Article review

"Predictability of high-impact weather associated with Mediterranean cyclones in ECMWF ensemble forecasts.
Part 1: Method and case studies"

Summary: In their article, [Katharina Hartmuth et al.] present a novel method to assess the forecast performance of the ECMWF ensemble in predicting extreme weather events associated with Mediterranean cyclones. The first part of their study explains the methodology and illustrates it with three case studies of impactful Mediterranean cyclones. The forecast performance is evaluated based on the ability to predict the occurrence of extreme precipitation and extreme surface winds (both defined by their exceedance of the local 99th percentile).

The authors addressed the questions comprehensively, and the revised version of the manuscript shows significant improvement. Below are some remaining comments and minor suggestions that may help further enhance the final version of the article. Previous comments are shown in italics, the authors' responses are in blue, and, where applicable, supplementary remarks are provided.

Minor revisions to consider

0.5°, 6 h: As discussed in your conclusion, using 0.5° may be limiting, especially if you look at small objects (such as medicanes). Also, 6 h is coarse for the Mediterranean, where storms evolve quickly. If the work is not too big, I strongly encourage you to take full advantage of the available resolutions. Another way (if increasing resolutions is not possible) could be to use products like "accumulation of precipitation within the 6 h" or "maximum wind gust within the 6 h" if they are available.

We start with the second part of the comment. Thank you for bringing this up, because we already use 6-h accumulated precipitation and maximum wind gust within 6 h as part of the standard output of the IFS ensemble. We will explain this more clearly in the revised version. Thank you for the precision.

With regard to the temporal and spatial resolution: output from the IFS ensemble is available "only" every 6 h for the entire 15-day forecast range (higher-frequency output would be available during the first six days (every 3 h) and the first 90 h (every 1 h), but using an inhomogeneous temporal resolution would make our study even more complicated.

The predictability signal for a cyclone may be extremely weak after a week. I think that your current resolution of 6 h will be a critical limitation in your part 2, probably not if you focus on large PV structures, but very important if you look at smaller scale phenomena. I would strongly encourage keeping the 6 first days with 3 h time resolution, and if it is impossible for the current work, to consider this point for future research within this framework.

Regarding the spatial resolution, this is clearly a compromise. As outlined in the general remark above, doing our data processing at a higher resolution (e.g., 0.25°) would render it unfeasible. Furthermore, although we retrieve and evaluate the forecast data on 0.5°, some of the improved information of the higher native resolution with which the forecast has been run should still be retained in our coarser dataset. I agree with this argument.

Percentile calculated each season: I do not think that this is relevant when looking at impacts. Indeed, high wind or precipitation do not impact differently following the season but following their strength. I encourage you to recalculate the results based on a fixed threshold for the whole year. Also, generally the 98th percentile has been used for wind gusts [Klawa and Ulbrich (2003)], as it was shown to fit well the observed losses. Finally, you could use the so- called Storm Severity Index to draw conclusions on the prediction of the impacts.

Thank you for your suggestions. We agree that when looking at impacts, annual percentiles might be the best option. However, since one of our long-term goals is to compare the predictability of high impact cyclones across different seasons, we introduce a seasonal threshold in this study.
I do not fully understand this argument. Would it not be simpler to compare cyclones occurring in both seasons (SON and DJF) using a common threshold? If the intention is to retain season-specific thresholds—which, from my point of view, is debatable—then a clear and comprehensive justification should be provided in the manuscript, similar to the nice one given for the distance threshold (line 178).

Regarding the percentile itself, we argue that using the 98th percentile is as subjective as using the 99th percentile.
I do not fully agree with the argument, since [Klawa and Ubricht, 2003] saw in the 98th percentile a link with insurance losses in Germany. Another argument is that even though your dataset of operational forecasts may contain many members and initialisations, it would not include a large number of different intense cyclones; therefore, the 98th percentile may be more appropriate in a statistical sense.

Minor suggestions for the revised paper
The lines given below correspond to those of the revised article (not the track-changes document).

Title: The new title is much clearer. You use "high-impact weather" while using "extreme surface weather" or "extreme objects" in every other parts of the manuscript. It may be preferable to unify the terminology throughout the manuscript.

l.27: I think the sentence "90/100 of heavy rainfall events in the western Mediterranean are attributed to cyclones" is not exactly what [Jansà *et al.*, 2001] says. "In around 90/100 of all cases of heavy rain in the Mediterranean [...] there is a cyclone centre located within 600 km of the heavy rain site or the MCS centre." It surely exists a convective system within a range of 600 km of a cyclone centre that is not dynamically linked to this cyclone.

l.41: Change "poor" by "poorer"

l.41: Since [Doiteau *et al.*, 2024] do not use a "skill score" and to be consistent throughout the article, use "performance" here.

l.61-67: This paragraph may be better placed after line 48 (or after line 32), which deals with cyclone predictability, rather than after the predictability of extreme events.

l.66: "the relevance of such storms for infrastructure and human safety". Should be reworded "*e.g.* the need for accurate predictions of such storms.."

l.71-76: "Given..methodology". The introduction was truly pleasant to read until these lines, which are unnecessary and are more appropriately placed in part 2.2. The reader should be able to appreciate the amount of work by reading the paper; therefore, I strongly suggest removing this part.

l.81: "quasi-climatological". If you do not plan to study predictability within several decades, keep "multi-year" instead of "quasi-climatological".

l.84: Precise the object of "predictability" here (of weather extremes?).

l.92: Change the sentence order: ”We discuss our results and conclude the study in Sect. 5.”

l.102: Point 4. is unclear, please reword it. On the opposite, the structure with bullet-points is very easy to read and enjoyable.

l.105: ERA5 is already available at 1 h resolution. Reword the sentence, the reader may understand that you interpolated ERA5 every 1 h. Also, since you use 6 h data, it may be appropriate to mention it here.

l.111: Check here if the physical parametrisations are also chosen randomly (I think it is the case). If it is exact, include it here along with perturbed initial conditions.

l.112: As you said in your first reply, data are available every 3 h the 6 first days. Maybe reword to say that you choose to keep only the 6 h resolution until 15 d.

l.144 and Fig. 1a.: ”illustrated in Fig. 1a”. While Fig. 1b is usefull to get what you did for the merging, the matching is already documented in [Flaounas et al., (2023)] and does not require a figure in this article.

l.168: Reword ”this is not practical given the challenge” or add a comma.

l.172: I do not understand the threshold values here. Are they your 99 th percentile? If it is the case, either reword it to explicitly say it, or remove the sentence. Since the values are quite small, it may not be the case, and if those values are indeed below the 99 th percentile, they are in all cases floored to 0.

l.185-188: This part does not improve the scientific objectivity of the paper. I strongly suggest removing it.

Table 1: Precise if the SLPmin is from ERA5 or not. Indeed, it seems that the reanalysis underestimates the ”true minimum” mean sea level pressure of cyclones.

l. 210: ”On 22 November” add 2022.

l. 226: A sentence could be added to show the coherence between a deeper

cyclone and stronger winds.

l. 243: The sentence could be changed to: "Extreme objects were diagnosed only when the storm reached the Mediterranean ..."

Fig. 9.: "seasonal climatological cyclone frequency averaged over Mediterranean" I do not understand this. Is it the probability to found a Mediterranean cyclone within the members of the ensemble at any time? Please simplify this sentence or remove the additional information.

Fig. 10c and d.: The average probability pobj is still not clear. If I understand, ploc is the proportion of members that found an extreme object at grid point x. Is pobj quantifying "how much" of the predicted object is located within the extreme object of ERA5? Please clarify this point. I would also avoid drawing full lines and dashed ones, since it is not visible in Fig. 13, and since there is already much information to process. If cyclogenesis and cyclolysis refer to ERA5, clarify it in the legend. Finally, it would be very enjoyable to have a scientific "tutorial" on the use of ploc and pobj in the text *e.g.* "The greater ploc the greater/better predicted X", "the greater pobj the greater/better predicted Y".

l. 463: remove quasi-climatological.

l. 471: "the coverage of different operational cycles" — Here, or alternatively in the Methods section, you could add a brief description of how you intend to quantify the impact of the different model versions on predictability. Otherwise, you may state explicitly your underlying hypothesis, namely that the predictability signal is expected to be stronger than the effects of model improvements.