Review – “Estimating return periods for extreme events in climate models through Ensemble Boosting” by Bloin-Wibe et al.

This paper reports a study of a way in which ensemble boosting may be used to estimate the probability of an event more extreme than a reference event by conditioning the boosted ensemble on the occurrence of the reference event. Under suitable conditions, this allows estimates of the probability of the extreme event of interest with a specified level of uncertainty at a lower computational cost than would be possible using a raw, unboosted ensemble. I like this very much because it brings statistical importance sampling concepts to bear on the analysis of storylines, thereby creating the possibility of associating estimates of the probability of recurrence to the cases considered in storyline studies and thus answering a key user question – what are the odds that a damaging event like the one that is considered in the storyline will happen again.

The main issue that the authors struggle with in this paper is how to generate the boosted ensemble in a way that allows them to produce unbiased estimates of the probability of the extreme of interest without incurring excessive computational cost while using a climate model that appears to be relatively expensive to run. (The authors report that the climate model they use is capable of simulating just over 1-year per wall clock day on their computing systems – which is slow compared to simulation rates, such as 5-years per wall clock day, that would be more amenable to this type of research). The conditions necessary to demonstrate that these estimates are unbiased, which are set out at around line 195 in the paper, essentially require the occurrence of the extreme of interest to be independent of the occurrence of the reference event. This leads to a delicate trade-off in which the authors try to use boosted ensembles where the reference event still has some influence on the occurrence of the extreme of interest while not discernably violating this assumption.

I find the trade-off and the way it is considered a bit unsatisfying because there are elements that seem ad-hoc. Also, the resulting boosted sample of short simulations leads to a sample of extremes the are conditional on the reference event that may have different properties than the extreme of interest. The extreme of interest (such as the maximum 5-day mean temperature observed during the heat dome event) can be thought of as a block maximum where the block is either the Northern Hemisphere summer season (JJA maximum) or, equivalently, a year (annual maximum), because we know now that the event was the annual maximum event in that region. Such block maxima are often assumed to have GEV distributions, based on asymptotic theory and a set of idealized mathematical conditions. In contrast, the sample of maxima obtained from the boosted ensemble are

something like the maxima of 18-day (or so, depending on the length of the boosted runs) blocks, with blocks representative of only part of the summer. The facts that the blocks are shorter alters their statistical properties relative to those of annual or summer maxima, even if after 18-days, the short timescale atmospheric behaviour of the boosted simulations is essentially independent of that of the unboosted simulations. An 18-day block maximum is substantially farther from the "domain of convergence" to the GEV distribution than a 90-day block maximum, particularly in cases when the parent distribution is near Gaussian (convergence to the GEV occurs only very slowly for the block maxima of Gaussian samples). Also, note that we would not expect complete independence from the reference event due to the influence of slower evolving parts of the climate system such as the temperature of the ocean mixed layer, sea and land ice, land surface moisture content, and perhaps stratospheric state.

With these complications in mind, while less computationally efficient, if the question concerns an extreme of an annual maximum temperature index where the annual maximum is presumed to occur sometime in JJA, then a simpler way to proceed would simply be to produce a boosted ensemble of MJJA simulations, perhaps conditioned on some aspect of the lower boundary conditions, such as the initial SST state. This would nevertheless produce an additional 3 annual maxima for the cost of a single additional large ensemble simulation given that the latter must run continuously through full annual cycles. It is recognized, however, that if the complications can be dealt with adequately, then using a strategy based on short "forecasts" (e.g., ensembles of 21-day boosted simulations) would clearly be much preferred.

A second issue that is also considered, but requires further thought, is the selection of the reference event, particularly in cases where the reference event is itself very extreme, as in the 2021 western North America heat dome event. This is because selection bias (e.g., see Miralles and Davison, 2023, https://doi.org/10.1016/j.wace.2023.100584) could have a serious impact on the estimate of the event probability in such cases.

Although the methods used to estimate the conditional probabilities are different, the problem considered in this paper is nevertheless somewhat similar to that of producing probability of precipitation forecasts, as is performed operationally at numerical weather prediction centers as part of each forecast cycle (these forecasts are clearly conditional on assimilated antecedent observations). This suggests that tools used for the verification of short-term probability forecasts could perhaps also be used to test the ensemble boosting approach and thus help determine whether departures from the conditions under which boosting can be used to estimate the probabilities of extreme events impede their interpretation. Note that I've framed this thought in terms of the impacts of departures

from the idealized conditions because it is unlikely that those conditions would ever be satisfied fully, just as the postulates that underpin standard extreme value theory leading to the GEV distribution are never completely satisfied in real applications.

Some editorial and specific comments (listed by line number where appropriate)**:**

The overall readability of the paper could be improved. I think the authors would be well advised to further polish the paper with the needs of the reader in mind, which means aiming for text that is somewhat more tutorial that supports reader comprehension as explanations are often very terse. In this same vein, I find the notation a bit awkward, with both superscripts and subscripts denoting time in different contexts. For example, i and n are used to indicate time (as a superscript), i is used to grid row (as a subscript), t is used to indicate time lags (as a subscript) and T means temperature. I don't have good suggestions for simplifying the notation, but complexity of the current notation system impedes reader understanding. The notation system seems a bit non-intuitive and gives the impression that it was the result of a series of quick additions each time something additional was required.

31: Also cite Philip et al. (2022, https://doi.org/10.5194/esd-13-1689-2022)

50-53: This characterizing of the reliability of inferences based on extreme value theory and approaches that "remain statistical" seems a bit disdainful, particularly given that this paper also relies heavily on statistical concepts.

95: Overall, I appreciate this well written introduction.

124: I think this is poorly stated. While the particular interest may be in the upper tail of the distribution of events that could plausibly follow a specified parent event at a given lead time, the boosted ensemble presumably has information about the entire distribution at that lead time conditional on the parent event.

127: "parent event will grow" → "parent event initially will grow" (as you show later, and as is well known, the magnitude of the errors does not grow forever).

135: Why did you choose to perturb Q rather than some other quantity?

161-162: It's not clear what would constitute a "good enough sample", particularly given that "good enough" forms the basis for a key assumption.

177: The assumption that $\mathbb{P}(T > T_{ext})$ is constant throughout the summer is not normally made when using statistical block maximum approaches for estimating upper tail probabilities; I think that assumption should be discussed more thoroughly here.

The presence of the annual cycle implies that daily surface air temperature (including daily values of the daily mean, maximum or minimum) are *not* identically distributed throughout the summer – which is what seems to be assumed here.

179 (eq. 9): Evidently, it is necessary to estimate all three terms on the right well, which has implications for reference event selection from an existing large ensemble and for the size of the boosted ensemble. It would be useful if these implications could be considered.

190: Rather than simply counting events, perhaps other, possibly more efficient methods of estimating these upper tail probabilities could be considered – for example, by computing these probability estimates from fitted extreme value distributions. Doing so would account not just for the frequency of exceedance above some value, but also the form of the distribution of those exceedances.

217: It is unclear which three parameters can be set by the experimenter (see my overarching comment about the notation).

220: I'm not sure I understand what exactly is set to 0.75 or 0.3, nor what that implies for the reference temperature or for the set AC_t^{ε}. See my overarching comment above – a more tutorial "handholding" approach would be beneficial if you want readers to understand well what this paper proposes.

226: I'm a bit confused by the various N's here and what they represent. It seems to me that the reference climate simulations must run through entire annual cycles over an extended period of time if you are going to simulate the entire variance spectrum (including internal variability associated with slow processes). Such long runs provide one annual extreme per year. Under the right conditions, a boosted 21-day simulation could hopefully provide an additional realization of an annual extreme at only a small fraction of the cost of running through an annual cycle (e.g., 21/365 or <6%). This would represent a speed-up of a factor of ~17 …

245-247: The 4000-year preindustrial control contains 80 non-overlapping 50-year segments, so if I were to choose two at random, as implied on line 247, it would be unlikely that they end up being exactly adjacent to each other (a simple combinatorial argument demonstrates that the probability of that occurring at random is 1/40 = 0.025). So, exactly how were those two periods selected? Also, how is time referenced – years since the start of the PI control?

272-281: I'm again a bit confused. First, there seems to be a strong implicit assumption that the 30-member large ensemble is indistinguishable from the 100-member ensemble (despite presumably having been produced at different times by different

groups using different computing hardware, etc., with each group learning by making their own mistakes). If they are indistinguishable, then pooling those two large ensembles together to form an even larger ensemble may allow some further improvement in tail probability estimation before boosting. Second, the idea of "detrending" extremes always raises some concerns because extremes are generally not symmetrically distributed. I don't immediately have suggestions for mitigating potential concerns, but I think at least you should flag to readers that there might be concerns with this kind of procedure in some instances.

282-286: I'm also confused about the numbers 7 and 13. Is 13 a typo (could you have meant 31 – highest value of TXx5d simulated across the 30-member ensemble for each of the 31 years considered?). And why boost only 7 of these events? Was that just a pragmatic decision made in the face of a limited computing budget?

282-290: Something else that is not said is to indicate the details of the time series of daily maximum temperatures is used to calculate TXx5d for the PNW case. Is this a time series of spatially averaged values across some region (if so, what region), or a times series of the maximum value observed over some period across all grid points in a region (again, what region), or simply based on temperatures at a selected representative location (if so, what location)?

374-375: This seems rather ad-hoc. I wonder if this could be formalized in some way to allow a more objective assessment of what the trade-off between relative sampling error and lead time should be. That trade-off implies a source of uncertainty for the estimate of unconditional event probability that you ultimately aim to provide. How that uncertainty depends on the rarity of the reference (parent) event would also be good to formalize, if possible.

394-396: How do you explain this? Note that I wouldn't necessarily accept that the estimates from fitted GEV distributions are "naïve". This seems naïvely dismissive in my view. As I mentioned above, there might be reasons to think that a boosted sample of extremes from short simulations may estimate something different from the thing that can be estimated with annual block maxima.

404-405: Suddenly, there is discussion of an event "E1" without a definition of the event. (please improve the notation …).

420: How do we know that it is a more robust estimate, or that it more realistic than other estimates that suggest that the heat dome was a very low probability event?

502-504: I would agree that extrapolation into the deep upper tail of the GEV distribution has many limitations, but the fundamental limitation is not uncertainty

quantification via bootstrapping, as suggested here. Rather, it is the assumption of max-stability (the notion that block-maxima of samples of block-maxima again have GEV distributions with the same shape parameter). This implies that there will be no “surprises” in the unobserved parts of the tail that do not conform with the behaviour seen in the sampled parts of the distribution. I think this implicitly assumes that there is only one physical process that generates extremes at a given location – while in reality, it might be reasonable to think that the PNW heat dome event reflects the impact of a process that is distinct from the process (or collection of processes) that produces the annual maximum temperature event in the region in most years.