Response to reviewers

We thank the Editor and both Referees for their detailed and constructive comments. We have revised the manuscript substantially in response. The revised manuscript includes updated Sentinel-1 preprocessing, a rewritten labelling section, removal of CRF post-processing, a revised four-scene held-out evaluation, updated figures and captions, standardised model names, and a more cautious interpretation of the results as evidence of label efficiency rather than operational readiness.

The main manuscript-wide changes are as follows:

- All Sentinel-1 scenes were reprocessed using a consistent SNAP workflow including precise orbit correction, thermal noise removal, GRD border-noise removal, radiometric calibration to sigma-nought, Lee Sigma speckle filtering, ellipsoid correction, conversion to dB, and GeoTIFF-BigTIFF export.
- The test-set framing was changed from two broad scene types to four held-out Sentinel-1 scenes described by dominant visible ice–water structure: sparse linear leads, larger curvilinear lead system, fragmented floe field, and spatially distributed thin lead network.
- The labelling section was rewritten to define the task as binary open water versus ice/non-open-water segmentation. Thresholding is now described only as an initial candidate-generation step, followed by manual refinement and quality control against HH/HV imagery.
- CRF post-processing was removed entirely from the experimental comparison.
- Model names and training regimes were standardised throughout the manuscript: U-Net (SSL), U-Net (SL), U-Net (Control), Random Forest, and Segment Anything Model (SAM).
- The Results section was rewritten around average performance, scene-level MCC, HH-binned MCC diagnostics, qualitative model-output figures, and U-Net (SSL) versus U-Net (SL) difference maps.
- Supplementary Table S1 now provides the full scene-level MCC, F1 score, and IoU values for all models across the four held-out test scenes.

In the response below, **black** refers to original reviewer comments, red is our response.

Referee 1 Major Comments

Comment 1: Lack of thermal noise discussion

(1) Lack of thermal noise discussion:

Figures 7 and 8 show the classification results for both test scenes, obtained from the UNet SSL and the four baseline models. The visually most striking feature in almost all results is the S1 thermal pattern. Yet, thermal noise is not explicitly mentioned at any point in the paper. Especially the results for test scene 2 (Figure 8) are dominated by thermal noise patterns (scalloping and sub-swath boundaries), and it appears that the classification results must be highly dependent on how well each model learns these noise patterns. Although this is a well-known issue for S1 sea ice mapping, the authors do not include thermal noise correction in their pre-processing chain. At least, these patterns and their influence on the scores should be discussed.

We agree with the reviewer. The original manuscript did not adequately address thermal noise artefacts in Sentinel-1 EW imagery. In response, we strengthened the Sentinel-1 preprocessing workflow used for the revised analysis. All scenes were processed consistently with thermal noise removal and GRD border-noise correction before radiometric calibration, reducing the influence of known Sentinel-1 EW scalloping and sub-swath boundary artefacts on the model comparison.

The revised preprocessing workflow is now described in Sect. 2.2. The same workflow was applied consistently to all scenes before labelling, training, validation, and evaluation. We also revised the interpretation of the results to avoid implying that model behaviour is controlled only by physical ice–water structure where residual radiometric ambiguity may still contribute to segmentation errors.

Sect. 2.2 now describes the updated SNAP preprocessing workflow, including thermal noise removal and GRD border-noise correction. The qualitative figures were regenerated from the revised processing workflow, and the Discussion now treats the results as a controlled label-efficiency comparison rather than as an operational mapping product.

Comment 2: Figure quality

(2) Figure quality:

The quality of some figures in their present form is not sufficient and should be improved (in particular Figures 2, 7, and 8). Suggestions for improvement are given in the detailed comments below. Generally, I suggest including all figures as vector graphics instead of pixel graphics, as this will highly improve the quality especially when zooming in on details.

We agree and have revised the figure set. The previous two-scene qualitative-output figures have been replaced with four updated qualitative comparison figures corresponding to the four held-out test scenes. Figure captions and in-panel model names have been standardised to match the revised model nomenclature.

Where possible, figures were regenerated at higher quality and exported in vector-compatible formats. For figures containing SAR image panels, the SAR imagery necessarily

remains raster data, but annotation layers, legends, text, and layout elements were prepared to preserve clarity at publication scale. The difference-map legend was also regenerated using the final model names and categories.

Figs. 7–10 now show qualitative model outputs for the four held-out test scenes. Fig. 11 shows HH-binned MCC diagnostics across the four scenes. Fig. 12 shows U-Net (SSL) versus U-Net (SL) difference maps. Captions were rewritten to match the revised four-scene evaluation and final model names.

Comment 3: Limited training and evaluation

(3) Limited training and evaluation:

While the authors explain that "the primary contribution of the study lies in demonstrating relative performance gains under reduced annotation budgets, rather than maximising accuracy on a specific benchmark", I am sceptical that the results ("MCC improvement of 24% and 44%") are transferable to "more realistic scenarios" with significantly more training data. Generally, all performances presented in Figures 7 and 8 are still quite poor and heavily affected by thermal noise (see comment (1)), and any algorithm that is moving closer to operations needs to produce much more accurate and reliable ice-water maps, requiring additional training for the UNet SSL as well as the baseline methods. While it may be worthwhile to demonstrate the potential improvements by the SSL approach (as done in this study), the authors should probably discuss the transferability/scalability of the improvements when a lot more training data is used to achieve overall good enough results (with any method).

We agree with the reviewer and have revised both the experimental design and the interpretation of the results. The revised manuscript no longer presents the results as evidence of operational readiness. Instead, the central claim is that SAR-specific self-supervised pretraining improves label efficiency under constrained annotation conditions.

The evaluation has been expanded from the original two-scene framing to four held-out Sentinel-1 test scenes. These scenes represent contrasting visible ice–water structures: sparse linear leads, a larger curvilinear lead system, a fragmented floe field, and a spatially distributed thin lead network. All models are evaluated on the same four held-out scenes.

We now emphasise that the MCC improvement is modest, while the improvements in F1 score and IoU are clearer. The revised Discussion therefore interprets the result as improved label efficiency and average segmentation overlap, not as a large absolute accuracy gain.

We also agree that the relative advantage of self-supervised pretraining may change as much larger labelled datasets become available. This is now discussed explicitly as a limitation. The revised manuscript states that broader validation across larger labelled

datasets, additional Arctic regions, seasons, sensors, incidence-angle distributions, and ice regimes is required before operational deployment.

Sect. 3.3 now describes the revised experimental design. Sect. 4 reports results across four held-out test scenes. Sect. 5.1 and Sect. 5.4 have been revised to emphasise label efficiency, limitations, and the need for broader validation before operational use.

Comment 4: Random Forest performance

(4) Random Forest performance:

Although it is only a "baseline" algorithm for comparison, the performance of the RF algorithm is noticeably bad and concerning. The presented RF basically maps everything as sea ice without even identifying the dark leads in the test scene 1 (Figure 7). Although the deep-learning approaches are probably expected to perform better than a "traditional RF", this exceptionally bad performance is still quite surprising, since RF have been used quite successfully for sea ice mapping in previous studies (e.g. Park 2020, Lu 2023). Without any additional explanation, this raises concerns that (a) the RF and the texture features for it are not well designed and/or (b) that the training data is simply not sufficient for any reasonable assessment. I think at some discussion/explanation of this poor performance compared to RF in previous studies is needed.

We thank the reviewer for raising this point. In the revised manuscript, the Random Forest baseline has been reworked and is now presented as a more informative classical machine-learning comparator rather than as a collapsed or uninformative baseline.

Random Forest is now included as a classical pixel-wise machine-learning baseline trained using the same 13 labelled training scenes and 2 validation scenes as U-Net (SL). The feature set is described explicitly: normalised HH and HV backscatter, HH–HV difference, Sobel-derived edge magnitudes for both channels, and local 7 × 7 mean and standard-deviation features for HH and HV. These predictors provide intensity, polarisation-difference, edge, and local texture information while retaining an interpretable pixel-wise classification framework.

Under the revised preprocessing, labelling, and four-scene evaluation, the Random Forest baseline performs substantially more reasonably.

The revised Discussion reflects this interpretation. We now present Random Forest as a useful classical baseline that can recover some local ice–water structure, while noting that it does not learn broader spatial organisation in the same way as patch-based U-Net models.

Sect. 3.1.3 now describes the RF feature stack, sampling strategy, and training setup. Sect. 5.2 now discusses RF as an intermediate classical baseline based on manually defined

SAR-derived predictors, rather than as a collapsed baseline or a fully optimised operational RF sea-ice classifier.

Comment 5: Applicational motivation of the study

(5) "Applicational" motivation of the study:

It is not entirely clear whether the main applicational focus (i.e., why do we need improved automated sea ice mapping) is on navigation or environmental studies (or both). The initial mentioning of "maritime safety" and explanation of operational ice charting suggests an operational focus on tactical navigation, but later this is repeatedly mixed with the importance of leads for ocean-atmosphere interactions, which suggest an environmental focus. Both applications are valid and important, but they should be clearly distinguished in their explanations.

This is relatively easy to fix, so more a "general" than a "major" comment.

We agree. The Introduction and Discussion have been revised to distinguish the immediate methodological contribution from broader applications.

The revised manuscript presents the primary contribution as a label-efficiency study for SAR ice–water segmentation. This has operational relevance for sea-ice mapping and maritime safety because timely and consistent ice–water information is important for navigation. We retain discussion of leads and ocean–atmosphere interaction because these features provide physical and scientific motivation for high-resolution mapping, but they are now framed as broader scientific relevance rather than as the primary operational objective.

The revised manuscript states that additional validation, broader labelled datasets, uncertainty treatment, and appropriate spatial aggregation are required before such methods could support operational forecasting or navigation workflows.

The Introduction has been revised to clarify the operational motivation while retaining the scientific relevance of high-resolution ice–water mapping. Sect. 5.4 has been retitled "Limitations and implications for label-efficient sea-ice monitoring" and now explicitly frames operational use as a longer-term implication requiring further validation.

Comment 6: Roughness scales

(6) Roughness scales:

Roughness is always a relative term, depending on scale. Whenever mentioning roughness throughout the manuscript, the authors should specify the roughness scales they are thinking of: small-scale (on the order of the radar wavelength) or large-

scale (on the order of meters). Large-scale roughness could also be referred to as large-scale deformation and is directly related to ice type (deformed FYI or deformed MYI), whereas small-scale roughness can in fact vary significantly within one ice type, especially young ice with/without frost flowers or finger rafting. It might be worth to explicitly mention somewhere that the interplay of both small- and large-scale roughness effects contributes to the challenging interpretation of sea ice in SAR imagery.

Also easy to fix and rather "general" than "major".

We agree and have revised the physical interpretation of SAR backscatter in the Introduction. The revised manuscript now distinguishes between centimetre- to wavelength-scale surface roughness, which affects C-band radar scattering, and metre-scale deformation features such as hummocks and ridges, which are associated with deformed first-year or multi-year ice.

We also revised the discussion of sea-ice backscatter ambiguity. The manuscript now explains that multi-year ice is often thicker, salt-depleted, and characterised by metre-scale deformation that can produce bright, heterogeneous radar returns, whereas first-year ice tends to be smoother, more saline, and less consolidated, except where wavelength-scale roughness, frost flowers, rafting, deformation, salinity, or flooding modifies the radar response.

The Introduction now specifies the relevant roughness scales and links them to overlapping SAR backscatter signatures and the difficulty of binary ice–water segmentation.

Comment 7: Description of the two test scenes

(7) Description of the 2 test scenes:

The two selected test scenes are throughout the manuscript described as "consolidated ice pack" (scene 1, Figure 7) and "MIZ" (scene 2, Figure 8). The MIZ is "traditionally" defined as the area with 20-80% SIC, and more recently physics-based approaches as the "area affected by waves" are more common. Based on visual inspection of Figure 8, neither of these definitions makes me think that this image is in the MIZ. The authors should explain the reasoning behind this description of the test scenes and maybe consider to change them.

We agree. The revised manuscript no longer describes the test scenes as "consolidated ice pack" and "MIZ". The evaluation has also been expanded from two test scenes to four held-out test scenes. These are now described according to dominant visible ice–water structure rather than formal ice-zone category.

The four revised descriptions are:

- Scene 1: sparse linear leads;
- Scene 2: larger curvilinear lead system;
- Scene 3: fragmented floe field;
- Scene 4: spatially distributed thin lead network.

The manuscript explicitly states that these descriptions are used to interpret qualitative model behaviour and are not additional label classes. This avoids implying that Scene 2 is a formally defined MIZ.

Changes made: Table 2, the Results subsections, qualitative figure captions, HH-binned figure caption, and Discussion have been updated to use the revised scene descriptions. The old MIZ framing has been removed from the Results.

Referee 1 Minor Comments

Title and abstract: The term "sea ice segmentation" sometimes refers to ice-water mapping and sometimes to sea ice types (or sometimes both). Please consider a slight adjustment of your title to indicate that you are working towards separation ice and water (not sea ice types). Even after reading the abstract, this remains unclear.

We agree. The title and abstract have been revised to specify that the study addresses Sentinel-1 SAR ice–water segmentation. The abstract now defines the task as binary ice–water segmentation and distinguishes it from broader sea-ice-type mapping.

The title was revised to "Self-supervised learning reduces labelling requirements for Sentinel-1 SAR ice–water segmentation". The abstract now describes the task as Sentinel-1 SAR ice–water segmentation using manual binary ice–water reference labels.

Lines 21 and following: You introduce the term "UNet SSL" here but then keep referring to the "BYOL-pretrained UNet". Unless I misunderstand, these two terms refer to the same algorithm/model in your study. Please consider sticking with one single term to avoid possible confusion.

We agree. To avoid confusion, the revised manuscript now introduces and explains each model name in the Methods section before using the abbreviated names consistently in the Results, Discussion, tables, and figure captions. The BYOL-pretrained U-Net is defined as U-Net (SSL), the fully supervised model as U-Net (SL), and the randomly initialised low-label baseline as U-Net (Control). Random Forest and Segment Anything Model (SAM) are also defined before their abbreviated names are used.

Lines 41-42: The cited numbers are from 2018, which is by now 8 years ago. Please consider presenting more recent numbers, especially since the decline in September extent has significantly slowed in the past years (see e.g.

https://www.meereisportal.de/en/maps-graphics/sea-ice-trends#gallery-1 or attached png)

We revised this section to avoid relying on a single dated numerical trend statement and instead frame the motivation in terms of sustained changes across the satellite record.

Lines 49-50: Time lag is one issue, but also subjectivity of the analyst and increased data availability overall with more sensors being launched (-> more analysts needed)

The Introduction was revised to include analyst subjectivity and increasing data volume as additional motivations for automated and label-efficient SAR ice–water mapping.

Lines 50-53: This statement should be more clearly formulated. The sea ice charts don't really lack details of leads or ridges because SAR products cannot resolve them, but rather because mapping individual leads manually is too time consuming in the manual operational production chain of most services. Hence, many ice charts include lead information in the form of young ice fractions in the egg codes each polygon in the chart. However, even rather simple automated products can in fact capture individual leads quite well, of course still limited by the sensor resolution (~90x90m for S1 EW) (e.g. Johansson 2018, Murashkin 2019, Lohse 2024).

Also, while the authors are right that leads or deformation zones are important for ocean-atmosphere interaction, this statement does not really fit into the context of ice charts. Here, the leads are important for safe and efficient for navigation and route planning.

We agree. The revised manuscript no longer frames the issue as a limitation of SAR imagery itself. Instead, it clarifies that the limitation is primarily associated with manual chart production and labelled-product generalisation. The manuscript now states that SAR imagery can resolve fine-scale ice–water features within sensor-resolution limits, but that manually delineating such features across large scenes is time-consuming.

Line 57: Please quantify the size of fine-scale features. Compared to other sensors, SAR very good at resolving fine spatial scales (although of course still limited by pulse and Doppler bandwidth, i.e. spatial resolution). If you are referring to the lack of individual leads in labelled data such as ice charts, consider specifying that this is an "ice chart issue" and not necessarily a "SAR issue".

The limitations section now discusses the 80 m spatial resolution and the uncertainty associated with narrow leads and mixed pixels.

Line 71: Please specify: Do they struggle with the separation between these two ice types, or with separating these ice types from other types? Consider explaining why.

We clarified this point. The revised Introduction now states that previous models struggled with separating young ice and first-year ice from other ice or water classes, particularly

where labels are coarse or inconsistent and where mixed SAR pixels occur near manually drawn chart-polygon boundaries.

The Introduction now links this difficulty to label quality, mixed pixels, and overlapping SAR backscatter signatures.

Line 103: Please specify roughness scale. I assume you are talking about "small-scale" surface roughness here. Maybe add "(cm-scale or wavelength-scale)" or somethings similar to avoid confusion with large-scale deformation (sometimes also called roughness).

The Introduction now specifies the relevant roughness scales and links them to SAR backscatter ambiguity in binary ice–water segmentation.

Lines 129-131: "heavily deformed", "smoother", "increases roughness". Please clarify roughness scales for the different statements.

The revised Introduction now distinguishes centimetre- to wavelength-scale roughness, which affects C-band scattering, from metre-scale deformation such as hummocks, ridges, and rafting, which is associated with deformed sea ice. We also clarify that both scales can contribute to ambiguous SAR backscatter.

Lines 134-135: See general comment (5): Until here I was under the impression that the main application focus is on navigation support. If you want to keep both navigation ("automated ice charting") and environmental studies ("lead detection to study energy balance") for your motivation, I suggest mentioning both of them quite early on and explaining that "accurate ice type mapping is required for a range of applications, including support of safe navigation as well as environmental studies of ocean-ice-atmosphere interactions" (or something along those lines).

The Introduction and Sect. 5.4 were revised to distinguish operational relevance from broader scientific applications.

Line 138: Overlapping backscatter signatures from which surface types?

The Introduction now explicitly lists the relevant surface types and conditions that can produce overlapping backscatter signatures.

Lines 145-150: If I understand correctly, this reads like it should be a list of 2 research questions, but the paragraph/line break between them seems strange. I suggest listing them as two bullet points or even numbering them as research goals (1) and (2) which you can then explicitly refer to later.

The end of the Introduction now lists the study's research questions in bullet form.

Lines 152-156: See general comment (3): This "relative" comparison of the different models makes sense to some extent. However, I am wondering if the relative improvement that you demonstrate later will also hold if you overall use much more

training data, which will be needed to achieve better results. In practice, you would probably never use a deep-learning approach for ice-water mapping trained on only seven images. I would like to see this commented on in the discussion.

Sect. 5.1 and Sect. 5.4 now discuss label efficiency, scalability, and the need for broader validation.

Lines 171-172: Something missing in the sentence; maybe a "-" after pack ice?

The relevant sentence was edited for grammar and readability.

Figure 1: Legend says "Sea Ice Concentration (m)", should probably be "(%)". I also find the legend entry "Label Extents" slightly confusing. I think you are showing the footprints of the S1 EW scenes used in the study? Please consider changing the label to "S1 footprints" or something similar.

Maybe also consider colour-coding footprints as "test scenes", "full training set (7 images)", and "small training set (3 images)" or similar.

Fig. 1 and its caption were revised to clarify the sea-ice concentration unit and the meaning of the Sentinel-1 scene footprints.

Sentinel-1 SAR imagery:

This entire section needs some clarification.

- **The presentation of the image size and spatial resolution in its current form is confusing. S1 EW in GRD format comes originally at 40x40m pixel spacing with an actual resolution of approximately 97x93m. A single scene covers about 400x400km. After geocoding to 80m pixel spacing (not the same as resolution, especially not since you speckle filter and multi-look during pre-processing, both of which reduces the effective resolution), you end up with an image size of 7000x7500 pixels, corresponding to 560x600km. This is not the swath width, which is fixed to ~400km by the acquisition geometry, but just the full extent of your geocoded image.**

- **In your pre-processing chain, you apply a Lee Sigma speckle filter followed by additional multi-looking. Please explain the need and benefit of doing both.**

- **What is GRD border noise? Do you perform thermal noise removal? If no, why not?**

- **Calibration to sigma_0 does not mitigate IA effects. In fact, sigma_0 remains significantly dependent on IA, with typical sea ice slopes between 0.1 and 0.3 dB/1deg and open water slopes up to 0.7dB/1deg for HH (e.g., Mäkynen 2017, Mahmud 2018, Lohse 2020, Geldsetzer 2023)**

- **Please clarify how the use of dual-pol data "implicitly handles residual IA effects". You are right in the sense that HV is much less dependent on IA than HH, due to predominantly different scattering mechanisms. But HH remains dependent on IA,**

with different slopes for different surface types. Consider adding this to your explanation.

- **The deep-learning approaches you are using inherently consider contextual information; hence I think you can get away with ignoring IA effects for these methods. "Traditional" approaches like the RF, however, should somehow account for IA effects.**

We revised and simplified the Sentinel-1 data-processing description. The manuscript now describes the final preprocessing workflow rather than mixing original sensor geometry, geocoded image extent, and effective resolution in a confusing way. The revised workflow includes precise orbit correction, thermal noise removal, GRD border-noise removal, radiometric calibration to sigma-nought, Lee Sigma speckle filtering, ellipsoid correction, conversion to dB, and GeoTIFF-BigTIFF export.

We also removed wording that implied sigma-nought calibration removes incidence-angle effects. The revised manuscript treats residual incidence-angle dependence as part of the SAR variability that affects ice–water segmentation. Incidence angle is incorporated directly into the initial label-generation threshold, and model performance is further examined using HH-binned diagnostics. We do not claim that dual-polarised inputs fully remove incidence-angle effects; rather, HH and HV provide complementary information.

Sect. 2.2 was revised to describe the final preprocessing workflow, including thermal noise removal and border-noise correction. Sect. 2.3 now describes incidence-angle-aware label initialisation. Sect. 4.3 and Fig. 11 present HH-binned MCC diagnostics.

Lines 229-230: Figures should be numbered in order of appearance. You refer to Figures 7 and 8 before Figure 2.

We agree. Figure references have been reordered and updated.

Lines 245-249: Please add acquisition date (season) for validation scenes. I am aware that they are in Table 1, but I think it is worth to repeat them in the text here.

Sect. 2.2 now states that the U-Net (SSL)/U-Net (Control) experiments used one validation scene acquired on 16 December 2022, while the U-Net (SL)/Random Forest experiments used two validation scenes acquired on 20 November 2022 and 15 November 2021. These scenes were used only for model development and were excluded from the four held-out test scenes.

Figure 2: I appreciate that you are showing the example of the training data and compare to other already published sets. While the advantages of your detailed manual labelling compared to the other methods become clear, the figure in its current form needs multiple changes/improvements:

- **Panels (a) and (d) (and respectively (b) and (e), (c) and (f)) should match exactly in scale and extent. Currently there is some offset in extent and maybe scale.**

Additionally, some of the floe shapes almost look distorted, e.g. the rather large floe in the centre bottom of (a) and (d).

- **You are not showing the full images, which is fine, but you should add a scale for reference.**

- **Add information on what channel we are looking at. Maybe consider a false-colour representation combining information from both HH and HV.**

- **Consider adjusting the dynamic range of the SAR intensity visualization. Currently all the sea ice in (a) looks very homogeneous, like one single ice type.**

- **Consider adjusting adjust the colour scheme for the segments in (e) – they are hard to distinguish**

- **A lot of the "leads" in the lower right part of (f) are likely rather deformed ice. Impossible to say for sure without knowing which channel we are looking. I am aware that this data set is not from this study and mostly shown for comparison, but from this presentation here I am doubtful of the quality of the labels.**

We agree with the reviewer and have revised Fig. 2 and its caption to address these points. The figure is now presented explicitly as a comparison of label provenance and spatial granularity across three datasets, rather than as a validation comparison between datasets. Each SAR panel is paired with the corresponding label panel directly below it, and the SAR/label pairs within each column now use the same crop extent and scale. Scale bars have also been added for spatial reference.

We have added explicit channel information to the figure caption. The SAR panels are shown as Sentinel-1 HH backscatter. The SAR display stretch was adjusted to improve visual contrast within the examples, while retaining the purpose of the figure as a label-comparison panel rather than a quantitative backscatter analysis. The label presentation was also revised to improve readability, particularly for the AI4Arctic chart-derived labels.

We have also revised the caption to clarify the status of the comparison datasets. The AI4Arctic and MOSAiC examples are included to illustrate differences in label source, spatial granularity, and dataset objective. They are not used as validation references in this study. The caption now states that AI4Arctic labels are derived from operational ice charts and therefore show spatial generalisation consistent with regional charting objectives, while MOSAiC labels are CNN-generated binary lead labels from the central Arctic Ocean. This wording directly addresses the reviewer's concern that some MOSAiC-labelled "leads" may represent deformed ice or ambiguous ice conditions, by making clear that the panel is illustrative of label provenance rather than evidence used for model evaluation.

Changes made: Fig. 2 and its caption were revised to show matched SAR/label crop extent and scale within each dataset column, include scale information, specify that the SAR panels show Sentinel-1 HH backscatter, improve SAR and label readability, and clarify that

the AI4Arctic and MOSAiC examples are included for label-provenance and spatial-granularity comparison rather than validation.

Figures 7 and 8: (commented here because they are referred to first here. Some of the comments below relate to the results part of the figures)

The figure quality in its current form is not good. Since you are showing a range of different panels, I do not see the need to rotate the figure 90deg, making the individual panels smaller and leaving half the page empty. Also please insert vector graphics to maintain better quality when zooming into details.

The manually selected labels look convincing.

The performance of the RF makes me question the training and design of the RF and whether it can be considered a fair comparison, please see general comment (4).

Finally, we see a lot of thermal noise effects across the classification results of all UNet approaches in Figure 8, please see general comment (1)

Figs. 7–10 now show qualitative model outputs for the four held-out test scenes. Figure captions and legends have been updated, and the old two-scene/MIZ framing has been removed.

Line 344-345: What exactly do you mean by selecting scenes based on "quality"?

We revised the dataset description to avoid vague language. The manuscript now describes the dataset in terms of its role in pretraining, training, validation, and held-out testing rather than using an undefined "quality" criterion.

Lines 374: Consider rephrasing "raw HH and HV backscatter" to "HH and HV backscatter intensities" or a similar more precise description. "Raw" backscatter in SAR usually refers to the unfocused image.

We agree. The RF description has been revised to refer to normalised HH and HV backscatter rather than "raw" backscatter.

Lines 373-377: Good to see that you are using texture features in the RF, this will make the comparison fairer. Please add information as to choices and design of texture features, e.g. GLCM parameters such as distance, angle, window size, discretisation. These choices are critical for good ice type separation (e.g. Zakhvatkina 2017, Karvonen, 2017, Park 2020, Lohse 2021, Khachatrian 2021).

The RF method description has been revised to specify the final feature set used in this study. The revised RF baseline uses normalised HH and HV backscatter, HH–HV difference, Sobel-derived edge magnitudes, and local 7 × 7 mean and standard-deviation features for HH and HV. These provide local intensity, edge, and texture information within an interpretable pixel-wise baseline.

Lines 386-388: Please specify the scaling (min/max values) when setting HH, HV, and HH/HV to 8-bit RGB channels.

We agree. The SAM preprocessing description has been revised to specify the pseudo-RGB conversion and clipping ranges.

Experiment design: Most commonly, I would think that you split the 9 labelled images, into 3 sets: train, test, validation. What you call the test set (the two images kept aside) would then be the validation set, whereas the remaining 7 images would be split into training (to fit the model weights) and testing (to avoid overfitting). Please comment/specify why you decide to only split into 2 sets and how you avoid overfitting.

Figure 3: The visualization seems to be missing some connections, e.g. the "labelled dataset" is also the input for the RF and SAM, not just the for the UNet (SL).

Also, colours for "SAM" and "Compare Models" appear very similar, please consider adjusting one of them.

Sect. 3.3 now describes the final training, validation, and held-out testing design. Table 1 summarises the dataset split by model configuration, and Table 2 lists the four held-out test scenes.

Figure 4: The label "HH, HV, 1024" for the input layer does not seem entirely accurate. I assume the three layers shown are HH, HV, and training labels, while the size of the input patches is 1024x1024?

Fig. 4 and its caption were revised to clarify that the model input consists of two-channel HH and HV patches of size 1024 × 1024 pixel

Figures 5 and 6: The main message that the reader is supposed to get from these figures is not entirely clear to me. You refer to the figures in the "model intercomparison" section on page 13 (lines 327-343), but I don't really understand what I am supposed to learn from these. I am sure there was some clear idea of why to show them, please consider explaining the main message more explicitly.

Sect. 3.1.2 and the captions for Figs. 5 and 6 now explain the purpose and limitations of the BYOL feature visualisations.

Model performance across ice types and HH backscatter: In addition to the different ice regimes, I think you need to associate backscatter bins with different IA regimes. E.g. in Figure 7, the overall decrease of HH sigma_0 across the swath is clearly visible. This should at least be included in the discussion of the results presented in Figure 9. (Please see also previous comment on IA sensitivity of sigma_0 in data section). Generally, you should be careful with any over-interpretation of this figure, since you do not account for HV at all in this analysis. However, based on the noise patterns in Figures 7 and 8, we see a clear influence/contribution of HV.

Sect. 4.3 and Fig. 11 were revised to present HH-binned MCC as a diagnostic complement to the scene-level results. The qualitative figures include both HH and HV panels for each test scene.

Lines 559-562: Please be careful to make sure that these interpretations are only valid for the shown test scene. I don't think you can generally associate the ice-water transition with a strong contrast in HH sigma_0, as sigma_0 is ice-type dependent and, more importantly, for open water highly wind-state dependent.

Sect. 4.3 was revised to use more cautious language and to avoid broad claims about general ice–water contrast in HH.

Lines 563 -572: The result description and discussion of MCC-vs-HH(dB) graphs for test scene 2 must include thermal noise, which is clearly visible in the results of all methods (Figure 8), both as scalloping in sub-swath EW1 and clearly at sub-swath boundaries. Due to its lower signal strength, HV is much more affected by thermal noise and should therefore be included in the visualization in Figures 7 and 8. A lot of the differences associated with the HH bins in Figure 9 may in fact be caused by noise effects or variation in HV, which is neither shown nor discussed here.

Figure 9: Better figure quality than many of the other figures. However, while this presentation of the MCC may contain useful information, I don't think it can be interpreted without accounting for the HV channel.

Sect. 2.2 now includes thermal noise removal and border-noise correction. Figs. 7–10 include both HH and HV panels. Sect. 4.3 and Fig. 11 present HH-binned MCC as a diagnostic complement to the main evaluation.

Lines 637-543: Although the limitations of RFs are pointed out correctly, I remain puzzled by the fact that the RF in this study almost completely fails to detect even the very dark (in HH) lead structures. Some additional knowledge and discussion of the parameters for the computation of textural features might help to explain this.

The RF baseline has been revised and now performs more reasonably as an intermediate classical baseline. The revised manuscript describes the RF feature stack, local-window features, sampling approach, and training configuration. The Discussion no longer treats the RF result as a collapsed model; instead, it describes RF as a useful classical comparator that remains below the U-Net-based models on average.

Subsection 5.3: If I understand this section correctly, I would not call it a "comparison". It rather provides reasoning why the alternative self-supervised models are not implemented and tested in this study. Maybe consider rephrasing it in the discussion, or whether it could be moved into the introduction and method section, strengthening the reasoning for the choice of BYOL.

Sect. 5.3 was revised to focus on the methodological rationale for BYOL and its relation to other self-supervised approaches.

Line 690: The Park (2020) study cited here is a good example of a RF classifier producing much better results than the RF in this study.

Lines 742-755: Please be careful to choose your wording such that this is only necessarily true for your two example images. The general statement of better contrast between ice (“bright”) and leads (“dark”) in the consolidated pack ice region compared to more overlapping signatures in the MIZ may be true for the 2 examples discussed here but does not necessarily hold generally. Even within the pack ice, wind-roughened leads may appear bright in HH and significantly overlap with sea ice backscatter signatures. HV will then be critical to distinguish ice and water. On the other hand, brash ice (heavily deformed -> strong backscatter) and calm water (in still wind conditions) can be easily separable in the MIZ and close to the ice edge. I think you should be more careful with general statements on the differences in model performance and ice-water separability based on the two examples selected here.

The old MIZ framing and broad two-scene contrast statements were removed. Table 2, Figs. 7–10, Fig. 11, and the Results subsections now use the revised four-scene descriptions.

Referee 2

Comment 1: Test labels may identify newly formed or young ice rather than open water

1. The labels on the test images are incorrect. They do not show open water, but new and young ice freshly formed in the leads. The pattern visible on the backscatter from ice in the leads clearly shows a cyclic evolution of freeze-up with several bands of lower and higher backscatter along the lead edge.

It is highly likely that the U-nets were trained on incorrect labels and are predicting not the ice/water classification, but the "older ice" / "younger ice" classification. In that case, both the Control and SL U-nets are performing better on the second image, where a clear distinction between older, rougher, brighter and younger, smoother and darker ice is visible. And the SSL U-net derives 'ice' which should be interpreted as 'old ice' in the pixels with obvious first-year ice. In this case, the main conclusion that unsupervised pretraining increases accuracy is actually compromised.

To prove that unsupervised training improves the classification accuracy, the training and the testing labels should be correct and consistent. It is highly recommended to include a sea ice expert in the team.

Moreover, two test scenes are obviously not enough to draw sound conclusions. As demonstrated by the authors, it is easy to create a solid argumentation based on false

results from a few samples. Large statistics reduce chances. It is also recommended to include scenes with true MIZ (see comment 3) and share the training / testing labels online (e.g., Zenodo) for simplifying the review process. Alternatively, the existing, widely used datasets with SAR imagery and ice-chart based labels (e.g., AutoIce) can be safely used.

We thank the reviewer for this important comment. We have substantially revised the labelling section to define the classification task and labelling workflow more clearly.

The revised task is binary open water versus ice/non-open-water segmentation. The manuscript now explicitly states that the labels are not intended to provide a multiclass separation of ice type, ice concentration, young ice, first-year ice, multi-year ice, melt ponds, leads, or marginal-ice-zone structure. In the final masks, class 1 denotes open water and class 0 denotes ice or non-open-water.

Initial open-water candidates were generated using a conservative thresholding procedure based on HH, HV, and local incidence angle. A pixel was initially classified as open water only where both conditions were satisfied: $HV \leq -25$ dB and $HH \leq 0.25 - 0.66 \times IA$, where IA is the local Sentinel-1 incidence angle in degrees. This threshold was used only to accelerate annotation and improve initial mask consistency. It was not treated as the final label.

All automatically generated masks were converted to polygons and manually inspected in QGIS alongside the HH and HV SAR imagery. Manual interpretation overrode the threshold rule where required. Particular attention was given to low-backscatter areas where calm open water, thin ice, newly formed ice, and young ice may overlap in SAR intensity. Final class assignment in such regions was based on HH/HV backscatter, local spatial context, lead geometry, surrounding ice texture, and consistency across the scene, rather than on the threshold rule alone.

This revision directly addresses the reviewer's concern by making clear that dark leads or low-backscatter pixels were not automatically labelled as open water. Ambiguous regions with spatial structure consistent with newly formed ice or young ice were treated conservatively and manually corrected where necessary.

Sect. 2.3 has been rewritten to describe the binary class definition, incidence-angle-aware threshold initialisation, manual refinement workflow, and treatment of ambiguous low-backscatter areas. Fig. 2 has been retained as a label-provenance and label-granularity comparison.

Comment 2: Per-image z-score normalisation, incidence-angle effects, and thermal noise removal

2. In the preprocessing step, the samples are normalised using a z-score statistic computed on a per-image basis (L. 431). This normalisation scales the sigma0 values from dB to relative units which are dependent on a range of values on each image. As a result, the networks never see the same values of backscatter for the same type of surface (water, young ice, older ice). That may explain why the SSL network works worse (or better, from the author's perspective). As it was trained on more images than the Control and SL, it probably learned to ignore the magnitude of the normalised signal, whereas the Control and SL networks rely only on the magnitude. As the contrast increased on the second image, the magnitudes of the signal from the older ice and younger ice became very different and the Control and SL networks easily distinguished them, whereas the SSL network failed. Figure 9, supports my hypothesis that the image-based scaling changes performance of the networks from image to image. If the images were normalized using a fixed scaling, the HH dependence of accuracy should be very similar on two images. It is recommended to normalise the images, but the z-score statistics should be computed from many images and kept fixed. Moreover, it is better to perform incidence angle correction and proper thermal noise removal to exclude these factors from the study. Otherwise, it difficult to say what is actually playing a larger role in the undertrained network performance - is it a super sensitivity to noise or other factors.

We thank the reviewer for this detailed comment. We agree that SAR normalisation choices can influence how absolute backscatter magnitude is represented to the models. In the revised manuscript, we clarify that the same preprocessing, valid-pixel handling, and normalisation strategy was applied across the U-Net experiments so that differences between U-Net (Control), U-Net (SL), and U-Net (SSL) reflect the training regime and pretraining strategy rather than differences in input preparation.

We have also addressed the reviewer's concern about thermal noise directly by reprocessing all Sentinel-1 scenes using a workflow that includes thermal noise removal and GRD border-noise correction prior to radiometric calibration. This reduces one of the major scene-dependent radiometric artefacts identified by the reviewers.

Regarding incidence angle, the revised manuscript treats residual incidence-angle effects as an important source of SAR backscatter variability. The labelling workflow now incorporates local incidence angle explicitly in the initial open-water candidate-generation step. Rather than applying a fixed HH threshold, initial candidates are generated using an incidence-angle-dependent HH condition together with an HV constraint. This does not claim to remove all incidence-angle effects from the imagery, but it makes the label-initialisation step physically more consistent across the Sentinel-1 EW swath.

In response to the reviewer's comment, we revised the normalisation procedure so that HH and HV inputs were standardised using fixed channel-wise z-score statistics computed from the training scenes and applied unchanged to validation and held-out test scenes.

Sect. 2.2 now describes the revised preprocessing workflow, including thermal noise removal and GRD border-noise correction. Sect. 2.3 describes incidence-angle-aware label initialisation. Sect. 3.3 now states that HH and HV inputs were normalised using fixed channel-wise z-score statistics computed from the training scenes and applied unchanged to validation and held-out test scenes.

Comment 3: No MIZ on the second test image

3. There is no MIZ on the second test image. Not because there is almost no open water, as discussed in comment 1, but because MIZ separates pack ice from the open ocean. Good examples of scenes with MIZ: A very wide MIZ. All ice is broken into 10 - 300 m floes:

S1C_EW_GRDM_1SDH_20260224T180149_20260224T180254_006501_00D19C_9D7E

A relatively narrow MIZ separates the pack with a mix of older and younger ice from open

water:

S1A_EW_GRDM_1SDH_20260224T073006_20260224T073106_063358_07F521_E549

It is highly recommended to include a sea ice expert in the team. It is also recommended to

test the networks in different ice conditions, including open water scenes.

We agree. In the revised manuscript, the scene previously described as a MIZ has been removed from the held-out test set. The revised evaluation no longer uses the old two-scene "consolidated ice" versus "MIZ" framing. Instead, it uses four held-out scenes described by dominant visible ice–water structure rather than by formal ice-zone category.

The four revised test-scene descriptions are sparse linear leads, larger curvilinear lead system, fragmented floe field, and spatially distributed thin lead network. The manuscript also states that these descriptions are qualitative descriptors used to interpret model behaviour, not additional label classes or formal ice-zone categories.

The MIZ terminology has been removed from the test-scene descriptions, Results subsection headings, figure captions, and Discussion framing. Table 2 now lists the four held-out test scenes using geometry-based descriptors.

Referee 2 Minor Comments

1. The Control UNet was trained on only two images, whereas the SSL on seven plus two images. That makes the comparison of these two networks unfair. I would rather train only two networks (SL and SSL) on the same amount of SAR images excluding labels from some samples for the SSL. That also enables an experiment - how many labels can be dropped to achieve a similar result.

The comparison with U-Net (SL) is now explicitly framed as a label-efficiency comparison rather than a same-training-set benchmark. U-Net (SL) uses a larger labelled-data configuration, with 13 labelled training scenes and 2 validation scenes. U-Net (SSL), by contrast, uses 39 unlabelled scenes for self-supervised pretraining and then only 9 labelled training scenes and 1 validation scene for supervised fine-tuning. This design tests whether SAR-specific self-supervised pretraining can achieve comparable or better performance with fewer labelled scenes. We also revised the U-Net (Control) experiment so that it uses the same 9-scene labelled configuration as U-Net (SSL), allowing the effect of BYOL pretraining to be isolated from the effect of the U-Net architecture itself. Sect. 2.2 and Table 1 now define the revised data split by model configuration. Sect. 3.1 and Sect. 3.3 explain the roles of U-Net (SSL), U-Net (Control), and U-Net (SL). Sect. 4 reports average model performance and scene-level MCC for the revised four-scene evaluation. The Results and Discussion now explicitly frame the study as a label-efficiency comparison rather than a same-training-set benchmark.

2. Although the training procedure description is quite verbose, it is difficult to understand without a flowchart or a concise algorithm description. I would recommend adding an Algorithm (https://www.overleaf.com/learn/latex/Algorithms) with notations of the online and target networks.

Sect. 3.1.2 was revised to clarify the BYOL online/target network training procedure. Fig. 3 and its caption were revised to summarise the training and evaluation pathways for U-Net (SSL), U-Net (Control), U-Net (SL), Random Forest, and SAM. Sect. 3.3 now provides a concise experimental-design description linking pretraining, training, validation, and held-out testing.

3. The CRF post-processing may significantly alter the results of classification depending on spatial distribution of probabilities returned by different networks. It is better to exclude it from the experiment, as it adds an unknown factor (similar to varying thermal noise, etc.) influencing the accuracy.

All CRF-related Methods text, Results interpretation, and Discussion claims have been removed. The revised Methods and Results now describe and evaluate the models without CRF post-processing.

4. Figure 3 needs improvement. The arrows are too long, leaving too much space between too small boxes with too small text (for my eyes).

Fig. 3 and its caption were revised to improve readability

5. Figures 7 and 8 need some improvements:

*** Labels on the images should be rotated to view the figure in the portrait layout,**

*** both HH and HV should be presented,**

*** the backscatter should be presented after normalisation with a colorbar**

Figs. 7–10 now show the qualitative model outputs for all four held-out test scenes. Each figure includes HH and HV input panels, manual ground truth, and model predictions. Captions were rewritten to describe the panel order, model training regimes, and binary mask convention consistently across all four figures.