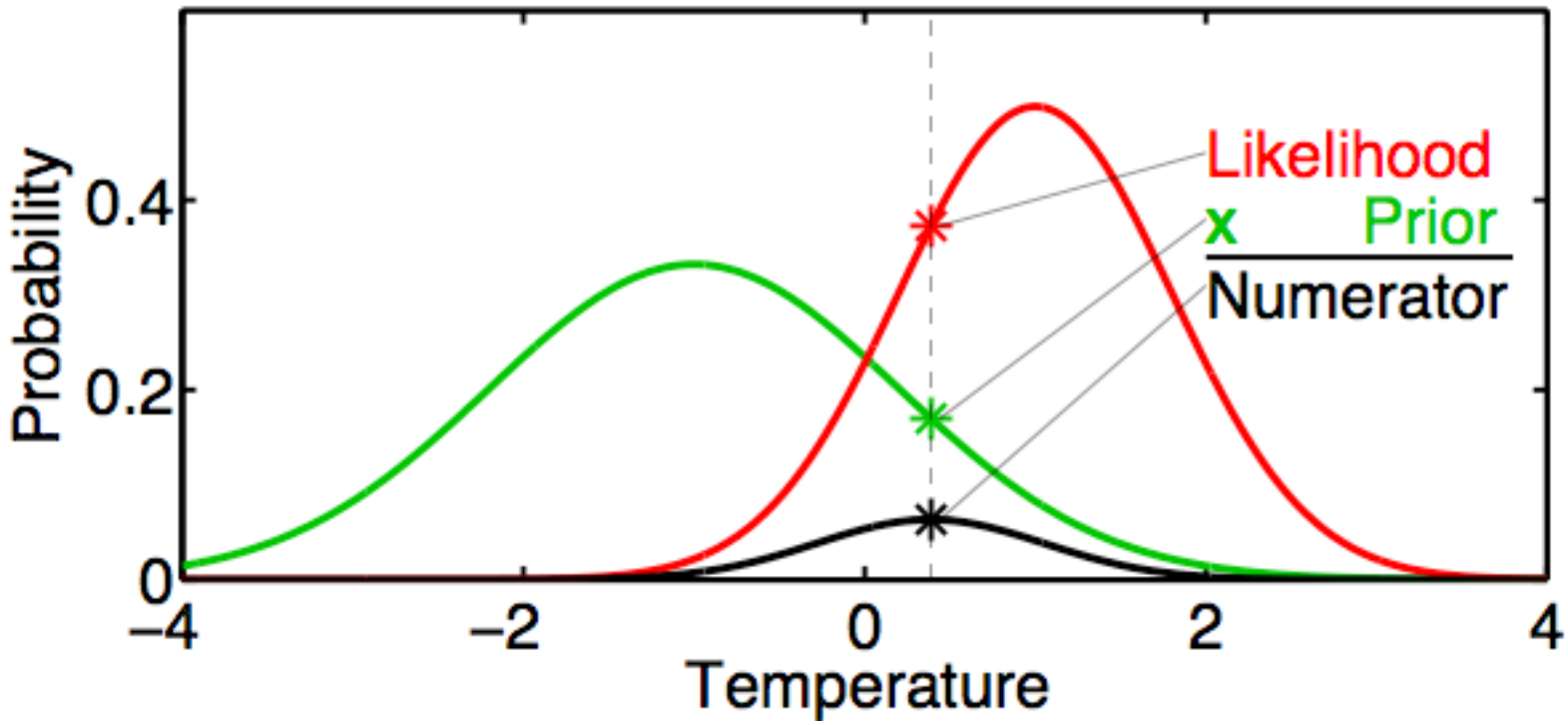
Combining the Prior Estimate and Observation

$$P(T|T_{O,1}, T_{O,2}) = \frac{P(T_{O,2}|T\ )P(T|T_{O,1})}{Normalization}$$



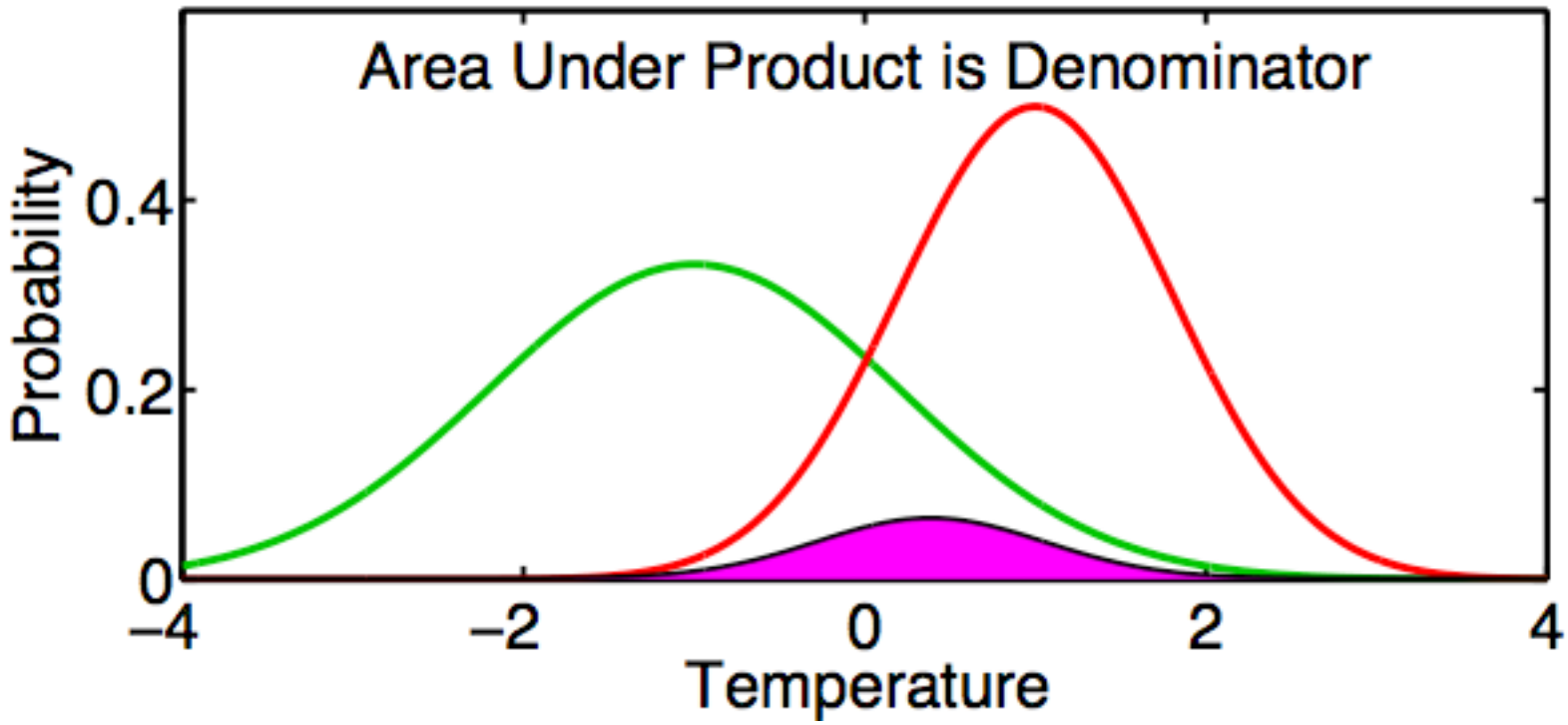
Combining the Prior Estimate and Observation

$$P(T|T_{O,1},T_{O,2}) = \frac{P(T_{O,2}|T\ )P(T|T_{O,1})}{Normalization}$$



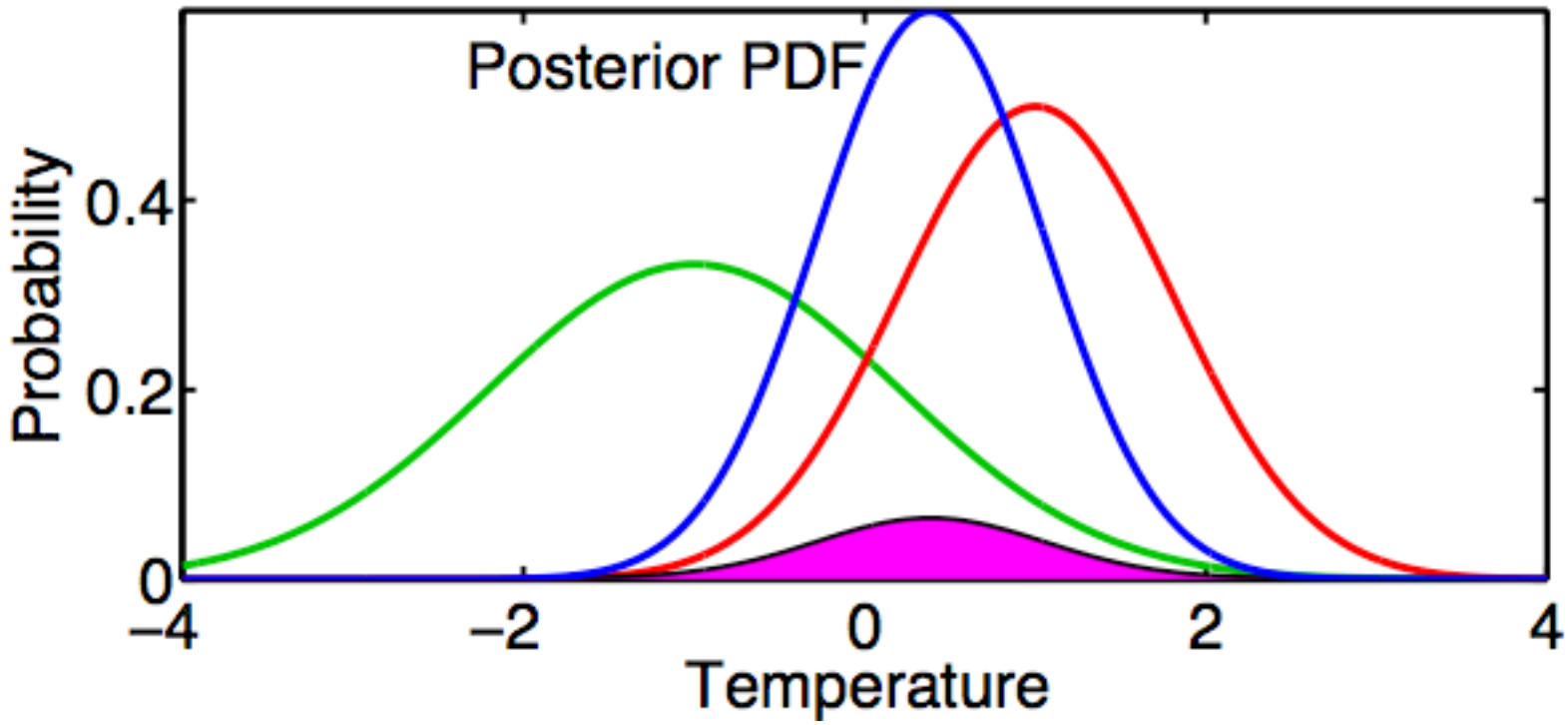
Combining the Prior Estimate and Observation

$$P(T|T_{O,1},T_{O,2}) = \frac{P(T_{O,2}|T\ )P(T|T_{O,1})}{Normalization}$$



Combining the Prior Estimate and Observation

$$P(T|T_{O,1},T_{O,2}) = \frac{P(T_{O,2}|T\ )P(T|T_{O,1})}{Normalization}$$



Color Scheme

Green == Prior

Red == Observation

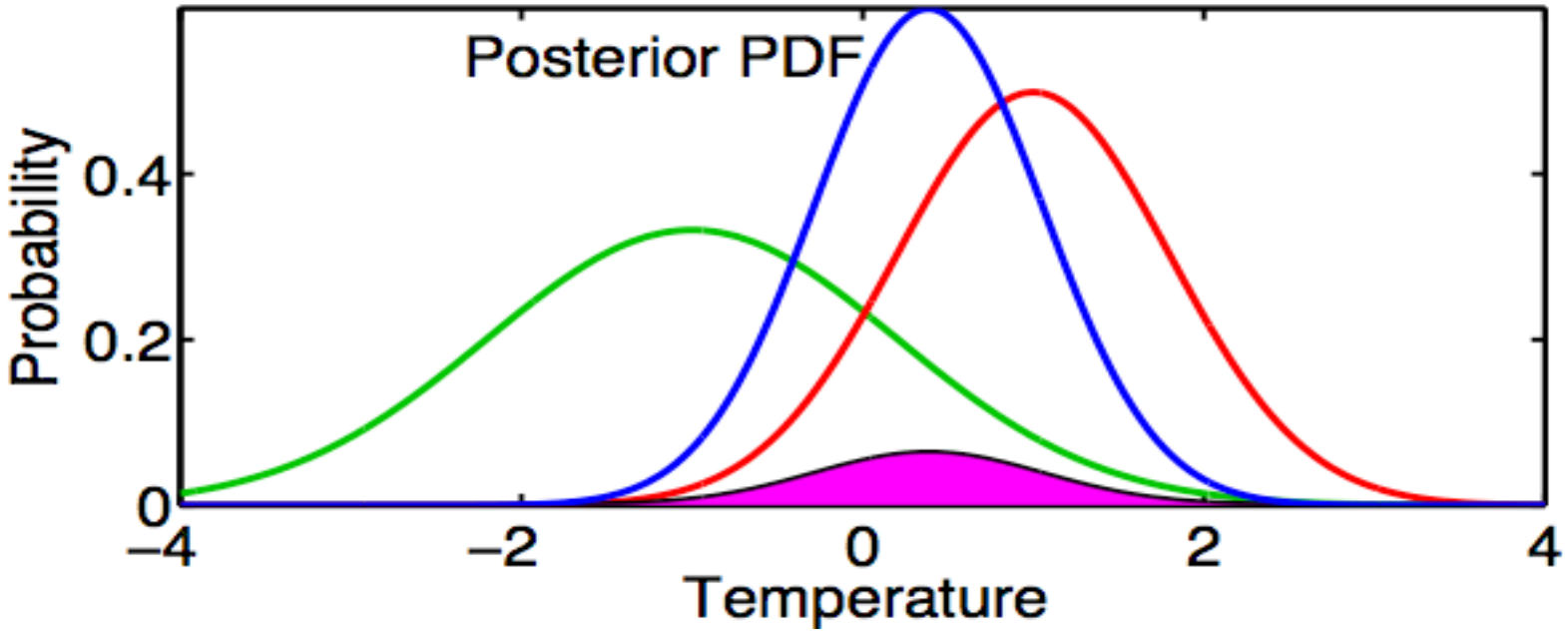
Blue == Posterior

The same color scheme is used throughout ALL Tutorial materials.

Combining the Prior Estimate and Observation

$$P(T|T_{O,1}, T_{O,2}) = \frac{P(T_{O,2}|T\ )P(T|T_{O,1})}{Normalization}$$

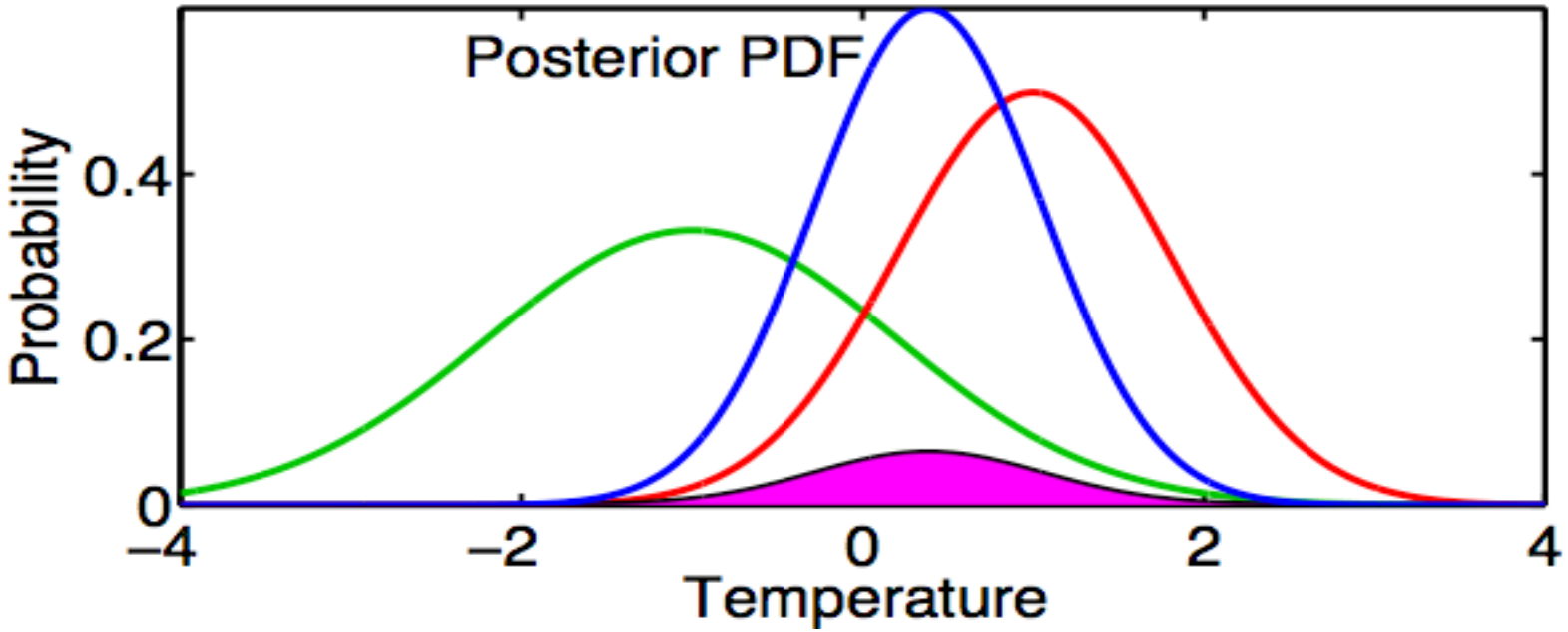
Generally no analytic solution for Posterior.



Combining the Prior Estimate and Observation

$$P(T|T_{O,1}, T_{O,2}) = \frac{P(T_{O,2}|T\ )P(T|T_{O,1})}{Normalization}$$

Gaussian Prior and Likelihood -> Gaussian Posterior.



Combining the Prior Estimate and Observation

For Gaussian prior and likelihood…

Prior $P(T|T_{O,1}) = Normal(T_p, \sigma_p^2)$

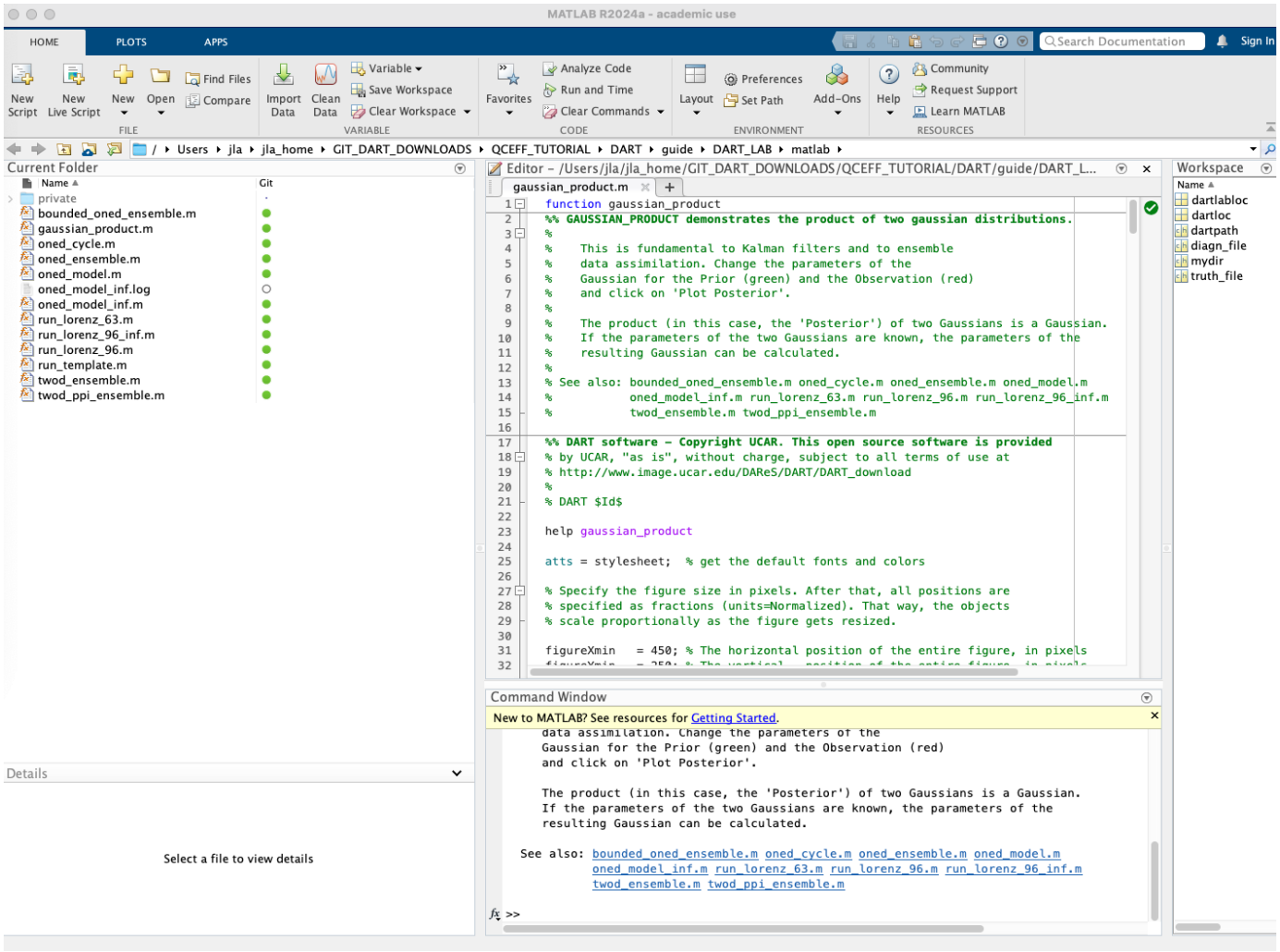
Likelihood $P(T_{O,2}|T) = Normal(T_O, \sigma_O^2)$

Then, Posterior $P(T|T_{O,1}, T_{O,2}) = Normal(T_u, \sigma_u^2)$

With $\sigma_u^2 = (\sigma_p^{-2} + \sigma_O^{-2})^{-1}$ and

$$T_u = \sigma_u^2(\sigma_p^{-2}T_p + \sigma_O^{-2}T_O)$$

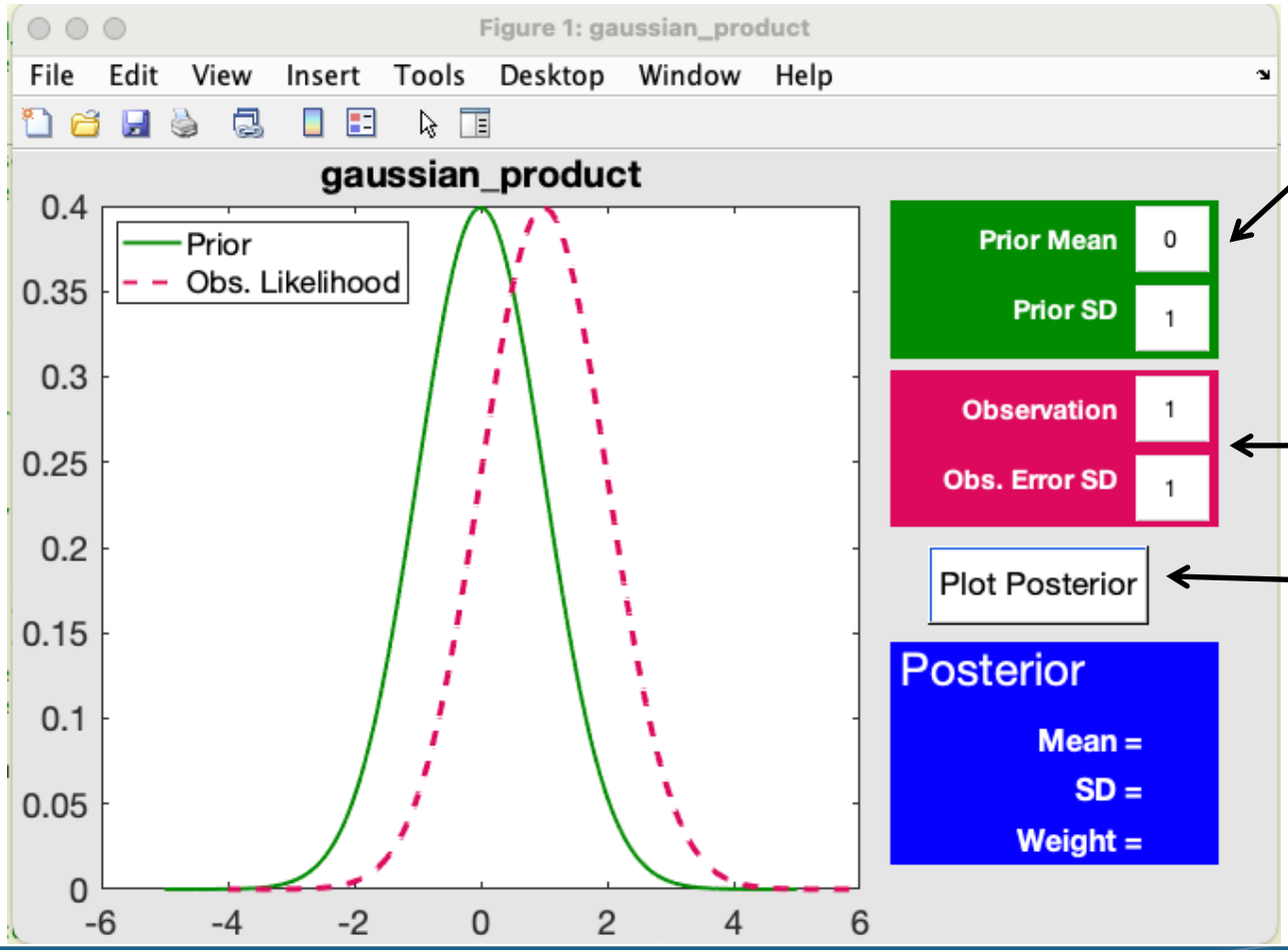
Matlab Hands-on: gaussian_product



This will also spawn a GUI that we will work with.

Matlab Hands-on: gaussian_product

Purpose: Explore the gaussian posterior that results from taking the product of a gaussian prior and a gaussian likelihood.



1) Set Prior Mean and Standard Deviation.

2) Set Observation Mean and Observation Error Standard Deviation.

3) Select Plot Posterior to Update the items in blue.

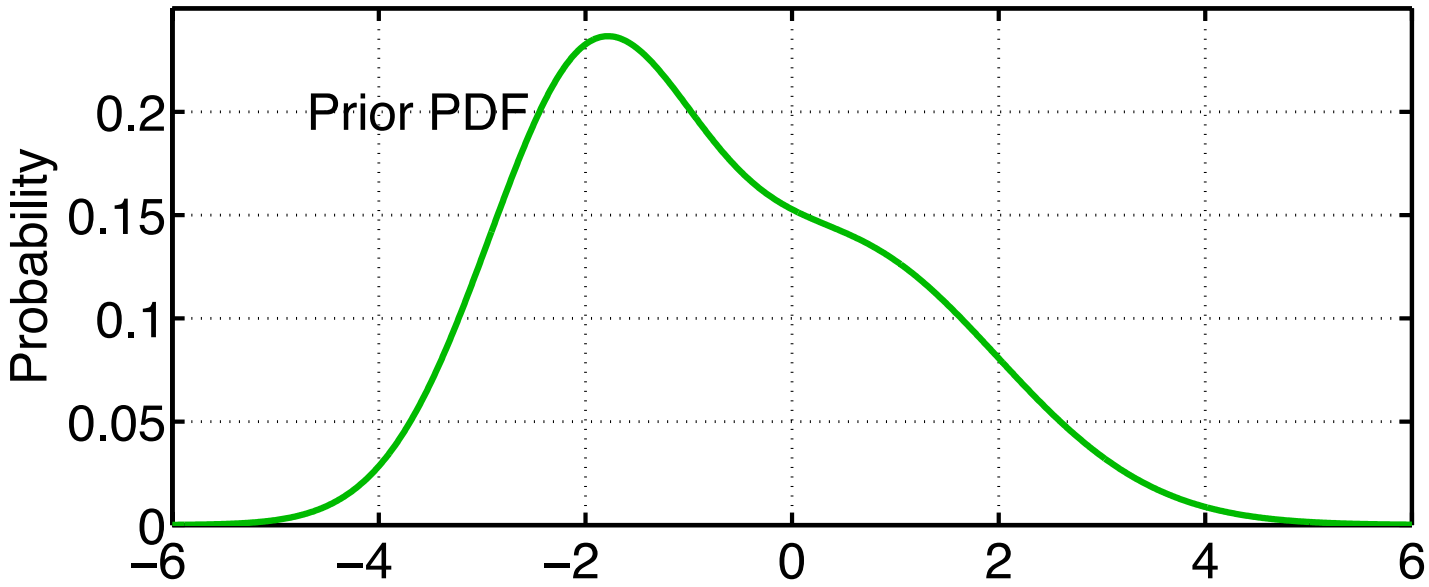
Matlab Hands-on: gaussian_product

Explore!

- Change the mean value of the prior and the observation.
- Change the standard deviation of the prior.
- What is always true for the mean of the posterior?
- What is always true for the standard deviation of the posterior?

Bayes' Theorem

$$p(A \mid BC) = \frac{p(B \mid AC)\,p(A \mid C)}{p(B \mid C)} = \frac{p(B \mid AC)\,p(A \mid C)}{\int p(B \mid x)\,p(x \mid C)\,dx}$$



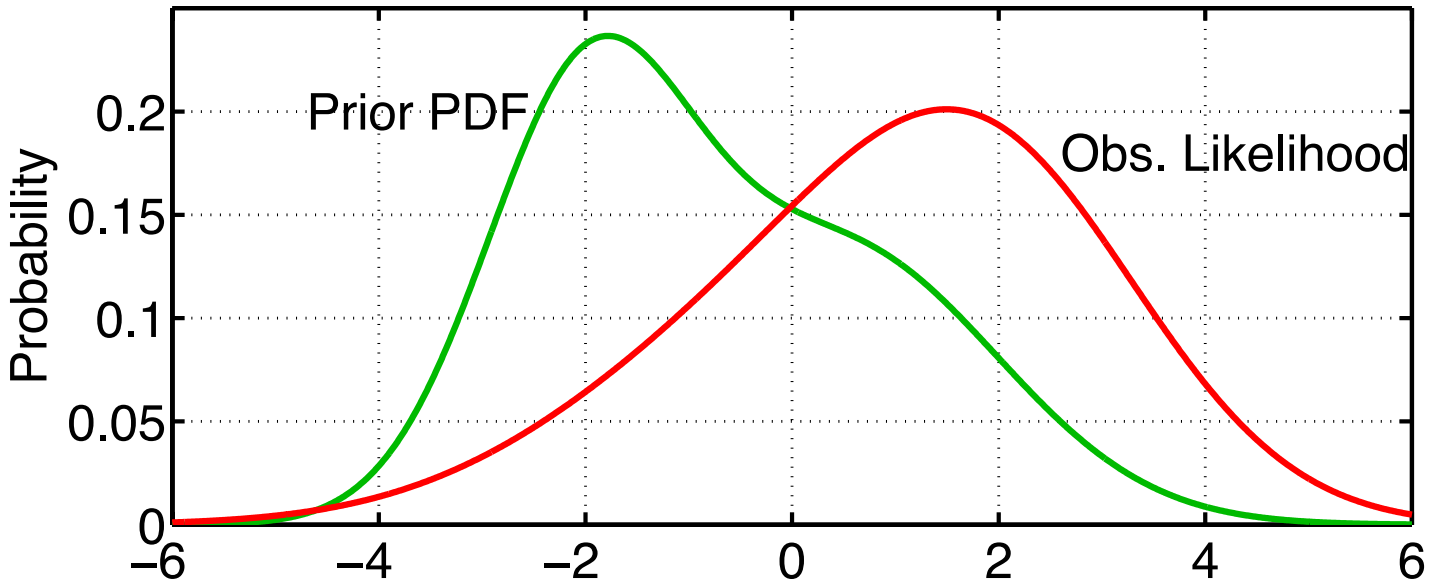
A : Prior Estimate based on all previous information, C.

B : An additional observation.

$p(A|BC)$: Posterior (updated estimate) based on C and B.

Bayes' Theorem

$$p(A \mid BC) = \frac{p(B \mid AC)\,p(A \mid C)}{p(B \mid C)} = \frac{p(B \mid AC)\,p(A \mid C)}{\int p(B \mid x)\,p(x \mid C)\,dx}$$



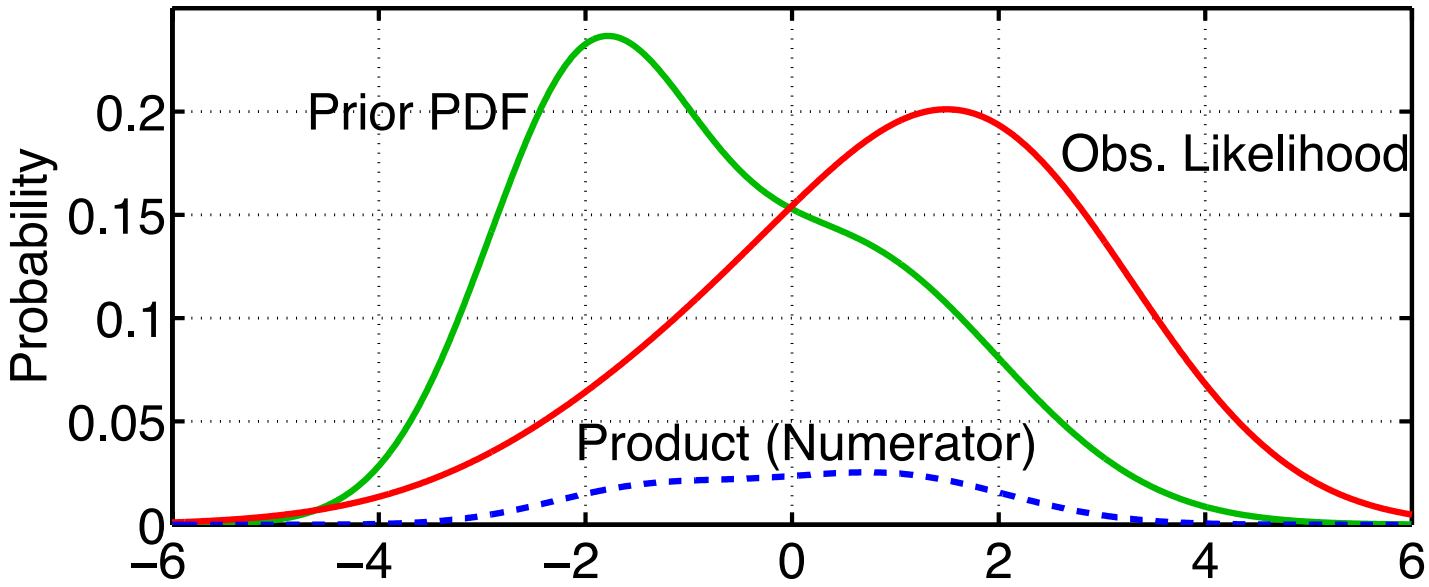
A : Prior Estimate based on all previous information, C.

B : An additional observation.

$p(A|BC)$: Posterior (updated estimate) based on C and B.

Bayes' Theorem

$$p(A \mid BC) = \frac{p(B \mid AC)\,p(A \mid C)}{p(B \mid C)} = \frac{p(B \mid AC)\,p(A \mid C)}{\int p(B \mid x)\,p(x \mid C)\,dx}$$



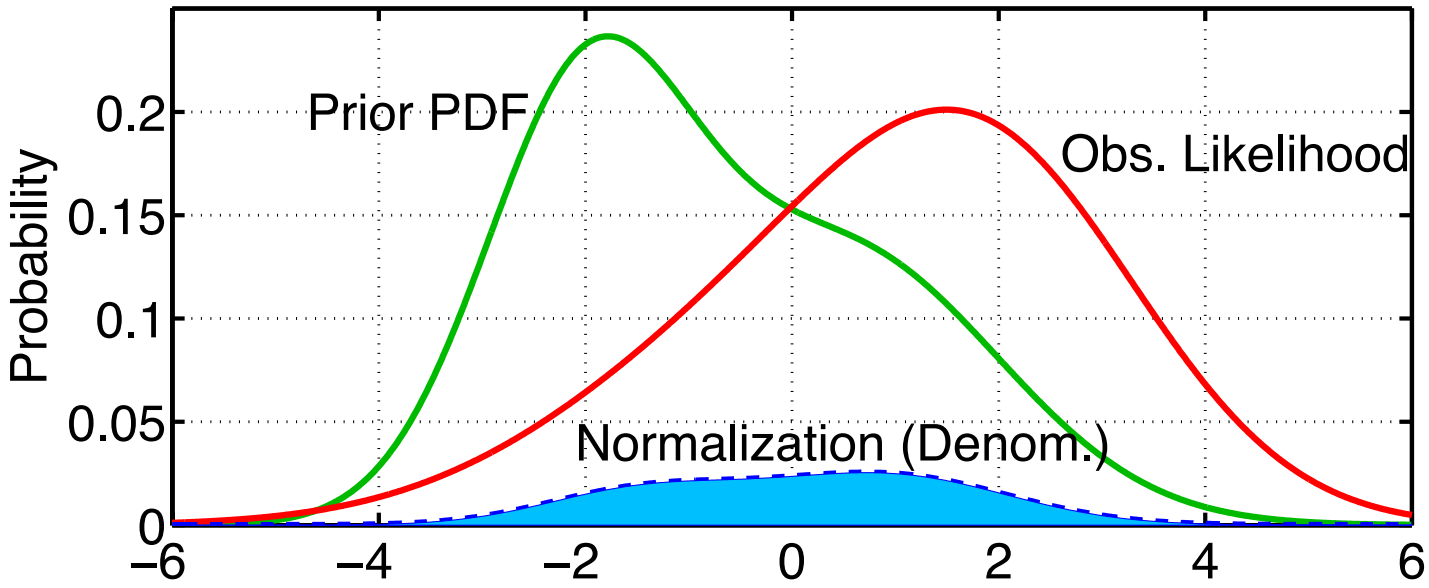
A : Prior Estimate based on all previous information, C.

B : An additional observation.

$p(A|BC)$: Posterior (updated estimate) based on C and B.

Bayes' Theorem

$$p(A \mid BC) = \frac{p(B \mid AC)\,p(A \mid C)}{p(B \mid C)} = \frac{p(B \mid AC)\,p(A \mid C)}{\int p(B \mid x)\,p(x \mid C)\,dx}$$



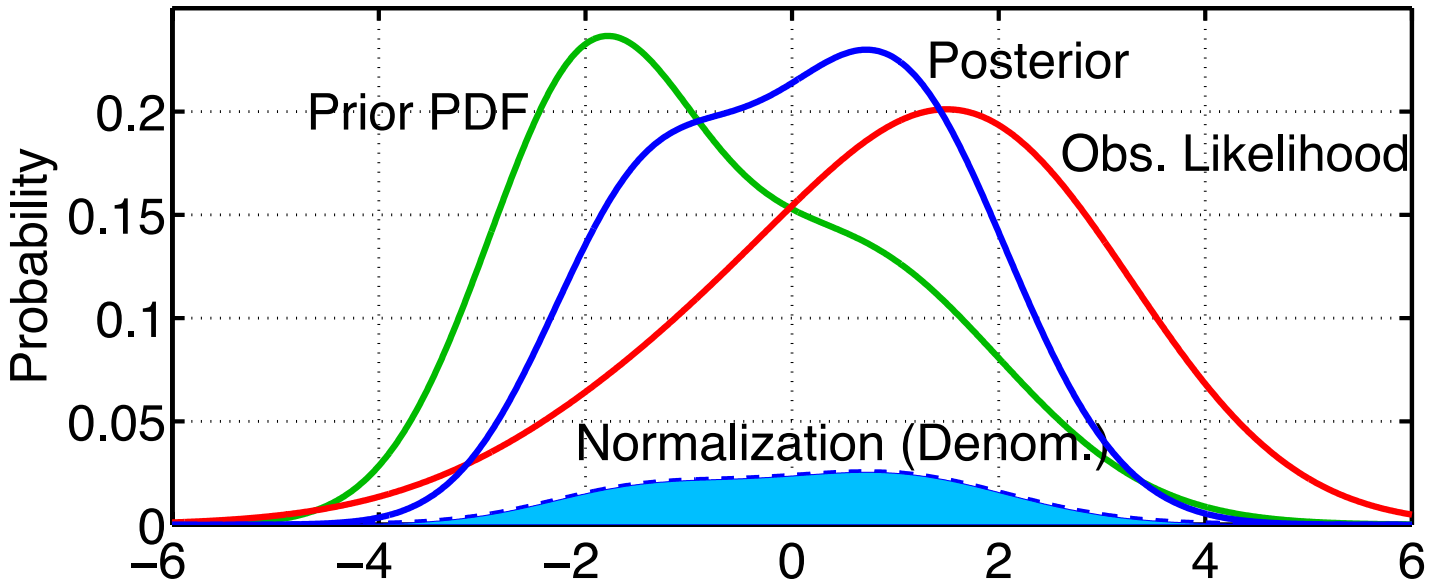
A : Prior Estimate based on all previous information, C.

B : An additional observation.

$p(A|BC)$: Posterior (updated estimate) based on C and B.

Bayes' Theorem

$$p(A \mid BC) = \frac{p(B \mid AC)\,p(A \mid C)}{p(B \mid C)} = \frac{p(B \mid AC)\,p(A \mid C)}{\int p(B \mid x)\,p(x \mid C)\,dx}$$



A : Prior Estimate based on all previous information, C.

B : An additional observation.

$p(A|BC)$: Posterior (updated estimate) based on C and B.

Back to Temperature Observations

Bayes Theorem:

Likelihood: Probability that $T_{O,2}$ is observed if T is true value.

Prior

$$P(T|T_{O,1},T_{O,2}) = \frac{P(T_{O,2}|T\ )P(T|T_{O,1})}{normalization}$$

Posterior: Probability of T given observation and prior. Also called **update** or **analysis**.

Denominator is a normalization so that posterior is a probability distribution (PDF).

Back to Temperature Observations

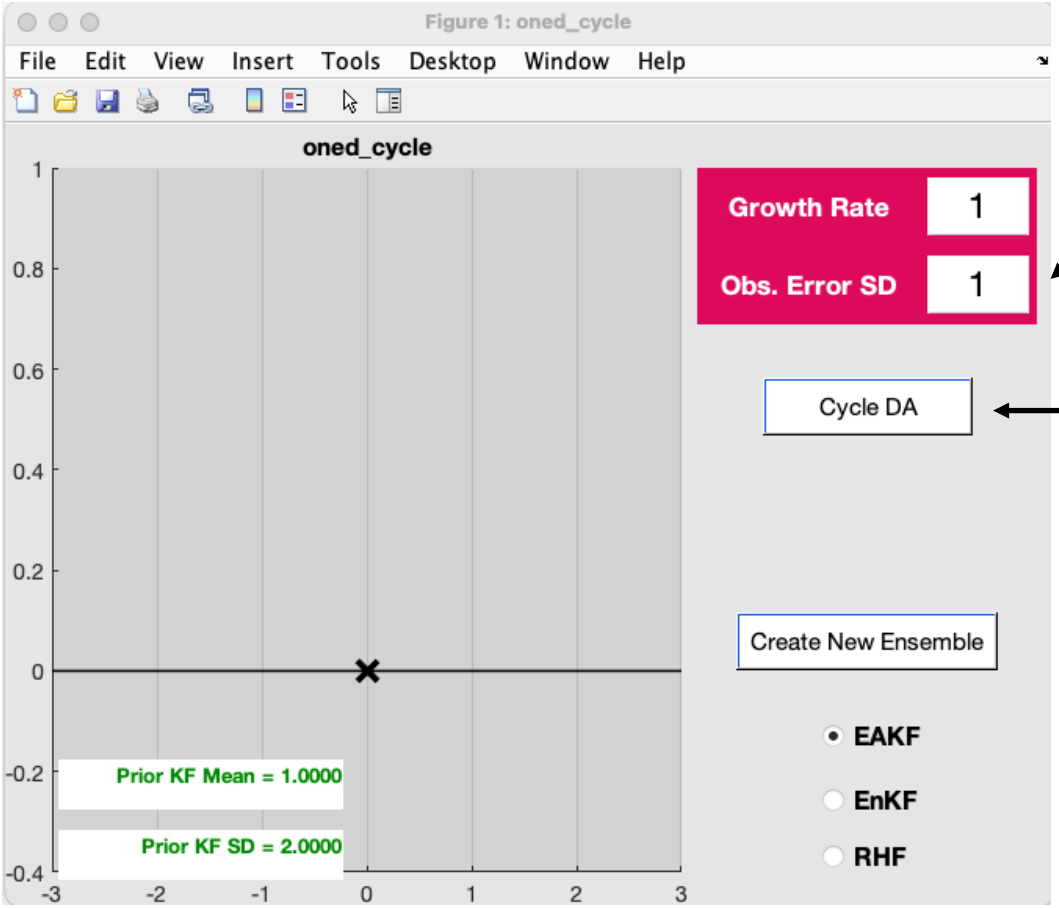
Bayes
Theorem:

$$P(T|T_{O,1}, T_{O,2}, \ldots, T_{O,n}) = \frac{P(T_{O,n}|T\,)P(T|T_{O,1}, \ldots, T_{O,n-1}\,)}{Normalization}$$

Can do a sequence of n observations, each time making the previous posterior into the new prior.
This will converge to the true temperature as a function of n.

Matlab Hands-on: oned_cycle (1)

Purpose: Use Bayes to 'assimilate' multiple observations of temperature at the same time.



Change the error standard deviation of observations.

Press to assimilate next T observation.

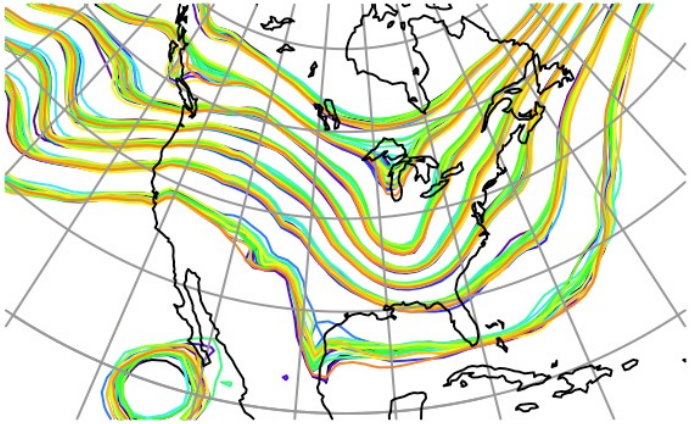
Matlab Hands-on: oned_cycle (1)

What happens as more observations are assimilated?

What is Data Assimilation?

Observations combined with a Model forecast …

…to produce an analysis
(best possible estimate).



The One-Dimensional Kalman Filter

1. Suppose we have a linear forecast model L.
   A. If temperature at time $t_1 = T_1$, then the temperature at $t_2 = t_1 + \Delta t$ is $T_2 = L(T_1)$.
   B. Example: $T_2 = G*T_1$

The One-Dimensional Kalman Filter

1. Suppose we have a linear forecast model L.
   A. If temperature at time $t_1 = T_1$, then the temperature at $t_2 = t_1 + \Delta t$ is $T_2 = L(T_1)$.
   B. Example: $T_2 = G*T_1$
2. If posterior estimate at time t_1 is $Normal\left(T_{u,1}, \sigma_{u,1}^2\right)$, then the prior at t_2 is $Normal\left(T_{p,2}, \sigma_{p,2}^2\right)$.

$$T_{p,2} = G * T_{u,1}$$

$$\sigma_{p,2}^2 = G * \sigma_{u,1}^2$$

The One-Dimensional Kalman Filter

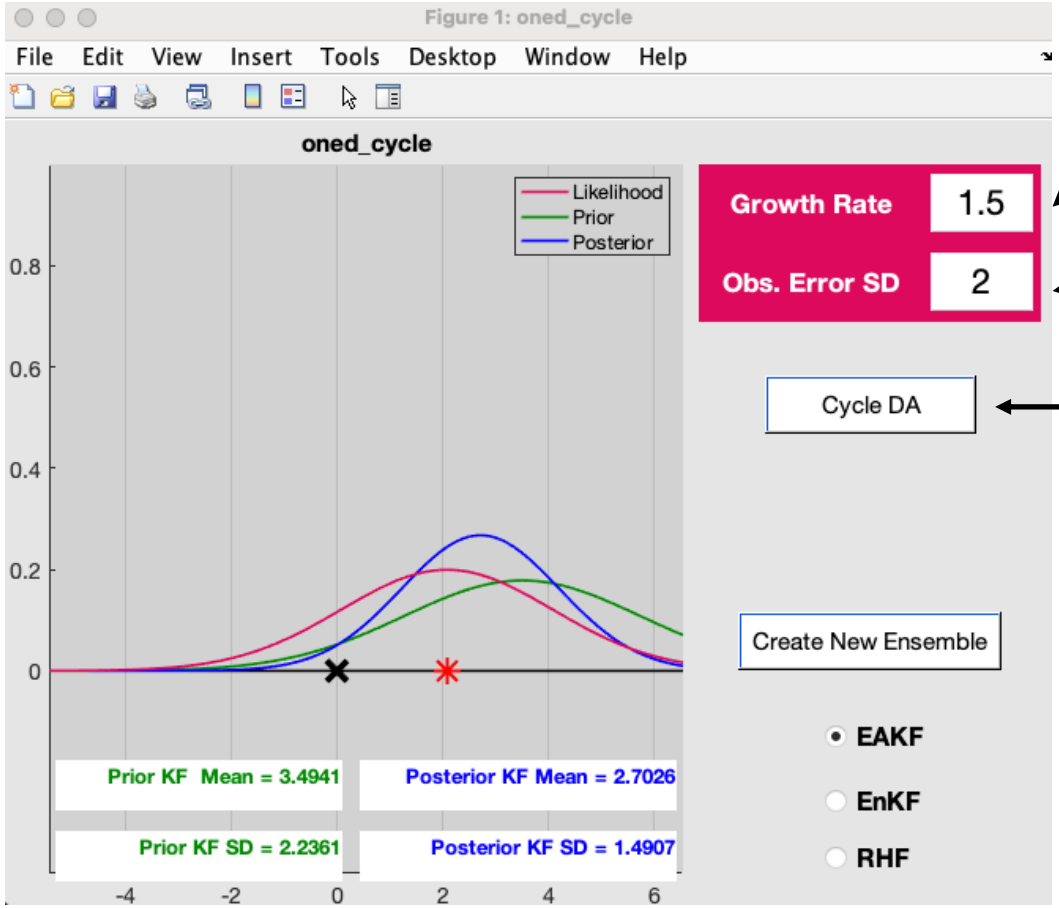
1. Suppose we have a linear forecast model L.
   A. If temperature at time $t_1 = T_1$, then the temperature at $t_2 = t_1 + \Delta t$ is $T_2 = L(T_1)$.
   B. Example: $T_2 = G*T_1$
2. If posterior estimate at time t_1 is $Normal\left(T_{u,1}, \sigma_{u,1}^2\right)$, then the prior at t_2 is $Normal\left(T_{p,2}, \sigma_{p,2}^2\right)$.
3. Given an observation at t_2 with observation distribution $Normal(T_O, \sigma_O^2)$ the likelihood is also $Normal(T_O, \sigma_O^2)$.

The One-Dimensional Kalman Filter

1. Suppose we have a linear forecast model L.
   A. If temperature at time $t_1 = T_1$, then the temperature at $t_2 = t_1 + \Delta t$ is $T_2 = L(T_1)$.
   B. Example: $T_2 = G*T_1$
2. If posterior estimate at time t_1 is $Normal\left(T_{u,1}, \sigma_{u,1}^2\right)$, then the prior at t_2 is $Normal\left(T_{p,2}, \sigma_{p,2}^2\right)$.
3. Given an observation at t_2 with observation distribution $Normal(T_O, \sigma_O^2)$ the likelihood is also $Normal(T_O, \sigma_O^2)$.
4. The posterior at t_2 is $Normal\left(T_{u,2}, \sigma_{u,2}^2\right)$ where $T_{u,2}$ and $\sigma_{u,2}^2$ come from page 29.

Matlab Hands-on: oned_cycle (2)

Purpose: One-dimensional Kalman Filter with linear growth model.



Change the growth rate G.

Change the error standard deviation of observations.

Press to assimilate next observation.

Matlab Hands-on: oned_cycle (2)

Make the growth rate > 1 to have a linear growth forecast model.

Cycle the data assimilation.

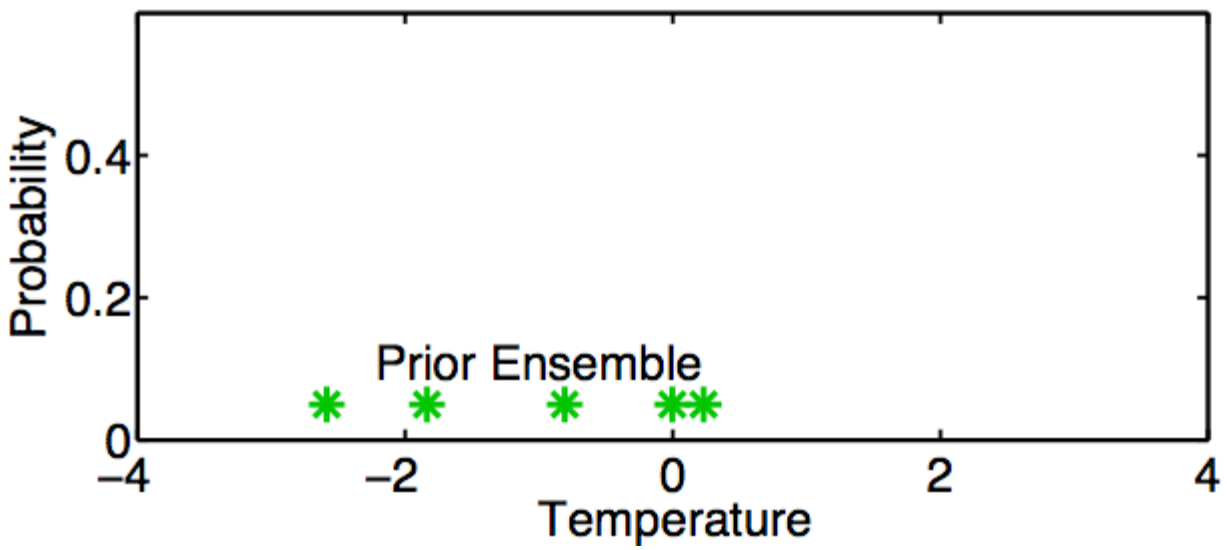
Cycle means do assimilation, do forecast, repeat…

What happens to the prior and posterior standard deviation as you cycle?

What happens to the prior and posterior mean?

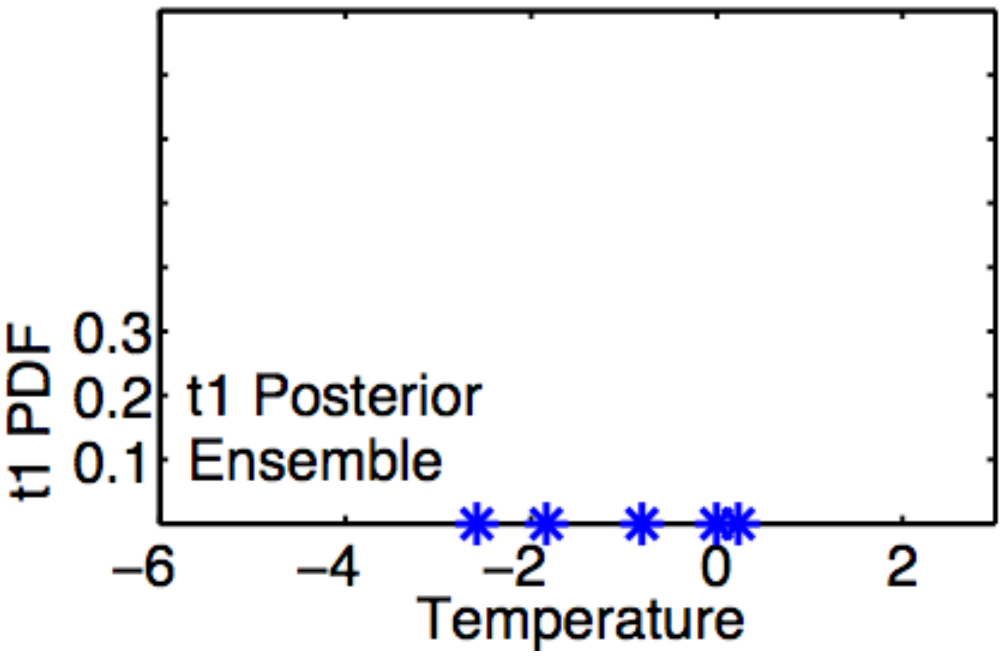
A One-Dimensional Ensemble Kalman Filter

Represent a prior pdf by a sample (ensemble) of N values:



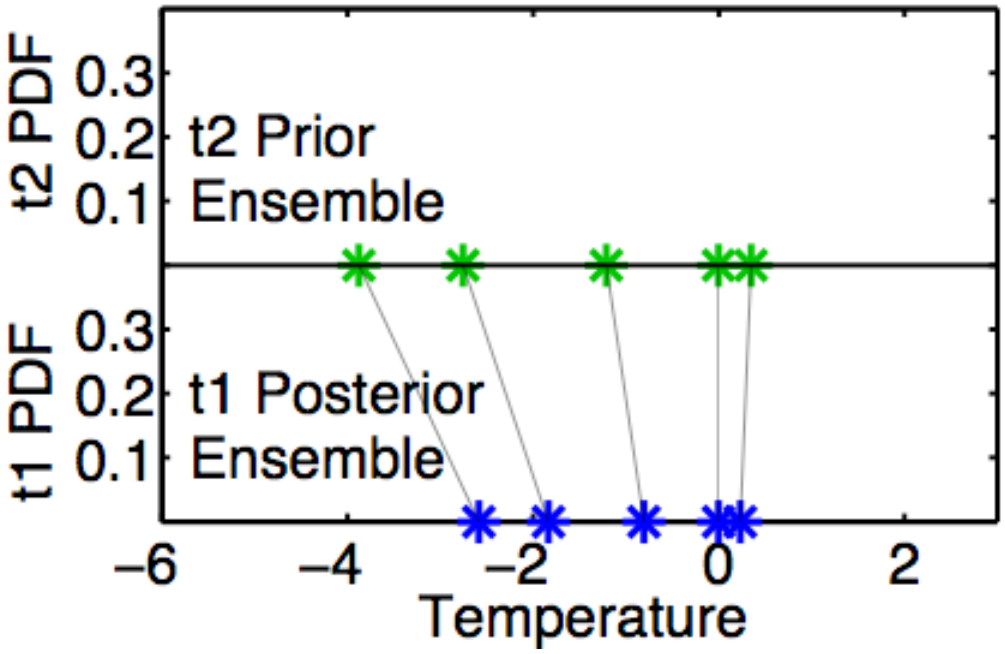
A One-Dimensional Ensemble Kalman Filter: Model Advance

If posterior ensemble at time t_1 is $T_{u,1,n}$, $n = 1, \ldots, N$



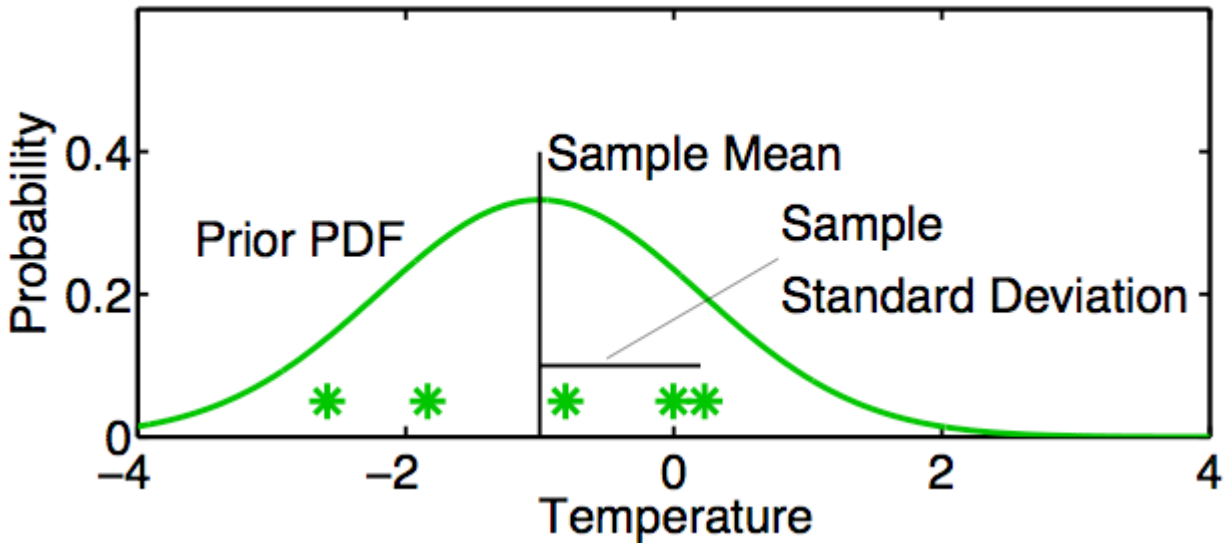
A One-Dimensional Ensemble Kalman Filter: Model Advance

If posterior ensemble at time t_1 is $T_{u,1,n}$, $n = 1, \ldots, N$
advance each member to time t_2 with model, $T_{p,2,n} = L(T_{u,1,n})$ $n = 1, \ldots, N$.



A One-Dimensional Ensemble Kalman Filter

Fit a continuous normal distribution to the prior ensemble (subscripts for prior and time omitted for clarity):

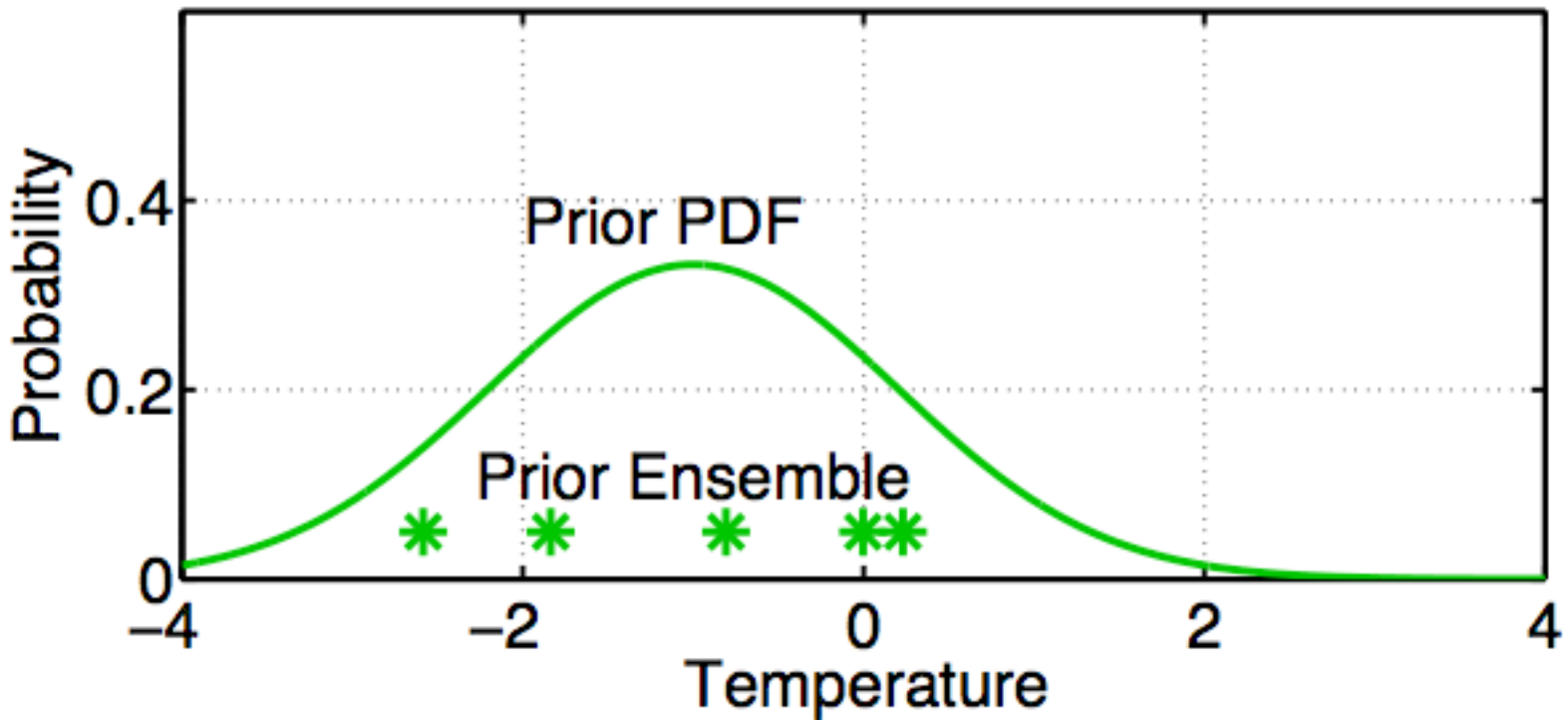


Use sample mean $\overline{T} = \sum_{n=1}^{N} T_n / N$

and sample standard deviation $\sigma_T^2 = \sum_{n=1}^{N} (T_n - \overline{T})^2 / (N-1)$ to determine a

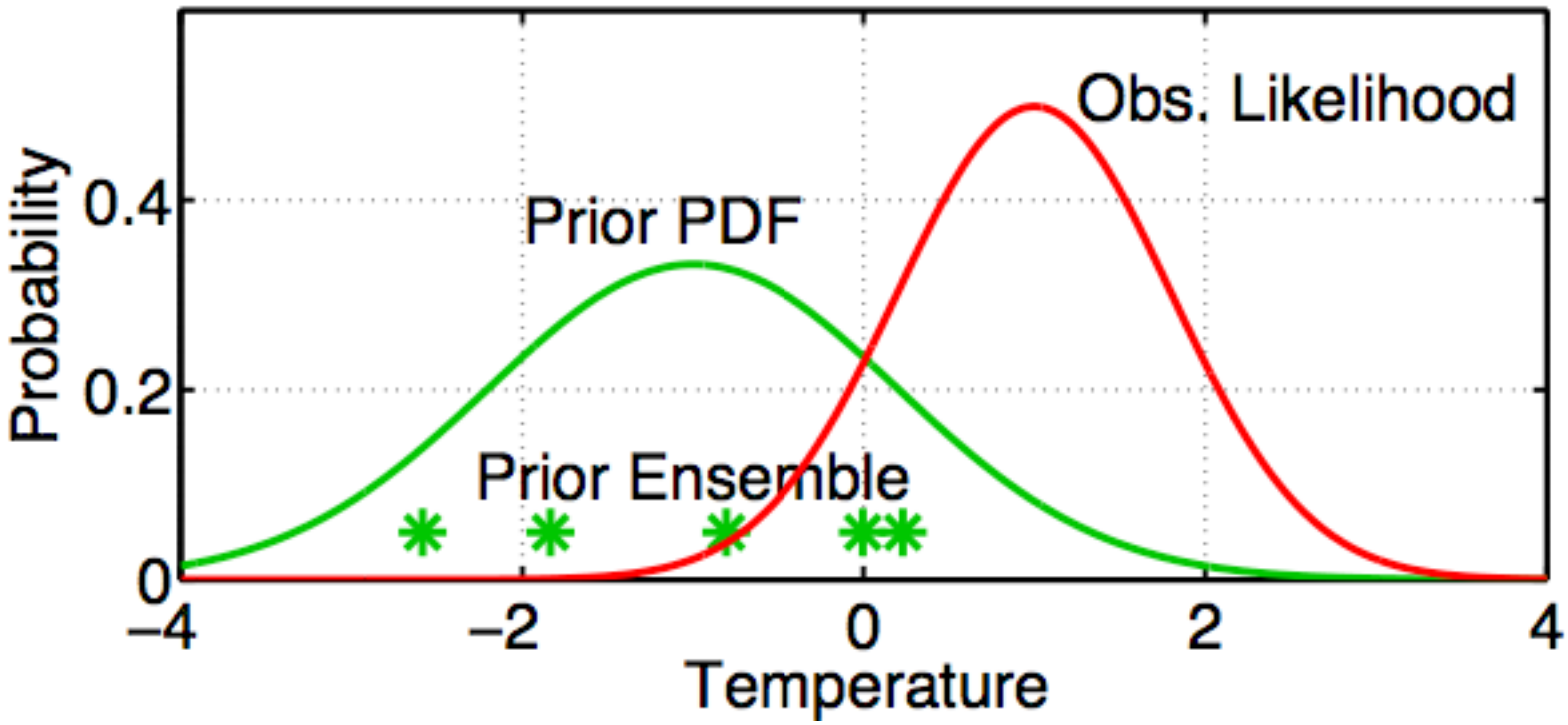
corresponding continuous distribution $Normal(\overline{T}, \sigma_T^2)$

A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation



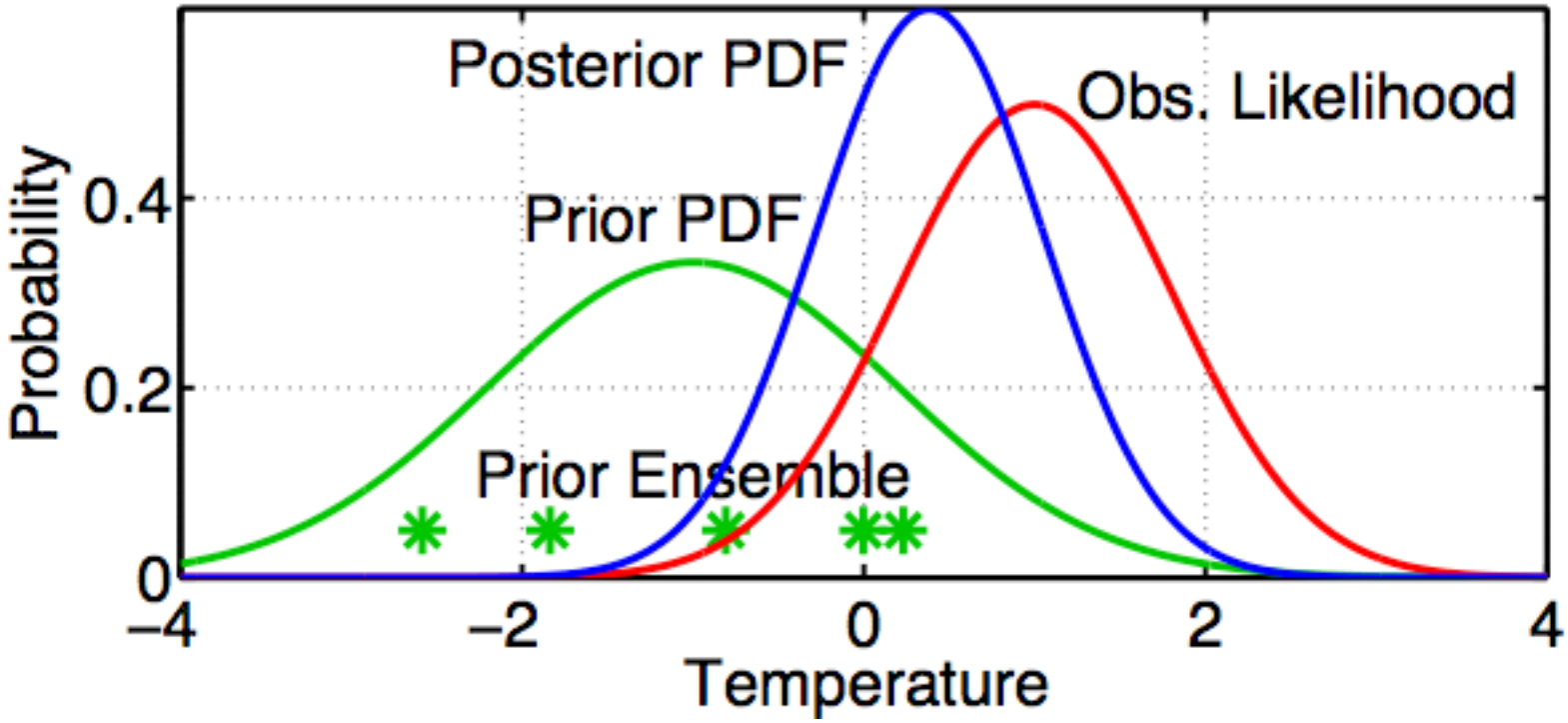
Fit a Gaussian to the sample.

A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation



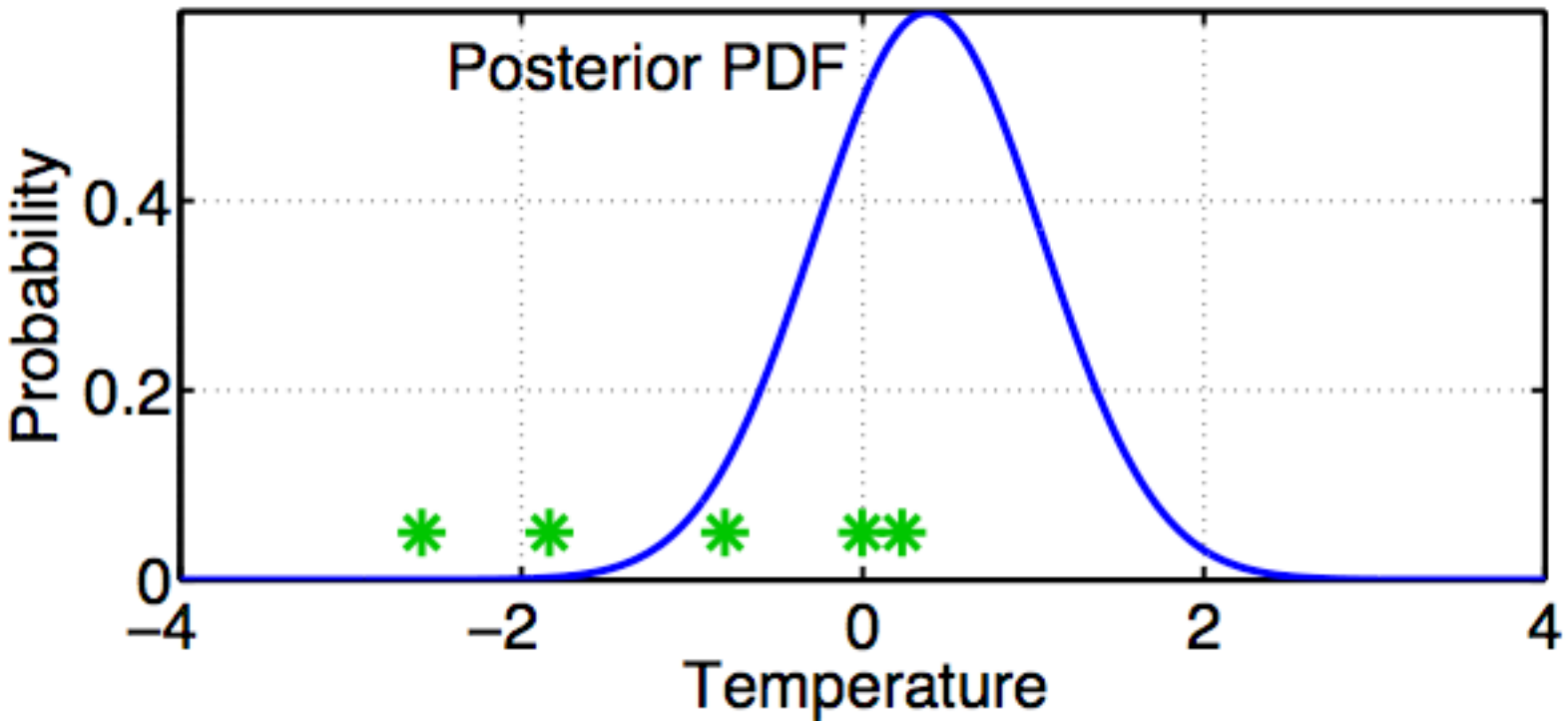
Get the observation likelihood.

A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation



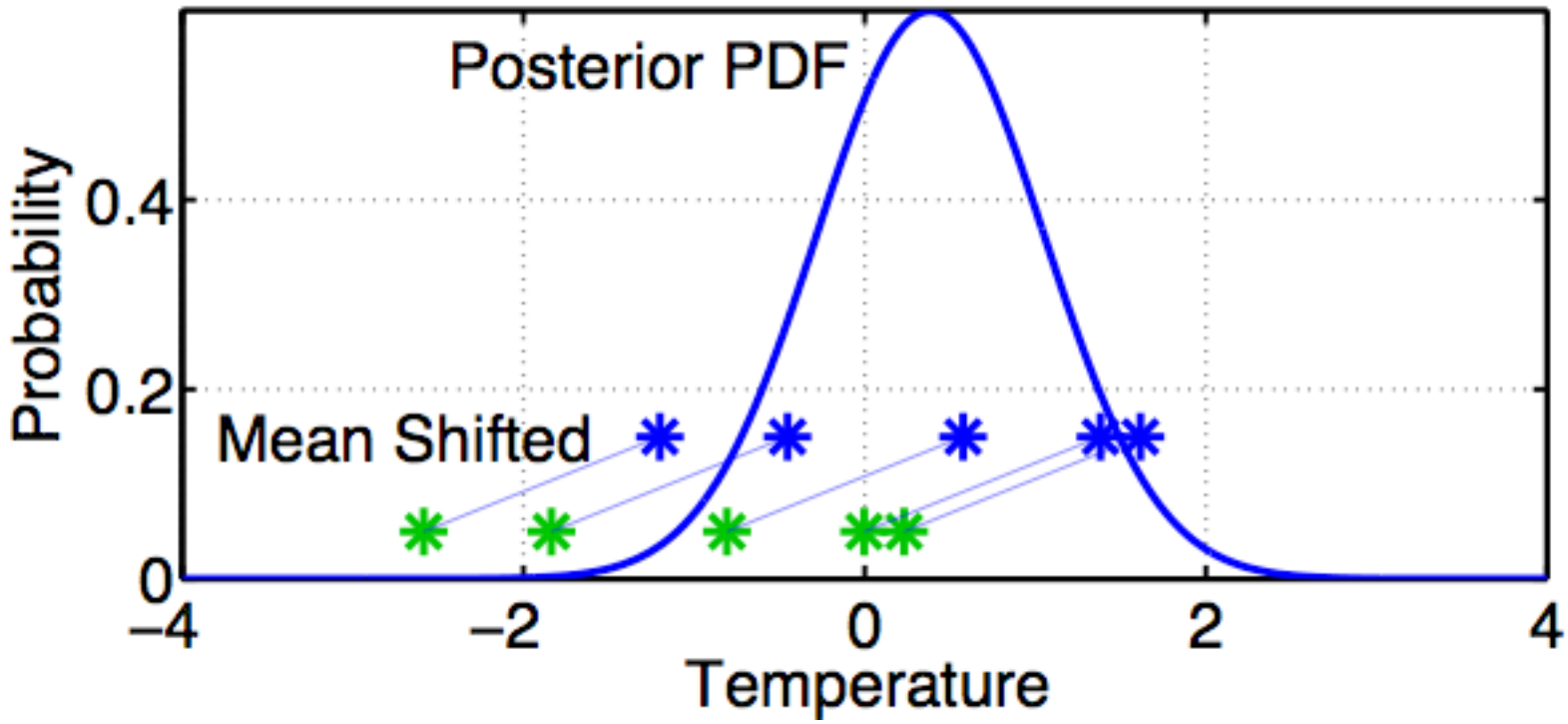
Compute the continuous posterior PDF.

A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation



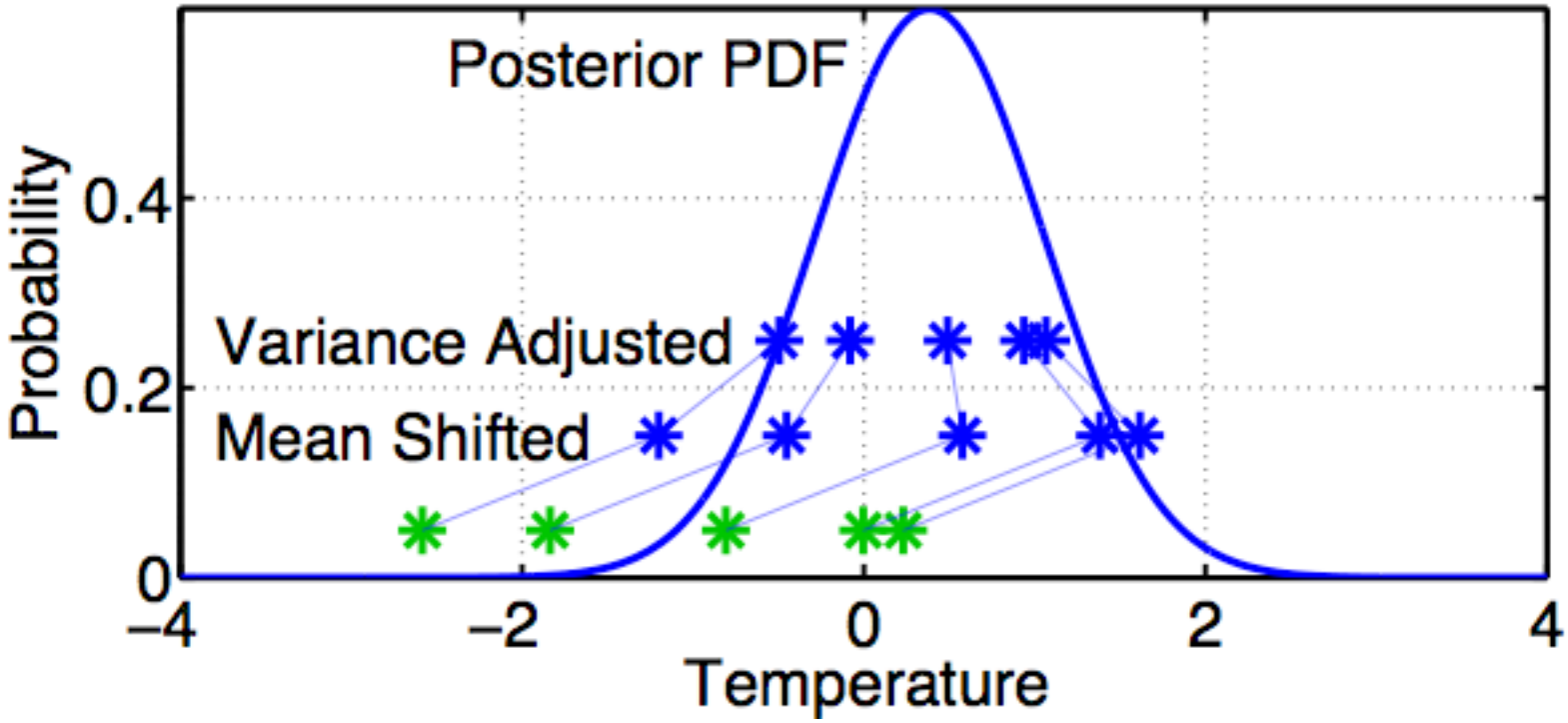
Use a deterministic algorithm to ‘adjust’ the ensemble.

A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation



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

A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation

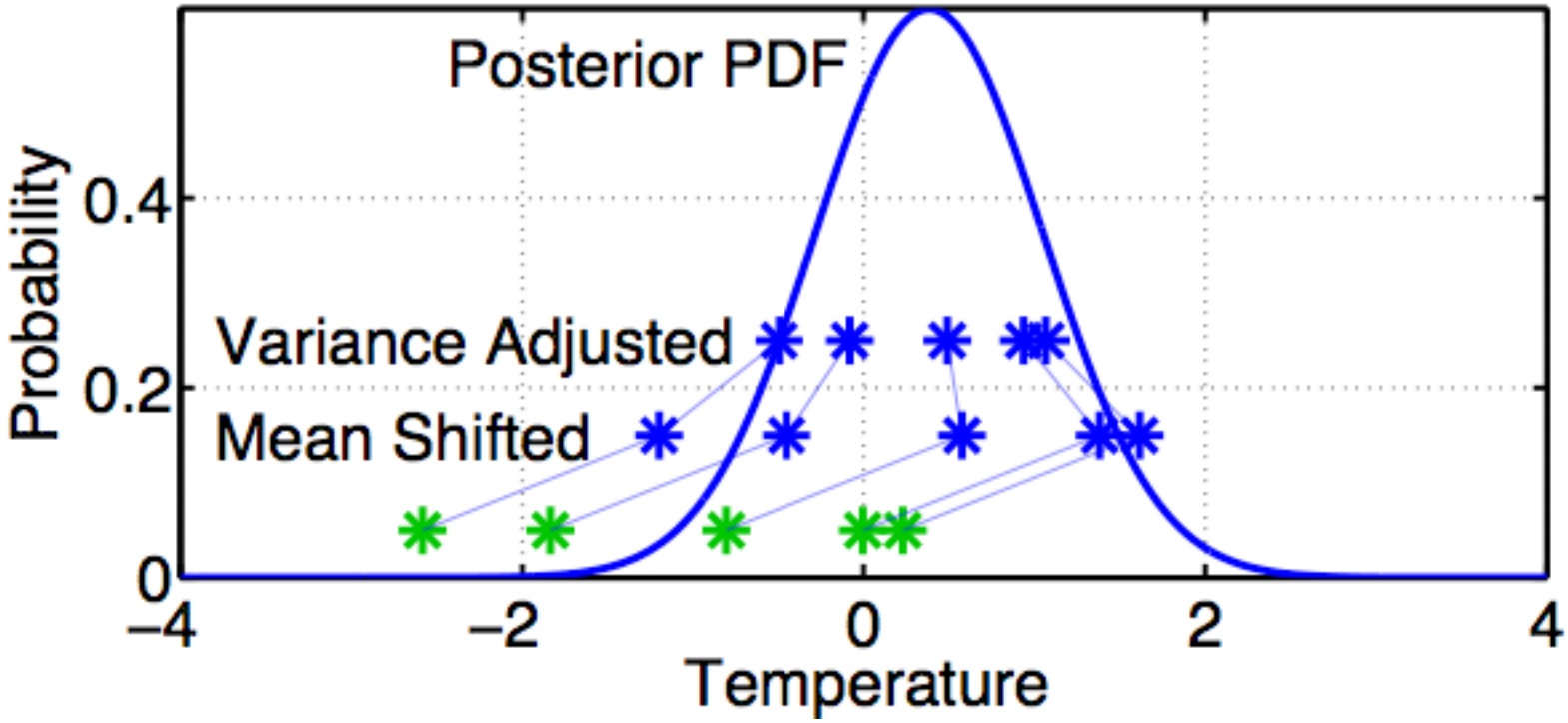


First, ‘shift’ the ensemble to have the exact mean of the posterior.

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

Sample statistics are identical to Kalman filter.

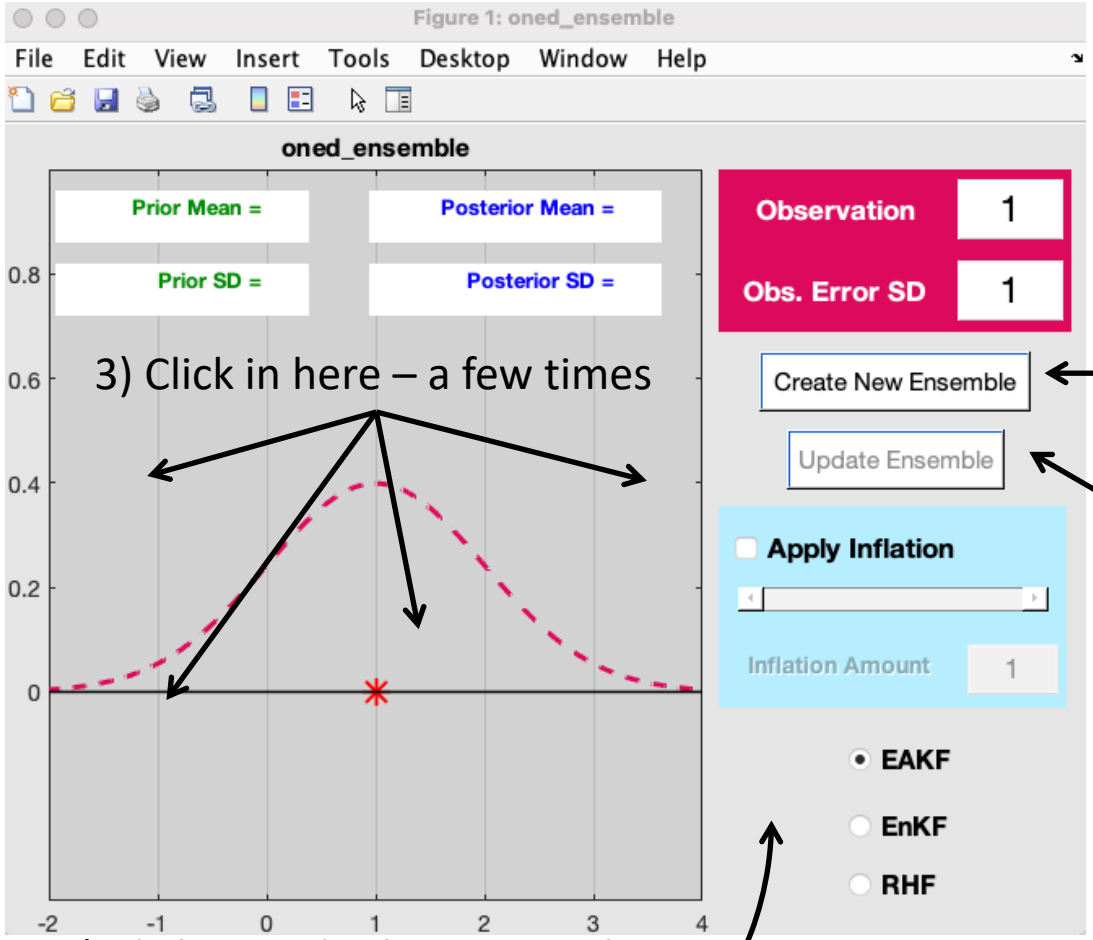
A One-Dimensional Ensemble Kalman Filter: Assimilating an Observation



As we'll discuss later, this is the same as conserving quantiles of the prior ensemble. This Ensemble 'Adjustment' Kalman filter is a type of Quantile Conserving Ensemble Filter.

Matlab Hands-On: oned_ensemble

Purpose: Explore how ensemble filters update a prior ensemble.



1) change these if you want to.

2) Click on Create New Ensemble

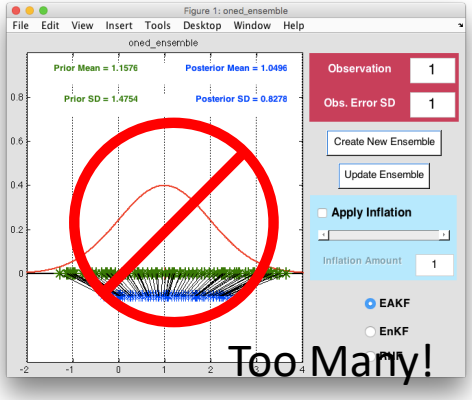
5) Click on Update Ensemble

Ignore the Inflation and EAKF menus for now.

4) Click outside the axis on the light gray (anywhere) to finish defining the ensemble.

Matlab Hands-On: oned_ensemble

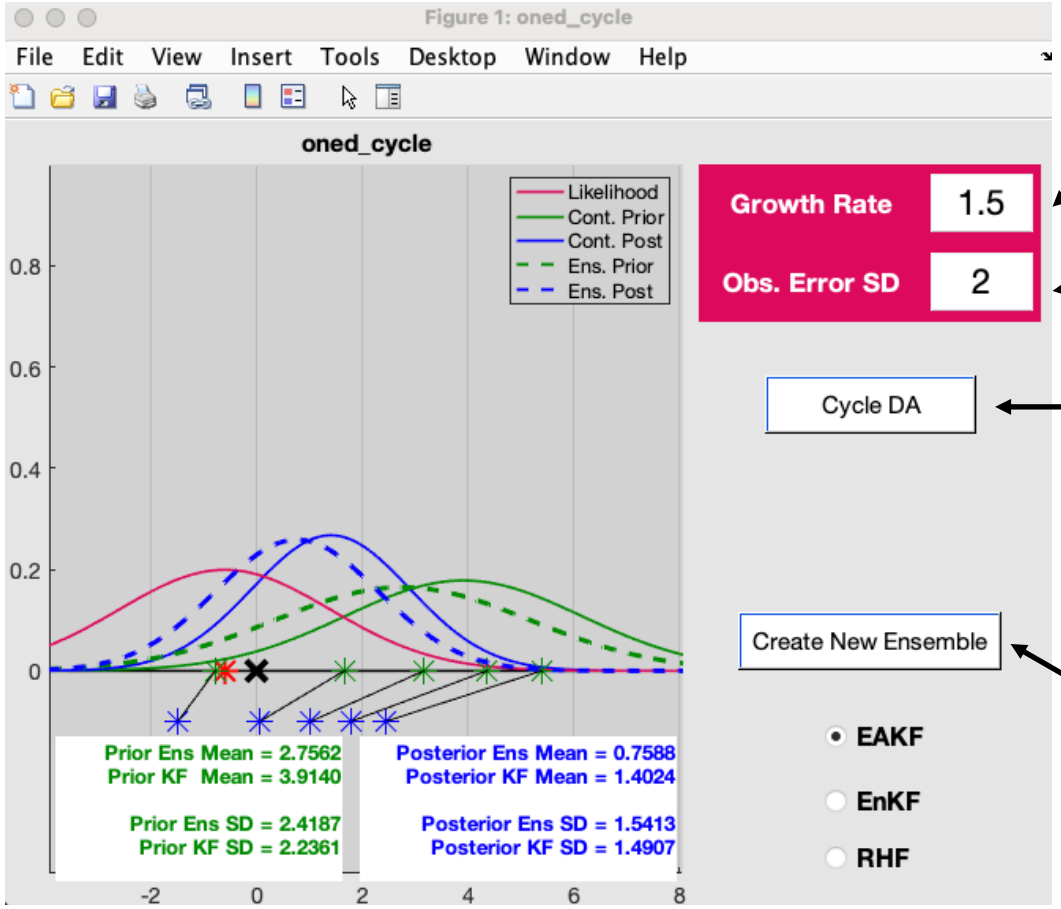
Explorations:



1. Keep your ensembles small, less than 10, for easy viewing.
2. Create a nearly uniformly spaced ensemble. Examine the update.
3. What happens with an ensemble that is confined to one side of the likelihood?
4. What happens with a bimodal ensemble (two clusters of members on either side)?
5. What happens with a single outlier in the ensemble?

Matlab Hands-on: oned_cycle (3)

Purpose: One-dimensional Kalman Filter with linear growth model.



Change the growth rate G.

Change the error standard deviation of observations.

Press to assimilate next observation.

Create an ensemble as in oned_ensemble.

Matlab Hands-on: oned_cycle (3)

See what happens when you create an ensemble and then cycle.

How do results compare to the continuous KF?

What elements of the ensemble vary with time?

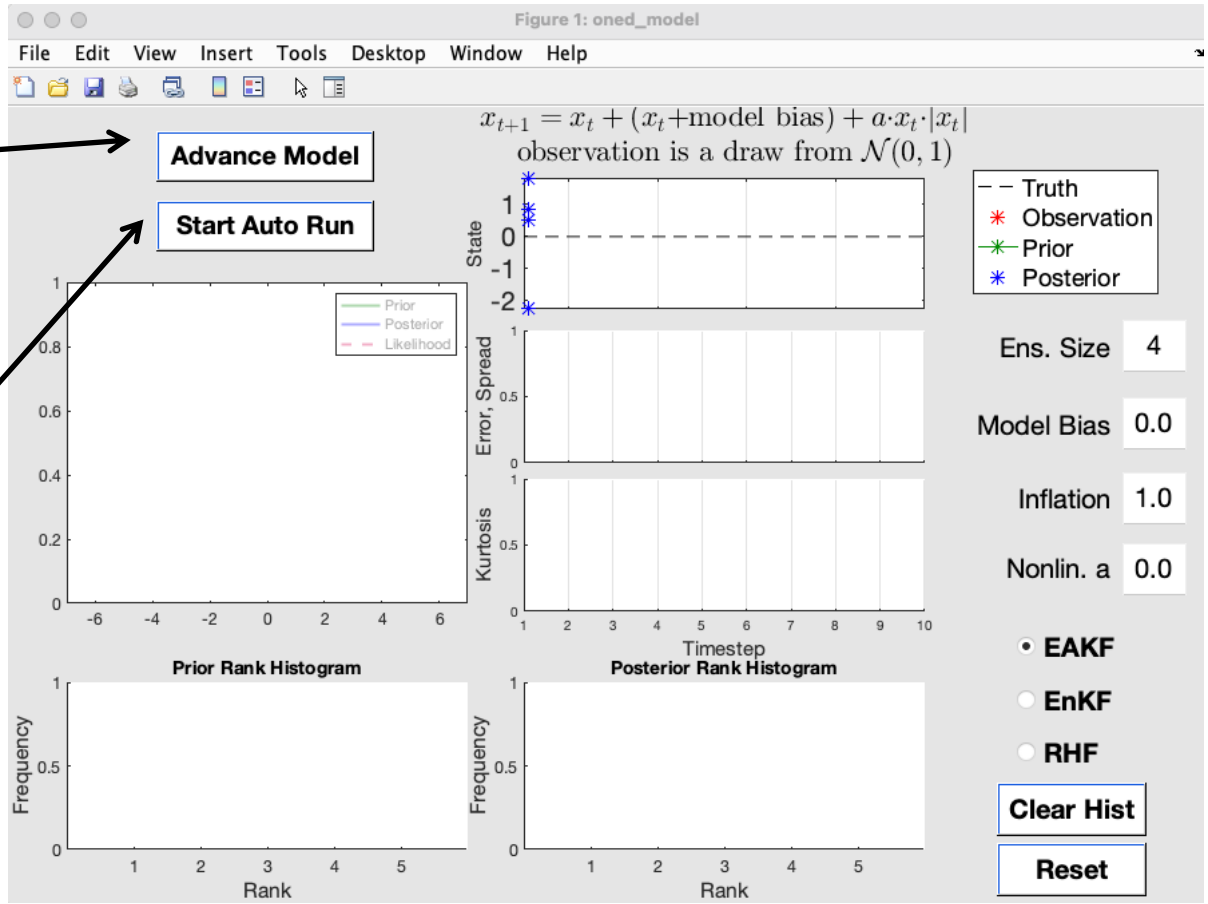
Matlab Hands-On: oned_model

Purpose:

- Explore behavior of a complete 1-D ensemble filter for a linear system.
- Look at the behavior of different ensemble sizes.

Top button allows alternating model advance and assimilation steps.

Or automatically sequence advances and assimilations.



Notes:

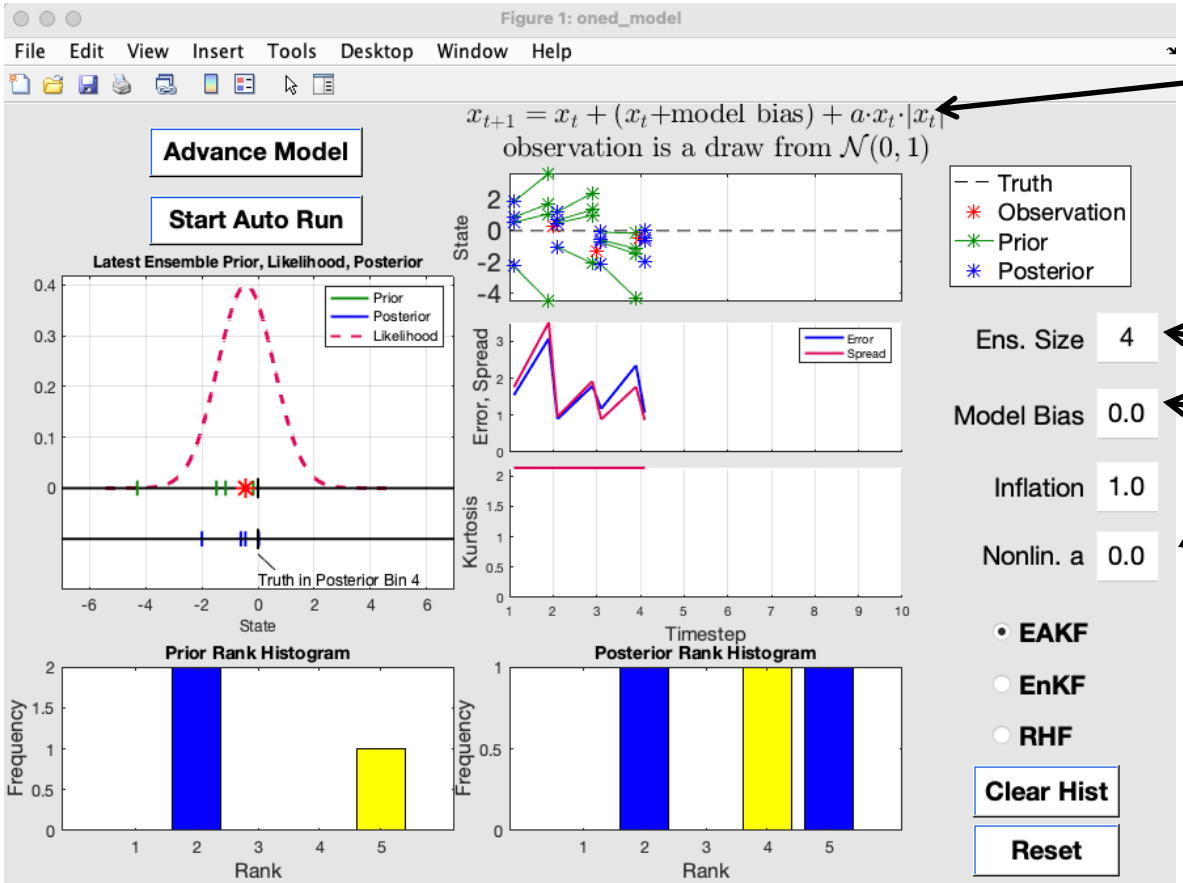
The 'truth' is always 0.

Observation noise is a draw from $Normal(0,1)$

Matlab Hands-On: oned_model

Purpose:

- Explore behavior of a complete 1-D ensemble filter for a linear system.
- Look at the behavior of different ensemble sizes.



This is the equation for the model time tendency.

Change the ensemble size or the model parameters.

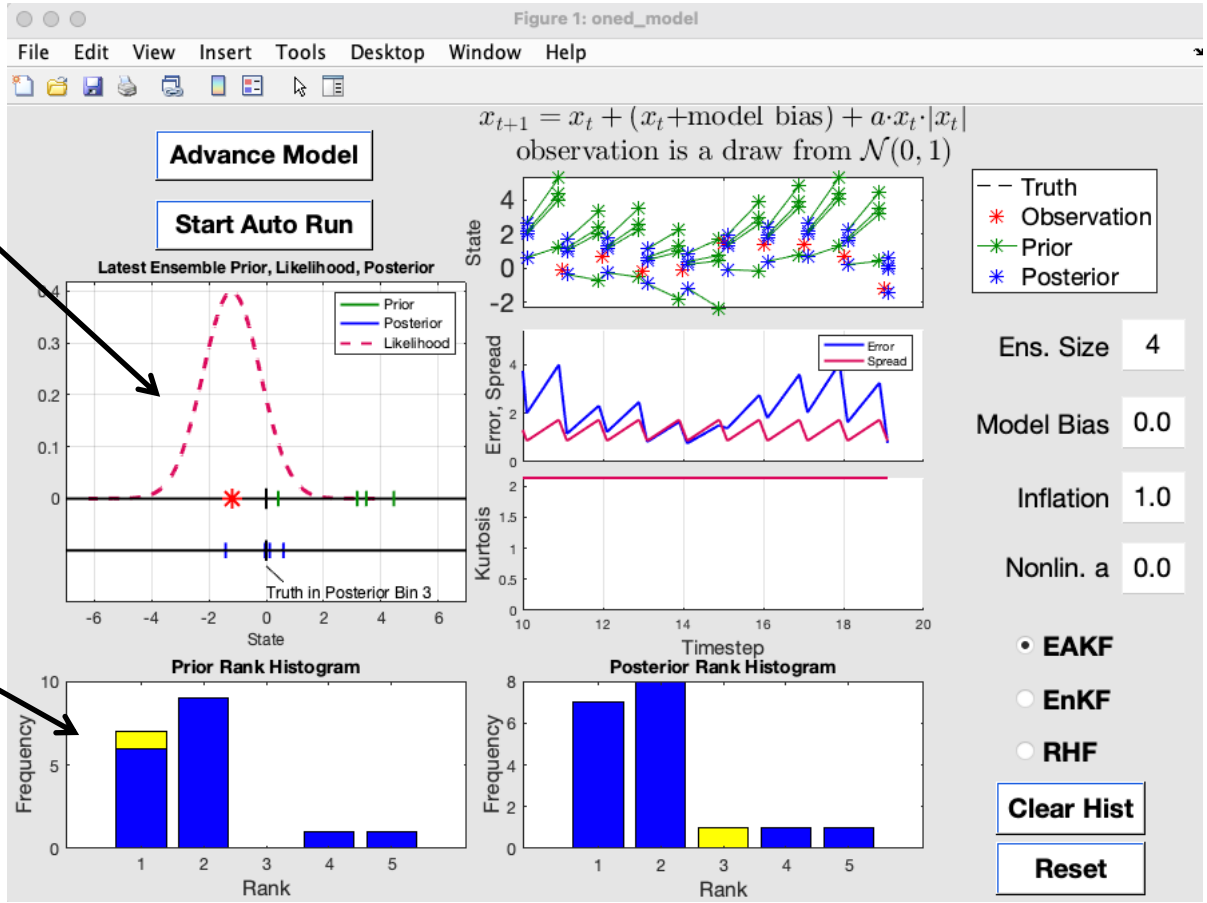
Matlab Hands-On: oned_model

Purpose:

- Explore behavior of a complete 1-D ensemble filter for a linear system.
- Look at the behavior of different ensemble sizes.

Prior ensemble members green ticks, Posterior blue, observation is *. Truth is always 0.

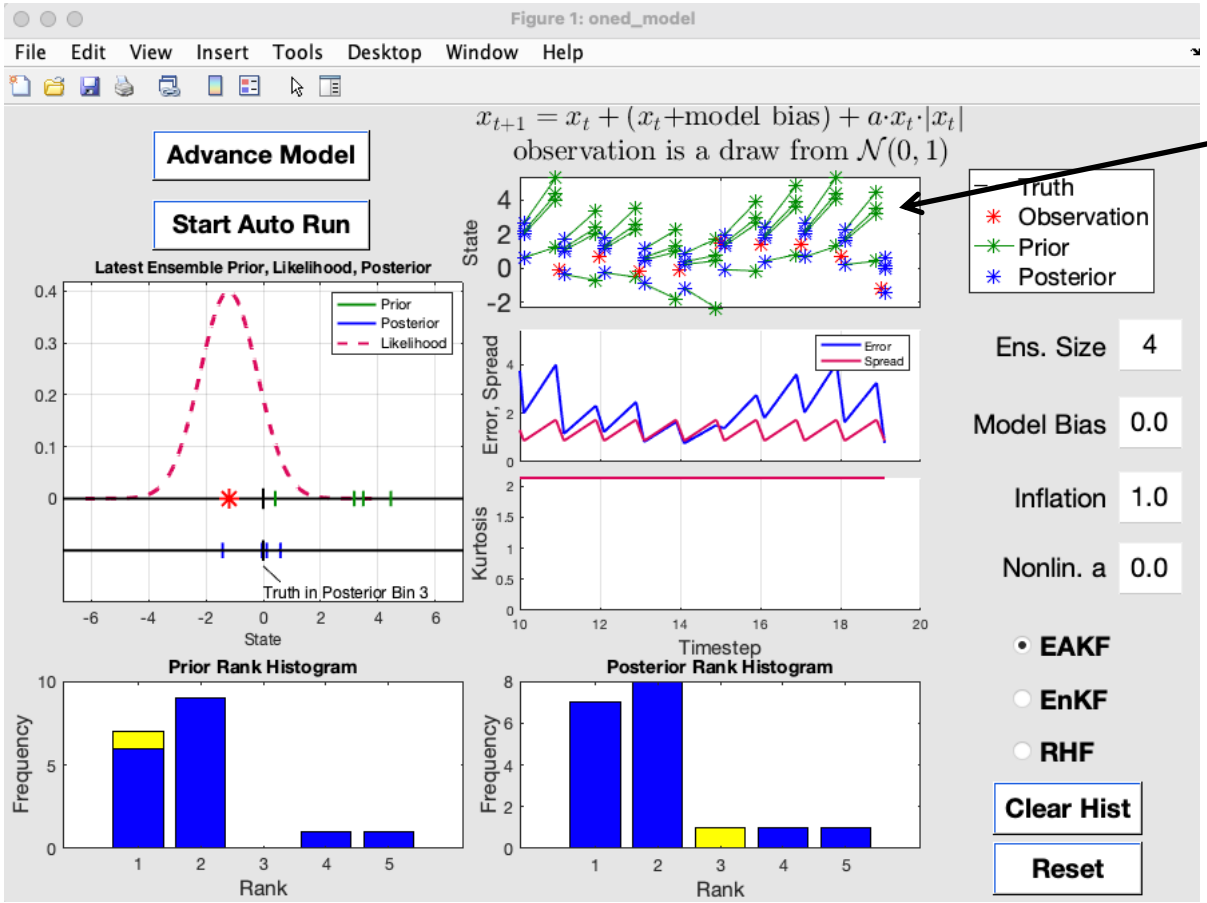
More on rank histograms later.



Matlab Hands-On: oned_model

Purpose:

- Explore behavior of a complete 1-D ensemble filter for a linear system.
- Look at the behavior of different ensemble sizes.

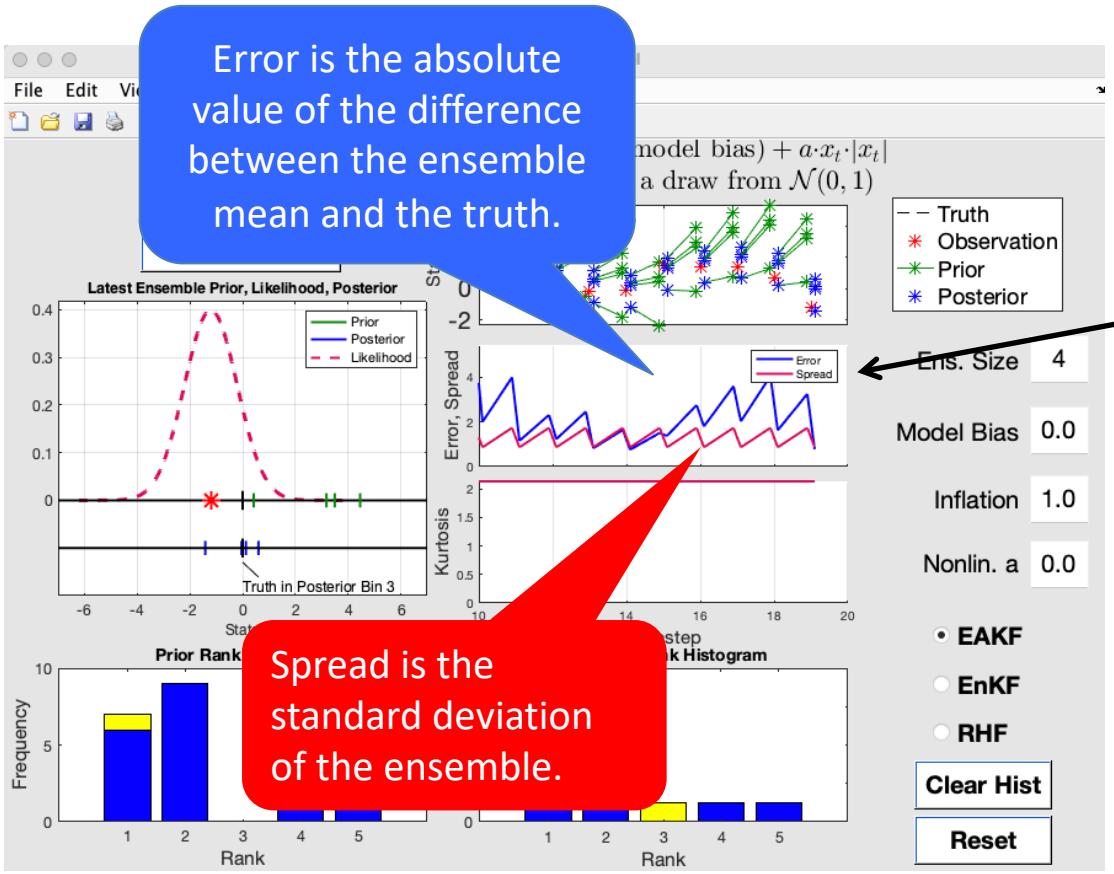


A time series of the assimilation. Line segments show forecast evolution. Most recent prior, observation, and posterior are same as in upper left window but plotted with a vertical axis.

Matlab Hands-On: oned_model

Purpose:

- Explore behavior of a complete 1-D ensemble filter for a linear system.
- Look at the behavior of different ensemble sizes.

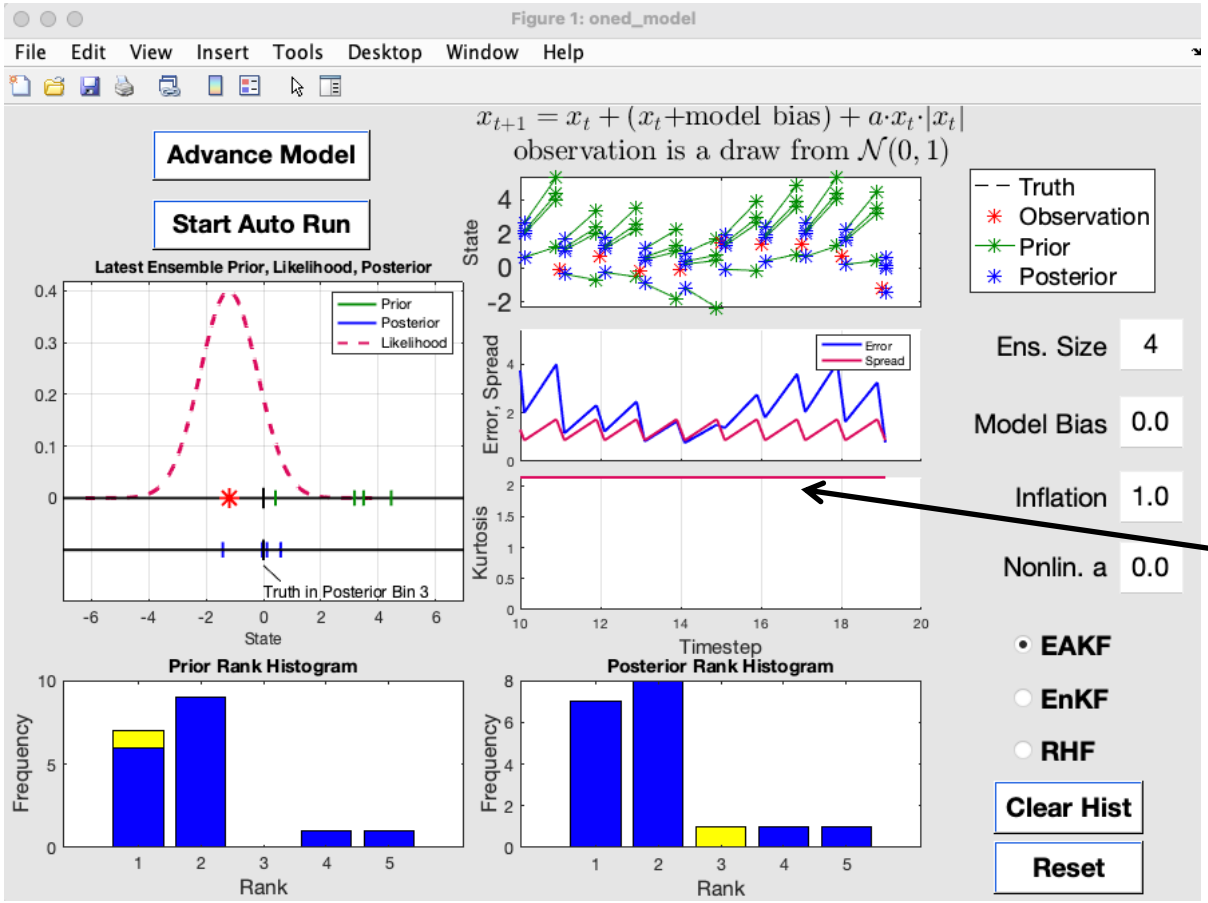


Time series of error and spread. Sawtooth pattern because prior and posterior values are shown for each time.

Matlab Hands-On: oned_model

Purpose:

- Explore behavior of a complete 1-D ensemble filter for a linear system.
- Look at the behavior of different ensemble sizes.



Kurtosis as function of time. Stays constant here.

Matlab Hands-On: oned_model

Explorations:

1. Step through a sequence of advances and assimilations with the top button. Watch the evolution of the ensemble, the error and spread.
2. How does a larger ensemble size (< 10 is easiest to see) act?

- Compare the error and spread for different ensemble sizes.
- Note the time behavior of the error and spread.

3. Let the model run freely using the second button.