We thank the reviewer for their thoughtful comments. We have replied to each point in italics below. Text in red indicates a change to the manuscript.

The manuscript outlines a set of protocols for multiple but standardized global climate modeling experiments to study the stratosphere-troposphere coupling under the umbrella called the Stratospheric Nudging And Predictable Surface Impacts (SNAPSI). The authors describe and outline an intercomparison modeling experiment to study the role of the Arctic and Antarctic stratospheric polar vortices in sub-seasonal to seasonal forecast models.

I appreciate that the authors have concerns for a nudging the stratosphere to the full observed state including eddies rather than the proposed zonally symmetric state. Still, I am concerned that only nudging to the zonally symmetric observed state may omit important stratospheric information or forcing on or coupling with the troposphere. I don't have a suggested solution but do want to raise the concern.

Similar concerns were also raised in RC2, and this was a topic of discussion amongst all of the authors when designing the protocol. There are merits and drawbacks to each approach, and so we discuss here some of the pros and cons to adopting a full-nudging versus a zonal-mean only nudging strategy.

There are two primary arguments for full-field nudging: first, that there may be some important role for stratospheric asymmetries in determining the tropospheric response to SSWs, and second, that models operating on a grid that is not aligned with the parallels may be unable to participate in the overall protocol.

There are two primary arguments for zonally-symmetric nudging: first, by leaving the wave field to evolve freely, the experiments will allow us to investigate the impact of stratospheric mean-state biases on the forecast of the planetary wave field, and second, that we have a deeper theoretical understanding of the consequences of zonal mean nudging.

Past work has demonstrated that much of the surface response is in fact captured by the zonal mean nudging approach alone. Planetary waves in the extratropical stratosphere are suppressed following SSWs, which means zonal asymmetries in the stratosphere are weak. This work has emphasized the time-mean jet shift component of the response rather than the shift in probability of extremes such as cold air outbreaks, so it is possible that the zonal mean nudging will miss some possible impacts arising from the asymmetric component of the stratosphere.

It is not clear, however, that full-field nudging will really provide an 'upper bound' on the downward impacts of the stratosphere. Nudging of any kind in the presence of strong balance constraints implies that there will be unintended, remote consequences of including an artificial forcing. Several model studies (Orbe et al. 2017, Chrysanthou et al. 2019) have pointed out poorly-understood dynamical and transport inconsistencies in specified dynamics integrations. Nudging the asymmetric component of the stratosphere will also introduce an effective and highly artificial reflecting layer for any large-scale Rossby waves that are not consistent between

the model forecast and the nudged reference state. The induced zonal asymmetries in the stratosphere may also act as an effective stratospheric source of waves. All of which may produce unintended biases in the surface flow; these effects have not been quantified. In contrast, the dynamical artifacts associated with zonally symmetric nudging are better understood (see Hitchcock and Haynes 2014).

Arguably, the zonally symmetric evolution of the stratosphere is an easier target for improved forecasts. Particularly following stratospheric sudden warmings, radiative processes dominate the evolution of the polar vortex. The zonally symmetric nudging ensembles might thus be better regarded as a plausible target for forecast models.

Finally, it is again arguable that relaxing the stratospheric zonal mean state to climatology provides a more natural control than relaxing the full field in the stratosphere. The zonal mean climatology is likely to be more dynamically consistent with the tropospheric state than the full field climatology, in which the planetary waves (which have long vertical wave lengths) will be constrained to their quasi-stationary climatology.

There is indeed one participating model (the GFDL SPEAR model) which cannot easily carry out zonal mean nudging and will only contribute the full nudging runs. However, the remainder of the models can and will carry out the zonally symmetric nudging. Given that several models may turn to more complex grids, it could be difficult to carry out model intercomparisons using zonally symmetric nudging in the future.

In summary, there are strong arguments for carrying out both symmetric and full-field nudging forecasts. Both are included in the protocol, and we expect some modeling centers carry out both, enough that we will be able to carry out some detailed comparisons between the two approaches. On balance, the additional science questions that can be addressed with the symmetric nudging approach was felt to be worth prioritizing.

Some of these arguments have now been added to the text, both in the introduction and in a new subsection at the end of section 3.

In Figure 2, I suggest that the plot of the observations be made consistent with the forecast pots? I found it hard to compare between observations and the forecasts.

We have modified the final panel of this figure so that the observations have the same color scale as the forecasts. We have also added some additional information in the figure caption.

Line 227 Not sure that I agree with the statement: “Comparisons between the nudged and control ensembles will provide a clear means of assessing the stratospheric pathway at play for those teleconnections that are active during the selected case studies.” There could be multiple forcings in play and nonlinear interactions that would make attribution complicated.

Yes, there are likely to be both multiple forcings and nonlinear interactions, and these experiments will not allow us to disentangle every such interaction. However, the experimental protocol provides a simple and causal way to assess the role of the stratospheric mean state in

any pathway. For instance, if teleconnections from the tropical Pacific play a role in the extratropical response, and if these depend on the state of the stratosphere (e.g. Domeisen et al. (2015)), they should be active in the nudged ensembles but not in the control ensembles.

We are now more explicit about how these comparisons will shed light on this scientific question.

3.1 Why are only temperature and zonal winds provided from ERA5. I would have thought to include geopotential height and meridional winds as well?

Only T and U are required to be nudged in the zonal mean; geopotential height is determined by the temperature field and V in the zonal mean is strongly constrained by hydrostatic balance and continuity (see Hitchcock and Haynes 2014).

Lines 256-258 This is a difficult balance to strike nudging the stratosphere towards observations without throwing the whole model simulation out of whack. I can understand imposing no nudging below 90hPa that accelerates to full nudging at 100hPa but I don't believe that we fully appreciate the importance and role of the lower-stratosphere separate from the mid-stratosphere in stratosphere-troposphere coupling. In fact, I believe that the lower- and mid-stratosphere could influence the troposphere somewhat independently. I am concerned that by imposing now nudging in the lower stratosphere will dampen the full influence of the stratosphere on the troposphere. One idea that I would suggest considering is applying the limit of the nudging to different levels.

We agree that the question of which levels within the stratosphere are most relevant for stratosphere-troposphere coupling is an interesting and important question, and one that is not fully understood. The lower stratosphere has been shown to be particularly relevant for understanding the impacts of stratospheric sudden warmings (e.g., Karpechko et al. (2018). On the other hand, the mid-stratosphere is thought to be more relevant for planetary wave reflection (e.g., Perlwitz and Harnik 2004). The impacts of imposing nudging at different levels has been considered in detail in a simpler model context by Hitchcock and Haynes (2016), who found that the surface impacts were stronger when the lower stratosphere was better constrained.

While this issue certainly warrants further research, our goal here was to constrain as much of the stratosphere as was feasible without directly impacting the troposphere. The choice to ramp up the nudging from 90 hPa to full strength at 50 hPa is similar to the lowest level of nudging considered by Hitchcock and Haynes (2016), remains well above the level of the extratropical tropopause (which is more than a scale height below), and is low enough to constrain the lower stratospheric QBO winds which are thought to be important for their tropical impacts.

Lines 340-343 – I felt that the discussion about the MJO and its possible influence on the NAM and Northern Hemisphere weather is an unnecessary distraction almost like "having your cake and eating it." The paper is about stratospheric influence and stratospheric nudging so why introduce that the MJO is needed to simulate the correct weather? I think better to leave tropical forcing and guidelines for modeling experiments to study tropical forcing for anther paper.

There is significant diversity in the surface response to stratospheric sudden warmings. However, a key question in the context of S2S predictability is the origin of this diversity. Does this diversity arise exclusively from synoptic-scale tropospheric processes that may only be predictable for a week or so in advance? In this case this diversity would essentially be `irreducible' on subseasonal timescales for any given event. However, if some of this diversity occurs because of other subseasonal drivers, we may be able to say in advance whether or not a given event will lead to a significant surface response. The results of Knight et al. suggest that the state of the MJO may impact the NAO in late February, which suggests a possible predictable control on the surface impacts.

If the diversity is due to unpredictable components, these should differ from ensemble member to ensemble member, and should be independent of nudging imposed in the stratosphere or of initial conditions. If, on the other hand, they are related to other forcings that can be predicted on sub-seasonal timescales, they should be present in ensemble means, but should differ between, e.g. the 2018 and 2019 events, or possibly between different initialization dates for the same events.

We do not know a priori which modes are most relevant to this diversity; we quote here the state of various potentially important modes of variability in part for the reference of future studies analyzing output from these runs.

We have added text to further clarify why we are quoting the state of these remote climate drivers.

Lines 361-365 I don't disagree that the tropospheric NAM response in 2019 was quite different than the tropospheric NAM response in 2018 to the stratospheric polar vortex split. However just by looking at Figures 3 and 5 it is not that obvious to me. I wonder if a different comparison might better highlight the difference.

The time-averaged NAM anomaly at 500 hPa for the one-month period following the central date after the 2018 SSW case (12 Feb 2018 through 12 Mar 2018) was -1.17σ. After the 2019 SSW (2 Jan 2019 through 2 Feb 2019) it was 0.02σ. We now quote these values in the text.

The composite average response following SSWs is negative, more consistent with the 2018 case. As discussed above, a central question to be investigated is whether this difference is predictable, and whether it can be attributed to any specific remote climate driver.

Lines 376-379 I agree that the tropospheric response differences to the SSWs in 2018 and 2019 are interesting and is worthy of model experiments. But again I do question introducing into the discussion the MJO and tropical forcing. Almost makes the role of the stratosphere seem like noise rather than a signal and therefore could be ignored. My opinion is to take out this mostly hand wavy discussion of the MJO and tropical forcing, which seems self-defeating in trying to motivate stratosphere-only sensitivity experiments.

As discussed above, these are potentially relevant climate drivers that may be shaping the details of the response to these specific events. We have left this discussion in, as justified above.

Line 400 I do winder why the September 2019 Austral minor warming was chosen over the September 2002 major warming? In fact the SAM was much more negative in October 2002 (I believe a record in fact) than 2019. If we assume that a reversal of the winds at 10hPa is necessary for the tropospheric response, how do you justify including a case where the winds never reverse? I do believe that the September 2019 austral polar vortex disruption is interesting but seems like not a good fit for the framework of this study. The first two words in the Abstract are "major disruptions." At a minimum justification of the choice is needed.

We disagree with the premise that the 10 hPa winds must reverse in order for there to be surface impacts. Minor warming events (in which these winds do not reverse) have been shown to impact the surface (Thompson et al. 2005; Lim et al. 2019, see their Fig. 4). The austral polar vortex was significantly and substantially disturbed throughout the stratosphere in the 2019 case (see Fig. 7) despite the fact that the winds at 10 hPa did not reverse. The anomalies in the 2019 case were in fact of very comparable amplitude to the 2002 case but occurred somewhat earlier in austral spring.

The 2002 event would also have been a valuable case study to consider. We chose to focus on the 2019 case for several reasons. Firstly, recent work has focused on this event from an S2S context. The 2002 case has also been highly studied, but not necessarily within the S2S context; moreover there are many open questions about the dynamical mechanisms that triggered the 2019 event and its surface impacts. Secondly, previous work has attributed the hot and dry extremes over Australia to these stratospheric anomalies, making this an interesting case to consider from the point of view of the dynamical attribution of extreme events. This has now been made more explicit in the text.

Thompson, D. W. J., Baldwin, M. P. & Solomon, S. Stratosphere–troposphere coupling in the Southern Hemisphere. J. Atmos. Sci. 62, 708–715 (2005).

Line 471 – I am surprised by the data being embargoed initially. Seems counterproductive to me.

The purpose of the embargo is simply to ensure that the modeling center participants receive credit and recognition for the resources and efforts that they put into the design, execution and post-processing of the experiments. Anyone interested in analyzing the output will have access to the data from the archive, but will be required to offer co-authorship to the modeling center participants and SNAPSI leads for any paper published within the embargo period. We feel this is a fair request that will not hinder community access to the dataset. We realize this was not made clear in the submitted draft; this has now been clarified.