Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2021-366-RC1, 2021

Comment on hess-2021-366

Francesco Marra (Referee)

Referee comment on "Enhancing the usability of weather radar data for the statistical analysis of extreme precipitation events" by Andreas Hänsler and Markus Weiler, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-366-RC1, 2021

This paper presents a new dataset of extreme precipitation return levels derived from radar estimates (RAD-BC). The methodology builds over the sampling approach presented by Goudenhoofdt et al. (2017) and improves the sampling strategy by using spatially-varying sampling probabilities which depend on both horizontal distance and terrain elevation. The topic is of interest to the readers of HESS, and the study is timely as it tackles a state-of-the-art problem.

While some results are encouraging as the spatial artifacts generated in the Goudenhoofdt methods are removed and the orographic influence on precipitation statistics is better represented in the product, the final product shows important systematic bias despite the application of a bias correction procedure.

Overall, I found some aspects of the study to be insufficiently robust, as highlighted in the "major comments" below. While I sincerely appreciate the efforts of the authors, these aspects prevent me from recommending acceptance of the paper in its present version. I'd invite the authors to consider my comments, and I'll be happy to discuss them further in the open discussion in case I misunderstood some parts.

Please do not consider the references below as recommended for inclusion in the paper, they are meant to be examples only.

Kind regards,
Francesco Marra

Major comments:

- The reference dataset is sometimes used to support the goodness of the new dataset and sometimes regarded as less accurate (e.g. in the patterns of sub-daily precip – see lines 16-18 in the abstract). Although the reasons behind this can be somehow understood, this is a problematic issue. On what bases is the dataset trusted as a reference (perhaps some durations are and some are not, some return periods are and some are not)?
  I think a proper evaluation should rely on a trusted dataset. For example, rain gauges could provide a quantitatively trusted reference to gather information on the quantitative accuracy of the method on some selected locations. This might allow us to understand what aspects the radar product is or isn't able to reproduce (orographic influence at different durations, different return levels, etc). Alternatively, the trusted parts of the available dataset should be defined a priori and used for the validation, while the parts which are not trusted should be only used for comparison and discussion.
- While I understand the need to avoid winter periods due to the known issues of weather radar monitoring with solid precipitation, it is not clear to me how it is possible to compare return levels derived from summer only (Apr-Oct as in this paper) with return levels derived from stations (the reference products) for durations up to 24 hours. The authors mention this at lines 82-84 (*"Since we are mainly interested in short to medium range storm events that are mainly of convective nature, we only use data for the (summer) months from April to October, representing the main season for these kind of storm events"*), but then durations up to 24 hours (e.g. see lines 244-255) are examined and discussed. This mismatch, which is not discussed by the authors, could also contribute to the overall bias found by the authors. I fear this might represent an important drawback of the presented product and of the presented comparison.
- I like the idea of sampling the surrounding pixels using probabilities, and I like the idea of basing the properties of the sampling pdf based on the typical size of convective rain cells in the region, but I am missing why the same mask is used for all durations. Since precipitation accumulations over longer durations are characterized by larger autocorrelation, my guess would be that 4 km might be good for short durations (even 1 hour could be border line according with what is said above), but too short for longer durations.
- Lines 107-111: this is presented in a confusing way. There is no guarantee that 100 years of data will provide perfect (or good for what matters) estimates of the 100-year return levels. Monte Carlo simulations run under realistic precipitation statistics show that empirical estimates will be subject to ~90% uncertainty (computed as the 90% confidence interval), while a simple GEV fit (method of the L-moments) will be subject to ~50% uncertainty. The advantage of using ~100 years of data instead of ~20 is clear, but should be presented in a better way.
- The results show an important systematic bias (as it can be inferred from fig. 4). This bias concerns most of the study area and cannot be seen as related to stochastic uncertainty, therefore the uncertainty quantification at section 3.4 cannot be accounted for explaining it. This is an important issue and I wonder what is the added value of such a quantitative information for the final user.
  To my view, this issue is related to a sub-optimal choice of the bias correction method (see details below), and addressing it should therefore be part of this study. The bias correction described in section 2.5 seems to me insufficient. Basically, this correction includes an additive adjustment to the data (changes the location parameter of the

GPD). Since radar errors are far from being only additive, the resulting product is necessarily biased. Eventually, the results presented in the paper confirm this: the underestimation increases with return period, meaning that the other parameters are wrongly represented by the product and therefore also need to be adjusted. While the authors mention these efforts as future directions, I think that the here presented results are not sufficient to justify this publication and that these additional efforts have to be invested here.

Moderate comments:

- It seems to me that larger ensembles could produce more accurate estimates (for example they could reduce the stochastic noise still present in the data and which required the smoothing of the maps). Why is a factor of 5 chosen? Are there only statistical-independence limitations or is it also a matter of computational time?
- Lines 40-41: this is an over-simplification. The short record length is indeed among the important drawbacks of weather radar archives, but other issues were highlighted in literature. The most important one is definitely estimation inaccuracy: large systematic over- and under- estimations were found due to measurement errors (e.g. Eldardiry et al., 2015; Haberlandt and Berndt, 2016, among others), but in a recent review on the topic we also highlighted the inadequacy of the adopted statistical methods (Marra et al., 2019). As these aspects are somehow addressed by the methodology in this paper, I think the introduction should better present them.
- Section 2.2: information on the extreme value methodology used in the reference products has to be provided. Something is said later in the text, but the information should be presented in an organized manner here. Also, the implications of these choices should be discussed. For example, distributions with different tail heaviness will unavoidably show different biases at different return levels. If indeed different methodologies are used, the impact of these aspects on the comparison and on the results have to be discussed.
- Lines 116-121: I am missing the relation between the typical size of the convective cells and sampling radius and normal distribution parameters.
- Line 132: similar to the previous comment, why is 4 km chosen here?
- Line 220: It would be nice to see the results also for 1-year or 2-yr return levels. Since the adjustment is basically done on the 1-yr event, they should well isolate the quality of the product in relation to the bootstrap sampling method.
- Line 232: why is the map smoothed? It seems this is to remove some noise. However, the noise we would see in these maps is a direct representation of the stochastic uncertainties affecting the overall methodology. I think the maps would be more informative without the smoothing.
- Line 310-313: I might agree on the fact that higher-order moments are more difficult to estimate and to rely on, especially from "indirect" datasets such as the ones used here as a reference. I however, think that this problem can be somehow addressed by using a more trusted reference and by using corresponding statistical methods.
- Although not a native speaker myself, I felt that the language level could be improved, in part due to missing use or misuse of technical terms.

Minor comments:

- Lines 16-18: this sentence is not completely clear. I could understand it only after reading the paper. Since this is the abstract, I suggest rewording it.
- Line 32: some change-permitting GEV methods allow for changes also of the scale parameter (e.g. see Prosdocimi and Kjeldsen, 2021)
- Lines 41-42: I personally disagree on this point. While this is very true for traditional methods based on extreme value analysis, there are some novel statistical methods which show promising results in this sense. They are now published since few years (the first papers are by Marani and Ignaccolo, 2015; Zorzetto et al., 2016), and many came after providing evidence (with applications to rain gauge data as well as satellite data) of the fact that 20 years might be sufficient for at-site estimates of even 100-year return levels. I believe it is time to recognize this by specifying that this limit concerns the traditional methods based extreme value analyses.
- Line 112: it is not clear to me what the authors mean with *"with underlying sampling probabilities"*
- Line 115: what does *"not necessarily present"* mean exactly? Is it a way to say "independent"?
- Line 127: I suggest to include this information on the elevation range earlier in the text. Perhaps a short section describing the study area could help also in the following discussion.
- Line 222: with *"lower time steps"*, do you mean "shorter durations"?

References

Eldardiry, H., Habib, E., Zhang, Y., 2015. On the use of radar-based quantitative precipitation estimates for precipitation frequency analysis. J. Hydrol. 531, 441–453. https://doi.org/10.1016/j.jhydrol.2015.05.016

Goudenhoofdt, E., Delobbe, L., and Willems, P.: Regional frequency analysis of extreme rainfall in Belgium based on radar estimates, Hydrology and Earth System Sciences, 21, 5385-5399, 2017

Haberlandt, U., Berndt, C., 2016. The value of weather radar data for the estimation of design storms an analysis for the hannover region. In: Schumann, A. (Ed.), The Spatial Dimensions of Water Management – Redistribution of Benefits and Risks Data. IAHS, pp. 81–85

Marani, M., Ignaccolo, M., 2015. A metastatistical approach to rainfall extremes. Adv. Water Resour. 79, 121–126. https://doi.org/10.1016/j.advwatres.2015.03.001

Marra F, EI Nikolopoulos, EN Anagnostou, A Bárdossy E Morin, 2019. Precipitation frequency analysis from remotely sensed datasets: A focused review., J. Hydrol. 574, 699-705, https://doi.org/10.1016/j.jhydrol.2019.04.081

Prosdocimi, I., Kjeldsen, T. Parametrisation of change-permitting extreme value models and its impact on the description of change. Stoch Environ Res Risk Assess 35, 307–324 (2021). https://doi.org/10.1007/s00477-020-01940-8

Zorzetto, E., Botter, G., Marani, M., 2016. On the emergence of rainfall extremes from ordinary events. Geophys. Res. Lett. 43, 8076–8082. https://doi.org/10.1002/2016GL069445