Thank you to the reviewers (Andreas Stockholm and Connor Shiggins) for their detailed, positive and constructive reviews of our manuscript. We have updated the manuscript accordingly. The main changes include the following:

Further clarifications:

- We clarified the cross-validation setup
- We clarified the strengths and weaknesses of the U-net further
- We clarified which steps are automated

Restructuring:

- Two paragraphs (one in introduction: L44 to L58, one in methods: L121 to L132) were merged
- We split “Data and Methods” into two separate sections
- One paragraph (L137-L153) was rewritten and restructured to make it easier to follow
- The first paragraph in the results section (L227 to L256) was moved to the methods section

Changes made to figures and tables:

- In Figure 1 the size of the black squares was increased, Scale bars were added to Figure 2 and in Figure 3 the arrow colours were swapped
- In all tables the best values were highlighted further using bold blue rather than just bold

Please see the responses to each individual reviewer comment in blue below. We believe these changes have substantially improved the manuscript.

Reviewer #1 Andreas Stokholm:

Review of the submitted manuscript; "*Mapping the extent of giant Antarctic icebergs with Deep Learning".* The manuscript investigates mapping giant icebergs around the Antarctic continent using the deep learning convolutional neural network architecture U-net for semantic segmentation and compares it with other approaches, namely the Otsu thresholding and K-means segmentation.

Thank you for a well-written manuscript with strong English and interesting results covering an interesting segmentation topic. To summarise, there are a few things that need clarification, particularly in relation to the data used for the training. Otherwise, the manuscript is of high quality.

Dear Andreas,
Thank you very much for your thorough review of our manuscript and many helpful comments and thoughts! We highly appreciate your efforts and believe that the changes we made have further improved the quality of this paper. Please see our responses to your comments below in blue. Line numbers in our response refer to the original pre-print (same line numbers as in your comments).

Major Comments

L86: Where do you have the resolution numbers from? Are you referring to SAR resolution or SAR pixel spacing because these are different and not identical to optical imagery? You should also mention the resolution type (high/medium etc..) Here is an overview of the Sentinel-1 products with resolution and pixel spacing.
https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-1-sar/products-algorithms/level-1-algorithms/products

We were referring to pixel spacing of the medium resolution. This has been rephrased throughout the paper and was added here:

L84-85: "We use the Level 1 Ground Range Detected (GRD) data at medium resolution."

L89: I think you should include more details on these preprocessing steps, e.g. I suspect you use the IPF (Instrument Processing Facility) denoising technique but it would be useful to know the version. I don't know much about radiometric calibration but a couple of words about it could be useful. Is the multilook 6x6 or 1x6? There is also already some multilooking applied if you are using GRD. Have you considered this?

For the thermal noise correction the SNAP user guide only mentions that it is done using look up tables with no further details.

The radiometric correction also uses look up tables with a range-dependent gain including the absolute calibration constant. A constant offset is also applied. (see SNAP)

We decided to apply multilooking, because it reduces speckle and we have to downscale the image size for our processing (RAM) and in order for the icebergs to fit in a 256 x 256 pixel image. As mentioned, the images of the biggest berg even have to be downsampled further

(L102-104). We select the “GR square pixel” option in SNAP, which means that we only specify the number of range looks (6) and the number of azimuth looks is computed based on the ground range spacing and the azimuth spacing. The window size is then determined by the number of range looks and the number of azimuth looks. As a result, images with approximately square pixel spacing on the ground are produced (SNAP User Guide).

In the paper we have summarised this as follows:

L92: “We also multilook the data with a factor of six to reduce speckle and image size, yielding a square pixel spacing of 240 m.”

Fig. 1: missing legend on colorbar (year). In addition, the black dots/squares are a bit difficult to see. I think you should consider adding either a larger indicator or a square with the four corner coordinates of your images. Particularly for the Eastern red line, it is difficult to see the indicators.

We have increased the size of the black dots and added ‘year’.

L204-209: reading this makes me confused about how your training, validation and test setup are divided. It may help to elaborate on what type of cross-validation you apply. As I understand it, you select a test set, which is one of the icebergs and all the images associated with it. But this changes somehow (at least the number of images varies). Between different models or epochs? In my view, the models should be tested on the same data?

Correct. Each iceberg is used as test data for one of our seven U-nets. The number of images per iceberg varies (L203-205). So, each of the seven U-nets has a slightly different number of images remaining as training data (121-152 images, L212).

The best performing hyperparameters were chosen using iceberg B42 as test data and these are the same for all seven networks.

To allow for a robust assessment (Table 2 shows how sensitive the results are to the choice of a single iceberg), we trained seven different networks, always retaining the images of one of the icebergs as test data. So, in the end we can evaluate the performance of our U-net across all seven icebergs.

Thank you for pointing out that this was not entirely understandable. We have added the following sentences to the paper and hope this clarifies any remaining uncertainties:

L204: “In the end, it allows us to evaluate the performance of our U-net across all seven icebergs, as each of them is used as (unseen) test data for one of the networks.”

L212: “Other hyper parameters like network architecture, number of layers, optimizer, initial learning rate, loss function and batch size are the same for all seven networks and were set using the B42 iceberg as test data.”

I also think it would be useful to present some information about the data distribution of icebergs, e.g. how many images of icebergs are in each iceberg group, i.e. *dark icebergs*,

coast etc. It would also be useful for the reader to understand how many images are available for each iceberg (so they are ready to examine whether there are any biases in the data).

Revisiting this comment, I realise that Table 2 has information about the number of images, I think it should be elaborated that the information is available here. I also see that Table 3 has similar information about the groupings

Correct. We have added cross-references at the end of Section 2.1 and 2.2., respectively, so the reader can find this information more quickly/early on:

"The exact number of images per iceberg is given in Table 2." and "The number of images per category is given in Table 3."

Minor Comments

L59-L70: This paragraph contains information about why you choose to utilise the U-net architecture. First, the benefits of it are highlighted followed by explaining that the paper Baumhoer et al. (2019) used to delineate ice shelf fronts with good success. The following sentence uses this as an argument for choosing the U-net. This makes it sound like it was only because of Baumhoer et al. (2019) that the architecture was chosen but I suspect this is not fully the case. I think I would rephrase it to something like "because of the many successful studies using the U-Net including one on a very similar topic (Baumhoer et al. (2019)), we choose to use the U-Net.

Thank you for the suggestion. This is indeed what we meant. We have rephrased the sentence to:

"Because of the many successful studies using U-net including one with similar challenges (Baumhoer et al., 2019), we decided to also employ a U-net. "

L63: For the authors information, there is a new entry of Stokholm et al. 2022 available, which also applies the U-net architecture. https://www.nature.com/articles/s41598-023-32467-x

Thank you. We have added this paper.

L71: I think you should consider splitting this section into a separate data section and a method section.

Good idea. We have split data and methods between 2.3 and 2.4, so 2.4 became 3.1, etc.

L82: which is frequent over the Southern Ocean? If it is clouds then it needs a reference. Alternatively, you could mention that clouds have a very similar albedo to ice (bergs) making them difficult to discriminate between in optical imagery.

We have deleted this last part of the sentence.

L90: I would like some more details on why only the HH channel was used. HV for instance is much better at differentiating between ice and water and is less affected by the measurement incidence angle. (I am not gonna ask you to redo your experiments with both channels).

We use HH, because HV is only available in some parts of the Southern Ocean and icebergs drift across these acquisition masks (see L89-90). We also added the following sentence:

"Should both modes become available across the Southern Ocean in the future, their collective use might be advantageous as icebergs and their surrounding cause different changes in polarisation, which could be exploited using e.g. the HH/HV ratio."

L99: I think you should reformulate the sentence to not include "cannot only".

Done. The sentence now reads:

"Therefore, we have a good estimate of where each of these giant icebergs should be and can firstly download targeted Sentinel-1 images containing these icebergs, and secondly crop the images around the estimated central position to a size of 256 x 256 pixels."

L101: I think you should explain briefly why icebergs are given certain names (e.g. B or C and number).

We have added a brief explanation:

"The names are determined by the NIC and indicate which quadrant in Antarctica the iceberg calved from (A-D) followed by a number (e.g. B30 was the 30th iceberg on their record that calved between 90-180° W)."

L103: what do you mean by "normal resolution"?

We mean 240 m. The text has been rephrased to clarify:

"B30 is the only iceberg that is initially longer than 37 km, so we rescale the first 27 images to 480 m resolution, until its length drops below 37 km. A further two images of this iceberg are then used at 240 m resolution (Figure 4 first column shows images of B30 at 480 m resolution and the last one at 240 m resolution)."

L121: You should add that the signal is depending on the **surface** properties, such as roughness, dielectric properties etc. Typically the angle of the satellite overpass is called the incidence angle of the measurement (also relevant in L130).

We have rephrased "angle of satellite overpass" and "viewing angle" to "(satellite) incidence angle" and "roughness" to "surface roughness".

L140: as I read this sentence, it seems like images that have coast are grouped into the *dark icebergs* category but there seems to be a group called *coast*. It is a bit confusing to read.

We have adjusted the sentence (see next comment).

L137-L153: this section was difficult to read. Lots of complicated sentence structures.

We have rewritten and restructured this section, always mentioning the groups first and then adding further explanation. The new text reads as follows:

"We visually group all input images into different categories to assess the performance in different potentially challenging conditions. These groups are *open ocean*, *sea ice*, *fragments*, *other icebergs*, *coast* and *dark icebergs* (Figure 2 shows one example each). We class an image as *dark iceberg*, if the iceberg appears as dark or does not stand out from the background (Wesche and Dierking, 2012). Images are grouped into the *coast* category, if they contain nearby ice shelves or glaciers on the Antarctic continent. Due to very similar physical conditions, ice-shelves and icebergs are hard to differentiate. The category of *other icebergs* was introduced, because in some cases, several giant icebergs drift very close to each other and both are (partially) visible in our cropped images. If another iceberg of similar size is present, the algorithms might pick the wrong iceberg and we class such images as *other icebergs*. There is also one case where a bigger iceberg is partially visible, but we are aiming to segment the largest iceberg that is fully visible (e.g. Figure 5h). We assign images to the *fragments* category if they exhibit fragments close to the iceberg. Fragments occur frequently in the vicinity of icebergs, as icebergs regularly calve smaller fragments around their edges. When the fragments are close to the main iceberg, they are easily grouped together (Koo et al., 2021). The last challenge is *sea ice*. Young and flat sea ice usually appears homogenous and dark, meaning it does not pose a problem. However, older, ridged sea ice and other cases where the background appears grey rather than black with significant structure (Mazur et al., 2017) are grouped into this category. Images are grouped into the *open ocean* category, if no obvious challenge is apparent to us. This includes cases where the sea ice is not visually apparent (i.e. young and flat) and the background appears as dark and relatively homogenous or where fragments are further away from the iceberg. If several challenges are present (e.g. if *coast* and *sea ice* are visible), we assign the image to the most relevant group. The number of images per category is given in Table 3."

Fig 3. Just for your consideration, this is a very large figure, I think it could be smaller e.g. by yellow/grey squares smaller. I would only consider swapping the red and green arrows as they are most often represented that way for the U-Net, e.g. Ronneberger et al., 2015 original paper.

Thank you for the suggestion. We have swapped the colours.

L171: Is the connected component analysis applied as part of the training, i.e. when calculating the loss and performing backpropagation? Or just as part of the results?

For the Otsu threshold (L171), there is no training needed. For the U-net, we use the greyscale probability maps without thresholding or connected component analysis during training. These two last steps are only applied before the final evaluation (L195-200).

L220-222: it says initially that the supplementary material shows the best visualisations of the results but the final bit of the sentence says all outlines for the 191 images. I guess it includes the outlines.

Yes, it contains all images with the resulting (predicted) outlines plotted on top. We have rephrased the sentence:

"The best visualisation of the results can be found in the supplementary animations (Braakmann-Folgmann, 2023), showing all 191 images with the predicted iceberg outlines from all methods plotted on top."

L235: It would be useful to report the accuracy for icebergs alone, i.e. True Positives / Total Positives. Though it will have some uncertainty associated with it as the manual delineations are not pixel-perfect.

The true positive rate (iceberg class accuracy/detection rate) TP/(TP+FN) is the same as 1-misses, so we did not list it separately (L233).

L241: seeing as I am a bit sceptical of the numbers in your pixel spacing, I would like to know what you define as the "pixel area".

We added "(240 x 240 m or 480 x 480 m)" for clarification.

Table 1. I think the boldness of the best results should be further highlighted (thicker bold) if possible.

We highlighted them further using bold blue (same for Table 2 and 3).

L293-295: I also suspect that increasing the receptive field of the network could have the model in these cases where the iceberg is relatively large.

We optimized all hyperparameters (including the number of layers) for B42, which is not the largest, but also a relatively long iceberg and found that the used architecture gave the best results. Theoretically, the receptive field of our network should also cover the largest bergs. Therefore, our suspicion is that the problem is rather that when the largest berg is kept as test data, the network never had to predict such a large number of iceberg pixels during training (L293-298). The following text was added for clarification:

L212: "Other hyper parameters like network architecture, number of layers, optimizer, initial learning rate, loss function and batch size are the same for all seven networks and were set using the B42 iceberg as test data."

Tab. 2, again I think the bold font could be increased in thickness.

We highlighted them further using bold blue

L444: I agree, a larger dataset with many more examples of icebergs would be very useful to advance the training. Considering larger receptive fields may also help. Using data augmentation could also significantly help to create more varied training data for the models.

We have also tried enlarging the receptive field further (see two comments above) and using data augmentation, but in our case it did not improve the performance. We added the following sentence in L212:

"We also tried to augment the data by flipping the training images vertically and horizontally, leading to a tripling of the training data, but we found slightly degraded performance (F_1 score for the B42 iceberg used as test data reduces from 0.88 to 0.79). We believe that this is because consecutive images already show a

similar iceberg shape and size in similar conditions, but with varying rotation and translation through the natural drift. Therefore, in this case data augmentation does not help but rather lead to overfitting."

L462: the link does not work.

We are very sorry about this. It works for us, so not sure how to fix it now, but we assume that typesetting will take care of this at the end anyway.

L494: the link does not work (https://doi.org/ appears twice)

We deleted the duplication, but for us the link actually even worked before.

L510: the link does not work

We are very sorry about this. It works for us, so not sure how to fix it now, but we assume that typesetting will take care of this at the end anyway.

L515: the link does not work (https://doi.org/ appears twice)

We deleted the duplication, but for us the link actually even worked before.

L545: the link does not appear to be a hyperlink

We are very sorry about this. It works for us, so not sure how to fix it now, but we assume that typesetting will take care of this at the end anyway.

In addition, there is a pdf document attached with grammatical suggestions.

We have adopted or considered them. Thank you for all your efforts, again.

Reviewer #2 Connor Shiggins:

The manuscript presented by A Braakmann-Folgmann and co-authors put forward a U-net approach to identify giant icebergs on different Sentinel-1 images which contain a variety of different environmental conditions. The resulting output from the U-net is then compared to manual delineations and other thresholding-based techniques (Otsu and k-means). I have made comments to try and not cross over with those made by reviewer 1.

The manuscript is generally well-written and structured, with only some tweaks required. The U-net performs well in some scenarios and provides scope for potential 'giant' iceberg identification on SAR imagery if training data is suitable. I very much enjoyed reading this manuscript and I am always excited in seeing new approaches to identifying icebergs on satellite imagery!

Dear Connor,

Thank you very much for taking the time to review our manuscript, for your careful assessment and helpful feedback! We highly appreciate your efforts and believe that the changes we made have further improved the quality of this paper. Please see our responses to your individual comments below in blue. Line numbers in our response refer to the original pre-print (same line numbers as in your comments).

While this manuscript is relatively close to publication and mostly requiring minor comments (below major), I have some key comments, the major ones stated below:

- I'm not convinced such a complex approach such as a U-net is necessarily required to identify the largest iceberg in SAR imagery as it fails (using F1 metric) in two of the six environments (dark and coast). The U-net also has lower F1 scores in two of the six environments (fragments and open ocean) when compared to Otsu and k-means. This means out of six image classifications, only two environments (sea ice and other bergs (notably)) require a U-net approach which yield the highest F1 score. In my opinion, this suggests that the U-net devised is fit for purpose in only certain SAR images and is not necessarily suitable in others (maybe due to the amount of training each environment had?), as well as other approaches yielding better F1 scores. I appreciate in other metrics (false alarms particularly) the U-net does a better job than the other approaches. In only one environmental condition in Table 3 ('other bergs'), the U-net outperforms the other approaches across the board (i.e. higher F1 score, with lower misses, false alarms, MAD). I think the narrative of the manuscript should consider that while the U-net can work suitably in some conditions, considering the amount of training it requires (and inability to identify larger icebergs than the ones identified in training) there should be caution made with regards to the success of this algorithm. I do however appreciate that the authors have been sensible with their claims and that the U-net requires more training data to be applied elsewhere.

  Thank you for bringing this up. We share your view, that it should always be assessed properly, whether a deep neural network is actually required for a certain task. Therefore, we compare U-net to two other machine learning techniques and outline where U-net is better and needed.

More specifically, looking only at Table 3 and different environmental conditions, we understand your hesitation. However, Table 2 draws another picture where U-net achieves the highest F_1 score across all icebergs except B31. For this iceberg we have discussed extensively that it represents a domain-shift and therefore conclude that U-net currently should not be applied to icebergs bigger than the training data or that the training data set should be extended (L293-298, L417-419). As Table 3 includes the results from B31, the two tables cannot be discussed separately (L367-369). A huge amount of the failures in the "fragments" category are due to the fact that U-net misses parts of B31, rather than fragments being classed as iceberg (as also seen by the high miss rate and low false alarm rate). This is discussed in L338-342. Also the open ocean category includes quite a few images from B31, which impact the average performance of U-net. For open ocean, we agree that any of the methods work fairly well. U-net only performs slightly worse overall due to the misses (11 % for U-net versus 2 and 4 % misses), but better in terms of false alarms (0.1 % for U-net versus 0.3 and 0.4 % false alarms). This is also discussed in L321-322.

Therefore, in our opinion, U-net scores similar in the open ocean category, performs as well as k-means in the fragments category (lowest false alarm rates) and scores better in the sea ice, other icebergs and coast categories. We consider this a significant improvement already. Dark icebergs are indeed still a problem for all the methods as you mentioned. We also agree that the remaining problem of missing parts of the largest bergs should be addressed and have suggested ways to do so throughout the paper and in the conclusion (L417-419, 422-424 + added sentence).

We have also made our point clearer in the paper by adding the following sentence:

L369: "Therefore, the fact that U-net misses parts of B31 also impacts its performance in mainly the fragments and open ocean category. Apart from theses misses, U-net scores at least as well as the other methods in the open ocean and fragments category (lower or same false alarm rate) and outperforms them in the sea ice, other icebergs and coast categories. Dark icebergs and larger areas of coast remain a problem for all methods."

L424: "For an operational application, in the short-term further post-processing could be implemented to filter outliers, but on the long run, we would suggest enlarging the training data set before applying it to icebergs that are smaller or larger than those currently covered by the training data."

- I'm hesitant to suggest that this approach is 'automatic'. While the segmentation is automated in the images, it requires training data which is still reliant on manual input, as well as downloading/pre-processing the SAR data used in the manuscript.

  Our method is indeed supervised (requiring training data), but the classification itself is automatic. Once a neural network has been trained, it is very quick to run (0.2 sec for 24 test images) and requires no manual input. Your point that the downloading/pre-processing of the SAR data is not automated yet is true. We believe that it wouldn't be too much effort to automate this with a script, though. We have clarified this in the paper, too:

  L106: "The current implementation still requires the user to manually find and download Sentinel-1 images, but in principle this could also be automated with a script. All pre-processing steps only rely on position and length estimates from NIC rather than actual decisions that a user has to make, paving the way for a fully automated end-to-end system."

L155: “Although the goal is to develop an automated segmentation technique, manual delineations of iceberg extent are required to train the U-net and for evaluation of all methods.”

L218: “Once trained, U-net can be applied without any user intervention and the prediction for 24 images takes 0.2 seconds.”

- Two paragraphs (one in introduction: L44 to L58, one in methods: L121 to L132) could be merged to help re-structure one of the introductory paragraphs and remove one of the methods paragraphs which is introductory information.
    - The paragraph contained in the introduction (L44 to L58) would really benefit with some re-structuring (see minor comments below to try and help) to best describe the challenges faced by automatically identifying icebergs. It is currently difficult to follow with study examples.
    - The methods paragraph (L121 to L132) is a really nice piece of text explaining the difficulty of identifying icebergs on SAR images and would complement the introductory context. In its current state, this paragraph does not describe the method at all, or even the grouping classification, as this happens in L137 to L153.

Thank you very much for this suggestion. The paragraphs have been merged and the new section in the introduction now is:

“Despite the quantity and variety of previous approaches, a range of limitations has so far hindered the operational application of an automated iceberg segmentation algorithm. One limitation is that previous studies have focused on smaller icebergs and perform worse for larger ones or are not even applicable there (Mazur et al., 2017; Wesche and Dierking, 2012; Willis et al., 1996). Our work extends previous studies with the goal to delineate specific giant icebergs. Giant icebergs make up a very small part of the total iceberg population, but hold the majority of the total ice volume (Tournadre et al., 2016), which makes them the most relevant for freshwater fluxes. Apart from iceberg size, there are many remaining challenges resulting from the variable appearance of icebergs as well as the surrounding ocean or sea ice in Synthetic Aperture Radar (SAR) imagery (Ulaby and Long., 2014): The appearance of icebergs versus the surrounding ocean or sea ice depends on their surface roughness, the dielectric properties (e.g. moisture of the ice) and the satellite incidence angle (Figure 2). Icebergs with dry, compact snow are usually bright targets in SAR images (Mazur et al., 2017; Wesche and Dierking, 2012; Young et al., 1998). While calm ocean appears as a dark surface in SAR images, wind roughened sea appears brighter depending on the relative wind direction versus the satellite viewing angle (Young et al., 1998). Therefore, many studies report degrading accuracies in high wind conditions (Frost et al., 2016; Mazur et al., 2017; Willis et al., 1996). Thin sea ice has a similar backscatter to calm sea (Young et al., 1998), but rougher first-year ice already exhibits higher backscatter and multi-year ice can reach backscatter values overlapping with the range of typical iceberg backscatter (Drinkwater, 1998). This explains why deformed sea ice or sea ice in general is also mentioned to lead to false detections (Koo et al., 2021; Mazur et al., 2017; Silva and Bigg, 2005; Wesche and Dierking, 2012; Willis et al., 1996). Surface thawing can reduce the iceberg backscatter significantly (Young and Hyland, 1997), meaning that those icebergs have the same or lower backscatter than the surrounding ocean and sea ice, and appear as dark objects (Wesche and Dierking, 2012; see our Figure 2, last column). Some of the existing techniques are therefore

limited to austral winter images (Silva and Bigg, 2005; Williams et al., 1999) and dark icebergs remain a problem for all existing methods using SAR images. Furthermore, giant tabular icebergs can exhibit a gradient (Barbat et al., 2019a) due to variations in backscatter with the incidence angle (Wesche and Dierking, 2012) or appear heterogeneous due to crevasses, (see Figure 2, third and last column), which also complicates segmentation and differentiation from the surrounding ocean and sea ice. And finally, clusters of several icebergs and iceberg fragments too close to each other have been found to pose a problem (Barbat et al., 2019b; Frost et al., 2016; Koo et al., 2021; Williams et al., 1999). Our work aims to delineate icebergs in a variety of environmental conditions as accurately as possible using a deep learning technique."

- On a similar note, the first paragraph in the results (L227 to L256) reads as methodological which describes the statistics and comparisons of the different approaches. I would suggest moving this paragraph into a methods sub-section and start the results section where currently L257 starts ('Comparing the performance of all three techniques…..').

  True. Thank you for these suggestions. We have also moved this paragraph.

Minor comments:

- Avoid using both 'iceberg' and 'berg' interchangeably. I would suggest sticking with 'iceberg' and change any existing 'berg'

  We have changed everything to "iceberg".

- L11: Could the size be put in brackets to classify what giant actually is, i.e. (> 20 km^2) - or is this just a general term?

  In the literature "giant iceberg" usually refers to those icebergs that are named and tracked by NIC, meaning longer than 18.5 km (10 nautical miles) or that encompass an area of at least 68.6 km^2 (20 square nautical miles) (L96-97). As this definition is a bit bulky though, we would rather not mention it in the abstract. We have clarified this later though:

  L96-97: "All icebergs that are longer than 18.5 km (10 nautical miles) or that encompass an area of at least 68.6 km^2 (20 square nautical miles) are named and tracked operationally every week by the National Ice Center (NIC). These are referred to as "giant" icebergs (Silva et al., 2006)."

- L12: Replace to 'second'

  Done

- L13: Put in brackets what the two approaches are for segmentation

  Done

- L26: Change 'derive' to 'delineate'

  Done

- L27: Add the references to first sentence if you're stating a number of methods have been used

  This is just an introductory/summary sentence for the paragraph. All methods and references are mentioned in the following. Citing all of them again would make the sentence very long and hard to read in our opinion.

- L27: It might be worth mentioning within the sentence the specific approaches used which will set the scene for the upcoming paragraph, i.e. 'A number of methods have been proposed, including thresholding (ref), edge detection (ref) etc.

  Done

  "A number of methods including thresholding, edge-detection and clustering techniques have been proposed to automatically detect and segment icebergs in satellite radar imagery."

  L28: Remove 'simple'

  Done

- L31: Does 'approach' need to be added after 'based'?

  No, based relates to "based on a K-distribution".

- L33: Would not include 'etc' --- either tell the reader what the parameters are or discard

  Done

- L36: Would avoid terms like 'sophisticated', 'simple', 'elaborate' as all approaches have their uses and people have different opinions on what algorithms should be classed as

  Rephrased "slightly more sophisticated" to "another"

- L38: Add 'similarly' before 'Collares et al' to connect the two sentences

  We rephrased it as follows, addressing also your next comment and added "similarly" before the next sentence:

  "A clustering technique was employed by Collares et al. (2018), who used the k-means algorithm (Macqueen, 1967) to segment icebergs, which are then manually tracked. Similarly, Koo et al. (2021) …"

- L38: Is it worth clarifying what k-distributions (as mentioned on L31) and k-means are? It currently reads as it is assumed the reader has prior knowledge of these approaches

As we are not using a k-distribution ourselves, we suggest knowing that it is a distribution is sufficient for the further understanding of the text. The interested reader is invited to take a look at the reference.

For k-means we have added a brief explanation that it is a clustering technique here (see above). As we are using it in the paper, we also added more details in the methods section on k-means:

"K-means is a clustering technique, which divides the data into k clusters iteratively. The initial cluster centres are chosen randomly and each observation is assigned to the nearest cluster. Then, in each iteration, new centroids (means) are calculated per cluster and all observations are assigned to the nearest cluster again."

- L39 to L42: Koo et al. paper is a very nice example again following the previous ones.... I think it is worth noting when discussing this paper the 'target' iceberg they track (B43 in this case) is originally delineated by a manual operator, limiting the application in terms of automation to other icebergs. I am pretty sure that is the method applied by that paper (just check), but it might be worth mentioning as it further bolsters your argument of limited approaches for 'true' automation

  Thank you for this comment. We have added it to the paper:

  "Calculating the similarity of the distance to centroid histograms of all detected icebergs to the first instance, which was manually digitized, they then track one specific giant iceberg (B43)."

- L42: Rephrase 'elaborate'

  Done:

  "Finally, Barbat et al. (2019) used a graph-based segmentation and Ensemble Forest Committee classification algorithm with a range of hand-crafted (selected by a human operator) features."

- L43: What does hand-crafted features mean? Are these manually delineated iceberg outlines?

  Hand-crafted features as opposed to neural networks, which determine the most relevant features themselves, are any features that human operators select, because they think they are useful to e.g. classify an object. This could be e.g. brightness, shape or edges.

  We have added "hand-crafted (selected by a human operator)" in brackets (see comment above).

- L44 to L58: This paragraph is the one mentioned in general comments about merging with the methods paragraph. I would suggest starting the new paragraph with a clear sentence stating what the exact problems are with 1) automated iceberg detection algorithms and 2) the application to SAR imagery. The examples can then be followed in the paragraph (i.e. L52 currently talks about 'dark icebergs' without explanation of what that term actually means --- this new re-structure will combine these approaches and hopefully clarify for the reader)

Done (see above)

- L60: Would combine the first two sentences, so replace full stop on L61 with ‘which can outperform classic…’

  “Which” would refer to the relationships then. By “They” we mean deep neural networks, though and have clarified this:

  “Deep neural networks can encode the most meaningful features themselves and are able to learn more complex non-linear relationships. Deep neural networks therefore outperform classic machine learning techniques in most tasks (LeCun et al., 2015; Schmidhuber, 2015).”

- L68: Beginning of sentence could be rephrased to: ‘As the ice-ocean interface provides similar environmental conditions to an iceberg-ocean boundary…’

  Yes, this sentence now reads:

  “The calving front to ocean boundary involves similar conditions and challenges to an iceberg to ocean boundary”

- L75: present in the image?

  Yes, added

- L80: SAR needs abbreviating in the introduction (is L50 first use?)

  True. We moved the explanation from the data to the introduction section.

- L91: The pre-processing aspect of the approach needs to include the sentence from L95 that it is conducted in SNAP

  The pre-processing also includes scaling and masking, which means that not all pre-processing steps are done in SNAP. Therefore, we mention SNAP in L95 to make it clearer that everything described before was done in SNAP and everything described after was not.

- L111: Put the temporal range of the dataset at the end of the sentence

  We would rather keep the time span and one month sampling together. Furthermore, Andreas Stokholm (reviewer 1) asked for a short explanation of the iceberg names to be added after L111. Therefore, we decided to keep this sentence as is, ending with the names, then added the explanation for the names, move on to spatial and then come to temporal coverage.

- L112: Are the first 27 images rescaled as they are within the time period of B30?

  They are rescaled because B30 is longer than 20 nautical miles during this time. Once its length drops below 20 nautical miles, we use the two remaining images of B30 at the “normal” 240 m resolution. (see L102-104, 112-114)

- L116: Consider moving temporal range to L111 as noted

  See above

- L118 to L119: Final sentence is saying basically the same as the sentence starting on L116 – either consider merging or remove

  Done. The new sentences are:

  "For each iceberg, the individual images are roughly one month apart. Far higher temporal sampling would be possible in terms of satellite image availability, but we aim to cover a wide range of environmental conditions, seasons and iceberg shapes and sizes, which are highly correlated in subsequent images."

- L121 to L132: Could be moved and merged with introductory information

  Done (see above)

- L134: Figure 2 could be placed under the new start of the section which actually describes the categories. Could scale bars also be placed on Fig 2 if possible, considering the size of these 'giant' icebergs

  Both done

- L140 to L141: Rephrase to remove use of brackets from this sentence

  Done:

  "Images are grouped into the *coast* category, if they contain nearby ice shelves or glaciers on the Antarctic continent."

- L146: Replace 'bits and pieces' with 'fragments'

  Done

- L148: Rephrase sentence to try and avoid double 'and' if possible

  Done:

  "Young and flat sea ice usually appears homogenous and dark, meaning it does not pose a problem."

- L155: The U-net requires the manual delineations, not 'we'

  Done

  "Although the goal is to develop an automated segmentation technique, manual delineations of iceberg extent are required to train the U-net and for evaluation of all methods."

- L156: Replace 'click' with 'digitise'

Done

- L156: How many iceberg outlines were manually delineated for training? This is a key part for training the network

  In all 191 images, the biggest iceberg was manually delineated. We added this in the paper, too:

  "We manually digitise the iceberg perimeter in all 191 images using GIS software to yield a polygon."

- L165: Replace 'click' with digitise

  Done

- L173: Are there any statistics used to determine the Otsu threshold was the best suited, or was it visual quantification? Particularly important as the following sentence states Otsu has never been used for iceberg detection. Could this please be clarified

  It was assessed statistically. We have added the following sentence in brackets:

  "(for the B42 iceberg we found F_1 scores of 0.58, 0.67 and 0.84 respectively)".

- L182: Apply or implement to be used instead of 'suggest'?

  Changed to "implement"

- L193: Dash for 'in-between'?

  Done

- L199: Rephrase 'would like to discard' to 'require the removal of small icebergs…'

  We changed the sentence to

  "As we are only interested in the largest iceberg, smaller icebergs and iceberg fragments are removed by also applying a connected component analysis and selecting the largest component (Figure 3)."

- L222 to L223: Struggle to follow this sentence, are you saying that your analysis combines both statistical and visual quantification to interpret the success of the approach? If so, could this please be rephrased

  Yes. This sentence now reads:

  "Our analysis in the following is based mainly on statistics, but we also show some examples to allow for a visual, qualitative assessment."

- L223: Consider rephrasing to ‘After an overall analysis, we assess the performance of the approaches for identifying each iceberg and…’

  We have rephrased it to

  “After an overall analysis, we assess the performance of the approaches on each iceberg and..”

- L224: Different environmental conditions in the scenes? Rather than ‘challenging’?

  Rephrased it.

- L227 to L256: Mentioned in general comments this is predominately methods which are important, but no results are provided

  True. We have moved this section to the methods as “3.3. Performance metrics” (data and methods were also split as suggested by Andreas Stokholm (reviewer 1), hence methods became chapter 3).

- L229: Add ‘an’ between ‘as iceberg’

  We rephrased it to “iceberg pixels”, as they are only part of the whole iceberg

- L227 to L256: I think it would benefit the study if it was clarified what an F1 score actually is and what it does (i.e. is 0 bad and 1 good? Is it a statistical comparison?) Would be easy to add a few sentences to describe the F1 score if the paragraph it resides in is moved to the methods above

  We have added the following explanation:

  “The F_1 score is a number between 0 and 1, where 1 is best and means that the model can successfully identify both positive and negative examples.”

- L239: Add references after ‘previous studies’

  Done:

  “However, some previous studies (Barbat et al., 2021; Mazur et al., 2017) have reported the MAE in area, but most (Silva and Bigg, 2005; Wesche and Dierking, 2012, 2015; Williams et al., 1999) have reported the area bias, …”

- L266: Considering the size of the icebergs which are being identified, it might be worth quoting the actual area difference between U-net and manual delineations, as well as the % difference as it will probably be a very large area difference (i.e. deviates by 12% which is X km^2)

  The numbers cannot be translated into absolute numbers, as they are averages or in the case of 12 % a 75%-quantile and each iceberg has a different size in each image. As stated in L112 the iceberg size varies greatly (from 54 – 1052 km^2), so relative numbers are a lot more meaningful.

We have added an explanation in the paper, too:

L241-242: "All area deviations are relative deviations and given in percent compared to the iceberg area in the manually derived segmentation map. Due to the large size range (54-1052 km^2) relative numbers are more meaningful."

- L275: Does 'dataset' mean the number of available images for this iceberg? I would suggest clarifying this

  Yes, done.

  "The number of images for this iceberg is smallest (15 images)…"

- L277: Consider rephrasing to: 'Furthermore, B41 remains in close proximity to its calving front for a significant period of time, which means…'

  Done

- L284: Casual phrasing 'fine', replace with U-net is 'suitable'

  Done

- L292 to L293: Does this therefore suggest that U-nets are not always required to identify icebergs, rather only in images with certain environmental conditions?

  It only suggests that U-nets struggle with icebergs larger than those used for training. This is discussed in the following (L293-298) and was clarified further in the conclusions (L422-424):

  "For an operational application, in the short-term further post-processing could be implemented to filter outliers, but on the long run, we would suggest enlarging the training data set before applying it to icebergs that are smaller or larger than those currently covered by the training data."

- L305: Table 2 – Good to see how the U-net performs for the test dataset

  Thank you

- L310: Figure 4 – Really nice figure and shows the U-net varies in terms of success, depending on the image conditions

  Thank you!

- L348: Is it not 'most' cases rather than 'some'? While U-net does better than the other two approaches for coasts, it only has an F1 score of 0.34. I'd suggest this therefore means U-net struggles in most scenarios which contain termini?

  We rephrased "some" to "certain". Our data set is too small to judge how often small or large patches of coast are visible in the images. What we observe is that U-net ignores smaller patches, but struggles if the area is too big (L 349-351).

- L361: I absolutely appreciate the fact land masks are temporally stagnant and therefore bergs in close proximity could be masked as well - however, it could be worth mentioning if the ice shelves frontal positions from Baumhoer et al. are available, they would provide a potential position to derive your own mask (for each scene and respective frontal position) which would be temporally dynamic and overcome this problem? A consideration which could be noted (I'm not saying do this by the way!)

  This is indeed a very good idea. We have added the following sentence to the paper:

  "A potential solution could be to always use the latest frontal positions from Baumhoer et al., (2019) as a dynamic land mask."

- L372: Replace 'not straightforward to compare' with 'not directly comparable'

  Done

- L408: Replace 'humans' with 'manual operators'

  Done

- L413: See key comment about the algorithm being automated with the word 'automatically' used here

  We have rephrased the sentence

  "We have developed a novel algorithm to segment giant Antarctic icebergs in Sentinel-1 images automatically" to

  "We have developed a novel algorithm to automatically segment giant Antarctic icebergs in Sentinel-1 images"

  to clarify that the segmentation is automatic, not the development of the algorithm.

- L413 o L425: I think it is worth clarifying in the conclusion that the U-net is currently fit for purpose in certain image scenarios and not currently scalable to larger (and potentially smaller?) icebergs, however as noted with more training data, there is at least scope to assess the potential of applying a similar U-net to more SAR imagery

  We have added the following clarification to the last sentence (L422-424):

  "For an operational application, in the short-term further post-processing could be implemented to filter outliers, but on the long run, we would suggest enlarging the training data set before applying it to icebergs that are smaller or larger than those currently covered by the training data."

- L429: It might just be me, but the Zenodo link doesn't seem to work (doesn't seem hyperlinked)

  Sorry! Fixed it.