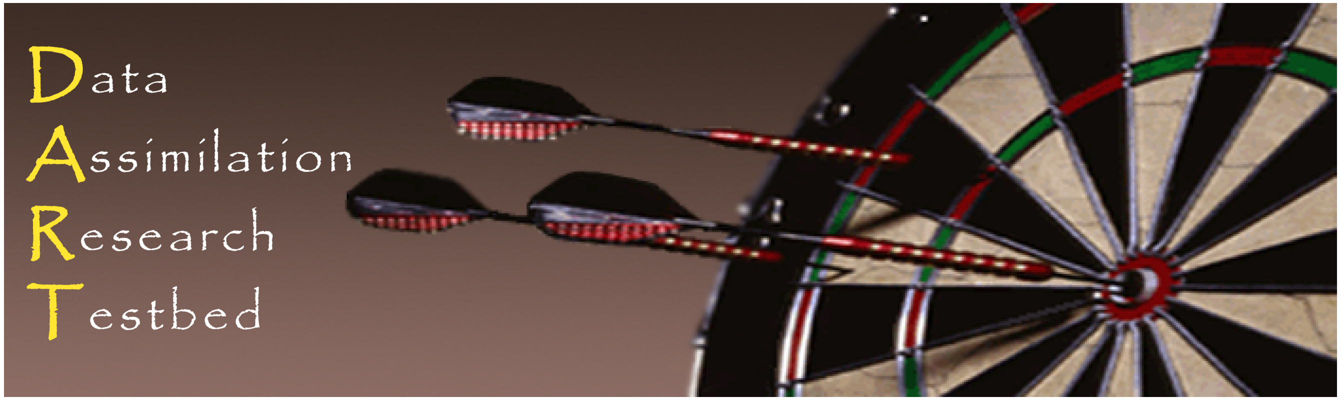
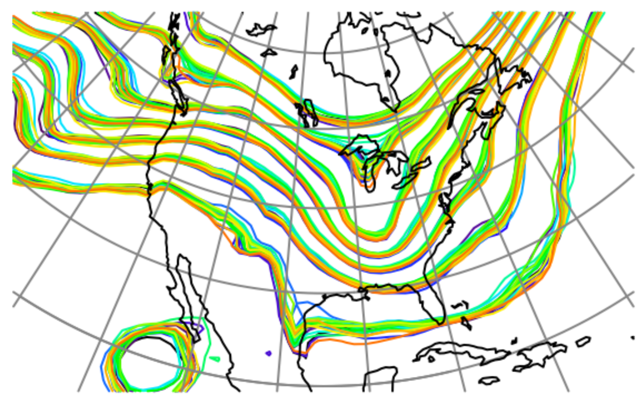


DART Tutorial Section 8: Dealing with Sampling Error



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.

NCAR | National Center for
UCAR | Atmospheric Research

Updating Additional Prior State Variables

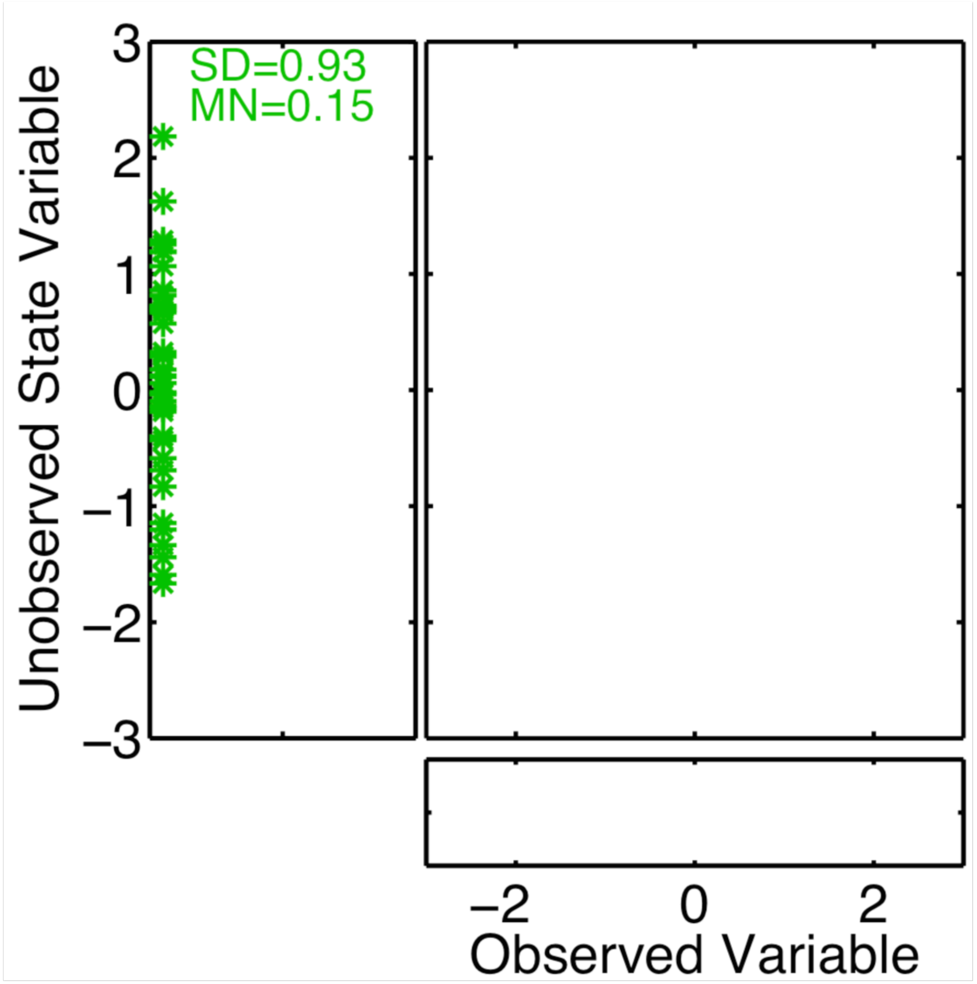
Two primary error sources:

1. Sampling error due to noise.
Can occur even if there is a linear relation between variables.
Sample regression coefficient imprecise with finite ensembles.

2. Linear approximation is invalid.
If there is substantial nonlinearity in 'true' relation between variables over range of prior ensemble. (see section 10).

May need to address both issues for good performance.

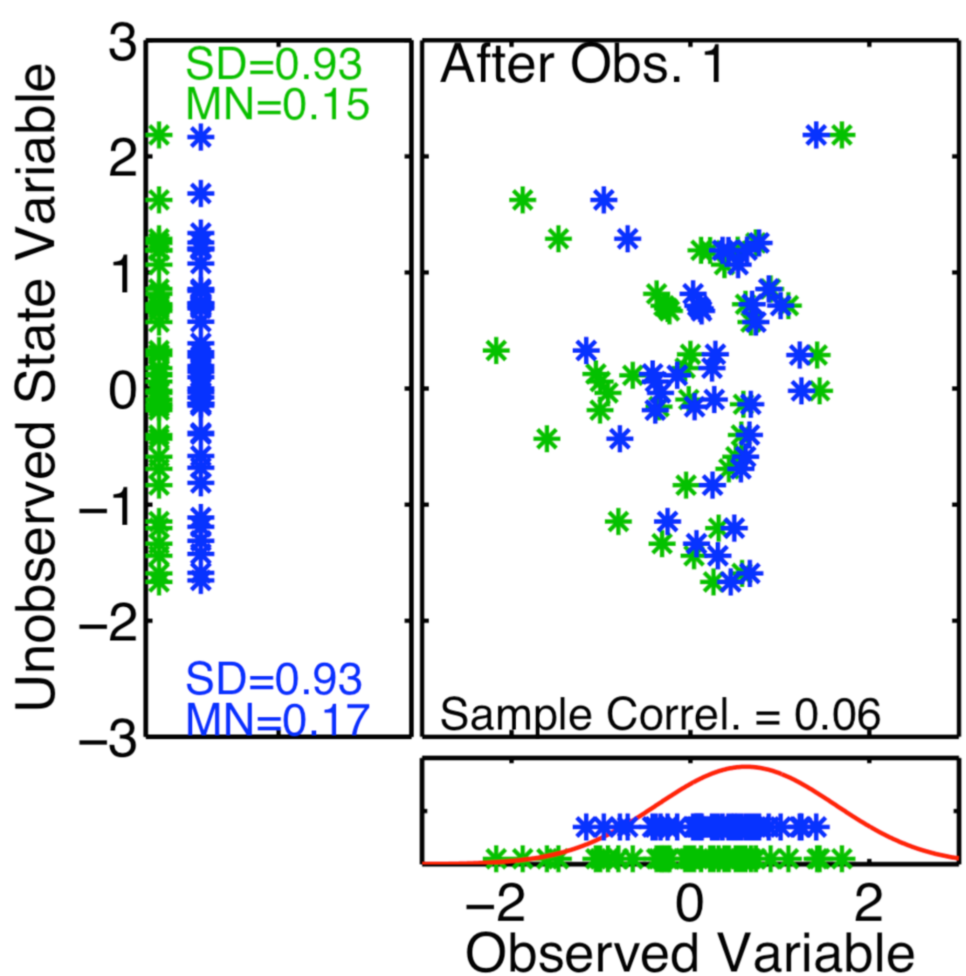
Regression Sampling Error & Filter Divergence



Suppose unobserved state variable is known to be unrelated to set of observed variables.

Unobserved variable should remain unchanged.

Regression Sampling Error & Filter Divergence

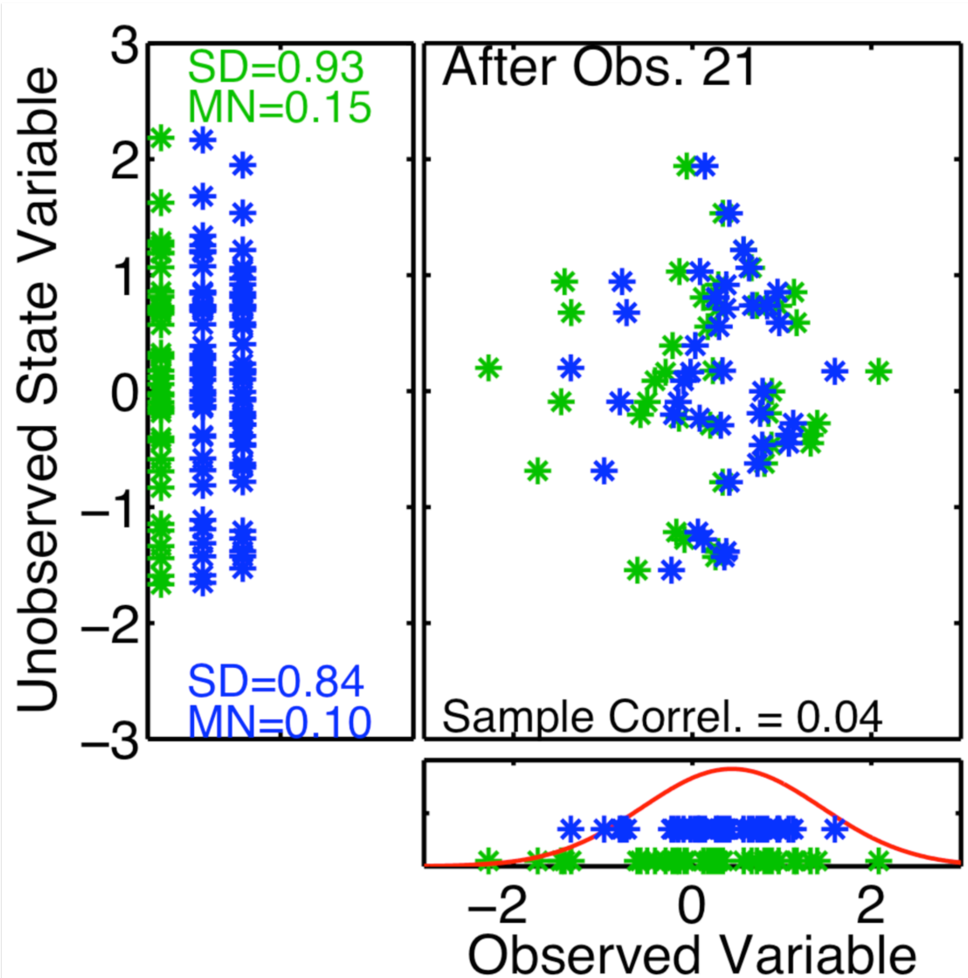


Suppose unobserved state variable is known to be unrelated to set of observed variables.

Finite samples from joint distribution will have non-zero correlation. Expected |correl| = 0.19 for 20 samples.

After one observation, unobserved variable mean, standard deviation change.

Regression Sampling Error & Filter Divergence

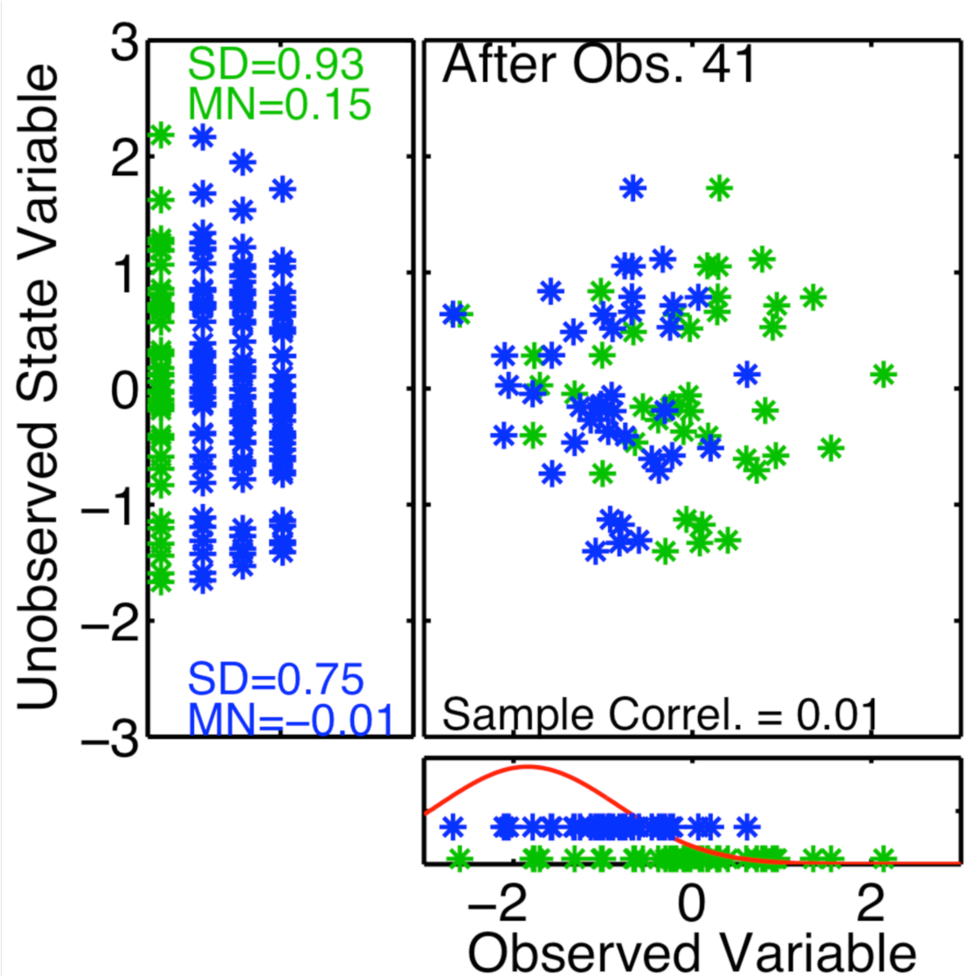


Suppose unobserved state variable is known to be unrelated to set of observed variables.

Unobserved variable should remain unchanged.

Unobserved mean follows a random walk as more observations are used.

Regression Sampling Error & Filter Divergence



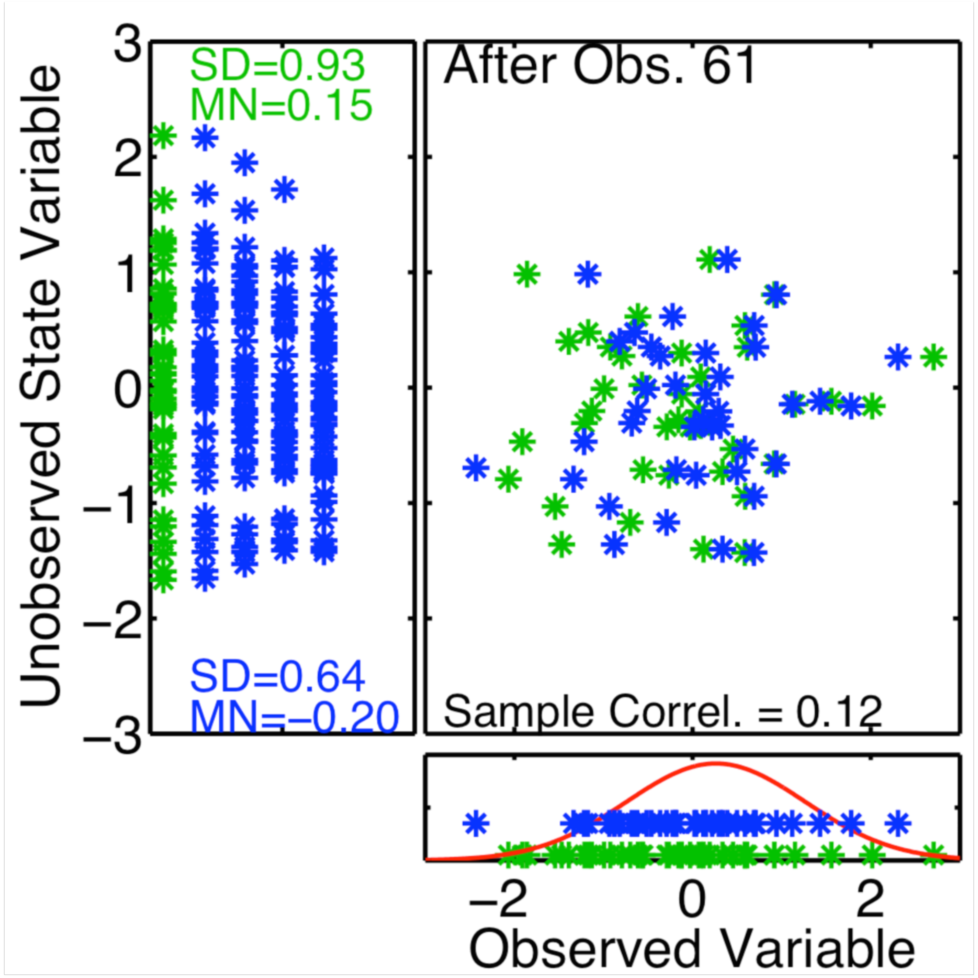
Suppose unobserved state variable is known to be unrelated to set of observed variables.

Unobserved variable should remain unchanged.

Unobserved S.D systematically decreases.

Expected change in |SD| is negative for any non-zero sample correlation.

Regression Sampling Error & Filter Divergence



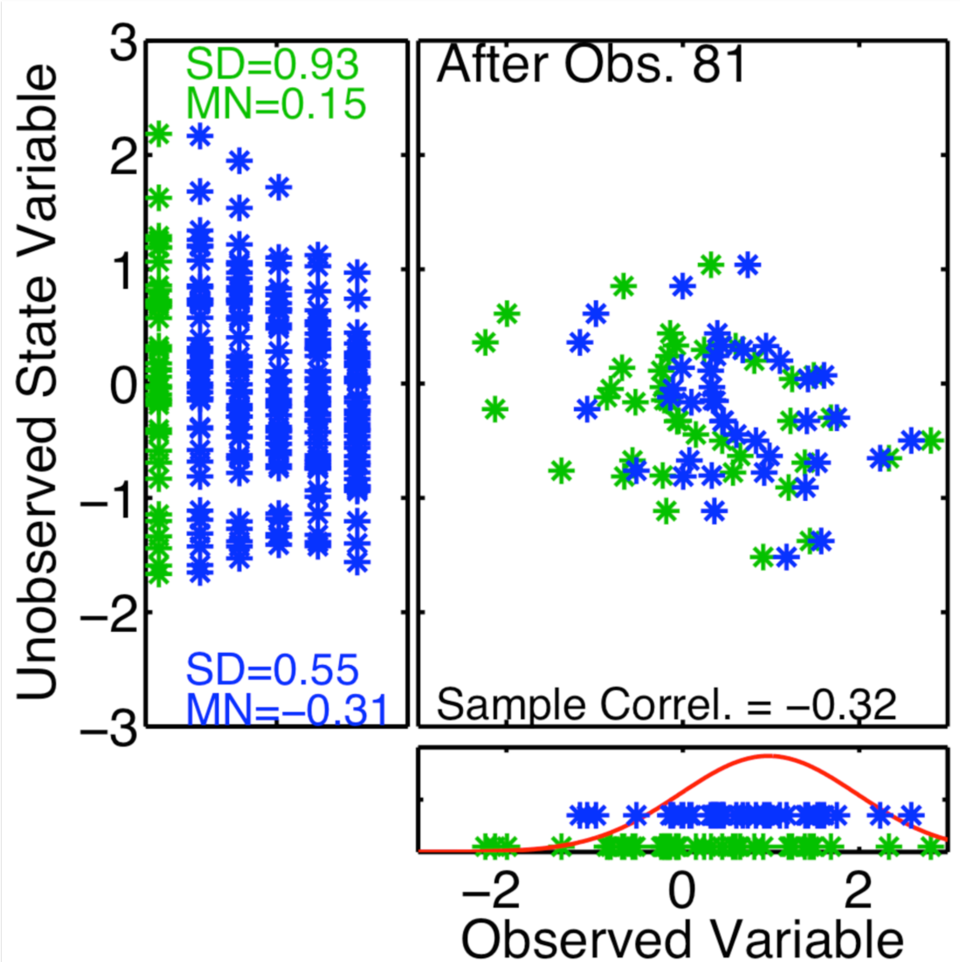
Suppose unobserved state variable is known to be unrelated to set of observed variables.

Unobserved variable should remain unchanged.

Unobserved S.D systematically decreases.

Expected change in |SD| is negative for any non-zero sample correlation.

Regression Sampling Error & Filter Divergence



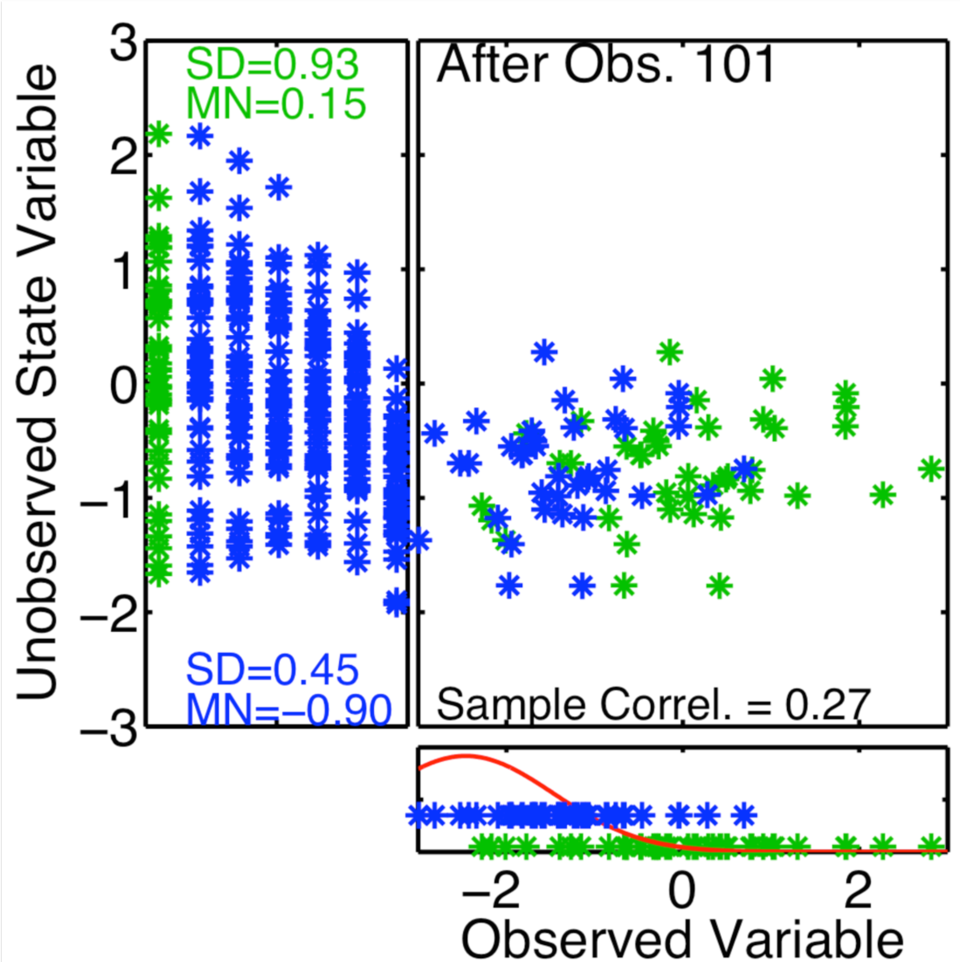
Suppose unobserved state variable is known to be unrelated to set of observed variables.

Unobserved variable should remain unchanged.

Unobserved S.D systematically decreases.

Expected change in |SD| is negative for any non-zero sample correlation.

Regression Sampling Error & Filter Divergence



Suppose unobserved state variable is known to be unrelated to set of observed variables.

Estimates of unobserved become too confident.

Give progressively less weight to meaningful obs.

Eventually, meaningful obs are essentially ignored.

Filter Divergence

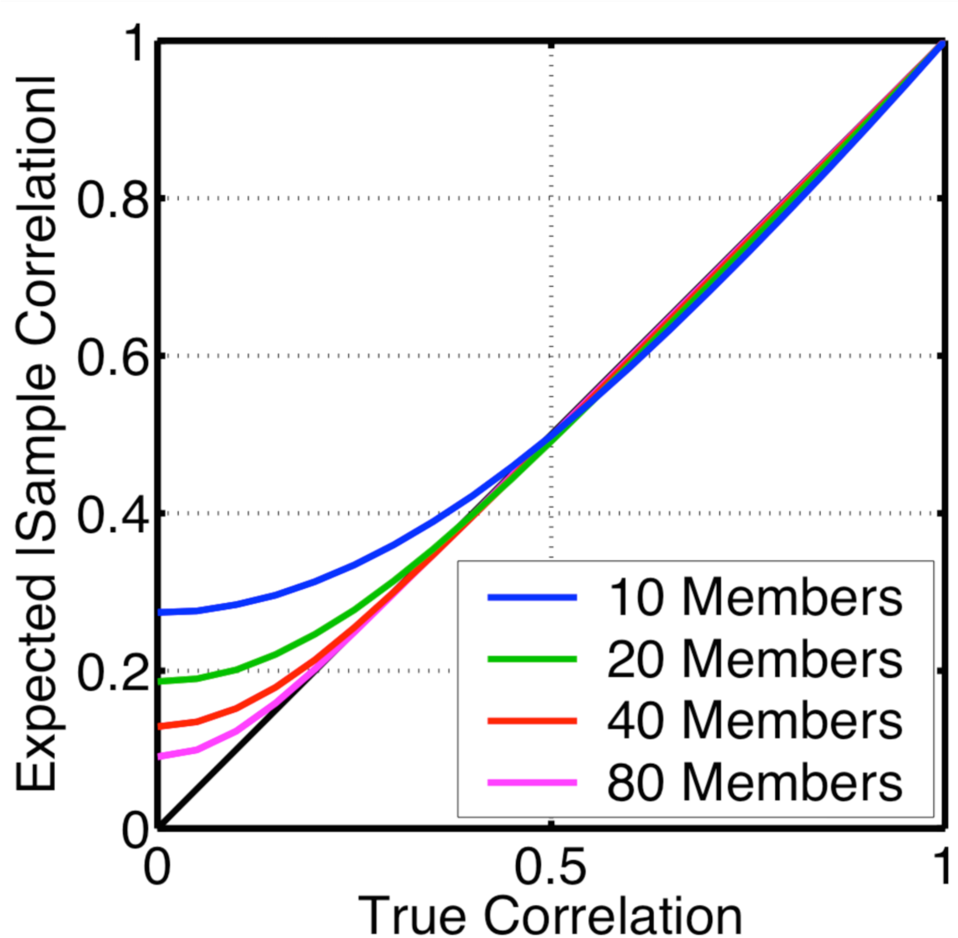
Ignoring meaningful observations due to overconfidence is a type of FILTER DIVERGENCE.

This was seen in initial Lorenz_96 (40-variable) experiment.

The spread became small => the filter thought it had a good estimate.

The error stayed large because good observations were ignored.

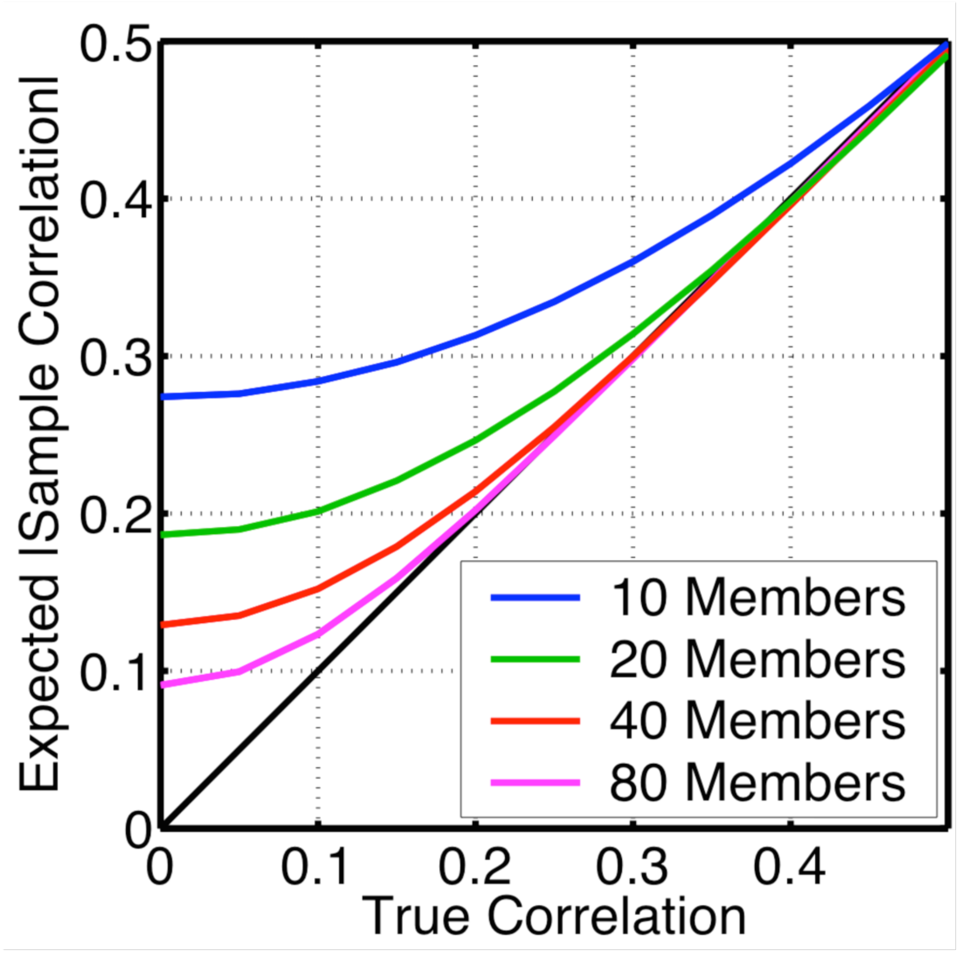
Regression Sampling Error & Filter Divergence



Plot shows expected absolute value of sample correlation versus true correlation.

Error decreases with sample size and for larger |real correlations|.

Regression Sampling Error & Filter Divergence



Plot shows expected absolute value of sample correlation versus true correlation.

For small true correlations, errors are still undesirably large even for 80 member ensembles.

Dealing with Regression Sampling Error

1. Ignore it: if number of unrelated observations is small and there is some way of maintaining variance in priors.
   We did this in the 3 and 9 variable models.

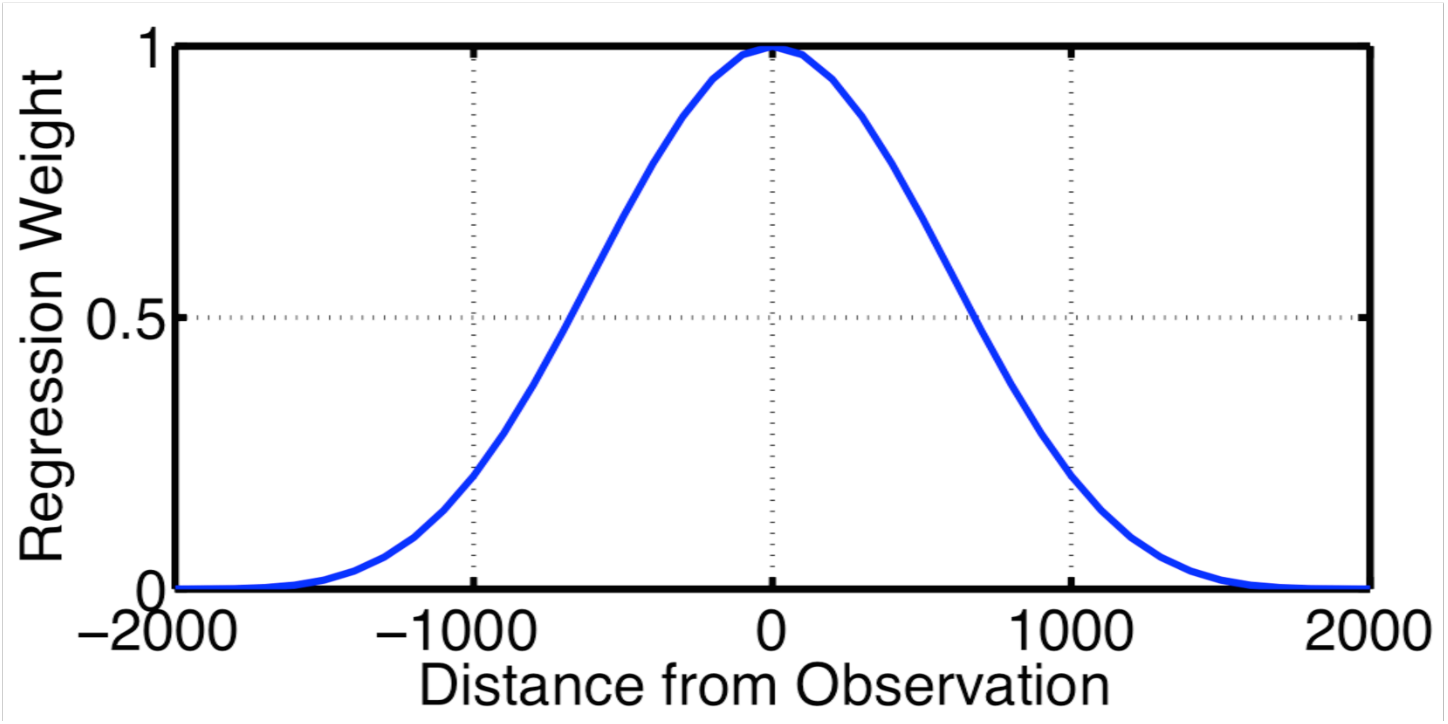
2. Use larger ensembles to limit sampling error (test in lorenz_96).
   This can get expensive for big problems.
   Try modifying *ens_size* in *filter_nml* (try 40, 80).

3. Use additional a priori information about relation between observations and state variables.
   Don't let an observation impact state if they are known to be unrelated.

4. Try to determine the amount of sampling error and correct for it.
   There are many ways to do this; some simple, some complex.

Dealing with Regression Sampling Error

3. Use additional a priori information about relation between observations and state variables.



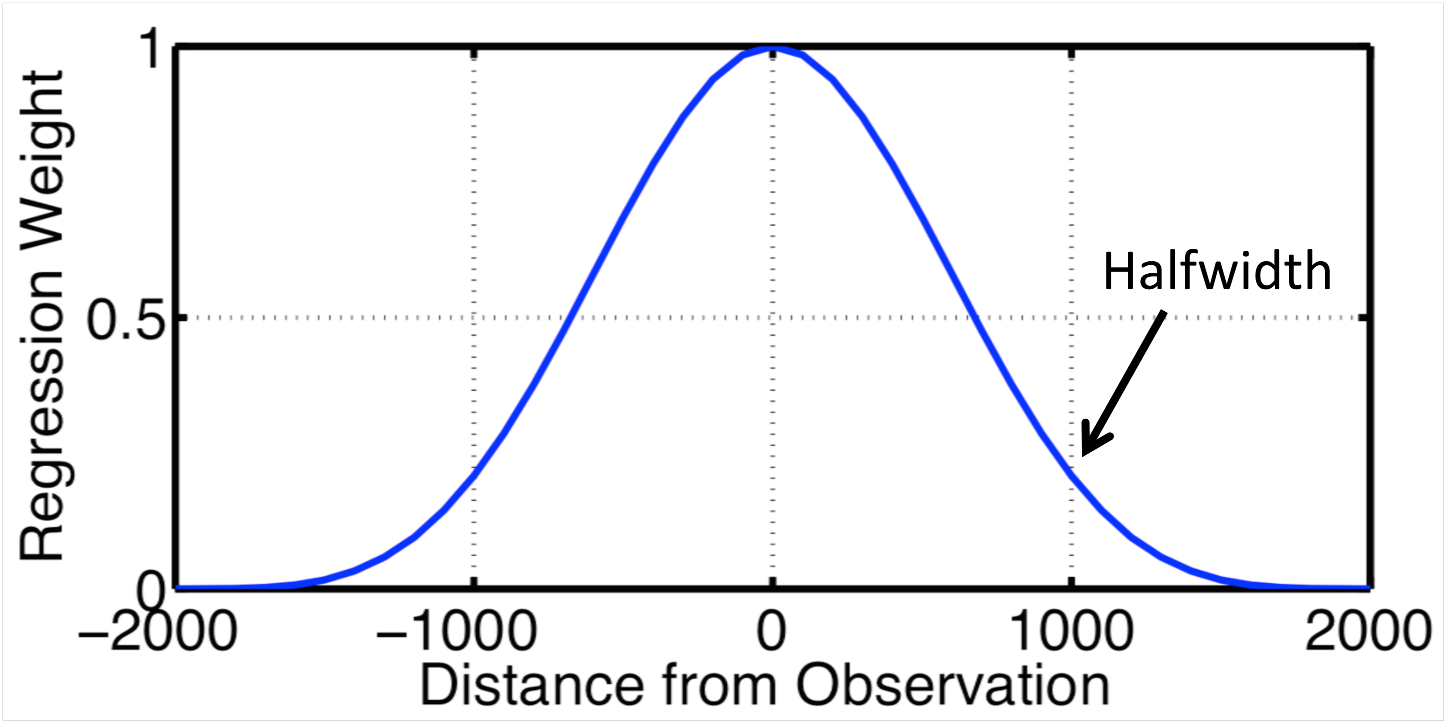
For atmospheric assimilation problems:

Weight regression as function of horizontal *distance* from observation.

Gaspari-Cohn: 5th order compactly supported polynomial.

Dealing with Regression Sampling Error

3. Use additional a priori information about relation between observations and state variables.



Can use other functions to weight regression.

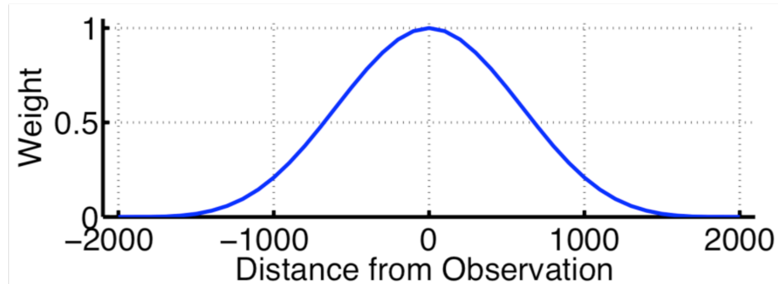
Unclear what distance means for some obs./state variable pairs.

Referred to as LOCALIZATION.

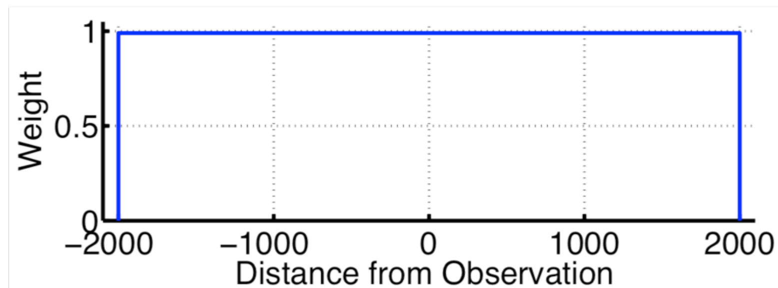
DART provides several localization options

1. Different shapes for the localization function are available. Controlled by *select_localization* in *cov_cutoff_nml*.

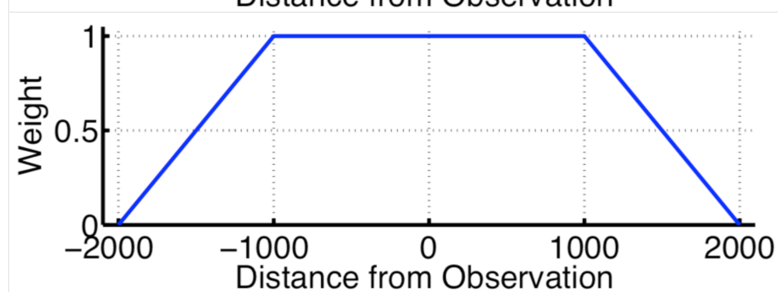
1=> Gaspari-Cohn



2=> Boxcar



3=> Ramped Boxcar



2. Halfwidth of localization function set by *cutoff* in *assim_tools_nml*

Experimenting with Lorenz96

The Lorenz_96 domain is mapped to a [0, 1] periodic range.

Try a variety of half widths for a Gaspari Cohn localization.
(First change *ens_size* in *filter_nml* back to 20)
Change *cutoff* in *assim_tools_nml.*

We already know that a very large localization half-width diverges.

What happens for a very small value?

What happens with intermediate values (say between 0.1 and 0.5)?

Can also try changing the shape, (best with 40 member ensemble):
Try option 2 or 3 for *select_localization* in *cov_cutoff_nml*.

Dealing with Regression Sampling Error

4. Try to determine the amount of sampling error and correct for it.

Many ways to do this. DART implements one naive way:

1. Take set of increments from a given observation,
2. Suppose this observation and a state variable are not correlated,
3. Compute the expected decrease in spread given not correlated,
4. Add this amount of spread back into the state variable.

The expected decrease in spread is computed by off-line Monte Carlo.
Results of off-line simulation are tabulated and applied.
(This can be a very useful technique when you're analytically clueless).

Try this algorithm: set *spread_restoration* in *assim_tools_nml* to true.
How does it work with 20 ensemble members, no localization?

DART Tutorial Index to Sections