To Reviewer 1

This work tries to estimate components of the observation error for radar reflectivity and radial wind measurements using two complementary techniques: the difference between high and low resolution forecasts as an indicator of representation error, and the well-known Desroziers approach that is intended to estimate the total observation error. There are substantial differences between the two estimates, which the paper attempts to explain. This is useful work on the tricky subject of assigning observation errors to precipitation-sensitive observations.

Answer: Thank you very much for your kind acknowledgment of our work.

However, there are major issues with the work as currently presented. The first is an important scientific issue: the results are given for a precipitation-affected sample based on a 5 dBZ reflectivity threshold; however little attention has been paid to how that threshold has been defined and how the definition might affect the sample of observations being examined. Given that the threshold definition must be different in the two different techniques, this might be a major cause of variation between the two. Second, there is a lack of high-level figures to help synthesise the results; some of the figures that are presented are given at a perhaps excessive level of detail. Hence, major revisions are recommended.

Answer: In the revision of the manuscript we hopefully clarified the reasons behind use of a threshold (see also answer to question 1 below). We also added figures that show results for two threshold values. In addition we added a new section (4.3) that focuses on comparison between the two methods of computing observation error statistics.

Major issues

1) Sample creation using the 5 dBZ threshold.

A first issue is that the precise application of the 5 dbZ threshold to identify precipitation is not described in detail. For the representation error calculation, the threshold described on line 128 may be applied to the low-resolution data, the high-resolution data, or both. Hence without further details, this is ambiguous. For the Desroziers calculation, the threshold may be applied to the observations or to the model simulations, or to both; again this is not specified. It is also not fully clear if the threshold is applied per observation location, and how the threshold relates to the ensemble if one is used (is it the ensemble mean, control or ensemble members being used?)

In the literature on all-sky passive microwave assimilation, it is well-recognised that sampling issues need to be treated with care. Given that the location of precipitation in the forecast and the observation may be different, a ”precipitation” sample based on the model precipitation alone will exclude many locations where the observations have precipitation but the model does not, and vice-versa. Depending on exactly how the thresholds are applied, very different bias and standard deviation characteristics may be observed. See for example:

"Observation errors in all-sky data assimilation", Geer and Bauer, 2011, https://doi.org/10.1002/qj.830

"Assessing the impact of pre-GPM microwave precipitation observations in the Goddard WRF ensemble data assimilation system", Chambon et al., 2014, https://doi.org/10.1002/qj.2215

Answer: The reflectivity Z has initially units of $[mm^6/m^3]$, however, because numerical values of Z may span several orders of magnitude, it is convenient to use a logarithmic scale in practice, defined as units of dBZ $= 10\log_{10}\left[\frac{Z}{1mm^6/m^3}\right]$. For instance, $10^5\ mm^6/m^3 = 50$ dBZ; $1\ mm^6/m^3 = 0$ dBZ; $10^{-5}\ mm^6/m^3 = -50$ dBZ, and $0\ mm^6/m^3 = -\infty$ dBZ, for which -99.99 dBZ is used to represent $0\ mm^6/m^3$ in the radar forward operator. It is noticed that for very small reflectivities, their differences in units of mm^6/m^3 are trivial but significant in units of dBZ. This can be problematic if assimilating those data, which could lead to unrealistically large increments and spurious convection (Zeng et al. 2021a). Therefore, it is very well established in radar data assimilation community to set a threshold value for small reflectivities, in the operational setup of KENDA, it is 0 dBZ, which means all reflectivities values lower than 0 dBZ are set to 0 dBZ, and we call 0 dBZ data as "no reflectivity data". The same threshold value is set to all observations and to all simulated reflectivities in each ensemble member. Second, due to setting the same threshold value to both observations and backgrounds, the innovations are reduced and the observation error variances are underestimated when computing Desroziers diagnostics (Zeng et al. 2021a). To have better statistics that are not affected by the operation of setting no reflectivity data, we calculate Desroziers diagnostics for reflectivities with positive values, for which 5 dBZ is chosen. We made those points more clear in the text. In addition, we added in conclusion section "since reflectivity data and no reflectivity data are associated with different error characteristics as for example the all-sky radiances (error standard deviations in clear sky are much smaller than in heavy cloud or precipitation, Geer and Bauer 2011; Chambon et al. 2014), one could consider treating them as different data during data assimilation".

In the current manuscript, a particularly striking difference is seen between the samples used in the representation error and Desroziers studies (Fig. 4c and 12c) at higher altitudes. Simulated reflectivity reaches 20 dBZ at 10km in the representation error sample but is just 12 dBZ in the observations in the Desroziers sample. This suggests that the sample at high altitudes in the representation error study is dominated by infrequent deep convection, since this is likely the only thing that can generate greater than 5 dBZ reflectivity at that altitude. The big difference to the observational sample could be explained by model error, but it could also just come from a major difference in the composition of the sample being analysed. These aspects need more attention.

Answer: High reflectivities in simulations at higher altitudes are probably caused by the inappropriate value for the slope intercept parameter N_0 in the particle size distribution function of the one-moment microphysical scheme. N_0 is empirically estimated by the surface measurements, which may be too small for anvil clouds at higher altitudes, leading too much large graupel over there and overestimated representation error. We added this into the text.

2) Need for higher-level figures to summarise the results.

Figures 6, 8, 13 and 15 give a possibly excessive level of detail. The text has to do a lot of description and further analysis of these figures. It presents many lists of numbers derived from these plots, such as the correlation length scales (see e.g. lines 159 - 164). The results derived from these figures would be better presented and analysed on higher level figures, ideally comparing the representation error study with the Desroziers study on the same figures. This would reduce the need to present long lists of numbers in the text.

Similarly, the most interesting aspects of the study are the comparisons between figures 4 and 12, and Fig 7. versus Fig. 14. It is somewhat inconvenient to have to compare these figures manually. A summary figure combining some of the lines from both could be useful.

Answer: In the revision of the manuscript, we moved some of the previous figure panels to Appendix, and added a new section that compares the methods. To this end, we also added Fig. 13 as the summary figure.

Minor issues

1) Line 22: the acronym ICON-D2 is not explained, nor its significance (presumably the first application of radar reflectivity assimilation in this framework?)

Answer: We rephrased as "The ICON-LAM (ICON-Limited Area Model) is the limited area version of the ICON model and is to replace the COSMO model in the operational forecasting system. The ICON-D2 (D: Deutshcland (Germany); 2: 2 km) is an ICON-LAM setup at approximately 2 km grid spacing, which is restricted to Germany and the neighboring countries and became operational for very short-range forecasting since February 2021". We also emphasized the significance of this study in the text.

2) Line 65: The $\mathcal{H}$ operators in equation 1 need some more explanation. Clearly they are not identical since the model inputs are on different grids. Any interpolation or coarse-graining within these operators needs to be explained. This is particularly important since the observation operator is nonlinear. Hence, the application of $\mathcal{H}$ to the mean of the model forecast may be strongly different to the mean of $\mathcal{H}$ applied to individual locations in the forecast. This maps onto the well-known beam filling effect

Answer: Reflectivities are first calculated on the model grid points and then interpolated onto radar coordinates. For radial wind, three wind components are interpolated onto radar coordinates and then radial winds are calculated.

3) Line 73: Although it's partly explained later on, it would be useful to have some words on how the model states x_a and x_b used in equation 2 relate to the model states defined in equation 1, given broadly as $M(x^T)$. Even better, homogenising the notation between these equations 1 and 2 would help define the precise differences in methodology between the two halves of this work (such as highlighting the resolution and / or model differences involved, and/or differences in the observation operators being used).

Answer: We added "In the following, we estimate statistics of the RE by using the method from Section 2.1 and statistics of the OE by applying the Desroziers method to an data assimilation experiment with a low resolution model, i.e., $\mathbf{d}_{o-b} = \mathbf{y^o} - \mathcal{H}(\mathbf{x}_b^L)$ and $\mathbf{d}_{o-a} = \mathbf{y^o} - \mathcal{H}(\mathbf{x}_a^L)$".

4) Line 76-77: "R_{est} is optimal in case of ..." - are there any further references to back this up or is it from Desroziers (2005)?

Answer: we added Reichle et al. 2002.

5) Figure 3, legend: "Scratch of..." is odd terminology - change the term or explain it.

Answer: We changed "Scratch" to "Illustration".

6) All figures in the manuscript, but particularly Figs. 3 and 4: the point markers (such as a square or a circle) are so frequently sampled that they change the width of the lines, making them very thick in places and making it hard to do detailed comparisons between the lines. Consider showing all these graphs with only lines, not lines and symbols.

Answer: Done.

7) Figures 2 and 3 are not linked to the text where they appear. In any case there needs to be some early description of the processes of superobbing, and the details of the observation operator, in section 3. Instead these details appear partially, and too late, in section 4.2.1, for example.

Answer: We switched the order of Figures 2 and 3. We added more details about the observation operator in section 3. Since the processes of superobbing is only done for data experiments for OE (not for RE), it is appropriately positioned in section 4.2.1.

8) Furthermore, there are some other introductory details missing on the model setups: for example for the models used in the representation error study, it is not clear whether data assimilation is applied to keep the forecast on track, whether (and where from) there are boundary conditions being applied to achieve the same result. An introduction to the models being used in both halves of the study, their similarities and differences, would be very useful around section 3.

Answer: For representation error, we are interested in its climatology instead of exact position and intensity of convection, therefore, no data assimilation is applied to the models used in the representation error study. But in both studies, models are driven by hourly boundary conditions.

9) Line 107: "standard deviation and horizontal correlation" ... of what?

Answer: We changed it to "standard deviation and horizontal correlation of OE and RE"

10) Line 113: "Around 10^3 ..." surely the authors mean "at 10^3 or below"?

Answer: We rephrased it.

11) Line 117: The training period concentrates on heavy thunderstorms. Somewhere, the authors should discuss the applicability of their results to other periods, such as wintertime cyclonic systems.

Answer: We added in the conclusion "Results presented here are based on the convective period in the summertime. The applicability of those results to other periods such as wintertime cyclonic systems needs further investigation. However, some studies have been done by the other centers. For instance for the Met Office UKV model with the 3D-VAR scheme, the estimated OE statistics (based on Desroziers' method) for radial wind are qualitatively similar to those in the summertime (Waller et al. 2016c), and for reflectivity, Kouroupaki 2019 shows that the estimated standard deviations of the OE in Winter are larger than in Summer and that they increase with reflectivity values".

12) Line 131-133: "It is noticed that the variations of standard deviations are very comparable to the simulated reflectivities of the (high resolution) model run in Fig. 4c, indicating a systematic error that is proportional to the true value". This does not make sense: Figure 4a and 4c do not look very similar, there is no strong similarity between the two. Also it is not supported why this should be explained by a systematic error.

Answer: We rephrased the text as "Standard deviations increase in the first few hundred meters and then slightly decrease for the next few hundred meters before increase to a local 160 maximum at around 3 km. Above 3 km, standard deviations decrease till 5 km and then increase to the top. The variations of simulated reflectivities of the model run exhibit a similar pattern although the decrease between 2 and 6 km is sharper. Overall, it can be said that standard deviations are approximately proportional to observed values."

13) Line 133-134: Comparison of Figure 4 to Figure 5 is not so helpful because the former is based on the < 5 dBZ sample, and the latter is based on all data.

Answer: We added Figures for reflectivity data ≥ 0 dBZ.

14) Line 208: "too big" - define?

Answer: We rephrased text as "a first guess check is carried out (i.e., the innovation of the deterministic run must be smaller than three times of the standard deviation of the innovation)".

15) Line 312-314: "the model produces too much ice" - this does not appear to have been much discussed or supported in the preceding text.

Answer: Thank you for noticing. It was a mistake, actually we meant graupel instead of ice. It is mentioned in Line 299-301.

16) Line 314-330: "this is mainly due to the application of scaling factor...". Unless I missed it, this scaling factor has not already been discussed in the text, and its effect on the Desroziers-based observation error estimate has not been established.

Answer: The explanation of the scaling factor is given in Line 230-233 and its effects are mentioned in Line 270-271.