Replies to reviews for:
Large ensemble assessment of the Arctic stratospheric polar vortex

WCD, Ales Kuchar et al.

1 Reply to reviews

Dear Daniela Domeisen,

thank you very much for helping in the publishing process of this paper. We appreciate the helpful comments of the reviewer #1. We revised the manuscript accordingly, such that we hope it can then be published in WCD. Please find our point-by-point answers to the comments below. Furthermore, we changed the title according to your own suggestion.

Best wishes
Ales, Roland, Maurice and Christoph

2 Ref #1

Dear anonymous reviewer #1,
thank you very much for your valuable comments. Please find our point-by-point answers (in blue) to your comments below.

2.1 MAJOR COMMENTS

1. My greatest concern regards the ROC curve analysis, since, from the explanations in the Methods and Supplementary Material, I still fail to understand how you build and interpret it. In the Supplementary you say «the false positive rate indicates the probability that the ensemble is simulating an event, even though there was no event observed in the reanalysis». Do you check this day by day, or year by year, or using some other selection criterion? I have in mind contingency tables for ensemble forecasts, where false positive rate corresponds to the ratio of forecasted positives in a period of time when a negative realisation was observed, the true positive rate is the ratio of forecasted positives when a positive realisation was observed, but cannot imagine what this represents for a climate simulation. In fact, in ensemble climate simulations, differently from initialised forecasts, you do not expect an exact temporal correspondence between simulated and reanalysis events (splits and displacements) because of internal climate and weather variability. So, what do you mean by true and false positive rates in

this case? How do you obtain the ROC curve, in the detail of the computation? What does it mean and how do you interpret it? And based on your answers to these questions, how are you then able to discuss over and under representation of SPV displacements and splits, and the model performance in distinguishing between SPV displacements and non-displacements? Because of my not understanding the above, all of Section 4 is still unclear to me.

If I may add a suggestion, a much less convoluted way of understanding the model fidelity in the frequency of split and displacement SSWs would be to compute the monthly frequency of the events across models and compare this with the reanalysis frequency. Hopefully, relations with the SPV diagnostics might become more apparent.

We check the forecast and observation probability of crossing split or displacement thresholds for daily data of the months from November through March in the NH. The detailed procedure how Fig. 2 and therefore true positive and false positive rates using a set of increasing probability bins from the upper right corner to the lower left corner are calculated can be found in the notebook ROC_values.ipynb (Kuchar and Öhlert, 2024). As shown therein the dichotomous contingency are calculated along the probabilities of whether the model simulates displacements or splits across its ensemble members. Therefore, there is no assumption about the temporal correspondence of these events but the probability distributions between ERA5 and model ensembles are assessed. To avoid that for readers these points are unclear, we added this sentence to the paragraph dedicated to the ROC curve in the manuscript.

As explained in the manuscript, we interpret higher AUC (arc of the corresponding ROC curve more closely to the upper left corner) as a model ensemble discriminates split or displacement SSW events better than a model ensemble with lower AUC (arc of the corresponding ROC curve more closely to the diagonal line). For further details about the application and interpretation of the ROC diagram we refer to Wilks (2011) or Jolliffe and Stephenson (2012). We added this sentence to the paragraph dedicated to the ROC curve in the supplement.

For example, SPV displacements are less frequent in models that simulate a bias towards higher values of centroid latitude. However, the ROC diagram is insensitive to certain types of biases (Kharin and Zwiers, 2003), as a biased model may still have good statistical resolution (e.g. CESM2). The ROC can still be considered as a potential skill (usefulness) when the model is correctly calibrated (Wilks, 2011). This fact can explain the differences between CanESM2 and CanESM5. While CanESM2 distinguishes between displacement events and stable SPV conditions better than CanESM5 according to the AUC, CanESM5 performs much better in the rank histogram analysis for the centroid latitude at 10 hPa.

Thanks for your suggestion to compute the monthly frequency of the events across models and compare this with the frequency of reanalysis. Nevertheless, we assess the model fidelity in the frequency of split and displacement SSWs in Fig. 9. The related discussion can be found in Section 4. Based on this, we concluded that the geometric SPV biases indicates also biases in split and displacement frequency as in Seviour et al. (2016) and Hall et al. (2021). However, this does not necessarily imply that bias-corrected models simulate split and displacement frequencies according to the reanalyses (Wu and Reichler, 2020).

2. Section 3 would benefit from a filtering of the most important information. The many details are difficult to follow.

Sect. 3 mostly is a mere description of the analysis results. Obviously not all points mentioned here are discussed later in the paper, but we do aim at a certain degree of completeness in this results description. As we already performed a fair bit of

condensing of this section in the process of the paper writing and the reviewers statement is very unspecific, we do not see the need for action here. Moreover, the reviewer was asking for additional descriptions and analyses for this section in the first round of reviews, which makes this statement appear somewhat ambiguous.

3. In the Abstract and in the Conclusions I suggest to remove the technical terms of the SPV diagnostics and to give an intuitive interpretation of the result. A reader should understand the main conclusions without having to browse the details of the Methods section.

We enriched the technical terms in the abstract and in the conclusions section with intuitive descriptions, or removing bits, or extending with the intuitive explanations.

4. In Table S1 you could evidence, for each diagnostic, the model/s with the weakest best performance, and refer to the table in the Conclusions Section.

Done.

2.2 TECHNICAL CORRECTIONS

Line 9:You should explain the results without mentioning the technical names of the different diagnostics, since these haven't been defined yet. (See Maj. Comm. 3)

Done, see answer to Maj. Comm. 3

Line 34: «mid-latitude zonal mean zonal wind reverses...».

Corrected accordingly.

Line 38: «into two separate vortices».

Corrected accordingly.

Line 74: Shift and merge sentence with that in line 67, preceeding the outlook paragraph.

Modified accordingly.

Line 92: Remove «, or in other words, ERA5 data serves as ground truth in our analysis.». The sentence is not useful and out of place.

Removed accordingly.

Line 94:As in previous comment, remove «however, clearly not as an absolute truth».

Removed accordingly.

Line 96: replace with «when satellite observations were assimilated in ERA5.». Lines 104-112: organise in a list of bullet points.

Replaced with:

'...when, among others, satellite observations were assimilated to ERA5...'

Paragraph 117: Shift this at the beginning of the Results' section. It is out of place in the Methods.

Corrected accordingly.

Line 128: I repeat a comment from my previous review. Is the sentence « The histogram counts of all bins greater or equal k are then increased by one » correct? From what I read from Sect. 2 in Hamill et al. 2001 (your ref.), only one of the bins is updated at each timestep, corresponding to the forecast ensemble bin where the observation falls.

Thanks for your comment. We modified the sentence accordingly.

Line 133: Not clear. Please be more precise in the wording.

We modified the sentence to increase clarity.

Line 140: How is the U-shape (spread) computed? The use of the word «spread» is confusing in this context, as it reminds of ensemble spread, which is not the case here.You could use «spread bias» or «dispersion bias» instead, here and in the rest of the manuscript. You could also mention that 0 values in the statistics indicate perfect model dispersion (if I understood correctly!).

The spread is computed according to the methodology by Jolliffe and Primo (2008). The code that was used to produce all plots in this study is available via Zenodo (Kuchar and Öhlert, 2024). We modified the sentence so that the context is clear.

Line 172: What is the methodology from Hall et al. (2021)?? Please specify.

We modified "the methodology from Hall et al. (2021)" to "the methodology to detect displacement and split SSWs from Hall et al. (2021)".

Line 176 and table S2: Please specify in the text and caption if this is the SSW list relative to your definitions, or if it's taken from literature. The interesting point would be to show two lists of dates and types, yours and that used in literature, and discuss the differences.

As mentioned in the previous sentence, it is determined with the method of Hall et al. (2021). We add "Using this methodology," in the beginning of l176. As we state in the manuscript, this method are within the uncertainty of other methods.

Line 184-187: Make shorter – you should summarise in a few words the Methods' description at the moment.

We removed the sentence:

"Exceptionally high aspect ratio values are of particular interest for SPV dynamics as they are often associated with SPV disturbances, in particular indicating splitting events (Seviour et al., 2013)."

Line 251-255:condence UKMO model results, and say in clearer language.

Condensed and made clearer.

Line 264-66: I am not convinced by this sentence. Remove if not adequately motivated.

We modified the sentence providing a reasoning.

Line 274: replace «too» with «anomalously».

Replaced accordingly.

Lines 276-277: replace «small» with «weak» and «large» with «strong».

Replaced accordingly.

Lines 385-401: The discussion is quite dispersive, because of alternating from literature to your own results multiple times. I would suggest to separate the two clearly within the paragraph.

Through minor restructuring we made the paragraph clear and non-dispersive.

Lines 402-413:Very long discussion on a topic that is outside your scope.You should filter and compress this paragraph.

We have compressed this part so that it meets the reviewers taste for such a discussion.

References

Hall, R. J., Mitchell, D. M., Seviour, W. J., and Wright, C. J.: Persistent model biases in the CMIP6 representation of stratospheric polar vortex variability, Journal of Geophysical Research: Atmospheres, 126, e2021JD034 759, https://doi.org/10.1029/2021JD034759, 2021.

Jolliffe, I. T. and Primo, C.: Evaluating rank histograms using decompositions of the chi-square test statistic, Monthly Weather Review, 136, 2133–2139, https://doi.org/10.1175/2007MWR2219.1, 2008.

Jolliffe, I. T. and Stephenson, D. B.: Forecast verification: a practitioner's guide in atmospheric science, John Wiley & Sons, 2012.

Kharin, V. V. and Zwiers, F. W.: On the ROC Score of Probability Forecasts, Journal of Climate, 16, 4145 – 4150, https://doi.org/https://doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2, 2003.

Kuchar, A. and Öhlert, M.: VACILT/reliability_LE: Fourth release of our code repository related to reliability of large ensembles, https://doi.org/10.5281/zenodo.10620326, 2024.

Seviour, W. J., Mitchell, D. M., and Gray, L. J.: A practical method to identify displaced and split stratospheric polar vortex events, Geophysical Research Letters, 40, 5268–5273, 2013.

Seviour, W. J., Gray, L. J., and Mitchell, D. M.: Stratospheric polar vortex splits and displacements in the high-top CMIP5 climate models, Journal of Geophysical Research: Atmospheres, 121, 1400–1413, 2016.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.

Wu, Z. and Reichler, T.: Variations in the frequency of stratospheric sudden warmings in CMIP5 and CMIP6 and possible causes, Journal of Climate, 33, 10 305–10 320, https://doi.org/10.1175/JCLI-D-20-0104.1, 2020.