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 1:

Reviewer 1 - General Comments:

This manuscript presents the Scaled Lagged Mean Technique (SMT), an ensemble generation method implemented in the NASA GEOS-S2S-3 forecasting system for subseasonal-to-seasonal (S2S) prediction. The topic is well aligned with the scope of GMD and is of interest to the S2S community. The manuscript provides a comprehensive description of the proposed methodology. However, I found that several aspects of the methodology, experimental design, and interpretation of the results require additional clarification. In particular, the manuscript relies heavily on previous publications for key methodological details and would benefit from a more self-contained presentation of the SMT methodology and a clearer demonstration of its impact on ensemble quality and forecast performance.

We thank the reviewer for the time and effort spent evaluating this manuscript, for the recognition of the contribution of this work and for the helpful comments and suggestions. Upon re-reading the manuscript after reading the reviews, we recognize the need for additional clarity regarding the details of the new perturbation scheme and its evaluation. We thank the reviewer for elucidating the gaps. As detailed below, we have revised the presentation of the methodology and results to better include the relevant information provided in the (non peer-reviewed) technical memorandum.

Reviewer 1 - Major Comments:

(*) The manuscript frequently refers readers to Schubert et al. (2019) for key aspects of the SMT methodology, including the perturbation scaling, theoretical basis, and implementation details. While referencing the earlier work is appropriate, the current manuscript would benefit from a more self-contained description of the key concepts and technical innovations. A concise summary of the underlying rationale, the main algorithmic steps, and the motivation behind the scaling approach would help readers establish the necessary background without repeatedly consulting the previous report. This would also better highlight the novelty and practical implementation of the SMT approach within the GEOS-S2S-3 system.

Additional details that previously appeared only in the Schubert et al. (2019) NASA internal technical memorandum were added to each section of the present manuscript. Specifically:

* The introduction was revised to more clearly state the motivation for each element of the new algorithm.

* 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.

Additional modifications to the manuscript to include a more self-contained description are detailed in the responses to the reviewer's other major comments below.

(*) The manuscript provides a detailed description of the coupled data assimilation system because the SMT perturbations are derived from AODAS analysis states. However, the relationship between the data assimilation system and the proposed SMT ensemble generation could be clarified further. Specifically, I have two questions. (1) How strongly does the SMT approach depend on the underlying data assimilation system? Could SMT also be applied in prediction systems that do not employ a coupled data assimilation system, or is it fundamentally tied to the availability of analysis states produced by AODAS? (2) The ocean data assimilation uses an LETKF-based ensemble framework. Is this assimilation ensemble also used for forecast initialization, or is SMT applied only after the assimilation cycle to generate a separate forecast ensemble? A brief discussion clarifying the relationship between the LETKF ensemble, the analysis states, and the SMT-generated forecast ensemble would help readers better understand the methodology. A simple workflow schematic describes the workflow of SMT within the GEOS S2S-3 system could be helpful to improve the clarity of the paper.

We have modified the text in section 2 to articulate the relationship between the AODAS and SMT, in particular to elucidate the two questions raised here. Since the SMT is applied only after the AODAS has been completed and before the forecasts are initiated, we deemed that the clarity of the manuscript would not benefit from a schematic of this simple work flow and have not added one. In answer to the two questions posed:

1. The SMT approach could be applied by using atmosphere-ocean and ocean-only analysis states, as the perturbations in each fluid are independant of each other. The choice of an AODAS for GEOS is based on other considerations not related to the perturbations for the ensemble forecasts.

2. The LETKF as applied in GEOS-S2S-3 (and S2S-2) uses a static ensemble, so there is no assimilation ensemble. Future plans (for GEOS-S2S-4) call for the inclusion of a dynamic assimilation ensemble, though the ensemble perturbations for the assimilation and for the forecast will remain independant. The introduction was revised to more clearly state the motivation for each element of the new algorithm.

(*) Line 135, Eq. (1): Although the derivation of Eq. (1) is provided in Schubert et al. (2019), it would be helpful to include a brief explanation of its physical and mathematical motivation. For example, the authors could briefly explain that the scaling is based on the temporal autocorrelation of the analysis differences and is designed to normalize perturbation amplitudes relative to the climatological variance, making them approximately independent of the chosen separation time. This would improve readability for readers unfamiliar with the original reference. In addition, the practical application of this method depends on the choice of ϵ. Is there an objective criterion or practical guidance for selecting its value? For example, how can one determine whether ϵ = 0.1 or 0.2 is sufficient to generate an appropriate

ensemble spread for S2S forecasts? It would be useful if the authors could discuss how ϵ was selected and relate this choice to the results presented in this section. Finally, is the scaling factor α variable dependent? Figure 3 appears to suggest that it is. If so, the statement in the abstract that "The scaling is also designed to produce perturbations that are small relative to the temporal variations of the quantity in question, retaining the dynamical balance of the unperturbed state" could benefit from further clarification. It is not immediately clear how the proposed scaling guarantees that dynamical and physical balances are maintained, particularly when different variables may be scaled differently. Moreover, this property appears to depend on the choice of ϵ. If ϵ is sufficiently small, any imbalance introduced by variable-dependent scaling may remain negligible. However, larger values of ϵ could potentially alter the multivariate relationships among variables (e.g., U , V , T , and q), thereby affecting the dynamical balance of the initial state. This raises the practical question of how ϵ should be determined or calibrated to minimize the introduction of model imbalance while still producing sufficient ensemble spread for S2S forecasts. A brief discussion of the assumptions underlying this design, together with any practical guidance or limitations regarding the choice of ϵ, would strengthen the manuscript.

The presentation of the algorithm in Section 3 has been reorganized, and the physical and mathematical motivation for different aspects of the SMT and clustering elements have been added. In particular:

* The section now begins with the description of eigenvectors of the ensemble perturbations based on examples of n-day apart differences, pointing out the qualitative resemblence to well known modes of variability that are important for prediction at subseasonal to seasonal scales. We believe that this provides some of the motivation for the decisions made as part of the scaling.

* The information about the motivation for the scaling is articulated more clearly. We have added text that explicitly states that the scaling should (i) normalize the amplitude of the perturbation relative to the climatological variance so that the perturbations are relatively independant of separation time, (ii) remain small enough so as not to generate a dynamical imbalance.

* Additional details regarding the parameter choices have been added. We now explain that the factor α depends on separation time and season alone, and that there are distinct values for atmospheric and oceanic fields. We have explained that we chose an ϵ=0.1 to minimize the initial shock (due to dynamical imbalance), and examination of the first days of a small sample of forecasts revealed no shock in either precipitation or ocean vertical motion. We have provided a reference for the character of the shock we saw no evidence of.

(*) Section 3: there are several questions that may need clarifications and extra analysis: o Figure 4 provides an interesting comparison between the leading EOFs of the 1-day and 5-day analysis differences. However, the physical interpretation of these results is not entirely clear. In particular, the manuscript states that the 1-day differences exhibit synoptic-scale structures, whereas the 5-day differences resemble teleconnection patterns such as the PNA,

but does not explain the underlying reason for this transition. I am somewhat confused by the interpretation of the PNA-like pattern, since the EOF is presented on a global map while the PNA is generally regarded as a regional teleconnection over the North Pacific–North American sector. It would be helpful if the authors could clarify why the leading EOF is described as PNA-like and, if appropriate, include a reference pattern or additional discussion to facilitate interpretation. Would it be useful for authors also show a reference plots to help reader obtain what is this for? It is not also clear why this transition occurs or why it is important for the SMT methodology. Some extra background and discussions should be helpful to help readers understand what the intention is here. o Figure 5: It is not entirely clear whether the results are derived from GEOS-S2S forecasts, analysis fields, or the MERRA-2 reanalysis. The figure caption states, 'Results are based on MERRA-2 for the years 1999–2016,' which I found somewhat confusing. A brief clarification would be helpful. o It is not entirely clear how the analyses presented in Figure 6 support the development or evaluation of the SMT methodology. A brief discussion explaining the purpose of this analysis and its relevance to the ensemble generation strategy would improve the presentation.

As stated above, Section 3 has been reorganized and additional explanations and clarifications have been added to better articulate the reasons for the discussion and presentation of the different EOFs. In Particular:

* We apologize for characterizing the EOFs of the 5-day apart differences as a PNA pattern. In the region of the PNA it does resemble that structure, but as the figure depicts a global pattern we have rephrased the manuscript to describe this as a planetary or stationary wave structure.

* The text in Section 3 has been modified to include an explanation of why we expect the 1-day, 5-day, 10-day apart differences to behave as they do, and the connection we are making between these structures and the 'fastest growing modes' we are aiming to perturb as they provide a set of initial states for the ensemble to generate a seamless and more reliable forecast.

* We have made clear from the start of this section where the data for all the figures are obtained from. The EOFs for the synoptic, planetary and oceanic eigenvectors were computed from our GEOS-S2S-2 coupled assimilation that spanned 2003-present. For the tropical larger scale structures we chose to use MERRA-2 as a longer period was deemed neccessary for these EOFs.

(*) Section 4: Similar to Section 3, some extra clarifications and extra analysis is needed o This section presents the experimental results without providing sufficient background information to help readers understand the methodology and the purpose of the experiments. While the manuscript cites the relevant references, it would be beneficial to briefly summarize the key aspects of the GEOS S2S-2 ensemble generation strategy and forecasting system that are directly relevant to this study, rather than relying solely on the cited literature. In particular, the authors should provide sufficient details on the experimental design used to generate the results shown in Figure 9. o Figure 9: I am somewhat confused by the interpretation of the results shown in Figure 9. Based on the definition of the metric, values

close to 1 indicate that the ensemble spread appropriately represents the forecast uncertainty, whereas values greater than 1 indicate an over-dispersive ensemble and values less than 1 indicate an underdispersive ensemble. Since one of the primary objectives of this study is to evaluate the SMT ensemble generation method, I would expect the discussion to focus on whether the new ensemble exhibits improved reliability (i.e., values closer to 1). From visual inspection, it is not immediately obvious that the GEOS-S2S-3 ensemble exhibits improved reliability relative to GEOS-S2S-2. It would therefore be helpful if the authors could provide a quantitative summary (e.g., a domain-averaged metric or the fraction of grid points within a specified range around unity) to support the interpretation. In addition, the current color scale makes the differences somewhat difficult to distinguish, and a revised color map or quantitative summary could improve readability. More importantly, the manuscript mainly describes the spatial features of Figure 9 but does not draw a clear conclusion regarding the performance of the SMT ensemble generation strategy. Later sections discuss differences in hindcast skill between GEOS-S2S-2 and GEOS-S2S-3; however, these differences may result from multiple factors, including model physics, resolution, and other system developments, in addition to changes in the ensemble generation strategy. Consequently, it is difficult to isolate the impact of SMT from the current presentation. Since the focus of this study is "Ensemble Generation for Seamless Prediction in the GEOS-S2S Forecast System," it would strengthen the manuscript if the authors could provide provide a more direct evaluation of the impact of SMT on ensemble reliability and forecast skill, or discuss to what extent the observed differences can reasonably be attributed to the ensemble generation strategy rather than other model developments.

Section 4 has also been modified to provide more background about the experiments used to generate the results in this section. We have added the information that the experiments used to evaluate the spread/error ratio were totally different from the experiments used in Section 3 (and appear in Table 1) to evaluate the sub-sampling strategy. The comparisons here were made using the complete retrospective forecast suites performed with GEOS-S2S-2 and GEOS-S2S-3. A limited suite of experiments for this more emcompassing purpose seemed inadquate. A more detailed description of the systems (GEOS-S2S-2 and GEOS-S2S-3) used to generate the experiments was added to Section 2.

The presentation of results in Section 4 has further been modified to include a new figure of the R-ratio. The new figure better illustrates the difference between S2S-2 and S2S-3 ratios, and points out where the new system's R-ratio results are closer to 1 or further.

As we present results in this section from suites of forecasts that include differences other than the ensemble perturbation strategy (model physics and resolution, data assimilation details, ocean data assimilated) an absolute and clear attribution of differences as being due to the ensemble perturbation strategy is not possible. Section now also includes some clear conclusions about the R-ratio and the implications for reliability of GEOS-S2S-3 as compared to GEOS-S2S-2, and some speculation about attribution of the differences based on differences in the components of the R-ratio (ensemble spread and ensemble mean error).

Reviewer 1 - Minor Comments:

1. Line 22: "correspond" should be "corresponds."

The sentence was modified. It now reads "...for the forecast probability to correspond to..."

2. Line 89: "temperqture" should be "temperature."
Done.

3. Line 162: there is no variables come with "<>" in Eq.(2).
The statement defining the angular brackets has been removed from Eqn. 2 and placed after Eqn 3.

4. Line 277: "Forecast are initialized..." should be "Forecasts are initialized..."
Done.

5. Lines 304–306: The sentence beginning with "These results with high skill of..." reads awkwardly (e.g., "These results with high skill of ... are results are also reflected in ..."). Please consider rephrasing it for clarity.
The sentence has been rephrased.

6. Line 276: The reference to "Section 33.2" appears to be incorrect. Similarly, "Section 33.1" in line 278 should also be checked.
The section numbering was corrected.

7. Line 279: Add a comma (or semicolon) after "OSTIA."
A comma was added.

8. Figure 4: the map plot are cutoff with a blank box in the lower middle. Also, the color bar are overlapped and hard to see clearly.
The figure has been modified to improve the clarity and quality.

9. Appendix: There appears to be a formatting issue in Eq. (A1), with an extra "(". In addition, Eqs. (A4) and (A5) appear to define the same quantity twice; please check whether this duplication is intentional.
Both equations have been corrected.

10. Author Contributions: Add a period at the end of the paragraph.
Done.