Deep learning based automatic grounding line delineation in DInSAR interferograms

Sindhu Ramanath Tarekere[1], Lukas Krieger[1], Dana Floricioiu[1], and Konrad Heidler[2]

[1]Remote Sensing Technology Institute, German Aerospace Center, Oberpfaffenhofen Germany

[2]Data Science in Earth Observation, Technical University of Munich, Germany

Correspondence: Sindhu Ramanath Tarekere (sindhu.ramanathtarekere@dlr.de)

Authors' response to review comments (https://doi.org/10.5194/egusphere-2024-223-RC2)

We thank the reviewer for providing their expertise and time to review our work. We plan to revise our manuscript after considering their comments and suggestions. The major changes that will be incorporated into the manuscript are summarized below:

1. We will improve Fig. 5 by enlarging the individual plots and Fig. 8, 9, 10, 11 and 12 by including subplots of the ROIs and zoomed-in insets of the double difference interferograms.
2. We will add the subpanels c-j of Fig. 2 to Fig. 6 and 7 in the manuscript.
3. We will restructure Section 5.1 Training scheme and Section 5.3 Spatial transferability and generalization to clarify the difference between in-sample and out-of-sample dataset splits. Additionally, we will rename the dataset splits:
    (a) The in-sample dataset split will be called a 'temporal' split.
    (b) The out-of-sample dataset split will be called a 'spatial' split.
4. We will quantify the uncertainty of networks 1 and 3 (Table 3 in the manuscript) by training an ensemble of five neural networks each. We will also include visualizations of the 95%
5. We noticed an error in Fig. 7 of the manuscript, wherein we had mistakenly included manual delineations of an independently processed set of double difference interferograms with the AIS_cci dataset. Fig 1 shows the differences between the two images. We will add the correct figure in the revised manuscript.

Below are point-by-point replies to the comments. Reviewer comments are in black, the authors' response in blue and snippets from the unrevised manuscript in red.

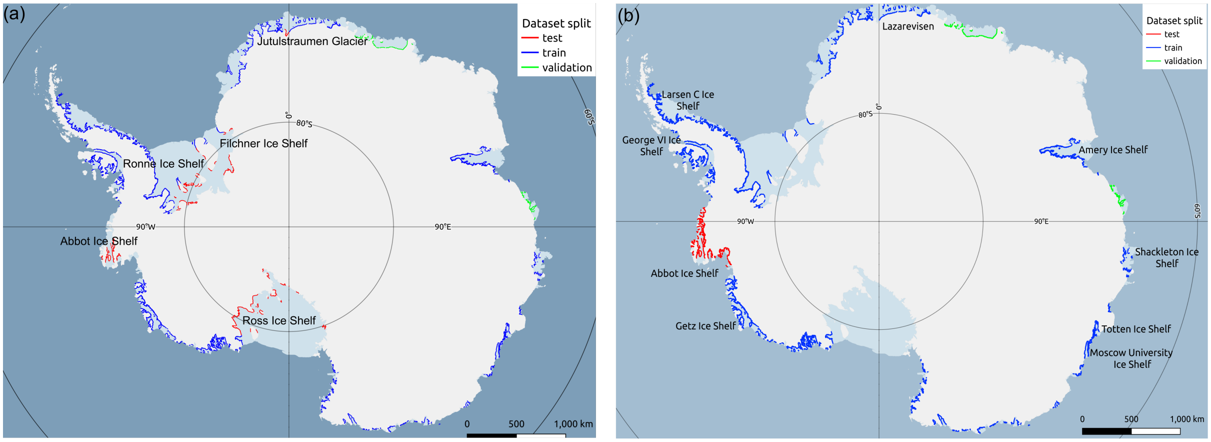

Figure 1. The spatial distribution of the out-of-sample dataset split shown in Fig. 7 (a) currently in the manuscript and (b) corrected figure. The corrected figure does not contain the lines from the Ross Ice Shelf, Ronne Ice Shelf, Filchner Ice Shelf and Jutulstraumen Glacier, which were processed independently and are not part of the AIS_cci dataset.

Major comments

1. Introduction and Related Work share many repetitive contents regarding non deep-learning remote sensing methods in detecting GL. I recommend merging these two.

   We agree that these sections can be merged. We will condense the content of Related Work and include it after line 48 in the manuscript.

2. Dataset:

   (a) The network-generated results have many spurious short line segments shown below (black lines – network GLs, red lines – AIS_cci GLs), any idea how to remove these inaccurate predictions when using the product?

   We have a python function *remove_outliers* (https://gitlab.dlr.de/rama_si/automatic_gll_delineation/-/blob/main/postprocess/vectorize.py?ref_type=heads) which filters out spurious lines which lie outside of a user given buffer (buffer distance in metres) around a reference GL (also user given). Please note that the accuracy of the routine is dependent on the chosen buffer distance and completeness of the lines in the reference GL dataset.

   Currently, this function is run as a part of our postprocessing procedure, resulting in unfiltered and filtered network delineations. The unfiltered lines were used for all the visualizations and metric calculations in the manuscript. We will provide both the unfiltered and filtered lines of the ensemble in our Zenodo repository (https://doi.org/10.5281/zenodo.11277696), along with a short description of the filtering routine.

   (b) Uncertainty: In Mohajerani et al. (2021), they used the width of the vectorized contours as mapping uncertainty. With the threshold (0.8) scheme in your postprocessing, the mapping uncertainty can be easily achieved by applying different thresholds in extracting the grounding line.

The last layer of the neural network proposed by Mohajerani et al. (2021) uses the Sigmoid function to produce scores in the range $[0, 1] \in \mathbf{R}$ for each pixel. To our understanding, the contours are constructed by choosing those pixels that the network scores above 0.3. We believe this to be a proxy for the actual neural network uncertainty, as these scores are not associated with the model uncertainty (Gawlikowski et al., 2023). Similarly, the threshold (0.8) in our postprocessing scheme is merely a criterion for choosing the pixels the neural network scores 0.8 and above. Therefore, varying this threshold would not quantify the uncertainty of our proposed neural network. Instead, we will train an ensemble of five neural networks and provide the mean network delineation as well as the 95% confidence interval for each sample. Please refer to our response to reviewer 1 for more details on uncertainty quantification https://doi.org/10.5194/egusphere-2024-223-AC1.

3. I believe it is unnecessary to spend extensive effort discussing calving front mapping in this paper, as the primary focus is on detecting grounding lines. While grounding line detection shares similarities with glacier calving fronts, such as both being line segments, the input data sources are fundamentally different. Consequently, methods effective for calving front detection may not be suitable for grounding line detection. It may be beneficial to mention that calving front edge detection inspired this research, but a detailed appendix reviewing various ML/DL methods for mapping calving fronts is unnecessary, especially since most referenced studies utilize UNET, unlike the edge detection approach in this research.

   As you suggest, we will cite a few important references for deep learning-based calving front delineation and remove the detailed survey from the appendix.

4. Additionally, you mention that Mohajerani et al. (2021) is the only study so far using a DL algorithm for mapping Antarctic grounding lines. However, there is no comparison between the models proposed in this study and those in Mohajerani et al. (2021) . What are the benefits of using edge detection algorithms compared to the encoder-decoder architecture in Mohajerani et al. (2021)? How does your model's performance compare to that of Mohajerani et al. (2021)?

   We have compared the performance of the network proposed by Mohajerani et al. (2021) to ours but chose not to add it to the first draft of the manuscript due to some limitations. First, The training dataset for the study was not published. Second, we could not reproduce their computing environment. We implemented their network architecture and trained it on our dataset, but did not achieve the performance stated in Mohajerani et al. (2021) and therefore do not think this is a fair comparison. Nevertheless, we will add the experimental results in our revised manuscript while mentioning the above-stated drawbacks. We also performed a few experiments in which we trained a UNet on the same features stack (Ramanath Tarekere, 2022) and found our targeted edge detection neural network to work better for this task. We will mention these experiments in the discussion.

5. In-sample and out-of-sample variants:

   (a) I am confused about creating two different variants of training/validation/test sets as in-sample and out-of-sample sets. I also wonder why these two variants are divided based on the spatial or temporal overlaps. The in-sample data are the datasets that the model has access to during training and validation, while out-of-sample data are used to test the model performance, so it is a testing set, as such I don't understand why both in-sample and out-of-sample sets contain three individual training/validation/testing sets and why you need to train two different networks on these two datasets according to Section 6.3.

Thank you for this comment. We agree that we must adequately clarify the distinction between in-sample and out-of-sample splits. We considered two different distributions of the AIS_cci dataset, i.e., they are simply two different ways of splitting the same dataset into training, validation and test sets. The samples for any ROI in the out-of-sample split belong exclusively to either the training, validation or testing sets. Conversely, the in-sample split contains training and test samples or training and validation samples with different epochs from the same ROI. However, there are samples for certain ROIs that belong exclusively to the training set. Figure 2 of this response shows examples of several ROIs.

The purpose of training the networks on these two splits was to determine how well they could delineate interferograms outside their training sets' spatial and temporal domains. Such evaluations have been performed for deep learning models that were developed for other cryosphere-related applications which deal with high spatiotemporal varying data, such as glacier mass balance (Guidicelli et al., 2023) and calving front delineation (Gourmelon et al., 2022), (Herrmann et al., 2023). Our results in Section 6.3 indicate a benefit in training the network with an in-sample dataset:

Line 284: "Even though the out-of-sample distribution did not contain any training sample that covered the Abbot Ice Shelf (Fig.7), the delineation of the respective HED variant is very similar to the network trained on the in-sample dataset, in which all but one interferogram was a part of the training set. Still, the out-of-sample HED GLs are more fragmented and spurious than those of the in-sample network. The latter network perhaps benefited from seeing several interferograms for the same region in the training set (Marochov et al., 2021), and therefore, finds application in producing a time series of GLs for regions with a sufficient number of coherent interferograms."

We used the term 'out-of-sample' to refer to a dataset split in which the test samples are spatially apart from the training and validation samples, as used in Gourmelon et al. (2022). We admit that the deep learning community primarily uses 'out-of-sample' to refer to the test set. We apologize for the confusing terminology. Instead, we will refer to the 'in-sample' split as a 'temporal' split and the 'out-of-sample' split as a 'spatial' split.

(b) In Table 4, I think the feature subset should be one of these interferometric/non-interferometric feature combinations listed in Table 3? Why here is In-sample or out-of-sample? When you train two networks for in-sample and out-of-sample datasets, which interferometric/non-interferometric features combination did you use?

Thank you for noticing this discrepancy; we will change the column heading for Table 4 from "Features subset" to "Dataset split". As mentioned in Section 5.3, the two networks were trained on the rectangular features of the samples of the respective dataset splits:

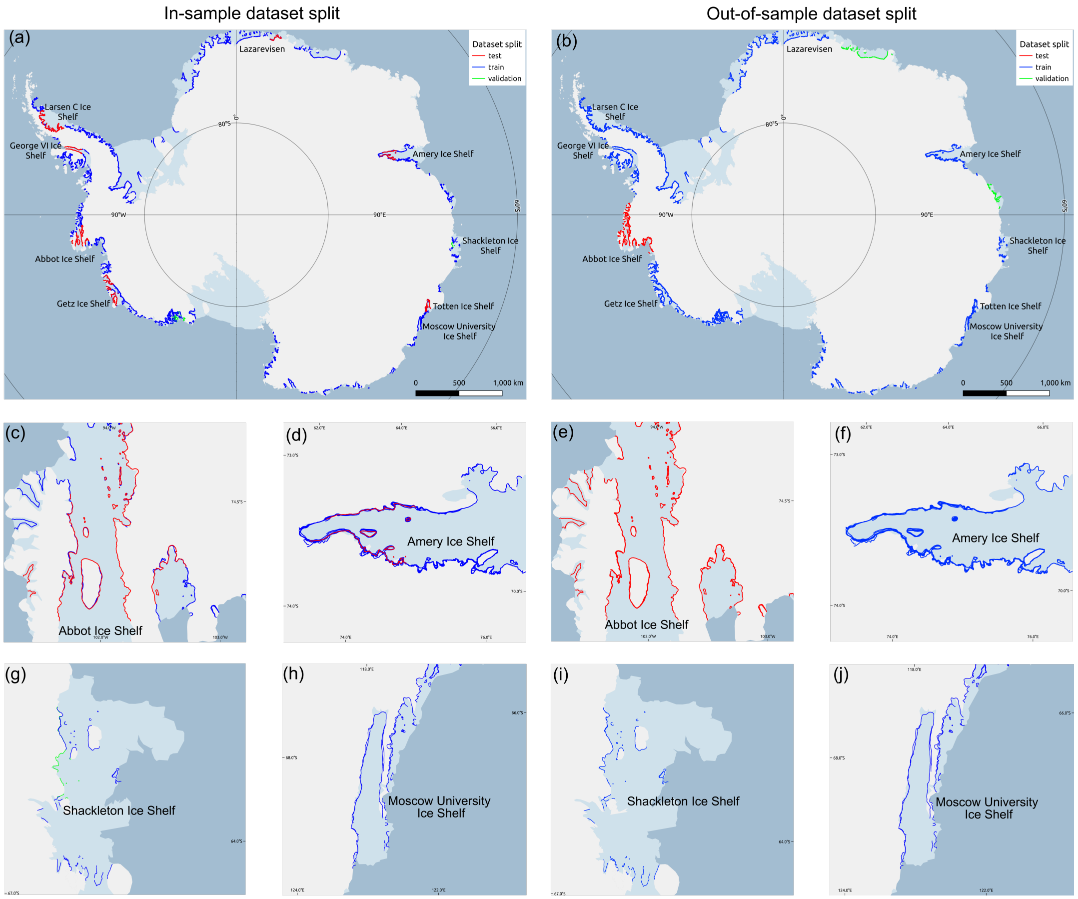

Figure 2. Spatial distribution of the AIS_cci lines into training (blue), validation (green) and test (red) sets (a) in-sample split (b) out-of-sample split. The in-sample data for (c) Abbot and (d) Amery Ice Shelf contain spatially overlapping but temporally separated training and test samples. In contrast, the out-of-sample data for the same ROIs contains only (e) test samples or (f) training samples. The in-sample data for the (g) Shackleton Ice Shelf contains training and validation samples, whereas the (h) Moscow University Ice Shelf samples contain only training samples. The out-of-sample data for the same ROIs (i) and (j) contain only training samples.

Line 208: "To investigate HED's generalization capabilities, we used the network trained with rectangular features of the training samples from the in-sample dataset to delineate the test samples from the out-of-sample dataset (Fig. 7). We compared these delineations to those generated by HED, which was trained with the training samples of the out-of-sample dataset."

(c) From the paper itself, it seems you mainly used the in-sample training dataset to train the model and then evaluate the model performance on the in-sample and out-of-sample test sets, then what is the point of generating the out-of-sample training and validation sets?

Please refer to the response to comment 5b. We trained one of the networks with the training samples of the in-sample set and the other with the training samples of the out-of-sample dataset. The respective validation sets were also used to adjust the hyperparameters of the networks.

(d) Table 3 shows the numerical results of different networks, however, here it only shows results for one test set, is it an in-sample or out-of-sample test set?

We evaluated both networks on the out-of-sample split test set to make a sample-to-sample comparison. We mentioned this in Section 5.3 in the manuscript, cited in 5b. We will reiterate this point in Section 6.3.

6. I am not convinced by Section 6.1. The importance of the interferometric features can only be proved by comparing them with networks trained with non-interferometric features. However, here you only compare networks 1 & 2, which are both trained with interferometric features.

   We intended to distinguish which among the two sets of interferometric features, rectangular or polar, are important for grounding line delineation: Line 193: "As mentioned in Section 3.2.1, we express the complex DInSAR interferograms as real and imaginary parts, i.e., the rectangular representation, as well as the pseudo coherence and phase components, i.e., the polar representation. We train two networks, one with rectangular interferometric features and the other with polar interferometric features in their training features stack, to determine the optimal representation of the interferogram for GL delineation." We did train the network with just the non-interferometric features and decided to discard this experiment as the resulting delineations were too fragmented and unusable. We will mention this in the manuscript.

7. Section 6.2 the importance of DEM (Line 270 and Figure 10):

   – Please include a detailed zoomed-in map of the interferogram inside the blue box. It seems the interferogram phase inside the blue box is decorrelated, so I won't be surprised that the network cannot map the correct GL. Also, only giving one example with a small spatial extent is not representative.

     We will provide a zoomed-in inset of the interferogram inside the blue box. Indeed, the interferogram in this region is partially decorrelated, making it likely that the network relies more on the DEM. While we agree that the manual GL in this region may not be 100% accurate, we do not believe that the real grounding line coincides with the elevation drop shown in the DEM. Ideally, we expect the network to not make any delineations at all for such cases. We will provide other examples as well.

- Have you checked the elevation change in Cabinet Inlet, is it a region undergoing significant elevation changes? If elevation is stable, I don't think you can attribute the wrong GLs to different DEM stacks

  Thank you for raising this valid concern. We will have a look to see if it is stable. You are right; if the region is stable, the break in slope proxy derived from the DEM could be close to the interferometric grounding line. However, the more significant point we were trying to make is that our DEM feature does not vary temporally; therefore, one must be cautious when training the network with the DEM.

- How to achieve the balance of including DEM to avoid over-reliance?

  We have yet to explore this aspect in our study, so, unfortunately, we cannot suggest a strategy to overcome this problem. It is of scientific interest and would be worth investigating in a follow-up study.

8. Section 6.3:

   (a) As mentioned above, I don't understand why compile two different in-sample/out-of-sample sets and train two networks. If you combine the in-sample and out-of-sample sets into one dataset, won't this greatly increase the training samples and improve the model performance?

   Please refer to the response to comment 5a.

   (b) You evaluate the in-sample trained model performance on the unseen Ross Ice Shelf interferograms by using Figure 12, however the discussion on the prediction quality is limited. Most GZ regions in Ross Ice Shelf are stable, I would like to see a distance deviation map between the AIS_cci GL and the network-generated GLs in Ross Ice Shelf to demonstrate the performance. If there are large deviations, please consider explaining 1) what are causing the large deviation? 2) which dataset is correct? 3) how can you further improve these results?

   Given the stability of the glaciers in the region, we agree that this is an 'easy' ROI. Still, the network was never trained with any interferograms from the Ross Ice Shelf. Fig. 12 was used only to show the spatial transfer capability of the network and gauge the quality of the delineations from a purely visual perspective, as these interferograms were never manually delineated. The black lines in Fig. 12 (a) correspond to a different set of double differences, for which we have manual delineations.

   Line 291: "Despite never having seen the interferograms during the training or validation stages, the network delineated the landward-most fringe and largely avoided delineating the decorrelated fringes of the Crary Ice Rise and Nimrod Glacier interferograms in Fig. 12c, d. The loose fringes in the interferograms of the Dickey Glacier and Nursery Glacier were not delineated. The network complemented existing manual delineations, reducing significant gaps and leading to a more complete grounding line in this area."

   Nevertheless, we agree that a quantitative performance assessment is more convincing than a qualitative one. We will add a figure showing the deviation between the existing manual and network delineations and augment the discussions as suggested.

(c) In addition, I am curious to know what new GL information you can provide by using your approach. What is the implication of using your model in mapping the GLs and improving our understanding of the GL migrations?

We have partially addressed this in the conclusion: Line 314: "We also demonstrated the ability of HED to delineate interferograms of previously unseen regions without retraining the network, which enables the timely delineation of new interferograms without manual intervention. Our delineation pipeline is independent of the source of the SAR scenes and, therefore, can be used to delineate coherent DInSAR interferograms provided by any spaceborne SAR mission."

Obtaining a dense time series of GLs enables migration analysis. We will add this point to the conclusion.

9. Figures:

(a) Please consider labeling all the subplots in each figure, and adding a subplot to show the ROI location in Antarctica.

(b) Figure 8, it is impossible to visually compare the differences between GL predictions from these two networks given the current presentation format. I suggest plotting the spatial deviations between the network predictions so we can directly see where and how much these two are deviating from each other. Again, there are multiple ways to visualize this difference.

(c) Same problem with Figure 9:

- Cross-referencing the three inset figures by just coloring the subplot figure frames is not helpful.
- On Larsen C Ice Shelf, it is impossible to see the details of network-generated GLs inside the green box in the first subplot.
- The plotting extent cut out the GLs in Totten main glacier stream, you need to expand the spatial extent.
- Why not also plot the three inset boxes in the second column?
- In the final column, you present the zoomed-in interferograms and show the manual GLs, why not plot the network-generated GLs from these two different networks so we will know the different performances of these two networks in Totten?

(d) Figure 11:

- It's difficult to compare these two outputs without putting them in the same figure or providing a distance deviation map.
- You have done an Antarctica-scale evaluation, why not include a comparison map for the whole ice sheet?

We will improve the visibility of our figures, as suggested. We will include an Antarctic-wide heat map of the deviations between AIS_cci and network delineations.

Technical Comments

Line 15: provide the mass change uncertainty for both ice sheets.

The interval of mass change uncertainty was mentioned in square brackets, as provided in Fox-Kemper et al. (2021). Line 15: "...the amount of ice mass loss from the Antarctic Ice Sheet (AIS) (4890 [4140–5640] Gt) and Greenland Ice Sheet (2670 [1800-3540] Gt)"

We will change these to ± standard deviations to improve readability.

Line 25-50: these three paragraphs need restructuring:

– Grounding line itself is a subglacial feature, please elaborate why detecting these two features is challenging and why different (surface) features can be used as proxies for the grounding line.

  We have mentioned the difficulties in detecting the grounding line in the manuscript: Line 26: "There are two main challenges when detecting the grounding line: its sub-glacial location and the short-term migration due to the tidal flexure of the ice shelf." We agree with you that we use "surface" observations, such as the tidal deformation of the ice, to detect a subsurface proxy for the grounding line, i.e., the hinge line. We will clarify this in the revised manuscript.

– You first cite Brunt et al., 2011 to say that existing methods detect grounding line proxies, then talk about using ice-penetrating radar in detecting true grounding line G which is a subglacial feature. The logic here is problematic.

  Thank you for pointing out the erroneous logic. We will restructure these sentences and bring them into the correct order.

Line 51: it is 'grounding line' not 'grounded line'.

Thank you for catching this typo. We will correct this.

Line 54: where is this research 'Ramanath Tarekere, 2022' published?

Thank you for bringing this to our attention. We will correct this reference.

Line 63-64: ICESat laser altimetry has also been used in generating grounding zone products manually by Fricker et al. (2006, 2009) and Brunt et al. (2010, 2011).

We will cite these references in the Introduction.

Line 65: I see what you are trying to say here – emphasizing DL method does not need manual intervention compared to other methods. However, I find it a bit confusing to follow the logic. Having read the first sentence, I would expect to know the research progress in using DL methods in detecting GZ, but here you directly dive into model inversion and ICESat-2 methods.

We mentioned both DL and classical techniques: Line 65: "Nonetheless, recent research endeavours have explored deep learning (DL) and classical model inversion techniques for automatic GL detection.". Nevertheless, this gives the impression that we will explain the DL methods first. We will rewrite this sentence and explain the classical methods first and the DL methods after that.

Line 74-79: In addition to laser altimetry, there are several studies that have used CryoSat-2 radar altimetry in mapping GZ automatically, such as Dawson and Bamber (2017, 2020), and Hogg et al. (2018).
We have indeed missed to cite these important references in our manuscript. Thank you for bringing this to our attention; we will include them in the revised manuscript.

Line 138: the pyTMD should be cited as Sutterley et al. (2017). Check https://pytmd.readthedocs.io/en/latest/getting_started/Citations.html
Thank you, we will cite the correct reference.

Line 175: how did you determine 0.8 as the threshold?
We performed a grid search over the thresholds $[0.5, 0.9] \in \mathbf{R}$ and chose the threshold which resulted in the smallest PoLiS distance between the AIS_cci GLs and the network delineations. However, the threshold value will likely change as we now use the delineations from an ensemble. We will mention the grid search method in the manuscript.

Line 278-279: can you explain more about this claim? Given the current evidence in this section, I don't follow how you can claim that HED relies more on the rectangular interferometric features or DEM than the non-interferometric features.
Thank you for pointing out this statement. Indeed, we cannot claim that the network relies more on the rectangular features or the DEM when non-interferometric features are present. The non-interferometric features have either a confounding or insignificant impact on the resulting delineations. We will change this sentence to the one stated above.

Figure 2: It should be differential tidal amplitude
Thank you for pointing out this inconsistency; we will correct the subplot heading.

Figure 5: I am confused about this figure:

- The subplot in the second row of the second column 'Resample Inputs', what are these two red boxes? Are these two different sampling locations that correspond to two different interferogram subsets in the third column? Also, what is the meaning of those three dots?
- I suggest replotting this figure to make it as clear as possible.

The red squares show two tiles/patches of one interferogram (the ROI being Amery Ice Shelf). The column 'Interferogram and Manual delineation tiles' shows the zoomed-in tiles highlighted by the red squares in the column 'Resampled Inputs'. The three dots indicate that the same process follows for each interferogram's tiles. We will include this explanation in the figure caption.

References

Fox-Kemper, B., Hewitt, H. T., Xiao, C., Aðalgeirsdóttir, G., Drijfhout, S. S., Edwards, T. L., Golledge, N., Hemer, M., Kopp, R., Krinner, G., Mix, A., Notz, D., Nowicki, S., Nurhati, I., Ruiz, L., Sallée, J.-B., Slangen, A., and Yu, Y.: Ocean, Cryosphere and Sea Level Change, Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, pp. 1211—-1362, 2021.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks, Artificial Intelligence Review, 56, 1513–1589, 2023.

Gourmelon, N., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Calving fronts and where to find them: a benchmark dataset and methodology for automatic glacier calving front extraction from synthetic aperture radar imagery, Earth System Science Data, 14, 4287–4313, 2022.

Guidicelli, M., Huss, M., Gabella, M., and Salzmann, N.: Spatio-temporal reconstruction of winter glacier mass balance in the Alps, Scandinavia, Central Asia and western Canada (1981–2019) using climate reanalyses and machine learning, The Cryosphere, 17, 977–1002, https://doi.org/10.5194/tc-17-977-2023, 2023.

Herrmann, O., Gourmelon, N., Seehaus, T., Maier, A., Fürst, J. J., Braun, M. H., and Christlein, V.: Out-of-the-box calving-front detection method using deep learning, The Cryosphere, 17, 4957–4977, https://doi.org/10.5194/tc-17-4957-2023, 2023.

Mohajerani, Y., Jeong, S., Scheuchl, B., Velicogna, I., Rignot, E., and Milillo, P.: Automatic delineation of glacier grounding lines in differential interferometric synthetic-aperture radar data using deep learning, Scientific reports, 11, 1–10, 2021.

Ramanath Tarekere, S.: Mapping the grounding line of Antarctica in SAR interferograms with machine learning techniques, Master's thesis, Technische Universität München, https://elib.dlr.de/189234/, 2022.