Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC3
https://doi.org/10.5194/gmd-2021-289-RC3, 2021

Comment on gmd-2021-289

Anonymous Referee #3

Referee comment on "Assessment of the data assimilation framework for the Rapid Refresh Forecast System v0.1 and impacts on forecasts of a convective storm case study" by Ivette H. Banos et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-289-RC3, 2021

General comments:

I appreciate the extended introduction and literature review of the state of the modeling and DA development for this application. The manuscript provides a useful snapshot with regards to the state of testing and evaluation of a case study using the RRFS during its development (v0.1).

For the benefit of the general data assimilation audience, some detail would be helpful to describe why the unconventional data assimilation cycle is used for the RRFS. For example, why is a cold start applied repeatedly throughout the DA cycles? Why can't the RRFS be run as a continuous cycled process like conventional DA applications? This could perhaps be connected with an improved discussion of the FV3 LAM and its limitations over other possible regional modeling approaches.

Please be consistent with the tense throughout. Present tense is appropriate, but occasionally it slips into the use of past tense.

Specific comments:

L 4-5:

"The current data assimilation component uses the Gridpoint Statistical Interpolation (GSI) system."

It would be helpful to mention here what DA method is being used. Is it 3D-Var?

L 6:

"Results show that a baseline RRFS run without data assimilation is able to represent the observed convection, but with stronger cells and large location errors."

How does the RRFS represent observed convection without data assimilation? Is it through the boundary/initial conditions coming from the global model? In that case, it would be using data assimilation indirectly through the global analysis. Could the authors please clarify.

L 8-9:

"using 75 % of the ensemble background error covariance (BEC)"

What does it mean to use only 75% of the BEC?

L 9-10:

"Decreasing the vertical ensemble localization radius in the first 10 layers of the hybrid analysis results in overall less skillful forecasts."

From what initial radius to what final radius?

Please change to:

"Decreasing the vertical ensemble localization radius [from X m to Y m] in the first 10 layers of the hybrid analysis results in overall less skillful forecasts."

L 97-99:

"Using hybrid 3DEnVar with 75 % of the ensemble background error covariance (BEC) showed storm structures in the 2 h forecast comparable to when using ensemble Kalman

(EnKF), although EnKF outperformed 3DEnVar in the first hour forecast."

(1) Does this mean that the BEC is weighted 75% toward the dynamic ensemble-estimated BEC and 25% to the static climatological BEC?

(2) If the EnKF outperforms the 3DEnVar in the first hour, and then they are comparable in the second hour, then when not use the EnKF instead of the 3DEnVar?

L 99-100:

"Both methods showed higher equitable threat scores (ETS) when compared to 3DVar and pure ensemble during the 4 h forecast analyzed."

Does "both methods" refer to the3DEnVar and the EnKF? What is a "pure ensemble"?

L 126-127:

"The FV3, originally a global model, features three types of local refinement capabilities:

stretching of the global grid (Harris et al., 2016), nesting within the global grid (Harris and Lin, 2013), and a LAM capability (Black et al., 2021).”

It would be useful to mention briefly how these types of local refinement differ.

L 143-144:

“Hence, the CCPP contains a set of physical schemes and a common framework that facilitates the interaction between the physics and a numerical model (Bernardet et al., 2020). ”

Perhaps it would be more clear to say “between the physics parameterizations and the dynamical core”.

L 166:

“3D[V]ar”

L 171:

"3DEn[V]ar"

L 177-178:

Some discussion should be given about what deficiencies this will have. For example, the lower-resolution global ensemble members (which are even lower resolution than the global deterministic forecast) may have significant biases, and will not resolve the error characteristics at the scale of the LAM. Cleary, the global ensemble statistics provides some useful information, but it is not ideal.

L 223:

"cold starts are performed every 12 hours and warm starts are performed at all other cycles using the 1 h forecast from the previous cycle as background for the analysis."

Why is the DA continually reset with cold and warm starts? What prevents the standard self-contained forecast-analysis cycle? Is there a model drift when the DA is run continually in the RRFS?

Figure 2:

It would help to have more annotation here. Where are the 18h forecasts coming from?

L 259:

“aircrafts”

Change to:

“aircraft”

L 263-264:

"The time window used is 1 hour, allowing for observations within 30 minutes before to 30 minutes after the analysis time to be assimilated."

I suppose this implies that the analysis time is at the middle of the 1-hour forecast window, but I didn't notice this mentioned earlier.

L 265-267:

" For the hybrid 3DEnVar analysis, the Global Data Assimilation System (GDAS) 80 member ensemble forecasts (9 h forecasts) are used to provide the ensemble BEC (e.g. Wu et al., 2017)."

Please mention the resolution of these ensemble members. How well do the lower resolution members resolve dynamics at the scale within the 3km resolution LAM? E.g. how many grid points of the low-res FV3 global ensemble member fall within the LAM region?

L 268-269:

"For example, the 9 h GDAS ensemble forecasts initialized at 00:00 UTC (valid at 09:00

UTC) are used for the cycles from 07:00 UTC to 12:00 UTC.”

How is the 9-hour forecast initialized at 0 UTC used at 12 UTC? Is this using the BEC fixed in time at 9 UTC for the entire window?

L 271-272:

“ In all experiments with data assimilation, two outer loops with 50 inner loops each are performed to minimize the cost function and find each analysis.”

I’m not sure I understand what is done in the outer and inner loops. Are the inner loops referring to the PCG solver? If so, then was is done in the outer loop?

L 274-275:

“for different [applications].”

L 277-279:

"This baseline experiment is called NoDA and uses the same cycling configuration as experiments with data assimilation but without the execution of GSI."

I understand that from your perspective this doesn't use the GSI, but for the entire procedure the GSI is used at the global scale. I think it would be helpful to explain this context in more detail.

L 281:

"experiments with different ensemble weights [are] conducted"

L 335:

"The analysis residuals (OmA) are also depicted in Fig. 4"

The OmA's are less useful than the OmF's for assessing the DA performance. The analysis can be drawn arbitrarily close to the observations (e.g. using complete replacement). It is more valuable to see that the forecasts are being drawn closer to the future observations.

L 343:

"while some MESONET observations have large analysis residuals."

What is the expected cause of this larger discrepancy in these observations?

L 350:

"This means the analyses fit closer to the observations, which is expected from a correctly executed data assimilation procedure."

This is partially true, but it is also not correct to fit the observations perfectly (due to

observational error), so care should be taken in such a statement.

L 351-353:

"There is a noticeable jump in the RMS error values of the OmB from 00:00 UTC (12:00 UTC) to 01:00 UTC (13:00 UTC) on 4 May 2020. This is because 00:00 UTC and 12:00 UTC are cold started from HRRR analyses. "

So why are you using the cold starts?

Figure 5:

It would be useful to add a panel showing the difference between 3dvar and the 75EnVar.

L 365:

"the EnKF filter method."

Change to:

"the EnKF method."

L 377:

"Figure 7 shows the 2, 4, and 6 h forecasts of …"

The analysis of these results seems very subjective - could the authors please provide some general statistics for each experiment result to help in the comparison, and provide more objectivity. Below, for example, I suggest adding RMSE statistics to each sub-plot in Figure 7.

Figure 6:

Where does the error covariance matrix come from for the 3D-Var? How was it computed?

L 405-407:

"The wind RMSE results do not clearly indicate which experiment is best, but in general 100EnBEC shows the lowest values when considering all vertical levels."

The hybrid methods generally underperform when the static BEC is inadequate. How is the BEC computed for these experiments? This may indicate that tuning is necessary for the B matrix in the hybrid, e.g. see:

Chang et al., 2020:

https://journals.ametsoc.org/view/journals/mwre/148/6/mwrD190128.xml

Also, the online estimation of the hybrid weighting parameter has been explored by De Azevedo et al. 2020, which may be worth mentioning:

https://www.tandfonline.com/doi/full/10.1080/16000870.2020.1835310

Figure 7:

It would be helpful if each one of these plots had an RMSE value appended (e.g. in the lower left corner) to make it easier to compare the methods.

Figure 8:

The Green RMSE is difficult to read. A slightly darker color would help. Also, perhaps you could order the RMSE in each plot from lowest to highest so that is it easier to see how each method performs in comparison to the others.

L 430:

"this study looked at"

Change to:

"this study looks at"

Figure 8 and 10:

I’m not sure that I see the value of the confidence interval shading. However are these confidence intervals computed? Can the authors justify that this statistic is meaningful for this application? (e.g., the confidence interval implies that this specific case can be extrapolated to all other relevant storm instances - perhaps more discussion of the computation and application of this method would be warranted.)

L 450-451:

“The impact of adding PBL pseudo-observations to the analysis based on surface temperature and moisture observations is evaluated in experiment PSEUDO.”

I may have missed it, but I don’t see the source of the pseudo-observations. What measurements are being converted to these pseudo-ops? Is there a reason the source measurements cannot be assimilating using an appropriate observation operator?

L 477:

"More tuning and testing of this function are needed before applying this technique in the RRFS."

The results (e.g. in Figure 12) seem to indicate the use of pseudo-observations may be a bad idea. Is there any reason why continued effort should be made to do this?

Figure 13:

Based on Figure 12, this seems to be over-fitting the data, which is then causing problems throughout the model column. It seems like this result would be better achieved via post processing, so that it doesn't have negative ripple effects throughout the cycled DA.

Figure 14:

Please state what the units and physical height correspond to model level 50.

L 559:

"Supersaturation clipping in GSI can improve specific humidity fields in the analyses, allowing for more realistic storm and precipitation forecasts at longer forecast lengths. At shorter forecast lead hours, it produces more spurious convection"

This sounds like the wrong modification is being made to correct a longer-term bias. It would be beneficial to track down the root cause that accumulates to cause the longer-term bias without degrading the short-term skill.

L 572:

"More extensive testing of RRFS, covering a wider variety of cases, larger domain, and longer period of time, is needed to demonstrate whether results found here are robust or may be case dependent."

I agree, which is why I’m confused by the presentation of the confidence intervals. Could the authors either remove these in the plots above or describe in greater detail how they were computed and why they are relevant in this case.