Response to Referees of "Stratospheric geoengineering impacts on El Niño/Southern Oscillation," by C. J. Gabriel and A. Robock

Referee comments are in black. Responses are in blue

Referee #3

1) The manuscript seeks to identify changes in ENSO frequency and amplitude under various historical, projected, and geoengineering scenarios. The paper is well motivated and clearly written. That said, there are major caveats associated with limitations in available experiments that go under-appreciated in the text and greatly limit what can be done. I find the results therefore somewhat unconvincing. I recommend strongly that the manuscript provide a more direct appreciation for the inherent limits of the simulations used and provide better context for what is needed to really address this questions with greater certainty.

We agree. We have added a discussion of the inherent limitation of our experimental design to the introduction. We hope that the additional context provided will not only clarify our results, but provide a road map for what types of future simulations would be most useful in detecting potential changes in ENSO variability under various geoengineering regimes.

"Detecting changes in ENSO variability is notoriously difficult. The use of lengthy simulations, multiple models, and ensembles is often employed. Cai et al. (2015) were able to detect a statistically significant change in the frequency of extreme La Niña events under RCP 8.5 as compared to a non-global warming control scenario. They selected 21 of 32 available CMIP5 models, because of their ability to accurately simulate processes associated with extreme ENSO events. Each model simulation lasted for a period of 200 years. The detectability of changes in ENSO variability in future SRM modeling experiments will likely be buoyed by the availability of more models and longer simulations. Additionally, future SRM experiments that attempt to offset or partially offset more extreme AGW scenarios, such as RCP 6.0 and RCP 8.5 improve detectability."

Cai, W., Wang, G., Santoso, A., McPhaden, M., Wu, L., Jin F-F., Timmermann, A., Collins, M., Vecchi, G., Lengaigne, M., England, M., Dommenget, D., Takahashi, K. and Guilyardi, E.: Increased frequency of extreme La Niña events under greenhouse warming. Nature Climate Change 5, 132-137, 2015.

2) The experiments used are useful but fail to provide a very tight constraint on the null hypothesis being posed because of the facts that 1) so few ensemble members are available for each model, and 2) ENSO is so poorly and variably simulated across the models (as alluded to in 3.3)- contributing to large error bars and hence coarse detectability of any potential change. There is considerable uncertainty associated with the changing mix models across the metrics being computed that is not adequately dealt

with. It would seem to be essential to me to make this weighting constant across any comparisons being made.

In the revised manuscript, each model included in each comparison group is weighted equally. Additionally, we have eliminated the aggregation of multiple experiments that produce large ensembles, but also mix experiments with dissimilar climates into the same ensemble. Please see the revised section 3.2. In all comparisons, each model is weighted equally in both the experimental and control runs. If there are three experimental ensembles for a particular model, there are also three control ensembles for that model.

3) I recommend that manuscript start with the most simple question: in a single scenario and for a single model, do any provide enough ensemble members to detect a change in ENSO? I presume the answer is 'no', since it is not dealt with - but pointing this out would be useful for motivating the need to create multi-model ensemble metrics. In cases where significant differences are identified - what is the role of changes in the model mix?

We agree that pointing this out would be valuable. We have added this discussion to start section 2 Methods (p9 lines 226-231):

"We begin with the simple question of whether or not, in a single GeoMIP participating model that simulates ENSO well, a difference in ENSO amplitude or frequency is evident. Unsurprisingly, given the large inherent variability in ENSO, such a change is not detectable in one model. Given that, we adopt an approach in which we use output from nine GeoMIP-participating GCMs, each running between one and three ensemble members, of each experiment G1-G4. "

4) Please put +- 1-sigma values on the model-mean numbers. I think this provides essential context

± sigma values have been added.

5) Detection of only two significant results in the context of the large # that have been done is at the limits of what may be expected by chance. An associated caveat should be added here.

As mentioned above, we have eliminated comparisons in which we combine experiments that depict climates that are dissimilar. This limits the number of comparisons performed. Only two significant results remain and they are at 90% confidence. However, a simple resampling with replacement technique revealed that the significant result was likely actually the result of chance.

6) Why would you put more significance on the RCP4.5 finding that ones that have assessed the question in a broader array of models?

We have rewritten the abstract so that it does not call attention to RCP 4.5. Not mentioning the other experiments with equal weight in the abstract was likely an oversight by the author. Please see section 3.2, as it has been rewritten. Any special emphasis on the RCP 4.5 finding is no longer present.

7) Doesn't the fact that SOI is not a useful ENSO proxy speak to the inherent deficiency of using a given model for this type of analysis? How can one expect to get a reasonable bearing on the dynamical-thermostat mechanism or other dynamical links of forcing to ENSO if the SOI relationship is so poor since essentially the dynamic component of ENSO (SOI) also so poor? Shouldn't this be an additional constraint on which models to use?

The area of high correlation (> 0.5) is suppressed in models relative to observations. However, the region of highest correlation is in the correct location. The spatial pattern is similar, but the value of the correlation coefficient is muted. Also, as can be seen in Figure 3, the temporal relationship between SST and SOI is realistically simulated. The strong ocean-atmosphere coupling that is evident in the heart of the immediate equatorial central and eastern Pacific shows that the models are depicting a plausible ENSO cycle, albeit over a smaller area. In models that were excluded, the SOI SST correlation was plainly unrealistic, rather than just muted in spatial extent. Further, we sampled several SST-SOI correlations from our analysis, looked up the maximum value of the correlation, and found that it was around 0.8 in both observations and the models sampled. Therefore, it seems that the most robust ocean-atmosphere coupling was occurring, just over too small an area. This distortion of the spatial pattern was substantial enough go that SOI could not simply stand-in for SST, but it was not cause to discredit the depiction of ENSO dynamics in the models we used.

8) How can one establish confident ENSO statistics from such a short duration/limited ensemble of runs? Model runs suggest that robust statistics of ENSO (particularly at its low frequency tails) require records of over a century. What has been done here (to group all of them together) might be justified if they all had the same ENSO statistics but clearly they do not.

We agree that detectability of changes in ENSO may be inhibited by the unavailability of more lengthy simulations. We have analyzed 150 years of historical simulations for all ensembles within each model. We have also now done further analysis to determine if differences in ENSO variability between geoengineering and control runs during the 40 year period is greater than or less than the differences between selected 40 year periods in the historical data. This has allowed us to determine that ENSO variability between experiments is less than that seen when comparing 40 year slices of the historical period with each other. Therefore, based on our findings, ENSO variability under geoengineering as compared to under AGW would not exceed the variability found within the historical record.

9) The fact that some models have unrealistic ENSO behavior is hardly a new result and I don't think it requires 2 figures. A mere sentence in the text would suffice. Moreover,

internal variability of ENSO could lead to periods of such low variability even with a reasonable ENSO and thus I'd base any such statement on multiple ensemble members or an extended control.

The results shown in the two figures mentioned are selected because they are typical of what is seen across most of the models used. The purpose is to show that the ocean-atmosphere coupling is muted in terms of its spatial extent, but similar to observations in terms of the maximum value of SST/SOI correlation. This negates the possibility of using SOI in place of ENSO 3.4 SST as an ENSO index. Also, the SST/SOI correlation result is relevant to how ENSO is depicted in the simulations used. The figures are intended to place the results about ENSO variability in the context of the underlying physical mechanism. We do not assert that the SST/SOI correlation demonstrates for the first time that some models depict unrealistic ENSO behavior.

Additionally, the figures augment the discussion of the ocean-atmosphere coupling that is an essential part of ENSO. A visual depiction of the dynamics of ENSO as provided in Figures 2 and 3 may clarify the discussion for some readers.

10) Maximum event magnitude (e.g. Fig 9) doesn't seem like a very robust metric to use given the limited length of these runs. Why not use total variance?

Maximum event amplitude, mean event amplitude and total variance were all considered for use in the figures. The results pertinent to event amplitude would have ranked the models and experiments similarly. The readers are likely familiar with the magnitude of the strongest and weakest ENSO events in the observational record and reporting results in terms of maximum event magnitude is done for clarity.

11) On the discussion: We already knew changes in ENSO were inconsistent across models (e.g. Guilyardi et al 2012). This is not new It is likely that additional model runs should have been rejected based on the importance of dynamics in the science questions being posed and the lack of SOI fidelity. It seems odd that the authors used this as a basis for rejecting the SOI rather than the models! Perhaps a dynamical validation combined with a power spectrum validation would be a more appropriate way to screen models.

The area of high correlation (> 0.5) is suppressed in models relative to observations. However, the region of highest correlation is in the correct location. The spatial pattern is similar, but the value of the correlation coefficient is muted. Also, as can be seen in Figure 3, the temporal relationship between SST and SOI is realistically simulated. The strong ocean-atmosphere coupling that is evident in the heart of the immediate equatorial central and eastern Pacific shows that the models are depicting a plausible ENSO cycle, albeit over a smaller area. In models that were excluded, the SOI/SST correlation was plainly unrealistic, rather than just muted in spatial extent.

12) The question of whether the 1966-2005 period is really adequate to validate modeled ENSO is never addressed but needs to be considered. ENSO statistics varied considerably through the course of the 20th C.

We have now analyzed the full historical period of approximately 150 years for each model. We took a number of 40-year time slices from the 150-year historical record and created ensembles to test the variability of 40 year ENSO statistics. Variability between 40 year periods in the historical record was at least as large as that seen between geoengineering and AGW simulations. (Please see 3.2 analysis, which has been rewritten) None of the time slices were significantly different from each other at a 95% confidence level on any of the metrics tested. ENSO frequency in the 1966-2005 period was very similar to that seen in the full 150-year record. However, the 1966-2005 time period failed to capture the strongest warm and cold events in many of the 150-year historical periods. It is somewhat reassuring that the maximum warm and cold event amplitudes in the observational record was almost identical to the average maximum warm and cold event amplitudes in the simulations.

Given the paucity of observations of SST over the Niño regions prior to 1960, we see less value in matching pre-1960 historical simulations to observations.