Reviewer 1

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.

The purpose of the paper is to estimate future return-values in circumstances when there are limited observations, where traditional methods are not applicable. The alternative that we present is calibrated on larger sample sizes (the mean climatology) stretching over longer time periods, which puts more weight on slow processes with long time scales. It is an estimate of the upper limit of the influence of temperature on precipitation in the sense that we assume that the seasonal cycle of the wet-day mean can be explained solely by variations in the temperature in the predictor domain. However, as one of the reviewers pointed out, other factors may also add to a precipitation increase, so it is a bit misleading to call it an upper bout. We use the term potential sensitivity to draw on analogous concepts such as climate sensitivity but include potential, since this assumes that all of the seasonal precipitation variations are related to seasonal temperature variations.

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.).

We used Overleaf and LaTeX for our revised version which has a better handling of mathematical symbols and equations. The greek letter μ is commonly used for the wet-day mean precipitation and f_w for wet-day frequency, and we would like to keep it that way. However, we have tried to make it easier to differentiate between observations and downscaled values by being more consistent in the use of hats to represent values estimated with the downscaling downscaling models.

The approach itself seems rather ad hoc. 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.

The point here with combining the 1-in-20 year events and annual maximum daily values is that we do not extend far into the tails of the distribution and in terms of those samples, we look at moderate extremes. This is different to the traditional methods where the outer parts of the tails are modelled. The point about the frequency, on the other hand, is a genuine caveat that we discuss in the paper. There is also additional discussion of this issue in the SM.

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.

Well, as the reviewer pointed out, we do discuss assumptions (a) in the paper and (b) in several previous papers. The robustness lies in the fact that we use a much larger sample size when calculating the mean climatologies that the downscaling model of mu is based on. We have changed the text to specify that it is "robust to outliers" (because of the larger sample size). It is true that the method is not robust to misspecifications (but no method ever is that) of the underlying distribution, but we do discuss the assumption of an exponential distribution.

As for point (c), our use of the ensemble was to represent local and regional variability of the climate system, which is strongly affected by internal variability. We discuss this in the supporting material, but then say "Nevertheless, the spread of downscaled annual mean temperature from ensemble experiments such as CMIP5 is often comparable to the magnitude of the observed year-to-year temperature variations..." and go ahead and use it as such anyways. We have added a sentence of caution in the main manuscript.

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.

The sentence has been changed to better reflect the content of the Munich Re report.

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.

This may be changing - there may be cases where RCMs are applied to larger ensembles of GCMs - but computational demands are still a limitation and certainly have been in the past. The text has been changed a little so that it refers more to past studies and doesn't exclude all possibility of ever applying RMCs to large ensembles.

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

The text has been changed here.

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?

This reanalysis extends back to 1948 and the surface air temperature is more comparable to output from GCMS in the CMIP experiment than observations-based SSTs. GCM output bias was not corrected and we used spatially aggregated estimates of e_s over a large region (100W-30E/0-40N) as predictors.

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).

It is simply a matter of finding the central 90% range in the ensemble of results obtained by using the available collection of climate simulations. We changed the term confidence interval to uncertainty range.

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.

The case for X=0 is trivial and is accounted for by the wet-day frequency 1-f_w. We change the notation 'f(.)' to 'Pr(.)'

P6, 15: Why this particular threshold for R2?

There has been a tradition that R^2 needs to be at least 0.6 for practical use in terms of skillful forecasts. This is of course subjective.

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".

The description of the estimate has been changed somewhat to emphasise that it is an approximate estimate of future return-values, and that the main advantage is that it is applicable in cases of limited data availability.

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?

The most obvious physical explanation of the spatial pattern is orographic effects as it is limited to the Alpine region and western Norway. The PCA analysis of the seasonal cycle of observed wet-day mean precipitation also pointed to a different precipitation regime in these areas. Since there is an obvious physical explanation and the PCA analysis and regression analysis both pointed to the same geographical pattern, we see no reason to suspect statistical or modeling artefacts.

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

We changed the wording and removed the term "worst-case".

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

The atmospheric rivers may be excluded here, but that depends on their frequency and the character of their seasonal appearance. The weights of the PCA for the seasonal cycle are more typical of convective events here. The atmospheric rivers and convective events represent different phenomena and one should not expect to have one statistical framework that fits all such.

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.

This approach is a hybrid between a physics problem-solving approach and statistical thinking, and therefore will appear as ad hoc to the pure physicist or statistician. The point is the practicality and our approach is more computationally efficient than many other methods. This can be seen because it is applied to a large ensemble of models. The calculations take a very short

time, and the study includes a number of validation exercises to evaluate its merit.

Reviewer 2

Benestad et al. present a methodology for statistical-empirical downscaling of precipitation time series to estimate upper limits for future return levels. The proposed methodology provides a considerable alternative in cases where more explicit estimates are not available. In general, the manuscript is carefully written and accompanied by very detailed supplementary material. In fact, my impression is that in some cases, the reader finds relevant information regarding the motivation and methodological details only in this supplement, and one might discuss if some of the supplementary material might fit better into the main paper.
In general, I recommend publication of this manuscript in NHESS after certain revisions have been made. Below, I provide a list of specific recommendations that the authors might wish to consider when preparing their final manuscript.

Some of the relevant discussions from the supplementary material have been included in the main text, but not the supplementary figures.

1. I acknowledge that the authors use well-studied data from the CMIP5 ensemble. In the context of the present work dealing with estimating future return levels, it would be advantageous if the authors could briefly summarize some information on potential known biases (if there are any) of the considered projections.
There is a number of biases, but we used aggregated results in time and space, and found that this then had little effect.

2. Referring to the statement that “heavy precipitation will become more severe in

already wet areas in the future “ (p.1, ll.25-26), I was wondering if this holds globally in all such regions.

We have not checked other parts of the world, as observations are sparse and missing. It is reasonable to infer that this may nevertheless be true if our selection can be considered a random sample from the planet that is representative of the entire system. Convective processes are more or less a universal process on Earth’s continents, but may be different over the oceans.

3. The proposed method relies on inferred statistical relationships between different climate variables. A few more words on possible limitations of these relationships (from both physical principles and empirical observations) would be useful.

We refer to the scaling between the two as ‘potential sensitivity’ exactly to communicate potential limitations.

4. The authors state several times that their estimates provide upper limits. From the presented material, I did not fully understand why this is the case. The argument seems to refer to the relationship between the variables used for empirical-statistical downscaling; however, the results would only be an upper limit if other (unconsidered) covariates would have exclusively opposite effects and could not enhance the considered relationship. Is it possible to rule out (from physical principles) possible “positive interferences” between different variables possibly influencing precipitation?

Good point. The text has been changed somewhat because the term upper limit can be misleading, and we use the term potential sensitivity. However, it still is an upper limit, but we try to explain this more carefully.

5. The proposed method is based on an exponential distribution of 24-hours precipitation sums, whereas I would naively expect a gamma distribution being a more common statistical model (even though the simple scaling from the behavior of the mean to that of arbitrary quantiles would not apply anymore in such case). I would be interested in some additional details on why the exponential distribution is justified here.

The exponential distribution as a description of precipitation is discussed in several previous papers, referenced in the paper. We used the exponential distribution for simplicity as it requires the estimation of just one parameter which is the mean.

6. The choice of the reference region in the North Atlantic appears to be motivated by general climatological considerations rather than statistical optimization. Could the results of the empirical-statistical downscaling be further improved by explicitly seeking for the strongest statistical relationships between predictor and predictand fields? Specifically, as the authors recognize, their predictions are not very convincing in regions with complex orography – could this be because the predictors are not appropriately chosen for these locations in terms of their geographical spread? Can the considered relationship be assumed to be essentially homogeneous over entire Europe?

Some tests of predictor domain were conducted on a few test sites, but no systematic optimization. The fact that the predictions are not useful in complex terrain is most likely due to different processes influencing the orographic precipitation (atmospheric circulation etc) than convective precipitation (temperature and moisture). (The same spatial pattern was seen also in the PCA of the seasonal cycle of the wet-day mean.) The predictor domain should be adapted to the predictand to reflect the main moisture source, but this domain should work ok in other parts of Europe as well.

7. In Eq. (1), is the considered noise term white or serially correlated?
White
8. The PCA in Section 2.4 predetermines mutually orthogonal annual cycle shapes in PC1 and PC2. It is not clear if and why this is desirable in the present case. Specifically, what the authors consider here in terms of the coefficients of PC1 (PC2) is closely related to the phase of the annual cycle, since both components essentially generalize the role of sine and cosine functions in case of a fully harmonic oscillation (PCA is commonly based on normalized time series, so amplitudes do not matter that much). It might be useful to directly refer to some corresponding phase variable to parameterize the shape of the seasonal cycle for each considered location.
The PCA is simple and not constrained by the shape (like sinusoids), and we wanted to identify the covariance structure in the mean seasonal variations. The orthogonality is nice when using them in regression analysis, but of course, the variations themselves are a superposition of several modes.
9. In a few figures (in both main manuscript and supplementary material), axis labels and labels/units to color bars are missing. This should be carefully revised. In Fig. 3, it is not clear if the inset shows relative or absolute changes.
We have improved the figures.
Technical comments:
* p.3, l.3: “Our approach...” would rather call for using present tense.
The sentence has been changed/removed.
* p.4, l.18: mathematical symbol missing after “referred to as”
A lot of mathematical symbols were missing because of a failed conversion to pdf. The revised manuscript is written in LaTex and all the symbols should be correct.*** p.4, ll.25: “the ration between explained variation... and the total variation...”
This sentence has been removed.
* p.4, l.25: “var() with the noise term is taken to be zero” is not quite understandable, please rephrase
This sentence has been removed and the R^2 calculations are now explained (and hopefully more understandable) when the results are mentioned later in the manuscript.

* p.4, l.25: “Principal component analysis”
* p.7, l.13: “constant value of the...”
* p.8, l.1: “dependency... on temperature”

There are also a few typos in the supplementary material that are not listed here for brevity. In general, in some of the R outputs embedded in the SM text, the meaning of the individual variables is not fully clear without consulting the full R code; at least identifying the variables in the corresponding text boxes would facilitate the reading. Also, I did not find a caption for Fig. SM14.

We have rearranged and rewritten parts of the text to be more easy to follow. Some of the supplementary figures that were not explicitly referenced in the main or supplementary text have been removed. The R-scripts included has been updated and some more explanation added so that it should be easier to follow it.