Review
Global estimates of 100-year return values of daily precipitation from ensemble weather prediction data

This paper's goal is to determine 100 year return values of annual maximum daily precipitation from a quasi-observational dataset (ensemble forecast) and compare these to 100-year return values from actual (semi-)observational products – both ground-based and satellite-based. In order to do so, generalised extreme value distributions are fitted to the annual maximum daily precipitation data (both the ensemble forecast and the observational data, it seems) and the 99% percentile value (1/100 years) is extracted. By comparing the results based on the large ensemble forecast and the observational data, the authors conclude that the estimated 100-year return levels are higher almost everywhere in the ensemble forecast data compared to observations, and the uncertainty range of the estimates, based on non-parametric bootstrapping, is smaller for the ensemble forecast data than for observations.

I think this study asks a relevant and useful question, and is thorough in its use of several observational products for comparison. The general choice of methods (extreme value theory) is appropriate for the purpose of the study. The manuscript is easy to read, albeit sometimes a bit too heavy on the details. I think this could become a good paper, but I see major methodical flaws that would need to be addressed first. In addition, the reproducibility is not guaranteed with the current level of detail in the method descriptions.

Below I first expand on these two major issues, followed by a list of more specific comments in order of appearance.

Major methodical flaws
The main, general problem is the lack of caution and thoroughness in the use of observational data. This manifests, firstly, in the lack of contextualisation of observational uncertainties for the data products used, and in the way statistical quantities based on very short observational records are presented and compared to EPS-ensemble results, without enough attention for the instrumental differences and the bias and uncertainties in observational results. As a consequence, the results and conclusions of the authors are likely to be misunderstood; they seem to suggest that observations show lower and more uncertain extreme precipitation than the ensemble forecast. In fact, the observations themselves are not what causes the difference, but the data processing and the "unfair" comparisons made.

1. My main concern is the use of observational data. As far as I can tell from the method description, observed 100 year return level estimates are based on the three observational datasets with lengths of 38 up to 65 values. These datasets are so short that a lot of caution is warranted when return periods much longer than the sample length are assessed. It is, even with much data, notoriously difficult to get the tail of extreme value distributions right. Furthermore, it is known that there is a systematic low bias when return levels are estimated from small samples (for shape parameters <0.5). This is all not taken into account enough in the generation and presentation of results.
   Furthermore, the authors compare results obtained from samples with N < 70 directly

to results obtained from samples with N > 1200, both for return levels as well as confidence intervals. The effects of sample size on this comparison are likely to be larger than any true systematic difference.

Lastly, as far as I can tell, confidence intervals for observational estimates are determined using non-parametric bootstrapping as for EPS data, i.e., re-sampling those small datasets. This produces a confidence interval, but if the original sample did not contain enough information to reflect the true underlying GEV tail accurately, any bootstrapped samples will not either. More importantly, comparing confidence intervals based on samples that differ in size by a factor of 15 (EPS:obs) is not very instructive, since the “real” differences are obscured by the differences caused by the different sample sizes.

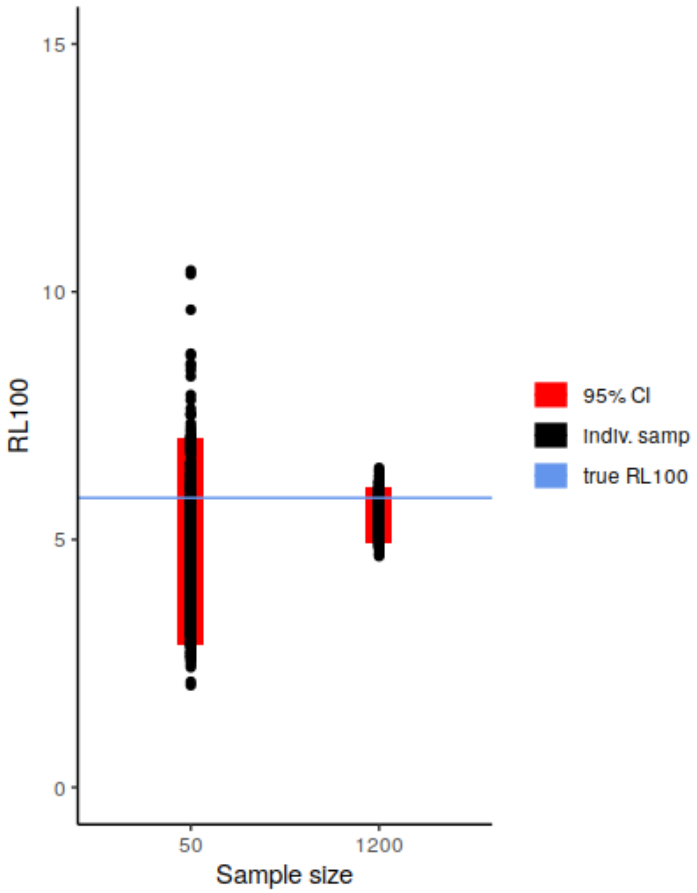
Just for illustration, here is a tiny R-script and the resulting plot. The script computes the RL100s of two random GEV-distributed samples, one of size 50 and one of size 1200, using 1000 bootstrap resamples. The original 50 and 1200 samples come from the same prescribed GEV distribution as you can see. The plot, showing RL100 values and the bootstrapped 95% confidence intervals, reveals that the n = 50 sample shows a smaller RL100 with a larger bootstrapped confidence interval, despite coming from the same GEV distribution. The fact that the results in the paper also show this, can thus not be ascribed with certainty to anything else than the fact that the sample sizes are so different. (If the n=50 sample contains one or more very extreme values, the result is vastly different, hence the caution needed when dealing with such small samples.)

```
bigs = revd(1200, 0,1,0.1) #define random GEV-distr samples
smalls = revd(50, 0,1,0.1)

funz<-function(data){return(return.level(fevd(data), 100))} #
    function to determine RL100 from bootstrapped samples

dtp = data.frame("n" = c(rep(1200,1000), rep(50,1000)),
                       "RL100" = c(booter(bigs, funz,
    1000)$results,  booter(smalls,funz,1000)$results)))

#dataframe longformat with individual resampled RL100 values as a
    function of sample size n
```



A possible way to address these issues, might be to generate independent subsamples of the EPS dataset of the length of the observational datasets, and use these subsamples to determine “quasi-observational” 100-year return levels (RL100), as well as a confidence range (the range spanned by all these subsamples). These can be directly compared to the RL100 values obtained from the observations. Subsequently, the EPS RL100 values based on these small subsamples, can be compared to other resampled EPS samples of progressively increasing sample size. In this way, the effects of the different data type/source can be separated from effects due to sample size.

In addition, a synthetic data study could be done to quantify the margin of error and potential bias in estimates based on small datasamples. E.g. one could generate 1000 independent samples of different lengths comparable to the observational datasets, based on random sampling from a known GEV distribution (or several, representative of several regions, for example). Then one could fit GEVs to these samples, and derive return levels from these GEVs. The “empirical” return levels as a function of sample size can be compared to the known true return level. See e.g. Zeder et al. (2023).

Naturally, there are other ways to make a justified comparison between the EPS data and the observational data, and to assess the sensitivities to the data properties. In

any case, the analysis should be much more careful around determining observed RL100 with the data at hand.

In the discussion it is indeed mentioned that the main cause for the differences in confidence interval is the sample size difference. This should get more attention earlier in the paper. If the main point of the paper is to show that the observational record length leads to biases, it should be restructured and backed up with more statistical analysis so that that point comes across more clearly.

2. Also on the general use of the observational data I have some concerns. Firstly, REGEN comes with a mask reflecting quality/trust in the interpolated values (for many locations, there are hardly any measurements). It would be good to acknowledge this fact, perhaps simply use the quality masked dataset, or at least show where confidence in the REGEN data is generally low, regardless of sample size. Also, REGEN provides Rx1d values (annual precipitation max) (e.g. on https://climdex.org), computed in a way that tries to best reflect actual annual maximum precipitation. I'd recommend using this product directly, instead of computing it from daily values.

3. Lastly, if CHIRPS and PERSIANN do not provide Rx1d products, they have to be calculated of course. In this procedure, the order of operations matters. As far as I can tell, the authors first regrid the data, after which they determine the annual maxima. This results in significantly lower values than the reverse order of operations due to spatial smoothing of extreme values. The preferred order of operations is this: first extraction of Rx1d, and then (conservative!) regridding. This order of operations conserves the intensity better. Given that the absolute magnitude of Rx1d return levels is central in the analysis, these details affect the results.

Major content flaws

4. Essential information from the methods is missing, especially on the way observations are processed to obtain 100 year return levels and confidence intervals. I assumed the same way as the EPS data, but this is not clearly stated.

Minor comments

General

1. Might it make sense to focus on precipitation over land? Two of the observational datasets have land coverage only anyway, and the ocean signal is so strong that it obscures the patterns over land due to the colourbar scaling being adjusted to the ocean mainly.

2. Some references are not properly displayed, e.g. "Organization, 2009".

Specific

L6-10: In light of my general comments above, these conclusions might need revision.

L99-100: I would think that another reason to not use all 10 days of the forecast, even if the members were uncorrelated, is that you would have the same day in the ensemble multiple times, because there'd be overlap between forecasts made on day n and day n+1, which would introduce more dependence between individual values.

L127-128: “regridded”/”interpolation scheme MIR”: more information is needed here. Is 1 by 1 degree the lowest resolution found in the original data? Is all data “upscaled” to lower resolution, or is some downscaled? What is MIR? What kind of interpolation scheme does it use (bilinear, conservative etc.)?

Sect. 2.2.1: as mentioned above, the data quality/confidence in REGEN is low for a large part of the global land, it would be good to mention and show this. I’d even recommend using the quality masked dataset.

L156: “interpolated”: how, which scheme/method? Also, see major comment 3 on order of operations.

L167: “interpolated”: see previous comment

L169: “set to 0”: this is not necessary and does not introduce problems if the annual maximum is used only. See comment below for L211

Sections 2.2.1-.2.2.3: It would be useful if, for each of the 3 obs datasets, the details were summarised, such as the length and coverage (land only, <60N/S only etc) of the record.

Section 3: Major comment 4: the methods for return value and confidence interval determination from observational data is missing.

L191-194: suggest to remove: the results for the river catchments do not matter here.

L208-211: Also here it is not necessary to report on what Ruff & Pfahl found in their previous study.

L211: “50th and 90th percentiles”. I wonder why this is done in this way, given that the analysis is about 100 year return levels of annual maximum precipitation, I do not think that the 50th and 90th percentile of daily precipitation really matter much. If there is good agreement at medium low percentiles, that does not guarantee that there is good agreement for the >99th percentile (where the annual max is location) as well. I would suggest assessing agreement in the distributions of Rx1d (the median and some measure of spread of the Rx1d distribution, for example).

L239: “for the observational datasets”: it is not clear to me how the trend is computed - Supplementary Fig. 2 suggests it is also the 99.9th percentile trend. Does that mean the 99.9th percentile is determined for each year based on the 365 daily precipitation values? Might it make more sense to simply assess the presence of a trend in the timeseries of annual maxima for both EPS and observational data?

L256-257: “sufficient for the Fisher-Tippett theorem” I am not sure what is meant here. The number of blocks does not determine whether the data is GEV-distributed. It needs to be i.i.d. and max stable (and a large enough dataset to contain enough information).

We already know the data are more or less i.i.d. based on the correlation study, but testing max stability (e.g. qq plots) would be good.

L258: “estimating the location, scale and shape parameters”: which method is used to estimate these?

L277-279: Perhaps L-moments (assuming MLE was used in the initial GEV fits) provides more robust results with less excessive parameters. For stationary GEVs, L-moments is generally the better way, see e.g. Hosking (1990).

Results section: using mm/day instead of mm would be more specific.

L285-292: This is a matter of taste perhaps, but I find the lengthy listing of absolute RL100 values not very useful. This also holds for lines L300-305, (partly) L326-355 (all the specific values can be seen in the figure, the text should contain interpretation rather than listing the values), L 370-381.

L293-298: In and of itself this section seems a bit lost. However, in L328, the authors mention “no clear pattern” in the relative CI-magnitudes, but there is: it is exactly the pattern in Fig. 5. This makes sense: where the spread is very large, and the tail long and thin, it is very difficult to estimate the tail percentiles. I would nonetheless suggest moving Fig. 5 to an appendix, and referring to that in L328.

L299-319: See major comment 1.

Fig. 4: The stippling is a bit too tight, so it becomes more like a gray haze.

L321-355: See major comment 1.

L337: What is p.p.?

Fig 7: I think the relative CI is the relevant quantity (Fig 6). Suggest to remove or move Fig 7 to appendix.

L364-365: It would be good to expand a bit on how the non-independence of the data in the tropical oceans and maritime continent might have affected the results.

L397: “substantially reduced in the EPS data”: see major comment 1.

References

Zeder et al. (2023): https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023GL104090
Hosking (1990): https://www.jstor.org/stable/2345653