We thank the reviewer for the insightful comments and feedback. Below are our replies, with the original comments in red and the reply in black.

General comments:

1. **Missing supplementary plots**: throughout the text, some results are presented accompanied by *not shown* instead of pointing to a figure. This is the case at least six times, which is too many, in my opinion. I believe that in most cases, the accompanying analyses have been done and would be valuable assets as supplementary figures, so I would encourage the authors to include them.

   We agree with the reviewer that including more figures would benefit the readership. Thus, we included figures showing the surface soil moisture trends in swl-5.0 (Figure A2) and the spatially averaged vegetation expansion timeseries in other regions other than North America (Figures A6-8).

   We decided to not include the figures showing the circulation analogue results conducted with slp rather than zg500 as we thought they would be redundant and might indeed confuse the readership. This allowed to also streamline the methods section and make it less convoluted. Evidence of how both these variables can be used to identify circulation analogues can be found in Merrifield et al., (2017) and Horowitz et al., (2022), thus we ultimately thought unnecessary for us to repeat this evaluation.

   For similar reasons we also decided to exclude TXx from the analysis and focus on TX5x. This metric offers a more robust evaluation of annual temperature extremes and better approximates the length of heatwave events of societal relevance (Russo et al., 2019). Thus, we believe that also presenting results for TXx might be redundant and little value for the readership.

   Russo, S., Sillmann, J., Sippel, S., Barcikowska, M. J., Ghisetti, C., Smid, M., & O'Neill, B. (2019). Half a degree and rapid socioeconomic development matter for heatwave risk. Nature Communications, 10(1), 136. https://doi.org/10.1038/s41467-018-08070-4

2. **Justify/explain the choice of some experiments or regions in analysis**: In some cases, some experiments or regions are chosen for part of the analysis, but without really justifying why they were chosen. For instance, three of the experiments are used in Figure 5, or only one in Figure 6. In some of these cases,

adding other experiments in the supplementary material would also help in showing that the displayed experiments were not cherry-picked.

We thank the reviewer for the comment. While the choice of experiments to include in the figures might have seemed arbitrary, it was mainly driven by either graphical reasons or for conciseness. In the case of Figure 5, the three runs displayed were chosen to highlight the differences between projections showing significant cooling trends (swl-4.0 and swl-5.0) and one that does not (swl-2.0). As we discuss throughout the manuscript, the other runs fall in a spectrum between these two cases so, to improve graphical clarity, we decided to only include these three examples. Similar reasons apply to Figure 6, in which we only show the trends in swl-5.0 as the other runs either had very similar patterns (swl-4.0) or shown no significant changes (all the others, as seen in Figure 7).

We made the reasons behind these decisions more clear in the text now (e.g., lines 179-180 and 200-201).

3. **Show dynamic contribution**: In the analysis of atmospheric circulation analogues, a small, but not negligible, dynamical contribution is found (Figure 4, second row, third column). Thus, I think it would be important to also show the slp and/or z500 fields, and discuss the changes in these when comparing the late and early periods.

   The zg500 trends during TX5x events and for the whole summer period (JJA) have been added to the Appendix (Figures A5 and A6). The discussion of their implication is added at lines 198-206.

   To further illustrate this, we calculated trends in zg500 and sea level pressure (slp) during the TX5x days and over the whole summer period (JJA) in swl-5.0 (Figures A4, A5). The geopotential height shows a slightly negative centennial trend (5-10 m/century) over the studied region, which is reflected in the slp with an increasing pressure trend over central Canada and upstream of the zg500 anomaly centre (Figure A4). While the zg500 and slp trends identified here could suggest an increasing dynamical contribution to the projected TX5x events (Merrifield et al., 2017), their magnitude seems insufficient to induce a noticeable change in the forcing associated with TX5x events. Moreover, even weaker trends are associated with the mean summer zg500 and slp fields (Figure A5), suggesting that no substantial changes in the mean atmospheric circulation occur in the run. Thus, it is unlikely that such small trends in the pressure fields can explain the sharp changes in TX5x projected throughout swl-5.0.

4. **Robustness of linearity and trends**: in many figures, linear fits are used, and trends are computed, but it is not often stated how well the fit actually explains the data. For example, trends in Figure 3 or Figure 5 would benefit from some mention of the reliability of the linear model (either directly on the plot or somewhere in the text).

   We added a mention of the statistical significance of the trends to both figures 3 and 5 captions.

5. **Add some more discussion**: the discussion of results is comprehensive and clear. Yet, a few more lines could be added. For instance, on the limitations of using stabilised warming levels compared to warming levels in a present-day/future transient climate.

   In this study, we do not use a global warming level approach, but we analyse trends within each of the net zero CO_2 emissions runs. The "warming level" identifiers refer to the warming level at the beginning of each zero emissions run and are used as labels for e.g. net-zero emissions at 2 degrees above pre-industrial, rather than warming levels in the strict sense. Thus, a comparison between climate responses under stabilised and transient conditions is not the intent of this study and lies outside the scope of this work. The limitations of the model protocol we used are discussed in the results and conclusions sections. We more explicitly acknowledged the difference between stabilised and transient runs at lines 27-28:

   Indeed, previous studies have shown how climate responses at comparable levels of warming can significantly diverge between stabilised and transient climate states (King et al., 2020; Dittus et al., 2024).

Specific comments:

1. Abstract, lines 7-10: three sentences in the abstract are used to present the results on mean temperature changes. Given that these are not the main focus of the article (and not particularly original considering the existing literature), I would recommend condensing them into one sentence or leaving them out entirely to focus on the changes in heat extremes.

   Shortened following suggestion.

   We find that the global average surface air temperature (GSAT) keeps increasing in all zero CO2 emissions UKESM1.2 projections. The increase is more *pronounced at higher warming levels and mostly occurs in the Southern Hemisphere, while the Northern Hemisphere land experiences a slight cooling trend.*

2. Line 23: Given the wide relevance of the temperature threshold, it should be clarified how *1.5°C above pre-industrial* is defined (which, in this case, I guess it is just one year above the threshold).

   To clarify the definition, we added "annual" to line 23.

3. Line 31: The wording *is long established* is a bit misleading. It seems to suggest that the post-net-zero global temperature evolution is better understood, and for a longer time, than the literature suggests (as mentioned in the next sentence). I would recommend providing a more nuanced statement and possibly mentioning some of the early work focusing on net-zero responses (e.g. Mathews & Caldeira, 2008).

   Wording changed to "well understood" and suggested reference added.

4. Lines 51-52: Two things here might lead to a minor misinterpretation of the TIPMIP setup. First, it reads as if 80 Gt CO_2 are imposed by the protocol, but it instead asks to find the emissions rate that leads to the desired warming trend. Second, while the goal is technically 2°C after a century, the protocol aims for 0.2°C/decade, which is a slightly different goal, as it carries a more specific guideline on the precision of the rate of warming.

   Wording corrected to make it more adherent to protocol.

5. Line 69: maybe mention the method to estimate the linear fit?

   Added "least squares".

6. Line 92: I cannot find a previous mention of the piControl run. It might be good to do that when explaining the analysed simulations, and mention the length of the experiment, which is relevant here.

   Added details at line 58, where piControl was first mentioned.

7. Line 106: the *residual* mentioned here can be confused with the residual contribution (thermodynamic) mentioned a few lines later. Also, how large is *e*, and does it affect the calculation of the dynamically induced temperature component?

   The terminology has been changed so that this term is now called "regression intercept". The size of this term varies depending on the level of resemblance of the analogue to the target pressure field. As stated at lines 111-112 and indicated in Horowitz et al., (2022), to ensure that only sufficiently good analogues are considered we filter for those with a dynamical temperature component above the 85th percentile.

8. Is it mentioned somewhere with respect to what period exactly the anomalies and trends are calculated? (i.e. is it just the value at emissions cessation, or some window around it?)

   Trends are calculated starting from the value at the branching date of each stabilised warming level. Specification added at line 75-76.

9. Line 118-121: a few intervals are mentioned here; are these covering the full range across all experiments, or some other measure of spread across the simulations?

   Ranges are expressed relative to all relevant ensemble members.

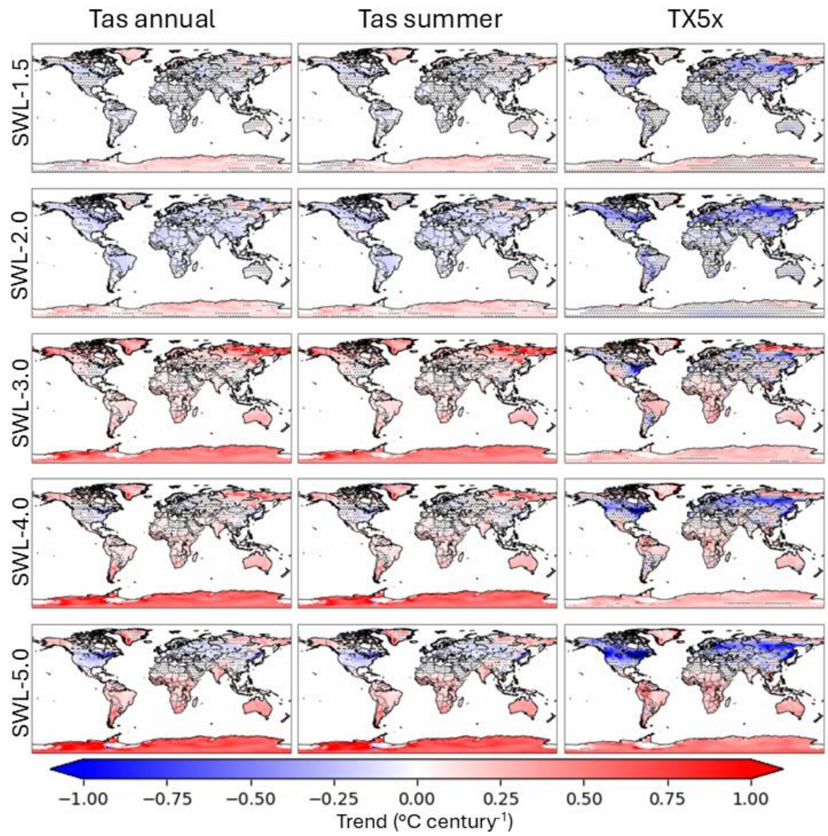
10. Lines 124-126: it may not be as clear as in the net-zero UKESM simulations, but the simulations from King et al. (2024a) also seem to show trend dependence on the global warming level. This is in part because the trends there are smaller, but also the ACCESS-ESM runs seem to cover a smaller range of global warming levels. I would rephrase the sentence here to better reflect that previous studies also show a similar tenden cy (i.e. stronger warming trend in warmer equilibrating climates).

    Rephrased to better reflect their result.

    However, while their experimental set-up did not exactly correspond to what we analysed here, they found a significantly weaker warming trend (0.03-0.05°C per century) and reported an apparently lower warming-level dependency on the trend magnitude.

11. Line 138: the *widespread increases over land masses* are barely noticeable in the presented figure, due to the colorbar chosen, as changes are often much larger over high latitude ocean regions. Maybe an alternative supplementary figure with only land masses (and a colorbar covering a smaller range) could help, but at least this (the differences in magnitude of change between regions) should be acknowledged more consistently in the text.

    We added a supplementary figure showing trends over the land masses with an enhanced colourscale.



12. Figure 2: I would label the subpanels and use the labels when referring to the results in the text (e.g. line 149), since it sometimes takes a while to find the plot that is being described.

    Changed throughout following suggestion.

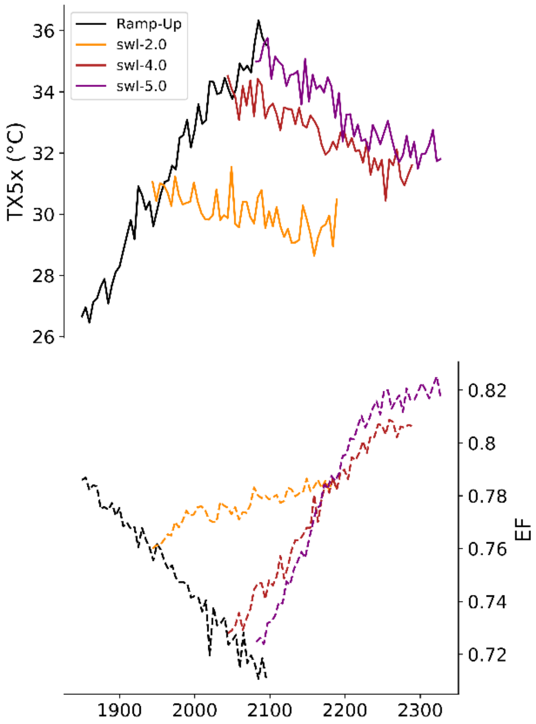
13. Figure 4: it is unclear to me how the repeated subsampling (30 times selecting 20 events in each period, as mentioned in the Methods section) is reflected in this figure. Would there be a way to show the robustness of the found patterns?

    The repeated subsampling is performed to reduce the error induced by sampling imperfect analogues and a small number of events, which is more likely to be "averaged out" with resampling. While this isn't easily shown in the final figure, the robustness of the methodology using these parameters (i.e., 30 times selecting 20 events and drawing from a piControl of 500 years) has been previously assessed by Deser et al., (2016) and Horowitz et al., (2022). The reader is referred to those works for further details on the sensitivity analysis of the methodology.

14. Figure 5: This figure is a little busy, so it might be worth considering separating some of the experiments into different panels. In any case, it might be important

to mention the large variability somewhere in the text, since it currently reads as if the relationship is much clearer than what the figure really shows.

We have changed the figure to improve clarity. Now TX5x and EF are plotted as individual timeseries to facilitate their comparison over time. The correlation between both variables is now more clearly visible.



15. Lines 211-213: Maybe a longer explanation on why this is the case might be relevant. Also, if winter temperatures are most relevant, I wonder if a supplementary plot for winter temperature changes might also be a nice addition.

    While the available literature indeed suggests that winter temperatures may be more important in driving biome shifts due to the vegetation sensitivity to yearly minimum temperatures, we discuss further down how we could not find a direct link between changes in local mean temperatures and the vegetation shift, as all regions show quite different temperature regimes (lines 251-261). As we also specify here, a deeper analysis of the drivers of this vegetation expansion in UKESM1.2 lies outside the scope of this study.

16. Line 214: I think that experiment *dn1.0* has not been properly introduced in the methods beyond saying that experiments with negative emissions are also used.

Added an appropriate introduction of this run at lines 63-64.

17. Lines 223-224: It reads at the end as if the negative emissions values go fully back to the values in the ramp-up, but the figure suggests that there is still a large fraction of hysteresis.

    Changed

18. Figure 8: the legend of the left panel is partly cut out.

    Corrected

Typos/language suggestions:

1. Abstract, line 6: comma missing after *pre-industrial*.

   Added

2. Line 29: both CO_2 *emission* and *emissions* are used throughout the text. I would stick to *emissions*.

   Corrected

3. Line 30: How are references ordered? I would recommend either temporally or alphabetically, but none seem to be used consistently in the manuscript.

   Changed so that they are now ordered temporally.

4. Line 37: I would say *could have* instead of *has*.

   Corrected

5. Line 53: *net-negative projections* can be misleading; use *net-negative emissions projections* instead.

   Changed

6. Line 70: The *them* could refer here to both the heat fluxes and the temperature extremes.

7. Line 113: change *events* to *event's*.

   Changed to events'.

8. Line 190: *observed* might be misinterpreted here, as if it would refer to observations.

   Corrected throughout the document.

9. Figure 3 caption: change *wit* to *with*.

Corrected

10. Line 199: correct *A2* to *A1* (and vice versa in line 250).

    Corrected throughout the document

11. Line 202: *striking* is a bit too subjective; I would prefer more neutral language.

    Changed

References:

- Matthews, H. D., & Caldeira, K. (2008). Stabilizing climate requires near-zero emissions. *Geophysical Research Letters, 35*(4). https://doi.org/10.1029/2007GL032388