Automated Estimation of Localization for Ensemble Kalman Filter Data Assimilation

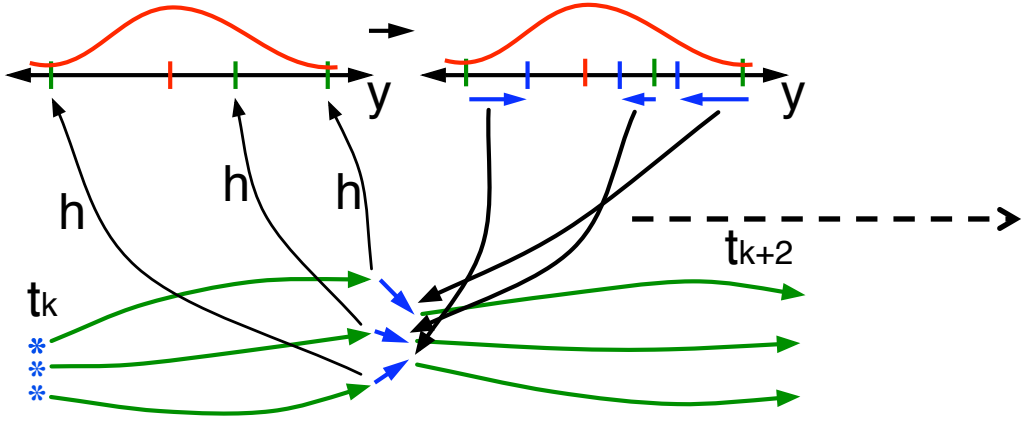
Lili Lei

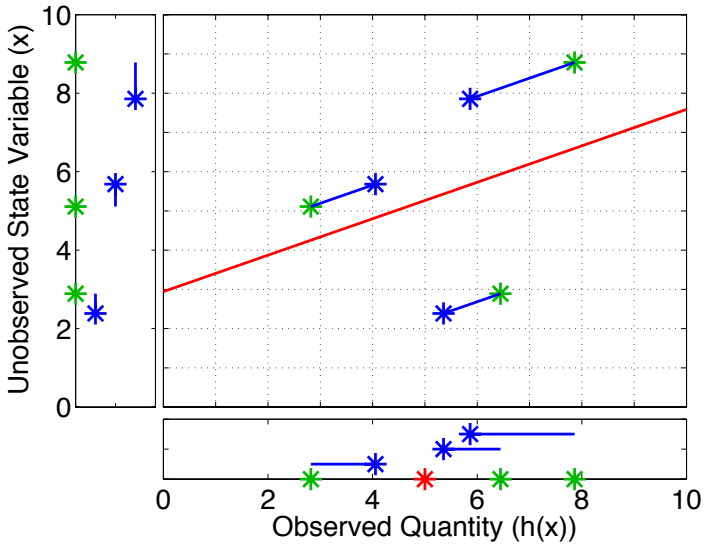
CIRES/CU Boulder and NOAA/ESRL/PSD

With acknowledgements to Jeff Anderson and the DART group

IMAGe Theme of the Year - Frontiers in Ensemble Data Assimilation for Geoscience Applications, Boulder CO, 3 - 7 August 2015

Sampling error and localization in an EnKF

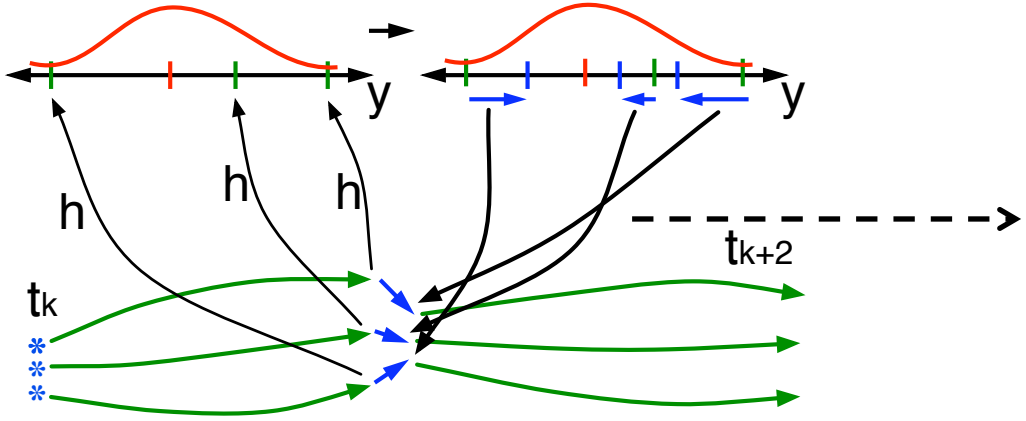


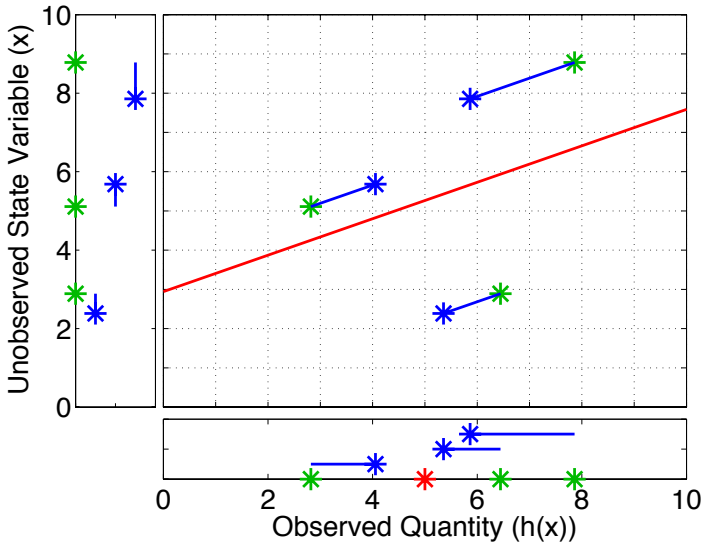


$$\Delta x_n = \hat{b}\Delta y_n$$

During the assimilation process, EnKF uses statistical properties of an ensemble model forecast to estimate the flow-dependent background error covariance to determine how an observation modifies the forecast background fields.

Sampling error and localization in an EnKF

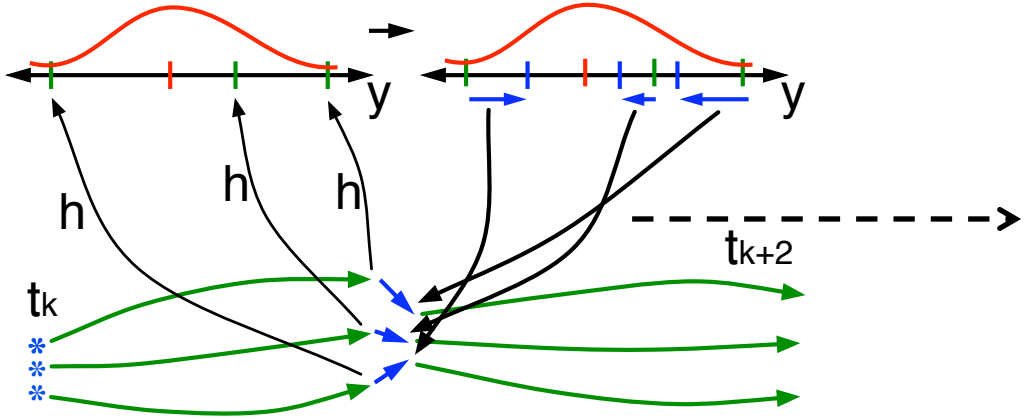


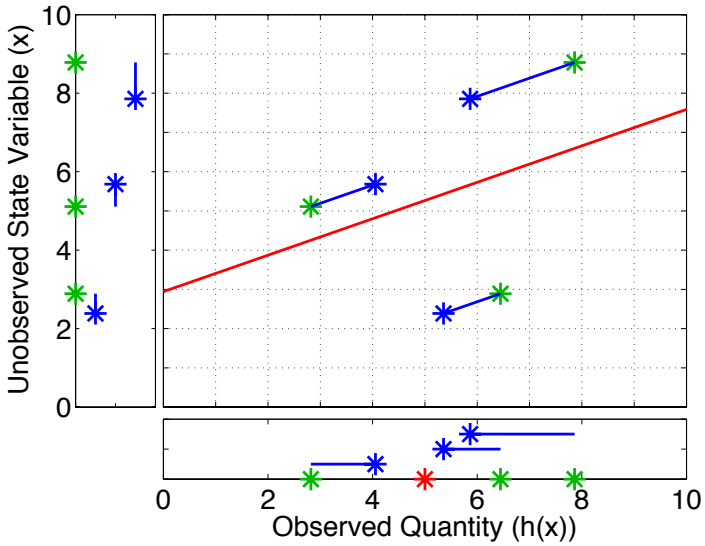


$$\Delta x_n = \hat{b}\Delta y_n$$

With limited ensemble size, there are spurious error correlations between an observation and a state variable, $\hat{b}$, especially when the distance between these two is large.

Sampling error and localization in an EnKF





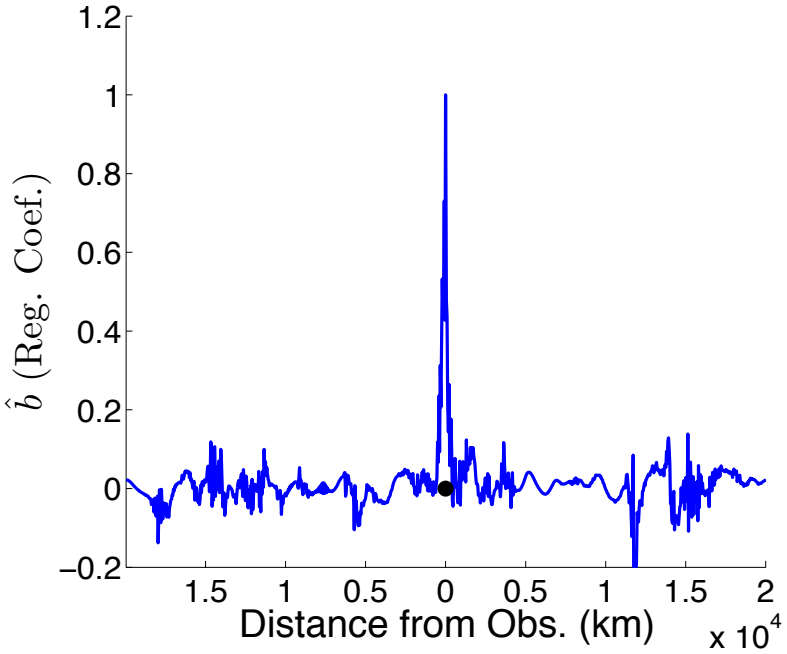
$$\Delta x_n = \hat{b}\Delta y_n$$

Remedy: Covariance localization (α) $\quad \Delta x_n = \alpha\hat{b}\Delta y_n$

Localization, a technique to 'localize' the impact of an observation to state variables that are close to the observation, is used to reduce spurious error correlations between the observation and distant state variables.

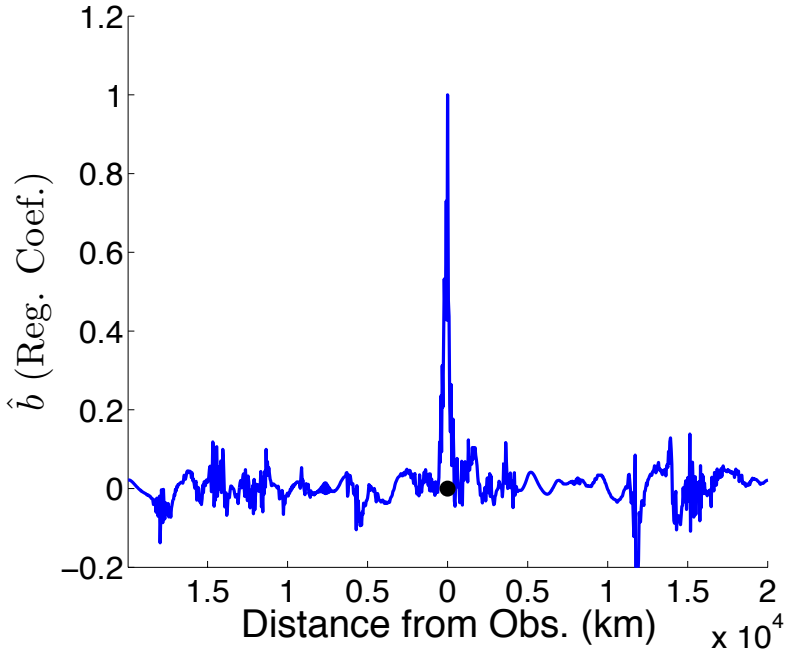
Sampling error and localization in an EnKF

Regression coefficient $\hat{b}$



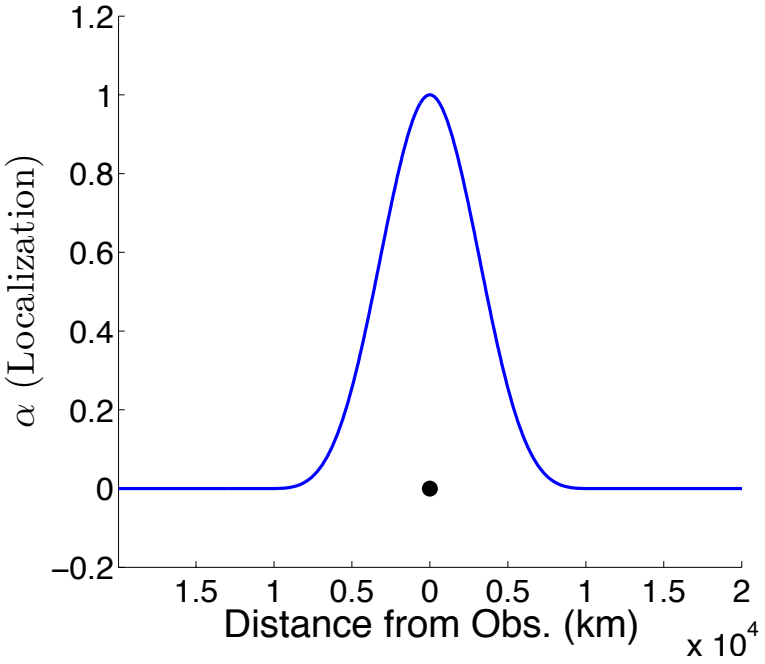
Sampling error and localization in an EnKF

Regression coefficient $\hat{b}$



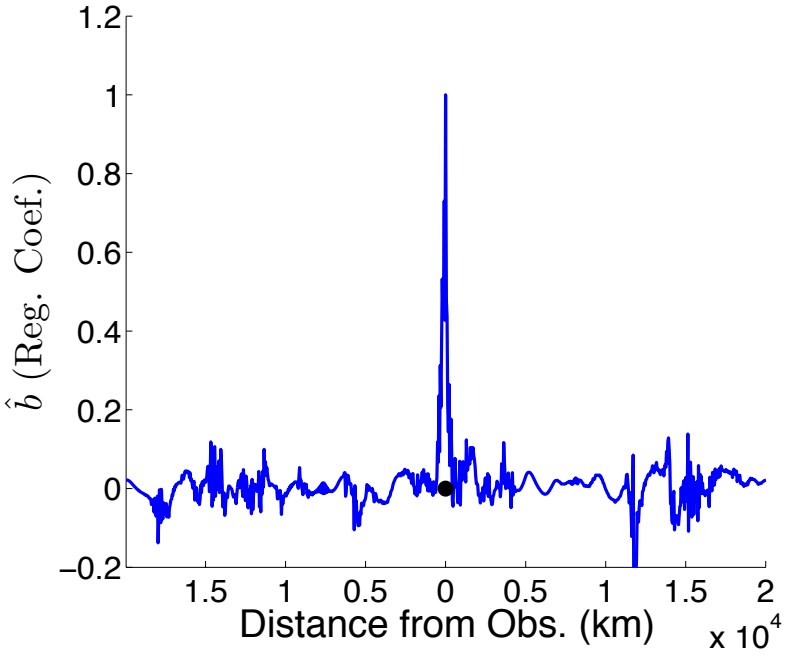
The most commonly used localization α :

GC localization (Gaspari and Cohn 1999)

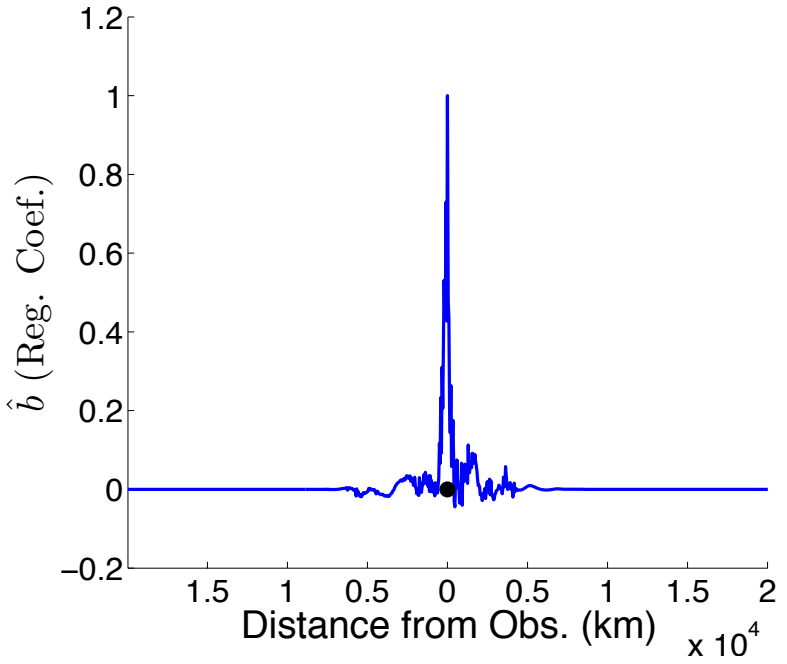


Sampling error and localization in an EnKF

Regression coefficient $\hat{b}$

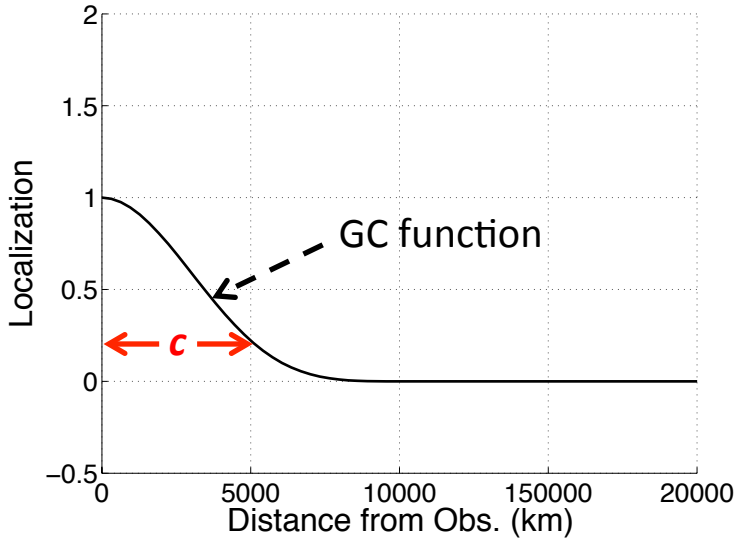


Localized regression coefficient $\hat{b}$



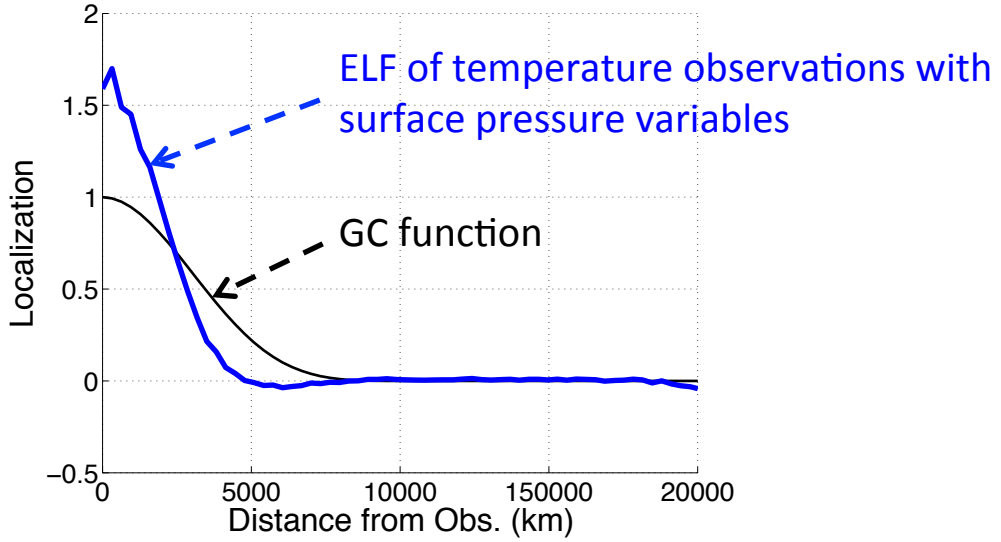
Motivation for an automated localization algorithm

- *The GC function has a single parameter that defines the width of the function.*
  But tuning even this single parameter can be computationally expensive.
- *The GC function is approximately Gaussian.*
  But different localization functions are needed for:
  - different observation types (Houtekamer and Mitchell 2005, Anderson and Lei 2013)
  - different state variable kinds (Anderson 2007, 2012)
  - different times (Anderson 2007, Chen and Oliver 2010)
  - different regions (Lei and Anderson 2014).



Motivation for an automated localization algorithm

- *The GC function has a single parameter that defines the width of the function.*
- *The GC function is approximately Gaussian.*
- Thus an automated localization algorithm, empirical localization function (ELF), is proposed.
  - ELF provides an estimate for the localization for any possible observation type with a state variable kind (at different times and for different regions).
  - ELF makes few a priori assumptions for the shape of the localization function.
  - ELF has computational cost advantage over tuning the GC halfwidth.
  - ELF can outperform the best GC function.

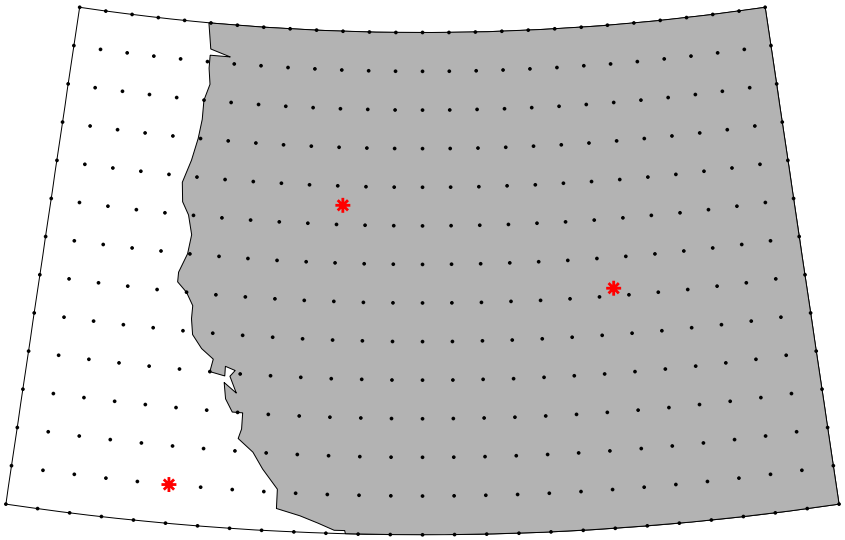


Empirical localization algorithm

1. Compute separation between each pair of an observation and a state variable;
2. Divide the set of all pairs into subsets using the separation;
3. Compute the localization for each subset.

Empirical localization algorithm:

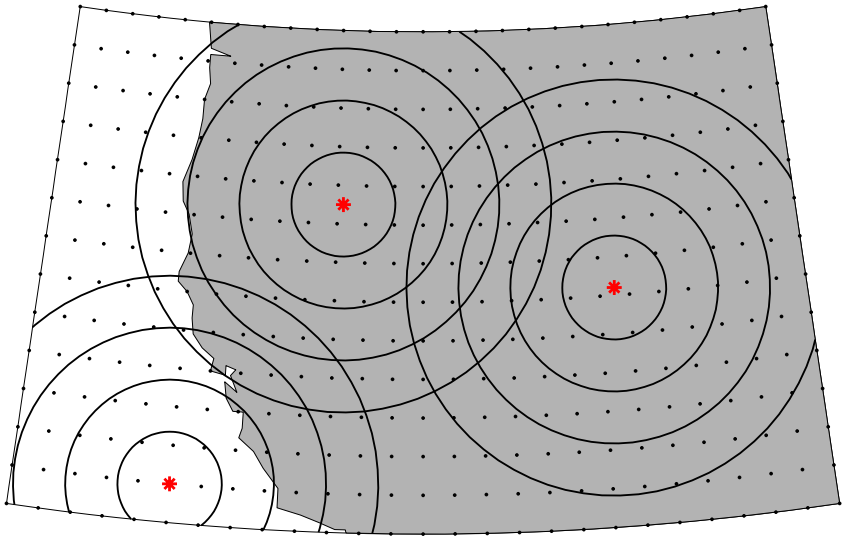
1. Compute separation between each pair of an observation and a state variable



- Black dots: grid points.
- Red stars: observations.

Empirical localization algorithm:

2. Divide the set of all pairs into subsets using the separation

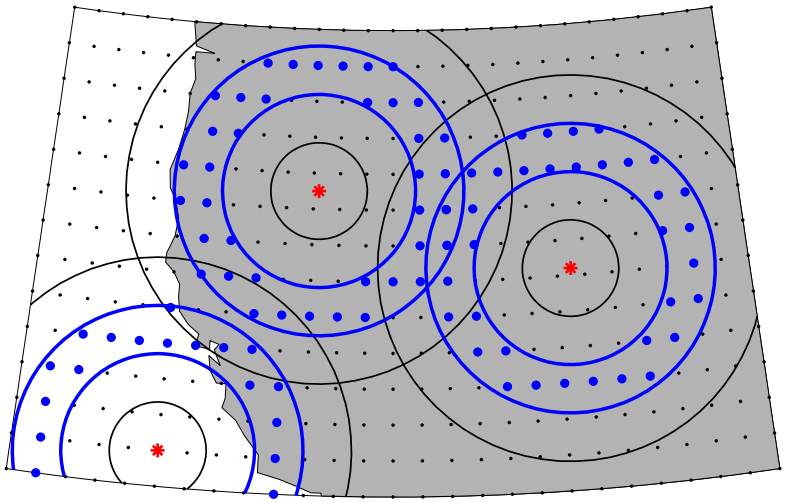


- Black dots: grid points.
- Red stars: observations.

  Circles: distance ranges from each observation.

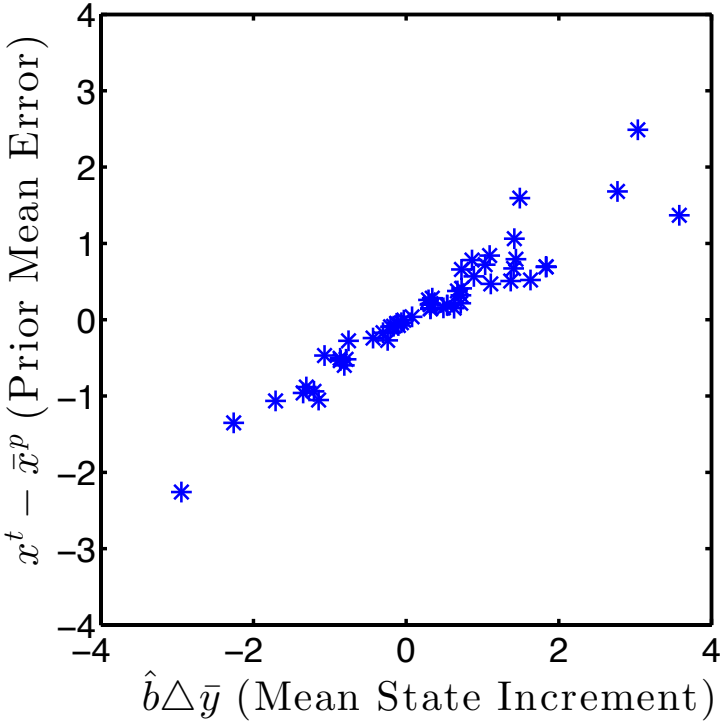
Empirical localization algorithm:

3. Compute the localization for each subset



Blue circles: the distance range chosen for one subset.

Blue dots: state variables in the chosen distance range for every observation.



The abscissa is the mean state increment, and the ordinate is the prior mean error.

These two quantities are plotted for each pair of an observation and a state variable.

Empirical localization algorithm:

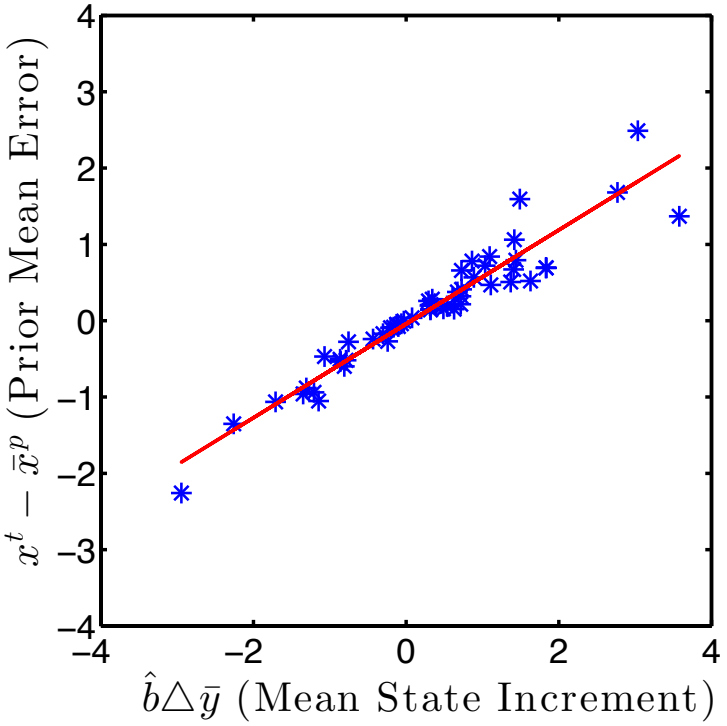
3. Compute the localization for each subset

The slope of the least squares fit is the localization α that minimizes the RMS difference between the increments and prior errors:

$$\alpha = \frac{\sum_{k=1}^{K}\left(x_k^t - \bar{x}_k^p\right)\hat{b}_k \Delta\bar{y}_k}{\sum_{k=1}^{K}\left(\hat{b}_k \Delta\bar{y}_k\right)^2}$$

The slope of the least squares fit is also the localization α that minimizes the RMS difference between the posterior ensemble means and true values of the state variable in the subset:

$$J = \sqrt{\sum_{k=1}^{K}(\bar{x}_k^u - x_k^t)^2} = \sqrt{\sum_{k=1}^{K}\left(\bar{x}_k^p + \alpha\hat{b}_k\Delta\bar{y}_k - x_k^t\right)^2}$$



The abscissa is the mean state increment, and the ordinate is the prior mean error.

These two quantities are plotted for each pair of an observation and a state variable.

Tests of the empirical localization algorithm

- The dynamical core of the GFDL B-grid global atmospheric model: Localization for different observation types and state variable kinds
- The Community Atmospheric Model version 5 (CAM5): Vertical localization and localization for different geographic regions
- The Weather Research and Forecasting Model (WRF): Localization for regions with and without precipitation

Tests of the empirical localization algorithm

- The dynamical core of the GFDL B-grid global atmospheric model: Localization for different observation types and state variable kinds
- The Community Atmospheric Model version 5 (CAM5): Vertical localization and localization for different geographic regions
- The Weather Research and Forecasting Model (WRF): Localization for regions with and without precipitation

ELF in B-grid Global Model

Experimental Design

Conduct Observing System Simulation Experiments (OSSEs).

B-grid global model:

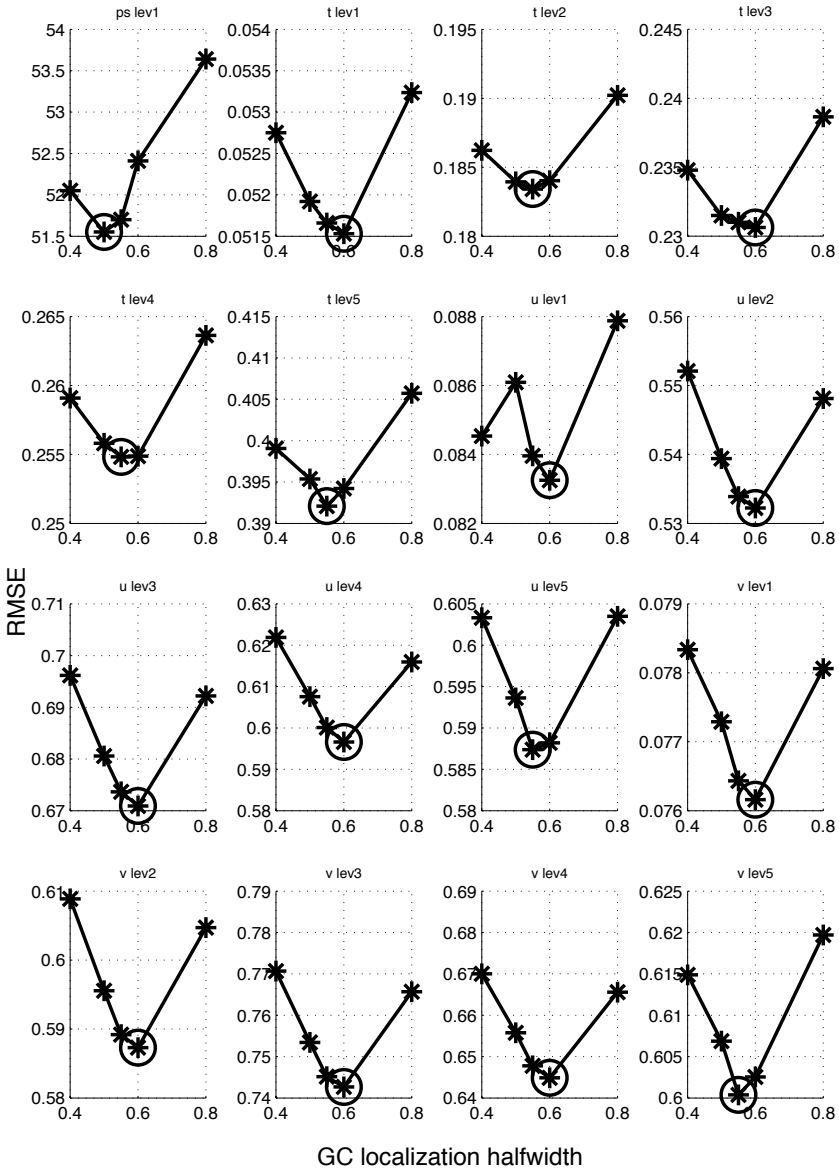
- State variables of surface pressure, temperature, and zonal and meridional wind components
- Horizontal model grid points 30 latitudes × 60 longitudes, 5 vertical levels, and model time step 1h

Data assimilation system:

- Ensemble Kalman filter with perturbed observation (Burgers et al. 1998, Houtekamer and Mitchell 1998) in the Data Assimilation Research Testbed (DART; Anderson et al. 2009)
- Time-varying but spatially uniform state space adaptive inflation (Anderson 2009)
- GC localization as the default

ELF in B-grid Global Model

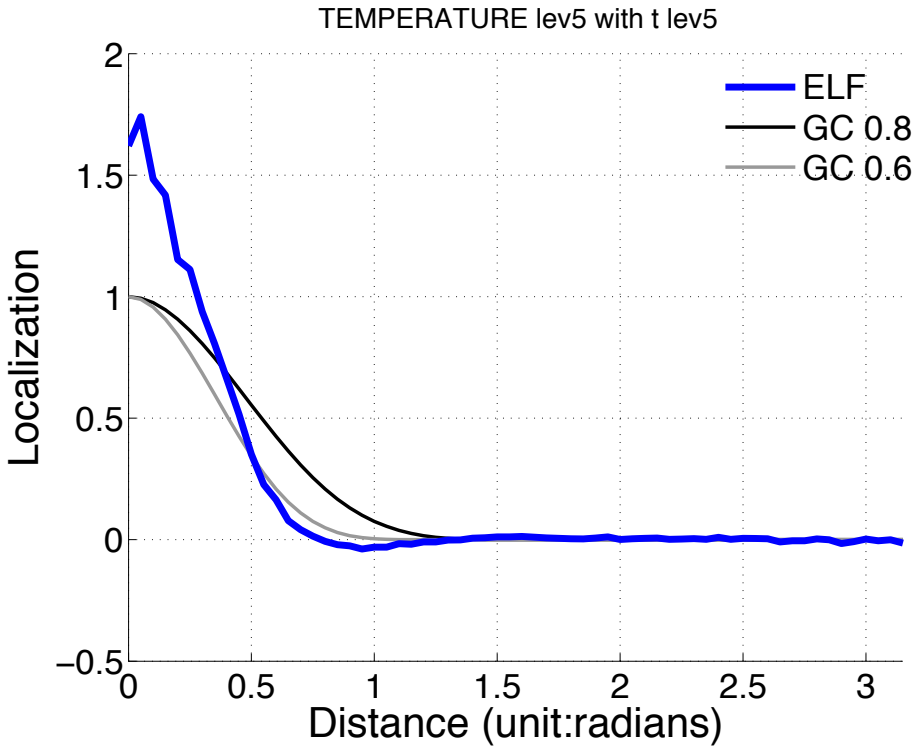
Baseline Test with GC Localization



On average the GC halfwidth of 0.6 radians gives the smallest RMSE, thus 0.6 radians is seen as the optimal GC halfwidth.

ELF in B-grid Global Model

Empirical Localization Function



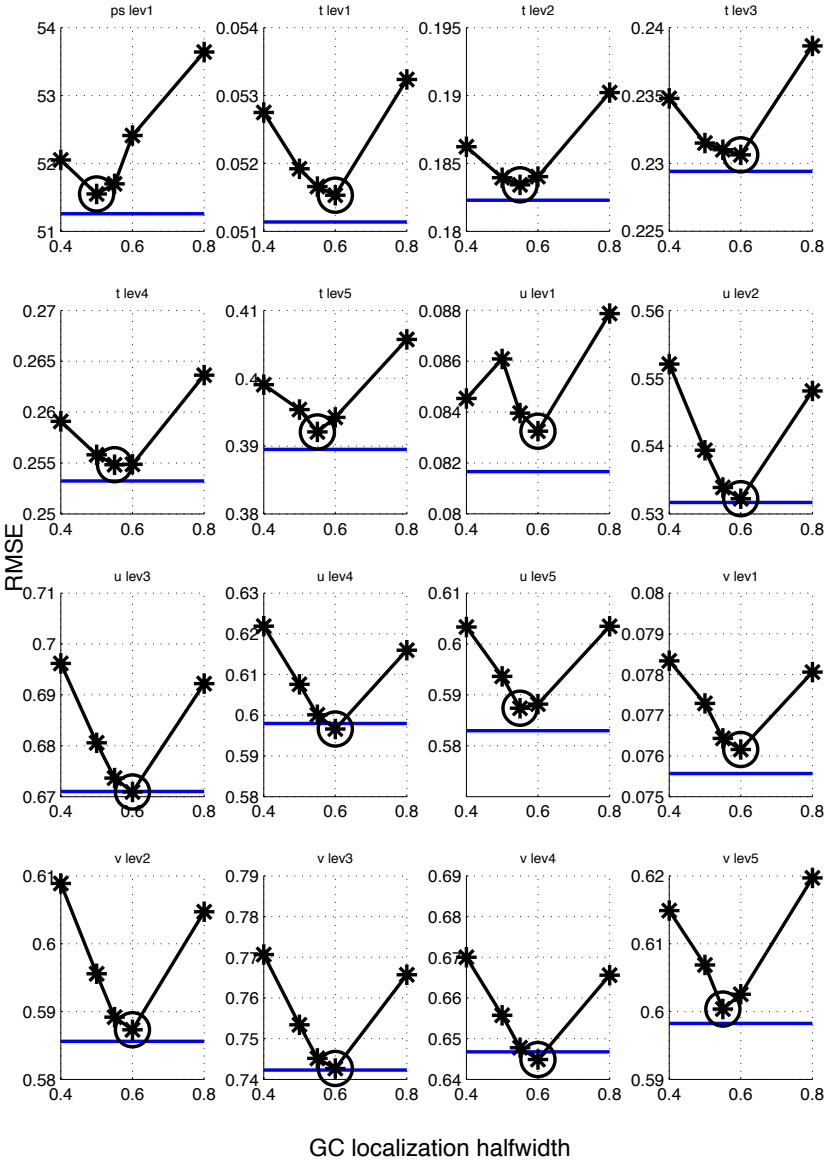
The ELFs are constructed from the output of an OSSE with GC halfwidth 0.8 (not optimal).

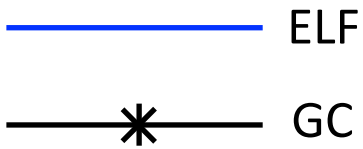
The ELF is narrower than GC0.8 and has better agreement with GC0.6 than GC0.8 at tails.

The ELF has values larger than 1.0 at small separations (< 0.3 radians), which indicates insufficient ensemble spread.

ELF in B-grid Global Model

RMSE of ELF and GC Halfwidths





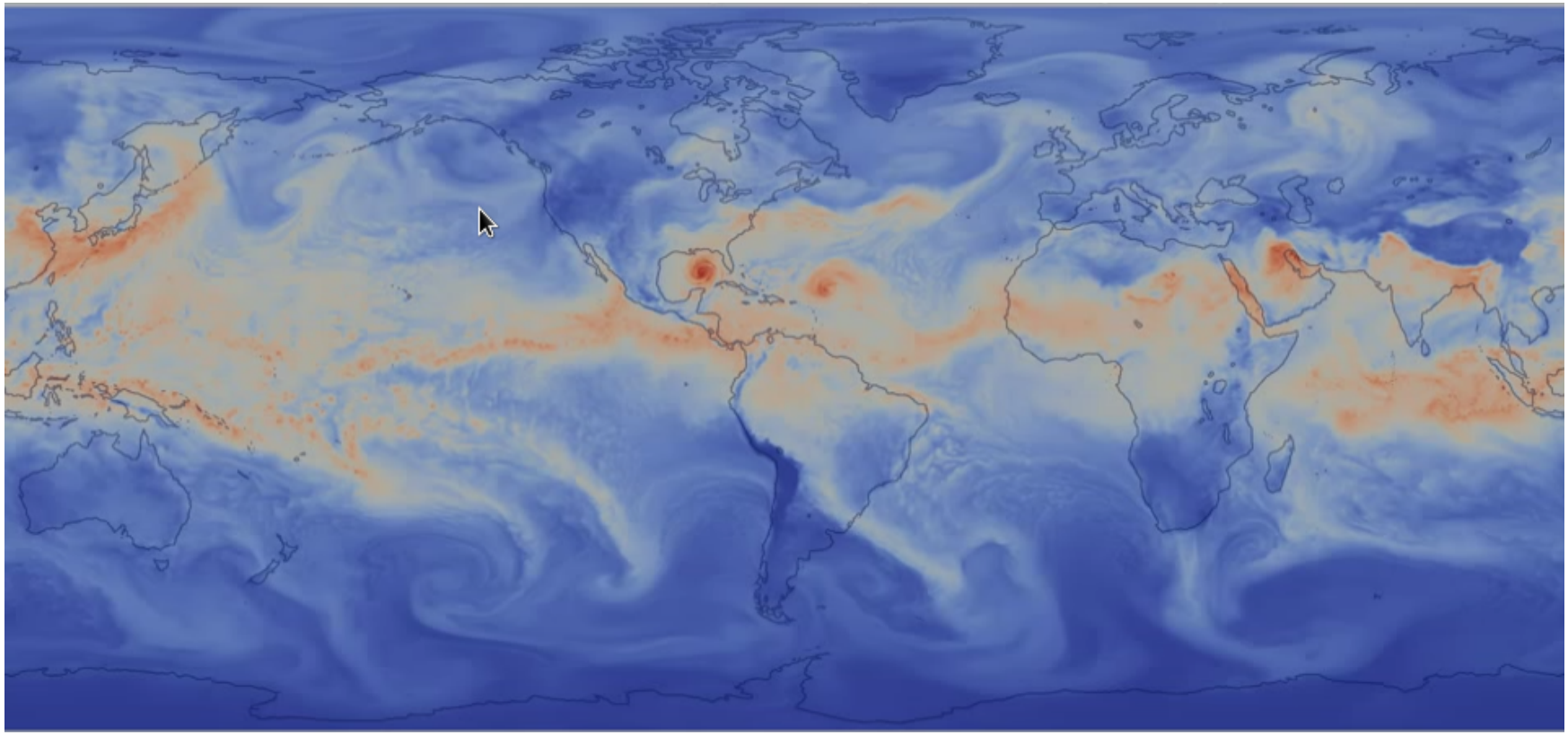
The ELF outperforms the best GC halfwidth (0.6) for most state variables and vertical levels.

Similar results are obtained from the ELF constructed from the output of an OSSE with GC halfwidth 0.4.

Tests of the empirical localization algorithm

- The dynamical core of the GFDL B-grid global atmospheric model: Localization for different observation types and state variable kinds
- The Community Atmospheric Model version 5 (CAM5): Vertical localization and localization for different geographic regions
- The Weather Research and Forecasting Model (WRF): Localization for regions with and without precipitation

A simulation of total precipitation water by CAM5



https://www.homme.ucar.edu

Experimental Design

Conduct OSSEs in DART/CAM system (Raeder et al. 2012).

CAM5 model:

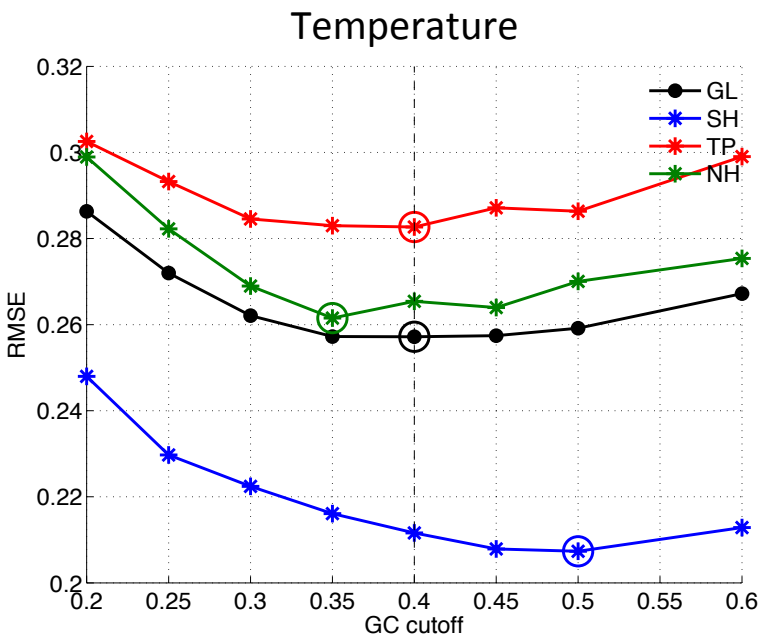
- Atmospheric component of the Community Earth System Model version 1 (CESM1; Gent et al. 2011)
- Finite volume grid with approximately 2° resolution (94x144) and 30 vertical levels
- Default configuration of the Atmospheric Model Intercomparison Project (AMIP; Gates 1992) protocol

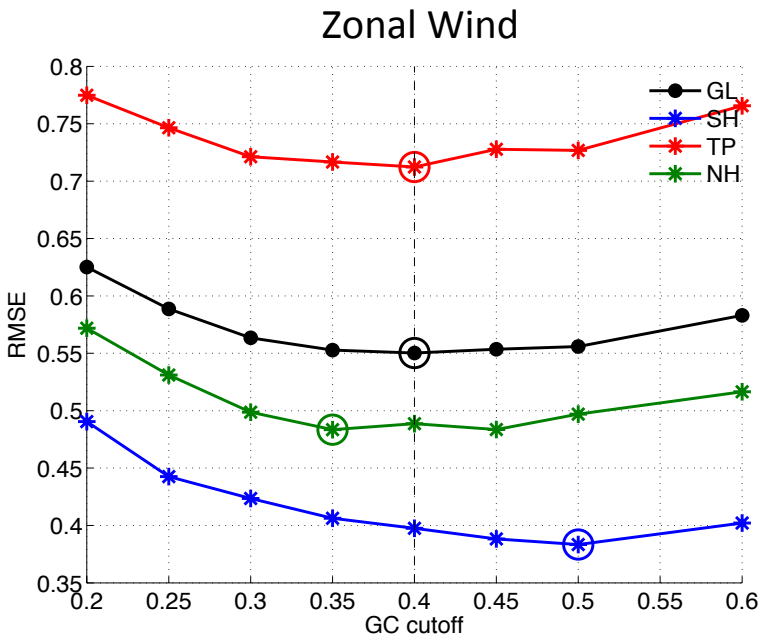
Data assimilation system:

- Ensemble adjustment Kalman filter (EAKF; Anderson 2001) in DART
- Spatially- and temporally-varying state space adaptive inflation (Anderson 2009)
- GC localization as the default

ELF in CAM

RMSE for Different GC Halfwidths





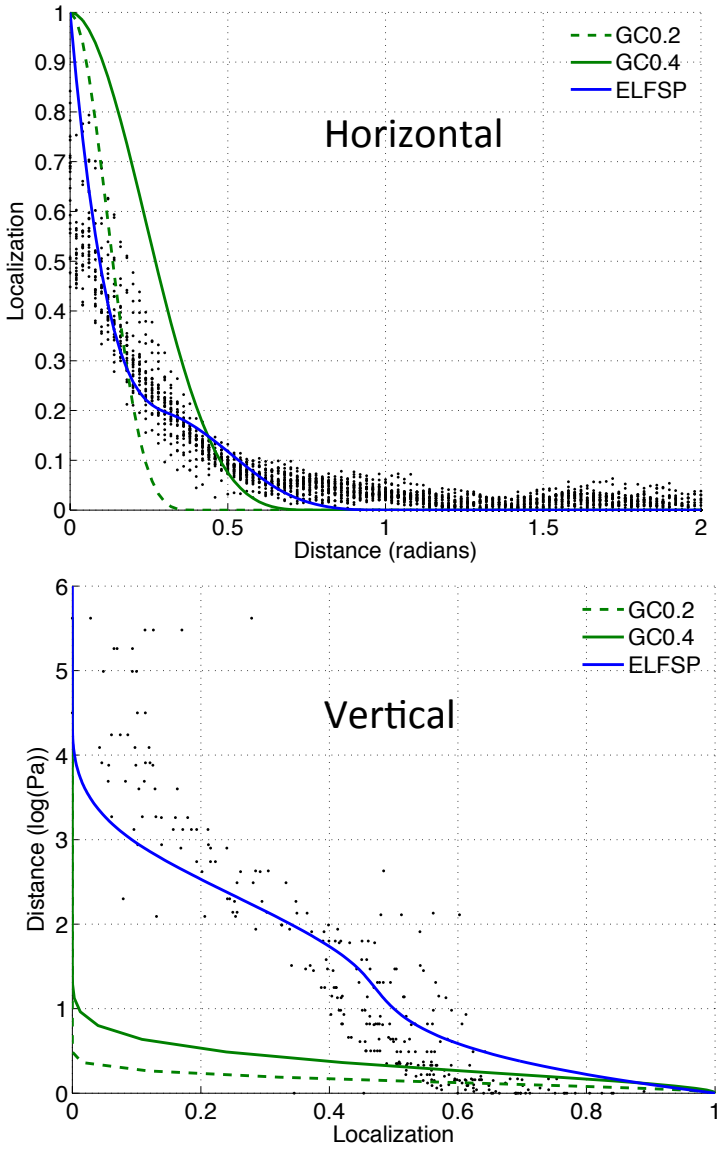
RMSEs for temperature and zonal wind are averaged globally (GL), in the southern hemisphere (SH), tropics (TP) and northern hemisphere (NH).

GC0.4 has smallest globally averaged RMSE, so 0.4 is chosen as the best halfwidth.

Some RMSEs computed for SH, TP and NH separately are smallest for other halfwidths; tuning the GC halfwidth is complex.

ELF in CAM

Horizontal and Vertical Empirical Localization Functions



Empirical localizations (black dots) are computed separately for temperature, zonal and meridional winds at ten levels (30 dots per separation).

A z-test is used to assess the significance of the empirical localization.

A cubic spline (blue line) is applied to the empirical localization to produce the final localization function (ELFSP).

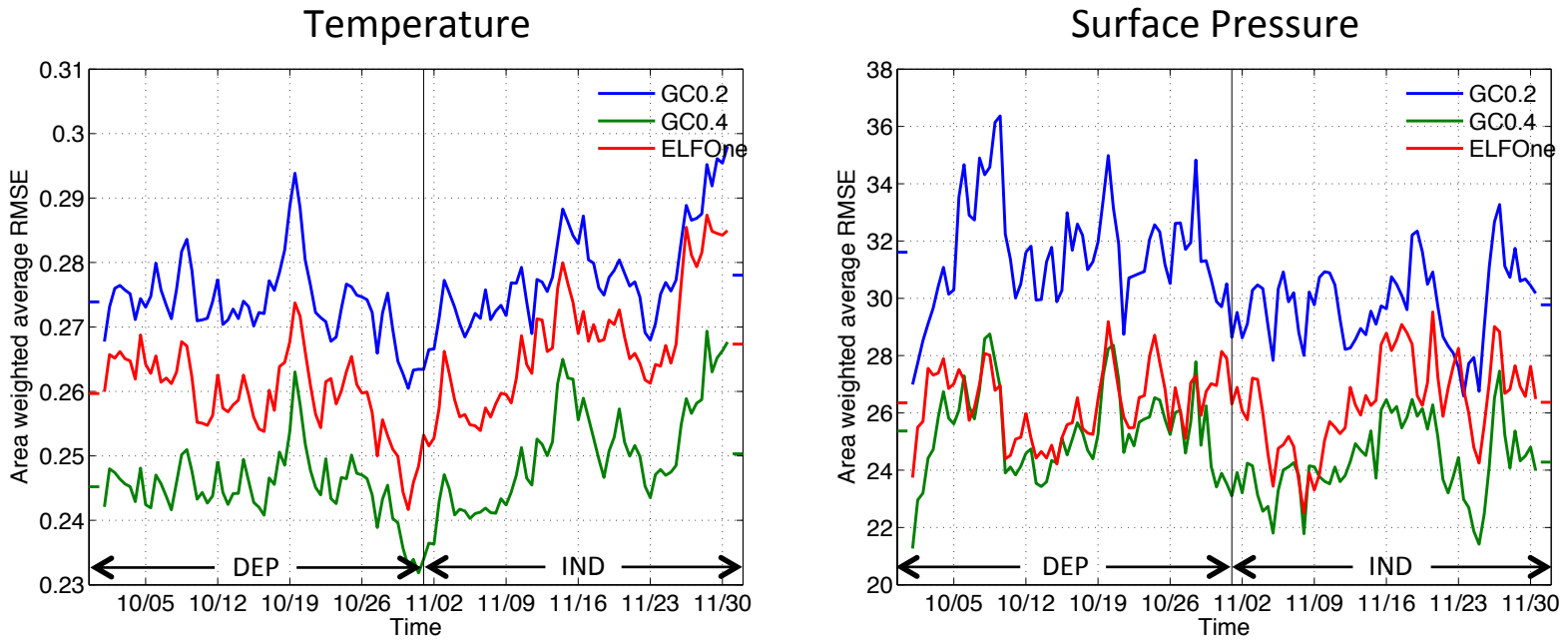
The horizontal ELFSP is smaller than the GC0.2 and GC0.4 at small separations and has a wider tail than GC0.2 and GC0.4.

The vertical ELFSP is much broader than the GC0.2 and GC0.4.

The horizontal and vertical ELFSPs are used in a subsequent OSSE (ELFOne).

ELF in CAM

Global Average RMSE for GC0.2, GC0.4 and ELFOne

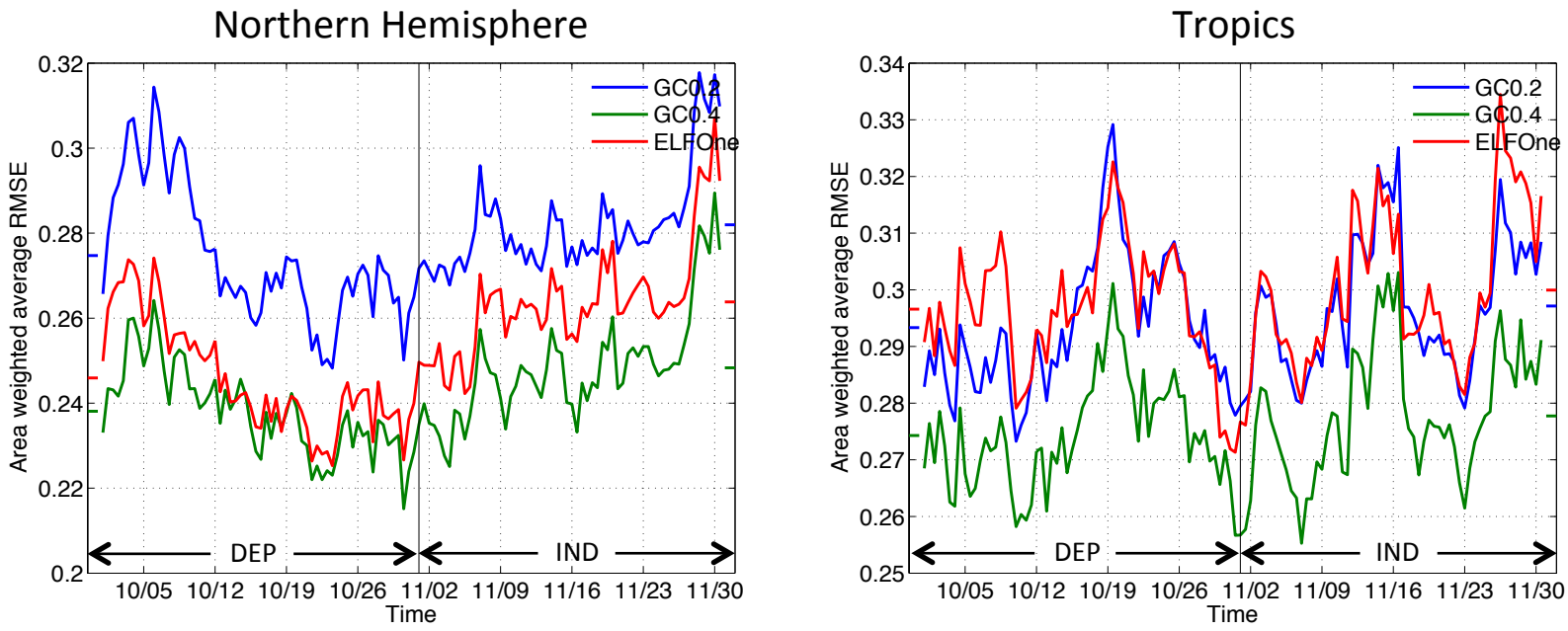


ELFOne has smaller temperature RMSE than GC0.2, but larger RMSE than GC0.4, the best GC.

ELFOne has smaller surface pressure RMSE than GC0.2, and slightly larger RMSE than GC0.4.

ELF in CAM

Temperature RMSE Averaged in NH and TP



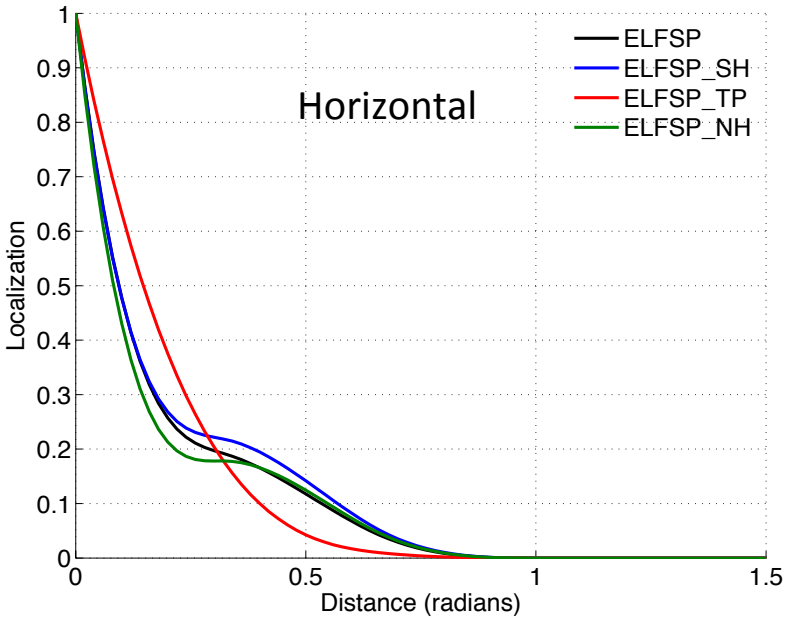
ELFOne has smaller temperature RMSE than GC0.2 in NH and SH.

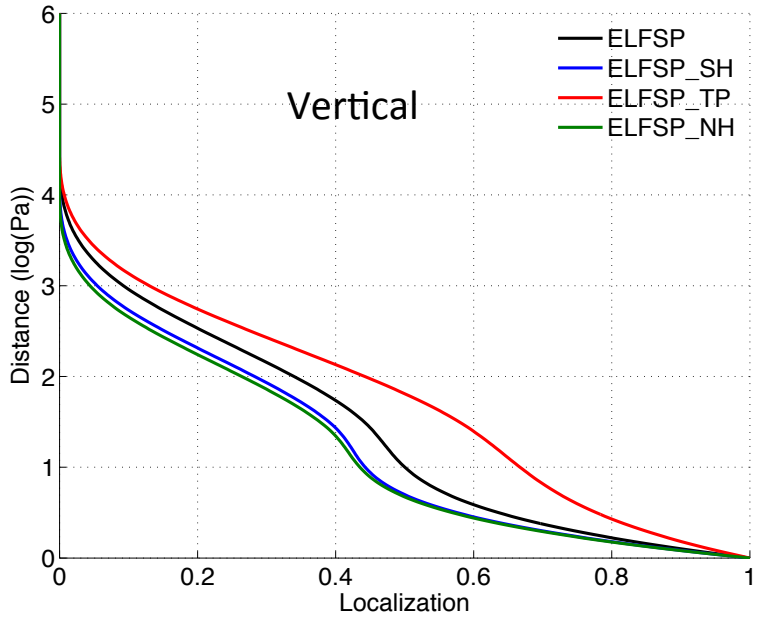
ELFOne has larger temperature RMSE than GC0.2 in TP.

Improvements of ELFOne over GC0.2 are mainly in SH and NH.

ELF in CAM

Horizontal and vertical ELFs varying by geographic regions





Horizontal and vertical ELFSPs are computed for the SH, TP and NH separately.

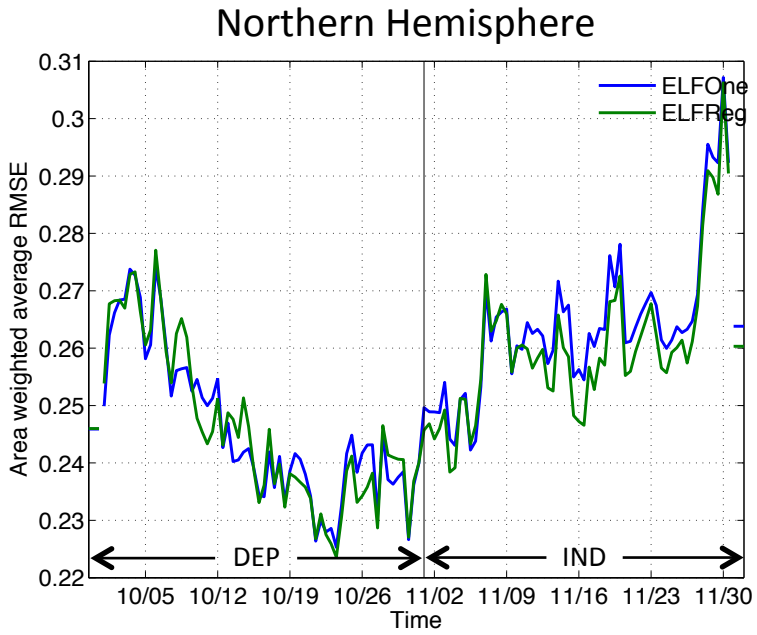
The horizontal ELFSP_SH and ELFSP_NH have similar shape to the global ELFSP. The horizontal ELFSP_TP has a more compact tail than the ELFSP, ELFSP_SH and ELFSP_NH.

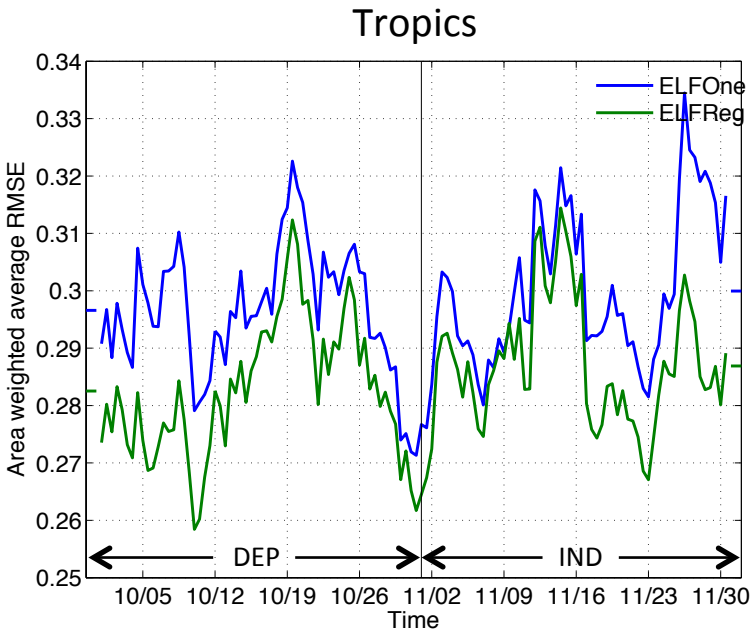
The vertical ELFSP_SH and ELFSP_NH are similar with smaller magnitude than the global ELFSP. The vertical ELFSP_TP is broader than the globa l ELFSP.

Horizontal and vertical ELFSPs varying by region are used in a subsequent OSSE (ELFReg).

ELF in CAM

Temperature RMSE Averaged in NH and TP





ELFReg has slightly smaller temperature RMSE than ELFOne in NH and SH.

ELFReg has smaller temperature RMSE than ELFOne in TP.

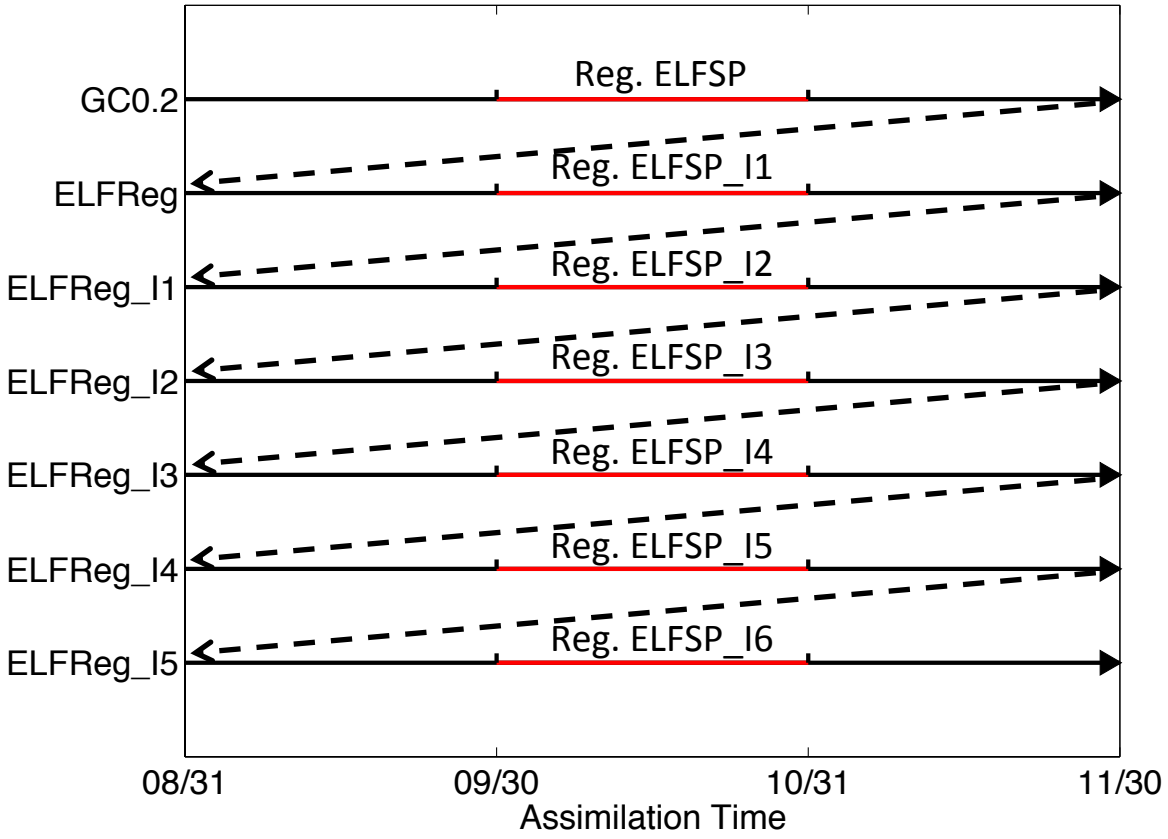
ELFReg has smaller globally averaged RMSE than ELFOne.

Convergence of the ELF with Cubic Spline Fit (ELFSP)

- The ELFSPs varying with geographic regions have advantages over the global ELFSP.
- The ELF appears to converge to a solution and lead to smaller error when the construction process of the ELF is iterated.
- Thus the convergence of the ELFSPs varying with region is examined.

ELF in CAM

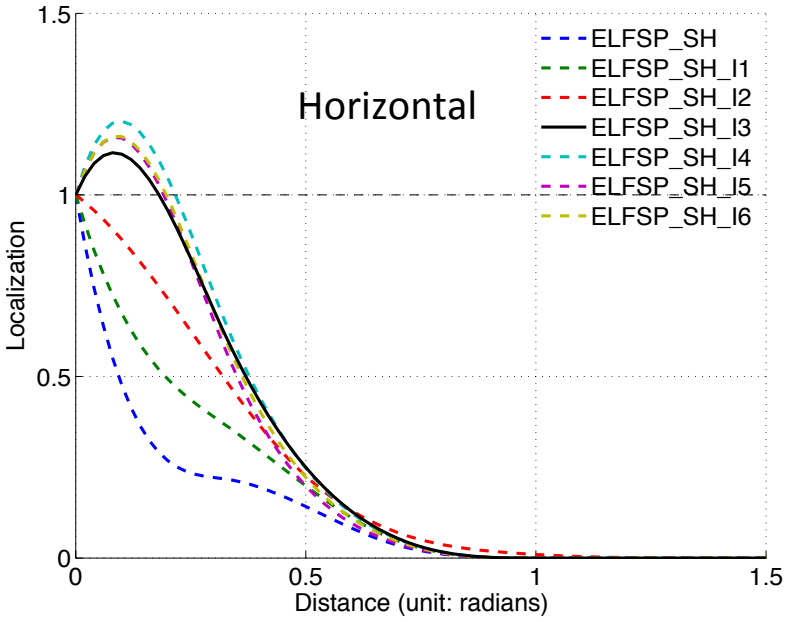
Convergence of the ELF with Cubic Spline Fit (ELFSP)

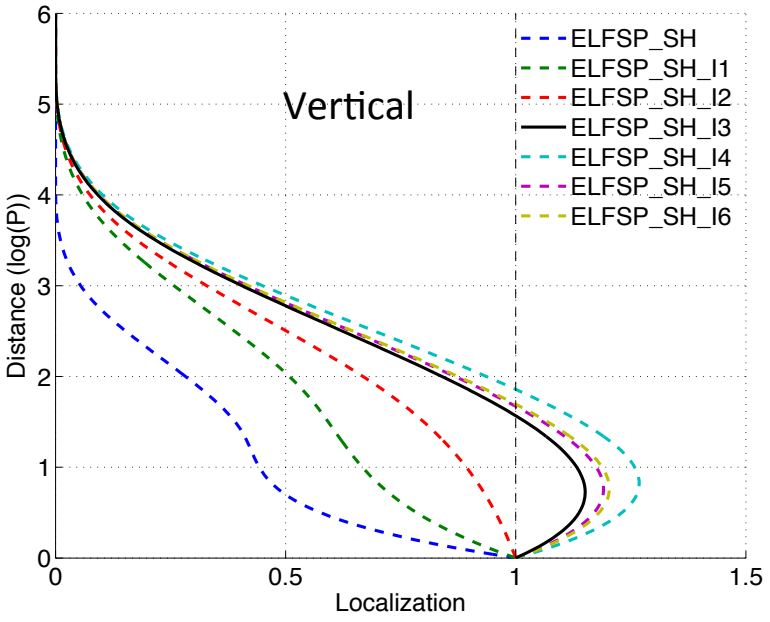


Five OSSEs (ELFReg_I#, #=1,…,5) are conducted iteratively. Each OSSE uses the regional ELFSPs computed from the output of the previous OSSE.

ELF in CAM

Convergence of the ELFSP: SH Example





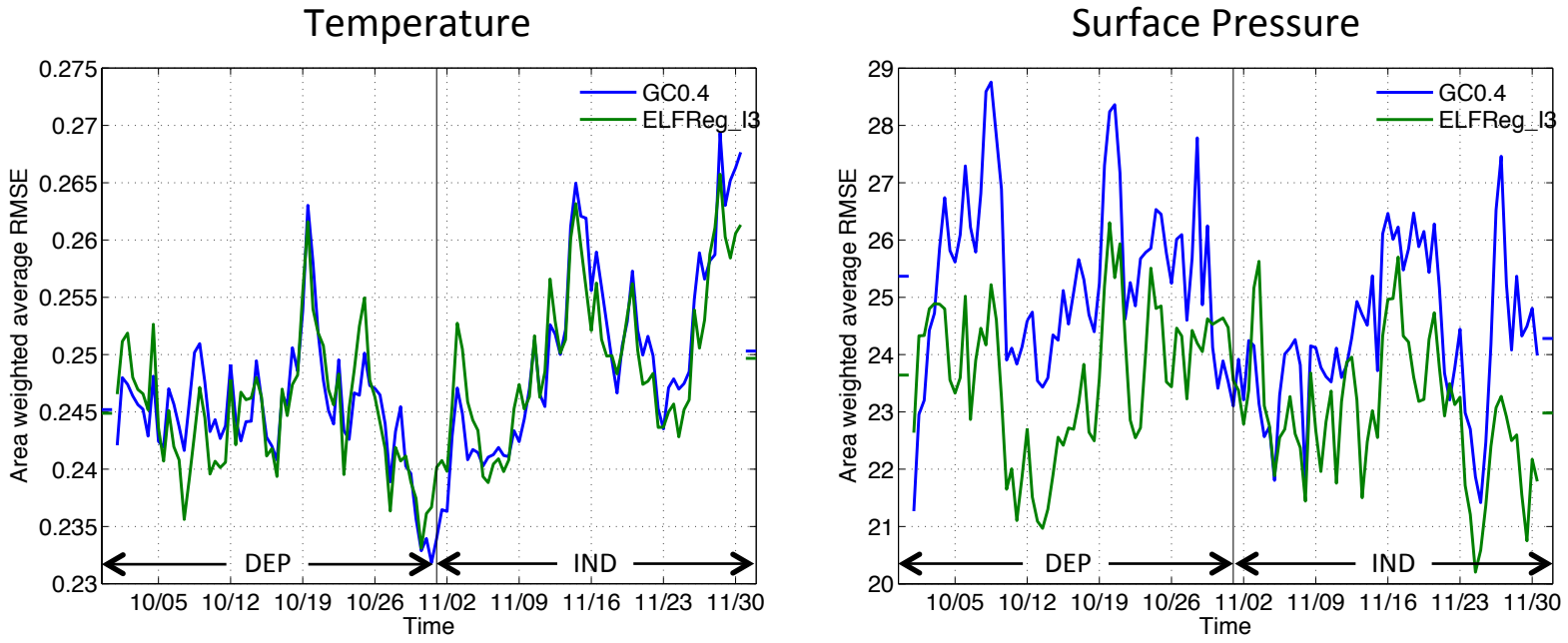
The ELFSP_SHs becomes larger with iterations.

The ELFSP_SHs appear to have mostly converged after 3 iterations.

The ELFSP_SHs from iterations 3 to 6 are larger than 1.0 at small separations. This indicates insufficient spread and the empirical localization acts as an inflation.

Empirical localization values larger than 1.0 are set to 1.0 when used in an OSSE.

Global Average RMSE for GC0.4 and ELFReg_I3

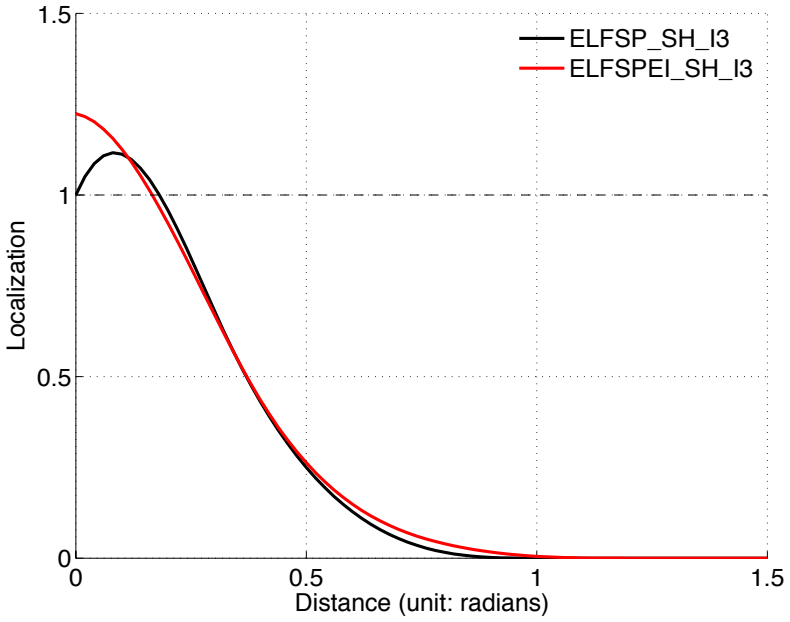


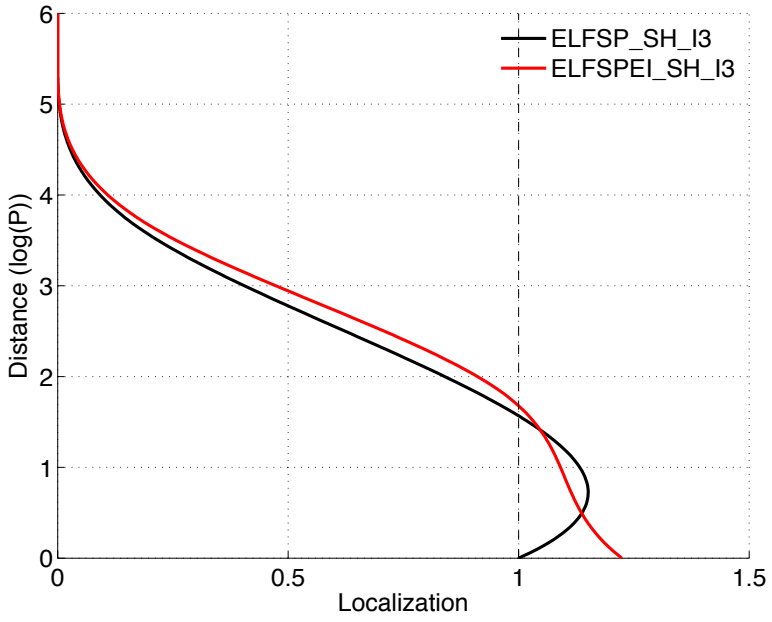
ELFReg_I3 produces slightly smaller temperature RMSE than GC0.4.

ELFReg_I3 has significantly smaller surface pressure RMSE than GC0.4

ELF in CAM

ELFSPs with Empirical Inflation



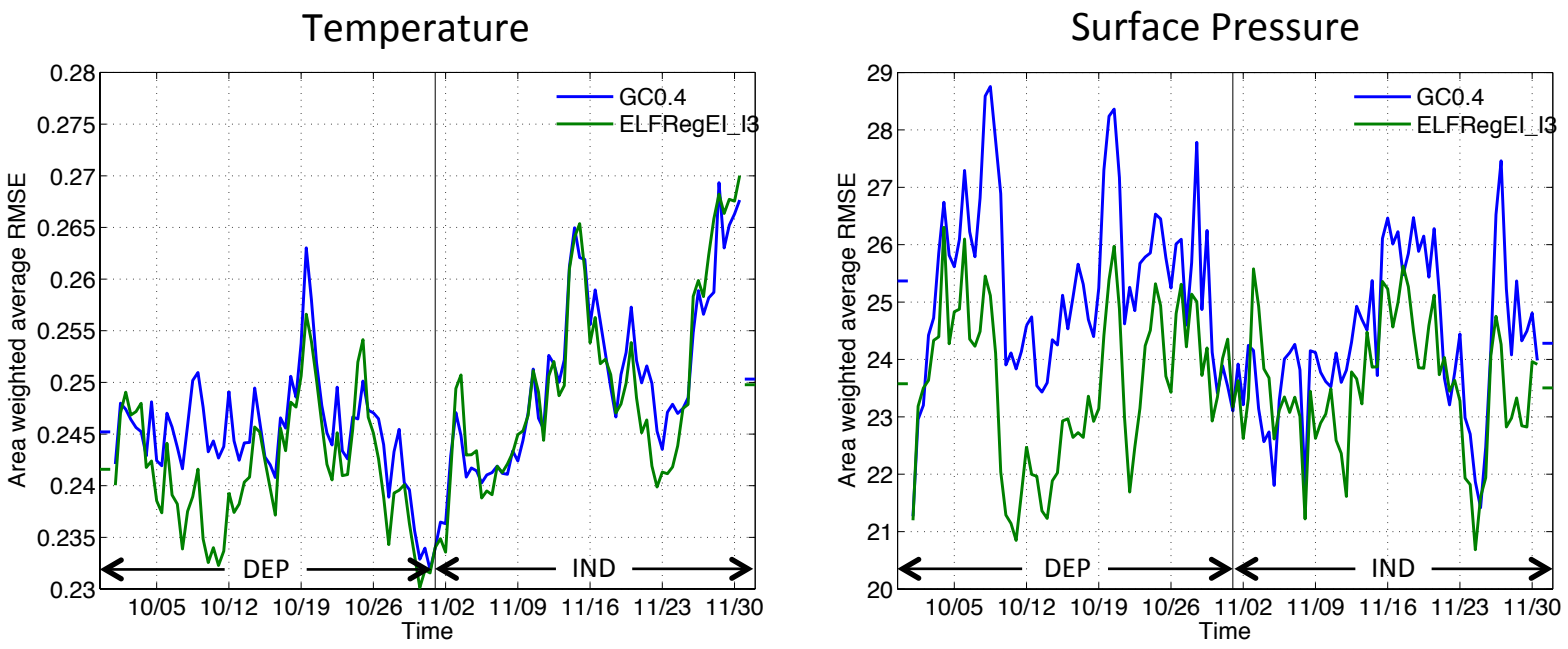


The cubic spline fit of ELF with empirical inflation (ELFSPEI) has values larger than 1.0 at small separations.

Horizontal and vertical ELFSPEIs are used in a subsequent OSSE (ELFRegEI).

ELF in CAM

Global average RMSE for GC0.4 and ELFRegEI_I3



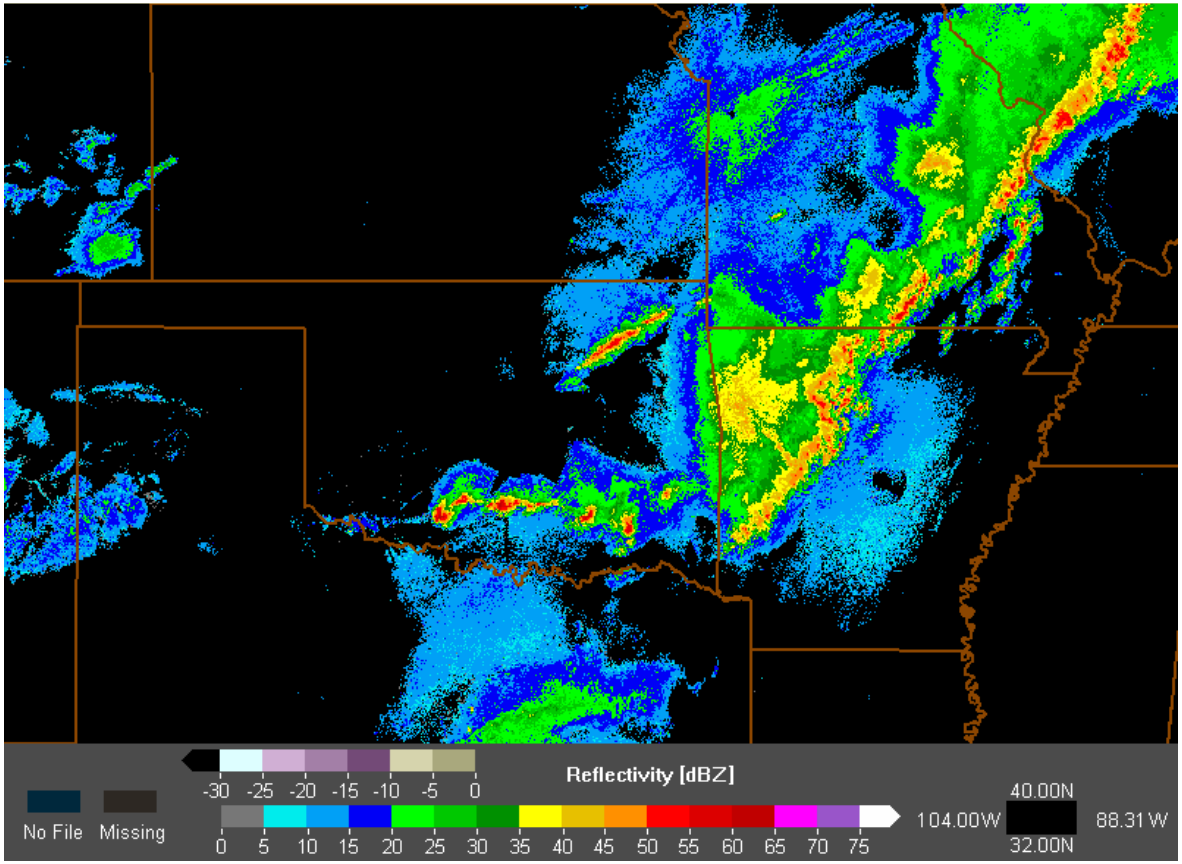
ELFRegEI_I3 produces smaller temperature RMSE than GC0.4.

ELFRegEI_I3 has significantly smaller surface pressure RMSE than GC0.4

Tests of the empirical localization algorithm

- The dynamical core of the GFDL B-grid global atmospheric model: Localization for different observation types and state variable kinds
- The Community Atmospheric Model version 5 (CAM5): Vertical localization and localization for different geographic regions
- The Weather Research and Forecasting Model (WRF): Localization for regions with and without precipitation

Is different localization needed for different weather?



(http://nmq.ou.edu/applications/qvs_2d_maps.html)

Experimental Design

Conduct OSSEs in DART/WRF system.

WRF model V3.3.1 :

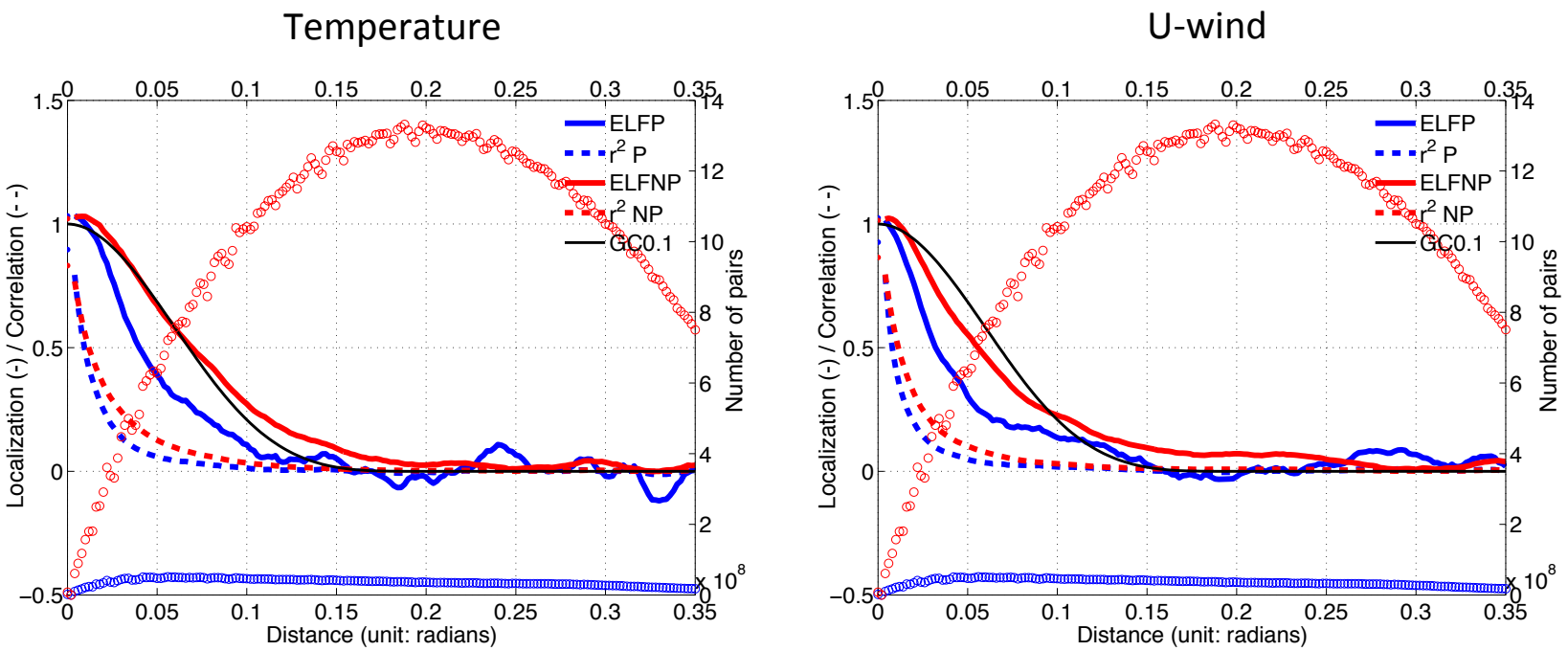
- CONUS domain with horizontal grid spacing 15 km, 40 vertical layers and model top at 50 hPa
- Model physics: RRTMG long wave and short wave radiation schemes, Thompson 2-moment microphysics scheme, Noah land surface model, MYJ PBL scheme, and Tiedtke cumulus scheme

Data assimilation system:

- EAKF in DART
- Spatially- and temporally-varying state space adaptive inflation
- GC localization of halfwidth 0.1 radians as the default

ELF in WRF

Horizontal Empirical Localization Functions

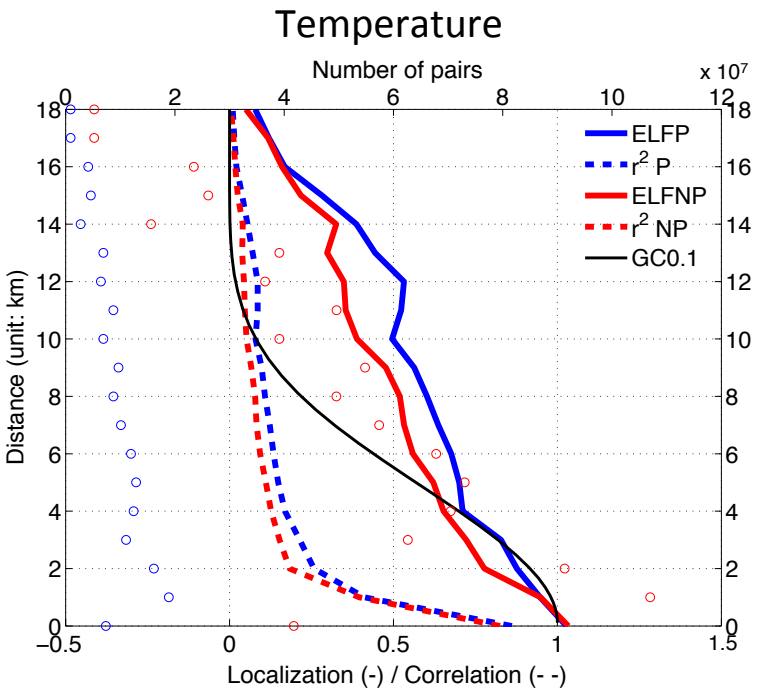


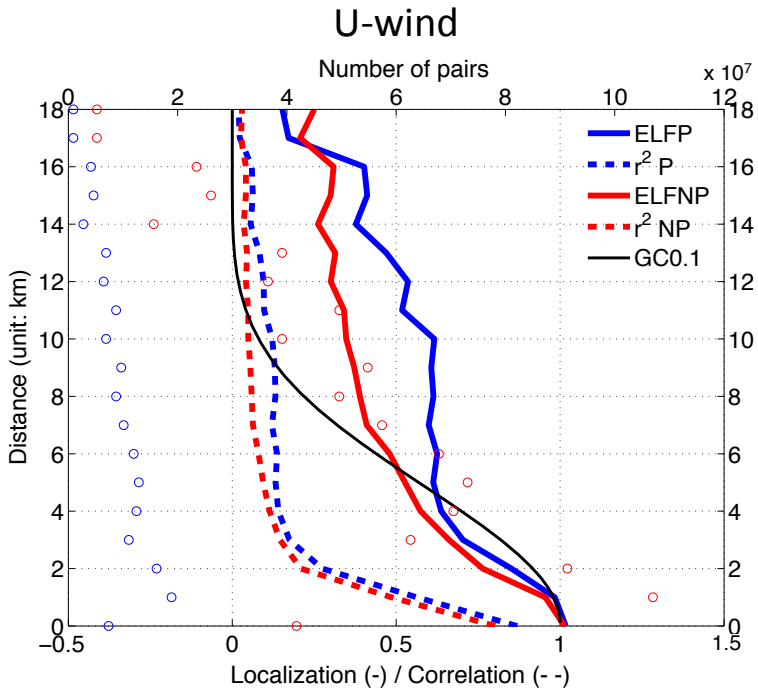
The ELFs for non-precipitating regions (ELFNP) have similar shape to GC0.1, but ELFNP of u-wind is smaller than GC0.1 for small separations.

The ELFs for precipitating regions (ELFP) are narrower than GC0.1 and ELFNP.

The correlation coefficient of ELF for precipitating regions decreases faster than for non-precipitating regions.

Vertical Empirical Localization Functions





The vertical ELFs for precipitating regions generally have larger localizations.

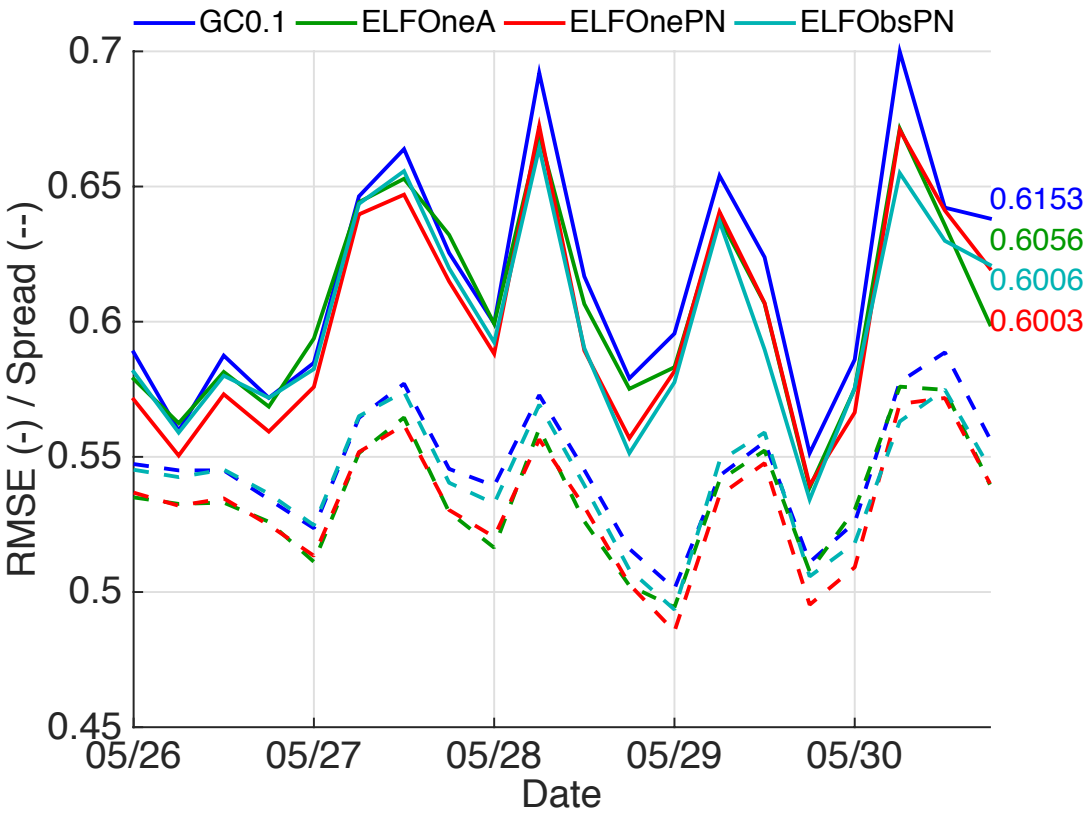
The vertical ELFP of temperature decreases more quickly with height than for u- and v-winds between 4 and 10 km.

The correlation coefficient of ELF for precipitating regions is larger.

Vertical Empirical Localization Functions

Exp. name	Applied localization function
GC0.1	GC localization function with half-width of 0.1 radians.
ELFOneA	One horizontal and one vertical ELFF that are computed from the output of GC0.1.
ELFOnePN	From the output of GC0.1, two horizontal and two vertical ELFFs that vary with precipitating and non-precipitating regions.
ELFObsPN	From the output of GC0.1, one horizontal and one vertical ELFF of temperature and one horizontal and one vertical ELFF of u- and v-wind for precipitating regions, and similarly four ELFFs for non-precipitating regions.

Average Temperature RMSE for GC0.1 and ELFs



ELFOneA yields slightly smaller (but statistically significant) RMSE than GC0.1.

ELFOnePN has slightly smaller (but statistically significant) RMSE than GC0.1 and ELFOneA, thus the advantages of varying localization for precipitation and non-precipitating regions are demonstrated.

But the localization functions varying by observation types (ELFObsPN) do not show additional benefits than ELFOnePN.

The performance of EnKF can be improved with the automatic localization algorithm ELF. Improved EnKF can lead to improved applications.

Can more frequent assimilation of surface pressure observations reduce the uncertainty of the entire troposphere?

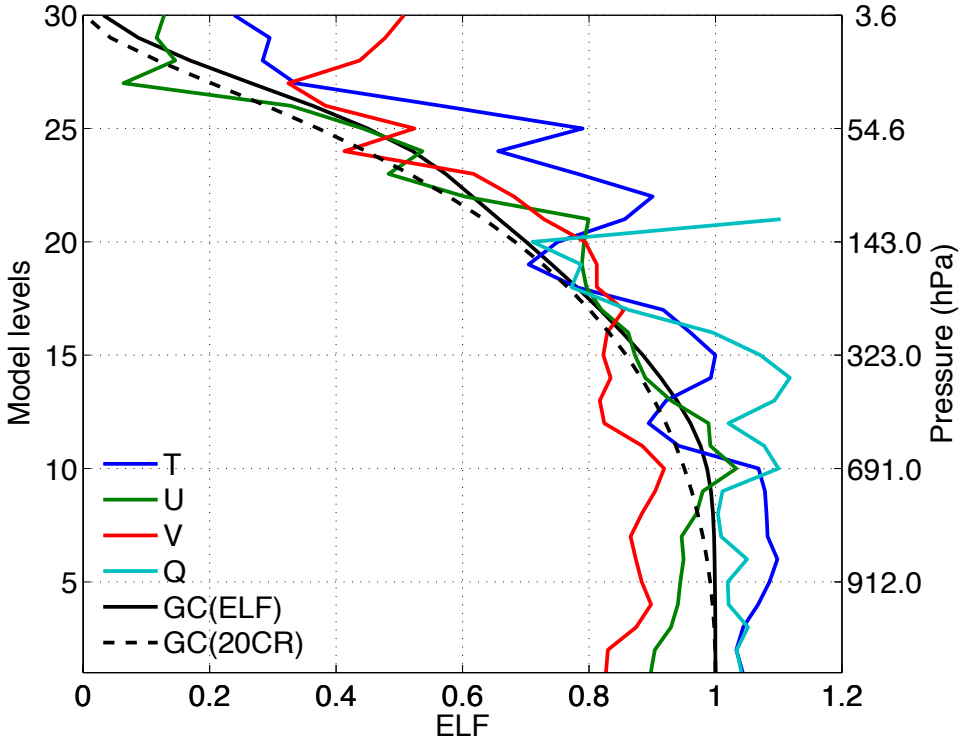
Conduct OSSEs in DART/CAM.

Assimilate uniformly distributed synthetic observations of surface pressure (7200 sites on the sphere).

Synthetic observations are available every 6, 3 or 1 hour.
(The 20th Century Reanalysis (20CR; Compo et al. 2011) assimilated only surface pressure observations every 6 hours.)

Two seasons, summer and winter, are examined.

Vertical Localization for Surface Pressure Observations

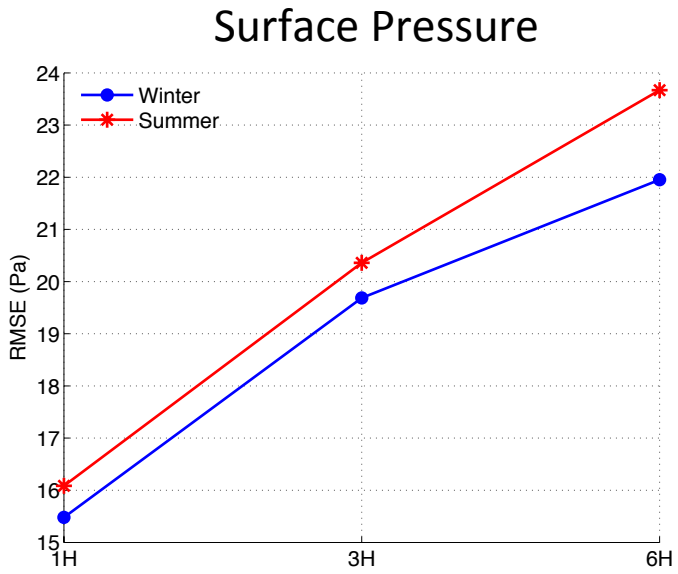


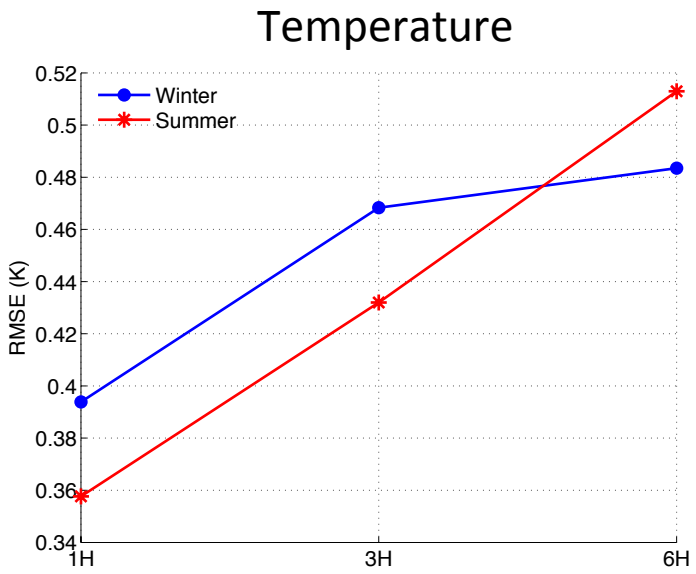
ELFs are computed for surface pressure observations with state variables of temperature, zonal and meridional winds, and specific humidity in the same column.

ELFs extend nearly vertically till model level 15 (~300 hPa), and then gradually decrease to 0 till model top.

A GC localization that fits the ELFs is used as vertical localization for surface pressure observations, which is similar to the GC function currently used in the 20CR.

Temporally and Spatially Average RMSE

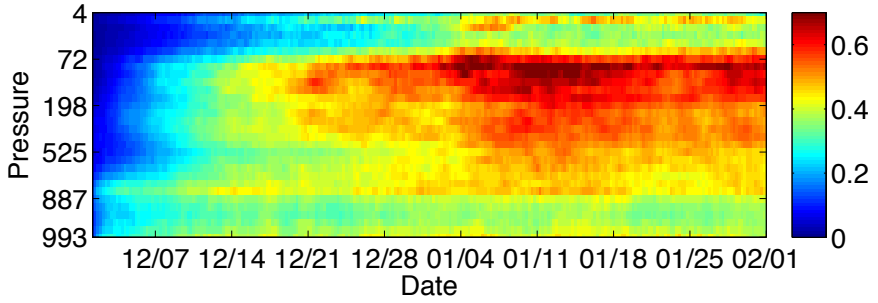




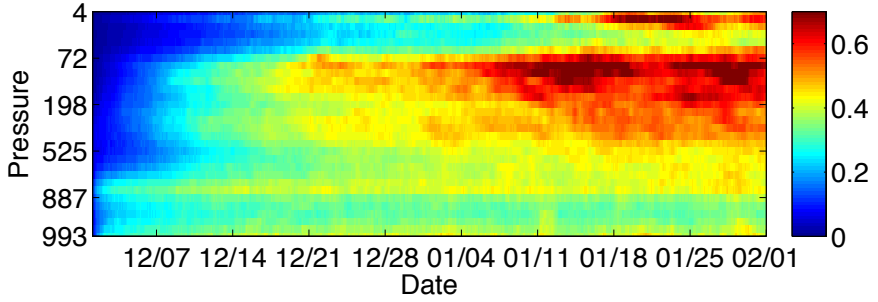
For both directly and non-directly observed state variables, the average RMSE decreases with increasing assimilation frequency.

Time series of Temperature RMSE Profile

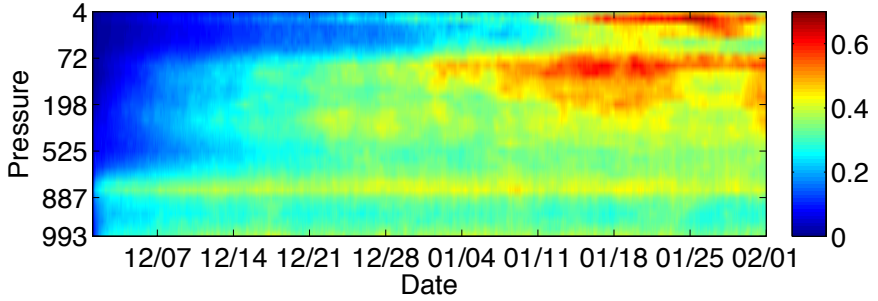
6-hour



3-hour



1-hour



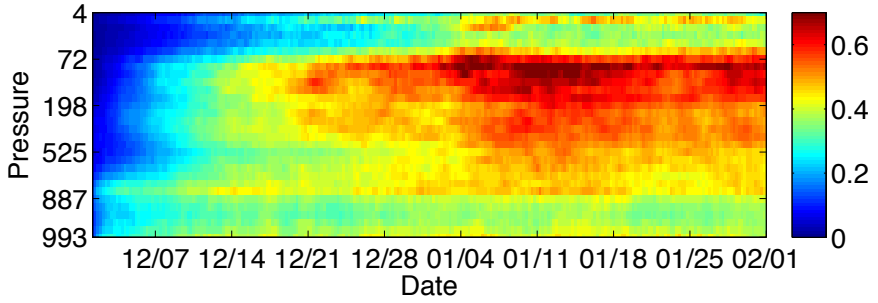
The uncertainty throughout the entire troposphere can be constrained by observing only surface pressure.

The information of surface pressure observations is spread upward more quickly with more frequent assimilation.

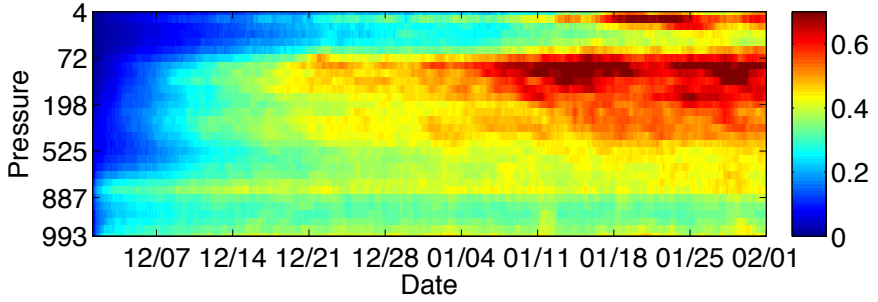
The error of the entire depth of the troposphere, especially the middle troposphere, can be better constrained with increased observation frequency (1 hour).

The frequent assimilation of surface pressure observations with the ELF could help to improve future versions of the 20CR.

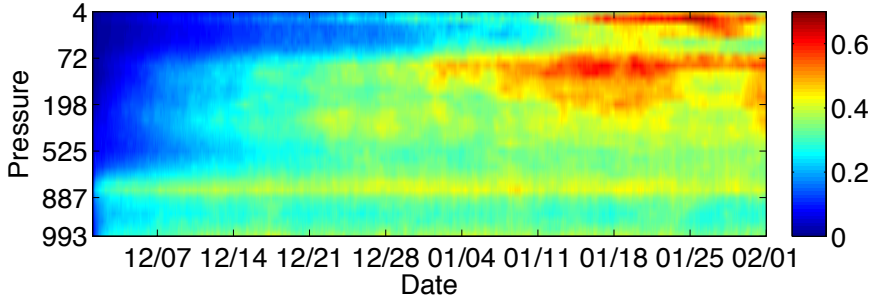
6-hour



3-hour



1-hour



The uncertainty throughout the entire troposphere can be constrained by observing only surface pressure.

The information of surface pressure observations is spread upward more quickly with more frequent assimilation.

The error of the entire depth of the troposphere, especially the middle troposphere, can be better constrained with increased observation frequency (1 hour).

Conclusions

- The empirical localization algorithm uses the output from an OSSE and constructs localization functions that minimize the RMS difference between the truth and the posterior ensemble mean for state variables.
- This algorithm can automatically provide an estimate of the localization function and does not require empirical tuning of the localization scale.
- It can compute an appropriate localization function for any potential observation type and kind of state variable, for different geographic regions and weathers.
- It plays the role of empirical inflation when needed.
- The empirical localization function generally outperforms the best GC localization in the GFDL B-grid model, CAM and WRF.