Dear Reviewer(s) and Editor,

Following the comments of the reviewers, the following changes were implemented in the updated version of the manuscript:

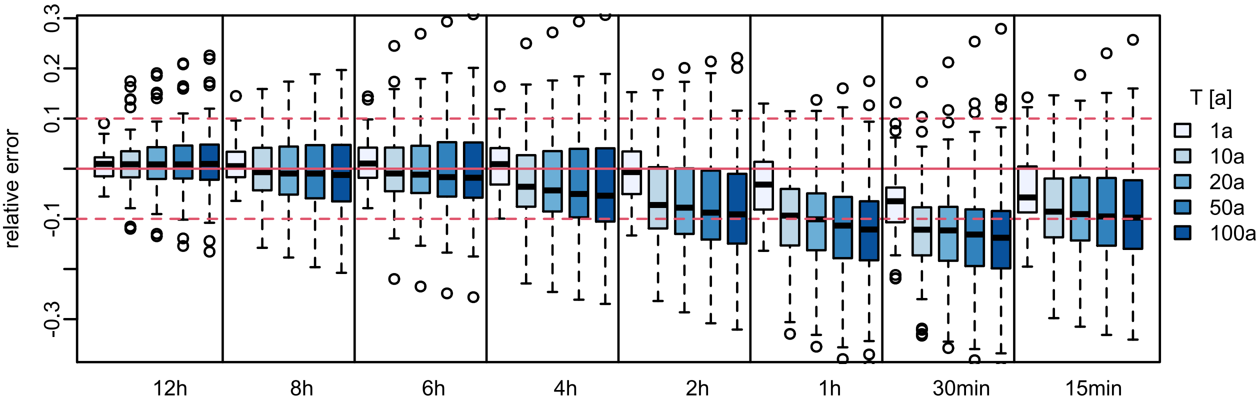
- We have updated the title of the manuscript to "*Regionalisation of Rainfall Depth-Duration-Frequency curves with different data types in Germany*", to give emphasize that we are investigating different data types in this manuscript.
- Line 9-13 in the Abstract has been substituted by the following text to emphasize again that we are investigating different data types in this manuscript:
  "For reliable estimation of such curves, long and dense observation networks are necessary, which in practice are seldom the case. Usually observations with different accuracy, temporal resolution and density are present. In this study, we investigate the integration of different observation data sets under different methods for the local and regional estimation of DDF curves in Germany."
- Following the comments of the reviewer one, the introduction is entirely changed and restructured to emphasize the novelty of this study:
  "Rainfall volumes at varying duration and frequencies are required for the design of water management systems and facilities, like dams or dikes, spillways, flood retention basins, urban drainage systems, etc. These design precipitation volumes are also known as IDF (Intensity-Duration-Frequency) or DDF (Depth-Duration-Frequency) curves, and are derived from an extreme value analysis (EVA) on observed rainfall. For sampling extreme values, either annual maximum series (AMS) or peak-over-threshold (POT) can be used, however for return periods greater than 10 years, there are hardly any differences between the two. Often the AMS are preferred over the POT because the methodology is more direct and easier, whereas the POT method needs a prior assumption on the threshold selection. Afterwards a theoretical probability distribution (PDF) is fitted to the extreme series of a certain duration, in order to extract design rainfall volumes at specific frequency (or return periods). Typically, a Generalized Extreme Value (GEV) distribution is fitted for the AMS series and a Generalised Pareto for the POT series extracted for a fixed duration level. Rainfall extremes of different durations are strongly related to each other, however if the parameter fitting is done independently to each duration level these relations may not be kept (Cannon, 2018). Therefore, generalised concepts as in (Koutsoyiannis et al., 1998), simple scaling (Gupta and Waymire, 1990) or multi scaling Van de Vyver (2015) approaches are used to smooth the extreme statistics over different duration levels. Finally, since the rainfall observations are mostly point measurements, a regionalisation procedure of the PDF parameters to un-observed locations is performed. Methodologically, a distinction can be made between two approaches: a) a direct regionalisation of quantiles, moments or parameters of distribution functions and b) a regional estimation of distribution functions for homogeneous regions. Borga et al. (2005) suggests the regionalisation of the parameters instead of the quantiles. For the direct regionalisation of parameters, regressions (Madsen et al., 2009; Smithers and Schulze, 2001), splines (Johnson and Sharma, 2017) or kriging methods (Ceresetti et al., 2012; Kebaili Bargaoui and Chebbi, 2009; Uboldi et al., 2014; Watkins et al., 2005) are applied. On the other hand, the estimation of regional distributions functions based on the index method proposed by Hosking and Wallis (1997), is one of the most used methods in the literature for the regionalisation of design precipitation (Burn, 2014; Durrans and Kirby, 2004; Forestieri et al., 2018; De Salas and Fernández, 2007). Since the analysis is performed on extreme values, first very long observations are required to ensure a proper fitting of the GEV parameters, particularly of the shape parameter which is of decisive importance for extremes of high return period (larger

than 20 years return period). For instance, Koutsoyiannis (2004a,b) showed clearly that short time series (less than 50 years) can choose falsely a shape parameter of zero (Gumbel distribution) and hide the true heavy-tail behaviour of rainfall extremes (also supported by Papalexiou and Koutsoyiannis (2013) and Papalexiou (2018)). Second, a dense observation network should be available to ensure an adequate estimation of extreme value statistics also at un-observed locations. A less denser network would cause for instance that the kriging interpolated values to be less accurate and the spatial features to be more smoothen in space (Berndt et al., 2014). On the other side, index-based regionalisation can provide more robust estimation at un-observed locations if larger samples (obtained from denser networks) are used (Requena et al., 2019). Third, a high-resolution observation network (with 1- or 5- time steps) is as well necessary to estimate extremes of short durations (at scales of minutes or hours) for catchments that respond quickly to rainfall events (i.e. urban or mountainous areas prone to flash floods). At the moment, no perfect observation network that fulfils these three criteria is available, however different networks or datatypes fulfilling two criteria co-exist. For example, daily observation networks are typically very dense (every 10km) and can have up to 100-150 years of observations, but don't capture the extremes at sub-hourly durations. Digital tipping bucket or weighting sensors can measure the rainfall at 1min time steps and can be dense (every 20-25km), however they are available mostly after 2000 and hence too short for EVA. Long observations at 1min time steps from analogous Hellmann or tipping buckets may be available from 1900-1950 only at some countries (i.e. Germany, Belgium) but are not as dense as digital or daily measurements (>50km). Alternatively, weather radar or satellite data can provide rainfall fields at 1- or 4-km^2 and 5min time steps, but offer short observations (less than 20 years) and suffer from high inaccuracies (Marra et al., 2019). To optimize the DDF estimation, different data types have been combined for instance; Madsen et al., (2017) regionalised extremes in Denmark from 1min observation with daily interpolated values as a co-variate, Bara et al. (2009) employed the simple scale principle to derive DDF curves for sub-daily duration levels (5min – 3h) from daily observations in Slovakia, Goudenhoofdt et al., (2017) used station observations (10min and varying lengths) to correct radar data and estimate the hourly and daily extremes, Burn (2014) pooled together long and short observations at 5min time steps to form the DDF curves in Canada. However, care should be taken when combining information from data types that differ in observation length, temporal and spatial scales. Holešovský et al. (2016) separated the historical data into groups when estimating DDF curves for Czech Republic (long series with 35-40 and short series with 11-15 years of observations), and concluded that the uncertainty at estimating parameters for the short time series is quite high, especially for high return periods. In the index-based regionalisation, regional L-moments are averaged based on the observation length, which may lead to more stable results (Burn, 2014; Requena et al., 2019), however the interpolated index may still suffer from high uncertainties from pooling together short and long time series. This may also be the case when interpolating local GEV parameters with the kriging theory. The regionalisation of the shape parameter may be not representative if short and long observations are pooled together with same importance, thus keeping a fix shape parameter may help to mitigate this problem. Nevertheless, further investigation should be done to ensure if long observations, as more reliable, should have more importance than the short ones when regionalising extreme value statistics. Regarding the temporal scale difference, a study from Paixao et al. (2011) performed in Ontario Canada concluded that the scaling factors should not be used for downscaling daily extremes to durations less or equal to one hour. This is because the extremes at such short durations are governed by other rainfall mechanisms then the daily extremes, and hence a low dependency exists between the two extreme groups. Alternative to *the scaling principle, disaggregation schemes can be applied to the daily data in order to obtain adequate extremes (with return period up to 5 years) for sub-hourly durations (Müller and Haberlandt, 2018). On the other hand, because of the spatial scale inconsistency between weather radar and gauge observations, the weather radar may not be appropriate to estimate directly extremes of short durations (Marra et al., 2019), however they can still be useful to extract sub-daily extremes if used to disaggregate daily observations as done by Bárdossy and Pegram (2017). More complex disaggregation procedures that take advantage of the radar information by implementing an extensive parameter-set as suggested by Lisniak et al. (2013), may also be used to disaggregate daily observation and estimate the extreme values at sub-hourly durations. Nevertheless, to authors knowledge, there is no study in the literature that investigates if disaggregated daily time series can be useful in regionalising extreme values statistics when high resolution data are present, and when so, if they should have the same weights as high-resolution data. Lastly, due to lack of data, in most of the literature, the combination of any two or alternative data types for EVA is validated on observations that are not dense or long enough (longer than 40-50 years). Therefore, it would be interesting to test different methods for estimation and regionalisation of DDF curves extracted from different datatypes, on a long and dense network. The German Weather Service (DWD) has a relatively dense observations network (every 50km) of 1min rainfall data available from 1950 (60-70years), that enables a proper validation of EVA for return periods up to 100 years. Additionally, denser digital observations (every 20km) at 1min time steps (mainly from 2000), very dense (every 10km) daily observations (10-120years) and weather radar observations (from 2000) at 1km^2 and 5min time steps are as well available. As multiple data types co-exist in Germany, it is important to investigate the suitability of methods and data types for the extraction and regionalisation of extreme statistics while validating only at the long and dense observations. In Germany, studies either use the Koutsoyiannis approach or multi/simple scaling approach of GEV 4 parameters to generalise the extremes over different durations. To authors knowledge there is no comparison of the two approaches in the literature. The Koutsoyiannis approach has been implemented in Germany by Ulrich et al. (2020), but on a shorter available 1 min dataset (up to 14 years), while Fischer and Schumann (2018) have implemented the multi scale approach only at a long station (~85 years). Here we investigate which of these methods gives more accurate and precise estimation of DDF based on the long and 1min rainfall data. The same is true also for the regionalisation approaches: to authors knowledge there is no comparison between kriging and index-based regionalisation. Naturally, it is interesting to see which of the methods is more appropriate when validated on a long and high-resolution network, and where lie the advantages and disadvantages of each method when different data types are integrated, and what combination brings the best outcome. For this purpose, we investigate here three competitive regionalisation methods (ordinary kriging, external drift kriging and index-based regionalisation) based on different combination of data types (long series, short series, disaggregated daily series from weather radar parametrisation), while validating only on the long and high-resolution observations. At the moment, a revision of the current design storm maps in Germany (KOSTRA-DWD) is required in order to use additional data and state-of-the-art methodology. Therefore, an additional aim of this study, is to give the basis for development of the new design storm maps in Germany (KOSTRA-2023)."*

- Line 154-156, we have added an explanation why we choose 30 ensembles instead of 100 for the disaggregation of the daily series:

“It was evaluated that the relative error doesn’t improve significantly for more than 30 realisations, as also reported in Müller and Haberlandt (2018), therefore only 30 realisations of disaggregated data were used in this study”.

- Figure 3 has been updated to accommodate as well the relative errors for the 15min durations.



- Lines 162 – 165, we have added an explanation about the relative error of the 15min extremes from the disaggregated time series:
  “Below the duration of 4 hours, there is a clear tendency to underestimate the extremes from LS, up to a median underestimation of 14% at the 30min duration level. At the duration of 15min, a weakening of the underestimation is observed, which is probably due to the instationarity in the original series identified in Section 2.4 below, which predominates only at duration levels up to 15min.”
- Lines 170-172, we have added the explanation for the scope of the 4 hours dry spell duration:
  “*A moving window with the length of each duration level is used to derive the annual maxima, considering a dry duration of 4 hours to ensure that the maxima selected in December and January of two consecutive years are independent from one another.”*
- Lines 194-195 we have added an explanation for the jump present in the data:
  “A possible reason could lie in the limited ability of analogue gauges to register abrupt intensity changes, so that the total amount of precipitation falling in a short time interval may not be fully detected by analogue sensors, leading to positive jumps at sensor changes from analogue to digital. However, as a counter-argument, the so-called "step-response-error" that occurs with digital sensors could also be considered (see e.g. Licznar et al. (2015)).”
- Figure 7 was updated because there were some grammatic errors in the figure.
- Lines 292 – 295 we have added an explanation for the chosen grid resolution:
  “This grid resolution was chosen for two reasons; first it is consistent with the HyRas product from German Weather Service that uses the same resolution, second it is a compromise between the coarsest and finest legible resolution computed from the given density of long series (LS) (the reference for this study) following the suggestions of Hengl (2006).”
- Lines 303-320 we have described the spatial structure of the parameters described by the variograms:
  “Different theoretical variograms were previously investigated, i.e. exponential, gaussian and spherical, with the spherical model together with a nugget effect showing the best fit for the case study. The fitting of the variogram model parameters for different data types and experiments is done automatically by weighted least square fit. Since the automatic fit relies on the initial values of the model parameters, we defined the initial values with trial and error, and accepted a fit that was adequate qualitatively. Figure 8 illustrates the empirical and theoretical normalised variograms for interpolation of the GEV and Koutsoyiannis parameters (after method KO.FIX shown in Table 5) estimated from the three main datasets available: long

series (LS), short series (SS) and 30 realisations of disaggregated daily series (DS). Note that the variograms are normalised in order to ensure a comparison between the different datasets. From this figure a clear difference between the spatial dependency of different datasets, due to different station densities and settings, is visible. The long and short series (LS and SS) exhibits similar relationship with each other for the GEV parameters (μ and σ) but distinguish either in the nugget value (co) or the range (a), whilst the daily disaggregated series clearly exhibit different nugget (c0), range (a) and even sill (c). The differences between the datasets are less visible in the spatial dependencies of the Koutsoyiannis parameters (ϑ and μ), where the three datasets differ slightly in nugget and range. Particularly the spatial dependency of the scale parameter is captured quite differently by the three datasets. Here, LS and SS are differing mainly at the nugget value, where LS has a smaller value than the SS series suggesting that the spatial structure of the scale parameter from SS is smoother than that of LS. On the other hand, the DS datasets exhibit a completely different variogram for the scale parameter, suggesting that the extremes of high return period (influenced mainly by the scale parameter) will have different spatial structures than those of LS and SS series."

- We have added Figure 8 which illustrates the empirical and fitted variograms to the parameters.

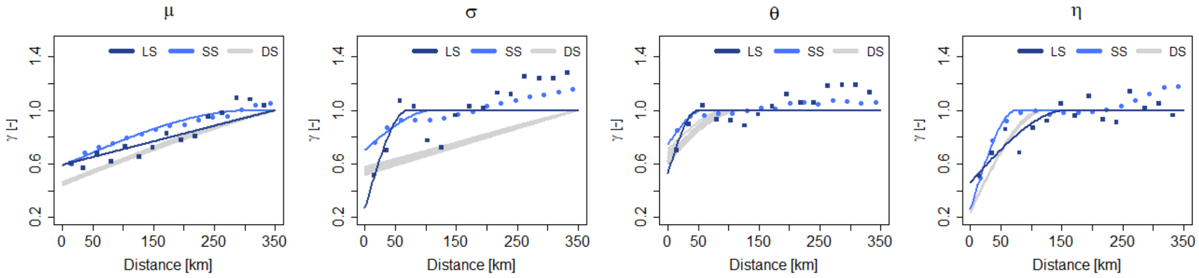


Figure 1 Empirical and fitted spherical theoretical variograms for the GEV parameters and Koutsoyiannis parameters estimated by three different datasets (LS in dark blue, SS in light blue and DS in grey).

- Lines 350 – 353: we have added an explanation about how we have chosen the homogeneous regions:
  "In order to find an appropriate number of clusters, different number of clusters between 2 and 20 are tested and compared based on the homogeneity indicator H1 and whether they were spatially continuous and physically reasonable. The maximum number of clusters of 20 was chosen to ensure a sufficient number of stations and thus a sufficient number of observation years per region (Hosking and Wallis, 1997)."
- Line 393, we have changed the name of the normalized 95% confidence interval with variable to *nCI95width* (also respectively in the text and in Figure 11) to have it in line with our uncertainty analysis study (Shehu and Haberlandt 2022).
- Line 543 – 544 we have introduced a new figure in the Appendix (Figure A5):
  "Expected deterioration in performance when the long series are not present in comparison to the best method selected for regionalisation (KED[LS|SS]) are given in Figure A5 in the appendix."
- Lines 546 – 555 we added a discussion and comparison of our results with other studies:
  "The three different data sets implemented here, distinguish from one another based on the parameter values (as shown in Figure A3 of the appendix) also on the spatial dependency, variograms, shown in Figure 8. When fixing the shape parameter to 0.1, the location and Koutsoyiannis parameters of LS and SS, are in similar range, and the main difference is seen at

the scale parameter (where the SS has high values of the scale parameter than LS). This gives a tendency of the short durations to estimate bigger rainfall volumes for higher return periods. This behaviour is also in agreement reported by Madsen et al. (2017) which used a Generalised Pareto distribution also with a fix shape parameter. Typically, this is treated by index-based regionalisation, where extremes within a region are pooled together to estimate the DDF curves at an unknown location as done in Requena et al. (2019). However, we show here that integrating the LS and SS with external drift kriging, hence accounting for the spatial dependency of the extremes, delivers better performance than grouping them together in the index-based regionalisation (also valid for the LS and DS integration)"

- Figure 13 and 14 were updated, in order to be easy assessible for everyone. Notice that we have selected a new colour palette but the results are the same.
- We have added discussion at the section 4.3 and also compared our results with other studies: *"Here the shape parameter is fixed to 0.1 for whole Germany, which is very similar to results obtained by Ulrich et al. (2021) (shape parameter as 0.11 from the annual GEV approach) and validates our approach. The spatial distribution of the location GEV parameter (μ) follows partly the elevation information, with higher values in the south east, where the German Alps are located. The scale GEV parameter (σ) values are independent of the elevation, with a high localised value near to Münster city. In 2014, there was a very extreme event in Münster which has affected the statistics of the station located in the vicinity. Currently it is not clear how to handle these singular extraordinary events in extreme value analysis in an optimal way. Both Koutsoyiannis parameters (ϑ and η) show similar spatial patterns with lower values in the Alp and other mountainous regions, as well as on the northern-west coast. These parameters exhibit higher variability in space than the GEV location or scale parameters. Overall, the spatial distribution of η parameter follows the spatial structure of the annual rainfall sum in Germany, the distribution of the location (μ) parameter follows the information from the elevation, while the scale (σ) and ϑ parameter don't seem to be influences by any climatologic or site characteristic. This is also seen at Van De Vyver (2012), where annual rainfall and elevation is concluded as important covariates, mainly for the location (μ) parameter, while the scale (σ) parameter didn't have meaningful covariates and the shape parameter didn't show any spatial structure but was kept constant over Belgium. These results agree to a certain extend with the results obtained here. However, the rainfall statistics extracted from short or daily series are considered as more important than the annual rainfall (which itself is an interpolation from point observation). Thus, interpolation of long datasets, should include extreme statistics from short or daily series rather than annual rainfall as an additional information.*

  With these 4 interpolated maps, together with the shape parameter fixed at 0.1, DDF curves can be obtained for any location in Germany. Few examples of design rainfall maps for duration levels 5min, 1 hour and 1 day, and return period Ta=1,10,100 years, are given in Figure 14. For short durations (i.e. D=5 min) the spatial distribution of rainfall extremes is independent from the elevation and becomes more erratic with higher return periods. This is in accordance with the fact that the convective extreme events can happen anywhere and are very low correlated with the orography. With increasing duration level, the relationship between orography and extreme rainfall becomes stronger. As for instance in D=1h, the influence of the alpine regions is visible, which becomes even stronger for the duration of D=1d. In the existing KOSTRA maps, all durations are dependent on elevation. Here, the elevation itself didn't show much effect on the scale (σ) and ϑ parameter, only to some extend on the location (μ) and η parameter. This means that the extremes of longer duration (affected by the η parameter) and of low return period (affected by the location parameter) will show a pattern resembling the elevation. This *is not true for short durations (affected by the ϑ parameter) and high return periods (affected by the scale parameter). This as well agrees with other studies, that report a weak dependence of short duration rainfall (shorter than 1 or 2 hours) with the elevation in Germany (Lengfeld et al., 2019). Lastly, the kriging interpolation as implemented here, opens the possibility to capture better the uncertainty – not only the sample uncertainty which is typically done by bootstrapping the points statistics, but accounting as well the spatial structure of extremes by considering spatial simulations. This results in estimates that will be more precise near to the location of long time series, and less precise in regions far from long time series (Shehu and Haberlandt, 2022)."*

- To make the text more consistent we have substituted the term “network” with “data types” or “series”.
- The number of figures has been updated as we have added Figure 8.
- The reference list has been updated to accommodate new reference used.

with kind regards,

Bora Shehu