# Reply to "Comments on 'Flash Flood Verification: Pondering Precipitation Proxies'"

RUSS S. SCHUMACHER[a] AND GREGORY R. HERMAN[a,b]

[a] *Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado*

(Manuscript received 17 November 2020, in final form 2 December 2020)

ABSTRACT: We applaud Gourley and Vergara for their thorough investigation of the relationship between precipitation and flash flood reports, as well as their inclusion of information from advanced hydrologic model output. We conducted some additional analysis to identify the reasons for the substantial differences between their findings and ours. The primary reason for the differences was found to be temporal sampling. The high temporal resolution of the MRMS dataset, as well as their use of "rolling" accumulation periods, explains most of the discrepancies. For guidance related to real-time warning decisions for flash flooding, Gourley and Vergara's analyses provide an important new guide and we recommend the use of their results for this purpose. For other applications, including model postprocessing and for precipitation datasets with lower temporal resolution, our results will continue to prove useful.

KEYWORDS: Extreme events; Flood events; Precipitation

## 1. Introduction

We appreciate the thorough and compelling analysis provided by Gourley and Vergara (2020a, hereinafter GV20) in their comment on our 2018 article (Herman and Schumacher 2018a, hereinafter HS18). The Multi-Radar Multi-Sensor (MRMS; Zhang et al. 2020) system represents a major advance in quantitative precipitation estimation (QPE) and in integration of atmospheric and hydrologic information. QPE is very challenging, and the need for accurate QPE for both research and operations is immense.

In an analysis of multiple QPE datasets in HS18, we found that relatively common rainfall accumulations, such as exceeding the 1- or 2-yr average recurrence interval (ARI) in 24 h, or exceeding 63.5 mm in 24 h, corresponded better to reports of flash flooding than rarer rainfall occurrences. These results were somewhat unanticipated, and in many ways, the findings of GV20 are more in line with what one might expect, with shorter accumulation periods and larger accumulations being associated more closely with flash flooding.

Curious about the reasons for the differences between GV20's findings and ours related to the proxies' correspondence to flash flood reports, in this reply we explore the potential reasons for the differences between the HS18 and GV20 findings and comment on the implications of these differences.

## 2. Data and methods

In this analysis, we use the hourly MRMS gauge-corrected QPE, which is the same MRMS dataset used in HS18. This differs from the version used by GV20, which is the radar-only dataset, includes additional polarimetric information, and has output every 2 min. We use the same study period, 31 May 2018–1 June 2019, as GV20, and the flash flood report (FFR) database provided by GV20 (Gourley and Vergara 2020c).

We generally follow the methods of HS18 here, which includes first regridding the MRMS analyses to the 4.75-km Hydrologic Rainfall Analysis Project (HRAP; Fulton et al. 1998) grid to identify exceedances of various thresholds. FFRs are mapped onto this grid using a 40-km radius of influence, projecting a single report onto numerous points on the grid. Then, both the exceedances and the flash flood reports are upscaled to a 0.5° × 0.5° latitude–longitude grid over the contiguous United States. A maximum of one "event" per 1200–1200 UTC "meteorological day" is recorded at each grid point, regardless of the number of individual exceedances at that grid point.

To identify reasons for differences between the HS18 and GV20 findings, some modifications are made to these methods; these will be described in more detail below.

## 3. Results

There are a few possible reasons for the large differences between the findings of HS18 and GV20 that we will investigate:

1) the different period of study;
2) differences in temporal resolution and temporal sampling;
3) the different MRMS dataset, including the use of radar-only versus gauge-corrected QPE, and the enhancements made in version 12 of the MRMS system; and
4) other factors such as the regridding methods.

### a. *Period of study*

The first possible reason for the differences would be the different time period analyzed; perhaps a different set of weather events would yield different results. To investigate this, we repeated the analysis of HS18 for a subset of ARI thresholds, over the period studied by GV20. Although quantitative differences arose between the two time periods, we reached the same overall

[b] Current affiliation: Amazon.com, Seattle, Washington.

*Corresponding author*: Russ Schumacher, russ.schumacher@colostate.edu

DOI: 10.1175/JHM-D-20-0275.1

conclusions as in HS18: generally lower ARIs had better correspondence with FFRs than higher ones, and 24-h totals were better than 6-h accumulations.

Specifically, we examined the 2-, 10-, and 100-yr ARIs for 24-h rainfall, and the 10- and 50-yr ARIs for 6-h rainfall. Of these, the highest equitable threat score (ETS) (0.126) was found for the 2-yr, 24-h ARI, and the lowest ETS (0.031) was found for the 100-yr, 24-h ARI. For the 10-yr ARI, the ETS for 24-h accumulations was slightly higher than that for 6-h accumulations. This suggests that the differences between the two studies were not simply a function of the specific weather events within the different time periods chosen.

### b. *Temporal resolution and sampling*

The primary reason why the rarest events corresponded poorly with FFRs in the HS18 study is that they are (by definition) very infrequent. For example, during the GV20 study period and after the upscaling to a common grid, we found that there were nearly 9 times as many FFRs as there were 100-yr, 24-h ARI exceedances. As such, it is unsurprising that the ETS for this threshold would be poor: the probability of detection is very low. However, GV20 found a slightly *greater* number of 100-yr, 24-h ARI occurrences than FFRs (Gourley and Vergara 2020b). Although GV20 pointed out that the improved radar-only version of the MRMS had increased precipitation at high thresholds, this seems unlikely to explain such a large discrepancy.

The explanation for the discrepancy instead comes from a subtle, but apparently significant, methodological difference between HS18 and GV20. In HS18, we were comparing multiple datasets with different temporal resolution, and all of the datasets were processed in the same way. For 24-h fixed threshold and ARI exceedances, we used the 24-h accumulation for the period ending at 1200 UTC. This was done in part because the NCEP Stage IV analysis receives different quality control procedures for 24- and 6-h analyses than for hourly analyses, and the sum of hourly grids does not necessarily equal the 24-h analysis (e.g., HS18). Likewise, 6-h exceedances were calculated four times per day at 0000, 0600, 1200, and 1800 UTC, owing to the 6-hourly resolution of the CCPA dataset.[1] Because ARIs are defined based on accumulation periods rather than periods ending at a particular time of day, this method of sampling necessarily underestimates the true number of occurrences. However, owing to the availability and temporal resolution of QPE data, it has been used in previous studies, including HS18.

In contrast, the radar-only MRMS data feature both a much higher temporal resolution (updates every 2 min), and "rolling" calculations of ARI exceedances. In other words, a 6-h ARI exceedance might be identified for the period from 0300 to 0900 UTC, or from 0302 to 0902 UTC, and so on. Such a rainfall event would not be fully sampled by either the 0000–0600 or 0600–1200 UTC time periods, and thus would likely be recorded at a lower ARI threshold. In other words, the higher-frequency sampling might find this hypothetical storm to exceed the 50-yr ARI from 0300 to 0900 UTC, but lower-frequency sampling might only find it to exceed the 5- or 10-yr ARI in either of the defined 6-h periods. As a result, the GV20 method is likely more closely sampling the "true" precipitation distribution in a way that has not previously been possible.

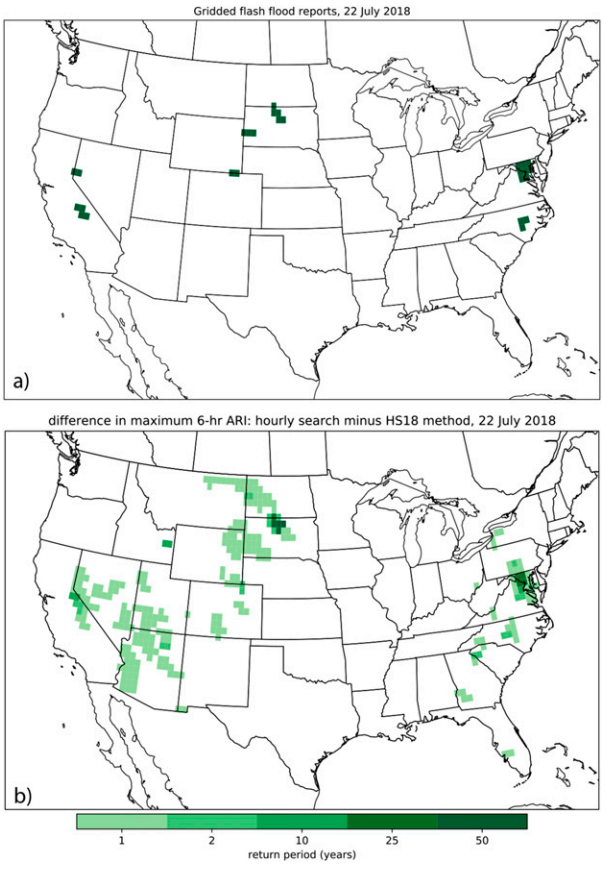


FIG. 1. (a) Flash flood reports on a 0.5° × 0.5° grid for the day ending 1200 UTC 22 Jul 2018, from the dataset provided by GV20. (b) Difference (in years) between the maximum 6-h ARI exceeded using rolling 6-h periods and using the HS18 method of 6-h periods ending at 0000, 0600, 1200, and 1800 UTC.

To test the sensitivity of the HS18 results to temporal sampling, we reanalyzed the gauge-corrected MRMS QPE, but searched for exceedances every hour for all thresholds, instead of only once per 3, 6, or 24 h. (Note that the version of MRMS we used has hourly resolution.)

This methodological change has the effect of greatly increasing the frequency of exceedances, and shifting the results to be more in line with what GV20 found. For the example of the 100-yr, 24-h ARI mentioned above, we found that using rolling 24-h periods approximately tripled the number of exceedances identified. The increase stems both from the more frequent sampling, and the fact that some events are counted on two separate days. For example, if a point is found to exceed the threshold for the period from 1000 to 1000 UTC, and also from 1300 to 1300 UTC, it would be counted on both meteorological days. This effect has potential disadvantages and advantages: heavy rain from a single weather system will often be counted

[1] Since the HS18 study was completed, the temporal resolution of the Climatology-Calibrated Precipitation Analysis (CCPA) dataset has been increased to hourly.

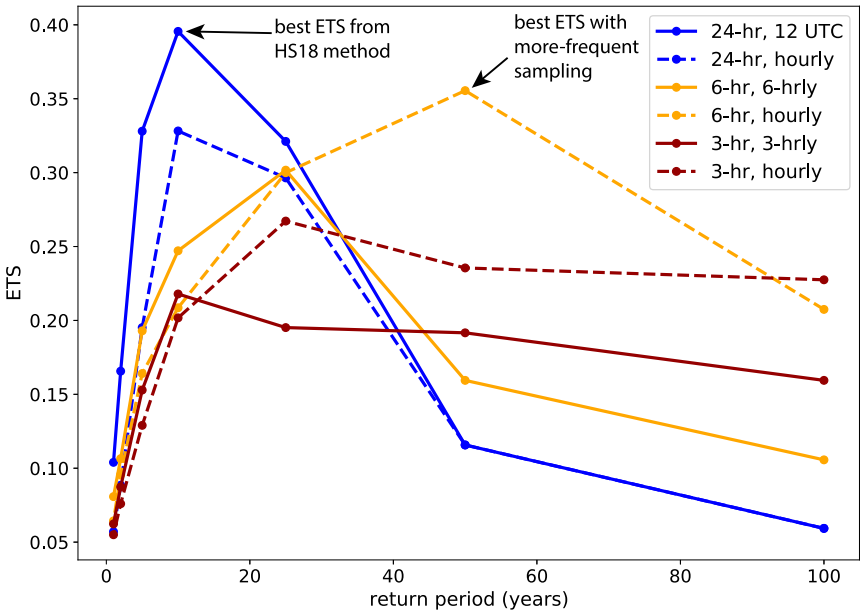


FIG. 2. Equitable threat score for the HS18 method (solid lines) and for rolling hourly searches (dashed lines) at the 24- (blue), 6- (orange), and 3-h (brown) accumulation durations for 22 Jul 2018.

more than once, but because flooding occurs after the causative rainfall, a flood report on one meteorological day may have been caused by rainfall on the previous day.

We will illustrate these effects using an example of a single day, 22 July 2018, which had numerous FFRs in different parts of the country (Fig. 1a). For 6-h precipitation, using the HS18 method, rain accumulations exceeded the 5–25-yr ARI in most places where flash flooding was observed, with areas in southern Wyoming, southwestern Utah, and western Nevada exceeding the 100-yr ARI (not shown). However, when using a rolling hourly search, locations in South Dakota and near Washington, DC, were found to exceed the 50-yr ARI for 6-h rainfall, whereas the HS18 method identified the same grid points at the 10–25-yr ARI. Figure 1b shows that the differences between the two methods in the maximum 6-h ARI that was exceeded were small in most areas, those two locations, which indeed had observed flash flooding, exceeded much higher ARIs with the rolling search than the fixed 6-h periods.

These differences, in turn, substantially alter the quantitative evaluation of the correspondence between rainfall and flash flood reports. For this day, the HS18 method found the highest equitable threat score (ETS) to be associated with the 10-yr, 24-h ARI (Fig. 2), with generally worse correspondence between higher ARIs and shorter accumulation periods. The conclusion using rolling hourly searches is quite different: the highest ETS is for the 50-yr, 6-h ARI (Fig. 2). Overall, more frequent temporal sampling shifts the distribution of ETS toward rarer events at shorter durations. Focusing on the 6-h duration (orange lines in Fig. 2), increasing the sampling frequency reduces the ETS at the 1–10-yr ARIs, but increases the ETS at the 50–100-yr ARIs. The same is true for the 3-h accumulation duration. The results for this single case are consistent with the subset of ARIs and durations that we reanalyzed over the full period of study.

Although we have only compared the difference between the HS18 method and *hourly* sampling here, the updated dataset presented by GV20 allows for sampling every *2 min*. This will identify even more exceedances, amplifying the effects described here. Thus, we conclude that the primary explanation for the disparate results found in the HS18 and GV20 studies is the difference in temporal sampling.

### c. *Other factors*

One other potential difference between the two studies is the gridding procedure. Although both studies ultimately use data on a coarse grid, HS18 regridded all precipitation analyses to a common grid (the HRAP grid), to allow for fair comparisons, before calculating the exceedances and upscaling. GV20 identified exceedances on the higher-resolution 1-km MRMS grid before upscaling. The higher spatial resolution also has the potential to identify more events, even if all else is equal. We conducted a small set of tests of the QPE data resolution, and in these limited results we found this effect to be nonnegligible, but smaller than the effect of temporal sampling.

Last, GV20 reported that the new version of MRMS they analyzed has increased QPE at higher rainfall amounts, which, as noted by GV20, explains some of the differences between the two studies.

## 4. Summary and recommendations

In their study, GV20 provide a detailed analysis of the correspondence between rainfall accumulation and flash flood

reports in the CONUS. Their results revealed some notable differences from what we found in HS18, and in this reply we explored the reasons for these differences. We found that the primary reason for the differences between the two studies was the temporal resolution and sampling. In particular, increasing the resolution and sampling shifts the correspondence between precipitation and flood reports from shorter return periods and longer durations, to longer return periods and shorter durations. In other words, a rainfall event that HS18 found to be a 10-yr, 24-h event might be identified as a 50-yr, 6-h event if using higher temporal resolution and more frequent sampling.

These results have numerous implications for different applications. GV20's analysis provides a quantitative and comprehensive study of the correspondence between flash flood reports and the suite of MRMS products. For operational forecasters assessing flash-flood threats in real time, with access to the full array of MRMS data, we strongly recommend using the proxies suggested by GV20: the higher thresholds are likely the best to use when considering the extremely high temporal sampling available with the MRMS, whereas the lower thresholds suggested by HS18 would likely yield many false alarms. Furthermore, GV20's analysis of FLASH output represents a major advance, by providing valuable hydrologic information, integrated with QPE, that has not previously been available.

However, this recommendation is more challenging for researchers to abide by. All other widely used precipitation datasets are only updated every 1, 6, or 24 h (e.g., the Stage IV, CCPA, or PRISM). Our study was originally motivated by determining which ARI thresholds work best for training machine learning algorithms for excessive rainfall prediction, which requires a long data record that does not yet exist for the MRMS; we have found that the 1- or 2-yr ARIs generally work well for this purpose (e.g., Herman and Schumacher 2018b) when using less-frequent sampling on datasets with lower temporal resolution.

Furthermore, it is becoming increasingly common to compare numerical model QPF to ARI, FFG, or fixed precipitation thresholds to identify the potential for heavy rainfall (e.g., Albright and Perfater 2018). Model output is generally available at hourly intervals, and the calculations used to identify exceedances do not generally use rolling accumulation periods (J. Correia 2020, personal communication). Thus, using lower ARI thresholds like the ones found in HS18 or the 5-yr ARI used by Erickson et al. (2019) may be more appropriate in these settings as well.

The MRMS system represents a major advance in QPE and a paradigm shift for hydrometeorological applications. There currently does not exist a publicly available archive of the full suite of MRMS products, and we strongly encourage the MRMS developers to make one available, so that researchers and practitioners can take advantage of these new advances and employ them in their work.

We appreciate the compelling analysis in GV20's comment, and thank them for the opportunity to further explore issues related to analyzing QPE in the literature. In all, the findings of HS18 and GV20, and their differences, highlight the importance of understanding the details of the datasets used in research and operations, and choosing the dataset(s) that best suit the desired application.

*Acknowledgments.* The authors thank J. J. Gourley and H. Vergara for making their data readily available online, which facilitated the analysis comparisons presented in this reply. The research in this reply was supported by NOAA Grant NA18OAR4590378.

*Data availability statement.* Flash flood report data come from the file provided by Gourley and Vergara (2020c). Archived gauge-corrected hourly MRMS analyses are available from the authors upon request.

## REFERENCES

Albright, B., and S. Perfater, 2018: 2018 flash flood and intense rainfall experiment: Findings and results. NOAA, 97 pp., https://www.wpc.ncep.noaa.gov/hmt/2018_FFaIR_final_report.pdf.

Erickson, M. J., J. S. Kastman, B. Albright, S. Perfater, J. A. Nelson, R. S. Schumacher, and G. R. Herman, 2019: Verification results from the 2017 HMT–WPC Flash Flood and Intense Rainfall Experiment. *J. Appl. Meteor. Climatol.*, **58**, 2591–2604, https://doi.org/10.1175/JAMC-D-19-0097.1.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, D. A. Miller, and T. O'Bannon, 1998: The WSR-88D rainfall algorithm. *Wea. Forecasting*, **13**, 377–395, https://doi.org/10.1175/1520-0434(1998)013<0377:TWRA>2.0.CO;2.

Gourley, J. J., and H. Vergara, 2020a: Comments on "Flash flood verification: Pondering precipitation proxies." *J. Hydrometeor.*, **22**, 739–747, https://doi.org/10.1175/JHM-D-20-0215.1.

——, and ——, 2020b: Data file with results supporting GV20 comments. GitHub, accessed 25 September 2020, https://github.com/HyDROSLab/FF_Products_Thresholds/blob/v1.0/outputs/general/hs18_regional_and_seasonal_All_weeks_contingency_stats_24H.ARI.csv#L615.

——, and ——, 2020c: Flash flood reports dataset. GitHub, accessed 25 September 2020, https://github.com/HyDROSLab/FF_Products_Thresholds/blob/v1.0/source_data/events_A4875356EB81004A1C6C9C960CF48888.csv.

Herman, G. R., and R. S. Schumacher, 2018a: Flash flood verification: Pondering precipitation proxies. *J. Hydrometeor.*, **19**, 1753–1776, https://doi.org/10.1175/JHM-D-18-0092.1.

——, and ——, 2018b: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Zhang, J., L. Tang, S. Cocks, P. Zhang, A. Ryzhkov, K. Howard, C. Langston, and B. Kaney, 2020: A dual-polarization radar synthetic QPE for operations. *J. Hydrometeor.*, **21**, 2507–2521, https://doi.org/10.1175/JHM-D-19-0194.1.