Author's response

1 Referee 1

We thank referee 1 for taking the time to re-assess our manuscript and providing valuable feedback.

The authors have taken on board both mine and the other reviewers comments - and I think the new manuscript is substantially improved. I am glad to see a renaming of the methods which makes it much easier to follow. Further clarifications in the methods section are useful.

I notice a minor typo in the brackets of line 92.

We can't find a typo on line 92 in either of the reviewed version of the manuscript or the pdf showing the tracked changes.

Only other comment is that the abstract contains references - this may not be allowed.

We removed the citation, but kept the footnote to WWA.

2 Referee 2

We thank referee 2 for taking the time to re-assess our manuscript and providing valuable feedback.

The authors have fully accommodated the suggestions made in the first round of reviews, and I think the paper is much stronger as a result. In particular, the comparison between the two methods using the same data is very illuminating, suggesting that the relationship between local and global temperatures may (at least in come cases) be well modelled using only a relatively short time series. There are still a few points that could be clearer, and I've highlighted these below. Generally these are fairly minor queries, but toward the end of the paper I became concerned that I may have misunderstood how the WWA-method is applied (see my comment on Figure A2), in which case more substantial revision of that section may be required to fully clarify what was actually done and avoid any potential for misinterpretation.

2.1 Specific comments

18-19. I think 'IPCC 2021' is redundant here – unless there's another chapter that could be cited specifically?

Fixed

44. Remove added 'had'

Fixed

50-54. I don't think this describes the WWA method very clearly – in particular, it's not clear that this approach is used for trend fitting (I initially took line 51 to mean a comparison of conditions in two distinct periods!). I'd suggest something like **'Instead, it is possible to build a statistical model in which the distribution of the variable(s) describing the event changes with global mean surface temperature (GMST), and to use this to estimate the magnitude of events in the preindustrial climate'**.

Great suggestion, we have incorporated this.

67. WWA studies often have found a fairly linear response, particularly for heat extremes, so 'is unlikely to capture' seems a bit strong – suggest 'may not always adequately capture'

Fixed, thanks!

83, 88. Unnecessary to redefine WWA-method after line 55.

Fixed, thanks!

97. Strange formatting, line starts with a comma

Only seems to be a problem in the diff/tracked changes.

101. It's not clear from the text how the models were evaluated – please briefly describe the four metrics used, and direct the reader to where in the text the results of the model evaluation are reported.

We have added a short list of the metrics and hopefully made it clearer what the reference is.

108-119. A minor comment, but this section would be easier to follow if you defined the 2018 heat event, then the 2021 rainfall, rather than flipping between the two as written.

This is a very good suggestion, we have re-organized this section.

141. It would be useful to define 'exceedance probability', maybe in line 133 – it may not be clear to all readers that you're still talking about p_0 and p_1.

This has been added.

145-159. I don't think this discussion of how to estimate p_0 belongs in the 'WWA attribution' section – maybe it could be moved to the introduction, around line 49.

This section goes into some details which requires knowledge of the method, and we think it would be tricky to include it in the introduction without extending this too much, borrowing what belongs to the method.

191. This description of the WWA method could still be clearer – suggest 'they estimate β, μ and σ, along with any other model parameters, directly from equ. 3-5, and would use the longest available reliable time series rather than only a subset as we use here'.
Fixed.

194-197. This section would be easier to follow if 'For all datasets, we used the regression coefficients... equations 3, 4 and 5' were moved to line 189, directly after 'time series of each index'.
This has been moved accordingly.

211. Change 'a station' to 'any stations'
Fixed.

234-245. This would fit better in the section on event definitions.
This has been moved.

254. The GEV could also have a positive shape parameter – suggest 'the GEV distribution has a finite upper bound when the shape parameter is negative'.
Thanks! We have incorporated this.

258. Please add a line addressing the potential criticism that the Gumbel/GEV is theoretically justified only for block maxima and not for count data.
This has been added.

292. I found this a bit hard to interpret – does this mean that temperatures were overestimated in the earlier part of the record? Please clarify in the text.
We have tried to clarify this.

297. Change to 'FAR similar to, albeit lower than'
Thanks!

309. Maximums → maxima
Thanks!

Fig 5 & 7 caption: change to 'The bar represents the percentiles and median of the...' 'Green bars denote the average for each method (PI and WWA)'.
Thanks!

Figures 5 & 7. Does the WWA average include the CORDEX runs? From the text and plot it's not clear. If not, perhaps move the CORDEX bar below the WWA average, to

make this absolutely clear.

It does. We have added a short passage that hopefully makes this clearer.

331. I think it's still right to say that uncertainties are higher for the stations, rather than for the PI method – suggest changing this back

We changed it so that it mentions both the PI-method and stations.

336. The 30-year period limits how well any time series can represent variability, especially when evaluating an event that may not occur within that 30-year period. I'm also not sure what is meant by 'over constrained distribution' here. I'd say instead that this may lead to unstable estimates of the return period/return level, and hence to unreliable estimates of FAR.

This is a good suggestion, we have changed the formulation of these sentences.

353. Again, it's not totally clear from the text what was actually done here. The PI-method has already been applied to the long-term observations; as I understand it, you now use the WWA- method to estimate the FAR using a trend fitted to only the current (1992-2021) climate, and compare the results. If this is correct, please update these lines accordingly.

We have tried to clarify what was done under 'Comparing the PI and ...'

361-362. While it is true that most of the regression coefficients seem to be close to zero, the northernmost point seems to be close to 10, but is still non-significant. Might this therefore also be because of high variability, as well as a weak mean trend?

366-367. I think you could go further with this conclusion: this result implies that the long-term relationship between GMST and local temperature extremes can be estimated using only a relatively short time series from the recent past, which in turn implies that the relationship between GMST and local temperature extremes – at least, in this area and for these extremes – remains fairly constant.

This is a good suggestion and we have added a sentence on this.

373. I'm still not sure what is meant by 'over constrained' in this context, please explain.

In line with the previous comment on a similar topic, we changed this to instead be framed as a lack of variability.

375. Please add a description of Figure 10 and explain what it is showing. How was significance determined? I'm surprised that a temperature-based index has so few significant points, can you suggest why this might be?

We have added two sentences to better explain figure 10. The regressions were computed using the ordinary least squares from the library statsmodels, here p-values are given by a t-test.

Figure 10. The colour scale here is a bit confusing - please redraw so that increased

rainfall is shown in blue (the BrBG colourmap in Python would be useful to distinguish precipitation from temperature results)

In general, when plotting precipitation, I would agree. However, for these figures, we think that it is the strength of the regression that is of importance. And keeping the colormap the same for the two plots makes this comparison easier.

385. One point that would be worth investigating is that the WWA approach assumes that the scale of the temperature distribution remains constant over time – Figure 3 suggests that this may not be a realistic assumption, because the PI temperature distributions are typically somewhat narrower than the 1991-2021 distributions. This would mean that the WWA-PI distributions are typically wider than their PI-PI counterparts, and that the event is therefore deemed more likely in PI than it really was – which would lead to the lower FAR seen in Figure 8. The effect is less clear in the precipitation series, and the situation is less straightforward because the scale parameter changes with the location: so it's not clear whether the scaled distribution would be likely to systematically misrepresent the PI distribution. However, this could be easily checked by comparing the scaled distributions to Figure 3.

This is an interesting comment, and as we interpret it touches on a somewhat larger topic: should the scale of the temperature distribution be kept constant or also be changed, as is done when scaling the precipitation distributions. As you mention, the current and pre-industrial observations seem to indicate that the scale does not remain. We consider a full investigation of this out of scope for this paper, but will add a few sentences on it in the results and conclusions.

393. I think you mean exceedance probabilities here?

Yes, fixed.

406. Or possibly the fixed scale parameter – which could be allowed to vary

Part of the comment to line 385 above.

406. The differences between the two results are fairly minor, as you highlight around lines 364- 367: this should be restated in the conclusion, along with a summary of my comment above about lines 366-367.

This has been adressed.

408. One potential advantage of the WWA method is that, if homogenised station data aren't available, it could be modified to accommodate a changepoint in the series and so to make use of all of the available data to estimate the effect of GMST.

419. Gridded datasets (worth highlighting: to distinguish from the station dataset you mention in the next line)

Agree, fixed.

423-425. It's not clear where in the main text this is indicated: is this a reference to Figure

10? If so, you should also draw this conclusion where the figure is discussed, perhaps at line 380. However, the next line suggests not. Please clarify.

This has been clarified in the relevant results section and in the conclusions.

432. You should also highlight here that WWA recommend using as much data as possible to estimate the model parameters, almost certainly more than 30 years of data. Although the mean trend in temperatures was fairly well estimated, the uncertainty probably was not: using a longer time series would be expected to give a better estimate of the variance of the distribution, which is critical for correctly estimating return periods and PR/FAR (see eg. Zeder et al. 2023, 'the effect of a short observational record on the statistics of temperature extremes')

Added a short section on amount of data.

Figure A2. At this point, I started to doubt whether I've misunderstood something in the methods. My understanding is that the WWA-method was implemented for station data by regressing these time series against GMST, and using Gumbel distributions corresponding to present-day and pre- industrial GMST to estimate the FAR. The tests reported suggest that there is no evidence of nonstationarity, but the estimated FAR is consistently strongly positive, which implies a nonstationary trend. How do you reconcile these two findings? (If I've misunderstood the methodology then that section needs to be rewritten to clarify what was actually done.)

Your understanding of how we implemented the WWA-method for station data sounds correct. As we interpret it, for the short snapshots, the trend test and regression to GMST are not directly comparable. The year-to-year variations within the trend stationary e.g. txge25 data can still follow the variations in the snapshot of the GMST data, which would give a positive regression. Furthermore, it is only the snapshots that appear trend stationary, if we were to do a similar test of the whole observational record, there likely would be evidence for non-stationarity. It is quite interesting that the regression between subsets of GMST and txge25 and the subsequent shifting according to the change in GMST can capture this.

Figure A2 & A3/A4 & A5 both have the same caption – please update to clarify that one is for the current period, and one historical.

Thanks, fixed.