We thank the Anonymous reviewer #2 for their insightful and very detailed review. Below we explain in detail how we intend to address the reviewer's comments during the revision process. Our replies are highlighted in italics.

In this study, the authors evaluate the ISMIP6 model experiment results to identify the dominant sources of uncertainties in sub ice shelf melting and dynamic sea level rise. To do so, the authors define two linear scaling factors: the melt sensitivity (MS), which is derived from an assumed linear relationship between shelf-average basal melt anomalies and average thermal forcing anomalies, and the dynamic sensitivity (DS), which characterises another assumed linear relationship between cumulative basal melt anomalies and dynamic ice loss. These factors, along with a categorical climate forcing variable and an observed/qualitative classification of modeled calving are used as the basis for decomposing variance in the model outputs across the ISMIP6 ensemble. Analysis of variance results in the conclusion that MS and DS factors are significant contributors to the uncertainty of the ensemble results, explaining about 75% of the variance in the dynamic sea level projections. The authors seem to present this as the primary result of their work. However, the results from the 3-way ANOVA deserve more of a highlight, especially the transition of the leading contributor to BMB uncertainties from melt sensitivity to calving groups.
Agreed, we will highlight the more the transition of uncertainties from BMB to calving.

General Comments:

- **Linear Regression:** The authors do a good job in acknowledging the limitations of using linear approximations to derive MS and DS, particularly in Discussions. However, several regression choices lack justification or explanation in the Methods, making it challenging to infer the motivation behind them. For instance, the purpose of the linear regressions shown in Figure 6 is unclear — they appear to serve as an evaluation of MS correlation with total melt change (is this total melt anomaly?) in sectors where predefined calving groups don't present a clear distinction in basal mass balance anomaly time series. I could not find any additional justification for this regression, or why a regression was chosen compared to a nonparametric statistical test suited to the data. It is recommended that the authors provide a consolidated description of all linear regression analyses in an expanded Methods section, including the purpose of each. This should also include a clear statement of the assumptions underlying each regression: for example, assumptions about residual independence, homoscedasticity, and linearity. This is particularly important for any regression analyses involving MS or DS and variables used in their derivation, where circularity in the relationships could affect interpretation.

    We thank the reviewer for this helpful suggestion. The purpose of the regression shown in Figure 6 is to quantify the relationship between MS and the total melt response, rather than solely to assess statistical significance. In this context, linear regression provides a simple and interpretable measure of how variations in MS relate to variations in melt response across models. We agree, however, that this approach implicitly assumes a linear relationship and may not fully capture more complex or non-linear behaviour. Nonparametric approaches, such as rank-based correlation measures (e.g. Spearman correlation) or group comparison tests, could provide complementary insights without relying on such assumptions.

To address this, we will clarify the purpose of the regression in the Methods section and discuss its limitations more explicitly. We will also include an additional nonparametric correlation analysis to demonstrate that the identified relationship is robust and not dependent on the assumption of linearity.

Regarding the reviewer's point on potential circularity, we acknowledge that MS is derived from the same underlying variables that influence melt change. We will clarify this relationship more explicitly and discuss its implications for the interpretation of the regression results.

· **Statistical Analysis Descriptions:** The Methods section on ANOVA lacks sufficient detail to support interpretation of the results, which has led to key analysis choices being scattered across the Results section. Notably, the definitions of the calving groups introduced in Section 3.1.2 and the definitions of the categorical variables used in Section 3.1.4 would be more appropriately placed in the Methods (or at minimum the Supplement). More broadly, the statistical analysis workflow would benefit from a clearer, unified description that establishes how ANOVA, the calving group classifications, and the categorical variable definitions fit together as a coherent analytical framework. Additionally, the Kruskal-Wallis H-test is introduced in the Methods but does not appear to be referenced again, leaving it unclear where its results are reported and interpreted. The authors should either clarify where these results appear and explain how this test relates to the broader analysis.

We agree that the current presentation of the statistical analysis is fragmented, with important methodological details appearing in the Results section rather than being introduced in a unified way in the Methods. In the revised manuscript, we will reorganise the statistical analysis workflow to provide a clearer and more coherent description in the Methods section. In particular, we will move the definitions of the calving groups (currently in Section 3.1.2) and the categorical variables (Section 3.1.4) to the Methods (or Supplementary Material), so that their role in the analysis is established before they are used in the Results. We will also expand the Methods section to provide a consolidated description of how the different components of the analysis—calving group classification, categorical variables, ANOVA, and related statistical tests—fit together as a unified framework. This will clarify the purpose of each step and how the statistical methods are used to interpret differences between models and regions.

Regarding the Kruskal–Wallis H-test, we agree that its current presentation is unclear. In the revised manuscript, we will explicitly state where the results of this test are reported and how they complement the ANOVA analysis. If necessary, we will either integrate its results more clearly into the main text or remove it if it does not add substantial additional insight.

· **Further Support for Calving Group Selection:** The authors present a thorough regional and sector-level analysis of their calving group classification (Section 3.1.2) in addition to admitting that this classification is qualitative, which is appreciated. However, the validity of these groupings would be more convincingly established if a formal test of statistical difference were applied directly to the BMB anomaly time series across groups, perhaps through a functional ANOVA or a similar trajectory-based test. Statistically significant differences in time series between the defined groups should also remove the need for the linear regression . This is particularly important given that the calving groups serve as a categorical factor in the subsequent variance decomposition: if the groups are not statistically distinguishable in their BMB anomaly trajectories, the interpretation of the ANOVA results rests on a qualitative classification.

We thank the reviewer for this insightful comment and agree that a more formal assessment of the differences between the identified calving groups would strengthen the analysis. The current grouping is based on the observed bimodal behaviour in the basal mass balance (BMB) anomaly time series, and was intended as a physically interpretable classification rather than a purely statistical one. However, we acknowledge that, as these groups are subsequently used as categorical variables in the ANOVA framework, it is important to demonstrate that they are statistically distinguishable. To address this, we will include an additional statistical analysis to quantify differences between the groups. We will incorporate these results in the revised manuscript and clarify how the statistical distinction between groups supports their use within the variance decomposition framework

· **Figure Organization and Styling:** Several figures in the main text present information that is either peripheral to the central discussion or duplicative of content covered in the narrative, while figures that are directly relevant to key results have been relegated to the Supplement. The authors should conduct a systematic review of figure placement, prioritising figures that support the main claims of the paper in the main text and moving ancillary material to the Supplement. In some cases, minor additions or removals of individual panels would substantially improve clarity. Specific instances are identified in the detailed comments.

We agree that the current figure placement can be improved. In the revised manuscript, we will conduct a systematic review of all figures to better align them with the central narrative of the paper. In particular, we will prioritise figures that directly support the main results and interpretations in the main text, while moving more peripheral or supporting material to the Supplement. We will also review individual figures and panels to remove redundancy and improve clarity, including consolidating or simplifying figures where appropriate. We will address the specific instances raised in the detailed comments and ensure that the revised figure structure more clearly supports the key messages of the manuscript.

Specific Comments:
Figure 1: I recommend using a different color scheme or changing the order of the groups being plotted, otherwise all three subplots are far too crowded to actually discern the differences from each climate model forcing. The authors can also consider splitting the plots by climate forcing groups.
We will find a way to improve visibility of the figure.

48-52: This sentence is far too long and contains too much information to be clear. Splitting this up into separate sentences is recommended.
Agreed, we will be splitting this up into separate sentences.

78: Descriptions for the abbreviations of initialization procedures should be included in the Table 1 caption, even if they are present in the main text of the section.
Agreed we will add the abbreviations on the table caption.

135: How do the authors anticipate that the PICO derived forcings are approximately similar to the extrapolated thermal forcings? Consider adding a reference or elaborate.
As the same concern was mentioned by reviewer 1 we will add a quantification for the thermal forcing actually used in PICO and the one we referred for an exemplary case. But the assumption is that due to the extrapolation of the CMIP ocean data under ice shelf cavities in the original forcing data set (Jourdain et al., 2020), temperature below the shelf and at the shelf front may not be too different. We will include this explanation in the new manuscript alongside the example quantification.

165: Shouldn't the term in the denominator be Shelfarea instead of delta Shelfarea?

The formulation of Eq. 1 was indeed misleading and will be corrected. In the actual MS derivation, we regress the simulated average shelf basal melt against average shelf thermal forcing as:

$\overline{BMB} = MS \cdot TF + c\overline{BMB} = MS \cdot TF + c$
Where $\overline{BMB} = BMB/Shelfarea$.

180-185: Figures S2 and S3 are far too small. We should be able to see the trends referenced in the text.
Agreed, we will improve figures.

188-189: Why was this choice made? Would it not make more sense to unify the derivation process of MS factors for each experiment? Since MS feeds directly into the variance decomposition, methodological inconsistency risks confounding physical differences in melt sensitivity with artefacts of the derivation approach. The authors should further justify this choice and report how the variance fractions change under a unified derivation scheme.

Our objective in deriving the MS factors is to obtain values that allow an accurate reconstruction of the simulated basal melt response. The use of different fitting approaches for MS reflects the need to ensure a good representation of this relationship across all models with differing behaviour, rather than arbitrary methodological choices. While the method used to estimate MS may vary depending on model characteristics, the criterion for evaluating its performance—how well B_predict reproduces B—is applied consistently across all cases. Using the same fitting method would be unfair to a few models and could lead to a misrepresentation of the melt sensitivity. We will stress this more in the new manuscript.

195-198: This paragraph would fit better in Section 2.1
Maybe elaborating a bit more on the different melt parametrization could be put in 2.1. but the MS factor is not introduced there so we rather keep the sentence here.

213–220: If a unified derivation (i.e. using the same sections of the experiment timeseries) method was applied for DS but not for MS, the authors should explicitly justify the inconsistency.
Because the DS was derived from a linear regression of <u>*cumulative*</u> *BMB and sea level contribution, there was not such a large variation between different models, hence the one derivation of DS was applicable for all models. We will add a sentence describing this in section 2.4.2.*

265-274: If this result is derived from the Kruskal-Wallis H-test for the MS factor, it should be stated explicitly. Same applies to any Kruskal-Wallis H-test results reported for the DS factor.
The results in lines 265-274 are derived from the Kruskal-Wallis H-test, which is a one-way ANOVA to investigate the relationships between various assumptions and the MS factor. So we used the KWH-test to arrive at these conclusions and will mention this now in the manuscript explicitly. We did a similar procedure for the DS factor and also did not find any statistically significant relationships, so we didn't include that result

344: The descriptions of the melt sensitivity bins are not very easy to understand. Authors should consider using/adding equations/inequalities to be more precise.
Agreed we introduce a more precise description of the sensitivity bin relying on mathematical formulations.

359: Is the unexplained contribution here the 3-way interaction? If so, the authors should be consistent with Figure 7.
Yes exactly we will refer to " Figure 7, (3-way interaction") in this sentence now.

Figure 10: The referenced observations from this plot would be more evident if the data encoded in the colormap was represented in two additional scatter plots showing any potential linear correlation of dSLR with MS and DS factors respectively.
Agreed, we will adjust the Figure to better represent the findings.