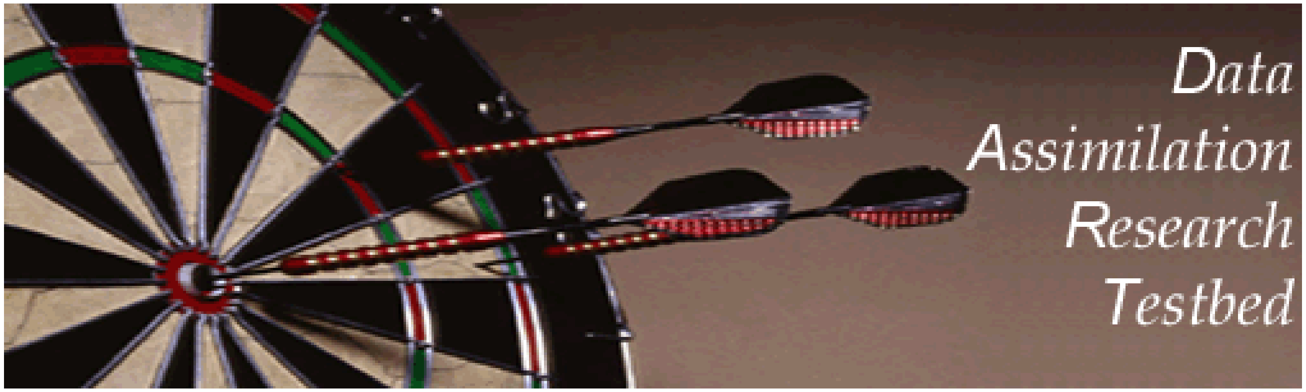
Data Assimilation Research Testbed Tutorial

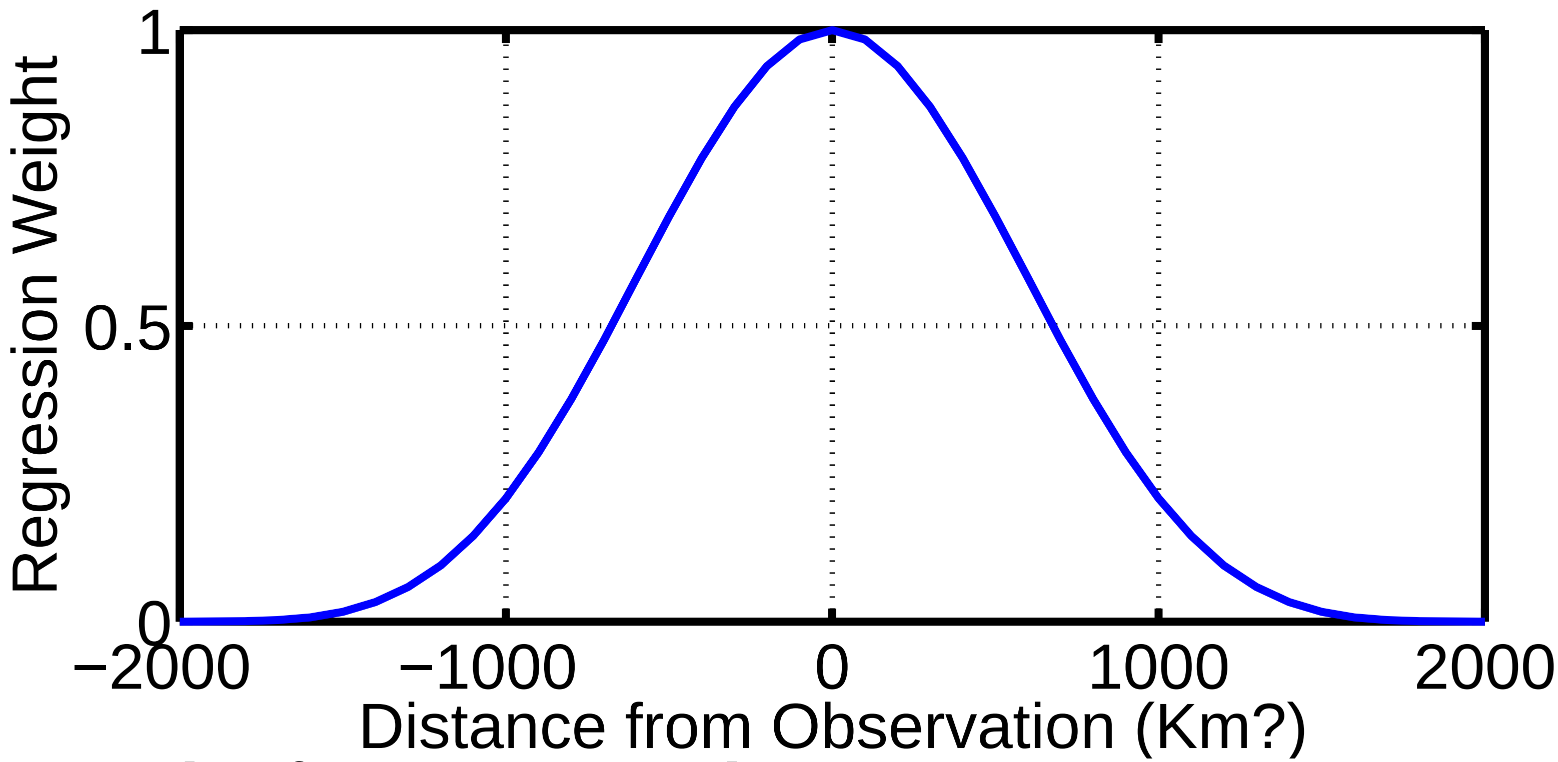


Section 13: Hierarchical Group Filters and Localization

Version 1.0: June, 2005

Ways to deal 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.

Localization is function of expected correlation between obs and state.

Often, don't know much about this.

Horizontal distance between same type of variable may be okay.

What is expected correlation for co-located temperature and pressure?

What about vertical localization? Looks pretty complex.

What about complicated forward operators:

Expected correlation of satellite radiance and wind component?

Note: DART does allow vertical localization for more complex models.

Ways to deal with regression sampling error:

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

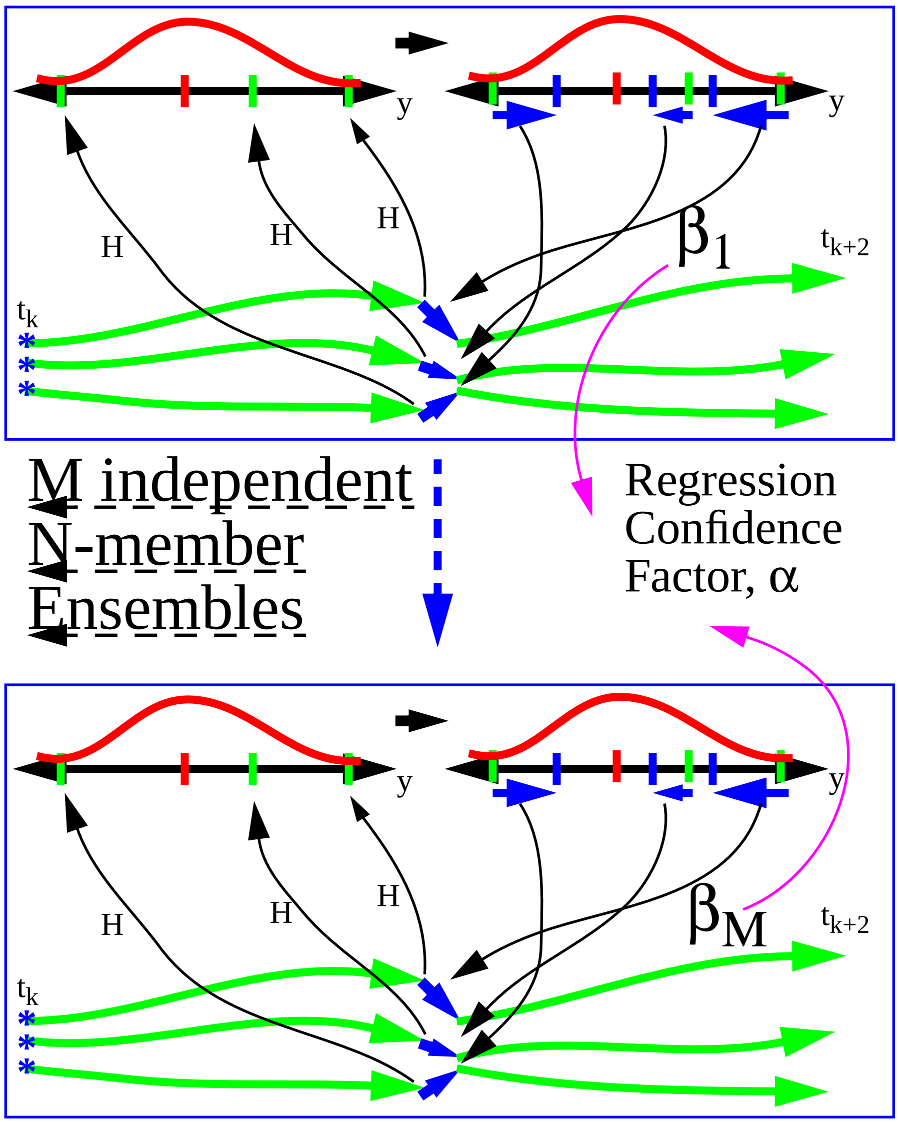
A. Could weight regressions based on sample correlation.
Limited success in tests.
For small true correlations, can still get large sample correl.

B. Do bootstrap with sample correlation to measure sampling error.
Limited success.
Repeatedly compute sample correlation with a sample removed.

C. Use hierarchical Monte Carlo.
Have a ‘sample’ of samples.
Compute expected error in regression coefficients and weight.

Ways to deal with regression sampling error:

4C. Use hierarchical Monte Carlo: ensemble of ensembles.



M groups of N-member ensembles.

Compute obs. increments for each group.

For given obs. / state pair:

1. Have M samples of regression coefficient, β.
2. Uncertainty in β implies state variable increments should be reduced.
3. Compute regression confidence factor, α.

4C. Use hierarchical Monte Carlo: ensemble of ensembles.

Split ensemble into M independent groups.

For instance, 80 ensemble members becomes 4 groups of 20.

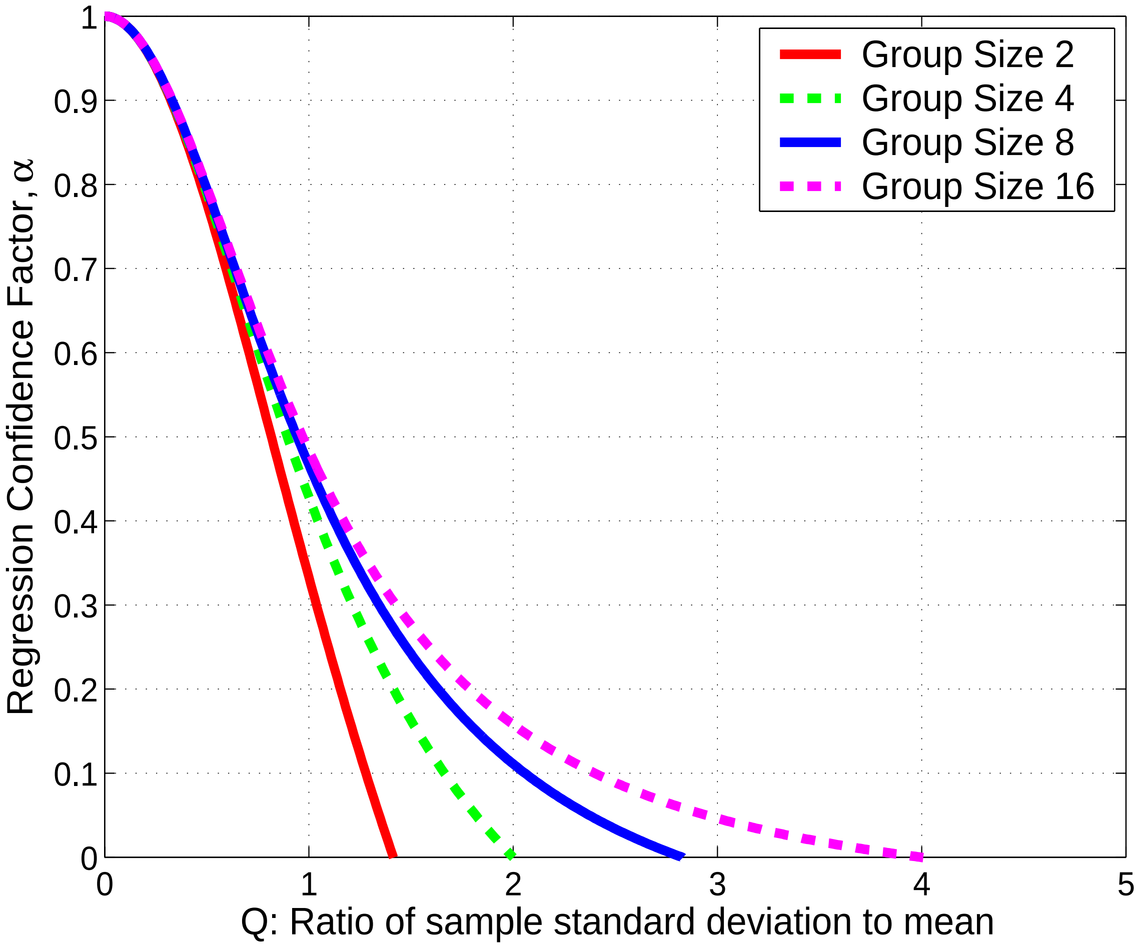
With M groups get M estimates of regression coefficient, β_i.

Find regression confidence factor α (weight) that minimizes:

$$\sqrt{\sum_{j=1}^{M} \sum_{i=1, i \neq j}^{M} [\alpha\beta_i - \beta_j]^2}$$

Minimizes RMS error in the regression (and state increments).

4C. Use hierarchical Monte Carlo: ensemble of ensembles.



Weight regression by α.

If one has repeated observations, can generate sample mean or median statistics for α.

Mean α can be used in subsequent assimilations as a localization.

α is function of M and $Q = \Sigma_\beta / \bar{\beta}$ (sample SD / sample mean regression)

4C. Use hierarchical Monte Carlo: ensemble of ensembles.

Hierarchical filter controlled by setting number of groups, M.
num_groups in *filter_nml*.

If we don't know how to localize to start with, can use groups to help.

Try splitting 80 ensemble members into 4 groups for Lorenz-96.
(4 groups of 20 each).
Use adaptive inflation (0.05 lower bounds) to make things nice.

Can look at the time mean and median value of α.

Essentially an estimate of a 'good' localization for a given observation.

<u>4C. Use hierarchical Monte Carlo: ensemble of ensembles.</u>

After running the 80 by 4 ‘group’ filter, look at plots of α.

Use *plot_reg_factor* in matlab.

Select default input file name.

Only observations 1, 2, 3, and 4 are available:
Located at: 0.39, 0.17, 0.64, 0.86

Think about value of time median vs. time mean.

Could use time mean or median as prior localization functions.

Play around with model error again. What happens to localization?

Lorenz 96 Experimental Design

Initial ensemble members random draws from ‘climatology’

Observations every time step

4000 step assimilations, results shown from second 2000 steps

Covariance inflation tuned for minimum RMS

4 groups of ensembles used unless otherwise noted

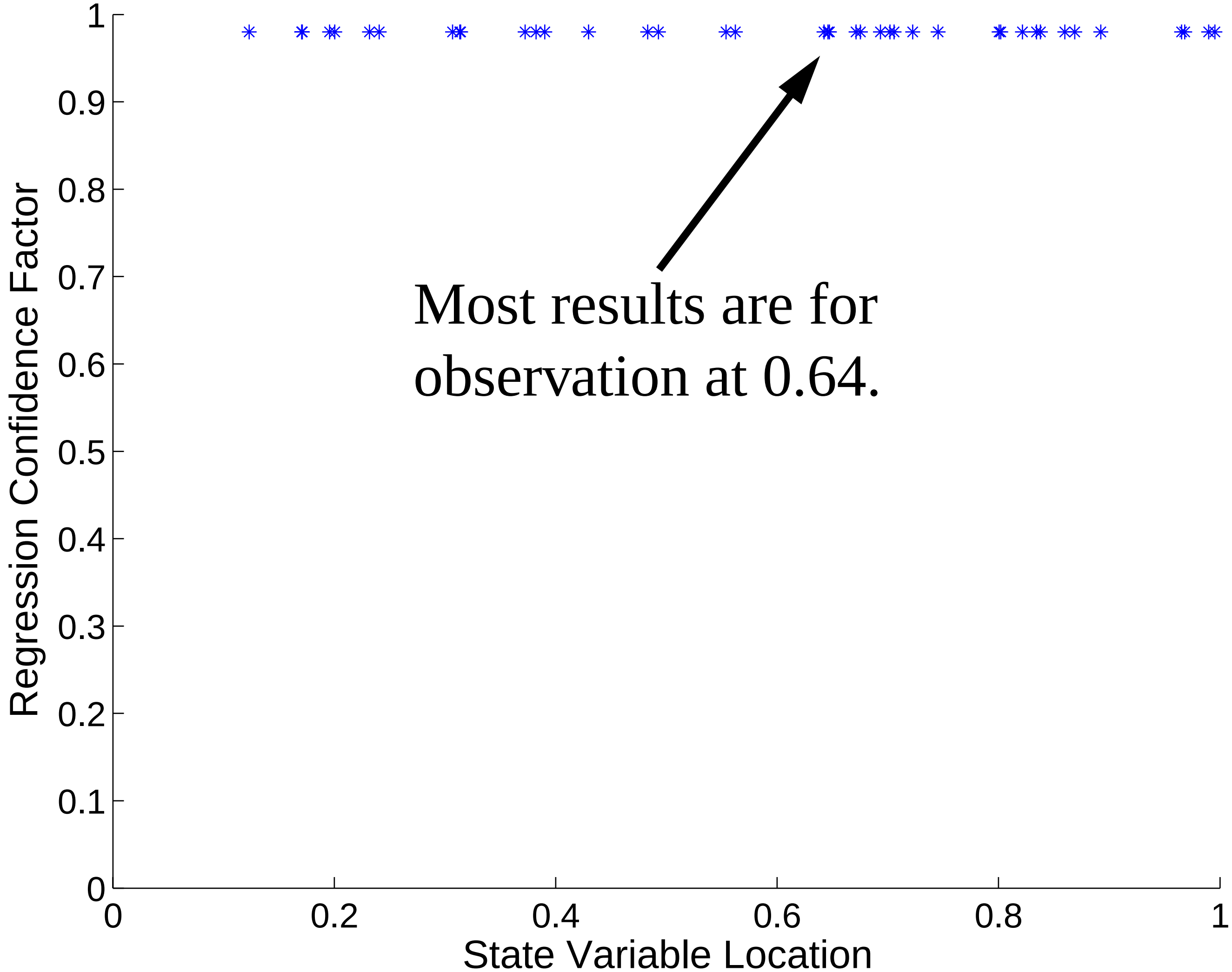
EXPERIMENT SET 1:

40 Randomly located observations

Error variance 10^{-7} (SMALL!)

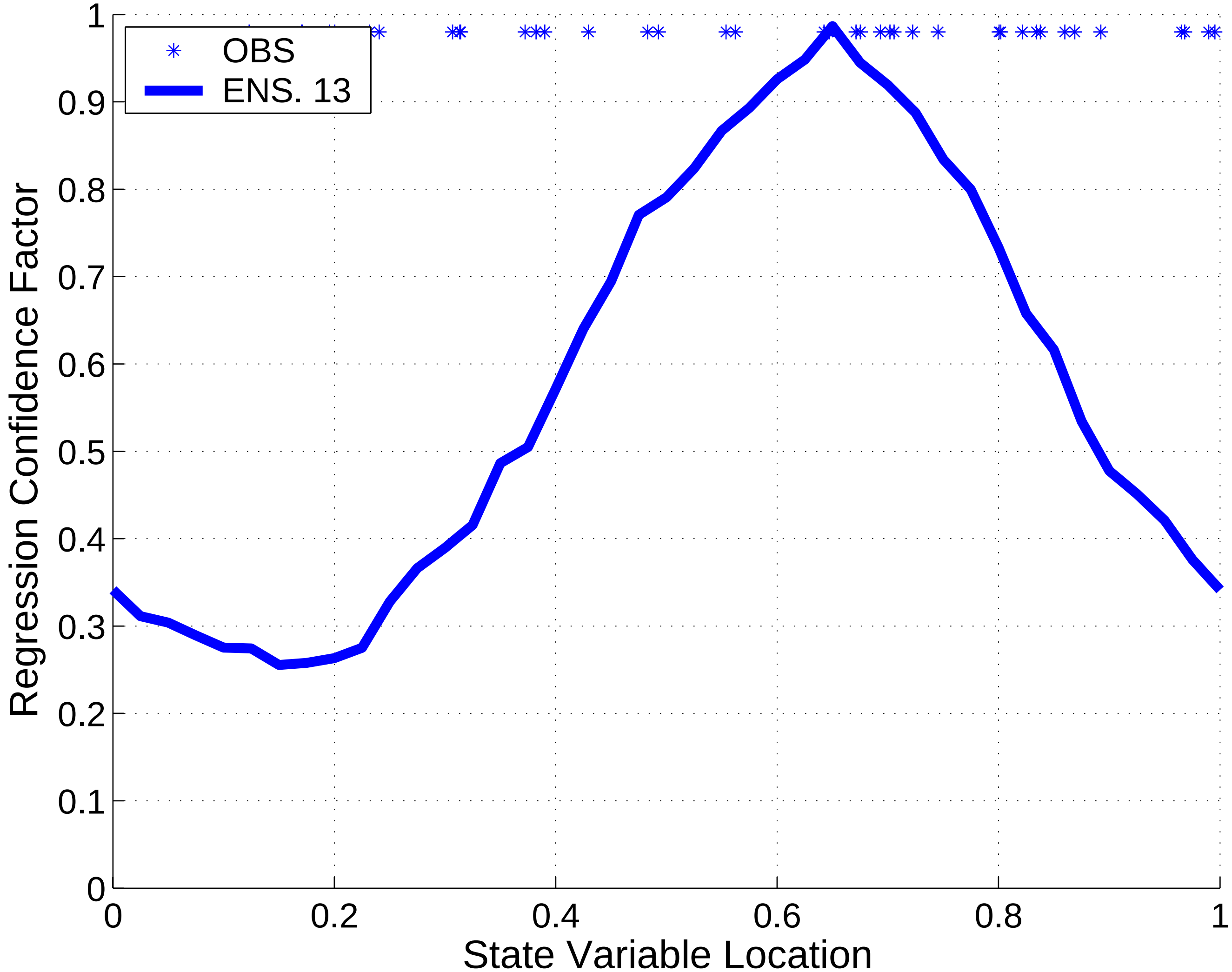
‘ERROR’ comes almost entirely from degeneracy of ensemble covariance
All errors shown are prior ensemble mean estimates

Time Mean Regression Confidence Envelopes: Small Error Limit



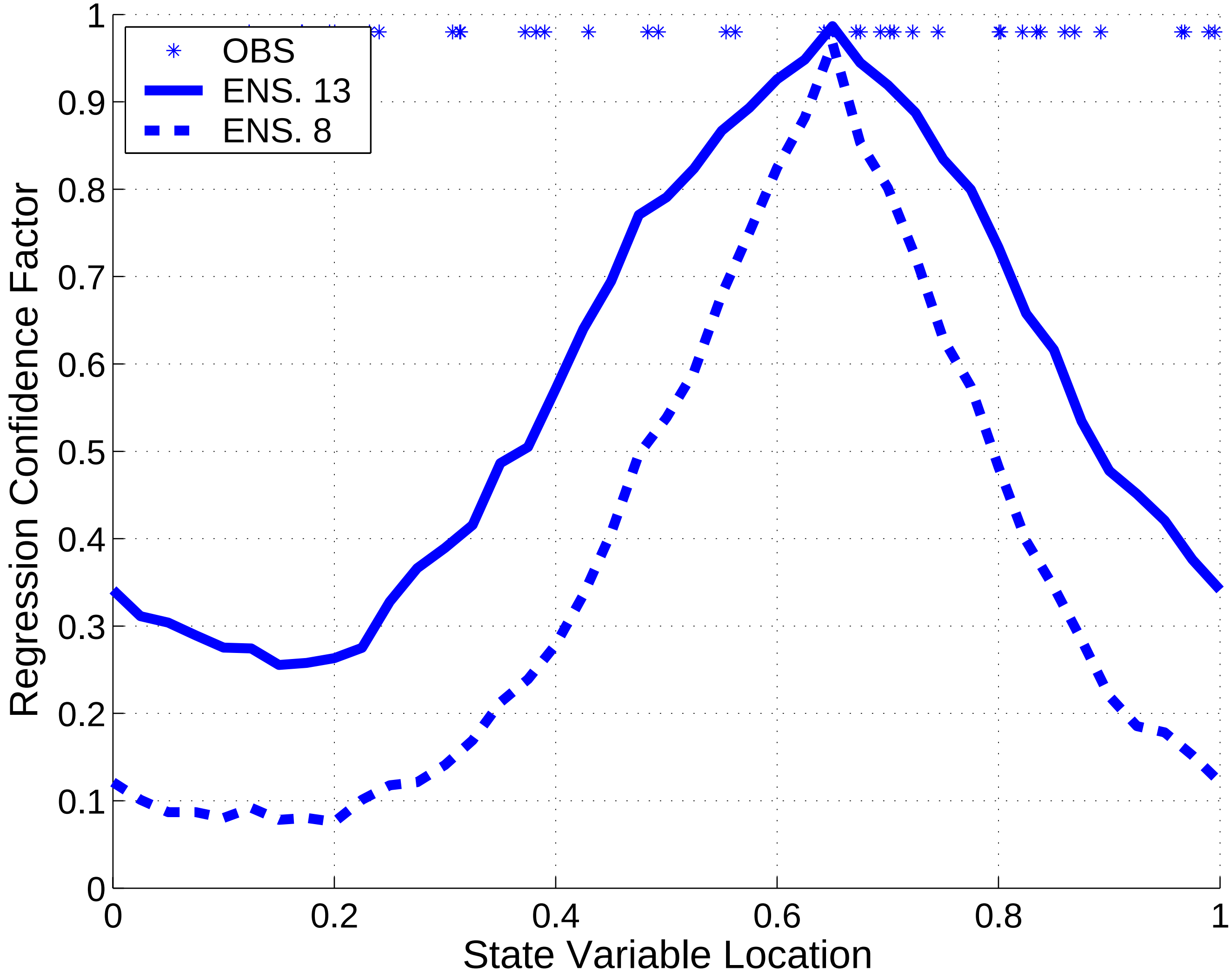
Location of 40 randomly located observations

Time Mean Regression Confidence Envelopes: Small Error Limit



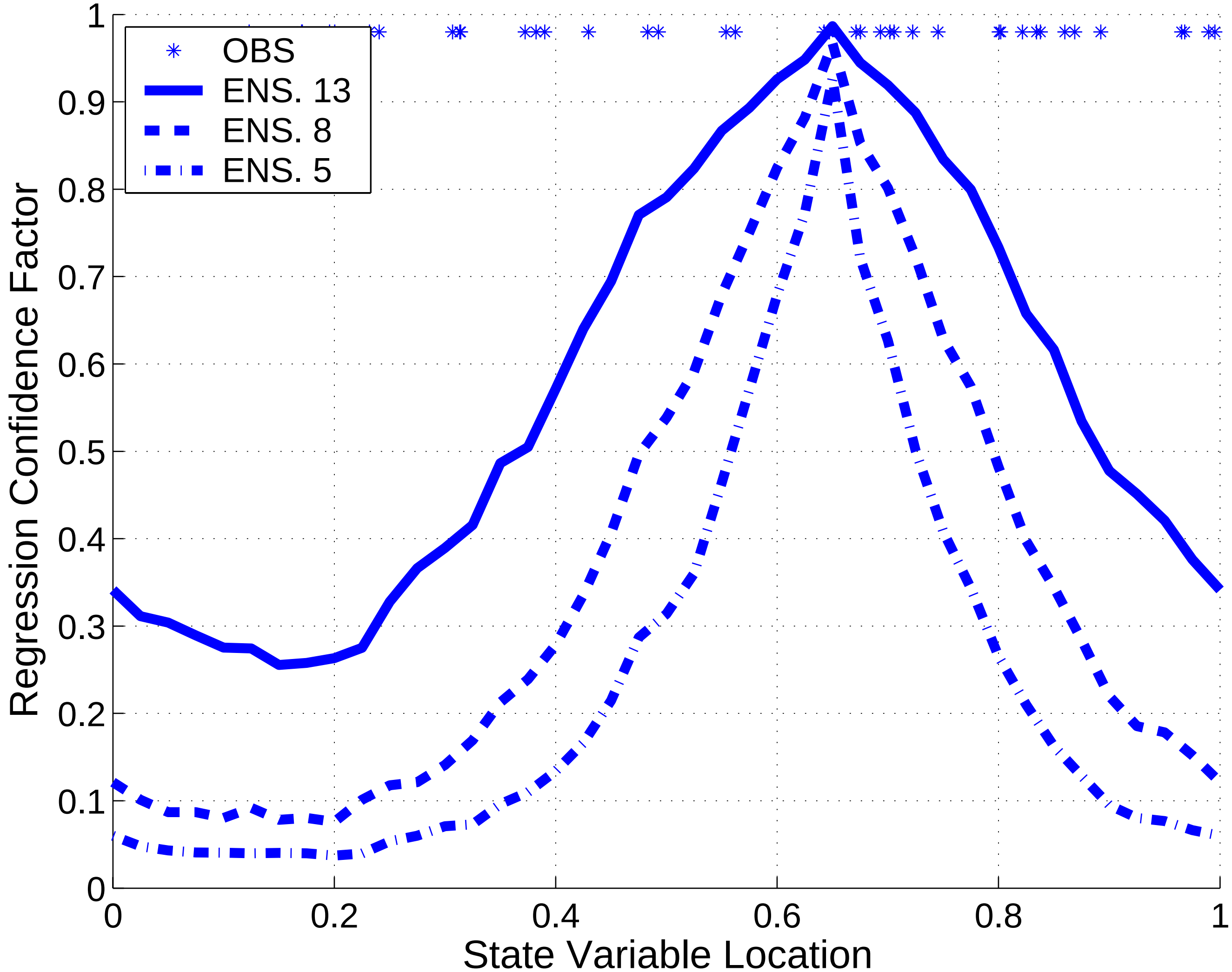
Envelope (localization) for 13 member ensemble (barely degenerate)

Time Mean Regression Confidence Envelopes: Small Error Limit



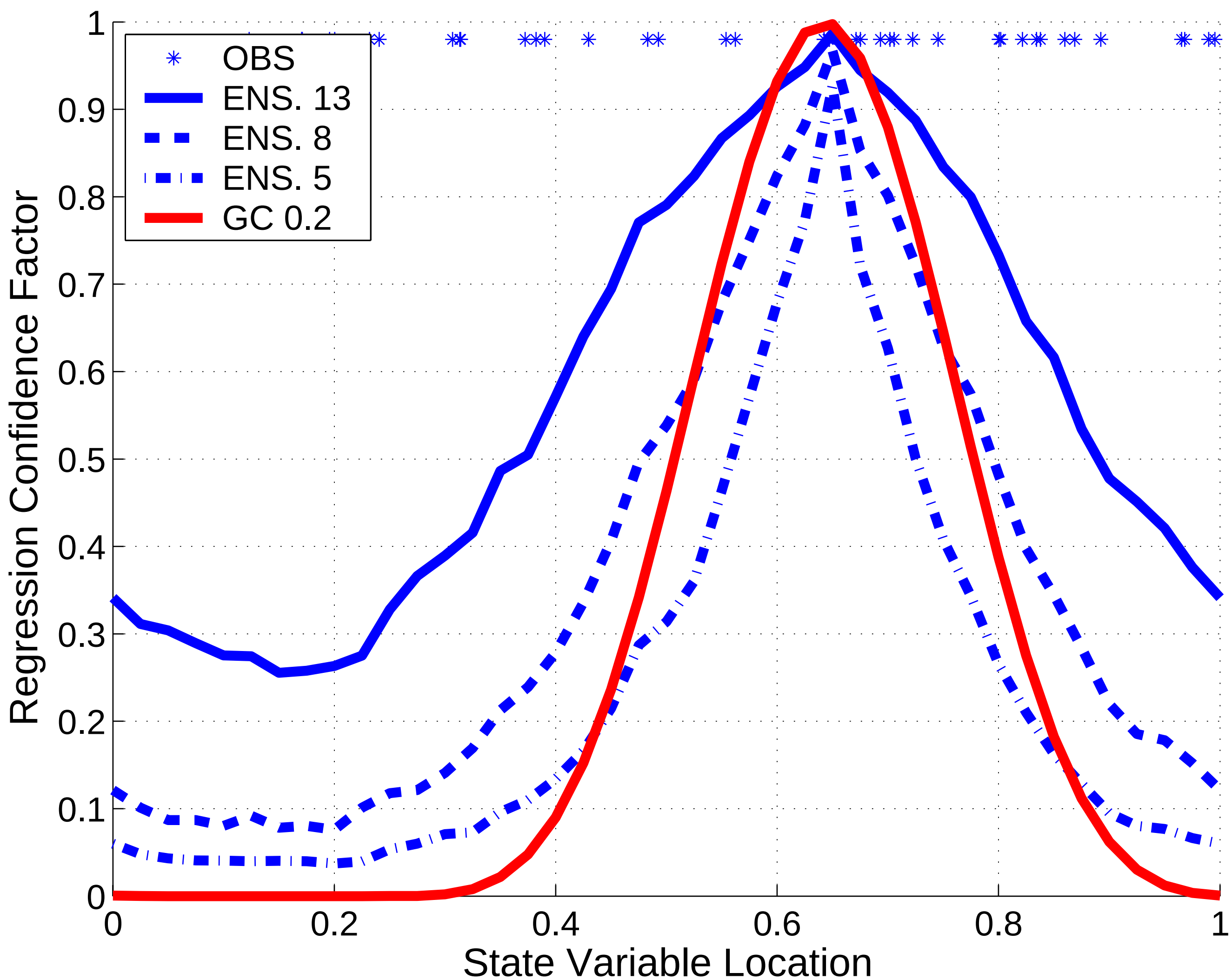
Envelope for 8 member ensemble

Time Mean Regression Confidence Envelopes: Small Error Limit



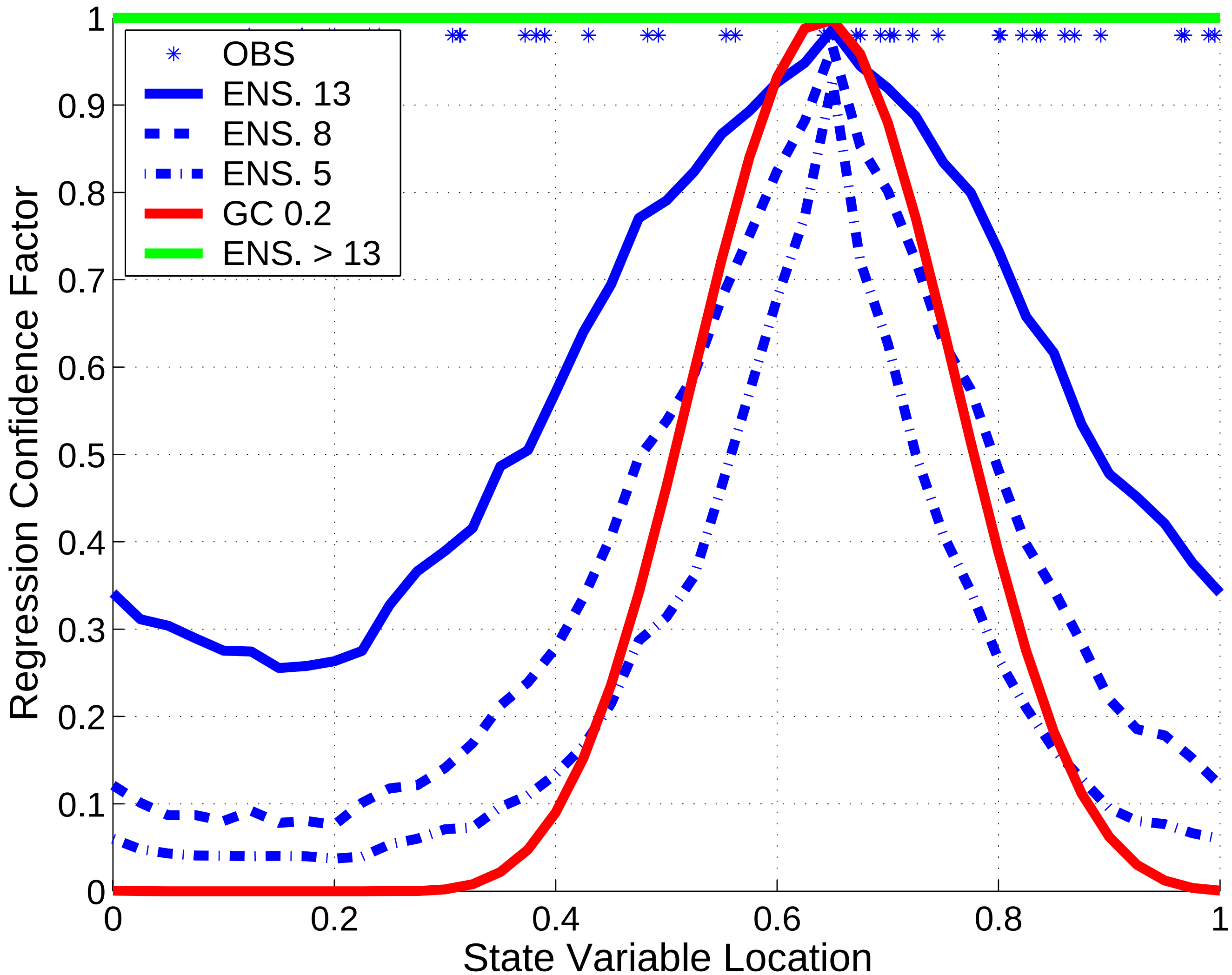
Envelope for 5 member ensemble

Time Mean Regression Confidence Envelopes: Small Error Limit



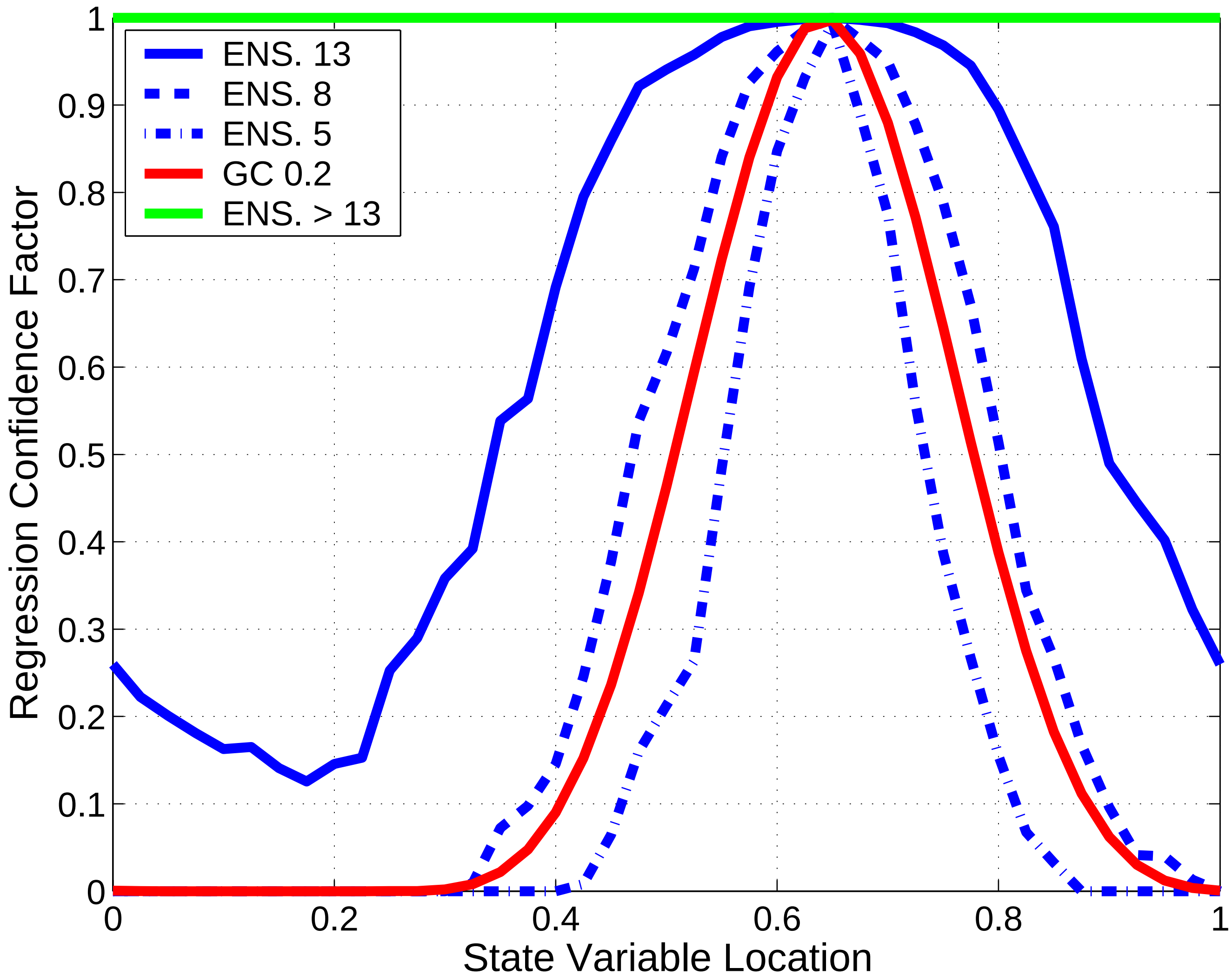
Compare to Gaspari Cohn with half-width 0.2

Time Mean Regression Confidence Envelopes: Small Error Limit



Ensemble sizes > 13 require NO LOCALIZATION (no significant error)

Time Median Regression Confidence Envelopes: Small Error Limit



Median reduces noise for small expected correlations

Additional Experiments: Experimental Design

1. Base case: plain ensemble, optimized Gaspari Cohn localization half-width

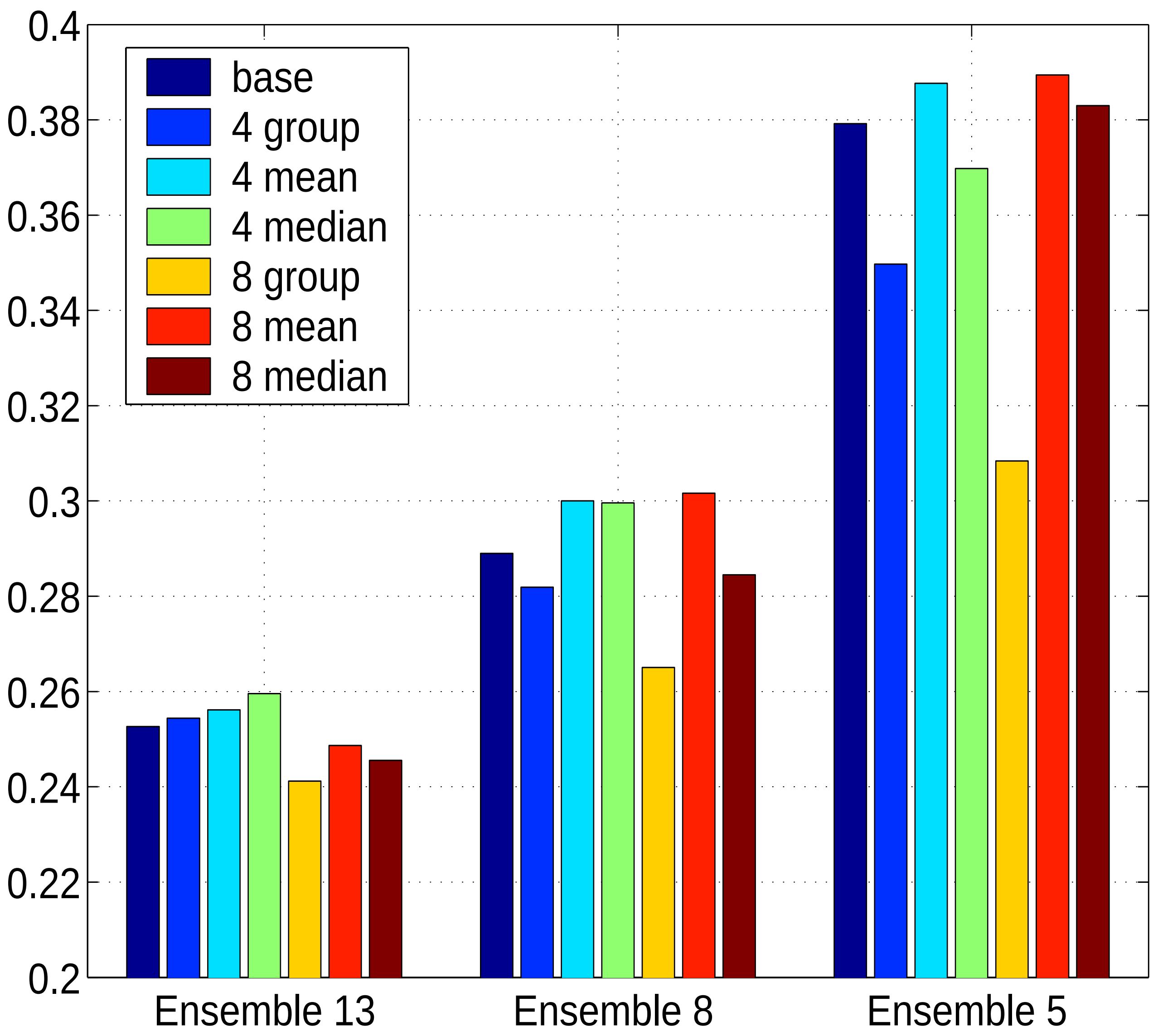
2. Time mean case: plain ensemble but with localization using time mean regression confidence envelope from group assimilation

3. Time median case: as above but with time median from group

All Start from group 1 ensemble at time step 2000

Covariance inflation tuned independently for each case

Time Mean global mean RMS Error: Small Error Results



Error grows with reduced ensemble size

8 Group filters always best

Time mean and median close to tuned base case

Mean/median cost same as base case

Experimental Design: Varying Observational Error Variance

Observation set as before

Observation error variance 10^{-5}, 10^{-3}, 0.1, 1.0, 10.0, 10^{7}

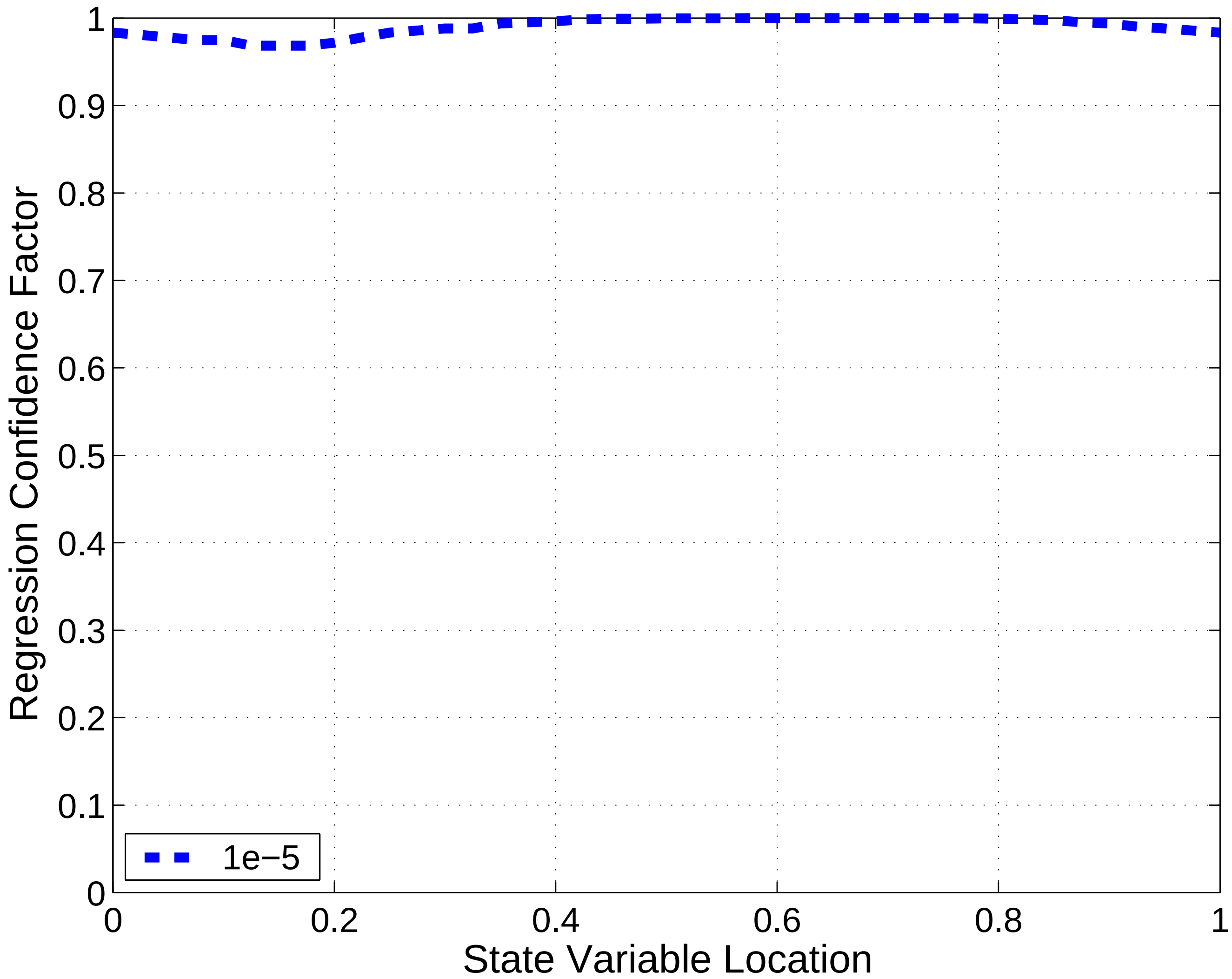
14 member ensembles; not degenerate

Error source is now from observation limitations, etc.

Claim: behavior is similar, error source is irrelevant for correction

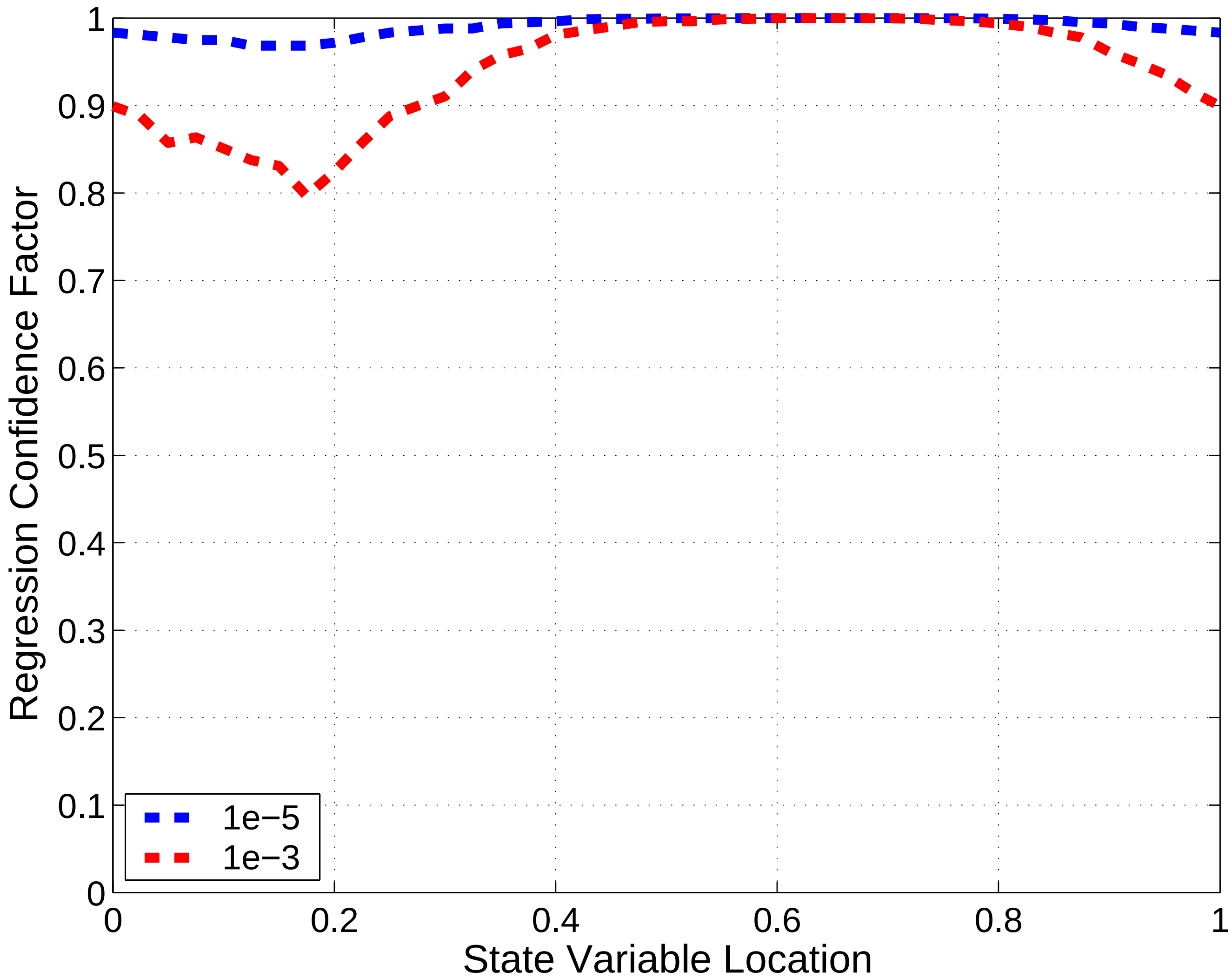
Understand degeneracy and other error sources as sampling error

Time Median Envelopes: Varying Obs. Error Variance

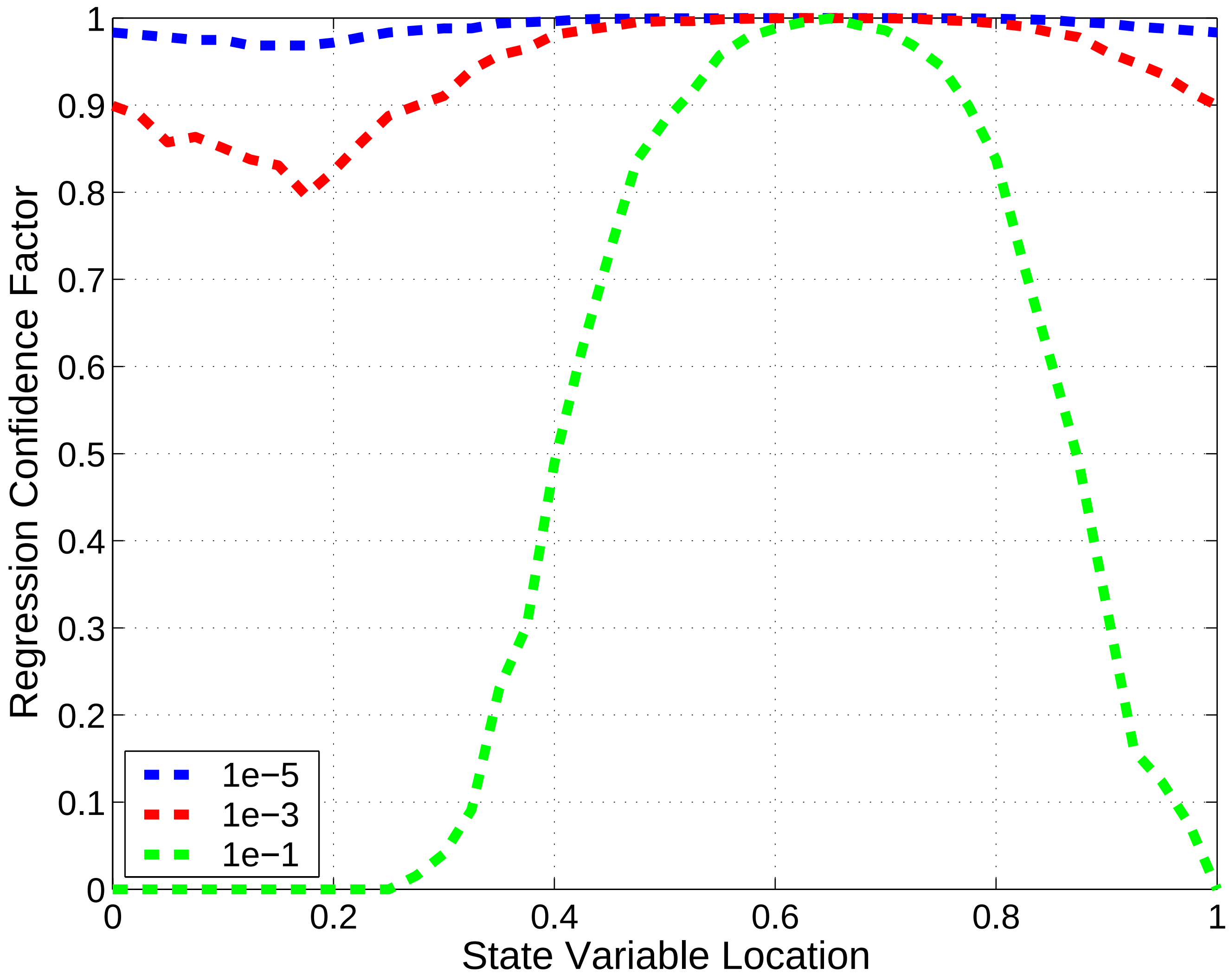


Small error implies no need for localization

Time Median Envelopes: Varying Obs. Error Variance

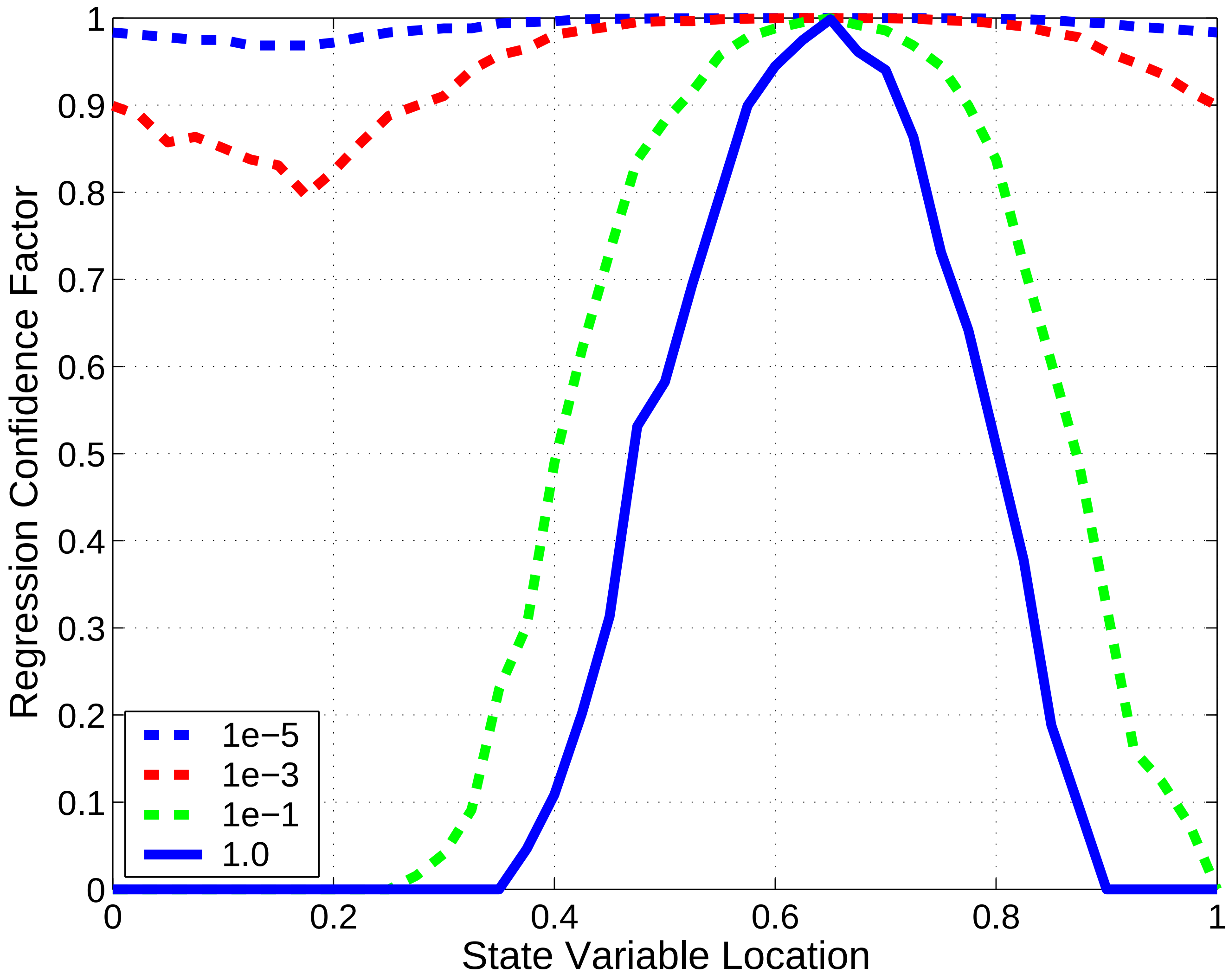


Time Median Envelopes: Varying Obs. Error Variance

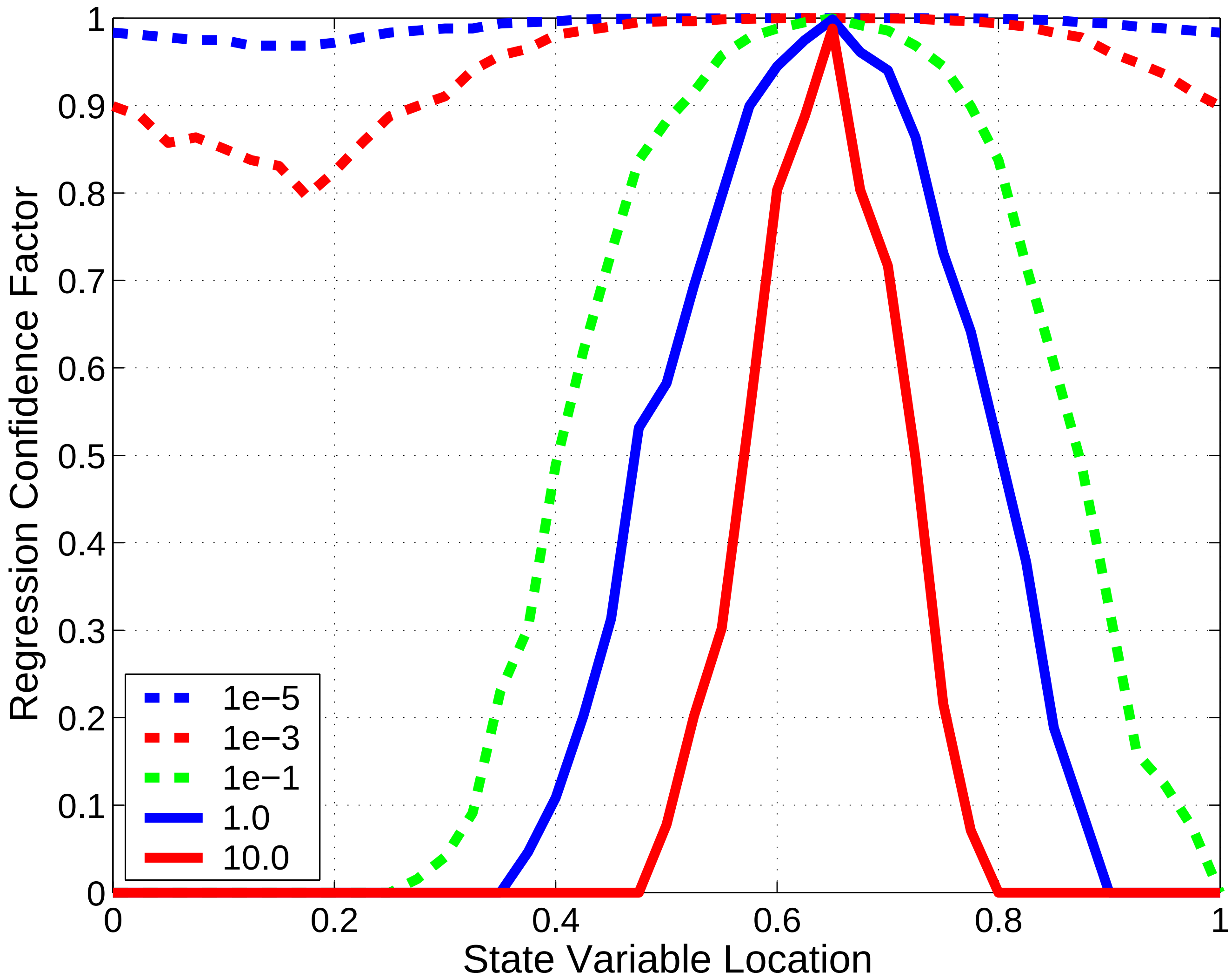


Increasing error implies increasing localization

Time Median Envelopes: Varying Obs. Error Variance

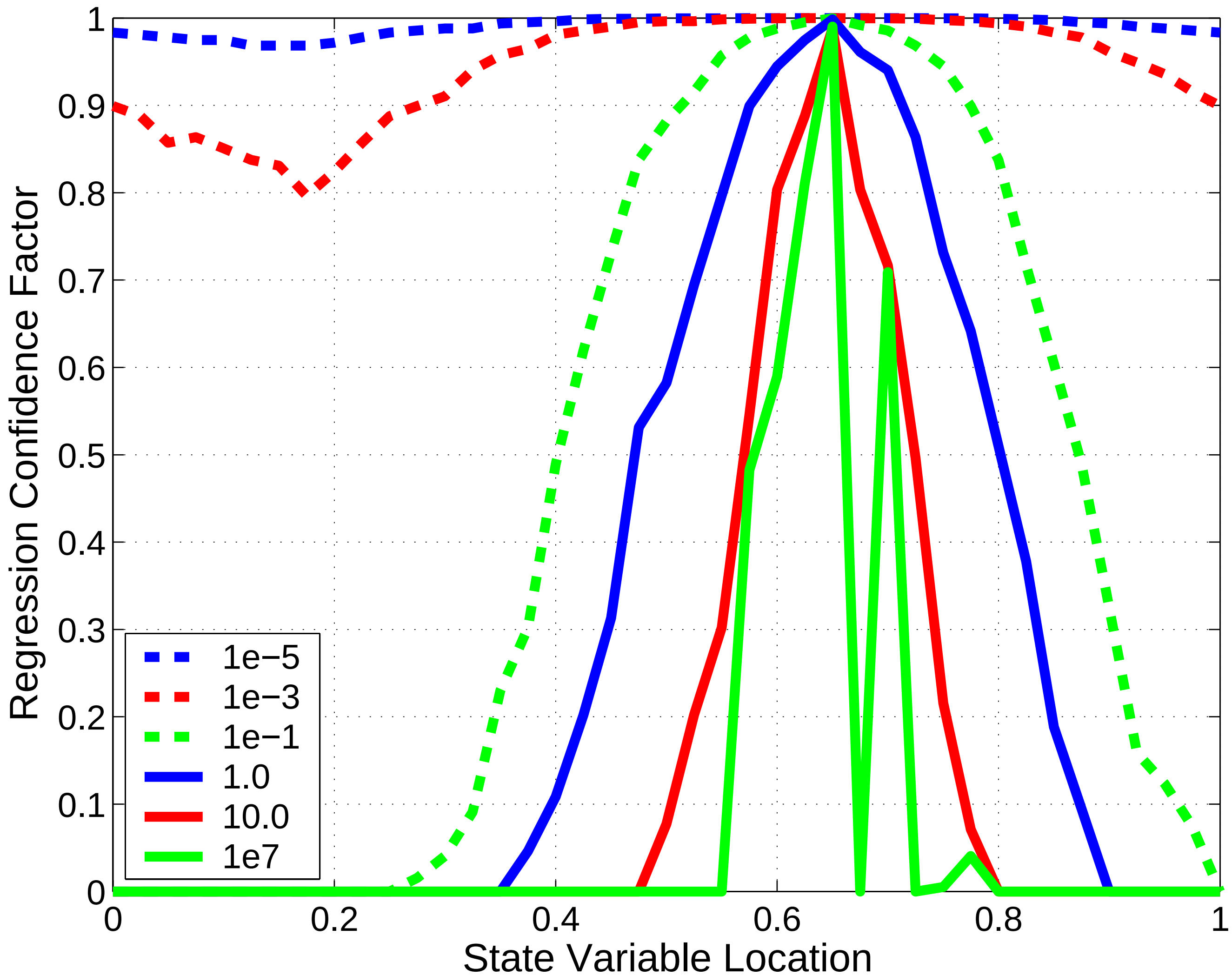


Time Median Envelopes: Varying Obs. Error Variance



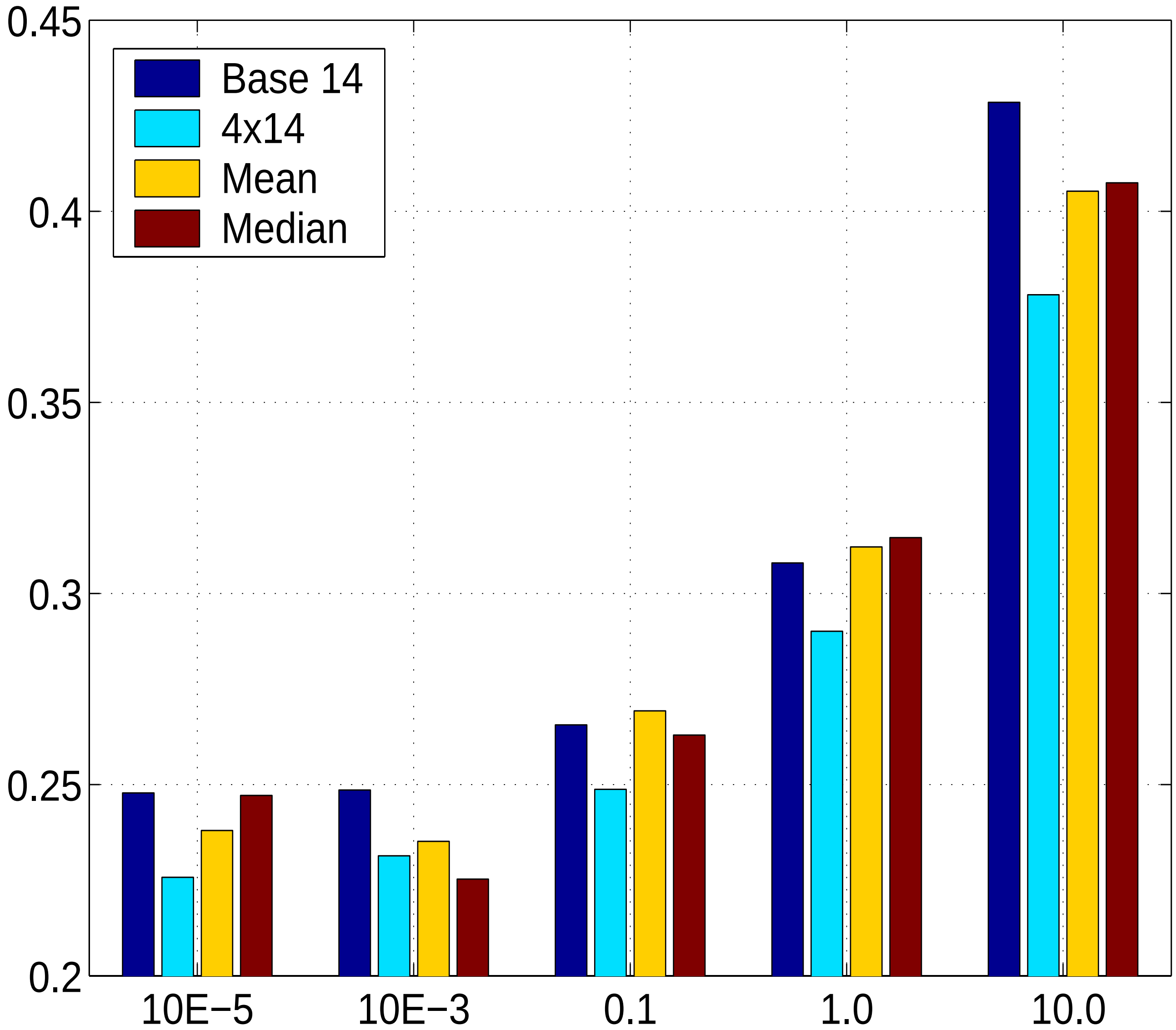
Single Gaspari Cohn half-width can't deal with this range of errors

Time Median Envelopes: Varying Obs. Error Variance



Climatological case is unique: Looks like time mean coherence

Time Median Envelopes: Varying Obs. Error Variance



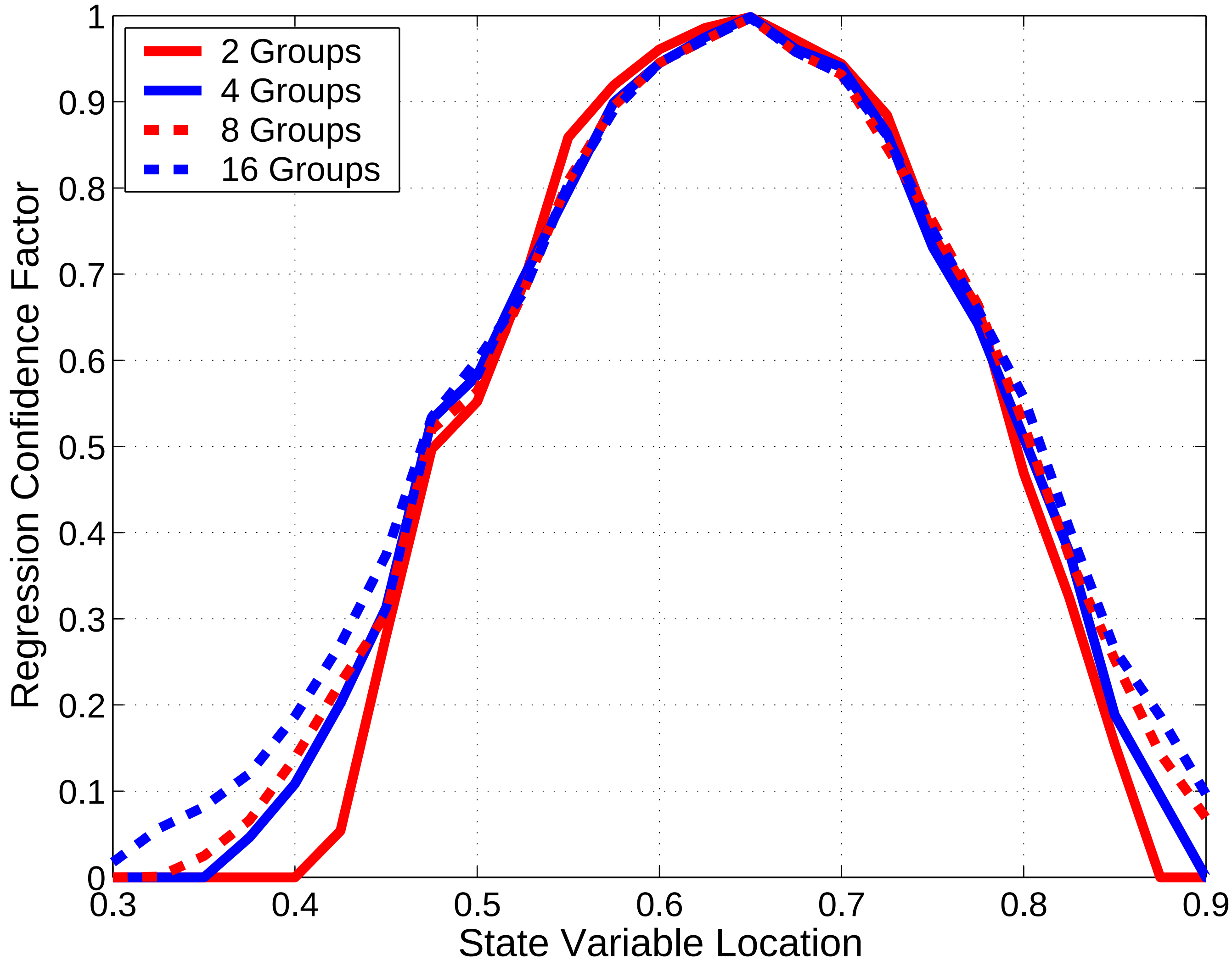
Error scaled by obs. error standard deviation

As error grows, filter becomes less 'efficient'

Things are more 'nonlinear' for large errors

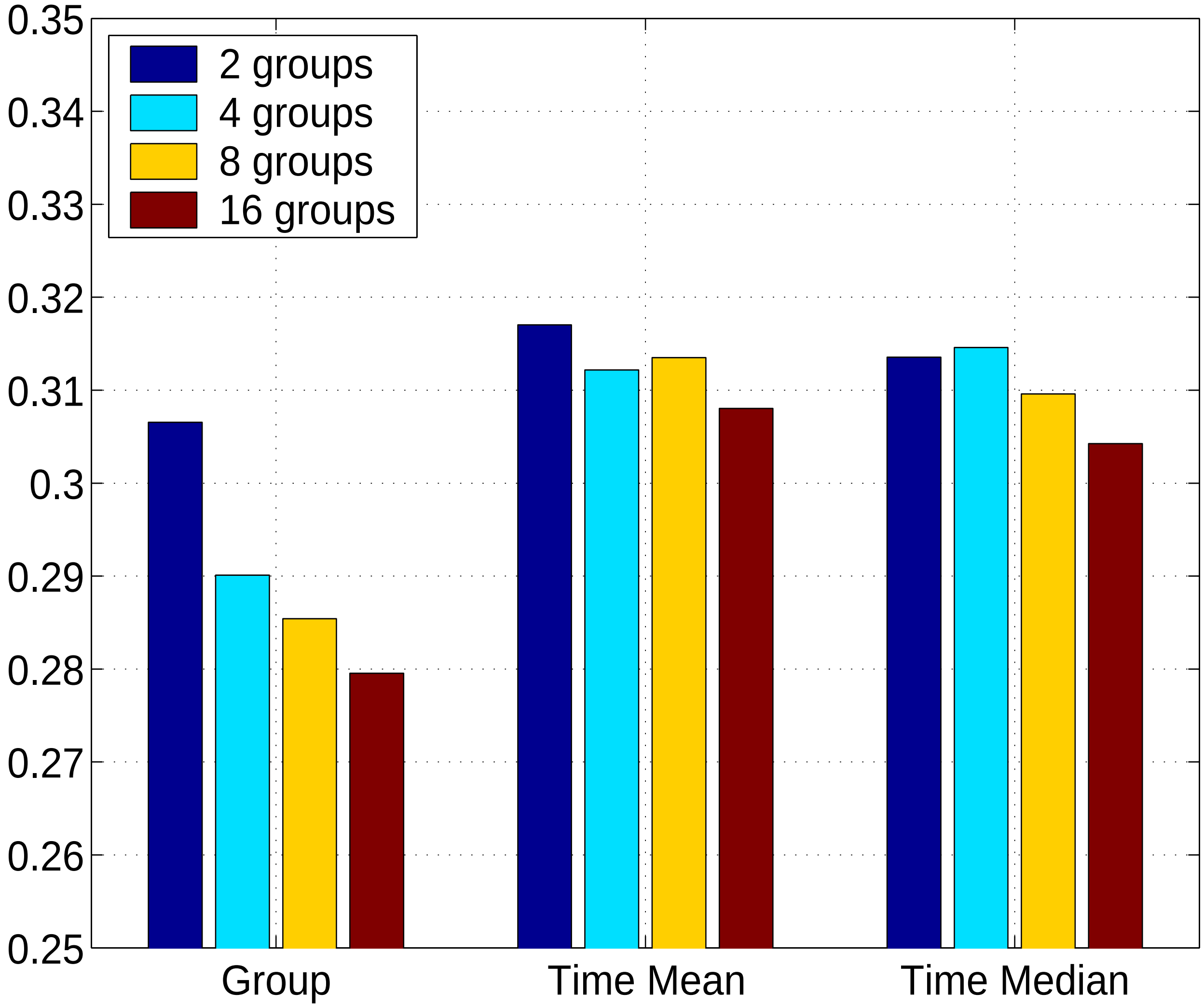
Group, mean and median compare favorably with tuned base for all cases!

Sensitivity of results to group size



Obs. error variance 1.0, 14 member ensemble case

Sensitivity of results to group size



Gradual improvement for increased group size

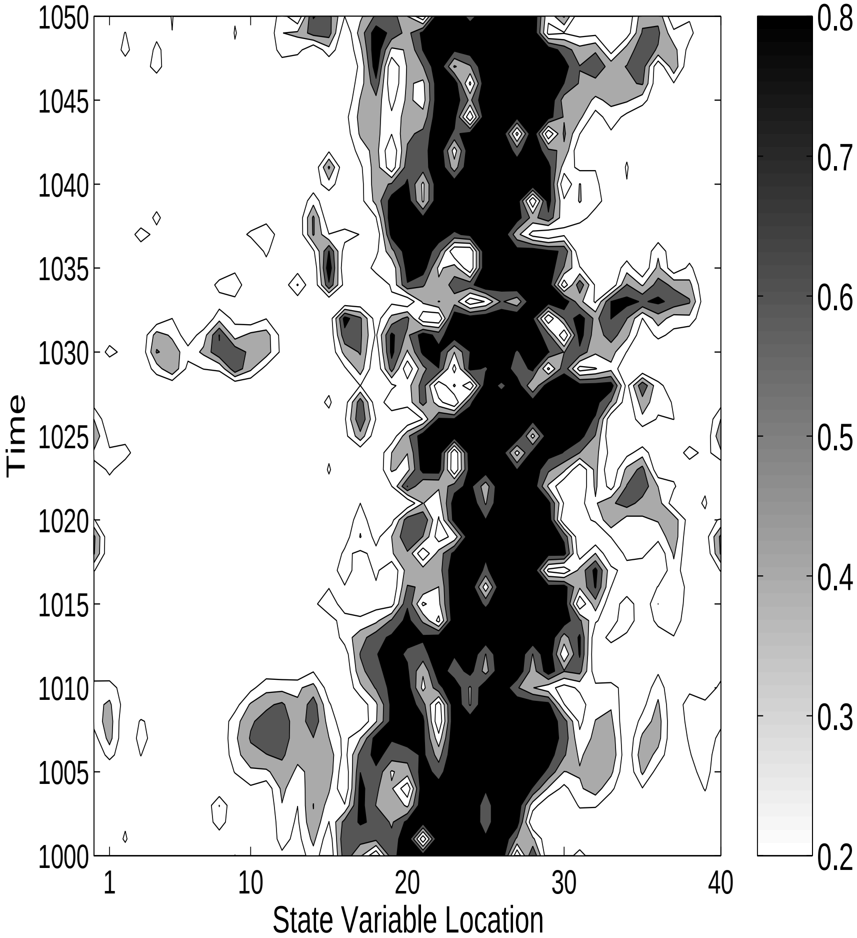
Time mean/median improve more slowly

Important that small groups give good results (for cost)

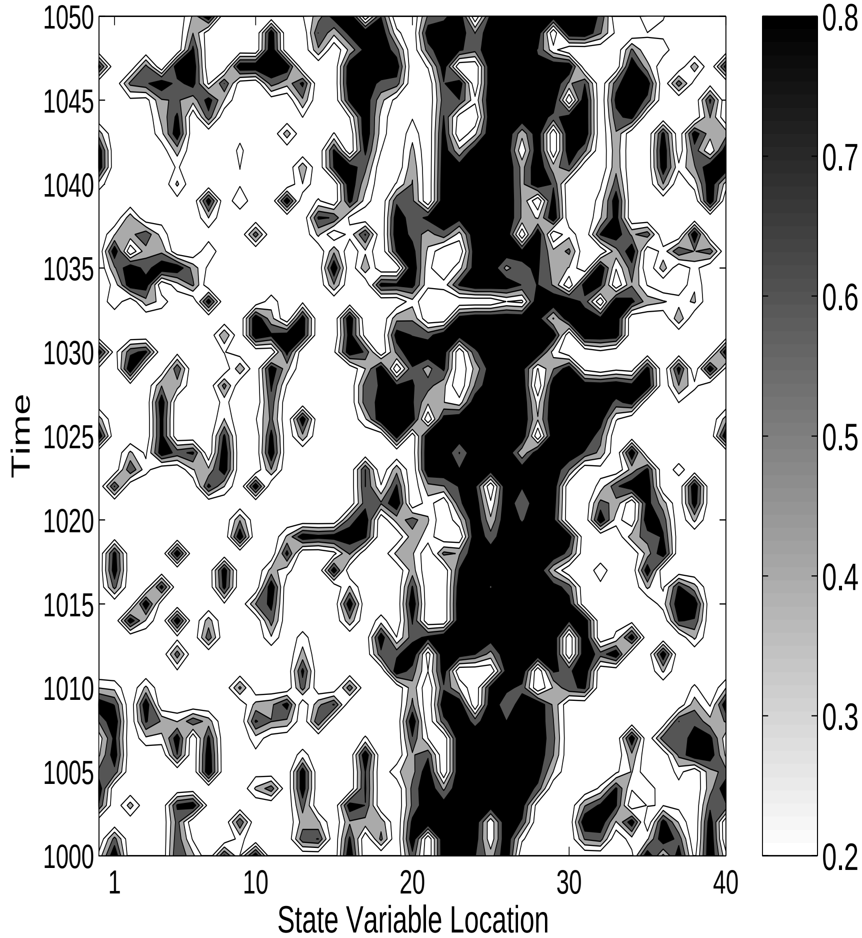
Group better than time mean implies some time varying information (time means should reduce noise)

Obs. error variance 1.0, 14 member ensemble case

Time variation of regression confidence factor



16 groups

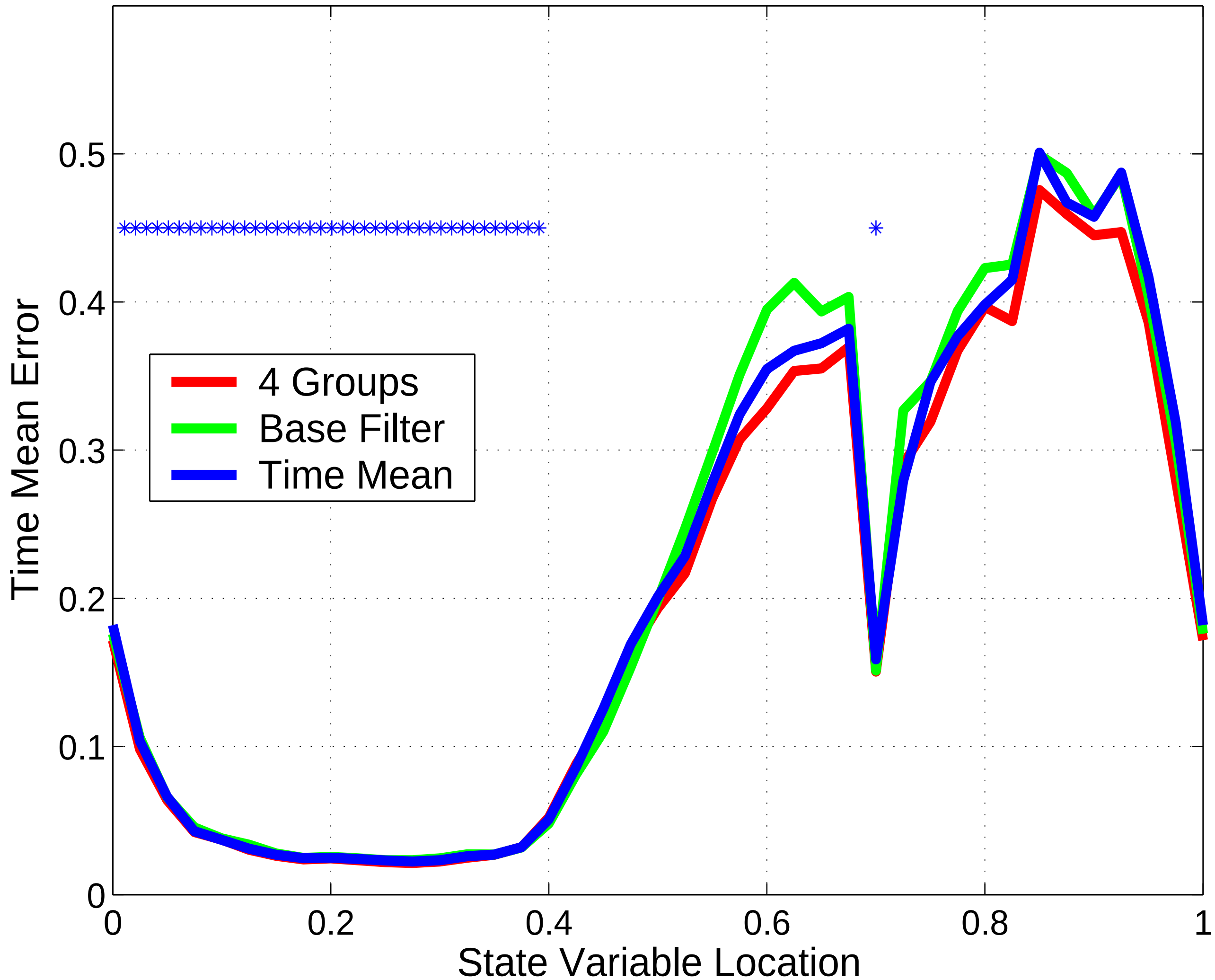


2 groups

Some consistent time variation; noise dominates far from obs.

Notice behavior around step 1033 for instance

Challenging Traditional Localization: Varying spatial obs. density



39 Dense obs

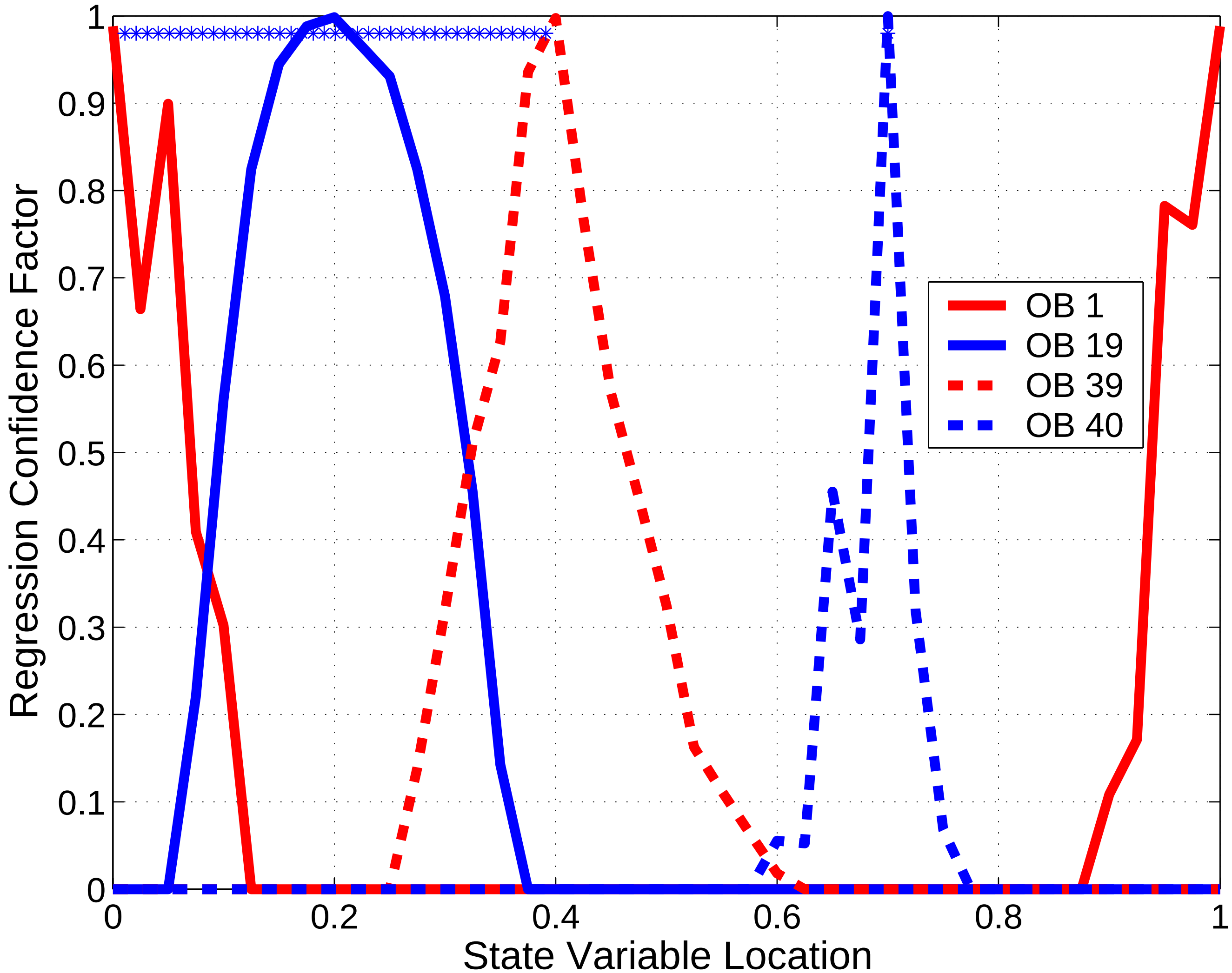
1 isolated ob.

Group filter does better in sparse region

Time mean filter perfor-mance in between

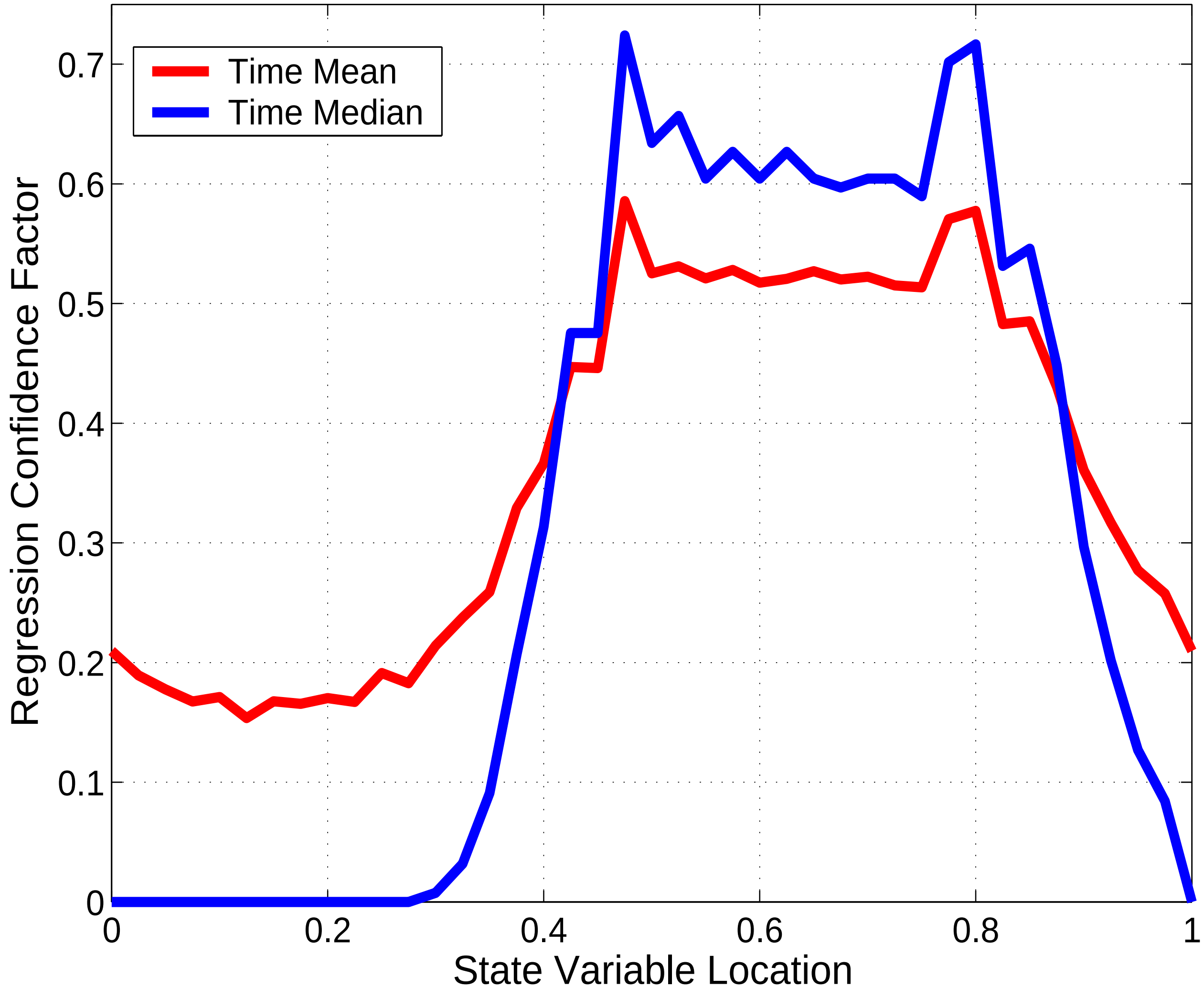
Time Mean Prior RMS Error as function of spatial location

Challenging Traditional Localization: Varying spatial obs. density



Time median envelopes vary lots with spatial location of obs.

Challenging traditional localization: Spatial-mean obs.



Forward observation operator is fifteen grid point mean

Envelope has become less Gaussian

In one-d L96 domain, everything stays pretty Gaussian

But..., <u>group filter does eliminate need for tuning localization!</u>

Assimilating observations at times different from state estimate

Ensemble smoothers: use future observations

Targeted observations: examine impact of obs. in past

Real-time assimilation: use of late arriving observations in forecast

Expect correlations to diminish as time separation increases

Need a ‘localization’ in time, too

Group filter can provide this

Time ‘localization’: Experimental design

4 group, 14 ensemble member filter

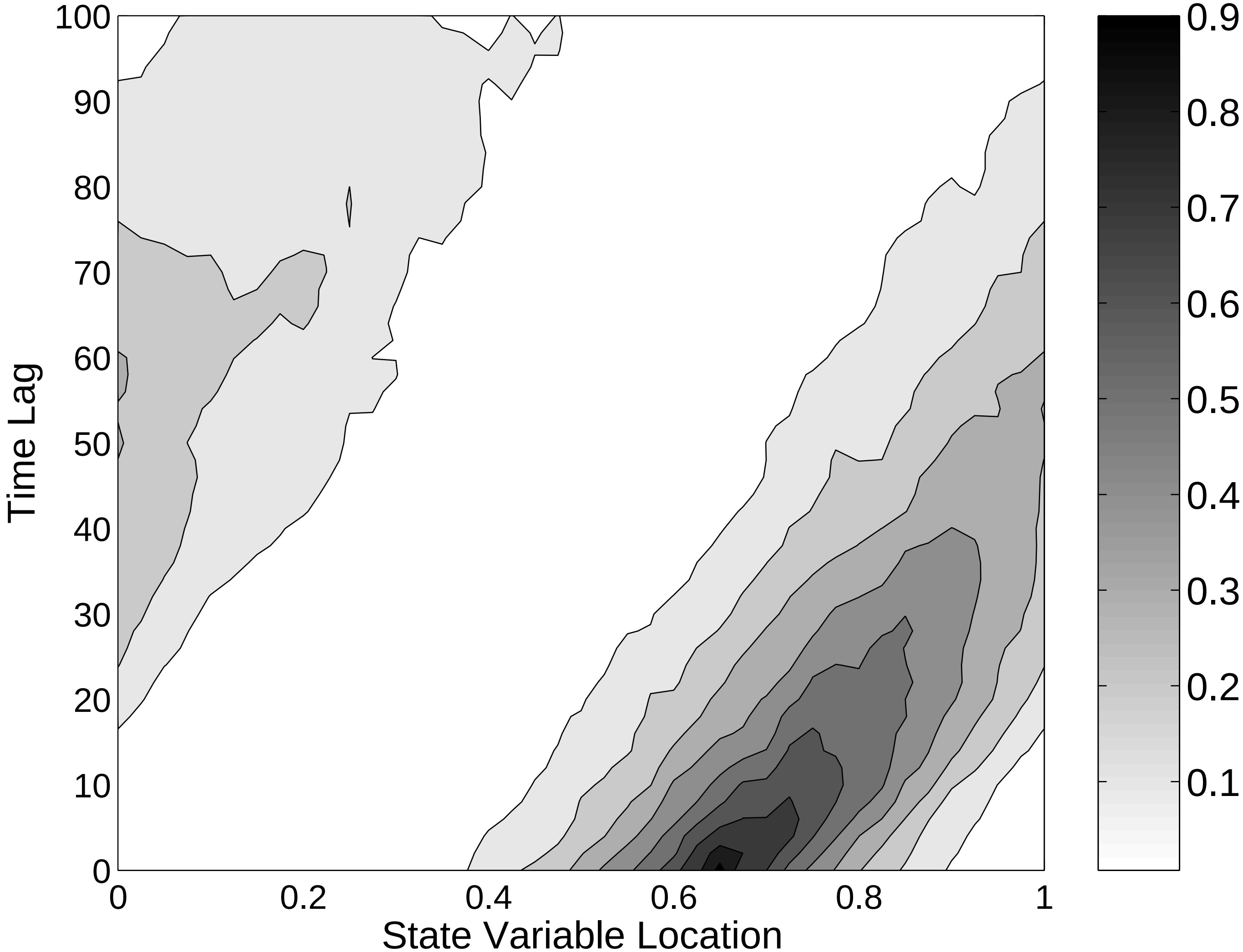
40 random obs. with 1.0 error variance

1 additional observation at location 0.642

The additional observation is from a prior time step

Time mean regression confidence envelope as function of time lag

Regression confidence factor as function of obs. lag time

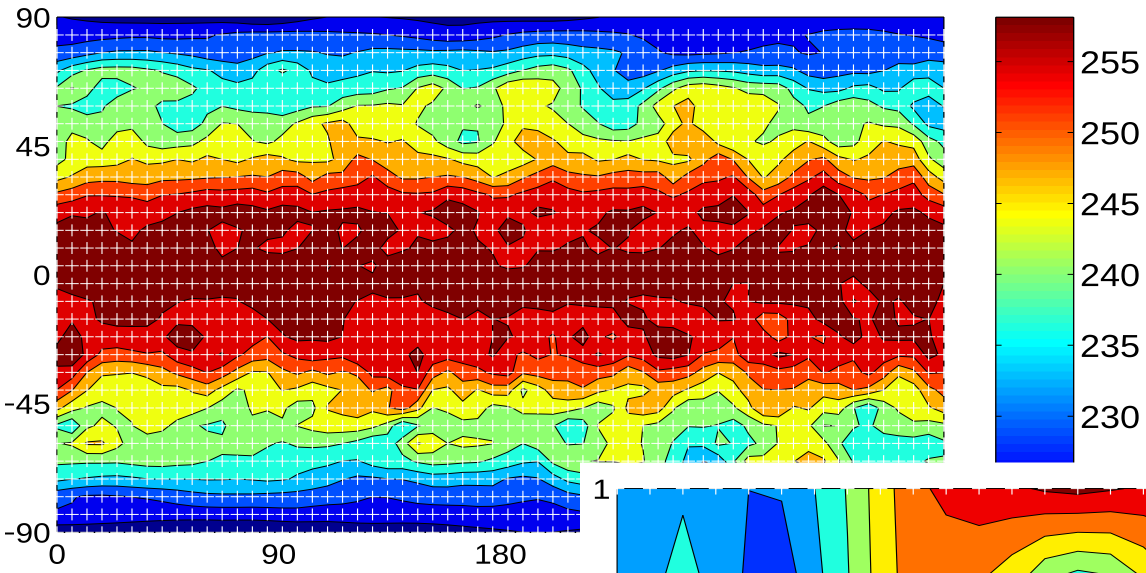


Moves with group velocity (approximately); dies off with lead

Assimilation in Idealized AGCM: GFDL FMS B-Grid Dynamical Core (Havana)

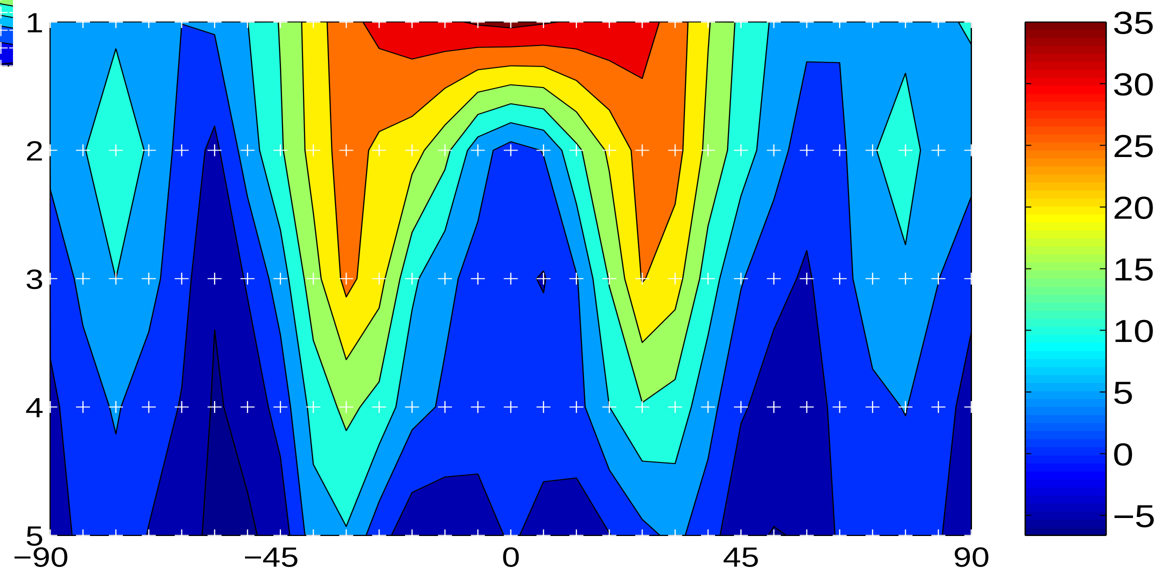
Held-Suarez Configuration (no zonal variation, fixed forcing)

Low-Resolution (60 lons, 30 lats, 5 levels); Timestep 1 hour

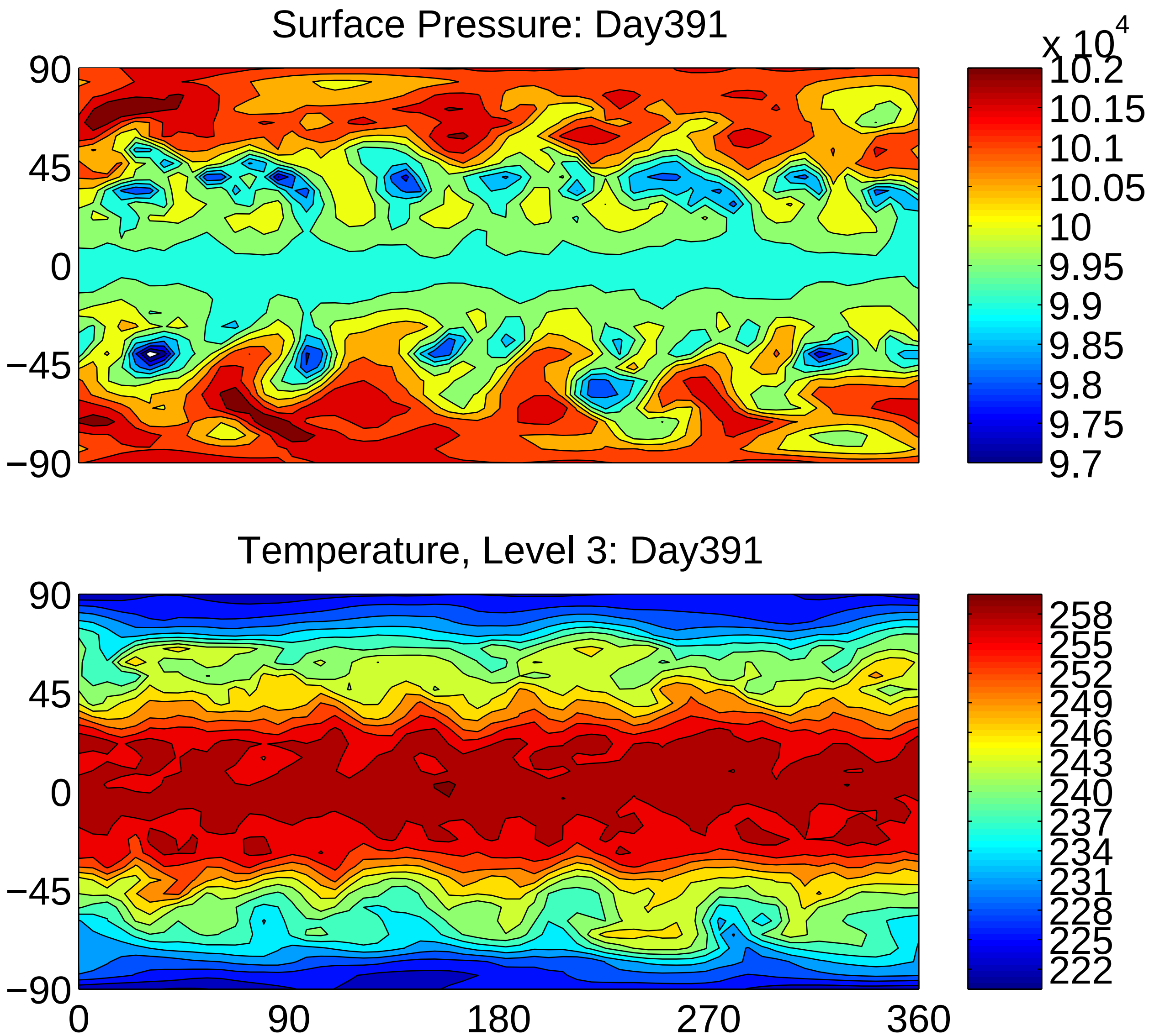


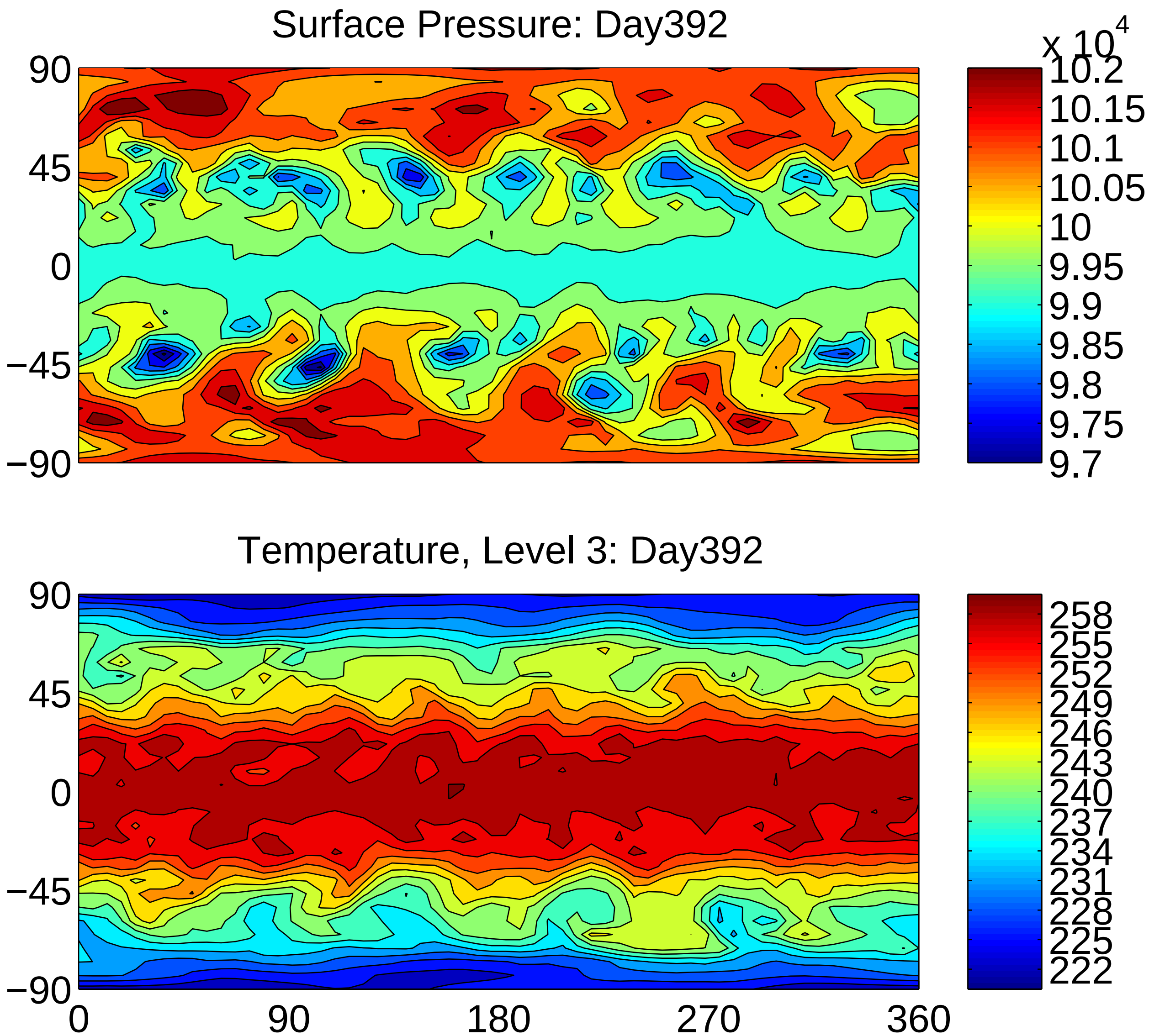
T at Level 3
(Middle of Atmosphere)

Cross Section of
Zonal Mean U

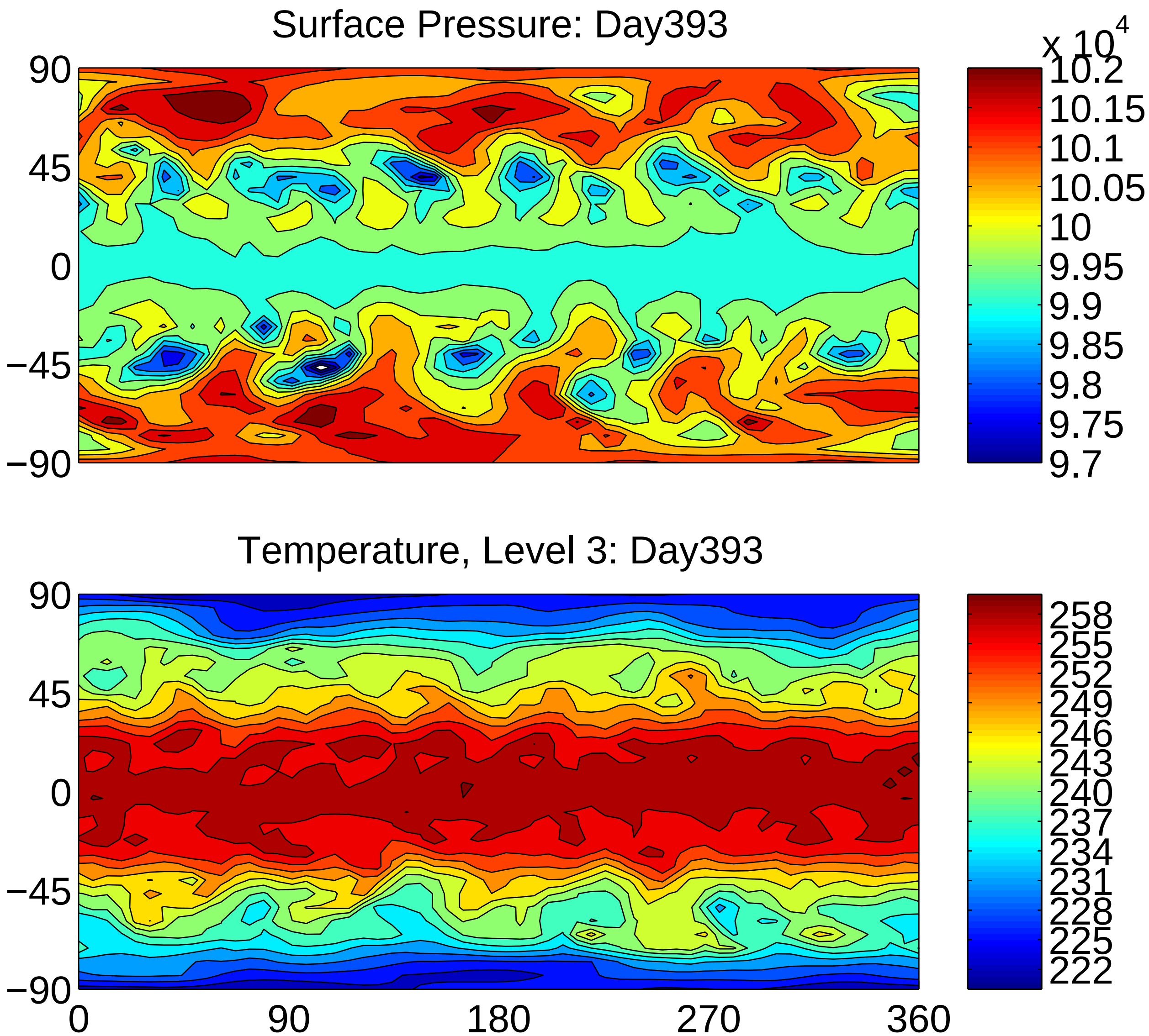


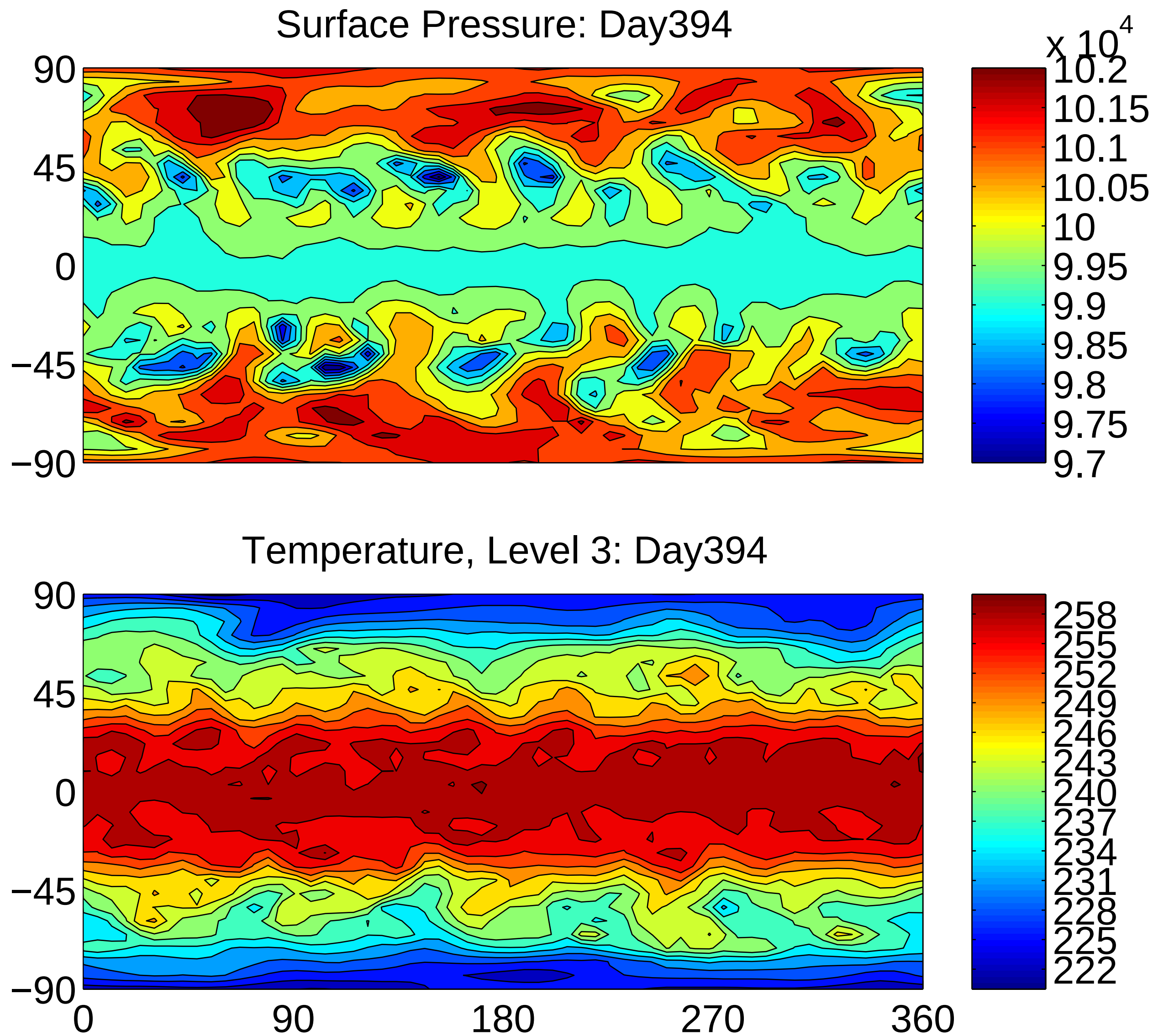
Has Baroclinic Instability

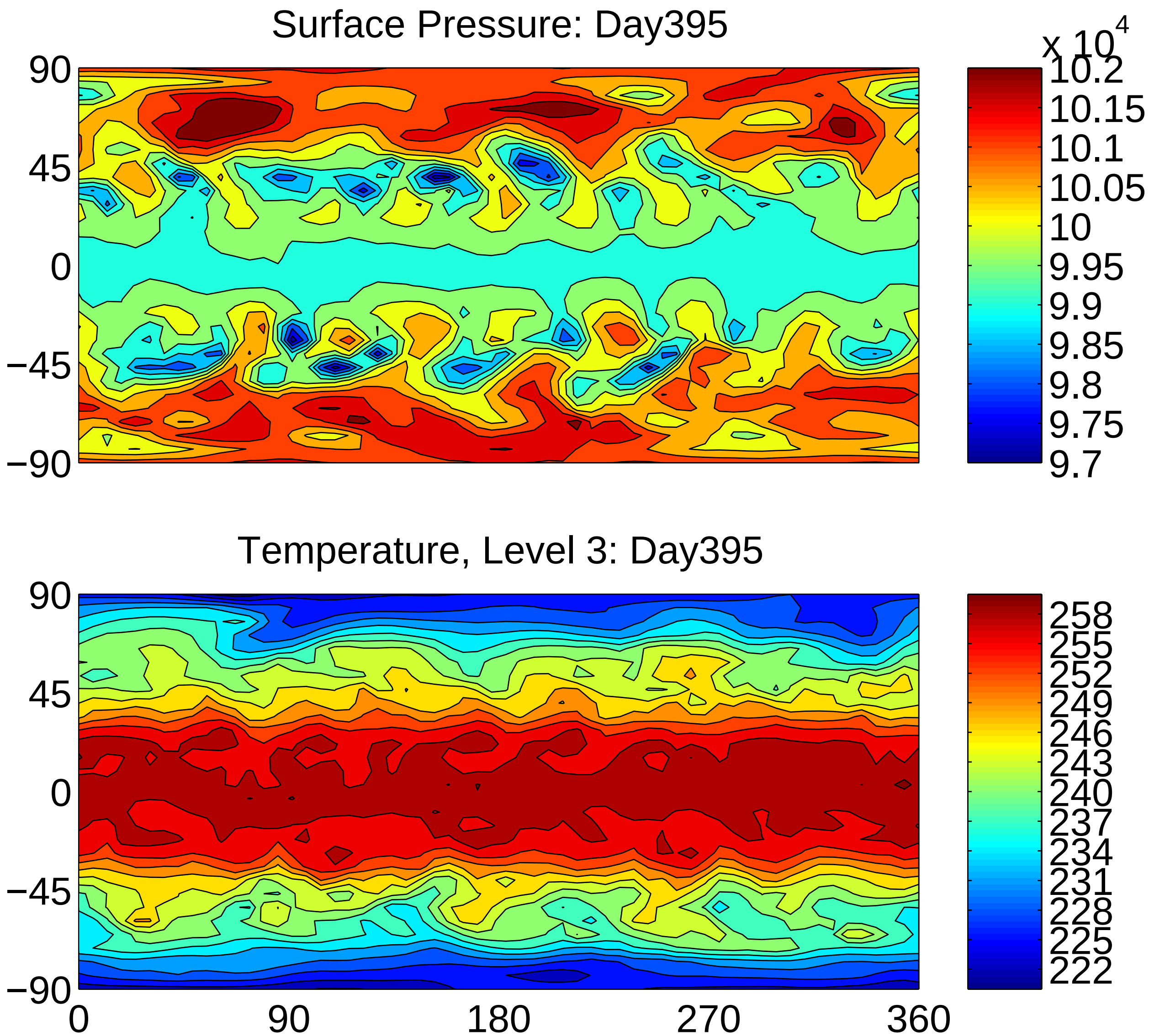

Surface Pressure: Day391
90
45
0
−45
−90
x 10^4
10.2
10.15
10.1
10.05
10
9.95
9.9
9.85
9.8
9.75
9.7
Temperature, Level 3: Day391
90
45
0
−45
−90
0
90
180
270
360
258
255
252
249
246
243
240
237
234
231
228
225
222

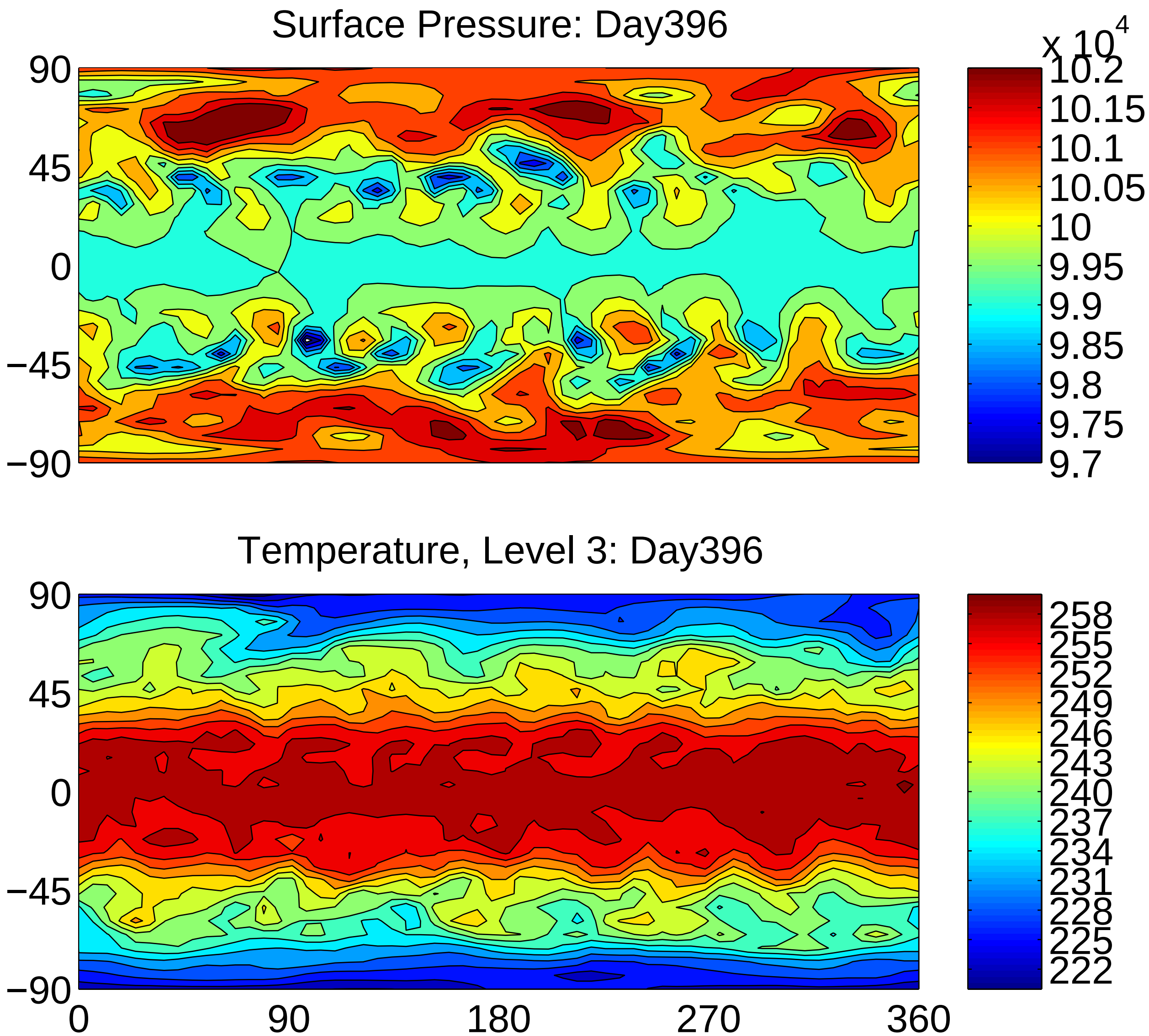


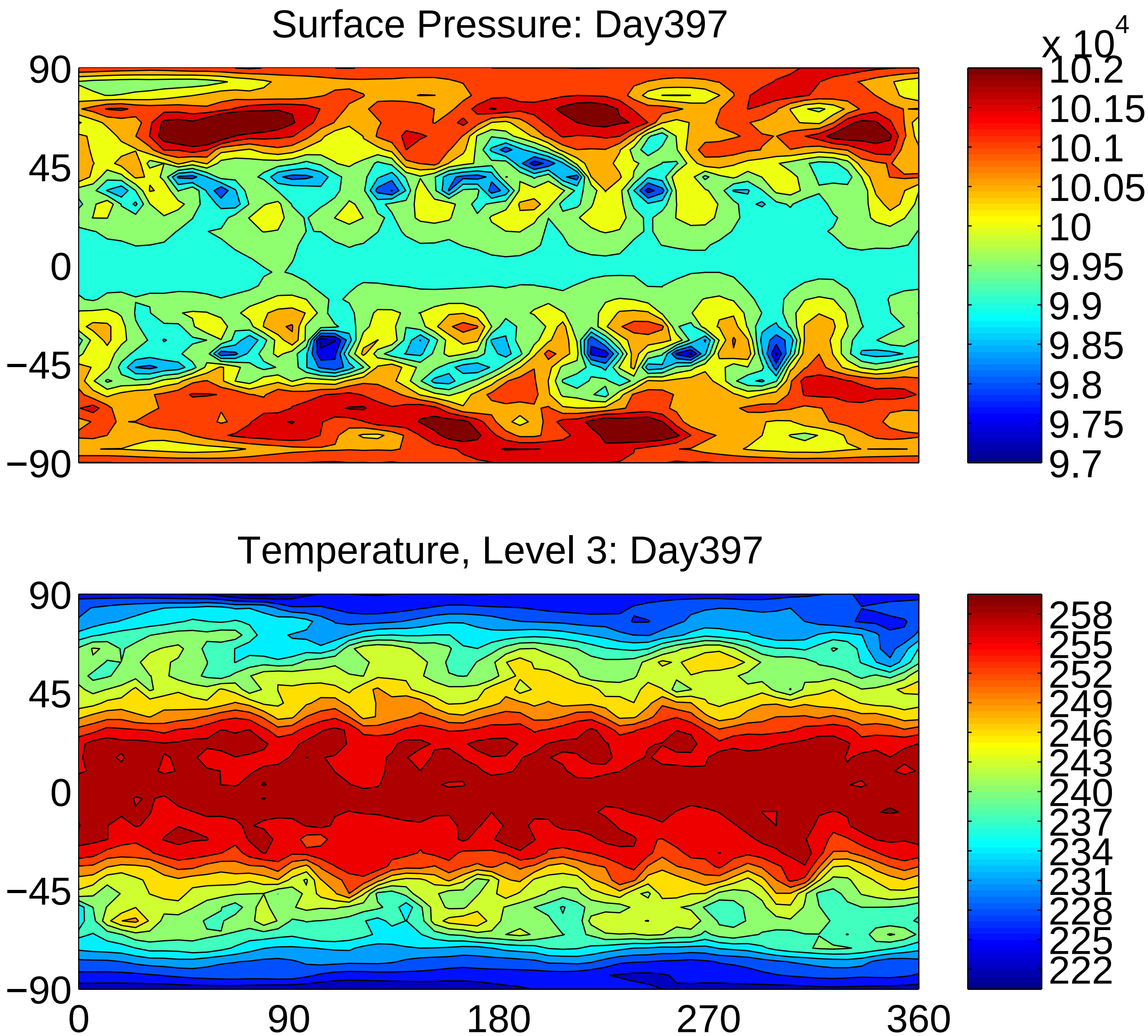
Surface Pressure: Day392
90
45
0
−45
−90
x 10^4
10.2
10.15
10.1
10.05
10
9.95
9.9
9.85
9.8
9.75
9.7
Temperature, Level 3: Day392
90
45
0
−45
−90
0
90
180
270
360
258
255
252
249
246
243
240
237
234
231
228
225
222

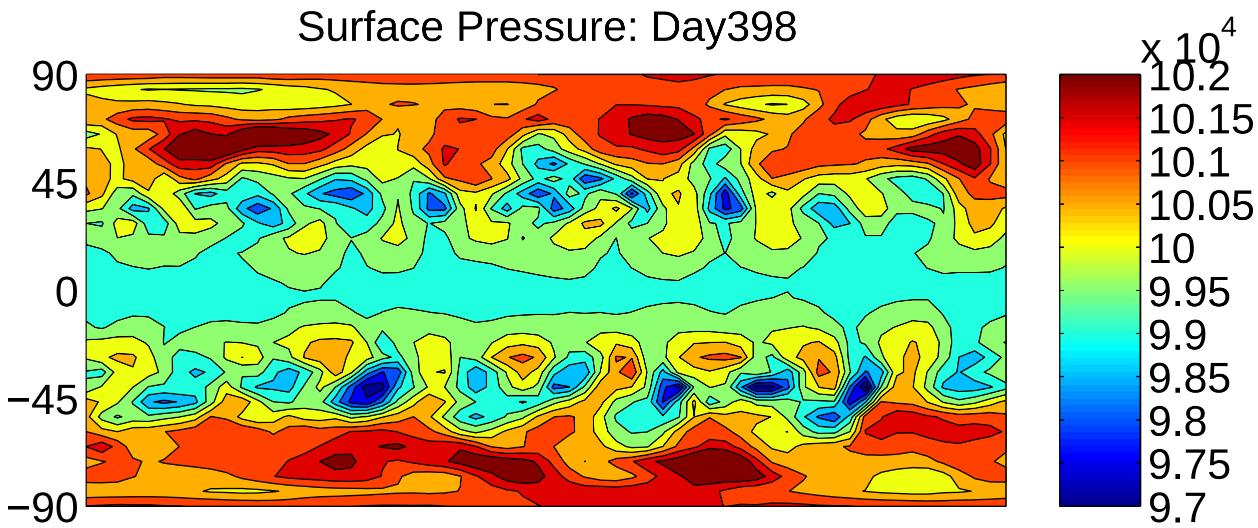

Surface Pressure: Day393
90
45
0
−45
−90
x 10^4
10.2
10.15
10.1
10.05
10
9.95
9.9
9.85
9.8
9.75
9.7
Temperature, Level 3: Day393
90
45
0
−45
−90
0
90
180
270
360
258
255
252
249
246
243
240
237
234
231
228
225
222


Surface Pressure: Day394
90
45
0
−45
−90
x 10^4
10.2
10.15
10.1
10.05
10
9.95
9.9
9.85
9.8
9.75
9.7
Temperature, Level 3: Day394
90
45
0
−45
−90
0
90
180
270
360
258
255
252
249
246
243
240
237
234
231
228
225
222

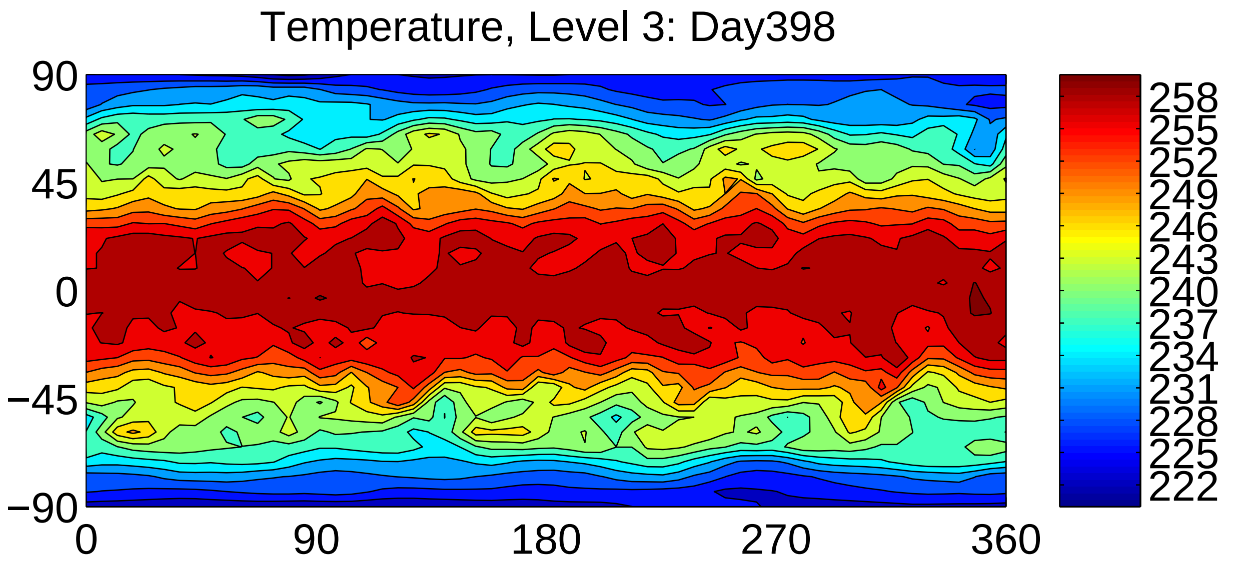

Surface Pressure: Day395
x 10^4
10.2
10.15
10.1
10.05
10
9.95
9.9
9.85
9.8
9.75
9.7
90
45
0
−45
−90
Temperature, Level 3: Day395
258
255
252
249
246
243
240
237
234
231
228
225
222
90
45
0
−45
−90
0
90
180
270
360


Surface Pressure: Day396
90
45
0
−45
−90
x 10^4
10.2
10.15
10.1
10.05
10
9.95
9.9
9.85
9.8
9.75
9.7
Temperature, Level 3: Day396
90
45
0
−45
−90
258
255
252
249
246
243
240
237
234
231
228
225
222
0
90
180
270
360

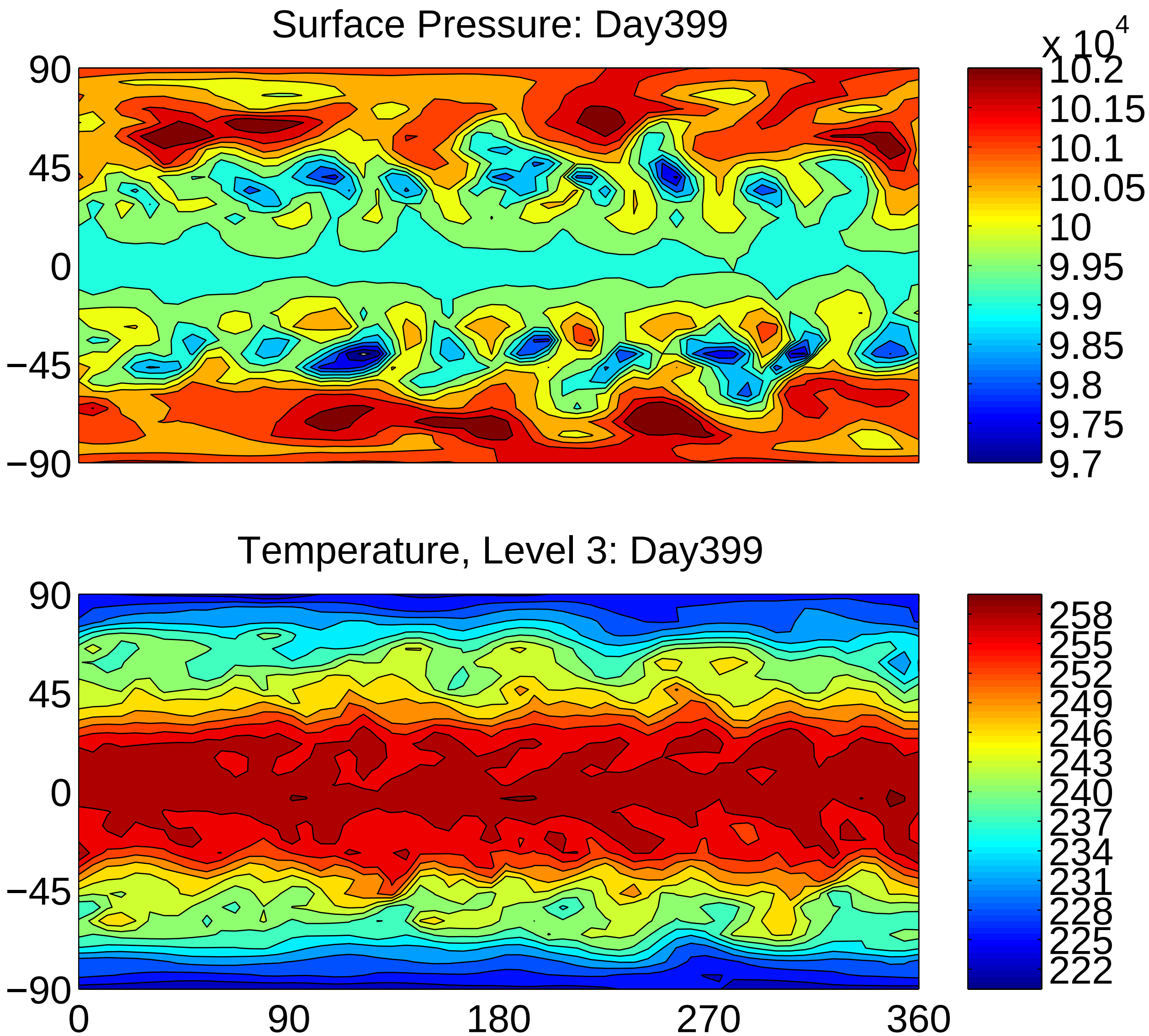

Surface Pressure: Day397
x 10^4
10.2
10.15
10.1
10.05
10
9.95
9.9
9.85
9.8
9.75
9.7
90
45
0
−45
−90
Temperature, Level 3: Day397
258
255
252
249
246
243
240
237
234
231
228
225
222
90
45
0
−45
−90
0
90
180
270
360

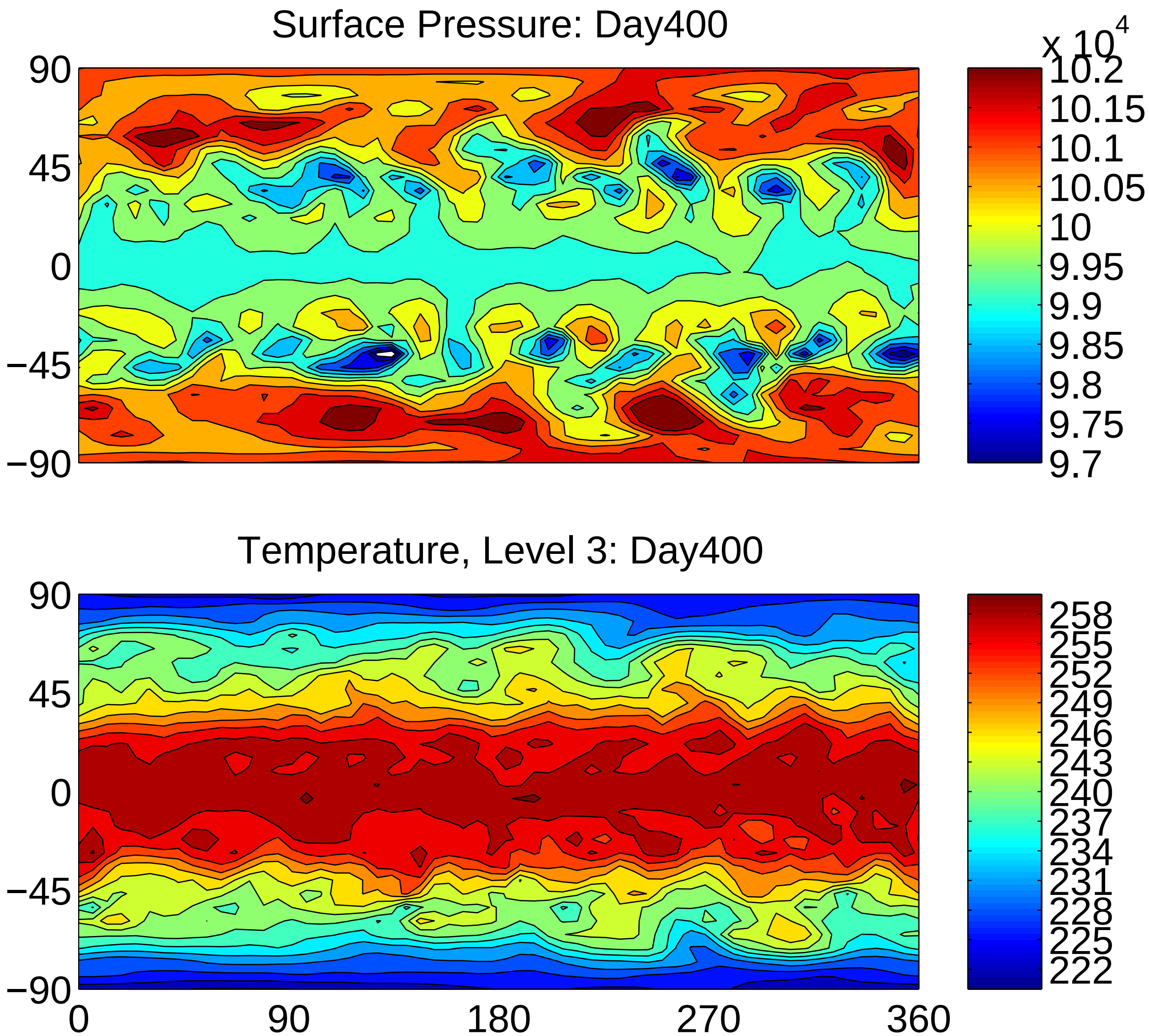


Surface Pressure: Day398
90
45
0
−45
−90
x 10^4
10.2
10.15
10.1
10.05
10
9.95
9.9
9.85
9.8
9.75
9.7



Temperature, Level 3: Day398
90
45
0
−45
−90
0
90
180
270
360
258
255
252
249
246
243
240
237
234
231
228
225
222


Surface Pressure: Day399
90
45
0
−45
−90
x 10^4
10.2
10.15
10.1
10.05
10
9.95
9.9
9.85
9.8
9.75
9.7
Temperature, Level 3: Day399
90
45
0
−45
−90
0
90
180
270
360
258
255
252
249
246
243
240
237
234
231
228
225
222



Surface Pressure: Day400
90
45
0
−45
−90
x 10^4
10.2
10.15
10.1
10.05
10
9.95
9.9
9.85
9.8
9.75
9.7
Temperature, Level 3: Day400
90
45
0
−45
−90
0
90
180
270
360
258
255
252
249
246
243
240
237
234
231
228
225
222

Experimental Design Details: Bgrid AGCM

Results for 4x20 group filter

Assimilation for 400 days; starting from climatological distribution

Summary results are from last 200 days

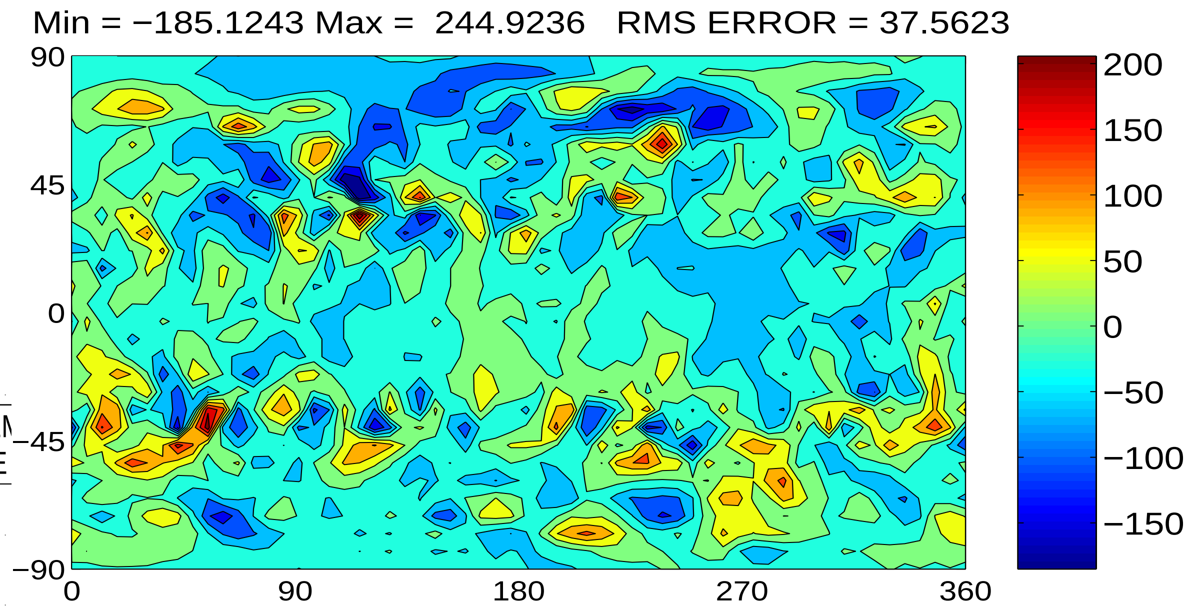
No covariance inflation

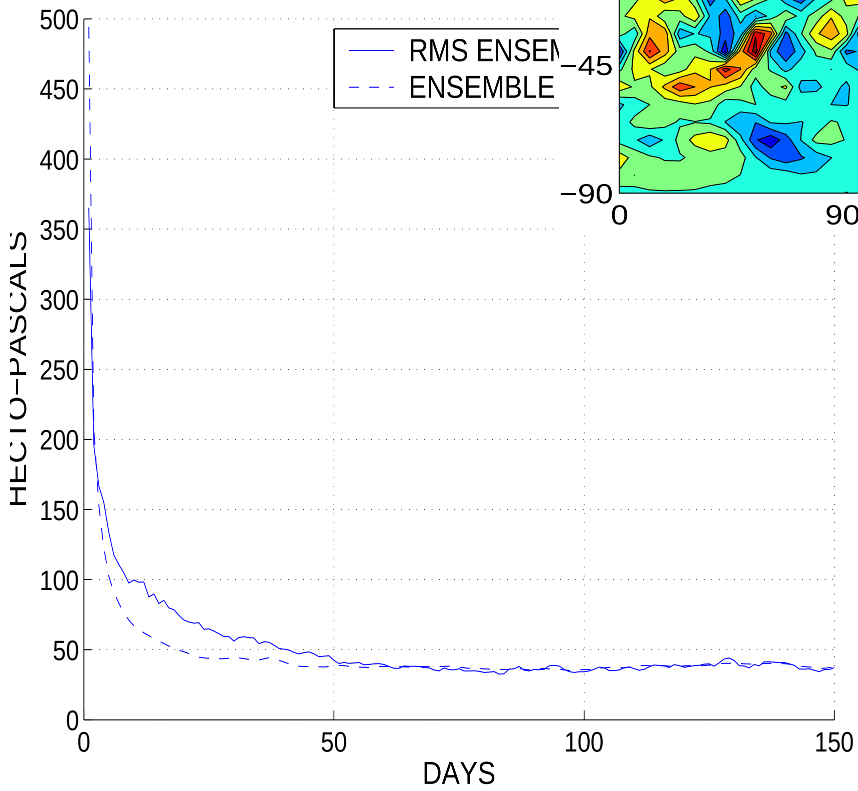
1800 randomly located surface pressure stations observe once every 24 hours

Observational error variance is 1 mb

Baseline Case: 1800 PS Obs every 24 hours

Largest error in mid-latitudes, 'synoptic' scales after 400 days





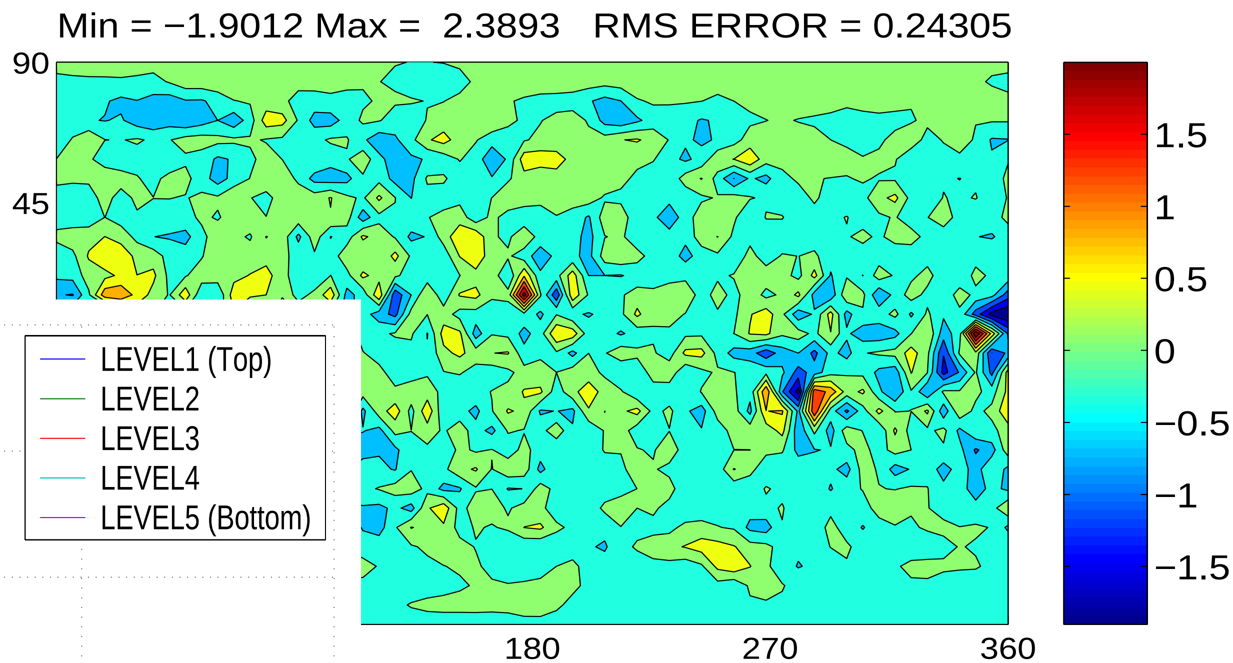
PS Error Reduces by about10x

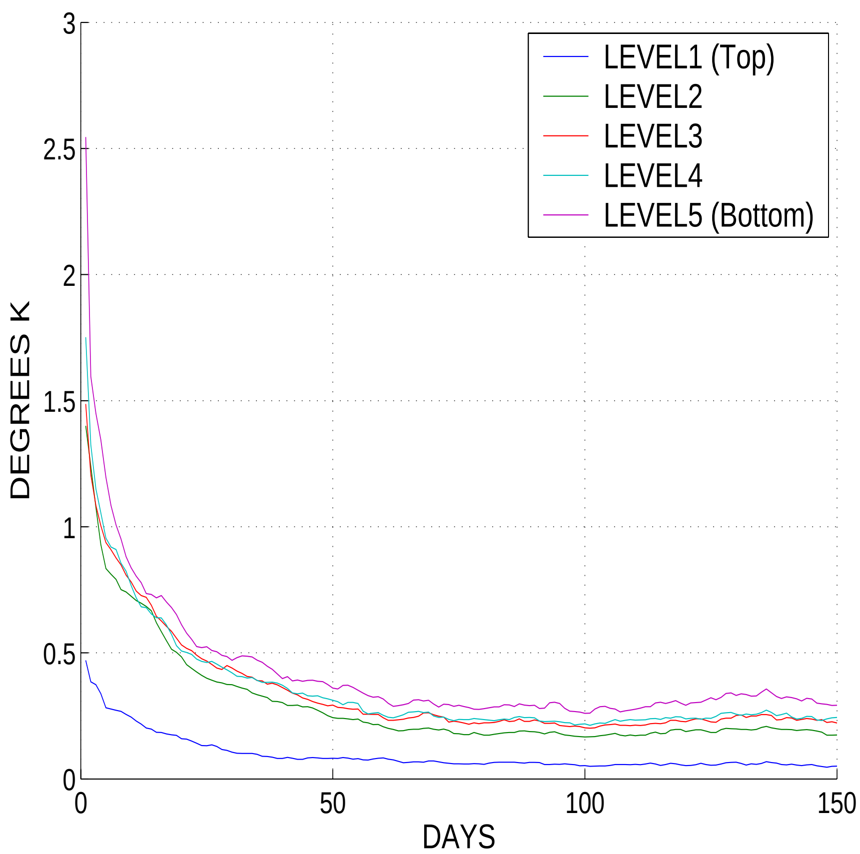
Asymptotes after about 50 days

Ensemble spread is approximately correlated with RMS error

Baseline Case: 1800 PS Obs every 24 hours

Largest T error in tropics for interior levels (level 3, day 400 shown)



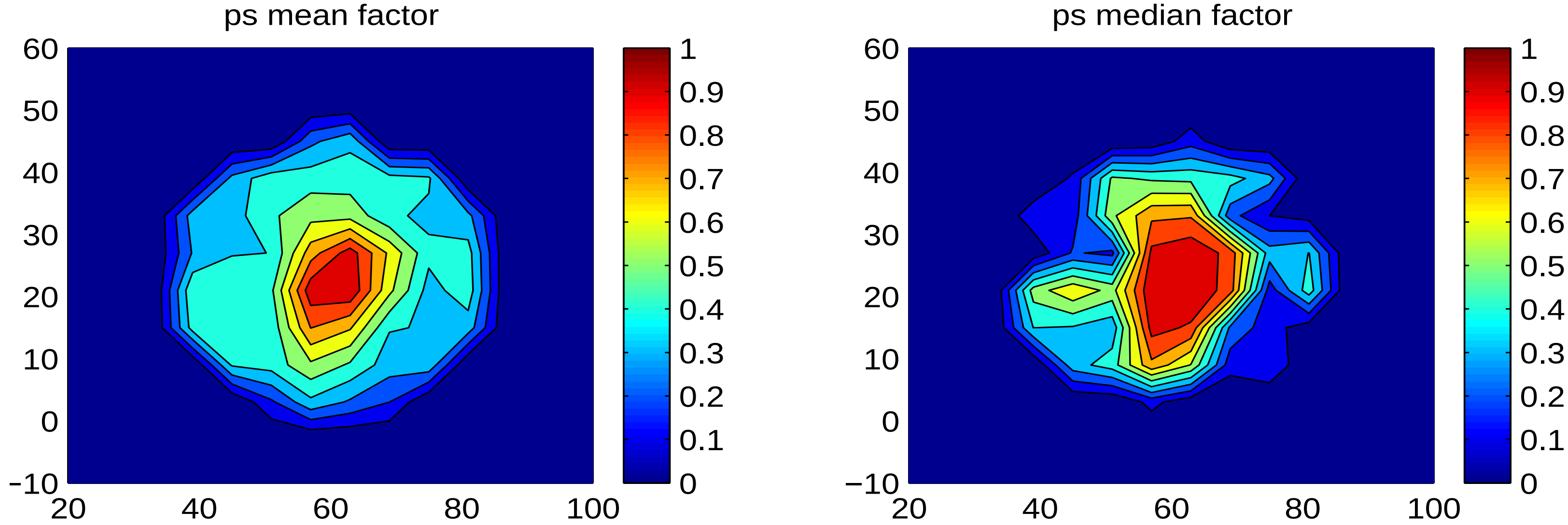


T Error also reduced by about 10x

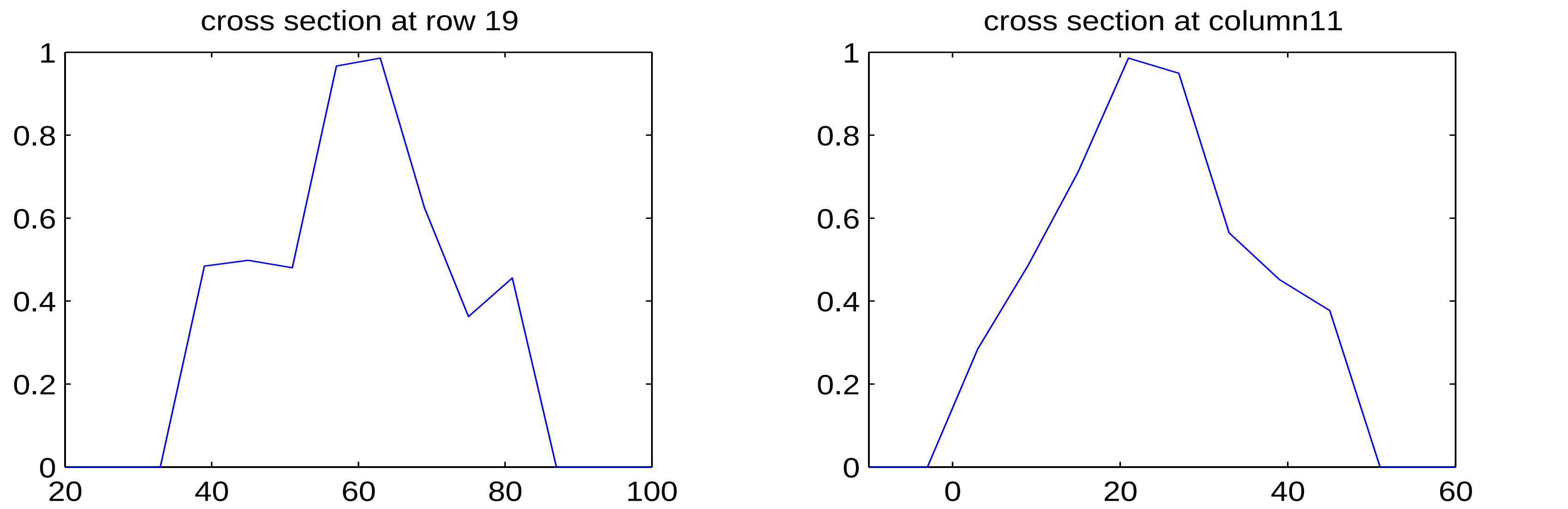
Asymptotes about day 70

Final error about 0.25 K for interior levels

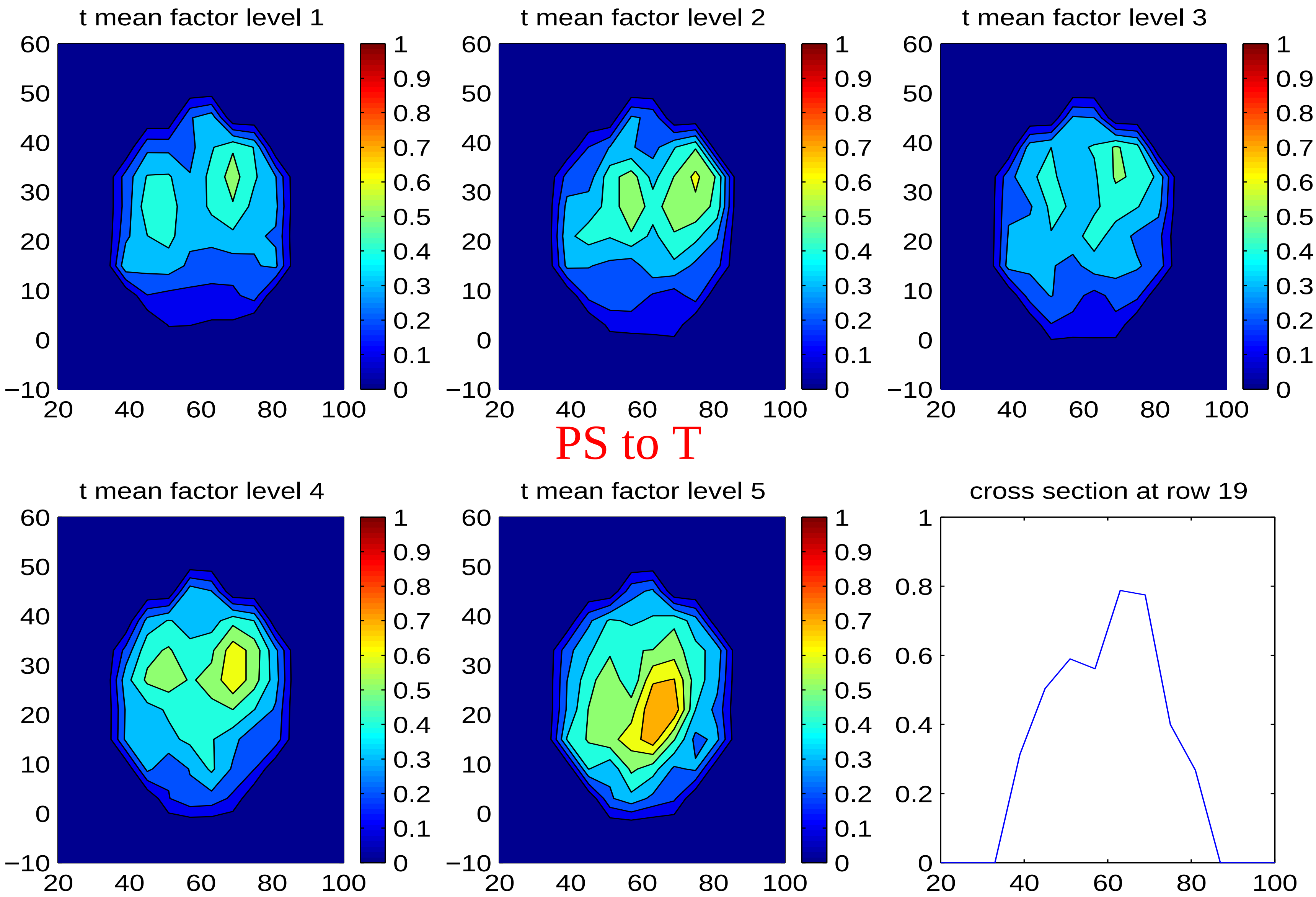
Hierarchical Filter Regression Confidence Factors: PS Obs. at 20N, 60E



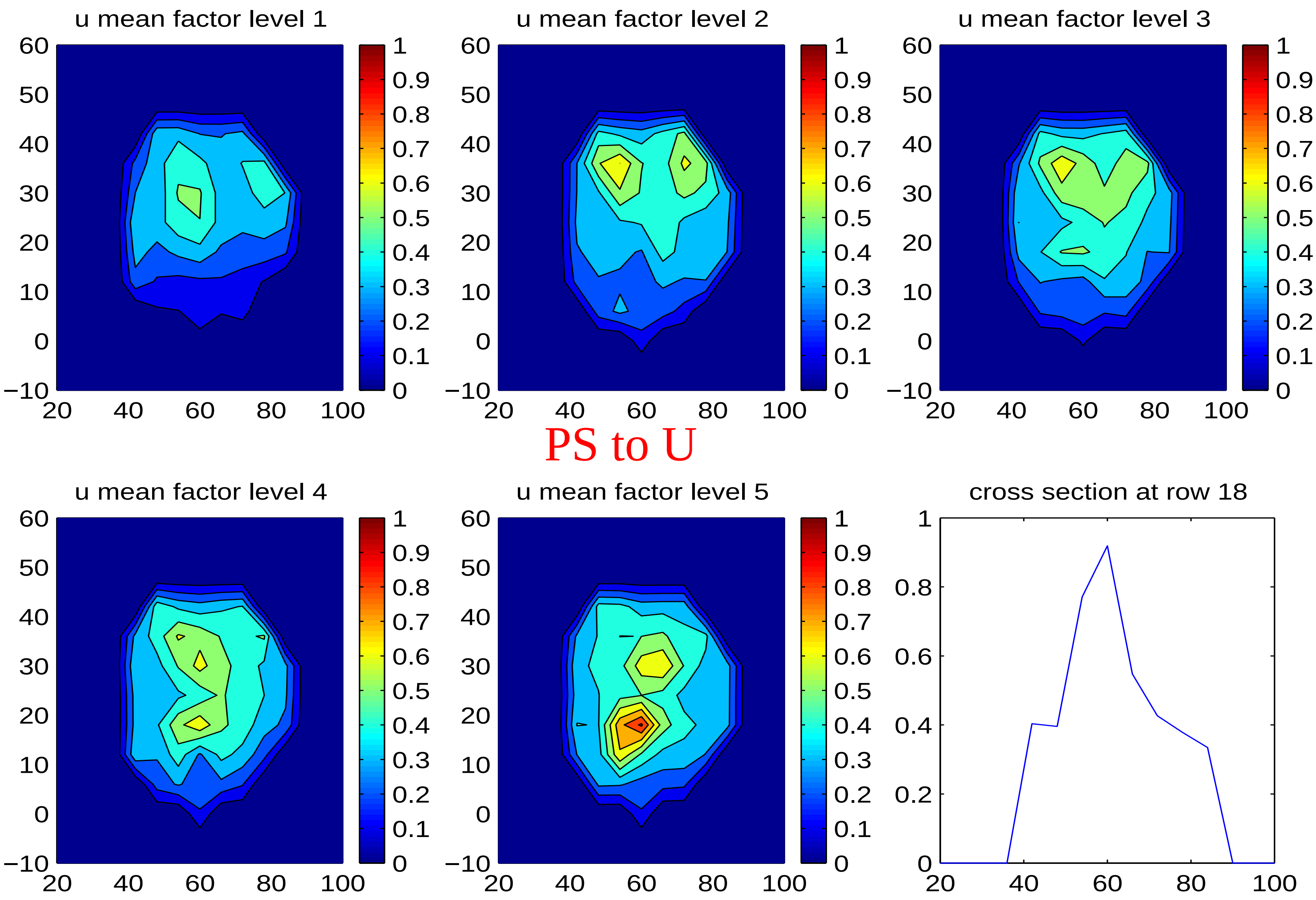
PS to PS



Hierarchical Filter Regression Confidence Factors: PS Obs. at 20N, 60E



Hierarchical Filter Regression Confidence Factors: PS Obs. at 20N, 60E



Hierarchical Filter Regression Confidence Factors: PS Obs. at 20N, 60E