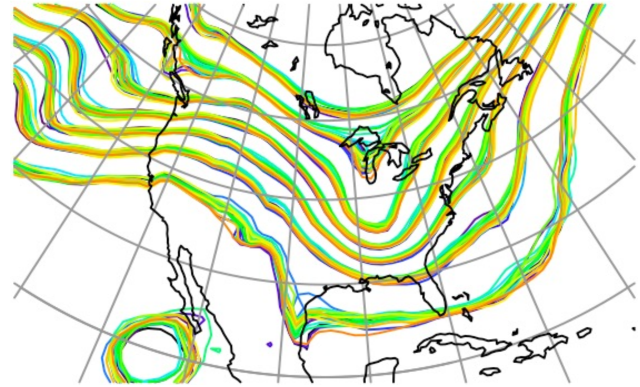
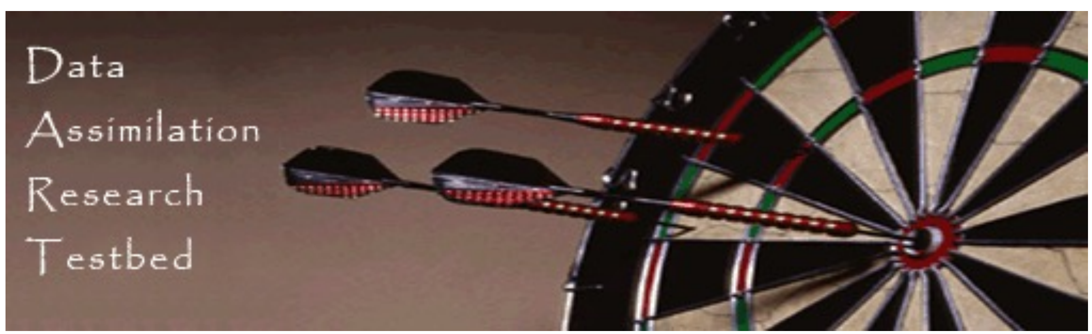


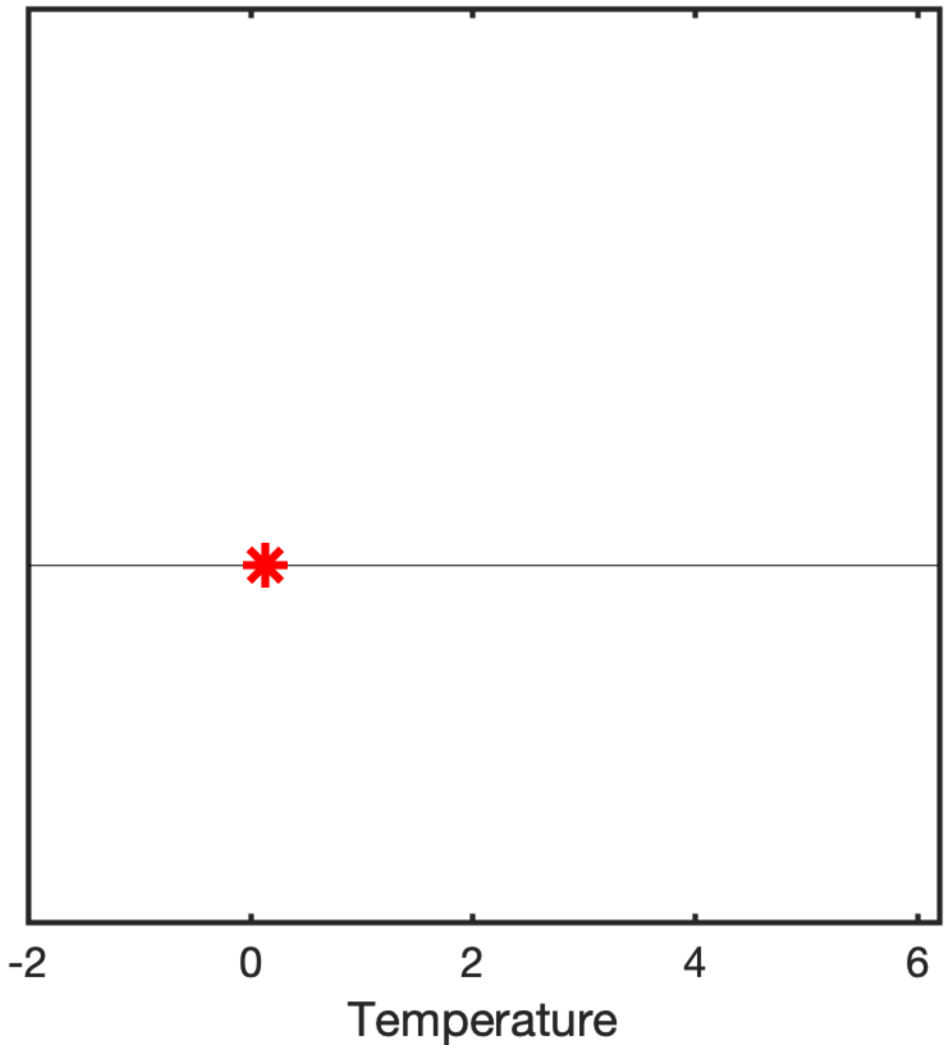
A General Ensemble Filtering Framework: Improved Data Assimilation (DA) for Tracers

Jeff Anderson, CISL Data Assimilation Research Section

Chris Riedel, ASP Fairuz Ishraque, SIParCS



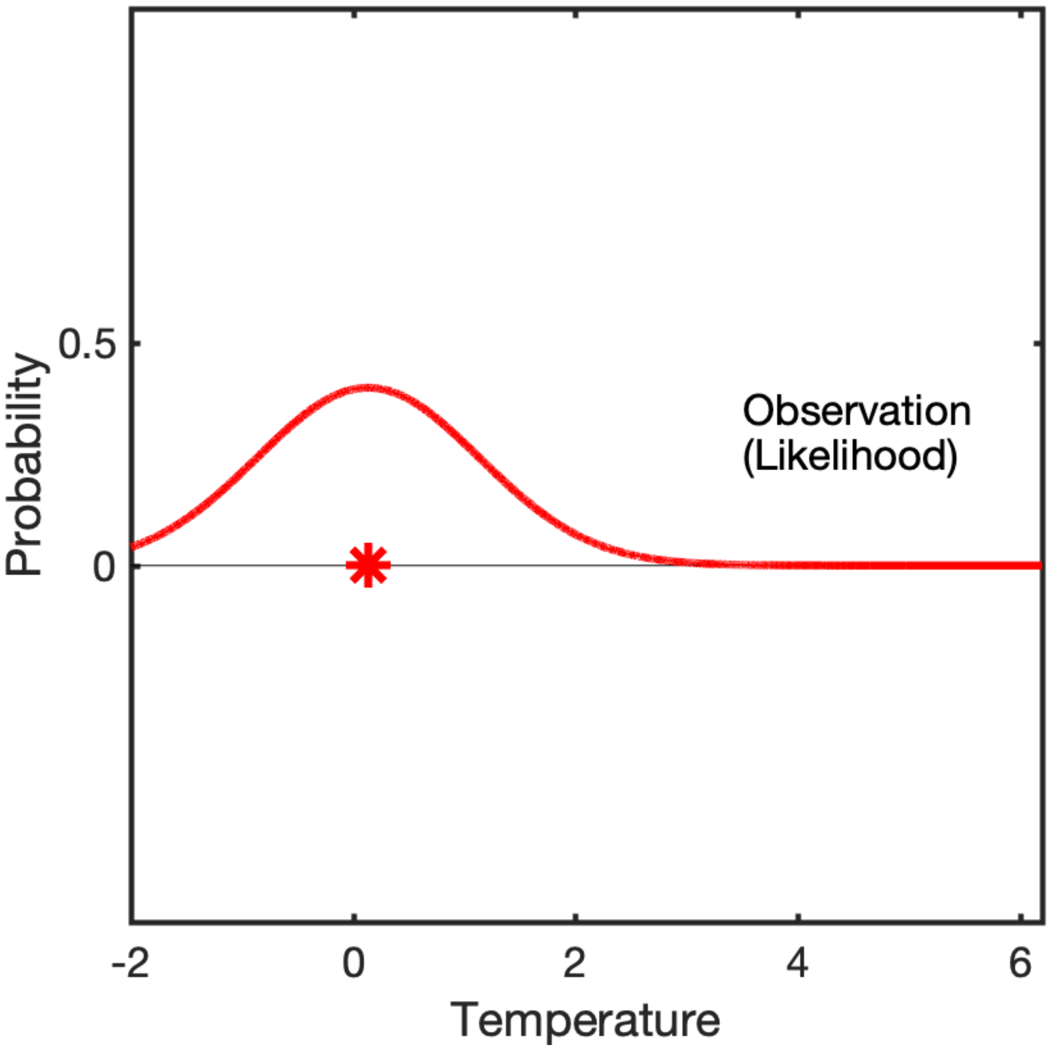
Should I Worry About Ice Going Down the Hill?



Check the ML weather station temperature.

Looks like it's above freezing, so it should be safe, but…

Should I Worry About Ice Going Down the Hill?



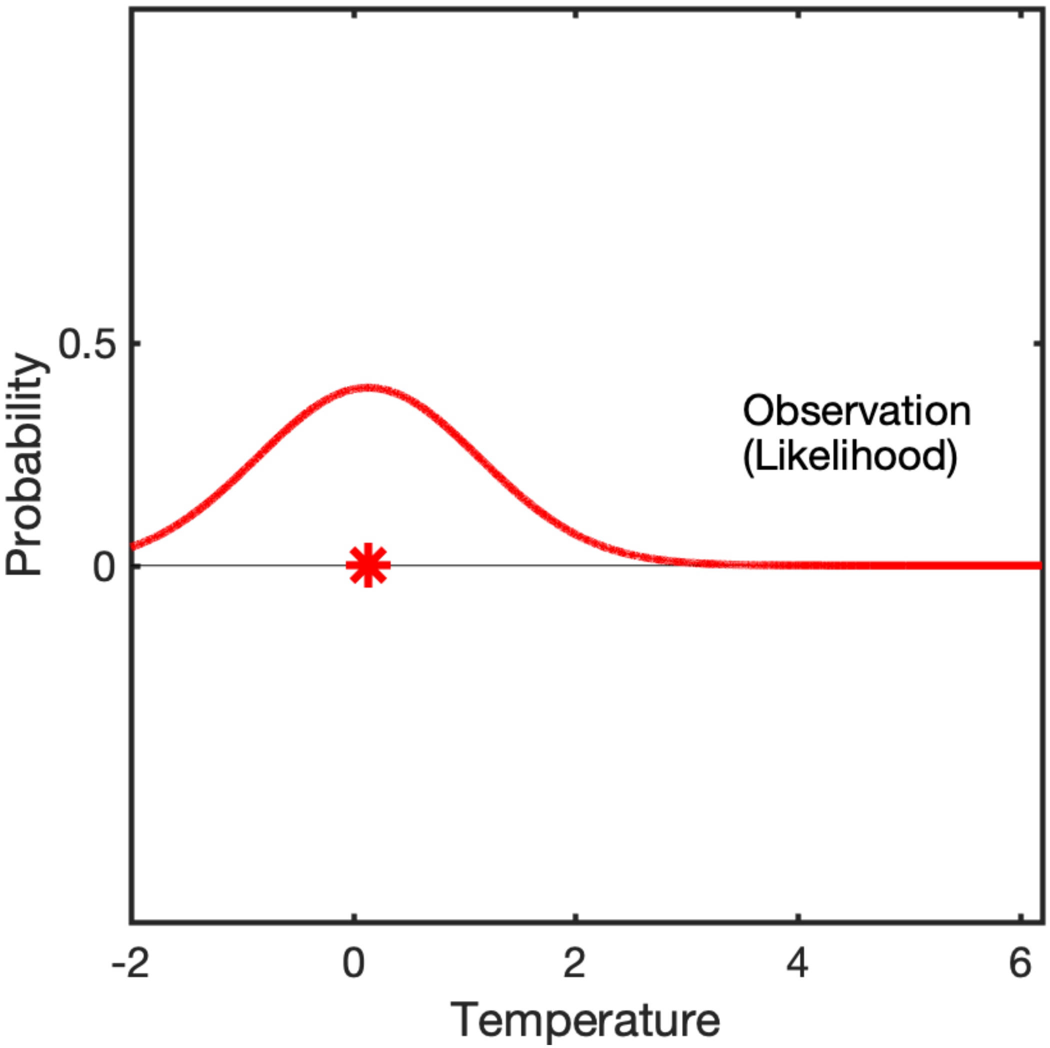
Check the ML weather station temperature.

Looks like it's above freezing, so it should be safe, but all instruments have errors.

Likelihood is probability that the real temperature has a certain value when 0.2 degrees is measured.

Comes from engineers.

Should I Worry About Ice Going Down the Hill?

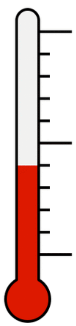


I need more information.

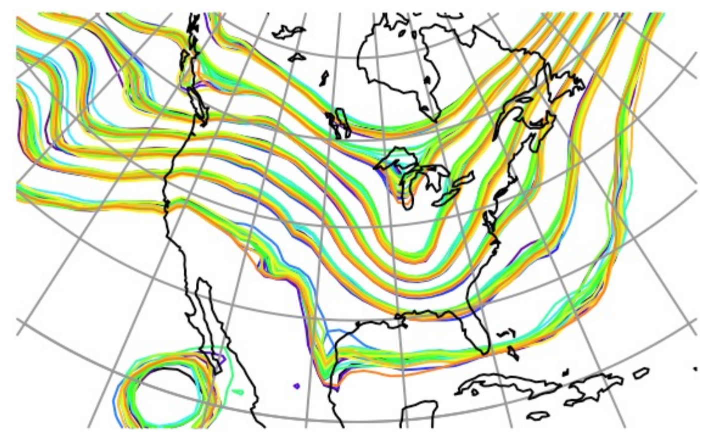
I'll use Data Assimilation (DA) to learn more about the ML temperature probability.

What is Data Assimilation?

Observations combined with a Model forecast…



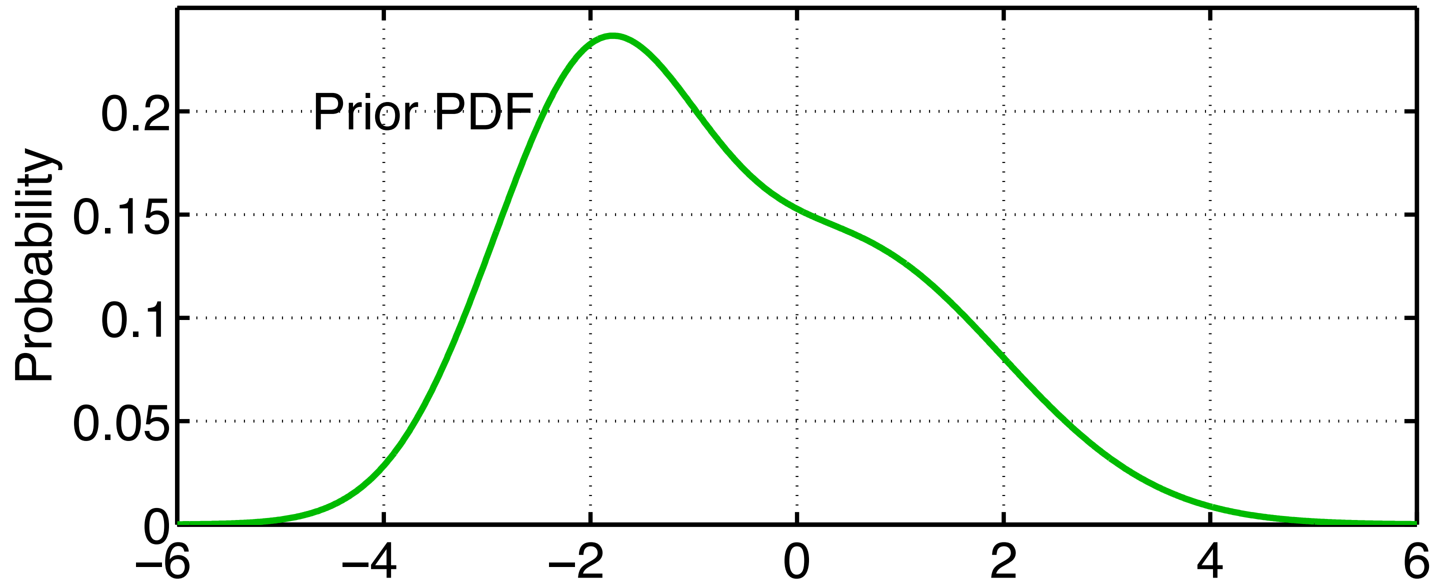
+



…to produce an analysis
(best possible estimate).

Bayes' Rule

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



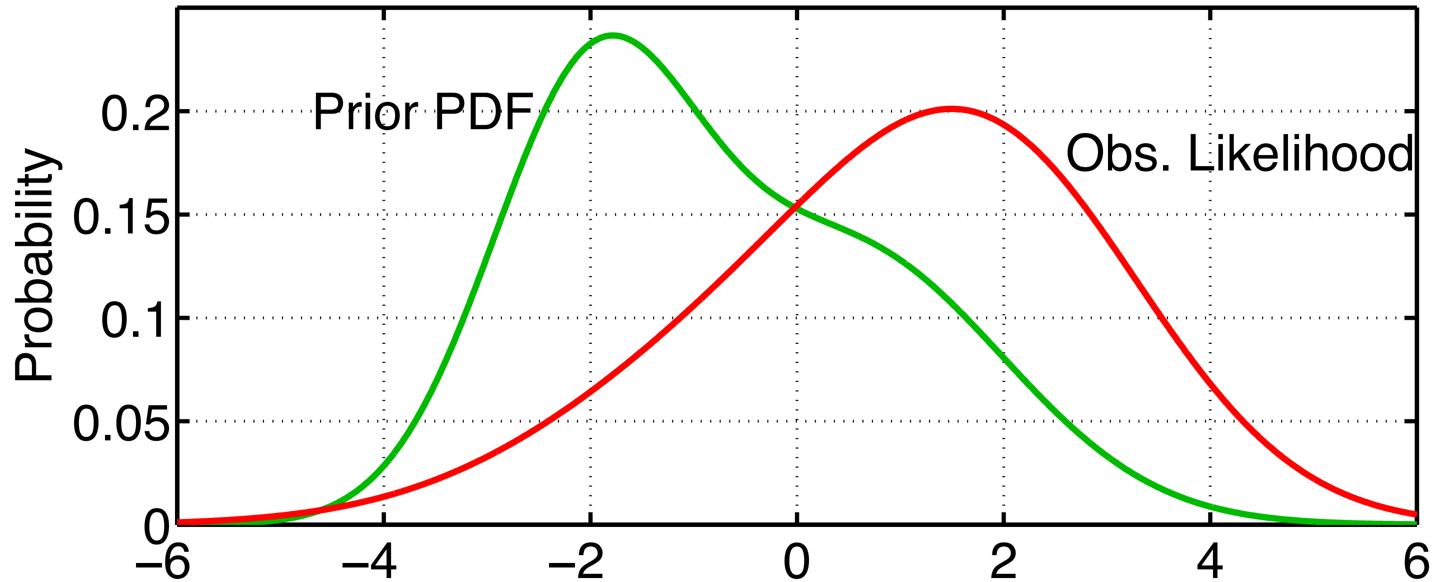
A : Forecast estimate from model (prior).

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

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

Bayes' Rule

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



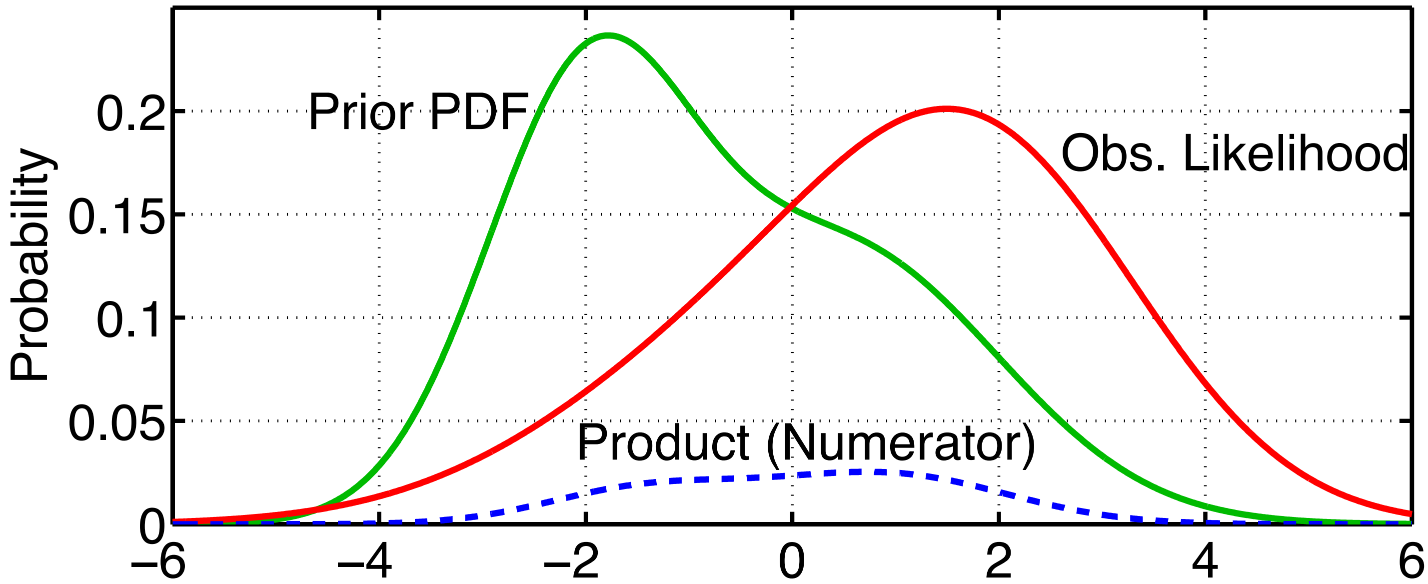
A : Forecast estimate from model (prior).

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

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

Bayes' Rule

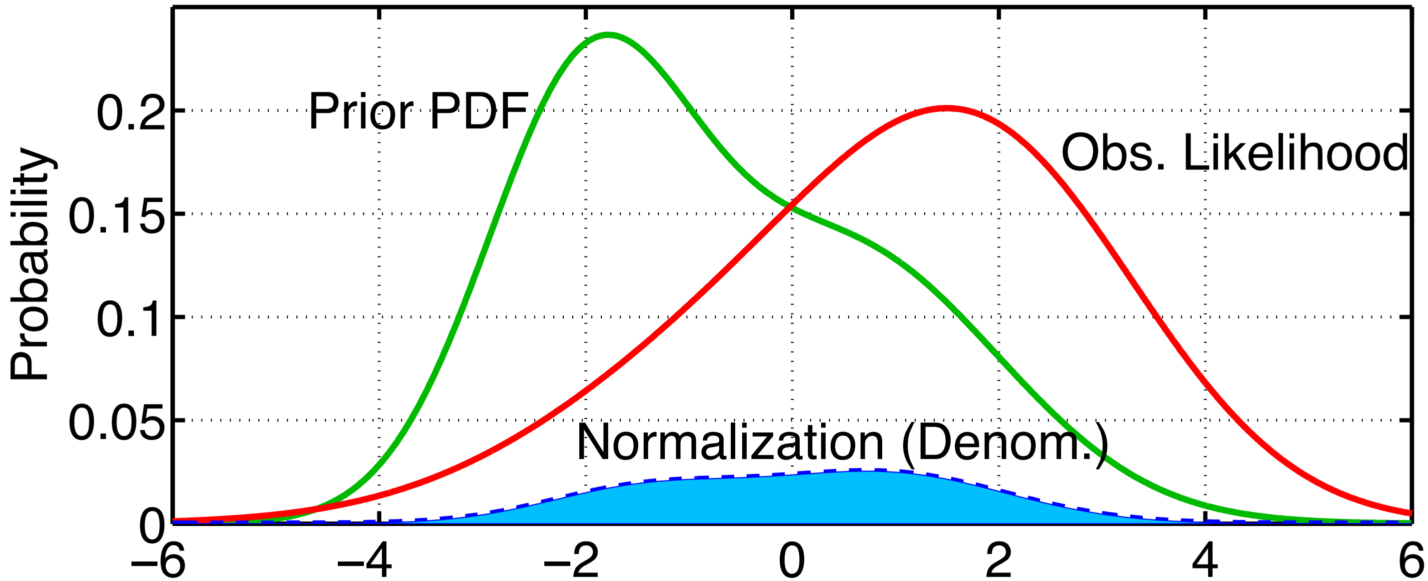
$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{Normalization}$$



A	: Forecast estimate from model (prior).
$p(B/A)$	: Observation likelihood.
$p(A/B)$	: Analysis (posterior) estimate combines A and B.

Bayes' Rule

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



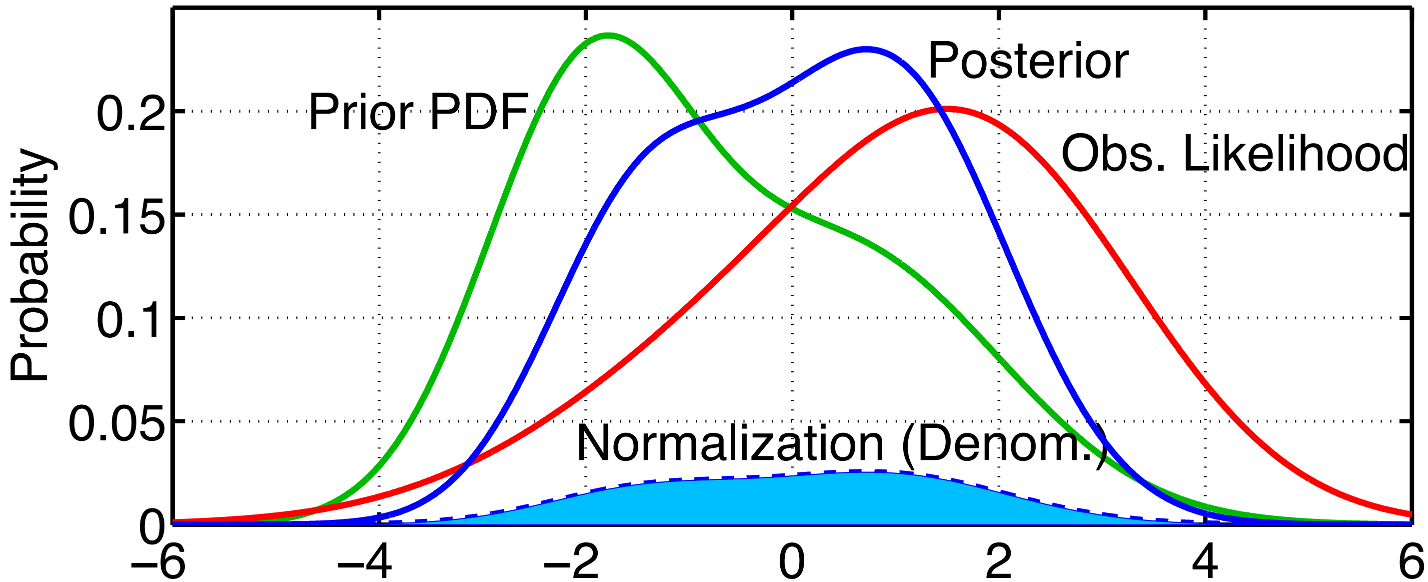
A : Forecast estimate from model (prior).

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

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

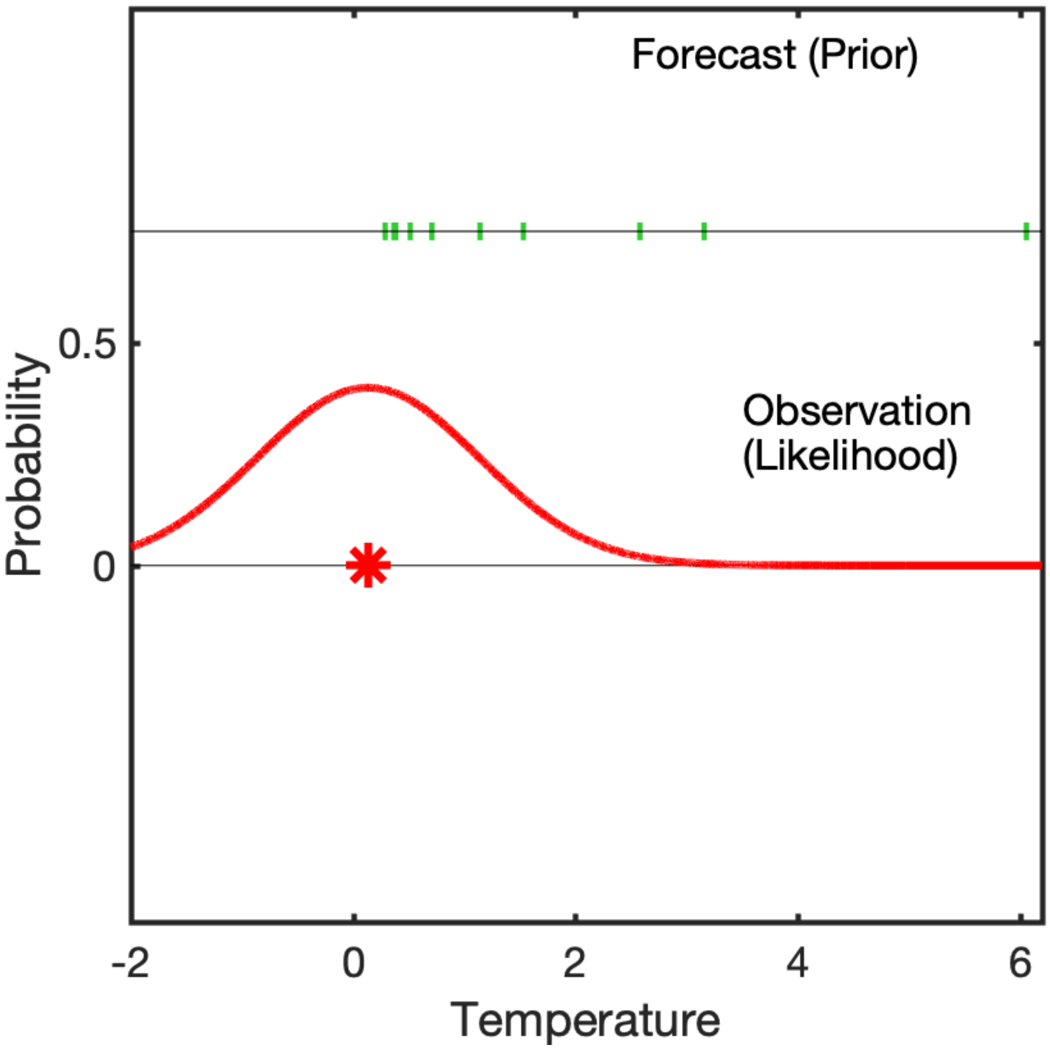
Bayes' Rule

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



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?

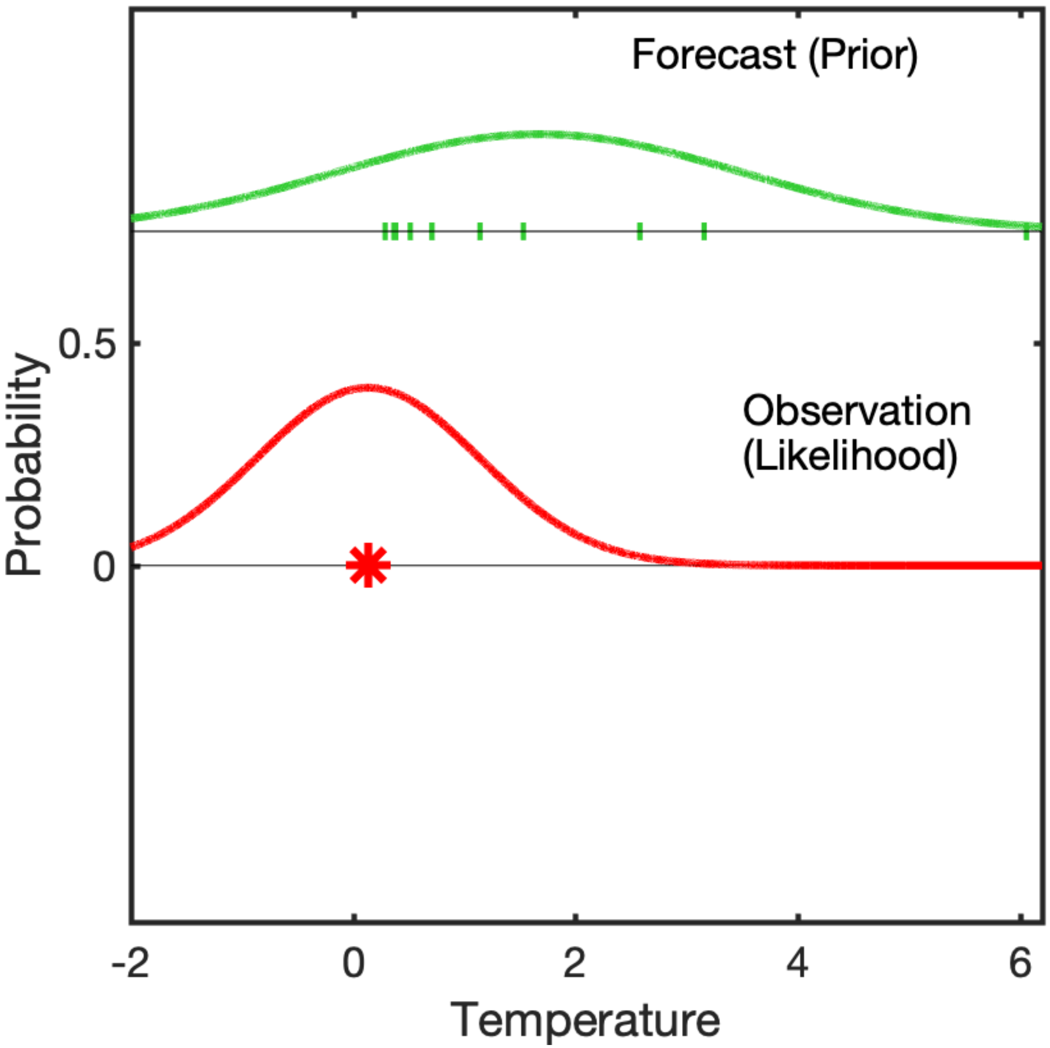


In DAReS, we do *ensemble* DA.

Run an 'ensemble' of forecasts (10 here).

Get 10 forecasts of the ML temperature.

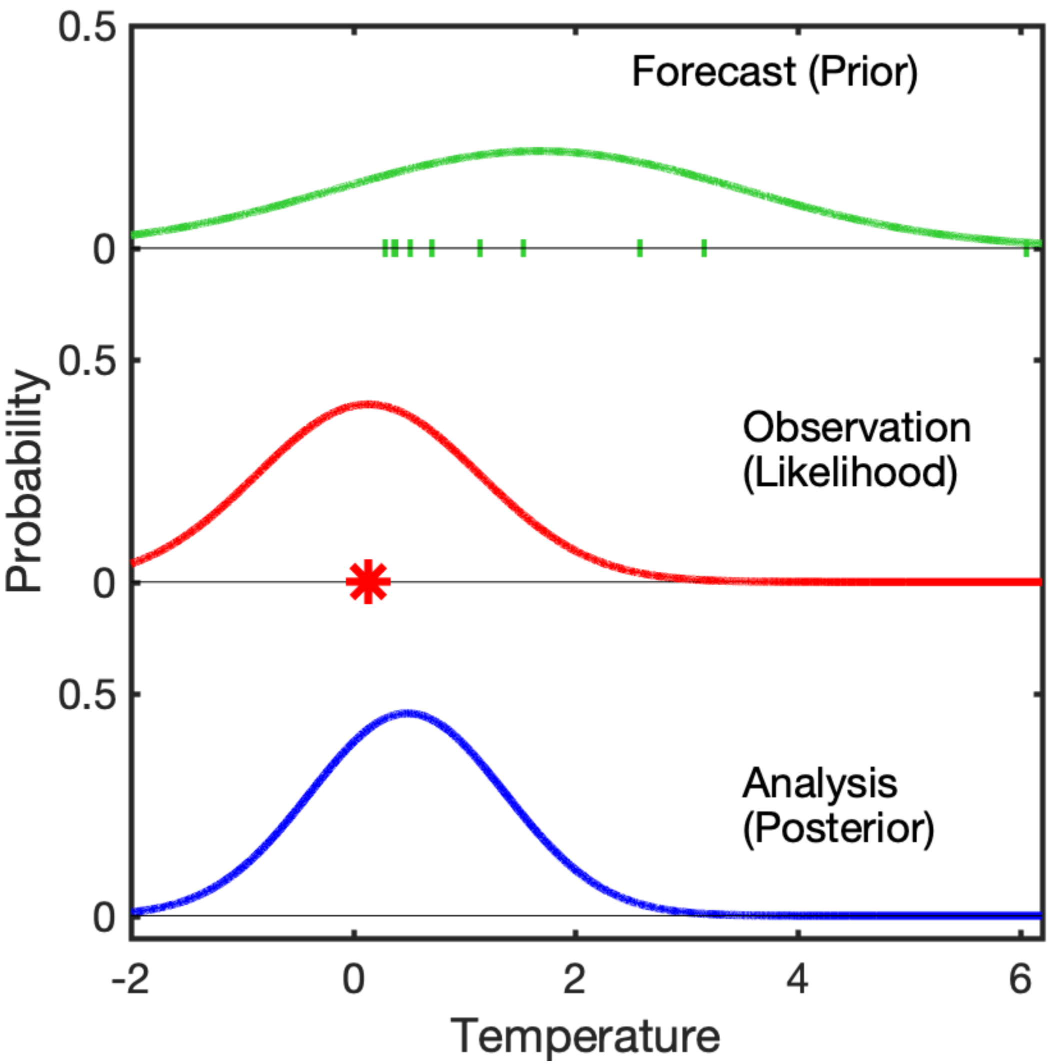
Should I Worry About Ice Going Down the Hill?



We get a continuous probability distribution by fitting some function to the ensemble.

This example fits a normal distribution using the ensemble mean and standard deviation.

Should I Worry About Ice Going Down the Hill?

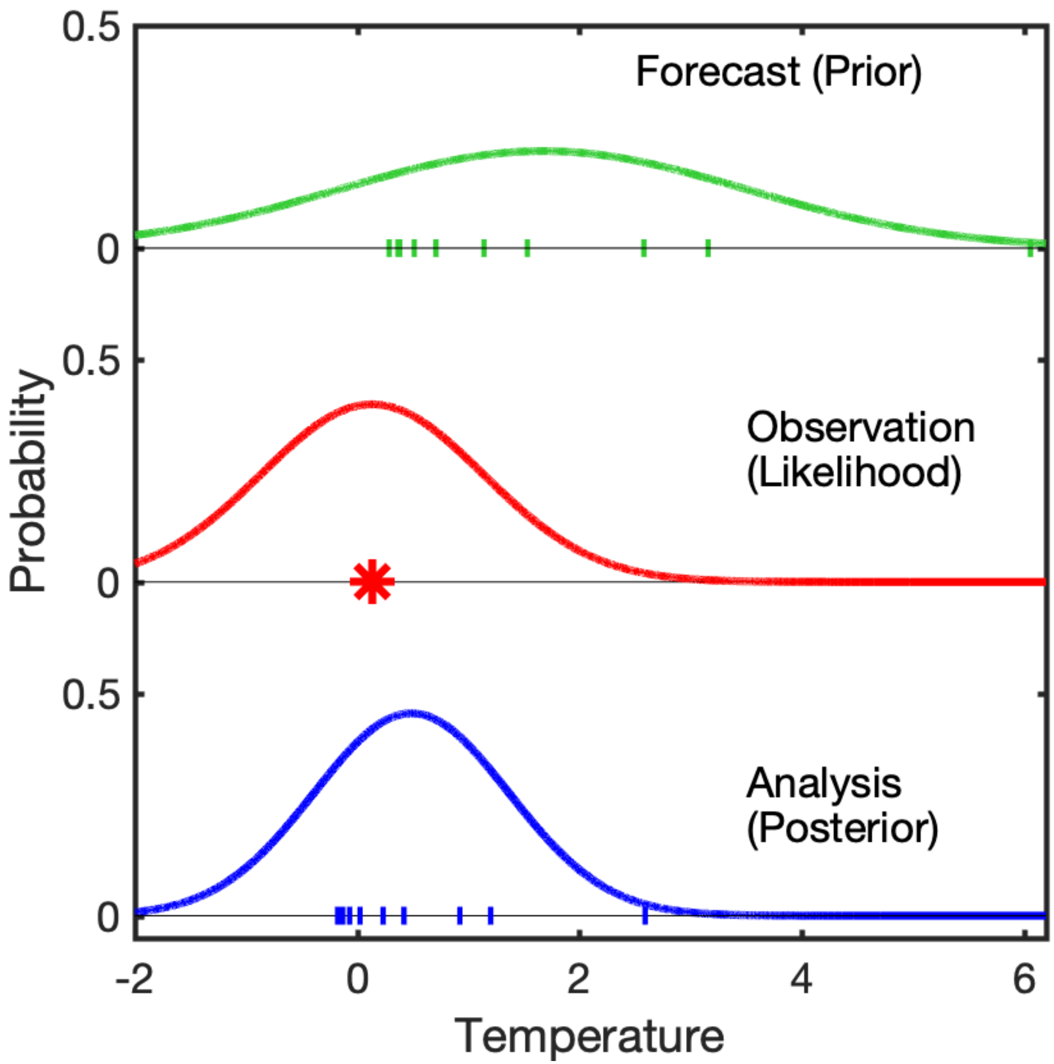


We then use Bayes theorem to get an analysis distribution.

Since both the forecast and likelihood are normal here, the analysis is normal, too.

This should be better than either the forecast or the observation.

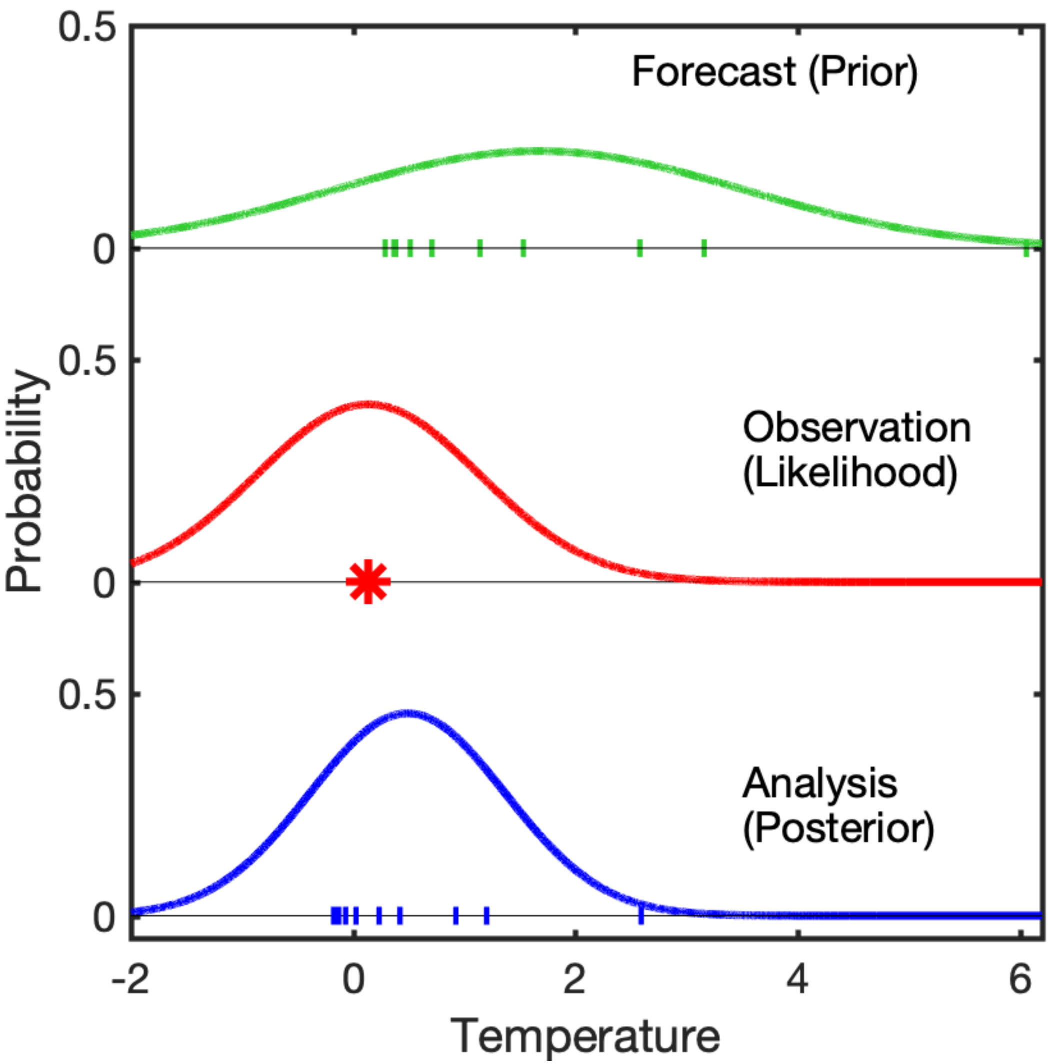
Should I Worry About Ice Going Down the Hill?



Finally, we need to get an analysis ensemble.

This is done with a standard method when the analysis is normal.

Should I Worry About Ice Going Down the Hill?

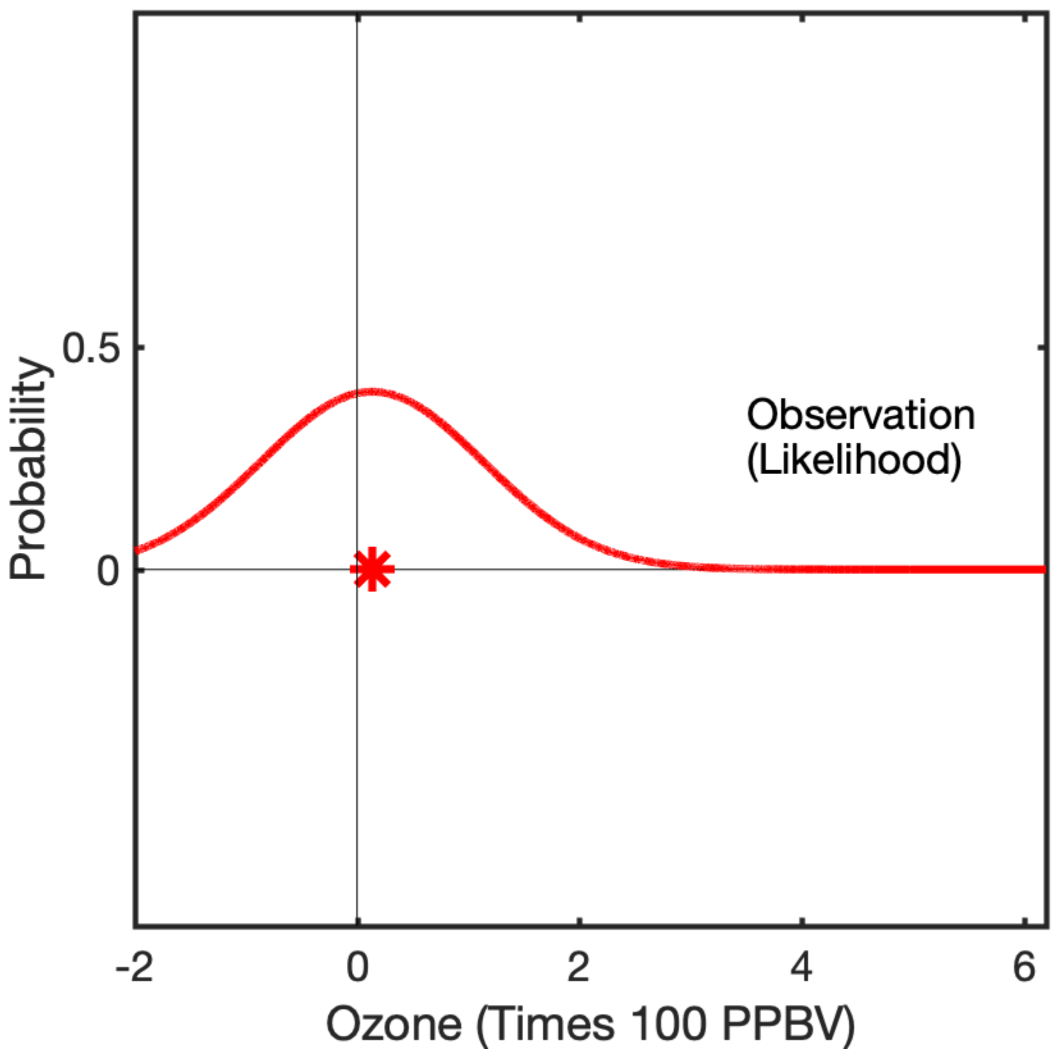


This method of using a normal likelihood with a normal forecast is an ensemble Kalman filter.

These work very well for numerical weather prediction.

This is the most common method used in DART.

Should I Worry About Air Quality Going Down the Hill?



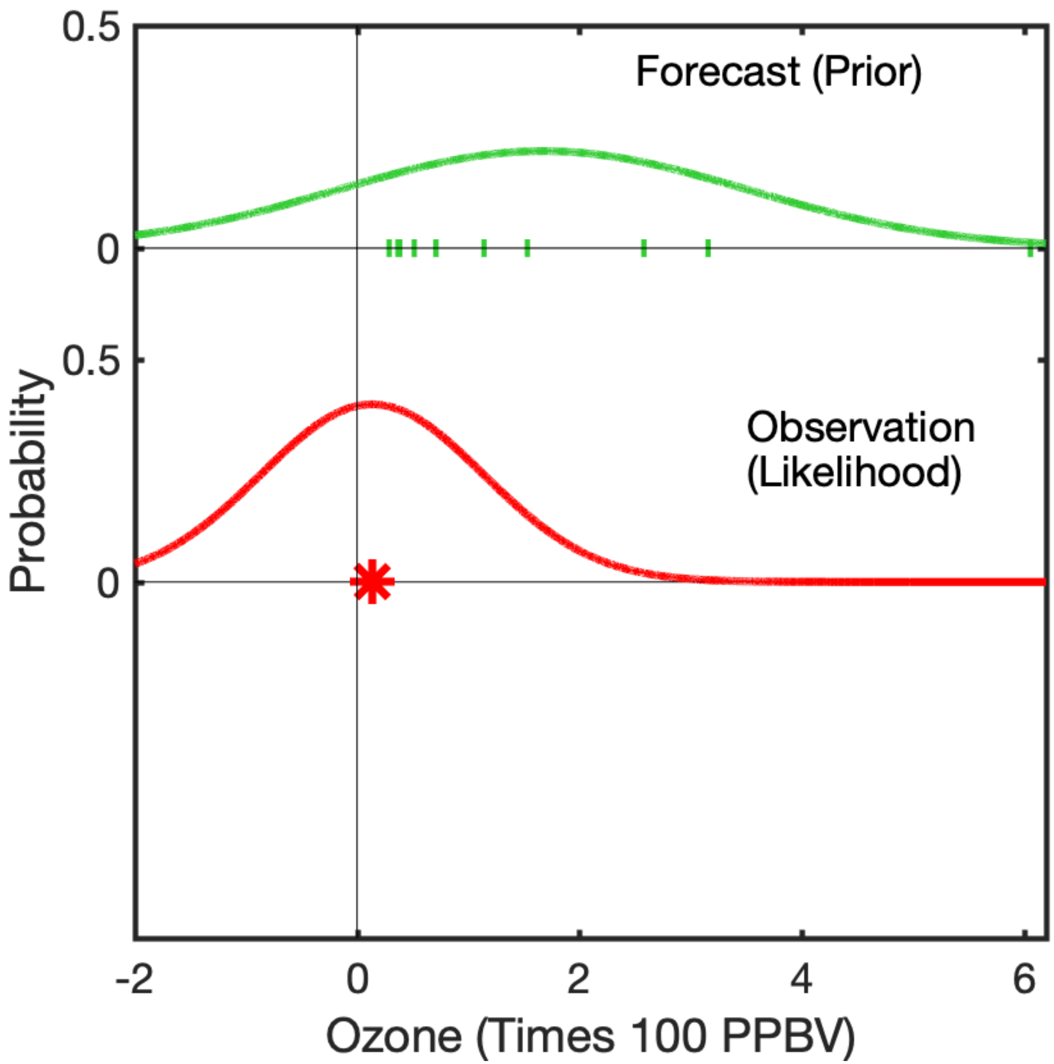
Check local air quality measurement.

Will find that information about the likelihood is limited. Mostly just a normal.

BUT, ozone concentration has to be positive.

All that probability of negative values is troubling.

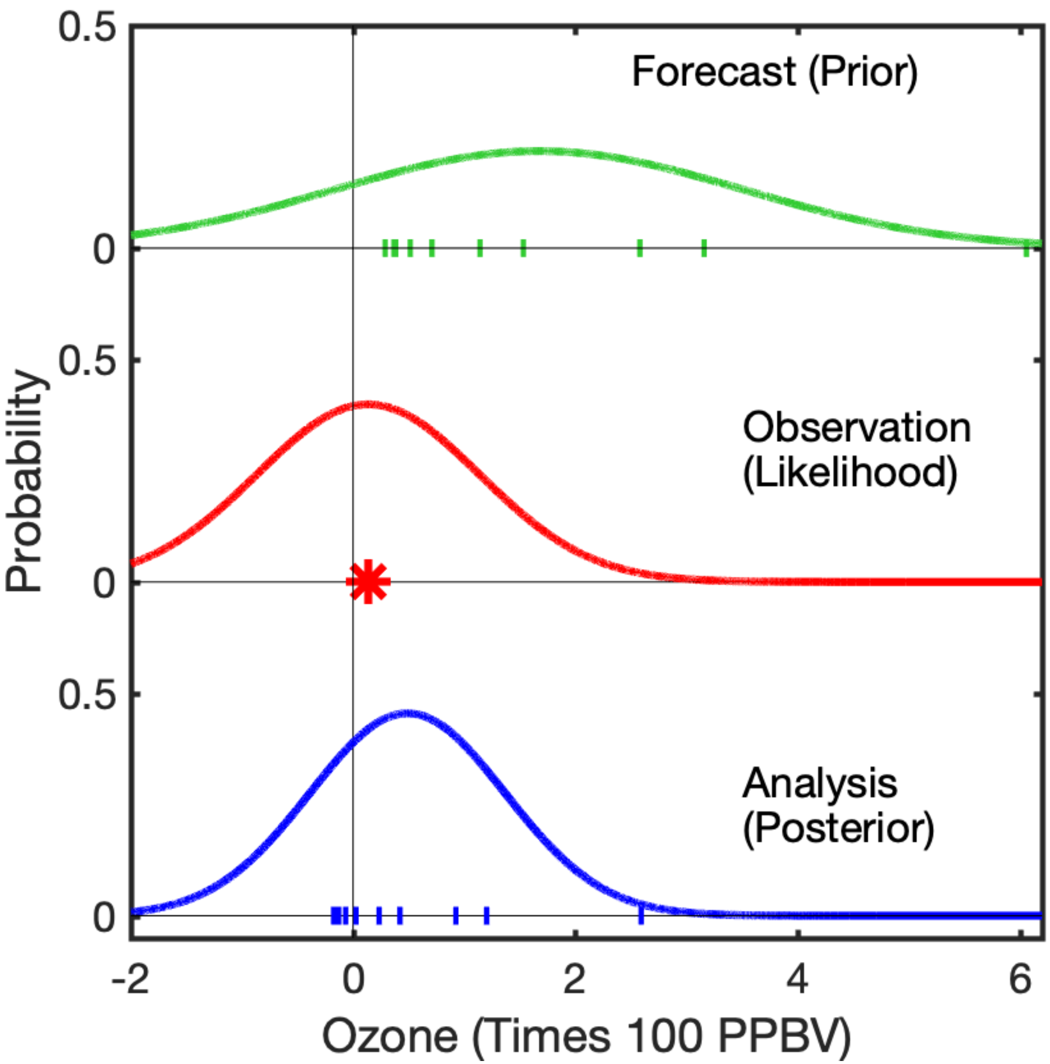
Should I Worry About Air Quality Going Down the Hill?



Forecast model knows ozone must be positive.

BUT, fitting a normal leads to probability of negative.

Should I Worry About Air Quality Going Down the Hill?

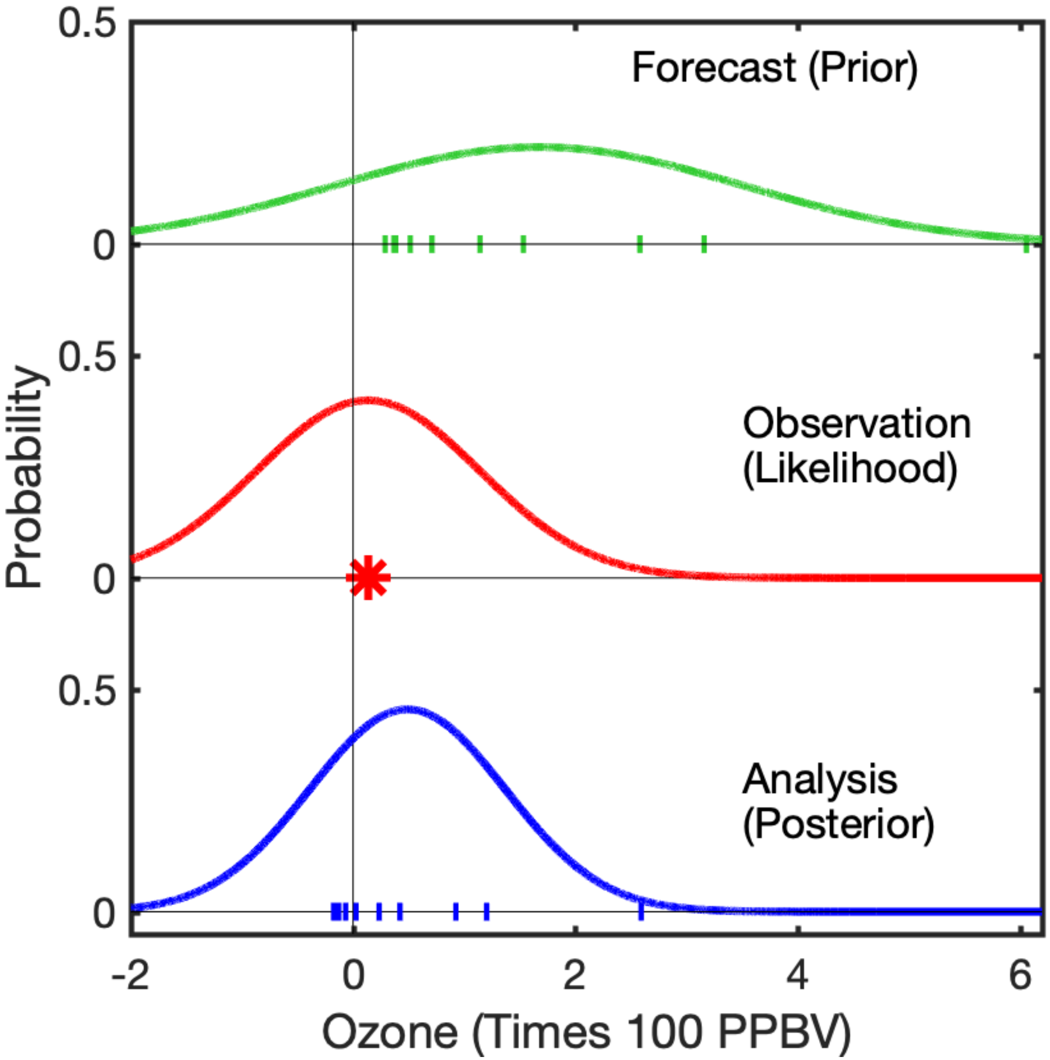


Doing the DA can lead to negative ensemble members.

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?

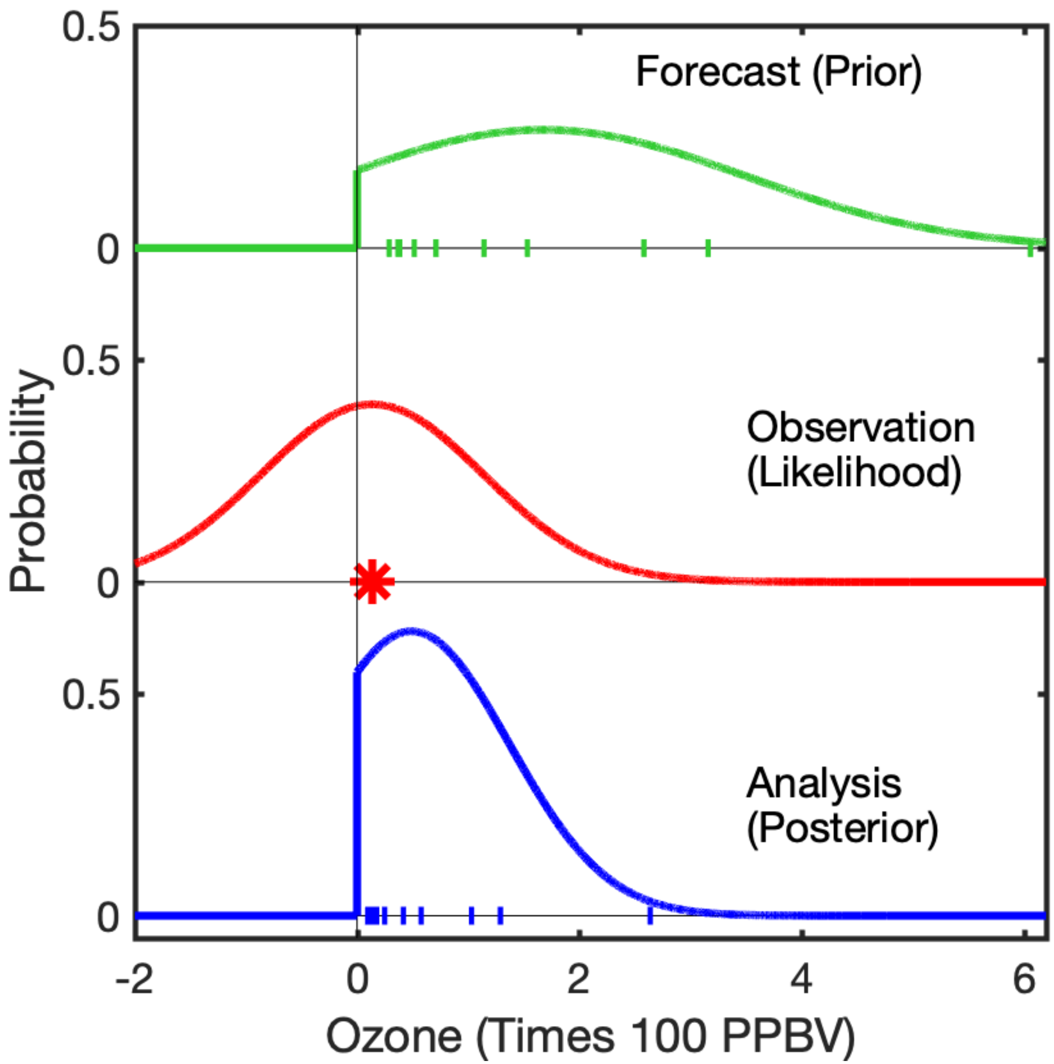


Normal distributions are good choice for unbounded quantities like temperature or winds.

Weather prediction works well because of this.

Many other applications may get poor results.

Should I Worry About Air Quality Going Down the Hill?



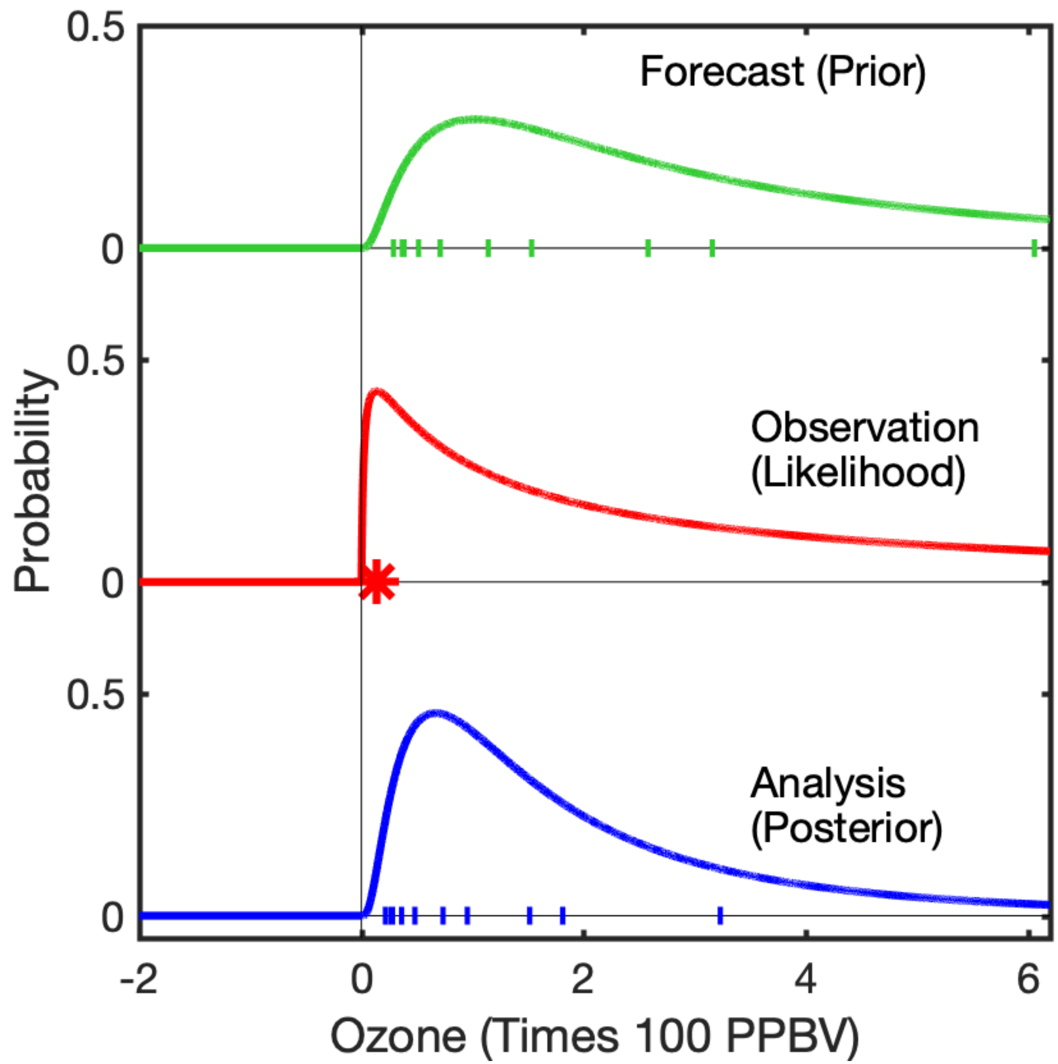
One solution: Fit a bounded distribution to the forecast ensemble.

Leads to bounded analysis.

There was no efficient general solution to get a posterior ensemble from such distributions.

New work in DAReS solves this problem!

Should I Worry About Air Quality Going Down the Hill?



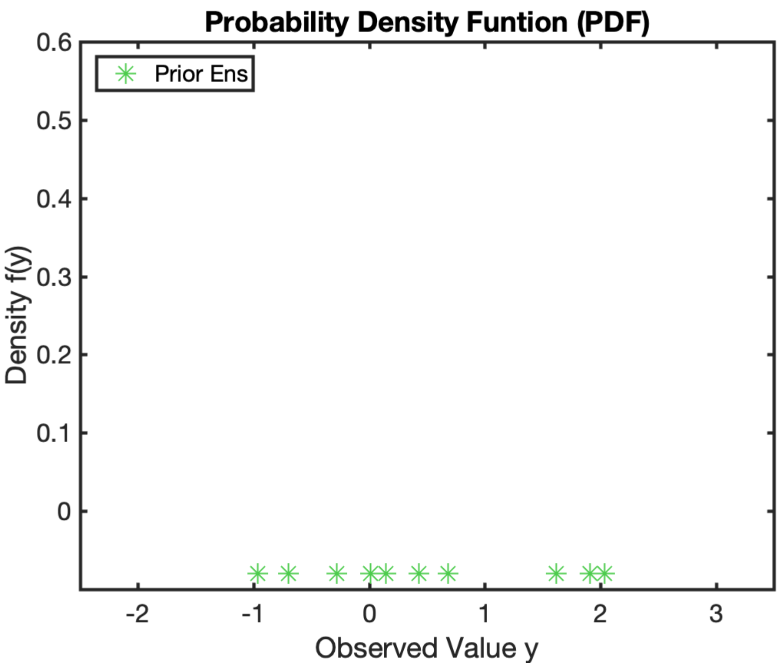
Existing information about likelihoods for bounded quantities is not generally sufficient either.

Instrument engineers had no motivation to explore this problem.

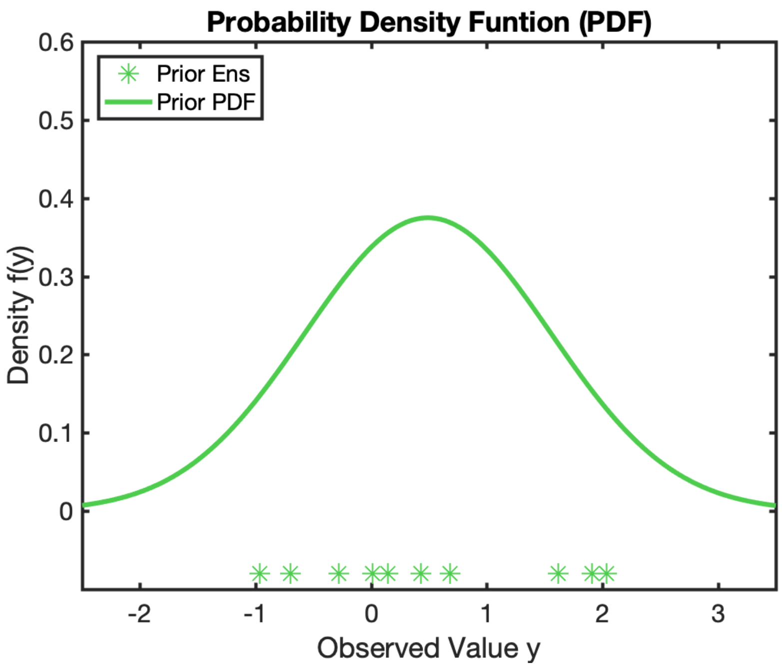
Cool collaborations with instrument builders are now possible to improve DA.

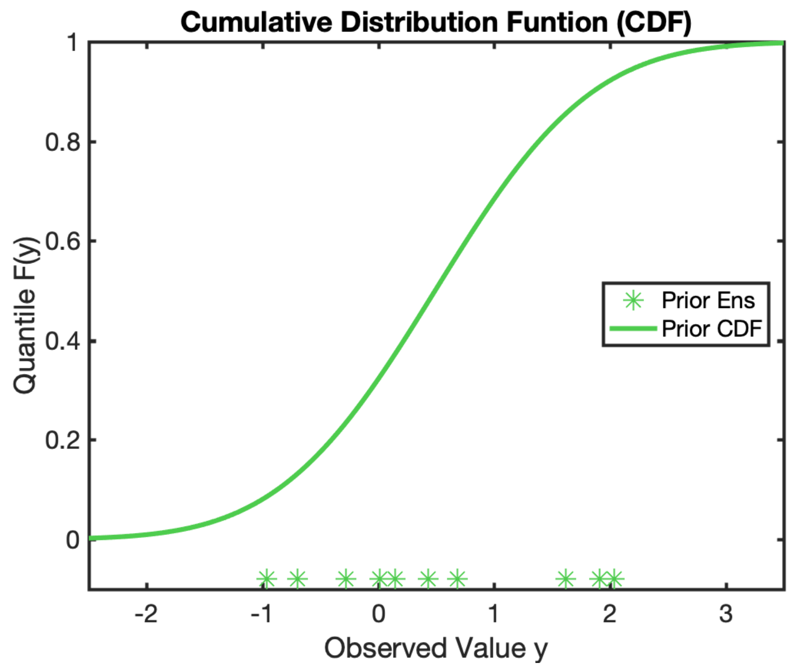
Generalized Ensemble Filter Framework using Quantiles

Given a prior ensemble estimate of an observed quantity, y



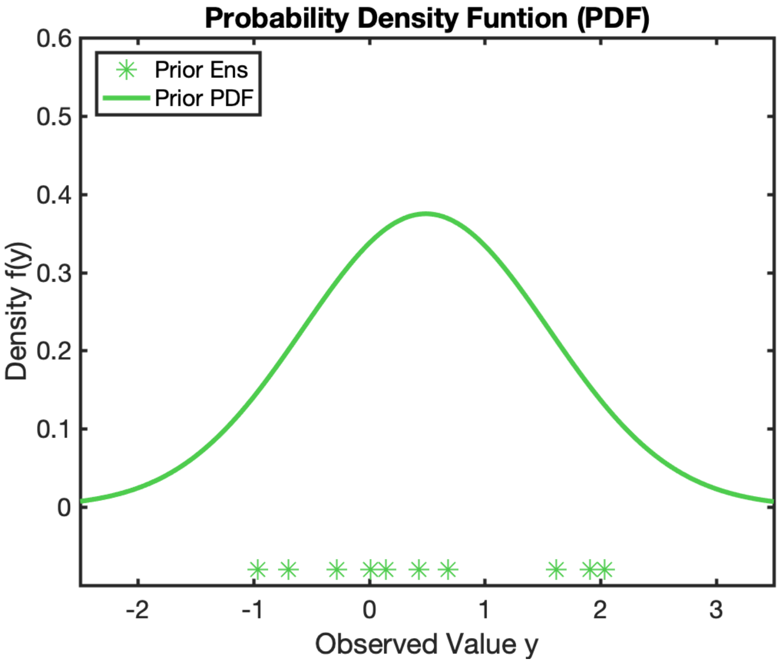
Fit a continuous PDF from an appropriate distribution family and find the corresponding CDF

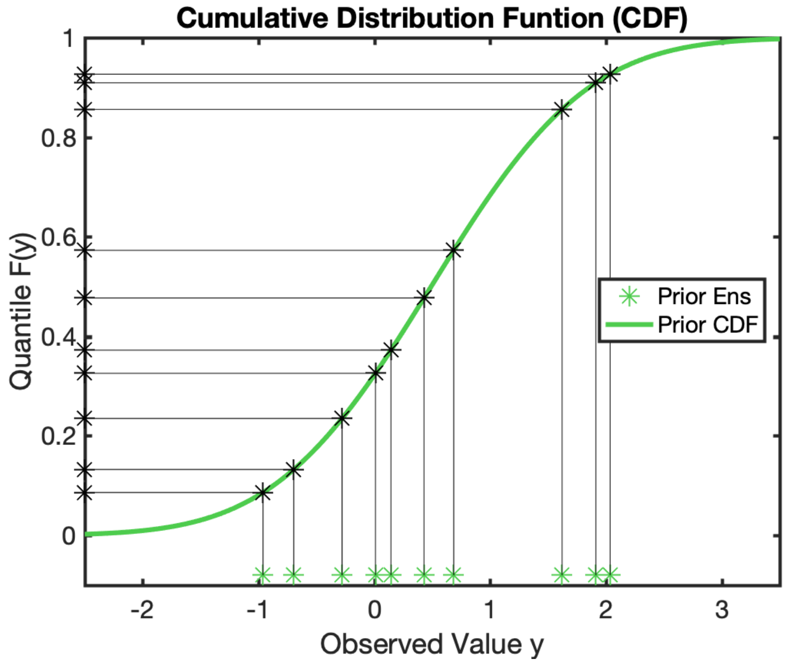




This example uses a normal PDF

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

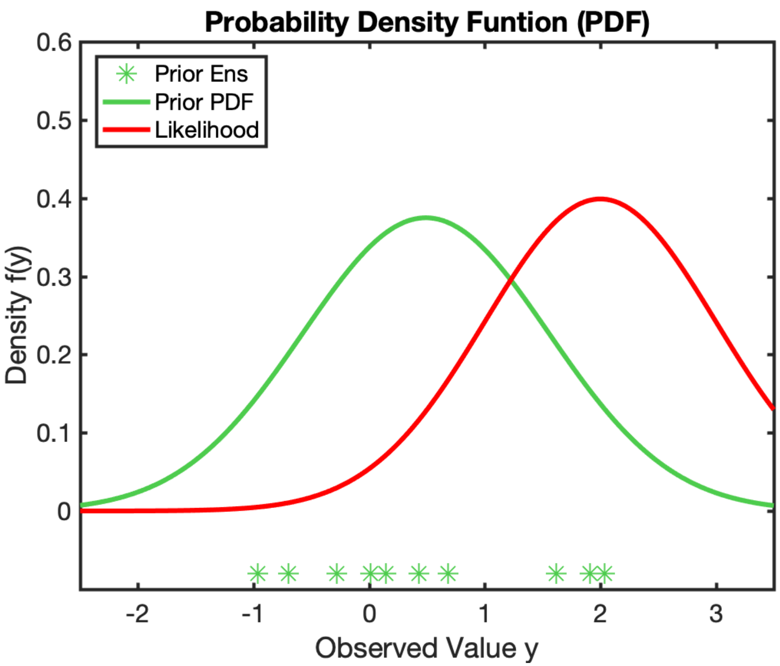


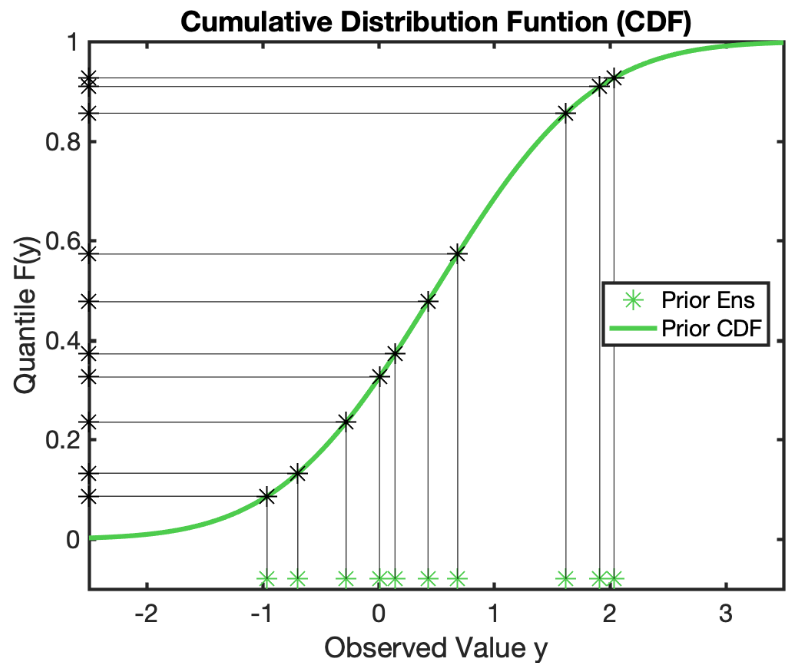


This example uses a normal PDF

Generalized Ensemble Filter Framework using Quantiles

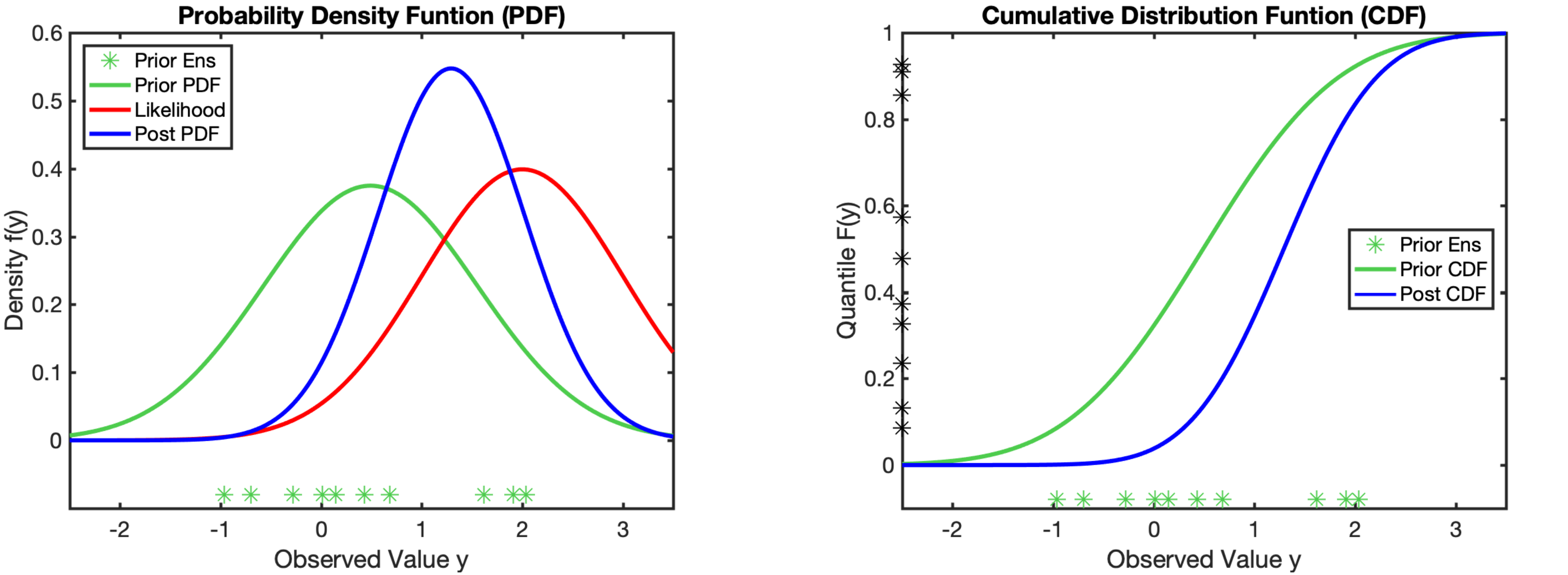
Continuous likelihood for this observation.





This example uses a normal PDF

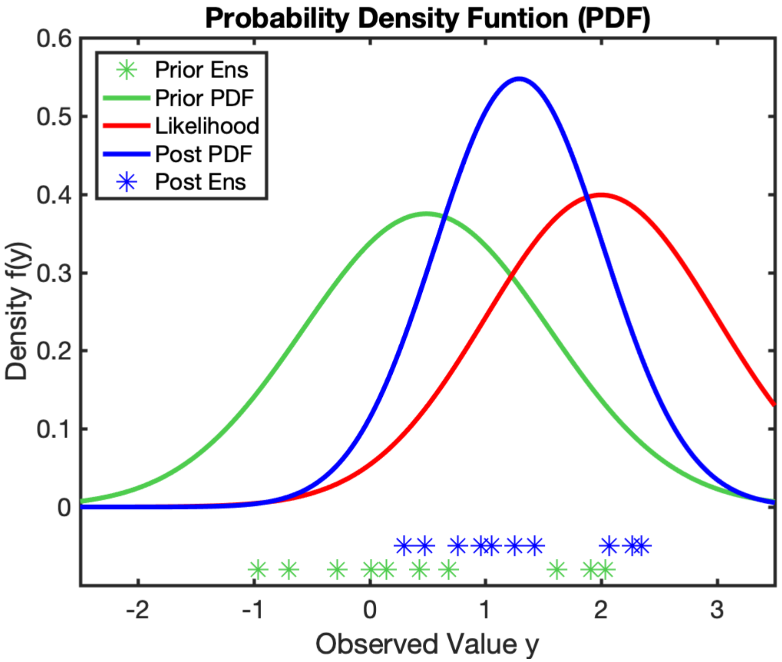
Bayes tells us that the continuous posterior PDF is the product of the continuous likelihood and prior.

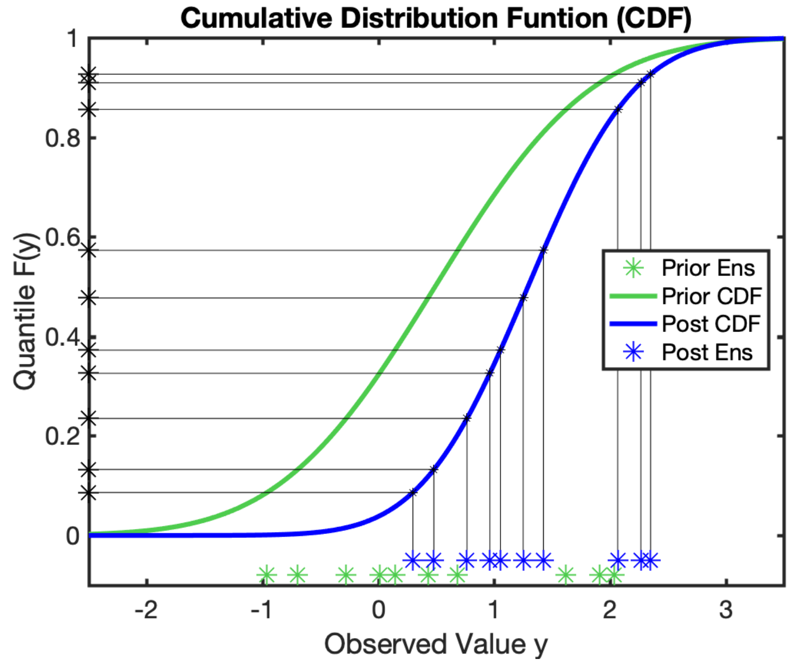


Normal times normal is normal.

Generalized Ensemble Filter Framework using Quantiles

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





This example uses a normal PDF

Useful families for continuous priors and likelihoods

Different families of distributions for continuous priors and likelihoods can lead to analytic continuous posterior.

A list of pairs of priors and likelihoods that may be useful for scientific application follows.

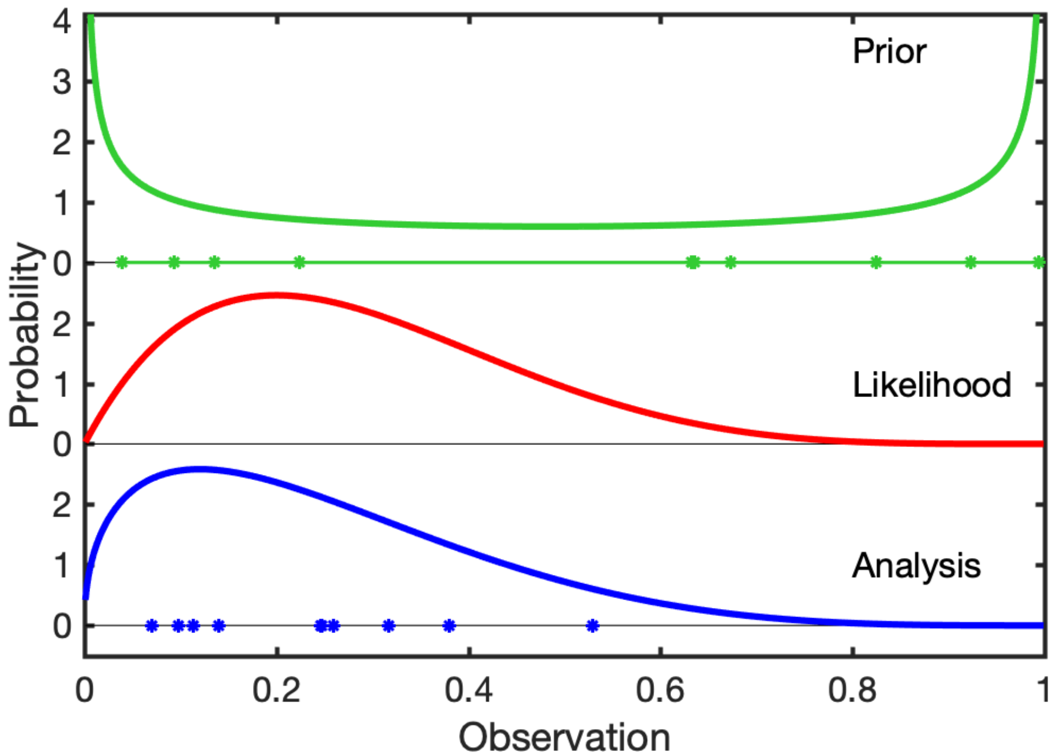
Useful families for continuous priors and likelihoods

Prior	Likelihood	Posterior
Normal	Normal	Normal
Lognormal	Lognormal	Lognormal
Gamma	Gamma	Gamma
Inverse Gamma	Inverse Gamma	Inverse Gamma
Beta	Beta	Beta
Beta prime	Beta prime	Beta prime
Exponential	Exponential	Exponential
Pareto	Pareto	Pareto
Genl. Gamma given p	Genl. Gamma given p	Genl. Gamma given p
Any	Uniform	Any
Gamma	Poisson	Gamma

Useful families for continuous priors and likelihoods (2)

Prior	Likelihood	Posterior
Delta function	Any	Delta function
Skew normal	Normal	Skew normal
Truncated normal	Normal	Truncated normal
Any	Piecewise constant	Piecewise weighted
Rank histogram	Any	Rank histogram (except tails)
Huber	Huber	Piecewise normal and exponential
Weighted sum of two normals	Normal	Weighted sum of two normals
Sum of N normals same variance	Normal	Weighted sum of N normals same variance
Jeffreys	Various	Various

Another Example: Beta Prior and Likelihood



Sea ice concentration is bounded between 0 and 1.

A beta distribution can enforce these bounds.

A beta likelihood leads to a beta posterior.

Extension to multivariate application

This talk has only discussed what happens for a quantity that is observed and forecast by the model.

The method can be extended to arbitrary model variables.

Many additional details to do that, but it works.

Summary

- DA uses Bayes' to combine forecasts with observations.
- Normal distributions have been widely used.
- Normal distributions aren't always appropriate.
- New numerical method allows us to work with any distribution.

DART is Uniquely Able to Use New Method

- Works with DART's sequential ensemble algorithms.

  Won't work in matrix-based ensemble facilities.

- No efficient comparable methods for variational DA.

- Compatible with existing DART parallel implementation.

Important for Many High-Impact DA Science Collaborations

- Estimating and Predicting Bounded Quantities
  - Atmospheric chemistry
  - Streamflow and flooding
  - Ocean biogeochemistry
  - Sea ice (Chris Riedel already pushing forward)
  - Snow and land ice
  - Land surface and biosphere
  - Source and sink estimation
    - CO_2, pollutants
    - Accidental/intentional releases
  - Model parameter estimation

- Exploring more appropriate distributions for other quantities; related to Moha's anamorphosis work

- More effective use of observations with better likelihoods

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