

We thank the two reviewers for reading our manuscript and for providing us with further comments that can lead to an improved version of the manuscript. We provide a point-by-point response to all the comments below. Reviewers' comments appear in blue and our response in black.

Reviewer 1

The authors have addressed my primary concern by including a paragraph discussing the differences with the study that introduced the original methodology, as well as by making minor revisions throughout the text that clarify the objectives. I think the aims of the study are now clearer, and these changes have strengthened the manuscript. I have a few remaining comments:

1. (Regarding the robustness of the results to internal variability): Some analogue-based studies assess modes of internal climate variability to quantify the extent to which the findings can be attributed to anthropogenic climate change versus internal variability. For example, if most analogues occur during El Niño conditions in the historical climate but during La Niña conditions in the future climate, ENSO could be explaining an important part of the observed changes. Finding a similar distribution of analogues across different phases of key internal variability modes would increase confidence that the results are primarily driven by anthropogenic climate change (if this is indeed the aspect being tested). I do not think such an analysis is required for the present study, although it would further strengthen the manuscript and represent an improvement over the original methodology proposed by Ginesta et al. At a minimum, I would recommend mentioning in the discussion that internal variability has not been assessed in depth and that future studies could explore this aspect.

We argue that the use of the large ensemble is enabling us to assess internal variability in depth, but we have to agree with the reviewer in that we have not assessed the role of large-scale modes of variability such as ENSO. We consider that such an analysis would constitute a major undertaking as it would need to be grounded for instance on a detailed analysis of the response of North Atlantic storms to ENSO in model-world. There is still a debate on what this response is in the real system and there are still many remaining questions as to the extent to which climate change affects teleconnections (e.g. Vos et al. 2026). Therefore, the answer is far from trivial. For these reasons, we consider this analysis to be out of our paper's scope. We have however added text in the final part of the discussion pointing to the need for this kind of analysis: "However, an aspect that has not been explored in the present analysis is that of the influence of from large-scale modes of variability, such as El Niño-Southern Oscillation or the North Atlantic Oscillation, on analogue occurrence to determine whether there are any changes in these relationships attributable to climate change."

Vos, E., Huybers, P., & Tziperman, E. (2026). Climate change alters teleconnections. *Geophysical Research Letters*, 53(2), e2025GL119307.

2. (Regarding the use of an absolute intensity threshold to identify extreme analogues based on ERA5): If the percentile used to define extreme analogues is adjusted separately for the historical and future climates, then I agree that the analysis essentially compares the total number of storms in each period. However, if the historical threshold is retained and future extreme analogues are identified relative to that fixed threshold, then the comparison is no longer equivalent. Again, I am not suggesting that the authors should adopt this alternative approach; rather, I encourage them to justify their methodological choices more thoroughly and to discuss the associated limitations clearly in the main text.

The purpose of selecting a high-percentile threshold value based on all ERA5 cyclones in the Northern Hemisphere is to ensure that the definition of an extreme storm indeed corresponds to an impactful storm in the real world. We already acknowledge that this choice means that the threshold corresponds to a higher percentile in the CANARI LE in the text. However, the difference in percentile values between datasets is not large. Based on Fig. 4, which corresponds to the North-Atlantic-Europe (NAE) rather than NH cyclones, we can see that for this region the chosen threshold corresponds to the 83rd percentile in ERA5 and to the 86th percentile in the CANARI LE historical and future periods. We have expanded the discussion along these lines in the description of the methodology.

3. Lines 475–477: Please clarify why a direct comparison with Ginesta et al. cannot be performed. The methodologies remain very similar, with the main difference being the way the number of analogues is identified. Although the studies examine different storms, I think the comparison is still informative and relatively straightforward.

We said that a direct comparison cannot be performed considering that the sets of storms in each work are different. We have tried to clarify by rewriting the first two sentences in this paragraph: “Since the storms investigated by Ginesta et al. (2024) and those included in this work are different, we do not attempt to perform a direct comparison between these two pieces of work. However, we can still contrast and compare the conclusions that are drawn from each study.”

4. Line 214: I do not see why this approach necessarily provides a “cleaner” insight into future frequency changes. Rather, it relies on a different set of methodological choices, each with its own assumptions and limitations, and therefore offers complementary insights rather than a cleaner perspective.

Thank you for pointing this out. We agree. The sentence has been removed as it was unnecessary.

Reviewer 2

We thank the reviewer for reading our manuscript so carefully and for the comments that indeed are leading to a better manuscript.

The authors have provided reasonable responses to some of my comments, while I find the replies to other points less convincing. I highlight below several remaining issues which I believe need to be addressed before the manuscript may be considered further. All line numbers refer to the tracked-changes version of the revised manuscript.

1) ll. 475 and following. I appreciate this addition, but would encourage the authors to revise the phrasing. For example, they first state that the number of analogues identified in CANARI and CESM is "very similar" yet that the differences reduce "the bias in the number of analogues from CESM to the CANARI-LE". These two statements appear in contrast with one another.

A related grammatical point is that there are a number of references to "both datasets", "each present period" and similar, which do not explicitly refer to any dataset or present period mentioned in the preceding sentence. This makes the text hard to follow.

To make sense of those two sentences, we need to put them into the context of the paragraph and the rest of the analysis. When analysing ERA5, the two teams, namely Ginesta et al. and us, find a different range of numbers. When analysing the two large ensembles (LEs), namely CANARI-LE and CESM, both teams find similar numbers. When we then compare each LE against ERA5 we find that the biases assessed by each study are different. We have rewritten this part to be more explicit.

In cases such as "both datasets" or "each present period" in this paragraph they always refer to the contrast between the work of Ginesta et al. and ours. These parts of the text have been rewritten to make this explicit.

2) Sect. 3.1 When the authors write that "ERA5 has been found unable to reproduce the highest observed precipitation amount in some cases studies, but it generally exhibits small errors in the extratropics" and that "it tends to underestimate near-surface offshore wind speeds (Gandoin and Garza, 2024), but it is generally capable of simulating near-surface winds" they are providing a partial view of the literature which is not aligned with the focus of their study. While the statements that the authors make are factually correct, this study is on cyclones, so the authors should concentrate on how the reanalysis reproduces extremes. The correlation between reanalysis and observations is generally very high, but a high correlation can hide large biases for extremes. Relevant papers that the authors could refer to (although there are many others) and that make this point clearly are:

Chen, T.-C., Collet, F., & Di Luca, A. (2024). Evaluation of ERA5 precipitation and 10-m wind speed associated with extratropical cyclones using station data over North America. *International Journal of Climatology*, 44(3), 1–19. <https://doi.org/10.1002/joc.8339>

The authors focus specifically on extra-tropical cyclones and find that, while ERA5 is generally good at reproducing wind, it has "a tendency to overestimate low winds and underestimate high winds. The ERA5 skill is worse for precipitation than for wind speed". They conclude that "results highlight some important limitations of the ERA5 reanalysis products for studies looking at the possible impacts of ETCs."

Flynn, C. M., Moemken, J., Pinto, J. G., Schutte, M. K., and Messori, G.: CLIMK–WINDS: a new database of extreme European winter windstorms, *Earth Syst. Sci. Data*, 17, 4431–4453, <https://doi.org/10.5194/essd-17-4431-2025>, 2025.

The authors show that ERA5 induces a heavy smoothing of cyclone-related wind gusts over Europe compared to higher-resolution numerical products.

Hénin, R., Liberato, M., Ramos, A., & Gouveia, C. (2018). Assessing the use of satellite-based estimates and high-resolution precipitation datasets for the study of extreme precipitation events over the Iberian Peninsula. *Water*, 10(11), 1688. <https://doi.org/10.3390/w10111688>

The authors find that “both reanalysis and multi-satellite datasets overestimate (underestimate) daily precipitation sums for the least (most) extreme events over the Iberian Peninsula”.

This is not to say that we should not use reanalysis for precipitation and wind extremes – in fact, I routinely use the ERA5 dataset myself for this purpose – but we should be open about its limitations instead of trying to hide them.

The only use of ERA5 in our study is in the assessment of the performance of the CANARI-LE at representing cyclone tracks, and not for any of the aspects dealt with in the suggested papers. Therefore, given this new opportunity to improve the paper, we have edited the text to make it clear but have chosen not to include any of the suggested references. However, we have added one that is more directly related to the performance of ERA5 in representing cyclone tracks (Gramscianinov et al. 2020).

3) I don't find the replies provided by the authors concerning their hazard analysis very convincing. The statement that “Precipitation and wind at 850-hPa, as an estimator of the maximum surface wind gusts, are both impactful quantities” is not substantiated by any evidence as for what concerns the quality of wind at 850-hPa as estimator of maximum surface wind gusts. In fact, the latter data are available for ERA5, so I do not understand the authors' reticence to use such variable instead of a proxy. In a later part of their replies, the authors then justify their use of 850-hPa winds by the fact that they mostly remove the surface effects that can affect wind gusts. This would in turn suggest that they are not a very good proxy of the observed (or reanalysed) wind gusts. Arguing that variable “A” is a good proxy for variable “B”, yet that it filters out a strong control on variable “B” is hardly a convincing argument.

What we are after here is a variable that can be used to compare the potential impact due to strong winds generated by analogue storms without penalising them for differences in their actual path, beyond those allowed by the methodology itself. We need a variable that can be used to assess the strongest wind speeds that can potentially be associated with a cyclone at the surface regardless of whether such wind develops over land or over ocean. 850-hPa wind speed satisfies such requirements. We have added further literature supporting the use of 850-hPa wind speed for this purpose. An alternative approach, which we have chosen not to follow in this work, is to assess the impact on specific land (or otherwise) points or regions, perhaps due to the population density or valuable infrastructure located at that point or region. Under this approach, using actual wind gusts would be key.

4) Previous comment on l. 323. The reply provided by the authors is far from clear. The authors state in their reply that “Maximum RV is not really being used to study hazards”, yet write in the paper that: “future changes in analogue dynamical intensity and associated hazards are assessed via composite maps, generated by averaging gridded fields from all analogue tracks at their time of maximum 850-hPa RV”. Thus, RV is used to decide when in the cyclone’s lifetime to study the hazards. I therefore reiterate my comment: What is the logic for this? Would it not be more informative to select the day of largest cumulative hazard instead of the day of maximum RV if there is an interest in hazards?

We have modified the assessment done in Table 1 and rather than looking at the time of maximum RV for the analysis of PR we now look at the time of the maximum value along the track for this variable (WS850 was already being evaluated at the time of maximum WS850). We have also corrected the range of values reported in the previous version of the paper. This has resulted in an enhancement of the statistical significance in PR changes.

Maximum RV is not being assessed as a hazard in itself; however it is still being used as a time-like variable to determine the stage in the cyclones’ life-cycle in our composite analysis. Studying cyclones at the time of maximum intensity is a common practice in the field, but as with any other methodology it is subject to choices to be made by the research team. Cyclone intensity can be assessed by different variables, such as RV, MSLP or others. We have chosen RV. This is the variable that is being used for tracking the cyclones and to match the historical ones to their analogues in the CANARI LE. The rationale for the use of RV in this work is grounded on the fact that it provides a dynamically consistent parameter to assess differences between cyclones at a given life-cycle stage under different climatic conditions.

5) Previous comment 8: There is an informative paper by Wilks on why reporting the results from large numbers of local significance tests is not a good idea – statistically speaking. The paper by Wilks is very widely cited, and is hardly an obscure publication. In saying that the stippling shows significance at the X % level, the authors are effectively providing an incorrect statistical significance level in the caption/text, and doing exactly what Wilks recommends not doing. This would only be justifiable if the authors were to interpret their results one gridpoint at the time. However, this is definitely not the case, as they are looking at well-defined atmospheric features with a clear spatial structure, and interpret their results in that context. I recommend that the authors refer to the study by Wilks and I reiterate the need to revise their statistical testing.

Wilks, D. (2016). “The stippling shows statistically significant grid points”: How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society*, 97(12), 2263-2273.

We have now implemented the methodology suggested by Wilks (2016) and modify Figs. 6-9 so that they show difference fields with clear labelling of each grid point’s statistical significance: not significant, locally significant, and globally significant according to the false discovery rate methodology. The captions and text have been modified accordingly and appropriate references added.

6) Data availability: Do the authors have an indicative timeframe for when the data will become available to the international scientific community, that they could specify in the paper?

We have modified the data statement to the following: "The CANARI LE dataset is publicly available through the Centre for Environmental Data Analysis archive (Schiemann et al. 2026b)." The reference in the paper (Schiemann et al. 2026b) includes the DOI to the data. If the reviewer is interested in gaining more direct access to the data than through the CEDA website, they can contact us or any of our colleagues listed as co-authors in that same reference item. We have also updated the reference to the paper documenting the CANARI LE (Schiemann et al. 2026).