Review: “A Bayesian statistical method to estimate the climatology of extreme temperature under multiple scenarios: the ANKIALE package” by Robin et al. (GMD manuscript egusphere-2025-1121, revised).

Overall, this manuscript is very much improved. Nevertheless, I do have some further comments on both the substance and the presentation of the paper that I hope the authors and editor will find useful.

General Comments:

1) My primary concern continues to be about influence of the simulated responses to future emissions on inferences about changes in the intensity and expected frequency of historical extreme temperature events when going from counterfactual to factual external forcing conditions. This apparently occurs because of the way in which SSP driven simulations are smoothed, which is performed in this case with B-splines, presumably still with the same, very small number of knots used previously (details are not provided anywhere in the paper). This degree of smoothing understandably affects the shape of the forcing response function during recent several decades of the historical period. The authors make this sensitivity appear to go away by incorporating climate simulations under 4 different forcing scenarios into their prior – but implicitly that means that they think that the resulting estimate of the response to anthropogenic forcing in recent decades is closer to being correct than we would get, for example, from simulations of the historical period that are extended only as far as the present using a single SSP such as SSP2-4.5. I think the authors should give this problem more thought and at minimum, discuss it more thoughtfully in the paper.

2) An additional concern is that I think the authors need to discuss the suitability of the GEV model much more carefully than they do. The paper contains essentially no cautionary words in this regard indicating to users of the ANKIALE package that they need to carefully justify their choice of extreme value distribution. In the paper, inferences are made based on the annual maxim of three-day running averages of daily maximum temperature without any concerns about whether the upper tail of the fitted distribution can adequately represent the intensity and frequency of rare heat events. What evidence is there that this sampling approach (annual maxima of three-day running means) places us deeply enough within the “domain of convergence” of the GEV distribution to trust inferences about events that are outside the support of the data and therefore very dependent on the assumption of

tail stability that is implicit in the approach that has been used? This is not a trivial issue that can simply be brushed aside because (a) daily maximum temperature generally has a unimodal distribution that is not all that far from being Gaussian (suggesting that convergence to the limiting GEV distribution will be slow), (b) smoothing the daily maximum timeseries creates something that is even closer to being Gaussian, and (c) doing so reduces the effective block length relative to that for daily maximum temperature, which itself has an effective block length that is substantially less than a year (clearly, annual temperature maxima do not occur in winter). Note that diagnosing the fit of the model within the support the observed annual maxima, as would be the case when applying the Kolmogorov-Smirnov test, doesn't help very much in providing confidence that the tail above the support of the data is well represented by the fitted distribution.

Detailed comments:

10-13: This sentence needs some clarification because it presently seems to suggest that ERA5 extends to 2100!

19: Not all "attribution studies" – what is being referred to here are *extreme event attribution* studies rather than long-term detection and attribution studies (sometimes called trend attribution studies).

29: Somewhere this paper needs to carefully discuss that assumption (see general comment 2 above).

33: "later" ➔ "latter"

51: "progresses" ➔ "progress"

55: "This code is" ➔ "This code was" (the rest of the sentence is in the past tense, so this should also be in the past tense).

58: These numbers are presumably for a grid of a specific size (~4000 points?). Other applications would have different computational requirements, so that should be mentioned I think.

76: "weather forecasting models" ➔ "a weather forecasting model" (ERA5 doesn't use multiple models).

78: Interpolated how? It would be good to at least give an indication of what is interpolated. I assume that what is meant here is that 2m air temperature is estimated from surface (skin) temperature and lower model level temperature.

79: “spatializing”?? There is no such word in English. I assume that you mean, “by spatially interpolating”.

82: “a global coverage” ➔ “global coverage”

92: “Change is this average temperature” ➔ “Changes in these spatially averaged temperatures”.

Note that there are many more minor editorial issues like this that can easily be corrected through careful proof reading, perhaps by enlisting the help of a colleague who is a native English speaker.

109: It would be helpful if the methods section, or perhaps an appendix, could provide details about the splines that are used and how the spline coefficients are estimated. They play a central role in the construction of the prior distribution, so it seems important to provide that information.

121: While ERA5 is of high quality, I think it is debatable whether it can be considered equivalent to (i.e., exchangeable with) in situ observations.

161: While the notation is much improved, it is still not entirely clear what some symbols are meant to represent. For example, exactly what is X^0 and what does X^N represent when a subscript is not present?

167: In equation (5), what is the time range for t? You refer to SSP’s, which implicitly indicates that t takes values from 2015 onwards, but I don’t think that’s what is intended. Some further adjustment of notation is presumably needed to distinguish between things estimated from historical simulations and their SSP driven extensions in the period beyond 2014.

211-212: See general comment 2 above – I don’t think we can be as confident in this assumption as you indicate here.

253-254 (positive shape parameter): I think it could be argued that this is physically implausible – which begs the question of whether physical understanding should play a role in constraining GEV parameter estimates to remain negative.

255-260: If there is no evidence that σ_1 differs from zero, wouldn’t it be better to simplify the model by assuming that $\sigma_1 = 0$?

284: This appears to be a notation failure (the note given here seems to end in tautology).

290-291: See general comment 1 above – this seems to be an artefact of an implementation choice (i.e., a subjective decision about how to represent the

estimated response to external forcing) rather than something that should be expected “in theory”.

292-299: This seems inadequate as a discussion of Fig. 4 (which consists of 30(!) figure panels). Also, it is not obvious to me what is being shown in these figures. The ordinate is labelled, but not the abscissa, and no distinction is made between the individual forcing results. Each panel seems to have many superposed QQ plots (in both red and blue), with no obvious way to distinguish between the individual QQ plots that are shown within an individual panel. In panel (a1) for example, there appear to be 6 red QQ plots, but there are only 4 different SSP scenarios. As you can see, I’m confused by this figure. Better labelling, a more complete caption, and further synthesis of the figure in the text would all help.

307-309: This is unclear, perhaps because the French verb “résumer” is used assuming that the English verb “resume” has the same meaning. In English, to resume something means to continue an activity (such as talking) after having paused that activity; it does not mean to summarize, as in French. To add confusion, however, there is a noun in English (resume – pronounced “resumé”) that refers to a document like a CV that summarizes a person’s career.

393-395: This seems a limitation given that many extreme events of interest, such as heat events, have large spatial extent.

396-401: I think the language in this paragraph could be tightened up and made more precise. What is being discussed here are estimated changes in the intensity of extreme events. A general reference to “increasing extremes” could mean the intensity of extreme events with a given probability of recurrence, but I think most often it would be interpreted by the public as meaning that “extreme events” (however defined) are becoming more frequent. Of course, the two are linked, but the discussion here is specifically about the parameters that control intensity.

410: “return periods” ➔ “estimated return periods”. See also the general comment 2 above about extrapolation into the tail. Uncertainty in the estimated recurrence frequency events with intensities that lie above the support of the distribution under counterfactual conditions would be affected by large sampling uncertainty and also by large (and unknown) extreme value model uncertainty.

421: Sentence formulation needs work. A weaker estimated intensification is noted in an area where the opposite might be expected (due to Arctic amplification related processes), but the wording seems to suggest that the estimated intensification

might nevertheless be greater than expected (even if the estimates are smaller than elsewhere).

427: Why “strange”?

434: Again, see general comment 2. The KS test might not indicate a problem, but I think visual inspection of the plots for Paris in Fig. S8 would suggest otherwise. It seems evident that the evolution of, say, the intensity of the 2-year event in ERA5 at the Paris location, does not follow the evolution of its intensity in the inferred (posterior) distributions – the intensity in ERA5 seems to increase more rapidly over the period since 1970 than inferred.

469: The discussion is this section is rosy and positive, envisioning further extensions and applications that could be pursued, but there is nothing here in the way of cautionary words, which I think is a shortcoming that should be corrected.