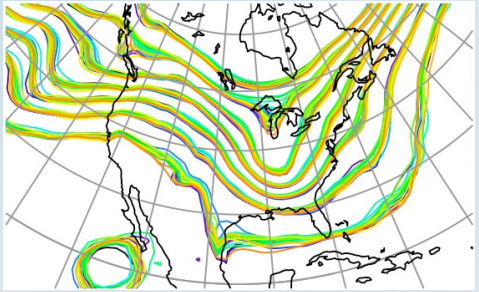


NCAR UCAR | CISL
Computational & Information Systems Lab



A General Nonlinear and Non-Gaussian Strongly Coupled Ensemble Data Assimilation Capability

Jeff Anderson representing NSF/NCAR Data Assimilation Research Section

Joint CP-TT/DA-TT Workshop, 28 October 2026

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.

DART Ensemble Filter Algorithm Implementation

1. Use model to advance ensemble members (3 here) to time at which next observations become available, and compute forward operators for all observations for each ensemble member.

Ensemble state estimate after using previous observation (analysis)

Extended ensemble state at time of next observation (prior)



DART Ensemble Filter Algorithm Implementation

1. Use model to advance ensemble members (3 here) to time at which next observations become available, and compute forward operators for all observations for each ensemble member.

Ensemble state estimate after using previous observation (analysis)

Extended ensemble state at time of next observation (prior)



Technical detail:

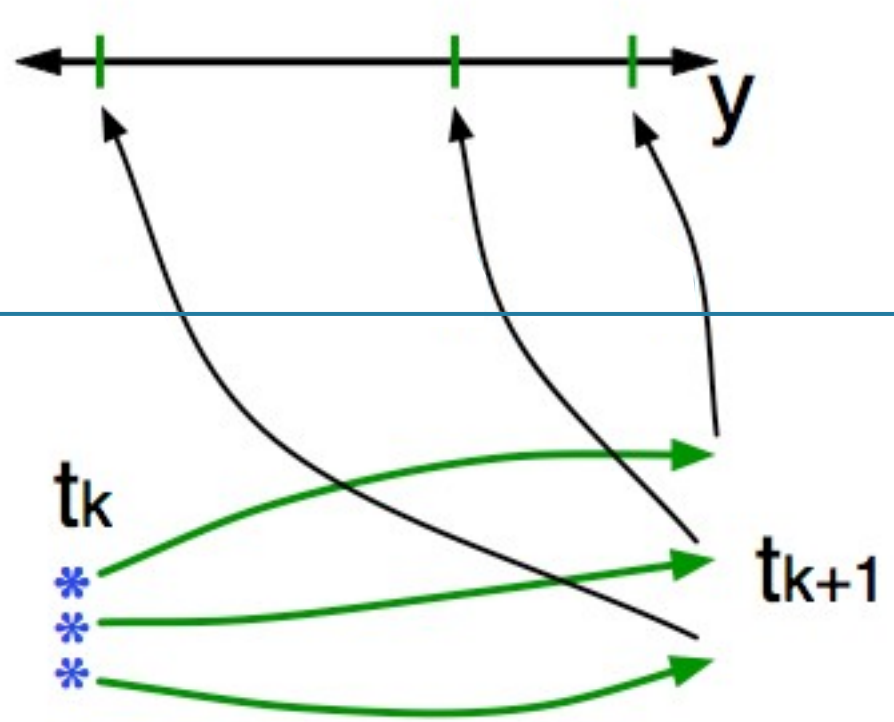
The extended state vector is the state vector plus an extra element for each observation.

If the state vector has M variables, and there are Q observations, the extended state vector has M+Q elements.

More on this in a while.

DART Ensemble Filter Algorithm Implementation

2. ***Get*** the ensemble of values of the first observation to be assimilated at this time (observation is of an extended state variable).



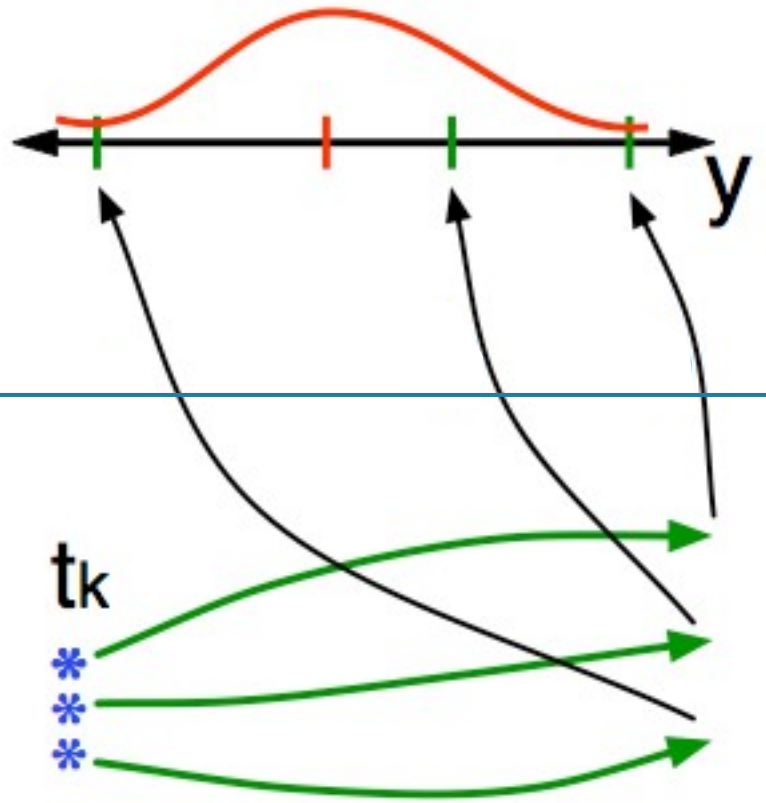
Theory: observations from instruments with uncorrelated errors can be done sequentially.

Scalar Observation Space

Model Space

DART Ensemble Filter Algorithm Implementation

3. Get observed value and likelihood from observing system.

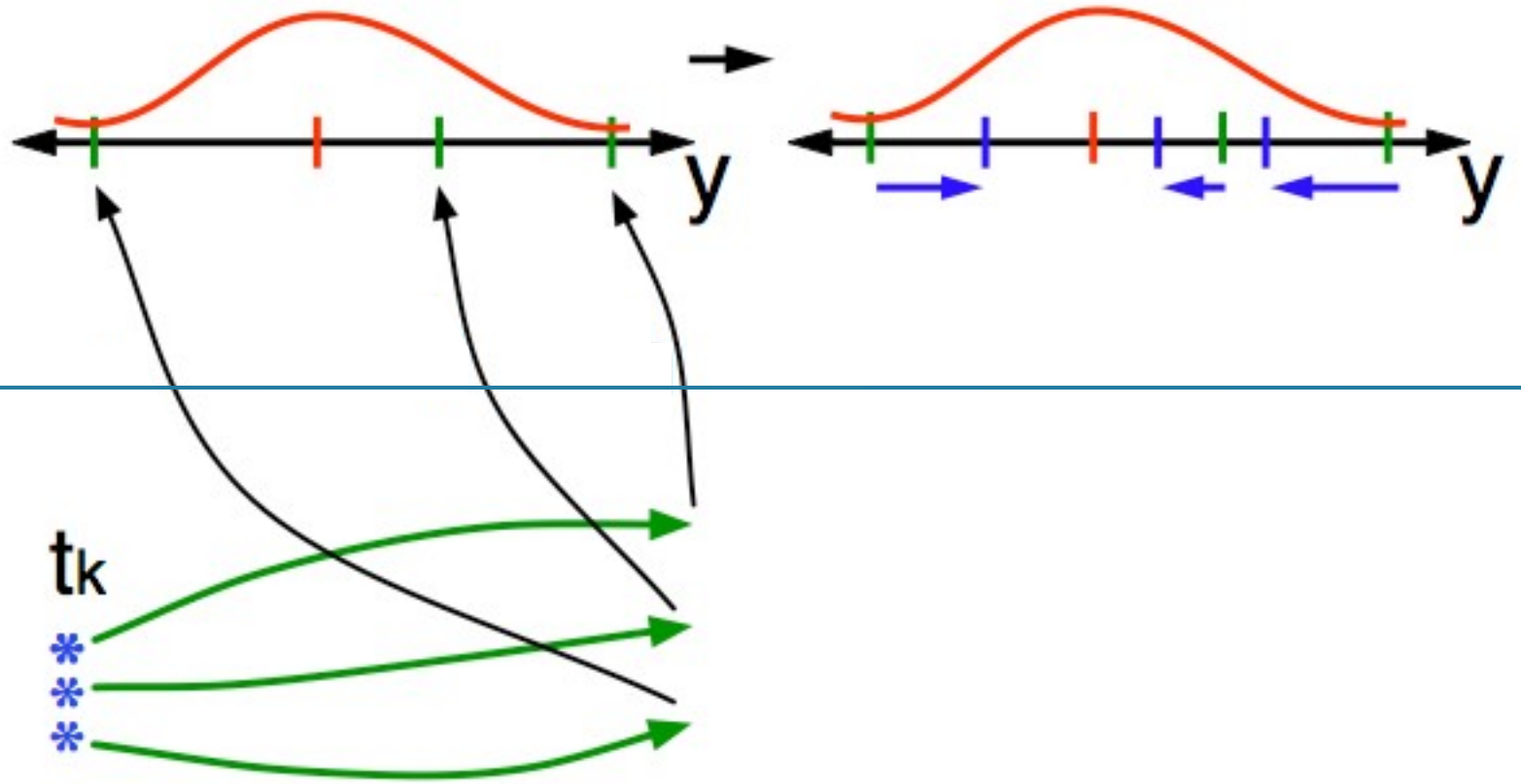


Scalar Observation Space

Model Space

DART Ensemble Filter Algorithm Implementation

4. Find the increments for the prior ensemble of observations, this is a scalar problem.

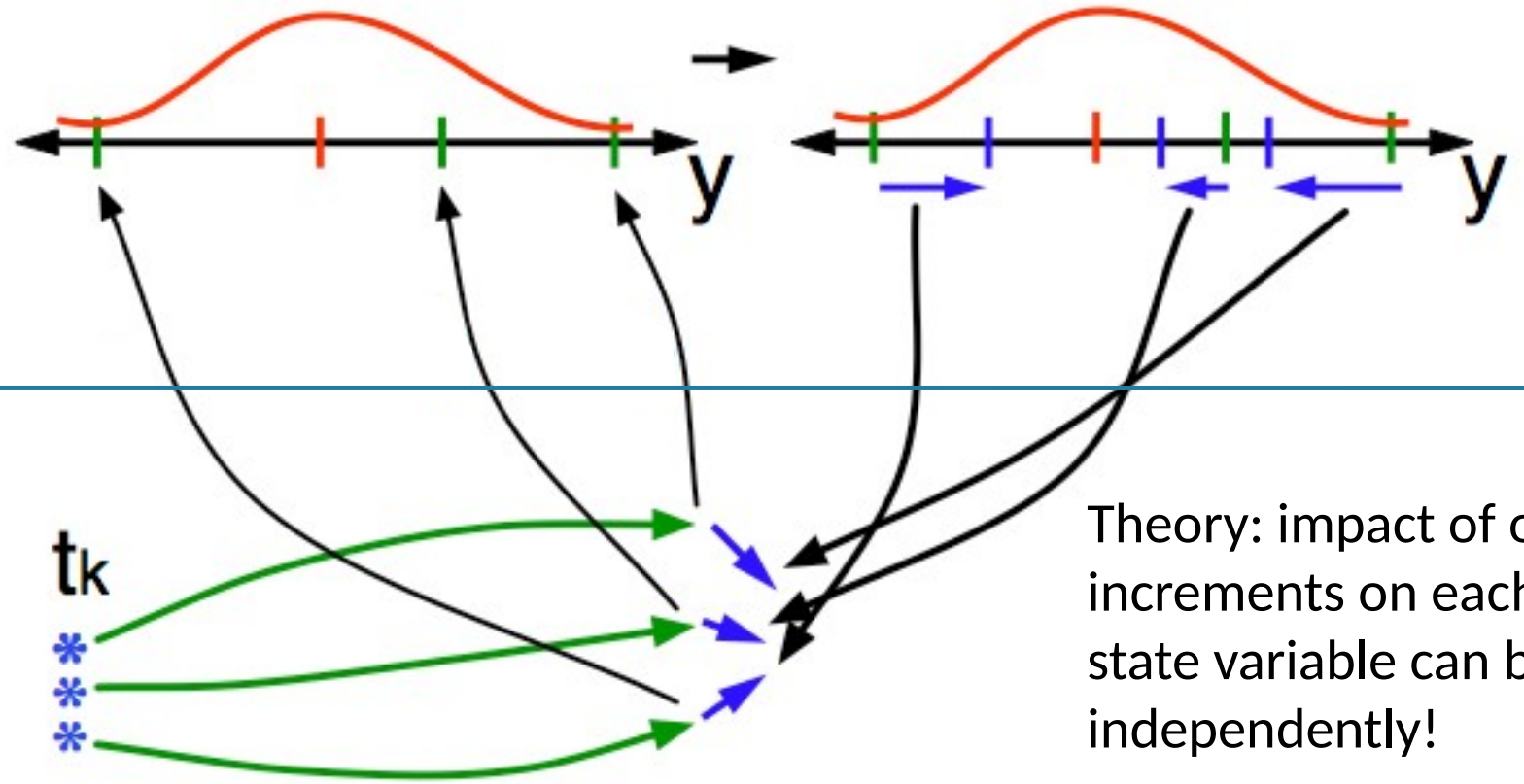


Scalar Observation Space

Model Space

DART Ensemble Filter Algorithm Implementation

5. Use ensemble samples of $\mathbf{y}$ and each state variable to linearly regress observation increments onto extended state variable increments.



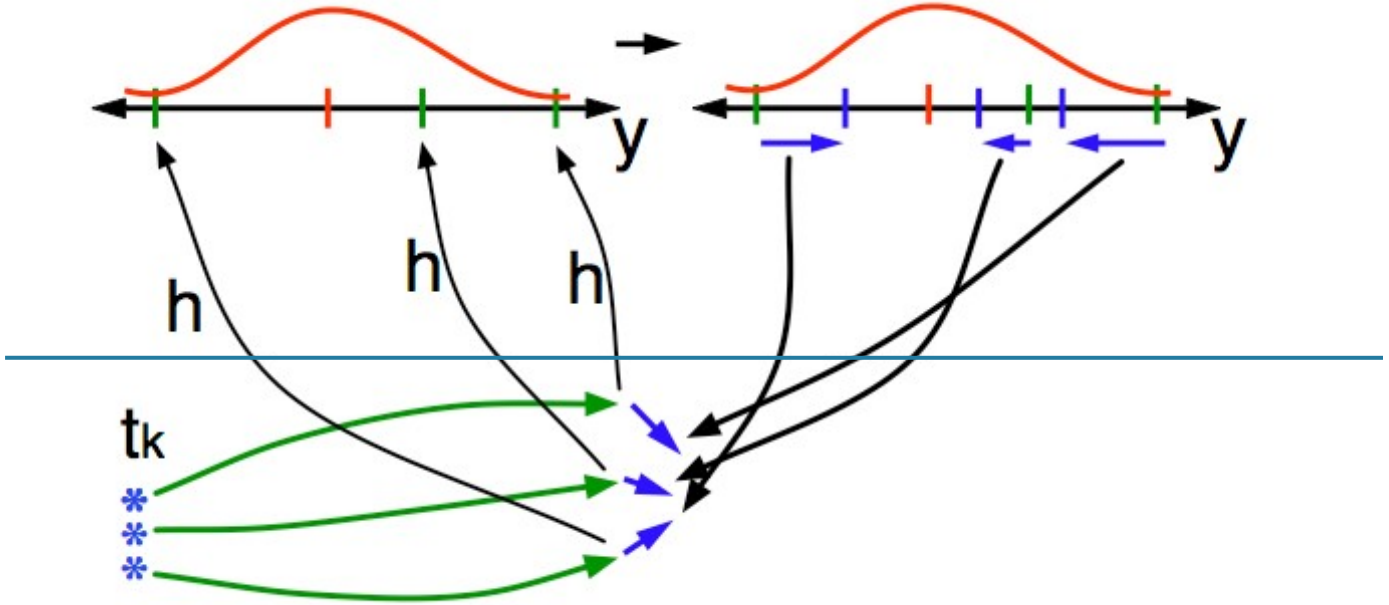
Scalar Observation Space

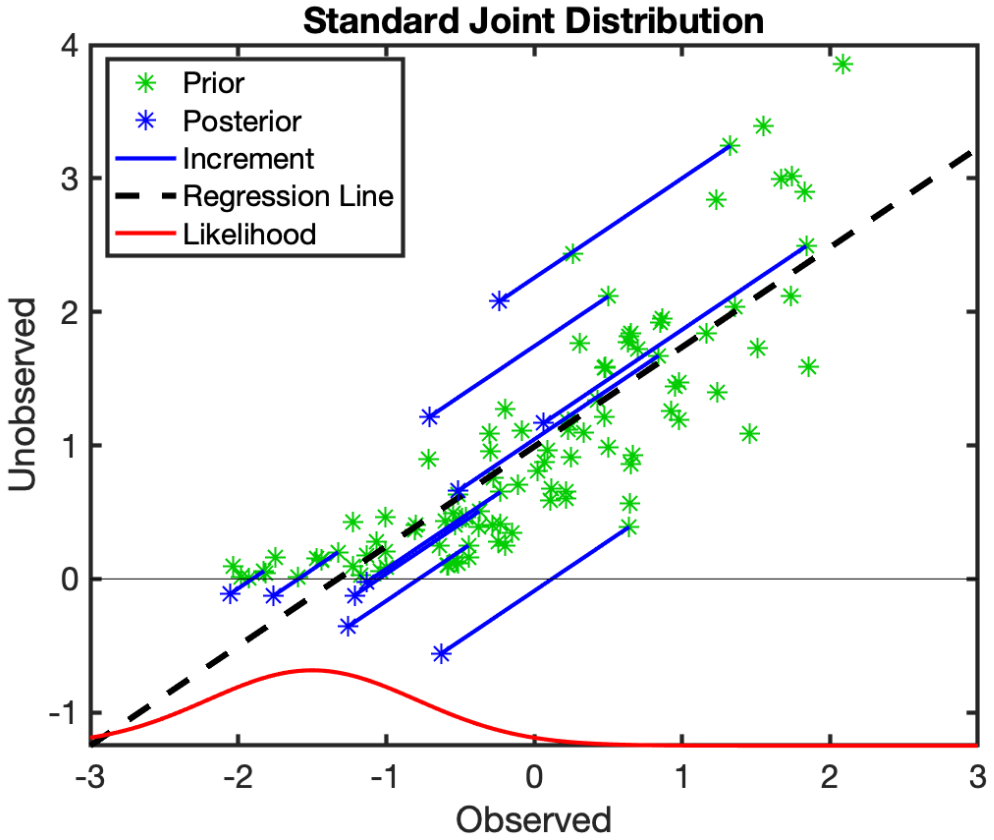
Model Space

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

DART Ensemble Filter Algorithm Implementation

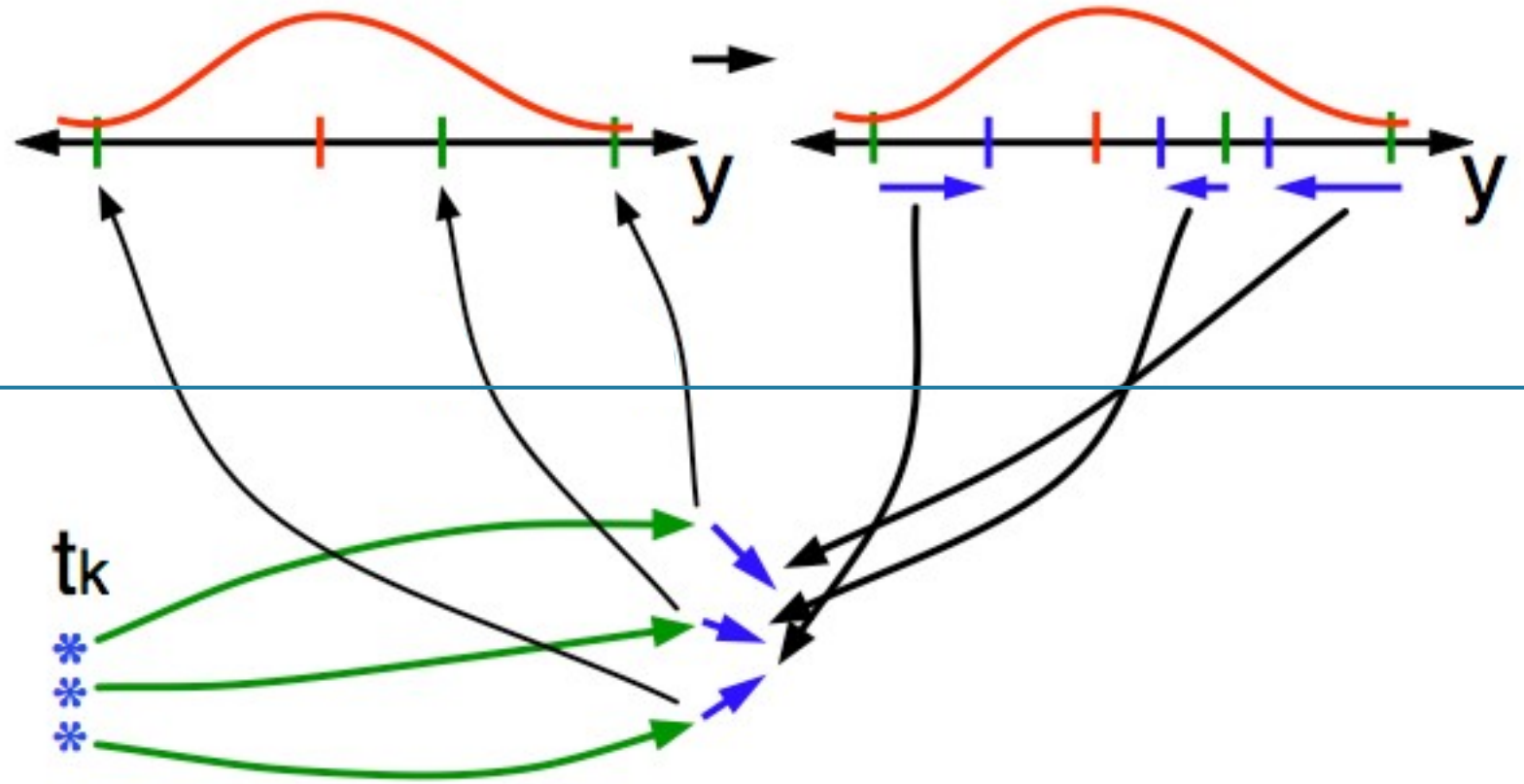
5. Use ensemble samples of $\mathbf{y}$ and each state variable to ***linearly regress observation increments*** onto extended state variable increments.





DART Ensemble Filter Algorithm Implementation

Repeat steps 2-5 sequentially for each observation at this time.

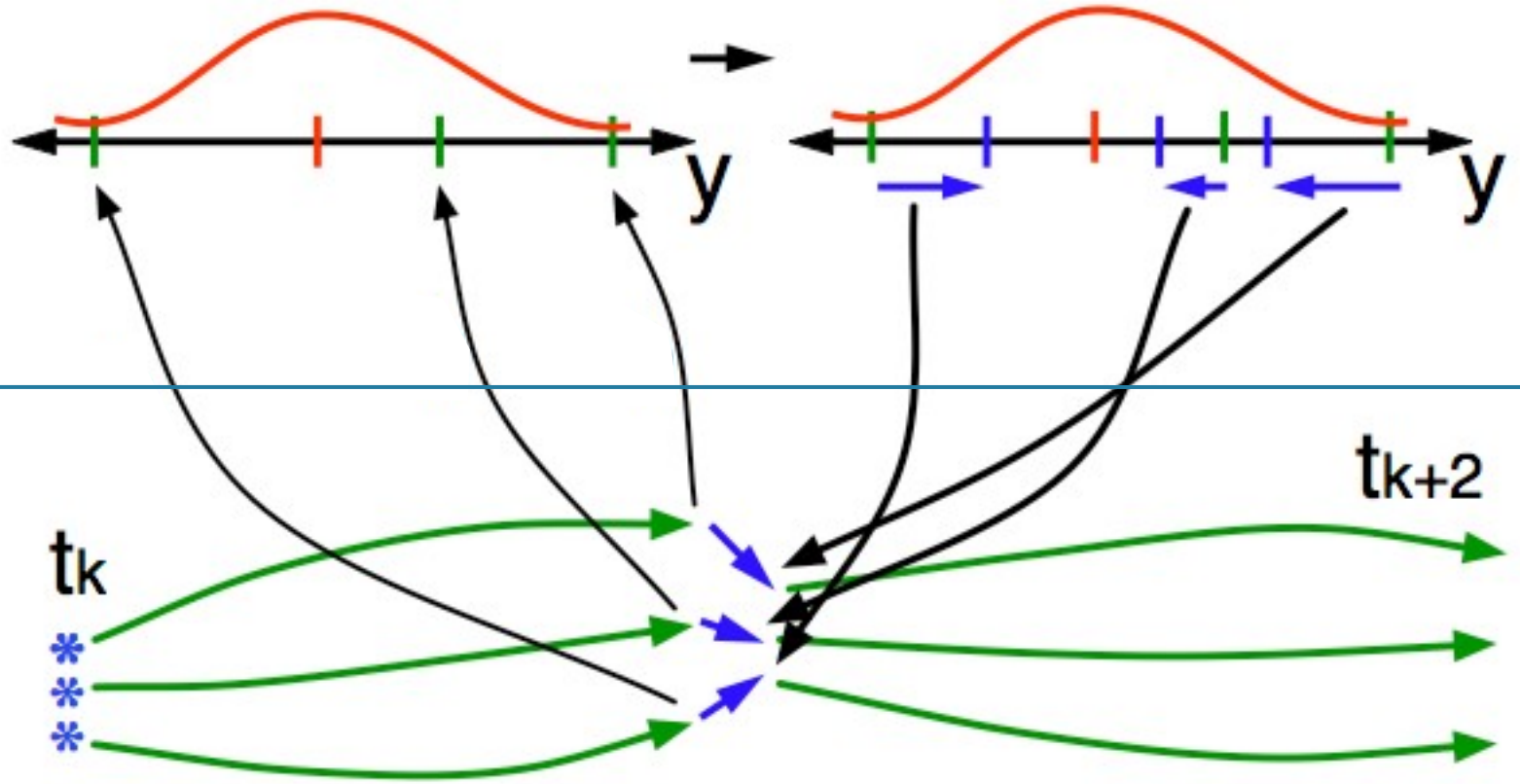


Scalar Observation Space

Model Space

DART Ensemble Filter Algorithm Implementation

6. When all observations are assimilated, integrate model state to the next time that has observations.

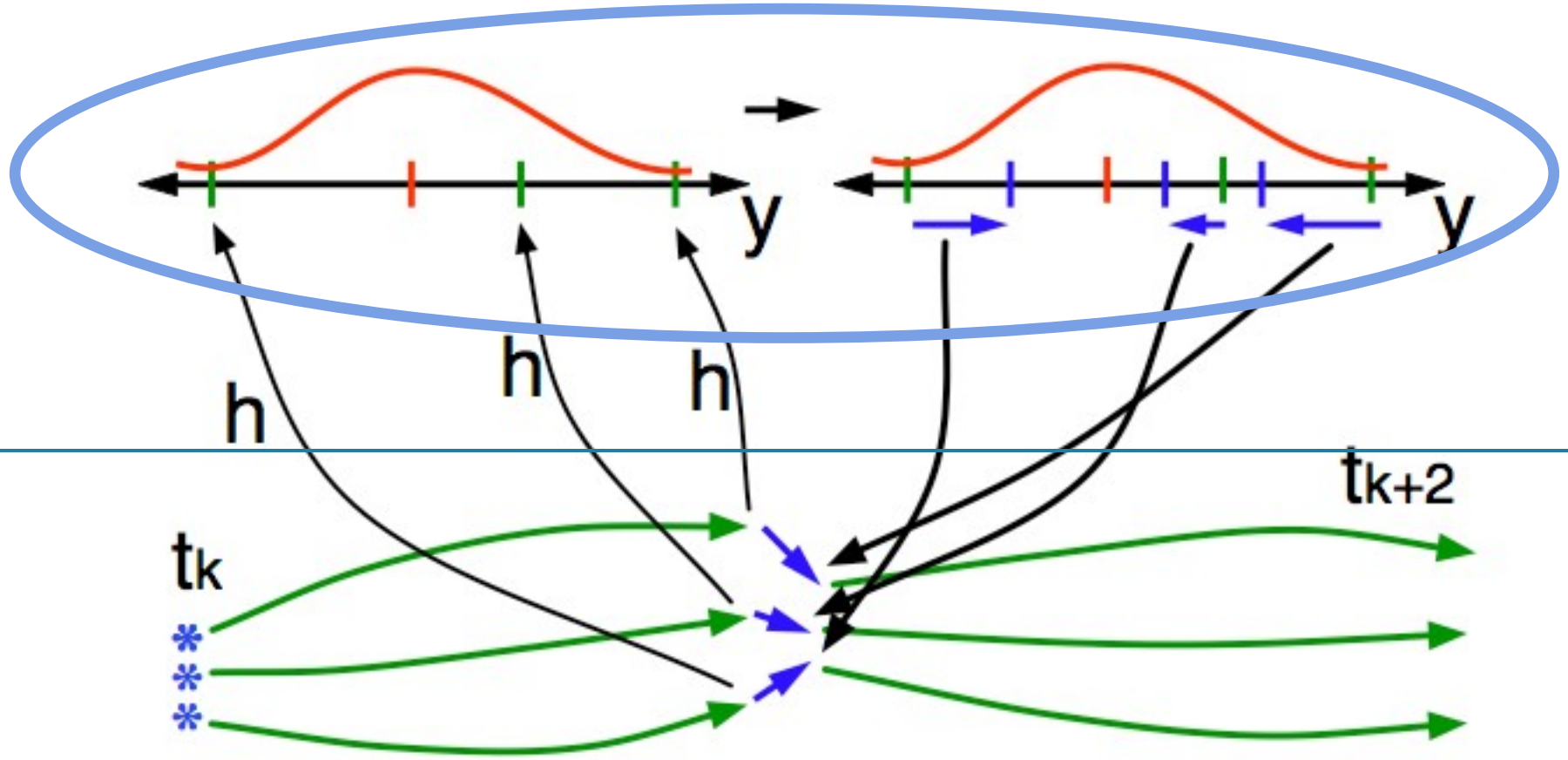


Scalar Observation Space

Model Space

Ensemble Kalman Filter Step 1: Observation Increments

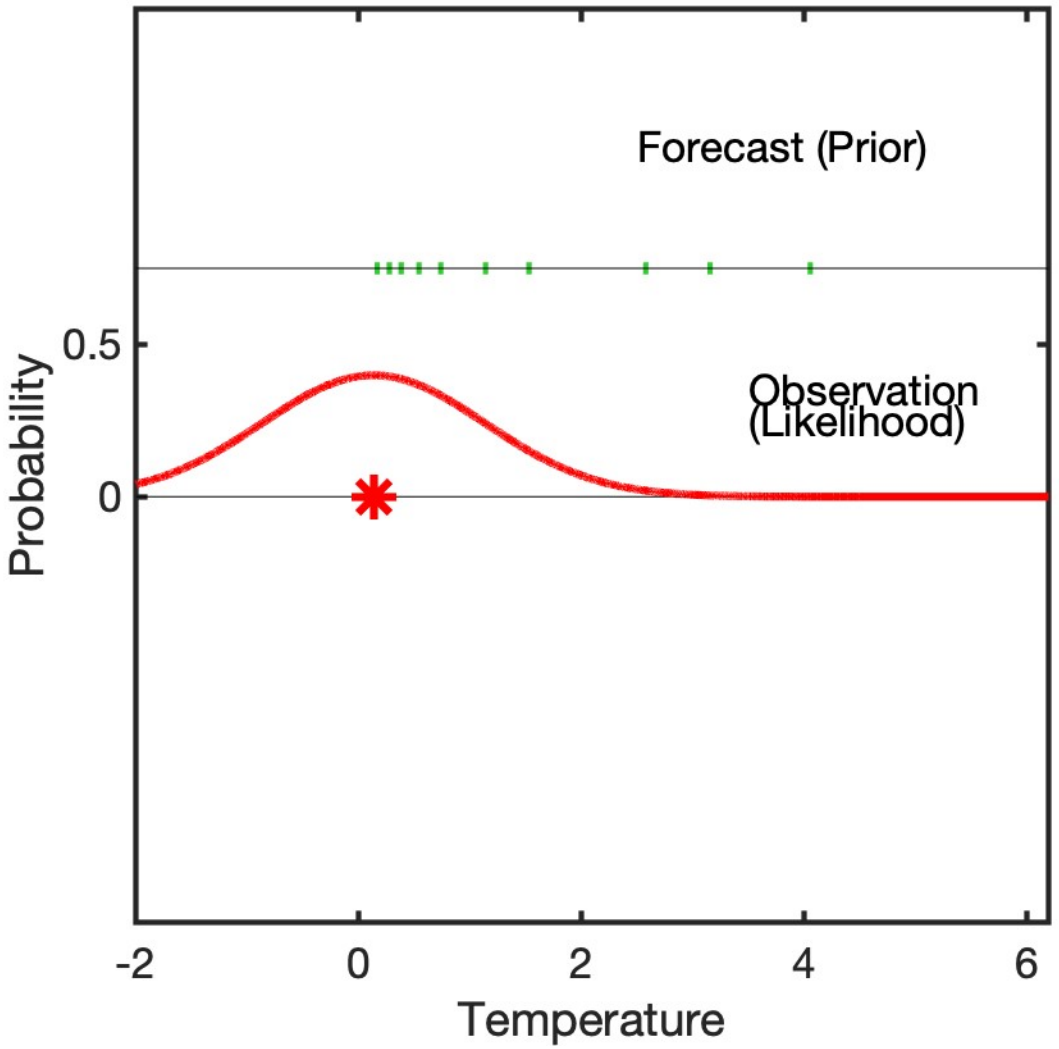
A nearly general solution for the observation space step:
(Anderson, 2022, MWR *150*, 1061-1074).



The Mesa Lab: Weather can Impact the Commute



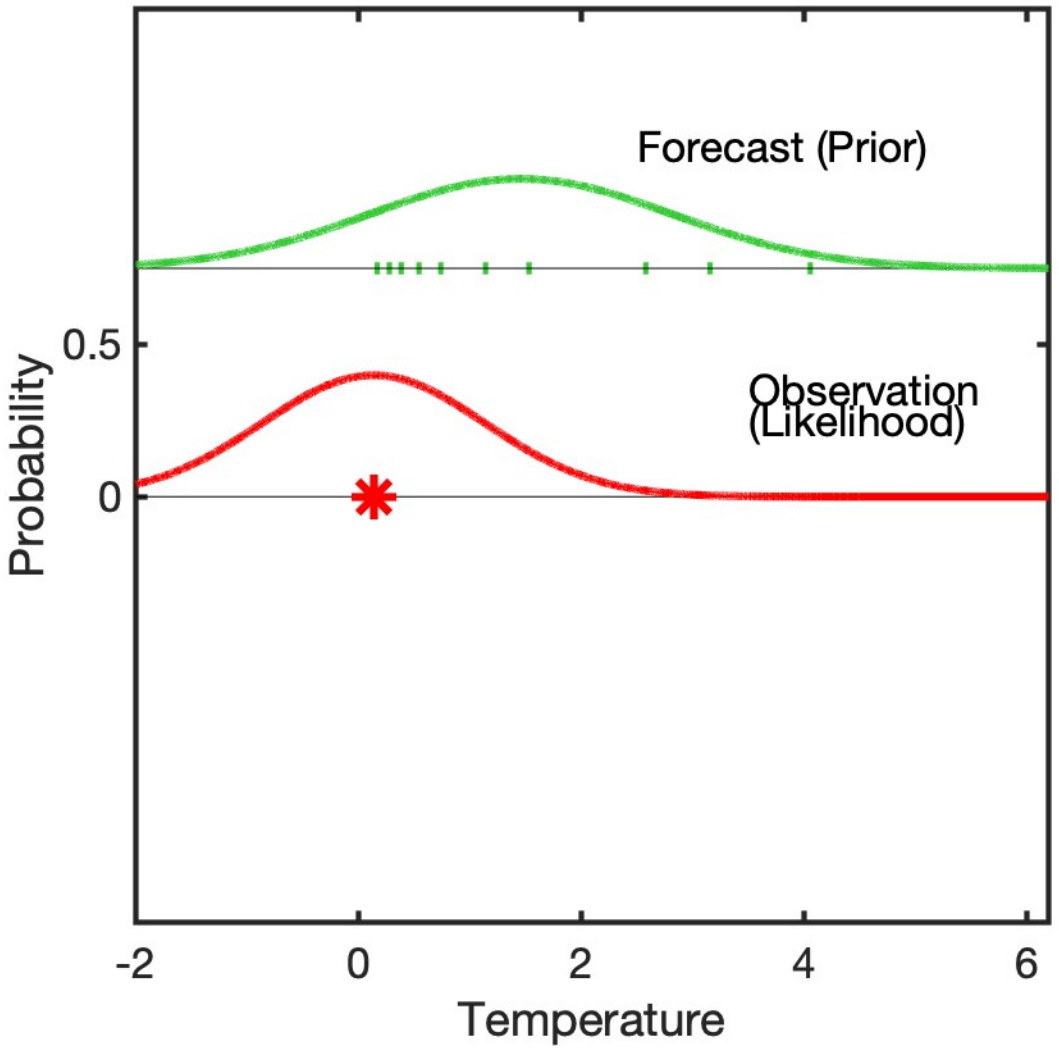
Should I Worry About Ice Going Down the Hill?



Have 10 forecasts of NCAR temperature.

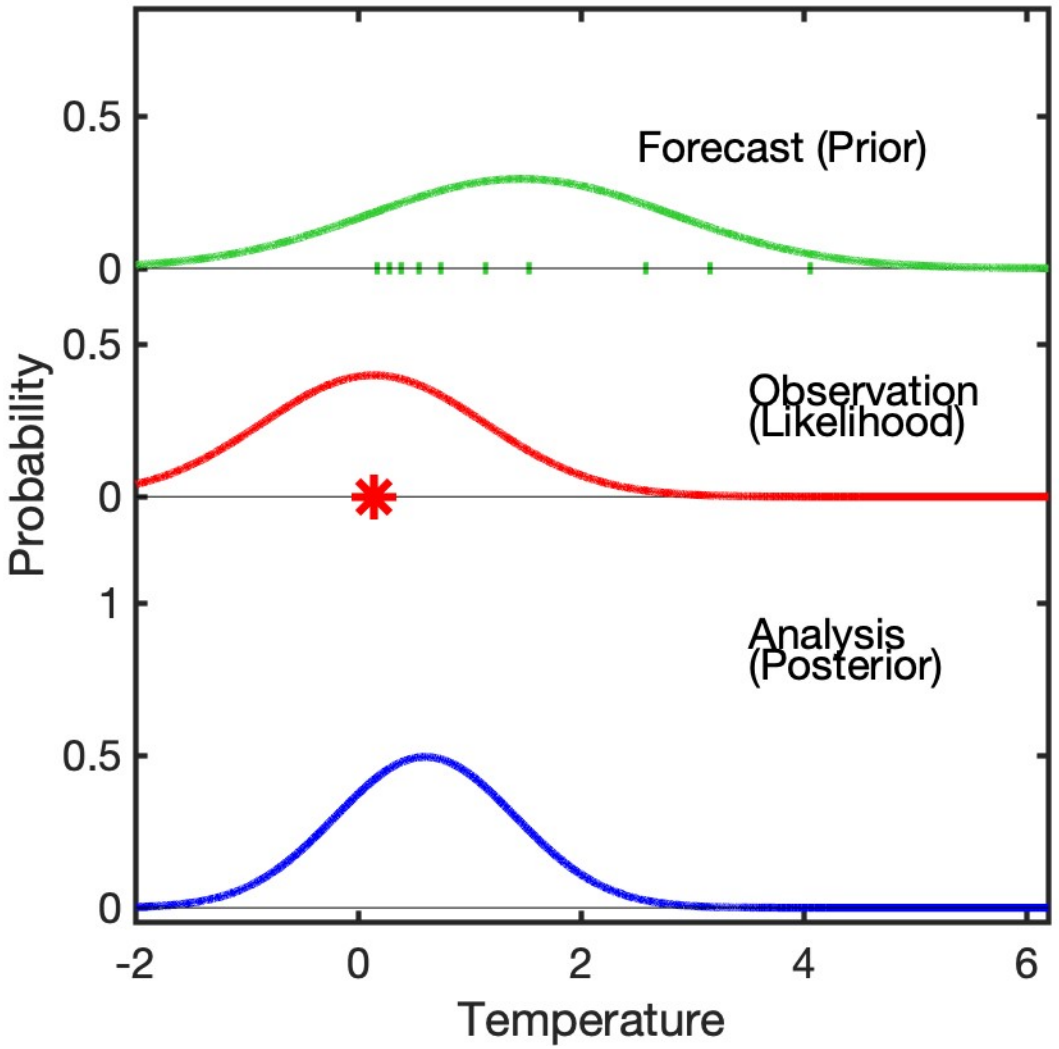
Use Bayes to combine with uncertain NCAR temperature observation.

Should I Worry About Ice Going Down the Hill?



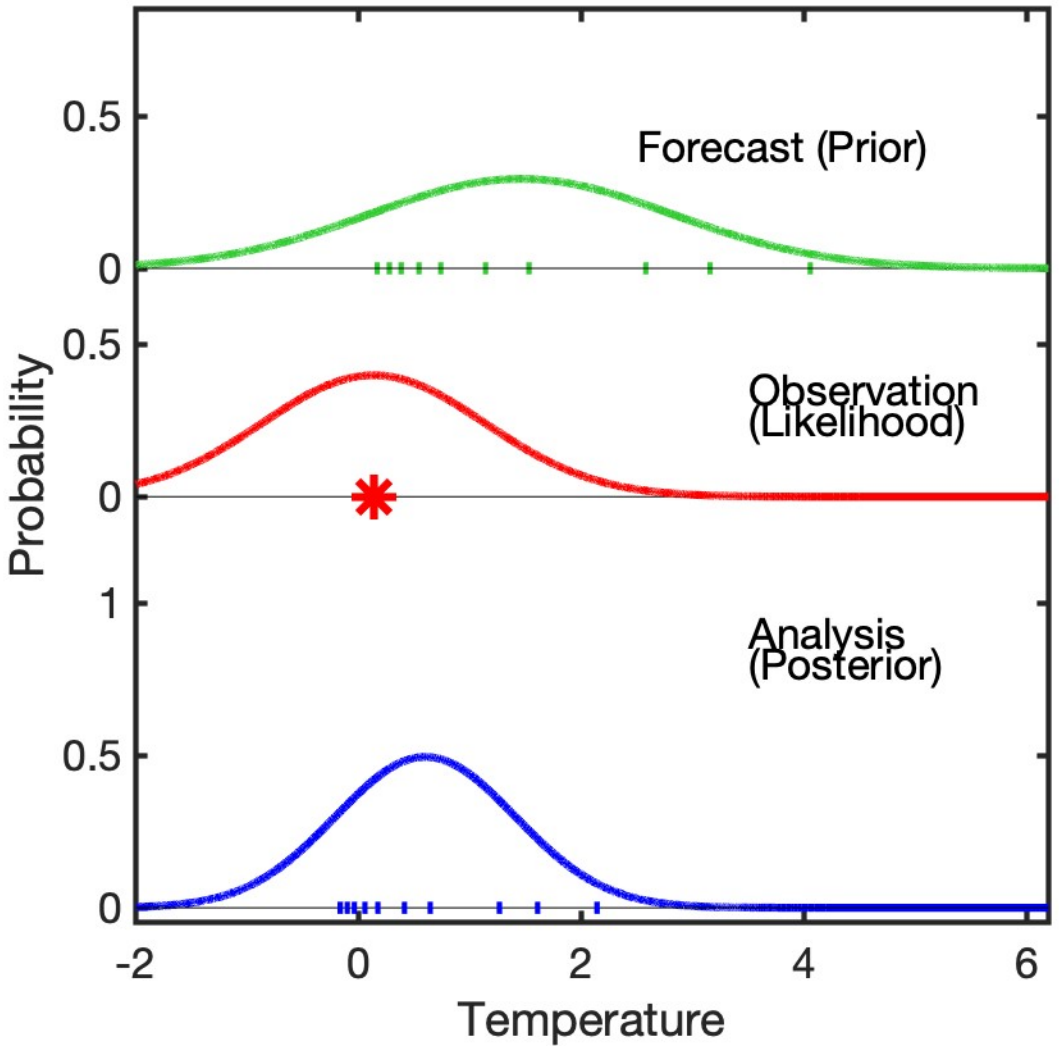
Standard Ensemble Filter:
Fit a normal to the forecast ensemble.

Should I Worry About Ice Going Down the Hill?



Bayes product gives continuous normal posterior.

Should I Worry About Ice Going Down the Hill?

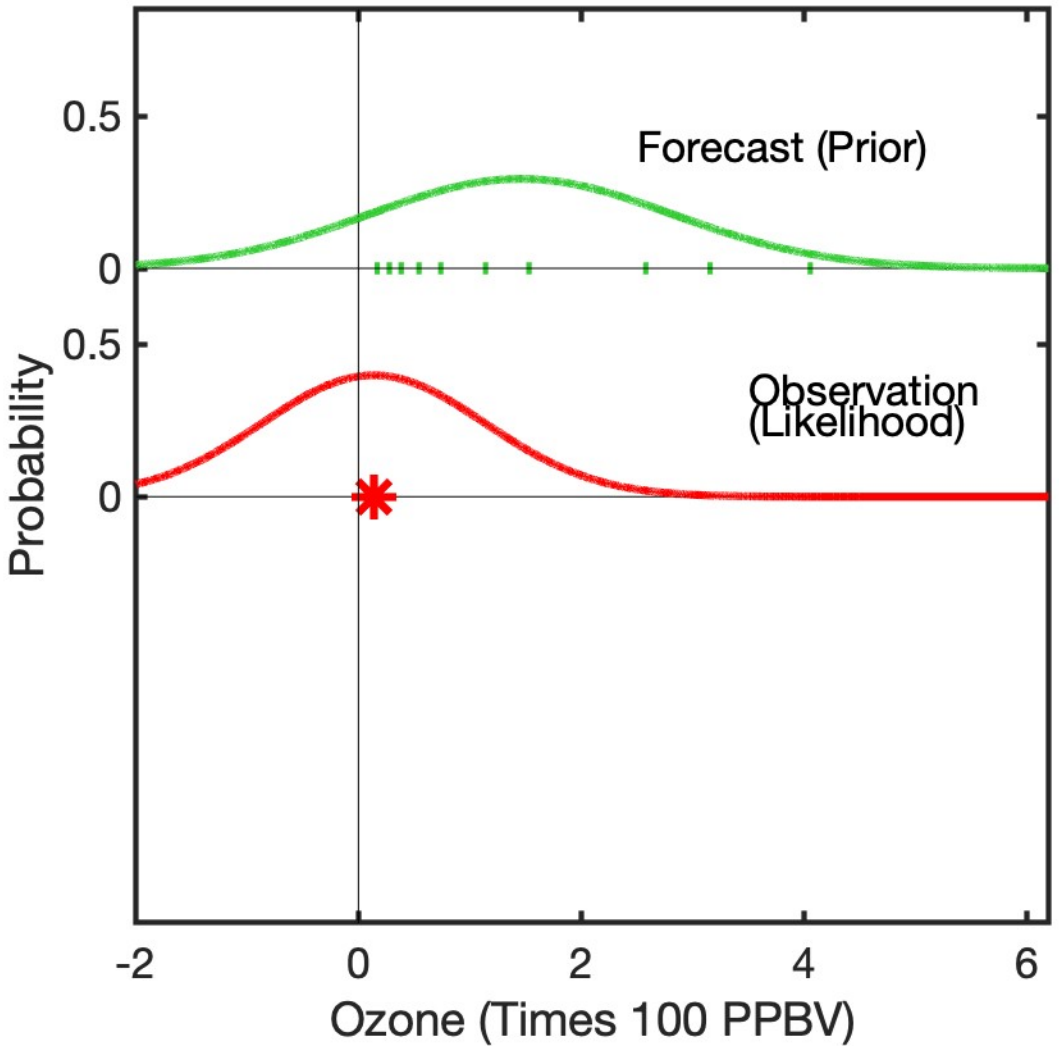


Get a posterior ensemble.

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?

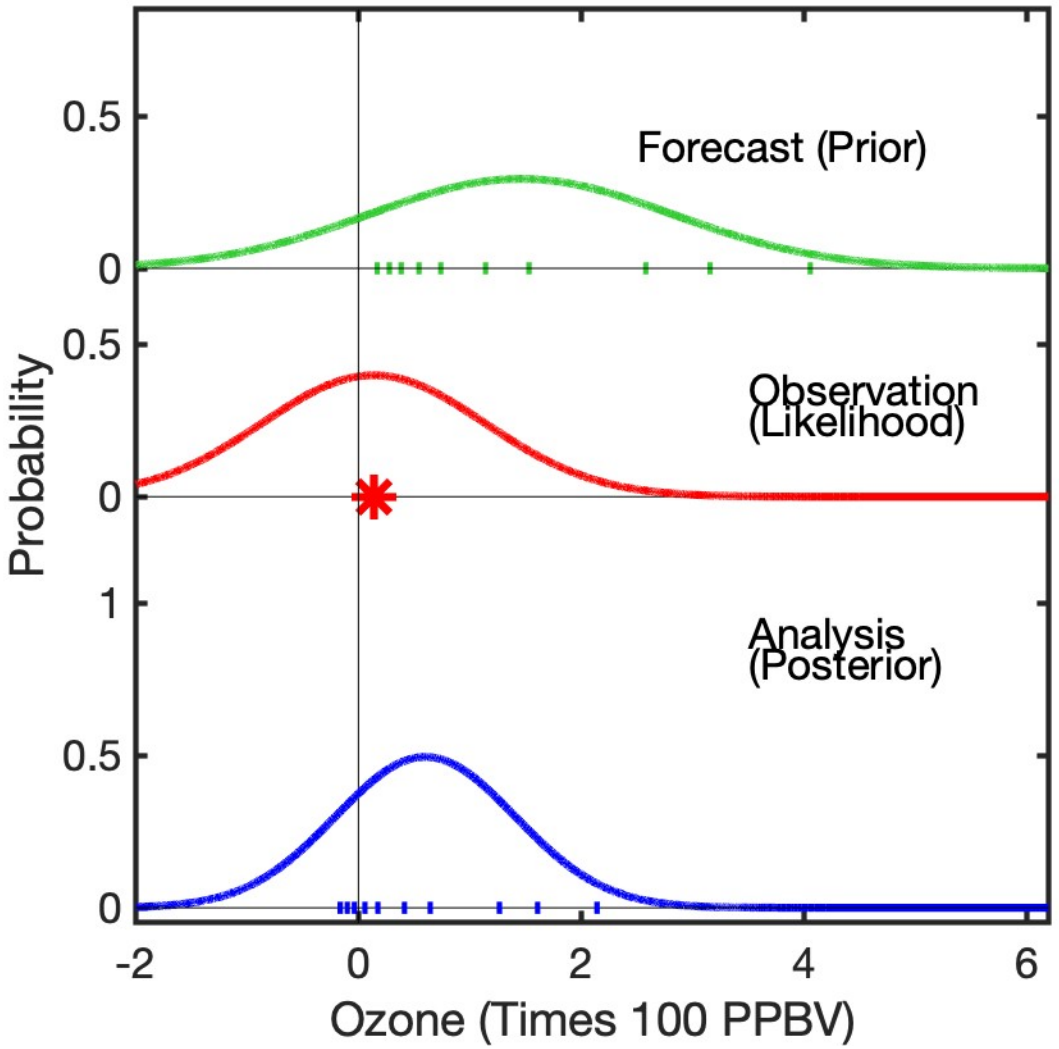


Forecast model knows ozone must be positive.

Fitting a normal leads to probability of negative.



Should I Worry About Air Quality Going Down the Hill?

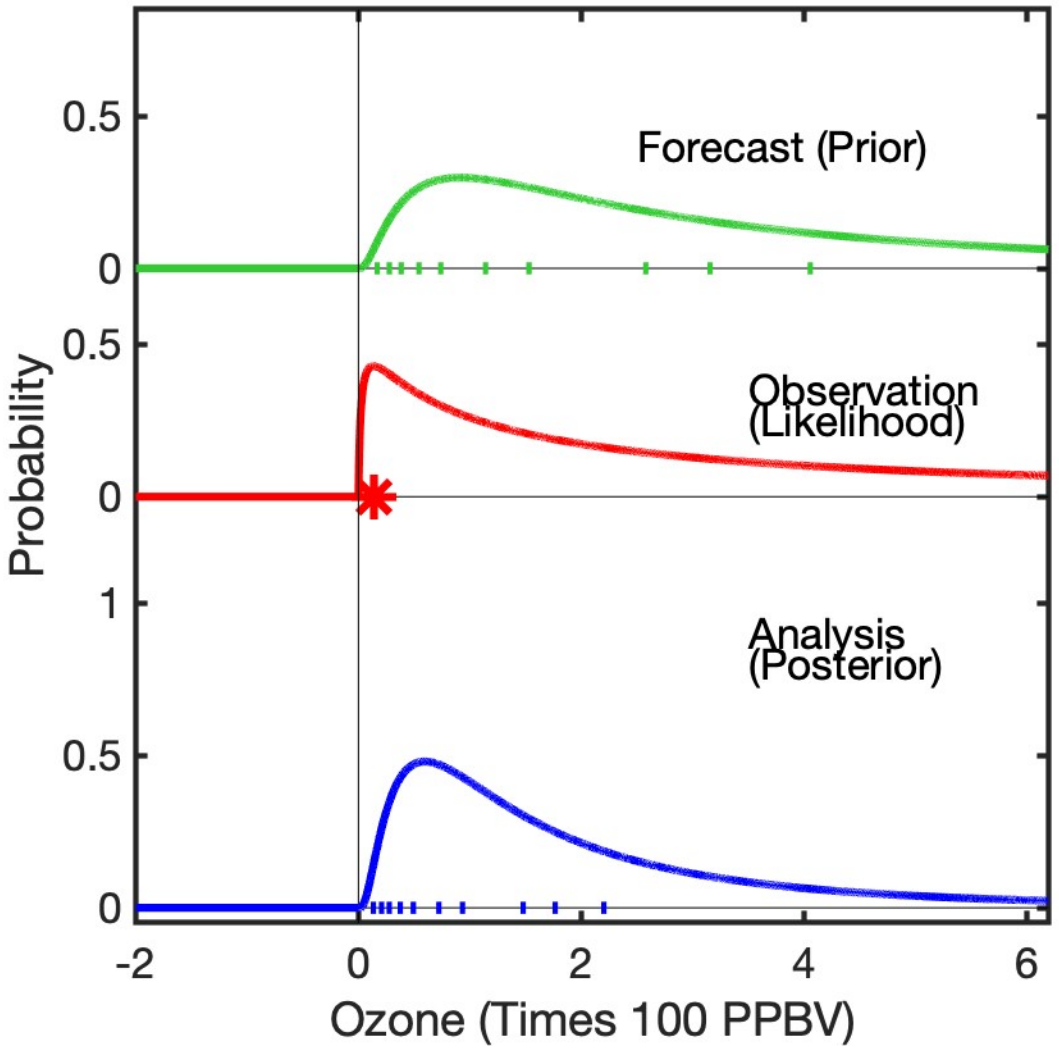


Doing the DA can lead to negative ensemble members.

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?

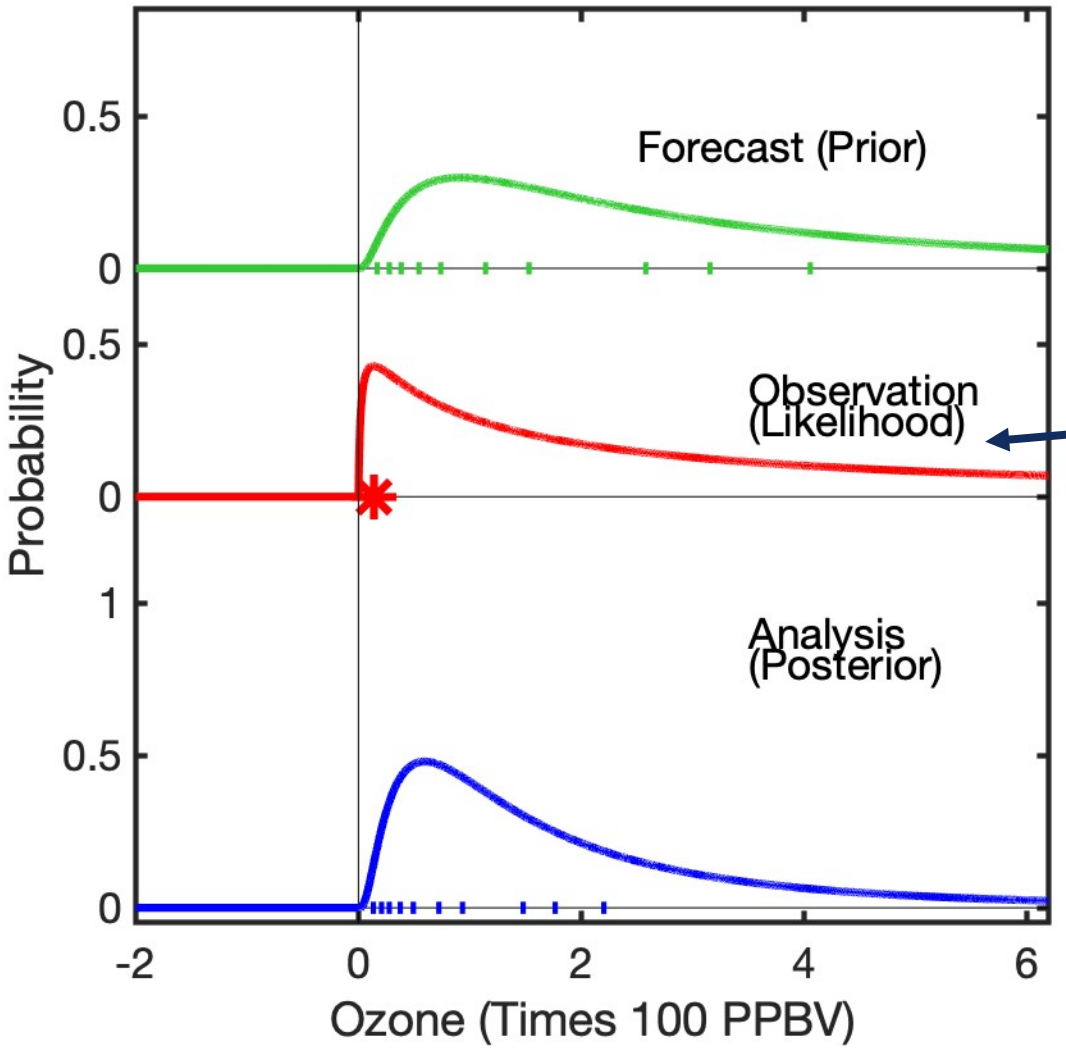


Now can do any distribution using quantile conserving ensemble algorithms.

Example: Gamma for bounded quantity like ozone.

Posterior ensemble no longer crazy.

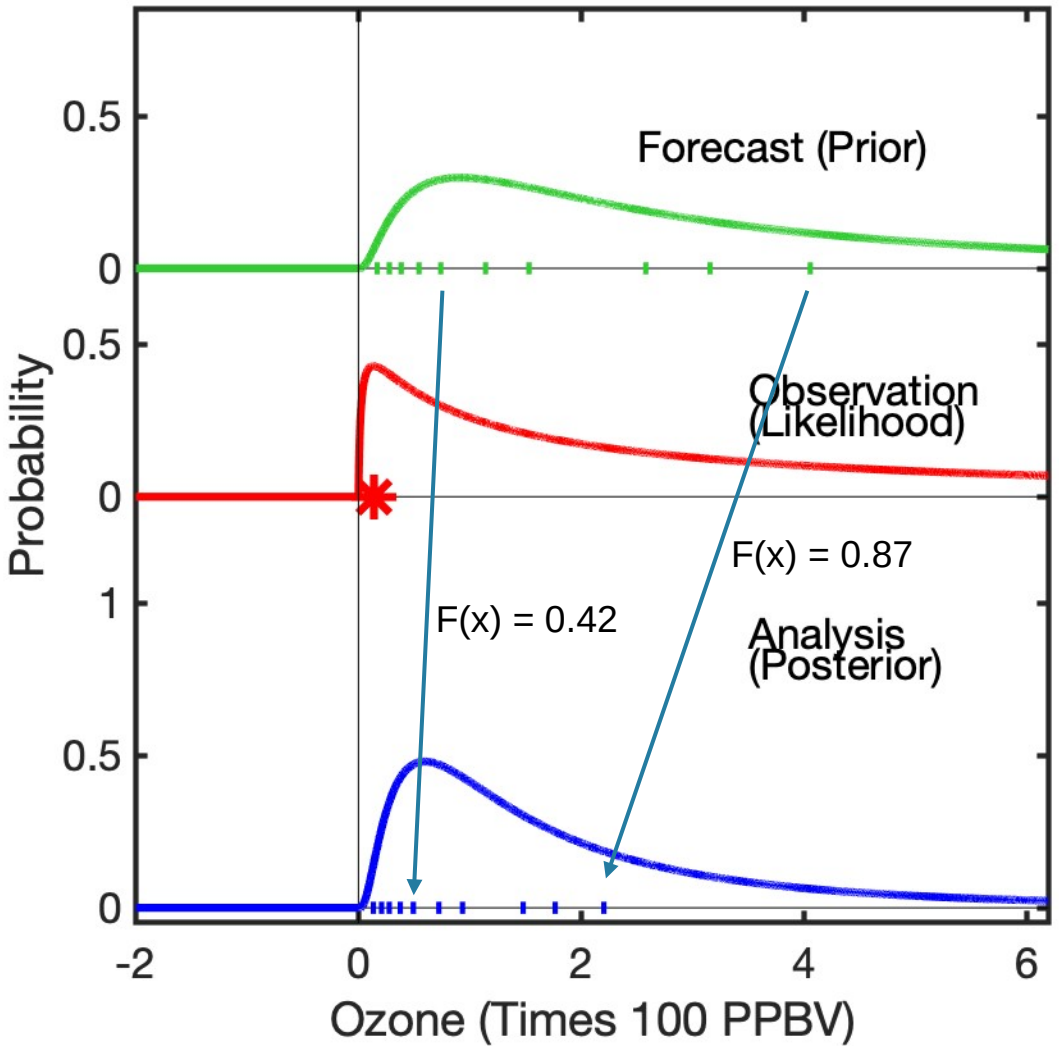
Should I Worry About Air Quality Going Down the Hill?



Now can do any distribution using quantile conserving ensemble algorithms.

Can now use much more general information about observation error from instrument experts. Nice collaborations are possible.

Quantile Conserving Ensemble Filter Framework



How to select ensembles for the analysis distribution?

Conserve quantiles from the prior ensemble.

F is the cumulative distribution function for the prior or analysis continuous distribution. It gives the quantile.

What if We Don’t Know the Right Distribution?

Can use any distribution family for prior in observation space.

gamma, inverse gamma, log normal, beta …

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

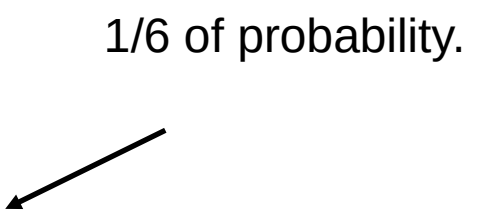
Use non-parametric continuous prior PDF.

Rank Histogram piecewise constant distribution.

The Bounded Normal Rank Histogram Distribution

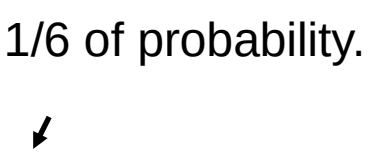
Have a prior ensemble for a state variable (like wind).

The Bounded Normal Rank Histogram Distribution



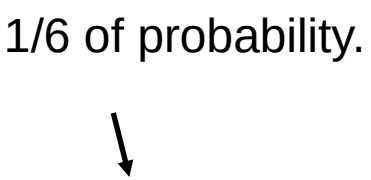
- Place $(ens_size + 1)^{-1}$ mass between adjacent ensemble members.
- Reminiscent of rank histogram evaluation method.

The Bounded Normal Rank Histogram Distribution



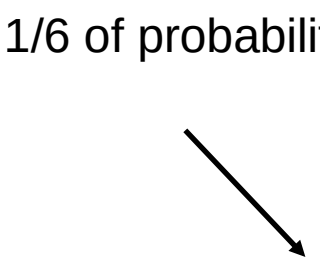
- Place $(ens_size + 1)^{-1}$ mass between adjacent ensemble members.
- Reminiscent of rank histogram evaluation method.

The Bounded Normal Rank Histogram Distribution



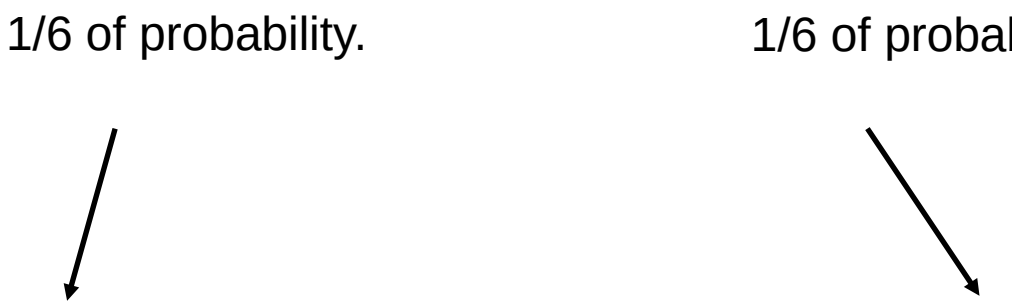
- Place $(ens_size + 1)^{-1}$ mass between adjacent ensemble members.
- Reminiscent of rank histogram evaluation method.

The Bounded Normal Rank Histogram Distribution



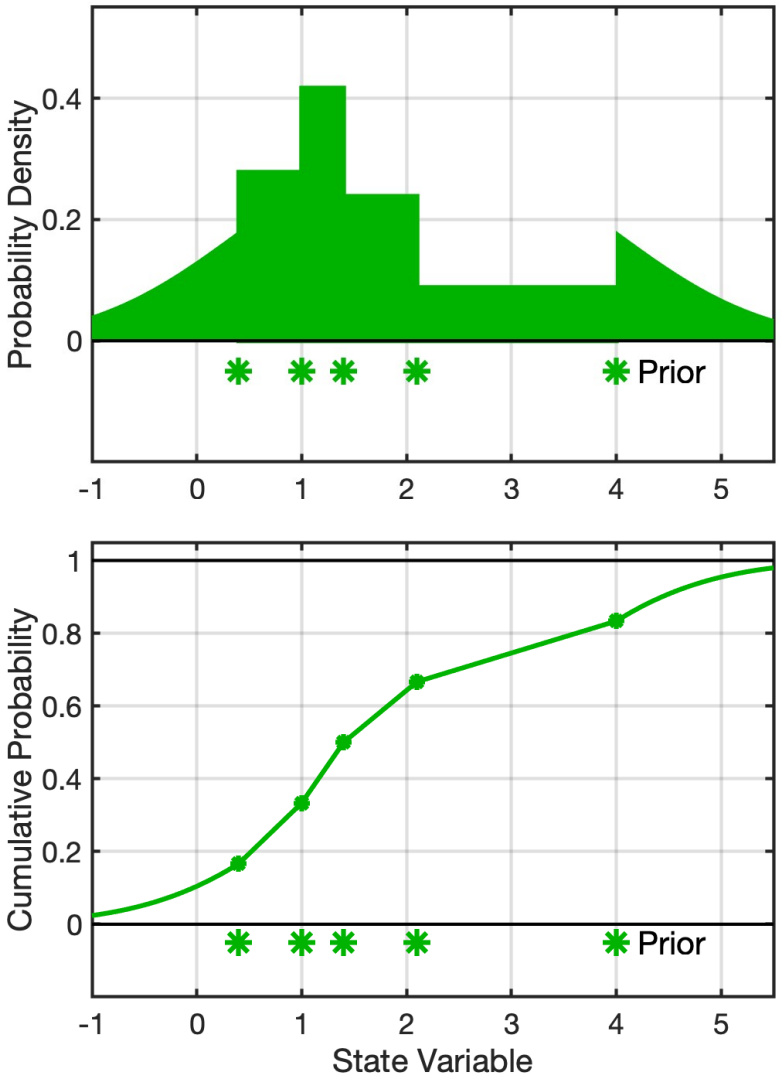
- Place $(ens_size + 1)^{-1}$ mass between adjacent ensemble members.
- Reminiscent of rank histogram evaluation method.

The Bounded Normal Rank Histogram Distribution



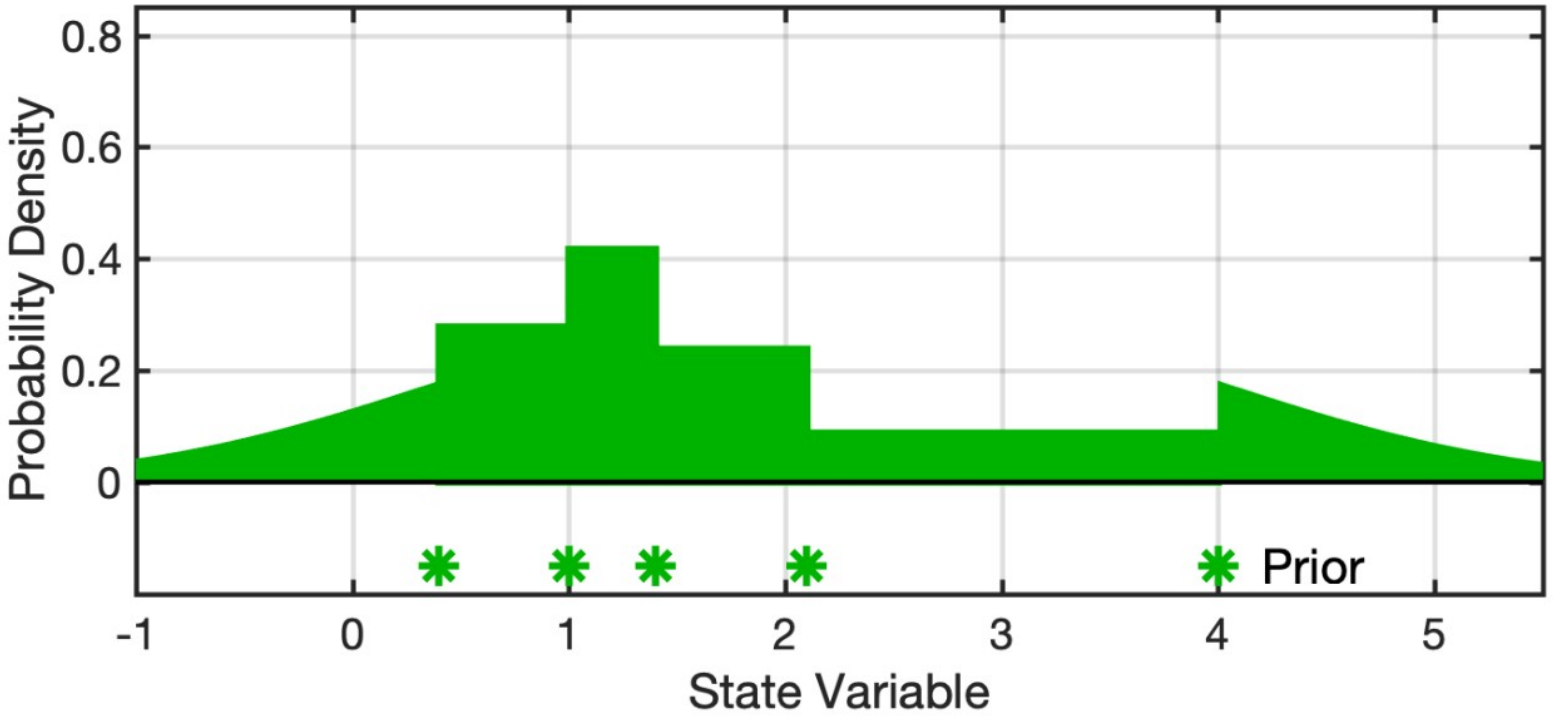
- Partial gaussian kernels on tails, N(*tail_mean*, *ens_sd*).
- *tail_mean selected so that (ens_size + 1)$^{-1}$ mass is in tail*.

Rank Histogram Prior PDF and CDF



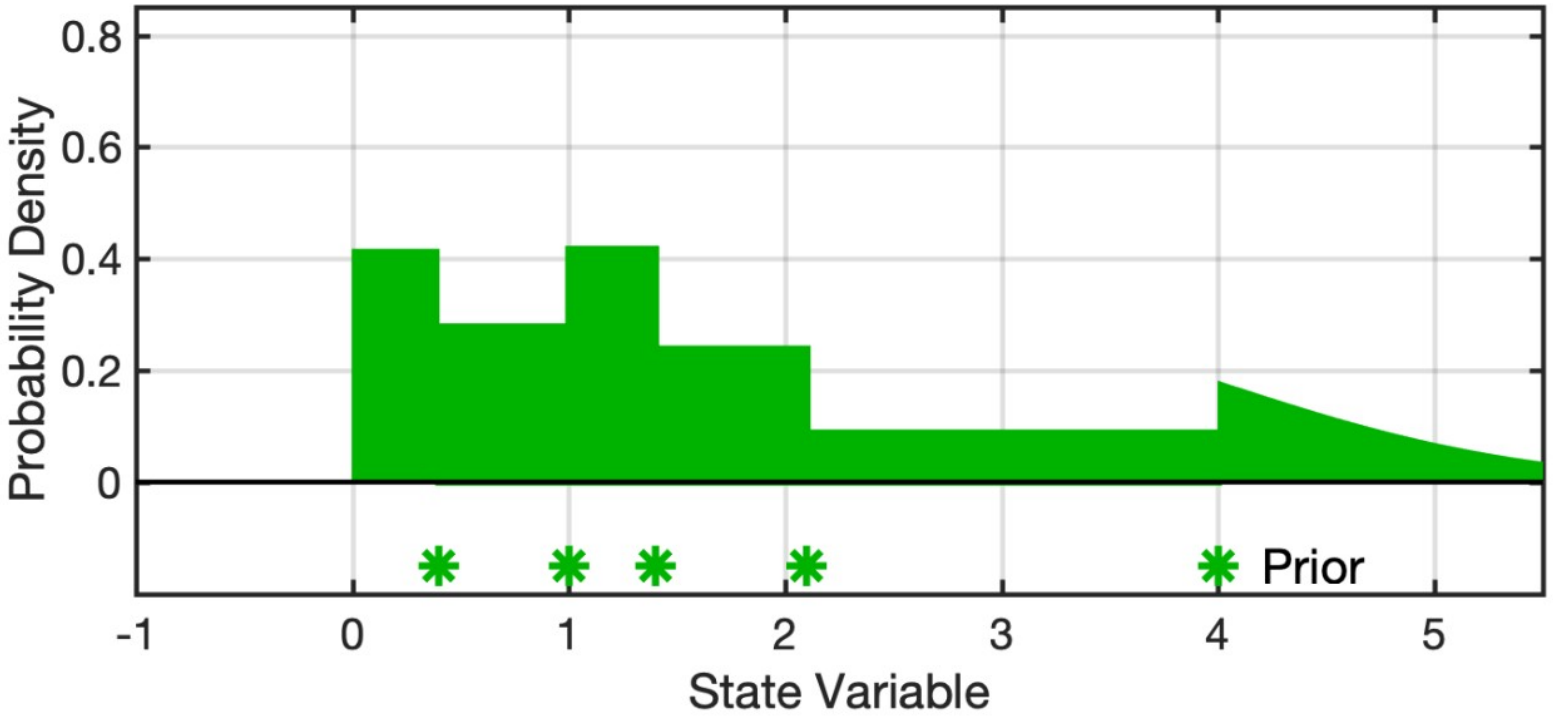
Rank Histogram Continuous Prior

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

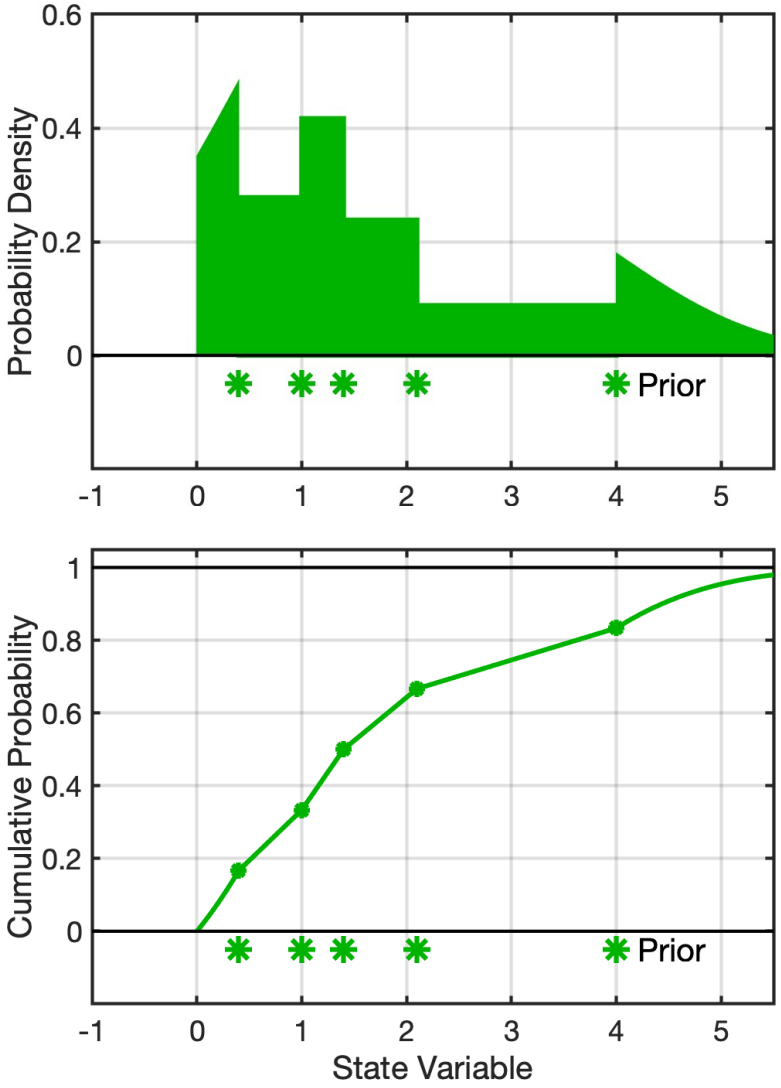


Bounded Rank Histogram Continuous Prior

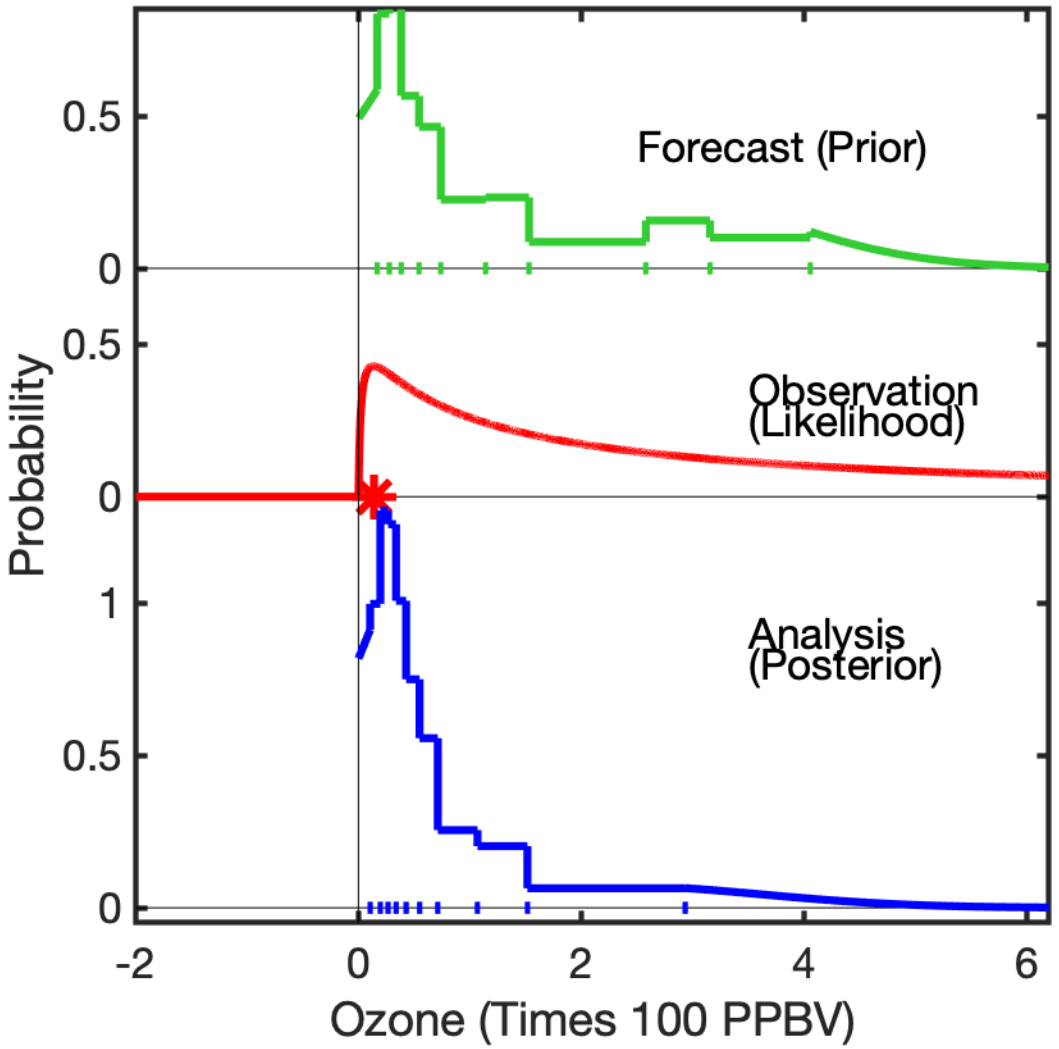
Bounded has truncated tail.
Quantiles are exactly U(0, 1) by construction.



Rank Histogram Prior PDF and CDF with Lower Bound



What if I Don't Know the Right Distribution?



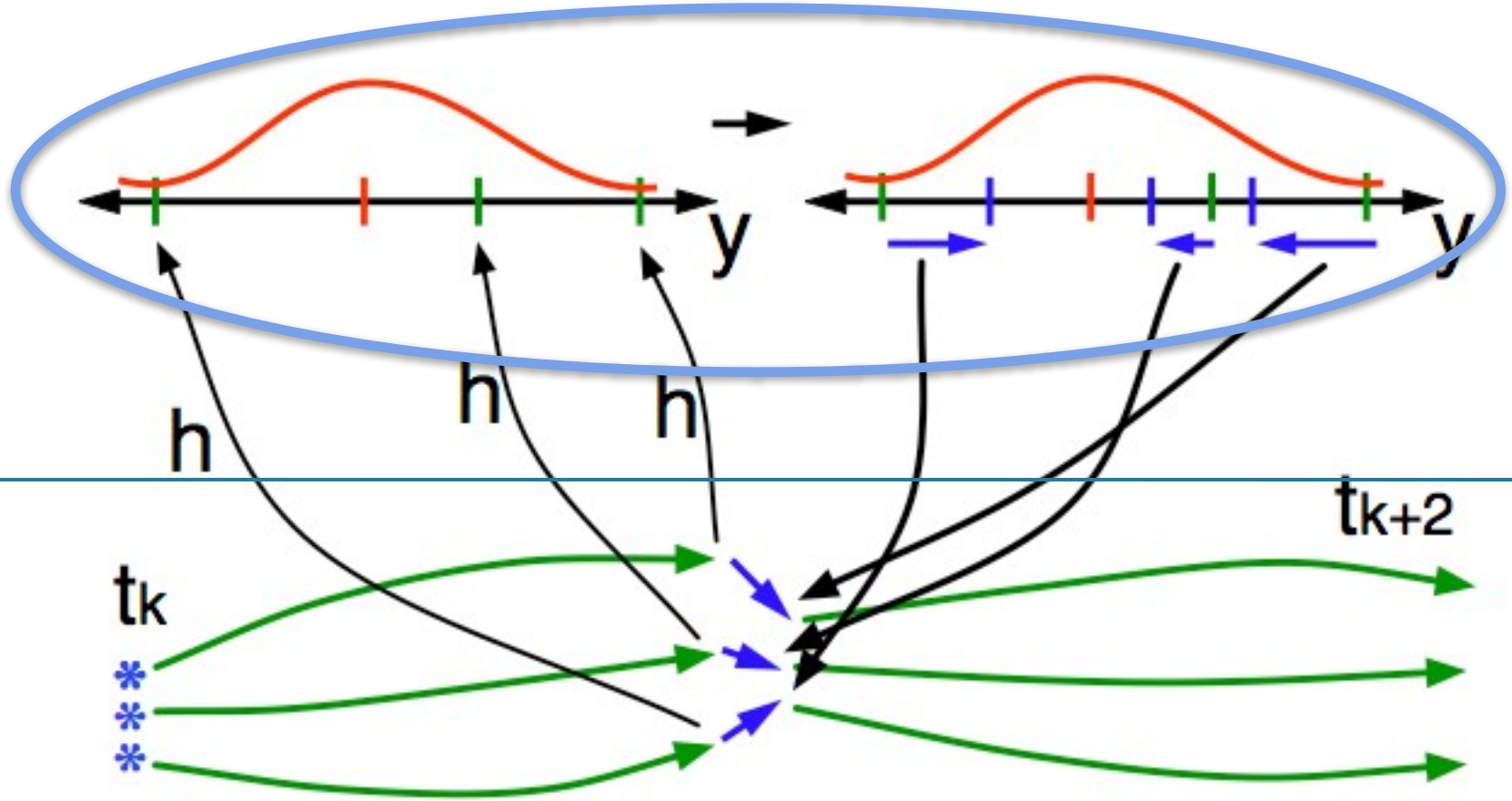
The Bounded Normal Rank Histogram Distribution works well for almost all cases.

Non-parametric. It builds a distribution from the ensemble.

All examples here use this distribution.

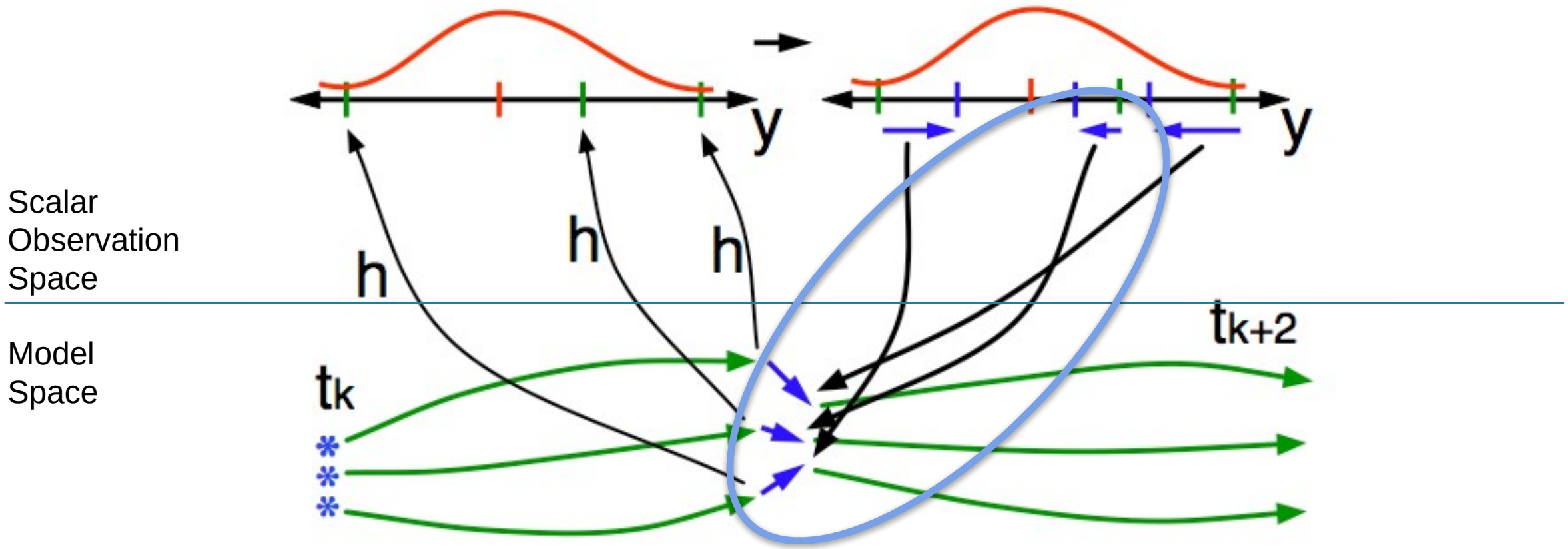
Quantile Conserving Ensemble Filters in Observation Space

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



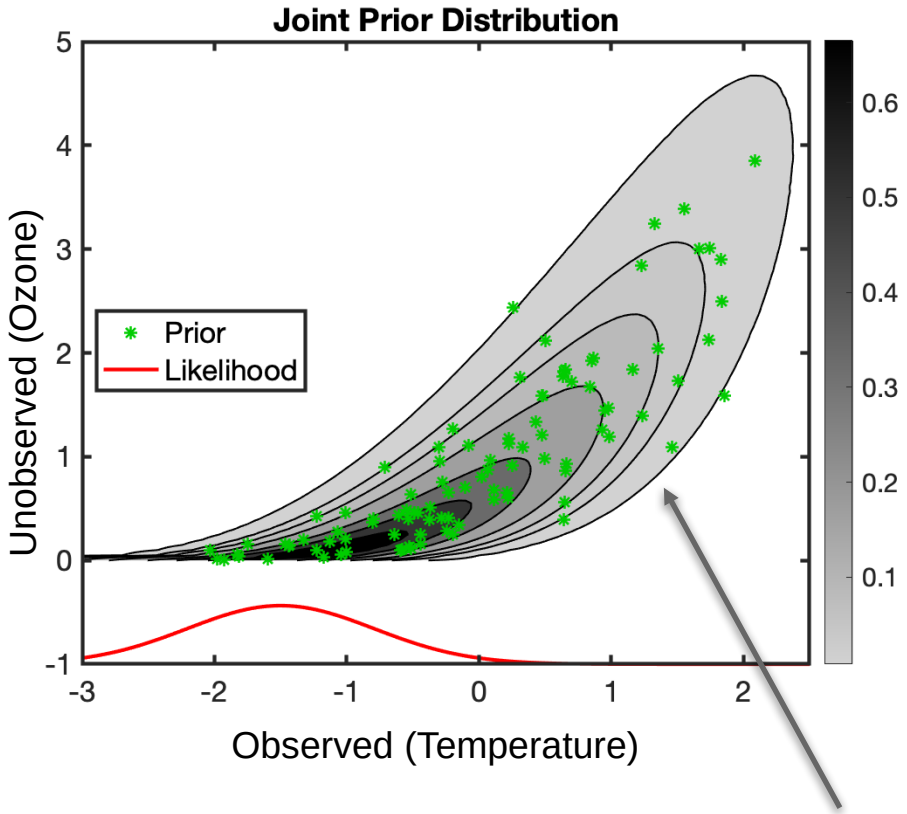
Linear Regression can Wreck Things

Linear regression can destroy benefits of new observation update method.



Standard EnKF: Challenged by Non-Gaussian and Nonlinear Relations

Prior for normal-gamma distribution with 100 member ensemble.

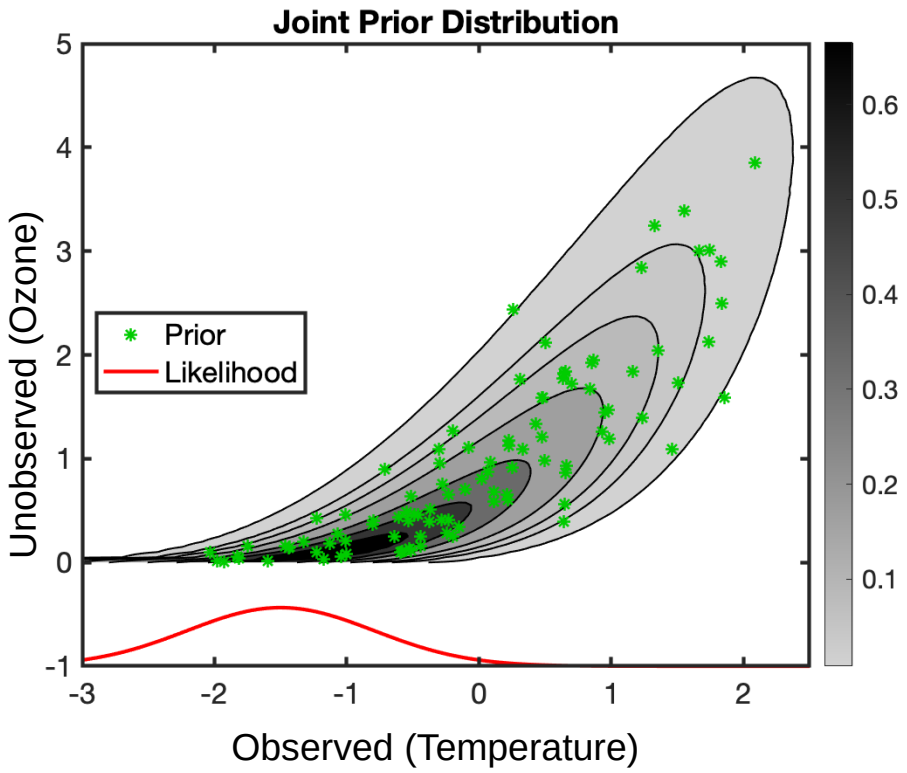


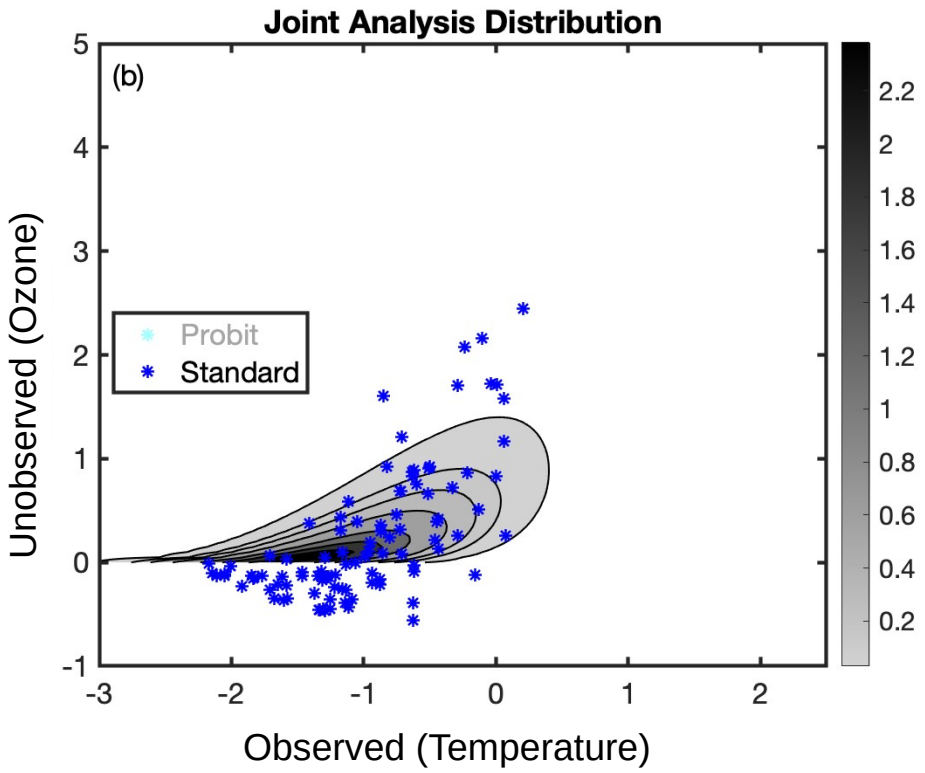
Correct distribution contours are 1, 5, 10, 20, 40, 60, 80% of max for all figures.

Standard EnKF: Challenged by Non-Gaussian and Nonlinear Relations

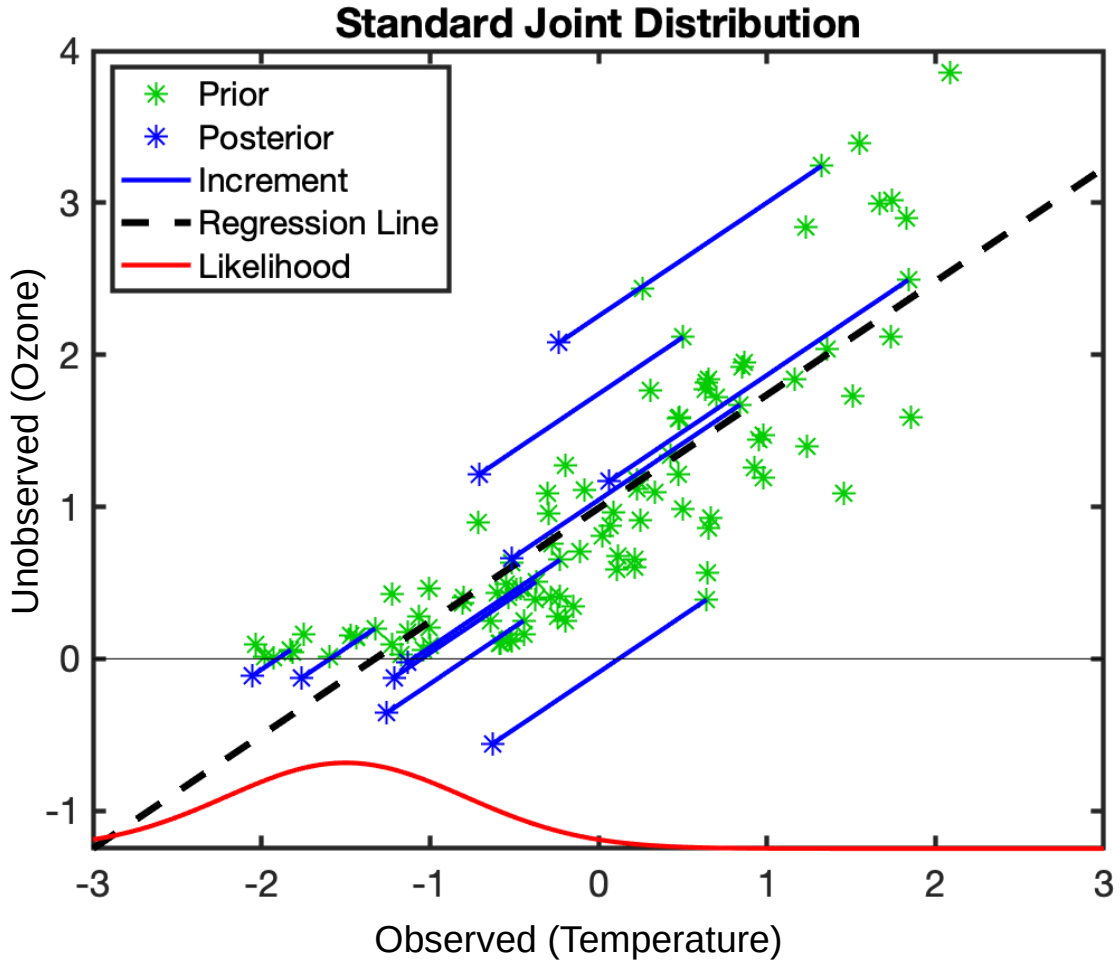
Prior for normal-gamma distribution with 100 member ensemble.

Posterior ensemble has problems.





Standard EnKF: Challenged by Non-Gaussian and Nonlinear Relations



Example regression increment vectors:

Don’t respect bounds,

Struggle with nonlinearity.

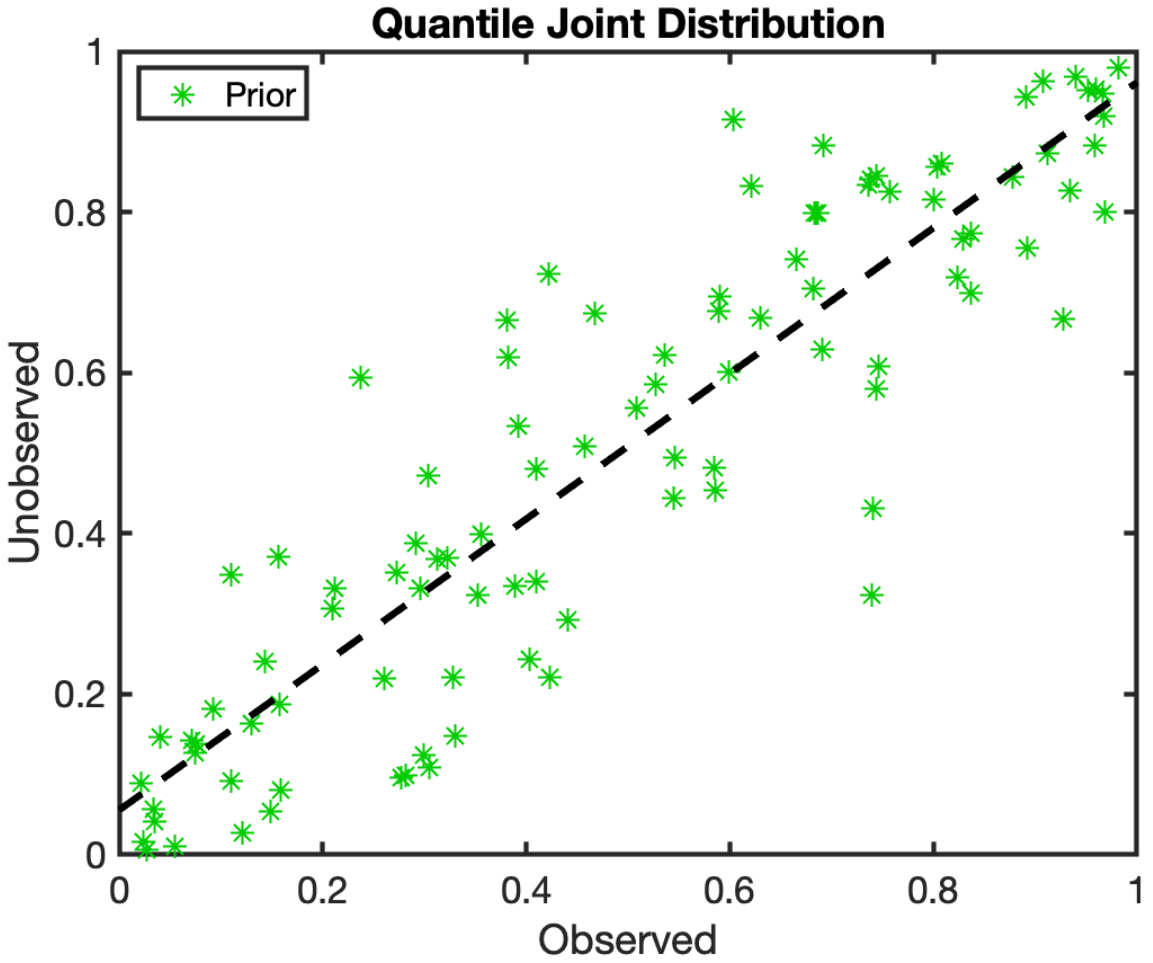
Solution, Transform Marginals: Step 1: Compute Quantiles

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.

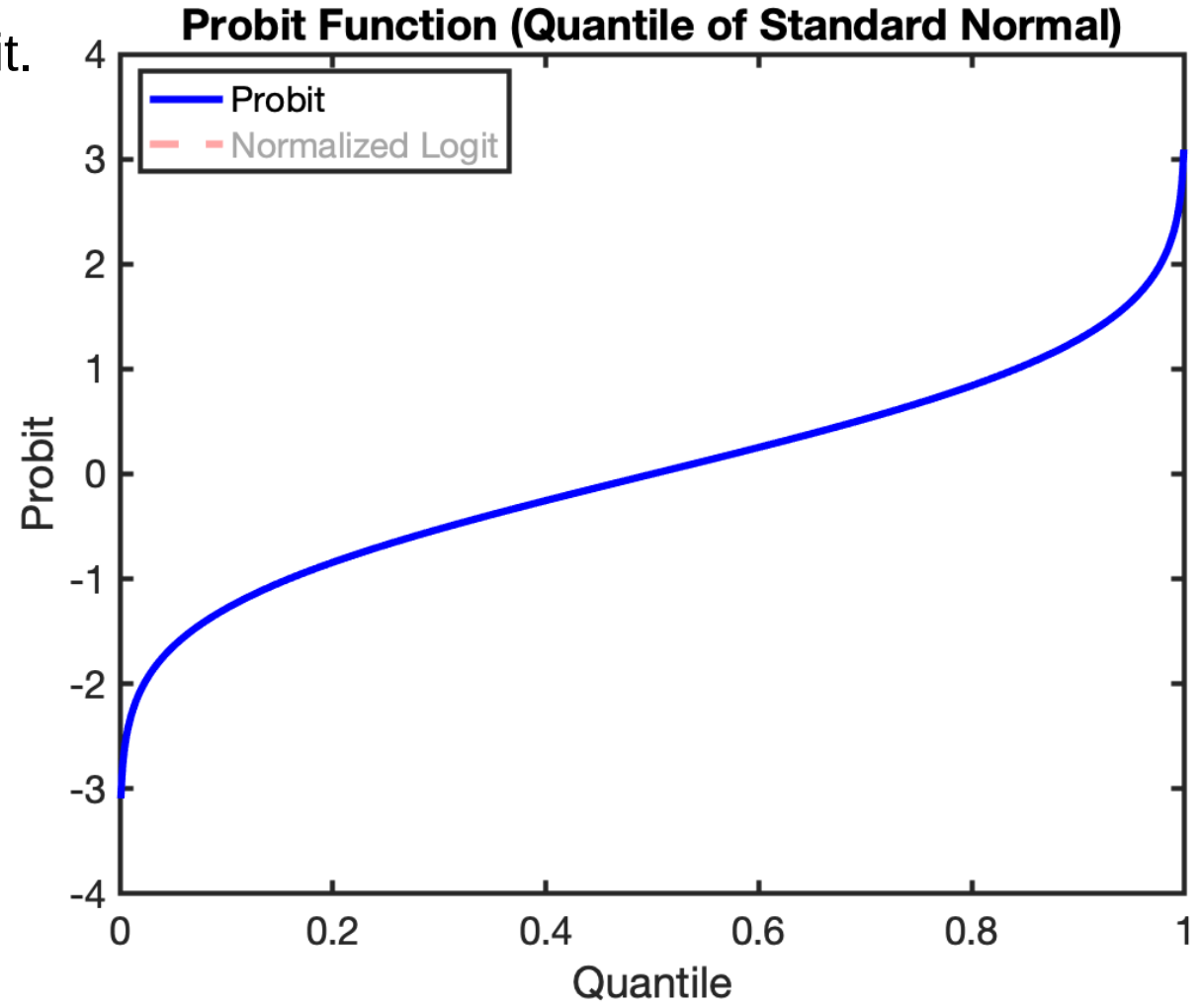


Solution, Transform Marginals: Step 2: Probit Transform of Quantiles

Quantile function for the standard Normal is probit.

Transforms U(0, 1) to unbounded.

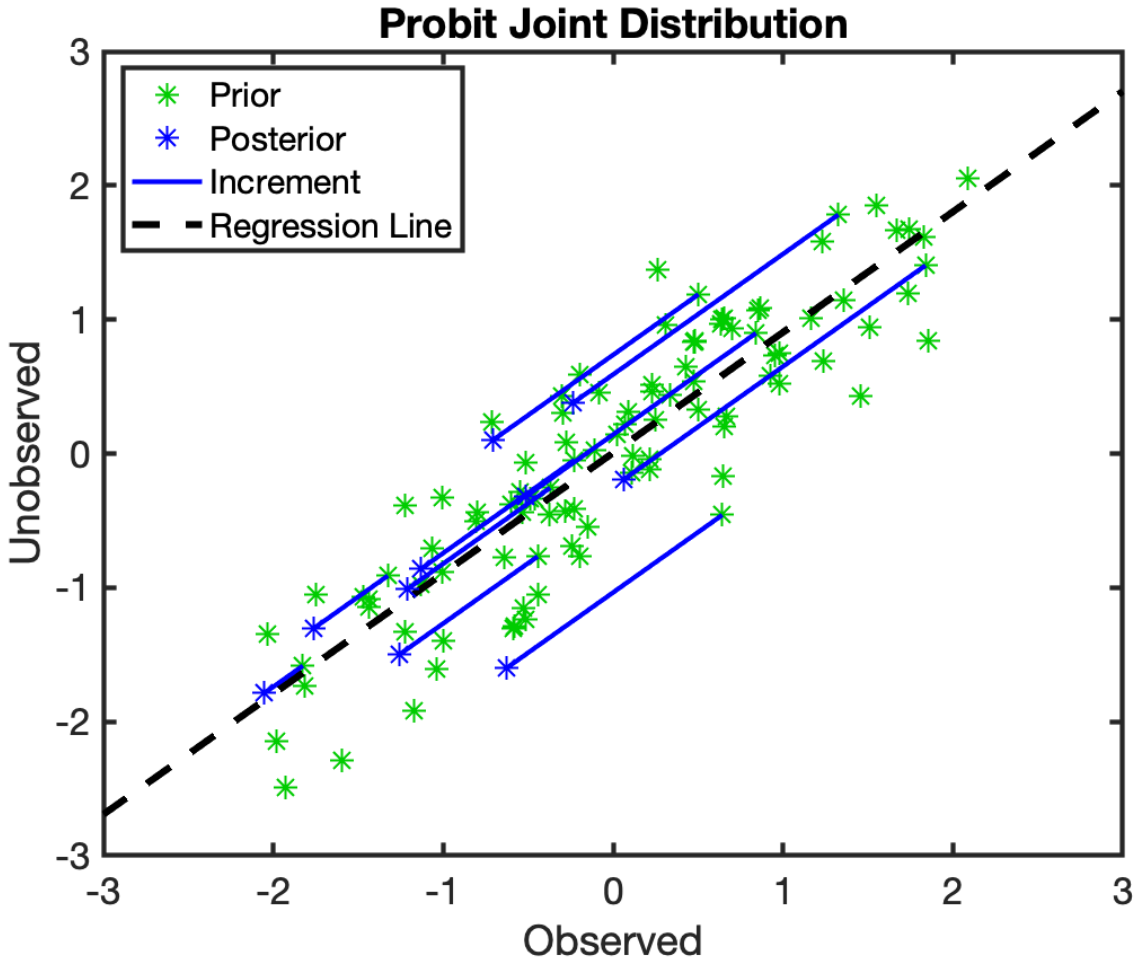
Marginal distributions should be N(0, 1).



Regression in Probit-Transformed Quantile Space

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

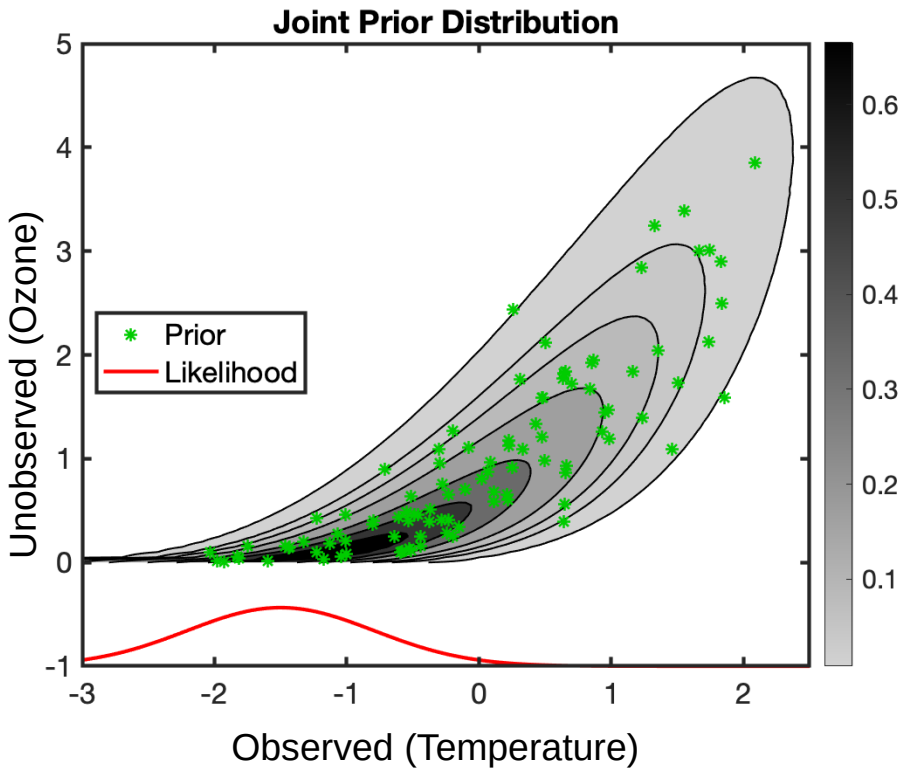
Linear regression is best unbiased linear estimator (BLUE) in this space.

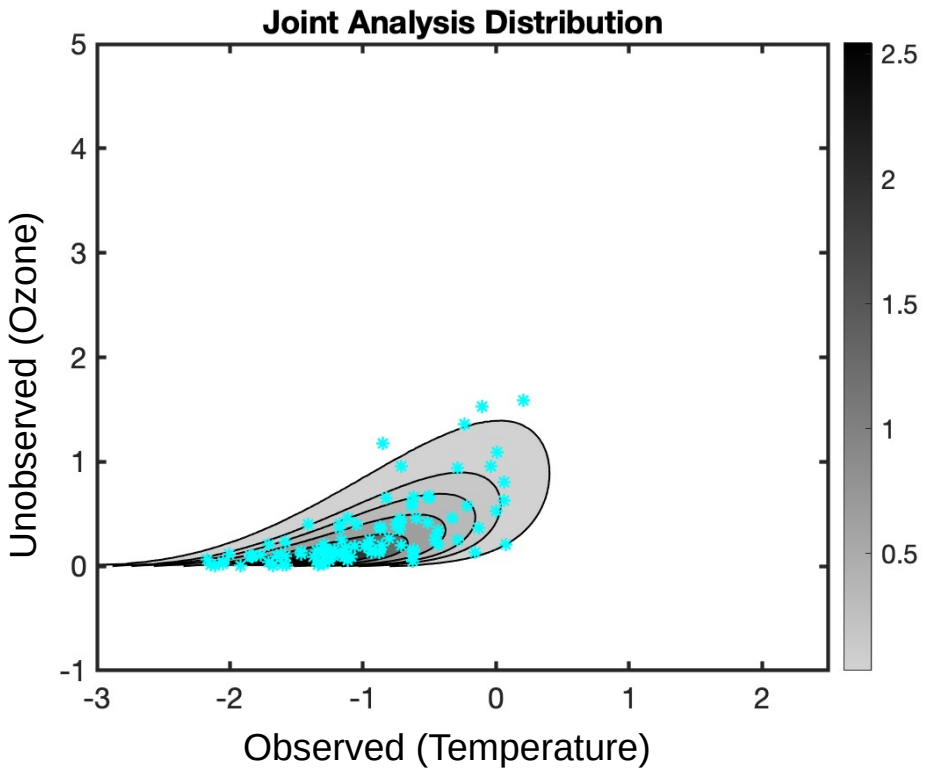


Novel, General Solutions for Nonlinear, Non-Gaussian Problems

Prior for normal-gamma distribution with 100 member ensemble.

Bounds enforced. Nonlinear aspect respected.

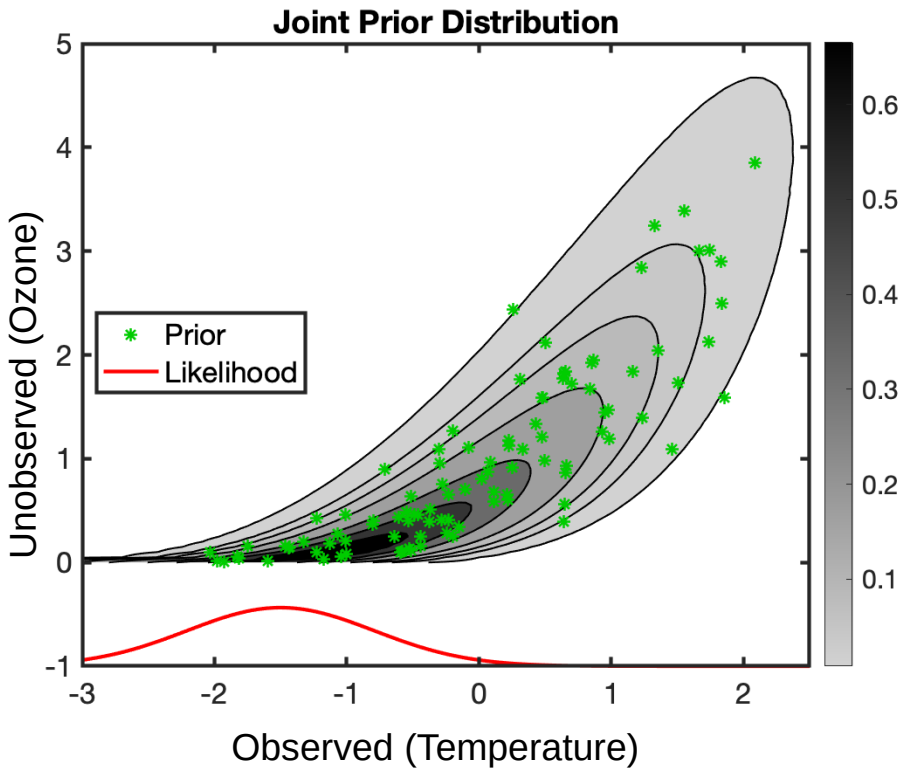


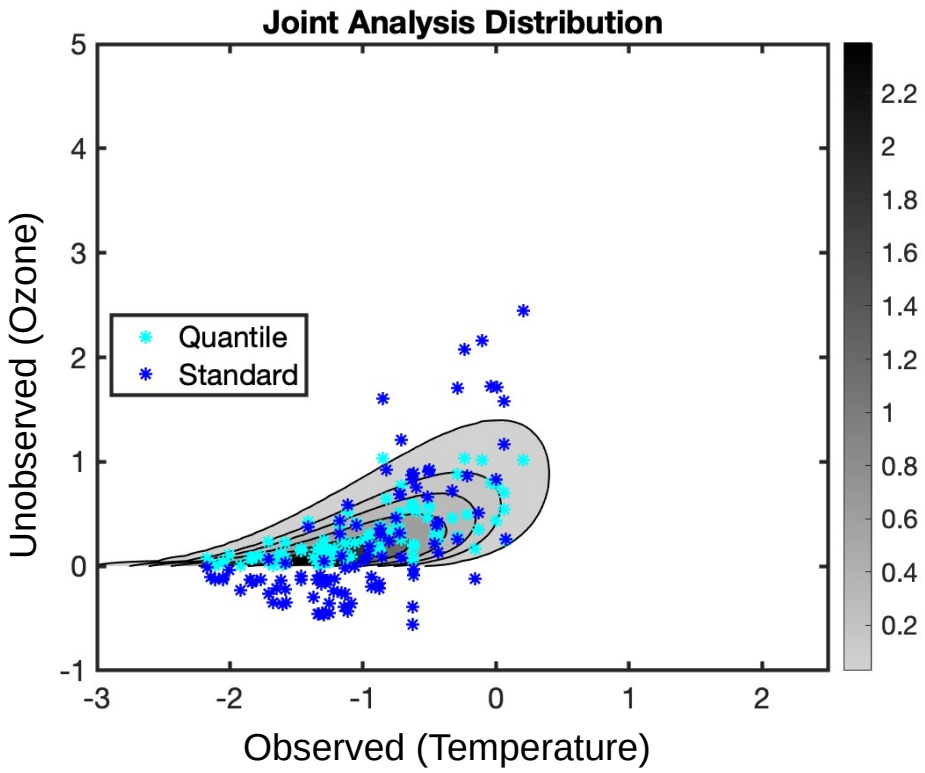


Novel, General Solutions for Nonlinear, Non-Gaussian Problems

Prior for normal-gamma distribution with 100 member ensemble.

Bounds enforced. Nonlinear aspect respected.

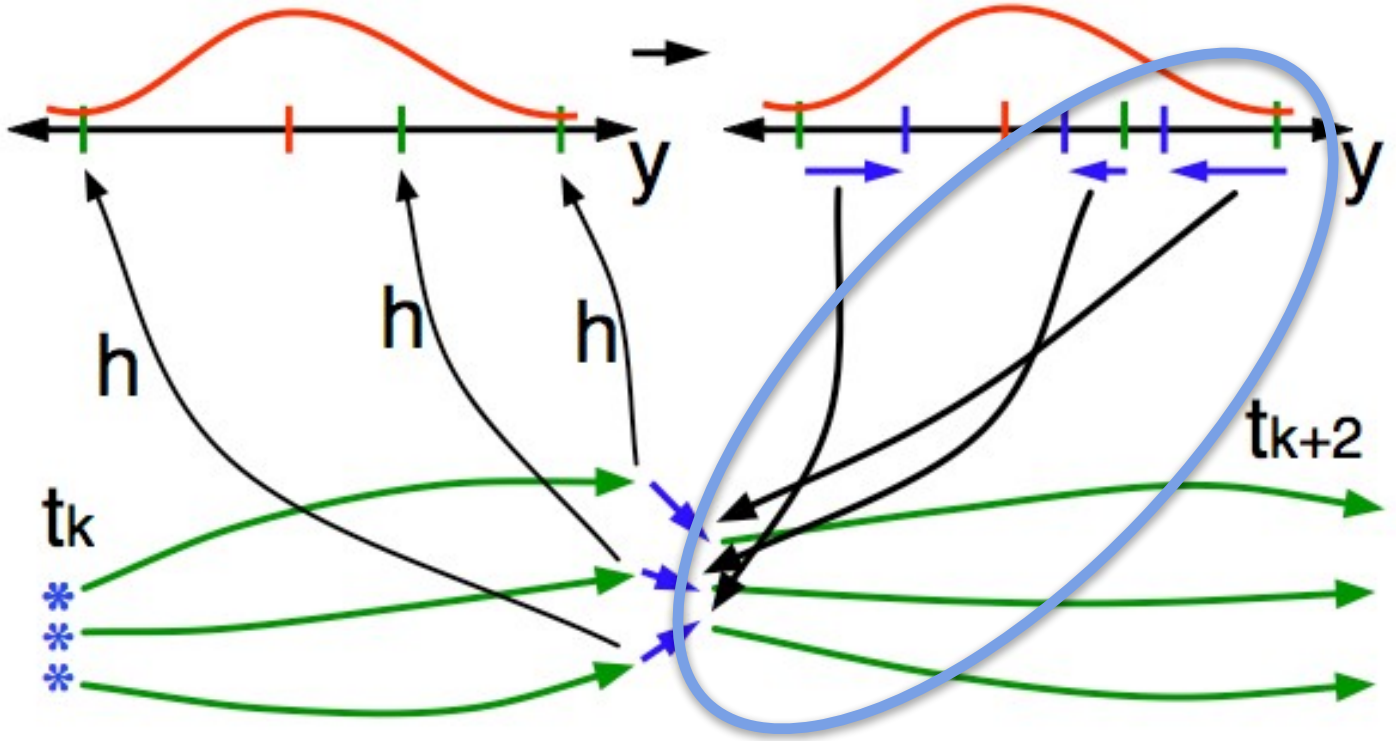




Implement Regression in a Transformed Space

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

, , , *n=1, …N* are prior and posterior (analysis) ensembles of observed variable y and

unobserved variable x

and are continuous CDFs appropriate for x and y

is the CDF of the standard normal, is the probit function

, and are probit space

is probit space observation increment

regress increments in probit space (eq. 5 Anderson 2003)

is posterior ensemble in probit space

is posterior ensemble

Mixed Distributions: A Challenge for Tracers and Sources

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

Sea ice concentration is an example:

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

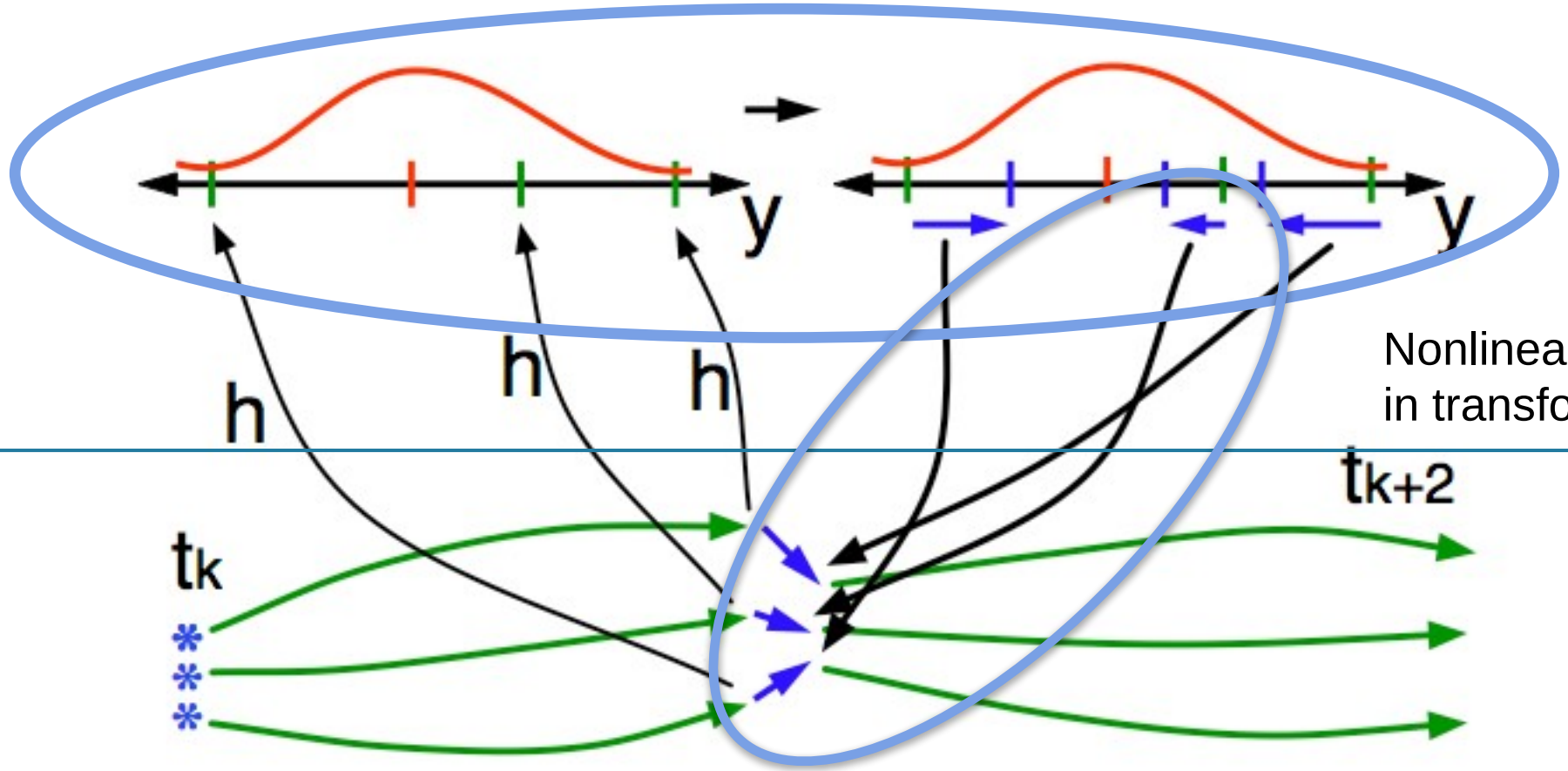
Important for some tracers and many sources,
Also crucial for good rainfall DA.

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

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

Have a General Nonlinear Non-Gaussian Ensemble Filter Framework

Arbitrary distributions including nonparametric BNRH for observation space.

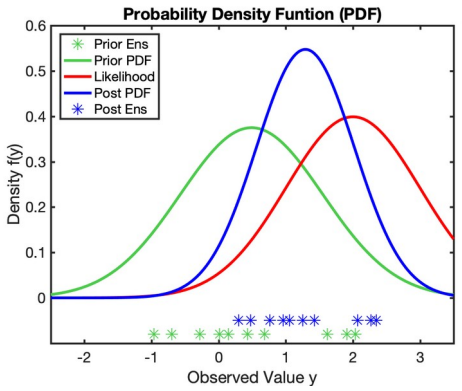


Scalar Observation Space

Model Space

Nonlinear regression in transformed space.

Quantile Conserving Ensemble Filters, Generalizing the EAKF



When a normal PDF is fit for the observation space update, and linear regression is used to update state variables, this is the existing Ensemble Adjustment Kalman Filter (EAKF).

The EAKF is equivalent to the Kalman filter for linear / normal problems.

Quantile Conserving Ensemble Filters (QCEF)

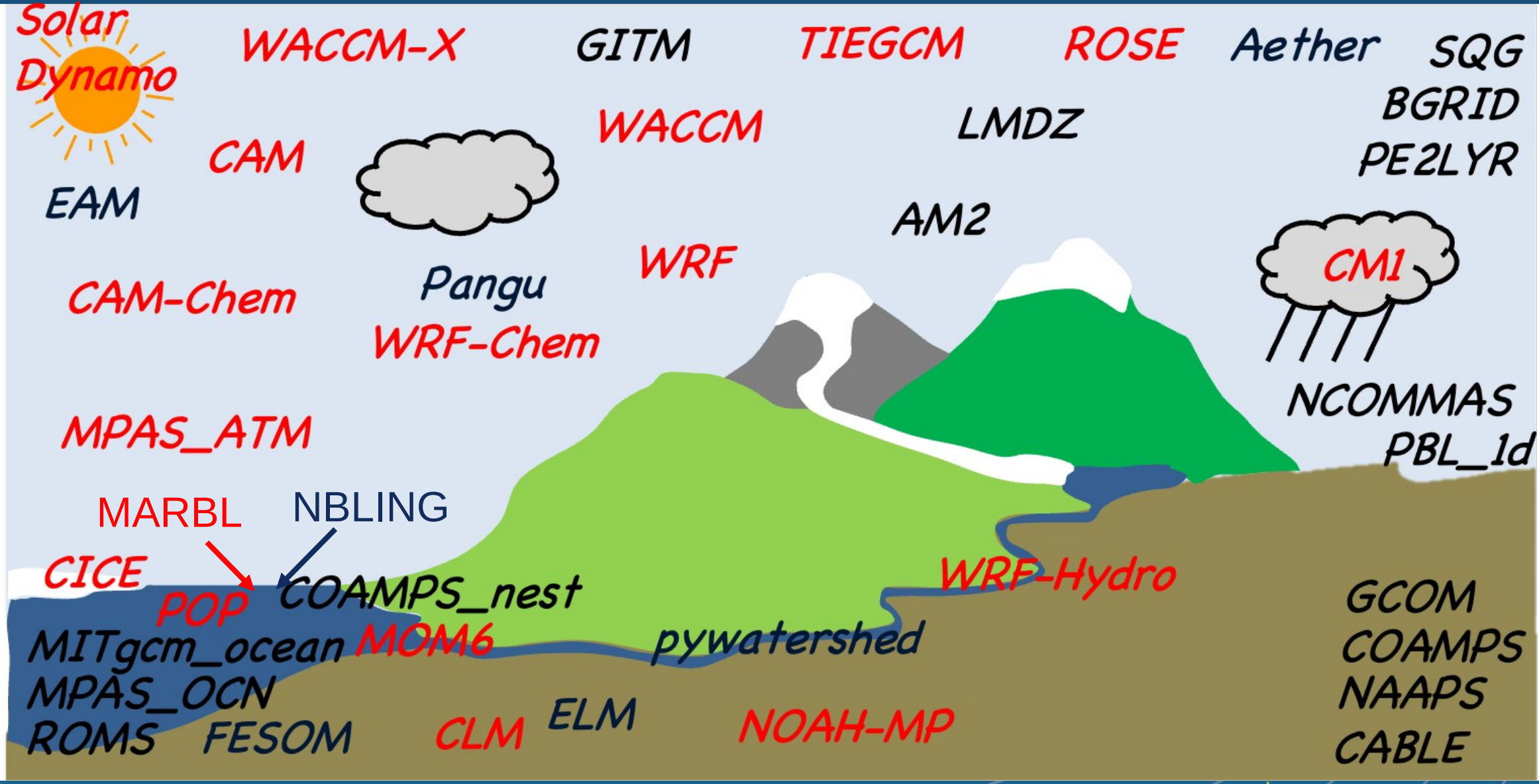
Improve performance for applications across the Earth system:

- RMSE and spread,
- Higher order moments,
- Full distribution.

- Reduce imbalance and spin-up,

- Use information from nonlinear relations more effectively.

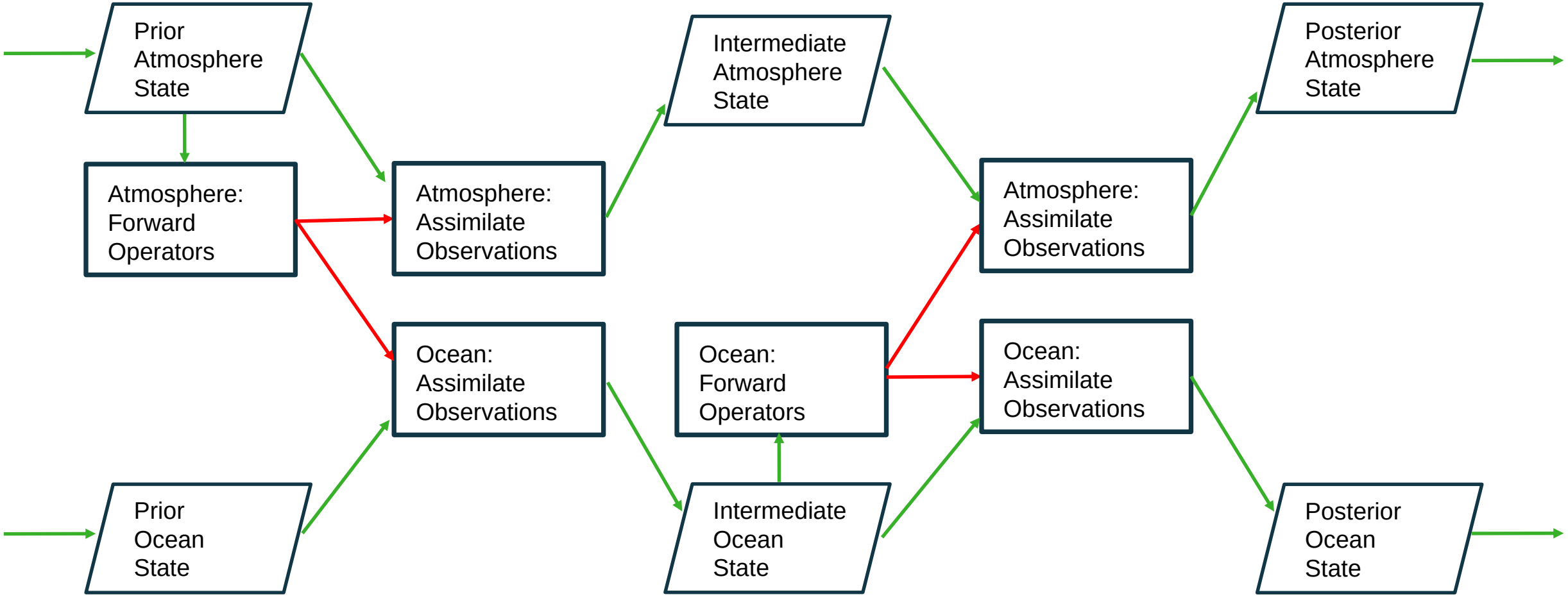
Geophysical Models Interfaced to DART (NCAR Models in Red)



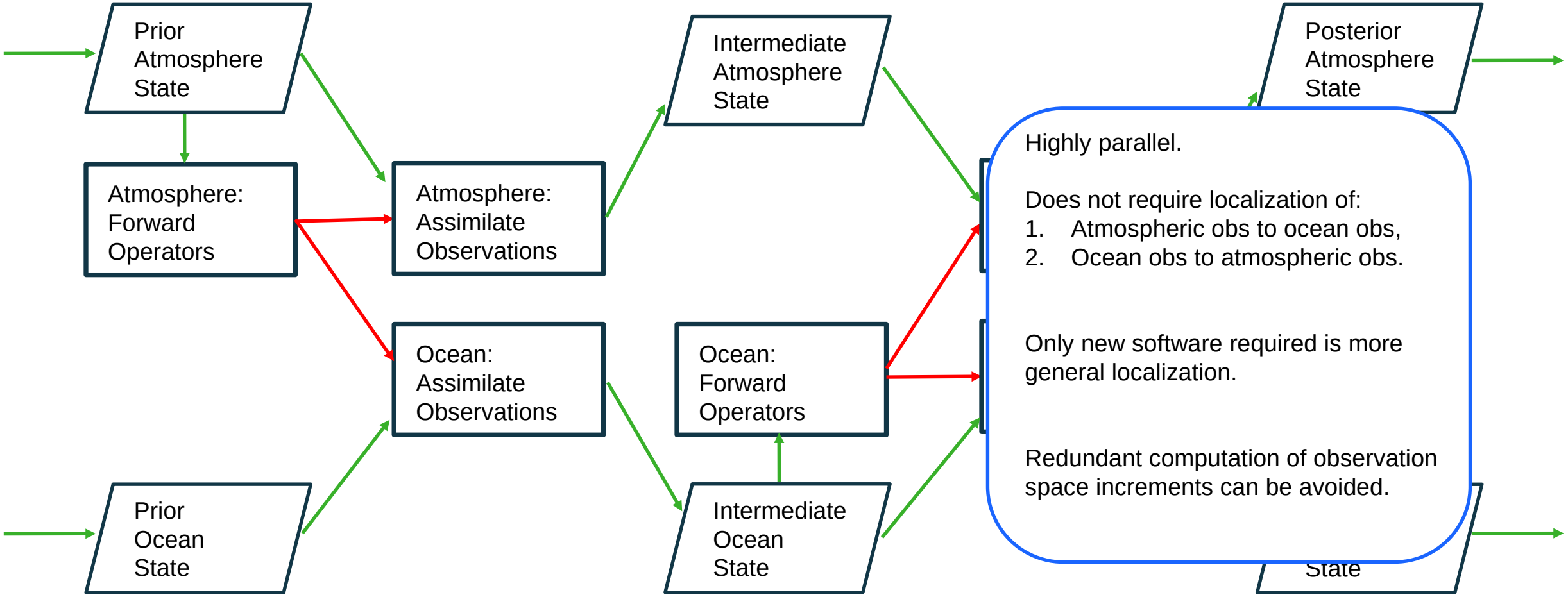
Strongly Coupled DA

Framework also allows general strongly coupled DA implementation.

Strongly Coupled DA



Strongly Coupled DA



Strongly Coupled DA

- Any models with DART interfaces,
- No additional assimilation code,
- Minimal scripting for file motion,
- Localization of obs impact on model state is still a key scientific challenge,
- It’s not a software challenge.

Strongly Coupled DA Challenges

1. Limited times and places with meaningful correlations between observations of one Earth system component and state of another.

2. Significant model errors in many of these places (boundary and mixed layers).

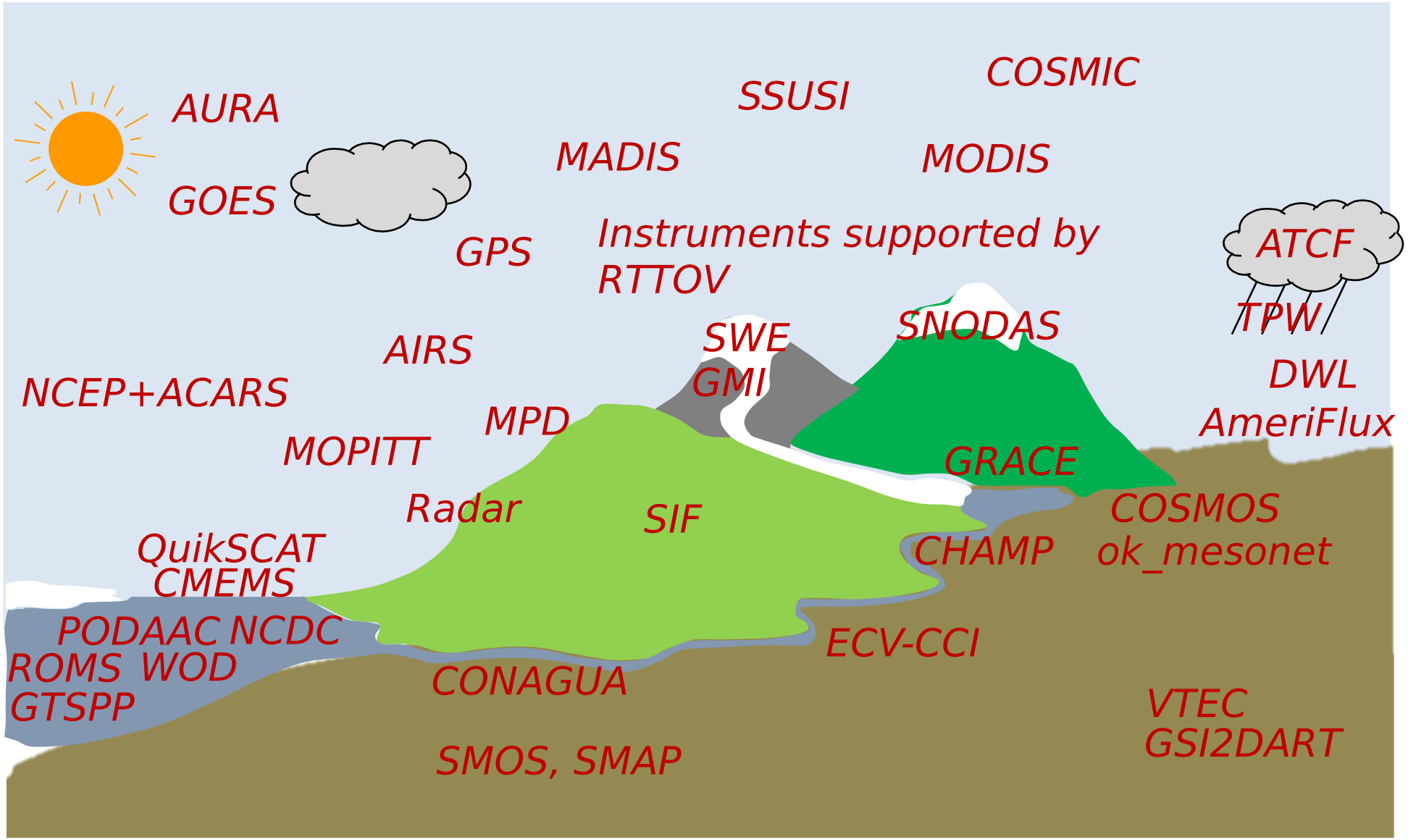
3. If both models are well-observed, cross model impacts may not add much.

4. If strongly-forcing model is well observed, don't need to do cross model DA.

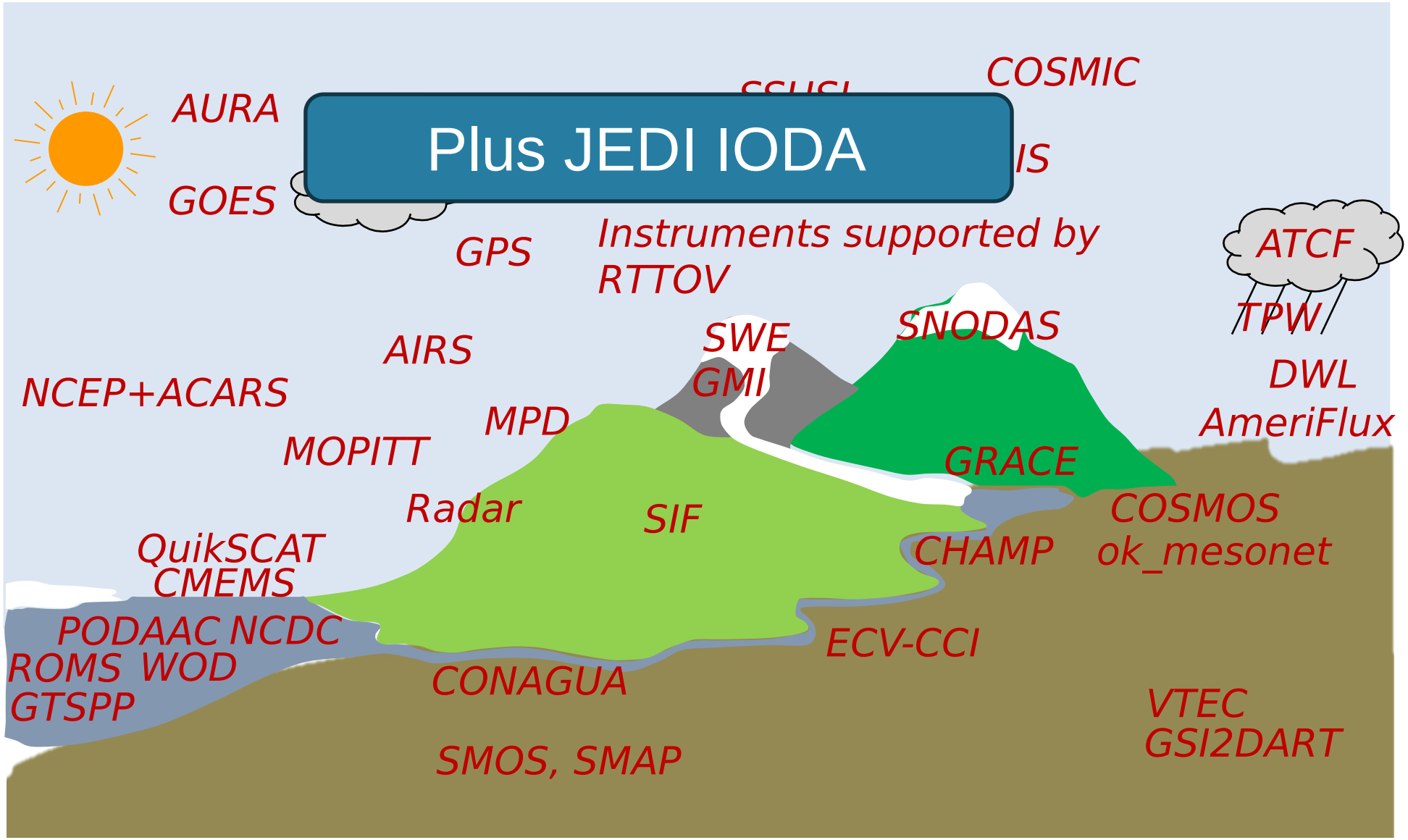
Best case:

- Component A strongly forces component B,
- B is well-observed, A is not,
- There is not significant model bias in this area.

Earth System Observations (Many Others Available)

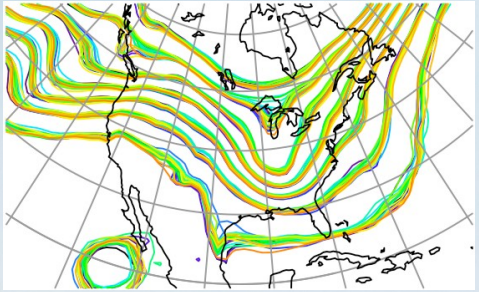


Earth System Observations (Many Others Available)





NCAR UCAR | CISL
Computational & Information Systems Lab



DAReS is Looking for Collaborators to Apply Strongly-Coupled DA for Earth System Prediction

https://dart.ucar.edu

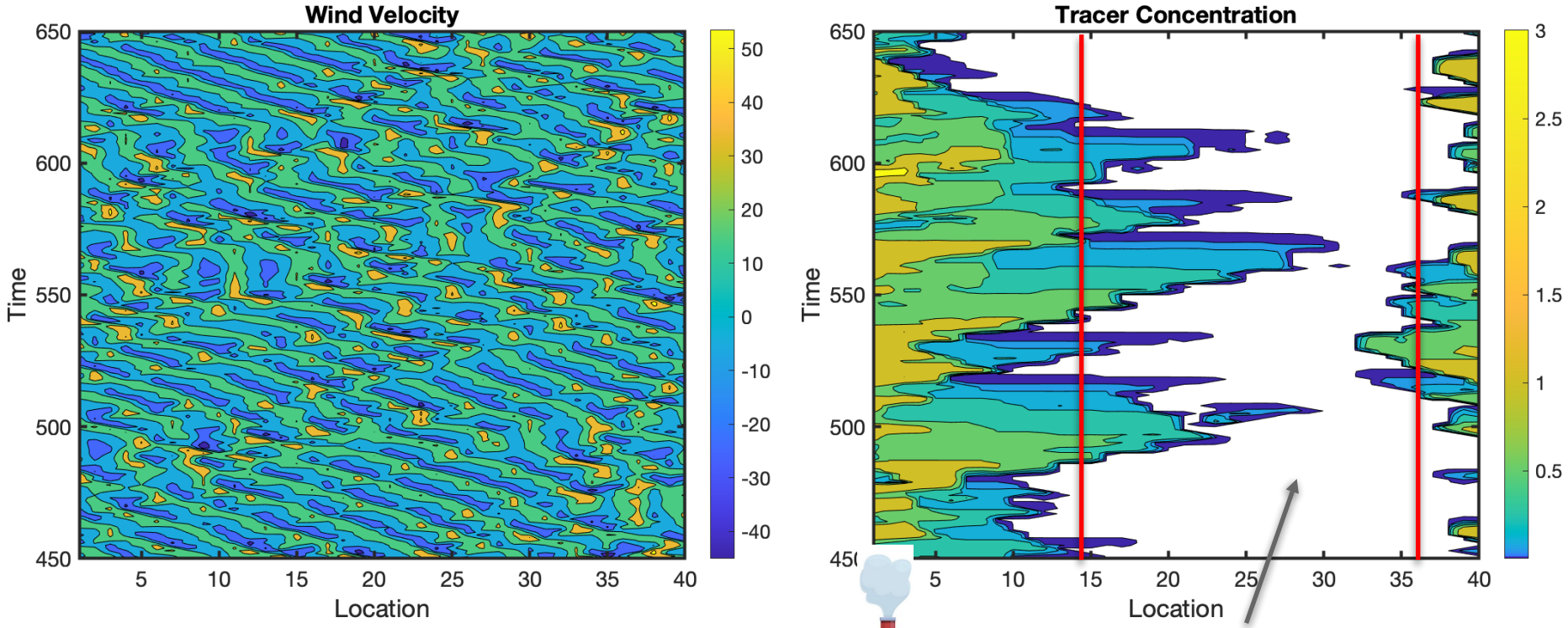
NCAR UCAR | National Center for Atmospheric Research

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.

Low-Order Tracer Advection Model Example

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.

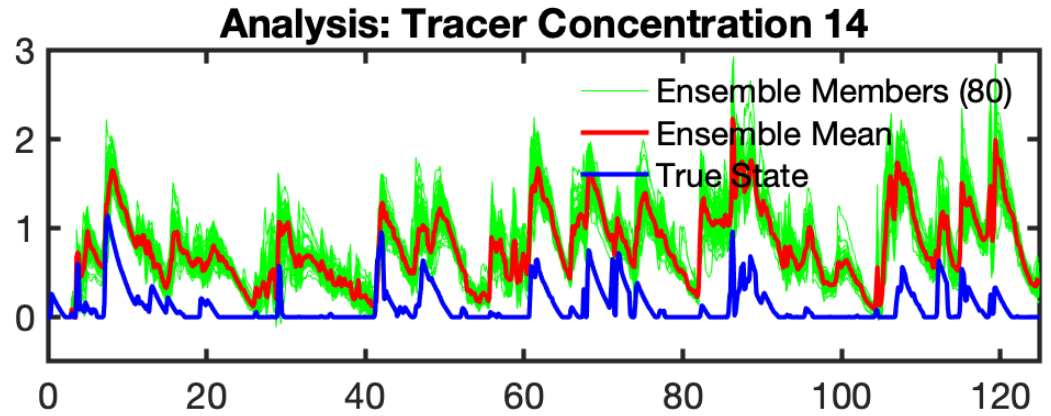


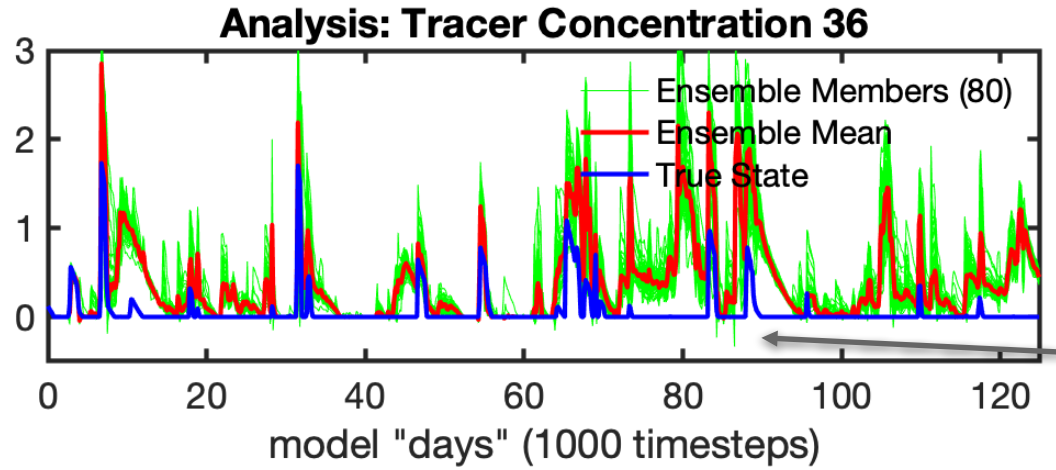
Constant Source.

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.

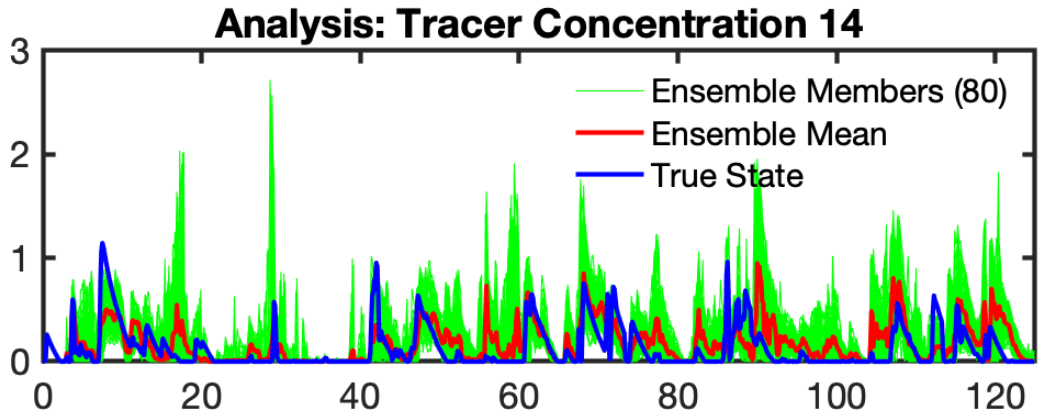
EAKF has large bias for tracers.

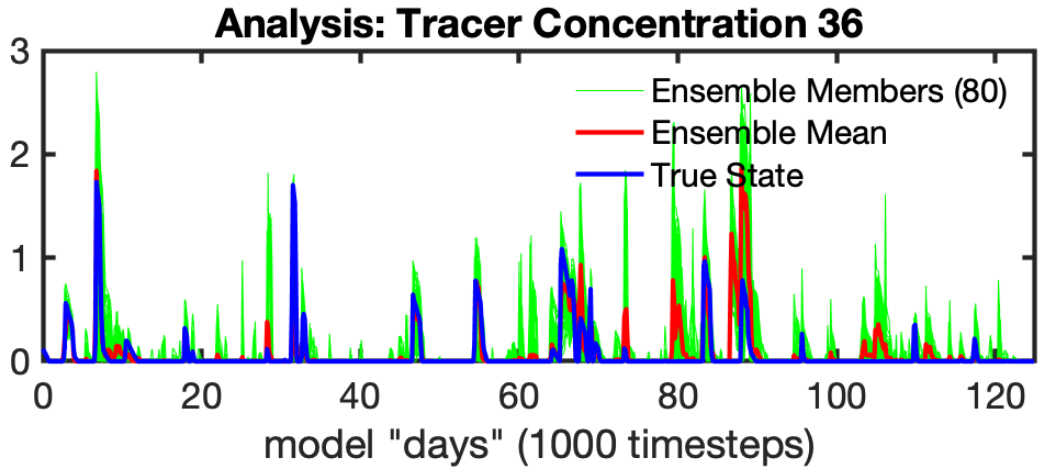
Can't go to all zeros.

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.

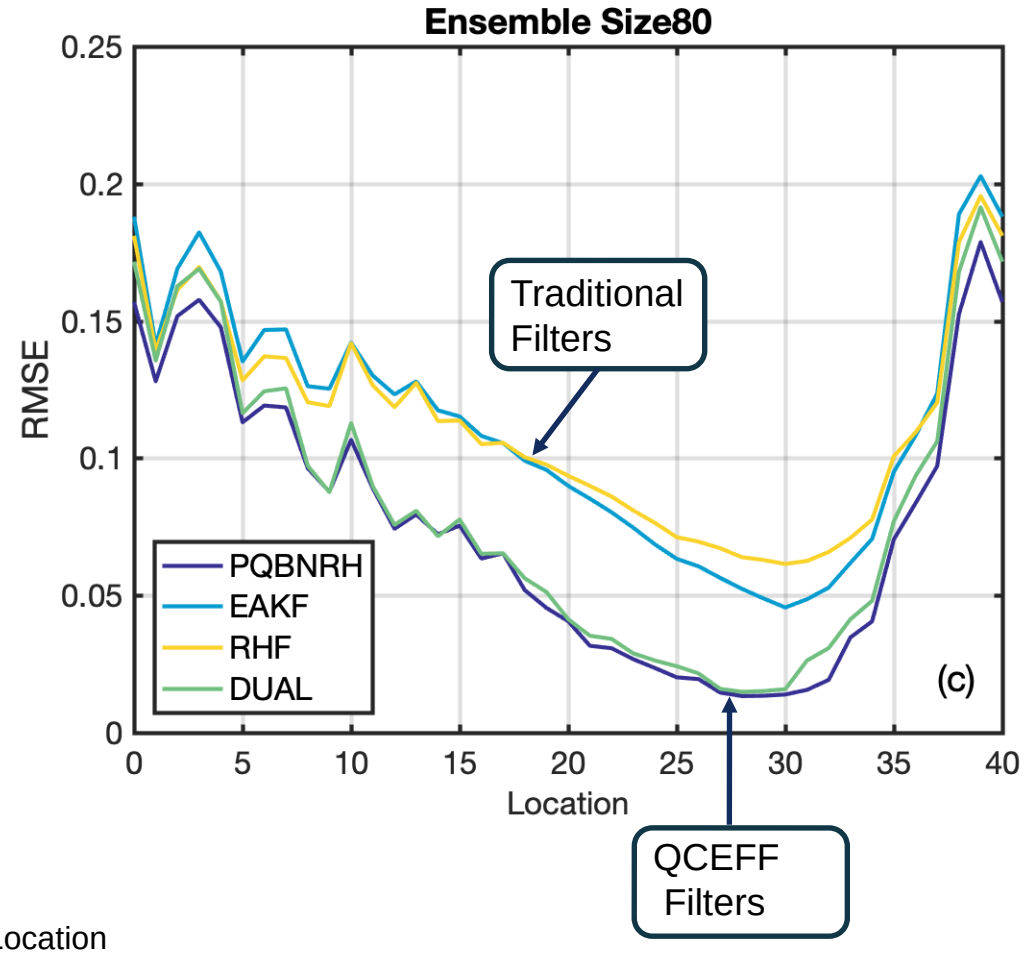
Bounded Normal Rank Histogram with quantile regression is unbiased.

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.



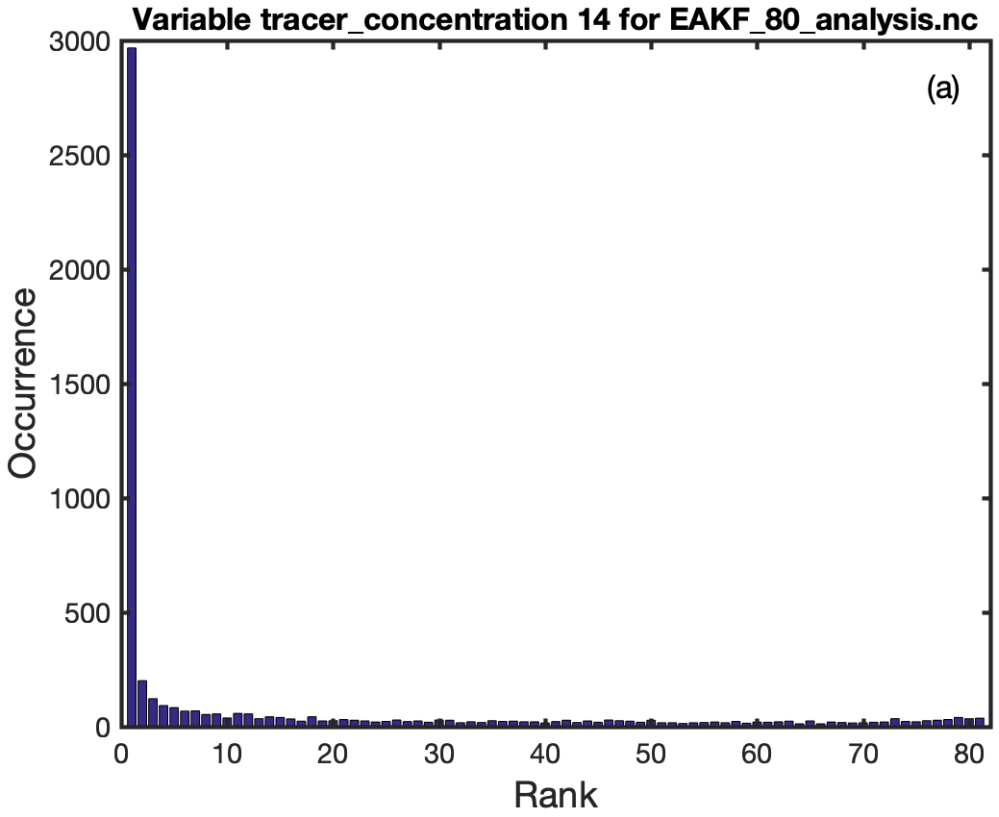
Location

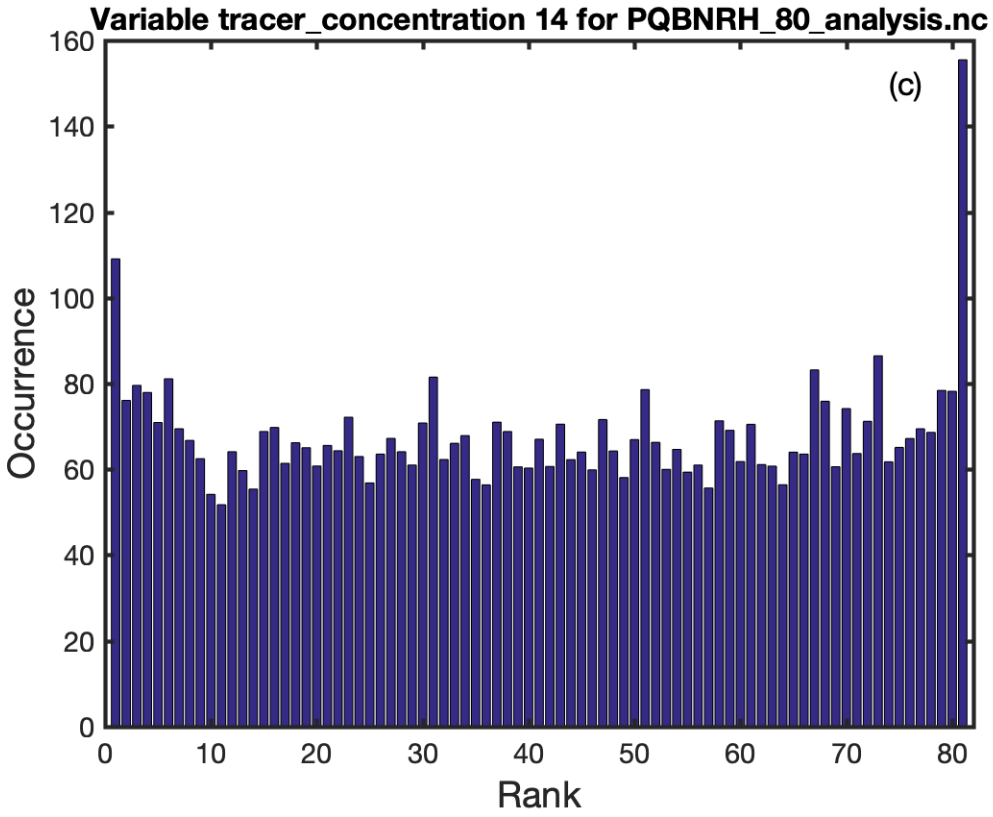
Low-Order Tracer Advection Model: Concentration Rank Histograms

Good ensembles have uniform rank histograms. Makes interpretation easy.

Normal is highly skewed.

QCEFF is close to uniform.



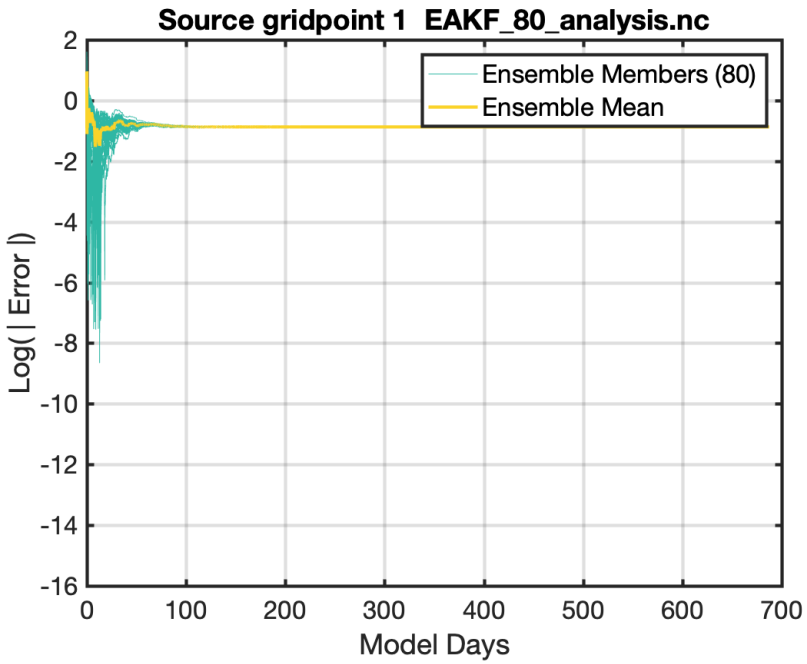


Low-Order Tracer Advection Model: Source Estimation

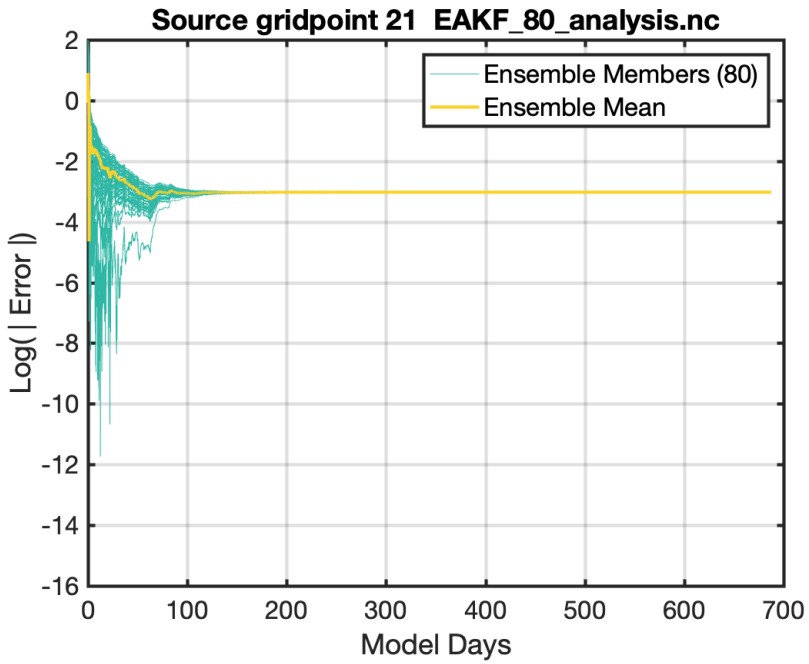
If sources are unknown, can also estimate them.

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

EAKF for nonzero source grid point 1.



EAKF for a zero source grid point 21.

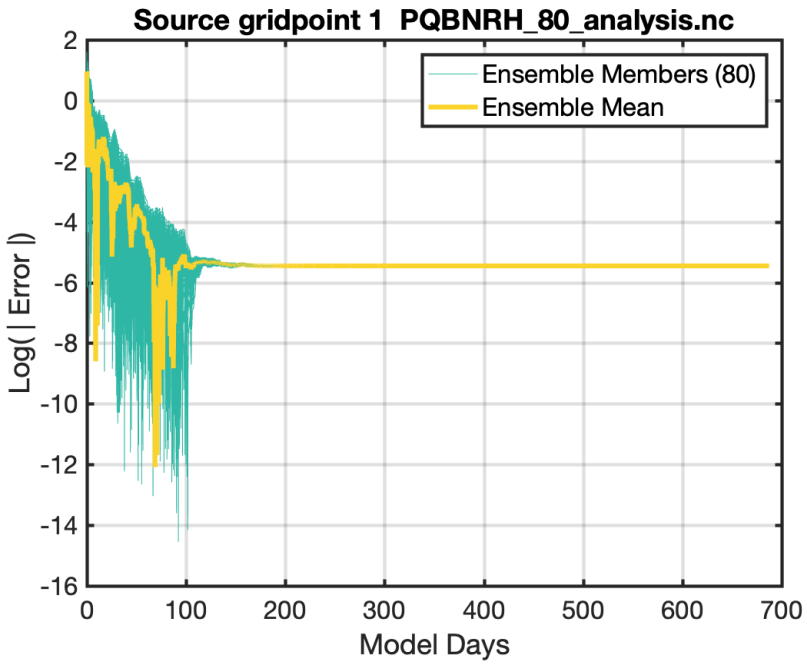


Low-Order Tracer Advection Model: Source Estimation

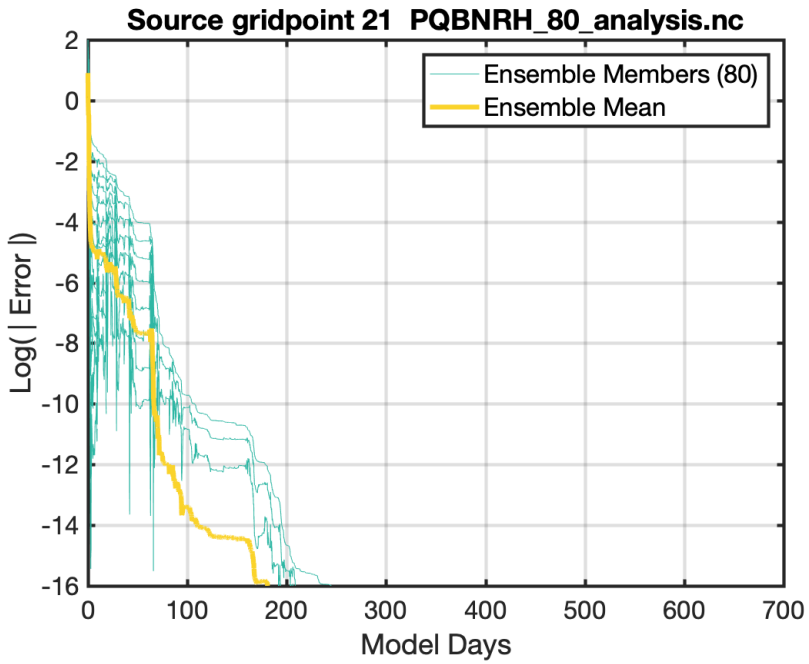
If sources are unknown, can also estimate them.

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

QCEFF for nonzero source grid point 1.



QCEFF for a zero source grid point 21.



Summary

The two step sequential ensemble assimilation algorithms in DART support general Nonlinear and Non-Gaussian data assimilation.

These algorithms can significantly improve results for Earth system DA with negligible added cost.

The same framework readily supports a general strongly coupled DA capability.

Any set of models that have interfaces to DART can be coupled with no additional code.

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

