Reviewer 1

This paper discusses the ability of long range forecast models to represent troposphere stratosphere coupling. The results come from a set of state of the art prediction systems and present an interesting set of results and metrics that could be used by operational centres developing future systems. I have most minor comments and suggestions.

We thank the reviewer for their constructive comments.

L30: The celebrated criterion of no wave propagation above an upper wind threshold (Charney and Drazin 1961) is a linear rather than nonlinear result.

We have added a citation to the end of the previous sentence about nonlinearity (Boljka and Birner 2020) to make it clearer that the sentence about Charney and Drazin is not a continuation of the previous idea.

L64-66: Suggest you remove this summary of results as it is repetitive of the Abstract and pre-empts the results section.

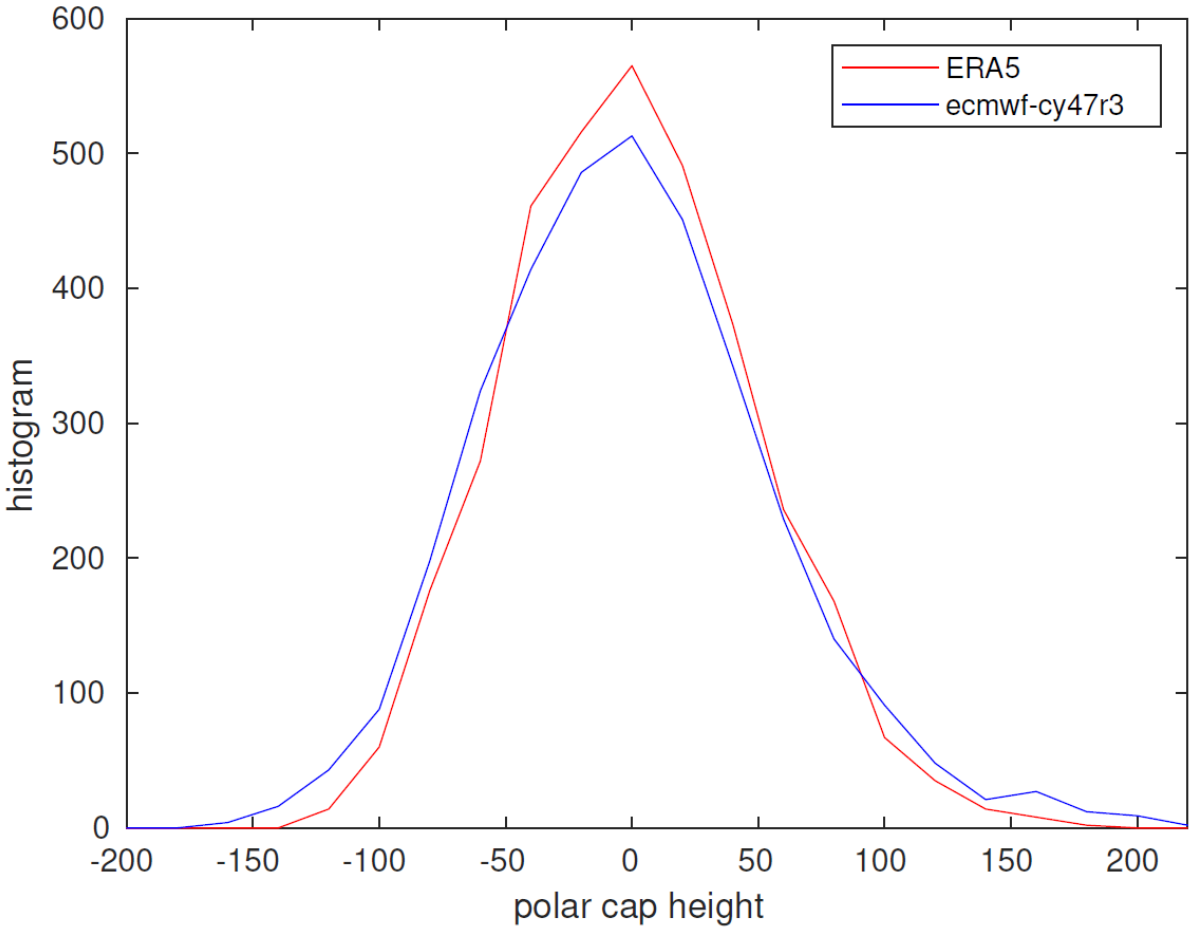
We have shortened this sentence.

L128: Why does an error in variance affect the correlations? After all, correlations are by definition insensitive to the amplitude of variability so do you mean regressions here? ;,

A model with a too-small coupling regression coefficient but a too weak variance bias that is even more dramatic, can have a too-strong correlation bias. We give several examples of this in the Results section (e.g., upward coupling for BoM). We have added "(the Results section provides several examples of such behavior)".

Fig.1 is very striking and suggests very large errors in the total variance of the models. Is it really correct that there are tens of percent errors in variance with too much in the troposphere and too little in the stratosphere? I have not seen this before and I think you should check and then emphasize this if it's robust.

We have computed the bias in the variance of polar cap geopotential height (Zcap) for additional tropospheric levels, and have confirmed that the models are systematically biased high at all levels up to 300hPa. We have confirmed this by creating histograms of Zcap for a few select models; the PDF of Zcap is indeed wider in the models than in ERA5. See below for Zcap500 for IFS.



Between 300hPa and 100hPa the models have a mix of biases, and then above 100hPa the variance is systematically too low. (This is compared to ERA5 subsampled to each model's available dates). We are not aware of any paper documenting the too-strong variance bias in the troposphere, and have now added mention of this into the methods section "We are not aware of previous work that has found such too-strong variance biases in the troposphere, and the causes and implications of these biases should be explored in future work.". (That models suffer from too-weak variance in the stratosphere is better known, and is worse in low-top models.)

Fig.2 would benefit from adding N Hem and S Hem labels.

The latitude range is indicated in the figure title

L155 and throughout the paper: In many cases it is really only some of the models that show the errors highlighted, for example in Fig.2a. PLease can the paper be phrased more carefully to say things like "models in general" or "models tend to" to avoid giving the impression that all models show the same errors?

We now use "most", "generally", etc.

L185-190, L245 etc: The paper tends to only reference very recent papers rather than giving a representative picture of current knowledge and following the *scientific convention of acknowledging those papers that first demonstrated ideas*. Some rewriting

is needed to better represent this. For example some wider discussion on the current knowledge of the effects of model lid height/degraded stratosphere would be welcome to put the results in wider context. Papers by Boville, J.A.S., 1984; Lawrence, J.G.R.,1997; Marshall and Scaife, J.G.R., 2010; Shaw and Perlwitz J.Clim., 2010.

These papers are now cited, though we prefer to include them in the introduction rather than in the results.

L265, L400: The underestimation of the heat flux variability in the stratosphere and upper troposphere is interesting. Is the underestimation of v*T* related to the underestimation of ENSO teleconnections reported in Garfinkel et al 2022 and Williams et al 2023? IS this also related to the so called signal to noise paradox in long range forecasts which appears to be clearer in the northern hemisphere than the southern hemisphere, just like the biases reported here? Perhaps some discussion would be useful on these points?

We indeed think that the lack of heat flux extremes are related to a poor simulation of teleconnection processes (and potentially related to ENSO), and we include suggestions to this effect near line 271. We have added to the discussion section near line 405 "(e.g., one possibility might be poor simulation of teleconnections arising from ENSO; Garfinkel et al 2022, Williams et al 2023)"

Garfinkel et al, in press finds no evidence for a signal to noise paradox in the stratosphere in the 7 S2S models they examined (7 of the 22 examined here). We are currently writing a follow-on paper which will examine the signal to noise paradox in all 22 of these model versions, but preliminary work indicates no S2N paradox in the stratosphere in any of these models. This follow-on work will include a much more detailed discussion of signal to noise characteristics of these models.

Figure 10: please provide a full caption for ease of reading.

We have provided a full caption (this is now Figure 11 in the revised version).

L362-364: Please again provide wider referencing for the surface impact of the stratosphere e.g. Baldwin and Thompson Quart. J Roy. Met. Soc. 2009, Kidston et al., Nat. Geosci., 2015.

added

L450: This is a potentially important point and should be moved to the earlier methods section.

We have added this to the methods section as well (near line 89)