Thank you for your efforts of going through the manuscript and the further suggestions. The language notes have been adopted and the larger comments are answered below. I believe the changes made and extra figure added help the clarity of the manuscript.

1. line 203 'inter-label' - this is different than in the reply to reviewers, where 'intra-label' was used. 'Inter' and 'Intra' label are quite different. Please clarify.

Inter is correct, as we mean the relationships between individual (pixel) labels.

2. The categorical cross-entropy vs KL divergence part is unclear. Looking at Table 1, the train/test accuracies and train/test KL divergence are quite different, and even for some models (e.g. Con vNext) the trends are different (in that training accuracy is higher than test for train while for KLD test accuracy is higher than train). Why is this the case? Then, when I look at the caption for the table it says 'standard deviation' has been shortened to 'std' - but there is no std in the table. Again on line 250 reference is made to the mean and spread of the KLD but on line 213 it is stated the the cross-entropy is minimized in training. Please clarify these points.

The Kullback-Leibler Divergence (KLD) is a measure for the distance of the distributions – considering not only the class with the highest predicted likelihood, but also describing how close the predicted label probabilities are to the target probabilities (derived from measurement). In this case the lower the difference is, the closer the models have fitted to the measured distribution. Thus, the trends are very similar to the accuracy trends, wherein models with higher disparities between train and test accuracies also have a larger difference between training and test set KLDs and the test KLDs are always larger (worse) than the train KLDs. We have added some text and rewritten the caption to clarify this

3. The wording for how the labels are generated I still find confusing. On line 136 it is stated a Gaussian distribution is assumed for the freeboard and reflection measurements. But in figure 2 freeboard and surface roughness are shown. Do the authors mean a Gaussian distribution is assumed for freeboard and surface roughness measurements? Continuing through this paragraph where it says the Gaussian distribution is integrated above and below the threshold - I am not sure the reader knows what this threshold is. You might mean (for example in the top right of fig2) where the color changes at a freeboard elevation of 0.4 so you are integrating the distribution to this threshold to determine the soft label for FYI etc. However, there are two PDFs shown on the right of Fig 2 (one for freeboard and one for elevation) - so presumably you integrate both to get the three classes considered (OW/YI, FYI and SYI shown in Fig 4). But this is not clear from the text. In addition the threshold mentioned in the text (searching for threshold) is for backscatter. This could be clarified by revising the text leading up to figure 2 and adding some subfigure labels (a,b,c etc). There are two sets of PDFs in the figure (one for the freeboard and surface roughness, and the other for the backscatter). Using subfigure labels would help by stating 'integrating the PDF (fig 2a etc)' so the reader knows what PDF and what threshold you are talking about.

Some detail has been added to the manuscript and a figure has been added to clarify the situation. The general procedure is to take the globally defined thresholds which define classes and then check locally – for each SAR pixel – how likely it is that the ice there lies above or below these thresholds, by assuming that all measurements of the ALS sensor which are mapped to that pixel admit an approximately Gaussian distribution. The refined text and figure I have included here for convenience:

“As detailed above, ice types are identified by thresholds in the reflectance, surface roughness or freeboard. The thresholds for the roughness and freeboard are indicated in the histograms in Fig. 4 by the different background colors. We can infer the probabilities of lying above or below a threshold for every pixel by assuming a Gaussian distribution of ALS freeboard and reflection measurements at each SAR pixel. From the 49 ALS measurements mapped to one SAR pixel, we compute the mean and standard deviation of the freeboard and can then compute the probabilities of lying below or above the globally defined freeboard thresholds by using the error function. Explicitly, we integrate the area under the curve of the estimated Gaussian probability density function, above and below the threshold. An example is shown in Fig. 3. Thus, we obtain 'soft labels' which give the probabilities of belonging to a certain class, rather than discrete classes. Assuming a Gaussian distribution allows us to also infer uncertainties of the surface roughness. One could have classified each of the 49 ALS measurements mapped to one SAR pixel and then used the relative occurrences as probabilities. However, this simplification to a Gaussian distribution leads to an inaccuracy of the probabilities (derived from freeboard) of only ~ 0.16% on average, but significantly increased computational efficiency.”

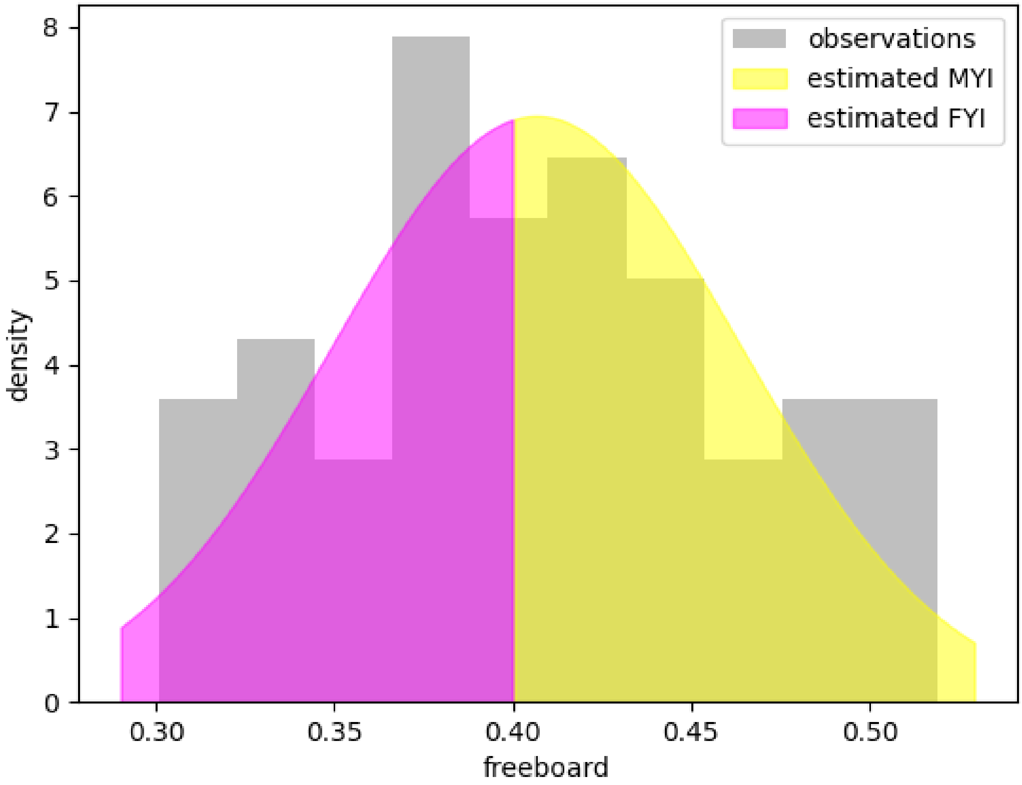

Fig: Soft labels are derived for one SAR pixel by assuming a Gaussian distribution (colored) of the 49 ALS observations (grey histogram) inside of it and then integrating the area under the pdf curve above and below the threshold. In the given example the probabilities are close to 50%.