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 paper 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:

Before addressing the specific comments, we would like to briefly clarify that several methodological refinements have been implemented following the suggestions of Reviewer #1. These updates affect all results presented below. Specifically, (1) the feature selection procedure has been fully embedded within each cross-validation fold, such that all feature filtering and ranking steps are performed using only the training subset, thereby eliminating potential data leakage; (2) the number of selected meteorological predictors has been reduced from 20 to 10 to improve model parsimony and stability; and (3) the previously implemented dynamic loss weighting strategy has been removed to enhance training robustness. Accordingly, all experimental results reported in the following responses are based on this revised modeling framework.

1. Comment: It would be helpful for the reader to know how the glacier-wide mass balance measurements were acquired. One can guess this comes from glaciological measurements, extrapolated spatially with the glacier hypsometry but this is not clearly stated in the manuscript.

   Response: Thank you for this helpful comment. We agree that the description of how glacier-wide mass balance measurements were obtained could be made clearer. The principle of the glaciological method has been briefly introduced in the Introduction section (Lines 29~31). To improve clarity, we have revised this description to more explicitly state that glacier-wide mass balance is derived from repeated in situ measurements (e.g., stake readings and snow pit observations), which are subsequently extrapolated over the glacier using its hypsometry.

Action: The corresponding descriptions have been revised in both the Introduction and Section 2.2.1.

sustainability of regional water resources under ongoing climatic change. In contrast to geometric indicators such as area or length, mass balance responds directly and immediately to climate forcing (Duan et al., 2009). ~~The glaciological method quantifies accumulation and ablation through repeated measurements of stake exposure heights and snow pit profiles, which~~

1

~~are spatially extrapolated to derive glacier-wide mass balance~~ The glaciological method quantifies accumulation and ablation through repeated in situ measurements, including stake readings and snow pit observations distributed across the glacier surface. These point-scale measurements are then spatially extrapolated using the glacier hypsometry to derive glacier-wide mass balance. (Kuhn et al., 1999; Thibert et al., 2008). Although this approach yields high-precision measurements, its long-

~~the National Cryosphere Desert Data Center (NCDC), the National Tibetan Plateau Data Center (TPDC),~~ and relevant academic literature. ~~These datasets originate from conventional glaciological measurements and are reported on an annual basis, with values expressed in meters water equivalent (m w.e.) or millimeters water equivalent (mm w.e.).~~ These datasets originate from conventional glaciological (in situ) measurements and are obtained by spatially extrapolating point observations (see Introduction). They are reported on an annual basis, with values expressed in meters water equivalent (m w.e.) or millimeters water equivalent (mm w.e.). For consistency and comparability, all records were converted to m w.e. The temporal coverage varies substantially among glaciers, ranging from 5 to 64 years, yielding a total of 180 annual observations. An overview of the mass balance records—including their sources and observation periods—is provided in Table 1.

2. Comment: From the feature selection study it seems that the meteorological features of the accumulation period contribute marginally to the annual mass balance in comparison to the ablation ones (Fig 3). This makes sense since most of the annual measurements are negative (Fig 2). A study of the performance for positive annual mass balance would improve the confidence in the learned model, for example by including some Karakoram glaciers.

Response: Thank you for this suggestion. We agree that including glaciers with positive annual mass balance would provide a more comprehensive evaluation of the model performance. We have made efforts to collect such data from regions where glaciers are reported to exhibit positive or near-balanced mass budgets, including the Eastern Pamir, Karakoram, and West Kunlun regions (Wang et al., 2025). However, due to the harsh environmental conditions and limited accessibility of these high-altitude areas, in situ glaciological observations remain extremely scarce. In addition, although some glaciers may have continuous mass-balance observations, these datasets are not always publicly available and therefore could

not be accessed for this study. We have acknowledged this limitation in the Discussion section and stated that future work will strive to expand datasets as they become available.

3. Comment: Taking as input features the longitude and latitude is questionable (L212). The whole point is to approximate the glacier-wide annual mass balance with a statistical regressor based on topographical, meteorological and albedo information. Why do the authors include the location of the glaciers? At best this does nothing, at worst this encodes spatial information into the model which can lead to overfitting.

   Response: Thank you for this insightful comment. We agree that including geographic coordinates (longitude and latitude) may introduce the risk of encoding spatial information and potentially lead to spatial overfitting. To evaluate this concern, we conducted additional experiments comparing model performance with and without geographic coordinates under both random cross-validation and leave-one-glacier-out (LOGO) cross-validation schemes. As shown in the figure below, the model performance remains nearly identical in both cases. In particular, under the more stringent LOGO cross-validation, which explicitly tests spatial generalization, the inclusion of geographic coordinates does not lead to any systematic improvement or degradation in performance. These findings indicate that the model does not rely on geographic coordinates as spatial identifiers and that their contribution is limited compared to physically meaningful predictors. Based on this evidence, we retain longitude and latitude as auxiliary inputs to represent large-scale spatial gradients that may not be fully captured by the available predictors. At the same time, we acknowledge their potential limitations and appreciate the reviewer's suggestion, which helped us to further validate the robustness of the model.

   Action: We conducted additional sensitivity analyses by comparing model performance with and without longitude and latitude under both random and LOGO cross-validation schemes.

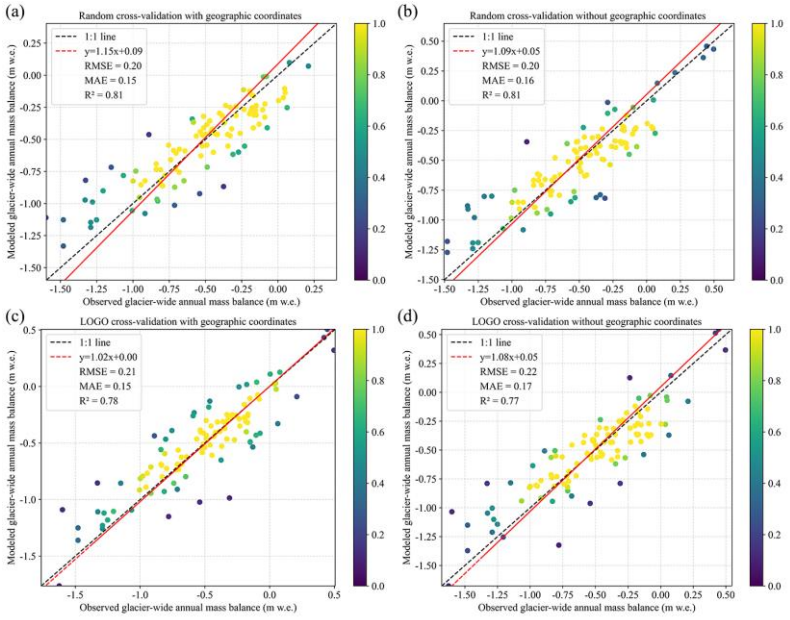


Figure Comparison of model performance with and without geographic coordinates under random and leave-one-glacier-out (LOGO) cross-validation schemes.

4. Comment: The size of the neural network seems very large given the limited amount of data. Depending if the authors use bias or not (not stated in the manuscript), the number of parameters is around 2100 or 2200. In contrast there are 27 predictors and only 109 samples in the reduced dataset. This learning problem is prone to overfitting and special attention should be paid to the validation of the model. They perform a cross validation which is the way to go. However, a grid search is also performed to tune the model architecture along with other hyper-parameters and they converge to the 40, 20, 10, 5 neurons combination. According to the test performance metrics from Figure 6 which are the same as the ones reported in Figure 5 (validation metrics), there is no independent test set that was kept aside from this grid search. From a statistical point of view, performing a grid search on the same data used for the test is equivalent to using the test samples for training. The direct implication is that the reported performance values are optimistic and a more robust test should be performed. If the authors have the opportunity to revise their work I would strongly recommend to keep some glaciers aside in a test set that is used only for comparison and not in the cross validation, nor in the grid search.

   Response: Thank you for this important and insightful comment. We agree that the

previous experimental design could potentially lead to optimistic performance estimates. In the revised manuscript, we have removed the previously described GridSearchCV procedure to avoid any potential ambiguity or risk of information leakage during model evaluation. The model architecture and hyperparameters are now determined through empirical tuning. Once the model configuration was fixed, cross-validation was performed exclusively for performance evaluation, with no further adjustment of model parameters. The corresponding descriptions in the manuscript have been revised accordingly.

In addition, we introduced a fully independent test set consisting of four glaciers with relatively long and continuous observation records. These glaciers were not involved in any stage of model development, including feature selection, cross-validation, or hyperparameter tuning. During the data collection process, we made efforts to compile all available continental glaciers in China with publicly accessible mass-balance observations. As no additional suitable glaciers were available within this region, we extended our selection to neighboring regions and identified four glaciers with relatively long and continuous mass-balance records. Specifically, we selected two glaciers from the Tien Shan—Glacier No. 354 (2012–2023) and Tuyuksu (2000–2023) glacier—and two from the western Himalayas—Patsio (2011–2017) and Chhota Shigri (2003–2019, 2022–2023) glaciers. The results indicate that the model achieves relatively strong performance for the Tien Shan glaciers ($R^2 \approx 0.58$–0.62), which share similar climatic characteristics with the training dataset (Table 3). In contrast, the performance decreases for the Himalayan glaciers, with R^2 values of 0.22 and -0.12, respectively. Despite the reduced R^2 values, the correlation coefficient remains consistently high ($R \approx 0.78$–0.81), suggesting that the model is able to capture interannual variability, although with noticeable biases in magnitude (Fig. 10). These findings indicate that while the model demonstrates reasonable generalization capability within predominantly continental glacier systems, its transferability to glaciers under distinctly different climatic regimes is limited. This behavior is consistent with the physical understanding of glacier–climate interactions and highlights the importance of

incorporating more climatically diverse training data in future work.

Action:

(1) The corresponding descriptions related to GridSearchCV in the manuscript have been revised accordingly (Section 2.3.3).

235 ~~Hyperparameters tuning was performed using GridSearchCV tool from the scikit-learn library (Pedregosa et al., 2011) combined with cross-validation to identify the optimal configuration of neuron numbers, hidden layer depth, activation functions, learning rate, and regularization strategies.~~ The model architecture and hyperparameters (e.g., neuron numbers, hidden-layer depth, activation functions, learning rate, batch size, optimizer, and regulation strategies) were determined empirically through iterative experimentation and preliminary sensitivity analysis. Once the model configuration was finalized, 240 cross-validation was performed exclusively for performance evaluation, without further adjustment of model parameters. The

(2) The corresponding analysis has been added to Section 3.4 of the revised manuscript, which is entitled "Out-of-sample evaluation on independent glaciers."

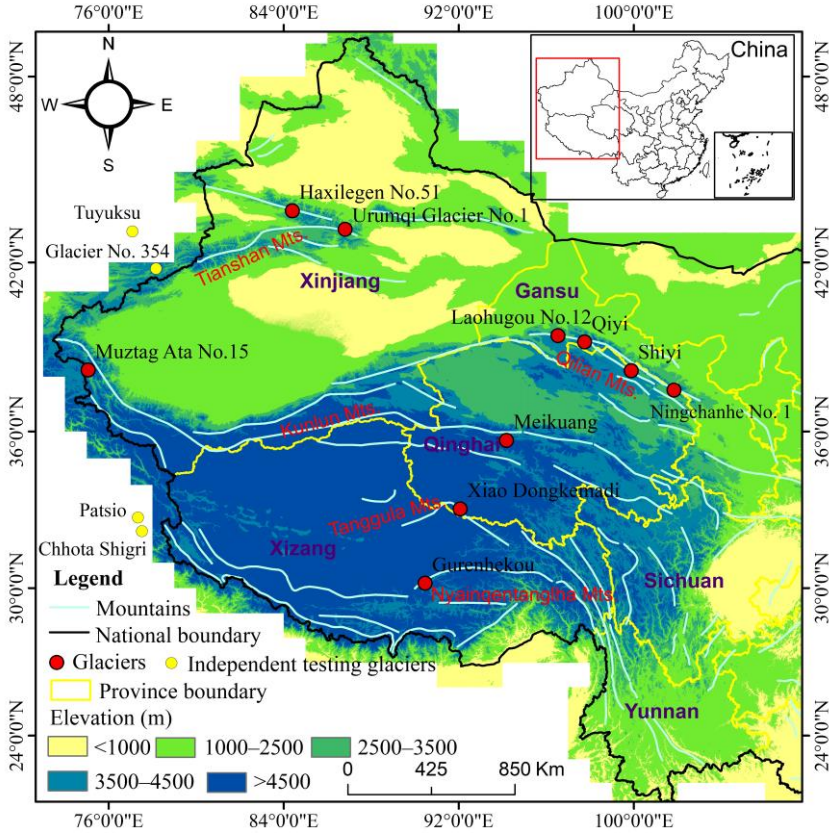


Figure 1 Spatial distribution of the ten continental-type glaciers in China and four independent testing glaciers.

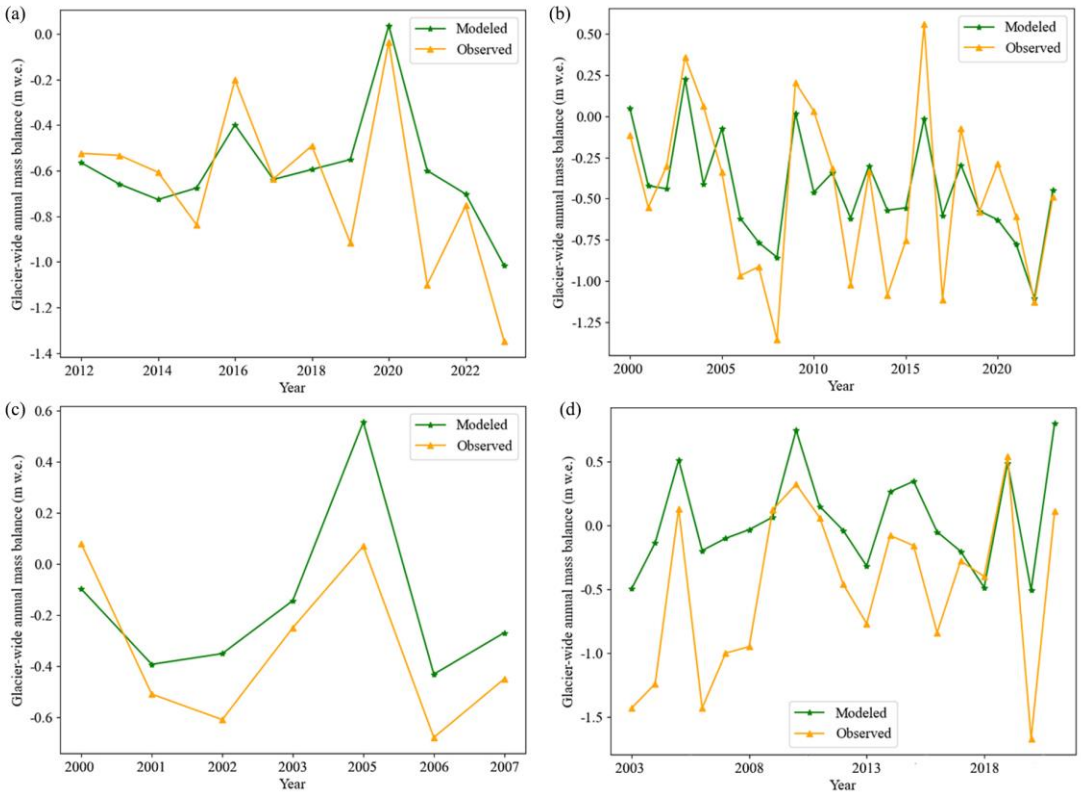


Figure 10 Modeled versus observed glacier-wide annual mass balance for independent test set: (a) Glacier No. 354; (b) Tuyuksu; (c) Patsio; (d) Chhota Shigri. Modeled values represent the average predictions across 10 random seeds.

Table 3 Predictive performance metrics on the independent test set.

Metrics / Glaciers	R^2	R	MAE (m w.e.)	RMSE (m w.e.)
Glacier No. 354	0.58	0.80	0.17	0.23
Tuyuksu	0.62	0.81	0.25	0.31
Patsio	0.22	0.81	0.23	0.25
Chhota Shigri	-0.12	0.78	0.56	0.68

3.4 Out-of-sample evaluation on independent glaciers

To further evaluate the generalization ability of the proposed model, four glaciers located outside the training regions were selected as a fully hold-out test set, including Glacier No. 354 and Tuyuksu in the Tien Shan, as well as Patsio and Chhota Shigri glaciers in the western Himalaya. To ensure statistical robustness, all reported metrics (R^2, R, MAE, and RMSE) are presented as average over ten independent simulations initialized with different random seeds. The results show that the model achieves relatively good performance for Glacier No. 354 (R^2 = 0.58) and Tuyuksu (R^2 = 0.62), indicating a stable generalization capability within similar continental glacier regimes (Table 3). However, its ability to reproduce high-magnitude peaks is notably limited for Tuyuksu glacier (Fig. 10(b)), suggesting that the underrepresentation of extreme values in the training data remains a primary constraint on accurately capturing extreme mass-balance variability. In contrast, the performance decreases for the western Himalayan glaciers, with Patsio yielding an R^2 of 0.22 and Chhota Shigri showing a poor R^2 of -0.12, although the correlation coefficients remain relatively high ($r \approx 0.78–0.81$), suggesting that the model can still capture the temporal variability but fails to accurately reproduce the magnitude of mass balance. The training dataset is dominated by continental-type glaciers in the arid and semi-arid regions of western China, whereas western Himalayan glaciers differ substantially in terms of moisture flux, precipitation seasonality, and monsoonal influence. These contrasts lead to distinct accumulation regimes and energy balance conditions, which likely constrain the model's regional transferability. This finding highlights the importance of incorporating more diverse training data to improve model robustness across heterogeneous glacier environments.

5. Comment: The performance study across different years of section 3.3.2 and the effort of the authors to explain the conditions under which the model performs poorly is appreciated. The approach is interesting in terms of machine learning as this is at the limit of what ML can support given the very limited amount of data. Providing more details on how overfitting was mitigated (e.g. regularization strategies L232) would be relevant for this work.

   Response: Thank you for this helpful comment. We agree that a clearer description of how overfitting is mitigated would improve the manuscript. In fact, several regularization and stabilization strategies were already implemented in the model, including Gaussian noise injection at the input layer, L1 regularization, and early stopping. However, we acknowledge that the original description may not have sufficiently emphasized their roles in preventing overfitting.

   Action: To improve clarity, we have revised the manuscript to provide a more detailed and explicit explanation of these strategies and their respective contributions to model robustness and generalization (Section 2.3.3).

~~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.~~ To mitigate overfitting and improve model generalization under the high-dimensional and small-sample setting, several regularization and stabilization strategies were incorporated. Gaussian noise (standard deviation = 0.1) was added to the input layer to emulate data augmentation effects and enhance robustness to input variability. L1 regularization was applied to constrain model complexity by encouraging sparsity in the network weights, thereby reducing the risk of overfitting. In addition, an early stopping criterion was employed to terminate training when the validation loss ceased to improve, preventing over-training on the limited dataset. 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

6. Comment: However, a comparison with other mass balance models would also be welcome to assess the advantages. Especially since the proposed approach stills relies on the outputs of energy-balance models (ERA5) for some of the variables (see technical corrections). I would recommend at the very least to compare the approach with a temperature-index model.

   Response: Thank you for this important suggestion. To provide a more comprehensive assessment of the proposed approach, we have conducted an additional comparison with a temperature-index (TI) model. Specifically, to ensure a rigorous comparison, the TI model was calibrated separately for each of the four test glaciers using their respective observed mass-balance data. This provides a glacier-specific empirical baseline for evaluation. The comparison results have been incorporated into Section 3.4, Figure 10, and Table 3 of the revised manuscript. The results show that the TI model generally achieves better agreement in terms of absolute mass-balance magnitudes due to local calibration, whereas the FF-FCNN model demonstrates a stronger capability to capture interannual variability, as reflected by consistently higher correlation coefficients. However, both models exhibit limitations in reproducing extreme mass-balance values.

   Action: Following the implementation of an independent test set (Comment 4), we further introduced a temperature-index (TI) model as a baseline for comparison. The TI model was calibrated separately for each test glacier, and the corresponding simulation results and evaluation metrics were added in Figure 10 and Table 3. Section 3.4 has been revised accordingly, including updates to the section title and result analysis.

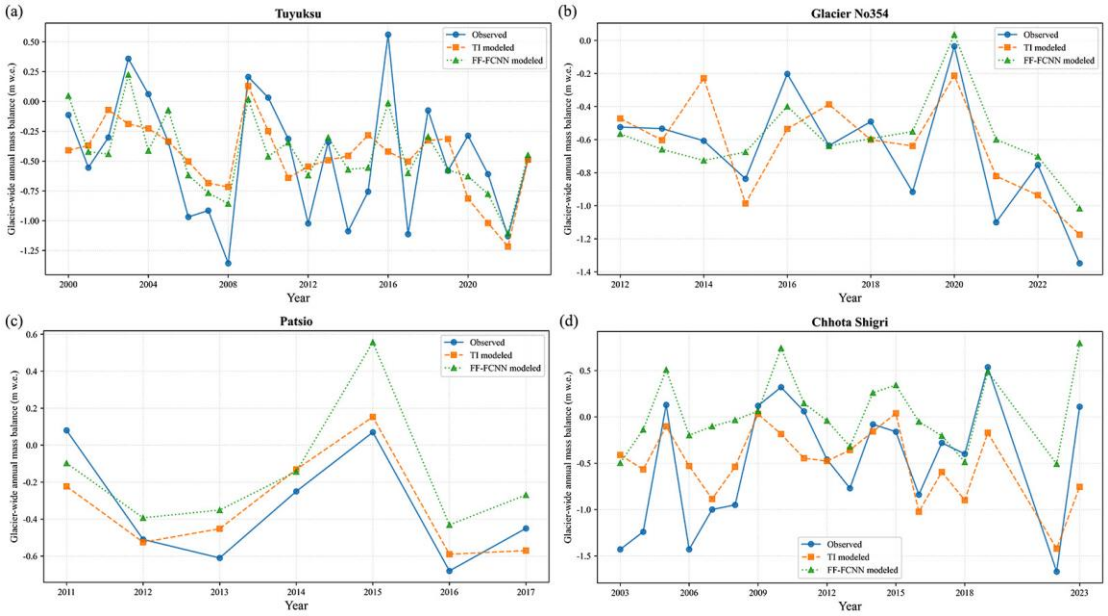


Figure 10 Modeled versus observed glacier-wide annual mass balance for independent test set: (a) Glacier No. 354; (b) Tuyuksu; (c) Patsio; (d) Chhota Shigri. The blue solid line represents observations, the orange dashed line corresponds to the temperature-index (TI) model, and the green dotted line shows the FF-FCNN model predictions.

Table 3 Predictive performance metrics on the independent test set.

Glaciers	Models	R^2	R	MAE (m w.e)	RMSE (m w.e)
Tuyuksu	FF-FCNN	0.62	0.81	0.22	0.31
	TI	0.30	0.55	0.35	0.42
Glacier No. 354	FF-FCNN	0.58	0.80	0.17	0.22
	TI	0.58	0.77	0.20	0.22
Patsio	FF-FCNN	0.22	0.81	0.22	0.26
	TI	0.72	0.85	0.13	0.15
Chhota Shigri	FF-FCNN	-0.12	0.78	0.56	0.68
	TI	0.37	0.61	0.42	0.51

3.4 Out-of-sample evaluation and model comparison on independent glaciers

To further evaluate the generalization ability of the FF-FCNN model, four glaciers located outside the training regions were selected as a fully hold-out test set, including Glacier No. 354 and Tuyuksu in the Tien Shan, as well as Patsio and Chhota Shigri glaciers in the western Himalaya. To ensure statistical robustness, all reported metrics (R^2, R, MAE, and RMSE) represent the average over ten independent simulations initialized with different random seeds. In addition, a temperature-index (TI) model was independently calibrated for each glacier to provide a baseline for comparison.

The results indicate that the FF-FCNN model generally outperforms the TI model for the Tien Shan glaciers, achieving comparable performance for Glacier No. 354 ($R^2 \approx 0.58$ for both models) and a substantial improvement for Tuyuksu (R^2 = 0.62 versus 0.30; Table 3). Time-series analysis (Fig. 10(a) and (b)) further shows that, although both models exhibit a shared limitation in accurately reproducing the magnitude of extreme mass-balance values, the FF-FCNN model demonstrates a superior ability to capture interannual fluctuations, as reflected by its consistently higher correlation coefficients ($R \approx 0.80$–0.81) compared to the TI model ($R \approx 0.55$–0.77). For the western Himalayan glaciers, the TI model achieves higher R^2 values for Patsio (0.72) and Chhota Shigri (0.37), whereas the FF-FCNN model maintains relatively high correlation coefficients ($R \approx 0.78$–0.81). These results reveal a fundamental difference between the two approaches. The TI model relies on glacier-specific parameter calibration, which enhances its capacity to reproduce local mass-balance magnitudes. In contrast, the FF-FCNN model—trained on glaciers from different regions without exposure to the test glaciers—demonstrates a stronger capability to capture interannual variability in an out-of-sample context. Moreover, by incorporating a broader set of physically-informed variables from ERA5-Land, the FF-FCNN framework is better able to represent complex energy-balance responses that cannot be captured by simple temperature–precipitation relationships. The reduced performance of the FF-FCNN model in the western Himalaya can be attributed to differences in climatic regimes. The training dataset is dominated by continental-type glaciers in the arid and semi-arid regions of western China, whereas western Himalayan glaciers differ substantially in terms of moisture flux, precipitation seasonality, and monsoonal influence. These contrasts lead to distinct accumulation regimes and energy balance conditions, which likely constrain the model's regional transferability.

Overall, the comparison demonstrates that the proposed FF-FCNN framework possesses reasonable generalization ability in capturing temporal variability across unseen glaciers, but its ability to reproduce absolute mass-balance magnitudes remains sensitive to the representativeness of the training dataset. This finding underscores the necessity of expanding the training dataset to include more diverse hydroclimatic regimes, which is essential for enhancing model robustness across heterogeneous glacial environments.

7. Comment: Finally the positioning of this work is ambiguous. It claims to be the first to develop a machine learning model that uses "more sophisticated energy-mass balance models" and that "this approach significantly improves predictive performance". For the first claim, [3] already developed a similar approach, although point-wise and at the monthly scale, with an extensive comparison to other models. The second claim is simply not supported at all by any comparison.

   Response: Thank you for this important comment. We agree that the contextualization and positioning of our study was not sufficiently clarified in the initial submission, particularly in the discussion presented in Section 4.3, where the description of novelty and performance could be interpreted as overly strong or insufficiently supported. We also acknowledge that previous studies have explored machine learning approaches incorporating energy-balance-related variables, although at different spatial and temporal scales (Sjursen et al., 2025; van der Meer et al., 2026). In the revised manuscript, we have therefore removed the statements

referring to “more sophisticated energy-mass balance models” and the claim that “this approach significantly improves predictive performance,” to avoid potential ambiguity in their interpretation. The discussion in Section 4.3 has been substantially revised by merging the original third paragraph with the opening paragraph. The revised version now provides a clearer description of the research background, focuses on the study objectives, and better clarifies the positioning of this work. In addition, the revised discussion emphasizes the scope and limitations of the proposed approach, and outlines potential directions for future research, rather than making strong claims regarding novelty or performance superiority.

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, t_{sm}, 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 t_{sm} 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

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.

8. Comment: L44: "they capture only discrete temporal snapshots and are therefore limited in resolving the continuous evolution of glaciers": Not true, technically many remote sensing methods could provide data at a frequency higher than annual; for example [1] (the Hugonnet et al. 2021 dataset) rely on a continuous time series through interpolation, CryoSat-2 provides monthly revisits, and ICESat-2 has a 91 days repeat cycle [2].

   Response: Thank you for this helpful comment. We agree that the original statement was somewhat overly generalized. Remote sensing techniques, including geodetic approaches and satellite altimetry, can provide observations at near-continuous temporal resolution, especially when combined with spatiotemporal interpolation

or time-series reconstruction approaches. In the revised manuscript, this section has been revised to better reflect the capabilities of modern remote sensing datasets, and appropriate references have been added. We have also clarified the associated limitations to ensure a more balanced and accurate description.

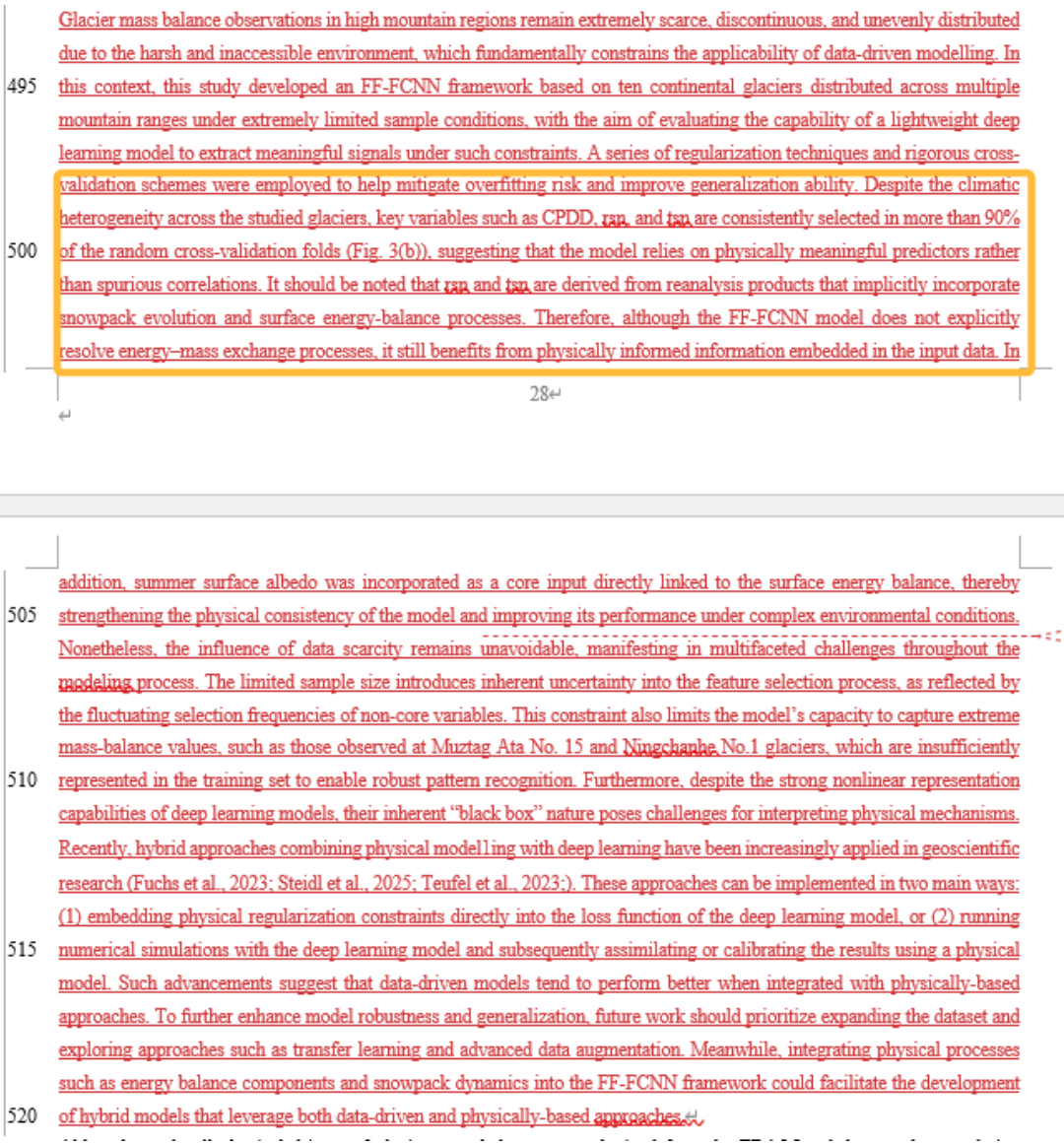
Action: The corresponding text in the Introduction has been revised accordingly.

Advances in remote sensing technology have established both the geodetic method based on differential DEMs (Bash et al., 2018; Rabatel et al., 2016) and the satellite gravimetry method based on temporal variations in Earth's gravitational field (Chen et al., 2007) as important techniques for observing glacier mass balance. ~~Although these approaches provide extensive spatial coverage, they capture only discrete temporal snapshots and are therefore limited in resolving the continuous evolution of glaciers.~~ In particular, the integration of multi-source satellite data and advanced spatiotemporal reconstruction has improved the temporal resolution of glacier observations, enabling near-continuous time series at regional to global scales (Hugonnet et al., 2021; Berthier et al., 2023). Nevertheless, these approaches remain limited in their ability to directly resolve continuous glacier evolution and process-level variability, due to reliance on temporally discrete observations and interpolation-based time series reconstruction. In parallel, numerical and data-driven modeling approaches have emerged as powerful tools for

9. Comment: L152: Snow evaporation, snow density and temperature of snow layer are all the outputs of a re-analysis model which incorporates snow packing and energy model components. I would make it clear since the authors make the distinction with energy-balance models (L57-59) but their model still relies on these energy-balance models.

Response: Thank you for this insightful comment. We agree that several variables used in this study, such as snow evaporation (es), snow density (rsn), and snow layer temperature (tsn), are outputs from reanalysis products that implicitly incorporate snowpack and energy-balance processes. In the revised manuscript, we have clarified this point in Section 4.3 (Discussion) by explicitly stating that these predictors embed physically based information from reanalysis data. We further emphasize that, although the proposed model does not explicitly resolve energy–mass exchange processes, it still benefits from physically informed inputs, thereby improving the physical consistency of the model.

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, rsp, and tsp 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 rsp and tsp 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.

10. Comment: Imprecision L220: The optimizer is not part of the neural network, this is an algorithm to solve the learning problem.

Response: Thank you for this helpful comment. We agree that the original statement was imprecise, as the optimizer is not a component of the neural network itself but rather an algorithm used to solve the learning problem. In the revised manuscript, we have corrected this description by distinguishing the network architecture and activation functions from the optimization algorithm used to update the trainable parameters.

Action: We have revised the Section 2.3.3.

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

11. Comment: L230: What is the search space of the grid search? This should be given for reproducibility.

    Response: Thank you for this comment. In response to the reviewer's earlier concern regarding potential information leakage, we have removed the GridSearchCV procedure from the revised manuscript. Therefore, the specification of a search space is no longer applicable. In the revised version, the model architecture and hyperparameters (e.g., neuron numbers, hidden-layer depth, activation functions, learning rate, batch size, optimizer, and regulation strategies) were determined empirically through iterative experimentation and preliminary sensitivity analysis. Once the model configuration was finalized, cross-validation was performed exclusively for performance evaluation, without further adjustment of model parameters.

12. Comment: L245: Vague statement "improving computational efficiency and training stability": remove computational efficiency.

    Response: Thank you for this helpful comment. We agree that the statement regarding "computational efficiency" was not sufficiently supported. In the revised manuscript, we have removed this term and retained only "training stability" to provide a more accurate description.

    Action: We have removed "computational efficiency" in the Section 2.3.3.

accelerated convergence. Batch normalization was applied before each activation function to stabilize intermediate feature distributions, thereby improving ~~computational efficiency and~~ training stability. During model training, the optimizer, loss function, and evaluation metrics must be specified. Given the limited and uneven distribution of samples, a loss-based dynamic

13. Comment: L246: A reference should be given about the "loss-based dynamic weighting strategy", or it should be explained. This seems like an important component of the approach but it is not explained

    Response: Thank you for this comment. We agree that the description of the "loss-

based dynamic weighting strategy" was insufficient and lacked proper explanation or supporting references. In the revised manuscript, this component has been removed from the model configuration to improve training stability and avoid unnecessary complexity. Accordingly, the corresponding description has also been deleted from the manuscript.

14. Comment: Fig 5a and 5b: A semilogy plot would be better to assess the convergence and the absence of overfitting.

Response: Thank you for this helpful suggestion. We have revised the loss curves by using a semilogarithmic scale on the y-axis, which provides a clearer visualization of the convergence behavior and facilitates the assessment of potential overfitting.

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

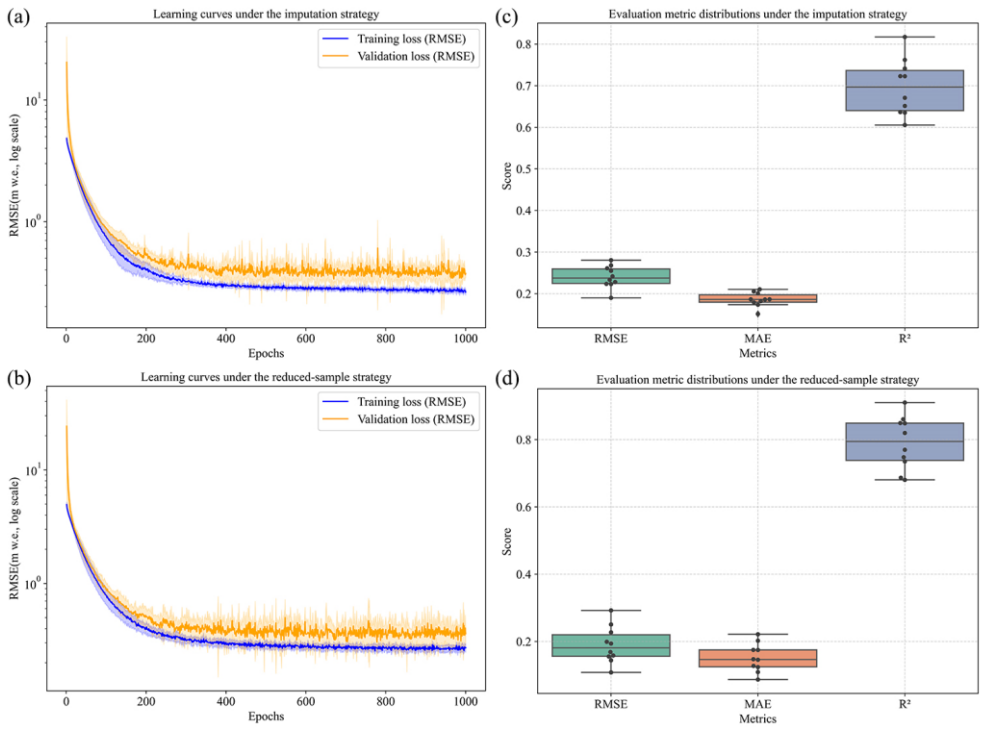


Figure 5 Overfitting assessment and model performance comparison under two dataset construction strategies. **(a)–(b)** Mean training and validation loss curves (RMSE, m w.e., log scale) from 10-fold random cross-validation. **(c)–(d)** Distribution of validation metrics (RMSE, MAE, and R^2) across folds.

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

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

15. Comment: L336 "Although the introduction of a dynamic loss-based weighting strategy enhances [...], the improvement remains limited under extremely small sample conditions.": This is not clear where the authors want to go with this. This is in contradiction with L246. If in the end it is not something that helps the training, I would suggest removing any mention to that dynamic loss-based weighting.

    Response: Thank you for this comment. We agree that the previous description of the dynamic loss-based weighting strategy was unclear and led to inconsistency with the earlier revision (L246). In the revised manuscript, this component has been entirely removed from the model, and all related descriptions have been deleted to ensure consistency throughout the text.

16. Comment: L355 on AUC: A reference would be helpful for non statistician glaciologists.

    Response: Thank you for this suggestion. We agree that a reference for the AUC metric would be helpful, particularly for readers without a strong statistical background. In the revised manuscript, we have added an appropriate reference and a brief clarification of the AUC metric in this context.

    Action: We have revised the Section 3.3.2.

validation periods using only the input features. Its performance was quantified using the area under the receiver operating characteristic curve (AUC) (Fawcett, 2006). An AUC value close to 0.5 indicates similar feature distributions between the two

17. Comment: L431-434: The proposed work is not novel in that sense. For example

[3] already developed a ML model that uses “more sophisticated energy-mass balance models". They did a complete comparison with other mass balance models.

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. In the original manuscript, the positioning of our work in this context was not sufficiently clear and may have led to an over-interpretation of its novelty. In the revised manuscript, we have reorganized Section 4.3 by consolidating the relevant paragraphs and removing the statement in question to avoid ambiguity.

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_{sm} 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_{sm} 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.

18. Comment: L434 "This approach significantly improves predictive performance": This is not supported by any experiment of the manuscript as no comparison at all was performed.

Response: Thank you for this comment. We agree that the statement "this approach significantly improves predictive performance" was not sufficiently supported by direct experimental comparison in the manuscript. In the revised version, this statement has been removed to avoid overinterpretation.

Reference

Wang Ying-Shan, Sun Wei-Jun, Ding Ming-Hu, Liu Wei-Gang, Du Wen-Tao, Qin Xiang, Zhang Dong-Qi, (2025). Characteristics of glacier mass balance changes and

response to climate change in the Qinghai-Tibet Plateau, China. *Advances in Climate Change Research*, *21*(2), 208.

Marijn van der Meer, Harry Zekollari, Alban Gossard, et al. Glacier mass balance modeling using a Long Short-Term Memory network. ESS Open Archive . January 20, 2026. DOI: 10.22541/essoar.176894225.56571551/v1.

Sjursen, K. H., Bolibar, J., Van Der Meer, M., Andreassen, L. M., Biesheuvel, J. P., Dunse, T., ... & Tober, B. (2025). Machine learning improves seasonal mass balance prediction for unmonitored glaciers. The Cryosphere, 19(11), 5801-5826.

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