The Developmental Testbed Center (DTC) Annual Operational Plan (AOP) 2015 Project for the National Oceanic and Atmospheric Administration (NOAA) Office of Oceanic and Atmospheric Research (OAR)

Final report documenting:

Regional Ensemble-based (and hybrid) Data Assimilation Testing and Evaluation

June 2016

Table of Contents

1. Abstract

As highlighted by the Developmental Testbed Center (DTC) Science Advisory Board (SAB), an effort needs to be made to build capability and support for ensemble-based (including hybrid) data assimilation (DA) for regional applications. The SAB recommends that the DTC emphasize convective-scale NWP and ensembles, focusing on the next generation Ensemble-Variational (EnVar) DA and the challenge of using dual dynamic cores. It is also of interest to the DTC partners, including NCEP's Environmental Modeling Center and the NOAA Earth System Research Laboratory (ESRL) to advance such a system through objective tests and evaluation. In response to this feedback, the DTC DA team proposed building capabilities and performing regional ensemble-based DA tests in the context of operational regional applications at NCEP. This work addresses future operational DA systems and complements the use of covariances from global models in current regional systems at the National Centers for Environmental Prediction (NCEP).

The DTC tested and evaluated 4D EnVar capabilities for regional mesoscale scale applications. This work started with an examination of current available capabilities of running ensemble-based DA using the Advanced Research Weather Research and Forecast (WRF) (ARW) forecast system. The DTC evaluated the performance of ensemble-based GSI and impacts of running such a system on regional forecasts, against GSI hybrid using global ensemble as in the current operational regional systems at the National Oceanic and Atmospheric Administration (NOAA) for its Rapid Refresh (RAP) system. The results show a great potential of using 4D EnVar to improve mesoscale analyses and forecasts. However, a few working areas were also identified for further improvement of this GSI capability. The DTC fixed some GSI code bug based on the findings and also suggests an overall examination of the GSI code and configurations for an optimal performance.

2. Single observation tests

In order to demonstrate the impact of the 4D EnVar using the ARW dynamical core, a series of experiments were conducted to assimilate a single simulated temperature observation. The resulting increment in each observation time window was studied to better understand how the 4D Envar capability was applied and ensure that the algorithm was working properly for ARW applications, with a specific focus on the RAP system.

2.1 Single observation experiment configurations

Pseudo-single observation tests (PSOT) were conducted using backgrounds from sample ARW cases as well as RAP cases, using an 80 member global ensemble. Experiments were conducted using the GSI 3DVar, 3D hybrid EnVar and 4D hybrid EnVar data assimilation methods. Table 1.1 summarizes the various PSOT

configurations tested. For each single observation experiment, a -2 K temperature innovation was placed at 500 hPa with a 1.0-degree observation error standard deviation. For the 3D EnVar and 4D EnVar experiments, the ensemble contribution was set 75% ensemble (25% static) for the background error covariance, with ensemble localization set to a horizontal length scale of 330 km and 3.0 scale heights in the vertical. Modifications for observation error, ensemble contribution to the background error covariance and frequency of global ensemble time levels were tested, as shown in table 1.1.

Table 1.1: Description of single observation experiments used for RAP configuration

Experiment	Background Error	Ensemble	Observation error
3DVar	Static = 100%	No ensemble	1.0
3DEnVar	Static = 25%, Ensemble = 75% Static = 0% Ensemble = 100%	Global (6 h window)	1.0
4DEnVar	Static = 0% Ensemble = 100% Static = 25% Ensemble = 75%	Global: –3, 0, +3 h (6 h window) Global: -1, 0, +1 h (2h window)	1.0, 1.0, 1.0 0.5, 1.0, 0.5 0.1, 1.0, 0.1

2.2 Results

An ARW case on 17 June 2014 was identified as an initial test case using the 4DEnVar method. As shown in Figure 2.1a, the PSOT was placed in a trough region with strong 500 hPa flow. Initial results when testing a sample ARW case indicated that the -3 and +3 hour time windows resulted in a lower magnitude increment, compared to the center observation time bin (Figure 2.1b).

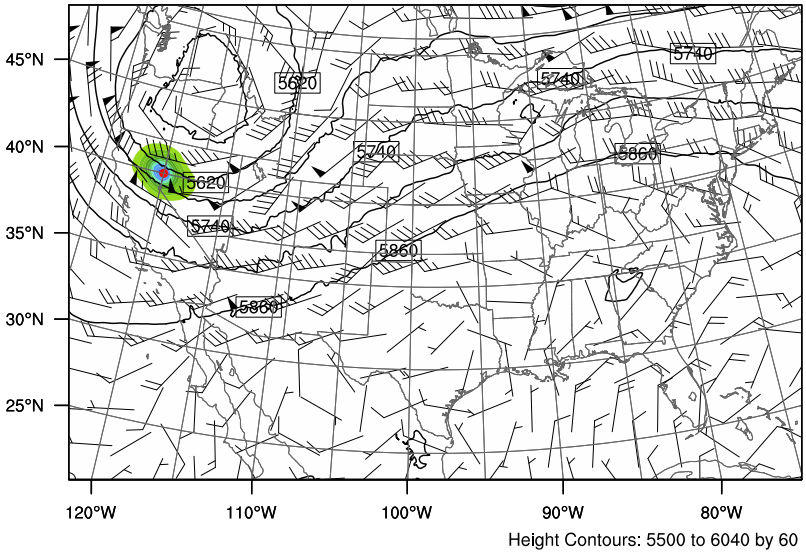

Figure 2.1a. 500 hPa temperature increment (shaded, K) resulting from a single observation test (location indicated by red dot). Geopotential height (contours, dm) and vector wind (m/s) valid are at the analysis time.

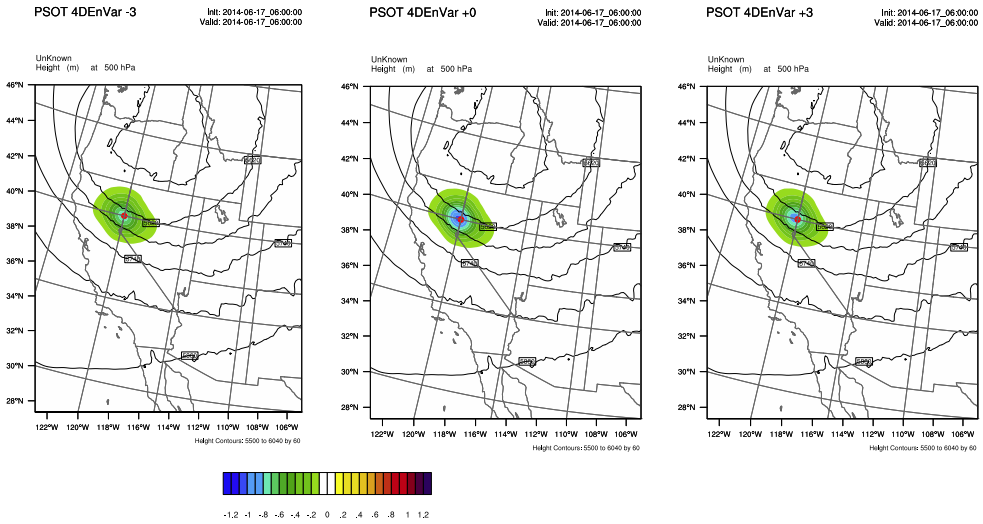

Figure 2.2b. Same as Figure 2.1a, except resulting 500 hPa temperature increment for -3 (left), 0 (center) and +3 (right) observation time bins.

In order determine if this region is representative of a location where the 4D EnVar increments differ from the 3D EnVar increments, the analysis increment difference between the 4D EnVar and 3D EnVar real data cases were compared (not shown), indicating an area over the great lakes region may be more suitable. A PSOT over this region further demonstrated the reduction in magnitude outside on the center time window (Figure 2.1c).

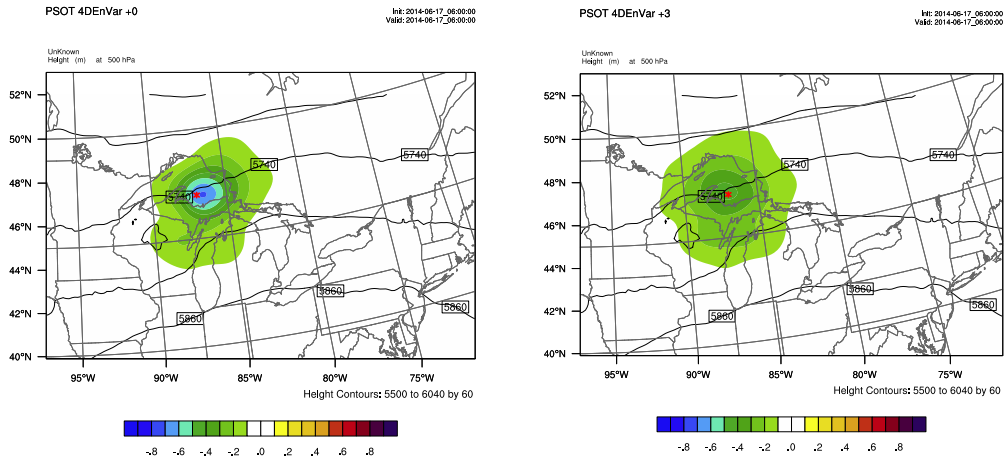

Figure 2.1c. Same as Figure 2.2b, except location of maximum 4D EnVar increment change for 0 (center) and +3 (right) observation time bins.

A RAP case valid on 2014 August 09 06Z was explored to study the behavior of the 3DVar, 3DEnVar, and 4DEnVar cases within the RAP configuration. An 80-member global ensemble with time levels -3, 0 and +3 hr within a 6 hr time window were used. The temperature single observation was placed once again in a trough at 500 hPa. The resulting analysis increments are shown in Figure 2.2.

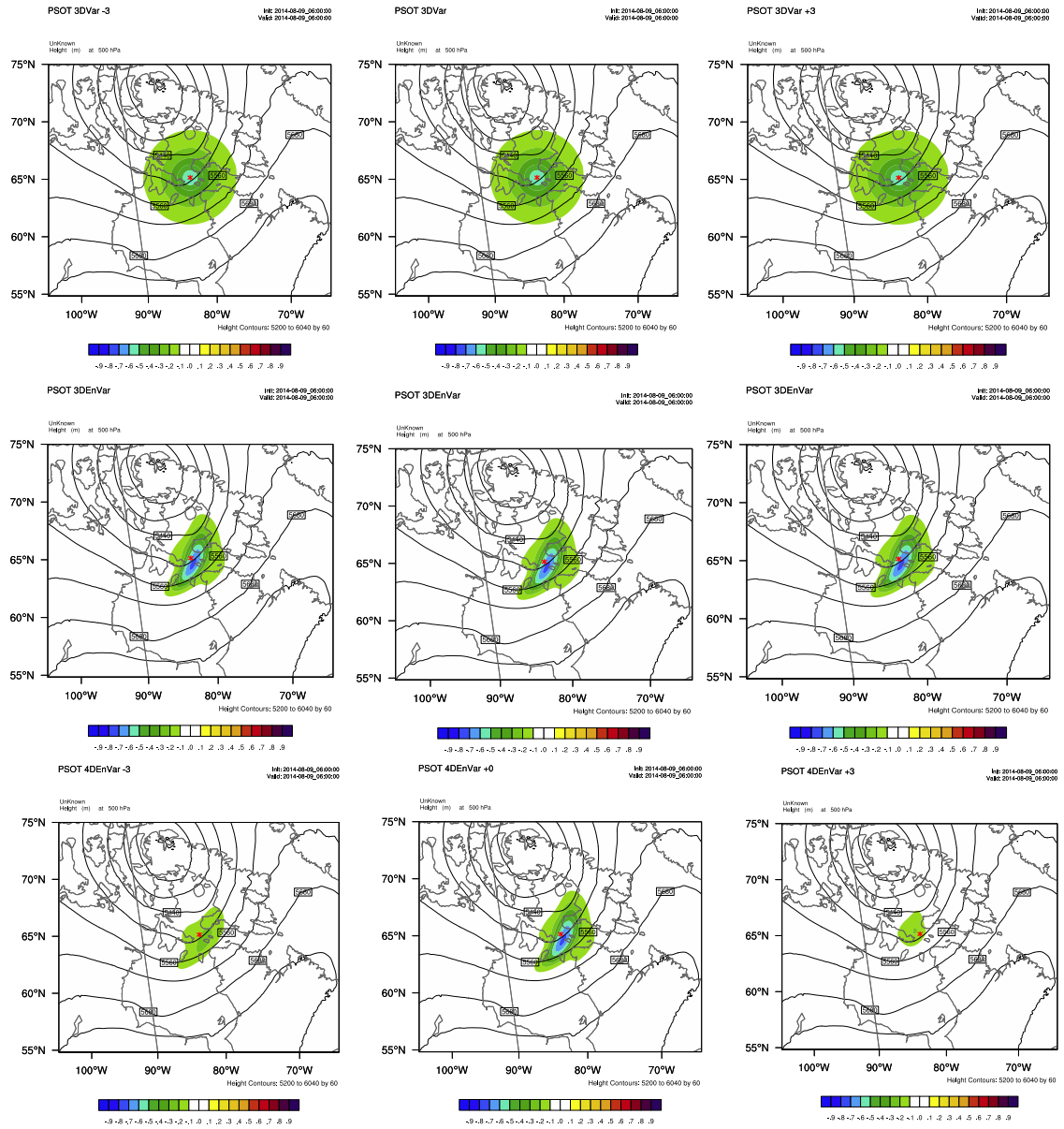

Figure 2.2. As in Figure 2.1c, except 3DVar increments (top), 3DEnVar increments (middle), and 4DEnVar increments (lower) for 3 time levels: -3 (left), 0 (center), +3 (right) within a 6 hr time window.

Figure 2.2 demonstrates the impact of the static BE in the 3DVar case, and the resulting flow dependence using the 3DEnVar method. Both 3DEnVar and 4DEnVar cases show a cyclonic wind response to the temperature increment. The impact of

utilizing the 4DEnVar methodology is evident with the +/-3 observation bins, as the impact for the 3DEnVar case does not produce time-variant information, as with the 4D EnVar case. Similar to the results shown for the sample ARW case, a lower magnitude impact at the non-center time bins is present for the 4DEnVar case. To further investigate the impact of the observation error and ensemble spread on the magnitude of the non-center observation time bins, additional PSOTs were generated using the 4D EnVar method. Figure 2.3 shows the impact of varying the observation error for the non-center time bins, which results in a noticeable increase in the magnitude of the temperature increment.

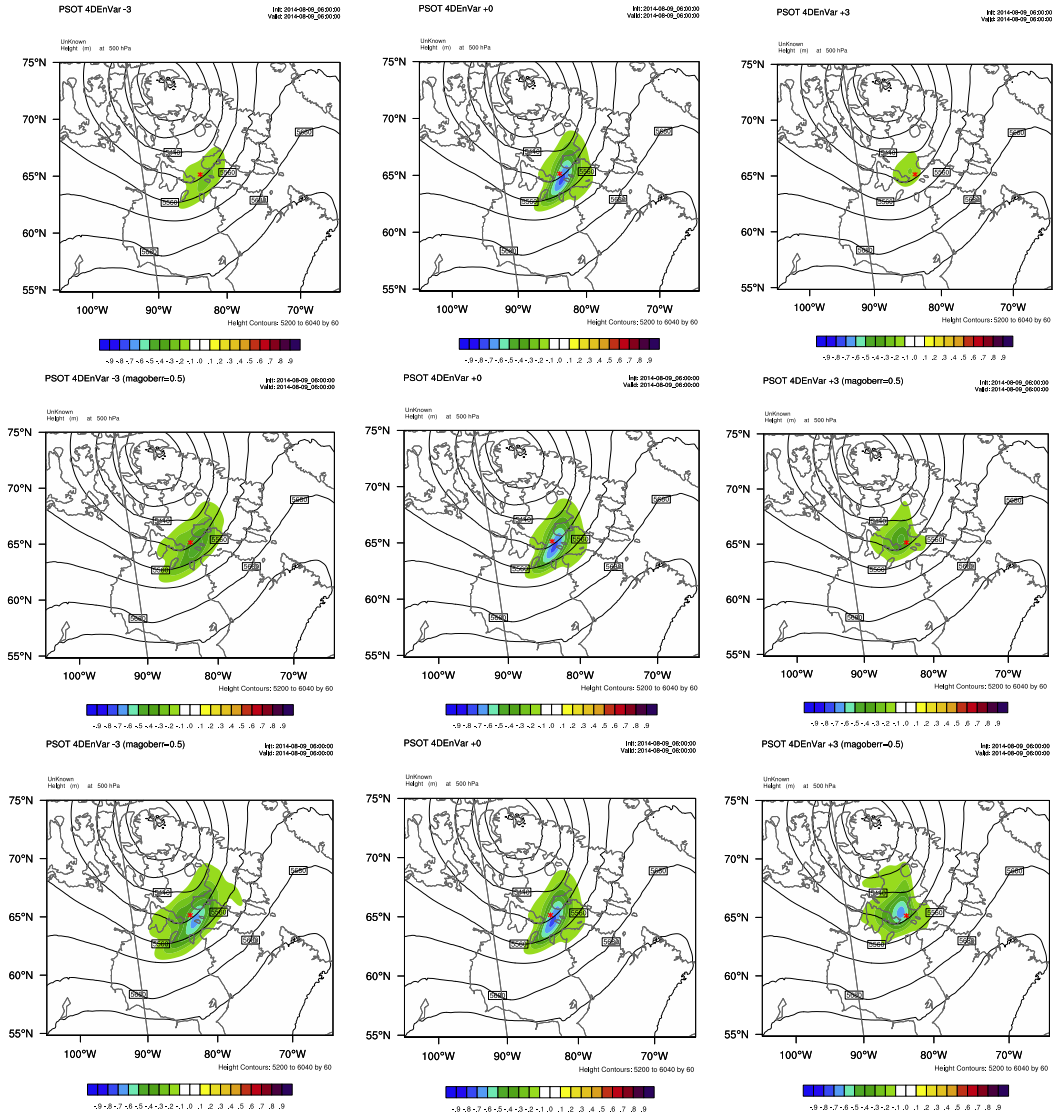

Figure 2.3. As in Figure 2.2, except 4DEnVar tests with 1.0 K (top row) observation error for -3 (left column), 0 (center column), +3 (right column) hr within a 6 hour time window, 0.5 K observation error (middle row), and 0.1 K observation error (lower row).

Although modifying the observation error for the non-center bins increased the increment magnitude as expected, the test was performed to investigate any unexpected reactions to modifications to the observation error in the non-center time bins. The impact of the ensemble on the non-center time bins was also investigated. Figure 2.4 demonstrates that the ensemble has a great impact on the magnitude and size of the increment for the 4DEnvar method by removing the contribution from the static background error. The minimal impact from the ensemble at the non-center time bins is very apparent for the 4DEnVar pure ensemble case.

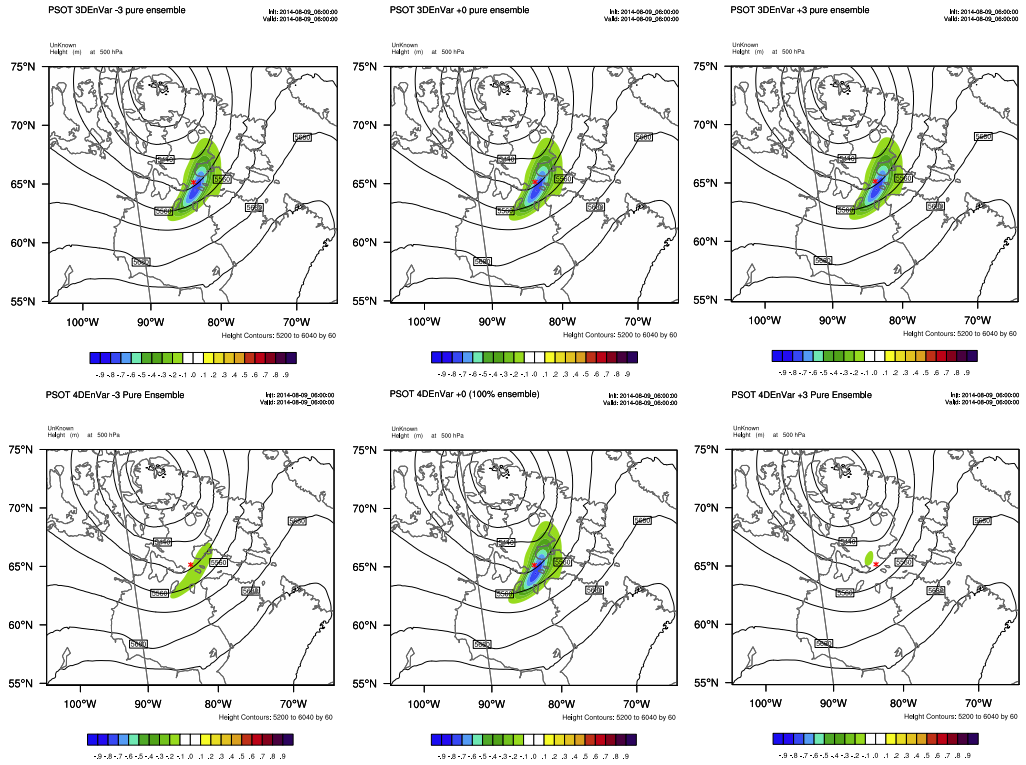

Figure 2.4. Same as Figure 2.3, except pure ensemble (100% ensemble, 0% static) PSOT temperature increments for 3DEnVar (top) and 4DEnVar (lower) methods.

Additionally, the utility of using a 1-hourly global ensemble was investigated for use in the RAP system. Figure 2.5 shows the 3 observation time bins for the -1, 0 and +1 hour time bins within a 2 hr time window. Although this region lacked strong flow dependent features, the focus is on the magnitude of the increment between the time bins. It is evident that the magnitude between the increments for the 1-hourly ensemble as compared to the 3-hourly ensemble shown in Figures 2.1-2.4 maintains a constant magnitude. This may be due to a more similar ensemble spread between the -1 to 1 time window, compared to the -3 to +3 time window.

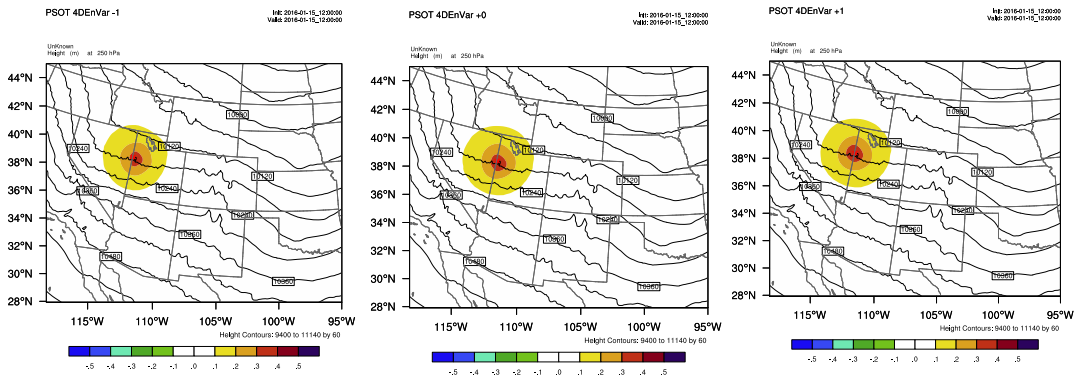

Figure 2.5. Same as Figure 2.4, except using a 1-hourly global ensemble with observation time bins of -1 (left), 0 (center), and +1 (right) within a 2 hour time window.

3. Cycling tests of the ARW and 4D EnVar GSI-hybrid data assimilation

For this part, 6-hourly ARW-GSI continuous cycling tests were conducted on the 13km RAP domain for the period of 2014080906-2014081600 to test the performance of the 4D EnVar hybrid capability. Three sets of experiments were conducted and are listed as follows:

- 3DVAR
- 3D EnVar GSI-hybrid, using the GFS 6-hour ensemble forecasts
- 4D EnVar GSI-hybrid, using the GFS 3-, 6- and 9-hour ensemble forecasts

For both 3D and 4D EnVar GSI-hybrid, the contributions from the static BE and the ensembles are 25% and 75%, respectively. The horizontal length scale for the GSI-hybrid is 110km and the vertical length scale is 3 (grid units). All the three experiments started from the RAP forecast initialized at 2014080900, which is used as the GSI background at 2014080906. The GSI analysis is then used as the initial condition for the 48-hour RAP WRF forecasts, in which the 6-hour forecast (3-, 6- and 9-hour forecast for 4D hybrid) is treated as the background for the next GSI analysis at 2014080912 and so on. Figure 3.1 shows the flow chart of the 4D EnVar hybrid cycling experiment, with the GFS ensemble being the 3-, 6- and 9-hour forecasts. For the experiments for 3DVAR and 3D EnVar GSI-hybrid, the flow chart is similar, with the only difference in the DA method. In all the three experiments, the assimilated observations include the Global Data Assimilation System (GDAS) GPS Radio Occultation, GDAS PrepBUFR data, GDAS radiance (AMSU-A, MHS, HIRS4). The RAP regional background errors are applied for the GSI analysis. The radiance bias correction coefficients from the previous RAP cycle was used for the GSI at 2014080906 and then cycled thereafter.

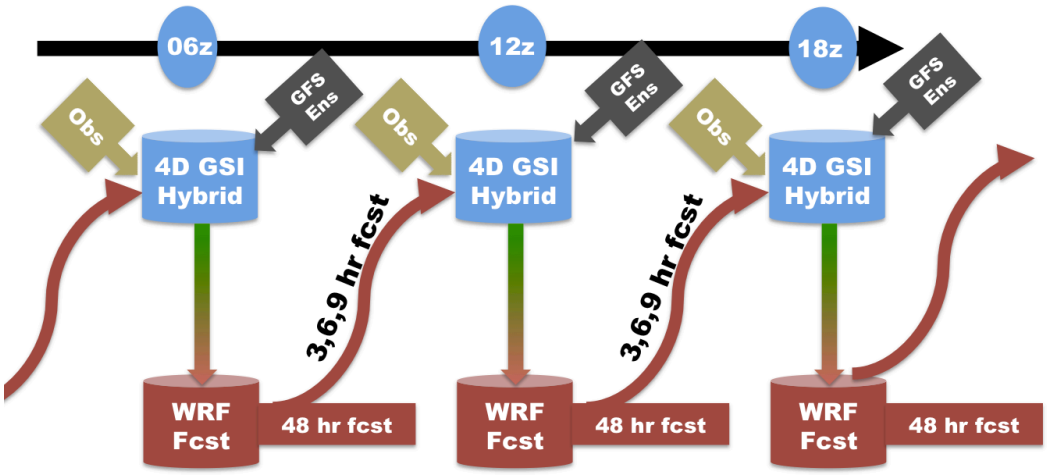

Figure 3.1. The flowchart of the 6-hourly cycling experiment using the 4D EnVar GSI Hybrid.

Before going into the cycling tests, some diagnostics are performed onto one single GSI analysis to compare the run-time features of the 4D EnVar hybrid against 3DVAR and 3D EnVar hybrid. Figure 3.2 shows the computation time for the GSI analysis at 2014080912 using 384 cores on NCAR supercomputer Yellowstone, in which the hybrid methods take more time than the 3DVAR method, and 4D hybrid takes about 1/3 more time than 3D hybrid, mainly due to the need to read more GFS ensembles. Figure 3.3 gives the contributions to the cost functions from different observation types for the three experiments for the 1st (left) and 2nd outer loop. As can be seen, the radiance data gives the largest contribution while moisture observations gives the smallest contribution. As the contributions vary between the observation types, the differences between the three experiments are small. This is reasonable since the numbers of used observations vary between different observation types for each experiment but the observation usage is similar between the three experiments for each observation type.

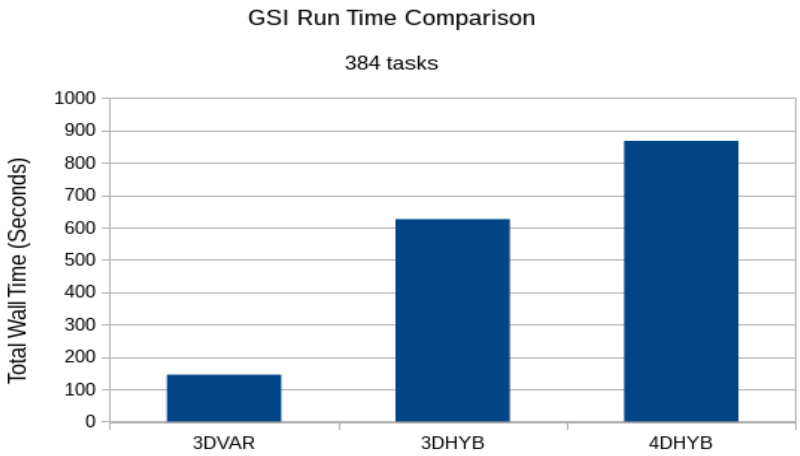

Figure 3.2. GSI run time comparison between the three experiments (3DVAR, 3D Hybrid and 4D hybrid) for the analysis at 2014080912.

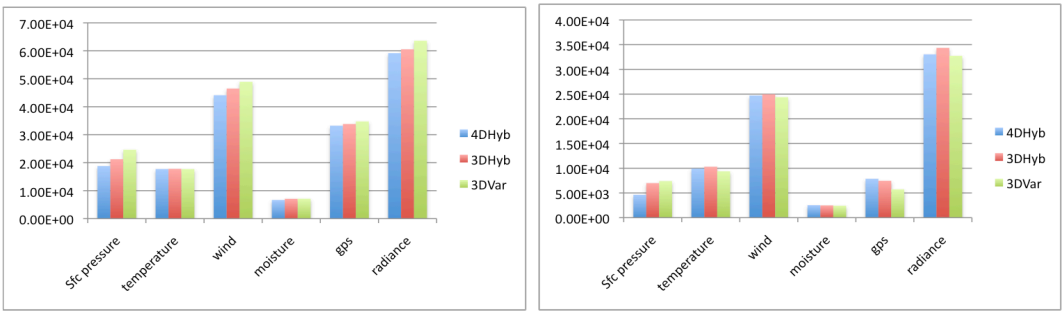

Figure 3.3. contribution to the cost function from different observation types for the first (left) and second (right) outer loop, for the three experiments at 2014080912.

As defined in the 4D hybrid, the observations are treated in more than one bin, in accordance with the time levels of the background files and the GFS ensemble forecasts. In this work, there are three time levels in the ensembles and background files (3-, 6- and 9-hour forecasts) and the observations are divided into three bins. Figure 3.4 gives the contributions from the three bins of each observation type for the 4D hybrid for the 1st (left) and 2nd (right) outer loops, in which the most contributions came from the bin 2, which corresponds to the center of the analysis time. This might be mainly due to the big observation density at the analysis time.

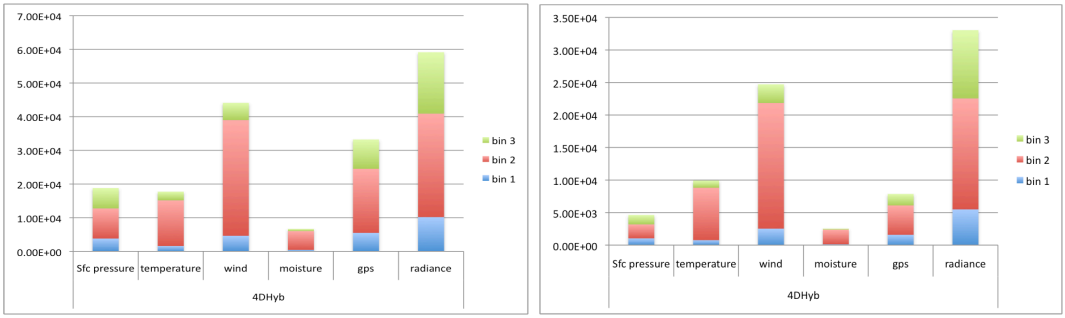

Figure 3.4. contributions to the cost function from different bins of different observation types for the first (left) and second (right) outer loop, for the 4D hybrid experiment at 2014080912.

Next we take a look at the aggregated statistics of the GSI background/analysis fit to the observations, O-B (observation minus background) and O-A (observation minus analysis), for the 28 cases over the period of 2014080906 to 2014081600, for the three experiments. Figure 3.5 gives the aggregated bias and rms of O-B (solid lines) and O-A (dashed lines) of the humidity field for the testing period. For both bias and rms, the values for O-A are reductions from O-B, suggesting the GSI analysis generally fit to the observations better than the GSI background for all levels. Looking at the rms of O-B, the 4D hybrid runs give the best fit of the background to the observations, while the background of the 3D hybrid runs gives better fit to the observations than the 3DVAR runs. Note that the GSI background in the three runs are actually the 6-hour forecasts from the previous GSI analysis. The GSI analysis fit to the observations for the 3D and 4D hybrid runs are very similar to each other, both better than the 3DVAR runs above 850hPa and worse than the 3DVAR runs below. Figure 3.6 gvies the time series of the rms of O-B (solid) and O-A (dashed) for the whole atmosphere and the total observation counts. As can be seen, there is a

clear diurnal cycle during the testing period, with rms values at 06Z and 18Z being smaller, and the observation counts for 06Z and 18Z being about 1/5 of that at 00Z and 12Z. And for 00Z and 12Z, the hybrid runs (red for 3D and black for 4D) have smaller rms than the 3DVAR runs (blue), and the difference tends to be larger than that at 06Z and 18Z, when the three experiments are very close to each other.

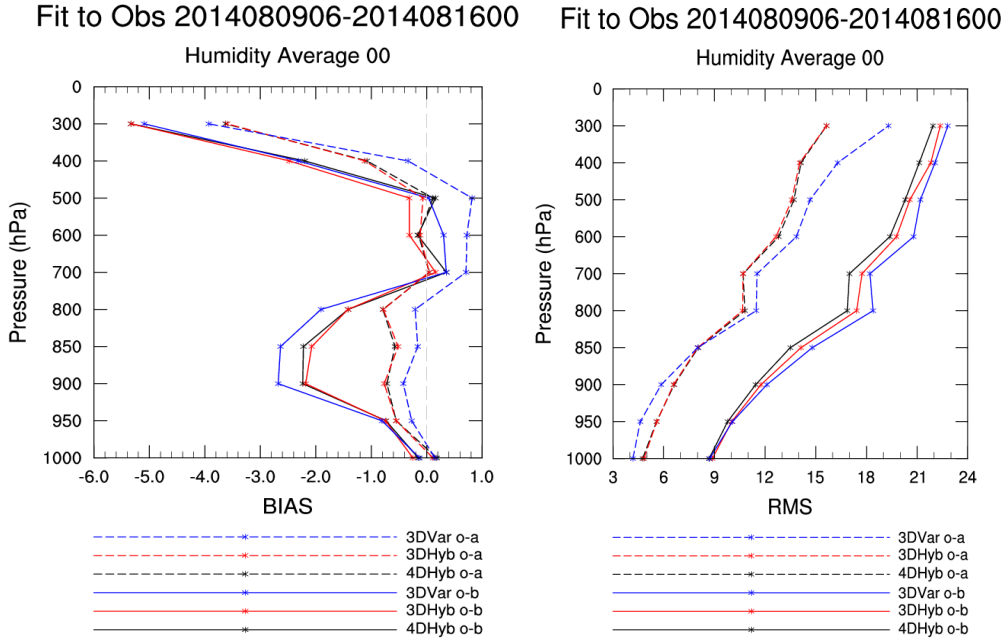

Figure 3.5. Vertical profiles of the aggregated bias and rms of the GSI background/analysis fit to the humidity observations, O-B (solid) and O-A (dashed), for the testing period of 2014080906 to 2014080916, for the 3DVAR (blue), 3D hybrid (red) and 4D hybrid (black) runs.

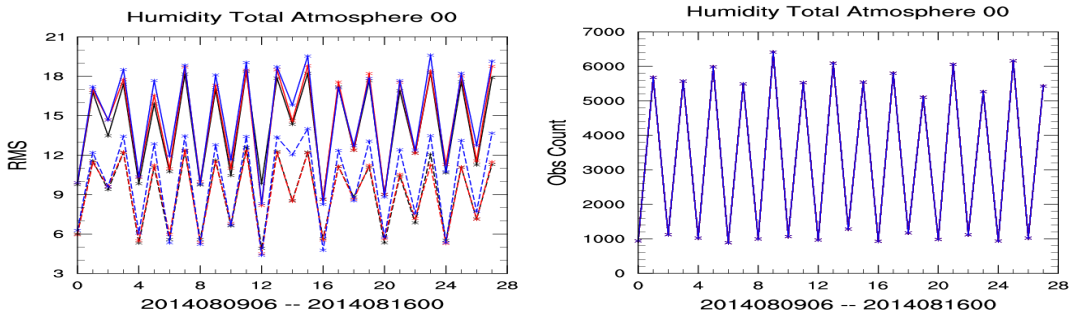

Figure 3.6. Time series of the rms (left panel) of O-B (solid lines) and O-A (dashed lines) and observation counts (right panel) for humidty for the three experiments: 3DVAR (blue), 3D hybrid (red) and 4D hybrid (black).

Figure 3.7 is similar to Figure 3.5, but for the temperature observations. Both the bias and rms for the temperature tend to be smaller for O-A than O-B, again suggesting the GSI analysis fits to the observations better than the background. It is hard to see which experiment is superior in terms of the bias but the rms from 4D hybrid is smaller than that of 3D hybrid. Figure 3.8 shows the time series of the rms of O-B (solid) and O-A (dashed) on the left, in which the 4D hybrid giving the smallest values among the three experiments, especially during the later part of the cycling tests. The right panel of Figure 3.8 shows the time series of the observation

counts, in which the number of temperature observations assimilated at 06 and 18Z is less than half of that at 00Z and 12Z.

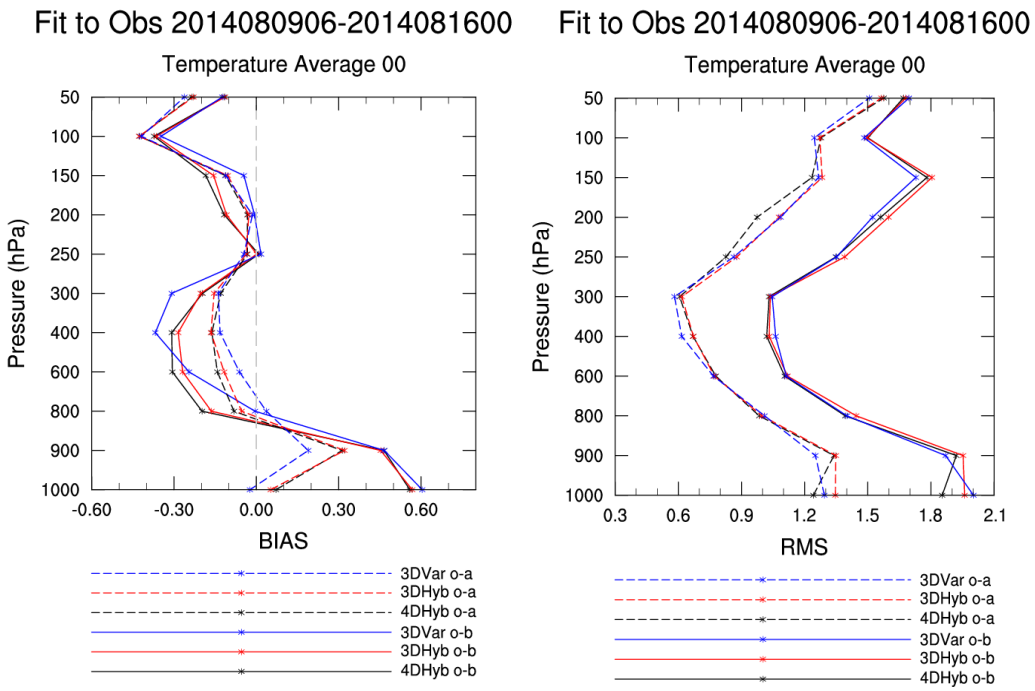

Figure 3.7. Similar to Figure 3.5, but for the temperature observations.

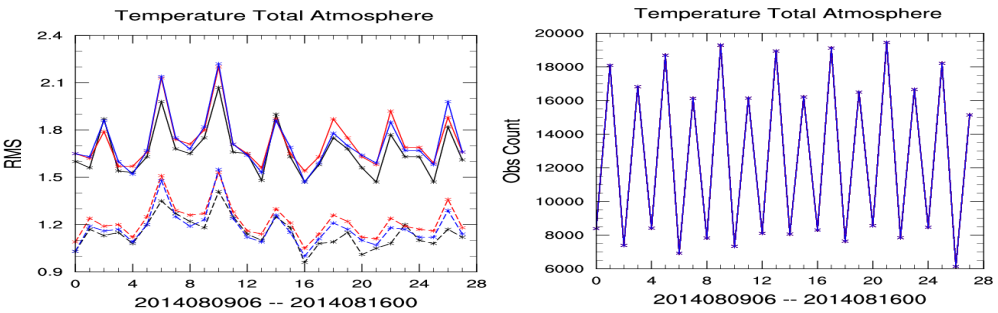

Figure 3.8. Similar to Figure 3.6, but for the temperature field.

Figure 3.9 gives the vertical profiles of O-B and O-A bias and rms for the wind field. The 3DVAR runs seems to produce smallest bias among the three experiments for the levels below 300hPa for both O-B and O-A. For the levels above, 4D hybrd runs have smallest O-B bias while for O-A the three experiments are very close above 300hPa. For the rms, the 4D hybrid gives smallest values for O-B except for the levels between 250hPa and 150hPa and O-A for all the levels below 100hPa. Figure 3.10 is simialr to Figure 3.8, giving the rms time series of the wind observations for the whole atmosphere on the left. For the O-B, consistent with the very large value at 200hPa for the 3D hybrid runs, the time series also show largest rms from 3D hybrid runs, mainly during the later part of the cycling tests. On the other hand, the O-A time series give smallest rms from the 4D hybrid runs. The rms in the 3D hybrid runs are largely reduced for the GSI analysis compared to the GSI background,

suggesting the improvements in the wind field analysis. The right panel of Figure 3.10 shows the time series of the observation counts. Different from the humidity and temperature fields in which there is barely any difference in terms of the observation counts between O-B and O-A, or among the three experiments, the observation counts for the wind O-A is larger than that of wind O-B, suggesting more wind observations being assimilated in the second outer loop. The difference between the three experiments is marginal and the difference between 06,18Z and 00, 12Z is not as big as that for the humidity and temperature fields.

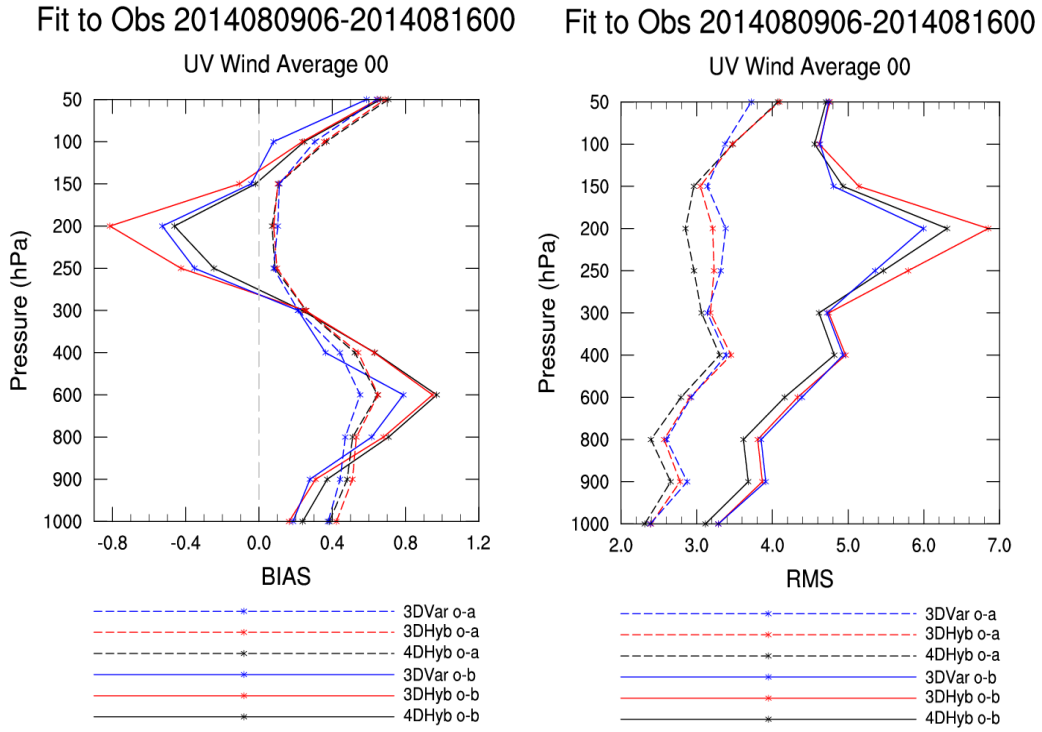

Figure 3.9. Similar to Figure 3.5, but for the wind observations.

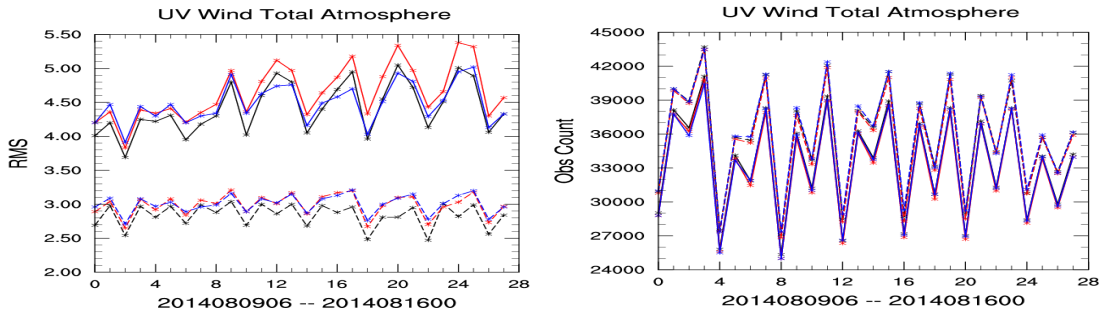

Figure 3.10. Simialr to Figure 3.8, but for the wind observations.

As suggested in the observation counts in Figure 3.6, 3.8 and 3.10, there tends to be less observations assimilated at 06Z and 18Z than at 00Z and 12Z. Figure 3.11 gives

the vertical profiles of O-B and O-A rms for the 00, 12Z group (left) and 06, 18Z group (right) for humidity (top), temperature (middle) and wind fields (bottom). As can be seen, the vertical profile of humidity and temperature O-B and O-A rms at 00, 12Z is very similar to the overall profiles in Figures 3.5 and 3.7, and different from that at 06, 18Z, suggesting the O-B and O-A statistics are mainly from 00Z and 12Z for these two fields. For the wind field, there is not much difference between the 06,18Z and 00,12Z statistics and it might be due to the observation counts being less different between the two groups.

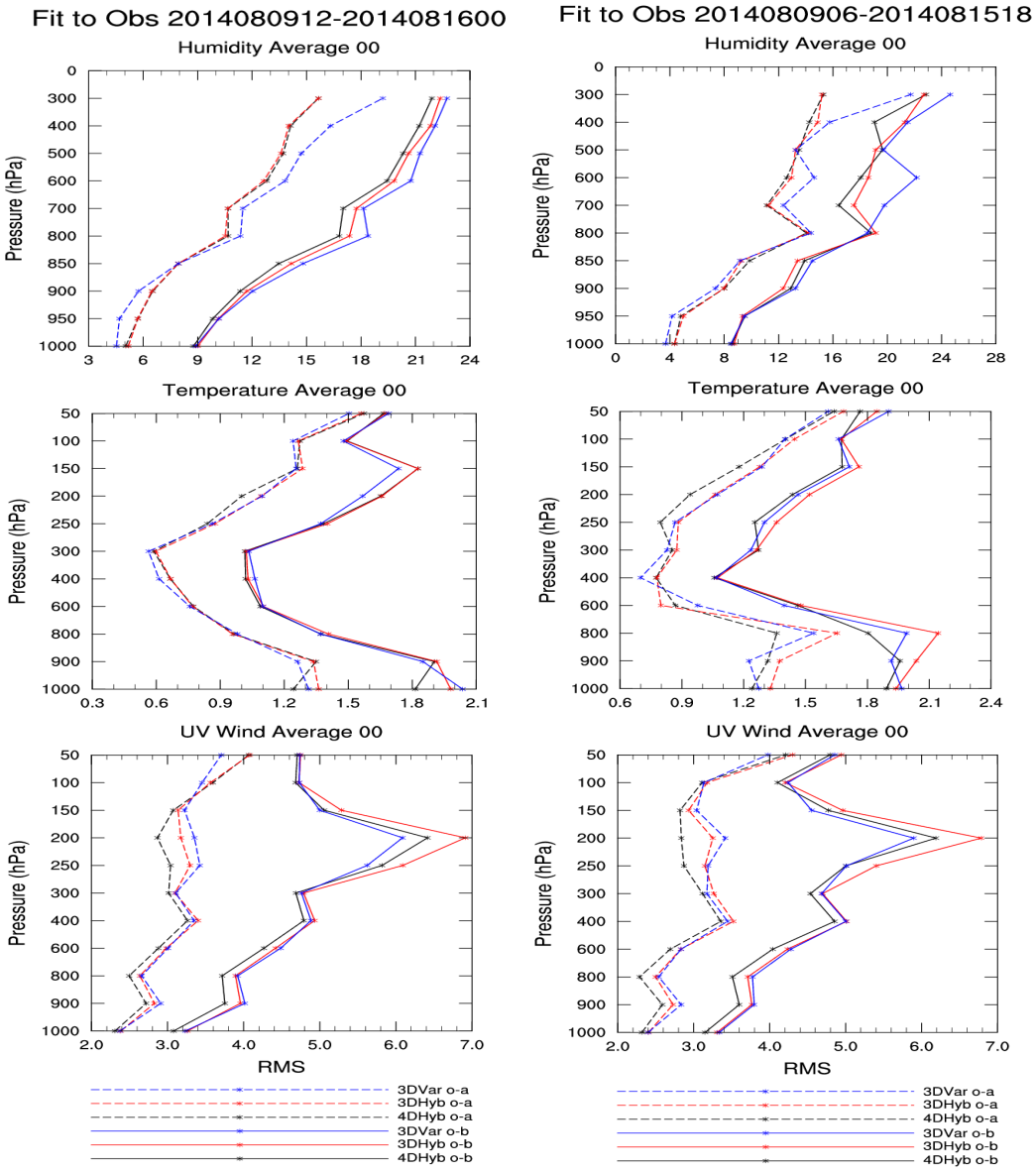

Figure 3.11. Vertical profiles of the aggregated O-B (solid) and O-A (dashed) rms for the GSI analysis at 00, 12Z (left panels) and 06, 18Z (right panels), for humidity (top), temperature (middle) and wind fields (bottom).

Figure 3.12 gives the vertical profiles of the humidity forecast rmse verified against the conventional observation using the GSI. At hour 6, the hybrid runs give smaller humidity rmse than the 3DVAR runs, altough it is hard to tell whether 3D or 4D hybrid is better than the other one. This rmse difference between the hybrid runs and the 3DVAR runs reduces with the forecast length, as can be seen for hour 12, 24 and until hour 36 it is hard to tell which one is superior.

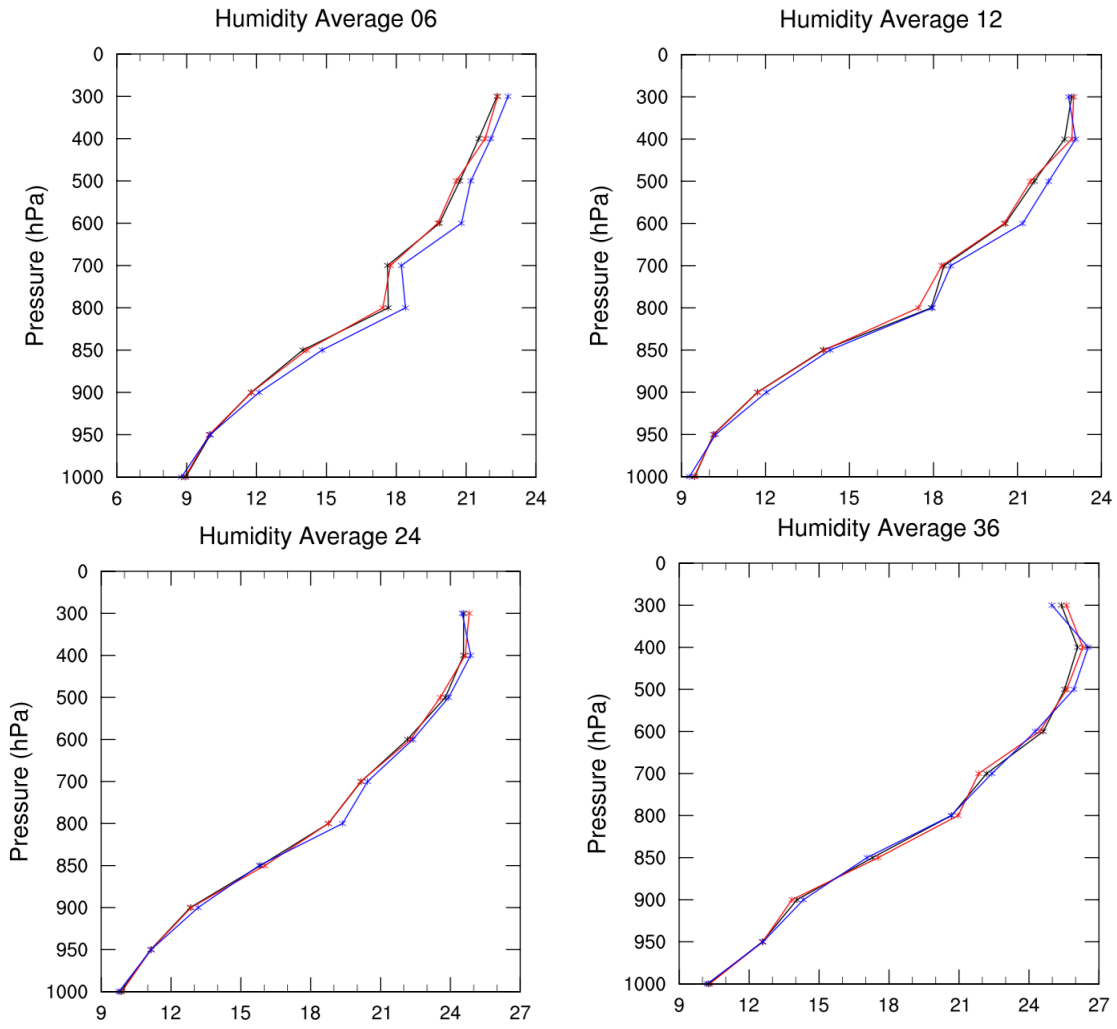

Figure 3.12. Vertical profiles of the humidity rmse at forecast hour 06, 12, 24 and 36 for the three experiments: 3DVAR (blue), 3D hybrid (red) and 4D hybrid (black).

The temperature and wind forecasts rmse verified against the conventional observations, however, don't suggest that the hybrid runs produce better forecast than the 3DVAR runs, nor does it tell whether 3D or 4D hybrid is better than the other one (figures not shown here).

Figure 3.13 shows the 24-hour accumulated precipitation from 2014080912 for the three experiments, compared to the stage IV precipitation record valid at

2014081012. It can be seen that all the three experiments captured the main rainfalls in Nebraska and between South Dakota and Minnesota but it hard to tell the differences between the three experiments.

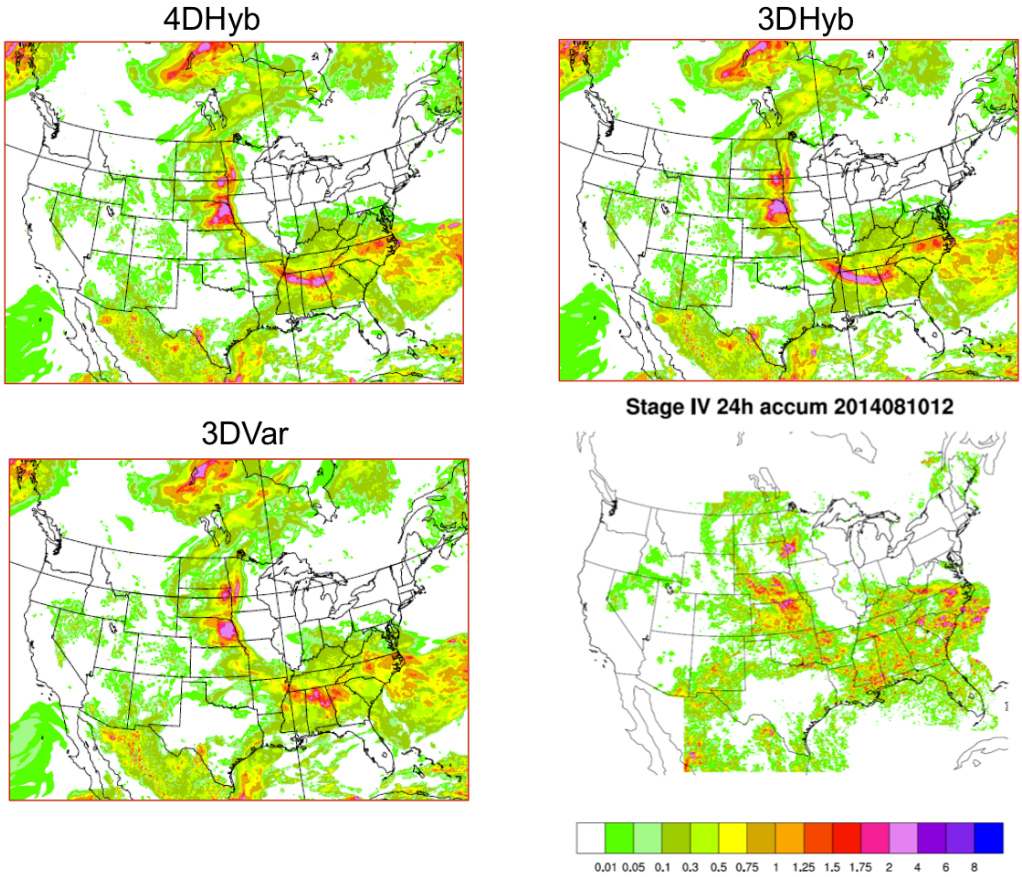

Figure 3.13. 24-hour accumulated precipitation from 2014080912 for the three experiments, compared to the stage IV 24h accumulated precipitation valid at 2014081012.

4. Test and evaluate RAP ensemble

Current RAP system is running GSI-hybrid with GFS EnKF 80 member ensemble. Using GSI-hybrid greatly improved upper air RAP forecast but gave little impact on surface, cloud, and precipitation forecast. This year's proposal is to run RAP ensemble and use them for the GSI-hybrid to see if the GSI-hybrid with the high resolution RAP regional ensemble can further improve the RAP forecast (especially surface, cloud, and precipitation).

4.1 Utility to downscale the GFS ensemble to generate the RAP ensemble initials

Because the GFS ensemble has been approved to have good ensemble perturbations in upper air through the RAP GSI hybrid, we think it is a good start to use the perturbations from the GFS ensemble forecast to initialize regional ensemble forecast. To this end, a downscaling tool was developed to calculate the GFS ensemble perturbations, interpolate the GFS perturbations to the RAP grid, and add

those perturbations to the RAP deterministic initial file to generate RAP ensemble initial files.

4.2 3-Day RAP ensembles

This utility was used to generate RAP ensemble forecast for 3 days from 18Z, August 08 to 00Z, August 12, 2014. Every 6-hour in those three days, 80 RAP initial ensemble members were generated from the 6-h GFS EnKF 80 member ensemble forecast and run for 12-h forecast with hourly output. This RAP ensemble forecast is called RAPens1X because the original GFS ensemble perturbation were used to initialize the ensemble forecast. Another same 3 days RAP ensembles were conducted initialed with the perturbations inflated twice larger than the original GFS perturbation values and this RAP ensemble is called RAPens2X.

For one 12-h 80 member RAP ensemble forecast, 1600 cores were used with 7 hour wall-clock time, and near 3T of disk space was used to save the results.

RAPens1X ensemble forecast initialized at 18Z, August 08, 2014 was checked below. Vertical profiles of horizontally averaged ensemble spread for T, Q, and U at every 3-h for the 12-h ensemble forecast are plotted in Figure 4.1. For T and Q, forecast spread is clearly larger than initial fields, especially for up level T field. But spread for T and Q at near surface level is reduced slightly during the forecast. This may indicate we need to figure out a way to increase low level spread in the future. The forecasted upper level U spread is larger than the initial field but the middle and low level forecast spread is smaller than initial ones except 3-h forecast.

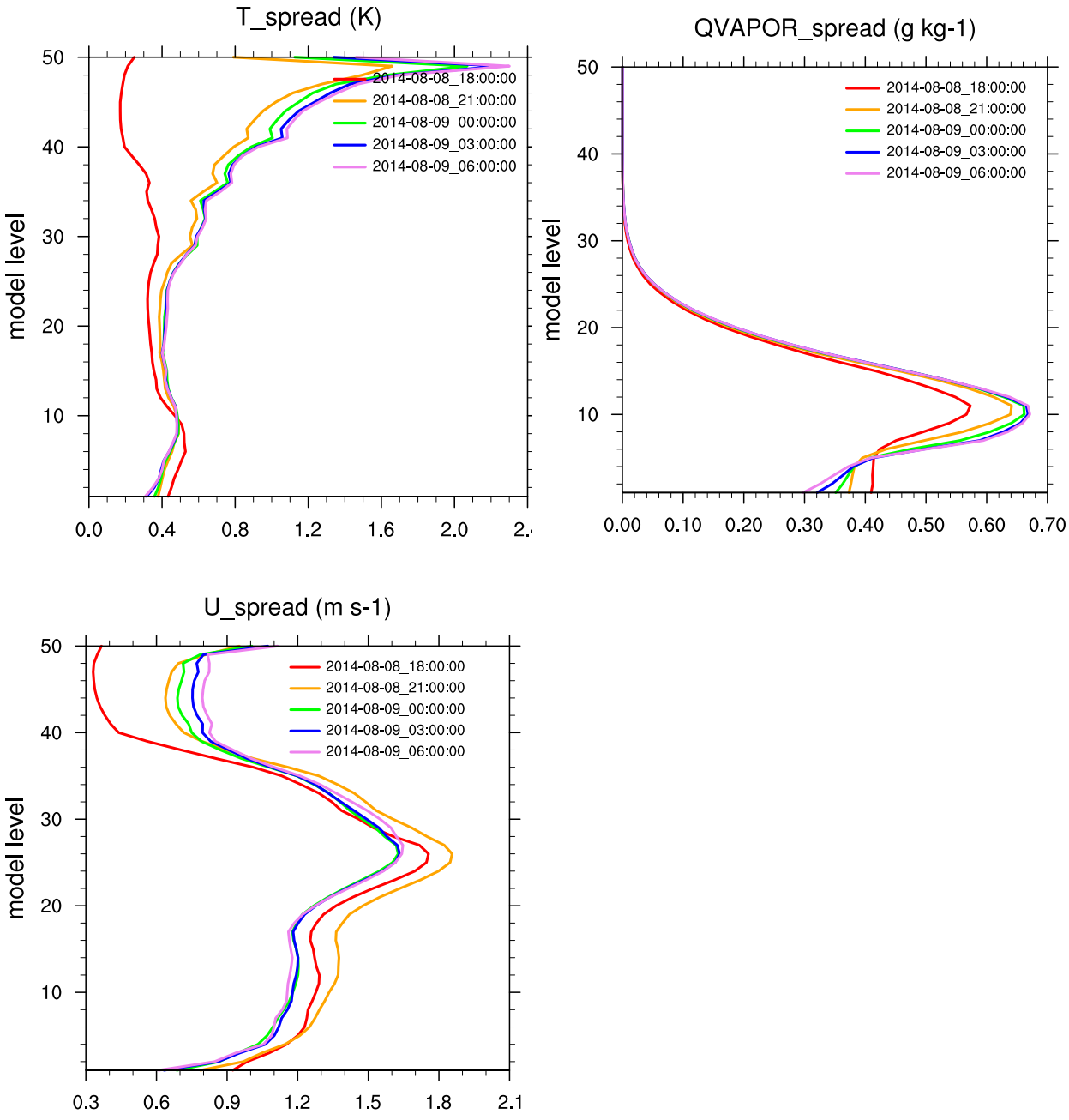

Figure 4.1. Vertical profile of horizontally-averaged ensemble spread for T, Q, and U, at the 0-h (18Z), 3-h (21Z), 6-h (00Z), 9-h (03Z), and 12-h (06Z) forecast lead times (valid times).

The ensemble spread at the 1st model level for moisture field at 0-h (12Z), 3-h (15Z), 6-h (18Z), 9-h (21Z), and 12-h (00Z) forecast lead times (valid times) are shown in Figure 4.2. We can see the clear patterns in the moisture spread field in Oklahoma and Texas and Florida, which may be related to the active convection later in those areas during this period.

The comparison of the ensemble spread from the RAPens1X and RAPens2X are shown in Figure 4.3. It can be seen that the ensemble spreads are clear larger across the domain from ensemble with inflated initial perturbations in temperature, moisture, and wind field after 9-h forecast. But the ensemble spread patterns associated with the weather system are similar in both RAP ensemble experiments. This indicates the enlarged initial ensemble perturbations are carried on well through the RAP model forecast but the larger perturbation does not change the weather system evolution in each member during the whole ensemble forecast.

4.3. RAP retrospective experiments

Four RAP retrospective experiments were conducted to test the impact of the RAP GSI hybrid with RAP ensemble (Table 4.1). When use regional ARW ensemble, the ensemble grid in the GSI-hybrid analysis has to be the same as ensemble forecast grid. So two experiments using GSI-hybrid with GFS ensemble were conducted: one using the ensemble grid with the same resolution as analysis grid (Cntl_GFSens_HiRes) and one using the operational low–resolution ensemble grid (Cntl_GFSens_LowRes). Two experiments using regional ensemble were conducted: one (retroRAPens1X) using GSI-hybrid with RAPens1X and one (retroRAPens2X) using GSI-hybrid with RAPens2X. The 7-12 hour RAP ensemble forecast were used for the GSI hybrid. Those GSI runs need about 400 cores and about 20 minutes to finish because of the high-resolution ensemble grid.

Table 4.1: RAP retrospective experiments for GSI hybrid with different ensemble options.

Experiment NAME	Ensemble For GSI hybrid	Ensemble Grid ratio	Analysis Grid ratio	GSI Ensemble perturbation grid
Cntl_GFSens_LowRes	GFS Hi-res	3	2	80 km
Cntl_GFSens_HiRes	GFS Hi-res	1	1	13 km
retroRAPens1X	RAPens1X	1	1	13 km
retroRAPens2X	RAPens2X	1	1	13 km

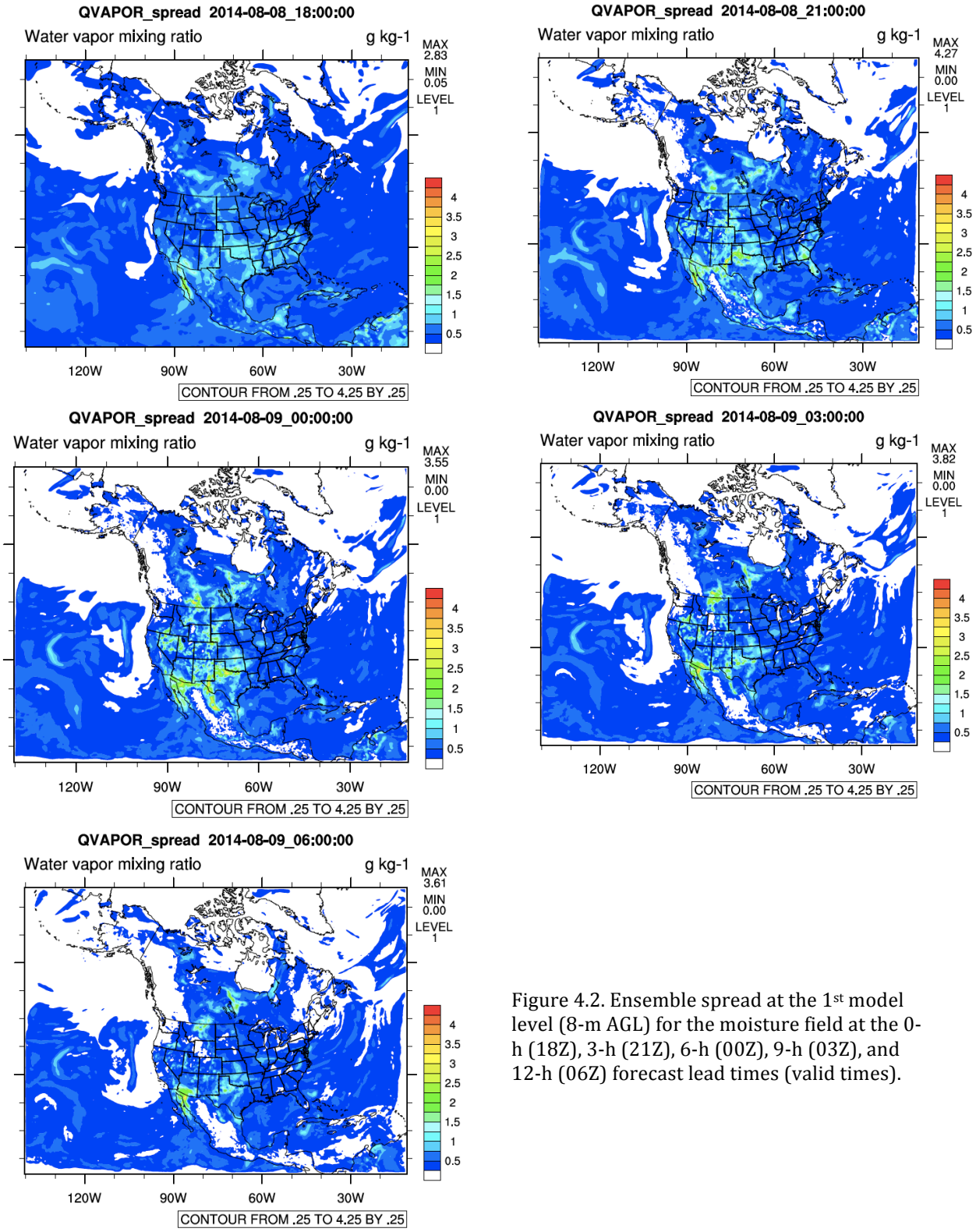

Figure 4.2. Ensemble spread at the 1st model level (8-m AGL) for the moisture field at the 0-h (18Z), 3-h (21Z), 6-h (00Z), 9-h (03Z), and 12-h (06Z) forecast lead times (valid times).

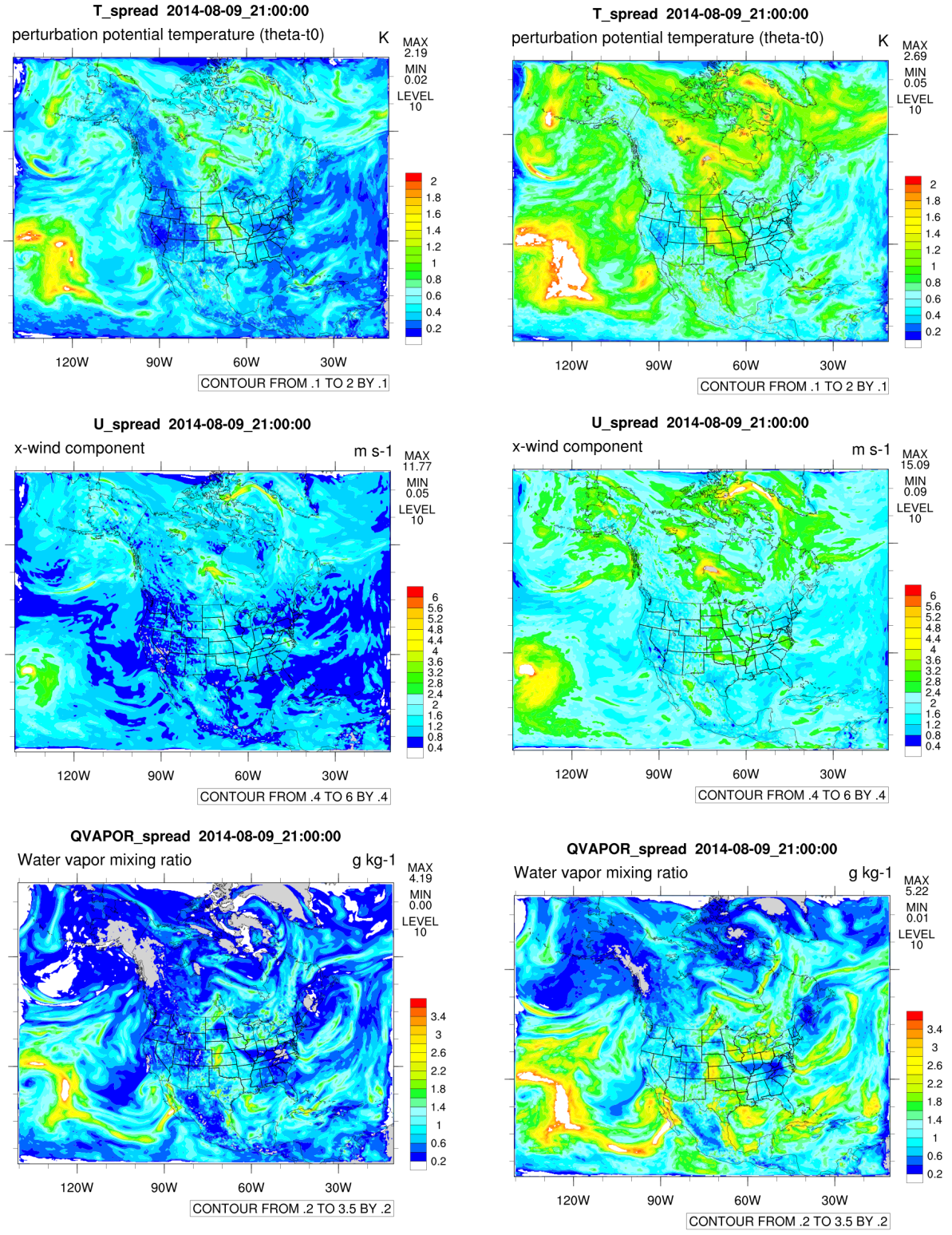

Figure 4.3. ensemble spread at level 10 (about 850 mb) for 9-h forecast initialized from 12Z August 9th, 2014. Left column is RAPens1X and right column is RAPens2X. The top row is potential temperature, the middle row is U-wind and bottom row is specific humidity.

The upper air and surface forecast error RMS time series from experiment Cntl_GFSens_lowRes (red line) and Cntl_GFSens_HiRes (blue line) are illustrated in Figures 4.4 and 4.5, respectively. We can see the blue line is consistently slightly smaller than red line, which indicate the high resolution ensemble grid are bring in more detailed information in the GSI analysis and can improve the analysis and forecast in both upper air and surface. But the high resolution significantly increased the GSI analysis wall time, which cannot be used in the operation. But we certain want to try high resolution ensemble grid if computer resource is not a problem.

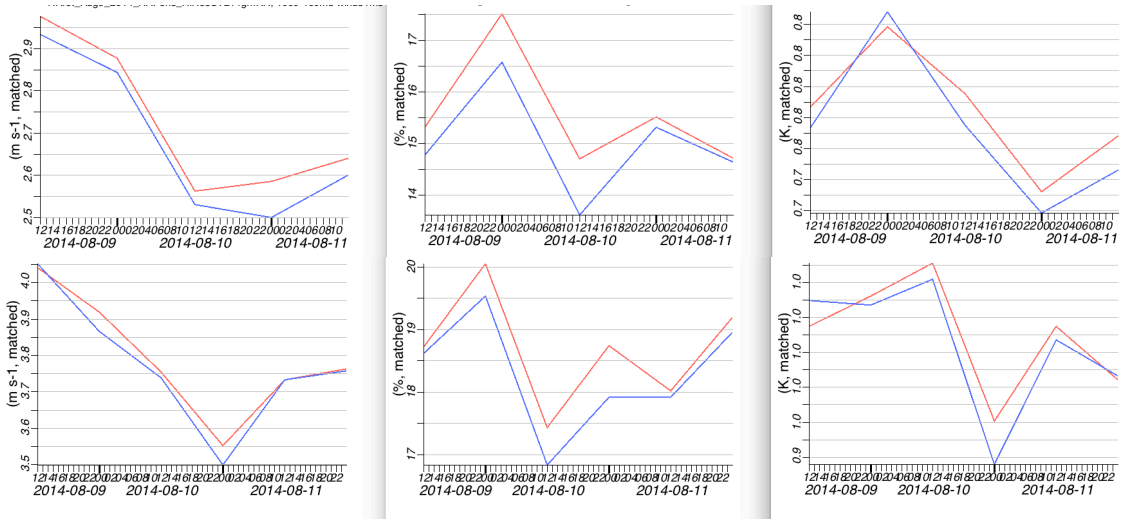

Figure 4.4: RMS time series of the forecast error against sounding observations for 0-h (upper row) and 3-h forecast (lower row). The left column is wind, the middle column is RH, and the right column is temperature.

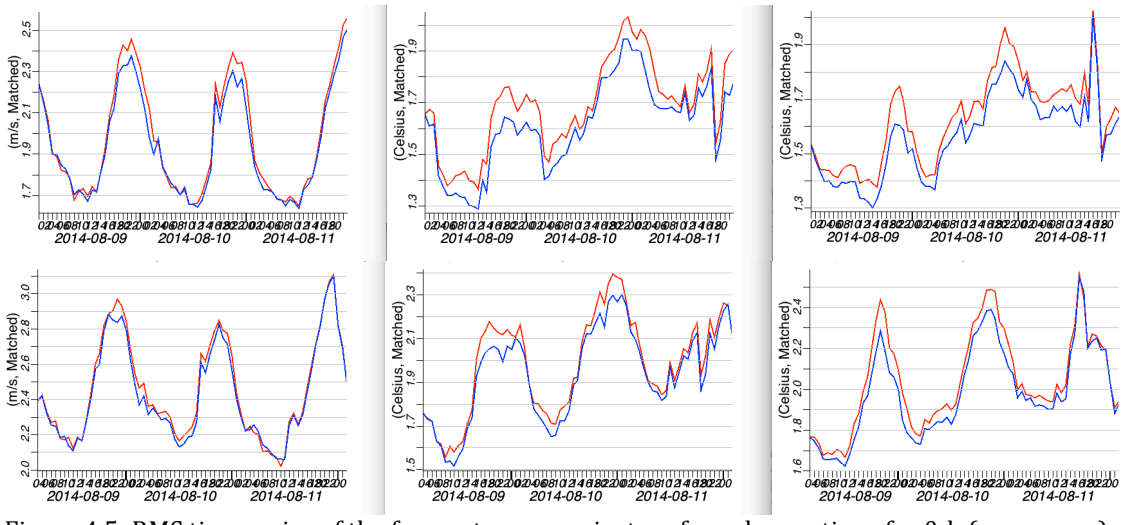

Figure 4.5: RMS time series of the forecast error against surface observations for 0-h (upper row) and 3-h forecast (lower row). The left column is 10-m wind, the middle column is 2-m dewpoint, and the right column is 2-m temperature

Figures 4.6 and 4.7 are like Figure 4.4 and Figure 4.5, but for the experiments Cntl_GFSens_HiRes (orange line), retroRAPens1X (red line), and retroRAPens2X (blue line). All three experiments use the same analysis ensemble grid but different ensemble forecast. Because the orange line and red line are almost identical, it illustrate the GSI hybrid with the RAP ensemble initialized from the original GFS perturbation are the same as the directly use of the GFS ensemble forecast. We can see the RAP model can successfully carry the GFS ensemble perturbation features through the 12-h forecast, although no additional details are added from GSI-hybrid analysis point of view.

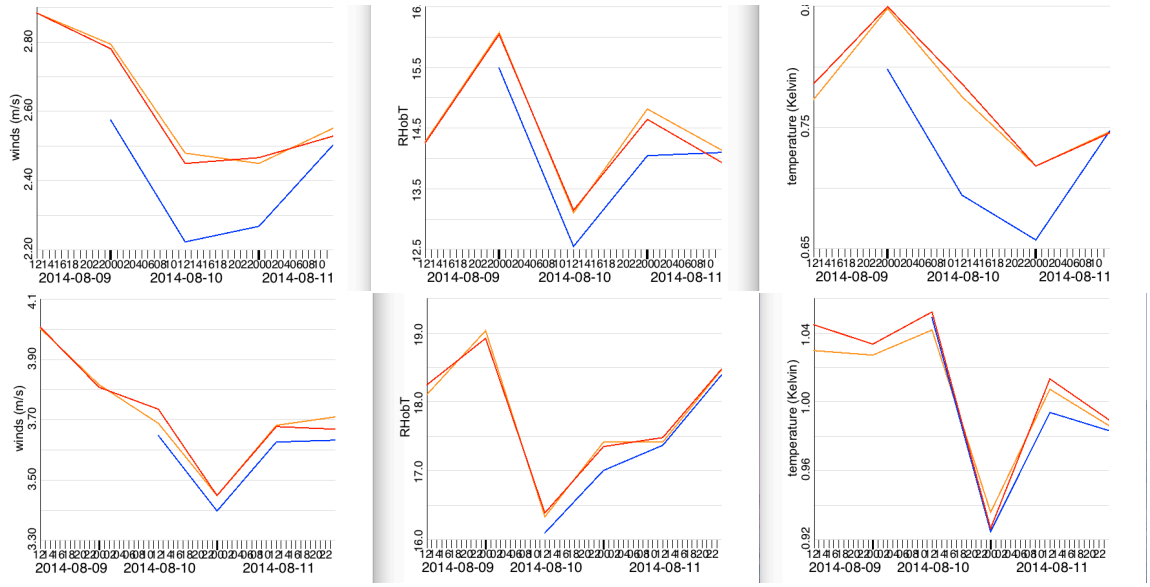

Figure 4.6: RMS time series of the forecast error against sounding observations for 0-h (upper row) and 3-h forecast (lower row). The left column is wind, the middle column is RH, and the right column is temperature. The Cntl_GFSens_HiRes is orange line, the retroRAPens1X is red line and retroRAPens2X is blue line.

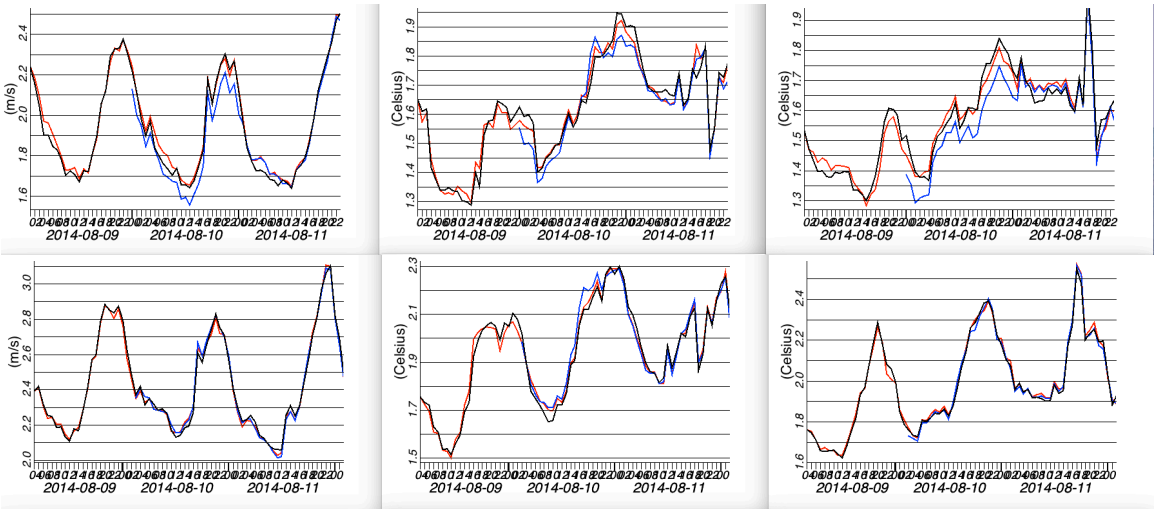

Figure 4.7: RMS time series of the forecast error against surface observations for 0-h (upper row) and 3-h forecast (lower row). The left column is 10-m wind, the middle column is 2-m dewpoint, and the right column is 2-m temperature. The Cntl_GFSens_HiRes is orange line, the retroRAPens1X is red line and retroRAPens2X is blue line.

Those experiments indicate we need to make RAP ensemble reflect more details of the local weather system by improving, the initial perturbation, perturbing model physics, and cycling the ensemble system. Our first effort is simplely double the GFS perturbation values when generate the RAP ensemble initial fields. That is the blue line in the Figures 4.6 and 4.7. Because the RAP ensemble RAPens2X does clearly increased the ensemble spread, we can see the analysis can fit to the observation slightly closer than the analysis use RAPens1X, but those close fit doesn't improve the forecast. The blue line (3-h forecast) mostly close to the other two lines, indicate the larger ensemble spread is not necessary improve the model forecast. More work needed in improving the RAP ensemble forecast.

5. Cycling tests of GSI 4D hybrid EnVar for RAP

The 4D hybrid EnVar has been developed in GSI and will be used in the next GFS operation upgrade with 7 hourly time bins from -3 to +3 hour of the analysis time. The RAP is hourly cycle, very different from the 6-hourly cycle system like GDAS. So we start some preliminary test and evaluation of using the GSI 4D hybrid EnVar in RAP hourly cycle. Because the previous RAP hybrid test using RAP regional ensemble didn't how the different from the analysis using the GFS ensemble, we decided to use the hourly GFS ensemble forecast, which is available in parallel testing in NCEP machine.

The 4D hybrid EnVar for RAP uses 3 time levels (3 hourly time bins). The background files use 0-h, 1-h, and 2-h forecast from previous cycle and the GFS ensembles use 7-h, 8-h, and 9-h forecast close to the RAP analysis time. The 8-h GFS ensemble matches the 1-h RAP forecast as analysis time background. Because the RAP partial cycles started from the GSI upper air, which has no hourly forecast, the RAP 4D hybrid EnVar test still use the 3D hybrid EnVar at 03Z and 15Z initial partial cycle. Several one day (Feb 11, 2016) retrospective experiments are conducted to evaluate the feasibility of the 4D hybrid EnVar for RAP system.

Figure 4.8 is RMS profile of the temperature forecast error against sounding observations for 0-h (left panel) and 3-h forecast (right panel) for experiments to debug the problems found in initial test of 4D hybrid EnVar for RAP. The blue line is the 1st experiment of the 4D hybrid EnVar for RAP but it is significant worse than the 3D hybrid EnVar (red line) in analysis and forecast for temperature and other fields (not shown). The big degradation of the analysis and forecast is in low level. A bug is found in the calculation of surface temperature observation operator using multi-time level 2-m background temperature. After bug fix (black line), the analysis and forecast are better but still much worse than the 3D hybrid EnVar. Additional test with the temperature terrain adjustment and T and Q PBL pseudo observations turned off (orange line) helped the analysis a little but doesn't solve the problem. When all the surface temperature observations are turned off, the analysis and forecast between 3D (red) and 4D (pink) hybrid EnVar are very close even though the 4D hybrid is slightly worse in the low level. We need to debug the 4D-hybrid EnVar to find why the surface temperature observations are making the trouble.

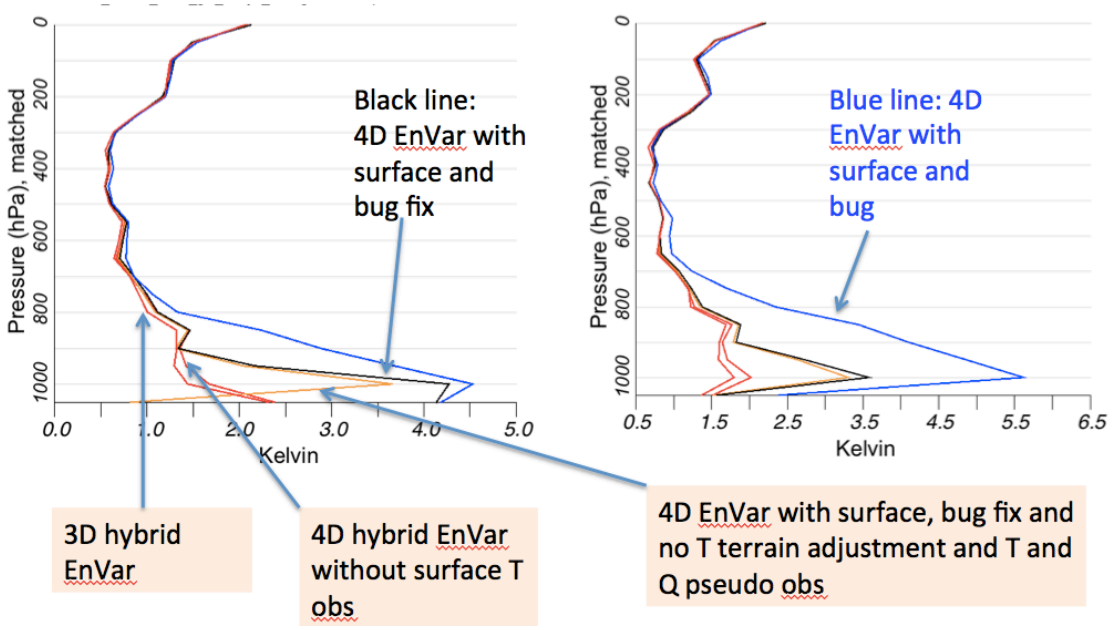

Figure 4.8. RMS profile of the temperature forecast error against sounding observations for 0-h (left panel) and 3-h forecast (right panel). The red line experiment with 3D hybrid EnVAR, the pink line is 4D hybrid EnVar without surface temperature observations, the orange line is 4D hybrid EnVar with surface temperature observations and bug fix and turned temperature terrain adjustment and T and Q PBL pseudo obs off, the black line is the 4D hybrid EnVar with surface observations and bug fix, the blue line is the 4D hybrid EnVar with surface observations and no bug fix.

The surface observations are very important observation types in the RAP system. Figure 4.9 showed the impact of the surface observations. When the surface observations are totally turned off in 3D hybrid EnVar, the RAP analysis and forecast

(blue line) at low level are clearly worse than the test using surface observation (red line). Again, the 4D EnVar (black) without the surface observations are close to the 3D-hybrid EnVar experiment without surface observations (blue line).

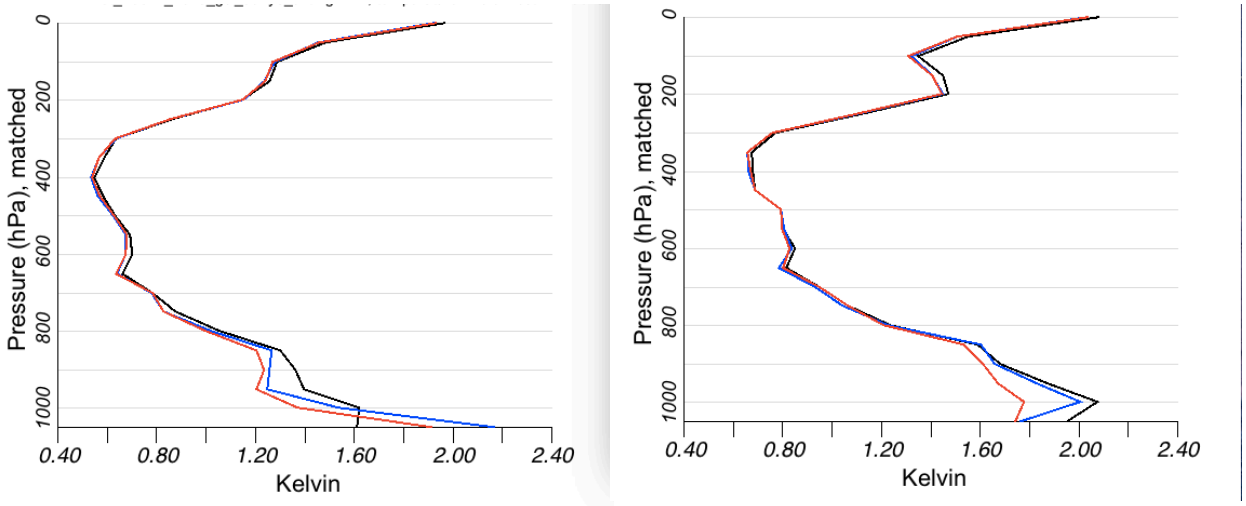

Figure 4.9. RMS profile of the temperature forecast error against sounding observations for 0-h (left panel) and 3-h forecast (right panel). The red line experiment with 3D hybrid EnVAR, the blue line is 3D hybrid EnVar without surface observations, the black line is the 4D hybrid EnVar without surface observations.