ASSESSING ANNOTATION ACCURACY IN ICE SHEETS USING QUANTITATIVE METRICS

Bayu Adhi Tama, Vandana Janeja, Sanjay Purushotham*

University of Maryland, Baltimore County, MD, USA

ABSTRACT

The increasing threat of sea level rise due to climate change necessitates a deeper understanding of ice sheet structures. This study addresses the need for accurate ice sheet data interpretation by introducing a suite of quantitative metrics designed to validate ice sheet annotation techniques. Focusing on both manual and automated methods, including ARESELP and its modified version, MARESELP, we assess their accuracy against expert annotations. Our methodology incorporates several computer vision metrics, traditionally underutilized in glaciological research, to evaluate the continuity and connectivity of ice layer annotations. The results demonstrate that while manual annotations provide invaluable expert insights, automated methods, particularly MARESELP, improve layer continuity and alignment with expert labels.

***Index Terms*—** Ice sheet annotation, quantitative metrics, automated annotation techniques, ice sheet structure analysis

1. INTRODUCTION

The escalating sea level rise, propelled by the ongoing climatic changes, underscores the vital need to understand the fundamental structure of ice sheets [1]. Acquiring this understanding is essential for enhancing the accuracy of future forecasts regarding the rise in sea levels, a topic of great concern for both coastal communities and climate researchers [2]. However, extracting essential information from ice sheet data is a substantial problem. The task is intricate and can be accomplished using several methodologies, such as manual, semi-automated, and fully automated procedures. Each of these solutions necessitates a substantial dedication of time and expertise from experts who meticulously annotate and assess complex data sets [3].

Considering the intricate nature and importance of these interpretations, validating ice sheet annotation methods becomes a vital element of glaciological research [4]. Errors or omissions in annotation might result in significant inaccuracies in comprehending and predicting ice sheet dynamics. In light of the crucial requirement for precision and dependability, this study presents a collection of new quantitative measurements. These measures aim to assess the accuracy of ice sheet annotation techniques thoroughly, enhancing our comprehension of glacial dynamics and their impact on sea level fluctuations. This study aims to leverage several metrics from the computer vision domain that have received limited attention in the existing literature. In addition, our work is centered around the utilization of automated annotation approaches, specifically ARESELP [5] and a customized version of ARESELP (e.g., MARESELP). These techniques are then contrasted with the manual labeling method conducted by a domain expert.

The evaluation and validation of layer-tracking performance in ice sheets have been subjects of significant research interest. A common approach to assessing the performance of layer-tracking algorithms is the use of synthetic age-depth profiles [6], which involves generating a synthetic age-depth relationship using a one-dimensional Nye model. This model is used to intersect picked isochrones with a known age-depth profile, thereby assigning an age to each pick and propagating this age-depth relationship across the ice sheet. The concept of isochrone connectivity is proposed by [7]. The metric assesses the degree of continuity and connectivity between detected ice layers, providing a quantitative measure of the uncertainty inherent in the tracking process. In addition, the metric is invaluable in highlighting areas where the interpretation consists of a high number of disconnected isochrones, which may indicate sensitivity to low amplitude signal anomalies.

2. MATERIAL AND METHOD

2.1. Radargram and Annotation Products

In this study, we harness the extensive data resources provided by the Centre for Remote Sensing of Ice Sheets (CReSIS), accessing a publicly available repository, to develop our annotation products. Our focus is concentrated on a selected set of 100 radargrams, all sourced from various locations across North Greenland. These radargrams, carefully chosen for their diverse characteristics and representativeness, include notable sequences such as 20120330_03_019-028, 20120404_01_004-011, 20120507_07_003-015, 20120508_04_002-018, 20120508_07_001-012, 20120508_07_015-023, 20120511_01_041-052, 20120511_01_061-067, and 20120516_01_080-091. Utilizing these specific radargrams, we embark on a detailed process of generating annotations for the ice

*Corresponding author

sheets.

2.2. Layer-Tracking Performance Metrics

2.2.1. Isochrones connectivity [7]

This method meticulously evaluates both the connectivity and continuity of the layers that have been identified. It achieves this by quantifying three key aspects: the total number of picked layers ($\#TL$), the number of layers that are continuous and uninterrupted ($\#CL$), and the number of layers that exhibit discontinuities or breaks ($\#DL$). An effective layer annotator is characterized by its ability to maximize connectivity while minimizing discontinuity, thereby ensuring a more accurate and cohesive representation of the layers.

Algorithm 1: Dip Estimation and Comparison

Require: $mask^a$, $mask^{gt}$, window_size
- Initialize dip_mask^a and dip_mask^{gt} as zero arrays with dimensions of $mask^a$ and $mask^{gt}$, respectively.

for each point in $mask^a$ and $mask^{gt}$
- Select a window_size.
- Compute transitions in the window.
- Calculate y and x differences of transitions.
- Compute angles using $arctan2$ of y and x differences.
- Calculate average dip as mean of angles.
- Assign average dip to the corresponding point in dip_results.

end for
- Calculate Pearson correlation coefficient between dip_mask^a and dip_mask^{gt}.

Return Correlation coefficient ρ.

2.2.2. Vision-based Metrics

- Pixel accuracy
  It quantitatively evaluates the accuracy of two binary masks by comparing them pixel by pixel. The calculation is the sum of all matching pixels divided by the total number of pixels in one of the masks. It is formally specified as $Acc. = \frac{\sum_i^N (mask_i^a == mask_i^{gt})}{N}$, where $mask_i^a$ and $mask_i^{gt}$ correspond to the values of the i^{th} pixel in the respective masks (i.e., ARESELP mask and ground truth mask), N is the total number pixels in the mask, and $\sum_i^N (mask_i^a == mask_i^{gt})$ represents the sum of pixels where the two masks have identical values (either both pixels are 1 or both are 0).

- Pearson's correlation of dip estimation
  It measures the similarity between the dip datasets obtained from two binary masks. The dip of each mask is estimated by calculating the dip angle for each pixel within a specified window size, and averaging the angles derived from transitions in the binary data [8]. Algorithm 1 shows a pseudocode of the metric calculation.

- Structural similarity index (SSIM) [9]
  It assesses the similarity between two binary masks. It considers changes in texture, providing a more perceptually relevant assessment of layer annotation similarity compared to simpler metrics like mean squared error. Algorithm 2 shows a pseudocode for calculating SSIM.

Algorithm 2: Structural Similarity Index Calculation

Require: $mask^a$, $mask^{gt}$
- Define parameters: window size, constants (C_1 and C_2) for stabilizing division with weak denominators.
- Initialize the SSIM map as a zero array with dimensions based on window size.

for each overlapping window in $mask^a$ and $mask^{gt}$.
- Extract corresponding windows from $mask^a$ and $mask^{gt}$.
- Calculate mean, variance, and covariance for these windows.
- Calculate y and x differences of transitions.
- Compute angles using $arctan2$ of y and x differences.
- Compute SSIM for the current window: SSIM(window) $= \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$
- Update SSIM map with computed SSIM value for the current window.

end for
- Calculate the mean SSIM over the entire image for the final SSIM value.

Return Final SSIM value.

- Recall IoU (IoU_r)
  The Recall Intersection over Union (IoU) metric is a method used to evaluate the accuracy of binary masks, specifically focusing on the overlap between a predicted mask ($mask^a$) and a ground truth mask ($mask^{gt}$). It calculates the ratio of the overlapping area (where both masks agree on positive pixels) to the total area covered by the ground truth mask, providing a measure of recall or how well the predicted mask captures the relevant areas of the ground truth. Algorithm 3 shows a pseudocode implementation of this metric.

- Layer-by-layer Recall IoU (IoU_r^l)
  It calculates the average recall for layer-by-layer comparison in binary masks. It first computes the IoU for

Algorithm 3: Recall Intersection over Union Calculation

Require: $mask^a$, $mask^{gt}$
- Compute the overlap as the sum of element-wise logical AND between $mask^a$ and $mask^{gt}$.
- Compute the total number of positive pixels in $mask^{gt}$.
- Calculate Recall IoU as the ratio of overlap to the total layers in $mask^{gt}$.

Recall IoU $= \frac{\text{Overlap}}{\text{Total Layers in } mask^{gt}}$.
Return Recall IoU.

each pair of layers between two masks. Then, it selects layer pairs with IoU scores above the average and calculates the recall for these pairs, which measures the proportion of actual positive samples (i.e., true layer matches) that are correctly identified. The average recall across these selected layers provides a metric for the overall accuracy of the layer identification in the masks. Algorithm 4 shows the outline for calculating layer-by-layer recall IoU metric.

Algorithm 4: Layer-by-Layer Recall IoU Calculation

Require: iou_scores, $mask^a$, $mask^{gt}$
- Initialize an empty list for recalls.
- Compute average IoU from iou_scores.
- Select layer pairs from iou_scores with IoU greater than or equal to the average IoU.

for each selected layer pair (i,j) in selected_pairs
- Extract corresponding layers from $mask^a$ and $mask^{gt}$
- Compute recall for the layer pair
- Append recall to recalls list

end for
- Compute average recall from the recalls list

Return average recall

Table 1. The mean quantitative score ($\pm$ standard deviation) of all the layer annotation techniques, such as manual approach by the expert, ARESELP, and MARESELP in terms of isochrones connectivity.

Method	$\#CL \uparrow$	$\#DL \downarrow$	$\#TL \uparrow$
Manual	7.63±6.75	144.03±58.75	151.66±59.99
ARESELP [5]	15.16±8.48	39.11±12.74	54.27±17.02
MARESELP	21.89±9.55	56.03±20.60	77.92±25.57

3. RESULT AND DISCUSSION

The implementation of all metrics is readily accessible online to ensure reproducibility and facilitate further research. Table 1 provides a comprehensive comparison of different layer annotation techniques - manual annotation by experts, ARESELP, and MARESELP - focusing on their performance in terms of isochrone connectivity. The metrics used for this comparison include the number of continuous layers ($\#CL$), the number of broken layers ($\#DL$), and the total number of layers ($\#TL$), where a higher number of continuous layers and a higher total number of layers are desirable, while a lower number of broken layers is preferred. The manual approach, traditionally considered the gold standard due to its reliance on expert interpretation, shows a moderate number of continuous layers but a significantly high number of broken layers, resulting in a total layer count ($\#TL$) of 151.6±59.99. This indicates the inherent challenges in manual annotation, where maintaining continuity across layers can be difficult, leading to a higher incidence of broken layers. In contrast, the ARESELP method shows a marked improvement in the number of continuous layers, more than double that of the manual approach. Notably, it also exhibits a substantial reduction in the number of broken layers, suggesting that this automated technique is more effective in maintaining layer continuity. The total number of layers identified by ARESELP is lower than that identified by the manual method, which may indicate a more selective layer identification process. MARESELP further advances these improvements, registering the highest number of continuous layers and a moderate number of broken layers, resulting in a total layer count ($\#TL$) of 77.92± 25.57. This suggests that MARESELP not only excels in identifying continuous layers but also strikes a balance in total layer detection, possibly offering a more nuanced and accurate representation of isochrones compared to the other methods. In summary, while the manual approach provides a substantial number of total layers, its high number of broken layers highlights the challenges of manual interpretation. ARESELP and MARESELP, on the other hand, demonstrate their strengths in automated layer annotation, particularly in maintaining layer continuity, as evidenced by their higher $\#CL$ and lower $\#DL$ scores. This evolution from manual to automated techniques underscores the potential of automatic approaches to enhance the accuracy and efficiency of ice sheet annotation.

In addition, Table 2 compares two automated annotation techniques, ARESELP and MARESELP, based on their performance in several vision-based metrics compared to expert labels (ground truth). These metrics include the Pearson correlation coefficient of the dip estimation (ρ), Structural Similarity Index ($SSIM$), Pixel Accuracy ($Acc.$), and Recall Intersection over Union for both global (IoU_r) and layer-by-layer (IoU_r^l) comparisons. The results demonstrate that while both ARESELP and MARESELP exhibit high accuracy and a

Table 2. The average quantitative score ($\pm$ standard deviation) derived from all pairwise comparisons between the expert labels (ground truth) and the automatic annotation techniques, specifically ARESELP and MARESELP, using vision-based metrics.

Method	$\rho \uparrow$	$SSIM \uparrow$	$Acc. \uparrow$	$IoU_r \uparrow$	$IoU_r^l \uparrow$
ARESELP [5]	0.527±0.190	0.822±0.059	0.973±0.010	0.633±0.185	0.313±0.125
MARESELP	0.547±0.184	0.827±0.056	0.974±0.009	0.692±0.154	0.325±0.120

strong structural resemblance to expert annotations, MARESELP shows a marginally better performance in aligning with expert labels, particularly in terms of IoU metrics. This indicates that MARESELP may offer a more refined and precise annotation than ARESELP, especially in complex layer-by-layer assessments. The superiority of MARESELP in terms of IoU metrics is particularly noteworthy, as it reflects a greater efficacy in capturing both the general and detailed aspects of ice sheet layers as annotated by experts However, annotation products from automatic approaches need to be further assessed, particularly when dealing with "hallucinated layers"- annotations that do not exist in the underlying radar imagery.

4. CONCLUSION

This study presents various quantitative metrics to validate the performance of ice sheet annotation techniques, specifically manual, ARESELP, and MARESELP. Our analysis reveals that while manual annotation provides valuable expert insight, it struggles with layer continuity, a challenge effectively mitigated by automated methods. ARESELP shows marked improvement in maintaining layer continuity, and MARESELP further excels by achieving the highest number of continuous layers and a balanced total layer count. With respect to vision-based metrics, MARESELP marginally outperforms ARESELP, especially in IoU measures, indicating its closer alignment with expert annotations. These findings highlight the potential of automated annotation techniques in enhancing the accuracy and efficiency of ice sheet analysis, which is crucial for understanding ice dynamics and their impact on sea level changes in the context of climate research.

5. ACKNOWLEDGMENT

This work is funded by the National Science Foundation (Institute for Harnessing Data and Model Revolution in the Polar Regions (iHARP) Award #2118285).

6. REFERENCES

[1] Nicholas R Golledge, "Long-term projections of sea-level rise from ice sheets," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 11, no. 2, pp. e634, 2020.

[2] Andrea Dutton, Anders E Carlson, Anthony J Long, Glenn A Milne, Peter U Clark, R DeConto, Ben P Horton, Stefan Rahmstorf, and Maureen E Raymo, "Sea-level rise due to polar ice-sheet mass loss during past warm periods," *science*, vol. 349, no. 6244, pp. aaa4019, 2015.

[3] Richard Delf, Dustin M Schroeder, Andrew Curtis, Antonios Giannopoulos, and Robert G Bingham, "A comparison of automated approaches to extracting englacial-layer geometry from radar data across ice sheets," *Annals of Glaciology*, vol. 61, no. 81, pp. 234–241, 2020.

[4] Naomi Tack, Bayu Adhi Tama, Atefeh Jebeli, Vandana P Janeja, Don Engel, and Rebecca Williams, "Metrics for the quality and consistency of ice layer annotations," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 4935–4938.

[5] Siting Xiong, Jan-Peter Muller, and Raquel Caro Carretero, "A new method for automatically tracing englacial layers from mcords data in nw greenland," *Remote Sensing*, vol. 10, no. 1, pp. 43, 2017.

[6] Joseph A MacGregor, Mark A Fahnestock, Ginny A Catania, John D Paden, S Prasad Gogineni, S Keith Young, Susan C Rybarski, Alexandria N Mabrey, Benjamin M Wagman, and Mathieu Morlighem, "Radiostratigraphy and age structure of the greenland ice sheet," *Journal of Geophysical Research: Earth Surface*, vol. 120, no. 2, pp. 212–241, 2015.

[7] Nanna B Karlsson, David M Rippin, Robert G Bingham, and David G Vaughan, "A 'continuity-index' for assessing ice-sheet dynamics from radar-sounded internal layers," *Earth and Planetary Science Letters*, vol. 335, pp. 88–94, 2012.

[8] Christian Panton, "Automated mapping of local layer slope and tracing of internal layers in radio echograms," *Annals of Glaciology*, vol. 55, no. 67, pp. 71–77, 2014.

[9] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.