We sincerely thank the two anonymous reviewers for their encouraging feedback and helpful comments, and we appreciate the time and effort dedicated by both the reviewers and the editor. We have addressed all comments with our responses indicated in blue text below. We could reply to all comments. Line numbers in our response refer to a revised manuscript assuming all changes are accepted. Text in italic refers to text added or modified in the revised manuscript, where changes to the original text are emphasized in darker font.

Response to Anonymous Referee #2

Summary

This study builds a Bayesian calibration technique to calibrate the surface energy and mass balance model COSIPY. The main goal of this study is to investigate a calibration approach enabling parameter tuning for unmonitored glaciers, relying on satellite-derived observational datasets only. Furthermore, sources of parameter equifinality and uncertainty are investigated. The calibration framework relies on Latin Hypercube Sampling to explore the model's parameter behavior and establish priors for a Markov Chain Monte Carlo calibration.

The study is of high relevance for the energy balance modelling community, as calibration of complex energy balance models is a large challenge. The motivation and relevance of this study are well outlined in the introduction. The manuscript is overall well structured, and the figures are well presented. Overall, I would suggest minor revisions: I have two bigger comments and several specific comments which need addressing and revision before publishing the article.

We appreciate the constructive feedback and encouraging comments on our manuscript. Please find our response to each point below.

Major comments

1. Description of methods

The applied methods and conducted methodological steps are pretty complex. Therefore, more details should be provided at several points, e.g.:

The description of the statistical methods applied in section 2.4.1 & 2.4.2 (e.g., PCA, what are half-normal priors and why are they used). TC is a journal with a lot of readers from the glaciology community, which might not all be very familiar with these methods.
The surrogate models: I miss a more detailed description of this approach. It is an important part of the calibration scheme. I would suggest including an entire subsection and explain the entire procedure in detail, including also the performance testing.

We thank the reviewer for this comment, which overlaps with comments 4 and 5 of Reviewer 1. We have added brief explanations of the statistical methods and explained missing details where appropriate without extensively lengthening the manuscript. We also added a separate section detailing the surrogate creation and testing. We kindly refer the reviewer to the answers for Reviewer 1 for more details and have highlighted here our new surrogate model section and the explanation of the PCA and half-normal priors.

L284f:

We estimate the standard deviation $\sigma_{\eta,i}$ as constant for all ablation season data points for SLA and ā. We apply small-magnitude half-normal priors, which are normal distributions restricted to positive values. This is required for a standard deviation, and we chose the priors to be minimally invasive: [...]

L334ff:

The PCA extends this analysis by investigating all seven parameters simultaneously. It reduces the high-dimensional parameter space and collapses it into a set of independent components that capture the dominant patterns of variability based on the standardised parameter vectors of the COSMO-forced LHS ensemble. By examining the weight each parameter contributes to a component, we can identify which parameters vary together or oppose each other, revealing the structural trade-offs associated with the observed equifinality in more detail.

The surrogate section (2.4.3) at L350ff now reads:

We circumvented the high computational cost of running the full COSIPY model for thousands of parameter sets by developing three surrogate models replicating the required COSIPY outputs. The surrogates were trained on the 2500 simulations of the second LHS stage (Sect. 2.4.2), which were randomly divided into a training set of 2000 and a validation set of 500 samples. The surrogate models take the seven standardised COSIPY parameters as inputs. For the time-dependent albedo and snowline branches, a sine and cosine encoding of the day of year is added to capture seasonality. They simultaneously predict the mean annual mass change rate of 2000 to 2010 (B_{geod}), the SLA and ā at their 58 and 98 observation dates, respectively. Since the forcing data, except for the precipitation field scaled by pf, stays constant between samples, we do not include time-lagged forcing fields, relying on the neural network to learn their effect as a stationary background forcing implicitly. The calibration is thus conditional on the COSMO-CLM forcing, and a new surrogate must be trained for each glacier and forcing dataset. Structurally, the mass balance branch consists of a multilayer perceptron (Rumelhart et al., 1986) with hidden layers of 64, 128 and 64 neurons. The snowline and albedo branches rely on two stacked bidirectional long short-term memory layers (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997) of 64 neurons, followed by dense layers of 128 and 64 neurons. The surrogates were trained with a mean squared error (B_{geod}) and Huber loss (SLA and ā) using the Adam optimiser and early stopping guided by the validation dataset for 300 epochs and a batch size of 32. On the validation set, the emulated values closely match the COSIPY simulations for all three observables (Figs. S10 to S12). Because the validation set directly guided early stopping and learning rate adjustments, we confirmed these results with a ten-fold cross-validation (Hastie et al., 2009), which iteratively retrains and evaluates the model on ten independent data folds. Across the folds, the surrogate shows an R2 of 0.997±0.001 (B_{geod}), 0.998±0.001 (SLA) and 0.998±0.001 (ā), with root mean square errors (RMSE) of 0.05 m w.e.a^{-1}, 11.34 m and 0.007, respectively. We thus deem the surrogate an acceptable substitute for COSIPY.

2. Parameters at edges of prior distribution

The posterior distributions of the model parameters (Fig. A1) are for some parameters very close to the prior's edges. This is the case for ice and fresh snow albedo and albedo depth. For ice and fresh snow albedo, these values would be well within the literature range (Table1). Is this an indicator that the prior distribution was not well selected for these parameters? Could the model performance be improved by a prior distribution extended on these edges? It might be worth to test this, or at least address the issue more deeply in section 4.1.

We thank the reviewer for this helpful comment! We report bounds and parameter values rounded to two decimals to facilitate the comparison. While our literature research (see Supplement) suggested that the value for the albedo depth parameter should be limited at 1.0, we determined the bounds for the albedo of fresh snow (α_{fs}) and ice albedo (α_{ice}) using histograms of the Landsat-derived per pixel glacier albedo values, so that these priors are observationally informed rather than arbitrary. This procedure is now described in more detail in the manuscript (Section 2.3, L238ff). The bounds correspond to the 2nd (0.12) to 50th (0.23) percentile of all summer (JJAS) albedo values for α_{ice} and to the 50th (0.89) to 98th (0.93) percentile of the winter (DJFM) albedo retrievals for α_{fs}.These bounds were chosen because in the albedo parameterisation, the two parameters represent the bounding values between which the surface albedo evolves. Thus, fresh snow should represent the brightest observed surface albedos while ice albedo should be among the darkest. Values beyond these bounds occur in the histograms and reflect the spatial and temporal variability of the glacier surface (see Figures S7 to S9). As for the albedo depth parameter, we hypothesize that the parameter pushing against its boundary represents a compensating effect for too much available Q_{ME} or too low summer snowfall in the forcing (e.g., Voordendag et al., 2024), as discussed in Section 4.1.

Generally, we argue that posterior distributions piling up near prior bounds are an expected feature of a highly overparameterised calibration with compensating parameters and known structural and forcing errors. The calibration cannot distinguish the error source and therefore tries to adjust the parameters in the direction favoured by the log-likelihood term. In the case of α_{ice}, it is driven towards brighter values by the mass balance term, trying to offset the positive melt bias of the forcing. In contrast, α_{fs} is driven towards darker values by the albedo log-likelihood term even though this acts against the mass balance log-likelihood. This directly shows the difficulty in finding a compromise solution for all log-likelihood terms given the forcing and structural errors. Widening the bounds would likely not resolve this, as the parameters would simply continue absorbing forcing and structural errors at values our observations do not support.

To assess the impact of widening the priors, we ran a test in which we fixed all parameters except α_{fs} and the α_{ice} to their posterior mean. We then perturbed each in nine steps centred on the posterior mean and extending beyond the prior range to still plausible lower and upper bounds (0.85 to 0.93 and 0.12 to 0.30), and evaluated the joint log-likelihood as well as the individual terms for mass balance, albedo and the snowline altitudes (Fig R2.1). For α_{fs}, the joint maximum lies within the posterior range and the individual terms show that while mass balance favours higher values, albedo favours lower values. The compromise lies within the posterior bounds, but extending the allowed albedo to lower values would create a more symmetrical normal distribution (Fig. A1 panel e).

For α_{ice}, the joint log-likelihood increases with increasing α_{ice} with only marginal gains compared to the same steps in α_{fs}. This gain comes exclusively from the mass balance term, confirming the compensating mechanism. While a higher ice albedo improves the mass balance estimate by offsetting the overestimated melt energy of the forcing, this occurs at values our Landsat histograms do not support at Hintereisferner. Furthermore, this test is conditional on the remaining now fixed parameters. In a new calibration with widened bounds, the other parameters would also adjust, likely reducing the gain displayed here. We therefore acknowledge the issue, but retain the observationally constrained priors.

This led to the following changes in the manuscript:
L238ff (Section 2.3) now details the percentile-based prior derivation:
The per-pixel seasonal distributions inform the prior ranges of the albedo parameters, and the scene means provide the glacier-wide mean albedo $\bar{a}$ as a third calibration target. For the prior ranges, we grouped all pixel values into winter (DJFM) and summer (JJAS) seasons and constrained each surface type to its published literature range (see Table 1). Since the albedo parameterisation of Oerlemans and Knap (1998) evolves the albedo between the two parameters αfs and αice as upper and lower limits, α_{fs} was constrained to the 50th to 98th percentile of plausible winter values (0.887 to 0.93), while α_{ice} and α_{firn} were constrained to the 2nd to 50th percentile of the plausible summer values, ranging from 0.115 to 0.233 and 0.506 to 0.692, respectively (see Figs. S7 to S9). The outer percentiles (2nd and 98th) aim to exclude potential outliers such as debris-covered or cloud-contaminated pixels.

We also extended the discussion at L636ff:
This compromise is also visible in the posterior distributions accumulating near their prior bounds for αice, αfs and αdepth (Fig. A1). While α_{ice} is driven towards brighter values by the mass balance log-likelihood, counteracting the negative mass balance bias, α_{fs} is driven towards lower values by the albedo log-likelihood term at the expense of the mass balance agreement (see Fig. S24). The satellite-derived priors for α_{ice} and α_{fs} then act as strict boundaries, keeping the compensation within the observational limits.

Addressing a comment by Reviewer 1 at the same time, we also added to the discussion at L710ff:
Furthermore, the per-pixel albedo retrieval is less reliable under snow-covered conditions, as the narrow-to-broadband conversion tends to saturate over highly reflective surfaces (Liang, 2001; Ren et al., 2021). In light of this uncertainty and the limitations of a constant parameter, widening the prior bounds could offer beneficial leeway. However, doing so risks allowing the parameters to compensate for other errors and lose their physical basis.

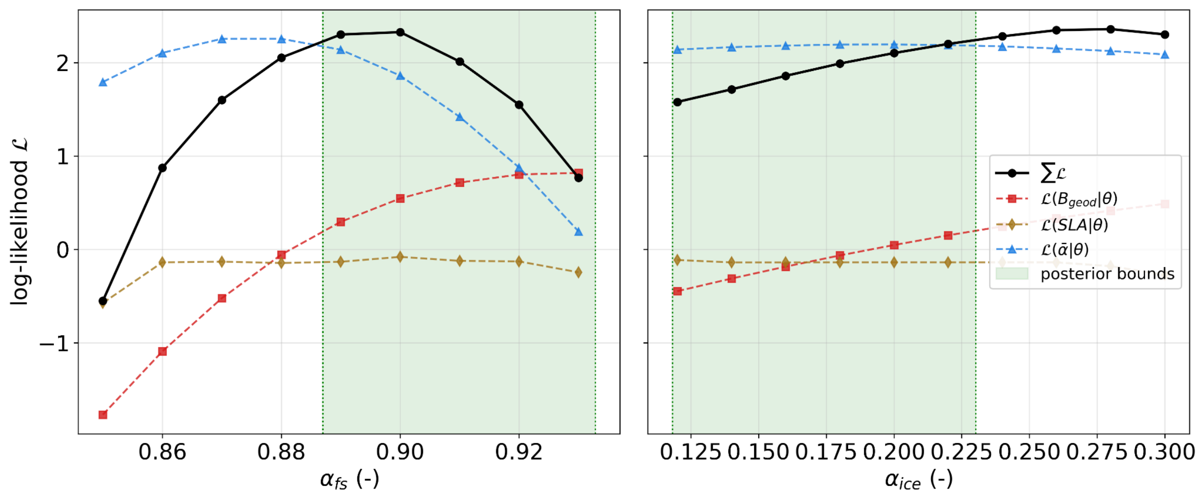

Figure R2.1:One-at-a time perturbation of the albedo of fresh snow (left) and ice albedo (right) parameters across nine steps extending beyond the prior range used for the MCMC calibration (prior bounds, green shading). All other parameters are fixed at their posterior mean. Displayed are the joint log-likelihood (black) and the individual mass balance (red), snowline (yellow) and glacier-mean albedo (blue) terms.

Specific comments

L39ff: Highlight the advances of SEB models (e.g., physical basis, enable process understanding) to make it clearer to the reader why it is important to use them

Thanks for this comment. In agreement with the request of Reviewer 1, we have adapted the sentences at L39ff to now read:
In contrast, surface energy (SEB) and mass balance (MB) models (e.g., Hock and Holmgren, 2005) resolve the individual energy fluxes at the glacier-atmosphere interface and enable the investigation of the drivers of glacier melt and mass balance. Because of this physical basis, they do not inherently suffer from the stationarity assumption (MacDougall et al., 2011), but they require a larger number of meteorological input variables. SEB model parameters may also be utilised to account for unresolved processes or limitations and biases in the forcing.

L71: remove "to"

Done.

L87: Add Hintereisferner

Done.

Fig. 1: I suggest including terrain information in the overview map in the lower right corner.

As suggested by the reviewer, we have added terrain information in the overview map.

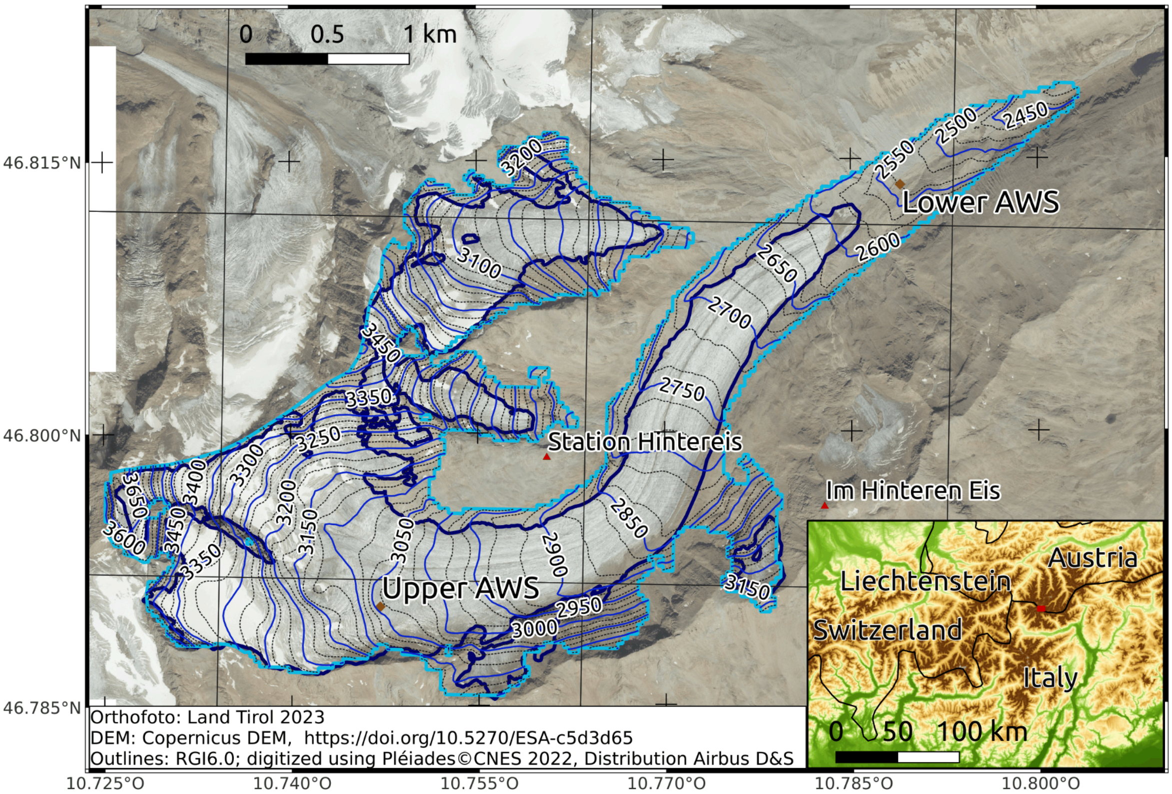

Figure R2.2: Adapted overview map with added terrain information. The caption will remain the same as in the previous version of the manuscript.

L51ff: What is the temporal resolution of the inner domain, 2.2km data?

The stored values are mostly at hourly resolution, except for surface pressure which comes at 6-hourly values. The timestep of the original simulations is unfortunately not stored anywhere in the metadata.

We have added the hourly to six-hourly resolution to the sentence at L154ff:

In this study, we used 2.2km grid spacing convection-permitting Consortium for Small-scale Modelling- Climate Limited-area Modelling Community (COSMO-CLM; Rockel et al., 2008; Baldauf et al., 2011) simulations produced as part of the World Climate Research Programme sponsored Coordinated Regional Climate Downscaling Experiment (CORDEX) Flagship Pilot Study on convection over Europe and the Mediterranean (Coppola et al., 2020) and conducted by the ETH Zurich (Leutwyler et al., 2017; Ban et al., 2021) at hourly to six-hourly resolution.

L161: What is the Earth System Grid Federation, and which data did you get from there?

Above, you state that the forcing data is from COSMO-CLM?

We apologise for the confusion. COSMO-CLM is indeed the climate model used to generate the forcing, while the Earth System Grid Federation is the public data repository

infrastructure where these simulations are stored. To make this clearer, we have rephrased the sentence at now L166ff as follows:
Except for snowfall (daily), air pressure (6 hourly) and incoming long wave radiation (not available), we obtained all stored COSMO-CLM fields at hourly resolution directly from the Earth System Grid Federation repository (ESGF; https://esgf-ui.ceda.ac.uk/search, last accessed: 04.11.2025).

L171ff: What is the resolution of your COSIPY simulations after distribution of the glacier surface?

Thanks for this comment. We forgot to mention here that COSIPY runs at hourly temporal and in 20 m elevation bands in this setup. We have moved the sentence mentioning the 20m elevation bands from L193 here to now L180f. The new sentences read:
Since most forcing fields are extrapolated using lapse rates, we applied a 20m elevation band setup in COSIPY. We applied hourly lapse rates across these 20m elevation bands for relative humidity, T_2, total precipitation and snowfall to distribute the forcing data over the glacier surface, and extrapolated surface pressure with the barometric formula.

L212: What is the rough frequency of SLA information? Is it daily, weekly, monthly?

Thanks for this question. The sampling is quite irregular, which is why we did not report it. The dataset comprises 58 observations over ten years, which roughly corresponds to a sample every second month. But owing to data availability and quality filtering, the between-observations interval ranges from 8 to 304 days with a median of roughly one month. This is almost twice as frequent for the albedo data (98 points). In light of the difficulty in reporting the frequency, we have added a sentence at L224f but choose to not expand on this further:
The SLA and $\bar{a}$ dataset comprises 58 and 98 observations over 2000 to 2010 at irregular intervals owing to cloud cover and quality filtering, respectively.

Fig. 2: This figure is not fully clear to me. It is difficult to follow the lines and understand which element belongs to which part. Maybe you could try to use different background colors, or outline colors? I would also get rid of the color gradient in the background.

We got rid of the background gradient, used thicker lines, increased font size and re-arranged the figure horizontally. The figure caption was adjusted accordingly.

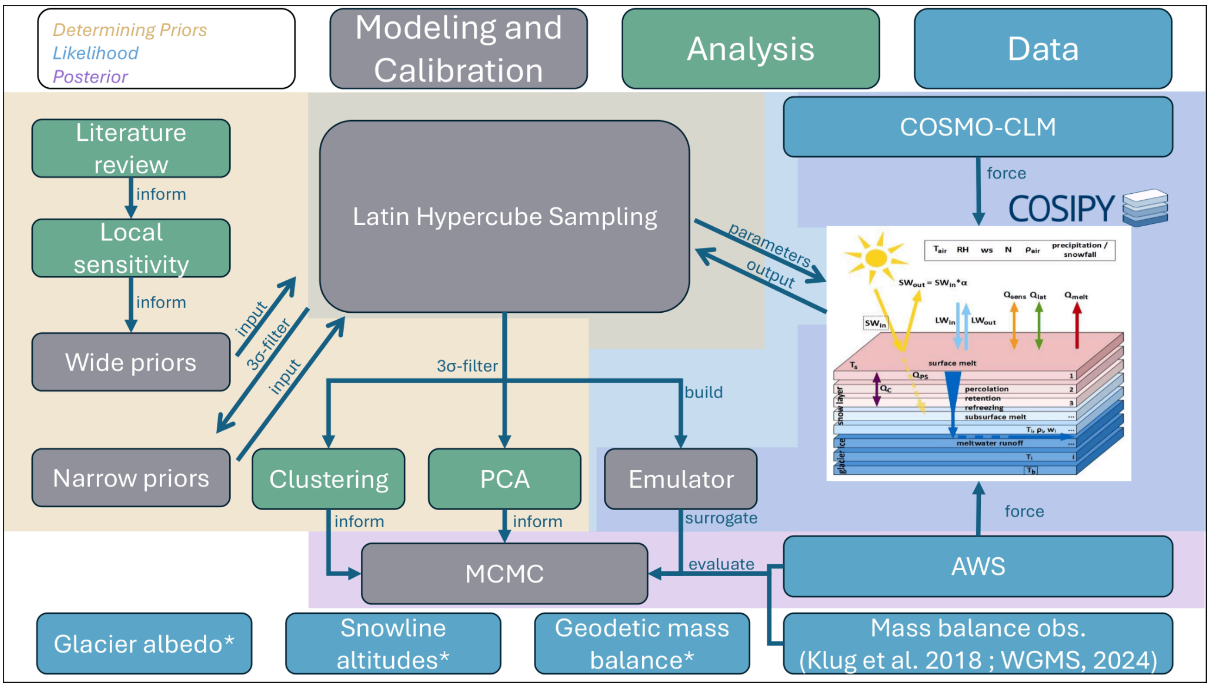

Fig R2.3: The adapted workflow schematic. The new in paper caption reads:
Overview of the workflow applied within this study. The beige, blue and lavender background shadings broadly outline the elements used for the priors, the likelihood calculation and the posteriors. The observational datasets, highlighted with a star, of the glacier-mean albedo, snowlines and geodetic mass balances were used throughout the whole workflow as described in the text. To maintain overview, we have not drawn an arrow from the automatic weather station (AWS) to the Latin Hypercube Sampling nor from the COSIPY model to the MCMC evaluation. We have roughly grouped the processes and tools into three categories, as outlined at the top of the figure, but these categories are not always clearly distinguishable. The COSIPY figure was adapted from Huintjes (2014).

L261ff: I suggest rephrasing for readability, e.g.: "This error was set to zero during the accumulations season (...), since the model matches well the fully snow-covered winter conditions, and set to XY for the snowline and albedo in the ablation season (...)."

Good idea. We changed it accordingly!

L302: Refer to Table 1

Done.

L305: Refer to Table 1

Done.

L314: Add one sentence what the PCA is used for and why it is needed. See major comment 1.

Thank you for this suggestion. We added the following lines at L334ff:

The PCA extends this analysis by investigating all seven parameters simultaneously. It reduces the high-dimensional parameter space and collapses it into a set of independent components that capture the dominant patterns of variability based on the standardised parameter vectors of the COSMO-forced LHS ensemble. By examining the weight each parameter contributes to a component, we can identify which parameters vary together or oppose each other, revealing the structural trade-offs associated with the observed equifinality in more detail.

L337ff: Give more details on the surrogate model. Potentially add a complete subchapter.

See major comment 1.

L349: What is meant by “posterior predictive checks”? Please add some details here.

Thank you for this comment. We have added a short description at L389ff:
Instead of comparing the output of the 300 simulations directly with the fixed observations, posterior predictive checks simulate replicated observations by propagating the observational and systematic uncertainties into the deterministic modelled values. This provides a more realistic assessment of whether the model chain can successfully reproduce the observed data as uncertainties can be explicitly accounted for.

L351: How do the results of Klug et al. (2018) fit to the Hugonnet (2021) data that you used for calibration? If these two datasets do not fit well together, the COSIPY data won’t fit well.

We agree with this comment. While the two datasets do not follow the same processing steps, their results are still in agreement with each other. We found that the singular value from Hugonnet et al. (2021) agrees well with the mean of the annual observed mass balance values reported by the WGMS (-1.06 m w.e.a^{-1}, 2000 to 2009) and is consistent with the geodetic survey of Klug et al. (2018) (-1.31 m w.e.a^{-1}, 2002 to 2010).
We have added a statement regarding their agreement at L394ff:
Importantly, the geodetic mass balance value from Hugonnet et al. (2021) agrees well with the mean of the annual observed mass balance values reported by the WGMS (-1.06 m w.e. a^{-1}, 2000 to 2009) and is consistent with the geodetic survey of Klug et al. (2018) (-1.31 m w.e. a^{-1}, 2002 to 2010).

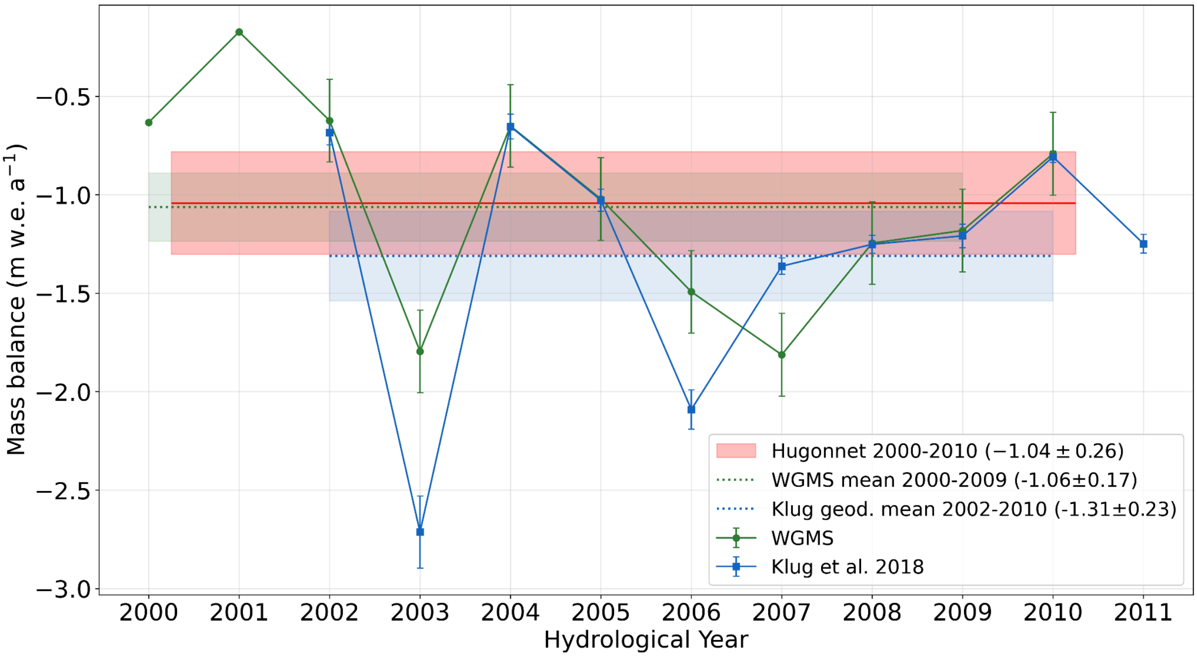

Figure R2.4: Comparison of mean annual mass balances reported over the hydrological year from the WGMS (green) and Klug et al. (2018) (blue) and the geodetic mass balance obtained from Hugonnet et al. (2021) in red. Also reported are the means of the WGMS values and the standard error of the mean, as well as those of Klug et al. (2018).

L370ff: I assume this whole paragraph refers to the AWS-assisted COSIPY simulations. If so, please mention it again at the beginning of the section to make it easier for the reader.

Done! We've added this at the start of the paragraph.
L414f now reads:
The AWS-assisted COSIPY simulations were initialised at the start of October 2002 with a snow height of zero and all other initial conditions taken from the COSMO-forced simulations.

L378: Remove "analysis" as it is part of the abbreviation

Done!

L391: Due to the limitations of the Spearman rank correlation, would there be a better method for this analysis?

We thank the reviewer for this question. Since the Spearman rank correlation can only capture monotonic bivariate relationships, we employed it as a simple and interpretable first screen for dominant relationships between parameters and also model scores. To fully resolve the often non-linear and interacting parameter effects would require dedicated and computationally intensive global sensitivity analysis or machine learning-based feature importance frameworks, which we consider beyond the scope of this first test.
We have made this first screening role clearer in the manuscript.

L330ff:
The Spearman rank correlation captures monotonic relationships, signalling whether inside the quality-filtered LHS ensemble parameters or parameter and model bias pairs increase consistently or decrease with one another. A strong correlation between a parameter and the model bias indicates that this parameter exerts direct control on the fit to the respective observations, while a strong correlation between two parameters indicates that they compensate each other. This serves as a first-order approximation of parameter interactions only, as the model system is highly non-linear (e.g., Johnson and Rupper, 2020).

L421ff:
This section first analyses the parameter identifiability, that is, the possibility of uniquely determining model parameters from the data, and their compensating effects based on the COSMO-forced LHS ensemble and the Spearman rank correlation as a first screening tool coupled with a PCA.

Fig.3: Is it really necessary to display the correlation between all parameters & biases here? I would suggest limiting the figure to parameters vs. observation biases.

We thank the reviewer for this suggestion. Retaining the internal parameter correlations is important to diagnose obvious structural equifinality, as for example the strong correlation between the precipitation factor and the albedo ageing parameter reveals a primary compensatory axis yielding equally plausible model outputs. Therefore, we prefer to show all correlations.

L402f: Do the three observational datasets fit together well? If the observations do not fit together, the model won't be able to satisfy all three constraints either. So the question here is: Does this difficulty really come from the model parameters, or from the observational datasets?

We thank the reviewer for this question. First, the transient snowline observations and the glacier-wide mean albedo are strongly correlated (r = -0.98), confirming their coherence. While we cannot directly assess this coherence for the mass balance calibration target, the Hugonnet et al. (2021) value aligns with the reported WGMS mean and is a bit more positive than the Klug et al. (2018) mean, though they are not directly comparable (see previous comment). The consistency of the transient observations and the mass balance is instead provided through our simulations. Our results revealed conflicts in the three data streams, but correcting the forcing with the measured AWS-derived biases largely closes the mass balance gap (new MB = -0.962 m w.e. a^{-1}) while not significantly degrading the fit to the albedo and snowline dataset (see Figure R2.5). We therefore attribute the calibration tension primarily to biases in the meteorological forcing and model shortcomings instead of inconsistencies among the observations.

We have adapted the manuscript in various places.
L224f: *We note that these two are strongly anti-correlated (r = -0.98), since they are both governed by the extent of snow cover.*
L643ff: *This likely stems from the forcing biases outlined below, as the albedo and snowline record are strongly anticorrelated, and the geodetic mass balance target agrees with the independent glaciological observations.*

L656ff: *Evaluated using the measured AWS data for the hydrological year 2003/04, the bias in Q_{LWin} amounts to 35.79 Wm^{-2}, that of wind speed to 1.51 ms^{-1} and specific humidity to 0.47 gkg^{-1}. When correcting for the biases in Q_{LWin} and wind speed and taking the posterior mean parameters, the simulated mass balance is closer to the geodetic target (absolute difference of 0.08 m w.e.a^{-1}) without significant degradation of the snowline and albedo fit (Fig. S25).*

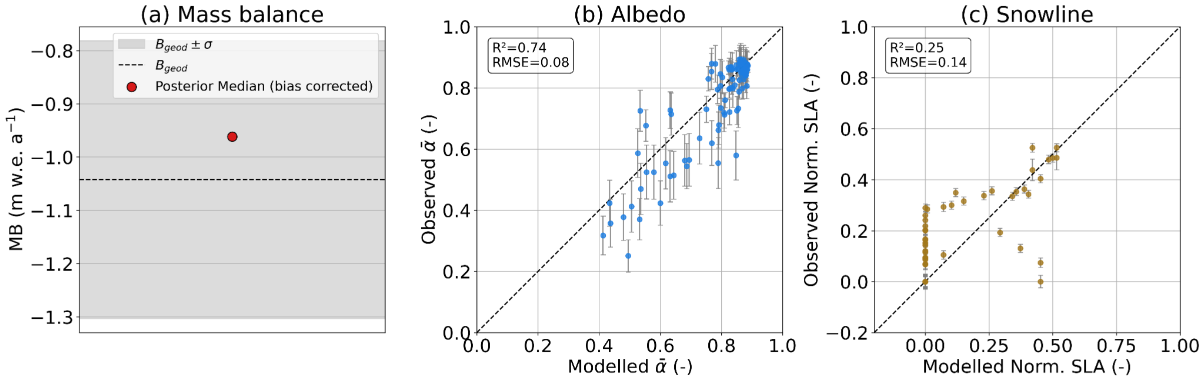

Figure R2.5: Evaluation of the mean annual mass balance (a), glacier-mean albedo (b) and normalised snowline altitudes (c) when run with the Posterior Median parameters and corrected wind speed and Q_{LWin} biases as reported by the evaluation for the hydrological year 2003/04 at the upper AWS station.

This Figure is now also implemented as Fig. S25.

Fig. 4: Panel b: Why is one of the boxes pale?

I would not put the legend as subpanel f in the middle of the plot. I would remove the panel letter and put it at the beginning, end or completely on the right or left side of the plot.

We thank the reviewer for these suggestions. We have fixed the pale box and shifted the legend panel to the end without a letter. The adjusted figure is given below.

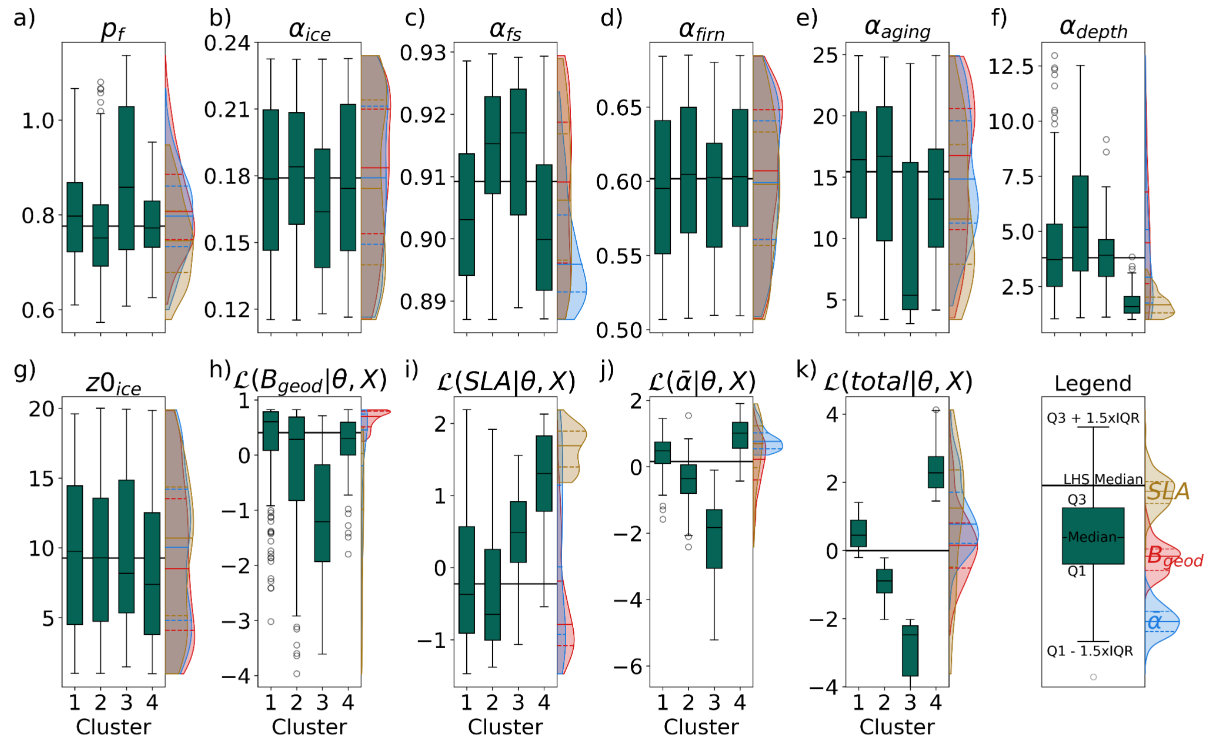

Figure R2.6: Updated Figure 4 from the manuscript. Caption will be updated according to the new order.

L415: The abbreviations for ESS, IAT and MCSE are all used only two times in the paper. I would just write them out.

Done!

Fig. 6: Panel e & f: There are dots in black and grey. What is the difference?

This seems to be a remnant of the transparency. In actuality, when dots are black, it means there are multiple dots close to each other. We considered reverting this transparency, but now it seems a nice addition to highlight point density in such a compressed plot. We have added a description of this phenomenon in the figure caption.

L455: R2 in the text is different to Fig. 6c.

Thanks for catching that! We fixed the mistake in the text.

L467ff: Do you trust the result from the glaciological mass balance here with more negative mass balance above 3325m? Could also be an artefact due to lack of observations above 3000m (L477)?

Thank you for raising this question. Indeed, the negative mass balances above 3325m are possibly exacerbated by the missing observations at such altitudes and the spatial extrapolation. When looking at Fig. 4 in Klug et al. (2018), that is comparing the WGMS-reported gradients with the geodetic measurements, the mass balance gradients do agree well with each other, although they may not directly have their most positive value at approximately 3325m above sea level. Instead, they remain roughly constant in elevation. It should be noted that the geodetic gradients incorporate ice dynamics, which are not accounted for by the glaciological method. Nevertheless, the overall value of the WGMS-reported mass balance gradient at high elevations closely follows the independent geodetic observations.

L469: Why is the simulated mass balance not getting more negative below 2550m? It should be warmer, less snow, … Where does this come from?

We thank the reviewer for this question. We propose three explanations summarised in the new response figure (Fig. R2.7).
Firstly, the number of grid cells per elevation band is very low at these elevations, so the band means are likely not robust and these bands are very susceptible to issues in the DEM quality.

Second, the tongue is heavily shaded. The mean shortwave correction factor, averaged over all hours of the ablation season (May to September), reaches its lowest values at these elevations (not shown). We confirm this by displaying the cumulative Q_{SWin} over May to September, averaged over the years 2001 to 2003 (Fig R2.7, panel d). The glacier tongue receives a lot less energy input over the average melt season, decreasing from 3160 MJ m^{-2} to 2900 MJ m^{-2}. We also verified that this is not an artefact of the one-dimensional setup: an additional distributed COSIPY simulation run at 30m grid spacing for the hydrological years 2001 to 2003 also features the near-constant mass balance at low elevations (Fig. R2.7 a). For the glacier-mean these values do not matter as much.

Thirdly, COSIPY does not account for processes that could counteract excessive shading, providing additional energy specifically at the glacier tongue such as debris cover. We quantified two of these as rough estimates. For the longwave emission of the surrounding terrain, which is not part of the present model design, we follow Prinz et al. (2016), who calculate the total incoming longwave radiation as:
$LWI_{cor} = s_f \cdot LWI + (1 - s_f)\ \varepsilon_R \cdot \sigma \cdot T_R^4$,
where LWI is the incoming longwave radiation from the forcing (sky), s_f the sky view factor, ε_R is the terrain emissivity set to 0.98, and T_R is the terrain temperature, defined as the absolute air temperature plus 0.01 K W^{-1} m^2 times the global radiation (Sicart et al. 2011). For our estimate, we derive T_R from the melt season (May to September) mean air temperature and mean global radiation of the COSMO forcing at the respective elevation band, and take LWI as the melt season mean incoming longwave radiation from the model output. Since COSIPY applies LWI over the entire area, it effectively treats the fraction that is occupied by the surrounding terrain as if it were the sky. The energy missing from the model is therefore the difference between the terrain fraction emission and the sky radiation (the forcing), which the model already defines for each elevation band. This difference is largest at the tongue, where the sky view factor is lowest and amounts to up to 50 MJ m^{-2} averaged over the mean melt season (Fig. R2.7 b).

Similarly, we approximated the effect of bare ice darkening towards the tongue, which the spatially constant ice albedo cannot represent, by prescribing a linear decrease of the ice albedo from the calibrated value of 0.229 at the equilibrium line altitude to the observed

minimum of 0.12 at the lowest glacier elevation. The additional absorbed shortwave radiation is then the albedo difference multiplied by the mean cumulative May to September insolation averaged over the years 2001 to 2003 for each band. This contributes up to 318 MJ m^{-2} over the mean melt season (Fig. R2.7 c), although such an effect is not very realistic.
To express these as mass equivalents, we assume that the additional energy acts on every melt season day for 12 hours per day and that the glacier is already at the melting point, so the entire energy is consumed by melting. This assumes that the surface cools below the melting point at night, but it also serves to partially compensate for the missing consumption for surface warming. Under these rather crude assumptions, both effects add a combined additional melt energy of up to 1.10 m w.e.a^{-1} at the tongue.

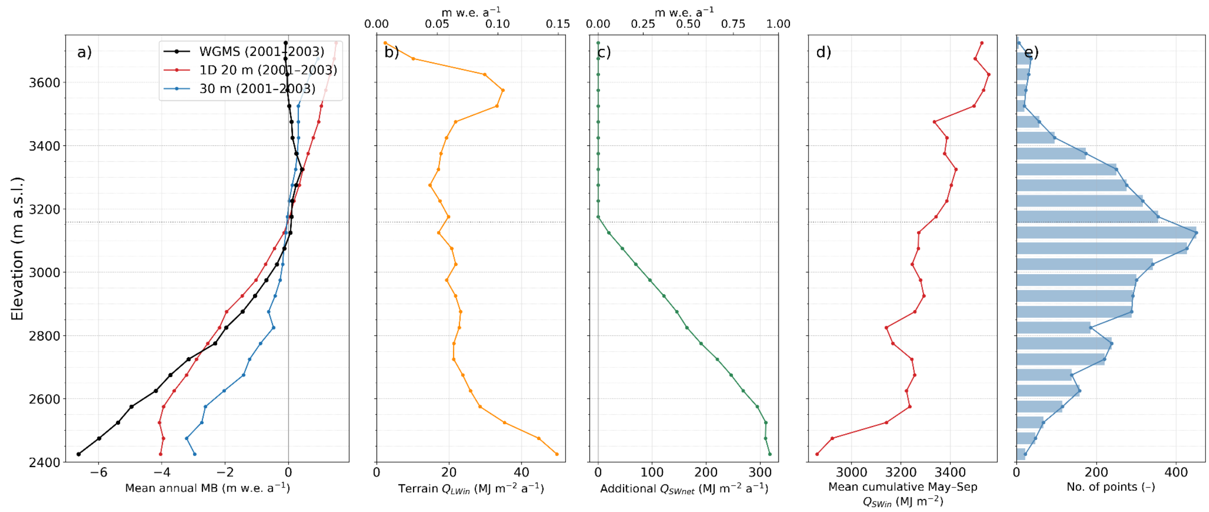

Figure R2.7: Elevation-band diagnostics showing the mass balance gradients (a) of the WGMS measurements (black), the elevation band COSIPY setup (red) and a 30m run forced with the posterior mean parameters (blue). Panel b) displays the potential terrain emitted longwave radiation calculated after Prinz et al. (2016) and employed for the elevation band setup. For a detailed description, we refer to the text. Panel c) displays the added net shortwave radiation when decreasing the ice albedo from the mean ELA of 2001 to 2003 towards its observed minimum at the glacier tongue. Panel d) displays the average cumulative May to September incoming shortwave radiation of the elevation band setup and panel e) shows the number of grid points per elevation band.

L474: Refer to Table 2

Done!

L475ff: If data above 3000 m is missing in 3 out of 8 years, how reliable is the WGMS curve in Fig. 7b? See comment to L467ff.

Thank you for raising this gradient-related question. As previously described, the gradients resemble quite closely the observations of Klug et al. (2018). While the uncertainty at higher elevations may still be larger, the overall pattern is valid. This is likely caused by processes of snow- and mass redistribution and drifting-snow sublimation (e.g., Voordendag et al., 2024). These are processes COSIPY currently does not account for. In addition, COSIPY assumes linear gradients which explain the ever-increasing mass balance at higher elevations.

L480f: Are the correlations given somewhere in a figure or table? If not, it might be useful to add correlation and MAE to Table 2.

We added the correlation and a mean value to the Table.

L483: Mention at the beginning of the section that you are comparing the COSIPY results from the posterior ensemble with the AWS-assisted ensemble, and that the latter is limited to the year 2003/04.

Thanks for this helpful suggestion. We have added a line at the start of the subsection. L528f now reads:
In this section we analyse the SEB conditions of the Posterior ensemble for the hydrological year 2003/04 and evaluate it against the AWS-assisted ensemble at the upper AWS station.

L493: Refer to Fig. 91

Done!

L494: Why is the glacier surface in May to June warmer than the air, making QH an energy sink? I would rather expect this in winter.

Thanks for catching this. While in theory a negative sensible heat flux can occur due to, e.g., cold air advection or the nighttime refreezing of the surface layer, it is incorrect as a monthly statement for May and June here. The sensible heat flux is an energy source throughout most of the year and especially from May onwards, as evidenced by Figure 9. We have deleted this erroneous statement.
The sentence now reads: Q_H is an energy source throughout most of the year.

L512: Shouldn't this be "overestimation of QLWnet"?

Thanks again! Indeed, it should be an overestimation. We have fixed it in the manuscript.

Discussion: I think, the manuscript would benefit from a comparison of the performance of the Bayesian calibration with other calibration approaches found in literature in the discussion. This way, it would be easier to understand if/how much improvement the presented method brings over more basic calibration schemes.

We thank the reviewer for this comment. We have added a short comparison with common parameter calibration techniques at L728ff:
SEB models are traditionally calibrated against a single target such as mass balance or surface height change (e.g., Mölg et al., 2009; Gurgiser et al., 2013). Such calibrations are prone to equifinality, where different parameter combinations achieve similar performance through compensating errors, often yielding physically unrealistic parameter sets. Multi-objective Pareto optimisation (e.g., Zolles et al., 2019; Khadka et al., 2024) and data assimilation (Landmann et al., 2021; Morlighem and Goldberg, 2023) provide alternative

ways to combine multiple observations and reduce this ambiguity. However, traditional and Pareto approaches typically treat observations deterministically, ignoring measurement errors and cannot probabilistically quantify parametric uncertainty. Sequential data assimilation, in turn, is primarily designed to adjust real-time model states rather than to infer parameter distributions. Our Bayesian calibration framework instead integrates glacier-wide albedo, snowline altitudes and geodetic mass balance into a single likelihood function that explicitly accounts for observational and systematic errors. Rather than eliminating equifinality, the resulting posterior distributions map and constrain it, directly propagating measurement errors into quantified parameter uncertainties and dependencies.

L536ff: It might be useful to repeat your posterior values for the comparison. At least, you should refer to Fig. A1 + the respective panel in all subsections so that the reader knows where to look up your values.

Thanks for this suggestion. We added a reference to the Fig. A1 and stated the values explicitly.

L560f: "This holds for many of our model parameters…" This is an important sentence, but more of a general discussion. Consider moving this sentence to the end of the section or only keep these considerations in section 4.3, where you also mention this issue (L636ff).

Thanks for this suggestion! We have moved this statement and it is now discussed in Section 4.3 (now L703ff).

L572: Replace the word "worsened", e.g. with "exacerbated" or "increased"

Done!

L650f: Where do we see this?

We thank the reviewer for this suggestion.
This refers to the sentence: "Indeed, inspecting the ensemble spread induced by the parameter choice versus the bias between observed and simulated net radiation emphasises the smaller role of parametric uncertainty" and was also picked up by reviewer 1.

This is only evident when referring to Fig. 9, where the narrow width of the shaded credible interval (representing parametric uncertainty) is much smaller than the difference between the simulated and the observed lines. We added this reference to the sentence. Furthermore, based on the comments of Reviewer 1, we added a separate first-order comparison, where we evaluate the 95% CI spread induced by the parametric uncertainty for the hydrological years 2002 to 2009 (see Table 2) with the shift in the mass balance invoked by running COSIPY with the posterior mean parameters and correcting for the forcing biases identified at the upper AWS station. We note that we apply these forcing corrections as constant over time and extrapolate them beyond the evaluation year 2003/04. See also our response to Reviewer 1.

We have added a section detailing this experiment at L722ff:
We support this claim by comparing the parametric ensemble spread introduced by the 95% CI (see Table 2) with the difference in simulated mean annual mass balances when running COSIPY with posterior mean parameter values with and without bias corrections to Q_{LWin}, wind speed and air temperature determined from the evaluation against the upper AWS station (Fig. A2). The 95% CI uncertainty spread averages 0.314 m w.e. a^{-1} over the hydrological years 2002 to 2009, whereas the shift induced by the forcing correction results in a difference almost twice as large at 0.535 m w.e. a^{-1}. Albeit only a first-order approximation, it confirms our claim that forcing uncertainty exceeds that of the parameter choices.

L690: Replace “too low” with “underestimated”

Done!

References:

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

Klug, C., Bollmann, E., Galos, S. P., Nicholson, L., Prinz, R., Rieg, L., Sailer, R., Stötter, J., and Kaser, G.: Geodetic reanalysis of annual glaciological mass balances (2001–2011) of Hintereisferner, Austria, The Cryosphere, 12, 833–849, https://doi.org/10.5194/tc-12-833 2018, 2018.

Prinz, R., Nicholson, L. I., Mölg, T., Gurgiser, W., and Kaser, G.: Climatic controls and climate proxy potential of Lewis Glacier, Mt. Kenya, The Cryosphere, 10, 133–148, https://doi.org/10.5194/tc-10-133-2016, 2016.

Sicart, J. E., Hock, R., Ribstein, P., Litt, M., and Ramirez, E.: Analysis of seasonal variations in mass balance and meltwater discharge of the tropical Zongo Glacier by application of a distributed energy balance model, J. Geophys. Res., 116, D13105, doi:10.1029/2010JD015105, 2011.

Voordendag, A., Goger, B., Prinz, R., Sauter, T., Mölg, T., Saigger, M., and Kaser, G.: A novel framework to investigate wind-driven snow redistribution over an Alpine glacier: combination of high-resolution terrestrial laser scans and large-eddy simulations, The Cryosphere, 18, 849–868, https://doi.org/10.5194/tc-18-849-2024, 2024

WGMS: Fluctuations of Glaciers Database, World Glacier Monitoring Service, https://doi.org/10.5904/WGMS-FOG-2024-01, 2024.