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 #1

Summary

In this study, the authors present the first Bayesian calibration framework for surface energy balance (SEB) modeling for unmonitored glaciers. The framework uses COSIPY calibrated against satellite-derived geodetic mass balance, transient snowline altitudes, and mean remote sensing glacier albedo. The authors employ a Latin Hypercube Sampling ensemble together with Markov Chain Monte Carlo (MCMC) simulations, supported by a novel surrogate model designed to emulate COSIPY outputs. Their analysis indicates that parameter equifinality primarily arises between the precipitation factor and albedo decay-related parameters. Model evaluation shows that the calibrated ensemble reproduces glacier-mean albedo and interannual mass-balance variability well, but shows a negative bias in mean annual mass balance and a delayed snowline rise.

Overall, this manuscript represents a very valuable contribution to the glacier modeling community. Bayesian calibration approaches for SEB modeling are long overdue, and I am pleased to see that the authors address this important and complex problem. The study represents an important first step toward more robust calibration of SEB models with uncertainty quantification and is highly relevant for advancing regional and global physics-based glacier modeling. The manuscript is very well written, the figures are clear and well presented, and the analysis is thorough. The authors also apply a wide range of carefully considered methods, analyses, and modeling tools. While Bayesian calibration itself is not my area of expertise, I am able to comment on the physics-based glacier modeling and machine learning components of the manuscript. There are some shortcomings that should be addressed before publication, but once these issues are resolved, I would support publication. Detailed comments and suggestions are provided below.

Thank you very much for your thorough review and your constructive feedback! Please find below our responses to each of your comments.

Major comments

1) Geodetic mass balance calibration data

The main concern I have relates to how the geodetic mass balance data is used for calibration. In L 204–205, the authors state that the model is calibrated against the “annual average specific mass change rate” for the period 2000–2010 derived from the dataset of Hugonnet et al. (2021). The reported uncertainty is −1.0425 ± 0.261 m w.e. a^{-1} (L 204).

While the dataset of Hugonnet et al. (2021) indeed provides estimates for shorter subperiods than the full 2000–2019 interval, the signal-to-noise ratio decreases substantially as the aggregation period shortens. In practice, this means that for most glaciers, periods shorter than ~10 years are basically just noise. Most current global or regional glacier modeling studies use the Hugonnet dataset only as a single 2000–2019 average value (i.e., one data point per glacier), precisely because shorter time intervals tend to have low signal-to-noise ratios.

Are the authors using the 1-year subperiods for calibration? If so, I think it should be carefully checked whether these values can actually be used, since their uncertainties might be too large to provide meaningful information. One way to verify this would be to compare the geodetic mass balance rates with the uncertainties reported by Hugonnet et al. (2021) (which are also known to be potentially optimistic) for different aggregation periods (e.g., 1-year, 2-year, 5-year, and 10-year) and calculate the corresponding signal-to-noise ratios. For example, Table 2 suggests that the observed mass balance values (though from WGMS and Klug et al., 2018 and not from Hugonnet et al.) exceed the uncertainty range (-1.0425 ± 0.261 m w.e. a^{-1}) in four (?) out of eight years. What implications would the use of longer aggregation periods (e.g., 5-year or 10-year averages from Hugonnet et al., 2021) have for the Bayesian calibration framework? I am aware that the primary focus of the manuscript is on the framework for Bayesian calibration. However, it is still very important that the calibration data are used with a clear understanding of their limitations and uncertainties.

We apologise for the ambiguous wording. We do not use the 1 or 5 year estimates from Hugonnet et al., (2021), which are very uncertain, especially at the glacier scale.

We calibrate against the single specific mass change rate averaged over 2000 to 2010 derived from the volume change estimates by Hugonnet et al., 2021. We agree that the signal-to-noise ratio for shorter periods is unsuitable for calibration. Nevertheless, we found the singular value 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 values of Klug et al., 2018 (-1.31 m w.e.a^{-1}, 2002 to 2010). Both of these averages were derived from hydrological years.

We have rephrased L204 in the manuscript:
L216: *We obtain the decadal-mean specific mass change rate and its uncertainty of [...].*
and added a statement:
L394f: *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).*

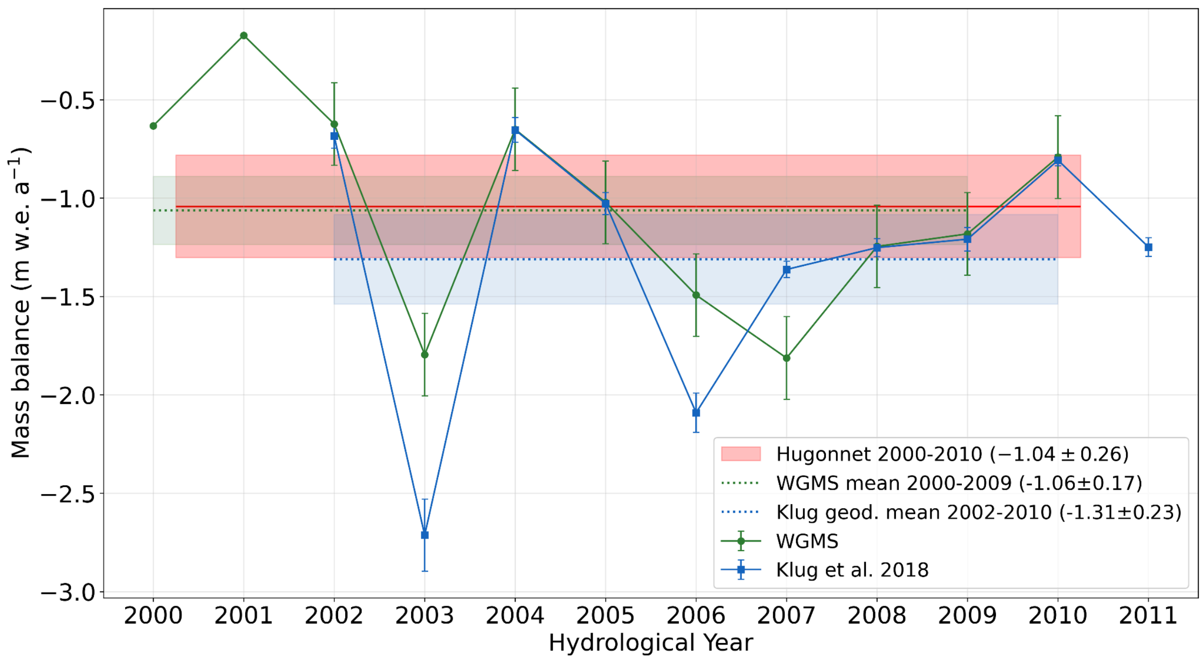

Figure R1.1: Comparison of mean annual mass balances reported over the hydrological year from the WGMS (2024) in green and Klug et al. (2018) in blue and the geodetic mass balance obtained from Hugonnet et al. (2021) in red. Also reported are the means of the WGMS values and the standard error of the mean, as well as those of Klug et al. (2018).

2) Climatic input data

The authors use COSMO as the climate forcing, produced as part of the CORDEX initiative with a grid spacing of 2.2 km. They state that these simulations are based on ERA-Interim (~80 km, L 156), which has been superseded by ERA5 (~31 km) for nearly a decade. ERA5 includes several updates compared to ERA-Interim that are relevant for glacier studies. Has the COSMO dataset been specifically evaluated over glacierized terrain? I am asking because regional climate model performance can be quite sensitive to the choice of physical parameterizations, and models optimized for other land surface types do not necessarily perform well in cryospheric environments.

I am also wondering why this dataset was chosen instead of, for example, ERA5-Land (~9 km) or ERA5 itself. Both have been shown to provide reliable forcing for SEB modeling studies, and higher spatial resolution does not necessarily translate into better forcing data. For example, Draeger et al. (2025) found that dynamical downscaling did not improve SEB model forcing for a sample of glaciers in Canada compared to ERA5, and that ERA5 even performed better than the higher-resolution ERA5-Land in their case. I also assume that the study period is limited to 1999–2010 (L155) because of the temporal availability of the COSMO dataset.

I am not suggesting that the authors redo the analysis with a different climate forcing dataset, but I do think it would be important to more clearly justify the choice of COSMO in this context, particularly since the manuscript puts a strong emphasis on uncertainties related to the forcing data in the discussion and conclusions.

We thank the reviewers for this comment. To our knowledge, the COSMO-CLM simulations have not yet been specifically evaluated over glaciers in the Alps, which is why we evaluated them against our on-glacier weather stations (Fig. A2), revealing the biases discussed in the manuscript. This evaluation showed a satisfactory performance of COSMO-CLM, and we deemed it a reasonable forcing dataset for this study. A systematic comparison of the impacts of different forcing datasets, similar to Draeger et al. (2024), would be an interesting addition, but the results could likely not be generalised across regions, and our target was a more transferable assessment of what is currently possible.

In summary, we forced COSIPY with the 2.2km convection-resolving COSMO-CLM simulations with three considerations in mind. Firstly, convection-permitting models have shown improved performance, especially in precipitation-related metrics compared to ERA5-Land or ERA5 (e.g., Ban et al., 2014; Lucas-Picher et al., 2021), which is particularly important for simulating the correct evolution of the spatio-temporal snow cover. Secondly, our goal was to assess the current state-of-the-art in simulating all terms of the energy balance. This is especially important, considering the compensating effects found in Draeger et al. (2024). In the absence of a dedicated evaluation over the Alpine glaciers, we extrapolated the findings from various convection-permitting model evaluations (e.g., Ban et al., 2021; Ou et al., 2023; Collier et al., 2024) to the Hintereisferner site and chose COSMO-CLM as the most suitable available dataset. Our evaluation against the in-situ observations nonetheless revealed significant biases in some of the forcing fields, which are likely present in other products as well. Overall, we deemed it an adequate forcing. Lastly, for future applications to unmonitored regions and glacier evolution studies, reanalyses such as ERA5 simply do not offer the possibility of running the model under climate change projections. For these reasons, we chose COSMO-CLM.

We have adapted the manuscript to highlight the reasons for the usage of COSMO-CLM more clearly:
L161f: *Beyond the advantages of CPM over reanalyses such as ERA5 in representing various surface fields including precipitation, air temperature and wind speed (e.g., Ban et al., 2021; Lucas-Picher et al., 2021; Ma et al., 2023), they can be applied to future climate scenarios and glacier evolution studies.*

3) Katabatic winds

In L170 the authors mention that a logarithmic wind profile is used. However, logarithmic profiles are known to poorly represent wind speeds under katabatic conditions and can lead to substantial underestimation of near-surface wind speeds over glaciers (L 598) For example, Shannon et al. (2019) addressed this issue by applying a calibration-dependent multiplicative factor (ranging from 1 to 4) to the 10 m wind speed from reanalysis data.

I wonder whether wind speeds during katabatic conditions might be systematically underestimated in the present setup, which would lead to underestimated turbulent fluxes. However, in L515 the authors mention an overestimation of turbulent fluxes in summer. Could the authors provide more detail on how this can be explained?

We thank the reviewer for raising this point. Contrary to the reviewer's concern, we find that COSMO-CLM systematically overestimates wind speed at both weather stations across all seasons (Fig. R1.2 and R1.3a). The overestimated turbulent fluxes are therefore the combined result of positive biases in the simulated humidity, air temperature and wind speed rather than a consequence of underestimated katabatic winds.

This wind speed bias is present year-round, which indicates that it is likely a consequence of topographic smoothing at the 2.2km grid spacing and the correspondingly reduced surface roughness. At Hintereisferner, the down-glacier katabatic flow aligns with the prevailing south-westerly synoptic direction (Goger et al., 2022; Nicholson et al., 2025) and we find no underestimation even in this katabatic-favouring wind section, although the biases are smaller (Fig R1.3b). The smaller biases in the westerly to south-westerly sectors suggest that the surface wind is then closer to the synoptically-driven flow represented by COSMO-CLM.

We appreciate the reviewer pointing out the wind-speed factor approach. Indeed, for a transferable application to other glaciers, a wind-speed correction factor as in Shannon et al. (2019) would be a useful addition, which we note for further applications. In the present study, given the positive rather than negative wind bias at our site, we proceeded with the standard logarithmic profile as in previous COSIPY applications (e.g., Arndt et al., 2021).

We changed the manuscript to mention this overestimation more clearly:
L603ff.: *The resulting biases in the turbulent fluxes are reinforced by a year-round positive wind-speed bias, presumably due to the smoothed topography in COSMO-CLM. Furthermore, these biases are compounded by an inaccurate representation of the roughness of the glacier surface and the choices of specific parameters within the land surface scheme, or, in the case of* Q_{LWin} *, by an overestimated simulated cloud cover.*

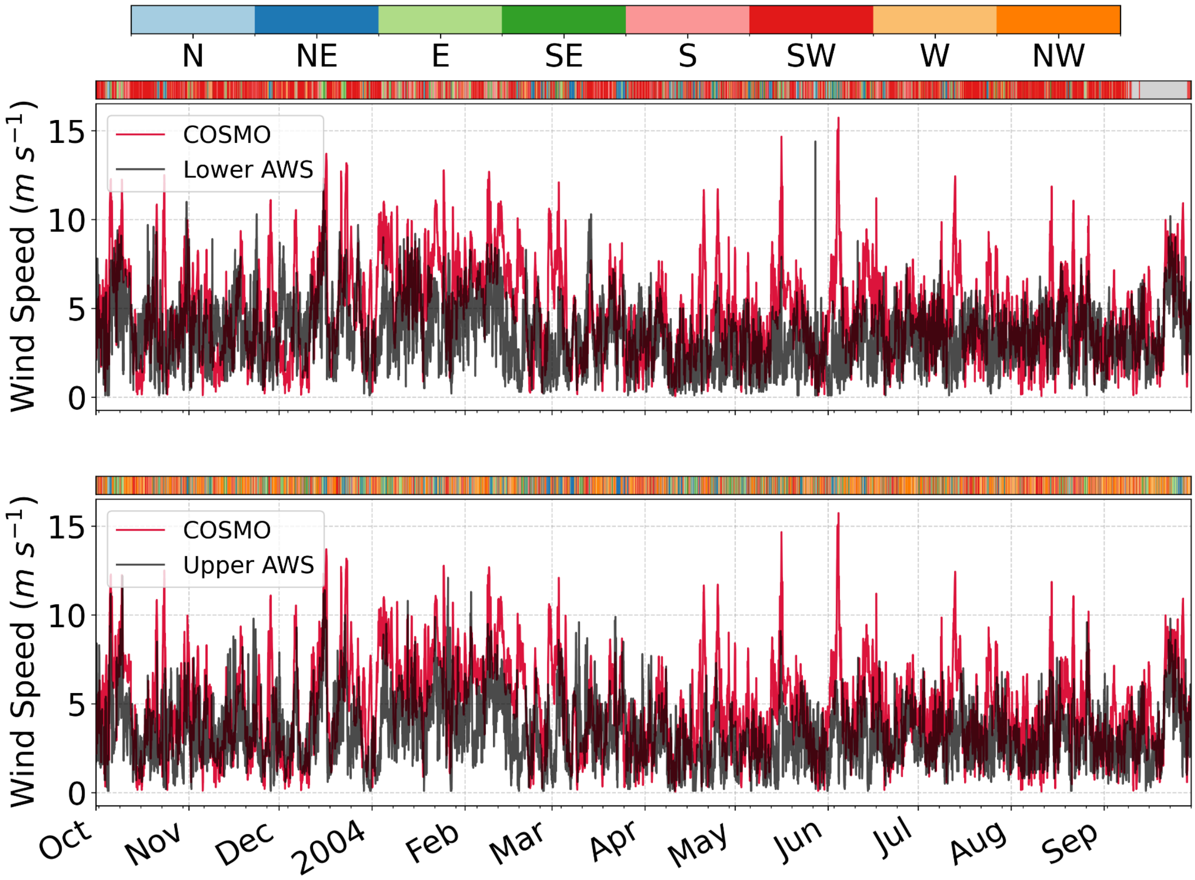

Figure R1.2: Comparison of two-meter wind speed values of COSMO-CLM (red) and the lower AWS (top) or upper AWS (bottom) for the hydrological year 2003/04. The coloured bar on top indicates the wind direction measured by the respective AWS station.

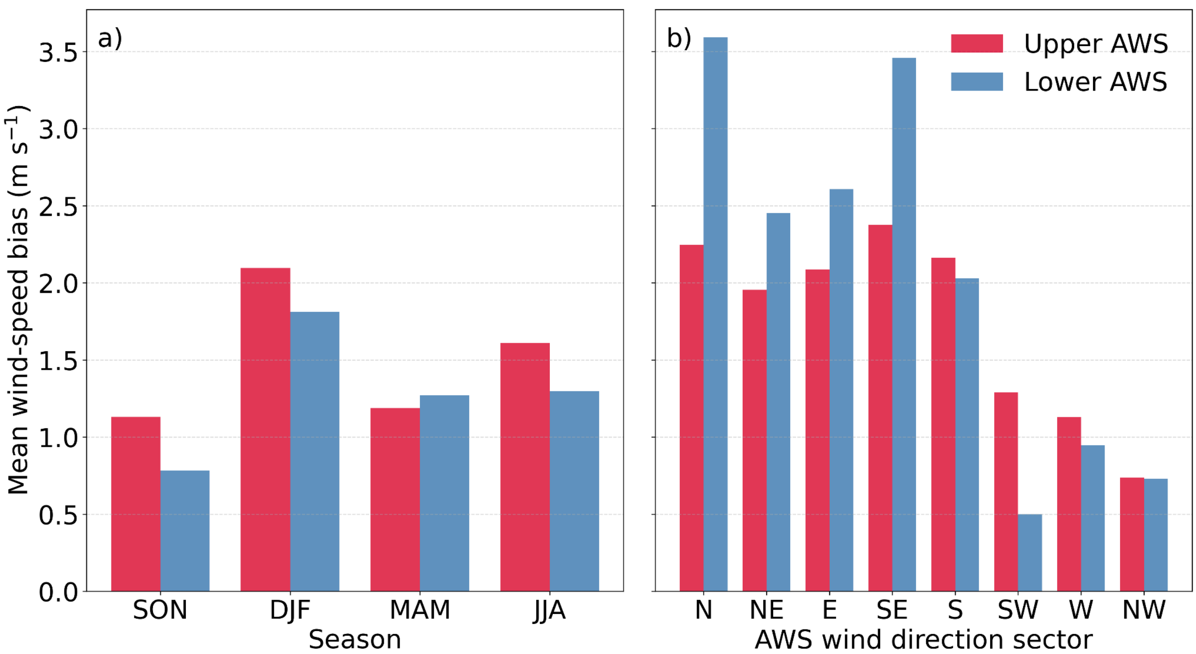

Figure R1.3: Evaluation of the wind speed bias (COSMO-CLM minus AWS station) across seasons (a) and across the different wind sectors (b) for the lower AWS (blue) and upper AWS (red).

4) Machine learning surrogate model

Using a machine learning surrogate model to avoid the computationally expensive COSIPY simulations is an interesting and valuable contribution. However, since this component is a key part of the modeling chain, I think the manuscript would benefit from providing more details.

For example, I am wondering how a test set was constructed (not mentioned in L339). It would also be helpful to include some technical details about the machine learning models (this could be placed in the SI), such as the number of epochs, batch size, loss function, optimizer, and the number of neurons in each layer. It would also help the reader if the input features and model outputs were stated more clearly. It is partly implied in L337, but there is already a lot of information in that section, which makes it somewhat difficult to follow. I might have missed it, but it is also unclear to me why the forcing data (except precipitation) are kept constant (L344).

We thank the reviewer for this helpful comment. We have created a separate subsection detailing the surrogate model creation, its inputs and outputs, training, and testing, and kindly refer the reviewer to the adapted text in Section 2.4.3.
To clarify the technical details, the surrogate models take the seven standardised COSIPY parameters as inputs, along with a sine and cosine encoding of the day of the year to capture seasonality. We have added the number of hidden layers and their neurons in the subsection and specified the usage of the Adam optimiser, the loss function and the number of epochs and the batch size.
Regarding the test set, the 2500 LHS samples were split into 2000 train and 500 validation samples. We call them validation samples, as they influence the surrogate generation in the early stopping and adaptive learning rate. To ensure robustness in our results and confirm the model generalises well, we additionally performed a ten-fold cross-validation, retraining the surrogate and evaluating the model on ten different held-out folds. This cross-validation confirmed the successful emulation of the simulated mass balance, snowline and albedo values yielding an R^2 of at least 0.997 across all variables and a low root mean square error.
Regarding the forcing data, it is kept constant because there was no functional need to include it in the emulator. Because the background forcing does not change from one parameter set to another except for the precipitation field which is influenced by the precipitation factor, the neural network can implicitly learn the forcing's effect as a background forcing. Trying to include time-varying meteorological variations alongside the parametric perturbations would drastically increase the training complexity and require a universal training dataset that is not feasible to generate within the scope of this study. Therefore, the calibration is conditional on this specific COSMO-CLM forcing dataset and has to be repeated for each new forcing dataset.

The new subsection for the emulator (Sect. 2.4.3, Line 350ff) 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 $\bar{\alpha}$ 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 $\bar{\alpha}$) 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 ($\bar{\alpha}$), 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.

5) Clarity and readability

The analysis is quite complex and involves several methodological steps, which at times makes it difficult for the reader to follow the technical details. For example, the PCA analysis is only briefly introduced in L314, and for me there are not enough details to clearly understand how it was implemented. It is also not entirely clear to me how the effect of the observed data on individual parameter values is isolated from the K-means clusters (L319).

I recognize that the manuscript is already quite long and that some additional details are provided in the SI. Nevertheless, I would encourage the authors to include a few brief, intuitive explanations of the more complex parts of the analysis in the main text. This would make the methodology much more accessible, particularly for readers in the glaciology community who may not be familiar with all the statistical techniques used.

We thank the reviewer for this comment. We agree that adding brief intuitive explanations make the text more accessible and have added these details for clarity where suggested by both reviewers. The Spearman correlation, the PCA and the K-means clustering are now each introduced more clearly. Regarding the K-means clustering, it is essentially a performance clustering and can be viewed as categorising simulations via their agreement with the observations. We performed the clustering separately per log-likelihood score which allows us to compare the distribution of the best performing cluster with the overall ensemble. The section now reads:
L328ff: *To identify which parameter combinations can produce similar model results (i.e., equifinality), we combined a Spearman rank correlation with a Principal Component Analysis (PCA) and a clustering analysis of the COSMO-forced LHS ensemble. 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).
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.
Finally, to understand which datasets constrain specific parameters, we applied K-Means clustering to the COSMO-forced LHS ensemble. We performed the clustering separately for the joint log-likelihood and for each individual log-likelihood score, in each case grouping the simulations into four performance clusters. The clustering on the joint score identifies compromise solutions across all datasets, whereas the clustering on an individual score groups the simulations by their performance for that dataset alone. Comparing the parameter distributions of the best-performing cluster of each individual score against the full ensemble isolates the constraint provided by each dataset, and a narrowed distribution shows strong observational constraints. The parameter distributions of the best-performing cluster were transformed into truncated normal distributions, which follow a normal probability density between a lower and an upper bound, and serve as the priors for the MCMC while preventing samples beyond the plausible range. An overview of the main methodology is provided in Figure 2.

We added further details for clarity where suggested by both reviewers. As acknowledged by both reviewers, the paper is already quite long and we refrain from adding even more methodological details that are not directly relevant for the paper's main novelties, the Bayesian calibration and the emulator. For the interested reader, our code is freely available and can be referred to where appropriate.

Minor comments

L39: To set the stage, I would briefly introduce SEB models here. It would help to mention that they are more physically based than simpler approaches but require substantially more input variables. It may also be useful to clarify the role of calibration parameters in SEB models, which differs from TI models. Unlike TI models, where the melt factor is an empirical parameter inherent to the model itself, many calibration parameters in SEB models are used to account for limitations in the forcing data (e.g., downscaling or unresolved processes).

Done. The new addition at L39ff reads:
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. [...]

L75: It would be helpful to provide a bit more detail on the Bayesian calibration frameworks introduced by Rounce et al. (2020) and Sjursen et al. (2023, 2025), as these are the only existing applications of Bayesian calibration in glacier modeling so far.

We thank the reviewer for this comment. To maintain the narrative and to keep the introduction concise, we prefer not to expand on the details of those studies here. Instead, the detailed Bayesian formulations are adopted and explained in Section 2.4.1.
To better phrase the context for the reader immediately, we have added a clarification at L78f:
Building on the pioneering Bayesian parameter calibration frameworks for TI models established by Rounce et al. (2020) and Sjursen et al. (2023, 2025) and capitalising on a recent suite of CPM simulations at 2.2 km grid spacing (Leutwyler et al., 2017; Ban et al., 2020) [...].

L71: Typo

Thank you!

L83: I would clarify here how (and with which data) the simulations are evaluated. This is mentioned at the end of the paragraph below, but it might fit better directly after the three main points.

Thank you. We have adjusted the last point at L85f to now read:
Evaluate the resultant SEB simulations using independent on-glacier measurements and glaciological mass balance records to quantify the accuracy and identify limitations of this "remote-only" approach

Figure 1: I believe the Copernicus WorldDEM-30 (https://doi.org/10.5270/ESA-c5d3d65) should also be cited in the references.

We have added the entry.

L156: I am a bit confused about the grid spacings mentioned. ERA-Interim has a resolution of ~79 km, but in L156 it is referred to as “12 km”. Could the authors please clarify this?

We apologise for the confusion. Indeed, the 12km refers to the grid spacing of the intermediate simulation. We have revised the sentence at L158f as such:
The simulations cover January 1999 to 2010 and are the result of a two-step, one-way nesting approach. ERA-Interim (Dee et al., 2011) provided six-hourly boundary and initial conditions for an intermediate 12 km COSMO-CLM simulation that in turn served as the input for the 2.2 km simulations.

L169: Could you provide a few brief details on the analysis of Jennings et al. (2018) to clarify why it is used as a justification here?

Of course. We have expanded the text to add more detail. L173ff now reads:
*We keep s_p at its default value of 1.0°C and set T_{rs} to 1.55°C, which is similar to the value employed in other **glacier** studies (e.g., Huss and Hock, 2015). This threshold is based on the analysis of Jennings et al. (2018), who constructed a 50% rain and 50% snow temperature threshold model and applied it to the gridded Modern-Era Retrospective analysis for Research and Applications (MERRA)-2 reanalysis data. The value corresponds to the threshold extracted from the closest MERRA-2 grid cell covering Hintereisferner.*

L175: Could you please provide details on the “large spacing between vertically stored levels in the COSMO simulation at standard pressure levels”?

Yes, available online are the pressure levels 200, 500, 700, 850, 925 and 1000 hPa. Assuming standard atmosphere, this corresponds to only three elevation levels above the glacier surface. We have adapted this in the text at L182ff:
Due to the coarse vertical resolution of the stored COSMO-CLM output at standard pressure levels (1000, 925, 850, 700, 500 and 200 hPa), we instead derived local lapse rates from the near-surface fields over a horizontal three-by-three stencil located over the glacier centroid.

L176–179: I am not sure I fully follow this part. Does this mean that, in the end, all 3x3 grid cells were used to calculate lapse rates, regardless of land surface type?

We apologise for the lack of clarity. All nine grid cells used for the lapse rate calculation are of the same “glacierised” surface type. This was one argument advocating for this size in comparison to, e.g., the 5x5 grid. We made this clearer in the revised manuscript at L186f:
*[...] The resultant lapse rates are thus the product of a nine-sample linear regression **based strictly on grid cells classified as glacierised**. [...]*

L195: Please provide a short description of HORZYZON. Also, please cite the original publication, as required when using the code (Steger et al., 2022).

Of course. We added the citation and changed the text accordingly. L180 now reads:
Since most forcing fields are extrapolated using lapse rates, we applied a 20 m elevation band setup in COSIPY.

L203ff:
To account for terrain- and self-shading, the high-resolution shortwave correction factors generated by the raytracing tool HORAYZON (Steger et al., 2022; Steger, 2022) at the native one-arcsecond resolution NASADEM (NASA JPL, 2020, last accessed: 19.03.2025) were averaged into their corresponding elevation bands. HORAYZON operates by calculating time-dependent cast shadows from surrounding terrain and self-shading from local slope orientation for each grid cell using a computationally efficient ray-tracing library.

L 229: I might have missed it, but it is unclear to me how you distinguish between alpha_ice, alpha_firn, and alpha_snow based on the Landsat-based pixel data. How does this vary temporally and spatially?

We thank the reviewer for this question. To clarify, we do not explicitly track or classify individual pixels into ice, firn or snow across space and time. Instead, we use a seasonal and statistical separation of the entire per-pixel albedo pool over the glacier to define prior distributions. From seasonal histograms that were filtered to literature-backed physical ranges for each surface type, we distinguish the parameters by applying percentile thresholds. This is described in more detail now in L238ff:

We use these retrievals in two ways. The per-pixel seasonal distributions inform the prior ranges of the albedo parameters, and the scene means provide the glacier-wide mean albedo $\bar{\alpha}$ 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 α_{fs} and α_{ice} 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.

L 254-258: Am I understanding correctly that "COSIPY model output" here does not only refer to mass balance, but also to other model outputs, such as the albedo simulated by the COSIPY albedo model (Eq. 5)? If so, the wording might be confusing.

We apologise for the confusion. Indeed, it refers to the simulated mass balance, albedo and snowline altitudes. We have addressed this by rephrasing L268ff:

*Following the approach from Rounce et al. (2020) and Sjursen et al. (2023, 2025), we assume that the calibration data represent the truth, and that the forward model (F; here COSIPY) predicts the observed quantities (B_{geod}, SLA and $\bar{\alpha}$), which deviate from this truth according to a normally distributed observational error and an unknown model error. Let **d** be the vector combining all N observations from our three observed data streams (B_{geod}, SLA and $\bar{\alpha}$). A single observation d_i is related to its modelled counterpart y^{mod}_i by: [...]*

L304: It would be helpful to already name LHS1 and LHS2 here (as defined in Table 1).

Done!

Figure 7: In the caption, it might improve readability to explicitly mention "ENS MEDIAN" and "95% CI".

Thanks for this suggestion. We have added ENS MEDIAN and 95% CI in the caption.

L 472: Typo

Thanks!

L510: Could you expand on what is meant by the "misrepresentation of surface albedo"? What do you think is the main issue here, for example, the simple decay-type albedo parameterization, limitations in the Landsat data (e.g., cloud cover leading to irregular observations or the 16-day revisit time), or something else?

Thank you for these questions. While the obstruction via cloud cover or the 16-day revisit time can limit the available information in the observations, at L510, we refer to the limitations of the decay-type albedo parameterisation. We have added more detail in the manuscript at now L555ff:
We attribute this difference to an underestimation of Q_{SWin} *(Fig. A2 and Fig. S17) and to limitations in the parameterisation of the albedo decay (Fig. 8).*

and in the discussion at 615ff:
More specifically, the albedo scheme does not explicitly represent processes like snow grain metamorphism or surface accumulation of meltwater, which accelerate albedo decline during intense ablation. While such a complex albedo parameterisation may not be necessary, the scheme of Oerlemans and Knap (1998) relies on a singular decay parameter to describe the albedo evolution, regardless of whether the surface is melting or not.

and L629ff:
Because satellite observations are less frequent during the melt season, the information available to constrain these albedo decay parameters was limited, as confirmed by the small positive bias in simulated ā. Furthermore, the repeat cycles of Landsat 5 and 7, combined with cloud cover, yielded a sparse observation timeline for both SLA and ā.

L 542: Brackets appear twice

Thanks!

L545–L547: It might be useful to briefly mention here that glacier albedo has been observed to decrease with climate change (e.g., Williamson et al., 2025). This would also highlight that capturing interannual variability in the albedo simulations is important, not only intra-annual variability.

Thanks for this suggestion. We have added a short statement at L596f:
Recent observations further indicate that glacier albedo has declined over recent decades, contributing to accelerated glacier melt (e.g., Williamson et al., 2025). This highlights the importance of accurately representing albedo evolution across spatial and temporal scales ranging from seasonal variability to long-term changes.

Section 4.3: I find this future work section very helpful and thorough.

Thank you very much!

L629: The mismatches between observed and simulated mass balance gradients are attributed here to linear lapse rates, but other factors could also contribute, so the wording may be somewhat misleading.

We apologise for the ambiguous wording. We have adapted Line 693ff:

Second, COSIPY is limited by structural simplifications, including the use of linear lapse rates and the lack of a representation for debris cover and snow redistribution (Voordendag et al., 2024; Zhu et al., 2024). The mismatch between observed and simulated mass balance gradients likely results from the combined influence of these structural simplifications, forcing biases, and unresolved small-scale variability in accumulation and surface energy fluxes (e.g., Sauter and Galos, 2016).

L640: I am wondering whether you applied any temporally-guided averaging of the Landsat observations (e.g., monthly or multi-month averages). This might help here.

We thank the reviewer for this suggestion. We agree that temporally-guided averaging is an interesting perspective for balancing the seasonal frequency bias. Such an averaging or, similarly, a weighted contribution, would prevent the clear-sky winter observations from dominating the log-likelihood scores. However, we did not apply this method here because observation density is already quite limited and averaging just a few data points over several months would not yield a robust average. More importantly, the temporally-guided averaging should not be applied to melt season values, as the evolution of the surface state (both snowline and albedo) is highly non-linear over the course of a single month and would create a synthetic state that is not representative. Since COSIPY actively tries to simulate this surface state transition, we prefer to keep the raw observations. In this paper, we partially account for the imbalance in the log-likelihood function as we average the scores by the number of data points. However, this does not fully resolve the internal seasonal bias within individual data streams.We added a statement at L708ff:
This issue is compounded by the higher availability of satellite observations under less cloudy and thus often winter and snow-covered conditions compared to the more information-rich ablation season, providing insufficient guidance for melt-related parameters. [...] Future applications may benefit from temporally-guided downsampling or adjusted weighting of winter scenes to reduce the seasonal frequency bias, while retaining the information content of ablation-season observations to capture non linear melt-out patterns.

L649: This is a very interesting observation: the “impact of parameter uncertainty is small compared to that of the forcing data.” I wonder whether there is a simple test or short analysis that could help quantify this effect.

We thank the reviewer for this comment. The original sentence was lacking a reference to Fig. 9, which partially shows this effect. In addition, we can evaluate the parametric spread reported in Table 2 (the 95% CI) to get a grasp of the mass balance uncertainty introduced by the parameter choices, and compare it to the change in mean annual mass balance introduced by forcing COSIPY with the posterior mean values with and without forcing corrections to Q_{LWin}, wind speed and air temperature for the hydrological years 2002 to 2009. We take these bias correction values from the comparison of the forcing against the upper AWS station for the hydrological year 2003/04 (see Figure A2) and apply them as constant over time.

The 95% credible interval spread stemming from parameter choices ranges from a minimum of 0.177 m w.e. (2001/02) to a maximum of 0.387 m w.e. (2002/03) with an average annual spread of 0.314 m w.e. a^{-1} (Table 2). Applying the posterior mean parameters without forcing corrections results in a mean annual mass balance of -1.724 m w.e.a^{-1} over the hydrological years 2002 to 2009. Applying the AWS-derived forcing corrections in Q_{LWin}, wind speed and temperature (taken from Figure A2), yields a mean annual mass balance of -1.189 m w.e.a^{-1}. This represents a shift of 0.535 m w.e.a^{-1}, which is nearly double the average parametric uncertainty. This correction also produces a mean annual mass balance that is closer to the reported means of the WGMS (-1.228 m w.e.a^{-1}) and that of Klug et al. (2018) (-1.374 m w.e.a^{-1}). While we acknowledge that this first-order test is not as accurate as a formal statistical variance decomposition, the contrast in the magnitudes supports our claim that the impact of forcing uncertainties exceeds that of the parameters.

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.

References:

Arndt, A., Scherer, D., and Schneider, C.: Atmosphere Driven Mass-Balance Sensitivity of Halji Glacier, Himalayas, Atmosphere, 12, 426, https://doi.org/10.3390/atmos12040426, 2021.

Ban, N., Schmidli, J., and Schär, C.: Evaluation of the convection-resolving regional climate modeling approach in decade-long simulations, Journal of Geophysical Research: Atmospheres, 119, 7889–7907, https://doi.org/10.1002/2014JD021478, 2014.

Ban, N., Caillaud, C., Coppola, E., Pichelli, E., Sobolowski, S., Adinolfi, M., Ahrens, B., Alias, A., Anders, I., Bastin, S., Beluši´c, D., Berthou, S., Brisson, E., Cardoso, R. M., Chan, S. C., Christensen, O. B., Fernández, J., Fita, L., Frisius, T., Gašparac, G., Giorgi, F., Goergen, K., Haugen, J. E., Hodnebrog, , Kartsios, S., Katragkou, E., Kendon, E. J., Keuler, K., Lavin-Gullon, A., Lenderink, G., Leutwyler, D., Lorenz, T., Maraun, D., Mercogliano, P., Milovac, J., Panitz, H.-J., Raffa, M., Remedio, A. R., Schär, C., Soares, P. M. M., Srnec, L., Steensen, B. M., Stocchi, P., Tölle, M. H., Truhetz, H., Vergara-Temprado, J., De Vries, H., Warrach-Sagi, K., Wulfmeyer, V., and Zander, M. J.: The first multi-model ensemble of regional climate simulations at kilometer-scale resolution, part I: evaluation of precipitation, Climate Dynamics, 57, 275–302, https://doi.org/10.1007/s00382-021-05708-w, 2021.

Collier, E., Ban, N., Richter, N., Ahrens, B., Chen, D., Chen, X., Lai, H.-W., Leung, R., Li, L., Medvedova, A., Ou, T., Pothapakula, P. K., Potter, E., Prein, A. F., Sakaguchi, K., Schroeder, M., Singh, P., Sobolowski, S., Sugimoto, S., Tang, J., Yu, H., and Ziska, C.: The first ensemble of kilometer-scale simulations of a hydrological year over the third pole, Climate Dynamics, https://doi.org/10.1007/s00382-024-07291-2, 2024.

Draeger, C., Radić, V., White, R. H., and Tessema, M. A.: Evaluation of reanalysis data and dynamical downscaling for surface energy balance modeling at mountain glaciers in western Canada, The Cryosphere, 18, 17–42, https://doi.org/10.5194/tc-18-17-2024, 2024.

Goger, B., Stiperski, I., Nicholson, L., and Sauter, T.: Large-eddy simulations of the atmospheric boundary layer over an Alpine glacier: Impact of synoptic flow direction and governing processes, Quarterly Journal of the Royal Meteorological Society, 148, 1319–1343, https://doi.org/10.1002/qj.4263, 2022.

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.

Lucas-Picher, P., Argüeso, D., Brisson, E., Tramblay, Y., Berg, P., Lemonsu, A., Kotlarski, S., and Caillaud, C.: Convection-permitting modeling with regional climate models: Latest developments and next steps, WIREs Climate Change, 12, e731, https://doi.org/10.1002/wcc.731, 2021.

Nicholson, L., Stiperski, I., Nitti, G., Prinz, R., Georgi, A., Groos, A. R., Shaw, T. E., Sauter, T., Haugeneder, M., Mott, R., Sicart, J.-E., Brock, B. W., Albers, R., Allegri, B., Barral, H., Biron, R., Charrondiere, C., Coulaud, C., Fischer, A., Reynolds, D., Richter, N., Schroeder, M., Vettori, P., Voordendag, A., and Wydra, C.: The second HinterEisFerner EXperiment (HEFEX II): Initial insights into boundary layer structure and surface-atmosphere exchange processes from intensive observations at a valley glacier, Bulletin of the American Meteorological Society, pp. BAMS–D–24–0010.1, https://doi.org/10.1175/BAMS-D-24-0010.1, 2025.

Ou, T., Chen, D., Tang, J., Lin, C., Wang, X., Kukulies, J., and Lai, H.-W.: Wet bias of summer precipitation in the northwestern Tibetan Plateau in ERA5 is linked to overestimated lower-level southerly wind over the plateau, Climate Dynamics, 61, 2139–2153, https://doi.org/10.1007/s00382-023-06672-3, 2023.

Shannon, S., Smith, R., Wiltshire, A., Payne, T., Huss, M., Betts, R., Caesar, J., Koutroulis, A., Jones, D., and Harrison, S. (2019). Global glacier volume projections under high-end climate change scenarios. The Cryosphere, 13(1):325–350.

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