The manuscript is substantially improved in terms of text, language and clarity. However, my recommendation for publication is conditional on overcoming my first major comment. A list of major and specific comments, referring to the line numbering in the marked-up version of the manuscript, follows.

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.

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

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.

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.

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)

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

Line 38 : « into two separate vortices ».

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

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.

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

Line 96 : replace with « when satellite observations were assimilated in ERA5. ».

Lines 104-112 : organise in a list of bullet points.

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

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.

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

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 !).

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

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.

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

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

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

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

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

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.

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