GENERAL COMMENTS

The manuscript assess the representation of the geometry of the winter SPV in a few model ensembles, and compares the model ensembles' SSW and SPV frequency to the observed. To persue this analysis, geopotential height data over different stratospheric levels and at daily frequency is taken from the historical simulations of climate models participating in CMIP5 and CMIP6. Based on their results, the authors filter the models that are best at representing the SPV spatial distribution and SSW frequency, and conclude that a good calibration of the SPV shape does not necessarily correspond with an adequate representation of split and displacement SSWs. Outlook on the possible reasons underlying inter-model differences is also provided.

Although the topic is of interest for the community and the results might be useful for the institutions developing the different models, I find that the present manuscript is of poor quality. This is because (i) the text is fragmented and difficult to follow, (ii) the language is often imprecise, (iii) the methods are not clearly presented, (iv) the many differences across models and resolutions are difficult to grasp. The whole manuscript needs to be generally and carefully revised before being fit for publication in WCD. In the following, some suggestions that should help improve the afore-mentioned issues are included.

SPECIFIC COMMENTS

1. To improve the text please work on fluency and on the precision of the wording. Often adjacent sentences repeat the same concepts, or are fragmented, or badly connected. For example, see some of the points in the TECHNICAL CORRECTIONS.

2. The methods need to be expanded and/or clarified. Specifically, the meaning of the geometric diagnostics can be better explained, the methods behind the rank histogram and ROC/AUC calculations are not clear.

3. It would be useful to compare the list of ERA5 SSW split and displacement events determined with geometric methods (i.e., those identified here) to those identified with more traditional methods (lists of events and their types are available online or in literature). Additionally, please provide the magnitude of the mean SSW frequency for each model, together with the AUC.

4. The Discussion and Outlook section is dispersive, and should be organised more clearly, without mixing your results (discussion and limitations) with the outlooks. Probably, some paragraphs in the Results section would fit better in the Discussion section (e.g., see TECHNICAL CORRECTIONS). Finally, it would be useful to add a table summarising the performances/biases of each model, by geometric diagnosis and by type of SSW event.

5. How does the map of the mean gh10 climatological bias look like for each model ? Is it relatable to the more specific rank histogram results ? If so, such a figure could be included at the beginning of the Results' section.

6. It would be useful to show an example of a ROC curve for one model, displaying both the split and displacement curve. This could go in the Methods or in the Results, and would be useful for explaining how the ROC curve is constructed.

7. Since you mention the diagnostics of the vertical structure of the SPV in the abstract, it is advisable to show a couple of meaningful plots (or statistics values) for the other stratospheric levels in the main manuscript.

8. You should mention any previous use and results of geometric diagnostics in your Introduction section. Or did I miss this ? On the other hand, SPV projections are not your main focus, so you should mention them briefly to motivate your work or more extensively only if directly related.

TECHNICAL CORRECTIONS

General

Please check the usage of the adjectives 'low'/'high' (e.g. related to bias), and replace with 'weak'/'strong' or 'negative'/'positive' when ambiguous.

Abstract

Line 1 : « each hemisphere ».
Line 1 : The word « phenomenon » is not adequate for describing the SPV. Please change.
Lines 5-13 : Abstract would gain in clarity if the description was separated into geometric diagnosis and split-displacement diagnosis (e.g. line 6 shifted after the geometric diagnosis). The fluency of the text would improve.

Introduction

Line 25 : « relate~~s~~ ».
Lines 25-26 : The term « windiness » is ambiguous. I would suggest to replace the first two elements in the list with « the position of the jet and the precipitation patterns over Europe and the Mediterranean region ».
Line 27 : « **where** the SPV is highly variable ».
Line 31 : « from **a** final warming».
Lines 32-33 : The sentence is difficult to read, could you please reformulate ?
Lines 37-39 : The sentence is out of place in the present position. Please fit adequately in the discussion of literature.
Lines 57-58 : Do you mean that a limitation of the single-member analysis is that one can't distinguish between natural and inter-model variability ? This is not the meaning emerging from the sentence, please correct or clarify.
Line 63 : « ~~will be~~ is », for consistency of the verb tense.
Lines 63-63 : The following sentence/content should be shifted in an adequate position, before the outline. « Due to the large SPV variability and the high SSW frequency in the NH, we limit our analysis to the NH SPV. »

Methods

Lines 70-72 : You can remove the reasoning here, as it is already discussed in the Introduction.
Line 76 : Please change to « climate models used **in** our analysis **are described** in table 1 ».
Line 78 : Inconsistency between the two sentences. Please modify.
Line 80 : The first part of the line is accessory, please remove or include in compact form as part of the previous sentence.
Line 91-92 : The last sentence gives a qualitative description of the different moment diagnostics, but the link with the diagnostics (two sentences back) is not obvious – I suggest giving this information just following the list of diagnostics. Moreover, an extension of their description, here or at the beginning of each section in the results, would be welcome for any non-expert reader.
Lines 95-99 : The description of rank histograms is confusing. Please describe precisely and clearly how the counts within each histogram are determined, and why the distribution should be uniform in a well-calibrated forecast. 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.
Line 119 : Mention where and how the perfect model range method is used within the present work.
Line 122-24 : It is not clear to me if your definition of displacements and splits corresponds to the description you give in these lines. Could you make it more explicit ?
Line 126 : Insert ref. for ROC and AUC.

Line 127 : Can you describe in this paragraph how the thresholds for drawing the ROC curve are computed ? And what you mean by true and false positive rates in a historical simulation (rather than in a forecast) ? An example of a ROC curve would also be welcome, including descriptive labels for x and y axes.

Results
Line 139 : It would be useful to specify again, within this sentence, that high values of the aspect ratio correspond to split events.
Line 143 : No need to repeat the aside « which is considered as ground truth for the analysis ».
Line 144 : « ~~are biased simulating~~ simulate ».
Line 145 : « ~~high~~ positive ».
Line 146 : How is the significance of the bias computed ?
Lines 148-49 : You could merge the last two sentences of the paragraph, there is a partial overlap.
Line 152 : Does « low » mean 'weak' or 'negative' ?
Line 153 : « model ensembles ».
Line 165 : Please specify pressure levels each time.
Line 192 : « high » means positive ?
Lines 200-01 : I don't understand the sentence starting by « However.. ». Please clarify or remove.
Lines 212-14 : Please specify large objective area or strong SPV, or describe what large SPV means. Check the same in the rest of the subsection.
Lines 221-228 : These lines are better fitted for the Discussion section.
Line 240 : Specify that the assertion is valid for displacements. Or else present only the displacement results in the first introductory sentence.
Lines 242-44 : I find the discussion quite vague. Could you clarify ?
Sections 4.1 and 4.2 : The second part of each subsection is better fitted for the Discussion section, specially when the manuscript's results are compared with previous literature.

Discussion and Outlook
Lines 314-17 : These lines are rather vague and un-related from the results and conclusions of the manuscript. Please remove if not otherwise justified.
Line 323 : It would be good to acknowledge that you are not analysing or describing GW parametrisation in this work.
Line 328 : Is this phrased correctly ? If so, it is difficult to understand what you are concluding.
Lines 348-352 : This discussion is too general – it should connect more with the results of your work if you want to include it.

Conclusions
Line 360 : What do you mean by « stability » ?
Line 370-71 : Can you tell why your results are different from previous literature ? Is it a different metrics ?
Line 375 : At some point you had some contradictory results regarding resolution and SPV representation (see e.g. Line 199). Is it worth taking this into account in the present assertion ?
Line 380 : « variability » instead of « behaviour ».
Line 380 : Replace « set the basis » with « can constitute/be a reference ».
Line 380 : Rephrase « with individual models changing individual model components ».
Line 382 : « the SPV spatial variability ».
Line 383 : Please rephrase « for finding the adjustment screws to improve their performance ».

Figures
Figure 1-5 : Increase font size of model name and statistics above each plot. Mention the different models and the respective letter labels in the caption.