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

**Comment**: The most interesting part of the paper to me is the network structure that generates the 3D fields. The U-Net structure is appropriate for the problem at hand, although it is not quite clear from the description how the transformation from 2D to 3D is performed.

**Answer**: The transformation process from 2D to 3D is illustrated in Figure 1, Section 2.1.3, and Section 2.2. The matching algorithm stacks the 11 MSG SEVIRI channels of one sample together to one 3D image of 11 x 128 x 128 [C x H x W] pixels. Within this sample, CloudSat crosses a horizontal transect. The location of the cross section is defined as described in lines 104-109. We aggregate the CloudSat pixels on this track by their maximum value to fit the resolution of MSG SEVIRI. As a result, we receive a 3D image of 90 x 128 x 128 [Z x H x W] pixels containing the reflectivities along the cross section. After the downsampling, the transect has a width of one pixel in the 3D image (Fig. 1.3). Pixels outside the transect are assigned as missing values. In that way, we can preserve the location of the CloudSat data during training and use this information to extract the pixels from the 3D predictions to calculate the loss. The loss function is the RMSE (lines 168-171). As we are restricted to calculate the loss only for predictions along the transect, the loss reflects the model performance on only a small subset of the data (defining it as a sparsely supervised network). Despite these limitations, the final output consists of a full 3D prediction of each tile. Each tile contains a contiguous cloud field and is visually analyzed in Section 3.3 and Figure 5. We add an extended explanation in the revised manuscript.

**Comment**: While the network is interesting, I wonder about the usefulness of the model outputs since the model clearly suffers from regression to the mean and blurriness of the results. Also, the uncertainty of the outputs is not estimated. Could you comment a bit on which applications would find the current results useful? And how the above mentioned issues could be improved?

**Answer**: As mentioned in lines 242-243 and 305-307, we agree that the network suffers from problems connected to the regression to the mean and the blurriness of results. In the revised manuscript, we will address these problems and possible improvements more straightforward.

The histogram of the CloudSat data shows a shift towards the background value of -25 dBZ (Fig. 6). To address this imbalance, and, to enhance the model performance, we use the UNet with additional residual connections (Res-UNet) and an automatic learning-rate optimizer. Using a full year of satellite data leads to the generation of approximately 30.000 samples. We reduce the percentage of cloud-free samples face the imbalance between cloudy and clear-sky samples (lines 116-119). Still, a cloudy sample can have a substantial proportion of background values. As reducing the size of the dataset affects the model performance, we randomly flip the samples during training for data augmentation. The goal is to include model invariance to the cloud orientation in the model and avoid overfitting. Although we use techniques to deal with the regression to the mean, the results show room for improvement. All predictions smear out. This is not caused only by the imbalance of the data, but can be led back to the loss function. The RMSE is known to enhance the blurriness of the predictions. An adapted loss may improve the results. Nevertheless, the study suffers from the resolution mismatch between the MSG SEVIRI and the CloudSat data. Aggregating the CloudSat pixels further increases the blurriness of the ground truth as well as those of the predictions (lines 109-112). An improvement can be achieved by using a deeper architecture along satellite data with a higher resolution (e.g., GOES-R, Himawari 8/9, or, in the future, MTG). As stated in lines 318 – 321, the goal of our study is to point out for the first time the possibility to derive continuous 3D clouds using DL. We see the innovation of our work not in the network architecture itself as it was kept simple on purpose. Nevertheless, using powerful architectures like the Vision Transformer as well as an adapted loss can improve the above mentioned issues.

Despite current performance issues, we see a vast potential for the method in applications of atmospheric and climate sciences. In Section 3.3, we exemplary compute the CTH to evaluate the quality of the predictions for analyzing cloud microphysics (lines 255ff.). Further use-cases comprise e.g., the analysis of cloud organizational patterns, the identification of lightning locations, or a detailed analysis of precipitation

onset. We use the CloudSat radar data as ground truth, but the approach can be transferred to different data on a 2D transect, e.g., aerosol measurements. The absence of artifacts between predicted tiles on the FD enables an investigation of 4D movement patterns through space and time (lines 249-254). The data can be used together with simulation results to improve the representation of large-scale cloud dynamics.

In the next section, we shortly address your specific comments that will be added in the revised manuscript.

- lines 24-25:
  - **Comment**: "Passive sensors such as geostationary satellites": the statement needs more precision, satellites are not sensors
  - **Answer**: Changed to “While observations from passive sensors on geostationary satellites are limited to monitor the uppermost atmospheric layer from space, [..]”.
- lines 36-37:
  - **Comment**: "The large-scale generability of these methods is expandable since their 3D results are limited to the cloud’s spatial vicinity": I don’t understand this sentence, please clarify.
  - **Answer**: We agree that this sentence might be unclear. We will omit it as it is not necessary for the central statement of the paragraph. The relevant information can be found in the sentences before and after this one.
- line 48:
  - **Comment**: "time efficiency and feasibility": it’s not clear to me what this means
  - **Answer**: DL networks do not require feature extraction like traditional ML methods. That reduces the time necessary to start the training and the user generated bias. Also, their performance on big data is better. We change the wording to avoid misunderstandings.
- line 80:
  - **Comment**: "orbiting the globe on a sinusoidal track": how does a satellite orbit on a "sinusoidal track"?
  - **Answer**: This refers to the radar track visualization in 2D. Changed to: "The ground truth of the study is derived from an active radar on board the CloudSat satellite which moves on a sun-synchronous orbit."
- lines 91-92:
  - **Comment**: "resampled to a geographic grid": what kind of grid, a lat-lon one?
  - **Answer**: Yes, data is resampled to a spatial grid in geographic coordinates with a resolution of 0.03° (lines 90-92). The missing information will be added.
- Section 2.1.2:
  - **Comment**: CloudSat is on a sun-synchronous orbit, meaning it sees every location at the same local solar time. This might introduce some diurnal bias to the data; this should be acknowledged.
  - **Answer**: We appreciate that remark and add following: “Since CloudSat follows a sun-synchronous orbit, it receives information on the cloud reflectivity at different locations along the globe always at the same local time. This reduces its ability to reflect the diurnal variability within each region of the AOI.”.

- line 109:
  - **Comment**: The use of "XY" is confusing, I read this initially as "X times Y" but apparently you mean a diagonal transect through the image? Or did I misunderstand?
  - **Answer**: “XY” refers to the diagonal transect of the CloudSat radar along the pixel dimensions X and Y of the MSG SEVIRI image. We change the naming to “transect” or "cross section" in the text (in figures as [Z,(H,W)]).
- line 125:
  - **Comment**: "smoothing" should probably be "filtering"
  - **Answer**: We agree and change to “filtering”.
- line 137:
  - **Comment**: The use of the word "delineate" here and a couple of other places in the places seems incorrect
  - **Answer**: Changed “delineate” to “predict” (line 7, line 137) or "estimate" (line 81).
- line 159:
  - **Comment**: I would like some more details on how the network structure maps the 2D input fields to the 3D output fields. Is the channels dimension used transformed to the Z dimension of the output?
  - **Answer**: See Figure 1 and Section 2.2. The MSG SEVIRI data consists of a 2D field for each satellite channel. During the matching scheme, these channels are stacked to one 3D image of 11 x 128 x 128 pixels. On the encoder side of the network, the channel dimension of the MSG SEVIRI input is expanded to the proposed filter size of 256. On the decoder side, these filters are reduced and mapped to a model output size of 90 x 128 x 128 pixels. As a result, the channel dimension of the output equals the Z dimension of the CloudSat ground truth. We will add an extended explanation in the revised manuscript.
- lines 163-164:
  - **Comment**: How is the 3D scene predicted by the network compared to the CloudSat data during training? CloudSat only gets a 2D vertical cross section of the scene. Is only part of the scene selected for comparison? Also, what loss function do you use for training?
  - **Answer**: Seems to be similar to the question dealing with the transformation from 2D to 3D in the upper part of the text. In summary, we transform the 2D cross section of CloudSat to a 3D image of 90 x 128 x 128 pixels during the matching scheme. This 3D image has the same size as the model output after training. For the CloudSat 3D image, only pixels along the transect hold finite values. These can be used to filter both images and calculate the RMSE between the observed and predicted cross section (sparse supervised network). We will add this information in the revised manuscript.
- line 176:
  - **Comment**: "Both models": unclear which models this refers to
  - **Answer**: Refers to the pixel-based methods mentioned in line 172. We rewrite the sentence for clarity.
- line 181:
  - **Comment**: "pictures" is used incorrectly here.

- **Answer**: Changed to “The RF is a supervised ML algorithm which provides robust results when working with environmental datasets in the natural sciences [. . . ].”.

- lines 189-190:
  - **Comment**: In the joined 2400 x 2400 pixel 3D prediction, is the field continuous at the borders of the 100 x 100 pixel tiles? Or do you see discontinuities or artifacts?
  - **Answer**: The joined 3D prediction consists of a continuous field without artifacts. As shown in Figure 5, each of the sub-figures (b) – (d) represents a combination of several 100 x 100 pixel tiles ( 2.5° on the Lat-Lon grid). The results are described in Section 3.3, lines 250 – 252 and will be explained in more detail in the revised manuscript.
- line 220:
  - **Comment**: "That said" seems out of place here - please revise.
  - **Answer**: Changed to: “[. . . ] The DL network indicates an underestimation of high reflectivities and an overestimation of low reflectivities for low-level clouds.”.
- line 231:
  - **Comment**: I don’t understand that "denominational structure" means here.
  - **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.
- lines 262-263:
  - **Comment**: High, thin ice clouds may also not be observed by CloudSat due to being under the minimum detectable reflectivity.
  - **Answer**: Thank you that remark. We will change the text in the revised version.
- Figure 7:
  - **Comment**: Maybe you could add a panel showing the difference of a and b to illustrate the biases better.
  - **Answer**: Interesting suggestion, we add that figure in the revised manuscript.
- line 288:
  - **Comment**: "Leaving out the affected channels downgrades the overall performance": it would be good to see something to demonstrate this.
  - **Answer**: We tried a model setup without VIS. Due to its poor results, we did not include it in the manuscript for brevity. In the revised manuscript, we will rewrite the text in this paragraph.
- lines 292-293:
  - **Comment**: "In contrast to pixel-based DL methods like the CNN or CGAN, the Res-UNet utilizes a larger receptive field preserving the spatial dimensionality and global context information during the training routine." This is not a correct statement regarding the CNN or CGAN architectures. CNN architectures can also achieve large receptive fields and global context using downsampling. In fact the UNet itself is a type of CNN - its distinguishing feature is the addition of skip connections to preserve resolution. As for the CGAN, it refers to a certain training setup of generative models that could be implemented with either normal CNNs or with (Res-)UNets.

- **Answer**: We agree and rewrite the sentence: “Like other CNN architectures, the Res-UNet preserves the spatial dimensionality and global context information during the training routine (Wang et al., 2022). Compared to pixel-based methods like the OLS, it reconstructs the spatial connectivity between the pixels more accurately and enhances the representation of the continuous cloud field. [. . . ]”.

- lines 312-313:
  - **Comment**: Approximately 1 km resolution is also already available from the GOES-R series and Himawari 8/9 satellites.
  - **Answer**: We know about the resolution of these satellites, but we decided to stick with the MSG satellite to fit our main study area. For a potential global composite of the 3D predictions, data from GOES-R and Himawari 8/9 may help to improve the results. We add this information and change lines 311-314: “At the moment, a compromise on the resolution is necessary to obtain predictions centered over Europe and Africa. However, newly emerging instruments offer an enticing prospect to tackle this information loss. While comparable data sources like the GOES-R series or the Himawari 8/9 satellites already provide data in a resolution of 1 km, the recently launched satellite Meteosat Third Generation by EUMETSAT (Holmlund et al., 2021) is expected to close the data gap. Together, they can be leveraged to investigate a 4D reflectivity field through space and time”.

- lines 321-322:
  - **Comment**: "Since it is independent of external or interconnected data sources, the bias within the data is reduced.": unclear sentence, I’m not sure how the latter follows from the former.
  - **Answer**: The operational CM SAF CTH is computed using the MSG SEVIRI satellite data as well as derived products and additional data. Each of them may bring their own bias, potentially multiplying their effects on the final CTH. In contrast, our CTH is based only on the predicted reflectivity. In that way, we can minimize the influence of additional data sources. We will rewrite this explanation in the revised manuscript.