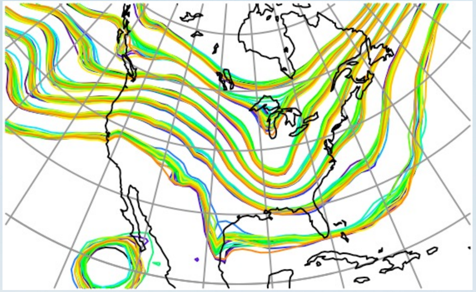
NCAR UCAR | CISL
Computational & Information Systems Lab



A Quantile Conserving Ensemble Filtering Framework for Non-Gaussian and Nonlinear Data Assimilation

Jeff Anderson representing NSF/NCAR
Data Assimilation Research Section

20th EnKF Workshop, 16 June 2025

NCAR UCAR | National Center for Atmospheric Research

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.

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.

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)

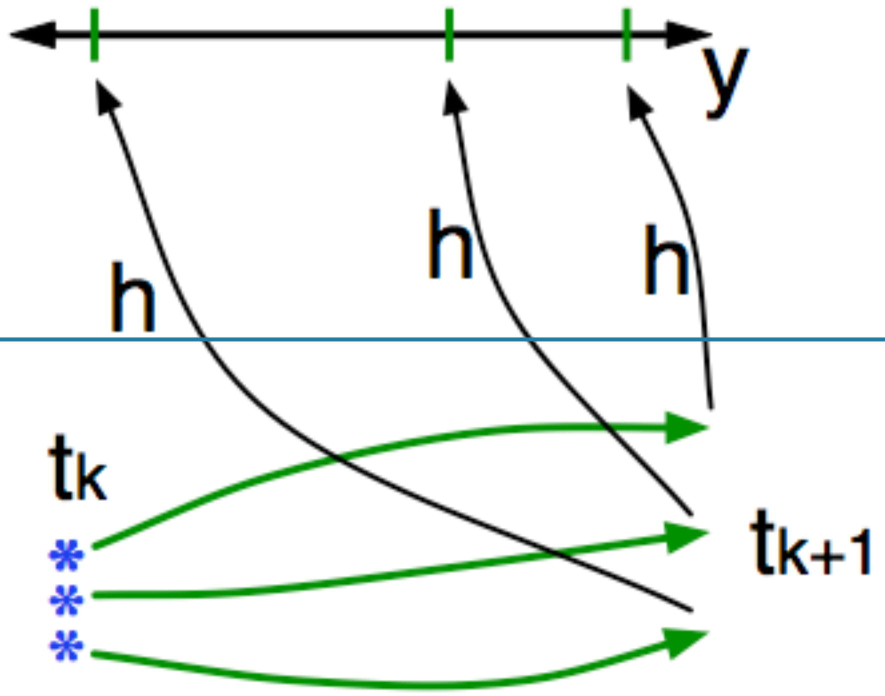
NCAR UCAR
Data Assimilation Research Testbed
U.S. National Science Foundation
WHERE DISCOVERIES BEGIN

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.

Scalar Observation Space

Model Space



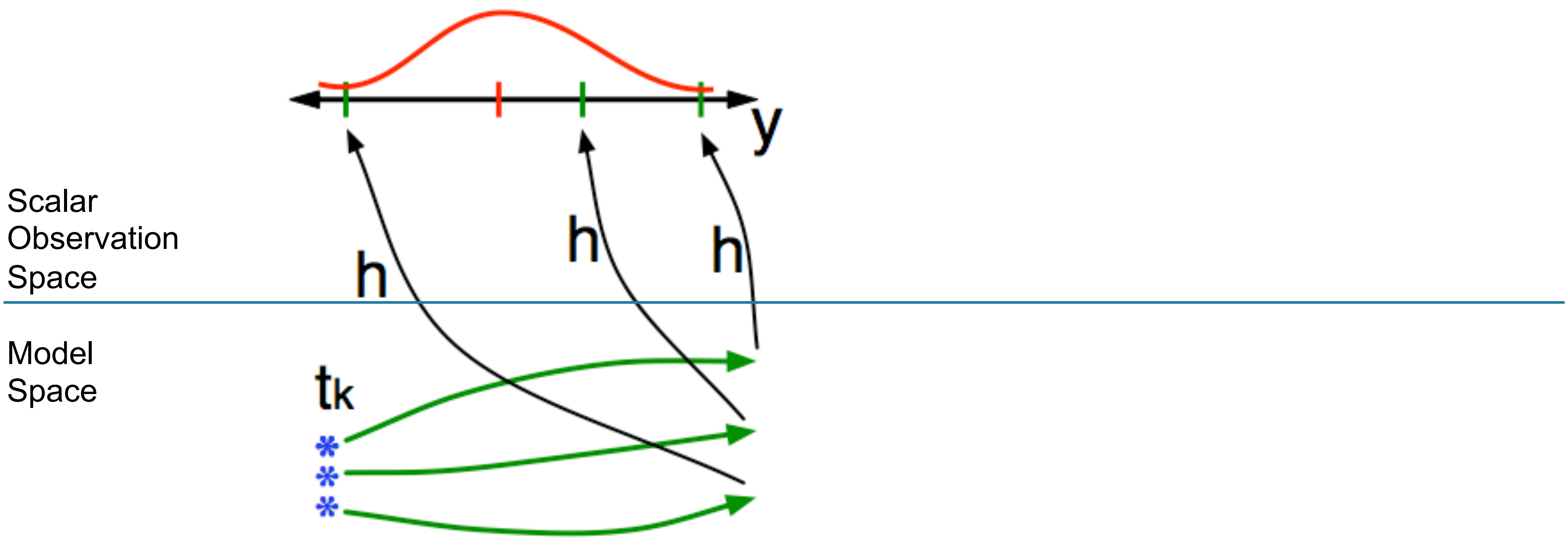
Theory: observations from instruments with uncorrelated errors can be done sequentially.

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

NCAR | UCAR
Data Assimilation Research Testbed
U.S. National Science Foundation
WHERE DISCOVERIES BEGIN

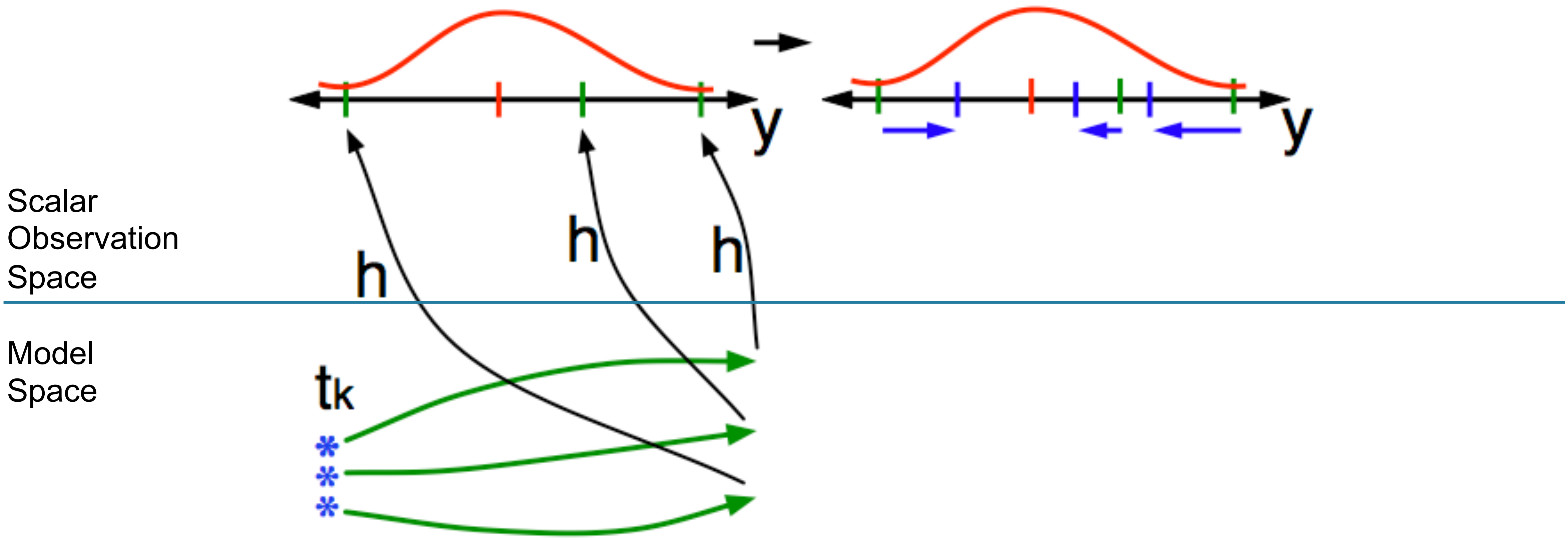
Schematic of a Sequential Ensemble Filter

3. Get observed value and observation likelihood 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).

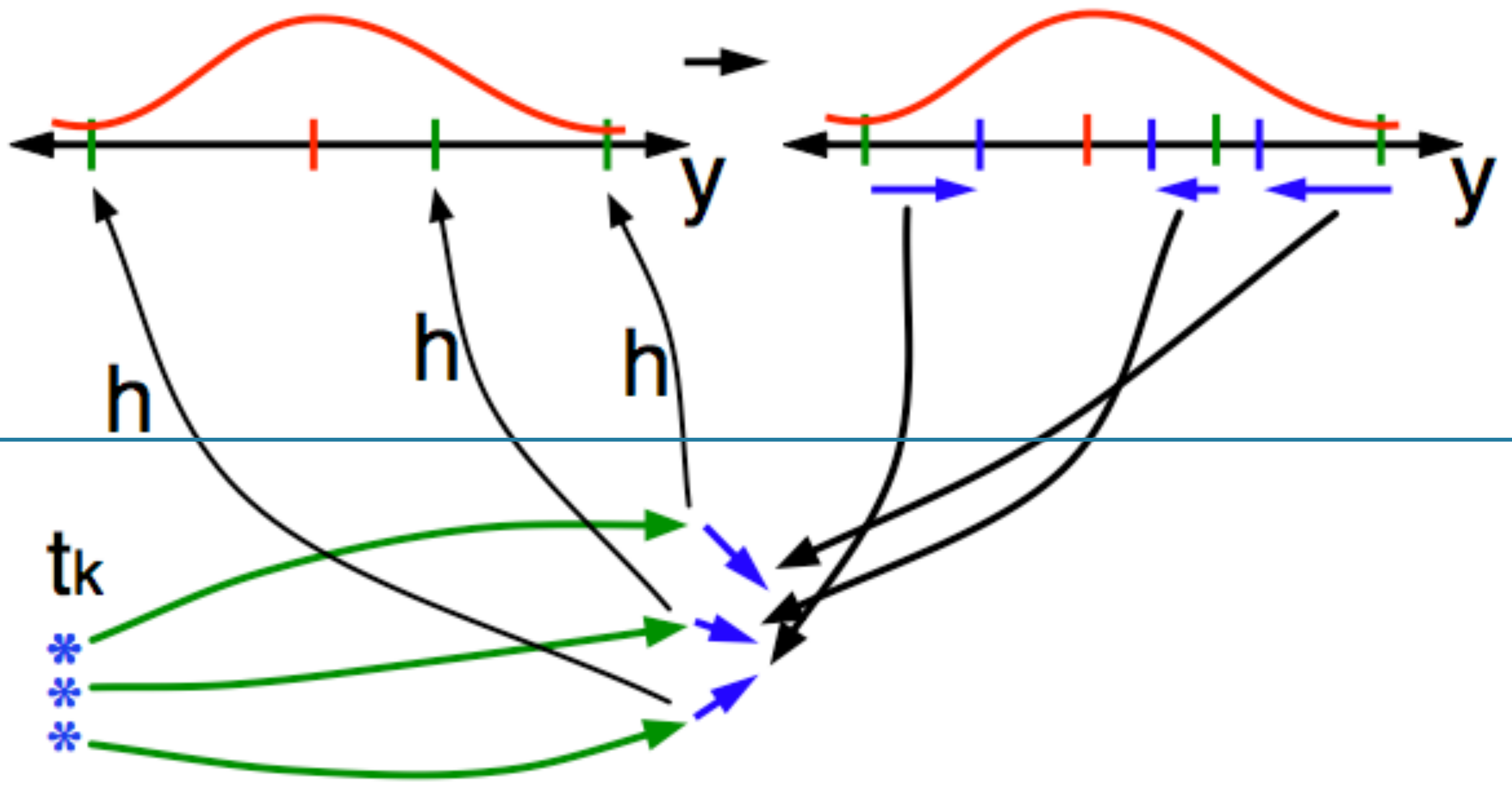


Schematic of a Sequential Ensemble Filter

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

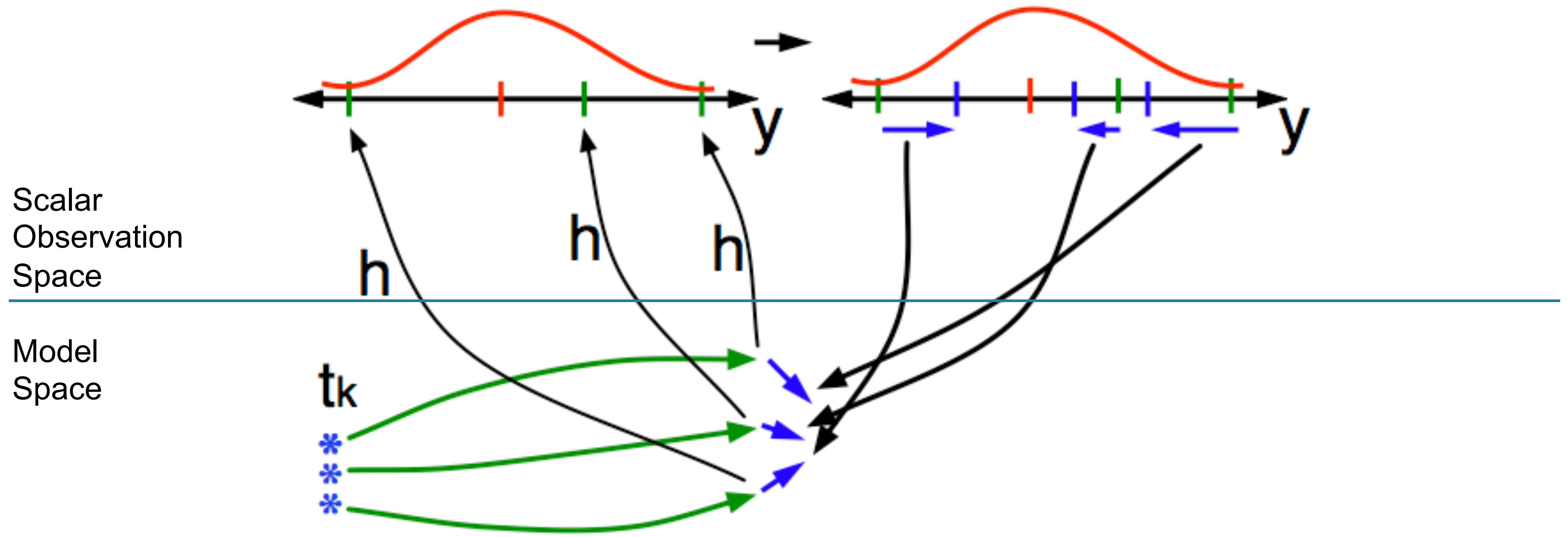
Scalar Observation Space

Model Space



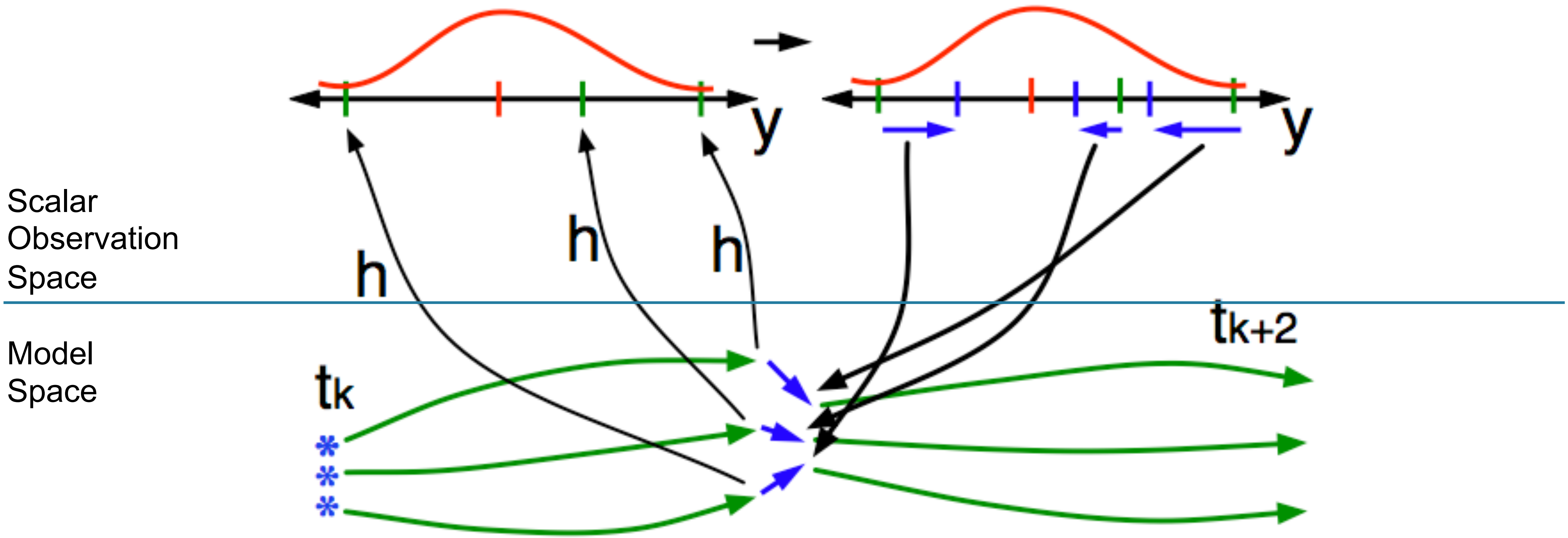
Schematic of a Sequential Ensemble Filter

Repeat sequentially for each observation available at this time.



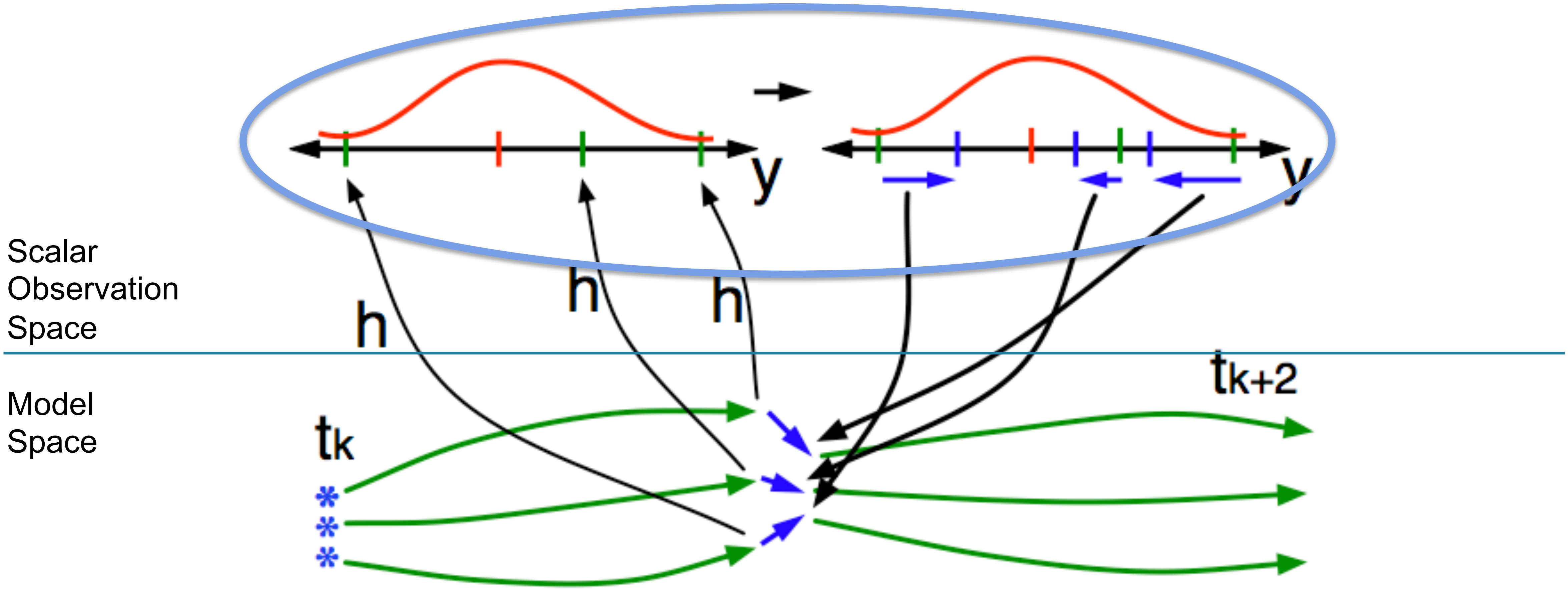
Schematic of a Sequential Ensemble Filter

When all observations have been assimilated, advance ensemble to next time with observations.



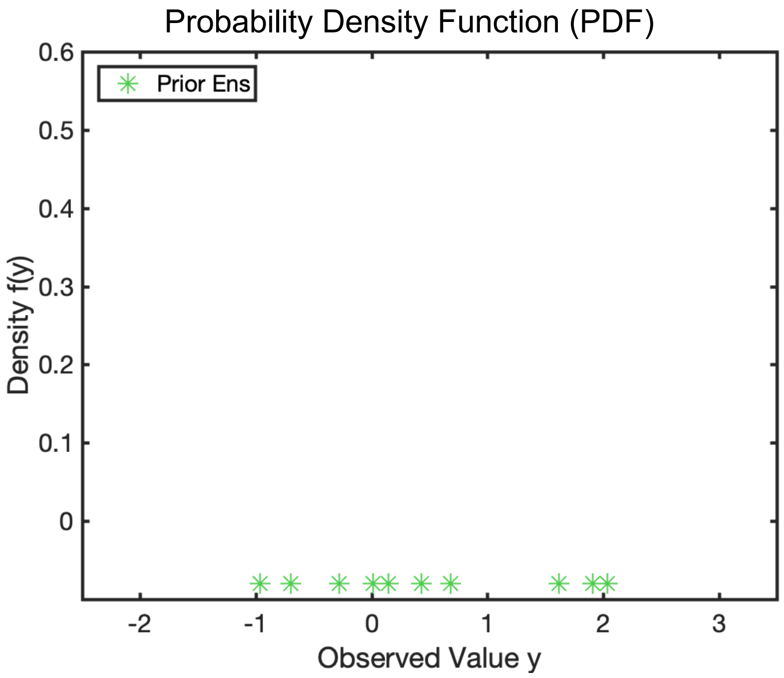
Quantile Conserving Ensemble Filters in Observation Space

A nearly general solution for the observation space step:
(Anderson, 2022, MWR *150*, 1061-1074).



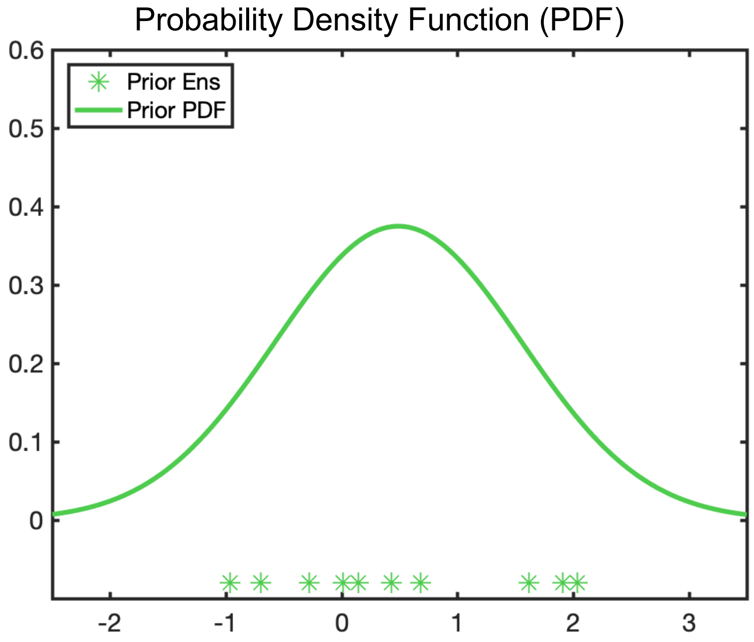
Quantile Conserving Ensemble Filter, Observation Update

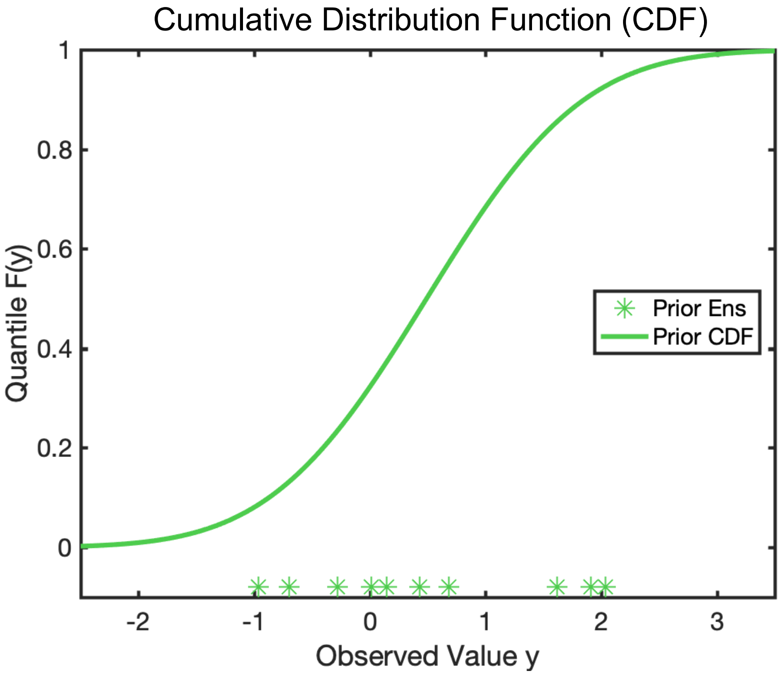
Given a prior ensemble estimate of an observed quantity, y.



Quantile Conserving Ensemble Filter, Observation Update

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

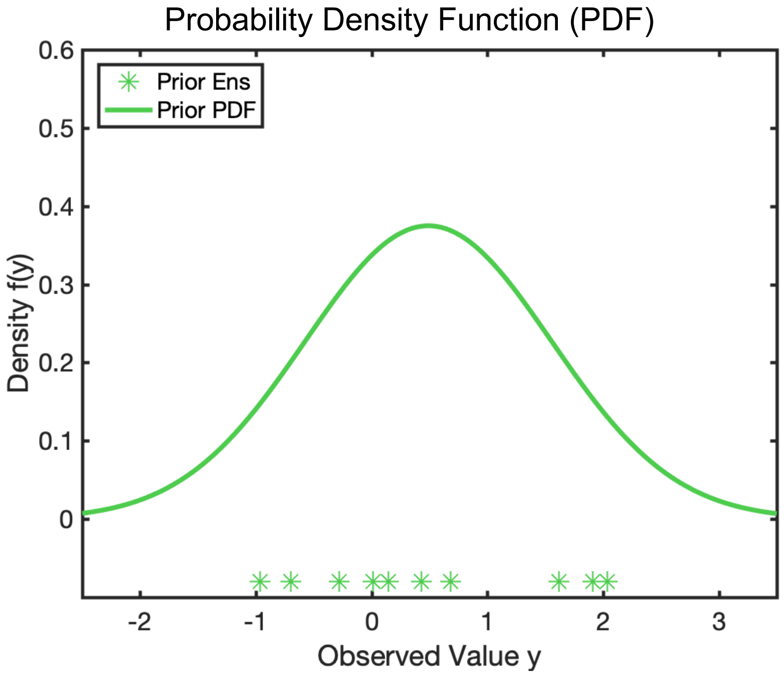


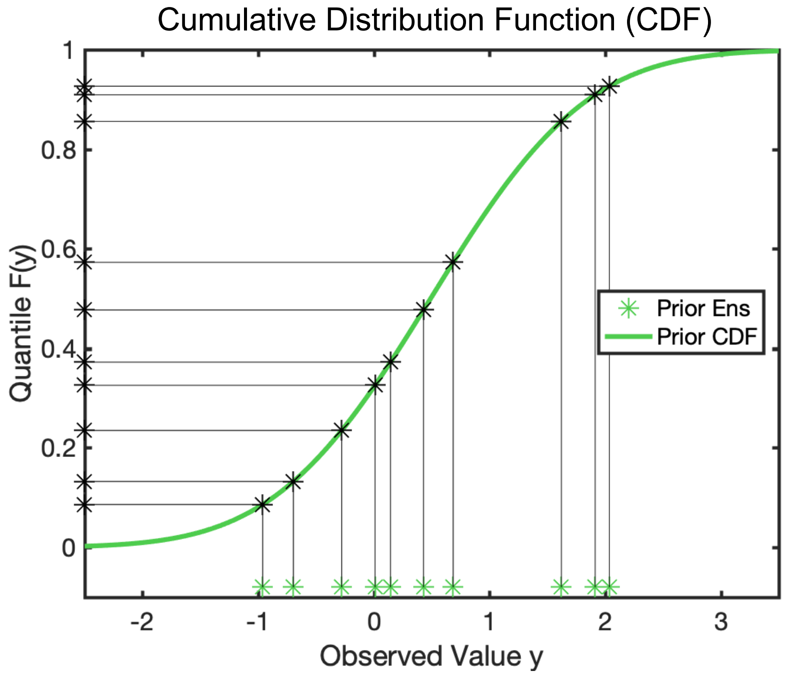


This example uses a normal PDF.

Quantile Conserving Ensemble Filter, Observation Update

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

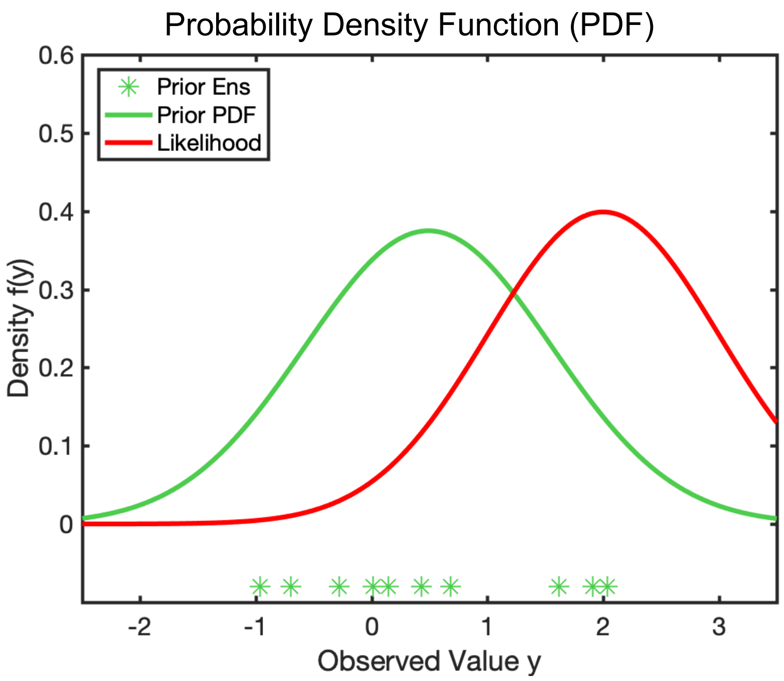


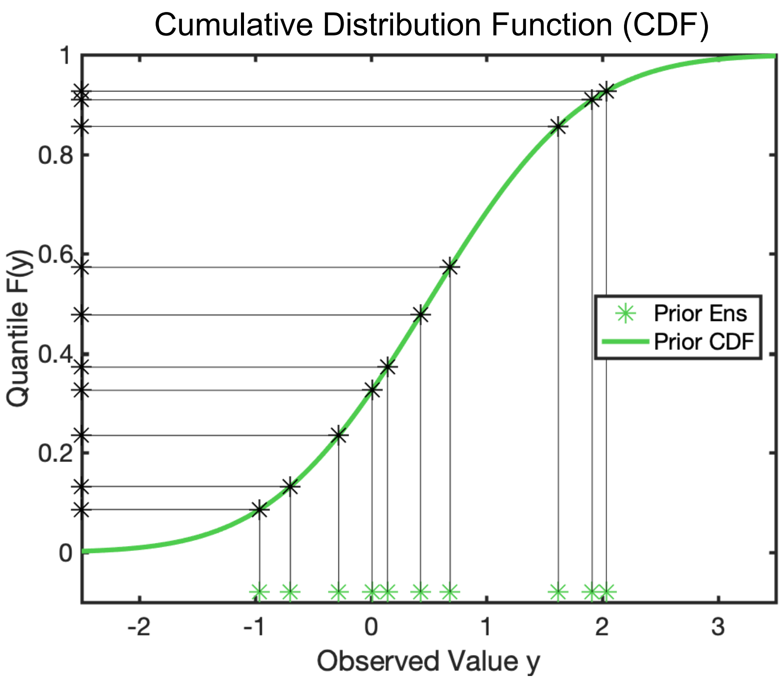


This example uses a normal PDF.

Quantile Conserving Ensemble Filter, Observation Update

Continuous likelihood for this observation.



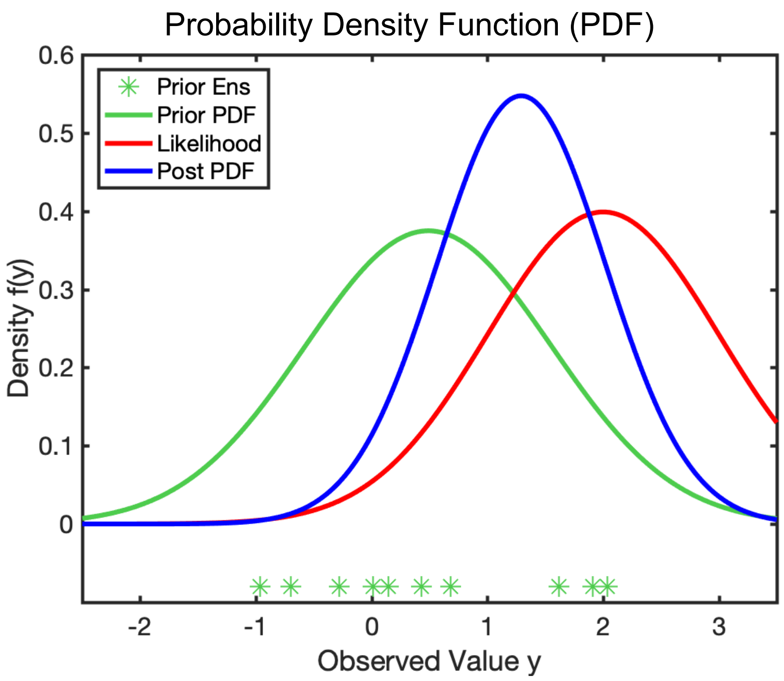


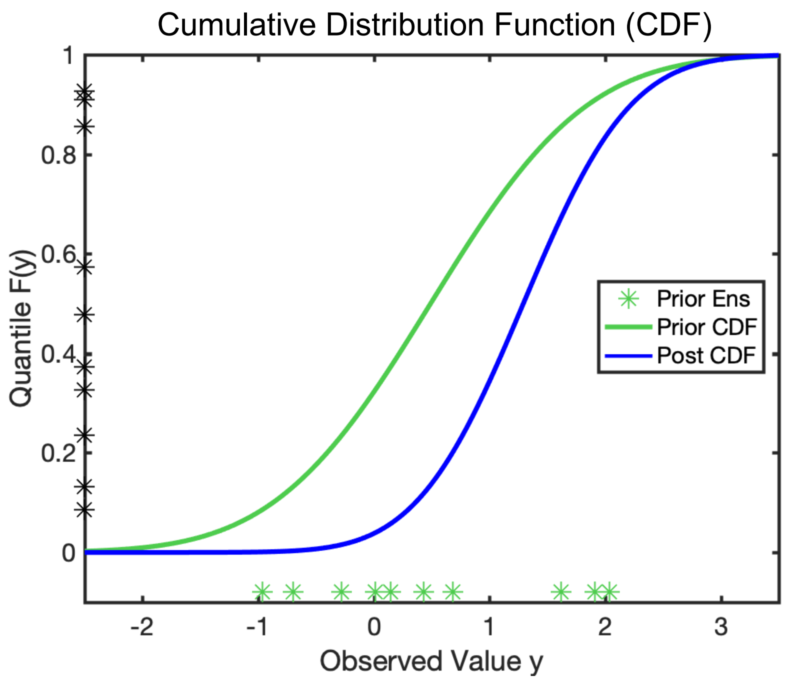
This example uses a normal PDF.

Quantile Conserving Ensemble Filter, Observation Update

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

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

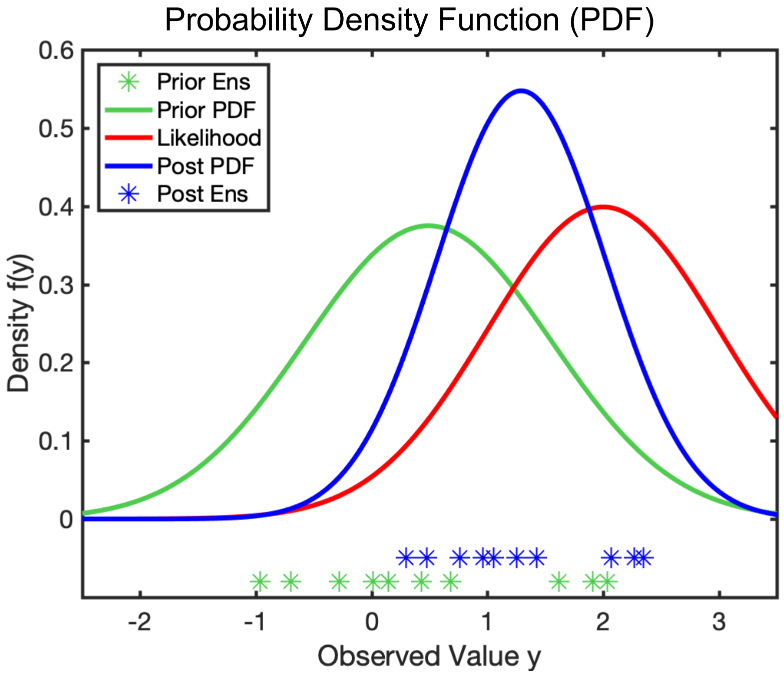


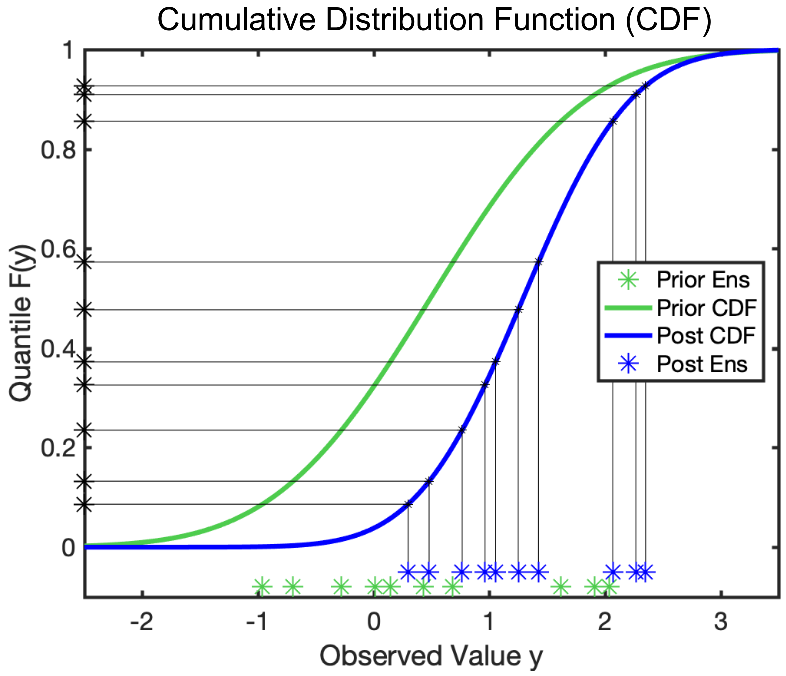


Normal times normal is normal.

Quantile Conserving Ensemble Filter, Observation Update

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

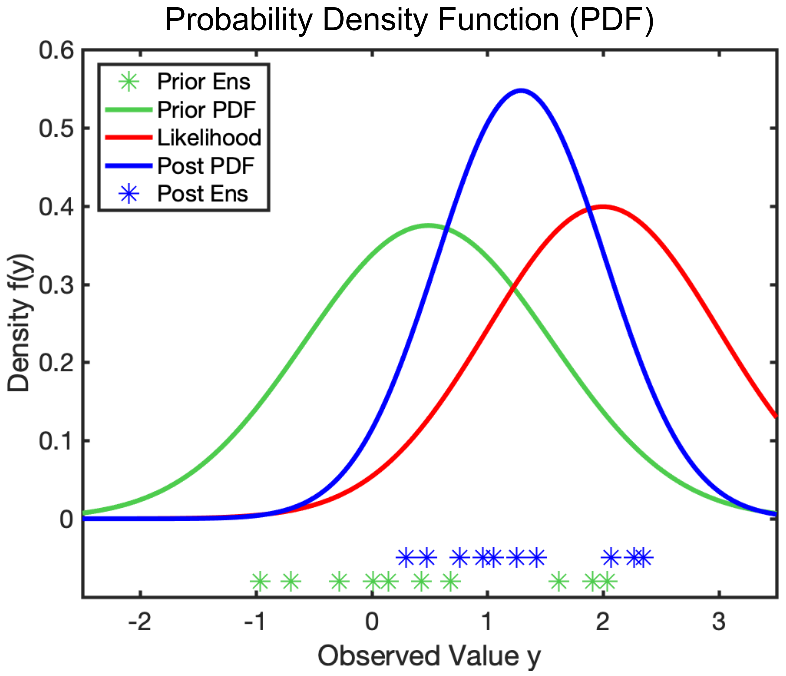


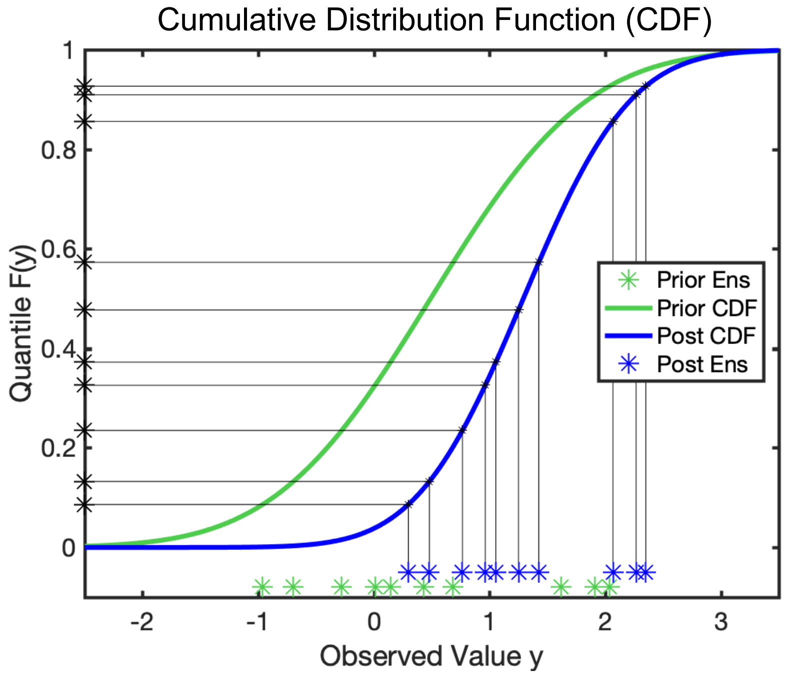


This example uses a normal PDF.

Quantile Conserving Ensemble Filter, Observation Update

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





The Bounded Normal Rank Histogram Distribution

Can use any distribution family for prior.

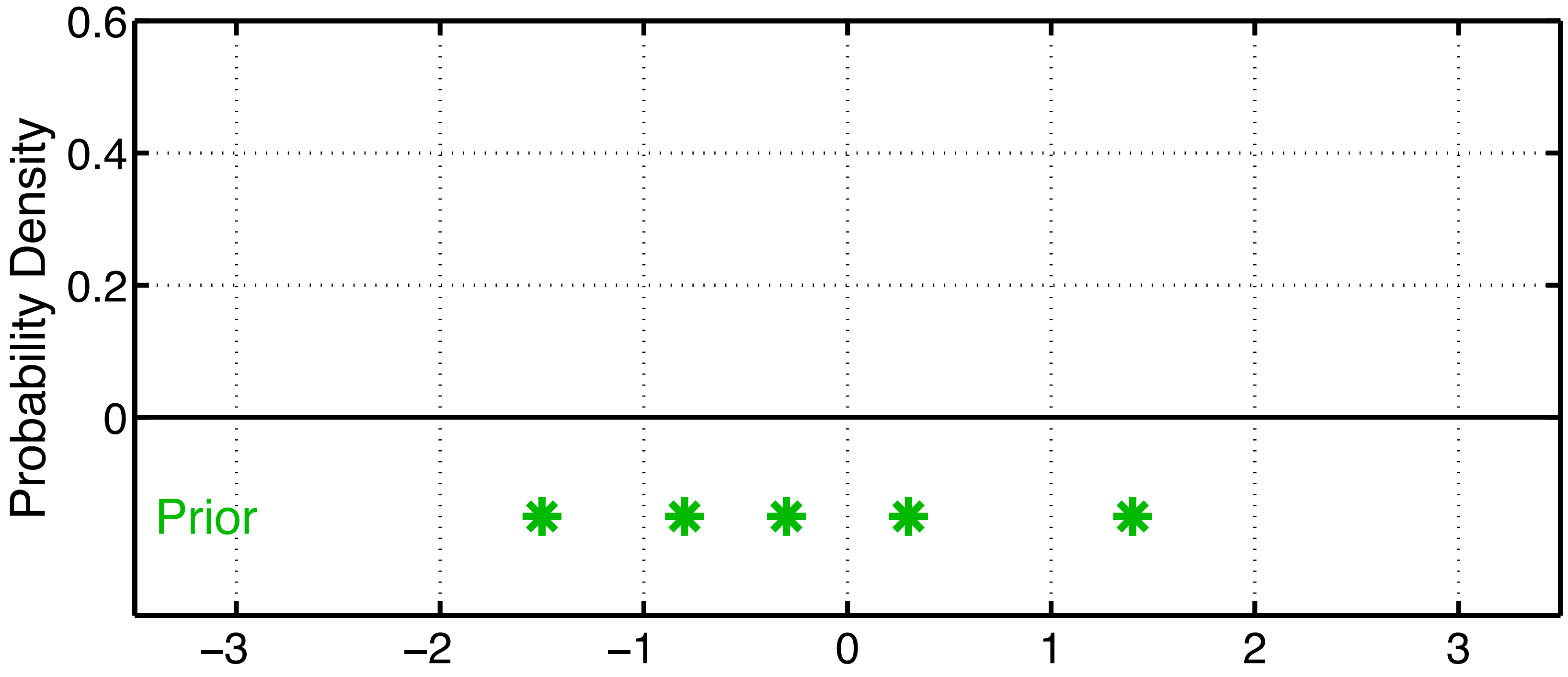
What if we don't know what family to use?

Use non-parametric continuous prior.

Rank Histogram piecewise constant distribution.

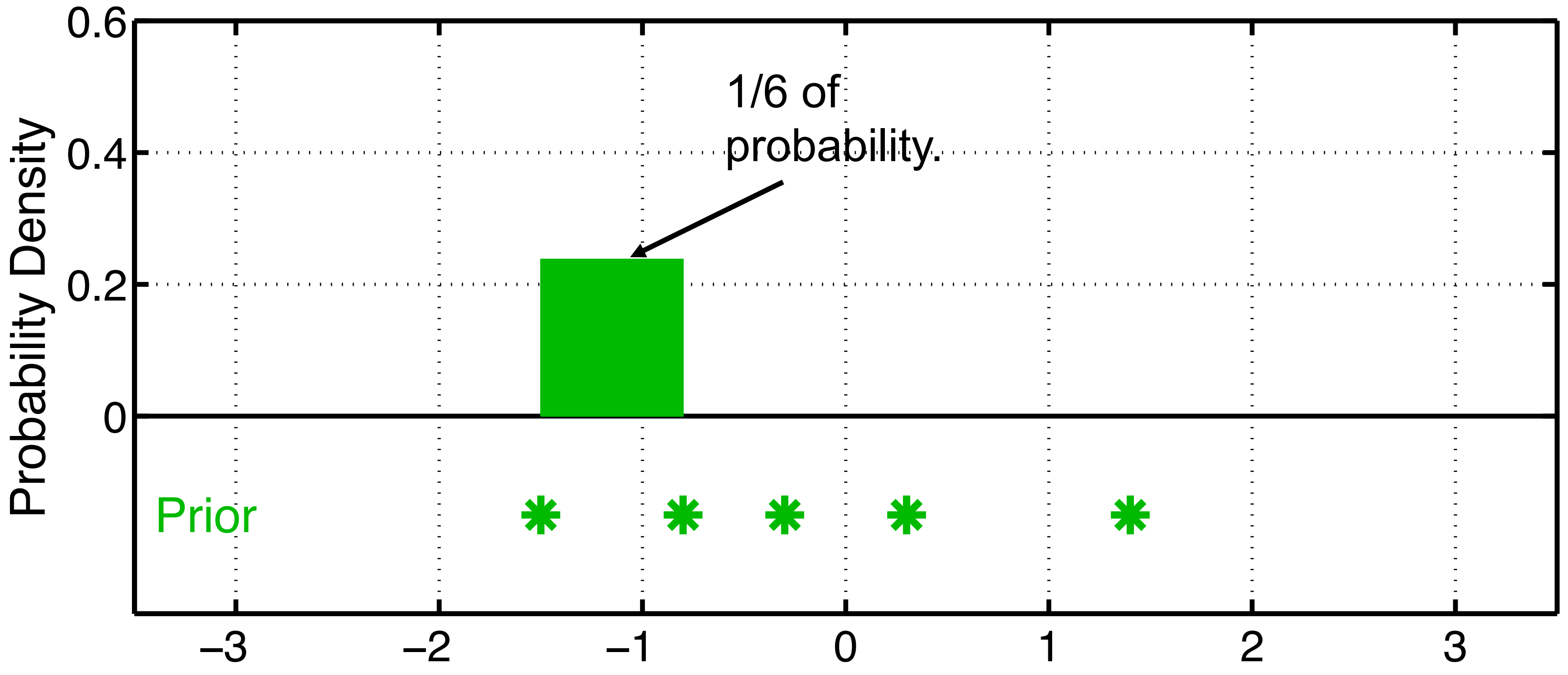
Could also use a variety of standard kernel density estimates, but these may have cost challenges.

The Bounded Normal Rank Histogram Distribution



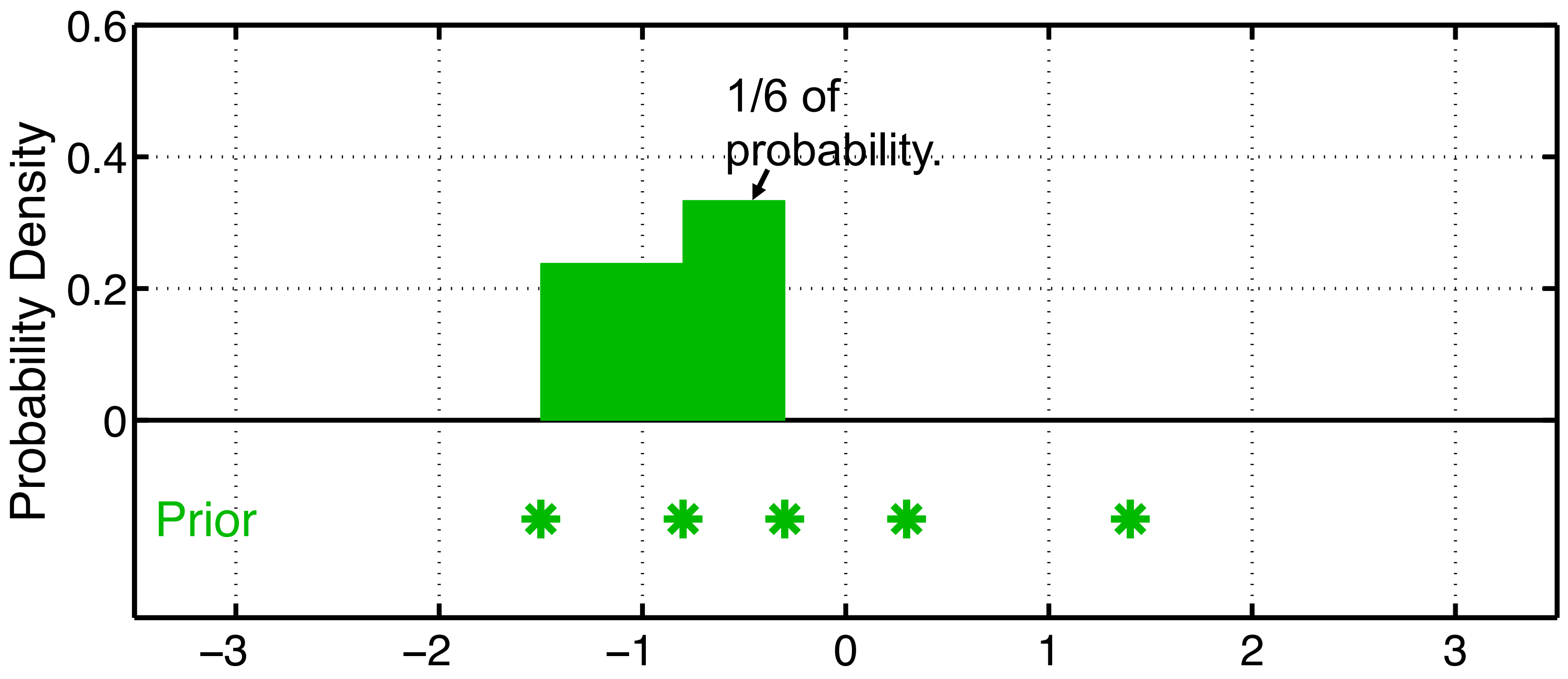
Have a prior ensemble for a state variable (like wind).

The Bounded Normal Rank Histogram Distribution



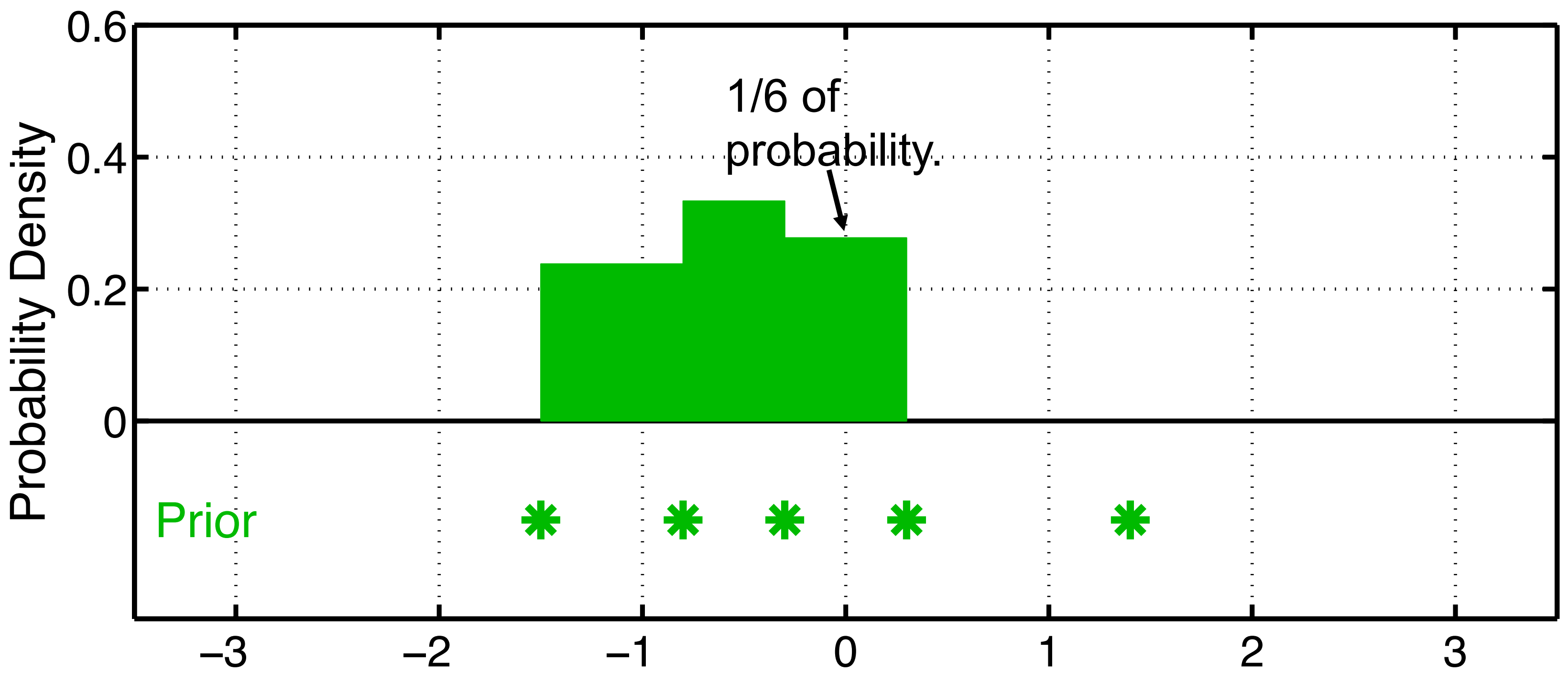
- Place (ens_size + 1)$^{-1}$ mass between adjacent ensemble members.
- Reminiscent of rank histogram evaluation method.

The Bounded Normal Rank Histogram Distribution



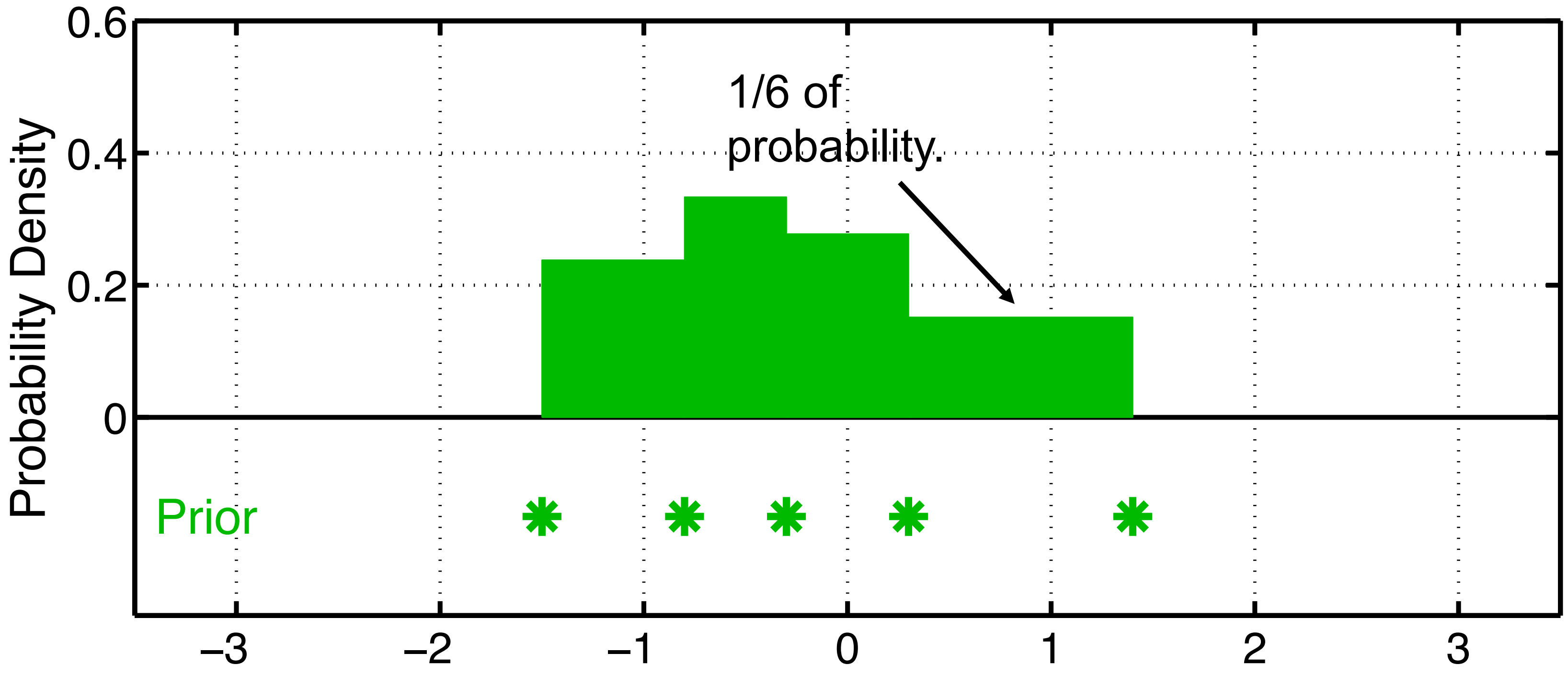
- Place $(\text{ens\_size} + 1)^{-1}$ mass between adjacent ensemble members.
- Reminiscent of rank histogram evaluation method.

The Bounded Normal Rank Histogram Distribution



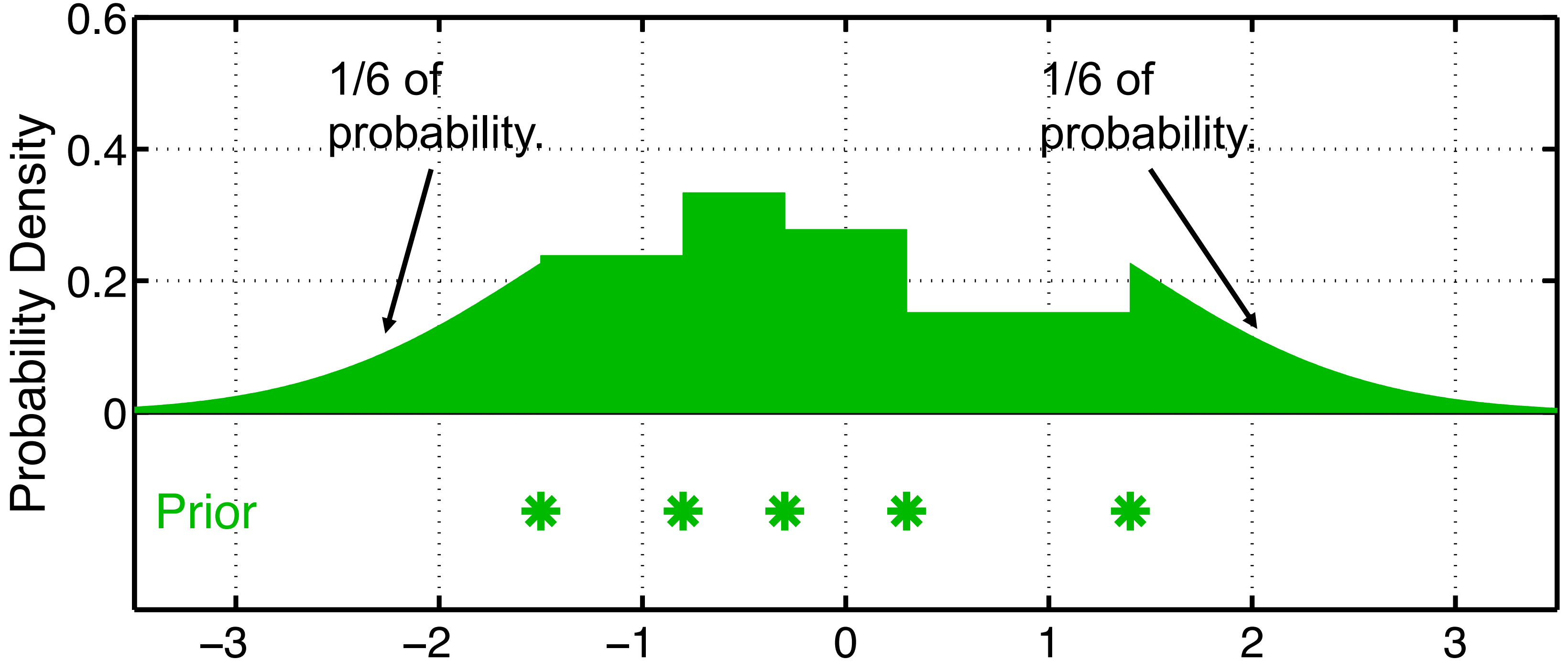
- Place (ens_size + 1)$^{-1}$ mass between adjacent ensemble members.
- Reminiscent of rank histogram evaluation method.

The Bounded Normal Rank Histogram Distribution



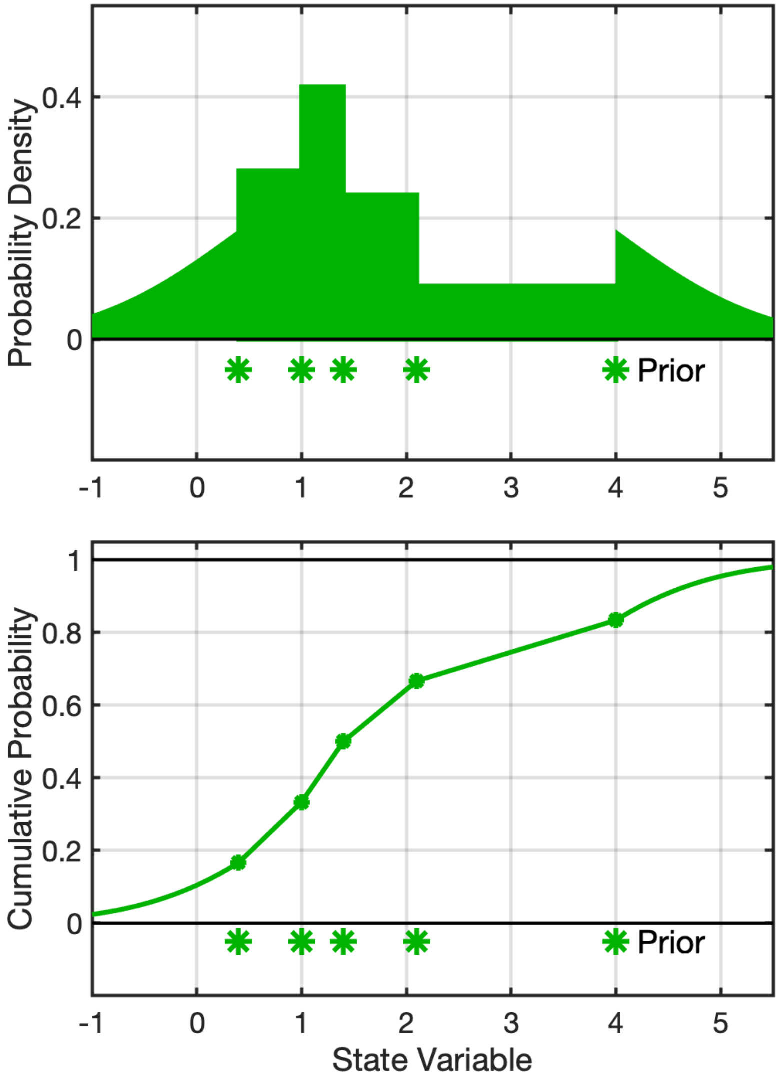
- Place (ens_size + 1)$^{-1}$ mass between adjacent ensemble members.
- Reminiscent of rank histogram evaluation method.

The Bounded Normal Rank Histogram Distribution



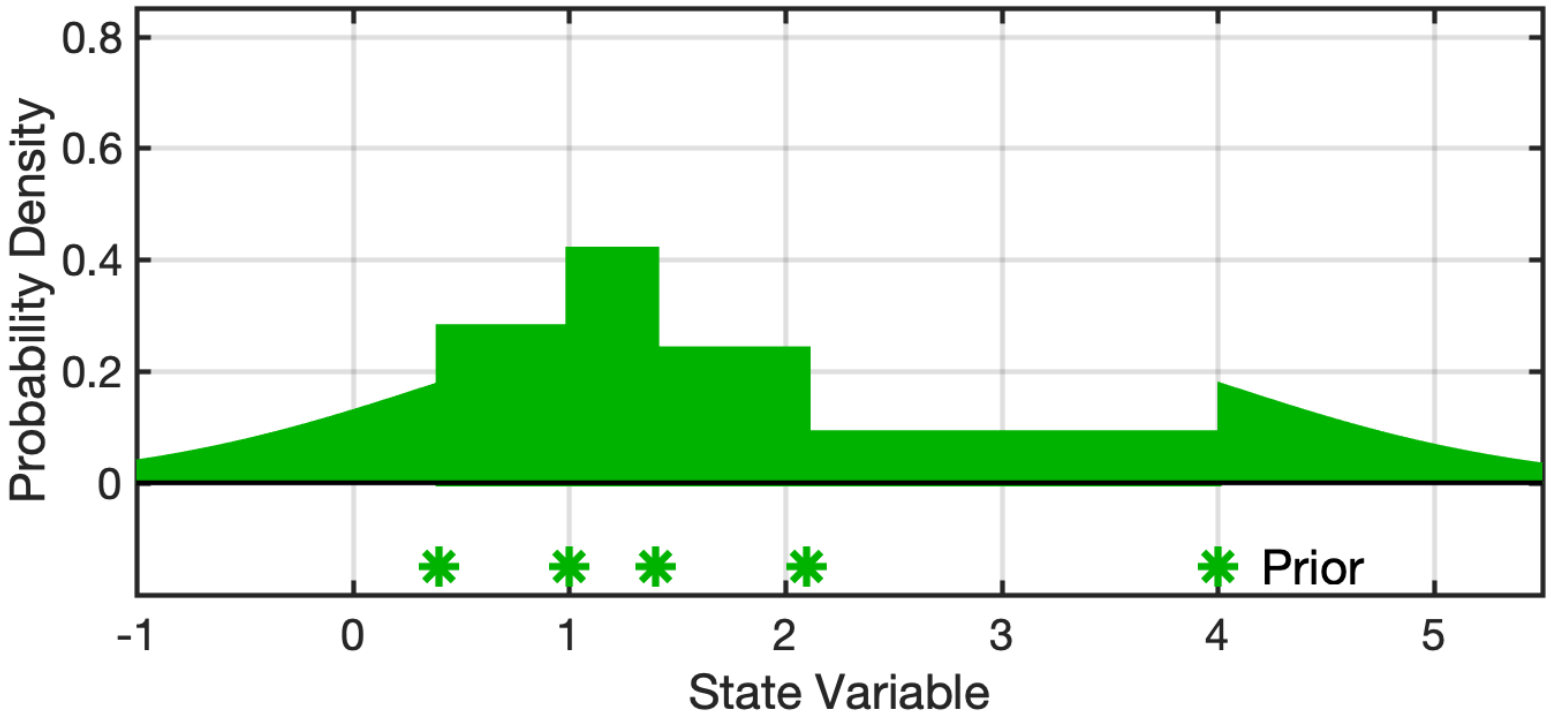
- Partial gaussian kernels on tails, Normal(*tail_mean*, *ens_sd*).
- *tail_mean selected so that (ens_size + 1)*$^{-1}$ *mass is in tail*.

Rank Histogram Prior PDF and CDF



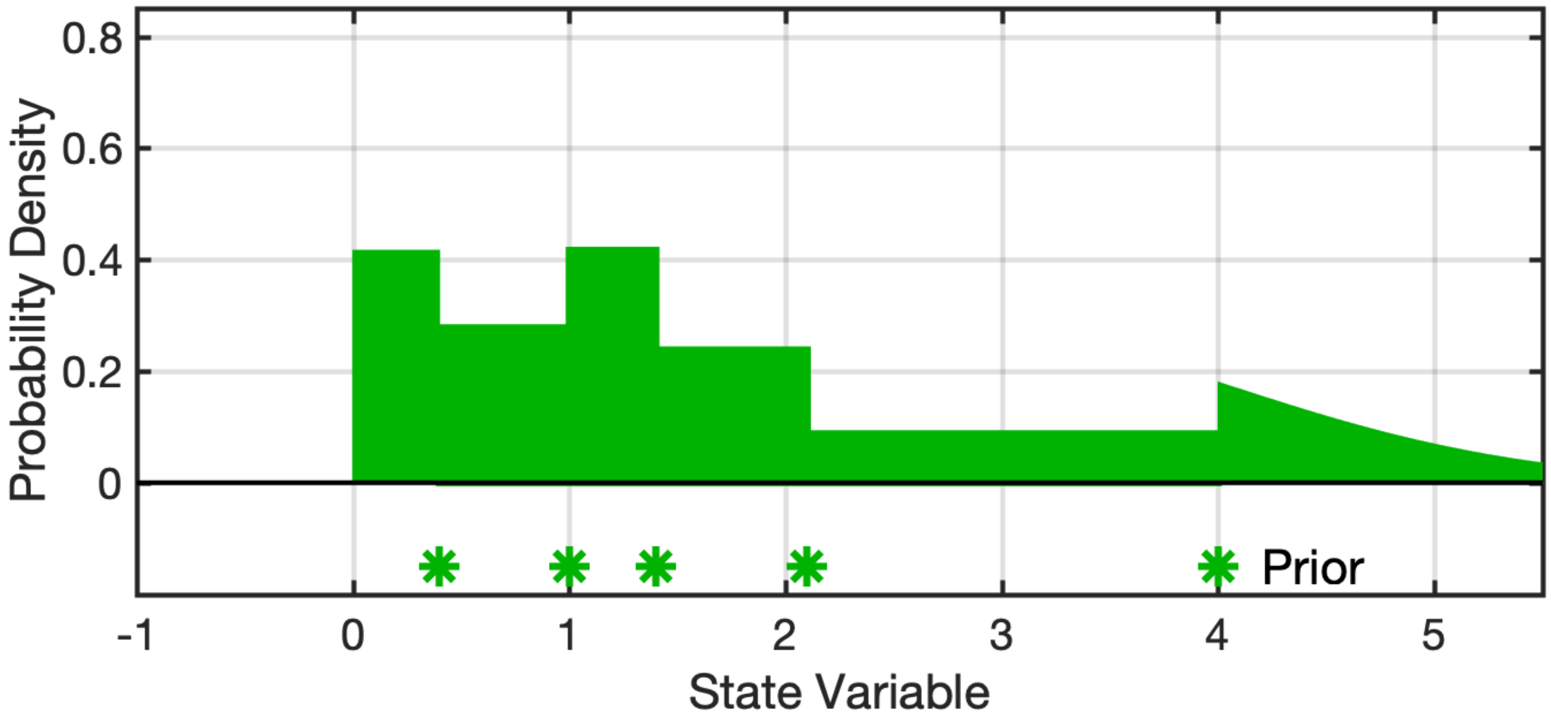
Rank Histogram Continuous Prior

Unbounded has normal tails.
Quantiles are exactly U(0, 1) by construction.

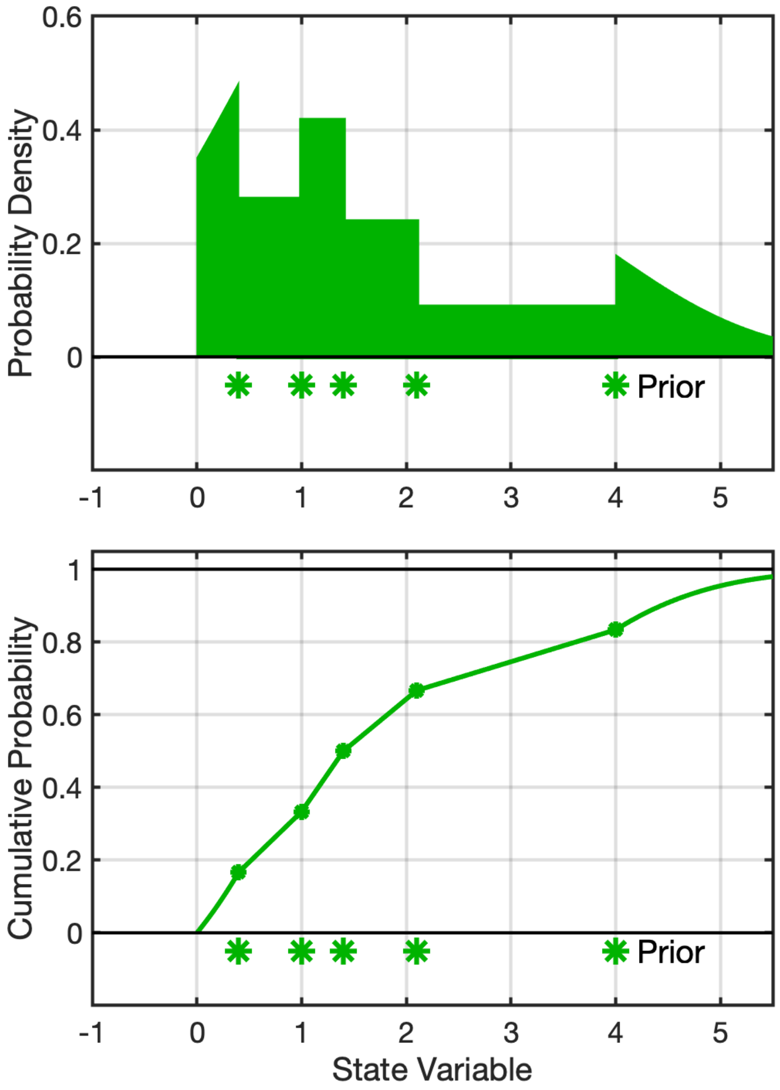


Bounded Rank Histogram Continuous Prior

Bounded has truncated tail.
Quantiles are exactly U(0, 1) by construction.

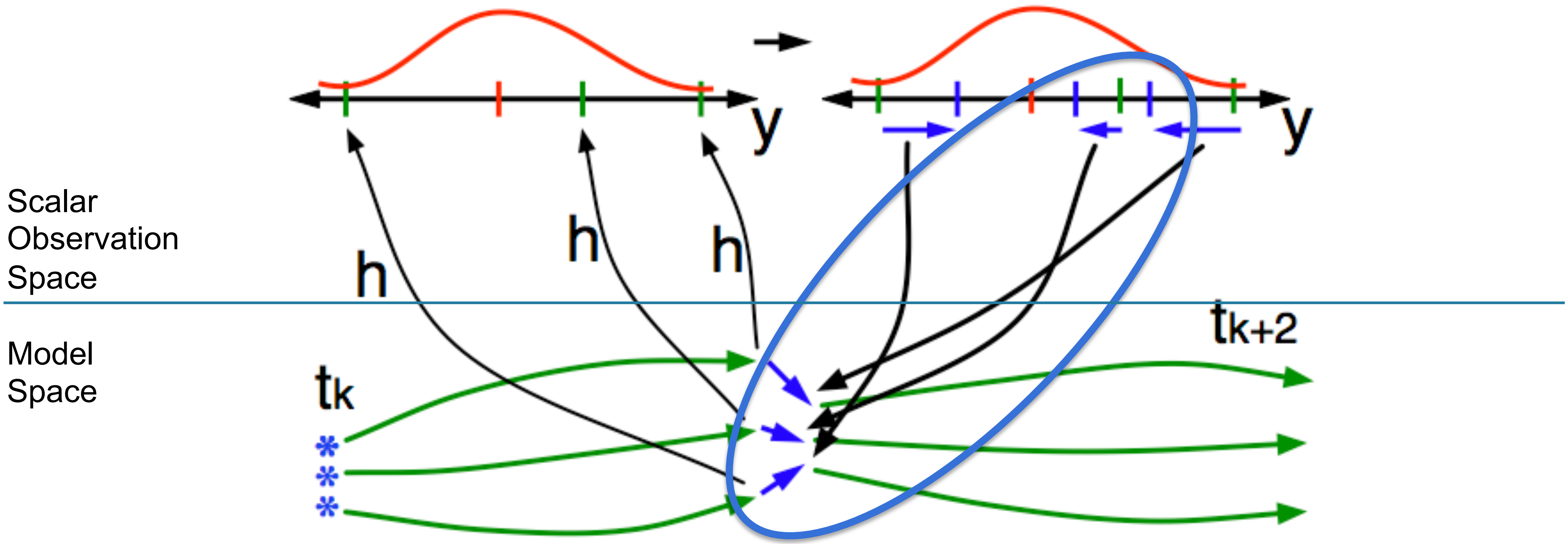


Rank Histogram Prior PDF and CDF with Lower Bound



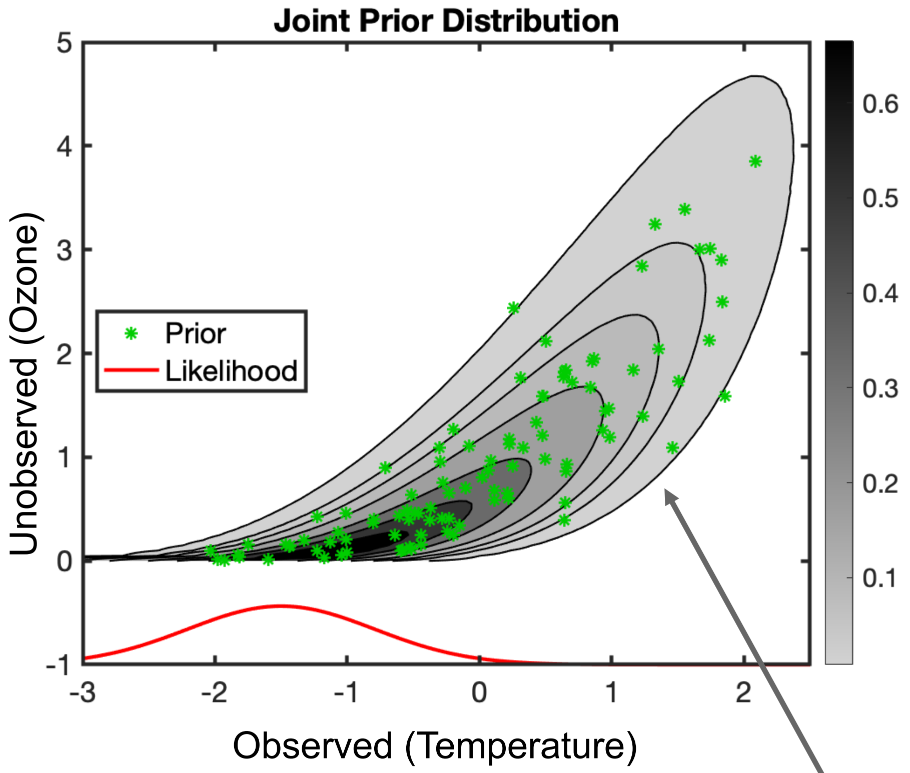
Linear Regression can Wreck Things

Linear regression can destroy benefits of new observation method.



Standard EnKF: Challenged by Non-Gaussian and Nonlinear Relations

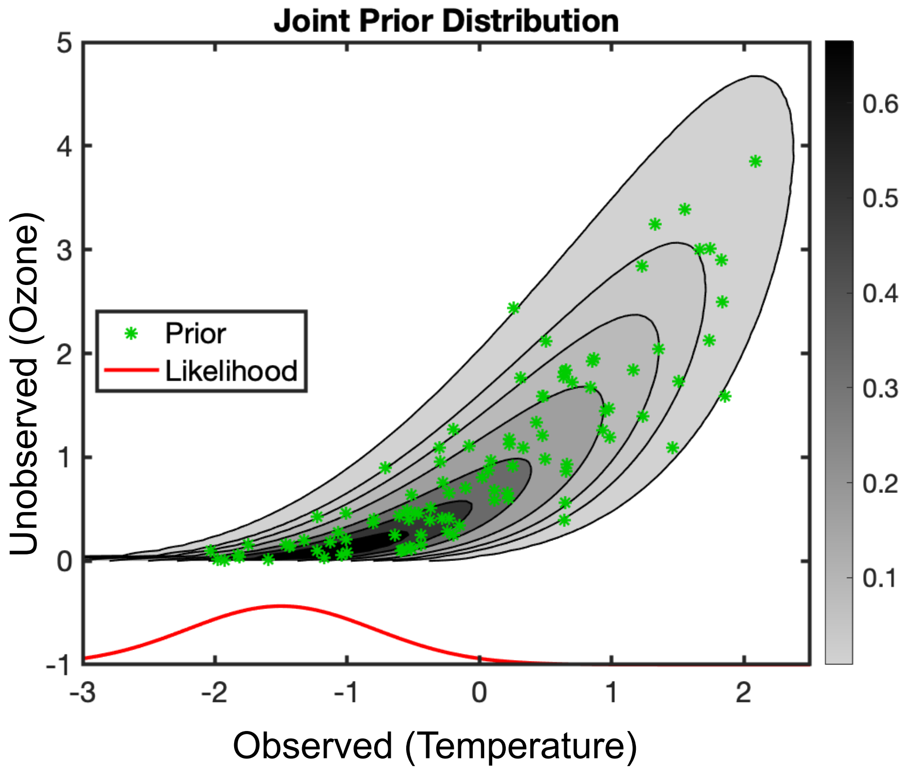
Prior for normal-gamma distribution with 100 member ensemble.



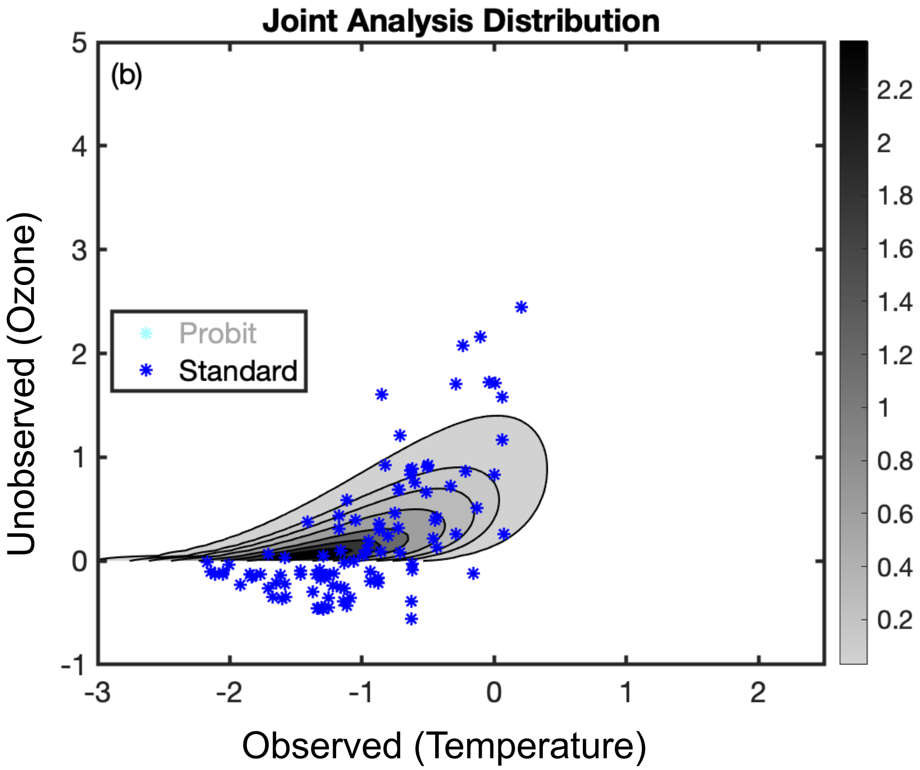
Correct distribution contours are 1, 5, 10, 20, 40, 60, 80% of max for all figures.

Standard EnKF: Challenged by Non-Gaussian and Nonlinear Relations

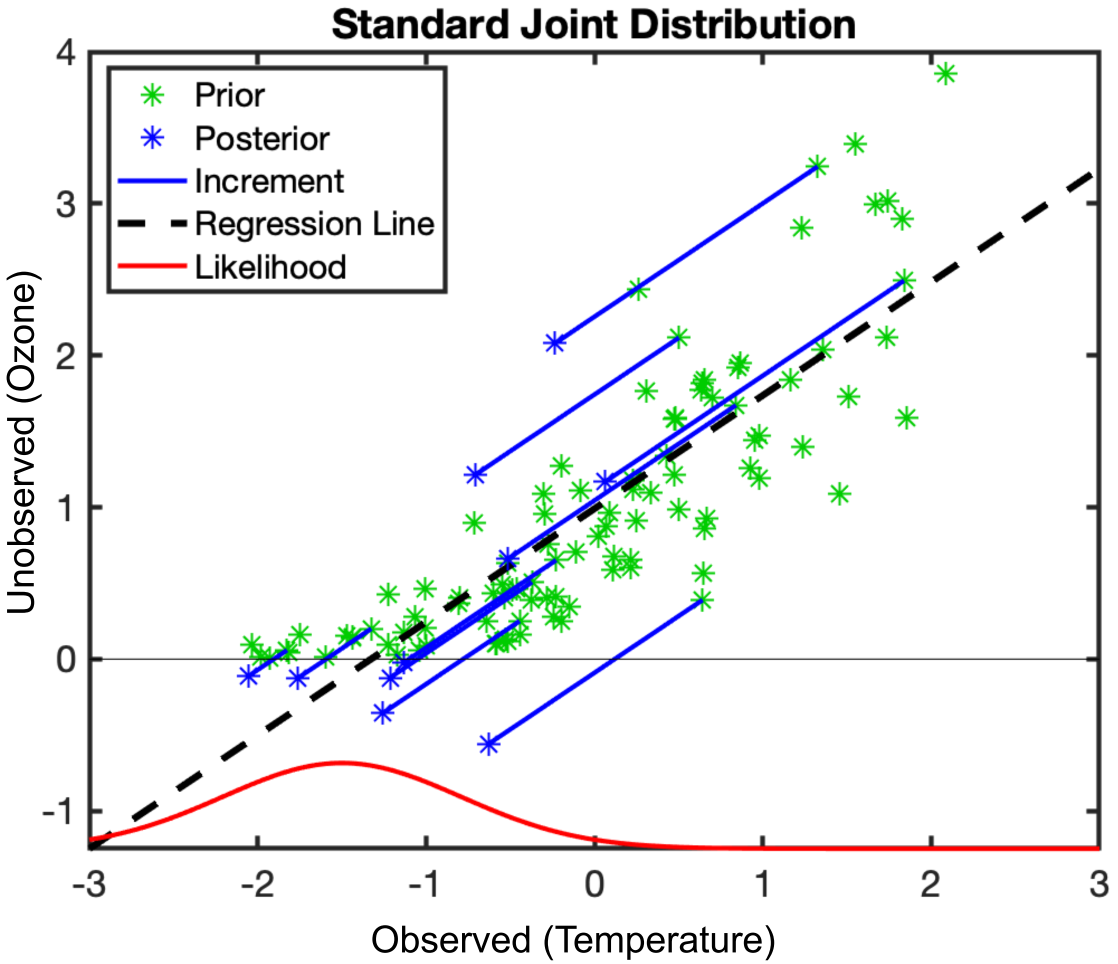
Prior for normal-gamma distribution with 100 member ensemble.



Posterior ensemble has problems.



Standard EnKF: Challenged by Non-Gaussian and Nonlinear Relations



Example regression increment vectors:
Don’t respect bounds,
Struggle with nonlinearity.

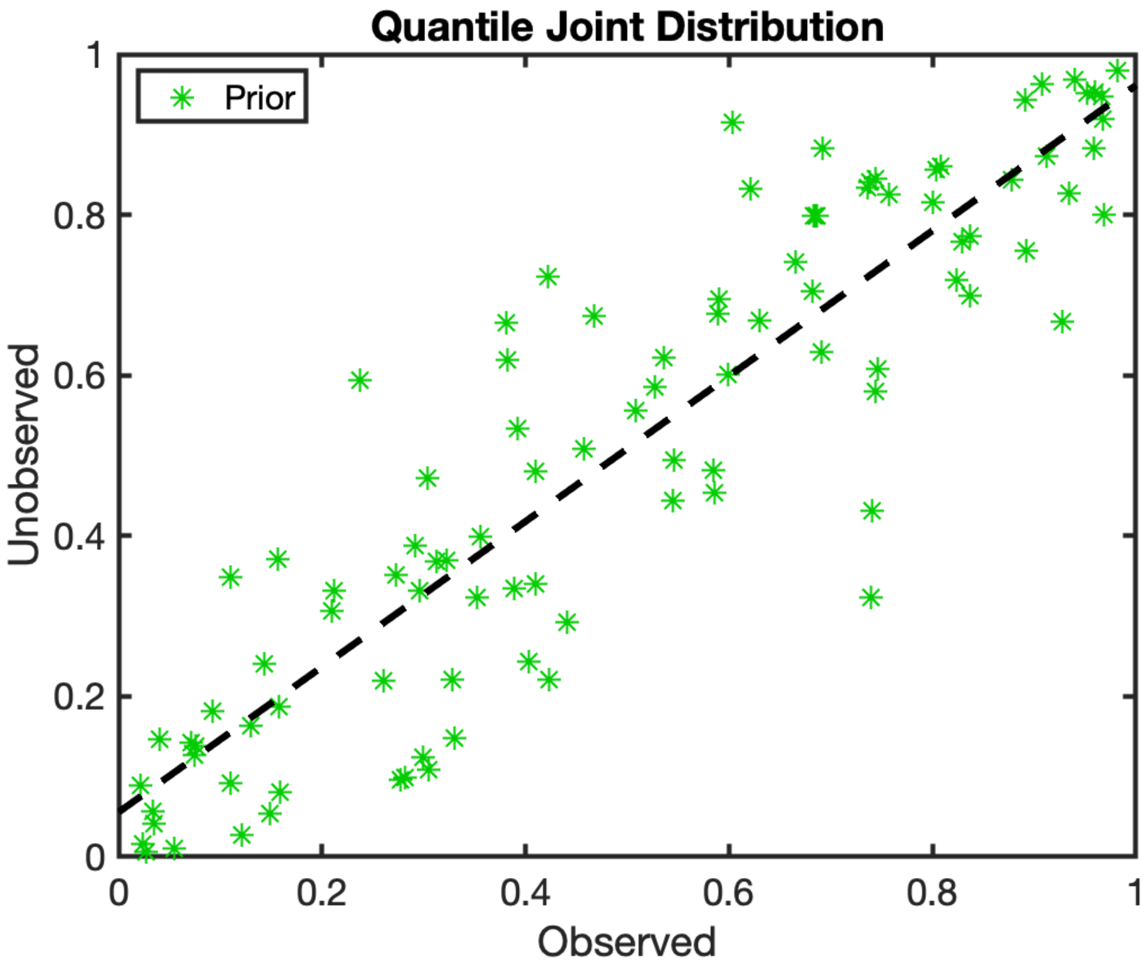
Solution, Transform Marginals: Step 1: Compute Quantiles

Pick an appropriate continuous prior distribution.

Compute CDF for each member to get quantiles.

Quantiles are U(0, 1) for appropriate prior.

This is the probability integral transform.



Solution, Transform Marginals: Step 2: Probit Transform of Quantiles

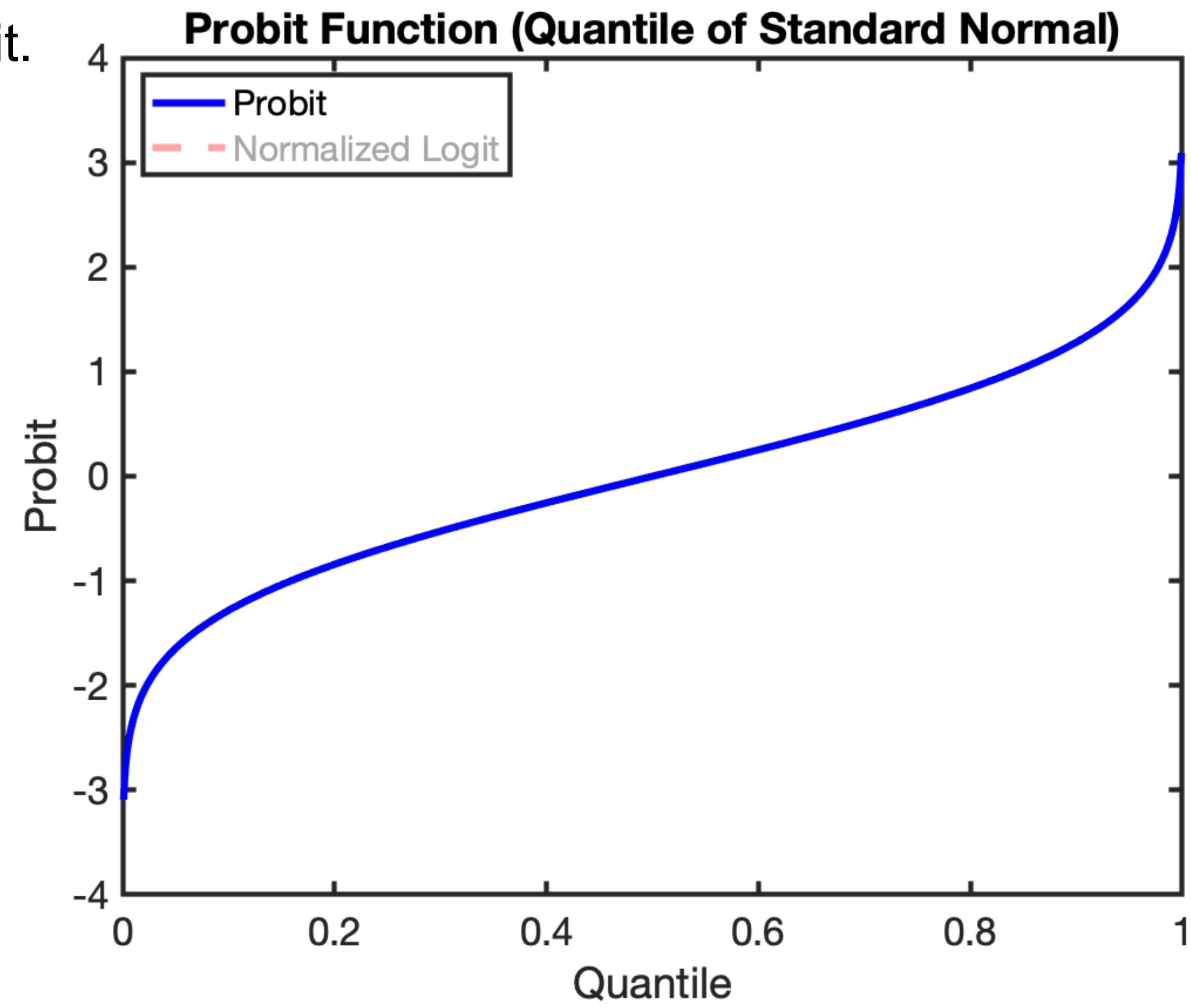
Quantile function for the standard Normal is probit.

Transforms U(0, 1) to unbounded.

Marginal distributions should be N(0, 1).

This is type of Gaussian anamorphosis.

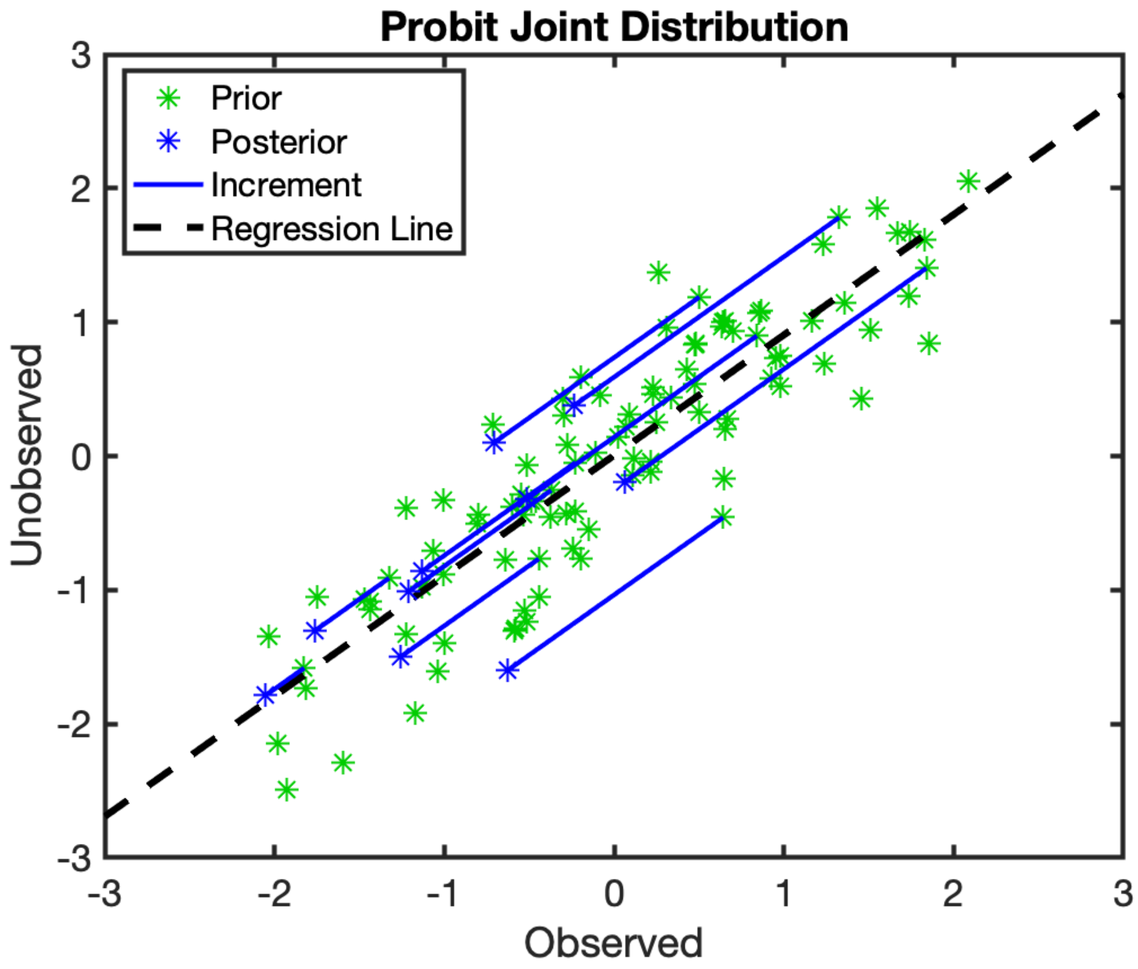
Novelty is what is transformed, not how.



Regression in Probit-Transformed Quantile Space

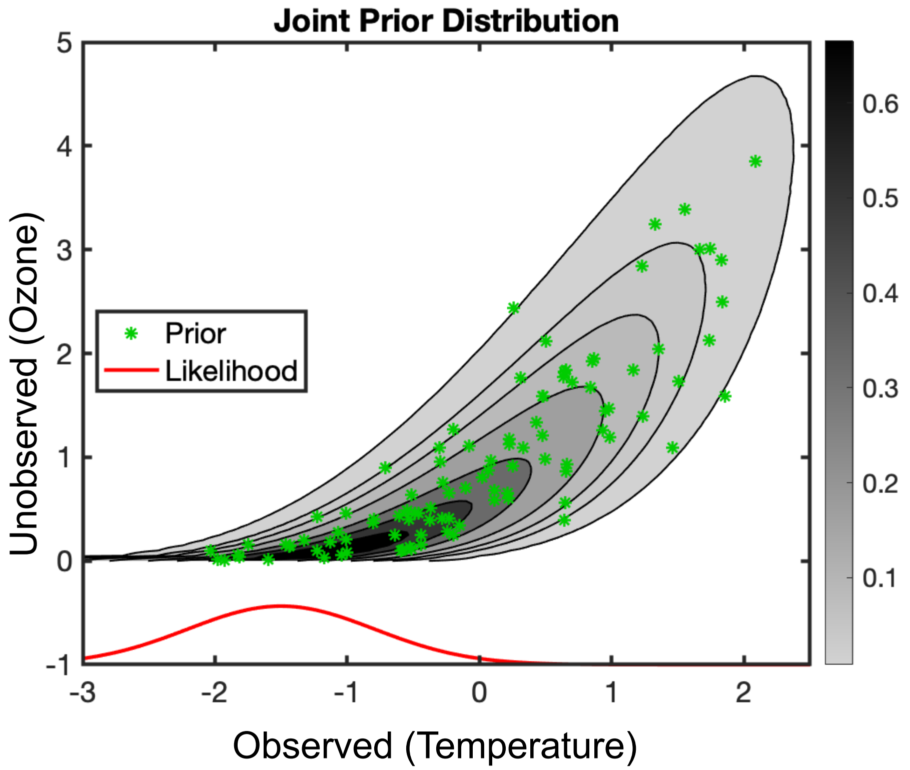
Do the regression of the observed probit increments onto the unobserved probit ensemble.

Linear regression is best unbiased linear estimator (BLUE) in this space.

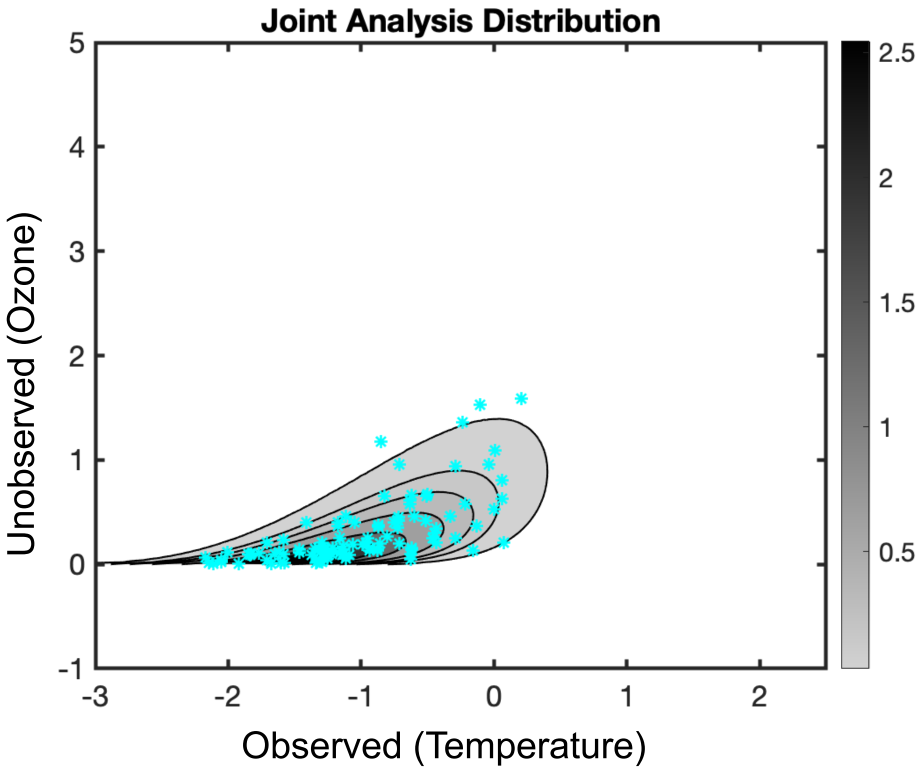


Novel, General Solutions for Nonlinear, Non-Gaussian Problems

Prior for normal-gamma distribution with 100 member ensemble.



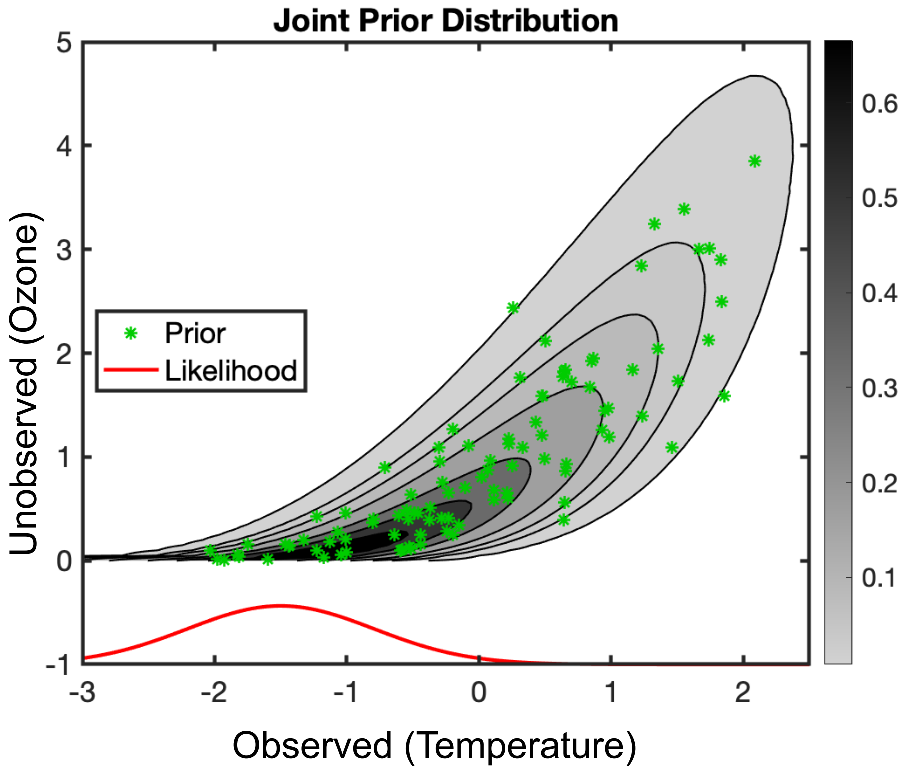
Bounds enforced. Nonlinear aspect respected.

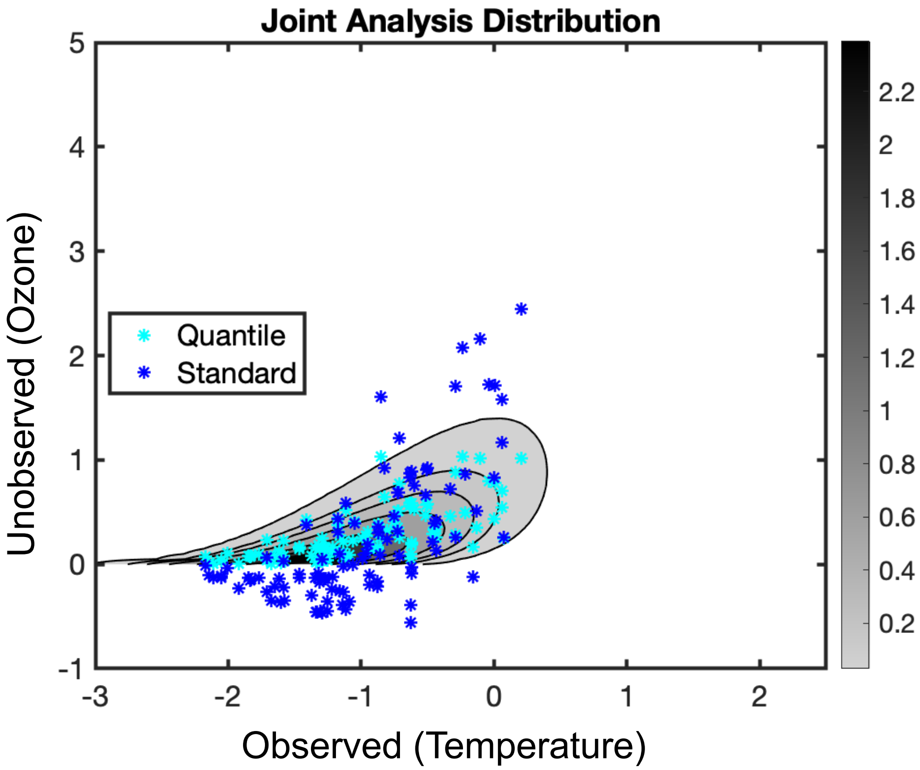


Novel, General Solutions for Nonlinear, Non-Gaussian Problems

Prior for normal-gamma distribution with 100 member ensemble.

Bounds enforced. Nonlinear aspect respected.

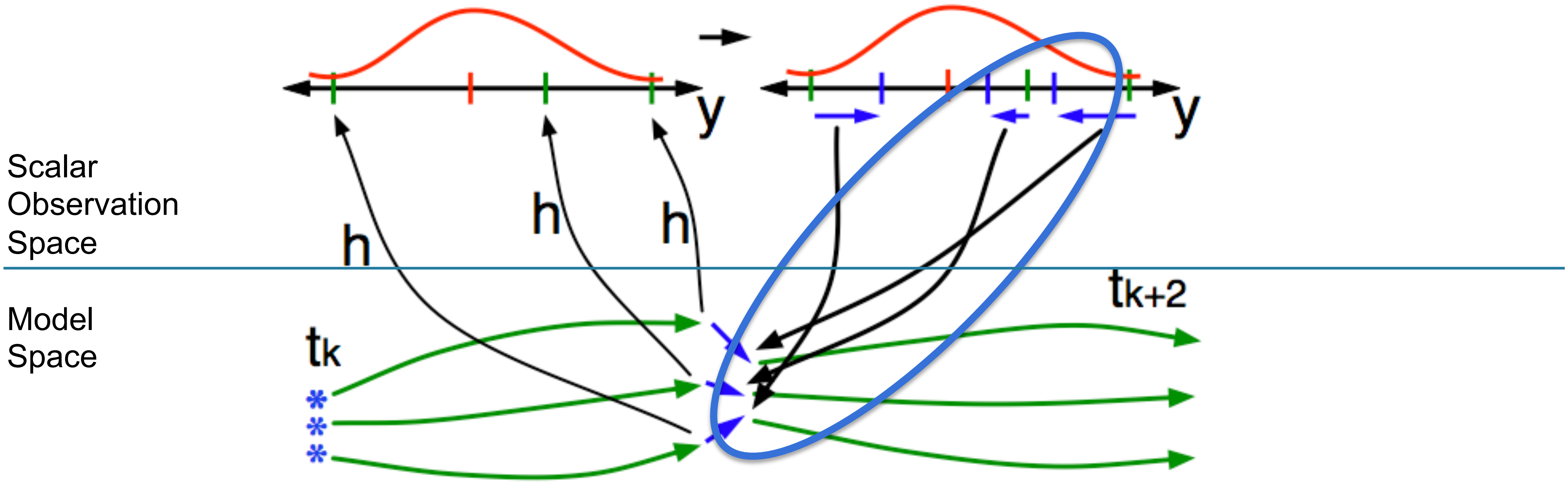




Implement Regression in a Transformed Space

Can update unobserved variables with regression in a transformed space for each state variable.

(Anderson, 2023, MWR *151*, 2759-2777)



Implement Regression in a Transformed Space

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

Mixed Distributions: A Challenge for DA

Mixed Distributions: Have both discrete and continuous probability distribution parts.

Precipitation forecast is an example:

Discrete probability of zero rain (50%),
Continuous distribution for all non-zero amounts,
(zero probability of exactly any given amount except 0).

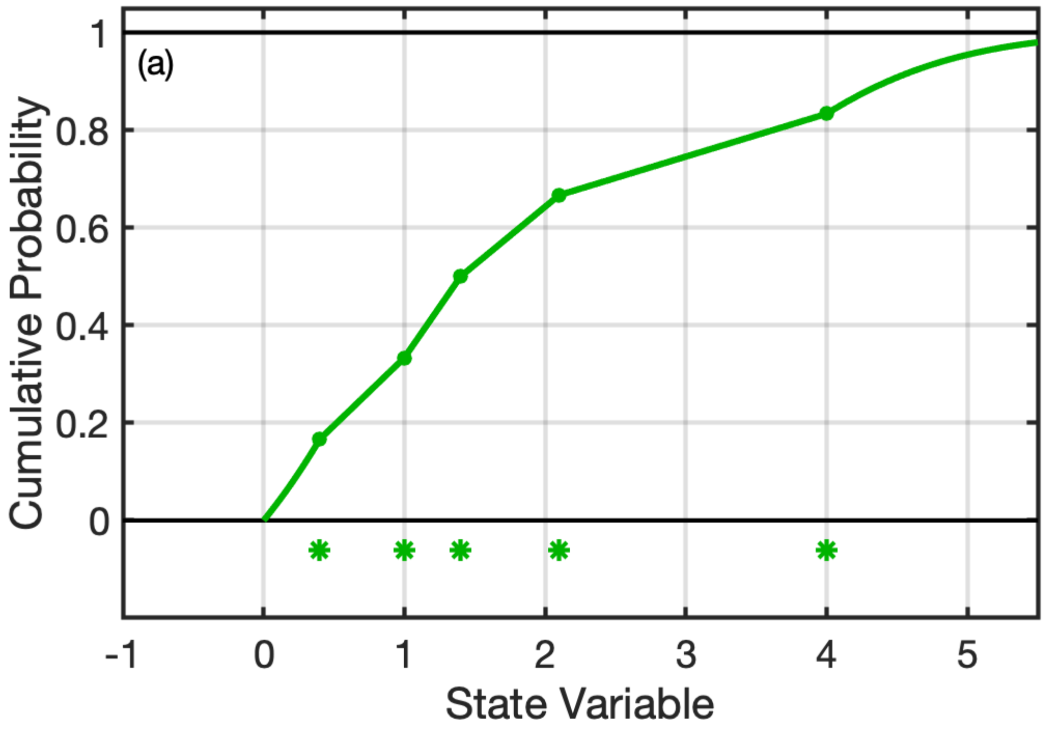
Important for some tracers and many sources (anthropogenic sources, wildfires, …).

Key: Define ‘continuous’ distributions that support duplicate ensemble members.

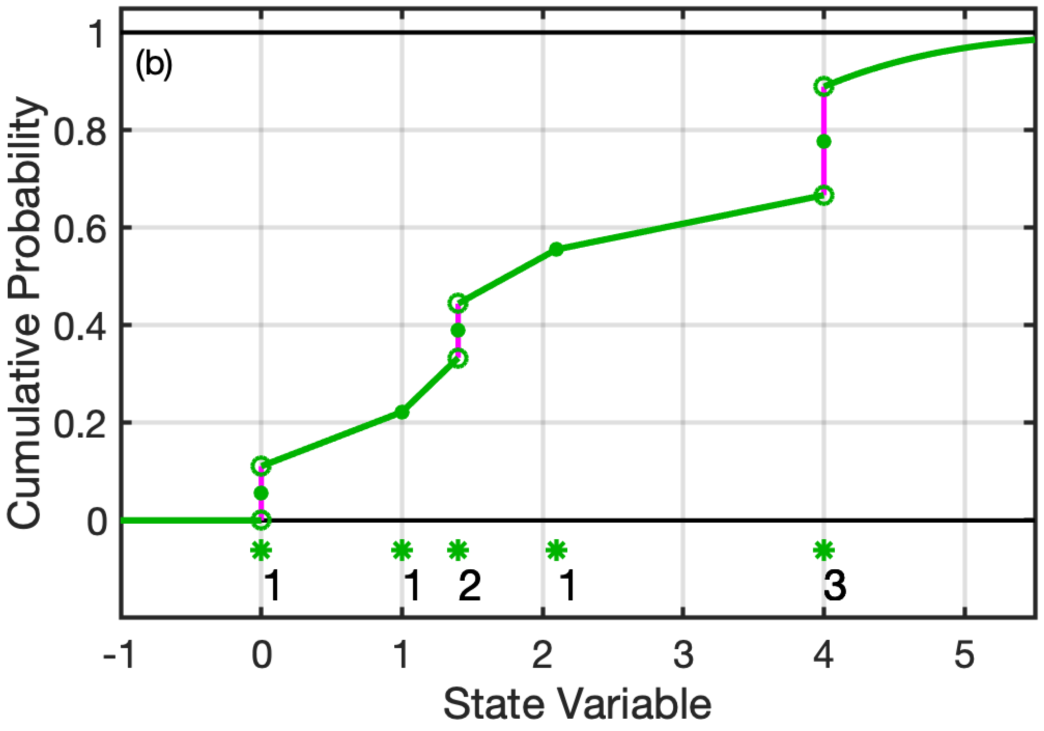
(Anderson et al., MWR *152*, 2111-2127)

Mixed Distributions: Bounded Rank Histogram Extension

Continuous CDF for 5 prior ensemble members, bounded below at 0.

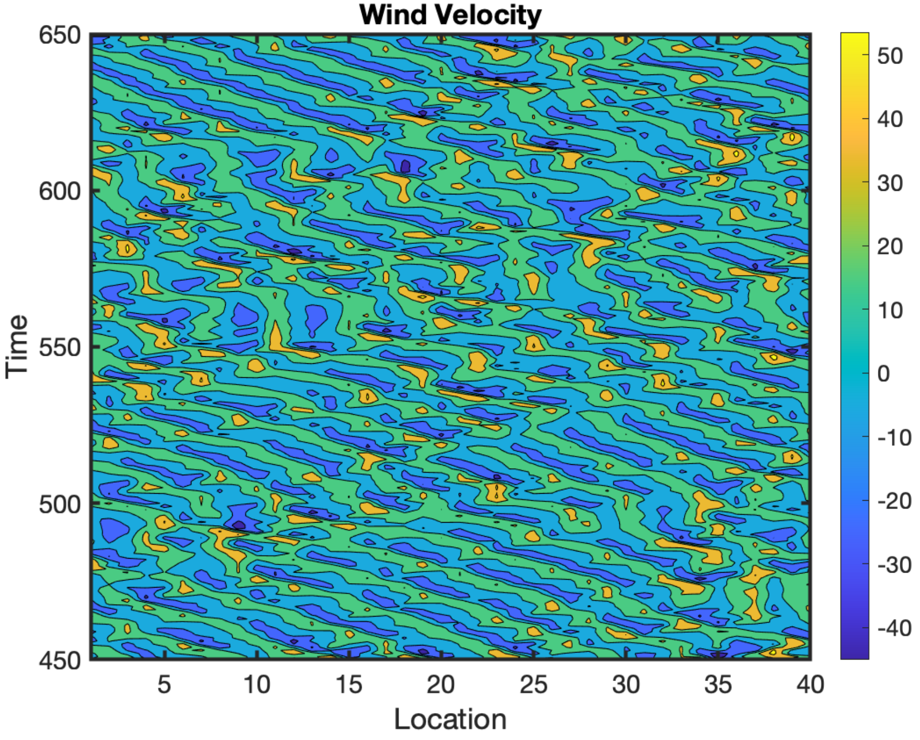


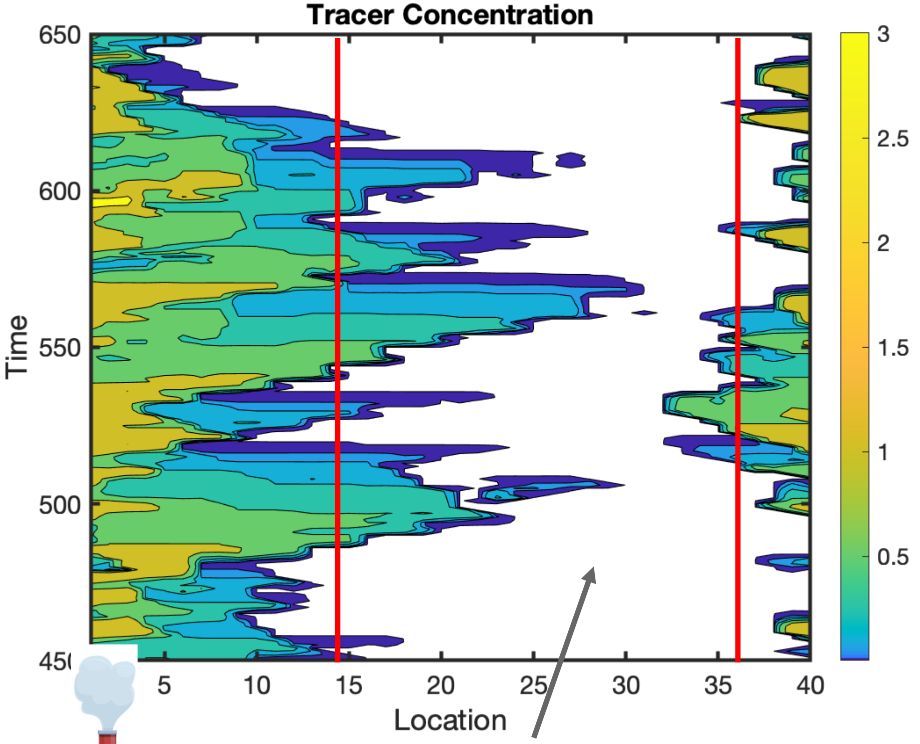
Mixed CDF for 8 prior ensemble members. An ensemble member at 0 and duplicate ensemble members at 1.4 and 4.0 lead to finite probability at those points. Modified CDF defines non-standard value at the jumps (the green midpoint on the magenta lines).



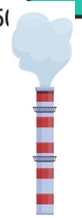
Low-Order Tracer Advection Model

Each grid point has Lorenz-96 state, tracer concentration, tracer source/sink.
Multiple of state treated as wind, conservatively advects tracer.
Example: single time constant source at grid point 1.





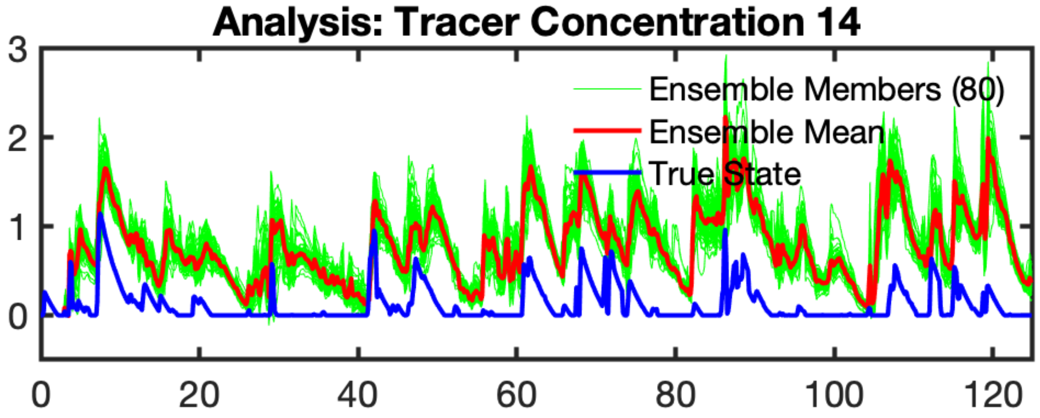
Constant Source.



Concentration can be zero far from source.

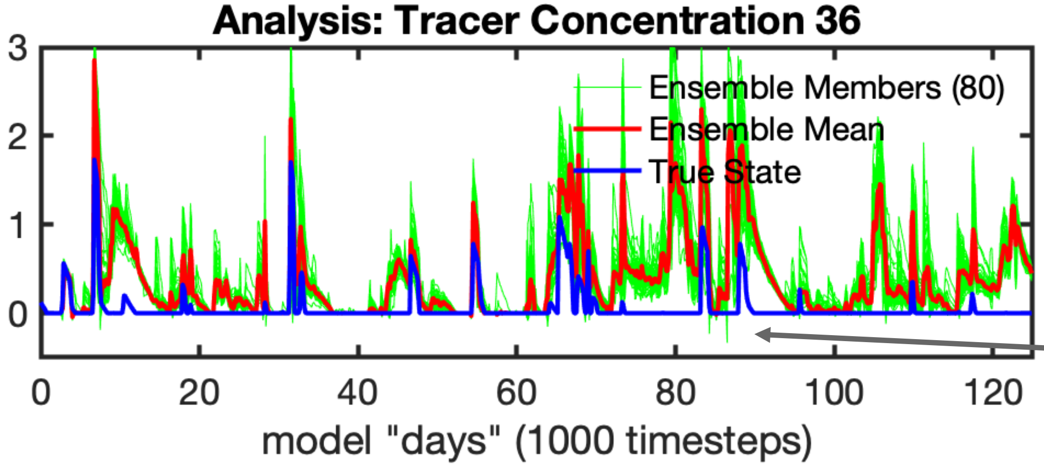
Low-Order Tracer Advection Model Example: EAKF (Normal)

Each grid point has Lorenz-96 state, tracer concentration, tracer source/sink.
Multiple of state treated as wind, conservatively advects tracer.
Example: single time constant source at grid point 1.



Observe state and concentration infrequently at each point.

Concentration error is truncated normal.



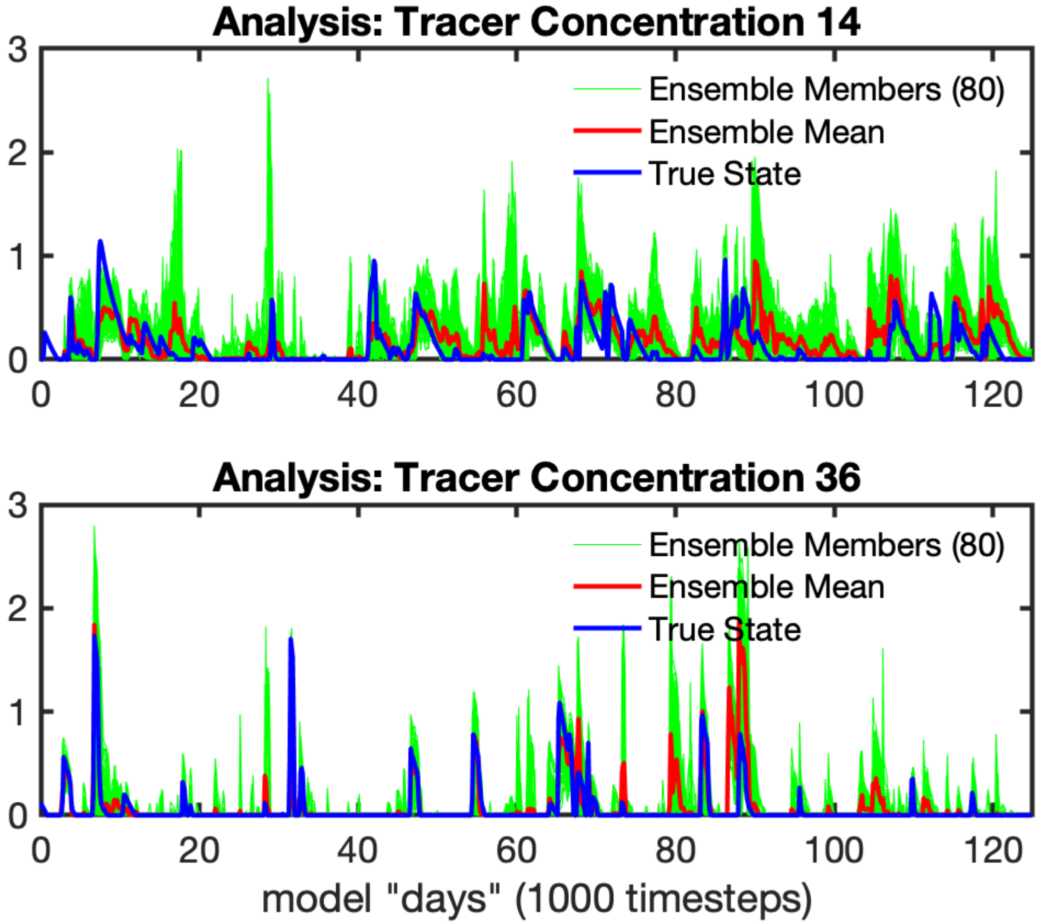
EAKF has large bias for tracers.

Can't go to all zeros.

Some negative values.

Low-Order Tracer Advection Model Example: QCEFF with BNRH

Each grid point has Lorenz-96 state, tracer concentration, tracer source/sink.
Multiple of state treated as wind, conservatively advects tracer.
Example: single time constant source at grid point 1.



Observe state and concentration infrequently at each point.

Concentration error is truncated normal.

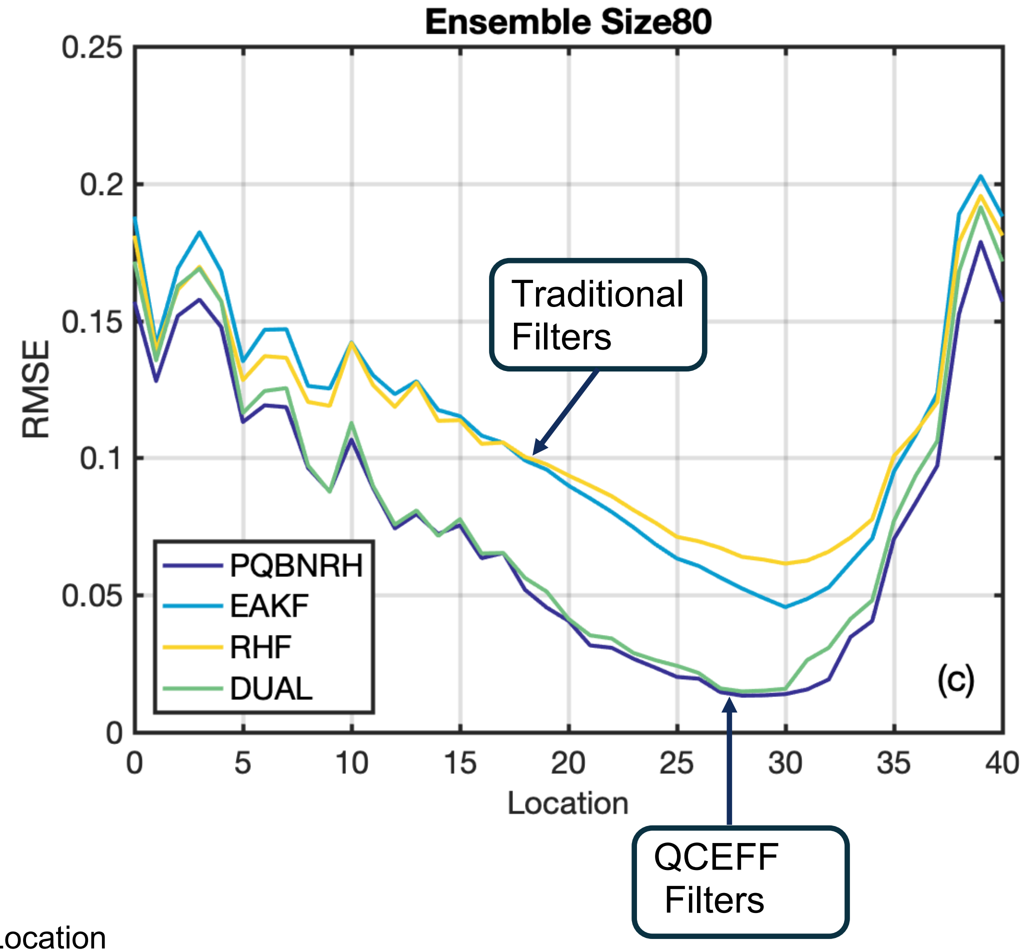
Bounded Normal Rank Histogram with probit regression is unbiased.

Can go to all zeros.

No negative values.

Low-Order Tracer Advection Model Example: Concentration RMSE

QCEFF Filters (Dark blue, Green) have smaller RMSE than traditional filters across the domain.



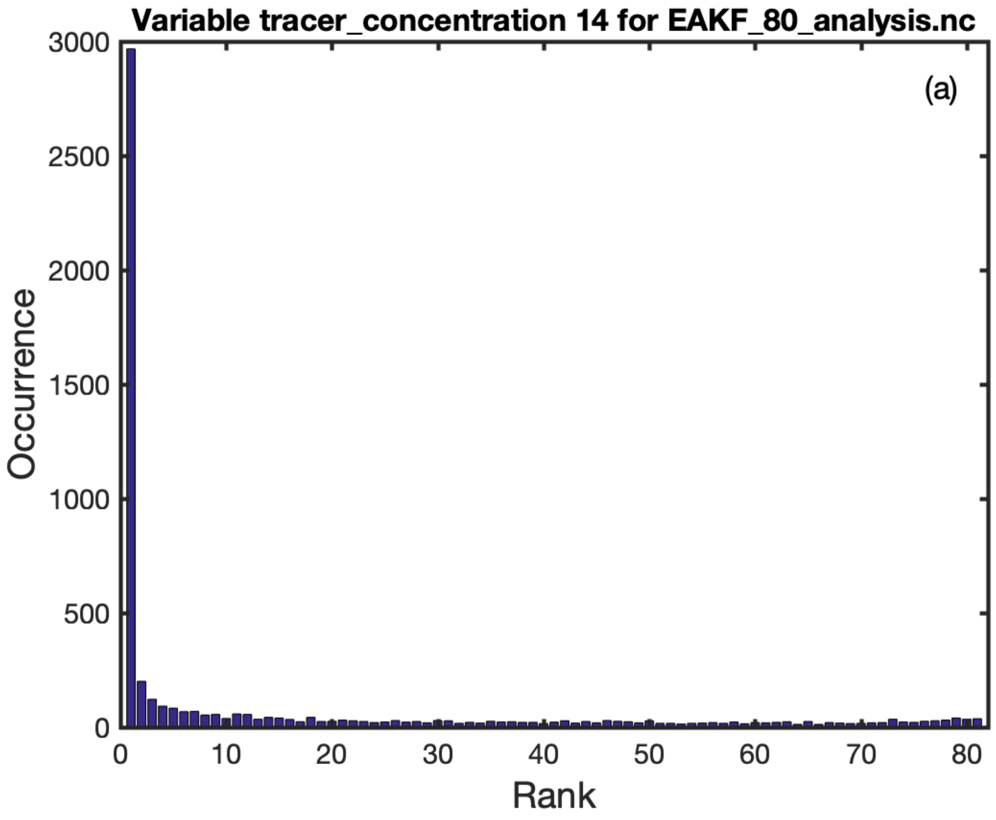
Location

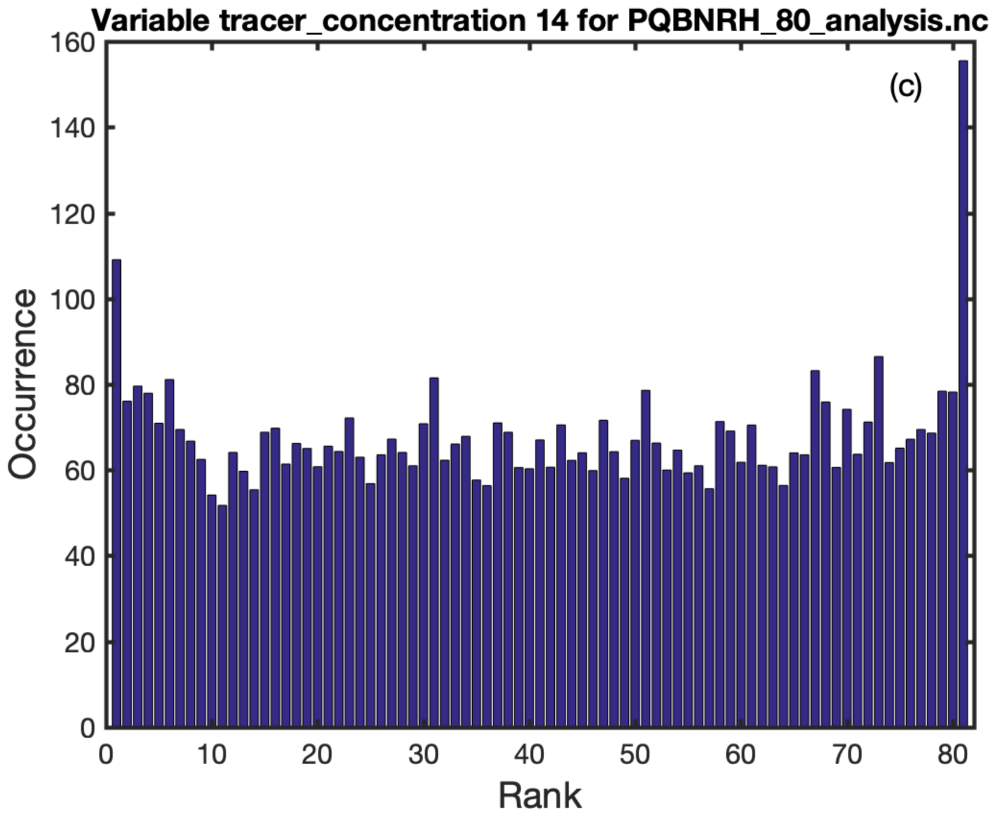
Low-Order Tracer Advection Model: Concentration Rank Histograms

Good ensembles have uniform rank histograms. Makes interpretation easy.

Normal is highly skewed.

QCEFF is close to uniform.



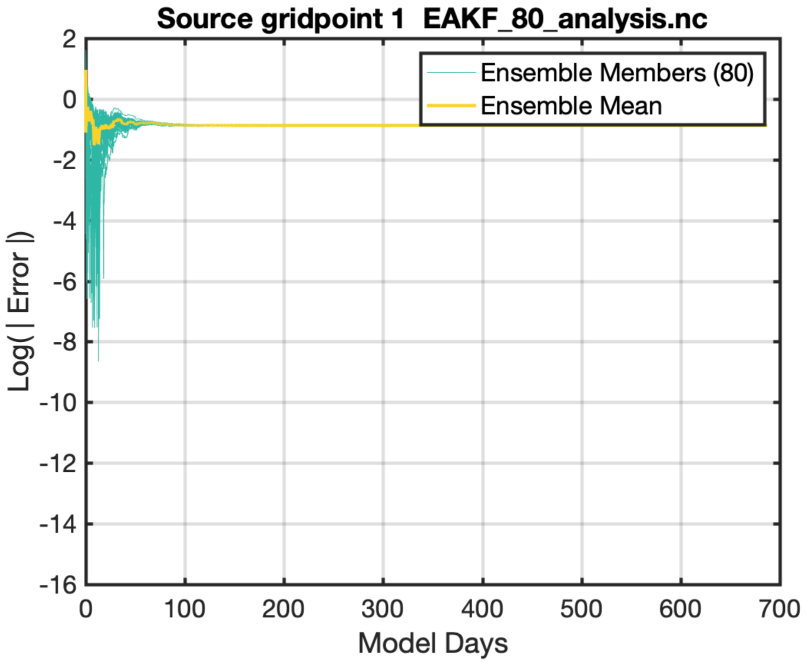


Low-Order Tracer Advection Model: Source Estimation

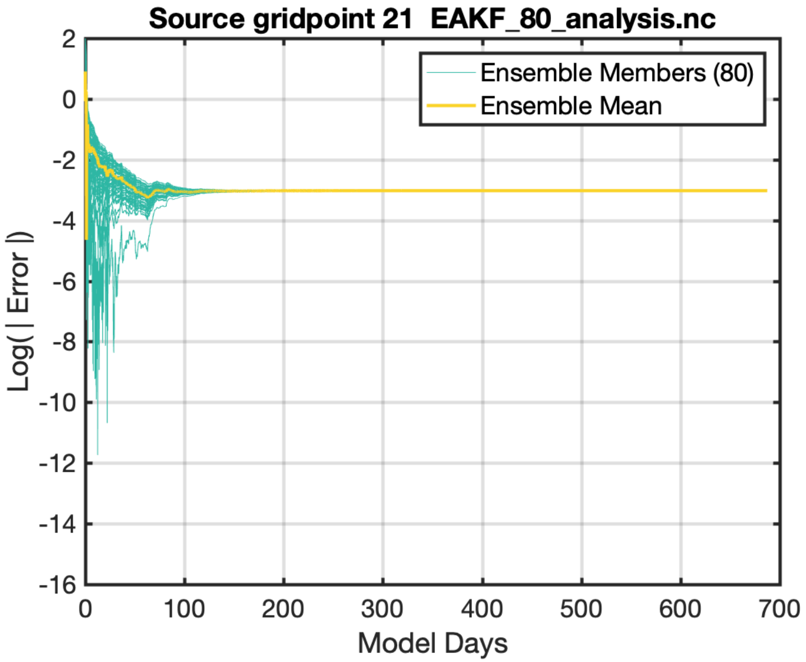
If sources are unknown, can also estimate them.

Example: single time constant source at grid point 1, zero source all other gridpoints.

EAKF for nonzero source grid point 1.



EAKF for a zero source grid point 21.

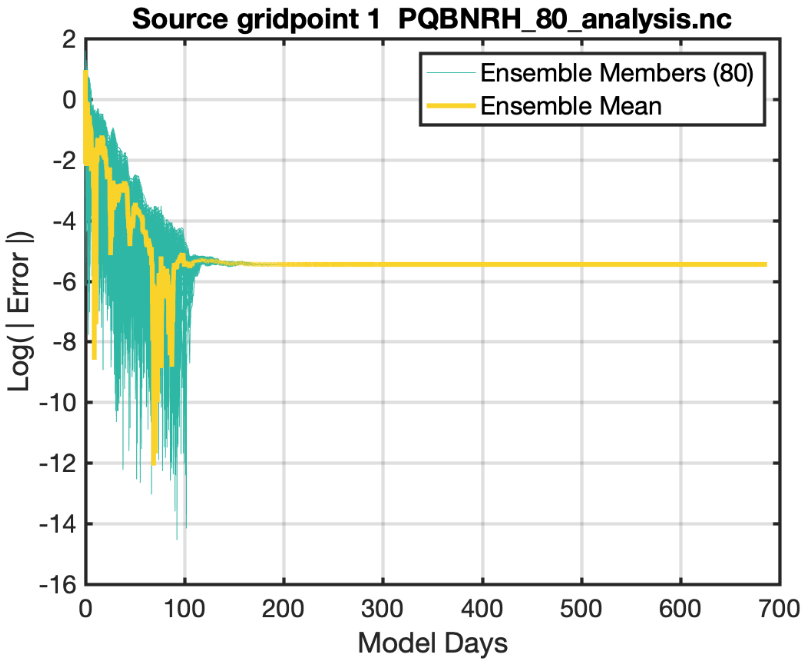


Low-Order Tracer Advection Model: Source Estimation

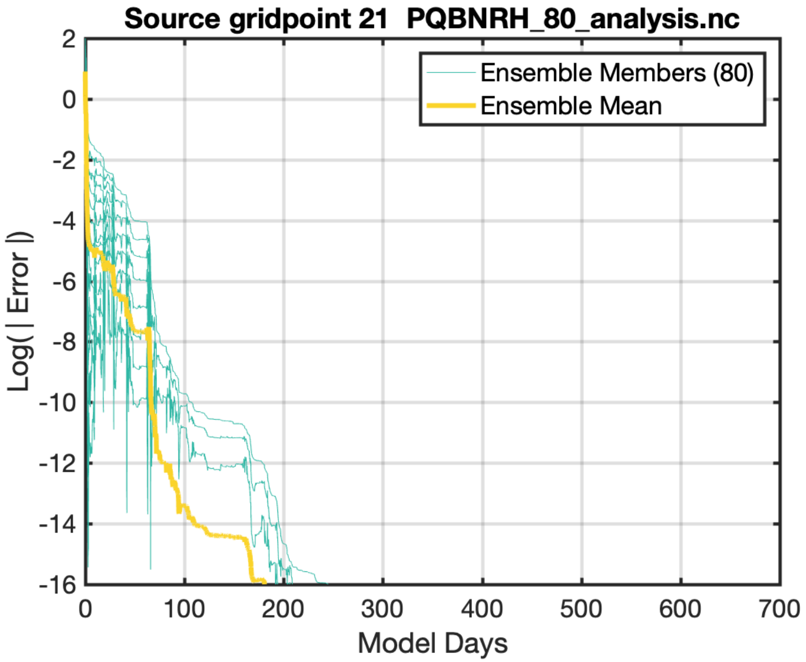
If sources are unknown, can also estimate them.

Example: single time constant source at grid point 1, zero source all other gridpoints.

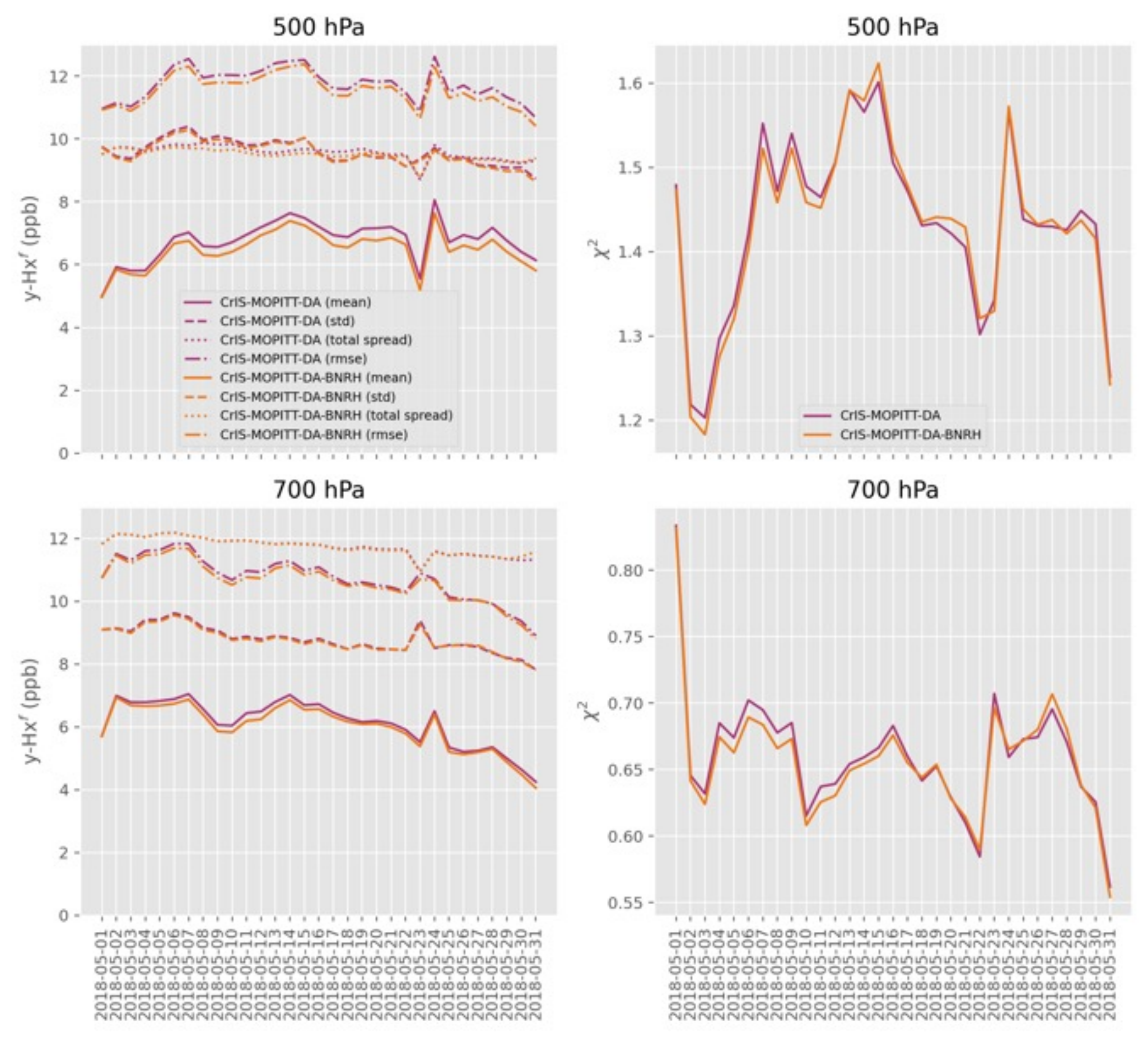
QCEFF for nonzero source grid point 1.



QCEFF for a zero source grid point 21.



First Results in Large Chemistry Model: Ben Gaubert’s Global CO DA

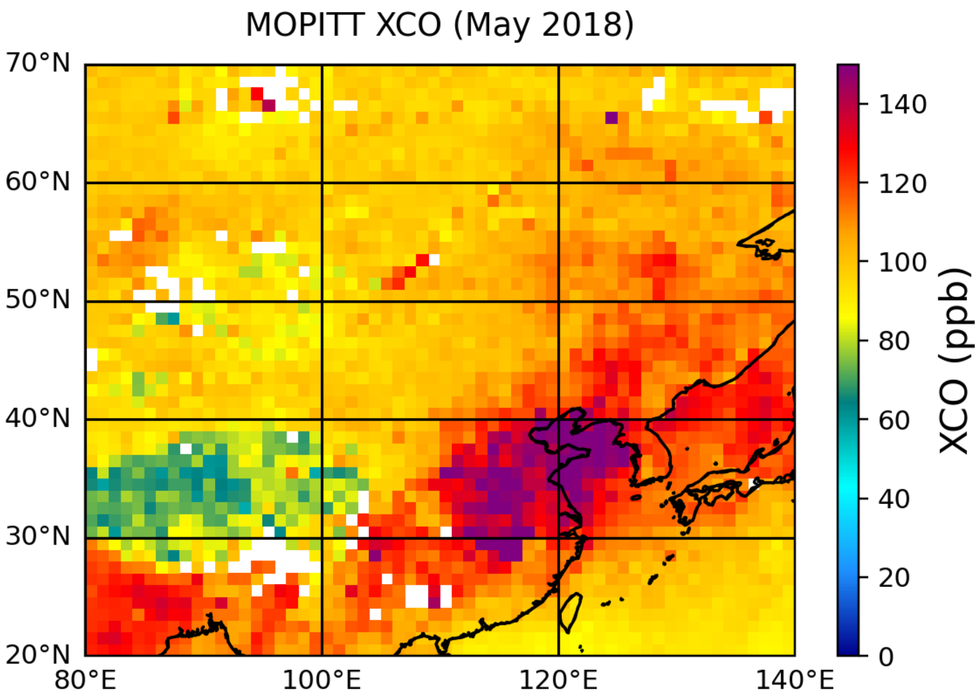


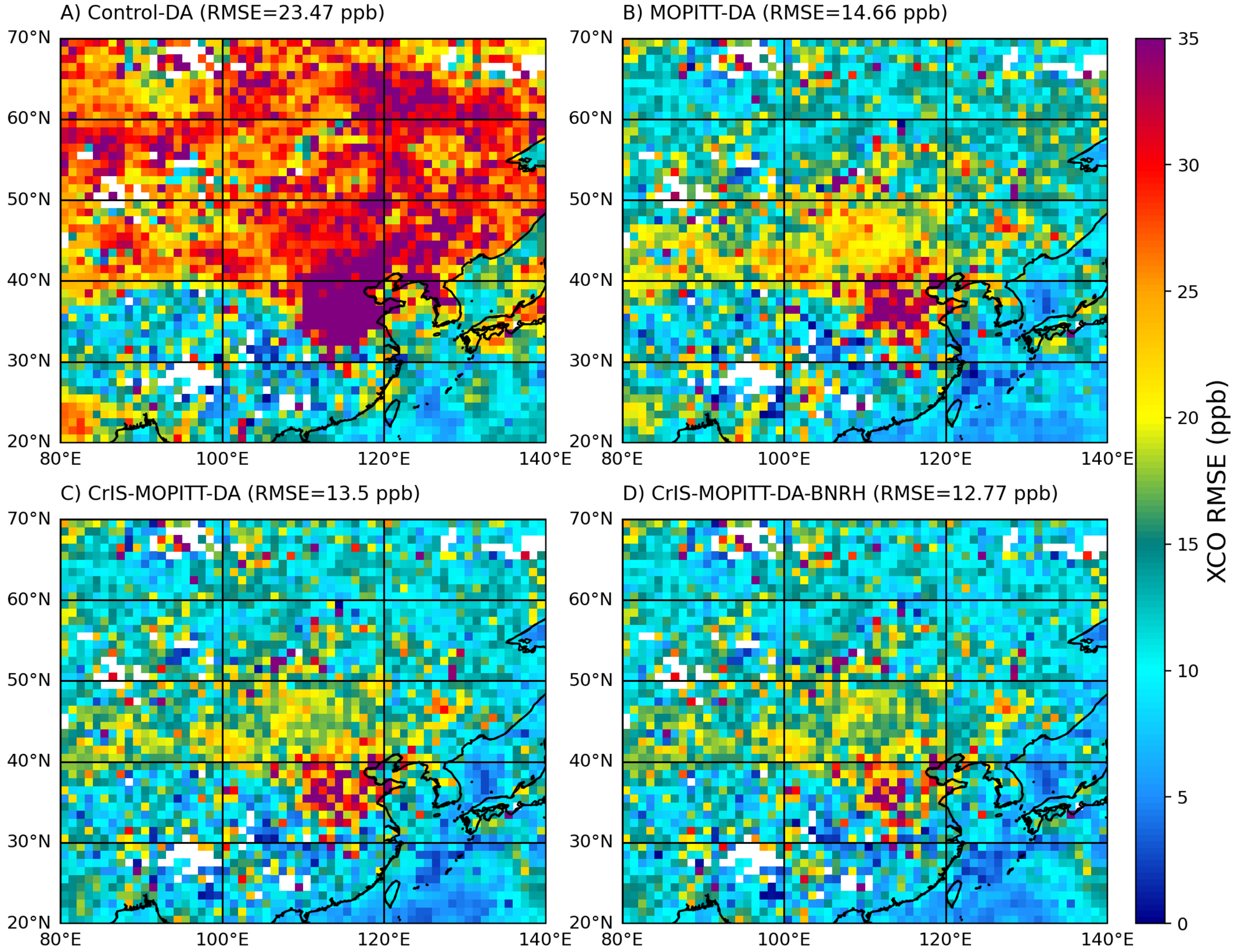
Consistent reduction in root mean square errors.

Cheyenne cost increment about 10%.

Improvement should be greater for larger ensembles.

Assimilation verification with MOPITT XCO

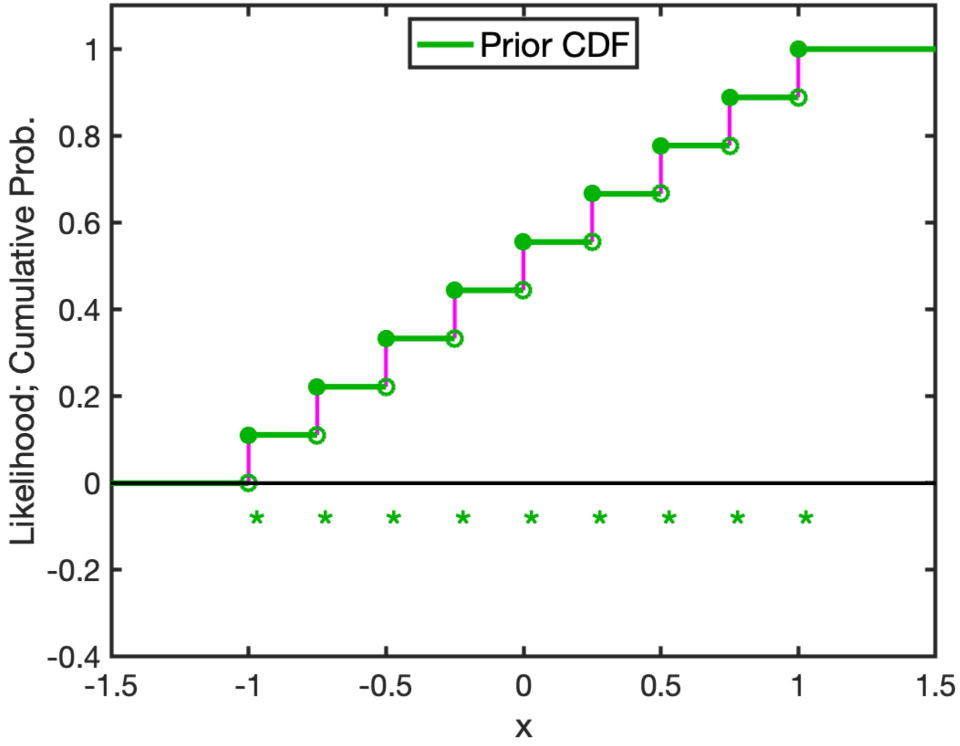




1. Control-DA
2. MOPITT-DA
3. CrIS-MOPITT-DA (EAKF)
4. CrIS-MOPITT-DA-BNRH (Bounded Normal Rank Histogram (BNRH)

QCEF Observation Update, Discrete Prior Distribution

9-member prior ensemble, symmetric around origin.

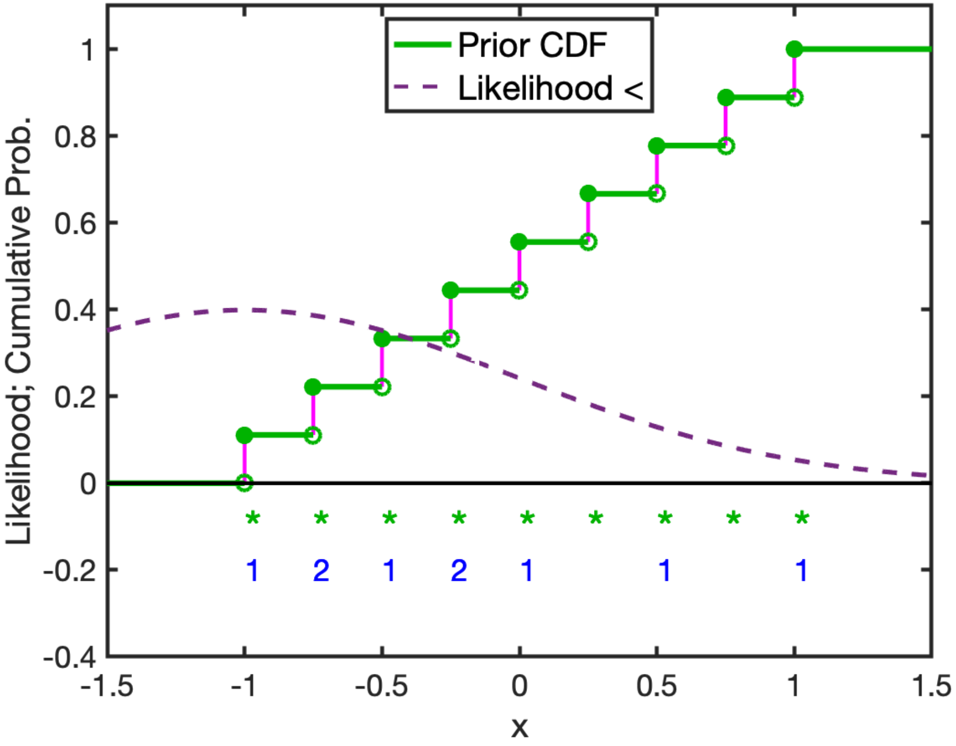


Definition of CDF:

$$F(x) = \begin{cases} 0 & if\ x < x_1 \\ \sum_{k=1}^{i} p_k & if\ x_i \leq x < x_{i+1}, \quad i \in \{1, \cdots, K-1\} \\ 1 & if\ x \geq x_K \end{cases}$$

QCEF Observation Update, Discrete Prior Distribution

9-member prior ensemble, symmetric around origin.
First example likelihood, Normal(-1, 1). Posterior ensemble counts shown in blue.

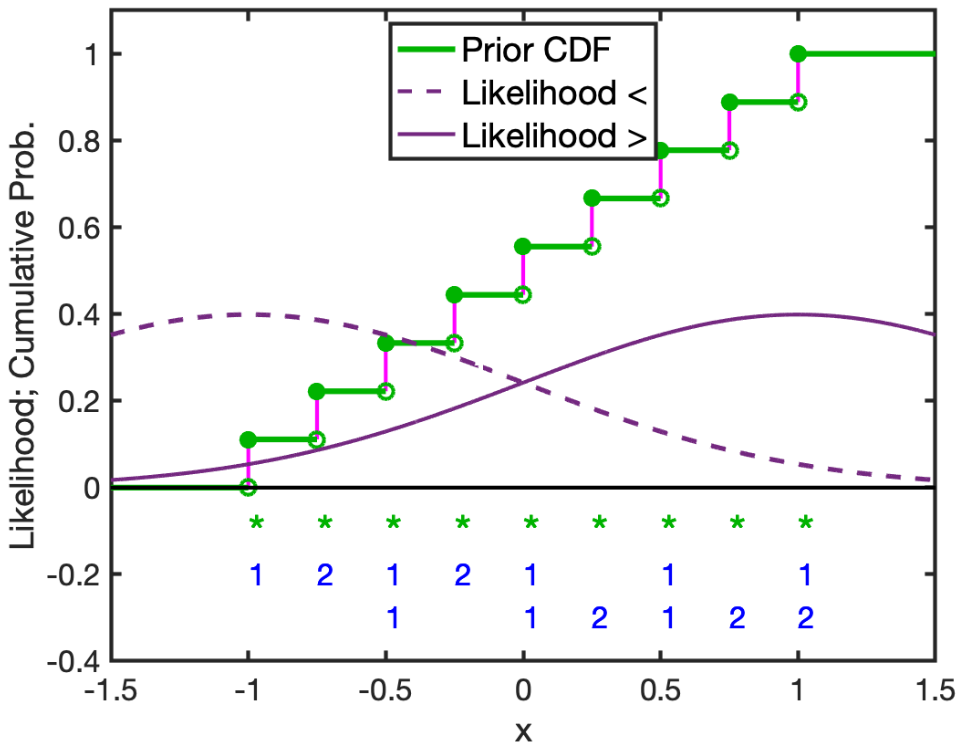


Definition of CDF:

$$F(x) = \begin{cases} 0 & if\ x < x_1 \\ \sum_{k=1}^{i} p_k & if\ x_i \le x < x_{i+1}, \quad i \in \{1, \cdots, K-1\} \\ 1 & if\ x \ge x_K \end{cases}$$

QCEF Observation Update, Discrete Prior Distribution

9-member prior ensemble, symmetric around origin.
First example likelihood, Normal(-1, 1). Posterior ensemble counts shown in blue.
Second example likelihood, Normal (1, 1).

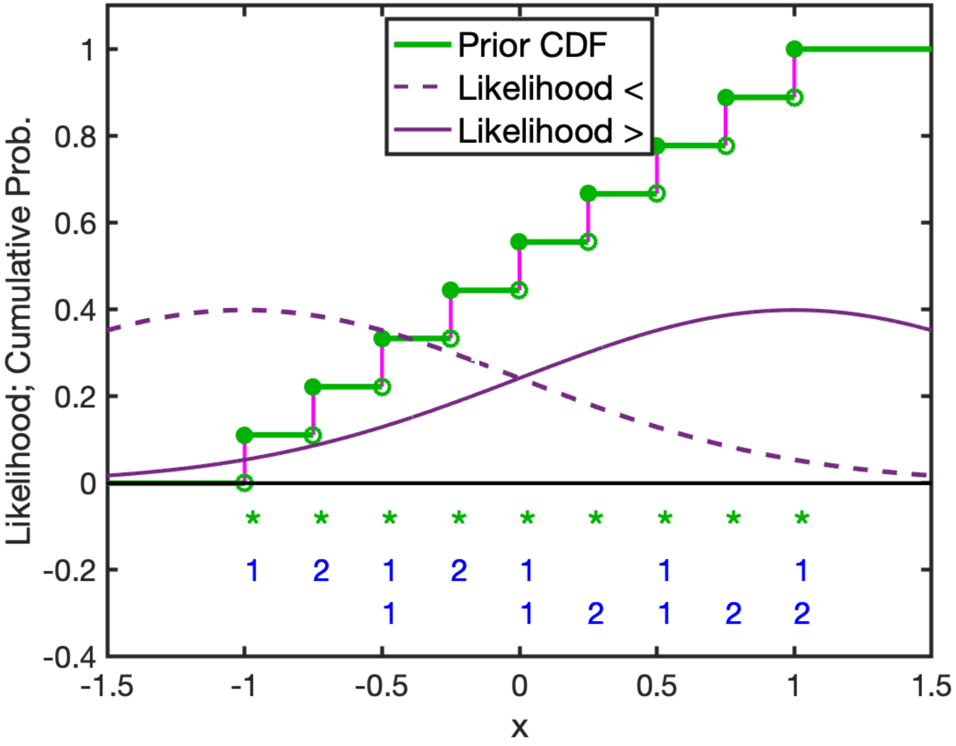


Definition of CDF:

$$F(x) = \begin{cases} 0 & if\ x < x_1 \\ \sum_{k=1}^{i} p_k & if\ x_i \le x < x_{i+1}, \quad i \in \{1, \cdots, K-1\} \\ 1 & if\ x \ge x_K \end{cases}$$

QCEF Observation Update, Discrete Prior Distribution

9-member prior ensemble, symmetric around origin.
First example likelihood, Normal(-1, 1). Posterior ensemble counts shown in blue.
Second example likelihood, Normal (1, 1).



Definition of CDF:

$$F(x) = \begin{cases} 0 & if\ x < x_1 \\ \sum_{k=1}^{i} p_k & if\ x_i \leq x < x_{i+1}, \quad i \in \{1, \cdots, K-1\} \\ 1 & if\ x \geq x_K \end{cases}$$

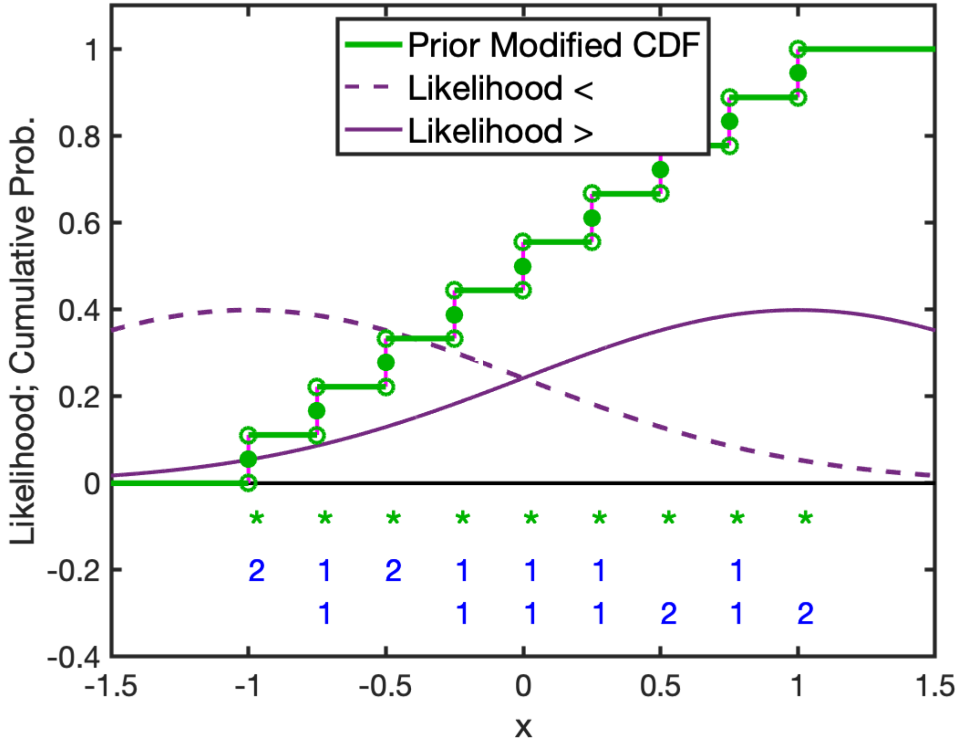
Posteriors should be antisymmetric around origin.

	Mean	SD
Posterior 1	-.2222	.6428
Posterior 2	.4444	.4965

Algorithm is biased to larger values.

QCEF Observation Update, Discrete Prior Distribution

Define a modified CDF to avoid the bias.



Definition of Modified CDF:

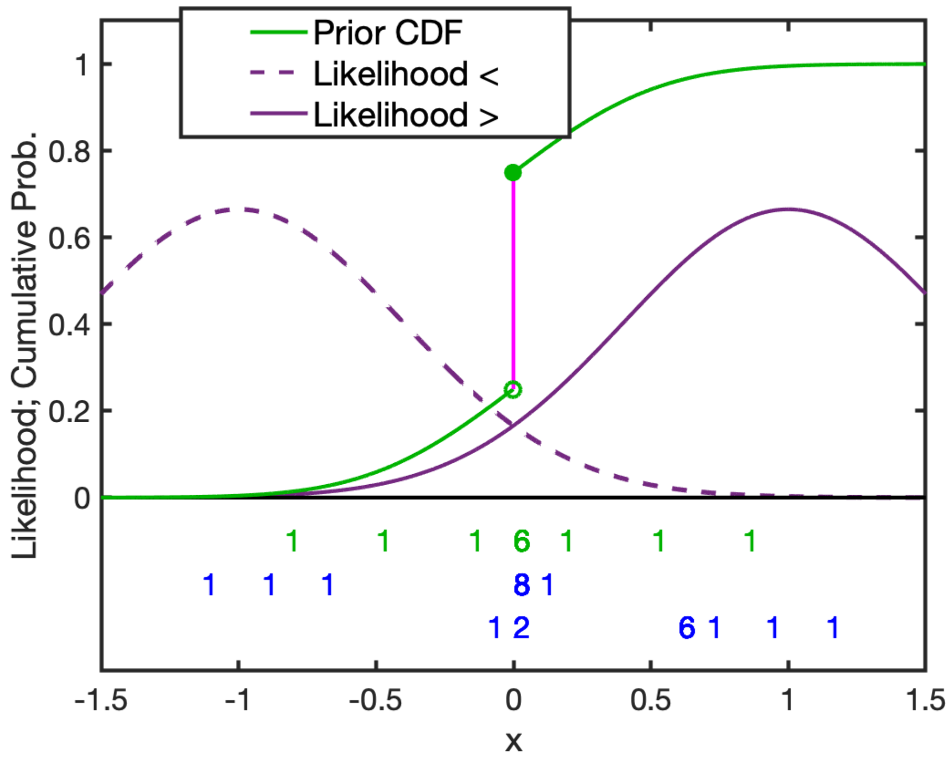
$$\tilde{F}(x) = \begin{cases} 0 & if\, x < x_1 \\ \sum_{k=1}^{i} p_k & if\ x_i < x < x_{i+1},\quad i \in \{1,\ldots,K-1\} \\ 1 & if\ x > x_K \\ \sum_{k=1}^{i-1} p_k + \frac{p_i}{2} & if\ x = x_i \end{cases}$$

Posteriors are antisymmetric around origin.

	Mean	SD
Posterior 1	-.3333	.5863
Posterior 2	.3333	.5863

QCEF Observation Update, Mixed Prior Distribution

A 12-member ensemble from a mixed distribution.
Duplicate members define a discrete point, 6 members at zero here.
Other members define a normal distribution; 50% of the total probability.

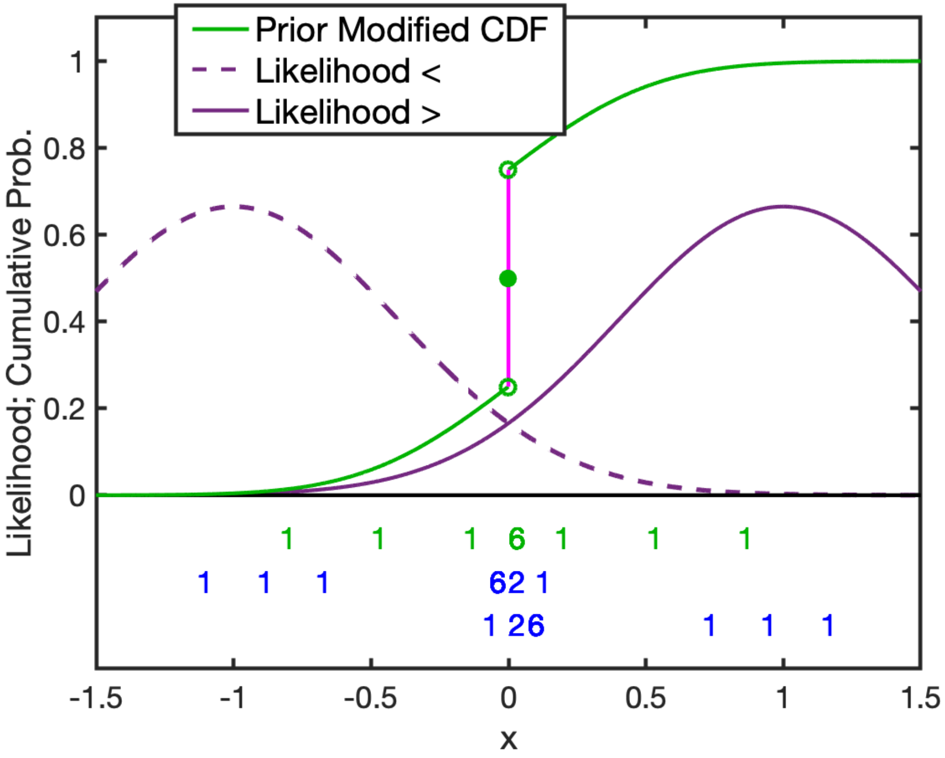


Traditional CDF leads to biased posteriors.

	Mean	SD
Posterior 1	-.2214	.4308
Posterior 2	.5281	.3727

QCEF Observation Update, Mixed Prior Distribution

A 12-member ensemble from a mixed distribution.
Duplicate members define a discrete point, 6 members at zero here.
Other members define a normal distribution; 50% of the total probability.

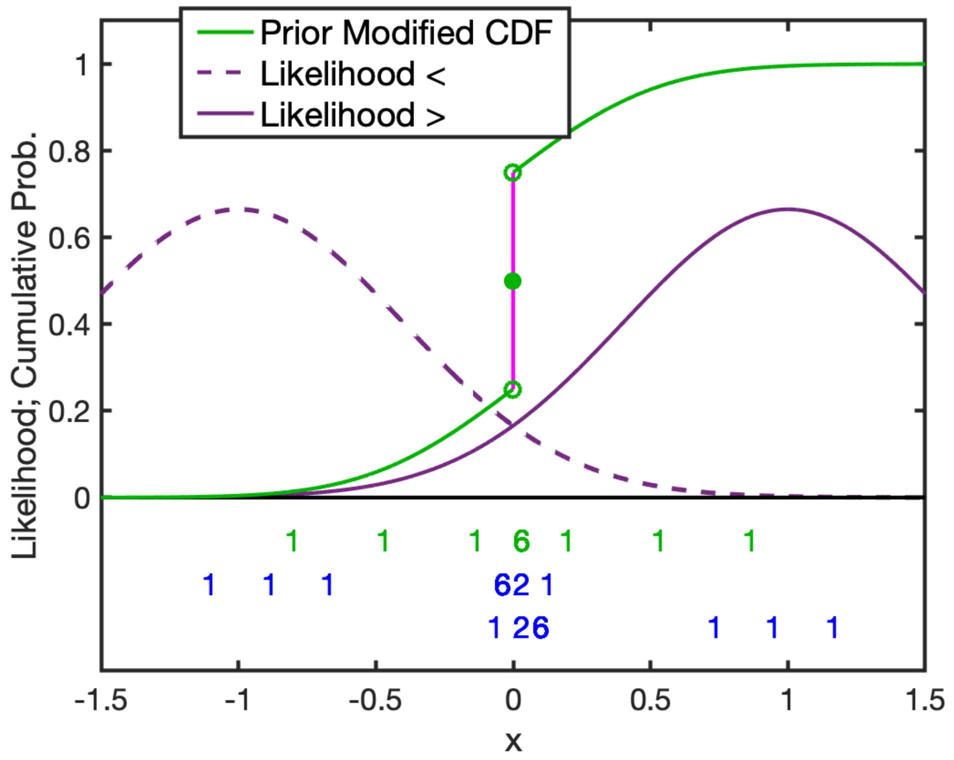


Modified CDF is unbiased.

	Mean	SD
Posterior 1	-.2564	.4123
Posterior 2	.2564	.4123

QCEF Observation Update, Mixed Prior Distribution

A 12-member ensemble from a mixed distribution.
Duplicate members define a discrete point, 6 members at zero here.
Other members define a normal distribution; 50% of the total probability.



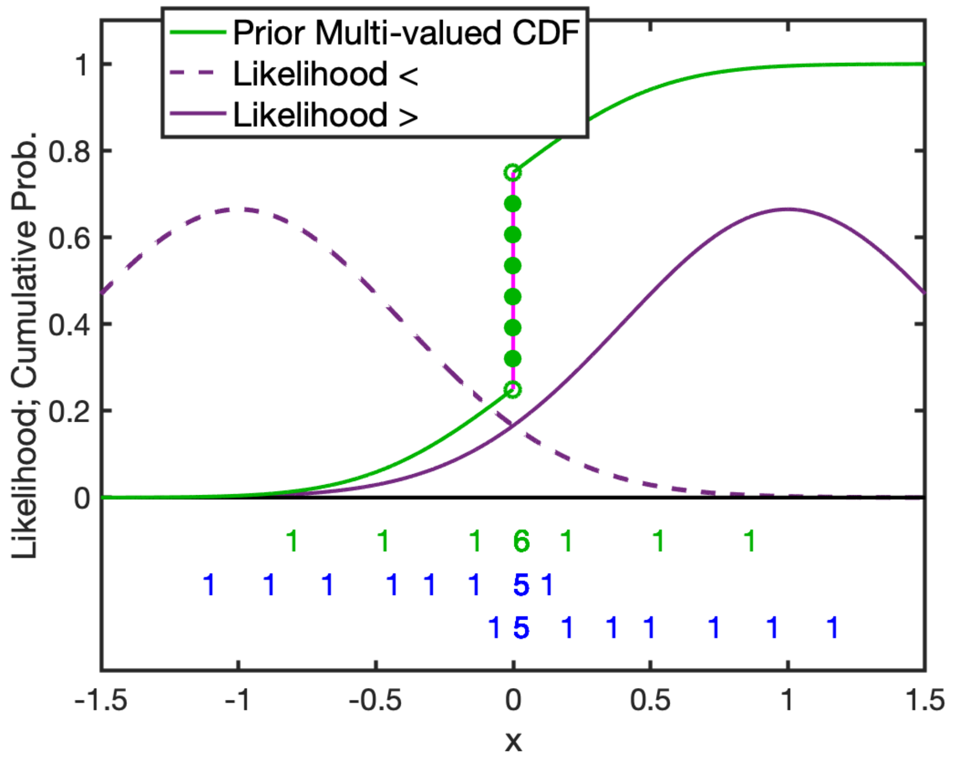
Modified CDF is unbiased.

	Mean	SD
Posterior 1	-.2564	.4123
Posterior 2	.2564	.4123

But all identical prior ensemble members must move together. Clearly incorrect for some applications.

QCEF Observation Update, Mixed Prior Distribution

A 12-member ensemble from a mixed distribution.
Duplicate members define a discrete point, 6 members at zero here.
Other members define a normal distribution; 50% of the total probability.



Multivalued CDF is also unbiased.

	Mean	SD
Posterior 1	-.3023	.4147
Posterior 2	.3023	.4147

Ensemble members with same prior value can have different posterior values. Can lead to balance issues.

QCEF Observation Update, Mixed Prior Distribution: Summary

Definition of the CDF is important for QCEFF applications in discrete and mixed distributions.

Standard definition leads to bias towards larger values.

Other definitions have pros and cons.

However, still much better than normal distributions for certain problems.

Anderson et al., 2024, MWR, 2111-2127.

Closing Thoughts

Earth system DA problems are nonlinear, non-Gaussian, have mixed distributions:

Tracer concentration and sources;
Parameter estimation;
Sea ice, snow, other depths and concentrations.

DART now provides QCEFF methods:

Arbitrary distributions in observation space;
Arbitrary univariate spatial transforms before updating state variables;
Support for duplicate ensemble members.

Initial application to large models is promising:

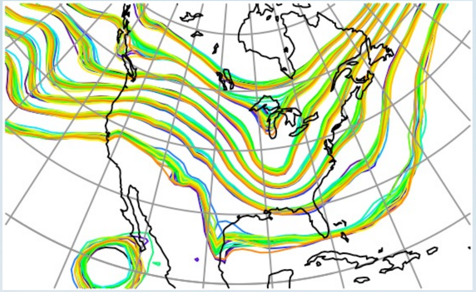
Generally improved forecast fit to observations;
Incremental computational cost generally O(0.1)

Try it out at https://dart.ucar.edu

NCAR
UCAR

CISL

Computational & Information Systems Lab



Questions?

NCAR | National Center for
UCAR | Atmospheric Research

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.

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.

Distributions for Important Forecast Quantities

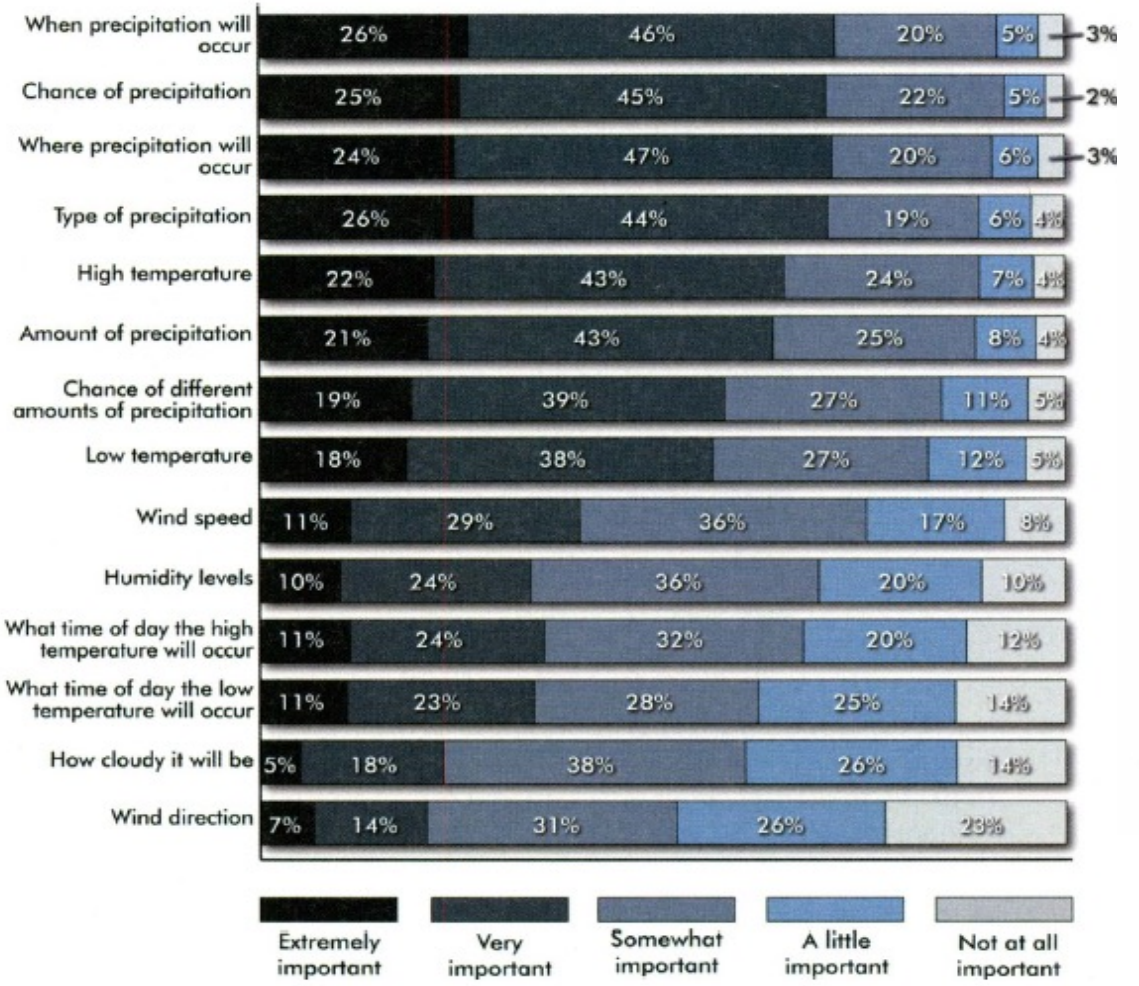


Fig. 5. Respondents' rating of importance for different potential components of weather forecasts (*n* = 1,465). The survey question asked "How important is it to you to have the information listed below as part of a weather forecast?"

300 BILLION SERVED

Sources, Perceptions, Uses, and Values of Weather Forecasts

BY JEFFREY K. LAZO, REBECCA E. MORSS, AND JULIE L. DEMUTH

A nationwide survey indicates that the U.S. public obtains several hundred billion forecasts each year, generating $31.5 billion in benefits compared to costs of $5.1 billion.

Every day, the U.S. weather enterprise collectively disseminates numerous weather forecasts to the U.S. public through various media. Considering the range of forecasts generated at a variety of spatial and temporal scales, the array of forecast providers and communication channels, and the size and diver-

Research on aspects of these issues has been conducted for specific geographical areas (e.g., Saviers and Van Bussum 1997; Lazo and Chestnut 2002), for specific events or weather phenomena and decision-making situations (e.g., Katz and Murphy 1997; Anderson-Berry et al. 2004; Stewart et al. 2004; Call

BAMS. 2009, 785-798.

Forecast Distributions for Important Quantities

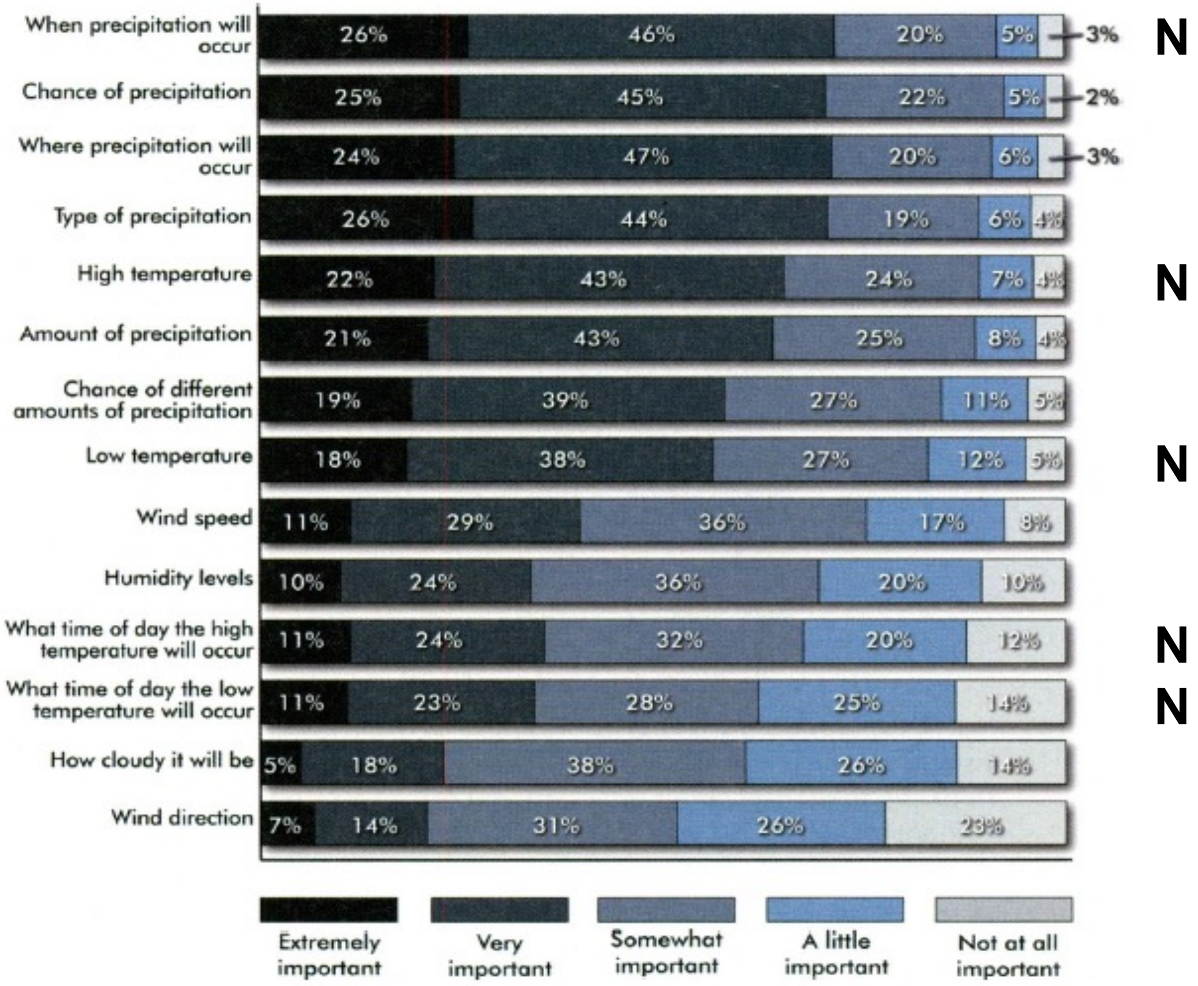
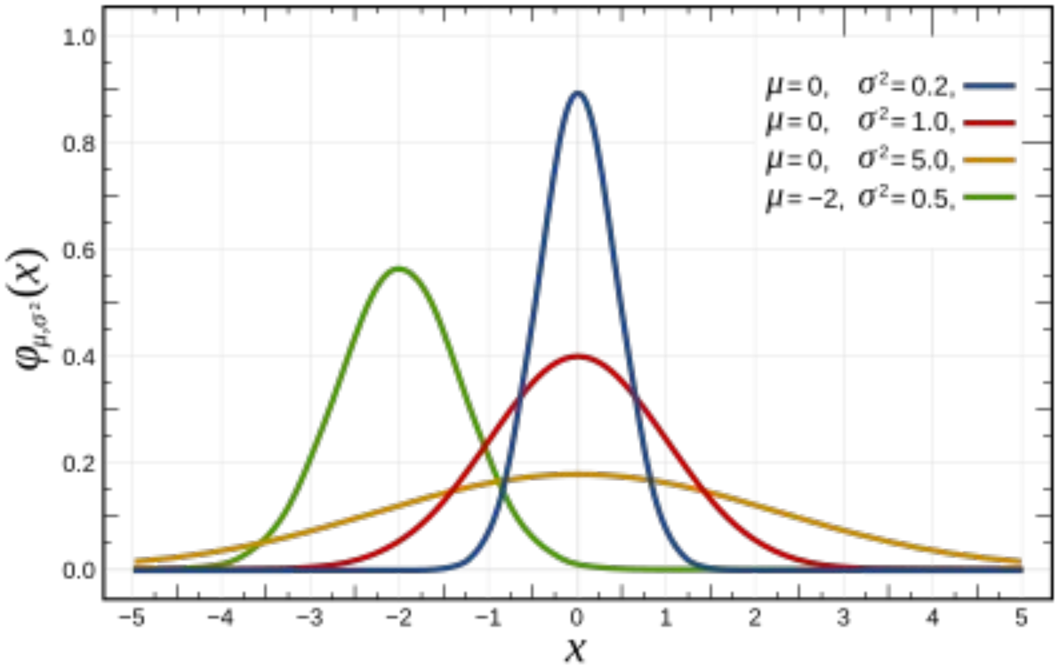


FIG. 5. Respondents' rating of importance for different potential components of weather forecasts (n = 1,465). The survey question asked "How important is it to you to have the information listed below as part of a weather forecast?"

Normal (N)



Forecast Distributions for Important Quantities

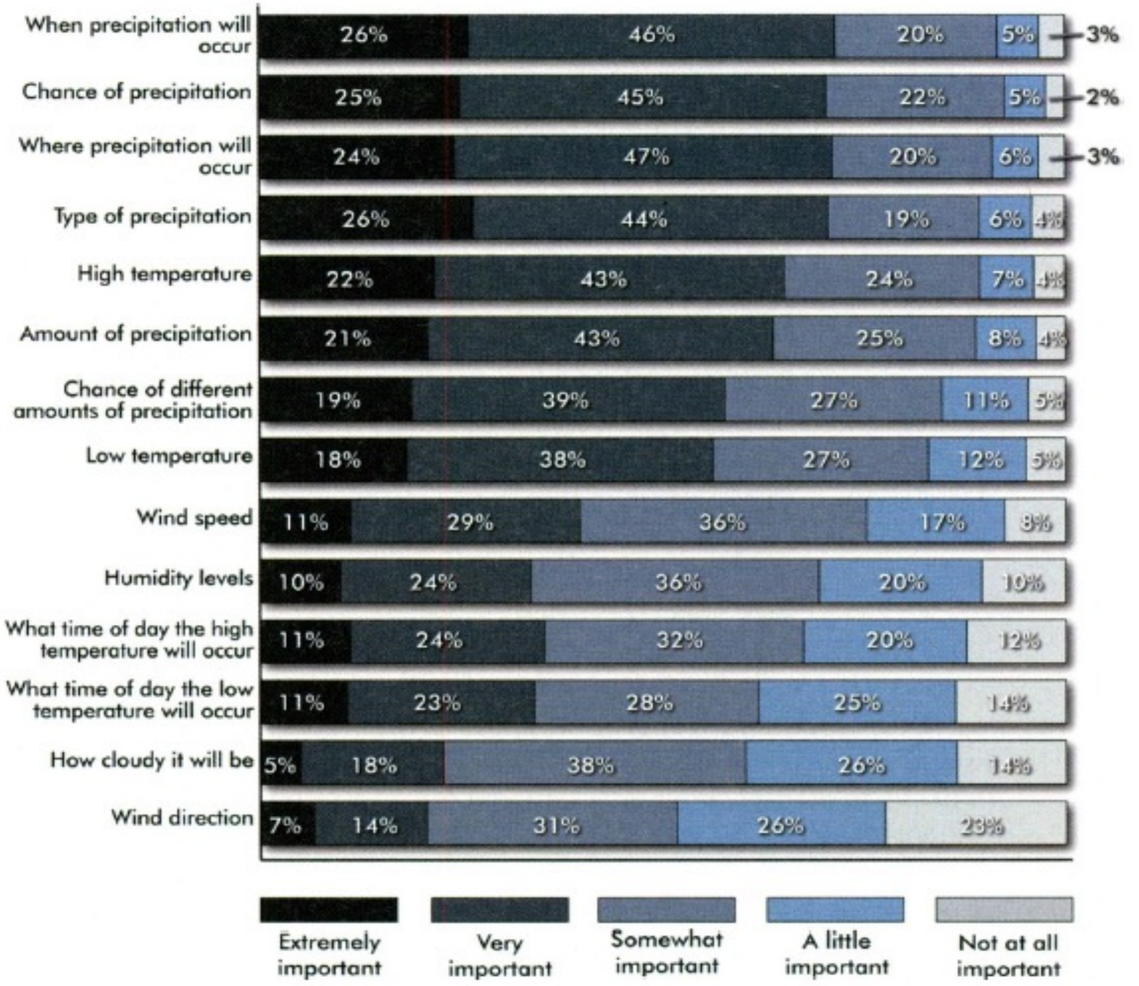
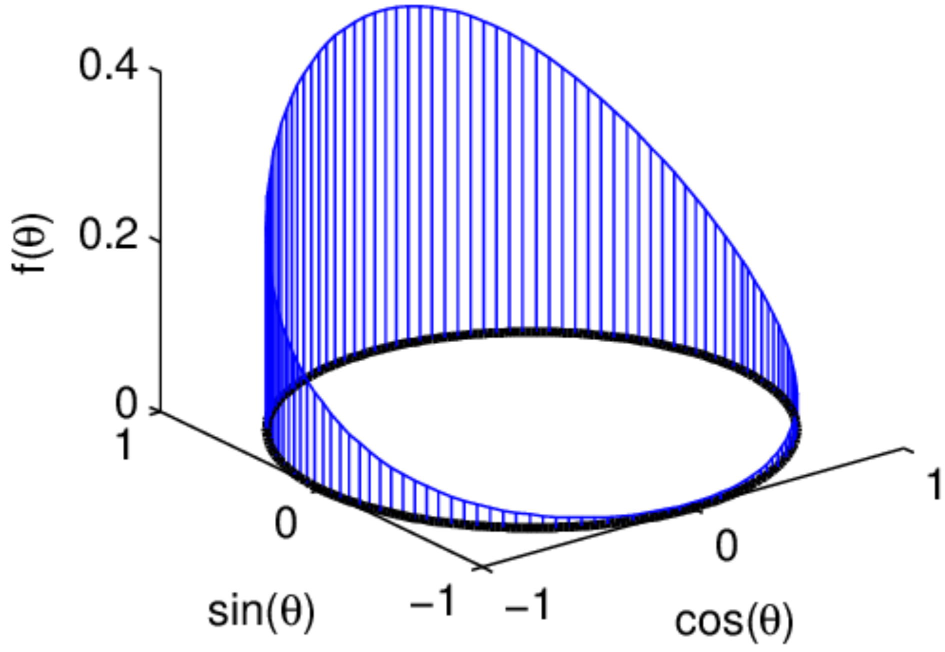


FIG. 5. Respondents' rating of importance for different potential components of weather forecasts (n = 1,465). The survey question asked "How important is it to you to have the information listed below as part of a weather forecast?"

WN

Wrapped Normal (WN)



Forecast Distributions for Important Quantities

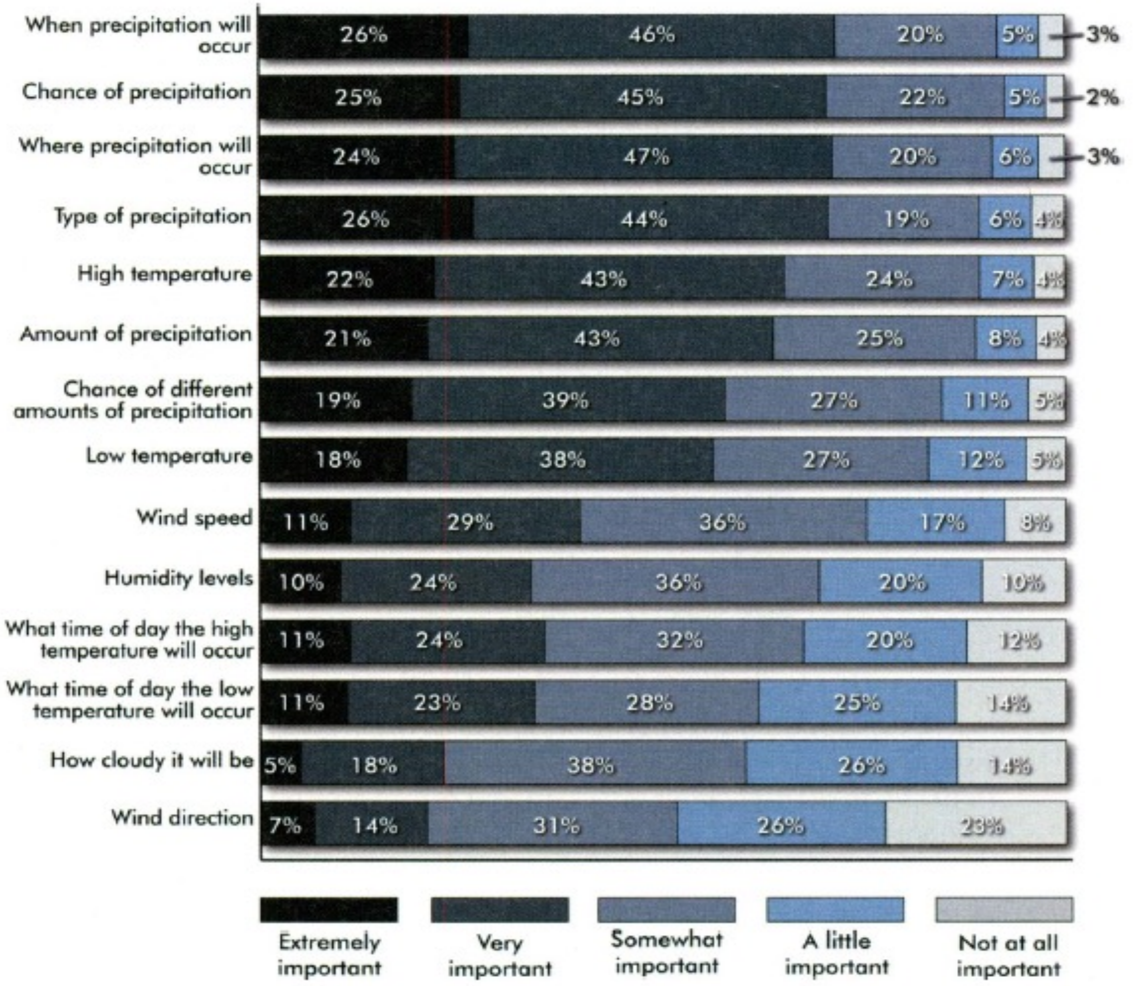
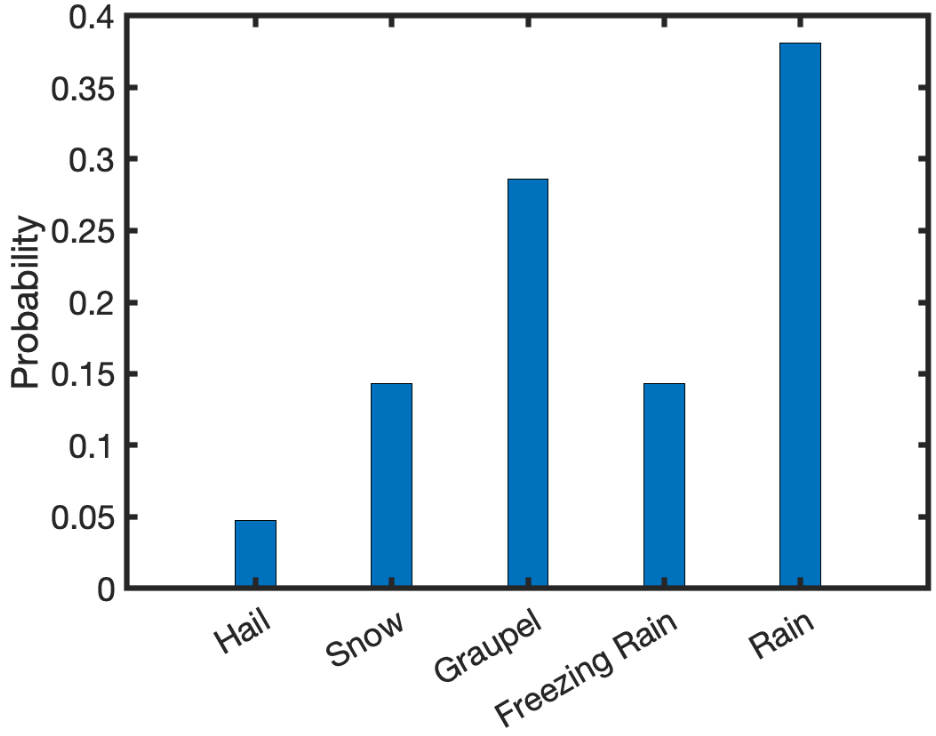


Fig. 5. Respondents' rating of importance for different potential components of weather forecasts (n = 1,465). The survey question asked "How important is it to you to have the information listed below as part of a weather forecast?"

D

Discrete (D)



Forecast Distributions for Important Quantities

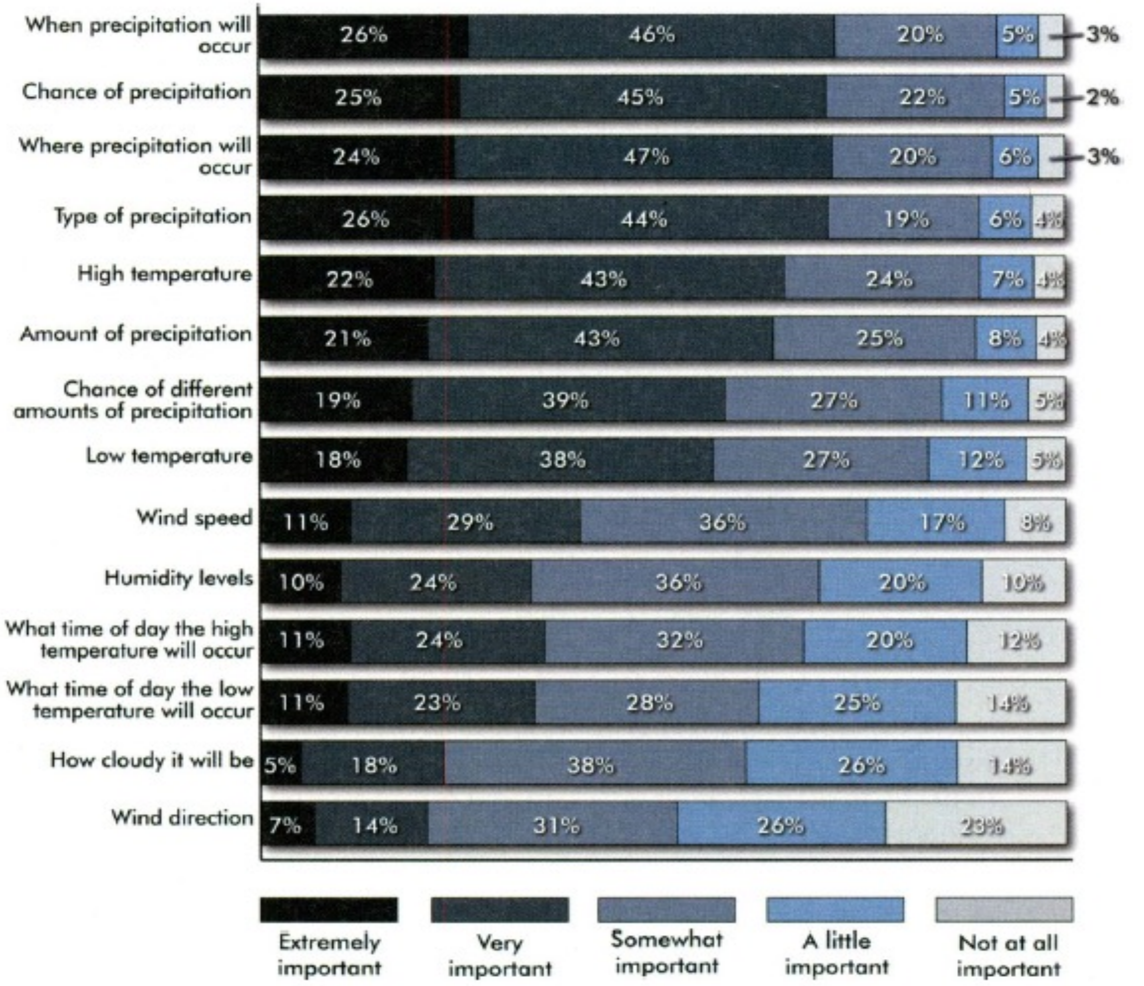
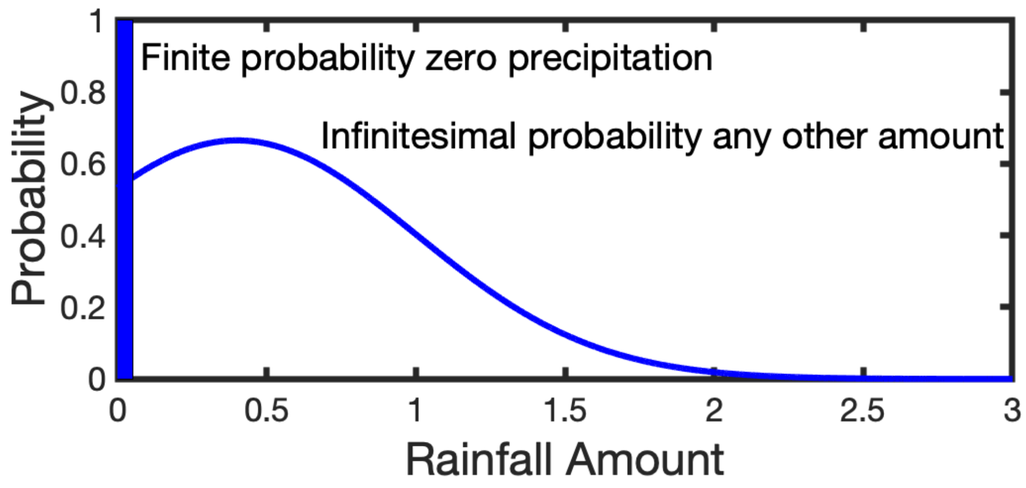


Fig. 5. Respondents' rating of importance for different potential components of weather forecasts (n = 1,465). The survey question asked "How important is it to you to have the information listed below as part of a weather forecast?"

MB
MB
MB

Mixed bounded (MB)



Forecast Distributions for Important Quantities

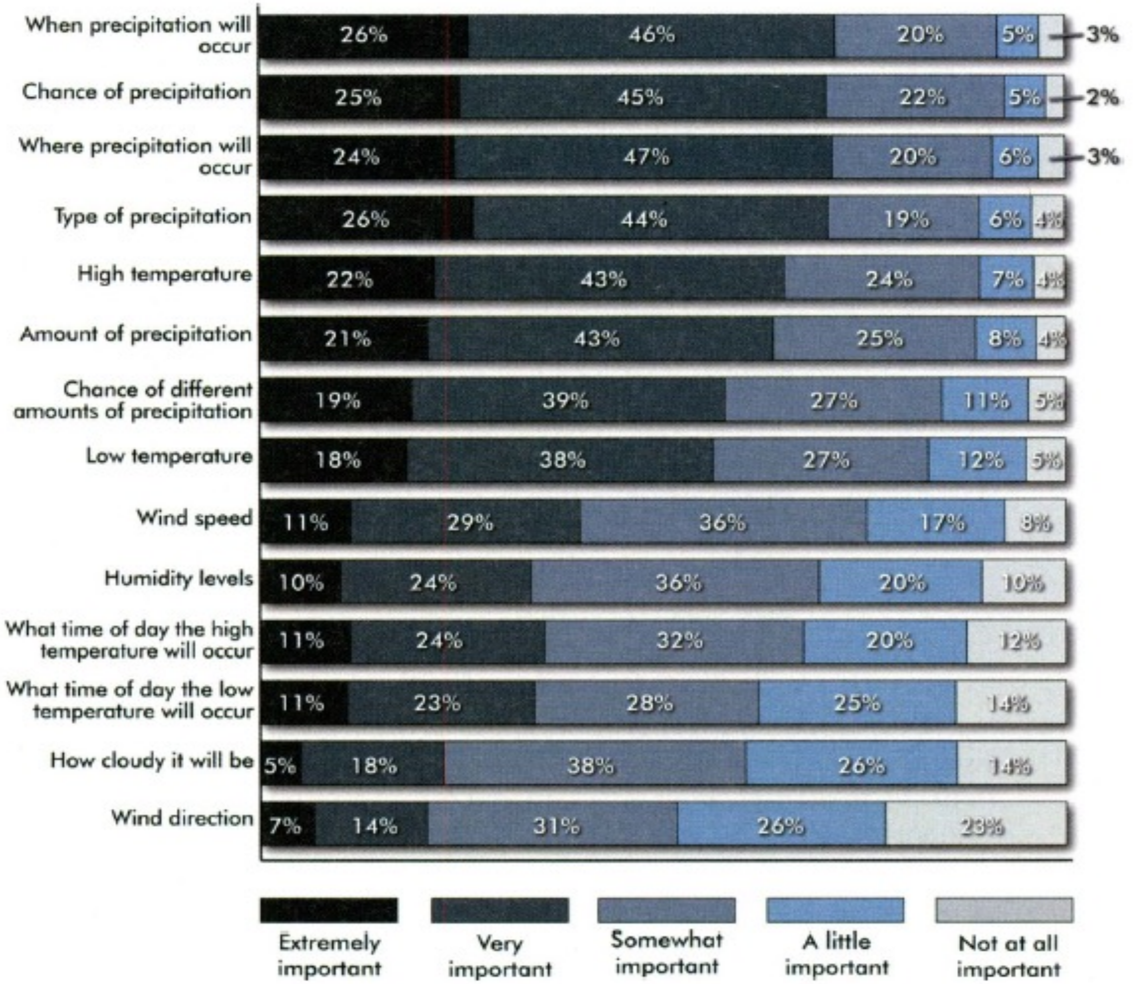
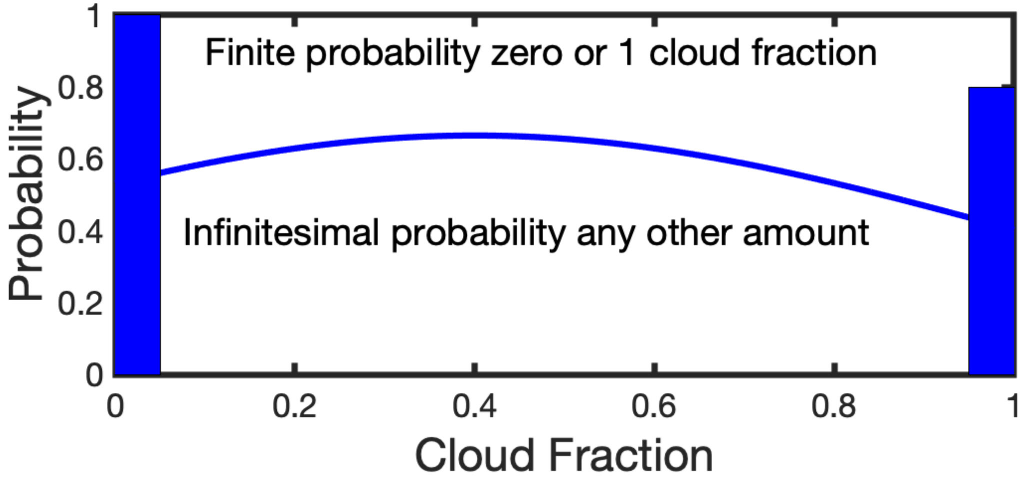


Fig. 5. Respondents' rating of importance for different potential components of weather forecasts (*n* = 1,465). The survey question asked "How important is it to you to have the information listed below as part of a weather forecast?"

MD

MD

Mixed doubly bounded (MD)



Forecast Distributions for Important Quantities

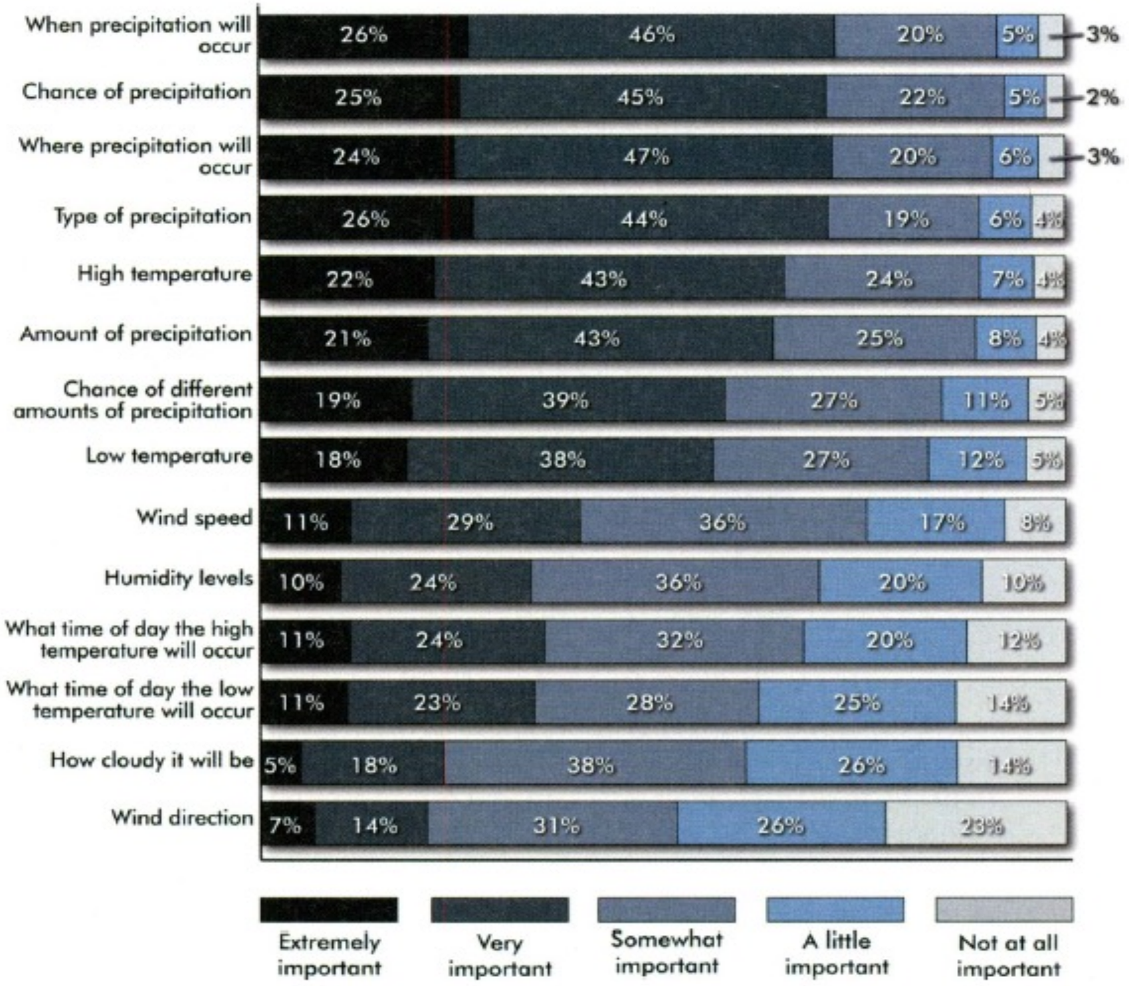


FIG. 5. Respondents' rating of importance for different potential components of weather forecasts (n = 1,465). The survey question asked "How important is it to you to have the information listed below as part of a weather forecast?"

N
D
N
MB
MB
N
MB
MD
N
N
MD
WN

Normal (N)

Wrapped Normal (WN)

Discrete (D)

Mixed bounded (MB)

Mixed doubly bounded (MD)