Enhancing the usability of weather radar data for the statistical analysis of extreme precipitation events

Andreas Hänsler[1], Markus Weiler[1]

[1]Chair of Hydrology, University of Freiburg, 79098 Freiburg, Germany

5   *Correspondence to*: Andreas Hänsler (andreas.haensler@hydrology.uni-freiburg.de)

Abstract. Spatially explicit quantification on design storms are essential for flood risk assessment and planning. Since the limited temporal data availability from weather radar data, design storms are usually estimated on the basis of rainfall records of a few precipitation stations having a substantially long time coverage. To achieve a regional picture these station based estimates are spatially interpolated, incorporating a large source of uncertainty due to the typical low station density, in

10   particular for short event durations.

In this study we present a method to estimate spatially explicit design storms with a return period of up to 100 years on the basis of statistically extended weather radar precipitation estimates based on the ideas of regional frequency analyses and subsequent bias correction. Associated uncertainties are quantified using an ensemble-sampling approach and event-based bootstrapping.

15   With the resulting dataset, we compile spatially explicit design storms for various return periods and event durations for the federal state of Baden Württemberg, Germany. We compare our findings with two reference datasets based on interpolated station estimates. We find that the transition in the spatial patterns of the design storms from a rather random short duration events, 15 minute) to a more structured, orographically influenced pattern (long duration events, 24 hours) seems to be much more realistic in the weather radar based product. However, the absolute magnitude of the design storms, although bias-

20   corrected, is still generally lower in the weather radar product, which should be addressed in future studies in more detail.

1 Introduction

In the light of flood risk preparedness preparation and climate change adaptation planning there is a rising need for reliable information on the regional to local impacts of urban and sub-urban storm flows (e.g. European Flood Directive: EC, 2007 or 'Guidelines of heavy rainfall management for the federal state of Baden Württemberg': LUBW, 2016 - in German only). This

25   information is usually provided based on data from hydrological and hydraulic modelling chains, which themselves need spatially homogenized information on the magnitude of design storms for various duration and frequencies as input data.

In order to be able to provide reliable information, design rainfall estimates have to be based on sufficiently long time-series of rainfall observations from climate stations at a high temporal resolution (e.g. Charras-Garrido and Lezaud, 2013). Especially for the estimates of rare events (Tr>=100a) this restricts the analyses usually to a rather limited number of precipitation stations,

30   hence requiring substantial spatial interpolation efforts in order to regionalize the information. A further issue when dealing

with station data with very long time-series is non-stationarity of the data which requires an adaptation of the extreme value analyses (e.g. temporally dependent location parameters of the Generalized Extreme Value distribution, Cheng et al., 2014) Apart from station data, however, temporally and spatially homogenized and station-adjusted precipitation data from weather radar become more and more available and have been used in the analysis of design storms (e.g. Overeem et al., 2009;

35 Haberlandt and Berndt, 2016; Panziera et al., 2016; Pöschmann et al., 2021). The main advantage of using weather radar data is the provision of a spatially complete picture of storm events on various temporal and spatial scales, as many short-term and small scale storm events are not captured by the typical network of precipitation gauges (Lengfeld et al., 2020). Hence, design storm estimates based on weather radar data are supposed to provide a more reliable spatial picture than interpolated station data.

40 One serious drawback of this approach, however, is the lack of long-term weather radar records as temporally consistent data is primarily available only for the recent two decades (e.g. Saltikoff et al., 2019). Although recent studies showed that statistical techniques are available to estimate design storms (with return periods in the range from 50 to 100 years) on the basis of shorter data series (e.g. Zorzetto et al., 2016), they still have larger uncertainties when compared to design storms based on data series equal/longer than the respective return periods.

45 In order to overcome short records (or ungauged sites), regional frequency analysis is often used for rainfall as well as for discharge records. Based on the so called region of influence (ROI) approach (Burn, 1990), the records of a target station are extended by pooling data from neighbouring stations located within a target-station specific region. While numerous applications of regional frequency analysis are reported for station data (e.g. Gaál and Kyselý, 2009; Requena et al., 2019), fewer examples are available for the extension of times series from weather radar. Goudenhoofdt et al. (2017) based a regional

50 frequency analysis over Belgium on pooled radar data time-series with a sampling scheme considering radar cells in a radius of 10 km around the target cell for the extension of the precipitation records. While in general the approach lead to promising results, the radial sampling scheme, however, lead to some artificial circular pattern in the final product and does account for the idea of similar regions based on distance alone.

A slightly different approach to conduct a regional frequency analysis is the spatial bootstrapping method (e.g. Uboldi et al.,

55 2014). For a specific station/cell a large number of samples are established by the repeated sampling of independent events from surrounding stations/cells. This approach was recently applied to 11 years of radar data (spatial resolution of 4 km x 4 km) over the state of Louisiana, US (Eldardiry and Habib, 2020). Also in this study, the cell specific ROI, out of which the samples were pooled, was defined by the distance to the target cell. For each cell they set up 500 samples with a sample size of 11 events (in order to equal the actual number of years), each. They found that the method can provide a robust representation

60 of extreme precipitation which is less affected by single outlier events than a non-regional pixel based approach. However, when compared with station based data, the re-sampled weather radar data has a tendency to underestimate the station records. Reasons for this could be on one hand that the definition of the target cell specific ROI based on the distance only might not sufficient, but other factors (e.g. elevation, climate) as it is usually done with station data (e.g. Uboldi et al., 2014) should be incorporated as well. Also the fact that each sample only considers 11 events could be a source of uncertainty.

65 On the other hand, a general 'bias' in the weather radar when compared with stations is visible, generally increasing with rainfall intensity (e.g. Schleiss et al., 2020, Kreklow et al., 2020) as the radar precipitation is an indirect product (derived from reflectivity) integrated over a larger area. This fact is certainly another serious drawback when using radar data for the estimation of design storms. A common approach to correct for such structural biases is the so called bias correction approach (see e.g. Maraun, 2016 for a review on bias correction) developed in climate impact research, but also already applied to

70 weather radar data (Rabiei and Haberlandt, 2015). The basic idea behind bias correction is that structural biases in the data are removed while the specific characteristics (either spatial or temporal) are kept.

A combination of regional frequency analysis and bias correction could be a promising approach in order to generate a robust radar based dataset for the spatially explicit estimation of design storm events. In our study we therefore apply a ROI based approach to extend a climatological record of 19 years of spatially and temporally homogenized weather radar data in

75 combination with a station based bias correction. We focus our study regionally on the federal state of Baden Württemberg (BaWu), Germany as we have two station based, regionally interpolated design storm products available for this region that can be used to evaluate the newly generated regional design storm product based on weather radar data. Furthermore, BaWu is topographically quite complex with an elevation range from 90 m to 1495 m in the study region (Fig. 1a), which leads to spatially rather inhomogeneous rainfall patterns (see Fig. 1a and Fig. 1c).

80 ## 2 Data and Methods

2.1 Radar-based rainfall estimates

We base our work on the spatially and temporally homogenized climatological precipitation radar product of the German Weather Service referenced as RADKLIM (Winterrath et al., 2017) that is available as quasi gauge-adjusted five-minutes precipitation product (RADKLIM_YW_V2017.002; Winterrath et al., 2018). This data consists of post-processed (artefact

85 and attenuation correction) and station adjusted (but only hourly values) precipitation rates on a 1km x 1km grid for the time period from 2001 to 2019. To be able to directly compare our data product to a station based spatially interpolated data product (see section 2.2.2 below) we only use data for the (summer) months from April to October. Furthermore, the increased uncertainty connected to the measurement of solid precipitation can be avoided when focussing on the summer season only. The original five-minute data is reassembled for different durations (e.g. 15, 60, 360 & 1440 minutes) via running sums and

90 analysed separately for each event duration.

2.2 Station based reference data

For an independent reference we use two spatially interpolated design storm estimates based on precipitation station data. Both datasets are frequently used by practitioners for the design and development of flood retention measures in Germany. It has to be noted that both datasets are based on a limited number of stations only and hence, a substantial spatial interpolation effort

95 was necessary in order to provide design storms on a regional scale. As mentioned previously, this is one of the major

uncertainties connected to the station based reference product and actually the main motivation for the estimation of design storms on a spatially homogenized precipitation product.

2.2.1 KOSTRA

The KOSTRA dataset (KOSTRA-DWD-2010R, Junghänel et al., 2017) was compiled by the German Weather Service and
100   can be seen as the national standard with respect to design rainfall in Germany. It provides design rainfall estimates for the whole of Germany for various return periods and event durations. KOSTRA is based on station data for durations below 24h and the raster based REGNIE (Regionalisierte Niederschlagshöhen, DWD,2017; in German only) daily precipitation dataset for longer durations. Design storms are locally estimated for four different event durations (D=15min, 1h, 12h and 72h) by applying a two parameter GEV distribution to the event data. Design storms for other durations are interpolated from these
105   four durations. In order to achieve a spatial picture for the whole of Germany, the local design rainfall estimates are spatially interpolated to a grid on the scale of about 8.2km x 8.2km. It has to be noted that due to the underlying methodological differences KOSTRA cannot be used for a direct one to one comparison but is rather seen as independent reference dataset.

2.2.2 BW-Stat

Due to the limited spatial resolution of KOSTRA an additional station based dataset (available on 1km x 1km) has been
110   recently compiled for the federal state of Baden Württemberg (subsequently referred to as BW-Stat; Steinbrich et al., 2016 - in German only). This dataset forms the basis for a coordinated effort of the state's environmental agency for the management of heavy rainfall and resulting pluvial floods in municipalities (LUBW, 2016; in German only). Since the focus of the state's heavy rainfall management activities is on short to medium range storm events that are mainly of convective nature only data for the extended summer season (April to October) were considered for the establishment of the BW-Stat dataset, representing
115   the fact that the extended summer season is the main season for these kind of storm events (e.g. Ruiz-Villanueva et al., 2012; Haacke and Paton, 2021). Nevertheless, the BW-Stat dataset represents design storm estimates for event durations from 5 minutes to 24 hours, since also heavy rain events of rather frontal nature, characterized by longer time durations but still substantial spatial variability can occur in-between the beginning of April and the end of October.

Like KOSTRA, this dataset is also based on station-specific local design rainfall estimates which were spatially interpolated
120   using a multi-linear regression approach. The finer resolution of BW-Stat when compared to KOSTRA could be achieved by incorporating data from more stations (however often with shorter time series) and other precipitation networks than in KOSTRA into the analyses. In order to set up a sound data base at each of the locations a ROI based events pooling approach (similar to the one described in this paper – see section 2.3) including neighbouring stations at similar altitudes was used. However, due to the limited station density and the fact that generally stations at similar altitudes are pooled together, the
125   horizontal distance between the pooled stations is generally much larger than in the RADKLIM case. Especially over the mountain regions of the Black Forest, Swabian Jura and Alpine Foothills (see Fig. 1a) the horizontal distance in-between the

can be up to 80km. It further has to be mentioned that in the final product all design rainfall values below/above the $5^{th}/95^{th}$ percentile have been set constant (to the $5^{th}/95th$ percentile) in order to prevent for extremely low/high outliers.

In order to estimate design storms the concept of partial series was applied to identify heavy rainfall events and a three

130   parameter Generalized Pareto distribution was applied. For details see section 2.4 below, since we use the identical approach in order to make our data set directly comparable to the BW-Stat data. Also BW-Stat design storms are available for different return periods and event durations (5 minutes to 24 hours). In order to allow a direct comparison with the radar based design storm estimates, the BW-Stat data was spatially re-interpolated (based on the original multi-linear regression based interpolation process and identical station data) to the radar grid.

135   **2.3 Regional sub-sampling**

We assume a storm event with a return period of 100 years to represent the upper end of our analysis. Therefore, we aim for a target length of the underlying time-series of about 100 years of rainfall data to meet the requirements for a profound extreme value analysis (EVA), although we are aware of the fact that a 100yr event is not necessarily present when analysing 100 years of data. Given the 19 years of RADKLIM data already available, we need to pool for each radar cell (cell of interest, COI) the

140   data from four additional radar cells to statistically extent the RADKLIM data series to a respective length (95 years).

Based on the ROI concept we defined for each radar cell a specific sampling area (with a specific sampling probability for each cell assigned) that has to fulfil two criteria. On the one hand, the specific sampling area has to be located in close proximity (in terms of horizontal as well as vertical distance) to the COI in order to be spatially representative. On the other hand, we also want to make sure that we sample additional rainfall events or intensities not present in the COI, so we have also to make

145   sure that the sampling happens not too close to the COI.

The underlying sampling probabilities of the specific sampling area are defined in a two-step procedure as depicted in Fig. 1b. First, a radial area around the COI is defined based on the distance to the radar cell of interest (Fig. 1, panel bI). Probabilities for this radial area are assigned based on a normal distribution with parameters mean and standard deviation of 9 km (cells) and 6 km (cells), respectively. The maximum sampling radius was set to 25 km (cells). These numbers are chosen in order to

150   reflect the typical size of a convective cell in Germany (~25 to 40 km for hourly events in the summer season in BaWu, Lengfeld et al., 2019) but still keep the spatial representation of the sampling region for the COI.

Second, spatial sampling probabilities are additionally defined by the respective altitude of the COI (Fig. 1, panel bII). Again, the probabilities are based on a normal distribution with the altitude of the COI as mean and a standard deviation of 50m. Note that for the few radar cells in BaWu with an altitude above 1150 m (70 cells, see also Fig. 1a) the mean of the normal

155   distribution was set to 1150m (instead of the altitude of the COI), in order to increase the number of possible sampling cells for these locations.

In a final step, both spatial sampling probabilities are normalized with the respective maximum probability, added together and again normalized by the maximum to generate the final spatial probability distribution for the sampling (Fig. 1, panel bIII). Random sampling based on the underlying probabilities is now conducted out of all cells with a probability above a certain

160 threshold (p>=0.8). In order to prevent that neighbouring cells are sampled (which actually would limit the number of additional rainfall events), the sampling probabilities of the cells in a radius of 4 km (cells) of the cell are reduced below the threshold value after each sample is drawn (Fig. 1, panel bIV).

Although the sampling is based on a cell specific spatial distribution of probabilities the random character of the sampling allows to sample different cells for a specific cell of interest in case the sampling is repeated. Hence it is possible to follow an
165 ensemble approach for the sampling in order to quantify the sampling uncertainty. For this purpose and also with respect to the computation and data efforts, the sampling is repeated 10 times. However, to minimize the effect of duplicated samples (cells) in the individual ensemble members, only the five members with the lowest number of cell duplicates are selected.

It has to be noted that the regional sub-sampling is not adapted for the different event durations, but instead data from the identical cells are analysed. However, as mentioned before, the different cells can contribute different numbers of events for
170 each of the event durations, depending on the actual rain amounts. The main reason behind keeping the sampling process constant is that we wanted to make sure that the change in the spatial patterns of the design storms between the different event durations is not affected by the sampling process, but by the precipitation data itself.

2.4 Event definition and extreme value analysis

175 In order to be able to directly compare our data product with the BW-Stat dataset we follow the same approach as applied in Steinbrich et al. (2016). After the sampling process is completed, for each radar cell a data series of 95 (5 x 19) artificial years is available for multiple durations and each of the five ensemble members. Each radar cell is hereby treated as an individual station. Although a time series of 95 years was generated, it has to be kept in mind, that it is actually based on 19 years of weather radar rainfall estimates, only. Hence, the concept of partial series (value over threshold concept) instead of annual
180 series is applied to identify the events for the EVA. The threshold value varies from cell to cell and is estimated to be the value that has a return period of 1 year using the approach of plotting positions T_k for each element k of the partial series (with k =1 representing the maximum event for the specific cell, duration and ensemble member within the 95 artificial years).

$$T_k = (L + 0.2/k - 0.4) \times (M/L) \qquad \text{(EQ 1)}$$

185

with M as the length of the time series in years (95 years in our case). L is the total number of independent events which is in our case estimated by e (Euler's number) times the number of years equals 258 events. This approach is identical to the method applied for the generation of the BW-Stat dataset (hence allows for direct comparison) and follows the guidelines for EVA given by the German Association for Water, Wastewater and Waste (DWA, 2012). Temporal independence of the individual
190 events is ensured if the events are at least 48 hours apart, starting from the maximum event. This time spacing is applied for all durations, although for short duration events this might be a rather conservative definition of independence.

For all events with rainfall rates equal to or above the threshold value, the Generalized Pareto distribution (GPD, see e.g. de Zea Bermudez and Kotz, 2010 for details on the parameters of the GPD) is fitted (individually for each event duration, radar cell and ensemble member) in order to be able to calculate precipitation rates for various return periods. The three (location, scale and shape) parameters describing the GPD are estimated using the L-Moment parameter estimation method. The application of the GPD and the fitting process is similar to the approach used for the generation of the BW-Stat dataset and enables the direct comparison between our dataset and the BW-Stat estimates.

In order to estimate the uncertainty of the parameter estimation a bootstrapping method is applied for each duration, cell and ensemble member to generate 1000 random samples of the events identified for the extreme value statistics. This results in a final total ensemble of 5000 parameter estimates for each cell and duration, hence allowing to explicitly assign confidence intervals to the estimated design storms. The chosen approach allows to eventually separate between the uncertainty range resulting from the different pooling (spanned in-between the five ensemble members) from the full range.

2.5 Bias-correction of RADKLIM Data

As introduced previously rainfall estimates from weather radar are known to frequently underestimate the magnitude of extreme rainfall events when compared to station data. This is usually triggered by the fact that radar measurements represent an integrated measurement of 1km x 1km while station data is a point measurement, but also other effects like an underestimation of high-intensity rainfall estimates using fixed Z-R relations for typical convective and stratiform events may play a role (e.g. Thorndahl et al., 2014). In order to compensate for such structural biases, we decided to match the magnitude of 1yr design storms (which can be derived in a rather robust manner) of the BW-Stat dataset and the radar data (and hence also improve the comparability of the two datasets). To achieve this match of datasets, a quantile mapping approach (e.g. Cannon et al., 2015) was applied. This approach has the advantage that it corrects bias for the whole probability distribution (either temporal or spatial) to match the distribution of the target dataset but keeps the respective spatial or temporal pattern of the data.

For each station within the analysis region we select the closest four radar cells. QM correction is applied to the distribution of the location parameter (which can be taken as a proxy for a 1yr event) of the GPD of all stations and their corresponding cells for each duration separately. The resulting correction function is then applied to the frequency distribution of the location parameter of the full radar data set (again separately for each duration). All design storms are then calculated based on the corrected location parameter, however the shape and scale parameters of the GPD have not been corrected in order to keep the consistency within the data. The design storm estimates form bias-corrected weather radar based GPD parameters is referred to as RAD-BC whereas the non-bias-corrected version is named RAD.

3. Results

3.1 Sampling statistics

In the perfect case each radar cell would exactly occur five times (once as COI, four times as additionally selected cells) in each member of the final dataset. However, as we do not prescribe how often a cell is drawn, the relative contribution of a single radar cell to the data set deviates from the optimal case (Fig. 2a). Overall about 90% of the radar cells contribute their events between 1 and 7 times to the final data set, with a maximum at 4 repetitions. Cells that occur more than 10 times are extremely rare and can be neglected, leading to the conclusion that the final dataset is not biased towards single cells. With respect to the regional characteristics of the sampling, about 90% of the sampled cells have a distance to the COI in the range of 5km to 14km with a maximum occurrence at a distance of about 8km (Fig. 2b). The inter-ensemble variability in the sampling statistics is rather low (depicted by the small error bars).

The spatial distribution of the effective ensemble size is depicted in Fig. 2c. Here we basically show how often for a COI the identical cell has been sampled in all of the five ensembles. Although we select our five ensemble members in a way that the number of duplicates is minimized (in an optimal case only the COI would be duplicated), duplicates partly still occur and hence reduce the effective ensemble size. So in case only the COI is duplicated in all members, we set the effective ensemble size to five. If two cells (e.g. COI and one of the four sampled cells) is duplicated the effective ensemble size is set to four, etc. While for the most parts of BaWu the effective ensemble size is five it is slightly reduced along the main slopes of the mountainous Black Forest (located in the west of BaWu) and Swabian Jura (located more to the centre of BaWu, see also Fig. 1a for regional specification) regions. But only 13 out of about 41000 cells have 4 duplicated cells in the ensembles and no cell with all 5 ensemble members having the identical sub-selection of neighbouring cells occurs. Based on these findings the number of ensemble members could have been enlarged, if more computational capacity would have been available, although over the mountain regions the added value of more ensemble members might be limited due to the increase in duplicated cells. Furthermore, it has to be kept in mind that even if some cells are duplicated in different ensembles, it doesn't necessarily mean that the EVA is based on the identical events, since the selection of the events is done subsequently to the cell sampling. Given the fact that the first event included in the EVA is the event that has the maximum rain rate per duration out of all of the five contributing cells and further having the prerequisite that two events have to be at least 48 hours apart to be considered in the EVA it is very likely that (depending on the previously identified event) the same cell combination contributes different events to the final EVA in the different ensembles. This is also supported by the fact that in each of the ensemble members each cell contributes roughly the same number of events to the final EVA. Considering all cells in BaWu, each cell contributes about 20% of events, with a range from 13% (5th percentile) to 28% (95th percentile).

3.2 Bias correction

The impact of the quantile based correction of the location parameter is depicted in the form of cumulative frequency distributions (CFD) in Fig. 3. While the uncorrected radar data substantially underestimates the 1yr design storms, the bias

corrected version overlaps (by purpose) almost perfectly to the station data when only the grid cells representing station points are included (upper row). Considering all of BaWu the comparison between interpolated station data and bias corrected radar data leads to slightly larger differences (bottom row) also partly resulting from the assumptions behind the spatial interpolation

260 of the station data. It has to be noted that both, BW-Stat and RAD-BC estimates, still show substantially lower rain rates for the 1yr design storms than the KOSTRA reference dataset, for most parts of the distribution. The overestimation of extremes in the case of shorter event durations can be attributed to the lower spatial resolution of KOSTRA. Linked to this is also the substantially lower variability of KOSTRA, when compared with the other two datasets.

What should be kept in mind is the fact, that the applied bias correction is not having the same effect for longer return periods.

265 Correcting 1yr design storms only basically means that a certain rain amount is added to all events included in the EVA, hence, the relative contribution of the bias correction decreases for less-frequent design storms (see the differences between RAD and RAD-BC in Fig. 5).

3.3 Comparison of design storms

The spatial patterns of a 100yr design storm for four different selected event durations (15, 60, 360 and 1440 minutes) for the
270 two station based reference datasets (KOSTRA, BW-Stat) as well as for the bias-corrected and re-sampled RADKLIM dataset (RAD-BC) are depicted in Fig. 4. Additionally, the absolute difference between BW-Stat and RAD-BC datasets is depicted. Note that the RAD-BC dataset represents the ensemble mean of the five individual sample products and that the data is spatially smoothed with a 3 by 3 cell filter to avoid single outliers. For comparison we compiled the identical figure for a 1yr design storm (see Fig. A1 in the appendix).

275 In the KOSTRA dataset orographic induced patterns with elevated storm intensities along the Black Forest mountains and the Swabian Jura as well as the Alpine foothills (see Fig. 1a for regional specification) in the far south east can be seen for short and long duration events. This pattern can be expected since the z-coordinate was incorporated (although with different weights for the different durations) in the interpolation of the station data (Junghänel et al., 2017): The 360-minute design storm in KOSTRA is actually interpolated from the 60 min and 12h (not shown) design storms and also the 24h design storm represents
280 an interpolated value (interpolation between 12h and 72h design storms). The Black Forest region is also characterized by high-intensity design storms in the BW-Stat dataset for both, short and long duration events. However, especially for events with longer duration the dataset shows very dominant, high-intensity design storms in a region located between the Lake of Constance and the Black Forest, usually known to represent rather a rain shadow area due to fronts moving in from the west (see Fig. 1c).

285

The spatial patterns in the RAD-BC dataset differ quite substantially from the patterns of the two station based reference datasets and also shows a distinct pattern change between short and long-duration events. While the spatial patterns of the 15 and 60 minute 100yr design storms show no relation to the orography or orographically induced rainfall patterns (but a slight north-south gradient) it changes in the case of the 1440 minute 100yr design storm events to a picture very similar to the April

290   to October mean rainfall distribution. This can also be proofed with a spatial correlation analysis with the mean rainfall

estimates from REGNIE resulting in an increase in the correlation coefficient from r=0.25 (15 minute events) to r=0.75 (1440

minute events). In the case of BC-Stat r remains below 0.6. Actually the spatial pattern of RAD-BC design storms is much

more in line with what is expected from the underlying processes (e.g. pure convection triggered, small scale and short duration

event versus more organized larger scale frontal systems for longer duration events (Lengfeld et al., 2019; Kaiser et al., 2021).

295   Interestingly, the spatially pattern in the BW-Stat dataset is following this behaviour in the case of a 1yr design storm (similar

to RAD-BC, see Fig. A1).

With respect to the absolute values, the direct comparison of BW-Stat and RAD-BC design storm intensities reveal that there

are regions with substantially larger intensities in the RAD-BC dataset (e.g. especially in the far south east for the 1440 minute

events) due to the difference in the spatial patterns. Also in case of the 1yr design storms (Fig. A1) RAD-BC shows generally

300   larger intensities than present in BW-Stat over the mountainous regions, although this is most probably largely affected by the

fact that in BW-Stat all values above the 95th percentile are set to the respective percentile value.

However, when integrated over the whole study region it becomes clear that although the radar dataset was bias-corrected for

the 1yr design storm events, it still shows lower rainfall magnitudes for 20yr and 100yr design storms than the two station-

based reference datasets. In Fig. 5 we depict the CFD of the different datasets for four different durations and two (20yr and

305   100yr) return periods. We additionally included the non-bias-corrected (but spatially resampled) radar dataset (RAD; green

line) in the figure to illustrate the effect of the initial bias correction. Additionally, the respective confidence interval for the

RAD-BC dataset (see section on uncertainty estimation below) is included. Apart from the very high and low percentiles, the

ensemble mean of the RAD-BC storm events is about 5 to 10mm lower than the respective rain rate of BW-Stat. Nevertheless,

the uncertainty range spanned within the two station based reference datasets is quite large itself. While there are cases where

310   the KOSTRA dataset lies within the confidence interval of the RAD-BC dataset (e.g. 100yr design storm with duration of 15

min), the difference to KOSTRA is sometimes even larger than to BW-Stat (e.g. 20yr design storm with duration of 360 min).

3.4 Uncertainty estimate

In order to be able to quantify the uncertainties for the newly developed RAD-BC dataset we conducted a twofold uncertainty

analysis based on an ensemble based cell-sampling approach and classical bootstrapping for the identification of parameter

315   uncertainty. The confidence interval in Fig. 5 is defined by the 10th and 90th percentile of the large data sample generated by

1000 bootstraps runs for each of the 5 ensemble members (so basically combining both sources of uncertainty). The confidence

band of the CFD spans about 5mm in the case of 20yr design storms and about 10mm in the 100yr case. The range of the five

ensemble members only (without bootstrapping) is defined by the stippled line and accounts already for a large amount of the

total uncertainty band demonstrating the importance of the ensemble based sampling approach.

320   The spatial patterns of the 10th and 90th percentile are rather similar to the patterns of the ensemble mean (see Fig. 6), and the

uncertainty range of the respective rain rate is for most regions between 15 and 20% in the case of 60 minute events and

between 10 to 15% in the case of 1440 minute events, with relatively larger ranges in regions with lower values for the mean

storm intensity. However, there are certain spots (e.g. the northern parts of the Black Forest in the case of a 100yr 1440minute design storm - framed with a dashed square in Fig. 6 - or various smaller regions in both examples) that have a slightly larger

325 uncertainty range, although the mean storm intensities are large as well. In order to reveal the uncertainty contribution resulting from the ensemble sampling we highlighted regions with a large (> 75% of the range) contribution of the sampling uncertainty. Generally, the contribution of the sampling uncertainty is larger in regions with a lower overall uncertainty range. However, there are various spots with relatively larger uncertainty that are dominated by the sub-sampling uncertainty. The previously mentioned enhanced uncertainty in the northern Black Forest case seems to be substantially influenced by sampling uncertainty

330 in its eastern parts, whereas the uncertain parameter fitting dominates in the central and western parts, indicating a rather inhomogeneous pool of heavy rainfall events in these regions.

4. Discussion

The major benefit of the RAD-BC dataset is certainly the possibility to derive spatially homogenized heavy rainfall estimates also for less frequent (up to 100 yr return period) events. A comparison with the station based spatially interpolated reference

335 products revealed that the spatial patterns of the design storm events with various durations fit much better to the theoretically expected spatial patterns (random pattern for short duration events versus large-scale, orography oriented patterns for longer duration rain events) than the interpolated station products. However, the tendency to underestimate the magnitude of design storms especially for less frequent events in comparison to the reference datasets is something which should be examined in further detail. One has to keep in mind, however, that a direct one to one comparison is only possible with the BW-Stat data

340 since the KOSTRA data has on the one hand a much lower spatial resolution and is based on a different set of rainfall stations and data from other seasons. On the other hand, KOSTRA uses a different EVA approach based on a two parameter GEV distribution (Junghänel et al., 2017).

When comparing the non-bias-corrected (scale and shape) parameters of the GPD of BW-Stat and of a single member of RAD-BC (Fig. 7) it can be seen that for the short durations (15 and 60 minute events) the scale parameter is lower in the RAD-BC

345 data. For the long (1440 min) events, however, the deviations in the magnitude of the design storms seem to result mainly from the shape factor which is lower in RAD-BC. The intermediate 360 minute events are affected by both effects. This finding is also true when looking at various topographic sub-regions and other ensemble members of the RAD-BC dataset.

The lower values for the scale/shape parameters of RAD-BC can partly be attributed to the fact, that for high rainfall intensities radar data is known to underestimate rainfall amounts due to the reflexivity bounds (e.g. Schleiss et al., 2020). A recent

350 comparison of the RADKLIM data with station data further revealed that fewer heavy rainfall events are detected in RADKLIM than in the station data. The average rainfall amount per heavy rainfall day (< 20 mm of rainfall) is, however, almost identical (Kreklow, et al.,2020).

The general underestimation of heavy rain events in RADKLIM is only partly corrected for by the applied bias correction of the location parameter as it is an additive correction which corrects less frequent events relatively less than the more frequent

355 events. Another approach would be to apply the bias correction in a multiplicative manner, by estimating a grid-specific multiplication factor on the basis of the uncorrected and corrected 1yr design storms. This would lead to a substantial increase in the rainfall amounts also for the less frequent events (see Fig. A2) but the spatial patterns of the RADKLIM data would be preserved. Nevertheless, one has to keep in mind that a multiplicative correction is disrupting the homogeneity of the sampled events of the radar data and the added rainfall amounts to the more intense rainfall events are rather large. It is also questionable

360 if a correction factor derived from the correction of 1yr events can be applied to events with a much lower frequency. Further it has to be kept in mind that the BW-Stat data itself is an indirect product with events pooled from surrounding stations with similar altitude. This leads to the fact that especially in the more mountainous areas of BaWu the events from the same set of stations are included in the EVA, which itself could lead to a biased picture in the magnitude of the derived design storms. While the location parameter still can be seen as rather robust, it is questionable if the derived scale and shape parameters

365 could be used with the same reliability for the bias correction of the radar data.

A promising way to proceed could be to only use a small subset of stations that have a reasonable long record to develop a frequency and duration specific correction function which could then be regionally applied to the radar data. However, for BaWu there are only two stations with high temporal precipitation records available with a data series length of more than 50 years (Steinbrich et al., 2016) proposing a major challenge for this approach. Another possible approach would be to base the

370 correction on the underling observed rainfall events itself instead of correcting design storms. This would have the benefit that the also the high-intensity events would be directly corrected and are not derived based on the correction factors estimated for less-intense events. A detailed comparison of the different bias correction approached and their implications on the derived design storms is currently in progress.

On top of applying bias correction methods, using a weather radar product that is compiled at a higher spatial resolution could

375 also be a promising approach. This could have the positive aspects that the difference between point versus area measurements become smaller. However, a real benefit would only be achieved if the deviations between rainfall estimates of weather radar and station data are not increasing with rainfall intensities, which basically calls for a non-static application of the Z-R relation in the weather radar product.

5 Conclusions

380 We present an ROI based approach to prolongate a 19yr climatological weather radar dataset of rainfall estimates in order to enhance its usability for the development of region specific design storm events. The established method has various positive aspects. The main improvement is the development of a spatially homogeneous dataset that allows for the calculation of rare extreme events that is not dependent on spatial interpolation methods that is often the main error source when building a regional dataset based on station data. Moreover, the chosen sampling approach allows on the one hand to control the sampling

385 region based on physical aspects. It further prevents that artificial structures previously reported in literature (e.g. development of circle structure, Goudenhoofdt et al., 2017) are dominant in the final dataset. Due to the combination of an ensemble-based

sampling approach and a bootstrapping based parameter estimation an explicit designation of associated uncertainty ranges is possible which is a major added value for the application by practitioners.

Nevertheless, the current version of the data still has some shortcomings that need to be addressed in the future. While the
390 applied bias correction approach substantially improved the outcome, the deviation to the two existing station based reference datasets in the case of the less frequent events is still something that has to be clarified in the near future. In order to improve the compatibility with the KOSTRA dataset it might be worthwhile to apply the KOSTRA EVA to the resampled event database which underlies RAD-BC. Furthermore, the previously proposed training of the RAD-BC dataset on some high-quality long-term temporally highly resolved station data could be a way forward to enhance the credibility of the RAD-BC
395 dataset.

Author contributions

AH and MW jointly designed the experiment. All data analyses have been conducted by AH. The interpretation of the results as well as the drafting of the manuscript was conducted jointly by AH and MW.

Acknowledgements

400 This work was conducted within the research activities on heavy rainfall at the Chair of Hydrology, University of Freiburg, Funding for these research activities are provided by the State Office for the Environment, Measurements and Nature Conservation of the Federal State of Baden-Württemberg (LUBW) as well as the Regierungspräsidium (governing council) Stuttgart.

405 **References**

Burn, D. H.: Evaluation of regional flood frequency analysis with a region of influence approach, Water Resources Research, 26, 2257-2265, 1990.

Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes?, Journal of Climate, 28, 6938-6959, 2015.

410 Charras-Garrido, M. and Lezaud, P.: Extreme value analysis: an introduction, Journal de la Société Française de Statistique, 154, 2013.

Cheng, L., AghaKouchak, A., Gilleland, E., and Katz, R. W.: Non-stationary extreme value analysis in a changing climate, Climatic Change, 127, 353–369, 2014.

de Zea Bermudez, P. and Kotz, S.: Parameter estimation of the generalized Pareto distribution—Part I, Journal of Statistical

415   Planning and Inference, 140, 1353–1373, 2010.

DWA: Starkregen in Abhängigkeit von Wiederkehrzeit und Dauer, Hennef, Germany: DWA: German Association for Water, Wastewater and Waste) (in German), 2012.

EC: DIRECTIVE 2007/60/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 October 2007 on the assessment and management of flood risk, Official Journal of the European Union, L 288, p. 27–34, 2007.

420   Eldardiry, H. and Habib, E.: Examining the Robustness of a Spatial Bootstrap Regional Approach for Radar-Based Hourly Precipitation Frequency Analysis, Remote Sensing, 12, 3767, 2020.

Gaál, L. and Kyselý, J.: Comparison of region-of-influence methods for estimating high quantiles of precipitation in a dense dataset in the Czech Republic, Hydrology and Earth System Sciences, 13, 2203-2219, 2009.

Goudenhoofdt, E., Delobbe, L., and Willems, P.: Regional frequency analysis of extreme rainfall in Belgium based on radar

425   estimates, Hydrology and Earth System Sciences, 21, 5385-5399, 2017.

Haacke, N. and Paton, E. N.: Analysis of diurnal, seasonal, and annual distribution of urban sub-hourly to hourly rainfall extremes in Germany, Hydrology Research, 2021.

Haberlandt, U. and Berndt, C.: The value of weather radar data for the estimation of design storms – an analysis for the Hannover region, PROC. IAHS, 373, 81-85, 2016.

430   Junghänel, T., Ertel, H., and Deutschländer, T.: KOSTRA-DWD-2010R - Bericht zur Revision der koordinierten Starkregenregionalisierung und -auswertung des Deutschen Wetterdienstes in der Version 2010, 2017.

Kaiser, M., Günnemann, S., and Disse, M.: Spatiotemporal analysis of heavy rain-induced flood occurrences in Germany using a novel event database approach, Journal of Hydrology, 595, 125985, 2021.

Kreklow, J., Tetzlaff, B., Burkhard, B., and Kuhnt, G.: Radar-Based Precipitation Climatology in Germany—Developments,

435   Uncertainties and Potentials, Atmosphere, 11, 217, 2020.

Lengfeld, K., Winterrath, T., Junghänel, T., Hafer, M., and Becker, A.: Characteristic spatial extent of hourly and daily precipitation events in Germany derived from 16 years of radar data, Meteorologische Zeitschrift, 28, 363-378, 2019.

Lengfeld, K., Kirstetter, P.-E., Fowler, H. J., Yu, J., Becker, A., Flamig, Z., and Gourley, J.: Use of radar data for characterizing extreme precipitation at fine scales and short durations, Environmental Research Letter, 15, 085003, 2020.

440   LUBW: Leitfaden Kommunales Starkregenrisikomanagement in Baden-Württemberg, 2016.

Maraun, D.: Bias Correcting Climate Change Simulations - a Critical Review, Current Climate Change Reports, 2, 211–220-, 2016.

Overeem, A., Buishand, T. A., and Holleman, I.: Extreme rainfall analysis and estimation of depth-duration-frequency curves using weather radar, Water Resources Research, 45, 2009.

445   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, Hydrology and Earth System Sciences, 20, 2317–2332, 2016.

Pöschmann, J. M., Kim, D., Kronenberg, R., and Bernhofer, C.: An analysis of temporal scaling behaviour of extreme rainfall in Germany based on radar precipitation QPE data, Natural Hazards and Earth System Sciences, 21, 1195–1207, 2021.

450 Rabiei, E. and Haberlandt, U.: Applying bias correction for merging rain gauge and radar data, Journal of Hydrology, 522, 544–557, 2015.

Requena, A. I., Burn, D. H., and Coulibaly, P.: Pooled frequency analysis for intensity–duration–frequency curve estimation, Hydrological Processes, 33, 2080-2094, 2019.

Ruiz-Villanueva, V., Borga, M., Zoccatelli, D., Marchi, L., Gaume, E., and Ehret, U.: Extreme flood response to short-duration

455 convective rainfall in South-West Germany, Hydrology and Earth System Sciences, 16, 1543–1559, 2012.

Saltikoff, E., Friedrich, K., Soderholm, J., Lengfeld, K., Nelson, B., Becker, A., Hollmann, R., Urban, B., Heistermann, M., and Tassone, C.: An Overview of Using Weather Radar for Climatological Studies: Successes, Challenges, and Potential, Bulletin of the American Meteorological Society, 100, 1739-1752, 2019.

Schleiss, M., Olsson, J., Berg, P., Niemi, T., Kokkonen, T., Thorndahl, S., Nielsen, R., Nielsen, J. E., Bozhinova, D., and

460 Pulkkinen, S.: The accuracy of weather radar in heavy rain: a comparative study for Denmark, the Netherlands, Finland and Sweden, Hydrology and Earth System Sciences, 24, 3157–3188, 2020.

Steinbrich, A., Stölzle, M., and Weiler, M.: Generierung von konsistenten Grundlagendaten zur Berechnung von Starkregenereignissen für eine Starkregengefahrenkartierung in Baden-Württemberg. Projektbericht an die LUBW, not published, 2016.

465 Thorndahl, S., Nielsen, J. E., and Rasmussen, M. R.: Bias adjustment and advection interpolation of long-term high resolution radar rainfall series, Journal of Hydrology, 508, 214-226, 2014.

Uboldi, F., Sulis, A. N., Lussana, C., Cislaghi, M., and Russo, M.: A spatial bootstrap technique for parameter estimation of rainfall annual maxima distribution, Hydrology and Earth System Sciences, 18, 981-995, 2014.

Winterrath, T., Brendel, C., Hafer, M., Junghänel, T., Klameth, A., Lengfeld, K., Walawender, E., Weigl, E., and Becker, A.:

470 RADKLIM Version 2017.002: Reprocessed quasi gauge-adjusted radar data, 5-minute precipitation sums (YW), 2018.

Winterrath, T., Brendel, C., Hafer, M., Junghänel, T., Klameth, A., Walawender, E., Weigl, E., and Becker, A.: Erstellung einer radargestützten Niederschlagsklimatologie, 2017.

Zorzetto, E., Botter, G., and Marani, M.: On the emergence of rainfall extremes from ordinary events, Geophysical Research Letters, 43,8076-8082, 2016.

475

Figures

480

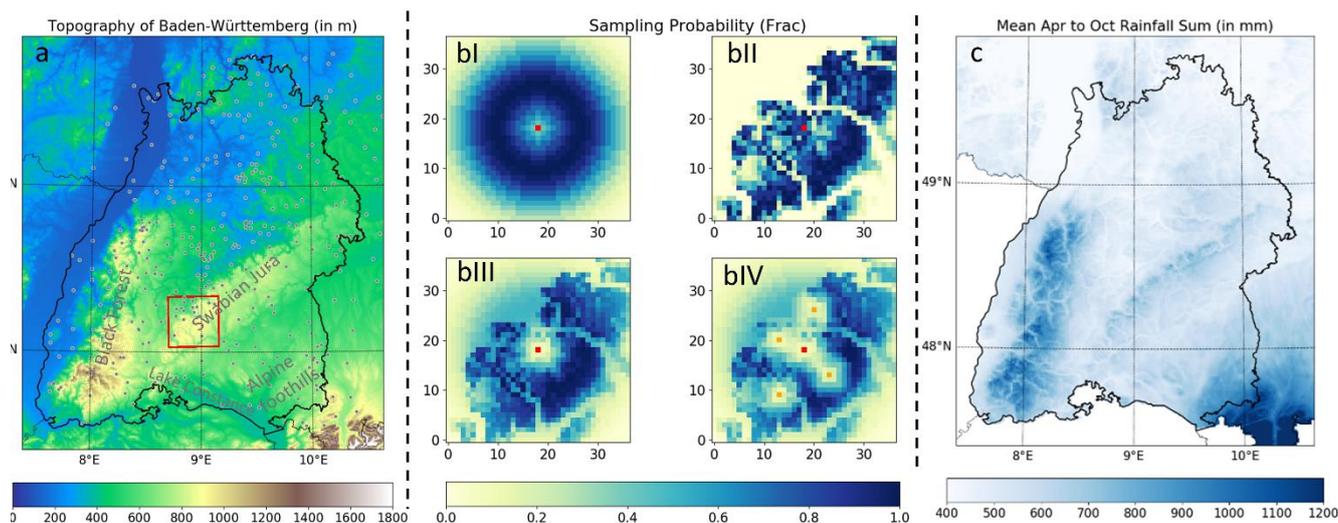
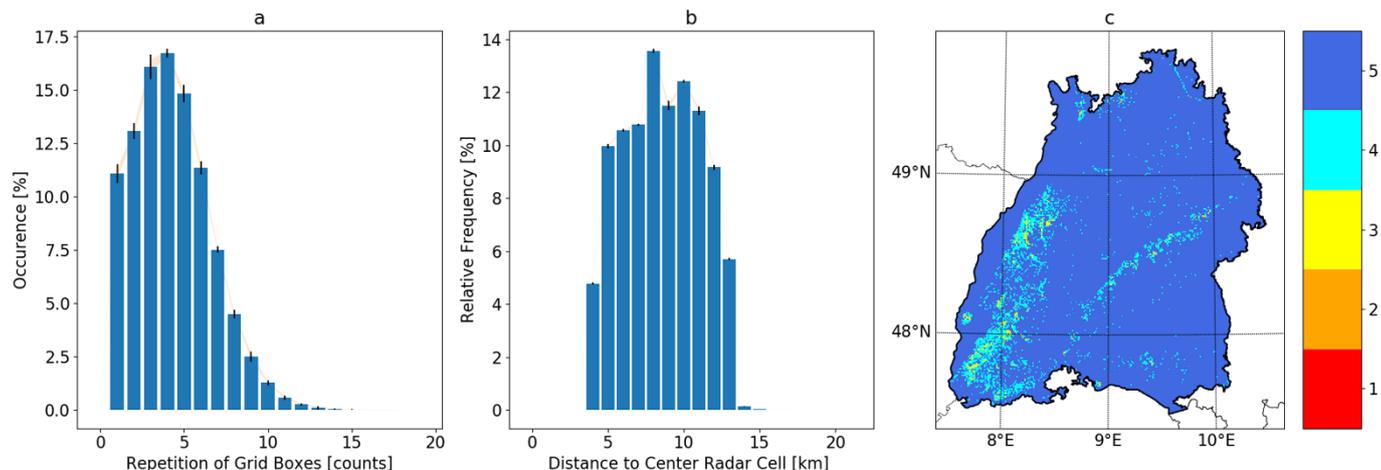



Figure 1: (a): Topography of Baden Württemberg (BaWu) as well as location of the precipitation gauges used in the BW-Stat dataset and some of the geographical regions referred to in the text. (b): Probability for a specific radar cell to be sampled based on the distance to cell of interest(bI), orography (bII) and orography and distance combined (bIII). Final sampled cells (orange) and
485 **reduced probabilities around the selected cells are depicted in panel bIV. All panels reflect the area indicated with a red square in the left part of the figure. The respective cell of interest is marked in red. (c): April to October rainfall sum (1991 to 2020) of the REGNIE (Regionalisierte Niederschlagshöhen) dataset compiled by the German Weather Service.**

490 **Figure 2: Distribution of the Fraction of (a) cumulative occurrences of individual radar cells in the final sample; (b) the distance of the sampled radar cells to the cell of interest. (c) Spatial distribution of the effective ensemble size. The distributions in a and b are for a single ensemble member, while the error bars indicate the rather small variation among the five ensemble members.**

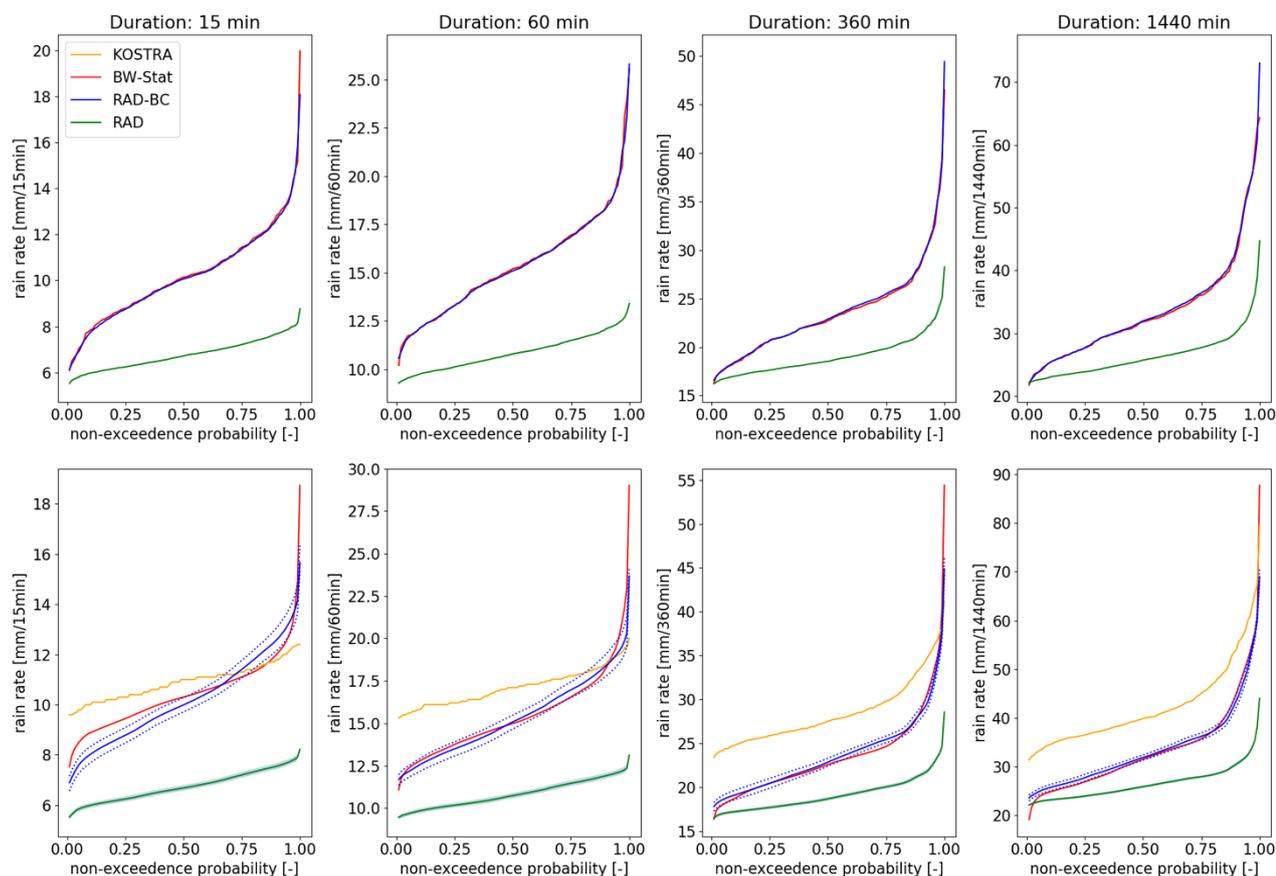



Figure 3: Cumulative frequency distributions (CFD) of the location parameter for four different event durations when comparing stations and radar data at the location of stations only (upper row) and integrated over the whole of BaWu (bottom row). The dotted blue lines in the bottom row represent the range of the five ensemble members (sampling uncertainty only, no bootstrapping).

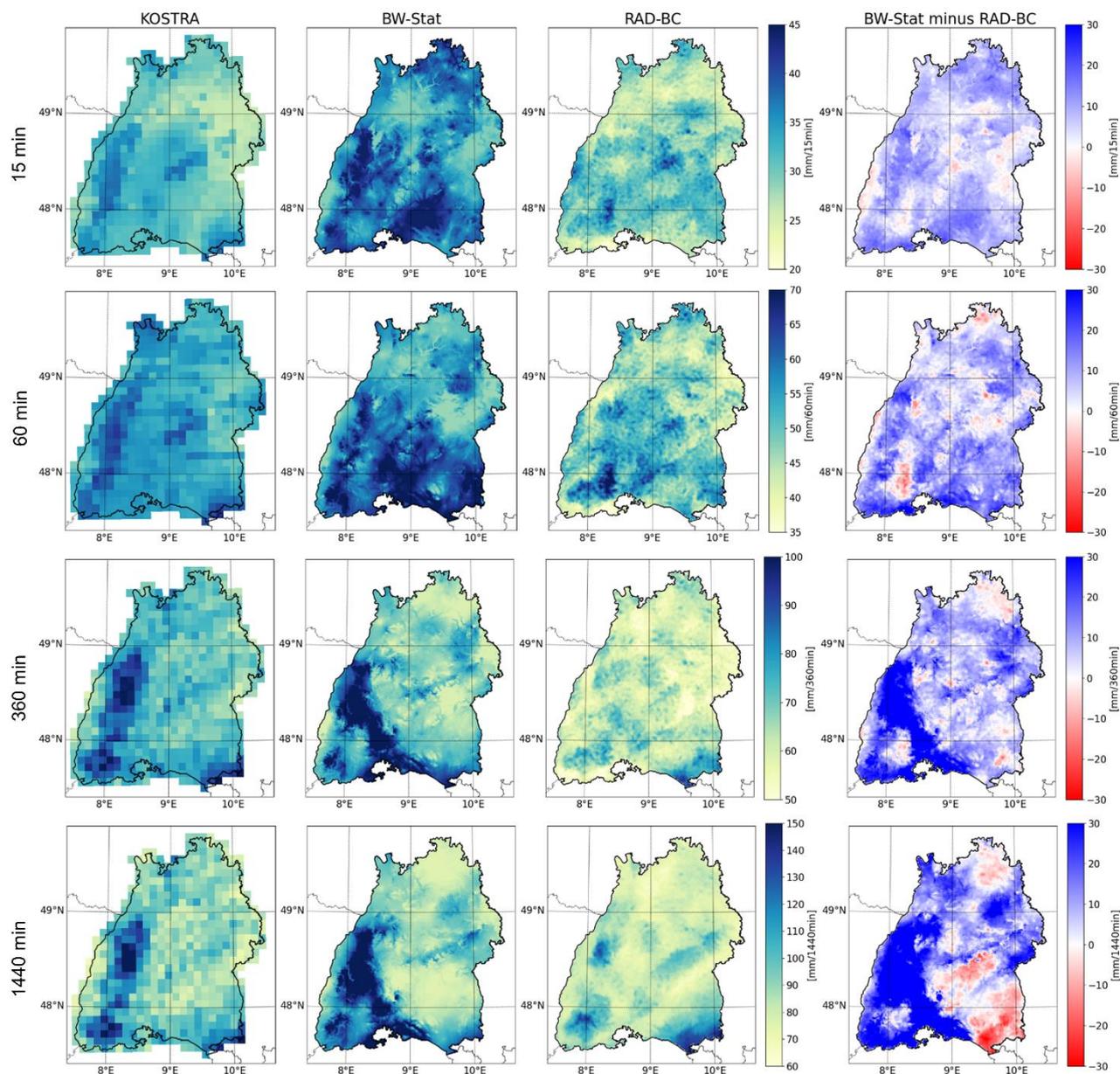



Figure 4: Magnitude of design storms with a return rate of 100 years for four different event durations (15, 60, 360 and 1440 minutes, depicted in rows) and three different datasets (KOSTRA, BW-Stat, RAD-BC, depicted in columns). Additionally, the difference between BW-Stat and RAD-BC is depicted (right column). Note that in the BW-Stat dataset all values below/above the 5th/95th percentiles are set to the respective percentile value.

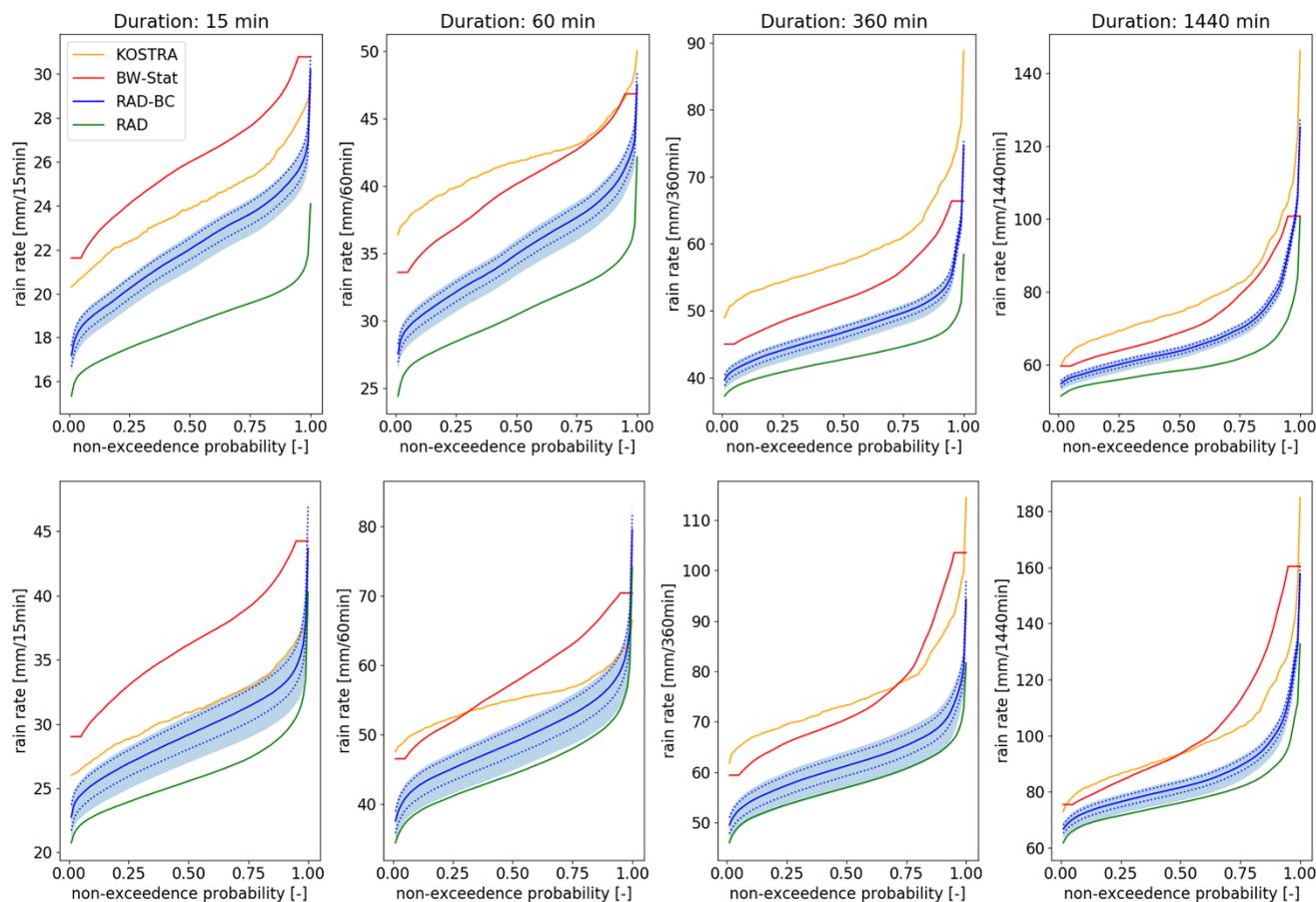



Figure 5: Cumulative frequency distributions (CFD) of the magnitude of 20yr (upper row) and 100yr (bottom row) design storms for four different event durations and different datasets. The shaded range depicts the ensemble uncertainty (10th and 90th percentile of the range from the 1000 bootstraps for each of the 5 ensemble members). The dotted blue lines in the bottom row represent the range of the five ensemble members (sampling uncertainty, no bootstrapping) only. Note that in the BW-Stat dataset all values below/above the 5th/95th percentiles are set to the respective percentile value.

510

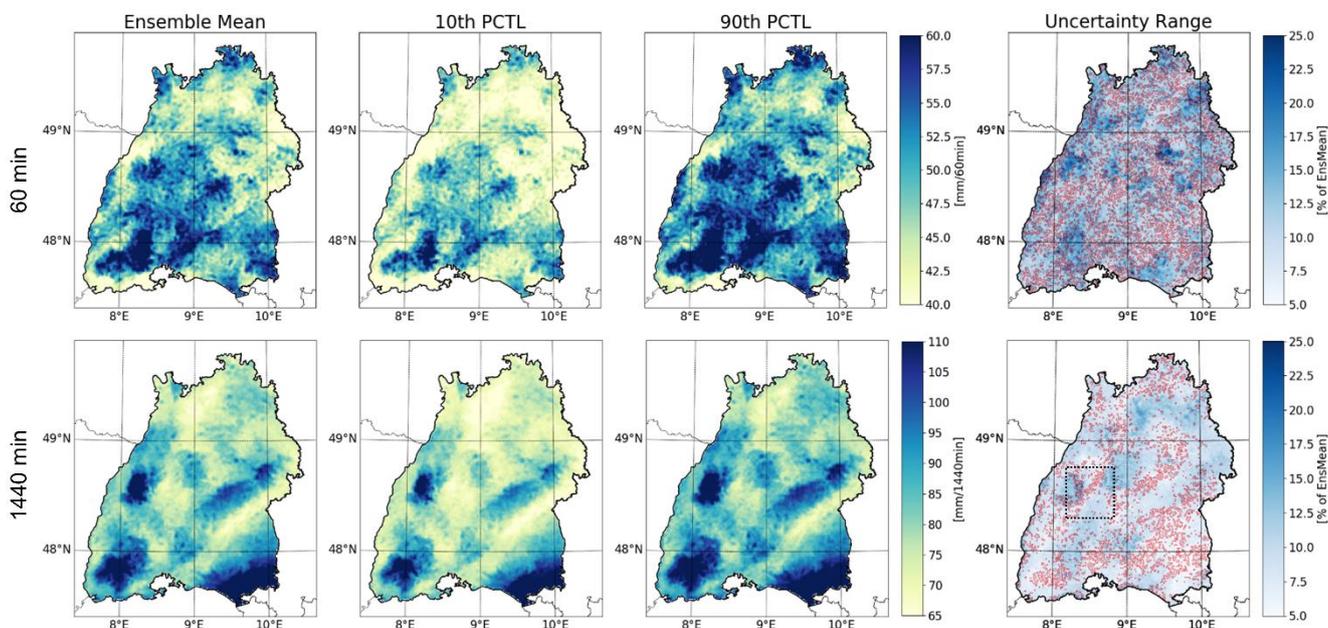



Figure 6: Ensemble mean (left column) and the 10th and 90th percentiles (two middle columns) of a 100yr design storm based for
515 **two durations (60 minute events – upper row; 1440 minute events – bottom row). Additionally, the ensemble uncertainty range (difference between the 90th and the 10th percentile of the full (bootstrapping & sampling) 5000 member) is depicted (right column). Regions with a large (> 75% of the range) contribution of the sampling uncertainty are marked with red. The black dashed square in the panel in the lower right defines the northern Black Forest region discussed in the text.**

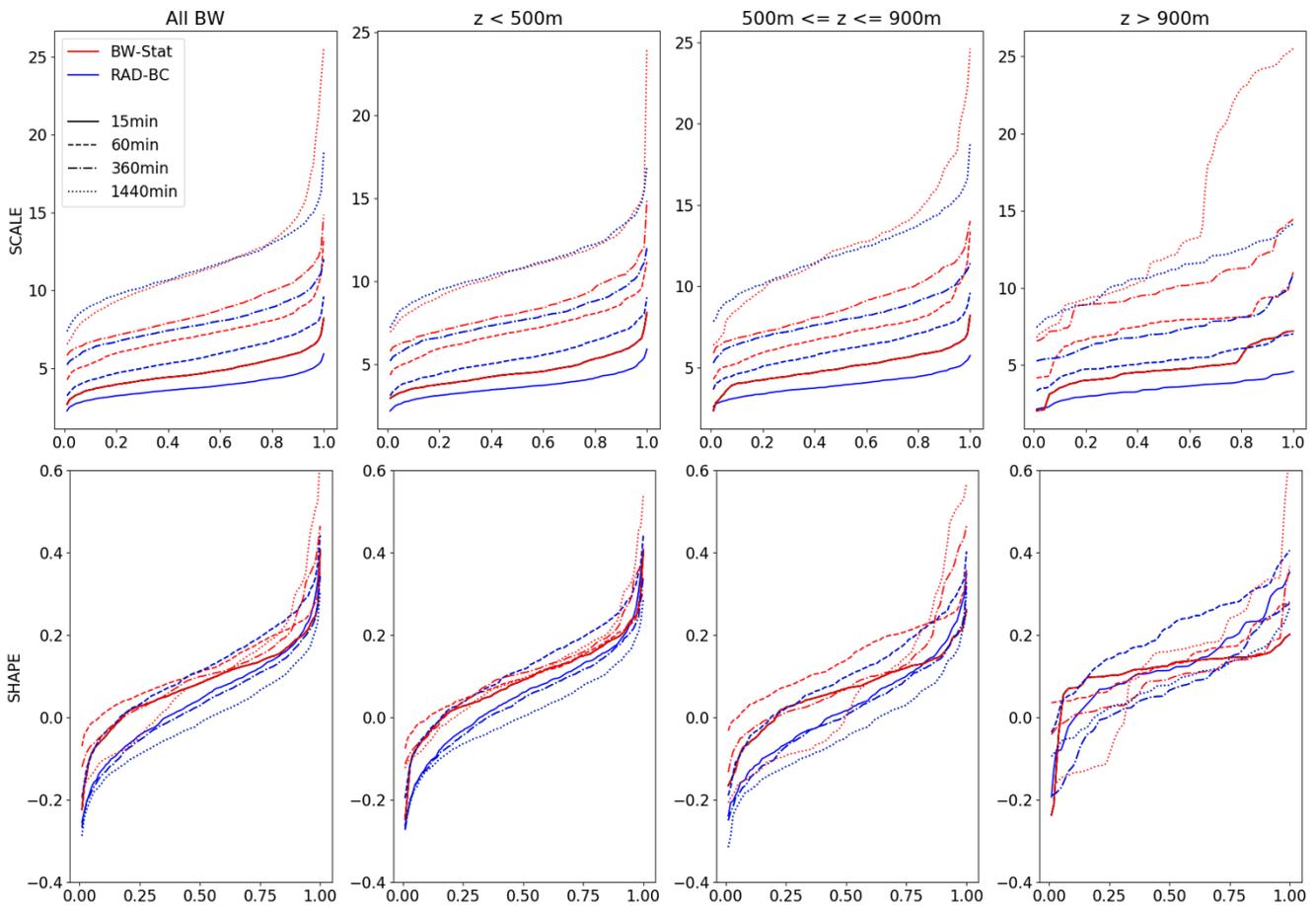



Figure 7: Cumulative frequency distributions (CFD) of the scale (upper row) and shape (bottom row) parameter for the BW-Stat and RAD-BC datasets, when comparing stations and radar data at the location of all stations (left column) and for three different subsets filtered by the altitude of the respective station locations.

Appendix

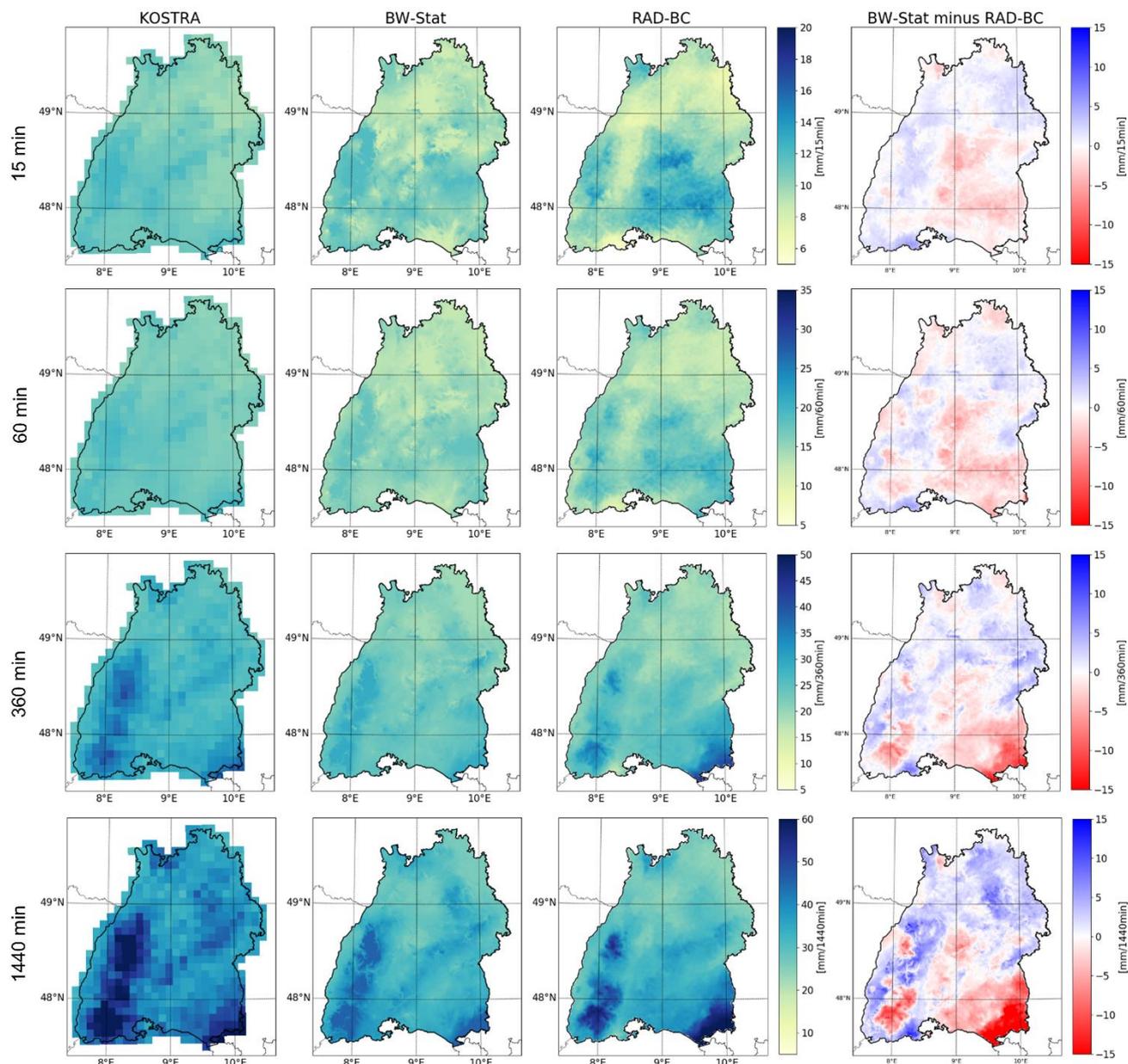



Figure A1: Magnitude of 1yr design storms for four different event durations (15, 60, 360 and 1440 minutes, depicted in rows) and three different datasets (KOSTRA, BW-Stat, RAD-BC, depicted in columns). Additionally, the difference between BW-Stat and RAD-BC is depicted (right column). Note that in the BW-Stat dataset all values below/above the 5th/95th percentiles are set to the respective percentile value.

540

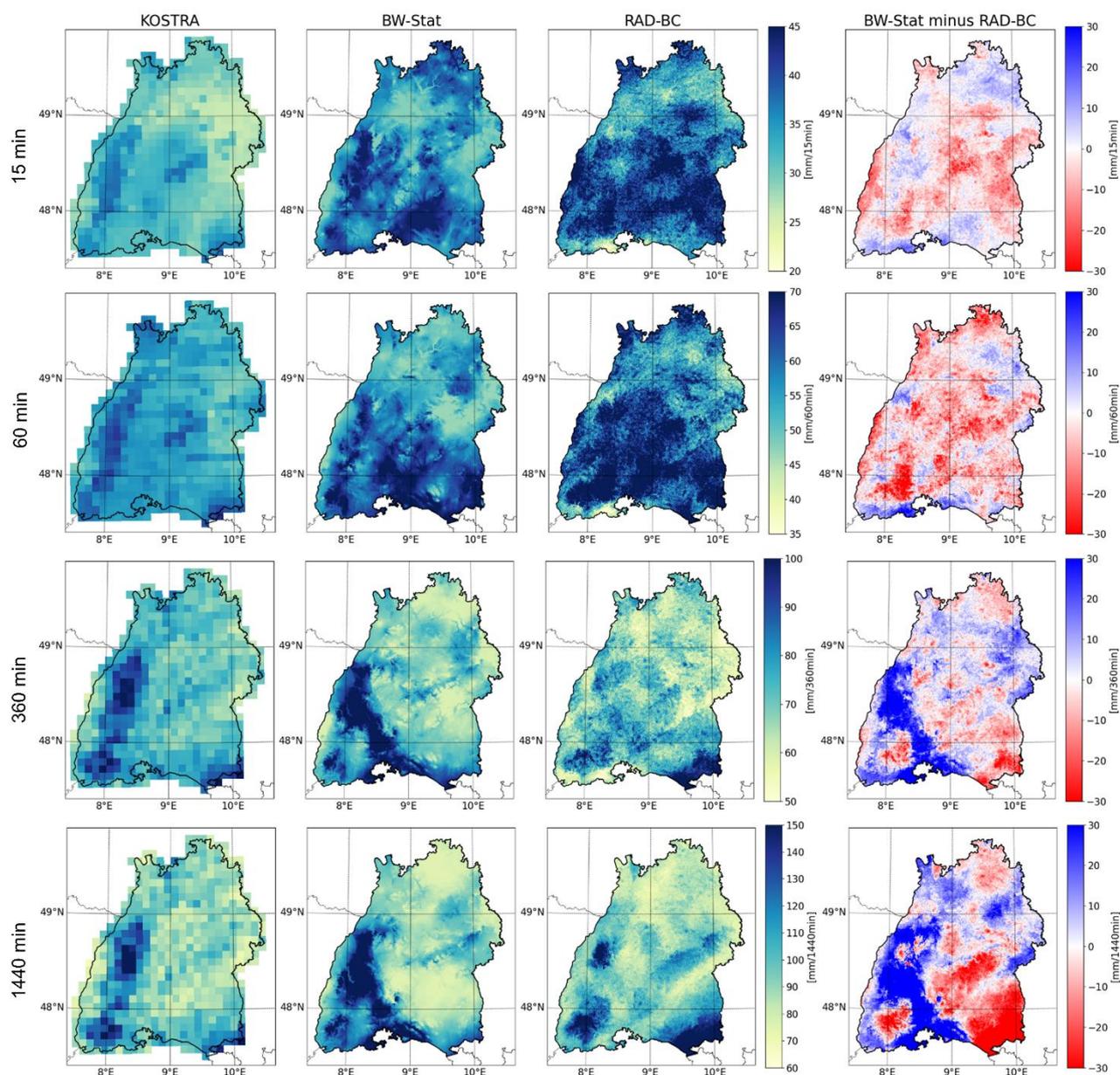



Figure A2: Magnitude of design storms with a return rate of 100 years for four different event durations (15, 60, 360 and 1440 minutes, depicted in rows) and three different datasets (KOSTRA, BW-Stat, RAD-BC, depicted in columns). Additionally, the difference between BW-Stat and RAD-BC is depicted (right column). In contrast to Fig. 4, RAD-BC is correct using a multiplicative
545 **bias correction approach. Note that in the BW-Stat dataset all values below/above the 5th/95th percentiles are set to the respective percentile value.**