Report 1

*I understand there are technical challenges for extending the DA experiments beyond 6 cycles; while the present experiments serve as a good demonstration for the developed operator capabilities, I maintain my position that extending the current set of experiments beyond 6 cycles is a necessary *minimum* requirement to show the DA system is stable. The operator developments are directly linked to these ob-space statistics and demonstrating stable performance is an important byproduct of your results*.

Reply: To further assess the stability of the DA system, we repeated experiment JEDI-New using the same 1900–0000 UTC DA window but with a 15-min DA interval instead of the original hourly interval. This yielded 21 analysis times, compared to 6 analyses in the original JEDI-New experiment.

The results indicate that the DA system remains stable under this more frequent cycling configuration. The RMSIs decrease substantially after the first analysis, but they do not continue to decrease monotonically during the later cycles. Therefore, the 15-min cycling experiment does not suggest that more frequent cycling necessarily provides additional improvement in this case. However, the analysis RMSI remains consistently lower than the corresponding background RMSI throughout the cycling period, indicating that each DA update continues to improve the fit to the observations and that there is no evidence of filter divergence.

One possible reason for the later increase in RMSI is that the covariance inflation was not specifically tuned for the 15-min cycling configuration. More frequent cycling may reduce the ensemble spread more rapidly, thereby requiring stronger covariance inflation. Retuning the inflation parameters for the 15-min cycling configuration and repeating all experiments would require substantial additional work and is beyond the scope of the present revision. Therefore, we use this experiment only as a supplementary stability check rather than replacing the main hourly cycling experiments.

We have added a brief statement in section 4 of the revised manuscript to mention this 15-min cycling sensitivity experiment. The detailed figure is not included in the manuscript because the experiment does not alter the main conclusions. We have also added a caption to the figure included in the response, defined HxB and HxA as the observation equivalents from the background and analysis, respectively, and used RMSI instead of RMSE for consistency with the manuscript.

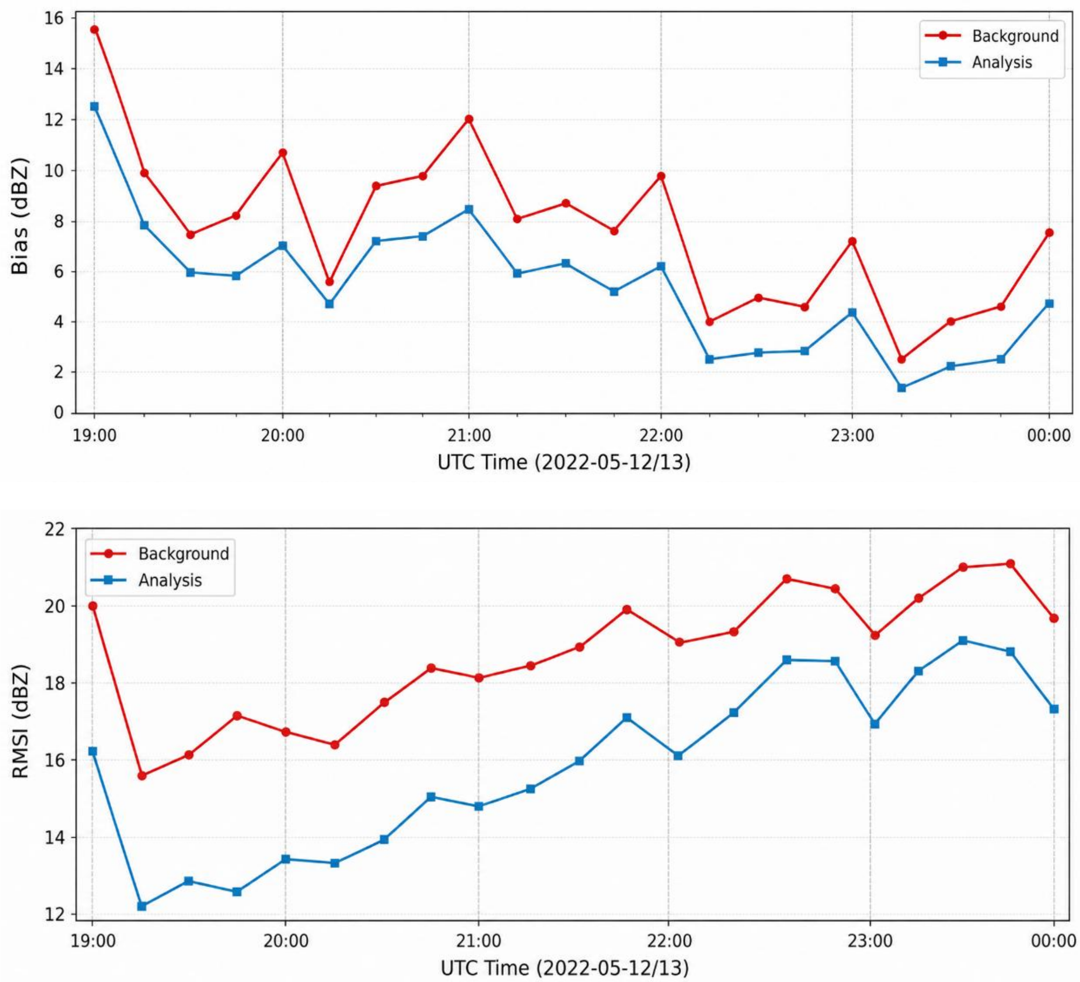


FIG. 1. Time series of radar reflectivity (top) bias and (bottom) root-mean-square innovation (RMSI; dBZ) for the En3DVar data assimilation experiment. The verification period spans from 1900 UTC 12 May to 0000 UTC 13 May 2022. Red lines with circles denote the background, and blue lines with squares denote the analysis.

For my minor comment related to Section 3.2.2, although equations are presented in previous papers, it will be good practice to write an appendix to make this paper self-contained.

Response: In the revised manuscript, we added an appendix that summarizes the incremental En3DVar framework used in this study. We also clarify that the experiments in this paper use a pure En3DVar configuration without static background covariance, so that the background error covariance is represented by the localized ensemble covariance.

The appendix is intended to be a concise summary of the formulations most directly relevant to this study, while full implementation details of JEDI EnVar are available in Liu et al. (2022b).

Report 2

(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.*

Reply: We agree that spatial displacement should not contribute to bias. Here bias is not the right description, and we have removed reference to bias and displace error, only point out that the positive mean innovation is related to the under-prediction of reflectivity.

The revised text after Fig. 5 now reads: “Without radar DA, NODA (Fig. 5e) fails to reproduce the observed convective organization. In particular, it depicts a more connected linear convective system, rather than two linear systems with larger areas of stratiform precipitation as in the MRMS observations.”

• *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.*

Reply: We agree that our original description did not sufficiently emphasize the primary structural deficiency in NODA. As shown in Fig. 5e, the most notable shortcoming of NODA is that it represents the convection as a more merged, narrower linear system, rather than reproducing the two larger linear convective systems evident in the MRMS observations and in the DA experiments. The reduced reflectivity coverage in western Minnesota and near the South Dakota–North Dakota border is also present, but it is secondary to the structural error in the organization of the convective system. We have revised the text accordingly.

Revised manuscript text:
"Without radar DA, NODA (Fig. 5e) fails to reproduce the observed convective organization. In particular, it depicts a more merged and narrower linear convective system, rather than the two larger linear systems evident in the MRMS observations. NODA also underrepresents the reflectivity coverage in western Minnesota and near the South Dakota–North Dakota border."

- *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.*

Reply: The reviewer is correct that this term should have been removed in accordance with our response to the previous review. It was an oversight during editing. We have now corrected this sentence and carefully rechecked the manuscript to ensure that the term "hook echo" is not used in this context.

Revised manuscript text:
"In contrast, the JEDI-Org analysis (Fig. 5g) shows less coherent storm organization, particularly in the southern segment, where the reflectivity enhancement is poorly represented."

- *Lines 331-332: Name these towers "eastern" and "western" for ease of reference later on. Example: "primary" tower (line 372) is unhelpful.*

Reply: We have revised the text accordingly and replaced less specific terms such as "primary" and "secondary" towers with these directional labels where appropriate.

Revised text:
"The MRMS observation (Fig. 6a) displays a well-developed deep convective system, featuring two upright convective towers: an eastern tower near 95°W and a western tower near 99°W."

- *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.*

Reply: We agree that the vertical extent of the precipitation region should be described more consistently across the discussion of Figs. 6 and 7. Accordingly, we added text in the Fig. 6 discussion noting that all three background fields slightly underpredict the 5-dBZ echo-top height relative to MRMS. We also revised the following paragraph to note that the analysis update increases the storm-top height in all experiments, bringing it closer to that shown in MRMS, and that this behavior is consistent with the positive snow increments at upper levels in Fig. 7.

We have added in the Fig. 6 discussion:
"… all three background fields slightly underpredict the top of the precipitation region, with the 5-dBZ echo top generally reaching only about 11–12 km, compared with about 12–13 km in MRMS."

Added in the Fig. 7 discussion:
"… these positive upper-level snow increments are consistent with the increased storm-top height seen in the analysis cross sections of Fig. 6."

- *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.*

Reply: We revised the captions of Figs. 5–7 to explicitly describe the labeled panels. For Fig. 7, we now describe the panel organization by rows and columns, which is more concise and consistent with the layout of the figure. We also clarified in the Fig. 5 caption that NODA has no analysis field because no data assimilation is performed in that experiment.

- *Figures 6 and 7: Clean up the x-axis tick labels to include fewer decimal points; the current appearance is excessive.*

Reply: corrected

- *Figure 7: Can you plot the y-axis ticks using geometric height to be consistent with Fig. 6?*

Reply: Yes. We changed the y-axis ticks to height coordinates.

- *Lines 374-377: the qr increments may be relatively weak compared to those of qs and qg, but the location of the positive qr 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.*

Reply: We agree that our original wording understated the role of the rainwater increments. Although the qr increments are weaker in magnitude than those of snow and graupel, their positive increments near the western tower are spatially collocated with the analyzed improvement of that tower in reflectivity. This suggests that the added rainwater contributed to the appearance and vertical development of the western tower, which was underrepresented in the background fields of all DA experiments. We have revised the text accordingly to clarify that, while the major differences among experiments are still associated with the treatment of snow and graupel, the rainwater increments also play an important role in restoring the western tower.

Revised text: .... However, the positive rainwater increments near the western tower appear to contribute to the restored reflectivity structure and vertical development of that tower in the analyses, which was underrepresented in the background fields of all DA experiments. Thus, while differences among experiments still arise mainly from the handling of snow and graupel rather than from rainwater, the rainwater increments also play a non-negligible role in improving the western tower.

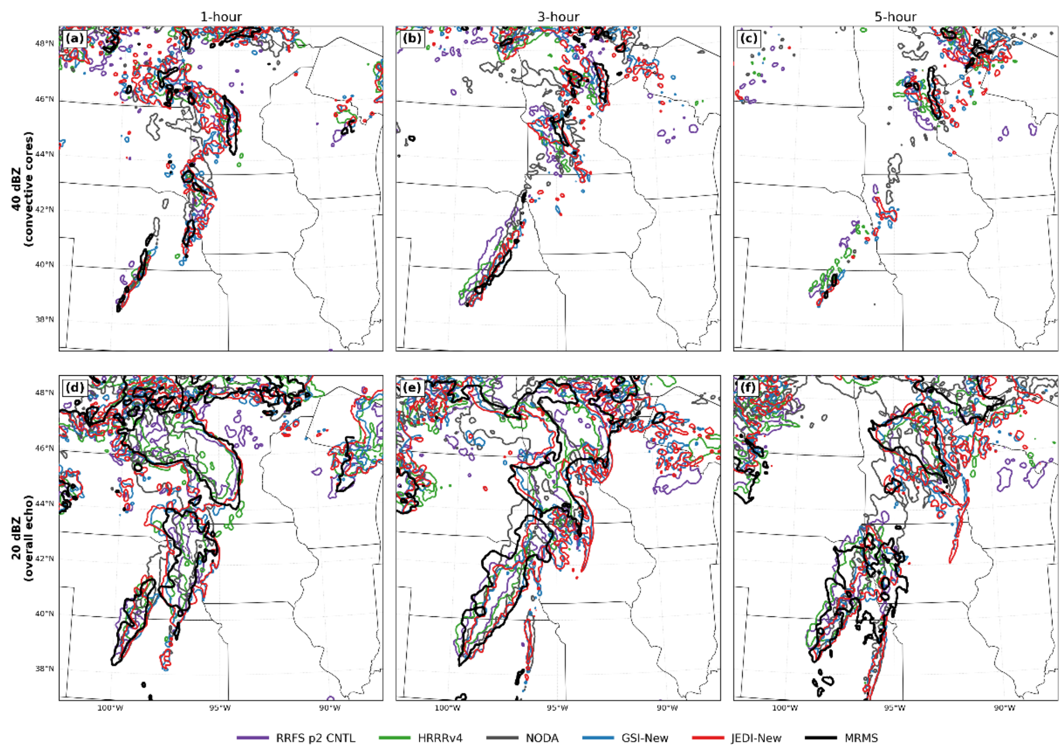
(b) Section 4.2:

- *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.*

Reply: We thank the reviewer for this thoughtful suggestion. We agree that the original Fig. 8 is visually dense and that a simpler comparison would be desirable. Following the reviewer's recommendation, we tested contour-overlay ("paintball") style plots using representative higher and lower reflectivity thresholds. However, for this case, overlaying multiple forecast systems still led to considerable contour overlap even for the higher threshold, while the lower-threshold contours representing the broader precipitation extent became especially cluttered (see figure below). Consequently, the contour-overlay version was not clearly easier to interpret than the original reflectivity maps(due to the complex structures, line contours are much harder to visualize that filled contours).

We therefore retained the filled-reflectivity format, but revised the figure presentation to improve readability. Specifically, the original figure has been split into two continued parts, with MRMS, NODA, and GSI-New shown in the first part and JEDI-New, HRRRv4, and RRFS p2 CNTL shown in the continued part. This allows larger individual panels while preserving the spatial structure and intensity information needed for comparison. In addition, the observed 40-dBZ composite reflectivity contour is overlaid in black in the forecast panels to facilitate comparison of the convective cores. We have also revised the caption to explicitly specify the valid times.

Revised caption:
Figure 8. Composite reflectivity forecasts and observations at 1, 3, and 5 h lead times, valid at 0100, 0300, and 0500 UTC 13 May 2022, respectively. Rows show MRMS observations, NODA, GSI-New, JEDI-New, HRRRv4, and RRFS p2 CNTL, and columns show the 1-, 3-, and 5-h forecasts. In the forecast panels, the observed 40-dBZ composite reflectivity contour is overlaid in black.



- *Line 438: I disagree. How are you defining the cold pool? I examined the areal extent of the θe = 332 K contour near the gust front gradient. To me the spatial extents of the cold pool between URMA and NODA look similar.*

Reply: Our intent was to describe the spatial coverage of the low-θe air associated with the cold pool as depicted in Fig. 9. In that sense, the low-θe region in NODA appears less expansive than in URMA, especially in the area behind the gust-front zone, while the leading edge of the cold pool is also displaced farther north. We have revised the text to make this interpretation more precise and less dependent on an implicit threshold definition.

Revised text: Figure 9 shows 2-m θe and 10-m wind fields from the UnRestricted Mesoscale Analysis (URMA; NOAA 2018) and the final analyses from the NODA, GSI-New, and JEDI-New experiments. Compared to URMA (Fig. 9a), NODA (Fig. 9b) exhibits a less expansive region of low-θe air associated with the cold pool, and its leading edge or gust-front zone remains closer to the South Dakota–Minnesota border. In contrast, URMA and the two DA experiments (Figs. 9a,c,d) show broader low-θe coverage and a farther southward extension of the near-surface cold-pool boundary.

- *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.*

Response: We agree that the improved near-surface θe and 10-m wind patterns in the DA experiments should not be overinterpreted as a direct result of radar DA updating these fields through cross covariances alone. In this case, the closer agreement with URMA is likely influenced substantially by the improved placement and organization of the convective system relative to the observations, as well as the model adjustment over the DA cycles. We have revised the text accordingly to avoid overstating the role of cross covariances in the near-surface analysis improvement.

Revised the text:
Compared to URMA (Fig. 9a), NODA (Fig. 9b) exhibits a less expansive region of low-θe air associated with the cold pool, and its leading edge or gust-front zone remains closer to the South Dakota–Minnesota border. In contrast, both DA experiments (Figs. 9c,d) show broader low-θe coverage and a farther southward extension of the near-surface cold-pool boundary in better agreement with URMA

- *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.*

Reply: We have revised the sentence to avoid this overstatement. The revised text now focuses on the more directly supported result that the DA experiments produce stronger surface winds behind the gust front than NODA, which is more consistent with the analyzed near-surface pattern in URMA.

Revised text: In the DA experiments, relatively stronger surface winds are present behind the gust front than in NODA, and this feature is more consistent with the analyzed near-surface pattern in URMA (Figs. 9c and 9d).

- *Line 447: It remains impossible to confirm this via inspection of Fig. 9. This must be ameliorated before I will recommend acceptance.*

Reply: To avoid overstating the comparison, we have removed the claim that the DA flow patterns are qualitatively closer to URMA in this region.

- *Lines 456-457: This was a methodological choice, right? If so, you only need to state that.*

Reply: To avoid unnecessary speculation, we deleted the final sentence.

- *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.*

Reply: The main purpose of Fig. 9 is to demonstrate that GSI-New and JEDI-New yield very similar near-surface analyses, rather than to emphasize small differences between them. We therefore retained the four-panel figure but revised the text to avoid over-interpreting differences between the two experiments. We also revised the caption to explicitly identify the 30-dBZ contours as observed MRMS composite reflectivity and removed the outdated reference to yellow contours. We retained the same observed contours in all panels because they serve as a common observational reference for comparing the analyzed near-surfaceE cold-pool and wind structures across the experiments.

- *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.*

Reply: We have revised the text to clarify that the comparisons apply only within the tested range of radii. We also acknowledge that the reference forecasts may achieve skillful FSS values at larger radii beyond those examined in this study.

- *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.*

Reply: We agree that reflectivity bias can affect FSS computed using a fixed threshold, especially at larger neighborhood radii. Rather than introducing an additional percentile-threshold analysis, we revised the text to better connect this point to the subsequent discussion of the performance diagrams. We now explicitly report the bias values for all forecast experiments in the precipitation performance-diagram discussion, so that the possible role of bias is more clearly linked to the interpretation of the FSS results.

- *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.*

Reply: We agree that, because these forecast systems differ in both dynamical core and DA method, the present comparison does not allow us to isolate the relative contributions of these factors to the bias differences. We have revised the sentence accordingly and removed the repetitive wording.

Revised text: Because these forecast systems differ in both dynamical core and DA method, the present comparison does not allow us to disentangle their respective contributions to the bias differences.

(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.*

Reply: Our intent was to describe the probability maxima in the vicinity of the observed severe reports, not the domain-wide probability magnitude. We have revised the sentence to clarify this point and to emphasize the improved spatial collocation of the forecast probability maxima with the observed reports.

Revised text: Compared to NODA and the reference forecasts, both GSI-New and JEDI-New show better spatial alignment between the forecast probability maxima and the observed severe reports.

- *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.*

Reply: We have revised the text to use the 50% threshold instead, which better represents the continuous probability swath in GSI-New.

- Lines 521-522: Redundant. Shorten or delete.

Reply: We have shortened the discussion by removing the repeated statement about both DA experiments and retaining only the more specific description of the slightly narrower probability core in JEDI-New.

- *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.*

Reply: inserted

- *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.*

Reply: We appreciate this clarification.

- *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.*

Reply: We appreciate this additional interpretation.

(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 Ze and where the 1018 comes from.

Response: We have revised the text surrounding Eq. (3) to explicitly state that Z_e is expressed in units of $mm^6\ m^{-3}$, and that the factor 10^{18} arises from converting drop diameters from meters to millimeters in the Rayleigh-scattering formulation.

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

Response: These terms are temperature-dependent coefficients in the moment-based power-law relation for dry snow from Field et al. (2005). We have revised the manuscript to clarify this point.

(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 °C and < 0 °C

Response: We have revised Eq. (7) into a single expression with temperature-dependent parameters and clarified their values for $T > 0°\mathrm{C}$ and $T \leq 0°\mathrm{C}$ in the text.

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

Response: The "New" data points are plotted as open red circles, matching the legend. The apparent discrepancy is likely caused by heavy point overlap in dense regions of the scatterplot, which visually obscures the open-circle style. We therefore did not make a substantive change to the marker definition.

(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?

Response: We thank the reviewer for this clarification. The data fitting was performed against model outputs from the snapshot of 2-hr WRF forecast (3 km grid spacing, valid at 20 UTC on May 20, 2019; centered on 34.32, -98.01), which differs from the case analyzed in the main experiments. We would like to emphasize, however, that the specific meteorological case used for the fitting is not material to the result. The fit captures the mapping between the model's hydrometeor fields and reflectivity, which is determined by the scheme's microphysical formulation and assumptions — both of which are independent of the weather regime. Any case that provides a sufficient range of hydrometeor values would yield an equivalent fit. The case used later in the manuscript is therefore an independent application of the same operator and is not affected by the choice of fitting sample.

(7) Language improvements

Response: We thank the reviewer for these helpful language and clarity suggestions. All of the points listed below have been addressed in the revised manuscript.

- *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.*