

Review of “Process-aware mixture modelling of precipitation-event distributions using event-type clustering”

Summary and general recommendation

This paper presents a method to model precipitation using a finite mixture of Gamma distributions, with one distribution for each identified cluster, where the clusters can be interpreted as precipitation types. Two clustering approaches are compared: k -means and model-based clustering based on a finite mixture of Gamma regression models. The proposed approach is applied to two weather stations in Austria.

I find the manuscript to be well written. I particularly appreciate the extensive discussion of the results of the data analysis in Section 3, and the overall pedagogical style of the paper. However, the paper claims to have a methodological interest, but I find the description of the methodology to be somewhat rushed. This is especially the case in Section 2.2.2, where the clusterwise regression is introduced without being explained in sufficient detail. In addition, I think that the paper would benefit from a few additions and from a revision of some less clear parts. My detailed comments can be found below.

Major comments

1. A literature review on existing approaches based on finite mixture models for the analysis of precipitation is missing in the Introduction. In addition, when discussing distributions that are preferable for modeling extreme rainfall, it is worth mentioning the extended generalized Pareto distribution (EGPD) of Naveau et al. (2016), which has also been used to model precipitation. There is also existing literature on finite mixtures for precipitation extremes (for instance Kjeldsen et al., 2018), which could also be mentioned.
2. The paragraph “Clustering Approach II: clusterwise regression, outcome-guided learning approach” contains one of the main methodological contributions of the paper, and it would benefit from a more explicit statistical formulation. The way clusters are identified would benefit from a clearer explanation, and the EM algorithm used for estimation could also be described in more detail.

It would also be helpful to establish a direct connection between the mixture formulation in Eq. 1 and the Gamma regression model, which is currently not written explicitly. For example, the authors could clarify how the parameters of the Gamma components are related to the regression coefficients. Moreover, Eq. 1 does not represent the model, but the likelihood contribution for event i . Presenting the mixture model connected to the regression formulation would make the statistical structure easier to understand. It would also be useful to state explicitly that the proposed framework is a dependent mixture model, as the component distributions depend on covariates through the regression structure.

3. Given the importance of accurately representing extreme rainfall events, the current diagnostic figures could be adapted to provide a clearer assessment of the right tail of the distributions. In Figure 5, differences in the upper tail are difficult to recognize from the CDFs. It would be more informative to complement or replace these plots with plots of the quantile functions. Analogously, for Figure C1, Q-Q plots may be more suitable than P-P plots for capturing differences in the tails.
4. To complete the methodological discussion, it would be useful to address the difference between the proposed approach, which divides the events into clusters and fits a distribution separately to the observations within each group, and a direct finite-mixture approach, in which the component distributions and cluster memberships are estimated jointly, for example through an EM algorithm.
5. For the clusterwise regression approach, the clustering is obtained from a finite mixture of Gamma regression models estimated by EM. It is therefore not entirely clear why an additional Gamma distribution is subsequently fitted to the observations assigned to each cluster and these fitted distributions are then recombined into a mixture. The authors could clarify the motivation for this additional step and explain how the resulting marginal mixture relates to the original mixture of Gamma regression components.

Minor comments

1. To complement Table A.1, it would be useful to add, for each variable, a single plot showing the overlaid histograms for the two stations.
2. Line 131: ; should be ,
3. Line 132: “the improvement of aggregated theoretical precipitation distribution fit”. I do not understand what the authors mean by this and how the improvement is quantified at this stage.
4. In the paragraph “Clustering Approach I: Partitioning clustering”, I do not understand why reason 1 is an argument to prefer k -means to PAM. It would be helpful if the authors clarified this point.
5. Line 167: EM acronym used before its definition (in the following sentence).
6. Both K and k are used to indicate the total number of clusters, and k is also used to indicate a single cluster in Eq. 1 and 3.
7. Paragraph “Chosen numbers of clusters k ”: it would be good to mention other criteria for choosing the number of clusters, e.g., Akaike information criterion (AIC), Bayesian Information Criterion (BIC).
8. Section 2.2.3: it would be useful to clarify that in the proposed approach this is done cluster by cluster.
9. Section 3 and onward: the variables are in teletype font only from this Section.
10. Section 3.2: please explain what the percentages in this section mean.

11. Figure 5: an analogous figure showing the histograms of the data with the overlaid densities (Gamma overall and Gamma cluster-aggregated) could be shown in the Appendix.
12. Table 1 should be after it is mentioned in line 255.
13. It would be helpful to provide clearer definitions of the error metrics used, in particular the RMSE of the CDFs, the MAE of the CDFs, and the MAE restricted to the empirical quantile band. It is not immediately clear why the MAE reported in Table 1 is referred to as the “quantile-band MAE” in line 264. A more explicit description of how each metric is calculated and how the terminology used in the text corresponds to that in the table would improve readability.
14. The Discussion could be streamlined, as some points are repeated, including references to literature that has already been covered in the Introduction. It may also be worth considering merging the Discussion and Conclusions sections, since the latter repeats several concepts and findings presented earlier.
15. Some capitalization appears to be inconsistent or missing in the References.

References

- T. R. Kjeldsen, H. Ahn, I. Prosdocimi, and J.-H. Heo. Mixture gumbel models for extreme series including infrequent phenomena. *Hydrological Sciences Journal*, 63(13-14):1927–1940, 2018.
- P. Naveau, R. Huser, P. Ribereau, and A. Hannart. Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769, 2016.