

Review of Rosin et al. (2026) - Radar-Derived Intensity–Duration–Area–Frequency Relations for Assessing Hydrological Hazards in Complex Terrain

Review 2 response

General assessment

This paper utilizes high spatiotemporal resolution radar quantitative precipitation estimation data from Switzerland during the summers of 2016–2024, combined with the areal SMEV framework, to establish the IDAF relationship for summer extreme precipitation with a return period of 2–100 years. Furthermore, it demonstrates its application value in hydrological hazard assessment in complex topographical areas through three recent high-impact flood events. The paper addresses a clear scientific question, provides valuable data resources, employs a highly targeted methodology, and the results have practical significance for multi-scale rainfall diagnosis of hazards such as mountain floods, urban flooding, and debris flows. I believe this paper has high publication potential overall, especially in applying areal SMEV to the complex Alpine terrain of Switzerland and linking multi-scale extreme precipitation statistics with actual hydrological hazard responses.

However, several key issues in the current manuscript still require further attention, particularly the reliability of extrapolating 9 years of radar data to a 50–100 year return period, the propagation of radar QPE systematic errors to the return level and IDAF, the impact of SMEV shape parameter spatial smoothing on uncertainty and terrain gradient, and quantitative comparison with existing Swiss IDF/IDAF or CombiPrecip related results. These issues do not diminish the overall value of this paper, but they do affect the interpretability and reliability of the results. Therefore, I recommend a major revision.

Response: We thank the reviewer for this constructive assessment and for recognising the value of applying the SMEV-based IDAF methodology of Rosin et al. (2024) to high-resolution radar precipitation data in complex Alpine terrain. In the revised manuscript we have limited the maximum return period to 50 years to reduce the degree of extrapolation, and we now state explicitly that the bootstrap intervals represent conditional sampling uncertainty rather than total uncertainty. We separate the sources of radar–gauge disagreement and clarify the role and limitations of the uniform bias adjustment. We added sensitivity analyses of the shape-smoothing window (Figure S2) and of the bootstrap ensemble size (Figures S5–S6), increased the ensemble from 100 to 800 replicates, and evaluated the dry-separation criterion used to define independent storms (Figure S1). We also define the scope of the 1 km² radar–gauge comparison and its limits for the areal estimates, clarify the interpretation of ARFs exceeding 1, and report the number of ordinary and upper-tail events used for fitting (Figure S7). The comparison with existing Swiss IDF/IDAF and CombiPrecip-based studies is addressed in the response to comment #2. Detailed responses follow.

Major comments

1) The paper uses summer radar data from 2016–2024 to estimate the multi-scale return level up to a 100-year return period. The authors reason that SMEV can utilize the tail distribution of ordinary events rather than relying solely on annual maximum values, thus making it more suitable for short radar sequences; the paper also cites existing research indicating that approximately 10 years of records can be used for stable estimation within the SMEV framework. This argument is reasonable and forms the core basis for the methodological choice in this paper.

However, from the perspective of engineering hydrology and risk management, extrapolating the 50-year and 100-year recurrence levels based on 9 years of data still exhibits significant model dependence. The uncertainties here include not only bootstrap sampling uncertainty, but also whether the tail of ordinary events follows a Weibull distribution, whether the assumption of event independence is sufficient, whether different weather patterns still exist within the summer, whether radar QPE errors affect the tail, and whether shape parameter spatial smoothing alters extreme tail behavior. It is recommended that the authors more clearly distinguish in the results and conclusions that: SMEV parameter estimation is statistically more stable than traditional GEV/POT; the 100-year recurrence level is still a significant extrapolation; and the reliability of this extrapolation depends on the SMEV model assumptions and the radar QPE error structure.

Response: Thank you. We agree that return levels substantially longer than the nine-year radar record should be interpreted with caution. To reduce the degree of extrapolation, we removed the 100-year return-level estimates throughout the manuscript and now limit the analysis to return periods up to 50 years. Following the reviewer's suggestion, the revised manuscript distinguishes more clearly between the statistical advantages of SMEV, the extrapolation involved in the 50-year estimates, and the assumptions on which these estimates depend.

First, regarding the statistical advantage of SMEV compared with GEV and POT approaches, the Introduction now reads:

Extreme precipitation has traditionally been analysed using asymptotic approaches based on extreme value theory. These include block maxima methods, where annual or seasonal maxima of asymptotically large blocks can only converge to the Generalised Extreme Value (GEV) distribution, and threshold exceedance methods such as peaks over threshold (POT), where Poisson exceedances of an asymptotically high threshold can only converge to the Generalised Pareto distribution (Coles, 2001). Both approaches utilise a small subset of available data, although POT generally retains more information than block-maxima methods, which use only one or a few observations per block. Both therefore require long observational records, typically obtainable only from gauge networks. In addition, these approaches heavily rely on asymptotic assumptions, which are likely not met in the case of precipitation (Marra et al., 2023, 2026).

More recent non-asymptotic approaches include the Metastatistical Extreme Value (MEV) framework (Marani and Ignaccolo, 2015) and its simplified variant, the SMEV (Marra et al., 2019), which estimate extreme value distributions from the bulk distribution of 'ordinary' events rather than from extremes only. They use substantially more information from short records and do not rely on the asymptotic convergence assumptions underlying GEV and POT models. This reduces parameter uncertainty and enables reliable estimation of rare return levels from short or error-prone records, making them particularly well-suited for short

radar archives (Marra et al., 2018; Zorzetto et al., 2016). The SMEV framework has been successfully applied to both point-based and spatial precipitation datasets worldwide, consistently showing improved performance over GEV-based models, especially when applied to short records (Vidrio-Sahagún and He, 2022). In particular, records of approximately 10 years have been shown to be sufficient for deriving stable return level estimates using the SMEV framework (Dallan et al., 2024; Marra et al., 2022; Rosin et al., 2024).

Second, regarding the extrapolation to long return periods and its dependence on the model assumptions, the Introduction now states that the adequacy of the record length is conditional on the Weibull-tail assumption:

Although short, this record length has been shown to be sufficient to derive stable extreme-value estimates using the SMEV framework, provided that the Weibull-tail assumption holds (Dallan et al., 2024; Marra et al., 2022; Rosin et al., 2024).

Sect. 3.5 now explains that the Weibull-tail assumption is supported by longer records rather than by the nine-year radar record alone, and that the reliability of the estimates is conditional on model adequacy and data quality:

Previous studies provide regional support for this choice. Dallan et al. (2024) identified suitable censoring levels of 90% for 1 h precipitation and 85% for longer durations in the Alpine-Mediterranean region, while Akbary et al. (2026) adopted a common 85% level for sub-daily areal precipitation over Switzerland after testing the Weibull-tail representation. Together with other analyses based on longer observational records (Peleg et al., 2025a, b; Marra et al., 2026), these studies indicate that the Weibull-tail assumption does not rely solely on the nine-year radar record. The radar record is used to estimate the model parameters, with SMEV drawing on multiple ordinary events per year rather than annual maxima alone. The reliability of the resulting return-level estimates nevertheless remains conditional on model adequacy and precipitation-data quality.

We added the following to Sect. 4.2 regarding the interpretation of the 50-year maps:

Differences between the 20- and 50-year maps arise largely from extrapolation of the fitted tails and inherit the associated model uncertainty. Regional contrasts that exceed the bootstrap interval widths (Figure S11b) are considered robust, whereas finer-scale differences at 50 years should be interpreted cautiously.

and the Conclusions now state:

The 50-year estimates are conditional on the SMEV assumptions and on radar-QPE quality and should be considered radar-based estimates rather than stand-alone design values.

Third, regarding the uncertainty that is not represented by the bootstrap, Sect. 4.3 now reads:

Uncertainty is shown as 90% confidence intervals derived from 800 bootstrap repetitions, computed by first calculating the spatial mean return level within each box for every bootstrap realisation and then taking the corresponding percentile range. These intervals quantify sampling variability conditional on the selected event definition, censoring threshold, Weibull-tail model, spatial-smoothing procedure, and radar-QPE product. They do not quantify structural uncertainty associated with possible departures from the independence or stationarity assumptions, variation in precipitation-generating processes, systematic radar-QPE errors, or the choice of smoothing window. In an idealised experiment,

Dallan et al. (2024) showed that, for a 10-year record, this block-bootstrap approach captured most (approximately 86%) of the true 90% confidence-interval width.

Regarding the specific assumptions listed by the reviewer, the Weibull-tail assumption was tested using the long-term rain-gauge records with the procedure of Marra et al. (2023). The analysis identified the minimum left-censoring threshold above which the null hypothesis of Weibull-tail behaviour is not rejected; all gauges and durations passed the test at 90% or at a lower threshold, so the common 90% threshold is applicable to all of them (see also the response to comment #7). This supports the Weibull-tail approximation, although it does not prove that the approximation is exact at every radar pixel or aggregation scale.

The independence of ordinary events is ensured by defining storms as wet periods separated by at least 24 h of dry weather. We added a sensitivity analysis of this criterion (Figure S1), and Sect. 3.4 now reads:

Its suitability was assessed by repeating the event identification using separation periods of 6, 12, 18, 24, 36, and 48 h at representative locations in the six regional analysis boxes. Figure S1 presents the sensitivity of the storm catalogue to the dry-separation criterion. Although the analysis did not identify a uniquely optimal threshold, it indicated that 18–24 h provides a reasonable balance: shorter thresholds substantially increased the number of identified storms, whereas longer thresholds increasingly merged rainfall episodes. We therefore retained the more conservative 24 h threshold, which is also consistent with the maximum accumulation duration considered in this study.

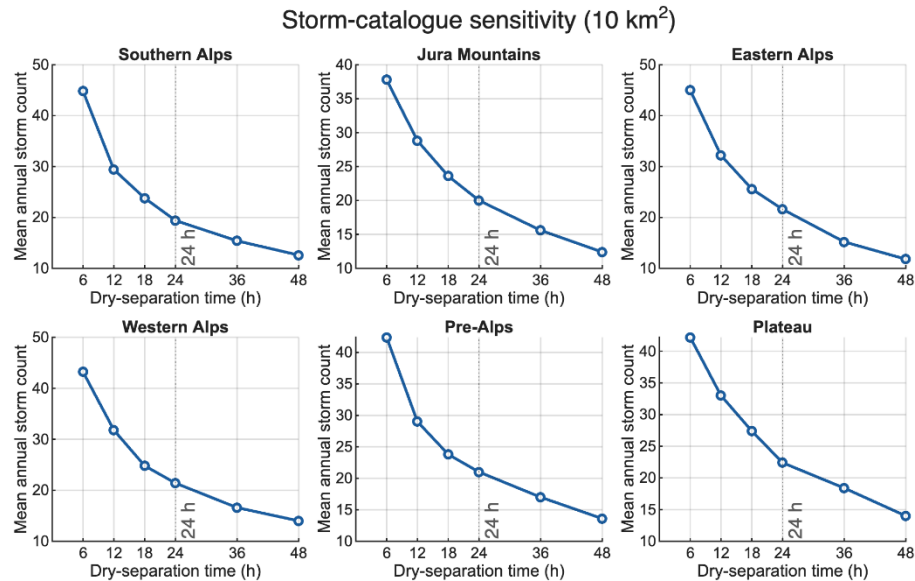


Figure S1. Sensitivity of the storm catalogue to the dry-separation criterion at the 10 km² spatial scale across the six study regions.

Regarding the presence of different weather patterns within the summer, Sect. 3.2 now acknowledges that restricting the analysis to summer reduces, but does not remove, process heterogeneity:

The SMEV does not explicitly account for the seasonal nature of ordinary rainfall events (Falkensteiner et al., 2023); restricting the analysis to summer reduces the pronounced seasonal heterogeneity of Swiss precipitation and focuses the analysis on the warm-season

precipitation regime, thereby improving the approximation that ordinary events are independent and identically distributed within the analysed sample, although some within-season heterogeneity in precipitation-generating processes may remain (Marra et al., 2026).

The effects of radar-QPE errors and of the shape-parameter smoothing on the tail are addressed in our responses to comments #2 and #3, respectively.

2) This paper uses five MeteoSwiss dual-polarization C-band radars with a data resolution of 1 km² and 5 min, and points out that complex terrain can bring problems such as ground clutter, mountain obstruction, bright bands, hail pollution, and uncertainties in the Z–R relationship. The authors also performed summer mean field bias corrections using a multiplicative factor of 0.87 and removed periods of high hail probability.

These quality control steps were necessary and reasonable. However, the current uncertainty estimate primarily relies on resampling years to estimate sampling uncertainty. The paper explains that 100 bootstraps were used, with the 5–95th percentile as the uncertainty range. This method does not adequately reflect the systematic error of radar QPE, especially for short-duration, small-area extreme rainfall events, where QPE error may be more significant than sampling error. This was particularly evident in the Lausanne 2018 event. The paper points out that even at the smallest scale of 1 km² and 30 min, the radar-derived maximum return period was only 33 years, while the return period at the city center rain gauge exceeded 3000 years. The authors attribute this difference to the point-area scale mismatch and the significant underestimation of the peak rainfall intensity of the event by the standard Swiss-tuned Z–R relationship.

It is recommended that the authors supplement at least one of the following analyses: conduct sensitivity tests on the 0.87 bias correction factor, such as $\pm 10\%$ or changing the correction factor by region/duration; conduct scenario tests for short-duration strong convective events with different Z–R relationships or raindrop spectrum assumptions; transform the radar-rain gauge bias structure into an uncertainty range for the recurrence level; and distinguish between sampling uncertainty and QPE-related uncertainty in the IDAF plot.

Response: Thank you for this important comment. We agree that the block bootstrap quantifies sampling variability associated with the nine-year record but does not propagate systematic uncertainty in radar QPE. We have revised the manuscript to distinguish these sources of uncertainty explicitly (see the Sect. 4.3 text quoted in the response to comment #1), and the Figure 6 caption now states that the shaded bands represent sampling uncertainty only.

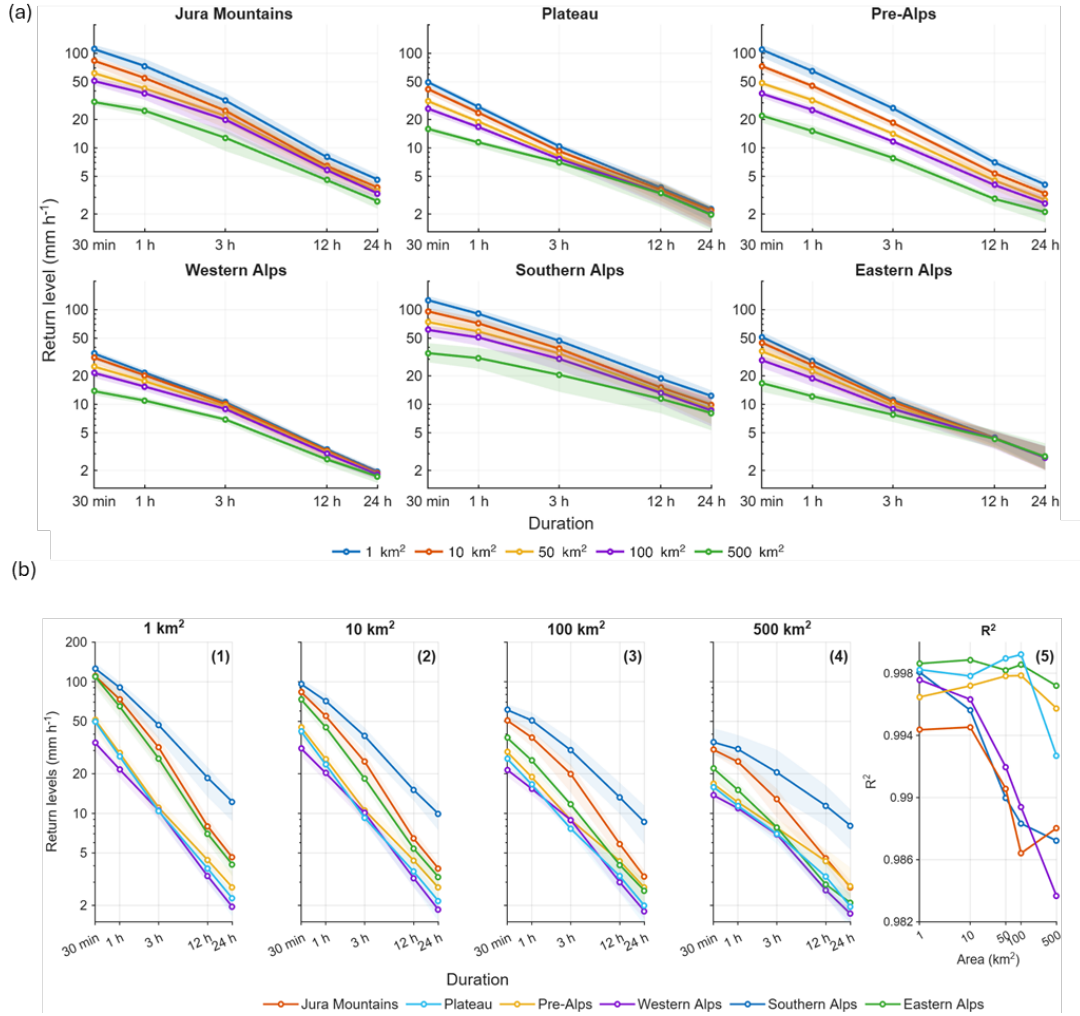


Figure 6. Twenty-year rainfall Intensity–Duration–Area–Frequency (IDAF) relationships for the six analysis regions. (a) Regional IDAF curves, with each panel showing one region and individual curves representing different aggregation areas. (b) Panels (1)–(4) show return levels as a function of duration for aggregation areas of 1, 10, 100 and 500 km², and panel (5) shows the coefficient of determination (R^2) of the linear regression between log-transformed return level and duration as a function of aggregation area. Shaded bands indicate approximate 90% sampling-uncertainty intervals derived from 800 bootstrap replicates; they do not include radar-QPE uncertainty. The locations of the regional analysis boxes are shown and labelled in Figure 1.

In addition, following a bootstrap-convergence analysis, we increased the block-bootstrap ensemble from 100 to 800 repetitions and recomputed all uncertainty intervals and associated figures. The following text was added to Supplement Sect. S2:

The sensitivity of the estimated return levels and their uncertainty intervals to the number of block-bootstrap repetitions was assessed for the six regional analysis boxes using the 10 km² aggregation area and 60 min duration. Figure S5 shows the sensitivity of the median 20-year return levels and the corresponding 5th–95th percentile interval widths to ensemble sizes ranging from 1 to 1000 repetitions. Figure S6 presents their percentage differences relative to the estimates obtained using 1000 repetitions.

We find that the median return-level estimates converged relatively rapidly across all six regions. For small numbers of bootstrap iterations, the estimates fluctuate, but beyond

approximately 100 repetitions, the median estimates are stable. Relative to the results obtained with 1000 repetitions, the median estimates are generally within approximately $\pm 2\%$ after around 100 repetitions.

The estimated uncertainty-interval widths are more sensitive to the number of bootstrap repetitions, and continue to fluctuate beyond 100 repetitions, particularly in regions with wider return-level distributions (Southern Alps and Jura Mountains). However, beyond approximately 200 repetitions the interval widths become progressively more stable. At 100 repetitions, the interval widths differ from the corresponding 1000-repetition estimates by more than 10% in some regions. By 600 repetitions, these differences are generally within approximately $\pm 5\%$, and by 800 repetitions they approach $\pm 2\%$ across the six regions.

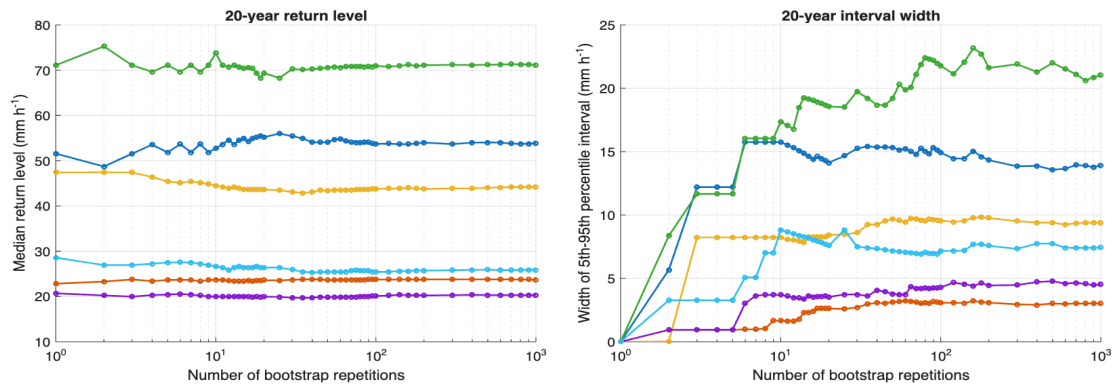


Figure S5. Sensitivity of regional rainfall return-level estimates to the number of block bootstrap repetitions. Results are shown for the six regional analysis boxes using a 10 km^2 area and 60 min duration. The figure shows the median 20-year return level and the corresponding 5th–95th percentile bootstrap interval width for ensemble sizes ranging from 1 to 1000 repetitions.

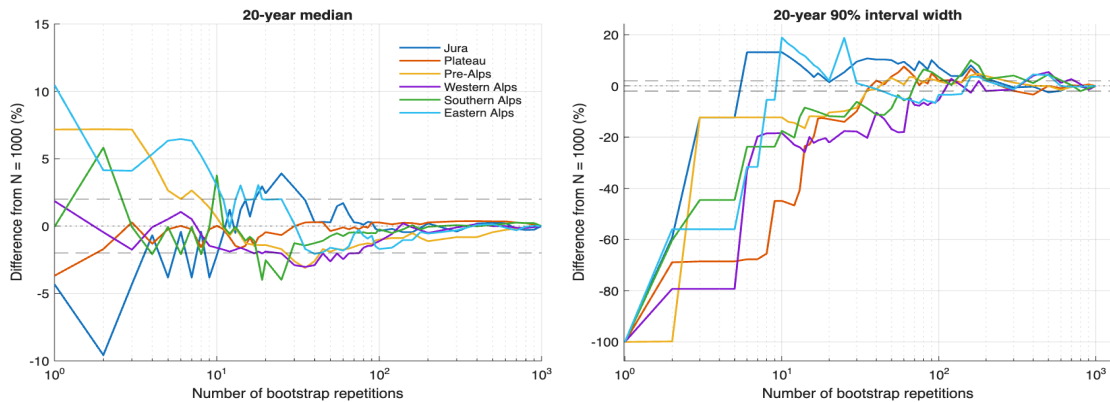


Figure S6. Relative sensitivity of regional 20-year rainfall return-level estimates to the number of block bootstrap repetitions. Percentage differences in the median 20-year return level and in the corresponding 5th–95th percentile interval width are shown for the six regional analysis boxes, using the estimates based on 1000 bootstrap repetitions as the reference. Results are for a 10 km^2 aggregation area and 60 min duration.

Regarding the radar–gauge comparison, we added a conceptual paragraph to Sect. 4.1 that separates the sources of disagreement and states what the comparison can and cannot establish:

The radar–gauge differences in Figures 2 and 3 combine several sources: (i) residual radar-QPE errors, including beam shielding and beam height, vertical variability of reflectivity,

uncertainty in the Z–R relationship, residual hail contamination, and the spatially uniform 0.87 adjustment; (ii) gauge measurement errors, notably wind-induced undercatch, which is largest at short durations and at exposed sites; (iii) the point-to-pixel representativeness gap, which is largest at short durations and long return periods; and (iv) sampling uncertainty in both SMEV fits, together with the different record periods of the gauges (full records) and the radar (2016–2024). Only (i) constitutes radar bias; (ii)–(iv) produce disagreement even with an error-free radar. Recomputing the gauge return levels for 2016–2024 alone preserved the bias pattern, indicating that the difference in record period is not the main driver.

Because these sources cannot be separated with the available data, the observed radar–gauge bias cannot be converted directly into a defensible QPE uncertainty range for the return levels without additional, largely untestable assumptions about their relative contributions. We therefore report the radar–gauge differences by duration, return period and region (Figures 2 and 3) rather than transforming them into an uncertainty band.

Regarding the 0.87 adjustment factor, Sect. 3.3 now specifies how the factor was estimated and what it can and cannot correct:

We apply a multiplicative mean-field bias adjustment of 0.87 to all radar pixels and summer timesteps. The factor was estimated from pooled summer radar–gauge comparisons at all 60 gauges. It corrects domain-mean bias only; regional and duration-dependent errors related to terrain, radar visibility, beam geometry, gauge exposure, and precipitation regime may remain. We did not apply regional or duration-specific corrections because these would use the gauge network to tune the local spatial and temporal structure of the radar estimates, increasing the risk of overfitting. Residual radar–gauge differences are therefore interpreted as region- and duration-dependent discrepancies, and further bias correction is beyond the scope of this study.

The radar–gauge return-level differences do not display a consistent domain-wide bias. Rather, the sign of the discrepancy varies with duration: radar estimates tend to be relatively high for short durations and relatively low for long durations (Figure 2). A uniform $\pm 10\%$ perturbation of the factor would shift return levels in the same direction at all durations, and so could improve agreement for one part of the IDAF curve while worsening it for another. The 0.87 factor was estimated directly from the radar–gauge comparison and was not selected retrospectively to optimise the return-level estimates. As shown in Figure 3, the radar–gauge differences also have a regional structure. Although a spatially or duration-dependent correction could be derived from these differences, we deliberately retained a single domain-mean adjustment that does not use the gauges to tune local spatial patterns, since the same gauges are used for the evaluation. A comparison with alternative precipitation products, such as the CombiPrecip radar–gauge combination, is beyond the scope of this study.

Regarding the comparison with existing Swiss IDF/IDAF and CombiPrecip results raised in the reviewer's general assessment, the revised manuscript compares our results with these studies in two places. The radar–gauge bias pattern is compared with the CombiPrecip-based IDAF analysis of Haruna et al. (2024) in Sect. 4.1:

These findings are consistent with Haruna et al. (2024), who estimated the relative bias of seasonal 20-year return levels over Switzerland from 17 years of CombiPrecip reanalysis data, likewise observing short duration radar overestimation in summer and decreasing biases for longer durations, despite the use of a gauge-adjusted product.

and the spatial patterns are compared with station-based IDF analyses and CombiPrecip climatologies in Sect. 6.1:

The duration-dependent return-level spatial pattern observed in Figure 4 is consistent with the station-based IDF analyses of Haruna et al. (2023) for 1 and 24-h 100-year return levels across Switzerland, and the climatology of warm season 1 km² extremes reported by Panziera et al. (2016, 2018), using CombiPrecip data (MeteoSwiss, 2018a; Sideris et al., 2014). Panziera et al. attribute the observed distribution to instability along the Alpine slopes which can be easily released through orographic forcing, a mechanism absent over the Plateau. Conversely, at higher altitudes near the Alpine crest convection is inhibited due to less moisture availability.

A direct quantitative comparison of return-level magnitudes is limited because these studies use different products (gauge-adjusted CombiPrecip or station data), periods, seasonal definitions and statistical models, so differences in magnitude cannot be attributed to any single factor. The comparison therefore focuses on the bias structure and spatial patterns, which are consistent across studies.

Regarding the Lausanne 2018 event, we agree that the discrepancy illustrates an important limitation for short-duration, small-area convective extremes. Sensitivity tests using alternative Z–R relations or drop-size-distribution assumptions are beyond the scope of this study, which uses the final operational radar-derived rain-rate product rather than the underlying reflectivity and polarimetric variables required to reprocess the event. Such tests would require a separate event-scale radar-retrieval analysis. The Lausanne storm was also highly unusual, producing the most intense 10 min rainfall accumulation recorded in Switzerland, with an exceptional drop-size distribution, so this single event should not be interpreted as evidence of a systematic bias across the nine-year radar archive. We have therefore clarified its interpretation as an illustration of event-specific QPE uncertainty and of the limitations of a 1 km² operational radar product for resolving sub-kilometre convective peaks. The text in Sect. 5 now reads:

The discrepancy probably reflects both spatial representativeness and event-specific radar QPE errors. The gauge measures rainfall at a point, whereas a radar pixel represents the mean areal rainfall over a 1 km² area and therefore attenuates highly localised peaks that can occur over scales of only a few hundred metres (Peleg et al., 2018). The IDAF estimates are consequently applicable only at spatial supports of 1 km² and larger and should not be used to reconstruct sub-kilometre urban rainfall peaks. In addition, Gabella et al. (2018) assert that deviations from the assumed exponential drop size distribution (DSD) also led to a severe underestimation of the 1 km² radar QPE peak rainfall during the event, when using the standard operational Swiss-tuned Z–R relationship. Under this relationship, the instantaneous radar-derived rate near the core is measured at 124 mm h⁻¹, whereas adopting a maritime-tropical Z-R relation yields 265 mm h⁻¹, in closer agreement with the gauge. This highlights the challenges radar faces in providing accurate quantitative estimates of rainfall amounts at the ground, particularly for short-lived, sub-kilometre tropical-like convective peaks in complex terrain. However, it should not be interpreted as evidence of a systematic bias throughout the nine-year radar archive, because the storm’s unusual drop-size distribution and highly localised convective core are unlikely to be representative of the broader summer precipitation record.

3) The paper uses a 15 × 15 km² moving window to smooth the Weibull shape parameters and then re-estimates the scale parameters under fixed shape conditions. The authors believe that this

treatment can reduce sampling noise from the 9-year short record while preserving physically reasonable spatial gradients. Furthermore, the paper mentions that shape smoothing significantly reduces estimation uncertainty without significantly changing the magnitude of the recurrence level.

I believe this is one of the key steps in the paper's method, but the current justification is still insufficient. Extreme precipitation in complex topographic regions often exhibits rapid spatial variations between windward slopes, leeward slopes, valleys, ridges, and rain shadow areas. A 15 km window may be reasonable in some regions, but it could also be overly smoothed in areas with strong topographic gradients, particularly affecting the tail of high return periods and the slope of small-scale IDAF. The narrow IDAF confidence intervals in Fig. 6–7 are explained by the authors as shape smoothing increasing the effective sample size and reducing sampling noise. However, statistically, spatial smoothing also introduces strong prior structures, potentially underestimating local uncertainties. The authors suggest supplementing the data with different smoothing windows, such as 5, 10, 15, and 25 km, to investigate their impact on 20-year and 100-year return levels.

Response: We thank the reviewer for this important comment. We agree that the choice of smoothing scale requires explicit evaluation, especially in complex terrain, where excessive spatial averaging could suppress physically meaningful local gradients and underestimate uncertainty. Spatial smoothing of distribution parameters is a common approach to reduce sampling noise in extreme-value analyses: Deidda et al. (2021) applied a 9×9 km moving average to interpolated GEV parameter fields, Srivastava and Mascaro (2023) smoothed radar-derived GEV parameter estimates using a 40 km radius moving-average window, and Lompi et al. (2025) applied a 3×3 spatial filter to grid-cell SMEV results. Other studies smooth the return levels themselves; Brajkovic et al. (2025) applied a 3×3 averaging kernel to climate-model return-level maps, and Kašpar et al. (2026) applied a weak Gaussian filter to radar-derived design-precipitation fields.

Following the reviewer's suggestion, we evaluated square windows of 5–50 km (5, 10, 15, 20, 25, 30, 40 and 50 km) for 20- and 50-year return levels (the 100-year return levels have been removed from the manuscript). The results are shown in the new Figure S2, and Sect. 3.5 now reads:

The scale and shape parameters are initially estimated independently at each pixel for all duration–area combinations using least-squares regression in Weibull-transformed coordinates (Marani and Ignaccolo, 2015). To reduce sampling variability associated with the nine-year record, the shape parameter is then smoothed using a $15 \text{ km} \times 15 \text{ km}$ moving window. The window size was selected through a sensitivity analysis designed to balance uncertainty reduction against preservation of spatial rainfall variability. Square windows with side lengths of 5–50 km were evaluated for 20- and 50-year return levels using paired bootstrap samples across all durations and the six regional analysis boxes (Figure S2). Increasing the window size generally narrowed the 90% bootstrap intervals but increased the difference from the unsmoothed estimates. The selected 15 km window reduced the interval widths while changing the regional median return levels only modestly in most regions, with larger changes in regions with strong precipitation gradients (Figure S2). It was therefore retained as a compromise across regions and return periods.

After smoothing, the scale parameter is re-estimated at each pixel using the original ordinary-event sample, conditional on the smoothed shape parameter. Smoothing the shape parameter rather than the observations avoids modifying the event distribution before fitting and potentially distorting high-quantile return levels. As shown in Figure S2, the procedure reduces estimation uncertainty while only modestly affecting the estimated

return levels. Comparable spatial smoothing has previously been applied to GEV and SMEV parameter fields and return-level maps (Deidda et al., 2021; Srivastava and Mascaro, 2023; Lompi et al., 2025; Brajkovic et al., 2025; Kašpar et al., 2026).

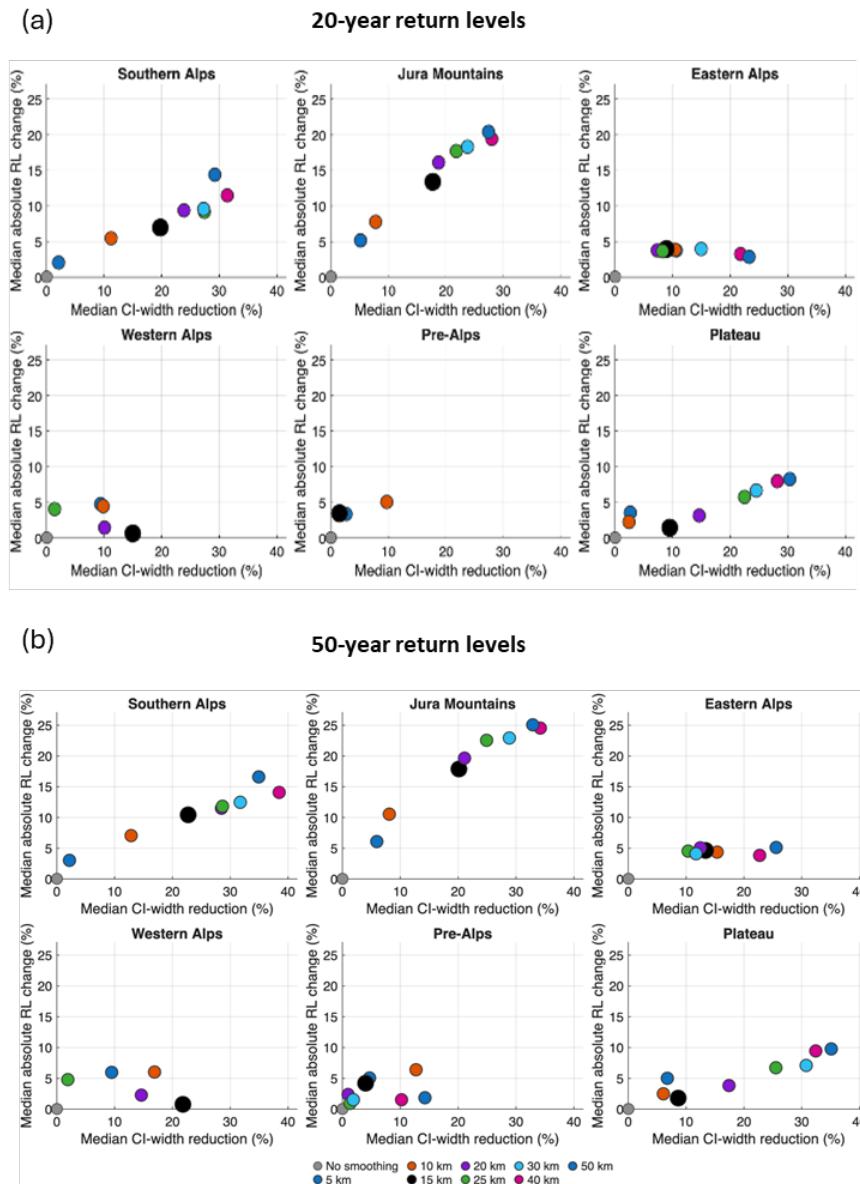


Figure S2. Median reduction in the relative 5–95% bootstrap confidence-interval width versus the median absolute change in (a) 20-year and (b) 50-year return levels for smoothing windows of 5–50 km, relative to the unsmoothed analysis, across the six study regions. Colours indicate smoothing-window size.

The trade-off differs among regions, which addresses the reviewer's concern about areas with strong topographic gradients (Figure S2). Return levels change only modestly in most regions, whereas larger changes occur in regions with strong precipitation gradients, notably the Jura. Because the 15 km window is small relative to the regional divisions, smoothing mainly affects locations close to sharp transitions. A single window was retained for consistency across regions and return periods.

Regarding the slope of the IDAF curves, Figure S3 compares the 20-year IDAF curves obtained with smoothed and unsmoothed shape parameters. The curves are similar in most regions and at most scales, with the largest differences in the Jura. Figure S4 compares the agreement with the gauge-derived return levels for the smoothed and unsmoothed estimates.

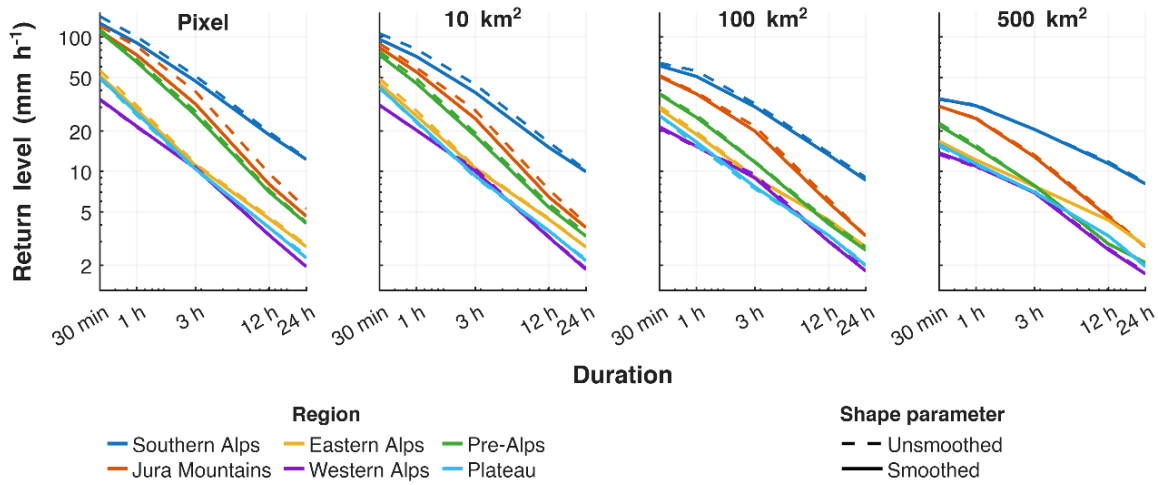


Figure S3. Comparison of 20-year IDAF curves estimated using smoothed and unsmoothed Weibull shape parameters. The regional analysis boxes are shown and labelled in Figure 1 of the main text.

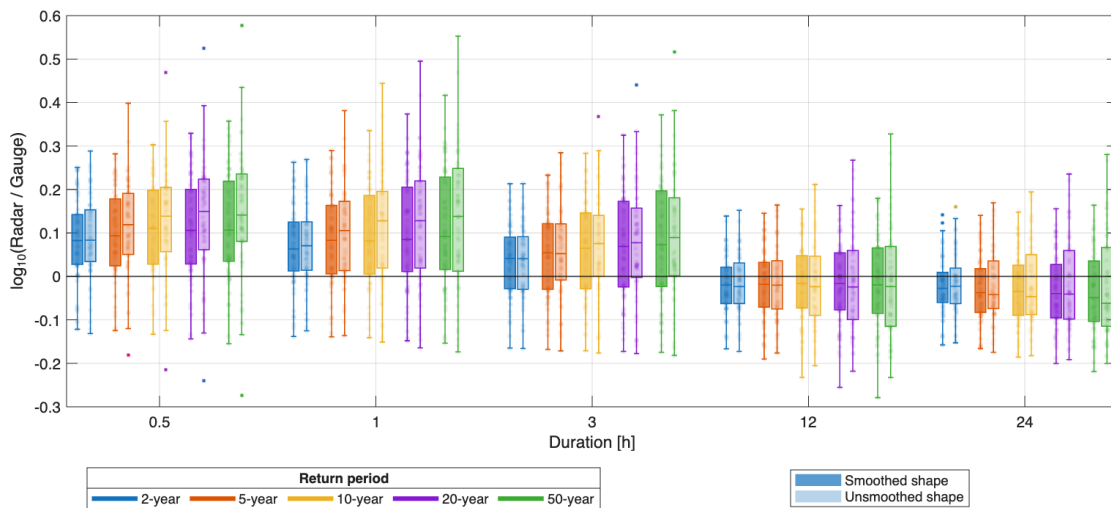


Figure S4. Bias between radar- (1 km² pixel scale) and gauge-derived return levels, expressed as Log10(Radar/Gauge), for estimates obtained using smoothed and unsmoothed Weibull shape parameters.

Regarding the reviewer's point that smoothing introduces a prior structure that may underestimate local uncertainty, we agree. The smoothing is re-applied in each bootstrap replicate, so the intervals include the variability of the smoothed estimates but are conditional on the chosen window. Sect. 4.3 now states that the intervals do not include the uncertainty associated with the choice of smoothing window (see the response to comment #1), and the narrow intervals in Figure 6 should be read in this light (see the response to comment #2).

4) The authors validated the radar-derived 1 km² pixel-scale return level using 60 long-term rain gauges. The validation results show overall consistency between radar and rain gauge data, but with a significant time-dependent bias: short-duration radar overestimates, while long-duration radar

underestimates. Furthermore, the bias exhibits regional structure, with more pronounced overestimation in the Jura and Central Alps, and more likely underestimation in the Plateau, Pre-Alps, and Southern Alps. The main innovation of this paper is areal IDAF, namely, the areal return levels and ARF at 10, 50, 100, and 500 km² scales. Due to insufficient rain gauge network density, the authors could not directly validate larger area scales, which is understandable. However, some statements in the current paper may lead one to believe that 1 km² validation is sufficient to support the reliability of all area scales. In fact, area aggregation changes the error structure: random errors may be averaged, but systematic underestimation or overestimation, terrain-related errors, storm core localization errors, and occlusion errors do not necessarily decrease simply with area. It is recommended that the authors add a clearer limiting statement to the validation section: current independent validation mainly supports radar-derived pixel-scale return levels, while areal return level validation is indirect.

Response: Thank you for this comment. We have revised the validation and the Conclusions to make the scope and limitations of the radar–gauge comparison explicit. We now state that the comparison evaluates the radar-derived estimates at the 1 km² pixel scale only and does not directly validate the areal return levels for 10–500 km². Sect. 4.1 now reads:

As the gauge network is not sufficiently dense to support a robust evaluation at larger areal aggregations, the validation is restricted to the pixel scale and does not directly validate the areal IDAF estimates derived for aggregation areas of 10–500 km². Areal comparisons are possible where gauges are sufficiently dense, as demonstrated for 100 km², 24 h return levels by Rosin et al. (2024). This provides methodological support for areal-scale radar–gauge evaluation, although differences in precipitation regime and dataset mean that their results do not constitute a direct validation of the 10–500 km² estimates presented here.

Following the reviewer's point that aggregation changes the error structure, we added:

Spatial aggregation is expected to reduce random radar errors, which supports the use of the radar data at larger scales; systematic errors, such as terrain-related biases, beam shielding or storm-core displacement, may however persist with aggregation, so the pixel-scale comparison provides indirect rather than direct support for the areal estimates.

The Conclusions now state:

Comparison of the pixel-scale return levels with 60 long-term gauges shows broad agreement, with duration- and region-dependent differences. The areal estimates (10–500 km²) are supported methodologically but cannot be validated directly with the available gauge network.

Specific comments

5) The paper points out that Western Alps show ARF > 1 over 24 hours, explaining that the reference pixel is located in an occluded valley, and as the aggregated area increases, high precipitation from adjacent exposed slopes is included, thus making the areal average rainfall intensity higher than that of the 1 km² reference pixel. This explanation is reasonable in complex terrain and consistent with

the specific behavior of fixed-area ARF in areas with strong spatial gradients. However, $ARF > 1$ is counterintuitive for many readers, especially in engineering applications where ARF is often understood as a reduction factor that decreases or equals 1 with increasing area. It is recommended that the authors provide a clearer explanation to avoid misunderstanding by engineering readers.

Response: We thank the reviewer for highlighting that ARF values exceeding 1 may be counterintuitive in an engineering context. Sect. 3.6 now defines the ARFs explicitly relative to the 1 km² radar pixel:

All ARFs reported hereafter are 1 km²-reference ARFs, rather than point-to-area ARFs.

and we added the following clarification to Sect. 6.1:

The Western Alps exhibit a 24 h ARF slightly exceeding 1 at the 500 km² scale, indicating that the estimated return level of area-averaged rainfall is slightly higher than the corresponding return level at the 1 km² reference scale. Although conventional engineering ARFs are generally expected to decrease with area and remain at or below 1, the ARF used here is defined as the ratio between the return level of area-averaged rainfall and the corresponding return level at the 1 km² radar reference scale, evaluated at the same return period. This location-specific ratio therefore differs from a conventional regional engineering ARF and is not inherently constrained to values at or below 1.

Values exceeding 1 have occasionally been reported empirically for fixed-area ARFs, particularly in regions with strong spatial precipitation gradients (e.g. Catchlove and Ball, 2003; Haruna et al., 2024). In the Western Alps, the 1 km² reference-scale return levels are relatively low within the sheltered inner-Alpine valleys, whereas adjacent, more exposed slopes exhibit higher 24 h return levels. As the aggregation footprint expands to include these higher-return-level areas, the area-averaged return level can slightly exceed that at the 1 km² reference scale. Thus, an ARF greater than 1 does not imply that spatial aggregation increases rainfall; rather, it reflects spatial heterogeneity in extreme precipitation and the comparatively low return level at the reference scale. A similar behaviour is observed in the Eastern Alps, where the 24 h ARF approaches 1 (0.9997) at 500 km².

6) It is recommended that the spatial variation of the number of common events, n , be more clearly explained in the methodology section. In SMEV, n is the average annual storm number; complex terrain and regional climate differences may cause significant variations in n across different regions of Switzerland. It is recommended that the spatial distribution or regional statistics of n be provided.

Response: Thank you for this suggestion. We agree that n varies spatially in response to regional differences in storm occurrence associated with climate and topography. We added Figure S7, which maps (a) the mean annual number of ordinary events, n , and (b) the number of upper-tail events used for SMEV fitting, for each aggregation area. The following text was added to Sect. 4.2:

Before examining the derived return levels, we first assess the number of ordinary events available for SMEV parameter estimation (Figure S7). The mean annual number of ordinary events, n , ranges from 17 to 22 per grid cell across all pixels and aggregation areas (Figure S7a), corresponding to approximately 150–200 ordinary events per grid cell over 2016–2024. After 90% left censoring, 15–20 upper-tail events per grid cell remain uncensored over the nine-year record (Figure S7b). The remaining ordinary events are retained in the likelihood as

censored non-exceedances, so the information used for SMEV parameter estimation exceeds the number of uncensored events. The domain-mean annual number of ordinary events also remained relatively stable among the nine years, indicating that event frequency did not vary strongly over the study period.

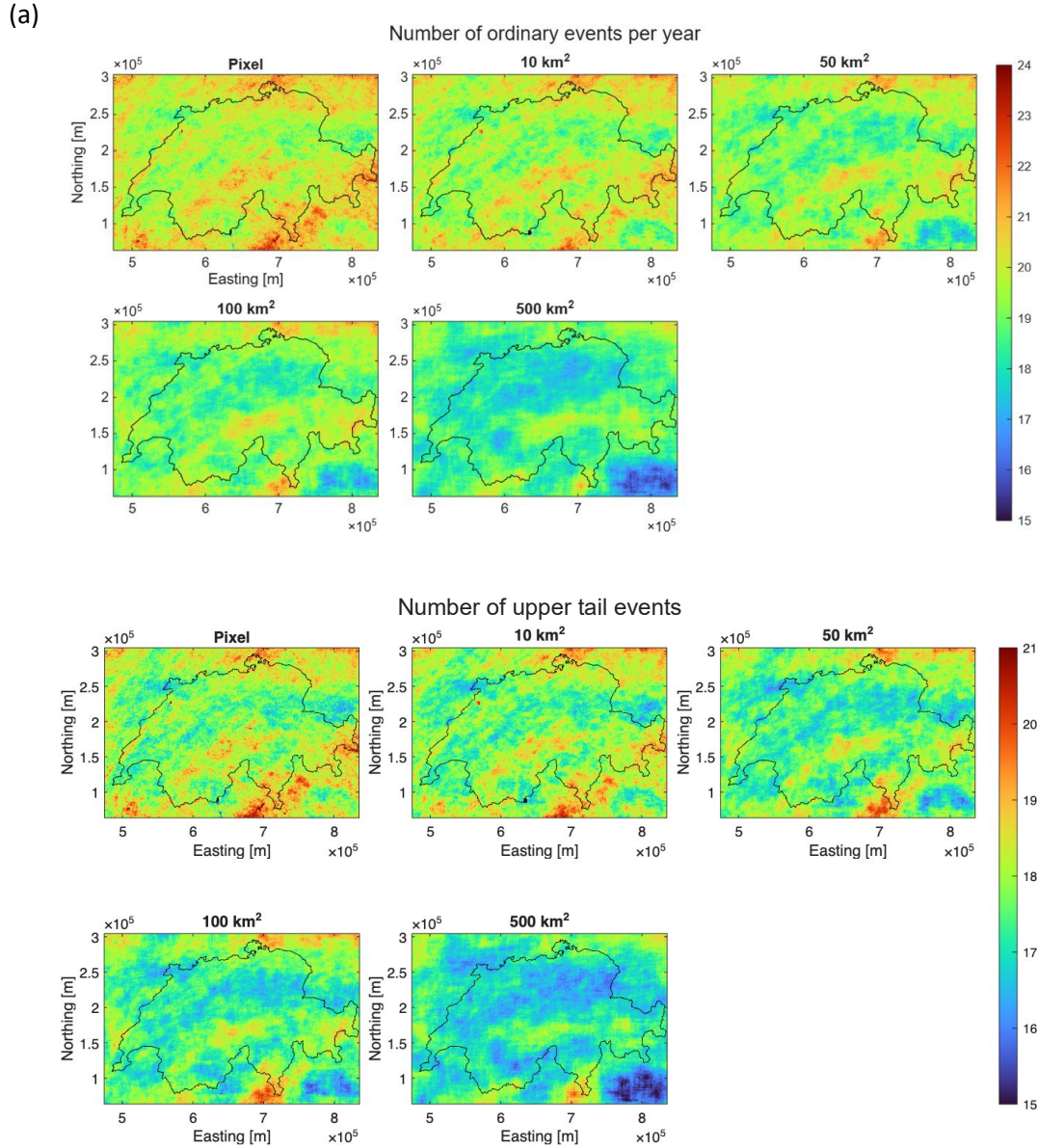


Figure S7. Spatial distributions of (a) the mean annual number of ordinary rainfall events and (b) the number of upper-tail events, defined as ordinary rainfall events retained after 90% left censoring and used for SMEV parameter estimation, for aggregation areas of 1 km² (pixel), 10 km², 50 km², 100 km², and 500 km²

We also added the definition of n to the methodology (Sect. 3.5):

where λ and κ are the scale and shape parameters of the Weibull distribution, respectively, and n is the mean annual number of storms. For each location and aggregation area, n is calculated by dividing the total number of locally wet independent storms by the number of years in the record. It varies spatially and with aggregation area but is common to all

durations because one ordinary event is extracted at each duration from every locally wet storm.

7) The left censoring threshold uses the lowest 90% of events as non-exceedances and is determined based on the sensitivity of rain gauge data. It is recommended that it be explained whether this threshold is equally applicable to radar areal samples, especially considering that the tail of the common event distribution may change after aggregation of different areas.

Response: Thank you, this is an important point. Because the gauge and radar events originate from the same regional storm population, we expect their upper tails to be governed by the same broad precipitation processes and to be representable by the same Weibull-tail family. However, we cannot assume that their intensity distributions or Weibull parameters are identical, as spatial averaging and differences in measurement support may change the scale and shape of the ordinary-event distribution. The same consideration applies to radar rainfall aggregated over different areas. We have revised Sect. 3.5 to explain these points and to acknowledge that the gauge-based test does not by itself demonstrate that 90% is the optimal threshold for every radar aggregation scale:

The Weibull distribution is intended to represent the upper tail rather than the full ordinary-event distribution. Parameter estimation therefore uses left censoring, with observations below the 90th percentile treated as censored. The censoring level was selected using the longer rain-gauge records and the statistical testing procedure described and implemented by Marra et al. (2023). This procedure identifies the minimum threshold above which the null hypothesis of Weibull-tail behaviour is not rejected. Although lower thresholds were acceptable at some gauges, 90% was the lowest common censoring level that provided an acceptable fit across all assessed durations and climatic regions. A common threshold was applied across regions, durations, and aggregation areas to facilitate comparison of the resulting uncertainties.

The gauge-derived threshold is applied to the radar data because gauge and radar events originate from the same regional precipitation regime and are expected to have upper tails governed by broadly similar processes. This does not imply that their intensity distributions or Weibull parameters are identical, given their different measurement supports and the effects of spatial averaging.

Previous studies provide regional support for this choice. Dallan et al. (2024) identified suitable censoring levels of 90% for 1 h precipitation and 85% for longer durations in the Alpine-Mediterranean region, while Akbary et al. (2026) adopted a common 85% level for sub-daily areal precipitation over Switzerland after testing the Weibull-tail representation. Together with other analyses based on longer observational records (Peleg et al., 2025a, b; Marra et al., 2026), these studies indicate that the Weibull-tail assumption does not rely solely on the nine-year radar record. The radar record is used to estimate the model parameters, with SMEV drawing on multiple ordinary events per year rather than annual maxima alone. The reliability of the resulting return-level estimates nevertheless remains conditional on model adequacy and precipitation-data quality.

In addition, provided that the selected threshold lies within the range where the Weibull-tail hypothesis is not rejected, higher thresholds are expected to give statistically equivalent return levels, apart from larger random estimation uncertainty from the smaller retained sample. The

common 90% threshold is therefore not expected to bias the estimates where a lower threshold would also have been acceptable (Marra et al., 2024).

Regarding areal samples specifically, Akbary et al. (2026) tested the Weibull-tail representation for sub-daily areal precipitation over Switzerland, which supports applying a Weibull tail to spatially aggregated rainfall in this region. A 90% censoring level has also been used in other SMEV applications, for example for daily precipitation by Miniussi and Marra (2021) and for sub-hourly and hourly precipitation in Switzerland by Marra et al. (2024) and Peleg et al. (2025b).

8) The text mentions that radar data was available from 2005, but due to system upgrades, it is limited to 2016–2024. It is recommended to more clearly state whether any remaining non-uniformities after system upgrades have been completely ruled out, such as algorithm version changes, radar maintenance, or missing data.

Response: Thank you for raising this point. Our MeteoSwiss co-authors confirmed that, from a radar perspective, 2016–2024 is an exceptionally stable period in terms of network maintenance, data availability and QPE processing, and more homogeneous than any earlier period of the archive; only minor changes occurred. Germann et al. (2022) assessed the radar-only QPE product through 2021 and found broadly consistent performance, and updated internal MeteoSwiss performance statistics show the same stability for the following years. Sect. 2.1 now reads:

Although radar data are available from 2005, major upgrades were implemented in subsequent years, including the introduction of dual-polarisation (2012) and the installation of additional radars (2014 and 2016), which substantially enhanced rainfall estimation in the inner valleys of the Western and Eastern Alps (Germann et al., 2022). The operational radar-QPE algorithm was run in a largely unchanged configuration from 2016 onwards. Over 2016–2021, the latest period assessed by Germann et al. (2022), radar-only QPE showed broadly consistent hourly summer performance, with scatter ranging from 2.5 to 2.8 dB. Given the homogeneity requirements of extreme value approaches, we restrict the analysis to 2016–2024.

In addition, the annual number of ordinary events remained relatively stable over the nine years (see the response to comment #6), which gives no indication of a strong temporal inhomogeneity in the record.

9) The 2018 Lausanne incident illustrates that 1 km^2 radar pixels may significantly smooth out sub-kilometer-scale peaks. It is recommended to explicitly state in the conclusion that the IDAF method described in this paper applies to spatial scales of 1 km^2 and above, and should not be used to interpret sub-kilometer urban storm peaks.

Response: We agree and have added this limitation explicitly to the Conclusions and to the Lausanne case study. The Conclusions now read:

This highlights the necessity of evaluating extremes across multiple dimensions, rather than relying solely on point-scale data. However, we also note that even the smallest radar estimate represents an average over a 1 km^2 pixel and thus cannot characterise sub-kilometre rainfall peaks within an urban catchment. The Lausanne event illustrates this

limitation, as the highly localised peak recorded by the city-centre gauge was substantially smoothed within the corresponding radar pixel.

and the case study in Sect. 5 now states:

The IDAF estimates are consequently applicable only at spatial supports of 1 km² and larger and should not be used to reconstruct sub-kilometre urban rainfall peaks.

Relevant references

Akbary, R., Dallan, E., Astagneau, P. C., Wood, R. R., Marra, F., Brunner, M. I., and Borga, M.: Scale-dependent biases in Alpine sub-daily areal precipitation extremes: added value of convection permitting models, *Hydrol. Earth Syst. Sci.*, 30, 4117–4139, <https://doi.org/10.5194/hess-30-4117-2026>, 2026.

Brajkovic, J., Fettweis, X., Noël, B., Van De Vyver, H., Ghilain, N., Archambeau, P., Piroton, M., and Doutreloup, S.: Increased intensity and frequency of extreme precipitation events in Belgium as simulated by the regional climate model MAR, *J. Hydrol. Reg. Stud.*, 59, 102399, <https://doi.org/10.1016/j.ejrh.2025.102399>, 2025.

Catchlove, R. H. and Ball, J. E.: A hydroinformatic approach to the development of areal reduction factors, in: *Proceedings of the 28th International Hydrology and Water Resources Symposium*, 1.9–1.15, 2003.

Coles, S.: *An Introduction to Statistical Modeling of Extreme Values*, Springer, London, UK, <https://doi.org/10.1007/978-1-4471-3675-0>, 2001.

Dallan, E., Borga, M., Fosser, G., Canale, A., Roghani, B., Marani, M., and Marra, F.: A method to assess and explain changes in sub-daily precipitation return levels from convection-permitting simulations, *Water Resour. Res.*, 60, e2023WR035969, <https://doi.org/10.1029/2023WR035969>, 2024.

Deidda, R., Hellies, M., and Langousis, A.: A critical analysis of the shortcomings in spatial frequency analysis of rainfall extremes based on homogeneous regions and a comparison with a hierarchical boundaryless approach, *Stoch. Env. Res. Risk A.*, 35, 2605–2628, <https://doi.org/10.1007/s00477-021-02008-x>, 2021.

Falkensteiner, M.-A., Schellander, H., Ehrensperger, G., and Hell, T.: Accounting for seasonality in the metastatistical extreme value distribution, *Weather Clim. Extremes*, 42, 100601, <https://doi.org/10.1016/j.wace.2023.100601>, 2023.

Gabella, M., Frassl, M., Germann, U., Joerg, P. C., and Wild, E.: UrbanRain18: High-resolution rainfall observations in urban environments, in: *Proceedings of UrbanRain18, 11th International Workshop on Precipitation in Urban Areas*, 2018.

Germann, U., Boscacci, M., Clementi, L., Gabella, M., Hering, A., Sartori, M., Sideris, I. V., and Calpini, B.: Weather radar in complex orography, *Remote Sens.*, 14, 503, <https://doi.org/10.3390/rs14030503>, 2022.

Haruna, A., Blanchet, J., and Favre, A.-C.: Modeling intensity–duration–frequency curves for the whole range of non-zero precipitation: a comparison of models, *Water Resour. Res.*, 59, e2022WR033362, <https://doi.org/10.1029/2022WR033362>, 2023.

Haruna, A., Blanchet, J., and Favre, A.-C.: Estimation of intensity–duration–area–frequency relationships based on the full range of non-zero precipitation from radar-reanalysis data, *Water Resour. Res.*, 60, e2023WR035902, <https://doi.org/10.1029/2023WR035902>, 2024.

Kašpar, M., Hulec, F., Müller, M., and Crhová, L.: Improving design precipitation estimates by combining estimates from high-resolution adjusted radar data and long-term ombrographic measurements, *Atmos. Res.*, 330, 108557, <https://doi.org/10.1016/j.atmosres.2025.108557>, 2026.

Lompi, M., Marra, F., Deidda, R., Caporali, E., Borga, M., and Dallan, E.: Non-stationary frequency analysis of long-term convection permitting simulations reveals sub-daily extreme precipitation changes in central-southern Europe, *Adv. Water Resour.*, 205, 105071, <https://doi.org/10.1016/j.advwatres.2025.105071>, 2025.

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

Marra, F., Nikolopoulos, E. I., Anagnostou, E. N., and Morin, E.: Metastatistical Extreme Value analysis of hourly rainfall from short records: estimation of high quantiles and impact of measurement errors, *Adv. Water Resour.*, 117, 27–39, <https://doi.org/10.1016/j.advwatres.2018.05.001>, 2018.

Marra, F., Zoccatelli, D., Armon, M., and Morin, E.: A simplified MEV formulation to model extremes emerging from multiple nonstationary underlying processes, *Adv. Water Resour.*, 127, 280–290, <https://doi.org/10.1016/j.advwatres.2019.04.002>, 2019.

Marra, F., Armon, M., and Morin, E.: Coastal and orographic effects on extreme precipitation revealed by weather radar observations, *Hydrol. Earth Syst. Sci.*, 26, 1439–1458, <https://doi.org/10.5194/hess-26-1439-2022>, 2022.

Marra, F., Amponsah, W., and Papalexiou, S. M.: Non-asymptotic Weibull tails explain the statistics of extreme daily precipitation, *Adv. Water Resour.*, 173, 104388, <https://doi.org/10.1016/j.advwatres.2023.104388>, 2023.

Marra, F., Koukoulou, M., Canale, A., and Peleg, N.: Predicting extreme sub-hourly precipitation intensification based on temperature shifts, *Hydrol. Earth Syst. Sci.*, 28, 375–389, <https://doi.org/10.5194/hess-28-375-2024>, 2024.

Marra, F., Dallan, E., Canale, A., Prosdocimi, I., Papacharalampous, G., Borga, M., and Papalexiou, S. M.: Apparent heavy tails of sub-daily precipitation explained by the coexistence of lighter-tailed processes, *Geophys. Res. Lett.*, 53, e2025GL119705, <https://doi.org/10.1029/2025GL119705>, 2026.

MeteoSwiss: Documentation of MeteoSwiss grid-data products: hourly precipitation estimation through rain-gauge and radar: CombiPrecip, Federal Office of Meteorology and Climatology MeteoSwiss, https://www.meteoswiss.admin.ch/dam/jcr:2691db4e-7253-41c6-a413-2c75c9de11e3/ProdDoc_CPC.pdf (last access: 12 October 2026), 2018a.

Miniussi, A. and Marra, F.: Estimation of extreme daily precipitation return levels at-site and in ungauged locations using the simplified MEV approach, *J. Hydrol.*, 603, 126946, <https://doi.org/10.1016/j.jhydrol.2021.126946>, 2021.

Panziera, L., Gabella, M., Zanini, S., Hering, A., Germann, U., and Berne, A.: A radar-based regional extreme rainfall analysis to derive the thresholds for a novel automatic alert system in Switzerland, *Hydrol. Earth Syst. Sci.*, 20, 2317–2332, <https://doi.org/10.5194/hess-20-2317-2016>, 2016.

Panziera, L., Gabella, M., and Germann, U.: A 12-year radar-based climatology of daily and sub-daily extreme precipitation over the Swiss Alps, *Int. J. Climatol.*, 38, 3749–3769, <https://doi.org/10.1002/joc.5528>, 2018.

Peleg, N., Marra, F., Fatichi, S., Paschalis, A., Molnar, P., and Burlando, P.: Spatial variability of extreme rainfall at radar subpixel scale, *J. Hydrol.*, 556, 922–933, <https://doi.org/10.1016/j.jhydrol.2016.05.033>, 2018.

Peleg, N., Koukoulou, M., and Marra, F.: A 2 °C warming can double the frequency of extreme summer downpours in the Alps, *npj Clim. Atmos. Sci.*, 8, 216, <https://doi.org/10.1038/s41612-025-01081-1>, 2025a.

Peleg, N., Koukoulou, M., Rajczak, J., Kotlarski, S., Dallon, E., and Marra, F.: Hotter summers, heavier showers: global warming and its impact on Swiss short-duration rainfall extremes, *Weather Clim. Extremes*, 50, 100829, <https://doi.org/10.1016/j.wace.2025.100829>, 2025b.

Rosin, T., Marra, F., and Morin, E.: Exploring patterns in precipitation intensity–duration–area–frequency relationships using weather radar data, *Hydrol. Earth Syst. Sci.*, 28, 3549–3566, <https://doi.org/10.5194/hess-28-3549-2024>, 2024.

Sideris, I. V., Gabella, M., Erdin, R., and Germann, U.: Real-time radar–rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland, *Q. J. Roy. Meteor. Soc.*, 140, 1097–1111, <https://doi.org/10.1002/qj.2188>, 2014.

Srivastava, A. N. and Mascaro, G.: Improving the utility of weather radar for the spatial frequency analysis of extreme precipitation, *J. Hydrol.*, 624, 129902, <https://doi.org/10.1016/j.jhydrol.2023.129902>, 2023.

Vidrio-Sahagún, C. T. and He, J.: Hydrological frequency analysis under nonstationarity using the Metastatistical approach and its simplified version, *Adv. Water Resour.*, 166, 104244, <https://doi.org/10.1016/j.advwatres.2022.104244>, 2022.

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