

FAU (Inf. 5) | Martenstr. 3 | D-91058 Erlangen

Computer Science Department 5

Pattern Recognition Lab

Marcel Dreier

Martensstrasse 3, D-91058 Erlangen
Room 09.156
Phone [+49 9131 85-28977](tel:+4991318528977)

Marcel.dreier@fau.de
<https://lme.tf.fau.de/person/madreier/>

Erlangen, April 30, 2025

Dear Editor and Reviewers,

We thank the reviewers and the editor for their helpful and thoughtful comments. Their time and effort spent reviewing our manuscript are greatly appreciated. Following their advice, we carefully adjusted the manuscript and added additional experiments to the appendix. In detail, we have:

- Expanded and clarified explanations that were previously difficult to grasp.
- Improved the coloring of Figs. 1 and 4 to 7.
- Added additional experiments to the appendix.
- Improved consistency in terminology and tone.
- Added additional citations to related work to give a broader context of previous work.
- Clarified claims about the novelty and impact of our work.
- Adjusted the technical explanations for a better reading experience.

With these changes, we believe we have significantly improved the quality of our manuscript, and we hope to meet the expectations of the reviewers.

We are looking forward to hearing from you.

Kind regards,
Marcel Dreier (on behalf of all co-authors)

The following pages contain a list of editor’s and reviewers’ comments followed by our replies. The comments are sequentially numbered and associated with the corresponding reviewer. The replies may contain references to changes in the original manuscript, which are identified by a label consisting of a running number and followed by the label of the original comment in parentheses. The label links back to the original reviewer’s comment within the manuscript. For instance, the reference

5 **C2 (1.3)**, which is typeset in the manuscript margin, refers to the second change stemming from the third comment of the first reviewer. In addition, please note that the reference section is not displayed correctly in this change document, for length and reference verification as well as correct table and figure numbering please refer to the “manuscript version”.

Comments of the 1st Reviewer:

1.1 *The choice to standardize all radargrams to a height of 1024 pixels requires further justification, especially given the reduction in resolution this causes, which could potentially affect the precision of the derived ice boundaries. The manuscript should provide a more detailed rationale (possibly linked to computational efficiency) for this choice, considering the capabilities of the U-Net-like to process varying input shapes.*

► Thank you for this comment. As the reviewer has correctly stated, the U-Net can technically handle any input resolution. However, having different input resolutions can lead to significant drops in accuracy. It also complicates and slows down processing as samples within a batch are usually assumed to be the same size. However, there are two common solutions to this problem. The first is patching, where the image is cut into patches of equal resolution. Since our networks need to view the full height to make a layer prediction, this method is not applicable to our case. The other option is resizing, where we resize the input to a standardized resolution. We chose this measure for our experiments. Resizing all radargrams to the maximum height would avoid inaccuracies due to downscaling but introduces its own issues: potential unnatural artifacts due to upscaling, a substantial increase in model depth needed to maintain an effective receptive field, and significantly higher computational costs. Considering all these reasons, we found a standardized height of 1024 pixels a good compromise. We have added a brief explanation of our reasoning in change 28.

1.2 *The decision to not use an exclusive flight for the AWI testing subset due to significant variability among the radargrams is questionable. The inherent variability could, in fact, provide a rigorous real-world test scenario, which is crucial for assessing the robustness and adaptability of the model to new and varied environments (which should be the eventual goal of any benchmark dataset and the models developed based on them). A reevaluation of the testing subset choice is recommended to potentially enhance the findings.*

► The reviewer raises an excellent point. In theory, we agree that having an exclusive test flight is preferable. However, all three AWI flights are from different campaigns with slightly different processing and different survey areas, introducing a domain shift. These changes, combined with the overall smaller size of the subset, make generalizations from one flight to another challenging without additional advanced techniques to mitigate the domain shift. Since these techniques would have been out of scope for this work and we still wanted to provide meaningful results for the AWI subset, we did not select an exclusive test flight for the AWI subset. Please note that generalization between different flights can still be tested with the other subsets.

1.3 *The omni model shows reduced performance in the FAU and AWI domains, which the authors attribute to domain shifts. Consideration of alternative approaches such as weighting samples by domain frequency or uniformly sampling training examples across domains could potentially mitigate this issue. An exploration of these methods would be valuable for enhancing model generalization.*

► This is an interesting point, which we investigated further with additional experiments in our appendix (see change 13). From the results, we can see a slight drop in performance for the CReSIS subset but substantial improvements for the AWI and FAU subsets compared to the original omni model. The model even outperforms its dataset-specific counterpart on the AWI subset. However, we still see a substantial performance gap in the FAU subset compared to the dataset-specific model. We reason that the FAU domain is naturally further away from the AWI and CReSIS domains, as it consists of undifferentiated radargrams. Thus, the uniform sampling strategy does help to mitigate the domain shift but does not fix it entirely when the processing of the radargrams differs significantly.

454 *The proposed U-Net uses two heads to separately predict the ice surface and bottom. Why is it better than a straightforward approach with one head simultaneously doing both? Softmax can be applied later in the column-wise manner to extract the boundaries as well, so it should not be a limitation.*

► Thank you for your comment. The idea of having a separate head for every task is a common approach for multi-task models in deep learning across various categories.

50 In our case, it allows every head to focus on a single layer instead of solving both simultaneously. As the reviewer correctly stated, the task can also be solved with a single head that projects onto multiple output channels depending on the specified prediction task. However, this also increases the size of the head proportionally to the output channels. In the case of the output head consisting of a single convolutional layer, the two approaches would be mathematically equivalent.

1.5 *The authors write in Section 5.1: "Depending on the chosen method, the metrics used to assess the quality of the predictions differ," which is not really true, as zone predictions are easily convertible to boundaries and vice versa, so there is no problem to providing the whole set of metrics.*

► Thank you for your comment. It is true that we can easily calculate the zones from the boundary prediction. However, a metric like the IoU does not give precise intel about the layer extraction task. Thus, we did not include it in our evaluation as we mainly focus on layer extraction.

60 However, going from zone predictions to layer predictions is a lot more difficult. In zone segmentation tasks, the model decides the zone class of each pixel separately. Hence, the resulting prediction does not need to have clearly identifiable layers. We give a small example in the figure below. In those cases, layer extraction and defining a corresponding error become ambiguous. Therefore, these tasks usually rely on segmentation metrics, as we cannot define layer metrics.

Air Layer

Ice Layer

Bedrock Layer

1.6 *The manuscript claims that confusion matrix-based metrics would perform poorly if predictions are, e.g., consistently off by a pixel. However, this statement is misleading as these metrics are typically used for zone predictions, not boundary delineations. A correction or further explanation is needed to resolve this confusion.*

► Thank you, we clarified this paragraph in the revised version of the paper (compare change 31).

1.7 *In Appendix A, it is stated that the authors used dropout layers inside the ResBlocks. Was it a regular dropout? If not, it should be specified. If yes, I would suggest also trying something like spatial dropout, as many practitioners found it more helpful in convolutional networks.*

► Thank you for the advice. We have utilized a regular Dropout Layer in our implementation. We have not experimented with spatial Dropout before, but it will be an interesting avenue for further optimizations. We added spatial dropout to future work and clarified the use of our dropout layer (compare addition 48 and change 11).

1.8 *Figures 6, 7, and similar graphics are challenging to interpret. I would suggest just plotting four curves—two groundtruths (surface and bottom) and two predictions on top (e.g. dashed).*

► Thank you for pointing this out. We adjusted the visualizations accordingly (compare Fig. 4, 5, 6, and 7).

Comments of the 2nd Reviewer:

2.1 *The manuscript uses varying terminology, such as “the air-ice layer and ice-ground layer” and “ice bottom and ice surface layer.” Maintaining consistency in terminology throughout the text would improve clarity and readability. ... Line 36: The phrase “radargram of the glacier” sounds somewhat awkward. Additionally, the manuscript does not always adhere consistently to standard glaciological terminology.*

► Thank you for this helpful suggestion. We unified our terminology (compare changes 3, 8, 4, 10, 19, and the update to the caption of Fig. 3).

2.2 *The manuscript claims to present the first benchmark dataset for ice boundary extraction, yet related datasets such as CReSIS data have been widely used. The authors should explicitly contrast IceAnatomy with existing datasets and justify why this dataset is uniquely valuable beyond just being a “benchmark.”*

► Thank you for raising this very important point. As correctly highlighted, IceAnatomy is not the first dataset in this field. However, it is the first benchmark dataset. Thus, it has to conform to higher standards, unlike previous datasets.

One such standard is reproducibility. While some works use the same dataset, they do not precisely define their training and testing subsets. Our evaluation on the CReSIS subset demonstrates that the composition of these splits can significantly influence model performance (see Table 1). For example, if the test set consisted only of radargrams from the Antarctic Peninsula, the model’s performance would appear substantially better. This emphasizes the necessity of standardized training and testing splits, such as those provided in a benchmark dataset, to ensure a fair and meaningful comparison between multiple models.

Another important factor is generalization. IceAnatomy provides data spanning multiple campaigns, radar systems, institutions, and diverse glaciological environments. Thus, researchers can evaluate their model in a multitude of scenarios to achieve good generalization.

Lastly, institutions like CReSIS may provide a large public database for radargrams with an annotated ice bottom layer, but their label quality is not guaranteed to have human-level accuracy. CReSIS themselves state that they utilize snake trackers, leading edge detectors, interpolation, and peak detectors based on the judgment of the operator picking the data, making the quality difficult to judge (compare https://data.cresis.ku.edu/data/rds/rds_readme.pdf). Although most labels are still correct, neural networks generally learn to imitate the annotation process, so models trained solely on automated labels may learn peak detection rather than more generalizable features. Thus, we need to have humanly annotated data. This might raise concerns about whether the CReSIS subset is an appropriate choice for a benchmark dataset. However, CReSIS has been the primary source of data in this field, and multiple works report human annotations for this specific dataset. Hence, we deem it appropriate to include it, even though some labels might be noisy.

We adjusted the manuscript to highlight the benefits of our new benchmark dataset (see addition 3).

2.3 *The radargram visualizations are useful but could benefit from additional annotations. Additionally, the color scheme makes it difficult to distinguish certain features, and the way different annotations are represented could be improved for better clarity*

► Thank you for pointing this out. We adjusted the visualizations to make them clearer and easier to grasp (compare the updated Fig. 4, 5, 6, and 7).

2.4 *The manuscript is highly technical and may be challenging for a glaciological audience. Since The Cryosphere primarily targets glaciologists, the extensive use of computer science jargon and technical terminology either requires more thorough explanations or suggests that a different journal may be a better fit. Ensuring the content is more accessible to the journal's primary readership should be a key consideration.*

115

...

Lines 240–245, 264–271, and other similar sections contain highly technical explanations. These should either be clarified and simplified for better accessibility to the journal's audience or, if the technical depth is essential, the choice of journal may need to be reconsidered.

▶ Thank you for raising this point. The editor initially asked us to add more explanations for the technical terms. As a result, some parts became a bit more technical than originally intended, as the explanations already require a certain level of expertise. We try to address this issue by simplifying our explanations and moving the more in-depth explanations in the Appendix. Please, see deletions 3 and 4, change 30 and the updated Appendix C and B.

2.5 *The scientific motivation of the study could be further elaborated. This is one of the aspects that might suggest the paper, in its current form, would be better suited for a more technical journal.*

125▶ Thank you for pointing that out. We have revised the manuscript to express our motivation and vision for the future more clearly (see change 1).

2.6 *The rationale for the baseline model choices (e.g., why U-Net with specific modifications) should be better justified. Why not test other architectures such as Transformers or hybrid CNN-RNN models?*

▶ The U-Net is a widely adopted approach for tasks such as ice boundary extraction and comparable tasks. While more recent architectures like the transformer or recurrent neural networks may offer better performance, they also come with increased computational costs, larger models, and other practical limitations. This might hinder the deployment in the field and other time-critical scenarios. For that reason, we chose the U-Net as our baseline model. We added our reasoning in change 29.

2.7 *The manuscript states that the dataset consists of manually labeled ice boundaries but does not provide sufficient details on the annotation process. What steps were taken to ensure label accuracy? Were multiple annotators involved? How was inter-annotator variability handled?*

135

▶ Thank you for your comment. We have clarified the picking process for the FAU subset in addition 5, for the AWI subset in addition 6, and added further details and references for the CReSIS picking procedure in change 25.

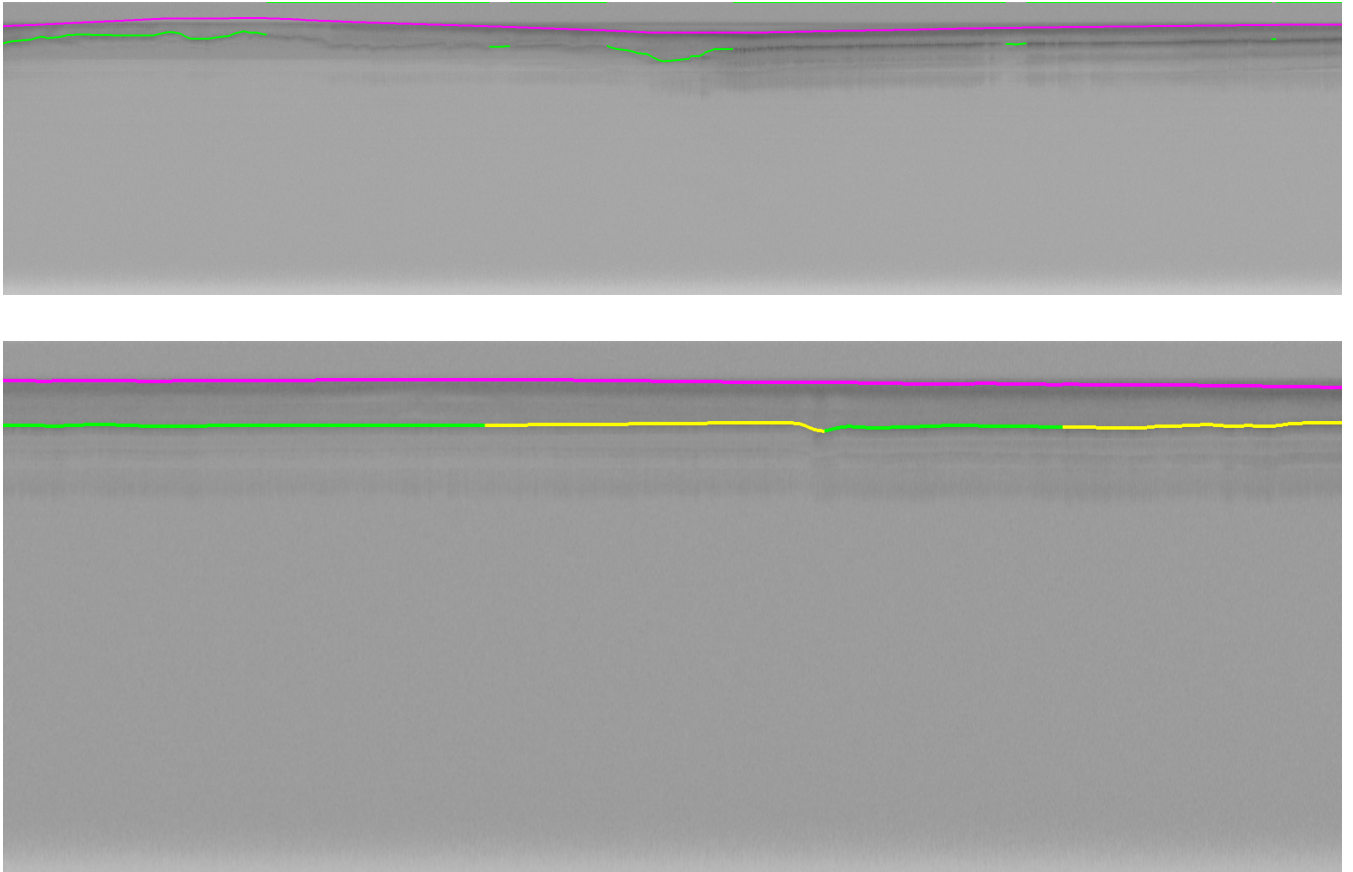
2.8 *The inclusion of noisy annotations from CReSIS data is acknowledged, but how does this affect training and evaluation? Have any data cleaning techniques been applied?*

140▶ Thank you for this comment. Several works have already pointed out that they consider the labels from the CReSIS subset noisy. Although CReSIS provided quality ratings for each label, we found several cases where the reasoning behind them was unclear to us. Below, we provide two example radargrams from the Abbott Ice Shelf. In the first radargram, several parts of the ice bottom are missing, although the original annotator has high confidence (green) in his labeled ice bottom. In the second example, the quality rating changes throughout the section between high confidence (green) and medium confidence (yellow) with no clear indication in the radargram. We included the CReSIS quality labels to stay consistent with the original. However, we do not think they necessarily reflect the mentioned noise. Thus, we could not evaluate how this noise affects the training or evaluation process.

145

2.9 *The dataset includes different radar systems and processing methods, which may introduce domain shifts. Are these shifts quantified? How do they impact model performance?*

150▶ Thank you for this question. As the reviewer has highlighted, the domain shift is a serious problem introduced by different radar systems, study sites, board electronics, and processing steps. As a result, it is nearly impossible to quantify the domain



shift in a precise mathematical manner. However, we have summarized the most critical processing steps in Section 3.2. From these steps, we can derive some qualitative characteristics, e.g., the AWI and CReSIS provide differentiated radargrams.

155 How does this impact the results? From our experiments with the Omni model, we can see that the data from different domains do not necessarily work well together. After conducting follow-up experiments in Appendix 13, we hypothesize that the issue might be connected to the differentiation of the radargram. However, we must also point out that the domain shift is very difficult to study as we cannot isolate this variable. We therefore deem it out of scope for this work and leave it to future studies to investigate this phenomenon in more detail.

2.10 *The AP-5% metric relaxes the error bounds, but why were 1% and 5% chosen? Would alternative thresholds (e.g., 2% or 10%)*
 160 *provide additional insights?*

- Thank you for this comment. While we agree that offering additional intervals could potentially help us analyze the error better, adding too many would also make the evaluation and interpretation of the error very confusing. Hence, we chose the AP-1% and AP-5% as they represent a near-perfect and a good pick.

2.11 *The "ice boundary collapse" issue observed in predictions is significant. Could this be mitigated with additional constraints in*
 165 *the loss function or post-processing techniques?*

- Thank you for your interest. We believe that modifying the loss function further is unlikely to significantly mitigate the boundary collapse, as our proposed distance-based loss already puts more emphasis on large outliers like the boundary collapse. However, additional post-processing steps could be an interesting approach for future work. The question here would be how

to prevent the algorithm from overcorrecting traces when the boundary collapse takes up the majority of the transit. As this would require more experiments and testing, it is outside the scope of this work. However, we included post-processing as a potential method for future work to explore (compare addition 10).

2.12 *The paper does not discuss the impact of hyperparameters in training. How sensitive is the model to learning rate, regularization, and architecture modifications?*

► Thank you for this comment. We believe that the focus of this paper lies in the benchmark dataset rather than the benchmark model. Accordingly, our ablation study focuses on factors connected to the data such as the study site, the domain shift, and the thermal regime. For that reason, we chose the hyperparameters like the learning rate close to standard values for this field. However, we added a small ablation study in addition 14 to investigate two of the hyperparameters more closely. From the results, we can see that different hyperparameter setups favor different subsets of IceAnatomy. However, there seems to be no universal optimal setup.

2.13 *Some terms, such as "depth resolution," "relative error," and "wave velocity assumptions," need clearer definitions in the main text rather than just appearing in equations.*
... Line 315: The phrase 'pass through a pixel' is unclear. At times, the radargram are treated as an image, and at other times as a matrix. However, it is important to note that a wave does not pass through a pixel. Line 324: The argument presented is not compelling.

► Thank you for pointing out the need for further clarification. We clarified the definitions of wave velocity and depth resolution in changes 33 and 32. We also rephrased our argumentation to make it clearer and more compelling. This also allowed us to further simplify the text and avoid the term relative error (compare change 35).

2.14 *Some parts of the manuscript have an informal tone.*

► Thank you for pointing this out. We adjusted several parts of the text to address this issue (see changes 1, 2, 5, 6, 7, 9, 4, 14, 18, 21, 20, 36, 37, 41, 42, and 46).

2.15 *The manuscript overstates the novelty and impact of its contributions. It describes the framework as the "first step" toward automated ice thickness mapping, despite acknowledging decades of prior research. Similarly, the claim that this work has "invited other scientists to start working" in this area overlooks longstanding studies. These statements should be revised to more accurately reflect the field's history. ...*

Line 448: *The statement "We believe that our framework is the first step towards a potential fully automated generation of ice thickness maps based on RES data" could be reworded for accuracy. As noted in the literature review, research in this area has been ongoing for nearly two decades. While this work is a valuable contribution, positioning it as the first step towards automation may not fully acknowledge prior advancements in the field.*

Line 462: *The statement suggesting that this work has "invited other scientists to start working in this research area" may overstate its impact. Given the examples of previous studies provided by the authors, it would be more accurate to acknowledge the long-standing research efforts in this field while highlighting how this study builds upon them.*

► Thank you for pointing this out. Our intention was not to discredit previous research in this area. We adjusted the wording of the sentences (compare changes 49 and 50)

2.16 *Line 98: Jebeli et al. 2023 have performed a very similar aim to this work in their study.*

Line 88: *Moqadam et al. 2024 (DOI: 10.22541/essoar.172987463.39597493/v1) also have done the tracking of internal layers.*

Line 98 – 105 : *The manuscript would benefit from citing additional relevant work to provide a more complete context for readers. For instance, Moqadam and Eisen (<https://doi.org/10.5194/egusphere-2024-1674>) offers a broad review of prior research on ice boundary extraction, making it a fitting reference at the end of the literature review.*

Line 102: *Where the use of CNN for automatic tracing of internal layers is mentioned, Jebeli et al. 2023*

(DOI: 10.13140/RG.2.2.23219.20007), Moqadam et al. 2024 (DOI: 10.22541/essoar.172987463.39597493/v1) directly addresses the application of deep learning to this task and would be valuable citations in the section discussing recent advancements in this area.

215 Including these references, along with other relevant studies would help situate the manuscript within the broader body of existing research and provide readers with a more comprehensive view of the field.

► Thank you for pointing this out. We included the missing references to previous work in Section 2.

2.17 Line 64: The statement, “however, a large portion of . . .,” should be supported with evidence. Importantly, the critique of automatically labeled bedrock seems contradictory, as the study itself aims to achieve this. Clarifying this point would strengthen the argument.

220► Thank you for highlighting this, as this is an important point. We do not intend to criticize datasets for utilizing automatic approaches to label their ice boundaries. However, we do criticize the use of such datasets in the context of training and evaluating deep learning approaches. Deep learning models essentially learn to imitate the labeling process. If all the labels are based on a peak detection algorithm, then our model learns peak detection. This defeats the purpose of employing deep learning models, which is to achieve more accurate predictions. Thus, we should avoid datasets where the labeling process was
225 significantly automatized or lacked the necessary transparency. We rephrased the text and added citations as an example. Please view changes 2 and 11.

2.18 Line 124: not clear what the authors want to say.

► Thank you for this comment. We have rephrased the text to make it clearer in change 13.

2.19 Line 134 – 137: More references are needed to support the claims

230► Thank you for pointing this out. We have added further references in this section to support our claim.

2.20 Line 141: “Hence, . . . “ it is not clear or accurate argument for the clearer signal of the thinner ice. The aim of the sentence is evident but the sentence should be reformulated.

► Thank you for bringing this to our attention. We have slightly adjusted the text in change 16.

2.21 Line 148: the sentence seems to be incomplete.

235► Thank you for bringing this to our attention. We have adjusted the manuscript accordingly in change 17.

2.22 Line 181: this process needs to be elaborated.

► Thank you for this comment. We have expanded on our explanation in change 22.

2.23 Line 226: “the” should be removed.

Line 251: hyphen needed between differently and sized.

240► Thank you for pointing this out. We have removed the extra ‘the’ and added the missing hyphen.

2.24 Line 381: The sentence needs to be rewritten for clarity.

► Thank you for bringing this to our attention. We rephrased the specified sentence in change 38.

2.25 Line 311: the authors mention that resizing changes the MAE so they introduce MME. It is not clear why they keep the MAE in the paper, if MME is a more suitable metric.

245► Thank you for this comment. We added an explanation in the text (compare addition 8).

2.26 Lines 392-399: These sentences need to be rewritten for clarity and flow.

► Thank you for bringing this to our attention. We have adjusted the text to improve clarity and flow (compare change 39).

2.27 Line 401: Please provide a more detailed explanation of the ablation study.

- Thank you for this comment. We expanded on our explanation of the ablation study in change 40.

250 Line 401: Please provide a more detailed explanation of the ablation study.

2.28 Line 414: the explanation of temperate ice this can appear much earlier in the manuscript

- Thank you for pointing this out. We added an earlier explanation in change 12.

2.29 line 418: the sentence should be rewritten.

- Thank you for pointing this out. We changed the sentence in change 43.

2.30 Line 421: It is obvious that the differences decrease when AP-5% is considered, and there is nothing surprising about this result. Please rewrite this statement or clarify the reasoning behind the argument.

- Thank you for mentioning this. You are correct that a decrease in error difference has to be expected, but our point was to highlight the strong decrease. We rephrased our point to make it clearer. Please view change 44.

2.31 Line 430: the sentence does not read well.

260► Thank you for this comment. We rephrased the sentence (compare change 45).

2.32 Line 437: Please provide further explanation. What exactly do you mean, and why is this the case?

- Thank you for this comment. We adjusted the sentence to make it clearer (compare change 47). Naturally, the ice surface and ice bottom are generally further apart in thicker ice.

2.33 Line 441: The authors mention that thicker ice is more challenging, but shouldn't it actually be easier, as less collapse would occur in thicker ice compared to thinner ice?

270► While we agree that the presence of shallow ice in the dataset first introduces the ice boundary collapse, once the model picks up on this characteristic, it also affects the predictions of thicker ice. The problem of the ice boundary collapse ultimately stems from the nature of differentiated radargrams, where we only visualize the change in amplitude. Here, peaks indicate a change in amplitude and are usually connected to the ice surface or ice bottom. When there is no or only very shallow ice, the peaks for the ice surface and bottom start overlapping, and it is often unclear to the model whether the two peaks overlap or whether it cannot find the peak for the ice bottom. Thus, it develops a bias to predict the ice bottom as the ice surface whenever it cannot find the second peak corresponding to the ice bottom.

275 Why does this affect thicker ice more than shallow ice? In thicker ice, the returned signal of the radar system is generally weaker than in shallow ice due to attenuation and more potential sources of interference. Thus, the depicted peak for the ice bottom becomes less clear. As a result, the model now faces considerable difficulty in determining whether the weak peak corresponds to the true ice bottom or if the ice surface and bottom overlap. We also added this explanation to the manuscript in addition 9.

Ice Anatomy: A Benchmark Dataset and Methodology for Automatic Ice Boundary Extraction from Radio-Echo Sounding Data

Marcel Dreier¹, Moritz Koch², Nora Gourmelon¹, Norbert Blindow², Daniel Steinhage³, Fei Wu¹, Thorsten Seehaus², Matthias Braun², Andreas Maier¹, and Vincent Christlein¹

¹Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany.

²Institut für Geographie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany.

³Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany.

Correspondence: Marcel Dreier (marcel.dreier@fau.de)

Abstract. The measurement of ice thickness is of great importance for the accurate estimation of glacier volume and the delineation of their bedrock topography. In particular, this is a crucial factor in forecasting the future evolution of glaciers in the context of a changing climate. In order to derive the ice thickness, the travel time of electromagnetic waves in radargrams acquired by radio-echo sounding (RES) systems is analyzed. This can only be achieved by identifying the ice surface and underlying ice bottom in corresponding radargrams. Manually identifying these two reflection horizons in RES data is a laborious and time-consuming process. Consequently, scientists are attempting to automate this task through the use of techniques such as deep learning. Such automation can significantly reduce the time between a field campaign and the calculation of the glacier’s ice thickness distribution. In this paper, we present the first benchmark dataset for delineating the ice surface and bottom boundaries in RES data, to facilitate [straightforwardstandardized](#) comparisons of deep learning models in the future. The “IceAnatomy” dataset comprises radargrams and the corresponding manual picks, amounting to a total of over 45,000 km of observations. The RES data originates from three sources: FAU, CReSIS, and AWI. The dataset comprises different RES systems as well as different pre-processing methods. In addition, the data was acquired over a large range of geographical and glaciological settings, featuring different thermal regimes present in Antarctica and the Southern Patagonian Icefield. This diversity ensures that the models’ behaviors can be analyzed in different scenarios. We define a standardized train-test split for each source in the dataset. This allows us to introduce not only a baseline model trained on the entire training set (the “omni” model), but also three source-specific baseline models. The source-specific models are trained exclusively on the subset of the training data acquired by the specified source. The baseline models provide an initial benchmark against which subsequent models can be compared. The source-specific models demonstrate more accurate results than the omni model. For the FAU, CReSIS, and AWI test sets, the source-specific models achieve Mean Meter Errors of 2.1 m, 23.1 m, and 4.9 m for the ice surface and 9.1 m, 78.2 m, and 29.3 m for the ice bottom. In relation to the mean measured ice thickness of the test set, these errors equate to 1.2%, 3.1%, and 0.3% for the ice surface and 4.9%, 10.4%, and 1.5% for the ice bottom. The dataset and implementation are available at <https://zenodo.org/records/14036897> (Dreier et al., 2024) and <https://doi.org/10.5281/zenodo.14038570> (Dreier, 2024).

1 Introduction

Glaciers and ice shelves are key indicators of global climate (Haeberli et al., 2007; IPCC, 2013). Knowing their volume and ice thickness distribution is crucial for assessing future cryospheric contributions to sea level rise. Moreover, data on the ice volume of glaciers and ice sheets is necessary for understanding their response to climate change. Ice thickness measurements enable the subsequent prediction of the rate and timing of glacier retreat or disappearance using different types of models. ~~That way, a glacier's contribution to regional hydrological cycles and subsequent influence on local to regional scales with associated socioeconomic impacts can be assessed~~ This enables the assessment of a glacier's contribution to regional hydrological cycles and its subsequent influence on local to regional scales with associated socioeconomic impacts. (Werder et al., 2020; Ayala et al., 2020; Farinotti et al., 2017). Several techniques to determine ice thickness exist, including seismic, gravitational, and magnetic methods, as well as radio-echo sounding (RES) (Bogorodsky et al., 2012; Kohler et al., 1997). While satellite gravimetry allows for a resolution in the range of kilometers, its spatial resolution does not allow for the interpretation of detailed subglacial features (Willen et al., 2024). Seismic measurements offer a high resolution, but widespread use in the Antarctic region is limited by high exploration costs or logistical unfeasibility (An et al., 2023). For this reason, RES is preferred over other methods when an accurate assessment of a subglacial topography is of interest. ~~After pre-processing the RES data, a cross-section, a so-called radargram of the glacier, becomes visible.~~ After pre-processing the RES data, we obtain an image commonly referred to as a radargram. It depicts the cross-section of the glacier along the flight path. Experts can then interpret the RES data by delineating the reflections of surfaces or internal glacial structures. Delineating the ice boundary, defined by the ~~air-ice layer and ice-ground layer~~ ice surface (air to ice transition) and the ice bottom (ice to ground/water transition), is necessary to obtain the glacier's thickness at each point in the radargram. However, it is a ~~laborious~~ time-intensive task, especially with large datasets (Sime et al., 2011). Several automated and semi-automated approaches to delineate the layers have been developed (Fahnestock et al., 2001; Gifford et al., 2010; Freeman et al., 2010; Rahnemoonfar et al., 2017a, b; Kamangir et al., 2018; Rahnemoonfar et al., 2019; Cai et al., 2020, 2022; Liu-Schiaffini et al., 2022b; Moqadam and Eisen, 2024; Moqadam et al., 2024; Jebeli et al., 2023b). However, these approaches are not comparable as they have been evaluated on different datasets or a different train-test split of the same dataset. In this paper, we present a publicly available, ~~ready-to-use~~ standardized benchmark dataset for ice thickness extraction. It is the first of its kind to be directly ~~conjured~~ designed for deep learning approaches, with a pre-defined train-test split, human-annotated labels, and different recording systems. It comprises radargrams from Antarctica and Patagonia with polythermal, cold-based, or temperate thermal regimes. The dataset is intended for supervised training and evaluation of deep learning models. Therefore, the dataset includes depth labels for both the ice bottom and ice surface layer. Together with the dataset, we present a baseline model that delineates the ice boundary in a given radargram. The model is based on the U-Net architecture (Ronneberger et al., 2015) and serves as a reference and a starting point for future improvements. We envision further development of this method in two main directions. First, once trained, our algorithm can be executed on virtually any modern laptop in the field. Combined with a pre-processing chain tailored to our approach, this allows for near real-time analysis of acquired data on-site. Since flight hours are costly and often limited by weather conditions, optimizing their use is crucial. If data can be processed in the field—e.g., between two flights—flight

335 plans could be dynamically adjusted to focus on areas of high interest within the same campaign. Second, and perhaps more
 importantly, the presented method can be further developed to handle more specialized tasks, such as delineating intraglacial
 water channel systems or identifying water tables within existing datasets. This would represent a step toward a comprehensive,
 quantitative, and standardized approach for interpreting radargrams, ultimately leading to fully automated products that could
 significantly benefit the cryospheric research community. In particular, the automated mapping of internal reflection layers
 340 remains a critical knowledge gap – one that deep learning is well – positioned to address (Moqadam and Eisen, 2024). ^{A1 (2.5)}

In summary, our contributions are as follows:

1. A novel benchmark dataset IceAnatomy for deep learning-based extraction of ice boundary from RES data is created.
2. A baseline deep learning model for the automatic delineation of ~~ice bottom and ice layer~~the ice bottom and the ice
 surface is proposed. ^{C8 (2.1)}
- 345 3. A thorough evaluation of individual models and an omni-model is conducted on the dataset.

The work is structured as follows: Section 2 provides an overview of datasets and algorithms previously used for automatic
 ice boundary extraction. Subsequently, Sect. 3 gives insight into the recording and processing of the dataset as well as relevant
 geographical and glaciological factors of the study sites. The baseline models are introduced in Sect. 4. An extensive evalu-
 ation of the baseline models and the benchmark dataset is presented in Sect. 5. Lastly, we summarize our research and draw
 350 conclusions in Sect. 6.

2 Related Works

Over the past decades, RES has been widely used in glaciology. A multitude of publications cover the extraction of ice boundary
 layers from RES data. In this section, we highlight related RES datasets and layer extraction approaches.

2.1 Datasets

355 ~~RES data on glaciers and ice sheets is abundantly available.~~Numerous RES datasets on glaciers and ice sheets are publicly
 available. However, a large portion of the associated ~~bedrock labels~~ice bottom labels are inaccurate, generated automatically,
~~or unavailable.~~unavailable, lack the necessary transparency in their creation, or do not have associated radargrams (Young ^{C9 (2.14)}
^{C10 (2.1)}
 et al., 2021; Blankenship et al., 2018; CReSIS; Dong et al., 2022; Corr, 2020; Corr et al., 2020). This makes them unsuitable ^{C11 (2.17)}
 for training or evaluating deep learning approaches, as they require human-annotated data. Hence, we focus our comparison ^{A2 (2.17)}
 360 on datasets that have been used to extract the ice boundary in previously published work and for which both radargrams and
 human-annotated labels are publicly available. These constraints significantly limit the number of related datasets.

The one RES system that has been used extensively to collect such data is the Multichannel Coherent Radar Depth Sounder
 versions 1-5 (MCoRDS) (Allen et al., 2012a), which was used, for example, in NASA’s Operation IceBridge (OIB) program
 on a McDonnell Douglas DC-8-72 jetliner (Shi et al., 2010a). The data acquired over Antarctica in 2009 are the most widely

used (Crandall et al., 2012; Lee et al., 2014; Rahnemoonfar et al., 2017a, b; Berger et al., 2018; Kamangir et al., 2018). However, also data from different years (Kamangir et al., 2018; Mitchell et al., 2013; Cai et al., 2020; García et al., 2021a, b; Cai et al., 2022, 2019; Ghosh and Bovolo, 2022; García et al., 2023; Donini et al., 2022; Ilisei and Bruzzone, 2014, 2015) and other locations like Greenland (Donini et al., 2022) and the Canadian Arctic Archipelago (Xu et al., 2017, 2018) were analyzed.

Only very few publications included data from RES systems other than MCoRDS. Gifford et al. (2010) extracted the ice boundary from data acquired by a predecessor RES system (Lohofener, 2006) during 2006 and 2007 in Greenland. Dong et al. (2022) featured data from the Chinese Academy of Sciences' Deep Ice Radar acquired during the 29th Chinese Antarctic Scientific Expedition. Lastly, Liu-Schiaffini et al. (2022a) used algorithm-assisted human-labeled data acquired in the Canadian Arctic and Antarctica by the University of Texas Institute for Geophysics' high-capability radar sounder (HiCARS). A major downside of these datasets is that they do not provide standardized training and evaluation splits, making inter-model comparison challenging. Furthermore, datasets usually only focus on a single area, e. g., Greenland or Antarctica, which makes generalization to other areas or glaciological settings difficult to verify. IceAnatomy addresses this issue by including data from multiple study sites, radar systems, and glaciological settings. It also provides standardized splits for training and evaluation to allow for an accurate and fair comparison between models. In conclusion, to the best of our knowledge, there is no comparable benchmark dataset for ice boundary extraction from radio-echo-sounding data. ^{A3 (2.2)}

2.2 Algorithms

RES has been employed to detect crevasses (Liu et al., 2020; Walker and Ray, 2019; Williams et al., 2012, 2014), the ice boundary (Crandall et al., 2012; Lee et al., 2014; Rahnemoonfar et al., 2017a, b; Berger et al., 2018; Kamangir et al., 2018; Mitchell et al., 2013; Xu et al., 2017, 2018; Cai et al., 2022; Gifford et al., 2010; Dong et al., 2022; Liu-Schiaffini et al., 2022a), to segment subsurface structures (Cai et al., 2020, 2019; García et al., 2021a, b; Ghosh and Bovolo, 2022; García et al., 2023; Donini et al., 2022; Ilisei and Bruzzone, 2014, 2015), and to track internal ice and snow layers (Crandall et al., 2012; Karlsson et al., 2013; Ibikunle et al., 2020; Rahnemoonfar et al., 2021; Varshney et al., 2020, 2021; Yari et al., 2019, 2020; Dong et al., 2022; Moqadam and Eisen, 2024).

To obtain the ice boundary, one can either directly delineate the ice surface and bottom or first segment different regions such as ice, bedrock, and air and then extract the two layers during post-processing. Most existing studies (Crandall et al., 2012; Lee et al., 2014; Rahnemoonfar et al., 2017a, b; Berger et al., 2018; Kamangir et al., 2018; Mitchell et al., 2013; Xu et al., 2017, 2018; Cai et al., 2022; Gifford et al., 2010; Dong et al., 2022; Liu-Schiaffini et al., 2022a) ~~to date~~ prefer direct extraction. ^{A4 (2.14)} ^{D1 (2.14)} Fewer studies (Cai et al., 2020, 2019; García et al., 2021a, b; Ghosh and Bovolo, 2022; García et al., 2023; Donini et al., 2022; Ilisei and Bruzzone, 2014, 2015) use the segmentation approach. The segmentation approach assigns a semantic class to each pixel in the radargram, from which the ice boundaries can be derived directly or after post-processing.

In terms of methodology, early studies mainly used classical image processing and machine learning techniques such as Hidden Markov Models (Crandall et al., 2012; Berger et al., 2018), Markov-Chain Monte Carlo (Lee et al., 2014), contour detection (Rahnemoonfar et al., 2017a), the level set approach (Rahnemoonfar et al., 2017b; Mitchell et al., 2013), Markov

Random Fields (Xu et al., 2017), edge-based and active contour methods (Gifford et al., 2010), Kullback-Leibler maps (Ilisei and Bruzzone, 2014), and Support Vector Machines (Ilisei and Bruzzone, 2015). After 2017, studies turned to Convolutional Neural Networks (CNNs) (Kamangir et al., 2018; Cai et al., 2020, 2019; García et al., 2021a; Cai et al., 2022; Donini et al., 2022; Dong et al., 2022; Liu-Schiaffini et al., 2022a; García et al., 2021b, 2023; Jebeli et al., 2023a, b; Matsuoka et al., 2021; Moqadam et al., 2024), combinations of CNNs and Recurrent Neural Networks (RNNs) (Xu et al., 2018), and combinations of CNNs and Transformers (Ghosh and Bovolo, 2022).

In comparison, we rely on the U-Net architecture from (Ronneberger et al., 2015) to evaluate our newly created dataset. Furthermore, we integrate Atrous Spatial Pyramid Pooling from (Chen et al., 2018) and the ResBlock design from (Esser et al., 2020) to improve the performance.

3 Dataset

In this section, we introduce the benchmark dataset “IceAnatomy” which covers several different geolocations and was acquired by multiple radar systems. We divide the dataset into three subsets based on the sources of the data: the AWI (Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research), CReSIS (The Center for Remote Sensing and Integrated Systems), and FAU (Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute of Geography) subsets. A summary of the most important information about the dataset is given in Tab. 1.

Table 1. A summary of details about the IceAnatomy benchmark dataset (Lippl et al., 2019; Shi et al., 2010b; Rückamp and Blindow, 2012; CReSIS, 2024a; Allen et al., 2012b; Steinhage, 2001, 2015).

	Study Sites	Depth-Reso.	Width-Reso.	Length	Year	Main Thermal Regime	Labeled Bottom %
FAU	James Ross Island	2.5 ns pixel ⁻¹	2 m pixel ⁻¹	275 km	2017/18	Polythermal	82.5 %
	Perito Moreno	2.5 ns pixel ⁻¹	2 m pixel ⁻¹	145 km	2022	Temperate	83.1 %
	Viedma	2.5 ns pixel ⁻¹	2 m pixel ⁻¹	140 km	2022	Temperate	46.2 %
CReSIS	Antarctic Peninsula	105 ns pixel ⁻¹	12 m pixel ⁻¹	20400 km	2009	Polythermal	63.9 %
	West Antarctica	105 ns pixel ⁻¹	12 – 30 m pixel ⁻¹	24400 km	2009	Polythermal	78.9 %
AWI	Antarctic Peninsula	12 ns pixel ⁻¹	62 m pixel ⁻¹	1490 km	2013	Polythermal	31.7 %
	East Antarctica	13.33 ns pixel ⁻¹	66 – 79 m pixel ⁻¹	1015 km	1997/99	Cold-based	73.7 %

3.1 Study Sites

3.1.1 Southern Patagonian Icefield

The Southern Patagonian Icefield (SPI) is the largest temperate ice body in the Southern Hemisphere. It is characterized by one of the highest mass loss rates in the world (Zemp et al., 2019; Marzeion et al., 2018; Hugonnet et al., 2021) and by its large outlet

glaciers that drain into lakes or the ocean (Aniya, 1999). Two of the largest eastward-flowing outlet glaciers in the region are the Perito Moreno and Viedma glaciers. The only way to obtain information over large areas about their bedrock topography is by helicopter-borne RES measurements. This is particularly applicable to the lower parts of the glaciers, which are surrounded by steep mountain flanks and have heavily crevassed surfaces. ~~The temperate nature of the glaciers, resulting in high water content in the ice, combined with the steep and deep glacier troughs, often makes analyzing radargrams challenging. As the glaciers are temperate, i. e., most of the ice is close to or at the pressure melting point, they contain a relatively high proportion of water. This characteristic, combined with the steep and deep glacier troughs, often makes analyzing radargrams challenging.~~ ^{C12 (2.28)}

~~Hence, they pose an adequate challenge to benchmark new machine learning systems. Hence, they pose a significant challenge to machine learning systems.~~ ^{C13 (2.18)}

Figure 1 shows the location of both glaciers on the east of the SPI.

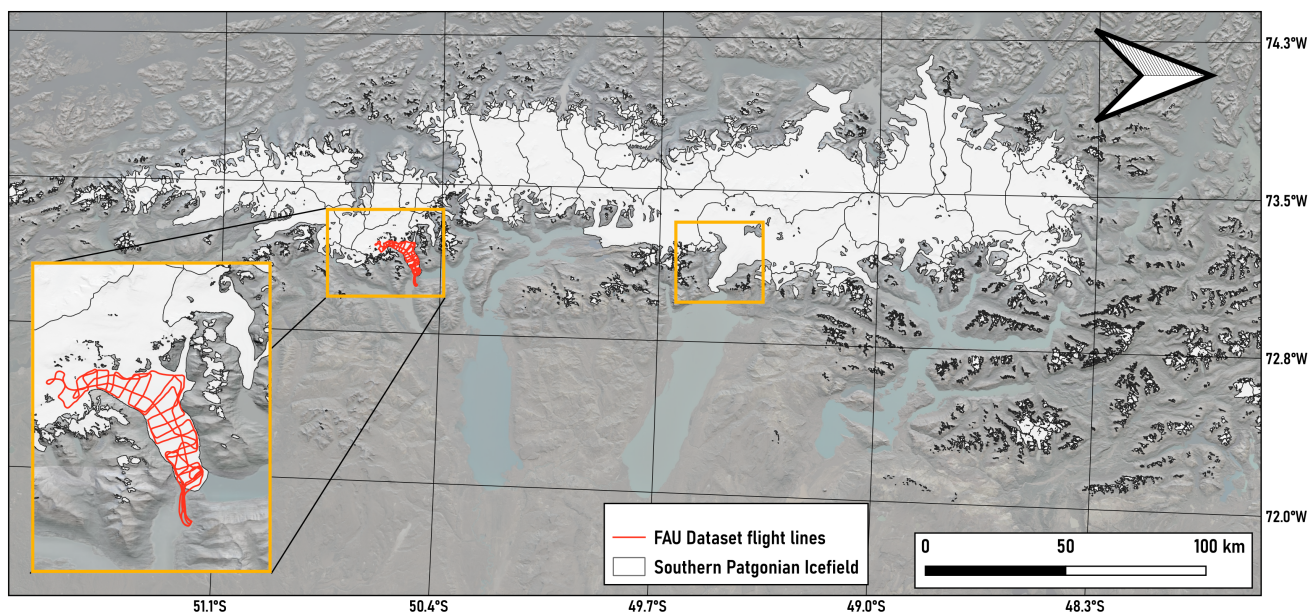


Figure 1. Overview of the Southern Patagonian Inland Icefield. Orange boxes indicate surveyed areas of Perito Moreno Glacier and Viedma Glacier. Black lines indicate flight paths over the Perito Moreno Glacier. The background is a hillshaded SRTM over ©Google Earth optical imagery (Consortium, 2017). Maps are rotated by 90 degrees.

3.1.2 Antarctica

As depicted in Fig. 2, the Ice-Anatomy dataset offers three major study sites in Antarctica: the Antarctic Peninsula (including James Ross Island (JRI)), West Antarctica, and East Antarctica.

The Antarctic Peninsula is the most represented region in the benchmark dataset, as it is present in all three RES subsets. It exhibits one of the milder climates in Antarctica, with an annual average temperature of -3.2°C (Morris and Vaughan, 1994). This is also reflected in the thermal regimes present in the region, as it contains temperate, cold-based, and polythermal ice. The

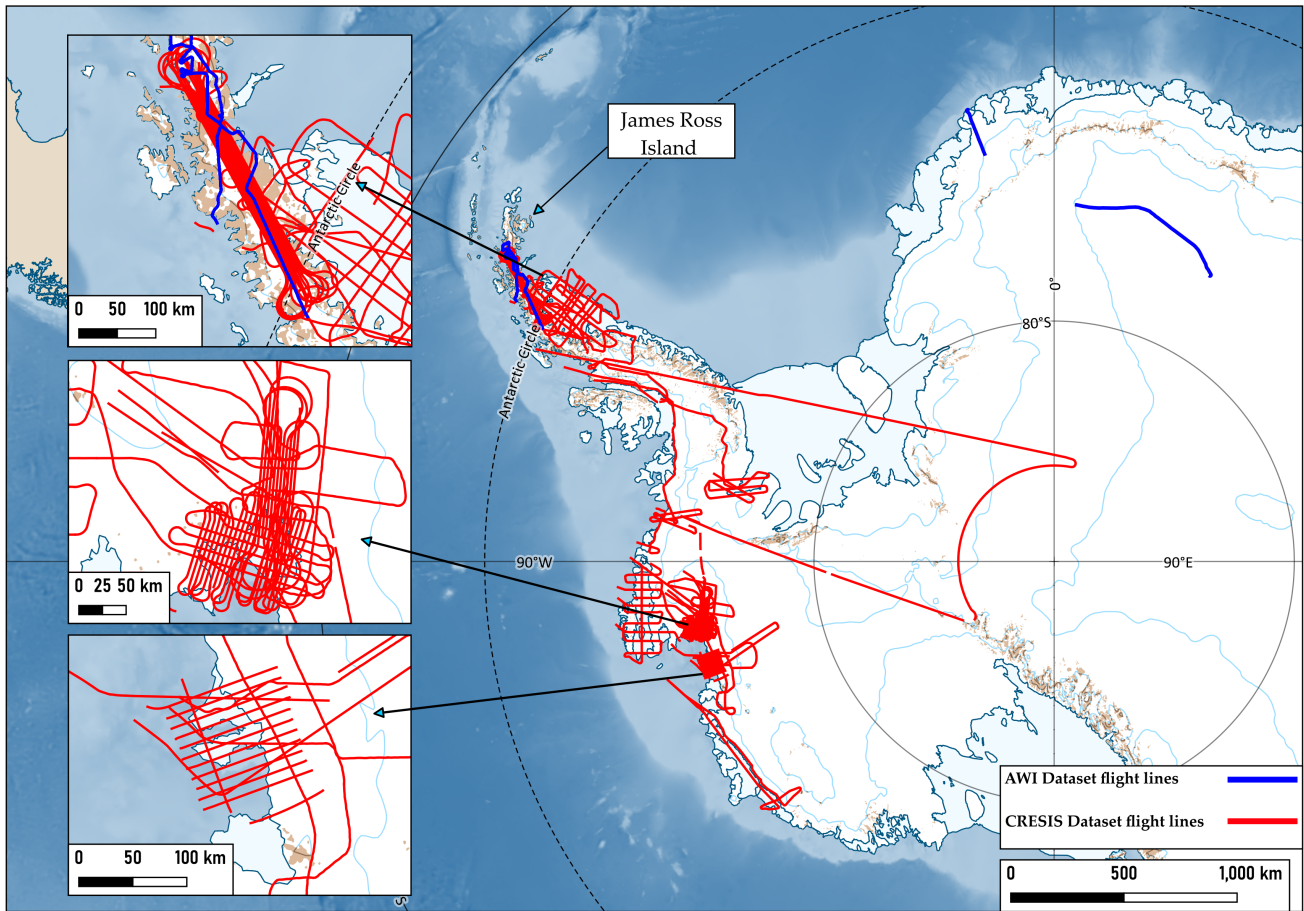


Figure 2. Flight paths of the AWI and CReSIS campaigns in Antarctica. The background is assembled with help from the Quantarctica QGIS project (Matsuoka et al., 2021).

temperate portions of the Antarctic Peninsula are frequently near the margins and at lower elevations, while the cold-based ice regions are generally found at higher elevations. The transition zones between higher and lower elevations commonly contain polythermal ice (Van Lieffering and Pattyn, 2013; Macelloni et al., 2019). However, elevation alone is often not sufficient to determine the thermal regime. If we compare the ice velocity maps of Rignot et al., we observe fast-moving ice even in higher elevation areas, which is atypical for cold-based areas. This leads us to the assumption that there is a significant amount of polythermal ice at higher elevations and that the main thermal regime is polythermal. A comparison with the ice velocity maps of Rignot et al. (2011) reveals that fast-moving ice is present even in higher elevation areas, which is atypical for cold-based areas (Park et al., 2024; Dawson et al., 2022). This suggests that there is a significant amount of polythermal ice at higher elevations and that the main thermal regime is polythermal. Another noteworthy significant characteristic of the Antarctica Peninsula is its relatively shallow ice sheet compared to the rest of Antarctica. On average, the ice sheet is estimated to be

C14 (2.14)
C15 (2.14)

610 m thick inland and 300 m in the ice shelves (Drewry et al., 1982). This results in a generally clearer signal because the signal has to travel through less ice. ~~Hence the returned amplitudes are usually higher and are less likely to be distorted by impurities in the glacier.~~ ~~and is less likely to be distorted by impurities in the glacier.~~ C16 (2.20)

West Antarctica is significantly colder than the Antarctic Peninsula, with an annual average temperature of approximately -28.1°C and a primarily polythermal thermal regime (Morris and Vaughan, 1994). Polythermal regions only commonly occur at the margins and the coastline, while cold-based zones are mainly present at higher elevations (Macelloni et al., 2019; Van Liefferinge and Pattyn, 2013; Rignot et al., 2011). West Antarctica also contains relatively thick ice with inland ice sheets estimated to be 1780 m thick and ice shelves around 375 m (Drewry et al., 1982).

The last subregion in Antarctica is East Antarctica. It exhibits the coldest climate of the three areas, with an annual average temperature of around -59.8°C and a primarily cold-based thermal regime (Morris and Vaughan, 1994). ~~Temperate areas are commonly only near the margins, while polythermal zones act as a transition~~ ~~Temperate areas are commonly found near the margins, while polythermal zones act as a transitional zone between the cold-based and temperate areas~~ (Macelloni et al., C17 (2.21)
2019; Van Liefferinge and Pattyn, 2013; Rignot et al., 2011). East Antarctica is also the region with the generally thickest ice. On average, its ice sheets are approximately 2630 m thick inland and 400 m in its ice shelves (Drewry et al., 1982).

3.2 Dataset Generation

3.2.1 FAU Data

The RES system of FAU is a broadband 25 MHz bi-static monopulse sounder designed as a sling load for helicopter use. It is a functional ~~clone~~ ~~replica~~ of the BGR-P30 system (Blindow et al., 2012). The antenna weighs roughly 280 kg and can be attached to any helicopter type that allows for the attachment of a sling load and has the required take-off capacity. The system is typically operated 20 m above ground at a nominal airspeed of 60 km h^{-1} . C18 (2.14)

The radar time series are collected at a 2.5 ns sampling rate using 256-fold stacking to improve sensitivity and signal-to-noise ratio. The traces are collected at a rate of 10 Hz, corresponding to approximately a 2-meter spatial sampling rate. The data are georeferenced by two Leica GS16 multifrequency Global Navigation Satellite System (GNSS) systems. The rover antenna is mounted on the radar antenna in a central position, while the base station is installed in proximity to the landing and starting area. After differential processing of the GNSS data, the positions are matched to the radar traces before further processing is applied. Then, the RES data is processed in REFLEX v8.5 software, developed by Sandmeier Geophysical Research. The processing flow comprises the following steps and is applied to subsections of each flight: equidistant trace interpolation, shift for time zero, subtracting special average, bandpass filter, amplitude regulation by gain function (cold ice) or energy decay (temperate ice), 2D migration, and static correction. To apply the 2D migration, it is necessary to derive a velocity model comprising an air and an ice layer. For the air layer, the wave travels at the speed of light, while for the ice layer, we assumed a speed of 0.168 m ns^{-1} (Johari and Charette, 1975). Especially in temperate ice, the migration helps to focus the scattered energy to enhance the ~~bedrock~~ ~~ice bottom~~ reflections. C19 (2.1)

475 The RES data of JRI was acquired during two different airborne ground penetrating radar campaigns in 2017 and 2018 (Lippl
 et al., 2019). Since Gourdon Glacier consists mainly of bare ice, no firn correction was applied for the outer parts of the profile.
 For the data on the plateau, a standard correction for firn and snow (+10 m, AWI/BAS Bedmap 1 mission summary) as used in
 the British-Argentinian survey was assumed (Lythe and Vaughan, 2001). The RES data of Perito Moreno Glacier and Viedma
 Glacier were acquired in March and April 2022. For these study sites in the FAU subset, the ~~radargrams go deeper than the ice~~
 480 ~~is thicker than the radar's maximum penetration depth of 700 m~~ the ice is thicker than the radar's maximum penetration depth
 of estimated 700 m (Blindow et al., 2011). The original depth of the radargrams is over 6000 pixels, which equates to over ^{C20 (2.14)}
 1300 m on average - the total depth in meters is not constant due to fluctuations in the flight height of the helicopter. We cut the
 radargrams to 4096 pixels, which corresponds to an average of about 800 m. ~~This way, we save computing power while keeping~~
~~all the necessary information.~~ This saves computing power while keeping all the essential information. ^{C21 (2.14)} To restore the full flight
 485 traces in the FAU dataset from their subsampled parts, we reassembled the radargrams according to their trace numbers. Any
 conflicting depths for the ice surface and bottom in overlapping parts were smoothened with Gaussian importance weighting.
~~Furthermore, we filled gaps of eleven pixels or less in the ice surface and bottom via bicubic interpolation.~~ Furthermore, in rare
 cases, the initially labeled ice surface and bottom had small gaps. To avoid such inconsistencies, we filled gaps of eleven pixels
 or less in the ice surface and bottom via bicubic interpolation, using the two nearest manually labeled points as a reference. ^{C22 (2.22)}
 490 The initial layer labels were annotated by a single interpreter to ensure consistency throughout the dataset. Surface reflections
 were generally straightforward to identify; however, in heavily crevassed areas, we increased the resolution to delineate the air-
 ice interface as accurately as possible across these features. Bedrock picks were conducted using the same approach. In regions
 with ambiguous reflections, ReflexW software enabled zooming into specific subsets of the radargrams, thereby enhancing
 the clarity of features of interest. Additionally, several intersecting profile lines provided cross-points for internal validation.
 495 These intersections were annotated independently by the same interpreter and subsequently compared. All cross-profile values
 fell within the expected margin of error, even in areas with steep slopes or greater depths (i.e., deviations < 10 %). At Glacier
 Perito Moreno, two control points from previous studies were available for comparison (Sugiyama et al., 2011; Stuefer et al.,
 2007). The first, along the 'Buscaini' profile, corresponds to a seismic survey conducted in 1996, which reported a maximum
 ice thickness of 720 m. The second, located nearer to the glacier terminus, corresponds to a borehole drilled in 2010, revealing
 500 an ice thickness of 515 ± 5 m. Both control points were in close agreement with our ice thickness estimates. ^{A5 (2.27)}

3.2.2 CReSIS Data

The CReSIS data was recorded during the 2009 campaign of Operation Ice Bridge in Antarctica, which comprised 21 missions.
 Three were sea-ice surveys and thus are not included in the CReSIS dataset. The remaining 18 missions can be split into
 two groups: six missions focusing on the Antarctic Peninsula (PEN1, PEN2, PEN3, PEN4, PEN5, and LVISPEN) and 12
 505 missions exploring West Antarctica (PIG1, PIG2, PIG3, PIG4, LVISPIG, LVIS86, GETZ1, ABBOTT1, TSK1, TSK2, TSK3,
 and TSK4) (Allen et al., 2012b). All 18 missions employed the Multichannel Coherent Radar Depth Sounder (MCoRDS) flown
 on a McDonnell Douglas DC-8-72. It has a center frequency of 195 MHz and an eight-channel-chirp signal to accurately assess
 the ~~groundice~~ (Rodriguez-Morales et al., 2014; Shi et al., 2010b). ^{C23 (2.1)}

To process the recorded data, the standard CReSIS L1B CSARP-mvdr (minimum variance distortionless response) processing steps were applied. These include pulse compression via a Tukey and Hanning Window, beam-forming, motion compensation, synthetic aperture radar processing in combination with f-k migration, channel combination, and waveform combination (CReSIS, 2024b). After the processing, the radargrams had a depth resolution of $105 \text{ ns pixel}^{-1}$ and a width resolution of $12 - 30 \text{ m pixel}^{-1}$ depending on the mission.

We obtained the fully processed CReSIS subset by downloading the CSARP-mvdr processed L1B product from the CReSIS website and taking the square root of the amplitudes. Likewise, CReSIS also provides downloads for the annotated [glacier and bedrock surface ice bottom and surface](#) layers on their website (CReSIS, 2024a). According to Lee et al. (2014) and Crandall et al. (2012), the rock-bed surface is humanly annotated but noisy. Although the noise might pose a problem for certain approaches, we chose not to alter the labels. The reason for this is that the dataset has been used previously in other publications, and in order to remain comparable, we use the same labels. ~~Nonetheless, we also include the CReSIS quality labels in the benchmark dataset to highlight the picker's confidence in each label.~~ However, to provide additional context regarding the quality, CReSIS provides a quality label for every pick. The label indicates the annotator's confidence, ranging from one (high) to three (low). We include these labels in the benchmark dataset for future research. The general picking procedure for CReSIS data is outlined in (CReSIS, 2024b).

3.2.3 AWI Data

The AWI subset was recorded during campaigns in Dronning Maud Land in 1997 and 1999 (Steinhage et al., 2023b, a) and in the Antarctic Peninsula in November 2013 (Steinhage, 2015). All three campaigns employed a version of the EMR radar system with a center frequency of 150 MHz and the toggle mode enabled. The toggle mode alternates the radar's pulse length between 60 ns and 600 ns periodically. Thus, the system can achieve a decent depth resolution while capturing deep internal layers of the ice. The processing of the recorded data was similar for all three campaigns. The data was differentiated, rescaled, high-pass filtered, and bandpass filtered. To reduce the amount of noise in a radargram, multiple traces were combined into a single trace. In detail, ten traces were combined for the 1997 and 1999 flights, and seven traces were combined for the 2013 flight (Steinhage, 2001; Nixdorf et al., 1999; Steinhage et al., 2001). Automatic gain control was used to normalize the amplitude values. After the processing, the radargrams had a depth resolution of $12 - 13.33 \text{ ns pixel}^{-1}$ and a width resolution of $66 - 79 \text{ m pixel}^{-1}$ depending on the campaign.

The ice surface and ice bottom were annotated by one person. To ensure consistency, plausibility checks were performed at crossing points with other profiles from the same or related campaigns. No systematic biases were observed. In the picks, gaps of eleven pixels or less were filled using bicubic interpolation. Finally, for the radargrams from 1997 and 1999, [everything all data](#) below 3600 pixels, which is about 4 km, was discarded because only noise was visible at these depths. The gathered data was processed with FOCUS, DISCO, LANDMARK, and Python.

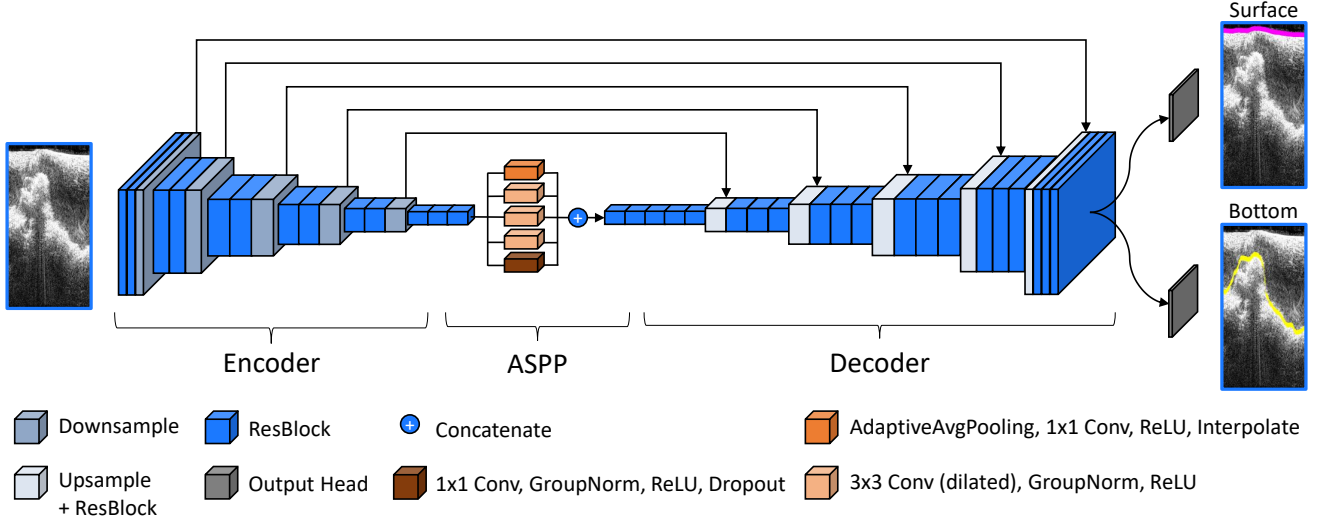


Figure 3. The architecture of the proposed deep learning model. It receives the normalized amplitudes of a radargram as input and predicts the ice surface and the ice bottom as two separate outputs. The atrous spatial pyramid pooling contains three dilated convolutional layers, one convolutional layer with adaptive average pooling, and a 1×1 convolutional layer. It utilizes the rectified linear unit (ReLU) activation function which is defined as $ReLU(x) = \max(x, 0)$.

575 *the bottom of the ice surface. In the bottleneck, we inserted an Atrous Spatial Pyramid Pooling (ASPP) (Chen et al., 2018) layer. ASPP processes the same feature map in parallel with differently sized fields of view. The sampling is uniform and based on a dilation rate. The chosen dilation rates in this model are 1, 4, and 6. Since the model is based on the U-Net architecture, it also includes skip connections. Skip connections directly forward the output of each down-sample step in the encoder to the corresponding parts in the decoder via concatenation. The increased channel dimensions in the decoder are solved by including an additional ResBlock for channel reduction after each up-sampling step in the decoder.*

585 To calculate the final prediction of the model, we first forward the feature map computed by the U-Net into two separate output heads, each consisting of a single ResBlock. Each output head then creates one probability map, resulting in two final probability maps. The first one represents the probabilistic prediction of the ice surface, while the second one represents the

probabilistic prediction of the ice bottom. The final prediction of the model is then the highest probable prediction of each column, which we compute by applying a column-wise argmax-operation.

590 To train the network, we employ a custom loss, a cost function that gives feedback to the network by measuring the difference between the prediction and the corresponding labeled ice boundary. The custom loss consists of two parts: a distance-based (L_{dist}) loss and a classification (L_{class}) loss:

$$L = L_{\text{dist}} + L_{\text{class}} \quad (2)$$

~~For the classification and distance-based losses, the probability maps of the ice surface ($\hat{Y}_s$) and ice bottom layer ($\hat{Y}_b$) are treated column-wise, i.e., per trace ($\hat{y}_{s/b}$). The classification loss is a smoothed cross-entropy loss (L_{CE}) that considers each pixel in a column C as a separate class c . The pixel closest to the corresponding labeled ice boundary layer is then the correct class. The loss gets smoothed with the smoothing factor ϵ_c . The distance-based loss sums up the probabilities in the column which are weighted with a distance map. The distance map contains the distance to the correct pick for each pixel. Hence, the further away the predicted pick is from the annotated layer, the greater the loss. The formulas of the classification and distance-based loss are as follows: For both the classification and distance-based losses, the probability maps of the ice surface ($\hat{Y}_s$) and ice bottom ($\hat{Y}_b$) are treated column-wise, i.e., per trace. The classification loss is a smoothed cross-entropy loss (L_{CE}) where each pixel in a column is treated as a separate class, and the pixel closest to the ground truth boundary is considered the correct class. The distance-based loss (L_{Dist}) sums up the probabilities in the column, which are weighted with a distance map. The distance map contains the distance to the correct pick for each pixel. Hence, the further away the predicted pick is from the annotated layer, the greater the loss. Section C provides a more in-depth overview of the loss function.~~

600

605 C30 (2,4)

$$L_{\text{CE}} = - \sum_{c \in C} x_c (1 - \epsilon_c) \log(p(x_c)) + \frac{\epsilon_c (1 - x_c) \log(p(x_c))}{|C|} \quad (3)$$

$$L_{\text{class}} = \frac{w_{s_class}}{|\hat{Y}_s|} \sum_{\hat{y}_s \in \hat{Y}_s} L_{\text{CE}}(\hat{y}_s) + \frac{w_{b_class}}{|\hat{Y}_b|} \sum_{\hat{y}_b \in \hat{Y}_b} L_{\text{CE}}(\hat{y}_b) \quad (4)$$

$$L_{\text{dist}} = \frac{w_{s_dist}}{|\hat{Y}_s|} \sum_{\hat{y}_s \in \hat{Y}_s} \langle d(y_s), \sigma(\hat{y}_s) \rangle + \frac{w_{b_dist}}{|\hat{Y}_b|} \sum_{\hat{y}_b \in \hat{Y}_b} \langle d(y_b), \sigma(\hat{y}_b) \rangle \quad (5)$$

~~With w_{s_class} , w_{b_class} , w_{s_dist} and w_{b_dist} the respective weights for a weighted combination of the single loss parts. $\langle \rangle$ is the dot product. σ is the softmax function that converts the model's outputs into probabilities, $|C|$ is the cardinality of C , and d is the function that creates a vector with the column-wise distance map given the respective column of the label.~~

610 D4 (2,4)

The annotations in the dataset have discontinuities in the labeled layers where the ice bottom dropped below the radar's penetration depth, the receiver flew over the edge of the glacier, or the signal was too ambiguous for experts to interpret. Tracks for which no pick is available for a layer are not included in the loss calculation and the evaluation.

615

5 Evaluation

5.1 Evaluation Metrics

Previous work either directly extracted the ice boundaries or deduced them from an intermediate segmentation, where they predicted a semantic class for every pixel in the radargram. Depending on the chosen method, the metrics used to assess the quality of the predictions differ. For segmentation approaches, most of these metrics are based on a confusion matrix that measures how accurately the model distinguishes between a chosen positive class and all the other classes, dubbed negative class. A confusion matrix contains four measurements: true positives (TP) (the number of correctly predicted pixels for the positive class), true negatives (TN) (the number of correctly predicted pixels for the negative class), false positives (FP) (the number of wrongly predicted pixels for the positive class), and false negatives (FN) (the number of wrongly predicted pixels for the negative class). Based on these four measurements, more sophisticated metrics are defined for the segmentation approaches. The most commonly employed one is the accuracy ($\frac{TP+TN}{TP+FP+FN+TN}$) (García et al., 2021a, b, 2023; Ghosh and Bovolo, 2022; Donini et al., 2022; Ilisei and Bruzzone, 2015). Less commonly used metrics include the Intersection over Union (IoU) ($\frac{TP}{TP+FP+FN}$) (Cai et al., 2019), precision ($\frac{TP}{TP+FP}$) (Ghosh and Bovolo, 2022), recall ($\frac{TP}{TP+FN}$) (Ghosh and Bovolo, 2022), the F1-score ($2 \frac{Precision \cdot Recall}{Precision + Recall}$) (Cai et al., 2020; Ghosh and Bovolo, 2022), sensitivity ($\frac{TP}{TP+FN}$) (García et al., 2023; Donini et al., 2022), specificity ($\frac{TN}{TN+FP}$) (García et al., 2023; Donini et al., 2022), and the error rate ($\frac{FP+FN}{TN+TP+FP+FN}$) (Ilisei and Bruzzone, 2014).

For direct extraction approaches, the mean column-wise absolute error also called mean absolute error (MAE) (Crandall et al., 2012; Lee et al., 2014; Rahnemounfar et al., 2017a; Berger et al., 2018; Mitchell et al., 2013; Xu et al., 2017, 2018; Gifford et al., 2010; Dong et al., 2022; Liu-Schiaffini et al., 2022a) is the most common metric. It measures the average pixel-wise distance between the annotated layer and the prediction. Other distance-based metrics include the median of the column-wise mean absolute error (Lee et al., 2014; Rahnemounfar et al., 2017a; Berger et al., 2018; Xu et al., 2017), the mean squared error (MSE) (Crandall et al., 2012; Mitchell et al., 2013; Dong et al., 2022), the root mean square error (RMSE) (Liu-Schiaffini et al., 2022a), and the largest under- and over-estimation (Gifford et al., 2010). ~~One problem with confusion matrix-based metrics like the precision is that they are not distance-weighted. We can also define confusion matrix-based metrics on the layer extraction task. In that case, we define each height pixel of the radargram as a separate class, and the closest pixel in each column to the corresponding layer as the correct class. However, a limitation of confusion matrix-based metrics, such as precision, is that they do not account for distance weighting. For example, if a prediction is always one pixel next to the~~ annotated layer, also known as ground truth (GT), the confusion matrix-based metrics will have the worst possible value, even though it is a near-perfect prediction. Therefore, some studies (Xu et al., 2017; Gifford et al., 2010; Liu-Schiaffini et al., 2022a) have relaxed these confusion matrix-based metrics by considering predictions a few pixels from the ground truth as still correct.

As metrics for our benchmark framework, we have chosen the MAE, two relaxed Average Precision (AP) metrics, and introduce the Mean Meter Error (MME). The MAE is calculated as the column-wise difference in pixels between the ground truth depth of a layer and the predicted depth. Resizing the radargram will change the value of this metric. Therefore, we

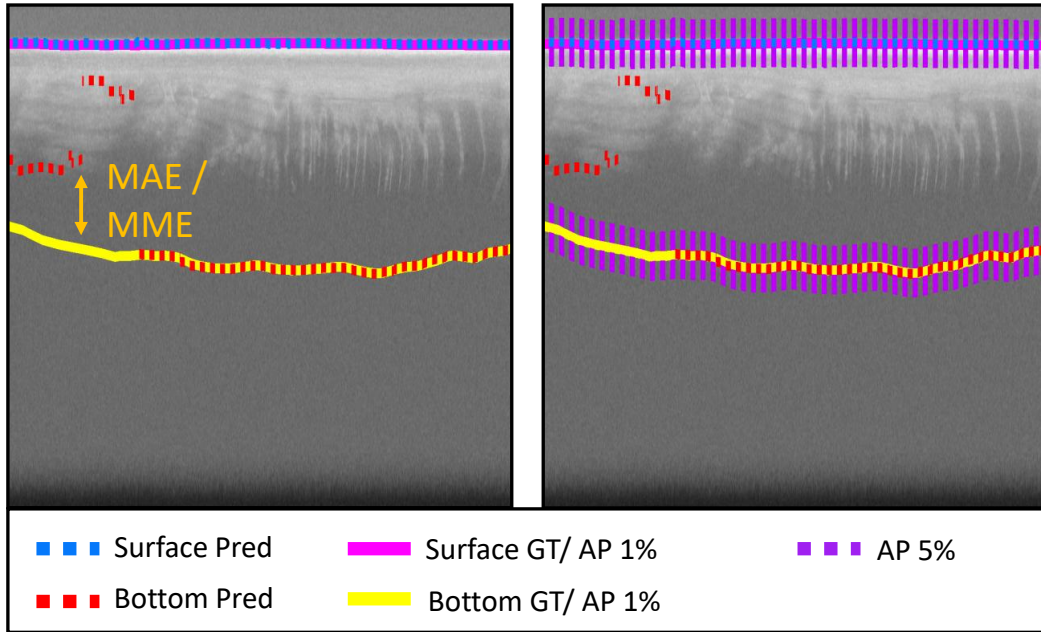


Figure 4. A visual representation of the four metrics used in this work. The left side of the figure depicts the MAE and MME respectively as the difference between prediction and ground truth (GT). Meanwhile, the right side of the figure features the AP-1% and AP-5% respectively as an interval around the ground truth. Note that the ground truth and the predictions are technically float numbers. However, we thickened the ground truth by 20 pixels to improve visibility.

also introduce the MME, which approximates the real-world error. We calculate the MME by multiplying the MAE with the product of the wave velocity in the medium and the depth resolution of the radargram. The speed of the wave in the medium is assumed to be constant with the speed of light ($c_{\text{air}} = 0.299792458 \text{ m ns}^{-1}$) for the air layer and $c_{\text{ice}} = 0.168 \text{ m ns}^{-1}$ for the ice layer. The wave velocity describes the speed of the electromagnetic wave of the radar through a medium. We assume it to be constant with the speed of light ($c_{\text{air}} = 0.299792458 \text{ m ns}^{-1}$) in air and with $c_{\text{ice}} = 0.168 \text{ m ns}^{-1}$ in ice (Johari and Charette, 1975). The depth resolution is the time it takes for the wave to pass through a pixel through the physical equivalent of a pixel in the radargram. Since the depth resolution is indirectly proportional to the y-dimension of the radargram, the MME stays consistent across different heights. Table 1 records the different depth resolutions for radargrams in the IceAnatomy dataset in their original height and equation 6 and 7 summarize the formula for the MME. Note that the MME is still highly dependent on the original depth resolution of the radargram. The MME will be naturally higher for a radargram where every pixel constitutes a 40 m change in height rather than a 4 m change, as even small mistakes lead to a drastic increase. Thus, we also record the MAE as it is more consistent over radargrams of the same image height but with different study sites and radar systems.

$$\text{MME}_s(\hat{y}_s, y_s) = \frac{c_{\text{air}}}{2|\hat{Y}_s|} \sum_{\hat{y}_s \in \hat{Y}_s} \text{Depth-Reso}^* \cdot \text{MAE}_s(\hat{y}_s, y_s) \quad (6)$$

$$\text{MME}_b(\hat{y}_b, y_b) = \frac{c_{\text{ice}}}{2|\hat{Y}_b|} \sum_{\hat{y}_b \in \hat{Y}_b} \text{Depth-Reso}^* \cdot \text{MAE}_b(\hat{y}_b, y_b) \quad (7)$$

(* Depth-Resolution after resizing the radargram.)

665 MME and MAE both describe the distance between two lines. A disadvantage is that they are not robust to outliers. As an outlier robust alternative, we also use a relaxed average precision (AP). To [normalize and standardize the relaxation](#), we count [everything below a 1 or 5 % error of the total height in pixels of the radargram as a hit \(AP-1% and AP-5%\)](#). [Choosing a relative error bound instead of fixing it to an absolute pixel value prevents the metric from changing when the radargram is resized.](#) For our chosen height of 1024 pixels, this would mean the AP-1% allows for an error of 10.24 pixels, and the AP-5% allows for an error of 51.2 pixels. Tying the average precision to the height of the radargram prevents the metric from drastically changing if future studies resize the radargrams differently. In addition, relaxing the metric alleviates the problem of uncertainties in the labels. Figure 4 shows a visualization of the employed metrics. C34 (2.13)

670 C35 (2.13)

5.2 Experimental Protocol

Since there are large differences between the subsets of the IceAnatomy dataset, we train one model for each subset, i.e., the FAU, CRISIS, and AWI subsets. The model for the AWI data is a special case, as the subset is very small. This would make the model prone to overfitting. To [counteract that mitigate this issue](#), the AWI model is first pre-trained on all three subsets of the IceAnatomy dataset and then finetuned on the AWI subset. In addition to the specialized models, we train one model on the full IceAnatomy training dataset and evaluate it on the test subsets separately to contrast it to the subset models. C36 (2.14)

For the FAU subset, we select one flight from each of the study sites as part of the test set: The third flight over Perito Moreno, the second flight over Viedma, and the flights from 2017 for JRI. The remaining flights are used for training and validation, where the validation set includes the second half of the first flight over Perito Moreno, the third section of the first flight over JRI, and the traces 5023 to 8077 for the flight over Viedma. For the CRISIS subset, we choose the TSK2, PIG4, PEN4, and PEN5 missions as the test set. This results in 7 flights in the test set, containing 3 over the Antarctic Peninsula and 4 over West Antarctica. From the remaining 25 flights, the flights from PEN3, PIG3, and GETZ1 missions are taken for the validation set. For the AWI subset, we decided not to pick an exclusive flight for testing as the differences between the collected radargrams are too big. Instead, we utilized the last 20 % of the 2014 flight over the Antarctic Peninsula and the 1999 flight over East Antarctica as our test set. For training, we picked the entirety of the 1997 flight over East Antarctica, the first 70 % of the flight over the Antarctic Peninsula, and the first 70 % of the 1999 flight over East Antarctica. The remaining 10 % of the 1999 and 2014 flights were used for validation.

690 We assess the model on the validation set after every iteration over the full training set and stop training when the AP-1% does not improve for 25 subsequent evaluations. We save the model with the highest AP-1% value on the validation set. The learning rate, a parameter that determines the strength of every network update, is set to $5e^{-4}$. As the optimizer, an algorithm

that updates the network weights based on the loss function, we use AdamW (Loshchilov and Hutter, 2019) with a weight decay of 0.05 and reduce the learning rate by a factor of 0.5 when AP-1% plateaus for ten subsequent iterations of the entire validation set. The batch size, a parameter that determines how many samples are used for every weight update, is 32 for all models. To increase variety in the data, we randomly modify the training data via data augmentations. In particular, we employ an additive Poisson noise scaled with Gaussian noise, brightness, contrast, gamma correction, and flipping horizontally.

5.3 Results

Table 2 provides quantitative results on all three subsets for the dataset-specific models and the omni model trained on the full dataset. Overall, the results are promising, with high AP-1% and AP-5% values and low MME and MAE values for most

Table 2. Overview of the performance of our presented deep learning model on the different subsets in our benchmark dataset. We distinguish the layer prediction into two classes: the ice surface (S) and the ice bottom (B). Furthermore, we split our experiments into two parts: The dataset specific models, which were trained only on a specific subset of the data, and the omni model, which was trained on the entire dataset. Note that for the AWI subset-specific model, we utilized the weights of the omni model as a starting point to stabilize training. We compare the model’s performance on the MME, MAE, AP-1%, and AP-5% as defined in Section 5.1. To contextualize the MME, we annotate the relative error to the mean measured ice thickness of the specified test set study site behind the MME. We conducted the evaluation on the test set and averaged the results over five runs to minimize statistical errors.

		Dataset specific Model				Omni Model			
	Layer	MME ↓	MAE ↓	AP-1% ↑	AP-5% ↑	MME ↓	MAE ↓	AP-1% ↑	AP-5% ↑
FAU	S	2.1 m [1.2 %]	2.0	98.8 %	100.0 %	2.4 m [1.3 %]	2.3	98.5 %	99.9 %
	B	9.1 m [4.9 %]	13.1	74.3 %	95.8 %	19.5 m [10.5 %]	27.3	68.3 %	90.5 %
CReSIS	S	23.1 m [3.1 %]	2.5	96.9 %	100.0 %	20.8 m [2.8 %]	2.2	97.9 %	100.0 %
	B	78.2 m [10.4 %]	15.2	87.9 %	94.1 %	66.5 m [8.9 %]	12.8	88.6 %	94.4 %
AWI	S	4.9 m [0.3 %]	0.7	99.3 %	100.0 %	12.0 m [0.6 %]	1.7	97.6 %	99.4 %
	B	29.3 m [1.5 %]	7.4	83.5 %	97.6 %	39.8 m [2.1 %]	10.0	75.7 %	95.6 %

combinations. Still, dataset and model-specific discrepancies exist.

5.3.1 Ice Surface Predictions

The predictions for the ice surfaces are nearly perfect for all subsets and all models. The three subset models even achieve 100 % accuracy for the AP-5%. Hence, the remaining discrepancies are likely significantly influenced by measurement inaccuracies, noise, and general model variance. Therefore, we will only consider the task of ice bottom delineation to assess model performance.

5.3.2 Ice Bottom Predictions

For the ice bottom predictions, the differences in the MME between the three subsets are more pronounced than for the MAE, which can be attributed to the different depth resolutions. The MAE difference between the FAU and CReSIS subsets is small, while the MAE on the AWI subset is substantially lower than both. The AP-1% is lower for the FAU subset than for the AWI and CReSIS subsets. Interestingly, this difference between subsets is relativized for AP-5%. This means that most incorrect predictions for FAU are in the 1 % to 5 % error range. The same is true for the AWI subset. For the CReSIS data, this effect is not as strong. Here, the AP only increases from 87.9 % for the 1 % error rate to 94.1 % for the 5 % error rate.

5.3.3 Omni Model

The omni model shows persistently higher MME and MAE values and lower AP-1% and AP-5% values for the FAU and AWI subsets than the dataset-specific models. In detail, it only achieves an MME of 19.5 m and 39.8 m and an AP-1% of 68.3 % and 75.7 %, respectively. We attribute the lower performance of the omni model to the substantial domain shift between the three subsets and the fact that the FAU and AWI subsets are significantly smaller than the CReSIS subset. For the CReSIS subset, the omni model outperforms the dataset-specific model. In particular, it achieves an MME of 66.5 m and an AP-1% of 88.6 %. These results suggest that there can be a benefit from more training data even with the domain shift. However, the domain shift makes the generalization to under-represented or new domains difficult.

5.3.4 Influence of Study Sites

Table 3. Overview of the influence of geographical and glaciological factors on the performance in detecting the ice bottom. We differentiate between the subset, the study site, and the general thermal regime. For the performance analysis, we compare the MME, MAE, AP-1%, and AP-5% as defined in Section 5.1. To contextualize the MME, we annotate the relative error to the mean measured ice thickness of the specified test set study site behind the MME. Note that for the AWI subset-specific model, we utilized the weights of the omni model as a starting point to stabilize training. We conducted the evaluation on the test set and averaged the results over five runs to minimize statistical errors. The analyzed models were the subset-specific models.

	Study Site	Main Thermal Regime	MME ↓	MAE ↓	AP-1% ↑	AP-5% ↑
FAU	Perito Moreno	Temperate	22.1 m [8.0 %]	26.3	54.9 %	91.1 %
	Viedma	Temperate	10.0 m [5.0 %]	12.0	68.5 %	96.8 %
	James Ross Island	Polythermal	3.9 m [2.7 %]	9.2	84.9 %	96.9 %
CReSIS	Antarctic Peninsula	Polythermal	31.6 m [4.5 %]	5.8	91.5 %	97.6 %
	West Antarctica	Polythermal	148.7 m [18.0 %]	29.4	82.5 %	88.8 %
AWI	Antarctic Peninsula	Polythermal	32.7 m [9.8 %]	8.1	87.3 %	96.4 %
	East Antarctica	Cold-based	27.3 m [1.0 %]	6.9	81.1 %	98.3 %

Table 3 divides the results of the subset-specific models by study site and thermal regime.

For the FAU subset, the Perito Moreno and Viedma predictions are quantitatively worse than the ones from JRI. A noticeable key difference between Perito Moreno, Viedma, and JRI is the thermal regime. The first two are temperate glaciers, while JRI contains polythermal ice. Besides the higher water content in Perito Moreno and Viedma, both are also substantially deeper than JRI in most areas. They even exhibit areas with ice too thick for the employed radar system to penetrate. They even have areas with ice thicker than the 700 m maximum penetration depth of the employed radar system. Viedma and JRI also feature several-meter-thick moraine material on the glacier surface. These rock and debris deposits are not penetrable by the wavelets and thus create radar shadows below them or substantially decrease the amount of reflected energy.

If we look at the associated radargrams, we can mostly see a relatively stable and clear prediction for JRI. On the other hand, Viedma and Perito Moreno have much stronger differences to the ground truth. Especially in deep and noisy regions, the models struggle. Figure 5 shows example traces for the three study sites of the FAU subset.

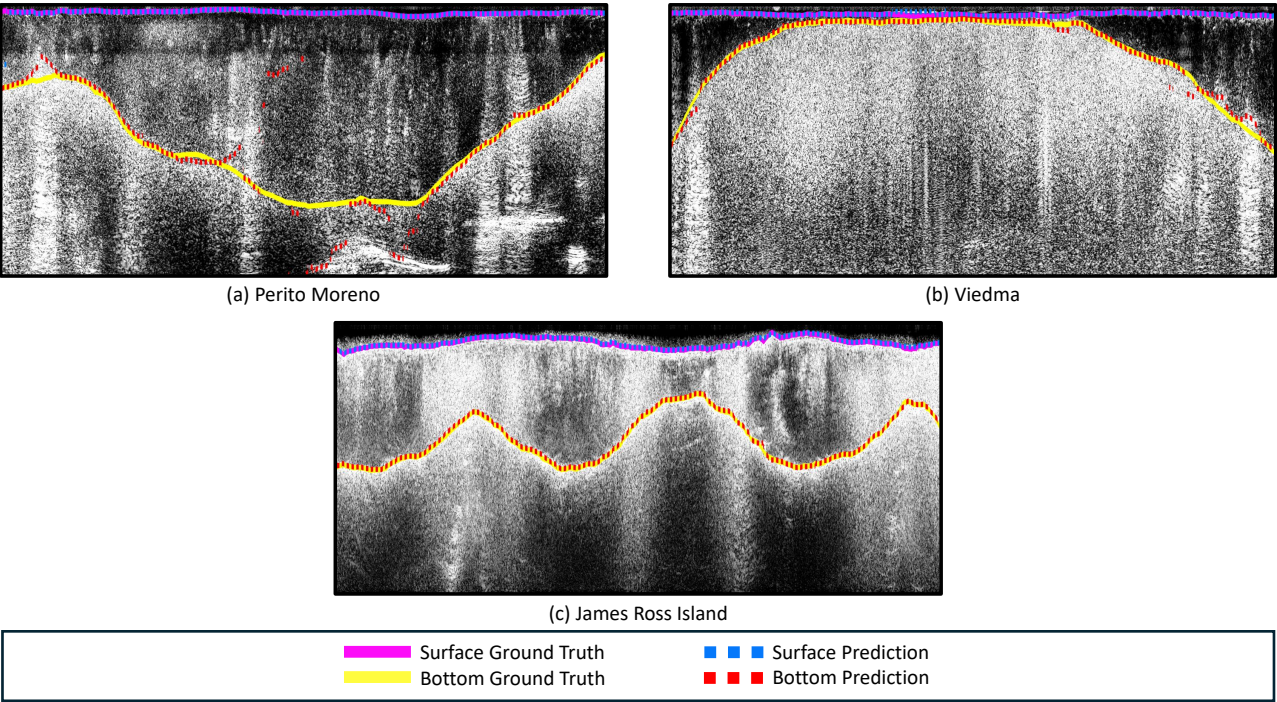


Figure 5. Visualization of the subset-specific model’s performance on the FAU subset. Figure (a) shows trace 3000-5500 of the third flight over Perito Moreno, Fig. (b) depicts traces 5000-7500 of the second flight over Viedma, and Fig. (c) presents traces 5000-7500 from the first Section of the 2017 flights over James Ross Island.

Between the Antarctic Peninsula and West Antarctica study sites of the CReSIS subset, there are strong differences in the quantitative analysis. The MME and MAE values exhibit a difference of approximately a factor of five, while the AP-1% and AP-5% are approximately 9 % apart. In the qualitative analysis, we can see that the predictions in both regions actually

follow the ground truth closely. However, sometimes the predicted ice bottom layer makes a jump and the actual ice surface is predicted to be the ice bottom. We call this “ice boundary collapse”. Examples of this phenomenon can be seen in Fig. 6.

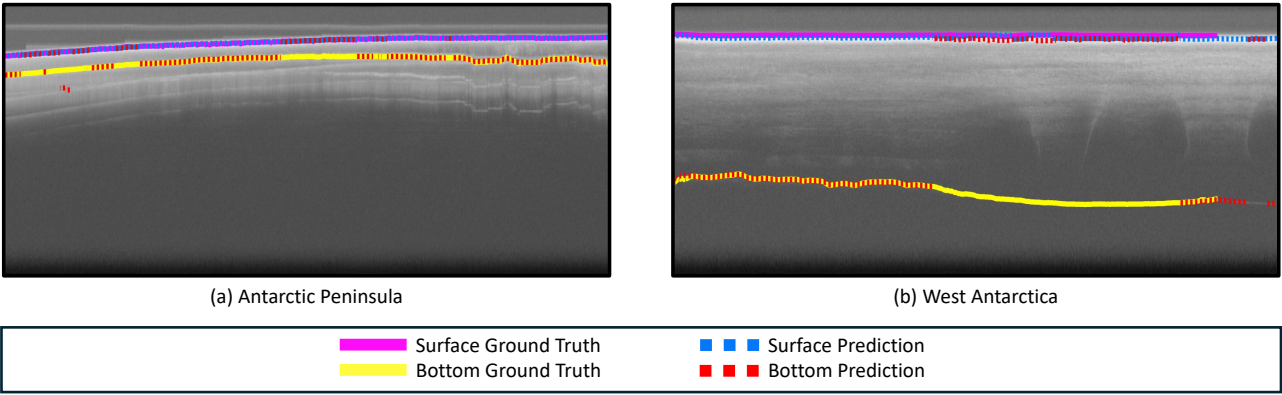


Figure 6. Visualization of the subset-specific model’s performance on the CReSIS subset. Figure (a) presents traces 2000-4500 from mission PEN4 in the Antarctic Peninsula (PEN4_01_001). Figure (b) presents traces 2000-4500 from mission TSK2 in West Antarctica (TSK2_07_003).

For the AWI subset, the results for East Antarctica are more favorable than those for the Antarctic Peninsula, with the exception of the AP-1%. This outcome is in line with the observation on the FAU subset, that the algorithm and radar system performs better for colder ice than for warmer ice. It is noteworthy that the Antarctic Peninsula exhibits a superior AP-1% compared to East Antarctica, yet a comparatively inferior AP-5%. This suggests that the inaccuracies inherent in the model’s results for East Antarctica are not large, which is further supported by the lower MAE and MME. This phenomenon is particularly evident in the qualitative analysis, where the prediction aligns closely with the ground truth in East Antarctica. Similarly, the prediction for the Antarctic Peninsula also appears to be relatively accurate, although there are occasional outliers, which contribute to the elevated MME. Example traces can be seen in Fig. 7. For the AWI subset, the results for East Antarctica are generally more favorable than those for the Antarctic Peninsula. This result is consistent with the observation on the FAU subset that the algorithm performs better for colder ice than for warmer ice. The only exception is the AP-1%, where the Antarctic Peninsula slightly outperformed East Antarctica. This result suggests that a large majority of the wrong predictions in East Antarctica are between the 1 % and 5 % interval and that our algorithm struggles to pinpoint the exact location of the ice bottom. We can confirm this behavior in the qualitative analysis, where the prediction is sometimes slightly above or below the ground truth line but follows it closely overall. Similarly, the predictions for the Antarctic Peninsula also appear to be very accurate but contain more occasional outliers. Figure 7 depicts both the predictions for East Antarctica and the Antarctic Peninsula.

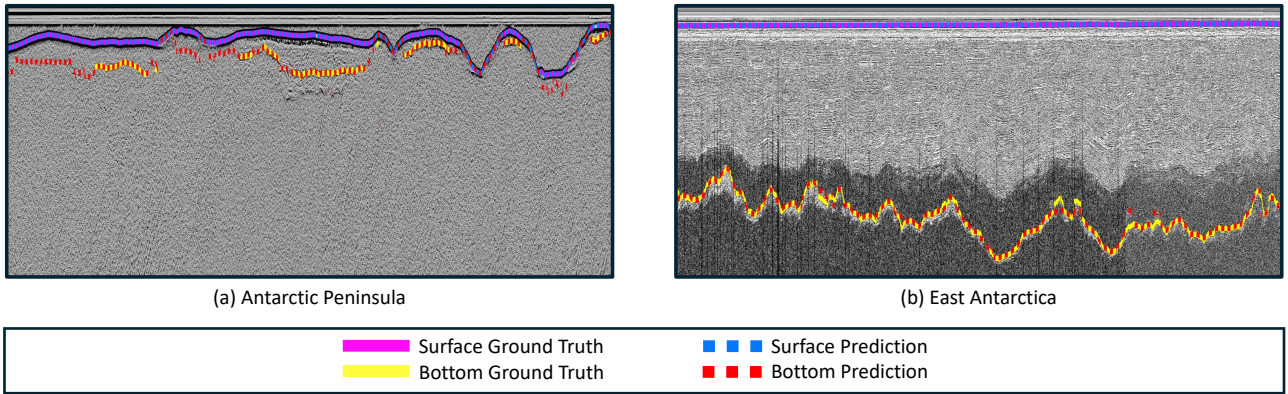


Figure 7. Visualization of the subset-specific model’s performance on the AWI subset. Figure (a) depicts traces 21000-23500 from the 2014 flight in the Antarctic Peninsula. Figure (b) presents traces 7837-9787 from the 1999 flight in East Antarctica.

Table 4. Summary of our ablation study regarding the proposed modifications to the loss function. For every variation of the loss function, we trained a subset-specific model and compared the performance based on the MME and AP-1% of the ice bottom layer. We conducted the evaluation on the test set and averaged the results over five runs to minimize statistical errors. To contextualize the MME, we annotate the relative error to the mean measured ice thickness of the specified test set study site behind the MME. Note that for the AWI subset-specific model, we utilized the weights of the omni model as a starting point to stabilize training, which was also trained with the specified loss function.

	FAU		CReSIS		AWI	
	MME ↓	AP-1% ↑	MME ↓	AP-1% ↑	MME ↓	AP-1% ↑
L_{CE}	13.9 m [7.4 %]	74.3 %	88.0 m [11.7 %]	88.6 %	29.7 m [1.6 %]	82.5 %
L_{Dist}	9.9 m [5.1 %]	72.3 %	119.5 m [15.9 %]	85.5 %	33.2 m [1.7 %]	81.8 %
$L_{CE} + L_{Dist}$	9.1 m [4.9 %]	74.3 %	78.2 m [10.4 %]	87.9 %	29.3 m [1.5 %]	83.5 %

5.3.5 Loss Function

To validate the performance of our combined loss function, we conducted an ablation study. Specifically, we trained a model using only one component of the combined loss (cross-entropy L_{CE} and distance loss L_{Dist}) and compare them to the results of the combined loss. The results are provided in Table 4. To assess the performance of our combined loss function, we conducted a small ablation study. Specifically, we evaluated two additional experiments in which we replaced the combined loss with each of its individual components: In the first setup, we trained the model with the cross-entropy loss, and in the second setup, we trained it only with the distance loss. We compare the results of these two configurations with the combined loss in Table 4.

For the FAU subset model, the distance loss improves the MME but not the AP-1%. ~~Meanwhile~~In contrast, the cross-entropy is better for the AP-1% but not for the MME. The combination of both losses results in an improved MME while AP-1% remains the same. The results for the CReSIS subset are less clear. It is evident that the distance loss alone does not enhance the MME or AP-1%. However, the combined loss demonstrates the most optimal outcome in relation to the MME, while the AP-1% is only slightly worse in comparison to CE alone. Similar to the CReSIS results, the distance loss alone does not improve the MME compared to the cross-entropy for the AWI subset. However, the combined loss ~~again delivers~~provides the best results with a higher AP-1% value.

5.4 Discussion and Outlook

One apparent influence on the quality of the ice bottom prediction is the primary thermal regime of the region. In general, the warmer the ice, the less reliable the prediction. The reason behind this is probably the influence of water on the signal, as well as the higher likelihood of a heavily crevassed surface. Temperate ice generally contains water, as most of the ice is close to or at the pressure melting point. Water absorbs the recorded signal, leading to higher noise with increased depth and strong attenuation. Hence, the model's performance naturally decreases as the associated radargrams are more challenging to interpret. Polythermal glaciers, contrary to temperate glaciers, do not exhibit ice at the pressure melting point everywhere. ~~Here, temperature is often induced into the ice due to strong frictional heat at regions of fast flow or close to the margins or the glaciers.~~Instead, elevated temperatures are usually confined to zones of fast flow driven by frictional heating or to marginal areas of the glacier. Hence, the effects are not as detrimental as for entirely temperate glaciers.

~~Another striking observation is the difference between temperate and polythermal ice regarding the AP-1%. The AP-1% of temperate ice is significantly lower than for polythermal ice. However, this difference becomes a lot smaller when comparing the AP-5%.~~Another interesting observation is the difference between temperate and polythermal ice regarding the AP-1% and AP-5%. The AP-1% of temperate ice is significantly lower than for polythermal ice. However, the AP-5% is relatively similar for both types of ice. While it is natural for the difference to decrease at higher error intervals, the change in this case is still very drastic. To put this into perspective, the Viedma and James Ross Island were 16.4 percentage points apart on the AP-1%, but on the AP-5%, only 0.1 percentage points. A possible explanation for this could lie in the meltwater at the base of the ice.

Temperate ice more commonly collects meltwater at its base than polythermal ice. Since water absorbs the signal, the exact position (AP-1%) becomes difficult to identify. However, the general position (AP-5%) is still clear because the water is only at the base. Besides the thermal regime and average depth, the presence of debris usually plays a significant role in radio-echo sounding. Interestingly, the quantitative results of JRI and Viedma indicate that the presence of debris did not play a major role in the model's performance compared to depth and thermal regime. However, we suspect that the numbers do not capture the effect of debris very well since the debris likely absorbed the signal entirely. Thus, the expert could not create ground truth labels for these parts, which makes the effect of debris on the model's performance not accurately measurable with numerical methods.

~~A notable feature of the CReSIS model predictions is ice boundary collapses.~~One of the more prominent and recurring phenomena in the CReSIS model's predictions is the collapse of ice boundaries. ~~When in doubt~~In ambiguous cases, the model

shows a bias toward predicting the ice bottom close to the or as the ice surface. One explanation could be that the CReSIS data is differentiated and thus represents only the change in amplitude. That makes it challenging to distinguish whether the peak of the ice surface and ice bottom overlap or the ice bottom is not visible. The problem gets further amplified by noise and artifacts, such as multiples. They can exhibit similar patterns as the ice bottom, making the model biased toward predicting the ice bottom as the ice surface when in doubt. **Thicker ice sheets are particularly affected by ice boundary collapse, as the radar signal returned from deeper ice is typically weaker than in shallower regions. This is due to increased attenuation and a higher chance of signal interference. Consequently, the peaks representing the ice bottom in the radargram become less distinct. As a result, the model encounters significant challenges in discerning whether a weak peak indicates the actual ice bottom or if it is just noise, and the signals from the surface and bottom overlap.**

Furthermore, we believe that the influence of the ice boundary collapse is also reflected by the quantitative analysis of the different CReSIS study sites. ~~As West Antarctica generally contains thicker ice sheets than the Antarctic Peninsula, the difference between the ice boundaries significantly increases.~~ **As West Antarctica generally contains thicker ice sheets than the Antarctic Peninsula, the average distance between the ice boundaries significantly increases.** Thus, a wrong prediction of the ice bottom as ice surface leads to a considerably higher MAE and MME for West Antarctica than the Antarctic Peninsula. However, the ice boundary collapse is likely not the only reason for this effect as the AP-1% and AP-5% are also lower for West Antarctica than the Antarctica Peninsula. Hence, thicker ice sheets might be naturally more challenging.

Nonetheless, future research should address ice boundary collapses as they tremendously affect performance. Larger contexts, **additional post-processing steps,** or recurrent neural networks could help stabilize the predictions as they incorporate more information. Another interesting problem to explore is the performance drop from subset-specific models to the omni model. Our results indicate that the domain shift between the subsets is too prominent for a simple omni model to catch up on all subset-characteristic features. Hence, models cannot utilize the full benefits of a larger dataset when they are recorded and processed differently. ~~In particular, domain shift techniques could help with this challenge.~~ **In particular, domain shift techniques could help with this challenge, but also more advanced regularization techniques, e. g., spatial dropout, could prevent the model from focusing too much on a single domain (Tompson et al., 2015). In appendix 13, we show that a uniform sampling strategy can also help mitigate the domain shift.**

~~We believe that our framework is the first step towards a potential fully automated generation of ice thickness maps based on RES data. Our presented work could lay the foundation for validating survey data while in the field.~~ **We believe that our framework is a significant step towards a potential fully automated generation of ice thickness maps based on RES data and that our work represents an important advancement toward validating survey data in the field.**

6 Conclusions

This paper presents the first benchmark framework for delineating the ice boundary in RES data. The included dataset “IceAnatomy” contains hundreds of kilometers of processed, labeled, and georeferenced RES data from three different sources (FAU, CReSIS, AWI). Since all sources employ a different radar system and processing methods, “IceAnatomy” offers a wide

830 range of varying amplitude spectrums, depth resolutions, and width resolutions, making it applicable to a multitude of settings. Furthermore, it also features different geographical factors, such as study sites and thermal regimes, allowing for in-depth analysis of the models and their behavior in different geographical scenarios.

To fairly compare different models in the future, we provide an official train and test split for each source of the dataset. This enables the development of not only an omni model trained on the entire dataset but also specialized subset-specific models
835 on one of the three sources. We trained and evaluated a baseline model for each of these scenarios. In our experiments, the subset-specific models provide the most promising results with MMEs of 2.1 m [1.2%], 23.1 m [3.1 %], and 4.9 m [0.3 %] for the ice surface and 9.1 m [4.9 %], 78.2 m [10.4 %], and 29.3 m [1.5 %] for the ice bottom depending on the source.

~~With this benchmark framework, we hope to encourage other scientists to start working in this challenging and important research area.~~ Previous work has already demonstrated the effectiveness of automatic approaches for ice boundary extraction
840 but lacked a common method for accurately comparing models. With this benchmark framework, we hope to address this issue by unifying and standardizing both training and evaluation schemes. We hope that this benchmark dataset will encourage more scientists to engage in this challenging and important research area. Deep learning models that extract the ice boundary ^{CS0 (2.15)}
can greatly speed up the processing of RES data. As a result, ~~ice depth~~ the ice thickness and, consequently, the subglacial ^{CS1 (2.1)}
topography can be determined more quickly after a field survey.

845 *Code and data availability.* The dataset is available at <https://zenodo.org/records/14036897> (Dreier et al., 2024) and the implementation at <https://doi.org/10.5281/zenodo.14038570> (Dreier, 2024).

Appendix A: Additional Hyperparameters

This section gives an overview of the hyperparameters in our employed U-Net from Chapter 4.2. The input dimension of our U-net is (1024,512,1) (H,W,1), which then gets scaled according to the depth level of the encoder or decoder. Inside
850 the network, we down- and upsample our feature map five times each while scaling the feature dimension according to the depth-level-dependant value of [8, 16, 32, 64, 64, 128]. To reduce the risk of overfitting, we also utilize dropout layers inside the ResBlocks with a probability of 10%. For the loss function, we employed our proposed combined loss function. Since the numerical value of the distance loss is significantly higher than that of the classification loss, we had to weigh the individual components. In detail, we chose the weights $w_{s_class} = 0.5$, $w_{b_class} = 1.0$, $w_{s_dist} = 0.05$, and $w_{b_dist} = 0.1$ as they performed
855 the best in preliminary experiments.

Appendix B: ResBlock Design

To provide a better understanding of the network architecture, this section examines one of its core components: the ResBlock from (Esser et al., 2020). Its structure, shown in Fig. B1, comprises several components. First, it starts with a group normalization layer (Wu and He, 2018) that normalizes the data in groups of channels to increase stability during training. Next, a swish

860 activation (Ramachandran et al., 2017) function adds nonlinearity to the ResBlock so the network can learn more complex patterns. The activation is followed by a two-dimensional convolution layer that processes and combines the visual features by applying convolutional operations. This is followed by another group normalization and swish activation function before a regular dropout layer (Hinton et al., 2012) is applied. The dropout layer randomly withholds information during training to improve generalization and prevent the model from overfitting – a process in which the model develops a strong bias towards

865 the training data. After the dropout layer, another two-dimensional convolutional layer is applied. Finally, a residual connection (He et al., 2016), a shortcut from the start of the ResBlock to the end through a convolution layer, is added to the output of this sequence of layers to improve the gradient flow in the network. A11 (1.7, 2.4)

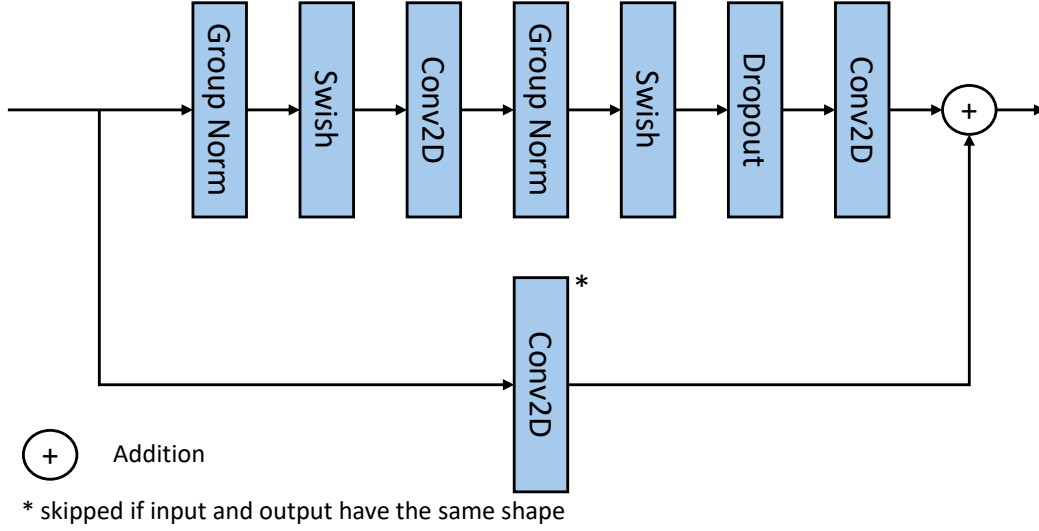


Figure B1. Structure of the residual block employed in our deep learning model. The arrangement is based on the design of Esser et al. (2020)

Appendix C: Loss Function Details

The formulas of the classification and distance-based loss are as follows:

$$870 \quad L_{CE} = - \sum_{c \in C} x_c (1 - \epsilon_c) \log(p(x_c)) + \frac{\epsilon_c (1 - x_c) \log(p(x_c))}{|C|} \quad (C1)$$

$$L_{class} = \frac{w_{s_class}}{|\hat{Y}_s|} \sum_{\hat{y}_s \in \hat{Y}_s} L_{CE}(\hat{y}_s) + \frac{w_{b_class}}{|\hat{Y}_b|} \sum_{\hat{y}_b \in \hat{Y}_b} L_{CE}(\hat{y}_b) \quad (C2)$$

$$L_{dist} = \frac{w_{s_dist}}{|\hat{Y}_s|} \sum_{\hat{y}_s \in \hat{Y}_s} \langle d(y_s), \sigma(\hat{y}_s) \rangle + \frac{w_{b_dist}}{|\hat{Y}_b|} \sum_{\hat{y}_b \in \hat{Y}_b} \langle d(y_b), \sigma(\hat{y}_b) \rangle \quad (C3)$$

w_{s_class} , w_{b_class} , w_{s_dist} , and w_{b_dist} are the respective weights for a weighted combination of the single loss parts, ϵ_c is the smoothing factor, C specifies the column, $\langle \rangle$ is the dot product, σ is the softmax function that converts the model's outputs into

probabilities, $|\cdot|$ is the cardinality of a set, and d is the function that creates a vector filled with the column-wise distance map given the respective column of the label. A12 (2.4)

Appendix D: Additional Experiments

Since the three subsets of IceAnatomy differ in size, we also investigate whether a uniform sampling strategy, where samples are drawn equally from each subset, could help the Omni-Model achieve the performance of the domain-specific models on the AWI and FAU subsets. From our results in Table D1, we can see that a uniform sampling strategy does lead to improvement for the AWI and FAU subsets. In the case of the AWI subset, the omni model even outperforms the domain-specific model. However, in the case of the FAU subset, we are still below the domain-specific model. We reason that the domain of the AWI and CReSIS subsets are significantly closer than the FAU subset as these two subsets contain differentiated radargrams. We, therefore, believe that domain shift remains an important area for future research. In addition to the uniform sampling, we also A13 (1.3, 2.9)

Table D1. Overview of the performance of our Omni Model with uniform sampling. We distinguish the layer prediction into two classes: the ice surface (S) and the ice bottom (B). We compare the model’s performance on the MME, MAE, AP-1%, and AP-5% as defined in Section 5.1. To contextualize the MME, we annotate the relative error to the mean measured ice thickness of the specified test set study site behind the MME. We conducted the evaluation on the test set and averaged the results over five runs to minimize statistical errors.

		Omni Model			
	Layer	MME ↓	MAE ↓	AP-1% ↑	AP-5% ↑
FAU	S	2.0m [1.1 %]	1.9	99.3 %	100.0 %
	B	14.0m [7.6 %]	19.0	74.1 %	94.1 %
CReSIS	S	23.1m [3.1 %]	2.5	97.2 %	100.0 %
	B	75.0m [10.0 %]	14.6	87.7 %	93.9 %
AWI	S	3.8m [0.2 %]	0.5	99.7, %	100.0 %
	B	23.9m [1.3 %]	6.0	86.1 %	98.3 %

investigated how different hyperparameter setups regarding learning rate and regularization would affect the benchmark model. From the results in Table D2 and D3, we can see that different hyperparameter setups favor different subsets of IceAnatomy. However, there seems to be no universal optimal setup. A14 (2.12)

Author contributions. **Marcel Dreier:** Conceptualization, Methodology, Software, Experiments, Project administration, Data Processing, Visualization, Writing - Original draft preparation. **Moritz Koch:** Data Collection, Data Processing, Data curating, Visualization, Writing - Original draft preparation. **Nora Gourmelon:** Conceptualization, Writing - Original draft preparation. **Norbert Blindow:** Data Collection, Data Processing, Writing - review & editing. **Daniel Steinhage:** Data Collection, Data Processing, Writing - review & editing. **Fei Wu:**

Table D2. Overview of the performance of our Omni Model with different learning rates and uniform sampling. We distinguish the layer prediction into two classes: the ice surface (S) and the ice bottom (B). We compare the model’s performance on the MME and AP-1% as defined in Section 5.1. To contextualize the MME, we annotate the relative error to the mean measured ice thickness of the specified test set study site behind the MME. We conducted the evaluation on the test set and averaged the results over three runs to minimize statistical errors. Note that for $lr = 0.005$ we averaged over five runs, as we had those values from previous experiments.

		$lr = 0.0001$		$lr = 0.0005$		$lr = 0.001$	
	Layer	MME ↓	AP-1% ↑	MME ↓	AP-1% ↑	MME ↓	AP-1% ↑
FAU	S	2.1m [1.1%]	99.0%	2.0m [1.1%]	99.3%	2.1m [1.1%]	99.1%
	B	14.1m [7.6%]	73.9%	14.0m [7.6%]	74.1%	14.3m [7.7%]	74.1%
CReSIS	S	26.2m [3.5%]	96.7%	23.1m [3.1%]	97.2%	21.9m [2.9%]	97.6%
	B	105.4m [14.0%]	87.2%	75.0m [10.0%]	87.7%	94.9m [12.6%]	87.9%
AWI	S	4.4m [0.2%]	99.5%	3.8m [0.2%]	99.7%	3.7m [0.2%]	99.6%
	B	26.9m [1.4%]	86.2%	23.9m [1.3%]	86.1%	21.5m [1.1%]	87.8%

Table D3. Overview of the performance of our Omni Model with different learning rates and uniform sampling. We distinguish the layer prediction into two classes: the ice surface (S) and the ice bottom (B). We compare the model’s performance on the MME and AP-1% as defined in Section 5.1. To contextualize the MME, we annotate the relative error to the mean measured ice thickness of the specified test set study site behind the MME. We conducted the evaluation on the test set and averaged the results over three runs to minimize statistical errors. Note that for $dropout = 0.1$ we averaged over five runs, as we had those values from previous experiments.

		$dropout = 0.0$		$dropout = 0.1$		$dropout = 0.2$	
	Layer	MME ↓	AP-1% ↑	MME ↓	AP-1% ↑	MME ↓	AP-1% ↑
FAU	S	2.0m [1.1%]	99.2%	2.0m [1.1%]	99.3%	2.0m [1.1%]	99.3%
	B	10.2m [5.5%]	74.2%	14.0m [7.6%]	74.1%	14.3m [7.7%]	73.5%
CReSIS	S	22.5m [3.0%]	97.0%	23.1m [3.1%]	97.2%	21.5m [2.9%]	97.6%
	B	79.0m [10.5%]	87.8%	75.0m [10.0%]	87.7%	69.2m [9.2%]	88.5%
AWI	S	7.3m [0.4%]	99.0%	3.8m [0.2%]	99.7%	3.8m [0.2%]	99.5%
	B	26.0m [1.4%]	85.7%	23.9m [1.3%]	86.1%	24.5m [1.3%]	86.6%

Writing - review & editing. **Thorsten Seehaus**: Supervision, Writing - review & editing. **Matthias Braun**: Supervision, Writing - review & editing. **Andreas Maier**: Supervision, Writing – review & editing. **Vincent Christlein**: Supervision, Validation, Writing - review & editing.

Competing interests. The authors declare none of the authors have any competing interests.

895 *Acknowledgements.* This research was funded by the Bayerisches Staatsministerium für Wissenschaft und Kunst within the Elite Network
Bavaria with the Int. Doct. Program “Measuring and Modelling Mountain Glaciers in a Changing Climate” (IDP M3OCCA)), as well as the
German Research Foundation (DFG) project “Large-scale Automatic Calving Front Segmentation and Frontal Ablation Analysis of Arctic
Glaciers using Synthetic-Aperture Radar Image Sequences (LASSI)” and the DFG project “Ice thickness, remote sensing and sensitivity
experiments using ice-flow modelling for major outlet glaciers of the Southern Patagonian Icefield”(ITERATE) grant DFG BR 2105/29-
900 1/FU 1032/12-1. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High
Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR projects
b110dc and b194dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the DFG
– 440719683. We acknowledge the use of data and data products from CReSIS generated with support from the University of Kansas, NASA
Operation IceBridge grant NNX16AH54G, NSF grants ACI-1443054, OPP-1739003, and IIS-1838230, Lilly Endowment Incorporated,
905 and Indiana METACyt Initiative. Furthermore, we also thank the support of Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und
Meeresforschung. (2016). Polar aircraft Polar5 and Polar6 operated by the Alfred Wegener Institute. Journal of large-scale research facilities
JLSRF, 2 (0), 87. doi: 10.17815/jlsrf-2-153. The authors would like to thank Aspen Technology, Inc. for providing licenses in the scope of
the Aspen Technology, Inc. Academic Program.

In order to improve the legibility of the manuscript, the authors have used ChatGPT (<https://chatgpt.com/>) and DeepL Write (<https://www.deepl.com/en/write>) to look for alternative phrases. The output of this service was reviewed and edited by the authors as needed. The
910 authors take full responsibility for the content of the presented manuscript.

References

- Allen, C., Shi, L., Hale, R., Leuschen, C., Paden, J., Pazer, B., Arnold, E., Blake, W., Rodriguez-Morales, F., Ledford, J., and Seguin, S.: Antarctic ice depthsounding radar instrumentation for the NASA DC-8, *IEEE Aerospace and Electronic Systems Magazine*, 27, 4–20, 915 <https://doi.org/10.1109/MAES.2012.6196253>, 2012a.
- Allen, C., Shi, L., Hale, R., Leuschen, C., Paden, J., Pazer, B., Arnold, E., Blake, W., Rodriguez-Morales, F., Ledford, J., et al.: Antarctic ice depthsounding radar instrumentation for the NASA DC-8, *IEEE Aerospace and Electronic Systems Magazine*, 27, 4–20, 2012b.
- An, J., Huang, S., Chen, X., Xu, T., and Bai, Z.: Research progress in geophysical exploration of the Antarctic ice sheet, *Earthquake Research Advances*, 3, 100 203, 2023.
- 920 Aniya, M.: Recent glacier variations of the Hielos Patagónicos, South America, and their contribution to sea-level change, *Arctic, Antarctic, and Alpine Research*, 31, 165–173, 1999.
- Ayala, A., Farías-Barahona, D., Huss, M., Pellicciotti, F., McPhee, J., and Farinotti, D.: Glacier runoff variations since 1955 in the Maipo River basin, in the semiarid Andes of central Chile, *The Cryosphere*, 14, 2005–2027, <https://doi.org/10.5194/tc-14-2005-2020>, 2020.
- Berger, V., Xu, M., Chu, S., Crandall, D., Paden, J., and Fox, G. C.: Automated Tracking of 2D and 3D Ice Radar Imagery Us- 925 ing Viterbi and TRW-S, in: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4162–4165, <https://doi.org/10.1109/IGARSS.2018.8519411>, 2018.
- Blankenship, D. D., Roberts, J. L., Greenbaum, J. S., Young, D. A., Van Ommen, T., Le Meur, E., and Beem, L. H.: *EAGLE/ICECAP II RADARGRAMS*, 2018.
- Blindow, N., Salat, C., Gundelach, V., Buschmann, U., and Kahnt, W.: Performance and calibration of the helicopter GPR 930 system BGR-P30, in: *2011 6th International Workshop on Advanced Ground Penetrating Radar (IWAGPR)*, pp. 1–5, <https://doi.org/10.1109/IWAGPR.2011.5963896>, 2011.
- Blindow, N., Salat, C., and Casassa, G.: Airborne GPR sounding of deep temperate glaciers — Examples from the Northern Patagonian Icefield, in: *2012 14th International Conference on Ground Penetrating Radar (GPR)*, pp. 664–669, <https://doi.org/10.1109/ICGPR.2012.6254945>, 2012.
- 935 Bogorodsky, V. V., Chebotareva, V., Bentley, C. R., and Gudmandsen, P. E.: *Radioglaciology, Glaciology and Quaternary Geology*, Springer Netherlands, softcover reprint of the hardcover 1st edition 1985 edn., ISBN 9789400952751, <https://doi.org/10.1007/978-94-009-5275-1>, 2012.
- Cai, Y., Ma, J., Li, H., and Hu, S.: Automatic Classification of Ice Sheet Subsurface Targets in Radar Sounder Data Based on the Capsule Network, in: *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition, ICCPR '19*, pp. 199–204, 940 Association for Computing Machinery, New York, NY, USA, ISBN 9781450376570, <https://doi.org/10.1145/3373509.3373585>, 2019.
- Cai, Y., Hu, S., Lang, S., Guo, Y., and Liu, J.: End-to-End Classification Network for Ice Sheet Subsurface Targets in Radar Imagery, *Applied Sciences*, 10, <https://doi.org/10.3390/app10072501>, 2020.
- Cai, Y., Wan, F., Hu, S., and Lang, S.: Accurate prediction of ice surface and bottom boundary based on multi-scale feature fusion network, *Applied Intelligence*, 52, 16 370–16 381, <https://doi.org/10.1007/s10489-022-03362-1>, 2022.
- 945 Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, *IEEE T. Pattern. Anal.*, 40, 834–848, <https://doi.org/10.1109/TPAMI.2017.2699184>, 2018.
- Consortium, R.: *Randolph Glacier Inventory - A Dataset of Global Glacier Outlines, Version 6*, <https://doi.org/10.7265/4m1f-gd79>, 2017.

- Corr, H.: Airborne radar bed elevation along flow lines covering the Evans, and Rutford Ice Streams, and ice rises in the Ronne Ice Shelf (2006/07), 2020.
- Corr, H., Ferraccioli, F., Jordan, T., and Robinson, C.: Antarctica's Gamburtsev Province (AGAP) Project - Radio-echo sounding data (2007-2009), 2020.
- Crandall, D. J., Fox, G. C., and Paden, J. D.: Layer-finding in radar echograms using probabilistic graphical models, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 1530–1533, 2012.
- 955 CReSIS: Radar Depth Sounder Data, <https://data.cresis.ku.edu/data/rds/>.
- CReSIS: Radar Depth Sounder Data-2009 Antarctica DC8, <http://data.cresis.ku.edu/>, 2024a.
- CReSIS: Radar Depth Sounder, https://data.cresis.ku.edu/data/rds/rds_readme.pdf, 2024b.
- Dawson, E. J., Schroeder, D. M., Chu, W., Mantelli, E., and Seroussi, H.: Ice mass loss sensitivity to the Antarctic ice sheet basal thermal state, *Nature Communications*, 13, 4957, 2022.
- 960 Dong, S., Tang, X., Guo, J., Fu, L., Chen, X., and Sun, B.: EisNet: Extracting Bedrock and Internal Layers From Radiostratigraphy of Ice Sheets With Machine Learning, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–12, <https://doi.org/10.1109/TGRS.2021.3136648>, 2022.
- Donini, E., Bovolo, F., and Bruzzone, L.: A Deep Learning Architecture for Semantic Segmentation of Radar Sounder Data, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14, <https://doi.org/10.1109/TGRS.2021.3125773>, 2022.
- 965 Dreier, M.: Implementation of Ice Anatomy: A Benchmark Dataset and Methodology for Automatic Ice Boundary Extraction from Radio-Echo Sounding Data, <https://doi.org/10.5281/zenodo.14038570>, 2024.
- Dreier, M., Koch, M., Gourmelon, N., Blindow, N., Steinhage, D., Wu, F., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Ice Anatomy: A Benchmark Dataset and Methodology for Automatic Ice Boundary Extraction from Radio- Echo Sounding Data, <https://doi.org/10.5281/zenodo.14036897>, 2024.
- 970 Drewry, D., Jordan, S., and Jankowski, E.: Measured properties of the Antarctic ice sheet: surface configuration, ice thickness, volume and bedrock characteristics, *Annals of Glaciology*, 3, 83–91, 1982.
- Esser, P., Rombach, R., and Ommer, B.: Taming Transformers for High-Resolution Image Synthesis, 2020.
- Fahnestock, M., Abdalati, W., Luo, S., and Gogineni, S.: Internal layer tracing and age-depth-accumulation relationships for the northern Greenland ice sheet, *Journal of Geophysical Research: Atmospheres*, 106, 33 789–33 797, 2001.
- 975 Farinotti, D., Brinkerhoff, D. J., Clarke, G. K. C., Fürst, J. J., Frey, H., Gantayat, P., Gillet-Chaulet, F., Girard, C., Huss, M., Leclercq, P. W., Linsbauer, A., Machguth, H., Martin, C., Maussion, F., Morlighem, M., Mosbeux, C., Pandit, A., Portmann, A., Rabatel, A., Ramsankaran, R., Reerink, T. J., Sanchez, O., Stentoft, P. A., Singh Kumari, S., van Pelt, W. J. J., Anderson, B., Benham, T., Binder, D., Dowdeswell, J. A., Fischer, A., Helfricht, K., Kutuzov, S., Lavrentiev, I., McNabb, R., Gudmundsson, G. H., Li, H., and Andreassen, L. M.: How accurate are estimates of glacier ice thickness? Results from ITMIX, the Ice Thickness Models Intercomparison eXperiment, *The Cryosphere*, 11, 949–970, <https://doi.org/10.5194/tc-11-949-2017>, 2017.
- 980 Freeman, G. J., Bovik, A. C., and Holt, J. W.: Automated detection of near surface Martian ice layers in orbital radar data, in: 2010 IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI), pp. 117–120, IEEE, 2010.
- García, M. H., Donini, E., and Bovolo, F.: Automatic Segmentation of Ice Shelves with Deep Learning, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pp. 4833–4836, <https://doi.org/10.1109/IGARSS47720.2021.9553610>, 2021a.

- 985 García, M. H., Donini, E., and Bovolo, F.: Transfer learning for the semantic segmentation of cryosphere radargrams, in: Image and Signal Processing for Remote Sensing XXVII, edited by Lorenzo Bruzzone and Francesca Bovolo, vol. 11862, p. 118620T, SPIE, <https://doi.org/10.1117/12.2600237>, 2021b.
- García, M. H., Donini, E., and Bovolo, F.: A Weakly Supervised Transfer Learning Approach for Radar Sounder Data Segmentation, IEEE Transactions on Geoscience and Remote Sensing, 61, 1–18, <https://doi.org/10.1109/TGRS.2023.3252939>, 2023.
- 990 Ghosh, R. and Bovolo, F.: TransSounder: A Hybrid TransUNet-TransFuse Architectural Framework for Semantic Segmentation of Radar Sounder Data, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–13, <https://doi.org/10.1109/TGRS.2022.3180761>, 2022.
- Gifford, C. M., Finyom, G., Jefferson, M., Reid, M., Akers, E. L., and Agah, A.: Automated Polar Ice Thickness Estimation From Radar Imagery, IEEE Transactions on Image Processing, 19, 2456–2469, <https://doi.org/10.1109/TIP.2010.2048509>, 2010.
- Haeberli, W., Hoelzle, M., Paul, F., and Zemp, M.: Integrated monitoring of mountain glaciers as key indicators of global climate change: the European Alps, ANN GLACIOL, 46, 150–160, <https://doi.org/10.3189/172756407782871512>, 2007.
- 995 He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- He, Y., Carass, A., Liu, Y., Jedynek, B. M., Solomon, S. D., Saidha, S., Calabresi, P. A., and Prince, J. L.: Fully Convolutional Boundary Regression for Retina OCT Segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, edited by Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A., pp. 120–128, Springer International Publishing, Cham, ISBN 978-3-030-32239-7, 2019.
- 1000 Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R.: Improving neural networks by preventing co-adaptation of feature detectors, <https://arxiv.org/abs/1207.0580>, 2012.
- Hugonnet, R., McNabb, R., Berthier, E., Menounos, B., Nuth, C., Girod, L., Farinotti, D., Huss, M., Dussaillant, I., Brun, F., et al.: Accelerated global glacier mass loss in the early twenty-first century, Nature, 592, 726–731, 2021.
- 1005 Ibikunle, O., Paden, J., Rahnemounfar, M., Crandall, D., and Yari, M.: Snow Radar Layer Tracking Using Iterative Neural Network Approach, in: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, pp. 2960–2963, <https://doi.org/10.1109/IGARSS39084.2020.9323957>, 2020.
- Ilisei, A.-M. and Bruzzone, L.: A Model-Based Technique for the Automatic Detection of Earth Continental Ice Subsurface Targets in Radar Sounder Data, IEEE GEOSCI REMOTE S, 11, 1911–1915, <https://doi.org/10.1109/LGRS.2014.2313858>, 2014.
- 1010 Ilisei, A.-M. and Bruzzone, L.: A System for the Automatic Classification of Ice Sheet Subsurface Targets in Radar Sounder Data, IEEE T GEOSCI REMOTE, 53, 3260–3277, <https://doi.org/10.1109/TGRS.2014.2372818>, 2015.
- IPCC: Climate Change 2013: The Physical Science Basis, Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK and New York, NY, USA, 2013.
- 1015 Jebeli, A., Tama, B. A., Janeja, V. P., Holschuh, N., Jensen, C., Morlighem, M., MacGregor, J. A., and Fahnestock, M. A.: TSSA: Two-Step Semi-Supervised Annotation for Radargrams on the Greenland Ice Sheet, in: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium, pp. 56–59, IEEE, 2023a.
- Jebeli, A., Tama, B. A., Purushotham, S., and Janeja, V. P.: Tracing Englacial Layers in Radargram via Semi-supervised Method: A Preliminary Result, in: Proceedings of the AAAI Symposium Series, vol. 2, pp. 85–88, 2023b.
- 1020 Johari, G. P. and Charette, P.: The permittivity and attenuation in polycrystalline and single-crystal ice Ih at 35 and 60 MHz, Journal of Glaciology, 14, 293–303, 1975.

- Kamangir, H., Rahnemoonfar, M., Dobbs, D., Paden, J., and Fox, G.: Deep Hybrid Wavelet Network for Ice Boundary Detection in Radra Imagery, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 3449–3452, <https://doi.org/10.1109/IGARSS.2018.8518617>, 2018.
- 1025 Karlsson, N. B., Dahl-Jensen, D., Prasad Gogineni, S., and Paden, J. D.: Tracing the depth of the Holocene ice in North Greenland from radio-echo sounding data, *ANN GLACIOL*, 54, 44–50, <https://doi.org/10.3189/2013AoG64A057>, 2013.
- Kohler, J., Moore, J., Kennett, M., Engeset, R., and Elvehøy, H.: Using ground-penetrating radar to image previous years’ summer surfaces for mass-balance measurements, *ANN GLACIOL*, 24, 355–360, <https://doi.org/10.3189/S0260305500012441>, 1997.
- Lee, S., Mitchell, J., Crandall, D. J., and Fox, G. C.: Estimating bedrock and surface layer boundaries and confidence intervals in ice sheet radar imagery using MCMC, in: 2014 IEEE International Conference on Image Processing (ICIP), pp. 111–115, <https://doi.org/10.1109/ICIP.2014.7025021>, 2014.
- 1030 Lippl, S., Blindow, N., Fürst, J. J., Marinsek, S., Seehaus, T. C., and Braun, M. H.: Uncertainty assessment of ice discharge using GPR-derived ice thickness from Gourdon Glacier, Antarctic Peninsula, *Geosciences*, 10, 12, 2019.
- Liu, Y., Li, H., Huang, M., Chen, D., and Zhao, B.: Ice Crevasse Detection with Ground Penetrating Radar using Faster R-CNN, in: 2020 15th IEEE International Conference on Signal Processing (ICSP), vol. 1, pp. 596–599, <https://doi.org/10.1109/ICSP48669.2020.9321072>, 2020.
- 1035 Liu-Schiaffini, M., Ng, G., Grima, C., and Young, D.: Ice Thickness From Deep Learning and Conditional Random Fields: Application to Ice-Penetrating Radar Data With Radiometric Validation, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14, <https://doi.org/10.1109/TGRS.2022.3214147>, 2022a.
- 1040 Liu-Schiaffini, M., Ng, G., Grima, C., and Young, D.: Ice thickness from deep learning and conditional random fields: application to ice-penetrating radar data with radiometric validation, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14, 2022b.
- Lohofener, A.: Design and development of a multi-channel radar depth sounder, Ph.D. thesis, United States – Kansas, ISBN 978-0-542-91978-7, <https://www.proquest.com/dissertations-theses/design-development-multi-channel-radar-depth/docview/305319328/se-2?accountid=10755>, copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works., 1045 2006.
- Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, in: International Conference on Learning Representations, <https://openreview.net/forum?id=Bkg6RiCqY7>, 2019.
- Lythe, M. B. and Vaughan, D. G.: BEDMAP: A new ice thickness and subglacial topographic model of Antarctica, *Journal of Geophysical Research: Solid Earth*, 106, 11 335–11 351, 2001.
- 1050 Macelloni, G., Leduc-Leballeur, M., Montomoli, F., Brogioni, M., Ritz, C., and Picard, G.: On the retrieval of internal temperature of Antarctica Ice Sheet by using SMOS observations, *Remote Sensing of Environment*, 233, 111 405, 2019.
- Marzeion, B., Kaser, G., Maussion, F., and Champollion, N.: Limited influence of climate change mitigation on short-term glacier mass loss, *Nature Climate Change*, 8, 305–308, 2018.
- Matsuoka, K., Skoglund, A., Roth, G., de Pomereu, J., Griffiths, H., Headland, R., Herried, B., Katsumata, K., Le Brocq, A., Licht, K., 1055 et al.: Quantarctica, an integrated mapping environment for Antarctica, the Southern Ocean, and sub-Antarctic islands, *Environmental Modelling & Software*, 140, 105 015, 2021.
- Mitchell, J. E., Crandall, D. J., Fox, G. C., Rahnemoonfar, M., and Paden, J. D.: A semi-automatic approach for estimating bedrock and surface layers from multichannel coherent radar depth sounder imagery, in: *Image and Signal Processing for Remote Sensing XIX*, edited by Bruzzone, L., vol. 8892, p. 88921E, SPIE, <https://doi.org/10.1117/12.2028992>, 2013.

- 1060 Moqadam, H. and Eisen, O.: Review article: Feature tracing in radio-echo sounding products of terrestrial ice sheets and planetary bodies, *EGU*sphere, 2024, 1–58, <https://doi.org/10.5194/egusphere-2024-1674>, 2024.
- Moqadam, H., Steinhage, D., Wilhelm, A., and Eisen, O.: Going deeper with deep learning: automatically tracing internal reflection horizons in ice sheets, *Authorea Preprints*, 2024.
- Morris, E. and Vaughan, D.: Snow surface temperatures in West Antarctica, *Antarctic Science*, 6, 529–535, 1994.
- 1065 Nixdorf, U., Steinhage, D., Meyer, U., Hempel, L., Jenett, M., Wachs, P., and Miller, H.: The newly developed airborne radio-echo sounding system of the AWI as a glaciological tool, *Annals of Glaciology*, 29, 231–238, 1999.
- Park, I.-W., Jin, E. K., Morlighem, M., and Lee, K.-K.: Impact of boundary conditions on the modeled thermal regime of the Antarctic ice sheet, *The Cryosphere*, 18, 1139–1155, 2024.
- Rahnmooonfar, M., Fox, G. C., Yari, M., and Paden, J.: Automatic Ice Surface and Bottom Boundaries Estimation in
1070 Radar Imagery Based on Level-Set Approach, *IEEE Transactions on Geoscience and Remote Sensing*, 55, 5115–5122, <https://doi.org/10.1109/TGRS.2017.2702200>, 2017a.
- Rahnmooonfar, M., Habashi, A. A., Paden, J., and Fox, G. C.: Automatic Ice thickness estimation in radar imagery based on charged particles concept, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 3743–3746, <https://doi.org/10.1109/IGARSS.2017.8127813>, 2017b.
- 1075 Rahnmooonfar, M., Johnson, J., and Paden, J.: Ai radar sensor: Creating radar depth sounder images based on generative adversarial network, *Sensors*, 19, 5479, 2019.
- Rahnmooonfar, M., Yari, M., Paden, J., Koenig, L., and Ibikunle, O.: Deep multi-scale learning for automatic tracking of internal layers of ice in radar data, *J GLACIOL*, 67, 39–48, <https://doi.org/10.1017/jog.2020.80>, 2021.
- Ramachandran, P., Zoph, B., and Le, Q. V.: Searching for activation functions, *arXiv preprint arXiv:1710.05941*, 2017.
- 1080 Rignot, E., Mouginot, J., and Scheuchl, B.: Ice flow of the Antarctic ice sheet, *Science*, 333, 1427–1430, 2011.
- Rodriguez-Morales, F., Gogineni, S., Leuschen, C. J., Paden, J. D., Li, J., Lewis, C. C., Panzer, B., Gomez-Garcia Alvestegui, D., Patel, A., Byers, K., Crowe, R., Player, K., Hale, R. D., Arnold, E. J., Smith, L., Gifford, C. M., Braaten, D., and Panton, C.: Advanced Multifrequency Radar Instrumentation for Polar Research, *IEEE Transactions on Geoscience and Remote Sensing*, 52, 2824–2842, <https://doi.org/10.1109/TGRS.2013.2266415>, 2014.
- 1085 Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, edited by Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., pp. 234–241, Springer International Publishing, Cham, ISBN 978-3-319-24574-4, 2015.
- Rückamp, M. and Blindow, N.: King George Island ice cap geometry updated with airborne GPR measurements, *Earth System Science Data*, 4, 23–30, 2012.
- 1090 Shi, L., Allen, C. T., Ledford, J. R., Rodriguez-Morales, F., Blake, W. A., Panzer, B. G., Prokopiack, S. C., Leuschen, C. J., and Gogineni, S.: Multichannel Coherent Radar Depth Sounder for NASA Operation Ice Bridge, in: 2010 IEEE International Geoscience and Remote Sensing Symposium, pp. 1729–1732, <https://doi.org/10.1109/IGARSS.2010.5649518>, 2010a.
- Shi, L., Allen, C. T., Ledford, J. R., Rodriguez-Morales, F., Blake, W. A., Panzer, B. G., Prokopiack, S. C., Leuschen, C. J., and Gogineni, S.: Multichannel coherent radar depth sounder for NASA operation ice bridge, in: 2010 IEEE international geoscience and remote sensing symposium, pp. 1729–1732, IEEE, 2010b.
- 1095 Sime, L. C., Hindmarsh, R. C., and Corr, H.: Automated processing to derive dip angles of englacial radar reflectors in ice sheets, *Journal of Glaciology*, 57, 260–266, 2011.

- Steinhage, D.: Contributions of geophysical measurements in Dronning Maud Land, Antarctica, locating an optimal drill site for a deep ice core drilling, *Ber. Polar-Meeressforsch.*, 384, 2001.
- 1100 Steinhage, D.: Links to master tracks in different resolutions from POLAR 6 flight ANT_1311070301, PANGAEA, <https://doi.org/10.1594/PANGAEA.174624>, in: Steinhage, D (2015): Master tracks in different resolutions during POLAR 6 campaign ANT_2013/14 [dataset publication series]. Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEA, <https://doi.org/10.1594/PANGAEA.856526>, 2015.
- Steinhage, D., Nixdorf, U., Meyer, U., and Miller, H.: Subglacial topography and internal structure of central and western Dronning Maud Land, Antarctica, determined from airborne radio echo sounding, *Journal of Applied Geophysics*, 47, 183–189, 2001.
- 1105 Steinhage, D., Franke, S., Eisen, O., Miller, H., and Helm, V.: Ice Thickness from western Dronning Maud Land (Antarctica) recorded with the airborne AWI EMR radar system in 1996/1997, PANGAEA, <https://doi.org/10.1594/PANGAEA.957056>, in: Steinhage, D et al. (2023): Airborne radar ice thickness measurements in Antarctica 1994-2007 [dataset publication series]. PANGAEA, <https://doi.org/10.1594/PANGAEA.957067>, 2023a.
- 1110 Steinhage, D., Franke, S., Eisen, O., Miller, H., and Helm, V.: Ice Thickness from western and central Dronning Maud Land (Antarctica) recorded with the airborne AWI EMR radar system in 1998/1999, PANGAEA, <https://doi.org/10.1594/PANGAEA.957059>, in: Steinhage, D et al. (2023): Airborne radar ice thickness measurements in Antarctica 1994-2007 [dataset publication series]. PANGAEA, <https://doi.org/10.1594/PANGAEA.957067>, 2023b.
- Stuefer, M., Rott, H., and Skvarca, P.: Glaciar Perito Moreno, Patagonia: climate sensitivities and glacier characteristics preceding the 2003/04 and 2005/06 damming events, *Journal of Glaciology*, 53, 3–16, 2007.
- 1115 Sugiyama, S., Skvarca, P., Naito, N., Enomoto, H., Tsutaki, S., Tone, K., Marinsek, S., and Aniya, M.: Ice speed of a calving glacier modulated by small fluctuations in basal water pressure, *Nature Geoscience*, 4, 597–600, 2011.
- Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C.: Efficient object localization using convolutional networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 648–656, 2015.
- 1120 Van Liefferinge, B. and Pattyn, F.: Using ice-flow models to evaluate potential sites of million year-old ice in Antarctica, *Climate of the Past*, 9, 2335–2345, 2013.
- Varshney, D., Rahnemoonfar, M., Yari, M., and Paden, J.: Deep Ice Layer Tracking and Thickness Estimation using Fully Convolutional Networks, in: *2020 IEEE International Conference on Big Data (Big Data)*, pp. 3943–3952, <https://doi.org/10.1109/BigData50022.2020.9378070>, 2020.
- 1125 Varshney, D., Rahnemoonfar, M., Yari, M., Paden, J., Ibikunle, O., and Li, J.: Deep Learning on Airborne Radar Echograms for Tracing Snow Accumulation Layers of the Greenland Ice Sheet, *REMOTE SENS-BASEL*, 13, <https://doi.org/10.3390/rs13142707>, 2021.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I.: Attention is All You Need, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, pp. 6000–6010, Curran Associates Inc, Red Hook, NY, USA, ISBN 9781510860964, 2017.
- 1130 Walker, B. and Ray, L.: Multi-Class Crevasse Detection Using Ground Penetrating Radar and Feature-Based Machine Learning, in: *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3578–3581, <https://doi.org/10.1109/IGARSS.2019.8899148>, 2019.
- Werder, M. A., Huss, M., Paul, F., Dehecq, A., and Farinotti, D.: A Bayesian ice thickness estimation model for large-scale applications, *Journal of Glaciology*, 66, 137–152, <https://doi.org/10.1017/jog.2019.93>, 2020.

- 1135 Willen, M. O., Horwath, M., Buchta, E., Scheinert, M., Helm, V., Uebbing, B., and Kusche, J.: Globally consistent estimates of high-resolution Antarctic ice mass balance and spatially resolved glacial isostatic adjustment, *The Cryosphere*, 18, 775–790, 2024.
- Williams, R. M., Ray, L. E., and Lever, J. H.: Autonomous robotic ground penetrating radar surveys of ice sheets; Using machine learning to identify hidden crevasses, in: 2012 IEEE International Conference on Imaging Systems and Techniques Proceedings, pp. 7–12, <https://doi.org/10.1109/IST.2012.6295593>, 2012.
- 1140 Williams, R. M., Ray, L. E., Lever, J. H., and Burzynski, A. M.: Crevasse Detection in Ice Sheets Using Ground Penetrating Radar and Machine Learning, *IEEE J SEL TOP APPL*, 7, 4836–4848, <https://doi.org/10.1109/JSTARS.2014.2332872>, 2014.
- Wu, Y. and He, K.: Group Normalization, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- Xu, M., Crandall, D. J., Fox, G. C., and Paden, J. D.: Automatic estimation of ice bottom surfaces from radar imagery, in: 2017 IEEE International Conference on Image Processing (ICIP), pp. 340–344, <https://doi.org/10.1109/ICIP.2017.8296299>, 2017.
- 1145 Xu, M., Fan, C., Paden, J. D., Fox, G. C., and Crandall, D. J.: Multi-task Spatiotemporal Neural Networks for Structured Surface Reconstruction, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1273–1282, <https://doi.org/10.1109/WACV.2018.00144>, 2018.
- Yari, M., Rahnemoonfar, M., Paden, J., Oluwanisola, I., Koenig, L., and Montgomery, L.: Smart Tracking of Internal Layers of Ice in Radar Data via Multi-Scale Learning, in: 2019 IEEE International Conference on Big Data (Big Data), pp. 5462–5468, <https://doi.org/10.1109/BigData47090.2019.9006083>, 2019.
- 1150 Yari, M., Rahnemoonfar, M., and Paden, J.: Multi-Scale and Temporal Transfer Learning for Automatic Tracking of Internal Ice Layers, in: *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6934–6937, <https://doi.org/10.1109/IGARSS39084.2020.9323758>, 2020.
- Young, D., Roberts, J. L., Blankenship, D. D., Van Ommen, T., Ritz, C., Caviite, M. G. P., and Frezzotti, M.: ICECAP radargrams in support of the international old ice search at Dome C - 2016, 2021.
- 1155 Zemp, M., Huss, M., Thibert, E., Eckert, N., McNabb, R., Huber, J., Barandun, M., Machguth, H., Nussbaumer, S. U., Gärtner-Roer, I., et al.: Global glacier mass changes and their contributions to sea-level rise from 1961 to 2016, *Nature*, 568, 382–386, 2019.