Response to Reviewer #1:

Line 96-98: Would you please be more specific about how the stationary waves are calculated?

Is the time mean geopotential height that is removed while calculating the stationary wave only a one-week average? Thank you. We have added to the text 'weekly mean'.

Is the climatological (multi-year) zonal mean geopotential height removed to calculate the stationary wave? Yes, this is mentioned now: "We define the stationary waves by first computing the weekly mean geopotential height over intializations during November-December-January-February (NDJF) for each model, then compute the climatology for each week, and finally subtract off the zonal mean height at each latitude."

Is a November to February temporal mean removed to calculate the stationary wave? If so, this definition of the stationary wave may not control for annual cycle variability. The stationary wave structure does evolve throughout the winter.

We use the mean for each week, and so we take into account the annual cycle in stationary waves.

Line 110: ERA-Interim is introduced to the reader on this line. I think a line should be added to the Methods section stating that ERA-Interim will be used as the "truth" that the hindcasts will be compared to. Please add a citation for the reanalysis as well. This has been added to the text.

Figure 1: This figure is nice. In its caption, please state what the contour intervals are for the wave-1 and wave-2 contours. Contour intervals have been added to the caption

Figure 1: On panels (i) and (l), the filled contours are not filled where the anomalies are lower than -60 meters. I have come across this as well while plotting. If you are using python and matplotlib, the …extend = 'both'… part of the code below will fill these contours:

contour = m.contourf(x, y, vort, latlon=True,cmap=cmap1,extend='both',levels=levs1,vmin=vmin,vmax=vmax)

Thank you, this is fixed now.

Figure 2: Please consider changing your contour colors to something that is more inclusive to people who are colorblind. This article provides guidance: https://www.nature.com/articles/d41586-021-02696-z Thank you, this has been done.

Figure 2a: ERA-I is shown with dashes. Why are there multiple dash contours? Is this ERA-I over different time periods? If yes, please state this in the Figure caption.

The dashed lines were for older versions of the ECMWF, UKMO and Meteo France models, while dots were reanalysis to match the dates for each model.

For simplicity, we decided to remove the old model versions in the revised version of this figure, and similar spaghetti plots. The caption has been revised accordingly.

Line 129: Please replace "observations" with "reanalysis." Replaced

Line 125 – 126: I like that the focus is on these three key regions. Figure 2b gives the impression that there is considerable variability amongst the modeled stationary waves at 200E also. Do you have any hypotheses on why this would happen? Yes, this is due to a more zonally confined NP trough in some models vs others. However this region is not near the maxima/minima of the pattern, and for 2b we prefer the minima.

When we consider the link between convection and North Pacific biases in Figure 9, we now focus on this region (195°E-215°E) instead of 160°E-170°E.

Figure 4: Typo in second line of Figure 4 caption. Thank you, fixed

Figure 4: Please list the contour intervals for the stationary wave and for the zonal wind.

 The contour intervals have been added to the caption.

Line 165 – 166: Is Figure 5c being referenced here? I cannot make out a PNA signal in panel 5c. This is indeed not a classic PNA pattern as it extends into the subtropics. This sentence has been removed from the text.

Line 180: There is a missing word in this sentence. Thank you, it is fixed now.

Line 191: Does either of these studies provide a physical explanation for what causes the biased ridge? If so, please add one line on this. It is difficult to infer a physical mechanism from analyzing the models, as we cannot control the model settings and due to limited data

availability at different pressure levels. For that, an idealized modeling work has to be performed, and this is a work in progress.

Line 284 – 285: Should we expect that the convection over the eastern Pacific is associated with the western North America ridge? The impression I have from Garfinkel et al. (2020) is that the ridge forms due to the nonlinear interactions amongst the "building blocks" that their study focuses on, not tropical convection. To the extent that models represent large-scale topography and land-sea contrast, they already have these two building blocks. Specifically, biases in land-sea contrast would have a rather obvious and immediate impact on surface temperature, winds, and moisture availability (subsequently precip), so we assume models are handling it as well as they can for reasons unrelated to stationary waves. Higher resolution helps resolve topography, but from contrasting the T42 vs T85 experiments in Garfinkel et al 2020 the added value in increasing resolution is not large. Hence our focus in this paper is on the role of tropical convection which varies qualitatively among the models, though we can't rule out other sources of bias.

More generally, we have lowered the degree of confidence implied when we discuss the role of convection for stationary wave biases.

Figure 9a: The connection between tropical convection, subtropical descent, and the East Asia trough is plausible. I feel like the investigation of if tropical convection/subtropical downwelling impacts the stationary wave pattern could be a little more thorough. I think this study could be improved by further investigating the sources of the stationary wave biases.

We have added a figure to the supplemental material showing correlations of these SW features with omega globally. Overall, we agree that demonstrating convincingly **the source** of a stationary bias from model output is a difficult task even if we had full access to model output rather than what is archived by the S2S project.

Thus, we have lowered the degree of confidence implied when we discuss the role of convection. Further, we have work in progress currently using idealized modeling trying to more closely pinpoint how convection and zonal wind biases lead to SW biases. For now, this study aims to identify the stationary wave biases in the models and suggest possible sources that are now further investigated using modeling work.

Have you considered the Rossby wave source? Scaife et al. (2017) did a similar analysis – analyzing the relationship between tropical precipitation and tropics-extratropics Rossby wavetrains. See their Figure 6. Here are some suggested plots: (1) subtropical Rossby

wave source as a function of tropical omega; (2) North Pacific trough bias as a function of subtropical Rossby wave source bias; (3) Rossby wave source maps; (4) North Pacific trough bias as a function of 200 hPa subtropical velocity potential bias.

Scaife, A. A., Comer, R. E., Dunstone, N. J., Knight, J. R., Smith, D. M., MacLachlan, C., ... & Slingo, J. (2017). Tropical rainfall, Rossby waves and regional winter climate predictions. Quarterly Journal of the Royal Meteorological Society, 143(702), 1-11.

Thank you for this suggestion. We had indeed already computed the RWS for three of the considered models (CMA, NCEP, UKMO), but we found omega at 500hPa more conclusive. Specifically, there exist significant differences in RWS between models, likely related to model biases, and hence, 500hPa omega provides a better illustration that is easier to compare between models. Hence, we have decided not to include the RWS results into the manuscript.

For the reviewer's interest, we have included figures of the RWS for these three models into the reviewer response, see Figure R1.

As an example, the Figure below shows the climatology of RWS at 200hPa, divergence at 200hPa, and omega at 500hPa in week 1 of NCEP as compared to ERA-I, for hindcasts initialized in NDJF. Note that the S2S database only includes the pressure levels 300hPa, 200hPa, and 100hPa, which limits the resolution with which we can compute RWS. The RWS and divergence at 200hPa are noticeably too weak in NCEP as compared to ERAI, even as omega at 500hPa is largely reasonable. Note that this is in week 1, when the initial conditions should still be playing a large role and one would hope that the models are doing a reasonable job.
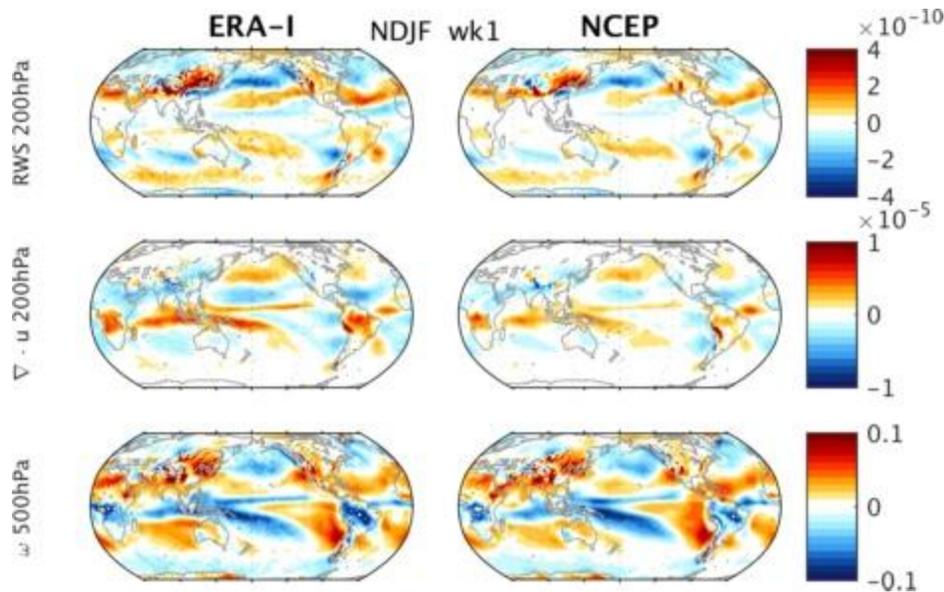
Figure R1. Left column, from top to bottom: week 1 200 hPa Rossby wave source, 200 hPa wind divergence, 500 hPa ω during NDJF; right column: same as left column but for the NCEP model

Our interpretation of this result is that too little of the divergent outflow in NCEP occurs at 200hPa, even though convection is occurring in the correct location with a reasonable mass transport, as the omega 500hPa climatology is reasonable. More generally, the amplitude of tropical and subtropical omega 500hPa is reasonable in most models, and hence we elect to focus on omega in the paper. Given the low vertical resolution available (and also the lack of diabatic heating) in the S2S archive, it is not possible to cleanly identify biases in the convective profile of each model, though we suspect that such biases exist. Future work should consider this issue in more detail assuming more detailed output data becomes available.

Figure 9a: Schwartz and Garfinkel (2020, JGR, MJO study) showed that there is more eddy heat flux entering the mid-latitude stratosphere 1-3 weeks after MJO phase 6/7, suggesting that there is an anomalous tropics-extratropics wavetrain producing the transient eddy heat flux. The convection center during phase 6/7 is between 140E and 180E. Figure 9a looks at subtropical omega between these same longitudes. Assuming that the subtropical descending branch of the meridional circulation between these longitudes is "feeling" what is taking place in the tropics, to what extent is MJO variability present in Figure 9a? Does the Figure 9a correlation improve by compositing by MJO phase?

The MJO is a transient variability on subseasonal timescales. By averaging over many years of data and many initializations within the NDJF season, variability associated with the MJO is filtered out.

Response to Reviewer #2:

1) Use and choice of small regions for bias characterization

The discussion around l125 suggests that the largest biases arise near the peaks and troughs of the observed stationary wave. In comparing Fig. 2a and b don't see this at all. In particular, I worry that focusing the discussion on these quite narrow (10 degree by 10 degree) regions can give quite an incomplete view of the nature of the biases across the S2S models. I worry that Figs. 3, 6, and 9 may be quite sensitive to these choices. At a minimum there should be some demonstration that the inferred connections between biases are not sensitive to these choices, and this should be in the manuscript, not just in the response to reviewers. It would also be very helpful to see maps of intermodel correlations in some cases (more on this below). I also wondered if the analysis might be more powerful if the focus was on amplitude and phase of the leading wavenumber components of the anomalies.

We now consider wider areas of 20 degrees by 20 degrees, and the results remained unchanged. If anything, the correlations are even stronger.

In general the phases of wv1 and wv2 are well captured by essentially all models, however the amplitude is more of a mixed bag. The amplitudes of wv1 and wv2 are already shown on figure 6, which connects these amplitudes to regional biases in Z*.

2) Connection to tropical convection

It is certainly very reasonable to hypothesize that these biases could be related to biases in tropical convection. But I again find the evidence presented to be pretty weak: I am not at all convinced that the first place a modeling group should turn to to correct these errors is the tropical mean convection. In part the correlations are relatively weak. Moreover, this is again based on correlations of very small regions. One way to make this connection more convincing may be to show inter-model correlation maps of omega versus geopotential height biases. This would indicate whether the biases have a teleconnection pattern.

We have made figures of the correlation between omega biases across models and Z biases across models as requested. See supplemental Figure S13 in the revised paper. The results from this figure support the paper. More generally, we have lowered the degree of

confidence implied when we discuss the role of convection for stationary wave biases, as we cannot demonstrate causality.

3) Connection between stratospheric bias and stratospheric resolution

This is a simple request (hopefully), but it takes a lot of effort to determine which symbol in a given plot corresponds to which model, and in particular, which symbol corresponds to a high resolution vs low resolution model. It would help to have a different kind of symbol for models in these categories; in particular this seems more useful than distinguishing model versions from individual models.

 Old model versions have been removed from figure 2 and similar spaghetti plots. We also added diamonds to low-top models.

A closely related question: Is the wave two component of low-resolution models in better agreement with observations than those of high-resolution models, or is it just that the wave one biases dominate in these cases?

For wave-2, biases in low-top models are comparable in magnitude to those of high-top models, especially in the troposphere (ISAC is an exception). For wave-1, biases in low-top models are more pronounced, therefore it is indeed wave-1 that dominates the biased mean state in the stratosphere.

4) Importance of outliers

In many of the inter-model correlation plots there are one or two models that are to some extent outliers and in some cases seem to be determining the overall correlation (at a quick glance: Fig 6d,e,g; 9b). Some discussion should be included about the sensitivity of these correlations to such outliers.

In figure 6, the outliers have been removed and the correlation coefficient has increased in most panels. Attached is figure 6 without the outliers:
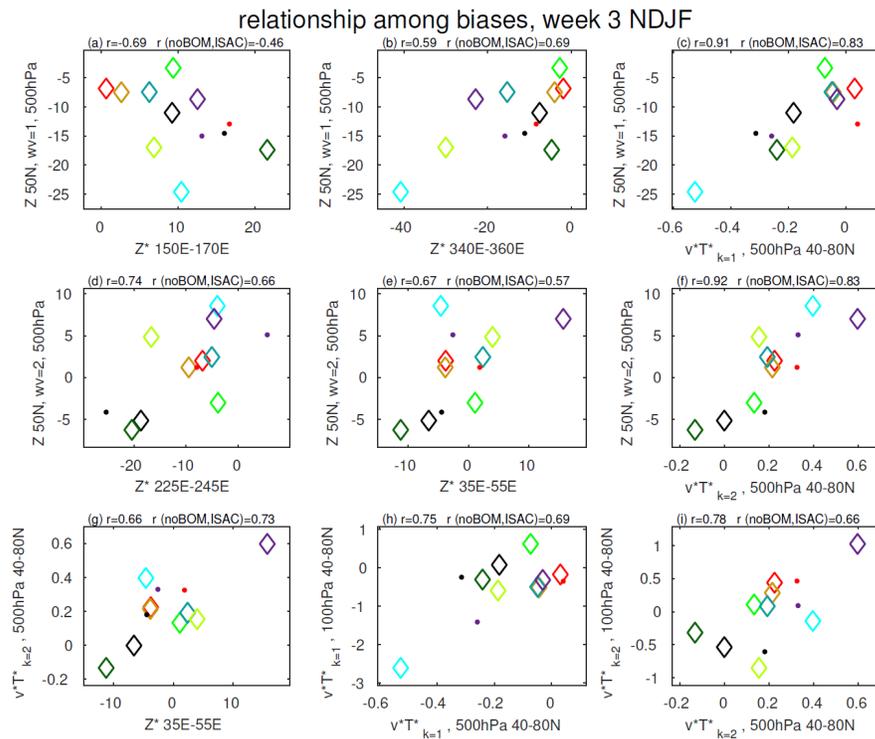
relationship among biases, week 3 NDJF

Further questions/comments

1) The authors choose to stratify forecasts by model version in some cases as a result of updates the forecast model over the course of the S2S project. Is there any evidence that these biases depend on model version and not just on sampling errors due to the different time periods? My impression from some single model studies was that the difference was fairly small (I could not easily find a reference for this). In any case, if this is clear it should be presented to justify the extra stratification; if not I would think it better not to stratify the results in this way (?)

Older model versions have been removed from figure 2 and other spaghetti plots. We choose to keep them for the correlation plots (figures 6 and 9), but the reviewer is indeed correct that biases are not substantially changed across model generations.

2) Figure 7 is quite interesting in that it suggests some connection between the stationary wave biases and the zonal mean state. One point of clarification - are the heat fluxes from the stationary component alone?

The heat fluxes are computed using daily data, and then we average over many initializations to get the time mean heat flux. So this isn't a true stationary wave heat flux (where one would generally take the time mean v and time mean T). However in the Northern Hemisphere the difference between the time mean of the daily heat flux and the heat flux computed using time mean v and time mean T is small. (See e.g. the ERA-40 atlas, though we have reproduced this result using other reanalyses. In the Southern Hemisphere this is not the case.)

This is important in that it provides a connection between these biases and other mean-state biases that could be of strong importance for accurately capturing the impact of the stratosphere on forecast skill, for instance. There are some interesting relationships - for instance, the heat flux forecasts of JMA seem to be about right, whereas the zonal mean wind speeds seem to systematically decay. Also, heat flux biases in the CMA forecasts are larger than those in the ISAC model, but the zonal mean state of the latter seems to diverge more quickly.

The JMA zonal mean winds at 10hPa decay as in reanalysis, so that agrees with the simulated eddy meridional heat flux in the lower stratosphere. As for the CMA and ISAC, for wave-1 the CMA indeed has larger biases, but for wave-2 the biases in ISAC are larger. In fact, ISAC is biased in both wave-1 and wave-2, so it somewhat agrees with its biased 10hPa zonal mean winds. For ECCC on the other hand, there doesn't seem to be a relationship between heat flux biases and U10hPa60N biases

Can the authors comment on the relative role of dynamical and radiative processes in determining the mean bias?

Given the present work, we can only comment on dynamical processes that may contribute to the mean bias. The SNAP subproject of SPARC is currently organizing a comprehensive overview of biases in the stratosphere in the S2S models, and this will include a discussion of the relative role of dynamical vs radiative processes. We have added "Radiative

processes can also contribute to mean-state biases in the stratosphere, and future work should consider the relative role of radiative vs. dynamical processes for mean-state biases."

3) Can the authors comment in the consequences of these biases? Do they correlate with forecast skill in any way? We added 1-2 sentences to the conclusions regarding predictability skill. However, this is not the focus of this work, and is discussed in Domeisen et al. 2020a, and will be further analyzed as part of the SNAP papers on stratospheric biases in S2S models.

Domeisen, D. I. V., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., et al (2020). The role of the stratosphere in subseasonal to seasonal prediction: 1. Predictability of the stratosphere. *Journal of Geophysical Research: Atmospheres*, 125, e2019JD030920. https://doi.org/10.1029/2019JD030920