Paper egusphere-2025-4111

Quantifying forecast uncertainty of Mediterranean cyclone-related surface weather extremes in ECMWF ensemble forecasts. Part 1: Method and case studies

by Katharina Hartmuth, Dominik Büeler, and Heini Wernli

We thank two anonymous referees for the feedback on the first revised version of the manuscript. Below, we address further comments by reviewer 1 (in black) with our replies in blue. Previous comments and replies from the first revision round are shown *in italics*. Please note that we always refer to the lines in the updated, revised manuscript (document without track changes). We supplement this document with a latexdiff-pdf showing changes since the last version of the manuscript.

Reviewer 1

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.

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.

Thank you for your comment, we agree that a higher temporal resolution would be beneficial, especially when looking at small-scale phenomena. We will consider this point for future research within this framework. For the current work, it is not feasible to manually retrieve this additional data for all cyclones from the MARS archive.

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).

If we applied a common threshold to define objects of extreme P and G10, we would lose the seasonal signal of these parameters. For example, at a grid point that experiences more precipitation in SON compared to DJF (and even more compared to MAM and JJA), a common threshold would in SON lead to larger objects of extreme P than a seasonal threshold, and relatively larger objects compared to DJF. Now comparing SON vs. DJF events would probably lead to the conclusion that objects in SON are larger and, thus, more extreme P occurs, which is – relatively spoken – not true since it just rains more in SON in general. Furthermore, we find that larger objects show a better probability compared to smaller objects. Using a yearly threshold could lead to the misleading result that P extremes in SON are better predicted compared to P extremes in DJF. However, since those SON objects would also include gridpoints with – in a seasonal framework – less extreme P, we could not do a proper comparison of predictability across seasons.

A second, less technical answer to this question is that, indeed, both approaches (seasonal vs. annual thresholds) have their pros and cons, and we think that, eventually, different stakeholders or even the same stakeholder considering different questions, might favour one approach or the other. We know that, e.g., large-scale P extremes have a strongly differing seasonality in the western vs. eastern Mediterranean (Raveh-Rubin and Wernli, 2015; their Fig. 3) and therefore using annual thresholds would lead to the identification of more cyclones associated with surface weather extremes in SON in the western and in DJF in the eastern Mediterranean. With our approach, we will have similar numbers of cyclones and extremes in both basins in both seasons, which enables more interesting comparisons across seasons and regions.

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 Ulbrich, 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.

Klawa and Ulbrich (2003) is an important pioneering study relating extratropical cyclones to damage-related losses. However, they looked at Germany, and being aware of the difficulties in robustly relating hazards to impacts across different regions, we doubt that necessarily the same percentile threshold would also link best with losses in the Mediterranean. We therefore still think that either threshold (98th or 99th percentiles) is equally meaningful and for pragmatic reasons we will keep the higher percentile. The 99th percentile is also often used when investigating heavy precipitation events in climate model simulations (e.g., Ban et al., 2021, Climate Dynamics, https://doi.org/10.1007/s00382-021-05708-w).

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.

Thank you for this suggestion. We are now only using the term “extreme surface weather” instead of “high-impact weather” for consistency 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 [Jans`a 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.

Thank you for this remark. We rephrased the sentence in L26f: “... found that over 90% of heavy rainfall events in the western Mediterranean occur within 600 km of a cyclone center.”

l.41: Change ”poor” by ”poorer”

Changed as suggested.

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

Changed as suggested.

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.

Thank you for this suggestion, this paragraph is now placed at L49f in the revised manuscript.

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..”

Changed to: “...is one example that emphasizes 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.

We disagree that these lines are unnecessary. They emphasize the relevance of focusing on the method in this part 1 which – given the feedback of all reviewers – was not immediately apparent from reading the initial version of the manuscript. However, we agree that this information would be better placed in Sect. 2.2 and moved it there.

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

We changed all “quasi-climatological” back to “multi-year”.

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

Rephrased to “...the link between the probability of these extremes and cyclone characteristics such as...”.

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

Changed as suggested.

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

Rephrased point 4 to: “Analysis of ensemble forecast probabilities of extreme surface weather objects (in a cyclone-centric framework)”

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.

Changed and extended the sentence by the following: “...forecast validation, featuring a 1-hourly temporal resolution and interpolated to a grid with a spatial resolution of 0.5°x0.5°.”

We further added the following sentence: “To make the dataset comparable to ENS (see Sect. 2.2) we only use 6-hourly data in this study.”

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.

The ECMWF switched in November 2024 (IFS Cycle 49r1) from the stochastically perturbed parameterization tendency scheme (SPPT) to the stochastically perturbed parametrizations scheme (SPP). Therefore, for the cases discussed in this study, SPP was not yet operational. However, we add information about SPPT and changed the sentence as follows: “... the ECMWF runs 50 medium-range ensemble members with slightly perturbed initial conditions and stochastically perturbed parameterization tendencies during the forecast integration.”

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.

Reworded the sentence to: “For each initialization time, we keep a 6-hourly forecast output that is available up to a maximum lead time of 15 d.”

l.144 and Fig. 1a.: ”illustrated in Fig. 1a”. While Fig. 1b is useful 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.

Given the importance of the cyclone track matching for this study, we think that both figures are useful to the reader to fully understand our methods. We agree that the matching is very similar to Flaounas et al. (2023), but readers often appreciate it when they are given the most relevant information in the study itself.

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

Added a comma: “this is not practical, given the challenge”.

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 99th percentile, they are in all cases floored to 0.

Yes, we refer to the 99th percentile here. Rephrased the sentence to: “In a next step, adjacent grid points that exceed P_{99} and $G_{10,99}$, respectively, are defined as extreme surface weather objects...”. The abbreviations used have been added in L167f above.

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

We added this paragraph since several reviewers commented on these choices and apparently it did not become clear from the previous version of the manuscript why we chose to set up the method this way. Furthermore, it has been the explicit wish of several reviewers to address our choices more explicitly in the revised version of the manuscript.

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.

Added to the table caption: “Characteristics of all three case study storms in ERA5,...”.

l. 210: ”On 22 November” add 2022.

Added as suggested.

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

Added the following: “Compared to the two other cases, the area affected by extreme winds is more than 3 times larger, which is coherent with Storm Denise showing the lowest minimum central SLP of all three storms (Table 1).”

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

Reworded the sentence to: “Extreme objects were diagnosed only when the storm reached the European continent as shown in Fig....”.

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.

As described in Sect. 2.3, “Mediterranean cyclones are identified as cyclones that reach their mature stage, i.e., their minimum central SLP, within a ``Mediterranean box'' extending from 10°W to 40°E and 30°N to 47°N (except for the Bay of Biscay in the northwestern corner). A seasonal climatological Mediterranean cyclone frequency is calculated as the spatial average of the seasonal cyclone frequency at each grid point in this box.” In the caption of Fig. 9, we refer to this averaged seasonal cyclone frequency. We simplified the sentence to the following: “Light blue line denotes an averaged seasonal cyclone frequency as detailed in Sect. 2.3.”

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”.

As stated in L344f “Figs.10c,d, 11c,d, and 12c,d show the probability of extreme objects in ENS averaged within the ERA5 object (p_obj; averaged over orange and black contour in panels (a) and (b) above)”, p_obj is the averaged p_loc within the ERA5 object contour. We added a reference to Sect. 4.2 in the caption of Fig. 10 and clarified that cyclogenesis/-lysis is referring to ERA5 in the figure caption.

We added the following to L335f: “From here on, these spatial probability fields are referred to as local probability p_loc, whereby a higher value of p_loc at a certain grid point represents a higher number of ensemble members predicting extreme objects at this grid point.” and furthermore the following in L346f: “... to further condense the information shown in the panels above. The greater p_obj at a certain timestep t_cyc, the better the prediction of the associated extreme object.”

l. 463: remove quasi-climatological.

Deleted as suggested.

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.

We added the following sentence in the method section 2.2 in L111f: “While the consideration of different IFS Cycles is unavoidable for this study, we expect that the predictability signal is generally stronger than systematic differences between the different cycles.”