General comments:

- **Comment**: In my opinion, any application of machine learning to scientific research questions requires some level of uncertainty quantification. I appreciate that this might go beyond the scope of the paper, but estimating model uncertainties would greatly improve the trustworthiness of the results, and likely help inform how (and for which scenarios) the model needs improving.
- **Answer**: We agree this is an interesting idea. We did not include a quantification of model uncertainties since it seemed out of scope of our initial submission. We will evaluate whether and how to integrate a form of uncertainty quantification in the revised manuscript.
- **Comment**: The writing is at times too verbose and unclear. I am including a number of specific comments below, but I would suggest going through the paper again and streamlining the narrative. At times, the section headings could be more specific. Some of them (e.g. “Comprehensive predictions”) are not particularly descriptive of the work done.
- **Answer**: We revise the manuscript with particular attention to this issue.
- **Comment**: Please add tables of the training and hyperparameters in the paper (or better Supplementary Information). Most importantly, please include information about the number of parameters of each of your models.
- **Answer**: The tables will be added in the revised manuscript.

Specific comments:

- Line 9:
  - **Comment**: Abstract: “average error” is a bit vague and not the same as RMSE. Please revise. Furthermore, it is not clear what “total value range” in this sentence is referring to.
  - **Answer**: Changed to: "The RMSE after training is 3.41 dBZ which equals a difference of 7.5 % compared to a reflectivity scale between -25 and 20 dBZ."
- Line 27:
  - **Comment**: “[...] spatially and temporally limited perspective [...].” Please expand on what the spatial and temporal characteristics of the different instruments used in this work are. Specifically, I think the authors could expand on the benefit of geostationary satellites when it comes to temporal resolution, as polar-orbiting satellites often require 1+ days to observe the same area again.
  - **Answer**: The spatial and temporal characteristics of the specific instruments are described in Section 2.1.1 and 2.1.2. We agree that some points may be missing. In the revised manuscript, we add a more detailed comparison.
- Line 29:
  - **Comment**: “While this analysis often rests upon subjective labeling or fixed thresholds [...].” Please add example citations to previous work.
  - **Answer**: We change the text adding examples for the retrieval of the cloud optical depth (https://doi.org/10.1175/2011JAMC2601.1) and the cloud effective radius (https://doi.org/10.5194/acp-20-1131-2020).
- Line 33:
  - **Comment**: “joined” ==> joint

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

- Line 36:
  - **Comment**: I am not sure what “generability” means.
  - **Answer**: This sentence is unclear, we omit it. The sentence before and after (lines 35, 37) are sufficient for our statement (that, to our best knowledge, no large-scale 3D interpolation of the cloud reflectivity from radar exists).

- Line 54:
  - **Comment**: I would suggest including other works, such as https://arxiv.org/abs/1911.04227, when discussing cloud classification.
  - **Answer**: We appreciate that remark and revise the citations.

- Line 70+:
  - **Comment**: The sentence is very long and hard to follow. Please rephrase.
  - **Answer**: Changed to: "Former studies had a focus on reconstructing the 1D or 2D cloud column. In contrast, we apply a DL framework to predict the radar reflectivity not only along the radar cross section, but on the satellite full disk (FD)."

- Line 78:
  - **Comment**: “[...] originates a geostationary satellite.” ==> “from” missing.
  - **Answer**: Changed to: "The input data for the network originates from a geostationary satellite [...]".

- Line 84:
  - **Comment**: The EUMETSAT abbreviation was already used in line 78, before it was defined here.
  - **Answer**: Moved the definition to line 78.

- Figure 1:
  - **Comment**: It would be great if you could add information about the size of the boxes shown in Part 1, to make it more clear how the matching algorithm works. Furthermore, it was confusing to see 90 height levels in the figure, when the text previously described CloudSat to have 125 height levels. Could you also comment on the missing data shown in the CloudSat profile?
  - **Answer**: We revise the figure adding an extended description of the matching algorithm. The boxes have a size of 128 x 128 pixels. The missing data in the profile represents every pixel with a value of -25 dBZ (lines 125-127). In Figure 1, these values are transparent to highlight the profile itself. We add this explanation in the figure caption. The reduction of the height levels is described in lines 123-124, but we restructure the section for clarity.

- Line 114:
  - **Comment**: “Extracted satellite samples display the physical predictors fed into the network [...].” Please rephrase, as it is not clear to me what you mean.
  - **Answer**: Changed to: "The matching algorithm extracts the image-profile pairs. We use the information of the satellite channels within these pairs to reconstruct the vertical cloud distribution."

- Line 115:
  - **Comment**: I would rename “samples” as “matched image-profile pairs” or similar to be more exact.
  - **Answer**: We appreciate that suggestion and revise the phrase.
- Line 117:
  - **Comment**: How come you have no test set? Validation sets are great, but susceptible to “human gradient descent”, since we usually justify model modifications by improved performances on the validation set. This doesn’t necessarily mean that the model is better, it just means that performance on these specific examples is improved.
  - **Answer**: The model is trained and validated on data from 2017 (lines 115-116). We use nine months for training (here January-September) and three months for validation (October-December). Due to limited resources, data from May 2016 (n=1500) is used as a test set to evaluate the model performance and to calculate the CTH. We will revise the text and the reported figures to ensure they represent the performance on the test set. In this study, we apply a standard architecture with slightly modified learning-rate and weight decay. In the manuscript, we add a more detailed training protocol.
- Line 123:
  - **Comment**: How was the reduction from 125 to 90 height levels done?
  - **Answer**: We analyzed the CloudSat quality flag for the radar profiles to identify height levels with a high proportion of noise (lines 124-127). The original radar scene contains a high amount of noisy pixels up to 2-3km height. That is why we crop off the lower 10 height levels. The reduction affects the representation of low clouds and the cloud base, reducing the model performance in these altitudes. We find an additional cloud free region > 20 – 30 km (https://doi.org/10.1029/2008JD009982, line 19). To reduce the proportion of cloud-free pixels, we cropped these upper height levels off. The final Z-dimension consists of 90 height levels (10 - 100, between 2.4 km and 24 km). We add an extended explanation in the revised manuscript.
- Line 140:
  - **Comment**: “By seeking non-linear approximations of between the input and the output data [...].” Please rephrase.
  - **Answer**: Move the sentence up to line 134 and changed to: "Neural networks have the potential to capture highly complex relationships between input and output data. The Res-UNet displays a modified framework designed for the use-case of remote sensing data. [...]"
- Line 163:
  - **Comment**: “As flipped images are perceived as new samples, we enhance the amount of training data by giving all samples a chance of 25% to be either vertically or horizontally rotated.” Please comment on whether these transformations are valid for the context of satellite measurements, especially CloudSat, with its ascending and descending orbits. Are the two acquisitions totally equal, in that flipping one can simulate the other?
  - **Answer**: The data augmentation used here is not meant to simulate ascending/descending orbits. Instead, by applying random flipping, we aim at including invariance to the cloud orientation in the model and avoid overfitting. A structured analysis of differences between ascending/descending CloudSat orbits is out of scope for this paper. Since data from both

orbits are included in the training set we don't expect a different performance of the model on either orientation.

- Equation 3:
  - **Comment**: Either define all variables in the equation, or remove from the paper, as the RMSE is a relatively standard quantity.
  - **Answer**: We remove the equation in the revised manuscript.
- Line 173-174:
  - **Comment**: Rephrase "horizontal diagonal".
  - **Answer**: Changed to "radar transect" or "radar cross section".
- Figure 3:
  - **Comment**: Please explain how many samples the joint plot was calculated over.
  - **Answer**: The plot was calculated over all samples of the test set (n=1500). We add the number of samples in the revised figure.
- Line 218+:
  - **Comment**: This sentence is very hard to follow. Please re-write.
  - **Answer**: We revise this paragraph. This sentence will be changed to: "In the joint plot, values around 0 represent a high agreement between the observed and predicted distribution. On these height levels, the observed reflectivities are most accurately reconstructed. We observe areas of high agreement in shape of a curved line reaching from high to low altitudes for all of the three models."
- Line 220+:
  - **Comment**: Do you have any idea why you observe different performance trends for your pixel-based and Res-UNet approaches?
  - **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.
- Line 226:
  - **Comment**: How were the four samples chosen?
  - **Answer**: Those four samples were randomly chosen from the test set. We add the missing information in the revised manuscript. We add this information in the manuscript.
- Line 228:
  - **Comment**: What do you mean by "transferability" in this context?
  - **Answer**: Refers to the ability of the network to better represent the horizontal and vertical position of the clouds along the transect. Wee see this is unclear and omit the sentence, as the information is more precisely expressed in the sentences before and after.
- Line 231:
  - **Comment**: "A denominational structure [...]." I am not sure what you mean by this.

  - **Answer**: Refers to the results for the OLS and RF in Figure 4. In contrast to the DL results, the reconstructed cross section is not as smooth but shows a rather fragmented structure.

- Figure 4:
  - **Comment**: Please label each subplot, and refer to the subplots as you discuss the results in the main text to make it easier to follow your arguments.
  - **Answer**: Thank you for the remark, we add this in the revised manuscript.

- Line 233:
  - **Comment**: “[...] the Res-UNet shows more robust results [...].” I am not 100% convinced that the RMSE and visual inspection of the results qualify this sentence. Have you looked at any other metrics, or quantitatively studied performance as a function of cloud type for more than a couple of samples?
  - **Answer**: We did not derive the cloud type as it was not within the scope of the study. But we used not only a few samples for the evaluation but we calculated the RMSE for all models on the test set (n=1500). Over all height levels, the RMSE for the Res-UNet decreases by 30-35% compared to the OLS & RF. This percentage is used to support our hypothesis.

- Line 235+:

  Comment: I am not sure I follow your discussion about quality flags. Please clarify. **Answer**: We refer to the internal CloudSat flag that describes the quality of the received reflectivity (lines 125-127). Reflectivity values are filtered by a minimum threshold of six and set to -25 dBz (our background value). As this affects almost all values in low altitudes, our network reconstructs only the cloud signal above 5 km height. We will revise this paragraph.

- Line 238-239:
  - **Comment**: “[...] this leads to a lower model uncertainty.” Do you mean “lower error”? Model uncertainties to me means quantification of how certain a model is in its predictions.
  - **Answer**: Yes, this refers to the model error and the difference between the observed and predicted distribution. We change the manuscript accordingly.

- Line 239:
  - **Comment**: “That said [...].” This sentence isn’t quite clear to me. Please rephrase/clarify.
  - **Answer**: Changed to: "Leaving out noisy pixels is needed to improve the model performance. At the same, this results in a loss of information in low altitudes. This issue is reflected within the results. [...]"

- Line 260:
  - **Comment**: “The first peak [...].” It would be really helpful if you could refer to the labels of the subplots when discussing the results.
  - **Answer**: We see this is confusing and revise the section to be more clear.

- Line 264+:
  - **Comment**: “These channels are identified as essential information [...].” It would be great if you could show evidence for this, or expand the discussion on what makes you draw this conclusion.

  - **Answer**: This statement is based on a different model setup without VIS. In the initial submission, we did not add this for brevity, but we revise the paragraph and add supporting information.

- Line 267+:
  - **Comment**: “Comparing [...] reveals an overall high agreement.” Do you have any quantitative comparisons between your model outputs and the CTH from CLAAS, or is this mainly from visual inspection?
  - **Answer**: The initial comparison was visual. In the revised manuscript, we add a quantitative analysis.
- Line 277:
  - **Comment**: “The approach offers [...].” Which approach are you referring to?
  - **Answer**: Refers to computing the CTH only by predicted reflectivities. We change that paragraph to be more precise.
- Figure 6:
  - **Comment**: Is the comparison calculated across the entire FD of the model predictions, or just the tracks that overlap with CloudSat? If it’s non-normalized frequencies, they should be calculated across the same number of datapoints to be comparable, if I am not mistaken? Please clarify in the caption and main text.
  - **Answer**: Thank you for that remark, we will check the figure concerning the number of datapoints and revise the corresponding text.
- Discussion section:
  - **Comment**: I found it at times hard to follow when you are referring to your own work, versus work done by other people. Could you please go through this section again?
  - **Answer**: We revise the manuscript with particular attention to this issue.
- Line 280:
  - **Comment**: Please clarify what you mean by “minimal architecture”.
  - **Answer**: This study is based on a small and rather standard architecture (UNet added by residual connections). We want to point out the possibility to interpolate continuous 3D clouds from the 2D data sources. That is why we stick to the UNet instead of using a more complex and powerful architecture (like ViT). We rewrite the sentence for clarity.
- Line 282:
  - **Comment**: “others” ==> others’ work
  - **Answer**: Changed in the revised version.
- Line 284:
  - **Comment**: “Nonetheless, defining those variables as additional predictors has a negligible effect on the model performance.” Are you showing evidence for this somewhere?
  - **Answer**: Similar explanation as for lines 264+. We tried a different model setup. Since it achieved poor results, we did not add it in the initial submission for brevity. In the revised manuscript, we change that paragraph.
- Line 288:

- **Comment**: “Leaving out the affected channels downgrades the overall performance.” Same here, do you have evidence supporting this statement?
- **Answer**: Same explanation as above, will also be changed.