Comments from Referee #3, followed by the authors' responses

C: Li et al. have developed a random forest classifier to produce a new set of glacier outlines over the southeastern Tibet region. One of the key issues with mass balance estimates is that they rely on a single set of glacier outlines, usually from the RGI, which does not account for glacier terminus changes. The paper produces a new set of outlines for 2000, 2005, 2010, 2015 and then 2016-2022 at annual resolution. The produced outlines were determined to be of high accuracy as determined by the kappa score in a confusion matrix. Furthermore, the authors compare mass balance estimates using a fixed outline and the changing outlines derived in this study and find that there is a 10% difference, although it was not clear from the paper if this was an under- or over-estimate. Finally, the authors find that the loss of glacier area in the region has been accelerating, although no discussion was made on potential drivers (although this was not the aim of this study).

R: We sincerely thank you for the constructive comments. In response, we have substantially revised the manuscript to improve clarity and rigor. The Introduction now gives a broader overview of machine learning and deep learning methods for glacier mapping, covering both traditional classifiers and advanced deep learning models, and explains why Random Forest was chosen for this study. We added a detailed analysis of input features, highlighting the roles of DEM, slope, and spectral indices, and confirmed model robustness through cross-validation. The Methods section has been streamlined with clearer descriptions of image preprocessing, multi-temporal selection, and data processing steps. Results and Discussion now provide a more thorough evaluation of classifier performance, including confusion matrices, feature importance, and fusion strategies, and emphasize the influence of glacier outlines on mass balance estimates. Figures and captions were refined for clarity, and text was revised to make variables, terms, and data periods more precise. Below, we provide a point-by-point response to each comment.

General Comments

C: The paper produces some useful results, particularly around the use of dynamic glacier outlines for quantifying glacier mass balance. The methods are thorough and

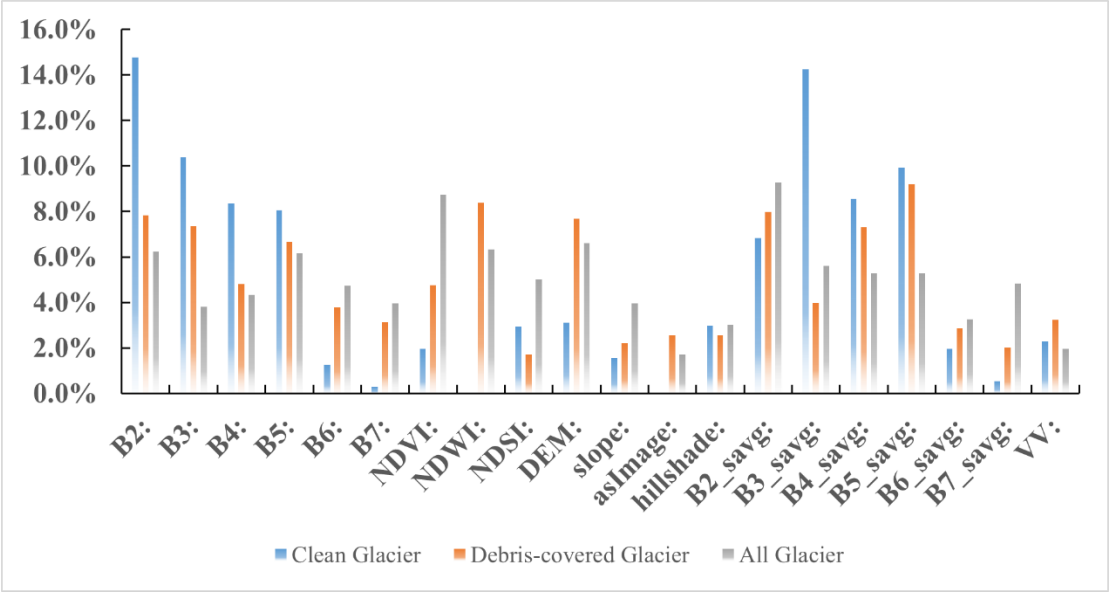
mostly well thought out with some caveats, and the findings appear to be of good quality, although some further information is required to improve understanding of these. There are several areas that require major improvement in a revised manuscript: The introduction requires a more detailed discussion of recent machine learning methodologies used to track glacier area and margin changes. More details are provided below in my technical comments, but the authors have missed a growing body of literature on this topic. This will help the authors better justify their choice of a random forest classifier used in this study.

R: Thanks for this valuable suggestion. In the revised manuscript, we expanded the introduction with a more detailed discussion of recent machine learning (ML) and deep learning (DL) approaches for glacier mapping. This now covers traditional ML classifiers (support vector machines, k-nearest neighbors, decision trees, gradient boosting, and random forests) as well as advanced DL architectures (U-Net, DeepLab V3+, attention-based CNNs, and Vision Transformers). We also highlight recent progress in automated global-scale glacier mapping with convolutional-transformer models such as GlaViTU. To explain our choice of the random forest (RF) classifier, we emphasize its proven robustness in classifying debris-covered glaciers and in cases with limited labeled data (Alifu et al., 2020; Lu et al., 2021). The revised text in the introduction now reads: *"In recent years, machine learning (ML) and deep learning (DL) have greatly advanced glacier remote sensing, enabling accurate mapping of glacier termini, area estimation, and surface feature analysis. AI-based automatic classification methods include support vector machines, k-nearest neighbors, decision trees, gradient boosting, multilayer perceptrons, artificial neural networks, and random forests (RF). Early DL models, such as U-Net (Ronneberger et al., 2015), have been applied to segment ice and ocean regions in Greenland and Antarctica (Baumhoer et al., 2019; Zhang et al., 2019). while DeepLab V3+ with atrous spatial pyramid pooling (ASPP) has been used for long-term glacier mapping (Cheng et al., 2020; Zhang et al., 2021). More recently, attention mechanisms (e.g., CBAM) and Vision Transformers (ViT) have further improved feature extraction in complex terrain and over large areas (Dosovitskiy et al., 2020; Chu et al., 2022). The Glacier-VisionTransformer-U-Net (GlaViTU) model enables automated, multi-temporal, global glacier mapping with accuracy approaching expert-level delineation, even in debris-rich regions(Maslov et al., 2025). Traditional ML remains effective in cases of debris-covered glaciers or limited labeled data. For example, Y. Lu et al. (2021) proposed a composite model that*

integrates RF and convolutional neural networks, while Xie Fuming et al. (2020) combined Otsu thresholding with ML algorithms on the Google Earth Engine to extract debris-covered glaciers in the Hunza Basin, achieving a Kappa coefficient of 0.94 ± 0.01 and an overall accuracy of 95.5 ± 0.9%. Alifu et al. (2020) further demonstrated that RF outperforms other classifiers for debris-covered glaciers, supporting its role as a robust core classifier. Collectively, these studies show that ML and DL approaches substantially improve the automation, accuracy, and scalability of glacier mapping compared with traditional index-based techniques."

C: A more critical review of the features used to train the random forest classifier is needed. In particular, a sensitivity analysis will assist in understanding which features in the model are dominating the training. Furthermore, the authors used a broad range of features in the random forest classifier, hence it would be interesting to see if the model is overfitting in some way due to the diversity of input data. Quantifying this would help improve reliability in the final results.

R: We are grateful to you for raising this point. In response, we have added a more critical review of the features used to train the random forest (RF) classifier. Specifically, we performed a sensitivity analysis based on feature importance scores derived from the RF model. As shown in Figure X, spectral bands (e.g., B2, B4, B5) and their spatially averaged values contribute most to the classification, followed by spectral indices (NDVI, NDWI, NDSI) and topographic variables (DEM, slope). Notably, DEM and slope are particularly important for debris-covered glacier mapping, while NDSI and NDWI dominate in clean glacier detection. SAR features (VV) exhibit relatively lower importance. To evaluate whether the inclusion of diverse input features could lead to overfitting, we further applied out-of-bag (OOB) error estimates and 10-fold cross-validation. The results indicate stable classification performance across different feature subsets, confirming that our RF model does not suffer from significant overfitting. Changes in the manuscript – We have added a new subsection in Methods describing the sensitivity analysis, a new figure showing feature importance for clean glaciers, debris-covered glaciers, and all glaciers, and a corresponding explanation in Results and Discussion.



C: The methods section is verbose and could be shortened significantly. This will allow for more space to discuss model performance later in the paper.

R: Thank you for this suggestion. In the revised manuscript, we have streamlined the Methods section by removing redundant details and condensing the text, thereby improving clarity and ensuring a better balance between methodology and results.

C: It would be useful to understand the performance of the random forest classifier in different contexts. In particular, how does it perform for different satellite images e.g. Sentinel-2, Landsat 7, Landsat 8 etc. Currently, the uncertainty is taken as 1 deriviative of the pixel size, but it should really reflect the accuracies of the glacier outlines which will vary with different data sets.

R: We thank you for this comment. In this study, Landsat 7, Landsat 8, and Sentinel-2 datasets were combined to reduce gaps caused by clouds, missing acquisitions, or seasonal limitations, resulting in more complete glacier maps. Although classification performance may vary across datasets, the main focus was on generating reliable annual glacier outlines. The robustness of the random forest model was validated using out-of-bag (OOB) error, and a more detailed analysis of dataset-specific performance and uncertainties will be addressed in future work.

C: A key outcome of the study is the impact of dynamic glacier outlines on mass balance calculations, but this is not explored sufficiently in the study. I would urge the authors

to present these results more fully and discuss the implications of this for mass balance studies in Tibet and the wider globe.

R: We appreciate your suggestion. We fully agree that the impact of dynamic glacier outlines on mass balance calculations is a key outcome of our study. In the revised manuscript, we have expanded the presentation of these results and provided a more detailed discussion of their implications, including the influence of multi-temporal glacier area changes on mass balance estimates for glaciers in the Tibetan Plateau and considerations for broader applications in other regions worldwide.

There are several typos and gramatical mistakes throughout the paper, some of which I have highlighted in my technical comments, but I would encourage the authors to thoroughly review the manuscript upon revision.

Technical Corrections (References to line (L) numbers in preprint)

C: L10: Better to say 'glacier area'? Also, the latter part of the sentence only applies to optical data.

R: W We appreciate your suggestion. We have revised the text to use "glacier area" for clarity. While persistent cloud cover mainly affects optical data, seasonal snow accumulation can impact both optical and radar observations.

C: L12: 'the Landsat satellite series'

C: L14: 'for this region'

C: L19: 'we calculated glacier mass balance'

C: L20: 'glacier areas calculated in this study, resulting in an annual mass loss of 6.20'

C: L36: 'hence the region is dominated by maritime glaciers'

C: L39: 'Glacier area mapping from satellite imagery'

C: L40: 'substantial time for human interpretation'

C: L64: ‘from optical satellite imagery’

C: L72: ‘Qinghua,2020), who’

C: L203: ‘Spectral reflectance alone is insufficient’

C: L204: ‘this study extracts spectral, terrain, texture, and radar interferometric features to train a Random Forest classifier for delineating glaciers in satellite imagery.’

C: L371: ‘4 Results’

C: L397: ‘It is noted that’

C: L430: Missing word, glaciers are losing mass at a rate of 6.20 Gt/y?

R: We have revised the manuscript to address all formatting, spelling, and wording issues, including those noted in lines L12, L14, L19, L20, L36, L39, L40, L64, L72, L203, L204, L371, L397, and L430. In addition, references and citation formats have been thoroughly checked and updated throughout the text.

C: L15: ‘integrating a three-year dataset’ isn’t clear to me- do you mean delineating glacier area for 3 years and then the median year is taken to be the time satmp?

R: Thanks for your comment. By “integrating a three-year dataset,” all available data from T − 1 to T + 1 were used for each target year, rather than only the median year, to reduce gaps due to missing early Sentinel observations.

C: L33-35: Does this sentence refer to the Tibetan Plateau specifically? If so, can the authors state this.

R: Thanks for pointing this out. The sentence refers specifically to the Tibetan Plateau, and we have revised the manuscript accordingly.

C: L37: What glacier changes? The natural cycle of accumulation/ablation or a longer term trend? This is not clear.

R: Thanks for pointing this out. The glacier changes mentioned correspond to long-term trends, with the fastest retreat and high accumulation and ablation rates. We have clarified this in the revised manuscript.

C: L39-52: The description of NDSI could be improved e.g. the use of a manual threshold is only mentioned at the end. The authors state a weakness is lack of automation, which is true, but there is a wider point that the application of NDSI varies in different geographic regions, which makes it hard to automate the process. This should be acknowledge.

R: Thank you for this suggestion. In the revised manuscript, we have clarified the description of NDSI by introducing manual thresholds earlier and emphasizing that these thresholds vary across geographic regions, which makes full automation challenging.

C: L43-44: Extracting what component of glacier and snow cover? Area changes? Differentiating between the two surfaces? Probably both.

R: Your suggestion is appreciated. The manuscript now explicitly states that NDSI is used to delineate the extent of glaciers and seasonal snow.

C: L53-63: This is quite a vague paragraph that misses a lot of important studies mapping glacirs with ML e.g. for terminus mapping, glacier area estimates and surface features (e.g. ????). There is a growing body of literature in this field and this should be acknowledged with a more detailed literatyre review in this section.

R: Thanks for your suggestion. In the revised manuscript, we expanded the literature review on machine learning (ML) applications in glacier mapping, covering glacier terminus mapping, area estimation, and surface feature classification. Additional references have been included to provide a more comprehensive overview.

C: L53: 'Recent developments in machine learning algorithms have enabled large volumes of satellite imagery to be used as training data for automated classification of glaciers'- or something like this. It's important to be clear what ML does and how it improves over the traditional techniques.

R: We appreciate your comment. The manuscript now clearly explains how ML enhances glacier mapping relative to traditional index-based methods. Specifically, we added: *"Collectively, these studies show that ML and DL approaches substantially improve the automation, accuracy, and scalability of glacier mapping compared with*

traditional index-based techniques."

C: L70: Define 'high temporal resolution'- either weeks, months, seasonal or years.

R: Thanks for your suggestion. In the revised manuscript, we clarified that "high temporal resolution" refers to annual to once-per-decade observations and updated the sentence accordingly: *"Consequently, generating glacier inventories with high temporal resolution (i.e., annually to once per decade) in the southeastern Tibetan Plateau remains a significant challenge."*

C: L72-77: What are the details of this inventory? What is their estimate of the number of glaciers, area etc.?

R: Your suggestion is appreciated. We clarified that here we introduce only existing glacier inventories and datasets, with detailed statistics provided in the Results section.

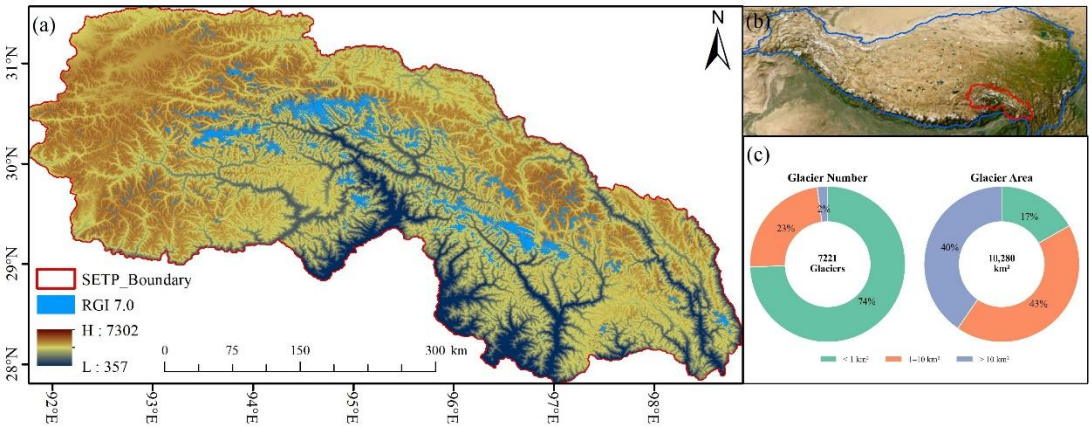
C: L91-98: I would like to see a bit more discussion of the important of glacier inventories (e.g. areas) for quantifying mass changes e.g. how do the GLAMBIE/IMBIE community estimates tackle this problem and what is the consensus approach when multi-temporal data sest aren't available? What is the impact on uncertainty estimates? This will naturally then lead onto the objectives in the paragraph.

R: Your suggestion is appreciated. Traditional altimetry-based methods rely on a static glacier boundary, overlooking area changes and risking systematic bias. To mitigate this, multi-temporal glacier inventories are incorporated, such as in the GlaMBIE approach, which uses RGI 6.0 as a baseline and adjusts mass balance with regional glacier area changes. We added the following sentence: *"Traditionally, altimetry-based methods calculate glacier mass change using a single, static glacier boundary, which ignores changes in glacier area and may introduce systematic biases. To address this limitation, current approaches increasingly incorporate multi-temporal glacier inventories to account for dynamic glacier areas. For example, the GlaMBIE community uses the Randolph Glacier Inventory (RGI 6.0) as a baseline and apply regional glacier area change rates to adjust mass balance calculations over time (Zemp et al., 2025). Regional studies further demonstrate the importance of this practice: in Peru, glaciers lost approximately 29% of their area between 2000 and 2016, with*

accelerated mass loss during 2013–2016 (−660 ± 178 kg m²/a) (Seehaus et al., 2019), and in Bolivia, glaciers in the Cordillera Real and Tres Cruces also experienced a 29% area reduction over the same period, with total mass loss of 1.8 ± 0.5 Gt and enhanced losses during 2013–2016 due to El Niño (−487 ± 349 kg m²/a) (Seehaus et al., 2020). When multi-temporal inventories are unavailable, static glacier boundaries are assumed, which can increase uncertainty. Collectively, these studies demonstrate that incorporating dynamic glacier areas into mass balance calculations is essential for accurate and robust estimates of glacier mass change."

C: L109: For those unfamiliar with this region, it might be worth zooming out a bit and placing an inset map to show the position of this region in the wider regional context.

R: Thanks for your suggestion. We added an inset map to Figure 1 showing the study region within the southeastern Tibetan Plateau, helping readers better understand its location and spatial context.



C: L113; What does 'glacier distribution area' mean?

R: We appreciate your comment. To improve clarity, we have revised the wording. Instead of "glacier distribution area," we now state that "t*he southeastern Tibetan Plateau is one of the major glacierized regions in China, containing a high concentration of glaciers and abundant ice reserves."* This avoids ambiguity and more accurately conveys the intended meaning.

C: L124: This section is not consistent- sometimes the sampling is describes, in other sections it is not. Either describe the sampling within each section or create a new

section where it is fully described.

R: Your suggestion is appreciated. Sampling details have been moved to the Methods section for a more systematic and consistent presentation.

C: L126: What does 'analysis-ready' mean? What processing has been applied before these images are provided on GEE?

R: Thanks for your suggestion. We clarified the term "analysis-ready" in the text. The Landsat Surface Reflectance Tier 1 datasets on GEE have undergone radiometric calibration, atmospheric correction, and geometric correction, making them directly usable for scientific analysis. We also streamlined the description of spectral bands and noted potential data limitations.

"This study utilizes the Landsat-5, Landsat-7 (pre-2003), and Landsat-8 Surface Reflectance Tier 1 datasets, provided on GEE in an analysis-ready format. These provides have undergone radiometric calibration, atmospheric correction, and geometric correction, ensuring that the reflectance data reliably represent surface features. The datasets include visible (VIS; 400–700 nm), near-infrared (NIR; 700–900 nm), and shortwave infrared (SWIR; 1400–2400 nm) bands at 30 m spatial resolution. Although Landsat offers a 16-day revisit cycle, data quality can be affected by cloud cover, seasonal snow cover, and sensor anomalies."

C: L139-145: Given the introduction focuses on the limitations of optical data, the authros should discus somewhere the pro's and con's of using SAR data as an alternative.

R: Thanks for your suggestion. We clarified the role of Sentinel-1 SAR data in the manuscript, emphasizing its use as supplementary information to improve glacier classification under cloudy conditions, given its high temporal resolution and reliability in adverse weather. *"The Sentinel-1 satellite is a synthetic aperture radar (SAR) mission launched by the European Space Agency (ESA). This study utilizes the COPERNICUS/S1_GRD dataset, accessed via the GEE platform with a six-day revisit interval. The VV polarization band (vertical–vertical) provides high temporal resolution and consistent multi-temporal observations. These data remain reliable*

under cloud cover or precipitation, making them valuable as supplementary inputs for training the classification model and enhancing robustness where optical data are limited. Nevertheless, glacier mapping with SAR can still be challenging due to signal saturation over wet snow, geometric distortions in mountainous terrain, and difficulties in distinguishing clean ice from debris-covered surfaces. Additionally, Sentinel-1 data are only available from 2015 onwards, so they do not cover the entire study period."

C: L146:153: What is the time stamp of the NASADEM? Or is it a dynamic data set?

R: Your suggestion is appreciated. We revised the manuscript to clarify the NASADEM dataset and its use for deriving elevation, slope, aspect, and hillshade, including information on resolution, sources, and processing.

C: L155-161: Time stamp of 2000 for RGI7.0.

R: Your suggestion is appreciated. We revised the text to specify that RGI 7.0 depicts glacier outlines for approximately the year 2000.

C: L167: Vague- define exactly in which period the data were acquired. If T is the sampling year, did you obtain all sutiable summer images in years T ± 2 years?

R: Thanks for your suggestion. We clarified in the manuscript that for 2000, 2005, 2010, and 2015, all suitable summer images within a ±1-year window around each target year were used. For 2016–2022, only images from the corresponding year were included.

C: L201: Do the image data cubes represent the 'image composites' described above? It would be useful to have consistent language throughout the manuscript to avoid confusion.

R: Your suggestion is appreciated. We revised the text to clarify that the image data cubes represent raw collections before composite generation.

C: L210: I'm confused here, how do Figures 4a-f represent cloud-free image composites?

R: Thanks for your suggestion. Figures 5a–f present the results after cloud masking and compositing, while the detailed workflow is shown in Figure 4.

C: L226: Which images are used to generate the NDVI image for each? Did you merge the NDVI values for a single year?

C: L233: Same point as for NDVI, not clear to me which images are being used to calculate this.

C: L242: Same as for L226 and L233.

R: Thanks for your suggestion. As shown in Figure 4, NDVI and other indices were first calculated from cloud-free, shadow-corrected individual images. These individual images were then composited to generate a single annual image for each year.

C: L248: I am confused by this figure- I assume each of the horizontal squares represents an image, so what do the colours represent? And what do the vertical boxes represent

R: Thanks for your suggestion. In the figure, each horizontal layer represents images from the same acquisition time, and each vertical column corresponds to the same spatial location. Colors indicate image values, while missing colors reflect gaps caused by clouds. For each location, cloud-contaminated observations were removed, indices (e.g., NDVI, NDSI, NDWI) were calculated from the remaining data, and these index images were composited to generate a single annual image.

C: L266-265: Image textures are better defined as the spatial arrangement of pixels in an image

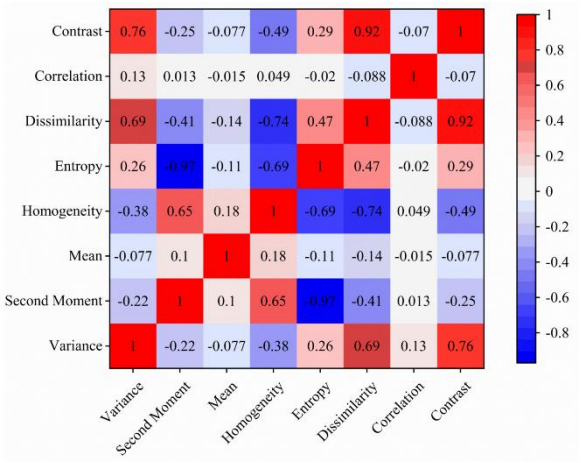
R: Thanks for your suggestion. We clarified that texture features capture spatial patterns independent of color or brightness, and that the gray-level co-occurrence matrix (Haralick, 1979) is used to quantify these patterns.

Haralick, R. M.: Statistical and structural approaches to texture, Proceedings of the IEEE, 67, 786-804, 1979.

C: L277: Is this the mean texture fro GLCM? It's not clear why this was chosen- the authors state that a previous study found it is 'consistent with other textures'- why would this mean it is the best feature to use? If it is consistent with other features, then any other texture feature could be used e.g. autocorrelation, entropy etc.?

R: Thanks for pointing this out. The mean texture from the GLCM was chosen based on previous studies (Lu et al., 2020), as it correlates strongly with other common texture

features. This provides a representative measure while reducing redundancy. It is not necessarily the "best," but balances information content, robustness, and computational efficiency. Other features like autocorrelation, entropy, or second-order moments could be used, but they either add redundancy or complicate glacier discrimination.



Correlation coefficients between texture features.(Lu et al., 2020)

Lu, Y., Zhang, Z., and Huang, D.: Glacier Mapping Based on Random Forest Algorithm: A Case Study over the Eastern Pamir, Water, 12, 10.3390/w12113231, 2020.

C: L287: What is a 'mean synthesis'? Also the 'salt-and-pepper noise' I assume is referring to 'speckle'- calling it noise is incorrect as speckle is a repeatable feature in SAR data.

R: Thanks for your suggestion. SAR imagery often contains speckle noise, which can reduce image quality. To address this, we averaged multi-temporal Sentinel-1 images on a pixel-by-pixel basis, stabilizing the data used to train the random forest classifier.

C: L291-304: Without a suitable figure (ie. Figure 5), it is difficult to interpret the feature layers described in this section. The inclusionf of RGI outlines would help, but also subtitles and a larger legend will help readability.

C: L292-293: This is not clear in Figure 5, see comment below.

C: L305: Figure 5: It's not clear what the values represent, the legend is way too small. One legend for all composites is sufficient, unless the values are significantly different between each panel. It would also be useful to overlay the RGU outlines here so the

reader can visually assess how wel each image feature matches the glacier area. Also, if this is referenced before Figure 3, it should also be first in the order of figures.

R: Thanks for your suggestion. We clarify that Figure 5 provides the processed feature layers used for glacier classification, and the RGI outlines are included to facilitate comparison. We have also updated the figure with clearer subtitles and an enlarged legend to improve readability and interpretation.

C: L314-319: RF has been widely used, although arguably it has been superseded by CNNs and foundation models. Can the authors comment on why they did not apply these other methods?

R: Thanks for your suggestion. In selecting the Random Forest algorithm, we focused on both data availability and computational efficiency. RF performs reliably even with limited labeled data and allows processing of multiple years across a large region. Although CNNs or other foundation models could potentially improve accuracy, RF offers a practical trade-off between performance, interpretability, and efficiency for multi-temporal glacier mapping.

C: L322: Are the labels used for all images or a subset? For the images labelled, are the labels shown in Figure 6 suitable for all images given the potential for changes in surface characteristics at different times of the year?

R: Thanks for pointing this out. To ensure accurate and representative labels, training samples were manually delineated separately for each year, meaning the sample points differ annually. While this approach preserves data quality, it naturally limits the maximum classification accuracy. Developing effective strategies for transferring or reusing samples across years remains an active area of our research.

C: L327-336: Are you discussing here the training data, validation data, or both? Subtitle is misleading, 'Selection of Classification Samples' doesn't really say anything here. How many images where the training data taken from?

R: Thanks for your suggestion. This selection applies to both training and validation data. All samples were manually delineated on the composite images, with 70% used for training the classifier and 30% reserved for validation.

C: L338-342: F1 score might be more suitable here if there is class imbalance- I suspect there is imbalance in the training data, but it is not stated.

R: Thanks for your suggestion. To prevent any class imbalance, we carefully balanced the number of samples for each land cover type in the training dataset. This approach reduces bias and ensures that metrics like overall accuracy, precision, recall, and F1 score accurately reflect performance across all classes. Using the 2022 confusion matrix, the glacier extraction achieved an F1 score of 95.5%, with precision 94.5% and recall 96.5%.

C: L358: What is meant by a 'decision-level fusion strategy'?

R: Thanks for highlighting this. The "decision-level fusion strategy" refers to merging the classification outputs from Sentinel-2 and Landsat individually. By doing this, we use the strengths of both datasets, enhancing the final glacier map's accuracy and robustness.

C: L372-380: This a surprisngly short section that only gives the headline figures. I would like to see a sensitivity analysis of the random forest classifier, particularly an understanding of which texture features were more important for classification than others. One possibility of using a diverse range of features is that the model could be overfitting, potentially leading to errors in the resultant classification maps. Furthermore, how do the accuracies compare for different data sets? I would expect there to be differences in Sentinel-2 vs Landsat, whislt Landsat 7 would likely yield different accuracies to Landsat 8. This information must be included to better understand the performance of the technique.

R: Thanks for your suggestion. Single datasets like Landsat 7, Landsat 8, or Sentinel-2 alone can't always provide full coverage because of clouds, missing data, or seasonal gaps. That's why we combined multiple datasets using a decision-level fusion, which merges the classification results from each dataset to produce more complete annual glacier maps. Although the classification performance differs slightly among sensors, our main goal was reliable glacier outlines. We confirmed the random forest model is robust using out-of-bag error, and future work will investigate dataset-specific performance and feature importance in more depth.

C: L373: Referrring to the ‘annual’ classification results, I assume you mean for the results after 2016 with the Sentinel / Landsat results? Furthermore, the authors should show here the confusion matrix to better highlight true positives, true negatives, false positives and false negatives. A single accuracy score may be misleading.

R: Thanks for your suggestion. Here, “annual” classification results refer to the outputs of the random forest applied to the combined Sentinel-2 and Landsat data from 2016–2022. To give a more detailed view of classifier performance, we include the 2022 Landsat confusion matrix (Table S1), showing true positives, true negatives, false positives, and false negatives. This complements the overall accuracy and F1 score, providing a clearer picture of performance across glacier and land cover classes.

Table S1 Confusion matrix of 2022 Landsat glacier classification

Actual \ Predicted	Bare Glaciers	Debris-covered Glaciers	Bare Ground	Water	Vegetable	Hillshade
Bare Glaciers	76	1	0	0	0	0
Debris-covered Glaciers	0	60	4	0	1	0
Bare Ground	0	8	49	0	2	0
Water	0	0	0	67	0	0
Vegetable	0	0	3	0	59	4
Hillshade	0	0	0	0	0	45

C: L386: Why is the mapping error based on half the image element? Surely the graph should be representing uncertainty calculated from the random forest model outputs?

R: Your suggestion is appreciated. In glacier mapping studies, half a pixel is often used to represent mapping error. Our focus is on the cartographic accuracy of the produced glacier maps. Because the classification outputs undergo extensive post-processing, directly using uncertainty from the random forest model does not fully reflect the accuracy of the final mapped products.

C: L395-405: This section is a bit confusing. It might be helpful to construct a table with the key results from previous studies to make it clear how the results in this paper

compare?

R: Thanks for your suggestion. In the revised manuscript, we reorganized the section and created a table highlighting key results from previous studies. This helps readers directly compare our results with existing work and enhances the clarity of the discussion.

C: L406-419: This reads like a results section- also, the inclusion of ICESat, ICESat-2, and CryoSat-2 should really be discussed in the methodology section. Are the data sets extracted simply just the time series as presented? Or did the authors process the data sets in some way? I also don't think thickness and area should be presented on the same graph, it might cause confusion- I would use 2 panels instead.

R: We acknowledge your point. We have moved the discussion of ICESat, ICESat-2, and CryoSat-2 data to the Methods section. The datasets were preprocessed to remove outliers and ensure temporal consistency before extracting the time series. To improve clarity, thickness and area are now presented in separate panels in the revised figures.

C: L423-424: The variables in the equations need to be stated.

R: Thanks for your suggestion. We revised the manuscript to define all variables in the equations clearly, specifying each symbol's meaning and units where applicable.

C: L432-434: This is an important result, but it is not shown graphically. Can the authors make a figure showing this key result? Although, what does the 10% refer to- an under- or an over-estimate compared to the fixed outlines?

R: Thanks for your suggestion. Compared with using static glacier outlines, the fixed-area approach underestimates glacier mass change by ~10%. We added a figure showing annual glacier mass change relative to the fixed RGI 7.0 outlines.

C: L420-447: These are results, and the methods described here should be presented in a methodology section. The small discussion towards the end should be expanded particularly focusing on the importance of updated glacier outlines for mass balance estimates, as this is key moving forward in future studies.

R: Thanks for the comment. We shifted the ICESat, ICESat-2, and CryoSat-2 data processing details to the Methods section. The revised text also stresses that using updated glacier outlines is important—static outlines can underestimate mass loss, so dynamic mapping is key for future glacier studies.

C: L473: Define the number of images used and over what time period

R: Thanks for your suggestion. In this study, we incorporated a comprehensive set of images from Landsat 7, Landsat 8, and Sentinel-2, covering the entire period from 2000 to 2022 to ensure complete temporal coverage.

C: L472-483: I would expect the conclusions to mention the performance of the random forest model as well

R: Thanks for your suggestion. We revised the Conclusions to underscore the random forest model's strong performance and reliability, validated through out-of-bag (OOB) error assessment.