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 1

The manuscript presents a new dataset of multi-mission SAR images of Antarctic ice shelf fronts, complete with manual annotations of the calving front. Five deep learning baseline models were trained on this dataset and evaluated on a separate test set. The dataset represents a valuable resource for developing and benchmarking deep learning models for calving front extraction, and the reviewer appreciates the considerable effort the authors have invested in its creation. To further strengthen the work, several aspects would benefit from clarification and refinement.

## General Comments

First, it would be helpful to more clearly define the intended objective of a model on this dataset. For example, is an MDE of 0 m considered the ideal target, or is there a more realistic performance benchmark? Providing this context would improve the interpretability of the results.

Thank you for raising this comment. Calving front delineation is a very subjective task and results vary between experts. Therefore, an MDE of 0 m is not the ideal target. We added to the manuscript:

“Because calving front delineation is inherently subjective, the MDE between expert delineations can vary substantially depending on image resolution and the difficulty of identifying the calving front. Reported MDE values range from 33 m (5.5 pixels) (Mohajerani et al., 2019) to 92.5 m (Zhang et al., 2019) and up to 183 m (4.6 pixels) (Baumhoer et al., 2019) between several expert delineations. Therefore, achieving an MDE of 0 m is not a realistic expectation.”

In addition, showing model performance on entire scenes (rather than patches) would give a more complete picture of real-world applicability.

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. “

Line 308: Did you perform hyperparameter tuning for the learning rate? Please clarify. If not, please consider doing so, as this can strongly influence model performance and may bias the comparison. The learning rate interacts closely with the architecture, optimizer, and parameter initialization, and different models (e.g., CNNs vs. Transformers, or pretrained vs. randomly initialized networks) typically require different learning rate regimes to converge properly. A learning rate that is too high may prevent convergence or lead to unstable training for some architectures, while a rate that is too low may result in underfitting or unnecessarily slow convergence. As a result, using a single fixed learning rate across all models can disadvantage certain architectures and lead to misleading performance differences that are not attributable to the model design itself but rather to suboptimal training settings.

We agree that hyperparameter tuning is important, and we have incorporated it into the revised analysis.

We added the following to the manuscript: “We conducted a hyperparameter tuning to obtain best performance for all models using a Bayesian optimisation sweep implemented through Weights & Biases. The same sweep configuration and tuning strategy were applied consistently across all architectures, including U-Net, DeepLabV3, DINOv3, ViT, and FPN. Validation loss was minimised over a fixed training duration of 200 epochs. The learning rate was sampled from a log-uniform distribution ranging from 1×10−5 to 1×10−3, enabling evaluation across both conservative and aggressive optimisation regimes. Weight decay was similarly tuned using a log-uniform distribution between 1×10−6 and 1×10−3 to assess the impact of regularisation on model generalisation. Batch size was treated as a discrete parameter with candidate values of 8, 16, and 32. The Adam optimiser was used for all experiments for consistency. 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. The hyperparameter sweep revealed that model performance was influenced by the learning rate, loss function, and batch size to a greater extent, while weight decay had a comparatively smaller effect within the tested range. Across all five architectures, the highest-performing configurations generally used moderate learning rates (≈$10^{-5}$–$10^{-4}$). DeepLabV3 showed the strongest overall model performance, where the best configuration used a learning rate of $3.49 \times 10^{-5}$, weight decay of $2.60 \times 10^{-5}$, batch size of 32, and the Dice–Focal loss. In general, the Dice–Focal loss slightly improved ice segmentation performance in comparison with the Dice–Cross-Entropy loss. Higher batch sizes of 16 and 32 produced stronger model performance than batch size 8, which was particularly apparent for lower segmentation metrics for FPN and DeepLabV3. The transformer-based models, DinoV3 and ViT, showed greater variation in model performance during hyperparameter tuning than UNet, FPN and DeepLabV3.”

Line 310: „We selected 150 epochs to ensure stable fine-tuning of model heads“ - Training length does not imply stability. Do you mean that all models converged within 150 epochs?

You are correct. We rephrased to “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. “

Furthermore, the statement about fine-tuning only the heads requires clarification. In standard practice, only foundation models are kept frozen with head-only fine-tuning because their internal representations are general-purpose. Other pretrained models, however, are not designed to perform well on entirely new downstream tasks without updating the full network. If in your experiments only the heads of these non-foundation models were fine-tuned, the reported results may not reflect their true potential. To obtain a fair and meaningful comparison, these models should be fully fine-tuned (updating all layers) except for the foundation model, which can remain frozen except for the head.

Thank you for raising this point. For clarification we added the sentence: “The UNet, FPN and DeepLabV3 models are fully fine-tuned during training to maximise task-specific feature adaptation, whilst the ViT and DinoV3 models are trained with frozen backbones to preserve pretrained features and reduce computational load.“ Additionally, we state the number of epochs each best model was trained in Table 3.

Line 441–442: „Additionally, the buffer used to clip the scene around the current coastline lies close to the image border and follows the calving of iceberg A-68, which may have further influenced model performance.“ – Please clarify how areas outside the buffer were handled by the models (e.g., treated as white or black). Additionally, why was clipping done in a way that some patches contain very little actual content? Would it be possible to use full patches when they lie partially on the buffer edge, rather than excluding areas outside the buffer?
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.”
In principle, a squared buffer with an outer no-data margin could have been generated instead of a rounded buffer to have patches fully covered. However, since the Sentinel-1 scenes were derived from previous studies and had already been clipped using the existing buffer, we retained this approach for consistency.

Line 521–522: „Even though, this did not result in lower prediction accuracies for specific regions because they weren’t represented within the training set.“ – I strongly recommend performing a quantitative analysis separating test scenes that are spatially covered by the training set from those that are not, in order to rigorously assess and demonstrate the model’s generalization performance.

Thank you for this valuable suggestion regarding a more explicit analysis of spatial generalization. In Antarctica, the coastline is highly dynamic and spatially complex, with rapidly changing calving fronts and heterogeneous shelf morphologies. As a result, spatial overlap between training and test scenes does not necessarily imply temporal consistency, since even identical locations can differ substantially between acquisition times.

To address generalization performance, we already provide a hardness-based analysis and visualization of the easiest and most difficult test scenes in Figure 7. This analysis shows that the most challenging cases (e.g., Crosson, Holmes, Thwaites, and Larsen C) are at least partially represented in the training data at different time steps, yet still yield high difficulty due to temporal and morphological changes. Conversely, several of the easiest scenes (e.g.,

Venable, Bach, and Quarisen) are not represented in the training set, indicating strong model generalization to unseen regions.

We therefore find no systematic dependence of model performance on spatial training coverage; instead, performance is primarily driven by scene-specific morphological complexity rather than geographic overlap. Given that Shelf-Bench already spans a broad range of ice shelf morphologies, this supports the model's ability to generalize across both seen and unseen regions under varying conditions.

To clarify this in the manuscript we added: "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. "

## Specific Comments

Figure 1 and Related Text:

• Several relevant works on Antarctic calving front delineation are currently missing and should be incorporated to provide a more comprehensive overview:
o https://ieeexplore.ieee.org/document/11296938 → added
o https://www.mdpi.com/2072-4292/15/21/5168 → excluded, see below.
o https://tc.copernicus.org/articles/17/3485/2023/ → corrected Zhang et al. 2021 to 2023
o https://tc.copernicus.org/preprints/tc-2023-52/ → same architecture as Loebel et al. 2021, but now cited in the text
o https://essd.copernicus.org/articles/14/4287/2022/ (baselines) → added
o https://arxiv.org/abs/2501.05281 (foundation model) → added
o 10.1109/IGARSS52108.2023.10281828 → added
o https://ieeexplore.ieee.org/document/10903382 → added
o https://arxiv.org/abs/2512.11560 → still a preprint
o https://arxiv.org/abs/2601.21663 → still a preprint
• Some references appear to be mislabeled: corrected
o Zhang et al. (2021) is labeled as AutoTerm (likely Zhang et al. 2023).
o Loebel (2021) should be Loebel (2022).
o "Transformer" should be extended to include Transformer–CNN hybrids.
o Gourmelon et al. (2025) should be specified as Gourmelon et al. (2025, TYRION).
• The definition of "edge-based methods" would benefit from clarification. It is currently unclear whether this includes methods trained on binary calving fronts (or additionally on binary calving fronts) or if a different definition is intended.

Thank you for completing Figure 1 with most up-to-date literature and references we missed. We added your suggestions to Figure 1 and updated the references accordingly. Except for reference https://www.mdpi.com/2072-4292/15/21/5168. I am aware of this manuscript; however, its methodological rigor appears limited. For example, Figure 7 suggests that the U-Net baseline may not have been implemented correctly, as it exhibits pronounced patch artifacts.Edge-based methods refer to models that are specifically designed to detect boundaries or edges rather than perform full semantic segmentation.

Line 57: „future SAR satellite mission“ - This claim is not fully supported, as the training set appears to include the same satellites as the test set. Demonstrating generalization to a new mission would require at least one unseen satellite in the test set.

We clarified that we meant with this upcoming Sentinel-1 missions by saying: “...across current and future satellite missions (e.g. Sentinel-1D). “

Table 1: coverage of Gourmelon et al. 2022 seems to be incorrect. Moreover, please define “benchmark dataset” more clearly (e.g., requiring preprocessed imagery and manual labels), otherwise additional datasets may need to be included.

Thank you for the correction, this was probably a copy paste issue in the wrong line. We corrected to five glaciers (Jorum, Crane, Mapple, DBE-Glacier, Sjogren Inlet). Furthermore, we now define what we understand as a benchmark dataset: “). However, the absence of a standardized benchmark dataset (consisting of pre-processed satellite imagery and manual labels) for Antarctic ice shelves prevents objective performance comparison across algorithms. “

Line 123: „in polar regions“ - Figure 1 specifies Antarctic; this should be made consistent.

We corrected “in Antarctica and Greenland” (e.g. work from Zhang or Loebel covers Greenland).

Line 158–160: State 353 m instead of 221 m, or clarify that this refers to the multi-annotator study of CaFFe. Additionally, it is 238 m on the CaFFe benchmark with additional prior knowledge and 75 m for the comparison to the multi-annotator study; please ensure this is clearly explained. „Transformer architecture“ → this appears to be a hybrid Transformer–CNN architecture.

Thanks for clarifying. We just report the values on CaFFe without the multi-annotator study to avoid confusion. We added hybrid Transformer-CNN architecture.

Line 168–170: „The test set was defined after expert evaluation of the images. It includes a representative selection of typical ice-shelf boundaries with varying levels of complexity to ensure comprehensive coverage of the problem space acquired by different sensors.“ - These sentences would fit better after the description of the training and validation sets.

Shifted.

Line 171–172: „the test set is based on 38 scene subsets especially focusing on the front“ - How much calving front is present in the training and validation sets if they are not front-focused? Please clarify.

The ERS and Envisat scenes are always front focused, as scene size is smaller. Sentinel-1 scenes cover large regions and hence include rock and grounded ice coastlines as well. Quantifying exactly is difficult as the delineation lines do not contain information on shelf vs rock but we made it more clear by adding: “. Whereas the training set includes parts of the

rock coastline due to the large scene size of Sentinel-1, the test set focuses explicitly on the front of ice shelves and therefore contains cropped scenes. “

Line 241: „multi-expert annotation protocol“ - This is a strong aspect of the dataset. However, please clarify for the reader that annotations are not independent and therefore not suitable for significance testing or uncertainty estimation.

To clarify, we added: “This makes the Shelf-Bench dataset in itself consistent, however, the annotations are not independent and hence not suitable for significance testing. “

Line 274–275: „Atrous Dilated Convolutions which enlarge the convolutions“ - Consider rephrasing to “which increase the receptive field of the convolutional filters.”
Rephrased to: “and utilises Atrous Dilated Convolutions which increase the receptive field of the convolutional filters, enhancing feature detection without increasing computation or the number of weights (Chen et al., 2017).”

Line 283: Please add Gourmelon et al. 2025a, as this work evaluates the foundation model SAM.

Added.

Line 285: „input each Shelf-Bench SAR image into the RGB channels“ - Please clarify this sentence, particularly for readers unfamiliar with modality mismatches.
Rephrased to: “and due to the requirement of optical three channel (RGB) images as inputs for DINOv3, we repeat the same SAR scene three times for each of the RGB channels.”

Table 3: Why is the inference time for DINOv3 higher than for ViT? A brief explanation would be helpful.
Having reevaluated the inference times for all models on cpu and gpu as they were highly influenced by usage of the compute cluster, we found different inference times to previously stated (these were for previous model versions). We have now decided to remove the inference time column due to the similarities of all models GPU times (1.07 - 1.17 ms per patch). Updated the sentence “The inference time per patch varied between 1.07 - 1.17 ms per patch between models run on GPU, which was ~200x faster than running on CPU for all models.”

Line 301: „clip method“ - Please provide the exact percentile values used.
Rephrased to: “Patches are normalized locally using a percentile clip method on individual patches, where values below the 2nd percentile and above the 98th percentile are clipped, improving contrast in glacial environments (Loebel et al., 2024)”

Line 319: „where the authors found this combination of loss functions improved performance“ - In this paper, focal loss does not appear to have been used; please verify.
Rephrased to “This combined loss function approach was inspired by Gourmelon et al. (2022), where the authors found a combination of loss functions improved performance (cross-entropy and dice loss in their case).”

Line 320: „class boundaries“ → class areas.

Corrected.

Line 320: „useful for learning the often complex and irregular shape“ - To the best of my knowledge, the standard Dice loss is not particularly suited for this, so please clarify or provide evidence for this claim, and indicate whether any modifications or weighting were applied to better handle such structures.
Rephrase to: “The Dice loss component measures the overlaps between true and predicted class areas, which optimises overlap between predicted and ground truth glacier masks.”

Line 321: „addresses class imbalance“ - Dice loss also already addresses class imbalance; please clarify the distinction.
Rephrased to “Alternatively, the Focal Loss component addresses boundary precision.”

Heading 4.3: consider renaming to „Baseline Results and Discussion“.
Done.

Table 4: „balances false positive and false negatives“ - More precisely, balances precision and recall (false positive rate and false negative rate).
Rephrased to “balances precision and recall”

Line 346: „Only DINOv3“ - Figure 4 appears to show similar behavior for ViT, U-Net, and DeepLabV3; please check consistency.
Rephrased to “Some models introduce holes in the ice prediction.”

Line 347: „U-Net delineated the front closest to the ground truth“ - Figure 4 suggests FPN may perform similarly or better; please clarify whether this refers to only Figure 4 or to the entire test set results.
Rephrased to “In the Figure 4 examples, FPN delineates the front closest to the ground truth whereas especially U-Net and DeepLabV3 detect the boundary but generalize the coastline and miss smaller details”

Line 350: „sea ice and fast ice challenges all models leading to slight discrepancies“ - Please elaborate on this interpretation. Would models not classify fast ice as part of the shelf?
We clarified by adding “By definition, fast ice is multi-year sea ice and hence not labelled as ice shelf in the training and test dataset. Still, especially in single-pol SAR data shelf and fast ice can have very similar backscatter characteristics and hence create a very challenging prediction task. “

Line 355: „reproducible segmentation performance“ - This would typically require multiple training runs of the same model; please clarify.
Rephrased to: “Overall pixel accuracy exceeds 0.906 across all models suggesting that the Shelf-Bench dataset supports stable convergence and consistent segmentation performance between models”

Line 359: „boundary ambiguity“ - As the boundary is defined between ice and ocean, ambiguity applies to both classes; consider clarifying.
Rephrased to: “This consistent gap suggests that the dataset exhibits greater intra-class variability and challenging ice–ocean boundaries for delineation”

Line 363: „the best performing model on ice recall (ViT) finds most of the relevant ice instances.“ – The differences in recall between the models appear relatively small. Therefore, this statement may be somewhat overstated, as “all relevant ice instances” would correspond to a recall of 1.0. Consider rephrasing this more cautiously to reflect the relatively close performance.
Rephrased to reduce the overstatement: ““In contrast, ice recall exhibits a wider spread (0.882–0.93), indicating marginal model performance differences, with ViT achieving the highest recall.”

Line 365–366: „The lower precision for ice indicates that the Shelf-Bench dataset includes a sufficient number of difficult ‘Ice’ samples to challenge model performance.“ - Relative differences in precision between classes can not support this conclusion.
Removed the sentence.

Line 382–384: „Satellite scene boundary artifacts were removed…“ - Please provide more detail on these processing steps.

To make this more clear we added: “Sometimes at the satellite scene boundary one-pixel wide misclassification along the boundary can occur and are removed by the binary erosion.“

Line 392: „The ViT delineates the coastline to be further inland than the ground truth and other models, highlighted by the higher MDE (Figure 5a).“ - Figures 5b, 5c, and 5e appear to show a different trend; please check consistency.

We created a new version of Figure 5 based on your suggestion below. Now the land inward prediction of ViT is clearer.

Line 396: „produce more stable front delineations“ - Please clarify whether this refers to specific panels (5c, 5e) or a general observation.

Removed as it no longer applies to the new figure.

Figure 5: The figure resolution appears somewhat low. Is this due to journal formatting, or could a higher-resolution version be provided to allow zooming in on details? Additionally, the SAR images look more grainy or lower in resolution compared to Figure 4; please clarify the reason for this. The use of different colors for each model may be unnecessary, as the column already identifies the model. Using a single, easily visible color for all models (e.g., avoiding yellow against bright ice) could improve clarity.
We redid the figure with only two colours (ground truth and prediction) now with higher zoom level and better resolution. We also edited the figure description in the text accordingly.

Line 408: „a more robust performance indicator“ - Please clarify what this refers to.

We meant “realistic” not “robust”. Changed in the text.

Line 450: „median“ - For the most challenging cases, the mean MDE is reported; please ensure consistency.
Changed to median for consistency.

Line 459–460: „The baseline models behave differently on Shelf-Bench, indicating that the dataset contains complementary challenges that no single architecture captures equally well“ - Please also acknowledge that the stochastic nature of neural network training has an influence on the baselines performance as well.
Rephrased to ““The baseline models behave differently on Shelf-Bench, indicating that the dataset contains complementary challenges that no single architecture captures equally well; however, the stochastic nature of neural network training may also influence the baselines' performance.”

Line 461: „the dataset spans multiple spatial scales“ - This is somewhat vague; consider specifying that objects of interest occur at different scales.

Added “objects” to make this clear.

Line 461–462: „... it also includes sufficiently clear fronts to evaluate pixel-level delineation accuracy.“ - This statement is unclear and would benefit from clarification.

We added the following to make our statement more clear: “...while U-Net's low median MDE (49.4 m) shows that it also includes sufficiently clear fronts with a very clear ice-ocean contrast to evaluate pixel-level delineation accuracy. “

Line 467–469: „The lower performance… suggests…“ - Please elaborate on how this conclusion is supported by the results.

We state now more clearly in line with the results: “The lower performance of all models based on IoU for the class 'Ice' compared to 'Ocean' (see Table 5) suggests Shelf-Bench contains scenes containing especially challenging cases for 'Ice' classification. This can be likely due to spectrally similar ice features (e.g. icebergs and shelf ice) challenging correct predictions.”

Line 499–500: : „Increasing the tile size and introducing overlap tiling with centred predictions would provide broader spatial context and help reduce such ambiguities.“ - Please cite prior work (e.g., on the CaFFe dataset) that has explored this.

Added Wu et al. 2023 using a local and a context patch. Feel free to mention further studies we might have overlooked.

Line 515: „most realistic“ - Given that the median MDE is lower than the mean, easier scenes may be more prevalent; consider revising.
Removed “realistic”

Line 525: „temporal“ → spatial?

Line 526: „temporal repetition“ - Please clarify (e.g., within the training set?).
Merged both comments above and added: “To make this more clear we added: “to assess whether temporal repetition in the training set over the same location improves model generalization “

Line 545: „targeted synthetic label generation“ - Please explain what is meant by this.

We added: (e.g. by creating synthetic labels of challenging mélange or fast ice conditions)

Line 574: „integrating temporal information“ - Please consider citing Dreier et al. 2025 (https://arxiv.org/abs/2512.11560), which explores this for the CaFFe dataset. Data availability section: Consider moving descriptive elements to the dataset section and keeping only the access link here.

We included the descriptive elements on purpose in this section to make it easy for users of the dataset to find key information.

Line 598: „locally z-score normalized“ - Please clarify what “locally” refers to and which statistics were used for normalization.
Rephrased to “Patches were additionally converted to decibel (dB) and z-score normalized ($\frac{x-\mu}{\sigma}$) across the dataset using the empirical mean (0.4768397510) and standard deviation (0.2779399157) of the train dataset”.

Table A3: Why are there missing entries? Please clarify.
We clarified by adding a better Table caption: “Table A 3 MDE values for the most difficult patches in the test set. Visual examples and number of available predictions are given in Figure 6. Missing values indicate no available prediction from the respective model. “

## Technical Comments

Table 2: introduce the abbreviations IMP, GRD, AMI, ASAR, EW. Please also use “multi-looked” consistently (with or without hyphen).

Add and used multi-looked with hyphen.

What does the number after “/” in the repeat cycle indicate?
The repeat cycle can vary e.g. depending whether one or two Sentinel-1 satellites are available. For ERS-1/-2 a one day repeat cycle was available during their tandem mission in 1995-1996. We added this to the caption of Table 2.

Line 205: introduce the abbreviation IM.
Added “Image Mode”

Line 410: Missing closing parenthesis.
Fixed