Referee 2

Summary

This study investigates whether variations in anthropogenic aerosol (AA) forcing may explain the observed multidecadal variations in property losses caused by European winter storms during the past decades. It is motivated by the result of a previous published study of the author, showing that European windstorm losses exhibited a notable positive anomaly during the 1980s and 1990s. The present study uses six CMIP6 climate models (with several ensemble members per model, in total 62 simulations) to isolate the effect of AA forcing on European windstorm losses, by comparing control simulations (with all external forcings set to preindustrial levels) to simulations with historical (1850-2014) AA forcing. A simple but well-established diagnostic for windstorm losses is applied to the wind data generated by the models, and differences between multidecadal averages of the losses ('AA forced' minus 'control') are investigated. The manuscript concludes, that AA forcing could have been a major driver of the observed positive anomaly of European windstorm losses during the 1980s and 1990s. To my understanding, this is the main conclusion of the manuscript.

Although the motivation and overall approach for this study are meaningful and interesting, the manuscript does, from my perspective, not present sufficient evidence for the main conclusion. Moreover, the Introduction section needs to be improved regarding terminology and structure, and also the description of the methodology requires some clarification. Details are given by the following comments.

- The author thanks Referee 2 for their comments which have led to improvements in the revised manuscript.
- There is a misunderstanding regarding the scope of this manuscript. The main aim of this study is to measure the AA forcing of European windstorm losses using climate models. In contrast, Reviewer 2 writes that the main question is to explain the extent to which AA forcing explains the total forcing of storminess, and that the manuscript requires analysis of the modelling of total forcing to answer their question. It is hoped that the responses below clarify the main aim of the manuscript.
- The text in the Introduction and the section on Data and Methods has been reorganised, and some of it re-written, to improve the clarity of the revised manuscript.

Major comments

(1) The main question to be answered by the present study is: Does AA forcing explain the observed positive anomaly in windstorm losses during the 1980s and 1990s? (Or at least: to which extent does it explain the anomaly?) However, this question can only be answered with a climate model that does actually simulate the observed anomaly. In particular, since the observed anomaly occurred under the full historical forcing (not just AA forcing), the above question can only be answered with a climate model that simulates the anomaly under full historical forcing (the so-called CMIP6 'historical' experiment). Then we can ask: Is this simulated anomaly driven by AA forcing or by another external forcing or may it have occurred by internal variability?

Unfortunately, the manuscript does not provide any information regarding this point. In my opinion, the most straight-forward, and probably easiest, way to address this issue, would be to repeat the analysis for the corresponding model simulations with full historical forcing. Then the obtained time series of the (low-pass filtered) windstorm losses could simply be added to those obtained with AA forcing only, already shown in Figure 2.

- The above feedback states that the manuscript addresses the question: '*Does AA forcing explain the observed positive anomaly in windstorm losses during the 1980s and 1990s?... Or, to what extent does AA explain the observed anomaly*'. The reviewer suggests that the extent to which AA forcing explains the full anomaly is found by comparing modelled AA forcing to modelled full forcing signals. This leads to their recommendation to include model simulations with full historical forcings in this study.
- The strategy of this study, and how it differs from the reviewer's approach, is now described.
- The main aim of this study is to measure the AA forcing of European windstorm losses using climate models, as described in the text (Abstract, Introduction and Conclusions). Then, the relative contribution of this AA forcing to the full observed anomalies is provided for context. There are two points of contrast with the reviewer's methodology which are now discussed.
    1. The manuscript uses observed anomalies merely to place AA forcing in wider context of multidecadal storm variations, whereas the reviewer considers full-forcing as part of the main question. The scope of the study was confined to AA forcing, because it is potentially a significant driver of past storm damage variations, and of relevance to many risk managers because AA is relatively predictable in the near future. In this framework, full-forced anomalies are not examined in this study, but are instead used to place the AA signal in its wider context.
    2. *Observed* anomalies place storm signals from AA forcing in the proper context, because model simulations of full-forcings are an approximation to what happened in the real world.
- Based on the above considerations, the revised manuscript retains the original approach: to measure modelled responses to AA forcing, then assess the quality of these model responses (Section 4), and to place AA-forced signals in the context of *observed* historical data.

I also suggest to compute the corresponding time series from the ERA5 reanalysis, as there the observed positive anomaly in windstorm losses should definitely be visible.

One could perhaps create a figure with six panels, one for each model. Each panel could show the time series of: the individual ensemble members and of the ensemble mean under AA forcing only, the same (in another color) under full historical forcing, and the ERA5 time series.

This may help to judge the realism of the individual models in simulating the anomaly, to be explained by this study – in terms of both: magnitude (expressed as relative change) and timing.

- A new Fig. 2c is included in the revised manuscript, containing a reconstruction of observed losses since 1950 from Cusack (2023). It is based on winds from ERA5, with an adjustment based on long-term records of gusts at weather stations which improves the validity of loss reconstructions.
- The full historical forcing simulations are outside the scope of this manuscript (see discussion above, and below).
- Individual ensemble members differ from each other in terms of their internal variability. However, the analysis in the manuscript is focused on modelled responses to AA forcing, and modelled internal variability is beyond the scope of analysis.

It may turn out that some of the six models do not simulate the observed anomaly at all, in which case those models cannot be used to explain the occurrence of the anomaly, simply because it does not exist in those "model worlds".

On the other hand, if all six models do actually simulate the observed anomaly to a sufficient degree of accuracy under full historical forcing, and if the AA forcing only simulations reproduce that anomaly (to

some extent), then this would be an explicit indication for AA forcing being responsible for the observed anomaly (to that extent).

- The above approach, whereby the accuracy of full-forcing model simulations is used to justify AA-forced results, does not provide enough information on the validity of model simulations of AA forcing, as now explained.
- Different forcings of storminess are associated with different climate processes and pathways. For example, volcanic forcing of storminess is considered to originate in the tropical stratosphere, whereas AA forcing of storminess originates mainly in the northern hemisphere mid-latitude troposphere, in the industrial period. The quality of climate model simulations of storm anomalies will, in general, vary between the forcings, therefore, the validity of simulations from a fully-forced model provides highly uncertain information on the validity of modelled AA forcing.
- This study uses a different method in Section 4, and assesses the quality of modelled responses to AA forcing using observed data on radiative forcing and poleward energy transports. It is considered a more suitable approach because it directly evaluates the quality of modelled AA-forcing, rather than assuming it has the same validity as modelled full-forcing simulations.

(2) The terminology and structure of the Introduction section are rather confusing.

First, the four possible drivers (lines 23-24) are ordered in a counter-intuitive way. The order is: volcanic eruptions, internal variability, solar variations, AA forcing, so the first is an external forcing, then follows internal variability, and then two external forcings again. I think, it makes more sense to first list the possible external forcings and then mention internal variability. The subsequent paragraphs repeat this counter-intuitive order, so they should be reordered accordingly.

- The four causes of multidecadal variations in storminess have been re-ordered in the revised manuscript following the reviewer's suggestion.

Moreover, the paragraph from line 34 to line 47 obviously is about internal variability, but line 45 mentions the solar cycle which is represents an external forcing.

- The reference to solar cycle in this paragraph was not necessary, and has been removed in the revised manuscript to reduce potential for confusion.

In addition, it is not entirely clear to me how the concept of internal variability is used in this manuscript.

- The Introduction of the manuscript contains a discussion of internal variability, to provide wider context on the causes of multidecadal variations of storminess. The main part of the study is focused on AA forcing, and does not analyse internal variability. For example, the presented results and uncertainties concern ensemble means, which have reduced amplitude of internal variability (the different phases of modes of internal variability in each ensemble member will largely cancel each other when forming ensemble means).

Second, internal variability is called a 'driver' in the manuscript. This may make sense if properly defined, but as such a definition of the term 'driver' is not provided, it may be misunderstood as something similar to a forcing. In general, in my opinion, the terminology is sometimes imprecise throughout the manuscript and I really suggest to sharpen the terminology and wording.

- Referee 1 also raised this issue, and changes were made in the revised manuscript to avoid confusion. These changes include describing how storm variations emerge as a result of internal variability, rather than driven by internal variability.

Third, I am not sure whether all those rather lengthy explanations in the Introduction section are really needed. If condensed to what is really needed for this manuscript, some of my above caveats may already disappear.

- The Introduction was designed to give a brief description of all the identified causes of multidecadal variations in European storminess, and is quite long because researchers have identified multiple causes, and provided insights into their mechanisms in many detailed studies. The author included this wider context in the Introduction partly because it is considered good practice to place the work in its broader context, and more specifically, it helps the reader understand why later results on AA-forcing are not expected to fully explain the full observed storm anomalies in history. Further, risk managers form part of the intended audience and they are stretched by a broad remit of hazards and regions, and a relatively compact description of multidecadal storm forcings is useful to them.
- On review, a summary of multidecadal storm forcings provides key context to understand why AA-forced signals do not explain full observed anomalies, and is helpful to a key part of the potential readership. The intention of including a summary of forcings has been described more clearly in the revised manuscript. Specifically, the second sentence of the second paragraph in the original Introduction has been replaced with two sentences in the revised manuscript: 'The following four paragraphs briefly describe the evidence linking each of these time-varying forcings to multidecadal anomalies in storminess. These summaries provide wider context, prior to conducting a more in-depth analysis of one specific multidecadal forcing in this study.'

Specific comments

(3) Line 105: Is the 98th percentile computed from October to March only or from all months of the year? And is it computed from all years of the control simulation (i.e., from several hundreds of years for each simulation) or from a shorter reference period? And is it computed from each ensemble member separately or from all members together (for a specific model)?

- The 98th percentile is derived from all October to April windstorm seasons in the model's Control integration (the Control for each model consists of a single run). The final sentence of the second paragraph in subsection 2.2 was revised: 'The first stage of processing defines the 98th percentile of the daily maximum wind in each grid cell ($v_{i,98}$). This quantity is defined uniquely for each model, based on all windstorm seasons in its Control simulation (see Table 1).'

(4) Line 109: What does "of up to three days" mean? This suggests the series may in some cases, that is, for some storm events be shorter than three days. It that true and, if yes, in which cases does that happen?

- The manuscript was revised to clarify this definition: 'an event is defined as a period of up to three days centred on the days with peak values of D_d, though may effectively be of shorter duration when daily maximum winds are below the 98th percentile throughout the domain'.

(5) Line 127: "…is computed using the full time series…" What does "full" time series mean? Does it mean "unfiltered" or does it mean "full length" (i.e., several hundreds of years) or does it mean both?

- The manuscript was revised to clarify this detail: '… using the entire timeseries of unfiltered annual means from each model's Control integration…'

(6) I am surprised that the t-test is explained in the Results section, rather than the Data and Methods section. Also, because the standard error is computed from the control simulations, which have constant

external forcings, this error precisely represents/quantifies internal variability. This could be mentioned together with the description of the t-test.

- The author is grateful for the suggestion to re-organise material. The descriptions of the t-test and standard errors have been moved to a new subsection 2.3 in the revised manuscript.
- The standard errors shown in the manuscript (Fig. 2a) are a measure of the sampling variability of multimodel means from the 62 AA-forced simulations. The standard error of a Control simulation is never used in the analysis. The text in the revised manuscript (the new subsection 2.3) was modified to clarify how these standard errors refer to AA-forced simulations, rather than Control integrations. Note that the sampling variability of multimodel means has an amplitude of $1/\sqrt{62}$ of the internal variability simulated by the Control run, because the former concern averages over 62 simulations.
- It is possible to compute standard errors for single model simulations (either AA-forced or Control), however, the analysis performed in this study is focused on AA forcing.

(7) Line 217: "These tabulated values contain a pattern…" As I understand this, the "pattern" refers to the relation between the strength of the AMOC changes and the magnitude of European windstorm loss changes. And this relation is later used to construct some arguments. However: Is there any indication that this relation is either statistically significant or at least physically plausible (i.e., it does not appear just by chance)?

- The values in Table 2 concern a sample of five models, and too small for reliable statistical testing.
- However, a chain of reasoning supports the conclusion that anomalies in AMOC and storm track intensity have a dependence, whereby larger anomalies in one imply smaller changes in the other.
- In essence, Needham and Randall (2023) described how AA perturbed the latitudinal gradient of outgoing radiation at the top of atmosphere (TOA), and how the total poleward energy transport (PET) responds to this change. In turn, Needham et al. (2024) showed how the total PET anomalies are manifested as changes in two quantities in northern mid-latitudes, namely oceanic heat transport (largely due to AMOC variations) and atmosphere eddies (storm tracks). Therefore, the validity of the modelled change in storm tracks can be inferred from assessments of the accuracy of historical changes in both simulated TOA radiation and AMOC.
- The author thanks the reviewer for drawing attention to the unclear description of this framework in the original version of the manuscript. The relevant text in Section 4 has been substantially re-written to provide a more detailed description of the reasoning, which leads to the conclusion that the reported bias in many CMIP6 models concerning an over-responsive AMOC to AA-forcing suggests these models underestimate storm track intensity anomalies in the late 20th century.

(8) Line 227: "…the magnitude of modelled AA forcing is smaller than best estimates of observed…"; versus line 231-232: "…it is possible that the models studied here have too strong AA forcing…" – which sounds contradictory to me.

- The sentence in the text has been revised for clarity: '… although modelled magnitudes of AA forcing (around -1.1 Wm^{-2}) are smaller than the best estimate, Forster et al. (2021) indicate it is very likely that global ERF due to AA forcing in 2014 lies in a range from -0.6 to -2.0 Wm^{-2}, therefore, there is some chance that the AA forcings in these models are stronger than the true value.'