Reviewer #2

We thank the reviewer for their time, effort, and suggestions to improve the manuscript. Below is our point-by-point response. Their comments are in black; our responses follow in blue.

General Comments

1) Identification of predictors of storminess: My first comment refers to the identification of predictors of storminess in the North Sea. First, the authors indicate that the spatial coherence of storm data is short, and they mention a global map of correlations between the number of January stormy days in the North Sea grid with other grid cells (but I could not find this map). Secondly, the temporal persistence of the monthly number of stormy days is shown to be very low. From these results, the authors infer then that SST anomalies (remote or in the North Sea) do not play a relevant role in the storminess due to their long persistence and large spatial scale. Although I agree with the large spatio-temporal scales of SST anomalies, I think it would be important to show SST plots confirming their lack of influence on North Sea storminess (for instance, correlation/regression maps of global SSTs (with different lags) on the storm index or composite maps of global SST anomalies for winters with a high or low number of storms (as in Figure 3),...). I think this is particularly important for ruling out the effects of remote SST anomalies, as they influence remote regional atmosphere through teleconnections that can change with time, even at intraseasonal time scales. For instance, ENSO influence on the North Atlantic circulation in early winter is different from that in late winter (e.g.: Ayarzaguena et al. 2018).
Another suggestion would be to include SST indices of specific well-known regions such the Tropical North Atlantic or El Niño 3.4 among others as predictors in the RFR model and see if results change significantly.

Response: The map we were referring to is Figure 2. This figure shows the correlation between storminess in our basis grid cell and that in other grid cells in the North Sea. The figure shows a coherence for the neighbouring grid cell that fades at the fringes of the North Sea. We will include, in a revised version, a global version of Figure 2 in three panels: simultaneous (January-January), lag +1 (December-January), and lag -1 (January-February). Essentially, only the simultaneous map shows correlations and only within the North Sea. From that result, we argue that SSTs are an unlikely predictor on average. However, there may be particular years in which SST may play a role - see the response in the following comments on the winter 2001-2002. We will also augment those maps with the corresponding correlation maps with SSTs at the global scale.

Note that, as per a comment of reviewer #1, we will also augment the analysis by including other North Sea grid-cells, located close to the UK coast and close to the German coast, and derive the corresponding composite maps of stratospheric temperature for those cells. The difference between those components can shed light on why storminess across the three cells is only weakly correlated.

2) ACE2 experiments: I have a couple of comments of different kind regarding ACE2 Experiments. First, although the authors provide a long description of ACE2 experiments, some key aspects or results are still missing. For instance, ensemble members are mentioned in L223, but there is no information about the number of ensemble members that the authors use in this study (maybe only 6?). Moreover, the authors indicate in the abstract that apart from the emulations initialised on December 1st, they have also performed others initialised on November 1st that fail to produce a meaningful response in the ensuing months. I think this result should be also shown in the results section.
Another aspect related to ACE2 experiments concerns the variables used for the perturbated emulations. Based on Figure 5 and considering that the stratospheric circulation is highly zonal, I am not sure if the authors obtain much additional information when including the meridional wind at 70hPa. Indeed, in the case of the RFR model, only the zonal component is used. Have the authors tried

geopotential height at 70hPa instead of the two wind components? I am also curious to see the results when using stratospheric variables at higher levels in both ACE2 experiments and the RFR model analysis. This could even provide a longer time window for predictability skill.

Response: For each case (i.e., each year), two runs were performed: a control run and a perturbed run. Reviewer 1 noted that the term *ensemble* was unclear, and we agree that it was not the most appropriate choice in this context. We therefore replace the term "ensemble" with "control" and "perturbed" runs throughout the manuscript.

The November 1st figures will be added to the results section.

The reason for initialising the models with both components of the wind, in addition to temperature, was to maintain physical consistency in the initial conditions. We are aware that this is not entirely possible, as we are disturbing only one level; this choice is a compromise to initialise the ACE2 model and investigate the drivers of storminess. Should we have used only the temperature to initialise the model, and should predictability have failed, it would have been unclear whether the initialisation was too unphysical, driving the model away from a physically reasonable trajectory. Our experiments with ACE2 are the first of their kind, so there is no prior experience with how ACE2 would behave.

The geopotential height is not a prognostic variable in the ACE2 model; rather, it is a diagnostic variable, meaning it is calculated from the prognostic variables. This is why we could not tamper with the geopotential height in the ACE2 model.

3) Storminess index: As previously mentioned, the North Sea storminess index is based on the wind in a single grid point. It would be important that the authors justify why they are selecting this specific point. Moreover, the index might not be very robust as it is based on a single point. Would it make sense to perform an average of the wind over the area? Or define stormy days when there is a specific number of grid points under stormy conditions? I think it is worth testing the sensitivity of the results to the definition of the storminess index.

Response: Also, as per the comment from reviewer #1 (see previous response here), we will expand our analysis to include the other two grid cells, which should give us more insight into why strominess has relatively low spatial coherence, especially at large continental-oceanic scales. We will show that the North Sea is, in this regard, disconnected from the North Atlantic, a finding that was initially unexpected.

Specific comments

L 49 and L 50 and later on: weaker/ stronger NAO à negative/positive phase of NAO.

Response: We will use negative/positive NAO instead of weaker/stronger NAO. However, we note that the sign of the NAO index depends on the period of reference, and so the term ‘positive NAO’ may not be well defined.

L62: SSWs are triggered by the sustained dissipation of waves. Although gravity waves are also important, it is Rossby waves who play the most relevant role in the generation of SSWs.

Response: We will also add the role of Rossby waves and add a reference to support this description.

L65: do the authors mean major SSWs?

Response: Yes, we will replace “strong” SSWs with “major” SSWs.

L69-70: It is not clear the study that the authors are referring to here.

Response: By "That study", we meant to refer to Kidstone et al. (2015). We will revise this sentence.

L73-75: Some studies such as Polvani et al. (2017) have shown that the high frequency of SSWs found during both phases of ENSO in observations was related to a short observational record. In model simulations with longer data record, only El Niño phase shows a higher frequency of SSWs than during the neutral phase.
Response: We will add this study to the revised manuscript.

L87: If it is the isobaric 70hPa surface, then please remove geopotential height level.
Response: We will remove it.

L112-115: Mouallem et al. (2025) is mostly focused on the generation of SSWs in an idealized model. Papers such as Kautz et al. (2020) or SNAPSI papers (Hitchcock et al. 2022, Day et al, 2025) fit better with the purposes of this study, as they use simulations where the stratospheric state is nudged to specific stratospheric conditions to assess their effects on the troposphere.
Response: We will add those papers to the revised manuscript.

Table 1: I guess none of these values are statistically significant, but it would be good to indicate It.
Response: We will add that the Pearson correlation values are statistically not very significant.

L176: In Figures 4 and 5, the authors are not only showing zonal wind but also meridional wind. Following my second major comment, I would focus only on zonal wind and remove Figure 5. Figure 6 shows a different area with respect to Figure 2 and since it does not have latitude and longitude labels, it is not easy for the reader to identify the region of study. I would suggest showing a larger area, as in Figure 1.

Response: Reviewer #1 requested a brief description of the meridional wind anomalies. In Figure 5, positive (southerly) meridional wind anomalies are consistently observed over the North Atlantic–Barents Sea region, representing a persistent feature across nearly all composites. To address this comment, we will retain the meridional wind fields in the figures while keeping the main focus of the analysis on the zonal wind, which is more central to this study's objectives.
Regarding Figure 6, we agree that the current presentation makes it difficult to identify the region of interest. Figure 6 will be replaced with a version showing a larger spatial domain.

L308- 309: Westerly anomalies at 70hPa around 55°-70°N are associated with a weak vortex. If the authors would like to refer to an intensified vortex, then I would suggest reversing the colorbar in Figure 10 to avoid confusion.
Response: We will fix the colorbar.

Figure 10 caption: Temperature at 70hPa is shown in the bottom row.
Response: We will change the caption.

L329-330: The strongest downward coupling occurs from December to February. Kidston et al. (2015) mention November to January as a common period of active stratosphere-troposphere coupling in both hemispheres.
Response: A related point was also raised by Reviewer #1. We thank both reviewers for bringing this to our attention. Indeed, our interpretation of Kidston et al. (2015) should have been more nuanced, as their results do not fully align. We definitely do not see late-winter predictability of storminess using the methods we have applied. Kidston et al (2015) show the persistence of the 70hPa temperature field, extending well into February. One reason for the partial discrepancy may be that the signal at the surface becomes already too weak statistically. Alternatively, the ACE2 model may be delivering incorrect results, but we also do not see any statistical predictability in the RFR analysis. So here ACE2

and RFR agree. In any case, this is indeed a relevant and interesting point, and we will expand our discussion on it.

Technical comments
L63 and later on: Baldwin et al. (2020) à Baldwin et al. (2021)
Figures 3 and 4 caption: Please remove "temporal mean"
L236: cool à cold
Response to all technical remarks: Will be corrected.