We thank both reviewers and the editor for their helpful comments to improve our manuscript. We have addressed each comment in detail below, but we would like to highlight the major improvements here:

- Hyperparameter tuning: We introduced hyperparameter tuning in order to determine the maximum achievable accuracy for each baseline model. All accuracy values for the best model were updated after hyperparameter tuning. The greatest improvement in accuracy was achieved for the U-Net model after hyperparameter tuning.
- Uncertainty: We provide the standard deviation (std) between the three best model runs during hyperparameter training to show the differences in model performance between runs (see Table 5). Furthermore, we provide confidence intervals for the performance metrics (see Table 6).
- Extended Figure 1 to include all model developments especially current ones including foundation models.

Kind regards,
Celia & Amy (in-behalf of all Co-Authors)

Review 2

1. The authors have introduced a new dataset which is very valuable, but the gap and novelty could be stated more clearly. It is required to explain what limitations of existing datasets are addressed and what new capabilities this dataset provides.

Thank you for raising this point. We emphasised the research gap more clearly and rephrased the paragraph as follows:

“Although open-access satellite data now enables continuous monitoring through automated methods, existing deep learning models for calving front delineation exhibit spatially variable accuracy, with significant error margins persisting in certain regions (Baumhoer et al., 2019; Heidler et al., 2021). Developing robust deep learning models for Antarctic coastline delineation requires high-quality, representative benchmark datasets to support robust model training and enable standardized algorithm comparison. While benchmark datasets now available for marine-terminating glaciers (Gourmelon et al., 2022; Lu et al., 2025), no equivalent standardized benchmark dataset currently exists for Antarctic ice shelf front delineation. The impact of such datasets is evident on other domains of polar research. For example, the AutoICE benchmark dataset fostered model development and community challenges to improve sea ice mapping (Chen et al., 2024; Jalayer et al., 2025; Stokholm et al., 2024). Also for Greenlandic outlet glacier termini, publicly available benchmark datasets have substantially accelerated methodological progress and model development in recent years (Gourmelon et al., 2025b; Wu et al., 2024; Zhao et al., 2025). In contrast, automated deep learning-based mapping of Antarctic ice shelf fronts remains limited (Baumhoer et al., 2019, 2023; Heidler et al., 2021), underscoring the urgent need for a standardized Antarctic benchmark dataset Such a dataset is particularly important given the vastly greater extent of the Antarctic coastline compared to Greenland’s marine-terminating glacier termini and the critical role of ice shelf retreat in future sea-level rise. “

The novelty is now better highlighted in the following paragraph we modified: “As the first standardized, large-scale benchmark dataset specifically designed for Antarctic ice shelf front delineation, Shelf-Bench addresses a major gap in both glaciology and Earth observation machine learning. Because the dataset is built on open-access satellite imagery, glaciologists can leverage models trained on Shelf-Bench to continuously monitor ice shelf front positions in future satellite acquisitions. These models are directly transferable to current and future C-band SAR missions, particularly Sentinel-1, with its consistent global coverage, enabling sustained monitoring of Antarctic ice shelf fronts well into the future. For the computer vision and machine learning community, the curated nature and accessibility of Shelf-Bench establish a novel benchmark for evaluating and adapting existing deep learning architectures to novel cryospheric applications. Furthermore, as foundation models become increasingly prevalent in Earth observation Shelf-Bench provides one of the first extensive annotated resources tailored to the Antarctic coastal environment. By providing a standardized, multi-purpose benchmark for both methodological development and operational deployment, Shelf-Bench enables more consistent model comparison, facilitates reproducible research, and advances progress toward continuous automated monitoring of Antarctic ice shelf dynamics in the era of persistent satellite observation. “

2. It would be helpful to add benchmark datasets from the Arctic, such as the AutoICE dataset and its associated challenge to section (2. Background and Related Work) and cite some references including (https://doi.org/10.5194/tc-18-3471-2024, https://doi.org/10.1016/j.rsase.2025.101538, https://doi.org/10.5194/tc-18-1621-2024). This would provide a more complete overview of existing benchmark efforts in polar regions.

Thank you for pointing us to the AutoICE benchmark dataset. We wanted to focus explicitly on ice shelf front datasets and mapping methods in the section “2. Background and Related Work”. To still highlight that benchmarks like AutoICE foster model development we added the sentence in the introduction: “For example, the AutoICE benchmark dataset fostered model development and community challenges to improve sea ice mapping (Chen et al., 2024; Jalayer et al., 2025; Stokholm et al., 2024) “

3. Need to clarify how the inputs are combined before being fed into the models. The authors have mentioned about resampling to a common resolution, but more detail on the interpolation method and data representation need to be included to improve clarity. In addition, it is not clear whether any fusion strategy (like early or late fusion across sensors) is used. A brief discussion or simple comparison could improve the study. Exploring different fusion approaches has been shown to improve performance and relevant works on multi-resolution data fusion for sea ice mapping (https://doi.org/10.1109/igarss55030.2025.11243075) can be cited here and used as a basis for additional experiments.

Sorry for the vague explanation regarding spatial resolution. We do not apply any form of sensor fusion, as all inputs consist of single-modality SAR imagery without integration of additional data sources. SAR pre-processing includes standard multilooking and geocoding steps as provided in the respective products. For Sentinel-1 GRD EW data, we use the preprocessed ground-range detected imagery with an effective pixel spacing of approximately 40 × 40 m. For ERS and Envisat data, the processed imagery used in this study has an effective pixel spacing of approximately 30 × 30 m.

Although the datasets therefore exhibit different spatial sampling resolutions, this does not constitute image fusion or multi-sensor feature integration. Training samples are extracted as fixed-size patches defined in pixel units. Consequently, the physical ground area covered by a patch varies between sensors. This introduces a scale difference in absolute spatial extent (e.g., icebergs appear at different physical sizes across sensors), which may influence model generalization but does not require explicit sensor harmonization beyond standard preprocessing and normalization. As no data fusion strategy is applied in this study, we have chosen not to further elaborate on this aspect.

To avoid confusion with the previous phrasing we changed the sentence to: “During pre-processing all images from ERS and Envisat were resampled to 30 m resolution using bilinear interpolation and were reprojected to the Antarctic Polar Stereographic projection (EPSG:3031). “

4. (Lines ~155–160) The authors have reported values such as 221 m, 75 m, and 38 m, but it is not clear how these should be interpreted. It would help to briefly explain at first mention that these correspond to mean distance error between predicted and ground-truth fronts, and to relate them to human annotation variability. This would make it easier for readers to understand what counts as good performance.

This was also mentioned by Reviewer 1 and we re-phrased to: “A hybrid Transformer-CNN architecture has achieved an average MDE (Mean Distance Error) of 353 meters for post-processed results on the CaFFe benchmark (Wu et al., 2024). The latest model, SSL4SAR, uses a transformer architecture with self-supervised learning, achieving a MDE of 239 m with prior knowledge (Gourmelon et al., 2025b) and outperforming the previous best model by 67 m. “

5. It would be helpful to clarify why the evaluation is performed at the patch level rather than on full scenes? Patch-based evaluation can introduce edge effects and not reflect the real performance. It is required to either include scene-level results or justify why patch-based evaluation is sufficient. It would also be useful to clarify whether the other metrics (IoU, F1, accuracy, etc) are computed at the patch level or aggregated at the scene level?

We thank the reviewer for raising this important point regarding patch-based evaluation. We are aware that both patch-based and scene-based evaluation have pros and cons. We decided to provide patch-based accuracies based on the following points:

This was also raised by Reviewer 1 and we repeat our answer here:

We thank both reviewers for raising this important point regarding patch-based versus scene-based evaluation. We agree that both evaluation strategies have advantages and limitations, and we have carefully considered this trade-off in the design of our benchmark.

We chose to report patch-based accuracies primarily to ensure consistency with the publicly released version of Shelf-Bench, which is also distributed in a tiled .png format to facilitate accessibility for the broader computer vision community, many of whom are less familiar with large-scale satellite scene processing and native SAR scene structures. Using patch-based evaluation therefore ensures that reported metrics are directly comparable to the dataset format used in practical model development and benchmarking.

To minimise potential artefacts associated with patch extraction, we take care to avoid artificial boundary effects by removing background scene margins and excluding metric calculations within an 8-pixel buffer zone around patch borders. This reduces the influence of edge discontinuities on the reported performance.

We mention this now also in the limitations section: “Results are reported at the patch level to ensure direct comparability between the tiled PNG version of Shelf-Bench and its scene-based counterpart, while also enabling the identification of the most challenging prediction tasks at a fine spatial scale. A scene-based evaluation would additionally have introduced inconsistencies stemming from the substantial size differences between Sentinel-1 scenes (~400 km) and ERS/Envisat scenes (~100 km). That said, accuracy metrics aggregated across entire scenes might have yielded a slightly different picture. Patch boundary artifacts were mitigated by excluding an 8-pixel buffer zone around patch borders, though transitions between patches remain a limitation of tile-based inference. Future work could address this by adopting larger patch sizes and overlapping patch strategies, which would provide greater spatial context and allow multiple prediction probabilities to be aggregated at patch edges, likely improving both accuracy and boundary coherence.”

Furthermore, scene-level evaluation would introduce a strong bias due to substantial differences in scene extents across sensors. For example, Sentinel-1 scenes cover approximately 400 km, whereas ERS and Envisat scenes are on the order of 100 km. As a result, scene-based metrics would be dominated by differing spatial coverage rather than model performance, potentially masking local-scale variability and leading to non-comparable results across sensors.

Most importantly, our analysis focuses on identifying challenging local-scale conditions, particularly related to coastline morphology, sea ice conditions, and ocean–ice interactions (e.g. icebergs). The patch-based evaluation framework is well suited for this purpose, as it enables fine-grained assessment of performance in precisely these heterogeneous and dynamically changing regions. As demonstrated in Figure 6, this

approach allows us to explicitly highlight areas of high prediction difficulty that would be diluted or obscured in a scene-level aggregation.

To improve clarity, we added: “To avoid potential artifacts at patch boundaries, we take care to avoid artificial boundary effects by removing background scene margins and excluding metric calculations within an 8-pixel buffer zone around patch borders. “ and “All evaluation metrics are reported on a patch-based level. “

6. It is better to include training/validation curves of loss and F1 score to show convergence behavior. The statement about “stable fine-tuning of model heads” also needs clarification, because training length alone does not ensure stability. Please clarify whether all models converged within 150 epochs and whether only the heads were fine-tuned or the full networks were updated. Clarification is important for a fair comparison, especially between different models.

Thank you for raising this point which also Reviewer 1 addressed. Training ran for 150 but with early stopping (patience of 60) until model convergence. We removed the misleading phrasing and now state:

“All five models ran for 150 epochs, with a learning rate of 0.0001 with the Adam optimizer (Kingma and Ba, 2014) and early stopping with a patience of 60 epochs imposed on the validation loss. We selected 150 epochs to ensure model convergence. “

As stated in the data availability section, the training curves can be accessed here: https://github.com/amymorgan01/Shelf-Bench/blob/main/Figures/training_curves_X100.png. Furthermore, we decided to add the number of training epochs to Table 3 for better transparency.

7. The authors mention about temporal/geographic separation between training and test sets (“The entire benchmark dataset guarantees that there exists either a temporal or geographical distinction between the test and training datasets to prevent overlaps”), but generalization is not evaluated clearly. It would be useful to include a brief analysis of performance across different regions and seasons, and discuss how well the models generalize spatially and temporally.

Thank you for this insightful comment regarding model generalization across spatial and temporal domains. We agree that evaluating transferability is an important aspect of benchmark design.

In this work, we do not include an explicit stratified analysis of performance across predefined spatial regions or seasonal subsets. Instead, we focus on an error-based analysis distinguishing the most challenging and most straightforward cases within the test set. This choice is motivated by the observation that model performance is primarily driven not by geographic separation itself, but by region-specific factors such as rapid morphological changes, calving activity, and low-contrast conditions, which can occur across multiple regions rather than being confined to a single area. This is supported by the

observation that the most challenging scenes are all included in the training set, whereas the easiest scenes are not.

To this end, we identify and analyse representative hard cases (e.g. dynamically active regions such as Pine Island Bay) and easier, more stable scenes, which we find to be more directly informative for understanding failure modes of the models than purely spatial or temporal partitioning.

Nevertheless, we acknowledge that spatial and temporal generalization remain relevant perspectives. We have therefore clarified this point in the manuscript to better reflect the factors that primarily govern model performance.

“Sample hardness is not linked to the spatial distribution of training and test data. The most difficult samples occur in regions covered by the training set but remain challenging due to temporal and morphological variability. In contrast, several of the easiest scenes (e.g., Venable, Bach, and Quarisen) are absent from the training data, indicating strong generalization to unseen regions. Given the strong temporal variability of Antarctic calving fronts, spatial overlap does not imply representativeness. Instead, performance is primarily driven by scene-specific morphological complexity rather than inclusion in the training set. ”

8. Please clarify how areas outside the 50 km buffer are handled during training (masked or zero-valued) and whether this leads to patches with limited useful content, especially given the patch-based setup?

This question was also raised by reviewer 1 and we added to the manuscript: “Areas outside the buffer have the value 0, and hence appear black when plotted. We clarify this in the added sentence: “Areas outside the buffer get assigned the value 0.”

Indeed, this creates patches that have large no-data areas. Therefore, we decided to exclude all patches with more than 80% no data area as also stated in the manuscript: “We discard patches which contain >80% no data. The two `Ice’ and `Ocean’ classes are roughly balanced. The validation set enabled monitoring of the model performance during training, before final evaluation on the test set.”

9. It would be helpful to clarify whether there is any spatial overlap between training and test scenes, especially given the use of patch-based sampling.

We understand that we have to better state the temporal and spatial overlaps to avoid confusion. Yes, in fast changing regions training and testing scenes overlap (but then they are temporally different). This is also clearly shown in Figure 2 showing training and test data extents. The extents can also be viewed in GIS when downloading the .gpkg files from the Zenodo repository of Shelf-Bench. The temporal coverage and temporal distinction between train and test is given in Figure 3. It can clearly be seen that there is no temporal overlap between training and testing data. We also improved the phrasing in the following sentence to make this more clear:

“The entire benchmark dataset guarantees that there exists either a temporal (in case train and test scenes overlap spatially) or geographical distinction between the test and training datasets to prevent overlaps. A detailed list of temporal and geographical coverage by each satellite mission can be found in Table A 1 in Appendix A. “

10. It is required to run the experiments with multiple random seeds (5 or 10) and report the average (with standard deviation) of the results or even ensemble. This can provide a better comparison between the models.

We fully agree that a common practice in machine learning benchmarks is to train models with multiple random seeds and report mean performance together with the standard deviation as a measure of stochastic uncertainty. However, performing full re-training with 5–10 seeds per model would require computational resources that are currently not available to us. The training and hyperparameter optimization of all baseline models already occupied approximately 48 GPU hours on the JASMIN cluster for one week, and extending this to multi-seed experiments would add a substantial additional compute burden (on the order of several further weeks of training time), which is not feasible within the scope of this study.

We would also like to emphasize that this work is a dataset benchmark paper rather than a model development study. The primary objective of the baseline experiments is therefore to provide a representative indication of achievable performance and to support dataset validation, rather than to exhaustively optimize or fully characterize model uncertainty.

Nevertheless, we fully agree that some form of variability estimate is important for interpretability. In this revision, we therefore provide an alternative analysis based on hyperparameter sensitivity. Specifically, we report performance statistics (mean and standard deviation) across the three best-performing configurations for each baseline model. These results are now included in Table 5 in Section “3.3 Baseline Results”. In addition, we include an extended table in the Appendix Table A3 that lists all evaluated hyperparameter combinations for the top-performing models to ensure full transparency.

To make this statement clear in the manuscript we added: “Due to computational constraints, we do not perform a full multi-seed ensemble analysis. Instead, in Table 5, we assess performance variability across the top-performing hyperparameter configurations obtained during tuning, which provides an estimate of model sensitivity to architectural and training choices. In addition, we report the primary results using the best-performing configuration. We acknowledge that this does not isolate stochastic training variance, which would require repeated training with identical hyperparameters under different random seeds.”

11. It would be useful to include some form of uncertainty estimations (such as prediction confidence, variability across runs, std, variance or entropy).

We agree uncertainty estimation is important. To make the uncertainty clearer, we shifted the previous Table A2 stating 95% bootstrap confidence intervals (CI) to the main text in Table 6.

11. The explanation of the loss function is unclear. Since Dice loss already helps with class imbalance, it would be helpful to explain why Focal loss is also used and how

they work together. A short explanation of how the two losses are combined would make this clearer. It would also be useful to report the performance when using each loss separately to better understand the benefit of combining them.

We agree that our previous information on used loss functions was insufficient. As this was also raised by reviewer 1 we added the following to the text: “We evaluated different loss function formulations for semantic segmentation, by comparing a dice loss and focal loss equally weighted combination, and similarly for a dice loss and cross entropy loss formulation.” and “The Dice loss component measures the overlaps between true and predicted class areas, which optimises overlap between predicted and ground truth glacier masks. Alternatively, the Focal Loss component addresses boundary precision”. We also show in the Appendix Table A3 the best three model runs based on hyperparameter tuning stating the loss function.

12. The authors mention about using multiple GPUs with a batch size per GPU, which suggests parallel training?, but this is not clearly explained. Need to briefly clarify how the model was trained across GPUs for better reproducibility.

We agree this was not clearly stated in the manuscript. Training was executed as five independent SLURM jobs on the Jasmin NERC GPU cluster using Hydra’s SLURM launcher. Jobs were submitted separately and queued by the SLURM scheduler, which assigned each run to available GPU resources. Models were trained independently on single GPUs rather than via multi-GPU distributed training. The batch size of 32 corresponds to the per-job batch size on one GPU.

To clarify this in the manuscript we added: “Batch size was treated as a discrete parameter with candidate values of 8, 16, and 32 (largest computationally batch size per single GPU without parallel training). “