Review, Robin et al., GMD

I think this is an important piece of work – it extends the use of observational constraints to the estimation of the characteristics of temperature extremes for unobserved past and future periods under both “factual” and “counterfactual” conditions, using a rigorous Bayesian framework in which extreme temperature distribution is described with a suitable extreme value distribution. The ANKIALE package that implements the method should help to make this sophisticated methodology relatively accessible to a broad range of users.

Unfortunately, however, the paper would be VERY challenging for the target audience to understand. For this work to be impactful, I think it will be necessary for the authors to think much more carefully about presentation issues, providing more complete and more accessible explanations for the choices that they have made in implementing the package. I think they also need to provide substantially more insight into choices users will have to make when applying the package, with an emphasis on physical considerations as well as statistical and pragmatic considerations. Also, to make the paper accessible to users will require a careful redesign of the notation that is used in the paper, which is hopelessly complex.

Some specific comments:

37: It is unclear why using different scenarios would result in different estimates of the counter-factual world.

It would help if a bit more were said here. Reading ahead, it turns out that this is discussed more beginning at line 113 and I can see from that discussion how this could arise. It is not made clear, however, whether the selection of a particular scenario would strongly affect inferences about observed events based on the posterior that results from using that scenario versus inferences that would be made with another scenario. Sensitivity to scenario choice, particularly if large (and especially under counterfactual conditions) would be of concern, but making that go away artificially by using information from all available scenarios doesn’t really solve the problem in a satisfying way. It would remain a concern that information from models about the future can somehow affect our understanding of the past unless there is a convincing physical argument about why that makes sense. On the other hand, if the sensitivity is small then there wouldn’t be a very compelling reason to bother with the added complexity of the prior and its dependence on the particular experimental design that was adopted in CMIP6. In summary, I think this is crucial point that needs clarification (and in each application, physical justification).

70: What is the point of the right-hand panel in Figure 1? It adds a bit of confusion by hinting that you will make inferences about the record event during 1940 to 2024 at every land point in the domain, irrespective of when that event happened or the spatial extent of the event that produced the record.

75-75: Comparison with a single long record that is almost surely inhomogeneous (e.g., due to instrument changes, observing procedure changes, development of the urban environment around the station, etc) is not going to do a great deal to increase confidence.

77: Usually, "external" forcing would mean external to the climate system (solar, volcanic, ghg's, aerosols …) rather than external for France ….

77-79: If read literally, the sentence could be interpreted as saying that you extract European mean temperatures from HadCRUT5 and global mean temperatures (more correctly, temperature anomalies) from GISTEMP…

Why do you use these two datasets rather than just using one?

In addition, Fig. S1 notes the use of the BEST dataset (Berkeley Earth) – why use yet another global dataset when consistent use of one, well regarded dataset, would probably suffice?

85-86: My understanding is that the historical forcing prescription used in CMIP6 is NOT part of the SSPs, which only cover the period from 2015 onwards.

90-91: I'm not aware that the IPCC assessed, in its synthesis report, that the current emissions trajectory is leading us towards SSP2-4.5. This is discussed by others, however, so I think you should provide a more suitable reference or delete this statement. Indeed, if you have high confidence in this statement, then it would seem that there would be no need to use the other scenarios.

101: The certainty expressed here that the variable of interest will be GEV distributed seems a bit of an overstatement. The GEV distribution is a limiting distribution for block maxima that is (sometimes) achieved as the block length grows without bound. Convergence to the limiting distribution (if it happens at all) can only be demonstrated theoretically under very idealized conditions. Nature, and climate models, do not comply with those conditions (we have awkward things like an annual cycle and the presence of multiple extremes processes that complicate life considerably, with the result that the upper tail does not always behave like that expected under idealized mathematical conditions. While we can't really look into the deep upper tail with observations, and can do so, albeit with some difficulty,

with climate models (e.g., see Ben Alaya et al, 2020, DOI: 10.1175/JCLI-D-19-0011.1). Experience shows that the GEV is nevertheless often useful for approximating the distribution of block maxima for blocks of even modest size (e.g., a year, which effectively only samples part of the year due to the annual cycle). The authors know all of this, and it would be good if some of this could be reflected in the paper, particularly as it is intended to introduce the methods and the ANKIALE package to a wide audience who are not as knowledgeable about the application of the GEV distribution and its limitations.

103: What is the time range considered? Also, I find the notation here somewhat confusing. Readers in a hurry will confound the index F (for factual) with "future", and might confound the index "0" (zero) with "O" (for "observed"). A further question is whether readers should think of the three components of X as being random or fixed.

109: The use of the * to indicate the reader should make a substitution for R or G is awkward and mostly just makes comprehension a bit more difficult for the reader.

114-115: Replace "supposed" with "assumed". Also, this assumption merits some discussion.

119-120: See the comment concerning line 37 above. This needs discussion – particularly why including different futures would affect our understanding of the past.

126-127: It might just be a French/English problem, but what this first step in the procedure entails could be better explained. This would include saying what the assumptions are that are implicit in calculating the uncertainty covariance matrix. It is not clear from the notation if there is one such covariance matrix that describes the spread amongst the different θ_m's (which I am guessing is the case) or whether each θ_m has its own uncertainty matrix.

131-132: It would be much better if this paper could be self-contained rather than sending readers off to another reference for the parts of the methodology that have not changed.

137, 139: While the paper is generally readable, there are many minor grammatical errors. Two examples are mentioned here. This is less excusable these days given the wide availability of tools for polishing text (assuming that GMD authors are permitted to use them).

At line 137, replace "3-days moving average" with "3-day moving averages".

At line 139, where the sentence seems unclear. In that sentence, rather than “are”, do you mean “are estimated with”?

147: The white noise assumption needs some justification. This might be roughly suitable for European regional mean annual surface air temperature anomalies, but the while noise assumption seems a bit less obvious for annual global mean surface temperature anomalies.

148-154: I find myself struggling to understand what is really done here, both because of the notation, which is increasingly complex, and because it is not obvious what model output is being used. If a model has 50 ensemble members, do you use all 50? And if so, do you treat that model differently from a model with only 1 ensemble member? What period is considered, how was the choice to use a smoothing spline with only 6 degrees of freedom made, how are the knots placed, do you worry about the fact that the knot placement is arbitrary and that this imposes wave-like fluctuations that are probably not part of the forcing response?

156: Figure S2 is referenced well before Fig S1 …

177: This statement is made with a lot of certainty and conviction, but whether an event would be judged to be impossible, even under anthropogenic forcing, is highly uncertain. It seems clear from Fig. S3 that the value of the shape parameter is driven by the extreme temperature that is farthest from the location parameter and hence must be very uncertain. This relation between the shape parameter and the most extreme temperature presumably occurs because the parameter estimation process enforces the feasibility of the fitted GEV distribution to a variable, temperature, that tends to have light-tailed extreme value distributions.

184-185: Why is this assumption needed to construct the prior? It’s a prior distribution (i.e., a proposal) that will be updated using the observations when the posterior is derived. It seems to me that this assumption is not needed to construct the prior. Given the way the prior is constructed, it would certainly be helpful if we can regard the models as being indistinguishable from each other (i.e., something hopefully like a simple random sample from model space), but even without that, couldn’t we construct a prior from the models, understanding that it may not do a good job of representing model uncertainty? The updating does require us to have a joint distribution for model simulated and observed quantities, and developing that joint model may require some additional assumptions – perhaps that’s where the “indistinguishable from the truth” assumption comes into play?

189-190: I think this needs discussion – in particular, why internal variability plays a role at all (haven't you filtered it out with the splines?) and what is being partitioned into two components. What is being referred to when you talk about the "common part" of internal variability and each model's additional internal variability??

200: I have no idea what is being referred to here (95% of the covariance matrix)…

202: Which model is excluded? Note that the UK models maybe be problematic due to a known problem in the coupling between the land surface and atmosphere that leads to extreme high localized daily maximum temperatures. The problem is documented at https://errata.ipsl.fr/static/view.html?uid=76b3f818-d65f-c76b-bfd8-cae5bc27825c. Note that the Australian ACCESS models, which also use a version of the UK MetOffice atmospheric model, are not affected but the Korean KACE model, which uses both the atmospheric model and the Jules land surface model, is affected (an erratum has not been published for the KACE model).

238-239: Why not use internal variability estimated from climate models rather than relying on the one, very limited realization we have been able to observe?