Overall assessment

The manuscript EGUsphere-2024-2876 proposes to inflate the covariance between radar reflectivity and specific humidity and reduce the declared error on radar measurements under specific conditions to ease the assimilation of observed storms that are not simulated in ensembles. The idea has merit, the work seems to have been executed competently, and the results are reasonably well described. The manuscript fails on one critical point that forces me to recommend rejection at this stage: The work, as described, is irreproducible. And because the basis of science dissemination is to provide enough information for others to reproduce the experiment described if they wish to, I judge that the manuscript cannot be published in its current form. But I believe that, with some relatively small efforts, the problem could be remediated.

Justification and major comment

A key to this work is the establishment of the relationship between radar reflectivity and specific humidity that was subsequently used for Targeted Covariance Inflation (TCI). How it was done remains unclear to me. The only information I could find is in L201-L204 (emphasis mine): "*Initially, <u>a raw data set</u> is constructed which contains all <u>relevant</u> simulated values of an ICON-D2 assimilation cycle with 40 ensemble members running for more than two weeks. Subsequently, a <u>particular filter</u> is applied to this raw data set such that only data representative for early-stage convective events, i.e., only spatial and temporal points in the direct vicinity of newly emerging convective cells are included.*". Nowhere is it explained from what or how the raw data set was constructed, what are the characteristics that make some of the simulated values "*relevant*", nor what is the "*particular filter*" in question. Alternatively, it is never clear what part of the description applies to the process used to initially select the data to establish the δZ- δq_v relationship, and the what part applies to the algorithm used to find the pixels on which TCI was applied. If the same approach, and possibly the same data, were used for both model creation and TCI application, this was never clearly stated, even though model creation and TCI application must have been two clearly distinct and sequential steps. Yet L58, stressing that the model "*has been trained on data exclusively found in the nearest spatio-temporal vicinity of early-stage convective events*", suggests that the two are different as the grid-point selection for TCI (13)-(17) does not appear to have any temporal check. And then we have unspecified "*internal details of the data selection process and the associated algorithm for the automatic detection of early-stage convective events*" (L208-209) that seem to constrain key aspects of model creation and algorithm development. This reader is left with the impression that, intentionally or not, the authors are trying to hide key information about the research undertaken, undermining its value. And I judged that this was unacceptable in a scientific publication.

The solution, therefore, is clear: The description of how the model was established must be more thorough and less opaque, not fearing to openly admit limitations that may have made this work not as ideal as the authors would have wished.

Other specific comments

1) A well-known key to make convection initiation possible in a numerical model is the presence of sufficient humidity to create convectively unstable conditions. Many researchers simulating convective storms have forced high humidity in regions where radar echoes are observed (as early as Lin et al. (1993)). In many ways, your work tries to do with TCI and taking advantage of covariances what others

have done: Saturate regions with echoes to allow convective motions to occur. I believe the introduction should include some more recognition of earlier efforts that were not undertaken in a context of ensemble forecasting is the sole focus of your introduction.

Lin, Y., P.S. Ray, and K.W. Johnson, 1993: Initialization of a modeled convective storm using Doppler radar derived fields. *Mon. Wea. Rev.*, **121**, 2757–2775.

2) L56-62: What do you define as early-stage?

3) L66: Can you briefly expand what you mean by "*accumulation effects*", and/or what you fear about them?

4) L135-144: This information does not appear very relevant to the work, and I believe it could be cut. A simple reference to Prill et al. (2024) or other relevant work seems sufficient to me.

5) The same applies to the latent heat nudging section 2.4. A simple reference in the fractional skill score discussion (4.3.2) would be enough.

6) Are you assuming a linear relationship between increments δZ and δq_v as stated by (10) or between values of Z and q_v as stated in the text on L189-190? I believe it is the former.

7) Aren't you concerned about assimilating radar observations from 3-4 km altitude where radar bright bands from melting hydrometeors could affect reflectivity estimates and make reflectivity simulation more difficult, including in convective storms?

8) L279-282 seems superfluous. Consider cutting them.

9) Fig. 4 and subsequent radar images: If your algorithm uses reflectivity from 3-4 km altitude, and since you stress that the algorithm is height-based and not elevation-based (L195), why plot reflectivity at a specific elevation angle instead of at a specific height level?

10) L333-336, on explaining why the simulated reflectivities of the TCI run are smaller than the observed ones: Two other explanations could include that 1) TCI was only applied over a small height range (3-4 km), limiting the region being convectively destabilized, and 2) the sluggish 2.1-km resolution model starts its convection well after it occurred in reality and hence generally cannot evolve as quickly.

11) Having positive skill when assimilating reflectivity in convection is challenging. I was hence puzzled how you could achieve such increases in skill scores (abstract, L363, Fig. 9…) by making changes over very limited areas (e.g., Fig. 6) resulting in visually modest changes (Fig. 7). I had to relook at and fully understood the very generous FSS formula (19) to figure out why: You simply test for the fraction of pixels exceeding a reflectivity threshold in boxes of a given size to declare success. After much thinking, I decided this is fair; but I believe that you should specifically mention that verification was made (L341-342) "by employing a special version of the fractional skill score (FSS) designed by Roberts and Lean (2008) to deal with highly structured fields such as reflectivity that are particularly susceptible to double penalty (Rossa et al. 2008)" (changes underlined)

Rossa, A., P. Nurmi, and E. Ebert, 2008: Overview of methods for the verification of quantitative precipitation forecasts. In: Michaelides, S. (eds) *Precipitation: Advances in Measurement, Estimation and Prediction*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77655-0_16

12) If (L373-L375) "*the negative impact of the TCI on the mean error of the TEMP relative humidity w.r.t. both the analysis and the first guess […] can be interpreted as the TCI introducing additional humidity into the simulation at those heights*", why are RH lower for the TCI experiments (top center of Fig. 8)? No attempt was made at explaining or even acknowledging the unexpected nature of this result.

13) L415-423: I believe that you are underselling the skill of your technique and being overly defensive regarding the results of Fig. 9: Why say it is neutral at 46 dBZ when it is not? Overall, the results seem as good as with the lower thresholds except that the skill evaluation is noisy because such echoes are rarer (note the very different scale of the FSS relative improvement).

Technical corrections

i) L51-52: "*The studies Yokota et al. (2018); Dowell and Wicker (2009); Vobig et al. (2021) could show that adding spread…*" sounds/looks funny. Do you mean "*Studies by Yokota et al. (2018), Dowell and Wicker (2009), and Vobig et al. (2021) suggest that adding spread…*"?

ii) L88-90 is a repeat of 10 lines above; consider cutting.

iii) Please improve the quality of Fig. 2, especially the right plot.

iv) (13)-(18) Radar purists will insist that the differences between two numbers in dBZ units as well as the standard deviations of reflectivities in dBZ have units of dB or of dB(*Z*), the latter being more specific. This comment also applies to L251, L253, and Fig. 8.

v) L253: "*… the system is significantly stronger pulled*" should be "*the system is pulled significantly stronger*".

vi) L275: For improved clarity, add a comma between "*implementation*" and "*the calculation*".

vii, and last) The sentence on L331-L332 needs to be edited: "At this point it is important to note that fig. 7 represents the general trend, that may be observed for other assimilation dates and lead times not shown here, very well". The first underlined comma should be cut. It is also unclear to me what you are trying to say in the second underlined section; I'll let you decide how to correct it to be clearer.