Review of "Stratospheric Nudging And Predictable Surface Impacts (SNAPSI): A Protocol for Investigating the Role of the Stratospheric Polar Vortex in Subseasonal to Seasonal Forecasts" by Hitchcock et al.

General comments

This manuscript describes an experimental protocol for multi-model assessment of the contribution of SSW events to surface predictability on sub-seasonal timescales. By adopting the nudging approach, this experimental protocol aims to reveal the influence from the "perfect" stratosphere explicitly. This experimental plan is coordinated by the SNAP working group of WCRP SPARC, and is a plan that the SNAP should have submitted and undertaken earlier. After the Phase-I multi-model experiment of SNAP (Tripathi et al. 2016), this community spared time for the "coordinated" (or dull self-nominated) analyses of S2S prediction data. However, as explained in sections 1 and 2, these Phase-II data analyses were almost impossible to disentangle the stratospheric influence on the tropospheric forecast skill with confidence in the causal relationship (I knew that before they did). Therefore, this kind of experiment is necessary to advance our understanding of the stratospheric influence on the tropospheric circulation and to build a common view of expectable skill contribution from the stratosphere in current prediction systems. I support the importance of this proposal.

However, as a reviewer, I feel a little concerned about achieving the purpose of multi-model inter-comparisons in the current proposed settings. In particular, the author's preference of the nudging only zonally symmetric component and the inclusion of the fourth purpose (about wave evolving process in the stratosphere) may prevent a sound comparison of tropospheric response to the prescribed stratospheric state among prediction systems. I could not convince the propriety of the settings, at least from this manuscript. Moreover, the treatment of tropical coupling seems to be inappropriate. Therefore, I recommend the authors reorganize the priority of scientific purposes and show the validity of experimental settings.

We thank the reviewer for their insightful comments and suggestions, in particular with regards to the scientific priorities and the design of the nudging. We have responded to these concerns in more detail below. Our comments are in italics; text in red indicates a change to the manuscript. In brief, we have added some preliminary analysis of model output using the given nudging settings to demonstrate the appropriateness of the nudging settings, and have provided further justification for the science priorities we have chosen.

Major Comments

(1) Is it possible to present some evidence for the validity of experimental settings?

I believe that the experimental protocol's main purpose is to share the fixed details of the setting after enough validations (which prevents others from laborious processes checking dependency on settings). In addition, the presentation of a typical (prototype) result would facilitate further participation by others (e.g., Held and Suarez (1994) presented results of two dynamical cores, and it helps the readers and following investigators to deduce the robustness of results). Since

this series of experiments depends largely on the nudging parameters, how the authors have fixed the parameters should be explained with enough reasoning. For example, p_b and p_t (and function) are different from those of Hitchcock and Simpson (2014). How did you tune these settings? I guess that the authors have conducted test experiments by using some operational system (the IFS?). How significantly affected the choice of a lower limit of the nudging on the tropospheric ensemble spread and mean difference? Proactive presentations of such information would prevent unnecessary future discussions in the step of inter-comparisons.

We have added two new figures showing preliminary output from one participating model (CESM2), along with further discussion of the choice of nudging parameters. These figures show the effects of the zonal symmetric nudging on the ensemble spread of zonal mean zonal winds, meridional heat flux (as a proxy for vertical wave propagation), and tropical temperatures.

(2) Isn't it too greedy to include the fourth purpose?

The zonally-symmetric nudging allows planetary waves to evolve freely (to some extent) even in the stratosphere. This enables the current protocol to address the fourth purpose. However, at the same time, it allows uncertainty of stratospheric state despite that the most important purpose of this experiment is to assess the contributions from the imposed "perfect" stratosphere and compare them among multi-model results. I feel that the well-tailored nudging of full stratospheric state (e.g., middle-to-upper stratospheric full-nudging with a wider buffer zone below) may be more appropriate to pursue the multi-model inter-comparisons. It would be better if the fourth purpose is placed as an additional scientific goal.

A similar concern was raised in RC1, and we refer readers to our response to that comment for a more complete discussion of the relative merits of zonally symmetric versus full-field nudging. In particular, there are good reasons to be concerned about introducing unintended artifacts within the troposphere when nudging to the full field within the stratosphere; these effects have not been quantified by previous work. In contrast the effects of zonally symmetric nudging are better understood, and previous studies have demonstrated that this approach produces much of the expected surface response. Moreover, the fourth science goal will be valuable for understanding limits to the predictability of the SSWs in the first place, which is fundamental for capturing in advance any associated downward impact of the stratosphere. We have added some further discussion of these points in the introduction and in the section on the nudging setup.

Nonetheless, we expect that we will have both zonally symmetric and full-field nudged experiments available to compare the results. The SNAPSI dataset may help to shed light on the tradeoff between the two nudging approaches.

(3) It may be better to change the nudging setting to discuss the tropical coupling.

Although this manuscript roughly touches the coupling between the tropical stratosphere and the troposphere as the secondary science questions (the fifth purpose), this topic has the potential to be a more important target than extratropical coupling. One of the ultimate purposes

of the multi-model inter-comparison is to attribute the model's performance to some particular model settings. As many modelers would agree, one of the most uncertain parts of atmospheric models is the representation of clouds. Therefore, it is natural that stratospheric influence on tropical convections should be placed at the highest priority of multi-model inter-comparisons. In such an investigation, the lower limit of nudging in the tropics should be set higher than that in the extratropics (not to interfere in the high tropical cloud directly). However, I am unsure whether the current setting (p_b = 90 hPa) is high enough to avoid the direct influence. It may be better to introduce latitudinal dependence in the nudging coefficient if the tropical ensemble spread shows an undesirable distribution. Otherwise, it would be better to plan the nudging experiment focusing on the tropical coupling separately. Sloppy spotlighting may ruin chances of further development.

The state of knowledge about extratropical coupling between the stratosphere and troposphere is much more mature than that of stratosphere-troposphere coupling in the tropics. The impacts of stratospheric sudden warmings on the surface are substantial and well-documented by many studies; moreover this methodology is well-established and understood theoretically, as is described in the methodology section. We are in the right position now to carry out this intercomparison exercise for extratropical coupling and our priorities reflect this.

In contrast, efforts to study the impacts of the QBO on MJO with this methodology have been mixed to date. Martin et al. 2021 carried out a similar experiment in a single model study with a range of nudging parameters, including one experiment with the nudging transition set from 100 hPa to 50 hPa; they did not find an MJO connection. In contrast, Noguchi et al. (2020), using full-field nudging but with a considerably higher transition region from 40 hPa to 1 hPa, found substantial impacts from SSWs on tropical convection more broadly. While there is no doubt that understanding tropical coupling between the stratosphere and troposphere is a key research topic, it is not at all clear what the optimal nudging strategy is to capture the details of the tropical coupling. Indeed the question of whether deep convection in the tropics might be somehow affected by the nudging is one aspect to consider. More single model studies are required to understand these kinds of concerns. Moreover, it is likely that this would differ depending on which aspect of coupling in the tropics one is interested in, as is suggested by the contrasting results from the Noguchi et al. (2020) and Martin et al. (2021) studies.

In the absence of clear guidance from previous studies, our approach was to set the lower boundary of the nudging sufficiently low to constrain the state of the QBO in the lower tropical stratosphere without introducing artificial constraints in the extratropical upper troposphere.

Nonetheless, the present experimental design has the potential to reveal aspects of two-way stratosphere-troposphere coupling in the tropics, and we feel it is appropriate to highlight this potential as a set of secondary science questions.

(4) Changing the priorities of the experiment will allow more models to participate.

Among the experiments listed in l.75-89, the "free" and "nudged-full" are free from the artificial relaxation procedure with shocks and relatively easy to conduct even by models with grids that are not necessarily harmonic with the zonally-symmetric nudging. I think it is better to set these

two experiments as the first step request. Then, other experiments ("nudge," "control," and "control-full") should be requested as the second step. Such a division would increase possible participants, at least for the first step. Since the "free" experiments would approach the model's climatological state if the initialization date is set far enough from the SSW onset date (although there are exceptions, of course), the purpose of deducing stratospheric contribution to the troposphere can be roughly achieved by just comparing the "nudged-full" and "free" experiments. I agree that there are large merits of conducting the zonally-asymmetric nudging and comparing it with the "control" experiment. However, I wonder which should we place the priority in the multi-model inter-comparison.

It is true that the nudged-full forecast is easier for models with grids that are not aligned with the parallels. There is one participating model that will carry out the full-field nudging and not the zonally symmetric case. However, it is not at all true that they are free from artifacts associated with the nudging. The boundary between the free troposphere and the nudged stratosphere will act as a strong, unphysical reflecting layer for any large-scale Rossby waves that are inconsistent between the model forecasts and the nudged stratosphere.

Minor Comments

Title: the Stratospheric Polar Vortex —> e.g., "Recent Weakening Events" of the Stratospheric Polar Vortex

Since this protocol covers just only 3 SSWs which are mainly touched by recent publications of quick S2S data analyses, it is inappropriate to use the term representative of various behavior of stratospheric polar vortices.

We have changed the title to "Stratospheric Nudging And Predictable Surface Impacts (SNAPSI): A Protocol for Investigating the Role of Stratospheric Polar Vortex Disturbances in Subseasonal to Seasonal Forecasts"

Is the "control-full" setting appropriate?

Unlike the "control" experiment, the "control-full" experiment would strongly damp the stratospheric wave components due to the sample-averaged smooth structure of the climatological state. Is this as you intended? I think the true "control-full" experiment should construct its ensemble by changing $T_C(t)$ to $T_{year}(t)$ (each year's state of ERA5). In this case, at least a 40-member ensemble can be obtained using the reanalysis data from 1979 to 2018.

Yes, the control-full specification will strongly constrain the stratospheric state. Like the control ensemble, it will provide an assessment of the effects of the zonally asymmetric stratospheric anomalies on the surface relative to a climatological state. It is also true that there will be artifacts associated with constraining the wave field in the stratosphere to something inconsistent with the tropospheric wave field; this is just as true of nudging to the observed stratospheric flow in other years, or of nudging to an observed state that is inconsistent with model dynamics. Using a different year for each ensemble member would also vastly increase

the size of the reference dataset. This is certainly an interesting idea, but it is not clear that it offers clear benefits compared to the technical challenges it introduces.

Figure 1:

Is it possible to arrange this figure as a more straightforward form for this manuscript? I think the histogram of the split SSW is unnecessary (e.g., Figure 11 of Maycock et al. 2020: Removing CTL_ADJ is more desirable...).

Figure 1 shows the effects of nudging the zonally symmetric component of the stratosphere on the ensemble distribution of NAO. The two nudged ensembles (SSWs and SSWs) are nudged to different events taken from a free running version of the model, much like the present protocol focuses on several specific case studies. Including the results from both nudged ensembles demonstrates that the result is robust across multiple reference cases, which is quite relevant to the present protocol. We added a brief comment to clarify this in the text, though a full discussion seems out of place in the discussion.

Figure 2:

This figure needs to be brushed up. The observation should be changed to the same format as the forecasts. It seems that the temperature anomalies of the forecasts are limited over the land. How many ensemble members are used to plot in each panel?

We have updated the observations panel to have the same color scale as is used in the forecast plots. We have also added text to the caption providing further information about the forecasts; specifically they include 40 ensembles initialized over a span of 10 days (four runs initialized per day). All panels are now masked to emphasize forecast temperatures over the land.

The caption of Figure 3:

In my understanding, Butler et al. (2020) does not describe the calculation method of NAM indices in detail. They have just cited Gerber and Martineau (2018). I do not really like such an inappropriate citation. It is better to write such as "ERA5 version of Figure 5(a) in Butler et al. (2020)."

We have changed the citation to Gerber and Martineau (2018).

Table 6 and Authorship:

I doubt the necessity of Table 6 and authors from operational centers since the numerical integrations are not performed, and any early results are not provided in this manuscript. They have just only expressed the intention to participate. The authorship of these people should generate when the data are submitted and the model settings are described in some data journals (e.g., ESSD?). Therefore, the contribution of these types should be noted in the acknowledgement.

The protocol was developed with significant input and feedback from the modeling center contacts. The choice of events and initial conditions, design of the nudging, data request and scientific priorities were all determined in consultation with these contacts to ensure they were reasonable from a technical point of view and that they were fit for the scientific purpose. Their authorship is well-justified in this protocol description paper.

Typos, etc.

l.320: 60 N --> 60^{o} N

Done.

Make consistency in the use of abbreviation terms (NAO, SAM, MJO, QBO). For example, l.396 and l. 403 uses “Southern Annular Mode” although the SAM is already defined in l.307. Also, “NAO” is used in l.129- before the “North Atlantic Oscillation” in l.308.

We have worked to improve our use of acronyms.

References

Gerber, E. P. and Martineau, P. (2018) Quantifying the variability of the annular modes: reanalysis uncertainty vs. sampling uncertainty, *Atmos. Chem. Phys.*, **18**, 17099–17117, https://doi.org/10.5194/acp-18-17099-2018

Held, I. M., and Suarez, M. J. (1994) A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models, *Bull. Amer. Meteorol. Soc.*, **75(10)**, 1825-1830. https://doi.org/10.1175/1520-0477(1994)075<1825:APFTIO>2.0.CO;2

Maycock, A. C., Masukwedza, G. I., Hitchcock, P., and Simpson, I. R. (2020) A regime perspective on the North Atlantic eddy-driven jet response to sudden stratospheric warmings, *J. Climate*, **33(9)**, 3901-3917. https://doi.org/10.1175/JCLI-D-19-0702.1

Tripathi, O. P., Baldwin, M., Charlton-Perez, and co-authors (2016) Examining the predictability of the stratospheric sudden warming of January 2013 using multiple NWP systems, *Mon. Wea. Rev.*, **144(5)**, 1935-1960. https://doi.org/10.1175/MWR-D-15-0010.1