Review for "*Deep learning based automatic grounding line delineation in DInSAR interferograms*".

This study introduces a novel deep learning framework utilizing Holistically-Nested Edge Detection (HED) to map Antarctic grounding lines (GLs), marking significant progress in the use of automatic algorithms for GL mapping from various remote sensing datasets. The network's demonstrated generalization capabilities highlight its potential for high-resolution temporal and spatial mapping of grounding lines, which is crucial for identifying grounding line migrations and understanding their dynamics. This timely study has the potential to make a valuable contribution to the field. However, several major concerns need to be addressed before I can recommend it for publication in The Cryosphere.

Major Comments:

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

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.

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)?
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 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.
    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 feature combination did you use?
    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?
    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?
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.
7. Section 6.2 the importance of DEM (Line 270 and Figure 10):
    a. 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.
    b. 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.
    c. How to achieve the balance of including DEM to avoid over-reliance?
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?

b. You evaluate the in-sample trained model performance on the unseen Ross Ice Shelf interferogram 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?

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?

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?

Technical Comments:

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

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.

- 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.

Line 51: it is ‘grounding line’ not ‘grounded line’.

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

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).

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.

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).

Line 138: the pyTMD should be cited as Sutterley et al. (2017). Check https://pytmd.readthedocs.io/en/latest/getting_started/Citations.html

Line 175: how did you determine 0.8 as the threshold?

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.

Figure 2: It should be differential tidal amplitude.

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.