Referee's comments are in red, our reply in black, quotes in the revised manuscript in blue.

Referee 1's comments

General Comments

The authors seek to address the question of how much apparent temperature in Beijing will vary under different future scenarios of climate change (including geoengineering). This includes an analysis of whether different downscaling methods – either statistical or dynamical – yield different results. They find that, although both methods (when applied to results from 4 global ESMs) yield roughly similar results for the present day, the same is not true when inspecting the effects that climate change and geoengineering will have. The study highlights the important issue of changes in human-relevant variables such as apparent temperature (and the number of times a threshold is crossed) rather than relatively abstract variables such as global mean surface temperature.

I struggled with this review because I could not clearly identify the core contribution. The base idea of whether statistical downscaling or dynamical scaling results in different outcome estimates is certainly important, but this question has been thoroughly discussed in a companion paper by the authors which looks at the same data for the same domain from the same models, and was submitted recently to this journal (https://esd.copernicus.org/preprints/esd-2022-35). The remaining question is whether apparent temperature is differently affected than more conventional meteorological variables, which is a relatively boutique concern. The methods used to address this questions are nonetheless appropriate, and the data produced generally support the conclusions. However, the existence of the companion paper (which I recognize the authors do cite) makes the contribution of this manuscript incremental.

The use of multiple downscaling techniques with multiple models is interesting and well executed, and it is particularly encouraging to see applications to health-relevant outcomes. The biggest issue is a lack of significant impact, although I also have some methodological concerns. I have laid these out in detail below, starting with major comments. If the paper can be focused more heavily on outcomes – in particular, the effect that downscaling has on health-relevant impacts – then I believe it could significantly improve its relevance and impact. This would also help to address the issue that the paper is not particularly interdisciplinary, which is a stated requirement of ESD. As such, in its current state I cannot recommend it for publication.

Reply: We thank your constructive comments, which help us clarify and improve the study vastly. There are two main problems, one is the lack of innovation, the other is the method of processing data. These two issues are also detailed in the major comments. We have responded to the major comments below one by one.

Major comments

The greatest issue is the lack of a clear and impactful outcome. The methods applied are interesting in large part because they look at interesting scenarios (RCPs versus geoengineering versus recent past) and include a significant problem (the performance of statistical versus dynamical downscaling). However, these issues are the focus of a paper which is already under review, and as such cannot be the major novelty of a second manuscript. I therefore assume that the major conclusions regard the question of change in apparent temperature, with the authors finding that changes in apparent temperature will be greater under RCP 8.5 than under a geoengineering scenario, and that this is mostly because of increases in temperature. The issues I perceive here are twofold. Firstly, apparent temperature – while an important metric – is just one metric of impact, and a relatively straightforward one which is (evidently) mostly just reflecting changes in temperature. The manuscript would be greatly improved if multiple outcomes were assessed rather than just one, to see if the different downscaling methods have different impacts on such outcomes. This could include, for example, regional air pollution (if reported in any of the ESMs). Alternatively, a deeper analysis of the likely consequences – for example by attempting to quantify the differences in health outcomes or costs, and the degree to which different demographics or sub-populations are affected – would help to improve the interdisciplinarity of the manuscript. Secondly, the current analysis is somewhat limited, being mostly observational (report differences) rather than explanatory. The manuscript would be greatly improved if the authors could provide mechanistic explanations for their findings; why, for example, does WRF-based downscaling seem to result in such a different seasonality in AP – T compared to statistical downscaling?

Reply: Thanks for these thoughts. We wanted to address a problem related to impacts of changes in the fundamental weather fields. The suggestion of looking at regional pollution is most relevant we felt. However, essentially the output from ESM on $PM_{2.5}$ simulations is not good (e.g. Ran et al., 2022). So we explored other ways of projecting air pollution. Based on our existing downscaling data, we further explored the impact of geoengineering on $PM_{2.5}$ in the Beijing-Tianjin region using the multiple linear regression model. We also explore the changes in $PM_{2.5}$ related relative risks of 5 main diseases.

References

Ran, Q., Lee, S., Zheng, D., Chen, S., Yang, S., Moore, J., and Dong, W.: Potential Health and Economic Impacts of Shifting Manufacturing from China to Indonesia or India, Sci. Total Environ., 855, 158634, https://doi.org/10.1016/j.scitotenv.2022.158634， 2023.

The details are as follows:

Introduction

In early 2013, Beijing encountered a serious pollution incident. The concentration of $PM_{2.5}$ (particles with diameters less than or equal to 2.5 μm in the atmosphere) exceeded 500 $\mu g/m^3$ (Wang et al., 2014). Following this event and its expected impacts on human health (Guan et al., 2016; Fan et al., 2021) and the economy (Maji et al., 2018; Wang et al., 2020), the Beijing municipal government launched the Clean Air Action Plan in 2013. The annual mean concentration of $PM_{2.5}$ in Beijing-Tianjin-Hebei region decreased from 90.6 $\mu g/m^3$ in 2013 to 56.3 $\mu g/m^3$ in 2017, a decrease of about 38% (Zhang et al., 2019), although this is still more than double the EU air quality standard (25 $\mu g/m^3$) and above the Chinese FGNS (First Grand National Standard) of 35 $\mu g/m^3$. The concentration of $PM_{2.5}$ is related to anthropogenic emissions, but also dependent on meteorological conditions (Chen et al., 2020). Simulations suggested that 80% of the 2013-2017 lowering of $PM_{2.5}$ concentration came from emission reductions in Beijing (Chen et al. 2019). Humidity and temperature are the main meteorological factors affecting $PM_{2.5}$ concentration in Beijing in summer, while humidity and wind speed are the main factors in winter (Chen et al., 2018). Simulations driven by different RCP emission scenarios with fixed meteorology for the year 2010 suggest that $PM_{2.5}$ concentration will meet FGNS under RCP2.6, RCP4.5 and RCP8.5 in Beijing-Tianjin-Hebei after 2040 (Li et al., 2016).

There are large uncertainties in projecting $PM_{2.5}$ concentration in the future due to both climate and industrial policies. Statistical methods are much faster than atmospheric chemistry models (Mishra et al., 2015), and different scenarios are easy to implement. We use a Multiple Linear Regression (MLR) model to establish the links between $PM_{2.5}$ concentration, meteorology and emissions (Upadhyay et al., 2018; Tong et al., 2018). We project and compare the differences of $PM_{2.5}$ concentration under G4 and RCP4.5 scenarios, and between different $PM_{2.5}$ emission scenarios. Accurate meteorological data are crucial in simulating future apparent temperatures and $PM_{2.5}$ because all ESM suffer from bias, and this problem is especially egregious at small scales. A companion paper (Wang et al., 2022) looked at differences between downscaling methods with the same 4 Earth System Models (ESM), domain and scenarios as we use here.

1. $PM_{2.5}$ concentration and emission data

In China there were few $PM_{2.5}$ monitoring stations before 2013 (Xue et al., 2021). However, aerosol optical depths produced by the Moderate Resolution Imaging Spectroradiometer (MODIS) have been used to build a daily $PM_{2.5}$ concentration dataset (ChinaHighPM2.5) at 1 km resolution from 2000 to 2018 (Wei et al., 2020). We use monthly $PM_{2.5}$ concentration data during 2008-2015 from ChinaHighPM2.5 to train the MLR model, and the data during 2016-2017 to validate it. Figure S1 shows annual $PM_{2.5}$ concentration over Beijing areas during 2008 (a) and 2017 (b).

Recent gridded monthly $PM_{2.5}$ emission data were derived from the Hemispheric Transport of Air Pollution (HTAP_V3) with a resolution of 0.1°×0.1° during 2008-2017, which is a widely used anthropogenic emission dataset (Janssens-Maenhout et al., 2015). $PM_{2.5}$ emissions over Beijing areas during 2008 (c) and 2017 (d) are shown in

Fig. S1.

Future gridded monthly $PM_{2.5}$ emissions to 2050 are available in the ECLIPSE V6b database (Stohl et al., 2015), generated by the GAINS (Greenhouse gas Air pollution Interactions and Synergies) model (Klimont et al., 2017). The ECLIPSE V6b baseline emission scenario assumes that future anthropogenic emissions are consistent with those under current environmental policies, hence it is the "worst" scenario without considering any mitigation measures (Li et al., 2018; Nguyen et al., 2020). Projected emissions are shown in Fig S2, with emissions plateauing at ~40 kt/year after 2030, so we assume 2060s levels are similar. These ECLIPSE projections are significantly larger than present day estimates from HTAP_V3. We therefore estimate 2060s emissions as the recent gridded monthly $PM_{2.5}$ emissions from HTAP_V3 scaled by the ratios of 2050 ECLIPSE emission to average annual emissions between 2010 and 2015. Before processing data, $PM_{2.5}$ concentration is bilinearly interpolated to the WRF and ISIMIP grids, while $PM_{2.5}$ emissions are conservatively interpolated to the target grids.

2. MLR model calibration

Previous studies have shown that wind and humidity are the dominant meteorological variables for $PM_{2.5}$ concentration in region we study (Chen et al., 2020). Hence, we generate an MLR model between $PM_{2.5}$ and temperature (T), relative humidity (H), zonal wind (U), meridional wind (V) and $PM_{2.5}$ emissions (E) at every grid cell as follows

$$PM2.5=\sum a_i X_i+b \qquad (1)$$

Where $X_{i(i=1,2,3,4,5)}$ are the five factors, a_i are the regression coefficients of the X_i with $PM_{2.5,}$ and b is the intercept, which is a constant. We assume that all factors should be included in the regression. All the meteorological variables are from the statistical and dynamical downscaling and bias corrected results during 2008-2017, with the first 8 years used for training model and the second 2 years used for validating model. We train the MLR for the 4 ESMs under statistical and dynamical downscaling in each grid cell separately, thus accounting spatial differences in the weighting of the X_i across the domain. Meteorological variables under G4, RCP4.5 and RCP8.5 during 2060-2069 are used for projection.

The contributions of meteorology and $PM_{2.5}$ emissions on future concentrations are examined by using recent $PM_{2.5}$ emissions (baseline) and future $PM_{2.5}$ emissions (mitigation), and the downscaled climate scenarios. Modeled $PM_{2.5}$ concentration using recent meteorology and PM2.5 emissions during 2008-2017 (2010s) is considered as our reference.

3. MLR model validation

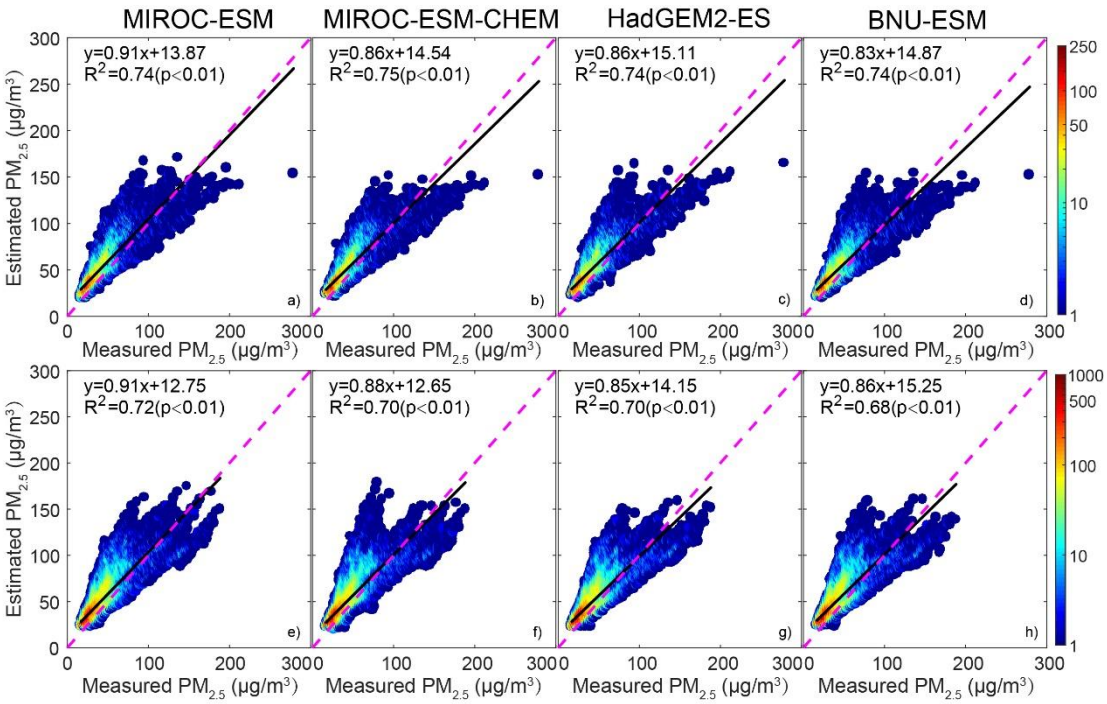


Figure 1. Scatter grams of $PM_{2.5}$ concentration derived by MODIS and estimated by MLR during validation period (2016-2017). Top figures (a-d) are the ISIMIP statistical downscaling results, and bottom figures (e-h) are the WRF dynamical downscaling results. R2 means the variance explained by the MLR, and color bar denotes the density of datapoints at integer intervals.

Figure 1 shows the scattergram of $PM_{2.5}$ concentration between ChinaHighPM2.5 dataset and MLR model during validation period based on ISIMIP and WRF results. Observations and MLR models have Pearson's correlations coefficients around 0.86 for ISIMIP results during the validating period, and the coefficient of determination of MLRs are 0.74-0.75 (Fig. 1a-d). WRF Pearson's correlations are slightly lower, 0.82-0.85, and explained variance ranges from 0.68-0.72 (Fig. 1e-h). These results are similar as found by Jin et al. (2022). We also compare the spatial patterns of observed and modeled $PM_{2.5}$ in Fig. S3. Both ISIMIP and WRF results can simulate the distribution characteristics of high concentration of $PM_{2.5}$ in the southeast and low concentration in the northwest.

4. Relative risks of mortality related $PM_{2.5}$

We estimate the effects of $PM_{2.5}$ on mortality by considering changes in the relative risk (RR) of mortality related to $PM_{2.5}$. We lack data on mortality rates in the study domain without which we cannot estimate numbers of fatalities, just the average population-weighted RR. Burnett el al. (2014) established the integrated exposure-response functions we use. The RR is non-linear in concentration, that is an initially low $PM_{2.5}$ region will suffer higher mortality and RR than an initially high $PM_{2.5}$ region if $PM_{2.5}$ is increased by the same amount. Ran et al. (2023) provide RR values for $PM_{2.5}$ concentrations up to 200 $\mu g/m^3$ that includes the 5 main major disease endpoints (Global Burden of Disease Collaborative Network, 2013) of $PM_{2.5}$ related mortality: chronic obstructive pulmonary disease, ischemic heart disease, lung cancer, lung

respiratory infection and stroke. We calculate the average population-weighted relative risks based on the gridded population dataset (Section 2.3) and $PM_{2.5}$ concentration in the Beijing-Tianjin province defined in the Fig. 1c-1d, following Ran et al. (2023):

$$RR_{pop,k}=\frac{\sum_{g=1}^{G} POP_g \times RR_k(C_g)}{\sum_{g=1}^{G} POP_g} \quad (2)$$

$RR_{pop,k}$ is the average population-weighted relative risk of disease k (k=1-5), POP_g is the population of gird g, and $RR_k(C_g)$ is the relative risk of disease k when $PM_{2.5}$ concentration is C_g in the grid of g.

4. Projection

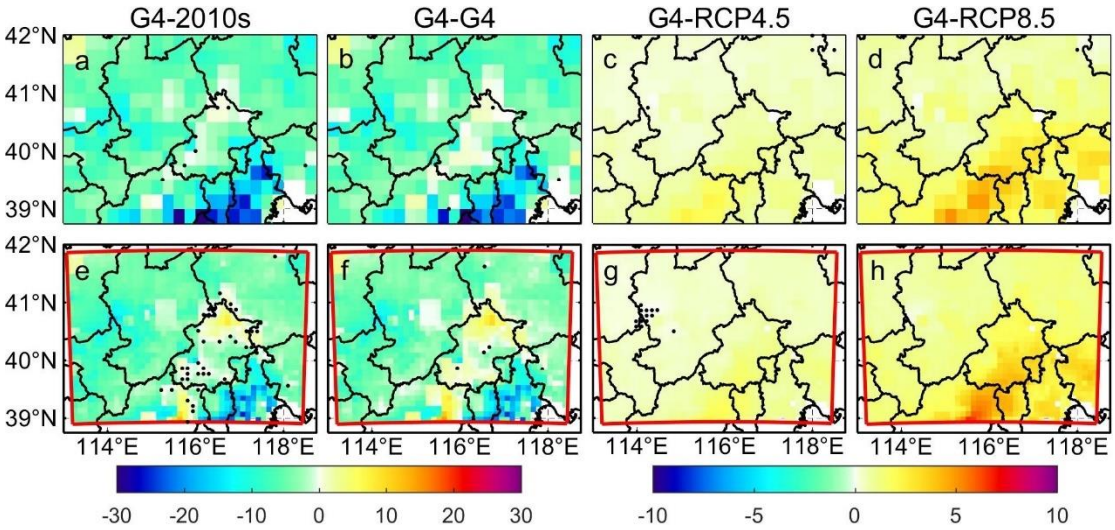


Figure 2. Spatial patterns of ensemble mean $PM_{2.5}$ concentration difference (μg/m^3) between "mitigation" under G4 in the 2060s and reference **(a, e)**, between "mitigation" and "baseline" under G4 in the 2060s **(b, f)**, between G4 and RCP4.5 under "mitigation" scenario in the 2060s **(c, g)**, and between G4 and RCP8.5 under "mitigation" scenario in the 2060s **(d, h)** based on ISIMIP **(a-d)** and WRF **(e-h)** results. Stippling indicates grid points where differences or changes are insignificant at the 5% significant level according to the Wilcoxon signed rank test.

We firstly project the change of $PM_{2.5}$ under G4 and the aerosol mitigation scenario in 2060s relative to 2010s (Fig. 2a, e). Both ISIMIP and WRF project $PM_{2.5}$ decreases in most areas, especially in Tianjin and Langfang, but $PM_{2.5}$ decreases more under ISIMIP than WRF. $PM_{2.5}$ concentration decreases by 6.5 μg/m^3 over Beijing-Tianjin province in ISIMIP, and decrease by 4.3 μg/m^3 in WRF (Table S1). $PM_{2.5}$ concentration is 0.5-8 μg/m^3 higher in northern Beijing under G4 ("mitigation") than that during the 2010s in WRF. To show the impact of emission reductions, we compare the $PM_{2.5}$ concentration between aerosol "baseline" and "mitigation" scenarios under G4 in the 2060s (Fig. 2b, 2f), and compare the "mitigation" $PM_{2.5}$ concentration under G4 and the RCP scenarios in the 2060s to clarify the effect of geoengineering compared with climate warming. Compared with "baseline" scenario, $PM_{2.5}$ concentration is less under "mitigation" scenario as expected in both ISIMIP and WRF under G4 (Fig. 2b, 2f), and has a similar spatial pattern with that in Fig. 2a and 2e. Compared with RCP4.5 and RCP8.5, $PM_{2.5}$ concentrations under G4 are higher in ISIMIP results (Fig. 2c-2d), but with large

differences between the 4 ESMs. G4 $PM_{2.5}$ is simulated greater than in RCP scenarios under HadGEM2-ES and BNU-ESM (Fig. S4k, l, o, p), but there are insignificant differences in most areas under the two MIROC models (Fig. S4c, d, g, h). $PM_{2.5}$ concentrations are larger between G4 and RCP8.5. WRF simulations shows similar changes in $PM_{2.5}$ between G4 and RCPs as ISIMIP (Fig. 2g-h).

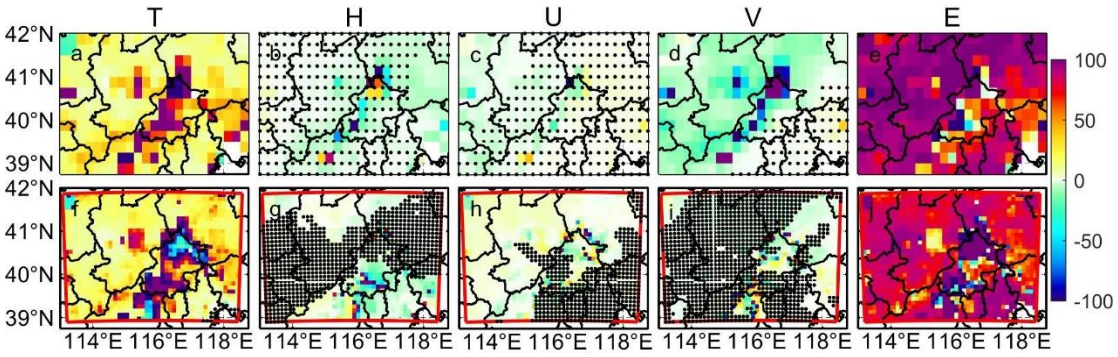


Figure 3. Contribution of climate factors (temperature/T, humidity/H, zonal wind/U, meridional wind/V) and emission (E) to changes in monthly $PM_{2.5}$ concentration (Δ $PM_{2.5}$) in 2060s under G4 ("mitigation") relative to 2010s. Top figures **(a-e)** are ISIMIP results, and bottom figures **(f-j)** are WRF results. Stippling indicates the changes are insignificant at the 5% significant level in the Wilcoxon test.

Next, we quantify the contribution of different meteorological factors and $PM_{2.5}$ emissions to $\Delta PM_{2.5}$ between G4 ("mitigation") in the 2060s and the 2010s (Fig. 3). Both ISIMIP and WRF results show that the increase of temperature and decrease of $PM_{2.5}$ emission play positive roles in reducing $PM_{2.5}$ concentration. ISIMIP results (Fig. 3a-e), suggest that the projected increase of temperature could explain 0-20% of the decrease of $PM_{2.5}$ concentration, and decrease of $PM_{2.5}$ emission could explain more than 90% of changes in $PM_{2.5}$ concentration differences in most of areas. Changes in humidity and westerly winds (positive U-wind) do not cause significant changes in $\Delta PM_{2.5}$, but projected increases southerly wind (positive V-wind) is detrimental to the decrease in $PM_{2.5}$ concentration, and has a 0-10% negative effect on $\Delta PM_{2.5}$ in Zhangjiakou. WRF results show similar spatial pattern in effect of temperature and emission on $\Delta PM_{2.5}$ with ISIMIP results. Although temperature is projected to increase over the whole domain (Fig. S7), there are negative contributions on $\Delta PM_{2.5}$ to the north of Beijing due to increase of $PM_{2.5}$ caused by the negative correlation between $PM_{2.5}$ and its emissions (Fig. S11). The ~1-2% wetter humidity has ~10% negative effect on decrease of $PM_{2.5}$ south of Beijing (Fig. 3g), and 0.2-0.3 m/s deceases of U-wind have 0-10% negative contribution on decrease of $PM_{2.5}$ in Zhangjiakou (Fig. 3h). The changes in each factor in ISIMIP and WRF results are shown in Fig. S6 and Fig. S7, respectively.

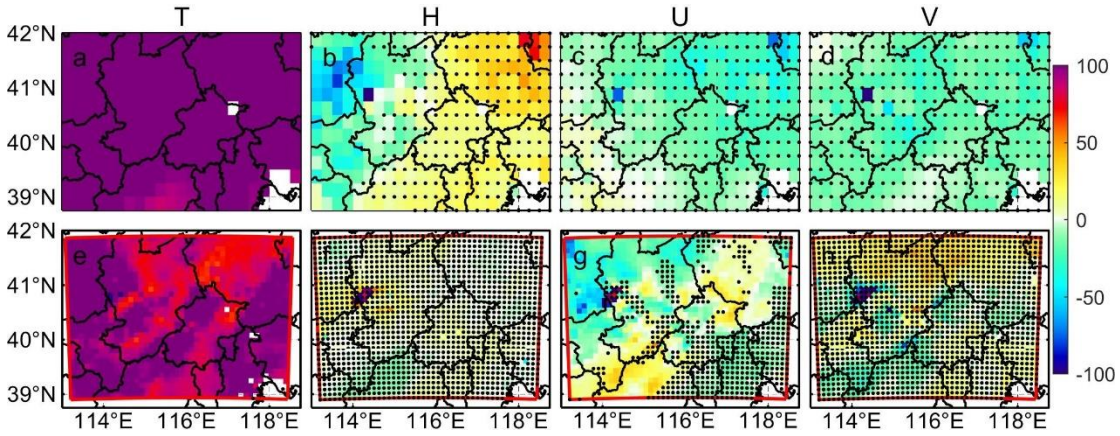


Figure 4. Contribution of climate factors (as in Fig. 3) to changes in monthly $PM_{2.5}$ concentration in 2060s under G4 with aerosol "mitigation" relative to 2060s under RCP4.5 with aerosol "mitigation". Top figures **(a-e)** are ISIMIP results, and bottom figures **(f-j)** are WRF results. Stippling indicates the changes are insignificant at the 5% significant level in the Wilcoxon test.

Now we explore the contribution of each meteorological factor to $\Delta PM_{2.5}$ between G4 ("mitigation") and RCP4.5 ("mitigation") in the 2060s (Fig. 4). The higher $PM_{2.5}$ under G4 is mainly caused by the lower temperature. In ISIMIP, lower temperature explains more than 90% (100% in some places) of the raised $PM_{2.5}$ relative to RCP4.5, although the increase of humidity is also helpful to lower $PM_{2.5}$ in the western domain (Fig. 4a-b). Humidity can increase suspended particle mass and coagulation, promoting deposition (Li et al., 2015). The contribution of differences in U-wind and V-wind on $\Delta PM_{2.5}$ is insignificant (Fig. 4c-d). In WRF, the projected lower temperatures explain more than 70% of the higher $PM_{2.5}$ under G4 relative to RCP4.5 (Fig. 4e). Although the increase of southerly (V) wind contributes 10-20% to the higher $PM_{2.5}$ in the northern domain under HadGEM2-ES and BNU-ESM (Fig. S9), it is insignificant in the ensemble (Fig. 4h). Decreased westerlies (U wind) explains about between +20% and -20% of $PM_{2.5}$ differences (Fig. 4g), since U-wind impacts vary spatially (Fig. S11).

Changes in RR of $PM_{2.5}$ for the 5 diseases under the geoengineering and global warming climate scenarios and different emission scenarios during 2060s relative to 2010s for the Beijing-Tianjin province are shown in Fig. 5. Present-day $PM_{2.5}$ related RRs are 1.32 (1.30), 1.37 (1.35), 1.46 (1.43), 1.83 (1.80) and 2.02 (1.99) for chronic obstructive pulmonary disease (COPD), ischemic heart disease (IHD), lung cancer (LC), lung respiratory infection (LRI) and stroke according to the ISIMIP (WRF) simulations (Fig. 5a). RR of LRI is the highest and COPD is the lowest in the five diseases, and WRF estimates of RR are 0.2-0.3 lower than those of ISIMIP. In both the "baseline" and "mitigation" emission scenarios, RRs will be lower under G4, RCP4.5 and RCP8.5 compared with the 2010s. Smaller RR reductions occur under G4 than under RCP4.5 and RCP8.5, and ISIMIP simulates larger reductions than WRF. This is because the $PM_{2.5}$ concentrations from ISIMIP are reduced more than with WRF (Table S1). Under the "baseline" emission scenario (Fig. 5b-d), the biggest reduction of RR for LRI is 0.047 under RCP8.5 in ISIMIP, and RRs for other diseases are projected to reduce by

no more than 0.02. Under the "mitigation" emission scenario (Fig. 5e-g), reductions in RRs are 3-6 times greater.

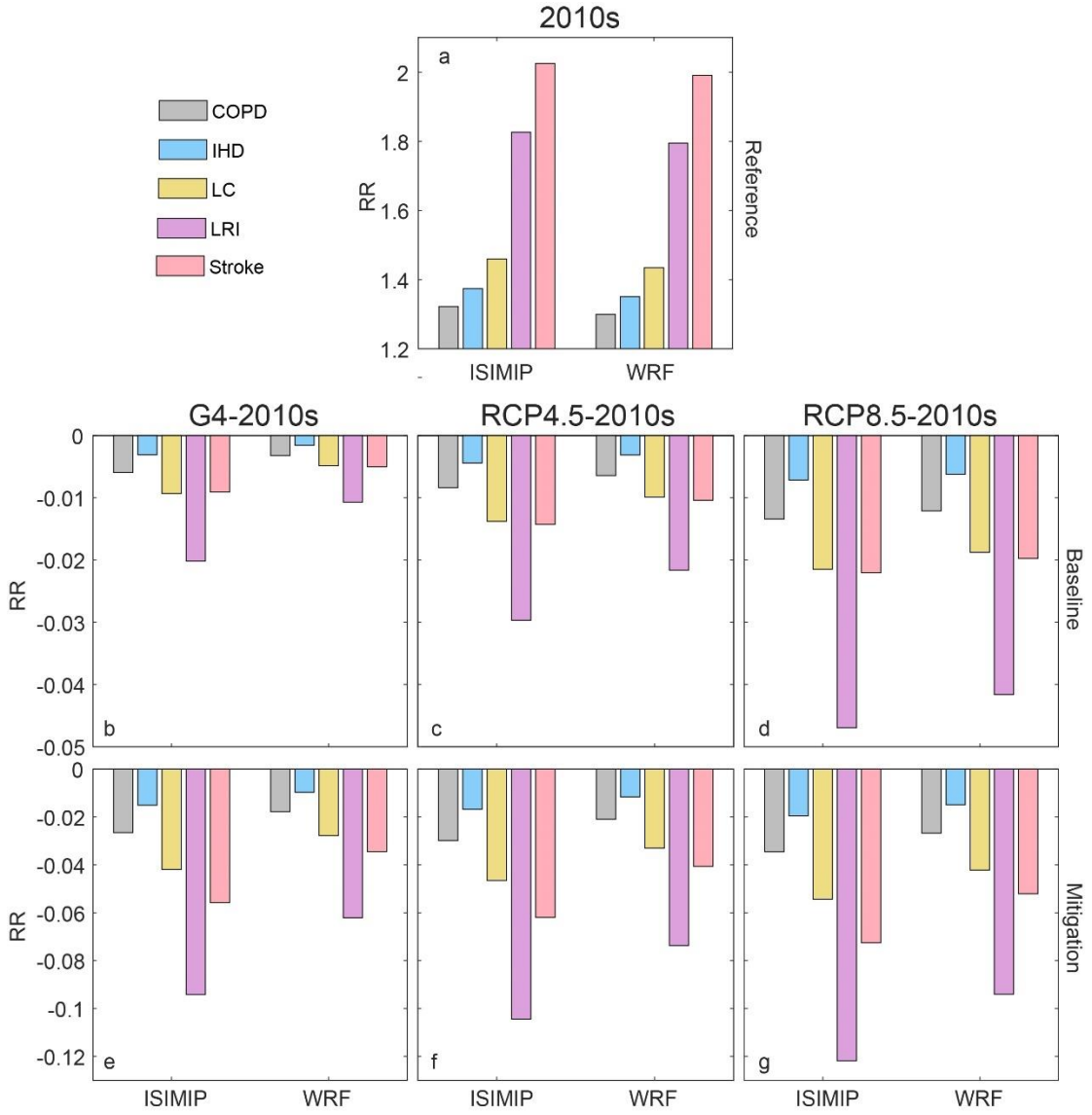


Figure 5. Average population-weighted relative risks of PM2.5 related 5 diseases in 2010s **(a)** and its changes between G4 and 2010s **(b, e)**, between RCP4.5 and 2010s **(c, f)** and between RCP8.5 and 2010s **(d, g)** in Beijing-Tianjin province based on the ISIMIP and WRF results, respectively. $PM_{2.5}$ concentration is based on the "baseline" emissions under G4, RCP4,5 and RCP8.5 in the middle 3 figures **(b-d)**, and it is based on the "mitigation" emissions under G4, RCP4,5 and RCP8.5 in the bottom 3 figures **(e-g)**.

5. Discussion and conclusion

We established a set spatially gridded MLR models based on the 4 ESMs downscaled variables under ISIMIP and WRF. The meteorological factors impact $PM_{2.5}$ in complex ways, but the simple spatially gridded MLR models display enough skill to make some illustrative projections of future $PM_{2.5}$ explaining about 70% of the variance during the historical period. $PM_{2.5}$ concentration is correlated with emissions and anti-correlated

with temperature in most parts of the domain (Fig, S10-S11). Increased turbulence increases diffusion of $PM_{2.5}$ (Yang et al., 2016), and higher temperatures increase evaporation losses (Liu et al., 2015) of ammonium nitrate (Chuang et al., 2017), and other components (Wang et al., 2006). Humidity may have both positive and negative effects on $PM_{2.5}$ (Chen et al., 2020). It causes more water vapor to adhere to the surface of $PM_{2.5}$, thereby increasing its mass concentration and facilitating aerosol growth (Cheng et al., 2017; Liao et al., 2017). However, when the humidity exceeds a certain threshold, coagulation and particle mass increases rapidly, promoting deposition (Li et al., 2015). So, the slope coefficients between $PM_{2.5}$ and humidity are positive in low humidity areas, including southern plain and the Beijing-Tianjin province, but negative in some north mountain areas (Fig. S10, S11).

There are large spatial differences in wind speed and direction impacts on $PM_{2.5}$. Yang et al. (2016) found that weaker northerly and westerly winds tend to increase the $PM_{2.5}$ concentration in northern and eastern China, respectively. The effects of wind direction depend on the distribution of emitted $PM_{2.5}$ and the condition of the underlying surface (Chen et al., 2020). Most sources of $PM_{2.5}$ lie to the south of our domain, relatively clean conditions prevail to the north, so northly winds tend to advect clean air, while southerlies bring high concentrations of aerosols. Weak winds tend to increase $PM_{2.5}$ and smog formation due to sinking air and weak diffusion (Su et al., 2017; Yang et al., 2017).

Emissions reductions are expected to play the dominant role in the decrease of $PM_{2.5}$ concentrations under G4 aerosol "mitigation" in 2060s (Fig. 3). Meteorological changes under the different future scenarios make much smaller changes as evidenced by the scenarios using "baseline" – that is present day $PM_{2.5}$ emissions, with decreases in mean annual concentration of 1.0 (1.3), 1.8 (2.0), 3.3 (3.2) $\mu g/m^3$ over Beijing-Tianjin province under G4, RCP4.5 and RCP8.5 with WRF (ISIMIP), (Table S1), which are mainly caused by the temperature increases (Fig. 4). The negative relationships between emission and $PM_{2.5}$ concentration result in the increase of $PM_{2.5}$ under G4 ("mitigation") relative to 2010s in the north of Beijing with WRF. This may be due to changes in $PM_{2.5}$ out of the domain being opposite to those in domain during the MLR fitting period, since relocation of polluting sources from the urban areas mainly to the west, was occurring over the calibration period. The accuracy of $PM_{2.5}$ emission data is also crucial for training MLR models, and $PM_{2.5}$ data was sparse before 2013, relying on reconstructions based on satellite optical depth estimates. Although both increase of temperature and decrease of emission explain more than 90% of the decrease in $PM_{2.5}$ in most areas, there are large spatial differences due to wind and humidity. On the one hand, there is uncertainty in the differences in changes of wind speed and humidity between different ESMs and downscaling methods; on the other hand, the complex physical relationship between them and $PM_{2.5}$ also increases uncertainties. Reductions in PM2.5 in the future are projected to decrease $PM_{2.5}$ related health issues, although its effect on different diseases are different. Changes in $PM_{2.5}$ related risk between G4 and RCPs are from 1-3%, with $PM_{2.5}$ emissions policy dominating differences over

climate scenario.

Eastham et al. (2018) deduced from experiments using 1 Tg/yr SAI in a coupled chemistry-transport model directly simulating atmospheric chemistry, transport, radiative transfer of UV, emissions, and loss processes, that per unit mass emitted, surface-level emissions of sulfate result in 25 times greater population exposure to $PM_{2.5}$ than emitting the same aerosol into the stratosphere. The G4 experiment specifies 5 Tg/yr injection rate, which over our domain would equate to 1450 t/yr if it was deposited uniformly globally (which it certainly would not be). Reducing this by the 1/25 factor amounts to 58 t/yr which can be compared with present $PM_{2.5}$ emissions of around 3.3×10^5 t/year in our domain. If we consider the aerosol deposition under G4 scenarios, $PM_{2.5}$ concentration will be 0-1 $\mu g/m^3$ higher than that without due to deposition of the SAI aerosols (Fig. S12), and RR is projected to increase by 0.01% for Beijing-Tianjin province (Table S2). This comparison suggests that tropospheric emissions will be much more important for human health in our domain than from the SAI specified by G4.

The most important change in $PM_{2.5}$ will come from emissions reductions, with the different weather conditions under both G4 and RCP scenarios making relatively little practical differences in concentrations. $PM_{2.5}$ concentration is expected to decrease significantly (ISIMIP: -6.5$\mu g/m^3$, WRF: -4.3 $\mu g/m^3$) in the Beijing-Tianjin province, but they will still not meet either Chinese or international standards. The temperature under G4 is lower than that under RCP4.5 and RCP8.5 scenarios, which makes the $PM_{2.5}$ concentration under G4 higher. But the difference in $PM_{2.5}$ between the two is small and even within uncertainty due to projected differences in humidity and wind. Potentially improved estimates from more complex models such as WRF-Chem, CMAQ and GEOS-Chem over the simple MLR methods used here will be of limited value unless the differences between the ESM driving these models is reduced. It can be confirmed that emission policies based on the 13th Five Year Plan are not enough, and higher emission standards need to be developed for a healthy living environment.

Supplement

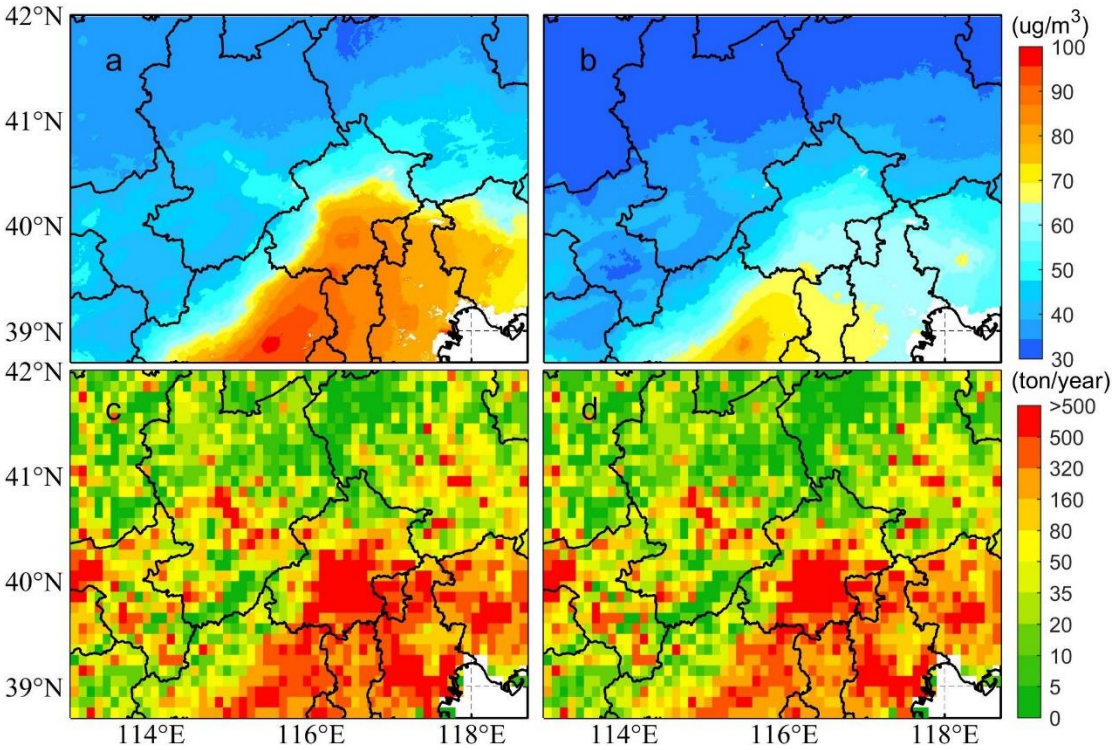


Figure S1. Annual mean $PM_{2.5}$ concentration (a, b) and $PM_{2.5}$ emissions (c, d) map for Beijing and surrounding areas during 2008 (a, c) and 2017 (b, d).

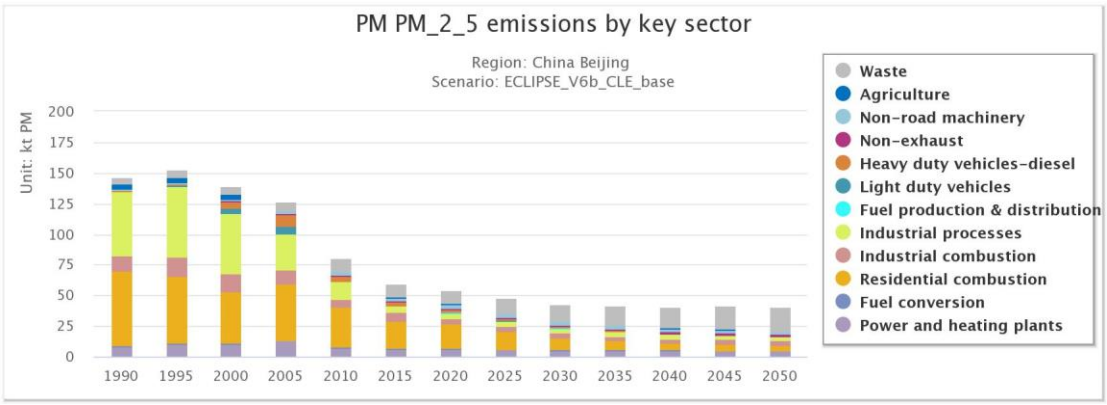


Figure S2. Annual $PM_{2.5}$ emissions from different sources in Beijing under the ECLIPSE V6b baseline scenario *(Source: GAINS East Asia online (iiasa.ac.at))*.

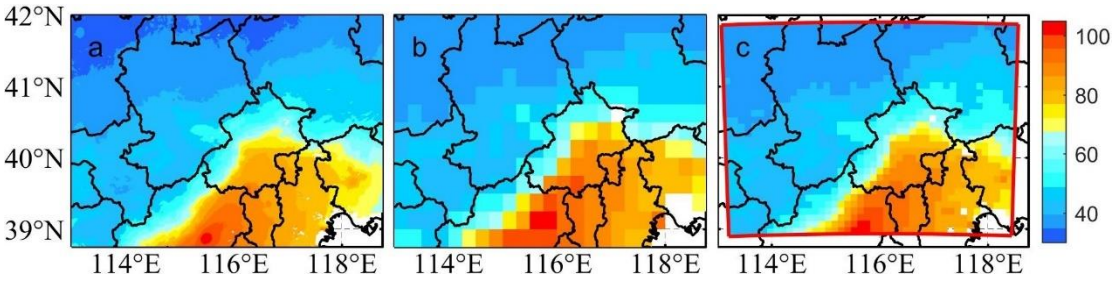


Figure S3. Distribution of observed $PM_{2.5}$ concentration (ug/m^3) from ChinaHighPM$_{2.5}$ **(a)** and estimated ensemble-mean $PM_{2.5}$ concentration from MLR under ISIMIP **(b)** and WRF **(c)** results for Beijing and surrounding areas during 2008-2017.

Table S1. Difference of $PM_{2.5}$ concentration between different scenarios for the Beijing-Tianjin province as defined in Fig. 1b during 2060-2069. The $PM_{2.5}$ emission scenarios used in each climate scenarios are in parentheses. Bold indicates the differences or changes are significant at the 5% significant level according to the Wilcoxon signed rank test. (Units: μg/m^3)

Model	G4 (mitigation) -2010s (reference)		G4 (mitigation) -G4 (baseline)		G4 (mitigation) -RCP4.5(mitigation)		G4 (mitigation) -RCP8.5(mitigation)	
	WRF	ISIMIP	WRF	ISIMIP	WRF	ISIMIP	WRF	ISIMIP
MIROC-ESM	**-4.5**	**-6.3**	**-3.1**	**-3.8**	0.5	0.7	**2.3**	**2.3**
MIROC-ESM-CHEM	**-6.0**	**-7.4**	**-4.9**	**-5.3**	0.5	-0.2	**1.9**	0.6
HadGEM2-ES	**-4.8**	**-6.8**	**-3.8**	**-6.8**	**1.4**	**1.3**	**2.6**	**2.6**
BNU-ESM	**-2.5**	**-5.5**	**-1.4**	**-5.0**	**0.8**	**1.1**	**2.4**	**2.2**
Ensemble	**-4.3**	**-6.5**	**-3.3**	**-5.2**	**0.8**	**0.7**	**2.3**	**1.9**

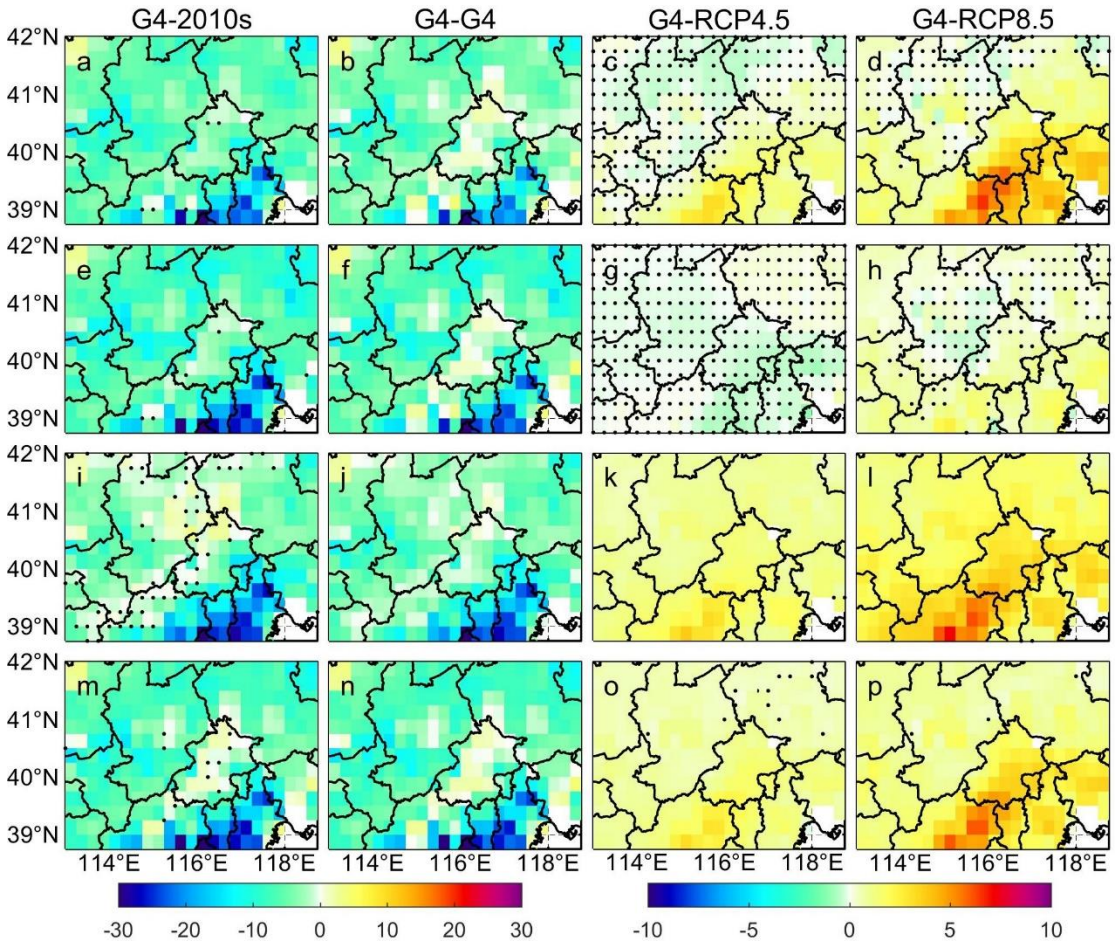


Figure S4. Spatial patterns of $PM_{2.5}$ concentration difference (μg/m^3) between "mitigation" in the 2060s under G4 and 2010s **(a, e, i, m)**, between "mitigation" and "baseline" under G4 **(b, f, j, n)**, between G4 and RCP4.5 under "mitigation" scenario **(c, g, k, o)**, and between G4 and RCP8.5 under "mitigation" scenario **(d, h, l ,p)** based on ISIMIP results. From top to bottom are MIROC-ESM **(a-d)**, MIROC-ESM-CHEM **(e-h)**, HadGEM2-ES **(i-l)** and BNU-ESM **(m-p)** respectively. Stippling indicates grid points where differences or changes are not significant at the 5% significant level according to the Wilcoxon signed rank test.

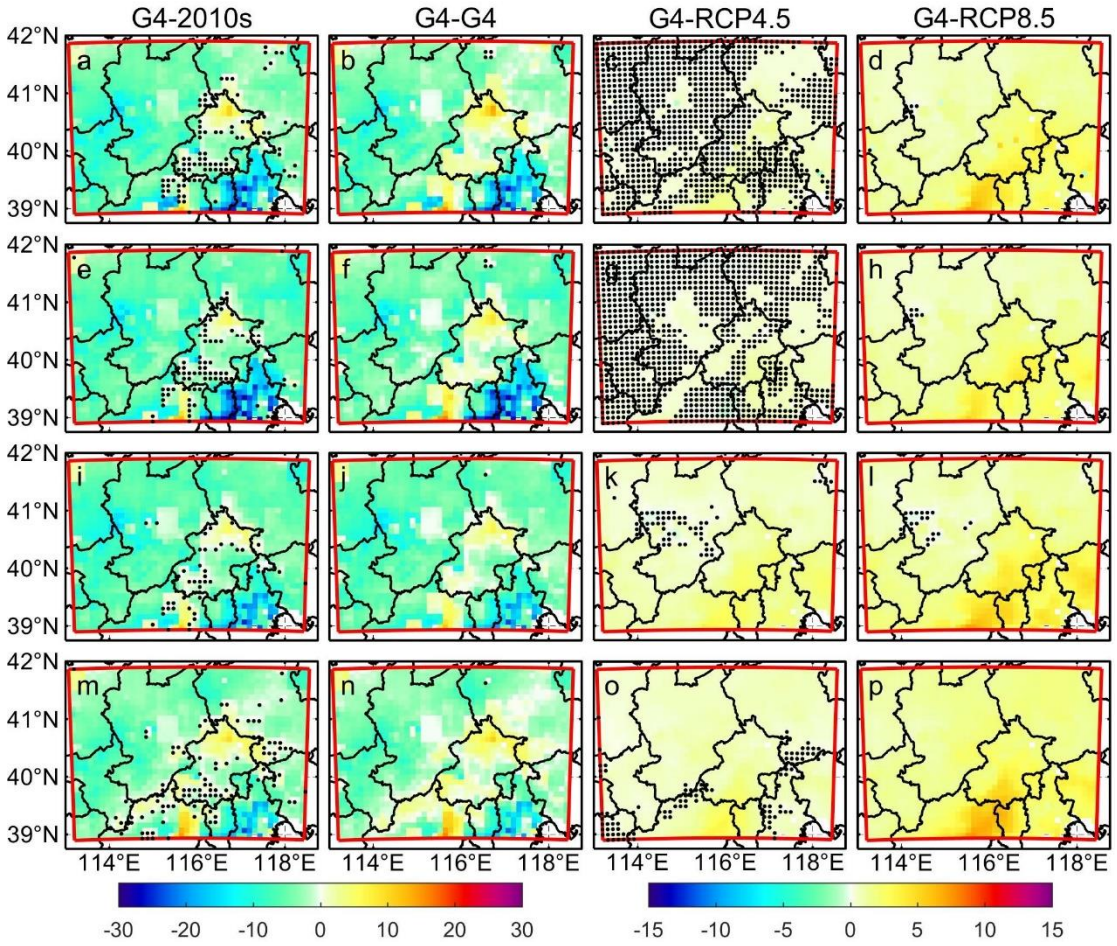


Figure S5. Same as Fig. S4, but by WRF.

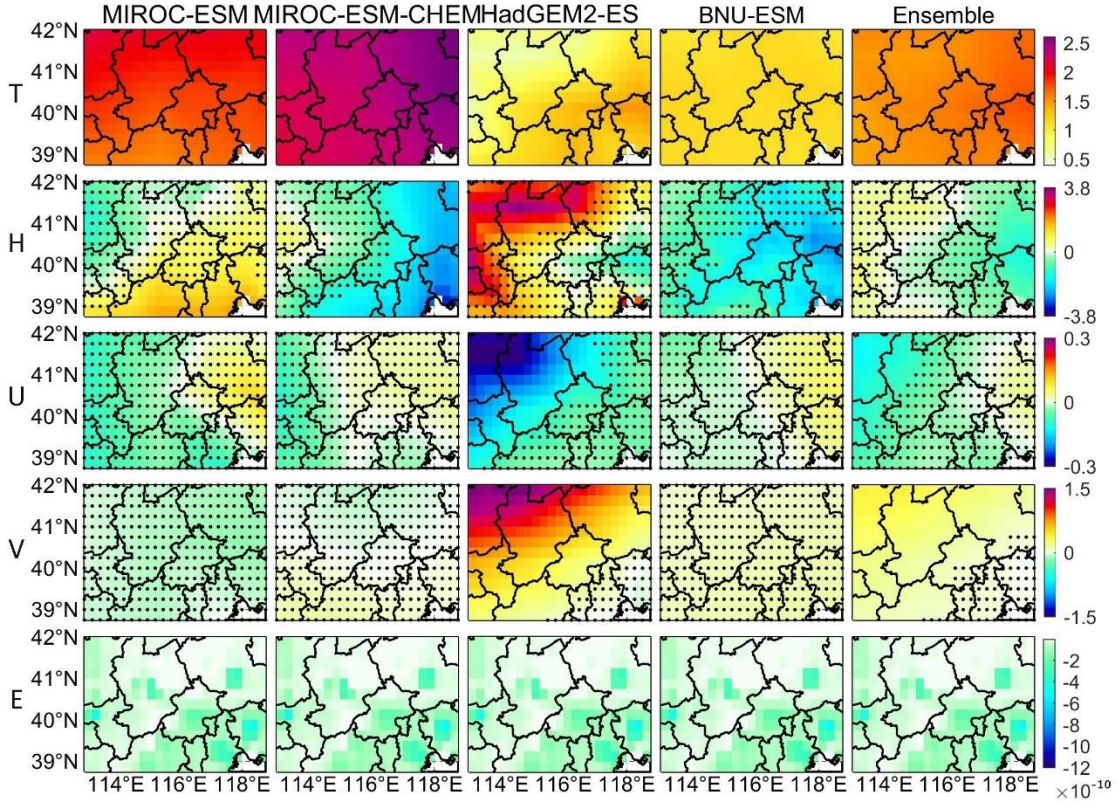


Figure S6. Spatial pattern of changes in temperature (T/°C), humidity (H/%), zonal wind (U/m s^{-1}), meridional wind (V/m s^{-1}) and $PM_{2.5}$ emissions (E/kg m^{-2} s^{-1}) under G4 ("mitigation") in the 2060s relative to 2010s in ISIMIP. Stippling indicates grid points where differences or changes are not significant at the 5% significant level according to the Wilcoxon signed rank test.

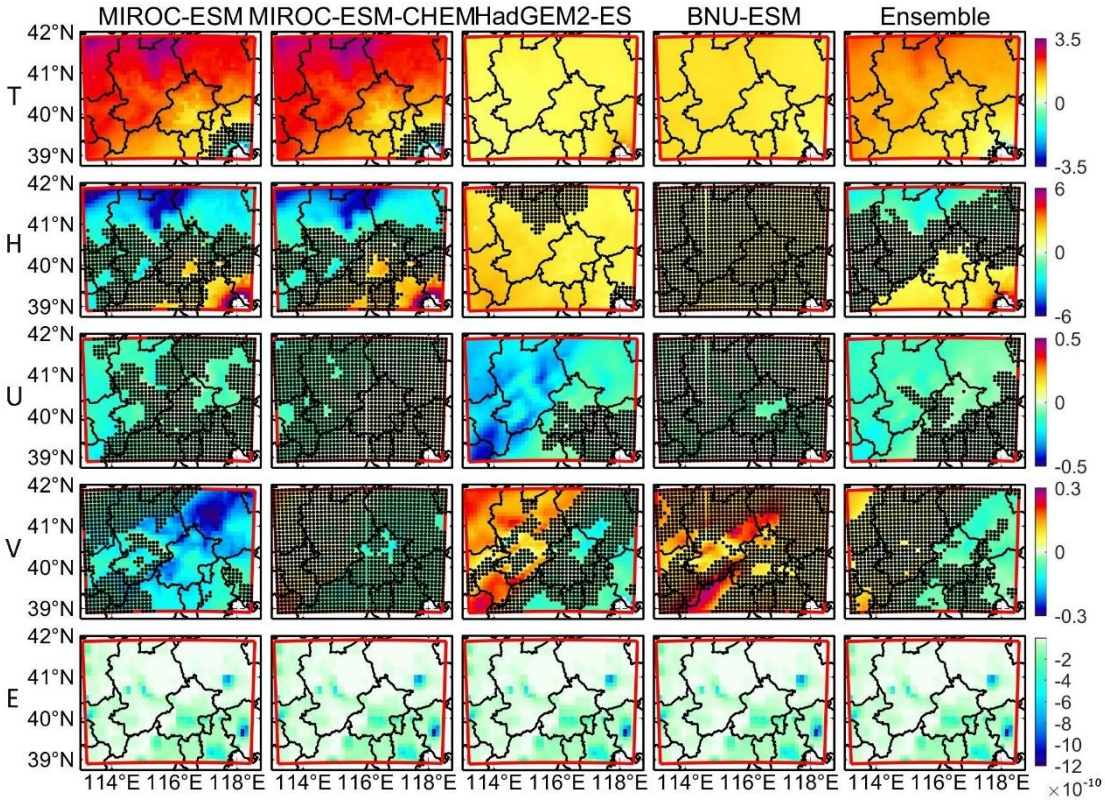


Figure S7. Same as Fig. S6, but by WRF.

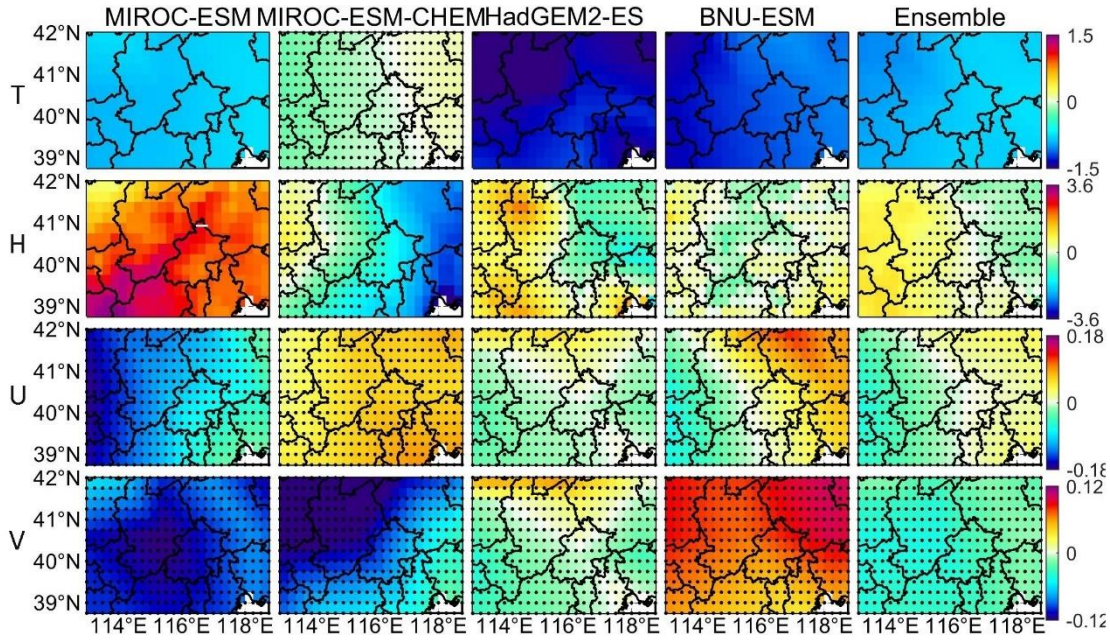


Figure S8. Spatial pattern of changes in temperature (T/°C), humidity (H/%), zonal wind (U/m s^{-1}) and

meridional wind (V/m s^{-1}) under G4 ("mitigation") relative to RCP4.5 ("mitigation") in the 2060s in ISIMIP. Stippling indicates grid points where differences or changes are not significant at the 5% significant level according to the Wilcoxon signed rank test.

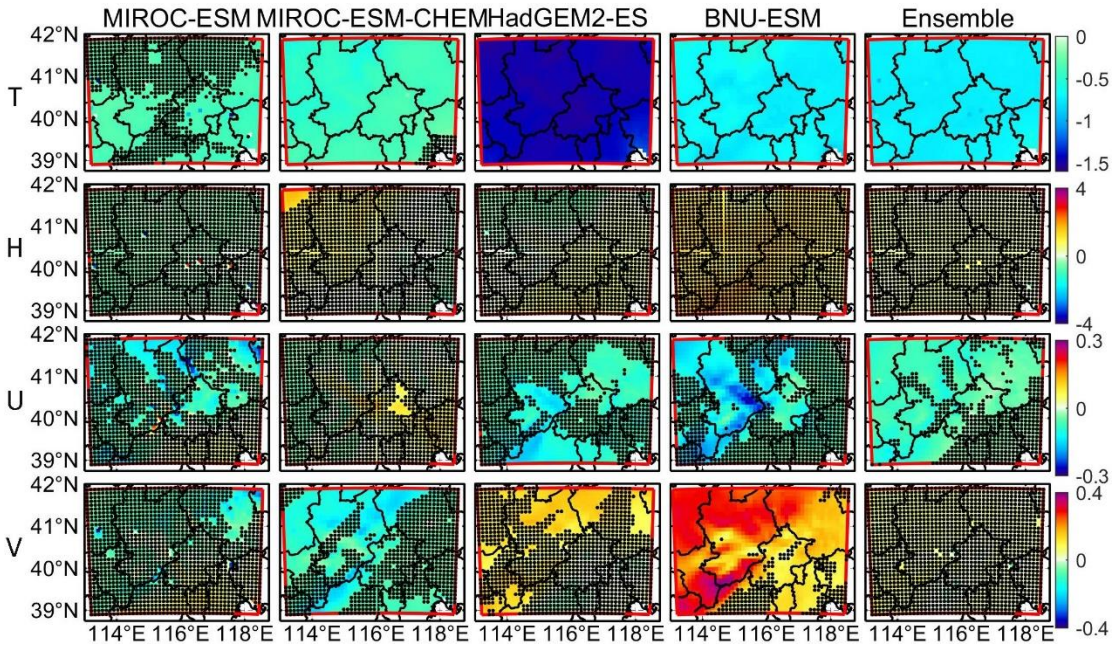


Figure S9. Same as Fig. S8, but for WRF results.

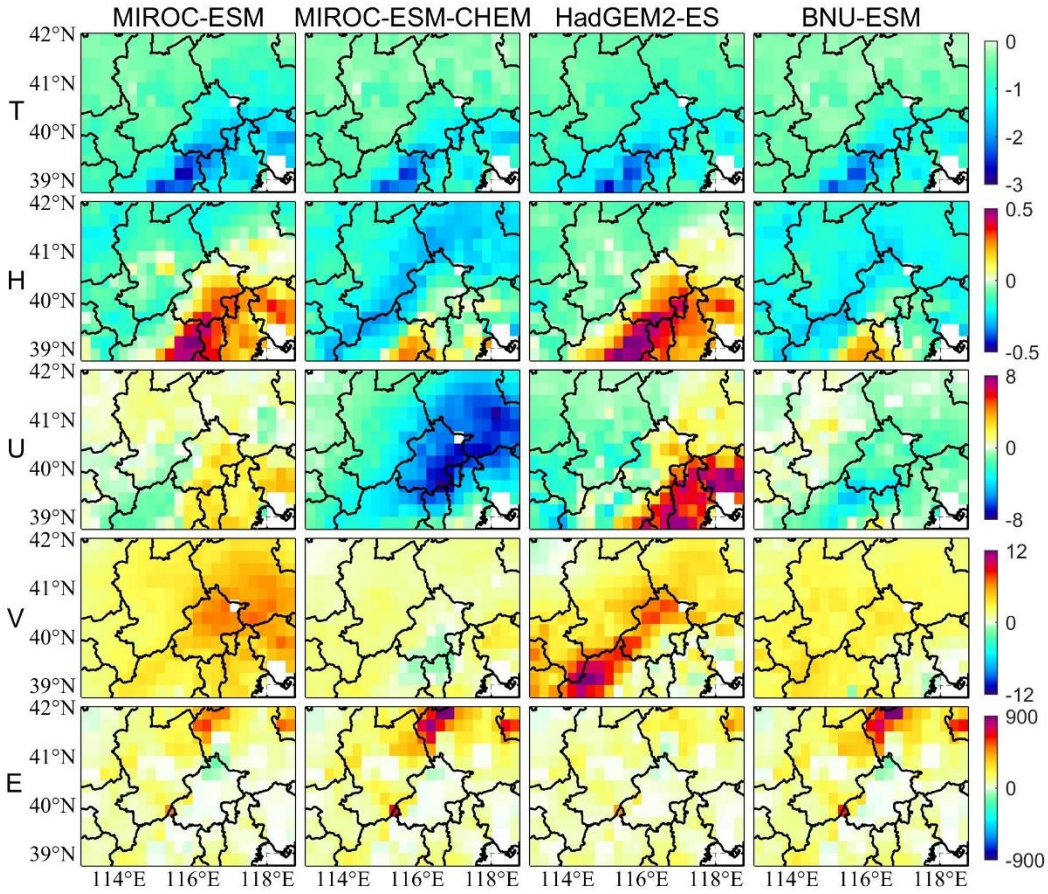


Figure S10. Slope coefficients of MLR of temperature, humidity, u-wind, v-wind and emission for ISIMIP results during training period.

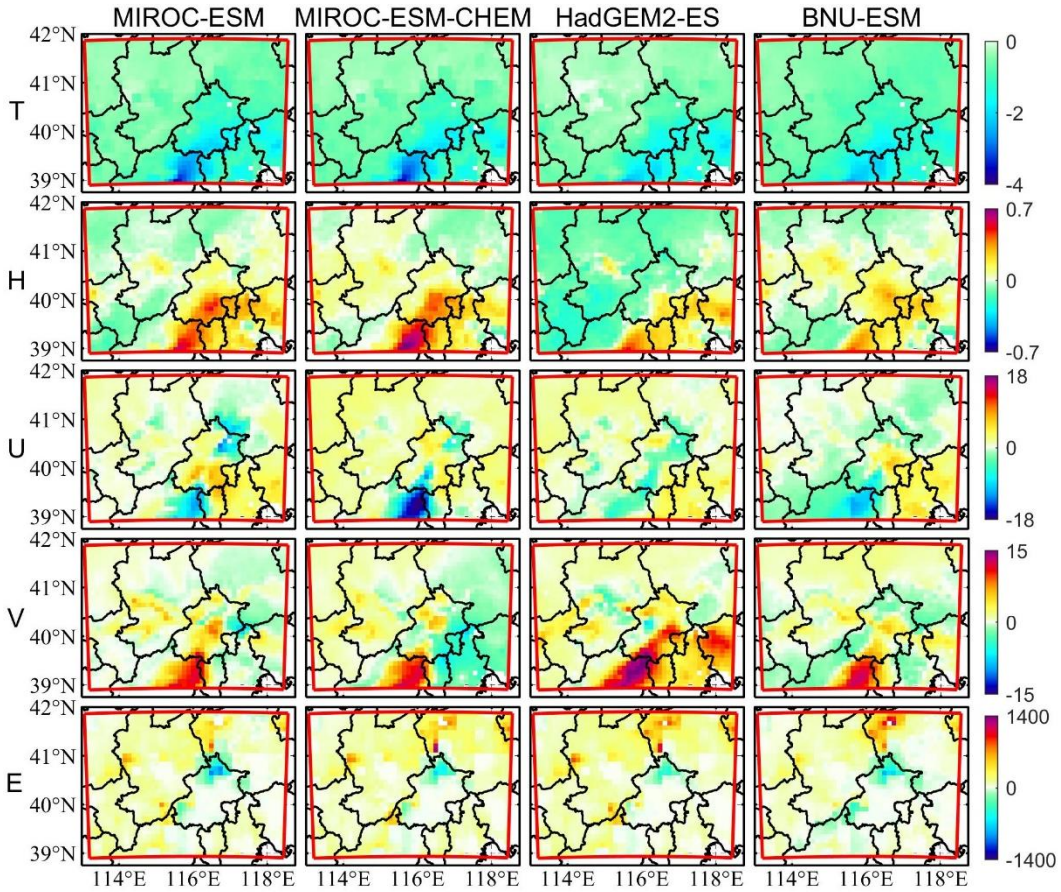


Figure S11. Similar as Fig. S10, but for WRF results.

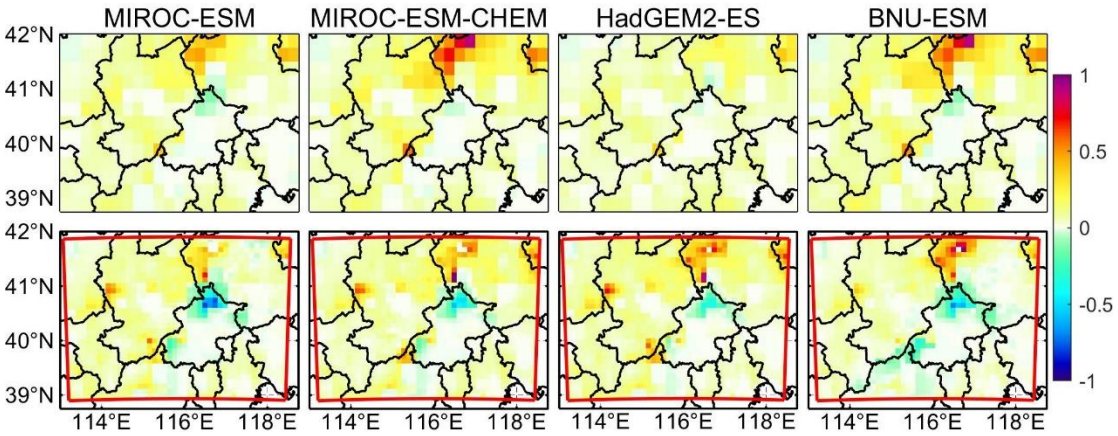


Figure S12. Spatial pattern of changes in $PM_{2.5}$ ($\mu g/m^3$) between G4 with and without considering aerosol deposition due to SAI specified by G4.

Table S2. RRs of the 5 mortality endpoints under G4 with and without considering aerosol deposition from the G4 SAI specification in both $PM_{2.5}$ aerosol "baseline" and "mitigation" scenarios.

G4			population-weighted RR				
			COPD	IHD	LC	LRI	Stroke
"baseline"	No deposition	ISIMIP	1.3166	1.3710	1.4505	1.8063	2.0161
		WRF	1.2968	1.3490	1.4299	1.7844	1.9857
	deposition	ISIMIP	1.3167	1.3711	1.4506	1.8064	2.0162
		WRF	1.2968	1.3490	1.4299	1.7845	1.9857
"mitigation"	No deposition	ISIMIP	1.2961	1.3590	1.4179	1.7323	1.9695
		WRF	1.2823	1.3408	1.4069	1.7331	1.9562
	deposition	ISIMIP	1.2961	1.3590	1.4180	1.7326	1.9696
		WRF	1.2823	1.3408	1.4070	1.7332	1.9563

Reference

Burnett, R., Pope III, C., Ezzati, M., Olives, C., Lim, S., Mehta, S., Shin, H., Singh, G., Hubbell, B., Brauer, M., Anderson, A., Smith, K., Balmes, J., Bruce, N., Kan, H., Laden, F., Prüss-Ustün, A., Turner, M., Gapstur, S., Diver, W., and Cohen, A.: An Integrated Risk Function for Estimating the Global Burden of Disease Attributable to Ambient Fine Particulate Matter Exposure, Environ., Health Perspect., 122, 397-403, https://doi.org/10.1289/ehp.1307049, 2014.

Chen, Z., Chen, D., Kwan, M.-P., Chen, B., Gao, B., Zhuang, Y., Li, R., and Xu, B.: The control of anthropogenic emissions contributed to 80 % of the decrease in $PM_{2.5}$ concentrations in Beijing from 2013 to 2017, Atmos. Chem. Phys., 19, 13519–13533, https://doi.org/10.5194/acp-19-13519-2019, 2019.

Chen, Z., Chen, D., Zhao, C., Kwan, M., Cai, J., Zhuang, Y., Zhao, B., Wang, X., Chen, B., Yang, J., Li, R., He, B., Gao, B., Wang, K., and Xu, B.: Influence of meteorological conditions on $PM_{2.5}$ concentrations across China: A review of methodology and mechanism, Environ. Int., 139, 105558, https://doi.org/10.1016/j.envint.2020.105558, 2020.

Chen, Z., Xie, X., Cai, J., Chen, D., Gao, B., He, B., Cheng, N., and Xu, B.: Understanding meteorological influences on $PM_{2.5}$ concentrations across China: a temporal and spatial perspective, Atmos. Chem. Phys., 18, 5343–5358, https://doi.org/10.5194/acp-18-5343-2018, 2018.

Cheng, L., Meng, F., Chen, L., Jiang, T., and Su, L.: Effects on the haze pollution from autumn crop residue burning over the Jing-Jin-Ji Region, China Environ. Sci., 37, 2801–2812, 2017.

Chuang, M., Chou, C., Lin, N., Takami, A., Hsiao, T., Lin, T., Fu, J., Pani, S., Lu, Y., and Yang, T.: A simulation study on $PM_{2.5}$ sources and meteorological characteristics at the northern tip of Taiwan in the early stage of the Asian haze period, Aerosol Air Qual. Res., 17, 3166-3178, https://doi.org/10.4209/aaqr.2017.05.0185, 2017.

Eastham, D., Weisenstein, D., Keith, D., and Barrett, A.: Quantifying the impact of sulfate geoengineering on mortality from air quality and UV-B exposure, Atmos. Environ., 187, 424–434. DOI: http://dx.doi.org/10.1016/j.atmosenv.2018.05.047, 2018.

Fan, M., Zhang, Y., Lin, Y., Cao, F., Sun, Y., Qiu, Y., Xing, G., Dao, X., and Fu, P.: Specific sources of health risks induced by metallic elements in $PM_{2.5}$ during the wintertime in Beijing, China, Atmos. Environ., 246, 118112, https://doi.org/10.1016/j.atmosenv.2020.118112, 2021.

Guan, W., Zheng, X., Chung, K., and Zhong, N.: Impact of air pollution on the burden of chronic respiratory diseases in China: time for urgent action, Lancet, 388, 1939-1951, https://doi.org/10.1016/S0140-6736(16)31597-5, 2016.

Han, J., Wang, J., Zhao, Y., Wang, Q., Zhang, B., Li, H., and Zhai, J.: Spatio-temporal variation of potential evapotranspiration and climatic drivers in the Jing-Jin-Ji region, North China, Agric. For. Meteorol., 256, 75-83, https://doi.org/10.1016/j.agrformet.2018.03.002, 2018.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Dentener, F., Muntean, M., Pouliot, G., Keating, T., Zhang, Q., Kurokawa, J., Wankmüller, R., Denier van der Gon, H., Kuenen, J. J. P., Klimont, Z., Frost, G., Darras, S., Koffi, B., and Li, M.: HTAP_v2.2: a mosaic of regional and global emission grid maps for 2008 and 2010 to study hemispheric transport of air pollution, Atmos. Chem. Phys., 15, 11411-11432, https://doi.org/10.5194/acp-15-11411-2015, 2015.

Jin, H., Chen, X., Zhong, R., and Liu, M.: Influence and prediction of $PM_{2.5}$ through multiple environmental variables in China, Sci. Total Enviro., 849, 157910, https://doi.org/10.1016/j.scitotenv.2022.157910, 2022.

Klimont, Z., Kupiainen, K., Heyes, C., Purohit, P., Cofala, J., Rafaj, P., Borken-Kleefeld, J., and Schöpp, W.: Global anthropogenic emissions of particulate matter including black carbon, Atmos. Chem. Phys., 17, 8681–8723, https://doi.org/10.5194/acp-17-8681-2017, 2017.

Li, D., Wu, Q., Feng, J., Wang, Y., Wang, L., Xu, Q., Sun, Y., Cao, K., and Cheng, H.: The influence of anthropogenic emissions on air quality in Beijing-Tianjin-Hebei of China around 2050 under the future climate scenario, J. Cleaner Prod., 388, 135927, https://doi.org/10.1016/j.jclepro.2023.135927, 2023.

Li, J., Chen, H., Li, Z., Wang, P., Cribb, M., and Fan, X.: Low-level temperature inversions and their effect on aerosol condensation nuclei concentrations under different large-scale synoptic circulations, Adv. Atmos. Sci., 32, 898-908, https://doi.org/10.1007/s00376-014-4150-z, 2015.

Li, K., Liao, H., Zhu, J., and Moch, J.: Implications of RCP emissions on future $PM_{2.5}$ air quality and direct radiative forcing over China, J. Geophys. Res. Atmos., 121, 12, 985-13, 008, https://doi:10.1002/2016JD025623, 2016.

Li, M., Klimont, Z., Zhang, Q., Martin, R. V., Zheng, B., Heyes, C., Cofala, J., Zhang, Y., and He, K.: Comparison and evaluation of anthropogenic emissions of SO_2 and NO_x over China, Atmos. Chem. Phys., 18, 3433–3456, https://doi.org/10.5194/acp-18-3433-2018, 2018.

Liao, T., Wang, S., Ai, J., Gui, K., Duan, B., Zhao, Q., Zhang, X., Jiang, W., and Sun, Y.: Heavy pollution

episodes, transport pathways and potential sources of $PM_{2.5}$ during the winter of 2013 in Chengdu (China), Sci. Total Environ., 584–585, 1056–1065, https://doi.org/10.1016/j.scitotenv.2017.01.160, 2017.

Lin, G., Fu, J., Jiang, D., Wang, J., Wang, Q., and Dong, D.: Spatial variation of the relationship between $PM_{2.5}$ concentrations and meteorological parameters in China, BioMed Res. Int., 2015, 684618, https://doi.org/10.1155/2015/684618, 2015.

Maji, K., Ye, W., Arora, M., and Nagendra, S.: $PM_{2.5}$-related health and economic loss assessment for 338 Chinese cities, Environ. Int., 121, 392-403, https://doi.org/10.1016/j.envint.2018.09.024, 2018.

Mishra, D., Goyal, P., and Upadhyay, A.: Artificial intelligence based approach to forecast $PM_{2.5}$ during haze episodes: a case study of Delhi, India, Atmos. Environ., 102, 239–248, https://doi.org/10.1016/j.atmosenv.2014.11.050, 2015.

Nguyen, G., Shimadera, H., Uranishi, K., Matsuo, T., and Kondo, A.: Numerical assessment of $PM_{2.5}$ and O_3 air quality in Continental Southeast Asia: Impacts of future projected anthropogenic emission change and its impacts in combination with potential future climate change impacts, Atmos. Environ., 226, 117398, https://doi.org/10.1016/j.atmosenv.2020.117398, 2020.

Ran, Q., Lee, S., Zheng, D., Chen, H., Yang, S., Moore, J., and Dong, W.: Potential health and economic impacts of shifting manufacturing from China to Indonesia or India, Sci. Total Enviro., 855, 158634, https://doi.org/10.1016/j.scitotenv.2022.158634, 2023.

Stohl, A., Aamaas, B., Amann, M., Baker, L. H., Bellouin, N., Berntsen, T. K., Boucher, O., Cherian, R., Collins, W., Daskalakis, N., Dusinska, M., Eckhardt, S., Fuglestvedt, J. S., Harju, M., Heyes, C., Hodnebrog, Ø., Hao, J., Im, U., Kanakidou, M., Klimont, Z., Kupiainen, K., Law, K. S., Lund, M. T., Maas, R., MacIntosh, C. R., Myhre, G., Myriokefalitakis, S., Olivié, D., Quaas, J., Quennehen, B., Raut, J.-C., Rumbold, S. T., Samset, B. H., Schulz, M., Seland, Ø., Shine, K. P., Skeie, R. B., Wang, S., Yttri, K. E., and Zhu, T.: Evaluating the climate and air quality impacts of short-lived pollutants, Atmos. Chem. Phys., 15, 10529–10566, https://doi.org/10.5194/acp-15-10529-2015, 2015.

Su, J., Brauer, M., Ainslie, B., Steyn, D., Larson, T., and Buzzelli, M.: An innovative land use regression model incorporating meteorology for exposure analysis, Sci. Total Environ., 390, 520-529, https://doi.org/10.1016/j.scitotenv.2007.10.032, 2008.

Tong, C., Yim, S., Rothenberg, D., Wang, C., Lin, C., Chen, Y., and Lau, N.: Projecting the impacts of atmospheric conditions under climate change on air quality over the Pearl River Delta region, Atmos. Environ., 193, 79-87, https://doi.org/10.1016/j.atmosenv.2018.08.053, 2018.

Upadhyay, A., Dey, S., Goyal, P., and Dash, S.: Projection of near-future anthropogenic $PM_{2.5}$ over India using statistical approach, Atmos. Environ., 186, 178-188, https://doi.org/10.1016/j.atmosenv.2018.05.025, 2018.

Wang, J., Zhang, L., Niu, X., and Liu, Z.: Effects of $PM_{2.5}$ on health and economic loss: Evidence from

Beijing-Tianjin-Hebei region of China, J. Cleaner Prod., 257, 120605, https://doi.org/10.1016/j.jclepro.2020.120605, 2020.

Wang, Y., Yao, L., Wang, L., Liu, Z., Ji, D., Tang, G., Zhang, J., Sun, Y., Hu, N., and Xin, J.: Mechanism for the formation of the January 2013 heavy haze pollution episode over central and eastern China, Sci. China Earth Sci., 57, 14-25, https://doi.org/10.1007/s11430-013-4773-4, 2014.

Wang, Y., Zhuang, G., Zhang, X., Huang, K., Xu, C., Tang, A., Chen, J., and An, Z.: The ion chemistry, seasonal cycle, and sources of $PM_{2.5}$ and TSP aerosol in Shanghai, Atmos. Environ., 40, 2935-2952, https://doi.org/10.1016/j.atmosenv.2005.12.051, 2006.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., and Cribb, M.: Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: spatiotemporal variations and policy implications, Remote Sens. Environ., 252, 112136, https://doi.org/10.1016/j.rse.2020.112136, 2021.

Xue, W., Zhang, J., Zhong, C., Li, X., and Wei, J.: Spatiotemporal $PM_{2.5}$ variations and its response to the industrial structure from 2000 to 2018 in the Beijing-Tianjin-Hebei region, J. Cleaner Prod., 279, 123742, https://doi.org/10.1016/j.jclepro.2020.123742, 2021.
Yang, S., Ma, Y., Duan, F., He, K., Wang, L., Wei, Z., Zhu, L., Ma, T., Li, H., Ye, S.: Characteristics and formation of typical winter haze in Handan, one of the most polluted cities in China, Sci. Total Environ., 613-614, 1367-1375, https://doi.org/10.1016/j.scitotenv.2017.08.033, 2018.

Yang, X., Zhao, C., Guo, J., and Wang, Y.: Intensification of aerosol pollution associated with its feedback with surface solar radiation and winds in Beijing, J. Geophys. Res. Atmos., 121, 4093-4099, https://doi.org/10.1002/2015JD024645, 2016.

Zhang, Q., Zheng, Y., Tong, D., Shao, M., Wang, S., Zhang, Y., Xu, X., Wang, J., He, H., Liu, W., Ding, Y., Lei, Y., Li, J., Wang, Z., Zhang, X., Wang, Y., Cheng, J., Liu, Y., Shi, Q., Yan, L., Geng, G., Hong, C., Li, M., Liu, F., Zheng, B., Cao, J., Ding, A., Gao, J., Fu, Q., Huo, J., Liu, B., Liu, Z., Yang, F., He, K., and Hao, J.: Drivers of improved $PM_{2.5}$ air quality in China from 2013 to 2017, PNAS, 116, 24463-24469, https://doi.org/10.1073/pnas.1907956116, 2019.

Zhang, Z., Gong, D., Mao, R., Kim, S., Xu, J., Zhao, X., and Ma, Z.: Cause and predictability for the severe haze pollution in downtown Beijing in November–December 2015, Sci. Total Enviro., 592, 627-638, https://doi.org/10.1016/j.scitotenv.2017.03.009, 2017.

Zhao, D., Xin, J., Gong, C., Quan, J., Liu, G., Zhao, W., Wang, Y., Liu, Z., and Song, T.: The formation mechanism of air pollution episodes in Beijing city: insights into the measured feedback between aerosol radiative forcing and the atmospheric boundary layer stability, Sci. Total Environ., 692, 371–381, https://doi.org/10.1016/j.scitotenv.2019.07.255, 2019.

My second concern regards the treatment of ERA5 data as an observational reference (L132-133). The paper would be significantly strengthened if the authors instead

compared their simulations of the recent past with monitor data. Even if this monitor data is used in ERA5, showing that the simulations are capable of reproducing truly observational data rather than a reanalysis would provide more convincing evidence of model performance.

Reply： We replaced ERA5 data with CN05.1 data as our observational data for validation. We replaced the sentences in line 131-134 using the following sentences and updated the figures and table.

To validate the downscaled AP from model results, we use the daily temperature, humidity and wind speed during 2008-2017 from the gridded observational dataset CN05.1 with the resolution of 0.25°× 0.25° based on the observational data from more than 2400 surface meteorological stations in China, which are interpolated using the "anomaly approach" (Wu and Gao, 2013). This dataset is widely used, and has good performance relative to other reanalysis datasets over China (Zhou et al., 2016; Yang et al., 2019; Yang et al., 2023; Yang and Tang, 2023).

Reference

Wu, J. and Gao, X.: A gridded daily observation dataset over China region and comparison with the other datasets, Chinese J. Geophys., 56, 1102–1111, https://doi.org/10.6038/cjg20130406, 2013 (in Chinese, data available at: http://climatechange-data.cn/resource/detail?id=228, last access: 8 February 2023).

Zhou, B., Xu, Y., Wu, J., Dong, S., and Shi, Y.: Changes in temperature and precipitation extreme indices over China: analysis of a high-resolution grid dataset, Int. J. Climatol., 36, 1051–1066, https://doi.org/10.1002/joc.4400, 2016.

Yang, Y., Tang, J., Xiong, Z., Wang, S., and Yuan, J.: An intercomparison of multiple statistical downscaling methods for daily precipitation and temperature over China: future climate projections, Clim. Dynam., 52, 6749–6771, https://doi.org/10.1007/s00382-018-4543-2, 2019.

Yang, Y., Maraun, D., Ossó, A., and Tang, J.: Increased spatial extent and likelihood of compound long-duration dry and hot events in China, 1961–2014, Nat. Hazards Earth Syst. Sci., 23, 693–709, https://doi.org/10.5194/nhess-23-693-2023, 2023.

Yang, Y., and Tang, J.: Substantial Differences in Compound Long‐Duration Dry and Hot Events Over China Between Transient and Stabilized Warmer Worlds at 1.5° C Global Warming, Earths Future, 11, e2022EF002994, https://doi.org/10.1029/2022EF002994, 2023.

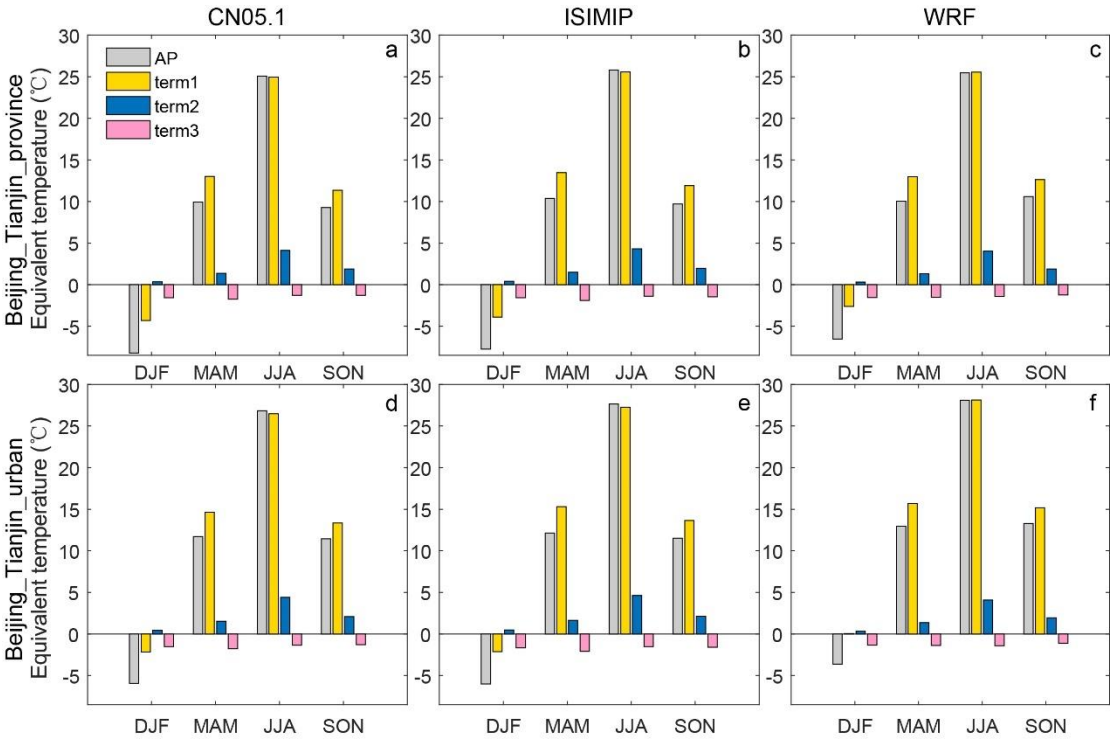


Figure 2. Seasonal averaged AP and equivalent temperature of each term in equation 1 for Beijing-Tianjin province **(a-c)** and Beijing-Tianjin urban areas **(d-f)** during 2008-2017 from CN05.1 **(a, d)**, 4-model ensemble mean after ISIMIP **(b, e)** and ensemble mean after WRF **(c, f)**. Term 1 is 1.04T, term 2 is 2P and term 3 is -0.65W.

Figure 2 shows the seasonally averaged AP and equivalent temperatures caused by temperature, relative humidity and wind speed in Beijing-Tianjin province and Beijing-Tianjin urban areas during 2008-2017. According to the CN05.1 results (Fig. 2a, 2d), AP and the separate 3 terms show similar seasonal patterns over the whole province and just the urban areas. Vapor pressure is higher in summer and wind speed is higher in spring. AP is lower than 2 m temperature in all seasons except summer, and especially lower in winter. AP, temperature, vapor pressure and wind speed are all higher in urban areas than in the surrounding rural region in any season. The ISIMIP results (Fig. 2b, 2e), by design, perfectly reproduce the CN05.1 seasonal characteristics of AP, temperature, vapor pressure and wind speed. WRF shows a similar pattern with that from CN05.1, but for the Beijing-Tianjin province, WRF overestimates both 2 m temperature and AP in winter by 2.1°C and by 1.7°C respectively relative to CN05.1 (Fig. 2c). In the Beijing-Tianjin urban areas, WRF overestimates the temperature and AP relative to CN05.1 in all seasons, especially in winter (Fig. 2f).

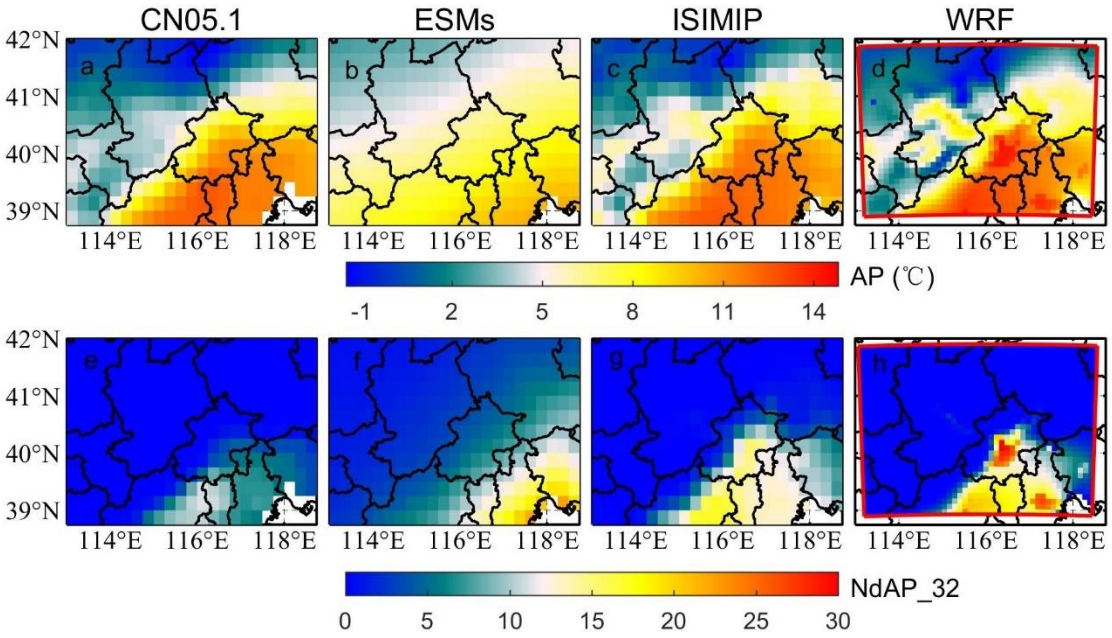


Figure 3. Top row: the spatial distribution of mean apparent temperature from CN05.1 **(a),** raw ESMs ensemble mean after bilinear interpolation **(b)**, 4-model ensemble mean after ISIMIP **(c)** and ensemble mean after WRF **(d)** during 2008-2017. Bottom row: the spatial distribution of annual mean number of days with AP > 32℃ from CN05.1 **(e)**, ESMs **(f)**, ISIMIP **(g)** and WRF **(h)** during 2008-2017. Fig. S1 and Fig. S2 show the pattern of AP and NdAP_32 for the individual ESM.

We compare the simulations of mean apparent temperature and NdAP_32 from both WRF dynamical downscaling with QDM and from ISIMIP statistical downscaling during 2008-2017 in Fig. 4. Both WRF with QDM and ISIMIP methods produce a pattern of apparent temperature which is close to that from CN05.1. While the raw AP from ESMs is overestimated in Zhangjiakou high mountains and underestimated in the southern plain, and shares a similar pattern with temperature from ESMs (Wang et al., 2022). The raw ESM outputs were improved after dynamical and statistical downscaling. The average annual AP from ISIMIP (9.6-9.7℃) is 0.5℃ higher than that from CN05.1 (9.1℃) over the Beijing-Tianjin province for all ESMs (Table 1). While WRF produces warmer apparent temperatures in the city centers of Beijing and Tianjin and lower ones in the high Zhangjiakou mountains than recorded in the lower resolution CN05.1 observations. There are also differences between different models after WRF downscaling. For example, apparent temperatures from the two MIROC models downscaled by WRF are the warmest. In contrast AP from all 4 ESMs after ISIMIP shows very similar patterns (Fig. S1).

ESMs tend to overestimate the number of days with AP>32℃ in southeastern Beijing and the whole Tianjin province. Both ISIMIP and WRF appear to overestimate the NdAP_32 in Beijing urban areas and the southerly lowland areas although NdAP_32 is close to zero in the colder rural areas at relatively high altitude for both downscaling methods. Some of these differences may be due to the WRF simulations being at finer resolution than the 0.25°× 0.25° CN05.1, leading to higher probabilities of high AP in

urban areas (Fig. 4d). ISIMIP results also show slight overestimations, especially in the tails of the distribution (AP>30℃) for urban areas (Fig. 4c). CN05.1 gives about 5 NdAP_32 per year in southern Beijing and Tianjin, but there are nearly 15 NdAP_32 from ISIMIP, and over 20 NdAP_32 per year from WRF downscaling in the Beijing-Tianjin urban areas during 2008-2017. NdAP_32 from WRF and ISIMIP downscaling of all ESM is overestimated relative to CN05.1. But there are differences in ESM under the two downscalings: with ISIMIP, HadGEM2-ES and BNU-ESM have more NdAP_32 than the two MIROC models, while the reverse occurs with WRF (Fig. S2).

Table 1. The annual mean apparent temperature and population weighted NdAP_32 in Beijing-Tianjin province and Beijing-Tianjin urban areas (Fig. 1b) from CN05.1, ISIMIP and WRF during 2008-2017.

Data Sources	AP (℃)				NdAP_32 (day yr^{-1})	
	Provinces		Urban		Population weighted for province (Fig. 1c, 1d)	
	WRF	ISIMIP	WRF	ISIMIP	WRF	ISIMIP
MIROC-ESM	10.5	9.6	13.6	11.4	22.2	10.1
MIROC-ESM-CHEM	10.5	9.6	13.6	11.4	21.9	11.0
HadGEM2-ES	9.5	9.6	12.0	11.4	12.3	11.1
BNU-ESM	9.4	9.7	11.8	11.5	10.2	12.7
CN05.1	9.1		11.1		2.4	

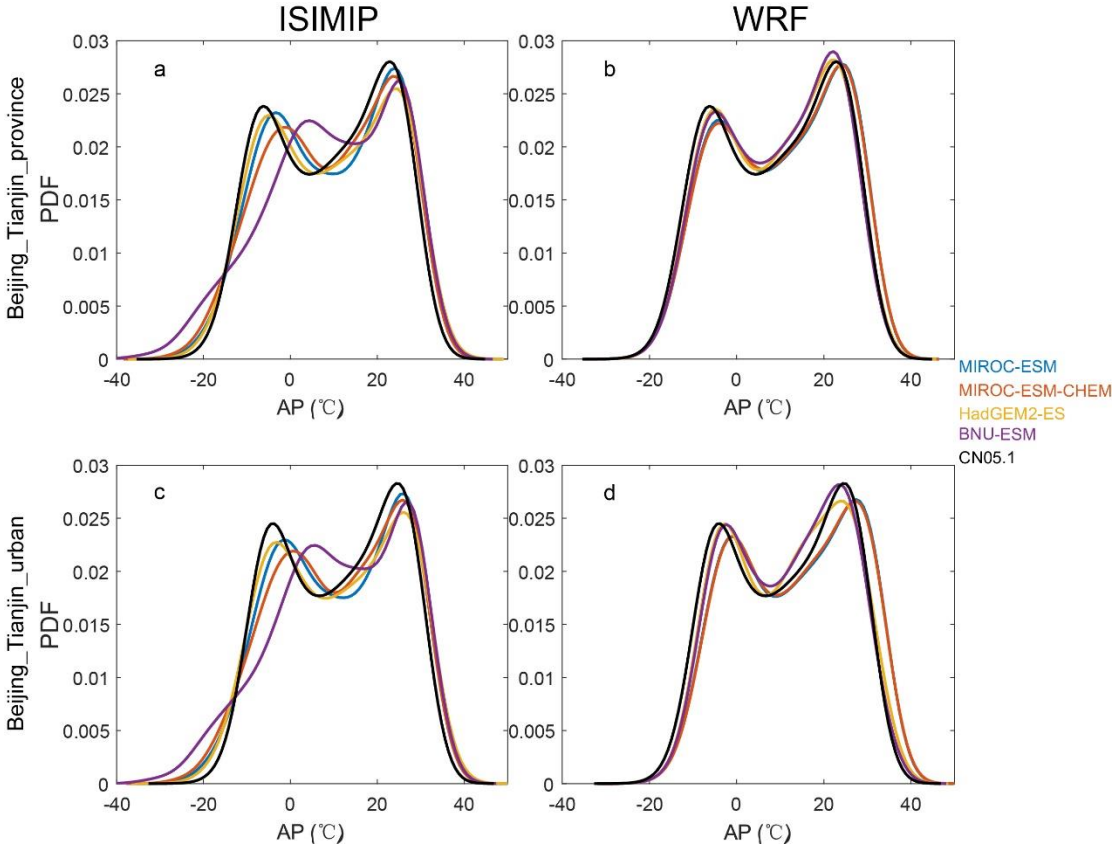


Figure 4. The probability density function (pdf) for daily apparent temperature under ISIMIP **(a, c)** and

WRF **(b, d)** results in Beijing-Tianjin province **(a, b)** and Beijing-Tianjin urban areas **(c, d)** during 2008-2017.

Figure 4 shows the probability density functions (pdf) of daily AP from the four ESMs under ISIMIP and WRF in Beijing-Tianjin province and Beijing-Tianjin urban areas during 2008-2017. ISIMIP overestimates the probability of extreme cold AP relative to CN05.1 (especially BNU-ESM), although all ESM reproduce the CN05.1 pdf well at high AP. WRF can reproduce the CN05.1 distribution of AP better than ISIMIP, but high AP is overestimated relative to CN05.1 and the urban areas perform less well than the whole Beijing-Tianjin province. In urban areas all ESMs driving WRF tend to underestimate the probability of lower AP and to overestimate the probability of higher AP, especially the two MIROC models (Fig. 4d). Fig. S7 displays the annual cycle of monthly AP, with ISIMIP proving excellent by design, at reproducing the monthly AP. While under WRF downscaling AP shows more across model differences, especially during summer and with greater spread for the urban areas.

I also have a methodological concern regarding the method used to try and separate out the roles of different meteorological variables in changes in AP. It is not clear to me why a linear regression is used. The expression for AP is a simple (albeit non-linear) combination of variables, which can be easily and explicitly broken down to find how each one contributes to changes in AP. I suggest the authors at least evaluate how their contributions change if they calculate them based on the degree to which excluding a factor changes AP (i.e. contribution of T to AP is estimated by calculating change in AP with no change in T, but including other factors). The authors could also consider defining the derivatives of AP with respect to each factor, given that these should be well defined (and include the Clausius-Clapeyron relationship directly).

Reply: The reason we used the regression approach is that this produces a least squares estimate of contributions. This is useful in many statistical applications and has desirable mathematical properties compared with, for example, absolute differences. However, it may not be the best choice here as the assumptions of Normality and homoscedasticity in the analysis are probably not true. The referee's suggestions are more localized estimators based around the mean values, which could be regarded as more statistically more robust, as they give less weight to outliers compared with minimizing squared anomalies. But these are reasonable alternatives and the gradient or Jacobian approach plays a role in statistical analyses. We compared the results using the regression method and the first method the referee suggests using. The detailed steps and results are as follows.

To determine the contribution of change in AP (ΔAP) for each meteorological factor under different scenarios, we first calculate the ΔAP caused by individual changes in three factors as follows:

$$\Delta AP(X_i) = \begin{cases} f(T_b, H_a, W_a) - f(T_a, H_a, W_a), i = 1 \\ f(T_a, H_b, W_a) - f(T_a, H_a, W_a), i = 2 \\ f(T_a, H_a, W_b) - f(T_a, H_a, W_a), i = 3 \end{cases} \quad (1)$$

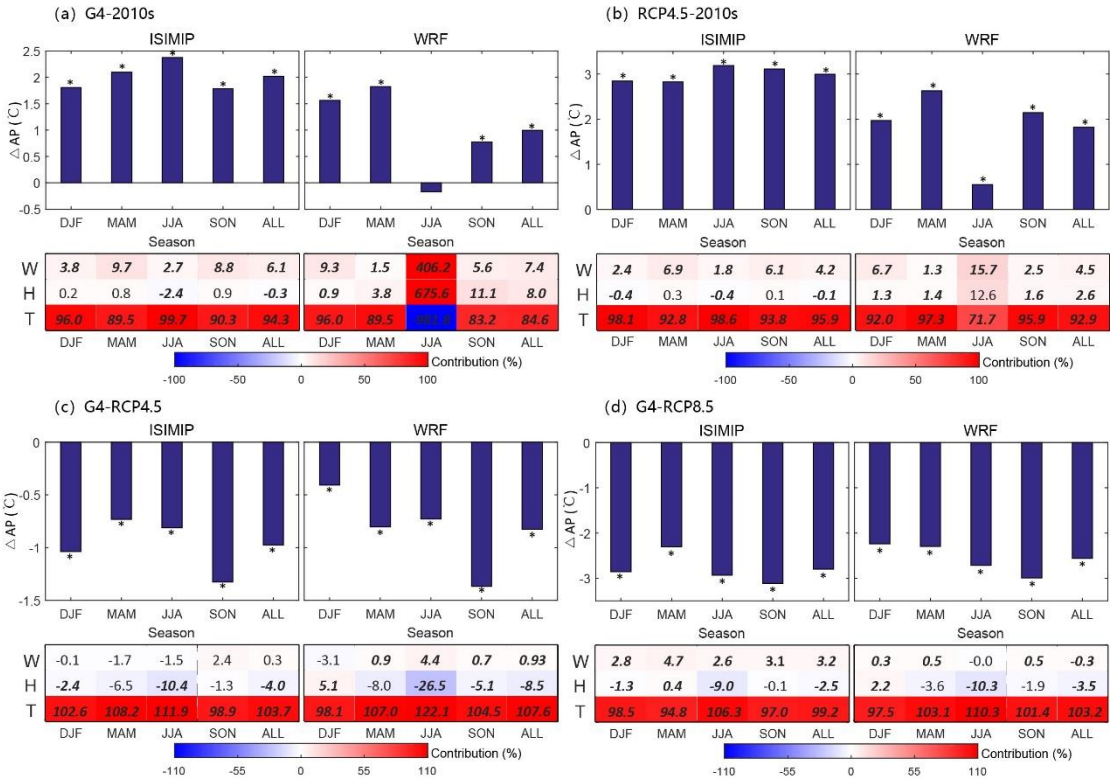
Where daily $\Delta AP(X_i)$ are the ΔAP caused by individual changes in three factors: temperature (X_1), humidity (X_2) and wind speed (X_3). $f()$ is the function of calculating AP. T, H, W are the daily temperature, humidity and wind speed respectively, and the subscripts a and b represent two different climate scenarios, respectively.

Then the contribution of each factor can be expressed as the ratio of the ΔAP caused by one factor alone to the total ΔAP caused by three factors.

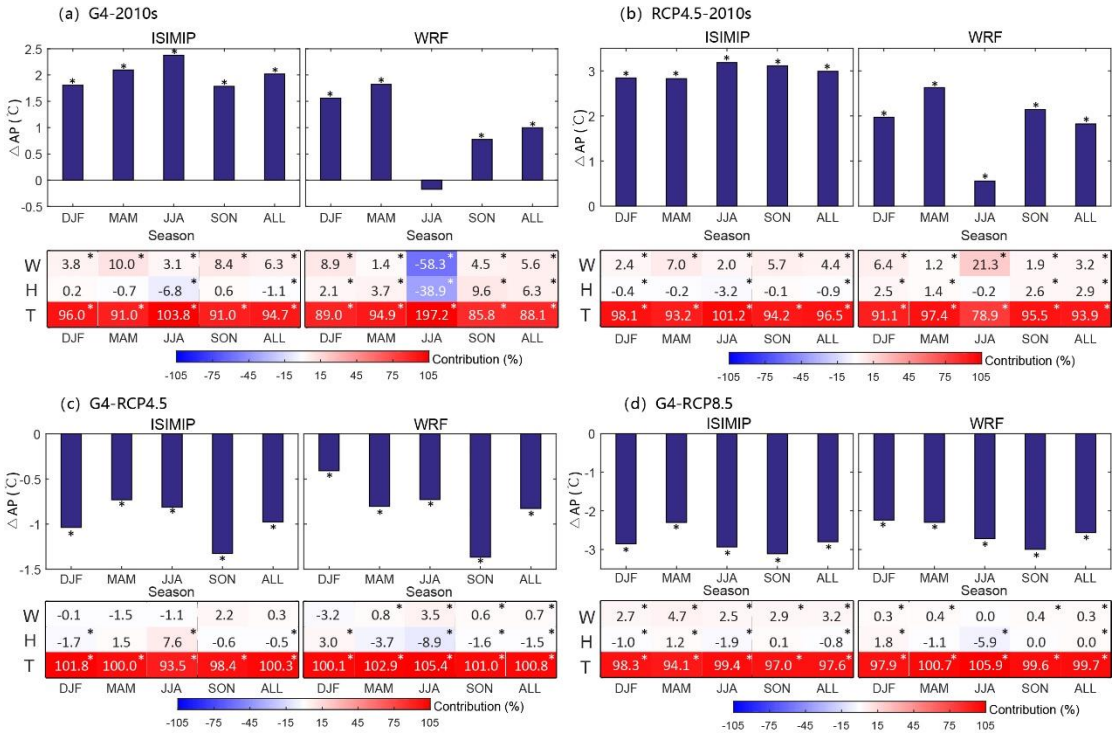
$$C(X_i) = \frac{\overline{\Delta AP}(X_i)}{\sum_{i=1}^{3} \overline{\Delta AP}(X_i)} \qquad (2)$$

Where $C(X_{i(i=1,2,3)})$ is the contributions from changes in each factor to the ΔAP, and $\overline{\Delta AP}(X_i)$ are the mean $\Delta AP(X_i)$. One thing to note is that due to the nonlinear relationship between factors, the total ΔAP caused by three factors is not strictly equal to the ΔAP itself.

We next replotted the figure 6 so we compare it with our previous plot.



Alternative Figure 6. The alternative method of calculating the contributions to seasonal changes of AP (ΔAP) and the seasonal contribution of climatic factors to ΔAP for Beijing and Tianjin urban areas under ISIMIP and WRF between G4 and 2010s **(a)**, G4 and 2010s **(b)**, G4 and RCP4.5 **(c)** and G4 and RCP8.5 **(d)** based on ensemble mean results. Bold italic numbers and "*" above the columns indicate differences are significant at the 95% under the Wilcoxon test.



Preferred original Figure 6. The seasonal changes of AP (ΔAP) and the seasonal contribution of climatic factors to ΔAP for Beijing and Tianjin urban areas under ISIMIP and WRF between G4 and 2010s **(a)**, G4 and 2010s **(b)**, G4 and RCP4.5 **(c)** and G4 and RCP8.5 **(d)** based on ensemble mean results. Colors and numbers in each cell correspond to color bar, and "*" above the columns and in the cells indicate differences are significant at the 95% under the Wilcoxon test.

We calculated the differences of contribution (%) of each meteorology on changes in apparent temperature between alternative suggested method (**Alternative Figure 6**) and preferred original method (**Preferred original Figure 6**), as shown in the figure below.

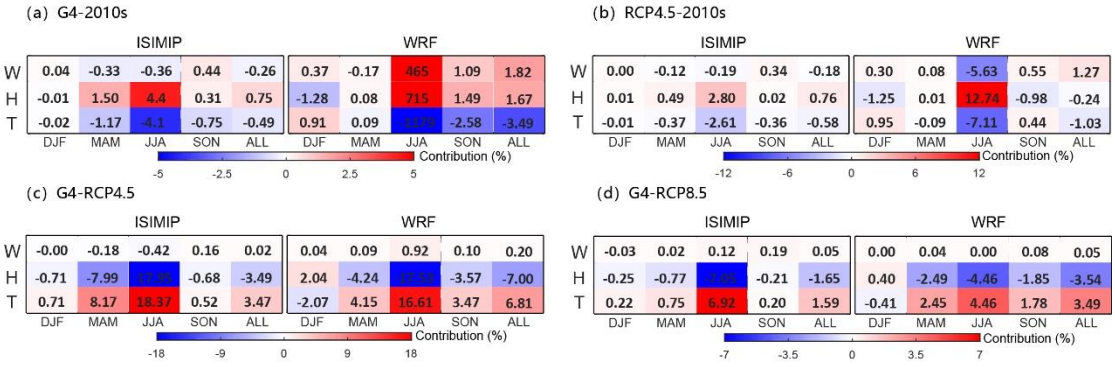


Figure. The differences of contribution (%) of each meteorology on changes in apparent temperature between alternative suggested method and preferred original method.

The contributions of temperature and humidity are different using different methods, but the contribution of wind speed shows little change due to the linear relationship between wind speed and AP under either method. Changes in contribution from humidity is significant. In the referee's suggested method, the contribution of humidity

is influenced by the hybrid effect of temperature, with big changes under higher temperature in JJA. As we all know, AP will change a lot under high temperature, although humidity changes little. In panel a and b, the contribution of humidity under referee's method is higher than that under previous method, but the opposite in the panel c and d. This is because different reference scenarios have different effects when calculating the contribution of humidity. For example, when we calculate the contribution of humidity on AP between G4 and RCP4.5, we can get the value of contribution A (we maintain the temperature in the G4 scenario and the humidity changes with the scenario) and B (we maintain the temperature in the RCP4.5 scenario and the humidity changes with the scenario), but A is not equal to B.

In summary, if we use the suggested method, the sum of changes in AP caused by three factors is not strictly equal to the absolute change in AP and the contribution of humidity and temperature is different when we select different reference between two scenarios. Actually, there is no best way to calculate contributions. Of course there are uncertainties between different methods. We prefer our original method, so we retain it unchanged in our paper.

We changed the sentence in line 178 using the followed sentences

We use an MLR approach, since this minimizes the square differences from the mean across the dataset, with the attendant assumption of independence between the data. Alternatives may also be considered that e.g. minimize the impact of outliers by considering the magnitude of the differences, but we prefer to keep the attractive properties of a least squares approach.

Finally, much of the analysis is rather subjective (e.g. lines 253-257 – "little difference", "slightly worse", "slightly better"). I would recommend that the authors revise the text to make use of quantitative statements, in particular from line 219 onwards. Furthermore, statements such as "There are no models with obvious regional differences" (line 287), "AP changes … are essentially the same" (line 296), "all ESM reproduce the ERA5 pdf well" (line 261), "striking differences" (line 318) and "ERA5… probably does not account for the broad overestimate" (line 234) lack rigor and are difficult to interpret or verify without some context (what counts as a broad overestimate, or an obvious difference? How big would a difference in the change in AP have to be to not count as essentially the same? Why?). A particularly significant example is on line 255, where it is stated that BNU-ESM's performance is "slightly worse" than the other three models when using the ISIMIP method to inspect the recent past. This seems like a significant understatement; BNU-ESM's performance appears to be significantly worse than the other three models (r ~0.85 compared to ~0.92 for the other three), predicting both too many extreme low temperatures and not enough moderately low temperatures (see Figure 4). This is central to the manuscript, since WRF appears to be able to "save" BNU-ESM, bringing its performance to at least be similar to that of HadGEM (albeit still worse than MIROC-ESM[-CHEM]).

Reply: Thanks for your suggestions. we edited them.

Lines 219-241 are edited as follows. It is the same with descriptions in our reply to your second major comments.

We compare the simulations of mean apparent temperature and NdAP_32 from both WRF dynamical downscaling with QDM and from ISIMIP statistical downscaling during 2008-2017 in Fig. 3. Both WRF with QDM and ISIMIP methods produce a pattern of apparent temperature which is close to that from CN05.1. While the raw AP from ESMs is overestimated in Zhangjiakou high mountains and underestimated in the southern plain, and shares a similar pattern with temperature from ESMs (Wang et al., 2022). The raw ESM outputs were improved after dynamical and statistical downscaling. The average annual AP from ISIMIP (9.6-9.7℃) is 0.5℃ higher than that from CN05.1 (9.1℃) over the Beijing-Tianjin province for all ESMs (Table 1). While WRF produces warmer apparent temperatures in the city centers of Beijing and Tianjin and lower ones in the high Zhangjiakou mountains than recorded in the lower resolution CN05.1 observations. There are also differences between different models after WRF downscaling. For example, apparent temperatures from the two MIROC models downscaled by WRF are the warmest. In contrast AP from all 4 ESMs after ISIMIP shows very similar patterns (Fig. S1).

ESMs tend to overestimate the number of days with AP>32℃ in southeastern Beijing and the whole Tianjin province. Both ISIMIP and WRF appear to overestimate the NdAP_32 in Beijing urban areas and the southerly lowland areas although NdAP_32 is close to zero in the colder rural areas at relatively high altitude for both downscaling methods. Some of these differences may be due to the WRF simulations being at finer resolution than the 0.25°× 0.25° CN05.1, leading to higher probabilities of high AP in urban areas (Fig. 4d). ISIMIP results also show slight overestimations, especially in the tails of the distribution (AP>30℃) for urban areas (Fig. 4c). CN05.1 gives about 10 NdAP_32 per year in southern Beijing and Tianjin, but there are nearly 15 NdAP_32 from ISIMIP, and over 20 NaAP_32 per year from WRF downscaling in the Beijing-Tianjin urban areas during 2008-2017. NdAP_32 from WRF and ISIMIP downscaling of all ESM is overestimated relative to ERA5. But there are differences in ESM under the two downscalings: with ISIMIP, HadGEM2-ES and BNU-ESM have more NdAP_32 than the two MIROC models, while the reverse occurs with WRF (Fig. S2).

Lines 248-257 are edited as follows.

The Taylor diagram of the daily mean apparent temperature in Beijing-Tianjin province and Beijing-Tianjin urban areas from 2008-2017 for the 4 ESMs shows that correlation coefficients between ESMs and CN05.1 are greater than 0.85 under both downscaling methods. Although there are differences between ESMs, the performance of WRF, with higher correlation coefficient and smaller SD (standard deviation) and RMSD (root mean standard deviation), is usually superior to ISIMIP (Fig. S3). Taking the Beijing-

Tianjin urban areas as an example (Fig. S3b), under the ISIMIP method, MIROC-ESM, MIROC-ESM-CHEM and HadGEM2-ES have the same correlation coefficient (0.92) and RMSD (5.4℃) with the CN05.1, while BNU-ESM has lower correlation coefficient (0.88) and higher RMSD (7.0℃). Under WRF simulations, MIROC-ESM and MIROC-ESM-CHEM have larger correlation coefficients and smaller RMSD with CN05.1 than HadGEM2-ES and BNU-ESM.

Lines 283-300 are edited as follows.

Figure 5 shows the ISIMIP and WRF ensemble mean changes in the annual mean AP under G4 during 2060-2069 relative to the past and the two future RCP scenarios. ISIMIP-downscaled AP (Fig. 5a-5c) shows significant anomalies ($p<0.05$), with whole domain rises of 2.0 ℃ in G4-2010s, and falls of 1.0 ℃ and 2.8 ℃ in G4-RCP4.5 and G4-RCP8.5 respectively. In WRF results, AP under G4 is about 1-2 ℃ warmer than that under 2010s, 0.8 ℃ and 2.5 ℃ colder than that under RCP4.5 and RCP8.5 over the whole domain. Individual ESM results downscaled by ISIMIP and WRF are in Fig. S6 and Fig. S7. For both ISIMIP and WRF downscaling results, the two MIROC models show stronger warming than the other two models between G4 and the 2010s. WRF-downscaled AP driven by HadGEM2-ES exhibits the strongest cooling, with decreases of 1.7 ℃ between G4 and RCP4.5 and falls of 3.0 ℃ between G4 and RCP8.5. Although different ESMs show different changes in AP between G4 and other scenarios, changes in AP are almost the same everywhere for a given ESM in the ISIMIP results (Fig. S6). WRF-downscaled AP anomalies driven by two MIROC models are larger in the Zhangjiakou mountains and smaller in the Beijing urban areas and Tianjin city between G4 and 2010s (Fig. S7). Changes in AP from ISIMIP results, whether across whole province or just the urban areas, are statistically identical given scenarios (Table 2), which is consistent with patterns in figure 6. AP under G4 is 0.8 ℃ (1.0 ℃) and 2.6 ℃ (2.8 ℃) colder than that under RCP4.5 and RCP8.5 in Beijing-Tianjin urban areas from ISIMIP (WRF) results. The warming between G4 and 2010s in urban areas is 1.0 ℃ in WRF results, while that is 2.0 ℃ in ISIMIP results (Table 2).

Lines 312-330 are edited as follows.

Figure 6 shows the ISIMIP and WRF ensemble mean changes in the annual mean AP anomalies G4 during 2060-2069 relative to the past and the two future RCP scenarios. ISIMIP-downscaled AP (Fig. 6a-6c) shows significant anomalies ($p<0.05$) across the whole domain, even for the relatively small differences in G4-RCP4.5. ΔAP by WRF is lower than that by ISIMIP. Between G4 and 2010s, AP are projected to have increases of 1.8 (1.6), 2.1 (1.8), 2.4 (-0.2), 1.8 (0.8) ℃ from winter to autumn in ISIMIP (WRF) results. In ISIMIP results, the contribution of temperature ranges from 91%-104%, and the contribution of wind speed ranges from 3%-10% in all seasons, while the contribution of humidity is negative or insignificant (Fig. 6a). However, the contribution of humidity is positive in WRF results (Fig. 6a). Between RCP4.5 and

2010s, annual mean AP is projected to increase by 3.0 ℃ and 1.8 ℃ in ISIMIP and WRF results respectively, which is higher than that between G4 and 2010s. The increase of temperature and decrease of wind speed have a significant impact on the annual average ΔAP contributed 97% (94%) and 4% (3%) in ISIMIP (WRF) results. The contributions of changes in humidity are significantly positive under G4 and RCP4.5 in WRF results, while it is the opposite in the ISIMIP results (Fig. 6a-6b).

Relative to RCP4.5 in the 2060s, AP is projected to decrease by 1.0 (0.4), 0.7 (0.8), 0.8 (0.7), and 1.3 (1.4) ℃ from winter to autumn under G4 in ISIMIP (WRF) results (Fig. 7c). In summer, the contribution from changes in temperature and humidity are 94% (105%) and 8% (-9%) in ISIMIP (WRF) results, respectively. There are insignificant contributions from wind speed under ISIMIP results, but a significant slight positive contribution (0.7%-4%) under WRF results (Fig. 6c). The annual mean AP under G4 is 2.8 (2.6) ℃ lower than that under RCP8.5 in ISIMIP (WRF) result. In this case, the contribution of changes in wind on ΔAP ranges from 3%-5% by ISIMIP, while it is close to 0 by WRF. As expected, ΔAP is mainly determined by the changes in temperature, with contributions usually above 90% between different scenarios.

Lines 362-366 are edited as follows.

In contrast WRF suggests that most areas do not show any significant difference between G4 and the 2010s, while the anomalies relative to RCP4.5 are similar as ISIMIP, the differences are insignificant over more area than ISIMIP. G4-RCP8.5 anomalies with WRF are smaller than that with ISIMIP, and differences are not significant in the Zhangjiakou high mountains.

Minor comments

L45-47: Need citations to support idea that apparent temperature is actually an important variable

Reply: Done. I added the references.

Apparent temperature (AP), that is how the temperature feels, is formulated to reflect human thermal comfort and is probably a more important indication of health than daily maximum or minimum temperatures (Fischer et al., 2013; Matthews et al., 2017; Wang et al., 2021).

References

Matthews, T., Wilby, R., and Murphy, C.: Communicating the deadly consequences of global warming for human heat stress, PNAS, 114, 3861-3866, https://doi.org/10.1073/pnas.1617526114, 2017.

Fischer, E., and Knutti, R.: Robust projections of combined humidity and temperature extremes, Nat. Clim. Change, 3, 126-130, https://doi.org/10.1038/nclimate1682, 2013.

Wang, P., Luo, M., Liao, W., Xu, Y., Wu, S., Tong, X., Tian, H., Xu, F., and Han, Y.: Urbanization contribution to human perceived temperature changes in major urban agglomerations of China, Urban Clim., 38, 100910, https://doi.org/10.1016/j.uclim.2021.100910, 2021.

Equation 3 is not exactly the Clausius-Clapeyron equation. It is an approximate form which fits some empirical data. Please provide the relevant citations for this relationship (most likely Tetens (1930), Murray (1967), and Monteith and Unsworth (2008)).

Reply: Thanks. I have added the citations.

P_s is calculated using the Tetens empirical formula (Murray, 1966):

$$P_s = \begin{cases} 0.61078 \times e^{\left(\frac{17.2693882 \times T}{T+237.3}\right)}, & T \geq 0 \\ 0.61078 \times e^{\left(\frac{21.8745584 \times (T-3)}{T+265.5}\right)}, & T < 0 \end{cases} \quad (3)$$

References

Murray, F.: On the computation of saturation vapor pressure, Rand Corp Santa Monica Calif, 1966.

L162: Citation needed for US NWS

Reply: Done.

This threshold does not lead to extreme risk and death, instead it is classified as requiring "extreme caution" by the US National Weather Service (National Weather Service Weather Forecast Office, *https://www.weather.gov/ama/heatindex*).

L164-166: The rationale for using NdAP_32 does not make sense to me. Since you are looking to identify an increase in the frequency of a rare event, why does the fact that it is rare mean that you should not use it? Similarly, why presume that the same outcome will apply for higher thresholds? I suggest revising the rationale.

Reply: We cannot simply use any threshold because the less frequent the threshold the more statistically uncertain is the estimate of its probability. For example the well-known estimate for the uncertainty in an estimate of uncertainty, s, is $s/\sqrt{(2n-2)}$. So if we have only a very small number of instances of s, (that is n) then its uncertainty is very high. So, we must compromise in having a measure of extreme that represents the tail of the distribution, while at the same time being common enough for a reasonable sampling of its likelihood in the 50 years or so of simulations available. This is why we choose NdAP_32 rather than say NdAP_27 or NdAP_36.

In regard of the second point - we do not necessarily think that rarer events will be changed by the same amount as NdAP_32, in fact, often extremes change more than central parts of the distribution. This is a consequence the "fat tails" seen in most real-

world climate distributions. The reason for the fat-tailed nature of real-world climate simulations probably relates to the long term spatial and temporal persistence (that is not simply autocorrelation) of processes rather than them being independent, and also to the presence of hysteresis behaviour (tipping points) in the system as it pushed further from the long term mean – for example in the fundamentally different physics at play on either side of the ice/water phase change. The fat tails implications for risk were examined in regard of economics by Weitzman's (2009) Dismal theorem which showed that since the likelihood of extreme fat tail probability distributions decay polynomially, the damage associated with them rises exponentially, thus leading to no bound when integrated to infinity. However, we do not think this is useful discussion in the manuscript. The issue is that we do not have the statistical power to discuss rarer extremes than NdAP_32 with the data available. This is what we tried to explain in the text in a simple way.

References

Weitzman, M.: On Modeling and Interpreting the Economics of Catastrophic Climate Change, Review of Economics and Statistics, 91, 1–19, https://doi.org/10.1162/rest.91.1.1, 2009.

We revised the text:

This threshold does not lead to extreme risk and death, instead it is classified as requiring "extreme caution" by the US National Weather Service (National Weather Service Weather Forecast Office, *https://www.weather.gov/ama/heatindex*), but carries risks of heatstroke, cramps and exhaustion. A threshold of 39°C is classed as "dangerous" and risks heatstroke. While hotter AP thresholds would give a more direct estimate of health risks, the statistics of these presently rare events mean that detecting differences between scenarios is less reliable than using the cooler NdAP_32 threshold simply because the likelihood of rare events are more difficult to accurately quantify than more common events that are sampled more frequently. While there is evidence to suppose that in some distributions, the likelihood of extremes increases more rapidly than more central parts of a probability distribution – such as larger Atlantic hurricanes increasing faster than smaller ones (Grinsted et al., 2013), a conservative assumption is that similar differences between scenarios would apply for higher thresholds.

References

Grinsted, A., Moore, J., and Jevrejeva, S.: Projected Atlantic tropical cyclone threat from rising temperatures, PNAS, 110, 5369-5373, https://doi/10.1073/pnas.1209980110, 2013.

L168: "Since health impacts are more important where there are more people": this seems like a value judgement, and not (I think) the intended meaning. I recommended simply stating that you calculated population-weighted changes.

Reply: No it is not. There are no value judgements here at all. The value of human life is exactly the same in the sentence, i.e. each life is the same. There are simply more lives in urban areas than rural ones. Hence the phrase “more important”. This is the same logic and values that suggest we should not be worried at all about human health impacts of climate change on Mars because there are no people there to be impacted.

The dark colors in Figure 6 make it nearly impossible to read the data.

Reply: We have changed the color bar in Figure 6.

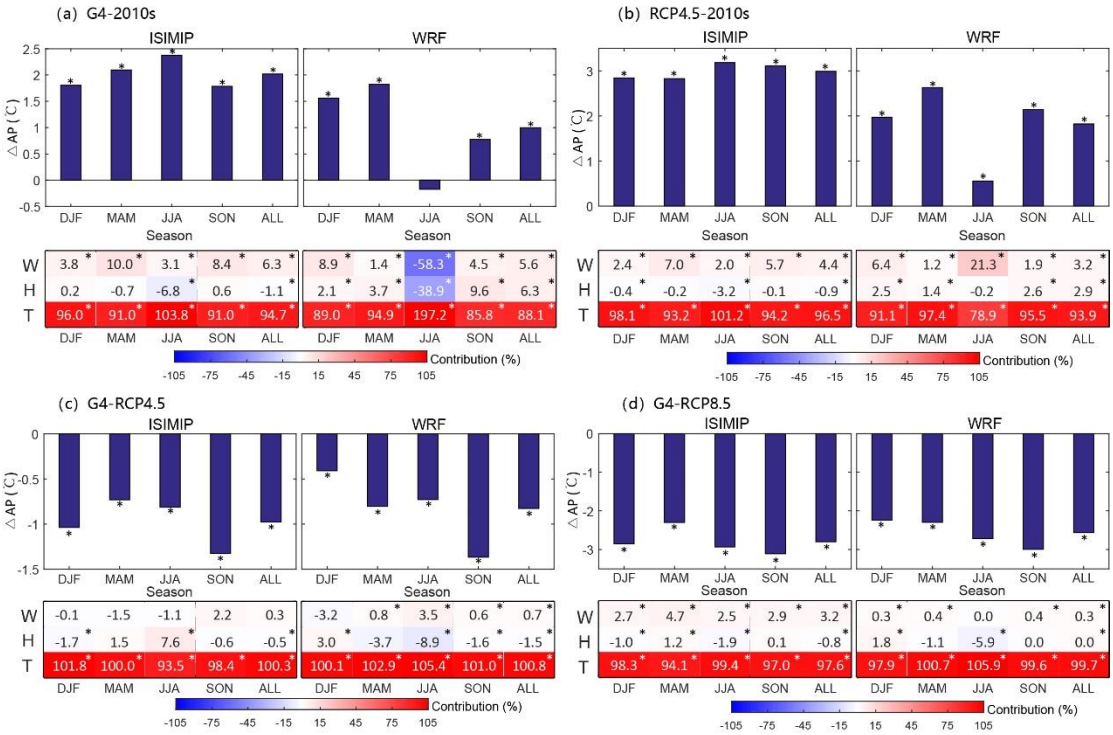


Figure 6. The seasonal changes of AP (ΔAP) and the seasonal contribution of climatic factors to ΔAP for Beijing and Tianjin urban areas under ISIMIP and WRF between G4 and 2010s **(a)**, G4 and 2010s **(b)**, G4 and RCP4.5 **(c)** and G4 and RCP8.5 **(d)** based on ensemble mean results. Colors and numbers in each cell correspond to color bar, and “*” above the columns and in the cells indicate differences are significant at the 5% significant level under the Wilcoxon test.

Figure 7: please label the months.

Reply: Done.

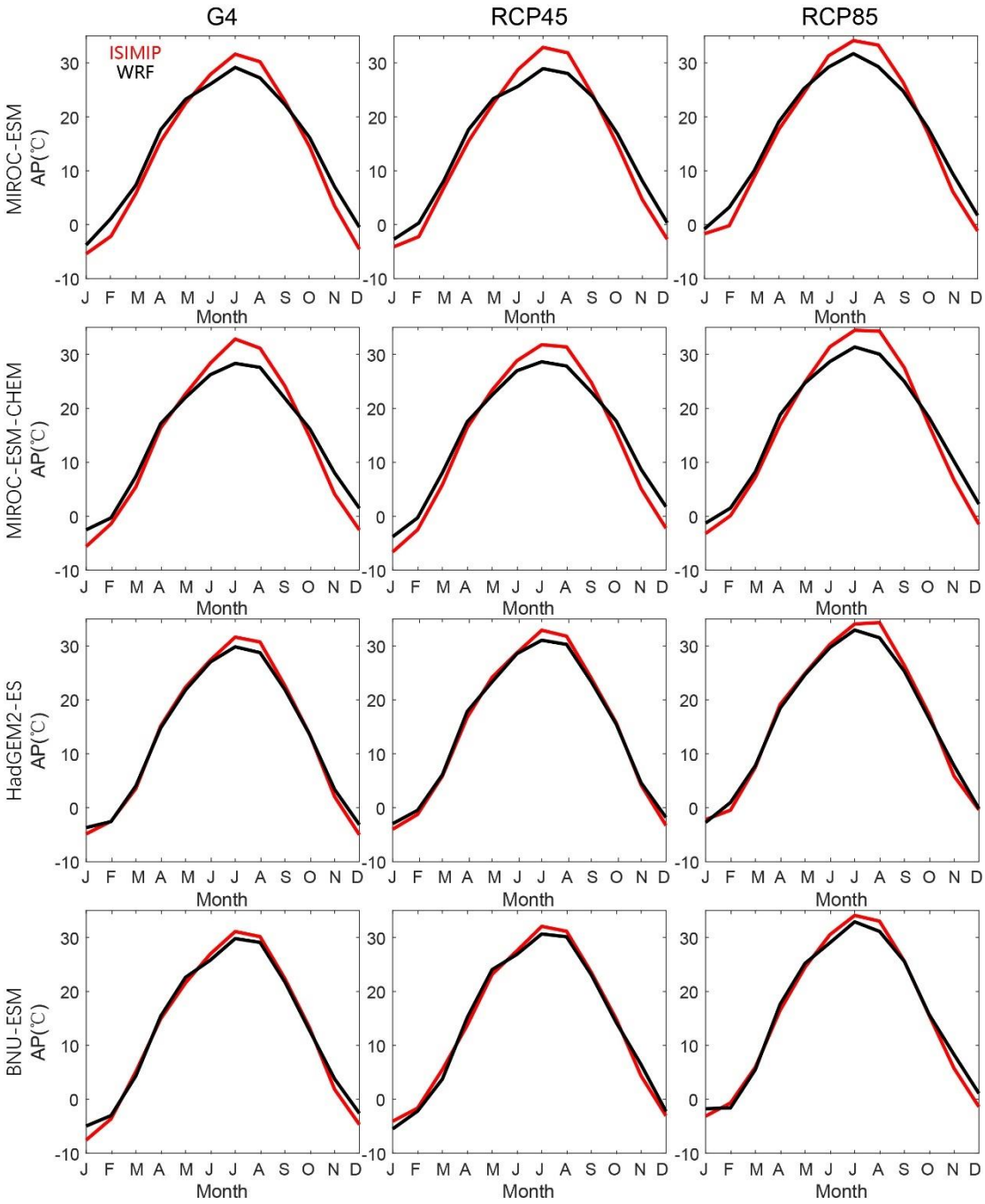


Figure 10. Seasonal cycles of apparent temperature from MIROC-ESM, MIROC-ESM-CHEM, HadGEM2-ES and BNU-ESM under G4, RCP4.5 and RCP8.5 in Beijing-Tianjin urban areas during 2060s based on ISIMIP (red) and WRF (black) methods.

Throughout: it would be helpful to see the baseline (undownscaled) results alongside the downscaled results, so that the readers might know how significant the differences between ISIMIP and WRF are compared to the differences between the original and downscaled outputs.

Reply: We plot the ESMs original AP in Fig. 3. We added two sentences after line 222 and 229.

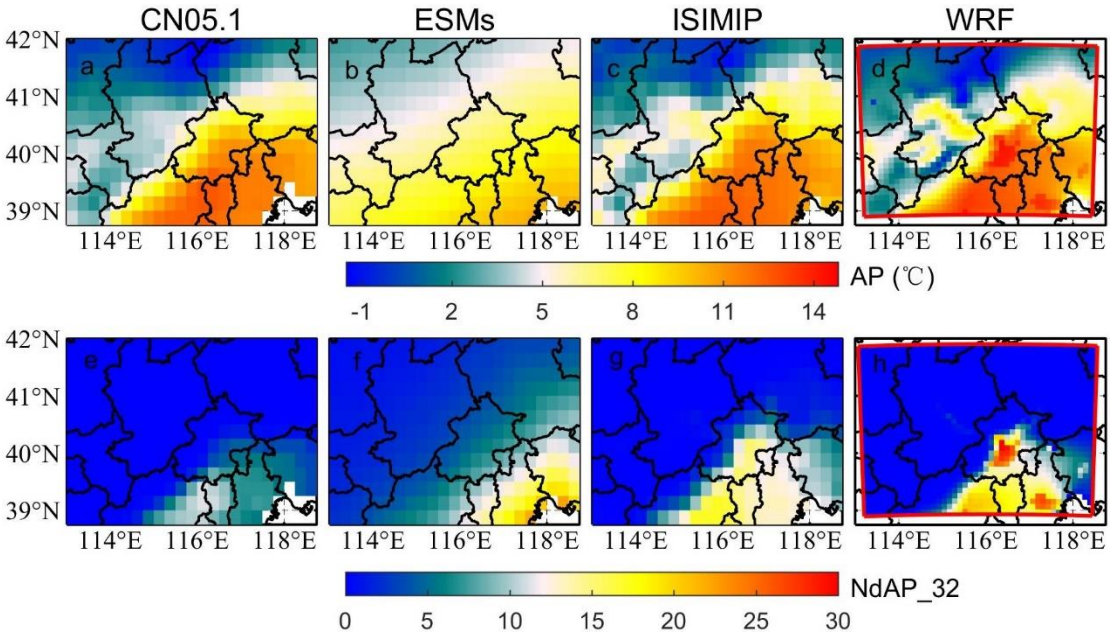


Figure 3. Top row: the spatial distribution of mean apparent temperature from CN05.1 **(a),** raw ESMs ensemble mean after bilinear interpolation **(b)**, 4-model ensemble mean after ISIMIP **(c)** and ensemble mean after WRF **(d)** during 2008-2017. Bottom row: the spatial distribution of annual mean number of days with AP > 32℃ from CN05.1 **(e)**, ESMs **(f)**, ISIMIP **(g)** and WRF **(h)** during 2008-2017. Fig. S1 and Fig. S2 show the pattern of AP and NdAP_32 for the individual ESM.

While the raw AP from ESMs is overestimated in the Zhangjiakou high mountains and underestimated in the southern plain, and shares a similar pattern with temperature from ESMs (Wang et al., 2022). The raw ESM outputs were improved after dynamical and statistical downscaling.

ESMs tend to overestimate the number of days with AP>32℃ in southeastern Beijing and for the whole Tianjin province.