Response to Referee 1

This study explores emergence of climate extreme indices in the CMIP6 ensemble. A robust early detection of temperature extremes, especially at low latitudes, is found. Also, some emergence of increasing precipitation extremes in northern high latitude regions is found.

Overall, this is a comprehensive study that is well written. My comments are mostly minor, however, I do raise a few methodological queries regarding the KS test and detection of changes in bounded/non-gaussian distributions (see specific comments below). I also think the novelty of this work isn't as clear as it could be and some edits to address this would help readers.

We thank the referee for their thorough read and encouraging comments on our work. We are particularly grateful for the constructive comments, which we believe have helped to improve the quality of the article. Below we reply to each comment in detail.

1. *Abstract: The first sentence describes extremes as "effects" which sounds strange. The last sentence uses the term "often quite high" which is vague. In general, while the abstract is short already the text that is already there could be more concise (e.g. removing "more and" in the first sentence) and additional space could be used to provide more insight into the results.*

   This is a good suggestion, and we have tried to both trim the abstract of superfluous wording and add more specificity on the results.

2. *Introduction: This is well written but is a bit too brief and doesn't specifically discuss what has been found in previous analyses on emergence of extremes (e.g. Bador et al., 2016; Harrington et al., 2016; King et al., 2015) and what this specific study adds to the literature. I note these references are mostly referred to later in the paper, although Bador et al. is hopefully a useful addition. On face value the main novel aspects of this work are its comprehensive nature (examining all ETCCDI indices), its use of CMIP6 models, and the model weighting approach, but I'm wondering if I'm missing something else. There's also been work to explore emergence of fire weather metrics which may be worth discussing (Abatzoglou et al., 2019).*

Thank you for this comment. We agree that the motivation and novelty of this paper could have been made clearer. We have rewritten the introduction to put our work more into context, added suggested references from both referees and stated the objectives of this study. We have also added a few more references to other relevant literature.

3. *Page 3, L65-71: I think some discussion of the fact that KS tests for emergence estimation have been employed previously would be worthwhile including (King et al., 2015; Mahlstein et al., 2011, 2012). There are other methods that could be applied to extreme indices, such as signal-to-noise ratios (Hawkins et al., 2020) or probability ratios (Harrington et al., 2016; King et al., 2016), so some discussion of why those methods aren't suitable are preferable here might be of use.*

   The methods section has been expanded by a short overview of other techniques for calculating emergence and relevant literature. We have added that our choice of the Kolmogorov-Smirnov test was a trade-off between having a non-parametric, versatile test and limited computing resources given the large and diverse data set.

4. *Page 3, L69-71: Could you clarify if the autocorrelation test is performed on the raw index values or if a detrending is performed first? I suspect it wouldn't make a large difference but it would be worth clarifying.*

   We have added that this was done without detrending the data first.

5. *Page 3, L82: Perhaps change "immense" to "high".*

   Done.

6. *Page 4, L97: The broadening of this distribution is interesting and I suppose could be to do with a higher rate of warming or increasing variability or a combination of the two. This is probably beyond the scope of your study to explore.*

   We agree that the reason behind this change would be interesting to explore, but we will leave this to future research.

7. *Page 5, L108: Too many parentheses on the reference.*

   Fixed.

8. *Section 2.2.2: If I understand correctly, the same model weighting is applied at each gridcell- this could be a bit clearer perhaps. I suppose this might mean a regional ToE analysis would involve apply a different model weighting and potentially have different results. I think this would be worth mentioning as a minor caveat.*

   This is a very good point, and we have added that weights are constant across all grid cells and that regional differences in model performance are thus not considered.

9. *Results: This is a very clear and well written section. I might have missed something with the TN10P calculation but I don't understand how results below -10% are possible (as shown in Figure 6a-c) given it is an index that is bounded and starts off with a climatological value of 10% and a possible minimum of 0%.*

   TN10p is calculated with respect to the base period 1981-2010 and denotes the percentage of cold nights in a year with night temperatures below the 10th percentile, which is derived from this base period. That means that over the entire the base period, 10% of nights count as cold nights (but in any individual year in the period, the number can be different from 10%). Since temperatures in the preindustrial period 1850-1900 are generally colder than during the base period, the percentage of cold nights here can be (and likely is) larger than 10%. The results in Figure 6 are all relative to the preindustrial period and thus changes of more than 10% can arise. We have added the following clarification to the text: "Note that reductions in TN10p greater than 10% are possible, because we present changes relative to the preindustrial period, which is colder than the 1981-2010 reference period used to define the 10th percentile (the same holds for TX10p)."

10. *More generally, I would wonder about KS test performance with the TX/N90/10P indices given the bounds and the fact that the KS test is sensitive to changes in the width of the distributions. This might be worth commenting on especially given earlier emergence relative to the absolute indices.*

    The KS test is generally sensitive to all changes, as it does not require any knowledge about the underlying probability distributions, apart from being continuous. As mentioned in the methods section, some indices are more likely to have a discrete distribution and the power of the KS test might therefore be lower, and the same holds for distributions with a bounded support. There is a range of variations of the general KS test for different types of data, intended to increase its power (e.g., Frey, 2020, Dimitrova er al., 2020 for the cases mentioned above). However, they usually require numerical

computation of the critical values based on a large number of simulations, which was not feasible in our study. If computing resources were not an issue, a permutation test would probably be the best universal option. We have added text to the methods section that also mentions the limitations of the KS test in these cases.

11. *Page 9, L184: "patters" should be "patterns".*

    Fixed.

12. *Seasonal results: It might be worth noting that the impacts/relevance of different indices may differ when performing seasonal analysis, e.g. JJA TNn ToE in Northern Hemisphere locations may be of less relevance to impacts than TXx or TX90P.*

    We agree that the potential impact and relevance of particular seasonal indices differ between regions or hemispheres. The main focus of this paper is on annual extreme indices and we only show two indices on seasonal scale. This is meant as a brief showcase of seasonal results and to compare our results to those of other papers, such as King et al. (2015). Relevance also differs considerably between different fields or applications. As such, we do not claim to have picked the most impactful indices and seasons, but they might be of interest to some. We have added a couple of sentence to this section explaining the choice of indices.

13. *Page 15, L310: I wonder if this finding is also related to the skewed nature of Rx1day and the lower bound of 0. Perhaps this makes detection of declines harder.*

    This is a very interesting observation and worth a deeper look in future research. As stated in point 10, the power of the KS test might be limited in these situations, so it would be interesting to investigate how other tests compare. However, annual values of Rx1day are likely to be considerably larger than zero (expect for arid regions or very dry years), so we wouldn't expect the lower bound to affect the test.

14. *Summary: This section is also well written but a little brief. Some discussion of consistencies with other ToE studies, including analyses of mean changes, may be helpful, as well as clearer elucidation of what has been learnt from this paper specifically.*

    We have expanded the summary, which now contains short comparisons and

references to other studies, both for emergence and mean changes. We have also added a paragraph highlighting that we intend to give a broad overview and encourage further research into the individual features identified in this study.

Response to Referee 2

The authors present future changes of extreme climate indices at the end of the 21. century 2070-2100 and they present the 20 year time slice, when these changes are starting to be robust [ToE]. But they do not make a connection between the two, which would be interesting and new.

The paper is very descriptive and does not give any physical explanation, why regional differences occur. An explanation would be helpful as well as an example for an impact on society which is mentioned in the introduction. The results are sufficiently presented, only interpretations and conclusions of the results are sparse in the paper.

Overall it is well written, but in some paragraphs additional text and explanation would be helpful for the reader.

We thank the referee for reading our paper in detail and providing a comprehensive review and suggestions, which helped us with improving the manuscript. Though we agree that physical explanations of regional differences would be very useful, they are not the focus of this paper and require detailed analysis of relative differences of climate variability and trends between regions. Instead, we try here to present and describe general patterns in the time of emergence for a large set of climate indices, which will hopefully motivate more detailed analysis of individual features in the future. The introduction and summary have been rewritten to make this clearer. Below we respond to detailed comments and describe adjustments we have made in accordance with them.

1. *Chapter 2.2.2: Why not explore the new ensemble from Merrifield, A. L., Brunner, L., Lorenz, R., Humphrey, V., and Knutti, R.: Climate model Selection by Independence, Performance, and Spread (ClimSIPS v1.0.1) for regional applications, Geosci. Model Dev., 16, 4715–4747, https://doi.org/10.5194/gmd-16-4715-2023, 2023.*

   Thank you for this suggestion. We now refer to this ensemble in the text and agree that model selection is a useful alternative or supplement to model weighting, however the data and computational work underpinning the current manuscript is in fact quite substantial and it is not feasible redoing it for this ensemble. The methodology presented here can be expanded to this or similar ensembles at a later date. We

anticipate that combining it with a corresponding ensemble of upcoming CMIP7 models will be an interesting application.

2. *The paper is missing connections to the previous studies concerning the results of extreme climate extreme indices of the CMIP6 ensemble. In general there are only a few references. Extreme climate indices based on the CMIP6 model ensemble have been publish already as well as the robustness of these changes for different time slices and different warming levels. (e.g. IPCC Atlas, Schwingshackl, C., Sillmann, J., Vicedo-Cabrera, A. M., Sandstad, M., & Aunan, K. (2021). Heat stress indicators in CMIP6: Estimating future trends and exceedances of impact-relevant thresholds. Earth's Future, 9, e2020EF001885. https://doi.org/10.1029/2020EF001885, Coppola, E., Raffaele, F., Giorgi, F. et al. Climate hazard indices projections based on CORDEX-CORE, CMIP5 and CMIP6 ensemble. Clim Dyn **57**, 1293–1383 (2021). https://doi.org/10.1007/s00382-021-05640-z)*

   We agree that the framing and context of the paper could have been made clearer and have therefore rewritten the introduction to put our work more into context, added suggested references from both referees and stated the objectives of this study. We have also added a few more references to other relevant literature.

3. *Page 5, Line 132: To expand the results of the research, it would be interesting to also use SSP3.7 and those have not been published as much.*

   We agree and have added the ToE results for SSP2-4.5 and SSP3-7.0 to the supplement.

4. *DTR , it could be helpful for the interpretation to look at it seasonally.*

   Indeed, it would be very useful to look at seasonal results for DTR, which might help to explain the patterns we see in the annual results. However, the main focus of the paper is on annual extreme indices and we only show two seasonal indices as a brief showcase and to compare our results to those of other papers, such as King et al. (2015). We hope to come back to exploring time of emergence on seasonal scales for more indices in future research.

5. *Precipitation: you can have a positive or negative change, how do you combine this with the ensemble results of ToE?*

   Time of emergence is computed for each grid point and each model/ensemble member

separately and only then aggregated over regions and models. This means that the KS test was able to detect changes related to both negative and positive changes and both are included in the overall ToE results (as well as changes in variability and so on).

6. *Concerning the change of precipitation extreme indices it could be helpful to additionally look at the monsoon season in the areas experiencing monsoon.*

   While we agree that it would be helpful to dive deeper into the physical explanations of the results we are seeing, such focus on particular regions and the monsoon season are outside the scope of this article. Our aim is to give a broad overview over a large set of climate indices, which we hope will inspire further work into individual features (we have highlighted this now in the summary). The underlying data set includes relevant extreme indices on monthly resolution, which can be used to go into depth on the topic of the monsoon season. For such a study one could even consider employing a model weighting scheme that focused more on the correct capture of monsoon patterns.

7. *The figures are too small, add median to the subtitle and longitude/latitude*

   Thank you for noticing this, we have scaled the figures up a bit, which should hopefully make them easier to read. As for the additional plot annotation suggestions, we think that they would likely lead to more cluttering of plots that are already relatively packed with information and that more information would defeat the readability improvement gained by making them larger. We thus prefer to keep the detailed specification in the figure caption, where it can be explained in full and is less likely to confuse the reader.

8. *Page 1, L 11-12: what do you mean by 'although between-model variability is often quite high' ? Could you add a little more information?*

   We have rephrased this to "… however there is substantial disagreement between models about whether or not emergence occurs elsewhere."

9. *Page 2, L 30-33: Start the sentence with precipitation to make it more clear.*

   We have reordered the sentence so that precipitation is mentioned earlier: "Internal variability also contributes significantly to the delayed emergence or lack of emergence of the anthropogenic signal in long-term regional mean precipitation changes (IPCC, 2021c), as demonstrated by multiple lines of evidence such as global projections from multi-model ensembles and SMILEs ."

10. *Page 4, L 90: add 'computer' to power limitations*

    This actually refers to statistical power rather than computing power. We have rephrased this sentence.

11. *Page 6, L 138: you explain the hatching, but it is not clearly written, use the text from figure 3.*

    This has been changed accordingly.

12. *Pages 6/8/10, L 143/176/196/ ….., do not use trend (this is also a statistical measure), better increase*

    We have edited the sentences to be more precise about whether we are talking about an increasing or decreasing trend, but we do feel that "trend" is a reasonable word to use, since the plots do show clear trends, even if they are not statistically quantified. To be more precise, we have amended the text as follows:
    Line 143: "this trend continues" -> "this warming trend continues"
    Line 176: "this trend again continues" -> "this decreasing trend again continues"
    Line 196: We have removed "trend"
    We have also clarified the language in several other places.

13. *Page 6, L 145: add median*

    We have added "for the weighted multi-model median" to the end of the sentence.

14. *Page 6, L 147: this still belongs to figure 3, why start a new paragraph, or add a break in line 140 too.*

    This has been fixed.

15. *Page 7, Caption of figure 3: Replace : 'available' with selected*

    We have removed the word "available" in all figure captions.

16. *Page 7, L 152: you talk about the Northern Central America region , it would be better to say that this is IPCC AR5 reference region and perhaps add the figure*

from, https://github.com/IPCC-WG1/Atlas?tab=readme-ov-file#new-reference-regions, than it is clear and not clear for the rest of the text e.g. line 256 I miss understood at the first time.

We now realize that it needs repeating here, thus we have amended the text at the beginning of the results section. We have also included a direct link to the figure.

17. *Page 7, L 153: add paragraph*

    We do not feel that adding a new paragraph here is appropriate - it would only be one sentence long and is a continuation of the theme in the previous paragraph.

18. *Page 9, L 186: I think something is missing in the sentence before the comma.*

    We have rephrased this sentence to "When comparing results for these percentile-based indices to the absolute temperature indices above, ..."

19. *Page 10, L 188: I do not agree, only the distributions differ earlier, but you are comparing two different indices*

    We have removed this sentence.

20. *Page 11, L 209: delete 'strong' it is not correct*

    We have rephrased the sentence as follows: "Results for ToE in Fig. 7(c) show that emergence occurs first (as early as the 1970–1990 period) in the NH mid-to-high latitudes and central Africa for both scenarios, and then later in ..."

21. *Page 11, L 214: what do you mean by 'but future changes could be even more drastic',*

    We are trying to say that early and historical emergence does not mean that changes stop or are less dramatic after emergence. We have expanded this sentence a bit to make this point more clearly: "However, it is important to note that early emergence does not mean further climate-induced changes do not occur, in fact they continue to occur and can be even more dramatic in the future."

22. *Page 11, L 220: perhaps rising instead of 'warming'*

    We have changed this.

23. *Page 11, L 226: 'strong': I do not agree*

    We respectfully disagree. Rx1day increases by more than 20% in SSP5-8.5 - this seems like a strong trend to us.

24. *Page 11, L 228: Sahara and south of it*

    We have amended this sentence as follows: "… with the largest percentage changes of more than 50% occurring in and a little south of the Sahara, and in northern Greenland (SSP5-8.5)."

25. *Page 11, L 235: What do you mean by 'but sign o model disagreement ….', please rephrase*

    We rephrase as follows: "However, there are many regions that do not show emergence for the weighted multi-model median, but nevertheless show high levels of disagreement between models about whether emergence occurs (hatching), which may hint at potential emergence after the end of our data set in 2100."

26. *Page 15, L 297: I think introducing CMIP5 is not helpful, you have a lot of results already and CMIP5 and your specific ensemble are not comparable. You could e.g. mention something like 'our findings are in line with previous studies by King et al. 2015'.*

    We have edited this section to make it more concise, nevertheless we feel that the King et al. (2015) study is the most similar to ours, as they use the same technique and a similar dataset, and it is worth comparing the broad results.

27. *Page 15, L 305: experience an decrease instead of 'see a negative change'*

    Edited as suggested.

28. *Page 15, L 306: delete 'yet'*

    We have amended this to "quite a different".

29. *Page 15, L 310: 'seem to translate into emergence', please revise*

    We have amended this to "seem to lead to".

30. *Page 17, L 333: delete 'trends'*

    We have amended this to "changes over time".

31. *Page 17, L 340: This does not make sense or I did not understand the method, this means that non of the 20 year time slice have the same distribution like 1850-1900. But I think this should be the case in 1850-1900. I did not see the code, how you calculated the TN10p I only found the download link. But it could be possible, that your sample size is too small for such kind of analysis.*

    You are correct, it would be impossible to detect emergence during the 1850-1900 time period, as the distributions would be identical (not considering the uncertainty of the KS test). Thus, the first of the rolling 20-year periods considered for emergence starts in 1901 (moving forward by one year in each step), which does not overlap with the base period and hence can have experienced emergence of a new distribution for the extreme index in question, from which the distributions never recover until 2070. The resulting time of emergence is the first year of the corresponding 20-year time period. For the plots, we have binned the results, this means that emergence during 1910-1930 can refer to emergence for a 20-year period starting somewhere between 1910 and 1930. In practice, the first instance of emergence occurred in the period starting in 1912 for TN10p. As we also described in the reply to Referee 1, there might be limits in the statistical power of the KS test for data with finite support and other tests might be better suited, but computationally very expensive to use.

References

Dimitrova, D. S., Kaishev, V. K., and Tan, S. (2020): Computing the Kolmogorov–Smirnov distribution when the underlying CDF is purely discrete, mixed, or continuous, J. Stat. Softw. 95, 1–10, https://doi.org/10.18637/jss.v095.i10, 2020.

Frey, J.: Refined asymptotic Kolmogorov–Smirnov tests for the case of finite support, Commun. Stat.-Theory Methods, 49, 5829–5841, https://doi.org/10.1080/03610926.2019.1622726, 2020.

King, A. D., Donat, M. G., Fischer, E. M., Hawkins, E., Alexander, L. V., Karoly, D. J., Dittus, A. J., Lewis, S. J., and Perkins, S. J.: The timing of anthropogenic emergence in simulated climate extremes, Environ. Res. Lett. 10, 094015, https://doi.org/10.1088/1748-9326/10/9/094015, 2015.