The very helpful comments and suggestions by Reviewer 1 are very much appreciated and we would like to thank him/her for the time and effort he/she put into this review of our manuscript. In the following, we would like to go through all comments/suggestions and reply to them or answer them point-by-point. Reviewer comments are put in **bold font**, our replies are colored and changes inserted to the manuscript are put in *italics*.

In this paper, the authors present a significant advance in the accurate detection of cloud in polar regions using advanced machine-learning techniques, and go on to illustrate the utility of this algorithm by retrieving much better estimates of thin ice thickness in the vicinity of the Brunt Ice Shelf, West Antarctica. The novel approach and subsequent advances in cloud detection are sorely needed for polar MODIS users, and as such I believe this paper is highly suitable for publication in TC, after minor modifications outlined below.

We appreciate the comment and are happy to reply to the made comments.

L22-29: It would be nice to point to other polar MODIS applications which would benefit from such a better cloud mask. Examples include composite image generation, landfast ice mapping, possibly sea ice motion retrieval using image cross-correlation.

We will follow this suggestion and add both previous studies that could have benefited as well as future potential applications. Examples include:

Ludwig, V., G. Spreen, C. Haas, L. Istomina, F. Kauker, & D. Murashkin (2019). The 2018 North Greenland polynya observed by a newly introduced merged optical and passive microwave sea-ice concentration dataset. The Cryosphere, 13(7), 2051–2073. doi:10.5194/tc-13-2051-2019

Reiser, F.; Willmes, S.; Heinemann, G. A New Algorithm for Daily Sea Ice Lead Identification in the Arctic and Antarctic Winter from Thermal-Infrared Satellite Imagery. Remote Sens, 2020, 12, 1957. https://doi.org/10.3390/rs12121957

Aulicino, G.; Sansiviero, M.; Paul, S.; Cesarano, C.; Fusco, G.; Wadhams, P.; Budillon, G. A New Approach for Monitoring the Terra Nova Bay Polynya through MODIS Ice Surface Temperature Imagery and Its Validation during 2010 and 2011 Winter Seasons. Remote Sens. 2018, 10, 366.

L29: This problem has also been reported in coastal leads, e.g. Fig. 6 from Fraser et al., 2009. DOI: 10.1109/TGRS.2009.2019726

We will add the suggested reference-

L40: Why was this study region chosen? Is this applicable for both flaw leads and nonlinear coastal polynyas (e.g., Terranova Bay)?

This point was brought up by both reviewers. The region was chosen because of earlier experience with the region through the corresponding author (see Paul et al. 2015a,b) as well as several benefits for the setup of the algorithm we would like to explain in the following. While this polynya is not a major player in, e.g., deep-water formation (as pointed out by Reviewer 2) it is one of the most active regions in the Weddell Sea and similar in ice production to the much larger Ronne Ice Shelf polynya (see Paul et al. 2015b). The high activity as well as the good spatio-temporal coverage through Sentinel-1 led to the decision to start-off with this region. However, the approach is assumed independent of the selected region as it mainly depends on the received satellite TIR signals, i.e., the temperature differences between surface types and clouds. Due to the investigation of a complete freezing period, this is assumed to be comparable to other polynya regions and should also be independent of polynya shape/size. A study applying the proposed procedure to all Antarctic polynyas is currently under preparation.

We will a clarification to the manuscript: e.g.

This region was chosen for its combination of high inter-annual polynya activity and high spatio-temporal coverage with Sentinel-1 data. Results are expected to be transferable to other polynya regions in the Antarctic.

L54: At this stage, It strikes me that it might be better to describe input data before discussing the methods (i.e., move section 2.2 to here). This may have just been my personal preference though.

We ill change the order.

L112: This parapgraph needs more explanation. E.g., the 29-23. . . metric. Also, why are some numbers in bold type? What are the 35 epochs? 100 whats in a batch? Huber needs a capital H too.

Due to overall changes in the manuscript related to comments by the Reviewer and a change in the processing software, this part will also change as described in a comment below. However, numbers remain also in the updated manuscript and refer to the number of hidden layers and associated neurons in the neural network architecture (here the autoencoder). For the above example, this refers to a total of seven layers with an input and output layer consisting of the 39 input variables; two hidden layers with 23 and 10 neurons respectively on each side of the dimensional reduced layer consisting of three neurons. The bold face numbers (on the left) highlight the encoder part of the autoencoder, which is used for the dimensional reduction. The decoder part is only used for the training of the autoencoder and not used afterwards. We added a clarification to the manuscript. For further reading we suggest the standard textbook by Goodfellow et al. (2016); https://www.deeplearningbook.org/.

Section 2.2: You need to describe the version of the MODIS products – particularly the MOD29 product. This determines which MOD35 version went into the product. There have been some significant improvements to MOD35 over the years, so it's important to document which version. There is no description of the gridding or MODIS destriping/de-bowtie here. These must have huge influence on the performance of the algorithm, so a description of these processes is needed, in my opinion. Many of the channels used suffer strongly from detector striping in particular.

We appreciate the reviewer for pointing these shortcomings out to us. The MOD/MYD29 version used is Version 6. We will also add this information to the manuscript.

The gridding is described in the manuscript (L145f), however, any additional pre-processing is not. The remainder of this comment is addressed in the process of following the suggestion by the reviewer concerning the cal/val.

L140: What is the resolution of the IST product?

Also 1km x 1km as the MOD021KM. We will add this information to the manuscript.

Section 2.2.1: The destriping description may fit better here.

Please refer to comment above.

L158: What about the increased atmospheric path encountered for high incidence angles – is that more important than the geometry distortion?

Please refer to our answer to your cal/val comment below.

L164: Does MOD/MYD29 also apply atmospheric correction to more accurately determine IST?

According to https://nsidc.org/data/mod29, the MOD/MYD29 product is derived from MOD/MYD021KM product, which is using the TOA radiances, so no atmospheric correction is applied.

L167: I guess you're developing this algorithm for coastal, latent heat polynyas. It might be good to make this clear here. I doubt it would work for offshore/sensible heat ones (which is fine)!

The success of the proposed cloud discrimination depends on the temperature regime and textural properties. To our knowledge, the thin-ice retrieval generally works for all thin ice areas. Problems could arise for rather fast changing shapes of offshore polynyas,as a set of input parameters depends on differences between swaths. However, the referenced line number solely relates on the selection of the SAR reference data.

L178: This sentence "Generally, " is somewhat ambiguous.

We will remove this sentence.

Fig 3: One sub-figure would benefit from including a distance scale.

We will add one.

Fig 3: "Examplary" typo.

We will fix that one. Thanks!

L219: The cal/val split was done on a point-wise basis? This seems a bit strange. Isn't the point of the cal/val split to ensure independence between the calibration and validation datasets by withholding at a more basic level, e.g., scene level? What I'm trying to say is, two neighbouring pixels are unlikely to be completely independent. So if there's a 75% chance of each pixel getting into the training dataset, then it's pretty much guaranteed that the validation dataset won't be particularly independent of the training dataset. Could you comment on this?

This is a very critical point and the reviewer is correct in his/her understanding and in bringing this up.

Due to this comment and the additional time we had due to the extensive discussion phase, we decided for a re-do of this part of the manuscript and the approach (as it also touches the time consuming part of manual categorization). Additionally, we also decided for a change towards a different machine-learning software environment inside the *R* environment. By implementing the *keras* package (a frontend for the well-known *tensorflow* backend), we also allow for a better distribution of the final neural network to potentially interested colleagues/collaborators as it can be easily transferred and adopted to/by other *R* as well as *python* users. We hope this rather big change is also in the reviewers and the editors interest and within the scope of the review process.

Here, we would like to briefly summarize the made changes to the manuscript work flow as bullet points. However, while the general procedure and overall results have not changed, we think these changes lead to an overall clearer description of our algorithm development as well as clarify/handle other brought-up comments by both reviewers.

- Instead of doing a point-wise randomized split of all our data, we settled for a swath-based split as suggested by the Reviewer. This ensures independence between both data sets. We now, in total, used **eight** combinations of Sentinel1-A/B/MODIS collocation data as **calibration** and **four** as **validation** (we took two examples that were previously shown in Figs. 4 and 5 and 'moved' them to calibration).

- We simplified and clarified the selection procedure of these swaths by selecting only the one swath closest in time to the Sentinel1 acquisition that **covers the study area by at least 90%** and features **incidence angles equal or below 35deg in 65%** of the study area. Both measures ensure high quality for manual categorization of the MODIS data.

- We generated in total 18 combinations per identified swath with surrounding swaths (90% coverage and at least 60% coverage with incidence angels of 50deg and below) to ensure high variability in the predictors used for the classification. To account for striping in the MODIS data and clarify and streamline the overall algorithm description, **we limit the data from MODIS to channels 20, 25, 31, and 33 in addition to IST data as well as the GLCM metric**s. This reduces the inut features to in total 34.

- With this we generally followed the same setup and procedure as before in the manuscript but used the *keras* R environment to train and use the Autoencoder as well as the initial and final Neural Network classifiers.

The resulting new classifier provides results as capable as before, however, it accuracy assessment is more realistic than before and likely less prone to overfitting.

L278: Again, the Fraser et al., 2009 reference which shows this in a flaw lead would be good to reference here.

We will add that one.

Fig 4, 5: There is a mismatch between the actual extent of the Brunt Ice Shelf and the masked version, based on the Rtopo product. This is due to ice shelf advection in the time between the creation of both products. In this case, there are both areas of ice shelf outside the mask, and areas of water/sea ice within the mask. Other highresolution coastal datasets have mitigated this by including a manually-updated ice shelf extent product on a regular basis (e.g., Fraser et al., 2020, ESSD Discussions,

https://doi.org/10.5194/essd-2020-99), but this level of mitigation is probably unwarranted here. However, could you comment on the effect this might have on the training algorithm?

The oftentimes very cold temperatures on the ice shelf would actually result in a similar difficulty to be reproduced by the neural network as we see with the wide temperature range for clouds. This is why we employed an at least rough estimate of the ice shelf to exclude these areas during the training process. While small scale effects exist, as pointed out by the reviewer, the effect on the training success appears to be negligible.

Fig 6: This is a great way of showing the seasonality in bias. However I'm still hanging out for a good old-fashioned scatterplot comparing these two datasets. This would show highly complementary information to your time series.

Agreed. We will add a scatterplot to the figure.

L286: I think the "average" metric you use here may not be the best way to highlight how much better your algorithm performs! Have you considered also using RMS difference?

L337: Again, the suggested RMS statistic would better highlight your improvement.

While we agree with the reviewer that RMS is a statistical value of interest, the daily coverage differences of the area due to different cloud screenings between the two methods would supposedly dominate the RMS in resulting polynya area. Therefore, we assume it is probably not usable as a quality measure of the method after all. However, we will give it a try and think of potential additional measures to use. Nonetheless, we think the result shown in the current manuscript – that also older studies without the benefit of a better cloud screening still hold valuable information at least on an annual basis – is of interest and value to the scientific community.

L303: Unclear which product that this statement corresponds to.

We agree and will change the sentence to clearly refer to the IST swath data.

L309: "Good agreement" between what and what?

We will clarify this by adding: e.g. *between the OSCD and the MOD/MYD29 product* to the end of the stated sentence.