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.

Mapping the extent of giant Antarctic icebergs with Deep Learning

Anne Braakmann-Folgmann[1], Andrew Shepherd[1,2], David Hogg[3], Ella Redmond[1]

[1] Centre for Polar Observation and Modelling (CPOM), University of Leeds, Leeds, LS2 9JT, UK
[2] Geography and Environment Department, Northumbria University, Newcastle upon Tyne, NE1 8ST, UK
[3] School of Computer Science, University of Leeds, Leeds, LS2 9JT, UK

Correspondence to: Anne Braakmann-Folgmann (anne.bf@gmx.de)

Abstract. Icebergs release cold, fresh meltwater and terrigenous nutrients as they drift and melt, influencing the local ocean properties and encouraging sea ice formation and biological production. To locate and quantify the fresh water flux from Antarctic icebergs, changes in their area and thickness have to be monitored along their trajectories. While the locations of large icebergs are tracked operationally by manual inspection, delineation of their extent is not. Here, we propose a U-net approach to automatically map the extent of giant icebergs in Sentinel-1 imagery. This greatly improves the efficiency compared to manual delineations, reducing the time for each outline from several minutes to less than 0.01 second. We evaluate the performance of our U-net and two state-of-the-art segmentation algorithms (Otsu and k-means) on 191 images. For icebergs~~,~~ larger than covered by the training data, we find that U-net tends to miss parts. Otherwise, U-net is more robust to scenes with complex backgrounds, ignoring sea ice, smaller ~~patches~~ regions of nearby coast or other icebergs and outperforms the other two techniques achieving an F_1 score of 0.84 and an absolute median deviation in iceberg area of 4.1 %.

1 Introduction

Icebergs influence the environment along their trajectory through the release of cold fresh water mixed with terrigenous nutrients (Duprat et al., 2016; Helly et al., 2011; Jenkins, 1999; Merino et al., 2016; Smith et al., 2007; Vernet et al., 2012). The more they melt, the higher the impact. However, this melting is not linear, but depends on the surrounding ocean temperature, current speed and many other variables that are hard to model or observe (Bigg et al., 1997; Bouhier et al., 2018; England et al., 2020; Jansen et al., 2007; Silva et al., 2006). Calculating fresh water input from satellite observations is possible and can partially be automated, but requires manual delineations of the iceberg outlines to calculate changes in iceberg area and to collocate altimetry tracks with a map of initial iceberg thickness to estimate basal melting (Braakmann-Folgmann et al., 2021, 2022). Here, we present an automated approach using a U-net (Ronneberger et al., 2015) to segment giant Antarctic icebergs in Sentinel-1 images and hence to ~~derive~~ delineate their outline and area.

A number of methods including thresholding, edge-detection and clustering techniques have been proposed to automatically detect and segment icebergs in satellite radar imagery. Early work by Willis et al. (1996) was based on a ~~simple~~ thresholding technique and limited to certain iceberg sizes of a few hundred meters and certain wind conditions. Later, the Constant False Alarm Rate (CFAR) thresholding technique has been applied to detect icebergs in the Arctic (Frost et al., 2016; Gill, 2001;

Power et al., 2001). Wesche and Dierking (2012) also used a threshold based on a K-distribution fitted to observed backscatter coefficients of icebergs, sea ice and open ocean followed by morphological operations. Mazur et al. (2017) developed an algorithm for iceberg detection in the Weddell Sea based on thresholds for e.g. brightness, shape and~~,~~ size~~, etc.~~ at five scale levels applied to ENVISAT ASAR data. Apart from thresholding, edge-detection techniques have been applied: Williams et al. (1999) used a standard edge-detection technique followed by pixel bonding (Sephton et al., 1994) applied to ERS-1 images during austral winter to detect and segment icebergs in East Antarctica. Silva and Bigg (2005) extended this to ENVISAT images and improved the algorithm by using another ~~slightly more sophisticated~~ edge detection technique followed by a watershed segmentation and a classification step that takes area and shape into consideration, but also requires manual interventions. 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) employ a built-in ~~segmentation~~ technique similar to k-means using Google Earth Engine to segment Sentinel-1 images and then apply an incidence angle-dependent brightness threshold to find icebergs. 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). ~~The most elaborate algorithm has been proposed by~~ Finally, Barbat et al. (2019) us~~ed~~ing a graph-based segmentation and Ensemble Forest Committee classification algorithm with a range of hand-crafted (selected by a human operator) features.

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. ~~Overall,~~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 (). 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). ~~Many~~Therefore, many studies ~~also~~ 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 d~~D~~eformed 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 , last column). Some of the existing

techniques are therefore limited to austral winter images ~~and still require manual intervention~~ (Silva and Bigg, 2005; Williams et al., 1999) and d~~. D~~ark 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 , 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.

~~Some of the existing techniques are therefore limited to austral winter images and still require manual intervention (Silva and Bigg, 2005; Williams et al., 1999). Dark icebergs remain a problem for all existing methods using SAR images. Many studies also report degrading accuracies in high wind conditions (Frost et al., 2016; Mazur et al., 2017; Willis et al., 1996). 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). And finally clusters of several bergs and berg 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.~~

Deep neural networks can encode the most meaningful features themselves and are able to learn more complex non-linear relationships. ~~They~~ Deep neural networks therefore outperform classic machine learning techniques in most tasks (LeCun et al., 2015; Schmidhuber, 2015). U-net is a neural network that was originally developed for biomedical image segmentation (Ronneberger et al., 2015). It has since been applied to many other domains including satellite images and polar science (Andersson et al., 2021; Baumhoer et al., 2019; Dirscherl et al., 2021; Kucik and Stokholm, 2023; Mohajerani et al., 2019, 2021; Poliyapram et al., 2019; Singh et al., 2020; Stokholm et al., 2022; Surawy-Stepney et al., 2023; Zhang et al., 2019). U-net works well with few training examples, trains quickly and still achieves very good results (Ronneberger et al., 2015). A comparison between three network architectures (Deeplab, DenseNet and U-net) for river ice segmentation found that U-net provided the best balance between quantitative performance and good generalization (Singh et al., 2020). Baumhoer et al. (2019) used a U-net architecture to automatically delineate ice shelf fronts in Sentinel-1 images with good success (108 m average deviation). ~~As t~~The calving front to ocean boundary ~~looks very similar~~involves similar conditions and challenges to an iceberg to ocean boundary~~ and both goals have to deal with comparable problems like near by sea ice and varying appearance of the ice, ocean and sea ice surfaces~~. 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.

2 Data ~~and methods~~

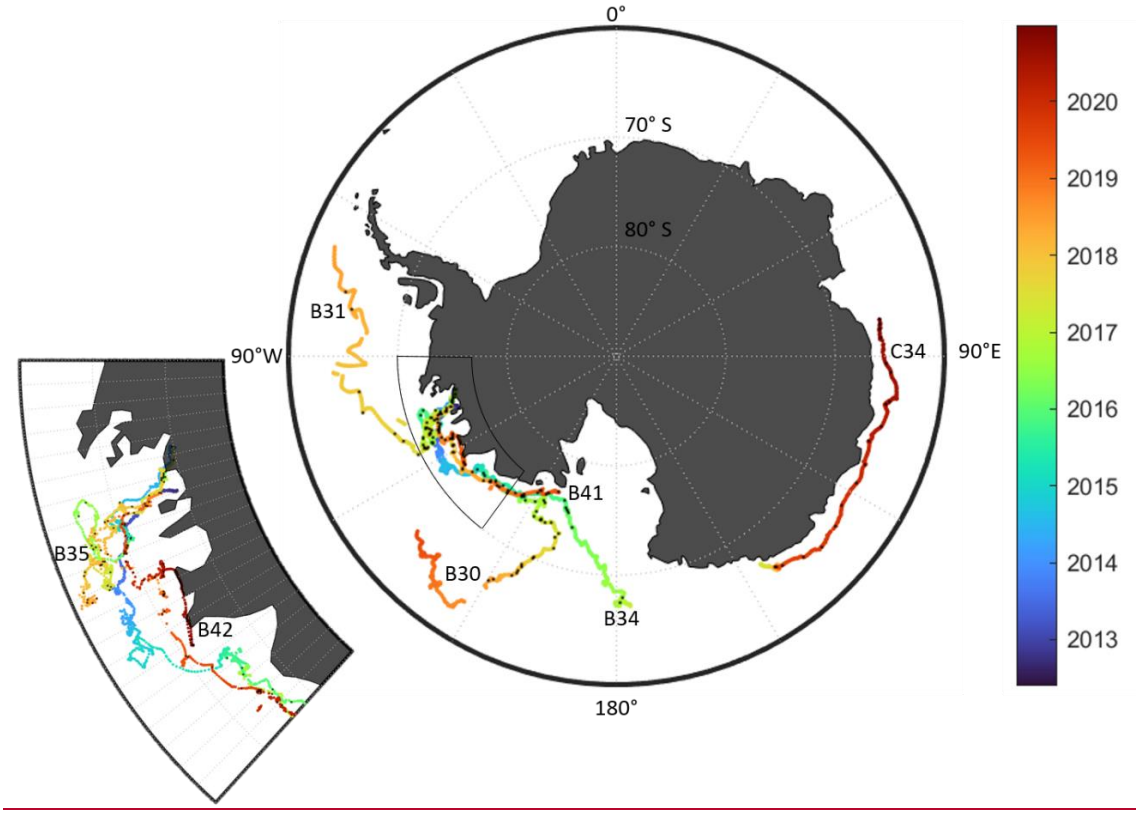
This section describes the Sentinel-1 input data and~~,~~ generation of the manually derived outlines for training, validation and testing. The goal is to derive the outlines of Antarctic icebergs, which are large enough to receive a name and to be tracked operationally. Therefore, we generate a binary segmentation map, where the biggest iceberg present in the image is selected and everything else – including smaller icebergs, iceberg fragments and adjacent land ice – is considered as background. This approach differs from most previous work, where the goal has been to find all icebergs and is targeted to monitor changes in area of these large icebergs, but also to track how the icebergs rotate and to use their outline to automatically colocate altimetry overpasses (Braakmann-Folgmann et al., 2022).~~, the implementation of two standard segmentation methods and our U-net architecture. The goal is to derive the outlines of Antarctic icebergs, which are large enough to receive a name and to be tracked operationally. Therefore, we aim to generate a binary segmentation map, where the biggest iceberg present is selected and everything else – including smaller icebergs, iceberg fragments and adjacent land ice – is considered as background. This approach differs from most previous work, where the goal has been to find all icebergs and is targeted to monitor changes in area of these large bergs, but also to track how the icebergs rotate and to use their outline to automatically colocate altimetry overpasses (Braakmann-Folgmann et al., 2022).~~

2.1. Sentinel-1 input imagery

The Sentinel-1 satellites measure the backscatter of the surface beneath them using ~~Synthetic Aperture Radar (~~SAR~~)~~. In contrast to optical imagery, SAR provides data throughout the polar night and independent of cloud cover (Ulaby and Long., 2014)~~, which is frequent over the Southern Ocean~~. The Sentinel satellites are an operational satellite system with free data availability (Torres et al., 2012). Sentinel-1a (2014-present) and Sentinel-1b (2016-2022) had a combined repeat cycle of 6 days (Torres et al., 2012), but the polar regions are sampled more frequently. We use the Level 1 Ground Range Detected (GRD) data at medium resolution. Depending on the geographic location around Antarctica, data are~~is~~ collected in either interferometric wide (IW) or extra wide (EW) swath mode. IW is a 250 km wide swath with 5 x 20 m native spatial resolution and EW is a 400 km wide swath with 20 x 40 m native resolution. We use both modes depending on availability. While HH (horizontal transmit and horizontal receive) polarised data are~~is~~ available across the Southern Ocean, HV (horizontal transmit and vertical receive) data are~~is~~ only available in some parts. As icebergs drift across these acquisition masks and HH has been found to give the best results for iceberg detection (Sandven et al., 2007), we use the HH polarised data only. 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.

We pre-process and crop the Sentinel-1 images before applying the segmentation techniques. First, we apply the precise orbit file, remove thermal noise and apply a radiometric calibration. We also multilook the data with a factor of six to reduce speckle and image size, yielding a square ~~spatial resolution~~pixel spacing of 240 m. Then we apply a terrain correction using the GETASSE30 (Global Earth Topography And Sea Surface Elevation at 30 arc second resolution) digital elevation model and

project the output on a polar stereographic map with true latitude of 71°S. These pre-processing steps are conducted in the Sentinel Application Platform (SNAP). 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). Also slightly smaller icebergs (longer than 6 km) are tracked by the Brigham Young University (Budge and Long, 2018), who release daily positions every few years. Therefore, we have a good estimate of where each of these giant icebergs should be and can~~not only~~ firstly download targeted Sentinel-1 images containing these icebergs, ~~but also~~and secondly crop the images around the estimated central position to a size of 256 x 256 pixels. Hence, every input image contains a giant target iceberg. Some images contain several icebergs and in this case, we are only interested in the largest one. To ensure that the largest icebergs fit within the image, we rescale images of icebergs with a major axis longer than 37 km (20 nautical miles). As the NIC also provides estimates of the semi major axes lengths, we apply the rescaling based on this. The rescaled images have a pixel ~~resolution~~ spacing of 480 m instead. For all input images, we scale the backscatter between the 1^{st} and 99^{th} percentile to enhance the contrast. In this step, we also replace pixels outside the satellite scene coverage with ones, and create a mask to discard the same pixels from the predictions. 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.


0°
70° S
80° S
90°W
90°E
180°
B31
C34
B41
B30
B34
B35
B42
2020
2019
2018
2017
2016
2015
2014
2013

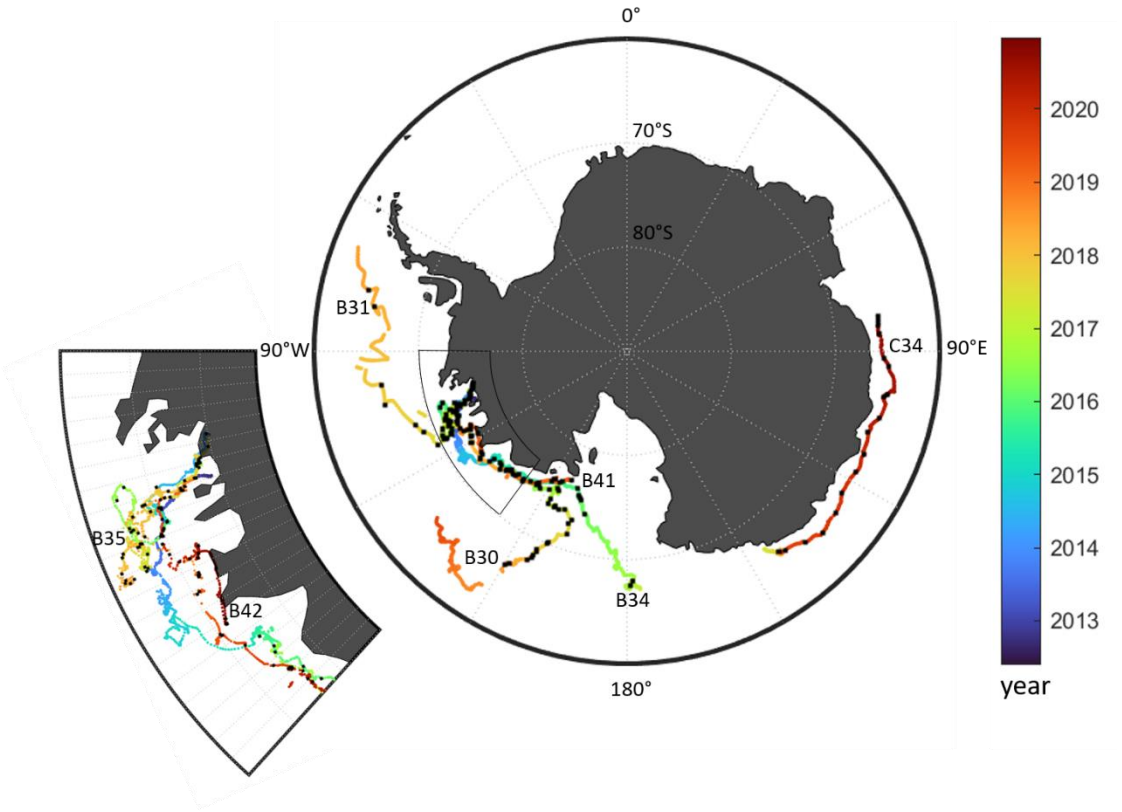
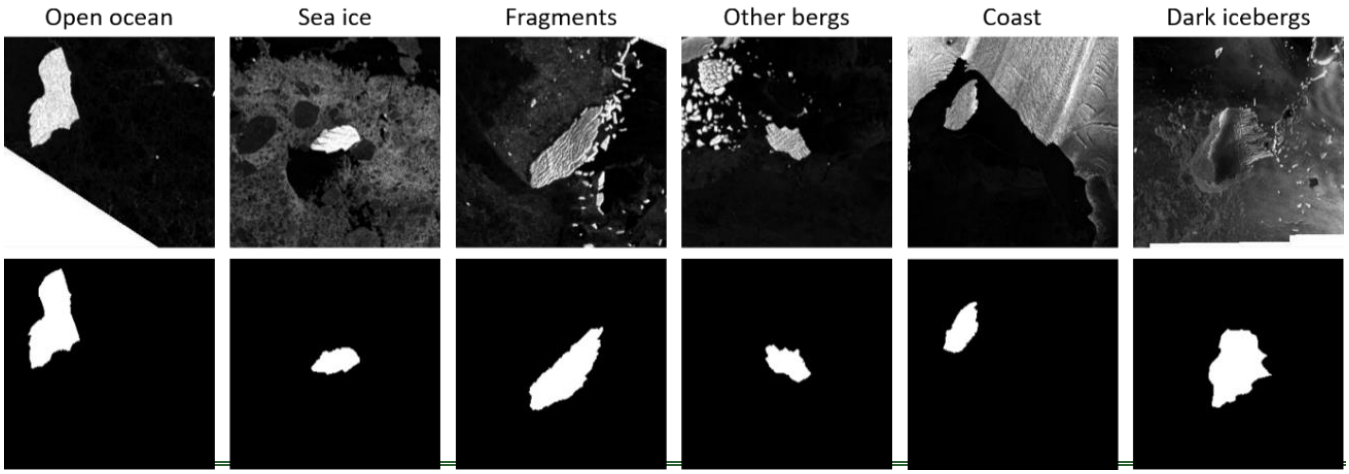


Figure 1: Spatial and temporal coverage of our dataset: The trajectories (by Budge and Long, 2018) of the seven selected icebergs are colour-coded according to time and black squares indicate the locations of the images used in this study.

The overall dataset consists of 191 images, showing seven giant icebergs: B30, B31, B34, B35, B41, B42 and C34. 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).The seven icebergs used in our study~~se~~ are between 54 and 1052 km^2 in size. 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 ~~normal~~ 240 m resolution (Figure 4 first column shows ~~rescaled~~ images of B30 at 480 m resolution and the last one at ~~normal~~ 240 m resolution). Spatially, we cover different parts of the Southern Ocean including the Pacific and Indian Ocean side with a focus on the Amundsen Sea (see Figure 1). Temporally, our images span the years 2014-2020 and are scattered across all seasons. 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~~. As these are~~ highly correlated in subsequent images~~, we decided to use only one image per month~~. The exact number of images per iceberg is given in Table 2.

2.2. Grouping of input images according to environmental conditions

~~The appearance of icebergs versus the surrounding ocean or sea ice depends on their roughness, the dielectric properties (e.g. moisture of the ice) and the angle of satellite overpass (Figure 2). 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). 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). Icebergs with dry, compact snow are usually bright targets in SAR images (Mazur et al., 2017; Wesche and Dierking, 2012; Young et al., 1998). However, 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). Furthermore, giant tabular icebergs can exhibit a gradient (Barbat et al., 2019a) due to variations in backscatter with the viewing 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.~~



~~**Figure 2: Examples of input images (top row) and segmentation maps based on manually derived delineations (bottom row) in different environmental conditions. From left to right these are B31 in open ocean, B41 surrounded by sea ice, B42 with nearby fragments, C34 and another similar-sized iceberg, B41 close to the coast and B30 appearing dark.**~~

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~~, because both have a similar intensity of grey, making it hard to pick out the berg~~ (Wesche and Dierking, 2012). Images are grouped into the *coast* category, if they ~~that~~ contain ~~coast (i.e.~~ nearby ice shelves or glaciers on the Antarctic continent~~) are grouped into this category~~. Due to very similar physical conditions, ice-shelves and icebergs are hard to differentiate. The category of *other icebergs* was introduced, because i~~I~~n 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 ~~therefore we introduce one~~we ~~group~~ class such images as ~~of~~ *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). ~~Fragments occur frequently in the vicinity of icebergs, as icebergs regularly calve smaller bits and pieces around their edges.~~ We assign images to ~~this~~ the *fragments* category if ~~the~~ they exhibit fragments ~~pose a challenge because they are so~~ close to the iceberg. Fragments occur frequently in the vicinity of icebergs, as icebergs regularly calve smaller ~~bits and pieces~~fragments around their edges.~~,~~ When the fragments are close to the main iceberg, ~~that~~ 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 ~~and~~ 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 ~~only contains~~where fragments ~~that~~ are further away from the iceberg ~~and hence there is no obvious challenge apparent to us, we class these images as *open ocean*.~~ .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.

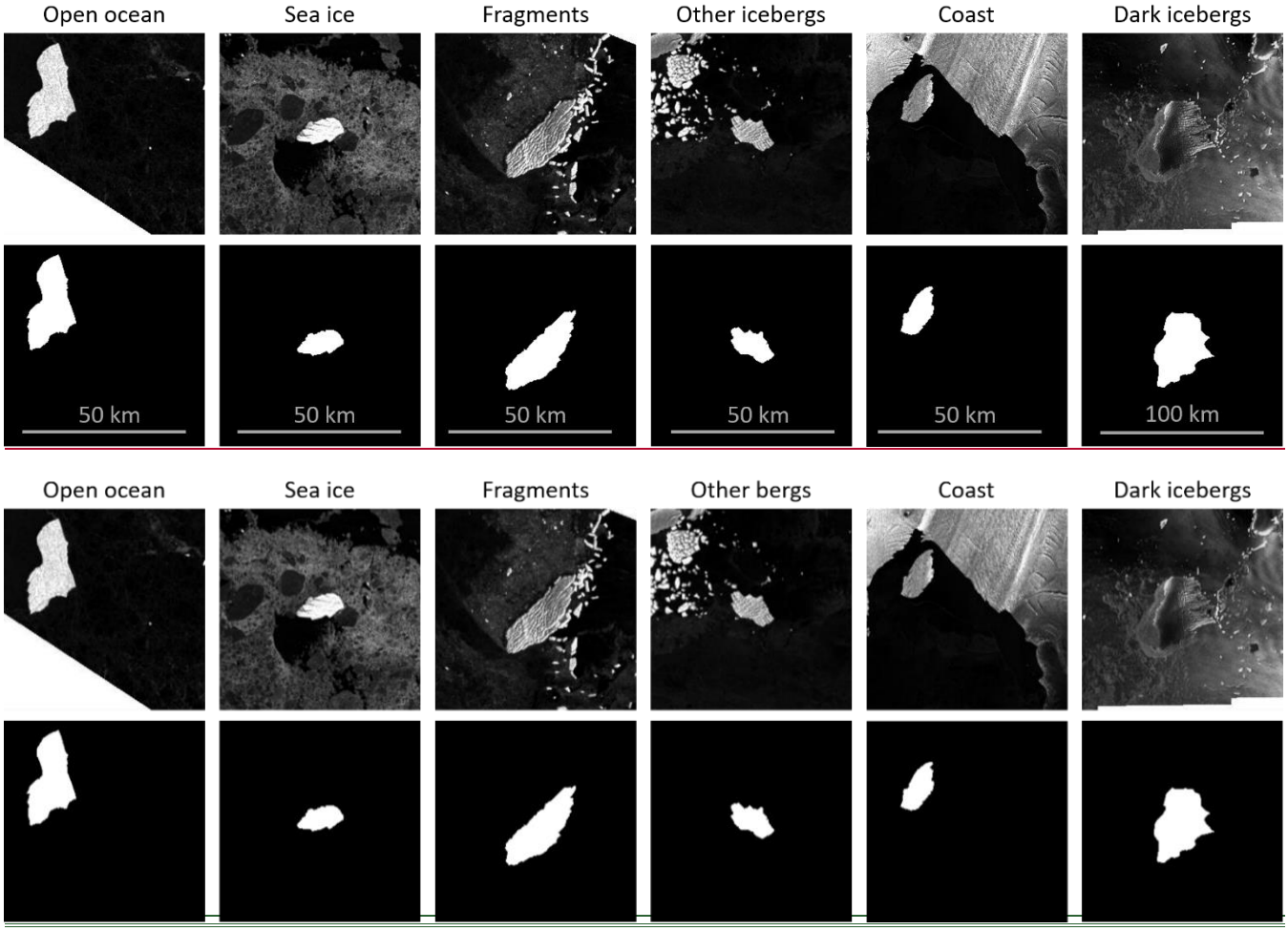


Figure 2: Examples of input images (top row) and segmentation maps based on manually derived delineations (bottom row) in different environmental conditions. From left to right these are B31 in open ocean, B41 surrounded by sea ice, B42 with nearby fragments, C34 and another similar sized iceberg, B41 close to the coast and B30 appearing dark.

2.3. Manual delineation of iceberg perimeters

Although the goal is to develop an automated segmentation technique, ~~we require~~ manual delineations of iceberg extent are required to~~for~~ train~~ing~~ the U-net and for evaluation of all methods. We manually ~~click~~ digitise the iceberg perimeter in all 191 images using~~in~~ GIS software to yield a polygon. The accuracy of such manual delineations is estimated to be 2-4 % of the iceberg area (Bouhier et al., 2018; Braakmann-Folgmann et al., 2021, 2022). We then create a binary map of the same size as the input image, where pixels within the manually derived polygon are defined as iceberg and everything else as background to allow a rapid evaluation of performance. Some examples of input images and their corresponding segmentation maps based on the manual outlines are shown in Figure 2. We regard the manually derived outlines as the most accurate and use these

binary maps to train our neural network and to evaluate all automated segmentation techniques. When the area deviation of our automated segmentation techniques drops below 2-4 %, their prediction might be more accurate than the manual delineation. In any case, automated approaches are advantageous over manual delineations – especially when rolled out for numerous icebergs or in operational applications, as each outline takes several minutes to ~~click~~ manually.

3 Methods

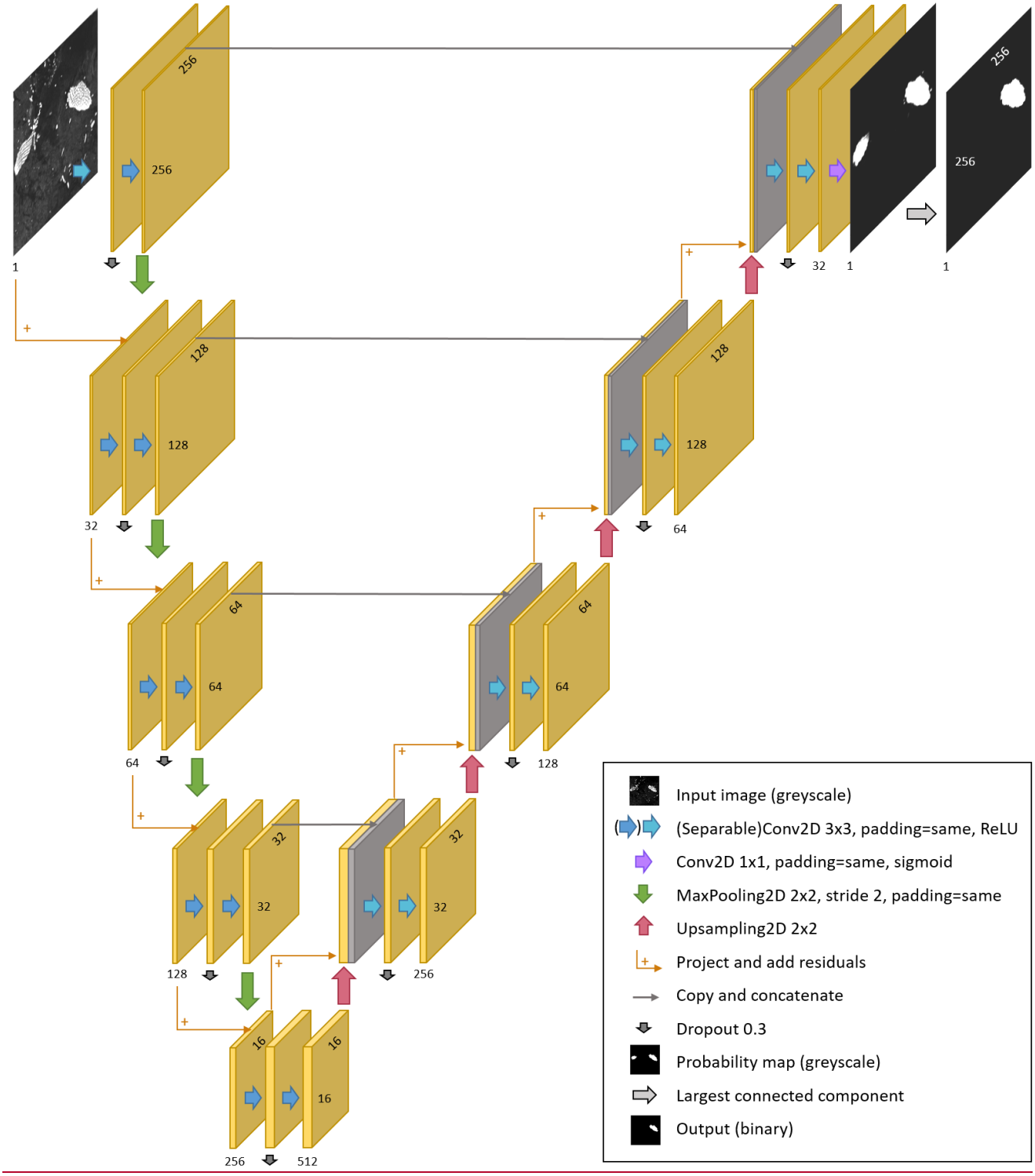
This section describes the implementation of two standard segmentation methods and our U-net architecture. We also introduce the different performance metrics used to evaluate our results.~~The goal is to derive the outlines of Antarctic icebergs, which are large enough to receive a name and to be tracked operationally. Therefore, we aim to generate a binary segmentation map, where the biggest iceberg present is selected and everything else – including smaller icebergs, iceberg fragments and adjacent land ice – is considered as background. This approach differs from most previous work, where the goal has been to find all icebergs and is targeted to monitor changes in area of these large bergs, but also to track how the icebergs rotate and to use their outline to automatically colocate altimetry overpasses (Braakmann-Folgmann et al., 2022).~~

3.1.~~2.4~~. Iceberg segmentation with k-means and Otsu

We implement two standard segmentation techniques as a baseline: Otsu thresholding and k-means. In both cases, we mask out the areas that had no satellite scene coverage by setting them to zero (black). For the first segmentation technique, we smooth the input image with a 5x5 Gaussian kernel. Then we apply the Otsu threshold (Otsu, 1979) yielding a binary image. The Otsu threshold is determined automatically based on the image's greyscale histogram so that the within-class variance is minimised. To find an iceberg, we apply connected component analysis to the binary image and select the largest component. We also experimented with other thresholding techniques including adaptive mean and adaptive Gaussian thresholding, but found that the Otsu threshold gave the best results (for the B42 iceberg we found F_1 scores of 0.58, 0.67 and 0.84 respectively).~~.~~ Although different thresholding techniques have been proposed for iceberg detection (Frost et al., 2016; Mazur et al., 2017; Power et al., 2001; Wesche and Dierking, 2012; Willis et al., 1996), to our knowledge none of them have used the Otsu method. The second technique is k-means (Macqueen, 1967) with k=2. 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. We ~~use random centre initialisation and~~ run the algorithm for 20 iterations. We also repeat this 50 times with different initialisations and take the result with the best compactness. Afterwards, we ~~also~~ perform a connected component analysis and select the largest component. K-means and a variation of it have also been applied to track selected icebergs by Collares et al. (2018) and Koo et al. (2021) respectively. Both our standard segmentation techniques are implemented using the OpenCV library (Bradski, 2000) for Python.

3.2.~~2.5~~. Iceberg segmentation with U-net

We ~~suggest~~ implement a U-net architecture to segment Sentinel-1 input images into the largest iceberg and background, which is based on the original U-net (Ronneberger et al., 2015) with some modifications. The input images are 256 x 256 one-channel backscatter images (as described in Section 2.1. and shown in Figure 2). The U-net is composed of an encoder that produces a compressed representation of the input image followed by a decoder that constructs a segmentation map from the compressed encoding with the same spatial resolution as the input (Figure 3). The encoder uses a number of convolutional and pooling layers to generate feature maps at increasing levels of abstraction and spatial scale. The decoder uses further convolutional layers and upsampling to construct the required segmentation map. Cross-links convey feature maps from different spatial scales in the encoder to the respective decoder stage, where they are combined with contextual feature maps from the decoder layer below. This allows U-net to produce accurate segmentations whilst also considering contextual features. We use padding in the convolutions and pooling operations, so that the feature maps remain the same size as the input at each level (spatial scale) and reduce by 50% in height and width between encoder levels. We also use depth-wise separable convolutions (Chollet, 2017), which are more efficient. Furthermore, we added dropout of 0.3 in ~~-~~between the two convolutions per level to avoid over-fitting (Srivastava et al., 2014) and residual connections to aid the learning process and increase the accuracy (He et al., 2016). The outputs are one-channel 256 x 256 arrays, representing the probability that each pixel belongs to the iceberg class. During training these output maps are compared with the segmentation maps from our manually derived outlines to alter the network parameters accordingly. When evaluating the validation and test data output, we convert the probability map to a binary output, where 1 corresponds to the iceberg class and 0 to background (everything else), by thresholding it at 0.5. As we are only interested in the largest iceberg, ~~and would like to discard other~~ smaller icebergs and iceberg fragments ~~around~~ ~~, we also~~ are removed by also applying a connected component analysis and selecting the largest component (Figure 3).


256
256
1
1
32
1
256
256
128
128
128
128
32
64
64
64
64
64
128
32
32
32
32
128
256
16
16
16
256
512
Input image (greyscale)
(Separable)Conv2D 3x3, padding=same, ReLU
Conv2D 1x1, padding=same, sigmoid
MaxPooling2D 2x2, stride 2, padding=same
Upsampling2D 2x2
Project and add residuals
Copy and concatenate
Dropout 0.3
Probability map (greyscale)
Largest connected component
Output (binary)

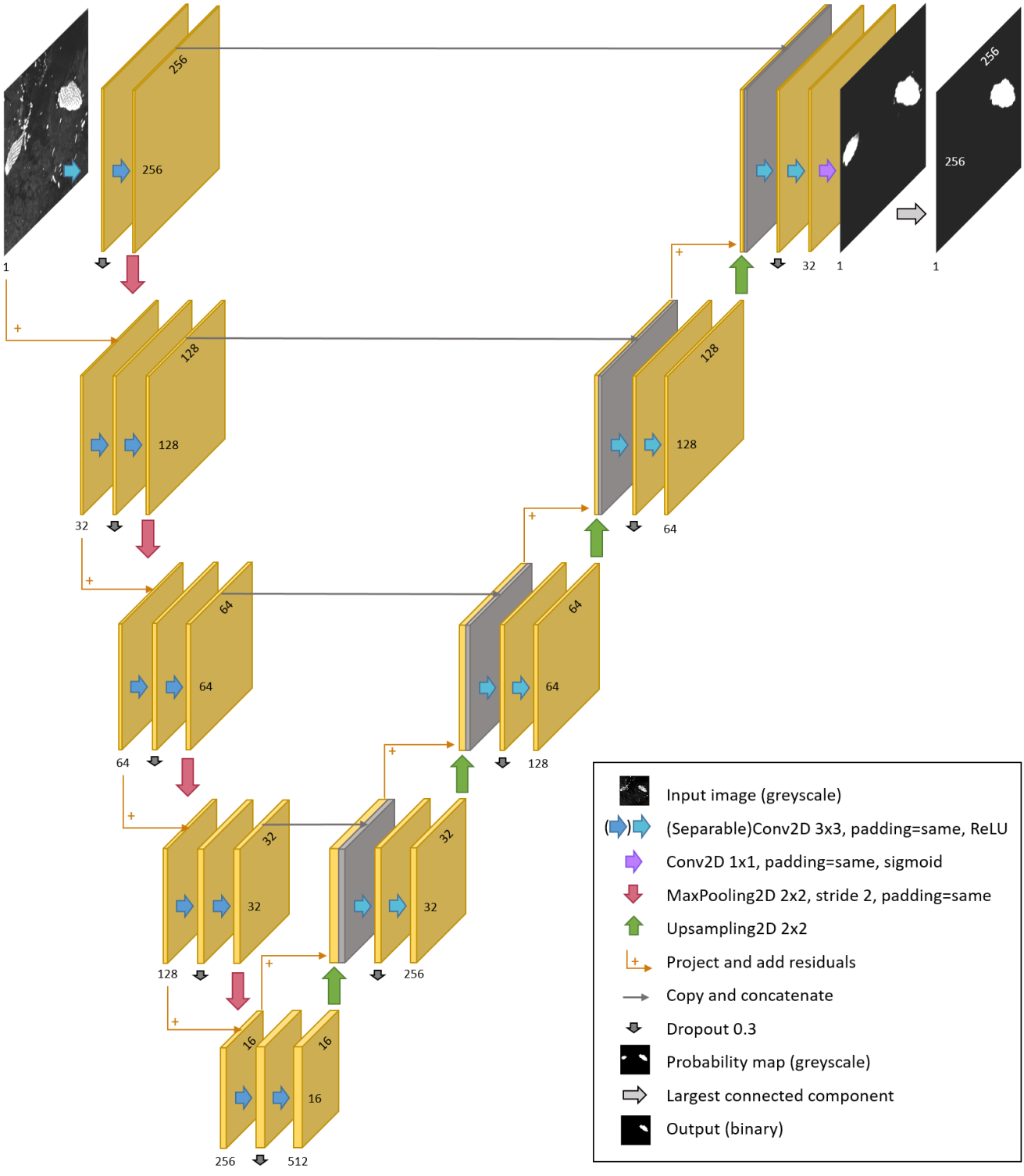


Figure 3: Modified U-net architecture as used in this paper

We train and evaluate the network using cross-validation. This means that we train seven different neural networks and always retain the images of one iceberg for testing as an independent dataset~~.~~. 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. The exact number of test images varies, as we have between 15 and 46 images per iceberg (Table 2). Although the images are roughly one month apart and cover a wide range of seasons and surroundings overall (e.g. near the calving front, surrounded by sea ice and within open ocean), we find that consecutive images of the same iceberg are often similar – both concerning iceberg shape, size and appearance as well as the surrounding. Therefore, we do not mix training and test data. On the other hand, and for the same reason, we find that it stabilises the training process, if we draw training and validation data from the same set of icebergs. 24 images are taken as validation data, which is used to ~~set the best performing hyperparameters (i.e. network architecture, number of layers, optimizer, learning rate, loss function and batch size). It also~~ determine~~s~~ when we stop the learning process to avoid overfitting. Depending on which iceberg was picked for testing, this leaves between 121-152 images for training. 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. 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. We train the network~~s~~ end-to-end using a binary cross entropy loss function and a batch size of one. Higher batch sizes had little impact on the performance and run time. The Adam optimizer (Kingma and Ba, 2015) is employed with an initial learning rate of 0.001. The learning rate is halved when the validation loss has not decreased for eight consecutive epochs. Training is stopped when the validation loss has not improved for 20 epochs. In practice, this means that the networks are trained for 57-193 epochs. The implementation is done in Python using Keras (Chollet and Others, 2015). Training takes up to 20 minutes on a Tesla P100 GPU with 25 GB RAM (Google Colab Pro). Once trained, U-net can be applied without any user intervention and t~~T~~he prediction for 24 images takes 0.2 seconds.

3.3. Performance metrics

We evaluate the performance of the three methods compared to the manual delineations using a range of metrics. True positives (TP) are all correctly classified iceberg pixels and true negatives (TN) are all correctly classified background pixels. False positives (FP) are pixels that were classified as iceberg pixels, but belong to the background according to manual delineations and false negatives (FN) are iceberg pixels in the manually derived segmentation map, which the algorithm has missed and erroneously classified as background. These are the basis for most evaluation metrics including the overall accuracy, the F_1 score (also known as dice coefficient), misses (also known as false negative rate) and false alarms (also known as false positive rate). The detection rate is equal to the iceberg class accuracy and can be derived from 1-misses; hence, we do not list it separately. 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. In the case of a large class imbalance, the F_1 score is much more meaningful than the overall accuracy. The iceberg class makes up only 5 % of all pixels, so we focus on the F_1 score, but list the overall accuracy for completeness. Except the F_1 score, all measures are given in percent. In addition to these metrics commonly used to evaluate segmentation algorithms, we also examine the accuracy of the resulting area estimates a_i. We calculate the mean absolute error (MAE) in area, the mean error (area bias) and the median absolute deviation (MAD) in area. We focus on the MAD, as it is robust to a few complete failures. 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, so we also list these for completeness. Areas a_i and α_i are calculated as the sum of all iceberg pixels in the prediction and manually derived segmentation map respectively multiplied by the pixel area (240 x 240 m or 480 x 480 m). 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. We also calculate the standard deviation for each metric. Only the MAD is given with the 25 % and 75 % quantiles instead.

$$F_1 = \frac{2\,TP}{2TP+FN+FP} \quad (1)$$

$$\text{Overall accuracy} = \frac{TN+TP}{TN+TP+FN+FP} \quad (2)$$

$$\text{Misses} = \frac{FN}{FN+TP} \quad (3)$$

$$\text{False alarms} = \frac{FP}{FP+TN} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|a_i-\alpha_i|}{\alpha_i} \quad (5)$$

$$\text{Area bias} = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i-\alpha_i}{\alpha_i} \quad (6)$$

$$MAD = \text{median}\left(\frac{|a_i-\alpha_i|}{\alpha_i}\right) \quad (7)$$

~~4~~3 Results and discussion

In this section, we present and discuss the results from the three different approaches (U-net, Otsu and k-means). 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~~and for all 191 images~~. There is one animation per iceberg. Our analysis in the following is based mainly on statistics, but we also show some examples to allow for a visual, qualitative assessment. After an overall analysis, we assess the performance of the approaches on~~for~~ each iceberg and evaluate the impact of iceberg size and different ~~challenging~~ environmental conditions in the scene~~s~~. Finally, we compare our results to previous studies.

43.1. Performance of the three methods

We evaluate the performance of the three methods compared to the manual delineations using a range of metrics. True positives (TP) are all correctly classified iceberg pixels and true negatives (TN) are all correctly classified background pixels. False positives (FP) are pixels that were classified as iceberg, but belong to the background according to manual delineations and false negatives (FN) are iceberg pixels in the manually derived segmentation map, which the algorithm has missed and erroneously classified as background. These are the basis for most evaluation metrics including the overall accuracy, the F_1 score (also known as dice coefficient), misses (also known as false negative rate) and false alarms (also known as false positive rate). The detection rate is equal to the iceberg class accuracy and can be derived from 1 misses; hence, we do not list it separately. In the case of a large class imbalance, the F_1 score is much more meaningful than the overall accuracy. The iceberg class makes up only 5 % of all pixels, so we focus on the F_1 score, but list the overall accuracy for completeness. Except the F_1 score, all measures are given in percent. In addition to these metrics commonly used to evaluate segmentation algorithms, we also examine the accuracy of the resulting area estimates a_i. We calculate the mean absolute error (MAE) in area, the mean error (area bias) and the median absolute deviation (MAD) in area. We focus on the MAD, as it is robust to a few complete failures. However, some previous studies have reported the MAE in area, but most have reported the area bias, so we also list these for completeness. Areas a_i and α_i are calculated as the sum of all iceberg pixels in the prediction and manually derived segmentation map respectively multiplied by the pixel area. All area deviations are relative deviations and given in percent compared to the iceberg area in the manually derived segmentation map. We also calculate the standard deviation for each metric. Only the MAD is given with the 25 % and 75 % quantiles instead.

$$F_1 = \frac{2\,TP}{2TP+FN+FP} \quad (1)$$

$$\text{Overall accuracy} = \frac{TN+TP}{TN+TP+FN+FP} \quad (2)$$

$$\text{Misses} = \frac{FN}{FN+TP} \quad (3)$$

$$\text{False alarms} = \frac{FP}{FP+TN} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|a_i-\alpha_i|}{\alpha_i} \quad (5)$$

$$\text{Area bias} = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i-\alpha_i}{\alpha_i} \quad (6)$$

$$MAD = \text{median}\left(\frac{|a_i-\alpha_i|}{\alpha_i}\right) \quad (7)$$

Comparing the performance of all three techniques, we find that U-net outperforms Otsu and k-means in most metrics. It achieves a significantly higher F_1 score (0.84 compared to 0.62, Table 1) and generates many fewer false alarms (0.4 % instead of 4.7 and 5.2 %). On the other hand, both standard segmentation methods have fewer misses than U-net (9 % and 13 %

compared to 21 %). On this metric Otsu scores best. In terms of iceberg area, the predictions by U-net are much closer to the manually derived outlines in terms of MAE and bias. Otsu and k-means clearly suffer from a few total failures with over 100 % deviation, which bias these metrics in their cases. The MAD, which is less sensitive to such outliers, is similar for the three methods, with Otsu scoring best (3.6 %), followed by U-net (4.1 %) and k-means (5.1 %). The 25 %-quantiles are very similar for all three methods (2.0, 2.1 and 2.2 % respectively). On the 75 %-quantiles, U-net achieves slightly better results (12.1 % area deviation, compared to 13.8 % and 14.9 % for k-means and Otsu). This means that 75 % of all U-net predictions deviate from the manually derived area by 12.1 % or less. -Overall, U-net scores better in most categories, but tends to miss parts and misclassify iceberg as background.

Table 1: Performance metrics with standard deviations of U-net, Otsu and k-means across all test data sets (191 images). The median absolute area deviation (MAD) is given with 25 % and 75 % quantiles instead of standard deviation. Arrows indicate whether high (up) or low (down) numbers are desirable. The best score per metric is highlighted in bold blue.

	F_1 score ↑	Overall accuracy [%] ↑	Misses [%] ↓	False Alarms [%] ↓	MAE in area [%] ↓	Area bias [%] ↓	MAD in area [%] ↓
U-net	**0.84 ± 0.30**	**99 ± 2**	21 ± 32	**0.4 ± 0.3**	**15 ± 26**	**-5 ± 29**	4.1 [2.1 – **12.1**]
Otsu	0.62 ± 0.34	95 ± 13	**9 ± 28**	5.2 ± 0.3	170 ± 490	170 ± 490	**3.6 [2.0** - 14.9]
k-means	0.62 ± 0.33	95 ± 12	13 ± 28	4.7 ± 0.3	150 ± 460	150 ± 460	5.1 [2.2 – 13.8]

43.2. Impact of iceberg size

Next, we evaluate how U-net performs for each of the seven different giant icebergs (Table 2, shaded in grey and Figure 4), to assess the impact of the chosen test data set and different iceberg sizes. Here, we find that B34 gives the best results. The ~~dataset~~ number of images for this iceberg is ~~the~~ smallest (15 images), meaning that there are more images left for training and the background is usually not too challenging. B41 gives the lowest F_1 score. This dataset is the largest one, containing 46 images, and hence leaves the least number of images for training. Furthermore, B41 ~~stays very~~remains in close proximity to its calving position for a ~~while~~significant amount of time, which means that the first 13 images contain a significant amount of coast – often directly next to the iceberg (see Figure 4 first three images or supplementary animation for all images). In these cases all techniques pick the coast rather than the iceberg (as discussed later). The highest MAD and miss rate occur for iceberg B31. Because the images of B30 – our largest ~~berg~~ iceberg – are resized, this means that B31 appears largest in the images. Therefore, we believe that the large size of the iceberg, which U-net has not seen in the training data, causes U-net to miss parts of the iceberg (Figure 4 and Figure 5b, f). This is supported by the fact, that U-net misses large parts of B31 in the beginning (first few images in Figure 4), then misses smaller parts and once the iceberg has decreased to a size similar to other icebergs, U-net ~~works fine~~is suitable (last four images of B31 in Figure 4). In general, we find quite variable performance depending on which iceberg is retained as test data. This is because the same challenges (e.g. iceberg size, shape, surrounding) occur in subsequent images of the same iceberg, even when they are one month apart (best seen in the supplementary

animations). It is also the reason why we decided to evaluate the methods using cross-validation, as this makes the analysis less sensitive to the choice of a single iceberg as test data.

Also for Otsu and k-means the performance varies a lot depending on which iceberg is chosen as test data. The F_1 scores for Otsu range from 0.20 – 0.91, being lowest for C34 and highest for B31. Similarly, k-means also reaches the lowest F_1 score of 0.23 for C34 and the highest for B31 of 0.93. Compared to that, U-net is more consistent reaching F_1 scores between 0.68 – 0.97, but still exhibits significant variability. The fact that Otsu and k-means score so well for B31, also indicates that this data set is not hard per se. We rather suspect that we are challenging U-net too much when the iceberg in the test data is bigger than any iceberg in the training data. Neural networks are known to struggle with a domain-shift, where the test data are~~is~~ from a shifted version of the training data distribution and even more with out-of-domain samples from outside the training data distribution (Gawlikowski et al., 2021). Both are caused by insufficient training data, not or barely covering these examples. Therefore, we recommend expanding the training data, before applying U-net operationally or to icebergs larger than covered by the current training data set. In contrast, iceberg B41, where U-net reaches the lowest F_1 score, poses an even greater problem to the other algorithms, meaning that this dataset is actually challenging. Finally, we observe that U-net achieves the lowest false alarm rate on each iceberg. Otsu generates most false alarms (highest rate for six out of seven icebergs), but also achieves the lowest miss rate for four out of seven icebergs. Except for B31, U-net consistently achieves the highest F_1 score. In terms of MAD in area, k-means and U-net score best on three out of the seven icebergs each.

Table 2: Performance of the three methods for each test data set (iceberg). The number of images per iceberg and their minimum and maximum size is also given. Note that most images of B30 are rescaled, so it appears smaller in the images. Arrows indicate whether high (up) or low (down) numbers are desirable. The best score per iceberg and metric are highlighted in bold blue.

		F_1 score ↑	Misses [%] ↓	False Alarms [%] ↓	MAD in area [%] ↓
B30	U-net	**0.90**	15	**0.3**	3.3
29 images	Otsu	0.77	**9**	3.2	2.7
463-1052 km^2	k-means	0.79	12	2.4	**2.4**
B31	U-net	0.79	34	**0.2**	13.6
32 images	Otsu	0.91	**5**	1.6	3.0
79-518 km^2	k-means	**0.93**	6	1.0	**1.9**
B34	U-net	**0.97**	2	**0.2**	2.1
15 images	Otsu	0.83	**1**	1.7	**1.2**
97-241 km^2	k-means	0.80	8	1.6	8.3
B35	U-net	**0.94**	**2**	**0.3**	6.9
21 images	Otsu	0.66	9	2.3	7.4
62-158 km^2	k-means	0.63	10	2.5	**4.0**
B41	U-net	**0.68**	33	**0.7**	**3.5**

46 images	Otsu	0.27	13	10.5	3.8
54-116 km^2	k-means	0.29	**11**	10.1	5.6
B42	U-net	**0.88**	13	**0.6**	**5.4**
24 images	Otsu	0.84	**6**	1.7	8.9
142-235 km^2	k-means	0.76	28	1.0	18.7
C34	U-net	**0.81**	**20**	**0.4**	**3.7**
24 images	Otsu	0.20	36	10.1	4.3
61-101 km^2	k-means	0.23	32	9.1	5.2

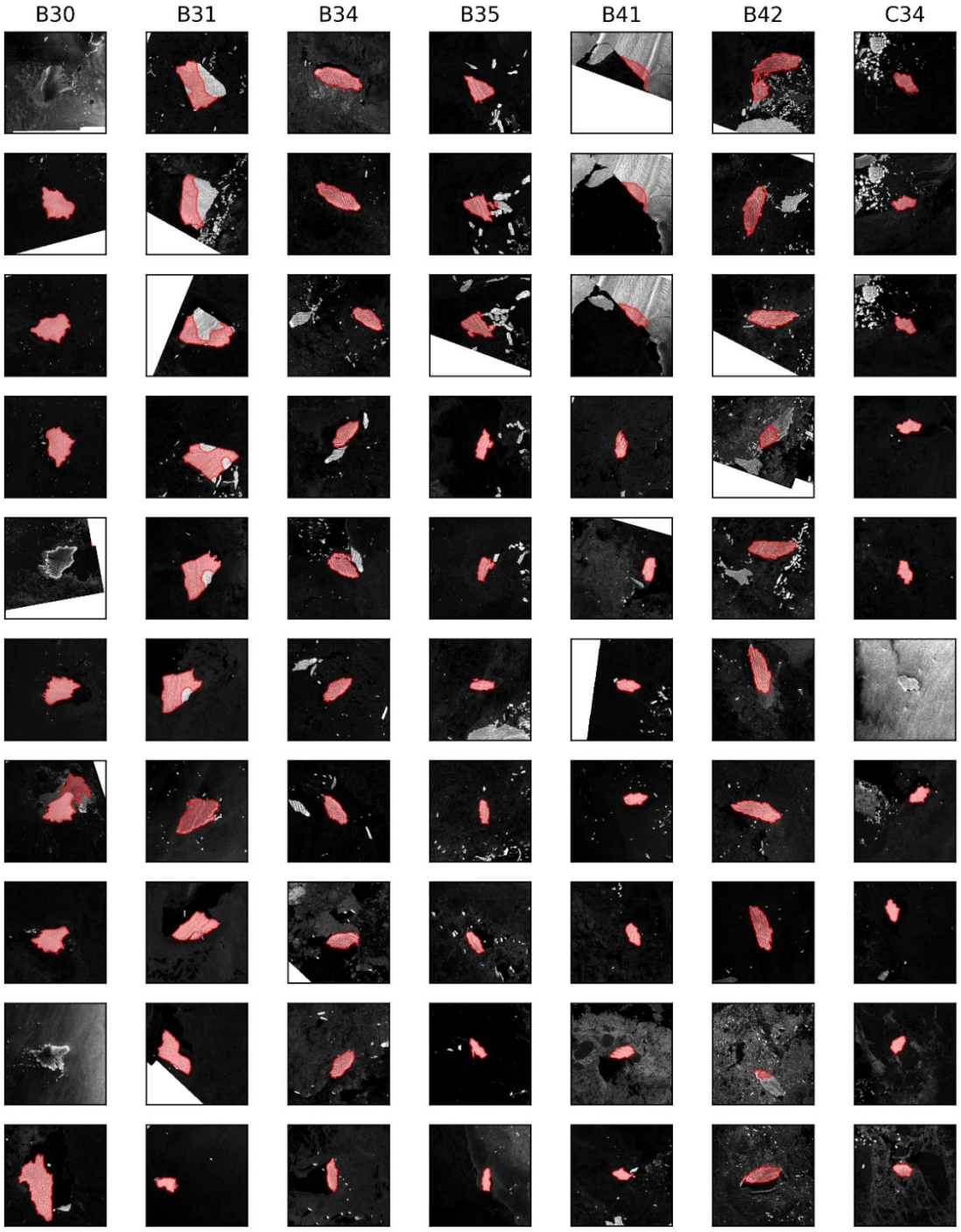
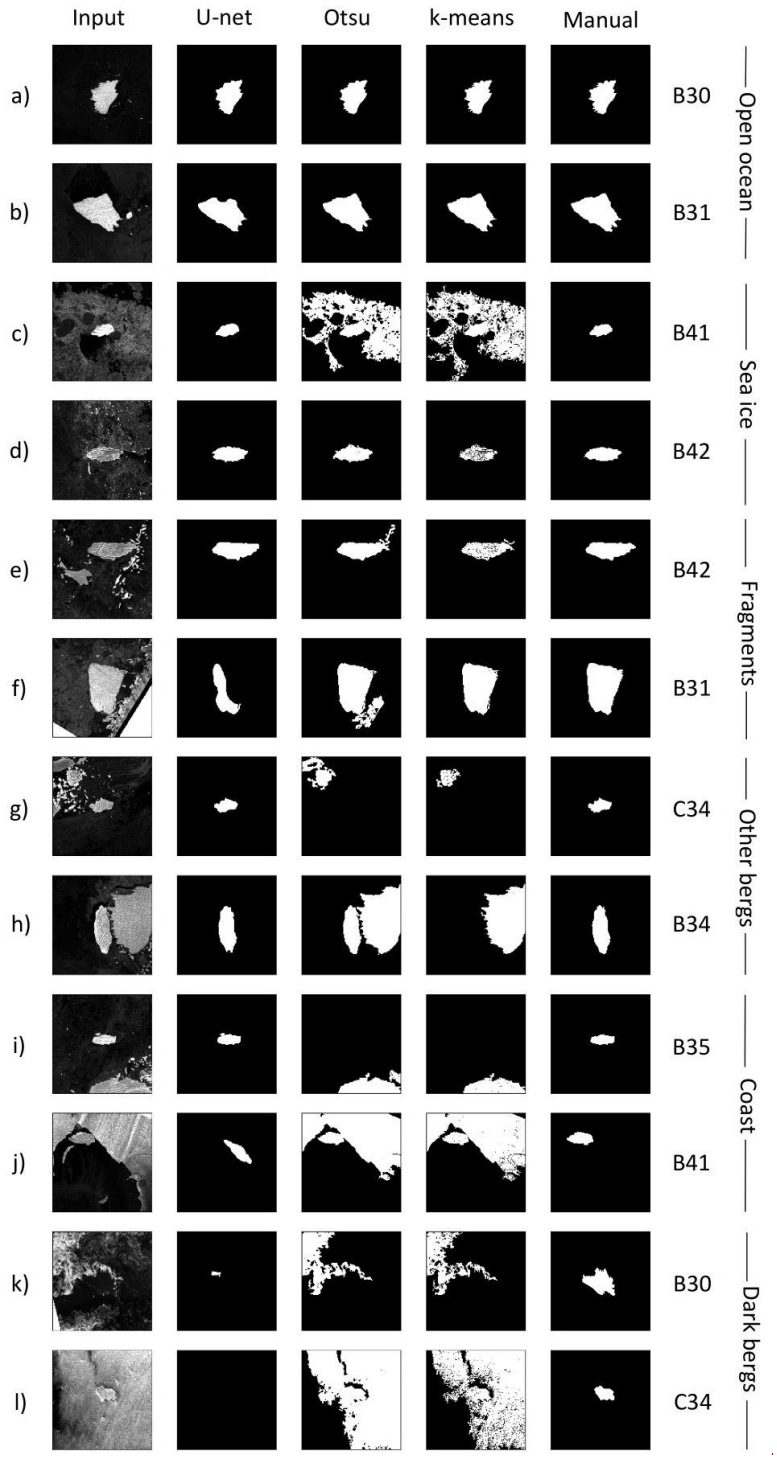
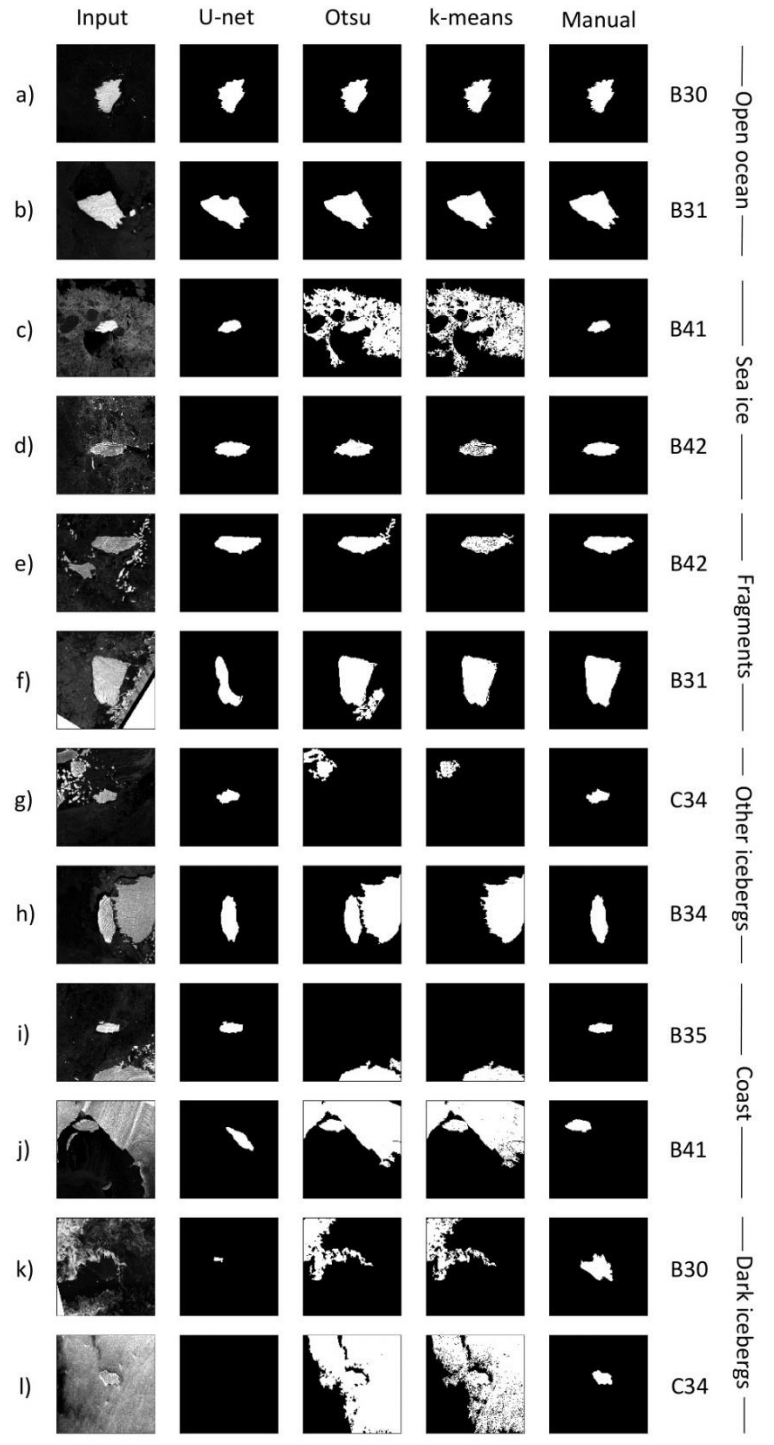


Figure 4: U-net derived iceberg outlines (red) plotted on top of the input images for 10 images per iceberg (columns). We always include the first and last image from each time series and sample the others equally in between. As the number of images per iceberg ranges from 15-46, this means that images of B34 are 1-2 months apart, while the images for B41 are 5 months apart in this figure. The full time series and results of all methods can be viewed in the supplementary animations (one per iceberg).


Input
U-net
Otsu
k-means
Manual
a)
b)
c)
d)
e)
f)
g)
h)
i)
j)
k)
l)
B30
B31
B41
B42
B42
B31
C34
B34
B35
B41
B30
C34
Open ocean
Sea ice
Fragments
Other bergs
Coast
Dark bergs



425 **Figure 5: Examples of input images (first column) and segmentation maps generated by U-net (second column), Otsu (third column), k-means (fourth column), and from manual delineations (last column). We picked these images for illustration to cover each category of environmental conditions twice and to include all icebergs (labelled on the right).**

43.3. Impact of different environmental conditions

Grouping the images according to the surrounding environmental conditions (see Section 2.2.) allows us to judge how well each method can deal with the respective challenge (Figure 5, Table 3). Open ocean makes up most of the images (46 %) and all methods perform very well with F_1 scores of 0.93-0.95 and MAD in area of 2.4-3.2 %. The Otsu threshold performs best, but the differences between the methods are very small. The two sample images (Figure 5a, b) also illustrate that the only problem in this category is rather that U-net generally tends to miss parts of B31 than open ocean in itself posing a problem. Sea ice occurs in 14 % of our images and overall U-net achieves the best F_1 score (0.88 compared to 0.72 and 0.74), but the Otsu threshold gives a slightly better MAD in area (4.3 % rather than 4.8 % and 5.4 %). Visually, the U-net predictions seem to be the most robust, as sea ice is discarded reliably. In contrast, the two other methods sometimes connect patches of sea ice to the iceberg (Figure 5c), but also work fine in other cases (Figure 5d).

Table 3: Performance of the three methods in different environmental conditions. The first column also indicates how often these conditions occur in our data set. Arrows indicate whether high (up) or low (down) numbers are desirable. The best values per category and metric are highlighted in bold blue.

		F_1 score↑	Misses [%] ↓	False Alarms [%] ↓	MAD in area [%] ↓
Open ocean (46 %)	U-net	0.93	11	**0.1**	2.8
	Otsu	**0.95**	**2**	0.4	**2.4**
	k-means	**0.95**	4	0.3	3.2
Sea ice (14 %)	U-net	**0.88**	14	**0.3**	4.8
	Otsu	0.72	**3**	2.4	**4.3**
	k-means	0.74	11	1.7	5.4
Fragments (24 %)	U-net	0.85	21	**0.4**	6.9
	Otsu	**0.94**	**2**	0.7	5.9
	k-means	**0.94**	7	**0.4**	**5.7**
Other icebergs (3 %)	U-net	**0.96**	**6**	**0.0**	**5.9**
	Otsu	0.18	66	7.7	110
	k-means	0.10	86	5.7	11
Coast (8 %)	U-net	**0.34**	68	**1.8**	**18**
	Otsu	0.12	**38**	29.5	1200
	k-means	0.11	44	28.6	1200
Dark icebergs (5 %)	U-net	**0.12**	92	**1.1**	**96**
	Otsu	**0.12**	**54**	34.3	450
	k-means	0.11	62	30.5	460

Iceberg fragments drifting in ~~the~~ direct proximity ~~of~~ to the target iceberg were found in 24 % of our images. Overall, k-means scores best in this category with a MAD of 5.7 % compared to 5.9 % and 6.9 %. In terms of F_1 score, Otsu and k-means both reach 0.94, whereas U-net only reaches 0.85. Visually, there are a few instances where Otsu connects more fragments to the iceberg than k-means and U-net (Figure 5e, f). This might be due to the Gaussian smoothing that we apply before the thresholding. We do not apply this step before k-means, and find that k-means tends to rather oversegment images, leaving small holes in the inside (Figure 5d, e). In the case of fragments, however, this turns out to be beneficial, as it allows k-means to reliably separate fragments from icebergs, even when they are very close by. The problem for U-net does not seem to be the actual fragments ~~itself~~themselves, as it rarely connects any fragments to the iceberg (Figure 5e, f). However, the images containing fragments are mostly from the large B31 and B42 icebergs, where U-net struggles due to their large extent. This can also be seen from the fact that U-net and k-means both generate only 0.4 % false alarms (fragments erroneously connected to the iceberg), but U-net has a much higher miss rate.

In 3 % of all images, another similar sized or bigger iceberg is (partially) visible. U-net scores best in all categories with a large margin, yielding an F_1 score of 0.96 compared to 0.12 and 0.11 and MAD in area of 5.9 % compared to 11 % and 110 %. Also visually, it becomes clear that U-net reliably picks the target iceberg and discards any other ice, while Otsu and k-means often pick the wrong iceberg or connect both with each other (Figure 5g, h). Considering iceberg shape and size in a tracking scenario could help mitigate this phenomenon, though (Barbat et al., 2021; Collares et al., 2018; Koo et al., 2021).

Coast is present in 8 % of all images and U-net outperforms the other techniques, but also struggles in ~~some~~ certain cases. The F_1 score is 0.34 for U-net and 0.12 and 0.11 for Otsu and k-means respectively. While U-net achieves a MAD of 18 %, the other methods yield over 1000 % each. Figure 5j illustrates what is happening in these cases: If too much coast is present, all algorithms pick the coast rather than the iceberg (and this is much larger than the iceberg, hence 1000 % deviation). However, U-net discards smaller parts of the coast around the image edges (Figure 5i). This is on the one hand because of the sliding convolution window and on the other hand, because U-net learns that the iceberg is usually in the centre (as we crop the images around the estimated position from operational iceberg tracking databases). Hence, U-net is able to correctly pick out the iceberg if not too much coast is present. For the same reason, it is easier for U-net to discard other icebergs at the image edges. Interestingly, even when a lot of coast is present, U-net does not pick the full coast, but predicts either nothing or a small – almost iceberg shaped – part of the coast (Figure 5j). This could indicate that U-net even learns that only ice that is fully surrounded by water is an iceberg. A possible strategy to avoid misclassifications due to large amounts of coast would be the inclusion of a land mask (Barbat et al., 2019; Collares et al., 2018; Frost et al., 2016; Mazur et al., 2017; Silva and Bigg, 2005). However, ice shelves and glaciers advance and retreat regularly and especially the calving of icebergs themselves significantly alters the land mask. Thus, just after calving, the iceberg would be within the former land mask and could not be picked up. A potential solution could be to always use the latest frontal positions from (Baumhoer et al., (2019) as a dynamic land mask.

The last category of dark icebergs is the hardest and makes up 5 % of the overall data set. In these cases, all methods fail with F_1 scores of 0.11-0.12 and the lowest MAD in area of 96 %. Again, it is interesting that U-net predicts either very small patches

or nothing at all in these cases (Figure 5k, l), while the other two methods segment large areas of brighter looking ocean. Potentially, U-net could learn to segment dark icebergs with a lot more training examples, but we only had ten such images in our overall data set. Finally, we would like to stress that the occurrence of these different environmental conditions will vary and our data set is not necessarily representative of all icebergs. We also find that the influence of iceberg size and environmental conditions cannot always be disentangled, as subsequent images of the same iceberg are often similar and the different environmental conditions are not spread equally across the different test data sets (individual icebergs). 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 categories (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."

43.4. Comparison to previous studies

Previous studies state different accuracy measures and due to the slightly different goal to detect all icebergs in a scene rather than finding one giant iceberg and accurately predicting its outline and area, they are not- directly comparable~~straightforward to compare~~. Two studies employ the k-means algorithm (Collares et al., 2018) or a variation of it (Koo et al., 2021), so we have indirectly compared U-net to them. None of them report any of our accuracy measures, though. Many of the previous approaches rely on some form of thresholding (Frost et al., 2016; Gill, 2001; Mazur et al., 2017; Power et al., 2001; Wesche and Dierking, 2012; Willis et al., 1996). We somehow covered these methods by comparing U-net to the Otsu threshold, but the exact approaches vary and none of them have applied the Otsu threshold. Two of the threshold-based methods report estimates for their area deviations. Wesche and Dierking (2012) state that iceberg area was overestimated by 10 ± 21 % with their approach. In a following study, they find that for the correctly detected icebergs 13.3 % of the total area was missing (Wesche and Dierking, 2015), meaning a bias in the opposite direction. Mazur et al. (2017) find positive and negative area deviations of ± 25 % on average. For edge-detection based algorithms, Williams et al. (1999) find an overestimation of iceberg area by 20 % and Silva and Bigg (2005)'s approach yields an underestimation of iceberg area by 10-13 %. These are biases again and both approaches are limited to winter images. For U-net, we find a bias of - 5.0 ± 29.1 %, which is lower than previous studies, but comes with a relatively high standard deviation due to some complete failures where the iceberg is not found at all. Previous studies only compare iceberg areas where icebergs were detected successfully. Barbat et al. (2019) report the lowest false positive (2.3 %) and false negative (3.3 %) rates, and the highest overall accuracy (97.5 %) of all previous studies. While their false negative rate is lower than our false negative rate (21 %), U-net achieves a lower false positive rate of 0.4 % and higher overall accuracy of 99 %. In a second study, Barbat et al. (2021) also analyse the area deviation of the detected icebergs and find average area deviations of 10 ± 4 %, which is also the best score reported so far. They only consider correctly detected icebergs in this metric, though. We find a MAE of 15 ± 26 % for U-net, which is slightly higher, but contains images where the iceberg was not found at all. These cases are not included in Barbat et al. (2021)'s estimates. Our MAD, which is less sensitive to such outliers, is 4.1 %, with 25 % and 75 % quantiles of 2.1 % and 12.1 %. These metrics compare

favourably to all previous studies. We also demonstrate in our study, that the performance varies depending on the chosen test data set and therefore, all measures and comparisons can only give an indication of the real performance. Judging from the data we have and comparing our results on this to previous studies as good as possible, U-net proves to be a very promising approach.

Qualitatively, previous studies have found degraded accuracies in challenging environmental conditions or excluded these from their datasets. Some studies report false detections due to sea ice (Koo et al., 2021; Mazur et al., 2017; Wesche and Dierking, 2012) or only applied their algorithm to sea-ice free conditions (Willis et al., 1996). Moreover, several previous studies have also encountered problems with clusters of several icebergs and iceberg fragments too close to each other (Barbat et al., 2019a; Frost et al., 2016; Koo et al., 2021; Williams et al., 1999). Also U-net shows slightly degraded performance in these situations (4.8 and 6.9 % MAD in area compared to 2.8 % in open ocean and F_1 scores of 0.88 and 0.85 compared to 0.93), but still achieves satisfying results in most of these cases. The challenge of other big icebergs does not occur in previous studies, since they were looking for all icebergs anyway. In terms of coast, many previous studies have employed a land mask (e.g. Barbat et al., 2019; Collares et al., 2018; Frost et al., 2016; Mazur et al., 2017; Silva and Bigg, 2005), but might miss newly calved icebergs due to that. Finally, the problem of dark icebergs has been described in several papers (Mazur et al., 2017; Wesche and Dierking, 2012; Williams et al., 1999), but was rarely mentioned in the evaluation. This is likely because most previous studies use visual inspection to identify misses and false alarms (e.g. Barbat et al., 2019; Frost et al., 2016; Mazur et al., 2017; Wesche and Dierking, 2012; Williams et al., 1999). However, dark icebergs are hard to spot in SAR images even for~~-~~ manual operators~~humans~~, so they might be missed by the visual inspection, too, unless in our case we know that there must be an iceberg of a certain size and shape that we are looking for. Others limit their method to winter images, when dark icebergs do not occur (Silva and Bigg, 2005; Williams et al., 1999; Young et al., 1998).

5~~4~~ Conclusions

We have developed a novel algorithm to automatically segment giant Antarctic icebergs in Sentinel-1 images~~ automatically~~. It is the first study to apply a deep neural network for iceberg segmentation. Furthermore, it is also the first study specifically targeting giant icebergs. Comparing U-net to two state-of-the-art segmentation techniques (Otsu thresholding and k-means), we find that U-net outperforms them in most metrics. Across all 191 images, U-net achieves an F_1 score of 0.84 and a median absolute area deviation of 4.1 %. Only the miss rate of Otsu and k-means is lower than for U-net, as we find that U-net overlooks parts of the iceberg appearing largest in the images, as in this case all training samples show smaller icebergs. We believe that this issue could be resolved with a larger training data set. U-net can reliably handle a variety of challenging environmental conditions including sea ice, nearby iceberg fragments, other icebergs and small patches of nearby coast. It fails when too much coast is visible and when icebergs appear dark, though. In these cases, all existing algorithms fail, but such obvious errors could easily be picked out in a tracking scenario. Also compared to previous studies, we regard our results as promising. For an operational application, i~~o~~n the short-term further post-processing could be implemented to filter outliers,

but on the long run, we would suggest ~~to~~ enlarg~~inge~~ the training data set before applying it to icebergs that are smaller or larger than those currently covered by the training data.

Code availability

The code is available from the authors upon reasonable request.

Data availability

Segmentation maps for all 191 images and from all three methods are shown in the supplementary animations (one animation per iceberg). DOI: 10.5281/zenodo.7875599 (Braakmann-Folgmann, 2023). The Sentinel-1 images are freely available from https://scihub.copernicus.eu/dhus/.

Author contributions

ABF, AS and DH designed the study. ER clicked most of the iceberg outlines, which are used as training data, during her internship, supervised by ABF. ABF also generated some of the outlines. ABF designed and implemented the U-net architecture, implemented the comparison methods, plotted the figures and wrote the manuscript. AS and DH supervised the work and suggested edits to the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgement

This work was supported by Barry Slavin and by NERC through National Capability funding, undertaken by a partnership between the Centre for Polar Observation Modelling and the British Antarctic Survey. The Antarctic Mapping Toolbox (Greene et al., 2017) was used. Thank you very much to Andreas Stokholm and Connor Shiggins for taking the time to carefully review our manuscript and their useful comments and suggestions, which helped to improve this paper. We would also like to thank the European Space Agency's ϕ-lab team for hosting Anne Braakmann-Folgmann during a three-month research visit and for several useful discussions and inspiration during this time and beyond. Thank you especially to Andreas Stokholm and Michael Marszalek.

References

Andersson, T. R., Hosking, J. S., Pérez-Ortiz, M., Paige, B., Elliott, A., Russell, C., Law, S., Jones, D. C., Wilkinson, J., Phillips, T., Byrne, J., Tietsche, S., Sarojini, B. B., Blanchard-Wrigglesworth, E., Aksenov, Y., Downie, R. and Shuckburgh, E.: Seasonal Arctic sea ice forecasting with probabilistic deep learning, Nat. Commun., 12(1), 1–12, https://doi.org/10.1038/s41467-021-25257-4, 2021.

Barbat, M. M., Wesche, C., Werhli, A. V. and Mata, M. M.: An adaptive machine learning approach to improve automatic iceberg detection from SAR images, ISPRS J. Photogramm. Remote Sens., 156(August), 247–259, https://doi.org/10.1016/j.isprsjprs.2019.08.015, 2019a.

Barbat, M. M., Rackow, T., Hellmer, H. H., Wesche, C. and Mata, M. M.: Three Years of Near-Coastal Antarctic Iceberg Distribution From a Machine Learning Approach Applied to SAR Imagery, J. Geophys. Res. Ocean., 124(9), 6658–6672, https://doi.org/10.1029/2019JC015205, 2019b.

Barbat, M. M., Rackow, T., Wesche, C., Hellmer, H. H. and Mata, M. M.: Automated iceberg tracking with a machine learning approach applied to SAR imagery : A Weddell sea case study, ISPRS J. Photogramm. Remote Sens., 172(December 2020), 189–206, https://doi.org/10.1016/j.isprsjprs.2020.12.006, 2021.

Baumhoer, C. A., Dietz, A. J., Kneisel, C. and Kuenzer, C.: Automated extraction of antarctic glacier and ice shelf fronts from Sentinel-1 imagery using deep learning, Remote Sens., 11(21), 1–22, https://doi.org/10.3390/rs11212529, 2019.

Bigg, G. R., Wadley, M. R., Stevens, D. P. and Johnson, J. A.: Modelling the dynamics and thermodynamics of icebergs, cold Reg. Sci. Technol., 26(2), 113–135, https://doi.org/10.1016/S0165-232X(97)00012-8, 1997.

Bouhier, N., Tournadre, J., Rémy, F. and Gourves-Cousin, R.: Melting and fragmentation laws from the evolution of two large Southern Ocean icebergs estimated from satellite data, Cryosphere, 12(7), 2267–2285, https://doi.org/10.5194/tc-12-2267-2018, 2018.

Braakmann-Folgmann, A.: Segmentation maps of giant Antarctic icebergs, , https://doi.org/10.5281/zenodo.7875599, 2023.

Braakmann-Folgmann, A., Shepherd, A. and Ridout, A.: Tracking changes in the area, thickness, and volume of the Thwaites tabular iceberg “B30” using satellite altimetry and imagery, Cryosphere, 15(8), 3861–3876, https://doi.org/10.5194/tc-15-3861-2021, 2021.

Braakmann-Folgmann, A., Shepherd, A., Gerrish, L., Izzard, J. and Ridout, A.: Observing the disintegration of the A68A iceberg from space, Remote Sens. Environ., 270, 112855, https://doi.org/10.1016/j.rse.2021.112855, 2022.

Bradski, G.: The OpenCV Library, Dr. Dobb's J. Softw. Tools, 2000.

Budge, J. S. and Long, D. G.: A Comprehensive Database for Antarctic Iceberg Tracking Using Scatterometer Data, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 11(2), 434–442, https://doi.org/10.1109/JSTARS.2017.2784186, 2018.

Chollet, F.: Xception: Deep learning with depthwise separable convolutions, Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, 1800–1807, https://doi.org/10.1109/CVPR.2017.195, 2017.

Chollet, F. and Others, &: Keras, https://github.com/fchollet/keras, 2015.

Collares, L. L., Mata, M. M., Kerr, R., Arigony-Neto, J. and Barbat, M. M.: Iceberg drift and ocean circulation in the northwestern Weddell Sea, Antarctica, Deep. Res. Part II Top. Stud. Oceanogr., 149(March), 10–24, https://doi.org/10.1016/j.dsr2.2018.02.014, 2018.

Dirscherl, M., Dietz, A. J., Kneisel, C. and Kuenzer, C.: A novel method for automated supraglacial lake mapping in antarctica using sentinel-1 sar imagery and deep learning, Remote Sens., 13(2), 1–27, https://doi.org/10.3390/rs13020197, 2021.

Drinkwater, M. R.: Satellite Microwave Radar Observations of Antarctic Sea Ice, Anal. SAR Data Polar Ocean., 145–187, https://doi.org/10.1007/978-3-642-60282-5_8, 1998.

Duprat, L. P. A. M., Bigg, G. R. and Wilton, D. J.: Enhanced Southern Ocean marine productivity due to fertilization by giant icebergs, Nat. Geosci., 9(3), 219–221, https://doi.org/10.1038/ngeo2633, 2016.

England, M. R., Wagner, T. J. W. and Eisenman, I.: Modeling the breakup of tabular icebergs, Sci. Adv., 6(51), 1–9, https://doi.org/10.1126/sciadv.abd1273, 2020.

Frost, A., Ressel, R. and Lehner, S.: Automated iceberg detection using high resolution X - band SAR images, Can. J. Remote Sens., 42(4), https://doi.org/10.1080/07038992.2016.1177451, 2016.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R. and Zhu, X. X.: A Survey of Uncertainty in Deep Neural Networks, , 1–41 http://arxiv.org/abs/2107.03342, 2021.

Gill, R. S.: Operational detection of sea ice edges and icebergs using SAR, Can. J. Remote Sens., 27(5), 411–432, https://doi.org/10.1080/07038992.2001.10854884, 2001.

Greene, C. A., Gwyther, D. E. and Blankenship, D. D.: Antarctic Mapping Tools for MATLAB, Comput. Geosci., 104, 151–157, https://doi.org/10.1016/j.cageo.2016.08.003, 2017.

He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2016-Decem, 770–778, https://doi.org/10.1109/CVPR.2016.90, 2016.

Helly, J. J., Kaufmann, R. S., Stephenson, G. R. and Vernet, M.: Cooling, dilution and mixing of ocean water by free-drifting icebergs in the Weddell Sea, Deep. Res. Part II Top. Stud. Oceanogr., 58(11–12), 1346–1363, https://doi.org/10.1016/j.dsr2.2010.11.010, 2011.

Jansen, D., Schodlok, M. and Rack, W.: Basal melting of A-38B: A physical model constrained by satellite observations, Remote Sens. Environ., 111(2), 195–203, https://doi.org/10.1016/j.rse.2007.03.022, 2007.

Jenkins, A.: The impact of melting ice on ocean waters, J. Phys. Oceanogr., 29(9), 2370–2381, https://doi.org/10.1175/1520-0485(1999)029<2370:TIOMIO>2.0.CO;2, 1999.

Kingma, D. P. and Ba, J. L.: Adam: A method for stochastic optimization, 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 1–15, 2015.

Koo, Y., Xie, H., Ackley, S. F., Mestas-Nunez, A. M., Macdonald, G. J. and Hyun, C.-U.: Semi-automated tracking of iceberg B43 using Sentinel-1 SAR images via Google Earth Engine, Cryosph., 15(May), 4727–4744, https://doi.org/10.5194/tc-15-4727-2021, 2021.

Kucik, A. and Stokholm, A.: AI4SeaIce: selecting loss functions for automated SAR sea ice concentration charting, Sci. Rep., 13(1), 1–10, https://doi.org/10.1038/s41598-023-32467-x, 2023.

LeCun, Y., Bengio, Y. and Hinton, G.: Deep learning, Nature, 521(7553), 436–444, https://doi.org/10.1038/nature14539, 2015.

Macqueen, J.: SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS, in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 233, pp. 281–297, California: University of California Press., , 1967.

Mazur, A. K., Wåhlin, A. K. and Krężel, A.: An object-based SAR image iceberg detection algorithm applied to the Amundsen Sea, Remote Sens. Environ., 189, 67–83, https://doi.org/10.1016/j.rse.2016.11.013, 2017.

Merino, N., Le Sommer, J., Durand, G., Jourdain, N. C., Madec, G., Mathiot, P. and Tournadre, J.: Antarctic icebergs melt over the Southern Ocean : Climatology and impact on sea ice, Ocean Model., 104, 99–110, https://doi.org/10.1016/j.ocemod.2016.05.001, 2016.

Mohajerani, Y., Wood, M., Velicogna, I. and Rignot, E.: Detection of glacier calving margins with convolutional neural networks: A case study, Remote Sens., 11(1), 1–13, https://doi.org/10.3390/rs11010074, 2019.

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, Sci. Rep., 11(1), 1–10, https://doi.org/10.1038/s41598-021-84309-3, 2021.

Otsu, N.: A Threshold Selection Method from Gray-Level Histograms, IEEE Trans. Syst. Man. Cybern., C(1), 62–66, 1979.

Poliyapram, V., Imamoglu, N. and Nakamura, R.: DEEP LEARNING MODEL FOR WATER / ICE / LAND CLASSIFICATION USING LARGE-SCALE MEDIUM RESOLUTION SATELLITE IMAGES, IGARSS 2019 - 2019 IEEE Int. Geosci. Remote Sens. Symp., (d), 3884–3887, 2019.

Power, D., Youden, J., Lane, K., Randell, C. and Flett, D.: Iceberg detection capabilities of radarsat synthetic aperture radar, Can. J. Remote Sens., 27(5), 476–486, https://doi.org/10.1080/07038992.2001.10854888, 2001.

Ronneberger, O., Fischer, P. and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 9351, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Sandven, S., Babiker, M. and Kloster, K.: Iceberg observations in the barents sea by radar and optical satellite images, in Proceedings of the Envisat Symposium, , 2007.

Schmidhuber, J.: Deep Learning in neural networks: An overview, Neural Networks, 61, 85–117, https://doi.org/10.1016/j.neunet.2014.09.003, 2015.

Sephton, A. J., Brown, L. M., Macklin, J. T., Partington, K. C., Veck, N. J. and Rees, W. G.: Segmentation of synthetic-aperture radar imagery of sea ice, Int. J. Remote Sens., 15(4), 803–825, https://doi.org/10.1080/01431169408954118, 1994.

Silva, T. A. M. and Bigg, G. R.: Computer-based identification and tracking of Antarctic icebergs in SAR Computer-based identification and tracking of Antarctic icebergs in SAR images, , (May 2018), https://doi.org/10.1016/j.rse.2004.10.002, 2005.

Silva, T. A. M., Bigg, G. R. and Nicholls, K. W.: Contribution of giant icebergs to the Southern Ocean freshwater flux, J. Geophys. Res., 111(July 2005), 1–8, https://doi.org/10.1029/2004JC002843, 2006.

Singh, A., Kalke, H., Loewen, M. and Ray, N.: River Ice Segmentation with Deep Learning, IEEE Trans. Geosci. Remote Sens., 58(11), https://doi.org/10.1109/TGRS.2020.2981082, 2020.

Smith, K. L., Robison, B. H., Helly, J. J., Kaufmann, R. S., Ruhl, H. A., Shaw, T. J., Twining, B. S. and Vernet, M.: Free-drifting icebergs: Hot spots of chemical and biological enrichment in the Weddell Sea, Science (80-.)., 317(5837), 478–482, https://doi.org/10.1126/science.1142834, 2007.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res., 15, 1929–1958, 2014.

Stokholm, A., Wulf, T., Kucik, A., Saldo, R., Buus-Hinkler, J. and Hvidegaard, S. M.: AI4SeaIce: Toward Solving Ambiguous SAR Textures in Convolutional Neural Networks for Automatic Sea Ice Concentration Charting, IEEE Trans. Geosci. Remote Sens., 60, https://doi.org/10.1109/TGRS.2022.3149323, 2022.

Surawy-Stepney, T., Hogg, A. E., Cornford, S. L. and Davison, B. J.: Episodic dynamic change linked to damage on the thwaites glacier ice tongue, Nat. Geosci., 16(1), 37–43, https://doi.org/10.1038/s41561-022-01097-9, 2023.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Navas, I., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., Abbate, M. L., Croci, R., Pietropaolo, A., Huchler, M. and Rostan, F.: GMES Sentinel-1 mission, Remote Sens. Environ., 120, 9–24, https://doi.org/10.1016/j.rse.2011.05.028, 2012.

Tournadre, J., Bouhier, N., Girard-Ardhuin, F. and Rémy, F.: Antarctic icebergs distributions 1992–2014, J. Geophys. Res. Ocean., 121(1), 327–349, https://doi.org/10.1002/2015JC011178, 2016.

Ulaby, F. T. and Long., D. G.: Microwave radar and radiometric remote sensing, The University of Michigan Press., 2014.

Vernet, M., Smith, K. L., Cefarelli, A. O., Helly, J. J., Kaufmann, R. S., Lin, H., Long, D. G., Murray, A. E., Robison, B. H., Ruhl, H. A., Shaw, T. J., Sherman, A. D., Sprintall, J., Stephenson, G. R., Stuart, K. M. and Twining, B. S.: Islands of ice: Influence of free-drifting Antarctic icebergs on pelagic marine ecosystems, Oceanography, 25(3), 38–39, https://doi.org/10.5670/oceanog.2012.72, 2012.

Wesche, C. and Dierking, W.: Iceberg signatures and detection in SAR images in two test regions of the Weddell Sea, Antarctica, J. Glaciol., 58(208), 325–339, https://doi.org/10.3189/2012J0G11J020, 2012.

Wesche, C. and Dierking, W.: Near-coastal circum-Antarctic iceberg size distributions determined from Synthetic Aperture Radar images, Remote Sens. Environ., 156, 561–569, https://doi.org/10.1016/j.rse.2014.10.025, 2015.

Williams, R. N., Rees, W. G. and Young, N. W.: A technique for the identification and analysis of icebergs in synthetic aperture radar images of Antarctica, Int. J. Remote Sens., 20(15–16), 3183–3199, https://doi.org/10.1080/014311699211697, 1999.

Willis, C. J., Macklin, J. T., Partington, K. C., Teleki, K. A., Rees, W. G. and Williams, G.: Iceberg detection using ers-1 synthetic aperture radar, Int. J. Remote Sens., 17(9), 1777–1795, https://doi.org/10.1080/01431169608948739, 1996.

Young, N. W. and Hyland, G.: Applications of time series of microwave backscatter over the Antarctic region, in Proceedings of the third ERS Scientic Symposium, 17-21 March 1997, Florence, Italy, pp. 1007–1014, Frascati, Italy: European Space Agency, SP-414, , 1997.

Young, N. W., Turner, D., Hyland, G. and Williams, R. N.: Near-coastal iceberg distributions in East Antarctica, 50-145°E, Ann. Glaciol., 27, 68–74, https://doi.org/10.3189/1998aog27-1-68-74, 1998.

Zhang, E., Liu, L. and Huang, L.: Automatically delineating the calving front of Jakobshavn Isbræ from multitemporal TerraSAR-X images: A deep learning approach, Cryosphere, 13(6), 1729–1741, https://doi.org/10.5194/tc-13-1729-2019, 2019.

720