Response to Reviewers

In the response to the reviews of the manuscript entitled: Ensemble Generation for Seamless Prediction in the GEOS-S2S Forecast System, we include here in italics each of the reviewer's comments and follow it in regular text with our response. We believe that we have addressed all of the reviewer's points, have modified figures and added a table in response to the reviewer's suggestions. We thank the reviewers for the comments and suggestions.

Reviewer 2:

Reviewer 2 - General Comments:

This study presents the Synchronized Multiple Time-lagged (SMT) approach to reduce the large forecast ensemble after a specified lead time, which involves performing a stratified sampling of the early larger ensemble in a way that accounts for the emerging directions of error growth. This is quite interesting and very useful for guiding the efficiency of ensemble forecast experiments in both research and operational practices. However, the authors emphasize that it is a follow-up study of Schubert et al. (2019), and it lacks a clear justification for the specific contribution of this work. It only tests the method applied to Niño3.4 SST, and remains to address the generalizability of this method when applied across/representative different latitude belts or large-scale weather teleconnection signals. Lastly, even though it is understood that the authors could only test/focus on the GEOS-S2S system by comparing GEOS-S2S-2 and GEOS-S2S-3 due to some logistical constraints, a more detailed analysis of the observed differences between both GEOS-S2S versions is lacking, which is important to address for the contribution of this study. Therefore, a major revision is recommended before considering publication in Geoscientific Model Development.

We thank the reviewer for the time and effort spent evaluating this manuscript and for the helpful comments and suggestions. As will be elaborated in the response to specific comments below, we have revised the introduction of the manuscript to make it clear that the manuscript presents two distinct aspects of the new ensemble perturbation method, the SMT that entails the perturbation patterns and magnitudes, and the clustering to reduce the number of ensemble members after 3 months of forecast. In addition, the manuscript has been revised to include the relevant information from the Schubert et al (2019) technical memorandum so that this presentation of the method can stand on its own. Finally, although a comprehensive comparison of GEOS-S2S-2 and GEOS-S2S-3 forecast skill is beyond the scope of the present manuscript (and is part of a manuscript in preparation), some relevant results of the comparison is reported here in the revised manuscript.

Reviewer 2 - Major Comments:

(1) The authors emphasize that it is a follow-up study of the work of Schubert et al. (2019), and repeatedly cite Schubert et al. (2019) in many places. For instance, in lines 105–107, the authors stated that "The motivation for improvement of this ensemble perturbation strategy and the development of the new strategy detailed here is described in Schubert et al. (Section 2.4 of 2019)." It is recommended that the authors reorganize this study and clarify the motivations and the specific contribution of this study.

The manuscript was revised to more clearly state the motivation for each element of the new algorithm, and now includes some details from the Schubert et al. (2019) NASA technical memorandum. In particular, the introduction was expanded and rearranged for clarity, and the section describing the algorithm was rearranged and the missing information about the motivation for the choice of scaling and the goals of the subsampling was added.

(2) The stratified sub-sampling strategy is exclusively tested using the Niño 3.4 SST index at extended forecast leads. To demonstrate the broader applicability of the SMT method, it is highly recommended—and necessary—to include evaluations against other prominent teleconnection processes or mid-latitude climate signals across different latitude belts.

As mentioned in response to this reviewer's general comments, the SMT method is the part of the new algorithm that includes the choice of perturbation eigenvectors and the choice of scaling, and the clustering is the part of the method that reduces the ensemble size after three months to exploit the longer lead predictability in the tropics. As the expectation, as now articulated more clearly in the introduction, is for mid- and high- latitude predictability to substantially decrease after 3 months of forecast, the evaluation of the clustering is performed in the tropics only. The evaluation of the SMT and the increased ensemble size will be reported in detail in another study, however, some salient results from this evaluation in the mid-latitudes are reported here, specifically as relates to the prediction skill of the North Atlantic Oscillation (NAO).

(3) Given the focus on the Niño 3.4 SST index, the authors should expand on how these stratified sub-sampling strategies can practically improve current long-lead ENSO forecasting (e.g., impact on ensemble spread and forecast error). This section should be supported by a deeper discussion of recent literature and technological advancements in the field.

The discussion of the Niño 3.4 index performance in GEOS-S2S-3 using the new ensemble perturbation method has been expanded. We have revised the R-ratio figure for clarity, and now discuss the initialization months and lead times for which there are both improvements and degradations in the R-ratio and the possible reasons for the change in behavior.

(4) The authors emphasize the distinctly different ensemble spread patterns as a function of forecast lead times and initialization periods in GEOS-S2S-2 and GEOS-S2S-3 (Lines 289– 296, Figure 9). I am wondering if this is an overly emphasized statement. As shown in Figure 9, the general pattern (e.g., over-dispersive in lead months 1–4 when initialized in Jan–Mar, and in lead months 5–9 when initialized Jul–Dec) is observed similarly in both model versions, though with a difference in magnitude. It is recommended that the authors provide a more in-depth comparative analysis of both versions.

As discussed in the response to this Reviewer's previous comment about the Niño 3.4 SST index results, the figure has been revised for clarity and the discussion has been revised to better reflect the differences between GEOS-S2S-2 and GEOS-S2S-3 behavior and now includes mention of possible reasons for the differences.

Reviewer 2 - Minor Comments:

1. Line 32: Please provide a brief explanation or context regarding the specific breeding approaches used in this study.

The line is part of a brief review with citations of the different perturbation methods used in the community. There is no breeding method used in the present study. The text in line 32 was modified to better elucidate the purpose of the references to the other methods in use in other centers.

2. Line 84: Please elaborate on what the "few key upgrades" entail to give the reader better technical context.

Text has been added to the manuscript to elaborate on the differences between GEOS-S2S-2 and GEOS-S2S-3. As now explained in the manuscript in Section 3, the important GEOS-S2S-3 system upgrades include an increase in ocean horizontal and vertical resolution, a more physically based Atmosphere-Ocean interface layer, a proper routing of glacial melt to the ocean, a change to a more dynamically stable "replay" method to constrain the atmosphere to the pre-computed assimilation, and a proper initialization of the aerosol fields. As suggested by this reviewer in the Minor Comments below, a table has been added to summarize these and other key differences.

3. Line 89: Please correct the editing error "temperqture" — it should be "temperature." Does the atmospheric state variable include mass? What components are included in the aerosol field?

Done

4. Lines 98–99: Does the phrase "last four start dates of the month" translate to the "last four days of the month" in the default GEOS-S2S-2 setup? This seems to conflict with lines 94–95, which state that forecasts are initialized on a fixed set of calendar dates (every 5 days starting January 1) across all GEOS-S2S systems to date. Please clarify if this applies to GEOS-S2S-2 or GEOS-S2S-3, and provide the justification for changing initialization dates.

Apologies for the confusion this statement caused. The phrase "last four start dates of the month" does not refer to the last four days of the month, but to the last four "start dates" based on our fixed calendar. We have modified the text to include the phrase "the last four start dates in each month based on our fixed calendar".

5. Model Comparisons: To improve readability, it is recommended to include a summary table outlining the key technical similarities and differences between GEOS-S2S-2 and GEOS-S2S-3.

A table has been added to list the key technical similarities and differences between GEOS-S2S-2 and GEOS-S2S-3.

6. Lines 146–147: Please provide an explanation for the seasonal variability observed in the average scaling factors in Figure 2. Additionally, how does the seasonal variability of these scaling factors behave within the ocean at various depths?

Although we have not determined the cause of the seasonality in the scaling factors, we

have added a brief note in the manuscript with the speculation that the varying strength and importance of the different modes of variability in different seasons may contribute to the seasonality of the scaling factors. We found that at depth (below approximately 300m depth) in the ocean there was little to no seasonality in the scaling factors, which would be consistent with the speculation that the observed variance of the different modes is relevant.

7. Line 161: Please correct or clarify the phrase "Over say".
The sentence was rephrased to avoid this colloquial syntax.

8. Lines 160 / 165: In Equation 2, τ is defined as the lead time, whereas in Equation 1 (line 134) it is defined as the separation time. Please clarify if these represent the same physical quantity or correct the notation.
The lead time and separation time refer to the same gap between analysis states used as the eigenvectors. For clarity, the manuscript now only refers to "separation time"

9. Line 174: Change "ie" to the standard abbreviation "i.e."
Done.

10. Line 191: Please specify which leading EOFs are included (e.g., EOF1, EOF2). What percentage of the total global variance do they account for, and are they statistically significant?
The information about what percentage of the total variance that each EOF shown represents was added to the text of the manuscript, along with the information about statistical significance. The figures all depict EOF1 and represent variance just under 20% in general, and the variance is statistically significant.

11. Lines 270–271: What physical or numerical mechanism causes the rapid drop in correlation (R) at the second forecast month when sub-sampled based on the second month of integration (Figure 8, right panel)?

The explanation of the metric presented in Figure 8 was expanded and better articulated in the revised manuscript. The metric is the ratio of the stratified sampling variance to the variance estimated based on random sampling. The goal of a successful stratified sampling is to minimize this ratio. So the right hand panel of Figure 8 shows a small drop in this ratio at the second lead month when the stratification is performed at that lead month, and a gradual increase, indicating that the ratio remains below 1 well into the forecast. This right hand panel shows a contrast to the left panel, with the lowest ratio in the stratified sampling month and a more rapid increase of the ratio, indicating a poorer quality of stratified sampling. This information is now contained in the text of the manuscript that describes Figure 8.

12. Lines 276, 278: Please correct the section numbering typos ("Section 33.2, 33.1").
The section numbering was corrected.

13. Lines 274–280: It is recommended to consolidate the experimental details for GEOS-S2S-3 into a structured format, potentially combining them with Table 1.

Table 1 was separated into Table 1a (Table 2a in the revised manuscript) to describe the experimental suite used to evaluate and test the SMT details, and Table (now Table 2b) to describe the experimental suite used to compare GEOS-S2S-2 and GEOS-S2S-3 in Section 4.

14. Lines 290–296: Please explicitly define the operational thresholds for what constitutes "shorter," "intermediate," and "longer" lead times.
We have rephrased the text to be more specific about lead time months, and removed this subjective vocabulary.

15. Figure 5: The text notes that Figure 5 is based on MERRA-2 data. Please provide the context or rationale for introducing MERRA-2 at this specific stage of the manuscript.
MERRA-2 is mentioned here due to its use to compute the lower frequency tropical EOFs (10-day lagged structures) relative to the synoptic scale EOFs that were computed from our GEOS-S2S-2 Atmosphere-ocean coupled assimilation. This difference was made clear in the text of the manuscript in Section 3.

16. Figure 8: Please verify that the mathematical equation displayed in the title matches Equation 6 in the main body of the text.
The mathematical equation was removed from the Figure.

17. Figure 9: The current caption makes it difficult to distinguish the specific ensemble datasets used for GEOS-S2S-2 and GEOS-S2S-3. Please expand the figure caption to include a detailed description of the exact datasets employed.
The expansion of what was originally Table 1 now includes a list of the forecasts used to generate Figure 9. The description was removed from the Figure caption.