Review of "Predictability of cyclones associated with heavy precipitation events in the Sahara"

Ling et al. investigate the predictability of surface cyclones associated with heavy precipitation events (HPEs) across the Sahara, using ERA5 reanalysis, satellite-derived HPE data, and ECMWF sub-seasonal reforecasts. The authors evaluate forecast skill at lead times ranging from 0.5 to 15.5 days using an area-based, feature-oriented verification framework and find that predictability varies strongly with season and geographic subregion. In particular, short-range skill is highest in winter for cyclones in the northern Sahara, while medium- to extended-range skill is higher in summer, especially in the southwestern Sahara. The authors also identify that the large-scale Rossby wave patterns associated with the winter cyclones are linked to both enhanced and reduced predictability. The manuscript addresses a relevant and under-explored topic: extreme precipitation predictability in arid regions, which has both scientific and societal importance given the growing frequency of high-impact flood events.

The study is generally well structured and the feature-oriented verification framework offers a useful methodological contribution. However, I have several comments that I would like to see addressed before possible publication.

GENERAL COMMENTS

Introduction:

1. Lines 33-34: "on the order of magnitude of the cyclone climatology" is unclear. Do the authors mean that the biases are of comparable magnitude to the climatological cyclone frequency? Please rephrase for clarity.

Methods:

2. The HPE catalog (Armon et al., 2024) uses IMERG V06. Since the publication of that paper, V07 has become available, which includes significant algorithmic improvements, particularly for arid regions where gauge-calibration data are sparse. Would it be possible to rerun the analysis for V07? If not, I think it should be explicitly mentioned that V06 is used and the limitations of the dataset should be discussed.
3. Lines 108–109: The authors state that the cyclone detection algorithm was applied to the 10 perturbed ensemble members, excluding the control run. What is the rationale for excluding the control run? Including it would modestly increase the sample size and may affect the skill statistics.
4. Lines 112–115: The HPE-associated cyclone identification method relies on selecting the cyclone whose center is closest to the HPE precipitation mass center at 12 UTC on the date of maximum precipitation volume. Several aspects of this procedure require clarification: (a) Why is 12 UTC specifically chosen rather than, for instance, the time of peak precipitation within the event? (b) What fraction of cyclone related HPEs are discarded due to the 2000 km and/or subregion criteria? This has implications for how representative the retained sample is. (c) If multiple cyclones are within 2000 km of an

HPE mass center (does this occur?), is it always the case that the nearest one is actually dynamically responsible for the HPE?

5. Lines 119–120: The ad-hoc study region is derived by expanding the observed cyclone mask by 6 degrees. How sensitive are the skill scores to this choice? Some discussion or sensitivity analysis would be beneficial.

6. Lines 144–147: Could the authors argue, why the threshold of 30% for defining "hit members" was chosen? How sensitive are the spatial distributions in Figures 5 and 6 to this threshold? This is particularly relevant for the interpretation of high- versus low-skill regions.

Results:

The text in Section 3, specifically Section 3.1, focuses primarily on describing the figures rather than interpreting them. I believe the manuscript would benefit from shifting the emphasis toward a discussion and physical interpretation of the results.

7. Figure 3: The black solid line representing climatological cyclone frequency is defined as the weighted cyclone frequency at each grid point of each cyclone area. This definition is not immediately intuitive. A clearer explanation in the Methods section of how this climatological baseline is computed would be helpful, as it is central to the interpretation of POD in summer extending beyond the climatological frequency at lead times greater than 10 days.

8. Lines 176-179: I don’t fully understand what the authors are communicating here. Specifically, they describe that POD and FAR values for spring and fall are close to those of summer for mid- and extended ranges, while I would argue that the values for spring are closer to winter? Could the authors maybe rephrase to clarify, and be more quantitative when comparing seasons with each other?

9. Figure 4 and Lines 181–187: The description of the black lines indicating the seasonal average MSLP standard deviation is missing in the caption. The MSLP RMSE is compared against this seasonal average MSLP standard deviation as a benchmark for predictability. MSLP standard deviation reflects variability across all dates in the season, not just HPE days. Could the authors explain why they do not compare against the MSLP standard deviation computed specifically over HPE days?

10. Lines 187-188: I do not agree with the interpretation that RSME stabilizes after 13.5 days lead time. In my opinion, three data points are not sufficient to make this statement.

11. Figures 7 and 8: The composite anomaly fields for GH500, MSLP, and T850 are computed by subtracting the monthly climatological mean. Statistical significance is assessed for GH500 using a Student’s t-test at $\alpha = 0.05$, but no significance testing is applied to the MSLP and T850 anomaly fields shown in the same figures. Either significance testing should be extended to all displayed fields, or the authors should explicitly acknowledge that the MSLP and T850 anomalies shown may not be statistically significant.

12. Figures 7 and 8: The discussion in Sect. 4 frequently refers to Rossby wave patterns as drivers of both high and low forecast skill. However, the composite analysis averages over many events and may obscure event-to-event variability. Is it possible that the composite wave patterns emerge primarily from a subset of extreme events? Some measure of within-group variability (or representative case studies, which I acknowledge would be out of scope) could strengthen the interpretation.

Discussion and Conclusions:

13. I am missing a discussion of the observational uncertainties and their impact on the study results. For example, it is known that the precipitation extremes are not well represented in IMERG in data sparse regions like the Sahara.
14. Lines 280-282: Could the authors elaborate, how this improved model accuracy could be achieved? I.e., how could improved understanding translate to improved model accuracy?
15. Lines 283–285: The discussion of predictability for southern Sahara cyclones is brief and primarily consists of directions for future work. Given that these cyclones show higher skill at extended lead times, a more substantive physical discussion of why thermally driven and monsoon-type systems might be more predictable at longer lead times would strengthen the manuscript. Could, for example, African easterly waves play a role here?
16. Lines 289-294: I think the motivation for the method presented here should be integrated into the first paragraph of Sect. 4, before the results are discussed.
17. Lines 298-300: “Moreover, since … complicated methods”. I believe this sentence is not needed.

In the Data Availability section, the authors reference the ERA5 and S2S datasets. However, the analysis scripts used to produce the results and figures are not mentioned. I would encourage the authors to make the analysis code publicly available, as this would improve reproducibility.

FORMALITIES

Text:

- Line 48: "normally dry on their poleward side cantrigger" appears to be a typo. "cantrigger" should read "can trigger".
- Line 115-116: "An example of this approach, showing the association of the nearest cyclone with a HPE during 20–24 November 2024 is in Fig. A1." However, in Fig. A1 the caption reads "20–24 November 2014." I believe the date in the text is incorrect, as the analysis period is 2000-2020, and should be verified.

Figures:

- Figures 3 and 4: The box plots would benefit from a brief explanation in the caption or methods of what the boxes and whiskers represent (e.g., interquartile range and 5th–95th percentile, or some other convention).

- Figure 4: The font size of the tick labels on the x and y axes are not the same.
- Figures 5 and 6: The color scale used to depict the number of hit ensemble members (ranging from 0 to 10) uses a sequential colormap that makes it difficult to visually distinguish high-skill from low-skill cyclones at a glance. A diverging colormap centered on 5 (i.e., half of the ensemble) or the use of distinct categorical colors for the groups defined in the text (e.g., hit count > 5 vs. ≤ 5) would improve readability.
- Figures 7 and 8: The contour intervals for MSLP and T850 anomalies are not stated in the caption. Please add this information. Also, the blue T850 contours are difficult to read over the GH500 color shading in some panels; maybe consider using a different line style or color.

I hope that my comments will be of some help to the authors.