Reviewer #1

Reviewer: The manuscript 'GlaUnTI: A hybrid physics–machine learning model enables transferable glacier surface mass balance estimation' is a very interesting read. The authors have tested the performance of (i) a differentiable temperature index model, (ii) a pure data-driven gated recurrent unit model, and (iii) a modified temperature index model with a neural network corrector. Two settings of (iii) are used, which are with and without a glacier facies map. This is quite an innovative approach, which attempts to allow the interpretability of temperature index models while allowing non-linear interactions between the meteorological variables at the grid level and the glacier level. The domain of hybrid machine learning and temperature index models is quite novel in studying glacier mass changes and is well aligned with the scope of this journal. That being said, I have a few suggestions that can strengthen the contributions and impact of this study. I divide my comments into Major and Minor Comments.

Response: We sincerely thank Dr Ritu Anilkumar for the careful reading of our manuscript, for the constructive feedback and for the overall positive assessment of the study. The comments were very helpful in improving several parts of the manuscript.

Major comments

Reviewer: The main selling points in this manuscript are as follows:

1. Development of the differentiable form of the temperature index model
2. Development of a machine learning based corrector to temperature index models
3. Spatial transferability of the models developed

However, some aspects regarding transferability and performance are not yet convincingly supported.

Response: We have addressed the major and minor comments in detail below and suggested the according revisions.

Transferability

Reviewer: Regarding the spatial transferability, the manuscript starts compellingly with descriptions of how differentiable models will aid transferability, but does not provide systematic quantification of it in the results. It would strengthen the paper to show that the differentiable TI model achieves comparable accuracy to conventionally calibrated (e.g., grid-search) TI models, ensuring that the smoothing approximations (softplus, sigmoid) do not degrade performance. The authors have described the advantages of the fully differentiable approach to include interpretability and transferability, which will be well justified with this comparison.

Response: Thank you for this valuable suggestion. To clarify, differentiable implementation does not directly lead to better transferability—this is achieved by the neural corrector in GlaUnTI. Nevertheless, we agree that comparing the autodiff-friendly TI formulation with a more classical non-smooth TI implementation is important, because it verifies that the smooth approximations do not degrade the performance of the TI backbone. To address this, we implemented a TI model without smooth approximations, reflecting Equations (6–10):

$$\widehat{\mathrm{SMB}}_{ijd} = \beta_1 P^{\mathrm{solid}}_{ijd} - \beta_2 \left(T^{+,\mathrm{snow}}_{ijd} + \kappa T^{+,\mathrm{ice}}_{ijd} \right) ,$$

$$P^{\mathrm{solid}}_{ijd} = \begin{cases} P_{ijd}, & T \le T_{\mathrm{left}} \\ P_{ijd} \cdot \dfrac{T_{\mathrm{left}} + L_{\mathrm{sr}} - T_{ijd}}{L_{\mathrm{sr}}}, & T_{\mathrm{left}} < T < T_{\mathrm{left}} + L_{\mathrm{sr}} \\ 0, & T \ge T_{\mathrm{left}} + L_{\mathrm{sr}} \end{cases} ,$$

$$T^{+,\mathrm{snow}}_{ijd} = \mathrm{FSC}_{ijd} \cdot T^{+}_{ijd} ,$$
$$T^{+,\mathrm{ice}}_{ijd} = (1 - \mathrm{FSC}_{ijd}) \cdot T^{+}_{ijd} ,$$
$$T^{+}_{ijd} = \max\left(0, T_{ijd}\right) ,$$

$$\mathrm{FSC}_{ijd} = \begin{cases} 0, & \mathrm{SWE}_{ijd} < \tau_{\mathrm{FSC,c}} \\ 1, & \mathrm{SWE}_{ijd} \ge \tau_{\mathrm{FSC,c}} \end{cases} ,$$

$$\mathrm{SWE}_{ijd} = \max\left(0, \mathrm{SWE}_{ij\,d-1} + \beta_1 P^{\mathrm{solid}}_{ijd} - \beta_2 T^{+,\mathrm{snow}}_{ijd}\right) ,$$

where $[\beta_1, \beta_2, \kappa, T_{\text{left}}, L_{\text{sr}}, \tau_{\text{FSC,c}}]^\top$ are the model parameters. T_{left} and L_{sr} define, respectively, the position and slope of the snow-to-rain transition, while the rest of the parameters correspond directly to our autodiff-friendly implementation.
We calibrated this classical TI formulation on the combined train+val glacier set using an iterative grid search procedure. The grid search was performed in *three* refinement iterations. In the first iteration, we explored broad parameter ranges. In the second and third iterations, the parameter ranges were narrowed around the best performing parameter set from the previous iteration. Within each refinement iteration, we fitted the accumulation-related parameters ($\beta_1, T_{\text{left}}, L_{\text{sr}}$ and the ablation parameters ($\beta_2, \kappa, \tau_{\text{FSC,c}}$) in separate grid search steps similar to, e.g., van der Meer et al. (2025) to make it computationally feasible.
For comparison, we also recalibrated the autodiff-friendly TI model on the same train+val glaciers with a comparable computational budget ($\sim$ 36 hours, corresponding to 150 epochs). To make the comparison more direct, we removed the literature prior regularisation by setting $\lambda_2 = 0$ and increased the learning rate to 0.1. This allowed the gradient-based optimisation to explore the parameter space more freely, which is important because the classical grid search calibration converged towards a one-DDF formulation ($\tau_{\text{FSC,c}} = 0$ and $\kappa = 1$), far from the priors.
The resulting point-level performance is nearly identical for the two TI formulations (Fig. 1). For annual SMB, the classical TI model reaches $r = 0.890$, bias $= 0.071$ m w.e. and RMSE $= 1.492$ m w.e., while the autodiff-friendly TI model reaches $r = 0.889$, bias $= 0.030$ m w.e. and RMSE $= 1.522$ m w.e.. The same conclusion holds seasonally. These differences are negligible relative to the overall model errors and show that *the softplus/sigmoid smoothing approximations do not degrade the performance of the TI backbone.*

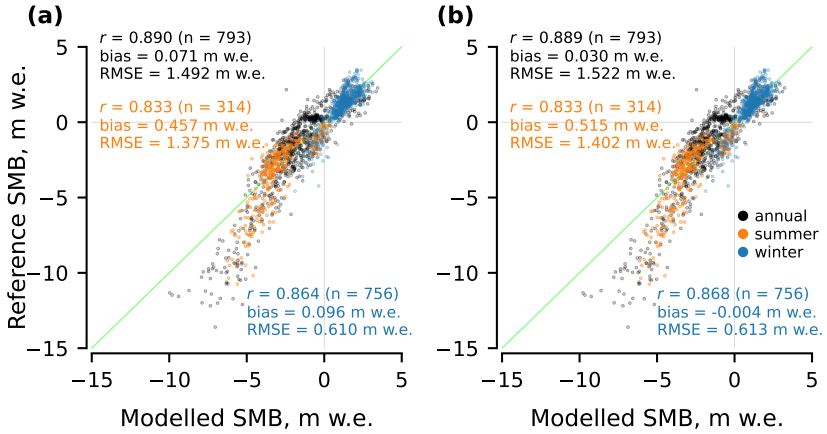

Figure 1: Point-level performance of (a) the classical TI formulation and (b) the autodiff-friendly TI formulation on test glaciers.

We suggest summarising this experiment as an additional Appendix in the revised manuscript, as it does not affect the conclusions of the study, which focus on comparing the spatial generalisation of TI, the GRU and GlaUnTI.

Validation Strategy

Reviewer: While the authors define six stratified folds, performance is reported for only one designated test fold. A full cross-validation framework where each fold serves as the test set in turn would yield more reliable and generalisable performance estimates, particularly given the limited test set size of 13 glaciers.

Response: We appreciate the reviewer raising this concern. We agree that a full fold-rotation cross-validation would provide an additional estimate of the sensitivity of the reported metrics to the particular split. Nevertheless, we consider the presented validation protocol to be robust for the main objective of the study, which is to assess spatial transferability to unseen glaciers. We use 13 glaciers or, to be more precise, 378 glacier-year pairs, 793/756/314 (annual/winter/summer) point measurements and 312/235/233 glacier-wide estimates for testing. Notably, our split is glacier-based, providing a completely independent spatial test set. For comparison, van der Meer et al. (2025) used nested cross-validation but considered 13 Swiss glaciers in total, and Sjursen et al. (2025) used a testing subset similar to ours with 14 glaciers withheld for testing in Norway. Other relevant regional SMB studies have also used validation protocols that do not take into account spatial autocorrelation between point measurements (e.g., Schuler et al., 2020; Anilkumar et al., 2023). Finally, performing full cross-validation would be computationally demanding in our case, as it would require retraining all four models for each fold ($\sim$ 864 hours or 36 days). Given the reasonable size of the dataset already used for final testing, we did not perform the full cross-validation. Therefore, our validation design is in line with recent practice in glacier SMB modelling, while also using a strict spatial split and *a multi-regional test set.*

Clarification on modelling framework

Reviewer: Further clarification is needed on both the model inputs and the SMB datasets used. It is unclear whether inputs are provided at daily time steps with the loss computed over aggregated annual or seasonal periods, and whether this aggregation follows hydrological year boundaries. Additionally, it is not specified how glacier-wide SMB estimates were derived—whether through spatial interpolation of point mass balance measurements or from geodetic estimates—nor how glacier-wide predictions are obtained from the models.

Response: We thanks the reviewer for raising this concern. All models use daily timesteps, and the loss is computed over aggregated periods that follow the dates of hydrological seasons reported in the datasets we compiled. We do not derive glacier-wide SMB estimates ourselves, instead we only use the ones reported in the original data sources. We hope these additional clarifications (here and below) eliminate the confusions, with some already being clear in our perspective:

> We harmonised all data to be largely compatible with the WGMS format. We distinguish between point SMB measurements, which are direct glaciological observations, and glacier-wide SMB estimates, which are derived *by the corresponding monitoring programmes* through spatial extrapolation of the glaciological observations over the glacier area. The latter are therefore used here as a weak constraint rather than as direct measurements in our modelling framework. All point measurements ...
>
> ...
>
> Regardless of the model, the total modelled glacier-wide SMB for a time period y (corresponding to a hydrological year or a season as defined in the dataset) is:
>
> $$\widehat{\mathrm{SMB}}_y^{\mathrm{total}} = \frac{\sum_{ij} A_{ij}\, \mathrm{GM}_{ij\,y-1}\, \widehat{\mathrm{SMB}}_{ijy}^{\mathrm{cell}}}{\sum_{ij} A_{ij}\, \mathrm{GM}_{ij\,y-1}},$$
>
> where $A_{ij} = 0.01$ km^2 is the area of one grid cell coming from the elevation grid (constant for all i and j), **GM** stands for the glacier outline mask, and $\widehat{\mathrm{SMB}}_{ijy}^{\mathrm{cell}}$ is the aggregated SMB for a particular grid cell ij over all days d in the simulation period y ($d|y$), given as a sum of daily $\widehat{\mathrm{SMB}}_{ijd}^{\mathrm{cell}}$:
>
> $$\widehat{\mathrm{SMB}}_{ijy}^{\mathrm{cell}} = \sum_{d|y} \widehat{\mathrm{SMB}}_{ijd}^{\mathrm{cell}}\,.$$
>
> Simulation periods y follow hydrological season bounds reported in the compiled dataset.
>
> ...
>
> All models, their training configurations and initial weight setups are summarised in Table 2. All four models are trained using daily temperature and precipitation inputs and the same loss definition. The training is done by minimising the following loss function on the glaciers from the training subset: ...

Glacier Facies Map

Reviewer: The descriptions of the glacier facies map included as a predictor take the focus away from the primary selling points and don't add much value to the manuscript. The fewer data points available for the validation constrain robust statistical analysis. Further, one can argue that it is a rather arbitrary inclusion, with more physically relevant ones possible, such as shortwave radiation, albedo, and wind.

Response: We agree that the statistical evidence for the added value of facies maps is more limited than for the main model comparison, because facies maps are only available for a subset of glaciers and years. At the same time, we do not consider their inclusion arbitrary. Nominally, point summer and glacier-wide annual/summer improvements are statistically significant ($p < 0.05$ with a paired t-test). Also, end-of-ablation-season facies maps contain information about the spatial position of the snowline region, accumulation area extent and surface type, all of which are closely related to SMB (Rabatel et al., 2017; Rabatel et al., 2005; Hock et al., 2007; Drolon et al., 2016). They also provide an observational proxy for surface albedo contrasts between snow, firn, bare ice and debris, and are thus particularly relevant for correcting summer melt.
Importantly, the effect of facies-map assimilation is not limited to the exact years in which facies maps are available. Because the model is run as a continuous daily trajectory and the SWE state is propagated through time, a correction informed by a facies map can influence subsequent melt partitioning, snow storage and SMB in later years as well. This is consistent with the observed improvement of Model D over Model C in summer

SMB, including in years without directly available facies maps. We agree that additional physically relevant predictors such as shortwave radiation, thermal infrared information, albedo and wind could further improve the model. Adding these variables, however, would move the formulation towards a more complete surface energy balance model, whereas the present study deliberately keeps the physical backbone close to a TI formulation and evaluates whether a neural corrector can improve its transferability. We therefore mention these variables as promising future extensions rather than expanding the present model.
We expanded our discussion to highlight these points:

> Inclusion of the glacier facies maps into GlaUnTI yields moderate improvement in predicting summer glacier-wide SMB, yet the benefits for other target variables remain marginal. Facies products likely act as surrogate indicators for multiple unresolved processes simultaneously, including late-summer albedo history, firn retention capacity, debris cover , ~~and~~ exposure of bare ice, and the position of the snowline region, rather than facies being a causal driver per se. This is physically motivated because accumulation area extent and snowline position are closely linked to SMB (Rabatel et al., 2017; Rabatel et al., 2005; R. Hock et al., 2007; Drolon et al., 2016), while the facies classes also provide an observational proxy for surface-albedo contrasts between snow, firn, bare ice and debris. This reframes facies assimilation as a pragmatic way to inject remote sensing constraints on the state of the surface into an SMB model when direct energy balance forcing is unavailable. Moreover, because GlaUnTI propagates SWE through time, facies-based corrections can affect subsequent snow storage, melt partitioning and SMB beyond the exact years in which facies maps are available. This is consistent with the improvement of Model D over Model C in summer SMB even for years without directly available facies maps. ~~Moreover, alternative remote sensing products might play similar roles, potentially with better temporal coverage than the end-of-summer facies maps, e.g., late summer albedo, transient snowline altitude, bare ice exposure duration or SAR melt intensity time series (as in Scher et al., 2021).~~ The weak statistical separation between Models C and D here is also affected by the limited sample size ($n = 32$ annual point measurements, $n = 27$ annual glacier-wide estimates), limiting the ability to resolve modest but potentially real gains. Alternative remote sensing products might play similar roles, potentially with better temporal coverage than the end-of-summer facies maps, e.g., transient snowline altitude, bare-ice exposure duration or SAR melt intensity time series (as in Scher et al., 2021). Future extensions could also include more direct energy balance predictors, such as shortwave radiation, thermal infrared information and wind, but doing so would move the present TI model closer to a complete surface energy balance formulation.

Minor comments

This section is divided into subsections that distinguish between suggestions in technical content and those associated with minor rephrasing and flow.

Sentences that can be rephrased for clarity and flow

Reviewer: Line 14-18: 'Including glacier facies... SMB trajectories.' These sentences are a bit abrupt and disconnected. The flow can be improved for readability.

Response: Thank you for your comment. We tried to smooth the phrasing:

> ...Including glacier facies maps from the end of the ablation season to the corrector yields moderate benefits in glacier-wide summer (11.0%) and annual (12.2%) SMB estimates. ~~We found that~~In contrast, the purely data-driven baseline model overall shows the weakest spatial transferability. ~~Also~~Beyond modelling accuracy, end-to-end differentiability enables efficient gradient-based calibration, transfer learning, inverse optimisation of effective forcing perturbations, formal model explainability and propagation of forcing-driven aleatoric uncertainty through long SMB trajectories...

Reviewer: Line 46 onwards: For the sake of completeness, it might be worth talking about how geodetic datasets are used for calibration as well.

Response: We thank the reviewer for this suggestion. We have added a short clarification that geodetic mass balance products are commonly used to constrain modelled multiannual glacier-wide mass changes, typically by tuning melt and accumulation-related parameters so that the simulated cumulative mass balance matches the geodetic estimate. We suggest the following modification:

> Despite their widespread use, the predictive accuracy of TI models depends critically on the calibration of a small number of empirical parameters, which are most often estimated independently for individual glaciers using local in situ or geodetic mass balance observations (Hock et al., 2003; Rounce et al., 2023). When geodetic datasets are used, DDFs and precipitation factors are commonly tuned, typically through

grid search, so that the simulated glacier-wide cumulative mass balance matches the geodetic estimate over the survey period. This glacier-specific calibration limits the transferability of TI parameter sets across glaciers and climate regimes, challenging the "one model works everywhere" paradigm. Moreover, global geodetic mass balance products can be noisy at the individual glacier scale and represent an integrated mass change signal that is not strictly equivalent to SMB (Fischer, 2011). ...

Reviewer: Lack of clarity and consistency in equation symbols. For example, `cell` superscript is used in certain equations like eq (5) and skipped in others. Perhaps it is not essential as i and j subscripts represent the cell uniquely? Is A_{ij} coming from the elevation grids? m and n are not defined in eq (17)

Response: We agree with the reviewer that our notation could be confusing in the initial submission. In the new submission, we will drop `cell` superscript in all equations as it is indeed unnecessary. To avoid potential confusions, we also suggest replacing SMB with y in Equation (18) and adjusting the description accordingly:

..... , where ~~SMB_i~~y_i is the reference SMB value, ~~and $\widehat{\mathrm{SMB}}_i$~~$\widehat{y}_i$ is the model output, and $\overline{\cdot}$ stands for the mean value.

Yes, A_{ij} comes from the elevation grids, which we clarify as follows:

..., where $A_{ij} = 0.01\ \mathrm{km}^2$ is the area of one grid cell coming from the elevation grid (constant for all i and j), ...

In Equation (17), m and n refer to the number of point SMB measurements and glacier-wide SMB estimates:

..., where $\mathrm{SMB}^{\mathrm{point}}_{ijy}$ are measured annual/winter/summer point SMB mapped to the corresponding grid cell, $\mathrm{SMB}^{\mathrm{total}}_{y}$ are glacier-wide SMB estimates, m and n are the number of point measurements and glacier-wide estimates, respectively, ...

Reviewer: Line 133: Should be 30th September, not 31st.

Response: Thanks for noticing this typo. We corrected it:

We extracted the elevation estimates referenced to the ~~31st~~30th of September for each year, corresponding to the end of the hydrological year.

Specific Technical Recommendations

Reviewer: Abstract: The ability of the models to be spatially transferable is not represented clearly here.

Reviewer: Line 9-10: What do you mean by heterogeneous regions?

Response: Thank you for pointing this out. We revise the sentence to clarify what we mean by "heterogeneous regions" and to highlight the focus on spatial transferability as follows:

Their ~~performance~~spatial transferability is evaluated on a held-out, spatially independent test subset of 13 glaciers across ~~heterogeneous regions~~climatically and geometrically contrasting European regions.

Reviewer: Line 94-98: Include the time periods during which the other studies were carried out as well.

Response: We thank the reviewer for this comment and suggest the following modification to the manuscript:

Our study area comprises 78 glaciers distributed across five European regions—the European Alps (38 glaciers), Scandinavia (21), Svalbard (9), Iceland (8) and the Pyrenees (2). In recent decades, glaciers across these study regions have shown persistently negative balances. In the European Alps and Pyrenees, regional geodetic assessments indicate mean losses of about -0.74 m w.e. a^{-1} for 2000–2016 (Davaze et al., 2020)

and -0.59 m w.e. a^{-1} for 2011–2020 (Vidaller et al., 2021), respectively. Typical mass balance magnitudes of order -1 m w.e. a^{-1} were observed in Iceland for 1994–2019 (Aalgeirsdóttir et al., 2020) and around -0.21 m w.e. a^{-1} in Svalbard for 2000–2019 (Schuler et al., 2020), consistent with generally negative mass balance in Scandinavia since 2000 (Andreassen et al., 2020) and recent global assessments of glacier mass loss in the last decades (Hugonnet et al., 2021; Zemp et al., 2025). Figure 1 shows ...

Reviewer: Line 165: How was the precipitation downscaling performed?

Response: We thank the reviewer for this question. No precipitation downscaling in the sense of an elevation-dependent precipitation lapse rate was applied. The ERA5-Land precipitation field was resampled to the 100 m modelling grid using nearest neighbour resampling:

> ..., while precipitation (**P**) was resampled using nearest neighbour interpolation to avoid artificial smoothing of precipitation amounts...

We chose not to impose a universal precipitation lapse rate correction because precipitation gradients over mountain glaciers are highly region-dependent and would introduce an additional assumption that is difficult to constrain consistently across the full pan-European dataset. Instead, precipitation magnitude is adjusted during calibration through the trainable precipitation sensitivity parameter (β_1) in line with Maussion et al. (2019), and, in GlaUnTI, through the learned multiplicative precipitation correction ($\Delta^{(1)}_{ijy}$).

Reviewer: Line 166: What are the uncertainties associated with Z? What are the sensitivities to the annual elevation maps for each of the models?

Response: We thank the reviewer for pointing this out. The annual elevation maps used in the study are provided together with a spatially distributed 1-σ uncertainty estimate, denoted as Z_σ. To assess the sensitivity of the four models to this source of uncertainty, we performed an additional uncertainty propagation experiment (similar to Appendix A3 but for elevation) for Grosser Aletsch as an example. We selected three grid cells representing different elevation zones of the glacier, corresponding to the cells at the 25th, 50th and 75th percentiles of the glacier elevation distribution, i.e. approximately the ablation, median-elevation and accumulation parts of the glacier.
For each annual elevation map, we introduced a scalar standard-normal latent variable ξ_y and perturbed the elevation field as

$$Z^*_{ijy} = Z_{ijy} + \xi_y Z_{\sigma,ijy}, \qquad \xi_y \sim \mathcal{N}(0,1)\,.$$

This formulation corresponds to a spatially coherent elevation-error realisation for each annual map. We consider this preferable to assuming independent cell-wise errors, because no defensible spatial covariance model for the elevation errors is available—this does not present a problem for model comparison. The perturbation was propagated through the complete model input pipeline. In Models A and B, this affects SMB only through the lapse-rate correction of near-surface air temperature. In Models C and D, the same perturbed elevation field was additionally propagated through the GlaUnTI corrector inputs, including the normalised elevation predictor and the auxiliary quantities derived from the perturbed temperature/elevation fields. The resulting 1-σ uncertainty bands are shown in Fig. 2. Models A and B show relatively weak sensitivity because elevation enters only through the lapse-rate temperature correction. Models C and D show stronger sensitivity, because these models also use elevation and TI-predicted SMB as learned predictors in the network, particularly outside the 2000–2019 period where the elevation maps are extrapolated.
This result is therefore expected and mainly confirms that the hybrid corrector models are more responsive to uncertain geometry inputs than the simpler baselines and rely on elevation as an informative feature for corrections. Since this experiment is a diagnostic sensitivity check rather than a central component of the modelling framework, and because the rebuttal itself will remain available and citable with a DOI, we prefer to report it here rather than adding another supplementary analysis to the manuscript.

Reviewer: Glacier facies map: It is not clear what the classes generated by the algorithm is? What prediction algorithm was used to generate it?

Response: We thank the reviewer for this comment. We suggest the following modificatons to the manuscript.

> Hence, in cases of available optical satellite imagery from Landsat or Sentinel-2 at the end of summer (31st July–29th September), we employed the ~~glacier facies classification models~~compact convolutional neural networks trained in the cross-validation folds of Maslov et al. (2026) to classify glacier facies and ensembled their predictions ~~as follows~~. This classification workflow yields five glacier facies classes (ice, snow, debris, firn and refrozen-like) and three auxiliary classes (shadow, water and cloud). For each pixel and class c,

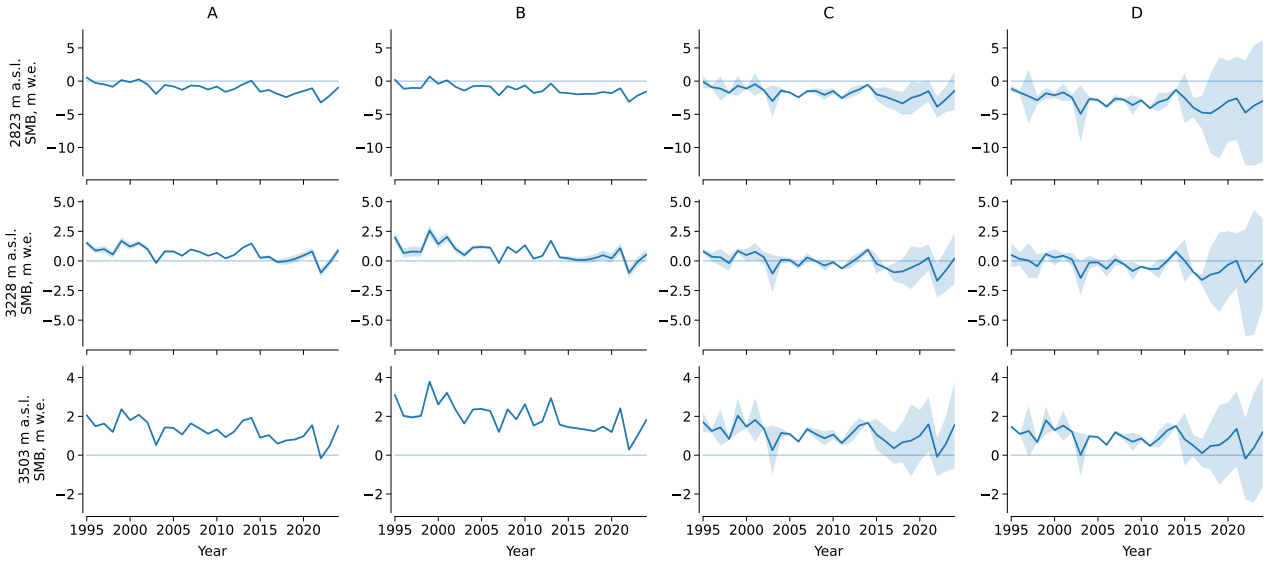

Figure 2: Sensitivity of annual point SMB predicted by all four models at three locations on Grosser Aletsch to uncertainty in the annual elevation maps. Rows correspond to points closest to the 25th, 50th and 75th elevation percentiles of the glacier, and columns correspond to Models A–D. Shaded bands denote $\pm 1\sigma$ uncertainty propagated from Z_σ.

> we obtained aggregated predictive confidence as: ...

We also presented several examples of the derived glacier facies maps and how they compare with the distributed annual SMB (see below).

Reviewer: What is the procedure by which the glacier-wide mass balance was estimated?

Response: We did not derive glacier-wide SMB estimation from the point measurements ourselves. Instead, we used only the numbers directly reported in the public datasets (WGMS, 2021):

> ... We harmonised all data to be largely compatible with the WGMS format. We distinguish between point SMB measurements, which are direct glaciological observations, and glacier-wide SMB estimates, which are derived *by the corresponding monitoring programmes* through spatial extrapolation of the glaciological observations over the glacier area. The latter are therefore used here as a weak constraint rather than as direct measurements in our modelling framework. All point measurements ...

To spatially aggregate modelled SMB cells into glacier-wide estimates, Equation (4) was used:

> $$\widehat{\mathrm{SMB}}_y^{\mathrm{total}} = \frac{\sum_{ij} A_{ij}\, \mathrm{GM}_{ij\,y-1}\, \widehat{\mathrm{SMB}}_{ijy}}{\sum_{ij} A_{ij}\, \mathrm{GM}_{ij\,y-1}},$$
>
> where $A_{ij} = 0.01$ km^2 is the area of one grid cell coming from the elevation grid (constant for all i and j), **GM** stands for the glacier outline mask, and $\widehat{\mathrm{SMB}}_{ijy}$ is the aggregated SMB for a particular grid cell ij over all days d in the simulation period y...

Reviewer: Line 215 to 220: What's the basis for this?

Response: Many classical two-DDF implementations use discrete snow/ice surface classes or threshold-based transitions, which can lead to abrupt changes in the applied melt factor (Huss and Hock, 2015; Rounce et al., 2023). This is not smooth and does not realistically represent grid cells with mixed snow and ice. Instead, we propose a switch mechanism based on fractional snow cover (FSC) in Equation (8), which changes from 0 to 1 smoothly based on the accumulated column of snow. For this, we track the snow column via SWE, and FSC depends on it, Equation (9):

$$\mathrm{FSC}_{ijd} = \frac{(\mathrm{SWE}_{ij\,d-1})^{\tau_{\mathrm{FSC,s}}}}{(\mathrm{SWE}_{ij\,d-1})^{\tau_{\mathrm{FSC,s}}} + (\tau_{\mathrm{FSC,c}})^{\tau_{\mathrm{FSC,s}}}} = \sigma\left(\tau_{\mathrm{FSC,s}}\left[\ln \mathrm{SWE}_{ij\,d-1} - \ln \tau_{\mathrm{FSC,c}}\right]\right),$$

where $\tau_{\mathrm{FSC,s}}$ and $\tau_{\mathrm{FSC,c}}$ are learnable parameters (steepness and centre) allowing to obtain a *very flexible* family of curves (see Fig. 3). All these curves are bounded between 0 and 1, satisfying the physical bounds of FSC, are monotonically increasing with $\mathrm{FSC}|_{\mathrm{SWE}=0} = 0$, which is also a property we want to have here as more snow should lead to higher FSC (and no snow means FSC = 0), and are smooth everywhere.

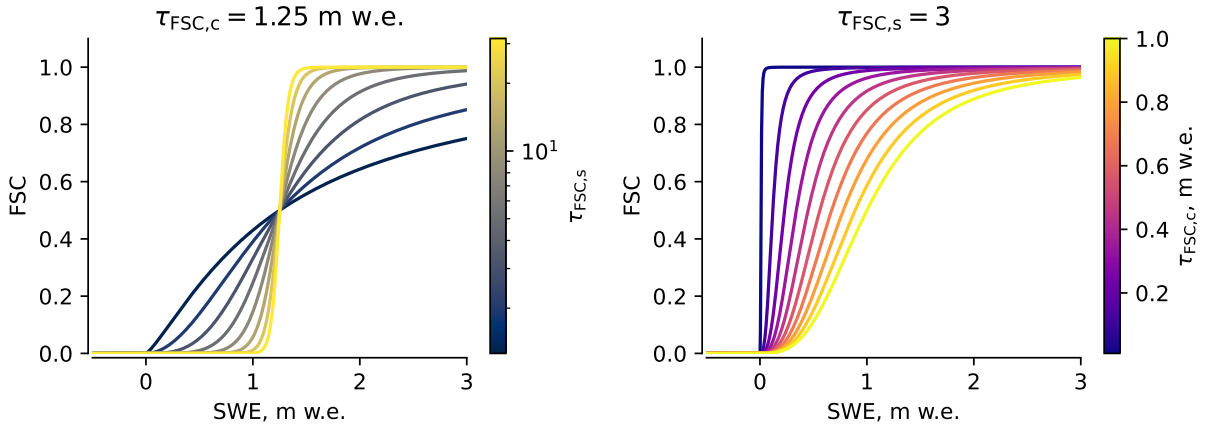

Figure 3: Hill curve shapes depending on (a) $\tau_{\mathrm{FSC,s}}$ and (b) $\tau_{\mathrm{FSC,c}}$.

We assume that many alternative formulations could have been explored, but their exploration remains out of the scope of our study. We suggest the following clarification in the revised manuscript:

> ..., where $\mathrm{SWE}_{ij\,d-1}$ is the snow water equivalent (SWE) at the previous modelling day, and $\tau_{\mathrm{FSC,s}}$ and $\tau_{\mathrm{FSC,c}}$ stand for, respectively, the steepness and the centre of the snow depletion curve. We use this functional form because it gives a flexible two-parameter mapping from SWE to FSC. The parameter $\tau_{\mathrm{FSC,c}}$ defines the SWE at which FSC = 0.5, while $\tau_{\mathrm{FSC,s}}$ controls the sharpness of the transition. The formulation is bounded between 0 and 1 and increases monotonically with SWE, ensuring physical consistency. In our base TI model, the FSC acts as a first-order proxy for the grid cell albedo. ...

Reviewer: Could authors confirm that normalisation was performed before training of the ML models?

Response: Yes, input normalisation was employed for the neural components. For the GRU baseline, the daily temperature and precipitation inputs were divided by fixed maximum-absolute scaling constants, 45.744850158691406 °C and 0.16851751506328583 m w.e., respectively, both specified in `constants.py`. For the neural corrector in GlaUnTI, the daily temperature and precipitation fields entering the TI backbone were not normalised, because these variables appear in the physical TI equations in their native units. However, the auxiliary inputs of the corrector were scaled: Z, Z_σ, T_M, and P_M were divided by their corresponding maximum absolute values, which are provided with the released dataset in `normalisation_factors.pkl` at `https://doi.org/10.4121/5ea53bc3-2c85-42bb-89d1-606c8ed1d80a.v1`. The initial TI SMB estimate (SMB_0) was divided by 10 m w.e. The perturbation-derived sensitivity fields, binary glacier mask, and facies/confidence predictors were kept in their original ranges. We clarify this in the revised manuscript as follows:

> ..., where $\boldsymbol{x}_{ijd} = [T_{ijd}, P_{ijd}]^\top$ is the input vector, $\boldsymbol{z}$ is the update gate, $\boldsymbol{r}$ is the reset gate, $\widehat{\boldsymbol{h}}$ is the candidate hidden state, $\boldsymbol{h}$ is the hidden state, $\odot$ stands for the Hadamard product, and $\mathbf{W}_z, \mathbf{W}_r, \mathbf{W}_h, \mathbf{U}_z, \mathbf{U}_r, \mathbf{U}_h, \boldsymbol{b}_z, \boldsymbol{b}_r, \boldsymbol{b}_h$ are trainable parameters. In the implementation, the GRU receives normalised inputs, corresponding to the temperature and precipitation values divided by their maximum-absolute values. Similar to the TI model, we initialise $\boldsymbol{h}_0$ as a trainable vector at the year $y-5$ and run the simulation for five consecutive years to obtain the starting conditions for the period of interest...
>
> ...
>
> ...This simple resampling is computationally efficient and provides a regularising effect due to the enforced smoothness of the outputs. Before entering the corrector, $\mathbf{Z}$, $\mathbf{Z}_\sigma$, $\mathbf{T}_\mathrm{M}$ and $\mathbf{P}_\mathrm{M}$ are divided by feature-specific maximum absolute values. $\mathbf{SMB}_0$ is divided by 10 m w.e., corresponding to the typical magnitude order of measured point SMB. The perturbation-derived sensitivity fields, glacier masks and facies predictors are kept in their original ranges. Notably, Eq. (6) remains the main backbone of the GlaUnTI model, and the overparameterised corrector only influences its core parameters, hence, keeping the interpretability of the model high by design...

Reviewer: It's unclear what the training process for the GRU is compared to the TI models. Is it daily inputs with losses computed the same way over time?

Response: All four models were trained similarly, using daily inputs and the same loss definition. We emphasise this moment in the revised manuscript:

> All models, their training configurations and initial weight setups are summarised in Table 2. All four models are trained using daily temperature and precipitation inputs and the same loss definition. The training is done by minimising the following loss function on the glaciers from the training subset: ...

Reviewer: Equation (18): Can the coefficient of determination be used in addition to or in place of Pearson's r? The Pearson r is a suitable metric if you aim to evaluate the linear association between variables. However, as a goodness-of-fit test, the coefficient of determination is the preferred metric as it explains how much of the variance in the observations (or reference set) is being represented by the model. Note that the coefficient of determination is not always the square of Pearson's r and can take negative values when the model performs worse than a model that naively predicts the mean.

Response: We thank the reviewer for this suggestion. We agree that R^2 is a useful summary metric in regression settings and note that, in predictive evaluation contexts, it captures a different aspect of performance than r. R^2 is given as:

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \overline{y})^2} = 1 - \frac{\mathrm{RMSE}^2}{\mathrm{Var}\,[y]},$$

where SSE is sensitive to both the scale of random errors and the systematic offset. In our manuscript, we report r, RMSE and bias jointly. This combination separates agreement in variability (r), magnitude of errors and systematic offsets (RMSE and bias), whereas R^2 tries to aggregate these effects into a single metric.
We report r, RMSE and bias for several reasons:

- These three metrics are more informative than R^2 when reported together
- R^2 fails in low-variance regimes, and can give negative values even when predictions remain meaningful (Onyutha, 2021)
- Such decomposition of errors follows general recommendations in the literature (Gupta et al., 2009; Mathevet et al., 2023)

Reviewer: Line 332-335: Might be worth checking if GPU is being used during the training. For 50 epochs on the small parameter size and dataset size, training is typically much faster. If GPUs are being used, check if jit pipeline breaks occur during training.

Response: We thank the reviewer for raising this concern. GPU is of course utilised during training. On average, the usage of GPU fluctuates around 70%, with $\sim 80\%$ for large grids and $\sim 30\%$ for small ones—this difference likely comes from an admittedly not-very-JIT-friendly implementation of the loss function. Nevertheless, the model calls are JIT-compiled, as well as some parts of the loss function.
While the parameter size remains small, the data volume is large, leading to considerable computational load. For reference, 30 years of daily time series of temperature and precipitation for a 442×575 grid (Bruarjokull) would require $30 \times 365 \times 2 \times 442 \times 575 / 2^{28} \approx 21$ GB of memory, when using float32 and uncompressed. Repeatedly running such data through GPU and (especially) calculating gradients takes time, no matter how simple the model is. We also rely on rematerialisation a lot to make gradient computation possible at all, which leads to extra recalculations during training.

Reviewer: Note on the number of epochs trained. 50-100 epochs used for training, as depicted in Table 2, and grid search training with a couple of epochs, as depicted in line 312, is generally not sufficient for most machine learning training, which can explain the poorer performance of GRU. This can be checked by plotting the loss vs epoch curve for training and validation. If both the training and validation losses continue to decrease, then training is not complete. If limited epochs were selected due to the time taken for training, see the previous comment. This may suggest that GPU acceleration or JIT compilation is not being fully utilised

Response: We thank the reviewer for raising this point. Of course, we monitor loss-epoch curves during training. Because of that, we used 100 epochs to train the GRU, as 50 epochs were not enough (unlike for the rest of the models). After 100 epochs, the GRU had almost reached the plateau, and we did not expect significant improvements after that. Nevertheless, as both reviewers raised the understandable concern of not having enough epochs to train the GRU, we increased the computational budget for the GRU to 300 epochs (~ 36 hours). This led to an improvement in the validation performance (Fig. 4).
Ironically, while the validation performance improved, the testing performance slightly dropped, indicating a lower tolerance of the purely data-driven baseline to domain shifts (Table 1).

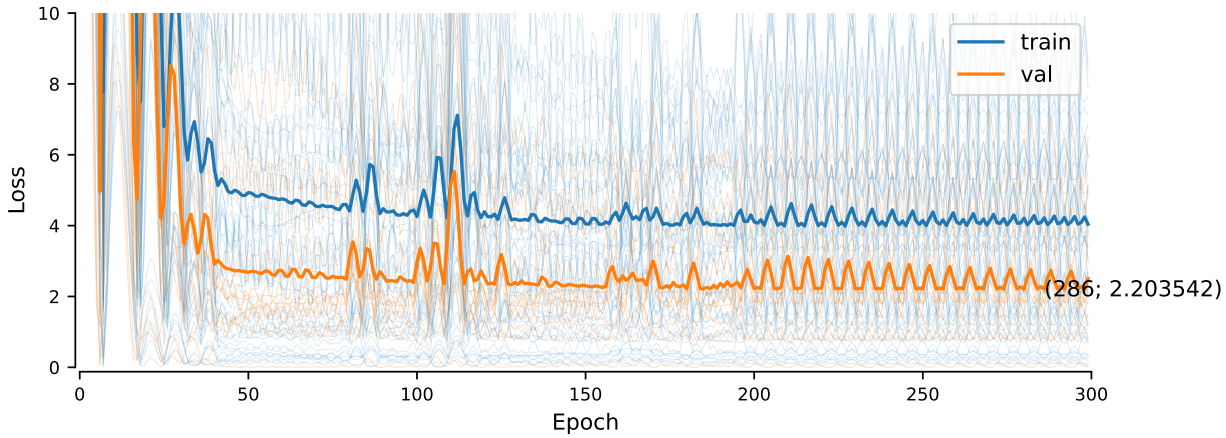

Figure 4: Loss function dynamics for Model B (the GRU). The plot area before Epoch 100 corresponds to the initial submission. Semi-transparent curves show loss dynamics for individual glaciers. The best validation loss value is labelled.

Table 1: RMSE of Model B (the GRU) on the test subset before and after training for more epochs.

Type	Period	Before, m w.e.	After, m w.e.
Point	Annual	2.032	2.066
	Summer	2.012	1.969
	Winter	0.869	0.875
Glacier-wide	Annual	0.930	1.009
	Summer	0.752	0.804
	Winter	0.904	0.921

Given that the validation loss is lower after longer training, we *have to* report the results for this new model in the revised manuscript. We therefore will update all figures for Model B in the manuscript. This, however, doesn't change any derived conclusions.

Reviewer: What do the model predictions look like over the glaciers? Can this be represented as a figure over a few of the test glaciers?

Response: Thank you for this valuable suggestion. We prepared such a figure. To make it more informative, we also provided visualisations of SWE states and glacier facies maps next to the modelled SMB maps. We suggest adding the following subsection to the Results section of the manuscript:

Qualitative assessment of modelled SMB

Fig. 5 provides qualitative examples of the spatial structure of annual SMB fields predicted by Model D for three glaciers from the test subset. The modelled SMB patterns are spatially coherent and follow the expected elevation-dependent structure, with negative SMB over lower ablation areas and positive SMB over upper accumulation areas. The modelled equilibrium line positions (indicated by the SMB = 0 contour) are also broadly consistent with the glacier facies maps acquired near the end of the ablation season: snow tends to occur above or near the modelled equilibrium line. The accompanying SWE fields provide an additional consistency check. The largest end-of-period snow storage is generally located in the upper accumulation areas and decreases towards the ablation zones. The point-level errors shown in the SMB panels indicate that the spatially distributed fields are locally consistent with available in-situ measurements, although individual deviations remain. An important nuance is that the SMB/SWE states and facies maps are not always perfectly co-temporal. This temporal mismatch is particularly visible for Hansbreen (Fig. 5c), where the facies map precedes the SMB end date by almost three months, meaning that subsequent late-season accumulation can affect the SMB and SWE states without being represented in the facies map.

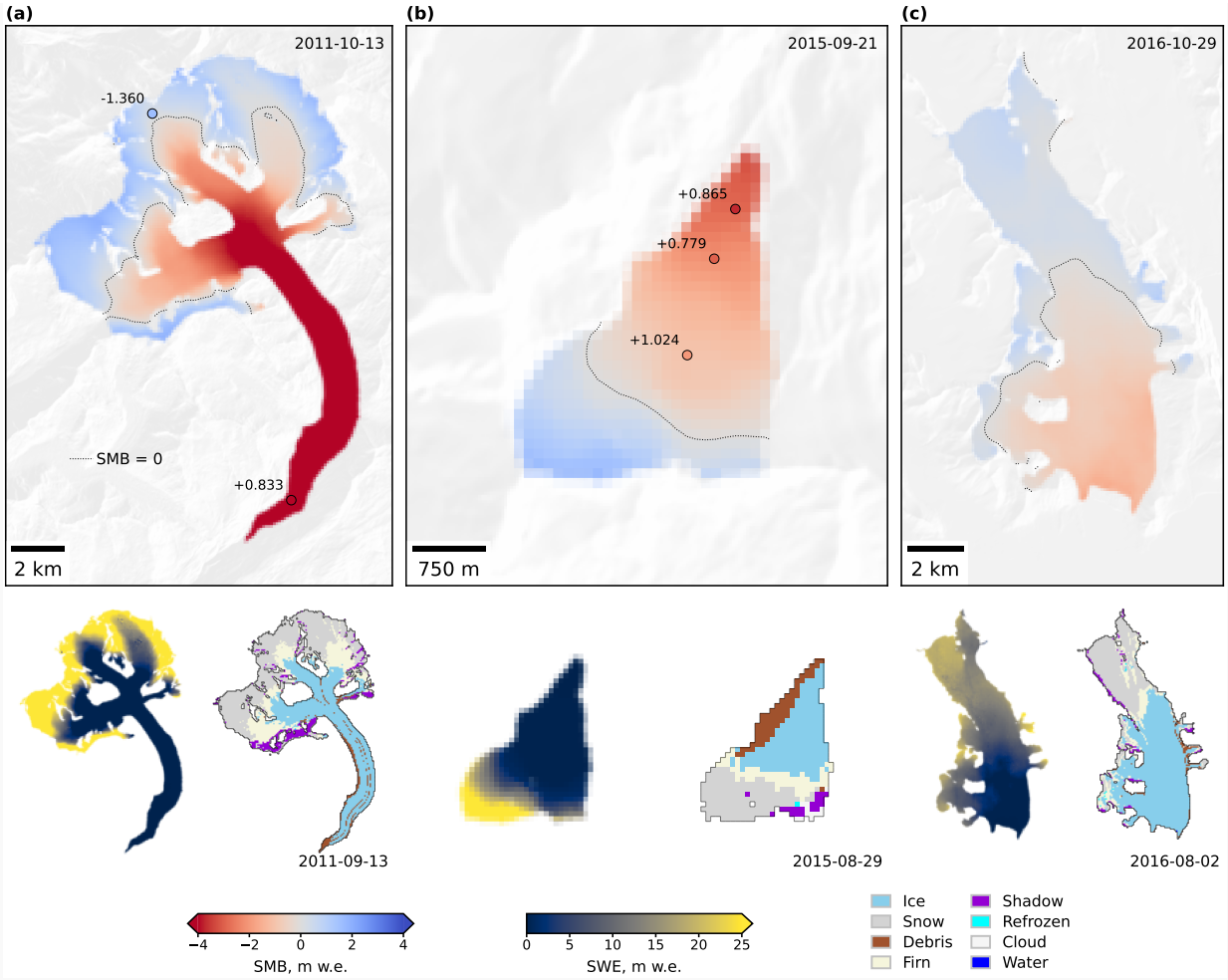

Figure 5: Annual SMB maps produced by Model D for: (a) Grosser Aletsch in 2011, (b) Schwarzberg in 2015 and (c) Hansbreen in 2016. Each panel shows the predicted annual SMB field (top), SWE (bottom left) and a glacier facies map derived from optical satellite imagery acquired near the end of the ablation season (bottom right). Dates in the SMB panels indicate the end date of the annual SMB measurement period, while dates in the facies maps indicate the acquisition dates of the corresponding satellite images. Point markers show available in-situ SMB measurements. Labels next to the points show point-level prediction errors in m w.e., calculated as predicted SMB minus measured SMB.

Reviewer: I also recommend that the integrated gradients-based explainability described in Appendix A2 can be included as a part of the main manuscript to highlight the advantages of this differential form of the temperature index model.

Response: We thank both reviewers for highlighting this point. We, however, decided to keep these experiments in the appendix. The central aim of the manuscript is to introduce and evaluate GlaUnTI as a transferable SMB model, with the main results focused on predictive performance, spatial transferability and the comparison with the TI and data-driven baselines. The differentiability experiments are intended as proof-of-concept demonstrations of additional capabilities enabled by the autodiff-friendly implementation, rather than as full-scale case studies evaluated across the complete glacier sample. Moving them into the main results would risk shifting the emphasis of the paper towards these demonstrations relative to the main transferability analysis. We think that keeping them in Appendix A optimally preserves the narrative flow of the manuscript, while the main text makes clear that differentiability is an important methodological implication of the overall framework.

References

Anilkumar, R. et al. (2023). "Modelling point mass balance for the glaciers of the Central European Alps using machine learning techniques". In: *The Cryosphere* 17.7, pp. 2811–2828. DOI: 10.5194/tc-17-2811-2023. URL: https://tc.copernicus.org/articles/17/2811/2023/.

Drolon, V. et al. (Oct. 2016). "Monitoring of seasonal glacier mass balance over the European Alps using low-resolution optical satellite images". In: *Journal of Glaciology* 62 (235), pp. 912–927. ISSN: 0022-1430. DOI: 10.1017/JOG.2016.78. URL: https://www.cambridge.org/core/journals/journal-of-glaciology/article/monitoring-of-seasonal-glacier-mass-balance-over-the-european-alps-using-lowresolution-optical-satellite-images/2AD9341322E5B40A9FDA03A6CB6B2E5E.

Gupta, Hoshin V. et al. (2009). "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling". In: *Journal of Hydrology* 377.1, pp. 80–91. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2009.08.003. URL: https://www.sciencedirect.com/science/article/pii/S0022169409004843.

Hock, R., D.-S. Kootstra, and C. Reijmer (2007). "Deriving glacier mass balance from accumulation area ratio on Storglaciären, Sweden". In: *7th Scientific Assembly of the International Association of Hydrological Science, IAHS - Workshop on Andean Glaciology and Symposium on the Contribution from Glaciers and Snow Cover to Runoff from Mountains in Different Climates*. 318, pp. 163–170. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-38549156622&partnerID=40&md5=abe3ce04fbedff2850dd16badab023de.

Huss, Matthias and Regine Hock (2015). "A new model for global glacier change and sea-level rise". In: *Frontiers in Earth Science* Volume 3 - 2015. ISSN: 2296-6463. DOI: 10.3389/feart.2015.00054. URL: https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2015.00054.

Mathevet, Thibault et al. (2023). "Multi-objective assessment of hydrological model performances using Nash–Sutcliffe and Kling–Gupta efficiencies on a worldwide large sample of watersheds". en. In: *Comptes Rendus. Géoscience* 355.S1, pp. 117–141. DOI: 10.5802/crgeos.189.

Maussion, F. et al. (2019). "The Open Global Glacier Model (OGGM) v1.1". In: *Geoscientific Model Development* 12.3, pp. 909–931. DOI: 10.5194/gmd-12-909-2019. URL: https://gmd.copernicus.org/articles/12/909/2019/.

Onyutha, Charles (Nov. 2021). "A hydrological model skill score and revised R-squared". In: *Hydrology Research* 53.1, pp. 51–64. ISSN: 0029-1277. DOI: 10.2166/nh.2021.071. eprint: https://iwaponline.com/hr/article-pdf/53/1/51/995215/nh0530051.pdf. URL: https://doi.org/10.2166/nh.2021.071.

Rabatel, A., J. P. Dedieu, and C. Vincent (2005). "Using remote-sensing data to determine equilibrium-line altitude and mass-balance time series: validation on three French glaciers, 1994–2002". In: *Journal of Glaciology* 51 (175), pp. 539–546. ISSN: 0022-1430. DOI: 10.3189/172756505781829106. URL: https://www.cambridge.org/core/journals/journal-of-glaciology/article/using-remotesensing-data-to-determine-equilibriumline-altitude-and-massbalance-time-series-validation-on-three-french-glaciers-19942002/7EE925C092A3DB1E15EBDAE02E769F68.

Rabatel, A. et al. (May 2017). "Annual and seasonal glacier-wide surface mass balance quantified from changes in glacier surface state: A review on existing methods using optical satellite imagery". In: *Remote Sensing* 9 (5). ISSN: 20724292. DOI: 10.3390/rs9050507.

Rounce, D. R. et al. (2023). "Global glacier change in the 21st century: Every increase in temperature matters". In: *Science* 379.6627, pp. 78–83. DOI: 10.1126/science.abo1324. URL: https://www.science.org/doi/abs/10.1126/science.abo1324.

Schuler, T. V. et al. (2020). "Reconciling Svalbard Glacier Mass Balance". In: *Frontiers in Earth Science* 8. ISSN: 2296-6463. DOI: 10.3389/feart.2020.00156. URL: https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2020.00156.

Sjursen, K. H. et al. (2025). "Machine learning improves seasonal mass balance prediction for unmonitored glaciers". In: *The Cryosphere* 19.11, pp. 5801–5826. DOI: 10.5194/tc-19-5801-2025. URL: https://tc.copernicus.org/articles/19/5801/2025/.

van der Meer, M. et al. (2025). "A minimal machine-learning glacier mass balance model". In: *The Cryosphere* 19.2, pp. 805–826. DOI: 10.5194/tc-19-805-2025. URL: https://tc.copernicus.org/articles/19/805/2025/.

WGMS (2021). *Fluctuations of Glaciers Database. World Glacier Monitoring Service, Zurich, Switzerland.* DOI: 10.5904/wgms-fog-2021-05.