The Developmental Testbed Center (DTC) Annual Operational Plan (AOP) 2017 Project for the National Oceanic and Atmospheric Administration (NOAA) Office of Oceanic and Atmospheric Research (OAR)

Testing DA advancement in observation aspects and evaluation for operational readiness

Final Report

June 2018

Data Assimilation Team

Table of Content

1. Introduction

In 2016, the Developmental Testbed Center (DTC) built a High Resolution Rapid Refresh version 2 (HRRRv2) functionally similar testing system on NCAR's supercomputer Yellowstone to provide the support for the research community to test and evaluate the possible contribution to operational storm-scale data assimilation and modeling. As part of contribution from DTC and initial application of such HRRRv2 operational functionally similar system, DTC Data Assimilation (DA) team conducted several experiments using this system on study to impacts of different Gridpoint Statistical Interpolation (GSI) hybrid Ensemble Variational (EnVar) options in HRRRv2. The 3-Dimensional and 4-Dimensional hybrid EnVar analysis with both hourly and sub-hourly cycling configuration were tested and results were reported in AOP 2016 T&E report.

In AOP 2017 T&E task for data assimilation, the HRRR version 3 (HRRRv3) operation system was proposed as the target of operational functional similar testing system and testing and evaluating radar radial wind analysis in HRRRv3 was proposed as community contributions from DTC DA team to the operational convective system. HRRRv3 is using radar radial velocity following the similar configuration used by Rapid Refresh (RAP) and DTC will investigate and tune the radial velocity analysis based on operational GSI configuration to help HRRR development team better understand and even configure the radial velocity data analysis in HRRR.

Section 2 of this report will describe the building of a HRRRv3 functionally similar testing system on NCAR Cheyenne for community users, including the system features and the verification of its functionally similar ability. The first step of data impact experiments is to understand the data and the operator of the data in the analysis. Section 3 explains several tools developed by DTC to check the radial velocity observations used in the HRRR and their superobed results. Also, in Section 3, we conducted a series of single case study to find the possible parameters for improving the GSI radial velocity analysis data. Section 4 introduces the experiments of testing and evaluating the impact of radial velocity on the HRRRv3 and section 5 describes the efforts to use MET as verification tool for the HRRRv3 Cheyenne system.

2. Building HRRRv3 functionally similar Testing System

2.1 HRRRv3

The NOAA operational HRRRv3 was developed by the NOAA Global Systems Division (GSD) on the NOAA development supercomputer platforms and is running in NOAA operational supercomputer in NCEP for operational testing and application. HRRRv3 will replace HRRRv2 to be US operational convective allowing modeling in middle July, 2018. HRRRv3 is an hourly

updated, cloud-resolving, convection-allowing atmospheric model running on 3-km grid covering US COUNS domain. The HRRRv3 uses the same code as the NOAA's Rapid Refresh system version 4 (RAPv4), including the Weather Research and Forecasting (WRF) Model version 3.8 with the hybrid vertical coordinates, WPS (WRF Preprocessing System) version 3.9 with some updates/modifications from GSD, and the GSD-modified version of the GSI (Gridpoint Statistical Interpolation) data assimilation and UPP (Unified Post Processing) systems.

2.2 Building HRRRv3 functionally similar system on Cheyenne

To access by the general research community, HRRRv3 needs to be runs and evaluated on the NCAR supercomputer Cheyenne, which was implemented in 2017 to replace the previous super computer Yellowstone. Cheyenne uses different compiling modules and environments from its predecessor Yellowstone and NOAA supercomputers, therefore a lot of efforts have been made to port the HRRRv3 system, including compiling the code with GSD options, debugging and fixing the WRF and GSI code in order for them to run successfully on Cheyenne.

In order for the experiments to be consistent with the runs conducted by NOAA colleagues on NOAA supercomputers, all the initial and boundary conditions (downscaled from the retrospective 13-km Rapid Refresh runs), and observations have been prepared by colleagues at NOAA/GSD and then transferred to Cheyenne. The run scripts and workflow were also changed for easily use by the research community. The details of preparing the HRRRv3 on NOAA machine for period from September 3th, 2016 to September 10th, 2016 is explained below:

Prepare observation data for experiments.

The RAPv4/HRRRv3 system will need the following observation files:

> *NCEP PrepBUFR files, satellite radiance files, satellite derived wind files, satellite clouds, snow observation files, national MRMS reflectivity mosaic files, NEXRAD level-2 radial wind files, GLD360 lightning files, aircraft observation files and Green Vegetable Fractions files.*

It will also need corresponding GFS forecasts and GDAS Ensembles. All these observation files, forecasts and ensembles for the experiment period were downloaded from different directories at the NOAA HPSS archive system.

Fill up missing radar data from the NCEP archive

It was found that quite a few MRMS reflectivity data (see below) were missing in the RAP archive at Theia HPSS. Since we will run HRRR retrospective experiments and reflectivity-based temperature tendency nudging is an important part of HRRR, considerable efforts were

made to fill up these missing radar reflectivity data from the NCEP operational data archive at Theia HPSS.

```
2016090318_30      2016090405_30      2016090414_45
2016090506_15      2016090505_30      2016090505_45      2016090505_60
2016090617_15      2016090618_15      2016090617_30
2016090618_30      2016090617_45      2016090618_45      2016090617_60
2016090618_60      2016090619_15      2016090705_30      2016090705_45
2016090705_60      2016090717_30      2016090717_45      2016090717_60
2016090706_15      2016090706_30      2016090706_45      2016090718_15
2016090718_30      2016090805_30      20160909_1715      20160909_1730
20160909_1745      20160909_1760      20160909_1815      20160909_1830
20160910_0415 ~2315      20160910_0430 ~2330
20160910_0445 ~2345   20160910_0460 ~2360
```

Run RAP to get ICs and BCs for HRRR experiments

Before running the HRRR retrospective experiments, we need to set up and run the RAPv4 system first to provide boundary conditions and initial conditions for HRRR runs. A full RAP domain retrospective experiment was conducted using the same parameter configuration as operational RAPv4 for the period. When the RAP retro was completed, the results were verified against radiosonde and METAR observations using the verification package developed by the ESRL/GSD verification group (MATS). The verification statistics were compared with GSD real-time RAPv4 runs to make sure the RAPv4 system was correctly installed and configured. The statistics can be examined by visiting MATS system at:
https://www.esrl.noaa.gov/gsd/mats/.

It was shown that the RAP retro was not significant different from the semi-operational RAPv4 running at the Global System Division. As an example, Fig. 2.1 show the temperature RMS profiles for the GSD real-time RAPv4 (blue) and our RAP retro (red)

Error Bars show 95% confidence
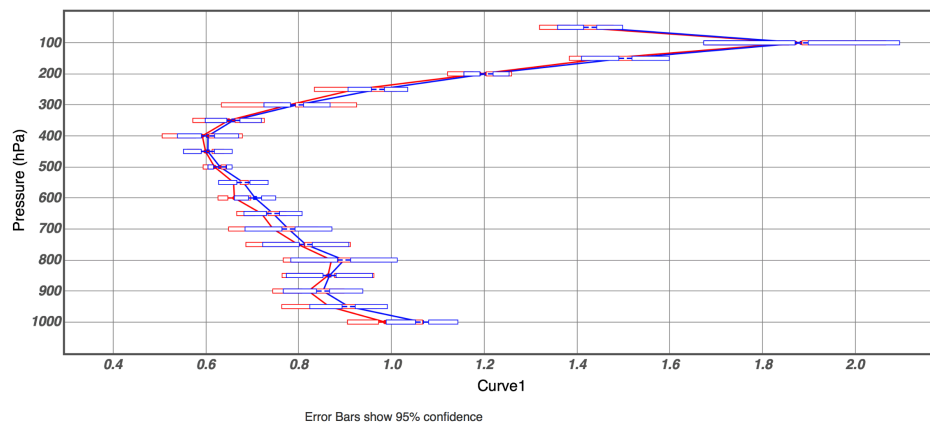
Figure 2.1, The temperature RMS profile for 0-h forecast of real-time RAPv4 (blue) and our RAP retro experiment (red)

Modify rocoto xml to best suit research uses

The RAP/HRRR system uses the ROCOTO software to control its workflow. ROCOTO utilize an XML file to define all running tasks and task dependencies. RAP/HRRR contains

hundreds of tasks and the semi-operational run at GSD wraps all tasks in one big XML file usually called *RAP_primary.xml* or *HRRR_primary.xml*. This is good for real-time running but not convenient for retrospective experiments in research. Retrospective experiments are characterized by more frequent changes to the XML files in order to test some specific tasks such as filling missing observations shown in the above, running a series of experiments with different data assimilation settings but with the same set of observational data. In this regard, we modified the all-in-one big XML file to a more structured format. Every task/meta-task in the RAP/HRRR workflow is described in an independent task file under the directory tasks/. Taking HRRR as an example, Fig. 2.2 shows the list of tasks in the HRRR workflow. This structured xml files make changing and checking given tasks much easier.

```
[Theia|tfe06 tasks]$ ls
conventional_gsi_early.xml   metgrid_RAP_ic.xml    radar_gsi.xml           ungrib_RAP_ic_early.xml
conventional_gsi.xml         ncl_HH.xml            radar_subh_tten.xml     ungrib_RAP_ic.xml
gsi_diag.xml                 nclpost.xml           real_arw_bc_long.xml    ungrib_SST.xml
gsi_hyb.xml                  ncltest.xml           real_arw_bc_short.xml   wrf_arw_long.xml
HRRR_soilsave_HH.xml         ncl_zip_HH.xml        real_arw_ic.xml         wrf_arw_pre.xml
lightning_gsi_MM.xml         post_00.xml           satellite_gsi.xml       wrf_arw_short.xml
metgrid_RAP_bc_long.xml      post_HH.xml           ungrib_RAP_bc_long.xml
metgrid_RAP_bc_short.xml     radar_gsi_MMSS.xml    ungrib_RAP_bc_short.xml
```

Figure 2.2. List of tasks in the HRRR workflow for community HRRRv3 system

With the above change, the main XML file becomes more readable. At the same time, we separate the HRRR retro workflow into two parts. The first part is named as "prep.xml" and included observation, initial conditions and boundary conditions processing tasks. It can be seen from Fig. 2.3 that it is very easy to add or remove specific tasks from the workflow. The second part is named as "prod.xml" and run data assimilation, WRF forecasts, post-processing and figure plotting tasks. For different retrospective experiments, we only need to run the second part of HRRR workflow, i.e. "prod.xml". All workflow related environmental variables are set in a file called "env.ent".

```
<cycledef group="12hr">00 00,12 29-31 03 2018 *</cycledef>
<cycledef group="01hr">00 01,13 29-31 03 2018 *</cycledef>
<cycledef group="02hr">00 02,14 29-31 03 2018 *</cycledef>
<cycledef group="03hr">00 03,15 29-31 03 2018 *</cycledef>
<cycledef group="04hr">00 04,16 29-31 03 2018 *</cycledef>
<cycledef group="05hr">00 05,17 29-31 03 2018 *</cycledef>
<cycledef group="06hr">00 06,18 29-31 03 2018 *</cycledef>
<cycledef group="07hr">00 07,19 29-31 03 2018 *</cycledef>
<cycledef group="08hr">00 08,20 29-31 03 2018 *</cycledef>
<cycledef group="09hr">00 09,21 29-31 03 2018 *</cycledef>
<cycledef group="10hr">00 10,22 29-31 03 2018 *</cycledef>
<cycledef group="11hr">00 11,23 29-31 03 2018 *</cycledef>

&ungrib_SST; <!-- every hour -->

&ungrib_RAP_bc_long; <!-- 03, 06, 09, 12 -->
&metgrid_RAP_bc_long; <!-- 03, 06, 09, 12 -->
&real_arw_bc_long; <!-- 03, 06, 09, 12 -->

&ungrib_RAP_bc_short; <!-- 01,02, ,04,05, ,07,08, ,10,11, , -->
&metgrid_RAP_bc_short; <!-- 01,02, ,04,05, ,07,08, ,10,11, , -->
&real_arw_bc_short; <!-- 01,02, ,04,05, ,07,08, ,10,11, , -->

&ungrib_RAP_ic; <!-- 02,03,04,05,06,07,08,09,10,11,12 -->
&ungrib_RAP_ic_early; <!-- 01 -->
&metgrid_RAP_ic; <!-- every hour -->
&real_arw_ic; <!-- every hour -->

&radar_gsi_MMSS; <!-- every hour -->

&radar_gsi; <!-- every hour -->
&satellite_gsi; <!-- every hour -->
&conventional_gsi; <!-- every hour -->
&conventional_gsi_early; <!-- 12 -->

</workflow>
```

Fig. 2.3 An example of the HRRR workflow "prep.xml" file

```
<cycledef group="12hr">00 00,12 30 03 2018 *</cycledef>
<cycledef group="01hr">00 01,13 30 03 2018 *</cycledef>
<cycledef group="02hr">00 02,14 30 03 2018 *</cycledef>
<cycledef group="03hr">00 03,15 30 03 2018 *</cycledef>
<cycledef group="04hr">00 04,16 30 03 2018 *</cycledef>
<cycledef group="05hr">00 05,17 30 03 2018 *</cycledef>
<cycledef group="06hr">00 06,18 30 03 2018 *</cycledef>
<cycledef group="07hr">00 07,19 30 03 2018 *</cycledef>
<cycledef group="08hr">00 08,20 30 03 2018 *</cycledef>
<cycledef group="09hr">00 09,21 30 03 2018 *</cycledef>
<cycledef group="10hr">00 10,22 30 03 2018 *</cycledef>
<cycledef group="11hr">00 11,23 30 03 2018 *</cycledef>

&radar_subh_tten;
&wrf_arw_pre;
&gsi_hyb;

&wrf_arw_long;
&wrf_arw_short;
&HRRR_soilsave_HH;

&post_00;
&post_HH;
&ncl_HH;
&ncl_zip_HH;

</workflow>
```

Fig. 2.4 An example of the HRRR workflow "prod.xml" file

The control and sanity check experiment

After prepared all the data and system for RAPv4/HRRRv3, ran two experiments to make sure the HRRRv3 functionally similar system can produce testing results that can be directly used to evaluate for operational implementations

The HRRRv3 functional similar system following the default configuration of the operational HRRRv3. Radar reflectivity data is assimilated every 15 minutes over a 1-hour pre-forecast period to generate background file that included with initialize storms. This updated field is then used for the GSI 3D EnVar analysis at 12-km grid and GSD cloud analysis at 3-km grid thereafter. The radar radial velocity data are assimilated at the 12-km grid, using the default superob parameters as listed in table 2.1, with the azimuth being 5 degrees and superob range being 5km (max range 100km), time window being +/-0.5 hour and minimum number of observations in the box being 50.

Table 2.1 Default superob parameters for the HRRR control run

Del_azimuth	Del_range	Range_max	Del_time	minnum	Del_elev
5 degrees	5km	100km	0.5	50	0.25

a. HRRR domain vs experiment domain

The full HRRR domain covers the full CONUS area and contains 1800x1100 horizontal grid points which is very computational expensive. In order to run experiments faster, we reduce the model domain to cover most interested part - the middle part of CONUS centered at great plain where spawns most convective storms in the experiment period. It contains 650x550

horizontal grid points (see Fig. 2.5). It is about 6 times smaller than the full HRRR domain and makes HRRR experiments run much faster.
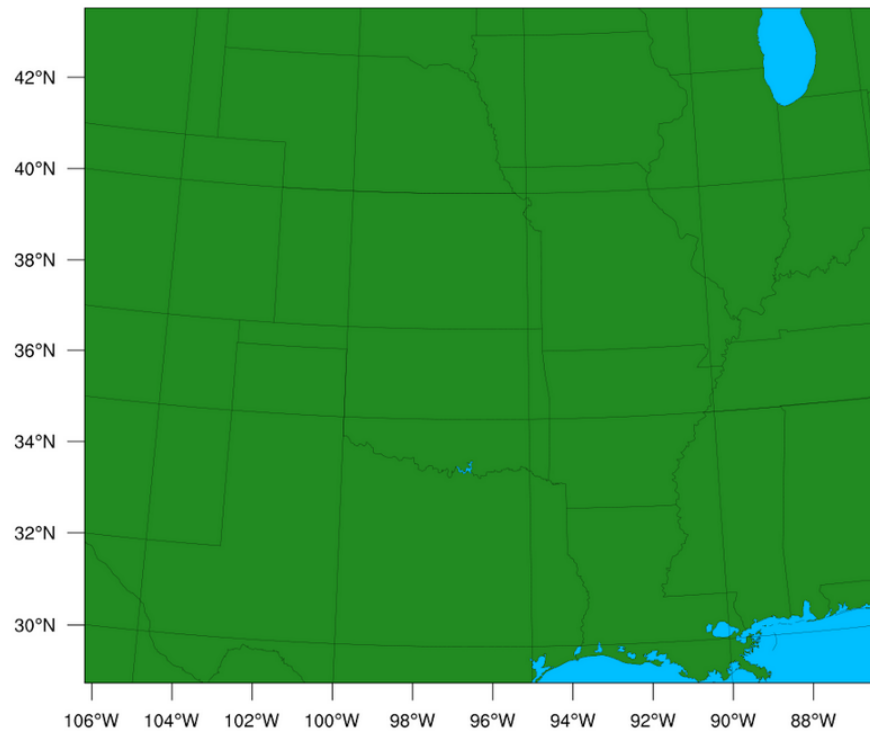

Figure. 2.5 The reduced HRRRv3 domain for experiments

To make sure the full domain HRRRv3 and reduced domain HRRRv3 are producing similar results, we ran a reduced domain HRRRv3 experiment for one day, generated verification statistics and compared them with real-time full domain HRRRv3 at GSD. Results show these are no significant differences between the two runs. This confirmed that the reduced domain HRRRv3 was successfully set up on Theia and can be used to replace full HRRR domain runs to save computation time.

b. Building HRRRv3 on Cheyenne (NCAR) and sanity check

After RAPv4/HRRRv3 was successfully setup at the NOAA supercomputer Theia and all observations processing, initial and boundary conditions were completed, the system were ported to the UCAR supercomputer Cheyenne. This makes HRRRv3 available for community users.

NCAR colleagues repeat the HRRRv3 control experiment, which is a one-week long HRRRv3 runs for September 3-10 2016 on Thiea, on the UCAR supercomputer Cheyenne. The post-processed model outputs were then transferred to the NOAA supercomputer Theia where the data can be loaded to the NOAA's verification system – MATS (Model Analysis Tool Suite) and then compared to the same control run conducted on Theia. This comparison is to ensure the functional similarity between the two computers and minimal divergence between the two control runs. Fig. 2.6 shows the comparison of the two control runs, in terms of the vertical

profile of the temperature root mean squared errors (RMS), for forecast hours 0, 1, 3, 6, 9 and 12. As can be seen from the almost overlapping curves and error bars (95% confidence), the two control runs perform very similar despite the differences from the computer platforms and compiling environment and etc. Same features can be observed in the vertical profiles of temperature bias (model minus observation), as in Fig. 2.7. This functional similarity of HRRR performance between the two platforms is observed across the variables, as shown in Fig. 2.8 for the vertical profiles of the wind RMS, Fig. 2.9 for the wind bias, Fig. 2.10 for the RMS of relative humidity (RH), and Fig. 2.11 for the bias of the relative humidity. Fig. 2.12 and 2.13 show the time series of the surface dewpoint RMS and bias respectively. It confirms the findings from the vertical profiles that the two control runs perform very similarly, in terms of both vertical profiles and surface time series. Therefore, it can be concluded that porting the HRRR system from Theia to Cheyenne doesn't cause the system to behave differently and it is ready for further experiments.
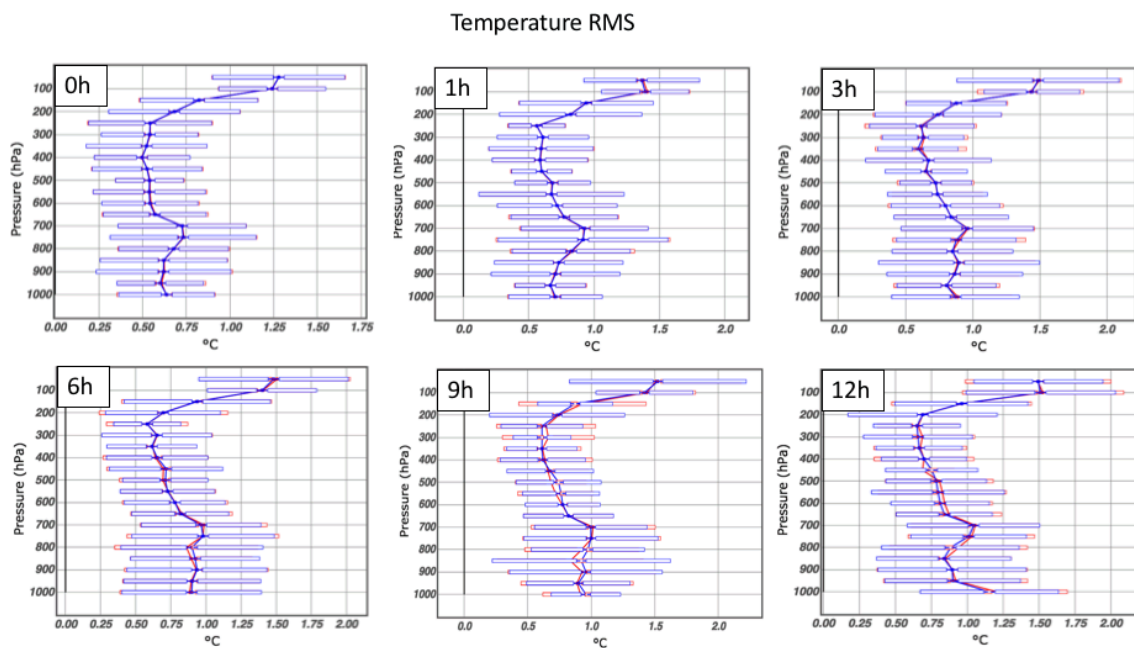

Figure 2.6 Vertical profiles of temperature root mean squared errors (RMS) from the control runs on Cheyenne (red) and Theia (blue) for forecast hours 0, 1, 3, 6, 9 and 12.
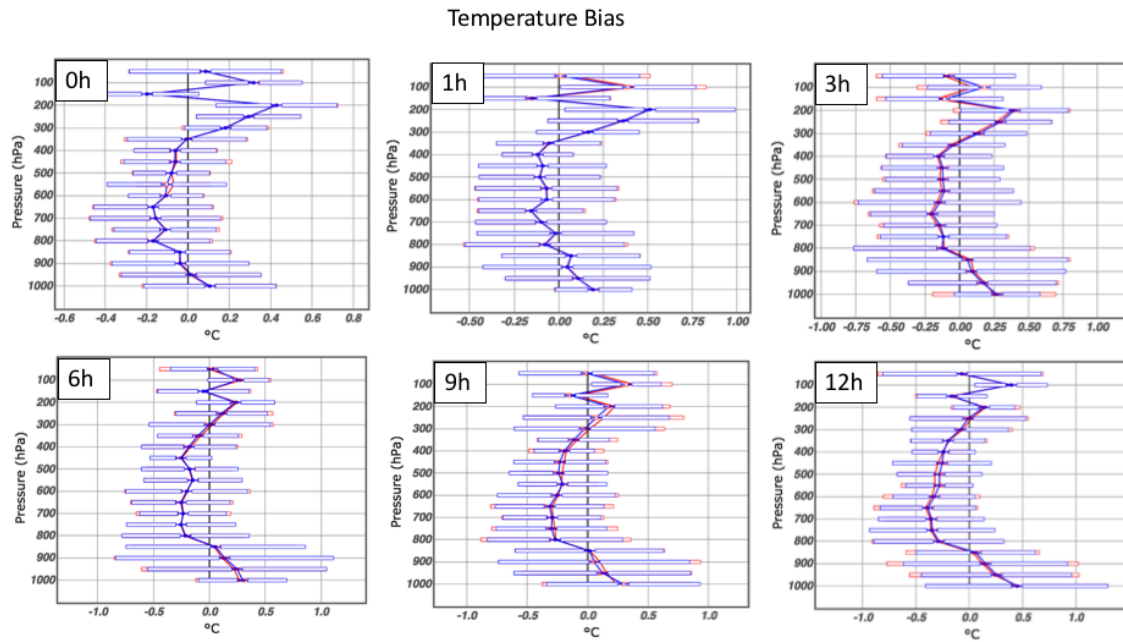
Temperature Bias

Figure 2.7 Vertical profiles of temperature bias (Model-Observation) from the control runs on Cheyenne (red) and Theia (blue) for forecast hours 0, 1, 3, 6, 9 and 12.
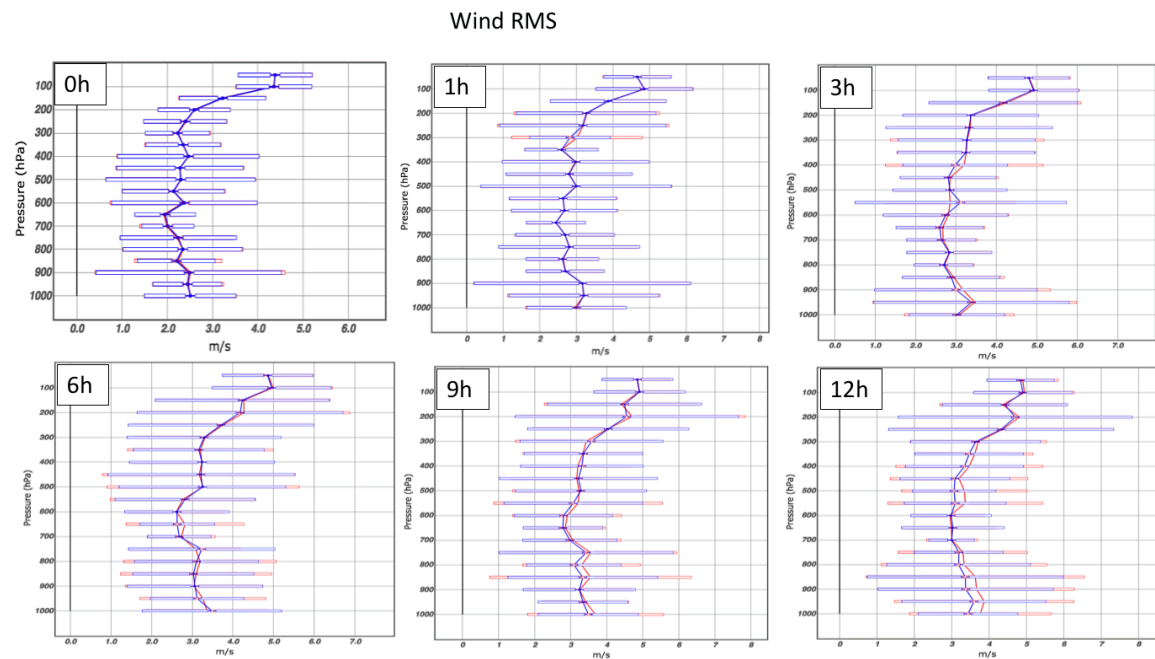
Wind RMS

Figure 2.8 Vertical profiles of wind root mean squared errors (RMS) from the control runs on Cheyenne (red) and Theia (blue) for forecast hours 0, 1, 3, 6, 9 and 12.

Wind Bias

Figure 2.9 Vertical profiles of wind bias (Model-Observation) from the control runs on Cheyenne (red) and Theia (blue) for forecast hours 0, 1, 3, 6, 9 and 12.

Cheyenne (red) and Theia (blue) for forecast hours 0, 1, 3, 6, 9 and 12.

RH RMS

Figure 2.10 Vertical profiles of RH root mean squared errors (RMS) from the control runs on Cheyenne (red) and Theia (blue) for forecast hours 0, 1, 3, 6, 9 and 12.
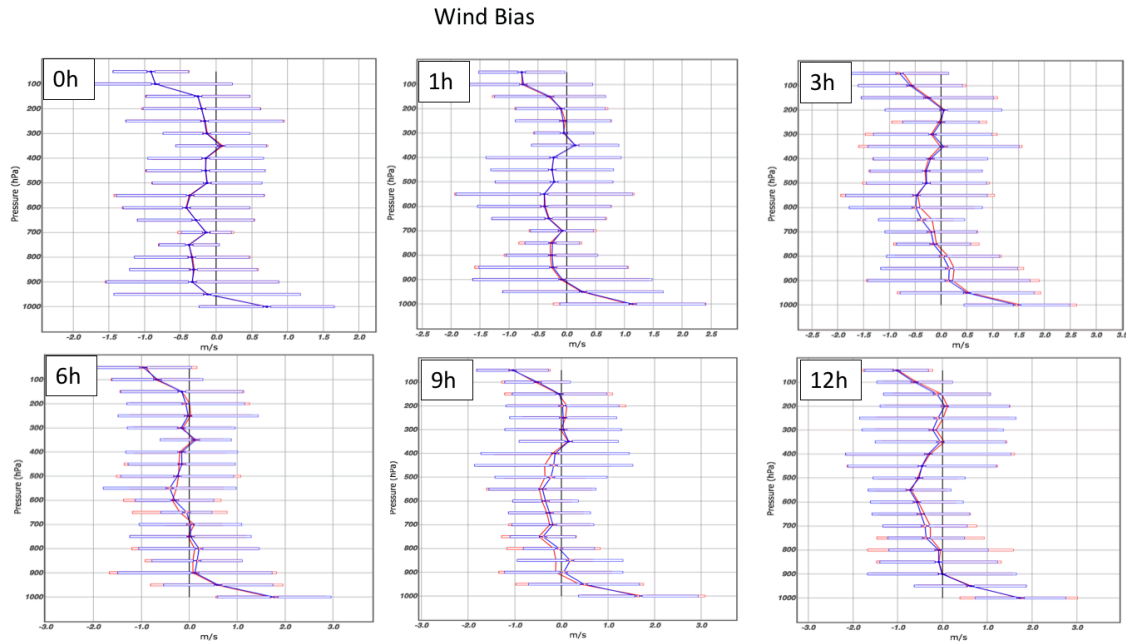
RH Bias

Figure 2.11 Vertical profiles of RH bias (Model-Observation) from the control runs on Cheyenne (red) and Theia (blue) for forecast hours 0, 1, 3, 6, 9 and 12.
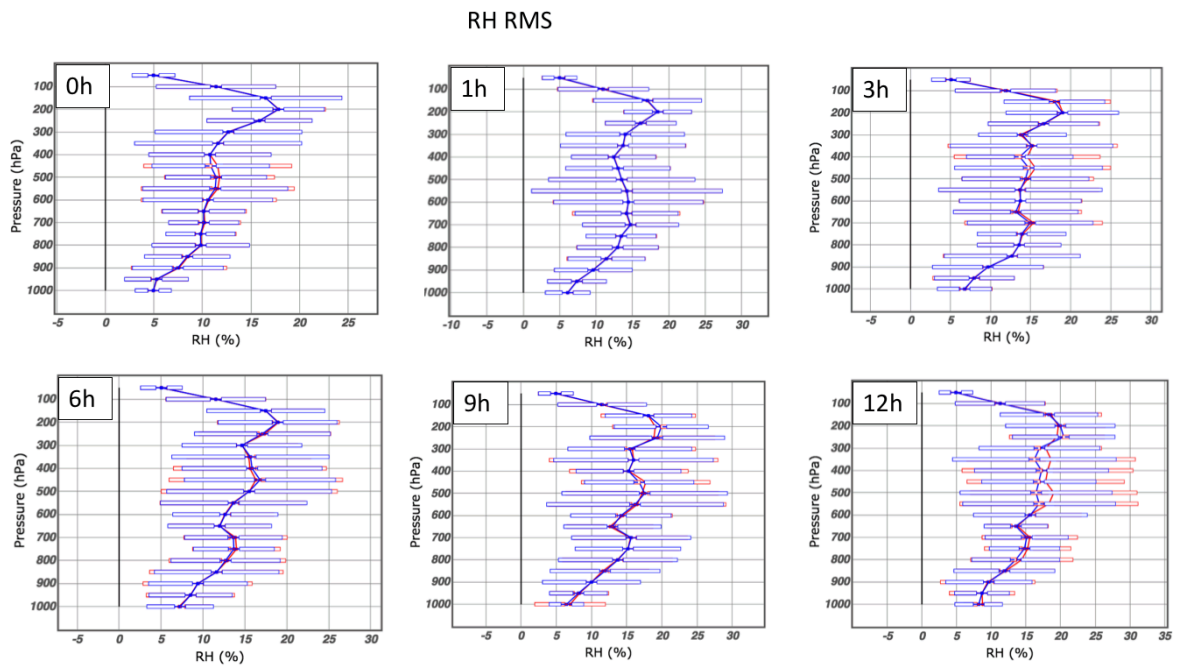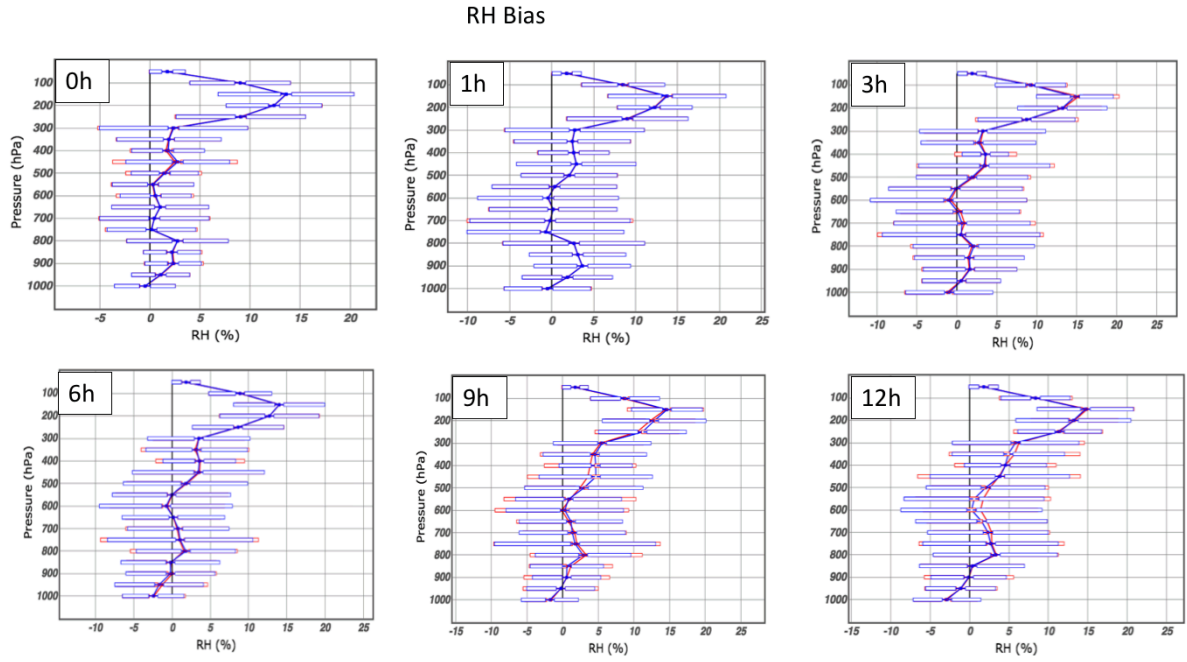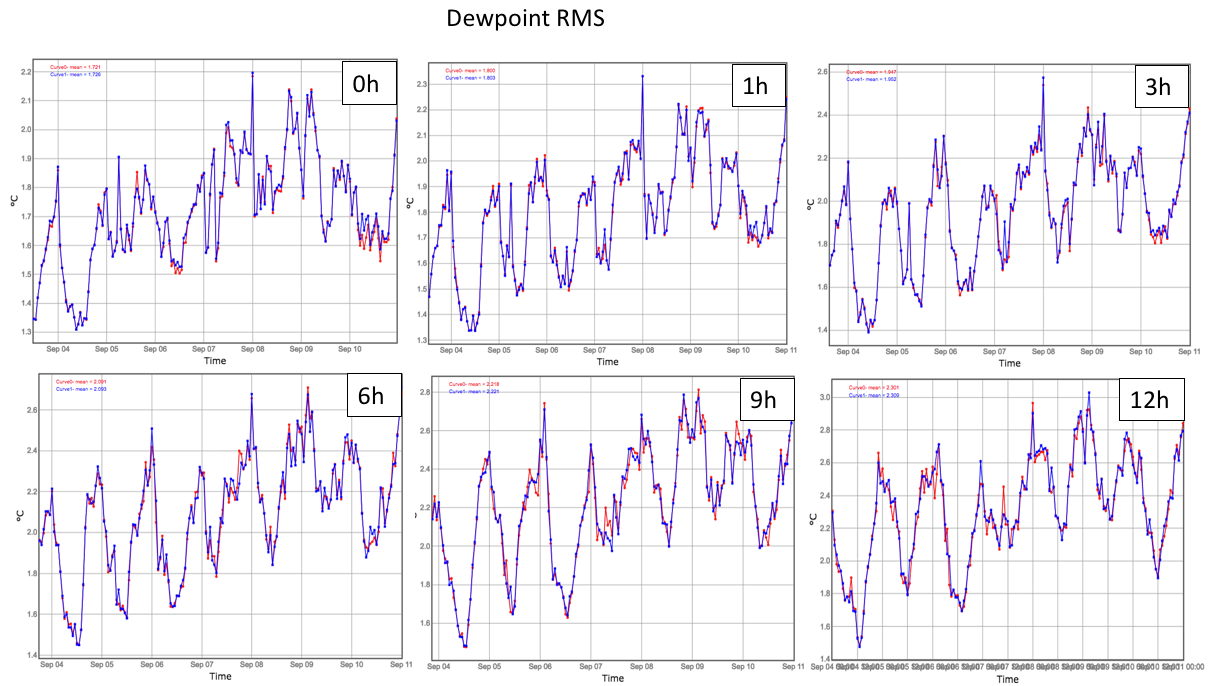
Dewpoint RMS

Figure 2.12 Time series of surface dewpoint root mean squared errors (RMS) from the control runs on Cheyenne (red) and Theia (blue) for forecast hours 0, 1, 3, 6, 9 and 12.
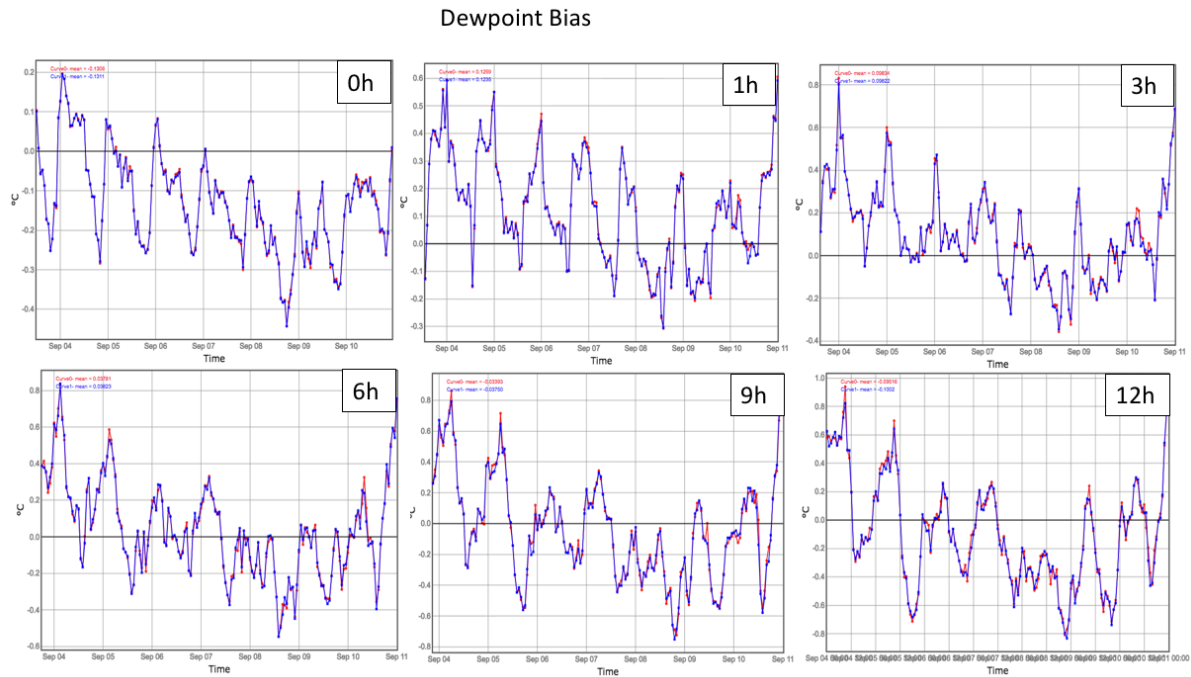
Dewpoint Bias

Figure 2.13 Time series of surface dewpoint bias (Model-Observation) from the control runs on Cheyenne (red) and Theia (blue) for forecast hours 0, 1, 3, 6, 9 and 12.

3. Radar Radial Velocity Data Check

The impact of radial velocity on HRRR forecast is the focus of AOP 2017 T&E tasks. Before doing data assimilation and forecast experiments, it is very helpful to take a close look at the raw radial velocity data use in the GSI analysis and understand the data coverage and quality.

3.1 Check NCEP level II radial velocity BUFR data examination

The radial velocity data used in the RAP/HRRR are provided hourly in BUFR format by NCEP. Each BUFR file contains radial velocity observation from all CONUS NEXRAD sites available in NCEP data tank within a time window [-30 minutes, 30 minutes] centered at the hour. A utility was developed to decode the BUFR file and write out radial velocity data in the gate, azimuth, elevation order to a text file for further examination. NCL scripts were also developed to plot BUFR radial velocity data in polar coordinates for visual checking. The corresponding raw NEXRAD Level II data was obtained from Amazon radar archive and plotted using the IDV software from Unidata.

The figures plotted from RAW files and NCEP BUFR files are compared against each other. It was found that the NCEP BUFR files missed a lot of radial wind data in some cases as

compared to the raw data files during the experiment period. Fig. 3.1 showed the radial winds at the 3.1˚ elevation angle for KVNX at 20160910 0000 UTC. The left panel is from the raw data file, the right panel is from the NCEP BUFR file. It is very clear that lots of radial velocity data were missed in the NCEP BUFR file. Fig. 3.2 showed the radial winds at the 0.5˚ elevation angle for KICT at 20160910 0021 UTC. Similar data missing issue showed for this case. We also checked many other cases and found the NCEP BUFR file does have a large portion of radial velocity data in some case, such as one showing in Fig. 3.3 (1.3˚ elevation angle for KICT at 20160910 0021 UTC).
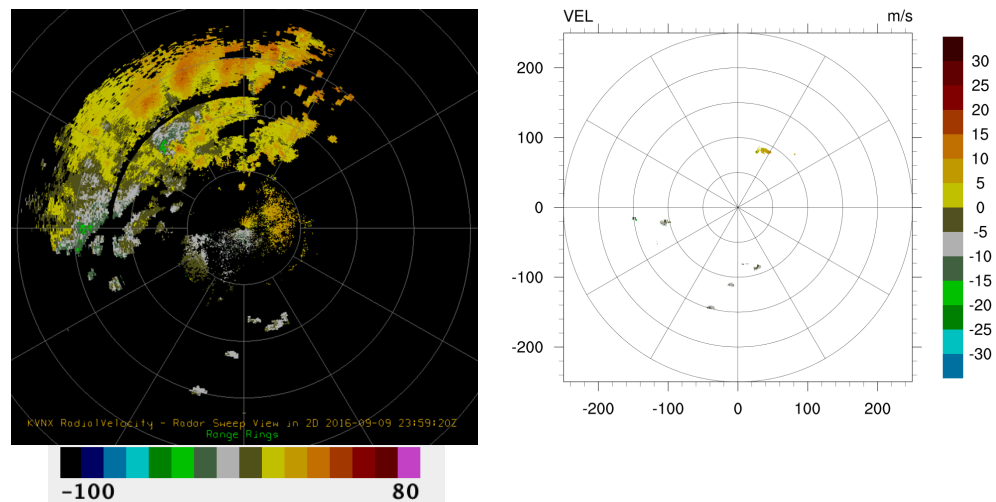

Figure 3.1 The radial winds at the 3.1° elevation angle for KVNX at 0000UTC 20160910. Left panel from raw data file and right panel from NCEP BUFR file.
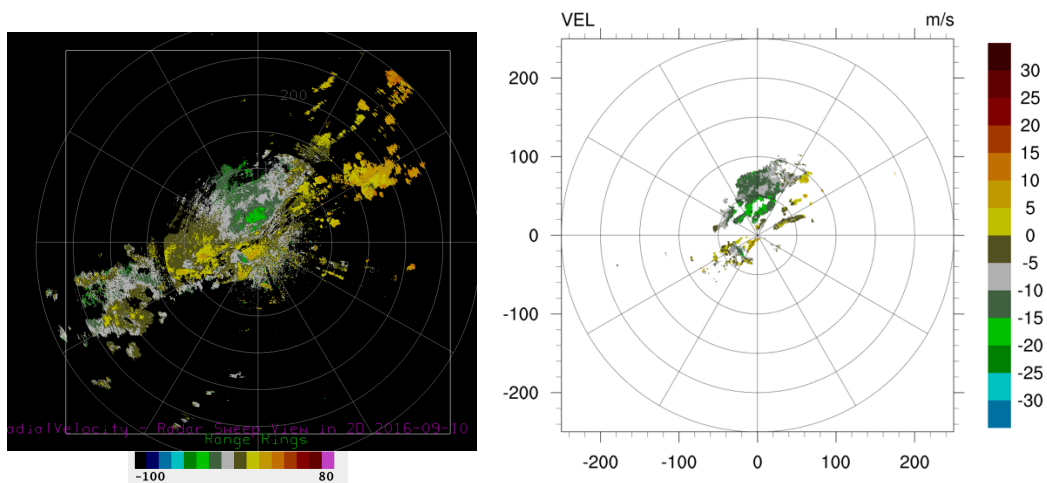

Figure 3.2 The radial winds at the 0.5° elevation angle for KICT at 0021 UTC, 20160910. Left panel from raw data file and right panel from NCEP BUFR file.

Figure 3.3 The radial winds at the 1.3° elevation angle for KICT at 0021 UTC, 20160910. Left panel from raw data file and right panel from NCEP BUFR file.

Because the cases we prepared for experiments were in September, 2016, we further examined recent NCEP radial velocity BUFR files in 2018 to check whether this data missing issue still exists. It was found the problem remains in many 2018 cases. Fig. 3.4 showed that at 20180228_1904 UTC, the NCEP BUFR radial velocity file missed a lot of observations as compared to raw format files. We discussed these results with NCEP EMC colleagues. They acknowledged that there were missing radial velocity observations especially in low elevation angles due to strict quality control procedures for operational applications.
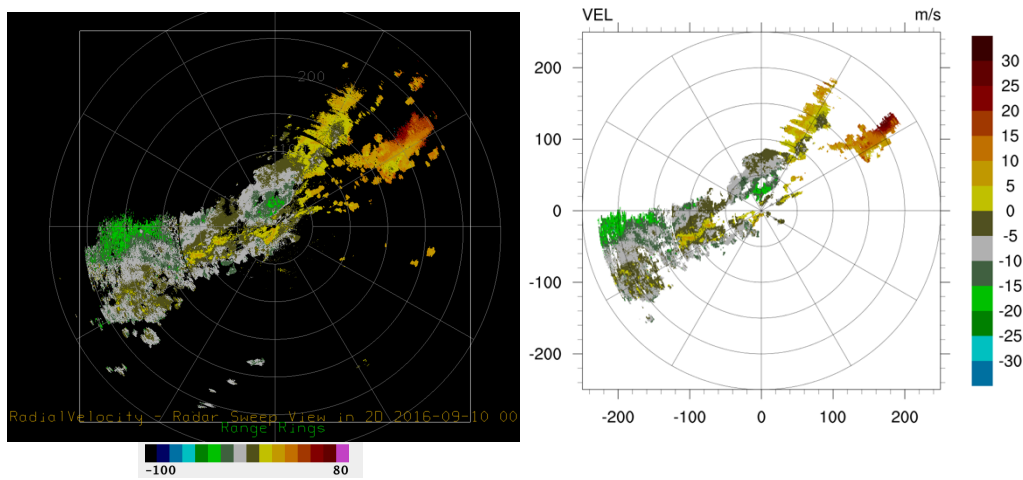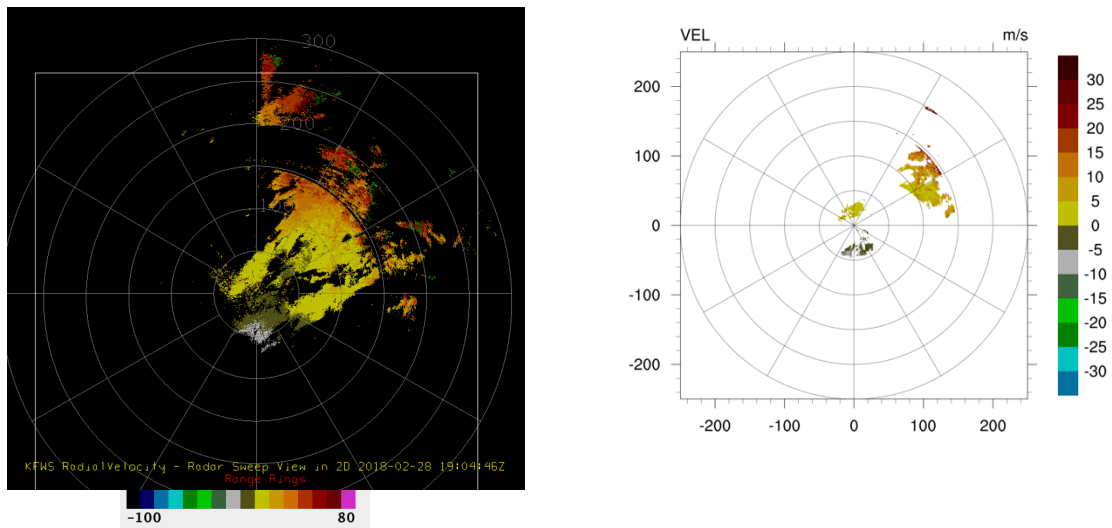

Figure 3.4 The radial winds at the 0.5° elevation angle for KFWS at 1904 UTC, 20180228. Left panel from raw data file and right panel from NCEP BUFR file.

Because the data availability is fundamental issue in operational data assimilation system, other comparison between raw radial velocity observations and the NCEP radial velocity

BUFR files with different display utilities were conducted. The new case checked is September 10, 2016, when there is strong convection along the front, crossing Kansas into northwestern Texas (as shown in Fig. 3.5, Satellite surface composite for 00:30Z of September 10, 2016).
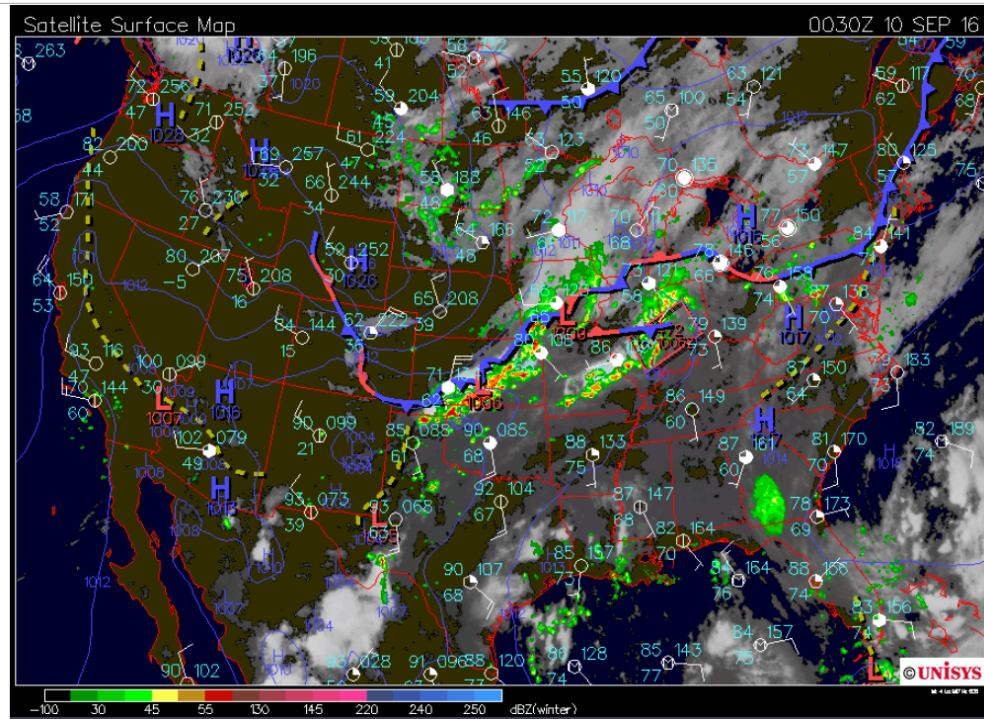
Figure 3.5 Satellite surface composite for 00:30Z of September 10, 2016 (downloaded from the NCAR/MMM image archives at http://www2.mmm.ucar.edu/imagearchive/).

Fig. 3.6 shows the radial wind data coverage within different layers, derived from the GSI diagnostic files at 00Z of September 10, 2016. The radial wind data seems to have a good coverage over the model domain, with most data between the pressure levels of 900hPa and 750hPa. The radial wind shows a northerly wind behind the front and southerly wind ahead of the front, and there are radar data observed near the convective system.
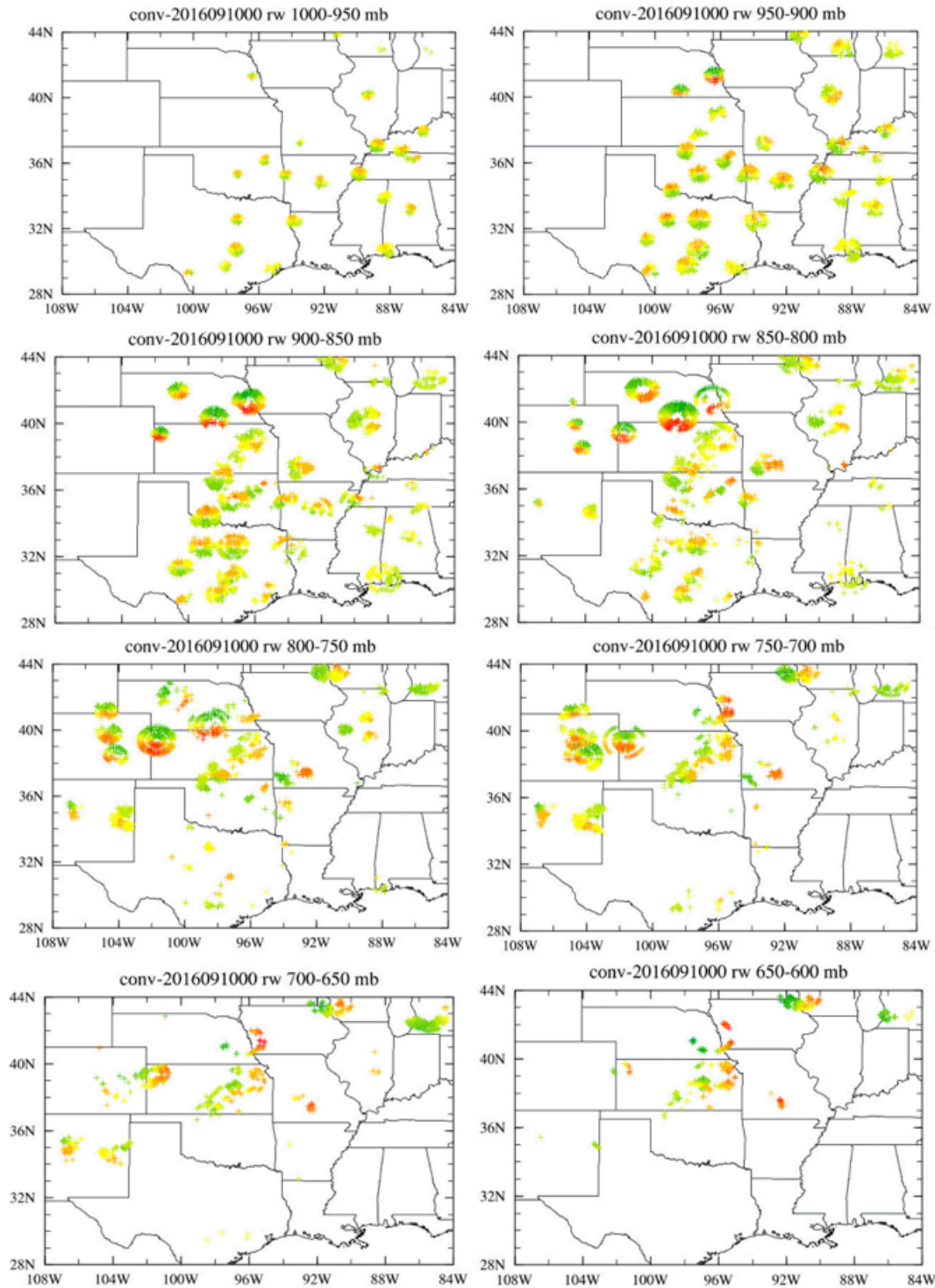
Figure 3.6 Radial wind data coverage within different layers, derived from the diagnostics of the GSI run at 00Z of September 10, 2016.

Fig. 3.7 shows the NEXRAD coverage map, focusing on the stations near the convective system as in Fig. 10. In the following section two stations in Kansas, KICT (Wichita) and KTWX (Topeka) were chosen to do more exam on the radar data quality and distribution.
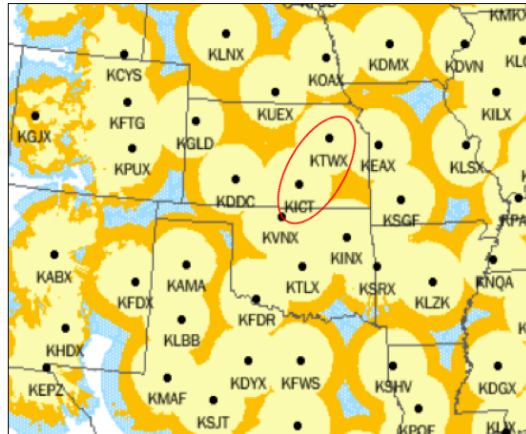
Figure 3.7. NEXRAD coverage map (from https://www.roc.noaa.gov/WSR88D/Maps.aspx)

Fig. 3.8 shows the raw data from NEXRAD Level II for the station KICT at around 00Z of September 10, 2016 for radial wind (upper panels) and reflectivity (bottom panels), for two elevation angles, 0.48 (left panels) and 1.32 (right panels), respectively. The northeast-southwest tilted convective system can be clearly seen from the KICT radar data, with a very good coverage in both radial wind and reflectivity.
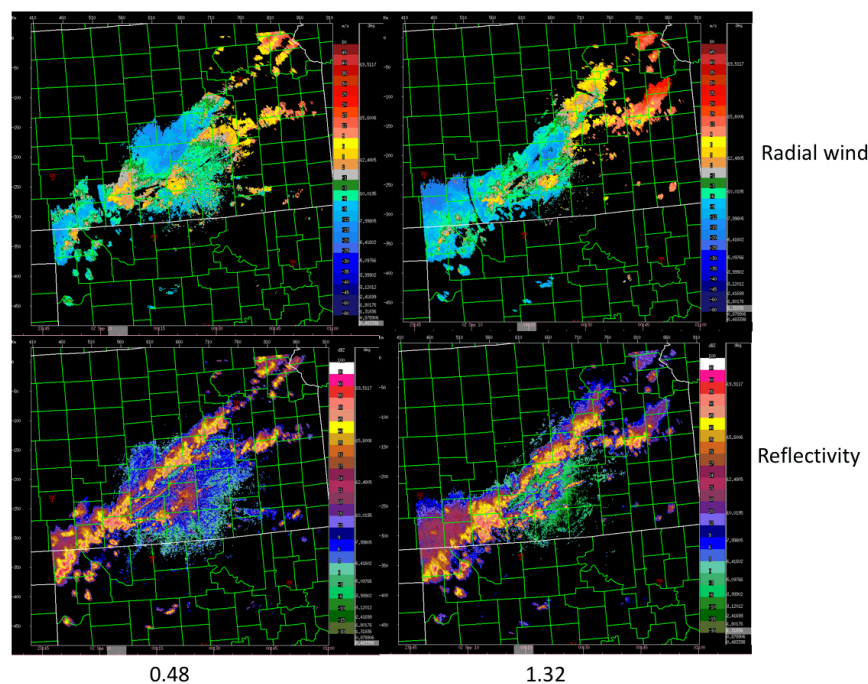
Figure 3.8, NEXRAD Level II raw data for the station KICT at around 00Z of September 10, 2016 for radial wind (upper panels) and reflectivity (bottom panels), for two elevation angles, 0.48 (left panels) and 1.32 (right panels), respectively.

Fig. 3.9 lists the data from RAP radar bufr file for KICT at 00Z of September 10, 2016, in which the 2nd column gives the time of scan in the format of DDHHMM (for example,

100004 mean September 10 00:04Z), the 3rd, 4th and 5th column gives the scan id, number of rays within the scan and elevation angle respectively. From the list, 4 scans are picked to show the radial wind distribution, as denoted in the red boxes in Fig. 3.9, with an intention to match the time and elevation angles in Fig. 3.8 and also maximize the data coverage by choosing the relatively large number of rays. Fig. 3.10 shows the radial wind plots of the above selected 4 scans (denoted in the red boxes). Comparing Fig. 3.8 and 3.10, it can be seen that quite some radial wind data have been missed from the raw observations during the NCEP processing, partly might be due to the cross check between the radial wind and reflectivity. Overall the NCEP BUFR data contains significantly less radial wind data compared to the raw observations and this limited data volume might limit the performance of the radar data assimilation.

1	100004	1	237	0.37
2	100004	2	216	0.76
3	100004	3	223	1.21
4	100004	4	211	1.58
5	100004	5	218	2.28
6	100004	6	199	2.97
7	100004	7	184	3.84
8	100004	8	178	4.92
9	100004	9	214	6.20
10	100004	10	273	7.87
11	100004	11	263	9.90
12	100004	12	261	12.33
13	100004	13	252	15.49
14	100004	14	212	19.38
15	100010	1	245	0.42
16	100010	2	253	0.73
17	100010	3	283	1.16
18	100010	4	273	1.58
19	100010	5	255	2.19
20	100010	6	244	2.98
21	100010	7	278	3.82
22	100010	8	264	4.93
23	100010	9	289	6.20
24	100010	10	307	7.82
25	100010	11	305	9.83
26	100010	12	281	12.31
27	100010	13	266	15.48
28	100010	14	200	19.35

28	100010	14	200	19.35
29	100015	1	285	0.42
30	100015	2	319	0.75
31	100015	3	320	1.16
32	100015	4	297	1.63
33	100015	5	269	2.22
34	100015	6	268	2.93
35	100015	7	281	3.87
36	100015	8	303	4.91
37	100015	9	331	6.21
38	100015	10	323	7.87
39	100015	11	340	9.90
40	100015	12	316	12.33
41	100015	13	289	15.46
42	100021	1	319	0.37
43	100021	2	328	0.73
44	100021	3	339	1.17
45	100021	4	334	1.59
46	100021	5	333	2.24
47	100021	6	322	2.94
48	100021	7	328	3.92
49	100021	8	333	4.93
50	100021	9	346	6.23
51	100021	10	358	7.85
52	100021	11	360	9.85
53	100021	12	336	12.37

Fig. 3.9 Lists of the data from RAP radar bufr file for KICT at 00Z of September 10, 2016, in which the 2nd, 3rd, 4th and 5th column gives the time of scan in the format of DDHHMM, scan id, number of rays within the scan and elevation angle respectively.
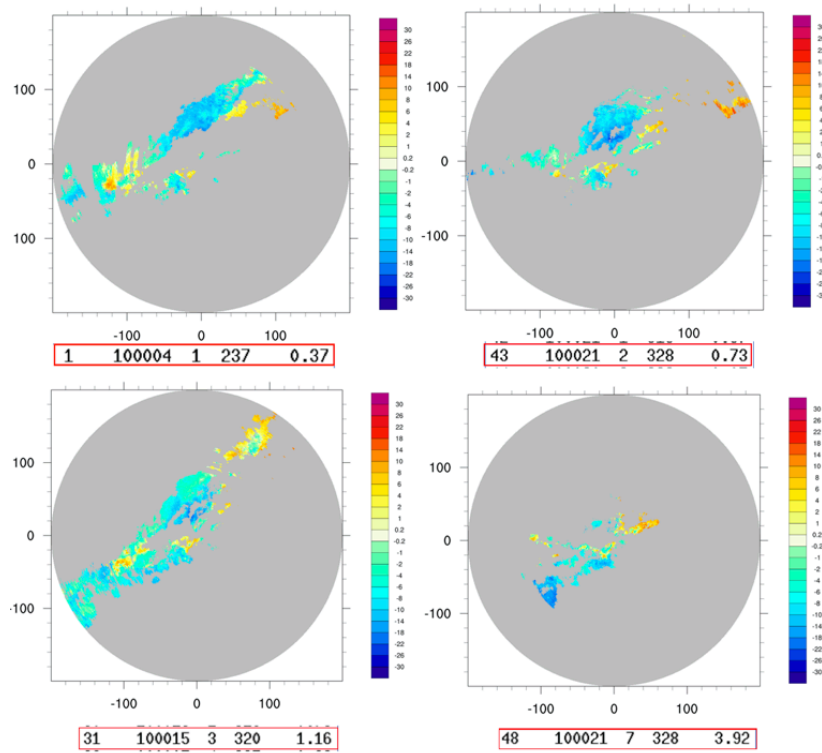
Figure 3.10 KICT radial wind plots of the above selected 4 scans (denoted in the red boxes) from the RAP radar bufr data at 00Z September 10, 2016.

We also take a look at another station, KTWX, which is also within the convective system. Fig. 3.11 is similar to Fig. 3.8 but showing the raw data from NEXRAD for elevation angle 0.48 only. It can be seen the radial wind covers the whole area around the station, suggesting a north-westerly wind at low level, while the strongest reflectivity shows the tilted convective system.

Figure 3.11 NEXRAD Level II raw data for the station KTWX at around 00Z of September 10, 2016 for radial wind (left panel) and reflectivity (right panel), for elevation angle 0.48.

Fig. 3.12 shows the RAP bufr data for KTWX at the same elevation angle (0.48). Surprisingly the data volume in the RAP bufr data is very different from the raw data. Compared to the full coverage in the raw radial wind data, the RAP bufr data only shows very sparse coverage and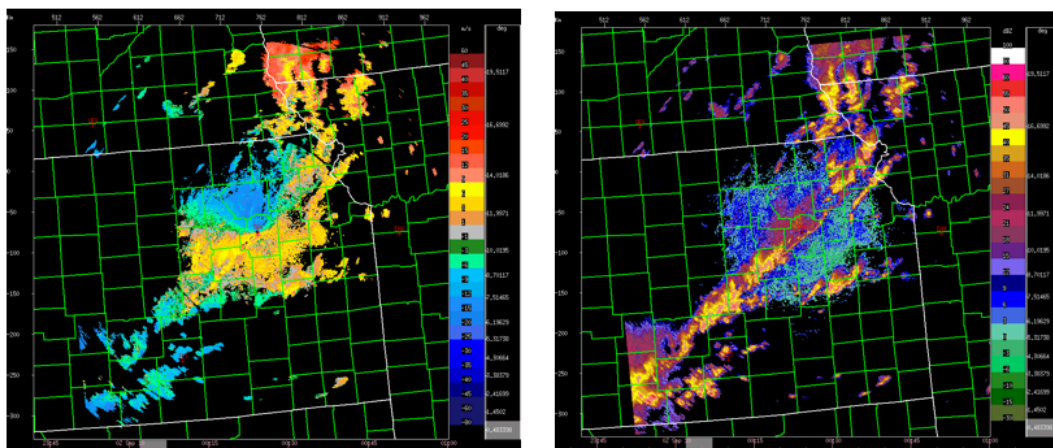 is even a huge reduction from the strongest reflectivity coverage. The bufr data is so sparse, making it almost useless in the data assimilation. It raises concerns about the procedure behind the NCEP radar data processing and what caused the large amount of data being removed, which otherwise might be useful.
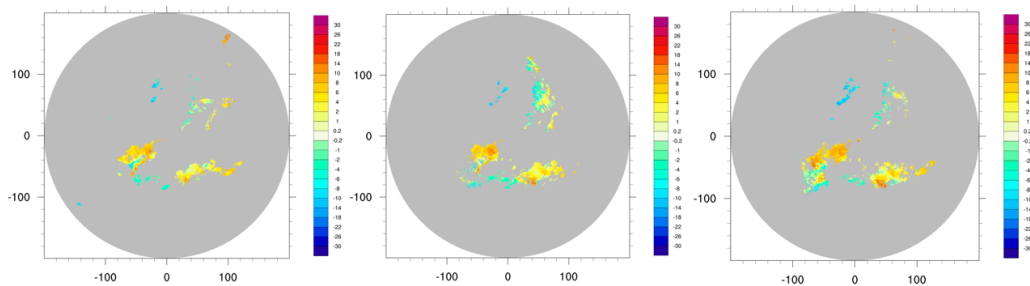

Figure 3.12 NCEP BUFR file for KTWX at 00:04Z (left), 00:09Z (middle) and 00:14Z (right) of September 10, 2016, for elevation angle 0.48.

This data missing issue indicates that we may need to revisit the radial velocity quality control procedures for storm scale implementations. However, this goes beyond the scope of AOP 2017 T&E tasks. On the other hand, although there are a lot of data missing, we still have enough radial velocity observations in a one hour or 30-minute time window to generate representative super-obbed radial velocity observations for meaningful data assimilation experiments.

3.2 Radar superob setting examination

GSI assimilates radar radial velocity in a superob fashion. It averages radial velocity observations in a three-dimensional cell, called the superob cell. The size of the superob cell is defined by range, azimuth, elevation and time intervals. In GSI, the superob setting is controlled by the "superob_radar" namelist section as shown below:

```
& superob_radar
del_azimuth=5.,
del_elev=.25,
del_range=5000.,
del_time=.5,
elev_angle_max=5.,
minnum=50,
range_max=100000.,
/
```

Where "del_azimuth" defines the azimuth interval, the default value is 5 degree; "del_elev" defines the elevation interval, the default value is 0.25 degree; "del_range" defines the range interval, the default value is 5 km; "del_time" defines the time interval, the default value is

0.5 hours, i.e. 30 minutes; "minmum" defines the minimum required number of observation points to generate a valid superob; "elev_angle_max" defines the maximum elevation angle, the default value is 5 degree; "range_max" defines the maximum range in the superob process, the default value is 100 km. The latitude and longitude of a superob are also computed as averages of available observation latitude and longitude in a superob cell. The observation error for a superob is the stand deviation of radial velocties in a superob cell.

To check how superobbed radial velocity looks like and how the GSI superob part works, we used tools developed in previous section to write out only one radar sweep from one radar into a small BUFR file and made superob's from this small BUFR file. The following discussions are based the 1.3° sweep from KINX radar at 20180227 2102z. Fig. 3.13 showed the comparison of original radial velocity observations in the NCEP BUFR file and the super-obbed radial velocity data using the default superob setting as in RAPv4/HRRRv3. It can be seen that superob process reduced data amount and kept important convective details. Fig. 3.14 showed the standard deviations associated with superob's in Fig. 3.13. Most standard deviations are smaller than 2 ms^{-1}, while a few cells with standard deviation of 3 ms^{-1}. All these indicate that the GSI superob process works well. Fig. 3.15 showed the superob results using smaller superob cells with the setting "del_azimuth=3, del_range=3, del_time=0.25, minmum=25". It can be seen that more observation details are kept when using smaller superob cells.
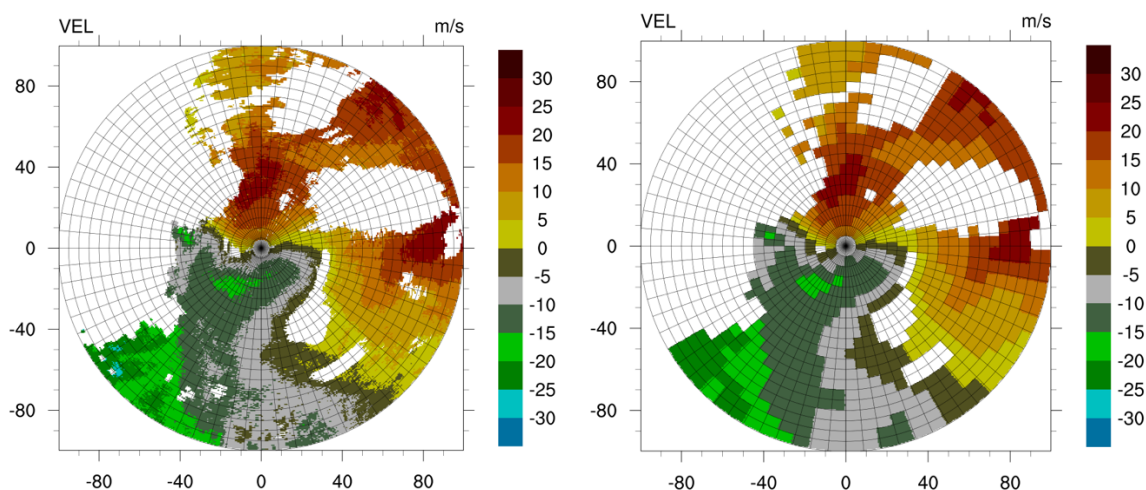

Figure 3.13. Original radial velocity data and super-obbed radial velocity data
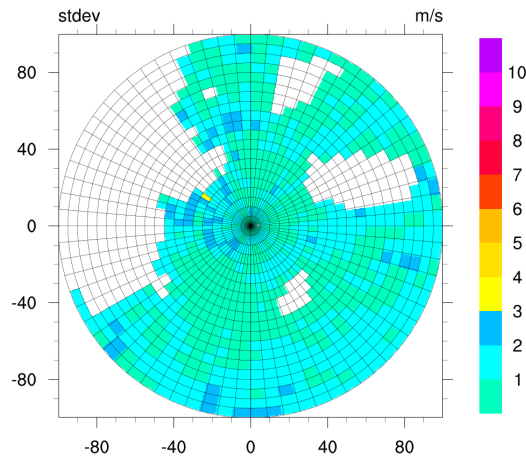
Fig. 3.14. Standard deviation of superob's in Fig. 3.13
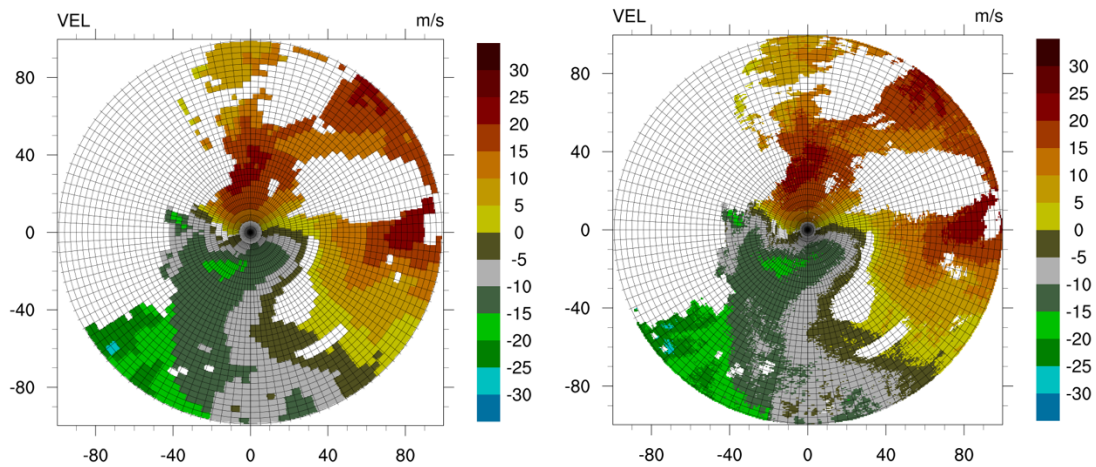
Fig. 3.15 . Original radial velocity data and super-obbed radial velocity

3.3 VAD checking

Besides the radial wind data being assimilated in the GSI, the VAD (Velocity Azimuth Display) wind profiles embedded in the NCEP PrepBUFR data is also being assimilated in the HRRR system. The VAD, a derived product from radial wind observations, is a plot of horizontal winds, as a function of height above a Doppler Radar. It depicts the change in wind with time at various heights and is useful for observing local changes in vertical wind shear. According to verbal communications with NOAA colleagues, past RAP/HRRR experiments have found negative impact from assimilating VAD wind, which is supposed to provide useful information on the storm environment. In the experiments to be conducted, the VAD products are also evaluated. Therefore diagnostics of the VAD wind data is performed for the same date. Fig. 19 shows the locations of the VAD observations and the wind barbs for each station. Similar to the radial wind distribution in Fig. 11, northerly (southerly) winds are observed behind (ahead of) the front. And there are some discrepancy between the radial wind and VAD stations.
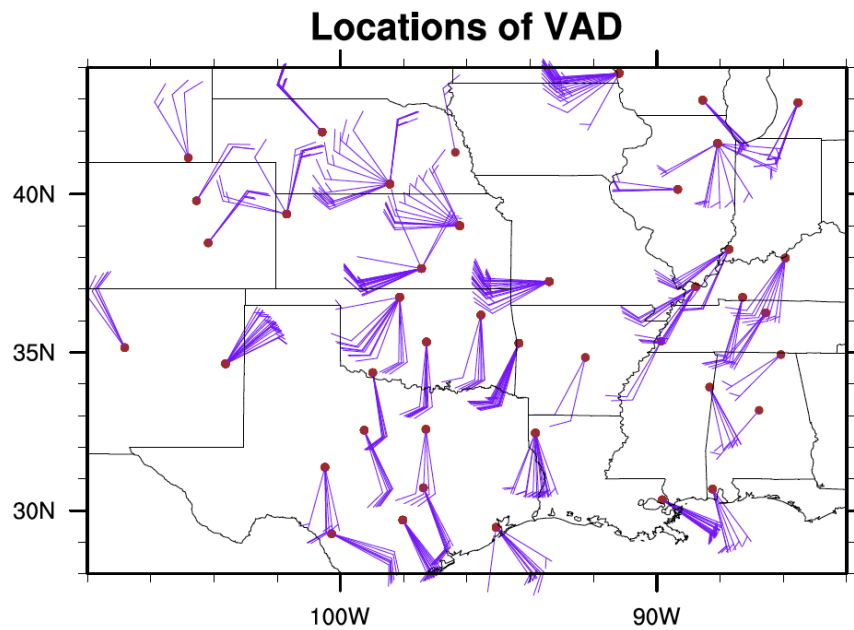
Locations of VAD

Figure 19 Locations of VAD wind data for 00Z September 10, 2016.

4. Understand the GSI radar radial velocity analysis with single case study

Before running one week long retrospective experiments, we conducted a series of single case studies to examine the impact of radial velocity assimilation on HRRR analysis and determined the optimal parameter settings for different experiments. Based on the SPC storm reports (Fig. 4.1), we selected 20160910 000UTC as the time to conduct our single case studies. At this time, a major square line moved across the Great Plain area (Fig. 4.2), yielded large hails, damage winds and a few tornadoes(Fig. 4.1).
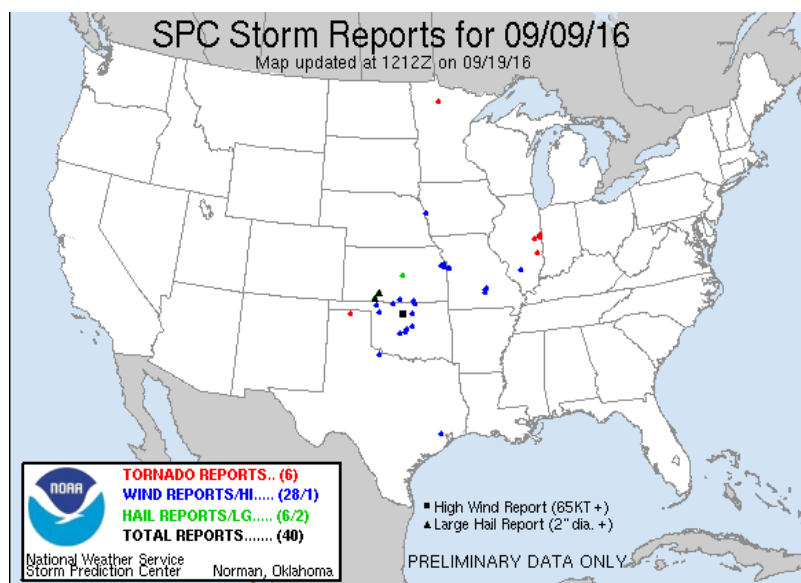
Figure 4.1 The SPC storm report around 20160910 0000 UTC
(http://www.spc.noaa.gov/exper/archive/event.php?date=20160909)

Figure 4.2 The composite reflectivity at 20160910 000UTC

The radar radial winds are in high spatial and temporal resolutions and contain convective-scale information. The range resolution of radial velocity data in NCEP BUFR files is 250 m, which is much higher than HRRR model horizontal resolution of 3 km. In order to correctly assimilate this convective-scale information into the model, a few modifications to current HRRR configuration should be made. First, the decorrelation scale of the static BE (background error covariance) and the localization scale of the ensemble BE should b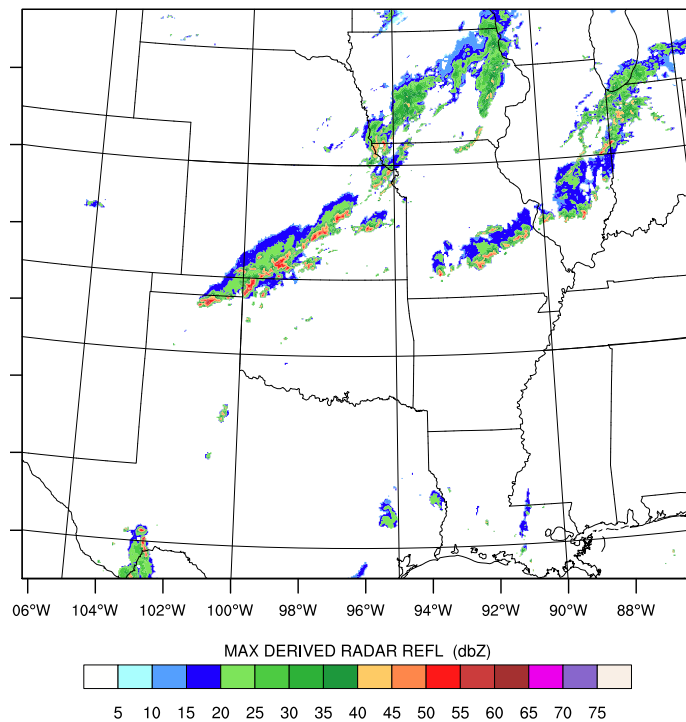e adjusted to match the high resolution of radial winds. Second, the default superob parameter setting is based on the NAM application which conducts data assimilation at 12 km, this setting may not be suitable for the 3 km data assimilation of this study.

4.1 BE horizontal decorrelation Length scale tuning

The static BE horizontal decorrelation scale is controlled by the parameter hzscl(3), which needs three values. For example, the default ones are "hzscl=0.373,0.746,1.5". It needs three values because the fat-tailed power spectrum of horizontal impact is achieved by combining three recursive filters with different impact scales respectively (Fig. 4.3). The above three values are the scale factors of each recursive filter. This poses a problem: the actual BE horizontal impact scale is not known straightforward. We hence conducted single observations tests by putting an observation at the model center and 500hPa height and then plotting the cross section through the location of the single observation. This way, we can know the rough impact scale of the combination of recursive filters. Fig. 4.4 showed that the

default "hzscl" setting corresponds to an impact scale of about 300 km. Through a few tries, we find a setting "hzscl= 0.02487, 0.04973, 0.1" which corresponds to an impact scale of about 20 km. This impact scale is considered to be more appropriate for 3 km convective scale radar data assimilation.
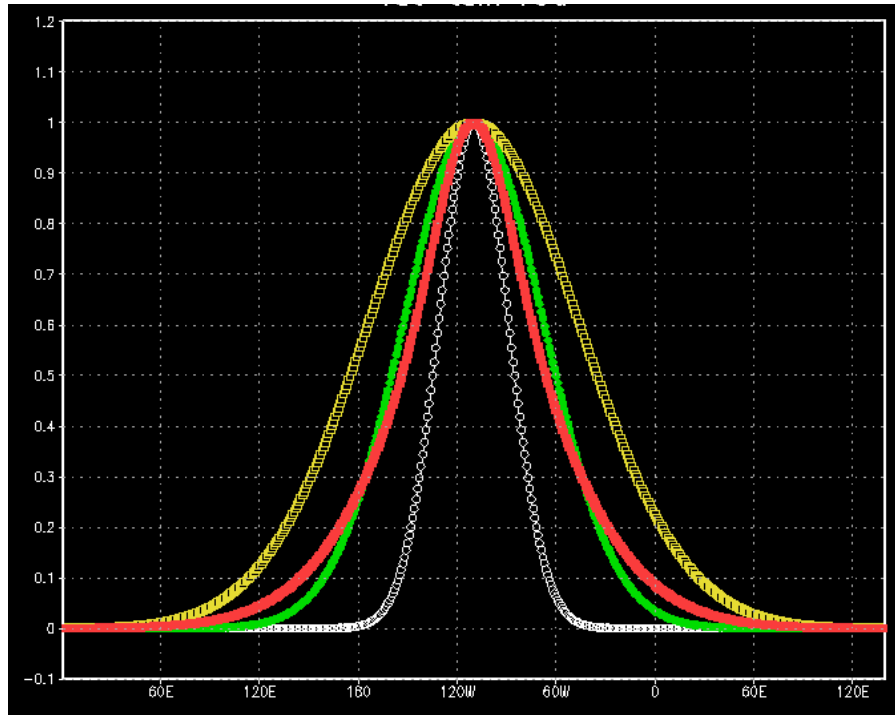

Figure 4.3 Fat-tailed Power Spectrum for horizontal impact in GSI static BE part (Courtesy of Wan-Shu Wu)
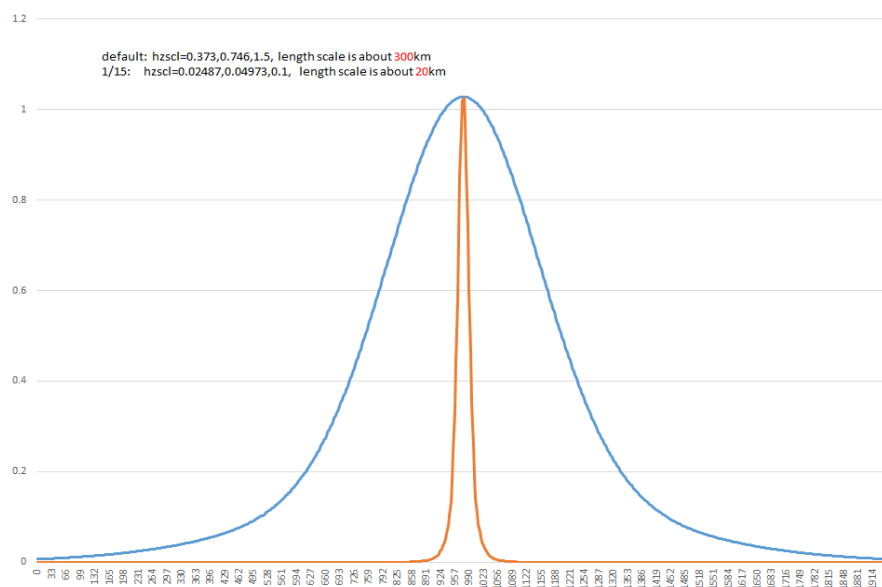

Figure 4.4. The horizontal impact of the GSI static BE part for two different set of hzscl parameters

The ensemble localization scale is controlled by the parameter s_ens_h and s_ens_v. Their settings are pretty straight-forward. The default values are 110 and 3, corresponds to 110 km in horizontal and 3 grid points in vertical respectively. To match the above static BE scale changes, we set the corresponding ensemble localization scale to 20 km in horizontal and 2 grid points in vertical.

4.2 Impact of super-ob tuning and length scale tuning on analysis

Three single analysis experiments were conducted. All three experiments were done at 20160910 00:00 UTC and they assimilated only radial wind observations in order to easily examine the impact of super-ob tuning and length scale tuning on analysis.

The first experiment is the control (CTRL) experiment, which assimilated radial winds at 12 km. The second experiment assimilated radial winds at 3 km in a separate second pass with smaller horizontal and localization scales, i.e. those corresponds to 20 km horizontally discussed in previous section. This experiment will be referred as HEX1 (mnemonics of hybrid experiment 1). The third experiment is similar to HEX1 except that it uses a new set of superob parameters, the changed ones illustrated in Fig 3.15 in section 3.2, which uses smaller superob cells. This experiment will be referred as HEX2.

Fig. 4.5 shows the analysis increments from the CTRL, HEX1 and HEX2 experiments. We can see that using default large impact scale, the analysis increments are very smooth in the CTRL experiment. When using smaller impact scale, the analysis increments become very local, showing more convective features in the HEX1 and HEX2 experiments. Fig. 4.6 showed the analysis increments by zooming into the Nebraska Area. It is very clear that using smaller impact scale yields evident convective-scale local analysis increments compared to smooth increments in the CTRL experiment. Comparing HEX1 with HEX2, we can see there are evident differences in the incremental wind field in the middle bottom part of Nebraska area. This indicates that more convective information kept from using smaller superob cells and it does affect the wind analysis. Fig. 4.7 and Fig. 4.8 showed the increment fields for temperature and water vapor specific humidity. We can see that GSI is able to balance the wind innovations with the temperature and water vapor fields. When using smaller impact scale, the increments get smaller and more local, similar behavior as in the wind field.
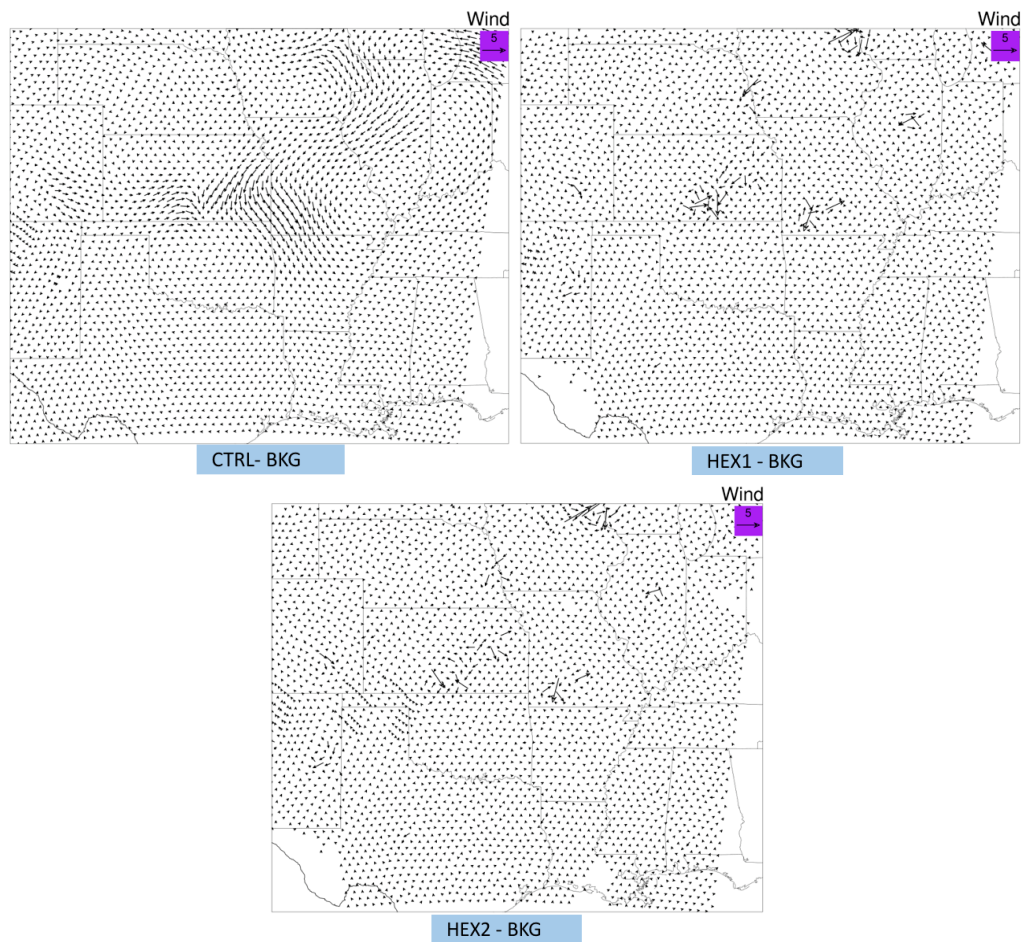
Figure. 4.5 The wind increments of the CTRL, HEX1 and HEX2 experiments at 00:00 UTC, 20160910
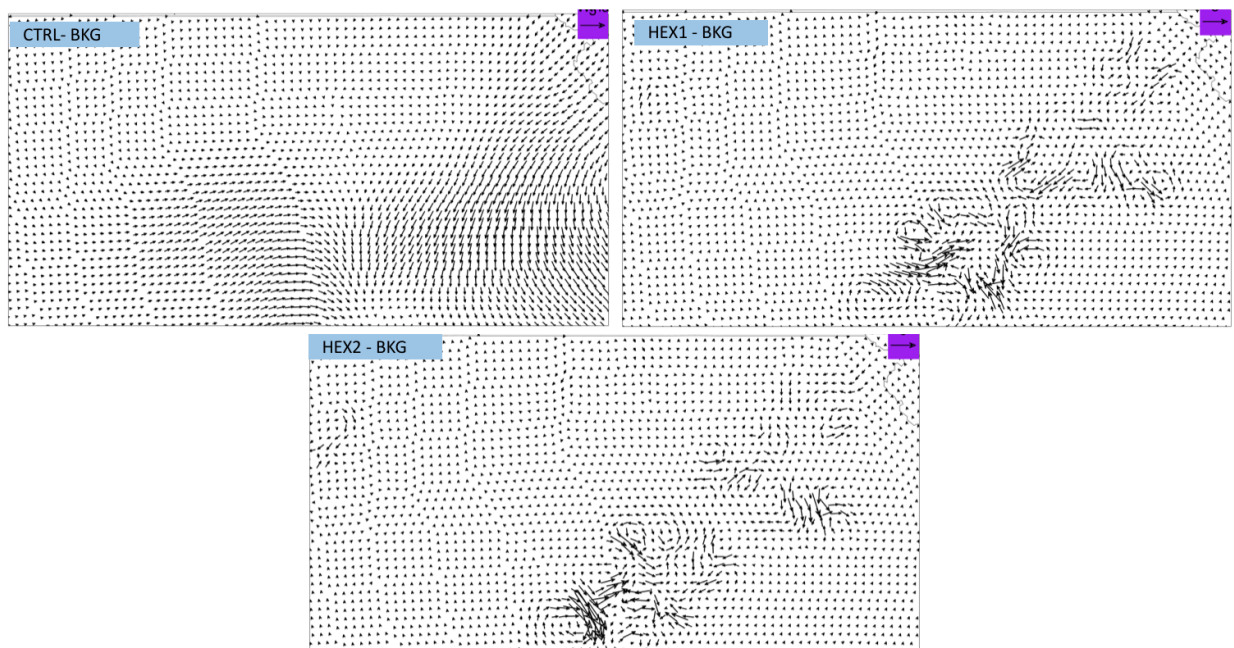
Figure 4.6. Similar as Fig. 4.5 but zoomed into the Nebraska area.
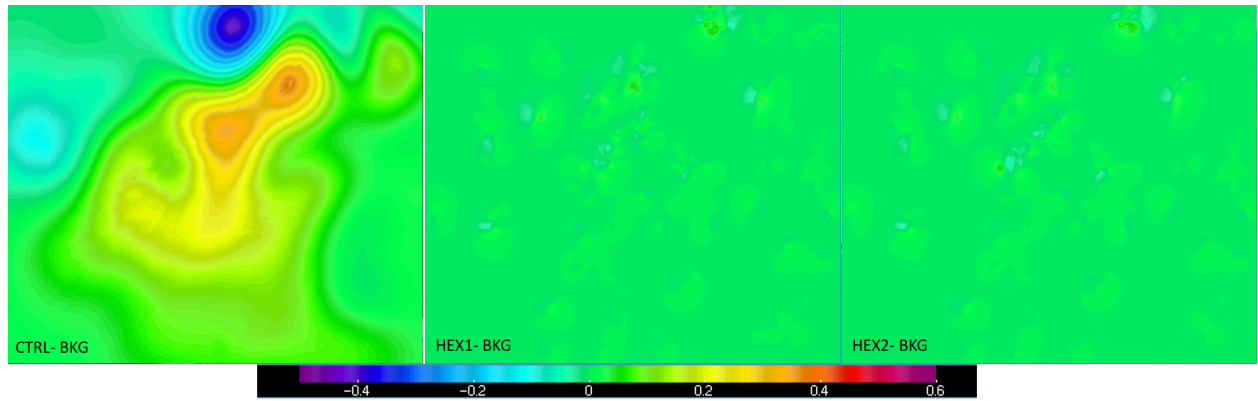
Figure 4.7 Temperature increments (K) of the CTRL, HEX1 and HEX2 experiments at 00:00 UTC, 20160910
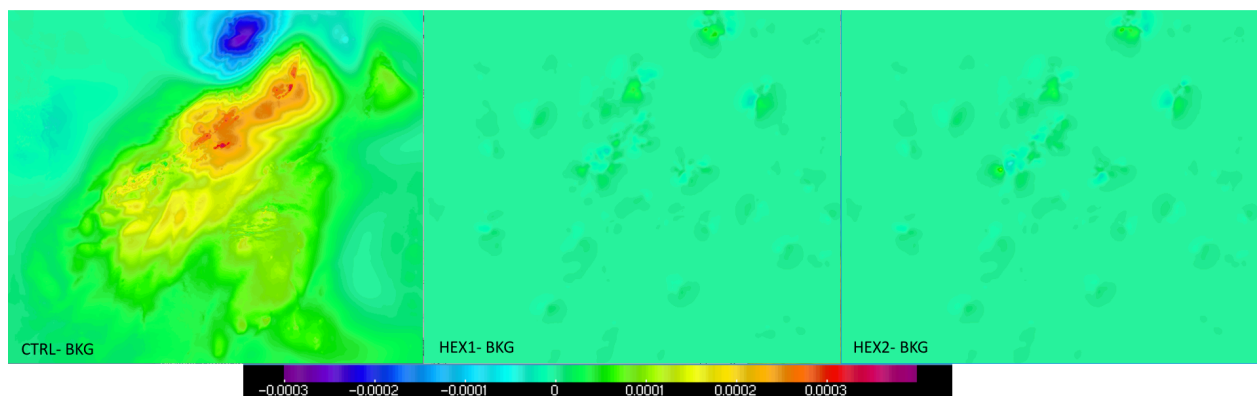
Figure 4.8 Specific Humidity increment (g/kg) of the CTRL, HEX1 and HEX2 experiments at 00:00 UTC, 20160910

The same single case is also used to study the possible tuning of the superob parameters for improving storm environment by increasing the size of the superob cells. As listed below, besides the control (using the default configurations, except for a shortened time window to +/-0.2 hours), there are 3 other single case studies which increase the superob box to be nearly 2, 4 and 6 times of the default one.

- Control: del_azimuth=5, del_range=5000
- 12km grid: del_azimuth=13, del_range=12000
- 24km grid: del_azimuth=27, del_range=24000
- 36km grid: del_azimuth=42, del_range=36000

Fig. 4.9 gives the wind increments (upper panels) and total wind fields (red: background; green: analysis) at model level 11, for the single case studies with superob range at 5km, 12km, 24-km and 36-km, from left to right. As can be seen, the wind increments tend to be smoother with increased superob ranges. Overall, the wind increments from the radial wind assimilation are quite small and differences between different superob boxes are small too. Similar features are observed in the temperature and humidity increment fields, as seen in Fig. 4.10, and main increments are observed near the convective system.

Figure 4.9 Wind increments (upper panels) and total wind fields (red: background; green: analysis) at model level 11, for the single case studies with superob range at 5km, 12km, 24-km and 36-km, from left to right.
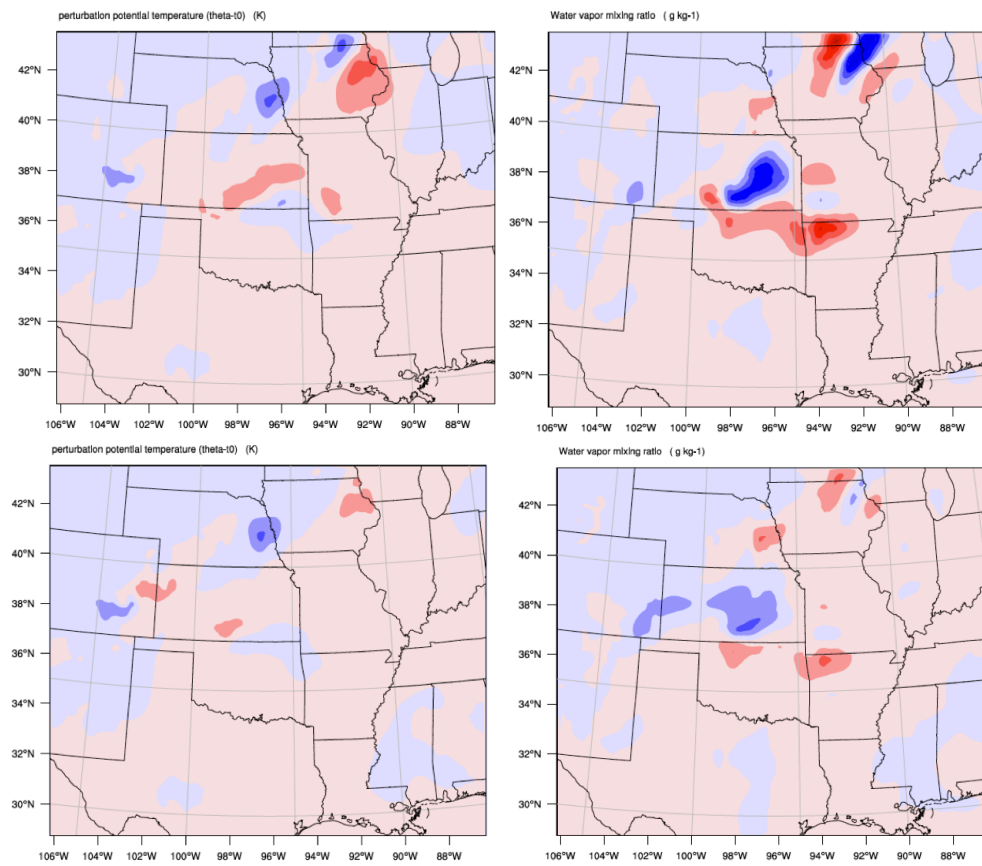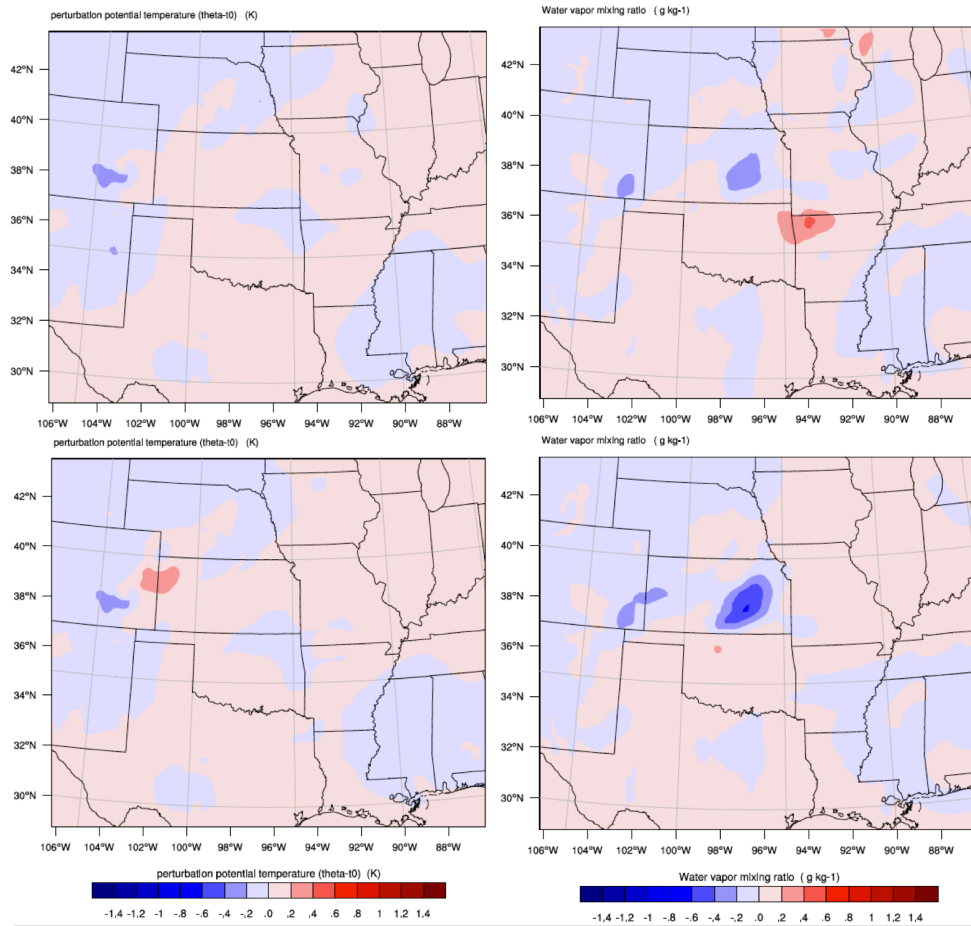
Figure 4.10 Temperature (left panels) and mixing ratio (right panels) increments at model level 11, for the single case studies with superob range at 5km (top row), 12km (2nd row), 24-km (3rd row) and 36-km (bottom row), from top to bottom.

4.3 The impact of VAD wind in GSI analysis

Besides the single case studies with tuned superob parameters shown above, another single case study is conducted for the same date, with only VAD wind assimilated. Fig. 4.11 shows the increments at model level 11 for temperature, humidity and wind fields. As can be seen the increments are quite different from the increments from assimilating radial wind only, in terms of both magnitude and distribution. And the increment plots are jagged at some areas, which remains of unknow causes yet. Fig. 4.12 shows the wind increment and total wind fields at model level 11 for the VAD only, showing very different response in the wind field compared to the runs with radial wind only. This is puzzling since the two are both radar data products and supposedly they should provide similar information. More diagnostics are needed.

Figure 4.11 Temperature (T, in K), humidity (q, in g/kg), U-wind (in m/s) and V-wind (in m/s) increments at model level 11 for the experiment with VAD only.
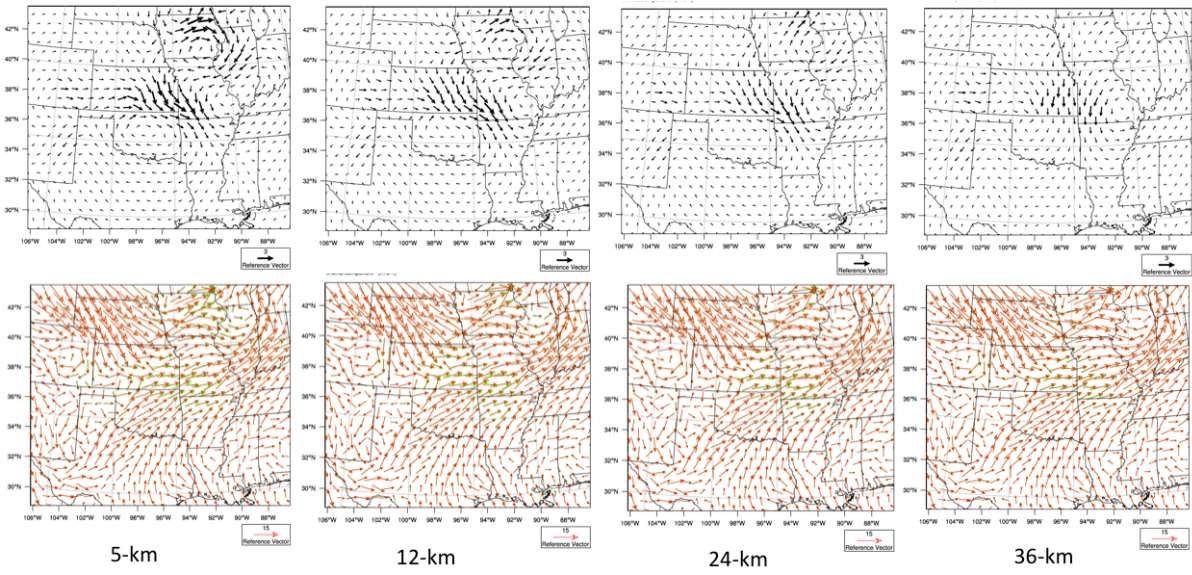

Figure 4.12 Wind increments (left panel) and total wind fields (right panel, red: background; green: analysis) at model level 11, for the single case study with VAD only.

5. September 2016 HRRR Retrospective Test for Radial Wind Impact
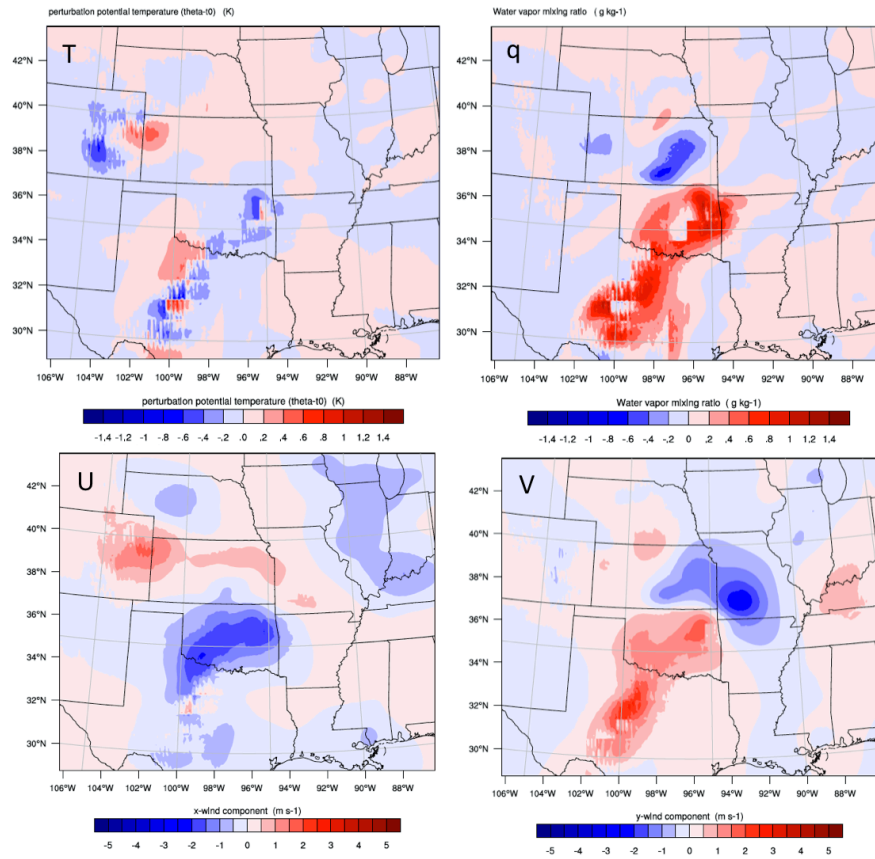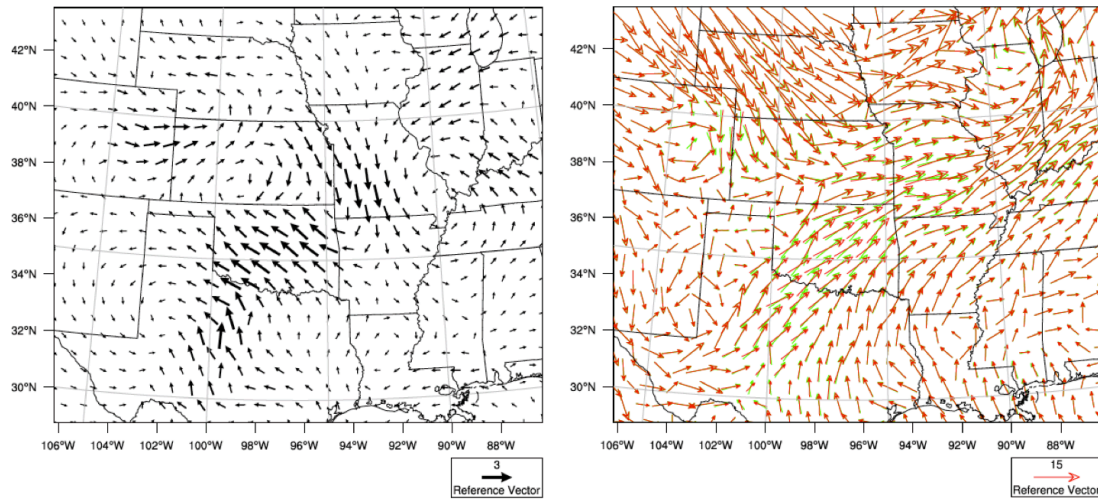
5.1 Experiment setups for high resolution analysis

Table 5.1 lists all one-week long retrospective experiments conducted to study the use of radial winds (RW) in high resolution 3-km analysis grid and small super-obbing cells. All experiments followed similar workflow and configuration as the real-time RAPv4/HRRRv3 system and assimilate all available conventional observations, radar reflectivity and lightning observations, NASA LaRC cloud products and aircraft observations. The JET was used for running the experiments and MATS was used for verification.

The CTRL experiment assimilated radial wind observations at a 12 km analysis grid using the hybrid 3DEnVar method. The horizontal impact scale, the ensemble localization scale and the superob settings followed the values used in the RAPv4/HRRRv3 system. The NoRW experiment did not assimilate radial winds and was used to study how much benefits we can get from additional RW assimilation using default parameter settings in the CTRL experiment. The HEX1 experiment assimilates additional RW observations at 3 km analysis grid using the hybrid 3DEnVar method. It is different from the CTRL experiment in the horizontal impact scale and ensemble localization scale parameter settings. The CTRL experiment uses a large horizontal length scale of about 300 km and an ensemble localization scale of 110 km horizontally and 3 grid points vertically, while the HEX1 experiment uses a smaller horizontal length scale of about 20 km and an ensemble localization scale of 20 km horizontally and 2 grid points vertically. HEX2 experiment is similar as HEX1 except that HEX2 uses a different superob setting which adopts smaller superob cells. NEX1 and NEX2 are similar to HEX1 and HEX2 respectively but NEX1 and NEX2 assimilate RW data using the pure 3DVar method. The purpose of NEX1 and NEX2 is to examine whether we can benefit from the hybrid 3DEnVar method for the assimilation of high resolution RW data by incorporating large scale GDAS ensemble covariance.

Table 5.1 List of experiments and different parameter settings for the assimilation of RW

Experiment	Assimilation strategy for RW (radial winds)	Horizontal scale for RW	Localization scale for RW	Superob setting
CTRL	Hybrid 3DEnVar at 12 km	0.373,0.746,1.5	h=110, v=3	5deg, 5km, 60min, minimum 50 points
NoRW	n/a	n/a	n/a	n/a
HEX1	Hybrid 3DEnVar at 3 km	0.02487,0.04973,0.1	h=20, v=2	5deg, 5km, 60min, minimum 50 points
HEX2	Hybrid 3DEnVar at 3 km	0.02487,0.04973,0.1	h=20, v=2	3deg, 3km, 15min, minimum 25 points
NEX1	3DVar at 3km	0.02487,0.04973,0.1	n/a	5deg, 5km, 60min, minimum 50 points
NEX2	3DVar at 3km	0.02487,0.04973,0.1	n/a	3deg, 3km, 15min, minimum 25 points

Table 2 listed for each experiment the mnemonics and the associated verification names in the NOAA/GSD MATS system (https://www.esrl.noaa.gov/gsd/mats/).

Table 2 List of experiments and the associated mnemonics and verification names in the NOAA MATS system

Experiment	Mnemonics	Verification Name in the NOAA/GSD MATS system
CTRL	Control	HRRRret_ctrl_20160903_gge
NoRW	No Radial Winds	HRRRret_gge_NoRW_Sep2016
HEX1	Hybrid experiment 1	HRRRret_gge_HEX1_Sep2016
HEX2	Hybrid experiment 2	HRRRret_gge_HEX2_Sep2016
NEX1	Non-hybrid Exp. 1	HRRRret_gge_NEX1_Sep2016
NEX2	Non-hybrid Exp. 2	HRRRret_gge_NEX2_Sep2016

5.2 Results for high resolution analysis

a. verifications against radiosonde and surface METAR observations

We computed the RMS and bias against radiosonde and surface METAR observations to examine the impact of different data assimilation experiments on synoptic scale environment forecasts. The statistical results are based on one-week long 00z and 12z forecasts. It showed that there is no significant difference among the six experiments in terms of synoptic storm environment forecasts. The conclusion is valid for all the wind, temperature and moist fields and all forecast hours. Detailed figures can be found in the NOAA/GSD MATS system. As examples, Figs. 5.1, 5.2, 5.3, 5.4 show some RMS and bias profiles for these experiments. We can see that although there are some differences among difference experiments, these differences are very small and not statistically significant.
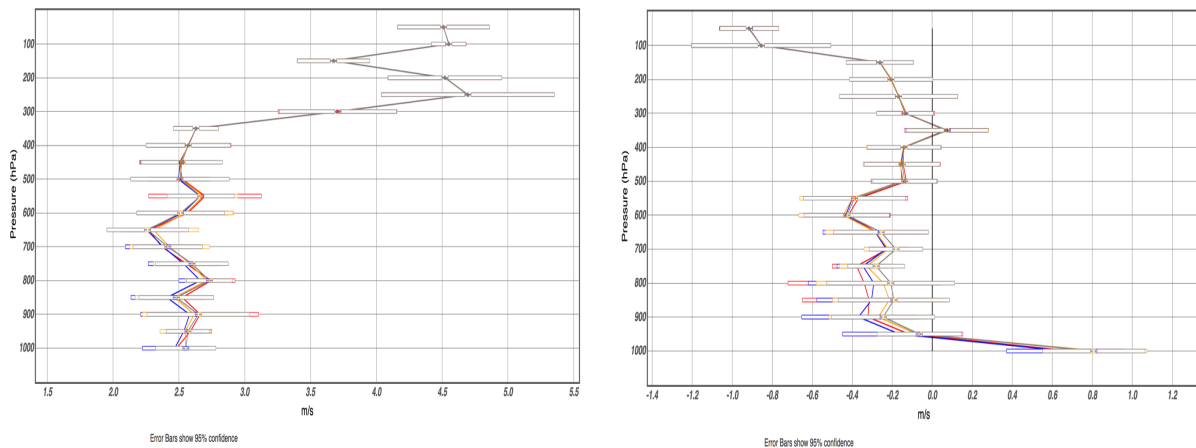
Figure 5.1. Wind RMS (left) and Bias (right) profiles for 0-hour forecasts for the CTRL(red), NoRW(blue), HEX1(orange) and HEX2(grey) experiments
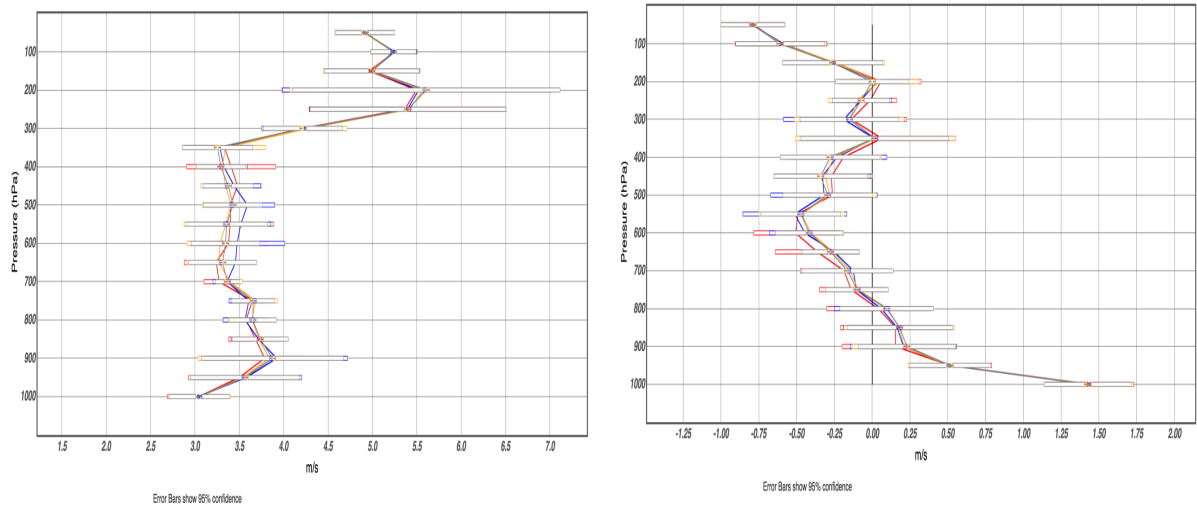
Figure 5.2. Wind RMS (left) and Bias (right) profiles for 3-hour forecasts for the CTRL(red) , NoRW (blue), HEX1(orange) and HEX2 (grey) experiments
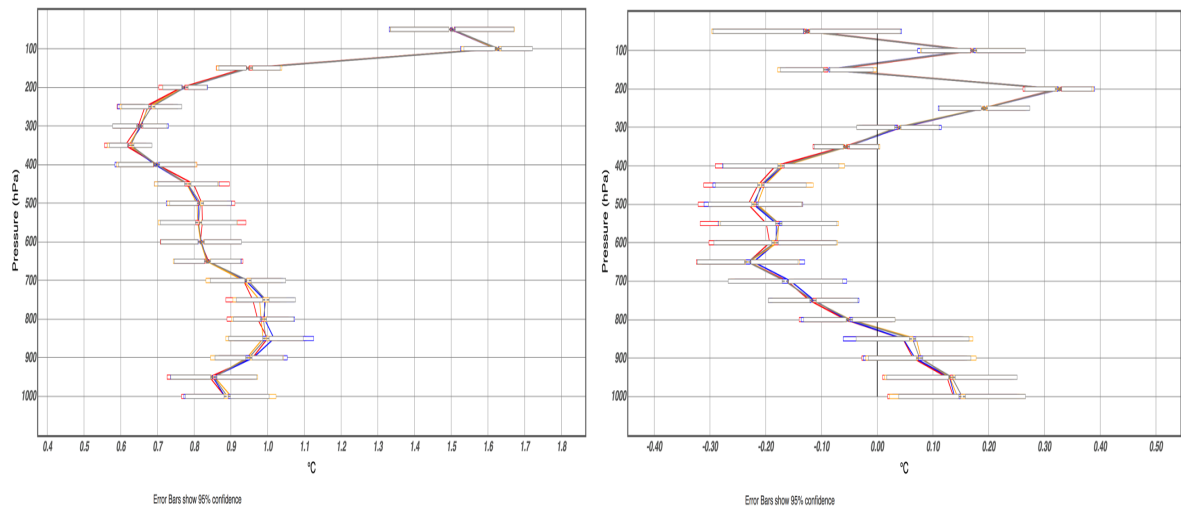
Figure 5.3. Temperature RMS (left) and Bias (right) profiles for 3-hour forecasts for the CTRL(red), NoRW (blue), HEX1 (orange) and HEX2(grey) experiments
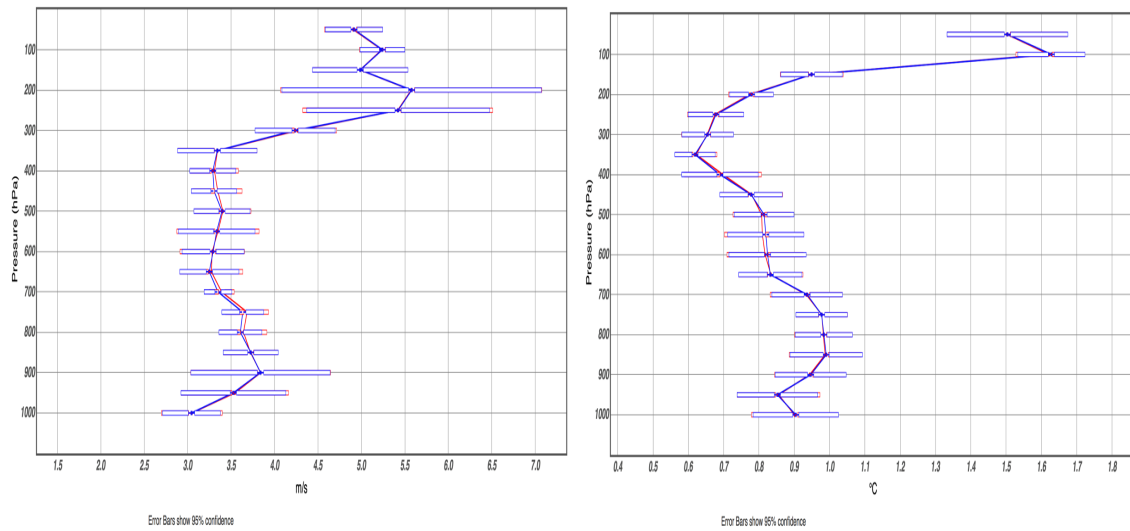
Figure 5.4. Temperature RMS (left) and Bias (right) profiles for 3-hour forecasts for the HEX1 (red) and NEX1 (blue) experiments

b. Verification of individual storms

The statistic verification showed there was no significant difference among the six experiments in this study in terms of synoptic storm environment forecasts. This indicate that although we assimilated lots of convective-scale radial wind observations, the large scale environmental forecasts did not change. This is a positive sign since our goal is to improve convective-scale storm forecasts and need to avoid bringing any negative impact to synoptic scale environment forecasts. In this section, we will investigate the forecasts of individual storms.

Checking the 0h, 1h, 3 h, 6h, 12 h storm forecasts valid at 00:00 UTC 20160910 as an example. Fig. 5.5 shows the observed composite reflectivity. Fig. 5.6 compared the 3 h forecasts between CTRL and HEX1. We can see that in the northeast corner of Nebraska (indicated by the blue circles in the figures), HEX1 predicted more storm cells and stronger storms than CTRL, matching observations better. In the southwest corner of Missouri (the red circles), CTRL missed a strong storm, while HEX1 predicted this storm very well. This means that when we adopt a smaller horizontal impact scale and ensemble localization scale in HEX1, the individual storm forecasts are improved.

MAX DERIVED RADAR REFL (dbZ)

5 10 15 20 25 30 35 40 45 50 55 60 65 70 75
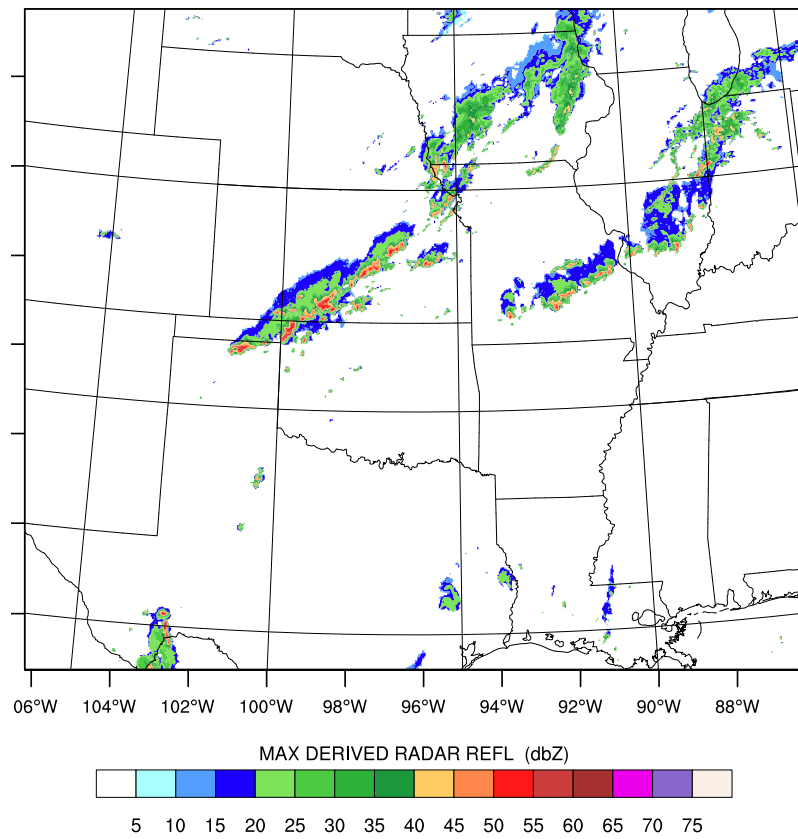
Figure 5.5. Observed composite reflectivity at 00:00 UTC, 20160910

Figure 5.6. simulated composite reflectivity from 3h forecasts for CTRL and HEX1 valid at 0Z, 20160910
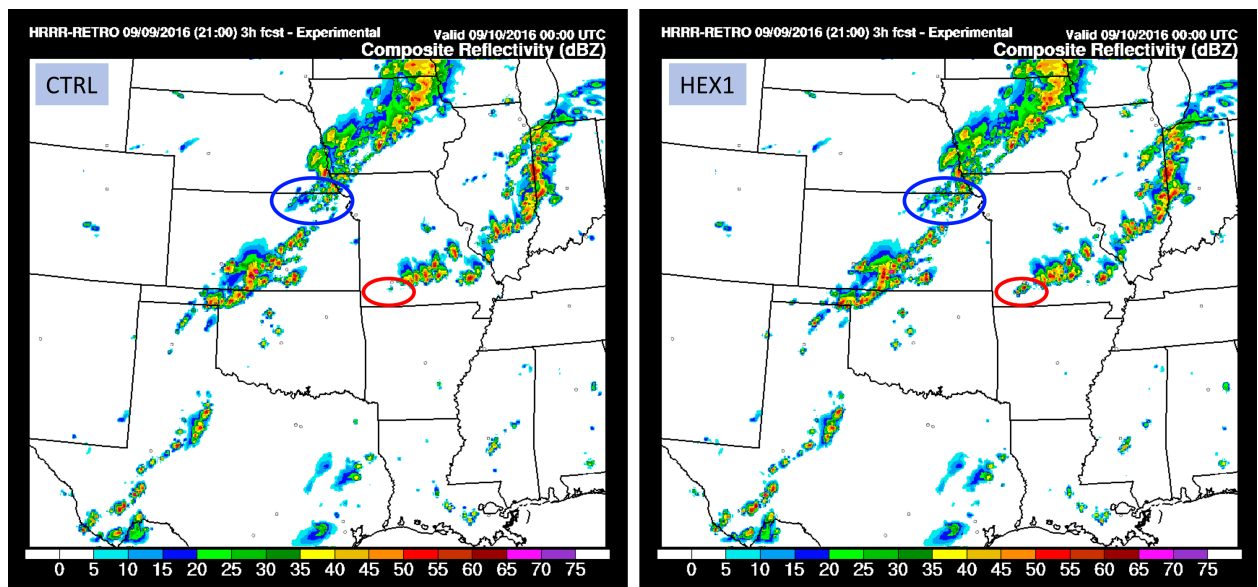
Fig. 5.7 compared the forecasts between NoRW and HEX1. We can see that HEX1 made a slight better forecast of storms in the northeast Kansas (the blue circles) and south Missouri (the red circles). It is interesting that NoRW successfully predicted the strong storm in the southwest Missouri and outperformed CTRL in this situation (Fig 5.7 and Fig 5.6). One possible reason for this is that in the CTRL experiment, using large horizontal impact scale (about 300 km) and large ensemble localization scale (110 km) smoothed too much the convective radial wind observations, lost small scale information and may generate some spurious large-scale wind field (see Figs. 4.5 and 4.6) which in turn degrades the individual storm forecasts.
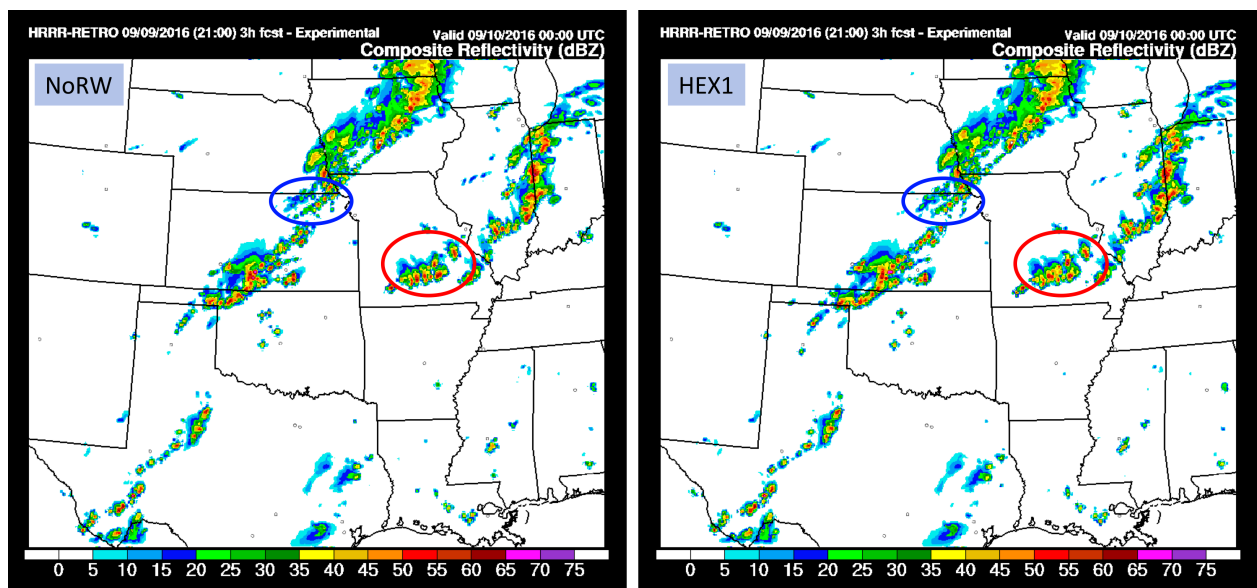
Figure 5.7. simulated composite reflectivity from 3h forecasts for NoRW and HEX1 valid at 20160910 00:00 UTC

Fig. 5.8 compared the forecasts between HEX2 and HEX1. We can see there is no evident difference between HEX2 and HEX1. This means that in current data assimilation configuration, the default superob setting provides enough convective scale information. Using smaller superob cells does not yield evident benefits. However, keep in mind that in HEX1 and HEX2, the impact scales are about 20 km. If in the future, we elect to use smaller impact scale, adopting smaller superob cells may bring extra benefits.

Fig. 5.9 compared the forecasts between NEX1 and HEX1. We can see that HEX1 made a better forecast of storms in the east Kansas (marked by the red circles in the figures) than NEX1. The storms were stronger and lined up better in HEX1, matching observations better. This indicates that although including the GDAS ensemble covariance in the hybrid 3DEnVar does improve the analysis evidently, but it does improve forecasts of individual storms. Effort should be made to use hybrid 3DEnVar as much as possible in high resolution convective-scale data assimilation.
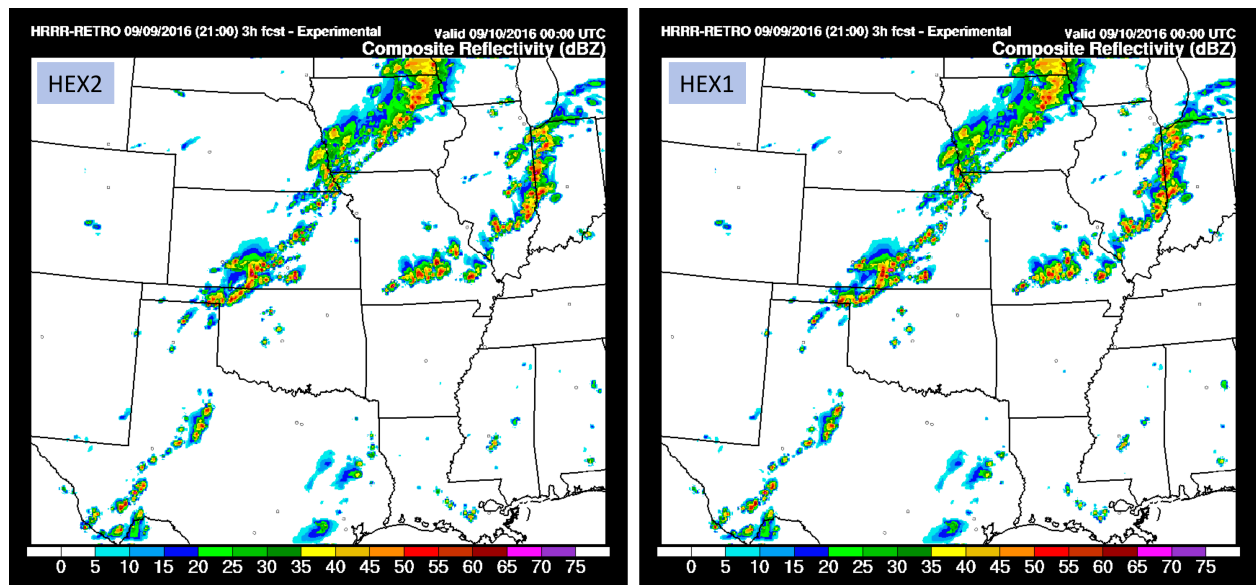
Figure 5.8 simulated composite reflectivity from 3h forecasts for HEX2 and HEX1 valid at 00:00 UTC, 20160910
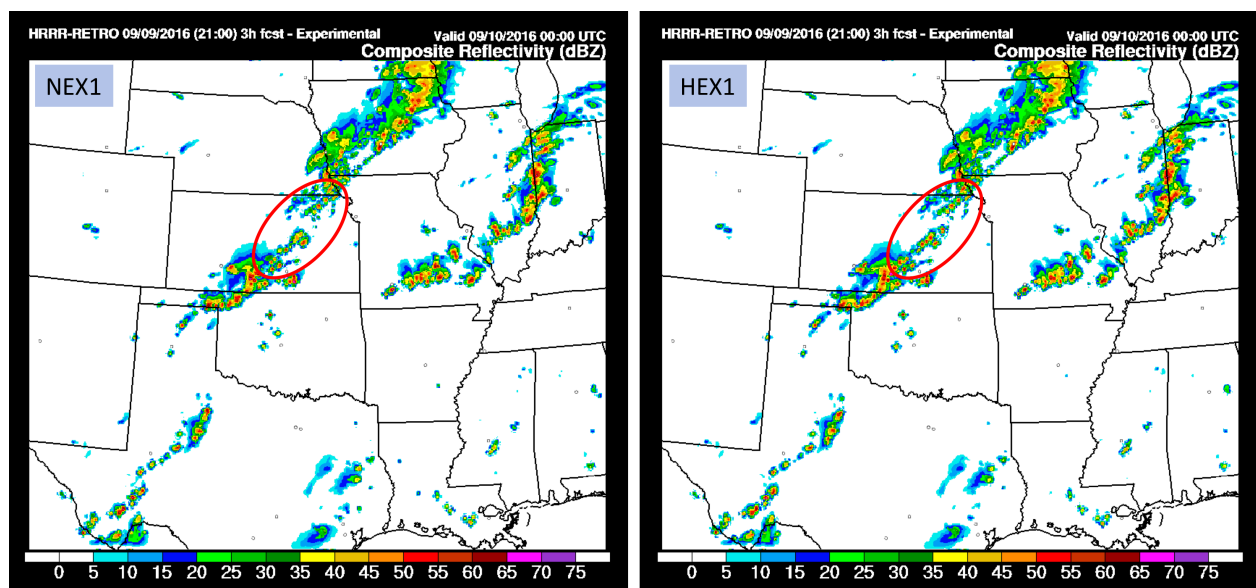
Figure 5.9 simulated composite reflectivity from 3h forecasts for NEX1 and HEX1 valid at 00:00 UTC 20160910

Fig. 4.10 compared the forecasts between NEX1 and NEX2. We can see that NEX2 made a better forecast of storms in east Kansas (the blue circles in the figures) and south Missouri (the red circles in the figures). NEX2 predicts stronger storms and better line up of storms. This indicate that although using smaller superob cells does not bring evident benefits when using the hybrid 3DEnVar method (comparison of HEX1 and HEX2), it does improve the forecasts when using pure 3DVar.
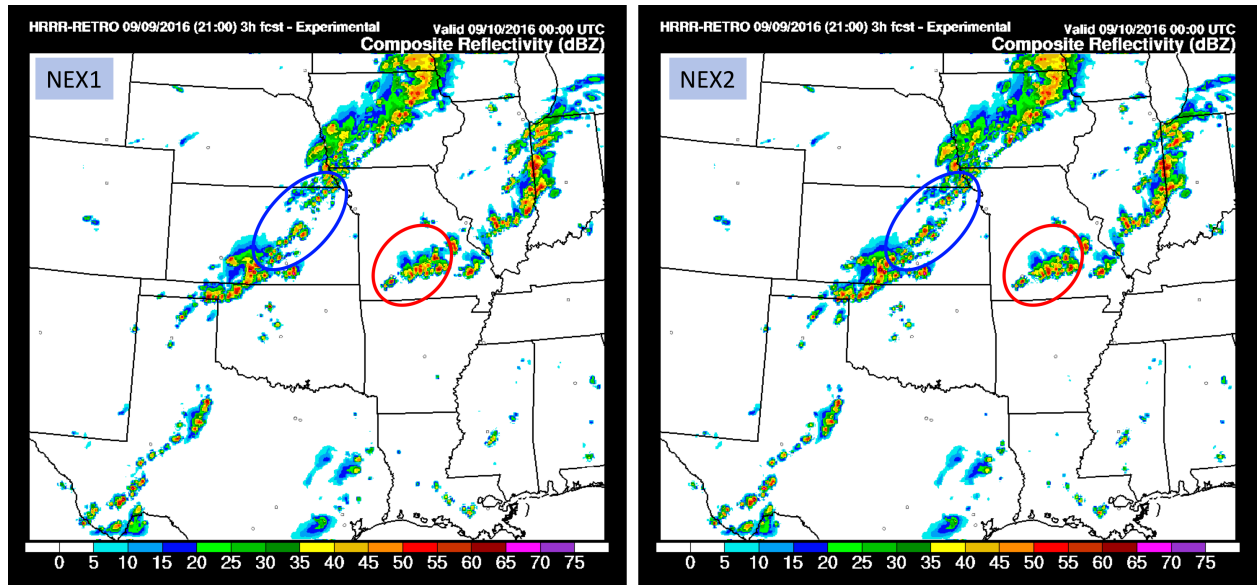
Fig. 5.10 simulated composite reflectivity from 3h forecasts for NEX1 and HEX1 valid at 00:00 UTC 20160910

The above findings are also valid for 1h and 2 h forecasts. After 3 hours, the difference between experiment are almost negligible.

5.3 Experiment setups for low resolution RW analysis

Table 5.3 lists the same one-week long retrospective experiments as section 5.1 to study the use of radial winds (RW) in low resolution 12-km analysis grid with coarse super-obbing cells. All experiments followed similar workflow and configuration as the real-time RAPv4/HRRRv3 system and assimilate all available conventional observations, radar reflectivity and lightning observations, NASA LaRC cloud products and aircraft observations.

Table 5.3 List of experiments and different parameter settings for the assimilation of RW

Experiment	Assimilation strategy for RW (radial winds)	Horizontal scale for RW	Localization scale for RW	Superob setting
CTRL	Hybrid 3DEnVar at 12 km	0.373,0.746,1.5	h=110, v=3	5deg, 5km, 60min, minimum 50 points
NoVAD	Hybrid 3DEnVar at 12 km	0.373,0.746,1.5	h=110, v=3	NO VAD wind
RW12km	Hybrid 3DEnVar at 12 km	0.373,0.746,1.5	h=110, v=3	13deg, 12km, 60min, minimum 50 points
RW36km	Hybrid 3DEnVar at 12 km	0.373,0.746,1.5	h=110, v=3	42deg, 36km, 60min, minimum 50 points

Three experiments, CTRL, RW12km and RW36km, are conducted to evaluate the forecast impact from radial velocity when superobed with coarse superob cells. The NoVAD is to see if the impact of VAD assimilations, which is a product retrieved from the radial velocity. All four sets of experiments were conducted over a limited domain covering the central U.S, and for a period slightly more than one week, starting at 12Z on 03 Sep 2016 and ending at

23Z on 10 Sep 2016. For each hourly initialization time in that period (180 total model cycles), hourly HRRR forecasts were made out to a lead-time of 12 h. To reduce the volume of verification data, the following lead times were examined: 0, 1, 3, 6, 9, and 12 h. The NCAR Cheyenne was used for running the experiments. The verification was primarily conducted using the Model Evaluation Toolkit (MET) verification system.

5.4 Experiment results for low resolution RW analysis

Both vertical verification against sounding and surface verification against METAR station were conducted with MET to compare the model performance for three experiments (NoVAD, RW36km, and RW12km) to see which configuration of the radar radial wind assimilation resulted in positive impact to the forecast.

a. Vertical verification

All of the vertical experiment comparison plots show very little visible differences. Figure 5.11 shows what may be the largest differences between the experiments out of the entire set, which occurred for RMS error for dewpoint temperature for the 6-h lead time. Some sort of anomaly seemed to have occurred at this forecast lead time, as some other plots show RMS errors as large or larger at 6-h lead time as at the 12-h lead time.
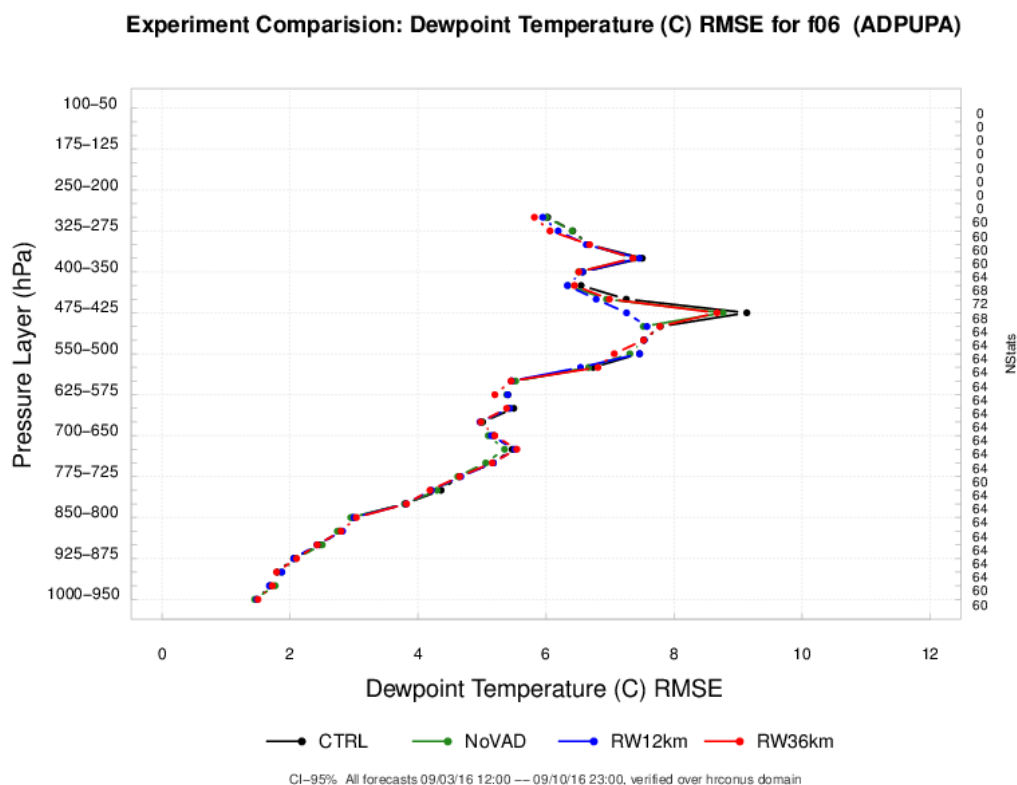
Figure 5.11: Comparison of three experiments and the control run for dewpoint temperature RMS error for the 6-h forecast lead time: control (black), NoVAD (green), RW12km (blue), RW36km (red).

But in general, for all variables and all lead times and all levels, the three experiments were largely indistinguishable from each other and from the CTRL experiment, as seen in Figure 5.12. This figure shows that all the confidence bounds for the three experiments overlap

each other and those of the control run. The experiment comparison plots for other variables for bias show similar results. For completeness, several additional variable/statistics are shown. Figure 5.13 shows the temperature RMS errors. Figure 5.14 shows the bias for zonal wind. Figure 5.15 shows the bias for meridional wind. Slight differences are evident, but again, none of the differences are statistically significant. Figure 5.16 shows the specific humidity bias. Figures 5.17 and 5.18 show the bias and RMS error, respectively for relative humidity. To summarize the results of this suite of plots, there is little basis to say that one experimental configuration performed better than any of the others (or control).

Experiment Comparision: Dewpoint Temperature (C) RMSE for f06 (ADPUPA)
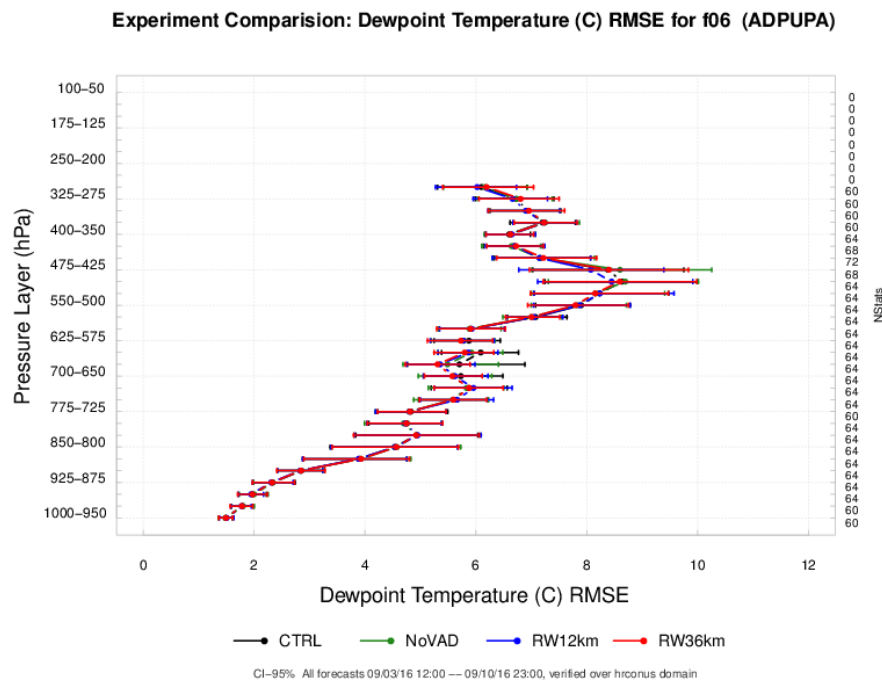
Figure 5.12: As in Fig. 5.11, but now with 95% confidence intervals.

Experiment Comparision: Temperature (C) RMSE for f12 (ADPUPA)
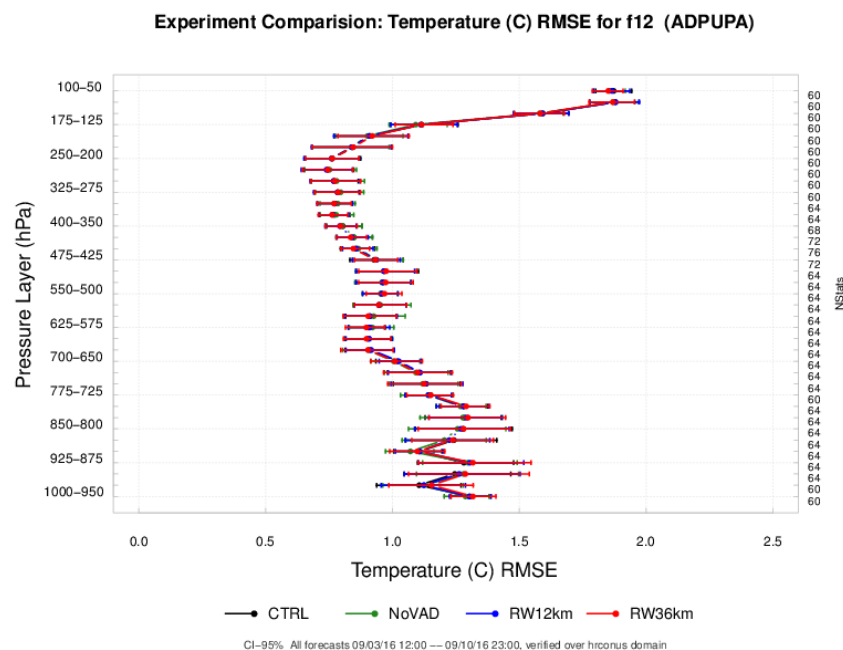
Figure .5.13: As in Fig. 5.11, but for temperature RMS error for the 12-h forecast lead time.

Experiment Comparision: U–component Wind (m_s) Bias for f12 (ADPUPA)
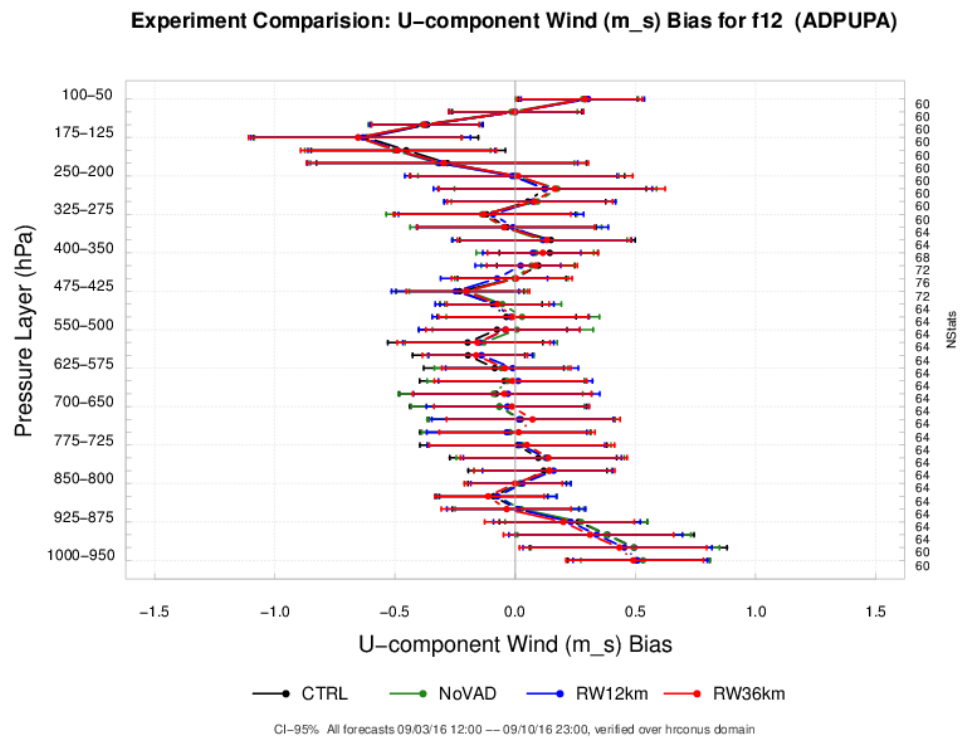
Figure 5.14: As in Fig. 5.11, but for zonal component wind bias for the 12-h forecast lead time.

Experiment Comparision: V–component Wind (m_s) Bias for f12 (ADPUPA)
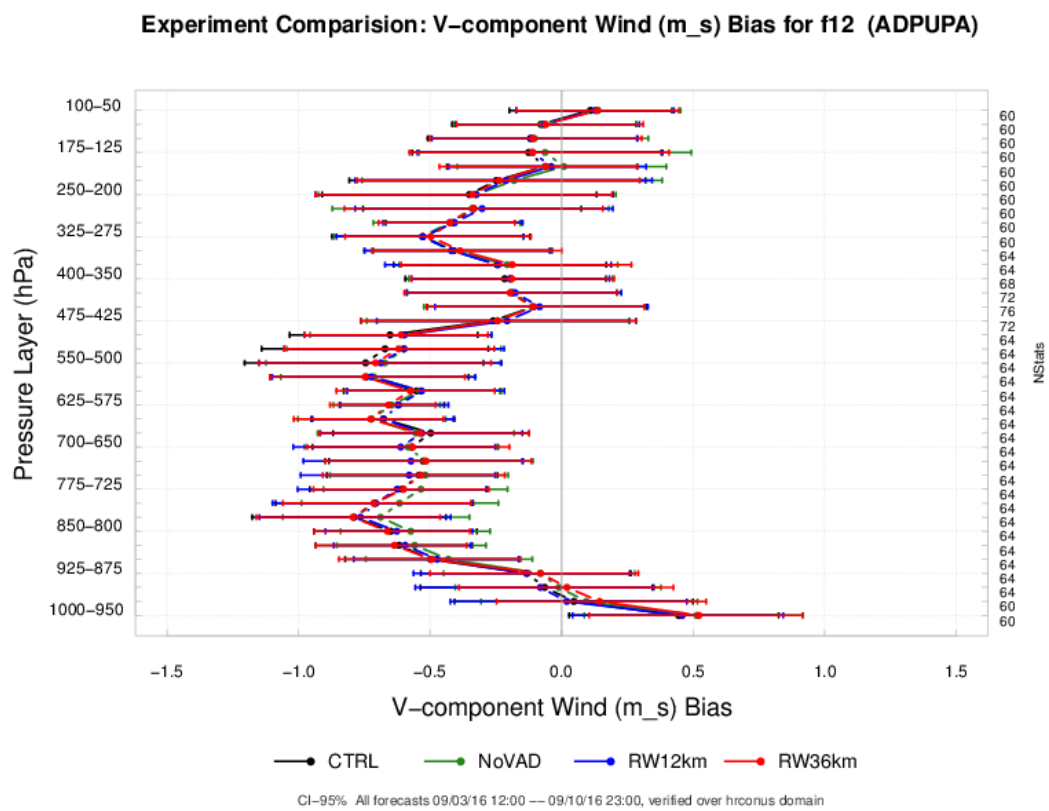
Figure 5.15: As in Fig. 5.11, but for meridional component wind bias for the 12-h forecast lead time.

Experiment Comparision: Specific Humidity (kg_kg) Bias for f06 (ADPUPA)
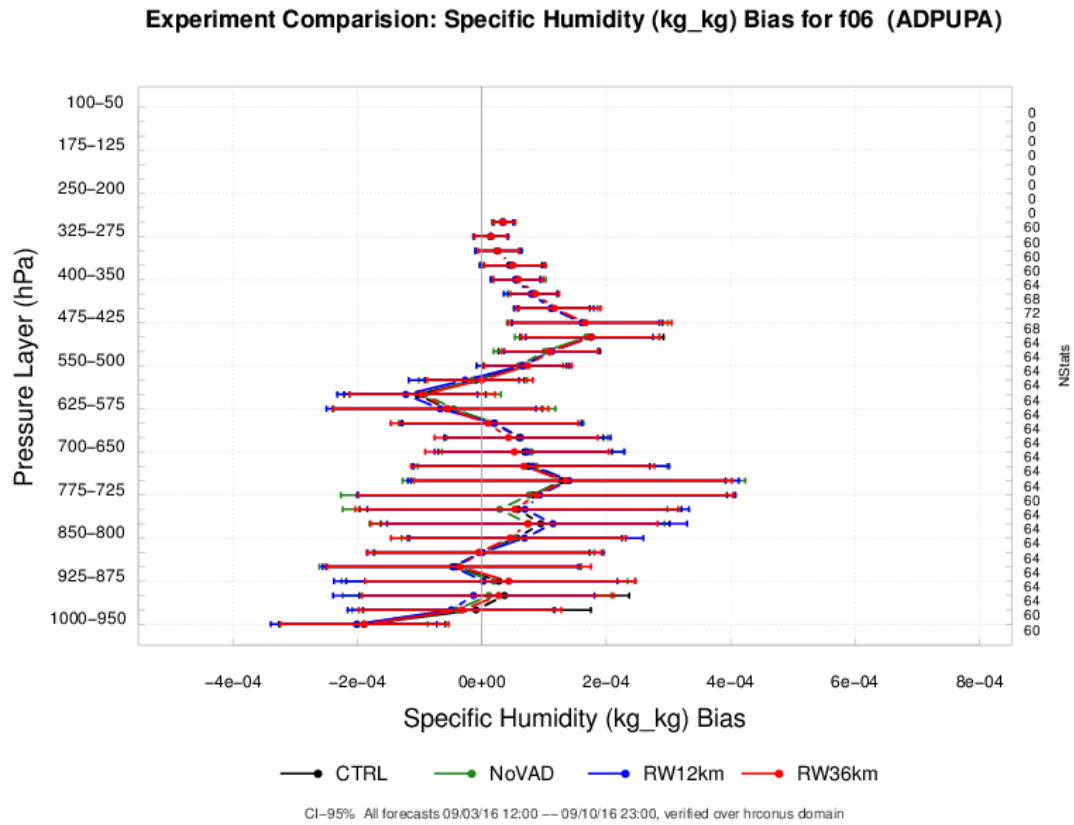
Figure 5.16: As in Fig. 5.11, but for specific humidity bias for the 12-h forecast lead time.

Experiment Comparision: Relative Humidity (%) Bias for f12 (ADPUPA)
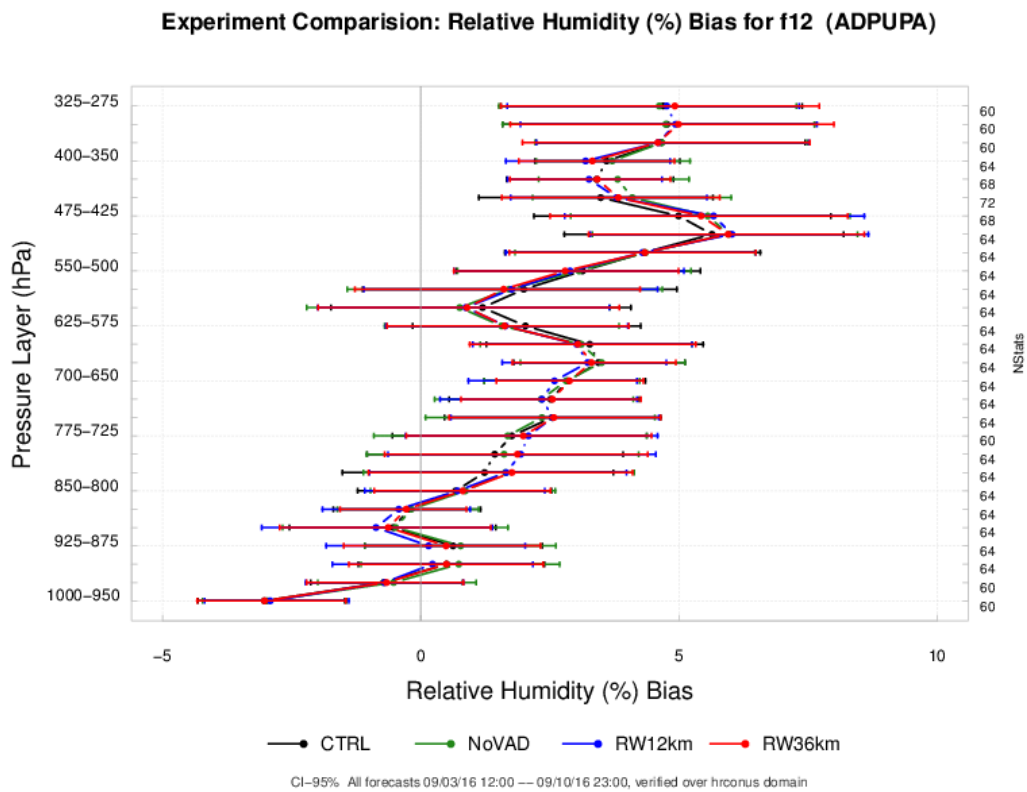
Figure 5.17: As in Fig. 5.11, but for relative humidity bias for the 12-h forecast lead time.

Experiment Comparision: Relative Humidity (%) RMSE for f12 (ADPUPA)
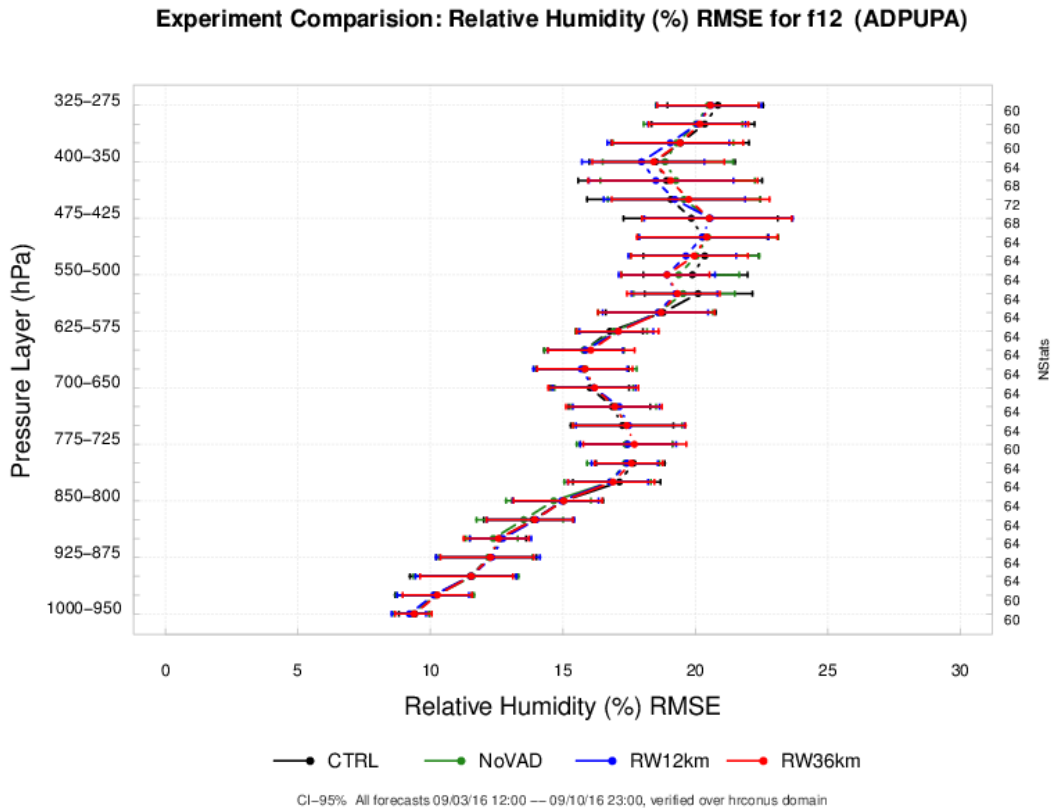
Figure 5.18: As in Fig. 5.11, but for relative humidity RMS error for the 12-h forecast lead time.

b. *Vertical Experiment Differences*

Because the differences between the experiments and control are so slight, it is helpful to plot the differences (e.g., experiment – control). These types of plots are particularly handy for seeing at a glance whether any differences are statistically significant. Any differences that are statistically significant will be indicated by confidence bounds that do *not* overlap the zero-reference line. Figure 5.19 shows an example for the relative humidity bias for the 12-h lead time for NoVAD – CTRL. At a handful of vertical layers, some statistical differences are seen (e.g., 450 hPa, 600 hPa, 800 hPa). Remembering that for the 95% confidence level, we expect to see false statistical significance 5% of the time, this example shows there is a possibility that there are some significant differences at certain levels, but no strong pattern of such differences. Figure 39 displays the NoVAD – CTRL experiment differences for RMS error. Again, there are a few statistically significant differences at a few middle and lower layers, but most layers are not statistically distinguishable. The difference for the RW12km – CTRL experiment and for the RW36km – CTRL are similar, no clear pattern and no statistically significant differences (Those plots do not show)

Experiment Differences: Relative Humidity (%) Bias for f12 (ADPUPA)
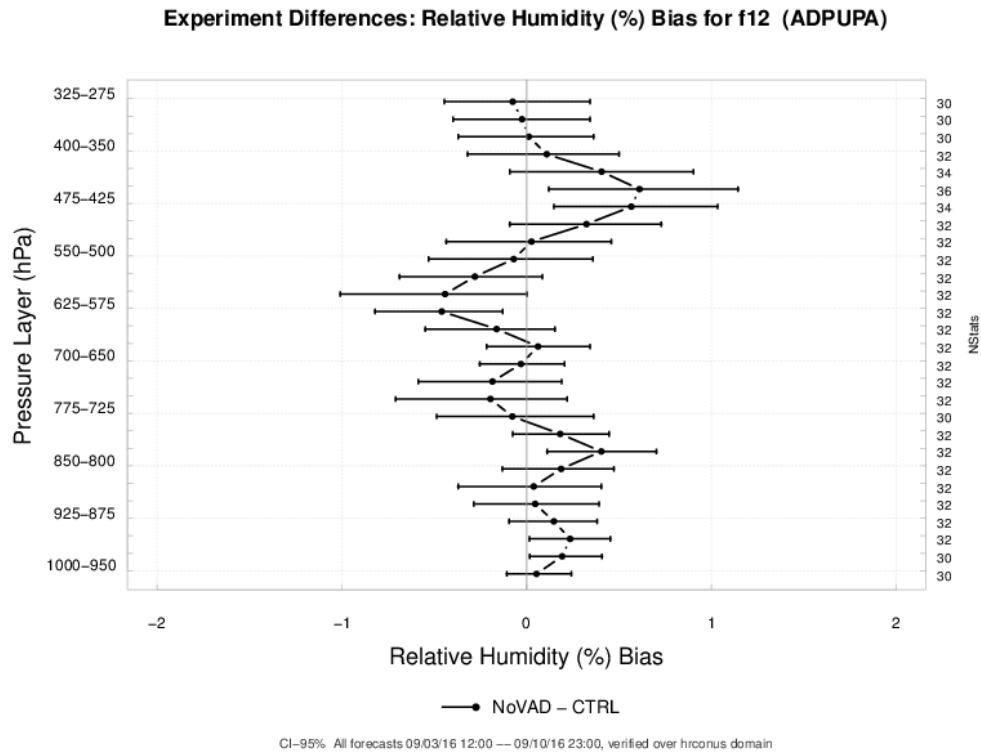
Figure 5.19: Plot showing the difference between NoVad and the CTRL experiments for the relative humidity bias for the 12-h lead time. Error bars indicate the 95% confidence intervals on the difference.

Experiment Differences: Relative Humidity (%) RMSE for f12 (ADPUPA)
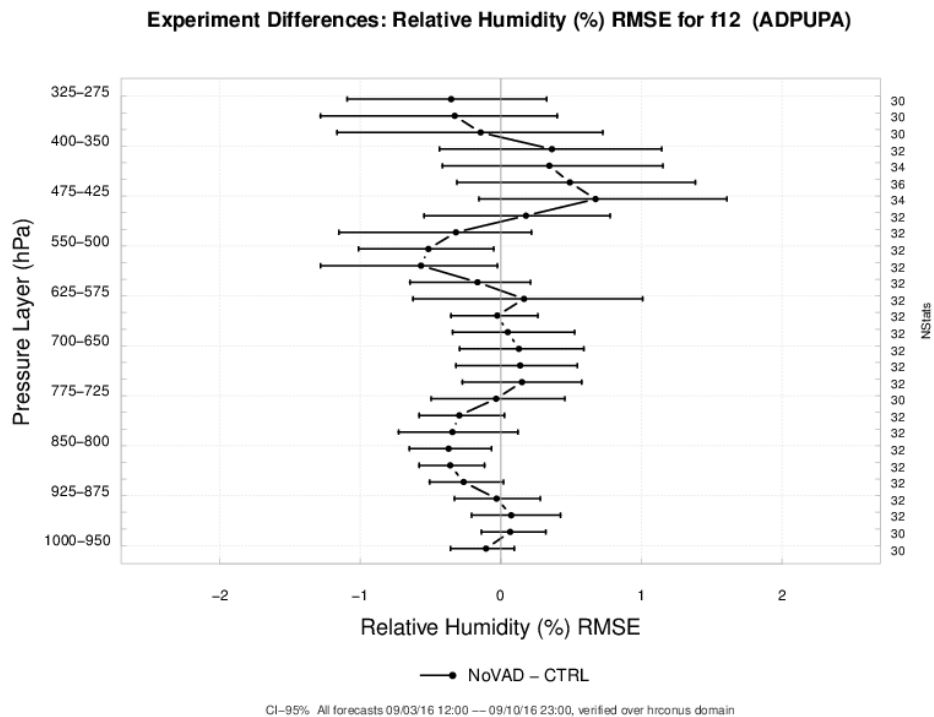
Figure 5.20: As in Fig. 5.19, but for RMS error.

To summarize the full suite of difference plots for each experiment:

- NoVAD: no systematic improvement or degradation over CTRL.
- RW36km: no systematic improvement or degradation over CTRL.

46

- RW12km: no systematic improvement or degradation over CTRL.

These results indicate that the innovations in the various experiments had neutral impact on the HRRR model's performance. While not showing any clear improvements over CTRL, the experiments also did not degrade the forecast compared with CTRL.

c. *Time Series Comparisons by Initialization Time*

To understand how the RMS error and bias characteristics vary across many model cycles, it is useful to plot a time series comparison of the verification statistics by initialization time for the various lead times. This allows one to see at a glance whether there are model "dropouts" when the model performance declines significantly, or whether there are any periods when a specific model configuration outperforms others. These types of plots can also show whether there are any diurnal characteristics to the model performance. Figure 5.21 shows the bias of dewpoint temperature across all initialization times in the experiment period for the 12-h forecast lead time for surface observations (ADPSFC). The first thing one may note is that there are distinct times during the day when model performance tends to be better or worse, however these trends are not completely tied a specific time. In general, the forecasts initialized at or just before 12Z tend to be worse, while the forecasts initialized around 00Z tend to be the best. This pattern breaks down some days however, as noted by relatively poor performance at 00Z on Sep 04. These fluctuations in performance may be related to large scale convective events occurring in the domain. The next thing to note about this figure is that by and large, the three experiments and control are normally very close to each other. In fact, no dropouts are evident in this first example.

Figure 5.21: Dewpoint temperature bias for all initialization times for the 12-h forecast lead time: CTRL (black), NoVAD (forest green), RW12km (blue), RW36km (red).

Comparison by Init. Time: Specific Humidity (kg_kg) RMSE for all init times, f12 (ADPSFC)
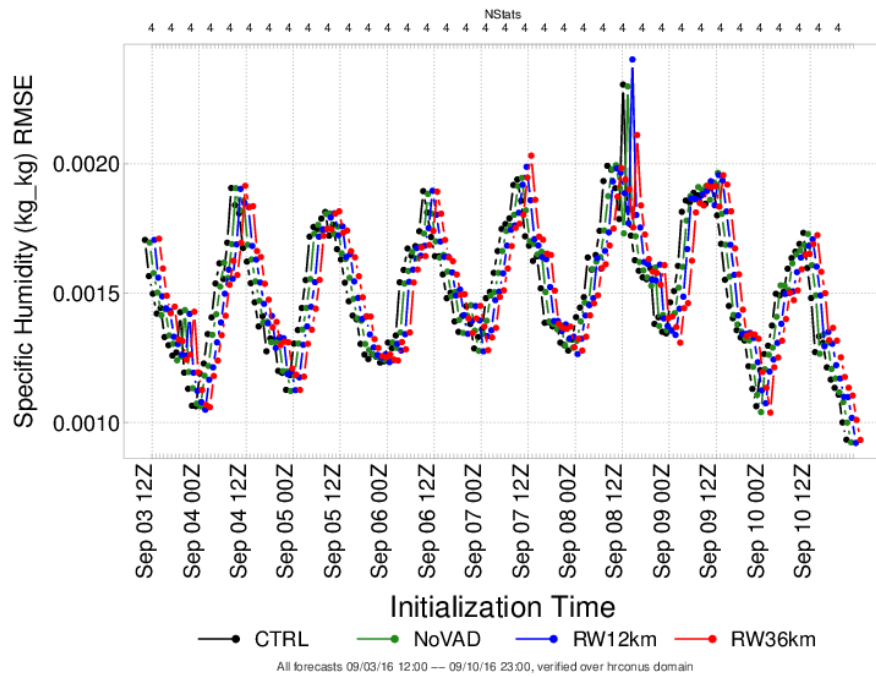
Figure 5.22: As in Fig. 5.21, but for specific humidity RMS error.

Figure 5.22 shows an example in which the RMS error spiked for one particular initialization time (around 12Z on 08 Sep). In this case, all three experiments and CTRL suffered a spike in RSM error at 12Z on Sep 08.

Comparison by Init. Time: U−component Wind (m_s) Bias for all init times, f06 (ADPSFC)
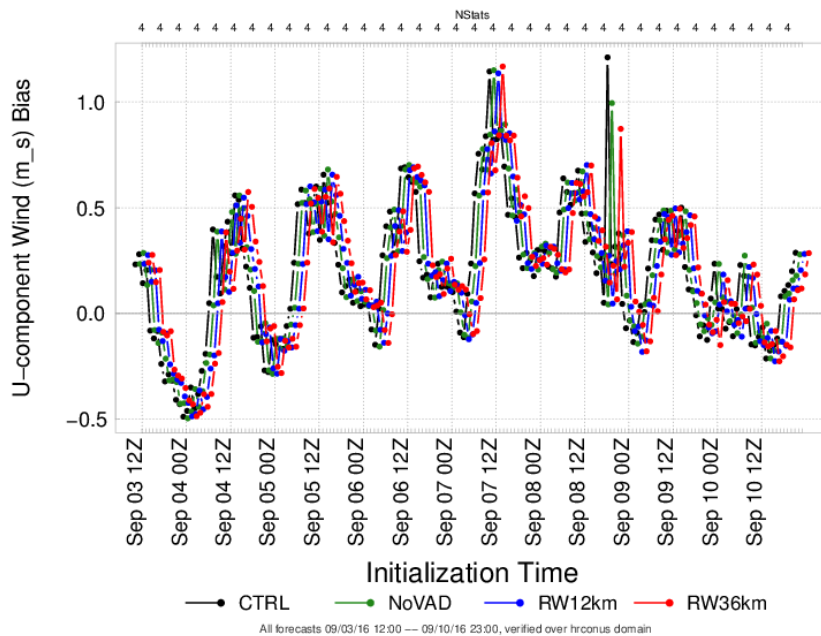
Figure 5.23: As in Fig. 5.21, but for zonal velocity bias for the 6-h lead time.

Figure 5.23 shows an example in which the RMS error spiked for one particular initialization time (around 12Z on 08 Sep). In this case, CTRL, NoVAD, and RW36km had a spike in the bias, but RW12km did not.

d. Time Series Differences by Initialization Time

A more detailed picture of the changes in performance across initialization times can be seen by taking the difference of the three experiments from the control simulation. This allows the relative performance of the experiments to be seen much more easily, although one must keep in mind that the vertical range on many of the plots extends over a very small set of values. This may magnify very small differences.

Figure 5.24: Experiment differences by initialization time for dewpoint temperature RMS error computed from surface observations (ADPSFC) for the 12-h lead time: NoVad – CTRL (green); RW12km – CTRL (blue); RW36km – CTRL (red).

Figure 5.24 shows how the experiment differences (experiment – control) vary by initialization time for the 12-h forecast lead time, computed from surface observations (ADPSFC). None of the three experiments offers any consistent performance advantage, with each performing better or worse at various times. The NoVAD and RW12km both have several notable degradation spikes, but RW36km does not.
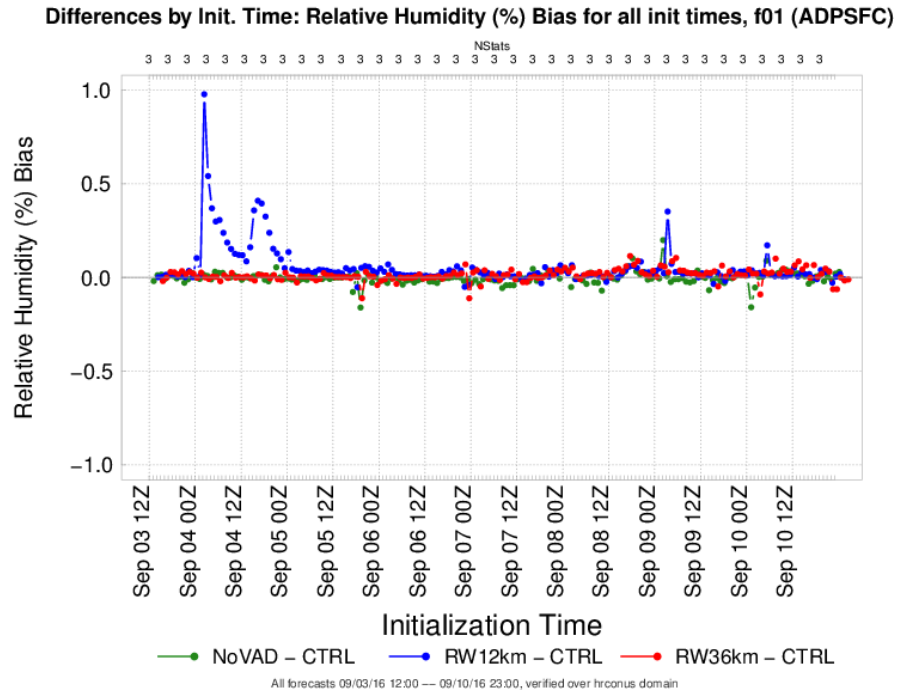
Figure 5.25: As in Fig. 5.24, but for relative humidity bias for the 1-h lead time.

Figure 5.25 shows the differences by initialization time for relative humidity bias for the 1-h lead time. This plot is interesting because it shows the RW12km experiment had some notable jumps in bias even at an early lead time.
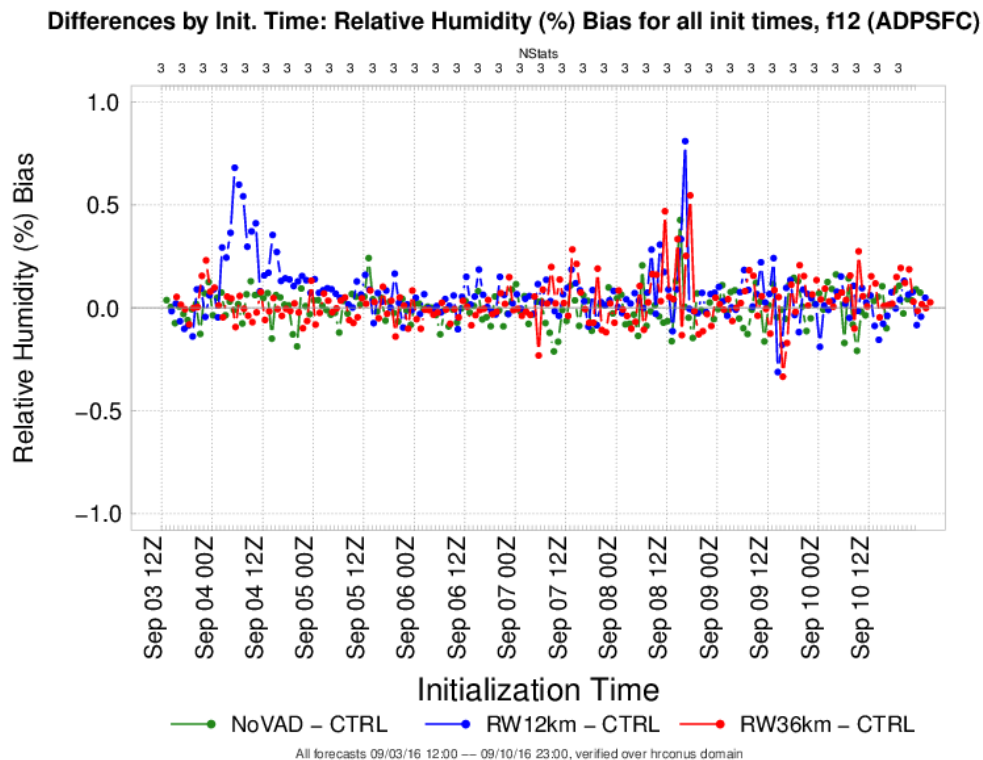
Figure 5.26: As in Fig. 5.25, but for the 12-h lead time.

Meanwhile, Fig. 50 shows the situation for the 12-h forecast lead time. The initial jump in relative humidity bias has persisted through the forecast, which taken at face value, might lead one to think that this indicate some sort of (negative) impact of the radar winds for those particular periods for the RW12km experiment. The RMS error for this lead time (not shown) does not show a similar degradation however, so it is likely that this may simply be due to the noise of looking at individual forecast.

e. Time Series Differences by Forecast Lead Time

The final battery of plots shows how the experiment differences vary by forecast lead time. Again, the experiment differences are computed by subtracting the CTRL value of the verification statistic from the corresponding value of the experiment (e.g., NoVAD – CTRL). For RMS error, the performance of a given experiment is statistically significant positive (negative) if the confidence intervals are entirely below (above) the zero-reference line.
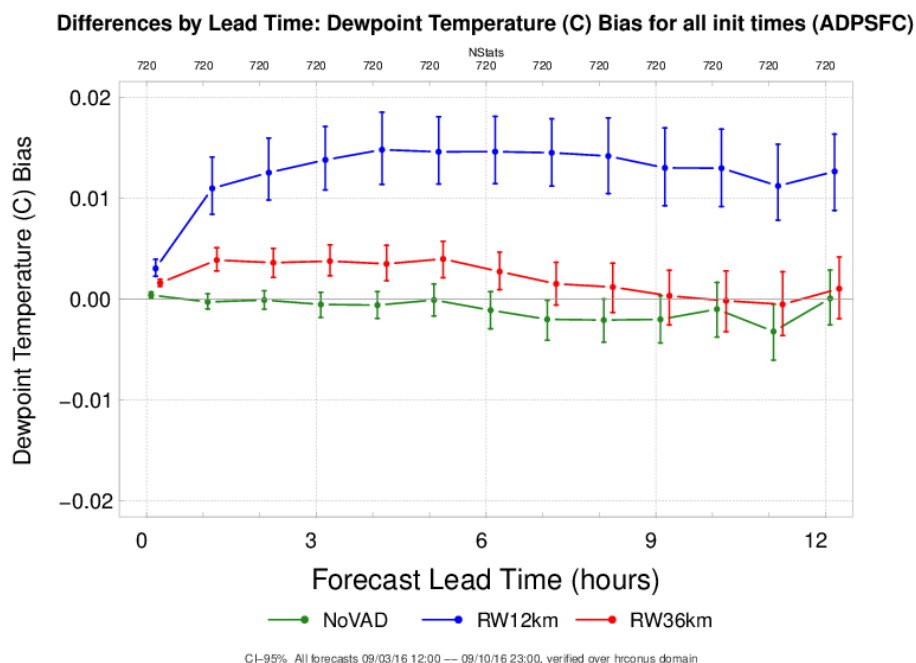
Figure 5.27: Variation of dewpoint temperature bias computed from surface observations (ADPSFC) by forecast lead time for the three experiments: NoVAD – CTRL (forest green), RW12km - CTRL (blue), RW36km – CTRL (red).

Figure 5.27 shows how the experiment differences varies by forecast lead time for the three experiments. Interestingly, the RW12km experiment has a larger bias across most forecast lead times, with statistically significant differences from the CTRL run. The RW36km also has statistically significant biases larger than CTRL at early and intermediate lead times. It must be noted, however, that these bias differences are very small.

Differences by Lead Time: Dewpoint Temperature (C) RMSE for all init times (ADPSFC)
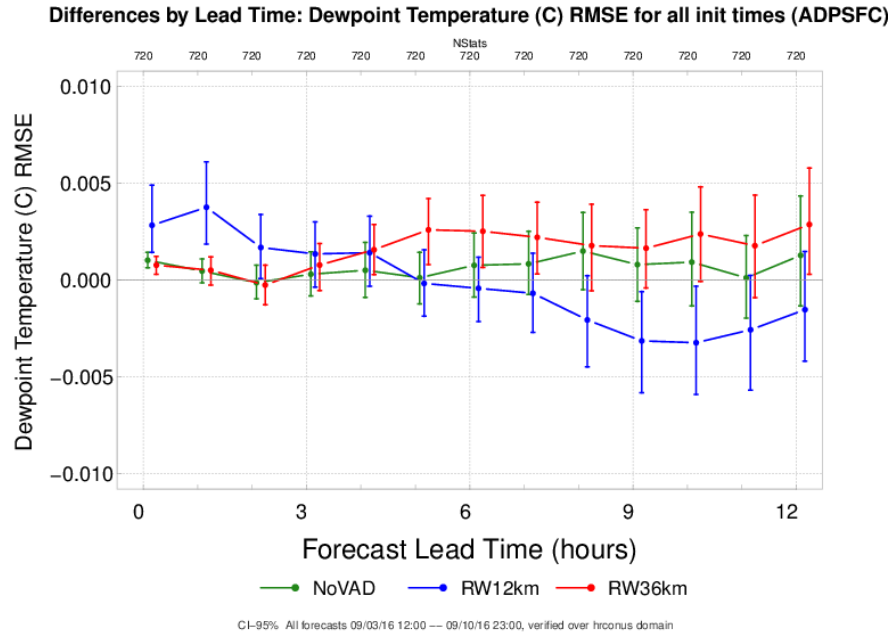
Figure 5.28: As in Fig. 5.27, but for dewpoint temperature RMS errors.

Figure 5.28 shows the picture for the RMS error of dewpoint temperature. The RW12km experiment possesses statistically significant differences (degradation) at a couple early lead times, but has statistically significant improved performance at the 9 and 10 h lead times. The RW36km experiment has statistically significant degraded performance relative to the CTRL run at the 4-7 and 12-h lead times. The NoVAD experiment does not have any statistically significant differences at any forecast lead times.

Differences by Lead Time: Temperature (C) Bias for all init times (ADPSFC)
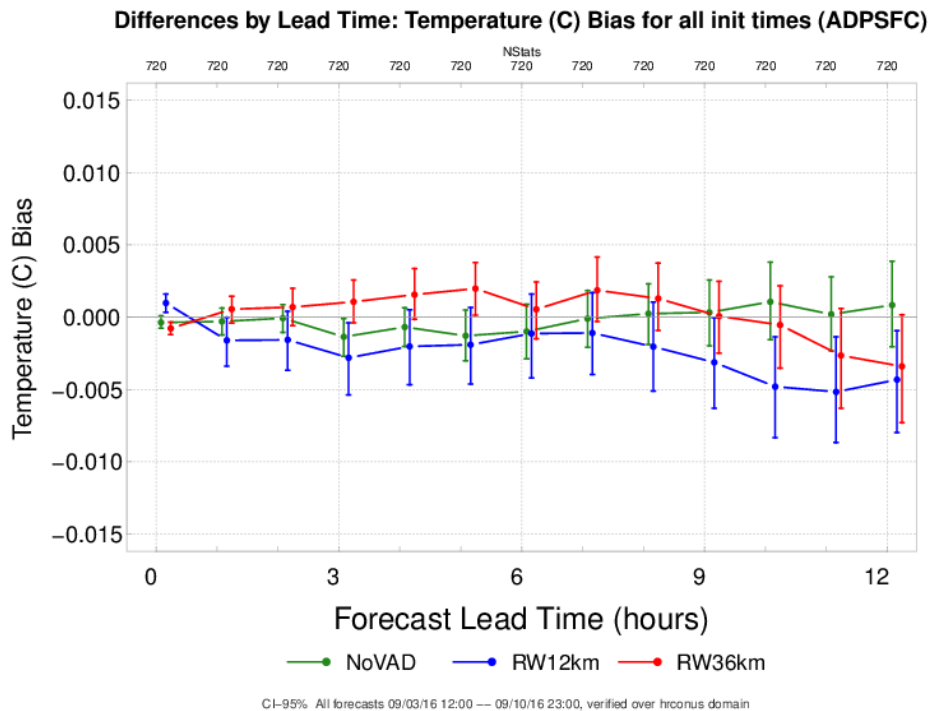
Figure 5.29: As in Fig. 5.27, but for temperature bias.

Figure 5.29 shows the experiment differences for temperature bias for the three experiments. While the biases for NoVAD and RW36km are not notably different from the

CTRL run, the RW12km experiment tends to have a tendency toward a low bias, with statistical significance at 3 h and 10-12 h.
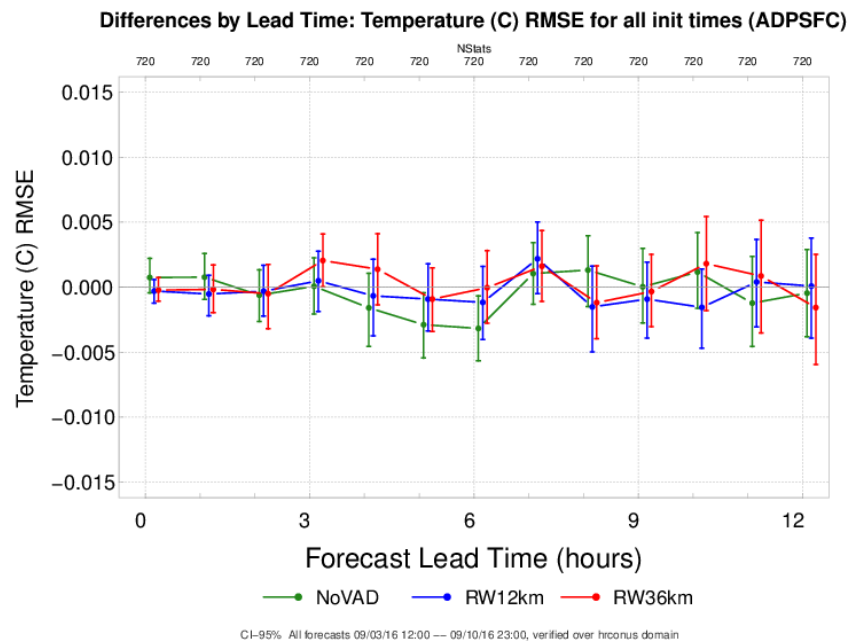
Figure 5.30: As in Fig. 5.27, but for temperature RMS error.

Figure 5.30 shows the experiment differences for temperature RMS error. All three experiments are quite close to the CTRL, expect for a couple periods of statistically significant improvement for the NoVAD experiment at 5 and 6 h.

To sum up the broad results of this section, most experiments at most forecast lead times were not statistically from the CTRL run, however the RW12km experiment did show some statistically significant improvements, mainly for the moisture variables at the longer lead times. This hints at the possibility that the innovations in the RW12km experiment had a positive impact, compared with CTRL. Beside this one finding for moisture, taken altogether with the other batteries of plots, we cannot say that any experiment showed a notable pattern of improved performance relative to CTRL. The preponderance of evidence points to neutral impact for all three experiments. At best, we can say that there is a hint of positive impact for the RW12km experiment for moisture.

f. FSS for precipitation

Besides the above verification against conventional observations using MET, verification against the Stage IV rainfall data, in terms of the FSS (Fractions skill score) method, was also conducted. The fractions skill score (FSS) method compares the forecast and observed rain fractional coverage within spatial neighborhoods. The fractions skill score is based on variation of the Brier score used to verify probability forecasts. The following formulas show how to calculate the FSS using the fractional forecast (Pfcst) and observed (Pobs) rain areas in each neighborhood. The FSS ranges from 0 for a complete mismatch to 1 for a perfect match.

$$\text{Fractional Brier Score (FBS)} = \frac{1}{N}\Sigma_N(P_{fcst} - P_{obs})^2$$

$$\text{FSS} = 1 - \frac{FBS}{\frac{1}{N}(\Sigma_N {P_{fcst}}^2 + \Sigma_N {P_{obs}}^2)}$$

Figure 5.31 shows the FSS score for the four set of experiments (CTRL, NoVAD, RW12km and RW36km) for different precipitation threshold (1mm and 2.5mm). Overall the FSS score decreases with forecast lead time for both 1mm and 2.5mm and there is no noticeable difference between the experiments. This is consistent with the findings from the MET verifications that there is no notable improvements for all three experiments, compared to the control runs.
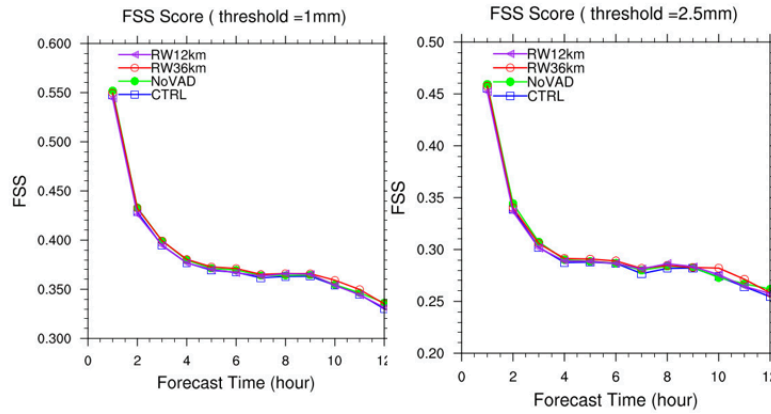
Figure 5.31 FSS score for the four set of experiments (CTRL, NoVAD, RW12km and RW36km) for different precipitation threshold (1mm and 2.5mm).

6.   February 20, 2017 Winter Retrospective Study

As discussed in the above sections, the differences between the experiments with difference radial velocity assimilation configuration and the control runs are small and the impact from tuning the superob parameters are neutral. One factor that might contribute to this minimal impact could be due to the limited data volume in the radial wind, especially after the strict quality control during the NCEP data processing. In this section we chose a winter retrospective case on February 20, 2017, when the radial wind data seems to have a better coverage. Since the convective system for this case is mostly in Oklahoma and Texas (Fig. 6.1), a new test domain has been set up (Fig. 6.2) and the initial and boundary conditions are generated for this new case.
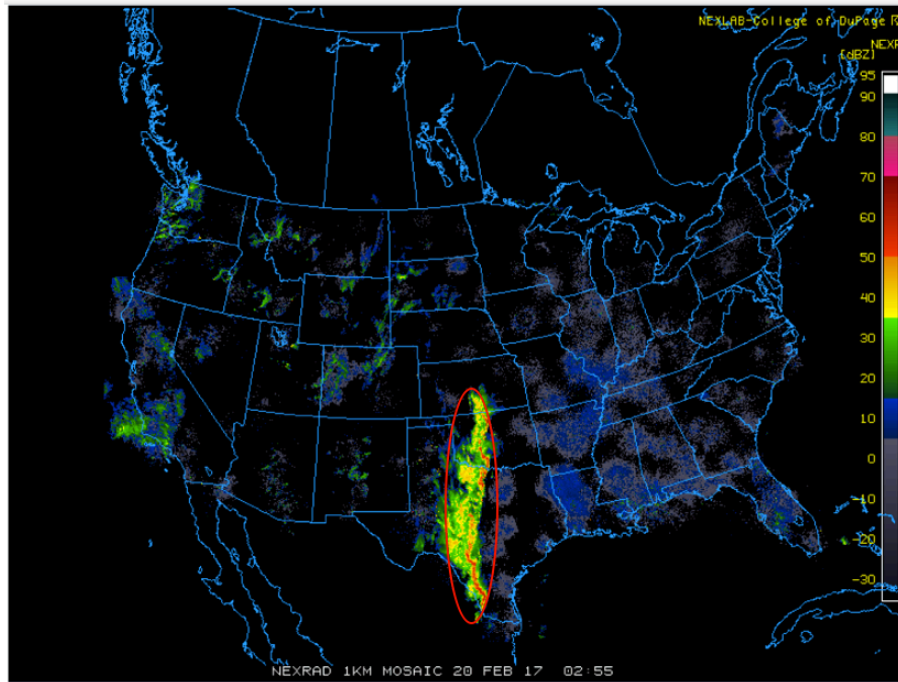
Figure 6.1 NEXRAD 1km radar composite at 02:55Z of February 20, 2017.
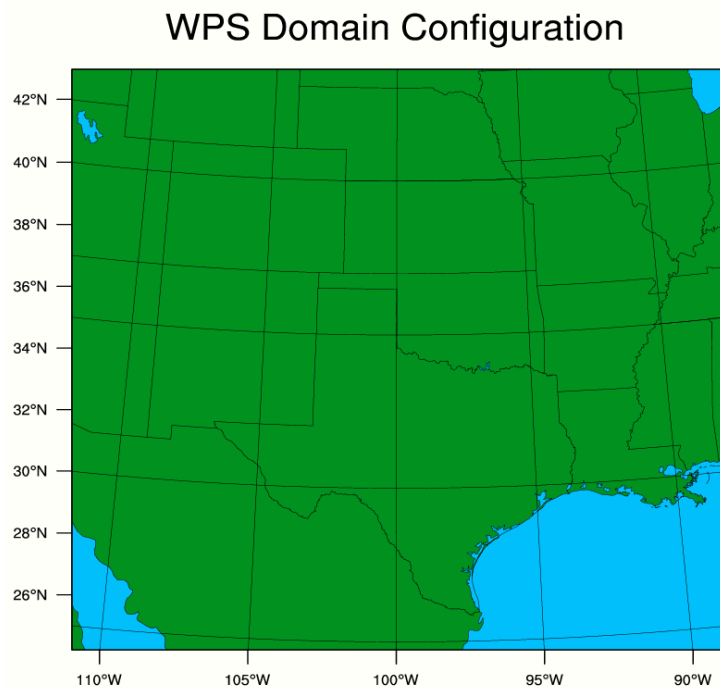
Figure 6.2 The new test domain for the February 20, 2017 case.

6.1 Radar data check

Along the convective system, two radar stations, KVNX (Vance Air Force Base in Oklahoma) and KDYX (Dyess Air Force Base in Texas), were chosen to show the radial wind coverage. Figure 6.3 shows the raw radar data from KVNX at 03Z of February 20, 2017 for the scan angle at 0.48, both showing a very good data coverage along the storm. Figure 6.4 gives the RAP bufr data coverage for scan angle 0.48 and 0.83. Comparing Fig. 6.3 and 6.4, it can be seen that the NCEP data processing kept most of the radial wind data from the raw

observations. Figures 6.5 and 6.6 are similar to Fig.6.3 and 6.4, but for KDYX. Comparing the raw data in Fig. 6.5 and RAP bufr data in Fig.6.6 suggests the radial wind data were kept during the NCEP data processing. The NEXRAD level II raw data and RAP bufr data from another radar station KSJT also show that good amount of radial wind data was kept after the quality control (not shown).
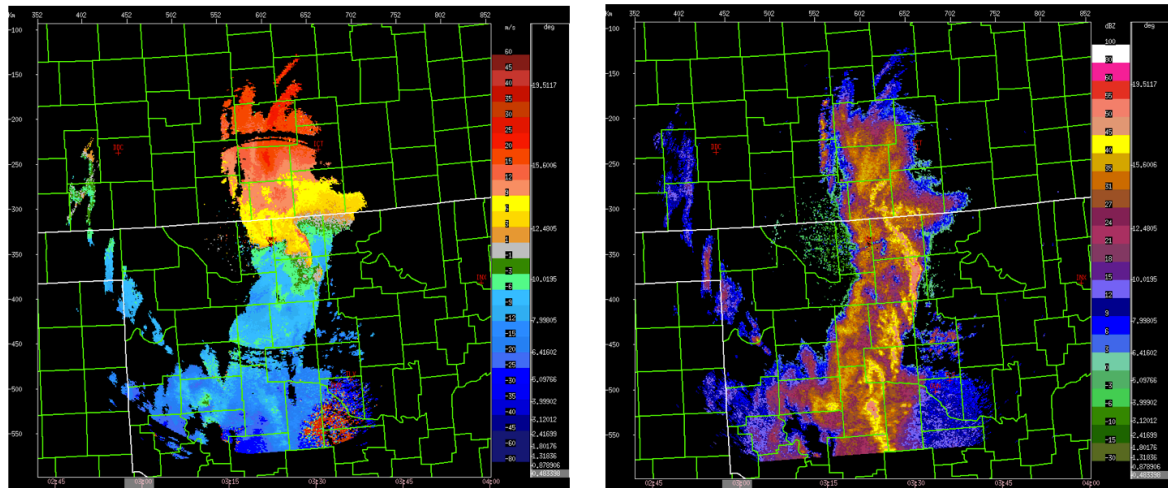
Figure 6.3 NEXRAD Level II radial wind (left) and reflectivity (right) data for KVNX at 03Z of February 20, 2017, for scan angle 0.48.
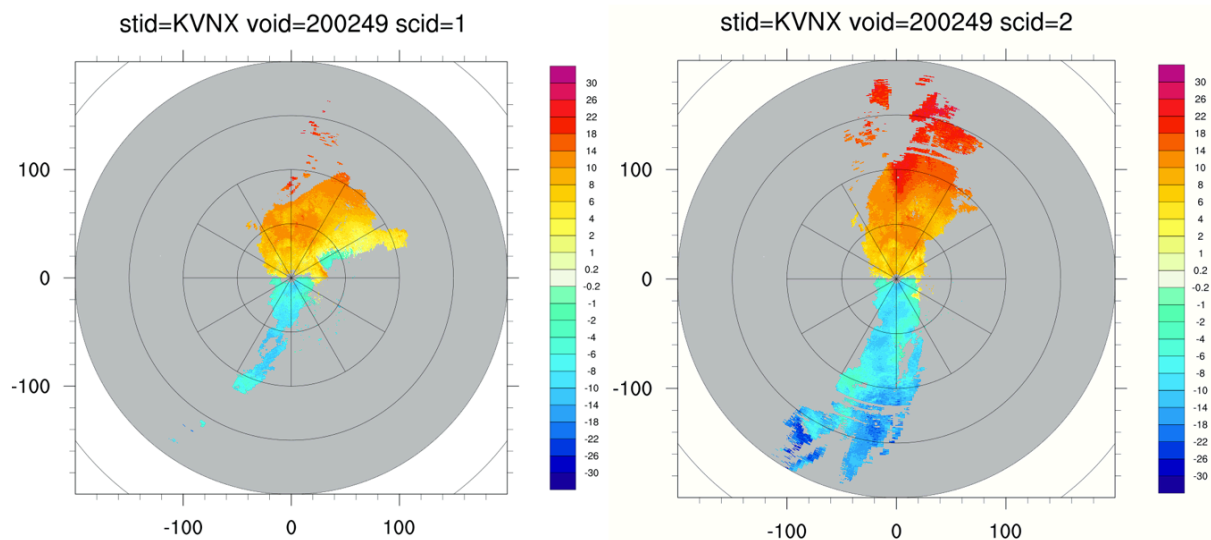
Figure 6.4 RAP bufr radial wind data from KVNX at 02:49Z of February 20, 2017, for the first (left, scan angle 0.48) and second (right, scan angle 0.83) scan.
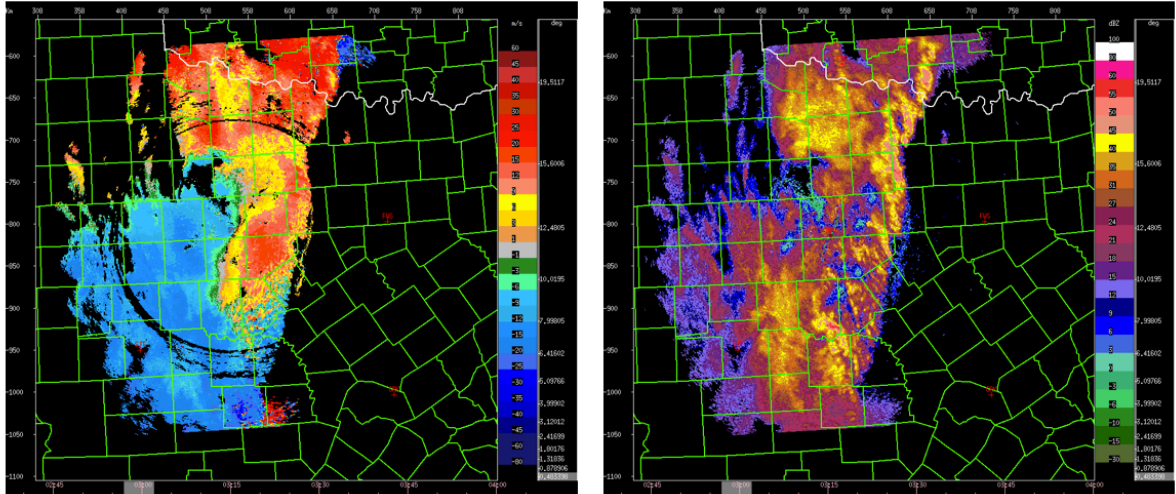
Figure 6.5 Similar to Fig. 6.3, but for KDYX at scan angle 0.48.
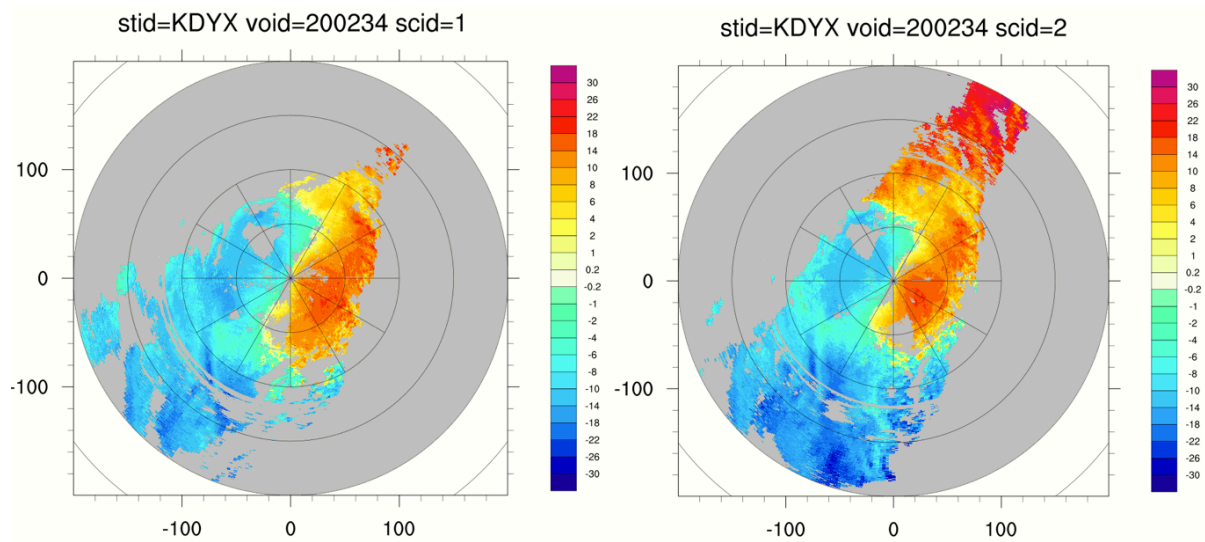
Figure 6.6 Similar to Fig. 6.4, but for KDYX at scan angle 0.53 (left) and 0.99 (right) at 02:34Z.

6.2 Experiment and results

The above data coverage analysis suggests that for this new convective case, the radial wind data has a much better coverage along the storm and therefore we wanted to see if this improved data volume would benefit the storm forecast. Four experiments were conducted, as listed in table 6.1 below.

Table 6.1 List of experiments and different parameter settings for the assimilation of RW

Experiment	Assimilation strategy for RW (radial winds)	Horizontal scale for RW	Localization scale for RW	Superob setting
CTRL	Hybrid 3DEnVar at 12 km	0.373,0.746,1.5	h=110, v=3	5deg, 5km, 60min, minimum 50 points
NoRW	n/a	n/a	n/a	n/a
PWER	Hybrid 3DEnVar at 12 km	0.373,0.746,1.5	h=110, v=3	Same as CTRL, but reduce observation error for radial wind to half, through changing erradar_inflate to 0.5 (default 1) in GSI namelist
HEX2	Hybrid 3DEnVar at 3 km	0.02487,0.04973,0.1	h=20, v=2	3deg, 3km, 15min, minimum 25 points

First, we look at the impact of with/without radial wind assimilation and tuning the BE scales by comparing CTRL with NoRW and HEX2. Figure 6.7 shows the 00-hour forecast of the maximum derived radar reflectivity for CTRL, NoRW and HEX2, compared to the observed, at 03Z of February 20, 2017. As can be seen that the 00-hour forecast of all the three experiments tends to produce a weaker than observed convective system at the initial time. As time goes by, the model tends to produce stronger than observed convections by hour 3 (Figure 6.8) but there is no notable difference among the three experiments.
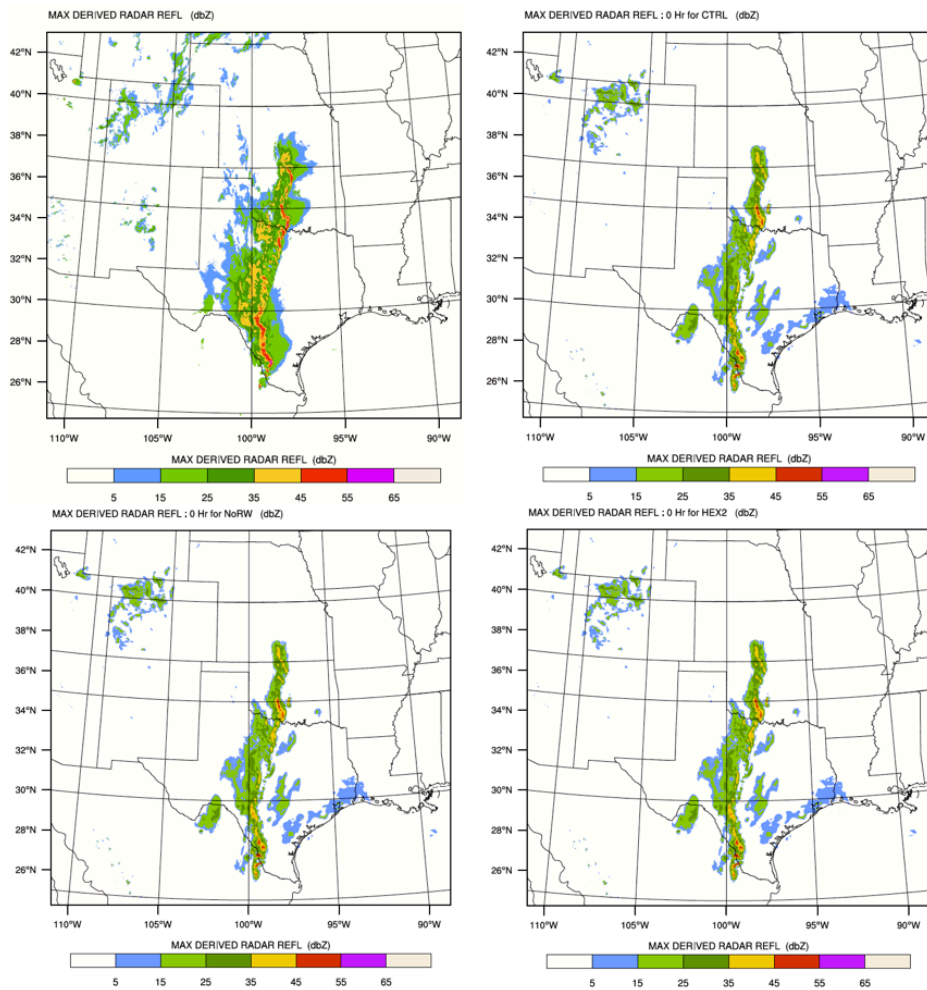
Figure 6.7 0-hour forecast of the composite reflectivity for CTRL (upper right), NoRW (bottom left) and HEX2 (bottom right), compared to the observed (upper left), at 03Z of February 20, 2017.
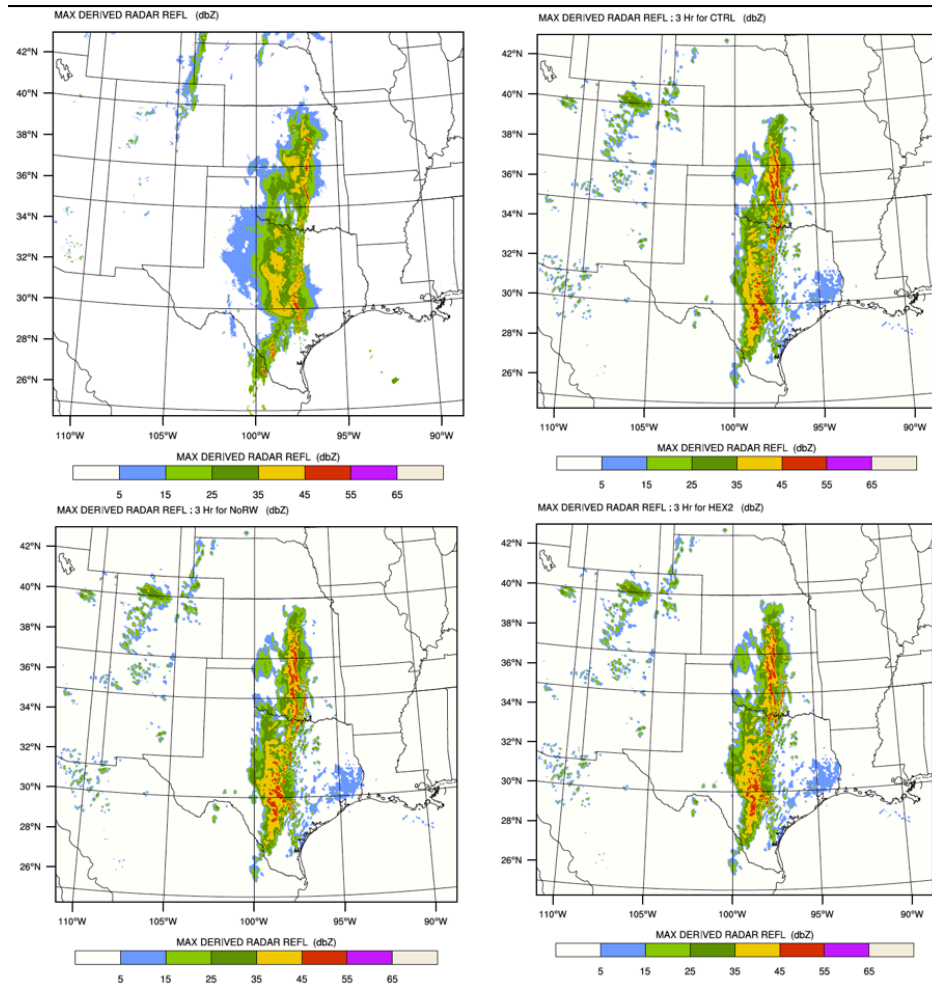
Figure 6.8 Similar to Fig. 69, but for 3-hour forecasts, valid at 06Z of February 20, 2017.

Next, we compare the experiment RWER with the control run, to see the impact of reducing radar observation errors, which supposedly would make more use of the observations. Table 6.2 shows the contribution of radial wind to the cost function for RWER and CTRL. The reduced observation error should increase weight in the cost function, suggesting the GSI code is working as expected with regard to the reduced radar observations errors. Figure 6.9 shows the u-wind (top panels) and v-wind (bottom panels) increments from CTRL (left) and RWER (right) at 03Z of February 20, 2017. The two experiments give generally very similar increments, with RWER adding a bit more details to the structure. However, these differences are too small to produce any notable difference in the storm forecast, as can be seen in the hourly precipitation forecast in Figure 6.10.

Table 6.2 Contribution of radial wind to the cost function

	Before and after 1st outer loop (2017022001)	Before and after 1st outer loop (2017022003)
CTRL	5.19E03 / 1.62E03	1.02E04 / 2.87E03
RWER	2.04E04 / 5.25E03	3.88E04 / 9.76E03
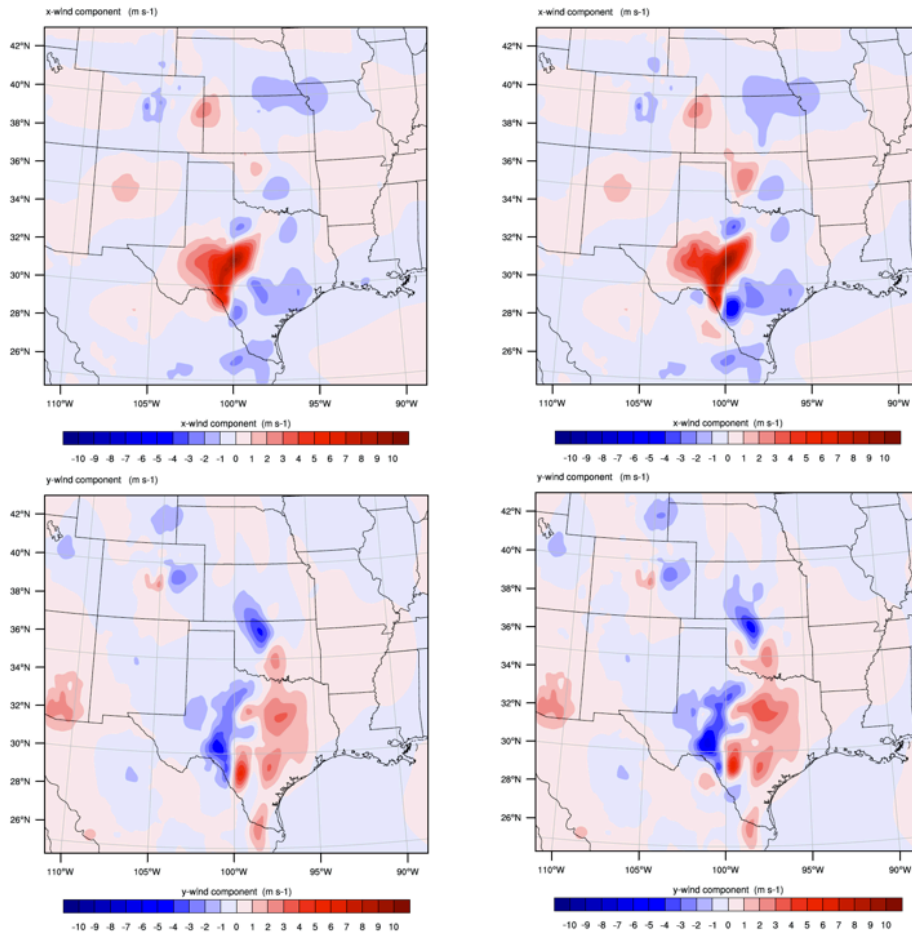
Figure 6.9 Wind increments from CTRL (left) and RWER (right) at 03Z of February 20, 2017. The top panels are u-wind and the bottom panels are v-wind.
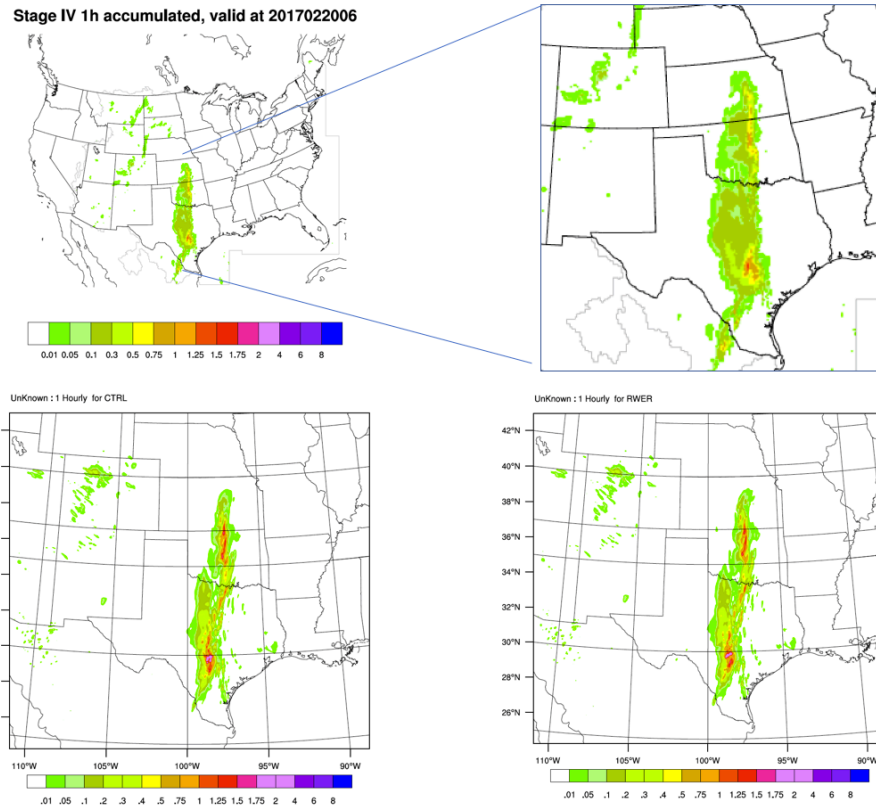
Figure 6.10 Hourly precipitation from the 3-hour model forecast initialized at 03Z of February 20, 2017. The top panels are Stage IV observations while the bottom left (right) is from CTRL (RWER) experiment.

Verification of the 1-day model forecast (starting at 01Z and ending at 23Z of February 20 2017, 23 runs) against Stage IV hourly precipitation was conducted using the FSS score, as seen in Figure 73. The four experiments show very similar scores, with RWER giving slightly better performance for some lead times for heavy rainfall (5mm). This might suggest potential benefit from further tuning the observation errors.
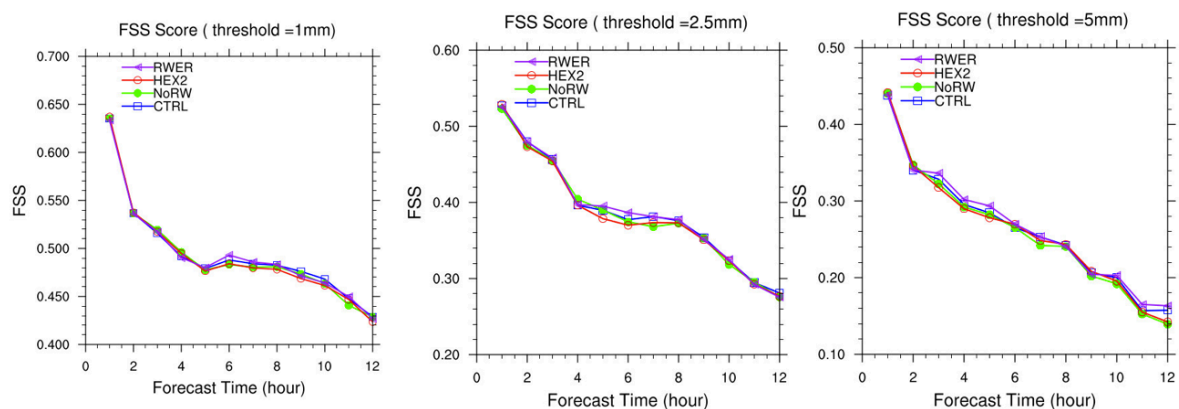
Figure 6.11 FSS score for the hourly precipitation forecast, verified against the Stage IV observations, for threshold 1mm (left), 2.5mm (middle) and 5mm (right).

7. Summary

In AOP 2017, we installed the operational RAPv4/HRRRv3 system on the NOAA supercomputer Theia. We also ported HRRRv3 on the UCAR supercomputer Cheyenne to make it available for community users to test their research ideas based on the operational HRRR system.

Before running the experiments, we made substantial efforts to prepare observations, fill up missing radar data and get boundary and initial conditions ready. We modified the operational RAPv4/HRRRv3 workflows to make them more suitable for research use. This will benefit the broad community, making users easier to setup and run the HRRR system for their research studies.

Radial wind data in NCEP BUFR files was carefully examined and compared to raw binary NEXRAD level II data. It was found that due to strict quality control procedure, lot of good radial wind observations were rejected in the NCEP BUFR files. This issue has been commuted with NCEP related parties. Currently HRRR GSI uses one-hour time window to average radial wind observations to generate radar superob's and then assimilates these superob's into the model. Hence, in spite of the data missing problem, we still got fair amount of radial wind observations to continue our experiments.

The background error covariance impact scale and the GSI superob algorithm was also examined in detail. It was found that the default impact scale in current HRRR is too large for convective scale radial wind data assimilation. A new set of impact scale parameters corresponding to a smaller impact scale was found and proved to be a good choice for future convective-scale HRRR applications. A new set of superob parameters was also found and proved to keep more small-scale observation details although current default superob parameter setting in GSI does not deviate much from convective-scale requirements.

Six one-week long retrospective experiments were conducted to investigate the impact of the radial velocity used in high resolution analysis grid with small superob cells. It was found that although we added lots of convective-scale radial wind observations into the model, the synoptic-scale environment forecasts did not degrade. When checking the individual storm forecasts, it was found that as compared to the experiment without radial wind observations, current HRRR degrades the individual storm forecasts a little bit due to its using too large BE impact scales for high resolution radial wind data. When we adopted a smaller BE impact scales and assimilated radial wind data at the 3 km analysis grid, the individual storm forecasts were improved as compared to the experiment without radial winds. The benefit of hybrid 3DEnVar using GDAS ensemble covariance was investigated and it was demonstrated that hybrid 3DEnVar can improve individual storm forecasts to some extent. Hence it is recommended to use hybrid 3DEnVAR as much as possible if computing resources is available. Using smaller superob cells does not bring evident benefits under our hybrid 3DEnVar settings although it improves individual storm forecasts under the pure 3DVar environment.

Another set of four week-long experiments were conducted to study the possible benefit of analyzing radial velocity over coarse analysis grid with large superob cells. This set of model

forecasts were verified using MET. The vast preponderance of verification results showed that the three experiments had neutral impact on the forecast performance. For the moisture variables, there was a slight hint of statistically significant improvement at several intermediate to later forecast lead times when the surface verification results were aggregated together. In the vertical, no clear advantages were seen for any experiment. This picture did not change notably when looking across individual forecasts (across all forecast initialization times). While certain periods had some dropouts of degraded performance, there were no clear picture that any experiment performed worse than any other. Results were largely consistent with the noise one would expect from comparing individual model forecasts.

A new one-day experiment for winter in February 2017 suggests that with good amount of radar data volume, tuning the radar observation errors might bring some potential benefit to the storm forecast. More work needs to be done in this area.

Research findings from the above study were shared through several conference presentations

- Guoqing Ge, and M. Hu, J. Beck, C. Zhou, H. Shao, S. Weygandt, S. Benjamin, and C. Alexander, 2018: Assimilating Radial Winds into the High-Resolution Rapid Refresh (HRRR) Model and Its Impact on Storm Forecasts. AMS 2018, Austin, TX in January 2018. 12.4
- Chunhua Zhou, and M. Hu, Y. Zhang, and G. Ge, 2018: Testing and Evaluation of the Radar Data Assimilation and Its Impact on Convective Forecast for the 3-Km High Resolution Regional Applications. 29th Conference on Weather Analysis and Forecasting/25th Conference on Numerical Weather Prediction, Denver, CO 06/03-06/08, 2018, 56
- Zhou, Chunhua, and Ying Zhang, Guoqing Ge, Ming Hu, Jonathan Vigh, 2018: Testing and evaluation of the radar data assimilation for high resolution convective forecasts. WRF Workshop, 06/11-06/15, 2018, Boulder
- Ming Hu, and G. Ge, C. Zhou, C. R. Alexander, and S. Weygand, 2018: Building Functional Similar Testing System for Community Researchers to Contribute Operational Storm-Scale Data Assimilation. 29th Conference on Weather Analysis and Forecasting/25th Conference on Numerical Weather Prediction, Denver, CO 06/03-06/08, 2018, 7B4
- Guoqing Ge, and M. Hu, C. Zhou, J. Beck, S. Weygandt, and C. Alexander, 2018. Further Improvements of Radial Wind Assimilation in the High-Resolution Rapid Refresh (HRRR) model. 29th Conference on Weather Analysis and Forecasting/25th Conference on Numerical Weather Prediction, Denver, CO 06/03-06/08, 2018, 7B5

Appendix A: Apply MET for HRRR functionally similar testing system

Verification is part of the HRRR functionally similar testing system. Because the NOAA HRRR test system use MATS, which is only available for NOAA machine, we employed MET as verification system for community HRRRR functionally similar testing system.

The following sections describe the attributes of the observations used in the verification, the verification protocol, the results of the T&E activity.

A1. Data

The HRRR experiment simulations were verified against the observations contained in PREPBUFR files from the RAP modeling system. The observations in these PREPBUFR files was also used as the basis for the routine observations (apart from the radar wind data) that were assimilated by the HRRR experiments that were conducted. To better understand the characteristics of the observational data, this section analyzes the availability of observations during the 24-h period beginning at 00Z on Sep 3.

Observations were reported for nine different message types in the PREPBUFR files during the 24-h period. Table A1 lists the mnemonic short-hand identifiers and a description of the types of data included. Because HRRR is initialized on an hourly cycle, it was prudent to select a time window centered on the analysis/verification time whose width was less than or equal to the time between initialization times (selecting a longer time windows longer would result in observations being used more than once). Thus, the observations availability was analyzed for two time windows: +/- 15 minutes from the analysis time, and +/- 30 minutes from the analysis time. Fig. A1.1 shows examples of plots that were created at 3-hourly intervals to display the spatial and temporal distribution of observations near each analysis time, with Fig. A1.1a showing the locations of all observations reported within +/- 15 min of the target time (00Z on 03 Sep 2016), and Fig A1.1b showing the observations within +/- 30 min of the target time. It should be noted that there is only one wind profiler site still in operation (near Houston, TX). The most noticeable difference between the two time windows is that a +/- 30 min time window captures significantly more aircraft observations.

The number of observations was then tabulated every 3 h and analyzed by message type, by time of day, and by the width of the observational time window. Table A2 summarizes this information through the 24-h period. In terms of the contribution to the total number of observations, MSONET contributed the most observations by far, followed by the VADWND, and AIRCAR. The ADPSFC and ADPUPA message types also contributed significant numbers of observations. In comparison, there were fewer observations for the SFCSHP, PROFLR, and RASSDA message types. Looking at variations over the 24-h period, the 12Z period had the most observations, followed closely by the 00Z period. The 18Z period had the least number of observations.

Mnemonic	Description
ADPUPA	**UPPER-AIR (RAOB, PIBAL, RECCO, DROPS) REPORTS**
AIRCAR	**MDCRS ACARS AIRCRAFT REPORTS**
AIRCFT	**AIREP, PIREP, AMDAR, TAMDAR AIRCRAFT REPORTS**
SATWND	SATELLITE-DERIVED WIND REPORTS
PROFLR	**WIND PROFILER AND ACOUSTIC SOUNDER (SODAR) REPORTS**
VADWND	**VAD (NEXRAD) WIND REPORTS**
SATEMP	POES SOUNDING, RETRIEVAL, RADIANCE DATA (TOVS)
ADPSFC	**SURFACE LAND (SYNOPTIC, METAR) REPORTS**
SFCSHP	**SURFACE MARINE (SHIP, BUOY, C-MAN/TIGE GAUGE PLATFORM) REPORTS**
SFCBOG	MEAN SEA-LEVEL PRESSURE BOGUS REPORTS
SPSSMI	DMSP SSM/I RETRIEVAL PRODUCTS (REPROCESSED WIND SPEED, TPW)
SYNDAT	SYNTHETIC TROPICAL CYCLONE BOGUS REPORTS
ERS1DA	ERS SCATTEROMETER WIND DATA (REPROCESSED WIND SPEED)
GOESND	GOES SOUNDING, RETRIEVAL, RADIANCE DATA
QKSWND	QUIKSCAT SCATTEROMETER WIND DATA (REPROCESSED)
MSONET	**MESONET SURFACE REPORTS**
GPSIPW	GLOBAL POSITIONING SATELLITE-INTEGRATED PRECIPITABLE WATER AND TOTAL ZENITH DELAY REPORTS
RASSDA	**RADIO ACOUSTIC SOUNDING SYSTEM (RASS) VIRTUAL TEMPERATURE PROFILE REPORTS**
WDSATR	WINDSAT SCATTEROMETER WIND DATA (REPROCESSED)
ASCATW	ASCAT SCATTEROMETER DATA (REPROCESSED)

Table A1: Current Table A Entries in NCEP's PREPBUFR mnemonic table. Bolded entries denote message types for which data were found in the observations analysis period on 03 Sep 2016. [This table was excerpted from NCEP's Table 1.a, which can be found at:
http://www.emc.ncep.noaa.gov/mmb/data_processing/prepbufr.doc/table_1.htm.]

Observations Availability Summary

	00Z		03Z		06Z		09Z		12Z		15Z		18Z		21Z	
	15 min	30 min	15 min	30 min	15 min	30 min	15 min	30 min	15 min	30 min	15 min	30 min	15 min	30 min	15 min	30 min
All Message Types	373,777	712,626	328,068	523,866	285,536	465,934	285,847	460,425	365,973	689,304	285,624	457,628	268,187	435,759	307,854	499,013
ADPSFC	16,832	23,144	16,431	22,329	16,194	21,955	16,411	22,270	16,215	22,633	16,504	22,347	16,838	22,287	16,424	22,206
ADPUPA	13,401	13,401	0	0	0	0	0	0	16,115	16,115	0	0	0	0	0	0
AIRCAR	31,060	66,667	31,662	48,142	3,590	5,521	10,839	15,946	21,166	44,060	36,326	59,158	29,844	47,788	29,229	46,765
AIRCFT	10	23	6	12	18	36	9	18	24	39	12	18	10	22	17	21
ASCATW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GOESND	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MSONET	252,166	485,330	200,991	317,417	199,019	314,324	197,859	312,613	247,023	473,780	201,466	317,154	191,026	310,277	203,960	323,143
PROFLR	234	542	236	480	256	468	232	432	138	582	246	498	214	444	212	430
RASSDA	12	39	9	27	12	18	9	21	9	27	24	48	30	57	18	39
SATWND	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SFCSHP	1,106	2,192	490	1,262	491	1,210	477	1,233	1,127	2,280	454	1,211	471	1,228	462	1,229
VADWND	58,956	121,176	78,243	134,155	65,956	122,402	60,011	107,874	64,156	129,746	30,592	57,146	30,224	53,620	57,532	105,108

Table A2: Summary of the number of observations that were reported in the PREPBUFR files for each of the message types at 3-h intervals. Green-shaded columns show values based on a +/- 15-min time window, and tan-shaded columns show values based on a +/- 30-min time window.

All message types: 373777 observations

All message types: 712626 observations
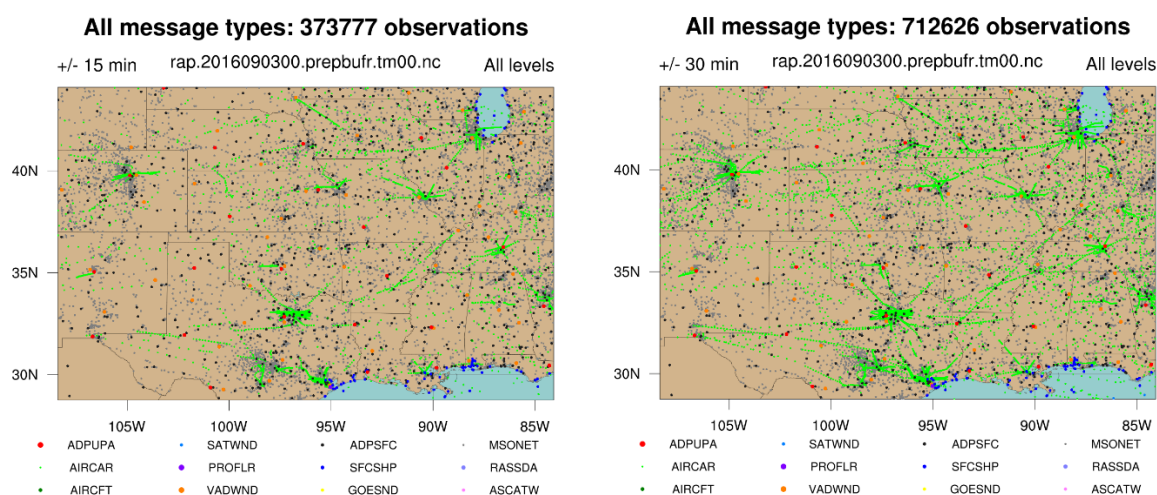
Figure A1.1: Locations of all observations available in PREPBUFR within the given time windows: a) (left panel) all observations within +/- 15 min of the target time; b) (right panel) all observations within +/- 30 min of the target time. Each observation is indicated by dots whose colors denote different message types: upper air soundings (APDUPA, red), aircraft observations (AIRCAR, lime green; AIRCFT, dark green), wind profiler and acoustic SODAR (PROFLR, purple), VAD (NEXRAD) wind reports (VADWND, orange), surface observations (ADPSFC, black; MSONET, grey), marine surface observations (SFCSHP, dark blue), and radio acoustic sounding system virtual temperature reports (RASSDA).

Some observations types displayed a distinct diurnal cycle in availability (Fig. A1.2). Interestingly, MSONET and SFCSHP had significantly more observations at 12Z and 00Z than at other times. From the full set of spatial and temporal plots (not shown), it is evident that some mesonets are only reporting twice a day. These include certain observing networks that appear to be situated along railroads. Apparently, some co-op observing network observations are also reported in MSONET twice a day (as seen over Colorado; not shown). As expected, the radar-based VADWND observations follow a cycle of availability that appears closely tied to convective and precipitation activity, with a diurnal maximum at 03Z and a diurnal minimum at 15Z and 18Z. The aircraft-related observations (AIRCAR and AIRCFT) show a distinct diurnal variation in observational availability that is linked with the airline schedules. From a distinct minimum during the overnight hours (06Z), the number of aircraft observations increases near daybreak and reaches a daily maximum at 15Z when the most planes are in the air. The number of aircraft observations holds relatively steady until 03Z when nearly all domestic flights cease. ADPSFC has less relative fluctuations throughout the 24-h period, but there were some diurnal variations, with the most observations occurring at 00Z and 18Z, and the fewest observations occurring at 0Z and 12Z. Finally, upper air rawinsonde observations (ADPUPA) normally only report at 12Z and 00Z; there were slightly more observations at 12Z than 00Z.
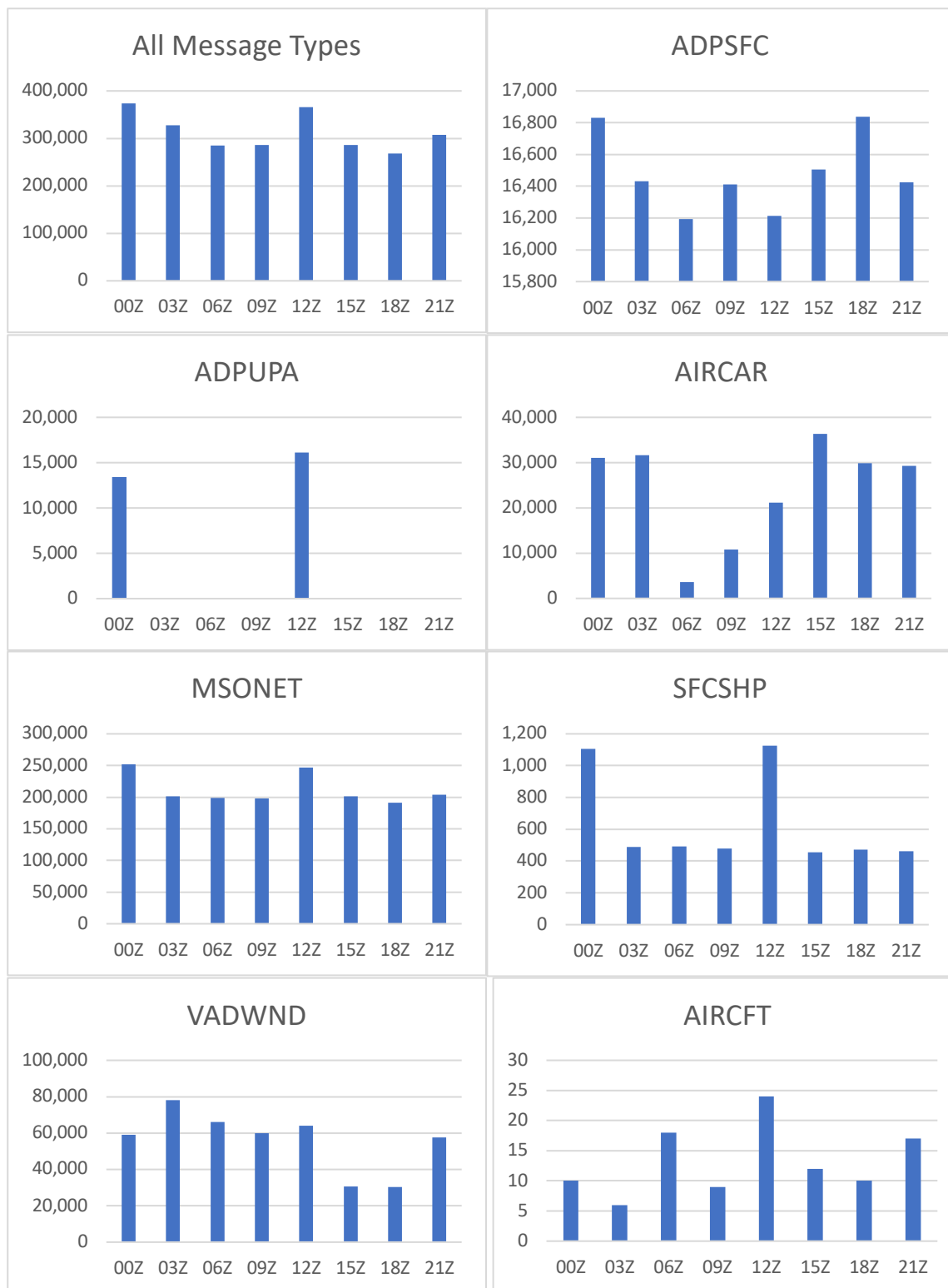
Figure A1.2: Diurnal variations in observational availability for the various message types.

Percent Increase in Observations for the 30 minute Time Window Compared With the 15 minute Time Window								
	00Z	03Z	06Z	09Z	12Z	15Z	18Z	21Z
All Message Types	+90.7%	+59.7%	+63.2%	+61.1%	+88.3%	+60.2%	+62.5%	+62.1%
ADPSFC	+37.5%	+35.9%	+35.6%	+35.7%	+39.6%	+35.4%	+32.4%	+35.2%
ADPUPA	-	-	-	-	-	-	-	-
AIRCAR	+114.6%	+52.0%	+53.8%	+47.1%	+108.2%	+62.9%	+60.1%	+60.0%
AIRCFT	+130.0%	+100.0%	+100.0%	+100.0%	+62.5%	+50.0%	+120.0%	+23.5%
ASCATW	-	-	-	-	-	-	-	-
GOESND	-	-	-	-	-	-	-	-
MSONET	+92.5%	+57.9%	+57.9%	+58.0%	+91.8%	+57.4%	+62.4%	+58.4%
PROFLR	+131.6%	+103.4%	+82.8%	+86.2%	+321.7%	+102.4%	+107.5%	+102.8%
RASSDA	+225.0%	+200.0%	+50.0%	+133.3%	+200.0%	+100.0%	+90.0%	+116.7%
SATWND	-	-	-	-	-	-	-	-
SFCSHP	+98.2%	+157.6%	+146.4%	+158.5%	+102.3%	+166.7%	+160.7%	+166.0%
VADWND	+105.5%	+71.5%	+85.6%	+79.8%	+102.2%	+86.8%	+77.4%	+82.7%

Table 3: Effect of increasing the observation window time width, by time and message type. The percent increase of the number of observations for a 30-min time window is shown, compared to a 15-min time window. Warm colors denote relatively small percentage increases (25-75%, representing less than a doubling of the number of observations); yellow shades denote increases near 100% (a doubling in the number of observations), and green shades denote significantly larger increases (>200%, or a tripling or more in the number of observations).

The impact of increasing the observation time window width was also analyzed (Table A3). Interestingly, doubling the window time did not typically double the number of observations. This is mainly a factor of the typical reporting interval of each message type. For many message types, the percentage increases for the 30-min time window relative to the 15-min time window were less than a double (e.g., +32-39% for ADPSFC, +52-90% for MSONET). For other message types, larger percentage increases were seen (+47-114% for AIRCAR, +71-102% for VADWND, +98-167% for SFCSHP, and +82 to +322% for PROFLR). In summary, this section investigated the availability of observations in the RUC PREPBUFR files, examining the relative contributions by each message type, the diurnal variations of each message type, and the impact of increasing the observational time window. In brief, MSONET, VADWND, and AIRCAR contribute the most number of observations by far, with significant contributions from ADPSFC and ADPUPA. Several message types display a distinct diurnal cycle in the availability of observations (MSONET, SFCSCHP, AIRCAR, and VADWND). Increasing the time window did not result in a uniform doubling of observational availability, but rather a spectrum of increases ranging from ADPSFC on the low side, to PROFLR, RASSDA, and SFCSHP on the high side.

A2. Verification protocol

MET is a widely used verification system developed through a long collaboration by the Development Testbed Center and NCAR's Joint Numerical Testbed. This verification study used MET v6.1, which although not quite as new as the very recent v7.0 release, is likely more stable with less bugs. Two MET utilities were used for this verification exercise: *pb2nc* and *point_stat*. Both of these were run on NCAR's Cheyenne supercomputer using a simple job submission script that was developed to facilitate rapid processing of the verification jobs. The statistics output was then uploaded into a mySQL database on a local NCAR server and where plots were generated using METViewer's batch processing engine. This section will describe the MET configurations used for these tools and provide links to the specific configuration files and codes that were used.

First, it should be noted that all codes and configuration files used in this verification activity have been placed under version control and placed in NCAR's institutional GitHub repository at https://github.com/NCAR/ral-jvigh-verification. By providing the full code and configuration files used, the full details of the verification protocol are preserved for future reference. This should facilitate more rapid verification studies for future T&E activities of the DTC DA Task. Public access is currently turned off. To request access to this repository, please e-mail Jonathan Vigh (jvigh@ucar.edu). If it would be useful, the repository can be made public in the future.

pb2nc configuration

The *pb2nc* utility converts the observations from PREPBUFR to NetCDF. To use this utility, any existing restrictions were removed for filtering based on message types, station IDs, observational mask regions. The number of vertical levels to retain was updated from 255 to 511. The configuration entry for the BUFR variables to be retained 'obs_bufr_var' was updated and the mapping to MET's variables ('obs_prepbufr_map') was correspondingly updated. Table A4 provides the correspondence between the PREPBUFR variable names and the MET variable names.

PREPBUFR variable name	MET variable name	Description
POB	PRES	Pressure
QOB	SPFH	Specific humidity
TOB	TMP	Temperature
ZOB	HGT	Geopotential height
UOB	UGRD	U-component wind speed
VOB	VGRD	V-component wind speed
D_DPT	DPT	Dewpoint temperature
D_WDIR	WDIR	Wind direction
D_WIND	WIND	Wind speed
D_RH	RH	Relative humidity
D_MIXR	MIXR	Mixing ratio
D_PRMSL	PRMSL	Mean sea level pressure

Table A4: Correspondence between the variable names in PREPBUFR and MET, along with a description. PREPBUFR variable names prefixed with 'D_' indicated that these quantities are derived from other variables.

This verification study focuses just on PRES, SPFH, TMP, UGRD, VGRD, WIND, and RH. Importantly, the quality mark threshold ('quality_mark_thres') was set to a value of 3, which uses PREPBUFR's quality flag to determine which observations to include in the verification. This means that MET will include all observations with quality markers of 0 through 3, which are as follows:

- 0: 'keep' observations (always assimilated)
- 1: 'good' observations
- 2: 'neutral' observations or not checked (default)
- 3: 'suspect' observations.

Typically, all observations with quality markers of 0 through 3 are assimilated, and all observations with higher quality marker values are rejected. Thus, our protocol is following

the standard verification practice of not including any observations which would normally be rejected for purposes of data assimilation.

The full configuration file for pb2nc can be viewed by those with repository access at: https://github.com/NCAR/ral-jvigh-verification/blob/master/MET_config/PB2NCConfig_HRRR.DA_task

point_stat configuration

Once the observations have been converted to NetCDF by *pb2nc*, the next task is to generate matched pairs using MET's *point_stat* utility. This step involves taking all available observations and interpolating from the simulation to the location and pressure of the observation to generate matched pairs. Then partial sums (SL12L1) lines are aggregated for all of these matched pairs according to the specifications of the user for the specified initialization times, forecast lead times, or vertical layers. Results can have aggregated for many types of observations or limited to just one observation type (e.g., surface observations, upper air soundings, aircraft data, etc.).

MET offers many options for how these horizontal and vertical interpolations are accomplished. This verification used bilinear interpolation in the horizontal direction (width=2 and valid_ threshold=0.5).

A rather critical option is the choice of time threshold that will be used to set the window of allowable observations, relative to the forecast valid time of the model output being verified. Based on the previous analysis of observational availability, it was deemed that sufficient observations were available to use a time window of +/- 900 s (+/- 15 min). This also happens to be MET's default.

For wind speed, count statistics are not computed for winds of less than 2.572 m s^{-1}; this is in accordance with long-standing practice since wind speeds below this threshold tend to have variable directions.

No censor threshold or censor values were used. The continuous statistics logic and wind logic were set to use 'UNION'.

MET allows users to specify the vertical layers or levels over which partial sums will be aggregated. This plays an important role in the design of the verification study, and the thickness of layers chosen must be optimized to balance the following factors:

- observational availability (a meaningful number of observations must be present within the layer over the timescales of interest; using levels too close together may result in adequate data)
- meaningful statistical significance (using layers that are too close together may lead to confidence intervals that are over-confident),
- accuracy (using levels too far apart will result in interpolation errors which will cloud the results), and
- vertical resolution (layers that are too far apart will not provide detailed enough picture of the vertical structure of errors).

Most of the DTC's past verification studies have used the upper air soundings and examined just the mandatory levels (e.g., 1000, 925, 850, 700, 500, 400, 300, 250, 200, 150, 100, 50 hPa). For a high-resolution model like the HRRR, it was decided to look in more detail at the vertical structure and to also examine the characteristics of other type of observational data that exist throughout the troposphere (e.g., aircraft observations, etc.). Thus, the *point_stat* configuration file was crafted to examine overlapping 50-hPa thick layers. The layer-centered pressures ranged from 975 hPa (near the surface) up to 75 hPa in the upper atmosphere. Specifically, the layers were interspersed at 25-hPa intervals: 1013.2-975, 1000-950, 975-925, 950-900, 925-875, 900-850, 875-825, 850-800, 825-775, 800-750, 775-725, 750-700, 725-675, 700-650, 675-625, 650-600, 625-575, 600-550, 575-525, 550-500, 525-475, 500-450, 475-425, 450-400, 425-375, 400-350, 375-325, 350-300, 325-275, 300-250, 275-225, 250-200, 225-175, 200-150, 175-125, 150-100, 125-75, 100-50 hPa. Given that the post-processed model output is available at 25-hPa intervals, this means that no observation was more than 12.5 hPa away from a model output level. While there will undoubtedly be some vertical interpolation errors, they should be relatively small. In this study, upper atmosphere verification was conducted for two sets of observational types: a) just upper air soundings (ADPUPA), and b) all upper observations types including upper air soundings (ADPUPA), both types of aircraft observations (AIRCRFT and VADWND), and radar-derived wind analyses (VADWND), and acoustic profilers (PROFLR). While verification statistics have been computed for both sets of upper air data, the inter-comparison of the different upper air observation types (to be shown later) suggests that only the upper air soundings and aircraft observations have error characteristics that are reasonable for use in this verification study.

Surface verification uses the following observation types: METAR and SYNOP reports (ADPSFC), mesonet observations (MSONET), and surface ship reports (SFCSHP). Temperature and moisture variables (TMP/DPT/RH/SPFH) are verified at the standard 2-m observing height above ground (which corresponds to the Z2 model level). Wind data (UGRD/VGRID/WIND) are verified at the standard 10-m observing height above ground (which corresponds to the Z10 model level).

Output was provided for the following line types: contingency table counts and statistics (FHO, CTC, CTS), multiple categorical thresholds (MCTC, MCTS), continuous statistics (CNT), partial sums (SL1L2, SAL1L2, VL1L2, and VAL1L2). Initially, the full matched pair lines (MPR) were also included, however this lead to a very large volume of data (510 GB per experiment) that took 3.5 hours to load into METViewer's XML Database. When *point_stat* was rerun without matched pair output lines, the total volume of data was much more manageable (3 GB per experiment) and was able to be loaded rapidly into the database.

The full configuration file for *point_stat* can be obtained for approved repository viewers at:
https://github.com/NCAR/ral-jvigh-verification/blob/master/MET_config/PointStatConfig_COMBINED_HRRR.DA_task_both_interpolation_and_exact_pressure_match (r17)

Plot Generation

METViewer's batch plotting engine was used to generate the various plots of verification results. This utility allows a user to define an XML file containing options which define the set of plots that will be generated. The batch engine, which is controlled by the mv_batch_script.sh, can loop over various dimensions of the defined workflow, thereby creating many plots from one XML file. For this project, an XML file was created for each major set of plots, resulting in 25 XML files. These files can be found in the following location:

https://github.com/NCAR/ral-jvigh-verification/tree/master/metviewer/xml/xml_generate_plots

Table A5 provides a list of the XML files that were created for this T&E activity, provides the status of the XML file (whether the file is considered 'final' or whether additional improvements could be made), and the number of plots in the set.

XML Template Name	Status	# of plots in set
Vertical Layers Plots		
observations-comparison		
template_da_radar_vert_obs-comparison_CTRL_ADPUPA-AIRCAR-AIRCRFT-PROFLR-VADWND_bootCI_layers	final	36
template_da_radar_vert_obs-comparison_CTRL_ADPUPA-AIRCAR_bootCI_layers	final	84
xml_template in template_da_radar_vert_obs-comparison_CTRL_ADPUPA_bootCI_layers	final	84
xml_template in template_da_radar_vert_obs-comparison_CTRL_AIRCAR_bootCI_layers	final	84
experiment-comparison		
template_da_radar_vert_experiment-comparison_CTRL-NoVAD-RW12km-RW36km_OBS_bootCI_layers	final	168
template_da_radar_vert_experiment-comparison_CTRL-NoVAD_OBS_bootCI_layers	final	168
template_da_radar_vert_experiment-comparison_CTRL-RW12km_OBS_bootCI_layers	final	168

template_da_radar_vert_experiment-comparison_CTRL-RW36km_OBS_bootCI_layers	final	168
template_da_radar_vert_experiment_comparison_CTRL-NoVAD-RW12km-RW36km_OBS_noCI_layers	final	168
template_da_radar_vert_experiment_comparison_CTRL-NoVAD_OBS_noCI_layers	final	168
template_da_radar_vert_experiment_comparison_CTRL-RW12km_OBS_noCI_layers	final	168
template_da_radar_vert_experiment_comparison_CTRL-RW36km_OBS_noCI_layers	final	168
experiment-differences		
template_da_radar_vert_experiment-differences_CTRL-NoVAD-RW12km-RW36km_OBS_bootCI_layers	final	168
template_da_radar_vert_experiment-differences_CTRL-NoVAD_OBS_bootCI_layers	final	168
template_da_radar_vert_experiment-differences_CTRL-RW12km_OBS_bootCI_layers	final	168
template_da_radar_vert_experiment-differences_CTRL-RW36km_OBS_bootCI_layers	final	168

Times Series Plots

experiment-comparison-by-lead-time

(an XML file for this plot type was not prepared)		

experiment-comparison-by-initialization-time

template_da_radar_time-series_experiment-comparison-by-init-time_CTRL-NoVAD-RW12km-RW36km_ADPSFC_noCI	could be rerun with fixed ranges	

experiment-differences-by-lead-time

template_da_radar_time-series_experiment-differences-by-lead-time_CTRL-NoVAD-RI12km_RW36km_ADPSFC_bootCI	final	14
template_da_radar_time-series_experiment-differences-by-lead-time_CTRL-NoVAD_ADPSFC_bootCI	final	14
template_da_radar_time-series_experiment-differences-by-lead-time_CTRL-RW12km_ADPSFC_bootCI	final	14
template_da_radar_time-series_experiment-differences-by-lead-time_CTRL-RW36km_ADPSFC_bootCI	final	14
experiment-diff-by-init-time		
template_da_radar_time-series_experiment-diff-by-init-time_CTRL-NoVAD-RW12km-RW36km_ADPSFC_bootCI	final	84
template_da_radar_time-series_experiment-diff-by-init-time_CTRL-NoVAD_ADPSFC_bootCI	final	84
template_da_radar_time-series_experiment-diff-by-init-time_CTRL-RW12km_ADPSFC_bootCI	final	84
template_da_radar_time-series_experiment-diff-by-init-time_CTRL-RW36km_ADPSFC_bootCI	final	84

Table A5: List of the XML files that were created for this T&E activity. First column: filename of the XML file; second column: status of the XML file (whether the file is considered 'final' or whether additional improvements could be made); third column: number of plots in each set.

A3. Verification Results

The major goal of this T&E verification activity is to compare the model performance for three experiments (NoVAD, RW36km, and RW12km) to see which, if any, method of treating the radar wind assimilation resulted in positive impact to the forecast accuracy. As detailed in Table 6, a number of suites of plots were generated. This section will summarize the findings from these. The vertical observations comparison plots are examined first, in which the impact of the observations type on the verification statistics are examined for the various levels to learn which observation types may be warranted for use in the verification, and whether there are any important differences between vertical regions of the atmosphere. Following this subsection, the vertical experiment comparisons are presented, in which the performance of the three experimental configurations are compared relative to the control run for the various vertical levels. Following that, vertical experimental differences are presented to determine if the close-up examination of differences relative to control tells any story. After that, the experiment performance will be examined across forecast lead times and initialization times to learn whether the performance of the experiments varied significantly across different forecast periods or initialization cycles.

Vertical Observations Comparisons

Before examining the relative performance of the three experiments, however, the vertical structure characteristics of the various upper air observations was compared. This was accomplished by plotting the verification statistics for the CTRL run for each observation type. One illustrative example is shown in Fig. A1.3 and A1.4. These plots show the Wind Bias for the 12-h forecast lead time for the ADPUPA, AIRCAR, AIRCFT, PROFLR, and VADWND observation types. One can immediately see that CTRL verified against ADPUPA and AIRCAR have a relatively small bias, while the other types have much larger biases. Notably, AIRCFT only has data in the layers between 350 and 150 hPa. The biases and RMS errors were significantly larger than for AIRCAR at these levels. Another striking feature is the very large sinuous bias structure in the vertical for the VADWND observation type. Similarly, RMS errors for the VADWND observations are much larger than any of the other observation types. Verification statistics computed using PROFLR observations are only available in the lower portion of the troposphere, however the verification statistics for this observation type match well with those computed using the ADPUPA and AIRCAR observations types. It is noted that most wind profiler stations were decommissioned prior to this experiment, so within the verification domain, there was only one profiler in operation (near Houston, TX). Finally, the vertical bias and RMS error structures computed using ADPUPA and AIRCAR observations match quite well, although the biases computed using ADPUPA were somewhat larger than those computed using AIRCAR. The plots also show bootstrap confidence intervals, which provide an estimate on the uncertainty of the verification statistics. In general, having more high-quality observations will result in smaller confidence intervals, while having fewer observations of lower quality results in larger estimates of uncertainty. From these figures, it is easy to see that VADWND, PROFLR, and AIRCFT all have relatively large uncertainties associated with their estimates of the verification statistic, while ADPUPA and AIRCAR have quite small confidence intervals (with AIRCAR being the smallest). This suggests that AIRCAR and ADPUPA are both suitable for use in this verification activity, but the other observation types have significantly large uncertainty that their use is not recommended. The plots for other lead-times and forecast parameters are similar and are not shown. Based on the findings from this set of plots, additional exploration of the impact of observation type was limited to just the AIRCAR and APDUPA observation types.

Observation Comparison: Wind (m_s) Bias for f12 CTRL
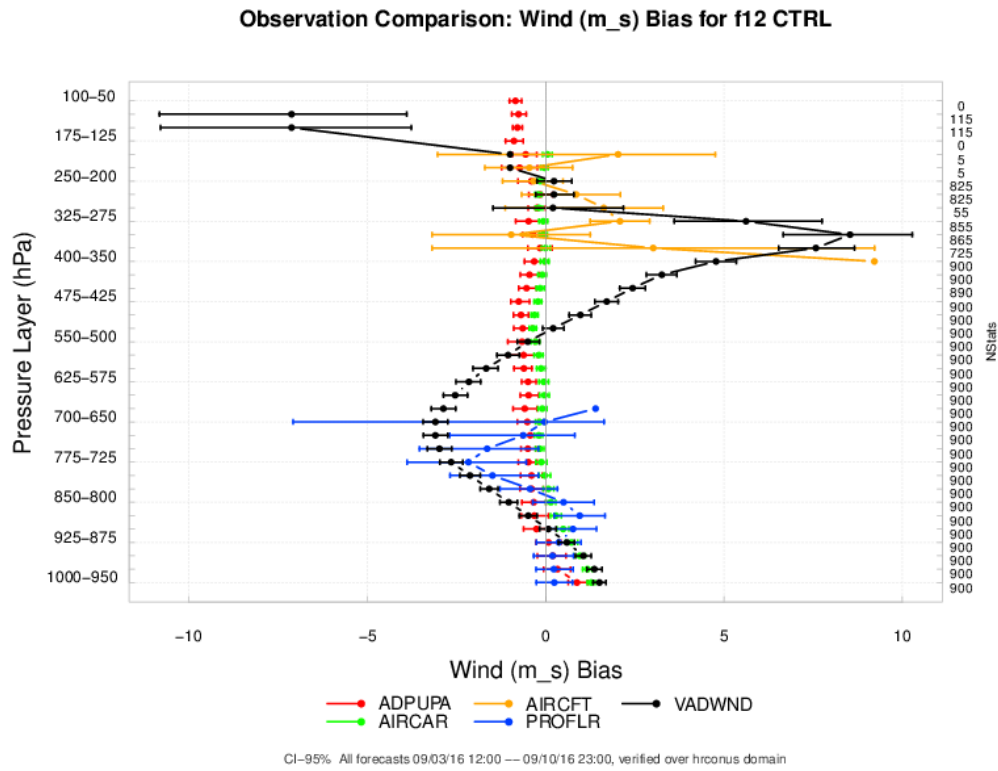
Figure A1.3: Comparison of the vertical structure of the wind speed bias computed for the CTRL experiment using five different observation types: ADPUPA (red), AIRCAR (green), AIRCFT (orange), PROFLR (blue), VADWND (black). The total number of aggregated SL1L2 lines is shown on the right axis. 95% confidence intervals on the estimates are also shown. This plot includes all of the forecasts which had verifying 12-h lead times in the period 12:00 UTC on 09/03/16 12:00 – 23:00 UTC on 09/10/16.

Observation Comparison: Wind (m_s) RMSE for f12 CTRL
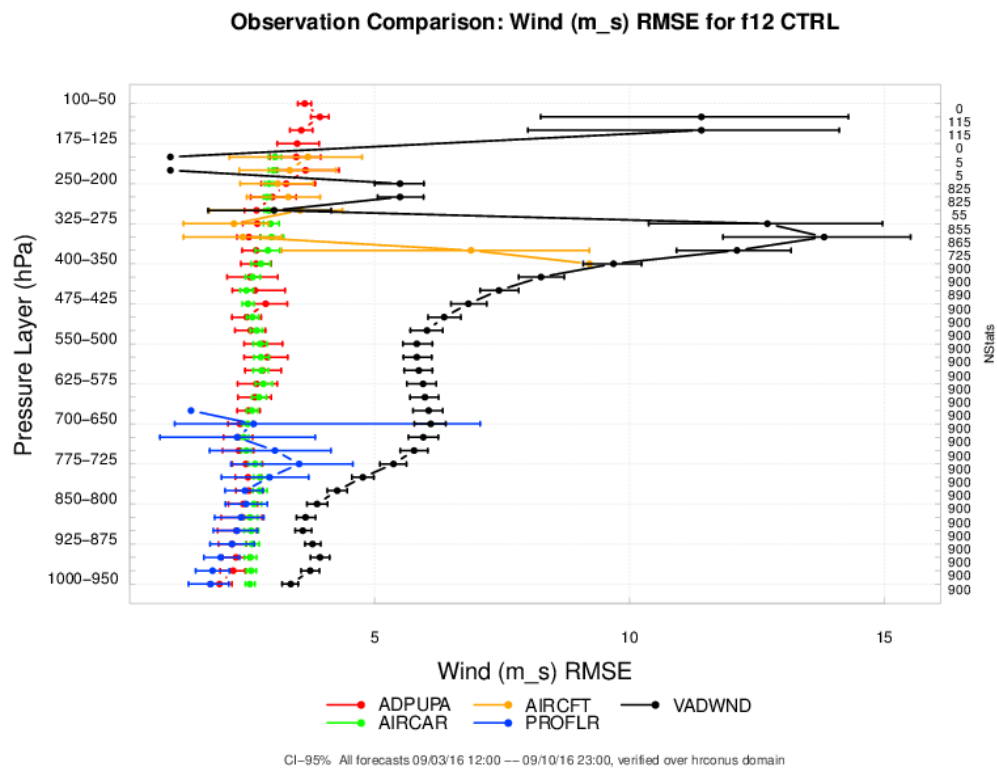
Figure A1.4: Same as Figure A1.3, but for the RMS error of wind speed.

Due to the large number of plots in the ADPUPA vs. AIRCAR comparison, just a few illustrative plots will be shown for a couple variables, with the rest of the results summarized qualitatively. Figure A1.5 shows the RMS error for the 12-h forecast lead time for temperature. For this variable and statistic, the results using either observation type are largely similar, however some differences are evident at upper levels, with the verification based on AIRCAR having somewhat larger errors. Also, of note, the results based on the AIRCAR observations do not extend above about 200 hPa, as commercial aircraft do not normally fly higher than this pressure level. One final feature to note in this plot is that the uncertainty on the RMS error is larger for ADPUPA than for AIRCAR. This is likely owing to the greater number of AIRCAR observations, compared to the relatively fewer ADPUPA observations. Figure 29 shows the temperature bias for the 12-h forecast lead time. The verification statistics using each observation type are very similar at lower and middle levels, with major differences only apparent above about 325 hPa. For dew point temperature, the two observations result in similar verification statistics except near the surface, where the verification statistic computed using ADPUPA has significantly smaller RMS errors than those computed using AIRCAR. At upper levels, the confidence bounds overlap at most levels, indicating that they are not statistically distinguishable. Similar patterns are seen for relative humidity and specific humidity, with statistically significant differences occurring largely at the lower levels. For wind, the two observation types result in largely similar results for the verification statistics at most levels, however for the meridional wind, significant differences are seen in the lower and middle troposphere for the bias statistic. This pattern also shows up in the statistics for total wind speed. All in all, this suite of observational comparisons suggests that the ADPUPA and AIRCAR are generally of high quality and produce verification results that are largely similar to each other. This means that either could be used in the remainder of the verification study. For simplicity, we will focus on the results using ADPUPA.
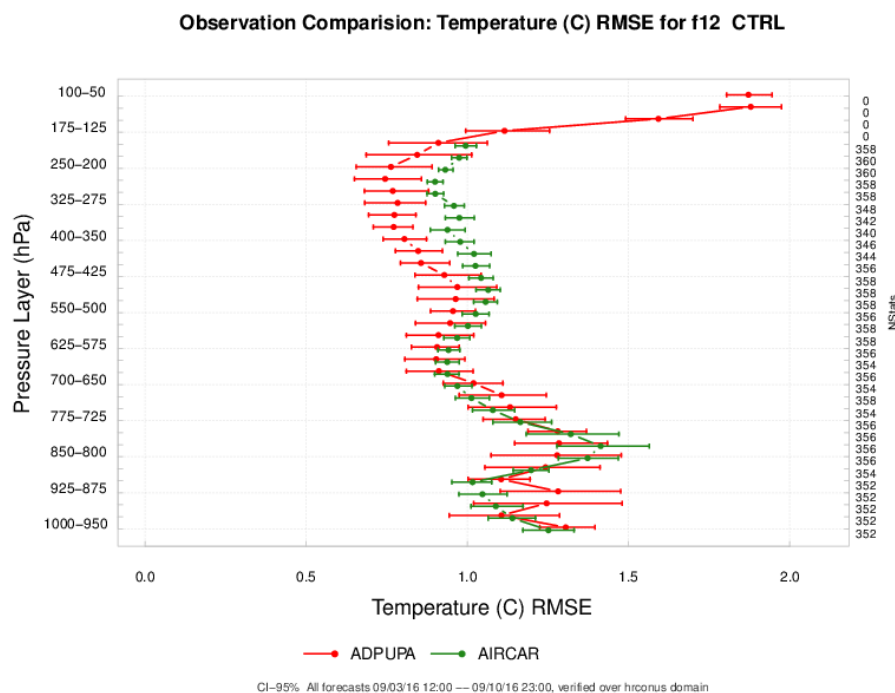
Figure A1.5: Comparison showing RMS error for temperature computed for the 12-h lead time using ADPUPA (red) and AIRCAR observations (green). The number of SL1L2 aggregations at each level is shown on the right axis. Error bars shown the 95% confidence intervals.

Observation Comparision: Temperature (C) Bias for f12 CTRL
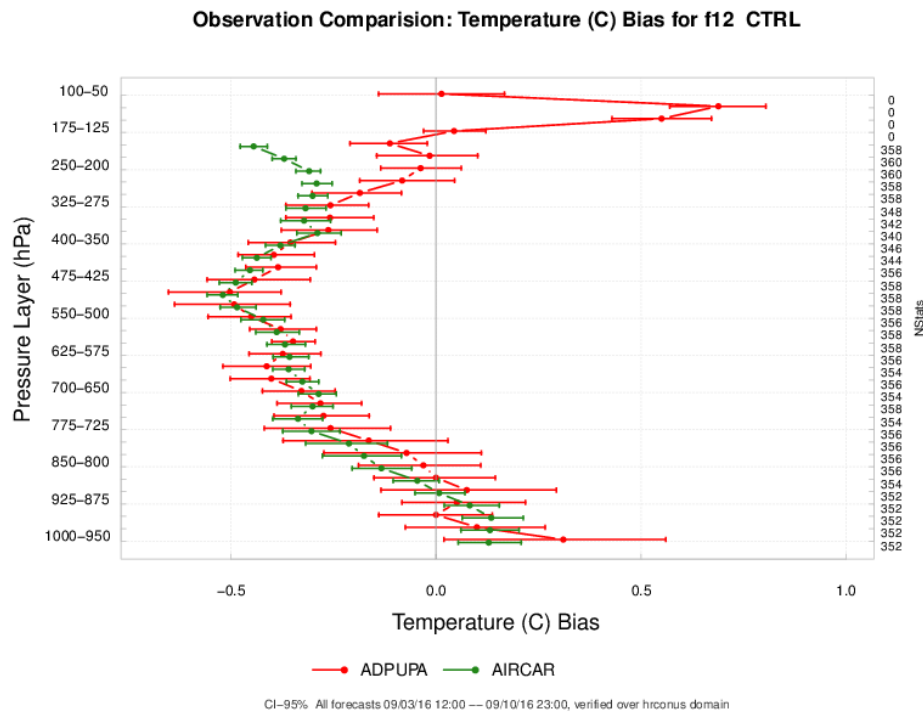
Figure A1.6: As in Fig. 1.5, but for temperature bias.

A4. Discussion of differences between results from MET and MATS

While MET was used as the primary verification system for this T&E activity, the verification results for the CTRL run were compared to those from GSD's MATS verification system. Some noticeable differences occurred, primarily for the RMSE metric.

To undertake this comparison, the output statistics of both verifications systems were captured as text files, then reformatted and aggregated in a spreadsheet so that they could be turned into clean ASCII text files that could be read by an NCAR Command Language (NCL) script. While each verification system utilized a different method and definitions for the vertical layers, this allowed the results to be plotted on the same plot and compared directly. Figure 1.7 shows the comparison for relative humidity bias for four different lead times. Note that MET did not computed verification statistics for the layers above 300 hPa due to the expected poor accuracy of relative humidity data in the upper atmosphere. These plots show that the relative humidity bias is very similar between two verification systems. Figure A1.8 shows the comparison for relative humidity RMS errors.
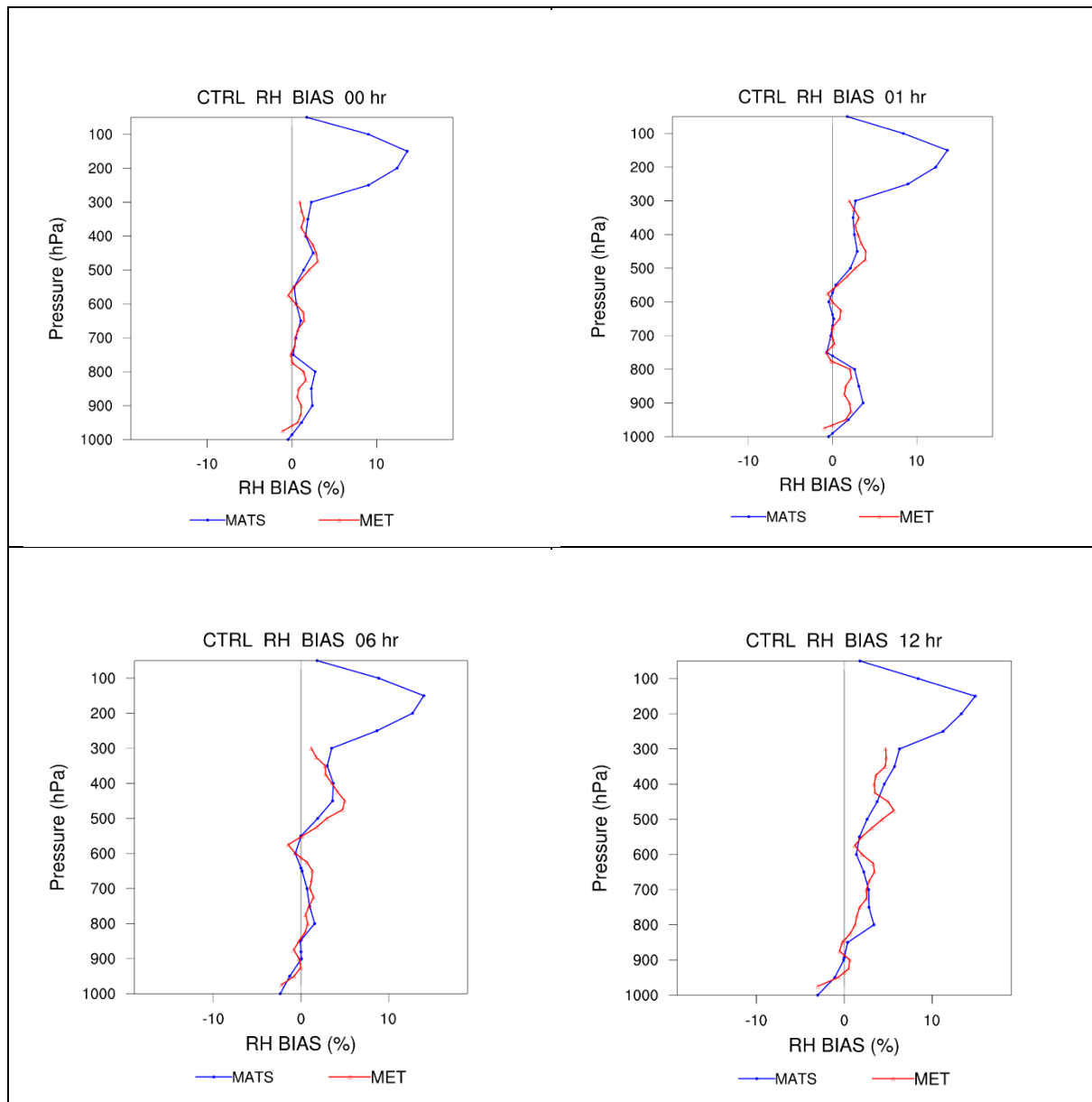
Figure A1.7: Comparison of MATS (blue) and MET (red) verification statistics for the various vertical layers for relative humidity bias at: 00h (upper left), 01 h (upper right), 06 h (lower left), and 12 h (lower right).
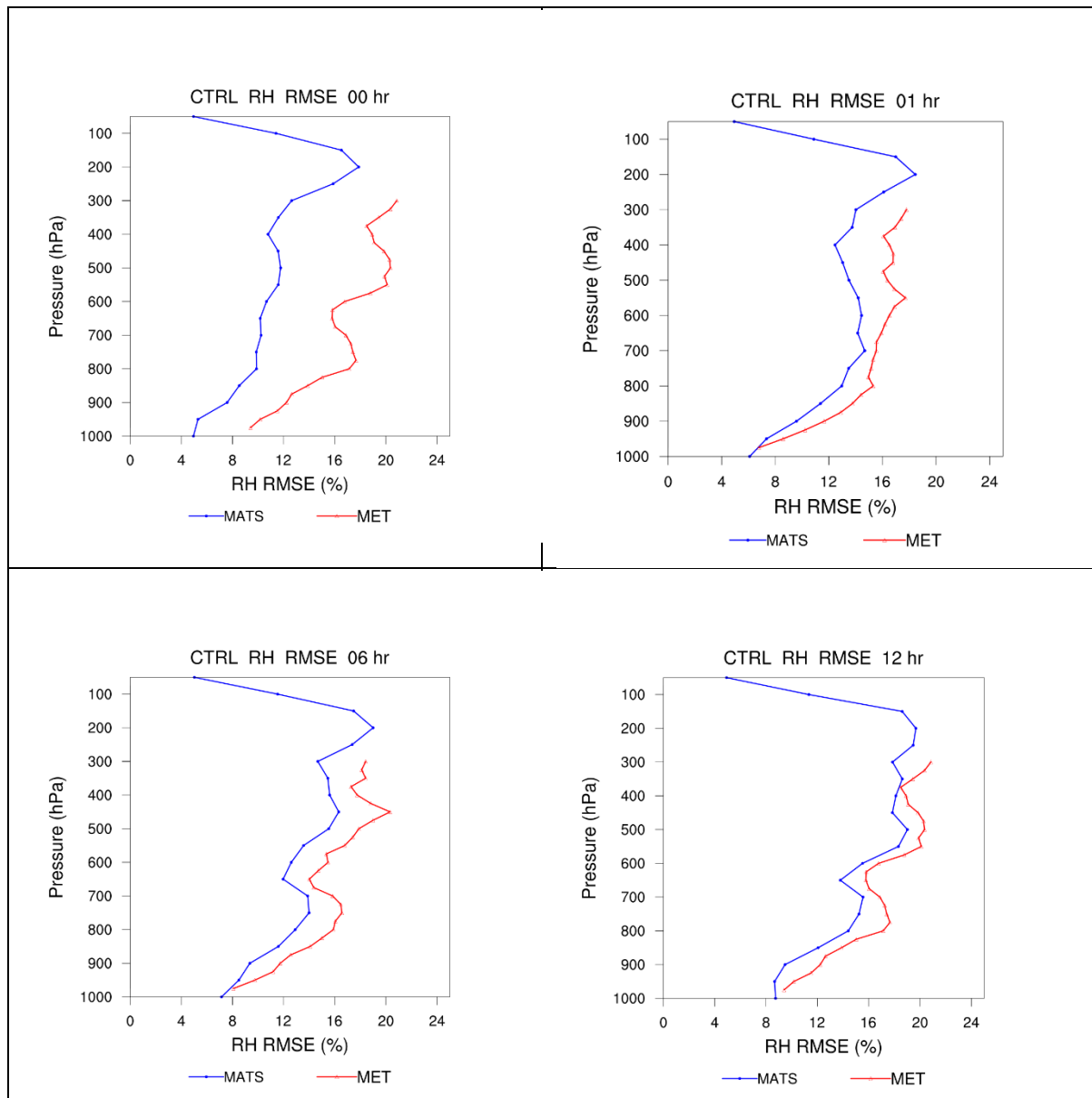
Figure A1.8: As in Figure A1.7, but for relative humidity RMS errors.

Now there are clear differences between the verification system, with MET giving larger RMS errors at all forecast lead times and for all vertical layers. A similar pattern is seen for the temperature biases (Fig. A1.9). The vertical structure of both verification systems interweaving together with height meaning that they are essentially indistinguishable from each other. But for the temperature RMS errors (Fig. A1.10), the MET verification system provides larger RMS errors than for MATS. The differences are not as great as were noted for relative humidity, but they are consistent across all vertical layers and for all lead times. For the wind speed biases (Fig. A1.11), again the biases are practically indistinguishable between the two verification systems. For wind speed RMS errors (Fig. A1.12) however, this time MATS has the larger RMS errors and MET has smaller errors. This is true for all vertical layers except one (600 hPa) at the 6-h lead time when they were essentially the same.
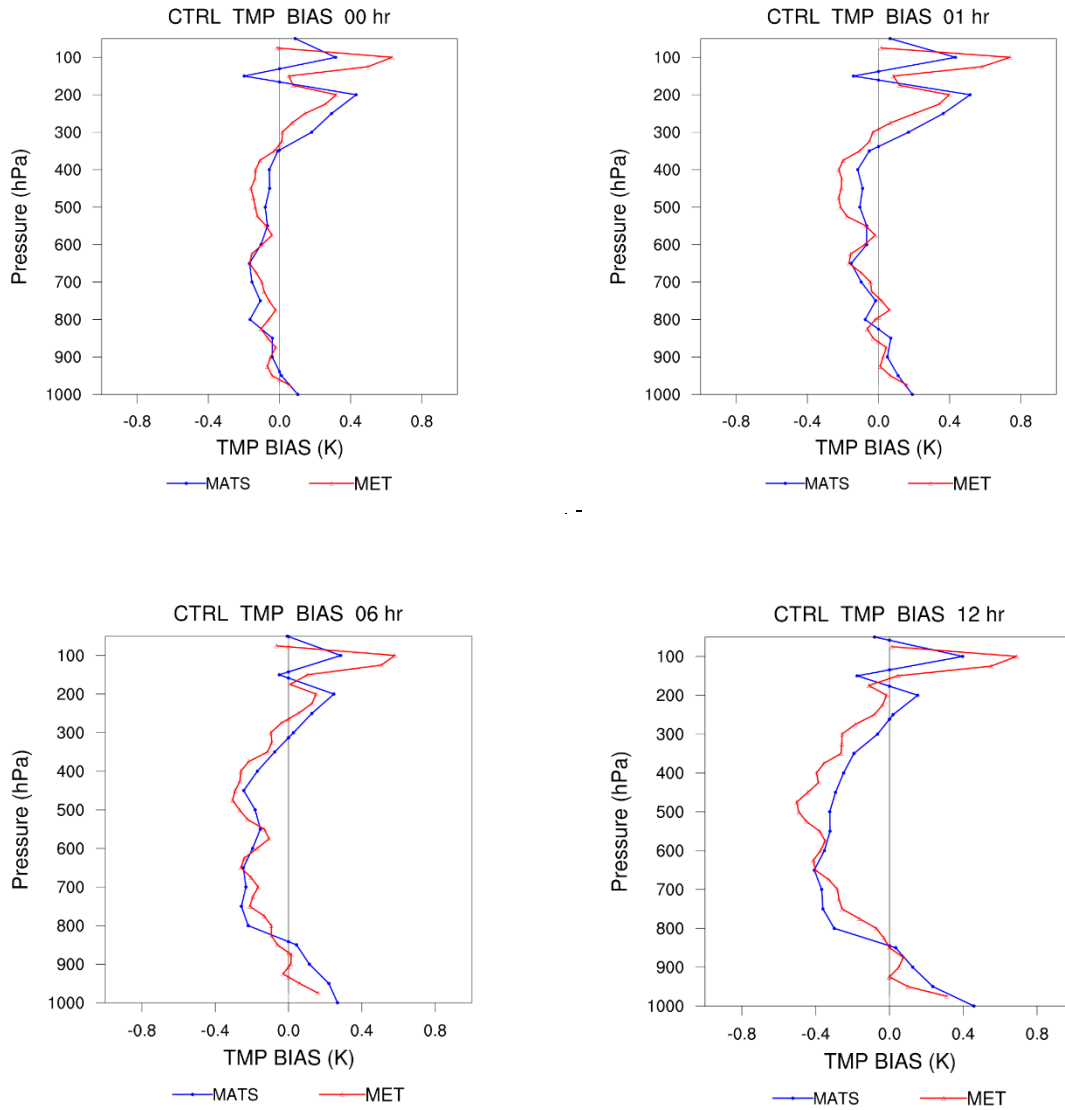
Figure A1.9 : As in Figure A1.8, but for temperature biases.
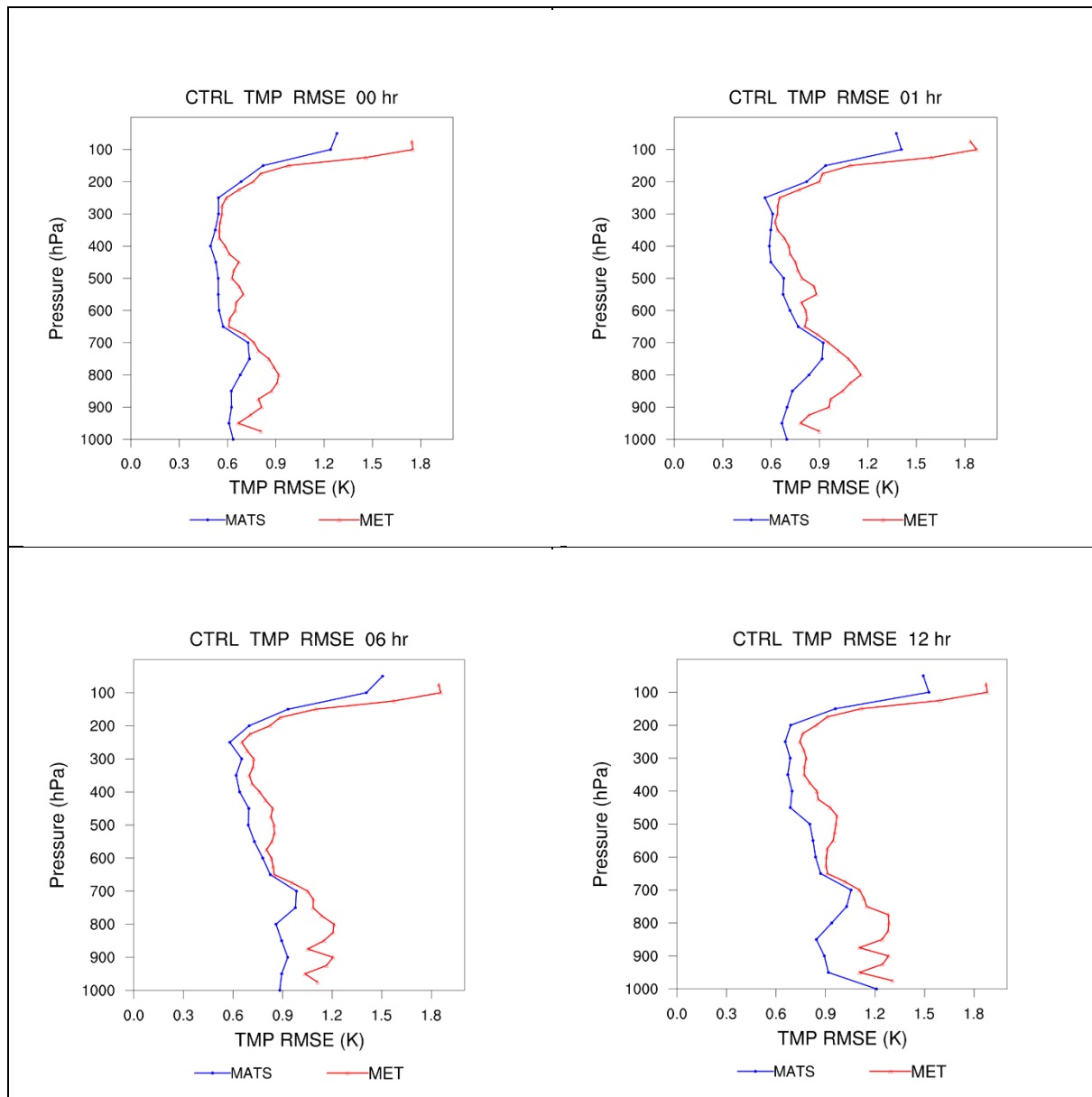
Figure A1.10 : As in Figure A1.9, but for temperature RMS errors.
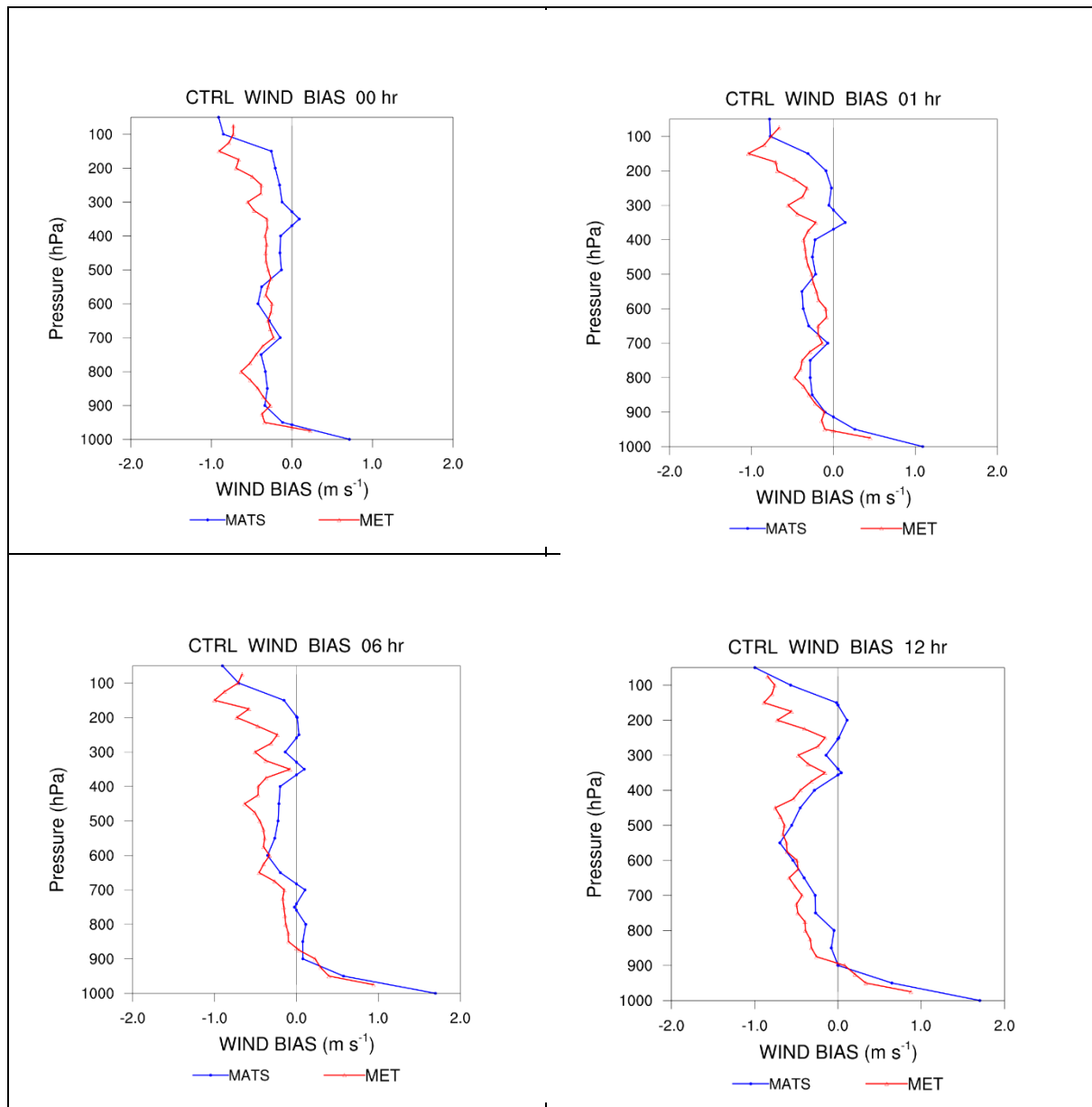
Figure A1.11: As in Figure A1.10, but for wind speed biases.
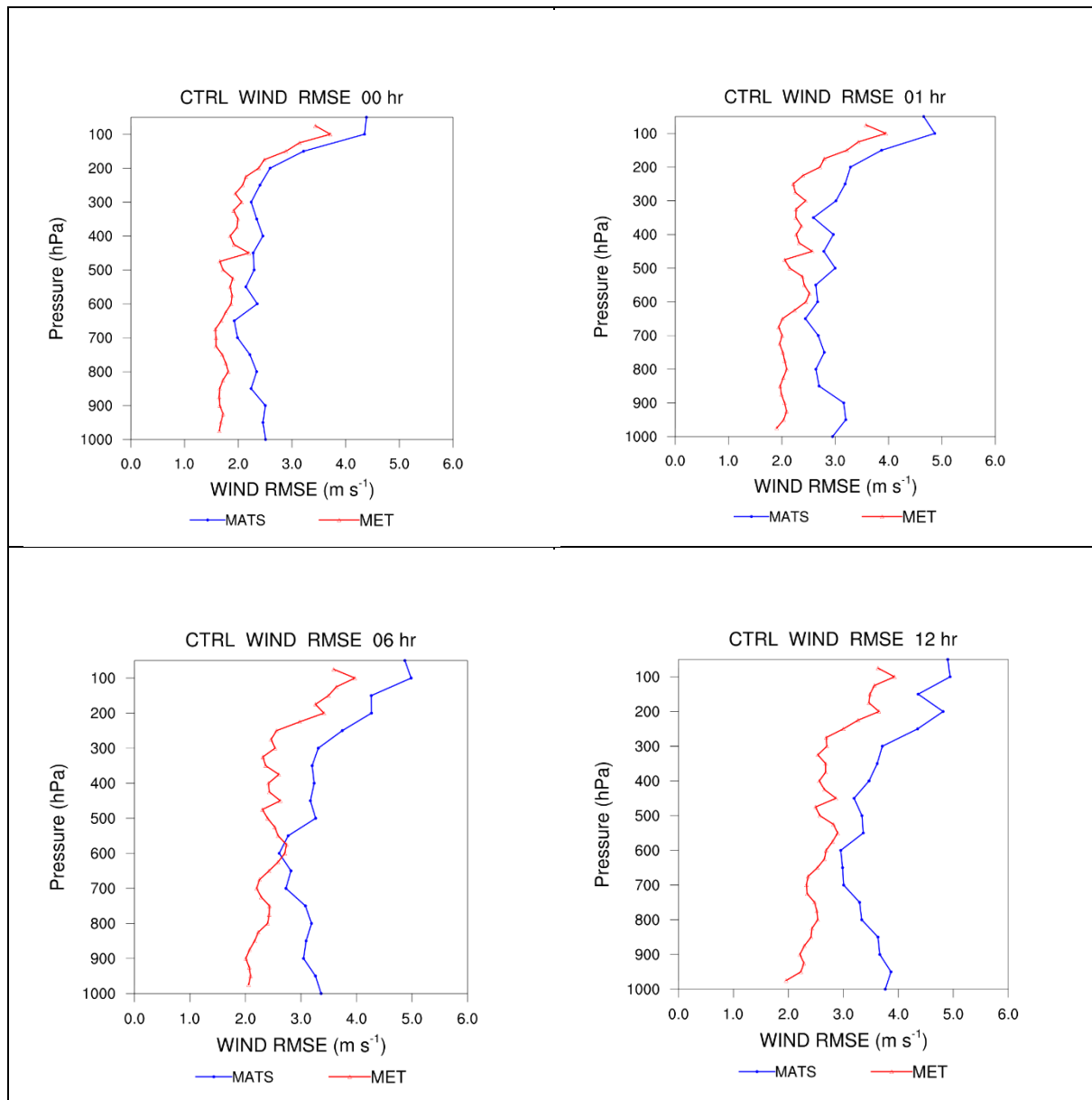
Figure A1.12 : As in Figure A1.11, but for wind speed RMS errors.

To attempt to understand the differences seen between the two verification systems, the MATS developers were contacted to find out if any methodological differences may be the cause. MATS uses a different approach to MET, in that it generates matched pairs by interpolating both the observations and the model data to the nearest 10 hPa level. From these matched pairs, MATS computes partial sums (SL1L2) lines an aggregate these for 50-hPa thick layers (every 50 hPa). These are the values seen in the plots. In MET, there is flexibility how the matched pairs can be generated. For this verification task, it was decided to use the observations at their native pressure level, and to vertically interpolate the model data to the level of the observation. Then the partial sums (SL1L2) lines are computed and aggregated for overlapping layers 50-hPa thick (every 25 hPa). The approached used in MET has the advantage that the observations do not have to be interpolated at all. The vertical distance over which the model data are interpolated depends on the vertical spacing of the model layers. In MET, the post-processed model output is available every 25 hPa. It is not

clear what model data are used in MATS (post-processed or native). One other possible significant difference between the two verification systems is that MET obtains the vertical sounding data from the PREPBUFR files that are used to initialize the model. MATS, on the other hand, uses an internal GSD database that has some quality control measures applied (occasionally, a bad sounding will be excluded).

Despite these methodological differences, it is still unclear exactly why the differences between the two verification systems occur. The fact that the vertical bias structures are essentially indistinguishable between the two systems suggests that whatever interpolation errors are occurring essentially cancel out by the time the aggregation to the 50-hPa layers occurs. The differences in RMS errors, however, suggests that the interpolation errors may not be entirely benign. The fact that MET has higher RMS errors for relative humidity and temperature, both of which may be strongly affected by interpolation errors, while MET has smaller RMS errors for wind speed (which might be expected to be less affected by interpolation errors), suggests the possibility that interpolation errors in MET's method and design could be causing these differences. However, the effect of using different input data cannot be ruled out. It is clear that more investigation will be needed to fully understand the differences between the two systems.

The investigation uncovered one other interesting fact. It had been thought that the PREPBUFR provided the high-resolution soundings, however after analysis, it became evident that the sounding data is quite coarse, sometimes with 50 hPa or more in between observations as the balloon ascended. To further analyze this, MET's *stat_analysis* utility was used to analyze the SL1L2 partial sums lines. Statistics from these was then analyzed in an Excel spreadsheet (Fig. A1.13). This turned up some interesting findings:

- At best, there are 9 obs per 50-hPa layer at ~100 hPa.

- At worst, there are 2 obs per 50-hPa layer around 700 – 800 hPa.

- Effective data sampling seems to be a function of sampling frequency and balloon rise rate.

To the last point, near the ground, the balloon rises quickly and the pressure levels are close together, so the sampling rate in pressure terms can be rather coarse. In the upper troposphere and lower stratosphere, the balloon rise rate may be slower and the pressure levels are far apart. This gives the appearance of having higher vertical density of observations in these regions, however it must be kept in mind that the picture may be quite different in absolute height terms (indeed, the vertical sampling rate may be relatively constant in height coordinates). Figure A1.14 shows a plot of the effective sampling computed from the MET-computed SL1L2 lines. This shows the variations of sampling rate with height just described.

This raise the interesting question as to how the verification results would change if high-resolution upper air data were used. For several years, the United Nations World Climate Research Program ran a program called Stratosphere-Troposphere Process and their Role in the Climate (SPARC), which took high vertical resolution radiosonde data (HVRRD) at sampling rates up to 1 Hz. Such soundings were taken for approximately four sites (possibly more). Interestingly, there are some vertical oscillations in the computed sampling rate. Possibly, this is due to aliasing and/or the presence of mandatory reporting levels.

Summary from MET's stat_analysis on how many data points went into each SL1L2 line for each pressure layer for the verification period (12Z on Sep 3 through 23Z on Sep 10)

Given that soundings are normally taken at 00Z and 12Z, there were 15 main raob observing periods (with the 16th period reported for some levels possibly from an 18Z sounding).

Level	Layer Center	# of SL1L2 lines analyzed	mean # of obs per SL1L2 line	total # of obs in period	estimated # of sounding sites for this pressure layer	# of obs per sounding site in period per layer	# of obs per sounding period per sounding site per layer	layer thickness (hPa)	Effective sampling interval of sounding data (hPa)	MEAN_NCL	MEAN_NCU	STDEV	MIN	P10	P25	P50	P75	P90	MAX	IQR
P100-50	75	15	155.4	2331	16	145.7	9.7	50.0	5.1	143.61008	167.18992	21.28984	132	133.8	140	156	164.5	168.8	216	24.5
P125-75	100	15	166.2	2493	21	118.7	7.9	50.0	6.3	158.84968	173.55032	13.27296	144	154.6	159	164	171.5	180.4	199	12.5
P150-100	125	15	188.93333	2834	21	135.0	9.0	50.0	5.6	185.5038	192.36286	6.19293	175	184.4	187	188	192.5	194.6	203	5.5
P175-125	150	15	108.53333	1628	21	77.5	5.2	50.0	9.7	103.99625	113.07042	8.19291	92	101	102.5	108	116	116.6	122	13.5
P200-150	175	15	110.73333	1661	21	79.1	5.3	50.0	9.5	106.07899	115.38767	8.40465	98	100.2	104	113	117.5	121	123	13.5
P225-175	200	15	88.06667	1321	21	62.9	4.2	50.0	11.9	82.2455	93.88784	10.51167	73	76.4	79	86	94.5	101	109	15.5
P250-200	225	15	104.8	1572	21	74.9	5.0	50.0	10.0	99.03657	110.56343	10.40741	94	96.4	97	101	111	118.8	129	14
P275-225	250	15	82.53333	1238	21	59.0	3.9	50.0	12.7	75.87851	89.18815	12.01705	69	71	74	79	85.5	98.8	111	11.5
P300-250	275	15	104.13333	1562	21	74.4	5.0	50.0	10.1	99.00261	109.26406	9.26489	91	95	98.5	103	107.5	112.4	129	9
P325-275	300	15	88.33333	1325	21	63.1	4.2	50.0	11.9	82.46643	94.20023	10.59425	70	72.2	81.5	91	96	98.6	104	14.5
P350-300	325	16	85.0625	1361	21	64.8	4.3	50.0	11.6	71.97031	98.15469	24.56954	1	73.5	83.75	88.5	99.25	101.5	106	15.5
P375-325	350	16	65.75	1052	21	50.1	3.3	50.0	15.0	55.47897	76.02103	19.2752	1	57	63.5	69	73	82	87	9.5
P400-350	375	17	75.82353	1289	21	61.4	4.1	50.0	12.2	60.61661	91.03045	29.57667	1	44.8	76	85	88	93.6	111	12
P425-375	400	18	71.61111	1289	21	61.4	4.1	50.0	12.2	54.75144	88.47078	33.90322	1	1.7	75	81.5	88.75	91.2	119	13.7
P450-400	425	19	66.21053	1258	21	59.9	4.0	50.0	12.5	48.49503	83.92602	36.75532	1	1.8	65	79	85.5	90.2	128	20.5
P475-425	450	18	53.61111	965	21	46.0	3.1	50.0	16.3	39.67019	67.55203	28.03388	1	1.7	49.25	58.5	62.75	82.9	104	13.5
P500-450	475	16	82.375	1318	21	62.8	4.2	50.0	11.9	69.04438	95.70562	25.01699	3	69	78.75	85.5	90.75	105.5	119	12
P525-475	500	16	85.125	1362	21	64.9	4.3	50.0	11.6	71.9846	98.2654	24.66002	6	71	81	87	99.25	106.5	111	18.2
P550-500	525	16	86.1875	1379	21	65.7	4.4	50.0	11.4	72.74469	99.63031	25.22755	7	72	78.5	92	99	108	119	20.5
P575-525	550	16	64.4375	1031	21	49.1	3.3	50.0	15.3	53.99019	74.88481	19.60602	4	54.5	58	65.5	77	80.5	94	19
P600-550	575	16	57.75	924	21	44.0	2.9	50.0	17.0	48.83882	66.66118	16.72324	3	49	53.75	60.5	67.5	72	75	13.7
P625-575	600	16	53.0625	849	21	40.4	2.7	50.0	18.6	45.08386	61.04114	14.97317	4	41.5	52	58.5	60	62.5	66	8
P650-600	625	16	52	832	21	39.6	2.6	50.0	18.9	44.05576	59.94424	14.90861	3	42	49	56.5	61	61.5	66	12
P675-625	650	16	49.75	796	21	37.9	2.5	50.0	19.8	41.5513	57.9487	15.38614	1	41	45.5	53	56.5	64.5	68	11
P700-650	675	16	66	1056	21	50.3	3.4	50.0	14.9	55.74465	76.25535	19.24578	1	56.5	63.25	72	74.25	80.5	84	11
P725-675	700	16	64.875	1038	21	49.4	3.3	50.0	15.2	55.04397	74.70603	18.44948	2	58	60	69.5	74	78	84	14
P750-700	725	16	64.1875	1027	21	48.9	3.3	50.0	15.3	54.97403	73.40097	17.29053	2	62	63.75	67.5	72	72.5	81	8.2
P775-725	750	15	46	690	21	32.9	2.2	50.0	22.8	42.70379	49.29621	5.95219	36	39.4	42	46	48.5	54.2	57	6.5
P800-750	775	16	40.5	648	20	32.4	2.2	50.0	23.1	34.13532	46.86468	11.94432	4	33	36	41.5	46	51.5	57	10
P825-775	800	16	38.625	618	20	30.9	2.1	50.0	24.3	32.72368	44.52632	11.07475	6	31	34	40	47.25	49	51	13.2
P850-800	825	16	55.5	888	20	44.4	3.0	50.0	16.9	46.16654	64.83346	17.51571	3	42.5	48	59	64.25	72.5	78	16.2
P875-825	850	16	49.25	788	19	41.5	2.8	50.0	18.1	40.13925	58.36075	17.09776	1	36	42.75	52	59.5	65	71	16.7
P900-850	875	16	43.5625	697	18	38.7	2.6	50.0	19.4	35.72388	51.40112	14.7104	2	35	36	46	49.25	58	70	13.2
P925-875	900	16	41.5	664	17	39.1	2.6	50.0	19.2	33.77564	49.22436	14.49598	3	31	33.75	44.5	48.5	54.5	68	14.7
P950-900	925	16	38.6875	619	16	38.7	2.6	50.0	19.4	31.08736	46.28764	14.26286	2	27	29.75	39.5	49.5	53.5	58	19.7
P975-925	950	15	40.86667	613	14	43.8	2.9	50.0	17.1	35.92737	45.80596	8.91921	29	29.4	33	45	48.5	50	52	15.5
P1000-950	975	15	38.26667	574	10	57.4	3.8	50.0	13.1	34.51385	42.01948	6.77671	29	31.4	32	39	43	48	48	11
P1013.2-975	994.1	15	27.66667	415	7	59.3	4.0	50.0	12.7	24.73882	30.59451	5.287	20	20.8	23.5	29	31.5	33.8	36	8

Figure A1.13 : A screenshot of an Excel spreadsheet showing an analysis of the SL1L2 statistics and a calculation of the effective sampling rate. The green shaded highlighted cells show the region with the highest effective vertical sampling rate (150 – 50 hPa layer), while the red shaded cells show the region with the lowest effective vertical sampling rate (825 – 775 hPa).
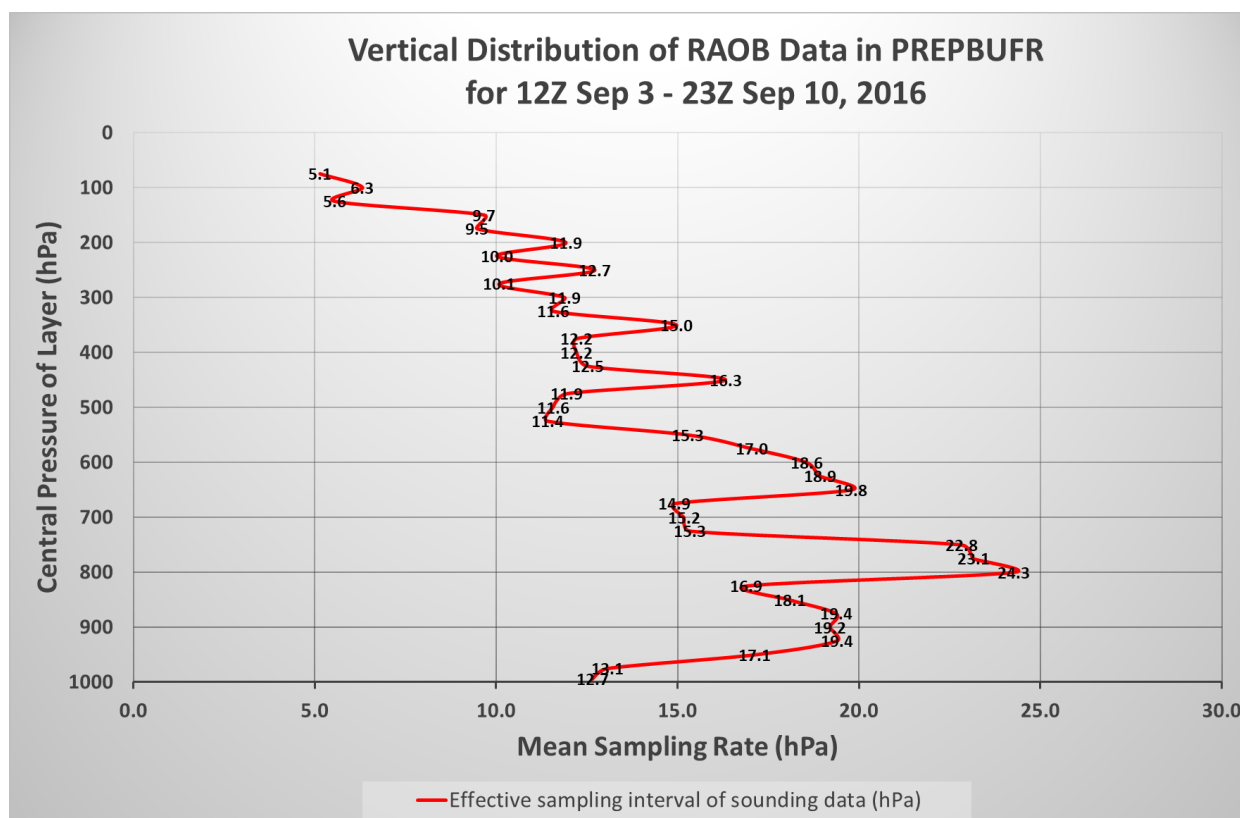
Figure A1.14 : A plot of the effective vertical sampling rate, based on the analyzed data shown in Fig. A1.13.

A5. Summary

A comparison of the characteristics of the observations suggests that both ADPUPA and AIRCAR are suitable for use in this verification study, however the other observation types

(PROFLR, VADWND, and AIRCFT) had data that were too few and/or had too large of uncertainty.

Some noticeable differences were noted in the verification results depending on which verification system was being used (MATS vs. MET). A small side study was conducted to try to ascertain the reasons for these differences. Results possibly point to methodological differences as the cause for the differences, but differences in the observational data cannot be ruled out.

It would be interesting to obtain the high-resolution sounding data from the SPARC program and repeat the verification to see what effect interpolation/sampling rates have on the verification results.

It would also be interesting to expand the scope of the effort to compare MATS and MET to learn whether the differences between two systems can be explained by methodological differences or by due to data differences. As a first step, MET could be run using the GSD sounding data rather than the sounding data from PREPBUFR. The MET configuration could also be setup to mimic the MATS approach. After repeating the verification, one would hope that the results will be quite similar. If noticeable differences still exist, an analysis of the matched pairs could then be undertaken. If there is a one-to-one correspondence of the matched pairs, this could point to the possibility of a bug in either system.

Finally, when computing confidence bounds, the data used should be independent, however it is not clear at what vertical spacing observations are truly independent. Failure to properly account for this factor may lead to confidence bounds that are overly confident if higher resolution vertical sounding data are used. There may be some ideal rate of vertical sampling that results in the most accurate confidence bounds. Further work could be done to investigate what the ideal sampling rate might be for sounding data.