We thank the editor, reviewers and the community commenter for their constructive feedback in elevating the standards of the manuscript. We have tried to carefully address all the comments as outlined below and incorporated into the revised manuscript, with changes reflecting in the updated numbering.

Reviewer – 1

> *We thank Reviewer 1 for their positive assessment of our study and the constructive comments below.*

Reviewer comment:

The authors indicate in Section 2.1 that they compute the elastic loading displacements in a centre-of-solid Earth (CE) reference frame at each GPS site location. However, GPS displacement time series are in the centre-of-figure (CF) frame. Although the CE frame closely approximates the CF frame, it is important to acknowledge this potential mismatch in the manuscript.

Author's response:

We have expanded the relevant sentence on CE-CF approximation in Section 2.1 to explicitly acknowledge the CE/CF frame distinction and its negligible impact (~2%; Dong et al. (1997)). The revised lines now read in Section 2.1, lines 101-105.

> *Lines 101-105: The GPS time series at non-secular timescales are expressed in a centre-of-figure (CF) frame, and thus a small inconsistency exists between the modelled loading displacements and the GPS displacement series. However, previous studies have shown that the CE frame approximates the CF frame to within ~2% for non-secular signals (Dong et al., 1997), and this difference is expected to be negligible relative to the magnitude of the loading signals.*

Reviewer comment:

The authors also mention that SMB anomalies are bilinearly interpolated onto a common regular grid of 2 km resolution. It would be very complementary to the discussion to address the effect of interpolation, specifically smoothing of the signal versus using the native grid resolution.

***Author's response*:**

We have explicitly added a discussion in section 2.1 (lines 77-82), clarifying the rationale for using a common 2km grid, the smoothing effects of bilinear interpolation and the limited impact of interpolation on the computed elastic loading response.

> *Lines 77-82: (Haran et al., 2021; Mottram et al., 2021). Interpolation of coarse SMB grids is important as SMB variability over the ocean or floating ice does not impose a crustal loading signal. Interpolation onto a common high-resolution grid avoids introducing loading computation errors due to low resolution coastline definitions, with our choice matching the native resolution of the highest-resolution product considered. While bilinear interpolation results in necessarily distributing signal smoothly, the elastic loading response at a GPS site integrates SMB contributions over spatial scales of several hundred kilometres (Fig. 1).*

Reviewer comment:

In line 170, it is also worth mentioning that different SMB models adopt different topography grids, which might potentially contribute to spatial coherence variability.

Author's response:

We agree with the suggestion. We have added mention of this point in Section 2.1 (lines 66-70).

> *Lines 66-70: In addition to different spatial resolutions, these models have different approaches to implementing large-scale climate forcing, have different parameterisations of SMB processes, and different underlying topographic models that are then sampled at the specific model's spatial resolution. The latter differences can lead to systematic divergence in orographically driven precipitation patterns, particularly over complex terrain. These all contribute to inter-model variability.*

Community Comment – 1

Reviewer comment:

Thank you for your work on this. Just a note about the spatial resolution of GEMB. It looks like you are using the zenodo version forced with ERA5. As opposed to the version in the GEMB paper, that was forced with high spatial resolution RACMO output, the ERA5 GEMB are directly forced with low-resolution reanalysis (~.25 degrees). This means the input is lower resolution than 10km. Even though the model is run, and reported, at a higher spatial resolution, in this version of GEMB there is no true "downscaling" involved. The ERA5 inputs are simply interpolated to the GEMB grid before the energy and mass balance is calculated. This is probably the reason that the results look smoother and more like a lower resolution output. While the current GEMB release includes the model just running the column snow model in the input it receives, downscaling routines will be part of a release in the near future, and we should expect to see a sharpening if gradients in areas like the margins. Happy to discuss more if you need more information.

Author's response:

We thank Nicole for this very helpful comment. This clarifies our finding that GEMB_10 performs comparatively weaker than RACMO_2DS and RACMO_11, particularly over the Antarctic Peninsula, despite its nominally comparable grid resolution. We have updated both the data section and the discussion as follows:

Section 2.1:

> *Line 65: …GEMB (10 km, monthly; (Schlegel and Gardner, 2023); forced directly with low resolution (~0.25°) ERA5 reanalysis fields interpolated to the model grid)…*

Section 4:

> *Lines 394 – 397: For example, although the GEMB model used in this study is run and reported at a nominal spatial resolution of 10 km, its effective input resolution is forced by atmospheric fields interpolated from the ~0.25° ERA5 reanalysis fields prior to the surface energy and mass balance calculations, without downscaling.*

Reviewer -2

We thank reviewer 2 for their through and constructive review. We address the major and minor comments below:

Reviewer comment:

The authors are clearly aware of this issue as it is discussed briefly in the text, but it is not really clear (and perhaps that is because it is not) on which dimension of variability is being evaluated here, and I'm assuming the answer is likely both variability in space and time. The title states, "Evaluating Surface Mass Balance Variability from Climate Models…", but it's obviously a convolution of space and time, which make its somewhat more difficult for the modelers to assess why their model performs as it does. This conundrum is central to the paper, and I believe deserves more exploration. A few additional thoughts to consider:

Author's response:

Yes, the variability occurs across both space and time, and GPS records from different sites often begin at different epochs, may end early, or contain data gaps. We have revised the manuscript in three places to make this explicit.

First, we note the heterogeneity of the GPS records in Section 2.2 (Line 134):

> *Line 134: GPS records in Antarctica begin at different epochs, and some end early or contain data gaps.*

Second, we have added an extended discussion of the space-time convolution of at the start of Section 4 (Lines 345 – 352)

> *Lines 345 – 352: Our GPS-based evaluation of SMB models provides an integrated observational constraint on SMB-driven elastic loading. Because the GPS loading signal reflects the combined effects of SMB variability over both space and time, our analysis evaluates the ability of each model to reproduce the integrated space-time variability recorded by the GPS network. While this complicates attempts to understand the origin of the mismodelled signal, careful analysis of the time series differences and the spatial pattern of SMB variability (e.g., Fig 1b) provides further key insights. Given the GPS network used here is concentrated in coastal and peninsula regions, our evaluation is therefore more informative in these regions. The conclusions about the model performance in the interior of Antarctica remain comparatively under-constrained by the current network, given the small number of interior sites governed by limited bedrock outcrops.*

Third, we clarify our aim of the study in the introduction (Lines 57 – 59)

> *Lines 57-59: We note that our objective is to evaluate the ability of GPS observations to test SMB model variability, rather than to diagnose the specific physical causes of inter-model differences, which would require a dedicated model intercomparison study.*

Reviewer comment:

Can one evaluate from each model the cells where the spatial signal exceeds the local signal? As in, where does the far field signal exceed the near field? Perhaps this would help us better understand the strengths/weaknesses of each model.

Author's response:

We thank the reviewer for this valuable suggestion. Rather than a full per-site, per-model computation across all 98 sites and seven models, which would be computationally extensive, we have extended the analysis presented in Figure 1 for three representative models. A table of figures is added in Supplementary material (Table S2) and this is referenced in Section 2.1 (Lines 118-120).

> *Lines 118-120: We extend this analysis to a selected set of GPS sites in Table S2, for RACMO_2DS, GEMB_10, MERRA2_12, illustrating how the balance of near-field and far-field contributions and resulting variance reduction varies across sites and models.*

Reviewer comment:

Related to the above, the GPS measurements are typically restricted to areas of high spatial SMB variability in coastal regions (although a few interior sites do exist). Because of this potential bias, the GPS sites that you use might be focused on regions where the far field signal exceeds the local scale, suggesting the GPS are more evaluating the spatial signal in the variability as opposed to the actual local variability through time, which would then bias performance metrics towards those models with finer spatial resolution.

Author's response:

We do not agree that the technique is evaluating the spatial signal as opposed to the temporal signal. Rather, the GPS is assessing the spatially integrated (a weighted spatial convolution) temporal mass variability. We have made changes to make this explicit in Section 4 (Lines 356-359). Also, we have acknowledged the GPS-based evaluation is most informative in the coastal and peninsula regions (Lines 349-352).

Lines 356-359 The observed site-to-site variability is consistent with the finite spatial sensitivity of GPS, whereby each site responds to spatially-weighted integral (i.e., Fig 1a) of temporally varying surface mass changes over near-field to far-field regions (e.g., Fig. 1c). Consequently, inaccuracies in far-field SMB can still exert a strong influence on estimated variance reduction and scaling.

Lines 349-352: Given the GPS network used here is concentrated in coastal and peninsula regions, our evaluation is therefore more informative in these regions. The conclusions about the model performance in the interior of Antarctica remain comparatively under-constrained by the current network, given the small number of interior sites governed by limited bedrock outcrops.

Reviewer comment:

And related to the above, was there consideration of weighting of the variance reductions across all sites for each model? One might attribute the best performance to a specific model but perhaps that is solely because the sites happened to be concentrated in their region of highest performance. Some additional discussion of the sparsity of sampling is worthwhile.

Author's response:

We confirm that spatial weighting was not applied in the analysis, and we have flagged this in the Section 4 (Lines 365 – 368)

Lines 365 – 368: We note that spatial weighting was not applied to the computation of median variance reductions. Consequently, the underlying site distribution particularly concentrated in Antarctic Peninsula and West Antarctica should be considered when comparing models' apparent overall performance (Fig. 6).

Reviewer comment:

It would be interesting to explore the site-by-site variance reductions more. Which sites are all negative or positive contributors to variance reduction? Which has the largest range of performances? Across all models there are a substantial number of sites where the variance increases (red dots); what does this imply? There is SMB signal that the model is not capturing? It's also important to note that the model with the highest variance explained had a median reduction in variance of 23%, which is still very small, suggesting there is a lot of unexplained SMB variability even in the highest performer.

Author's response:

We have expanded the discussion of variation in site-by-site variance in two places in Section 4 and have added a new Supplementary Figure S13 showing the range of variance at each site across all models and identifying the sites where variance increases consistently across all models. First, the possible reasons for this increase are discussed explicitly in Section 4 (lines 449 – 456).

> *Lines 449 – 456: Across all seven models, seven sites consistently show an increase in variance and are grouped in Category 5 (red) (see Fig.S13). Generally, an increase in variance at a given site can arise from: i) either the SMB loading signal at that site having poor temporal and/or spatial accuracy (see Fig. S2), (ii) the GPS record contains substantial low-frequency deformation that is not associated with SMBLs, such as those due to ice dynamic mass changes or volcanic activity, or (iv) the GPS time series is compromised at low frequencies due to snow intrusion into the radome or other equipment issues (Koulali and Clarke, 2020). In these cases, we note that the non-SMBL signal is dominant and hence interferes destructively with the application of SMBL time series (see Fig. S2).*

Second, regarding the interpretation of the median variance-reduction values, we have added a discussion of this interpretation in Section 4 (lines 404 – 408).

> *Lines 404 – 408: We note the best performing models achieve a median variance reduction of <~30%, implying that a substantial fraction of low-frequency GPS variance remains unexplained across all models. A substantial portion of this residual unexplained variance at GPS sites is attributable to non-SMBL geophysical signals common to all the models evaluated, including residual mis-modelled SMB signals and unmodelled non-SMB loading signals such as varying ice dynamics, viscoelastic deformation, as well as low frequency GPS noise. These signals are not expected to be reduced by any SMB loading correction.*

Reviewer comment:

Finally, is there any impact of the choice of the time span of the reference climate interval or the long-term mean SMB on your results? After generating the cumulative anomalies, the authors detrend the time series, so that likely accounts for much of the differences that would result from a variable choice of reference time interval. There could still be minor differences, however. Is it worthwhile to explore various time reference intervals to see if it impacts the ability of the SMB models to reduce the variance in the residuals?

Author's response:

We have added a brief statement to this effect in Section 2.1 (Line 73)

> *Line 73: The loading variability is not particularly sensitive to the choice of reference period provided it is multi-decadal.*

Reviewer comment:

Does the solid earth response have a memory of the historical mass loadings? The authors state that (while not perfect) there does appear to be some correlation between higher accumulation and higher accumulation variability in time with more variance explained. It could be that regions with higher mass loadings have a higher mass flux where the total mass signal has a "shorter" memory, resulting in a ~40-year record that can approximate most of the fluctuations whereas at lower accumulation sites there is a much longer memory that models cannot explain.

Author's response:

The discussion beginning with "We note that, as shown by Nield et al. (2025), vertical motion …" (presently starting at line 409) in the manuscript already discusses this point very strongly. Therefore, we have not made any changes.

Minor Comments

Reviewer comment:

Line 65, states the SMB anomalies are determined by using a refence climate computed over 1980-2022, but TableS1 shows different start and end dates. I'm assuming that indeed they are all calculated over their common interval and that the start and end dates are simply the temporal extent of the model. Also, I assume that the detrending is based on that same interval (1980-2022)?

Author's response:

The reviewer is correct on both the occasions. We have amended the Table S1 caption to remove this ambiguity and it now reads as follows:

Summary of characteristics of Surface Mass Balance (SMB) estimates used for the computation of SMB loads, including temporal coverage, temporal resolution and spatial resolution. The start and end dates listed reflect each model's native temporal coverage.

Reviewer comment:

For clarification, the MERRA-2 SMB product is provided at 12.5 km resolution, but the temporal variability in time comes from the MERRA-2 reanalysis, which is very coarse in space (0.5 degrees latitude by 0.625 degrees longitude). The precipitation magnitudes come from the 12.5 km high resolution M2R12K data product. Therefore, it is more of a blend between 12.5 km and several 10s of kms. Given that this paper focuses on the variability more, its representative resolution is likely much larger than 12.5 km given variability is driven by MERRA-2 at coarse resolution (see Medley et al., 2022).

Author's response:

We thank the reviewer for this clarification. We have added the following to section 4 (Lines 396 – 398).

> *Lines 396 – 398: Similarly, although the MERRA-2 SMB product is distributed at 12.5 km (MKR12K) resolution, it is driven by the underlying MERRA-2 reanalysis at a much coarser ~0.5° × 0.625° grid making the representative resolution much larger than 12.5 km (Medley et al., 2022).*