Manuscript EGUsphere-2025-1121 'A Bayesian statistical method to estimate the climatology of extreme temperature under multiple scenarios: the ANKIALE package' by Robin et al. Referee's report

The main purpose of this paper seems to be to improve on a methodology for extreme event attribution that was proposed by two of the authors. There are two main improvements: one is the ability to consider multiple emissions scenarios simultaneously, while the other relates to the adoption of faster algorithms for the analysis. In scientific terms, the first of these is much more interesting. Regarding the algorithms: it's useful to know that the software is faster and I'm aware that the implementation can be a lot of work, but I wouldn't regard it as a headline message for the paper, especially it seems to come just from adopting a more modern (but now fairly standard) MCMC method.

I'm not aware of other work that attempts to incorporate information from multiple emissions scenarios simultaneously within the framework of a single statistical representation, as is done here: most analyses treat individual scenarios separately which, as the authors correctly point out, leads to potential inconsistencies when focusing on quantities that are not scenario-dependent. As a first attempt to highlight this issue therefore — albeit in an application where the role of the scenarios needs to be explained more clearly (see general comments below) — this paper is welcome.

Unfortunately however, I think it needs a very major rewrite before it can be considered publishable. As far as I can tell, the methodology is *probably* consistent with many approaches to attribution — although I think it may be unnecessarily complicated and the approach to fitting the GAMs is possibly not optimal (see below). However, the manuscript is poorly written and, as a result, I find it hard to understand exactly what is being done and why. The scientific content is therefore hard to evaluate.

The main problem is that the paper doesn't contain all of the information needed to make sense of it. There are also problems with the organisation of the material; with terminology that is not properly defined or is used incorrectly; and with mathematical notation that is not always defined or explained. Conversely, the paper makes heavy weather of some things that are actually rather straightforward (for example, the estimation of GAMs with a common component across scenarios — see below).

Comments and questions follow. Some of them are probably due to my lack of understanding, for which I apologise. I have done my best, within the 12 hours or so that I'm willing to spend on a review.

General comments

1. The paper is not self-contained. The authors state explicitly that they will only describe the changes relative to their earlier paper (denoted as RR20); but a reader should not have to look at another paper in order to understand this one. A brief but complete summary of the relevant points from RR20 is needed.

2. There is no clear statement of what the analysis is for. The abstract (line 1) says that it's about estimating the statistics of temperature extremes; but the paper itself is framed around the methodology of attribution studies. Reading through the paper for the first time, I didn't know where it was going or why. To give just one example: at some point the aim seems to be the estimation of a quantity called κ which is described as "the random variable of a multi-model synthesis" (line 128): I still don't know what this represents, and from here on I started to get completely lost. It would

be helpful to include, early on, a clear statement of what you're trying to estimate and to explain, in broad terms, the information requirements (I don't mean the precise datasets used: rather, the need for real-world time series of TX3x and whatever else you need for the real world; and GCM simulations of [etc. etc.]). It might also be helpful to include a schematic diagram — again, early in the paper — showing how all of these quantities fit together in the steps of the analysis.

3. I don't understand the role of the scenarios in the context of an attribution study. Scenarios relate to *future* climate (albeit starting in 2015 for the CMIP6 runs), whereas attribution studies typically work with information on the past and present. It's possible that the scenarios allow more precise estimation of anthropogenic effects using the GCM outputs, but this isn't explained anywhere. It's also possible that I have missed the point — but in that case it hasn't been explained clearly enough.

4. In some way related to the previous two points: throughout, the paper attempts to describe *what* was done, but there needs to be more justification for *why* it's being done.

5. The approach uses the GCM outputs to place a prior distribution on real-world quantities, based on an assumption that the GCMs are centred on reality (lines 184–185). This is known to be unrealistic: climate models are collectively biased relative to the real-world, and the biases are not independent. I am aware that the authors' assumption is often made in the literature. However, the paper should at least provide an honest acknowledgement that it is known to be (very) wrong — this is obvious to anyone who understands CMIP, and Reto Knutti and co-authors have demonstrated it empirically beyond doubt. There is also literature that aims to address the problem by relaxing the assumption e.g. to a notion of co-exchangeability (defined by Rougier et al. 2013, doi 10.1080/01621459.2013.802963).

   I can imagine an argument that the precise assumptions about the GCMs don't matter too much because they are only being used to set a prior. My response to that would be: if the prior doesn't influence the results then you don't need it because the real-world data are already sufficiently informative; but if the prior *does* influence the results then it needs to be defensible. The work of Knutti et al. shows that this prior is not even *close* to being defensible. At the very least, there should be some investigation of sensitivity to plausible alternative prior choices.

   A minor point, related to this isssue, is that the formulation doesn't date back to Ribes et al. (2017) as claimed on line 185 — it goes back much further.

6. The proposed model seems needlessly complicated. I'm not 100% sure of this, because the presentation around equation (1) is unclear (e.g. in lines 106–107 we learn that this equation wasn't actually used). I am also aware that the proposed methodology arises from a line of reasoning that is accepted in the attribution literature. Nonetheless:

   - As presented, the model seems overparameterised. For example, there are constant terms in the model for $X_t^{R,F}$ in equation (1), but also in the models for μ_t and $\log \sigma_t$. These constant terms can presumably be merged: the only reason for distinguishing between them is the workflow which splits the analysis into a first stage analysing the global and regional temperatures, and a second stage in which the results are fed into the GEV estimation. It's not clear to me why the analysis should be split in this way: surely the Bayesian computational machinery allows you to do everything in one step? A clear justification for the two-step approach is needed.
   - More fundamentally perhaps: I am aware of the fashion for regressing on global and regional mean temperatures in the attribution literature. I am also aware that this literature tends

to focus on partitioning variation into natural and anthropogenically-forced components. If I understand correctly however, the partitioning here is derived (lines 140–147) from a regression on either the outputs of an energy balance model (EBM), or the net radiative forcings. It's not clear which option has been used to obtain the presented results (or, if it was an EBM, which one was used). However, the results in (for example) Figures S1 and S2 show a strong resemblance between the "naturally-attributed" component of global mean surface temperature and the natural component of the forcings (not shown in the paper, but I'm familiar with the forcing time series). I guess that the correlation between this component and the input forcings is at least 0.95, and possibly greater than 0.99 — indeed, it will be 1 if the forcings themselves have been used in the estimation procedure. Moreover, the estimated anthropogenic components of the global and regional temperatures will be smooth curves, just like the anthropogenic component of the forcings (or of the EBM outputs). I will therefore bet a moderate sum of money that one can obtain almost identical results to those reported in the paper by ignoring the global and regional mean temperatures altogether, and instead using the natural and anthropogenic components of the net radiative forcing in the models for the GEV parameters. This would simplify the analysis both conceptually and computationally (e.g. it would eliminate the need for GAMs).

The attribution community may consider this suggestion to be very oversimplistic. Nonetheless, I would be interested to see how it performs. With the ANKIALE software as described, this should not be hard to implement: just replace the estimates of $X_t^{*,N}$ and $X_t^{*,A}$ with the corresponding components of the net radiative forcing, and see what happens.

7. The split of figures and tables between the main paper and supplement seems strange. For example, Figures S1–S3 are helpful in understanding the process and the results, but they have been relegated to the supplement. On the other hand, Tables 2 and 3 in the main paper take up a lot of space and will be of interest to relatively few potential readers: these could reasonably be moved to the supplement.

Specific / detailed comments and queries

1. Line 1 "estimating the statistics of temperature extremes": as noted above, it isn't clear from this what you're *actually* doing.

2. Line 29: typo "compbining".

3. Line 30: a Bayesian *framework*?

4. Lines 33–34: I would remove the sentence "However ... inconsistencies". It is redundant given the content of the subsequent paragraph, and only one of the issues is actually an inconsistency.

5. Line 47: the time taken to "process the entire domain" presumably depends on the size of the domain! To put this in context, it would be helpful to indicate the approximate number of grid cells being referred to here.

   More generally: if there are any real applications in which it's necessary to carry out an analysis over a domain of this size, then a time scale of up to a week is perhaps still not fast enough. It is certainly too slow for the kinds of real-time attribution studies carried out by World Weather Attribution as

described in lines 18–20, given that the time available for data analysis in such studies is only a fraction of the total time available. For those kinds of applications though, the focus is typically on specific events in much smaller regions. It's certainly helpful to give an indication of computational cost; but it would be more relevant to indicate the cost for realistic applications of the methodology. Alternatively, give (e.g.) an indicative cost per thousand grid cells, so that readers can figure out for themselves the likely cost for their own applications.

6. Line 51: what's the relevance of the number of countries and their partial inclusion? What does "exact ratio" refer to?

7. Line 64: "refer to" $\Rightarrow$ "referred to".

8. Lines 73–74 "we use observations from a weather station ...: But you just said you were using ERA5. Perhaps the station data are used to illustrate the methodology and then, as a separate study, the ERA5 data are used to examine the entire European area. If this is the case, probably it would be helpful for the reader if the paper is reorganised clearly along these lines.

9. Lines 75–76: how does the use of a longer dataset allow you to "better verify the contribution"? Given more data, *any* reasonable method is going to do better. And, if you need an unusually long dataset to identify the benefits of the new approach, then the implication is presumably that the benefits aren't obvious for datasets of more realistic size! (I'm playing devil's advocate, but you need to think more carefully about what you're trying to say).

10. Lines 77–78: as noted above, I believe it is fairly common to use global or regional mean temperatures in attribution studies. Nonetheless, it would be helpful to clarify this — perhaps as part of an introductory summary of the basic steps / logic of the approach (see general point 2 above).

11. Line 85: do the emissions scenarios really cover the historical period? I think the intended meaning is probably that historical estimates of emissions were used for the period 1850–2014, while scenarios based on alternative SSPs were used for the period 2015–2100.

12. Lines 89–90: is there a reference to support the claim that the current warming trajectory is close to the SSP2-4.5 scenario? Or is IPCC (2023) intended to be this reference? As written, it seems to be a reference just for the 2.8K estimate.

13. Line 92 (also line 137): the GCM simulations of TX3x are broadly comparable with the ERA5 estimates because the latter are gridded. However, they are not comparable with the Paris station-based observations. How is this handled?

    While on the subject of gridded data: each GCM has its own grid, and many of them are different from the grid used in the ERA5 dataset. How was this dealt with?

14. Lines 86–97: see general comment 1 above.

15. Line 104: this is where the paper starts to become hard to follow. There is no statement of what the covariate $X_t^{R,F}$ actually *is*: all we're told is that it's a proxy for something else. And how do the quantities in the final line of equation (1) relate to the regional covariate X_t^R mentioned immediately beforehand? And why are these quantities needed, given that the GEV parameters are modelled using the original quantity X_t^R?

The answers to some of these questions emerge implicitly later in the paper. If I understand correctly, equation (1) and the subsequent paragraph do not correctly describe what was done. My best guess is that μ is represented $\mu_0 + \sum_{j\in\{G,R\}} \left(\mu_{N,j}x^{N,j} + \mu_{A,j}x^{A,j}\right)$ with a corresponding expression for $\log\sigma$, where $x^{N,j}$ and $x^{A,j}$ represent respectively the 'natural' and 'anthropogenic' components of variation in the global or regional temperature series. Setting the $\{\mu_{A,j}\}$ and $\{\sigma_{A,j}\}$ to zero then allows an estimate of the GEV parameters in the absence of anthropogenic influence. If this *is* correct, then the presentation in the manuscript seems very over-complicated.

16. Line 105: typo "anthropogenice".

17. Lines 106–107 "We also add . . .": this confirms that equation (1) does not describe what was done. See above.

18. Lines 109–10: it is not correct that the model can be seen as a linear model. It's not linear, and linear models have additive noise terms. Suggest deleting this sentence.

19. Line 112: "construct" ⇒ "estimate"? Also, the 'factual / counterfactual' vocabulary seems unnecessary and confusing. I know that it comes from the literature on causal inference, but the present paper doesn't really deal with causality. If my summary in point 15 above is correct, then none of it is really needed — and other explanations can be simplified as well.

20. Line 116: it's not clear to me that all elements of $\boldsymbol{\theta}$ should be described as 'parameters'. The Xs would be more precisely described as latent or hidden variables. Arguably this is a minor point, but it does affect the readability of the paper.

21. Line 128: what does κ represent? See general point 2 above.

22. Lines 131–132: this repeats lines 96–97 (but see general comment 1).

23. Line 143: 'Energy Balance Model' appears twice. And which EBM are you using?

24. Line 145: residuals from what? (actually I don't think they *are* residuals).

25. Line 147: how is the variance of ε^* represented? Is it the same for all SSPs?

26. Lines 148–151: This presentation is needlessly complicated. If GAMs are needed at all (see general point 6), the model could be fitted directly using the `gam()` function in the `mgcv` package in R, e.g. using the `by()` argument to fit different smooths for the SSPs while retaining the common natural forcing term. The required setup is actually quite straightforward: it may even be available somewhere in python, although I appreciate that there's nothing in python that gets close to the capabilities of `mgcv`. We probably don't need to know the details of backfitting either: this is a completely standard approach to fitting GAMs.

27. Line 153: if it *is* necessary to retain the details on how the GAMs are fitted, then the expression in line 154 is a sum rather than an average as described: either the expression or the description is wrong. And in line 154, I assume both terms are summed: they should be enclosed in parentheses to make this clear, therefore.

28. Line 156: as noted above, it would be helpful to move Figure S2 into the main paper. Also however, there is no indication as to whether this was obtained using radiative forcings or EBM outputs.

29. Line 157 "We can see that ..."': how can we see that? What are we supposed to be looking at? How many model realisations are there? If more than one, this hasn't been mentioned in the text (or how you have dealt with it) — it is noted in the caption to Table 1, but if affects the interpretation of results then it needs to be discussed in the text as well.

30. Line 158 "peaks are due to the volcanoes"': they are troughs I think, not peaks.

31. Line 159: when referring to the "constant" signal, do you mean constant over time or constant across SSPs? If over time: why is this relevant? If over SSPs: why do the volcanic and solar effects differ between them as suggested at the end of the line?

32. Line 161 "we propose the following method ..."': I have no idea what is being done here, or why.

33. Line 162 "resampled"': what does this mean? Resampled from what?

34. Line 164: what "method described above" is being referred to here?

35. Line 165 "the four smoothed covariates"': what are these? I thought the previous step (described in lines 138–155) was designed to estimate these covariates.

36. Lines 184–188: it is not correct to describe the formulation here in terms of models that are "statistically indistinguishable from reality", which implies that reality and the models are exchangeable. The representation set out here is sometimes called the 'truth plus error' representation, I think.

37. Lines 202–203: I don't know what 'weights' are referred to here, but this sentence suggests that outlying GCMs are excluded from the analysis. No clear justification is given for doing this but, if what's written is correct, it is very worrying because it suggests that the analysis uses only those models that agree with each other. This is very poor scientific practice, and will also lead to underestimation of uncertainty in the final results.

38. Line 207 'Recall that we have ...': when I read the paper, I didn't recognise that this information had been provided previously. All of the relevant information needs to be provided, clearly and unambiguously, at the outset.

39. Line 208 "Our goal is to estimate the distribution ..."': really? I was unaware of this on reading the paper. And there has still been no clear statement of what κ actually is, or why we might want to learn about it.

40. Line 210: the 'double conditioning' notation in this equation does not exist in conventional practice, as far as I'm aware. I think the denominator is incorrect, as well: what's being done here is Bayes' theorem conditional on $X_t^{*,o}$, which should lead to

$$\mathbb{P}\left(\kappa|T_t^o, X_t^{*,o}\right) = \frac{\mathbb{P}\left(T_t^o|\kappa, X_t^{*,o}\right)\mathbb{P}\left(\kappa|X_t^{*,o}\right)}{\mathbb{P}\left(T_t^o|X_t^{*,o}\right)} \ .$$

    I suspect there is also an unstated assumption that $X_t^{*,o}$ is irrelevant for T_t^o once κ is known, so that $\mathbb{P}\left(T_t^o|\kappa, X_t^{*,o}\right) = \mathbb{P}\left(T_t^o|\kappa\right)$. Apart from the denominator, this then gives equation (5) in the paper.

41. Lines 211–212 "In other words ..."': this makes very heavy weather of quite basic material, while at the same time failing to provide enough information as to what's being done and why.

42. Line 212: "are constrained" by what? Why?

43. Line 216: why do you want "the posterior of global and regional temperatures"?

44. Line 218: in this equation, the left-hand side depends on t but the right-hand side doesn't. I have no idea what is intended. What are the dimensions of the subsequent matrices? What are you doing?

45. Line 224: again, what are the dimensions of these quantities?

46. Line 227: N^{SSP} isn't defined.

47. Line 228: what do you mean by 'null values'? If you mean zeroes, say so.

48. Lines 234–235: what do you mean by "scenarios" and "observations" here? Are the "scenario" data from the climate models? If so, what's the basis for this assumption? After all, the available scenarios (assuming you mean the SSPs that are included in the analysis) are arbitrary: the observations can't be the average over *all* collections of scenarios.

    If the scenario data here are not from the climate models, where *are* they from?

49. Line 237: how is it 'more general' to take the average of the scenarios? All you're doing is to making one choice instead of another.

50. Line 240: what does 'this constraint' refer to?

51. Line 241: 'the covariate Europe'?! At this point, I'm starting to wonder whether all of the authors checked the manuscript before it was submitted.

52. Line 263: use of the NUTS is sensible, but there's no need to provide details of, e.g. the generation of proposal distributions or anything else in the paragraph. The explanation provided isn't enough for readers who are not unfamiliar with the algorithm, and they don't need to understand how NUTS works in any case. All a reader needs to know is that this is a a more modern MCMC method that can often overcome the difficulties described in the previous paragraph.

53. Line 280: what are "the laws"?

54. Line 281: again, 'recall' implies that the reader has been told already.

55. Line 284: what's an 'intensity'? Don't you just have a value of T here?

56. Line 291: I don't think 'deduce' is the right word here. Rather, you can construct measures of the change.

57. Line 295: Figure S4 is another that belongs in the main paper.

58. Lines 297 'we calculate ...': why? **NB** the answer to this may be obvious to someone who understands what the authors are trying to do, but I lost track of that long ago.

59. Lines 301–303: why are you calculating the Wasserstein distance? Few readers will have a good intuition for what a given Wasserstein distance looks like, in terms of the distributions being compared. And most will be interested primarily in whether the distributions differ to an extent that would change any substantive conclusions of interest.

60. Line 312: 'slightly ahead' in what sense?

61. Line 320: is the 'simultaneous analysis of several thousand grid points' often required? Given that you're just doing one grid point at a time (I think), attribution studies would usually focus on specific events that affect a subset of the grid points. More importantly though: in these kinds of applications, the focus is usually on weather that is simultaneously extreme over a large area. As far as I can see, the present paper doesn't address that problem: some discussion of this is needed, as part of the context for the work.

62. Line 326: I wonder whether this kind of "package manual" material is appropriate for inclusion in a journal article? This query is perhaps for the journal editor.

63. Line 377: what does 'since 1940' mean in 'the maximum observed in 2024 since 1940'?

64. Line 389: typo 'conter-factual'.

65. Line 451: what would be the challenges in extending the methodology to other variables such as wind and precipitation? It seems to me that it would be basically the same: it's all just based on the GEV distribution for annual maxima.

66. Caption to Table 2: what are these values averaged over? Grid cells, I assume — it would be helpful to clarify this. See also the earlier suggestion for moving this table (and Table 3) to the supplement.

67. Algorithm 1: this is a completely standard algorithm, and there is no need to include it (but if you do include it, the cross-references at the end are broken — similarly for Algorithm 2 on the next page).

68. I don't find Appendix A very helpful. The material on GAMs in Appendix A1 is mostly fairly standard, but seems more complicated than necessary; while I don't understand Appendix A2 because I still don't know what the purpose is.

Finally: out of interest, before submitting my report I looked at the comments from the other reviewers. There is quite a lot of overlap with my comments above, particularly from Anonymous Reviewer 1 who also found the paper hard to follow. To be completely clear therefore: in this report, everything except this paragraph was written independently and without looking at the other review reports.

Richard Chandler
University College London