Responses to Referee #2:

Summary: *The manuscript presents aerosol forcing sensitivity simulations from 2GCMs. It focuses on the China, and asks what effects changes in aerosols may have on future air pollution/haze events via their impact on circulation change. It is found that reductions in aerosols promote circulation patterns associated with haze events, however the intensity of such events is decreased due to the reduction of the particulate load.*

Recommendation: *Air quality in China is of major concern to public health officials, impacted citizens, and environmental scientists. The focus on the influence of aerosols on meteorology is novel and interesting. Generally speaking, projections of future air quality over China have primarily focused on the influence of GHGs on meteorological conditions. In the below I suggest greater engagement on the authors' parts with the subject of internal variability and its potential ramifications for the result presented here and the robustness thereof.*

Response: 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.

Specific Comments:

- On Line 140 the authors indicate 2 models are used to assess the robustness of results. However, some modeling groups have made available single forcing large ensembles, including simulations in which the single forcing is aerosols. Both CESM and CanESM have made available these data sets. This is a big ask, but the claims of the paper are substantial and require rigorous testing. I would have much more confidence in the claims presented here, if the authors were to test their hypothesis using these data:

– Relevant CESM publication: Deser et al 2020 https://journals.ametsoc.org/jcli/article/33/18/7835/353234/Isolating-the-Evolving-

Contributions-of – Relevant CanESM publications: —Swart et al 2019 https://gmd.copernicus.org/articles/12/4823/2019/ —Santer et al 2019 https://www.pnas.org/content/116/40/19821

Responses: Thanks for recommending those individual anthropogenic forcing large ensemble simulations from CESM, CanESM2 and CanESM5. This study aims to estimate changes in Beijing haze events under different anthropogenic aerosol emission scenarios in the future. Thus, we used two climate models forced by different anthropogenic aerosol forcings but the same greenhouse gas emissions following the RCP4.5 scenario. However, the experiment designs for single forcing large ensemble simulations are different from this study. The relative role of individual anthropogenic forcing can be estimated from the single forcing simulations, while the impact of different anthropogenic forcing scenarios can not be estimated. In addition, the greenhouse gases scenario in CESM and CanESM2 are under the RCP8.5 scenario, different from this study as well. Thus, the single forcing large ensemble simulations cannot be used here to answer the scientific questions of this study. So, we didn't use those single forcing experiments in the revised manuscript. We also agree that the single forcing experiments provide good database to test the relative role GHG, AA and internal variability in HWI changes both in historical and future projection. This is discussed in the end of the revised manuscript as follows (P23 L484-487).

"But internal variability may not be fully sampled because of the limited number of realizations and models used in this study. In the future, single forcing experiments and large ensembles simulations are useful ways to confirm the relative role of greenhouse gases and anthropogenic aerosol forcing on haze event."

- I am likewise concerned with the authors' lack of consideration/discussion of model simulated internal variability. Beijing's haze events have received a lot of attention, to include work done by researchers that have articulated the role of internal variability on past (Zhang et al 2020 https://acp.copernicus.org/articles/20/12211/2020/; Callahan et al 2019 https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JD029738) and future

(Callahan & Mankin 2020 https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020GL088548) air quality-relevant circulation changes. I would like to see the authors contextualize their results in light of the findings presented in these papers. Do the sample sizes studied in this manuscript approach those required to account for signals rising above noise?

Response: Thanks for this insightful comment. To answer this questions, we used two methods to estimate internal variability, then tested the significance of the changes of mean HWI and the frequency with HWI≥1 under the two scenarios. Both methods support the robustness of our previous results.

Method 1: A bootstrapping approach. We estimated internal variability by performing bootstrapped samples. This resampling-based procedure involves three steps. First, we randomly select 75-month (135-month), i.e. 25-yr (45-yr) winters, from His (projections), and calculate the mean HWI change of the 75-month relative to His or frequency of month with HWI≥1 of the 75-month. The 75-month and 135-month are selected to mimic any 25-yr in the period 1980-2004 and 45-yr in 2016–2050, respectively; We repeat the first step 2000 times, and the 2000 bootstrapped samples can be viewed as internal variability of His or future projections. We then compare the results of ensemble mean of each model with those of the 2000 bootstrapped samples. If it falls outside the top 5% of the distribution, we then claim that the projected changes in mean HWI or frequency of month with HWI≥1 are statistically significant at the 5% level and beyond the variability of internal variability. (P10-11 L210-222 in the revised manuscript).

The difference in mean HWI between CLE vs His, MTFR vs His, and CLE vs MTFR, are also statistically significant at the 5% level in both models (Fig.A1c-d here, Fig.4a-b in the revised manuscript). The frequency with HWI⩾1 in CLE and MTFR are both statistically different from that in His in the two models, while only in HadGEM3-GC2 simulations is the frequency in MTFR statistically significant from that in CLE at the 5% level (Fig.A1c-d here, Fig.4c-d in the revised manuscript). (P14 L291-296 in the revised manuscript)

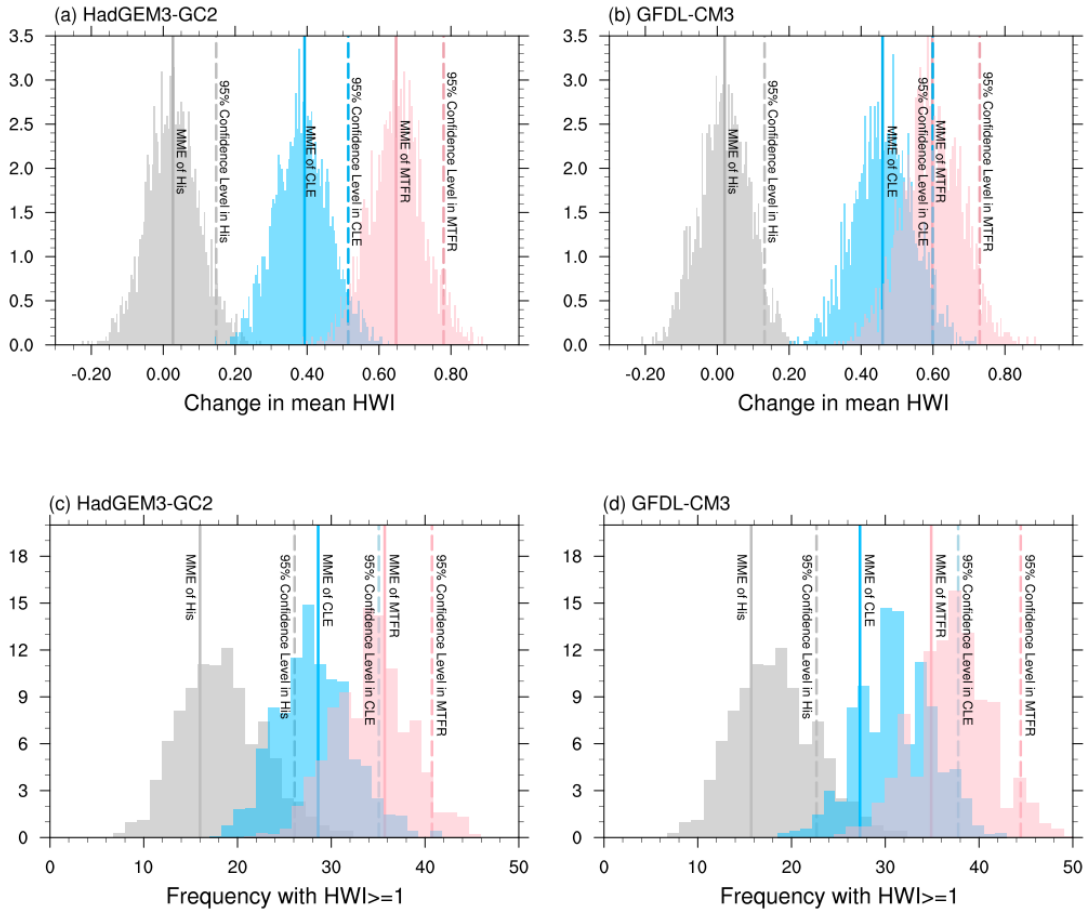

Fig.A1 Histogram plots of (a) changes in winter mean HWI, and (c) frequency of month with HWI≥1 in HadGEM3-GC2 for the 2000 bootstrapped samples, and (b), (d) similarly for GFDL-CM3. The grey, blue and pink shadings are the results estimated from His, CLE and MTFR respectively. Solid (dashed) grey, blue and pink lines are the results of multi-member mean (95% confidence level) in His, CLE and MTFR, respectively. (Fig.4 in the revised manuscript)

Method2: We used historical simulations of multi-model output from CMIP5 to estimate internal variability of His. 108 realizations of CMIP5 for the period 1961-2005 are used in total (model list is shown in Table A1). Internal variability of changes in mean HWI is estimated as ranges of the difference between any running 25-year mean HWI and the baseline (1985-2004) of each model. Internal variability of frequency of month with HWI≥1 is the range of frequency of month with HWI≥1 in any running 25-year in historical simulation. The results are shown in Fig.A2. The range of internal variability estimated from method 2 is larger than that from method 1, as it also includes a degree of structural uncertainty between the CMIP5 models. As for the winter mean

HWI changes and frequency of month with ≥1 in CLE and MTFR projected by HadGEM3-GC2 and GFDL-CM3, they are also statistically significant at the 5% level using this method, consistent with *Method 1*. A shortcoming of *Method 2* is that it cannot estimate internal variability of future projection. Thus, we used *Method1* in the revised manuscript rather than assuming that internal variability remains the same in future.

Table A1. List of CMIP5 models used to estimate internal variability

Model	Institute, country	Atmosphere Resolution (lat×lon)	realizations
ACCESS1-0	CSIRO-BOM, Australia	1.2°×1.9°	1
bcc-csm1-1-m	BCC, China	2.8°×2.8°	3
bcc-csm1-1	BCC, China	1.1°×1.1°	3
BNU-ESM	BNU, China	2.8°×2.8°	1
CanESM2	CCCma, Canada	2.8°×2.8°	5
CanCM4	CCCma, Canada	2.8°×2.8°	10
CCSM4	NCAR, USA	0.9°×1.2°	8
CESM1-CAM5	NSF-DOE-NCAR, USA	0.9°×1.2°	3
CMCC-CN	CMCC, Italy	3.7°×3.7°	1
CNRM-CM5	CNRM, France	1.4°×1.4°	10
EC-EARTH	EC-EARTH, Europe	1.1°×1.1°	6
FIO-ESM	FIO, SOA, China	2.8°×2.8°	3
GFDL-CM3	NOAA/GFDL. USA	2.0°×2.5°	5
HadCM3	MOHC, UK	2.5°×3.7°	10
HadGEM2-CC	MOHC, UK	1.2°×1.9°	3
HadGEM2-ES	MOHC, UK	1.2°×1.9°	4
inmcm4	INM, Russia	1.5°×2.0°	1
IPSL-CM5A-LR	IPSL, France	1.9°×3.7°	6
IPSL-CM5A-MR	IPSL, France	1.4°×2.5°	3
MIROC5	AORI-NIES-JAMSTEC, Japan	1.4°×1.4°	5
MIROC-ESM	AORI-NIES-JAMSTEC, Japan	2.8°×2.8°	3
MPI-ESM-MR	MPI-M, Germany	1.9°×1.9°	6
MRI-CGCM3	MRI, Japan	1.1°×1.1°	5

NorESM1-M	NCC, Norway	1.9°×2.5°	3

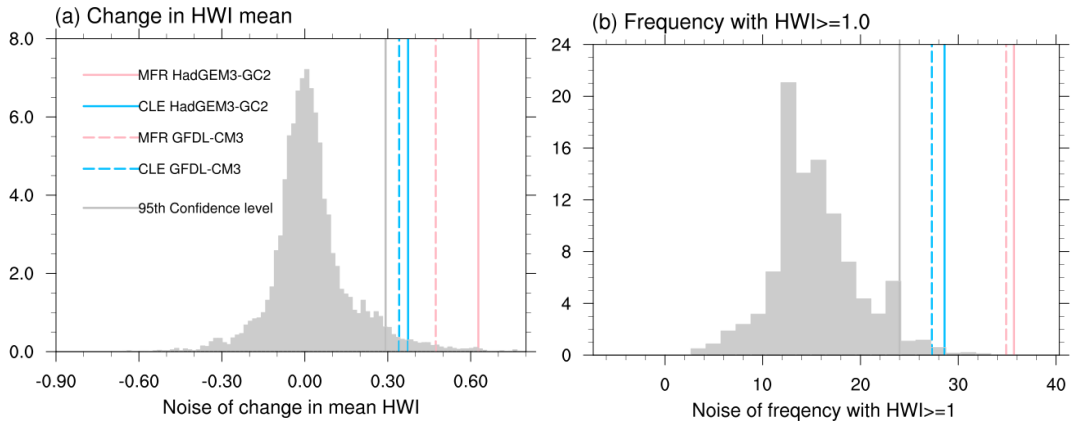

Fig.A2 Histogram plots for (a) winter mean HWI change and (b) frequency of month with HWI≥1 in His estimated from 108 realizations of CMIP5 multi-model output. The solid grey lines are the 95% confidence level in His. The solid (dashed) blue and pink lines are the results from CLE and MTFR simulated by HadGEM3-GC2 (GFDL-CM3), respectively.

- The presentation of the results gets muddled beginning on Line 210. Up to this point we've been discussing HWI greater than 0, as defined by Cai, but now we've switched to greater than 1.0. We are also now talking about reanalysis data, but we only know that from the figure caption.

Response: Thanks for pointing this out. It should be HWI-daily greater than 0.0. We've corrected it in the revised manuscript. To make it clearer, we revised the title of section 3 into "Favorable climatic conditions for Beijing haze events in reanalysis", and moved the model evaluation to section 4.1.

- JRA55 renanlysis is used to assess the ability of the models to simulate key synoptic features relevant to air quality/haze events. However, I am curious if JRA55 is able to simulate historical poor quality conditions over China. Haze data in China goes back a few decades and there are some notable examples or extremely poor air quality. Does the JRA55 capture these events? Can they be identified on Figure 3?

Response: Thanks for this good question. To answer your question, we tested the consistency between JRA-55 and NCEP-NCAR, and examined the relationship

between HWI-month derived from JRA-55 and observed haze occurrence and haze visibility. The observed haze is calculated from the station observations from the National Climatic Data Center (NCDC) Global Surface of the Day (GSOD) database for 1974-2013. The main results are as following:

(1) Consistency between JRA-55 and NCEP-NCAR reanalysis.

HWI-daily is defined based on the synoptic circulations associated with observed PM2.5 in Beijing by using NCEP-NCAR reanalysis (Cai et al. 2017), which has proved that HWI-daily can well represent the poor air-quality conditions in Beijing. Here we used JRA-55 because of its higher resolution and better quality over East Asia. We compared the HWI derived from JRA-55 and NCEP-NCAR reanalysis from daily time scale and long-term changes (Fig.A3). We found that the two reanalysis datasets are almost identical with each other in depicting the daily variation of HWI.

A short discussion is added in the revised manuscript (P7 L140-142) as: "The variation of haze index derived from JRA-55 are highly consistent with those from NCEP-NCAR reanalysis (not shown). We only use JRA-55 in this study."

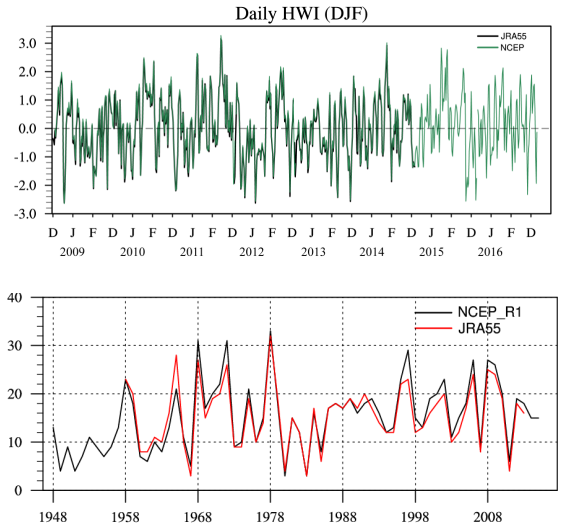

Fig.A3 (a) daily changes of HWI from 1st December 2009 to 31st December 216 in JRA-55 (black) and NCEP-NCAR (green). This period is same as Figure 2 in Cai et al. (2017). (b) Number of days (unit: days) with daily HWI>0 in the winter (DJF) from 1948 to 2013 in JRA-55 (red) and NCEP-NCAR (black).

(2) **Relationship between HWI-month derived from JRA-55 and observed poor air-quality event in China**

This question is also one of the main concerns from Referee #1. We used observed daily visibility, relative humidly and wind speed from 1974 to 2013 from the National Climatic Data Center (NCDC) Global Surface of the Day (GSOD) database (Fig.S1a in the revised manuscript). The positive values of HWI-month are associated with more haze days and less visibility over Beijing and the surrounding region (box in Fig.A4c-d). It proved the good relationship between HWI-month and observed haze occurrence and haze intensity (Fig.A4 here, and Fig.2 in the revised manuscript). Please refer our responses to the first question of Referee #1 for further details.

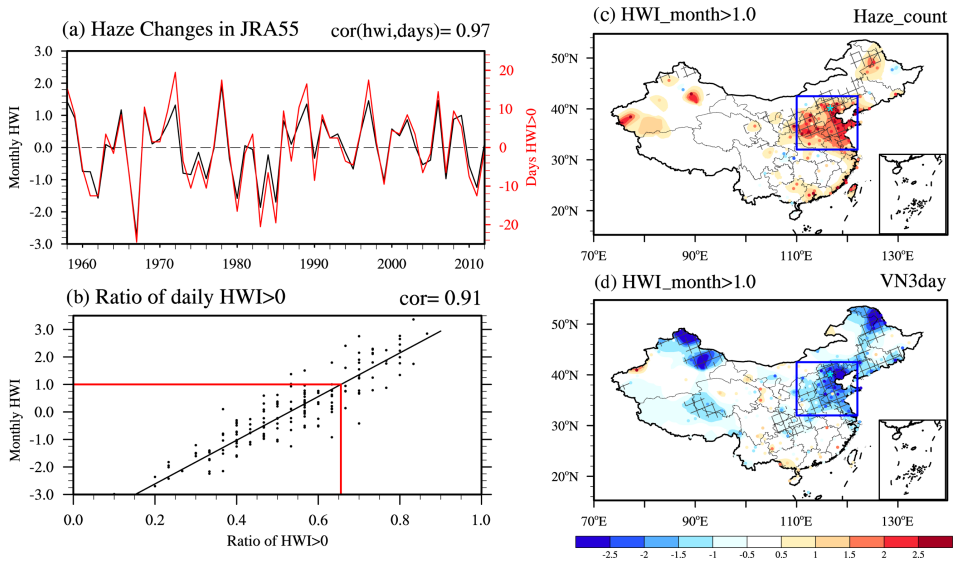

Fig. A4 Changes in winter HWI from 1958 to 2013 in JRA-55 reanalysis relative to 1958-2013 winter mean. (a) DJF mean monthly-based HWI (HWI-month, black line) and the anomalous days with daily based HWI >0 (HWI-daily, red line, unit: day), (b) scatter plot of HWI-month of December, January and February (y-axis) and the ratio of days with HWI-daily>0 (x-axis) in each winter month. HWI-month and HWI-daily are the HWI calculated from monthly data and daily data, respectively. (c)-(d) are the anomalies of haze occurrence and the VN3day when HWI≥1, where VN3day is the minimum 3-day consecutive visibility. Cross area in (c)-(d) is statistically significant at the 10% level using a Student's test. (Fig.2 in the revised manuscript)

- Figure 3 a&b: The blue CLE data obscures the MTFR data. If showing the data to your readers is important, please do so.

Response: We deleted Fig.3a-b since there is no significant trend here. Instead, we added CDF distributions of HWI in the revised manuscript (new Fig.3 and P14 L289-296).

- Figure 3 c & d: This may be personal preference, but I've always had trouble reading discontinuous box plot pdfs. Could simple linear pdfs be used here? I assume the relatively differences in chape is more important to convey than the numbers at each gradation of HWI? I would also appreciate the statistical analysis indicated on each distribution comparison, i.e., at what statistical threshold are these distributions significant?

Response: Thanks for this suggestion. I modified this plot by using linear pdfs based on non-parameter kernel method. The significant threshold is also added. (please see new plots in the revised manuscript)

- I would reiterate the above comment for all of the plots of this style in the manuscript.

Response: All plots are modified as suggested.