Thank you for taking your time to write the detailed review.

Main points:

- **Comment**: This looks to me as if this study has the characteristics of an ML-based retrieval, where to properties being retrieved from the SEVIRI radiances are the vertical profile of radar reflectivity. This has some similarities to other algorithms, such as the GPROF retrieval (Kummerow et al, J. App. Met., 2001), which could be discussed. Perhaps the method in this paper could produce improved precipitation retrievals in a similar style to the GPROF retrieval (although not in this publication)?
- **Answer**: That is an interesting suggestion worth to be investigated. We see the similarity to reflectivity-based methods for precipitation retrieval and thought about this aspect before. A detailed analysis seemed out of scope for this paper. Nevertheless, we think this would be an interesting follow-up.
- **Comment**: Although it doesn't seem to be at the level where I could be comfortable using the retrieved reflectivity profiles yet, but perhaps there are certain conditions where the data could be more trustworthy? Would limiting the study to a smaller area (e.g. close to nadir for SEVIRI) produce better results?
- **Answer**: We agree with you on this issue. In our paper, we state the current problems with the method and its limits in terms of accuracy (Section 4). Resampling the satellite data on a spatial grid with geographic coordinates affects the accuracy of the input data with greater distance to the equator (Section 2.1.1). We think limiting the area to e.g. the tropics has the potential to improve the results. In the revised manuscript, we add a visualization of the zonal error to display the geographic differences.

Line by Line comments:

L9 -**Comment**: Bias in retrievals based on approximations and cloud physical properties L9 - It is later stated that RMSE varies significantly by cloud situation - how is this value derived? **Answer**: We did not classify the cloud type, but we investigated the RMSE on the test set with 1500 samples. We rephrase that sentence for clarity.

L10 -

- **Comment**: receive high accordance -> find high agreement?
- **Answer**: Thank your that remark, will be changed.

L30 -

- **Comment**: at risk of bias. I would argue this is almost always an issue (and would be for a ML method too)
- **Answer**: You are right, no method is probably completely free of bias. We will rephrase that sentence.

L30 -

- **Comment**: If passive sensors lack this information, how can it be recovered from this method?

- **Answer**: Information on the cloud column below the cloud top can be approximated using the satellite channels at different wavelengths (lines 27-29). While they enable an approximation, their information density is lower compared to the radar reflectivity received from active radar. The hypothesis is that the information on the vertical resolution is hidden within the satellite channels. While previous research reached their limits, we assume that these representations can be learned from a neural network better than by human analysis. We rephrase that paragraph for clarity.

L33 -

- **Comment**: joined -> combined
- **Answer**: Changed to: "A combined use of different instruments to derive comprehensive 3D structures [...]".

L88 -

- **Comment**: The near IR channels also include reflected solar radiation (depending on channel wavelength and time of day)
- **Answer**: We rephrase that sentence to clarify the channel differences.

L109 -

- **Comment**: How is parallax addressed? SEVIRI has an off-nadir view for most points on the disc, while CloudSat always views at nadir. This could cause registration issues between different cloud layers. Changes in resolution might also become an issue closer to the edge of the disc.
- **Answer**: We recognize the importance of the parallax effect for the correction of the geometry, especially when dealing with multiple cloud layers. In this work, we estimate only a small impact on the overall results. First, we assume that neighbouring pixels are more similar to each other than those in far distance. Second, we face the need to downsample the resolution of the CloudSat data to match the MSG SEVIRI resolution. That is why we don't expect the estimated differences of up to 3 pixels to substantially impact the predicted reflectivities. Overall, we estimate the error resulting from geometric inaccuracy to be lower than current model limitations. We will add a remark in the manuscript.

L110 -

- **Comment**: I found the last sentence difficult to follow. What does the 'factor to coarse grain the data' mean here?
- **Answer**: The resolution mismatch between CloudSat and MSG SEVIRI requires an aggregation of CloudSat pixels. We use the maximum reflectivity to reduce the resolution of the radar data. This blurs the sharp contrasts within the radar cross sections (as seen in Figure 4, CloudSat CPR cross sections are blurred compared to the original data) . We rephrase that sentence for clarity.

Fig. 2 -

- **Comment**: Why not low cloud layers? This is explained later on, but could have been included in the description of the radar data (unless I missed it?)
- **Answer**: As stated in lines 123-125, we analyzed the CloudSat quality flag to identify height levels that were affected by noise. This applies especially to the near-ground height levels influenced by the topography and connected radar attenuation. Those height levels were excluded from the model framework. The reduction affects the representation of low clouds

and the cloud base, reducing the model performance in these altitudes. We add an extended explanation in this section in the revised manuscript.

Fig. 3 -

- **Comment**: How is this normalisation done? Is it across the whole image?
- **Answer**: For the joint plot, we computed the value distribution for observed and predicted reflectivities on the test set. The normalisation is done across the whole image using the respective distributions.

L220 -

- **Comment**: Why is the DL model different here? Is this due to a regression-to-the-mean effect for the Res-UNet?
- **Answer**: After revising, we think this was due to an error in the code. The results could differ as a result of the regression-to-the-mean. Since the DL network is fed a lot of nearly cloudfree images, they can lead to a shift of the distribution. In contrast, extreme values have a higher influence on the OLS and RF. This can affect the prediction of positive values. Nevertheless, we change the plot to all models pointing towards the same direction.

L262 -

- **Comment**: I think thin clouds are also difficult to detect from CloudSat, given they might have a small radar reflectivity. Both SEVIRI and CloudSat on their own are able to produce this second, higher peak in cloud top height, suggesting they can detect these thinner clouds.
- **Answer**: For the first sentence, we agree and rephrase that statement. The second part I found to be contradicting. CloudSat in particular can produce this peak whereas our model fails to appropriately reconstruct high reflectivities in higher altitudes (Fig.3). The missing peak might be connected to the imbalance within the data and resulting underestimation of high reflectivities. We rephrase that paragraph for clarity.

Fig. 4 -

- **Comment**: Perhaps blur the Cloudsat data to match the DL resolution? It would also be nice to have 5km marked at the bottom of the y-axis (if that is the case), to highlight this doesn't go to zero.
- **Answer**: Figure 4 shows the blurred CloudSat data after downsampling. The models predict the cross section with a higher blurriness, as a result of the regression-to-the-mean. We will add the lower y-axis label.

L270 -

- **Comment**: So it is (or could be) a simultaneous retrieval, due to the inclusion of the water vapour channels?
- **Answer**: We agree the water vapour channels could lead to this distortion over water bodies. There are only few observations in these regions. Extreme values have a higher influence on the mean. This could be a further source for distortion.

Fig. 7 -

- **Comment**: Much more structure on the Res-UNet results - why is this? Also, the Res-UNet image seems to have almost all the cloud tops at the same altitude, other than a small fraction of low-level cloud in some regions.

- **Answer**: Both are aggregated to a monthly mean, whereas the datasets have a resolution mismatch (0.05 ° for CLAAS-V002E, 0.003 ° for the predictions). The aggregation of the derived CTH contains a higher small-scale variability. The results reflect the lacking model ability to for predict low level and high level clouds (Fig.3). Using a fixed threshold of -15dBZ to identify clouds may not be accurate for all regions of the FD. Combining the threshold with the shifted reflectivity distribution can affect the distribution of the CTH in, smearing out regional differences. As a result of this underestimation of high reflectivities, the mean of the monthly aggregate is shifted towards a lower CTH than for the CMSAF product. CloudSat information is retrieved at the same local time for every region (due to its sun-synchronous orbit) and only applicable to daytime predictions. This can increase the diurnal representation of the reflectivities.

L299 -

- **Comment**: I guess this is much like satellite retrievals in general? Presumably there is actually just a lack of information there?
- **Answer**: Apart from the inaccuracy of the model results itself, we agree this could be traced back to a general lack of information. We rephrase the sentence to clarify.

L300 -

- **Comment**: 5km is quite high from the ground to deal with clutter. It also cuts out a lot of the lower level clouds with a more challenging cloud top height retrieval, which might make the DL method seem better overall? I am not suggesting they should be included (working with just high clouds is an important task), but could be worth mentioning.
- **Answer**: We are aware that low level clouds are important features needed to be represented within the model results. CloudSat suffers from attenuation in low height levels. After analyzing the CloudSat quality flag, we observe that most of the reflectivity values between 0–2.4 km are affected by noise, so we set these values to -25 dBZ. This affects the predictions between 2.4 and 5km height. Due to the regression-to-the-mean and the imbalanced reflectivity distribution, we see a further shift of the distribution towards low reflectivities, especially up to 5 km. Using a modified network architecture and loss function can help to improve the representation of clouds between 2.4–5km, and the calculation of the CTH. We include a statement in the revised manuscript.

L321 -

- **Comment**: I think this reduction in bias should be shown if it is claimed. Those external data sources might decrease bias themselves in some situations (e.g. with a better estimate of water vapour than is available from the SEVIRI IR channels).
- **Answer**: The statement is drawn from the comparison of the CMSAF CTH and the predicted CTH, but we revise the paragraph and add a further quantification.

L325 -

- **Comment**: maybe “remote oceanic regions” instead of “secluded regions above the sea surface”?
- **Answer**: We appreciate that suggestion and change the wording here.