Review Report on egusphere-2026-1340

Summary

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.

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.

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.

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.

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.

Minor Comments

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

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

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?

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.

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.

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?

7. Line 161: Please correct or clarify the phrase *"Over say"*.

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.

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

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?

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)?

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

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.

14. Lines 290–296: Please explicitly define the operational thresholds for what constitutes "shorter," "intermediate," and "longer" lead times.

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.

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

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.