Reply to the Reviewers' comments: "Predictability of cyclones associated with heavy precipitation events in the Sahara" by Ling et al.

Dear Editor,

We would like to thank the reviewers for their useful comments. Please find our detailed responses to the reviewers' comments and suggestions below *(in blue)*.

In the revised version, we have addressed the reviewers' concerns regarding the methodology and interpretation of the results. All changes have been included in the revised manuscript and are indicated in the annotated manuscript.

The main changes to the manuscript include:

- We have added more explanation on the Methodology used in this study.
- We have thoroughly revised the results section for better clarity, and extended the interpretation of the results and the discussion.
- We have revised Fig. 3, Fig. 5, and Fig. 6 to allow clearer interpretability of the temporal and regional variability of cyclone skill.
- We performed sensitivity tests for different thresholds and parameters used in this study.

As requested, we state that all figures in the manuscript are our own.

Kind regards,
Moshe Armon (on behalf of all authors).

Reviewer 1:

In this paper an attempt is made to investigate the predictability of high precipitation cases related to cyclones and the dynamic drivers. The topic is very interesting for the Mediterranean region weather and climate. I recognize the huge amount of processing data and the complexity of the methodology. However, I have some concerns:
Thank you.

- lines 85-90: The method used to retain from 42,000 HPEs only around 12,500 cases in which cyclones 90 were associated with HPEs, «according to a Monte Carlo cyclone-association test performed in Armon et al. (2024) that determines whether HPEs occur closer to a cyclone than would be expected by chance, based on repeated comparisons with randomly selected cyclone dates» is not clear to me. I think that the authors should be more specific.

We thank the reviewer for this comment and agree that the description of the Monte Carlo cyclone-association test required clarification.

We have revised the manuscript to describe the method more explicitly. In brief, for each HPE we compute the minimum distance to the nearest cyclone mask, and directly classify overlapping cases as associated. For non-overlapping cases, the observed distance is compared to a null distribution derived from 100 cyclone fields sampled from randomly selected dates. If the distance falls below the 5th percentile ($\alpha = 5\%$) of this distribution, the proximity is considered unlikely to occur by chance.

Applying this procedure across all HPEs yields a representative threshold distance of ~180 km, below which HPEs are classified as cyclone-associated. Because the random samples include cyclones from all dates (including rainy conditions), this approach provides a conservative estimate and reduces false associations. The manuscript has been revised accordingly (Sect. 2.2.1).

- Line 115: " Cases where the detected cyclone is located at a distance ≥2000 km..". How this threshold value derived? It seems rather arbitrary

We thank the reviewer for this comment. Indeed, the 2000 km threshold is not representing a specific physical constraint, but rather a pragmatic constraint introduced by our simplified attribution framework. In this study, we do not track cyclones continuously in time, but instead examine their proximity to HPEs once per day (12 UTC). Therefore, a threshold is required to avoid attributing HPEs to remote cyclones that are unlikely to be related.
The 2000 km threshold reflects a compromise between allowing some flexibility in attributing an HPE to a cyclone (given the single daily snapshot) and preventing systems that are clearly too distant to be physically related. To assess the sensitivity of this choice, we repeated the analysis using alternative thresholds (1000–3000 km, 500 km intervals).

While reducing the threshold from 2000 to 1000 km reduces the number of cyclones attributed quite substantially (~18%), increasing it to 3000 km results in only a minor increase (~4%). This indicates that increasing the threshold much beyond 2000 km has little impact on the results, whereas decreasing it would substantially reduce the sample size, which has its own drawbacks. We have clarified this point in the manuscript: “The 2000 km threshold represents a pragmatic upper bound to avoid HPE attribution to remote cyclones; sensitivity tests (1000–3000 km) indicate that results are only weakly affected by increasing the threshold beyond this value (~+4% at 3000 km), whereas smaller thresholds would substantially reduce the sample size (~−18% at 1000 km).”

- Line 125: how do you define cyclone mask? The MSLP does not measure directly the intensity. The Laplacian of p is such a measure.

We thank the reviewer for the comment and pointing the problem out. Cyclone masks, based on the methodology developed by Wernli and Schweirtz (2006) and adapted by Sprenger et al. (2017), are defined using the outermost closed contour (interval = 2 hPa) around only one MSLP minimum. MSLP is indeed not a direct measure of cyclone intensity, and is not used for this purpose, but only for the detection algorithm, and for the verification of MSLP field in the forecast (against ERA5). We have, therefore, removed the “intensity” terminology from the manuscript and revised the text accordingly (L108–111).

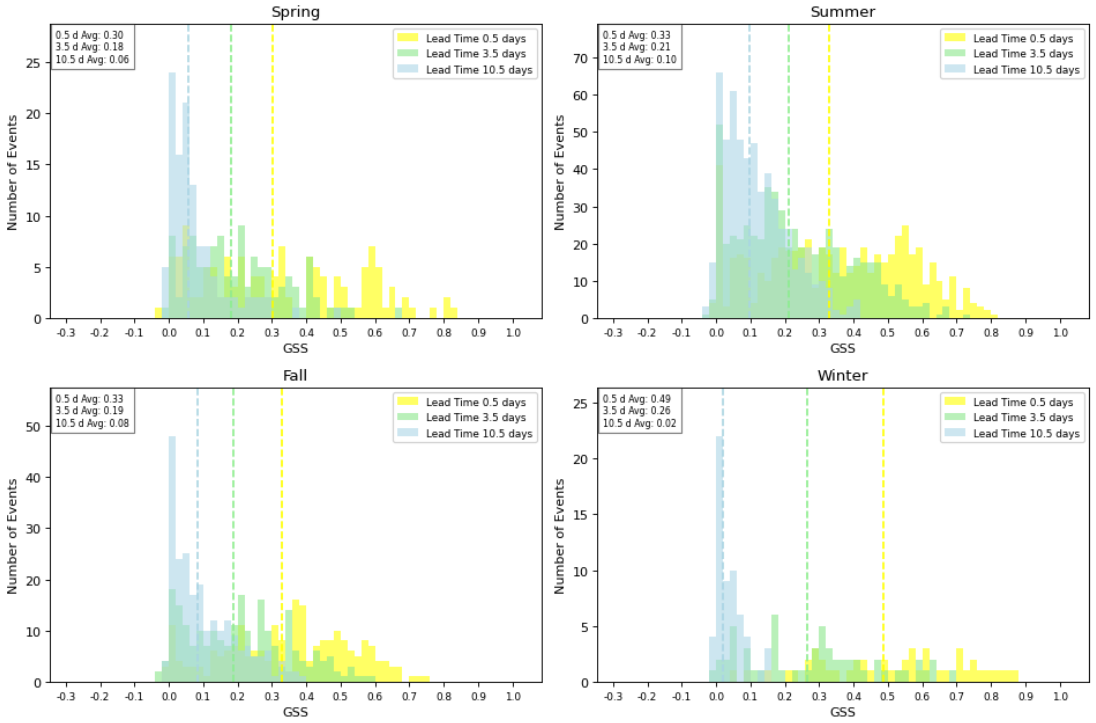
- Line 146: Although the authors state that the 30% value is arbitrary, they should document this value (e.g based on operational experience, statistical analysis, sensitivity tests)

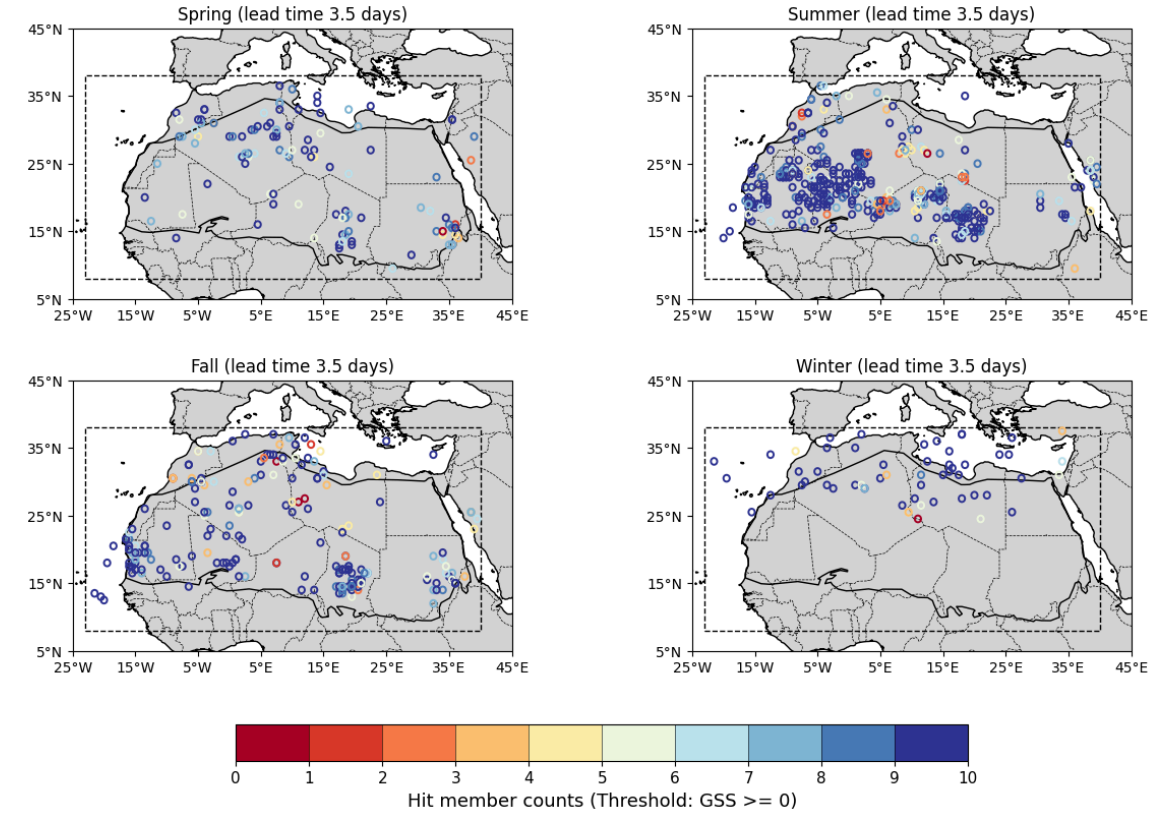
We agree with the reviewer that the 30% threshold is not uniquely defined and therefore should be justified. Our purpose in using this threshold was not to distinguish whether the forecast is merely better than climatology (as in using a 0% GSS threshold), but rather to highlight regions and seasons with relatively higher versus lower predictability. While a threshold of 0% GSS has a clear interpretation, since it marks where the forecast outperforms a random forecast, in practice it produces a largely binary picture with many events classified as equally skillful and therefore provides less contrast for identifying meaningful spatial and seasonal differences in forecast quality.

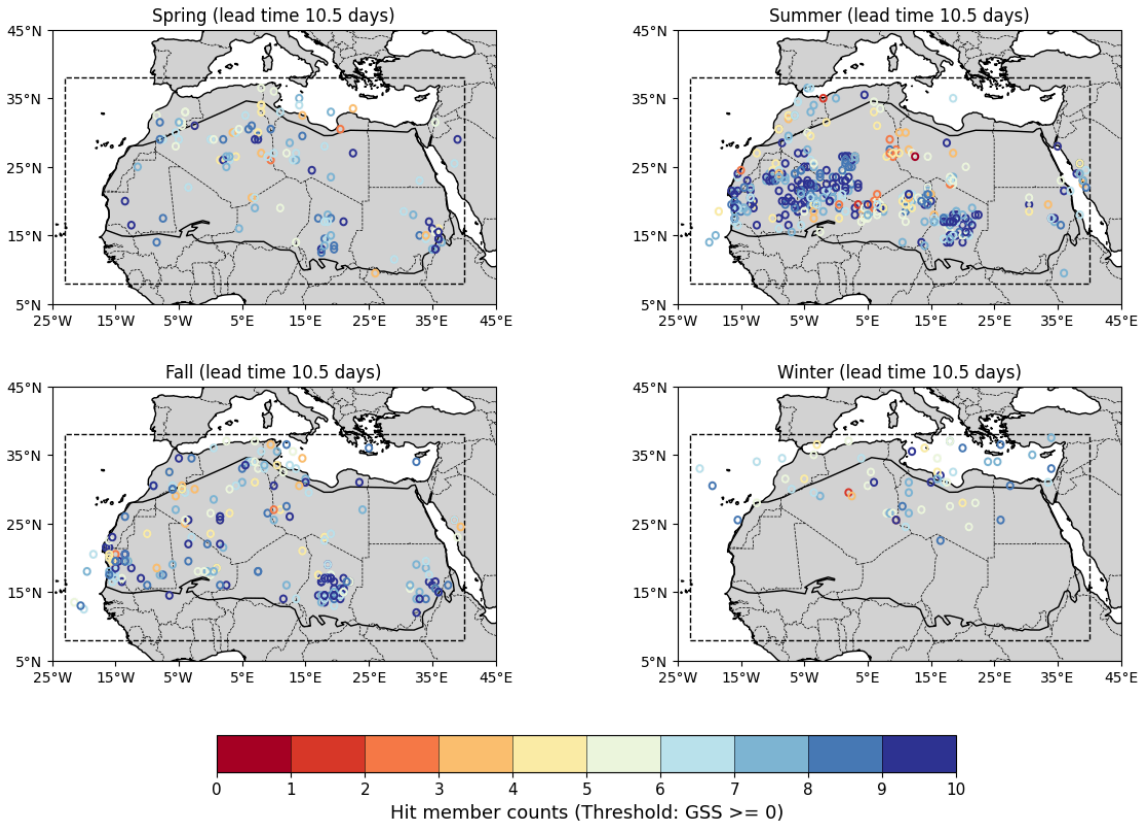
Therefore, we retained the 30% threshold, which better emphasizes where the model outperforms climatology by a substantial margin and helps distinguish regions and seasons with better or worse predictability. To assess the sensitivity of this choice, we repeated the analysis using alternative thresholds of 0%, 20%, and 40% (see figs. below). The resulting spatial patterns are qualitatively similar, indicating that the main conclusions do not depend strongly on the exact threshold value, while the 30% threshold provides, in our opinion, the clearest visualization of the spatial structure of forecast skill. Furthermore, to emphasize the binary nature of the 0% GSS threshold, we show here the distribution of GSS value for each season,

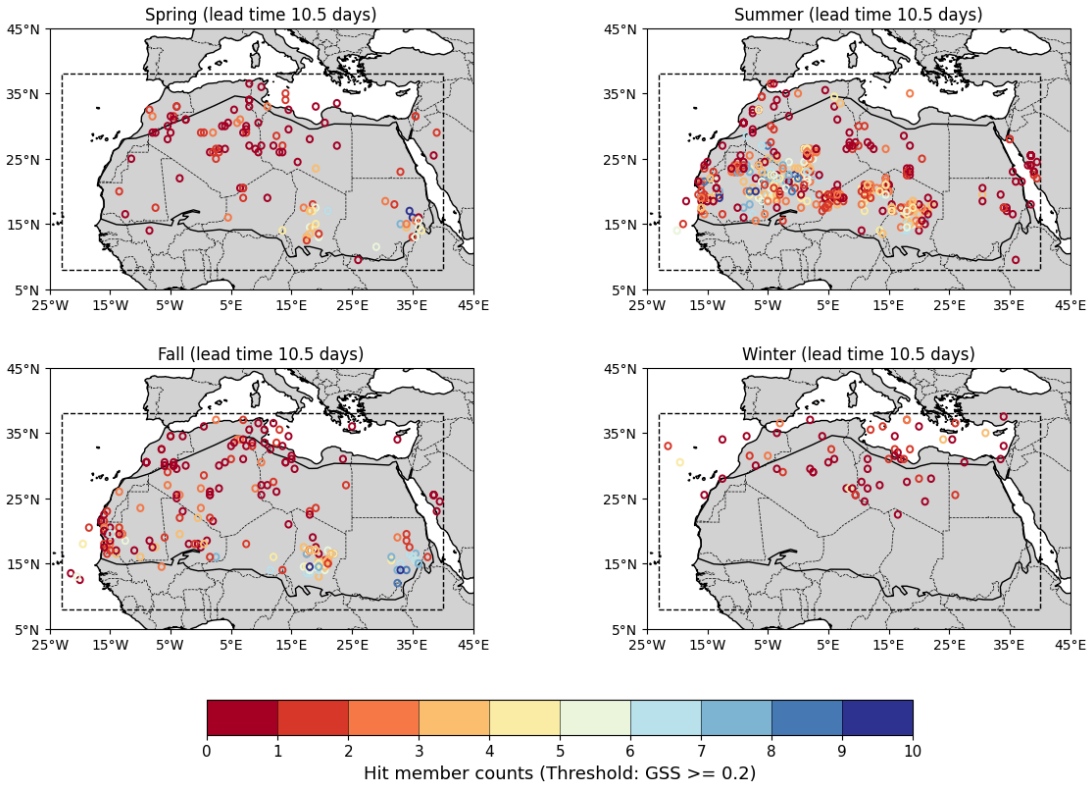
for three different lead times, which shows very few forecasts dropping to <0%. We have clarified this point in the manuscript (L168–171, and Sect. 3.2).

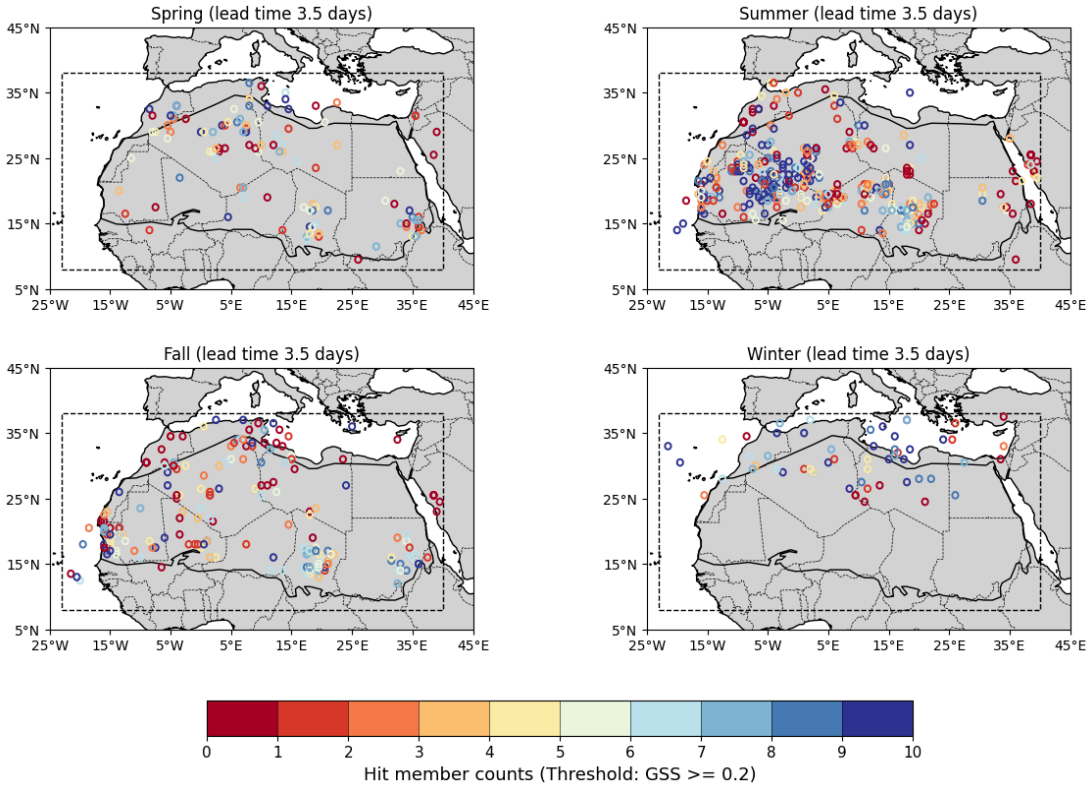
In addition, we referred to a suggestion from the second reviewer and replaced the sequential categorized colors with gradient colors for better visualization of spatial distributions of forecast skill.

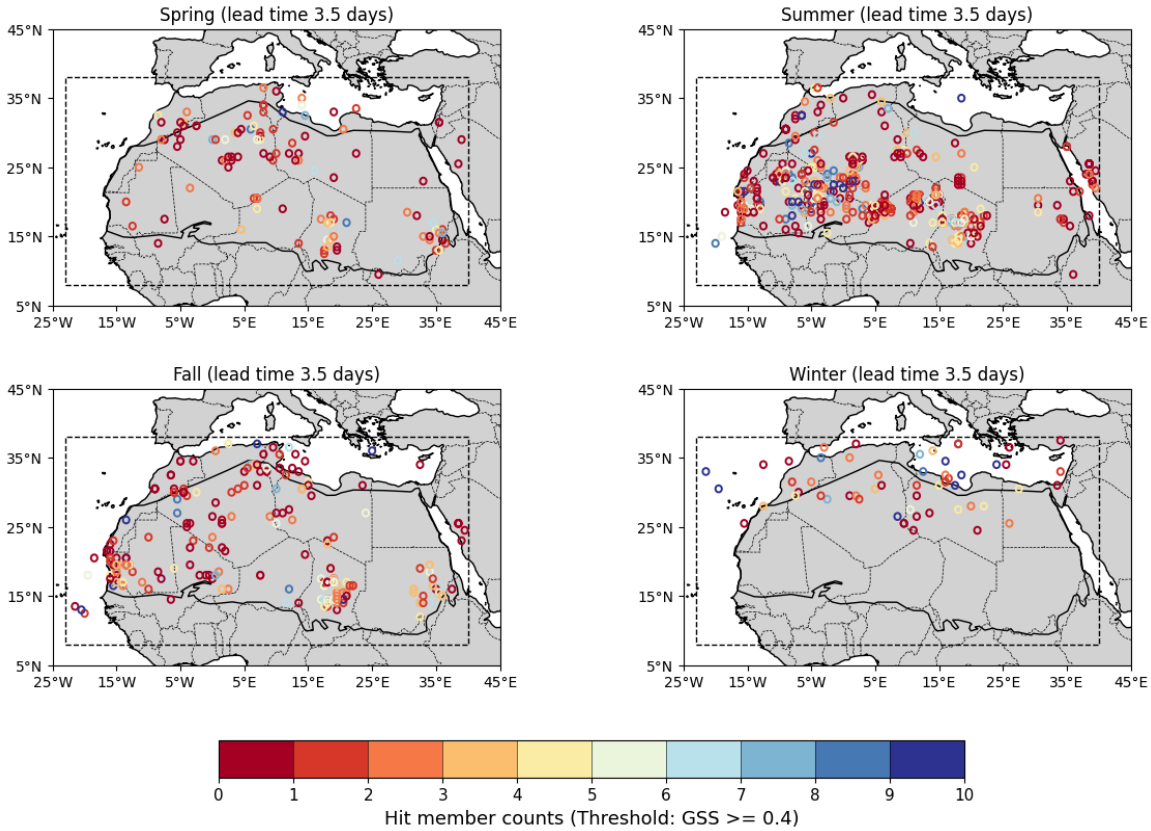

Spring
0.5 d Avg: 0.30
3.5 d Avg: 0.18
10.5 d Avg: 0.06
Lead Time 0.5 days
Lead Time 3.5 days
Lead Time 10.5 days
Number of Events
GSS
Summer
0.5 d Avg: 0.33
3.5 d Avg: 0.21
10.5 d Avg: 0.10
Fall
0.5 d Avg: 0.33
3.5 d Avg: 0.19
10.5 d Avg: 0.08
Winter
0.5 d Avg: 0.49
3.5 d Avg: 0.26
10.5 d Avg: 0.02

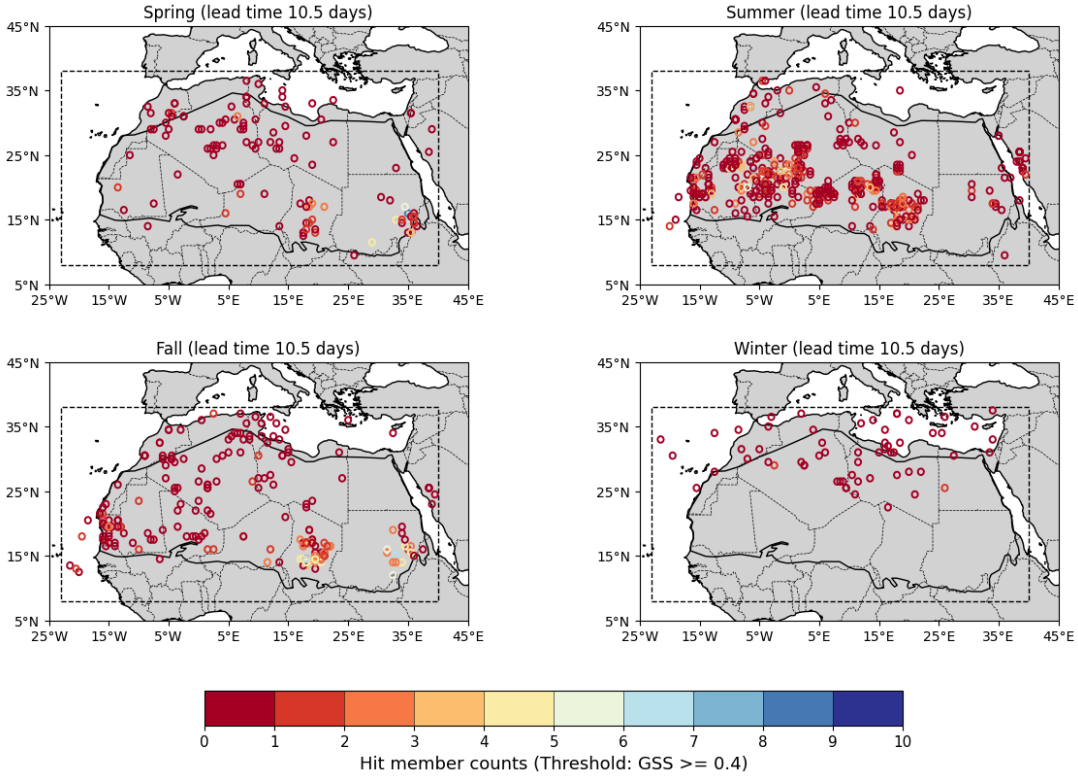

Spring (lead time 3.5 days)
Summer (lead time 3.5 days)
Fall (lead time 3.5 days)
Winter (lead time 3.5 days)
45°N
35°N
25°N
15°N
5°N
25°W
15°W
5°W
5°E
15°E
25°E
35°E
45°E
0
1
2
3
4
5
6
7
8
9
10
Hit member counts (Threshold: GSS >= 0)


Spring (lead time 10.5 days)
Summer (lead time 10.5 days)
Fall (lead time 10.5 days)
Winter (lead time 10.5 days)
45°N
35°N
25°N
15°N
5°N
25°W
15°W
5°W
5°E
15°E
25°E
35°E
45°E
0
1
2
3
4
5
6
7
8
9
10
Hit member counts (Threshold: GSS >= 0)


Spring (lead time 10.5 days)
Summer (lead time 10.5 days)
Fall (lead time 10.5 days)
Winter (lead time 10.5 days)
45°N
35°N
25°N
15°N
5°N
25°W
15°W
5°W
5°E
15°E
25°E
35°E
45°E
0
1
2
3
4
5
6
7
8
9
10
Hit member counts (Threshold: GSS >= 0.2)


Spring (lead time 3.5 days)
Summer (lead time 3.5 days)
Fall (lead time 3.5 days)
Winter (lead time 3.5 days)
45°N
35°N
25°N
15°N
5°N
25°W
15°W
5°W
5°E
15°E
25°E
35°E
45°E
0
1
2
3
4
5
6
7
8
9
10
Hit member counts (Threshold: GSS >= 0.2)


Spring (lead time 3.5 days)
Summer (lead time 3.5 days)
Fall (lead time 3.5 days)
Winter (lead time 3.5 days)
45°N
35°N
25°N
15°N
5°N
25°W
15°W
5°W
5°E
15°E
25°E
35°E
45°E
0
1
2
3
4
5
6
7
8
9
10
Hit member counts (Threshold: GSS >= 0.4)


Spring (lead time 10.5 days)
Summer (lead time 10.5 days)
Fall (lead time 10.5 days)
Winter (lead time 10.5 days)
45°N
35°N
25°N
15°N
5°N
25°W
15°W
5°W
5°E
15°E
25°E
35°E
45°E
0
1
2
3
4
5
6
7
8
9
10
Hit member counts (Threshold: GSS >= 0.4)

- Line 142: why the lead time extends so far to 15.5 days?

The focus of this study is on the short- and medium-range predictability of Saharan HPE-related cyclones. Focusing on a range of 15.5 days allows for assessment of predictability limits of weather systems in the subtropics and over arid regions - a topic that has received less attention in the literature. At longer lead times, better predictability can contribute to early warning and help to reduce devastating societal impacts due to cyclones. Additionally, beyond 15 days, the horizontal resolution of the ECMWF reforecasts is reduced from 16 km to 32 km, making predictability of local factors even more challenging. We have added an explanation to the manuscript (Lines 116-118).

- Figure 2 presents an example of the four category verification methodology. It is very confusing for me and I need more clarification.

To verify the occurrence of cyclones in ECMWF subseasonal reforecasts against reanalysis, we develop a feature-oriented, area-based approach, and implement this method for the Sahara region. We have revised the description of this method, and added a specific example to the manuscript text:

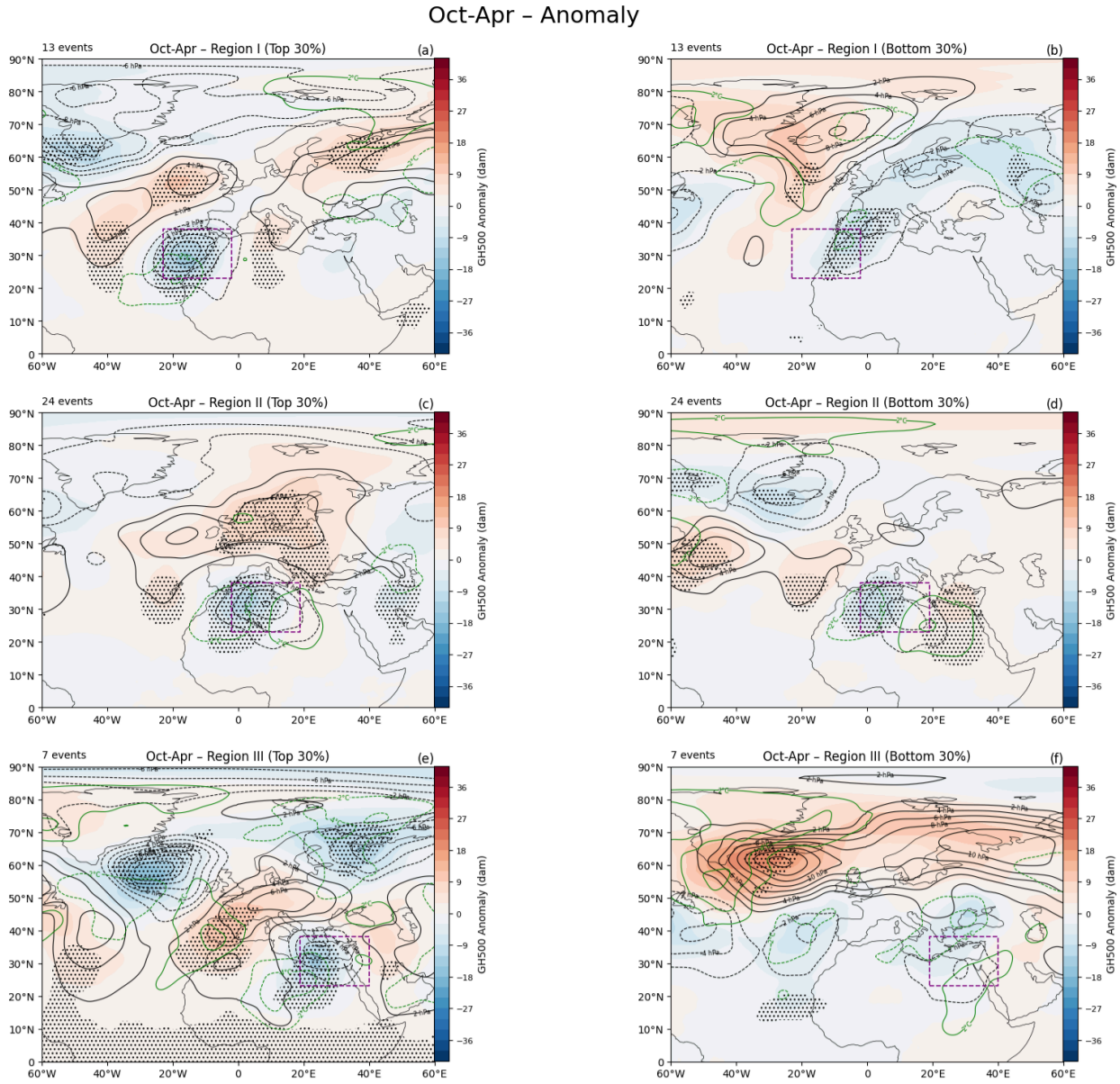
"For the areal storm coverage, we use a four-category contingency table (Table 1 and Fig. 2). This contingency table is constructed by summing the area of the grid points that meet each category in the table within the corresponding ad-hoc study region for each ensemble member and forecast lead time. Such a contingency table is explained for the following example: an HPE-associated cyclone was observed on 21 Nov 2014 over the northwestern Sahara and the nearby area in the Atlantic Ocean (red line in Fig. 2). An ad-hoc study region is constructed using a 6° buffer around its location (purple line). Ensemble member 10, at a 3.5-day lead time, predicted the presence of a cyclone to the southwest of the observed cyclone (blue line). The contingency table for this case is composed of hits (pixels where both the forecasted and observed cyclone are present; blue area), misses (observed only; yellow area), false alarms (forecasted only; purple area), and correct negatives (cyclone is not observed nor forecasted; orange area)."

- Line 155: "The upper 40% and lower 40% of these events were classified as high- and low-skill..". Again, how this percentage derived?

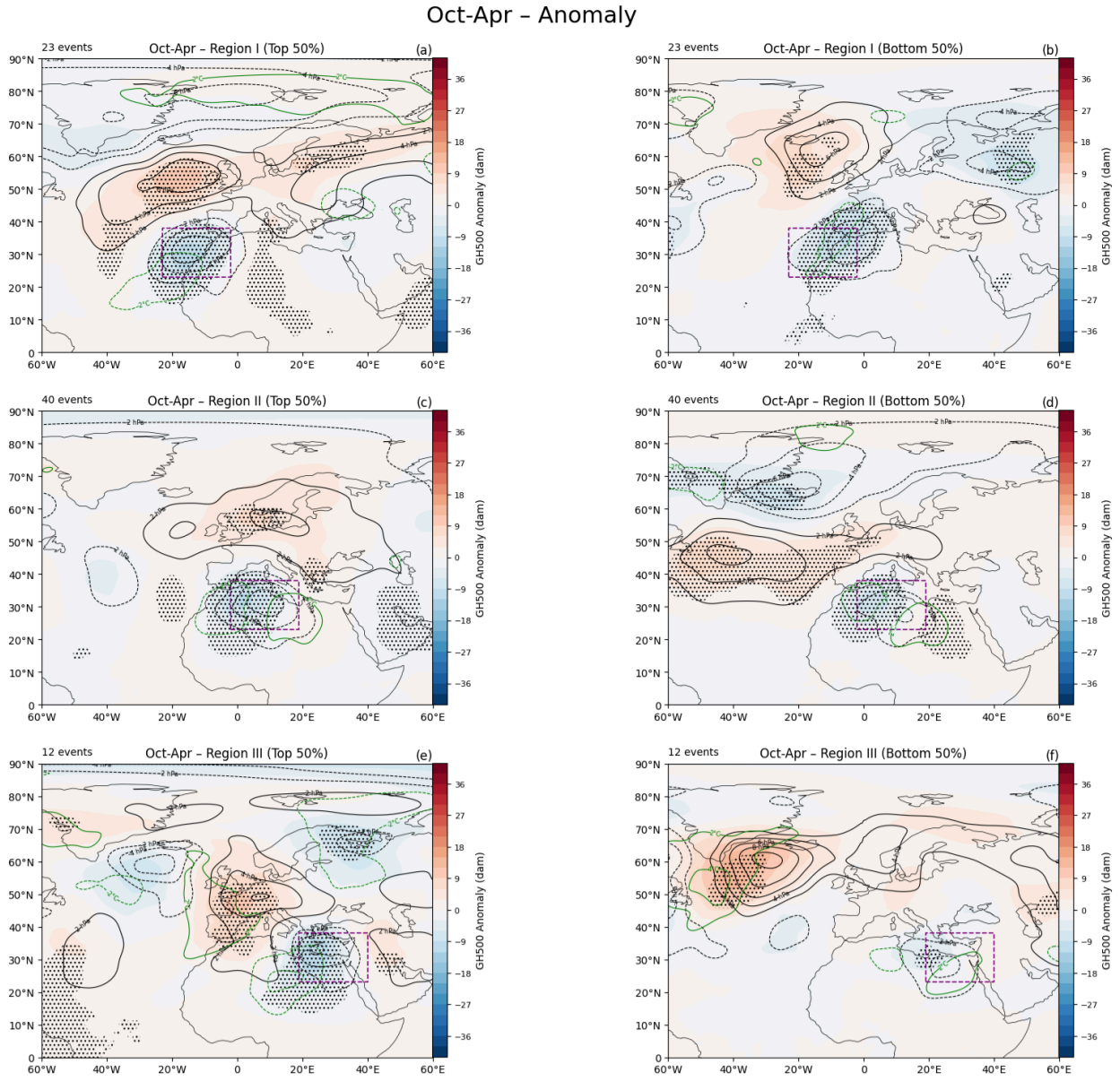
The choice of the 40% threshold represents a compromise between retaining a sufficient sample size for composite analysis and ensuring a clear separation between high- and low-skill events. In some subregions (e.g., the northeastern Sahara), the total number of cases during Oct-Apr with reforecasts at the 5.5-day (10.5-day) lead time is relatively small (~20–22 events), and therefore selecting too small a fraction would lead to insufficient sampling for the composite analysis. At the same time, including a larger fraction (e.g., 50%) would incorporate events with intermediate forecast skill, which tend to reduce the contrast between high- and low-skill conditions.

To assess the sensitivity of the threshold choice, we repeated the analysis using two alternative thresholds (30% and 50%). The resulting large-scale patterns are qualitatively similar to those obtained with the 40% threshold (see figures below).

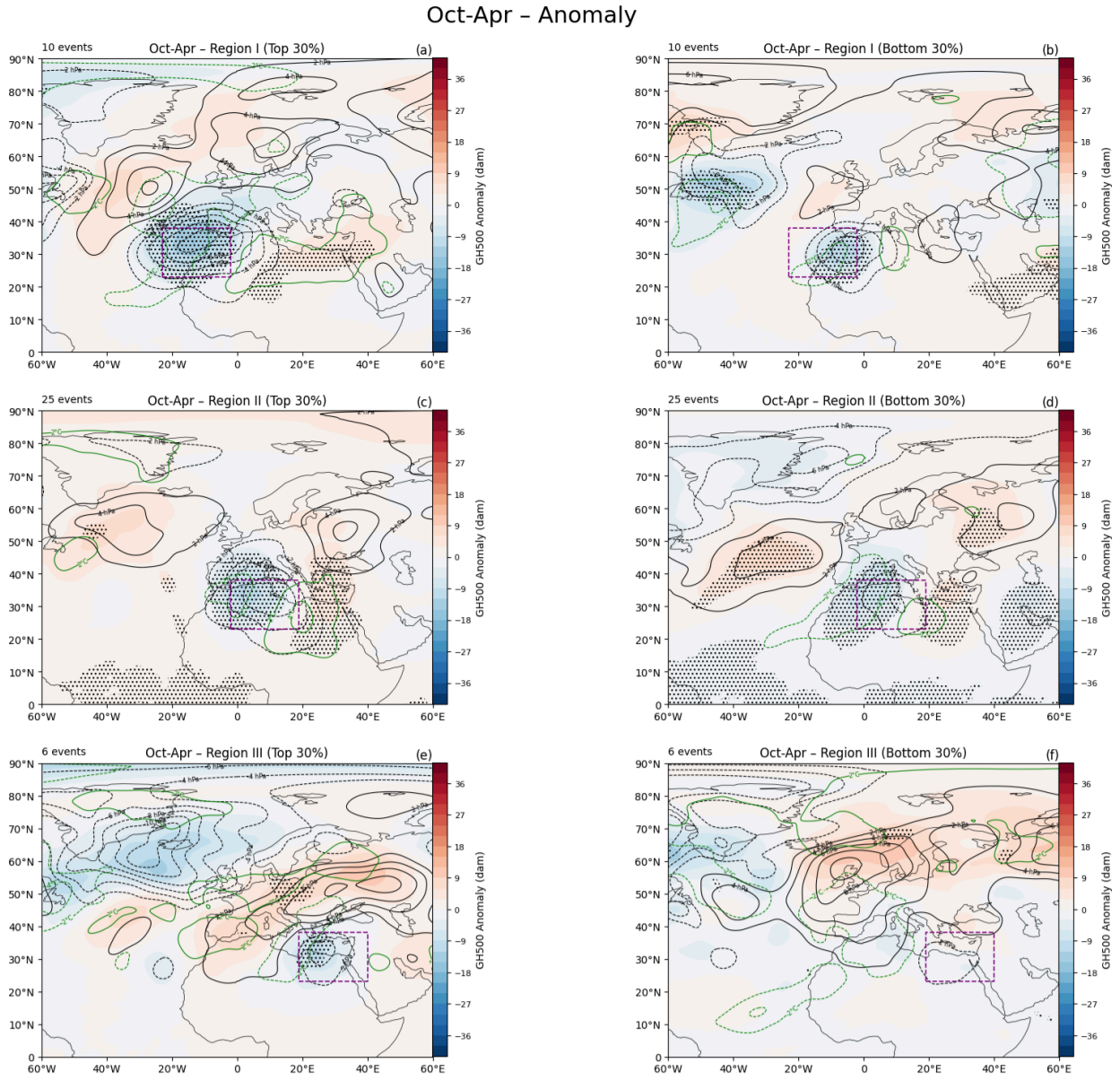
We have clarified this point in the manuscript (“The 40% threshold represents a compromise between maintaining sufficient sample size for robust composites and ensuring a clear separation between high- and low-skill events; sensitivity tests (30% and 50%) indicate that the resulting large-scale patterns are not strongly dependent on this choice.”).



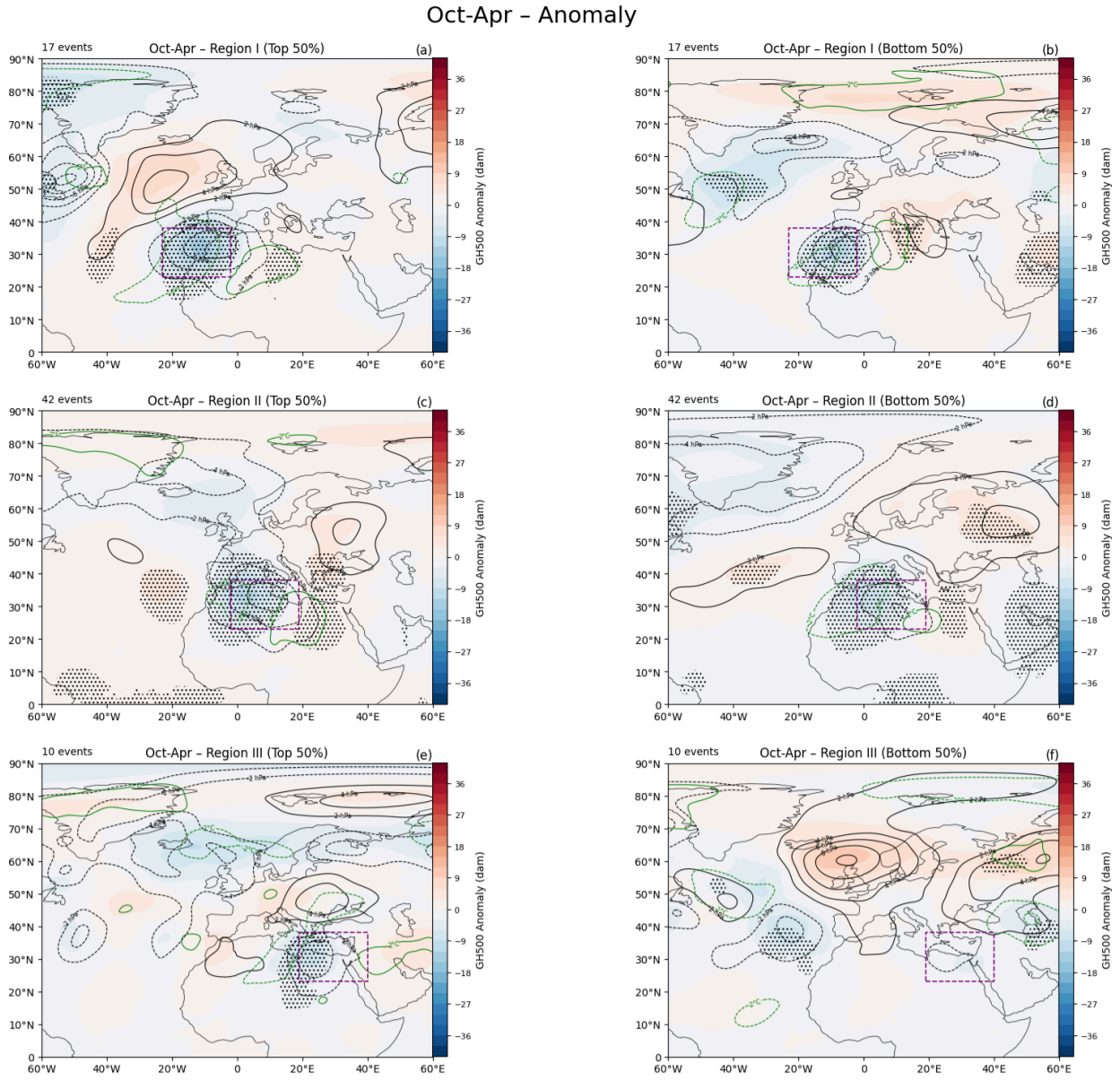
Similar to Fig.7, but with lead time = 5.5 days, percentage = 30%.



Similar to Fig.7, but with lead time = 5.5 days, percentage = 50%.



Similar to Fig.5, but with lead time = 10.5 days, percentage = 30%.



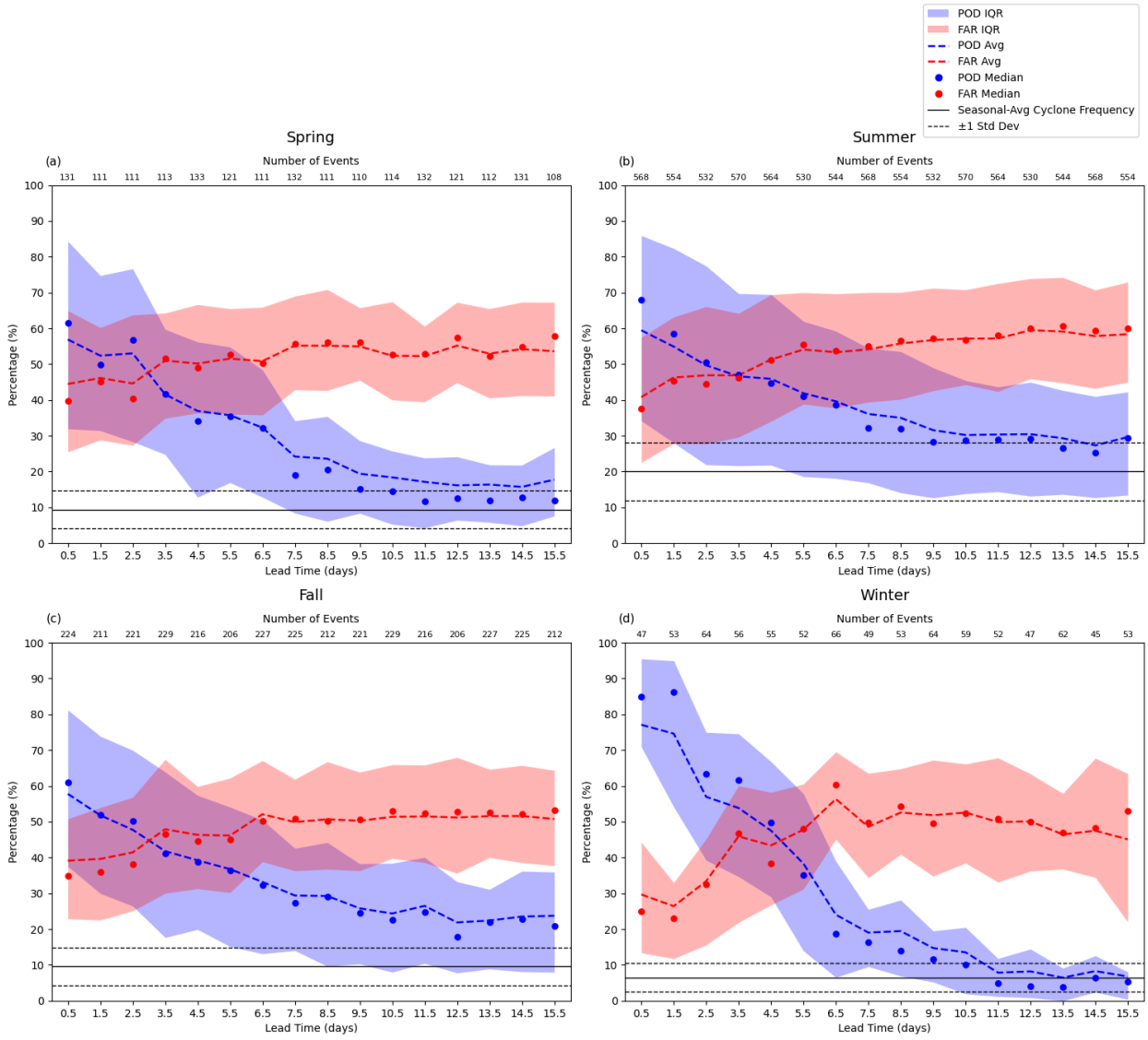
Similar to Fig.5, but with lead time = 10.5 days, percentage = 50%.

- Line 157: why specifically at 12 UTC?

The choice of 12 UTC is consistent with our simplified attribution framework, in which each HPE is represented by a single reference time corresponding to the date of maximum precipitation. Using a fixed time allows a consistent comparison between observed cyclones and reforecasts across all events and lead times. While evaluating cyclones at multiple time steps during their evolution would be more comprehensive, this would substantially increase the computational cost given the large number of cases (>3000 HPE-associated cyclones). The use of a single representative time therefore provides a practical compromise, while still capturing the large-scale conditions associated with peak precipitation.

- Figure 3: These four plots are very confusing and hard to follow. The authors might find another manner to present their findings. In the legend it is stated “The black solid line shows the average climatological frequency of cyclone coverage, computed as the weighted cyclone frequency at each grid point of each cyclone area.” What cyclones are considered? How they are identified?

We revised Fig. 3 and used shading instead of boxplots to show interquartile ranges of POD and FAR for better clarity of the results. To compute the climatological background with which we compare the POD results we search for the mean cyclone frequency over all ad-hoc study regions (i.e., a 6-degree extended buffer region around each of the observed cyclones). Then, for each of these regions we compute the mean grid-box-area-weighted cyclone frequency. Finally, to obtain the climatological frequency, we average the frequency based on the ad-hoc regions per season over all ad-hoc regions. We revised the text accordingly (L163-166).



- Section 3.1 is very wordy and tiring in terms of numbers and parameters. I think that the focus should be a physical interpretation and discussion related to the implications of the findings

Following the reviewers' suggestion, we have revised the entire Section 3 to make it more concise and readable. In the revised section, we focus on the physical interpretation of the results, distinguishing three main topics:

(1) Skill evolution and error growth with lead time, across seasons

(2) Regional variability of cyclone forecast skill

(3) Prevailing large-scale circulation patterns for high- and low-forecast skill over the northern Sahara during the cold season.

Overall, the analysis of Saharan cyclones associated with heavy precipitation events (HPEs) reveals that predictability is heavily dictated by seasonal dynamics and geographic

location rather than a uniform decay in skill. While winter systems are characterized by higher predictability (Hit Rate) in the short-range (up to 4.5 days), summer systems demonstrate a longer predictability horizon, with error growth remaining below climatological thresholds until day 9.5. However, this extended summer skill is accompanied by a high False Alarm Ratio (~60%), indicating that cyclones are too often predicted in areas where they are not observed.

Finally, the difference between FAR and RMSE at extended lead times suggests that while the forecast's false alarm rate remains steady at medium-to-extended lead time, cyclones' MSLP magnitude and structure characteristics become increasingly difficult to resolve beyond the 0 to 5.5-day window.

Therefore, to focus on the physical interpretation and implications of the results, we have thoroughly revised this section as suggested.

Reviewer 2

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.

We thank the reviewer for the constructive comments.

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.
   We have rephrased this sentence in the Introduction (lines 33-37) to clarify this point. The forecast biases in cyclone frequency over the Sahara show an underestimation of cyclone frequency, on the order of magnitude of the cyclone climatology. While cyclones occur about 4.5% of the time over the Sahara in winter, the range of the forecast bias during this time of the year is ~4%. We have clarified this in the text.

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.

We agree with the reviewer that IMERG V07 includes algorithmic improvements compared to V06, particularly in regions with sparse gauge calibration such as deserts. However, our analysis does not directly use IMERG data, but rather the HPE catalog produced by Armon et al. (2024), which is based on IMERG V06 and forms the basis of this study.

To assess the potential impact of using V07, we performed an independent comparison by rerunning the HPE identification algorithm using IMERG V07 over a slightly different Saharan domain and over a longer span (1998–2025). This analysis shows good overall agreement between the two products. In particular, the main climatological features are preserved, including the strong dominance of summer events (JJA fraction: 69% in V06 vs. 64% in V07). The spatial clustering of the largest events in the northwestern Sahara is also consistent, and the top events are recovered with similar locations and volumes (typically within ~2–25%). We therefore expect that using V07 would not qualitatively change the conclusions of this study.

Some systematic differences are, however, observed, including a lower mean event depth in V07 (~−25%) and slightly larger event areas, as well as a higher fraction of winter events (DJF fraction: 3% in V06 vs. 7% in V07), and a decrease in the rate of HPE occurrence (~-30%). While these differences are interesting on their own, and probably reflect the change in algorithms and satellite-data availability, they are outside the scope of our manuscript.

In addition, a new IMERG version (V08) is currently being implemented, highlighting the evolving nature of the dataset. We have therefore retained the established V06-based catalog for consistency, and clarified in the manuscript that the analysis is based on V06 together with a discussion of its potential limitations.

To address this in the manuscript, we have added: “Although newer versions of IMERG (e.g., V07) introduce some changes in event statistics, applying the same detection approach to V07 yields consistent climatological patterns and extreme-event characteristics (not shown).”

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.

The control run was excluded to ensure that the calculated skill statistics were derived only from the perturbed members. A similar approach has been implemented

for simplicity in Afargan-Gerstman et al., 2024. We have added this clarification to the manuscript (section 2.1.3).

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?

   The choice of 12 UTC follows directly from the use of a daily-based HPE catalog, in which each event is represented by a single reference day (the date of maximum precipitation volume). Since sub-daily information on the timing of peak precipitation is not retained in the catalog, a fixed time is required to consistently associate HPEs with cyclones. We therefore selected 12 UTC as a representative synoptic time, ensuring a uniform and reproducible comparison between observed cyclones and reforecasts across all events.

   (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.

   In our analysis, 10016 HPEs out of 12160 HPEs were retained, which indicates around 82.4% of events were retained. Therefore, the results in this study are quite representative for cyclone-related HPEs.

   We thank the reviewer for this comment. Indeed, the 2000 km threshold is not representing a specific physical constraint, but rather a pragmatic constraint introduced by our simplified attribution framework. When using these criteria we lose ~18% of cyclones. However, since in this study, we do not track cyclones continuously in time, but instead examine their proximity to HPEs once per day (12 UTC), a threshold is required to avoid attributing HPEs to remote cyclones that are unlikely to be related.

   The 2000 km threshold reflects a compromise between allowing some flexibility in attributing an HPE to a cyclone (given the single daily snapshot) and preventing systems that are clearly too distant to be physically related. To assess the sensitivity of this choice, we repeated the analysis using alternative thresholds (1000–3000 km, 500 km intervals). While reducing the threshold from 2000 to 1000 km reduces the number of cyclones attributed quite substantially (~18%), increasing it to 3000 km results in only a minor increase (~4%). This indicates that increasing the threshold much beyond 2000 km has little impact on the results, whereas decreasing it would substantially reduce the sample size, which has its own drawbacks. We have

clarified this point in the manuscript: “The 2000 km threshold represents a pragmatic upper bound to avoid HPE attribution to remote cyclones; sensitivity tests (1000–3000 km) indicate that results are only weakly affected by increasing the threshold beyond this value (~+4% at 3000 km), whereas smaller thresholds would substantially reduce the sample size (~−18% at 1000 km).”

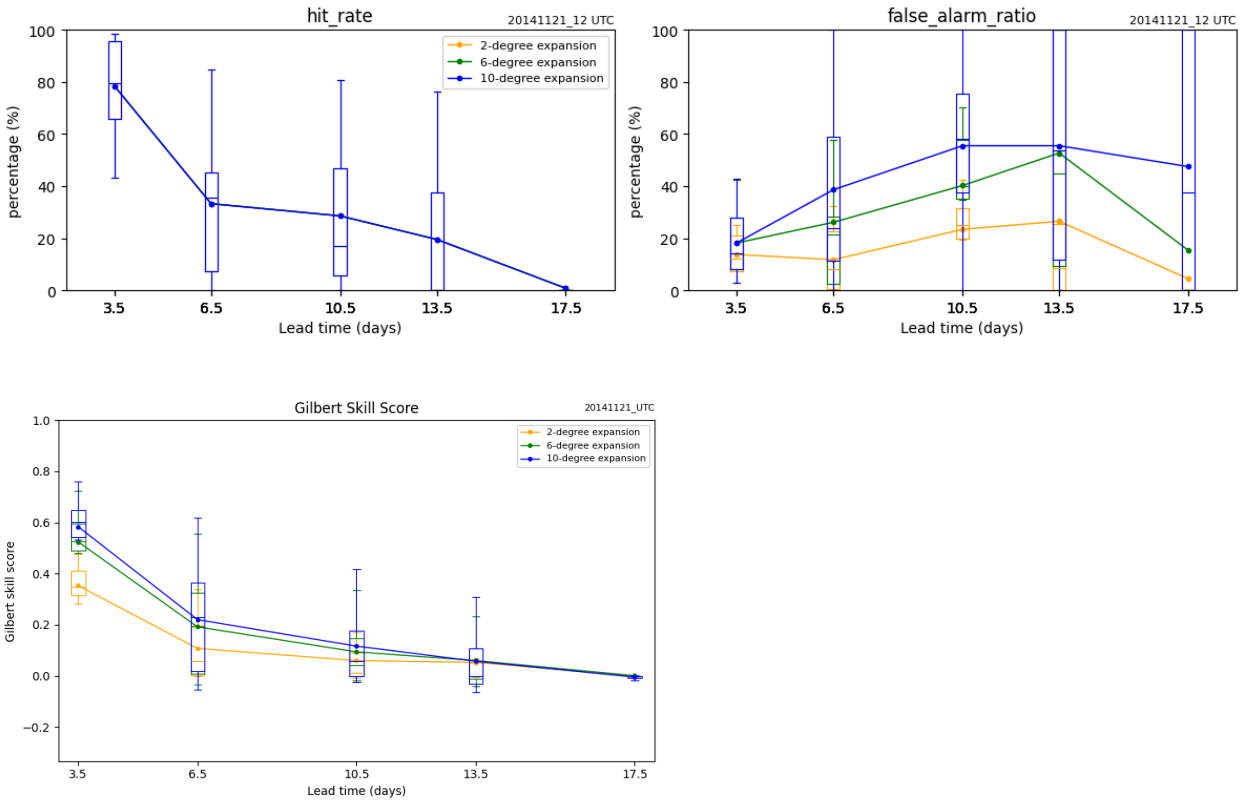
(c) If multiple cyclones are within 2000 km of an HPE mass center (does this occur?), is it always the case that the nearest one is actually dynamically responsible for the HPE?

We acknowledge that proximity alone does not guarantee a causal dynamical link between a cyclone and an HPE. However, in the absence of a universally accepted method for attributing precipitation to specific cyclones, spatial proximity is commonly used as a practical proxy (e.g., Pfahl and Wernli, 2012; Armon et al., 2024). Selecting the nearest cyclone maximizes the likelihood that the identified system is the most relevant to the precipitation event, while avoiding more subjective or computationally intensive tracking approaches. We note that our analysis does not aim to establish detailed dynamical attribution, but rather to provide a consistent and reproducible framework for associating HPEs with nearby cyclones.

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.

   In the figures shown below, we evaluate the sensitivity in hit rate, false alarm rate and GSS for a case-study on the 21 November 2014 (due to the high computational cost we focus on one event). As shown in figs below, hit rate is not sensitive to the size of the expansion area by definition and GSS is only weakly sensitive to it. The false alarm rate is indeed reduced for smaller expansion areas, yet these changes are relatively small compared to FAR variability. Therefore, we choose the 6-degree expansion to study for a large number of verification cases. The reason for defining a 6-degree buffer zone around each cyclone (as part of the verification process) is to reduce the likelihood of over-estimating forecast errors in cases where cyclones are forecasted at the correct lead time and region, but not at the exact location and size. Thus, this methodology aims to improve the accuracy in which the location bias is evaluated, by seeking for the predicted cyclone in a larger area than the observed cyclone location.

   We have added an explanation to the manuscript (L131–135).



Sensitivity of the area-based skill scores (hit rate, false alarm rate and GSS) to the width of the expansion zone.

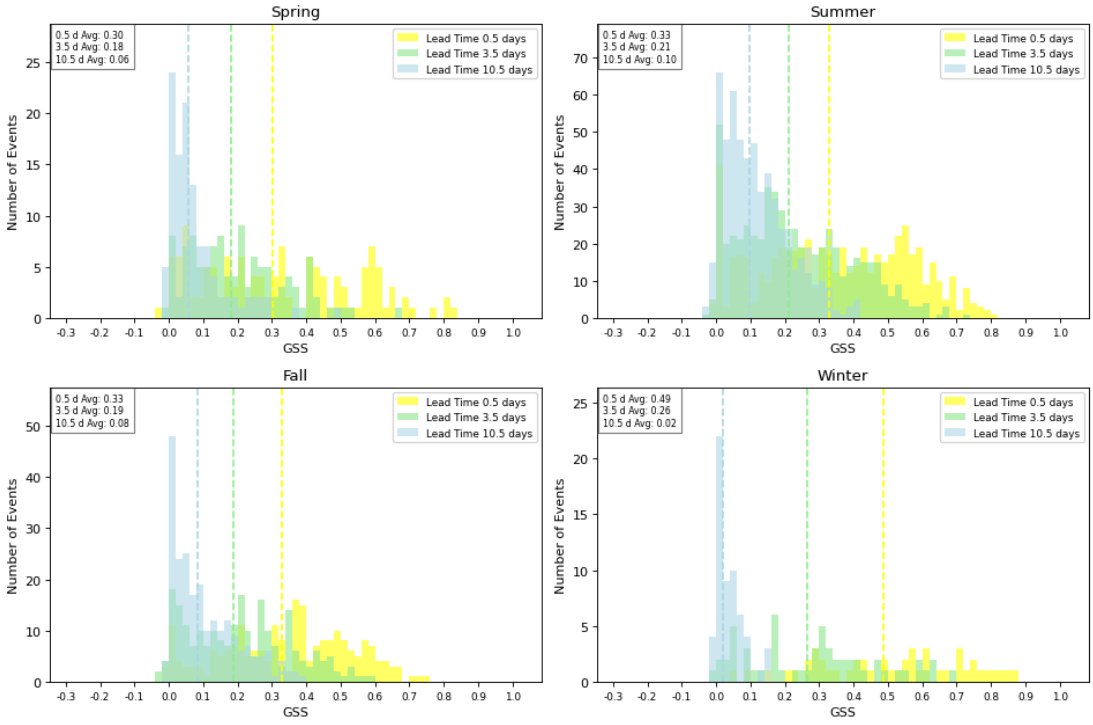
6. Lines 144–147: Could the authors argue why the threshold of 30% for defining "hit members" was chosen? How sensitive are the spatial distributions in Figures 5 and 6 to this threshold? This is particularly relevant for the interpretation of high- versus low-skill regions.

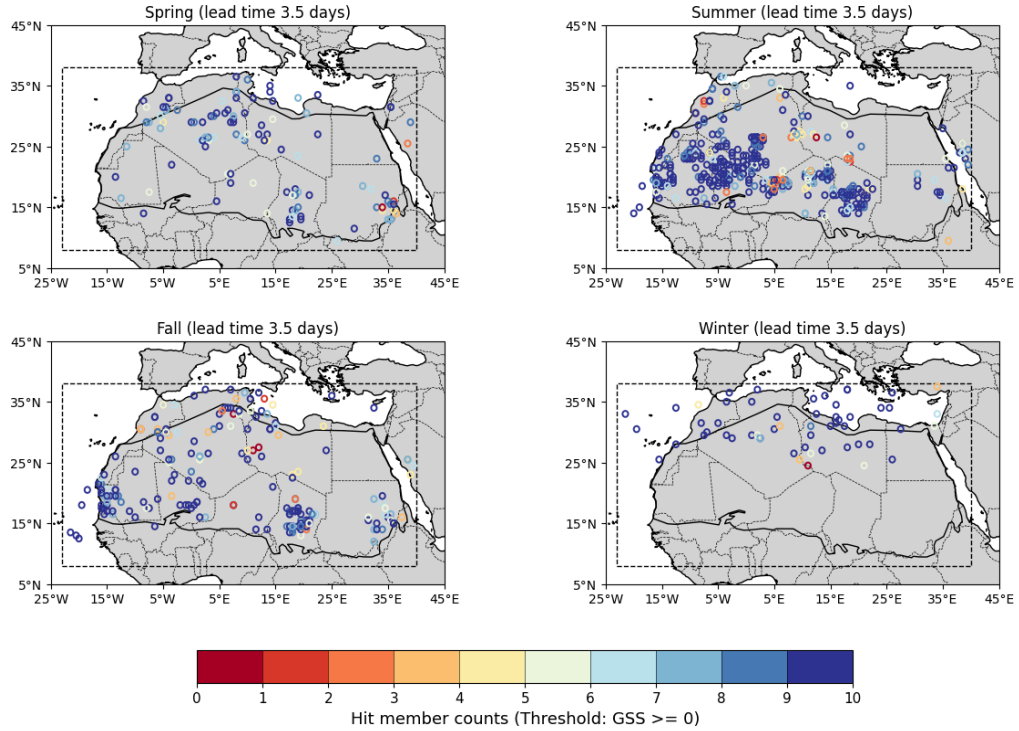
We agree with the reviewer that the 30% threshold is not uniquely defined and therefore should be justified. Our purpose in using this threshold was not to distinguish whether the forecast is merely better than climatology (as in using a 0% GSS threshold), but rather to highlight regions and seasons with relatively higher versus lower predictability. While a threshold of 0% GSS has a clear interpretation, since it marks where the forecast outperforms a random forecast, in practice it produces a largely binary picture with many events classified as equally skillful and therefore provides less contrast for identifying meaningful spatial and seasonal differences in forecast quality.

Therefore, we retained the 30% threshold, which better emphasizes where the model outperforms climatology by a substantial margin and helps distinguish regions and seasons with better or worse predictability. To assess the sensitivity of this choice, we repeated the analysis using alternative thresholds of 0%, 20%, and 40% (see figs. below). The resulting spatial patterns are qualitatively similar, indicating that the main conclusions do not depend strongly on

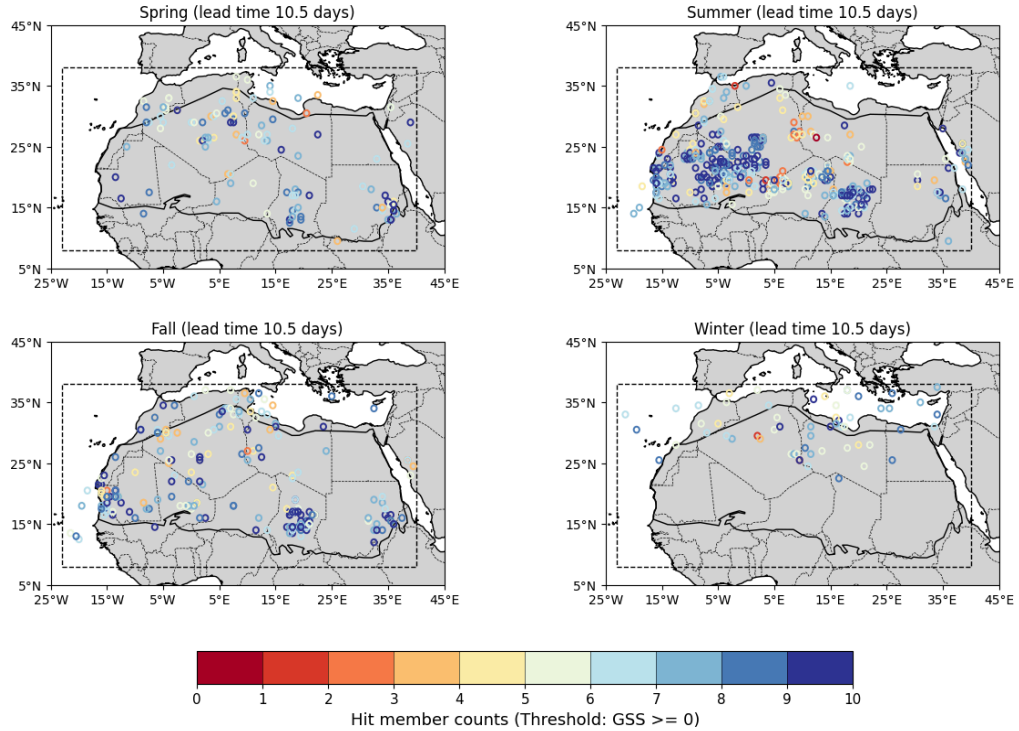
the exact threshold value, while the 30% threshold provides, in our opinion, the clearest visualization of the spatial structure of forecast skill. Furthermore, to emphasize the binary nature of the 0% GSS threshold, we show here the distribution of GSS value for each season, for three different lead times, which shows very few forecasts dropping to <0%. We have clarified this point in the manuscript (L168–171, and Sect. 3.2).

In addition, we referred to a suggestion from the second reviewer and replaced the sequential categorized colors with gradient colors for better visualization of spatial distributions of forecast skill.

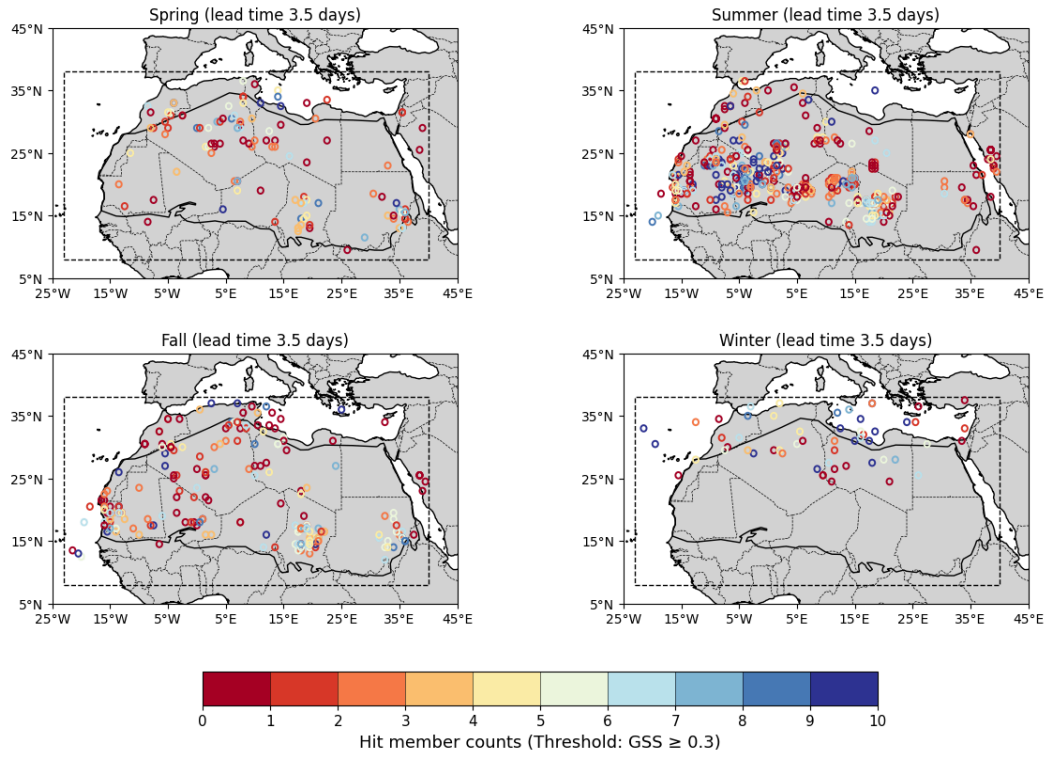




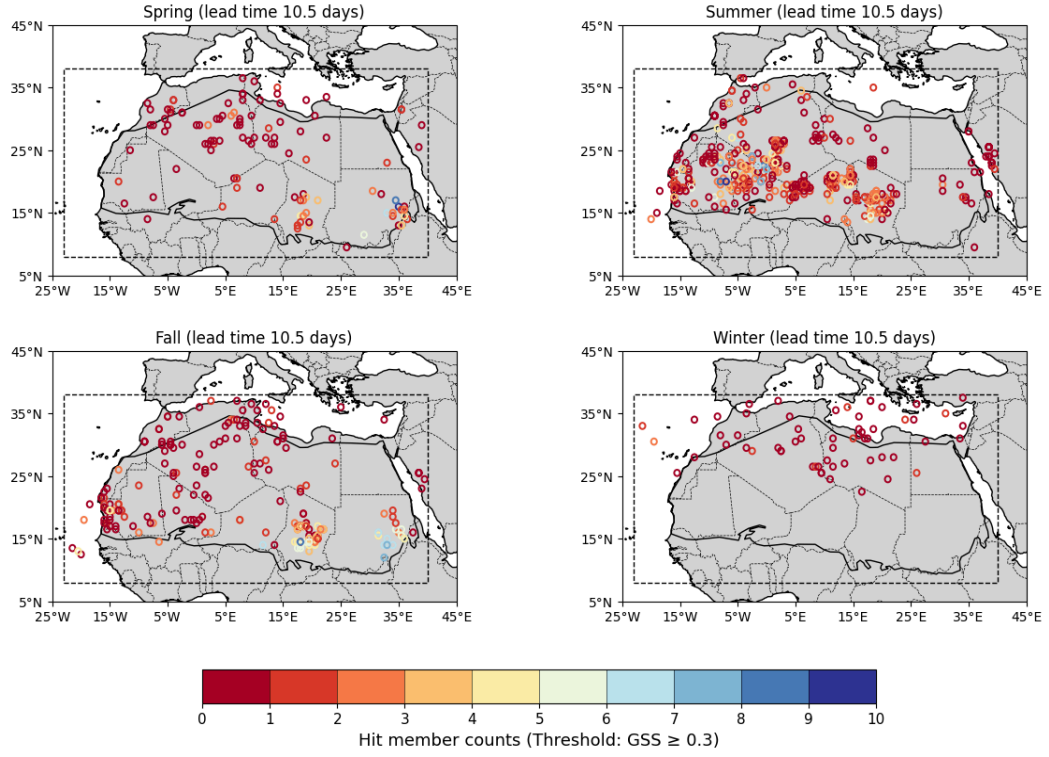
Similar to Fig. 5 in the manuscript but for a threshold of 0



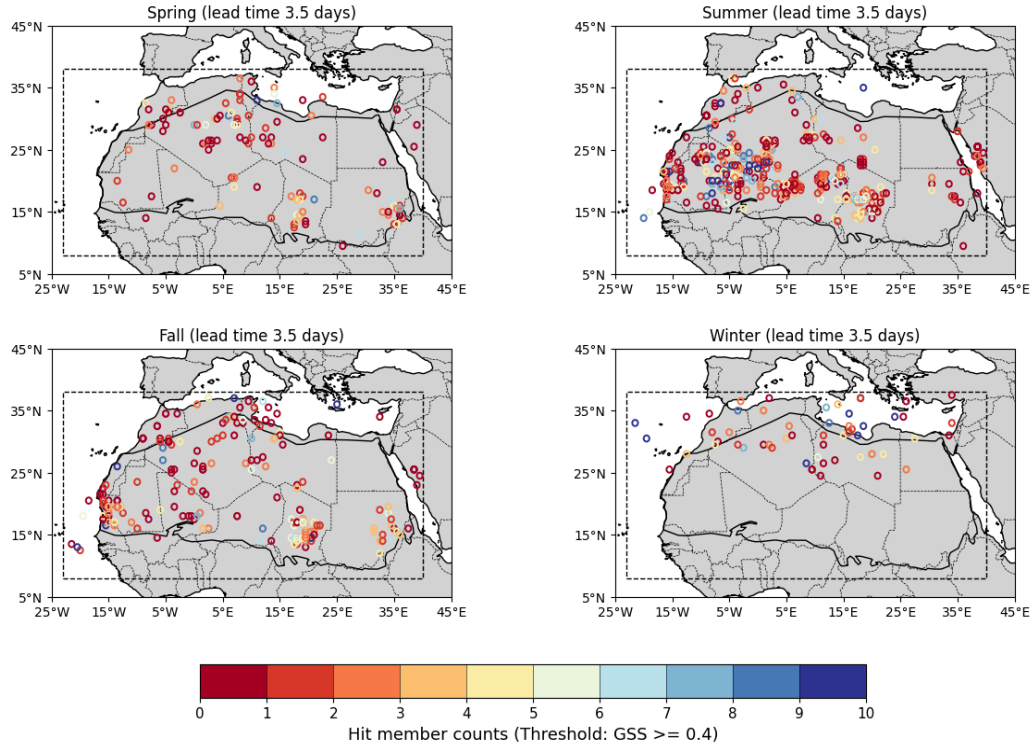
Similar to Fig. 6 in the manuscript but for a threshold of 0



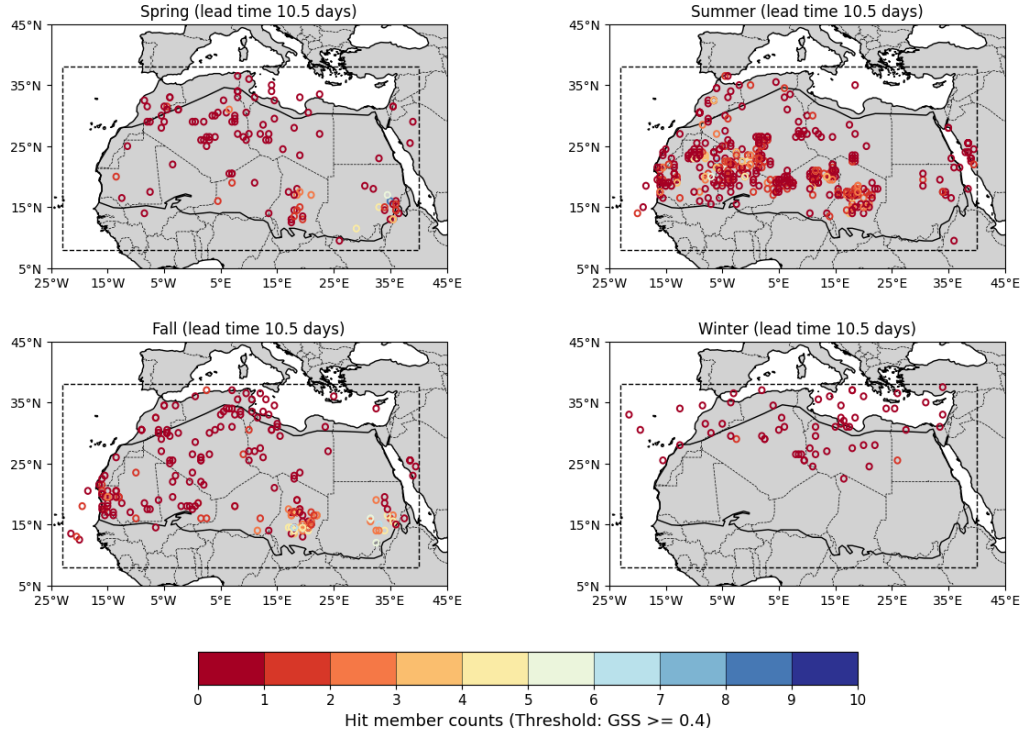
Similar to Fig. 5 in the manuscript but for a threshold of 0.3



Similar to Fig. 6 in the manuscript but for a threshold of 0.3



Similar to Fig. 5 in the manuscript but for a threshold of 0.4



Similar to Fig. 6 in the manuscript but for a threshold of 0.4

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.

Following the reviewers' suggestion, we have revised the entire Section 3 to make it more concise and readable. In the revised section, we focus on the physical interpretation of the results, distinguishing three main topics:
(1) Skill evolution and error growth with lead time, across seasons
(2) Regional variability of cyclone forecast skill
(3) Prevailing large-scale circulation patterns for high- and low-forecast skill over the northern Sahara during the cold season.
Overall, the analysis of Saharan cyclones associated with heavy precipitation events (HPEs) reveals that predictability is heavily dictated by seasonal dynamics and geographic location rather than a uniform decay in skill. While winter systems are characterized by higher predictability (Hit Rate) in the short-range (up to 4.5 days), summer systems demonstrate a longer predictability horizon, with error growth remaining below climatological thresholds until day 9.5. However, this extended summer skill is accompanied by a high False Alarm Ratio (~60%), indicating that cyclones are too often predicted in areas where they are not observed.
Finally, the difference between FAR and RMSE at extended lead times suggests that while the forecast's false alarm rate remains steady at medium-to-extended lead time, cyclones' MSLP magnitude and structure characteristics become increasingly difficult to resolve beyond the 0 to 5.5-day window.
Therefore, to focus on the physical interpretation and implications of the results, we have thoroughly revised this section as suggested.

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.

To compute the climatological background with which we compare the POD results we search for the mean cyclone frequency over all ad-hoc study regions (i.e., a 6-degree extended buffer region around each of the observed cyclones). Then, for each of these regions we compute the mean grid-box-area-weighted cyclone frequency. Finally, to obtain the climatological frequency, we average the frequency based on the ad-hoc regions per

season over all ad-hoc regions. We revised the text accordingly and added a clearer and detailed explanation (L163-166).

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?

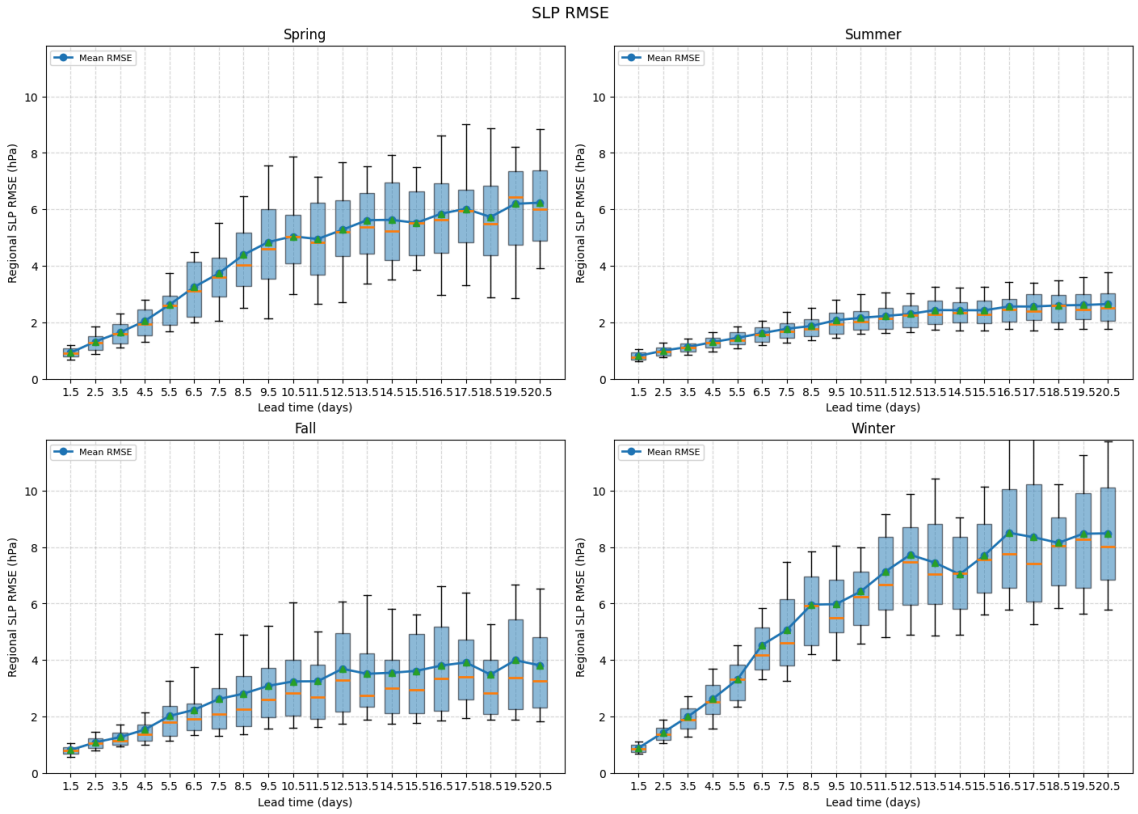
   We have revised section 3.1 to clarify our findings, and corrected this sentence. The interquartile range shown in each panel provides a way for quantitative comparison between the different seasons, as it shows the range of skill variability that characterizes each season. Additionally, due to the large regional variability between cyclones within each season, the comparison between seasons is focused on the regional dependence of cyclone skill.

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?

   Thanks for pointing this out. We have revised the reference line for MSLP RMSE in this figure. RMSE values are compared against the standard deviation of MSLP averaged over all ad-hoc study regions for all **cases** within every season. We have added a clarification regarding the reference definition to the Methods sect. and revised Fig. 4 caption.

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

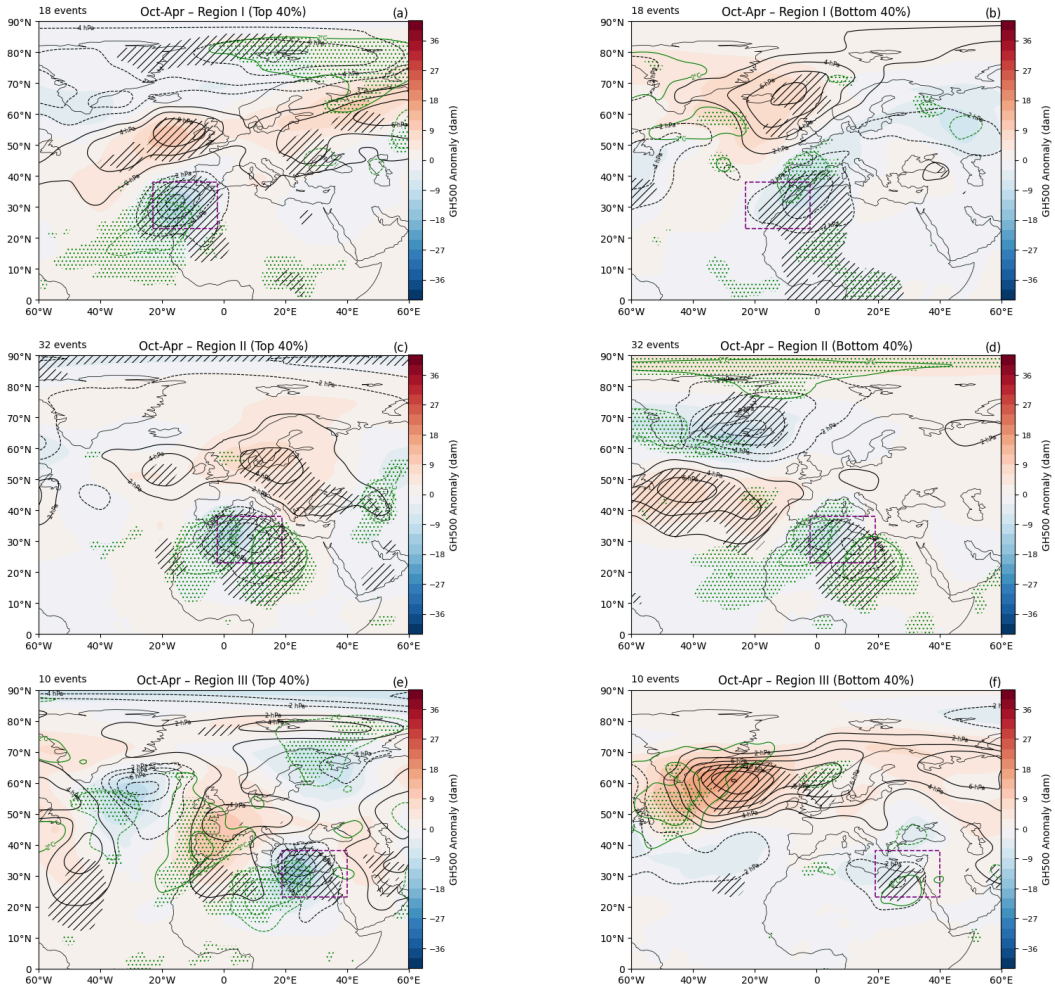
   We have examined the RMSE on longer lead times (see figure below). It appears that the error growth is slower on the longer lead times, although it is indeed difficult to determine if this is indeed due to stabilization. We rephrased this sentence as suggested (Lines 208–209).



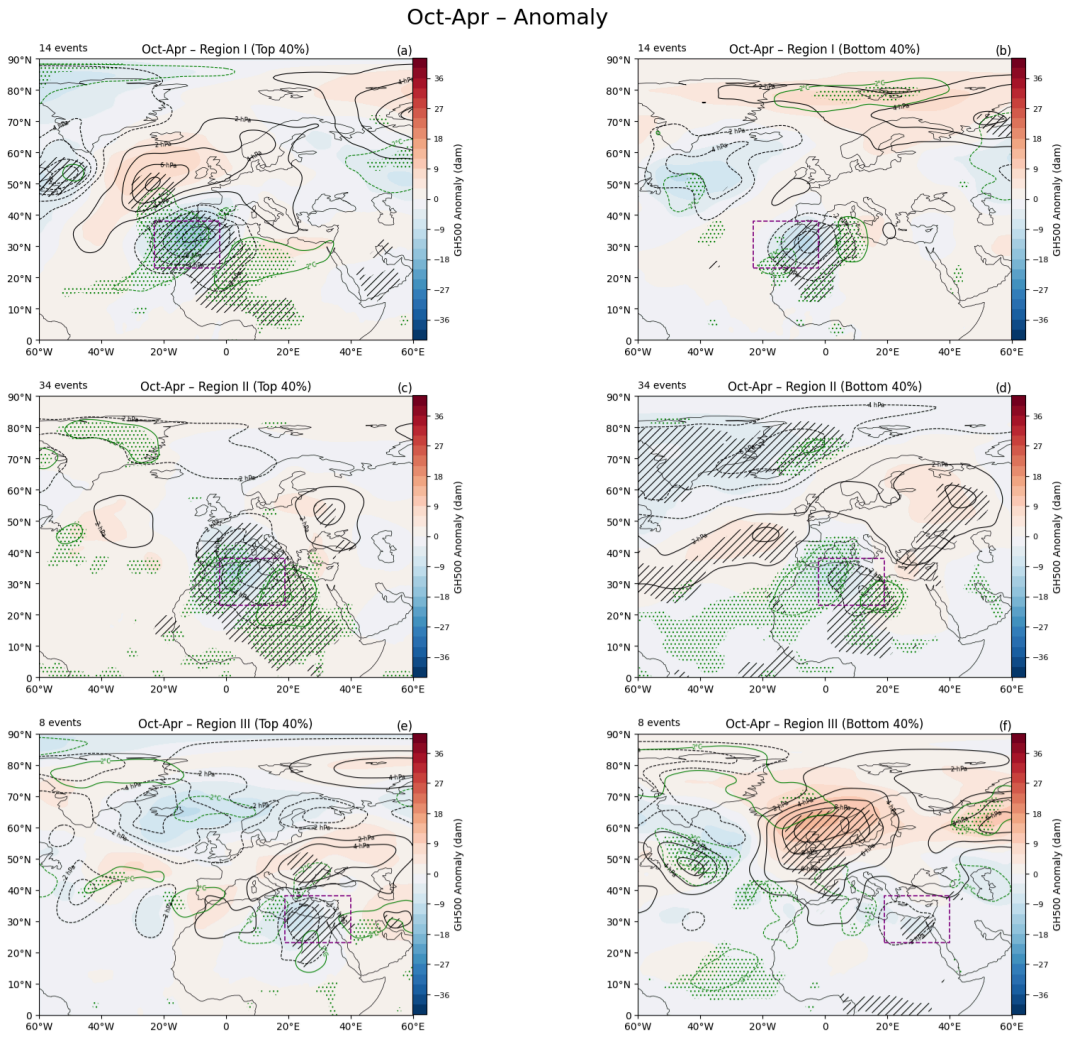
MSLP RMSE for lead times 1.5 to 20.5 days from initialization.

10. 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.

    Visualizing the significance of three different variables in one panel is difficult and might reduce the interpretability of the results. Therefore, we focus in the manuscript on GH500 significance. Below, we included two plots with significance for MSLP and T850, separately for lead times of 5.5 and 10.5 days. GH500 anomalies are shown in black hatching and green stippling shows T850 anomalies. Overall, there is a general agreement between significant areas of MSLP and GH500.



Similar to Fig. 7 in the manuscript, but with green stippling for T850 anomalies.



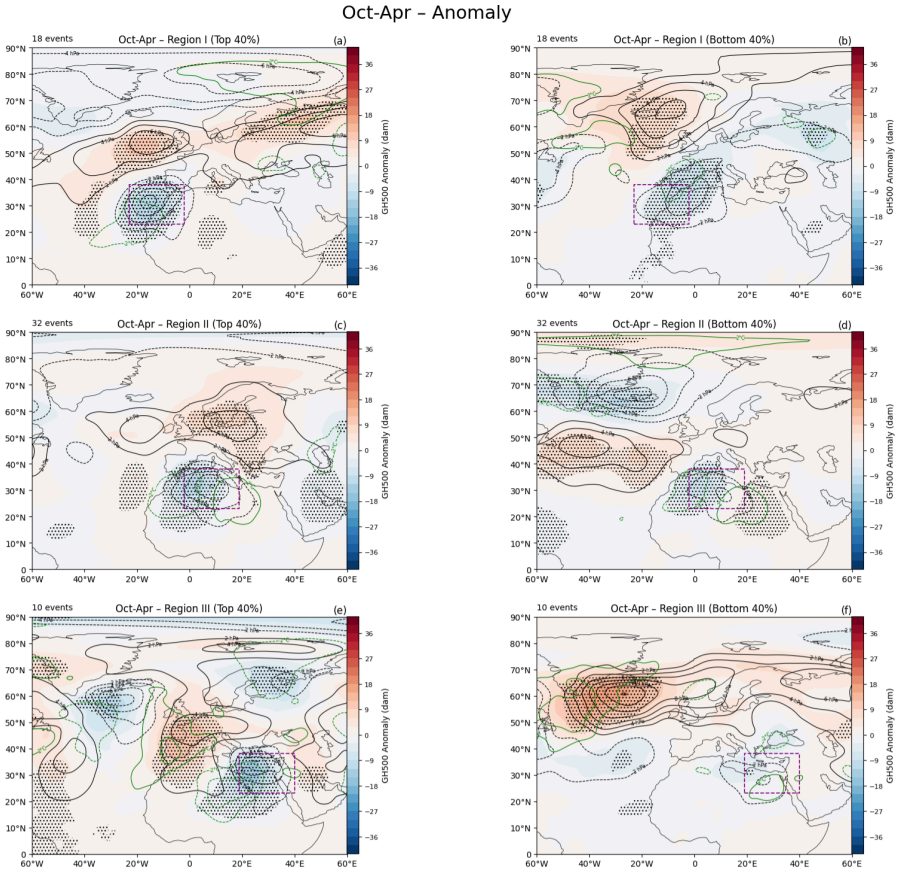
Similar to Fig. 8 in the manuscript, but with green stippling for T850 anomalies.

11. 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.

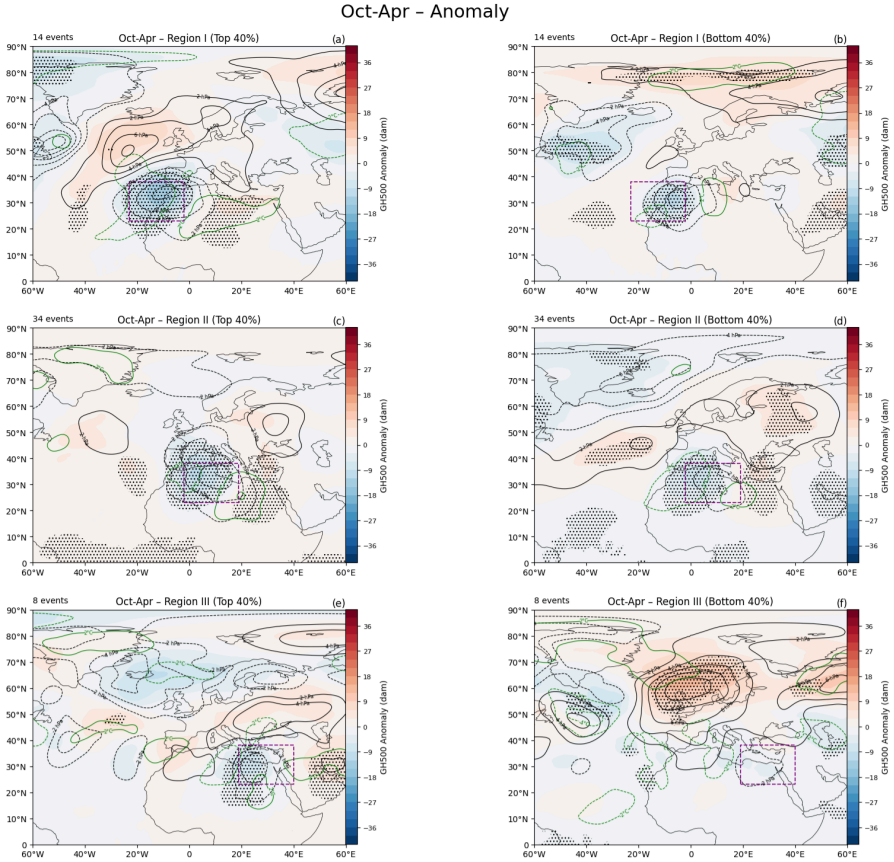
We thank the reviewer for this valuable comment. As the Student's t-test is sensitive to extreme samples, we replaced the t-test with a bootstrap resampling test (2000 resamples) to reproduce Figures 7 and 8 with a robust analysis regarding the significance of the patterns. Grid points were marked significant when zero lay outside the 95% bootstrap confidence interval of the mean. The significance-test results shown below for the bootstrap resampling method are rather similar to those

for the t-test. The original versions of figures 7 and 8 are replaced with these two figures in the text.

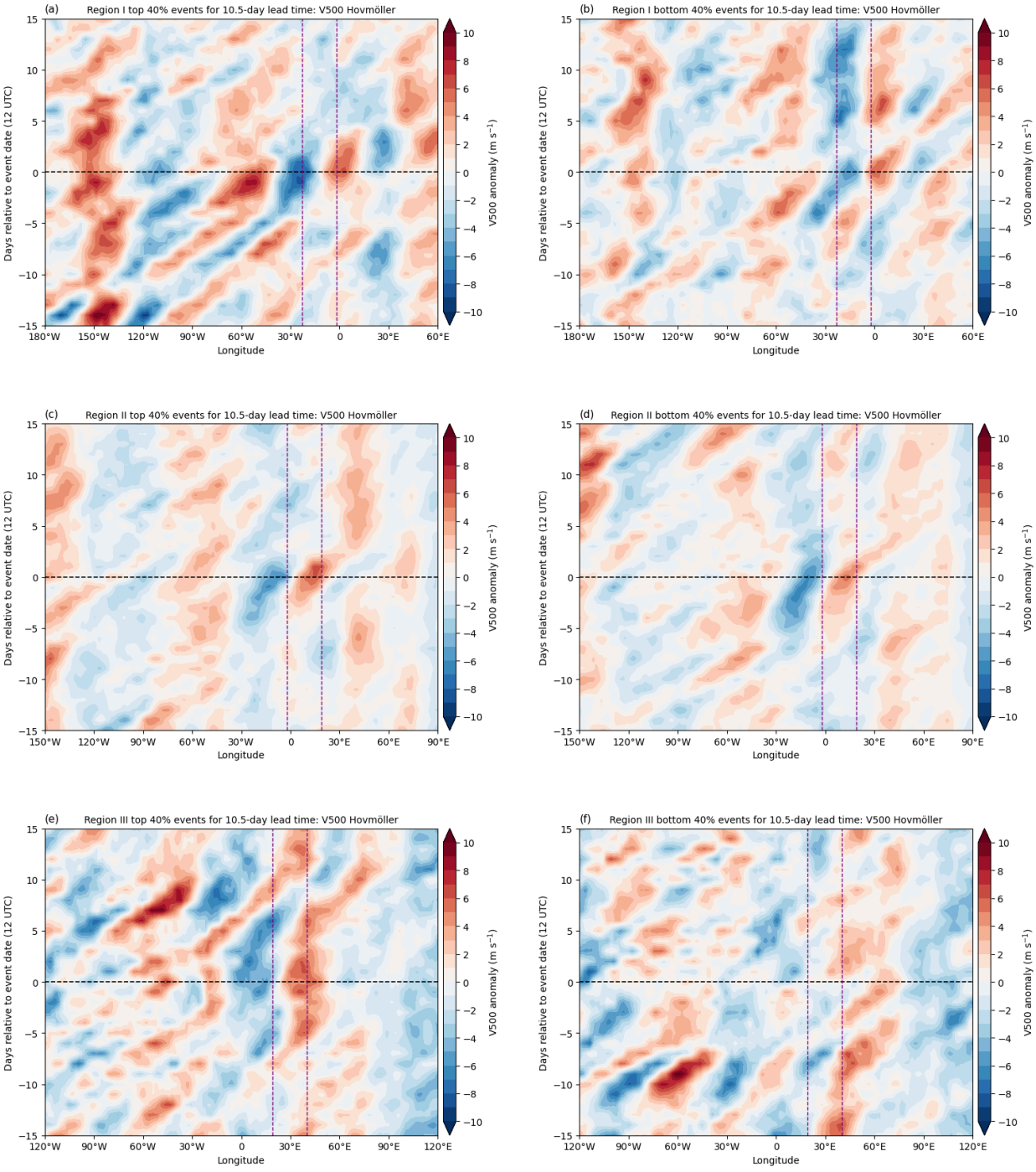
In addition, to examine whether the composite analysis obscures an event-to-event variability, we added new plots below with Hovmöller diagrams for V500 for sub-regions in the northern Sahara. These plots demonstrate the differences in the large-scale atmospheric patterns between the high- and low-skill cases.



Same as Fig. 7 in the manuscript, but using a bootstrap significance test.



Same as Fig. 8 in the manuscript, but using a bootstrap significance test.


(a) Region I top 40% events for 10.5-day lead time: V500 Hovmöller
(b) Region I bottom 40% events for 10.5-day lead time: V500 Hovmöller
(c) Region II top 40% events for 10.5-day lead time: V500 Hovmöller
(d) Region II bottom 40% events for 10.5-day lead time: V500 Hovmöller
(e) Region III top 40% events for 10.5-day lead time: V500 Hovmöller
(f) Region III bottom 40% events for 10.5-day lead time: V500 Hovmöller
Days relative to event date (12 UTC)
Longitude
V500 anomaly (m s^{-1})

Hovmöller diagrams for V500 for subregions I , II, and III at the lead time of 10.5 days.

Discussion and Conclusions

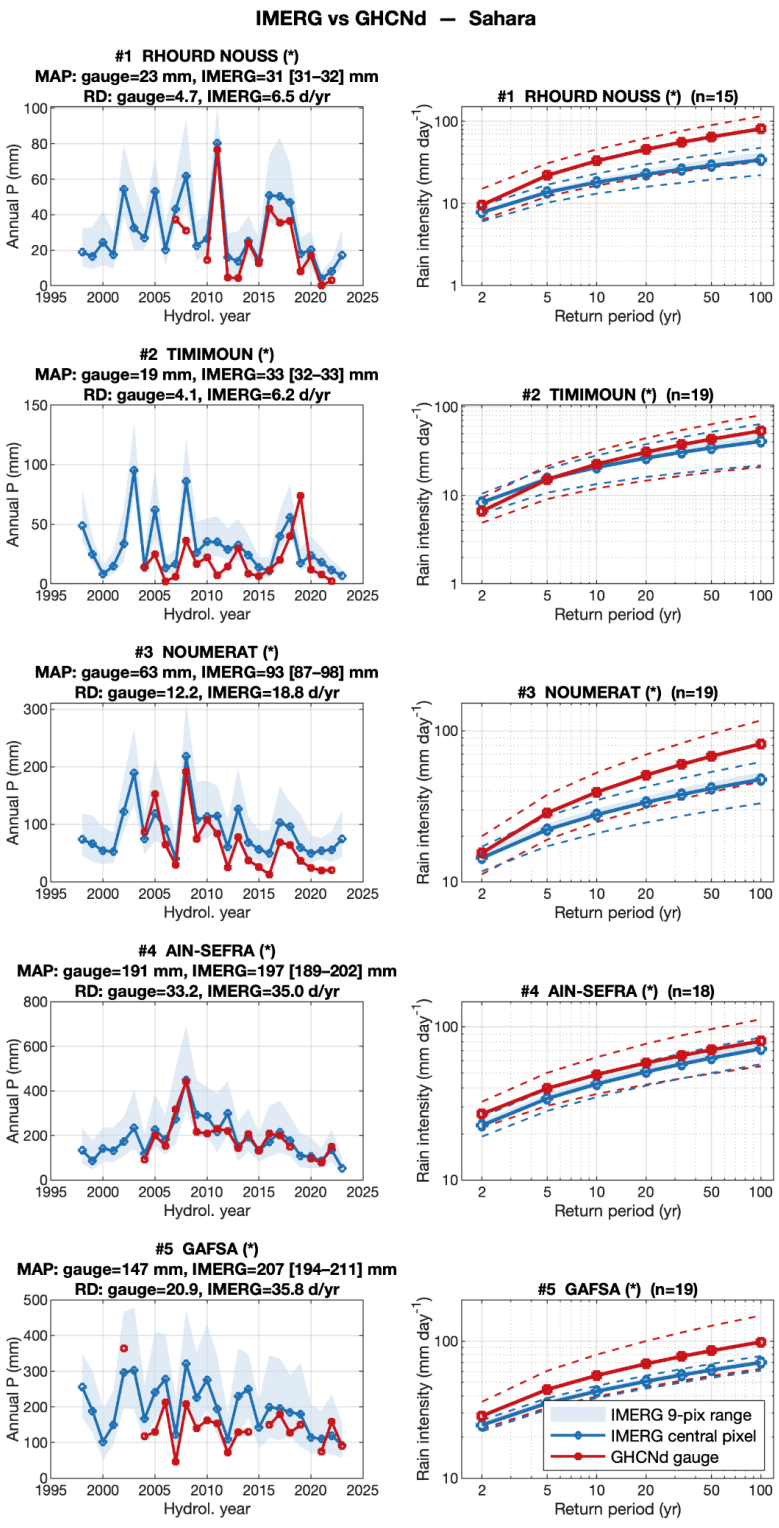
13. I am missing a discussion of the observational uncertainties and their impact on the study results. For example, it is known that precipitation extremes are not well represented in IMERG in data-sparse regions like the Sahara.

    We agree with the reviewer that observational uncertainties in precipitation products are an important consideration, particularly in data-scarse regions such as the Sahara where satellite estimates are less constrained by gauge observations. A detailed evaluation of IMERG uncertainties, however, is beyond the scope of the present study, which focuses on the relationship between cyclones and HPEs based on an existing event catalog.

    We note that such uncertainties were partially addressed in Armon et al. (2024), where the robustness of the HPE catalog was evaluated using independent approaches, including comparisons of return period of events with gauge-based analyses (see their supplementary material: Fig. A4).

    In addition, we have begun evaluating IMERG V07 against available gauge data for a different study, which indicates reasonable agreement, particularly for multi-day accumulations and event-scale precipitation, while larger uncertainties remain at the daily scale (see multi-year comparisons in the figures below). These results provide additional confidence that the main characteristics of HPEs used in this study are reasonably captured.

    We have added a brief discussion of these limitations and their potential implications in the manuscript (Corresponding method subsections, and L333–345)



This figure shows initial results from validating IMERG V07 vs. selected (daily-based) GHCNd gauges across the Sahara. Validation stations are numbered from driest (#1) to wettest (#5). Left panels show annual precipitation totals (mm) for hydrological years (September–August, 1998–2025). Blue circles represent the IMERG V07 central pixel, and the shaded area indicates the range across the surrounding 3×3 IMERG pixel neighborhood. Right panels show return period curves for daily

precipitation over the same stations. Solid lines and markers correspond to the datasets shown in the left panels. Dashed lines indicate 5th–95th percentile bootstrap confidence intervals derived from resampling hydrological years (1000 iterations), with red for gauge uncertainty and blue for IMERG.

14. Lines 280–282: Could the authors elaborate how this improved model accuracy could be achieved? In other words, how could improved understanding translate to **improved model accuracy**?

    Identifying *when* and *where* the forecast model struggles to predict storms is a crucial step towards improving forecast accuracy. Specifically, model developers can use a variety of approaches for achieving such an improvement: examine different parametrizations, tune existing parametrizations, test if increased horizontal or vertical resolution is needed in certain region, improve assimilation of initial conditions or add more sources of observations – all of these steps can help to translate improved understanding of Saharan cyclones to improved forecasting.

    For example, our finding that summer cyclones possess a longer predictability horizon, yet suffer from a ~60% False Alarm Ratio (FAR), suggests a systematic model bias. Improved or tuned parametrization of convection and boundary layer schemes for the arid Sahara can help with such a bias.

    In addition, understanding the limits of weather predictability, and storms in particular, can help with identifying “regimes” of predictability for the Sahara. Specifically, understanding that winter skill is "initial conditions-limited" while summer skill is "physics scheme-limited" allows for seasonally-optimized forecasting strategies.

    Lastly, the strong regional dependency identified in our study points toward localized errors, possibly linked to Saharan surface properties (surface albedo, soil moisture) or complex topography (e.g., the Atlas Mountains or Ahaggar Mountains) can help model developers and decision makers when planning their next model development. We have added this discussion to the Discussion section (Lines 346–354).

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?

Indeed, predictability of southern Saharan cyclone, especially in summer, shows higher skill at extended lead times. Thermally-driven cyclones (e.g., heat lows) and monsoon-driven troughs, as well as African easterly waves, play a dominant role in driving HPEs in this region. Such factors can act as potential sources of predictability for cyclones at medium-to-extended lead times. However, due to their complex interactions, involving both tropical and extratropical influences, linking southern Sahara HPE-associated cyclones to large-scale dominant drivers is not discussed as part of this study. We have added a short discussion on this (L325–329).

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.

    We integrated this part into the second paragraph of Sect. 4 (Lines 289–296)

17. Lines 298–300: “Moreover, since … complicated methods”. I believe this sentence is not needed.

    We removed this sentence.

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.

We thank the reviewer for this advice. We uploaded a code for a verification method and for producing the corresponding figures in Zenodo for reference (DOI 10.5281/zenodo.19557134).

FORMALITIES

Text

• Line 48: "normally dry on their poleward side cantrigger" appears to be a typo. "cantrigger" should read "can trigger".

Thanks for pointing the typo out. We corrected it.

• Lines 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.

It should be 2014. Thanks for pointing the typo out. We revised it.

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).

We have revised the captions of Figures 3 and 4 and added more explanation regarding the box plots.

• Figure 4: The font size of the tick labels on the x and y axes are not the same.

We unified the font sizes as suggested.

• 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.

We adopted the suggestion and used a diverging colormap to visualize the spatial distributions of the forecast skill. The revised figures are shown as Figs. 5,6 in the manuscript.

• 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.

We followed the suggestion and added descriptions for MSLP and T850 contours. We also changed the colors of T850 contours to green for better visibility.

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

Thank you! They were indeed very helpful, and we believe the manuscript is much improved in the revised version.