Review – Benestad et al.

I think there is considerable substance in this paper, but confusing presentation and imprecise language gets in the way of making a convincing case for the approach that is proposed. This is apparent even from the title. As soon as one combines the words "upper-limit" with "precipitation" in the same sentence, one evokes implicitly the idea of probable maximum precipitation – ie, an amount that is a physically plausible upper bound. The language in this paper is sufficiently imprecise that confusion about whether the aim is to estimate an upper bound for precipitation arises at several places. This is reinforced even within the abstract, where we read that the proposed method "utilises … to estimate the maximum effect that temperature change can have on precipitation …". Since the paper is about extreme precipitation, and the title refers to "upper-limit estimation", one could be excused for thinking that the interest is in estimating absolute upper bounds for 24-hour precipitation accumulations.

Awkward mathematical notation and poor conversion of the typescript to a pdf document also contribute confusion. At various places the reader has to infill his/her own guess of what symbols are meant to be present in the text, apparently because they simply have disappeared, or have been misplaced, in the process that rendered the review document. More importantly, symbols such as the Greek letter μ seem to be used in multiple ways, with the reader needing to infer the correct interpretation from the context in which the symbol appears. For example, μ is used as the parameter of a probability distribution, as a quantity that varies from month to month, and as a quantile. It would be wonderful if standard statistical conventions could be used for notation. Generally, this means using Greek letters to represent unknown parameters that are to be estimated from observations, Greek letters with hats to denote parameter estimates, upper case Latin characters to represent random variables, and lower case Latin characters to represent realizations of those random variables (either observed or simulated), etc.).

The approach itself seems rather adhoc. The fundamental working assumptions are apparently that

a) an exponential distribution fitted to all non-zero daily precipitation amounts will nevertheless represent most of the upper tail of the precipitation distribution reasonably well, including as far into the tail as 1-in-20-year 1-day events, and that
b) precipitation frequency will not change.

The first, (a), is actually a very strong assumption given the existing body of observational evidence from multiple sources that suggests that extreme daily precipitation is heavy-tailed (that is, Frechet distributed rather than Gumbel distributed). A consequence of assuming the exponential distribution for daily precipitation is that block maxima (such as annual maxima) will converge to the Gumbel distribution (GEV distribution with shape parameter equal to zero) rather than the Frechet distribution. The second assumption also seems a strong assumption given the sensitivity of the interpretation of changes in quantiles of non-

zero precipitation amounts to changes in frequency (e.g., see Schar er al, 2016, doi: 10.1007/s10584-016-1669-2). To be fair, the authors defend (b) in this paper, and have discussed (a) in previous papers. Nevertheless, these are strong assumptions that work against the claim that the method provides robust estimates of an upper (uncertainty) bound for 20-year return values.

A third key assumption is that it is assumed that

c) to the extent that 20-year return values are sensitive to temperature, uncertainty in projected changes in 20-year return values can be bounded by using the 95th percentile of changes in the parameter of the exponential distribution that are obtained using predictors from an ensemble of opportunity of climate models.

Again, I don't understand how this assumption would make this number "robust". Robustness is a specific concept in statistics – an estimator is robust if it is insensitive to outliers or misspecification of the underlying distribution. The working assumption (briefly discussed in the supplement), is that the available ensembles of opportunity represent model uncertainty adequately (uncertainty arises from structural and parametric differences, and from internal variability), and that all simulations from all models are equally representative of variations that can be associated with model uncertainty. Trimming the "sample" at the 95th percentile would, under that assumption, add a measure of robustness from a statistical perspective, but this relies on the strong assumption that the available ensemble of opportunity can be conceived of as a random sample of climate change simulations that is representative of a well-defined population of plausible representations of the climate system.

Some specific comments (numbering refers to page and line number within page):

1, 24: This statement incorrectly describes the information that is presented in the Munich Re report. They count loss events that are weather related, not weather events that are loss-relevant. While the difference is subtle, it is important to understand that the focus of the Munich Re report is losses.

2, 7-9: This is changing; I'm aware of at least one large-ensemble experiment similar to the NCAR large ensemble in which an RCM is being used to construct a large ensemble by downscaling a large ensemble of global simulations.

2, 29-31: Why would this be a more tractable question and have greater prospects of overcoming the problems cited in the preceding lines?

3, 15-18: Why use NCEP/NCAR-1 in preference to some other source of North Atlantic surface air temperature or SSTs. How is GCM output bias corrected?

4, 3-5: It's unclear how this 90% "confidence interval" is constructed. Is it really solved as a parametric problem that involves a distributional assumption, or is this

simply a matter of finding the central 90% range in the ensemble of results obtained by using the available collection of climate simulations? As an aside, I don't think trying to say that the two approaches produce similar results increases "confidence". Also, I would recommend simply referring to an "uncertainty range" rather than a confidence interval, since the statistical basis for calculating a confidence interval seems unclear (see my comments concerning robustness above).

P5, 9-24: It would be useful to start here by giving a complete description for the cumulative distribution function for X, which would include a statement about the probability that X=0, as well as a description of the probability that X>0. Note that standard statistical notation usually reserves lower case letters for probability density functions, and uses upper case letters for cumulative distribution functions.

P6, 15: Why this particular threshold for R^2?

P6, 18-19: I'm not convinced that the strategy has achieved the goal that is stated here. With enough caveats, you might convince readers that you have estimated an upper bound for a 90% uncertainty interval, but that's a far cry from evaluating "the maximum potential effect of temperature changes on the wet day mean".

7, 2-5: What is the physical explanation for the spatial variability that is seen in Fig. 3? Can we consider the spatial variation to be "robust", as opposed to being the result of statistical or modelling artefacts?

7, 22-23 "worst case": See previous comments on the communication of what this paper is trying to do.

8, 11-14: What about mid-latitude coastal regions affected by atmospheric rivers?

8, 24: See previous comments about robustness. Furthermore, because a formal statistical framework is not used to perform this work is adhoc, I think it is unclear whether estimates produced from this approach are more efficient (ie, "make the most out of the available information") than competing estimates. In general, robustness and efficiency can be somewhat opposed to each other, although one objective of statistical robustness is to limit losses of efficiency due to a change or misspecification in distribution. Perhaps one could demonstrate "robustness" by showing that there is only modest sensitivity to excluding or adding the "outlier" model according to some measure of model performance vis-à-vis extreme precipitation.