The Developmental Testbed Center (DTC) Annual Operational Plan (AOP) 2016 Project for the National Oceanic and Atmospheric Administration (NOAA) Office of Oceanic and Atmospheric Research (OAR)

Testing and Evaluation of the Hybrid 4D EnVar GSI for 3-km High Resolution Regional Applications

---

Final Report

8/31/2017

Table of Contents

Section I: Introduction
Section II: First objective - testing and evaluation of hourly update 4D EnVar data assimilation
Section III: Second objective - testing and evaluation of sub-hourly update 4D EnVar data assimilation

Section I: Introduction

1. Motivations

NCEP's transition from the GSI three-dimensional (3D) variational data assimilation (DA) system to the GSI-EnKF 3D hybrid DA system for its Global Forecast System (GFS) produced significant improvements in forecast skill. Following on the global application success, a number of regional applications have transitioned to one-way 3D hybrid ensemble-variational systems that use the GFS ensemble (e.g., NAM, HWRF, RAP). A number of operational centers (e.g., UK Met Office, Environment Canada) have implemented four-dimensional Ensemble-Variational (4D EnVar) DA toward further improved initial conditions for their global models. Such a DA system is slated to be included in NCEP's 2016 GFS implementation. The DTC Science Advisory Board (SAB) recommended the DTC continue to work with operations on 4D EnVar (including 4D hybrid EnVar), as this work is needed to support R2O transitions. Conducting EnVar DA tests is critical to DTC staff being able to make effective contributions to this effort. Such testing and evaluation (T&E) activities have the potential to provide critical insight into future operational applications at NCEP/EMC and NOAA/ESRL. Given the DTC's current focus on DA for regional applications and NCEP's goal to move toward high-resolution regional guidance, focusing its DA T&E activities on convective-scale NWP and ensembles provides the best alignment with its operational partners.

2. Project Description

The DTC proposed to test and evaluate 4D hybrid EnVar capabilities for high-resolution regional DA. This work began with a benchmark system based on the operational HRRR system, which uses global ensembles (~30 km) for its hybrid DA system and one-hour ARW forecasts from interpolated RAP analyses (13 km) for its 3-km initial conditions.

The first objective of this activity is to demonstrate whether the 4D hybrid EnVar system improves upon the performance of the benchmark HRRR system. The DTC set up an experimental 4D hybrid EnVar system, including setting up a workflow based on the operational NOAA HRRR system. Then the DTC tested hourly cycled 4D hybrid EnVar for comparison to the control runs using the 3d hybrid system with current HRRR DA configuration. The results generated from this effort are presented and discussed in Section 2. NCAR node of the DTC took the lead on the first objective and authored this part of report.

The second objective of this activity is to estimate feasibility and then impacts of fast cycling of 4D hybrid EnVar. Based on the above-mentioned system and experiments, the DTC increased the analysis update frequency from hourly to sub-hourly. The DTC collected and adapted appropriate observation and background files for such a fast cycling system and performed experiments to evaluate its performance. The impacts of increasing update frequency on analysis and forecasts are presented and discussed in

Section 3. GSD node of the DTC took the lead on the second objective and authored this second part of report.

Although having separate responsibilities, during this task, NCAR and GSD discussed results interactively and help improve the performance of the system in a collaborative effort.

Section II: First Objective - Testing and Ealuation of Hourly Update 4D EnVar Data Assimilation

1. Background and Setup

As a number of operational centers have implemented four-dimensional Ensemble-Variational (4D EnVar) Data Assimilation (DA) for their global models, the Developmental Testbed Center (DTC) continues to conduct testing and evaluation of the GSI (Gridpoint Statistical Interpolation) 4D hybrid EnVar system for regional 3-km High Resolution Rapid Refresh (HRRR), as part of the efforts to improve the convective scale and cloud resolving numerical weather predictions at the National Oceanic and Atmospheric Administration Earth System Research Laboratory (NOAA/ESRL).

Due to computational constraint, the operational HRRR 3km domain has been reduced from the CONUS (Continental US) to the central US. The period for the testing and evaluation is set to be September 3-9 of 2016, with hourly update of the initial and boundary conditions from the retrospective 13-km NOAA Rapid Refresh (RAP) runs. In addition to the experiments with the hybrid GSI three-dimensional EnVar (3D EnVar) as in the operational HRRR configurations, the hybrid 4D EnVar is applied to the reduced HRRR domain to investigate whether the 4D hybrid EnVar system improves upon the performance of the benchmark HRRR. Two-hours pre-forecast runs from each cycle are conducted to provide GSI backgrounds at three time levels (t-1hr, t and t+1hr) to match the observations in three time bins in 4DEnVar. The 80-member global ensemble forecasts at three time levels provide time-variant, flow dependent background errors to the GSI in addition to the climatology background errors. Preliminary results suggest that the hybrid 4D EnVar GSI analysis gives better fit between the observations and the HRRR background than the 3D EnVar in wind and humidity but limited impact on the forecasts. Case studies with hourly cycled data assimilation show accumulated positive impact in forecasting the fast evolving convective systems in both 3DEnVar and 4DEnVar. And cycling the hybrid 4DEnVar also brings out more benefit from the time-variant and flow-dependent background information, as compared to the flow-dependent but not time-variant information in 3DEnVar. Additional experiment with regional 3-km ensemble forecasts in the hybrid 3D EnVar as compared to the 3D EnVar using the low-resolution GFS ensemble suggests the potential benefit of applying high resolution regional ensembles in the 4D EnVar for better background error representative and therefore better forecasts.

2. Single Observation Tests

Before conducting real case experiments, a series of pseudo single observation tests (PSOT) were conducted to check the machinery of the GSI code. These single

observation tests were conducted for 06z August 9, 2014 on the 3-km domain around the Hudson Bay, where there is a mid-latitude trough. 80 member global ensemble forecasts were used in the 3D hybrid EnVar and 4D hybrid EnVar data assimilation, in addition to the 3DVar tests. 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 grids in the vertical.

The resulting increment in each observation time window was examined to better understand how the 4DEnVar capability was applied and ensure that the algorithm was working properly for high-resolution regional applications (Figure 1). Both 3DEnVar and 4DEnVar tests show a flow-dependent response to the temperature increment, compared to the isotropic response in the 3DVar. The observation impact for the 3DEnVar case does not change regardless of the observation time difference. The impact of utilizing the 4DEnVar methodology is evident with the +/- 3 hour time bins, where the observational impact is not as strong as the observations right at the analysis time, giving the analysis increment time-variant information. The above single observation tests confirm that the 3DEnVar and 4DEnVar methodology are working properly for the regional applications and therefore we can proceed to the real data assimilation and examine their impact on the regional forecasts.

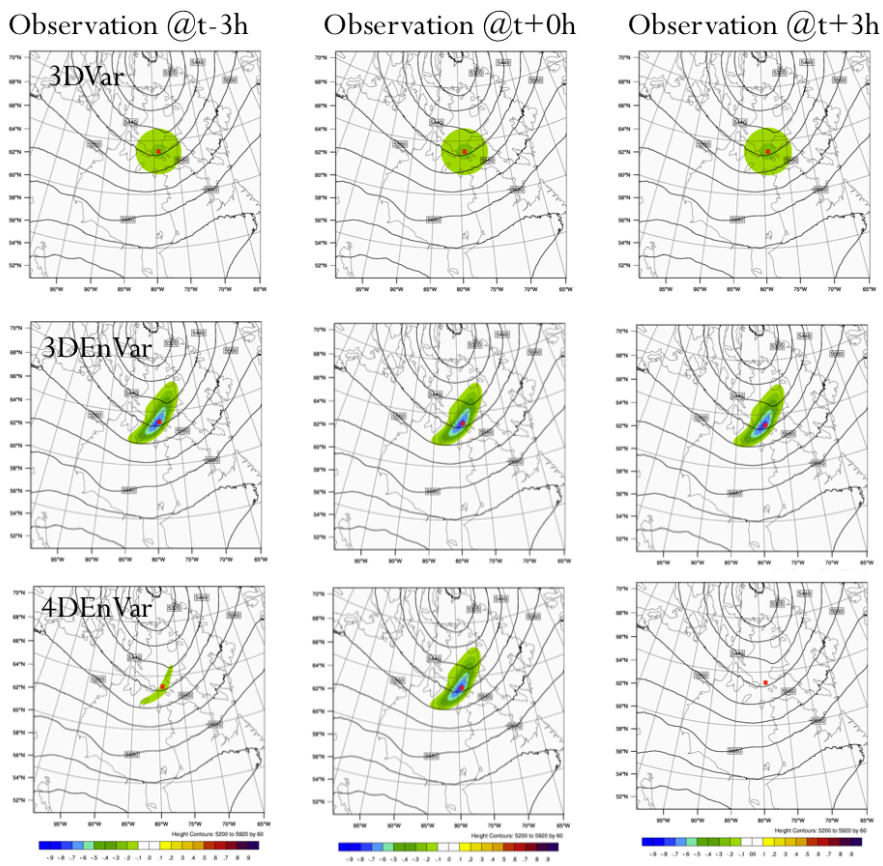

Figure 1. 500 hPa temperature increment (shaded, K) resulting from a single temperature observation test (location indicated by red dot) and the background geopotential height (contours, gm) at the analysis time. The left, middle and right columns correspond to the single observation at t-3h, t-0h and t+3h. The top, middle and bottom rows correspond to 3DVar, 3DEnVar and 4DEnVar.

3. High-Resolution 3km HRRR Experiments

3.1 Experimental Design

The goal of this task is to build and, then, test the latest hybrid four-dimensional ensemble-variational (4DEnVar) Data Assimilation technique for operational NOAA High Resolution Rapid Refresh (HRRR) applications. For that purpose, the High Resolution Rapid Refresh model system is used, which is a NOAA real-time 3-km resolution, hourly updated, cloud-resolving, convection-allowing atmospheric model, initialized by 3-km grids with 3-km radar assimilation. The initial and boundary conditions come from the 13-km Rapid Refresh (RAP) system of NOAA. GSI hybrid 3DEnVar is used for the operational data assimilation. In operational configuration of HRRR, radar data is assimilated every 15 min over a 1-hour period adding further details to that provided by the hourly data assimilation from the 13-km radar-enhanced Rapid Refresh (see Figure 2 for the RAP/HRRR workflow). In this work, a reduced HRRR domain of the central United States is utilized due to the

computational constraints, while other model configurations (model resolution, physics package, initial and boundary conditions and etc) remain the same as the operational. Another caveat is that the radar data assimilation (nudging) for the 1-hour pre-forecast period is removed to isolate the pure impact from the EnVar data assimilation methods.

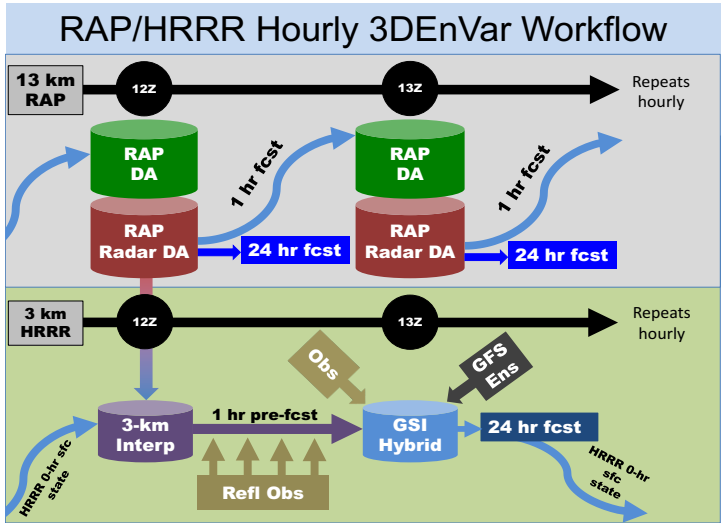

Figure 2. Workflow of the operational RAP/HRRR.

Two sets of experiments are conducted for the period of September 3-9 2016, based on the two hybrid data assimilation schemes: 3DEnVar and 4DEnVar, both applying the 80 member GFS ensemble forecasts at around 30-km horizontal resolution and assimilating the archived RAP real-time observations.

- **3D**: 3D hybrid EnVar (operational configuration), control experiment
- **4D**: 4D hybrid EnVar

For both hybrid 3D and 4D EnVar, the contributions from the static BE and the ensembles are 25% and 75%, respectively. The horizontal length scale for the GSI hybrid is 110 km and the vertical length scale is 3 (grid units). The assimilation time window is set to be +/- 1 hour from the analysis time. All the experiments are initialized from the retrospective RAP forecast and then 2-hour pre-forecast were performed to provide background fields for the GSI data assimilation. For 3DEnVar, the 1-hour HRRR forecast field from 1 hour earlier is the background field to be updated through the data assimilation, and then used to initialize the 18-hour WRF forecast. For 4DEnVar, besides the 1-hour HRRR forecast, 0-hour and 2-hour HRRR forecasts from 1 hour earlier are also needed in order to provide backgrounds for the observations of three time bins (t-1, t, t+1, relative to the analysis time).

3.2 3DEnVar and 4DEnVar Experiments

Firstly, a series of GSI 4DEnVar tests with a single observation prescribed at different observation times were conducted, which revealed that the three

observation bins are not evenly distributed in the 2-hour assimilation window (not shown here). Within the 2-hour time window, observation bin 1 corresponds to t-1 hour to t-0.5 hour, bin 2 corresponds to t-0.5 hour to t+0.5 hour, and bin 3 corresponds to t+0.5 hour and after. The RAP real-time observations are mainly between -1 to 0.5 hours from the analysis time. Apparently the central bin has the most of the observations, with fewer observations for bin 1 and barely any observations available for bin 3. This is confirmed by Table 1, which shows the observational contribution to the cost function from the three time bins in one 3DEnVar and 4DEnVar assimilation. As can be seen, the observational contribution to the cost function from 3DEnVar has only one value for each variable and they tend to be bigger than that from 4DEnVar. For 4DEnVar, the central bin tends to dominate the contributions except for the precipitable water, among the three bins. This uneven observational distribution would potentially limit the performance of 4DEnVar, as compared to the ideal situation of three evenly distributed observation bins.

Table 1. Contributions to the cost function from different observations

observations	**3denvar**	**4denvar**		
		Bin 1	Bin 2	Bin 3
Surface pressure	5.370E+03	0	2.84E+03	0
temperature	2.425E+03	2.93E+02	1.63E+03	0
wind	6.973E+03	1.11E+03	4.32E+03	0
moisture	1.619E+03	1.07	1.94E+03	0
Precipitable water	4.407E+01	1.48E+01	0	0

Hourly-updated runs were conducted for both 3DEnVar and 4DEnVar and there were more than 150 runs for each experiment for the testing period. A first look at the results is to see how the analysis fields fit to the observations. Figure 3 gives the domain averaged background and analysis fit to wind observations. The left panel shows the time series of the root mean squared (RMS) difference between the background/analysis and the observations, in which the analysis of both 3DEnVar and 4DEnVar fits to the wind observations better than the background, suggesting the GSI analysis improves upon the GSI background fields. For both background and analysis, 4DEnVar fits better than the 3DEnVar through the testing period. The right panel gives the vertical profile of the fit to the wind. Again, the analysis fits to the wind observations better than the background fields for all levels and the 4DEnVar gives better fit than the 3DEnVar, for both the background and analysis. Figure 4 is similar to Figure 3 but for humidity field. Similar features are shown for the

humidity, with 4DEnVar background and analysis fields fit better than the 3DEnVar, for both time series of the total atmosphere and vertical profiles. Figure 5 gives the background and analysis fit statistics for the temperature field, suggesting an overall better fit in 3DEnVar for the total atmosphere and worse fit in 4DEnVar for levels below 400hPa.

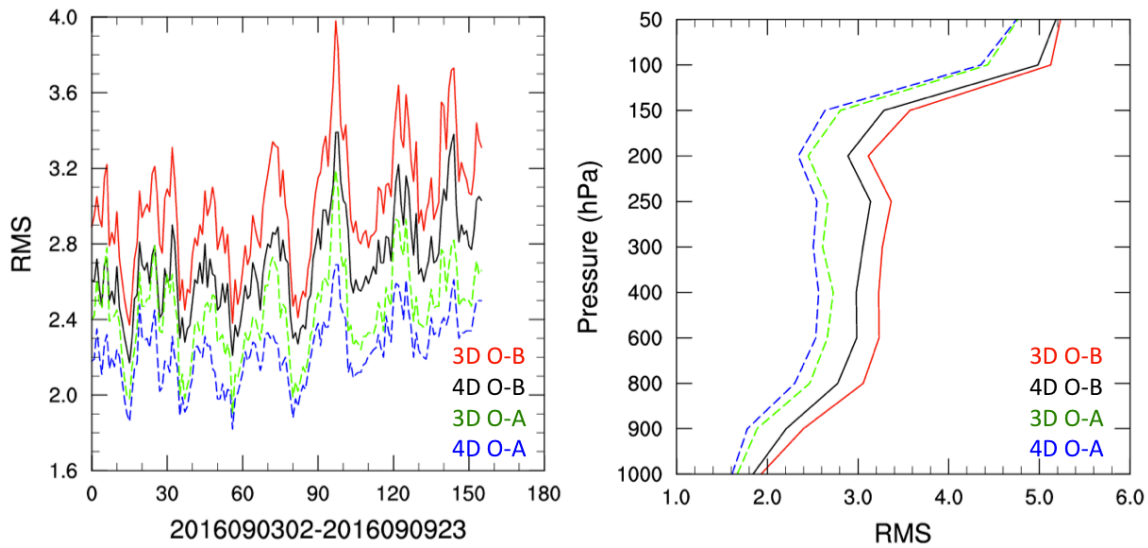

Figure 3. Domain averaged RMS of the background and analysis fit to wind observations (m/s) for both 3DEnVar (red: background; green: analysis) and 4DEnVar (black: background; blue: analysis). The left panel gives the time series of the fit and the right panel gives the vertical profiles.

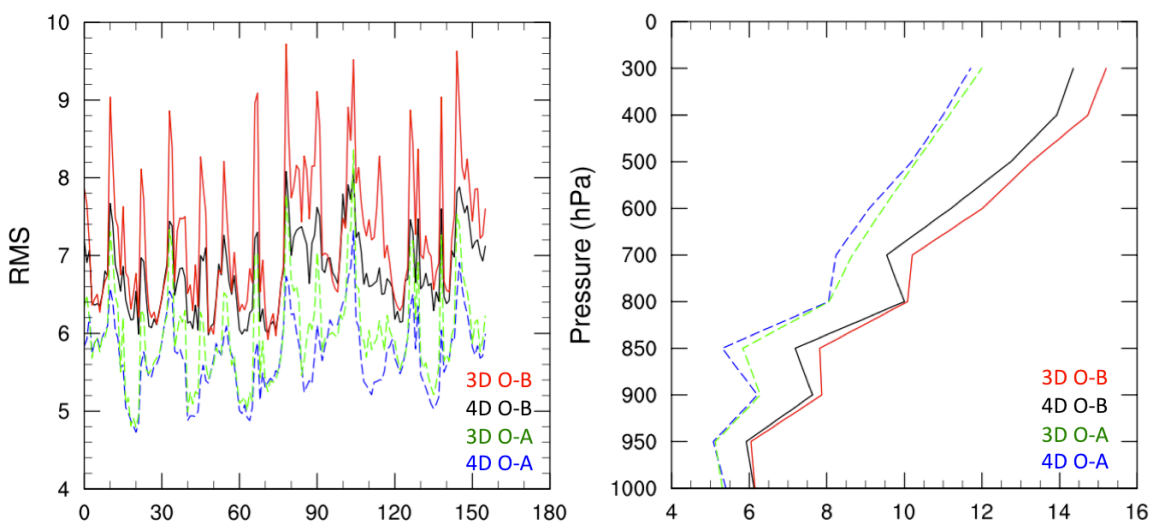

Figure 4. Similar to Figure 3, but for humidity (unit: percent of guess q_{sat}).

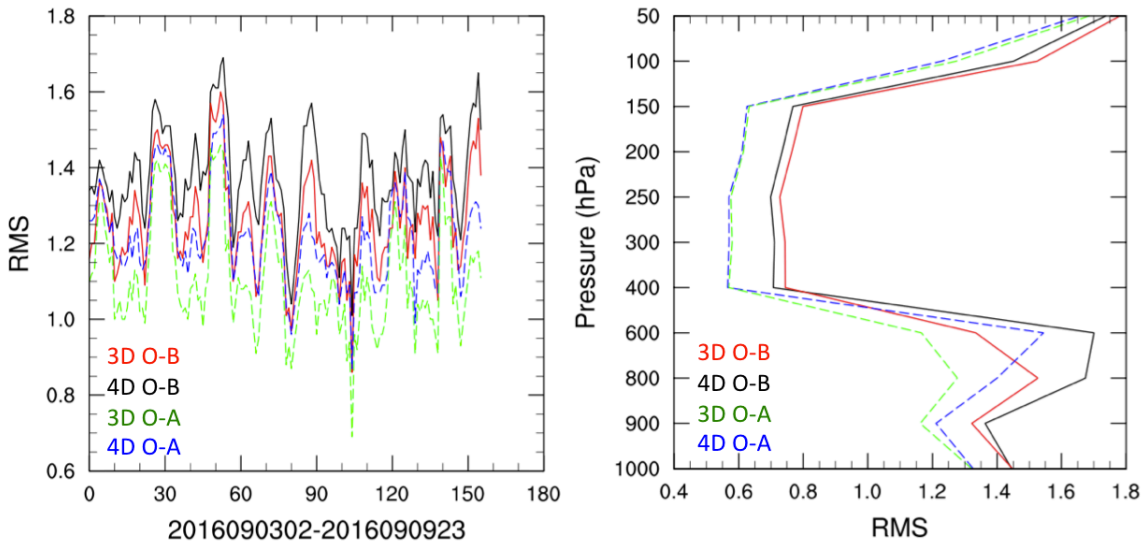

Figure 5. Similar to Figure 3, but for temperature (unit: K).

Next we look at the forecasts from the two sets of experiments. The forecast verification against the humidity observations at 00, 03, 06 and 09 hours is presented in Figure 6, with 3DEnVar being red and 4DEnVar being black. At 0 hour forecast time, the RMS of humidity in the 4DEnVar experiments is comparable or smaller than that in 3DEnVar for levels 500hPa and below. Compared to the background and analysis humidity fit to the observations as in Figure 4, the benefit of applying 4DEnVar is marginal after the WRF forecast is initialized. By forecast hours 3 and after, the difference in the humidity RMS is negligible. This suggests that regardless of the better fit of the analysis to the observations, it didn't persist for the later forecast in terms of the overall statistics.

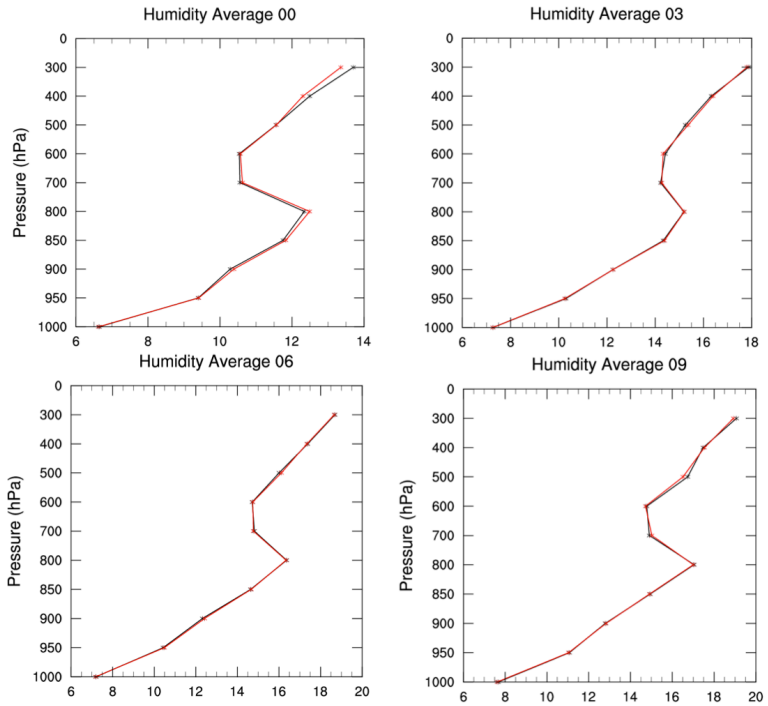

Figure 6. Vertical profiles of the 3DEnVar (red) and 4DEnVar (black) humidity fit to the observations for 00, 03, 06 and 09 forecasts.

Figure 7 shows the column maximum radar reflectivity from the observation (top), the analysis in 4DEnVar (bottom left) and the difference between 4DEnVar and 3DEnVar (bottom right) at 12Z September 8 2016. As can be seen, there is barely any noticeable difference between the two DA schemes. Comparing the observed and the 4DEnVar analysis, we can see that both 4DEnVar and 3DEnVar capture the observed convective system very well at the analysis time. As a diagnostic variable, although the difference of the radar reflectivity between the two DA schemes is trivial at the analysis, it grows with time (Figure 8) and the difference mainly lies in the connective features. Figure 9 gives the Equitable Threat Score (ETS) of the radar reflectivity verified against the observed for the testing period for 3DEnVar (blue) and 4DEnVar (red) for the first 6 hours. For both all scatters (>5dBZ) and all rain/snow (>15dBZ) reflectivity, there is barely any difference at 0-hour but the two experiments diverge with longer forecast time, with 4DEnVar having higher ETS than 3DEnVar in the first 6 hours except for hour 3. For convective precipitation (>30dBZ), we see mixed results and it is hard to say which DA scheme produced better forecasts.

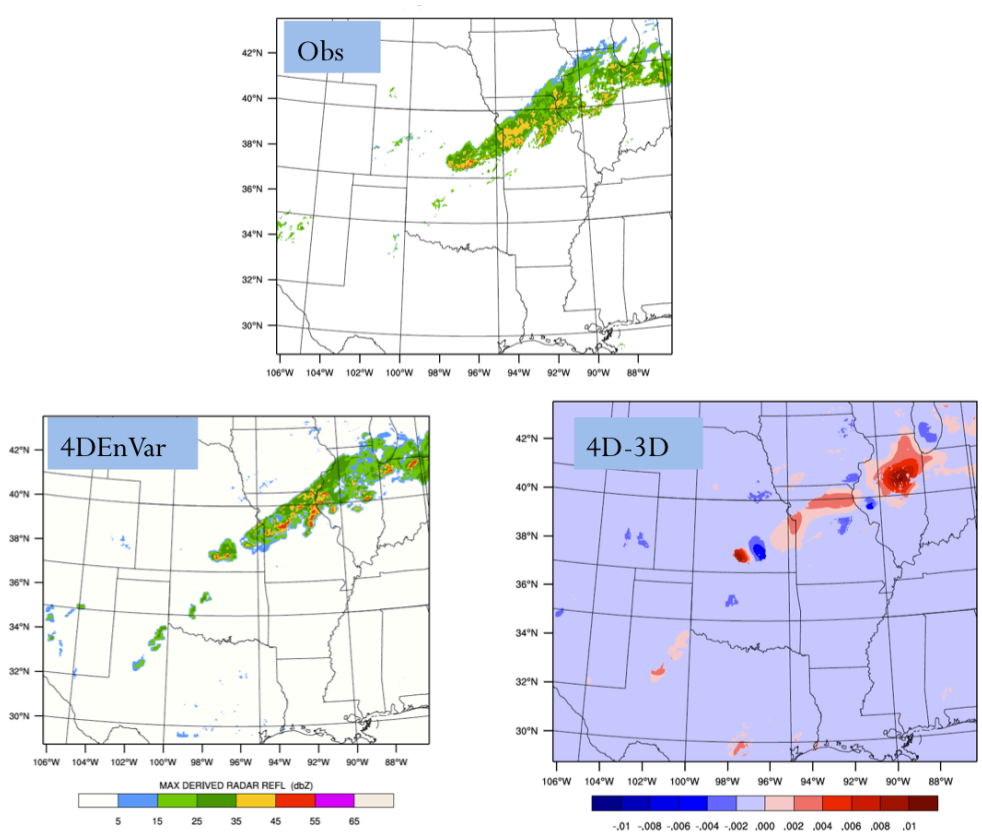

Figure 7. Column maximum radar reflectivity (dbZ) valid at 2016090812. Top panel is the observed. The bottom left panel is the 4DEnVar analysis and the bottom right is the difference between the 4DEnVar and 3DEnVar analysis.

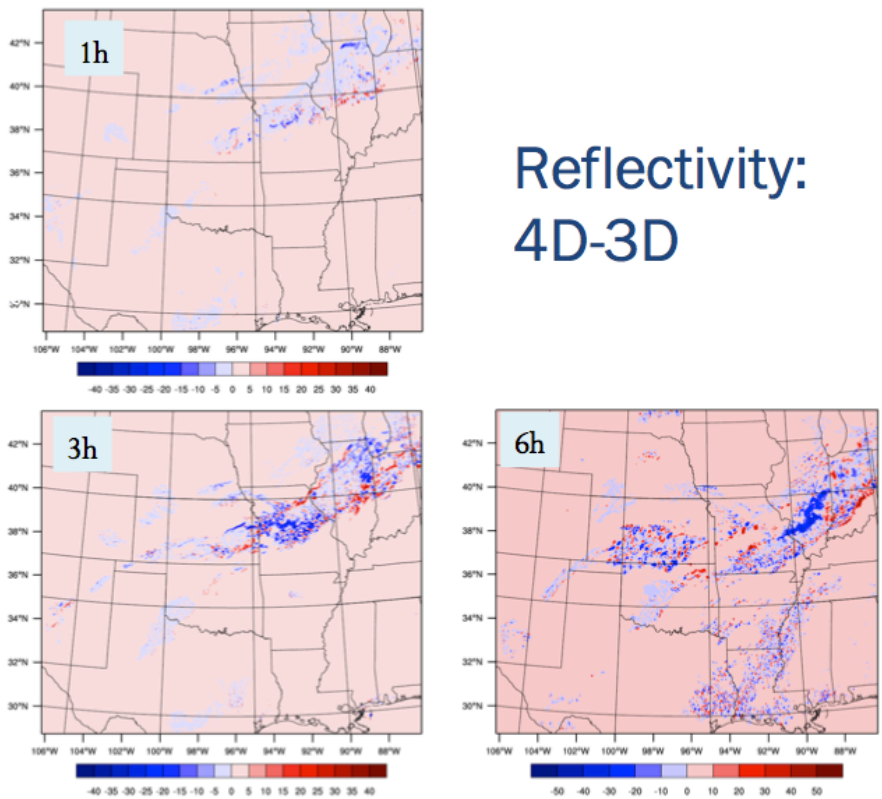

Figure 8 Radar Reflectivity difference between 4DEnVar and 3DEnVar for 1-, 3- and 6-hour forecast, initialized at 2016090812.

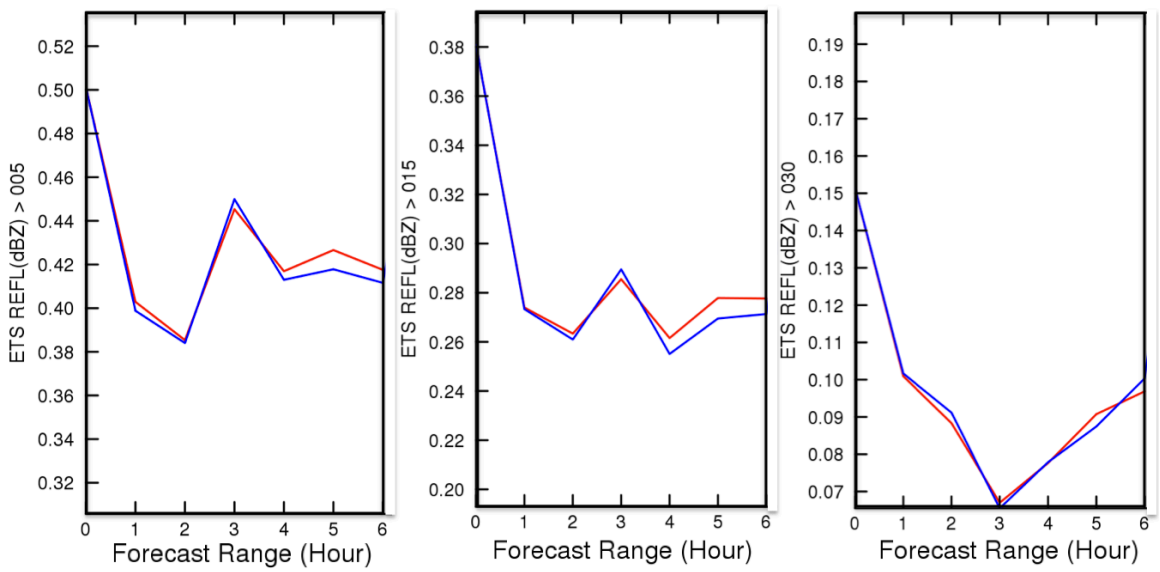

Figure 9. Equitable Threat Score (ETS) of the radar reflectivity verified against the observed for 3DEnVar (blue) and 4DEnVar (red) for the first 6 hours. From left to right: reflectivity >5dBZ, reflectivity >15dBZ and reflectivity >30dBZ.

The above results show some benefits from the 4DEnVar, but the impact is very limited. When looking at the configurations of the data assimilation, we find the following work areas that might contribute to further improvements:

1) All of the above experiments are initialized from the 1-hour pre-forecast from the RAP retrospective runs. The impact from the 3-km 4DEnVar data assimilation only last for the run and therefore there is no accumulated impact from the data assimilation schemes.
2) The GFS ensemble forecast has a much coarser horizontal resolution (~30km) than the model grid (3km) and can't resolve the convective structure on the fine grids, besides the GFS ensemble being very under-dispersive.
3) The unevenly distributed observations from the real-time observation feed in the three bins might affect the performance of the 4DEnVar.

In the following sections, case studies are conducted to address work area 1 and 2 in the above.

4. Cycling 3DEnVar and 4DEnVar

Two additional experiments were conducted for the convective events on September 8 2016:

- **3D_cycling**: hourly cycling 3DEnVar data assimilation
- **4D_cycling**: Hourly cycling 4DEnVar data assimilation

The workflow is slightly different from what is shown in Figure 1. Same as in Figure 1, the boundary conditions come from the 13-km RAP forecast. However, the initial field from RAP is used for only the first cycle, which is 00Z of September 8, 2016. For the cycles starting from 01Z of September 8 2016, the initial condition is from 1-hour forecast from previous HRRR forecasts and then used as the background fields for the 3DEnVar and 4DEnVar analysis.

The reason to run cycling experiments for September 8 2016 is that there is a fast evolving convective system within the domain, as can be seen from the precipitation and radar reflectivity observations (shown in later plots). The goal is to see the impact of cycling the two different data assimilation schemes on the convective structures.

Figure 10 gives the 6-hour accumulated Stage IV precipitation (inches) valid at 18Z September 8, in which a northeast – southwest tilted rain band extended from Indiana, Illinois, Missouri to southeastern Kansas with the strongest precipitation centers in the state of Missouri, as in the red oval. Figure 11 shows the 6-hour accumulated rainfall (inches) from the 3DEnVar experiment (right) and cycling 3DEnVar (left). As can be seen, the 3DEnVar with only one-hour pre-forecast did produce the northeast-southwest titled rain band but didn't capture the strong precipitation in the state of Missouri. Instead, the main convection center is in the state of Illinois, heavily overestimated for that area compared to the observed. On

the other hand, after continuously cycling the 3DEnVar data assimilation for 12 hours, the rainfall forecast initialized at 12Z September 8 2016 gives the strongest precipitation in the state of Missouri within the northeast-southwest tilted rainband, which is consistent with the observed, with some overestimation though. The unrealistic strong precipitation in the state of Illinois has been corrected through the accumulated effect of cycling the data assimilation. Figure 12 is similar to Figure 11, but for 4DEnVar cycling (left) and 4DEnVar (right). Similar to the case of 3DEnVar cycling, the strongest precipitation is in the state of Missouri. Compared to the overestimated rainfall in Missouri in 3DEnVar cycling, the magnitude of the rainfall in 4DEnVar cycling is reduced and much closer to the observed.

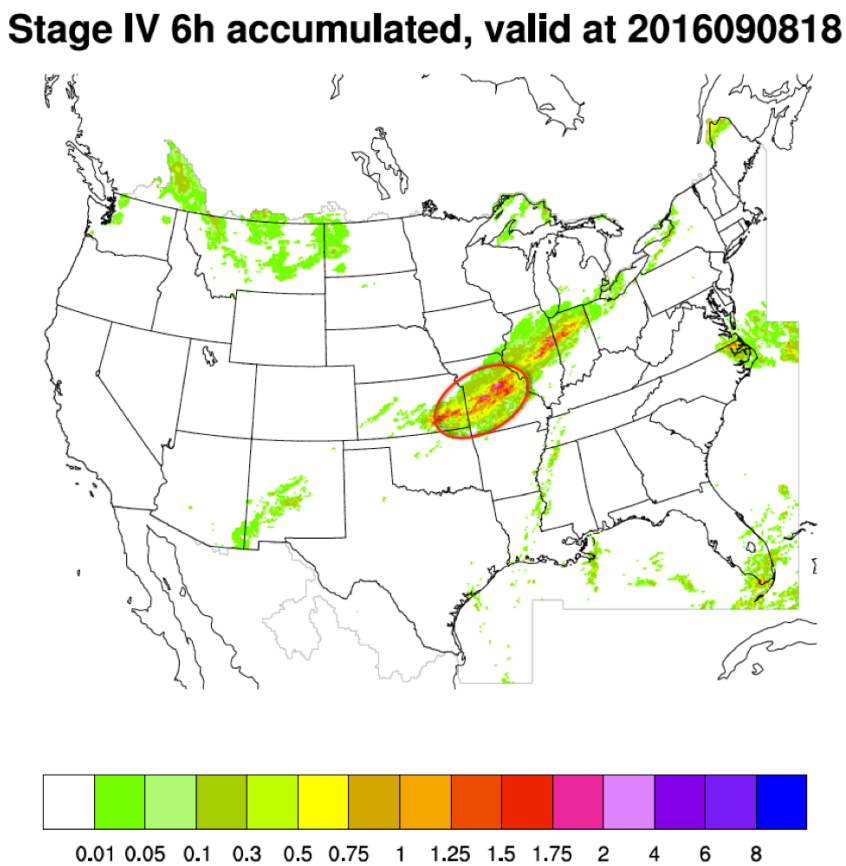

Figure 10. Stage IV 6-hour accumulated precipitation (inches) valid at 18Z September 8 2016.

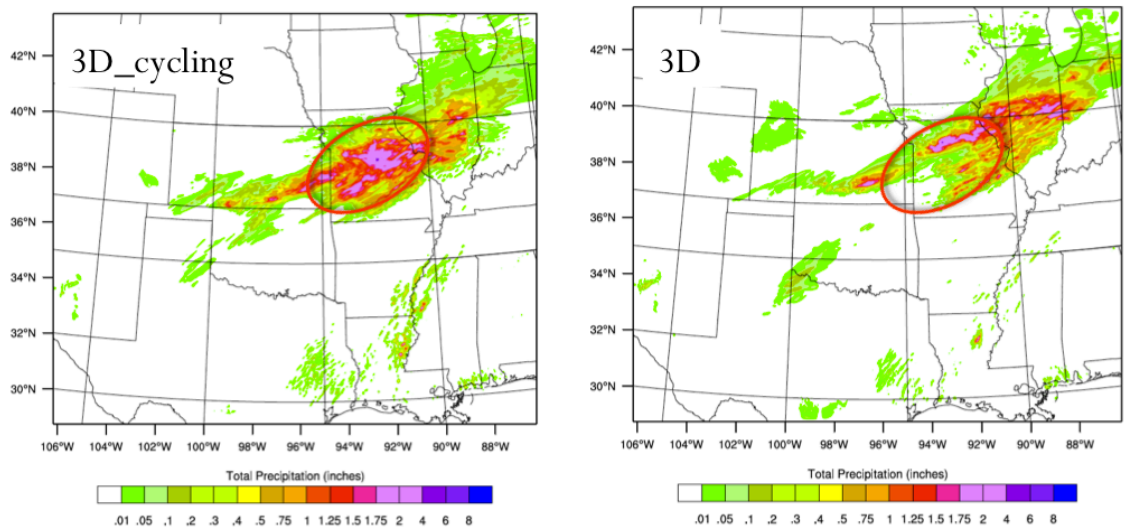

Figure 11. 6-hour accumulated precipitation (inches) valid at 18Z September 8 2016 from the experiment 3DEnVar cycling (left) and 3DEnVar (right).

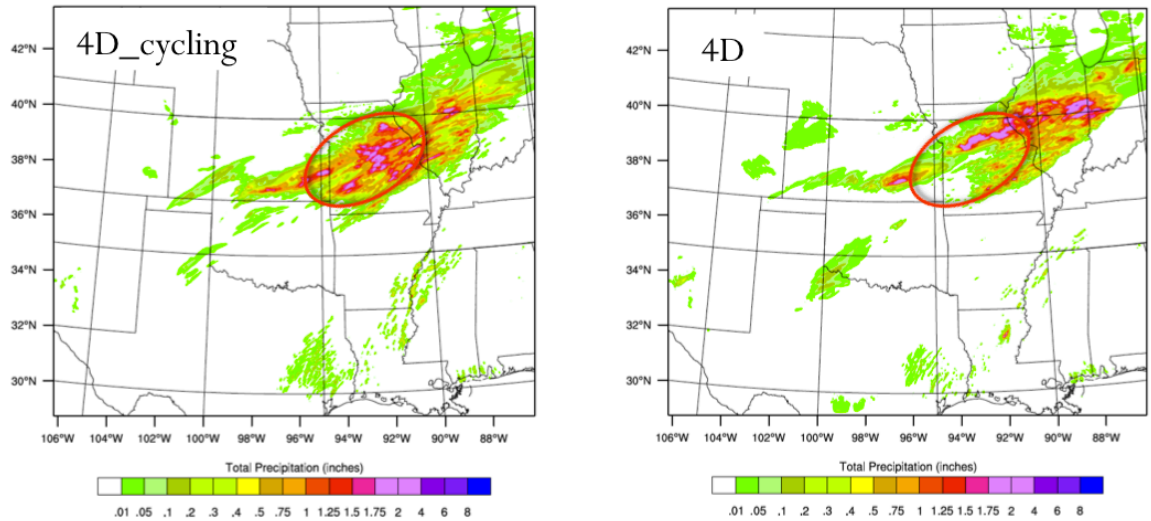

Figure 12. 6-hour accumulated precipitation (inches) valid at 18Z September 8 2016 from the experiment 4DEnVar cycling (left) and 4DEnVar (right).

Figure 13 presents the radar reflectivity at 00Z September 9 2016, which corresponds to the 12-hour forecast from 12Z September 8 2016. The observed is in the upper left, with the two main convective centers circled in red (at the intersection of four states: Missouri, Kansas, Oklahoma and Arkansas) and blue (crossing the southern part of Illinois and Indiana). The upper right panel gives the 12-hour forecast radar reflectivity from the experiment 4DEnVar, valid at the same time as the observed. Compared to the observed, the 4DEnVar well captured the convective system at southern Illinois and Indiana. But the convective system observed at the intersection of the four states is forecasted to be mainly located in Kansas and western Missouri, and it is too broad compared to the observed. The forecast from the 3DEnVar experiment is very similar (not shown here) and didn't

capture the convection at the intersection region. With 12 hours of data assimilation cycling, the convection at the borders of the four states, denoted as in the red oval, is now more localized at the borders and much closer to the observed. The intensity of the convection in 4DEnVar cycling is closer to the observations compared to the overestimated in cycling 3DEnVar.

The above results suggests that cycling 3DEnVar and 4DEnVar shows an accumulated positive impact compared to the "warm" start data assimilation as in 3DEnVar and 4DEnVar, and the cycling brings out more benefit from the 4DEnVar too.

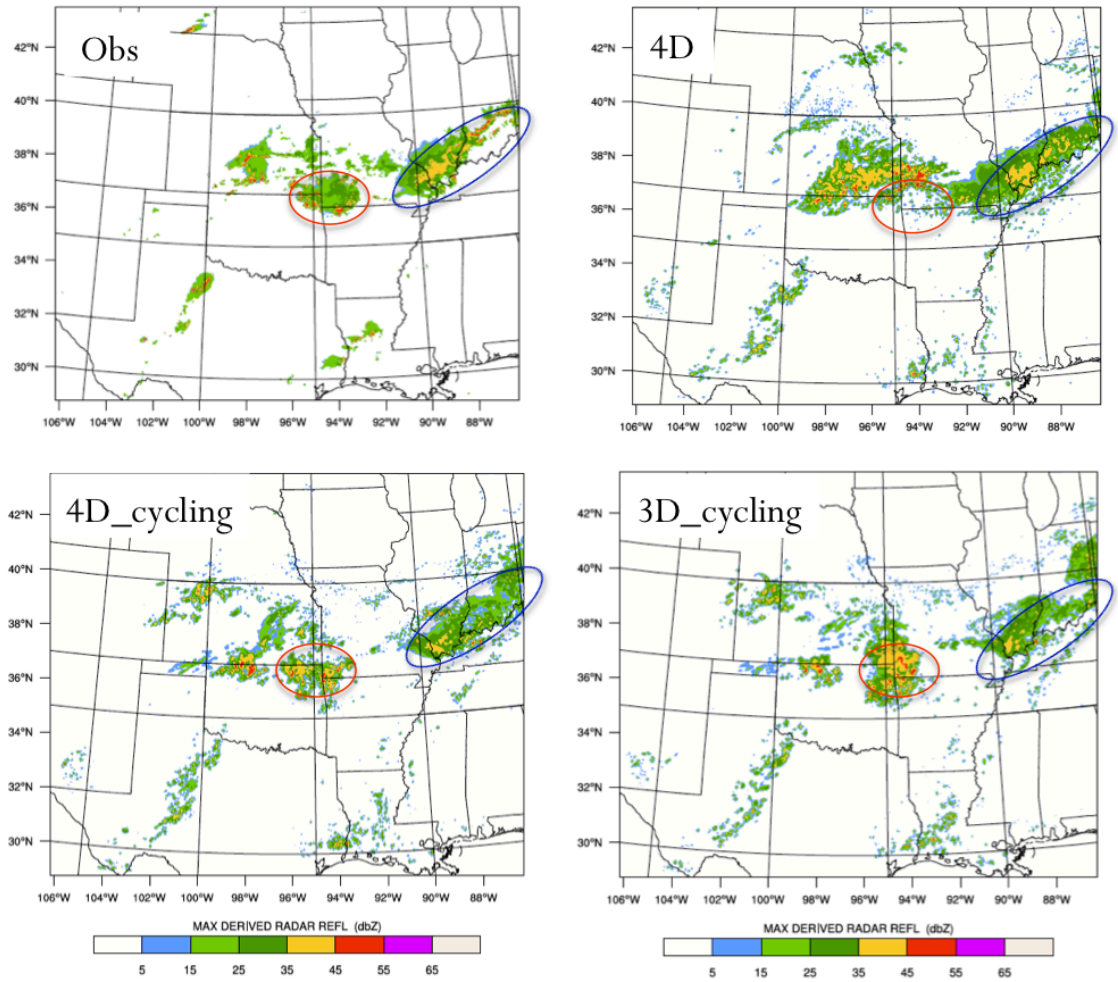

Figure 13. Radar Reflectivity (dBZ) valid at 00Z September 9 2016 from the observed (upper left), 12-hour forecast from: 4DEnVar (upper right), cycling 4DEnVar (lower left) and cycling 3DEnVar (lower right).

5. 3DEnVar using 3-km ARW ensembles

As discussed in section 2, the GFS ensemble forecast used in this work has a much coarser horizontal resolution (~30km) than the model grid (3km) and can't resolve the convective structure on the fine grids. This adds to the limitations of the hybrid

EnVar data assimilation. To explore the potential benefit of applying a regional high resolution ensemble forecast, an additional experiment is conducted in 3DEnVar using the regional 3-km ensemble forecast, which is generated by dynamically downscaling from the GFS ensemble. The reason we conducted 3DEnVar using the regional ensemble is that the GSI code applying regional ensemble for 4DEnVar is not ready yet.

Using the tool from NOAA/ESRL Global Systems Division (GSD), we calculated the perturbations from the 80-member GFS ensemble forecast at 06Z September 8 2016 and added these perturbations to the 3-km HRRR background, and then generated 80-member WRF initial fields. 6-hour WRF forecast is then performed for the 80 member HRRR ensembles. Figure 14 gives the ensemble spread from the 80-member 6-hour HRRR forecast (right column) and the GFS 6-hour ensemble forecast (left column), valid at 12Z September 8 2016. As can be seen, the ensemble spread from the GFS is very small while the HRRR ensemble gives a much stronger ensemble spread in all the four variables (T, q, u and v). Compared to the smooth structure in the GFS ensemble spread, the regional ensemble gives more details about the convections, as seen in all the variables.

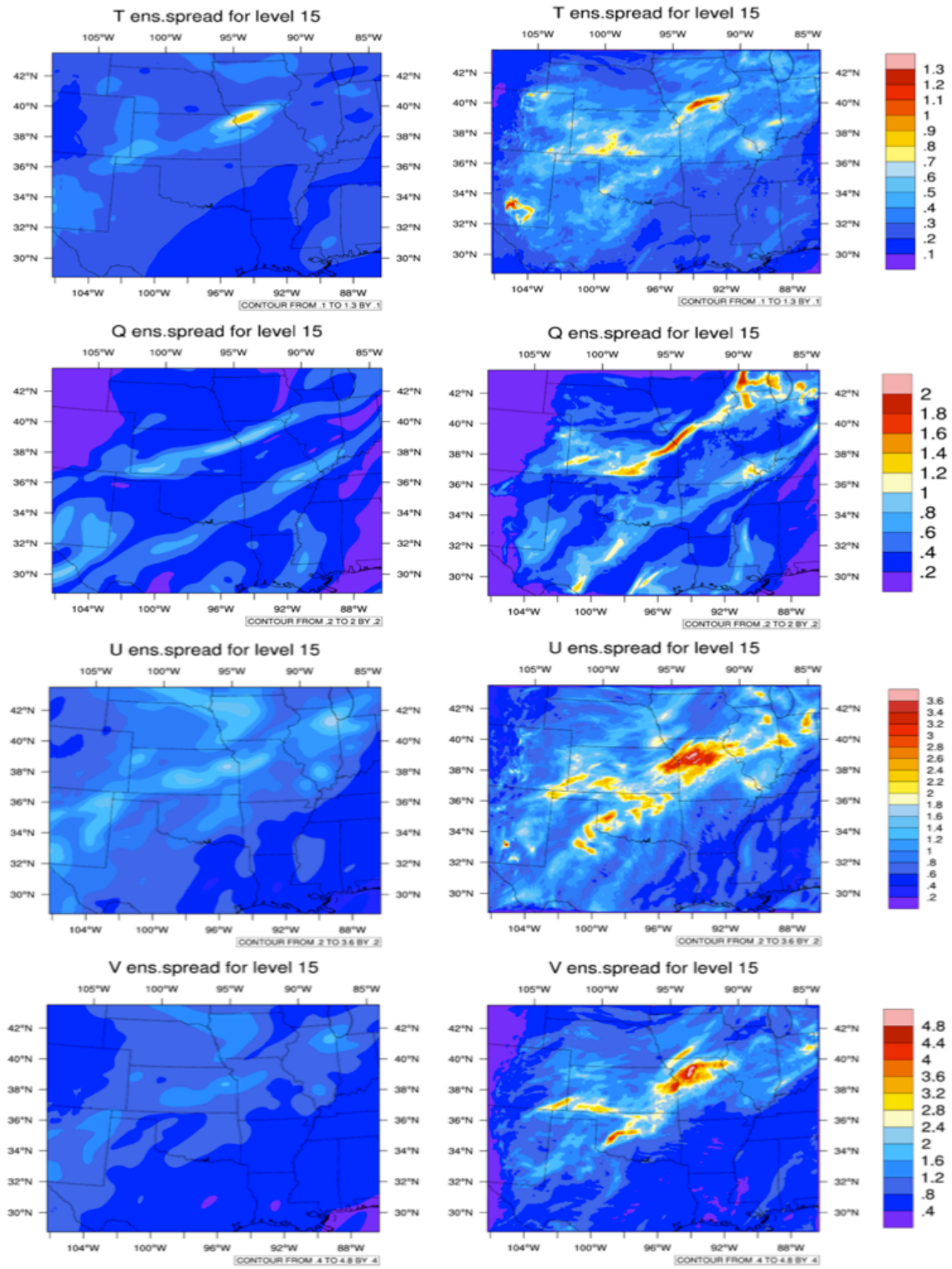

Figure 14. Ensemble spread from the GFS (left) and 3-km HRRR (right) at model level 15 for 2016090812. From the top to bottom: temperature (K, 1st row), humidity (g/kg, 2nd row), U wind (m/s, 3rd row) and V wind (m/s, 4th row).

A 3DEnVar experiment with the above regional ensemble forecast (3D_reg) is then conducted for 12Z September 8 2016 and the forecast is compared to the experiment with the GFS ensemble (3D). Preliminary results showed little difference in the 5-hour radar reflectivity forecast (Figure 15). The top panel in Figure 15 is the observed and bottom panels are the 5-hour forecast in the experiment 3D_reg (lower left) and 3D (lower right). Both 3D and 3D_reg captured the convective system, with slight location change, and the difference between the two experiments are negligible. Another look at the impact from applying the high-resolution regional ensemble forecast is to compare the difference between 3D_reg and 3D against the difference between 4D and 3D, as in Figure 16. As discussed in section 2, the radar reflectivity difference between 4DEnVar and 3DEnVar is very small at the analysis time but grows fast with forecast lead times. Figure 16 gives the 6-hour radar reflectivity forecast difference due to the regional ensemble (left, 3D_reg – 3D) and the data assimilation method (right panel, 4D-3D). The difference introduced in 3DEnVar by the 3-km regional ensemble is comparable to the difference brought by 4DEnVar.

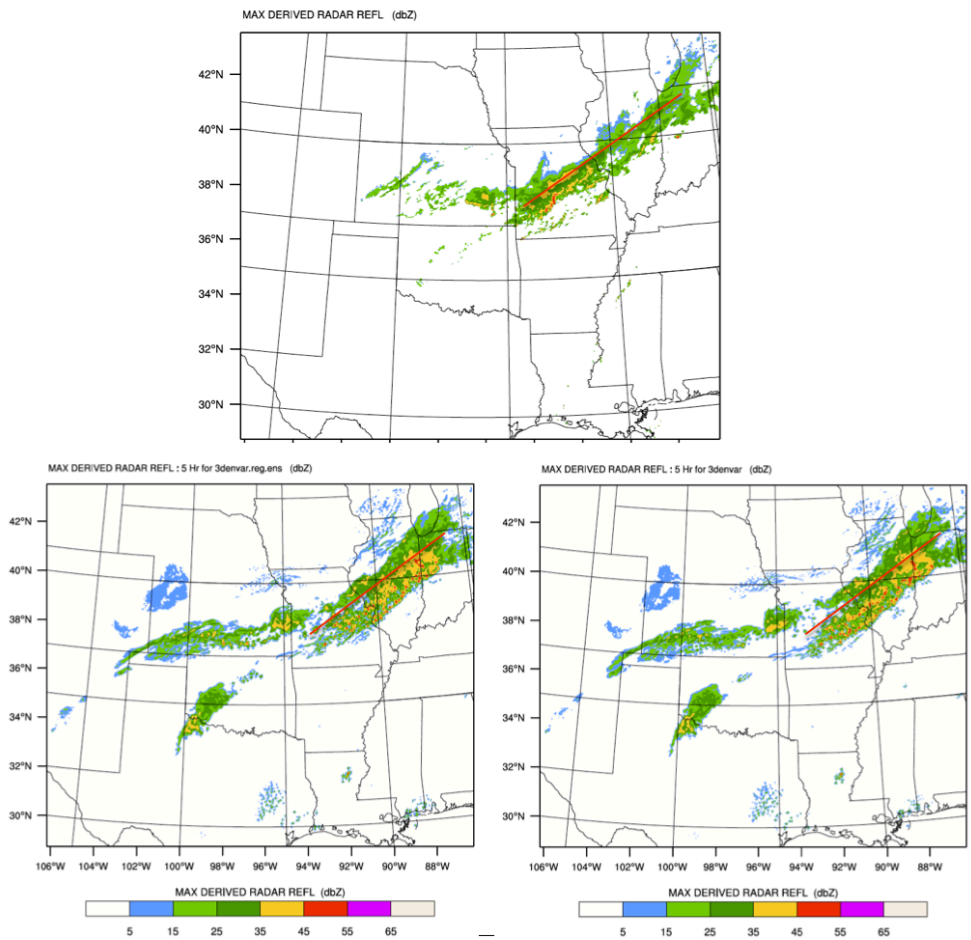

Figure 15. Radar reflectivity (dBZ) at 17Z September 8 2016. The top panel gives the observed. The bottom panels give the 5-hour forecast for 3D_reg (left) and 3D (right).

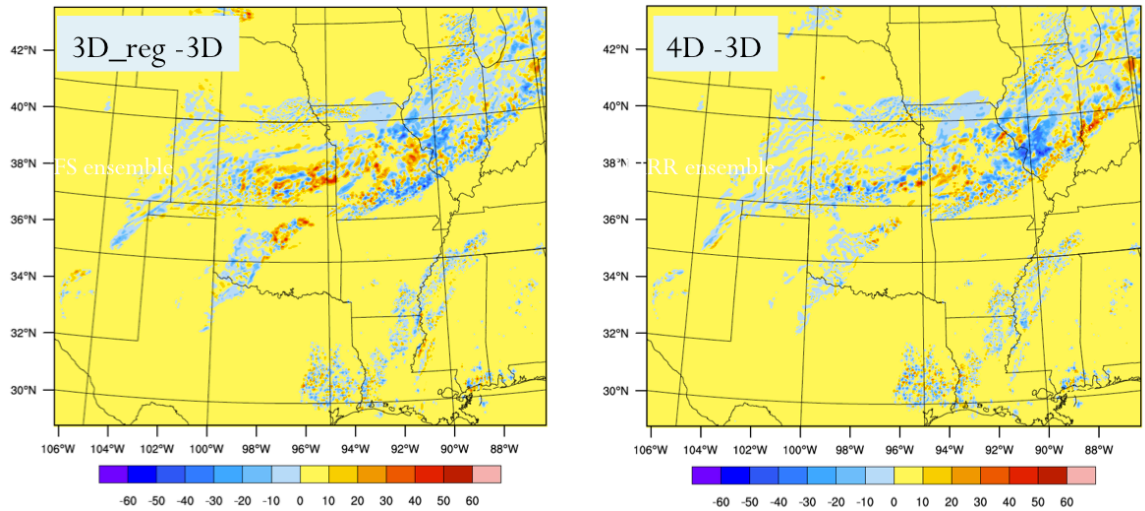

Figure 16. 6-hour forecast composite radar reflectivity difference (dBZ) between 3D_reg and 3D (left) and 4D and 3D (right) at the 9th model level, valid at 18Z September 8 2016.

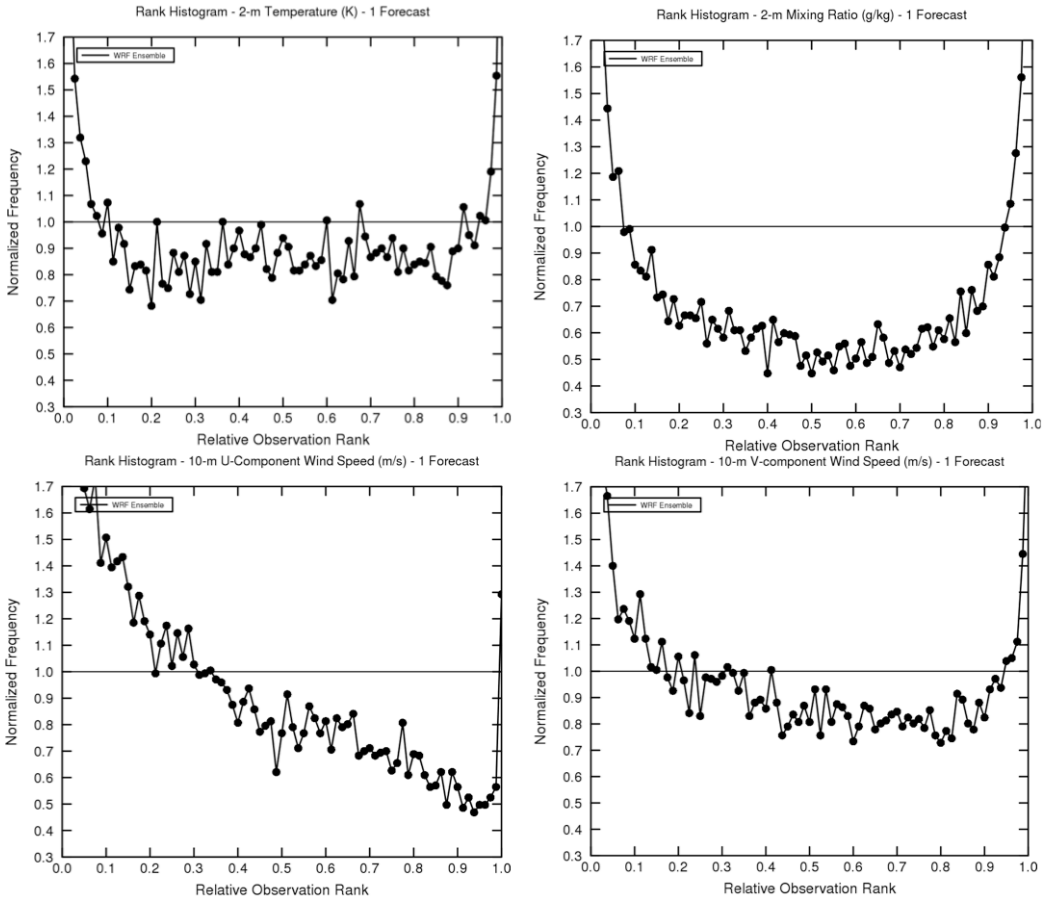

Figure 17. Rank histograms of 2-m temperature (upper left), 2-m mixing ratio (upper right), 10-m u-wind (lower left) and v-wind (lower right) for the 3-km regional ensemble forecast at 12Z September 8 2016.

Although the impact from applying the regional ensemble seems to be comparable to the impact from 4DEnVar, the forecast impact itself is still very small for this experiment at 12Z September 8 2016. On one hand, this is just one case study with one set of regional ensemble. More case studies need be performed in order to gain more meaningful insights. On the other hand, the regional ensemble, although improved upon the GFS ensemble, is still very under-dispersive, as shown in Figure 17 and 18. Figure 17 shows the rank histograms of the surface fields: 2-meter temperature and mixing ratio, and 10-meter winds. Figure 18 gives the rank histogram of the upper air temperature, mixing ratio and winds. Both figure 17 and 18 suggest that the regional ensemble at 12Z September 8 2106 is very under-dispersive and therefore limited the performance of the 3DEnVar experiment with these ensemble forecasts.

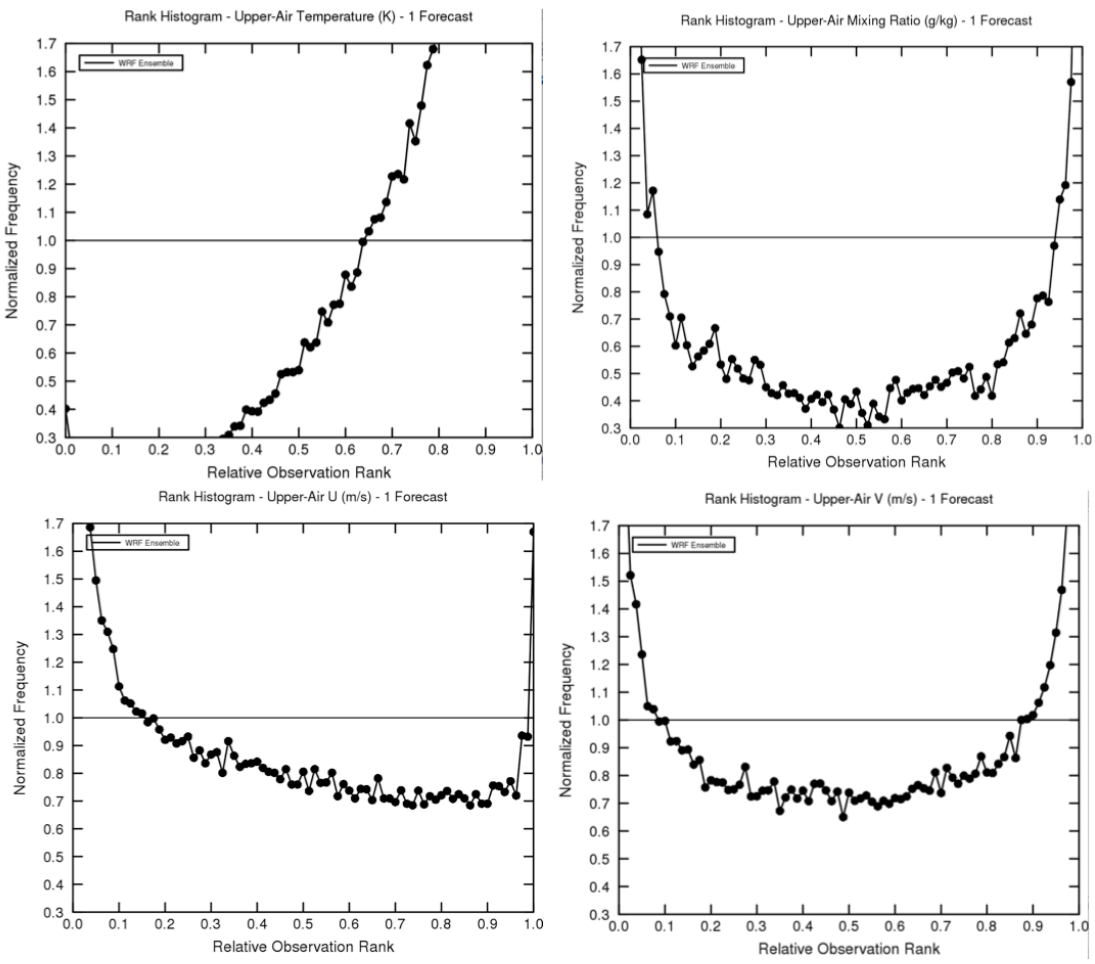

Figure 18. Rank histograms of the upper air temperature (upper left), mixing ratio (upper right), u-wind (lower left) and v-wind (lower right) for the 3-km regional ensemble forecast at 12Z September 8 2016.

As discussed, there might be some potential impact from applying regional high-resolution ensemble forecast in the hybrid EnVar data assimilation but there are

still more work to do to improve the ensemble representation. And more case studies would definitely help understand the potential impact from the regional ensembles.

6. Discussion

In this work, a reduced HRRR 3-km domain, following the operational HRRR configurations, is adopted. Hourly updated experiments are conducted for the period of September 3-10 2016, for the 3DEnVar and 4DEnVar data assimilation schemes. Preliminary results showed limited impact from applying the 4DEnVar data assimilation. Further case studies suggested that cycling the data assimilation for both 3DEnVar and 4DEnVar improved the convective scale forecasts. And the benefit of applying 4DEnVar is enhanced through the accumulated impact from cycling the data assimilation. Another case study of 3DEnVar with regional 3-km ensemble forecast brings comparable impact to that from 4DEnVar. The potential benefit of regional ensemble on the convective forecasts could be improved by further improving the ensemble representations.

Section III: Second Objective – Testing and Evaluation of Sub-Hourly Update 4D EnVar Data Assimilation

1. Background and Setup

The ultimate objective of the DA2 task for AOP 2016 was to test and evaluate sub-hourly 4D hybrid EnVar capabilities in the context of the operational HRRR, currently running hourly at 3-km resolution. Multiple milestones existed for the GSD portion of this T&E activity, including the production of a benchmark, week-long 3DEnVar HRRR retrospective simulation, running a sub-hourly 3DEnVar HRRR test, and finally, a sub-hourly 4DEnVar HRRR retro. All milestones for this task are now complete. A description of all systems involved in this T&E activity is presented herein, along with GSD node results, plus analysis and conclusions.
The first step in assessing sub-hourly 3D or 4D hybrid EnVar within the HRRR was to produce a week-long benchmark simulation using the operational configuration of the HRRR (Figure 1). Run on an hourly basis, the HRRR uses downscaled RAP analysis fields to run a one-hour pre-forecast, during which time radar data are used to produce temperature tendencies, implemented during a 60-minute period. Following this pre-forecast, GSI is run in 3D hybrid EnVar mode, using the 6-hourly GFS ensemble and assimilating conventional and aircraft observations, with a hydrometeor analysis for cloud building/clearing conducted afterward. Once finished, a 24-hr simulation is produced, and the surface field information is cycled and combined with the downscaled RAP for the next hour. This configuration of the HRRR was implemented in retrospective mode on Theia and a benchmark simulation was run from 3-10 September, 2016, during which time 1-hr GFS ensemble data were archived for later use in the sub-hourly testing.

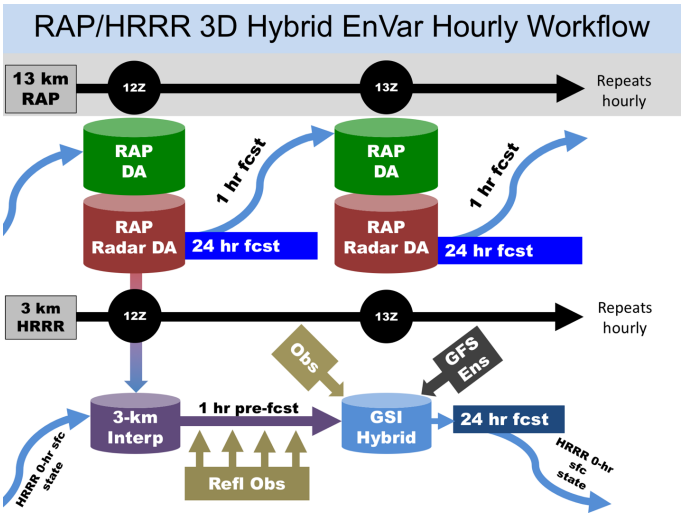

Figure 1. Hourly 3DEnVar workflow diagram showing the simultaneous evolution of the operational RAP and HRRR.

After verification was conducted for the benchmark HRRR retro, the sub-hourly 3DEnVar HRRR simulation was configured. A workflow diagram for the sub-hourly 3D hybrid EnVar HRRR is shown in Figure 2. The main difference between this new set-up and the operational design of the HRRR is the sub-hourly implementation of GSI run in 3DEnVar mode (a reference to the operational configuration is shown at the bottom of Figure 2). Fifteen-minute HRRR forecasts, with sub-hourly conventional, radar, and aircraft observations are combined with 8-h GFS ensemble information every 15 minutes until the top of the hour, when a new 24-hr forecast is conducted. Being computationally more expensive than the hourly-updating HRRR, a subset of the seven-day period was chosen from 8-10 September, 2016, during which time active weather was rapidly evolving over the continental United States. In the hopes that this period would effectively illustrate the ability of sub-hourly rapid updates to improve HRRR forecasts, the sub-hourly 3DEnVar HRRR retro was run during these three days.

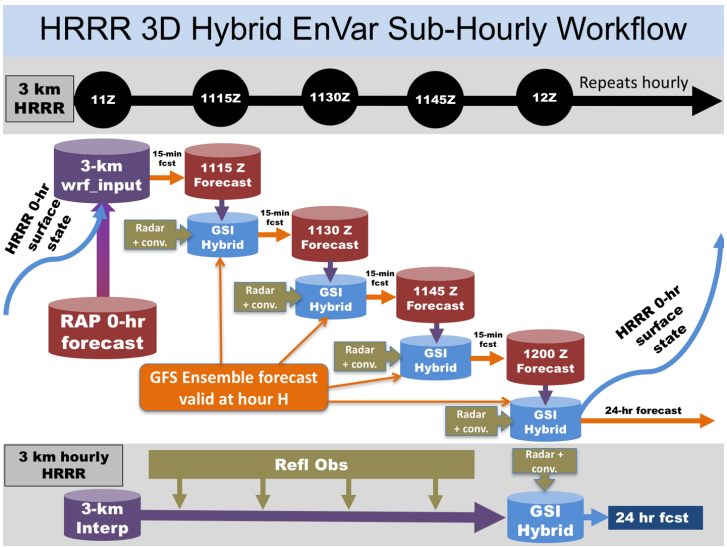

Figure 2. Sub-hourly 3DEnVar workflow diagram showing the evolution of the 15-minute assimilation cycles prior to the 24-hour forecast issued on the hour. Note the single-hour GFS ensemble forecasts and 15-m HRRR forecasts at each cycle needed for 4DEnVar assimilation at three time levels. The operational 3-km hourly HRRR workflow is at the bottom in gray for reference.

.

One way to evaluate how these sub-hourly cycles impact the data assimilation as well as the forecast as a whole, is to assess the number of observations assimilated during these sub-hourly cycles. Table 1 details assimilated observation totals from METAR and ACARS data for the sub-hourly cycles leading up to the 24-hour forecast on 8 September 2016 at 1800 UTC. Surface-based stations and flight-level aircraft observations represent the majority of data available at sub-hourly intervals and can provide a wealth of information on quickly evolving meteorological phenomena which can be assimilated during each 15-minute cycle.

Table 1. Assimilated observations for each 15-minute cycle of the sub-hourly 3DEnVar 8 September 2016, 1800 UTC run. METAR and ACARS temperature, winds, and moisture observations are shown.

Observation Type	**Analysis Time During Pre-Forecast (17Z-18Z)**			
	45	**30**	**15**	**0**
Surface Pressure (METAR)	922	968	2788	2196
Temperature (METAR)	906	955	2754	2176
Temperature (ACARS)	13162	8149	7761	547
Surface Winds (METAR)	876	930	2595	2012
Winds (ACARS)	13083	8074	7808	8012
Moisture (METAR)	892	944	11787	2158
Moisture (ACARS)	730	491	574	585

A final retrospective run was conducted using the sub-hourly 4DEnVar HRRR configuration (Figure 3). In this set-up, the 15-minute HRRR forecasts from the sub-hourly 3DEnVar test were extended to 30 minutes, since three time levels are needed at each 15-minute interval for GSI to be run in 4DEnVar mode. Therefore, at each cycle, information from HRRR forecasts at the analysis time, 15 minutes, and 30 minutes was used. Additionally, GFS ensemble forecasts at 7, 8, and 9 hours were incorporated into the variational GSI analysis. This retrospective HRRR configuration ran for 8-10 September, 2016, and was compared with both the sub-hourly 3D hybrid EnVar and operational, hourly 3D hybrid EnVar configuration.

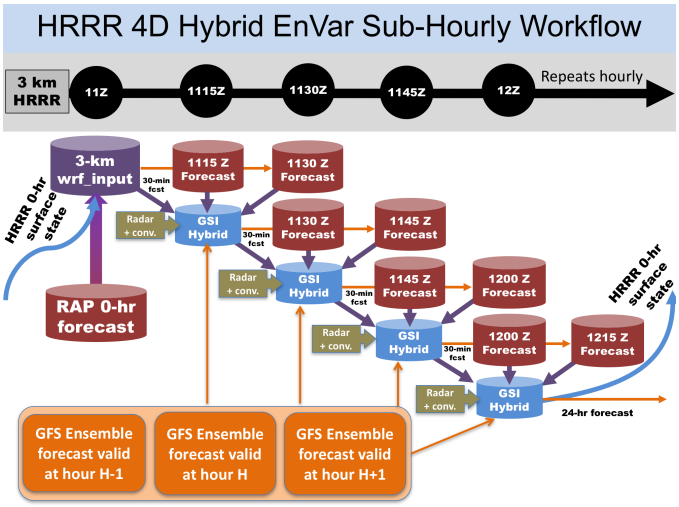

Figure 3. Sub-hourly 4DEnVar workflow diagram showing the evolution of the 15-minute assimilation cycles prior to the 24-hour forecast issued on the hour. Note the multiple-hour GFS ensemble forecasts and 30-m HRRR forecasts at each cycle needed for 4DEnVar assimilation

2. Evaluation and Analysis

Verification for the hourly and sub-hourly 3D hybrid EnVar, and sub-hourly 4D hybrid EnVar HRRR experiments was conducted at the top of the hour, for the 24-hr forecast cycle. Therefore, comparison between the three configurations took place after the four, 15-minute rapid update cycles had finished for the sub-hourly runs

(and after the hour-long pre-forecast for the operational configuration). RMSE, bias, and differences for vertical profiles of temperature, relative humidity, and winds from 8-10 September 2016 for all three experiments can be seen in Figures 4 to 7. Verification of 2-m temperature, 2-m dew point, and 10-m winds from the same time period is shown in Figures 8 to 11.

For the analysis-time vertical profile comparisons (Figure 4), the rapid-update cycles have favored a closer fit to the observations by the free-forecast, with the sub-hourly 3D and 4D hybrid EnVar experiments showing generally lower RMSE and bias compared to the hourly, operational HRRR setup. However, by the 1-hr forecast (Figure 5), these improved scores have diminished, and all three experiments exhibit much more similar RMSE and bias values. Some of the lower level profiles even show a very small statistical advantage of the operational setup over the sub-hourly 3D and 4D hybrid EnVar experiments; however, wind bias for the sub-hourly experiments appear slightly better than for the hourly 3DEnVar experiment. Similarities continue between the two setups at the 6-hr (Figure 6) and 12-hr (Figure 7) forecast period, with very few statistically significant differences in upper-air profile verification between the two experiments.

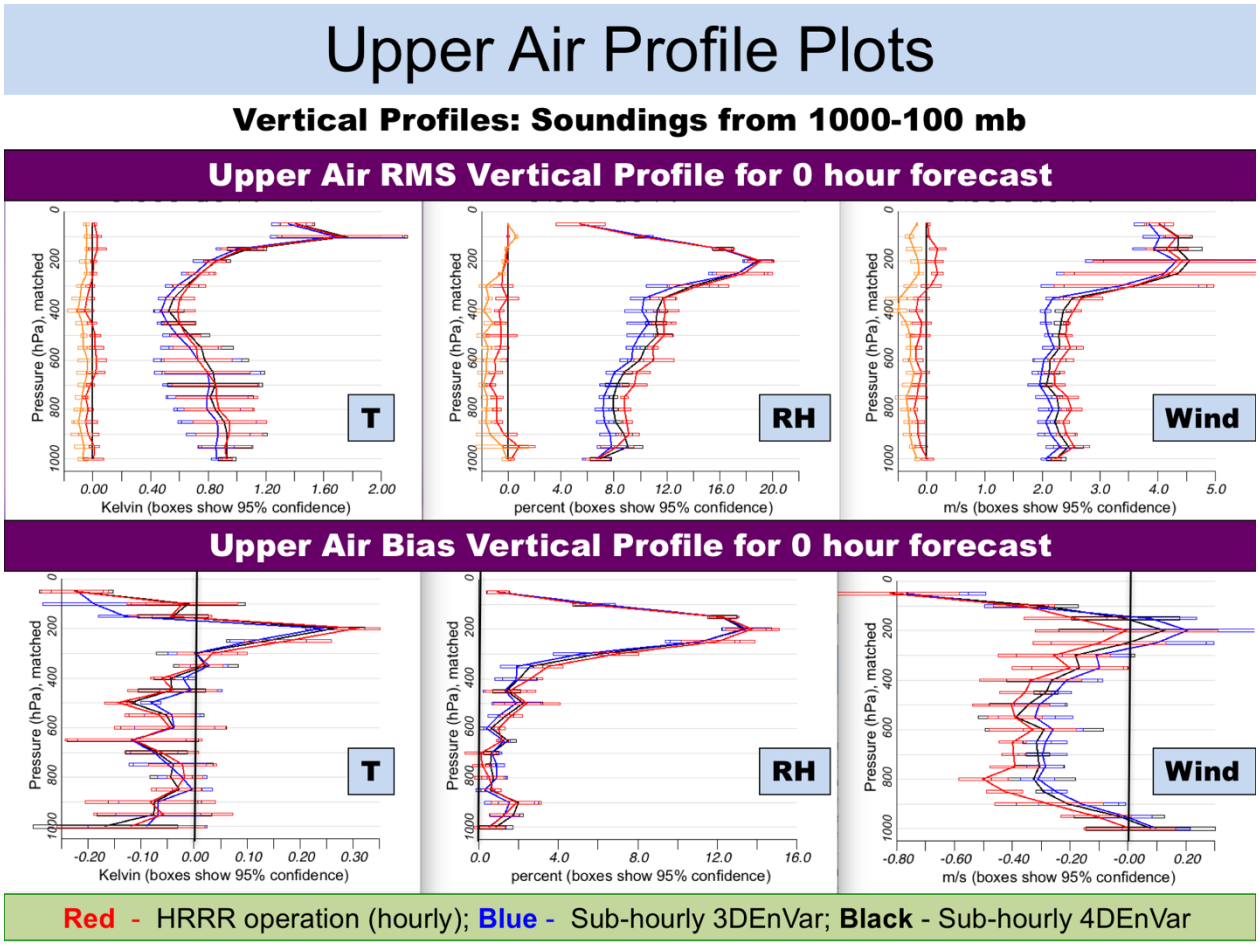

Figure 4. Upper-air profiles of temperature (left), relative humidity (center), and wind (right) mean RMSE (top) and mean bias (bottom) for hourly 3DEnVar (red), sub-hourly 3DEnVar (blue), and sub-hourly 4DEnVar (black) for 0-hr forecasts. RMSE difference values are plotted in orange (sub-hourly 3DEnVar minus hourly 3DEnVar) and red (sub-hourly 4DEnVar minus hourly 3DEnVar). Error bars are also shown for all three curves.

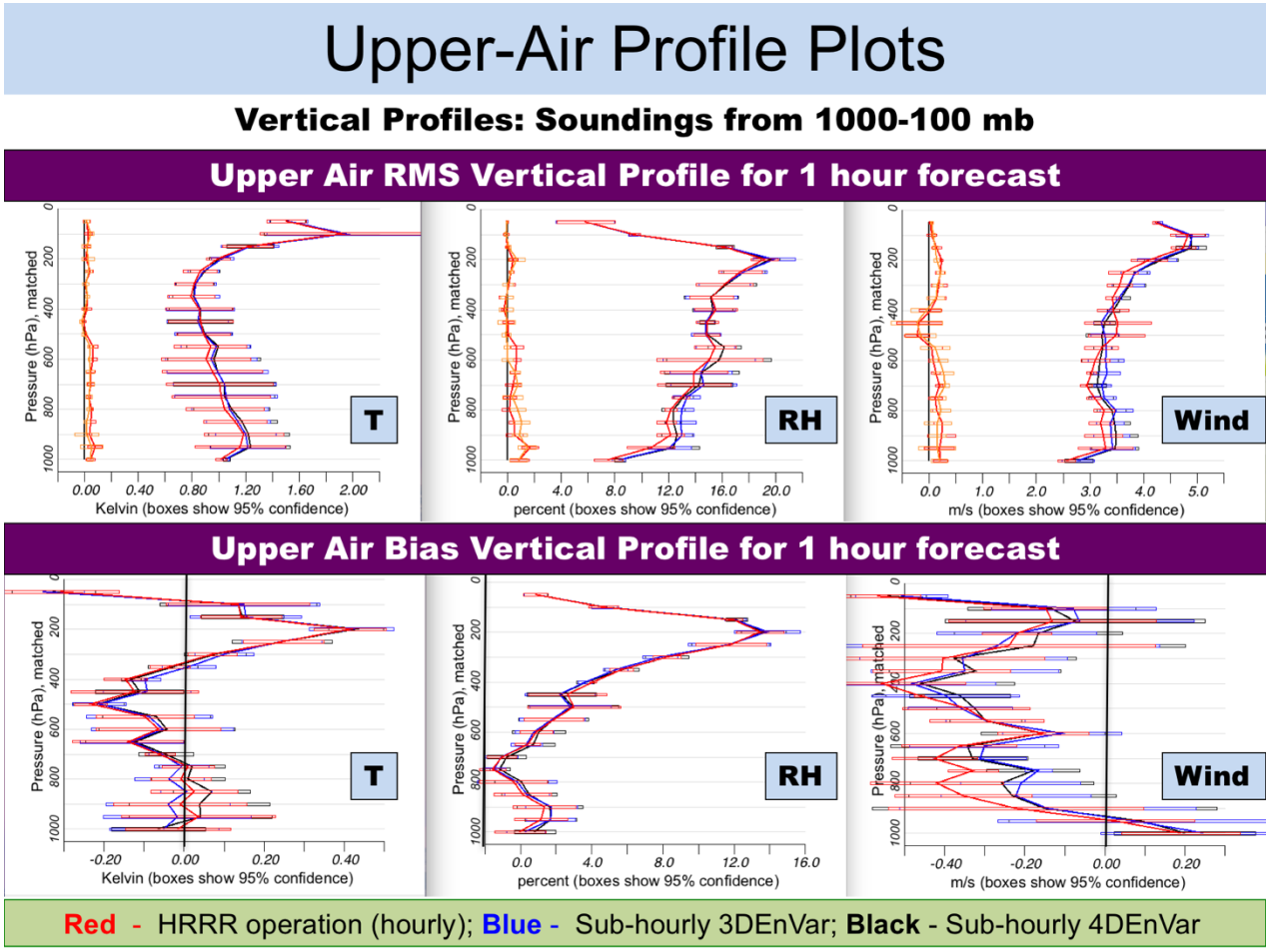

Figure 5. Upper-air profiles of temperature (left), relative humidity (center), and wind (right) mean RMSE (top) and mean bias (bottom) for hourly 3DEnVar (red), sub-hourly 3DEnVar (blue), and sub-hourly 4DEnVar (black) for 1-hr forecasts. RMSE difference values are plotted in orange (sub-hourly 3DEnVar minus hourly 3DEnVar) and red (sub-hourly 4DEnVar minus hourly 3DEnVar). Error bars are also shown for all three curves.

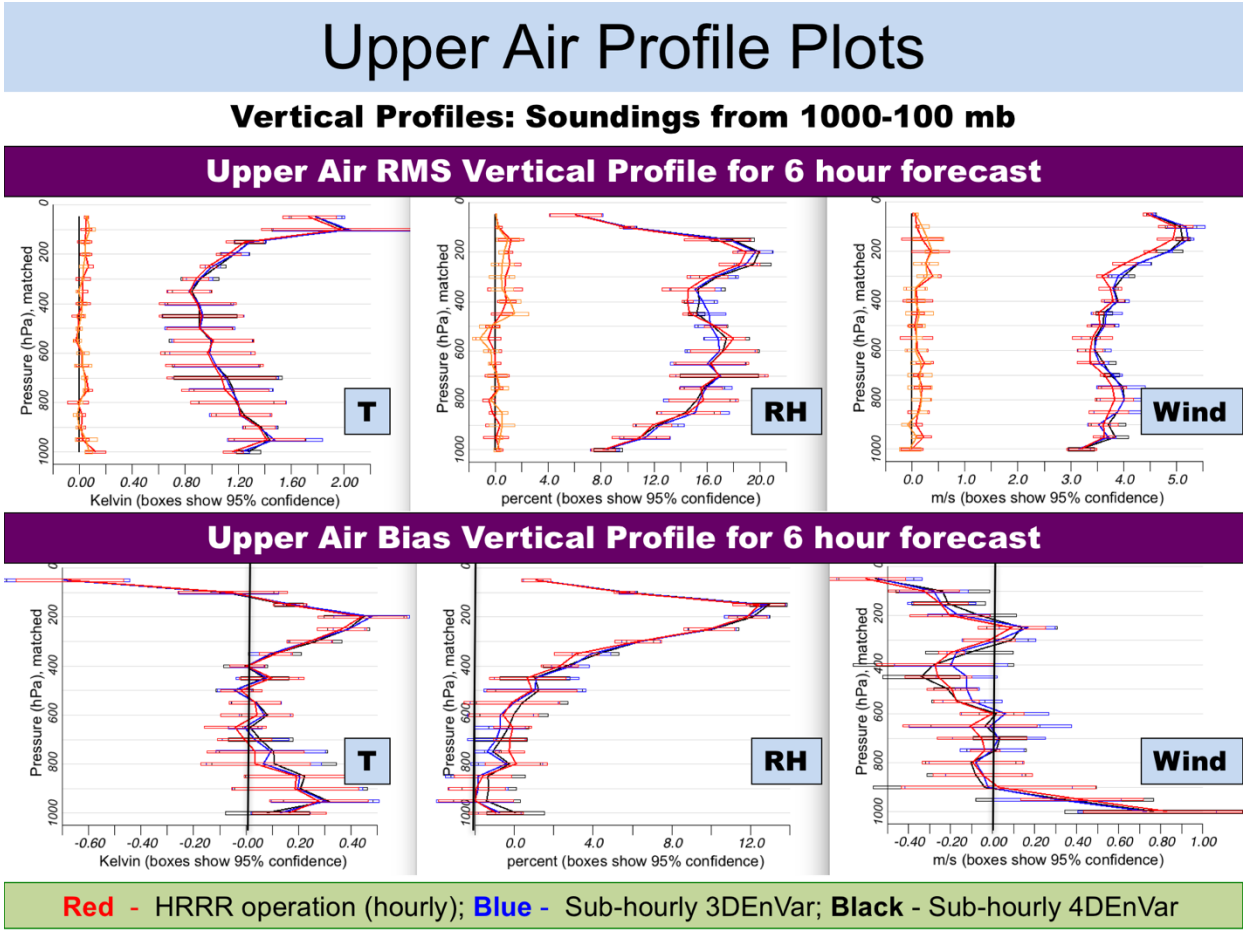

Figure 6. Upper-air profiles of temperature (left), relative humidity (center), and wind (right) mean RMSE (top) and mean bias (bottom) for hourly 3DEnVar (red), sub-hourly 3DEnVar (blue), and sub-hourly 4DEnVar (black) for 6-hr forecasts. RMSE difference values are plotted in orange (sub-hourly 3DEnVar minus hourly 3DEnVar) and red (sub-hourly 4DEnVar minus hourly 3DEnVar). Error bars are also shown for all three curves.

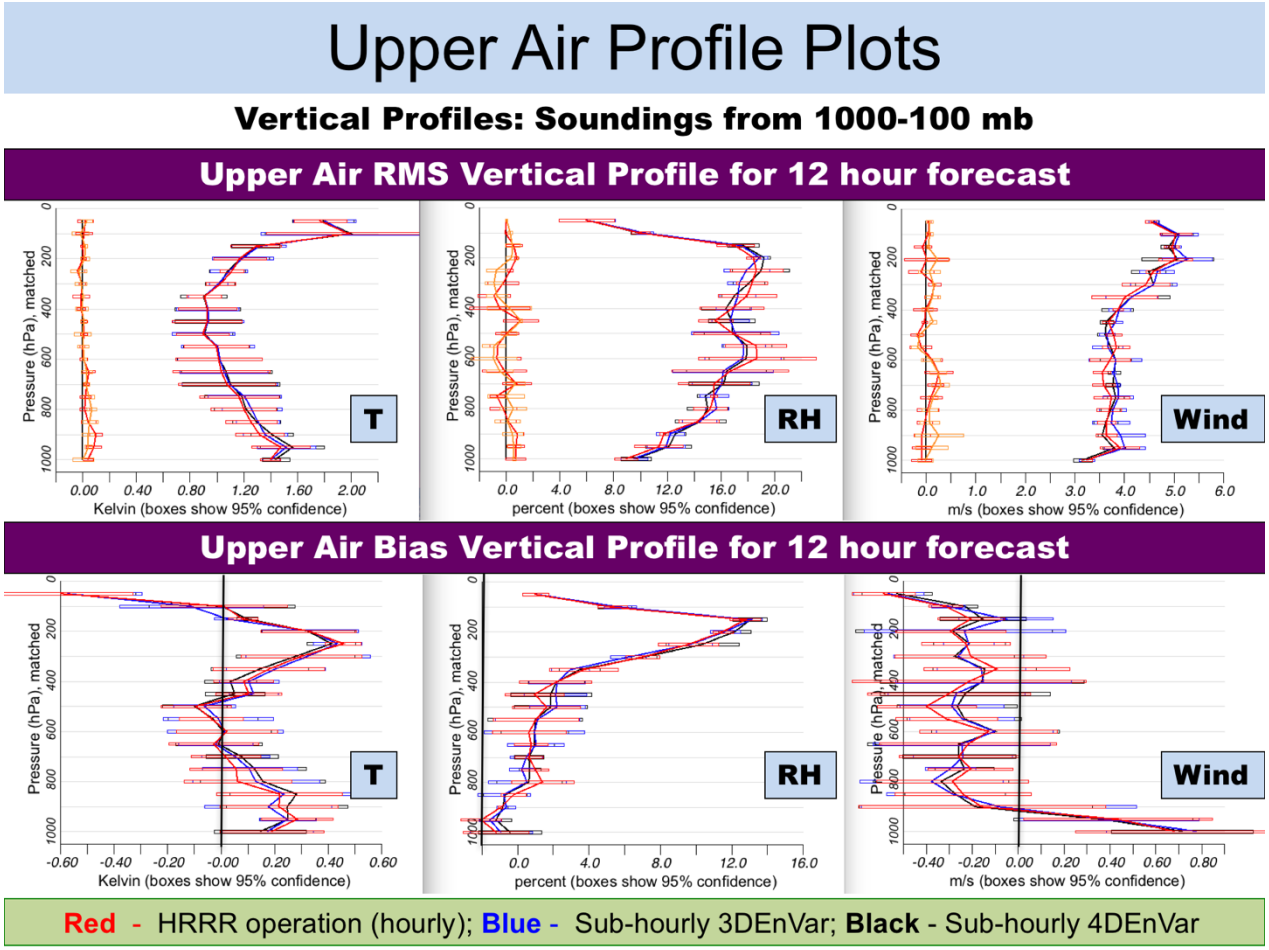

Figure 7. Upper-air profiles of temperature (left), relative humidity (center), and wind (right) mean RMSE (top) and mean bias (bottom) for hourly 3DEnVar (red), sub-hourly 3DEnVar (blue), and sub-hourly 4DEnVar (black) for 12-hr forecasts. RMSE difference values are plotted in orange (sub-hourly 3DEnVar minus hourly 3DEnVar) and red (sub-hourly 4DEnVar minus hourly 3DEnVar). Error bars are also shown for all three curves.

Given the evolution to very similar results after the analysis time, and the large error bars present, a decision was made to analyze layer-averaged, upper-air time-series to assess whether these results were stable with time. Overall, the analysis period upper-air time series showed consistent differences with time with the sub-hourly experiments having generally lower RMSE and bias. These differences are found to decrease markedly for longer forecasts, with RMSE nearly identical between the experiments. However, throughout the different forecast hours, the experiments illustrate consistent trends with time, suggesting that a longer verification period would yield similar results.

At analysis time, surface-based evaluation reveals that the 3D hybrid EnVar sub-hourly configuration results in reduced RMSE for 2-m temperature and relative humidity compared to the operational configuration (Figure 8), while 10-m wind verification is mostly similar between the two experiments. For the 4D hybrid

EnVar simulations, RMSE results are comparable to the operational system for 2-m temperature and relative humidity, but notably worse for 10-m winds for the 0-hr forecast. A strong diurnal cycle is evident in the bias plots, but the sub-hourly experiments result in more stable (and lower) bias characteristics for 2-m temperature and relative humidity. However, the 4D hybrid EnVar experiment exhibits enhanced bias for 10-m winds compared to the other two experiments, in line with the 4D hybrid EnVar RMSE errors for 10-m winds. For 1- (Figure 9), 3- (Figure 10), and 6-hour (Figure 11) forecasts, differences between the sub-hourly experiments reduce considerably, but both have slightly higher RMSE for all three variables than the operational configuration. Generally speaking, bias for the sub-hourly simulations are comparable to each other at extended forecast hours and are more stable than the operational setup. However, for 10-meter winds, all three experiments show very similar bias and diurnal cycles.

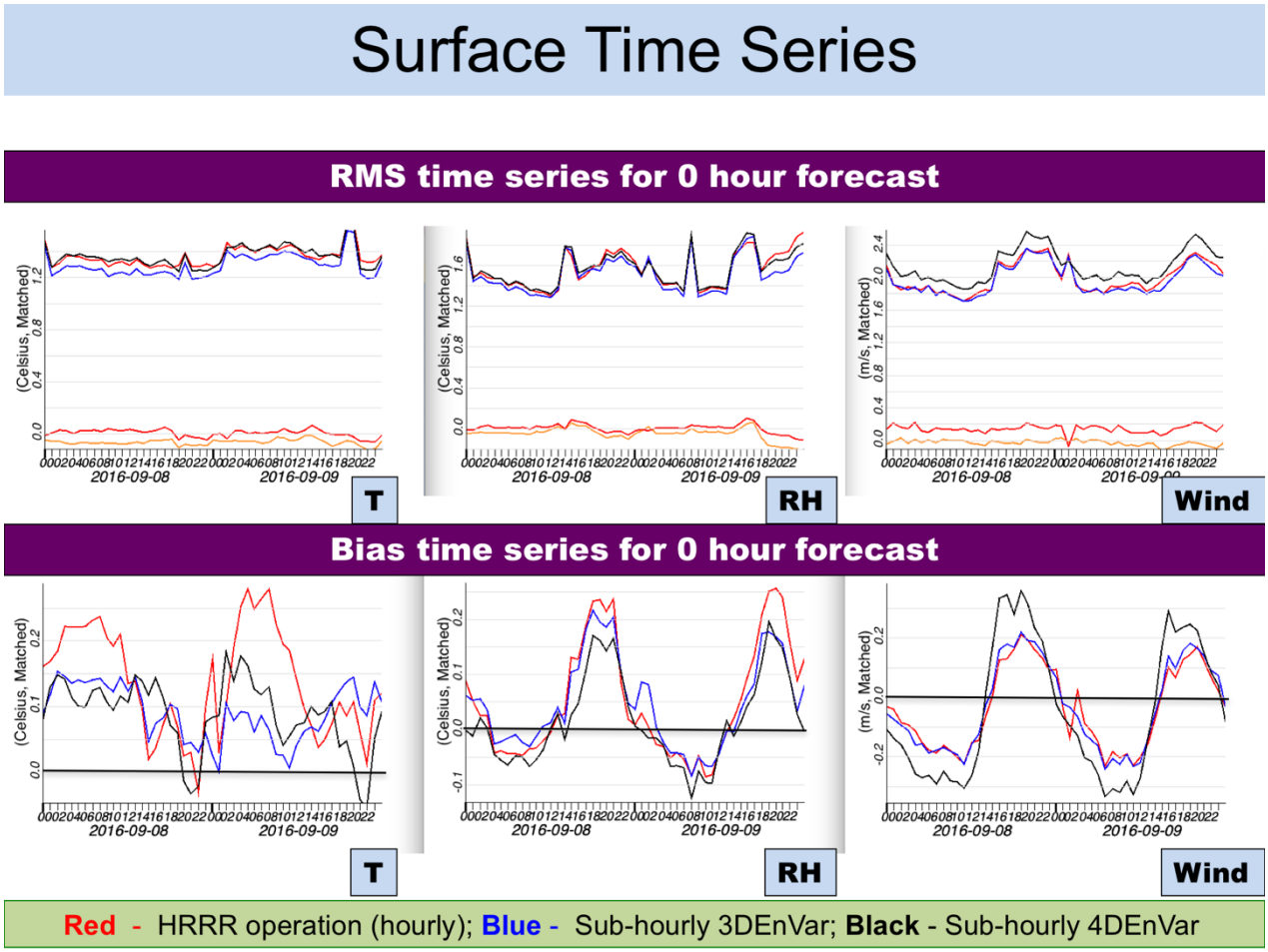

Figure 8. Time series of 2-m temperature (left), 2-m relative humidity (center), and 10-m wind (right) mean RMSE (top) and mean bias (bottom) for hourly 3DEnVar (red), sub-hourly 3DEnVar (blue), and sub-hourly 4DEnVar (black) for 0-hr forecasts. RMSE difference values are plotted in orange (sub-hourly 3DEnVar minus hourly 3DEnVar) and red (sub-hourly 4DEnVar minus hourly 3DEnVar).

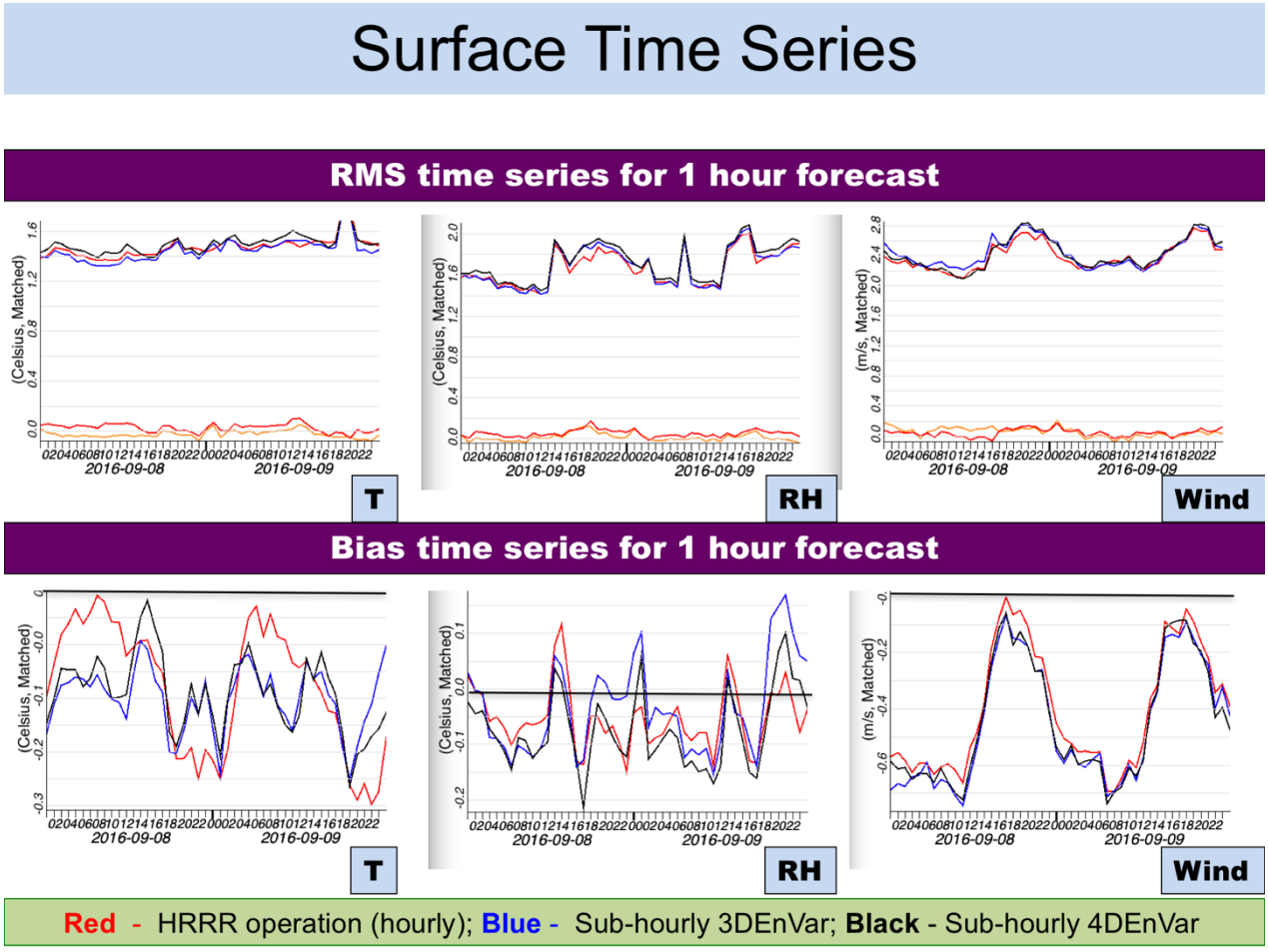

Figure 9. Time series of 2-m temperature (left), 2-m relative humidity (center), and 10-m wind (right) mean RMSE (top) and mean bias (bottom) for hourly 3DEnVar (red), sub-hourly 3DEnVar (blue), and sub-hourly 4DEnVar (black) for 1-hr forecasts. RMSE difference values are plotted in orange (sub-hourly 3DEnVar minus hourly 3DEnVar) and red (sub-hourly 4DEnVar minus hourly 3DEnVar).

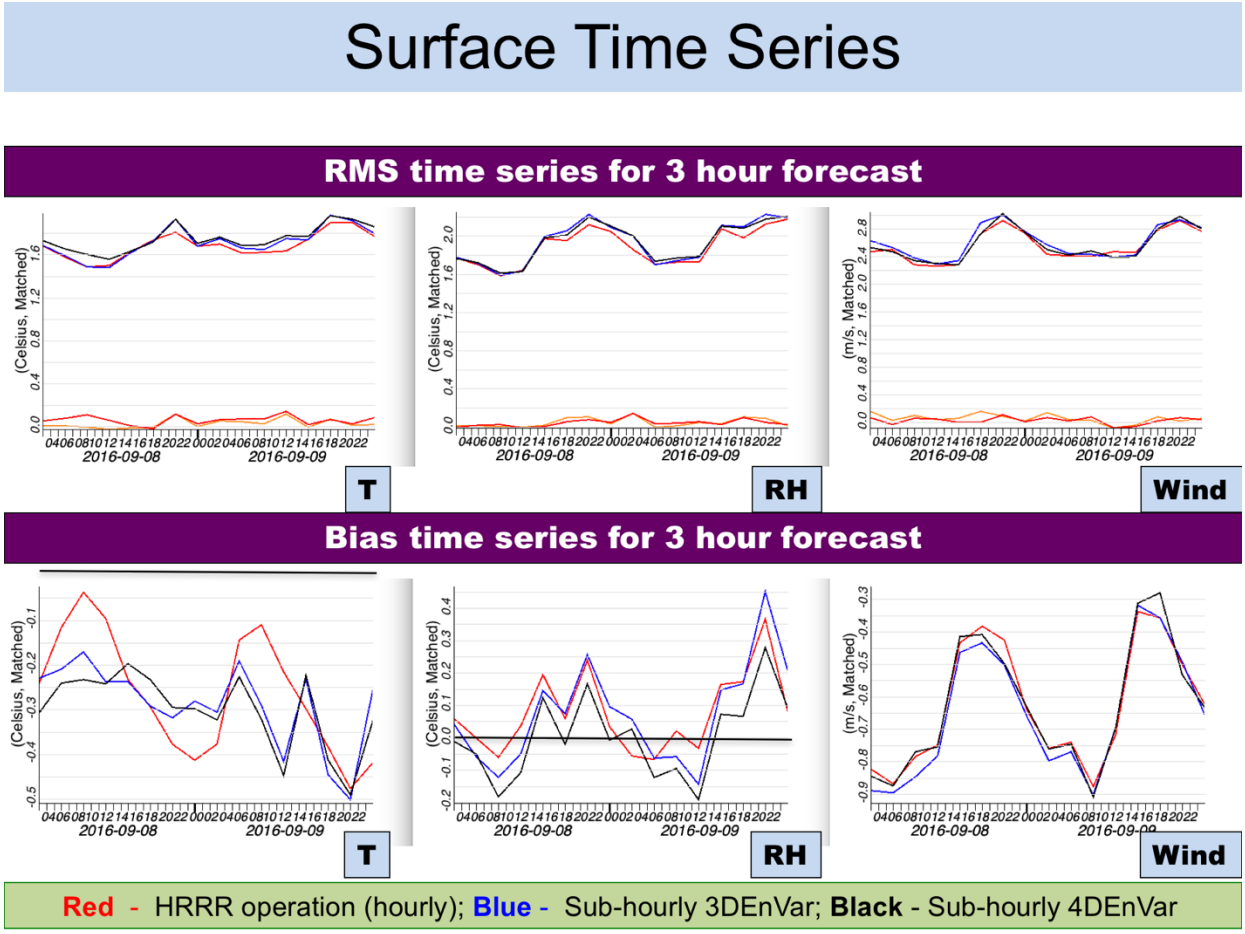

Figure 10. Time series of 2-m temperature (left), 2-m relative humidity (center), and 10-m wind (right) mean RMSE (top) and mean bias (bottom) for hourly 3DEnVar (red), sub-hourly 3DEnVar (blue), and sub-hourly 4DEnVar (black) for 3-hr forecasts. RMSE difference values are plotted in orange (sub-hourly 3DEnVar minus hourly 3DEnVar) and red (sub-hourly 4DEnVar minus hourly 3DEnVar).

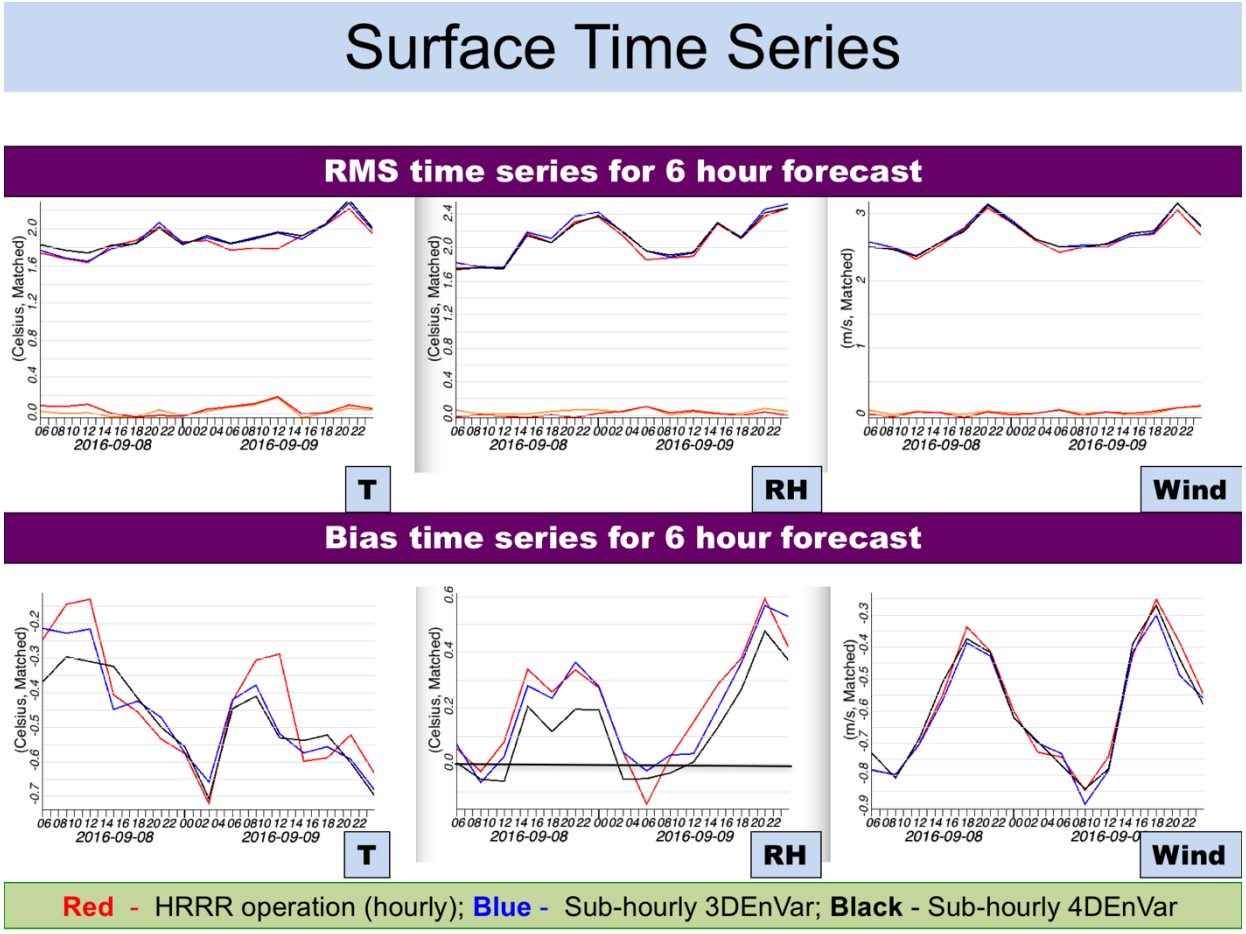

Figure 11. Time series of 2-m temperature (left), 2-m relative humidity (center), and 10-m wind (right) mean RMSE (top) and mean bias (bottom) for hourly 3DEnVar (red), sub-hourly 3DEnVar (blue), and sub-hourly 4DEnVar (black) for 6-hr forecasts. RMSE difference values are plotted in orange (sub-hourly 3DEnVar minus hourly 3DEnVar) and red (sub-hourly 4DEnVar minus hourly 3DEnVar).

Since ASOS and radiosonde observations are mainly influenced by synoptic patterns, mesoscale perturbations can often be missed when evaluating numerical models using standard verification. Therefore, it is also important to assess the ability of the sub-hourly simulations to correctly forecast convection and other small-scale features through other means. One first-order option is to compare forecasts of radar reflectivity with the observed radar observations. Therefore, a particularly active forecast was chosen for analysis within a selected area of the HRRR domain when a convective line was forecast.

Figure 12 shows the initialized radar reflectivity fields for 8 September 2016 at 1200 UTC for the three different experiments and the observed radar observations at the same time. Very little difference can be seen at initialization between the experiments, as well as when compared to the observations (aside from an underlying positive bias in the reflectivity values). A very close fit to the observed

reflectivity is therefore achieved. However, once the forecast is launched, differences can quickly be found between the experiments. By 1700 UTC (Figure 13), differences along the convective line become apparent, in particular in southwest Missouri (circled). The sub-hourly experiments both appear to qualitatively improve upon the operational, hourly simulation in this area, with increased reflectivity coverage. At 0000 UTC on the following day (12-hour forecast, Figure 14), improvements can be seen in southern Illinois (circled), with the sub-hourly experiments indicating a better match with the radar observations. The 4D hybrid EnVar experiment provides the best qualitative comparison with the observed reflectivity at this time. Overall, for most forecast hours, it was found that the sub-hourly experiments provided qualitative improvements of both coverage and intensity of radar reflectivity when compared with the operational simulation. In an effort to understand the impact of assimilating three different time levels within the sub-hourly 4D hybrid EnVar forecasts, plots of ensemble spread were generated for the GFS ensemble data used to initialize the 8 September 2009, 1200 UTC forecast (Figs. 15-17). Two pressure levels were plotted, ~475 and ~975 hPa, in order to assess the upper-level and near-surface meteorological features (the same domain was used as in previous plots analyzing this forecast). Figure 15 shows the GFS ensemble spread for the u-component of the wind, with enhanced areas of spread evident along the area where the squall line was seen at 1200 UTC in radar reflectivity (Figure 12). These areas are particularly apparent at ~975 hPa, with GFS ensemble forecast uncertainty (enhanced spread) also seen in the temperature field (Figure 13) over Missouri. The last GFS ensemble spread plot is for mixing ratio (Figure 17), and reveals a sharp increase in spread along the convective line at ~475 hPa. At ~975 hPa, the largest spread in mixing ratio is collocated with the same area of low-level enhanced temperature spread (Figure 16) over Missouri. These areas are likely related to ongoing convection with some or all of the members, but not necessarily in the same location, which produces uncertainty in mixing ratio/temperature values. These regions of enhanced spread inherently result in a larger impact for the observations within GSI, as there is larger uncertainty with the background forecast. In addition, the influence of the observations is spread along these regions of larger spread, augmenting the impact they have upon the final analysis field. The evolution of this uncertainty/spread is also a key difference between the two sub-hourly simulations, since it is captured within the three time levels assimilated with 4D hybrid EnVar.

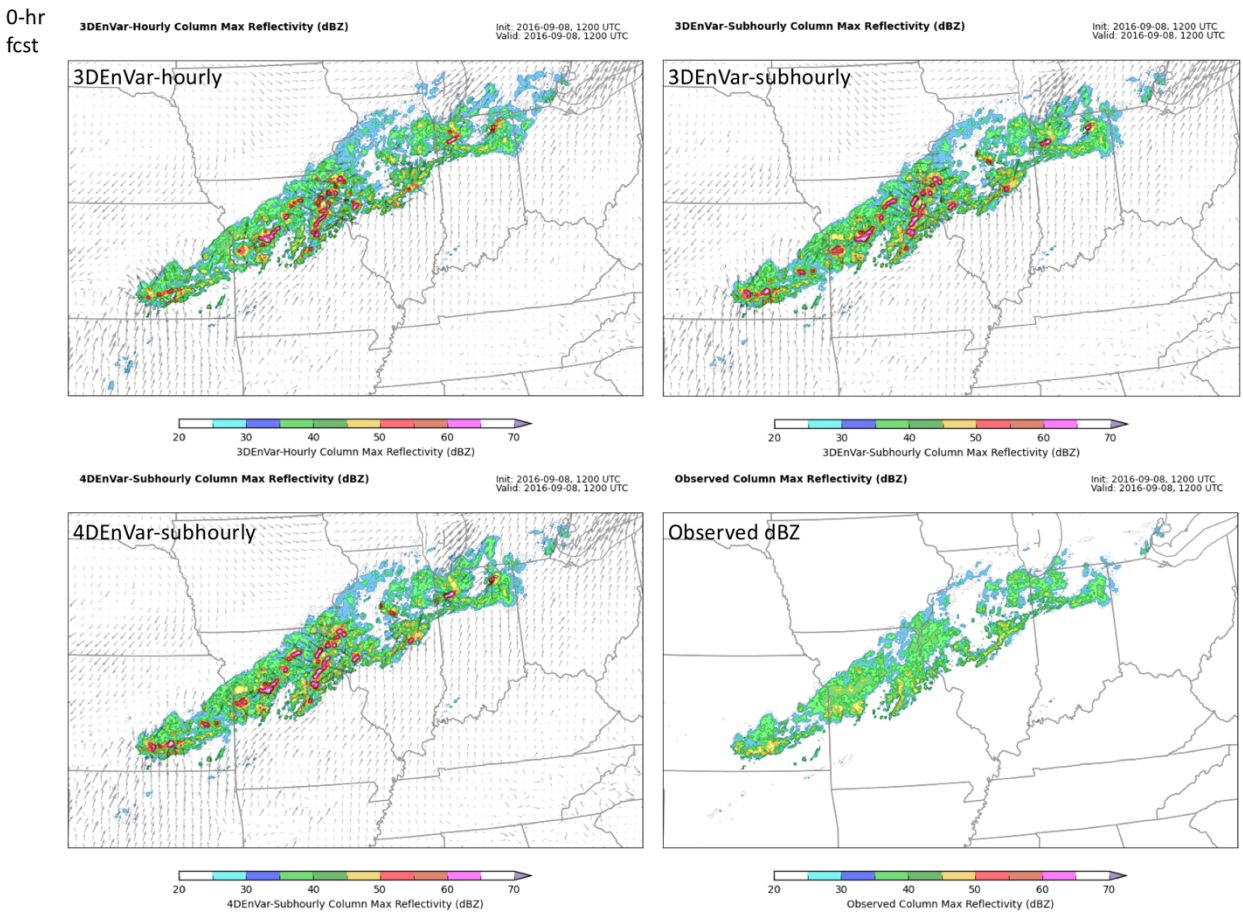

Figure 12. Analysis time (8 September 2016 at 1200 UTC) plots of column max reflectivity (dBZ) for hourly 3DEnVar (top, left), sub-hourly 3DEnVar (top, right), and sub-hourly 4DEnVar (bottom, left), with 10-m wind vectors for a selected region of the HRRR domain. Observed column max reflectivity (dBZ) is also shown (bottom, right).

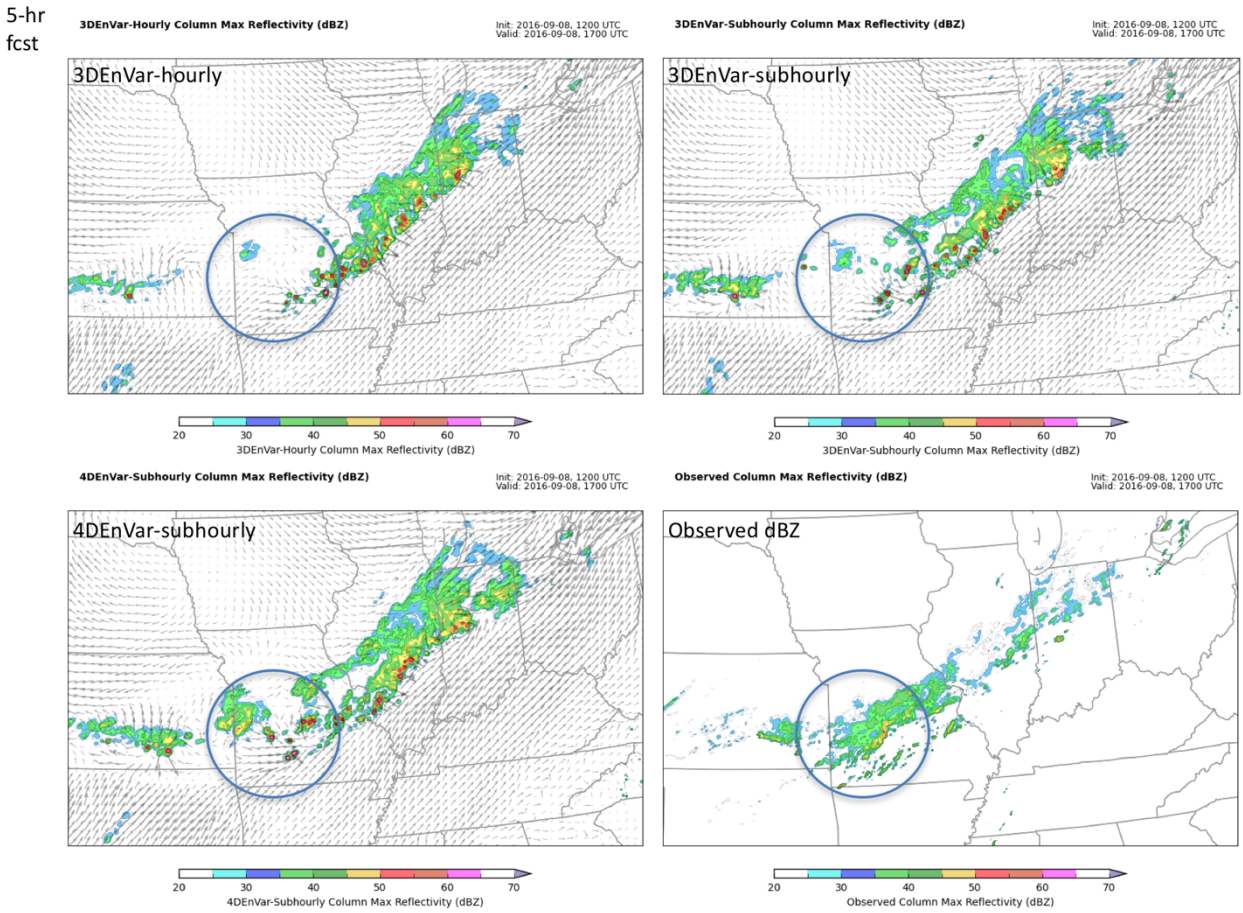

Figure 13. Five-hour forecast (8 September 2016 at 1700 UTC) plots of column max reflectivity (dBZ) for hourly 3DEnVar (top, left), sub-hourly 3DEnVar (top, right), and sub-hourly 4DEnVar (bottom, left), with 10-m wind vectors for a selected region of the HRRR domain. Observed column max reflectivity (dBZ) is also shown (bottom, right).

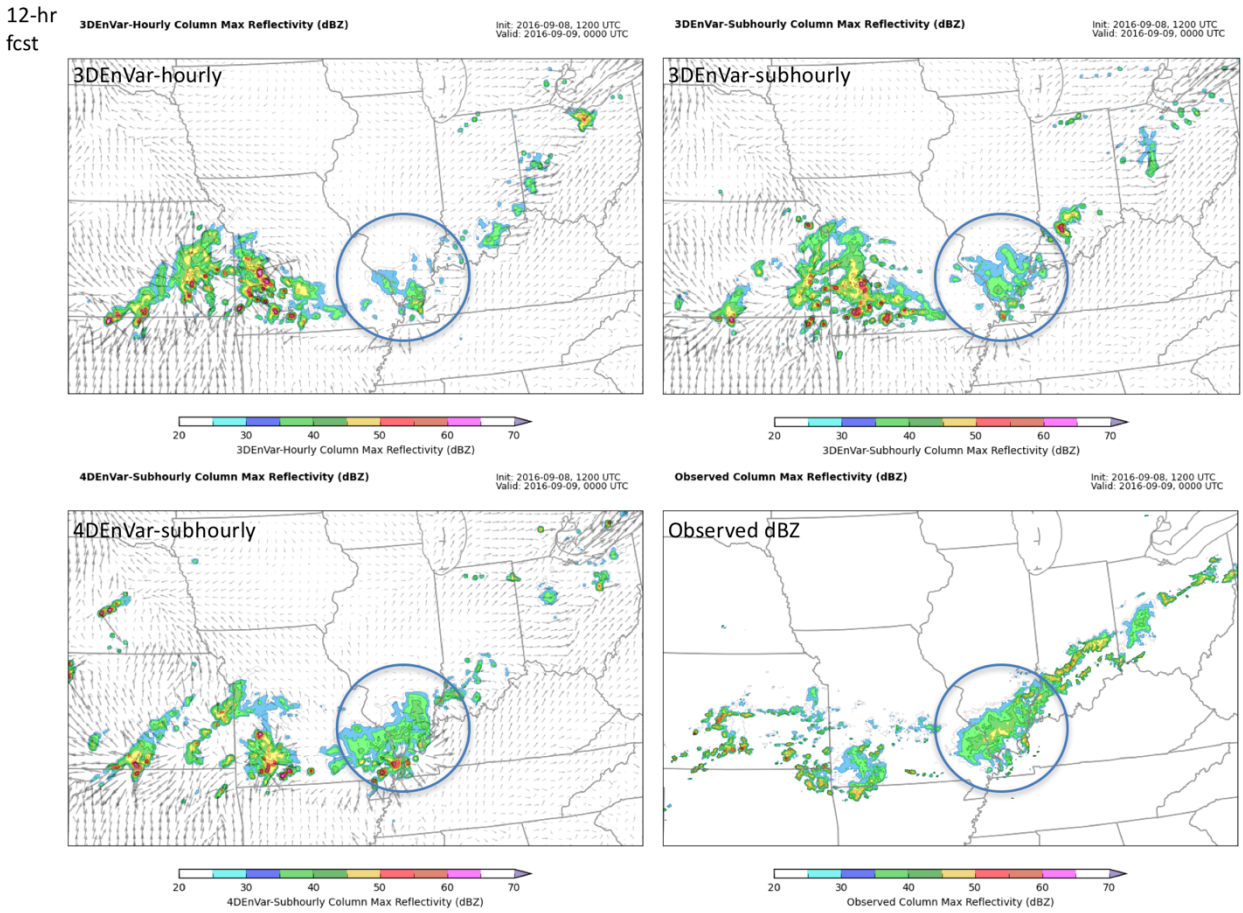

Figure 14. Twelve-hour forecast (8 September 2016 at 1700 UTC) plots of column max reflectivity (dBZ) for hourly 3DEnVar (top, left), sub-hourly 3DEnVar (top, right), and sub-hourly 4DEnVar (bottom, left), with 10-m wind vectors for a selected region of the HRRR domain. Observed column max reflectivity (dBZ) is also shown (bottom, right).

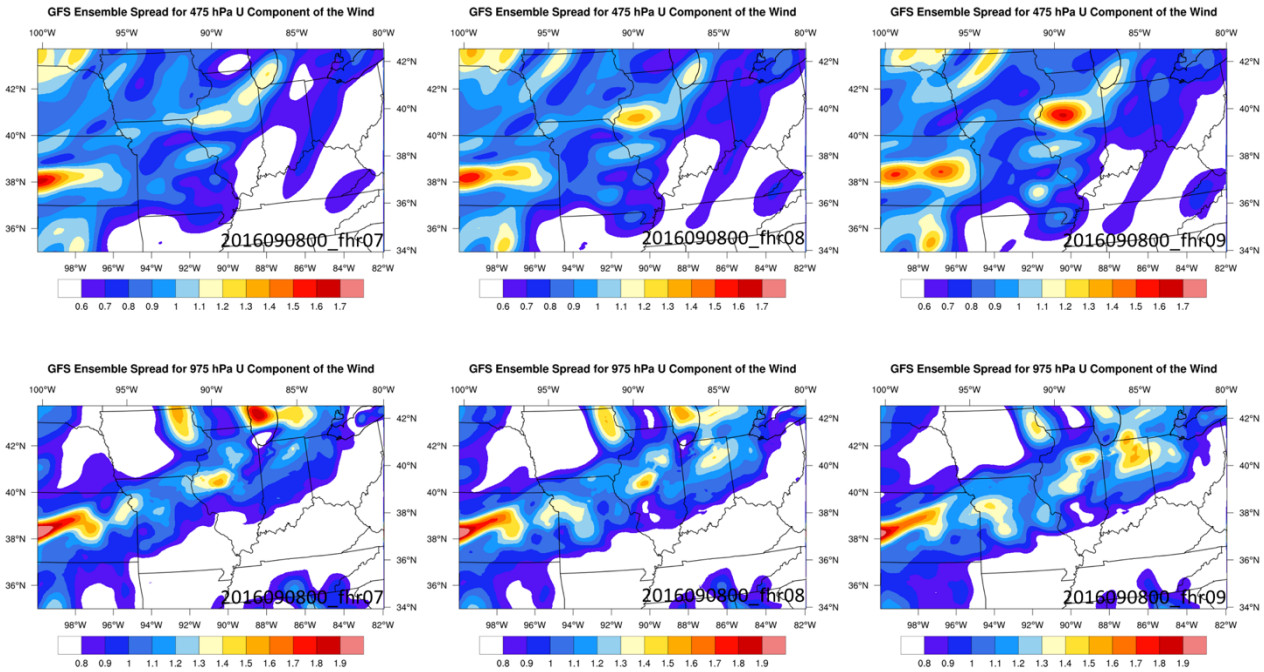

Figure 15. GFS ensemble spread for the U-component of the wind for the 7-h (left), 8-h (center), and 9-h (right) forecasts from the 8 September 2016 0000 UTC initialization. Plots are shown for ~475 hPa (top) and ~975 hPa (bottom) over the same selected region of the HRRR domain as in Figs. 12-14.

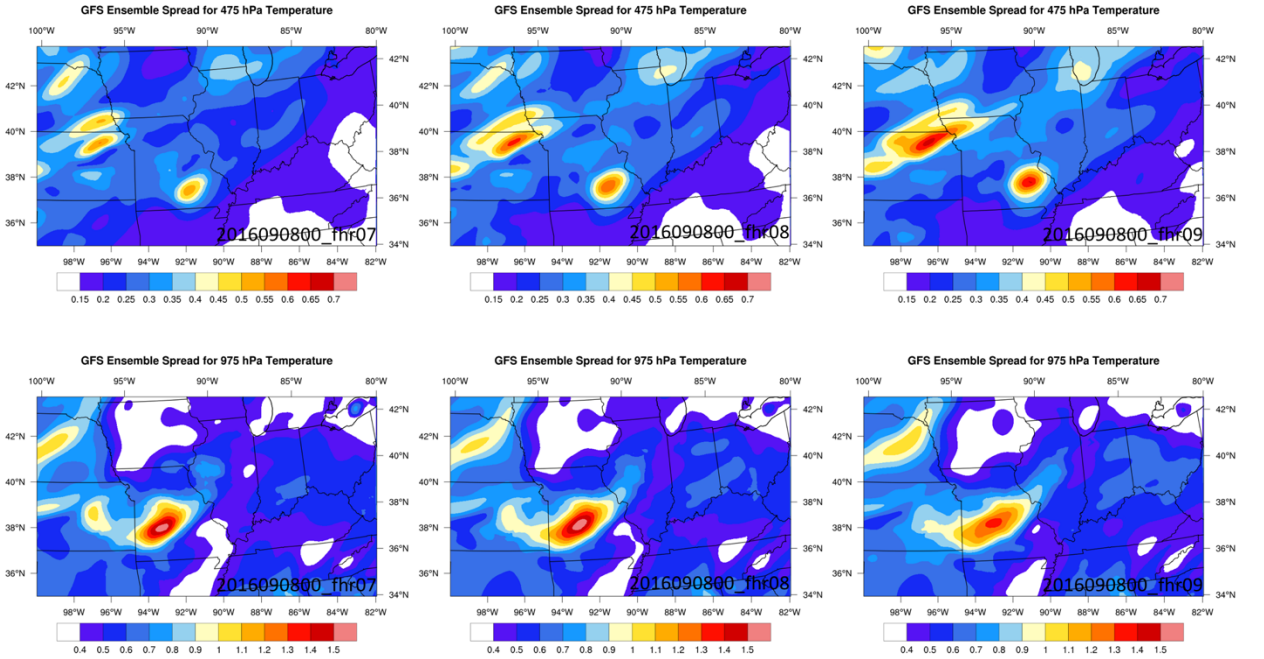

Figure 16. GFS ensemble temperature spread for the 7-h (left), 8-h (center), and 9-h (right) forecasts from the 8 September 2016 0000 UTC initialization. Plots are shown for ~475 hPa

(top) and ~975 hPa (bottom) over the same selected region of the HRRR domain as in Figs. 12-14.

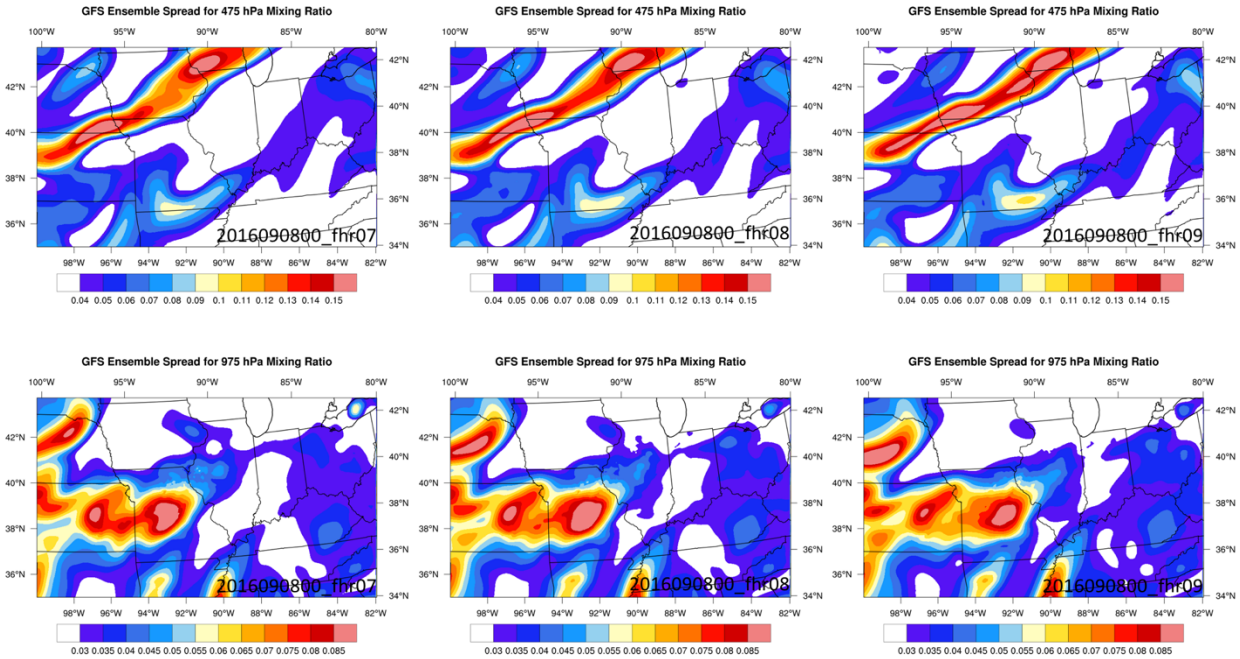

Figure 17. GFS ensemble mixing ratio spread for the 7-h (left), 8-h (center), and 9-h (right) forecasts from the 8 September 2016 0000 UTC initialization. Plots are shown for ~475 hPa (top) and ~975 hPa (bottom) over the same selected region of the HRRR domain as in Figs. 12-14.

Finally, ensemble spread from a regional WRF ensemble was analyzed over the same domain (Figs. 20). This ensemble was used by NCAR DTC colleagues to assess the impact of hourly cycling with a higher-resolution ensemble. Similarities can be seen between the GFS and WRF ensembles, however, much more detail is present in the WRF ensemble, with multiple filaments of spread likely associated with the convective line. Enhanced spread is also seen in pockets aloft for the u-component of the wind over northern Missouri (Figure 18), for temperature spread over Iowa (Figure 19), and for mixing ratio spread over Nebraska (Figure 20). It is possible that the ability to resolve these smaller-scale uncertainties in the WRF ensemble could provide a quantifiable impact on initialization and subsequent forecast fields, and would be a good candidate to test with the sub-hourly 3D and 4D hybrid EnVar simulations.

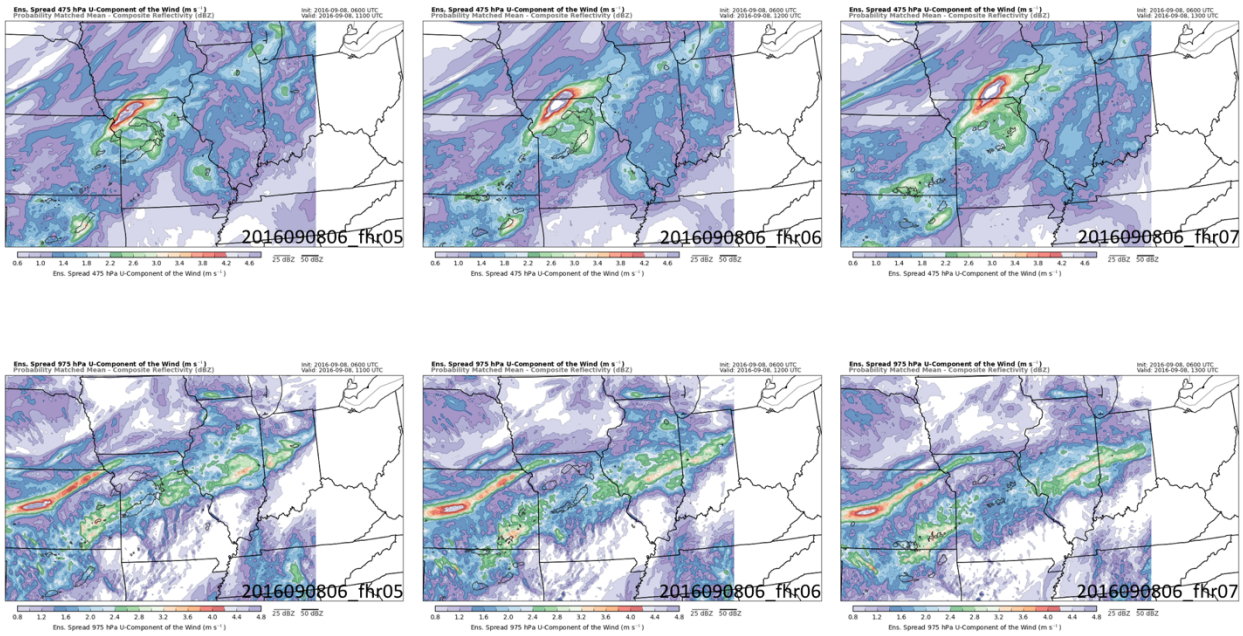

Figure. 18. WRF ensemble spread for the U-component of the wind for the 5-h (left), 6-h (center), and 7-h (right) forecasts from the 8 September 2016 0600 UTC initialization. Plots are shown for ~475 hPa (top) and ~975 hPa (bottom) over the same selected region of the HRRR domain as in Figs. 12-14.

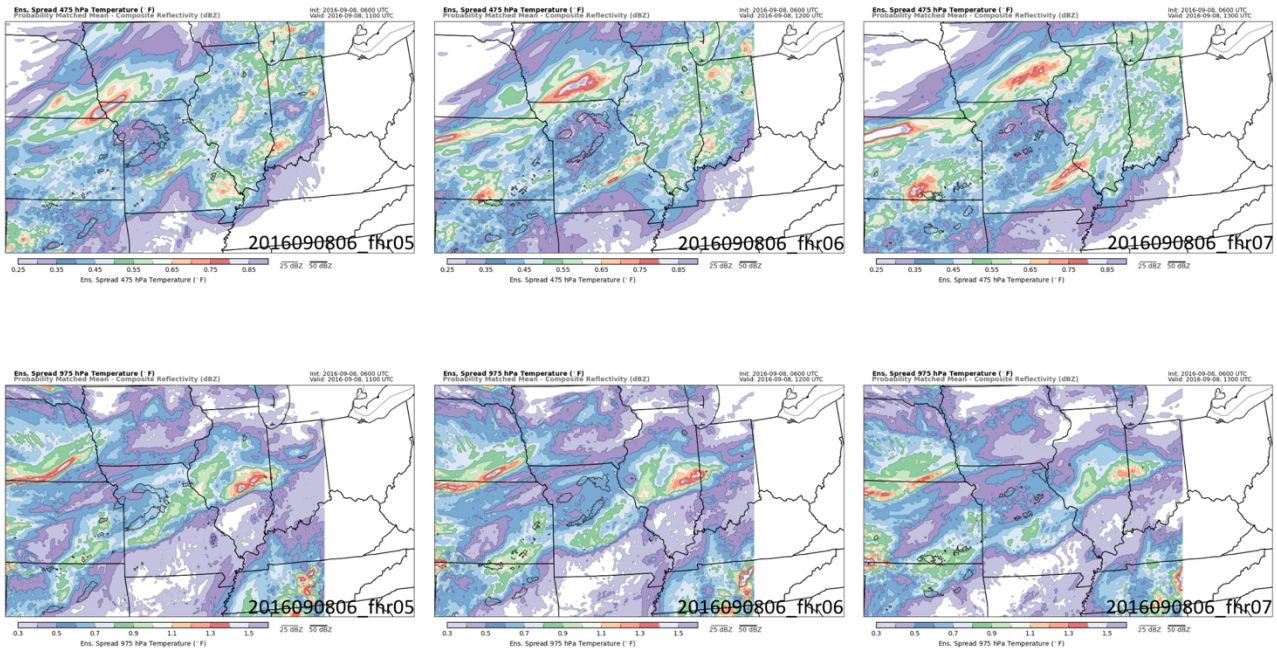

Figure 19. WRF ensemble temperature spread for the 5-h (left), 6-h (center), and 7-h (right) forecasts from the 8 September 2016 0600 UTC initialization. Plots are shown for ~475 hPa (top) and ~975 hPa (bottom) over the same selected region of the HRRR domain as in Figs. 12-14.

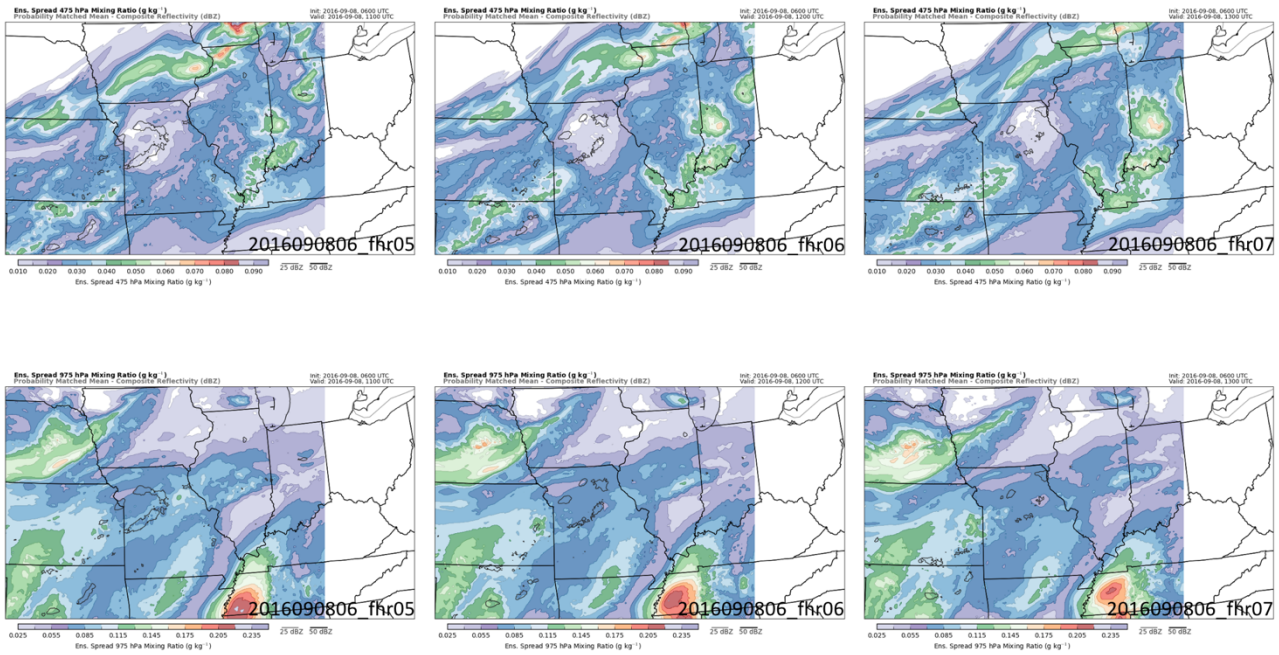

Figure 20. WRF ensemble mixing ratio spread for the 5-h (left), 6-h (center), and 7-h (right) forecasts from the 8 September 2016 0600 UTC initialization. Plots are shown for ~475 hPa (top) and ~975 hPa (bottom) over the same selected region of the HRRR domain as in Figs. 12-14.

3. Conclusion

Overall, implementation of the sub-hourly cycling has a positive impact on verification for initialization time and a more-or-less neutral impact on HRRR surface and upper-air verification compared to the operational hourly configuration. Both sub-hourly experiments show the potential for improved vertical-profile wind bias, but neutral to negative vertical profile temperature and relative humidity bias. The sub-hourly retrospective runs also indicate more stable surface-based error characteristics for bias, and are not as affected by the diurnal cycle as the hourly configuration (with the exception of winds).

Qualitative radar reflectivity comparison of the three experiments indicates a very close match to the observations at initialization time, with differences developing between the three with forecast lead time. In particular, certain areas of the analyzed convective line are better represented by the sub-hourly simulations in terms of coverage and intensity, with the 4D hybrid EnVar set up providing the best results at times. Many of these forecast improvements in qualitative radar reflectivity were found at lead times of 12 or more hours.

Regarding the neutral to slightly negative impacts of the sub-hourly simulations in the traditional verification, it is possible that the sub-hourly 3D and 4D hybrid EnVar configurations need to be tuned in terms of observation time windows for

assimilation, or the adjustment of observation error for certain fields. Modification of observation error for data collected between hours, particularly from surface stations and aircraft, could potentially benefit the sub-hourly configurations the most. It is also possible that such a rapidly updating assimilation cycle would benefit from a higher-resolution regional ensemble with more frequent forecast intervals, such as the WRF ensemble discussed above. While these hypotheses could be investigated, it is clear that some advantages to the sub-hourly configurations are apparent, particularly in terms of convective structure, coverage, and intensity, when comparing forecast results to the observed radar reflectivity.