First of all, we would like to thank the reviewer for the expertise and time taken to review the manuscript. In general, we have considered all comments carefully and made changes to the manuscript to incorporate the improvements. In the following responses are given to comments where some additional detail is needed.

**Broad Comments:**

The manuscript covers an exciting topic that it is an open problem in the sea ice and SAR community. However, the text itself is currently lacking some much-needed details that would help the reader identify the full contributions of this work. The authors state in the Introduction that they are assessing existing deep learning approaches for sea ice classification in SAR imagery by testing them on a more reliable data set. The goal seems to be to advance our understanding of which deep learning approaches are actually advantageous for sea ice classification, leading to a standard for the community. However, there wasn't enough detail provided about the previous studies nor this study's methods to know if enough was kept consistent when repeating the analysis to have comparable results (e.g., did the previous studies also use X-band data? Were model parameters and frameworks kept consistent? etc.). The authors provide some discussion towards their differing results, especially in terms of resolution, but more information would be useful.

Establishing a new standard for the community is a little bit too lofty a goal for this study, as the best data driven solutions will always depend on the data, which will not be the same as in this study. The aim is to disseminate as much information as possible from the unique opportunity provided by the dataset created in this study to inform algorithm choices in the future and too offer a less idealised perspective of sea ice retrieval from SAR, than what is typically shown with human annotations used as ground truth – where labels are coarse and heavily influenced by human interpretation. In these scenarios typically regions are separated, because they look different on SAR, whilst in this study regions are separated because of different measured ice properties. As a result, the performance of the models in this study is more indicative of true dissemination of ice properties and not influenced by human labelling bias.

Concerning the consistency of methods: We have opted to use architectures that have been used for sea ice classification, but cannot go to the full length of effort to entirely reproduce many of the different variations that have been published on in the past. This would not be fruitful anyways, however, because of the difference in datasets that they were optimised for. The question we are trying to answer, and this will be posed more pointedly in the manuscripts next iteration, is: How do different CNN architectures perform on labels derived from measurements (not human interpretation) and how does that influence future algorithm choices if the aim is to produce classifications near the resolution of the sensor?

The answer to that question in the manuscript is, that real measurement derived classes are much more difficult to disseminate than human annotations – showing that human bias has had a large influence on studies of the past and the resulting models cannot be assumed to perform near the resolution/fidelity of the SAR sensor. Additionally, centre-pixel classifiers are significantly inferior at this task than semantic segmentation models, that can make use of intra-label dependencies.

**Specific Comments:**

**Section 1. Introduction**

1. Page 2, Figure 1: The caption mentions a red circle, but it was not visible to me. Perhaps check printed color (and consider using a different color for those who are colorblind).

This has been corrected.

2. Page 3, lines 47-48: Traditionally, these types of datasets tend to be too sparse to provide robust training sets for image-based ML models. How does this dataset differ?

Because of the efforts made during the MOSAiC expedition and the subsequent collocation for this work, the dataset used here is far larger than any previously used data derived from measurements. However, it still suffers from a loss of generality from being constrained to certain region. It is none-the-less most likely the most complete (as in large) collocated dataset that we will be able to synthesise at least until another expedition of the scope of MOSAiC comes along (which could be decades).

3. Page 3, line 57: How close is near-coincident? An example in parentheses would be helpful here.

This has been updated in the text and is 7 hours on average, with a range of 0-24 hours.

## Section 2. Methodology

1. Page 3, first paragraph: More information is needed here on how the TSX data is processed (what corrections were performed, filtering, etc.) and what the effective resolution of the data is after such processing. Were you considering backscatter as your data? If so, what was it normalized to (e.g., , etc.)?

Thank you, this has been added to the next version of the manuscript.

2. Page 3, first paragraph: It is stated that only co-polarized channels are used. How do you mitigate effects from wind-roughened waters and other confounding factors which are more prevalent in co-polarized data?

The models have full access to both channels and contextual data to mitigate this as best as possible. Both configurations have advantages and disadvantages that a retrieval model will have to deal with.

3. Page 3, line 68: The ALS dataset seems to focus on the winter seasons. Is any comparison done for seasonality transfer?

Unfortunately, in summer the ship (used as coordinate system origin) moved significantly more during the warmer seasons and thus accurate geolocation here is more difficult. The same trained models would certainly not work in the warmer seasons, but I would expect the discrepancy between centre-pixel and semantic segmentation models to persist if both were retrained on summer data.

4. Page 3, second paragraph: What was the regional coverage of the ALS dataset? Did it cover all representative portions of the Arctic, or was it constrained to a particular region?

It was constrained to the region (~15km radius) around the central observatory (floe) of the MOSAiC Expedition (this is what enabled the colocation).

5. Page 3, second paragraph: A verbal comparison between ALS dataset footprint size versus the footprint of a TSX image would be useful for context.

This has been added to the data section (TSX 50x15km, ALS 10x5km).

6. Page 4, line 87: Was a distribution analysis of the freeboard and roughness per SAR pixel done to ensure that the mean and standard deviation are representative statistical measures for these data? If the distribution is heavily skewed, it may be more appropriate to use the median, for example.

The median and mean of the distribution, where on average within a percent of the span of the respective variables, so that a Gaussian description seems adequate. We have added some text in the manuscript mentioning this.

7. Page 4, line 92: Again, it would be helpful to have an example of these time differences (minutes, hours?).

It is typically on the order of hours. Within the dataset we did not see a drastic event that significantly impacted the ability to collocate the data. The relevant variables also should be largely unchanged within this time span.

8. Page 4, line 95: It would be good to explicitly name the conventions you are pulling from and provide an associated reference.

Conventions was the wrong word to be used here. We have changed this to 'names' it is just to allow an easier mental abstraction of the classes rather than to keep mentioning the thresholds used to distinguish them.

9. Page 5, line 111: It is unclear to me why you are assuming a Gaussian distribution when the density functions in Figure 2 are non-Gaussian. Your metrics (e.g., standard deviation) are likely to be heavily influenced by outliers.

The Gaussian distribution is only assumed locally to estimate the uncertainties in the derived classes. Whilst the global distributions are non-gaussian (as they are bounded), locally the gaussian description seemed adequate (see answer to comment 6 above) to achieve this. It is worth mentioning that the effect of this simplification on the loss function is minimal, so the model training is hardly affected by the small error we are introducing by simplifying the uncertainties to a standard deviation.

10. Page 6, Figure 3: Please add the mathematical notations for freeboard and SAR backscatter to the caption here so it is clear what the figure axes are referring to.
11. Page 6, Figure 4: Similarly, it would be good to note in the caption that PDF refers to probability density function for the general reader.
12. Page 7, line 123: Does favoring equal class performance affect how the model will perform operationally, where classes are almost always not equal?

Absolutely, the way it is set up here gives the most representative performance for unseen regions. If the aim was to make the most accurate classifications of the regions, the balancing would have to be approached differently.

13. Page 7, line 125: This sentence is unclear to me. Perhaps the second mention of "backscatter" should actually be "topography"?

Thank you, you are completely right. This has been corrected.

**Section 3. Results**

1. Page 8, first paragraph: Referring to the VGG16, etc. as pixel-wise classification approaches is confusing here, especially since under the Section 2.3 you describe these classification approaches as predicting over patches as a whole (or just the center), and the segmentation approaches as predicting a label for every pixel in a patch. Given that, seems backwards to refer to the center-pixel classification approaches as pixel-wise classification. I would try using a different descriptor if possible.

This is a very good idea and we have adopted it for the manuscript. Thank you!

2. Page 9, Figure 5: Do you have ground truth labels for this example? If so, it would be helpful to include them in the figure.

We will choose a different scene where ground truth is available. Thank you for pointing this out.

**Section 4. Discussion**

1. Page 10, line 204: Can you elaborate more on how your results compare to the results from previous studies that you are replicating?

As no specific study is being replicated, we need to be careful with our wording. We have added some extra discussion to establish the differences and conclusions to be drawn in comparison with existing studies relying on manual annotations.

2. Page 11, Figure 7: A reference for the misclassification of water and old ice being a common issue would be helpful to include here.

Yes, thank you for pointing this out, we have added a reference.

3. Page 12: It would be great to see some discussion on how the temporal span of the dataset may affect the results, as well as how you expect the results to change (or not change) when applied to different seasons. For example, do you think certain models would be more robust to the existence of melt ponds on older ice during the summer months?

Good idea! We have added some more discussion. To briefly recap I think that the core result of the usefulness of intra-label relationships for ice classification will be particularly relevant for surface types with characteristic shapes, such as melt ponds (in high resolution data) and leads. In general, the backscatter signatures themselves will be even less reliable in the summer months, so therefore I would expect the segmentation models to have an even bigger advantage there.

4. Page 12: Likewise, a discussion on the spatial coverage of the ALS dataset and how that could affect your results, especially when compared to previous studies, would be useful. Did this dataset only cover a particular region of the Arctic, and would you expect results to differ if you had data from other regions?

We have added some more discussion about this. The core takeaway is, that of course this dataset is not representative of the entire Arctic, but the distributions and shapes of different classes will have the same complexity in other regions. Thus, the core messages concerning the strong bias of manual labels translating to classifiers persists. The discrepancy between centre-pixel classification and semantic segmentation models is also expected to be representative for the entire ice classification domain regardless of region.

Finally, we would like to once again thank the reviewer for a thorough and insightful review of the manuscript. The comments are very helpful and have/will significantly improve/d the manuscript from both a scientific and a didactic standpoint in our opinion. Sincerely, thank you!