Reply to Reviewer #2

Thank you for your insightful comments and detailed instruction on how to improve the manuscript. The quality of the manuscript has been greatly improved based on your comments. In the following, the texts with *italic font* are your original comments, and the texts with normal font are our responses. Below, we reply point-by-point, highlighting the changes we have implemented.

Summary: *I acknowledge that my previous ask: to assess whether the changes observed herein are present in the aerosol-only single forcing experiments are somewhat incongruent with the simulations described herein, i.e., different radiative forcing. However, I don't think the authors should be so dismissive of their ability to test a similar hypothesis using these simulations. I feel that the manuscript would be substantially strengthened if the authors discussed this potential in a bit more depth in their discussion.*

Response: Thanks for your suggestion. We tested the roles of greenhouse gases and anthropogenic aerosols in driving the HWI changes during 2015-2050 using "all-but-one-forcing" initial-condition large ensembles (LEs) with CESM1 (Deser et al. 2020). Four experiments were used, i.e., a 40-member ensemble with all forcing for 1920-2015 (ALL), a 40-member ensemble for 2015-2080 under RCP8.5 (RCP8.5), a 20-member ensemble with fixed GHGs at 1920 (XGHG) and a 20-member ensemble with fixed industrial aerosol (XAER) for 1920-2080. The large number of ensemble members enables an estimation on internal variability, and an estimation on the signals of regional response to anthropogenic aerosol (AA) and GHGs forcing from the noise of model's internal variability. The baseline and time period of future projection are set to 1984-2013 and 2015-2049, respectively, which are the same as the two models used in this study.

The difference between XGHG (XAER) for 2015-2049 and XGHG (XAER) for 1984-2013 is used to estimate the role of AA (GHGs) in the projected changes of HWI (blue and red boxes in Fig.A1). Consistently, increase in winter mean HWI and frequency of months with HWI≥1 under RCP8.5 relative to the baseline of ALL are also projected in CESM-LEs (red box). Both decrease in AA and increase in GHGs contribute to the

higher time-mean HWI and more frequent HWI≥1.0 under RCP8.5. It confirmed the findings of this study. Corresponding discussion is added in P23 L480-488 in the revised manuscript as follows:

"We thus further tested their roles in driving the HWI changes during 2015-2050 using "all-but-one-forcing" initial-condition large ensembles (LEs) with CESM1 (Deser et al., 2020; Key et al., 2015, Table S2 and Fig.S10 in Supplementary). The large number of ensemble members enables an estimation on internal variability, and an estimation on the signals of regional response to AA and GHGs forcing from the noise of model's internal variability. Comparing the winter mean HWI of the baseline, it increases under RCP8.5, and both decrease in AA and increase in GHG contribute to the projected higher HWI and more frequent HWI⩾1.0 (Fig.S10)."

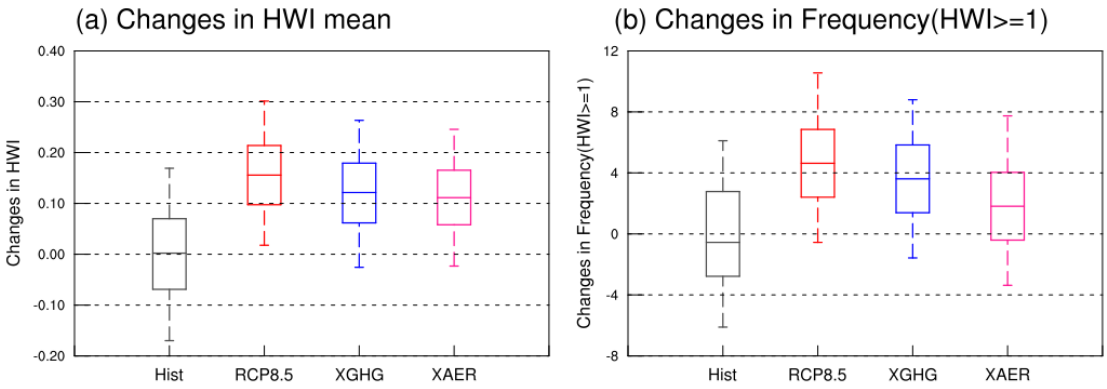

Fig.A1 Box plots for 5000 bootstrapped samples of changes in (a) winter mean HWI, and (b) frequency of month with HWI≥1 from CESM-LEs. Hist (grey boxes) is estimated from the baseline of ALL forcing experiment. RCP8.5 (red boxes) denotes the difference between RCP8.5 and Hist. XGHG (XAER) denotes the difference between 2015-2049 and 1984-2013 when GHG (AA) is fixed at 1920, and it is used to estimate the role of AA(GHG) forcing under RCP8.5. Boxes show the interquartile ranges of the 5000 bootstrapped samples, and black lines show the median. End points are the 5th and 95th percentiles. Significant difference is seen when the median from one experiment falls outside the interquartile range of another (Wilcox et al., 2020).

• *The above comment is a piece of a larger comment/suggestion: the paper in general would be strengthened with a discussion of the various ways/future directions to prove the result. Such a discussion can be presented as caveats, or future work, but*

substantively discussing the issues of your signal v. internal variability and isolating the drivers of change with single-forcing experiments will strengthen the paper and provide some inspiration to future researchers that tackle this subject.

Response: We calculated the signal to noise ratio (SNR) for the projected changes defined as the ratio of changes in MME relative to spread of ensemble members. The SNR for winter mean HWI (frequency of HWI≥1) is 1.30, 1.44 and 1.17 (1.22, 0.93 and 0.54) in RCP8.5, XGHG and XARE, respectively. The signal of changes in winter mean HWI is larger than internal variability in all experiments, consistent with the results derived from HadGEM3-GC2 and GFDL-CM3. Using the same method of this study, we estimated ranges of internal variability in Hist, RCP8.5, XGHG and XAER, respectively. Large internal variability is shown in both present-day simulation and future projection. The medium of the projected changes in winter-mean HWI and frequency of month with HWI≥1 under RCP8.5, XGHG and XAER during 2015-2049 still fall in the ranges of internal variability, but fall outside the upper quartile except for changes in frequency (HWI≥1) in XAER (Fig.A1). It gives additional support for the substantial impacts of aerosol forcing for future changes in the atmospheric conditions favoring haze events.

Corresponding discussion is added in P24 L488-498 in the revised manuscript as follows:

"The response to decrease in AA is significant, as seen from the medium of changes in the projected winter-mean HWI and frequency of month with HWI≥1 falling outside the upper quartile of internal variability (Fig.S10). The signal to noise ratio (SNR), defined as the ratio of changes in MME relative to spread across the changes of ensemble members, is higher than 1.0 (1.44) for HWI change when only AA forcing changes in the future (XGHG), consistent with the results derived from HadGEM3-GC2 and GFDL-CM3. The results from CESM-LEs give additional support for the main findings of this study, highlighting the substantial impacts of aerosol forcing for future changes in the atmospheric conditions favoring haze events. A detail examination on the role of single anthropogenic forcing and on the impact of internal variability is needed in the future."

• *I'm confused by the differences in date range choices. The NCDC & JRA analyses*

end in 2013. Models were run from 1965 to 2014 and 2016 to 2050. 1980 to 2004 is used as the baseline and 2016-2050 is used as the future. Please add explanation/justification for the lack of overlap in your analyses time choices.

Response: Thanks for spotting this. The historical simulation is from January 1965 to December 2014. In the revised manuscript, the reference period is for 1984-2013, covering the winter months from 198412/19850102 to 201312/20140102, and future projections period is for 2015-2049. All plots are updated in the revised manuscript. The results are highly consistent with previous study but with some changes in the magnitude.

• *L94-97: Investigations of global air stagnation changes using the CMIP framework (3 and 5), with relevant projections over China, should be cited here https://iopscience.iop.org/article/10.1088/1748-9326/7/4/044034 & https://doi.org/10.1038/nclimate2272*

Response: Cited as suggested (P5 L96 in the revised manuscript)

• *L210-222: Do you have a citation that indicates your bootstrapping procedure provides a reasonable estimate of the internal variability? If this is a method that you've devised, please provide evidence that this procedure adequately captures the underlying internal variability.*

Response: The method is similar to that of Zhang and Delworth (2018) who used piControl simulations to estimate internal climate variability. Here we aimed to estimate internal variability of the baseline, so we did resampling based on the baseline simulations instead here. Citation is added in the revised manuscript (P10 L211-213).

• *The use of CMIP5 realizations to constrain internal variability does not make sense. The authors note that the structural uncertainty of the individual models confounds the attempt to isolate internal variability uncertainties when using a multi-model ensemble. This method did not end up in the manuscript, but its addition to the response in support of the bootstrap method is not justifiable.*

Response: Thanks for your question. Because the two models and limit realizations in this study can't well estimate the projection uncertainties, we wanted to use CMIP5

realizations to see the ranges of uncertainty of the changes in time-mean and frequency of HWI, no matter the uncertainty is caused by model or internal variability. We thought the 107 realizations from the 24 CMIP5 models can give a relative better estimation on the uncertainty in the historical simulations. If the changes in the two models fall outside the top 5% range of the uncertainties in CMIP5 simulations, the projected changes are regarded significant. In the revised manuscript, we used a Monte Carlo method and CESM-LNs to verify the robustness of the changes. Their results are consistent with each other. Please refer our responses to your first comment and Fig.4 in the revised manuscript.

• *L210-219: As written, the bootstrap method seems to imply that 75 months of a possible 75 months are randomly sampled and averaged 2000 times. Would this not provide the same mean value 2000 times over? Should the method be listed as "bootstrapping with replacement"?*

Response: The method is resampling with replacement. The mean value of the 2000 times resampling doesn't change the mean when changing resampling times. Please note, in the revised manuscript, resampling is selected from the baseline simulation (1984-2013), while it was from the whole period of historical simulations (1965-2005). Thus, the mean value in the revised manuscript is different from the previous version. In the revised manuscript, we used 5000 times resampling (Fig.4 in the revised manuscript). Here we also show the results of 2000 times resampling in Fig.A2, the mean value of which is the same as Fig.4.

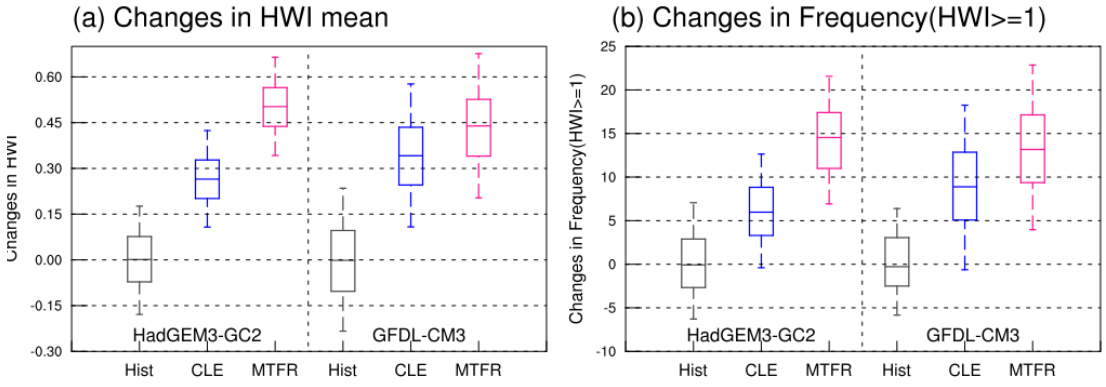

Fig.A2 Same as Fig.4 in the revised manuscript, but for the results from 2000 times resampling.

• L484-486: I identify these lines, and return to my first two bullet points regarding added discussion: I suggest turning these lines into a full discussion with citation from the literature working as a roadmap for further consideration. In my first review, I provided some examples of relevant citations, but here are a few additions: single forcing (https://doi.org/10.1175/JCLI-D-20-0123.1), internal variability & large-ensembles (https://doi.org/10.1038/s41558-020-0731-2)

Response: Citations are added.