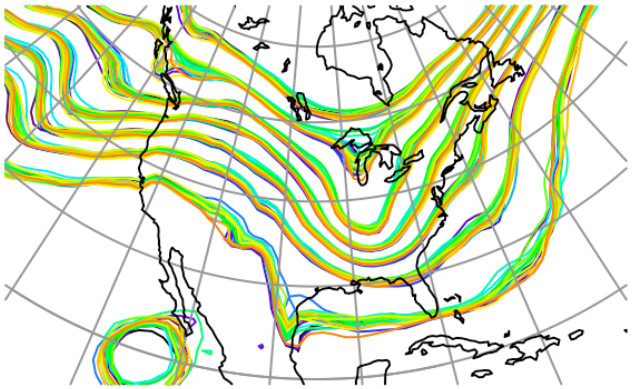
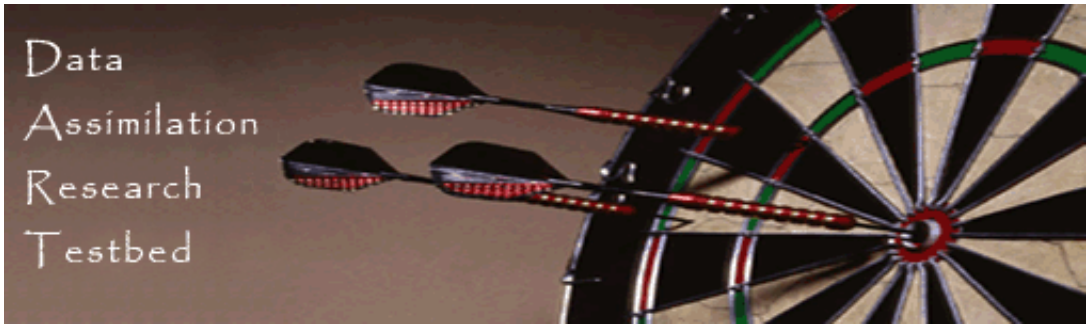


Nonlinear, Nongaussian Ensemble Data Assimilation with Rank Regression and a Rank Histogram Filter

Jeff Anderson, NCAR Data Assimilation Research Section



Schematic of a Sequential Ensemble Filter

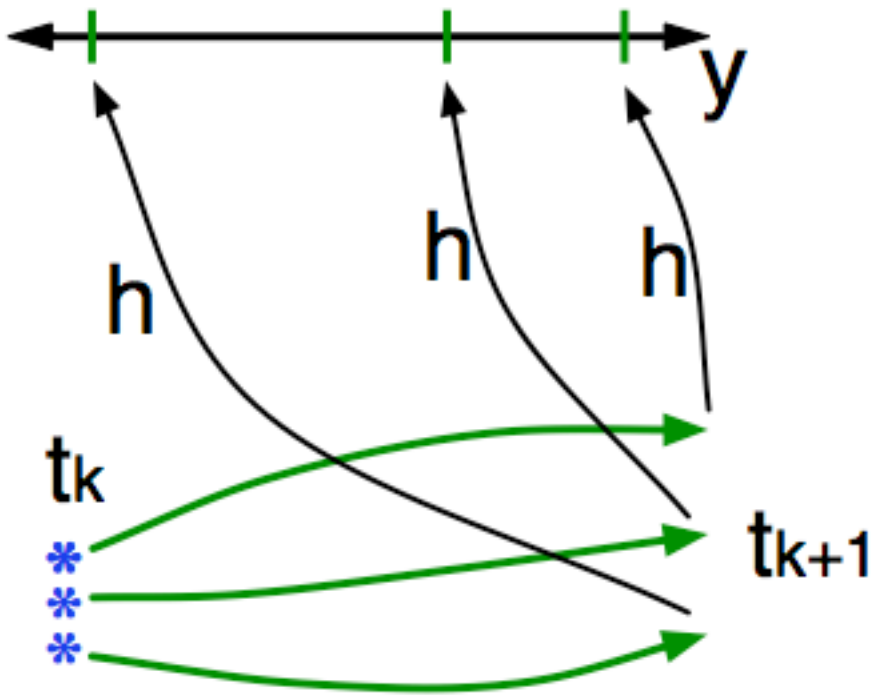
1. Use model to advance ensemble (3 members here) to time at which next observation becomes available.

Ensemble state estimate after using previous observation (analysis)

Ensemble state at time of next observation (prior)

Schematic of a Sequential Ensemble Filter

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

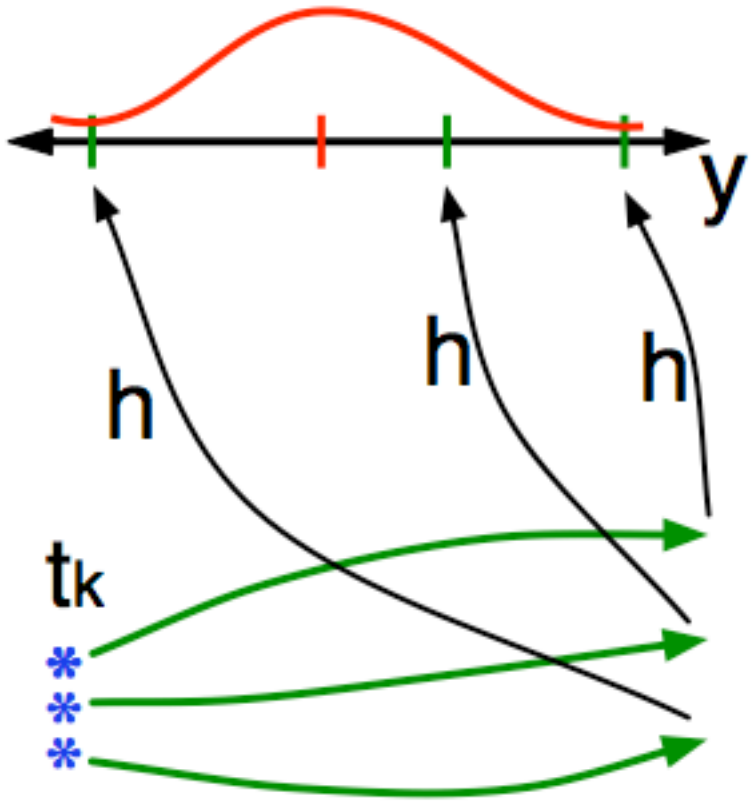


Theory: observations from instruments with uncorrelated errors can be done sequentially.

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

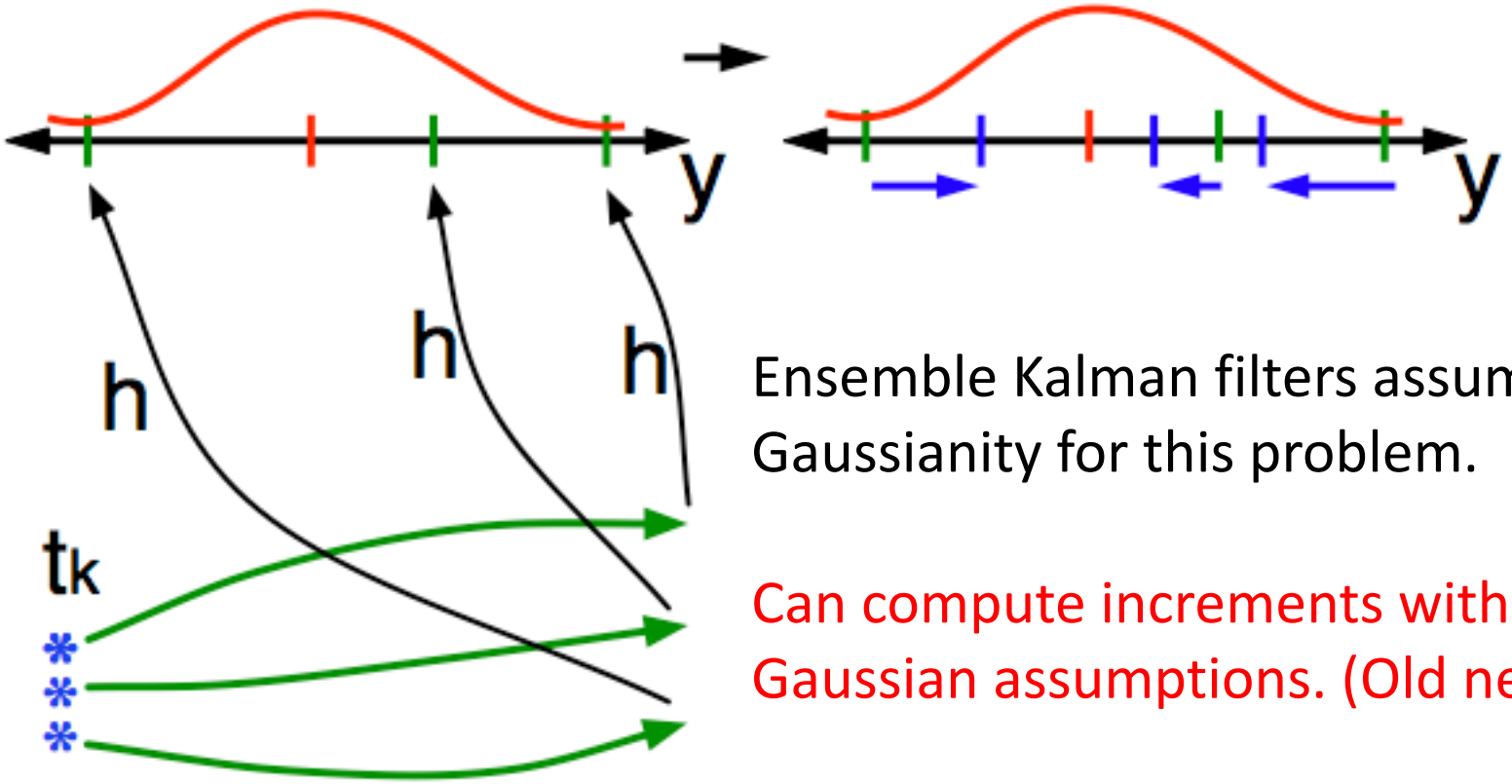
Schematic of a Sequential Ensemble Filter

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



Schematic of a Sequential Ensemble Filter

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

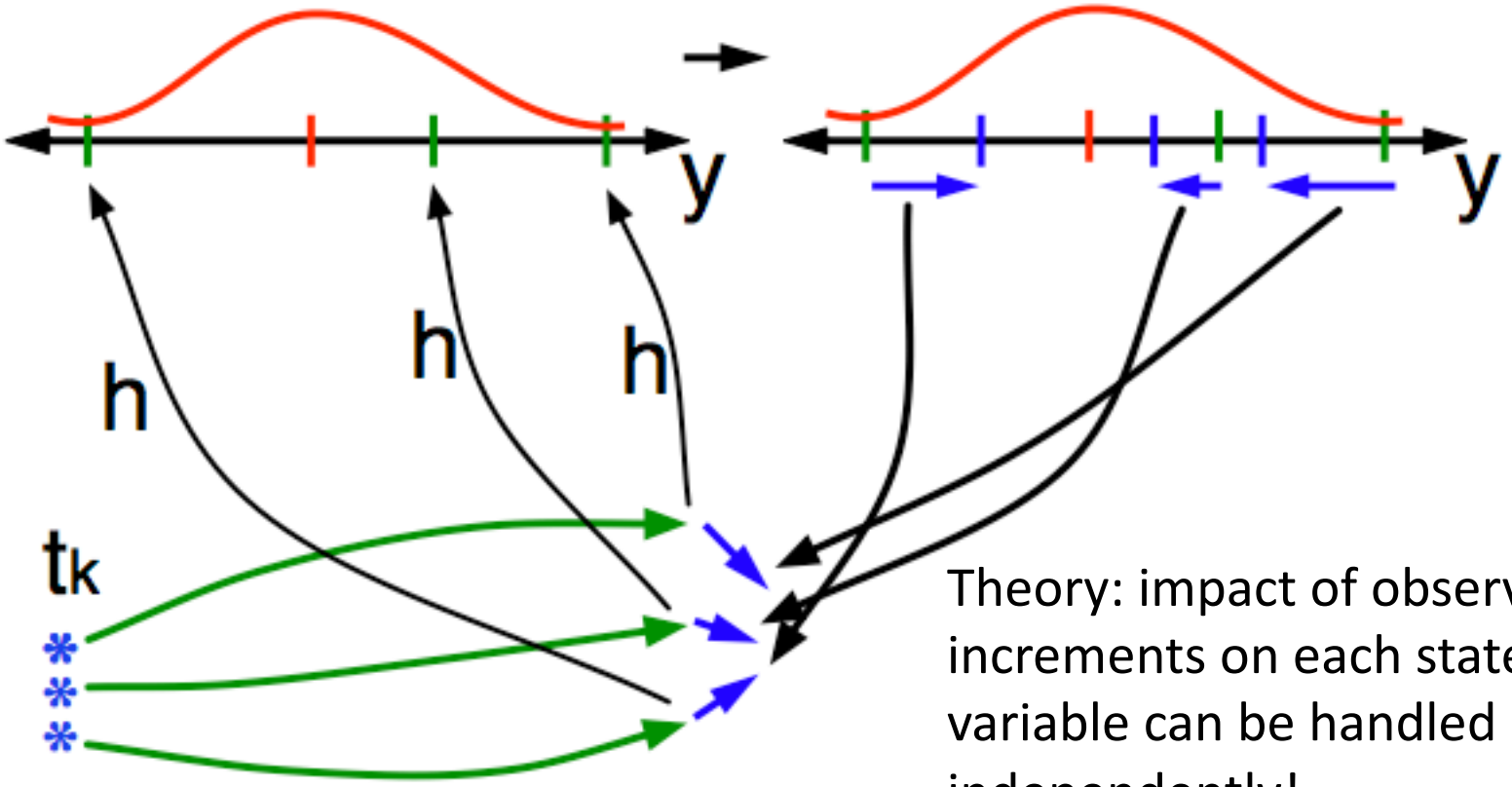


Ensemble Kalman filters assume Gaussianity for this problem.

Can compute increments without Gaussian assumptions. (Old news).

Schematic of a Sequential Ensemble Filter

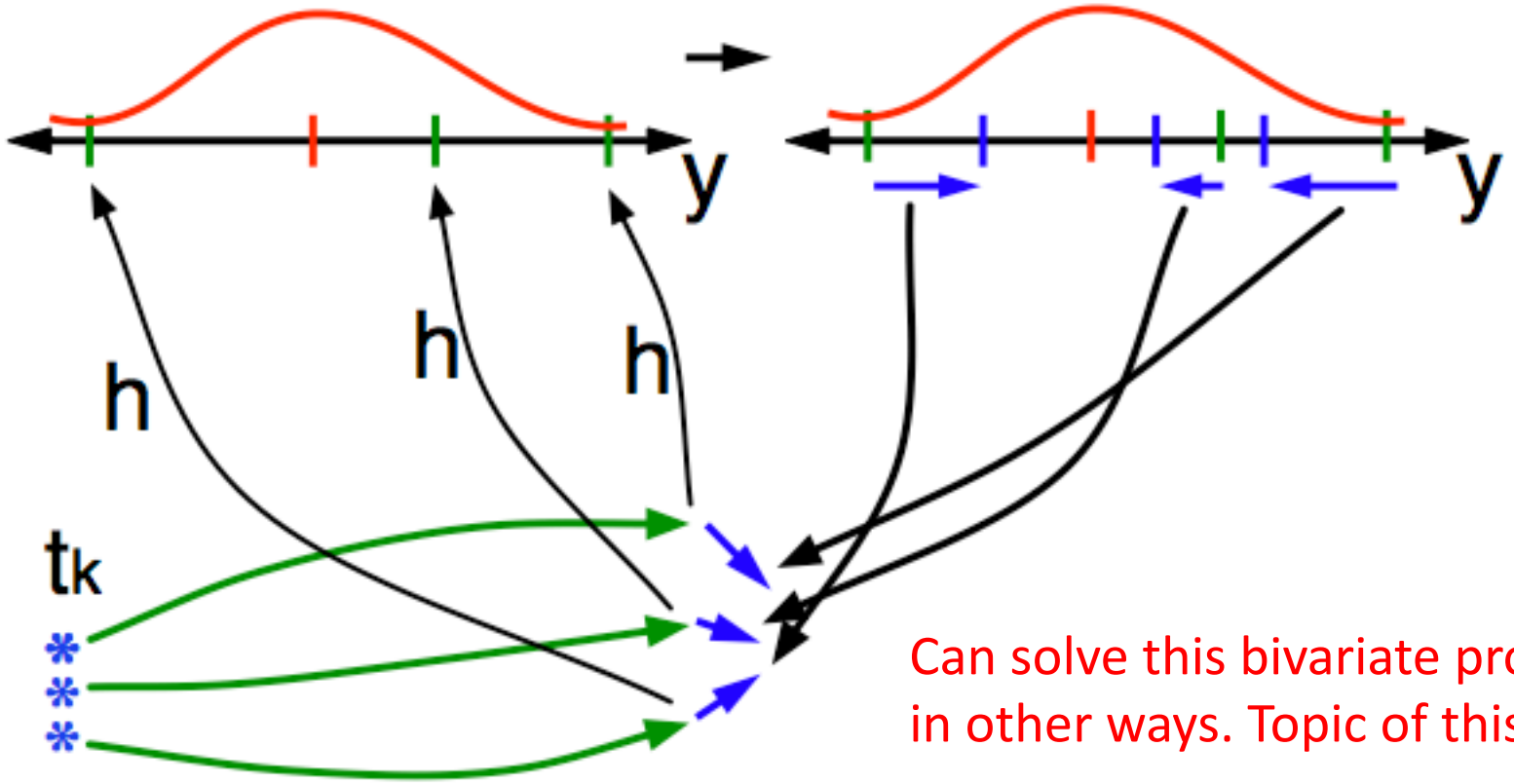
5. Use ensemble samples of y and each state variable to linearly regress observation increments onto state variable increments.



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

Schematic of a Sequential Ensemble Filter

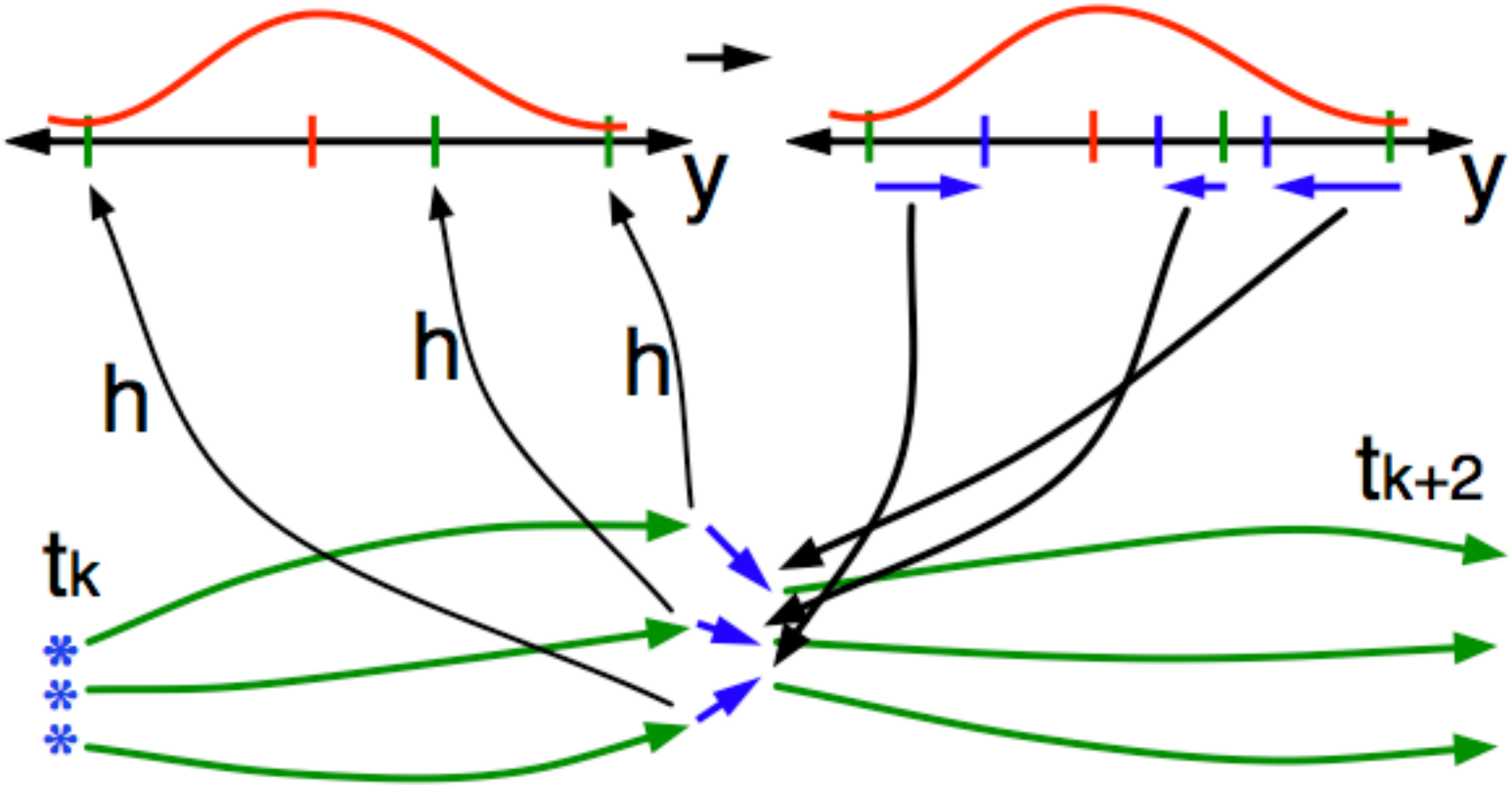
5. Use ensemble samples of y and each state variable to linearly regress observation increments onto state variable increments.



Can solve this bivariate problem in other ways. Topic of this talk.

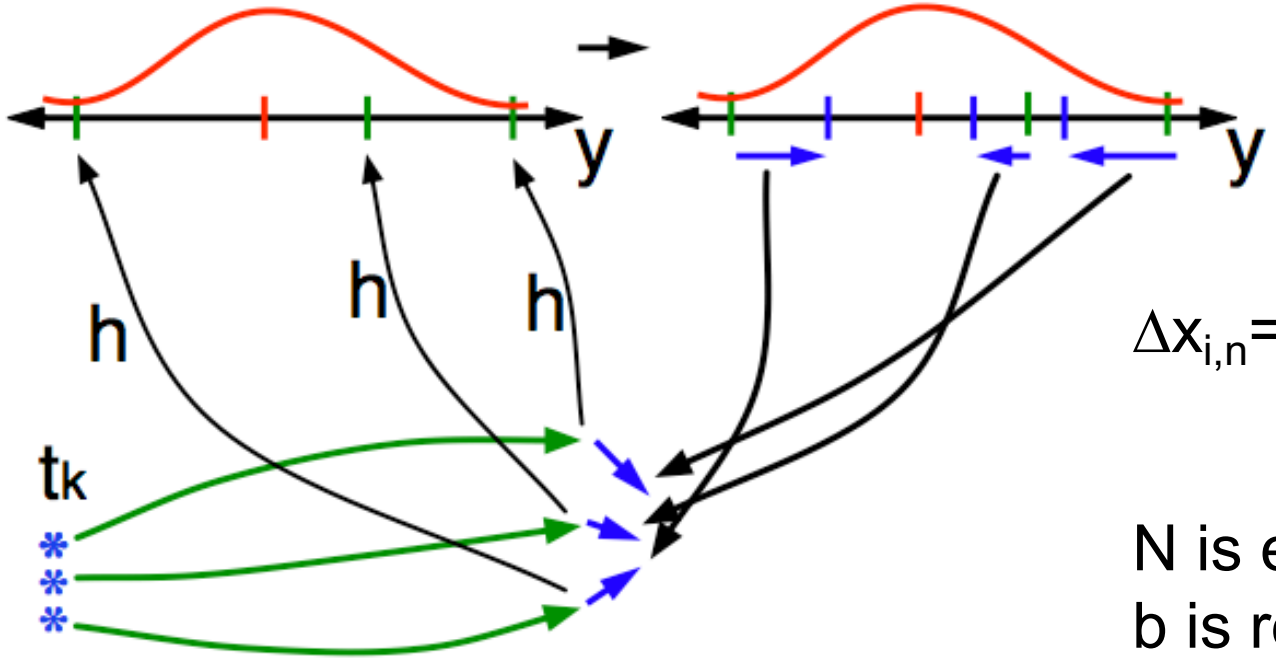
Schematic of a Sequential Ensemble Filter

6. When all ensemble members for each state variable are updated, there is a new analysis. Integrate to time of next observation …



Focus on the Regression Step

Standard ensemble filters just use bivariate sample linear regression to compute state increments.

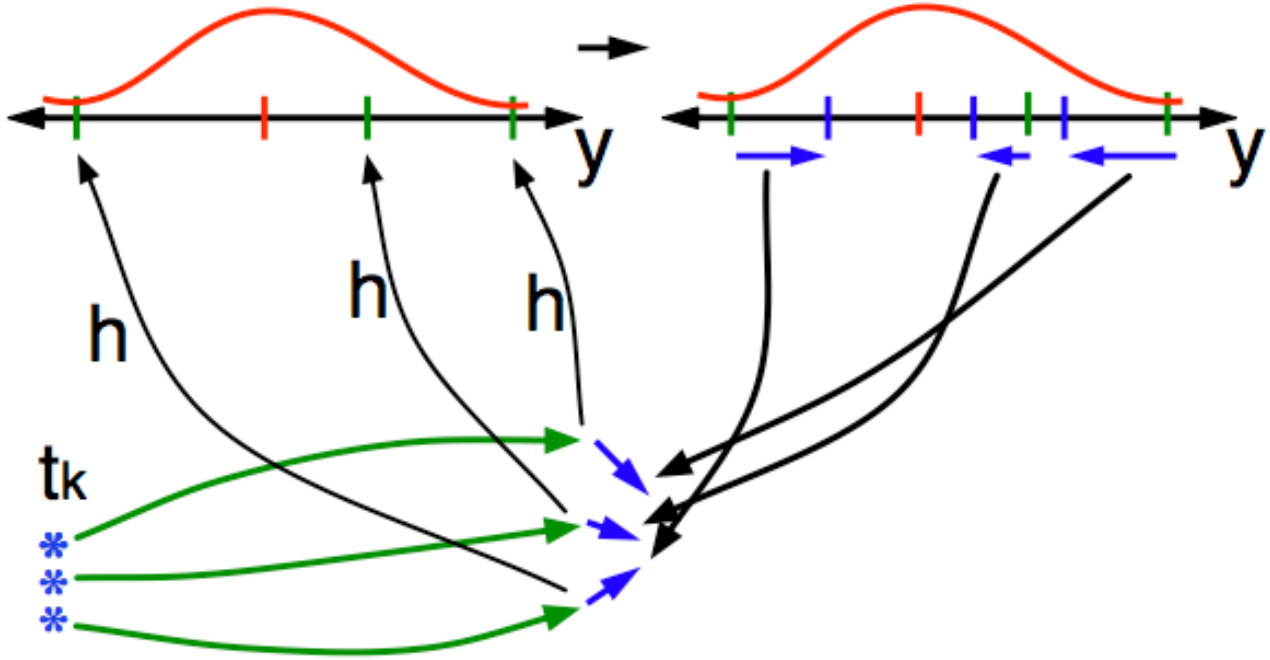


$\Delta x_{i,n} = b \Delta y_n$, $n=1,\ldots N$.

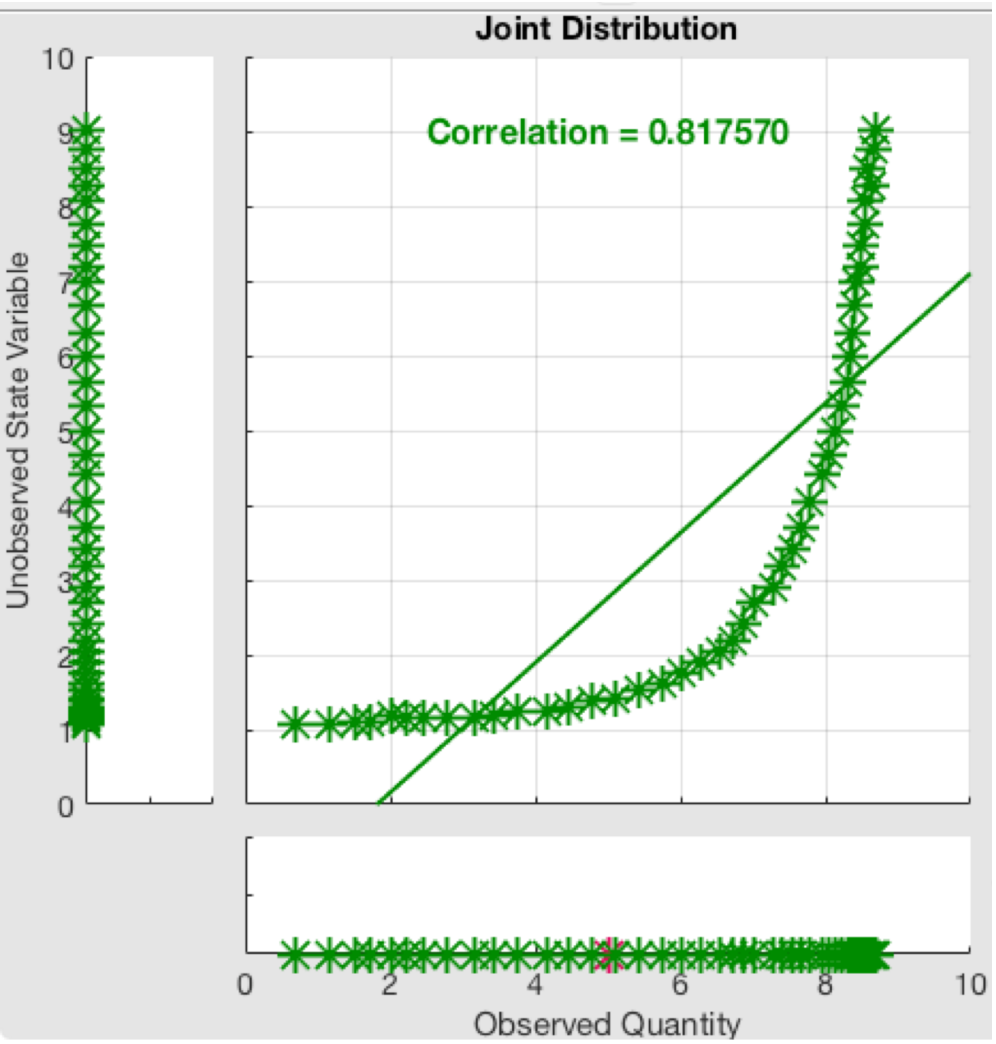
N is ensemble size.
b is regression coefficient.

Focus on the Regression Step

Examine another way to increment state given observation increments; still bivariate.



Nonlinear Regression Example

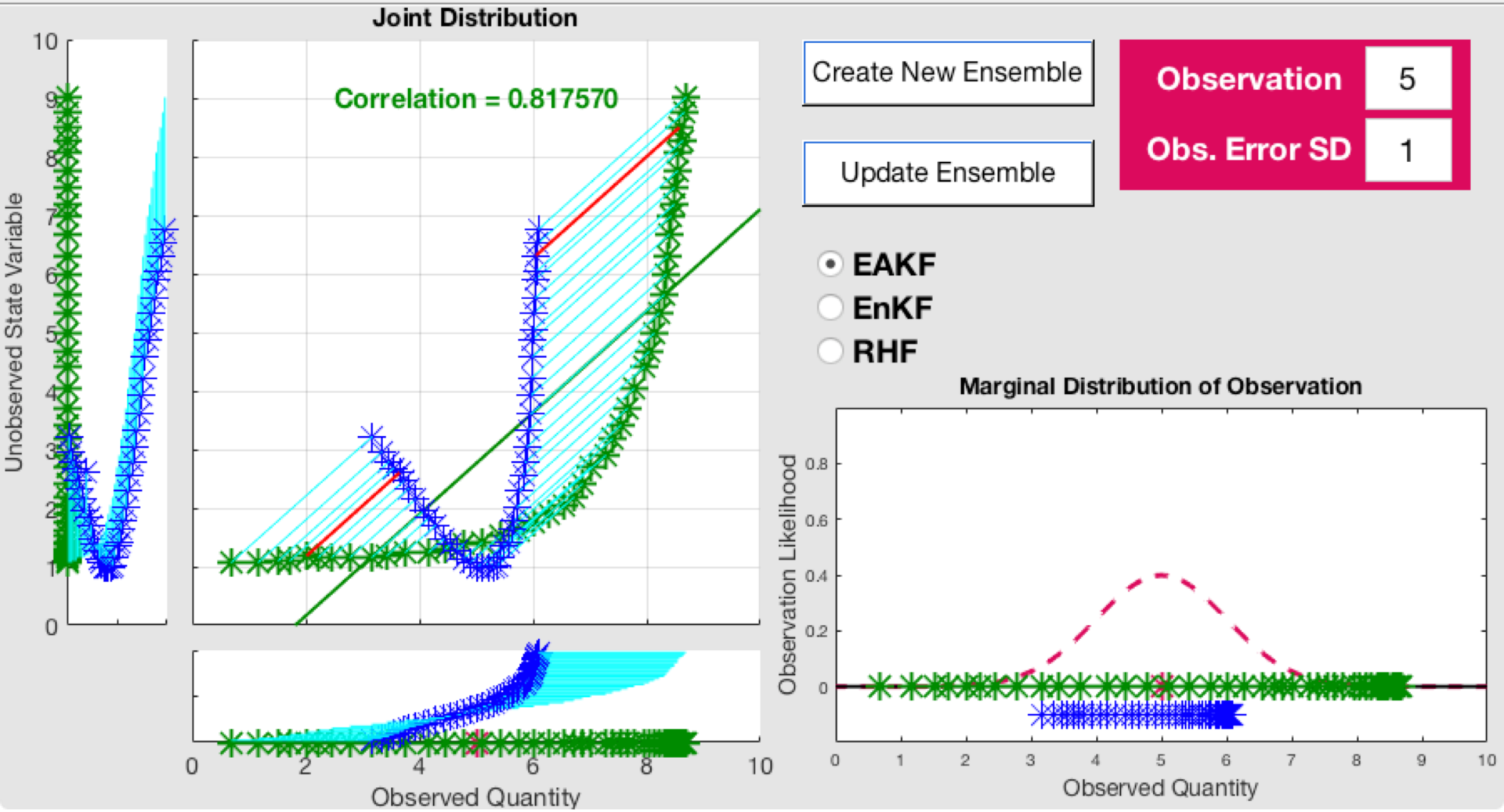


Try to exploit nonlinear prior relation between a state variable and an observation.

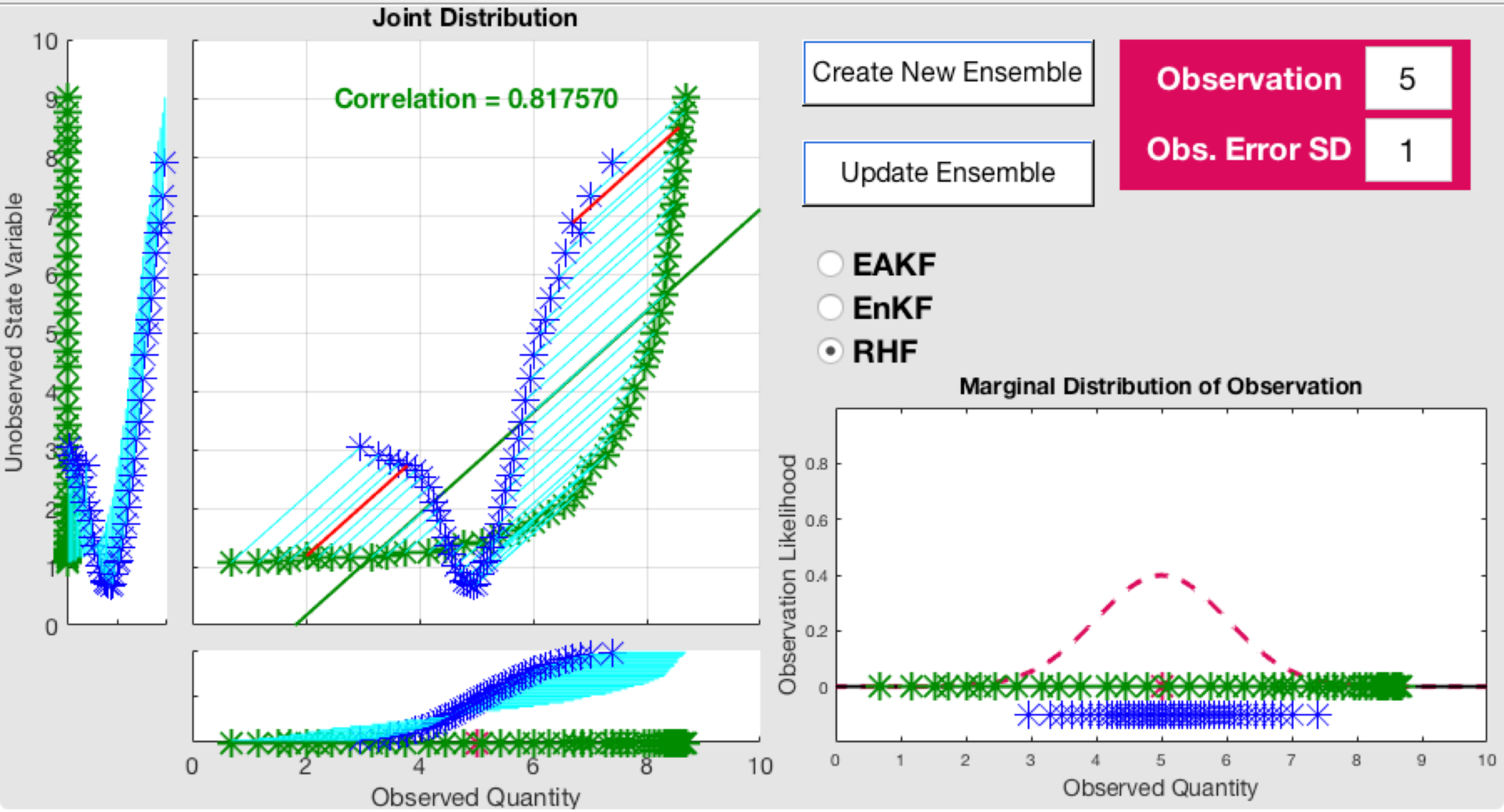
Example: Observation y~log(x).

Also relevant for variables that are log transformed for boundedness (like concentrations).

Standard Ensemble Adjustment Filter (EAKF)

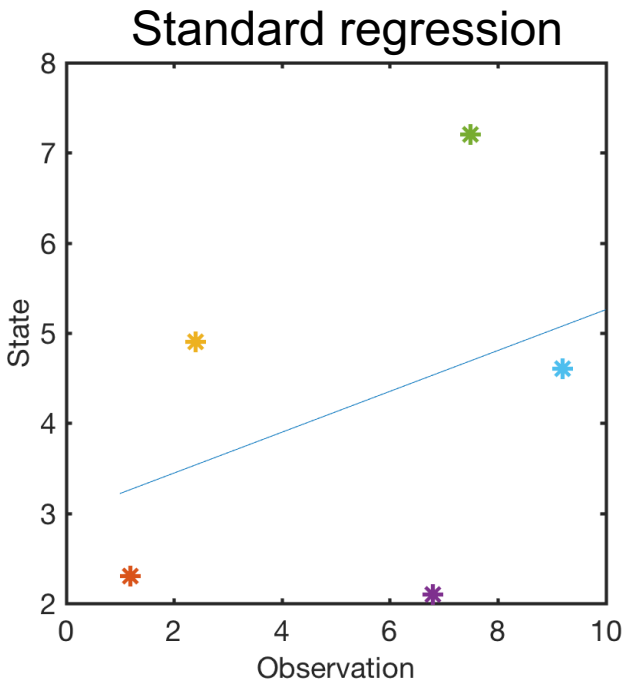


Standard Rank Histogram Filter (RHF)

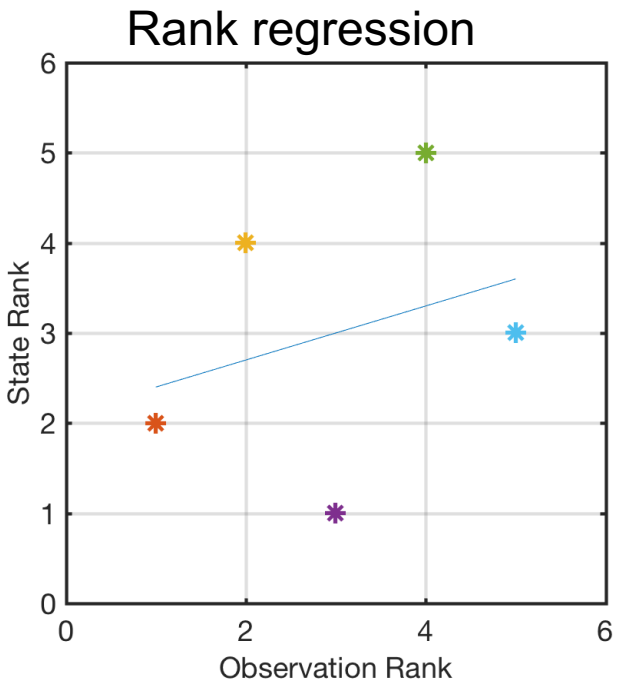


Rank Regression

1. Convert bivariate ensemble to bivariate rank ensemble.
2. Do least squares on bivariate rank ensemble.

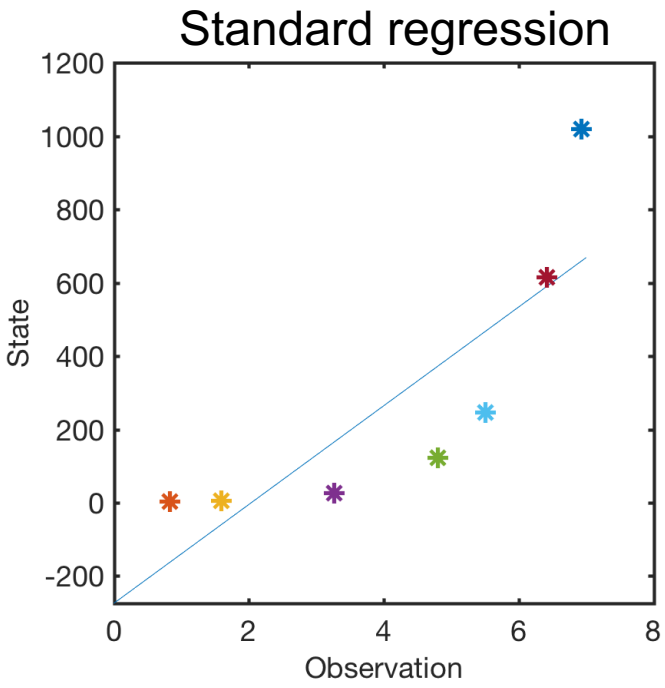


Noisy Relation.

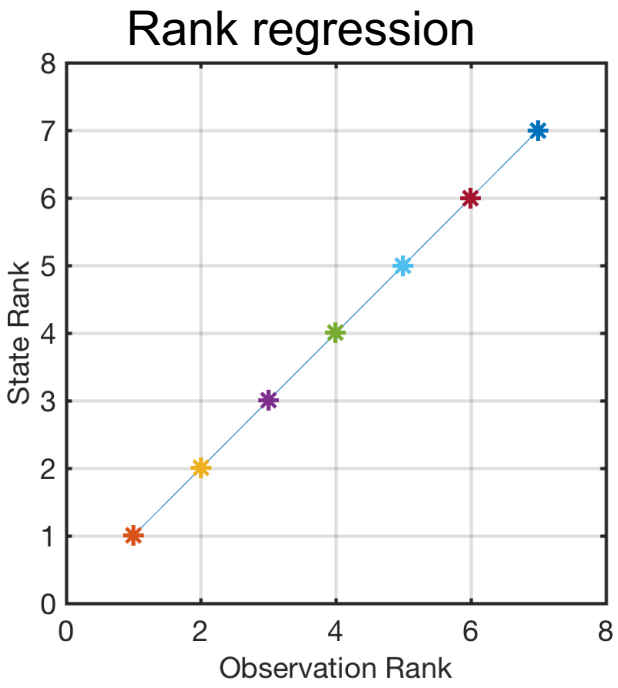


Rank Regression

1. Convert bivariate ensemble to bivariate rank ensemble.
2. Do least squares on bivariate rank ensemble.



Monotonic relation.



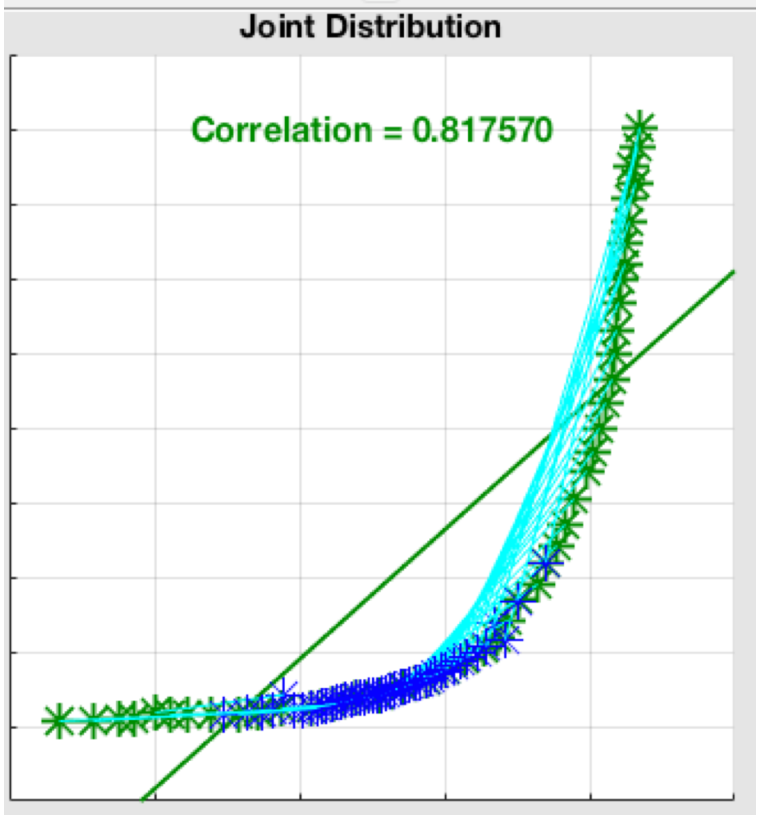
Rank Regression

1. Convert bivariate ensemble to bivariate rank ensemble.
2. Do least squares on bivariate rank ensemble.
3. Convert observation posteriors to rank.
4. Regress rank increments onto state ranks.

Rank Regression

1. Convert bivariate ensemble to bivariate rank ensemble.
2. Do least squares on bivariate rank ensemble.
3. Convert observation posteriors to rank.
4. Regress rank increments onto state ranks.
5. Convert posterior state ranks to state values.

Rank Regression Example



Rank regression with RHF for observation marginal.

Follows monotonic ensemble prior ‘exactly’.

Compare Six 'Treatments'

Each 'treatment' pairs an observation space method:

1. Deterministic Ensemble Adjustment Kalman Filter (EAKF),
2. Stochastic Ensemble Kalman Filter (EnKF),
3. Non-Gaussian Rank Histogram Filter (RHF)

With a regression method:

A. Standard linear regression,

B. Rank regression.

Empirically Tuned Localization and Inflation

Select the Gaspari Cohn localization half-width:

0.125 0.15, 0.175, 0.2, 0.25, 0.4, infinite.

and fixed multiplicative inflation:

1.0, 1.02, 1.04, 1.08, 1.16, 1.32, 1.64.

that gives minimum time mean posterior RMSE.

(7x7=49 possibilities checked).

Find best localization/inflation pair for each of the six treatments.

Lorenz-96 Tests

40 state variables.

40 randomly located 'observing stations' around unit circle.
(x_o is state interpolated to a station location).

Explored 4 different forward operators:

Identity: $y_o = x_o,$

Square root: $y_o = sgn(x_o)\sqrt{|x_o|}$,

Square: $y_o = x_o^2$,

Cube: $y_o = x_o^3$.

Lorenz-96 Tests: Statistical Significance

Select rank regression treatment with lowest RMSE:

This picks inflation, localization, and observation space filter (EAKF, EnKF, or RHF).

Select linear regression treatment with lowest RMSE.

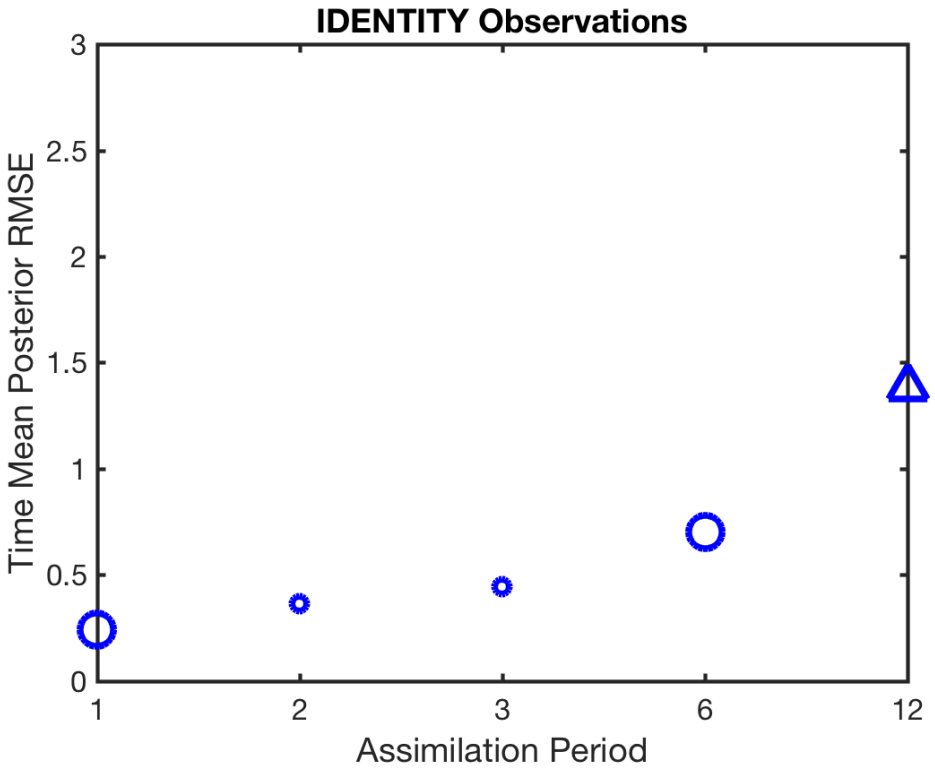
Apply each to ten 5000-step assimilation cases.

Result is 10 pairs of RMSE for rank versus linear regression.

Paired T-test used to establish significance of RMSE differences.

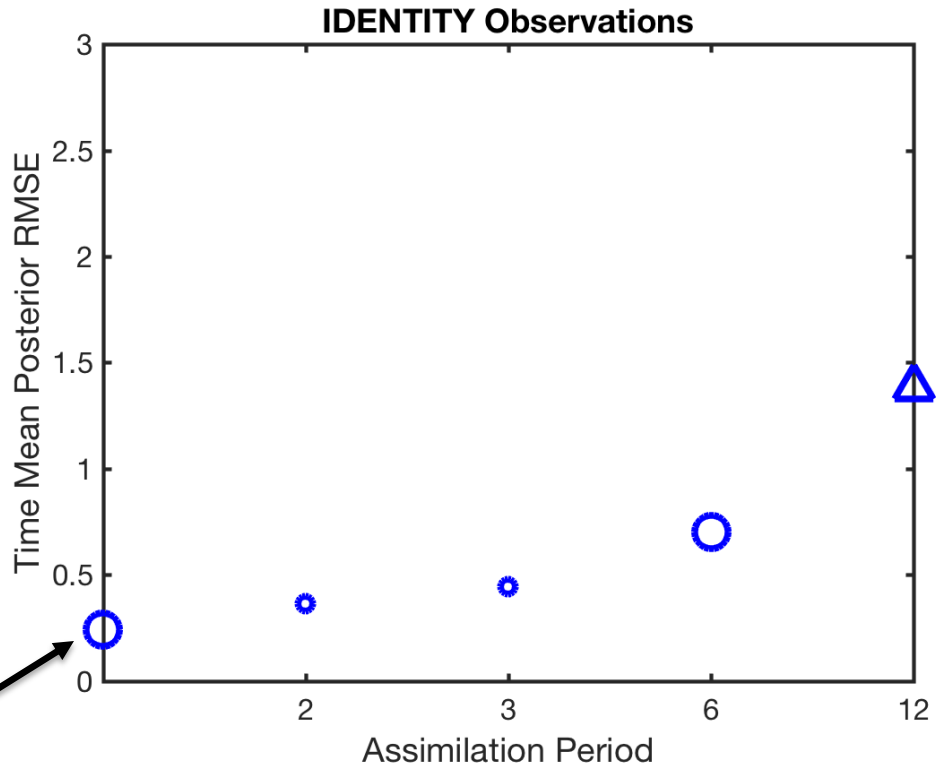
Identity Forward Operator Results

Observation error variance 1.0



Identity Forward Operator Results

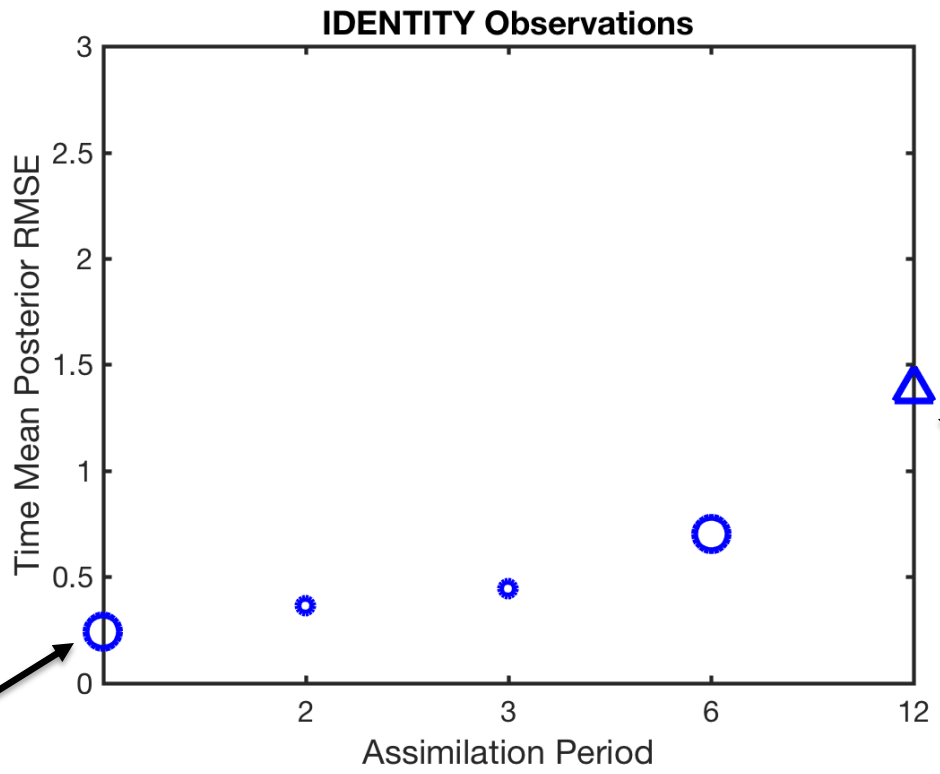
Observation error variance 1.0



Circle:
EnKF best

Identity Forward Operator Results

Observation error variance 1.0

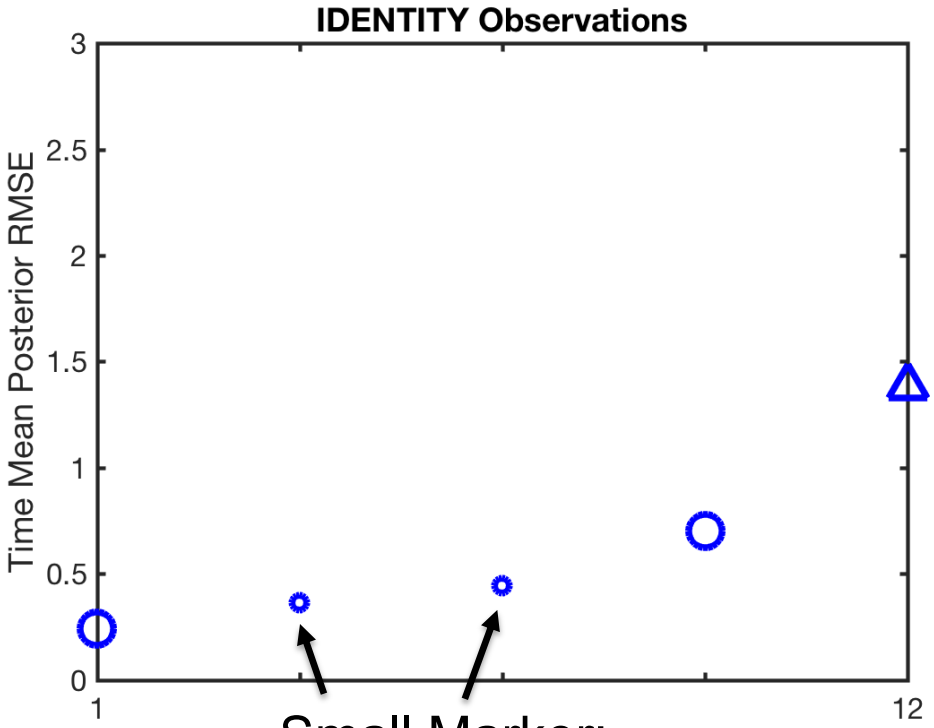


Triangle:
RHF best

Circle:
EnKF best

Identity Forward Operator Results

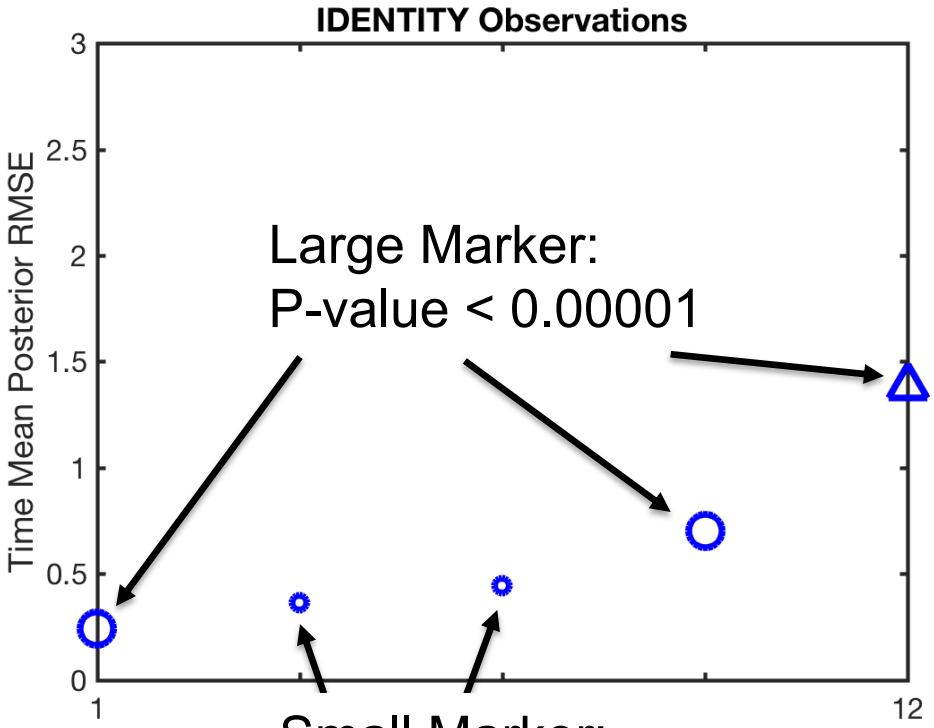
Observation error variance 1.0



Small Marker:
P-value > 0.01,
Not significant

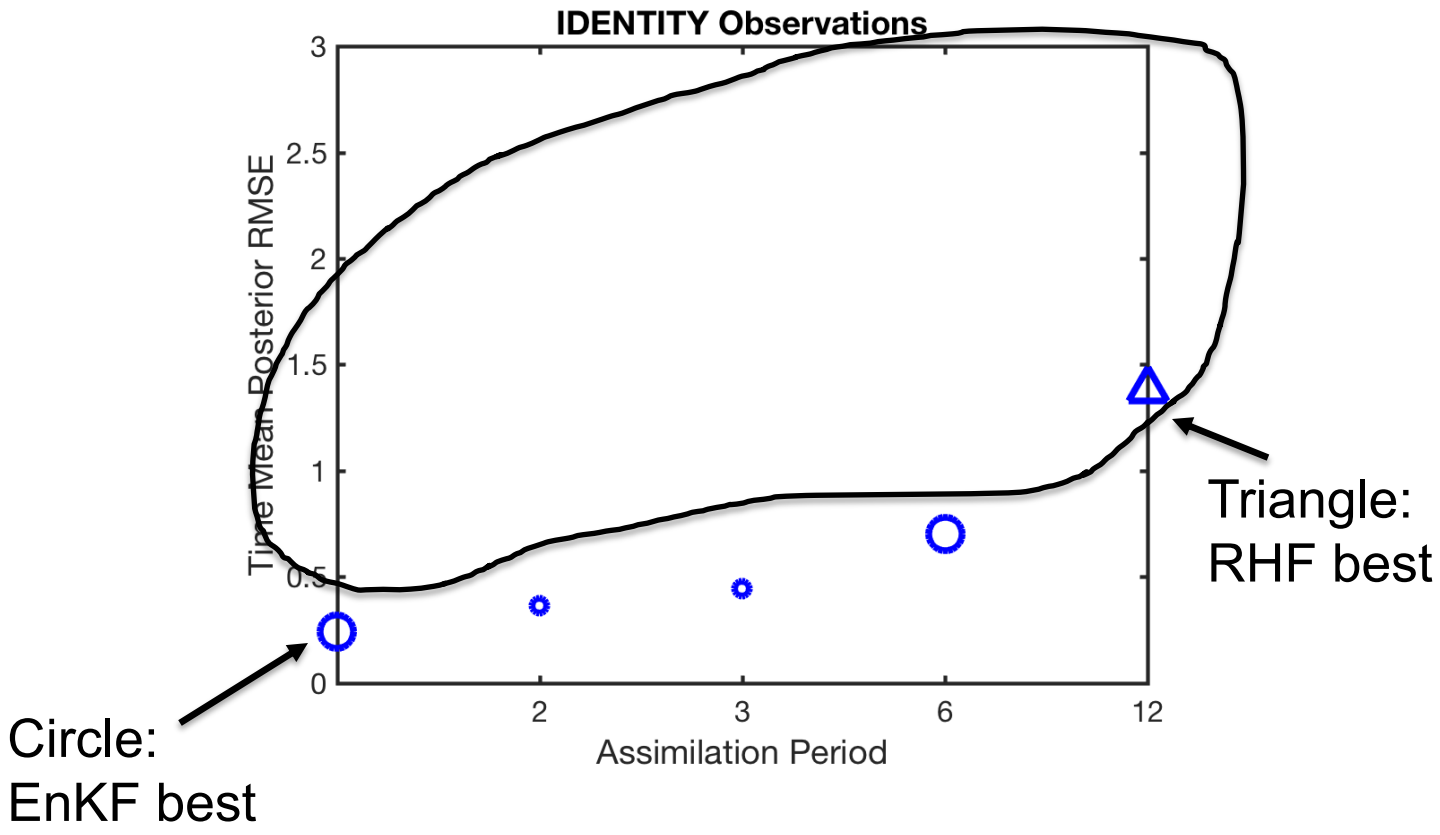
Identity Forward Operator Results

Observation error variance 1.0



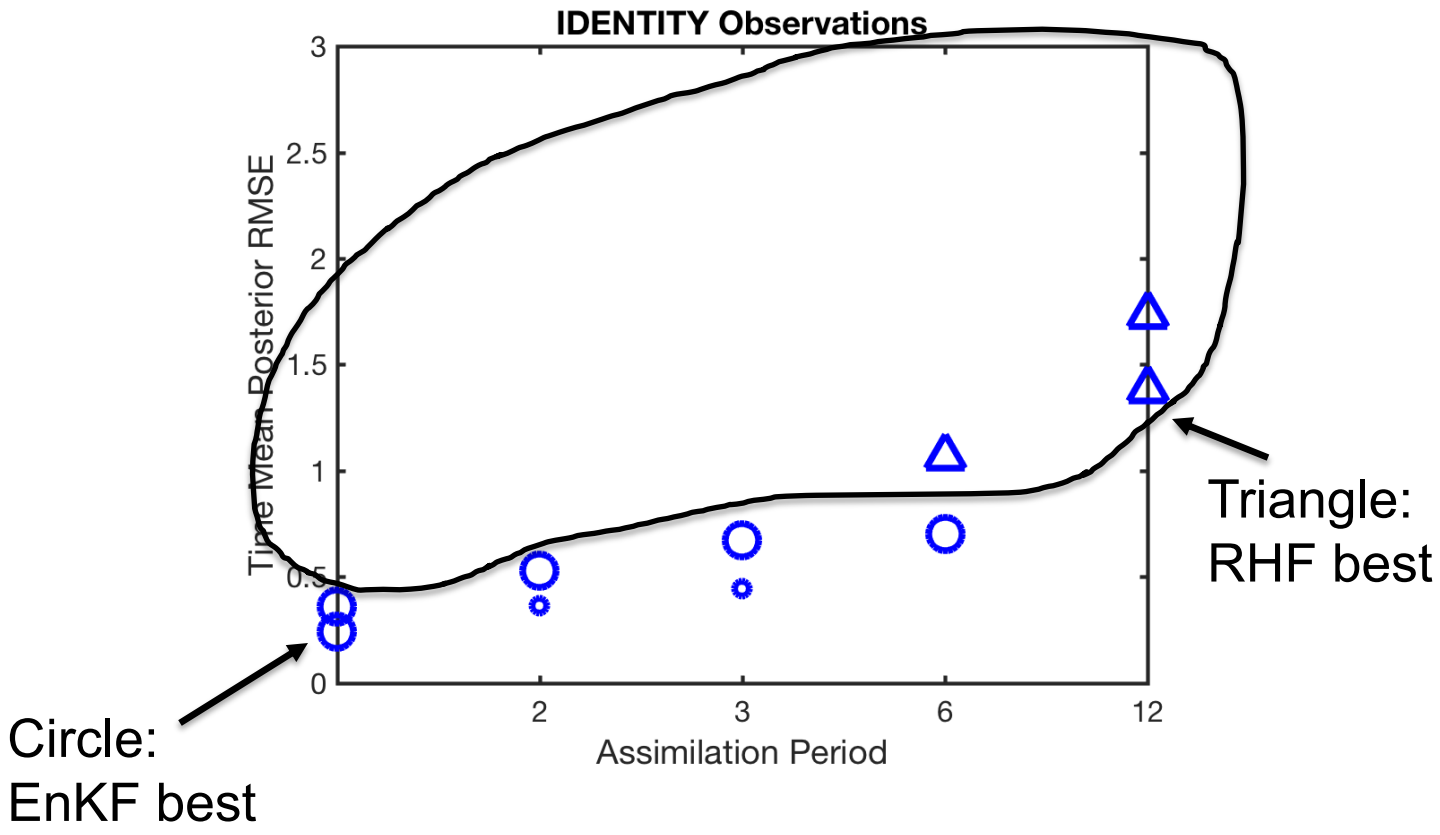
Identity Forward Operator Results

Observation error variance 1.0



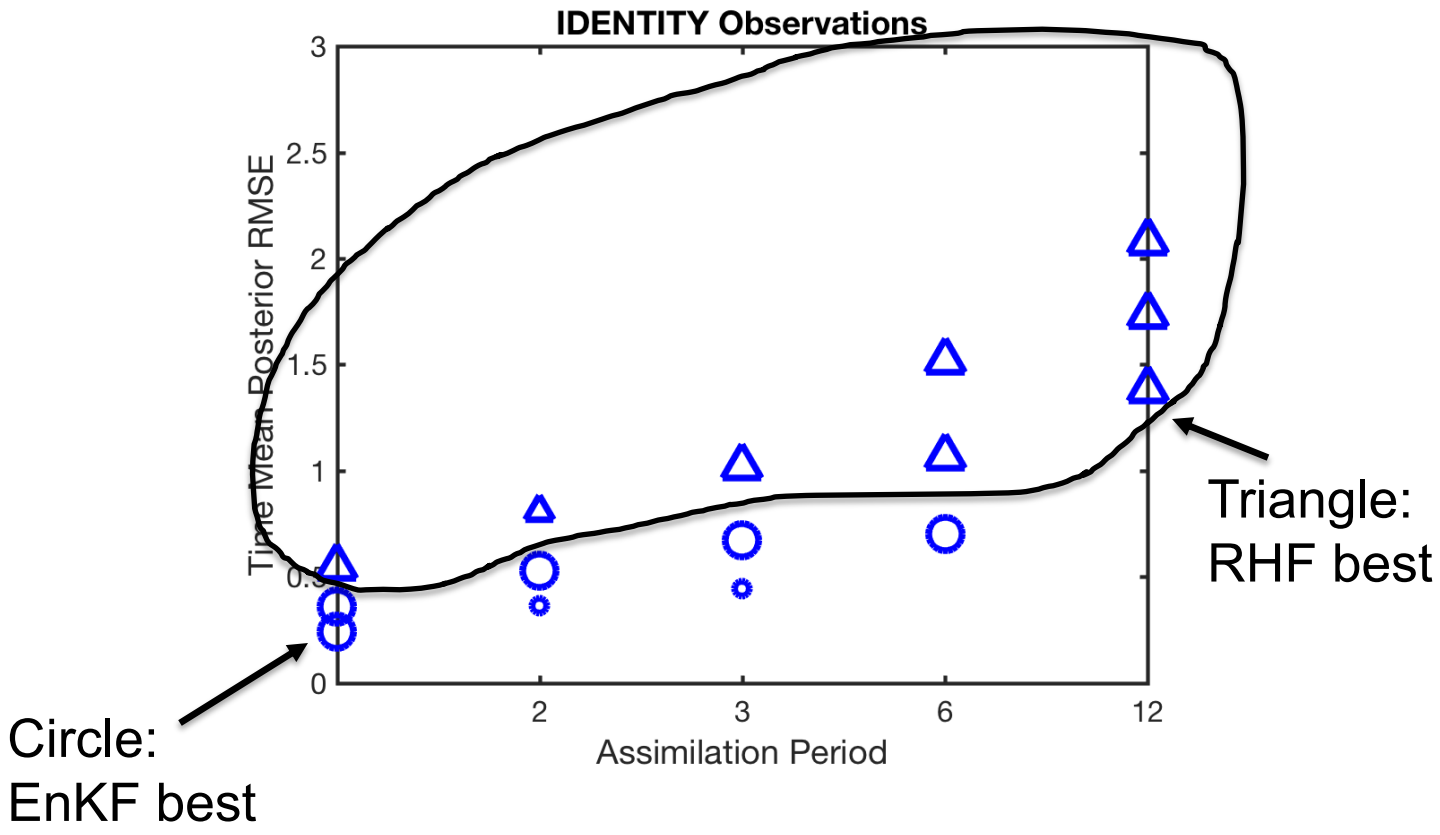
Identity Forward Operator Results

Observation error variance 2.0



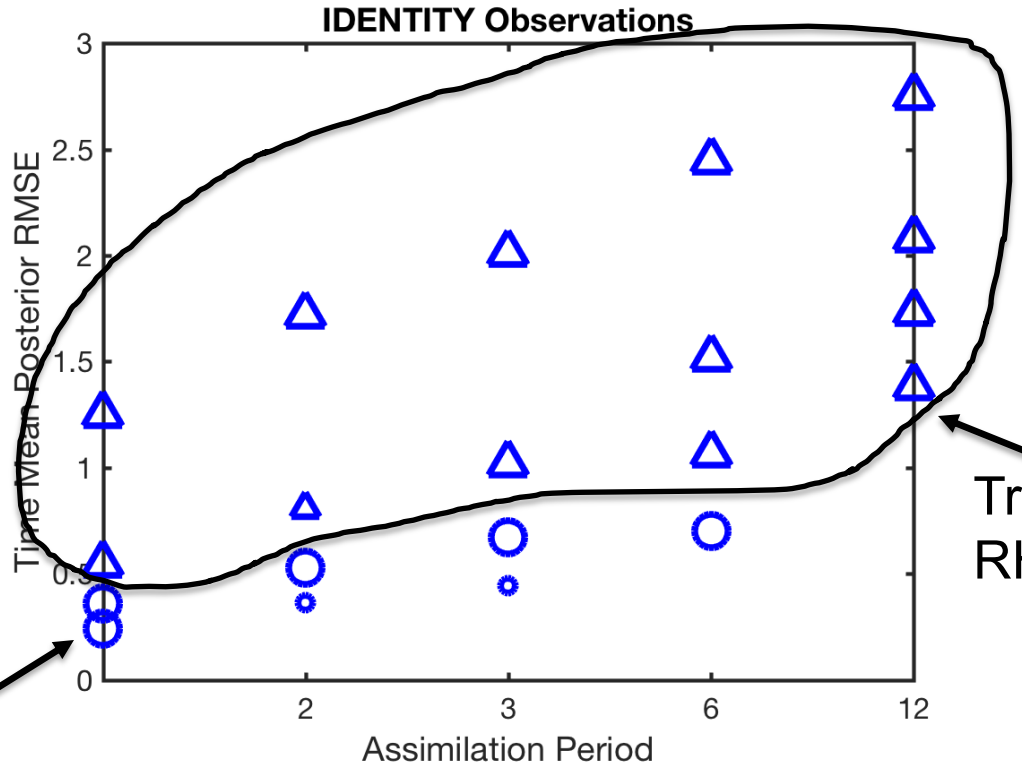
Identity Forward Operator Results

Observation error variance 4.0



Identity Forward Operator Results

Observation error variance 16.0

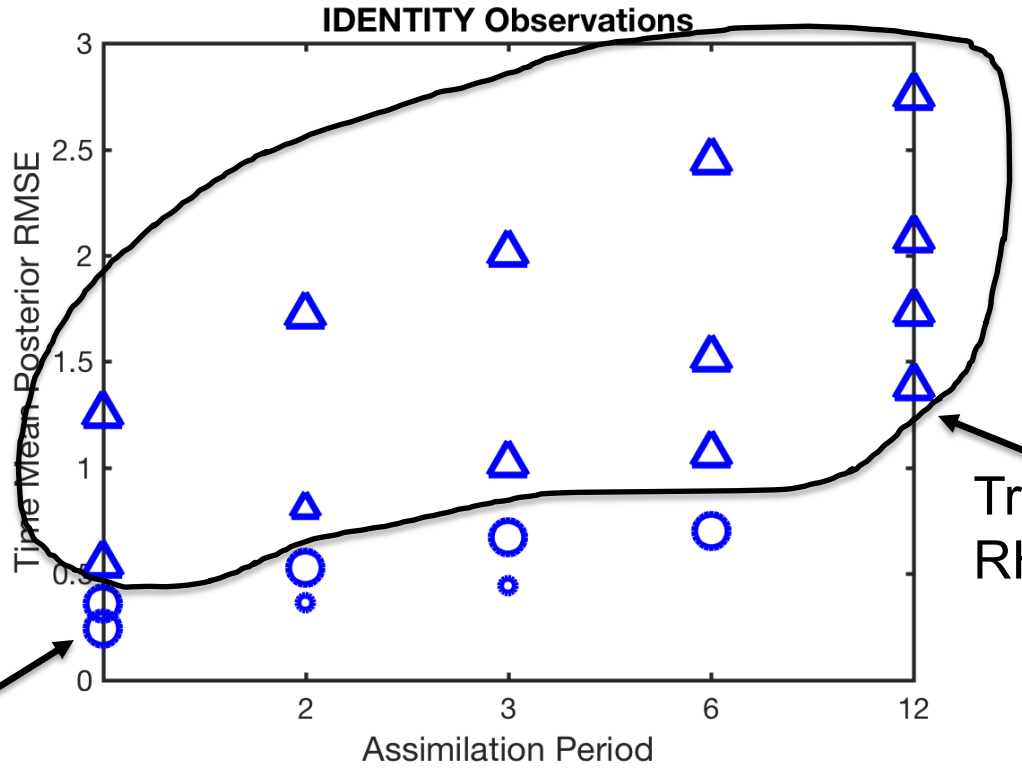


Triangle:
RHF best

Circle:
EnKF best

Summary: Identity Forward Operator Results

RHF better for larger RMSE, EnKF for smaller.
Linear regression always better than rank regression.

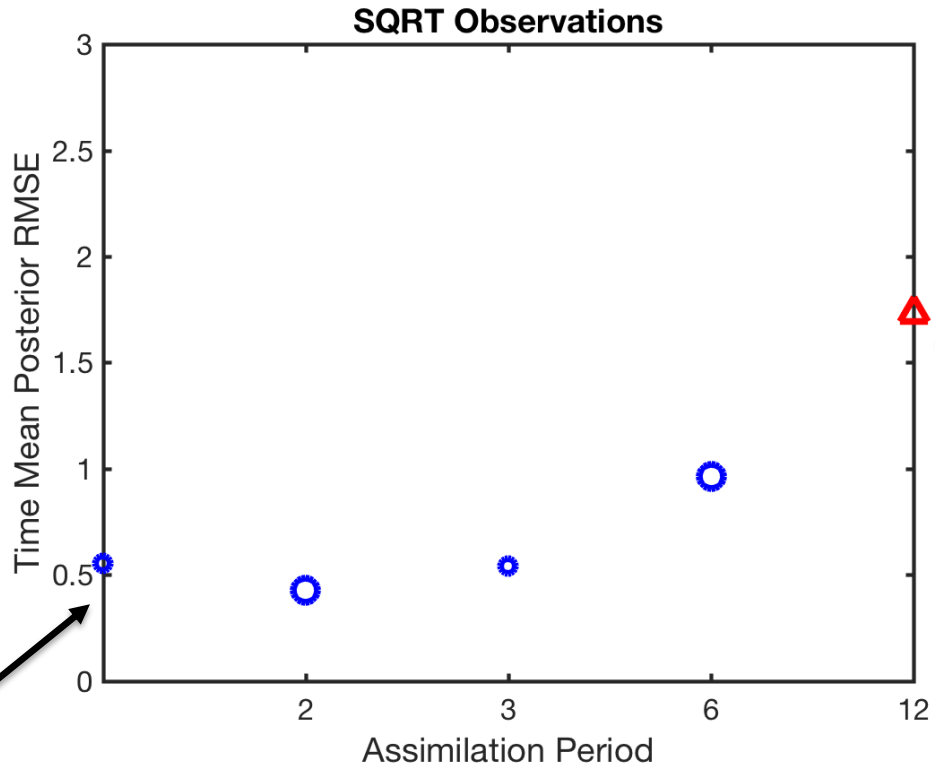


Triangle:
RHF best

Circle:
EnKF best

Square Root Forward Operator Results

Observation error variance 0.25

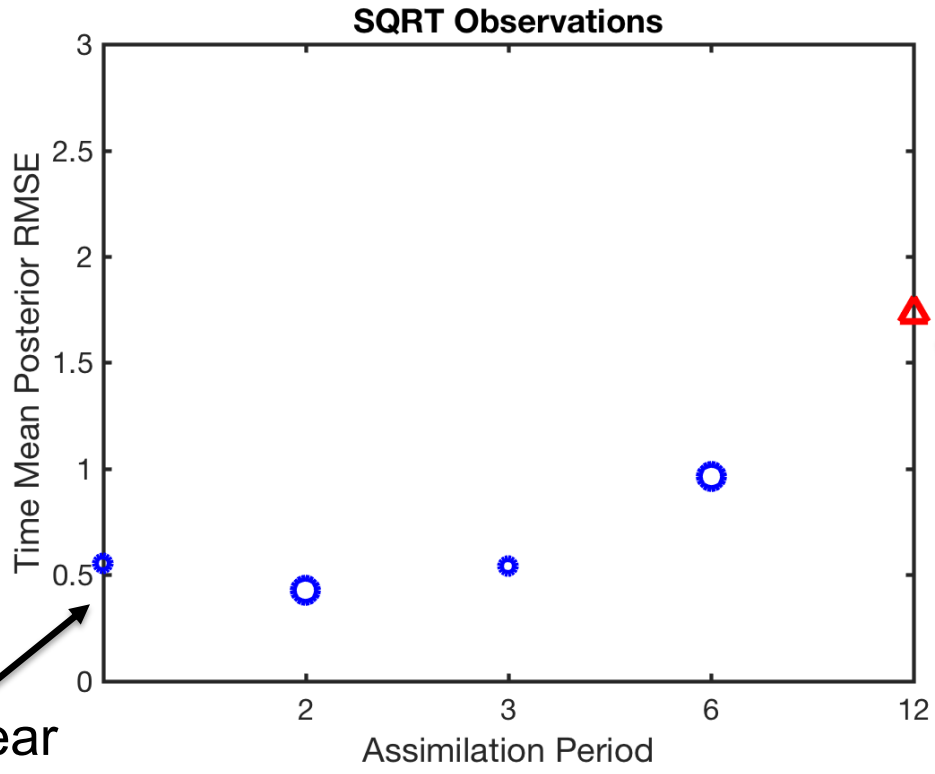


Triangle:
RHF best

Circle:
EnKF best

Square Root Forward Operator Results

Observation error variance 0.25

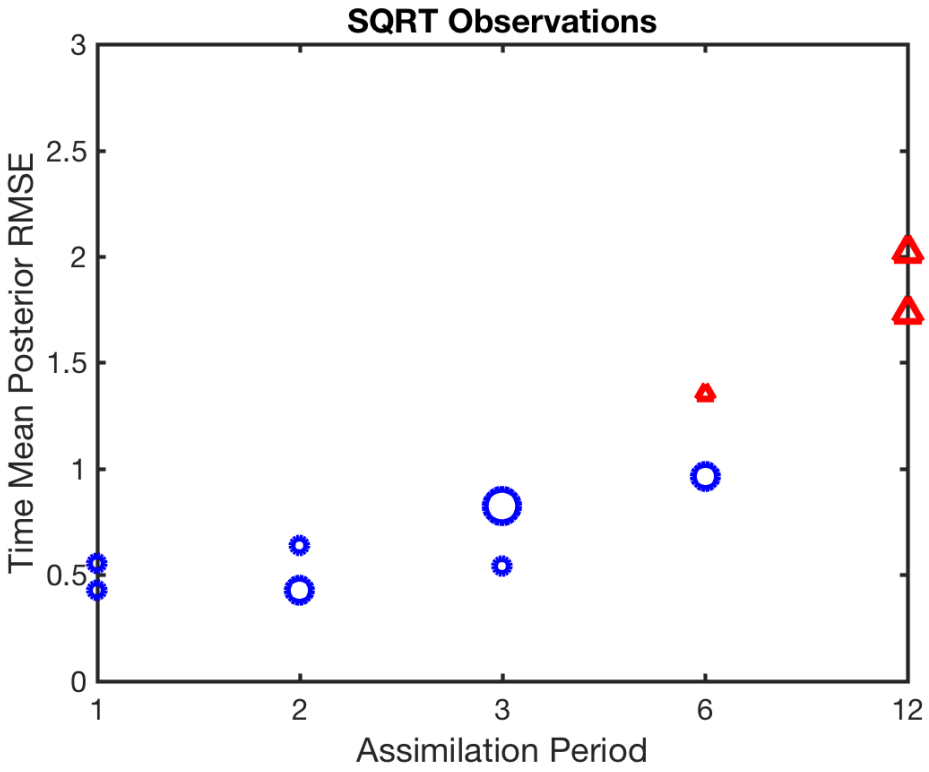


Red: Rank regression best

Blue: Linear regression best

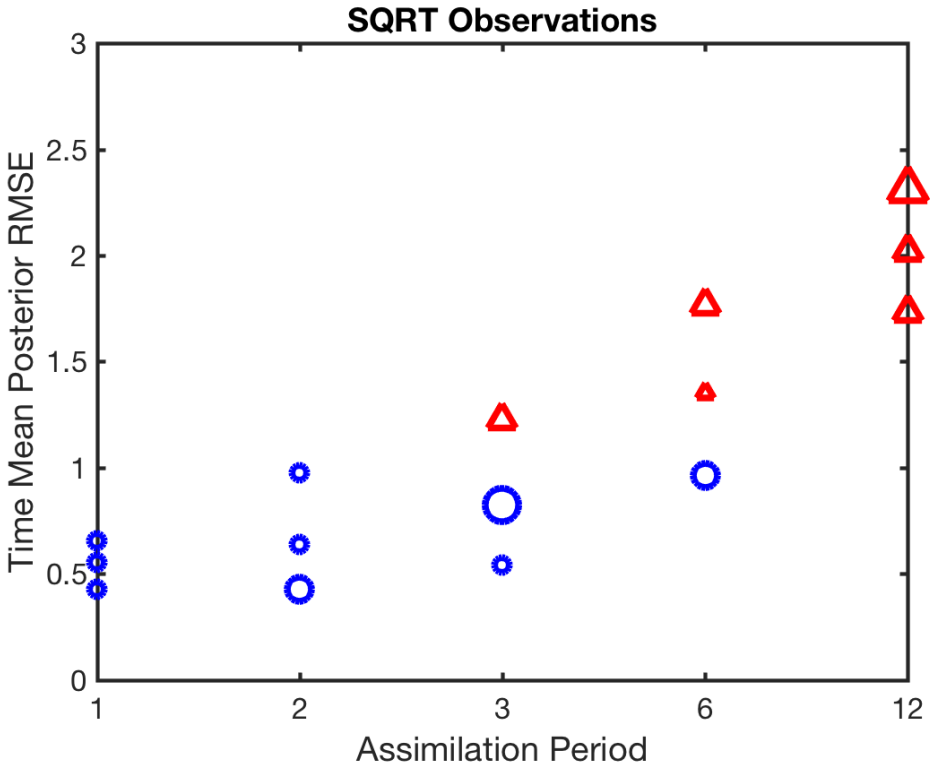
Square Root Forward Operator Results

Observation error variance 0.5



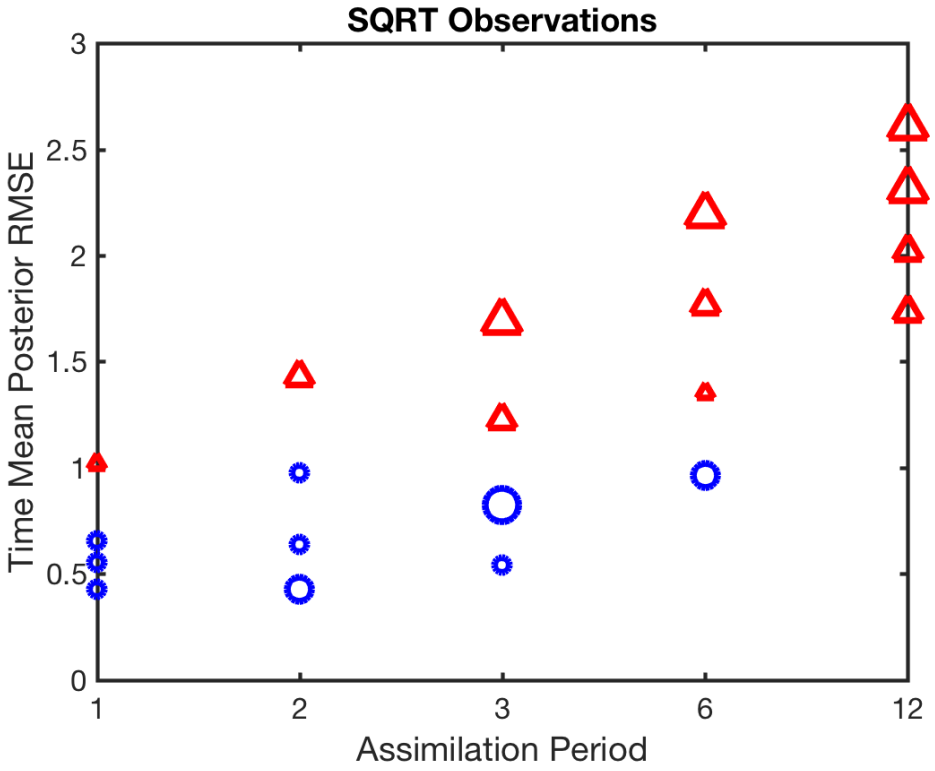
Square Root Forward Operator Results

Observation error variance 1.0



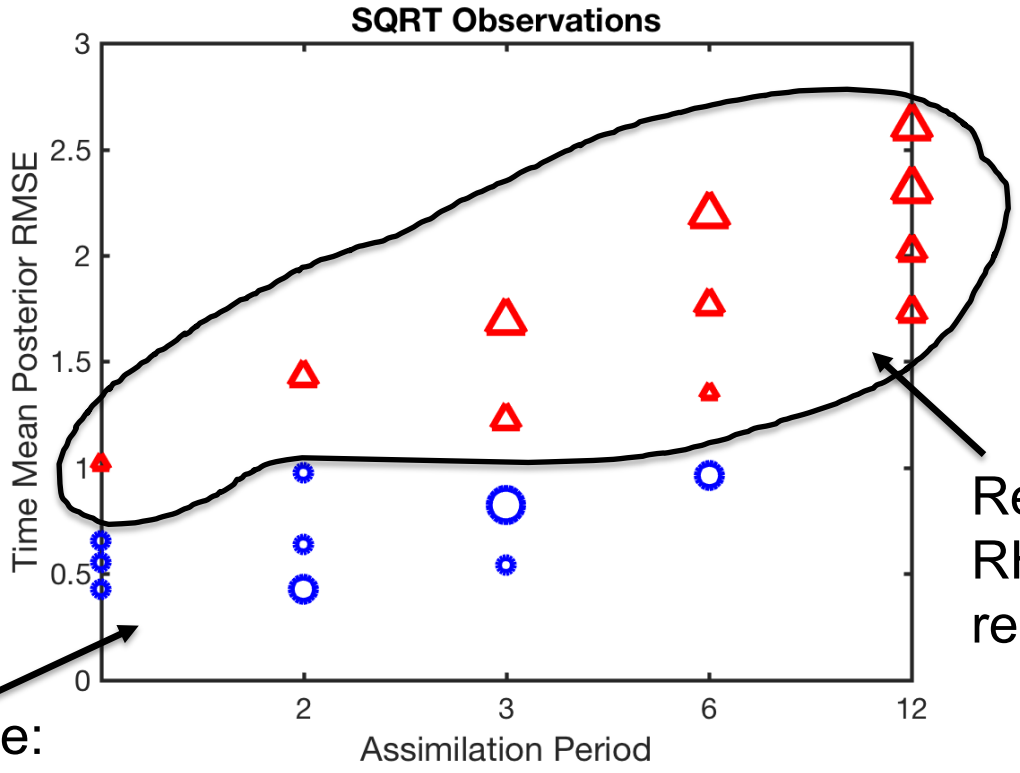
Square Root Forward Operator Results

Observation error variance 2.0



Summary: Square Root Forward Operator Results

RHF with rank regression better for larger RMSE.
EnKF with linear regression better for smaller RMSE.

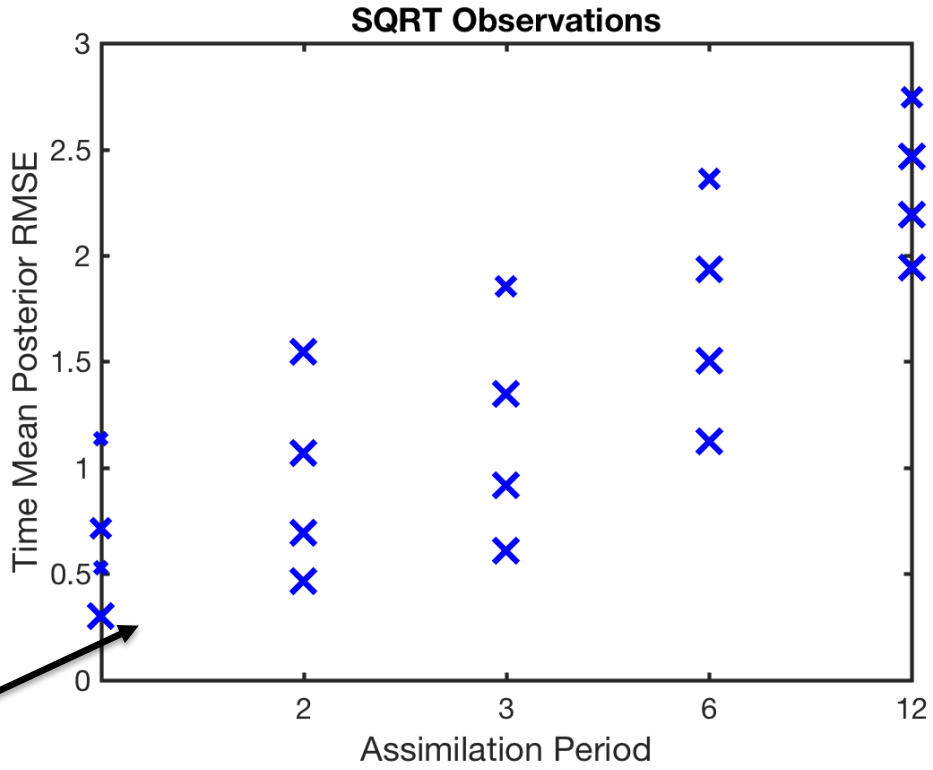


Red Triangle:
RHF with rank regression best

Blue Circle:
EnKF with linear regression best

Square Root Forward Operator: Ensemble Size

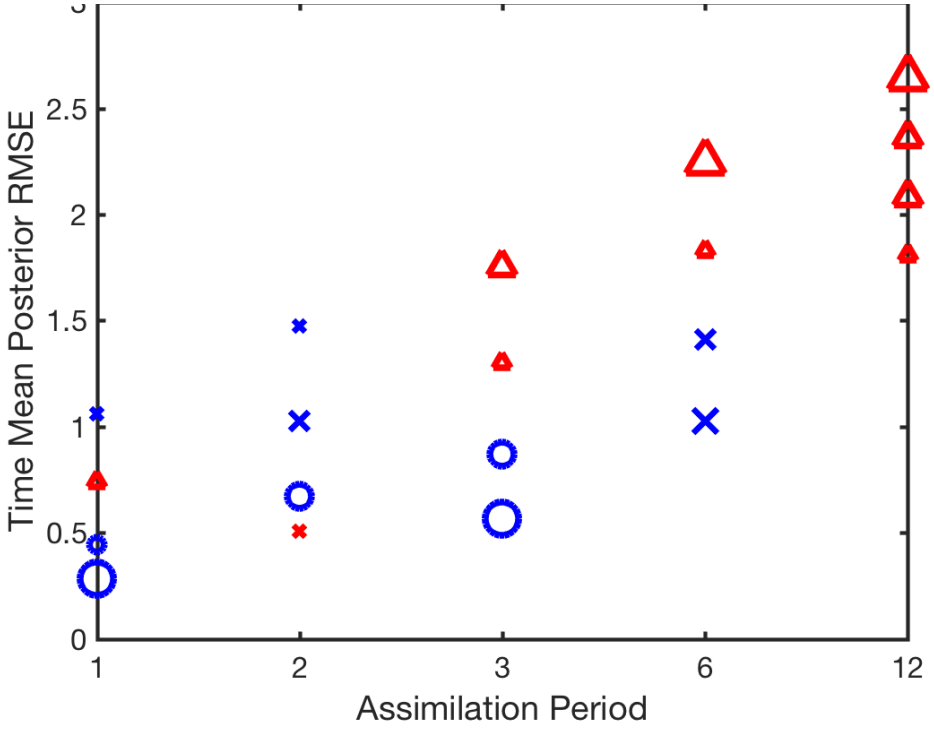
20 Members: EAKF with linear regression always best.



Blue ‘x’:
EAKF with linear regression best

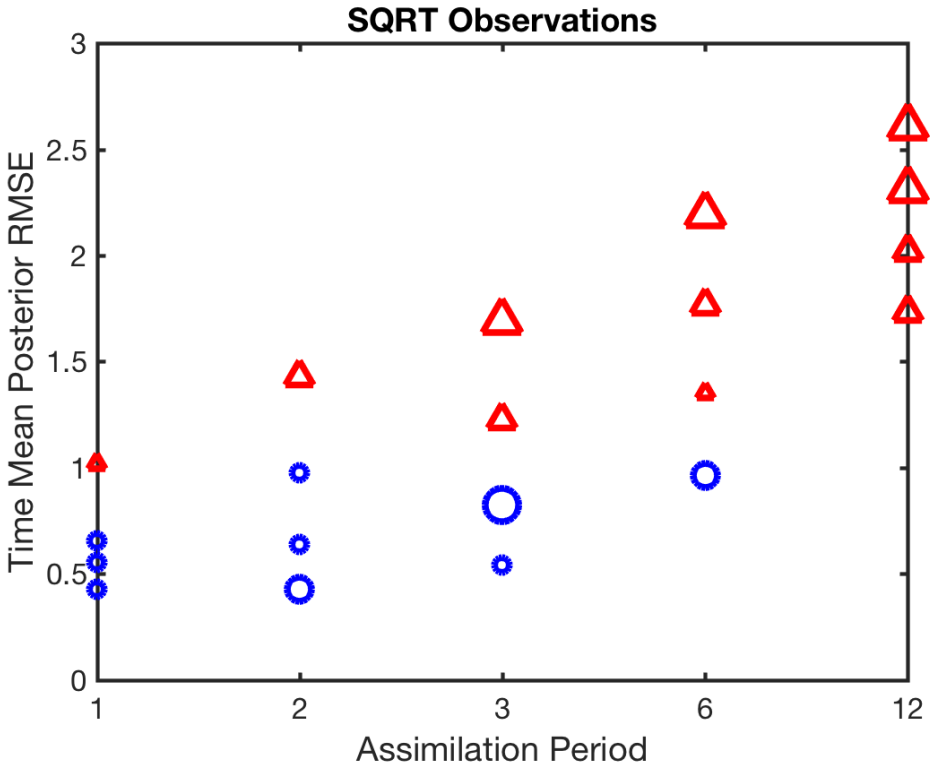
Square Root Forward Operator: Ensemble Size

40 Members: RHF with rank regression for large RMSE.
EAKF/linear regression for intermediate RMSE.
EnKF/linear for smaller RMSE.



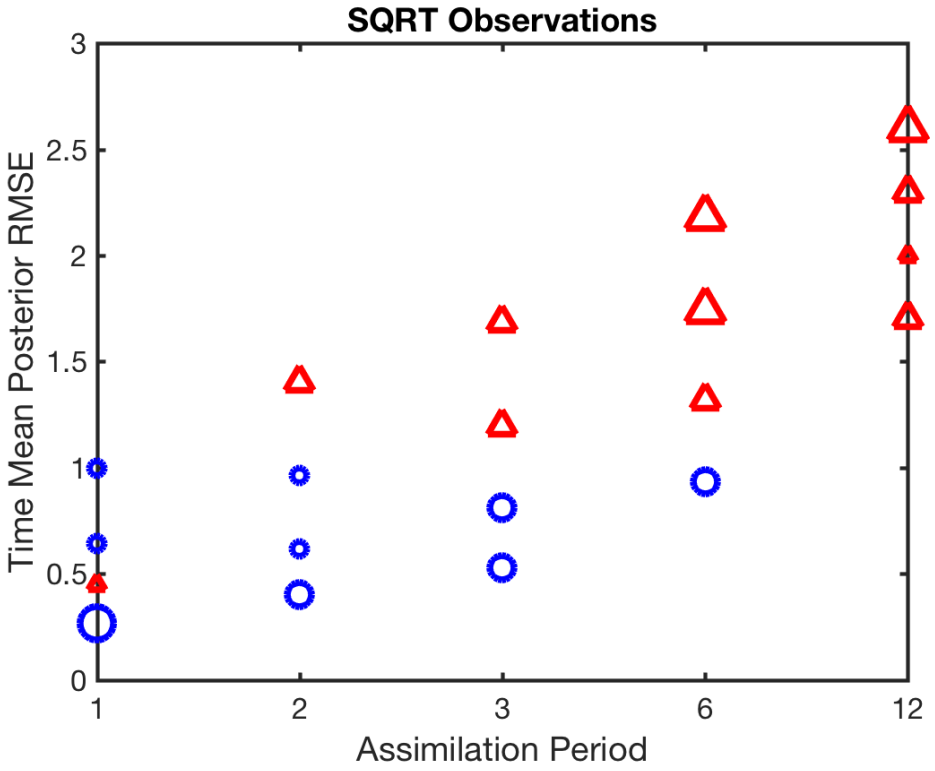
Square Root Forward Operator: Ensemble Size

80 Members: RHF with rank regression for large RMSE.
EnKF/linear for smaller RMSE.



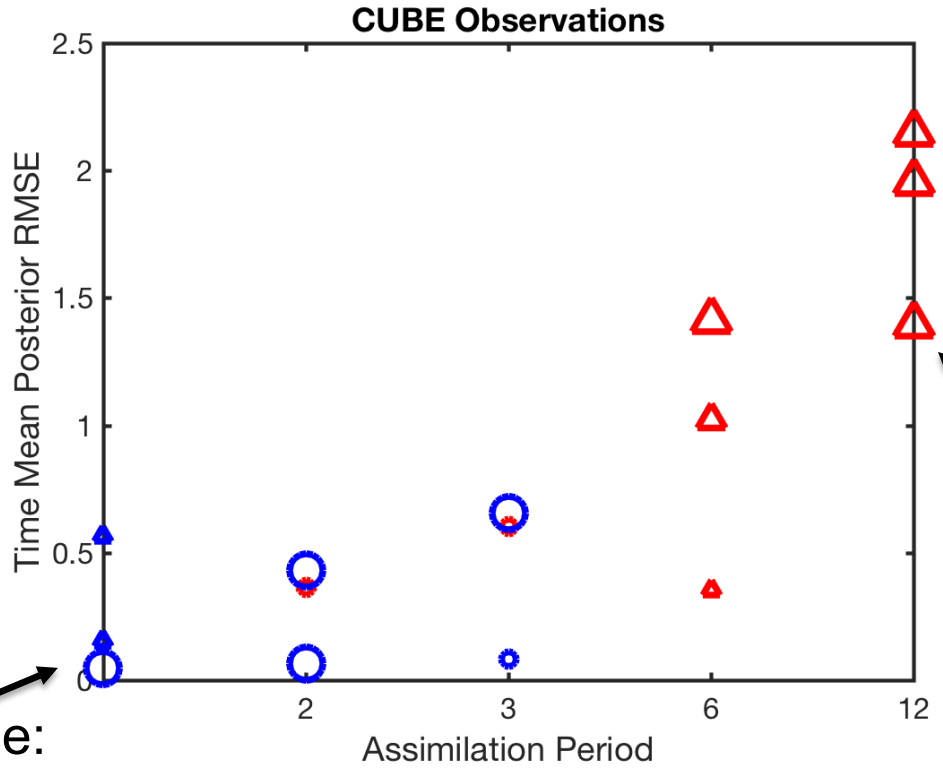
Square Root Forward Operator: Ensemble Size

160 Members: RHF with rank regression for large RMSE.
EnKF/linear for smaller RMSE.



Summary: Cube Forward Operator Results

RHF with rank regression better for larger RMSE.
EnKF with linear regression better for smaller RMSE.

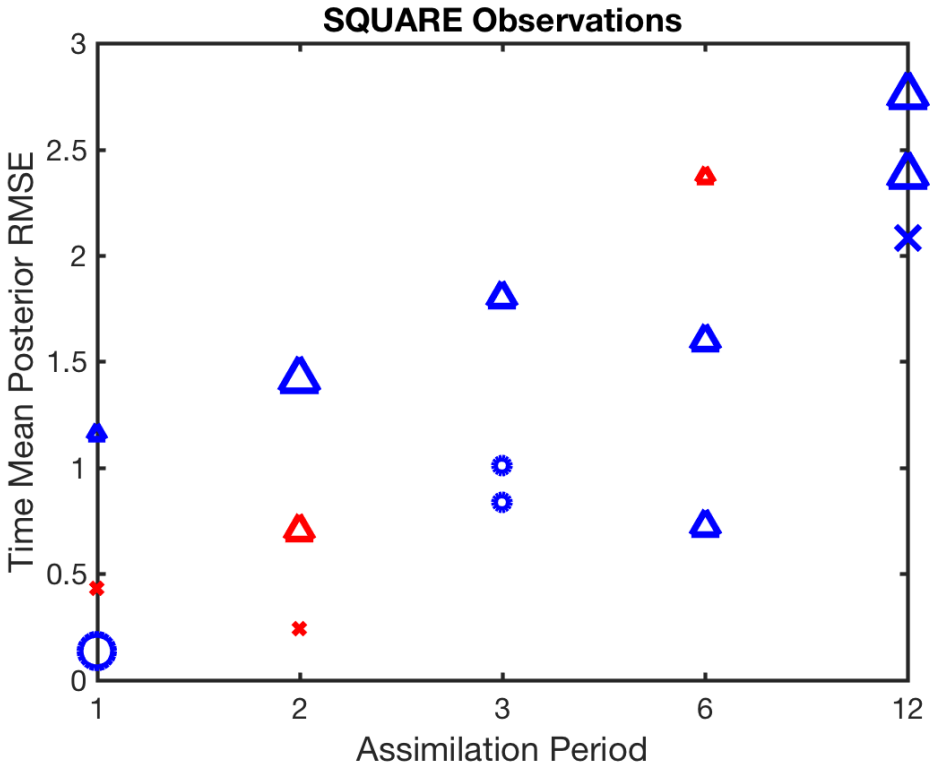


Red Triangle:
RHF with rank regression best

Blue Circle:
EnKF with linear regression best

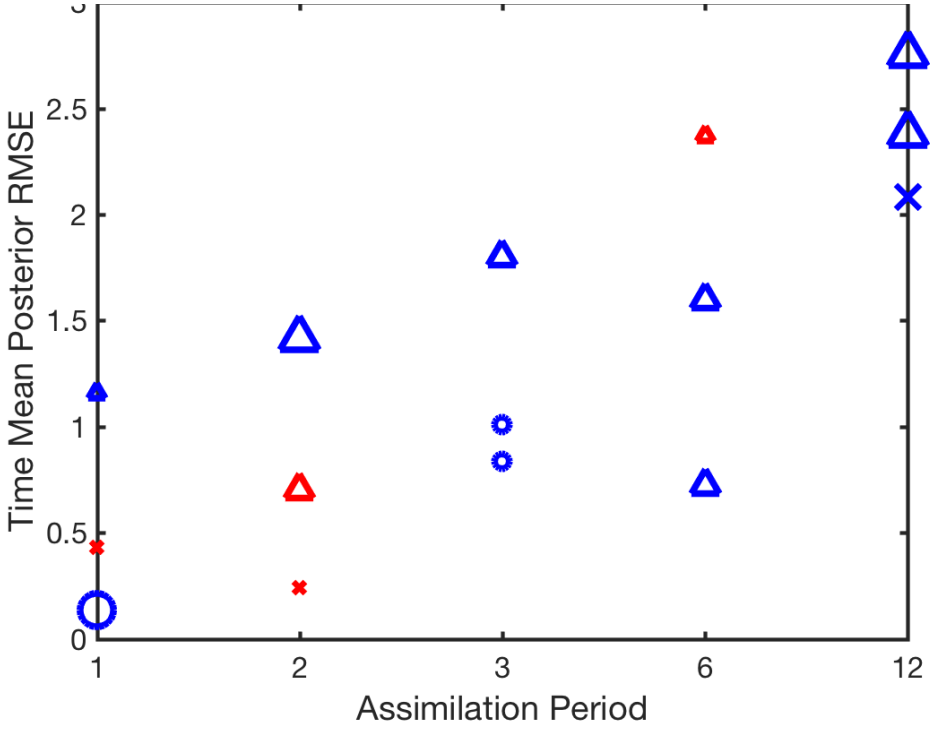
Summary: Square Forward Operator Results

RHF/linear better for larger RMSE.
Mixed for smaller RMSE.



Summary: Square Forward Operator Results

RHF/linear better for larger RMSE.
Mixed for smaller RMSE.
Observation operator is not invertible!



Computational Cost

Computational cost for the state variable update:

Base regression: $O(m^2n)$
Rank regression: $O(m^2n\log n)$

m: sum of state size plus number of observations,
n: ensemble size.

Average rank cost can be made $O(m^2n)$ with some work (smart sorting and searching).

Good for GPUs (more computation per byte).

Conclusions

Non-Gaussian observation space methods can be superior to Gaussian.

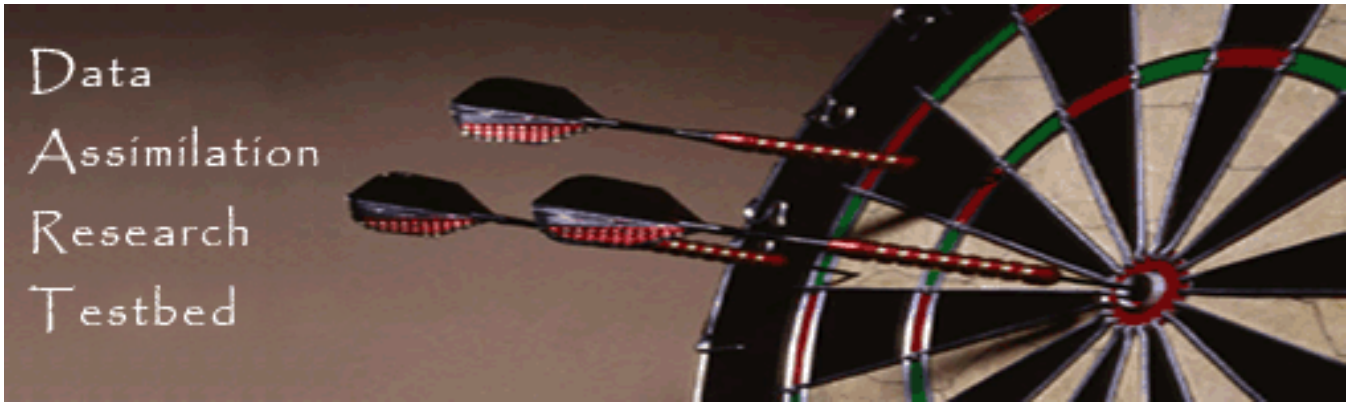
Nonlinear regression can be superior in cases of strong nonlinearity.

Other nonlinear regressions (e.g. polynomial) can be effective.

Ensemble method details make significant differences in performance.

NWP applications should be carefully developed.

All results with DART/DARTLAB tools
freely available in DART.



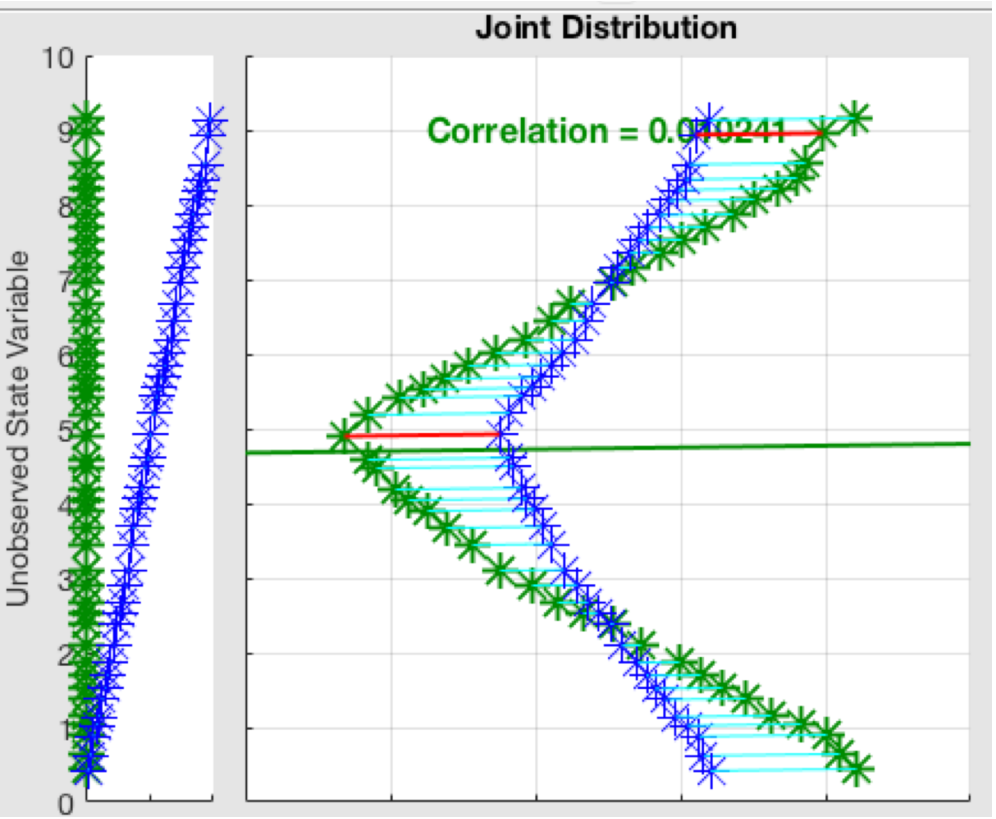
www.image.ucar.edu/DAReS/DART

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

Multi-valued, not smooth example.

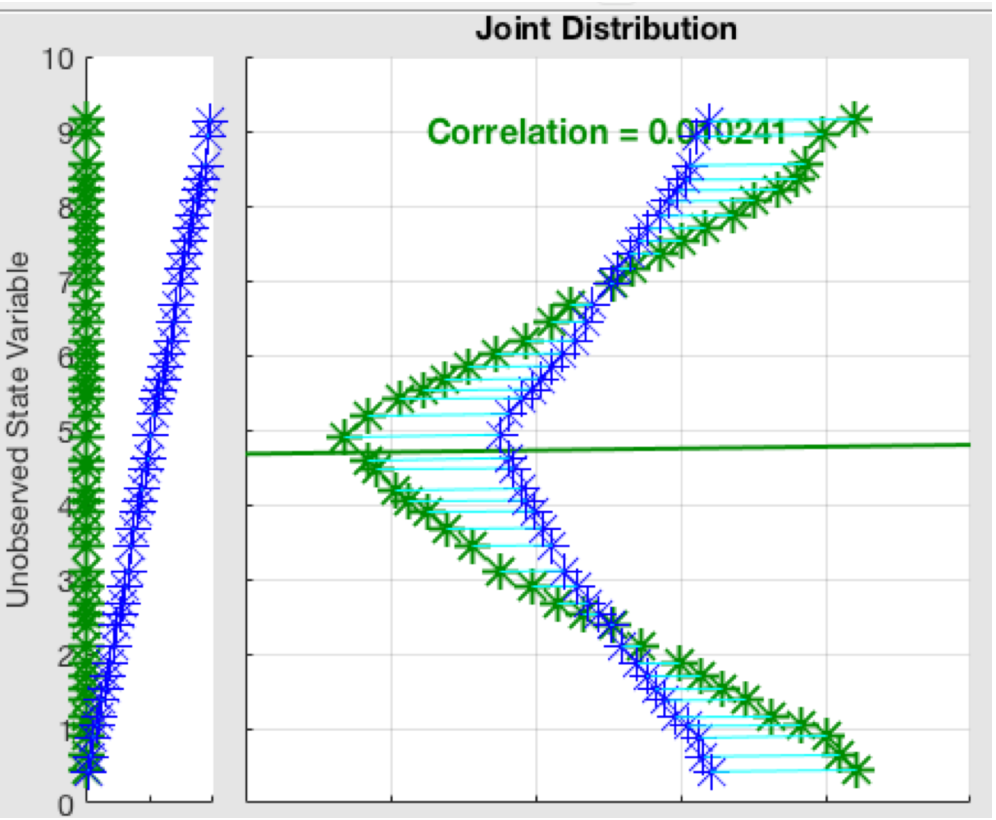
Similar in form to a wind speed observation with state velocity component.

Multi-valued, not smooth example.



Standard regression does not capture bimodality of state posterior.

Multi-valued, not smooth example.



Rank regression nearly identical in this case.