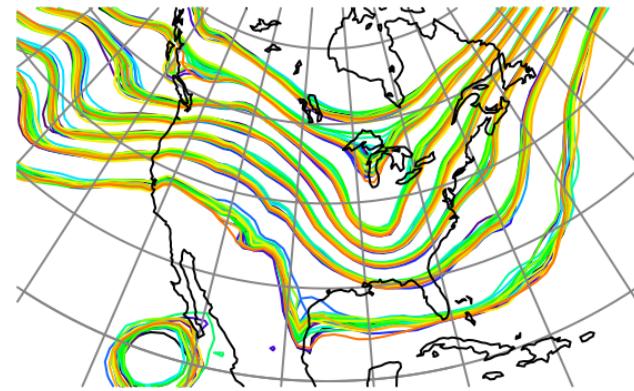


D  
A  
R  
T

ata  
ssimilation  
esearch  
estbed



# DART Tutorial Part III: Dealing with Sampling Error



©UCAR 2014

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.

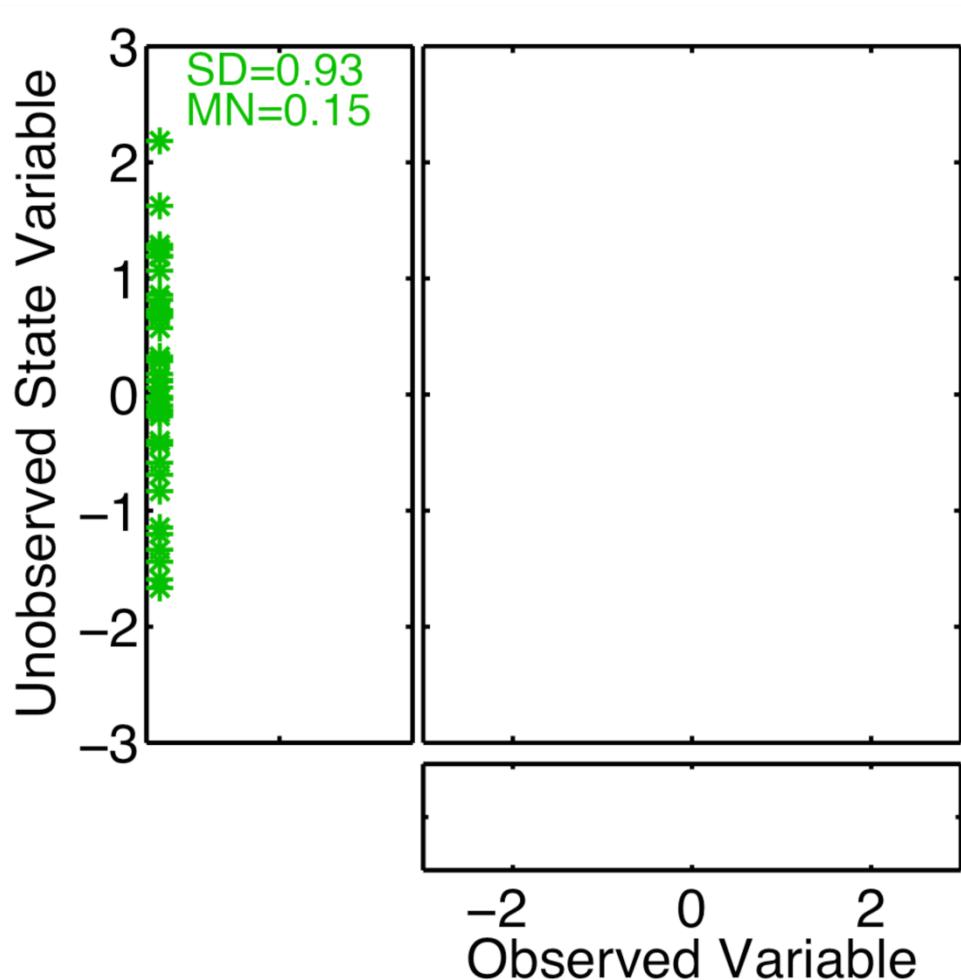
Sample regression coefficient imprecise with finite ensembles.

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.

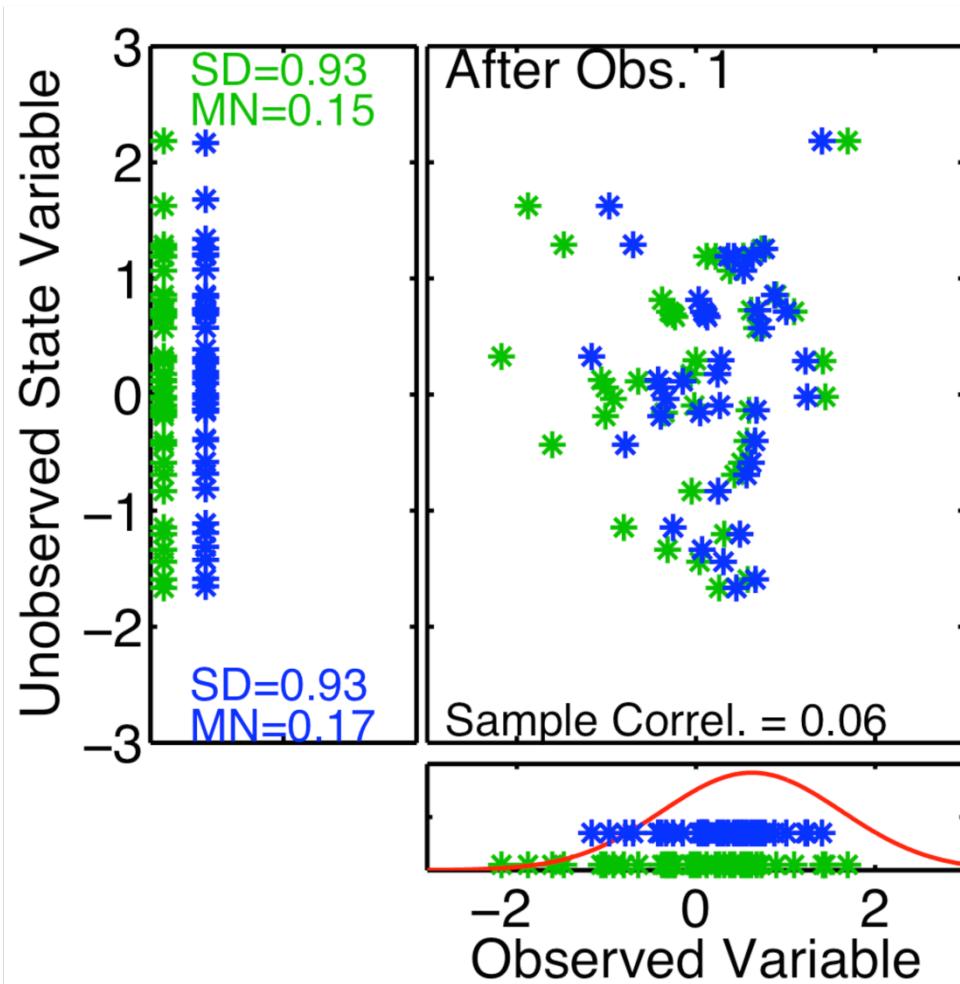
# Regression Sampling Error & Filter Divergence



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

Unobserved variable  
should remain unchanged.

# Regression Sampling Error & Filter Divergence

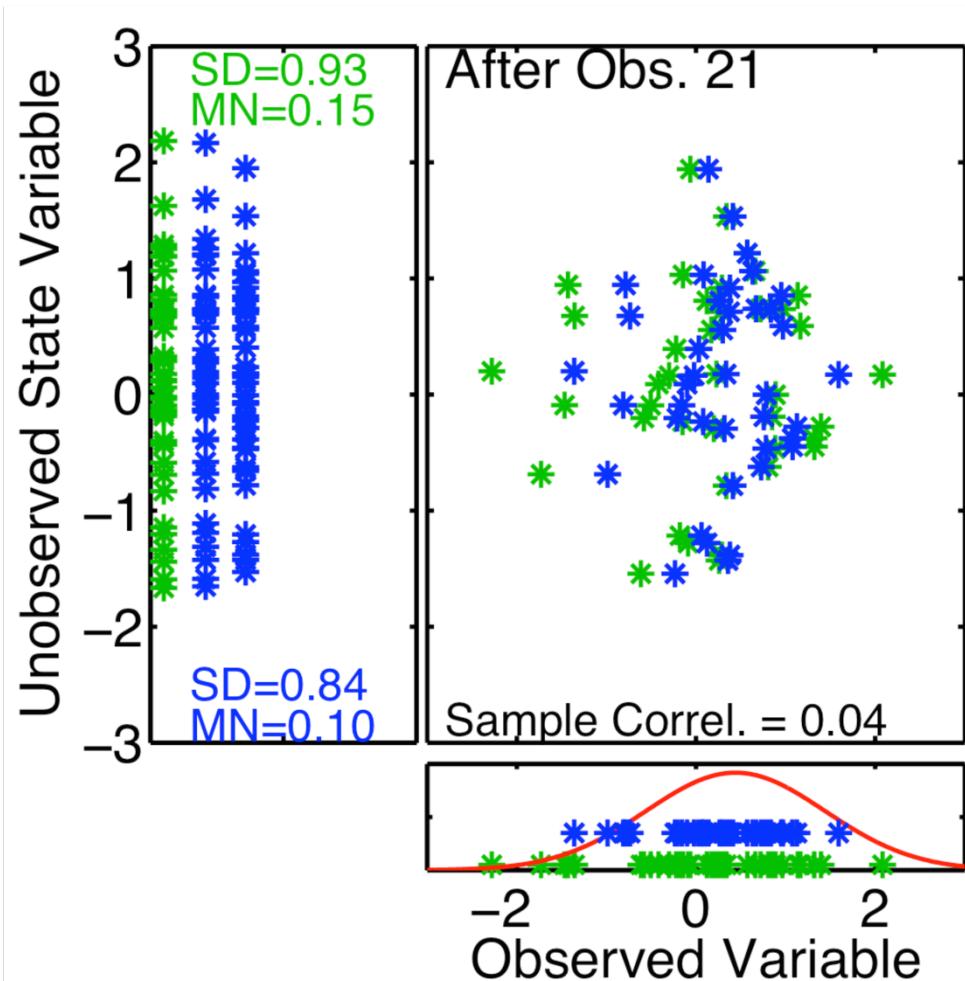


Suppose unobserved state variable is known to be unrelated to set of observed variables.

Finite samples from joint distribution will have non-zero correlation. Expected  $|\text{correl}| = 0.19$  for 20 samples.

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

# Regression Sampling Error & Filter Divergence

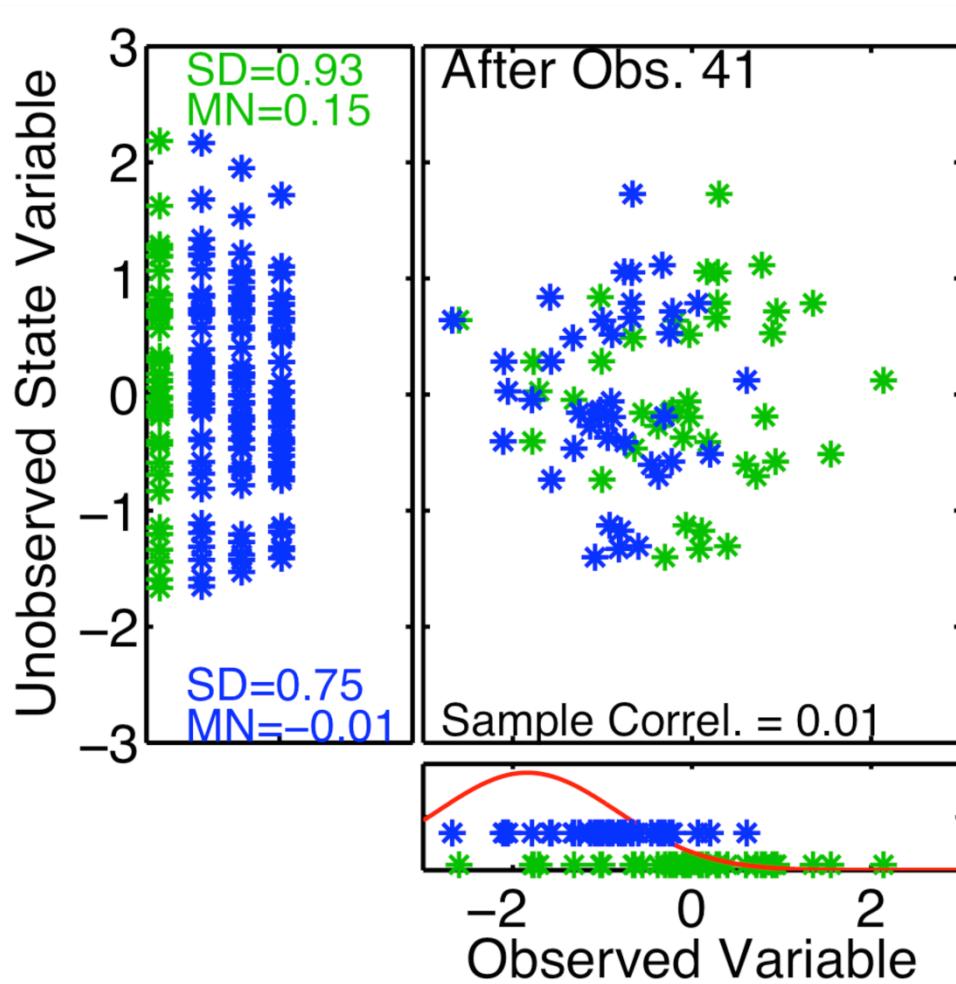


Suppose unobserved state variable is known to be unrelated to set of observed variables.

Unobserved variable should remain unchanged.

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

# Regression Sampling Error & Filter Divergence



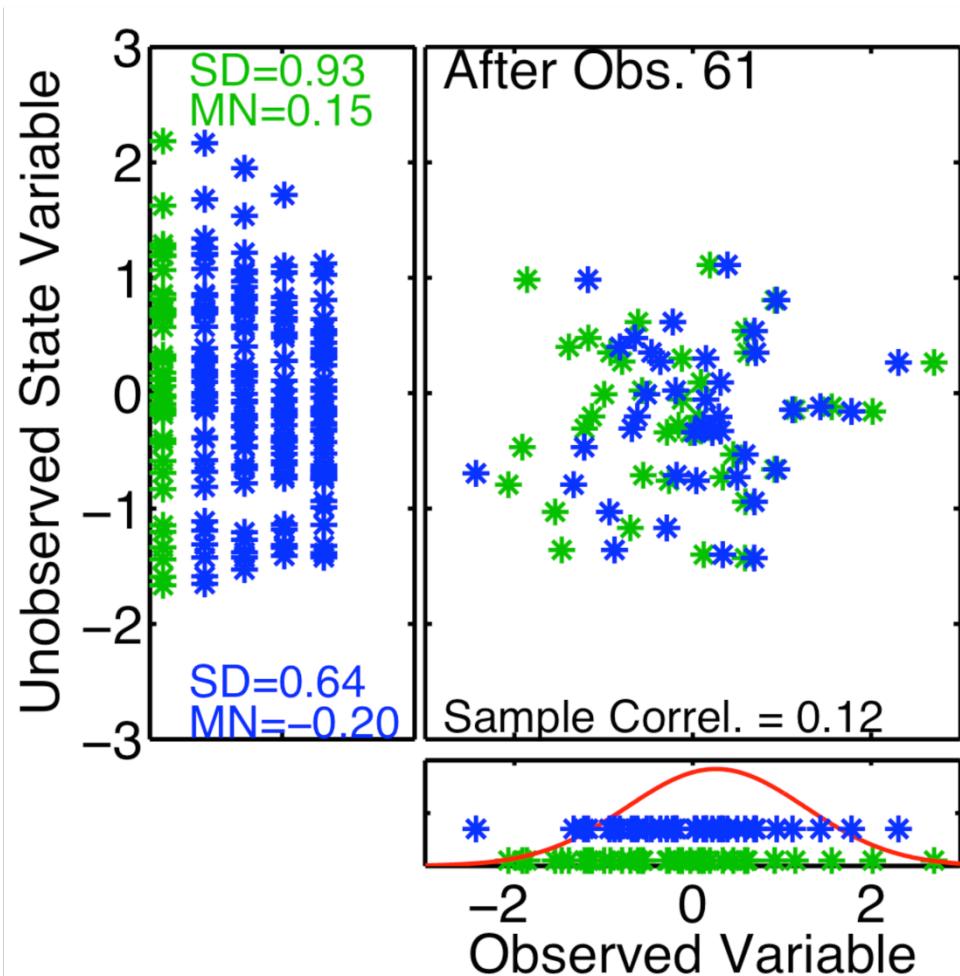
Suppose unobserved state variable is known to be unrelated to set of observed variables.

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



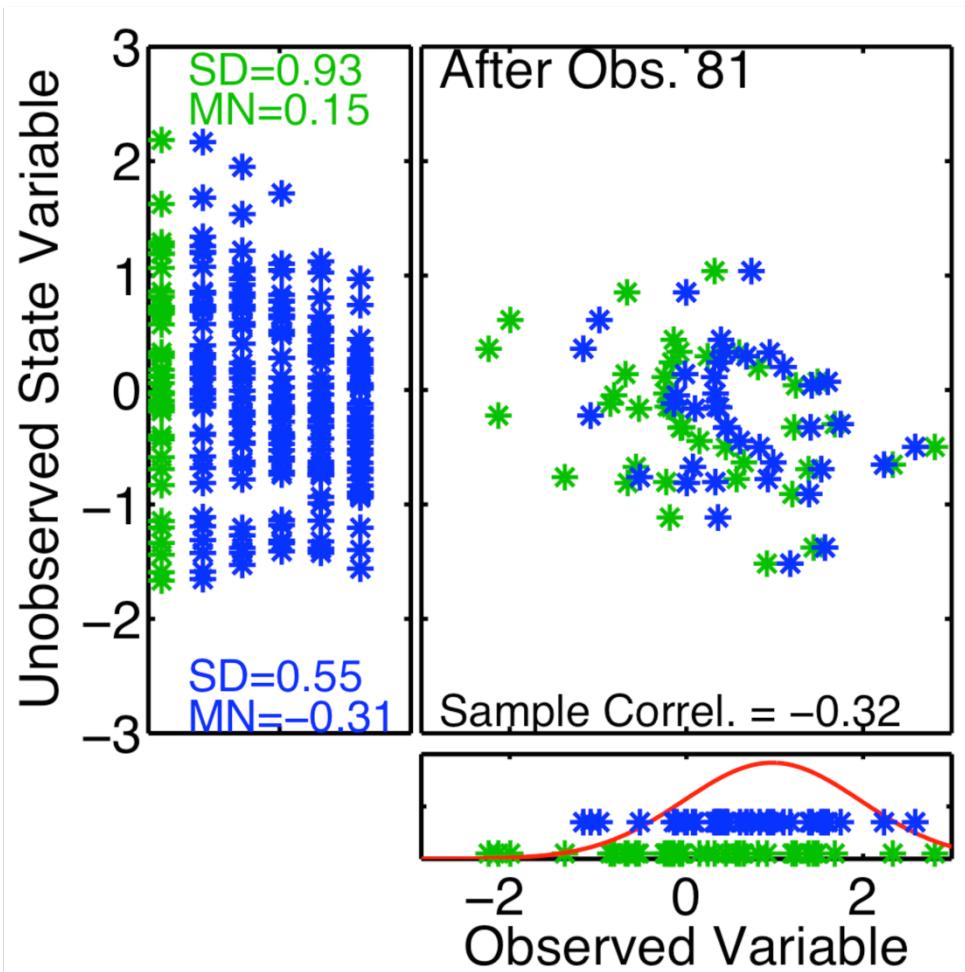
Suppose unobserved state variable is known to be unrelated to set of observed variables.

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



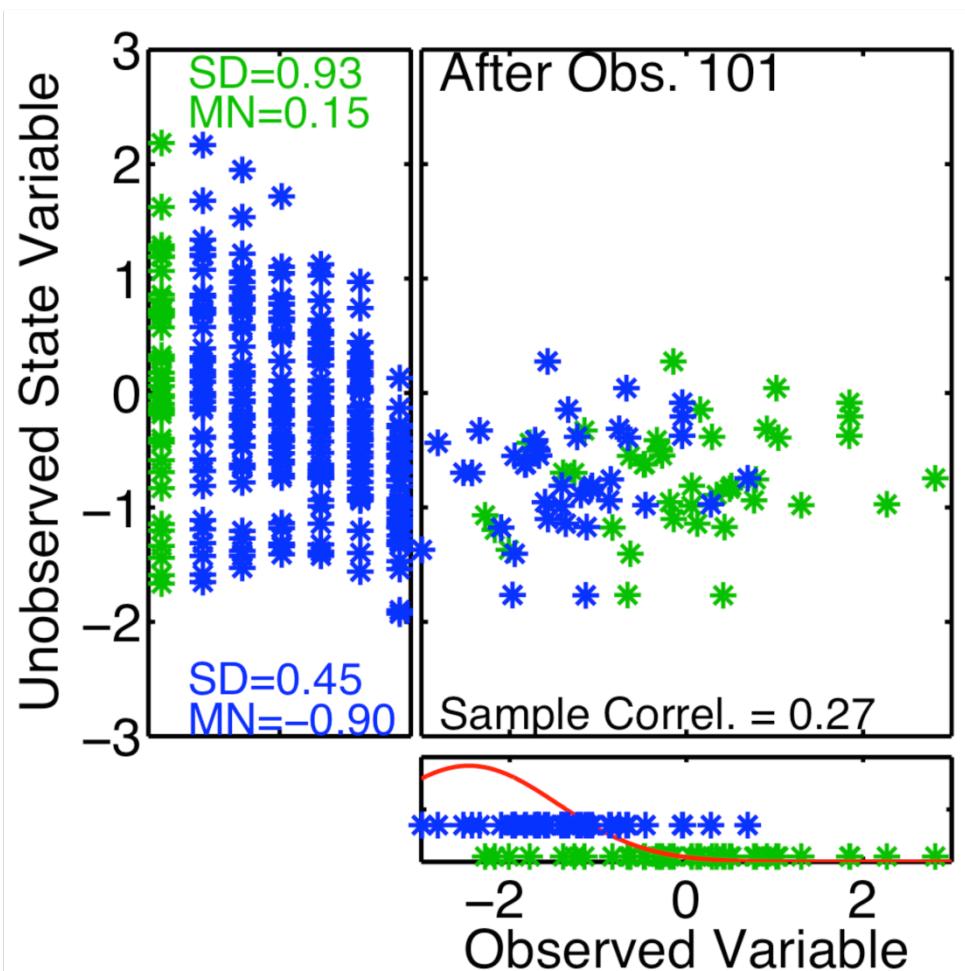
Suppose unobserved state variable is known to be unrelated to set of observed variables.

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 any meaningful observations that enter. Eventually, meaningful obs are essentially ignored.

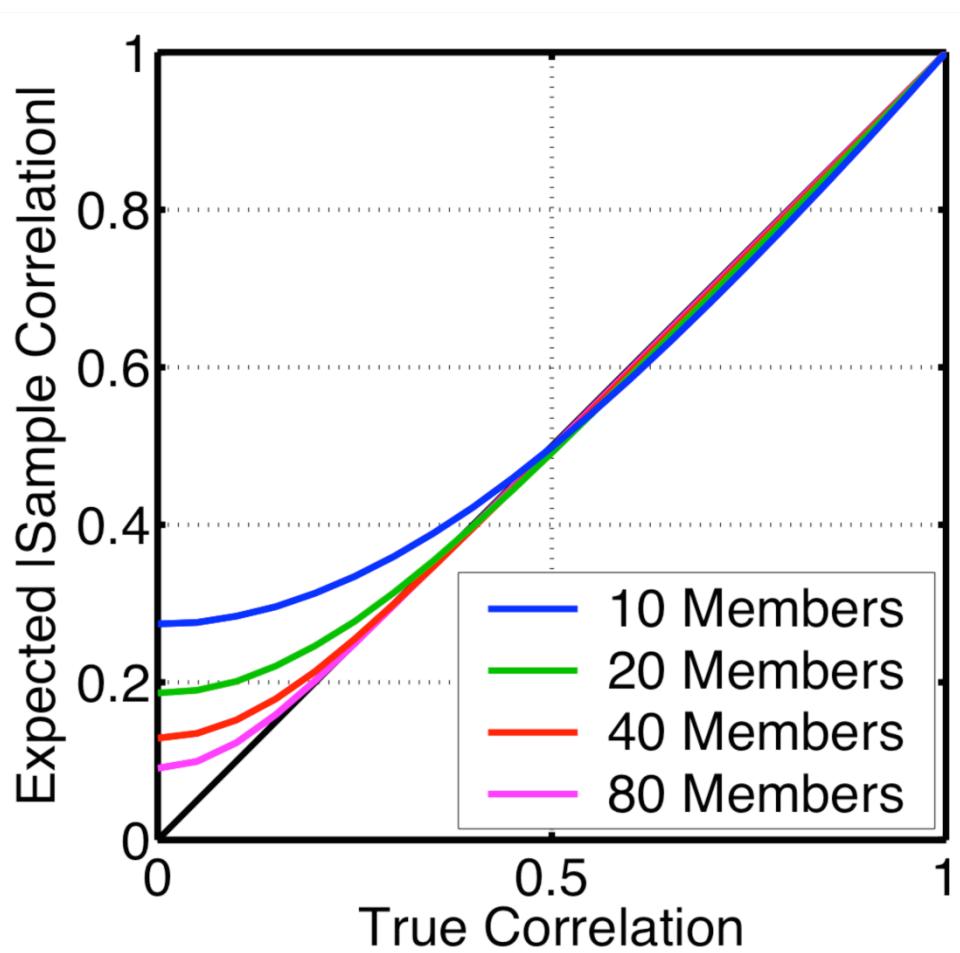
# Filter Divergence

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

The spread becomes small as the filter thinks that it has a good sample estimate.

The error stayed large because good observations were ignored: recall, the weighting given to the observation is small when the standard deviation of the prior is small!

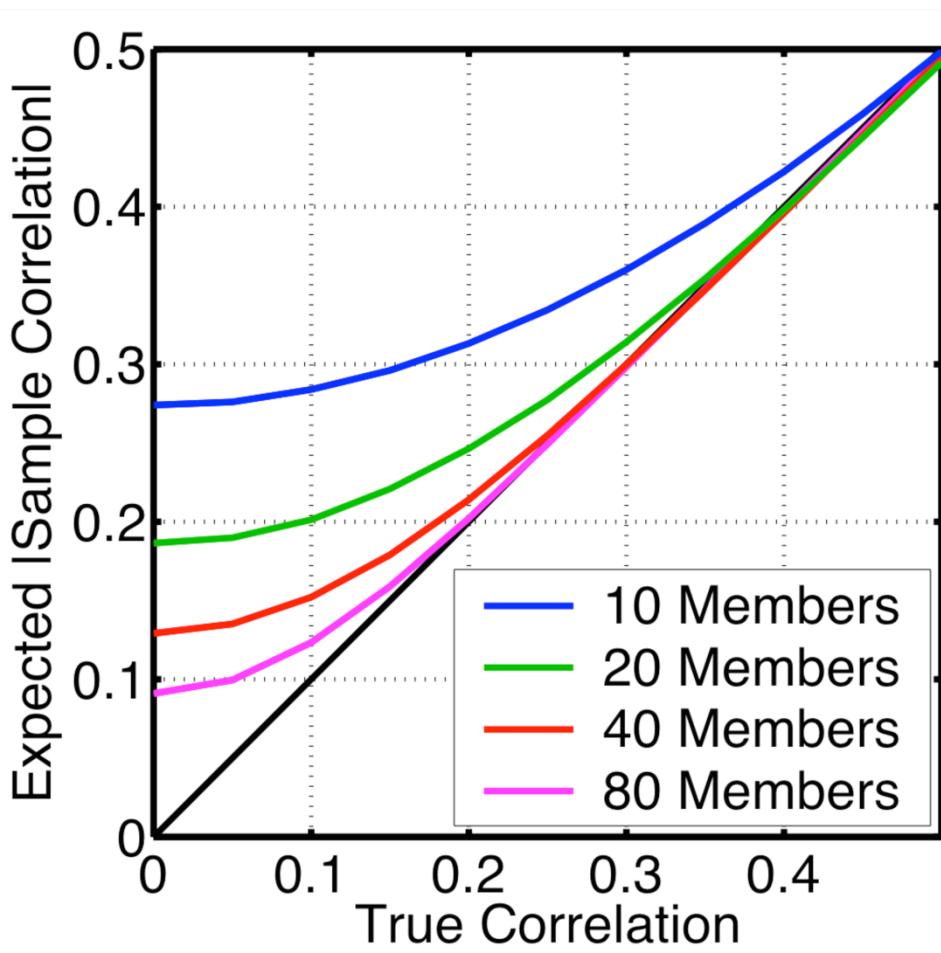
# Regression Sampling Error & Filter Divergence



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

Error decreases with sample size and for larger |true 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.

# Ways to deal with Regression Sampling Error

1. Ignore it if number of weakly correlated observations is small and there is a way to maintain prior variance.

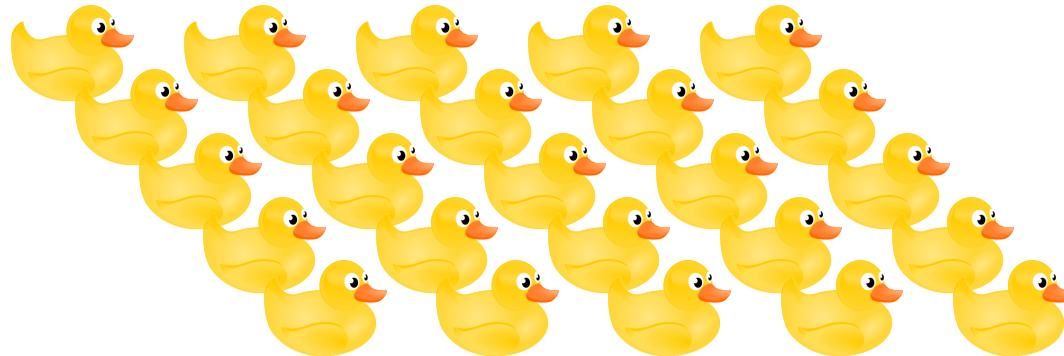


?

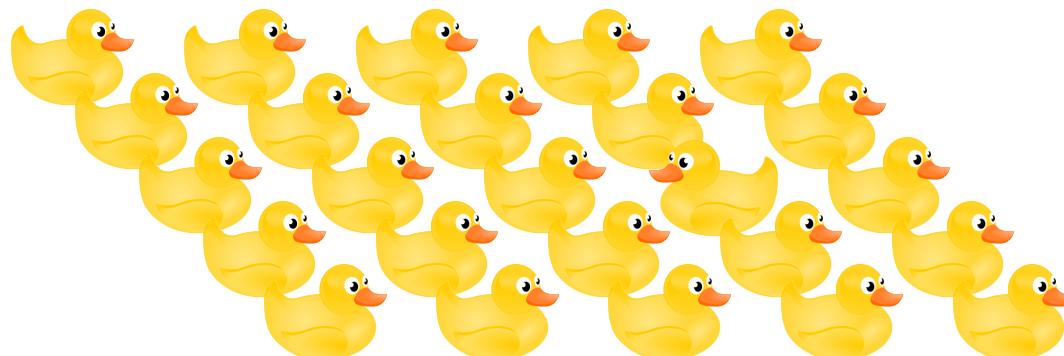


It may work.  
It may not.

# Ways to deal with Regression Sampling Error

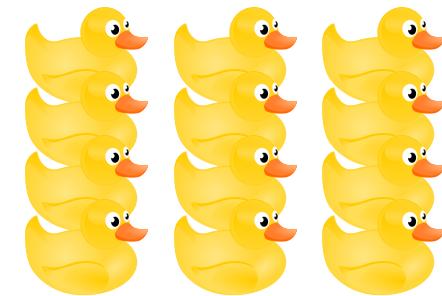


2. Use larger ensembles; computationally expensive for large models.



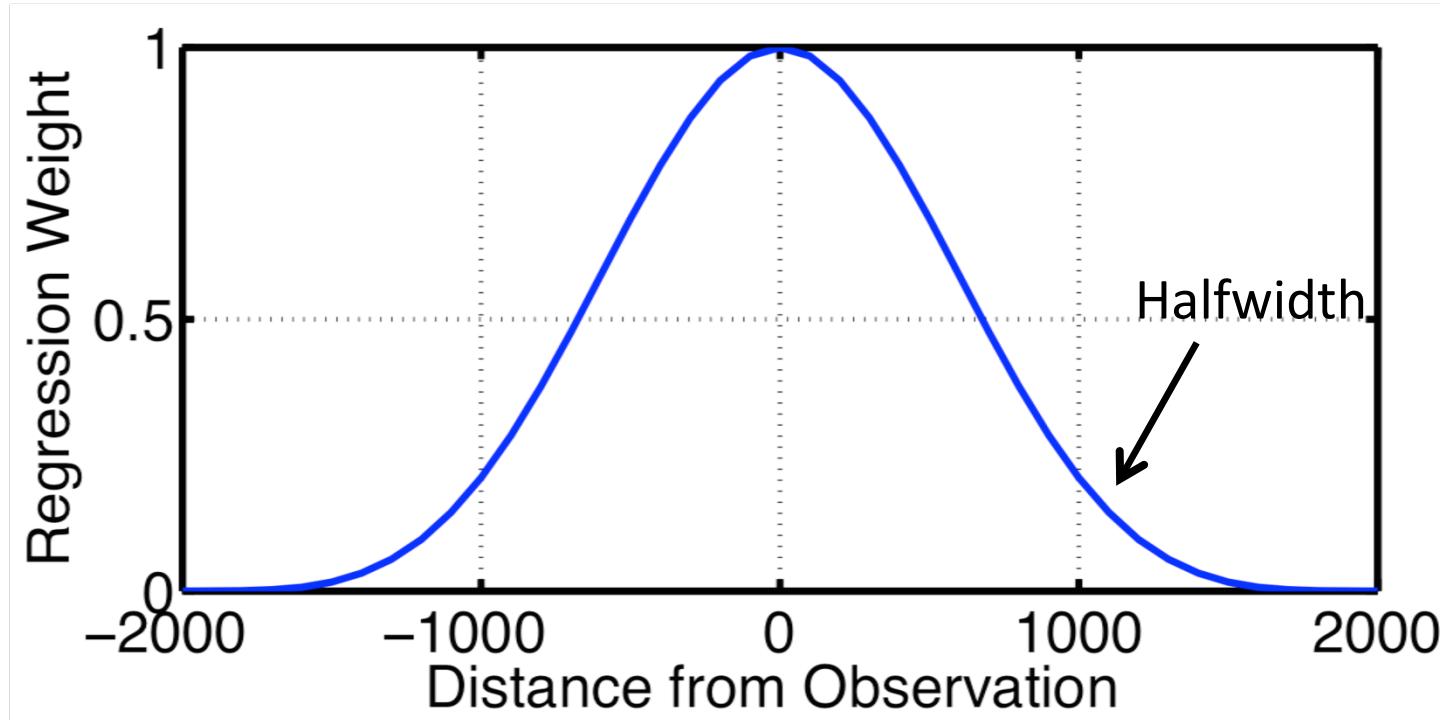
# Ways to deal with Regression Sampling Error

1. Ignore it if number of weakly correlated observations is small and there is a way to maintain prior variance.
2. Use larger ensembles; expensive for large models.
3. Use additional a priori information about relation between observations and state variables.  
*Don't let an observation impact a state variable if they are known to be unrelated.*



# 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.

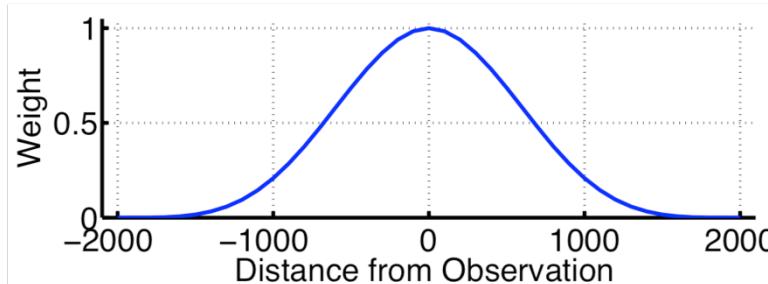
Gaspari-Cohn: 5th order compactly supported polynomial.

This is known as **localization**.

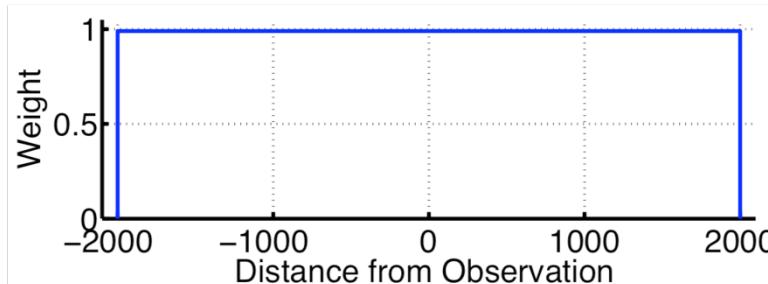
# DART provides several localization options

1. Different shapes for the localization function are available.  
Controlled by *select\_localization* in *&cov\_cutoff\_nml*.

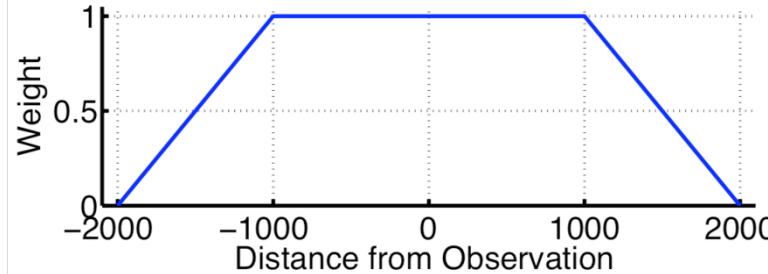
1=> Gaspari-Cohn



2=> Boxcar



3=> Ramped Boxcar



2. Halfwidth of localization function set by *cutoff* in *&assim\_tools\_nml*

# 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.

The results of off-line simulation are tabulated and applied.

(This can be a very useful technique when you're analytically clueless).

To use, set

```
&assim_tools_nml: spread_restoration = .true.
```

# Dealing with Regression Sampling Error

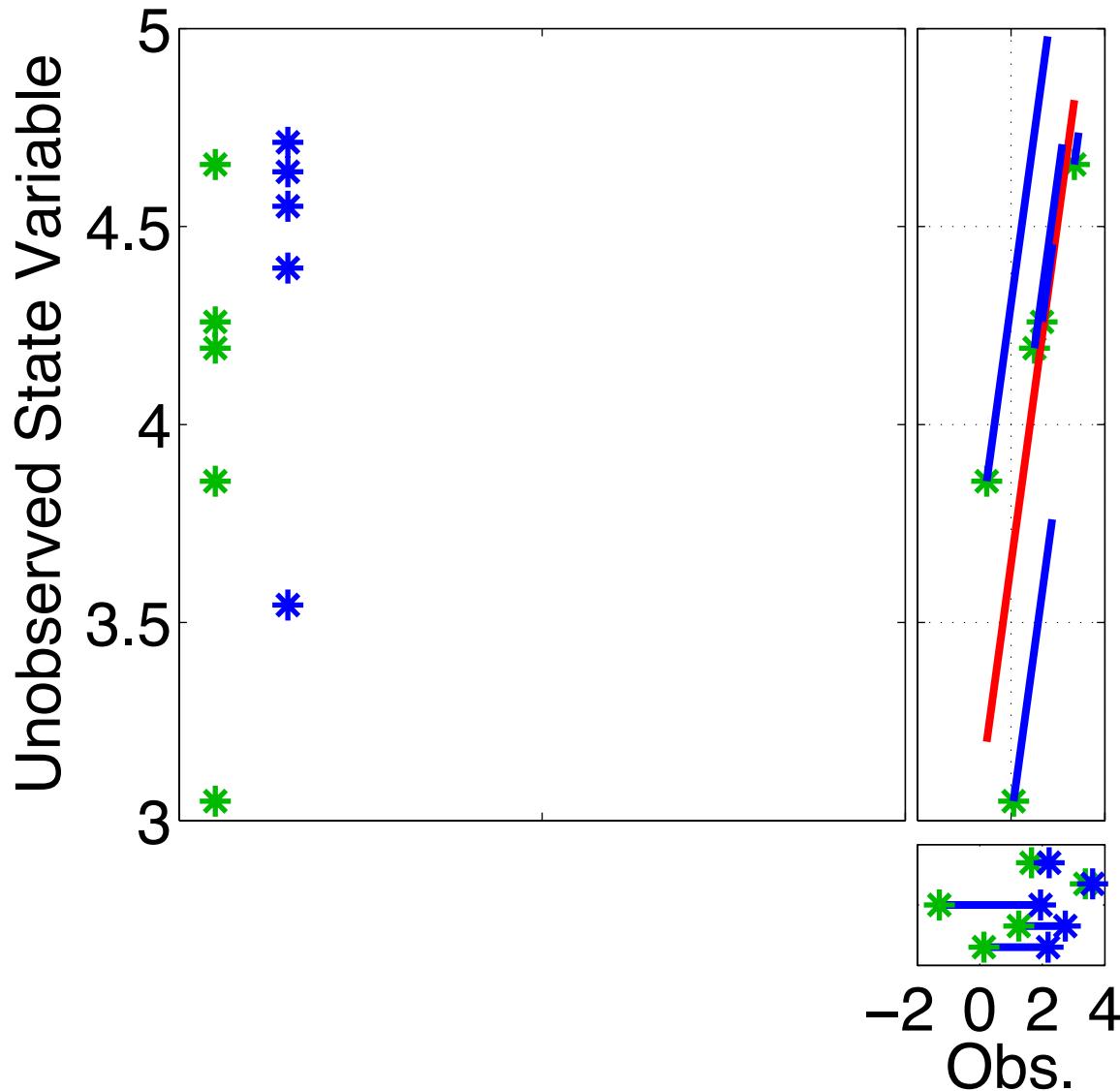
4. Try to determine the amount of sampling error and correct for it.

DART also implements a sampling error correction algorithm that can reduce but not eliminate problems. This algorithm ALMOST ALWAYS IMPROVES large (geophysical) model results.

Try this algorithm: set

```
&assim_tools_nml: sampling_error_correction = .true.
```

# Dealing with Regression Sampling Error



Recall our example of updating an unobserved variable in the previous lecture.

The misalignment of the blue stars at left and lines at upper right is the result of DART's sampling error correction method. This method scales the regression line's slope *downward* by a correction factor that is related to the confidence in the regression (which is small for small samples and near-zero linear correlation coefficients).