Review of Menthon et al. "Comparison of calibration methods of a PICO basal ice shelf melt module implemented in the GRISLI v2.0 ice sheet model"

Reviewer: Xylar Asay-Davis

I wish my name to be relayed to the authors, as I feel I am always a better reviewer when I am not anonymous and I encourage others to consider reviewing non-anonymously whenever they feel able.

General Comments:

This paper presents the implementation of the PICO parameterization for Antarctic basal melting in the GRISLI v2.0 ice sheet model. Then, the authors analyze a series of potential methods for parameter calibration, settling on a binning technique together with the mean absolute error (MAE) as the most robust approach. They probe the robustness of the approach in several ways, including modifying the ocean forcing, changing melt-rate dataset use as the observational target, calibrating regionally rather than Antarctica-wide, and using different ice-sheet resolutions (also the resolution of the parameterization). A significant finding is that their calibration method requires no correction of the input ocean temperature field, in contrast to previous work. In addition to the calibration, they ran an ensemble of projection simulations to the year 2300, showing the impacts on total melt flux, floating ice area, and Antarctic mass change from different PICO calibrations.

The paper is quite detailed and for a fairly niche audience. Even so, I think the level of detail is justified because the analysis performed here is potentially applicable not just to the PICO parameterization but to all Antarctic basal melt parameterizations. The results present a strong case that the binning technique used for calibration leads to parameter choices that are robust to the various uncertain inputs to the parameterization and which represent the statistics if not the detailed spatial patterns of Antarctic basal melting.

I see a potentially significant issue in the equation presented for the pressure at the ice shelf-ocean interface that could impact the results. As I expand on in my specific comments below, the pressure as presented is computed from the ice density but the ice draft (the ocean thickness, so to speak), whereas the correct pressure must involve either the ocean density and the ice draft or the ice density and full ice-shelf thickness. This could amount to a 10% error in PICO melt rates and affect the analysis in non-trivial ways. If the mistake is just a typo, this would be easy to fix. If it is an error in the implementation, the magnitude of the impact will need to be assessed and addressed in the paper. Aside from that one issue, there are a few other small technical issues in the paper that I think can easily be addressed, and which I point out in more detail below.

The manuscript is well written and the figures are clear and well formatted. I have made several suggestions about possible improvements in formatting and grammar below that I think might improve the manuscript. For the most part, I found the organization of the paper to be quite good. The minor exception is that I found that parts of the Discussion section read as if they belong more in the Results, as I will expand on below. And I felt like the Discussion section could dig a bit more into the implications of the choice in the GRISLI implementation of PICO to treat each IMBIE basin as if it were a single ice shelf, rather than using the original PICO design of treating ice shelves individually. The text mentions that this choice is made for computational efficiency but doesn't go into the potential this has for homogenizing melt patterns across regions with a lot of spatial variability between ice shelves (I'm thinking of the Bellingshausen and Amundsen Seas in particular, where both spatial variability in ocean temperature and in bathymetry depth at calving fronts likely relate to the spatial variability in melting).

My hope is that correcting the calculation of pressure at the ice draft and the reorganization I suggested to the Results and Discussion amount to relatively minor revisions.

Specific Comments:

l. 3 and 29 "As a result of limited understanding of these ice-ocean interactions…" and "With our current understanding and computational resources…" My view is that Antarctic ice shelf-ocean interactions are parameterized primarily because of the limited computational resources you point out and also because of difficulties of a more technical (rather than scientific) nature in implementing ice sheet-ocean coupling. A tertiary reason might be the biases in the ocean components of Earth system models that can be corrected as part of applying a parameterization. While I won't deny that there are things we don't understand about ice shelf-ocean interactions, I don't think it is due to a lack of (scientific) understanding that we are using parameterizations. If "limited understanding" was meant to encompass "limited technical understanding", I think it would be important to lean into that a bit more explicitly because I read it as implying "limited scientific understanding" without further clarification.

l. 13 "...we can avoid using ocean temperature bias corrections." That's really great! As I said, above, this is a big advantage of your approach!

l. 69: "$p_k = \rho_{Ice} * g * z_{IceDraft}$": I'm afraid this is my main concern in the paper. This equation need to be either one of these two equivalent forms:

$$p_k = \rho_w \, g \, z_{IceDraft}$$
$$p_k = \rho_I \, g \, h_i$$

where h_i is the full ice thickness of the ice shelf. The form you have here will be too low by a factor of ρ_i / ρ_w or about 0.89 (the value for this ratio given in the text). The error in pressure of about 10% will likely propagate to an error of about 10% in the thermal forcing and therefore the melting from PICO, since the pressure term is a dominant one in the linearization of the freezing point. Please check your implementation to make sure you don't have this error, and correct the text if this is just a typo. If you have implemented this incorrectly, you will need to see what the implications are by rerunning at least a small subset of the calibration with the correction and seeing what the impacts are. This seems large enough that it could change the optimal parameter values.

l. 83: "we decide to make use of the drainage basins defined in (Mouginot and Rignot, 2017) and used in (Rignot et al., 2019)": Please change "make use of" to something more specific. It seems that you treat all ice shelves in each basin as if they were a single ice shelf in PICO.

l. 82-89: Each basin uses the same number of boxes, then, even if it contains some large and some small ice shelves? Did this cause any issues for small shelves with many boxes? I think we found that this gave us trouble in the original PICO and that was the reason for different numbers of boxes for large vs. small ice shelves. I believe we were prone to unrealistic levels of freezing or oscillating thermal forcing between boxes or both. But maybe aggregating the shelves so they act as one makes the difference or perhaps a different calibration is responsible for not seeing this behavior.

l. 105-108: It wasn't clear in this part of the paragraph whether what you describe applies to GRISLI-PICO or PISM-PICO (or both). Could you try to clarify that a bit better?

l. 109: "...outside of the ice shelves in the open ocean…": I don't understand why any ice-sheet model calculation is needed in the open ocean. Shouldn't the growth of ice shelves toward the open ocean be limited by calving, not by parametrized melt? I think more explanation is needed for why you need this additional melt parameterization. It is also unclear where the ocean temperature T_o is being taken from.

l. 120-122: Can you explain how you decided on the upper and lower bounds for these parameters? The results indicate that it's a good range but it would still be helpful to know how you decided what was a reasonable range since too broad a range leads to imprecision in the final calibrated values, while too narrow a range could meant the true optimum is outside the bounds you searched.

l. 124: "...with a 30 years relaxation with GRISLI." Could you say more about how you did the 30-year relaxation with GRISLI? How were the surface and basal fluxes determined, for example?

l. 142-143: "...are up-scaled to the same resolution as the ice sheet model.": So far, you haven't said what that resolution is. Since you say "upscaled", it might be worth stating that that's to either 16 or 40 km resolution.

Also, what method is used for upscaling? Bilinear interpolation? Conservative remapping?

l. 151 and 153: "pixel-to-pixel": elsewhere in the text, you refer to "points" (e.g. l. 102) or "data points" (e.g. Table 1) rather than "pixels". I think "points" is better than "pixels", since this is not really image data. But I think the best terminology might be "cells", "grid cells" or "model grid cells" to make it clear that the cells of the ice-sheet model grid are what are being referred to in each case. Here for sure, I would use "cell-to-cell" or (less preferable) "point-to-point" rather than "pixel-to-pixel". Elsewhere, I do think "cell" would be better than "point" or "data point" but I'll leave that up to you.

l. 155-156: I am familiar with these norms being called the Euclidean and Absolute-value error norms, respectively, or also the L-2 and L-1 norms, respectively. The terms (MAE and RMSE) you use are fine, too. But since they are pretty broadly used in statistical and numerical methods, I don't think you need this sentence. The reader can see the difference between them in the table.

l. 156-157: This is great! I appreciate you providing this context here, because it dovetails well with your findings later on that the MAE is a more robust choice than RMSE.

Table 1: “data points”: Again, I think it’s worth clarifying that these are GRISLI model grid cells, since “data points” is a somewhat vague term.

l. 158-159: “...we do not manage to constrain the method to pick systematically the ensemble members with the best fit to the target.”: This phrasing is a little awkward but more importantly I don’t think it’s clear what you mean by “best fit to the target”. It seems like you always find the best fit to the target according to the given metric. I think you need to describe more precisely here what these metrics fail to achieve.

l. 163-164: “...the 20 points with the lowest values…”: Should this be “lowest melt-rate values” for clarity?

l. 212-213: “The first of the three, the ADA of bins (panels (g) and (j)), selects almost the same members as the ADA without binning.”: It seems to me like the ADA method with and without binning should be (almost) mathematically identical. The average of the cells should be the same as the average over the bins. The only explanation I can come up with for why they might be different is that some bins contain slightly different numbers of grid cells than others, meaning that they cover slightly different areas but still get exactly the same weighting in the binning approach. But can you explain why you think the ADA method with and without binning are not the same? Please include this in the text, not just in your response to my review.

This has implications for your later analysis where you often treat the two ADA results as functionally identical because they choose the same ensemble member. It might be cleaner to just state that the ADA method with binning is so close to the ADA method without binning—any differences are due to poor statistical sampling in each bin—that it isn’t worth exploring as a separate approach, and drop it from the rest of the paper.

l. 226: “Nonetheless, being able to match the distribution can also mean spatial compensation between different locations.” This spot on, and get at the heart of why many previous calibration methods have not done a great job!

l. 248: “...despite the 169 members of the ensemble…” I think I understand why you bring it up but I do not think the number of ensemble members is a particularly relevant factor here. The only reason the number of ensemble members would play a role is if you had badly undersampled the parameter space and missed a region with a local minimum entirely. The large number of parameter values makes it less likely that you did this, but the difficulty in finding a good set of parameters for BA compared with FRIS is largely for the reasons you get to later in the paragraph. I would leave this phrase out or explain better why it is relevant.

l. 249-250: “This can be explained by the difference in the number of data points.”: I doubt very much that this is correct. The smaller number of data points could make the value of the error metric noisier as a function of parameter values but it should not be related to the steep gradient in the error metric with respect to parameter values. That is simply related to the fact that the melt rates are a much stronger function of the parameters in this region as you point out below. I feel strongly that this sentence should be removed or it should be rephrased to indicate that the smaller number of data points means the error metric will be noisier (but that it is not related to why the metric is a strong function of parameter values).

l. 254-255: Yes! I agree that this is the main reason for the strong gradients in the error metric: at thermal forcing values, the melt rates are much more sensitive to the parameters.

l. 259-261: “The seven cases correspond to the six best members according to the six methods applied Antarctic-wide. It corresponds to five members since the best ADA and best ADA of bins give the same best member.” I found this wording quite confusing. Would this be a correct restatement?

> “Five of the seven cases correspond to the best members according to the six methods applied Antarctic-wide. We get five cases rather than six because ADA and ADA of bins give the same best member.”

l. 274-275: “The calibration 34, considered as the best one in the calibration process, show very steady ice floating area with a small decreasing trend.”: First, I would suggest avoiding subjective words like “very”. But the bigger issue here is that the sentence as written seems contradictory. Either the floating area is (very) steady or it has a decreasing trend, and I would say more the latter. How about something like this? “The calibration 34, considered as the best one in the calibration process, shows a small decreasing trend in floating ice area over the course of the simulation compared with most other cases.” If you want to say that the trend is small compared to the overall floating area, that would be fine, too. But it should be small with respect to something.

l. 279: It is important to state what the sea level contribution is “low” with respect to. In this case, I think you mean that it is low compared with the other cases you are plotting.

Discussion section: As stated in my general comments, I think this section contains several subsections that would fit better under results. These are 4.1, 4.2 and parts of 4.3 (which part is discussed further below). Each of these sections introduces new calibrations together with error-metric plots similar to those in Section 3. I think these subsections should simply continue section 3 as 3.3, 3.4 and 3.5.

l. 285: When I look at the resolution in plot S57, it looks like Pine Island and Thwaites ice shelves cover only 2 grid cells each (at least in the Paolo data set, not sure about in GRISLI itself). What are the implications of this for PICO? It seems like 2 grid cells for a given ice shelf isn’t enough to provide meaningful results? Do you think the parameterization becomes meaningful when you aggregate across basins at this resolution? While I think this subsection should be moved to results, I think these questions could provide the basis for discussion here, after my suggested reorganization.

Figure 8: Why not just include Figure A2 in the text. As far as I can tell, Figure A2 has extra information beyond Figure 8 and there is nothing in Figure 8 that is not in Figure A2. Did you feel that Figure A2 was too complex?

Section 4.3: I would move lines 316 to 326 to the results section for sure. By line 329, you are firmly back in discussion territory. I would probably argue for keeping lines 326 to 329 in the discussion section as well, but that's less clear to me. In any case, this is far beyond my prerogative as a reviewer. I would merely advise that some of the material in this section move to a Results subsection and some remain here as discussion.

l. 326-327 and Figure 10 legends: "The average difference between the two datasets is about 25% of the average value of Adusumilli et al. (2020) at 16 km.": I think the average of the difference is not the right metric to use here. Even if the average of the difference happened to be quite small (and I think that maybe should be the case here), that would not tell you very much about how different the datasets are. Instead, you should use the standard deviation of the difference here. That is a good measure of how different the data sets are from one another. Since that is about triple the mean melt rate (in either data set), it tells you there's a **lot** of disagreement.

l. 329: "...by using the binning methods **we give spatial freedom to the datasets** and constrain them by their values." I think you need to state more clearly what "spatial freedom to the datasets" might mean. I also think it would be good to flesh out the discussion here even more about why you think it's important to get a good statistical representation of melt rates but why the details of the spatial pattern don't matter. More below…

l. 329-331: This is a great argument!

I think it would also be worth pointing out that the Joughin finding isn't a universally accepted result in the field, and instead some researchers (Gudmundsson 2013, doi:10.5194/tc-7-647-2013; Reese et al. 2017 https://doi.org/10.1038/s41558-017-0020-x for two) think that buttressing plays an important role and melt in more strongly buttressed areas has an outsized effect.

Figure 10: In the caption, could you make clear that these results are for the 16 km GRISLI mesh? It wasn't clear to me until I looked at the supplementary material.

Also, could you help me understand why the difference between the means in panel a would be about 0.02 m/yr but the mean of the difference is 0.16 m/yr? My guess is that there are many cells where only one of the datasets has data, and that accounts for the difference but wanted to make sure. Otherwise, I would have expected the difference of the means to be the mean of the differences.

Section 4.4: This is the first section within the discussion that I think fully belongs here. It is full of some really great points.

l. 350-362: Be careful that you present these differences in a neutral tone. You can't know for sure whether the PICO or the quadratic nonlinear results are more accurate since we don't know the right answer for future response. The only room there is to criticize one parameterization in relation to another is based on the physics each includes or excludes.

l. 350-351: “PICO includes overturning circulation under the ice shelves, which tends to reduce the basal melt rate. This freshwater negative feedback is not included in the QuadNL parameterisation.” I think this is a misunderstanding of the quadratic parameterization. Like PICO, the quadratic parameterization has one term involving the thermal forcing that is meant to represent how warmer temperatures near the ice-ocean interface lead to higher melt rates. But the second term in the product involving the thermal forcing is meant to represent the stronger overturning that results from warmer conditions. This is essentially equivalent to the PICO term involving the overturning coefficient C. See, for example, Holland et al. (2008, doi:10.1175/2007JCLI1909.1), Jourdain et al. (2017, doi:10.1002/2016JC012509), Burgard et al. (2022, doi:10.5194/tc-2022-32). So I do not agree with the statement on this line and think it needs to be heavily revised or removed.

l. 352-359: These are 3 excellent points!

l. 360-362: I really like the point about sensitivity to oceanic forcing in QuadNL. But I take issue with “therefore this QuadNL calibration could overestimate basal melt rates”. I don’t think this follows. The lack of refreezing certainly will mean that, if the method is calibrated to get the right mean melt, it will also underestimate melt maxima. But it doesn’t follow that it will overestimate melt on the whole, this depends on the detail of the calibration approach. I would remove this phrase.

l. 363-364: I would think this would be fairly obvious to anybody doing ice sheet modeling. There is not necessarily any relationship between the area of floating ice (the ice shelves) and grounded ice, which is the only part that contributes to sea level. Even the relationship between the grounded area and the volume (or mass) above flotation can be complicated. But I have not generally seen the area of floating ice presented as a quantity of major dynamical significance for ice sheet models in the past. (But I will add that ice-sheet modeling is less my field.)

l. 364-366: I’m afraid I don't quite follow this argument. It’s not obvious to me why different time series of ice-shelf area would imply different pathways. Ice shelf area could be reduced along the same pathway just by the presence of more (or less) calving relative to grounding-line motion.

l. 368-369: “This methodology can potentially be applied to modules in other models that benefit from existing observational datasets.” This is an excellent point! I think this work has exciting potential applications to other methods and models.

l. 379: “...spatial contrast…”: I find this phrase a little inexact. How about “range of special variability, if not the details of spatial patterns”?

l. 390-391: This is an excellent point! The paleo record is one good way to constrain this work over longer times. Perhaps high fidelity coupled ice sheet-ocean simulations will provide another such method in the future.

l. 398-399: I appreciate you providing this recommendation! I think many modelers will find it useful.

l. 401-402: “But also, as the present study has the limitation of targeting a given basal melt rate for a given ocean temperature rather than the sensitivity of the basal melt rate to changes of ocean temperatures.” I think this is a great point but it isn’t clear how it follows from the previous sentence. How will more modelers testing the calibration method address this limitation?

l. 404-406: “Alternatively, the low sensitivity of PICO in comparison to QuadNL could also be adjusted by developing a quadratic dependence to thermal forcing that would give a quadratic parameterisation that also accounts for overturning circulation under the ice shelves.” Presumably, it will come as no surprise given my statements above that I do not agree with this statement. I believe the QuadNL parameterization does already take the overturning into account.

Typographical, Grammatical and Formatting Suggestions:

l. 22: “Sub-surface melt of the ice shelves **on the other hand**…” It’s not clear what “on the other hand” is contrasting here. Maybe something else like “in turn” is more appropriate?

l. 39: “...distribution of values from the observations **as best as possible.**” The phrase “as best as possible” is often used informally but probably should be replaced with something like “to the extent possible” in scientific writing.

l. 45: “...whether it matters to calibrate PICO at a smaller scale than Antarctic wide…” This took me a little time to understand. Maybe rephrase this for clarity as something like “...whether regional calibrations of PICO produce different results than those applied over all of Antarctica…”

l. 65 and throughout the subsequent text and figures: “$Sv.m^3.kg^{-1}$”: You consistently use periods to separate units. I imagine this is something the typesetter will handle but my experience with TC is that they use half-spaces (`\,` in latex) between symbols. I have seen a multiplication dot (`\cdot` in latex) used occasionally in other journals but never a period. I would advise changing to a half space.

l. 68: “under-burden pressure”: This may just relate to our difference in point of view but I had not heard of this term used in this context before. I guess from an ice-sheet modeling perspective, this is the pressure of the ocean pushing up on the ice? From my ocean modeling perspective, this is the overburden pressure of the weight of the ice pressing down on the ocean. Maybe just call it the “pressure at the ice-ocean interface” or something similar to be agnostic to these two perspectives?

l. 69: “$p_k = \rho_{Ice} * g * z_{IceDraft}$”: I would suggest using multiplication dots (`\cdot` in latex) here or just spaces but “*” is not the right thing to use in scientific writing. For consistency with what you have on l. 74, ρ_{Ice} should be ρ_I.

l. 75-77: “The coefficient a is the salinity coefficient of the freezing equation…pressure coefficient of the freezing equation and equals $7.77 \times 10^{-8} \circ C\ Pa^{-1}$.” This simplified equation is typically referred to as a linearization of the “equation of state for the freezing point of seawater". This might be better than "freezing equation." The coefficients are often called the liquidus slope, liquidus intercept and liquidus pressure coefficient, respectively (see, for example, Asay-Davis et al. (2016, doi:10.5194/gmd-9-2471-2016) – no need to cite this, just pointing to the terminology there).

l. 83: "...we decide to make use of the drainage basins defined in (Mouginot and Rignot, 2017) and used in (Rignot et al., 2019)": These citations should be outside of parentheses (`\citet` instead of `\citet` in latex).

l. 95: "All the four criteria are here not followed…": This would be significantly clearer as something like, "Here, we do not follow any of these criteria…"

l. 98: "...show in some areas refreezing…" should be "...show refreezing in some areas…"

l. 102-103: "...we consider as grounding line any ice points that is surrounded by grounded ice and not grounded ice…": I found this a little hard to follow and I think it would be clearer as something along the lines of, "...we consider any ice points to be on the grounding line if it has some neighbors that are grounded and others that are floating…"

l 108, 113, 210: "...enables to have…" and "...enables to limit..." should be something like "...enables us to have…" or "...enables one to have…" (less preferable). This particular passive form is not grammatically correct in English.

l. 124: "...30 years relaxation…" should be "...30-year relaxation…"

l. 146 and 292; "We present first the three ones…" and "...the methods RMSE of bins and MAE of bins are the two ones..": It would be more appropriate for scientific writing to replace "three ones" and "two ones" with "three methods" and "two methods", respectively. (I know it sounds redundant, but there doesn't seem to be a good way to avoid that.)

l. 151 and 153: "two Dimensions…" should be "two-dimensional"

l. 159: "...we decide to proceed a binning on the dataset…": I'm not entirely sure what you are saying here but maybe, "...we decide to bin each dataset…"

l. 160 "method" should be "the method"

l. 161: Here and later in the text, I believe the standard is to use a capital F in "Figure" (and T in "Table") even within a sentence when referring to figures by number. I think you do this later in the text in a few places, but mostly you use lowercase f, and I think you should use capital F throughout.

l. 176-177: "For that…" isn't super clear or standard grammar. How about, "to accomplish this…"

l. 183: "...in line with…": This is super picky but I think this should just be "...using…" because "in line with" implies that the results are consistent with what you're about to refer to, not that they were produced using what you're about to refer to.

l. 198, 200, 266, 271, 291 and elsewhere: “on panel” or “on figure” should always be “in panel” or “in figure”.

l. 207: “...will not be systematically in the same order of magnitude of values…” might be easier to follow as something like “...will not systematically be of the same order of magnitude…”

Figure 3: I would put the panel letters before the text they refer to, not after: “...to each methods, (a) ADA of bins, (b) RMSE of bins and (c) MAE of bins…” This is more standard notation.

l. 229: “...they even do not have values…” should be “...they do not even have values…”

l. 235: “...the possibility to compensate…” should be “...the possibility that there is compensation…”

l. 279: “result all” should be “all result”

l. 283: “...discuss further about the ISMIP 2300 results obtained using PICO” should be “...discuss the ISMIP 2300 results obtained with PICO further.”

l. 290, 291, 313: “cumulating”: Although “to cumulate” appears to be a valid English verb, as a native speaker I had never heard of it before. I would suggest using "accumulate," "aggregate," or "combine" instead.

l. 294-296: This could be cleaned up a little as: “This consistency can matter to inter-compare: i) different parameterisations ; ii) when the same parameterisation **is** used in different ice sheet models ; iii) results at different resolutions.” All 3 use “inter-compare” so it can be brought out front and “is” is missing.

l. 306: An end parenthesis “)” is missing before the end of the sentence.

l. 307: “...we present in figure 9 results…” should be “...in Figure 9, we present results…”

l. 318: “Hence, we could argue to choose a different target…” would be more grammatical as something like , “Hence, we could make an argument for choosing a different target…”

l. 321: “They are overall similar..” should be, “Overall, they are similar…”

l. 328: “...important to justify…” should be “...important in justifying…”

l. 331: “...the resolution **of** the observational datasets…”

l. 334-335: “...also consider to compute…” should be “...also consider computing…”

l. 333-335: These are also great points of discussion.

l. 345: I would suggest: “Elmer-ice with PICO **even suggests** a negative sea level contribution from Antarctica **throughout** the simulation.”

l. 347-348: “...with PICO as shown with the simulation with PISM.” That’s a lot of “withs”. I’d suggest, “...with PICO as shown **in** the simulation with PISM.”

l. 348: “But **here** we want to provide…”

l. 358: “...does not peak that high at grounding lines”: again I take issue with the qualitative and subjective nature of this phrase. How about something like “...does not show significantly higher melt rates at grounding lines, as seen in satellite-derived fields”?

l. 367: “fits best” should be “best fits”

l. 377: “...forces to fit the target values also at the low and high extremes…” should be something like “...that also fit the low and high extremes in the target histogram…”

l. 378: “systematic” should be “consistent”

l. 380: “...matches well the entire distribution of…” could be something like “...does a good job at matching the entire distribution…”

l. 388-389: “ice shelves extent” should be “ice-shelf extent” and “ice shelves changes” should be “ice-shelf changes”

l. 394-395: “...could avoid modellers to use temperature corrections…”: How about this? “...could avoid the need for modelers to use temperature corrections…”