

(1) In general, the analysis needs to be better presented. There remain many points of contention and disagreement between what is stated and what is evident from figures or logic. There are also useful points of discussion that are omitted. Many improvements to the writing and flow of the text would make section 4 read much more cleanly and the logic flow better. This includes cleaning up many of the figures (esp. Figs. 5-9).

(a) Section 4.1:

- Lines 285-287: Explain, as that is not what a "mean" innovation traditionally measures. If the model simply has displacements, then the mean should be zero.
- Lines 309-311: Based on cf. Figs. 5a,b I would say the primary shortcoming of NODA is that it depicts a single/merged linear convective system rather than the two larger linear systems evident in MRMS and all other experiments. Yes, the coverage is also reduced compared to observations, but that is secondary in importance.
- Line 323: "hook echo" appears. Your response to my comment #20 from the previous review said this was fixed. *Apparently there were mis-copies between the responses and the formal manuscript file.*
- Lines 331-332: Name these towers "eastern" and "western" for ease of reference later on. Example: "primary" tower (line 372) is unhelpful.
- Lines 338-345: Add that the three DA experiments also slightly under-forecast the top of the precipitation region (5 dBZ contour generally around 11-12 km vs. 12-13 km in MRMS). In the next paragraph note that the DA increased the height of the top of the convective system in all experiments, bringing them close to that shown in MRMS. This messaging will increase consistency in the discussion between here and the next few paragraphs, including Fig. 7, which show positive snow increments at high levels in the middle of the lateral span.
- Figures 5-7: Named/labeled panels should be described in the caption. Also explain why there are no curves for NODA in the analysis panels.
- Figures 6 and 7: Clean up the x-axis tick labels to include fewer decimal points; the current appearance is excessive.
- Figure 7: Can you plot the y-axis ticks using geometric height to be consistent with Fig. 6?
- Lines 374-377: the  $q_r$  increments may be relatively weak compared to those of  $q_s$  and  $q_g$ , but the location of the positive  $q_r$  increments in the western tower make it pretty clear that what little rainwater was added made the difference in the appearance and height of the western tower (based on reflectivity), which, as you discussed previously, was underdone in the backgrounds of all DA experiments. This is an important point that should not be omitted.

(b) Section 4.2:

- Figure 8 is too difficult to inspect to compare the forecasts and check the narrative associated with it. The panels are too small and there is too much detail to cover. I propose a different type of plot to analyze: instead, plot paintballs (contour outlines only, don't fill) of one or two thresholds of interest such as 35 or 40 dBZ for the

convective cores (to compare the quality of the squall line) and 10 or 20 dBZ to represent the overall size and position of the entire scenario. Use one panel for each time and plot all forecast types on the same plot for quick and easy comparison. This will also allow for larger images. Additionally, in the caption, please specify the valid times.

- Line 438: I disagree. How are you defining the cold pool? I examined the areal extent of the  $\theta_e = 332$  K contour near the gust front gradient. To me the spatial extents of the cold pool between URMA and NODA look similar.
- Lines 439-441: This is more likely an effect of the convective system being displaced from observations than any direct result of DA impacting these fields via cross covariances.
- Lines 442-443: (overall cold pool structure) I remain dubious of this claim. You need to provide additional evidence to make it apparent. Plot a difference plot of surface wind speeds or temperatures, for example.
- Line 447: It remains impossible to confirm this via inspection of Fig. 9. This must be ameliorated before I will recommend acceptance.
- Lines 456-457: This was a methodological choice, right? If so, you only need to state that.
- Figure 9: GSI-new and JEDI-new differences seem imperceptible. It would be more effective to plot a difference field between these two. In the caption you need to specify the data source of the 30-dBZ contours and consider plotting contours for each forecast system instead of the same one. Finally, the remark about the now-removed yellow contours remains.
- Lines 463-477: since you only plot radii up to 72 km, which is far smaller than half the domain size (the maximum reasonable neighborhood size over which to compute FSS), you don't know what the FSS would look like at larger radii and so it is misleading to state "at all radii" for this analysis. You can state something like "at radii < 72 km" or "all tested radii" instead. It is likely that both HRRRv4 and RRFS p2 CNTL attain skillful FSS values at 45 dBZ at radii larger than 72 km.
- Line 475-493: Bias impacts on FSS can be addressed by substituting a percentile threshold for the fixed threshold. You can try that, or you should link this comment better to the discussion of the performance diagram in the next paragraph using connective language. In the subsequent paragraph, specifically state these bias figures. You currently only do that for GSI-New.
- Lines 500-503: Unfortunately, as these forecast systems exhibit two changes (dynamical core and DA method) you cannot untangle the contributions of each to the different biases. You should state this instead of repeating what has already been stated earlier in the manuscript, as the current language is tired and repetitive.

(c) Improving the discussion of neighborhood UH probabilities:

- Line 518: "higher neighborhood probabilities": ambiguous – what part of the forecast region are you referring to? NODA shows the absolute largest overall probability magnitudes of all forecast systems.

- Line 519: I think you have to say 50% since the 60% region is technically broken into two pieces with a large gap between 60% contours.
- Lines 521-522: Redundant. Shorten or delete.
- Line 524: insert “including a region with a peak probability in SE North Dakota associated with few severe reports and < 10% probabilities associated with a cluster of hail and wind reports in eastern Nebraska.” At the end of this sentence.
- Line 525: (changes to manuscript not required) this suggests the low UH probabilities in HRRRv4 are a bias problem rather than a total forecast miss. Imagine inflating the probabilities in HRRRv4 by 50% or so (not percentage points...multiplicatively). The resulting plot would look fairly similar to GSI-New and JEDI-New, with one exception being in E/SE Nebraska where the HRRR appears to have simply missed the potential for severe storms there.
- Lines 526-527: (changes to manuscript not required) It is interesting that RRFS p2 does capture a weak secondary (perhaps "tertiary" is a better description) region of UH probabilities in southern Nebraska, in agreement with GSI-New and JEDI-New, but something that HRRRv4 does not represent.

*(Less-significant comments)*

(2) Equation 3: When describing this equation, please include the information in your response to my previous comment (number 6) about the units of  $Z_e$  and where the  $10^{18}$  comes from.

(3) Equation 5: please define or qualitatively describe  $a(T)$  and  $b(T)$

(4) Equation 7: Since only two factors of the six in this equation differ between the two temperature regimes, I suggest using a single equation with variable names for the two factors that change, and then state in the text what the values are for  $T > 0^\circ\text{C}$  and  $< 0^\circ\text{C}$

(5) Figure 1: The symbol for "New" in the legend does not correspond to the actual data points plotted on the axes.

(6) Lines 167-170: What meteorological case were these data based on? The same as the one analyzed later in the manuscript or something else?

(7) Language improvements

- Line 41: replace “ensembles that...forecasting systems” with “ensembles of initial condition perturbations for ensemble forecasting systems”
- Line 60: “GSI” has not yet been defined
- Line 64: “next generation DA framework” is too vague. Be specific.
- Line 70: missing a space after the reference

- Lines 96-97: The way this sentence starts suggests that the point of this sentence should be to enable a fair comparison between the implementations of the TL/AM in GSI and JEDI. Rephrase.
- Line 100: change “convective storm case” to “high-impact severe weather event”
- Lines 101-102: repeated text from lines 67-69. You can summarize this by stating that you're comparing the JEDI 3DEnVar approach to HRRR and RRFS forecasts that use the previously-discussed alternative approaches to assimilating radar reflectivity for variety.
- Line 117: “The” Thompson
- Line 212: include “gnomonic” when naming the Schmidt projection
- Lines 222-223: re-write so that "ensemble" doesn't appear so frequently. You can replace "ensemble" with "member" or eliminate the word somewhere.
- Line 236: insert “mixing ratios of” before “rainwater” in the list.
- Lines 306-309: Delete – you have already stated this (line 188).
- Line 318: delete everything starting at “in this case...”
- Line 341: delete “compared to...(end of sentence)”
- Line 342: delete “associated with the Thompson microphysics scheme”
- Line 346: insert “partially” before “recovers”
- Line 348: replace “differences remain relative to MRMS” to “the tower is too short in all experiments compared to MRMS.”
- Line 349: insert “additionally” before “underestimate its strength”
- Line 349: insert “GSI-New and” before “JEDI-New” and change the verb plurality subsequent to this addition.
- Line 350: insert “, but the resulting structure looks arguably better in JEDI-New compared to GSI-New” at the end of the sentence.
- Lines 378-382: none of this seems necessary to state; it is a remix of what has already been stated.