Technical Note: Space-Time Statistical Quality Control of Extreme Precipitation Observations

Abbas El Hachem[1], Jochen Seidel[1], Florian Imbery[2], Thomas Junghänel[2], and András Bárdossy[1]

[1]Institute for Modelling Hydraulic and Environmental Systems, University of Stuttgart, D-70569 Stuttgart, Germany
[2]Deutscher Wetterdienst, Offenbach, Germany

Correspondence: Abbas El Hachem (abbas.el-hachem@iws.uni-stuttgart.de)

Abstract. Information about precipitation extremes is of vital importance for many hydrological planning and design purposes. However, due to various sources of error, some of the observed extremes may be inaccurate or false.~~Precipitation extremes form the basis of many engineering design decisions. Extremes are rare events which may differ strongly from "normal" observations. Unfortunately some of the observed extremes may be inaccurate or false.~~ The purpose of this investigation is to present a quality control~~-check~~ of observed extremes using space-time statistical methods. To cope with the highly skewed rainfall distribution a Box-Cox transformation with suitable parameter was used. ~~As a first step the biggest values for each observation location and event duration are selected. For each of these, the observed values of all other stations corresponding to the same time steps are collected and transformed using a Box-Cox transformation, which factor was derived from fitted truncated normal distribution.~~ The value at the location of a potential outlier is estimated using the surrounding stations and the calculated spatial variogram and compared to the suspicious observation~~The value at the extreme location is estimated using the surrounding stations and the calculated spatial variogram, and this estimated value is compared to the observed extreme~~. If the difference exceeds the threshold~~critical value~~ of the test, the value~~extreme~~ is flagged as ~~a~~ possible outlier. The same procedure is repeated for different temporal aggregations in order to avoid singularities caused by convection. ~~The flagged extremes are then compared to the extremes of the surrounding stations using the same procedure – interpolation and subsequent comparison of the interpolated and the observed values.~~ Detected outliers~~Flagged extremes~~ are subsequently compared to the corresponding radar and discharge observations and finally, implausible extremes are removed. The procedure is demonstrated using observations of sub-daily and daily temporal resolution in Germany.

1 Introduction

A clear definition of an outlier might be intuitive to many but it has been formulated differently by several researchers. In the work of Barnett and Lewis (1994) an outlier was defined as an observation showing an inconsistent behavior compared to other data values. Hawkins (1980) described an outlier as being an observation that differs substantially from other observations as if it might have been produced by an alternating mechanism. More precisely, for Iglewicz and Hoaglin (1993) an outlier is an observation that arouses suspicion to the analyst and does not belong to the same data distribution. ~~Johnson et al. (2017) defined~~ In general, there are two types of outliers, ~~namely~~ those associated with an error and those associated with a real observation. The reasons for an observation being erroneous could ~~either~~ be due to instrumental errors (e.g. use of false instrument, equipment malfunction, false equipment operation) or/and human errors (false reading or recording or even computation of observations). Moreover, errors can occur if the measuring site is falsely chosen, providing a false representativeness of the observed process.

Hydroclimatological data are of unique nature as they occur in a non repetitive manner. If an observation is not registered correctly reconstructing it~~such a measurement~~ is very challenging, especially for precipitation values. ~~Due to the high spatial and temporal variability of such events, surrounding rain gauges can often not be used for outlier detection or data plausibility checks~~ However, reliable information about precipitation extremes are essential for many design purposes such as flood analysis, extreme value statistics and stationarity analysis to name just a few. ~~The presence of outliers in a data set can lead to under- or overestimation of design values.~~

~~Precipitation observations have a space-time dimension. Observations are taken at different locations in space and in discrete time intervals. Some precipitation events occur over a local spatial scale with short duration and high intensity. These events are outliers defined as 'single' events. They were correctly observed but strongly differ from surrounding observations. These observations play, for example, a major role in urban hydrology.~~

~~Even in normal conditions, false observations can occur especially when considering high-resolution temporal data (1-minute frequency).~~ Many quality control (QC) algorithms have been developed and are being used by weather service agencies to minimize and detect false measurements. Durre et al. (2010) established a comprehensive QC algorithm for daily surface meteorological observations (temperature, precipitation, snowfall~~snow fall~~ and snow depth). For precipitation data, the QC method for detecting false observations consisted of several steps. ~~from which are~~ A climatological outlier check is used for flagging values exceeding a certain temperature-dependent threshold and a spatial consistency check based on comparing the ~~by which the~~ target observation ~~is compared~~ to neighboring ones. An observation is eventually flagged if the difference exceeds a certain climatological percent ranks threshold. Qi et al. (2016) implemented a QC algorithm to identify erroneous hourly rain gauge observations by using additional information as radar quantitative precipitation estimates (QPE). A common practice ~~approach~~ for detecting outliers is to use an interpolation method to estimate the probability of a local ~~suspected~~ observation using the surrounding locations. If this probability is very low then the observation is suspicious. This was mathematically formulated by Ingleby and Lorenc (1993). ~~When only comparing the observed to the estimated value, many observations might not be correctly captured depending on the event spatial extent.~~ In the work done by Hubbard et al. (2005) a QC method was developed for daily temperature and precipitation values consisting of four steps. Observations are flagged if they do not fall within $\pm$ 3 standard deviation of the long-term mean and if they differ from the estimated value using a spatial regression technique. Some other QC methods are available but are often limited to time series analysis and tend to disregard the temporal-spatial extent of precipitation.

Precipitation observations have a space-time dimension. Observations are taken at different locations in space and in discrete time intervals. ~~Moreover,~~ Due to the presence of non-negative and many no-precipitation (zero) values, precipitation data (especially at daily and sub-daily resolution) have a positively non-normal skewed distribution with heavy tails (Klemeš, 2000) and fall under the zero-inflated data. Therefore, an adequate transformation of the data should be performed to reduce the effect of the data skewness. A relatively simple approach to normalizing~~normalize~~ a variable is the~~to apply a~~ Box-Cox transformation (Box and Cox, 1964). ~~To cope with the positive nature of precipitation a transformation to a truncated normal distribution can be used.~~

The following work proposes a statistical space-time methodology based on interpolation in a cross-validation mode to find possible outliers in ~~the~~ precipitation ~~data~~ observations across several temporal aggregations. An outlier is defined here as an observation that strongly differs for a certain temporal aggregation from its~~it's~~ spatial neighboring locations. A difficult task while working with outliers in general and especially in hydrology is distinguishing between correct and false observations. Therefore, to validate detected outliers, the suspected values are additionally compared to independent~~external~~ information such as discharge and radar measurements.

This paper is organized as follows: after the introduction, the data and methodology ~~to find possible outliers in the data~~ are presented. Afterward, the results of the outlier detection~~the quality control procedure~~ are presented and four~~two~~ examples of verification via subsequent comparison to radar or discharge data are shown. In the final section of the results, the number of identified outliers for every year (and month) and for each location is mapped and presented.~~The paper ends with a discussion and a conclusion.~~

2 Study Area and Data

This study was done using the German-wide precipitation data set from the German Weather Service DWD which covers an area of approximately 357.000 km^2. The average annual rainfall in Germany is around 800 mm and can reach up to 2100 mm in the higher elevations of the Alpes in the South. Currently, the DWD operates a network of rain gauges with different temporal resolutions ranging from minutely to daily. Hourly and sub-hourly data are available from the 1990s onwards. The number of these stations has been continuously increasing since then. ~~As for the daily data, some stations are available since the year 1900.~~ On the other hand,~~After the year 1990,~~ the number of stations with daily observations started to decrease since then as~~while~~ they were replaced by automatic rain gauges. Rain gauges near the border (30-kilometer inland buffer) ~~without neighbours from other countries~~ were not included in this analysis. ~~A 30-kilometer inland buffer was used to select all rain gauges that are located within this region. This is relevant for the used methodology relying on neighbouring observations.~~

~~Figure ?? shows the change of the number of stations in Germany and the corresponding distribution of available observation years for daily and sub-hourly station data.~~

~~In scope of this study, data have been collected for most available temporal resolutions (1 minute, 1 hour and 1 day) and time periods.~~ In the 1990s, most DWD rain gauges were tipping buckets or drop counters. From 2000 onwards, these were replaced by weighing gauges (OTT Pluvio) and since 2017 these are being replaced by combined tipping bucket and weighing rain gauges (Lambrecht rain[e]).

Precipitation data from the recent DWD observation network go through several quality control steps. The first step is a quality control directly at the automatic stations. Since this is an automatic test, relatively wide thresholds are applied. It includes tests for completeness, thresholds, temporal and internal consistency. Based on these tests, a quality flag is assigned to the data. The data is then submitted to a database. Another test with tighter thresholds is then performed, based on the QualiMet software (Spengler, 2002). This phase of the quality check also tests for completeness as well as climatological, temporal, spatial, and internal consistency. Questionable values are manually checked and corrected and~~respectively~~ the quality label is adjusted. A final quality check step occurs after all of a month's data are available, focusing on aggregate values. The quality flags are stored in the database and are also made available to users. ~~, e.g., when the data are made available on the Internet.~~ DWD quality assurance also includes the identification and correction or description of errors in the historical data (Kaspar et al., 2013). Appropriate procedures have been developed~~implemented~~ for the quality control of historical data. ~~and are applied to the data. The focus is on the daily values.~~ In general, the quality of these values can be considered reasonably good, but there are still doubtful values on the order of about 0.1-1%, especially for the pre-1979 data. The user must keep in mind that the data can be affected by certain non-climatic effects, such as changes in instrumentation or observation time. With few exceptions, the data are reported "as observed", i.e., no homogenization procedure was applied.

As independent data for verification~~verification data~~, radar derived rainfall QPE~~rainfall derived radar images~~ and discharge observations from the state of Bavaria were used. The radar data used is the product RADOLAN-RW that is provided by the DWD in hourly and daily resolutions starting the year 2005 (DWD Climate Data Center (CDC), 2021). These products have been gauge-adjusted with the observed hourly station data. The occurrence (or absence) of precipitation observation in the radar data over the target location is an indication to the quality of the observation. The discharge data were quality checked and provided by the environmental agency of Bavaria with hourly and daily resolutions for approximately 400 gauges within the region of Bavaria (Bayerisches Landesamt für Umwelt, 2022). Different headwater catchments were derived and selected for validating the results. A reaction (within few hours) in the headwater catchment discharge is expected after the event occurred in case of correct rainfall observations.~~The aim is to identify if a reaction in the discharge was noted after the event or not.~~

Figure 1 illustrates the location of the ~~sub-~~daily and sub-daily rain gauges as well as their spatial density. For the daily data all available locations (historical and present) are displayed. ~~This is because all available data were investigated.~~ The stations do not have a homogeneous spatial distribution over the country where some locations have a higher network density than others. The spatial density was calculated using a kernel density estimation (KDE) with a Quartic shape and a radius of influence of 30 kilometers. The estimated density value depends on the separating distance between the known and unknown locations and the kernel parameters. These are the bandwidth (h) which is reflected by the radius of influence and the weighting function or kernel function (K). The latter defines the contribution of each point as a function of the separating distance. Further detail regarding KDE estimation can be found in Yu et al. (2015). ~~The stations do not have a homogeneous spatial distribution over the country some locations have a higher network density than others. Figure **??** describes the number of available station over time with the corresponding available data duration.~~

~~Figure**??**~~

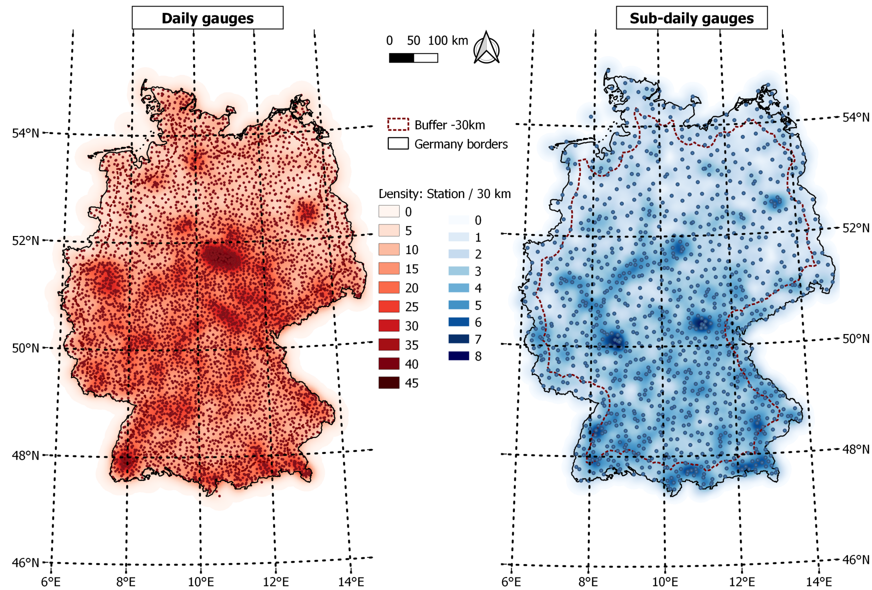

Figure 1. Map of the study area showing the location and density (number of stations per 30 km) of the DWD gauges with daily (left) and sub-daily (right) resolutions.

3 Methodology

3.1 Data transformation

As an initial step, a Box-Cox transformation equation 1 was applied for every variable X and temporal aggregation t to reduce the effect of the skewed precipitation distribution (Box and Cox, 1964).

$$X_t^* = \begin{cases} \frac{(X_t^\lambda - 1)}{\lambda} & \text{if } \lambda \neq 0 \\ \log(X_t) & \text{if } \lambda = 0 \end{cases} \tag{1}$$

Where:

X^* = transformed precipitation data at location u and temporal aggregation t

X = original precipitation data at location u and temporal aggregation t

λ = transformation factor for temporal aggregation t

To find which transformation factor λ is most suitable, several simulated lower truncated standard normal distribution (sampling space bounded by $[-\infty < a = p_0,\ b = +\infty]$) ~~(truncated at p_0)~~were fitted to the original data (Burkardt, 2014). The probability of having a value above or below p_0 is then derived (p_0 probability of having 0 mm precipitation value) .

~~equation p0~~

From this probability (denoted p_{norm}) a new standard normal distribution is generated where $(x < p_{norm} = 0,\ x >= p_{norm} = x)$. From this distribution the skewness γ_{norm} is calculated. The goal now is to find which transformation factor minimizes the difference between the original data skewness and γ_{norm}. This was done for each station separately and for all aggregations. Eventually an average transformation factor (denoted hereafter λ) was derived for each temporal aggregation. The results of this procedure can be seen in Table 1.

~~Figure 2~~

Table 1. Average transformation factor λ used to transform the original data to the truncated normal space with reduced skewness.

	60 min	**12 0 min**	**180 min**	**240 min**	**360 min**	**720 min**	**1440 min**
Average transformation factor λ	0.097	0.155	0.219	0.262	0.318	0.427	0.499

Once λ was calculated, the original precipitation data were~~was~~ transformed as in equation 1, and in the ~~the~~ newly truncated normalized space the following approach was implemented to find outliers in the precipitation data over several temporal resolutions.

~~equation Z*~~

3.2 Outlier detection

The proposed method was initially tested for identifying outliers in groundwater quality data (Bárdossy and Kundzewicz, 1990). In this paper, a similar method was implemented to identify unusual precipitation data and is extended by a validation of the results using external information such as radar or discharge observations. For detecting ~~precipitation records that are~~ possible outliers the concept of jackknifing is used, a method initially developed by Quenouille (1949, 1956). The main idea is based on removing one (or each) observation from the data and estimating its value again. In this study, ~~for checking the quality of intense precipitation values,~~ the four largest ~~four~~ annual observations for every station are compared to the estimated values at the same location. Each cross-validated value is estimated using the nearest 30 neighboring locations with valid observations. ~~For example, if a station has 5 years of data, 20 events were investigated (for every aggregation).~~ Since many possible faulty observations can only be detected ~~if they are inspected~~ at lower temporal resolution, the procedure was applied over several temporal aggregations. ~~and on not on the temporal scale at which they were observed.~~ For example, when looking at sub-daily and sub-hourly values a single observation might not be unusual but the accumulation of many values reveals suspicious sums. Furthermore, single events might occur on high temporal scales (e.g. hourly) and are not detected on lower aggregations (e.g. daily).

For estimating the target value Ordinary Kriging (OK) is used as an interpolation technique. It is a regionalization method initially introduced by ~~the french mathematician Matheron the statistician Krige~~ Krige (1951) and Matheron (1962) to estimate an unknown value at a target location by solving a linear equation system by minimizing the estimation variance and maximizing the accuracy (no systematic error). Each cross-validated value is estimated using the nearest 30 neighboring locations with valid observations. The spatial correlation structure is reflected by the variogram which is derived in the rank space domain and rescaled to the variance of the data. This allows a for variogram calculation in a more robust manner (Lebrenz and Bárdossy, 2019). The target location is calculated by solving the kriging equation and the estimation variance is noted. For identifying unusual observations the ratio between the absolute value of the difference between the observed and the estimated values and the estimation variance is calculated. This Criteria Ratio, ~~denoted hereafter as~~ (CR) describes the relative agreement/disagreement between the observed value and the spatial surroundings for the corresponding time step. Larger CR values reflect high spatial-temporal disagreement and low values denote greater agreement. Based on the CR value, different types of events can be identified, namely those occurring on a local scale with high CR values and other on a regional scale with low CR values. ~~As in~~Following Bárdossy and Kundzewicz (1990) a CR value of three is initially used to identify suspicious observations. The CR value is derived for every cross-validated event. Eventually the CR value is related to all of the observed (interpolated) data establishing a possibility to find a suitable CR value for identification of precipitation outliers.

$$CR_i(u) = \frac{|Z_i^*(u) - Z_i(u)|}{\sigma_i(u)} \tag{2}$$

Where:

$Z_i^*(u)$ =	estimated value at location u and timestep i
$Z_i(u)$ =	observed value at location u and timestep i
σ_i =	kriging standard deviation at location u and timestep i

Since precipitation events occurring on a local scale might represent ~~a single event and not a false one,~~an actual small scale event to validate the first or the second case, the suspicious events are compared to the observed radar QPE or discharge values in the corresponding catchment. Despite having their own drawbacks the radar and discharge observations are used here as a qualitative decision support tool. ~~As a final step, the time series before and after the occurrence of the observed value and its neighbors is investigated. If the observed value is due to a 'single' peak or many 'continuous' small peaks the possibility of being an outlier increases. Some outliers can be easily identified as erroneous values others are more challenging, this is where the additional information is valuable.~~

The flowchart in figure 2 describes ~~Following is~~ the implemented space-time precipitation outlier detection scheme.

~~for every station select the largest 4 yearly values~~

~~for every selected observation~~

~~transform the target and surrounding locations using equation ??~~

~~calculate the estimated value at this point using the surrounding stations and their spatial structure~~

~~for the estimated value find the corresponding estimation variance (or standard deviation)~~

~~calculate the CR value~~

~~compare the CR value to the~~

~~find all selected observations that have high CR values and are within the upper 1% quantile~~

~~compare these events to the corresponding radar image or discharge values~~

~~repeat the procedure for different aggregations~~

~~find events that are suspicious for single or several aggregations~~

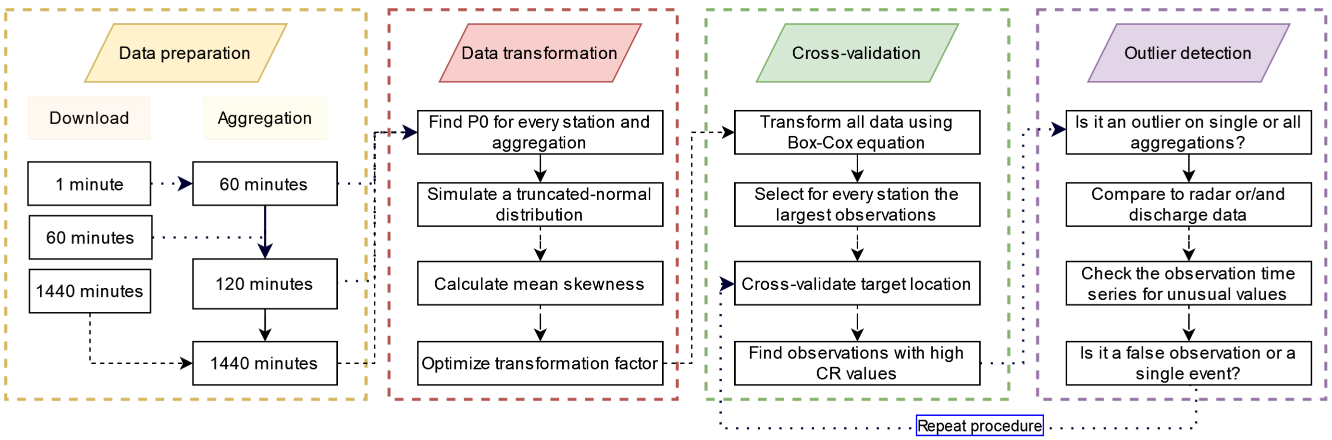

Figure 2. Flowchart summarizing the described method starting with the data download procedure and ending with the identification of suspicious observations.

~~For interpolating the selected events Ordinary Kriging (OK) is used as an estimation technique. The main concept behind Geostatistics is the consideration of the data as spatially dependent random numbers with a variance that increases with increasing separation distance. The observed data at the corresponding location u are seen as a realization of the regionalized variable of the random space function. Since for every location u in the domain D there is many (infinite) random variables $z(u)$ of the random variable $Z(u)$ describing each $Z(u)$ using it's own distribution function $F_Z(u)$ using the available single realizations is practically impossible. For simplifying the problem different hypotheses are considered. The first hypotheses which is a central one in Geostatistics is stationarity. Simply said, the whole domain D is represented by a single distribution function regardless of the location of the points u in D. A further simplification is introduced with the second-order stationarity. For this the expected value of the random function $E(Z(u))$ is constant over the domain and the covariance of two random variables corresponding to two locations u_i, u_j depends only on the separating vector $h = u_i - u_j$ between the two points. This means that the covariance depends on the spatial configuration of points and not their exact values. The second-order stationarity hypotheses requires that a covariance function exists. Since for a separation distance of $h = 0$ the covariance is same as the variance, the existence of a finite variance for D is required. This is often not the case in many natural processes (such as rainfall) where the variance increases with the distance. To solve this problem, the final hypothesis known as the intrinsic hypothesis was introduced. Same as the second-order stationarity the expected value is constant all over the domain D and the increment of the variance between two locations depends only on the separating vector h. The intrinsic hypothesis is a simplification of the second-order stationarity that is not constrained on the variance but on the variance of the increments.~~

~~The (semi-) variogram is defined by:~~

~~Properties of the variogram: The experimental variogram is derived from the observed values and their spatial distribution. A theoretical variogram model (spherical or exponential) was then fitted to the experimental one. The variogram reflect the change of correlation as a function of the separating distance between the considered values. Once the variogram was estimated OK could be performed. Note that for estimating the variogram the cross-validated location is not used.~~

~~For the case where all neighboring stations had a zero precipitation value, as theoretical spherical variogram with a range of 30 km was used.OK is a regionalization method to estimate an unknown value at a target location by solving a linear equation system through minimizing the estimation variance and maximising the accuracy (no systematic error). The estimation of the target location Z^* using the surrounding observations Z_i at the measurement locations n is defined by a linear estimation equation: The kriging estimation variance at the target location is formulated as: The weights λ_i are solved by guaranteeing the unbiased property of OK. Namely, the expected value of the estimation value should be equal to the expected value of the field Z. For this the Lagrange multiplier μ is introduced and the following linear system is solved: Note that γ refers to the (semi-) variogram which can also be replaced by the covariance function with the respected modifications. These two functions reflect the change of correlation as a function of the separating distance between the spatially distributed values. Note that for estimating the variogram the cross-validated location is not used.~~

~~Figure2~~

3.3 Data corruption

To test further the validity of the method, 20 stations without any detected outliers were randomly selected and their data (same events as before) were 'artificially' manipulated such that~~contaminated~~ the transformed observations of each target location were decreased and increased by several percentages (from 25~~5~~ to 100~~150~~ %) and the outlier detection method was tested. ~~The main idea here is to check if the method is able to identify such previously non-existent false values.~~ The results of this procedure can be seen in table 2. By decreasing the observed value until reaching a false zero observation the method was able to identify on the hourly scale around 60% and on the daily scale 94% of the cases as being outliers. On the other hand, by increasing the error value to up to 100%, almost all values were detected on all temporal aggregations. This emphasizes the validity of the method especially regarding identifying false high observations.

Table 2. Number of newly detected events after corrupting by different percentages the cross-validated observations of 20 randomly selected stations with no previous outliers.

	Temporal aggregation	60 min	120 min	180 min	240 min	360 min	720 min	1440 min
	Number of events	**150**	**150**	**150**	**150**	**150**	**150**	**150**
	Minimum of the minima [mm]	5.12	5.16	5.17	5.13	5.26	5.6	5.17
	Average of all averages [mm]	11.05	12.67	14.41	14.41	16.85	19.8	24.03
	Maximum of the maxima [mm]	51.2	50.1	53.47	63.93	71.92	73.6	76.37
Percentage of error	**-100 [%] (false zero)**	88	115	102	125	100	124	141
	-50 [%]	10	29	38	41	33	46	65
	-25 [%]	2	3	13	9	8	8	4
	0 [%]	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	+25 [%]	23	45	48	52	65	46	55
	+50 [%]	74	88	118	121	119	113	116
	+100 [%]	149	150	150	148	150	149	149

4 Results

~~Transformation to truncated normal~~

~~Table 1~~

~~Figure ??~~

4.1 Outliers vs single events

~~Based on the CR value, different events can be identified.~~ The center panel of figure 3 represents the CR value versus the ratio between the interpolated and observed values. All values denoted in red have a CR value above 3. This figure allows identifying the events that are of interest and relating the CR value to the interpolated and observed data. Note that the observed and interpolated values are in the original non-transformed space, only the CR values are calculated from the interpolation of the transformed values. ~~The suspected outliers (high CR values) are further inspected.~~

The values in the plot having a ratio of interpolated to observed of 5, are values obtained ~~by interpolating with the original values where~~ when a neighboring station (or stations) had simultaneously recorded an outlier (in this case a false high observation). This leads to detecting a false outlier. This can be accounted for by running the method again after all neighbours have been checked.

In the left and right panels of figure 3 the cumulative distribution function (CDF) from all investigated observations was calculated and the location of the detected outlier is~~are~~ marked. The events that were detected as being outliers spread over the curve showing that the method can detect not only high values but as well relatively small values that differ highly from their neighboring space. The left panel of figure 3 shows the results for the original hourly observations. The right panel shows those for the aggregated minutely observations. By comparing the two, the quality control procedure of the DWD can be investigated. ~~The minutely data are available without undergoing a thorough quality control as the hourly data.~~ Spatial consistency is checked more intensively by the DWD for higher aggregated precipitation data ($\geq$ 1 h) than for high temporal resolution data (e.g. 1 min). For example, in the hourly data, none of the largest values ($>$ 60 mm h^{-1}) is detected as an outlier and only one observation is larger than $>$ 80 mm h^{-1}. In the aggregated minutely data, many values above 80 mm h^{-1} exist and are mostly all detected as being suspicious. There are even values above 200 mm h^{-1} which can be caused by several faulty 'small' measurements or a few single large spikes in the data.

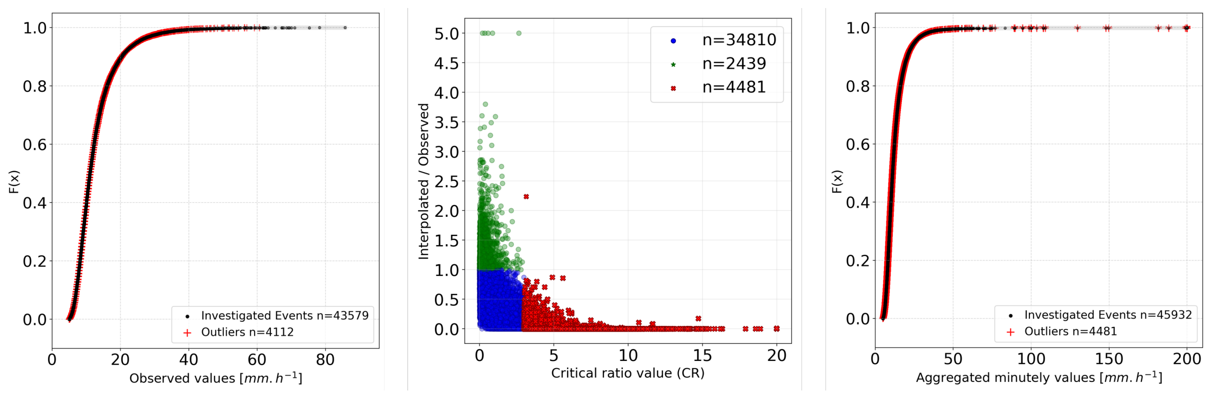

Figure 3. The left panel shows the CDF of all investigated hourly events with the detected outliers marked in red. The center panel shows for the minutely aggregated data the CR values versus the ratio of interpolated and observed hourly values ~~in the original data space~~. The right panel shows the CDF of all investigated hourly events with the detected outliers marked in red. The hourly data in the center and right~~left~~ panel were aggregated from the minutely values. Note that an upper limit of 200 mm h^{-1} was set.

4.2 Selected case studies

~~In this following part 4 selected events, that were identified as outliers are presented.~~ The first example in panel a) of figure 4 shows the presence of unusual values in the minutely data of the cross-validated station ($>$ 8 mm min^{-1}). The radar data for that hour ~~are used for result verification and~~ do not show such a high-intensity event ~~that occurred~~ above the investigated location. The second example in panel b) of figure 4 shows a similar case in the minutely data but the radar image confirms the occurrence of the event.

~~In small river catchment, discharge data can also be used to identify the occurrence or absence of an identified outlier. To this purpose,~~ Discharge data from small~~smaller~~ headwater catchments in the federal state of Bavaria with one (or many) rain gauge stations within the catchment were analysed. If a rain gauge observation was identified as being suspicious the discharge values for the next hours following the event were checked. An example for this~~is~~ is shown in the upper Pegnitz catchment which is

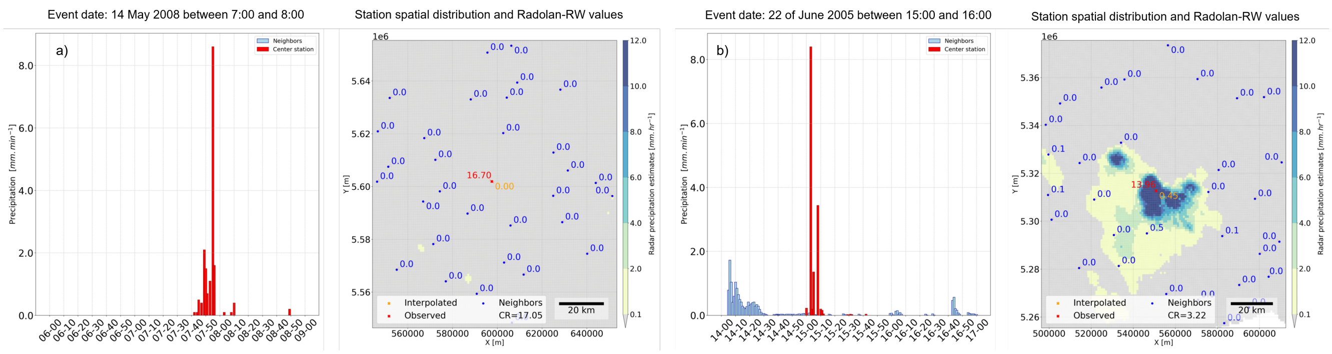

Figure 4. Panels a) and b) show two examples of an identified false observation and a single event~~outlier~~ in the minutely~~hourly~~ values that were verified using the Radolan-RW data.. ~~that occurred on the 14 May 2008 between 7:00 and 8:00~~. ~~Left panel shows the minutely observations of the target and neighboring stations. The right panel shows the spatial distribution of the stations and the corresponding radar image from the hourly RADOLAN-RW values. Panel b) shows an example of an identified outlier in the hourly values that occurred on the 22 of June 2005 between 15:00 and 16:00. Left panel shows the minutely observations of the target and neighboring stations. The right panel shows the spatial distribution of the stations and the corresponding radar image from the hourly RADOLAN-RW values.~~

located on the northern part of the Bavaria (Fig. 5). Panel b) in figure 5 shows an hourly outlier observation that resulted in a reaction in the corresponding headwater catchment. On the other hand, panel c) in figure 5 shows the opposite case, i.e. an hourly outlier that did not cause any reaction in the Pegnitz~~corresponding~~ catchment. ~~Note that both cases are selected for the same gauging station in the Pegnitz headwater catchment.~~

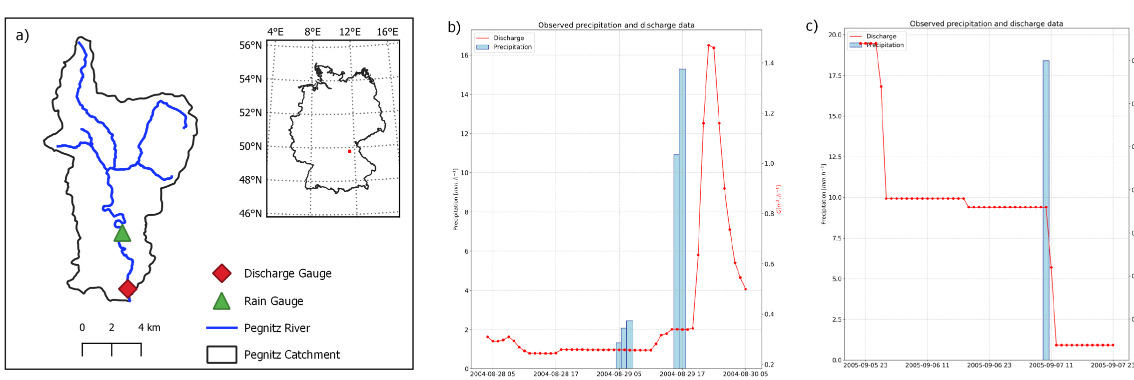

Figure 5. a) Location of discharge and rain gauge station within the Pegnitz headwater catchment (panel a) and observed discharge and precipitation data (+/- 1 day) for detected outliers with b) a discharge increment and c) without a discharge increase

4.3 Results over all stations and aggregations

The method was applied over several temporal aggregations (hourly to daily) and events that are suspicious over single or several aggregations were identified. The result of this can be seen in Table 3. The diagonals show events that are common over the corresponding test and reference temporal aggregation. Some observations are only suspicious until a temporal aggregation is reached or exceeded beyond which they are not detected anymore. The result of this can be seen in the values above and below the diagonals in Table 3.

Table 3. The diagonals show the number of unique days with identified outliers. The ~~other~~ values above the diagonals reflect the number of different days between the reference and test aggregation. For example, there are 358 days in the reference 60 min aggregation that are not in the test 120 min aggregation.

		Test aggregation						
		60 min	**120 min**	**180 min**	**240 min**	**360 min**	**720 min**	**1440 min**
Reference aggregation	**60 min**	**1581**	358	392	414	498	762	898
	120 min	218	**1441**	210	237	354	646	787
	180 min	344	302	**1533**	240	341	657	825
	240 min	437	400	311	**1604**	329	657	837
	360 min	539	535	430	347	**1622**	559	762
	720 min	771	795	714	643	527	**1590**	439
	1440 min	889	918	864	805	712	421	**1572**

The number of active stations (and device quality) affects the number of detected outliers. The red curve in figure 6 represents the ratio of detected outliers to the number of active stations (for every hour) which is shown by the blue curve. As the number of active stations increases the number of detected outliers decreases which is an indication that the quality of the observations is improving with time. In the center panel of figure 6 the effect of seasonality was inspected. The detected outliers were grouped by the month in which they occurred. The results show that in the summer period the number of detected outliers is much larger than in the winter period. This is related to convectional rainfall processes occurring in the summer period leading to more small scale single events. Finally, the percentage of outliers in the investigated events of every station for the hourly aggregated data is presented in the right panel of figure 6. The map does not present any clear structure related to elevation and topography. Moreover, the map shows that outliers can happen everywhere meaning this is not a systematic problem.

–section*.193

–section*.195

–section*.197

5 Conclusion

In this study, a methodology to identify outliers in~~Through this work a methodology to identify outliers in intense~~ precipitation data was presented. Due to the high spatio-temporal variability of precipitation, the quality control of precipitation data

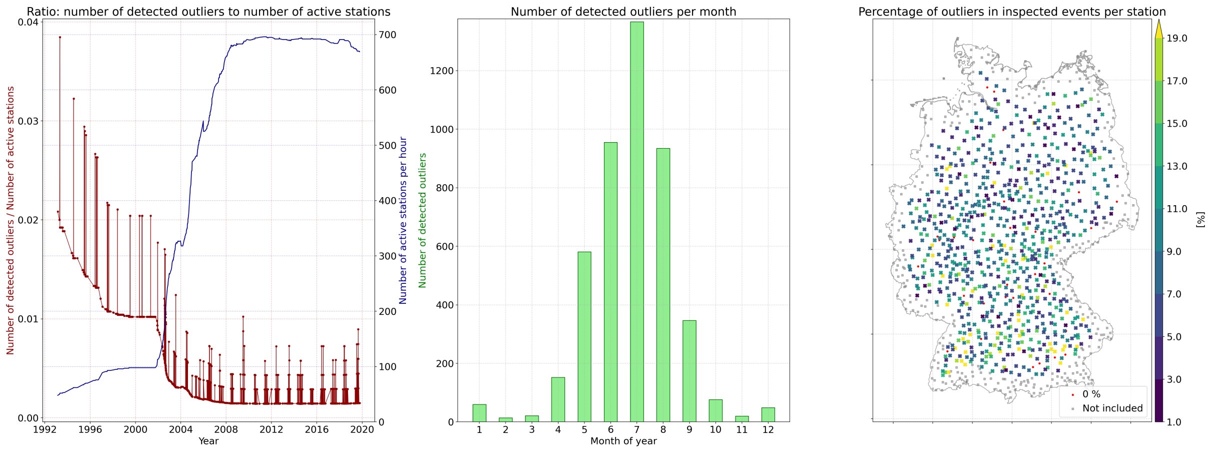

Figure 6. Left figure shows the number of hours with outliers within the investigated hourly events (aggregated from 1 minute) of all stations per~~pro~~ year. The center~~right~~ figure shows the number of detected outliers within the investigated events for every month within the hourly data using a CR value of 3. The right figure shows the percentage of possible outliers in the investigated events of every station.

takes on a special task among the meteorological parameters. Therefore, a transfer of the procedures developed in this paper to other variables is not necessarily reasonable. From an hourly to a daily temporal aggregation the largest four yearly values for every station were inspected~~identified and analysed~~. ~~The method is based on a relatively simple interpolation technique in a cross-validation mode.~~ To cope with the skewed rainfall distribution a Box-Cox transformation with a suitable parameter was applied. ~~The transformation factor was derived by fitting a truncated normal distribution function to the original data and by optimizing the factor to have similar skewness values. The factor is averaged from all stations and is calculated for each aggregation. The difference between the median and the mean skewness reveals that some stations have very high skewness values affecting the mean on the sub-hourly scale. Once the factor has been derived, the transformation of the data and the subsequent cross-validation were applied.~~ The results revealed different outliers~~events~~ throughout various temporal aggregations ~~that strongly differed from their surroundings at the same observation time. Some events were identified as being outliers over several temporal aggregations, while others events appeared/disappeared with change of temporal aggregation. Since several datasets were present, namely the minutely, hourly and daily data, a cross-investigation was done to test if unusual outliers in the different datasets are similar. For events were neighboring stations had false extreme observations, the interpolated value in the original data space (without transformation) is influenced by these values and often exceeds by much the observed value, leading to detecting a false outlier. This can be accounted for by running the method again after all neighbours have been checked. Depending on how many neighboring observations have simultaneously false observations (which is rarely the case) the kriging standard deviation in the transformed space is relatively low. For several stations no outliers were detected in their data.~~ To test the robustness of the method,~~Their~~ data from stations with no outliers were corrupted with several percentages and checked again. The method was able to identify most events as outliers as the value of the added error increased. ~~By decreasing the observed value until reaching a false zero observation the method was able to identify~~

~~around 60% of the cases as being outliers on the hourly scale. With increasing temporal aggregation the identification of false zeros increased reaching 94% on the daily scale. By increasing the error value to up to 100%, almost all values were detected as being an outlier on all temporal aggregations. Note that in this case, different events on each timescale were investigated but there is the possibility to investigate each event on all temporal aggregations.~~ Seasonality was seen to play a major role in the number of detected outliers and as the quality of the observations improved, the number of detected outliers (namely the false measurements) decreased.~~When dividing the data based on the month in which they occurred, the number of identified outliers in the summer period is much larger than in the winter one. This was noted on all temporal aggregations and was presented for the hourly and daily aggregations. This is related to convectional rainfall procedures occurring especially in the period between June and August where several events occur on a local scale. Such events are detected as outliers and are not necessarily false observations.~~ ~~The detected events are denoted as being suspicious as they can either be a false observation or a single event.~~ For distinguishing between a false observation and a single event~~the two possibilities additional~~ external data was used~~has been used~~. Discharge gauge data of corresponding headwater catchments and radar rainfall images were used when available. ~~A change in the water level (or discharge value) within a time interval after the event date revealed the event as being a single event and not a false value. The radar images for the corresponding time step (or for the accumulated sum) were used to find the presence or absence of a rainfall event over the station location.~~ A final choice regarding flagging an observation is done carefully and individually for every location. Eventually, the flagged observations are kept aside and investigated before being used in ~~the~~ further analysis.

The current method needs to be extended and modified for temporal aggregations below the hourly scale. Especially the kriging methodology should include time as a third dimension to account for advection and correlation between subsequent steps. Moreover, many events are identified as being an outlier when part of the neighboring stations had zero precipitation values. This can happen in the case of directional-dependent events driven by a frontal system. These cases could be further handled by including anisotropy in the interpolation method.

~~The aim of this study was to develop a relative simple method to check the intense observed rainfall values and identify unusual observations that should be carefully handled.~~

Data availability. The precipitation data was obtained from the Climate Data Center of the Deutscher Wetterdienst (https://opendata.dwd.de/climate_environment/CDC). The discharge data are provided by the environmental state agency of Bavaria LfU (https://www.gkd.bayern.de).

Code and data availability. The corresponding code is available upon request from the contact author.

Author contributions. AEH developed and implemented the algorithm for the study area. JS assisted in the analysis and description of the results. FI and TJ provided valuable information regarding the data processing. AB designed and supervised the study. Moreover, all authors contributed to the writing, reviewing and editing of the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This study is part of the project B 2.7 "STEEP - Space-time statistics of extreme precipitation" (Grant No. 01LP1902P) of the "ClimXtreme" project funded by the German Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF). The authors thank the German Weather Service DWD for providing the precipitation and weather radar data and the Bavarian Environment Agency for providing the discharge data. Moreover we acknowledge all developers of different python core libraries (e.g. numpy, pandas, matplotlib, cython, scipy) for providing open source code. The authors thank the University of Stuttgart for funding this open-access publication and the two anonymous reviewers for their valuable remarks that helped improve the quality of this manuscript.

References

Bárdossy, A. and Kundzewicz, Z.: Geostatistical methods for detection of outliers in groundwater quality spatial fields, Journal of Hydrology, 115, 343–359, https://doi.org/10.1016/0022-1694(90)90213-H, 1990.

Barnett, V. and Lewis, T.: Outliers in statistical data, John Wiley and Sons Location, Hoboken, NJ, 1994.

Bayerisches Landesamt für Umwelt: Wasserstand und Abfluss, https://www.lfu.bayern.de/wasser/wasserstand_abfluss/index.htm, 2022.

Box, G. E. and Cox, D. R.: An analysis of transformations, Journal of the Royal Statistical Society: Series B (Methodological), 26, 211–243, 1964.

Burkardt, J.: The truncated normal distribution, Department of Scientific Computing Website, Florida State University, 1, 35, 2014.

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., and Vose, R. S.: Comprehensive automated quality assurance of daily surface observations, Journal of Applied Meteorology and Climatology, 49, 1615–1633, 2010.

DWD Climate Data Center (CDC): Historical hourly RADOLAN grids of precipitation depth (GIS-readable), version 2.5, 2021.

Hawkins, D. M.: Identification of outliers, Springer Dordrecht, https://doi.org/10.1007/978-94-015-3994-4, 1980.

Hubbard, K., Goddard, S., Sorensen, W., Wells, N., and Osugi, T.: Performance of quality assurance procedures for an applied climate information system, Journal of Atmospheric and Oceanic Technology, 22, 105–112, 2005.

Iglewicz, B. and Hoaglin, D.: The ASQC basic references in quality control: statistical techniques, in: How to detect and handle outliers, edited by Mykytka, E., vol. 16, pp. 1–87, ASQC Quality Press Milwaukee, WI, 1993.

Ingleby, N. B. and Lorenc, A. C.: Bayesian quality control using multivariate normal distributions, Quarterly Journal of the Royal Meteorological Society, 119, 1195–1225, 1993.

Kaspar, F., Müller-Westermeier, G., Penda, E., Mächel, H., Zimmermann, K., Kaiser-Weiss, A., and Deutschländer, T.: Monitoring of climate change in Germany – data, products and services of Germany's National Climate Data Centre, Advances in Science and Research, 10, 99–106, https://doi.org/10.5194/asr-10-99-2013, 2013.

Klemeš, V.: Tall tales about tails of hydrological distributions. I, Journal of Hydrologic Engineering, 5, 227–231, 2000.

Krige, D. G.: A statistical approach to some basic mine valuation problems on the Witwatersrand, Journal of the Southern African Institute of Mining and Metallurgy, 52, 119–139, 1951.

Lebrenz, H. and Bárdossy, A.: Geostatistical interpolation by quantile kriging, Hydrology and Earth System Sciences, 23, 1633–1648, 2019.

Matheron, G.: Traité de géostatistique appliquée, 14, Editions Technip, 1962.

Qi, Y., Martinaitis, S., Zhang, J., and Cocks, S.: A real-time automated quality control of hourly rain gauge data based on multiple sensors in MRMS system, Journal of Hydrometeorology, 17, 1675–1691, 2016.

Quenouille, M. H.: Approximate tests of correlation in time-series, Journal of the Royal Statistical Society: Series B (Methodological), 11, 68–84, 1949.

Quenouille, M. H.: Notes on Bias in Estimation, Biometrika, 43, 353–360, http://www.jstor.org/stable/2332914, 1956.

Spengler, R.: The new quality control and monitoring system of the Deutscher Wetterdienst, in: Proceedings of the WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation, Bratislava, vol. 330, 2002.

Yu, W., Ai, T., and Shao, S.: The analysis and delimitation of Central Business District using network kernel density estimation, Journal of Transport Geography, 45, 32–47, 2015.