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

This manuscript presents a benchmark dataset, "IceAnatomy," designed to support and standardize the development and evaluation of deep learning models for extracting ice surface and bottom boundaries from radio-echo sounding (RES) radargrams. The dataset includes over 45,000 km of RES observations from multiple institutions and systems across diverse glaciological settings, along with baseline models and standardized train-test splits. Overall, the work addresses a pressing need in the cryosphere and remote sensing communities for reproducible, large-scale datasets that can accelerate progress in automated ice thickness estimation.

I commend the authors for their thorough and careful revisions, which have significantly improved the clarity, completeness, and overall quality of the manuscript since the previous review round.

Below, I provide detailed comments regarding the strengths and areas where the manuscript could be improved.

- In the sentence *"It is the first…"* line 45, the term *"human-annotated labels"* could be clarified further. Does this refer to fully manual annotations or semi-automated labels subsequently verified or correctd by humans? Given the importance of label quality in training and benchmarking deep learning models, this distinction is relevant for understanding the dataset's reliability.
- Line 51: please rewrite the sentence
- lines 52-75: The authors suggest that near real-time identification of the ice bottom boundary during RES data acquisition could allow for dynamic adjustments of flight plans to focus on areas of high interest. While this is an interesting idea, I wonder how often knowledge of the **ice bottom alone**, without broader context (e.g., basal conditions, surface conditions, prior survey goals), would justify altering flight plans during a campaign. Some clarification or examples from field experience would strengthen this claim and help the reader better understand its practical relevance.
- Line 57: "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." please clarify that " interpreting radargrams" is **only in terms of ice surface and ice bottom boundaries**.
- Line 75 – 77: The statement regarding the limitations of existing ice bottom labels is broad and lacks sufficient specificity. To strengthen this important critique, the authors should clearly separate and elaborate on each claimed issue, such as inaccuracies, automatic generation methods, data unavailability, lack of transparency, and missing radargrams, and provide concrete examples or citations of datasets where these problems have been documented. Without this clarification, the claim risks appearing vague and unsubstantiated, which weakens the justification for the need and novelty of the IceAnatomy dataset. More precise and evidence-backed discussion is necessary to convincingly demonstrate the dataset's advantages over existing resources.

- Line 77: In support of the statement regarding the limitations of existing ice bottom labels (e.g., inaccuracy, lack of transparency, or missing radargrams), the authors cite several references. However, Dong et al. 2022 use **synthetic radargrams**, which may not be directly relevant to a critique of real RES datasets or their associated manual/automatic annotations. I recommend revisiting this citation and ensuring that each reference clearly supports the specific issue being discussed. This would improve the precision of the argument and strengthn the manuscript's positioning.
- Line 102 – 108: In the list of references for works that track internal ice and snow layers, the citation *Moqadam and Eisen (2024)* is included alongside algorithm-focused studies. However, this is a **review article** rather than a method paper, so it may be better to distinguish it from the rest. Consider adding a sentence such as *"For an overview of methods used in this domain, see Moqadam and Eisen (2024)"* instead. This would clarify the nature of the citation and improve the precision of the literature summary.
- Line 141: The statement *"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"* would benefit from a supporting reference. Please consider citing glaciological studies or datasets that characterize the thermal regime and water content of these specific glaciers to substantiate this claim.
- Line 143: The authors state that the glacier characteristics *"pose a significant challenge to machine learning systems."*. That is very good intuition, I appreciate that. This is an important and plausible point, but it would be helpful to clarify whether this is based on **prior research**, **quantitative comparisons** in the current study, or anecdotal experience. If other studies have demonstrated lower model performance on temperate glaciers or radargrams from deep/steep troughs, please cite them. Otherwise, consider softening the language or providing some evidence from the dataset or baseline results presenteed in this paper.
- Line 154: The reference to *Rignot et al. (2011)* for ice velocity maps is valid, but more recent and higher-resolution velocity datasets are now available. I recommend updating or complementing this citation with a more recent source to ensure the comparison reflects the current state of ice velocity mapping.
- Line 203 – 213:
    - The authors provide a commendably detailed description of the annotation process, including the use of a single interpreter for consistency, cross-profile validation, and comparison with control points. This level of detail strengthens confidence in the dataset quality, good job.
    - Since the authors mention using *ReflexW* software for zooming and clarifying radargram features, it would be helpful to include a formal citation or reference for this commercial software to guide readers unfamiliar with it.
    - The description suggests that the labeling involved some degree of software-assisted (semi-automatic) annotation rather than purely manual picking. For clarity, please

specify whether the labels were created fully manually, semi-automatically with manual corrections, or a combination thereof. This clarification is important for users evaluating the dataset and its annotations.

- Line 268 – 270: The description of the U-Net–based model for ice boundary extraction is clear and well supported by relevant citations. However, I suggest including an additional recent relevant work in this context: *Moqadam et al. (2024),* which presents a closely related approach with U-Net for ice boundary extraction using deep learning. Also, as the cited version of this work is a preprint, please update the citation to the published version to ensure readers have access to the finalized paper.
- Line 341: “depth resolution is the time it takes for the wave to pass through the physical equivalent of a pixel in the radargram,” is not scientifically accurate. Depth resolution refers to the minimum vertical distance between two subsurface reflectors that can be distinguished as separate features in the radargram. It is a spatial (distance) parameter, not a temporal one, and depends on the radar wave velocity and the system’s temporal (time) resolution. I recommend revising this sentence for clarity and accuracy.
- Line 487: Tone of self-evaluation. The sentence claiming the work is "a significant step" and "an important advancement" could benefit from more objective framing or clearer support from the results. Consider revising this statement to maintain a more neutral tone in line with scientific conventions. While the impact of the work is indeed notable, I suggest moderating this language unless further evidence is provided to substantiate such claims in comparison to existing datasets or methods.