We thank the editor and both reviewers for their thoughtful suggestions of this study. Below find reviewer comments colored blue and our responses colored **black and bolded**.

Reviewer 2

In `The Greenland Ice Sheet Large Ensemble: Simulating the future of Greenland under climate variability,' Verjans and co-authors use a stochastic variant of the Ice Sheet System Model to explore the sensitivity of the Greenland Ice Sheet to variability in oceanic and surface mass balance forcing. In particular, they aim to quantify the relative importance of such so-called `aleatoric' uncertainty relative to other types of uncertainty derived from imperfect or unresolved modeling assumptions and initial conditions. Through a detailed comparison of ensemble experiments meant to represent both a continuation of contemporary forcing alongside a potential high end warming scenario, they find that the influence of stochastic climate is non-negligible over the coming two or so decades (in terms of total predicted mass change), while these stochastic effects become relatively unimportant over century-scales.

This result is interesting (albeit not particularly surprising) in that it illuminates a principal challenge for short term sea-level prediction, while providing some important guidance as to whether short time-scale variability represents a source of uncertainty that needs to be better quantified for long term projection (thankfully not, it seems!).

This manuscript represents an impressive and insightful culmination of several methodological threads that seem to have been `in the works' for a few years -- the development of StISSM and its ensemble generation tools, the statistical characterization of climate variability in a generative sense, and the coupling of ice dynamics to downscaled surface mass balance and frontal ablation paramterizations. The current work is undoubtedly at the vanguard of ensemble methods for ice sheet uncertainty quantification, and a big step forward for understanding Greenland's sensitivity to climate noise. I have no issues with the paper's general methodology. I have included below a few comments that I hope can improve the manuscript's clarity and utility.
We thank the reviewer for their suggestions and address them each in turn below.

L42: `are performed' should be `have been performed' for consistent case.
Fixed

L46: Perhaps here, perhaps elsewhere, it's maybe worth providing a higher level overview of where climate stochasticity comes from (and where it does not). In particular, it's worth noting that climate is very likely not actually random, but rather appears that way due to the chaotic dynamics characteristic of the atmosphere (and ocean, to a lesser extent). Ice sheets do not exhibit such ostensible stochasticity (EISMINT2 and ice streams notwithstanding), so the irreducible uncertainty in the ice sheet context is derived solely from the forcing term.

These are good points, we have added additional sentences here to make the origin of this idea, that stochastic climate forcing can be used to approximate the deterministic chaotic behavior of climate variability, more clear. Here is the revised text:

“For any given scenario of anthropogenic forcing, there remains an ``irreducible'' (also called ``aleatoric'') uncertainty associated with the ice sheet response to internal climate variability, due to the limited predictability of the chaotic climate system (Lorenz 1969). Indeed, nearly 50 years ago, in work that would later garner a Nobel Prize, Hasselman (1976) posited that such chaotic deterministic climate variability could be equivalently represented by stochastic random forcing when simulating slow ``integrators'' in the Earth system, such as ice sheets. However, in the last five decades, there have been no sustained attempts to build such a stochastic large-scale ice sheet model.”

L84: What question is being referred to here?
Changed to *“We apply a novel stochastic modeling approach to produce this ensemble.”*

L143: A qualitative description of what EN4 is, and why it's helpful for bias correcting the ocean thermal forcing would be very helpful.
We have added additional details about the EN4 product here and why it is helpful for the bias correction used here (and refer again to Verjans et al. 2023 for more extensive discussion of this point).

L202: I was expecting a similar interpretation of the moving average component of the fits. Do these exhibit any interesting patterns? Does the MA component even matter?
We’ve added a sentence addressing this issue: *“The fact that very few of the best-fitting parameterization include a moving-average component (q > 0) is not surprising and follows considerable prior work (Gilman et al., 1963; Hasselmann, 1976) showing theoretically and empirically that climatic variables are well described as purely stochastic autoregressive processes due to the memory implicit in systems with finite heat capacity.”*

L203: I spend more time than most glaciologists thinking about covariance, and yet I'm still unclear as to what's going on here. In particular, after fitting the ARMA model to each time series of TF, SMB, and runoff independently, how are spatio-temporal correlations between them calculated. Reading the appendix, it seems that there are three layers to this model: Fitting a piecewise linear function, fitting an ARMA model to each basin/variable, and then computing a big covariance matrix between the residuals for all? Okay, I guess, but I would like a more centralized and coherent justification for why this is a reasonable way to control the spatial relationships.
Yes, the reviewer is correct in their understanding of our spatio-temporal statistical modeling. The spatial unit level is the glaciological catchment. We have 253 catchments. For each of these, we have a time series of SMB, of runoff (for those catchments where runoff is non-zero), and of TF (for those catchments with a marine outlet). These time series are computed from the climate model outputs. Residuals from the variable- and catchment-specific ARMA models are correlated to compute the covariance matrix. Our covariance matrix is of dimension 676×676. We acknowledge that such a large covariance matrix is subject so spurious correlations when constrained using time series of only 354 yr (1850-2203). This limitation is discussed in the main text:

“The residuals are obtained after fitting the optimal ARMA model to a given time series, such that the stochastic component ε_t (see Eq. A2) is isolated. Isolating residuals allows to first

remove potential spurious correlations between catchments and variables that appear in the raw time series caused by temporal autocorrelation. This autocorrelation is removed from the ARMA residuals, and the remaining correlations found capture cross-spatial and cross-variable dependencies. However, because the number of entries in the correlation matrix to be estimated (1/2 × 676 × 675 = 228 150) is large compared to the number of yearly samples (354), we compute a sparse correlation matrix (Hu and Castruccio, 2021) with the commonly-used graphical lasso method (Friedman et al., 2008)."

The sparse covariance matrix is shown in Figure D1. We understand the concern of validity of this approach. However, we justify our choice of using the glaciological catchment as a compromise between (1) using a large enough spatial unit, and (2) maintaining spatial detail at a level that is physically meaningful for glacier flow. Indeed, using the individual grid-point as a spatial unit would not only drastically increase the computational complexity of covariance estimation, but would also lead to increased risk of generating spurious correlations. By averaging over glacier catchments, we increase the signal-to-noise ratio of correlations between spatial units. Averaging at a coarser level would be possible, but at the expense of losing some of the details in SMB, runoff, and TF that partly explain different behaviors of glaciers located within a same region.

The motivation for fitting the covariance matrix to the ARMA residuals is the following: we need to separate temporal autocorrelation (within each catchment and variable) from the spatial cross-correlation (between different catchments and variables). Raw time series often exhibit strong temporal dependencies, which can artificially inflate estimates of spatial correlation if not accounted for. By modeling and removing the temporal structure via ARMA fits, the residuals represent the remaining component of each time series after accounting for its own past behavior. Computing covariance on these residuals ensures that the estimated spatial covariance matrix captures true cross-spatial and cross-variable dependencies, after having removed the temporal autocorrelation. It also enhances the interpretability and stability of the covariance matrix used downstream in the model. We have added this explanation in the manuscript, as given in the text snippet above (*Isolating residuals* (…) *cross-variable dependencies*).

This approach follows the framework laid out in Hu and Castruccio (2021), where controlling for temporal structure improves estimation of spatial dependencies in spatio-temporal datasets. It has subsequently been applied to Greenland SMB by Ultee et al. (2024) and Greenland TF by Verjans et al. (2023).

Sec 2.2.3: I'm not completely sure that this is the right thing to do, but it might be helpful to lead the section with this (which is essentially the `physics'), so that the reader will have a better idea of what the TF, etc. is going to be used for. Similarly, you might include here the way that lapse-rates and such enter the SMB calculation.

This subsection has been moved to the beginning of section 2 and a description of the SMB parameterization has been added.

L264: I'm sympathetic to the need to use SSA for computational reasons, but it would be worthwhile to briefly describe the implications -- Greenland has a lot of ice that is very much not consistent with the assumptions of that model after all.
Two sentences have been added addressing this assumption: *“Using this approximation over the entire ice sheet may neglect deformation ice sheet flow, particularly in the ice sheet interior where ice is more likely to be frozen to the bed. However, over the centennial time scales considered in this study, these errors are unlikely to be significant on the scale of the entire ice sheet where most ice transport near the margins occurs via basal sliding.”*

Eq. 3 and lines after: Am I missing a previous point at which N is defined? How is it computed here? Constant fraction of overburden?
We have added a definition and explanation of N, effective pressure, here.

L271--273: Would it be possible to provide some additional justification with respect to the linear regression step described here? This isn't something I've seen before, so it would be nice to understand a little bit better how/whether this works.
Added citations and a justification for this approach: *“This is a common approach in ice sheet model simulations (Åkesson et al., 2018; Cuzzone et al., 2022) where advance onto currently ice-free portions of the bed may occur. It captures the general pattern that deep portions of the bed are likely to have accumulated deformable marine sediments when they were covered by ocean rather than ice.”*

Sec. 2.3.2: I am confused as to the technical approach for performing this calibration. Is this done by manually fiddling with σ_{max} until the eyeball norm is minimized, or is there an objective (and automated) procedure that is taking place?
For calibration, we performed a large sequence of calibration runs. Between every run, we evaluated the retreat rates of all glaciers, and increased/decreased σ_{max} of individual glaciers if their retreat had a positive/negative bias. This was performed until all glaciers reached a retreat rate within ±1 km of the observed retreat rate. For a subset of glaciers, this observational constraint could not be met while keeping σ_{max} to physically acceptable values (see Figure 3). Once we achieved this objective, we compared the total ice sheet mass loss to the IMBIE mass loss. We then increased/decreased σ_{max} at those glaciers where it was possible to still remain within the ±1 km individual glacier constraint in order to better match the IMBIE mass loss. The process was semi-automated: evaluation of retreat rates, of ice sheet mass loss, and adjustment of σ_{max} values were automated. However, calibration runs had to be manually re-configured and re-launched. That process was quite tedious and work-intensive.

L399: I am surprised that the assertion that the small deviation of the ensemble mean from the deterministic run is a result of noise-induced drift is not backed up by a statistical test. It would strengthen the argument to include a test of significance here.
You’re right that it is not significantly different, which is what was implied by the following sentence. We have re-arranged and modified these sentences to be clearer that this difference does not represent a statistically significant difference. This is because the deterministic simulation can be considered as a single “sample” which falls well within the distribution of the stochastic ensemble. To be clear about our thinking: a statistical

significance test (e.g., t-test) involves creating a statistic (e.g., t-statistic, if we are interested in the mean of a sample) that can then be compared to the expected distribution of such a statistic (e.g., t-distribution). In this case, we already have a sample (ice mass change in the deterministic run) and a distribution (ice mass change sampled by the stochastic ensemble members), which can be compared directly in a statistical inference problem, as we do in the current, now clarified, text. Thus, it should not be necessary to pursue more complicated statistical testing than this.

Here is the revised text:
"By 2203, the mean ensemble ice loss is 4% greater than in the deterministic simulation. This difference is just 1 standard deviation of the ensemble final mass change from the ensemble median, and 17 members have less ice loss than the deterministic run. Thus, while this difference with the deterministic run may be suggestive of potential noise-induced drift in the stochastic ensemble (Tsai et al., 2017; Hoffman et al., 2019; Robel et al., 2024), it is not large enough to be statistically significant (i.e., differences of this magnitude or larger occur by chance in 17% of ensemble members). We disentangle the mechanisms of this drift more fully in Section 3.3 with comparison to small ensemble experiments."

Fig. 4d: This is a challenging metric to use in order to assert the relative importance of uncertainty because the denominator gets so very small close to the start of the simulation period. I am not sure what the alternative is, but it might be helpful to acknowledge that.
Yes, this is true. We have added a sentence acknowledging this: *"We note that the denominator in this ratio starts near zero and so should be interpreted with caution, though it does help us to understand the relative importance of ensemble spread as compared to ensemble mean change."*

L465: delete `briefly'.
Deleted

L544: This is a pretty awkward sentence -- suggest rephrasing.
Split into two sentences are reworded

Discussion: I appreciate the comparison to both Tsai (for the forcing uncertainty comparison) and ISMIP6 (for the model uncertainty comparison), but it might be useful to also compare to some of the previous works that explore parametric uncertainty -- which seems to be of similar size to model uncertainty in some cases. Would using randomly sampled climate-to-SMB parameters drown out the influence of the stochastic climate? This would be important to know in making a decision about whether to include stochastic forcing in, say, ISMIP7.
This is a nice suggestion. Aschwanden et al. (2019) is probably the most comparable single-model, parameter perturbed ensemble study. We have added a paragraph in the discussion drawing out this comparison, particularly with respect to what it means for designing ensembles to quantify uncertainty in future Greenland ice loss:

"Aschwanden et al. (2019) provides another useful point of comparison to GrISLENS. In that study, parameters for ice flow and forcing parameterizations were perturbed over 500 ensemble members with an ice sheet model of sufficiently high-resolution (1 km) over

Greenland to resolve individual outlet glaciers. They found consistently much greater mass loss compared to WARM-LE, with median sea level contribution of 22 cm in 2100 (compared to 11 cm in WARM-LE) and 103 cm in 2200 (compared to 25 cm in WARM-LE). The ensemble spread in Aschwanden et al. (2019) is also much greater than in WARM-LE, with a standard deviation of 10 cm SLE in 2100 and 50 cm SLE in 2200, about 2 orders of magnitude greater than the WARM-LE ensemble spread. Later work to calibrate these ensembles using observations reduced the median and spread of their ensemble (Aschwanden and Brinkerhoff, 2022), but the broader higher sensitivity and spread compared to WARM-LE remain. Decomposing the parametric uncertainty quantified in their ensemble using Sobol indices, they find that at 2100, uncertainty ice flow and surface melt parameters contribute the most to uncertainty in total ice sheet mass loss. At 2200 and beyond, uncertainty in the sensitivity of mean atmospheric temperatures to emissions forcing (i.e., climate sensitivity) dominates uncertainty in total ice sheet mass loss. Consistent with our study, they find that uncertainty in ocean forcing plays a relatively minor role in driving uncertainty in total ice sheet mass loss. We thus conclude that even for a given large-scale climate forcing, uncertainty in parameters that govern how SMB is calculated from large-scale climate models is currently large enough to substantially exceed uncertainty from variability in SMB in terms of the resulting influence on total ice sheet mass loss."