Original reviewers comments are in *italic and black*, our answers are in blue.

1 Summary and Main Points

This paper presents a probabilistic forecast for mass change at Upernavik Ice Stream in West Greenland. In very broad terms the forecast is produced by running an ensemble of predictions over different model parameters using the ice sheet model Elmer/Ice, and then weighting those ensemble members based on their agreement with a variety of observations. The paper is remarkable in that performs a very detailed methodological exploration of different schemes for weighting ensemble members. The paper also presents an index based sensitivity analysis, allowing for an interesting temporal discussion of the influence of variance in model parameters on the variance in predictions.

Overall, I think that this paper is exceptional and represents an important advance in data-constrained forecasting for Greenland. In particular, the paper makes (and justifies) a rather significant claim, which is that at centennial scales, the dominant source of uncertainty for projection remains elements of climate forcing and that contemporary observational constraint on ice sheet models exhibits diminishing returns. I have no major methodological issues with this work and I find the science to be sound. I do think that the paper could be made more accessible to a broad audience – in particular one that isn't complete versed in the language of probabilistic forecasting, and for whom this paper would still be a very useful read – by some clarification in exposition. In particular, I think that the use of acronyms could be reduced, some sections could be shortened, and the lanauage of Bayesian inference modified to be in closer accordance with standard usage.

We thank Doug Brinkerhoff for his very positive comments. It is truly gratifying to receive such positive feedback about our paper. As suggested, we will add clarifications to make the paper more accessible to a broader audience. In particular, we will include figures that summarize our methodology, which will help to streamline some sections.

2 Line-by-line comments

L23 *'limited' used twice.*

Thanks, we will change the term.

L24 *What is the difference between 'limited understanding of ice dynamics' and 'uncertainties in ISMs' ?*

Agreed, we will change "ice dynamics" by "initial state".

L30 *It's worth noting that in Aschwanden (2021) the authors state that ISMIP6 actually already does a really good job with quantifying uncertainty in model structure! For the other considerations, yes, those must still be better accounted for.*

Yes, we will change the sentence to add this mention.

L34 *This would be an appropriate reference for sensitivity studies in Greenland as well: https://doi.org/10.5194/tc-13-1349-2019*

Thanks for the reference. It seems that this paper is on Antarctica, so we will add it with the reference to Hill et al.

L47 *For item (i), I think it makes sense to characterize this as 'establishing prior distributions' over uncertain model parameters.*

Agree, we will change the sentence.

L50 *It might be worth noting that this lack of cross-validation is often done because there's just not that much data to work with usually, but the authors' point generally a very fair one. I hope that future studies incorporate the authors' suggested cross-validation framework.*

We partially agree with this comment. Despite the limited data availability, we maintain that dividing the dataset into calibration and validation subsets is always feasible and should not detract from the robustness of the results. However, in scenarios where there are no significant variations, such as unchanged ice dynamics during the validation period, this approach might not significantly enhance the robustness of the findings. To address the concern of data scarcity, we will add a mention to it.

L111 *The specification of the calving front position rather than it being a prediction from the model was and remains one of the most contentious aspects of the ISMIP6 experiments. While the position of the calving front may be specified with precision by the parameterization, that doesn't necessarily mean that the position will be accurate in the future, and successfully simulating calving rates and front positions remains one of the largest open challenges in glaciology. It is worth noting here that by ignoring this source of uncertainty in model projections, the authors' are making a large and potentially critical assumption. I don't think it's a problem, but it really does need to be discussed.*

The sentence to which the reviewer refers is for the historical period, and so the position comes from observations whose uncertainties are unknown. However, as described below for the future (Ppr), this uncertainty is taken into account (as in ISMIP6) by adjusting the sensitivity parameter of the parameterisation. This choice of parameterisation is also discussed in section 4.2.2 where we agree that it remains a major challenge.

2.1.1–2.1.3 *I struggle a bit with the semantic separation of the historic ensemble (hpr) from the two future ensembles (cpr and ppr), given that it is the initial period for both. I don't think it's too important but the authors for clarity may wish to refer to the historic ensemble as the 'shared hindcast period' or similar for the two prognostic ensembles.*

We will change the term Hpr to that suggested by the reviewer. To clarify this point, we will also add a figure to summarise our framework.

L141–150 *I find this section to be a little bit confusing. It might help to have a more complete discription of the parameterization and specifically the role of κ.*

We will add a description of the parameterisation.

L195 *'Ice Discharge' should be lower-cased.*

Thanks, we will correct it

L197 *Given that this producet is based on a model result (RACMO), is there the potential for this product to contain systematic bias?*

Yes, it could, even though we have shown in our previous paper that members using RACMO outputs lead to less bias in surface reconstruction than those using MAR (Jager et al. 2024). Potential bias of the input-output method can be find in Mouginot et al. (2019), so we will not change the text.

L203 *'ensemblist' → ensemble*

Thanks, we will correct it

L225 *I think that M needs to be understood as a random vector of model parameters, with P (M) its prior distribution. It will be the case that a sample will be drawn from this distribution, which will be used to create the ensemble, but M is not the ensemble itself.*

Agreed, we will change it "by considering a model ensemble M consisting of n_m members traversing the parameter space Σ" to "by considering a vector of model parameters M from the parameter space Σ"

Eq. 6 *I think that some of these components are mislabelled. In particular, Eq. 6 is not Bayes' theorem, so I'm not sure it makes sense to refer to a posterior and prior as such. Rather, Eq. 6 is the definition of the 'posterior predictive distribution' (which is what the left-hand side should be labelled), which is the distribution of future sea level outcomes conditioned on data. This is decomposed into the two terms on the right side,*

$$P(M|B),$$

which is what would usually be called the 'posterior distribution', 'parametric posterior distribution', or 'calibration' (as it already is in the paper) to disambiguate, and

$$P(SLR|M, B),$$

which is the distribution of model predictions given a particular parameter value (perhaps called 'model prediction' or 'projection').

Yes you are right, we will change the terms to "Posterior prediction", "Model Prediction" and "Calibration".

Eq. 7 *An important condition here is that*

$$Mi \sim P(M),$$

which is to say that the realizations of the particles need to be drawn from the prior distribution. If that's not the case, then the prior needs to appear in the numerator and denominator of the term in L. 237.

Thanks for the comment, we will change the sentence before the equation and add an equation to define P(M), as suggested by the second reviewer.

L240 *'gaussian' should be capitalized.*

Correct

Eq. 8 *This intersection notation is weird – I think it would be better to just start with Eq. 9.*

Agreed, we will remove this line.

Eq.11 *A should be Anobs or a new constant should be used.*

Agreed, we use C instead of A.

L248 *Be explicit as to what this measurement operator is, e.g. the evaluation of the Elmer/Ice FEM basis representation of the velocity field at the desired locations in space and/or time.*

We will add this information.

L255 *It is also worth noting that even if observational uncertainty were IID, model error definitely is not, which is what ultimately leads to the egregiously peaked distributions over ensemble members and heavily weighting only a single one. Ultimately the problem is that – priors aside – we don't know an appropriate likelihood function for comparing models to data! As such, the more ad hoc methodology described in this work is justified.*

Thanks for the comment, we will write two sentences at the end of this paragraph to add this information.

L265–282 *This section is really great. It has significant similarities to lots of previous work on Bayesian inference in the presence of model misspecification, and it might be worthwhile to frame the discussion in terms of that. This is a good reference: https://doi.org/10.1146/annurev-statistics-040522-015915*

Thanks for the reference, but as the article is very generic, we don't see where we can quote it without having to modify the discussion too much.

L280 –282 *I'm not sure I understand this statement.*

We will try to reformulate the sentence to clarify it.

L304 *It would be worth describing whether this weighting scheme is more or less permissive than full-period weighting - I don't have a good intuition. It might also be helpful to introduce an equation for each of the weighting schemes to be concrete.*

On the sub-period weighting, we will add sentences at the end of the paragraph and some details to precise the number of RMSEs used for each weighting.

L308 *I don't really understand the introduction of fparam weighting. This is very much tied to the particular parameterization and behavior of the authors' existing model setup (thoroughly described in a separate paper) and it's challenging for someone not so involved with that work to understand what this specific experiment is trying to capture. Can this be expanded to provide more substantial justification?*

We will reverse the position of the 2 paragraphs of this sub-section to start with the justifications and add some justification of the use of this weighting.

L332 *This sentence changes from passive to active voice in the middle. Probably best to stick with active voice.*

Agreed, we will change it.

Sec.2.3.3 *While I appreciate the desire to include SSP as a random variable, I also think that doing so sort of obscures the influence of all the other aspects of the model since this is expected (and*

turns out to be) a dominant factor in determining predictive variance. Is it possible in later plots to also present ensemble ranges conditioned on SSP (i.e. the sub-ensembles of particles using just SSP2.6, SSP4.5, SSP8.5)? That would be helpful for comparison against existing similar work and would also facilitate a 'best-case versus worst-case' analysis for climate change impacts.

We will add in the supplementary the results conditioned on SSP for the prior ensemble and the weighting. In the main text, we will also add a figure and a paragraph to discuss the effect of weighting on SSPs sub-ensembles.

L343 *It would be super helpful to reiterate here what the difference is between the Ppr and the Cpr. I had initially thought to suggest more instructive names, but I can't think of any, so at the very least a brief reminder of the assumptions of each would be great.*

Agreed, we will add sentences to reiterate the difference between the 2 ensembles.

L356 *Is the agreement between the Hpr median and mass loss observations by design or a happy accident? Fig. 2 and 3 Perhaps consider changing the symbology scheme to something friendly for greyscale/colorblindness, e.g. cross-hatching one of the two shaded regions.*

It is not by design, and therefore a coincidence. For the symbology, we will add hatchs for Cpr and Ppr. For consistency, we will also add hatching to all other figures. Most of the figures will be modified for greater clarity and legibility.

L361 *It might be worth contextualizing this with respect to Robel, 2019: https://doi.org/10.1073/pnas.190482211. The skew in the distribution is perhaps not surprising.*

We will add a reference to this paper at the end of the sentence: "[…], which is similar to other results in glaciology (e.g. Robel et al. (2019))."

Sec. 3.3.1 *It's a little cumbersome to start a section with a reference to another section. I understand shunting the methodological details to the appendix, but a recapitulation of the methodological approach would be helpful here.*

Agreed, we will start the section with a summary of the cross-validation method previously introduced.

Sec. 3.3.1 *More generally, all four points introduced here seem a bit ad hoc. Do there exist references that could help place the current procedure on more sound probabilistic footings? Seems like this problem has to have been studied before.*

We agree that the four points are a bit ad hoc, but it is because this section just summarize our results. We discuss this in section 4.4.1 Use of Bayesian Calibration, where we show that our results are similar to those of Jiang and Forssén (2022).

L446 *Where is factor mapping previously established?*

It is introduced in section 2.1.1. We will add a reference to this section.

L448 *The parameter flaw sometimes appears throughout the manuscript as just law. Please revise for consistency.*

Thanks, we will change all terms to f_{law}.

Fig. 5 *The overlying transparent bars aren't really readable.*

As in Figures 2 and 3, we will add hatch patterns.

Sec 3.4.1 *I think that this section would benefit from a bit of extra subdivision. I think it would be helpful to break this into individual subheadings describing the historical period and the forecasts. I think it would also be helpful to separate the principal conclusions about the relative insensitivity of long term forecasts to ISM parameters from the details of weighting. I also don't think that it makes sense to refer to the changes in ranges described around L506 as 'notable' – the more notable thing is that they're almost exactly the same!*

Agreed, we will add sub-headings to help the reader distinguish between the Hindcast ensemble, the Control ensemble, and the Predicted ensemble. We will also change the term 'notable' to 'little'.

Fig. 6 *The font in this figure is too small.*

Agreed, we will changed it.

Sec 3.4.2 *Again, I would like to reiterate that presenting ensemble results which each of the SSPs held fixed would be useful here, and would help to ameliorate some of the challenges associated with trying to guess the probabilities of future human carbon emissions (which is why previous works have treated these as fixed hypotheses rather than as random variables).*

As mentioned in a previous comment, we have added a figure and a paragraph to demonstrate the effect of weighting when separating the different SSPs.

Sec. 4.1 *I am not quite sure I understand the reference to ISMIP6 here. How is that relevant to the present model being able to reproduce observations?*

These papers explain the reason behind the use of a control run in ISMIP6. We will rewrite the sentence.

Sec. 4.2 *Again, I think that this section would be clarified by adding some more sub-headings. e.g. at L589, this paragraph could be called 'reducing uncertainty through ISM calibration), whereas at L598, this could be called 'reducing uncertainty through climate forcing calibration', or something like that.*

Agreed, we will added the suggested sub-headings.

L576 *The referenced compensatory effect is not clear to me from Fig. 4 or elsewhere. Could this please be clarified?*

We will add a figure in the supplement to illustrate this compensatory effect, and we will change the sentence.

L585 *If the front parameterization has such as significant influence, then perhaps this calls into question the validity of imposing the front at all. Would it be possible to make a statement about how the predictions might be influenced if the front were allowed to evolve freely or based on a different parameterization?*

We agree that imposing the front using this parameterisation is not the optimal method for projecting calving in Greenland tidewater glaciers. We believe our paragraph on this specific issue ('In the context of front retreat parameterisation, [...]') demonstrates the limitations of this approach

and highlights future developments needed in the ice sheet modeling community. For Elmer/Ice, implementing calving laws is currently under development but not yet available.

L602–604 *This is a very significant assertion that would have major implications for how ice sheet modeling proceeds in the future! What should the community be doing if improving ice dynamics isn't likely to help? (note that I don't disagree with the assertion – I am genuinely curious where effort should be allocated instead).*

We want to clarify that this assertion is mainly valid for Greenland ice sheet (we will add a mention to Greenland ice sheet in this sentence), and may differ significantly for Antarctica. Besides this, we believe there are numerous challenges to address:

- firstly, we demonstrate here that front parameterisation has a significant influence. Part of this influence stems from ocean processes (such as the transport of warm water into the fjord and melting at the front), but another part is due to uncertainties surrounding calving itself. This aligns with the previous comment: developing and validating calving laws for our Elmer/Ice ice sheet configuration appears to be a priority;
- secondly, in my opinion, we need to better account for uncertainties associated with the bed. Here, we have neglected this uncertainty, as in our previous article, but tests conducted in the previous article indicate a strong influence of this uncertainty on the ice discharge obtained. This factor should also play a crucial role in calving, as the stabilization of the front at a given point heavily depends on the bed. Since Bayesian calibration cannot be applied to this bed calibration, other transient data assimilation techniques will need to be used (e.g., Gillet-Chaulet, 2020);
- thirdly, it is possible that phenomena currently unknown could alter future outcomes. New discoveries often lead to higher predictions of mass loss;

We will add a mention that this assertion concerns the Greenland ice sheet and that some uncertainties have not been explored. We will add the discussion paragraph of this front parameterisation directly after this paragraph on ice dynamics.

L618 *Its foundation in observational data is sort of the problem – no data available in the future.*

Partially agreed, as this statistical downscaling has been partially validated through cross-validation: a training set composed of some glaciers and a validation set with the remaining glaciers.

L703 *There are other possibilities than the Gaussian or T.*

Agreed, we will rewrite the sentence.

Sec. 4.4 *There is a lot of good in this section, but there is also a lot of material that is only applicable to the authors' own modeling setup (which has already been covered) it might be worthwhile to take a critical read to assess which of these insights are going to be generally applicable, and which are more like notes to guide the authors' own continuing work.*

We will add two sub-headings: the first one to highlight insights in terms of Bayesian calibration ('Use of Bayesian Calibration'), which will interest anyone using such data assimilation techniques; and the second one to emphasize insights in ice sheet modeling, particularly focusing on friction ('Friction Law'). We agree that the first paragraph mainly concerns ISMs using data assimilation to calibrate friction fields like ours, but other ISMs using similar techniques may also find these results valuable. Regarding the second paragraph, we also agree that the parameterization mainly applies to our model. Therefore, we will try to rephrase the beginning of this paragraph to emphasize that it is crucial to consider sub-hydrology, at least in a parameterized manner as demonstrated here.

Sec. C2 *The student-t distribution has an additional degree of freedom, namely the number of degrees of freedom. What was used for this, or how was it estimated?*

As in Aschwanden and Brinkerhoff 2022, we used 2 degrees of freedom. We will include this information in this section C2.