Dear Sanggyun Lee,

Thank you very much for taking the time to review our manuscript and your constructive and supportive comments. They certainly helped us improve the quality and clarity of the paper. Please see our responses in blue below.

This study follows Dawson et al. (2022) and Landy et al. (2022), extending their work by adding more training samples and improving the CNN architecture. While the contributions are clear, there remain parts of the Methods section that require further elaboration to improve readability and reproducibility. I therefore recommend publication after a major revision.

General comments:

P4, L97: Since the addition of training samples is a key contribution of this study, I recommend providing a table summarizing the newly added samples, including their geographic region, month, and satellite source.

We agree that this is a key contribution and that all the listed information is relevant. You can find the novel training dataset on zenodo (as csv file) and linked to the paper in the assets. This table contains a binary flag indicating whether the sample was newly added or not, the coordinates, time, class label, names of the corresponding image and CryoSat-2 files and all input variables within the 11-point windows.

P4, L117: Could the authors clarify why a 7 km window was chosen?

As explained in the text, we average the airborne observations along 7-km sections to mirror the length scale used in obtaining a single freeboard estimate from CryoSat-2 for our summer CryoSat-2 processor (L117-118, L243). In the summer CryoSat-2 processor this window size was chosen, because it "[...] included a sufficient number of floe elevation samples to obtain realistic freeboard measurements even when there was high local variability in elevation." (Dawson et al. 2022, end of page 5).

P8, L172: Apparently, the lead and ice samples were manually extracted using visual inspection. However, as noted in Section 4.2, distinguishing between thinned floes and good floes is not straightforward in Sentinel-1 imagery. Could the authors comment on the potential impact of human error during the manual extraction of training samples, and how such uncertainty was mitigated?

We added the following explanations at the end of the section in the paper. "There is the possibility for human error in the manual extraction of training samples; however, several steps were taken to mitigate this. Most care was taken in the identification of leads, since the other ice floe classes are eventually combined in the freeboard processing. Leads were identified where there was any evidence for variation in CryoSat-2 SSH around the sample, and where the lead could be clearly visualized in the coincident image, from both a contrast in intensity and – importantly also – by the shape

of the feature. There is more chance for human error in the separation between ice floe classes, but these are radar classes: for all floes the sample was clearly an ice floe in the coincident satellite image, but it was only based on the patterns in the along-track CryoSat-2 parameters that they were separated into thinned floe, good floe and noisy floe classes."

P8, L175: If I understand correctly, Istomina et al. (2016) does not explicitly demonstrate that melt ponds and leads can be spectrally distinguished. It may be helpful for the authors to clarify how this reference supports their statement on spectral separation.

We replaced the reference with Istomina et al. 2025, where Figure 8 shows the albedo per wavelength, making it clear that melt ponds are generally more reflective, but especially in the blue wavelengths.

L8, 187: Since distinguishing leads from melt ponds is the most critical challenge in summer, could the authors clarify whether melt ponds are assumed to be entirely included in the noisy floe class? In addition, are refreezing ponds in August and September also included in this class? From a sea surface height (SSH) perspective, it may be difficult to separate thinned floes from refreezing ponds. It would be helpful if the authors could explain how refreezing ponds are treated in their classification.

Thank you for this comment and pointing out this confusion. We now made it more explicit throughout the paper, that only the differentiation between leads and floes is physically meaningful. All ice classes (thinned floe, good floe, noisy floe) should be regarded as radar detectable classes that might not have a clear physical meaning or where the mapping to their corresponding ice class is yet to be understood. To explicitly answer your question: Ponds and refrozen ponds could in principle be included in any of the floe classes, as they cannot be reliably detected in the CryoSat-2 or image data and therefore do not form a separate class in the training dataset. We separated leads and floes from the image data and assigned the three floe classes based on their unique signatures in the CryoSat-2 data. After classification, all floe classes are merged to calculate freeboard.

Figure 3: It would strengthen the manuscript if the authors could provide quantitative evidence that the addition of the thinned floe class improves the classification performance.

We did not detect a significant quantitative difference between using three or four classes (i.e. adding the thinned floe class) but decided to treat the thinned floe samples as a separate class, since the class reveals unique patterns in the along-track waveform parameters and even if we do not yet fully understand it physically, it opens up novel research avenues. Anyone using the CNN is free to experiment with the new class, to simply treat it as one of three ice classes or to discard it.

Figure 6: Since surfaces with diverse geophysical conditions also occur in July, August, and September, it would be helpful to include additional examples of CryoSat-2 tracks with coincident imagery from these months. This would further support the robustness of the classification.

Thank you for this suggestion. We added two additional examples from July and September.

Minor comments:

P1, L16: One reference is missing the publication year. Please correct this.

Done

Figure 1: The caption of Figure 1 is unclear as currently written. Please rephrase it to improve clarity.

Done

P8, L163-164: The text refers the reader to Dawson et al. (2022), but without further explanation it may be difficult to understand Figure 3 and the use of the 11-point window. A short description would be helpful.

Good point. We added a short explanation here: "For each sample, the anomalies of five parameters are extracted along an eleven-point window centered around the labeled sample to capture along-track anomalies."

P10, L209: Please specify whether pooling refers to max pooling or mean pooling.

We mean max pooling (l.207) and explicitly rephrased pooling to max pooling everywhere in the text now.

P10, L221: Did the authors also test the Adam optimizer? If so, was there any improvement in performance compared to RMSProp?

Yes, we also tried the Adam optimizer and found no improvement, so we stuck to RMSProp, which was also used by Dawson et al. (2022).

Table 1 and 2: Please place the captions above the tables rather than below.

Done

Table 1: For the ice user/producer accuracies, does this metric include both good and noisy floes together? Please clarify.

Yes, we clarified this in the text.