Review of "Strategies for Regional Modelling of Surface Mass Balance at the Monte Sarmiento Massif, Tierra del Fuego"

By Temme et al.

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.

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

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".

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.

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.

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.

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.

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.

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.

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)?

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?

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?

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.

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

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.

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

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

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.

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

L369: what does a "profound" estimate mean?

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

L376: unclear how this "transfer" is done.

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.

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.

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.

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

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.

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.

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.

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

L434: what’s the difference between “huge” and “very huge”. Suggest deleting “very”.

L437: cite study or show in supplemental figure?

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?

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.

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.

L446: “22-year”

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.

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.

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.

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?

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.

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

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

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.

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).

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

L594: “is strongly dependent” perhaps?

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

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

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

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.

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

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

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

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

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.

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.

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?