Author Response to Referee Enrico Mattea

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.

The study by Temme *et al*. employs models of various complexity level (from degree-day to full energy-balance) to simulate glacier mass balance at the Monte Sarmiento Massif (MSM), Tierra del Fuego. The models are calibrated against geodetic mass balance estimations, testing three different calibration strategies, and evaluated using an objective aggregate score. The Authors conclude that regional geodetic observations are the better calibration target to improve model transferability, compared to single-glacier mass balance; the addition of a snowdrift model increases overall model performance. At the same time, no single model clearly out-performs the others, and comparison to *in situ* ablation measurements shows very poor agreement for all tested approaches.

The research questions addressed by the Authors are relevant and of current interest – especially calibration of physical mass balance models and assessment of the benefits of added complexity compared to parametrized approaches (e.g., Brun *et al*., 2022). The investigated location is important for an improved coverage of diverse climatic and topographic settings in mass balance modeling. Furthermore, I appreciate the Authors honest presentation of the challenges facing model calibration, validation and transferability.

Still, in the current form the study raises methodological concerns about the input datasets processing and the model calibration choices. These could potentially lead to significant differences in the reported results, and need to be discussed by the Authors. Presentation of the methods and results also needs to be improved, both to ensure reproducibility and to better substantiate the conclusions. Thus, the manuscript clearly needs major revisions. My review includes three Major and some Minor comments that should be addressed in the Authors response, and several Technical comments which refer to individual statements, tables and figures.

Major comment 1: model sensitivity and the choice of calibration parameters

One stated focus of the study is the calibration of surface mass balance (SMB) models of various complexity. As such, the choice of which model parameters are subject to calibration (and of the explored ranges of values) is crucial and must be informed by a well-documented sensitivity analysis – all the more so when models are run in a setting with scarce *in situ* observations like the Cordillera Darwin. In fact, sensitivity of physically-based glacier mass balance models like COSIPY is an important topic of current research (e.g., Brun *et al*., 2022, and reviewer comments therein; Mattea *et al*., 2021). Comprehensive sensitivity analyses from diverse glacierized regions are needed to assess the benefits of increased model complexity, which is one of the stated purposes of the present study.

The Authors select some parameters for calibration (Table 1, ll. 268-269), without showing nor discussing the associated sensitivity testing; further on, there is no more discussion of the consequences of leaving other parameters at their default values. Such values are either arbitrarily chosen, or calibrated by previous studies in settings potentially very different from the MSM study area.

In fact, the best-performing model runs all achieve very similar skill scores for each model type (Fig. S1, S2): as such, the choice of a best-performing parameter set can certainly be affected by the values selected for the other parameters (not considered for calibration). In other words, multiple combinations of physically plausible values can produce very similar results for glacier-averaged mass balance. With little *in situ* data available (notably a complete lack of accumulation measurements), the simulation is therefore largely under-constrained; calibration choices made by the Authors must be better discussed.

Find our answer at your next but one comment.

In particular, the correction of precipitation under-catch is set at 20 % throughout the simulations (l. 137), with no supporting evidence. Such a parameter is known to be highly uncertain and time-dependent (e.g., Sevruk, 1997; Barandun *et al*., 2015; Buisán *et al*., 2017), and exerts a direct control on modeled mass balance – so much that it is the one parameter of choice for model calibration to geodetic mass balance in Huss *et al*. (2009). While the claimed focus of the manuscript is more on calibration of the melt model (l. 14), several sections refer to *SMB* model performance and transferability, clearly including also accumulation (ll. 199-214). Moreover, the snowdrift module used in calibration strategy C is allowed to alter snowfall totals by ± 10 % (l. 320). Given the relatively small 20 % precipitation correction, such a potential bias is significant and should be discussed.

Thank you for this comment. We agree that precipitation is highly variable both in time and space. The assumption of an under-catch of 20% of precipitation is related to annual average and, thus, only applied to the average annual precipitation amount, not as an addition to single precipitation events along the time series. Typically, bucket-based precipitation measurements show under-catch due to wind and snow, specifically if - as in our case - the bucket is not heated.

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, which leaves us with +20%.

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).”

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.”

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 %.

We use the average annual value (+20%) as a constraint for the OPM to guarantee that the modelled annual amounts are in the same order of magnitude as observed at the AWS. We know that the measured precipitation is underestimated. It is correct that we do not know the exact relative value for such under-catch from measurements. It might be even larger, as seen above. However, to account for the large uncertainties related to precipitation, our calibration strategy comprises precipitation fallout timescales τ. These timescales have a large influence on precipitation amounts and distribution at higher elevation. The range of τ-values results in total accumulation variations that exceed ±10/±20%. Therefore, uncertainty associated to the under-catch assumption is considered secondary.

With the inclusion of the τ and the temperature lapse rate in the calibration, we additional have control on the solid precipitation and with it on the actual accumulation. This way, also the accumulation is calibrated by varying the temperature and precipitation field. We agree that the formulation of “melt model calibration” or “SMB model calibration” is not uniform throughout the manuscript. We will adapt that in a reworked version of the manuscript. Furthermore, we will clarify the section about the usage of the AWS precipitation measurements in this study as explained above.

Other parameter choices which should be addressed in the manuscript include atmospheric transmissivity (l. 242); fresh snow albedo in COSIPY (as *DDF$_{snow}$* is indeed calibrated in the PDD model); the threshold temperature for solid/liquid precipitation; and the temperature at which melt can occur in the PDD model. For each of these parameters, the Authors should provide supporting evidence for the used values; or at least comment on the consequences of them being somewhat arbitrarily chosen.

Focusing on the COSIPY model, as acknowledged at l. 549, the best performing set of calibrated parameters lies on the margin of the tested ranges – for all three parameters (Table 1, Fig. S2c). I commend the effort by the Authors to not introduce physically implausible values in the simulations (l. 550); nonetheless, such a result confirms that the value of one or more other parameters (not considered for calibration) is not optimal. This should be discussed, since the purpose of calibration is usually to find a local maximum of model skill within the tested parameter ranges – not outside. In particular, the best parameter set appears to minimize energy inputs to the glacier (highest albedo, slowest albedo decay, smallest roughness length in a warm and moist setting). A well-documented examination of the simulated energy fluxes may yield some insights into the causes of the observed model behavior.

Thank you for your comment. We will give more explanation and rationale for the choice of the calibration parameters in the beginning of Section 3.5.1 based on the following arguments: We chose the parameters in a way to cover all relevant contributions to the SMB. The snowfall and temperature-dependent melting are controlled by the temperature lapse rate and τ. For the PDD and the SEB model variants, we deliberately limit ourselves to calibrate the model-specific parameters (the $DDF_{ice/snow}$ for the PDD and the $C_{0/1}$ for the SEBs). For COSIPY, we have to constrain the number of calibration parameters to limit the computational effort in a feasible dimension. Based on discussions with the COPSIPY-developers in the team, we decided for the ice albedo and roughness length of ice, which address both the radiative and the turbulent energy fluxes. The albedo time constant controls the firn- and snow-covered part of the glaciers and is rather uncertain.

Furthermore, we did intense sensitivity testing during the preparation phase of this study. Parameters we considered in these sensitivity tests are: i) the temperature threshold of transfer rain - snow; ii) the albedo of snow, and iii) the methods of stability correction available in COSIPY (Monin-Obukhov similarity theory and bulk Richardson-Number). For the latter methods (iii), we found that the similarity theory clearly outperforms the bulk approach. Therefore, we constrained our simulations to the former method. For the temperature threshold (i), a redundance with temperature lapse-rate (TLR) tuning was experienced and we prescribed a single value for all model variants. Concerning the snow albedo (ii), a clear preference for high values was identified so we kept the maximum value of 0.90.

Still, we share the reviewers concerns on the COSIPY calibration that the best scores are achieved for parameters at the margin of the accessible parameter space. Therefore, we discussed again with the COSIPY-experts among the co-authors, which important parameters might have been neglected so far that would cause a lower energy input to the surface. We decided for a two-fold strategy:

1) We expanded the range of the ice albedo in COSIPY to a value up to 0.467 to see if the value of 0.4 is the local maximum, or if we can improve the results with a higher value. From our personal experience in the field, we know that the ice surfaces in the Monte Sarmiento Massif are very clean which might justify such a high ice albedo.

2) We added the firn albedo (0.50, 0.55, 0.60, 0.65) as an additional calibration parameter, which has shown to be an important calibration parameter for COSIPY before (e.g., Arndt et al., 2021). This parameter controls the energy balance in the higher elevated, firn-covered part of the glaciers. We had the suspicion that the current value (0.50) might have been too low, which would explain a high-bias in the ice-albedo calibration. To limit computation costs, we therefore excluded one other calibration parameter, i.e. the time constant of snow albedo aging. We decided to fix the albedo time constant at 22 days which is i) the optimum value from our current analysis and ii) the default value from literature (Oerlemans and Knap, 1998).

Extending the calibration to the firn albedo and increasing the ice-albeod range, we were surprised to exactly end up with the same optimal parameter combination as before: α_{ice} = 0.40, α_{firn} = 0.50, z_{ice} = 0.3 mm, (t_{albedo} = 22 d as fixed now). This means that for the ice albedo we found a local maximum and do not lie at the margin anymore. However, for the other two parameters, values remain at the margin of the parameter space. For the roughness length, lower values seem unjustifiable with literature. For the firn albedo, expanding the range to values below 0.5 would make it comparable or even identical with albedo values for ice. Thus, we also refrain from allowing a larger parameter range.

Further reasons for the difficulties with the COSIPY-calibration, other than the model-inherent parameters, might lie in the input dataset. We trust the temperature and precipitation fields determined in the PDD-calibration. The TLR fits well with the TLR calculated from ERA5 data. But there are several variables that are only considered in COSIPY and not in the other models, which might cause the issues for COSIPY. These are for example wind velocities and relative humidity, which both affect turbulent heat fluxes and thereby impact the choice of ice roughness length. We will include these thoughts to the Discussion in L548-550.

Looking at the new aggregated skills, we again see that several parameter combinations, also some not at the margin, perform very similar (see Fig. 1), with close results. Thus, if we would for example chose #2 (α_{ice} = 0.367, α_{firn} = 0.55, z_{ice} = 0.3 mm), we would be away from the margins for two of the three parameters. Respective model performances would not drop much not only in terms of calibration but also with respect to validation. However, we want to follow the skill ranking to select our optimal model setup and thereby exhaust the parameter ranges.

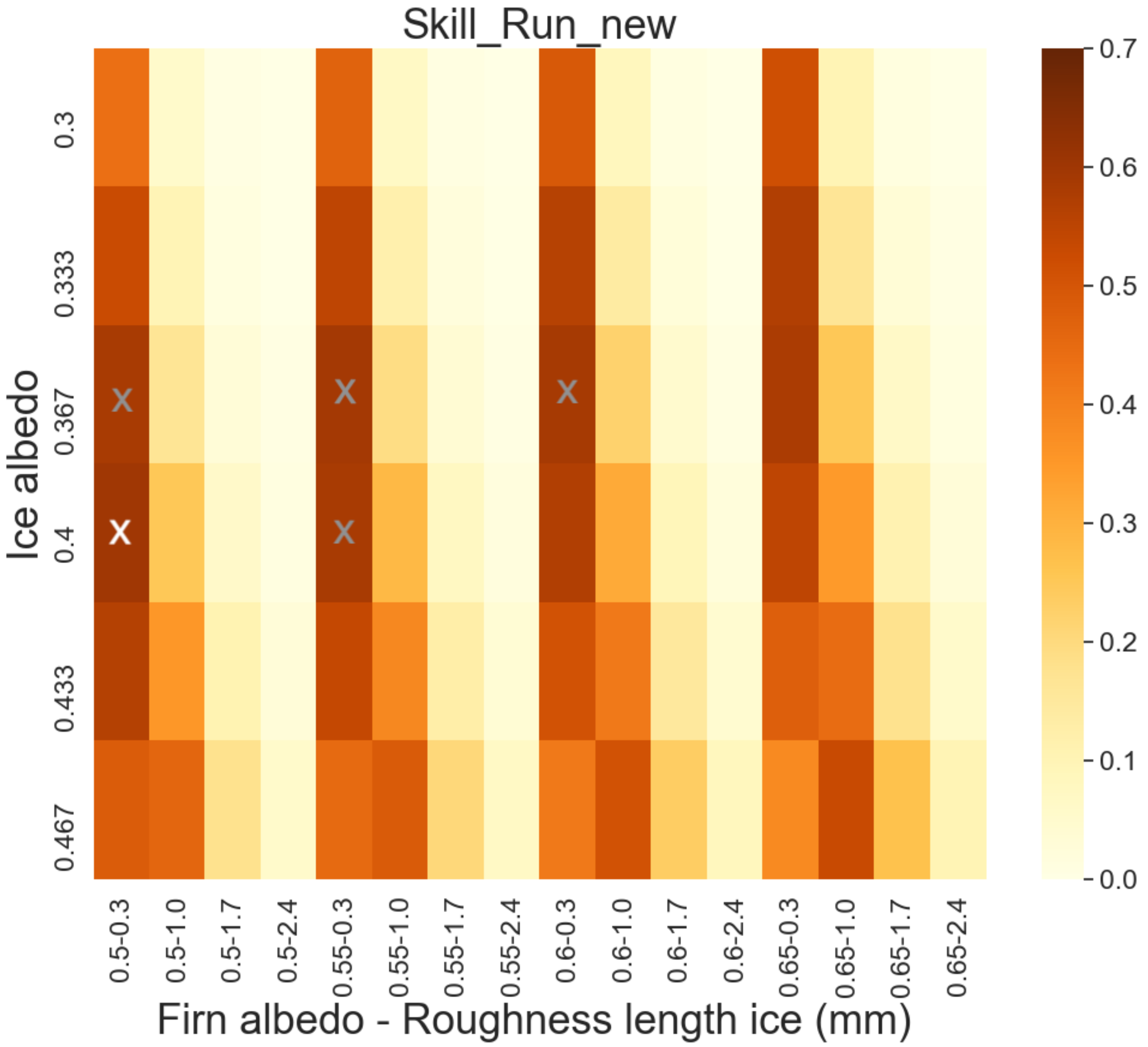

Figure 1: Heat plots of the model-specific calibration showing the aggregated model skill for COSIPY. The highest perfoming run is highlighted with a white cross, position 2-5 with a grey cross.

Note that the final results in the revised manuscript might still differ, since we did not yet include the outline changes (major comment 2) here. This optimization will improve the calibration further and might determine another optimal parameter combination in the end.

While all models achieve quite a low RMS error compared to the glacier-wide geodetic estimations (Table 2), agreement with the *in situ* ablation measurements is very poor (Table 3) and should be better discussed. Importantly, model biases appear to persist (for a given observation period) across stake locations (Fig. S4). The presence of large, spatially coherent biases should be investigated. It could indicate an enduring model miscalibration (Mattea *et al*., 2021), or the input meteorological series could include biases or major outliers – although the effect of the latter could be partly mitigated by the use of downscaled reanalysis data. Some questions that could be addressed include: if stakes are “spread over the ablation area” (l. 144), why are melt amounts almost the same at all stake locations according to the PDD model (Fig. S4)? Is the drop in modeled melt over 2016 (Apr-Oct) supported by a drop in PDDs? If yes, what is then the role of incoming radiation? (2016 Apr-Oct is notably the only instance in Fig. S4 where the SEB_Gpot simulates more ablation than SEB_G). It would also be interesting to calculate the cumulative sum of positive degreedays estimated at each stake location over each period shown in Fig. S4, and also to compare it against the value computed from AWS data only.

Thank you for your questions regarding the ablation stake measurements. We will include more information about the stake network 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 better reflects the actual survey network.

Overall, we share your concerns about the poor agreement between the in-situ ablation measurements and the model results. However, two major points led us to the decision to accept the results as they are:

a) Although for a certain period there seems to be a model bias, this bias is not persistent over the whole period. Until November 2018, all models tend to underestimate the ablation (except for COSIPY, which is giving closer results). However, afterwards all models (except for COSIPY) match the measurements quite well, and only COSIPY overestimates the ablation. With this contrary behavior we were not able to infer a single reason/explanation for the poor agreement.

b) Generally, these measurements have to be treated with caution. If you look at individual measurements more closely, some questions arise. One example you have named yourself is the Apr-Oct 2016: This is a (and unfortunately the only) measurement spanning exactly the WINTER period. However, the measured ablation (roughly 7 m w.e.) is in the same order of magnitude as the measurements Oct 2016-Mar 2017 spanning the SUMMER period. Although we know that melt occurs all year round the high value of this winter observation seems unrealistic and we assume a measurement issue. There are more examples of stake measurements that appeared flawed, and we therefore scrutinized all stake measurements very carefully. Extremely unrealistic measurements were already excluded in our analysis, and we decided to further drop the Apr-Oct 2016 stake measurement in the revised version of the manuscript. The measurements kept for analysis are the best we have in the whole area of the Monte Sarmiento Massif (and to our knowledge the whole Cordillera Darwin).

The reason why the period Apr-Oct 2016 is the only period where SEB_G drops below SEB_Gpot is probably connected to the fact that this is the only winter season. Here we observe different conditions regarding cloud cover and shading.

We did calculate the DDF_{ice} directly from the measured ablation at the stakes and the positive degree-day sum at the stake location (see l. 518-520). The values calculated are close to the calibrated DDF_{ice} of 5.0 mm d^{-1} $°C^{-1}$. For the individual stakes we get an average DDF_{ice} of 6.0 mm d^{-1} $°C^{-1}$, for the automatic ablation sensor an average DDF_{ice} of 5.0 mm d^{-1} $°C^{-1}$.

Finally, the aggregate model score (l. 334) is suitable for model ranking, but the actual sensitivity (i.e., the impact of a parameter change on modeled mass balance) is arguably more interesting for inter-model comparisons, uncertainty assessments and the design of future studies. It should therefore be briefly summarized for each parameter, and possibly reported in extended form in the supplementary materials.

The sensitivity of each model to the calibration parameters is visible in the supplementary figures S1 and S2, and also discussed in the uncertainty assessment (Section 4.3 and 5.3), where we analyze the results that we get by varying different parameter combinations. From referee 2 we received the comment, that this assessment is not really an uncertainty quantification but more a sensitivity analysis, which we agree on. We will move these sections in the supplement and rewrite it to a more comprehensive sensitivity analysis.

Major comment 2: reference-surface mass balance compared to geodetic mass balance

SMB in the four models is computed over 2000-2022 (and sub-periods) using the constant glacier outlines of Barcaza *et al.* (2017) and presumably a constant digital elevation model (DEM). This approach is commonly referred to as the reference-surface mass balance (RSMB; Elsberg *et al*., 2001), as opposed to the so-called conventional mass balance (CoMB), which is calculated taking into account the temporal evolution of glacier extent and hypsometry (Huss *et al*., 2012).

Glacier retreat – taking place mostly at the terminus, where specific mass balance is more negative – provides a stabilizing (negative) feedback, which reduces mass losses. As such, over the years the cumulative CoMB of a retreating glacier will accumulate an increasingly positive bias compared to the RSMB (Fig. I). The magnitude of such a bias is related to the extent deglacierized during the study period, especially increasing (on a retreating glacier) if the reference surface is measured at the start (Elsberg *et al*., 2001; Mukherjee *et al*., 2022). A larger bias is also possible on glaciers with steep mass balance gradients, as in Tierra del Fuego.

The RSMB is arguably more useful than the CoMB for climatic interpretations (e.g. Harrison *et al*., 2005); but unlike the geodetic mass balance it does not simply reflect mass change at the considered glaciers (Thomson *et al*., 2017). As such, the two values are not directly comparable for model calibration.

The order of magnitude of the discrepancy can be roughly quantified, using the land-terminating Pagels glacier (Fig. 1) as an example. Reported glacier-wide RSMB is -0.49 m w.e. yr^{-1} (Table 2, PDD model, Strategy C), over an area of 18.59 km^{2} (Table 2). The 2004-2019 area loss (as per the 2022 inventory: https://dga.mop.gob.cl/estudiospublicaciones/mapoteca/Documents/IPG2022.zip) is 0.67 km^{2}, in a region with strongly negative specific mass balance (Fig. 5). If the average SMB over the 2004-2019 deglacierized area is e.g. -6 m w.e. yr^{-1}, glacier-wide SMB (modeled over the 2004 area) could then be decomposed in the following area-weighted average:

$$-0.49 \cdot 18.59 = X \cdot (18.59 - 0.67) + (-6) \cdot 0.67$$

X being the average mass balance over the 2019 glacier extent.

The result is *X* = -0.28 m w.e. yr^{-1}, which is 0.21 m w.e. yr^{-1} less negative than the reported value of -0.49 m w.e. yr^{-1}.

The actual numbers will depend on the spatial distribution of specific mass balance and on the spatial patterns of glacier retreat, but clearly the mass balance discrepancy has the same order of magnitude as the reported RMSE values (Table 2). As with the model parameter choices, this can certainly affect the best parameter combinations which are computed by calibration. As such, the results of Table 2 – including the relative performance of models and calibration strategies on individual glaciers – may be inaccurate, and statements such as l. 409 ("further tuning is neither required nor justifiable") and l. 521 ("Going from a single-glacier calibration (Strategy A) to a regional calibration (Strategy B), only the TLR needs changing") may no longer hold true. The rough calculation shown above refers to a single glacier (Pagels); still, the argument is readily transferable to all glaciers in the MSM, which are undergoing rapid (but uneven) area changes at their termini.

In order to properly compare model output to geodetic mass change, the models should be run on up-to-date input grids for each year (Barandun *et al*., 2015). Alternatively, the CoMB could be computed from the RSMB with the methods of Elsberg *et al*. (2001), or in a post-processing stage as in Kronenberg *et al*. (2022).

We agree that there is a difference between the reference and conventional mass balance and that the outlines should be updated regularly as suggested. In order to produce as accurate results as possible, we improved the outlines from Barcaza et al. (2017) further and identified additional outlines for the year 2013. We will compute the SMB with updated outlines from 2004, 2013 and 2022. We will use the same updated outlines to calculate the specific geodetic MB (2000-2013) again over the average area.

Overall, we expect mainly small changes over the 2000-2013 period that is used for calibration based on the geodetic MB. Changes in 2013 outline are relatively small for most of the glaciers (see Figure 2). However, an impact on the calibration and skill ranking is likely. Changes to the 2022-outlines are larger for some glaciers, which might indeed influence the reported glacier-wide SMB results of the whole 22-year period significantly.

We will include these outline changes to the SMB in a post-processing stage as suggested, thank you.

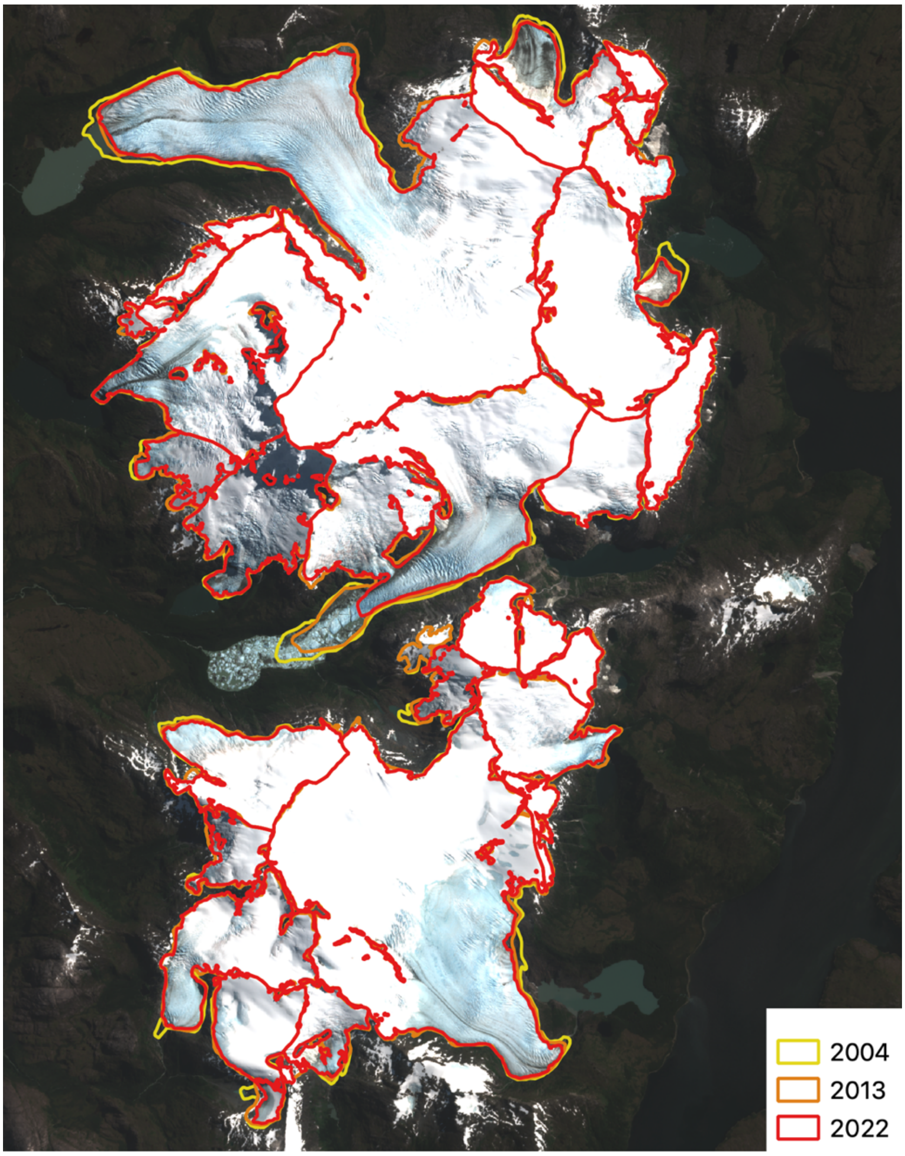

Figure 2: Glacier outlines for the years 2004, 2013 and 2022.

Major comment 3: geodetic data processing

I tried to reproduce the computed geodetic mass balances (Fig. 2), from the glacier outlines of Barcaza *et al*. (2017) and the grids of surface elevation change of Braun *et al.* (2019), downloaded respectively from https://dga.mop.gob.cl/estudiospublicaciones/mapoteca/Documents/Glaciares.zip and https://doi.pangaea.de/10.1594/PANGAEA.893611.

The elevation change grids contain patches of large absolute values near the edges of the glaciers (Fig. II), which are likely outliers and can significantly alter geodetic mass balance estimations. Moreover, large data voids are visible in the accumulation areas of several glaciers.

Indeed, recomputed geodetic mass balance (Fig. IIIa) does not match the result in Fig. 2 of the manuscript. Filtering out the 2nd and 98th percentiles of elevation changes (as mentioned by Braun *et al*., 2019) yields a closer result but not quite a match (Fig. IIIb); if anything, it shows that the study results can again be very different following relatively minor methodological choices. Since geodetic mass balances are a key input of the present study, it is important to detail all processing steps (filtering, gap-filling, etc.) applied to the initial datasets – possibly in an appendix or supplement.

Moreover, uncertainties in the geodetic mass balances (quickly mentioned at l. 174) must be shown, both per-glacier and for the entire study area (in Fig. 2 and/or Table 2).

We thank the reviewer for raising this point and we apology if it was not very clear so far. We have used part of the DEMs generated by the study of Braun et al. (2019). They estimated the geodetic mass balance for the entire Tierra del Fuego region using SAR DEMs from 2011 to 2015. For this study, we only selected the SAR DEMs that cover the Monte Sarmiento Massif in one ablation season (in this case 2013) (Fig. 3). It is worth mentioning, the data provided by Braun et al. (2019) (Pangaea dataset) are the un-filtered dh/dt fields, which means no further post-processed had been made. This is why mean values of the grid are different from Braun et al. (2019) and from our numbers.

The methodology employed in this study has been described by several authors (Malz et al., 2018; Braun et al., 2019; Farías-Barahona et al., 2020; Seehaus et al., 2019; Sommer et al., 2020). Nonetheless, in general terms, (1) TanDEM-X (TDX) DEMs are produced using SAR interferometry approach (see Braun et al., 2019). Once the DEM are generated, the (2) TDX DEMs need to be precisely horizontally and vertically coregistered to the respective reference DEM (i.e. SRTM) using stable areas. (3) Then the elevation changes differencing is estimated (un-filtering dh/dt fields). As the reviewer mentioned, there are some gaps in the elevation changes fields. In order to be filled, (4) we apply an elevation change versus altitude function by calculating the mean elevation change within 100 m height bins across the entire glacier area. (5) To avoid artificial biases introduced by outliers we do not include steep slopes (>50°) (Seehaus et al., 2019; Sommer et al., 2020) and filter each elevation band by applying a quantile filter (1%–99%). All these patches observed by the referee are therefore not included in the estimation. We will explain these processing steps in more details in a revised version of the manuscript.

Regarding the uncertainty estimations, we agree with the reviewer comments. We will include the uncertainty estimation for each glacier as well as for the entire massif using the below error propagation equation. The uncertainty estimation is in accordance with Braun et al. (2019) and Seehaus et al. (2019), in which the uncertainty estimations of the geodetic mass change ($\frac{M}{\Delta t}$) considered the following factors:

- Accuracy of the elevation change rates ($\delta_{\Delta h/\Delta t}$) (considering spatial autocorrelation and hypsometric gap filling)

- Accuracy of the glacier areas (δ_A) (for this study we will include the accuracy of the two glacier inventories)
- Uncertainty from volume to mass conversion using a fixed density (δ_ρ)
- Potential bias due to different SAR signal penetration ($\frac{V_{pen}}{\Delta t}$).

$$dM = \sqrt{\left(\frac{M}{\Delta t}\right)^2 * \left(\left(\left[\frac{\delta_{\Delta h/\Delta t}}{\frac{\Delta h}{\Delta t}}\right]^2 + \left[\frac{\delta_{A1}}{A1}\right]^2 + \left[\frac{\delta_{A2}}{A2}\right]^2 + \left[\frac{\delta_\rho}{\rho}\right]^2\right) + \left(\left(\frac{V_{pen}}{\Delta t}\right) * \rho\right)\right)}$$

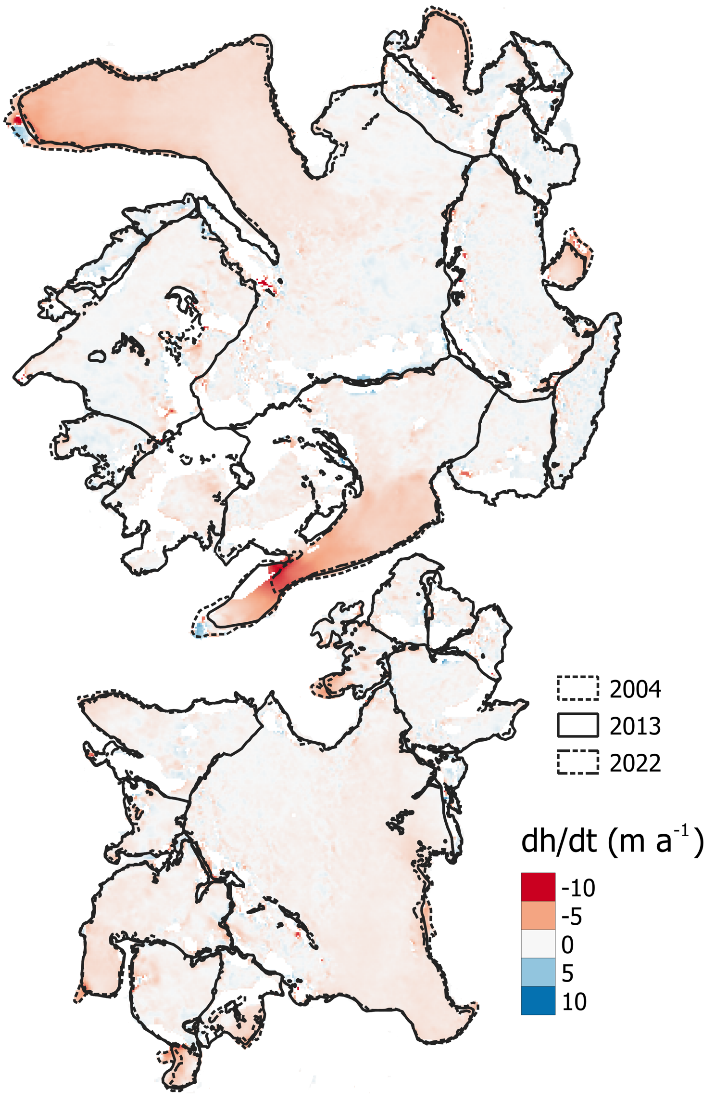

Figure 3: Elevation changes (m yr^{-1}) for the Monte Sarmiento Massif between 2000 and 2013 (unfiltered).

Minor comments

1. Presentation of mass balance models

Introduction and description of mass balance models should be improved. At ll. 64-70, the text needs to cover previous work on temperature index models, with more references than just Six *et al.* (2009) and Gabbi *et al.* (2014). Such models are mentioned here for the first time – not just in the Methods section; thus the relevant references should also appear here. Not all empirical models simply assume a linear relationship between temperature and melt rates (l. 65) – the most relevant variants and enhancements should be briefly mentioned. As the paper is about calibration strategies, it would be useful to also cite (and possibly compare in the discussion) other approaches at calibration of PDD models, like the use of snow line positions of Barandun *et al*. (2021). Calibration of full energy-balance models has also been extensively tackled in previous work, which should be appropriately referenced (e.g., van Pelt *et al*., 2012; Gilbert *et al*., 2014; Mattea *et al*., 2021; and references therein).

Thank you for this suggestion. The references of the used SMB models are given in l. 74-78. We will work on this section of the introduction to give more details on the four different SMB models used and on calibration approaches in literature.

2. Presentation of the input data

All data mentioned in Sect. 2 should be shown in greater detail. Specifically, an ablation stake network is mentioned – it should be displayed on a map (possibly an inset of Fig. 1). The same applies to the automatic ablation sensor and the location of ground-penetrating radar tracks. The final meteorological series is also a key input, as such it should be either made publicly available, or shown in a figure (possibly in the supplementary materials).

Thank you for your comment. We will give more details on the datasets and especially on the ablation stake network. For more details on this, see Major Comment 1, paragraph 4.

We will upload the model forcing and final SMB results in a public repository.

3. Description of the methods

The methods should be presented in enough detail to enable reproducibility of the study. Below, I list some instances where more information is needed.

- Numerical model setup: some information on the actual model setup is missing, such as the elevation grid used and the grid cell resolution. Did the Authors re-implement their own version of a PDD model? If yes, it would be good (for reproducibility) to make the code publicly accessible online. Moreover, does the time resolution listed for the PDD model (24 / 8 = 3 hours, l. 230) apply also to the other models used? COSIPY also has several parameters related to the vertical subsurface layers (l. 253) – were these left at their default values? Recent evidence indicates potentially large impacts of the numerical setup on computed melt amounts (Brun *et al*., 2022, and reviewer comments therein). For reproducibility and future comparisons, it would be beneficial to add a table (possibly in the supplementary material) of the main parameter values used in the models setup.

Thank you for pointing out this lacking information. We will include a table with the model setup and all parameters in the supplement.

- The accumulation model should be better explained. Specifically, how is precipitation partitioned into solid and liquid components? How are the AWS measurements used to “inform the statistical downscaling” (l. 143) of precipitation? In the orographic precipitation model, the “timescales of hydrometeors” should be briefly explained (since they are explicitly referred to). The sensitivity tests mentioned at l. 211 should be better explained – what is the “optimal relative humidity threshold”? Optimal in relation to what, according to which metric?

Thank you for your comment. We will describe the partition of precipitation in solid and liquid parts in the SMB model description (section 3.2). The orographic precipitation model is not an accumulation model. It calculates precipitation. The partition in solid and liquid component is done in the SMB models. The PDD and SEB models distinguish between solid and liquid precipitation at a hard temperature threshold of 1.0°C. COSIPY uses a logistic transfer function snowfall from precipitation scaling around a threshold temperature of 1.0 °C.

The total precipitation is calculated adding the large-scale precipitation (without the orographic part) given from ERA5 data to the orographic precipitation calculated in the model. We use the annual precipitation amounts from AWS Rock to constrain the relative humidity threshold (90%) above which orographic precipitation can occur. This way, we guarantee that the annual total precipitation at the AWS location agrees with the observed amounts. We will extend the explanation of the OPM and the configuration in a reworked version of the manuscript.

See also Major Comment 1, paragraph 2.

- The snowdrift model described (Eq. 6) does not match the cited Warscher *et al*. (2013, Eq. 10) – there is an additional factor *U* giving linear dependence of accumulation on wind speed. If this is indeed the case, the change is major and should be explained.

We will stress the fact that we added a small modification to the parametrization of Warscher et al. (2013) more clearly and give a more detailed explanation. The reason, why we added the velocity here is that we observe different wind directions with different velocities. The linear dependence of snowdrift on wind velocity is indeed highly simplified, but so is the entire snowdrift scheme.

- I could not find which glaciers exactly contribute to the B_{MSMnc} (massif-wide mass balance used for calibration). Are these all the glaciers of Table 2 except all the lake terminating ones? It should be made more clear in the table caption.

Thank you for pointing out this lack of information. We will include this information in the manuscript. The B_{MSMnc} comprises all glaciers > 3km^2 that have no significant calving losses. The only glacier with significant calving losses is Lovisato Glacier, which is, thus, the only lake-terminating glacier excluded here. We decided to include the other lake-terminating glaciers because calving losses are negligible for those, and we would otherwise lose a large part of the glacierized area for calibration. This way, we include 71% of the glacierized area.

- At ll. 319-320, it should be made clear how the Authors "defin[e] the regional massif- wide amount of accumulation". Is it simply the output of the Orographic Precipitation Model, partitioned into solid and liquid precipitation according to local air temperature?

Thank you for your question. The massif-wide amount of accumulation is the sum of snowfall over the massif, which is the solid part of the output of the OPM according to local air temperature. What we wanted to say with this sentence is, that we first determine the total amount accumulation and ablation over the massif in Strategy B, and subsequently optimize the snowfall distribution with the snowdrift model in Strategy C. We will reformulate the sentence to make it clearer.

- At l. 437, the Authors claim that snow line altitudes from satellite observations support their computed spatial patterns of Equilibrium Line Altitude (ELA). While I believe the Authors, I still suggest to either remove the statement or show supporting evidence.

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.

- At l. 515, it is not clear how the Authors "calculate a rough estimate of τ from ERA5 data". The method should be described (possibly in the supplementary material), or a reference should be provided.

To assess the rough range of τ, we calculated it according to the following equation from Jiang and Smith 2003, J. Atmos. Sci.:

$$\tau_f = (\boldsymbol{H_w} + \boldsymbol{H_b})/\boldsymbol{2V}$$

With the water vapor depth $\boldsymbol{H_w}$, the cloud base height $\boldsymbol{H_b}$ and the mass-weighted average falling speed $\boldsymbol{V}$, which is suggested to vary between 1 and 2 m s^{-1}. Assuming τ_c = τ_f = τ, we can estimate τ. This calculation is performed with the upstream ERA5 data as used in the OPM model. However, due to the wide spread of $\boldsymbol{V}$, this method only provided us with a rough estimate of τ = 1050 ± 350 s.

We will add the above reference, thank you.

4. Quantification

Throughout the manuscript, several statements should receive quantitative support. Some examples:

- l. 436, “Equilibrium line altitudes tend to be lower in the east of the massif” – by how much, and what is the spread? The ELA is one of the fundamental quantities in mass balance studies, and its spatial patterns are certainly of interest for comparisons and future studies.

See comment above (Minor 3, l. 437).

- l. 494, “the differences between both models are overall minor” – it would be good to mention here the relevant values from Table 2, such as the global mass balance and RMSE differences.

We will give the demanded values from Table 2 in the text as suggested, thank you.

- l. 644, “surface velocities of around 402 m yr^{-1}” – 402 is quite a specific number, which suggests an uncertainty (and/or variability) affecting only the units place, all across the glacier calving front. Is this the case? If not, could the Authors provide an estimation of the spatio-temporal variability and uncertainty of the values? Else, the number should be given as an order of magnitude only.

Thank you for this comment. We will change the exact values to order of magnitude only.

- ll. 653-654, “the uncertainty in the observed elevation change rate is large […] we assume an increased uncertainty [...]” – the Authors mention estimating these uncertainties (l. 174); the numbers should be provided, to support the given explanation of the mass balance discrepancy (is the uncertainty at glacier 138 20 % times larger than for the other glaciers? Or 100 times larger?).

We will provide the numbers in a revised version of the manuscript, thank you for the comment. The uncertainty is given mainly by accuracy of the inventories (30m resolution) and the voids in the dh/dt field in the upper part of this small glacier.

5. Benefit of increasing the complexity level

Research question Q3 (l. 99) states: "Can the performance of the SMB model be improved by increasing the complexity level regarding included processes?". The inclusion of a snowdrift module is indeed shown to reduce the overall model error. But at the same time, the addition of a physical model for incoming radiation (SEB_G, l. 244) also represents an increase in the complexity level; and the Authors observe (l. 495) that it does not improve the performance of the SMB model. As such, the conclusion at l. 694 should be revised to reflect these contrasting results.

Thank you for this comment. We will include the lack of improvement by increasing the complexity level of the models in this paragraph.

Technical comments

- ll. 31 and 53: "2000-2011/14" is not fully clear, please explain the date range.

  The data range stems from the satellite images that were chosen over a period from 2011-2014 for the analysis. See the reference given in the text for more details.

- l. 33: I suggest adding *in situ* to "scarce observations of glacier MB", as remote sensing observations appear to be plentiful.

  Changed as suggested, thank you.

- l. 49: please specify the time range of the Little Ice Age – is it the same period as commonly understood in the European Alps?

  The Little Ice Age spans roughly the same period as commonly understood. Maximum advances in southern Patagonia have been observed in the 16th to 19th century. We will add this information to the sentence.

- ll. 53-54: the two estimates of annual thinning rates appear to be in stark contrast. It would be useful to mention whether they have been reconciled, or they refer to different areas, or the more recent study has superseded the previous results.

  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.

- ll. 58-63 are a description of the study site, partially repeated from line 107 in section "Study site and data".

We will shorten the paragraph about the study site in the introduction, thank you.

- ll. 391 and 395: if I understand correctly, mass balance in Strategy B is calibrated solely to the regional value (l. 385). As such, it is not surprising that the value of B_{MSMnc} is reproduced perfectly and the bias is no longer discernible – it is the only expected outcome of a successful single-target calibration. If that is the case, I would then suggest rephrasing these sentences.

  Thank you for this comment. We agree that the perfect agreement of the B_{MSMnc} is not surprising here. We will rephrase the sentence.

- l. 411: this is a methodological choice which should be mentioned already in the methods.

  Thank you. We do explain this methodological choice in l. 321-326 and Figure 3. We will revise the paragraph to make it clearer.

- l. 442: summer and winter together amount to 46 % of snow accumulation – then at least one other season should contribute the single largest amount over the year. Could the Authors please provide some information on the occurrence of the other 54% of snowfall?

  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 clear that the largest part of snowfall (65%) is accumulated in these two seasons, whereas only a small part (13%) accumulates in summer.

- l. 456: this appears to be an exact repetition of l. 352.

  We will reformulate the whole section about uncertainty quantification. See response to Major Comment 1, last paragraph.

- ll. 467-468: it is not immediately clear what is a negative range of uncertainty, please explain.

  We will reformulate the whole section about uncertainty quantification. See response to Major Comment 1, last paragraph.

- Table 3: here the BIAS (mean signed difference) should be shown alongside the RMSE. Moreover, the simple (unweighted) arithmetic average of RMSE at multiple stakes and at one automatic ablation sensor does not appear to be a very relevant metric.

  Thank you for this comment. We will remove the average value and include the bias instead.

- l. 563: winter ablation is mentioned here (in the Discussion) for the first time. Its quantification is a result and should appear already in the corresponding section if it is to be compared to previous studies.

  We apologize for the confusion and thank you for this comment. It is supposed to state "winter accumulation".

- ll. 595-604: these are objective results, I would recommend moving them to Sect. 4.

  We will reformulate the whole section about uncertainty quantification. See response to Major Comment 1, last paragraph.

- l. 613: could the Authors formulate here a hypothesis as to why the exclusion of Schiaparelli Glacier from the results significantly alters the relative performance of the models? This would be beneficial for a deeper understanding of the models intercomparison and applicability to other geographic settings.

  Schiaparelli Glacier is the largest glacier in the area, thus, it has a significant impact on the area-weighted RMSEs. The agreement with observations at Schiaparelli Glacier is overall not too good. But it seems that with COSIPY the agreement at Schiaparelli Glacier cut along the fluxgate (Schiaparelli_FG) is better (also seen in Table 2). Thus, including it decreases the RMSE for COSIPY and increases it for the other models, which is changing the ranking.

  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 paragraph, thank you.

- l. 701: the PDD approach is by now well established and known to produce robust results, “surprisingly good” may not be the best wording here.

  We will rephrase the sentence, thank you.

- Fig. S1a/b/c: add white crosses as in Fig. S1d/e.

  Will be added.

- Fig. S1e: it is quite hard to compare the different values. I would suggest placing the two *DDF* values on different axes, to see if a more readable (smoother) result can be achieved.

  We will try the suggested changes, thank you.

- Fig. S5: the Y axis is likely wrongly labeled – ablation rates are too low compared to e.g. Table 3.

  The y-axis is labeled correctly. Please note that these are the absolute, measured values in m w.e. over a certain time period (given on the x-axis) (same for Fig. S4) whereas in Table 3 we provide values in m w.e. yr^{-1}.

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.

Braun, M., Malz, P., Sommer, C., Farias, D., Sauter, T., Casassa, G., Soruco, A., Skvarca, P., and Seehaus, T.: Constraining glacier elevation and mass changes in South America, Nat. Clim. Change, 9, 130–136, https://doi.org/10.1038/s41558-018-0375-7, 2019.

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.

Farías-Barahona, D., Sommer, C., Sauter, T., Bannister, D., Seehaus, T., Malz, P., Casassa, G., Mayewski, P.A., Turton, J.V., Braun, M. Detailed quantification of glacier elevation and mass changes in South Georgia. Environmental Research Letters 15, 034036. https://doi.org/10.1088/1748-9326/ab6b32, 2020.

Jiang, Q. and Smith, R. B.: Cloud timescales and orographic precipitation, J Atmos Sci, 60, 1543–1559, https://doi.org/10.1175/2995.1, 2003.

Malz, P., Meier, W., Casassa, G., Jaña, R., Skvarca, P., and Braun, M. H.: Elevation and Mass Changes of the Southern Patagonia Icefield Derived from TanDEM-X and SRTM Data, Remote Sensing, 10, 188, https://doi.org/10.3390/rs10020188, 2018.

Oerlemans, J. and Knap, W. H.: A 1 year record of global radiation and albedo in the ablation zone of Morteratschgletscher, Switzerland, Journal of Glaciology, 44, 231–238, https://doi.org/10.1017/S0022143000002574, 1998.

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.

Sommer, C., Seehaus, T., Glazovsky, A., and Braun, M. H.: Brief communication: Increased glacier mass loss in the Russian High Arctic (2010–2017), The Cryosphere, 16, 35–42, https://doi.org/10.5194/tc-16-35-2022, 2022.

Seehaus, T., Malz, P., Sommer, C., Lippl, S., Cochachin, A., and Braun, M.: Changes of the tropical glaciers throughout Peru between 2000 and 2016 – mass balance and area fluctuations, The Cryosphere, 13, 2537–2556, https://doi.org/10.5194/tc-13-2537-2019, 2019.

van Pelt, W. J. J., Oerlemans, J., Reijmer, C. H., Pohjola, V. A., Pettersson, R., and van Angelen, J. H.: Simulating melt, runoff and refreezing on Nordenskiöldbreen, Svalbard, using a coupled snow and energy balance model, The Cryosphere, 6, 641–659, https://doi.org/10.5194/tc-6-641-2012, 2012.

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.