Weather Clim. Dynam. Discuss., referee comment RC2
https://doi.org/10.5194/wcd-2022-12-RC2, 2022

Comment on wcd-2022-12

Anonymous Referee #2

Referee comment on "Quantifying stratospheric biases and identifying their potential sources in subseasonal forecast systems" by Zachary D. Lawrence et al., Weather Clim. Dynam. Discuss., https://doi.org/10.5194/wcd-2022-12-RC2, 2022

The study by Lawrence et al provides a comprehensive evaluation of the models contributing to the subseasonal to seasonal (S2S) hindcast database with respect to their representation of the stratosphere. This evaluation exercise is relevant because there is increasing focus on analyzing the impact of the stratosphere on (sub-)seasonal predictability; the evaluation of the stratosphere-troposphere coupling, which is in this context even more relevant, is planned for a follow-up paper by the group of authors. A number of common biases is identified here, often consistent with long-standing biases known from climate model evaluations. As such, the study provides a useful compendium of stratospheric diagnostics that both users of the data as well as modeling centers can refer to in future applications and model development. The study mostly falls short to reveal the (possible) reasons for the biases, but this is a task that cannot be expected from such an evaluation effort. Overall, the study is well written and I can recommend publication after some (mostly minor) issues are addressed, as detailed in the specific comments. In particular I have concerns with some of the diagnostics and/or their interpretation, that should be addressed by the authors.

Specific comments:

- page 5, line 128: The comment on excluding the newer cycles of the ECMWF model is confusing - as such the modification of the model would rather be a motivation to include the data in order to assess its impact on the stratospheric representation. I guess this has to do with avoiding to mix different model cycles, and not having the data available for the relevant periods, but this could be explained better ibn the text.

- page 7, line 175: Why "approximately" the 80th percentile, but still this is a very specific value (41.2 m/s) ? I would expect to either use a rounded value (40 m/s), which is "approximately" the 80th percentile, or this value is indeed very close to the 80th

percentile, in which case you can drop the "approximately".

- page 9 / Figure 2: I wounder whether the averaging over the groups of high- versus low-top models makes much sense - E.g. in the group of the high-top models, there are ~4 models with global mean warm biases (dominated by GEFSV12 and NCEP), and ~4 models with (less strong) cold biases at 10 hPa, so in the average there is some compensation going on, and the total bias is likely dominated by the two models with strongest biases (GEFSV12 and NCEP). Also in the group of low-top models, there is some compensation between models with strong warm and cold biases. So excluding a specific model with a strong bias from the composite of models would likely strongly change the picture, thus making the analysis somewhat meaningless. If the aim is to show that the group of high-top models generally has smaller biases (plus reveal the regions of biases), it could be better to calculate a metric like mean of the absolute error, or a root-mean square error across the groups of models.

- page 11, line 265: I expect the GW parameterization will be relevant here, too.

- Fig. 3: Could it make sense to show the difference of a model's drift (panels a,c,e,g,i,k) minus the respective subsampled ERAI value (panels b,d,f,h,j,l). As is, the Figure requires quite a bit of comparison in the head between the panels to interpret the model differences (e.g. BoM seems to be a clear outlier, but at in some cases (e.g. U50 QBO) this seems to be at least partly due to the sampling, as also to ERAI sampled at BoM is offset to the other models).

- Fig. 3 /4: in both instances, the ERAI subsampled data to the BoM model appears to an outlier. I first suspected that this has to do with limited sample size, but according to the numbers in the Figures and from checking the table 1, BoM is actually the system with most members. Can you comment on the reason for the difference, and possibly add a sentence to the manuscript? Does this mean the other models are sub-sampled (but why do they tend to agree, then?)

- page 12, line 286: I expect you mean to say that you need to consider initialization dates before the Holton-Tan effect is established, so it is not prescribed in the initial conditions, but can freely evolve in the models? As is, the sentence sounds a little confusing.

- page 13, line 295: "generally fail": consider changing to "mostly fail" or such, since some systems (ECMWF, NCEP) do show some signature of the effect.

- page 13, line 300: I would change the term "upward wave driving" to either just "wave driving" or "upward wave fluxes that disturb the polar vortex" or "upward wave fluxes that lead to wave driving of the polar vortex". The wave fluxes can be described as

(propagating) upward, but the wave driving is not "upward". Likewise, in like 301, the meridional eddy heat flux is not a metric of the wave driving, but of the upward wave flux (being the main component of the upward Eliassen-Palm flux).

- page 13, line 303: consider re-wording "such wave-driving should be resolved" to be more precise: the planetary waves, which we know are the major contributor to stratospheric wave driving, are definitely resolved in the models, also synoptic wave activity, which plays a role around e.g. the subtropical jet is also resolved. Maybe you rather mean "well represented" rather than "resolved"? Could change to e.g. "the wave driving by planetary and synoptic waves is resolved, and its representation is dependent upon... "

- page 14, line 306: what does "long-term coupling" refer to here? I suspect coupling on the sub-seasonal time-scales?

- page 14, line 309 ff / Fig. 5: The analysis of heat fluxes as a proxy of wave activity in the stratosphere is a good metric and valuable addition to simply looking at the polar vortex in order to get at the process representation. However, I have some concerns with how the diagnostics are used / discussed here. In particular, the "map" of heat fluxes reveals "two centers of action" (line 313). This pattern simply emerges because (as is well known) the wave activity is dominated by planetary waves, in particular it is wavenumber 2 that is showing up here. Therefore, I would argue that the representation of the fluxes as a map, and even more so the averaging over only one segment of the map is rather meaningless. Indeed, heat or momentum fluxes are only meaningful when averaged over the scale of the dynamical feature (the wave) they are connected with (here the planetary wave 1 and 2), i.e. the zonal average should be considered. If the geographic information (latitude of maximum amplitude, or phase of the wave) of the wave is of interest, a better quantify would be geopotential height anomalies. Therefore, I'd strongly recommend to drop the averages over the segments, and focus on the wavenumber 1 and 2 zonal averages, as it is done in the lower panel. In this light the statement in line 327 ("have small biases near 0, indicating that on a global scale, the regional biases tend to cancel") is misleading: yes, the values over the zonal mean are smaller compared to the regional average, but this is because you average over the phases of the wave (as it should be done). Whether the values are "small" is not a question of comparing them to the values from the segments above, but how they compare to the mean value from ERAI.

- page 16, line 351: As stated, in particular for the strong vortex events, much of the deviations might arise from the fixed threshold. I wounder whether this metric then rather measures the mean vortex biases then the ability of a model to produce extremes, in particular for the strong vortex events (for SSW it is different, because a "0 m/s" threshold is a dynamically meaningfull value - it inhibits planetary wave propagation. For a strong vortex, no such fixed dynamical threshold exists). Have you considered using a model-dependent threshold that might be based on a specific percentile?

- page 18, line 383: I think you have to be careful here with the formulation on statements on prediction of vortex events: Yes, "the prediction systems are generally not

forecasting extreme vortex events..." at lead times of 3-4 weeks - and they shouldn't, because we know that the predictability horizon of such events is around 2 weeks (or less) due to the chaotic nature of the atmosphere (and in particular the high non-linearity around vortex events), as you state in one of the next sentences. Consider rewording to make it clear from the beginning that this is not a shortcoming of the models, but an expected result given the nature of the system.

- page 18, line 395 ff.: I wonder whether the analysis of the vortex geometry adds much value to the already comprehensive evaluation. The paper is already very long, and this parts appears to add little to the understanding / quantification of relevant biases, and as stated in the last sentence, the sample size is maybe too small to make any robust statements. I also do not understand the reasoning given in line 396, in that the "shape or location would affect vertical wave propagation" - isn't it the other way round, i.e. the waves themselves lead to distortions of the shape and location? (In the absence of wave activity, wouldn't the polar vortex be a circular feature, with its location simply given by the radiative constrains on the maximum temperature gradient?). Therefore, my suggestion would be to consider to move this part to the appendix.

- page 20, line 434 ff, Fig. 9: does the composition of low- and high-top models make sense here, given the limited number of low-top models (two), which is completely dominated by the BoM model (as stated in line 441)?

- Fig. 11: The analysis of the final warming date in the SH is certainly an important quantity to consider, but I have to admit I don't fully understand the analysis presented in Fig. 11. In particular I guess it comes down to the question whether the "violins" are weighted in some way by the fraction of members that do predict a final warming in the given time frame. E.g. as is written, some of the early initialization dates do already predict final warmings, but as far as I can see it is impossible to tell from the Figure how many those would be. Likewise, I suspect that for the ini. dates in mid-November, for a number of instances the final warming has already happened. So overall it is, if I'm not mistaken, not possible to infer from the Figure at which times the model most likely predicts a final warming. Maybe some scaling of the "violins" by the relative number of members that do predict a final warming at this ini. date could solve this. Further, in the caption you could expand the sentence "... initialized closet to the beginning or middle of each month" with ", given on the x-axis" or so, otherwise this information has to be guessed by the reader.

- Page 28, line 520ff: I was a bit surprised by the statement that the polar cap temperature bias is specifically assigned to param. gravity waves (only). What leads you to this statement, why not planetary waves?

- page 26, line 560ff: As stated above, the number/percentage of those early final warmings from the present analysis is not clear to me (indeed the sample size is, I think, not mentioned in this specific analysis?) - possibly it is not in contradiction with observations, given that the sample size of modern satellite observations is only ~ 40 years (with essentially 2 "early final warming", counting 2002 and 2019).

- page 26, line 566: As above, I'd caution the authors on the formulation here - "the failure of the model to predict SSW beyond 2 weeks" implies that the model should be able to predict this, which is (to the best of our knowledge) not true.