Reply to Referee #2:

I have read over the paper. I think it is very well written and if I was reviewing this I would only ask for moderate or minor revisions.

The authors developed a storm tracking algorithm (using a combination of existing algorithms in a novel way) and use it to create a dataset of large storm events which they perform frequency analysis on. Despite the novel storm tracking algorithm, performing frequency analysis on a storm dataset (as opposed to gauge records) is not in itself novel. This paper's key contribution comes from the way the authors uses copula based multivariate analysis on atmospheric variables from ERA5 to develop a way to stochastically generate annual maxima series representative of the observed storm catalogues.

Major comments

My first major concern with the paper is that the authors do not make enough attempts to validate their method or at least compare it to external data sources.

The only comparison to other methods they make is with a GEV fitted to storm annual maxima in Figure 6-7. I believe there is also opportunity to compare DAD curves in Figure 8 to an external data sources. The authors mention the relationship between DAD curves and ARFs, so any ARF information available for the Mississippi basin could be used to formulate a comparison here.

I appreciate that these comparisons may be difficult to facilitate because of the authors have taken a storm-centred approach while the majority of other datasets are based on gauge-centred data. Still for their approach to be applied outside of research we need to understand how it compares to existing approaches.

We thank the reviewer for generally positive reception, and we hope our response and revisions can address any remaining concerns.

We agree that comparing and validating our approach to other data sources are important, particularly if the goal is for future development of the method. We would highlight several places in the original manuscript where we did compare our approach with "external" observation-based datasets. The spatiotemporal properties of ERA5-simulated storms were validated against the storm tracking results based on the IMERG satellite-based precipitation estimates (Section 4.1). The bias in ERA5 extreme precipitation was also evaluated by comparing the 99th percentile daily precipitation against the gauge-interpolated nClimGrid dataset in the Mississippi Basin (Section 5.2 and Figure B1).

Comparing DAD curves is indeed rather difficult because most previous studies were based on gauge-centered data and had limited data length, smaller/different areas, and other differences. For example, the ARF data from Technical Papers No. 29 (US Weather Bureau, 1958) and No. 49 (Miller, 1964) in the Mississippi Basin were limited to watersheds less than ~1,000 km^2, much smaller than the minimum area of 5,000 km^2 in our study.

A recent study by Kao et al. (2020) provides ARFs for 10-year precipitation with durations of 2-72 hours and areas of 10-100,000 km^2 in the Ohio River Basin, using a watershed-based approach and gauge dataset. We calculated new DAD curves by converting the hourly precipitation depth at 5,000 km^2 from the vine copula model to precipitation depths at larger areas based on the ARFs. We then compared these new DAD curves with the original DAD curves that were purely estimated from vine copulas (see Figure C1 below). The two DAD curves agree well for durations between 6 to 72 hours, while for 2-hour storms the vine copula estimates are more conservative, i.e., the precipitation depth reduces much slower than the ARF estimates when area increases. Such discrepancies may be attributable to the ARF estimation of Kao et al. (2020) being a “fixed-area” approach, i.e., the point precipitation depth is related to areal depth in a watershed. Nevertheless, the number of large watersheds in the Mississippi Basin (e.g., watersheds greater than 50,000 km^2) is limited, which may limit the approach’s ability to identify truly areal maxima, especially for short-duration large-area storms. This suggests that our storm searching algorithm may provide more conservative DAD relationships for storms with short durations. An alternative explanation, however, could be that these differences highlight the limits of ERA5 in depicting extreme convective rainfall at small space-time scales. We propose to include Appendix C “Validation of vine copula DAD relationships” (see draft below) and a short discussion in Section 5.3 to describe the above comparison:

Appendix C: Validation of vine copula DAD relationships

The vine copula DAD curve was compared against the ARFs in Kao et al. (2020). The ARFs were estimated for 10-year precipitation with durations of 2-72 hours and areas of 10-100,000 km^2 in the Ohio River Basin, using a watershed-based approach and gauge-based dataset (DSI-3240, National Climatic Data Center, 2003). We calculated new DAD curves by converting the vine copula hourly precipitation depth at 5,000 km^2 to the depths at larger areas based on the ARFs. The new DAD curves were then compared with the original DAD curves estimated from vine copulas, as shown in Figure C1. More discussion of this figure can be found in Section 5.3.

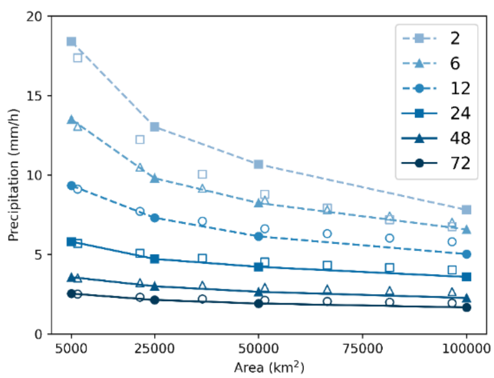

Figure C1 Comparison of DAD curves estimated from vine copulas (solid markers) and ARFs from Kao et al., 2020 (empty makers) for Ohio River Basin with 10-year ARI and 2-72 hours durations.

The following short discussion will be added to Section 5.3 (line 510):

"…The ability to derive storm-centered DAD relationships using our method can in principle obviate the need for ARFs entirely, something that has been advocated for previously (Wright et al., 2014). To support this point, we compared vine copula DAD curves with those estimated by the ARFs from Kao et al. (2020) in the Ohio River Basin at 10-year ARI (see Appendix C and Figure C1). The vine copula DAD estimates agree well with those ARF estimates for storm duration between 6 and 72 hours, while for 2-hour storms the vine copula estimates are more conservative, i.e., the precipitation depth reduces much slower with increasing area. Such discrepancies can be attributable to the ARF estimation of Kao et al. (2020) being a "fixed-area" approach, i.e., the precipitation depth is compared to areal depth in a watershed. Nevertheless, the number of large watersheds in the Mississippi Basin (e.g., watersheds greater than 50,000 km^2) is limited, which may limit the approach's ability to identify truly areal maxima, especially for short-duration large-area storms. This suggests that our "storm-centered" approach may provide more conservative DAD relationships for storms with short durations. An alternative explanation, however, could be that these differences highlight the limits of ERA5 in depicting extreme convective rainfall at small space-time scales. Another contention within the ARF literature is whether or not such ratios are independent of recurrence intervals (Greener and Roesch, 1997; Osborn et al., 1980; Pavlovic et al., 2016). Relevant to this debate, we found that DAD appears to be independent of recurrence intervals for 5,000-100,000 km2 scales in the Mississippi Basin."

My second major concern is about the use of empirical CDFs for all atmospheric variables except the divergence term for which a GEV is fitted. While divergence shows the highest correlation to precipitation I find this insufficient justification for why only this variable is modelled using a GEV. I also note that other terms such as the residual also have non-negligible contributions to rainfall. I would be interested to know if there is any change in results if similar extreme value distributions are used for other atmospheric variables.

It is feasible to use parametric distributions for all atmospheric variables. Based on our testing, the impacts on results are minor if the parametric distribution fits well to the corresponding atmospheric variable. The choice of parametric distribution depends on the variable's statistical properties. For example, the distribution of the residual term can be fitted by a t-distribution, while the time derivative term and evapotranspiration term can be fitted with a beta distribution. Note that they need not be extreme value distributions to achieve a good fit.

In the study, we used empirical cumulative distribution functions (CDFs) for the remaining variables to reduce additional parameters and errors introduced by fitting parametric distributions; this is common practice in vine copula modeling. However, empirical CDFs can constrain the simulated variables to their maximum in the original sample data, leading to unrealistic upper-bounded tail behavior. Therefore, we used GEV distribution to fit the dominant component (i.e., the convergence term) to allow our model to generate extreme precipitation that exceeds the original maxima. Another advantage of using parametric distributions is that nonstationarities (e.g., changing location and scale) in atmospheric variables can be modeled using distribution parameters that vary with time or other climatic covariates. Indeed we did this while preparing the manuscript, but decided that the "story" became too complicated to present due to the difficulty of showing and evaluating nonstationary return levels.

We propose to add the above discussion at the end of Section 4.4.2 (line 453):

“The histograms along the diagonal show the marginal distributions of each water balance component used in the vine copula model; the divergence term’s histogram is smooth due to the use of a GEV marginal distribution, while the other three components used empirical CDFs. We used empirical CDFs for the remaining variables to reduce additional parameters and errors introduced by fitting parametric distributions; this is common practice in vine copula modeling. However, empirical CDFs can constrain the simulated variables to their maximum in the original sample data, leading to unrealistic upper-bounded tail behavior. Therefore, we used GEV distribution to fit the dominant component (i.e., the convergence term) to allow the model to generate extreme precipitation that exceeds the original maximum. Note that it is feasible to fitting parametric distributions to all atmospheric water balance components. The influence on the results is rather minor if the parametric distribution fits well to each component. For example, the distribution of the residual term can be fitted by a t-distribution, while the time-derivative and evapotranspiration terms can be fitted with beta distributions. Another advantage of using parametric distribution is that nonstationarities (e.g., changing location and scale) in each atmospheric water balance component can be modeled with distribution parameters that vary with time or other climate indices (see Section 5.5).“

Minor comments

I’d prefer the use of spelling gauge to gage, I think it’s most common in modern literature.

Thanks for this suggestion. The spelling has been changed in the manuscript.

Section 3.1: I believe the authors could draw more attention to their storm search method being a novel combination of existing approaches

We propose to rearrange the first paragraph in Section 3.1 (line 130) to highlight the novelty of our approach.

Original text:

“…we developed the Storm Tracking and Regional Characterization method (STARCH, publicly available at https://github.com/lorenliu13/starch). The method can track storms based on successive two-dimensional precipitation fields and create catalogs of extreme storm events with specific areas and durations within a chosen region. The storm identification and tracking portions of STARCH combine two prior storm tracking algorithms: 1) double-threshold identification from the Thunderstorm Identification, Tracking, Analysis, and Nowcasting (TITAN) algorithm (Dixon & Wiener, 1993) and 2) “almost-connected component labeling” from the Storm Tracking and Evaluation Protocol (STEP; Chang et al., 2016)…”

Proposed text:

“…we developed the Storm Tracking and Regional Characterization method (STARCH, publicly available at https://github.com/lorenliu13/starch). The method is a novel combination of two prior storm tracking algorithms: 1) double-threshold identification from the Thunderstorm Identification, Tracking, Analysis, and Nowcasting (TITAN) algorithm (Dixon &

Wiener, 1993) and 2) "almost-connected component labeling" from the Storm Tracking and Evaluation Protocol (STEP; Chang et al., 2016). An area-duration selection algorithm is also developed to search storms with user-defined duration and area. STARCH can not only track storms based on successive two-dimensional precipitation fields, but also create catalogs of extreme storm events with specific areas and durations within a chosen region…."

Line 170-175: I think the explanation of the 'binary search' is not clear and could be improved

We propose a more detailed description of the binary search algorithm in this paragraph (line 173):

Original text:

"…To do this, a binary search is implemented on the total precipitation map of the storm to find a precipitation contour whose area value is closest to but less than the desired area *A*. Thereafter, the area selection algorithm recursively expands…"

Proposed text:

"…To do this, a binary search was implemented to find a precipitation threshold whose corresponding contour area on the total precipitation map is closest but less than the desired area *A*. We began with a threshold at the midrange of the precipitation interval, i.e., (maximum+minimum)/2, and computed the contour areas, i.e., areas of precipitation regions above the threshold. The largest contour area was compared with the desired area *A*. If the contour area is less than *A*, we narrowed the precipitation interval to the lower half, i.e., from the minimum to the midrange. Otherwise, we narrowed the interval to the upper half. We then repeatedly calculated the midrange of the new interval as the next threshold and compared the contour area with the desired area *A*. All the thresholds and associated contour areas were recorded through iterations. The binary search stops if the difference between the contour area and desired area is less than one pixel, or the selected contour area does not change in consecutive 3 iterations. From the search record, we found a threshold with a contour area that is closest but less than the desired area *A*. Thereafter, the area selection algorithm recursively expands…"

Line 255: The GEV scale parameter must be greater than zero ($\sigma>0$)

The range of the scale parameter has been corrected from $(-\infty, +\infty)$ to $(0, +\infty)$.

Line 354: Should reference Figures A1-2?

Yes. The reference should be Figures A1-2 and has been corrected.

Additional - uncertainty in the ERA data is not accounted for in the approach here. Alternate reanalyses do not agree with each other, even for atmospheric moisture - see Moalafhi, D. B., Evans, J. P. & Sharma, A. Influence of reanalysis datasets on dynamically downscaling the recent past. Climate Dynamics 49, 1239-1255 (2017). Some discussion on the impact of this uncertainty and how it could be included in the GEV modeling may be helpful.

We agree that uncertainties exist in atmospheric water balance components and can influence the precipitation estimates. Uniquely among reanalyses (to our knowledge anyway), ERA5 includes coarser-resolution (3-hour, 0.5° grid scales) ensembles that can be used to examine some forms of uncertainty. To assess these uncertainties, we computed the water balance components in the annual maximum storm catalogs based on 10 ERA5 ensembles. These ensembles estimate the uncertainties of observations in DA and model parameterizations (Hersbach et al., 2020). All the atmospheric water balance components showed certain variations, especially for precipitation and water vapor flux convergence. Nevertheless, the coarse resolution of the ensemble can smooth out high precipitation regions and periods and result in different storm tracking and search results, making it difficult to be directly used to quantify the uncertainty in our precipitation estimates. However, the variations in ERA5 ensembles can still qualitatively reflect the uncertainty in atmospheric variables across different subbasins and different storm spatial-temporal scales. Other reanalysis or numerical models, such as MERRA-2, can also be alternative sources to evaluate the uncertainty in ERA5 reanalysis. We need to use the common period of these data (e.g., 1980-2020) and transform them to the same spatial-temporal resolution to perform storm tracking and vine copula fitting. An ensemble distribution can be generated from extreme precipitation estimates based on each dataset. It is expected that the variability of the precipitation estimates is likely to increase by incorporating multiple reanalysis/numerical model sources. We propose to add the above discussion at the end of Section 5.2 (line 492) and change the section name to “Uncertainty and Bias in ERA5 Reanalysis.”

References

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Kao, S.-C., DeNeale, S. T., Yegorova, E., Kanney, J., & Carr, M. L. (2020). Variability of precipitation areal reduction factors in the conterminous United States. *Journal of Hydrology X*, *9*, 100064. https://doi.org/10.1016/j.hydroa.2020.100064

Miller, J. F. (1964). *Two- to Ten-Day Precipitation for Return Periods of 2 to 100 Years in the Contiguous United States* (Technical Paper No. 49; p. 32). U.S. Weather Bureau.

National Climatic Data Center. (2003). *Data documentation for dataset 3240 (DSI-3240) hourly precipitation*.

US Weather Bureau. (1958). Rainfall intensity-frequency regime. *Technical Paper*, *29*.