Rebuttal Letter to Reviewer 1

First of all, we would like to thank the reviewer again for the thorough feedback and appreciate the time invested for the critical comments. In the following, the comments will be in normal text formatting, our answers in bold format.

General Comments

1. With respect to the evaluation of ACE2, it appears that the authors conduct a couple of analysis approaches without a clearly founded and explained hypothesis behind it. One specific question is if a different version of the emulator ACE2 is used than what is cited. According to the publication of Watt-Meyer et al. (2025) and many locations in the text, ACE2 is using SSTs as input. However, the authors state that they use skin temperature because this parameter is needed by ACE2 (section 8.2). The authors could not convince me that the emulator can provide additional information on the searched links of storminess pattern and pressure variability patterns, and thus avoid the effort of running full numerical models. It is clearly not independent from the ERA5 reanalysis, and the fact that the emulator's training with 6 hourly reanalysis data is no guarantee that it will produce independent data (avoiding circularity) for a quantification of relationships in the aggregated winter seasons. Isn't the essential result or ACE2 merely a re-shuffling of SST and weather situations in ERA5, combining such situations which fit to each other? Given the steering of ACE2 by SSTs (or is it global skin temperatures?) and other patterns of variables (section 7.1.1) , I would not consider this a basis for independent free climate simulations.

This is just a misunderstanding of the terminology. One of the forcings used by ACE2 is indeed SST, but in the technical paper where the authors explicitly describe their forcing (Watt-Meyer et al. https://arxiv.org/abs/2310.02074) the SST are referred to as skin temperature over the ocean.

We use the version as described in Watt-Meyer et al., 2025 (see references of manuscript).

2. The reviewer criticises that in the ACE2 analysis, we are looking into other seasons than ONDJFM, different aggregations, and spatial extents of analysed regions than in the ML Prediction setup (ch6) and questions the purpose of the ACE2 validation and choice of tests.

The ML prediction setup was initially prognostic, but the score metric (R^2) was not sufficiently high. Even when moving to a diagnostic setup, this did not change. The ML method attempted to predict the PCA score. This means the spatial scope of that prediction was hemispheric (e.g. landmasses in the NH). The predictors used in this setup were global, such as the surface-temperature forcing pattern used to drive ACE2 in the chapter on sensitivity analysis. The

purpose of section 6 is to inform the scientific community that we tried this avenue, but skill metrics were not sufficient. Hence, we switched our method and proceeded to use ACE2 and conduct a sensitivity analysis.

We used quarterly seasons in the validation chapter of ACE2 (section 7) instead of only the extended winter season (ONDJFM) to verify the consistency of ACE2 regarding the wind speed variable throughout the year. The purpose of this section is to show that ACE2 can reproduce the wind speed variable, its percentiles and possible driver patterns such as the NAO. This analysis is done to extend the validation of ACE2 already done by Watt-Meyer et al., 2025. If ACE2 would have not been able to reproduce the wind speed pattern and possible predictors sufficiently well, it would not have been suitable to use it for the sensitivity analysis later on.

3. While the overall initial idea of the paper is interesting, I see a couple of aspects of the approach and its realization that must be re-thought by the authors. I wonder, for example, why the authors regard it a good idea to leave the role of extratropical cyclones out in their study, which are the local and regional link between windstorms and large-scale pressure pattern variability. Another question is the definition of the threshold wind speed for months of the year. Wind damage thresholds for buildings will depend on wind force, but be largely independent of the month, so that storm relevance will be overestimated for the autumn and spring events. While the authors claim that the resulting patterns look similar (line 259), I still see no good reason for the chosen monthly separation.

We agree that extratopical s cyclones are the primary dynamical drivers of these wind extremes. Our formulation, however, is not to exclude the extratropical cyclones as individual entities, but to identify the large-scale patterns of spatial coherence of the 10m winds. We chose a field-based approach (PCA) rather than an event-based approach (cyclone tracking) to capture these modes of coherence / synchrony in the wind-extremes. So, extratropical cyclones are not excluded per se in the analysis

In our previous response (interim response) we argued that the choice of defining extremes as exceedances of the locally defined wind percentiles is dictated by our research question. Are there large-scale coherent patterns of extreme wind variability? If we had chosen to set a constant global threshold for all locations to define wind extremes, the resulting patterns would probably just cover regions with climatological high winds. This choice would not have led to large-scale coherent patterns of winds thar are locally extreme. Similar choices are also made, for instance, when defining climate modes, such as the North Atlantic Oscillation: for instance, the NAO-index based on Azores and Akureyre time series uses standardised time series; approaches based on PCA of the SLP field usually (not always) use normalised SLP time series.

Regarding the reviewer's concerns about the suitability of the ACE2 emulator, it seems to us that the reviewer is, in this case, biased against all weather and climate emulators that have been published over the past 5 years. The reviewer's comment is not directed against specific properties of the ACE2 model. Whereas such a position probably reflects a general mistrust towards machine-learning-based emulators as lacking 'physical basis' , which admittedly is shared by others in the community, we do not believe it is a sufficient reason to dismiss them all. The ACE2 emulator, and actually all others are

indeed trained with ERA5 , and attempt to replicate the properties of the reanalysis, but they are not a 'simple reshuffling'. The ACE2 models capture the statistical autoregressive properties of ERA5 in a complex manner, and our study does show that it is able to replicate the response to changing SSTs through time, not only for thermodynamical variables, but also for circulation indices.

Specific Comments

1. Definition of the storm variability index.
   It is not clear if the different signs of variable storm activity in Fig. 2 represent a real dipole, or if it is really a monopole of variable activity which appears as a dipole because of the applied PCA. A correlation map (teleconnection analysis) showing the areas with strongest negative correlations may be elucidating in this respect.

We agree with the reviewer in this point and we will provide a simpler analysis of the local-to-global correlation between the storm index time-series at a local grid and remaining grid-cells to verify the dipole pattern of the PCA

2. Given the hemispheric scale of the PCA applied, different regional variability patterns (for example, related to the NAO or alternatively to the PNA) enter the PCA. The methodology can have the effect of putting different variability patterns of storm activity into one (see Ambaum, Hoskins and Stephenson, 2001, for a related discussion on NAO, PNA and AO). Thus, it is not clear if the hemispheric approach hides the existence of regional patterns which could have a much stronger relationship to the pressure patterns. Fig. 8 points at an intermittent link between the patterns, which has been documented in some studies some time ago, for example in conjunction with the so-called "storm track"

The reviewer is right that a more regional analysis of the links between large-scale circulation modes and wind extremes would have yielded clear relationships. For instance, the NAO is related to storms in Northern Europe. But this is not our research question. As we tried to explain in the introduction, our research question is specifically to identify hemispheric-scale coherent patterns of wind extremes, and to answer the question, for example, whether particular configurations of SST may be conductive to large-scale extreme wind variability. These configurations could be, for instance, driven by ENSO or the AMOC. There is already a vast literature regarding wind extremes at regional scales. While regional PCAs (e.g., NAO-focused) should yield higher explained variance, they would fail to capture the inter-basin teleconnections that are central to our hypothesis.

3. Chapter 5 produces some handwaving reasons for variability without actually looking into the relevant data (e.g., AMOC, SSTs and baroclinic instability). These could be explicitly checked as the respective data are available.

In this section, we correlate global SST and MSLP fields with the storm index derived by the PCA. These maps show well-known patterns indicating the influence of ENSO (strong correlation in the Eastern Pacific), AMOC (dipole fingerprint in Greenland and along the Gulf Stream), and NAO (dipole pattern in the Atlantic for MSLP). So we do use relevant SST data. Baroclinic stability would be a local driver of extremes,ma, and it would have been suitable for a regional analysis, but again, our research question is different. Additionally to those variables, we also repeated this correlation task for other variables and indices such as the geopotential height, NAO-Index, AAO-Index ENSO-Index, WYMI-Index since we wanted to use them as possible predictors for our initial ML-setup. That analysis led to the suggestion that SST and MSLP are worthwhile predictors. Hence, they are shown here.

4. I would have expected that physical concepts like geostrophic or thermal wind as well as baroclinicity are used in the analysis and the interpretation.

Parts of this can be found when we relate SSTs to the Score pattern of the PCA (section 5.1). For example, in line 289, we discuss an increase in the North Atlantic SST-signal when moving the lead time closer to ONDJFM as a possible driver of the loading pattern due to a stronger baroclinic instability since meridional temperature gradients in the ocean increase. In Line 601 we comment on the influence of the AMOC and ENSO on baroclinicity and its connection with jet-stream strength.
The reviewer's suggestion of using geostrophic or thermal wind is relevant to explain, for instance, an increase or shift of the position of extreme winds at a regional scale, but we think it is not that relevant when trying to explain the existence of large-scale coherent patterns of extreme winds, for instance whether or not a season with extreme winds in North America is also a season with extreme winds in Europe. In this case, other types of explanations are needed, such as the large-scale modification of the jet-stream. This is what is aimed at.

Nevertheless we can provide a simple correlation analysis between the mentioned variables of the reviewer and the PC-score a_1.

Other Comments

1. Line 82-86: There seems to be a surprising misconception of the authors with respect to the basic mechanisms of tropical and extratropical cyclones. I wonder how this can have made it into the submitted text.

This is indeed a big error, and we apologise for that. We will amend this sentence to something similar as: “While TCs are primarily driven by latent heat release from moist convection under barotropic conditions, ETCs are driven by baroclinic instabilities, i.e. meridional temperature gradients”

2. Line 119: What is the basis of the speculation that “regions with similar extreme wind variability may experience comparable shifts due to climate change”

The reasoning for this hypothesis is that these regions of similar variability might be (currently) coupled to the same large-scale drivers, for example, the position of the jet stream. If these drivers shift due to climate change (e.g. a poleward shift of the Jet Stream as discussed in some literature) this might lead to the comparable shifts in the storm variability of the connected regions.

3. Section 2.3: What is the pattern of the NAO-pattern associated with the downloaded index? For which months was it computed from the monthly indices at https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml? What is the role of the seasonal changes for the analysis conducted?

We downloaded the NAO-Index from here; https://www.ncei.noaa.gov/access/monitoring/nao/. The associated NAO Patterns are depicted on that page. For the correlation analysis with the storm index (which has a temporal seasonal ONDJFM timestep) we averaged the NAO-Index values to a seasonal index for seasons ONDJFM (diagnostic) and JJAS (predictive). Since our analysis focuses on ONDJFM storms we expect the NAO-Index to be an important predictor close to the beginning of the season

4. Section 3: The section is very difficult to understand, and the sense of the procedure applied with linear trend subtraction and polynomial fit does not make apparent sense to me. How does the pattern of exceedances and the time series typically look like for subsections of the time series? It is necessary to have an idea of the data eventually going into the correlations. It is not clear in how far known storms are found in the time series.

The reasons for detrending the data and subtracting the long-term trends is to avoid artificial correlations between the strom indices and possible drivers, and to avoid missing an interannual correlation due to the presence of long-term trends. For instance, SST displays a strong warming, whereas atmospheric circulation is not that sensitive to global warming. Without detrending, correlations between a storm index and SST would appear smaller than the possible interannual link between SST and wind extremes. Therefore, we computed the linear trend within the training period only and extrapolated it into the validation period in order to have the data ready for ML. In ML it is necessary to divide the timeseries into a training and validation dataset to avoid

data leakage. If we would subtract the linear trend of the whole calibration period (i.e. training + validation period) the training dataset would get information of the validation period indirectly via the trend subtraction. This leads to an erroneous performance of the ML-prediction schemes in the validation period.

The storm index is designed to remove seasonality. Hence, it allows the analysis of anomalously stormy months as it is computed relative to the percentile of a given month. Comparing this index to *absolute* extremes like cyclones resolved in IBTrACS or other sources is not sensible. Nevertheless, we did compare the IBTrACS data to a different storm index based on the same concept as the one used in our study.

To validate the idea of the storm index, we computed one percentile over the whole period (e.g. one threshold for the storm index per grid-cell for all days, not as in the paper one threshold for each month specifically). This index includes the seasonality of wind speed, making it useful for validation against observations such as IBTrACS. Comparing IBTrACS events to our index, we found a hit rate of around 80%. We can provide the details of this in the supplementary material.

5. Line 259: This is a manuscript related to meteorology. The word "trough" is linked to a specific atmospheric feature and should not be used here for describing a time series

We will change the word "peak" to maximum and "trough" to minimum

6. Section 4: It appears that the outcome of the procedure from section 3 and 4 results in rather useless patterns, with just 8% of the variance explained by the largest 5 (!) modes

The first 5 modes explain 24% of the variance (see line 243). The 8% explained variance was obtained for the setup when applying the PCA to the global grid (see lines 234-236), not the restricted grid to the NH landmasses (lines 236ff). We agree that it is confusing, since the cumulative explained variance of our first approach is 8%, similar to the explained variance of the first mode only when restricting the index to the NH landmasses.

However, the reviewer raises an important point that helps clarify the objectives of the study. The research question is whether coherent large-scale patterns of wind extremes exist. We find that these patterns do exist, but they do not contain a large portion of the overall variability in wind-extremes. The fact that the explanation at large scales is low is a result in itself, which was not known prior to the study. We, therefore, find it unfair to criticise this particular point.

7. The citation of Bellomo et al on line 292 is obviously misleading, as these authors look into 4CO2" experiments which do not represent small AMOC reductions.

We agree that the scale of the forcing in Bellomo et al. might lead to different physics compared to the natural variability of the AMOC. Our intention was to highlight the potential dynamical coupling between the AMOC and jet position. The reviewer's indicated paper (i.e., Ambaum 2001) might be a better fit here to highlight this relationship.