Review of Burgard et al. "An assessment of basal melt parameterisations for Antarctic ice shelves"

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:

For this first time, this paper performs thorough evaluation of a large number of ice-shelf melt parameterizations in a pan-Antarctic context. This is a topic of significant importance to the community, as melt parameterizations allow for Antarctic ice-sheet modeling on time scales that are not currently possible using ice-sheet models coupled to ocean or Earth system models. A number of aspects of this work are novel. First and most importantly, the evaluation is performed on a large number (36) of Antarctic ice-shelves that are fully representative of the continent and which use realistic topography and climate forcing. Second, the rich NEMO model ensemble provides over 500 years of cumulative simulation results, a far more robust dataset than current observations are able to provide. Third, new variants of melt parameterizations (including various ways of including ice-shelf slope dependence and a new approach to plume parameterization) are introduced for inclusion in the evaluation. Fourth, the paper carefully evaluates each parameterization for continental-scale "biases" (relative to the NEMO model), "biases" for individual ice shelves, melt patterns (with an emphasis on grounding lines) and the behavior of parameterizations at high resolution. Finally, the paper includes invaluable recommendations on which parameterizations show the most promise (in a few different categories) and the caveats and best practices for using each, suggested by the evaluation.

The paper is well written and arranged. The figures are quite effective, well crafted, and easy to follow. They give the reader a good, intuitive sense of the relative strengths and weaknesses of the different parameterizations in a variety of ways. The numerous tables will help future users of these parameterizations to know the correct coefficient values to use for each, including recommended values of the tunable parameters that best fit the NEMO simulation results.

I do have a few major concerns about the methodology, and addressing these concerns may entail a significant effort by the authors. For that, I am sorry. But I also feel a duty to be particularly careful in evaluating the methodology and suggesting improvements because of the way that the paper is framed: it aims to provide recommended best practices, including preferred coefficients for use in pan-Antarctic modeling using the evaluated parameterizations.

The first concern is one I share with the other reviewer. It would be preferable to tune the parameterizations with one subset of the available data and evaluate it with another, independent subset. Given the more than 500 years of simulation data, I agree that this would seem to be fairly easy to do by using a random subset of years for tuning and the remainder for evaluation. It may be worth adopting this approach in revising th paper.

The second major concern is that the metrics used in the paper for evaluating the root-mean-squared "error" (RMSE) relative to the NEMO results, Eqs. (32) and (33), do not seem robust to me, nor are these definitions the standard ones used in Earth system modeling. This is especially unfortunate because the vast majority of the results show in the paper (from section 2.4 onward) are dependent on these metrics, including the process for identifying the best-fit parameters, the evaluation of parameterizations on the continental scale (both over each ice shelf and near grounding lines), and the methodology for evaluating parameter uncertainty. I am confident that the adoption of a more standard metric is unlikely to change the broad conclusions of the paper but it is hard to see how it would not change specific results such as the recommended values of coefficients used in the parameterizations. For this reason, I ask the authors to strongly consider redoing their analysis with a more robust and widely accepted version of the RMSE metric.

In my own work and I think in climate modeling broadly, we always use the area-weighted variance to compute the RMSE as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(m_i - m_{o,i}\right)^2 A_i}{\sum_{i=1}^{N} A_i}}, \tag{1}$$

where, in the context of this paper, m_i is the melt rate (in whatever units you choose) from the parameterization for sample i (which might be one location and year); $m_{o,i}$ is the "observed" melt rate from NEMO (at the same location and year); A_i is the area associated with the *ith* sample (this could be a whole ice shelf or an individual NEMO grid cell, depending on how the evaluation is performed); and N is the number of sample points. This metric is robust to different ways of sampling the evaluation points. It works just as well to evaluate the "error" over a single ice shelf as for any number of ice shelves aggregated together into a region. The RMSE values for individual ice shelves can be computed by sampling the pattern at locations across the ice shelf, and then these RMSE values can be combined to get regional- and continental-scale values of the RMSE. This approach is also robust in the sense that including or excluding any number of small ice shelves would have a small impact on the RMSE.

Unfortunately, the same is not true for either of the RMSE definitions, Eqs. (32) and (33), proposed in the paper. Instead of weighting by area, these metrics treat each ice shelf (independent of its area) as having equal weight. This approach has at least three problems. First, the metrics are significantly more sensitive to the number of ice shelves under consideration. Considering Eq. (32) first, repeated here:

$$RMSE_{\mathrm{int}} = \sqrt{\frac{\sum_{i}^{N_{\mathrm{isf}}}\sum_{t}^{N_{\mathrm{years}}}\left(M_{\mathrm{param}}[i,t] - M_{\mathrm{ref}}[i,t]\right)^2}{N_{\mathrm{isf}}N_{\mathrm{years}}}} \tag{2}$$

if we were to add several ice shelves with both negligible bias in melt flux and negligible area, the $RMSE_{int}$ would decrease (because we aren't adding significantly to the numerator but we are adding new ice shelves to the denominator). Now, considering Eq. (33):

$$RMSE_{\mathrm{GL}} = \sqrt{\frac{\sum_{i}^{N_{\mathrm{isf}}} \sum_{n}^{N_{\mathrm{simu}}} (m_{\mathrm{GL,param}}[i,n] - m_{\mathrm{GL,\,ref}}[i,n])^2}{N_{\mathrm{isf}} N_{\mathrm{simu}}}} \tag{3}$$

In this case, if we were to add several tiny ice shelves with negligible area but significant biases in melt rate, they could significantly increase the $RMSE_{GL}$. In this way, these metrics are likely fairly sensitive to the number of ice shelves (36) chosen for the study.

Second, the metric in Eq. (2) here gives significantly more weight to ice shelves with large areas (since $M_i = m_i A_i$, the numerator in this equation is getting weighted by A_i^2), while the metric in Eq. (3) here gives significantly *less* weight to large ice shelves (because A_i does not appear in the numerator at all).

Third, the definition of an ice shelf is not particularly robust. The Filchner and Ronne Ice Shelves are sometimes considered distinct (as in this paper) and sometimes are combined together. Similarly, the Eastern and Western Ross are sometimes considered separately (though Ross is a single ice shelf here). As grounding lines move, ice shelves that were once distinct may merge, or a single ice shelf may be divided into two. The RMSE metrics from the paper are sensitive to the authors' definition of the 36 ice shelves used in the paper, whereas the metric in Eq. (1) here would not change if a nice shelf were divided into 2.

In summary, I do not think robust statistics can be performed when treating ice shelves as distinct entities given equal weight regardless of size. I do not see how a meaningful RMSE in the integrated melt flux can be computed between ice shelves: it would only be meaningful if totalled over all ice shelves in a region (but averaged over samples in time and across ensembles).

My preference would be that the results of the paper be re-evaluated with the metric in Eq. (1) here instead of Eqs. (32) and (33) in the paper. I realize this is a monumental task. If the authors do not feel that this is necessary, I am willing to hear you out. But I would need to see a strong case for why the existing metrics are more robust and defensible than I currently believe.

I want to reiterate that I think this is amazing work! If one takes the metrics for evaluating "biases" as given, the rest of the results follow nicely: the evaluation of each parameterization is remarkably thorough and well executed. The service that this work does for the community is truly commendable, particularly because it gives future model developers not only the tools to make an informed choice among the existing parameterizations but also a good choice of coefficients to use.

Below, I have a number of specific comments and suggestions for typographical or grammatical changes. Please do not be intimidated by the number of comments. I have been told by editors in the past that I seem to review papers as if I were a coauthor, which may be true.

Specific Comments:

l. 7-8: “The box, PICOP parameterization and quadratic parameterizations with slope dependency…” In the context of the abstract (in which these different parameterizations have not yet been introduced), I found it hard to understand what was being described here. The previous sentence does a nice job of introducing what the quadratic parameterization means and what the slope dependence is about. A similar introduction for the box parameterization and PICOP here would be helpful.

l. 19-20: “...interacting with land or shallow rocks…”: I’m not sure I follow what you mean by interacting with shallow rocks. Maybe take this out?

l. 38: “...and do not resolve ice-shelf cavities…”: Should this be “include” instead of “resolve”? It seems like you are referring to model configurations that don’t have the cavities at all, rather than that they are there but at too coarse a resolution.

l. 66-69: When you refer to a “step” here, does each of these correspond to a section of the paper? If so, it might be clearer to replace “step” with “section”.

l. 115 “The top boundary-layer thickness is set to 30 m.” This likely needs a little more explanation: this is referring to the boundary-layer thickness over which temperature and salinity are averaged for use in the parameterization of sub-ice-shelf turbulent fluxes.

l. 147: “basal mass loss” might be better as “basal mass flux” or something like that because usually “mass loss” is used in our field to refer to the uncompensated loss of mass from the ice sheet that contributes to sea-level rise.

l. 188-191: Here, a bunch of physical constants and variables are introduced for the second time (they were introduced on l. 104-115). I think the only new ones here are Γ_{TS}, T_i and U. Also, it seemed to me that U might be the same as u_{oc} introduced above. If so, maybe use the same symbol for both. If not, maybe clarify the distinction.

l. 213: “where the subscript *oc* denotes the far-field temperature and salinity that is advected into the cavity”. This seems like a different definition than on l. 105: “T_{oc} and S_{oc} are the temperature and salinity averaged over a boundary layer below the ice shelf.” It would be helpful if the *oc* subscript meant the same thing throughout the manuscript. Maybe another subscript like *ff* could be used for far-field quantities (or *bl* for boundary-layer quantities)?

l. 217: The liquidus slope has already been defined on l. 108.

l. 240-240: “However, through the resulting circulation, the temperature at one point of the ice shelf cavity is not necessarily independent from the temperature at the other points of the ice shelf cavity.” This may be correct, but I understood the quadratic dependency a little differently. I thought the concept was that the overturning itself was driven by the aggregate effect of the thermal forcing (so that the strength of the overturning circulation would be proportional to the mean thermal forcing in the cavity). Thus, the local velocity is represented in terms of the amount of heat available for melting and there isn't necessarily any implication that the temperatures at different parts of the ice shelf are correlated to one another.

Eq. (17): It seems like some explanation of the bracket < > notation is needed.

Eqs. (20) and (21): I think these should be moved after Eq. (24) (with Eq. (20) after Eq. (21)) because they don't currently have any context where they are.

Eqs. (23) and (25) are identical, right? If so, I don't think it's necessary to have Eq. (25), you can just state that M_2 is still given by Eq. (23).

l 311-312: "If one of the boxes has an area of zero…": It isn't clear to me how this could happen. Is this just for an ice shelf that is poorly resolved so that there aren't enough grid cells for each box to get some?

l. 313: "are non-zero": Should this be "have a non-zero area?"

l. 349-350: "We use yearly mean profiles as the residence time of water in ice shelf cavities might be longer than a month for some cavities." You explain this more in the discussion but it might be worth giving it more context here, too. The parameterizations do not include the finite advection time from the ice-shelf front to the ice-shelf base, meaning they are missing important delays that are present in the physical system. This means there is little hope of capturing seasonal melting in the proper phase.

l. 431: "We therefore only show results for heterogeneous boxes **hereafter**." This is not true in the next subsection when you talk about PICOP, so maybe change "hereafter" to "in the remainder of this subsection."

l. 458-460: "To put these RMSE in context, the integrated ice-shelf melt averaged over all ice shelves and simulation years in the reference is 38 Gt/yr, which means that even for the better-performing parameterisations, the RMSE remains very high compared to the reference melt as such."

I have concerns about this concept, related to my concerns about the metrics in Eqs. (22) and (23). I do not think it is meaningful that the total Antarctic melt flux (~1500 Gt/yr) averaged over 36 ice shelves (an arbitrary number related to the resolution of the NEMO simulations used for reference) comes out to ~38 Gt/yr. The different ice shelves have such different areas and melt rates that this average isn't informative. Given that I have problems with the metric from Eq. (22) that you're given context to, this might be a moot point but if you want to give context to what ~40 Gt/yr is like, maybe just point out an ice shelf by name that has about that much melting.

l. 504-505: "...likely because they are all forced by yearly ocean temperatures." This seems to assume that the parameterizations would be able to get the Mode 3 melt if they were forced with seasonal or monthly temperatures. As you state later, they are unlikely to have the correct processes or the correct seasonal phasing to get these melt rates right even if temperature forcing were provided on a shorter time interval.

Figs. 4 and 7: I appreciated this visual representation of the RMSE – it was very intuitive.

Figs. 5 and 6: These are wonderfully done and very clear!

l. 575-576: "...and is composed of 36 random samples, with replacement, of the different ice shelves." Do I understand correctly that you double or triple (or whatever) count some ice shelves and omit other ice shelves in this process? That seems like an uncommon practice and that the approach (suggested by reviewer 1) of using only some of the years for training and holding back others for validation would be a more standard method for testing the robustness and parameter uncertainty.

Fig. 9a. On the light gray curve it says "fit on pairs with E_0 < 50" but in the text describing Eq. (34), I didn't see anything stating that it was for E0 < 50. Is (34) not for the gray curve? Is this description simply missing from the text?

l. 617-619: "Note that some of the higher values of $C_d^{1/2}\Gamma_{TS}$and E_0 are several orders of magnitude higher than expected (see e.g. Table 12), which we cannot explain." This seems odd. Don't you have constraints on the parameters to prevent them from varying outside a physically likely range?

l. 648: "The slope-dependent parameterisation is nonetheless relatively **good for cavity melt rates…**" Do you mean "...good for capturing melt patterns" or something like that?

l. 744-745: "...with RMSEs on the same order as or even larger than the reference value." This is another reference to ice-shelf-by-ice-shelf statistics that I think needs to be rethought in terms of area-weighted statistics.

Appendix D: Could you say how the RMSE for the heatmaps are defined? Are these the same as Eq. (32) but with $N_{isf} = 1$?

Typographical and Grammatical Suggestions:

l. 11: "Additionally to…" should be "In addition to…"

l. 28: "uncertainty source" should be "source of uncertainty"

l. 140: Probably remove "For example", since this sentence didn't clearly follow (at least for me) as an example of the changed parameters not having a significant impatct on the physical ocean.

l. 204: "from an ice shelf to another" should be "from **one** ice shelf to another".

l. 252: "(based on Sec. 2.2.2)" might be more natural as "(as described in the next section)".

l. 288-291: This is a lot in one sentence and gets a bit confusing. It might be clearer if you break the details of each subscript into its own sentence.

l. 308, 314, 315, 437, 444, 446, 575, 589, 666, 667, 747, Appendix C, Table C1: The word “amount” is used in several places where “number” is correct (because the object being referred to is countable, rather than indefinite): “number of boxes”, “number of data points”, “number of high-resolution ocean simulations”, etc.

l. 381-382: “two constraints are additionally needed” would be better as “two additional constraints are needed”

l. 399, 434: “(2020)(Table 4, 1st and 3rd columns)” and similar: This is very picky on my part but if there’s a way to avoid back-to-back parentheses like here, I would prefer it.

l. 404-405: “The new tuning only slightly achieves to reduce the RMSE further” would be better as “The new tuning achieves only a slight ruther reduction in RMSE”

l. 414-415 and 449: “input properties are **of** a similar order of magnitude” and “all **of** a similar order of magnitude”

l. 438: “...about twice as high **as** the…”

l. 454: “Slightly above…” should be “At slightly above…”

l. 460, 517, 565: “... to the reference melt **as such**.” and “...the melt patterns **as such**”. The phrase “as such” is used in a few places in the paper where I don’t fully understand it. Maybe you mean “itself”? Maybe this can just be omitted?

l. 506: “additionally” should be “in addition”

l. 519: “This can be explained by **punctual** very strong melt”. I do not know for sure what is meant here by “punctual”. This usually means “on time”.

l. 521: “should therefore be kept” would be a bit better as “is important to keep” and “On the opposite” should be “In contrast”.

l. 527: “...in a more consistent **way** and with larger…

l. 543: “emulates well NEMO” should be “emulates NEMO well”

l. 549-550: “...and therefore ~~rather~~ is a result of the circum-Antarctic tuning **rather** than of the resolution”

l. 550 “...on the contrary” should be “in contrast”

l. 568: “...require to account for…” should be “...require accounting for…”

l. 616: “This relation is different **from** Eq. (34)...”

l. 627: “On the opposite” should probably be “In contrast”

l. 637: “Especially” should be something like “In particular”

l. 648: “...used as end members” might be “...used to generate end members…”

l. 660-661: “We therefore **suggest, when possible, trying** out…”

l. 694: “...any improvement in **either** of the two…”

l. 716: “...take into account**, to some extent,** the horizontal…”

l. 719: missing a comma after “geometry”