Rebuttal Letter to Reviewer 2

We appreciate the time invested by the reviewer and thank the reviewer for the constructive and well structured feedback.
A point-by-point commentary/answer section is provided below. Normal text is the reviewer's comment, our answers are bold.

General Comments

1. The reviewer has concerns with the low explained variance of the PCA approach and suggests to keep 52 modes that explain 90% of the storm index variance, and apply a cluster analysis to those modes in order to obtain results that are easier to grasp.

We agree that a cluster analysis over 52 modes would retain more regional information compared to only looking at the loading of PC1. The goal of this study though, is to find large-scale coherent modes across the NH and its possible large-scale drivers. A clustering approach would quite possibly identify coherent small-scale regions, without assuring whether those regions also display a temporally coherent behaviour. While a cluster-analysis would indicate regions that are covarying, it lacks the information of asynchronous regions - which is inherent in the PCA loadings.
Additionally the timeseries of the PCA-approach (a_1) is well suited for the correlation analysis to find possible drivers, such as SST, MSLP and SKT, of the obtained mode. This is the very same approach that is used to identify large-scale coherent patterns of the atmospheric circulation (e.g. NAO, AO, AAO). All those patterns are identified and defined through PCA.
Reviewer 1 raised a similar concern about the robustness and physicality of the first mode's loading pattern. In response, we plan to apply a local-to-NH correlation analysis between a local storm index in a high loading region and all other grid-cells in the NH. If the loading pattern is physical, this should be reflected in the correlations and justify the use of the PC1 loading pattern and its scores.

2. The use of ACE2 needs further support. The reviewer is asking for information on what kind of climate the ACE2 data represents and how ACE2 works in detail.

The ACE2 model is always initialised with the ERA5 data from that given day of the simulation starting point. The forcing variables are described in line 314f. We

discussed the relevant details for our study in lines 314-331 and refer to the actual publications of Watt-Meyer et al. 2023, 2025.

3. The creation of the counterfactual and factual climate is just not clear enough. Why is it not possible to use e.g. ERA5 to provide initial conditions for the emulator?

We agree that the initial conditions might be confusing, but the set up is just a classical sensitivity analysis of the impact of changed SSTs on the simulated trajectories of the atmosphere. In our setup, the initial conditions serve only as a starting point for simulating the ONDJFM season. The impact of particular initial conditions, say on October 1st, fades away after a couple of weeks, as in any atmospheric model. Thus, the precise initial conditions are actually not relevant. We focus on the impact of the forcing. This way we produce two sets of simulations, both with the same initial conditions but with different forcings.
We do use ERA5 to provide initial conditions for ACE2. The important variables that are changed are the forcings (in our case, the SST forcing for the counterfactual). Since we took the forcing variables from observations, the ACE2 simulation will produce different trajectories for each year in the simulation period, even when the initial conditions are similar. Since we only want to apply a statistical analysis of the changes between the counterfactual and factual world (and not a simulation of the observations, e.g. a hindcast), we do not need to prescribe the exact initial condition of that year. We are only interested in the statistical changes, and these are driven by the forcing, not the initial condition.

4. The reviewer finds a poor connection between sections and mentions that in Section 5.1 we use prognostic SST and MSLP correlation maps while in the ACE2 setup we use diagnostic lead times.

We use prognostic correlations between SST, MSLP and the Scores of the storm index to increase the lead time of prediction schemes (such as linear regression or random forest). Unfortunately, these schemes did produce a reasonable predictive skill at seasonal timescales(see Section 6). Nevertheless, these correlation maps show an emerging signal of ENSO and NAO, especially when decreasing the lead time (which is expected). We comment on the use of diagnostic lead times in the ACE2 setup in lines 418-426.

Specific Comments

1. L84-85: While I agree that extratropical cyclones are primarily driven by baroclinic instability the same is not true for tropical cyclones, which develop

over tropical oceans with much weaker horizontal temperature gradients and therefore much weaker baroclinicity.

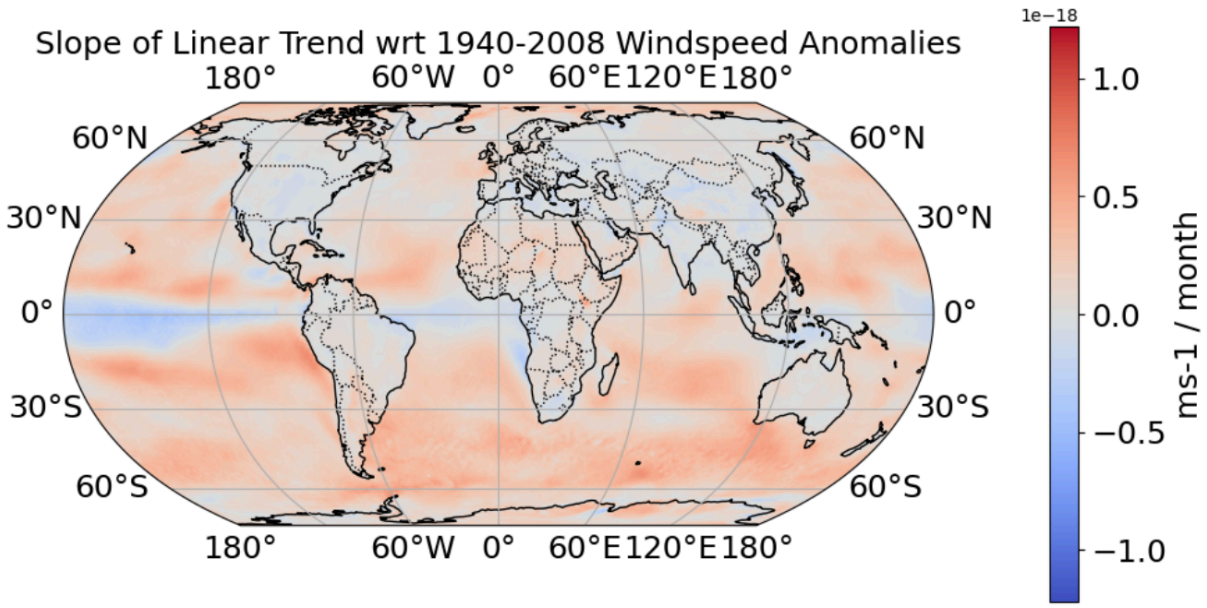
We agree and apologise for this error. We will will amend this sentence to something similar as: "While TCs are primarily driven by latent heat release from moist convection under barotropic conditions, ETCs are driven by baroclinic instabilities, i.e. meridional temperature gradients"

2. L86-87: The predominance of extratropical cyclones in boreal winter rather than summer is not due to a strong land-sea temperature gradient, but an enhanced temperature gradient between the subtropical and polar regions.

This sentence is also mistaken, and we agree that the main driving force is the meridional temperature gradient. The land-sea temperature gradient contributes to this, especially in winter, and aids the storm tracks (see von Storch et al. 2024, not in von Storch et al. 2021; we will update this).

3. L186: Comment on the validity of extrapolating the trend found on one period to a different period. How strong is the assumption that the trend will remain constant?

The main reason we did this is to avoid data leakage when applying ML methods for seasonal prediction. The reviewer is right that the underlying assumption is that the long-term trend will remain unchanged, but in a predictive scheme, in which information leakage should be avoided, there is no other way than this type of assumption (L186f). We can add global plots of the slope and intercept of the linear trend subtracted in a supplementary. Since wind speed trends are small (slope of max 1e-18 ms/month) we saw no issue in extrapolating this trend.



4. L189: ‘Due to increased computing speed…’ Do you actually mean ‘To increase computing speed…’?

Yes

5. L250: Is one mode sufficient to represent variability given the small values of the variance explained by each mode (only 24% with 5 modes)?

Given that we analyze wind extremes (and not mean statistics similar to the NAO MSLP pattern), which are by nature rare events the variance is naturally lower. Additionally, we look at the whole NH, which increases the complexity and expectedly lowers the explained variance compared to a PCA applied more regionally. Our goal was explicitly to capture the large-scale modes. This naturally comes with the caveat of a decrease in the explained variance when analyzing extremes.

6. L282-283: You talk about robust or not robust signals but it is not clear what you mean by that. Do you mean that they are persistent in time? All the correlations are below 0.4, which can be considered low.

Yes, we mean robust as in “persistent in time” and significant at the 5% confidence level (see dashed lines). The signal strength increases with decreasing lead time.

The low correlation of 0.4 is due to the nature of the correlated variables containing internal noise, i.e. wind extremes (high internal noise) and more stable drivers such as SSTs.

7. L297: I don’t see a signal over Europe and Asia, but over Africa and Asia in January.

The reviewer is correct. The signal extends into Africa and we will adjust this sentence. Given the loading pattern, the main influence of this pattern should come from the dipole contrast in MSLP close to the 50°N latitude, which is in Eurasia. We had this dipole in mind, when phrasing the sentence but agree that the signal extends to Africa.

8. L298: There is a north-south dipole but I’m not sure how much this should be identified with the NAO. In January the dipole is not at all over the Atlantic, and in summer the NAO itself is less useful.

We agree that we should generalize this to a dipole pattern between the polar and subpolar regions. These patterns occur close to the Atlantic though and are

reminiscent of the known climate modes in that region, such as the NAO and Scandinavian Pattern. We can rephrase this sentence this.

9. L326-327: ACE2 might be a dynamical system sensitive to initial conditions, but how much is that sensitivity similar to that of the atmosphere or that of ERA5?

We are not sure if we understand this comment. The real atmosphere is of course chaotic, but 'its sensitivity to initial conditions cannot be evaluated. The real atmosphere has only one trajectory. Perhaps the reviewer means to what extent similar initial conditions in the atmosphere tend to diverge. But, there are actually no two similar initial conditions in the atmosphere, globally speaking.
Regarding the sensitivity of ERA5 to initial conditions, again we are not sure if we understand this comment. ERA5 is a product of 4-dimensional data assimilation. Therefore, there are not two free-running ERA5 ensemble members differing only in the initial conditions. The ERA5 trajectories are closely constrained by observations, and there is only one observed trajectory. So the question ' how sensitive is ERA5 to initial conditions' cannot be answered either.

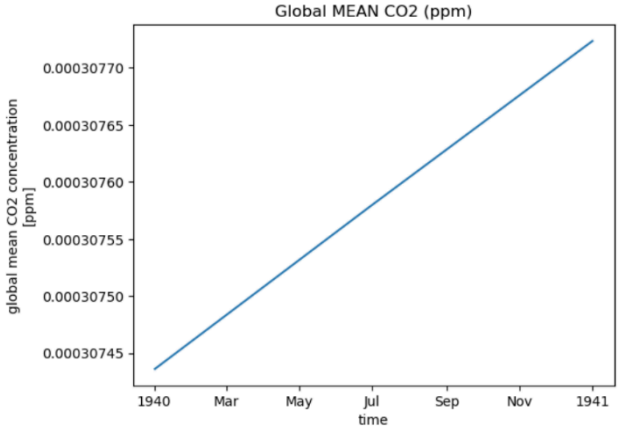
We interpret that the reviewer's comment is related to the amplitude of internal variability generated by ACE2, i.e. to what extent two trajectories started with different initial conditions diverge. This can only be compared to the internal variability generated by a full-fledged GCM. In our simulations with ACE2, the within-ensemble difference of , say, the global mean annual temperature is of the order of 0.1K, which is close to the standard deviation of the global mean annual temperature of control runs with GCMs or the standard deviation of the detrended observed global mean annual temperature. Of course, the evaluation of the internal versus external variability of the ACE2 model, considering many other variables, seasons, etc, is a whole study in itself

10. L341: What are the standard forcings obtained through downloading ACE2? Do they represent some climatic conditions in particular? Are the SSTs based on observations? If so, on what observations?

Yes, the external forcings are derived from observations. They have a temporal resolution of 6 hours, and are the same as used in ERA5. We can comment on the standard forcings in more detail but essentially, they are based on ERA5 (find specifics here: https://huggingface.co/allenai/ACE2-ERA5). As mentioned in our comment earlier, the initial conditions are not necessarily important for our usage of ACE2 to create sensitivity experiments.

11. L367: What CO2 concentrations are used in ACE2?

Global mean CO2-concentrations (see example for 1940-1941). We can add this figure to the supplementary. For more details we refer to Watt-Meyer et al. 2025, where they introduced the CO2-forcing for the first time.



12. Figure 6: Why does the time series in the left panel both end in a minimum completely discordant to the rest of the time series or their tendencies? And what is the red dashed line? In general, add labels, such as (a), (b), etc., to panels in figures to distinguish between them.

We will add panels and clarify the issues.
The cutoff exists at the boundaries of the period because we apply a seasonal aggregation. At the boundaries not all months of a seasonal window are included, hence the dropoff.

The dashed line represents a 1:1 correspondence between ACE2 and ERA5. Data above it suggests an overestimation, below underestimation.

13. L387: Why was it necessary to compute the NAO index using boxes when it was already computed using an EOF-based approach?

We wanted to compare the NAO-index between ACE2 and ERA5 in a consistent manner. Previously, we downloaded the NAO-Index to perform a correlation analysis with the PC-score to identify potential drivers. In this section, we want to validate the ACE2 model. Hence, we need to compute the NAO-Index for the ACE2 output again and compare it against ERA5. To ensure methodological consistency in the comparison we used the same boxes for both.

14. Figure 7: It is not clear what is the difference between the second and third column? What are the ERA5 statistics?

We tried unsuccessfully to clarify this in L389. We standardized the data of the climate modes by either using the ERA5 statistics (standard deviation and mean) or ACE2 statistics.
We standardized ACE2 by the ERA5 statistics (right column) to compare it to a "ground truth" and check for over or underestimation of the patterns.
We standardized ACE2 by its own statistics (central column) to check for internal consistency, e.g. if the shape of the climate mode relative to the ACE2 climatology looks similar to ERA5.
This figure shows that independent of the standardization, the patterns look quite similar, essentially validating the use of ACE2.

15. Figure 8: Why is the ERA5 correlation never negative? This leads to a big difference between ERA5 and ACE2 around 1980 involving a difference in sign. And what does the green curve represent and how is it different to the orange curve?

The ERA5 correlation does become negative in some parts. It is correct that we did not investigate why the correlations differ in 1980. Rather, we focused to show that the ACE2 is able to capture the general trend of the ERA5 coupling between the PNA and NAO. We agree that there is a difference in 1980 but the overall temporal correlation is approx. 70%, despite ACE2 being a free-running simulation, only driven by the external forcing. The agreement is therefore remarkable.
Especially, because the realistic atmospheric dynamics are not only driven by the SSTs but also for instance by aerosol forcing. Given that ACE2 is only forced by SST and CO2 (as dynamic forcings, e.g. changing in time) it is therefore remarkable that the multidecadal behaviour as derived from ERA5, is reasonably captured by ACE2.

The difference in curves is again due to the normalization by ERA5 or ACE2 statistics (see comment 14.)

16. L420-421: From the description of ACE2 it sounds as it just requires SSTs (or surface temperature) as boundary conditions. Why can these not be obtained e.g. from ERA5 for the months previous to the extended winter season?

We interpret that the reviewer suggests starting an ACE2 simulation some time prior to the extended winter season, say on June 1st and letting the model run into the winter season. Yes, the SSTs could be obtained from the ERA5 in the months previous to the extended winter season, but ACE2 is just an atmospheric emulator, it does not simulate the evolution of the ocean. Thus, a simulation with ACE2 started in the prior months would need to use the SSTs of the extended winter season anyway, and those simulations would be the same as the ones we conducted (only the initial

conditions on October 1st) would be different, but the impact of the initial conditions faded away very rapidly).

17. L428: Where are the initial conditions taken from?

The initial conditions are derived from ERA5, vertically interpolated to match the settings and levels of ACE2. These initial conditions are available on the download page of ACE2

18. L432: What is the source of the ACE2 emulator's boundary and initial conditions?

They are all derived from ERA5. These data have been interpolated by the authors of ACE2 to match their emulator resolution
We downloaded the ACE2 model based on ERA5, hence both conditions are sourced from ERA5. Find more details here (https://huggingface.co/allenai/ACE2-ERA5)

19. L441: I don't understand the restriction leading to use the same 1 October 1940 for all years. Why is it not possible to use ERA5 as before? What are the implications for the emulator's climatology? How would it differ from e.g. the more realistic ERA5 climatology?

As explained in one of the previous comments, the initial conditions used for October 1st are not really relevant. The simulations with ACE2 are an SST- sensitivity study, comparing two ensembles of simulations with the same initial conditions but different forcings. The impact of the initial conditions fades away after a few weeks, as in any GCMs

20. L446: How is SKT related to MSLP?

This is precisely one of the main open questions that we wanted to address: to what extent large-scale SST patterns may affect coherent
Patterns of extreme circulations. The reviewer's question cannot be answered in general: MSLP is tightly coupled to SST in the tropics, but not so much in the extratropics. It also depends on the temporal scale: at short time scales, the coupling is weak, however, at long decadal time scales the coupling becomes stronger.
For instance, a large-scale gradient in SKT leads to a gradient in air-density and an alteration of the Jet-Stream position. The Jet-Stream alters MSLP by sucking air up from the surface, creating low pressure.
We will add this as a comment on the connection.

21. L453: Why is M added to F and not to climatological conditions?

Because ACE2 is driven by boundary conditions, e.g. the forcing variables. It is predicting the next step (t+1) given the current step (t) altered by the forcing variables. Hence, if we want to analyze the causal effect of the surface temperature on wind speed we need to add this to the forcing F.

22. L495: I don't see the point in comparing only the magnitudes of the zonal wind disregarding the sign. The wind could be moving in opposite directions but the effect will be obscured by the authors' approach.

This is again the point of our research question, and why it differs from previous approaches. We try to identify coherent patterns of extreme winds, independently of the direction of those extreme winds. The choice suggested by the reviewer would be needed to answer a different research question, and that question would be closer to a classical analysis of large-scale atmospheric circulation patterns.

23. L501: The authors talk about a dipole but I see a tripole.

We will clarify this. Indeed the pattern is not a strict dipole but rather of alternating meridionally.

24. Section 8.4.5: The visual analysis of the sort of 'higher signal in red regions of the loading pattern and lower signal in the blue regions' is very subjective. I don't really see the signals described by the authors. Perhaps a more in depth correlation analysis would be useful here.

We acknowledge that the results are not quantified but rather by observation. We want to emphasize the differences in the wind speed when looking at the maximum and minimum composites, especially around the 50°N latitude in Eurasia. In the max composite wind speeds above this meridional threshold are emphasized, while in the min composite they are diminished and vice versa.

However, we can provide a correlation analysis between this pattern and the loading pattern to address the reviewer's comment more directly

Technical Comments

1. L143: ERA5 does not only hold daily estimates of atmospheric variables. In fact it has an hourly temporal resolution.

This is correct, we will rephrase this and note that hourly temporal resolution exists but we only use the daily aggregates of it.

2. L306: Why not mention the unmentioned variables?

Due to the scope of the paper and the fact that we later correlated the SST and MSLP. We see the point though and can also give a list of all tested variables, such as geopotential height, gradients, stratospheric temperature and PCA scores of those.

3. L340: Check whether the journal has a specific format for dates.

We will check for that and adjust it accordingly.

4. L390: It should read 'ratio' rather than 'ration'.

Point taken.

5. L466-469: Why are those very specific pressure values used and not simply 600 hPa (or 500 hPa) or 250 hPa?

These are the native levels in the ACE2 emulator.

6. L475: Add that the metric shows the *mean* differences between factual and counterfactual.

This is already mentioned in L475: "To visualize this metric spatially, we compute the temporal mean […]". The metric itself as in equation (4) is not a mean! It's the spatio-temporal difference between the counterfactual and factual world.

7. L480: Nomenclature: Why are the variables named wind_speed_5 and wind_speed_3? Where do 3 and 5 come from?

This is due to the output of ACE2. They output the meridional and zonal wind components at different vertical levels (numbered from 0 to 7). From those outputs we computed the wind speed, hence we stuck to this nomenclature to make clear that we computed the wind speed based on the zonal and meridional wind speed at that specified vertical level given by the number (the naming conventions can be found on their github: https://github.com/ai2cm/ace/blob/5494f4c67c9af6a38ec0b33ca7d0739941bf28c6/fme/ace/inference/default_metadata.py#L59)

Watt-Mayer et al include a table with the approximate heights of the ACE2 levels, but we can include this information in our manuscript as well

8. Figure 1: Figures in general require work. Increase the size of the map relative to the colour bar.

Point taken.

9. Figure 9 (and others): It is not necessary to repeat the colour bar.

Point taken

10. Figure 15: Adjust the colour scale to the values of the field. Otherwise the colour bar might get saturated as it is happening in the right column.

Point taken.