We thank this reviewer for their useful review of our manuscript. Below we provide a point-by-point response.

Major issues.

2 major methodological choices that need to be addressed because the processes destroy physical relationships or modeling is directly outside the scopes of calibrations.

I. Bias Correction of non-linear, thus non-stationary variables.

1) Bias Correction Motivation

The authors do not motivate why they are bias correcting their data. They cite a handful of manuscripts that show research groups bias correcting data. But they don't actually explain why they need to bias correct their own data. Bias corrections are necessary when the output being used is incompatible with a tool that it is being applied for. For example: precipitation from a GCM is on a 100km x 100km grid and the rainfall fields produced are often a constant drizzle. This output is required to drive a hydrological catchment model, however, the rainfall from the GCM does not represent any catchment scale stochastic processes. Therefore a correction is required to be able to continue the research. Within the context of this manuscript, I don't see any motivation for requiring some sort of bias correction. Population data is interpolated to the GCM grid, or diagnostics are executed on the the GCM outputs. Nothing that warrants utilizing a bias correction that would be imperative for interpreting the results.

We thank the reviewer for pointing out this issue. We agree with the reviewer's opinion, and the bias correction has been deleted from the manuscript. The entire analysis has been re-done with the original output from the model without bias-correction, including the reviewer's further suggestion on calculating the wet-bulb temperature.

We have replaced the bias correction with a sensitivity test that evaluates potential uncertainty in our results (described in Sect. 2.1 and throughout the paper). Overall, we find that the biases lead to changes much smaller than the changes in temperature and wet-bulb temperature due to climate change.

2) Bias correction methods.

Bias correction of a covariance of temperature-humidity is extremely difficult to produce reliable results that are not physics breaking …

As discussed above, we have removed this from the paper.

II. Choice of heat stress algorithm.

The authors use wet bulb temperatures as their primary heat stress indicator. There are various reasons why this is good and bad, and a battery of metrics would probably be a better approach (see Buzan et al., 2015). I think adding multiple more metrics to the manuscript would reduce the

clear language and systematic approaches in the analysis. However, what is of major concern is the use of Stull 2011 for wet bulb temperatures. Much like how the statistical bias correction methods are only valid for modern climate, Stull wet bulb temperature, too was specifically calibrated for modern climate, which limit its capacity in global warming applications (Buzan et al., 2015). Figure 1 Buzan et al., 2015 demonstrates the increasing growing errors that occur as temperature increases. A better method is the Davies-Jones 2008 wet bulb temperatures. Specifically equations, 4.8-4.11 using Bolton 1980 eqn. 39. for equivalent potential temperature inputs (Davies-Jones 2009 evaluates various different equivalent potential temperature calculation methods and demonstrates that bolton eqn. 39 is the best). The easiest way to calculate all of these variables is with the HumanIndexMod (Buzan et al., 2015). Python enabled:

We have replaced the Stull calculation in the paper with the Davies-Jones calculation.

It is difficult to determine if the wet bulb temperature errors are coming from the Stull or the Bias correction (likely both). But these errors have serious consequences for the results: I am suspicious that line 327 states that 5% of the Earth's population is exposed to 180 deadly days and 302 tropical nights. Just a quick peak at 4x daily JRA55 shows the 1986-2005 climatology the value of Tw 25°C does not appear until the ~60th Percentile, i.e. less than half of the available deadly days (if am understanding the definition of days properly). I am not sure how the authors were able to generate 302 tropical nights deadly for modern climate, which is more than 9 months a year. My JRA55 climatology only starts to have 25°C appearing at the 25th percentile. I recommend using the Davies-jones eqn. 4.8-4.11 as in the HumanIndexMod

As mentioned in previous points, we have replaced our calculation with Davis-Jones 2009. However, we still find many Mora et al deadly days (daily maximum w2m over 24°C) and tropical nights even in our reference period (Fig. R1). While we have not confirmed the results with the JRA data, it is certainly possible that the difference is due to different time periods — the reviewer is looking at data from 1986-2005, while our reference period covers 2003-2017.

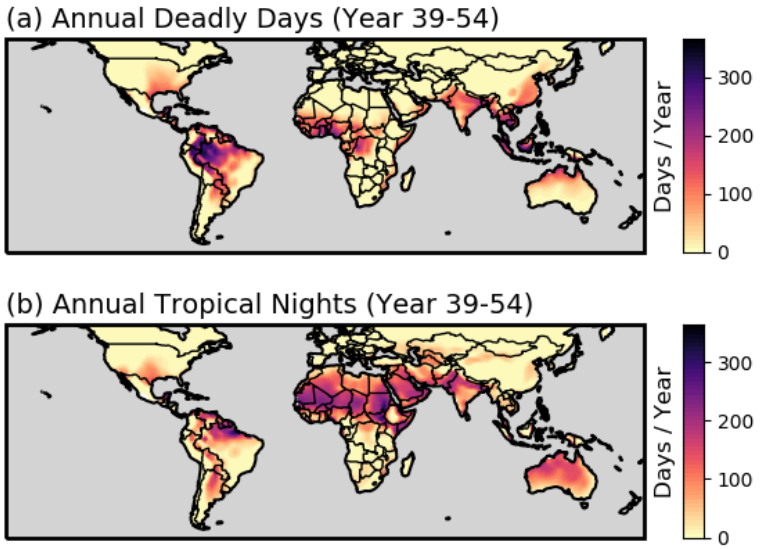

Figure R1. Number of ensemble-averaged annual (a) deadly days (daily maximum over 24°C) and (b) tropical nights (daily minimum t2m over 25°C) in 1% CO_2 experiment, in same global average temperature as present day (0.87°C, year 39-54)

To investigate this in more detail, we plotted the 15-year median value of daily maximum w2m (metric for deadly days) and daily minimum t2m (metric for tropical nights) for present day (2003-2017 in ERA-I, year 39-54 in MPI) as in figure R2.

As seen in the figure, in both ERA-I and 1% CO_2, we find median values of daily maximum being over 24°C (threshold for deadly days) in central Africa, Southeast Asia, and Northern part of South America. This translates to over 180 days of deadly days per year. As for 15-year median values of daily minimum t2m, we find values over 25°C (threshold for tropical nights) in Southeast Asia, India, central Africa, Northern Australia, and Northern South America. This also translates to over 180 days of tropical nights per year.

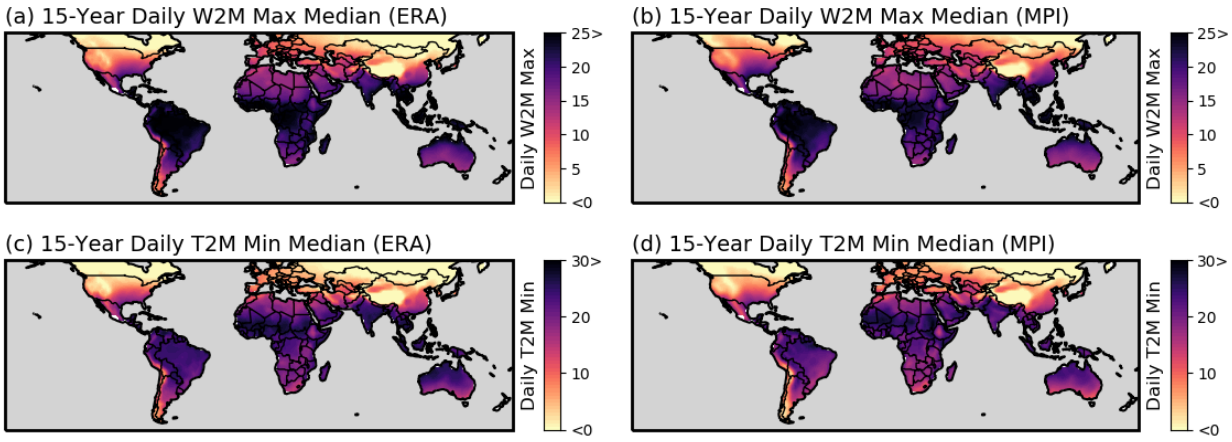

Figure R2. 15-year median values of daily maximum w2m and daily minimum t2m for ERA-Interim and 1% CO_2 experiment in present day warming (2003-2017 for ERA Interim, year 39-54 for 1% CO_2).

Minor issues:

Mora et al., 2017 really puts a low threshold for deadly heat stress. The world exposure to these conditions is fairly high, yet we don't have people dying all over. More likely, there are epidemiological reasons for people suffering heat stress, i.e. health, socio-economic, etc (which Mora states). But it really makes it difficult to use these values as a realistic driver of impacts on humans. Harder limits, such as Tw 32°C, where all laborers cannot work anymore for sustained amounts of time (Brunt 1943; Liang et al., 2011), at least have clearer thresholds for populations being impacted. Something to consider in analysis…

We agree that Mora et al., 2017 suggested a low threshold, although it represents a temperature where mortality begins to occur. So, we revised our definition of "Deadly Days" to daily maximum wet bulb temperature above 26°C.

We have changed the text to read: [L188-196] We therefore also count the number of days each year with daily maximum w2m above 26°C, which we refer to as "deadly days". We note that other values could be chosen, with higher values occurring less frequently but having more significant impacts. This value is based on the analysis of Mora et al. (2017), who demonstrated that w2m of about 24°C is the threshold which fatalities from heat-related illness occur. However, since we find that there are some regions that already experience over 9 months of 24°C w2m events per year, we increase this threshold to 26°C in our analysis. We could have chosen higher w2m values, but any choice in this range is associated with negative impacts, so we have chosen a value near the bottom of the range where mortality occurs in order to maximize the signal in the model runs.

The CDD methods are fine, but I worry about using the bias corrections. The 28 member ensemble should characterize the variability of that 18°C threshold without the need of bias correcting the results.

We have removed the bias-correction from the analysis and used the original output of the model.

Line 319 is stating rapid increase in wet-bulb extremes. I am not quite following the language. Is this due to large population living in the tropics? Some clarity here would be useful.

Explanation is added for clarity.
[L330-334] Given that the planet has already warmed about 1°C above pre-industrial, this suggests that the world should presently be experiencing a rapid increase of wet-bulb extreme frequency, particularly in the tropics. This is related to the increased slope in Figure 6, in which cluster 1 and 2's values of HWD_{w2m} and HWF_{w2m} increase rapidly until 3.0°C and 2.0°C of global warming.

Figure 6 K means should be listed. a) is not described in the caption (no colorbar). I think the colors match figure 7, but I am unsure. Ah, I think it is in Table 2. I think there just needs to be a reference in Figure 6 and 7 to Table 2 for the descriptions.

Thanks for pointing this out. Figure captions have been revised for clarity.

Figure 5. (a) Clustered regions via K-means clustering. Characteristics of each cluster are listed in Table 2.
Figure 6. Evolution of each index averaged over each cluster. Colors are consistent with Figure 5 and Table 2.

Recommended citations:

Buzan JR, Oleson K, Huber M. 2015. Implementation and comparison of a suite of heat stress metrics within the Community Land Model version 4.5. Geosci. Model Dev. 8:151–70

Brunt D. 1943. The reactions of the human body to its physical environment. Q. J. R. Meteorol. Soc. 69:77– 114

Liang, C., Zheng, G., Zhu, N., Tian, Z., Lu, S., and Chen, Y.: A new environmental heat stress index for indoor hot and humid envi- ronments based on Cox regression, Build. Environ., 46, 2472– 2479, 2011.

Davies-Jones, R.: An efficient and accurate method for computing the wet-bulb temperature along pseudoadiabats, Mon. Weather Rev., 136, 2764–2785, 2008.

Davies-Jones, R.: On formulas for equivalent potential temperature, Mon. Weather Rev., 137, 3137–3148, 2009.

Zhang et al., 2021 Projections of Tropical heat stress constrained by atmospheric dynamics.

Schwingshackl et al., 2021 Heat Stress Indicators in CMIP6: Estimating Future Trends and Exceedances of Impact-Relevant Thresholds

Pierrehumbert RT. 1995. Thermostats, radiator fins, and the local runaway greenhouse. J. Atmos. Sci. 52:1784– 806

Williams IN, Pierrehumbert RT. 2017. Observational evidence against strongly stabilizing tropical cloud feed- backs. Geophys. Res. Lett. 44:1503–10

Williams IN, Pierrehumbert RT, Huber M. 2009. Global warming, convective threshold and false thermostats. Geophys. Res. Lett. 36:L21805

Mauren 2016 Bias Correcting Climate Change Simulations - a Critical Review

We thank this reviewer for their useful review of our manuscript. Below we provide a point-by-point response.

Major points

#1 The authors motivate much of their work by referring to climate impacts as a function of a population's vulnerability. Given this context, it would be good if they could acknowledge the vast amount of literature (see e.g. IPCC WGII's work) there is that conceptualises climate risk and vulnerability, use terms such as >risk<, >exposure<, and >vulnerability< accordingly with more care, and back up their claims on the strong dependency of vulnerability of GDP with references.

We thank the reviewer for helpful comments. Following sentences are revised and added to the manuscript.

[L46-53] This was discussed in various assessment and reports such as US National Climate assessment and those by IPCC (Melillo et al., 2014;Wuebbles et al., 2017;Hoegh-Guldberg et al., 2018;Masson-Delmotte et al., 2018) and it is expected to have significant impacts on human society and health. More importantly, previous studies have analyzed the risk (Quinn et al., 2014;Sun et al., 2014;Lundgren et al., 2013), exposure (Dahl et al., 2019;Ruddell et al., 2009;Liu et al., 2017;Luber and McGeehin, 2008), vulnerability (Chow et al., 2012;Wilhelmi and Hayden, 2010) and susceptibility (Arbuthnott et al., 2016) of population in the current and warmer climates.

[L77-79] The greater impacts of extreme heat in economically less developed region in a warmer climate has been discussed in multiple studies (Marcotullio et al., 2021;Russo et al., 2019).

Related, minor text comments: 67 that is only one factor. Please avoid misunderstand by saying that the impact of climate extremes on different populations depends on a range of factors, including… 116 "risk" – likelihood? 351-2 "it is well-known that not everyone is equally vulnerable to extreme weather, with richer developed countries having more resources to deal with extreme events" – 408-9 "given underdeveloped countries' lack of ability to endure climate extremes" – problematic phrasing? Firstly – 'least developed' maybe? 'underdeveloped' may sound like the country is deficient and you're judging it for that but check e.g. ipcc wgII terminology or united nations. Secondly, it seems that many developing countries are already enduring more climate extremes than developing countries, and showing much more endurance than any developed country has in recent decades had to show. Thirdly it's too unspecific anyway – are you referring to resilience or adaptive capacity maybe? See general point.

[L71-74] The effect of climate extremes on different populations depends on numerous factors, including the level of economic development, with impacts of heat extremes being more severe in less economically developed countries (Diffenbaugh and Burke, 2019;Harrington et al., 2016;King and Harrington, 2018).

Terminology is replaced with less developed or relatively more developed.

#2 I wonder about the choice of cities (e.g. line 223) and how transferable your results regarding those are to other places. You talk of (line 244) representative cities – representative in which respect? And how do you know? Please specify or don't claim this. Further, a different choice of cities, including e.g. Australia with high GDP >and< high exposure to extremes, for instance, might give a slightly different picture; or including more continental cities might have impacted the impact of ENSO; etc. Also, I worry that these 15 cities give very few degrees of freedom for your EOF analysis. If the authors want to claim these to be representative, they should do a sensitivity study, repeating the respective analysis with 15 different cities, drawn randomly out of a suitable selection based on size and with some coverage. Or with all cities of a certain size, or the largest city in some region, or similar? If not, the authors should be careful not to overgeneralise the results. In any case, a description needs to be added as to how you have chosen these cities, so that the reader can assess the potential impact of selection bias on the results.

In the previous version, representative cities were selected based on population, but excluding the cities in the same county (e.g., Shanghai and Beijing are both included in top 15 cities, but Beijing is excluded since both cities are in China). However, in our current analysis, we changed this into the 15 most populated cities in the world to simplify the argument.

[L211-212] To investigate the impact of unforced variability on more regional heat extremes, we take the 15 largest cities by population (Fig. 2a)

We also thank the reviewer for suggesting a sensitivity test. We agree that this selection of 15 cities could generate uncertainty in our analysis. In that context, next 15 most populated cities around the world. Selected cities are Kinshasa, Lagos, Manila, Tianjin, Guangzhou, Rio De Janeiro, Lahore, Bangalore, Moscow, Shenzhen, Bogota, Paris, Jakarta, Lima, and Melbourne. Melbourne is not actually in top 15 (or 30), but we included this to reflect the reviewer's suggestion. Figure R1 and R2 shows the same analysis with the EOF analysis done in the manuscript, but for the 15 newly selected cities.

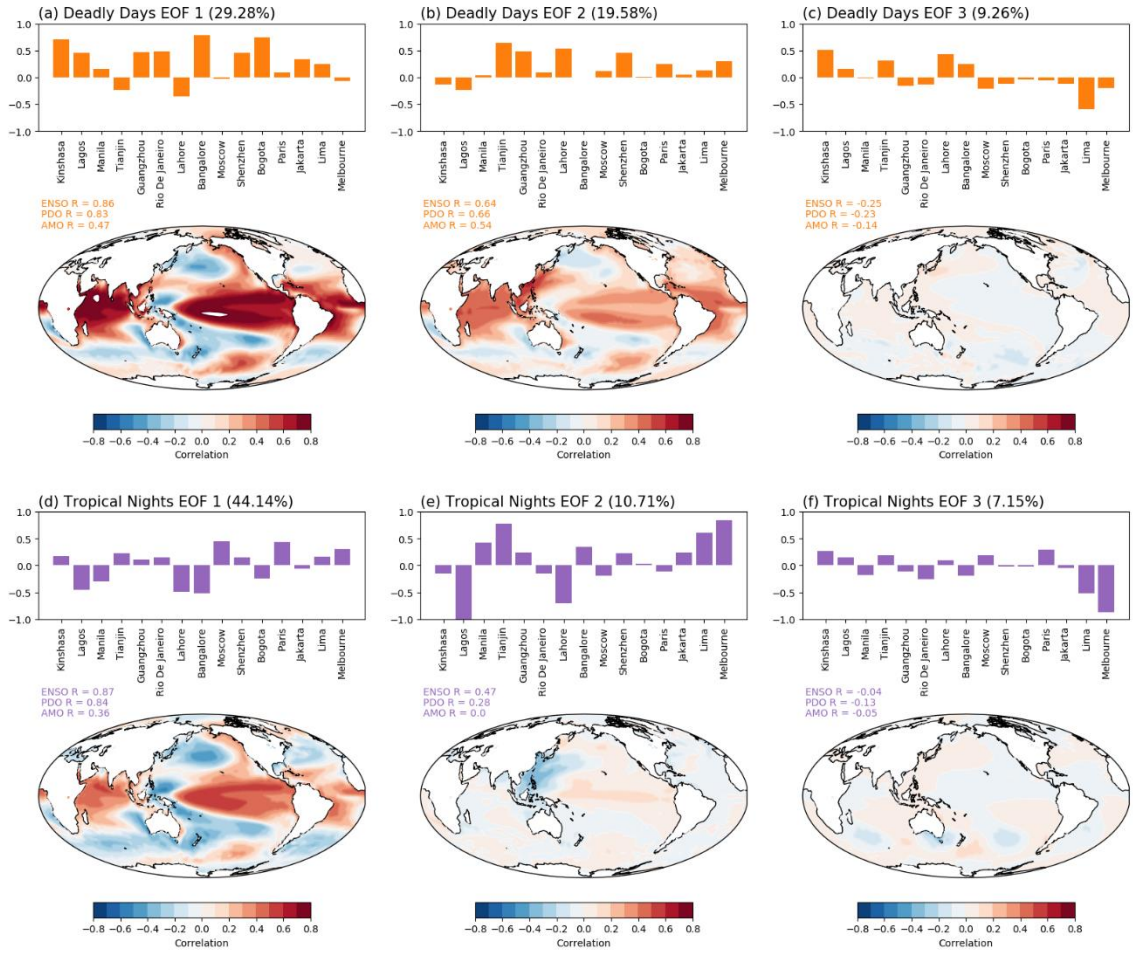

Figure R1. Same as Figure 3 (in manuscript), bur for the newly selected 15 cities.

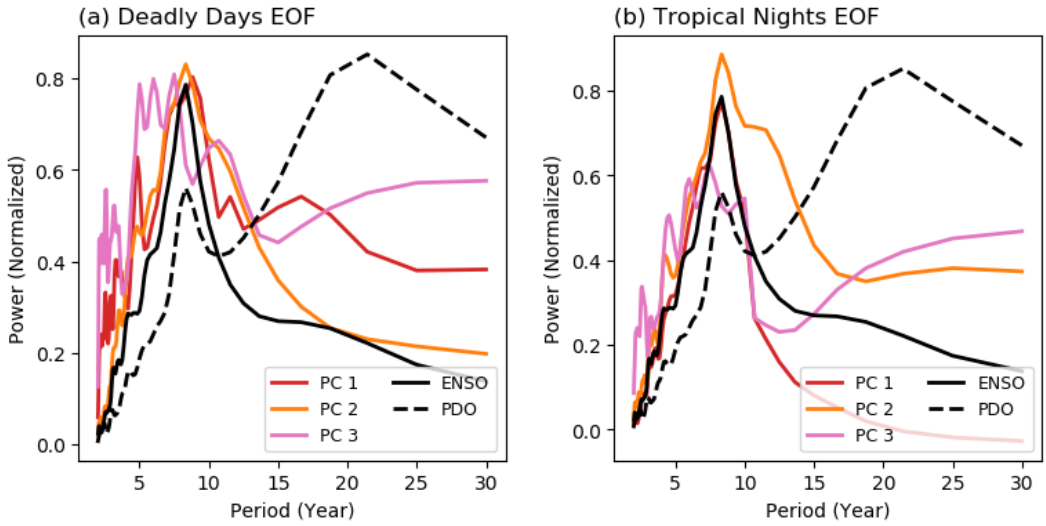

Figure R2. Same as Figure 4 (in manuscript), bur for the newly selected 15 cities.

As seen in the figure R1, first 2 modes of SST show ENSO-like pattern for both deadly days and tropical nights, similar to the current analysis in the manuscript. Figure R2 shows the

spectral analysis, which shows that frequency of pattern in Figure R1 is similar to that of ENSO, also similar to current analysis in the manuscript. As sensitivity test with another 15 cities showed similar results, we will keep the statement that the dominant driver of internal variability that impacts extreme heat event is ENSO.

Related – you explain the low absolute spread of days of extreme heat in Moscow with low absolute values in the mean, which makes sense of course. However, Moscow is at the same time the only city with truly continental climate, as far as I can see. It's also the only city at a latitude higher than 45deg. And most of the cities you consider are in the Tropics. All of these aspects will also shape how much each of these cities will be impacted by internal variability and on which timescales (atmospheric vs. SSTs). I think some of this should be discussed.

We agree with this point. Discussion is added.
[L269-270] Added: This may be a consequence of the fact that these large cities are mostly located near ocean and at lower latitudes.

#3 I am unsure about the bias correction, and I am not sure I can assess the method's correctness given unclear text. Generally: Given the authors have historical model simulations available, why not compare those with observational(ly derived) data for the same period, they should be much more comparable than the CO2 runs. The authors might otherwise end up correcting a bias that is due to the model not having all the forcings (ghgs, lu, aerosols, solar, etc), not due to the model systematically over/underestimating temperature extremes. 15 years also seems rather short.

Thank you for pointing this out. Due to the concern of bias-correction breaking the relationship between t2m and w2m, we have removed the bias-correction step in the present analysis, and instead just use the original model output in our analysis. In its place, we perform a sensitivity test to estimate how important biases in the model are for our results (Section 2.1).

Minor/further - the authors show the area-averaged bias and show it's near zero – (line unclear whether that is the bias of the averages or the average of the bias? T2m and d2m are corrected separately, and then combined, I'm not sure that is physical, the separate bias-correction might break the physical link between the variables, but that might not be an issue for the subsequent analysis. The bias-correction should further apply the same correction to each member, calculated from the whole ensemble, since it's a single-model w unperturbed physics; I think the authors follow this but not all clear from line 155-6.

The bias correction has been removed from the manuscript.

I don't understand what the authors mean in lines 177-8. Fig. 2e-f shows only data where the bias is greater than a threshold (lines 150-1); I don't see why? That will impact the average bias quoted. Can the authors further be kind enough to explain to me how the maximum value of what they show in those figures can be around a third. That means that the ensemble-mean bias is at every grid point (much) larger than the ensemble spread, right? Also for t2m; how does that go with their Fig. 1? That wouldmean really rather large bias? I think I am misunderstanding, so the manuscript could do with clarifications there.

The bias correction section has been removed from the manuscript.

#4 I think different baselines are used in the different parts of the study, w/o motivating this. e.g., 200-201 heatwaves defined based on PI. 380 CDD/HDD days also PI. Fig. 9 present-day baseline (description missing as to how the values compared to the present-day baseline is derived). I see that both baselines are useful, they are just pointing out different aspects (emphasis on all changes due to anthropogenic cc (well, co2 here) vs. changes still to come), but the differing usages should be noted & motivated in the text. Would the conclusions regarding CDD/HDD changes furthermore be more meaningful too for a PD rather than PI baseline?

We agree with this point we have changed the analysis on CDD and HDD to use a PD baseline.

[L394-396] Fig. 9 shows the percent change of CDD and HDD at 1.5°C, 2.0°C, 3.0°C, and 4.0°C relative to the reference period CDD and HDD values.

#5 The conclusions section is not comprehensive enough in terms of the limitations, uncertainties, and caveats (e.g., line 430 etc). Firstly, the results are from 1 model only, and another model may give different results, and the reality be yet again different. An obvious uncertainty here is the sensitivity of the model to CO2 (measured e.g. by TCR); please discuss this see e.g. Mauritsen, T. & Roeckner, E. Tuning the MPI-ESM1.2 global climate model to improve the match With instrumental record warming by lowering Its climate sensitivity. J. Adv. Model. Earth Syst. 12, (2020). Other aspects will be the model-simulated internal variability, that also varies substantially between models (spectral characteristics, magnitude, couplings). Then there might be uncertainty that should be discussed in how much the real 1.5/2/3/4degC given other climate forcings will be different from the 1.5/2/3/4-CO2 world, which might also be scenario-dependent. Then the paper would gain too from a discussion on what the assumption of unchanged socioeconomic and population (e.g. 113-4) distribution means for the results (I guess for instance that Fig. 8 will show even larger share of the global population exposed to heat extremes considering projected population changes. Please discuss this in the text? Of course there might be feedbacks too but they also will have important implications).

We thank the reviewer for suggesting this. The following points have been added.

[L434-441] Added: Uncertainties in this analysis include our use of gridded 6-hourly climate model output. More detailed analysis could be done with climate simulations with higher temporal and spatial resolution. The model has biases relative to measurements, potentially due to the fact that there are no aerosols in the forcing, which is another source of uncertainty. This was tested by adding the difference between the ensemble average and the reanalysis data to the model fields and recomputing the heat wave indices. In general, the impact of this bias was not important. In future analyses, this could be better resolved with use of multi-model ensembles or detailed bias-correction of the model.

[L449-450] Added: Also, this study could gain further insights by considering changing population and socioeconomic distribution in the future.

#6 It's a rather trivial and much documented fact in the literature that larger-scale aggregation reduces variability, and I don't see why/how this is worth as much

highlighting as done by the authors (in the text, e.g. key point #1; 312 “Notably” – Unsurprisingly? 334 “notable”.) More importantly, though, the authors conclude (line 339) that “this [int variabilty averages out over large averages] indicates that int var will play a minor role in determining global exposure to temp thresholds”, but they write earlier that ‘no one lives in the average’ as a motivation for their study, emphasising hence the need to look at smaller scales. Can they please explain (in the manuscript) how that goes together?

We don’t see what the issue here is. All of the statements we’re making are correct and do not (to us) contradict each other. However, we have added the statement below to hopefully clear up any misunderstanding.

[L408-409] Added: We find that both forced and unforced variability play a key role in extreme heat events, highlighting the necessity of considering both contributions to extreme heat.

#7 Many methods come as a surprise in the results section w/o being mentioned in the earlier methods section, and are then not sufficiently explained. Some variables, regions, indices, are not introduced either (see specific comments below). Please can the authors make sure to remedy this. Examples are 223/Fig 3a, info on how the values for cities are calculated (surrounding 3x3 grid boxes) should not just be in the caption but also in the text; 242, EOF should be in the methods too; 257-9, it is not explained in the text (partially in the figure caption only) how you calculate ENSO, PDO, and AMO. Etc.

Explanations of methods are added in more detail as the reviewer suggested.

[L212-214] determine the number of deadly days and tropical nights over time by averaging the 3×3 grid points surrounding the city, only including the land grid points.

[L391-394] Both CDD and HDD are calculated by averaging the CDD and HDD values of 3×3 grid points surrounding each city, including only land grid points. CDD and HDD values are then averaged for 5 years after global warming reaches each levels of threshold.

EOF analysis is relatively well-known method, so will add reference and only discuss how the EOF analysis is used, and what implication it will have.

[L239-242] To probe the statistical modes of variability affecting this ensemble spread and to identify the underlying physical mechanisms, empirical orthogonal function (EOF) analysis (North, 1984) was performed on the detrended and normalized time series of deadly days and tropical nights for the 15 cities. For each city, the 28 ensemble members are concatenated together (total of 28×150 years) in order for all ensemble to share the same EOF. In this way, we aim to find the dominant drivers of unforced variability that impacts heat extremes in the largest cities around the world.

Reference is added for ENSO, PDO, and AMO calculation.

[L256-259] Characteristic patterns for ENSO (Trenberth, 2020), PDO (Deser and Trenberth, 2016), and AMO (Trenberth and Zhang, 2021) are calculated for each ensemble using all 150-year of SSTs, and the pattern is averaged over ensembles to come up with a single ENSO, PDO, and AMO SST pattern for the ensemble.

Minor/other/text

There are very many typos/minor language errors that need to be fixed (e.g., line 42 “efforts and to”, 45 “value(H”, 47 “various assessments”, 48 “assessments and those by the IPCC”, 63 “links”, 65 “the PDO” and “the AMO” (following conventions in the literature; also should not be “the ERA-Interim” but in turn “the ECMWF”, “the CDF-t method” etc), 66 “investigated in” etc.) Some language is a bit too colloquial too (e.g., 72 “run”, 73 “runs”, 91 “go”, 106 better “provide” than “has”, 398 “stats” etc.). The tenses should be consistent too (e.g., 136 compare 137 compared), also 130 “Figure 1a and 1b” 143 “Figures 2a and 2b” etc.

All suggestions the reviewer made is applied in the manuscript.

- Title. I think w/ ‘forced variability’ you mean the long-term change in response to CO2? The variability itself too could change in response to forcing, so I find this not ideal

Title changed: The effect of forced change and unforced variability.

- Key Points. Need to add ‘in one model’ or ‘in the MPI-ESM model’ to all. #1 w/o reading the paper, I think that sounds like it’s the int var rather than changes in co2 that are driving extreme events. #2 ‘by using the large ensembles’ confusing here – that refers to ‘is shown’, not ‘to reduce’, I think, but everything you do is with a large ensemble anyway.

Key points are revised.
- Unforced variability of the climate system, primarily ENSO, plays a key role in the occurrence of extreme events in a warming world.
- Uncertainty of unforced variability becomes smaller as one looks at larger regions or at a global perspective.
- Increases of heat wave indices are significant between 1.5°C and 2.0°C of warming and the risk of facing extreme heat events is higher in low GDP regions.
-
- 46 regional extreme heat events and heat waves – isn’t a heatwave also an heat event?

[L45-46] Revised: Previous studies have reported that regional extreme heat events will not only be more frequent, but also more extreme in a warmer world.

- 52 risk ratio – of what? I think it can’t be listed like this, it’s a comparison method not an index in itself

[L55-56] Revised: risk ratio of population’s exposure to heat (Kharin et al., 2018)

- 55 probably unclear to the reader less familiar to heatwave indices how the mean of a heat wave is different from its amplitude and how the mentioning of consecutive-day differs from duration; so maybe this can be clarified

Thanks for pointing this out. However, this terminology is from the previous paper cited and are explained in the methods section so we will leave the text unchanged.

- 60-62 I don't follow this reasoning, can that be clearer please. Is it that the future change in risk is due to forcing, so to assess that it's important to tell how much was due to forcing in the first place, and not due to variability?

We are trying to make a point that both long-term forcing and the internal variability of the climate is important on analyzing extreme heat events. The text has been clarified to reflect this.

[L64-67] Climate extremes are always a combination of long-term forced climate change acting in concert with unforced variability (Deser et al., 2012). Thus, characterizing and quantifying both long-term change due to external forcing and the unforced variability of the climate system is crucial in assessing the future risk of extreme events.

- 66 side note - there is a large body of literature discussing to degree to which the observed AMO is actually unforced or not…

We thank the reviewer for pointing this out. We will not discuss point in detail, but put in a reference: Mann et al., 2021.
[L70-71] … the Atlantic Multidecadal Oscillation (AMO) (Zhang et al., 2020;Mann et al., 2021) have been …

- 72 install and operate? It also needs buildings in the first place.. not sure everyone has them everywhere, thinking of the most vulnerable in a society

[L75-79] Revised: For example, as temperatures go up, increased energy demand to cool buildings will be required (Parkes et al., 2019;Sivak, 2009) in metropolitan area. But this requires resources to both install air conditioning and then operate it. The greater impacts of extreme heat in economically less developed region in a warmer climate has been discussed in multiple studies (Marcotullio et al., 2021;Russo et al., 2019).

- 73/74 Which model? What model experiment? This information is not complete

We will briefly talk about this in the introduction, and later discuss in detail in section 2.1.

- 77/78 please clarify, could also mean the economic status within societies or the economic status >during< extreme events. Similar line 116 – sounds like the wealth defines the extreme event hazard like the global-mean temperature levels do

[L83-85] We also utilize per capita gross domestic product (GDP per capita) data to investigate how climate change impacts extreme heat events on different levels of economic status.

- 94 was that the model resolution? Then specify that please, that is what matters more I think that what data your analysis starts with

[L96-97] Revised: We analyze 6-hourly output with 1.875° × 1.875° spatial resolution, which is the native resolution of the model output, for land areas between 60°N and 60°S.

- 98 climate projections

[L101-103] Revised: Unforced variability in the climate system generates uncertainties in the projection of the climate by impacting the dynamic component of the climate, especially for extreme events (Kay et al., 2015;Thompson et al., 2015)..

- 100/101 can you please explain to the ignorant reader why using the ensemble allows you to estimate the effect of unforced variability (initial-condition ensemble, a range of realisations of internal variability, etc)

[L103-108] Added: One way to analyze the impact of unforced variability in climate system is to use an initial-condition ensemble. Each members of initial-condition ensemble are generated by perturbating the initial conditions of single climate model. This perturbation will then propagate to generate different sequence of climate, such as ENSO, PDO, etc. (Deser et al., 2012;Kay et al., 2015). In this paper, we use the ensemble to allow us to estimate the impact of unforced variability on temperature extremes.

- 103 how many RCP runs?

RCP runs have been deleted form the manuscript.

- 111 average or sum – confusing. 118-120 confusing too, total or per capita in the end?
-
[L141-144] Revised: The data represent the population in year 2015 at 30"× 30" spatial resolution, and we re-gridded to the 1.875° × 1.875° grid of the MPI model by summing the values in grid boxes surrounding the MPI grid centers.

[L147-150] Revised: These data are re-gridded from the original 5"× 5" spatial resolution to the MPI model's resolution of 1.875° × 1.875° by averaging the GDP inside the grid box. When doing this average, per capita GDP was weighted by population and also averaged over the 1990-2015 period.

- 127 should repeat annual her (or why repeat global?) - important since it makes of course a difference for the timing of reaching a warming level whether you look at monthly or annual exceedences
-
[L156-157] Revised: Global warming is defined as the global and annual average temperature increase compared to the average of first 5 years of the 1% run.

- 128 why not cite 3degC too? That's shown in Fig 3 etc

3°C added.
[L157-159] We find that ensemble- and global-average t2m reaches 1.5°C, 2°C, 3°C and 4°C occur in years 59, 76, 108, and 133 years, respectively, and reaches 4.6°C at the end of the 150-year run.

- 129 4.6. 236, 54, 53, 57, 50,50,51,52 etc. (be consistent)

Revised to be consistent.

- 140 near-land needs to be defined

This section about bias correction has been deleted.

- 143-5 syntax; difference in what?

This section about bias correction has been deleted.

- 151 why median, not mean? should make a negligible difference, but i think the mean makes more sense when it's about averaging out random variability

This section about bias correction has been deleted.

- 152 Do the authors mean grid-point with 'region'? Please avoid, that's confusing. Or clarify otherwise

This section about bias correction has been deleted.

- 153 13% - average across what? Which line in the plot is that referring to? unclear

This section about bias correction has been deleted.

- 161 realistically – 'reasonably'? 'reliably'? will be more correct

This section about bias correction has been deleted.

- 163 rh is not introduced. w2m not introduced.

This section has been deleted, however rh and w2m are introduced in section 2.1 in the current version.

- 167 i think the 'runs' is in common usage the 'experiment' itself, not the data created. And you don't correct the experiments so I would think 'bias-corrected data from the xx run' is more correct. But I also deem the word 'run' rather colloquial.

This section about bias correction has been deleted.

- l169 estimated by these >runs<, not by these >models< - the model can be run in other configurations with those other forcings. (I know anything can be a 'model' but here I think it will be understood to refer to the MPI-ESM.) Also, what about other forcings like land use change and GHGs other than CO2?

This section about bias correction has been deleted.

- 175 regions? Which regions? Not explained anywhere. Also – a difference of 0.5degC seems >very much< compared to 1.5, 2, 3 degC, not "very small" as you write!

This section about bias correction has been deleted.

- 176 the renalaysis data contains some >response to< aerosol forcing, not the forcing itself, I think. And 178 the effect of aerosols fixed at/to >that of the> 2003-2017 period

This section about bias correction has been deleted.

- 181-186 I find this more confusing than necessary. Please clarify. Also, your definition means that you have not only a grid-point dependence, but also a seasonal dependence, so a heatwave in summer must be hotter than a heatwave in winter to be a heatwave, right? Worth noting I'd say.

This section about bias correction has been deleted.

- 201-203 this is not very well explained. I think you mean that heat wave thresholds are different … because they are based on the respective values in pre-industrial times, this means that heat waves with the same index value will refer, region- and season-dependent, to heatwaves with different absolute values, which is from an adaptation/impacts perspective meaningful because there will exist some degree of adaptation to the existing climate conditions, but that there are on the other hand also physical limits that depend on the absolutely values of heatwaves, which is why you additionally look at the other indices? Then maybe that could be clarified.

Thanks for pointing this out. Following point is added.
[L185-189] Heat wave thresholds are different for each grid point because they are based on pre-industrial temperatures at that grid point. Combined with regional differences in the ability to adapt, this means that heat waves in different regions may have different implications for human society. We therefore also count the number of days each year with daily maximum w2m above 26°C, which we refer to as "deadly days".

- 238 another conclusion would be that in in this metric there is no feedback of CC on int var, right? Maybe noteworthy too.

This could be true, but we think more detailed analysis is needed to come up with this conclusion. So, we will not include this in current study.

- 239-241 distinguish what from what? Origin from mechanisms (what's the difference anyway) or unforced from forced (not written) or temperature (mean) from temperature (extremes); really not clear what you want to say.

[L237-239] Revised: Previous work has attempted to distinguish the origin and mechanisms of unforced variability of temperature and temperature extremes (Meehl et al., 2007;Zhang et al., 2020;Birk et al., 2010).

.

- 241 EOFs don't show physical mechanisms, please avoid reiterating this common misconception. 268 also, higher modes very unlikely to refer to a clear mode due to the orthogonality constraints etc

[L239-242] Revised: To probe the statistical modes of variability affecting this ensemble spread and to identify the underlying physical mechanisms, empirical orthogonal function (EOF) analysis (North, 1984) was performed on the detrended and normalized time series of deadly days and tropical nights for the 15 cities.

[L272-273] Added: Also, higher modes of EOFs are unlikely to refer to a single mode of climate due to the orthogonality constraints between each mode.

- 241 ensemble spread? 285, 342, 343, 344, 346, 367?, etc the lowest etc ensemble

>member< (or 'run'). It's only one ensemble.

Revised as suggested.

- 242 what does that 'separately' refer to?

[L242-243] Revised: For each city, the 28 ensemble members are concatenated together (total of 28×150 years) in order for all ensemble to share the same EOF.

- 244 is it a really driver or maybe rather a manifestation?

[L243-245] Revised: In this way, we aim to find the dominant drivers of unforced variability that impacts heat extremes in the largest cities around the world.

- 246 I gather you mean the bar charts with 'EOF patterns'? Slightly confusing here since traditionally one would expect to see a map (and you show maps too)

[L246] Clarified: The first three EOF patterns are plotted in Fig. 3 as bars.

- 261 shown above the map plot in each panel ('lower panel' sounds like d,e,f to me)

[L255-256] Clarified: Maps of correlation coefficients are plotted in Fig. 3.

- 281 so this classification groups based both on climatology (from the intercept) and magnitude of the response to CO2 forcing (slope), right? I think that would be worth clarifying somewhere

We thank the reviewer for pointing this out.
[L284-285] Added: With slope and intercept, we can characterize the heat indices of each grid point with response to CO_2 forcing (slope) and climatology (intercept).

- 282 observations? Isn't it model data? 290 'observed' too, please avoid to avoid confusion

[L290] Revised to "grid point".
[L291] Revised to "shows".

- 284 have you tried numbers of clusters other than 6? not suggest you need to do a sensitivity study now, but if you have, you should report on the results.

With 5 clusters, we found a distribution of cluster where cluster 5 and 6 are merged, and with 7 clusters, we found that cluster 6 is divided into 2 clusters.

[L286-289] Added: When using 5 clusters, we find that two clusters (the light and dark blue regions in Figure 5a) merge, and when using 7 clusters, we find that one cluster (the dark blue region in Figure 5a) divides into two separate clusters.

- 287 "as might be expected" – as a consequence of the methodological approach?.

[L292-294] Clarified: As might be expected from how we calculated the 16 variables for clustering, each cluster shows a different evolution of heat extremes in warmer world (Figure 6).

- 292 how do you know that that is the only reason or whether there is a differing response to the co2 forcing too?

[L296-297] Clarified: This is mostly due to low variability in these regions compared to polar regions, making it easier for a trend to exceed the heatwave threshold.

- 294 3.5 and 2.2., what do those numbers refer to?

Each refer to HWA_{t2m} and HWA_{w2m}. Manuscript is clarified.

[L303-309] For HWA and HWM, the rate of increase is similar for all clusters, with increases of HWAt2m and HWAw2m of 1.45°C per degree of global average warming and 0.85°C per degree of global average warming, respectively, and HWMt2m and HWMw2m of 0.66°C per degree of global average warming and 0.47°C per degree of global average warming, respectively (Figure 6e-h). The exception is HWAt2m in cluster 6. The large increase of HWAt2m in this region is connected to the strong global warming signal in high latitudes that has been predicted for decades and now observed (Stouffer and Manabe, 2017).

- 296-297 "and now observed (…, 2017)" – potentially misleading because Polar (Arctic) amplification has been observed since much longer than 2017

Stouffer and Manabe 2017 mentions that it was observed before 2017. We also mention that it has been predicted for decades.

[L307-309] The large increase of HWA_{t2m} in this region is connected to the strong global warming signal in high latitudes that has been predicted for decades and now observed (Stouffer and Manabe, 2017).

- 302 I struggle to see that. Isn't it rather cluster 6 that shows a more rapid increase beyond ~1.5degC, but not cluster 4, in Fig. 7j? 'these regions' would be 1-4 I think

Thanks for pointing this out. This sentence is deleted from the manuscript.

- 306 could you briefly discuss how the clusters you find link to existing knowledge about climate zones; do they make sense from a physical point of view?

These climate zones are discussed in Table 2 as cluster names. This is clarified in Figure captions.

Figure 5. (a) Clustered regions via K-means clustering. Characteristics of each cluster are listed in Table 2.
Figure 6. Evolution of each index averaged over each cluster. Colors are consistent with Figure 5 and Table 2.

- 308 please consider adding ‘global’ to population so it’s easier to understand you are now going from regions/zones to global

[L317] Added: We also generated indices weighted by global population.

- 310 ‘heatwaves lasting 131 days’ – not really, it’s 131 heatwave days, isn’t it, that’s not the not the same as 1 or more heatwaves of 131-day length. 64- days same. Also, can you please give a range for these values? Lastly, ‘shows that … will experience’ IN THE MODEL – are projected in the model to experience or similar

These are HWD, which is the duration of longest day of heat wave. So, it is heatwaves lasting for that days. Number of heatwave days are HWF, which is in Figure 7c, d. Otherwise, manuscript is clarified as reviewer suggested.

- 327 now you cite the number for 5%, elsewhere in the text you cite numbers for 10%. What motivates that? Doesn’t make for good comparison and seems arbitrary

We revised the manuscript to discuss 5%, 10%, and 50% for other metrics.

- 340 no one is exposed to thresholds

[L359] Revised: temperature above thresholds

- 341 I don’t understand. ‘climate realisations’ do you mean, depending on how internal variability will play out, people in different regions will >temporarily< be differently affected? Confusing as written

[L359-360] Revised: although different people may be affected in different realizations of unforced variability.

- 375, 384 “the” 15 cities (otherwise unclear whether you maybe looked at more and the statement applies to only 15 of them) or ‘the cities considered’, even better, not need to iterate it’s 15. Lots of typos/missing articles/wrong prepositions on this page, please fix!

This paragraph has been clarified.

- 348 I think this statement is to broad and needs to be more specific to the context (that it includes the share of the global population etc). As it is it sounds like a mechanistic statement about physical limits the magnitude of extreme events.

[L366-368] Clarified: Thus, this model predicts that the occurrence of extremes will soon be able to exceed values likely possible in our present climate for these metrics.

- 383 “also large compared to other cities” – please add "(not shown)" to avoid confusion?

This has been deleted from the manuscript.

- 384 4062 days make for a very long year!?

The unit for CDD and HDD is [days times Δ°C]. So 4062 days°C makes sense. However, this part has been deleted from the manuscript.

- 391 20degCs

Revised

- 389 IN A MODEL. Needs to be added

This point is mentioned throughout the paper.

- 397 significant regionally >and temporarily<? Int var impact won't change the cc-induced change in their likelihood profiles, will it?

[L413-415] But while the impact of unforced variability might be significant regionally and temporarily, it becomes less important when one looks at larger aggregate regions.

- 399 the global population

[L416] Added: global population-weighted statistics.

- 411 investigate >some< economic impacts of increasing heat extremes (heating and cooling is only part of that)

[L425-427] Revised: To further investigate some economic impacts.

- 417 the high relative cost? Or how are energy costs globally, would need a reference. And: high demand? The demand is much lower in the poorest countries than in the richer countries currently.

[L431-433] Revised: increasing CDD in a warmer world could be one of the factors driving increased economic inequity from global warming related heat extremes, due to relative high cost and need for energy in poorest countries.

- 419 how does the 6-hourly climate model output contribute uncertainty? Which? What temporal resolution would be better? and aren't all the indices in tab. 1 for daily (>6h) min/max anyway, isn't that a model output (tmin, tmax), calculated by the model over its time step (that will be less than 6h perhaps)?

[L434-441] Uncertainties in this analysis include our use of gridded 6-hourly climate model output. More detailed analysis could be done with climate simulations with higher temporal and spatial resolution. The model has biases relative to measurements, potentially due to the fact that there are no aerosols in the forcing, which is another source of uncertainty. This was tested by adding the difference between the ensemble average and the reanalysis data to the model fields and recomputing the heat wave indices. In general, the impact of this bias was not important. In future analyses, this could be better resolved with use of multi-model ensembles or detailed bias-correction of the model.

- 427 the model you used is already bias-corrected? In which way? I don't understand.

Also, I think the model data is corrected, not the model.

This sentence has been deleted.

- Fig. 1 Units missing x and y axis labels. the figure caption should be improved to include all the necessary information. Also: land and near-land ocean area: both together? How is near-land defined? And: why do you choose those areas?

This figure has been deleted.

Fig. 2 unit x axis. Why now call it MPI run and in Fig. 1 only run? If anything, turn around. I think the label 'internal variability %' on the x axis is an incomplete name of your variable. a)-d) can you have the same y axis range?

This figure has been deleted.

- Figs. 3 & 10: I would consider plotting the ensemble spread around the ensemble mean, to allow inferences at least by eye about the relative variations implied by this (e.g., Moscow low).

Figure revised as suggested.

- Fig. 4 Can you add the percentage of variance explained by each mode? Unit missing x axis barcharts. Please specify in the caption that you look at annual values.

Percentage of variance explained is now plotted in the title. Bar charts are unitless since the EOF analysis is conducted on normalized deadly days and tropical nights values. Caption revised to reflect reviewer's suggestion.

- Fig. 5 Why is the AMO not shown? Caption should be in methods. Is the model lacking multidecadal variability or is that just not impacting enough deadly days in the model? Should be discussed in the text.

Since the spatial pattern of PC-projected SST resembles that of ENSO and PDO, rather than AMO, we plotted only ENSO and PDO. Methods section has been revised.

[L262-264] All of the projections of deadly day PCs and projections of the first two modes of tropical nights shows patterns similar to El Niño-Southern Oscillation (ENSO) and Pacific Decadal Oscillation (PDO).

- Fig. 6 I would add that it's arbitrary colouring/see Tab. 2. "observed" – in the model I guess, so I'd avoid that term. Crossed?

Revised to reflect reviewer's suggestion.

Figure 5. (a) Clustered regions via K-means clustering. Characteristics of each cluster are listed in Table 2. (b) Zonal average of temperature increases at the time of 0.87°C (our reference period), 1.5°C, 2°C, and 4°C of global warming compared to pre-industrial baseline in the 1% runs. Temperatures are averaged over a 5-year period after each warming threshold is exceed in the model.

- Fig. 8 Don't you want to show the 95% also for a-d? I think that would be interesting too. The purple lines in e-f carry IMO a very powerful message. Languagewise, the second sentence of the caption is particularly unclear.

95th percentile line added. Caption revised to reflect reviewer's suggestion.

Figure 7. Changes of population-weighted heat wave indices as a function of global average warming. Each line denotes one ensemble member for different percentiles of population.

- Fig. 9 absolute number .. above present day? Sounds like relative, but think you refer to 'warming'. Please clarify in the text.

Colors represent the relative increase compared to present day, and numbers represent the absolute number of deadly days and tropical nights (not subtracted by present day). Caption revised to reflect reviewer's suggestion.

Figure 8. Increase in (a) deadly days and (b) tropical nights compared to the reference period (0.87°C warming), binned by percentile of GDP per capita at selected levels of warming compared to reference climate (calculated by subtracting reference values, shown as heatmap), averaged over the population within the GDP percentile (for example, averaged over population in 0~10 percentile of GDP), and over all ensemble members for 5-year window after each level of warming first occurs. Green text inside the heatmap represent the absolute number of deadly days and tropical nights in each level of warming.

- Fig. 10 Again, a more complete caption would be helpful. Like, change (in percentage) since pre-industrial in the model-simulated number of cooling degree days (CDD; blue) and heating degree days (HDD; red) in the 1%CO2 experiments after the time they cros the global-mean temperature thresholds of (a) 1.5degC, (b), …., (d) 4degC, respectively. Error bars ...

Figure caption revised.
Figure 9. Change (in percentage) of ensemble averaged cooling degree days (CDD; red) and heating degree days (HDD; blue) compared to the reference climate (0.87°C) in the 1% CO2 experiments at the time they reach the global mean temperature thresholds of (a) 1.5°C, (b) 2.0°C, (c) 3.0°C, and (d) 4.0°C, respectively. Error bars represent the standard deviation of CDD and HDD values between the ensemble members.