Dear Reviewer,

Thank you for your comments concerning our manuscript entitled "Modelling glacier-wide annual mass balance of continental-type glaciers in China using a deep neural network" (ID: egusphere-2026-333). These comments are very helpful for the revision and improvement of our manuscript and provided important guiding significance for our research. We have carefully revised the manuscript based on the comments, and our point-to-point responses to your comments are shown below. The comments are in **black,** our responses are in **blue** and actions in **green**.

Response to Reviewer:

1. Comment: The manuscript is generally readable but would benefit from careful editing. Several acronyms are defined more than once, and a number of passages repeat information already stated earlier in the text. Some material also appears in sections where it does not belong, as flagged in the commented text, and would read more naturally if reorganised.

   Response: We sincerely thank the reviewer for the careful reading of our manuscript and for the detailed and constructive comments regarding clarity, organization, and writing quality. We have carefully reviewed all annotations and suggestions provided in the annotated manuscript. In the following, we address each comment point-by-point and describe the corresponding revisions made in the manuscript.

   (1) Comment on Page 3:

      (a) More recently see the papers on the Mass Balance Machine: Sjursen et al. 2025 and van der Meer et al. 2026 have developed xgboost & deeper neural networks for glacier mass balance in Europe based on energy balance variables.

      Response: We thank the reviewer for this helpful comment. We have carefully reviewed the recommended studies (Sjursen et al., 2025; van der Meer et al., 2026) and have incorporated them into the Introduction. We note that the study by van der Meer et al. (2026) is currently available as a preprint in ESS Open Archive. Following the journal's citation guidelines for unpublished or preprint articles, we have cited it using the format "van

der Meer et al., 2026" in the text, with the full reference provided as: Marijn van der Meer, Harry Zekollari, Alban Gossard, et al. (2026), ESS Open Archive, DOI: 10.22541/essoar.176894225.56571551/v1.

Action: We have revised the Introduction:

simulating point glacier mass balance in the Alps. Similarly, Ren et al. (2024) examined the applicability of various ML approaches for modeling annual glacier-wide mass balance of both maritime and continental glaciers in High Mountain Asia. Sjursen et al. (2025) proposed the Mass Balance Machine (MBM), an XGBoost-based model that leverages sparse observations to achieve transferable, regional-scale glacier mass balance predictions and improves seasonal estimates compared to current large-scale evolution models. van der Meer et al. (2026) further developed the Mass Balance Machine (MBM) using a Long Short-Term Memory (LSTM) architecture, showing that incorporating temporal dependencies enhances predictive skill for winter and annual mass balance and enables robust generalization to unseen glaciers across different regions. Both studies

(b) This too me feels already too detailed for the introduction and more like belonging to methods.

Response: We thank the reviewer for this helpful comment. As suggested, we have revised the text to make it more concise and appropriate for the Introduction by removing excessive methodological details and emphasizing the main contributions and relevance of these studies.

Action: We have revised the Introduction:

it is therefore essential that the study glaciers belong to the same type. In China, continental glaciers account for approximately 77.8% of total glacier area and dominate the long-term observations (Li et al., 2018). Accordingly, ten continental glaciers distributed across different mountain ranges in western China with publicly available long-term mass balance observations were selected as representative samples to investigate the potential of data-driven approaches under extremely small-sample conditions. A lightweight deep neural network framework is developed to simulate glacier-wide annual mass balance by integrating multi-source predictors, including meteorological variables at multiple temporal scales, terrain factors, and summer surface albedo. The framework integrates feature selection for dimensionality reduction of meteorological predictors (using Pearson correlation analysis and permutation-based feature importance from the random forest algorithm) and adopts two

dataset construction strategies (reduced-sample and imputation) to ensure temporal consistency, with a focus on overfitting risk, model stability, predictive performance, and spatiotemporal generalization. This study aims to improve the reconstruction of glacier-wide annual mass balance and to provide a robust evaluation of the applicability and generalization capability of deep learning approaches for large-scale glacier studies under data-scarce conditions. ~~Accordingly, ten continental glaciers~~

(2) Comment on Page 5: For this and Zhang et al. 2021 did you take the glacier-wide mass balance calculated from raw measurements or the reconstructed mb ? Because I know that they used a temperature-index model to reconstruct mb for a bigger time-frame and that would not be a good mass balance target for the

model :) based on the years you mention here I guess not but just double-checking.

Response: We thank the reviewer for this helpful clarification. In this study, the glacier-wide annual mass balance used as the target variable is derived from in situ glaciological measurements rather than reconstructed datasets. For Zhang et al. (2021), although a temperature-index model was used to reconstruct the mass balance of Haxilegen No. 51 glacier for the period 1999—2015, the study also provides a subset of mass balance data based on direct field observations. The data used in our study correspond exclusively to these observed glaciological mass balance records, rather than the model reconstructed series.

(3) Comment on Page 6: How many measurements of the 180 are you then left with? Also in 2.3.2 you're still talking about taking the data from 2000 so this is inconsistent.

Response: We thank the reviewer for this important comment. The apparent inconsistency likely stems from insufficient clarity in our original description. In fact, the post-2000 data mentioned in the two sections serve different purposes. In Section 2.2.1, the focus is on analyzing the spatiotemporal variability of glacier-wide annual mass balance. Although the full dataset comprises 180 observations with varying temporal coverage across glaciers, most records are concentrated after 2000. Therefore, only data from 2000 onward are used to ensure comparability among glaciers, resulting in a subset of 109 observations for this analysis. In contrast, Section 2.3.2 concerns model input construction. The restriction to post-2000 data is due to data availability, as the summer surface albedo used as an input variable is only available from 2000 onward. To ensure temporal consistency among all input features, two dataset construction strategies were developed: a reduced-sample strategy that restricts the training data to samples after 2000, and an imputation-based strategy that retains earlier samples by addressing missing input features. To avoid ambiguity, we have revised the relevant descriptions in Section 2.2.1.

Action: We have revised the Section 2.2.1:

sources and observation periods—is provided in Table 1.

~~Because most observations were collected after 2000, this study used data from 2000 onward to minimize temporal inconsistencies and analyze the spatiotemporal variability of annual mass balance.~~ Given that the majority of observations are concentrated after 2000, this study focuses on data from 2000 onward to ensure comparability among glaciers when analyzing the spatiotemporal variability of annual mass balance, resulting in a subset of 109 observations used for this analysis. Temporal and spatial heterogeneity among the glaciers is illustrated in the boxplots and heatmap (Fig. 2). The median mass balance values of all glaciers except Muztag Ata No. 15 are below 0 m w.e., indicating a predominant ablation trend during the

(4) Comment on Page 7: more recently: van der Meer et al. 2026, Sjursen et al. 2025, van der Meer et al. 2025.

Response: We thank the reviewer for this suggestion. The recommended recent studies (van der Meer et al., 2026; Sjursen et al., 2025; van der Meer et al., 2025) have been carefully reviewed and incorporated into the revised manuscript.

Action: We have revised the Section 2.2.2:

climate variables from 1950 to the present with a spatial resolution of ~9 km (Muñoz-Sabater et al., 2021). This dataset was chosen for its comparatively high spatial resolution and its proven utility in glacier mass balance simulations (Anilkumar et al., 2023; Ren et al., 2024; Arndt et al., 2023; Draeger et al., 2024; Sjursen et al., 2025; van der Meer et al., 2025; van der Meer et al., 2026). The selection of climatic variables was guided by the physical mechanisms governing glacier ablation and accumulation processes (Réveillet et al., 2017; Gabbi et al., 2014). Nineteen variables were extracted for each glacier's

(5) Comment on Page 8: Per glacier-wide MB target?

Response: We thank the reviewer for this clarification. The 271 climatic variables are defined for each glacier-wide annual mass balance (MB) target, i.e., each MB observation is associated with a corresponding set of meteorological predictors. Given the inherent uncertainties in ERA5-Land data, a relatively comprehensive set of meteorological variables related to glacier accumulation and ablation processes was considered to better capture the relevant climatic controls. The manuscript has been revised to clarify this point. In addition, we agree that the original wording of Section 2.2.4 may be misleading, as these six variables refer only to the geometric and topographic features, which constitute one component of the full set of model input variables. The sentence has been revised to clearly indicate that these features represent a subset of the overall input variables.

Action: We have revised the Section 2.2.2:

each variable. Cumulative Positive Degree Days (CPDD), a key indicator of melt energy (Braithwaite et al., 2000), were calculated from temperature data and included as an additional predictor. ~~In total, 271 climatic variables were generated.~~ In total, 271 climatic variables were generated for each glacier-wide annual mass balance target, representing multi-temporal meteorological conditions associated with each observation.

Action: We have revised the Section 2.2.4:

Elevation Model Version 3 (ASTGTM_003) provided by NASA's Land Processes Distributed Active Archive Center (LP DAAC). Mean elevation, slope, and aspect were derived for each glacier, while glacier longitude, latitude, and area were obtained from RGI 6.0. ~~In total, six geometric and topographic features were included as input variables for the model.~~ In total, six geometric and topographic features were included to represent the terrain factors of the model input features.

(6) Comment on Page 9: Ok so this process has a data leakage problem: feature selection was performed on the full dataset prior to any cross-validation split. Specifically, both the Pearson correlation filtering and the Random Forest importance ranking, were conducted using all available samples, including those later designated as validation data in each crossvalidation fold. This constitutes data leakage: the features chosen for model training were implicitly selected based on information from the validation folds, meaning the crossvalidation no longer provides a truly independent assessment of generalization performance. The issue is compounded by the fact that the Random Forest used for importance ranking is itself a learned model, capable of capturing nonlinear patterns across the full dataset rather than just linear associations as in the Pearson step.

Response: Thank you for pointing out the potential data leakage issue in our original feature selection procedure. We have addressed this issue by embedding the entire feature selection procedure within each cross-validation framework. Specifically, both the Pearson correlation filtering and the RF-based importance ranking are now performed independently using only the training subset in each fold, ensuring that no information from the validation data is used during feature selection. This ensures a strict separation between training and validation data throughout the entire modeling pipeline. In addition, while the original manuscript retained the top 20 predictors, we have refined this to the top 10 in the revised version. The associated changes in model performance and stability are discussed in detail in our response to the "Comment on Page 15". To

examine the stability of the feature selection process under this revised framework, we provide the selection frequency of each feature across random cross-validation folds (Fig. 3(b)).

Action: The corresponding description in Section 2.3.1 has been updated to reflect these changes. We believe this modification effectively eliminates the data leakage issue and provides a more rigorous and reliable assessment of model generalization performance.

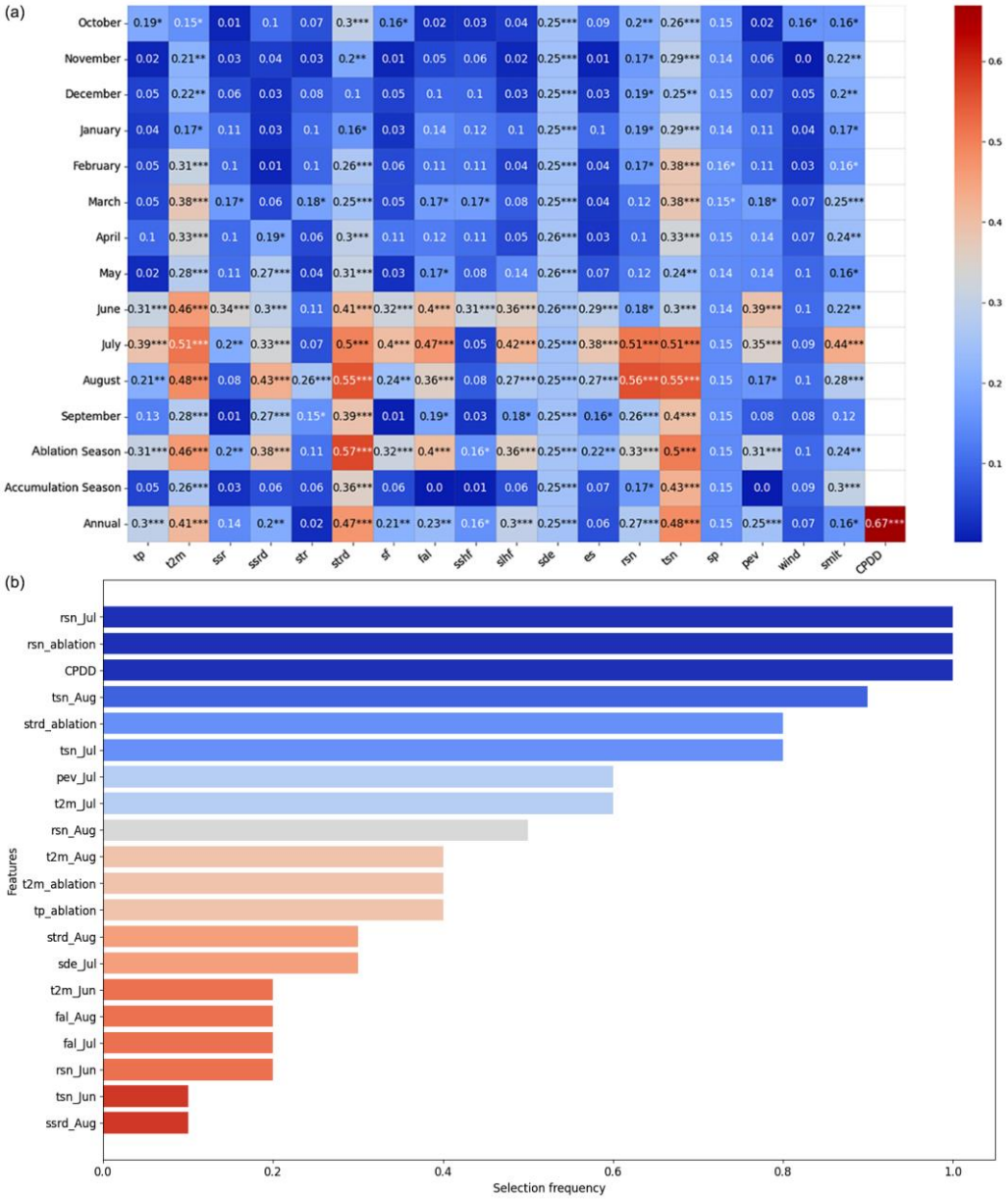


Figure 3 Feature selection results based on Pearson correlation coefficients and RF-derived variable importance. **(a)** Pearson correlation coefficients between climatic variables and glacier-wide annual mass balance. One, two, and three asterisks indicate significance at $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively. **(b)** Selection frequency of the top predictors across cross-validation folds.

al., 2003; Liang et al., 2018; Sun et al., 2018). ~~Based on the Pearson correlation analysis, 145 climate variables showing statistically significant relationships (P < 0.05) were retained for RF-based feature importance ranking. As illustrated in Fig. 3(b), the top 20 predictors were ranked in descending order of relative contribution.~~ Based on Pearson correlation analysis, 145 climate variables with statistically significant relationships ($P < 0.05$) were retained, from which the top 10 predictors with the highest relative importance were further selected using the RF model. The stability of feature selection process was

evaluated by computing the selection frequency of each predictor across random cross-validation folds (Fig .3(b)). The results indicate that a core subset of variables (e.g., CPDD, rsn, and tsn) is consistently selected in more than 90% of the folds, whereas others exhibit greater variability, likely due to the limited sample size and the heterogeneity of glacier–climate interactions. Overall, this strategy effectively identifies stable and physically meaningful predictors, while also reflecting the inherent uncertainty associated with data-driven modelling under sparse observational conditions. Applying the same feature-selection process to the reduced109-sample dataset yielded 10~~20~~ meteorological predictors (see Section 2.3.2).

(7) Comment on Page 10: Here when a variable is noted as "melt" do you mean it's the average/sum over the ablation season? Btw in the text you use "ablation season" and here in the figure "melt season". Make sure to be consistent so that the reader can easily follow.

Response: We thank the reviewer for this helpful comment. In Figure 3, the term 'melt' refers to meteorological variables aggregated over the ablation season (May-September), represented as either seasonal averages or cumulative values depending on the variable. We acknowledge that the use of both "melt season" and "ablation season" may lead to ambiguity. To improve consistency and readability, we have replaced "melt" and "melt season" with "ablation" and "ablation season" throughout the manuscript and figures.

(8) Comment on Page 11:

(a) This difference is not clear in the next paragraph. Do you just mean the additional binary input?

Response: We thank the reviewer for this clarification. The structural difference between the two dataset construction strategies is indeed limited to the input layer. Specifically, for the imputation-based strategy, an additional binary indicator variable is introduced to explicitly represent the presence or absence of albedo data, whereas this variable is not included in

the reduced-sample strategy. We acknowledge that this distinction was not clearly stated in the original manuscript. The relevant descriptions in Sections 2.3.2 and 2.3.3 have been revised to explicitly clarify this point.

Action: We have revised the Section 2.3.2:

2000, reducing the dataset to 109 samples. As a result, the deep neural network architectures corresponding to the two dataset construction~~s~~ strategies differ slightly in their input configuration, ~~; the structural configurations are~~ as described in Section 2.3.3. Static features—including longitude, latitude, glacier area, mean elevation, slope, and aspect—were incorporated into each glacier's time-series feature set as auxiliary predictors. The 20 selected meteorological variables, together with summer

Action: We have revised the Section 2.3.3:

layers, and one output layer. ~~The input layer contains a number of neurons equal to the total selected predictors, while the hidden layers follow a progressively compressed structure (40, 20, 10, and 5 neurons), which enhances abstraction, stabilizes feature extraction, and reduce overfitting. The output layer contains a single neuron to generate regression-based predictions of annual glacier mass balance. Missing albedo values were imputed using the mean of available observations, and a~~

~~corresponding binary indicator variable was introduced to denote the presence or absence of albedo data. Both the filled albedo values and the corresponding missing indicator were incorporated into the neural network as input features.~~ The input layer contains neurons corresponding to the selected predictors. For the imputation-based strategy, missing albedo values are replaced by the mean and an additional binary indicator of missingness is included as an input feature, whereas the reduced-sample strategy uses only complete observations without this indicator. The hidden layers follow a progressively compressed structure (40, 20, 10, and 5 neurons), which enhances abstraction, stabilizes feature extraction, and reduces overfitting. The output layer contains a single neuron to generate regression-based predictions of glacier-wide annual mass balance. To further

(b) I think this is too much detail.

Response: We thank the reviewer for this suggestion. With respect to the opening description of Section 2.3.3, we acknowledge that the original text was somewhat overly detailed. Accordingly, the text has been revised to improve conciseness.

Action: We have revised the Section 2.3.3:

▪ **2.3.3 Deep neural network construction**

~~Artificial neural networks (ANNs) are nonlinear statistical models inspired by biological neural systems, capable of storing experiential information and applying it to interpret and solve complex problems (Hastie et al., 2009). A neural network is generally characterized by three essential components: 1) network architecture, defined by neuron connectivity and the number of layers; 2) the optimizer, which iteratively updates trainable parameters based on a specified loss function; and 3) activation functions, which introduce nonlinear transformations and enhance the model's expressiveness (Fausett, 2006). A typical ANN comprises multiple interconnected layers—an input layer, several hidden layers, and an output layer—through which raw data are progressively transformed into increasingly abstract representations before generating task-specific predictions (Goodfellow et al., 2016).~~ ANNs are nonlinear statistical models inspired by biological neural systems, capable of storing experiential information and applying it to interpret and solve complex problems (Hastie et al., 2009). A typical ANN comprises an input layer, multiple hidden layers, and an output layer, enabling the progressive transformation of input data into higher-level representations for task-specific predictions (Goodfellow et al., 2016). A neural network is generally characterized by its architecture, activation functions, and an optimization algorithm that updates its trainable parameters (Fausett, 2006). The feed-forward fully connected neural network (FF-FCNN) developed in this study represents a lightweight

(c) This acronym has been defined before. Make sure to also check the others.

Response: We thank the reviewer for this suggestion. we have carefully checked all acronyms throughout the manuscript and ensured that each is defined only once.

(9) Comment on Page 12:

(a) Start new paragraph.

Response: We thank the reviewer for this suggestion. The manuscript has been revised accordingly.

Action: We have revised the Section 2.3.3:

contains neurons corresponding to the selected predictors. For the imputation-based strategy, missing albedo values are replaced by the mean and an additional binary indicator of missingness is included as an input feature, whereas the reduced-sample strategy uses only complete observations without this indicator. The hidden layers follow a progressively compressed structure (40, 20, 10, and 5 neurons), which enhances abstraction, stabilizes feature extraction, and reduces overfitting. The output layer contains a single neuron to generate regression-based predictions of glacier-wide annual mass balance.

To further improve robustness and generalization, multiple optimization strategies were incorporated. Gaussian noise (standard deviation = 0.1) was added to the input layer to emulate data-augmentation effects and enhance resilience to input variability. L1 regularization constrained the effective number of active parameters and mitigated overfitting. Early stopping was employed to halt the training process when validation performance ceased to improve. The combined use of He-normal weight initialization and the LeakyReLU activation function facilitated stable gradient propagation and accelerated convergence. Batch normalization was applied before each activation function to stabilize intermediate feature distributions, thereby

(b) I'm a bit confused about this RF model. Is that the one used to make the feature selection?

Response: We thank the reviewer for this comment. The RF model referred to here is not used for feature selection. Feature selection is conducted separately prior to model training. The RF model is instead trained using the

selected predictors and evaluated under the same random cross-validation scheme. It serves as a benchmark model to compare predictive performance with the proposed FF-FCNN.

(c) Acronym already defined.

Response: We thank the reviewer for this suggestion. we have carefully checked all acronyms throughout the manuscript and ensured that each is defined only once.

(d) van der Meer et al 2025, Sjursen et al 2025.

Response: We thank the reviewer for this suggestion. The recommended recent studies (Sjursen et al., 2025; van der Meer et al., 2025) have been incorporated into the revised manuscript.

Action: We have revised the Section 2.3.4:

2016). RF models excel at handling complex variable relationships and capturing nonlinear characteristics and they have proven valuable for glacier mass balance prediction (~~Ren et al., 2024;~~ Anilkumar et al., 2023; Ren et al., 2024; Sjursen et al 2025; van der Meer et al 2025). In this study, the RF model was implemented using the sklearn package in Python to simulate glacier-wide annual mass balance within a regression framework. GridSearchCV, combined with cross-validation, was

(e) the test set (make this clear).

Response: We thank the reviewer for this suggestion. The description has been clarified by explicitly indicating that the excluded fold is used as the test set in each iteration.

Action: We have revised the Section 2.3.5:

evaluate the model's ability for data reproduction as well as its temporal and spatial prediction performance. Fig. 4(d) displays the cross-validation strategies schematically. Each strategy first split the data into k folds and then iteratively trains the model k times, each time using all folds except one for training, with the remaining fold used as the test set. In the random cross-

(f) Due to the differing observation periods among glaciers, the test set in

Response: We thank the reviewer for highlighting this point. The sentence has been revised to clarify that, due to differing observation periods among glaciers, the test sets in temporal cross-validation may not include all glaciers, and model evaluation is therefore conducted on temporally varying subsets rather than the full glacier ensemble.

Action: We have revised the Section 2.3.5:

validation, samples are split into folds by their observation year, with each fold representing a distinct time period. Due to the differing observation periods among glaciers, the test set in each fold does not necessarily include data from all glaciers, meaning that model evaluation in temporal cross-validation is performed on subsets of glaciers rather than the full glacier ensemble. The R^2, MAE, and RMSE metrics are used to evaluate model performance and are calculated using the following

(10) Comment on Page 14:

(a) This is repetitive and has been said twice now already. If you really want to keep it, make it a short reminder only.

Response: We thank the reviewer for this comment. We agree that this description was repetitive. The text has been revised to provide a concise reminder and to refer to the earlier section for detailed information.

Action: We have revised the Section 3.1:

In this study, the models were driven by meteorological, albedo, and topographic data. Because albedo data are available only from 2000 onwards, two strategies were considered for building the training dataset (see Section 2.3.2 for details). ~~In the first strategy, all 180 samples were retained by imputing missing albedo values using mean substitution and introducing a corresponding binary indicator variable as an additional input feature in the FF-FCNN. The second strategy excluded samples prior to 2000, reducing the dataset to 109 samples.~~ Given the limited amount of data, it was necessary to evaluate potential

(b) agreement between modeled and observed values. 2hese findings suggest that masking missing albedo values does not fully eliminate the negative effects of incomplete features. Instead, it increases data noise and introduces distributional inconsistencies, ultimately degrading model performance. In contrast, restricting the training dataset to samples with complete and homogeneous feature sets yields a cleaner input, enabling the FF-FCNN model to learn more robust and generalizable relationships. 3herefore, all subsequent model training was conducted on the reduced-sample dataset.

Response: Thank you for highlighting this point. We agree that our original wording overstated the impact of missing albedo masking on performance differences. In the revised manuscript, we have moderated this statement to avoid implying a direct causal relationship and instead describe the observed differences more cautiously.

We also acknowledge that selecting the reduced-sample dataset for all subsequent analyses may introduce implicit optimization bias. In response, we have revised both the experimental design and the presentation. Specifically, rather than relying exclusively on a single dataset construction strategy, we now report and compare the performance of both strategies under a consistent cross-validation framework in the model comparison experiments (see Comment on Page 17 for details). For the spatiotemporal generalization analysis, we focus on the reduced-sample dataset only. This is because it provides a relatively consistent and complete set of input features

across glaciers, whereas the full dataset contains substantial variability in temporal coverage and missing variables (e.g., albedo), which would introduce additional heterogeneity and confound the assessment of generalization performance. This choice is therefore motivated by data consistency rather than performance considerations. These revisions improve the transparency and robustness of the analysis and avoid overinterpretation of the comparative results.

Action: We have revised the Section 3.1:

yields better predictive performance. This conclusion is further supported by the test scatterplots (Fig. 6(a) and (b)) and the corresponding mean cross-validation metrics, both of which clearly show improved agreement between modeled and observed values. ~~These findings suggest that masking missing albedo values does not fully eliminate the negative effects of incomplete features.~~ These results suggest that masking missing albedo values may not fully mitigate the impact of incomplete features.

350 Instead, it increases data noise and introduces distributional inconsistencies, ultimately degrading model performance. In contrast, restricting the training dataset to samples with complete and homogeneous feature sets yields a cleaner input, enabling the FF-FCNN model to learn more robust and generalizable relationships. ~~Therefore, all subsequent model training was conducted on the reduced-sample dataset.~~ ↵

(11) Comment on Page 15: Can you remind us in the figure what the loss and its units are? E.g. RMSE in m .w.e; Could we also skip the first epochs that have a very high RMSE? Because it's hard to see to what the validation & training loss converge too in this plot as the first epochs are way way higher.

Response: Thank you for this helpful suggestion. We agree that the clarity of the loss curves can be improved. First, the loss is now explicitly defined as RMSE, and its unit (m w.e.) has been added to the y-axis label. Second, a logarithmic scale has been applied to the y-axis to better visualize the convergence behavior across epochs. These changes allow for a clearer comparison between the training and validation curves and make the convergence behavior more interpretable.

In addition, for clarification, we note that reducing the number of meteorological predictors from 20 to 10 leads to slightly smoother loss curves and reduced model variability, while consistently achieving slightly improved overall performance. For clarity and conciseness, only the results based on the 10-predictor configuration are presented in the revised manuscript. The

corresponding results for the 20-predictor configuration are provided here solely in response to the reviewer's comment.

Action: Figure 5 and the corresponding description in Section 3.1 have been revised accordingly.

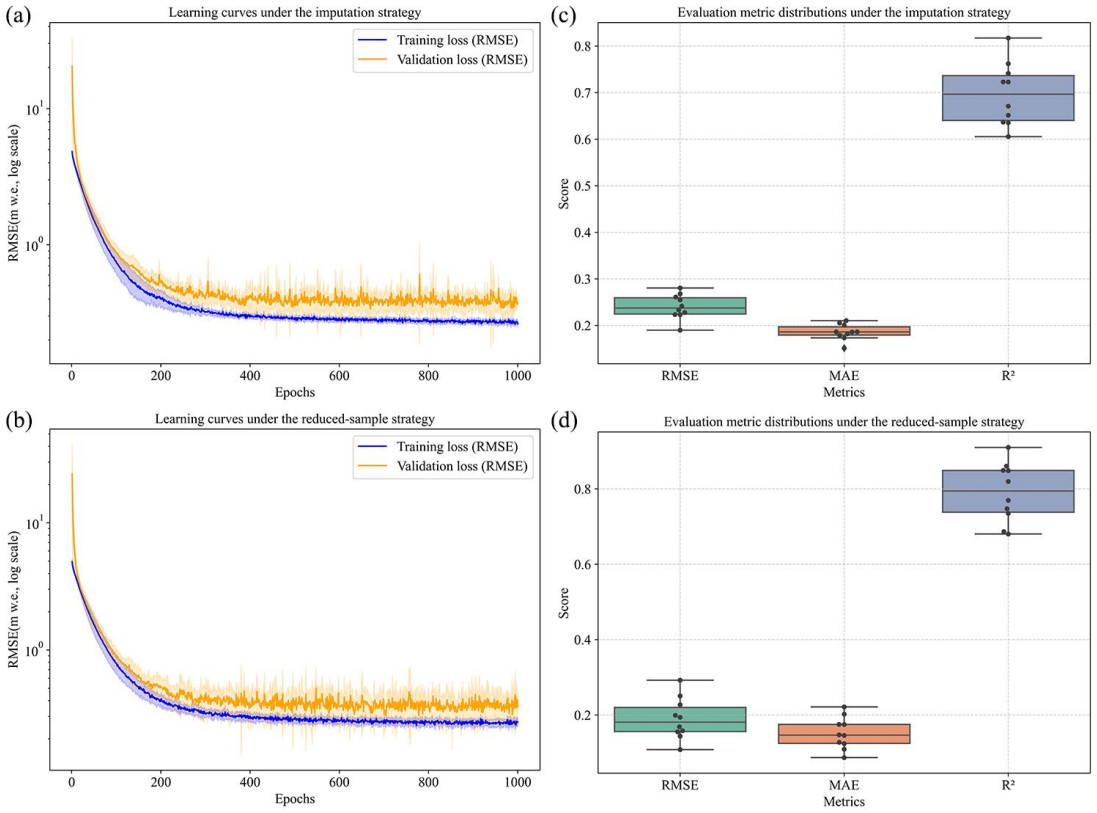


Figure 5 Overfitting assessment and model performance comparison under two dataset construction strategies. **(a)–(b)** Mean training and validation loss curves (RMSE, m w.e., log scale) from 10-fold random cross-validation. **(c)–(d)** Distribution of validation metrics (RMSE, MAE, and R^2) across folds.

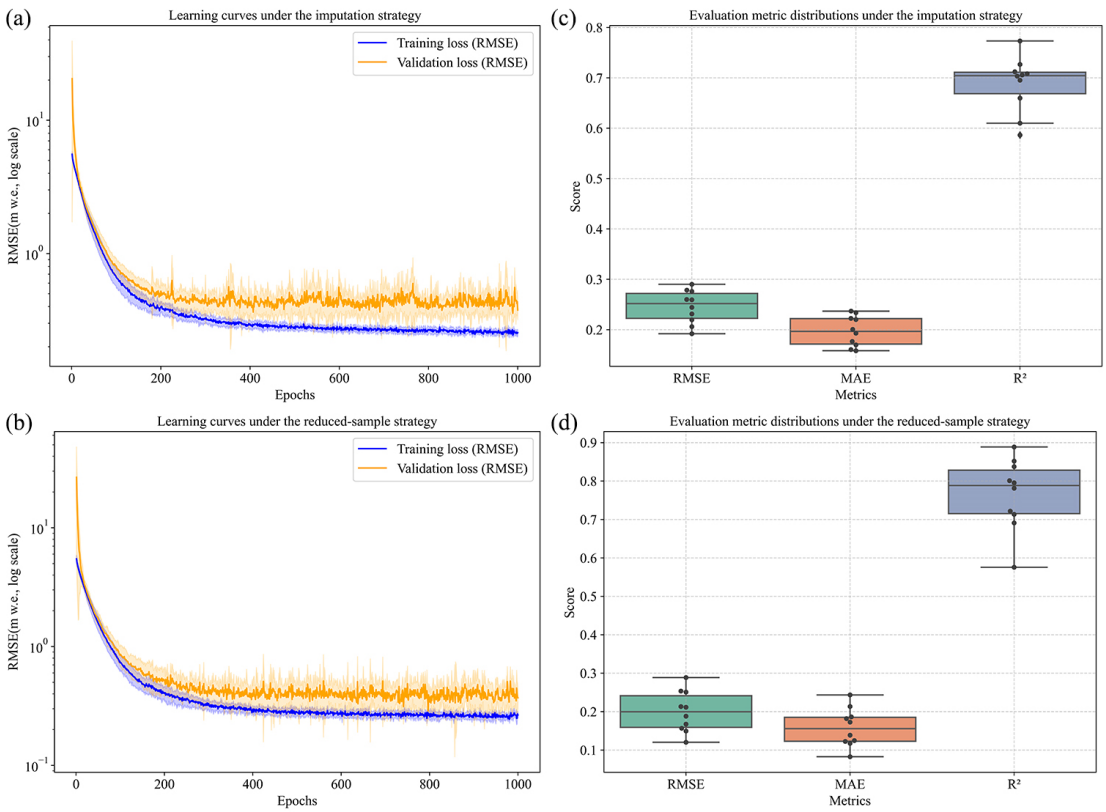


Figure 20 meteorological predictors

320 ~~prior to 2000, reducing the dataset to 109 samples.~~ Given the limited amount of data, it was necessary to evaluate potential overfitting risks to ensure model robustness and reliability. Mean learning curves from 10-fold random cross-validation were examined to compare training and validation losses, and boxplots of validation metrics (RMSE, MAE, and R^2) were used to assess model stability under both strategies. As shown in Fig. 5(a) and (b), the training and validation mean loss curves (log

325 scale) exhibit a consistent decline during the early epochs and gradually converge without noticeable divergence, indicating stable training behavior. Although a small gap between the two curves emerges in the later stages, it remains stable, and the validation loss does not show any increasing trend, suggesting that overfitting is effectively controlled. The slight fluctuations observed in the validation loss are likely attributed to the stochastic nature of mini-batch training and the intrinsic heterogeneity of glacier-wide annual mass balance data. ~~show a consistent downward trend and eventually converge, with no obvious~~

(12) Comment on Page 17: I'd be curious to see also a comparison with XGBoost, a powered up version of RF which has been shown now also several times to outperform neural networks for small datasets.

Response: Thank you for this valuable suggestion. We agree that XGBoost is a strong baseline model and has been shown in several studies to perform well on small datasets. Following this recommendation, we have included XGBoost as

an additional benchmark model and conducted a comparative analysis alongside the RF and FF-FCNN models. Importantly, all three models were evaluated under both dataset construction strategies (i.e., the imputation-based 180-sample dataset and the reduced 109-sample dataset) using the same input features and a consistent cross-validation framework. The results are presented in the revised manuscript (Fig. 6).

The comparison shows that XGBoost achieves slightly better predictive performance than RF across both strategies, confirming its effectiveness for this type of problem. However, the FF-FCNN model consistently attains higher R^2 values and lower prediction errors (e.g., $R^2 = 0.81$, RMSE = 0.2 m w.e., MAE = 0.15 m w.e.). These results suggest that, despite the limited sample size, the proposed FF-FCNN framework is capable of capturing more informative nonlinear relationships among meteorological, topographic, and albedo-related variables.

Action: We have added this comparison and corresponding discussion in Section 3.2 of the revised manuscript.

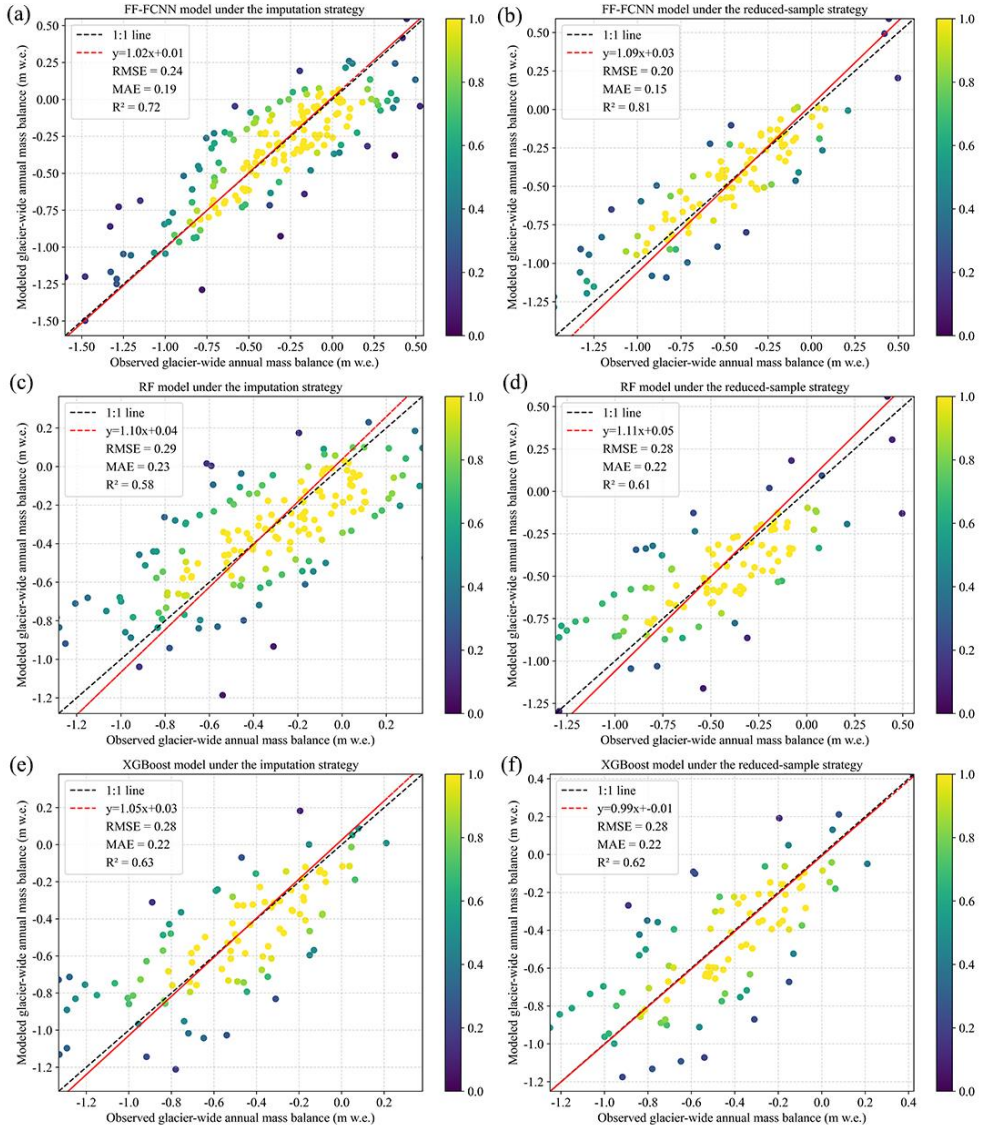


Figure 6 Comparison of modelled versus observed glacier-wide annual mass balance for three models (FF-FCNN, RF, and XGBoost) under two dataset construction strategies. The left column (a, c, e) corresponds to the imputation strategy, while the right column (b, d, f) corresponds to the reduced-sample strategy.

~~networks for glacier mass balance modeling, even under limited data conditions.~~

To comprehensively evaluate the performance of the FF-FCNN model in simulating glacier-wide annual mass balance, RF and XGBoost were employed as benchmarks. Both models are widely used for nonlinear regression tasks and are well-suited for small-to-medium-sized datasets, thereby providing robust baselines for assessing the proposed framework. Both dataset construction strategies are retained and systematically compared under a consistent cross-validation framework in the model comparison experiments to ensure an unbiased evaluation. All models were trained and validated using a 10-fold random cross-validation scheme. Under the reduced-sample strategy (Fig. 6(b), (d), and (f)), the FF-FCNN model achieves the best performance, with RMSE = 0.20 m w.e., MAE = 0.15 m w.e., and R^2 = 0.81. In comparison, the RF model yields RMSE = 0.28 m w.e., MAE = 0.22 m w.e., and R^2 = 0.61, while the XGBoost model shows comparable performance (RMSE = 0.28 m w.e., MAE = 0.22 m w.e., and R^2 = 0.62). A similar pattern is observed under the imputation strategy (Fig .6(a), (c), and (e)). Overall, although all three models capture the general variability of annual mass balance, the FF-FCNN exhibits superior predictive accuracy and stronger agreement with observations across both dataset construction strategies. This comparison highlights the robustness of the FF-FCNN framework and suggests that its architecture is effective in capturing physically meaningful patterns from glacier mass balance data, even under sparse and heterogeneous conditions.

(13) Comment on Page 22: Actually, the mass balance machine has already been trained on more sophisticated energy mass-balance models... so it's not really novel in that sense.

Response: Thank you for this important comment. We acknowledge that previous studies have already developed machine learning models incorporating energy-balance-related variables and performed comprehensive comparisons with other mass balance models. We have reorganized Section 4.3 by consolidating the first and third paragraphs and have removed the sentence in question.

Action: We have revised the Section 4.3:

Glacier mass balance observations in high mountain regions remain extremely scarce, discontinuous, and unevenly distributed due to the harsh and inaccessible environment, which fundamentally constrains the applicability of data-driven modelling. In this context, this study developed an FF-FCNN framework based on ten continental glaciers distributed across multiple mountain ranges under extremely limited sample conditions, with the aim of evaluating the capability of a lightweight deep learning model to extract meaningful signals under such constraints. A series of regularization techniques and rigorous cross-validation schemes were employed to help mitigate overfitting risk and improve generalization ability. Despite the climatic heterogeneity across the studied glaciers, key variables such as CPDD, r_{sp}, and t_{sp} are consistently selected in more than 90% of the random cross-validation folds (Fig. 3(b)), suggesting that the model relies on physically meaningful predictors rather than spurious correlations. It should be noted that r_{sp} and t_{sp} are derived from reanalysis products that implicitly incorporate snowpack evolution and surface energy-balance processes. Therefore, although the FF-FCNN model does not explicitly resolve energy–mass exchange processes, it still benefits from physically informed information embedded in the input data. In

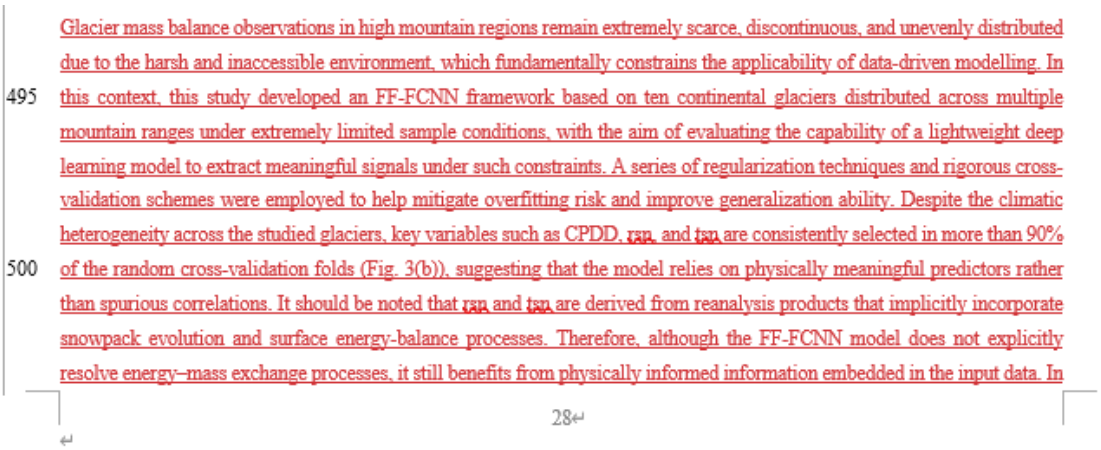
addition, summer surface albedo was incorporated as a core input directly linked to the surface energy balance, thereby strengthening the physical consistency of the model and improving its performance under complex environmental conditions. Nonetheless, the influence of data scarcity remains unavoidable, manifesting in multifaceted challenges throughout the modeling process. The limited sample size introduces inherent uncertainty into the feature selection process, as reflected by the fluctuating selection frequencies of non-core variables. This constraint also limits the model's capacity to capture extreme mass-balance values, such as those observed at Muztag Ata No. 15 and Ningchanhe No.1 glaciers, which are insufficiently represented in the training set to enable robust pattern recognition. Furthermore, despite the strong nonlinear representation capabilities of deep learning models, their inherent "black box" nature poses challenges for interpreting physical mechanisms. Recently, hybrid approaches combining physical modelling with deep learning have been increasingly applied in geoscientific research (Fuchs et al., 2023; Steidl et al., 2025; Teufel et al., 2023;). These approaches can be implemented in two main ways: (1) embedding physical regularization constraints directly into the loss function of the deep learning model, or (2) running numerical simulations with the deep learning model and subsequently assimilating or calibrating the results using a physical model. Such advancements suggest that data-driven models tend to perform better when integrated with physically-based approaches. To further enhance model robustness and generalization, future work should prioritize expanding the dataset and exploring approaches such as transfer learning and advanced data augmentation. Meanwhile, integrating physical processes such as energy balance components and snowpack dynamics into the FF-FCNN framework could facilitate the development of hybrid models that leverage both data-driven and physically-based approaches.

2. Comment: The authors are tackling a genuinely difficult problem. Glacier mass balance observations in High Mountain Asia are extremely sparse, discontinuous, and unevenly distributed in both space and time, and the need for scalable modeling approaches is real and well-motivated. However, the very scarcity of data that motivates this study also fundamentally limits what a purely data-driven approach can achieve here. With only 109–180 samples spanning ten glaciers across highly heterogeneous climatic settings, a deep learning model faces an inherently ill-constrained problem. The authors invest considerable effort in mitigating overfitting through architectural choices, but the core issue is that the available data may simply be insufficient to reliably train and evaluate a model of this complexity. Alternative approaches, such as physically-constrained models, transfer learning from better-observed regions, or hybrid physical-statistical frameworks, may be better suited to this data regime, and the authors should more explicitly acknowledge and engage with this fundamental limitation rather than treating it primarily as an engineering problem to be solved through regularization.

   Response: Thank you for these insightful and constructive comments. We fully agree that the extreme scarcity and heterogeneity of glacier mass balance observations in High Mountain Asia impose fundamental constraints that cannot be resolved solely through methodological adjustments like regularization. In the revised manuscript, we have explicitly acknowledged this limitation and reframed the positioning of our study. However, we would like to emphasize that the selection of key variables (e.g., CPDD, rsn, and tsn) through cross-validation-based feature selection and the inclusion of physically-grounded predictors like summer albedo demonstrate that the framework captures robust glaciological relationships rather than stochastic noise, providing critical physical anchors in this data-sparse environment. These inputs act as critical physical anchors in an otherwise data-sparse environment. Furthermore, we concur that hybrid frameworks and transfer learning represent the next frontier for this field, as discussed in our original manuscript. Following your guidance, we have substantially refined Section 4.3 to critically engage with these inherent data limitations. We now explicitly position

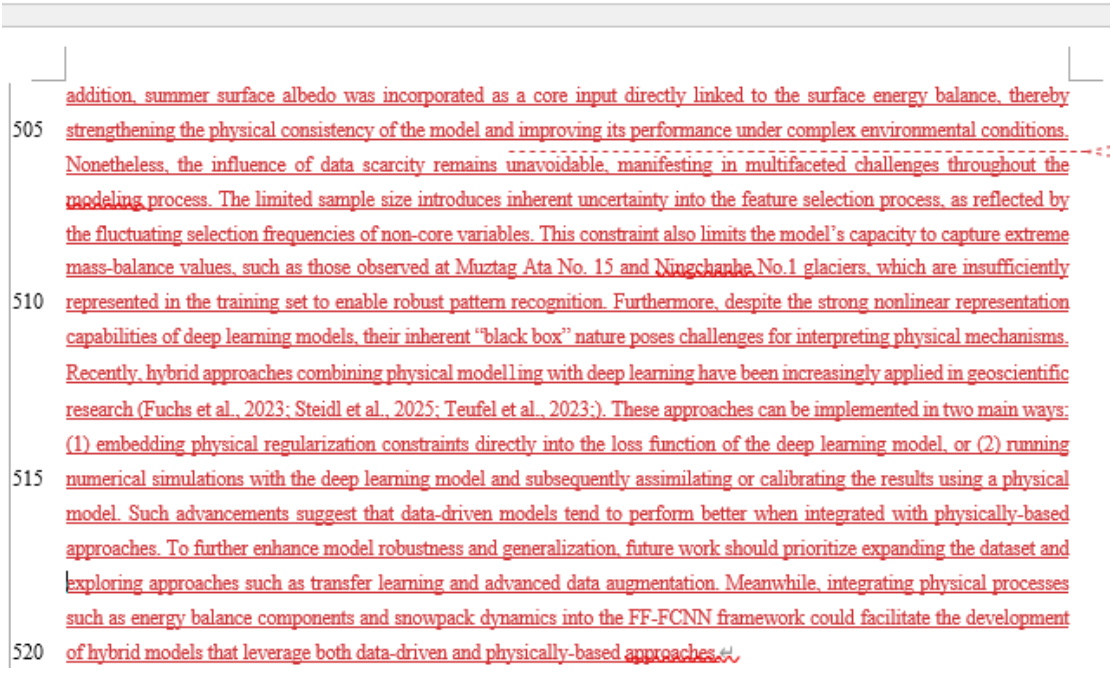
our study as a foundational benchmark that demonstrates the potential of lightweight data-driven architectures, providing a necessary reference for future development of physically-constrained or hybrid modeling frameworks.

Action: We have consolidated and restructured the first and third paragraphs of Section 4.3 to provide a more coherent discussion of the study's limitations and the prospects.



Glacier mass balance observations in high mountain regions remain extremely scarce, discontinuous, and unevenly distributed due to the harsh and inaccessible environment, which fundamentally constrains the applicability of data-driven modelling. In this context, this study developed an FF-FCNN framework based on ten continental glaciers distributed across multiple mountain ranges under extremely limited sample conditions, with the aim of evaluating the capability of a lightweight deep learning model to extract meaningful signals under such constraints. A series of regularization techniques and rigorous cross-validation schemes were employed to help mitigate overfitting risk and improve generalization ability. Despite the climatic heterogeneity across the studied glaciers, key variables such as CPDD, r_{sn}, and t_{sn} are consistently selected in more than 90% of the random cross-validation folds (Fig. 3(b)), suggesting that the model relies on physically meaningful predictors rather than spurious correlations. It should be noted that r_{sn} and t_{sn} are derived from reanalysis products that implicitly incorporate snowpack evolution and surface energy-balance processes. Therefore, although the FF-FCNN model does not explicitly resolve energy–mass exchange processes, it still benefits from physically informed information embedded in the input data. In

28



addition, summer surface albedo was incorporated as a core input directly linked to the surface energy balance, thereby strengthening the physical consistency of the model and improving its performance under complex environmental conditions. Nonetheless, the influence of data scarcity remains unavoidable, manifesting in multifaceted challenges throughout the modeling process. The limited sample size introduces inherent uncertainty into the feature selection process, as reflected by the fluctuating selection frequencies of non-core variables. This constraint also limits the model's capacity to capture extreme mass-balance values, such as those observed at Muztag Ata No. 15 and Ningchanhe No.1 glaciers, which are insufficiently represented in the training set to enable robust pattern recognition. Furthermore, despite the strong nonlinear representation capabilities of deep learning models, their inherent "black box" nature poses challenges for interpreting physical mechanisms. Recently, hybrid approaches combining physical modelling with deep learning have been increasingly applied in geoscientific research (Fuchs et al., 2023; Steidl et al., 2025; Teufel et al., 2023;). These approaches can be implemented in two main ways: (1) embedding physical regularization constraints directly into the loss function of the deep learning model, or (2) running numerical simulations with the deep learning model and subsequently assimilating or calibrating the results using a physical model. Such advancements suggest that data-driven models tend to perform better when integrated with physically-based approaches. To further enhance model robustness and generalization, future work should prioritize expanding the dataset and exploring approaches such as transfer learning and advanced data augmentation. Meanwhile, integrating physical processes such as energy balance components and snowpack dynamics into the FF-FCNN framework could facilitate the development of hybrid models that leverage both data-driven and physically-based approaches.

3. Comment: In the discussion, the authors claim that their approach is novel in training a machine learning model with energy-balance-type variables (radiation, turbulent fluxes, albedo), contrasting it with prior work they characterise as relying primarily on temperature and precipitation. However, this claim overlooks the Mass

Balance Machine framework (Sjursen et al. 2025, van der Meer et al. 2026), which already adopts a broadly comparable philosophy of driving a data-driven model with generally the same variables drawn from the energy balance rather than simple temperature-index approaches. This work should be cited and discussed in the introduction, where the landscape of existing data-driven mass balance models is reviewed. As it stands, the authors overstate the novelty of their input feature design, and readers familiar with the literature will notice the omission. The discussion should be revised to situate the FF-FCNN more accurately relative to existing approaches that similarly go beyond temperature-index inputs.

Response: Thank you for this valuable suggestion and for bringing these recent and relevant studies to our attention. We agree that our previous wording overstated the novelty of the input feature design and did not adequately acknowledge existing data-driven frameworks such as the Mass Balance Machine (Sjursen et al., 2025; van der Meer et al., 2026). In the revised manuscript, we have incorporated these studies into the Introduction to provide a more comprehensive and accurate review of existing data-driven glacier mass balance modelling approaches. In addition, the corresponding discussion in Section 4.3 has been revised to remove the inappropriate claims of novelty and to better position the proposed FF-FCNN framework within the broader context of existing methods.

Action: We have revised the Section 4.3:

Glacier mass balance observations in high mountain regions remain extremely scarce, discontinuous, and unevenly distributed due to the harsh and inaccessible environment, which fundamentally constrains the applicability of data-driven modelling. In this context, this study developed an FF-FCNN framework based on ten continental glaciers distributed across multiple mountain ranges under extremely limited sample conditions, with the aim of evaluating the capability of a lightweight deep learning model to extract meaningful signals under such constraints. A series of regularization techniques and rigorous cross-validation schemes were employed to help mitigate overfitting risk and improve generalization ability. Despite the climatic heterogeneity across the studied glaciers, key variables such as CPDD, t_{sp}, and t_{sp} are consistently selected in more than 90% of the random cross-validation folds (Fig. 3(b)), suggesting that the model relies on physically meaningful predictors rather than spurious correlations. It should be noted that t_{sp} and t_{sp} are derived from reanalysis products that implicitly incorporate snowpack evolution and surface energy-balance processes. Therefore, although the FF-FCNN model does not explicitly resolve energy–mass exchange processes, it still benefits from physically informed information embedded in the input data. In

addition, summer surface albedo was incorporated as a core input directly linked to the surface energy balance, thereby strengthening the physical consistency of the model and improving its performance under complex environmental conditions. Nonetheless, the influence of data scarcity remains unavoidable, manifesting in multifaceted challenges throughout the modeling process. The limited sample size introduces inherent uncertainty into the feature selection process, as reflected by the fluctuating selection frequencies of non-core variables. This constraint also limits the model's capacity to capture extreme mass-balance values, such as those observed at Muztag Ata No. 15 and Ningchanhe No.1 glaciers, which are insufficiently represented in the training set to enable robust pattern recognition. Furthermore, despite the strong nonlinear representation capabilities of deep learning models, their inherent "black box" nature poses challenges for interpreting physical mechanisms. Recently, hybrid approaches combining physical modelling with deep learning have been increasingly applied in geoscientific research (Fuchs et al., 2023; Steidl et al., 2025; Teufel et al., 2023;). These approaches can be implemented in two main ways: (1) embedding physical regularization constraints directly into the loss function of the deep learning model, or (2) running numerical simulations with the deep learning model and subsequently assimilating or calibrating the results using a physical model. Such advancements suggest that data-driven models tend to perform better when integrated with physically-based approaches. To further enhance model robustness and generalization, future work should prioritize expanding the dataset and exploring approaches such as transfer learning and advanced data augmentation. Meanwhile, integrating physical processes such as energy balance components and snowpack dynamics into the FF-FCNN framework could facilitate the development of hybrid models that leverage both data-driven and physically-based approaches.

4. Comment: Wang & Zhang (2026) evaluate their FF-FCNN model using cross-validation, but feature selection was performed on the full dataset prior to any cross-validation split. Specifically, both the Pearson correlation filtering and the Random Forest importance ranking, which together reduced 271 candidate variables to a final set of 20 meteorological predictors, were conducted using all available samples, including those later designated as validation data in each cross-validation fold.This constitutes data leakage: the features chosen for model training were implicitly selected based on information from the validation folds, meaning the cross-validation no longer provides a truly independent assessment of generalization performance. The issue is compounded by the fact that the Random

Forest used for importance ranking is itself a learned model, capable of capturing nonlinear patterns across the full dataset rather than just linear associations as in the Pearson step. Furthermore, the authors evaluate two dataset construction strategies and select the better-performing one (the 109-sample reduced strategy) for all subsequent analysis. Without a fully held-out test set that is untouched by any model or strategy selection decision, this introduces an additional layer of implicit optimization; the reported results reflect the best outcome across tested configurations rather than an unbiased estimate of model performance. Combined with the feature selection leakage, this suggests the model is likely still overfitting to the available data despite the anti-overfitting measures incorporated into the architecture, and that the reported metrics meaningfully overestimate true generalization performance. The correct approach would be to nest the entire feature selection pipeline inside the cross-validation and to evaluate all strategic choices on a fully independent holdout set. As implemented, the reported performance metrics are likely optimistic, though the degree of inflation is difficult to quantify without re-running the analysis with a properly nested procedure.

Response: Thank you for this careful and technically rigorous assessment. We acknowledge that performing feature selection on the full dataset prior to cross-validation introduces data leakage and compromises the independence of model evaluation.

To address this concern, we have substantially revised our methodology. The entire feature selection pipeline, including both Pearson correlation filtering and RF-based importance ranking, is now fully nested within the cross-validation framework. Specifically, feature selection is performed independently within each fold using only the training data, ensuring that no information from the validation subset is used and thereby eliminating data leakage. In addition, we have revised the experimental design to avoid implicit optimization across dataset construction strategies. Rather than selecting a single best strategy, we now report and compare the performance of both strategies under a consistent cross-validation framework in the model comparison experiments, and interpret the results more cautiously.

Furthermore, we have refined the model configuration to enhance robustness and further mitigate overfitting risks. Specifically, the number of selected meteorological predictors has been reduced from 20 to 10, yielding a more parsimonious and physically interpretable feature space. We also deprecated the dynamic loss weighting strategy to avoid introducing additional sources of implicit optimization. Combined with the nested cross-validation framework, these adjustments ensure a more conservative and reliable evaluation of model performance.

Following these revisions, several notable changes in model performance are observed. First, the overall performance across all cross-validation strategies shows a slight decline, confirming that the previously reported metrics were somewhat optimistic due to the non-nested procedure. Second, across both dataset construction strategies, the FF-FCNN model consistently demonstrates better predictive performance than the RF and XGBoost models, indicating that its architecture effectively captures the underlying physically meaningful relationships under sparse and heterogeneous data conditions. Third, reducing the number of meteorological predictors from 20 to 10 yields comparable and slightly improved, validation performance, while reducing model variability. This indicates that a more parsimonious feature set helps mitigate redundancy and enhances model stability under data-sparse conditions. Fourth, with the removal of the dynamic loss weighting strategy, the model exhibits improved training stability and reduced fluctuations in validation error. However, it tends to underestimate large positive mass-balance values, likely due to the inherent class imbalance, as positive samples account for only approximately 10% of the dataset and are therefore underrepresented during training.

These changes have been incorporated into Section 2.3.1 (feature selection) and the corresponding Results and Discussion sections. We thank the reviewer for highlighting this important issue, which has significantly improved the methodological rigor of the study.

We hope that our responses have adequately addressed your comments and concerns, and that the revisions made accordingly have improved the clarity and robustness of the manuscript.

Best regards,

Lili Wang