Author Response to Referee David Rounce

We would like to thank you very much for the detailed and constructive review of our manuscript. In the following, you find our point-by-point list of answers to the raised comments. We are convinced that our actions will significantly improve the quality of the manuscript. We sincerely hope you find our response satisfactory, and we are able to overcome your methodological concerns. Referee comments are reproduced in blue font color. Our response and the undertaken actions are formulated in black font color.

General Comments

This study evaluates the performance of using different calibration data and different glacier mass balance models for glaciers in the Monte Sarmiento Massif. Specifically, the study uses a temperature-index model with three different calibration datasets/strategies that vary based on using ablation stake data, regional geodetic mass balance data, and both ablation stake and regional geodetic mass balance data. After the calibration is performed, three other models based on a simplified surface energy balance with and without a more complex radiation scheme as well as a full energy balance model (COSIPY) are calibrated, although some of the model parameters from the ablation stake and regional geodetic mass balance data are assumed constant for all four of these models. As more data becomes available, it is important for detailed studies to evaluate the effect of using different calibration datasets and different models in order to improve how well we can estimate present-day and future mass change, glacier runoff, etc. for different regions. Thus, I find this to be an important study that could be useful in guiding modeling efforts in the future.

While I enjoyed reading this study and overall the study was fairly well-written, there were a number of elements that I found reduced readability or hindered my ability to understand the study fully. For example, I would suggest avoiding acronyms to improve readability as much as possible, especially if the world limit is not a problem. For example, common acronyms like “MB” versus “mass balance” only saves a single word and little space but is much more readable. Less common acronyms such as OPM for orographic precipitation model drastically reduce readability. Additionally, the calibration methods should include more detail to support reproducibility and an understanding of how the methods were actually applied. The terminology used is a bit hard to follow (e.g., specific referring to the total mass balance and use of geodetic versus surface mass balance) and I would thus recommend using standard terminology from Cogley et al. (2011) to improve readability and understanding.

Overall, the study does a good job at referencing the current literature and stating their research questions and their findings. I find this study somewhere between major and minor revisions; thus, I am recommending this to be reconsidered after major revisions, although I will note that I believe these changes are relatively easy to accomplish.

General Comments

The use of standard terminology (e.g., Cogley et al. 2011) is needed to fully understand what is meant. The mixed used of terms such as specific, dynamical losses, mass balance, total

mass balance, etc. made it challenging to fully understand what was being stated.

Thank you for pointing out that shortcoming. We will clarify the terms of specific mass balance, surface mass balance, geodetic mass balance and dynamical losses accordingly following the suggested standard terminology.

The calibration methods and how model parameters were "transferred" to uncalibrated glaciers needs to be stated more clearly.

Thank you for this comment. We will explain the calibration strategies in more detail. For more details, see the response to the specific comments on this section.

The different types of uncertainty evaluated were unclear and I'm not convinced that they are achieving what they mean to. I may be wrong or misunderstood the experimental setup, but it should be very clear what uncertainty is being captured, which I think can be achieved by making the language more explicit as opposed to its current form where the text reads as "generalizable".

Thank you for pointing that out. We agree that the uncertainty assessment is more an analysis of model sensitivity to individual parameter combinations, and not of such a high relevance that it is necessary in the main body of the paper. We will move these sections (3.5.3, 4.5 and 5.5) in the supplement and state more clearly what is being analyzed. The idea is to analyze how the results vary with i) changes in the climatic input, which we see by varying the TLR and τ; ii) changes in the model-specific parameters (DDF_{ice} and DDF_{ice} for the PDD; C_0 and C_1 for the SEB models; and α_{ice}, z_{ice} and t_{albedo} for COSIPY); iii) different types of SMB models.

The one thing that remains unclear to me after reading this study is now that geodetic mass balance data is available for every glacier, why was this not included as a model option? Why do the calibration strategies go from ablation stake data to regional data as opposed to utilizing this intermediate scale dataset? I think this would add great value to the study; however, I accept that the authors may consider this beyond the scope of the study as it would require considerable modeling work and thus an adjustment of the results/discussion and interpretation.

Thank you for this question. The main reason why we did not use the specific geodetic mass balances of the individual glaciers for calibration, is that we deliberately withheld this 'intermediate' dataset for model validation. It is the only observational dataset covering each individual glacier in the study site. In-situ measurements are limited to Schiaparelli Glacier only. The regional mean geodetic mass balance, which we used for calibration strategies B and C, did not qualify as an appropriate validation dataset as well since local differences are not captured. We agree that utilizing this intermediate scale dataset could add value to the results. Suppose we pursued a glacier-specific calibration exclusively using the intermediate dataset, then the validation was mostly limited to the single regional average specific mass balance which was anyhow indirectly calibrated. For this reason, we decided to solely target the region-wide geodetic mass balance with the idea that this allows a general calibration of the magnitudes of the input and output to the SMB.

It is striking that this simple calibration target results in drainage-basin RMSE values comparable to uncertainties in specific geodetic mass-balance observations.

Specific Comments

L53-57: This is a remarkable difference and its unclear if this is meant to reflect a high amount of uncertainty in the estimates (i.e., if they covered the same area since they cover nearly the same time period) or if the next sentence is implying a reason for these differences since they covered different areas that have different precipitation patterns. Please clarify.

Thank you for this question. The two studies do indeed not cover the exact same area, since Melkonian et al. (2013) are focusing on the Cordillera Darwin itself, whereas Braun et al. (2019) consider Tierra del Fuego, thus a larger area. However, the main difference between the two studies is the methodological approach in the calculation of the elevation changes: Melkonian et al. (2013) assume penetration into the firn and compensate these effects by adding 2 m to each SRTM elevation over ice.

We will rephrase this section in a reworked version of the manuscript.

L97: The question (Q1) posed is a bit unclear as little background has been provided on calibration and transferability. I suggest reframing the question to make it clearer or add a little background.

Thank you. We will formulate Q1 more clearly in a reworked version of the manuscript.

L137: Is this 20% adjustment based on the sensor measurements, the studies mentioned by the previous studies, or is it just an assumption with no prior support as the sentence currently indicates? If the latter and precipitation is as important as specified in the introduction, then a sensitivity analysis of this assumption seems warranted.

Thank you for this question. It is almost certain that the measured precipitation suffers from some underestimation due to heavy winds and snowfall, since the bucket is not heated. It is correct that we do not know the exact relative value for under-catch. It might be even larger than the undercatch of 20% that we use as best guess based on the references given in the text and the conditions we experienced in the field.

This assumption is supported from literature, e.g.:

Schneider et al., 2003:

“During situations with high wind speed, unshielded rain gauges typically underestimate rainfall because of the deformation of the wind field around the collecting bucket (Yang et al., 1999). However, vibrations of the tipping gauge may produce extra counts during storms, thus overestimating rainfall. […] and for the multiple systematic errors of the precipitation measurement we consider this estimate to be good only within ±20%.”

At Schiaparelli, we do not have the issue of over-estimation from vibration of the gauge due to the solid installation of the device.

Weidemann et al., 2018b:

“Precipitation is measured at 1 m above the ground using unshielded tipping-bucket rain gauges. This type of measurement underestimates precipitation by up to 30% at wind speeds of 1.5 ms−1 and even up to 50% at wind speeds of 3.0 ms−1 (Rasmussen et al., 2012; Buisán et al., 2017). Windspeed induced deviations increase during snowfall due to an

intensified drifting of snow (Rasmussen et al., 2012; Buisán et al., 2017).”

The location of our AWS is less exposed to extreme winds as in other cases in Patagonia because it is located at low elevations and shielded by the valley slope at one and the glacier at the other side. Snowfall accounts for only a minor portion of precipitation at this location due to the low elevation. Thus, the under-catch due to wind will probably not be as high as 50 %.

Buisán et al., 2017:

“[…] wind is the most dominant environmental variable affecting the gauge catch efficiency, especially during snowfall events. At wind speeds of 1.5 ms^{-1} the tipping bucket recorded only 70% of the reference precipitation. At 3 ms^{-1}, the amount of measured precipitation decreased to 50% of the reference, was even lower for temperatures colder than -2 °C and decreased to 20% or less for higher wind speeds.”

However, to account for the lack of knowledge of the precipitation amounts in the MSM, we include the precipitation field (both amount and distribution) in the calibration via the parameter τ that is varied to four different values. The four different precipitation fields included here result in a wider range of changes in massif-wide accumulation than the 20% precipitation at the AWS.

Furthermore, the only value we take from these measurements is the average annual precipitation amount. We use this annual value as a constraint for the OPM to guarantee that the annual amounts are in the same order of magnitude as our measurements at the location of the AWS. We will reformulate this part in the manuscript and give more details about how we make use of what the precipitation measurement.

L199-214: I assume there is some temperature threshold or temperature threshold range used to distinguish snow versus rain? If so, this should be stated somewhere.

Thank you for indicating this missing information. The distinguishment between snow and rain is done in the SMB models. The threshold temperature is set to 1.0 °C. For the PDD and the two SEB variants, this is a hard threshold. In COSIPY, a logistic transfer function is used to derive snowfall from precipitation. The proportion of solid precipitation scales between 100% and 0% $\pm$ 2.5 °C around the threshold temperature of 1.0°C.

We will include this information in the description of the SMB models (Section 3.2).

Table 1: Unclear what Column 1 is. Value for atmospheric forcing should be negative. TLR appears in Table prior to being stated what it’s an acronym for within text.

We will include the suggested changes in Table 1, thank you, thank you. The acronym TLR has been explained in L 193.

Section 3.5.1 Calibration Strategies: The description of the calibration strategies is fairly broad and additional detail is warranted. For example, was a minimization algorithm used? How were the scores used to select model parameters (L335-340)?

We will give more detail of the calibration strategies in this section. The optimal setting for each calibration strategy was determined based on the mentioned model skill score. It depends on the misfit between model and observation, which is given by the mean squared error. The run

(combination of parameters) with the highest score gives us the optimal parameter combination. This calculation is performed for each strategy because other measurements are considered, and for each model.

For Strategy A, how were the glacier-specific parameters "transferred" to regional scales; was there just a single value that was determined that was assumed to be the same for every other glacier? Or was there some sort of transfer function? Additionally, what "ablation stake measurements" were used? How many ablation stakes are there (L144 only states "several")? What elevation range do these stakes span? Were they measured seasonally, annually, or something else? Was the calibration performed for the entire period (8/2013 to 03/2019) or was the higher temporal resolution data used? How was the calibration performed if there were multiple observations or different time periods and thus discrepancies between the model and observations that do not allow perfect agreement?

Thank you for your questions. For Strategy A, the forcing-related (fall-out timescales, temperature lapse rates) and melt-model-specific parameters (degree-day factors) are determined based on the measurements of Schiaparelli Glacier (stakes and mass budgeting). Subsequently the parameter set identified best was transferred to the entire study site without additional modifications. This way, we investigate if it is feasible to directly transfer the model parameters determined at one glacier to surrounding glaciers. We will reformulate the description of Strategy A to make this clearer in a revised version of the manuscript.

We will reformulate and include more information about the stake network in section 2.2 and include the stake locations of one measurement period in Figure 1. Since the stakes have not been installed at the same position every time, we think this is the best solution to give an idea about the locations. The formulation "spread over the ablation area" will be changed to "concentrated on the lowest part of the ablation area", which describes the situation much more realistic with the stakes being all quite close together. Thus, the altitude range is very limited to between around 150 and 200 m a.s.l.. Unfortunately, this is the only accessible area of the glacier. The time periods of stake reading are also not regular ranging from several months to almost one year. All stakes and the respective time period are shown in the supplement figures S4 and S5. The calibration was performed comparing modeled and measured ablation for each individual time period and stake.

For Strategy B, how were the parameters for the glaciers with significant calving losses selected? Were the model parameters varied to get perfect agreement between the modeled and observed specific mass balance or was some amount of uncertainty deemed acceptable? What constitutes "larger uncertainties" (L315)? If the model parameters are being calibrated to the regional specific mass balance, what is the reason that this cannot be done for smaller glaciers as well? Given that the smaller glaciers aren't calibrated, how are their model parameters determined? What percentage of the glaciers (by area) are actually calibrated using this approach?

Thank you for your question. The only glacier with significant calving losses is Lovisato Glacier (> 3 km^2), thus it is the only one excluded in this measure. We will write that more clearly in the manuscript.

Glaciers with an area < 3 km^2 are excluded because specific mass balances from the elevation changes cannot be determined accurately for small glaciers. Having voids in the satellite

products, the accuracy decreases rapidly for small glaciers, thus including those in the calibration is rather disadvantageous. We aim for perfect agreement between model and observation because otherwise we would end up with many parameter combinations hitting the rather large range of uncertainty of the geodetic mass balances. The aim is to determine model parameters suitable for the whole massif. Thus, glaciers excluded in the calibration (Lovisato and glaciers < 3 km^2) are modelled with the same model parameters. 71% of the total glacierized area is included in the B_{MSMnc}.

The model parameters determined in the calibration are fixed for the entire study site.

Figure 3 suggest that Strategy C includes the mass budget (assuming this is the elevation change data); however, it is not clear from the text (L318-326) how this data is incorporated as it does not appear to be mentioned. The text of which parameters is calibrated for which models and how the calibration is done (L323-326) is similarly very vague.

The mass budget included in Strategies A and C is the total mass budget of Schiaparelli Glacier. It is the combination of the SMB and the mass lost through a flux gate parallel to the glacier front. This value is comparable with the elevation changes of Schiaparelli Glacier in the area above the flux gate. This is explained in Section 3.4. We will make it clearer, to which exact calibration targets we refer in Figure 3.

We will also give more explanation on the set and choice of calibration parameters for all four models at the beginning of Section 3.5.1.

L347 – state what "where we have measurements between 2013 and 2019" means.

We have ablation stakes measurements between 2013 and 2019. We will reformulate the whole section according to the next comment.

Section 3.5.2: Is this "model evaluation and intercomparison" an independent validation step or is this more detail on the calibration? The datasets described appear to be used in the calibration, so it's unclear how this is used for model evaluation as well. Please clarify.

Thank you for the question. The section "3.5.2 Model evaluation and intercomparison" describes the model validation and the intercomparison of the four SMB models. For validation we use the specific geodetic mass balances of the individual glaciers (land-terminating, > 3 km^2 only). The ablation stakes are considered as an additional measure for intercomparison of the four models. Since the ablation stakes have only been considered in calibration Strategy A, they are an independent dataset for model intercomparison, where we followed Strategy C.

We will reformulate this paragraph accordingly in a revised version of the manuscript.

L352: lower-case "c" for "climatic" forcing-related …

Thank you. We will change that in a revised version of the manuscript.

L354: is rainfall also important for COSIPY given that it considers refreezing?

Thank you for your question. Rainfall is considered in COSIPY, but the impact on the energy

balance is overall negligible. However, we will reformulate “snowfall” to “precipitation” in this sentence to be more accurate.

L354-356: could you clarify the difference between uncertainties “related to process parameterizations” which falls under model-inherent uncertainties versus “model type” which fall under model type-related uncertainties as they sound the same? It appears that the second type of uncertainty is primarily focused on the calibration procedure and methodological choices as opposed to the physical parameterizations.

See response to General Comment 3.

L367: Does TLR not also affect the melting? Seems overstated that this solely effects the amount of snowfall.

Thank you for this question. We wanted to highlight that with considering the total mass budget of Schiaparelli Glacier, we are able to constrain not only ablation, but also accumulation. The formulation is indeed misleading. We will rephase in a revised version of the manuscript.

L369: what does a “profound” estimate mean?

With the formulation “profound estimate” we want to emphasize that the precipitation estimates are based on a scientifically sound comparison with observations (total mass budget), which has not been done in previous studies in the area (e.g. Weidemann et al., 2020), where the precipitation amounts are one of the main uncertainties due to the lack of observations. We will reformulate to “well-informed estimate”.

L372: Doesn’t it also tell you that there is a stronger melt gradient with respect to elevation?

Thank you for this question. We agree that a stronger TLR not only implies more snowfall but also a stronger melt gradient with respect to elevation. We will adjust the paragraph accordingly.

L376: unclear how this “transfer” is done.

See response to specific comment 7 on the calibration Strategy A.

L377: the “specific” mass balance is merely an area-averaged mass balance (see Cogley et al. 2011, https://wgms.ch/downloads/Cogley_etal_2011.pdf). The “surface mass balance” is thus technically the “specific surface mass balance” and the “mass balance” in this case is referring to the “total mass balance”. I suggest modifying this use of specific and “total” throughout to be consistent and properly use standard terminology.

We will adjust these terms following standard terminology, thank you.

L381: what do you mean by “dynamical losses”? Frontal ablation? Additionally note the inclusion of “annual” here, but all results shown are “annual”. If you’re going to specify annual, then this should be added to each time this is stated; otherwise, suggest deleting it here for

consistency.

Thank you for your comment. We will change the word “dynamical” to “calving”.

L385: “second step” implies that this expands upon Strategy A, but Strategy B I thought was independent on Strategy A. Please clarify here or in the methods.

We will rephrase the sentence, thank you.

L393: suggest listing a few of the names where this agreement has increased here.

We will list the glaciers where the agreement has increased, as suggested, thank you.

L395: I’m confused as to where this negative SMB bias from calilbration Strategy A is shown. Table 2 suggests that Strategy A results in a positive SMB and thus a positive bias, not negative? Please clarify.

Thank you for pointing out this typo. It is indeed a positive bias.

Figure 4: see my comment about L377. It is thus unclear what Figure 4 is actually showing. I assume it’s showing the difference between the surface mass balance and total mass balance, i.e., the amount of frontal ablation, which was stated in L378 for Schiaparelli. However, L378/379 states that Figure 4 is showing the difference between the surface mass balance and geodetic observation, which is very different. Please clarify as I currently don’t know how to interpret the results of Figure 4.

Thank you for your comment. What is shown in Figure 4 is the difference between the modelled specific surface mass balance and the observed specific geodetic mass balance. As you assumed, this gives us the calving rates for the calving glaciers (dotted). And the absolute error between model and observation for the land-terminating glaciers, where the difference between both dataset would ideally be zero.

We will adjust the terminology in the figure caption accordingly to make it better understandable.

Table 2: Is Schiaparelli the only glacier with frontal ablation? Or is frontal ablation included her by some other means given that what is reported is the “specific mass balance”. Is the third column the observation? If so, this should be stated clearly.

All lake terminating glaciers are marked by asterisk. Frontal ablation is not included in the surface mass balance of these glaciers but in the geodetic mass balances. Column 3 gives the specific geodetic mass balances, thus, the observation that modelled SMB (columns 4-9) is compared to. We will state that more clearly in the table, thank you.

L427: consider using “accumulation” instead of “snowfall” because COSIPY technically also includes internal accumulation from refreezing.

Thanks for that suggestion. It is true that COSIPY also includes refreezing and deposition in the accumulation. Here we, however, explicitly want to address snowfall.

L434: what's the difference between "huge" and "very huge". Suggest deleting "very".

We will change that as suggested.

L437: cite study or show in supplemental figure?

We will include a table with the average end of summer snow line altitudes (2003-2022) of the four largest glaciers in the MSM in the supplement.

L439: Figure S3c is showing the snowfall, which is primarily showing that there is no snowfall and thus that the temperature is above the snow/rainfall threshold for almost the entire glacier. If showing the mass balance for the summer, perhaps at a 3rd column of subfigures to Figure S3 to make this clear?

We are not sure if we understood the suggestion correctly. We understood that you were asking for a 3rd column in Figure S3 with the winter and summer SMB. We would agree that this a good idea to include the winter and summer SMB in this figure, and will do so in a revised version of the manuscript.

L440: Is the "largest part" referring to a specific area of the MSM or is this meant to state that almost the entire area of the MSM has a positive mass balance? I assume it's the latter, so suggest clarifying.

Thank you for the question. The "largest part" refers to majority of the MSM area showing positive MB. We will formulate that more clearly.

L442: If only 33% accumulates in winter and 13% in the summer, then does the remaining 54% accumulate in the spring in fall? It'd be good to specify what time periods "winter" and "summer" refer to to make this and the figures clear.

Thank you for this question. Spring and fall contribute 31% and 22% to the annual snowfall, respectively. Thus, we have the largest contribution to annual snowfall in winter. However, since the amounts for winter and spring are in a similar order, we will rephrase the sentence to make it clear that the largest part of snowfall (65%) is accumulated in these two seasons, whereas only a small part (13%) accumulates in summer. We will also give the exact months for each season in brackets.

L446: "22-year"

Will be changed as suggested.

Figure 6: COSIPY appears to show more negative MB during negative years and at times more positive MB in positive years (e.g., 09-10 and 10-11), so it doesn't seem to be as consistent of a signal as sated on L451; albeit COSIPY is more negative on negative years as stated.

We will rephrase the sentence to put the statement more accurately.

L447: It's unclear to me how varying the TLR and tau allows one to assess the uncertainty related to the climatic forcing given that these are calibrated model parameters? This sensitivity analysis instead seems to look at the model parameter uncertainty. If one were to analyze the uncertainty of the climatic forcing, I would have expected a different climate product/reanalysis dataset to be used, which is not the case.

See response to General Comment 3.

L462: The "model-specific" uncertainty seems to have similar issues as my previous comment, since TLR, tau, and snowdrift parameters are assumed to be the same as the PDD (L411); thus, it's really only looking at a subset of the model parameters across the models, no? It would be good to make this explicitly clear.

See response to General Comment 3.

L462-469: I'm not sure what value of information this adds because the subset of model parameters that is being modified have specific ranges. Hence, whether one model has a higher or lower range is merely dependent on the range of values selected. What information is gained by this analysis? Is it that the ranges reflect the values used in literature and thus when you use those values different models are more sensitive than others?

We analyze the sensitivity of the four individual models to the model-specific parameters that are calibrated. We agree that the range and sample size of calibration parameters impact the analysis as stated in L 593f.

L476-482: Is there a reason why Strategy A, B, and C are being discussed given that Strategies A and B were only applied to the PDD model, yet the sentence before and after refer to all the models? This seems to be out of place and is confusing since it's also not explicitly stated that these sentences only refer to the PDD model.

Thank you for this comment. We agree that this formulation is possibly causing more confusion than adding content, and will delete the sentence in a revised version of the manuscript.

L479: Can you provide an explanation for why this change in performance occurs?

With Strategy C, we are able to simulate the SMB of all glaciers in the MSM satisfactory. However, the agreement with observations at Schiaparelli Glacier is overall not too good with Strategy C. But it seems that with COSIPY the agreement at Schiaparelli Glacier above the fluxgate (Schiaparelli_FG) is better (also seen in Table 2). Including the Schiaparelli budgeting, the RMSE for COSIPY decreases whereas an increase is seen for the other models. Since Schiaparelli Glacier is the largest glacier in the area, it has a significant impact on the area-weighted RMSEs here.

The reason, why Schiaparelli Glacier is showing so different behavior, might be in its geometry: It has an extremely large ablation area and covers an extreme altitude range (from almost sea level to 2200 m). Looking at the SMB results from COSIPY (Fig. 5), we see a stronger gradient with more intense ablation in the lowest parts of the massif compared to the other models, and

a more positive SMB in the highest parts of the glacier related to the additional processes considered in accumulation (mainly refreezing). If we cut Schiaparelli along the flux gate, the lowest part of the glacier with the most extreme ablation is cut off, and the SMB in the area above the flux gate is less negative than for the other models, and thus closer to observations.

We will add this hypothesis to the discussion, section 5.4, thank you.

L483-485: Why changing from 10 to 5 best ranked runs? Was there something wrong with ranked runs 6-10 in this case?

Thank you for this question. We give the explanation for this in L 590-593 at the moment, thus later in the text. To correct this, we will move this explanation to Section 3.5.2.

L532: Suggest changing “model” to “model parameters” since the model can clearly be run at other sites, but it’s the parameters that are assumed constant that is the issue.

Will be changed as suggested, thank you.

L583: Again, are dynamical losses referring to frontal ablation? Otherwise, with these glacier-wide values, the total mass balance should equal the surface mass balance (assuming internal and basal mass balance is negligible).

We will adjust the word “dynamical” to “calving”.

L592: This should be stated at L483-485 (see comment above).

See comment above to L483-485.

L594: “is strongly dependent” perhaps?

Will be changed as suggested, thank you.

Section 5.3: This section did not add much value beyond reiterating what seemed to already be stated in Section 4.3.

See response to General Comment 3.

L605-606: citation is needed for “previous studies”

We will add the references here, thank you.

L629: what’s the difference between “strong” and “very strong” correlation?

We will remove the word “very”.

L632: I don't understand how these models overestimate the mass balance when the MSM mass balance is specifically used for calibration. If agreement is not matched, then it seems to highlight a problem with the model calibration as opposed to the model performance itself; unless independent datasets are being used. I also note that "overestimate" the B_MSMnc is a bit hard to understand whether this is more or less mass loss; hence, I would suggest stating more positive or more negative mass change to make this clear.

The models are calibrated towards two datasets: the B_{MSMnc} and the mass budget of Schiaparelli Glacier (see Fig. 3). The latter is pushing towards lesser ablation because the simulated total mass budget is more negative than the mass budget based on the observations (elevation changes and mass flux through the flux gate). This causes the less negative values of B_{MSMnc} in the end.

We will reformulate the "overestimation of the B_{MSMnc}" to make it easier understandable as suggested, thank you.

L636: see comment above. This line is very hard to understand given the terminology being used.

See response to mentioned comment above.

L641: Unclear what "the question" refers to as no question was given.

Thank you for this comment. We will rephrase the sentence.

L648: consider removing the double negative and changing "not unrealistic" to "realistic"

Will be changed as suggested.

Supplementary Figures 1,2,5: the text in these figures' scales and labels are too small to read.

We will adjust the scales and labels to make them better readable, thank you.

Figure S1: The x-axis appears to show the difference between the ddf_ice and ddf_snow; however, it now seems like the "-" is meant to show the two different values. This is rather unclear and I suggest making it easier to read perhaps in a list format if this is what's actually being shown. The same thing for the y-axis. This also suggests that grid search was conducted for the calibration as opposed to any minimization/maximization. This should be specified in the methods.

Indeed the "-" means to separate the two different values. We will adjust the axes to prevent confusion, thank you.

Figure S3: the scale label suggests this is showing snowfall/melt, which is not the case. I suggest clarifying this perhaps by putting labels above left and right figures of "accumulation" and "ablation" or changing the "/" to "or" to make this clear.

We will adjust the axes to prevent confusion, thank you.

Code and data availability – It's surprising to see the "Meteorological and ablation stake observations are available on request." What is the reason for these not being deposited in a permanent archive thus ensuring the data is publicly available?

Thank you for this comment. We will upload the observations of the automatic weather stations and ablation stakes used in this study to a publicly available data repository (Pangaea) to follow good scientific practice. Furthermore, we will upload the model forcing and final SMB results of this study.

References

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

Buisán, S., Earle, M., Collado, J., Kochendorfer, J., Alastrué, J., Wolff, M., Smith, C.G., and López-Moreno, J.I.: Assessment of snowfall accumulation underestimation by tipping bucket gauges in the spanish operational network. Atmos. Meas. Tech., 10, 1079-1091. https://doi.org/10.5194/amt-10-1079-2017, 2017.

Cogley, J. C., Rasmussen, L. A., Arendt, A. A., Bauder, A., Braithwaite, R. J., Jansson, P., Kaser, G., Möller, M., Nicholson, M., and Zemp, M.: Glossary of Glacier Mass Balance and Related Terms, IACS Contrib. No. 2, 2011.

Rasmussen, R., Baker, B., Kochendorfer, J., Meyers, T., Landolt, S., Fischer, A. P., Black, J., Thériault, J.M., Kucera, P., Gochis, D., Smith, C., Nitu, R., Hall, M., Ikeda, K., and Gutmann, E: How Well Are We Measuring Snow: The NOAA/FAA/NCAR Winter Precipitation Test Bed, Bull. Am. Meteorol. Soc., 93, 811–829, https://doi.org/10.1175/BAMS-D-11-00052.1, 2012.

Schneider, C., Glaser, M., Kilian, R., Santana, A., Butorovic, N., and Casassa, G.: Weather Observations Across the Southern Andes at 53°S, Phys Geogr, 24, 97–119, https://doi.org/10.2747/0272-3646.24.2.97, 2003.

Weidemann, S., T. Sauter, R. Kilian, D. Steger, N. Butorovic and C. Schneider: A 17-year Record of Meteorological Observations Across the Gran Campo Nevado Ice Cap in Southern Patagonia, Chile, Related to Synoptic Weather Types and Climate Modes, Front Earth Sci, 6, https://doi.org/10.3389/feart.2018.00053, 2018b.