The authors thank the referees for their useful and constructive comments. Replies are provided below in red.

Referee #1:

General comments:

Thank you for making the changes. Here are some newer details that I have spotted. The minor comment is very minor, but the major comment may require some minor checking.

Major Comment:

Pg 12, lines 19-22: You could probably do some quick areal reduction factor (ARF) calculations to show how good is to assume high resolution gridded data can be compared with point data without correction. I do not know the ARF guidelines for US or Canada, but UK ARF guidelines (Kjeldsen 2007; see Chapter 4, free PDF is available at the referenced website) suggest with WRF, CMORPH and NSWEP grid at 39N (Washington DC) to have ARF ~ 0.9, 0.85, 0.8 respectively for 1hr precipitation. For 24h precipitation, ARF > 0.95. Of course, ARF themselves are approximations and region specific, but it is probably a good idea to comment on that how may affect your results. For example for your Fig 5, failure to account for areal averaging may push your MLAR curves downward; WRF's MLAR for 1h is negative about -0.08 (?); assuming UK ARF, log10(0.9) ~ -0.05, so your WRF results may actually look better than the plot indicates. Overall, your comment (Pg. 14, lines 23-24) that WRF is probably better than CMORPH in the representation of the current-climate rainfall extremes remains true (in fact WRF may be better than you think).

An analysis of area-to-point corrected MLAR and MALAR statistics has been added to the text and supplemental material. Specifically, Figures S3-S5 provide analogues of Figures 5, 7, and 8 (MLAR and MALAR plots) but with prior adjustment of the gridded data quantiles using inverse areal reduction factors based on Kjeldsen (2007). As the referee surmised, the WRF results now outperform CMORPH and are essentially unbiased for durations from 15-min to 6-hr.

Minor Comment:

Equation 6: Define the meaning of U (uniform distribution) and N (normal distribution) and their bracketed parameters; some readers may not be familiar with that notation style.

Done.

Kjeldsen, T.R.. (2007). The revitalised FSR/FEH rainfall-runoff method. Centre for Ecology & Hydrology. Wallingford. Retrieved from https://www.ceh.ac.uk/services/flood-estimation-handbook

The authors have definitely improved the manuscript and have addressed most of my comments in the revised manuscript. The revised manuscript can be accepted for publication subject to two minor suggestions.

1. The authors had replied to my previous comment #3 in the discussion. But I do not see any mentioning of it in the manuscript. I apologize if I missed it in the manuscript. I request authors to make it clear if they have included the response to the revised manuscript. It should be clearly seen in the text if a significance test for MLAR and MALAR have been added. If no such significance test has been added, the authors should make it clear as to why a significance test cannot added?

Confidence intervals have been added to the MLAR/MALAR plot for the empirical quantiles (Figure 5), with the following explanation added to the text:

Values are accompanied by 95% confidence intervals estimated based on 1000 bootstrap samples drawn from the series of annual maxima at each location.

Similarly, 95% credible intervals have been added to the MLAR and MALAR plots for the GEVSS estimates (Figures 7 and 8), with the following explanation added to the text:

In all cases, posterior means and 95% credible intervals for the MLAR and MALAR statistics are estimated from the posterior distributions of the GEVSS parameters and the resulting return levels.

Results are used to infer statistical significance.

2. The authors have added the text “Grid points in each dataset were matched with the nearest neighboring station.” in line 25 of page 12. As far as I understand, the grid point nearest to the station is picked for computing “RE”, “MLAR” and “MALAR”. However, text in the last paragraph on page 13 mentions “station and grid box annual maxima were pooled for each given location”. This confuses me. Did you include “all” grid points surrounding the station to pool with the station data or just “one” grid point nearest to the station? Please make it clear in the text.

One grid point nearest to the station was used in all cases. The text has been reworded to make it clear that “pooling” in this case means combining durations for a single location, rather than any pooling of different locations in space:

Under this null hypothesis, station and grid box **annual maxima at a given location are combined into a single sample**. The combined data are then randomly reassigned to two permutation resamples (i.e., two series of shuffled annual maxima) having equal length as the original, unpermuted station and grid box samples. For each station-grid box pair, the distribution of the RE statistics is approximated using 5000 random permutation resamples and the p-value is computed as the fraction of resamples generating RE absolute values equal to or larger than those observed on the original annual maxima samples.