

Referee 1

Summary

The manuscript has been, in my view, revised in the right direction. Although the link between aerosol forcing and European storminess remains elusive, as some models do show this connection whereas other models do not, the manuscript now provides a more nuanced view of the hypothesis presented in the first version. I think it now presents a reasonable case that can be further explored in future studies, and therefore it can be interesting for some readers. I found the manuscript more clearly written and more streamlined than the first version. I have a few minor comments on this version:

- The author thanks Referee 1 for their feedback on the revision, and further comments that improve the manuscript.
- The author agrees that the evidence is presented more clearly in the revised manuscript: four of six climate models simulate AA forcing of European storminess at the 0.1% significance level, and the validation based on AA radiative forcing and AMOC indicate those models simulating stronger storm responses are more valid. Further, the physical link between AA forcing, AMOC and storminess, and associated uncertainties in quantifying this link, are clearer in the new Section 4.

Comments

- 1) L30 *'Line Multidecadal storm variations have also been reported to emerge from internal variability of the climate system..'. I think a reference is needed here to back up this statement.*
 - The sentence that follows in the text contains seven references to support this statement. This second sentence has been revised to begin with 'For example, ...' to more clearly couple the sentence highlighted above to the one that follows.
- 2) L 45 *'Reliable observational records of the European storm climate consist of one complete multidecadal cycle..'. This sentence assumes that multidecadal variability is cyclical (?), but actually it can be purely stochastic. A peak in the power spectrum does not mean that there is an oscillation. I found this sentence a bit misleading.*
 - Thanks, the sentence has been adjusted in the revised manuscript: 'Reliable observational records of the European storm climate are several decades in length, ...'
- 3) L 269. *'We measured the AA-forced contribution to multidecadal variations in European windstorm losses..'. The use of the word 'measure' here may imply a link to observations, whereas the study is mostly an analysis of model data. This may be just a language issue, but I would suggest to use 'we estimated'*
 - Agreed: 'measured' has been replaced with 'estimated' in the revised manuscript.
- 4) *Finally, the quality of the figures is not very good. The resolution is poor, and optically they look unappealing. This is admittedly a matter of taste, but at least the resolution should be improved.*
 - Thanks, all figures are plotted at higher resolution in the revised manuscript, since the information in the plots should be clear.

Referee 2

Summary

The revised version of the manuscript is a major improvement over the previous version. In particular, the confusion regarding the overall motivation of the study has been removed, the introduction section is quite clear now, and also the arguments in section 4 are easier to follow. I appreciate the revisions done by the author.

I suggest publication of the manuscript after another, but minor revision according to the following comments.

- The author thanks Referee 2 for the positive feedback on the first major revision, and is grateful for their new set of comments.

Comments

(1) L109-110: The fact that $v_{i,s}$ is obtained by applying the maximum-function in time rather than an integration in time, implies that the applied loss measure L_s is independent of the (local) duration of the storm event. I suggest to add a short comment on this, so the reader is aware of this property of the applied loss measure L_s .

- A new sentence has been inserted into the end of the 3rd paragraph of subsection 2.2 in the revised manuscript “As mentioned above, published studies of observed windstorm damages to properties, such as Klawe and Ulbrich (2003) and Prahm et al. (2015), found that event-maximum winds per location lead to valid estimates of total event losses.”

Moreover, I am curious whether the author has a specific motivation for using a duration-independent loss measure, given that, in principle, various aspects of mid-latitude storms may change (and affect losses) in response to AA forcing, like frequency, intensity and duration, the latter one being ignored by the applying a duration-independent loss measure.

- The choice of loss measure was guided by observational evidence.
 - The manuscript uses a method to convert storm winds to damages which has been extensively validated by storm losses (e.g. Klawe and Ulbrich, 2003; Prahm et al., 2015; Cusack, 2023).
 - In contrast, the author is unaware of an index with a dependency on storm duration which has been validated over a large sample of European event losses. A study was published recently which created a new index of storm strength with linear dependence on storm duration (the cumulative wind exposure index, Ørsted et al., 2026 <https://doi.org/10.5194/nhess-26-1039-2026>), but they presented no evidence on the validity of their new index to represent European storm damages.

(2) L124: I suggest to change “using the appropriate control simulation” into something more specific like “by dividing through the time mean loss of the entire timeseries of the corresponding control simulation” or something similar.

- The phrase “using the appropriate control simulation” was removed, and the paragraph was revised to describe data processing methods as clearly as possible in the new manuscript.

(3) L131: Here the subscript “s” means simulation, but in the previous subsection the subscript “s” was already used to indicate a specific storm event. I suggest to change one of them to another subscript, perhaps use “e” for event in the previous subsection.

- Thanks, the revised manuscript uses subscript “e” for event in subsection 2.2

(4) L134: If the standard error of the multi-model mean is really calculated like this, then it is mathematically incorrect. The averaging has to be done over the variances, that is, over the squares of the individual standard errors, rather than over the standard errors themselves. Specifically, it should be $\sqrt{\text{mean}[(\text{eps}_s, N)^2]} / \sqrt{M}$, where $\text{mean}[]$ indicates the average over the M simulations. This is not the same as what is actually described in the manuscript, namely: $\text{mean}[(\text{eps}_s, N)] / \sqrt{M}$. In particular, the correct expression will give larger values for the multi-model mean standard error and, thus, lower significances. Even though this may not change the arguments and conclusions of this study, it should definitely be corrected. The confidence interval plotted with the blue line in Fig. 2a might become a little bit wider when using the correct expression.

- The standard error was not calculated as described, and the author is grateful to the referee for drawing attention to this aspect, and apologises for the confusion in this aspect of the study. This new revision contains a clearer description, and reverts to the correct calculation used in the original manuscript, with changes to Figure 2a. The conclusions drawn from the results do not change.
- The revised manuscript contains a new description of the method to estimate the size of sampling errors in multimodel means: “... uncertainties in multimodel mean signals were estimated as follows. Low-pass filtered storm anomalies for historical year y were extracted from each of the 62 model simulations, then the standard deviation of this set was computed, and subsequently divided by the square root of the total number of simulations to estimate the standard error of the filtered ensemble mean for year y . This was repeated for all historical years, to yield the timeseries of standard errors of the filtered multimodel ensemble mean.”

Actually, an even better way to compute the multi-model mean standard error would be to first apply the above correct expression to each model ensemble separately, and subsequently apply it across the models. This would avoid unequal weighting of models with different numbers of ensemble members.

- The referee’s preferred method was tested, and found to produce standard errors which were about 10% smaller amplitude than those which weight all 62 simulations equally. In addition to making no difference to the conclusions, the author is uncertain as to whether the referee’s preferred method is more accurate, since it has an artefact which is difficult to interpret, namely that some individual simulations are given more weight than others in the estimate of the multimodel standard error, solely based on the size of the ensemble from each model. Given how this second method does not change conclusions, and the author’s doubts about the method (which may be unfounded), the simpler method is retained.

(5) L141-142: Because the t-test is applied to ensemble mean timeseries, and those timeseries mainly reflect the response to AA forcing, as the internal variability of the individual ensemble members is largely averaged out, those timeseries will definitely have less effective degrees of freedom (N_{eff}) than years, i.e., $N_{\text{eff}} \ll 50$, because they are auto-correlated in time, even though they are unfiltered. In the manuscript, however, there is no information regarding the value of N_{eff} used for the t-test. If $N_{\text{eff}} = 50$ has been used, then the obtained p-values are too small, that is, the results will actually be less significant than what is claimed in the manuscript (in which case the anomalies of the MPI and the CMCC model might perhaps turn out to be not significant anymore). Please, indicate which value of N_{eff} has been used and, if $N_{\text{eff}} = 50$ has been used, please, re-perform the t-test with the updated value.

- The t-test uses $N_{\text{eff}}=50$, and is considered the appropriate value to use.

- The t-test is designed to detect whether the model ensemble mean is significantly different from a null hypothesis that the storm response to AA forcing is zero. If the null hypothesis was true then $N_{eff}=50$. If, on the other hand, AA forcing has an impact on storm loss, then we expect this impact to be manifested as a series of anomalies with positive autocorrelation due to the persistence of AA forcing. The test statistic should reflect how positive anomalies with autocorrelation are a sign that the null hypothesis can be rejected, and AA forcing of storminess is non-zero.
- The referee's t-test seems to be based on a null hypothesis that the timeseries of ensemble mean values should contain AA forcing (and autocorrelation), whereas the t-test used in this study is based on the null hypothesis of no AA forcing of storm losses.
- Reduced values of N_{eff} would be appropriate in some tests, such as assessing the significance of differences between two models with AA-forced storm impacts. However, this is not analysed in the study.

(6) L156: I am wondering whether it is really 4 models having a significant loss anomaly; see point (5) above.

- There is no change to significance levels, for the reasons given in the reply to point (5) above. Four of the six models simulate an AA forcing of storminess which is different from zero at the 0.1% level.

(7) L249-250: "...the zonal-mean framework ... is consistent with the ... zonal scale...". I suggest to improve the wording to make it easier to understand this sentence.

- Thanks, this sentence was split into two smaller sentences for clarity. The sentence was placed in a wider discussion of uncertainty in validation results, which has also been revised, for clarity.

(8) L257-258: How can a difference be expressed in %?

- The AMOC trends were expressed in % in Hassan et al. (2021), their subsection 2.1, hence the difference is in the same units of %.

(9) L258-259: "These data contain a pattern". I still find this quite a claim. If I correlate these to vectors of length 5, the correlation is roughly -0.5, but its p-value is only about 1/3, so there is no significance at all. I am not questioning that entire line of arguments, but I would suggest to use a more careful wording here, something like "there is some hint that..." or so.

- The sentence was adjusted to: "The data indicate how models with greater AMOC changes have smaller European windstorm loss changes in the late 20th century..."

(10) Line 137: "signal-to-noise" -> "signal-to-noise ratio"

- Done.

(11) L196-197: Is it really "forcing" what is meant in these two sentences, or is it "response to forcing"?

- The two sentences were re-written for clarity.

(12) L241: "and may be" -> "are maybe", I guess?

- Thanks, change made.