Review of "*Assessment of the data assimilation framework for the Rapid Refresh Forecast System v0.1 and impacts on forecasts of a convective storm case study***"**

The authors use a relatively new modeling system to produce a set of forecasts of a convective-scale event. Sensitivity experiments are performed to assess the impact of different data assimilation choices with standard verification metrics. The authors do a nice job of explaining the new system, and providing justification for their choices, as well as incorporating a diversity of verification approaches. My major concerns with the study are the overlap with prior work (e.g., Tong et al. (2020) used similar model and DA systems) and the lack of additional cases for analysis. I've provided some minor comments below that I believe need to be addressed before recommending acceptance.

Minor comments

- I suggest the authors provide more detail about how this study differs compared to Tong et al. (2020), especially the model configuration and design choices. There are a lot of similarities, including the use of FV3 (termed the FV3-SAR in that study, which I believe is the same model that is the core component of the UFS-SRW), variational and hybrid DA with GSI, similar physics choices, and a similar convective-storm case study approach (although for a different case). Can the authors describe how that work ties in with the current set of experiments?

- The RRFS will be an ensemble-based system, so generating and verifying ensembles seems like a good choice to assess the benefits of the various approaches. It may be useful to clarify why the authors only performed deterministic forecasts somewhere in the text (sorry if it's there and I missed it!).

- I really think this study would benefit from additional cases, especially when the authors argue at many points in the paper they are using the results to guide future configuration decisions. Some of the differences between the experiments seem very small, and may become more evident with a larger sample size. This could be considered a "fatal flaw" by some, but I think there's some merit in providing documentation of ongoing work leading up to the implementation of the future RRFS system in the form of this manuscript.

Specific comments

Lines 40-42: This sentence implies that the addition of the WaveWatch model into the operational forecast somehow improved the low-level cold temperature bias observed in a prior version of the GFS. I don't think that's possible and I don't think the change notice referenced supports that claim. Please revise.

Line 87-91: I recommend removing these sentences. The number of studies that examine data assimilation for convection-allowing applications is too numerous to mention here, so describing these two specific studies is necessary, unless they are aspects of the work that are especially relevant to the current work.

Line 104: The authors should make clear that the eventual RRFS implementation will produce ensemble forecasts and not just a single deterministic forecast.

Line 153-161: I suggest moving the list of these parameterizations into a Table that can be referenced in the future, including names of schemes and associated studies that describe each scheme.

Section 2.6: What do the authors mean by “workflow”? As written, the term is used rather generically, but I’m guessing that there is specific workflow software that is used that should be described in the text (this may be described later in the text, but the authors should bring this up earlier).

Line 260-262: How were the MLCAPE and shear diagnostics computed? The text states they were “observed”, but there are no routine soundings typically available between 19-20 UTC in northeastern Oklahoma. If these values are from a model analysis, that should be stated (e.g., “The RAP analysis contained MLCAPE values of…).

Line 367: Are Oklahoma Mesonet observations assimilated?

Line 543: Is model level 50 really located around 850 mb? Does that mean that there are 50 levels below 850 mb and only 14 levels above 850 mb (64 levels total)?

Lines 554-556: How are the MMI values so different at 21 UTC in CLIPSAT and 75EnBEC? To my eye, the figures look almost identical. The differences at 23 UTC look more significant, but the MMI values are more similar at this time. Can the authors explain why this is the case?

Figure A1: Why is this included as figure A1 and not Figure 11?