Response to Reviewers comments about the article "*A Bayesian Statistical Method to Estimate the Climatology of Extreme Temperature under Multiple Scenarios: the ANKIALE Package*"

Robin, Y., Vrac, M., Ribes, A., Barbaux, O. and Naveau, P.

February 16, 2026

Note In this document, the text in regular format corresponds to the reviewers questions. The answers from authors are given in the grey blocks.

Contents

1 Reviewer 1

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.

> We thank you for taking the time to review this new version and for recognizing the important work that has been done. We are pleased that this new version is much clearer, and you will find our answers to your questions below.

1.1 General comments

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).

> You are absolutely right to point out that the analyses performed are highly dependent on how the counter-factual world is constructed, this partly explains the variability in results that can occur between different studies. The World Weather Attribution (WWA), for example, typically taking a 15-year moving average to smooth out global temperature and and they define their counterfactual with a constant covariate, which takes as its value the current GMST minus the observed change.
>
> We also note that, after describing in detail in the supplementary material how the smoothing is constructed, we completely forgot to give the exact parameters used in our study. Historically, in the original paper by Ribes et al. (2020), the authors used a natural spline basis (in other words, the knots are the 251 annual time steps), smoothed so as to have only 6 degrees of freedom. This value was obtained by cross-validation on CMIP5 simulations (historical and RCP8.5, van Vuuren et al., 2011; Taylor et al., 2012). This value was then retained in subsequent papers (see, e.g., Robin and Ribes, 2020; Ribes et al., 2021). In this new work, we favored *B*-splines, with a number of knots more in line with the number of degrees of freedom. We settled on 10 uniformly distributed knots, this time with 8 degrees of freedom regardless of the scenario in order to better represent the bell curve of SSP1-2.6. These parameters are easily modifiable in ANKIALE.
>
> We have amended the supplementary material to incorporate the above elements.

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.

We do not eliminate the sensitivity of the response to forcings; we remove what we consider to be natural variability (the smoothing can be configured), and the use of several scenarios only serves to ensure that the historical response (partly pre-2014) is the same regardless of the scenario. Indeed, the period between 2014 and today no longer corresponds to any of the scenarios used, but amounts to using the average of the four SSPs over this time interval. We believe that this is not a problem, as the four scenarios are extremely close over this period (especially when their uncertainty is taken into account, which we do).

We have added a discussion clarifying these points in the supplementary material in the Sec. S.1.1.4 and S.1.3.1.

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.

The question of whether a variable actually follows a given distribution is a classic problem in statistics, made even more difficult in our context because the statistical model must correspond to each of the climate models, the prior, and also the posterior (and therefore to the observations). This question, although only briefly touched upon in the article (we simply state that we assume that the variable TX3x follows a GEV distribution), has been addressed in many previous studies. For example:

- In the work of Robin and Ribes (2020), the choice of statistical model is studied to best fit climate models in the case of a GEV distribution. The results are very mixed, showing that the chosen model appears to be a compromise.
- In the work of Ribes et al. (2020), the normal distribution is used because a monthly average is better represented by a Gaussian distribution.
- The WWA studies, based on the protocol in Philip et al. (2020), include the search for the most

appropriate distribution. Annual maximum variables over n days (with n small) are generally better represented by GEVs than normal distributions (the tail is too strong to be consistent with the normal distribution, which drops to 0 much too quickly), see for example Pinto et al. (2024), Zachariah et al. (2023), Barnes et al. (2023), and van Oldenborgh et al. (2019). Variables averaged over longer periods are more Gaussian (Ciavarella et al., 2021).

With this in mind, several statistical models are offered in ANKIALE (currently, GEV, Gaussian, and GEV adapted to minimums), and since the list cannot be exhaustive, adding a statistical model is reasonably simple.

It should also be noted that the objective of our work here is not to choose a statistical model (even though we give GEV as an example), but rather to focus on the inference method, which can be adapted to different statistical models. The choice of GEV for TX3x is sufficiently validated in the literature that we do not consider it necessary to include a warning with each estimate we provide, particularly in Sec. 5.

A discussion along these lines has been added in the introduction of the Sec. 3.1 (*Definition of the statistical model*), highlighting the importance of choosing the right statistical model. Several new examples have been added to ANKIALE dealing with the normal distribution and the GEV for minimums.

1.2 Detailed comments

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

We have rephrased it.

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).

We have replaced "attribution studies" by "extreme event attribution studies".

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

A discussion has been added, see the response to the general comment 2.

33 "later" & "latter"

Thanks, corrected.

51 "progresses" & "progress"

Thanks, corrected.

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).

> Thanks, corrected.

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

> We also give the calculation time for a grid point, which in theory does not depend on the size of the grid. In practice, a larger data set will take longer to load (or may not fit in memory and may need to be split up, which ANKIALE allows), which may affect calculation times. We have modified the text of the introduction to reflect this.

76 “weather forecasting models” & “a weather forecasting model” (ERA5 doesn’t use multiple models).

> Thanks, corrected.

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.

> The source of our sentence is the ERA5 documentation, which says that *2m temperature is calculated by interpolating between the lowest model level and the Earth’s surface, taking account of the atmospheric conditions* without further details. The key point for us here is that we wanted to draw attention to why ERA5, particularly in terms of extremes, can differ significantly from observations (and especially EOBS).
>
> We added the additional elements to this sentence and sourced it directly.

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

> Thanks, corrected.

82 “a global coverage” & “global coverage”

> Thanks, corrected.

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.

> Thanks, corrected.

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.

> The supplementary material contains all the equations and methods used to find the spline coefficients. In particular, we show how this problem can be rewriten as a least squares problem.

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.

> It is true, so we have replaced by the word "similar". Note that Sec. 2.1 already contains a discussion urging caution against confusing observations with reanalyses.

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?

> The t was missing for X_t^N, this has been corrected. The term X^0 is a constant, while X_t^N is the response to natural forcings. This allows us to write the response to forcings as the sum of the natural and anthropogenic response X_t^A, modulo a constant X^0.

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.

> Indeed, notations with SSP give the impression that we are only using the future part, whereas we are using both the historical data and the SSP scenario simultaneously (meaning that the time axis spans from 1850 to 2100). We have clarified the text by explicitly stating that the historical scenario is being used at this stage. Ideally, we would have a HIST+SSP notation, but we believe that this would significantly reduce readability.

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

> A discussion has been added in the supplementary material, and we clarified the text in the Sec. 3.1, see our response to the general comment 2.

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.

> Even if we agree on the principle, the assumption of physical implausibility is not necessarily sufficient to force a negative shape. For example, precipitation is physically limited (absurdly, it cannot rain more than all the water available in the atmosphere), but the value is so large that the shape becomes positive in this case. An approach such as that of Noyelle et al. (2026),

which imposes a bound calculated on physical variables, allows the shape to be forced to a negative value. A perspective is to integrate this type of approach, which has yet to be tested.

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 $\sigma_1 = 0$?

Indeed, the posterior shows that σ_1 is very close to 0. Unfortunately, this is not the case throughout the procedure, where some climate models show a σ_1 significantly different from 0 (Robin and Ribes, 2020). We are therefore forced to retain the hypothesis of $\sigma_1 \neq 0$.

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

In the formula $I^{\mathrm{F}}_{t=2019} = I^{\mathrm{F}}_{2019}$, the term on the left is the value of the function I^{F}_t, while the term on the right is the value of the intensity of the event, defined from the data. It is obviously intentional that I^{F}_t is equal to I^{F}_{2019} in 2019, but technically these two objects are constructed differently.

The explanation has been added to the text.

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".

If by "artefact of implementation" you mean "choice of statistical model" (spline, separate inference of scenarios rather than simultaneous), then we are actually saying the same thing. It is indeed theoretical statistical choices that lead to different results, where physical theory would suggest that they should be the same (the counterfactual and historical parts of the simulations).

We have reformulated.

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.

There are indeed 6 QQ plots per colour, as these are QQ plots *between* scenarios. As there are 4 scenarios, this gives $4 \times (4-1)/2 = 6$ QQ plots.

We have modified the figure caption to better describe the axes (as this is a quantile-quantile plot, the x and y axes are in fact the same, and given in the first column). We have also made it clearer that the four scenarios are shown in the same colour, with blue and red distinguishing between an approach where the scenarios are analysed simultaneously or independently. The text has also been modified accordingly.

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.

> There may be a problem with the English, but that's not the issue here. Here, "to resume" means that we are going to redo two attributions for two specific events (the median and the 99.9% quantile). We have reworded it.

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

> Events that have a spatial structure are generally analyzed by taking the average or maximum value for the area in question, which produces one (or more) time series that can also be analyzed by ANKIALE. We therefore do not really see this as a limitation.

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.

> That's right, thank you for pointing that out. We have reworded it.

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.

> Although we understand the limitations inherent in choosing a statistical model (see our response to your general comment), we would like to point out that we attach crucial importance to calculating uncertainty (and therefore confidence intervals), particularly the uncertainty of global, regional, and local change. Perspective elements (such as estimating the upper limit of temperatures, including counterfactuals) would undoubtedly improve these estimates, but this is beyond the scope of our article.
>
> We have replaced "return times" by "estimated return times".

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).

It would appear that the Arctic amplification phenomenon affects higher latitudes, and rather the average. We can see, for example, in Fig. 4 of Zhang et al. (2024) the change in intensity in TX1x, and the changes are indeed very strong over Greenland, but consistent with our results for Europe.

We have also reproduced Fig. S8 for three random grid points in Northern Europe in Fig. RR.1, and we find no errors in the inference, even though in the case of Iceland there is clearly a homogeneity issue in ERA5.

427 Why "strange"?

We changed "which may seem strange" to "which appears to show breaks."

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.

We regret to disagree with your interpretation of Fig. S8. For a draw of (μ_t, σ_t, ξ_0), a KS test is performed between the residuals.

$$R_t := \frac{T_t^o - \mu_t}{\sigma_t}$$

and the law GEV$(0.1, \xi_0)$. This process is repeated 1000 times, giving 1000 p-values. If there were a significant disagreement in the median between T_t^o and the GEV(μ_t, σ_t, ξ_0) distribution, this would result in a significant difference between the empirical median of R_t and the theoretical value from the GEV$(0, 1, \xi_0)$ distribution. Ultimately, the difference between the empirical and theoretical CDFs would be significant, and therefore so would the KS test (since it is the maximum of the absolute values of the differences). Even though it is true that the KS test is rather conservative and tends (especially for small samples) not to reject the hypothesis of equality, we believe that the 1000 draws reinforce the credibility of the test (99% non-rejection). We also checked using a Cramer-Von Mises test, and obtained similar results.

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.

A new paragraph has been added to the conclusions to put our results and choices into perspective in relation to other studies.

2 Reviewer 2

Firstly may I commend and thank the authors for efforts they have made to make the paper more readable and accessible to a wider audience and for the much improved notation too. Reading the new draft was so much easier and enjoyable even.

> We thank you for taking the time to review this new version and for recognizing the important work that has been done. We are pleased that this new version is much clearer, and you will find our answers to your questions below.

2.1 General comments

The only major pushback I would have is that while the authors have made considerable efforts to put their work into context in the introduction they still fail to do this with their results. This is most notable with other attributions studies such as Vautard et al. (2020) which seems a bit odd given the overlap in authors. For example their estimated return periods seem to be way smaller (more frequent) than those presented here for the same locations (fig 6b) even when considering your lower bounds (S6b). But also a other studies are returning numbers much more frequent for the recent hot events, particularly Northern Europe and the UK. Similarly there are other studies that calculate climate change terms for EV parameters (your μ_1 & σ_1 and Fig 5b & 5d) e.g Brown et al. (2014) - yes they look at TX1x rather than TX3x and use regional models but one would expect the impact on EV parameters not to be that different. And you also mention the use of regional models in your Perspectives (section 6) as something that would be useful to do - it would be good to acknowledge that this has already been done at least to some degree.

> It is particularly difficult to compare parameter values directly. In the study you mention, the statistical model is certainly a GEV, but
>
> - The external forcing used is not the same: in the studies you mention, it is global temperature, whereas we use regional temperature, which impacts μ_1 and σ_1.
> - The inference approach can also be very different (observations and models separately, or mixed using bias correction).
> - The variable itself is not the same (TX1x vs. TX3x).
> - We use 10 more years of observed data than a study such as Brown et al. (2014).
> - We use simulations from CMIP6 and no CMIP5 as in Brown et al. (2014).
> - Different constructions of the counterfactual for attribution studies.
>
> All these differences lead to estimated parameters that may be different, and different values for return periods.
>
> We have added a paragraph describing these elements in the conclusion.

I think it would be much better for the Paris results of S8 column 1 to be placed in the main text, perhaps as part of Figure 3.

We wondered whether this figure would be better placed in the main text, but we decided that seven figures were already enough. Hence, we preferred to leave it in the appendix because it illustrates a very specific phenomenon of spatial discontinuity.

Line 419: "which are supposed to be records (and therefore rare), are becoming the norm". Surely we as a community need to be moving away from such terminology. The whole point of your work is to show that the climate is warming and thus we should expect records to be broken - it is now normal and expected for records to be broken, it is not exceptional. Perhaps there was a time when this was a useful communication tool but surely that has now passed. Now it looks more like headline grabbing and scare mongering. It is my view that record breaking is no longer a useful metric or tool and we need to move to a risk based perspective, that is quantification of risk, which this paper does admirably and which the "record braking" trope does a disservice to.

We agree, we deleted the sentence.

2.2 Minor

399 "generally constant across the map" - well this is a matter of perspective and more a feature of the colour scale. The UK being 1 and Spain 2 seems rather big to me in terms of actual changes in temperature being double in Spain what it is un the UK and other more northerly regions. Also with Fig 5b it looks like all numbers are positive so it is a bit of a waste having a diverging colour scale. A colour scale more like 5a might be more appropriate.

It is true that constant is somewhat exaggerated; we based this statement on an average value of $\mu_1 \sim 1.14 \pm 0.3$, but in reality 95% of the values are in the range $0.6 - 1.7$, with min/max values of $0.2 - 2.7$.

We narrowed the color bar to the interval $[-2, 2]$, while keeping the colormap (which is designed to also highlight the sign, which is always positive here). We also adapted the text.

Fig 5d could the scale be narrower please as it does not really show anything at the moment.

The scale is chosen here to show the absence of signal on σ_t after observational constraint. We could use a smaller scale, but we would only see noise with values $\sigma_1 \sim 0.008 \pm 0.03$.

We have added the value in the text.

Fig 6c There seems to be a slight "cubism" in the shades of red. Is this an artifact of the printing or does this mean something?

The slight cubism that we can see at the end is the remainder of the signal from the GCMs cells that may remain. It may be stronger or weaker depending on spatial variability and quality of the observations.

494 - "their uncertainties are lower" - I think this misses the point. The uncertainty in future climate (given a forcing scenario) comes from a lack of understanding in the physical processes that produce the various feedbacks. Using well defined forcings does not mitigate these uncertainties, it just ignores them and will thus be misleading.

We agree, we deleted the sentence.

Bibliography

Barnes, C., Y. Boulanger, T. Keeping, P. Gachon, N. Gillett, J. Boucher, F. Roberge, S. Kew, O. Haas, D. Heinrich, M. Vahlberg, R. Singh, M. Elbe, S. Sivanu, J. Arrighi, M. Van Aalst, F. Otto, M. Zachariah, F. Krikken, X. Wang, S. Erni, E. Pietropalo, A. Avis, A. Bisaillon, and J. Kimutai (Aug. 2023). *Climate Change More than Doubled the Likelihood of Extreme Fire Weather Conditions in Eastern Canada.* Report. World Weather Attribution. DOI: 10.25561/105981.

Brown, S. J., J. M. Murphy, D. M. H. Sexton, and G. R. Harris (Nov. 2014). "Climate Projections of Future Extreme Events Accounting for Modelling Uncertainties and Historical Simulation Biases". In: *Clim Dyn* 43.9, pp. 2681–2705. ISSN: 1432-0894. DOI: 10.1007/s00382-014-2080-1.

Ciavarella, A., D. Cotterill, P. Stott, S. Kew, S. Philip, G. J. van Oldenborgh, A. Skålevåg, P. Lorenz, Y. Robin, F. Otto, M. Hauser, S. I. Seneviratne, F. Lehner, and O. Zolina (2021). "Prolonged Siberian Heat of 2020 Almost Impossible without Human Influence". In: *Clim. Change* 166.1, p. 9. ISSN: 1573-1480. DOI: 10.1007/s10584-021-03052-w.

Noyelle, R., Y. Robin, P. Naveau, P. Yiou, and D. Faranda (Jan. 2026). "Integration of Physical Bound Constraints to Alleviate Shortcomings of Statistical Models for Extreme Temperatures". In: *Journal of Climate* -1.aop. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-25-0112.1.

Philip, S., S. Kew, G. J. van Oldenborgh, F. Otto, R. Vautard, K. van der Wiel, A. King, F. Lott, J. Arrighi, R. Singh, and M. van Aalst (Nov. 2020). "A Protocol for Probabilistic Extreme Event Attribution Analyses". In: *Adv. Stat. Clim. Meteorol. Oceanogr.* 6.2, pp. 177–203. ISSN: 2364-3579. DOI: 10.5194/ascmo-6-177-2020.

Pinto, I., R. C. Odoulami, K. A. Lawal, E. Olaniyan, W. A. Ibrahim, K. Guigma, M. Vahlberg, D. Heinrich, C. P. Marghidan, M. Vogel, J. Arrighi, C. Barnes, F. Otto, S. Philip, M. Mistry, S. Sengupta, S. Kew, and J. Kimutai (2024). *Dangerous Humid Heat in Southern West Africa about 4°C Hotter Due to Climate Change.* Report. World Weather Attribution. DOI: 10.25561/110082.

Ribes, A., S. Qasmi, and N. P. Gillett (Jan. 2021). "Making Climate Projections Conditional on Historical Observations". In: *Sci. Adv.* 7.4, eabc0671. DOI: 10.1126/sciadv.abc0671.

Ribes, A., S. Thao, and J. Cattiaux (2020). "Describing the Relationship between a Weather Event and Climate Change: A New Statistical Approach". In: *J. Clim.* 33.15, pp. 6297–6314. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-19-0217.1.

Robin, Y. and A. Ribes (2020). "Nonstationary Extreme Value Analysis for Event Attribution Combining Climate Models and Observations". In: *Adv. Stat. Clim. Meteorol. Oceanogr.* 6.2, pp. 205–221. ISSN: 2364-3579. DOI: 10.5194/ascmo-6-205-2020.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (Apr. 2012). "An Overview of CMIP5 and the Experiment Design". In: *Bull Am Meteorol Soc* 93.4, pp. 485–498. DOI: 10.1175/BAMS-D-11-00094.1.

van Oldenborgh, G. J., S. P., K. S., R. Vautard, O. Boucher, F. Otto, K. Haustein, J.-M. Soubeyroux, A. Ribes, Y. Robin, S. I. Seneviratne, M. M. Vogel, P. Stott, and M. van Aalst (2019). *Human Contribution to the Record-Breaking June 2019 Heat Wave in France.* Report. World Weather Attribution (not peer reviewed).

van Vuuren, D. P., J. Edmonds, M. Kainuma, K. Riahi, A. Thomson, K. Hibbard, G. C. Hurtt, T. Kram, V. Krey, J.-F. Lamarque, T. Masui, M. Meinshausen, N. Nakicenovic, S. J. Smith, and S. K. Rose (Aug. 2011). "The Representative Concentration Pathways: An Overview". In: *Clim. Change* 109.1, p. 5. ISSN: 1573-1480. DOI: 10.1007/s10584-011-0148-z.

Vautard, R., M. van Aalst, O. Boucher, A. Drouin, K. Haustein, F. Kreienkamp, G. J. van Oldenborgh, F. E. L. Otto, A. Ribes, Y. Robin, M. Schneider, J.-M. Soubeyroux, P. Stott, S. I. Seneviratne, M. M. Vogel, and M. Wehner (2020). "Human Contribution to the Record-Breaking June and July 2019 Heatwaves in Western Europe". In:

Environ. Res. Lett. 15.9, p. 094077. ISSN: 1748-9326. DOI: 10.1088/1748-9326/aba3d4.

Zachariah, M., R. Vautard, R. Chandrasekaran, S. T. Chaithra, J. Kimutai, T. Arulalan, K. AchutaRao, C. Barnes, R. Singh, M. Vahlberg, J. Arrgihi, E. Raju, U. Sharma, A. Ogra, C. Vaddhanaphuti, C. S. Bahinipati, P. Tschakert, C. Pereira Marghidan, A. Mondal, C. Schwingshackl, S. Philip, and F. Otto (May 2023). *Extreme Humid Heat in South Asia in April 2023, Largely Driven by Climate Change, Detrimental to Vulnerable and Disadvantaged Communities.* Report. World Weather Attribution. DOI: 10.25561/104092.

Zhang, X., T. Huang, W. Wang, and P. Shen (Nov. 2024). “Change of Global Land Extreme Temperature in the Future”. In: *Glob. Planet. Change* 242, p. 104583. ISSN: 0921-8181. DOI: 10.1016/j.gloplacha.2024.104583.

Figures

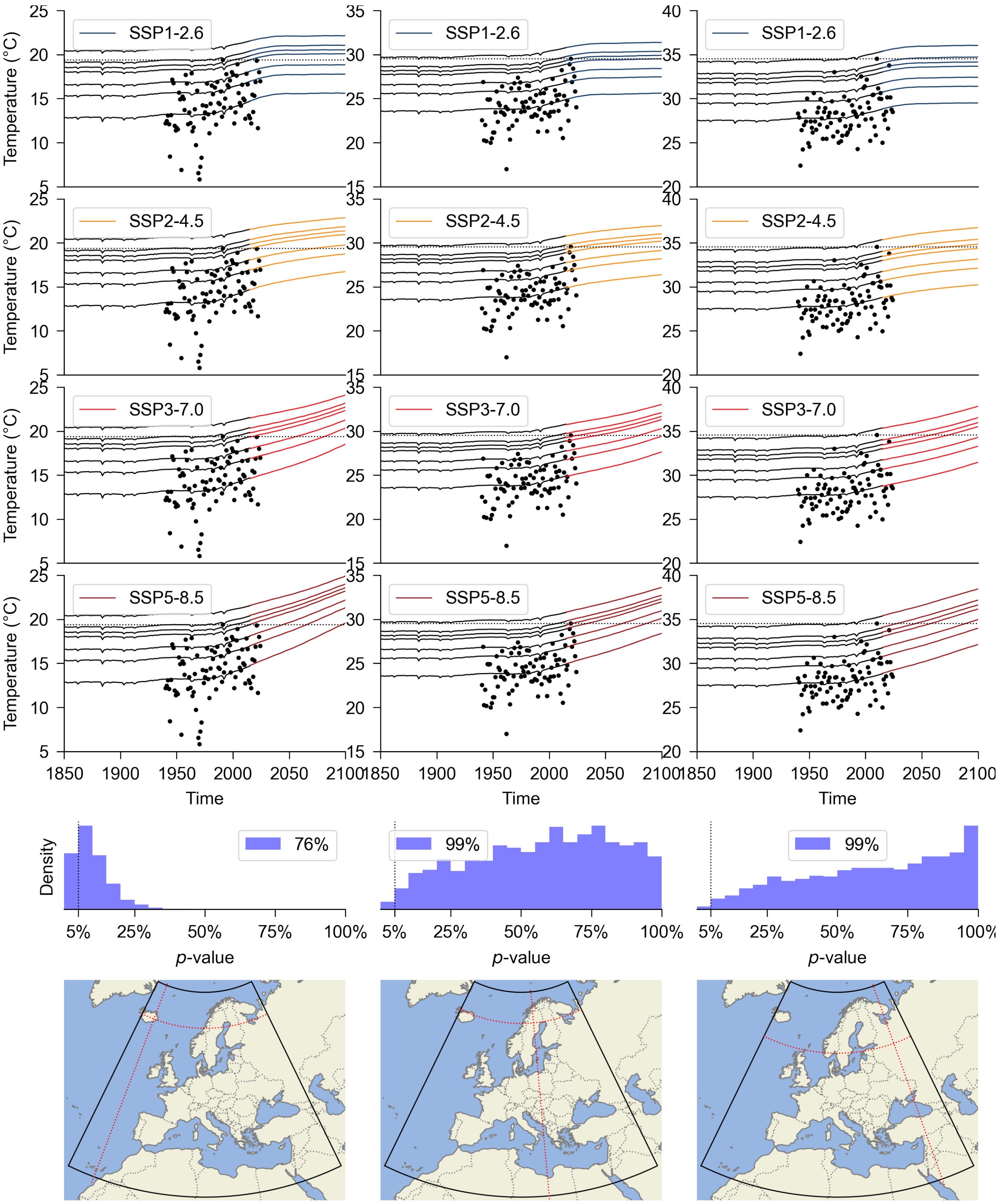

Figure R.1: Comparison between observations and the inferred GEV distribution for three grid points (one per column). The first 4 lines (representing, in order, the 4 scenarios SSP1-2.6 to SSP5-8.5) show ERA5 (black dots), the maximum value of ERA5 (black dotted line), as well as the following return levels: 2, 5, 10, 30, 50, 100, and 1000 years. The fifth line shows the histogram of the p-values of the KS-test of 1000 samples compared to ERA5. The probability indicates the number of tests where the p-value is greater than 5% (threshold where we do not reject that the observations follow the inferred GEV law).