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

Referee 2's comments

General Comments

The purpose of this work was to compare the changes in apparent temperature across three future scenarios using two different downscaling techniques. The authors find that although both downscaling methods using ISIMIP and WRF reproduce historical observations, projections to future scenarios produce differing results. In general the authors conclude that changing temperature contributes most to changes in apparent temperature which is driven by a combination of 2-m temperature, relative humidity and windspeed to more accurately capture the physiological impact of a warming climate. They find that the occurrence of days exceeding a 32 deg C apparent temperature threshold across the Beijing-Tianjin megalopolis will increase in frequency under RCP 4.5 and RCP 8.5. They draw attention to GeoMIP scenario G4, designed to test SAI for its ability to mitigate risks from an RCP 4.5 scenario, finding that individual ESMs show no statistically significant differences in the number of days exceeding 32 deg C apparent temperature between G4 and RCP 4.5.

This paper is companion to a manuscript which focuses on the impacts of using different downscaling methods in the Beijing-Tianjin region. Therefore, my interpretation is that the core contribution of this paper is to quantify the effect of proposed SAI for this region, and to compare this effect across two downscaling methods. However, in some ways this comparison seemed like an afterthought, and the paper discussion centered on the differences in apparent temperature across the two downscaling methods. Therefore, I would agree with reviewer 1 in claiming that this paper is providing only an incremental contribution with this manuscript. If the authors were to reframe this piece to focus on key differences in SAI forcing vs the RCP 4.5 scenario using downscaling to identify sources of uncertainty in response, then I believe this would provide novel insight into SAI as a proposed technique.

Reply: Thanks for your constructive suggestions. As noted we mainly explore the impact of geoengineering on apparent temperature, which is affected by the combined effects of temperature, humidity and wind speed. It is clear that the change in temperature dominates the change in apparent temperature. To study more impacts of changing meteorological conditions under SAI on human health, we develop an analysis of air pollution, using $PM_{2.5}$ since this is the dominant component at present in the region in our revised manuscript. The specific revisions are shown in response to referee 1.

Major Comments

Use of the WRF and ISIMIP downscaling techniques across the four ESMs used was technically interesting and well executed. The use of apparent temperature was also useful, but given the results were largely driven merely by changes in 2 m surface temperature, it left the reader waiting for more of an understanding of the impact of the work.

1. To better frame this piece I believe the manuscript would be more clearly a departure from the companion piece if the framing of the paper was towards understanding the inter-scenario responses vs. the inter-method responses in apparent temperature. This would also make the work more appropriate for submission to the special issue on solar geoengineering.

Reply: We developed a novel analysis of the role of SAI in air pollution from $PM_{2.5}$ part in our revised manuscript. This is the first such analysis made under geoengineering scenarios, and is of interest since $PM_{2.5}$ plays a serious role in health in the region.

2. I would encourage the authors to include more than apparent temperature in these results. Given this is a monsoon region is there a reason why precipitation was not a variable included with apparent temperature? Soil moisture is also a useful metric when understanding SAI; and could give the reader more insight into expected agricultural outcomes.

Reply: Apparent temperature is mainly related to the temperature, humidity and wind speed. In the formula of calculating AP, we can also know the relationship between AP and three meteorological factors. Precipitation itself has little impact on apparent temperature, which has little to do with whether it is located in the monsoon region. Crop yield is also an important index. Many studies show that the crop yield is projected to change under geoengineering scenarios, (Zhan et al., 2019; Fan et al., 2021). However, our main focus is the effect of geoengineering on human wellbeing. This is an expansive topic and we cannot deal completely with it in a single paper or reasonable length. So we chose to focus here on looking at changes that may occur as the climate is modified in the emissions and distribution of $PM_{2.5}$ aerosol. This is a novel analysis that could be of interest to a wide community.

References

Zhan, P., Zhu, W., Zhang, T., Cui, X., and Li, N.: Impacts of sulfate geoengineering on rice yield in China: Results from a multimodel ensemble, Earths Future, 7, 395-410, https://doi.org/10.1029/2018EF001094, 2019.

Fan, Y., Tjiputra, J., Muri, H., Lombardozzi, D., Park, C., Wu, S., and Keith, D.: Solar geoengineering can alleviate climate change pressures on crop yields, Nat. Food, 2, 373-381, https://doi.org/10.1038/s43016-021-00278-w, 2021.

3. The authors use AP>32 deg C as a metric citing that "similar differences between scenarios would apply for higher thresholds." I would be curious for the author to provide us with results based on a 32 deg C threshold as well as a higher limit instead of speculating here. This could also help the author tie these findings to tangible impacts, such as mortality, or even economic outcomes.

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_39.

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.

4. I would encourage the author to more clearly tie the variable changes to tangible impacts; specifically mortality or economic outcomes. Or at least make this a larger portion of the discussion. The results of the two downscaling methods employed can then provide a measure of uncertainty in the expected response to SAI or future warming.

Reply: We add the $PM_{2.5}$ part and discuss the contribution of changes in three variables to changes in $PM_{2.5}$ concentration.

Stylistically I found much of the results section difficult to decipher and was confused by qualitative descriptions of changes and the use of subjective adverbs. This section should be reconsidered with some rigor to provide the reader with a clearer quantitative description of each piece of analysis. I was also unable to see any equations in this manuscript – and based on the comments of reviewer 1 I would second concern regarding the use of linear regression to quantify contribution of wind, RH and T to the apparent temperature.

Reply: The wording has been changed where possible. This results in more repetitive wording.

The reason we used the regression approach is that this produces a least squares estimate of contributions. This is preferable in many statistical applications. However it may not be the best choice here as the assumptions of Normality and homoscedasticity in the analysis are probably not true. Using the referee #1's suggestions are more localized estimates around the mean values, which could be regarded as more statistically more robust, giving less weighting to outliers. But these are reasonable alternatives and the gradient or Jacobian approach plays a role in statistical analyses.

Please see the more detailed analysis we provide in response to Referee #1 on this issue. The conclusion we reached was:

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.

Minor comments:

(88) : I was confused by "Beijing experienced an increasing trend of 12.7% or 2.07 days per decade in extreme warm nights (Wangetal.,2013) …" does this mean they experienced a percent increase of 12.07% per decade in the number of extremely warm nights – I was unable to confirm this based on the citation provided.

I was also just a bit confused why nights was the most useful metric here. Has anyone done a study of the increase in warm days from 1978-2008. It seems in line with your paper it could be useful to provide information regarding the historical increase in surface temperature vs. apparent temperature in this region.

Reply: We changed the sentence.

Over the period of 1971-2014, apparent temperature rises at a rate of 0.42℃/10 years over Beijing-Tianjin-Hebei region, with urbanization having an effect of 0.12℃/10 years (Luo and Lau, 2021).

(105) Define AP before using as an acronym

Reply: Done. We define the AP in line 45.

(109) Figure 1, Panels C and D color-bar labels should be added to specifiy units. I would also specific in the label what the red line is in figure 1, D – it seems that this is where the WRF domain terminates?

Reply: We pointed out in the annotation that the units of panels c and d refer to the number of people within the gird cell. The red line is the south boundary in WRF domain as is clear from panel b, and the that there are no blue cells south of the red line.

(127) Consider changing "climate forcing comes from 4 ESMs)" to something like climate simulations were performed by 4 ESMs – for clarity. As written it sounds like the radiative forcing from each model was extracted or somehow used separately.

Reply: We changed it:

Climate simulations are performed by 4 ESMs: BNU-ESM (Ji et al., 2014), HadGEM2-ES (Collins et al., 2011), MIROC-ESM (Watanabe et al., 2011) and MIROC-ESM-CHEM (Watanabe et al., 2011).

(151-158) I am unable to see any equations in the pdf preprint view of this document (I presume this is not an author issue but rather a technical issue!)

Reply: Ok. I am sure equations can be displayed normally.

(159-162) It would be useful to supply the read with values of the various thresholds for context as they read. I'm finding myself curious – what is the physiological maximum of the apparent temperature that humans can tolerate? What is the dangerous level? I would explain this before diving into your threshold value of 32 deg C to give the reader greater context. I would also provide citations of this empirically based scale.

Reply: We add the table in the supplementary information which gives the various thresholds and potential impacts according to by the US National Weather Service (National Weather Service Weather Forecast Office, *https://www.weather.gov/ama/heatindex*). We select the threshold AP of 32 as a trade off between rarity and hence uncertainty of its likelihood in each scenario, and the threat to health.

Table S1. Apparent temperature thresholds and its health impact (National Weather Service Weather Forecast Office, *https://www.weather.gov/ama/heatindex*).

US NWS Classification	AP threshold	Effect on the body
Caution	27-32°C	Prolonged exposure and/or physical activity can cause fatigue
Extreme caution	32-39°C	Prolonged exposure and/or physical activity can lead to heatstroke, heat cramps, or heat exhaustion
Danger	39-51°C	Heat cramps or heat exhaustion may occur, and prolonged exposure and/or physical activity may cause heatstroke
Extreme danger	>51°C	Very likely to suffer from heat stroke

We add more text to contextualize the threshold.

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.

Figure 2: I would ask the author to revise the labeling of the terms – I am not following the utility of the bar chart here. These terms should be telling the reader information about the contribution to AP from each of three terms, however the bar chart makes it seem like terms 1-3 should add to the given AP; but on close inspection they do not? Unless my pdf view is not showing me, the coefficients on each term are also not defined directly in the text.

Reply: Maybe it's because the equation isn't shown in your downloaded preprint. In this figure, we show the equivalent temperature caused by three variables in Equation (1). The bars show the level of the equivalent temperature under each factor, and also shows the performance of downscaling on three variables.

Figure 3: I consider labeling the color bars of the bottom panels with # of days > 32 or something equivalent for clarity.

Reply: The color bar is labelled NdAP_32 and described in the caption as annual mean number of days with AP > 32℃

(234) : Missing "of" in the sentence "across most the North China.."

Reply: Done. We reedited this sentence.

Figure 4: consider modifying the titles of the lower plots for clarity to also read ISMIP and WRF, since all four plots are showing AP.

Reply: We have increased the font size of the title and widened the spacing between the top two panels and bottom two panels.

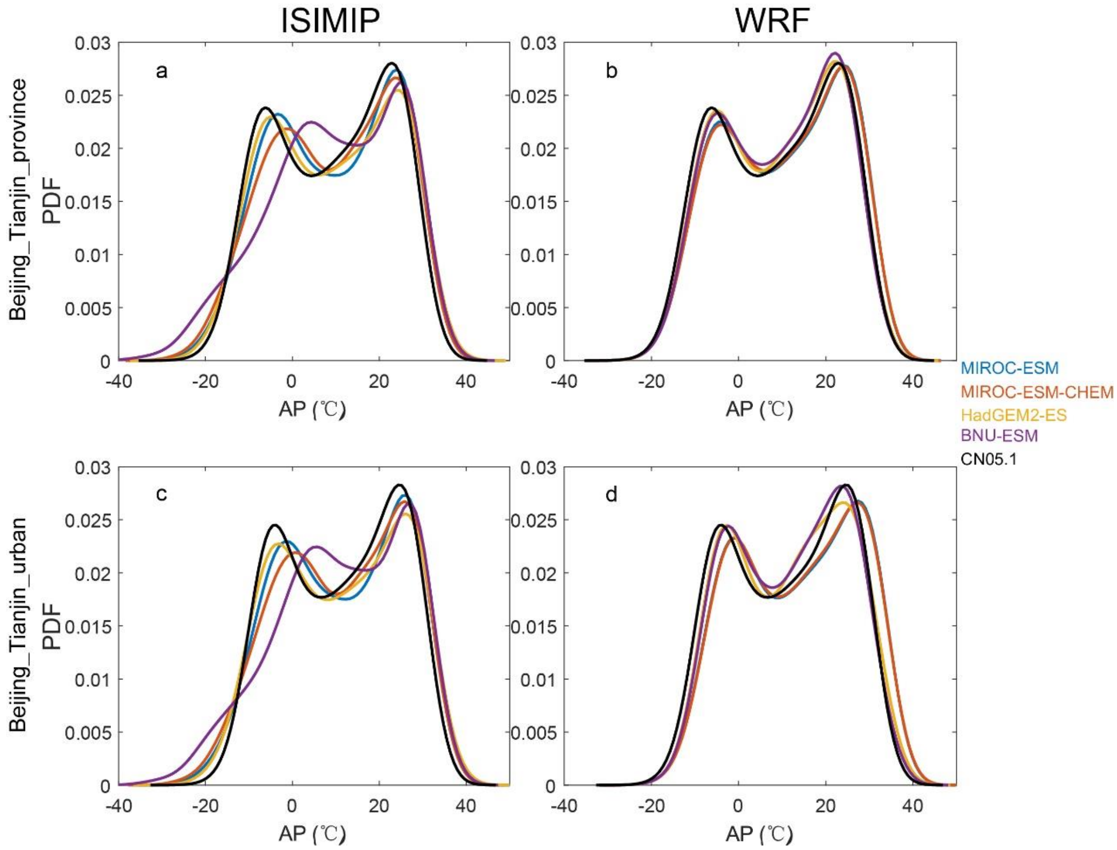

Figure 4. The probability density function (pdf) for daily apparent temperature under ISIMIP **(a, c)** and WRF **(b, d)** results in Beijing-Tianjin province **(a, b)** and Beijing-Tianjin urban areas **(c, d)** during 2008-2017.

Figure 5: The purpose of this figure is to compare the downscaling across WRF and ISIMIP – however the colorbar is constrained to give the reader cross scenario information. I would consider using different colorbars to for each scenario to better highlight differences in downscaling method, otherwise it seems these results cannot be well resolved by the reader.

Reply: Done.

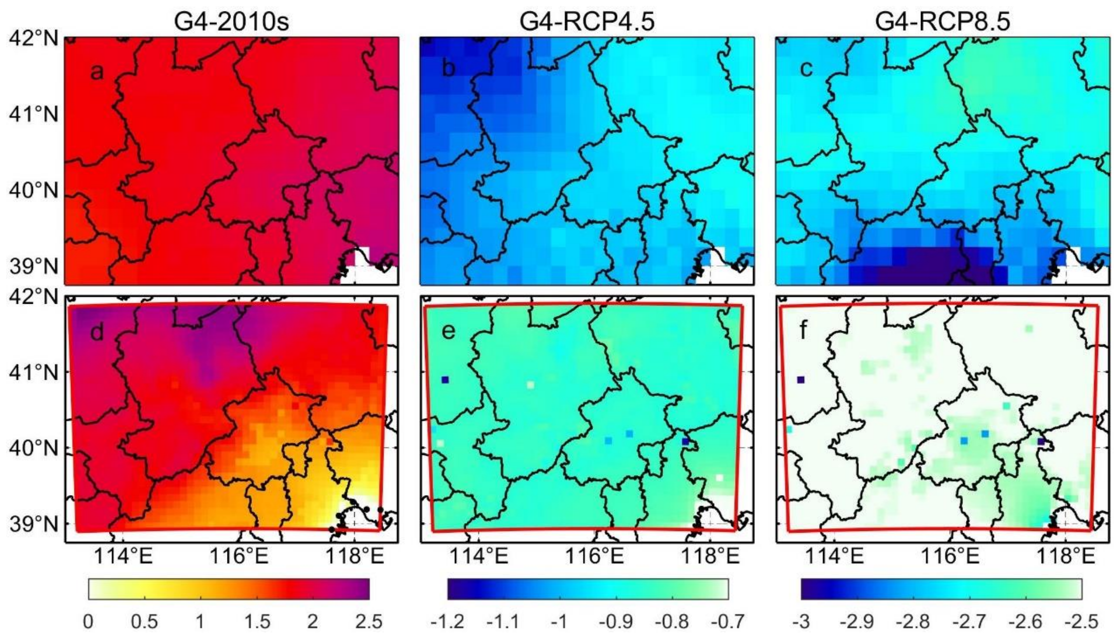

Figure 5. Spatial pattern of ensemble mean apparent temperature difference (°C) under different scenarios over 2060-2069: G4-2010s (left column), G4-RCP4.5 (middle column) and G4-RCP8.5 (right column) based on ISIMIP and WRF methods. 2010s refers to the 2008-2017 period. Stippling indicates grid points where differences or changes are not significant at the 5% level according to the Wilcoxon signed rank test.

Also, this figure is called back in line 285 as Fig. 5a-5c; please add alphabetic denomination in the figure paneling.

Reply: I am sure that there are alphabetic denomination in each panel, maybe it is a display problem. In the new figure, I again added the alphabetic denomination.

(320-331): in expressing results in this section be more direct in statements and consider breaking into smaller sentences. Be more quantitative in these expressions and consider modifying the descriptor "significantly" as it is unclear if this means humidity is changed at a statistically significant amount in this context. In several sentences in this section the author quantifies the contribution of humidity or wind to changes in AP – make more clear to the reader what this percent is referring to (e.g. amount to over 3% of the total change in delta AP in summer).

Reply: We reedited this section.

Figure 7 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. 7a-7c) 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) °C 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. 7a). However, the contribution of humidity is positive in WRF results (Fig. 7a). Between RCP4.5 and 2010s, annual mean AP is projected to increase by 3.0 °C and 1.8 °C 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. 7a-7b).

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) °C 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. 7c). The annual mean AP under G4 is 2.8 (2.6) °C 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.

Figure 8: Consider changing stippling to a cross grid hatching to allow the reader to perceive values in the bottom WRF panel that are not statistically significant but still provide context to the reader.

Reply: I reduced the size of the stippling points and updated the figure. The values which are covered by black points are nearly zero in the panel d-f.

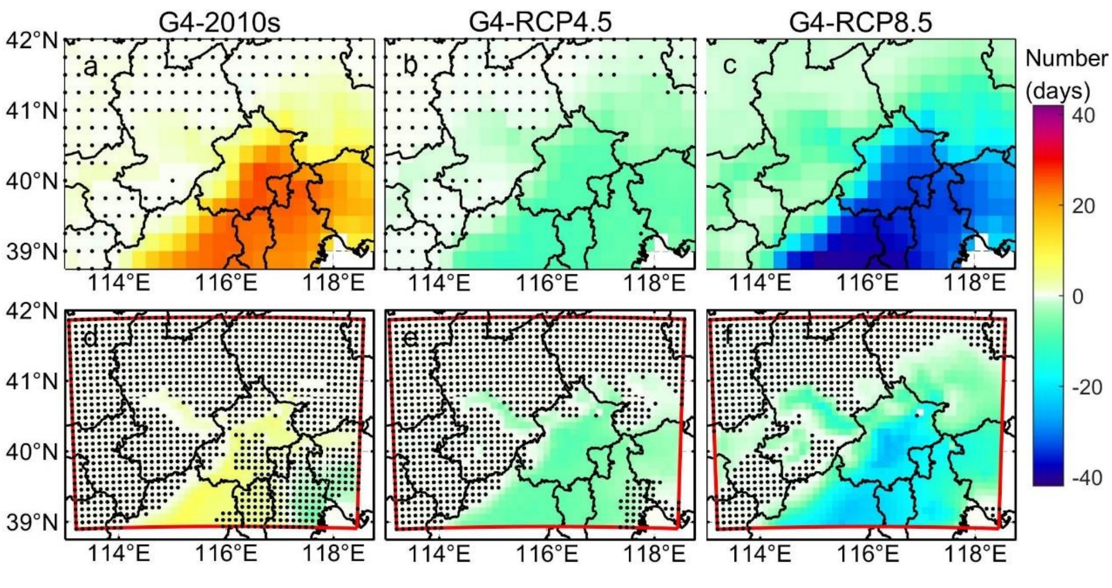

Figure 8. Ensemble mean differences in annual number of days with AP > 32°C (NdAP_32) between scenarios for 2060-2069: G4-2010s (left column), G4-RCP4.5 (second column) and G4-RCP8.5 (right

column) based on ISIMIP method and WRF. 2010s means the results simulated during 2008-2017. Stippling indicates grid points where differences or changes are not significant at the 5% level according to the Wilcoxon signed rank test. Corresponding ISIMIP results for each ESM are in Fig. S11, and WRF results in Fig. S12.

(433) Change “warmer that 2m” to warmer than 2m.

Reply: Done.

AP is about 1.5°C warmer than 2 m temperature over the Beijing and Tianjin urban areas in summer.