We thank the reviewer, Brian Kyanjo, for the effort in assessing our manuscript and for the positive feedback and suggestions for improvement. We have considered the feedback carefully and made several changes to the manuscript in accordance with the comments provided. The main improvement is the addition of a feature importance analysis in Appendix C (including two new figures, Fig. C1 and C2), where we consider and discuss feature importance based on different metrics. In addition, we have made several updates to the text in relation to the specific comments. Please see our detailed responses below to each of the specific comments. In the following response, reviewer comments are indicated in black and our responses are indicated in blue italic font.

Reviewer 2:

Overall Assessment

The paper "egusphere-2025-1206" (https://doi.org/10.5194/egusphere-2025-1206) presents the Mass Balance Machine (MBM), a machine learning model for predicting glacier surface mass balance, trained on a robust dataset of 4,201 point mass balance measurements from 32 Norwegian glaciers (1962–2021) using the XGBoost algorithm. This work is a significant advancement in glaciology, offering high-resolution predictions that outperform traditional models like GloGEM, OGGM, and PyGEM, particularly for seasonal mass balances. Its potential for applications in climate change research and water resource management is substantial. However, minor refinements in data resolution, model transparency, and uncertainty analysis, along with clarifications in Appendix A, could elevate its impact. Below, I provide a detailed review of the paper's strengths, areas for improvement, and a focused analysis of Appendix A, including specific corrections and broader recommendations.

Strengths of the Paper

1. Robust Dataset: The dataset, sourced from the Norwegian Water Resources and Energy Directorate (NVE) database, spans nearly six decades and includes 4,201 point mass balance measurements across 32 glaciers. The thorough cleaning process, detailed in Appendix A, ensures reliability by addressing missing coordinates and outliers, resulting in 3,910 annual, 3,929 summer, and 3,751 winter measurements.

2. Effective Methodology: The use of XGBoost is well-suited for capturing complex relationships between weather, terrain, and glacier mass balance. The independent glacier-based train-test split enhances the model's generalizability, making results trustworthy.

3. Superior Performance: MBM demonstrates lower RMSE and bias compared to established models, particularly for seasonal predictions. This is evident in figures like Fig. 6 and Table D1, which effectively support the text.

4. Practical Applications: The model's high-resolution predictions at point and monthly scales are valuable for water resource planning and glacier flow modeling in a warming climate.

Areas for Improvement

1. Data Resolution: The ERA5-Land data (9 km resolution) may be too coarse for smaller glaciers. Exploring higher-resolution datasets or downscaling techniques could improve local accuracy.

We agree that the resolution of ERA5-Land is coarse compared to the size of the glaciers in the dataset. Our intention with using a globally available dataset is that the model can easily be adapted to other regions. In addition, ERA5-Land is relatively high resolution compared to other globally available datasets. Some higher-resolution climate datasets exist for Norway (e.g. the NORA datasets), but these do not cover the entire time period of the mass balance measurements. However, in terms of resolution, we believe that MBM can, at least to some degree, implicitly downscale the meteorological data to the elevation of the point measurements by using the elevation difference feature to distinguish between points within the same ERA5-Land cell. This is evidenced by MBM's performance on such data (Figs. 5 and 6) and supported by the newly added feature importance analysis. We have elaborated on this in Section 6.1.1:

400 ~~MBM effectively~~ The performance of MBM on point mass balance and the apparent importance of the elevation difference feature (see feature importance analysis in Appendix C) suggests that MBM implicitly downscales and bias-corrects relatively coarse meteorological data to the point scale. In addition to the spatio-temporal transfer of mass balance information across glaciers, MBM's apparent downscaling capacity is crucial for generating accurate high-resolution predictions. For instance, ~~as accumulation is primarily governed by precipitation and temperature,~~ MBM's strong performance in reconstructing winter
405 mass balance at the stake level (Fig. 5a and b)~~shows its ability to downscale these variables~~, together with a high importance of precipitation and elevation difference features in winter months (Fig. C2a–c and k–l), suggests that it is able to downscale precipitation locally. The ~~key to MBM's downscaling abilities~~ same is true for temperature in the summer months (Fig. C2e–i). The key to this ability lies in using the elevation difference between the stake and the climate model as a feature (Fig. 3) which enables MBM to effectively map the relationship between climate and elevation.

It would be interesting to compare the performance of the model on other climate datasets and different resolutions, and we expect that the performance of both MBM and the other models to increase with increasing climate data resolution. We mentioned the use of higher-resolution meteorological data already in Section 6.2.1 as an option that may improve MBM's predictions (and limit reliance on high-resolution topographical features). We consider this to be out of the scope of the current study, but have elaborated some more in Section 6.2.1:

can resolve smaller-scale variations. Artefacts in the topographical ~~data may , therefore , influence predictions, for example,~~ ~~the high~~ features may therefore influence predictions. For example, MBM predicts high summer melt rates along the ~~border on~~ eastern border of the tongue of Tunsbergdalsbreen (Fig. 11c). ~~Here, MBM predicts more negative mass balance for~~ We believe this is due to the combination of steep and ~~south-facing~~ south-west facing slopes (Fig. 11g and h). However, these ~~artefacts likely~~ steep, south-west facing slopes are likely topographical artefacts. They result from the ~~calculation of slope and aspect arising from the influence of the steepterrain surrounding the~~ surrounding terrain influencing the calculation of these variables from the DEM, specifically a steep, south-west facing wall that borders the glacier tongue. The issues outlined here may be mitigated by extracting meteorological variables from a single ERA5-Land cell closest to the glacier centre~~or~~. Another option would be to train MBM using higher-resolution meteorological data, which may also elucidate MBM's downscaling capabilities. Regardless of these challenges, ~~MBM's~~ our results show that MBM excels in reconstructing local winter mass balance, which indicates implicit downscaling and bias correction of meteorological variables ~~excel in reconstructing local winter mass balance~~ (Figs. 6 and 7). This suggests, in line with other findings (Guidicelli et al., 2023), that ML models are valuable tools to assess spatio-temporal biases in precipitation estimates in mountain regions.

2. Model Transparency: XGBoost's complexity warrants feature importance analysis or partial dependence plots to clarify key drivers of predictions, enhancing interpretability.

We added a new appendix (Appendix C: Feature importance, please see additions below) with a discussion of feature importance based on different methods, including two new figures showing overall feature importance in terms of weight and gain on the trained model, and monthly permutation feature importance on the test dataset. The analysis provides additional insights into the importance of different monthly features in seasonal and annual predictions, and we believe that many of the findings support the current assessment of MBM's capabilities.

Appendix C: Feature importance

We performed a feature importance analysis on MBM to investigate the importance of different variables on MBM's performance. Since feature importance is complex to interpret and is not adequately represented by any single metric, we based our assessment on different metrics. We calculated weight and gain scores, which represent the total number of times a feature is used in splitting the data in a node and the average improvement in model performance (sum of loss change for each split) in splits where a feature is used, respectively. To complement this analysis, we computed monthly permutation importance for each feature. This involves consecutively permuting (shuffling) the values of each feature, breaking the relationship between the feature and prediction, and assessing the resulting change in model performance. For a given feature and month, the performance change thus represents the effect of feature permutation on the seasonal and annual predictions.

Temperature is overall the most frequently used feature in the trained model (t2m; Fig. C1a). It also scores highest in terms of gain, followed by elevation difference and downward surface solar radiation (elev_diff and ssrd, respectively; Fig. C1b). The importance of temperature according to the weight and gain scores is not surprising given that both accumulation and melt are strongly influenced by this variable. The combination of lower gain but relatively similar weight of the remaining features may suggest that these are generally used at lower levels of the tree structures, e.g. to distinguish between smaller variability in mass balance for points on the same glacier.

Considering monthly permutation feature importance, elevation difference is an important feature in all months (Fig. C2). In mid-winter (Dec–Mar) total precipitation is the most important feature (tp; Fig. C2l and a–c) and also relatively important compared to other meteorological variables in the transition months April, October and November (Fig. C2d, j and k, respectively). This aligns with the fact that solid precipitation is the main contribution to accumulation on glaciers in Norway. In addition, precipitation is likely a key variable in explaining the substantial differences in winter mass balance rates across climatic regions in Norway.

Temperature is the main influence on model performance in the summer season (May–Sep; Fig. C2e–i). In addition, downward solar radiation and forecast albedo are important in May and June (Fig. C2e and f, respectively), which is consistent with the onset of snowmelt and subsequent changes in albedo. Although albedo is coarsely resolved, it may provide larger-scale geographical information about changes in snow cover, which may be why it is also considered somewhat important in mid-winter months. The transition months April and October show less clear importance between meteorological variables (Fig. C2d and j, respectively). This may be because the timing of transitions between seasons varies with latitude, e.g. glaciers in northern Norway may receive a fair amount of snow in April and October.

We caution against placing too much emphasis on the specific details of the feature importance analysis. For example, when assessing permutation importance, correlated features (i.e. skyview factor and slope) may appear to be less important since, even if one feature is permuted, the model can rely on a second correlated feature. However, the main findings of the feature importance analysis presented here are consistent across metrics and physically meaningful with respect to the main meteorological drivers of mass balance on Norwegian glaciers.

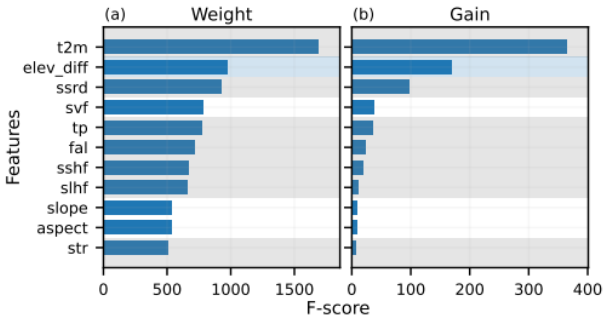

Figure C1. Feature importance on trained model in terms of (a) weight and (b) gain (t2m: 2 m air temperature, sshf: surface sensible heat flux, slhf: surface latent heat flux, ssrd: downward surface solar radiation, fal: forecast albedo, str: net surface thermal radiation, tp: total precipitation, elev_diff: elevation difference between climate model and stake, svf: skyview factor). Weight represents the total number of times a feature is used to split the data, summed over all trees. Gain represents the average improvement in model performance (sum of loss change for each split over all trees) in splits which use the given feature. Shaded grey, white and blue background indicates meteorological features, topographical features and elevation difference feature, respectively.

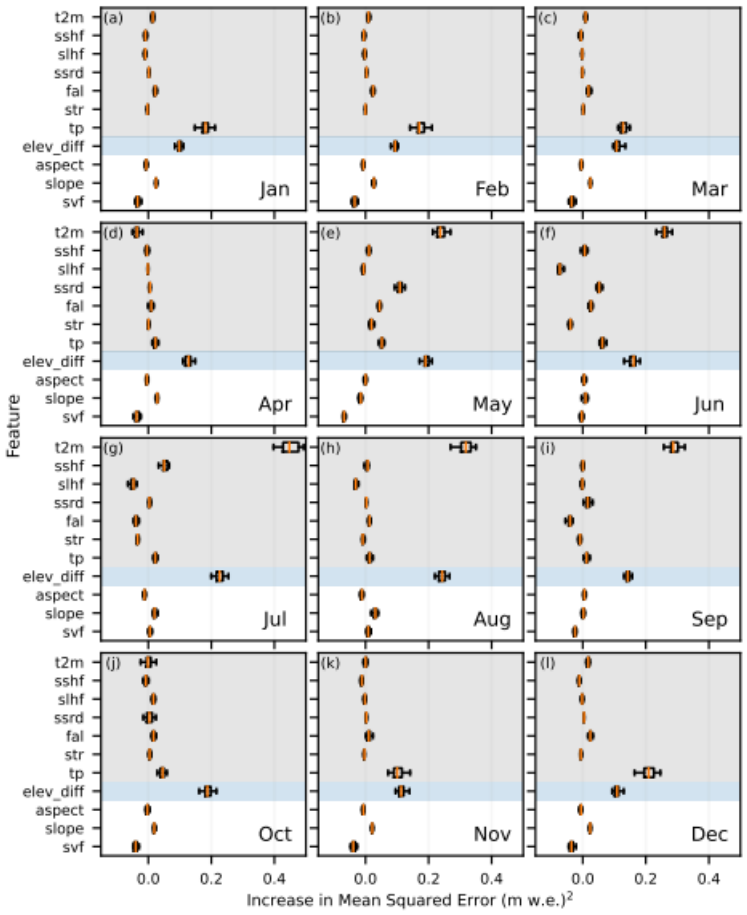

Figure C2. Monthly permutation feature importance on the test dataset (t2m: 2 m air temperature, sshf: surface sensible heat flux, slhf: surface latent heat flux, ssrd: downward surface solar radiation, fal: forecast albedo, str: net surface thermal radiation, tp: total precipitation, elev_diff: elevation difference between climate model and stake, svf: skyview factor). Each feature is permuted on a monthly basis and the resulting change in model performance is computed with respect to the seasonal and annual targets. Shaded grey, white and blue background indicates meteorological features, topographical features and elevation difference feature, respectively.

3. Uncertainty Quantification: While measurement uncertainties (0.08–0.26 m w.e. a−1) are noted, their impact on model outputs is unclear. A sensitivity analysis would strengthen confidence in predictions.

Measurement uncertainties are noted in Appendix A to underline our confidence in the quality of the dataset. Since these uncertainties are relatively small, we do not expect them to have a major impact on model results or the conclusions in this study. However, using other datasets that may be afflicted with substantial uncertainties, such as geodetic mass balance based on remote sensing, considering uncertainty in observations could be increasingly important (e.g., using uncertainty-aware learning; Diaconu et al. (2024)). We added a comment on this in Section 6.3.2:

for the reliability of the data by weighing the observations in the loss function according to their confidence levels or using uncertainty-aware learning (Diaconu and Gottschling, 2024). Incorporating diverse and complementary datasets could provide
530 reconciled estimates of glacier mass balance across multiple observational datasets.

4. Global Applicability: Testing MBM in diverse regions like the Alps or Himalayas would broaden its relevance. A discussion of transferability challenges would be valuable.

We agree that testing MBM in other regions would clarify its potential and limitations. We partly discuss the transferability challenges already on L500-503. Since MBM is specifically trained for Norwegian glaciers in the current study, we do not expect it to perform equally well in other regions where conditions differ. We thus expect the transferability of the current application of MBM (trained on Norwegian glaciers) to be limited. We would expect that for larger regions, it would be preferable to retrain MBM using additional data. However, in applications it would be interesting to investigate the limits of the models transferability to clarify how it can be expected to perform in regions with limited data. We consider this to be out of the scope of the current study, but ongoing research using MBM is aimed at addressing this particular issue. We have added the following to expand on this discussion and highlight needs for future research:

period. However, since the model is trained on meteorological conditions specific to Norway and designed for interpolation within this context, we expect its performance be limited in regions with significantly different climates. Future research into the transferability of ML approaches could clarify the extent of such limitations, for example by testing MBM on glaciers in other regions. For larger, or climatically different regions, we expect MBM to benefit from additional training data. Since in
520 situ observations are not readily available for many regions, the diversity of spatio-temporal analogues and extent of MBM's generalisation capabilities on larger scales remain to be investigated.

5. Future Directions: The mention of remote sensing data is promising but vague. Specifying datasets (e.g., satellite-derived albedo or surface temperature) would clarify future enhancements.

Here, we are referring to mass balance observations from other sources, such as satellite-derived geodetic mass balance. We have amended the sentence to specify this. We already provide a specific example on line 509-511, but have now also added references to additional datasets: "On the other hand, the purely data-driven nature of

ML approaches makes them uniquely suited to take advantage of the increasing availability of remote sensing-based mass balance datasets (e.g., Belart et. al (2017), Pelto et al. (2019), Hugonnet et al. (2021), Falaschi et al. (2023)).” Please see references at the bottom of the document.

6. Presentation Polish: Minor typos and awkward sentences, particularly in Appendix A, need correction. Additionally, Fig. 10 requires clearer labels for improved readability.

Please see our response to the comments on Appendix A below. We have checked the labels in Fig. 10, but are not sure how these need to be clarified and have thus not made any changes to these.

Detailed Review of Appendix A

Appendix A details the data quality and cleaning processes for the MBM dataset, critical for establishing its reliability. It describes the handling of 4,201 point mass balance measurements from the NVE database (accessed 12 October 2022), including the removal of erroneous entries and verification of stake locations. Below, I identify specific typos and awkward sentences with approximate line numbers (based on sequential sentence or paragraph counting) and provide broader recommendations to enhance clarity.

Identified Typos and Awkward Sentences

1. Line 570: “The total contribution of such uncertainties have been quantified 0.080.26 m w.e. a−1…”
   • Issue: Subject-verb agreement error; “have” should be “has” for the singular subject “The total contribution.”
   • Suggestion: Revise to “The total contribution of such uncertainties has been quantified as 0.08–0.26 m w.e. a−1 for five glaciers in our dataset.”
   Done.

2. Line 575: “Prior to training MBM, we performed a thorough cleaning and quality check… including removal of erroneous values and points with missing location, and a quality check of stake locations.”
   • Issue: Redundant use of “quality check.”
   • Suggestion: Streamline to “Prior to training MBM, we performed thorough cleaning and quality checks on the raw point mass balance dataset (4,201 entries, NVE database, accessed 12 October 2022), removing erroneous values, points with missing locations, and verifying stake location accuracy.”
   We removed the redundant “quality check” and revised the sentence according to point 1 under “Additional recommendations for Appendix A”. Please see our reply to this comment.

3. Line 578: “Approximate locations are based on the approximate position and elevation of a given stake ID…”
   • Issue: Repetition of “approximate” is awkward.
   • Suggestion: Revise to “Approximate locations are derived from the estimated position and elevation of a given stake ID, whereas exact locations use GPS-measured position and elevation at the time of measurement.”
   We revised the sentence to: “The approximate location is based on the estimated position and elevation of a given stake ID, whereas the exact location is the actual position and elevation of the stake at the time of measurement (e.g. measured using GPS).”

4. Line 581: “Seven and 23 entries that were missing both exact and approximate elevation or geographical coordinates, respectively, were removed…”
   • Issue: Ambiguous phrasing regarding elevation and coordinates.
   • Suggestion: Clarify to “Seven entries missing both exact and approximate elevations and 23 entries missing both exact and approximate geographical coordinates were removed from the training dataset.”
   We thank the reviewer for the suggestion and have amended the sentence accordingly.

5. Line 585: “The mean ± standard deviation of the absolute difference between the exact and approximate coordinates and elevations is 166 ± 498 m and 24 ± 71 m, respectively.”
   • Issue: Dense phrasing combines measurements, reducing clarity.
   • Suggestion: Split to “The mean ± standard deviation of the absolute difference between exact and approximate coordinates is 166 ± 498 m, while for elevations, it is 24 ± 71 m.”
   We thank the reviewer for the suggestion and have amended the sentence accordingly.

6. Line 589: “For stake locations where both summer, winter and annual mass balance measurements were available…”
   • Issue: List lacks an Oxford comma for clarity.
   • Suggestion: Revise to “For stake locations where summer, winter, and annual mass balance measurements were all available for a given year…”
   Done.

Additional Recommendations for Appendix A

1. Improve Transitions: The shift from uncertainties to data cleaning is abrupt. Add a bridging sentence, e.g., “Ensuring dataset quality is crucial for MBM’s accuracy, leading to the following cleaning procedures.”

To improve the transition, we modified the original sentence to: "To ensure the quality of MBM's training data, we performed a thorough cleaning and quality check of the raw point mass balance dataset (4201 entries, NVE database accessed on 12 October 2022) prior to training MBM. This consisted of removing erroneous values and points with missing locations, and verifying stake locations."

2. Define Technical Terms: Define "point mass balance" (e.g., "measurements of mass change at specific glacier locations") in a footnote or glossary for accessibility.

This is defined in the introduction (line 35) and we do not find it necessary to introduce the term again here. No changes were made.

3. Clarify Data Sources: Specify that NVE is the Norwegian Water Resources and Energy Directorate to aid international readers.

This is defined on line 93. We find it unnecessary to repeat again in the appendix. No changes were made.

4. Quantify Cleaning Impact: State the total entries removed, e.g., "After cleaning, the dataset was reduced from 4,201 to 4,170 stake locations (99.3% retained)."

The number of entries after cleaning is already summarized on L592. No changes were made.

5. Explain Coordinate Conversion: Justify the UTM to latitude/longitude conversion, e.g., "This conversion ensured compatibility with MBM's input requirements."

We amended the sentence to: "Finally, we converted geographical coordinates from UTM to latitude and longitude format for compatibility with the feature datasets."

6. Justify Erroneous Values: Explain the removal of the 9.99 m w.e. measurement, e.g., "This value was unrealistically high for typical regional winter mass balances."

We amended the sentence to: "One measurement with erroneous winter mass balance (unrealistically high; 9.99~m~w.e.) was removed."

7. Quantify Corrections: If available, note the number of rounding error corrections, e.g., "In [X] instances, annual mass balances were corrected by summing summer and winter components."

We included the number of corrections and magnitudes of rounding errors in the following sentence: "For stake locations where both summer, winter, and annual mass balance measurements were available for a given year, we corrected for rounding errors where these were present by replacing annual mass balance values by the sum of seasonal values (magnitudes between 0.01–0.03 m w.e.; 255 instances)."

Conclusion

The “egusphere-2025-1206” paper is a compelling contribution to glaciology, with MBM offering high-resolution, accurate predictions for glacier mass balance. Its robust dataset, effective methodology, and practical applications make it a valuable tool. Minor revisions, including addressing typos in Appendix A, improving data resolution, and enhancing model transparency, will further strengthen its impact. Appendix A effectively supports the dataset’s reliability but can be polished with clearer transitions, defined terms, and quantified impacts. These changes require minimal effort but will significantly enhance the paper’s clarity and global relevance.

Recommendations

• Accept with Minor Revisions.

• Specific Actions:

- Correct the six typos and awkward sentences in Appendix A as suggested.
- Explore higher-resolution weather data or downscaling for smaller glaciers.
- Add feature importance or partial dependence plots for model transparency.
- Conduct a sensitivity analysis to quantify uncertainty impacts.
- Discuss testing MBM in other regions for global applicability.
- Specify remote sensing datasets (e.g., albedo, surface temperature) for future work.
- Add a transitional sentence in Appendix A between uncertainties and cleaning.
- Define “point mass balance” in a footnote or glossary.
- Clarify NVE as the Norwegian Water Resources and Energy Directorate.
- Quantify total entries removed during cleaning (e.g., 4,201 to 4,170).
- Justify UTM to latitude/longitude conversion and the 9.99 m w.e. removal.
- Note the number of rounding error corrections, if available.
- Improve Fig. 10 labels for clarity.
- Include missing DOIs or URLs in the reference section.

Please see our replies above to the overall assessment and to each of the comments.

Additional references:

Belart, J. M. C., Berthier, E., Magnússon, E., Anderson, L. S., Pálsson, F., Thorsteinsson, T., Howat, I. M., Aðalgeirsdóttir, G., Jóhannesson, T., and Jarosch, A. H.: Winter mass balance of Drangajökull ice cap (NW Iceland) derived from satellite sub-meter stereo images, The Cryosphere, 11, 1501–1517, https://doi.org/10.5194/tc-11-1501-2017, 2017.

Diaconu, C.-A. and Gottschling, N. M.: Uncertainty-Aware Learning With Label Noise for Glacier Mass Balance Modeling, IEEE Geoscience and Remote Sensing Letters, 21, 1–5, https://doi.org/10.1109/LGRS.2024.3356160, 2024.

Falaschi, D., Bhattacharya, A., Guillet, G., Huang, L., King, O., Mukherjee, K., Rastner, P., Yao, T., and Bolch, T.: Annual to seasonal glacier mass balance in High Mountain Asia derived from Pléiades stereo images: examples from the Pamir and the Tibetan Plateau, The Cryosphere, 17, 5435–5458, https://doi.org/10.5194/tc-17-5435-2023, 2023.

Hugonnet, R., McNabb, R., Berthier, E., Menounos, B., Nuth, C., Girod, L., Farinotti, D., Huss, M., Dussaillant, I., Brun, F., and Kääb, A.: Accelerated global glacier mass loss in the early twenty-first century, Nature, 592, 726–731, https://doi.org/10.1038/s41586-021-03436-z, 2021.

Pelto, B. M., Menounos, B., and Marshall, S. J.: Multi-year evaluation of airborne geodetic surveys to estimate seasonal mass balance, Columbia and Rocky Mountains, Canada, The Cryosphere, 13, 1709–1727, https://doi.org/10.5194/tc-13-1709-2019, 2019.