RC1: 'Comment on essd-2024-513', Anonymous Referee #1, 30 Jan 2025

The authors present a QPE data for a large number of watersheds throughout the Appalachians developed using inverse method for correcting radar/gage QPE based on observed streamflow and a hydrologic model. I still start by disclosing that I also reviewed 2024WR038446 (Water Resources Research), which is among the as-yet unpublished studies by the same authors. As with that manuscript, I found this study to be intriguing and somewhat challenging, and ultimately have the same general concern as I expressed (not as clearly as here) in my review for that prior work.

Thank you for reviewing this manuscript. The replies are in blue fonts.

Basically, my interpretation is that the authors have developed an approach for adjusting QPE based on back-trajectories of simulated streamflow such that the water budget closes at the event scale. This makes sense; I toyed with similar ideas myself years ago (though never did any real work on it). It does raise a potential concern, that I don't feel the authors did a great job addressing in either study. Specifically, if you use this approach, it seems to me that the appropriate way of judging success is whether the adjusted rainfall looks "better," i.e. closer, in terms of amount and spatiotemporal pattern, to some reference precipitation. Obviously, that's hard to do, since we don't have good reference precipitation in many locations. Instead, the authors show that the simulated hydrographs using the corrected precipitation have improved. This doesn't seem convincing—of course they have improved. You've adjusted the rainfall specifically to make sure that the hydrographs improve; then used the improved hydrographs as evidence that you have fixed the rainfall problems. But does that mean that the rainfall is more accurate? If the hydrologic model is good (and I trust that the authors' model is good) then the answer is "probably." If the model is not good, the answer is "probably not." The authors don't really answer the question.

Thank you for pointing out issues regarding this approach (i.e. IRC) and the need to validate post-IRC QPE. As mentioned in your comment, many places don't have good reference precipitation data, which is particularly true for mountainous regions. In response to this comment, we downloaded the Multi-Radar/Multi-Sensor (MRMS) data, and conducted a comparison between post-IRC QPE and MRMS QPE. Note MRMS data suffer from a relatively low radar quality index (RQI) for radar gaps in the mountains. For comparison, basin-averaged event total QPE is calculated for each dataset (i.e. StageIV$_D$, StageIV$_D$IRC, MRMS) and for each basin and each event. The results are shown in Figure S1.

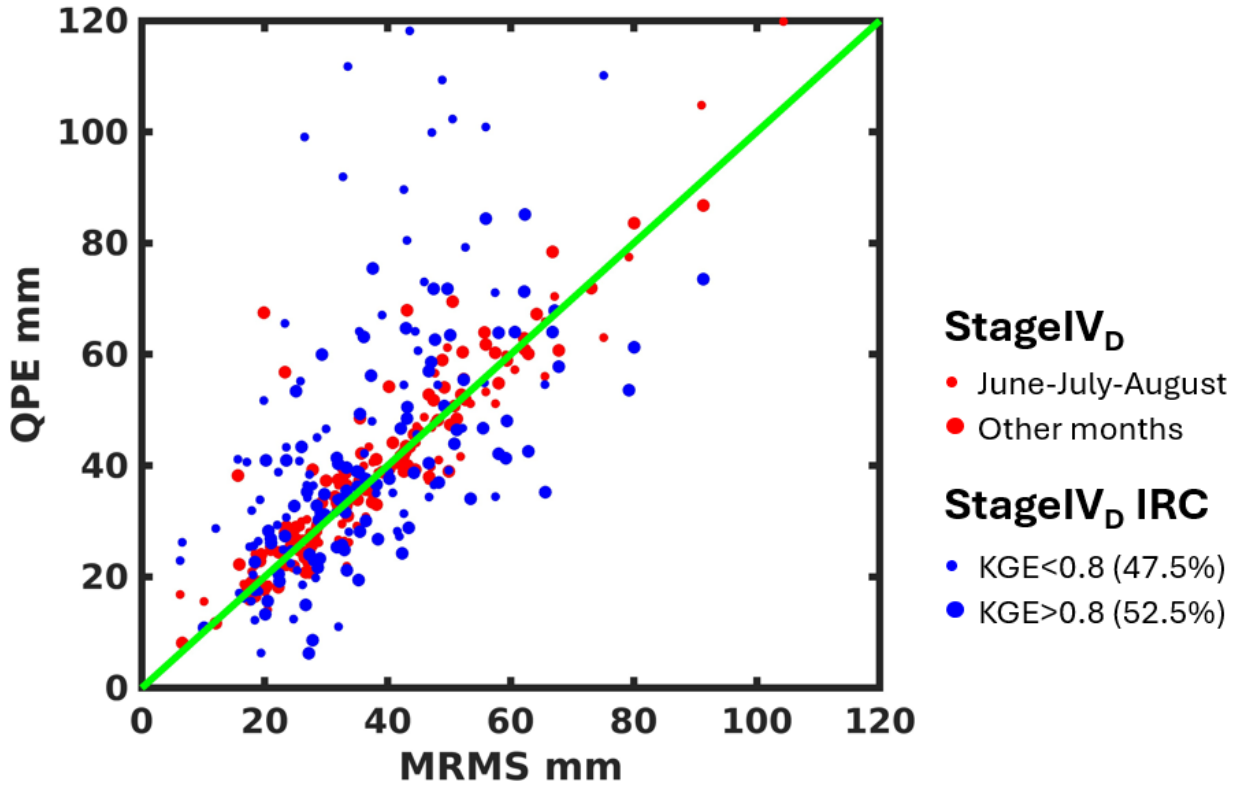

Figure S1 – Comparison of different QPE products for event total precipitation estimation (basin-averaged). Each point represents one event.

MRMS is often considered more advanced than StageIV in the scientific community because it incorporates much more data from various data sources, making it potentially more accurate in precipitation estimation than StageIV. Figure S1 shows that when post-IRC products yield high KGE values (>0.8, accounting for 52.5% of the events), StageIV$_D$ IRC has better agreement with MRMS. When post-IRC event rainfall generates relatively bad hydrological simulation (with KGE<0.8), it is expected that StageIV$_D$ IRC and MRMS have a relatively large discrepancy in this scatter plot. Figure S1 also points out that MRMS generally agrees well with StageIV for intense rain rate that occurs in June, July and August. This information is also reported in other studies.

A closer look at the outliers of StageIV$_D$ IRC in Figure S1 indicates a dependency of IRC performance on basin size. Relatively larger basins (mostly >200 km^2) usually produce lower KGE values (<0.8), indicating the IRC is not as effective in larger basins as in smaller headwater basins. This is also illustrated in the manuscript because the current version of IRC only uses shallow-layer travel time distributions up to 24 hours. For larger basins, it is expected that slower hydrological response from deeper layers becomes increasingly

important simply because of larger areas of relatively flat floodplains. Water travel time distributions from deeper layers should be considered for larger basins (>200 km^2), and long-lasting precipitation events (>24 hours).

Besides comparison against MRMS data, the authors also investigated available raingauge data. However, only one raingauge at Mill Gap in Virginia from COOP v2 dataset is available. This raingauge is unfortunately located in Basin 14, which is one of the two basins studied (Basin 13 and 14) that have complicated subterranean structures (i.e. Karst terrain), where DCHM performs poorly due to the lack of a Karst terrain module. Therefore, the post-IRC QPE is not reliable in these two basins, and this is discussed in the manuscript, thus no comparison against raingauges is executed in this study. However, raingauge comparison is done in the very original method paper (Liao and Barros, 2022) using a network of raingauges at high elevations in the Cataloochee Creek Basin.

Furthermore, MRMS data are downscaled to 250m resolution using nearest neighbor interpolation for hydrological simulation. The histograms of KGE distributions for various QPE data products are demonstrated in Figure S2.

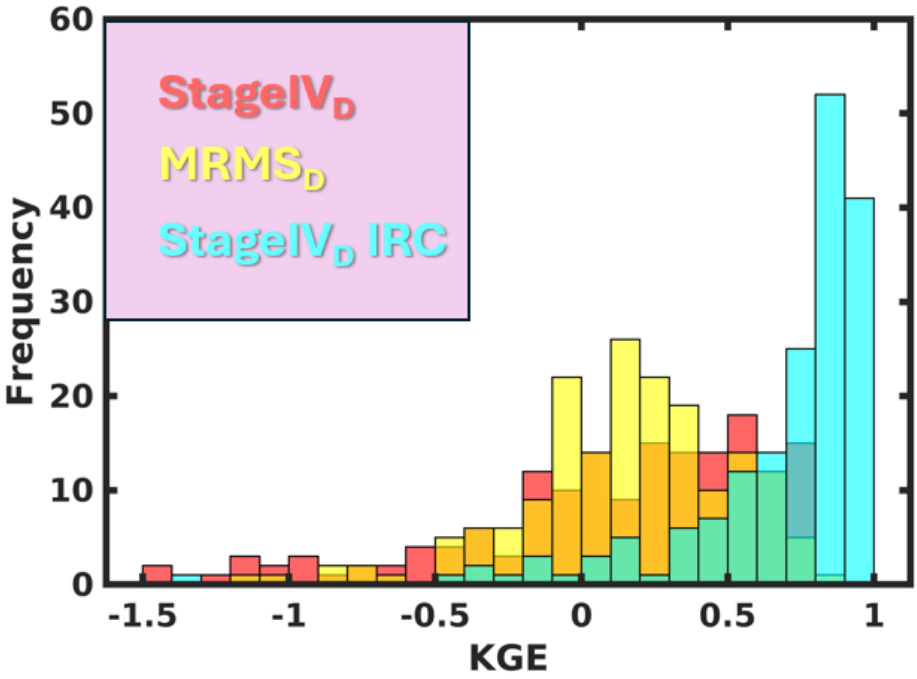

Figure S2 – Histograms of KGE values of studied events for various QPE products.

Figure S2 shows that the hydrological performances of StageIV$_D$ and MRMS$_D$ are not largely different, which is expected because they are using the same radars. The median KGE for both datasets are below 0.2. However, the median KGE for StageIV$_D$ IRC is above 0.8 for these extreme events.

I would appreciate the authors' response to that criticism. In addition, the authors need to state more clearly the differences between this study and others, particularly Liao and Barros (2024a), which I have reviewed, and Liao and Barros (2024b), which I have not. I guess this study is essentially a "scaling up" of the methods from those papers to more watersheds over a larger region? Fine, but please state it clearly.

Yes, this paper is a 'scaling up' of the methods from previous papers. Previous papers focus on a couple of headwater basins in the Southern Appalachians. This paper focuses on 30 headwater basins with over 200 extreme precipitation events in the entire Appalachians. Weather and climate regimes are dramatically different between Basin 01 (located in Gorgia) and Basin 30 (located in Maine). The headwater basin sizes range from 40km^2 to 450km^2. We also plan to submit another paper (Liao and Barros, 2025, in preparation) that includes 500 mountainous basins with over 10,000 extreme events including the Alps, the Andes, the Himalayas and the Brazilian Highlands, where the IRC method is demonstrated with significant success.

I will add that, similar to the earlier manuscript, the writing and overall presentation quality of figures should be improved. There are a large number of minor grammatical problems, especially with run-on sentences (three in the abstract alone) and missing articles (mainly "the") and some verb tense problems. These didn't make it impossible to understand the paper but do distract from the study's strengths. The figures should include legends, readable font sizes,

Thank you for the comment. We attached a revised manuscript along with the replies. Specifically, font size is increased in Figure 3. Legends are added for Figure 5-11. Excessive examples in Figure 8-10 are removed to accommodate font sizes.

The statement "uncertainty from the model and model parameters is assumed to be negligible" is not really reasonable, and is inconsistent with the following sentence that states that these have "secondary importance." "Secondary importance" is ok (at least for flood events, since as you note, forcing uncertainty will be large), "negligible" is not. Indeed, you aren't neglecting them in your method. Instead, you transfer the calibration effort from model parameters to rainfall. That is perhaps a reasonable thing to do for flood simulations with a high-quality distributed hydrologic model, as in this study. But it is an issue that should be more clearly acknowledged in your studies. And returning to my first concern, it is problematic if you are unable to quantify whether that produces improved rainfall—if your rainfall is practically incorporating model structure and parameter error,

there is the risk that it produces unrealistic rainfall outcomes. But in your study, we are left to wonder.

Thank you for improving the rigor of this article. The authors acknowledge that parameter uncertainty and model uncertainty can have a large impact on model performance, especially for simulating the hydrological response of moderate and less significant precipitation events. Therefore, the study samples of this study only include rainfall that produces flood peaks greater than 95 percentiles of streamflow measurements, where precipitation uncertainty should dominate over other sources of uncertainties.

I realize that I might be misunderstanding the issue entirely. If so, please clarify.

Thank you again for reviewing this manuscript. The replies above should address the issues raised in the reviews.