SFB 823

Homogeneity testing for skewed and cross-correlated data in regional flood frequency analysis

Jona Lilienthal, Roland Fried,
Andreas H. Schumann

Nr. 16/2016

Discussion Paper

Homogeneity testing for skewed and cross-correlated data in regional flood frequency analysis

J. Lilienthal*, R. Fried*, A. Schumann†

April 13, 2016

Abstract: In regional frequency analysis the homogeneity of a group of multiple stations is an essential pre-assumption. A standard procedure in hydrology to evaluate this condition is the test based on the homogeneity measure of Hosking and Wallis, which applies L-moments. Disadvantages of it are the lack of power when analysing highly skewed data and the implicit assumption of spatial independence. To face these issues we generalize this procedure in two ways. Copulas are utilised to model intersite dependence and trimmed L-moments as a more robust alternative to ordinary L-moments. The results of simulation studies are presented to discuss the influence of different copula models and different trimming parameters. It turns out that the usage of asymmetrically trimmed L-moments improves the heterogeneity detection in skewed data, while maintaining a reasonable error rate. Simple copula models are sufficient to incorporate the dependence structure of the data in the procedure. Additionally, a more robust behaviour against extreme events at single stations is achieved with the use of trimmed L-moments. Strong intersite dependence and skewed data reveal the need of a modified procedure in a case study with data from Saxony, Germany.

Key words: homogeneity test; regional flood frequency analysis; TL-moments; copulas

1 Introduction

The estimation of frequencies of extreme events like high floods is an important task in hydrology. One of the main difficulties there is the lack of data. To date most gauging stations in Germany provide data of the last 30-50 years. The calculation of high return levels is therefore highly

*Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44221 Dortmund, Germany

†Institute of Hydrology, Water Resources Management and Environmental Engineering, Ruhr-University Bochum, Universitätsstr. 150, 44801 Bochum, Germany

imprecise. To handle this problem, regional flood frequency analysis can be applied, which combines the information from a group of several similar stations.

The popular index-flood procedure (Dalrymple, 1960) assumes that the quantile function of the variable of interest (like the annual flood peak) at the i-th station, Q_i, can be written as

$$Q_i(p) = \mu_i \, q(p), \quad i = 1, \ldots, d,$$

which means that the distribution is the same for each station, except for one site-specific scaling factor μ_i, which is usually estimated by the sample mean or another location parameter. The common factor $q(p)$ is called regional growth curve. A group of stations which fulfils this assumption is called homogeneous.

The assessment of the homogeneity of groups of stations is an important step in regional flood frequency analysis, because heterogeneous groups lead to biased estimations of the regional growth curve (Lettenmaier et al., 1987). Several procedures to test homogeneity have been proposed. Viglione at al. (2007) compared the L-moment based Hosking-Wallis procedure (Hosking, 1990; Hosking and Wallis, 1993) with two rank based tests and concluded that L-moment based tests are better for little skewed distributions while rank based tests are better for higher skewness.

The procedures analysed by Viglione at al. (2007) assume intersite independence, an assumption which is often not valid. Castellarin et al. (2008) investigated this problem for the Hosking-Wallis procedure. Their result is that cross-correlation reduces the power of the test. This means that the heterogeneity of a group of stations is detected less often and therefore heterogeneous cross-correlated groups misleadingly tend to look homogeneous. Their proposed solution is an empirical corrector that is added to the test statistic.

This study generalises and improves the Hosking-Wallis procedure for highly skewed data as well as for the existence of intersite dependence. Trimmed L-moments (Elamir and Seheult, 2003) are used instead of regular L-moments. Hosking (2007) already showed that asymmetrically trimmed L-moments can be preferable to regular L-moments in parameter estimation of generalized Pareto distributions of high skewness. To incorporate the intersite dependence structure between a set of stations, copula models are utilised.

In Section 2 the methodology needed to generalize the Hosking-Wallis procedure is explained. The changes are motivated and displayed in Section 3. In Section 4 simulation studies are performed to compare the new method to existing approaches and to assess the quality of the modifications. A small case study follows in Section 5, in which the need for modifications of the original procedure is revealed. Thereafter the main results are summarized and their relevance is discussed.

2 Methods

This section gives the basics about the Hosking-Wallis heterogeneity measure, as well as the trimmed L-moments and some copula theory, which are later on used to generalize the homogeneity measure.

2.1 Hosking-Wallis heterogeneity measure

The original heterogeneity measure of Hosking and Wallis (1993) is based on the comparison of the observed sample variability of L-moment ratios (Hosking, 1990) and the expected variability under the assumption of homogeneity. It can be divided into three parts: the calculation of a statistic, the calculation of coefficients to normalize that statistic and finally the decision about homogeneity.

First L-moment ratios have to be calculated at each site, as well as regionalized averages of them. Let there be d samples of length $n_1, \dots, n_d$ and let $\hat{\tau}_{(i)}$, $\hat{\tau}_{3(i)}$, and $\hat{\tau}_{4(i)}$ be the empirical L coefficient of variation (L-CV), L-skewness, and L-kurtosis of sample $i = 1, \dots, d$, respectively. The regional averaged versions are defined by

$$\hat{\tau} = \frac{\sum_{i=1}^{d} n_i \hat{\tau}_{(i)}}{\sum_{i=1}^{d} n_i}, \quad \hat{\tau}_3 = \frac{\sum_{i=1}^{d} n_i \hat{\tau}_{3(i)}}{\sum_{i=1}^{d} n_i}, \quad \hat{\tau}_4 = \frac{\sum_{i=1}^{d} n_i \hat{\tau}_{4(i)}}{\sum_{i=1}^{d} n_i}.$$

Hosking and Wallis (1993) defined three different measures of dispersion using L-moment ratios. The two most often applied ones are

$$V_1 = \sum_{i=1}^{d} n_i (\hat{\tau}_{(i)} - \hat{\tau})^2 / \sum_{i=1}^{d} n_i, \tag{1}$$

$$V_2 = \sum_{i=1}^{d} n_i ((\hat{\tau}_{(i)} - \hat{\tau})^2 + (\hat{\tau}_{3(i)} - \hat{\tau}_3)^2)^{1/2} / \sum_{i=1}^{d} n_i. \tag{2}$$

These statistics have to be normalized afterwards. This means that

$$H_i = \frac{V_i - \mu_i}{\sigma_i}, \quad i = 1, 2, \tag{3}$$

is calculated with appropriate values of μ_i and σ_i. To select these, a parametric bootstrap is performed. Using the regional averages of L-CV, L-skewness and L-kurtosis a kappa distribution (Hosking, 1994) is fitted. A large number of datasets is then drawn from this kappa distribution, each of them consisting of d samples with corresponding sample lengths $n_1, \dots, n_d$. For each bootstrap dataset the value of V_i is calculated and the mean and standard deviation over all of these values are inserted for μ_i and σ_i, respectively.

According to Hosking and Wallis (1993) the set of samples is called "acceptably homogeneous" if $H_i > 1$, "possibly heterogeneous" if $1 \leq H_i < 2$ and "definitely heterogeneous" if $H_i \geq 2$. Note that Hosking and Wallis (1993) did not formulate this as a test, but rather as a recommendation. However, in the following sections we will treat this procedure like a test, saying that the test rejects the null hypothesis of homogeneity if the test statistic exceeds 2.

2.2 Trimmed L-moments

Trimmed L-moments (TL-moments) are a generalization of L-moments (Elamir and Seheult, 2003). Like L-moments they are defined as linear combinations of expectations of order statistics, but trimming of some upper and lower order statistics is possible. Let $X_{1:n} \leq \ldots \leq X_{n:n}$ be order statistics of a conceptual sample of size n which is drawn from the distribution of X. The TL(s,t)-moment of order $r \in \{1, 2, \ldots\}$ is defined by

$$\lambda_r^{(s,t)} = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E(X_{r+s-k:r+s+t})$$

with trimming parameters $s, t \in \mathbb{N}$. The ordinary L-moments are a special case of TL-moments when choosing $s = t = 0$. Like other moments, TL-moments characterize the underlying distribution. The first TL-moment is a measure of central location, the second one describes dispersion, and so on. The series of TL-moments $\{\lambda_r^{(s,t)} : r = 1, 2, \ldots\}$ characterizes the underlying distribution uniquely, if it exists (Hosking, 2007).

TL-moment ratios are defined by

$$\tau^{(s,t)} = \frac{\lambda_2^{(s,t)}}{\lambda_1^{(s,t)}}, \quad \tau_3^{(s,t)} = \frac{\lambda_3^{(s,t)}}{\lambda_2^{(s,t)}}, \quad \tau_4^{(s,t)} = \frac{\lambda_4^{(s,t)}}{\lambda_2^{(s,t)}} \tag{4}$$

and are called TL coefficient of variation (TL-CV), TL-skewness, and TL-kurtosis, respectively. These statistics generalize the corresponding L-moment based ones called L-CV, L-skewness, and L-kurtosis.

Sample TL-moments of order $r \in \{1, 2, \ldots\}$ can be calculated as a linear combination of the ordered data $x_{1:n} \leq \ldots \leq x_{n:n}$ by

$$\hat{\lambda}_r^{(s,t)} = \frac{1}{r\binom{n}{r+s+t}} \sum_{i=s+1}^{n-t} w_i \, x_{i:n}$$

with $w_i = \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \binom{i-1}{r+s-k-1} \binom{n-i}{t+k}$ (Hosking, 2007). Sample TL-moment ratios are denoted by $\hat{\tau}_r^{(s,t)}, r \in \{2, 3, \ldots\}$, and are calculated by replacing the theoretical moments with the corresponding empirical ones in equation (4). Alternative methods to calculate TL-moments are given and compared among each other in Hosking (2015).

2.3 Modelling and generating multivariate extremes

Copulas are a convenient and flexible way to model multivariate distributions. In the recent decade the interest in copulas has rapidly increased in the hydrological community (see e.g. Genest and Favre, 2007).

The following Theorem of Sklar is the main theorem of copula theory: Let $X_1, \ldots, X_d$ be random variables with joint distribution F and marginal distributions F_i. Then there exists a decomposition

$$F(X_1,\dots,X_d) = C(F_1(X_1),\dots,F_d(X_d)) \tag{5}$$

which splits the joint distribution into marginals $F_1,\dots,F_d$ and a copula $C : [0,1]^d \mapsto [0,1]$ describing the underlying dependence structure (Sklar, 1959). The reverse of this is also true: with a given d-dimensional copula C and marginal distributions $F_1,\dots,F_d$ formula (5) defines a joint distribution of $X_1,\dots,X_d$. This enables us to separate the modelling of the marginal distributions and of the dependence structure.

Now let each X_i be a maximum of k independent and identically distributed random variables. The Fisher-Tippett-Theorem (see for example de Haan and Ferreira, 2007) states that in this setting for increasing k the univariate limiting distribution of F_i has to be a generalized extreme value distribution (GEV), which is given by

$$F(x) = \begin{cases} \exp\left(-\left(1+\gamma\,\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right), & \gamma \neq 0, \\ \exp\left(-\exp(-\frac{x-\mu}{\sigma})\right), & \gamma = 0, \end{cases}$$

if $1+\gamma(x-\mu)/\sigma > 0$ with μ, σ, and γ being parameters of location, scale, and shape, respectively (Coles et al., 2001).

To construct a multivariate extreme value distribution, we need to combine marginal extreme value distributions with an extreme value copula (Durante and Salvadori, 2010). Gudendorf and Segers (2010) stated that a copula C is an extreme value copula if and only if it is max-stable, that means that

$$C(u_1,\dots,u_d) = \left(C(u_1^{1/t},\dots,u_d^{1/t})\right)^t$$

holds for all $u \in [0,1]^d$ and $t > 0$.

When analysing multivariate block maxima, for example the maximal flood peaks at several stations over one year, it is therefore nearby to model the marginal distributions by the GEV and the dependence between the stations by an extreme value copula. Because the GEV is only the limiting distribution for a fixed block length, other distributions, like the more general kappa distribution used in the Hosking-Wallis procedure, can also be utile.

There are only a few high-dimensional (meaning $d > 2$) extreme value copulas implemented in popular statistical software. The following result comes in useful since it allows to expand the model complexity by combination of multiple simple copulas. If C_1 and C_2 are copulas and $a_i \in [0,1],\ \ i = 1,\dots,d$,

$$C(u_1,\dots,u_d) = C_1(u_1^{a_1},\dots,u_d^{a_d})\; C_2(u_1^{1-a_1},\dots,u_d^{1-a_d}) \tag{6}$$

is a copula as well. If C_1 and C_2 are extreme value copulas, so is C (Durante and Salvadori, 2010). This model introduces new parameters $a_1,\dots,a_d$ which define the mixture between the two copulas for every dimension and enable even modelling of asymmetrical dependence structures.

Fitting the copula parameters to the data is commonly done through a maximum pseudo-likelihood approach, but depending on the copula model relations between the parameter and

Spearman's rho or Kendall's tau can be exploited to derive the parameter (Genest and Favre, 2007).

Being multivariate distributions, copulas can be used to generate data sets which feature specific dependence structures. There are different algorithms to generate random numbers from copulas. One general approach is to begin with one independent uniformly-distributed vector $(v_1, \ldots, v_d)$ and recursively transform each component using the conditional distribution depending on the former components (Bouyé et al., 2000). To generate random numbers from archimedean copulas like Gumbel or Clayton, a more efficient approach is described in Marshall and Olkin (1988). To generate random numbers from a copula constructed as in (6), we apply an algorithm presented by Durante and Salvadori (2010).

Initially each margin of the copula-generated data follows a uniform distribution on $[0, 1]$. To get the desired marginal distributions the inverse probability integral transform is applied at each margin. This way we obtain synthetic data sets featuring a dependence structure described by the copula model and specific marginal distributions.

3 Motivation & construction of the generalized Hosking-Wallis procedure

In this section we describe the modifications which we propose to the original Hosking-Wallis procedure. For this we first review the papers of Castellarin et al. (2008) and Viglione at al. (2007), which dealt with two different drawbacks of the original procedure and we explain the differences to our approach. The modified procedure is summarised afterwards.

3.1 Consideration of intersite dependence with copulas

The original Hosking-Wallis homogeneity measure assumes intersite independence, meaning that the observations of each station are independent of the other stations' observations for the same year. In practice this is a strict assumption, which is often not fulfilled. Stations in the same river network feature a natural dependence, because floods at upstream stations affect floods downstream. Additionally and more generally, all nearby stations are simultaneously influenced by events like snowmelt or synoptic rainfalls.

Castellarin et al. (2008) investigated this problem. Their result is that cross-correlation reduces the power of the test. This means that heterogeneity of a group of stations is detected less often and therefore heterogeneous cross-correlated groups misleadingly tend to appear as homogeneous. Their proposed solution is an empirical corrector. They first calculate the test statistic H_1 (see formula (3)) under the assumption of intersite independence. After this they calculate an adjusted test statistic

$$H_{1,\text{adj}} = H_1 + C \times \overline{\rho^2}(d - 1), \qquad (7)$$

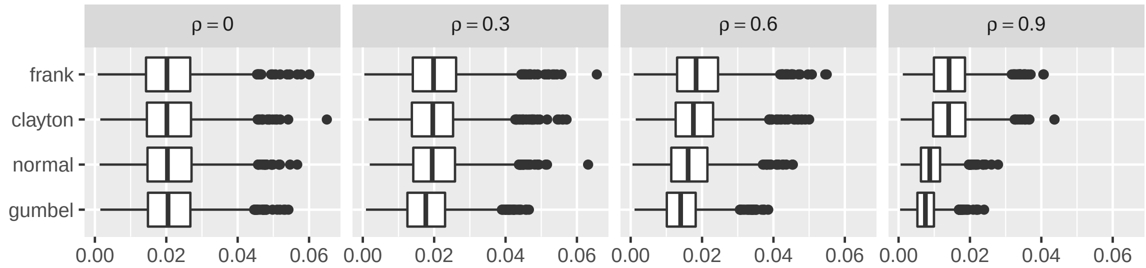

Figure 1: Empirical distribution of V_1 when normal, Gumbel, Frank, or Clayton copulas generate data, grouped by increasing rank correlation.

with the average squared correlation $\bar{\rho^2}$, the number of stations d, and a constant correction coefficient C, which has to be derived by simulations. Castellarin et al. (2008) computed it as $C = 0.122$, but they noted that "the coefficient C is inevitably associated with the Monte Carlo simulation experiments performed in this study".

A drawback of this approach is that the dependence structure is only taken into account through the correlation ρ. Figure 1 shows the empirical distribution of the dispersion of L-skewness (V_1 in formula (1)) when generating data with four common copulas and given rank correlations $\rho = 0, 0.3, 0.6, 0.9$. One can observe that the distribution changes more with increasing correlation and thus would require different corrections. Hence, following the above approach, we would need to calculate a specific correlation coefficient C for each dependence structure.

Our solution to the problem of cross-correlated data differs from the above approach. Instead of ignoring the assumption of independence and correcting for that afterwards, we dispose the assumption by allowing cross-correlated data. Therefore the procedure of generating bootstrap data to calculate the test statistic has to be altered.

First we fit a suitable copula model to our observed data. We then utilise this model to generate data which features the same dependence structure like the observed data (see 2.3). All marginal distributions are modelled to represent the same kappa distribution, whose parameters are determined through L-moments like in the original procedure.

Assuming that our copula model along with the kappa marginals sufficiently describe the distribution of the data, there is no further need to correct the test statistic afterwards and the same decision rules like in the original procedure can be applied.

3.2 Trimmed L-moments instead of L-moments

Another weakness of the original procedure arises when analysing skewed data. Viglione at al. (2007) compared the Hosking-Wallis procedure to two rank-based test statistics, a generalization of the Anderson-Darling goodness of fit test and the Durbin and Knott test (Durbin and Knott, 1972). Viglione at al. (2007) concluded that L-moment based tests are better for little skewed distributions while rank-based tests are better for higher skewness. Their final recommendation is to use the Hosking-Wallis procedure if the regionalized L-skewness is lower than 0.23 and the Anderson-Darling test otherwise.

Our approach is to improve the Hosking-Wallis procedure by using trimmed L-moments (see 2.2) instead of regular L-moments. This is done by substituting the L-moment ratios $\hat{\tau}, \hat{\tau}_{(i)}, \hat{\tau}_3, \hat{\tau}_{3(i)}$ in formula (1) and (2) with corresponding TL-moment ratios $\hat{\tau}^{(s,t)}, \hat{\tau}_{(i)}^{(s,t)}, \hat{\tau}_3^{(s,t)}, \hat{\tau}_{3(i)}^{(s,t)}$ with trimming parameters $s, t \in \mathbb{N}$, which have to be specified. The choice of suitable trimming parameters will be discussed in the simulation studies in Section 4.

3.3 Generalized Hosking-Wallis procedure

The two previous subsections described the changes applied to the Hosking-Wallis procedure. Now we summarise these changes and provide the generalized Hosking-Wallis procedure.

To calculate the generalized Hosking-Wallis procedure the following steps are performed:

1. Analyse the data to identify suitable trimming parameters (s, t) and copula model C.
2. Calculate a TL-moment based statistic:

$$V_1 = \sum_{i=1}^{k} n_i (\hat{\tau}_{(i)}^{(s,t)} - \hat{\tau}^{(s,t)})^2 / \sum_{i=1}^{k} n_i, \tag{8}$$

$$V_2 = \frac{\sum_{i=1}^{k} n_i \sqrt{(\hat{\tau}_{(i)}^{(s,t)} - \hat{\tau}^{(s,t)})^2 + (\hat{\tau}_{3(i)}^{(s,t)} - \hat{\tau}_3^{(s,t)})^2)}}{\sum_{i=1}^{k} n_i}. \tag{9}$$

3. Fit the copula $\hat{C}$ and the marginal kappa distribution $\hat{K}$ to the data.
4. Generate N_{Sim} datasets using the copula approach with copula $\hat{C}$ and equal kappa margins $\hat{K}$ and calculate (8) and/or (9) in each of them.
5. Calculate $\mu_i = 1/N_{Sim} \sum_{j=1}^{N_{Sim}} V_j$ and $\sigma_i = 1/(N_{Sim} + 1) \sum_{j=1}^{N_{Sim}} (V_j - \mu_i)^2$ to get

$$H_i = \frac{V_i - \mu_i}{\sigma_i}, \quad i = 1, 2.$$

6. Classify the group as "acceptably homogeneous" if $H_i < 1$, "possibly heterogeneous" if $1 \leq H_i < 2$, and "definitely heterogeneous" if $H_i \geq 2$.

The original procedure can be obtained by choosing L-moments (setting $s = t = 0$) and the independence copula $C(u_1, \ldots, u_d) = \prod_{i=1}^{d} u_i$. Hence, this new procedure truly generalizes the original one.

4 Simulation studies

In this section simulation studies are carried out to investigate the advantages and drawbacks of the modifications to the Hosking-Wallis procedure presented above.

First of all we compare the new approach directly to the results of Castellarin et al. (2008) and Viglione at al. (2007). Therefore simulations of these works are partially replicated. Additional new simulation studies examine the influence of the trimming parameters and the copula model in more detail. An analysis of the different abilities of heterogeneity detection between L and TL-moments and a small study about robustness against outliers at individual stations complete this section.

We always calculated both test statistics H_1 and H_2, but the results of H_1 were superior to the results of H_2 in nearly every aspect (Viglione at al. (2007) experienced similar results). The statistic H_2 is therefore neglected in the following simulations.

In all simulation studies the proportion of rejections of the null hypothesis over all replications, i.e.

$$\frac{1}{B}\sum_{i=1}^{B} \mathbb{1}(H_1^{(i)} \geq 2) \tag{10}$$

is calculated, with $\mathbb{1}(\cdot)$ denoting the indicator function, which takes the value 1 if $H_1^{(i)} \geq 2$ and the value 0 otherwise. $H_1^{(i)}$ indicates the test statistic in replication i and B denotes the amount of replications. In case that the given group of observations is truly homogeneous the proportion of rejections of the null hypothesis (10) describes the probability of making a type-I-error and is subsequently called size of the test. Otherwise, if the group is heterogeneous, this proportion measures the power of the test.

4.1 Comparison to Castellarin

Castellarin et al. (2008) proposed an empirical corrector applied afterwards to the Hosking-Wallis test statistic to adjust for cross-correlation. We replicated their simulation with 20 stations of which 19 follow a $GEV(1, 0.4, 0)$ and one station follows a $GEV(1, 0.7, 0)$ as marginal distribution. Each station consists of $n = 25$ years of measurements. Because the authors used a multivariate normal distribution to generate cross-correlated data and then transformed the margins to the above-mentioned distributions, we utilised a Gaussian copula to generate similar dependence structures. Rank correlations of $\rho = 0, 0.1, \ldots, 0.8$ were considered. To simulate the case of homogeneity, data sets in which each margin follows a $GEV(1, 0.4, 0)$ distribution are considered as well.

For each setting 25,000 data sets are simulated. We apply the original Hosking-Wallis procedure, the corrected version (with correction coefficient $C = 0.122$ calculated by Castellarin et al. (2008)), and the generalized Hosking-Wallis procedure using L-moments and the Gaussian copula model.

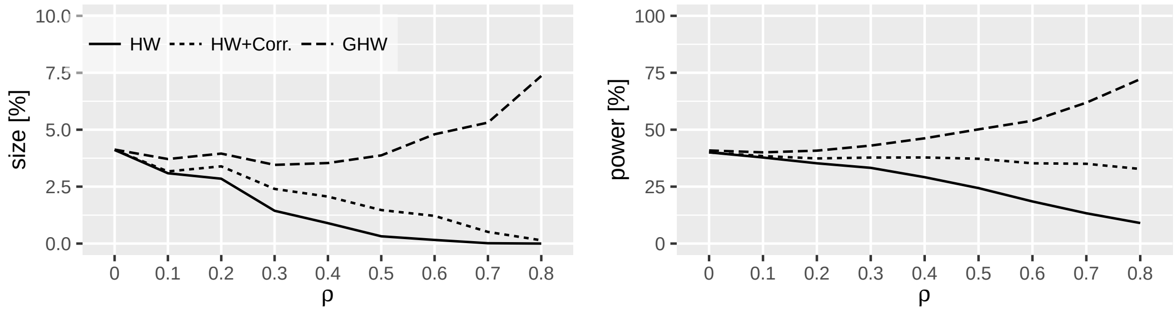

Figure 2: Comparison of size and power between original, corrected, and generalized Hosking-Wallis procedure for different correlations.

Figure 2 contains the size and power of the three procedures depending on the rank correlation coefficient. The size and power rates are similar in the independence case ($\rho = 0$), but differ with increasing correlation. The size of the original and corrected version shrinks with higher correlations, while the size of the generalized version stays higher and exceeds 5% at the highest correlation. The power of the original procedure decreases, but the corrected version is able to compensate and leads to a stable curve. The generalized procedure, however, is able to increase the power with increasing correlation.

This shows that in the settings considered here the generalized version can be more suited to face cross-correlation, because it not only compensates the power reduction but instead can incorporate the dependence to increase the power. It has to be noted that in this simulation we specified the copula model accurately, which is difficult in practice. The problem of misspecification of the copula model will be analysed later.

4.2 Comparison to Viglione

Viglione at al. (2007) compared the Hosking-Wallis procedure to two rank-based procedures, the Anderson-Darling test and the Durbin-Knott test. Their final recommendation is to choose the Hosking-Wallis procedure if the L-skewness coefficient is below 0.23, and the Anderson-Darling-test otherwise. We redid most of their simulation study including our new approach. For this, several combinations of L-CV and L-skewness are considered as the mean of a group of stations. Sets of 11 stations are built with varying either L-CV, L-skewness, or none (which corresponds to the homogeneous case). Each station consisting of $n = 30$ measurements is simulated using the generalized extreme value distribution with parameters corresponding to the specific L-CV and L-skewness.

Besides the original Hosking-Wallis procedure and the Anderson-Darling test, we included the generalized Hosking-Wallis procedure using TL(0,1)-moments and an independence copula to the simulation.

Figure 3 (top panel) gives the size when analysing homogeneously built data sets. Comparing the new procedure (GHW) to the others, it can be observed that the size is comparable or smaller, and that there are no regions in which the new procedure exceeds 5% drastically.

In the bottom panel of Figure 3 the power rates are given for the case that the station's L-CV τ varies equally over a span of $\Delta\tau = 0.5\tau^R$, with τ^R denoting the groups mean L-CV. The generalized Hosking-Wallis procedure is the best among these tests when L-skewness is roughly larger than or equal to 0.2. The original Hosking-Wallis procedure outperforms the other procedures when L-skewness is lower than 0.2. The Anderson-Darling test provides the highest power rates when the groups centre lays on the upper edge of L-CV, but in this region the size (see upper panel) is increased simultaneously.

Besides variation in L-CV, Viglione at al. (2007) examined variation in L-skewness τ_3. With $\Delta\tau_3$ denoting the group's spread in L-skewness, Figure 4 gives the power rates at four specific points in the $\tau - \tau_3$-grid depending on the relative spread. As we can see, the new procedure can compete with the Anderson-Darling test, which was superior to the original Hosking-Wallis procedure in this setting.

In summary, the new procedure outperforms the Anderson-Darling test when analysing highly skewed data sets. When the skewness is low and variations at L-CV scale are expected, the original Hosking-Wallis procedure is still to be recommended.

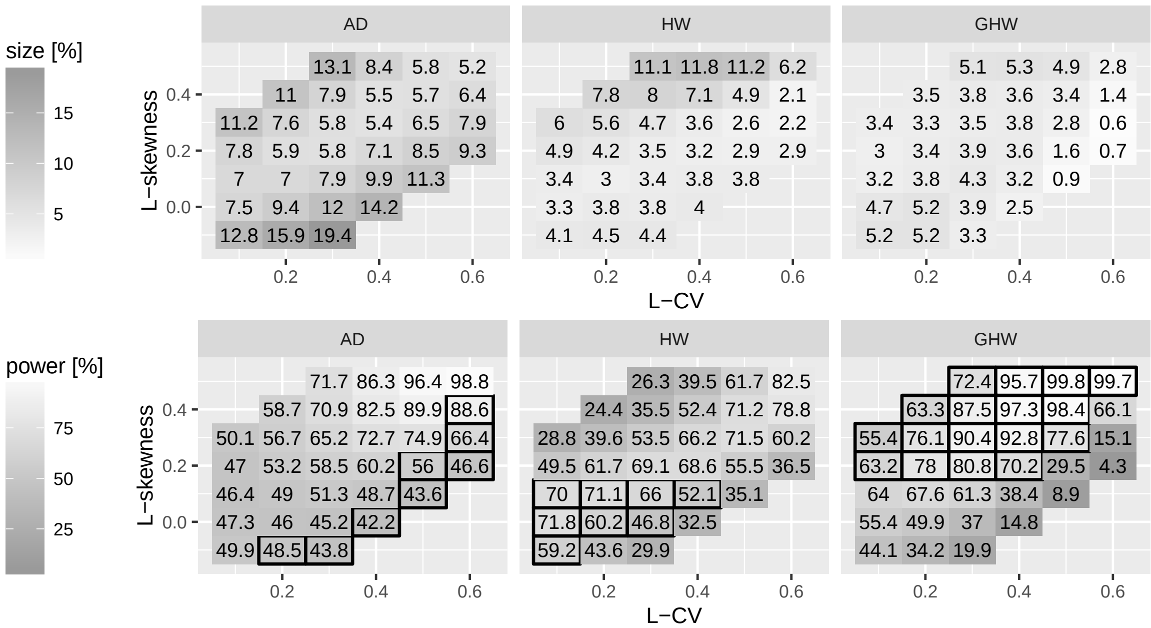

Figure 3: Size and power of the Anderson-Darling, Hosking-Wallis, and generalized Hosking-Wallis procedure in the $\tau - \tau_3$-space. Position on the grid determines the group's mean L-CV and L-skewness. Heterogeneity is constructed varying L-CV. Bordered tiles indicate the procedure with the highest power at each position in the bottom panel.

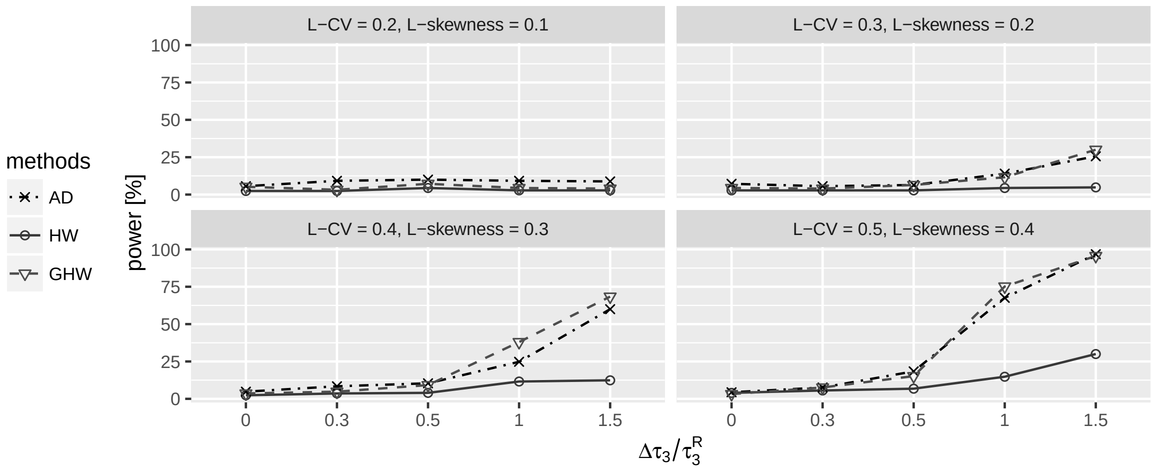

Figure 4: Rejection of the null hypothesis using the Anderson-Darling, Hosking-Wallis, and generalized Hosking-Wallis procedure at specific points of L-CV and L-skewness. Heterogeneity is constructed varying L-skewness.

4.3 Analysis of the new procedure by a simulation study

After these direct comparisons to existing approaches we want to investigate some characteristics of the new procedure. First we want to examine the choice of trimming parameters and of the copula model prior to the analysis. Afterwards a sensitivity analysis is performed to show the differences in detection of heterogeneity between L- and TL-moments and a final study investigates the robustness against outliers at single stations.

In all subsequent studies a situation with $n = 50$ years of measurements for each of five stations is considered. The rather small number of five stations is chosen due to computational reasons. The fitting of complex, high dimensional copula models can be very time-consuming, therefore the analysis of large groups is impractical in simulation studies. In less comprehensive simulations we ensured that our results are valid for larger groups as well.

The simulated network of stations can contain intersite dependence and can either be homogeneous or heterogeneous. The joint distribution of the sites is constructed with the copula approach, which means that the marginal distributions and the copula are specified separately.

The generalized extreme value distribution is selected for every marginal distribution. Four of the stations always have the same GEV parameters, while the parameters of the last station are allowed to vary. Similarly to Viglione at al. (2007) we do these modifications in the space of L-moment ratios, which means that we either modify the L-CV τ or the L-skewness τ_3 of the fifth station. These modifications are indicated as $\tau\downarrow$, $\tau\uparrow$, $\tau_3\downarrow$, and $\tau_3\uparrow$. Figure 5 shows the $\tau-\tau_3$-combinations chosen as well as the corresponding GEV parameters. We consider three parameter combinations as a base point (filled circles; parameters of station 1-4) and vary them at the τ- and τ_3-scale (unfilled circles; possibly modified parameter of station 5 in the

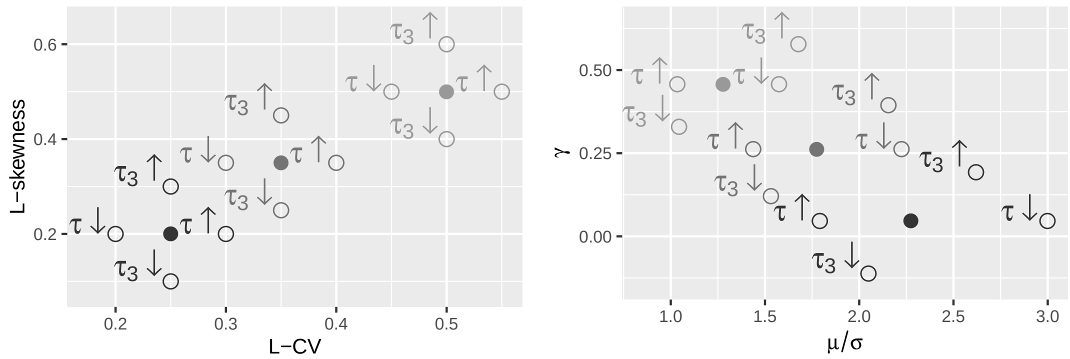

Figure 5: Distributions employed in the simulation studies characterized by their L-CV and L-skewness and their corresponding $GEV(\mu, \sigma, \gamma)$ parameters.

main study). These three combinations represent common distributions we found in real data and feature high, medium, or little skewness.

To model the intersite dependence of these stations copula models are utilised. Table 1 contains all copula models of the following simulations, their parameters and if they are extreme value copulas (EVC) and capture upper tail dependence (UTD). The independent, Gumbel and mixed Gumbel copula are extreme value copulas of increasing complexity (the independent copula is a special case of the Gumbel copula which is a special case of the mixed Gumbel copula). The mixed Gumbel copula is constructed using two Gumbel copulas and the result of formula (6) of Section 2.3. The Clayton copula is neither an extreme value copula nor captures upper tail dependence. It is included to simulate a severe case of copula misspecification.

Each of the subsequently described settings is replicated $B = 5000$ times. Size and power are calculated like in the previous studies (see formula (10)).

Table 1: Copula models applied in the simulation studies, their parameters and if they are extreme value copulas (EVC) and capture upper tail dependence (UTD).

copula model	parameters	EVC	UTD
Independent	-	yes	no
Gumbel	$\theta = 1.5$	yes	yes
MixedGumbel (see formula 6)	$C_1 = \text{Gumbel}(3)$, $C_2 = \text{Gumbel}(1.5)$, $a = (.9, .7, .5, .3, .1)^T$	yes	yes
Clayton	$\theta = 1.5$	no	no

4.3.1 Influence of the trimming

This simulation study assesses which trimming parameters are appropriate, especially in the presence of intersite dependence. A Gumbel copula with parameter $\theta = 1.5$ generates the data, and the Gumbel copula model is also applied in the generalized Hosking-Wallis procedure. Fitting of the copula parameters is done through the maximum pseudo-likelihood approach. Results for L-, TL(0,1)-, TL(0,2)-, TL(0,3)- and TL(0,4)- moments are given below. Other trimmings like TL(1,1), TL(1,2) or TL(2,2) were calculated as well, but lead to inferior results and are not reported here.

The results are depicted in Figure 6. The left panel contains the size and the right one displays the power of the test for the different modifications. Both graphics are grouped row-wise by the three grades of skewness. The different trimming parameters are indicated through different grey scales.

The size using L-moments (TL(0,0)) exceeds 5% noticeably in the medium and highly skewed setting. Application of trimmed L-moments can reduce this to acceptable levels. TL(0,1)-moments leads to a low size, but for higher trimmings the rate raises again.

Because we want to ensure that the size does not exceed the significance level substantially, these results suggests the use of TL(0,1)-moments. Higher trimmings lead to an increased size and should therefore not be considered. L-moments can lead to an increased size when intersite dependence is present.

Looking at the power of the test for the heterogeneous cases, the first finding is that there are differences between the modifications. In the little skewed setting changes of L-CV τ are more likely to be detected than changes of L-skewness τ_3, while in the highly skewed setting the detection rates are more similar.

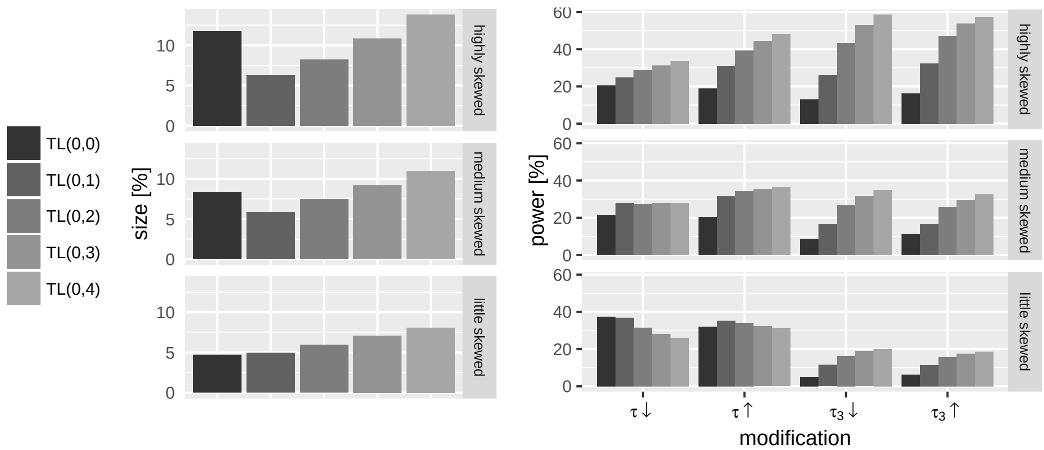

Figure 6: Size and power of the generalized Hosking-Wallis procedure with different TL-moments and Gumbel copula. Synthetic data is generated with Gumbel(1.5) dependence structure.

In the little skewed setting higher trimmings can lead to a decrease in power when changing L-CV, and to an increase when changing L-skewness. In the more skewed settings higher trimmings always lead to an increased power rate, but the gain decreases with increasing trimming.

Our simulation suggests that TL(0,1)-moments are superior when analysing dependent data, especially when the data has high skewness. In comparison to regular L-moments the size is reduced and the power rate is mostly increased. The use of higher trimmings can lead to better detection rates but this comes along with higher sizes. It can be noted as well that, in our settings, little can be gained by applying higher trimmings than TL(0,2).

With these results in mind, we will choose TL(0,1)-moments in the remaining studies to calculate the modified Hosking-Wallis statistics, unless stated otherwise.

4.3.2 Influence of the choice of the copula model

This section deals with the influence of the chosen copula model. The independence copula or the mixed Gumbel model were utilised to generate data sets. Then the generalized Hosking-Wallis procedure was applied to each data set using either the independence copula, the Gumbel copula, or the mixed Gumbel copula, whereat parameters of the latter two copula models had to be estimated. As previously stated TL(0,1)-moments were chosen to calculate H_1.

First we have a look at the results when there was no intersite dependence, depicted in Figure 7. There are no big differences between the fitted copula models. All sizes are around 5% and all power rates are quite similar between the different copula models. This demonstrates that there is little harm in the assumption of a dependence structure when in fact there is no intersite dependence present. Neither raises the size nor decreases the power. Of course, one reason for this is that both the Gumbel copula and the mixed Gumbel copula include the independence copula as a special case. It can be assumed that the estimated parameters are close to the parameter values under independence.

Now we have a look at the opposite case, meaning that the real dependence structure is more complex than the assumed ones (Figure 8). The independence copula is not able to provide good results in this case of dependent data. Both the size and the power are low. The reason for this is that neglecting the dependence structure in the bootstrap procedure leads to wrong standardisation coefficients. Apart from that it is noteworthy that there are no big differences between the Gumbel and the mixed Gumbel model regarding size and power. The Gumbel model is much simpler than the mixed Gumbel model, but suffices to adjust the test statistic in this setting. It seems that an exact modelling of the dependence structure may not be necessary to adjust the test statistics adequately.

To investigate the case of complete misspecification of the copula model, data was generated in the medium-skewed situation using a Clayton(2) copula. The Clayton copula describes dependencies with a lower tail dependence and is therefore very different to the Gumbel model, which features upper tail dependence.

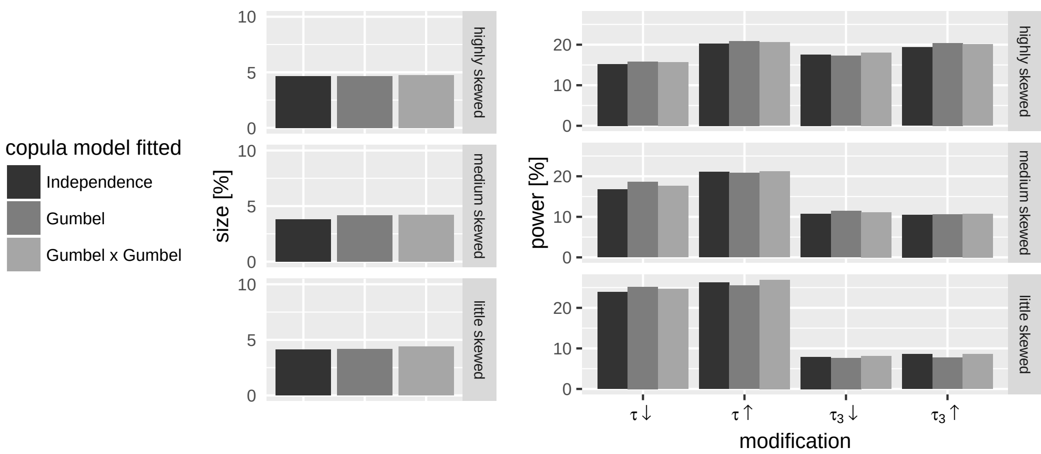

Figure 7: Size and power of the generalized Hosking-Wallis procedure with different copula models and TL(0,1)-moments. Synthetic data is generated without dependence structure.

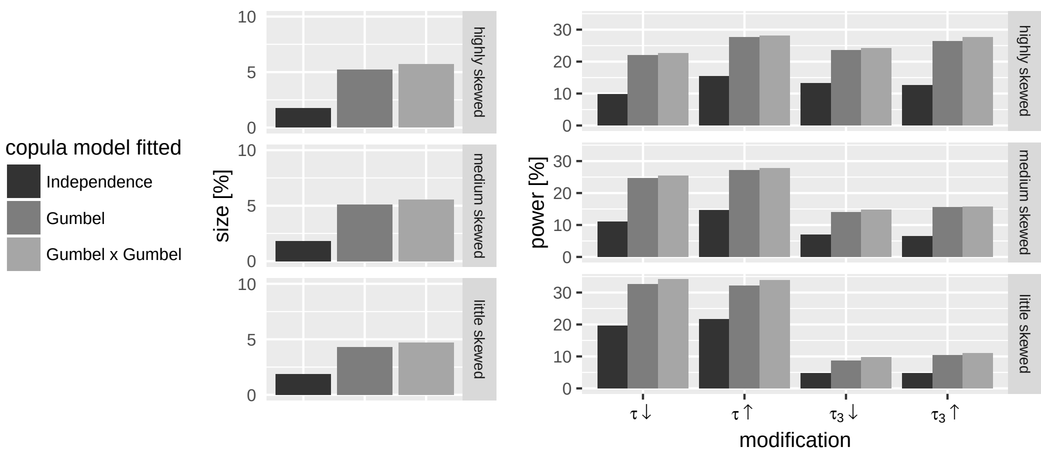

Figure 8: Size and power of the generalized Hosking-Wallis procedure with different copula models and TL(0,1)-moments. Synthetic data is generated with mixed Gumbel model.

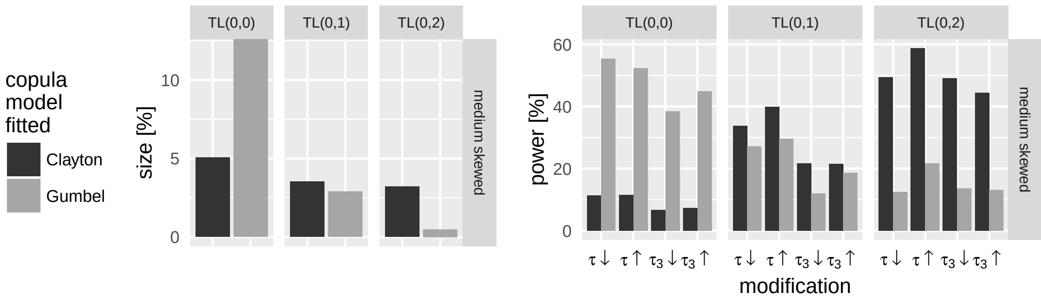

Figure 9: Size and power of the generalized Hosking-Wallis procedure in the case of copula misspecification. Real copula is Clayton(2); Gumbel and Clayton copulas are fitted and different TL-moments are used.

Figure 9 shows the results of this situation, in which the H_1 statistic was calculated under the assumption of Clayton or Gumbel copulas. To analyse the role of the trimmed L-moments in this setting, L, TL(0,1)- and TL(0,2)-moments were chosen.

The Gumbel copula is not to be recommended in the case of L-moments (TL(0,0)) because of a very high size due to the copula misspecification. Fitting a Clayton copula leads to reasonable size but poor power. The results change with trimmed L-moments in the test statistic. The size is lower than 5% for both copula models. The best power can be achieved by the Clayton copula, but a Gumbel copula can also lead to reasonable results (especially with TL(0,1)-moments). Larger trimming leads to very conservative tests in this situation, but it prevents from committing a type-I-error while giving at least some power.

These findings indicate that the application of trimmed L-moments leads to some robustness against misspecification of the copula model. Even in this highly constructed case, in which the copulas for generating the data and for fitting them differ a lot, the test maintains some power while not exceeding the size if trimmed L-moments are used.

Three main conclusions can be drawn from these simulations. The first one is that there is no big drawback when the sites are independent but a Gumbel or mixed Gumbel model is fitted. The power as well as the size are comparable. The second conclusion is that for these copula models it is sufficient to fit the ordinary Gumbel model, even when the true model is of the more complex mixed Gumbel type. For computational reasons this is very convenient, especially when analysing groups of higher dimension. The final conclusion is that the application of TL-moments seems to reduce the impact of misspecification of the copula model. Using TL(0,1)-moments we can get reasonable size and power even when we completely misspecified the copula model.

4.3.3 Sensitivity Analysis

We have seen that the test's power is influenced depending on whether the discordant station differs in direction of L-CV τ or L-skewness τ_3. It is important to investigate this further to be able to decide if an assumed heterogeneity can be discovered by the procedure.

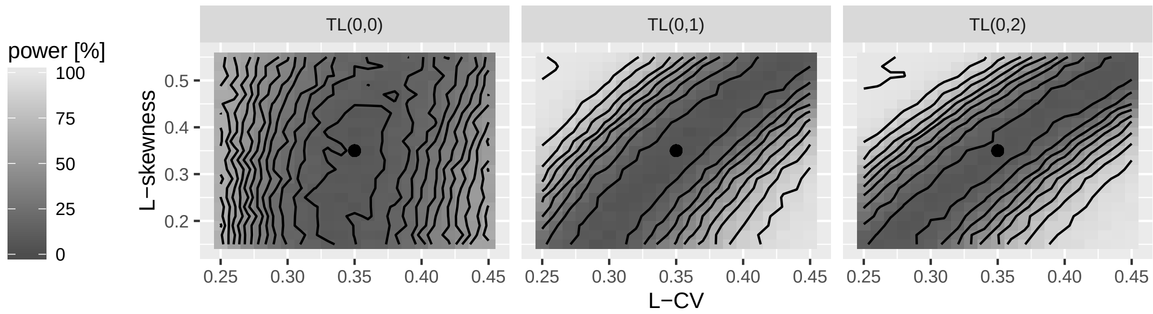

Figure 10: Detection rates of the generalized Hosking-Wallis procedure depending on the position of discordant site grouped by TL-moments.

For this, the medium skewed situation was regarded as base distribution ($\tau = \tau_3 = 0.35$). Now the modified station was not only varied to the four earlier modifications but to all possible combinations of the grid $(\tau, \tau_3) \in \{.25, .26, \ldots, .45\} \times \{0.15, .17, \ldots, 0.55\}$. A Gumbel(1.5) copula described the intersite dependence.

Figure 10 contains the rejection rate of the test statistic in the $\tau - \tau_3$-space of the modified station separated for different TL-moments. Differences in the structure of the rejection rates between L and TL-moments become obvious. With L-moments, the power raises mainly when the L-CV varies. The test statistic only incorporates the empirical L-CV, so this behaviour is very expectable. With TL-moments a variation of lower L-CV and higher L-skewness or inversely can easily be detected, but when L-CV and L-skewness changes in the same direction nearly no detection is possible. The reason for this is that the modified statistic now incorporates TL-moments and therefore can only detect variations in the TL-moment space.

4.3.4 Robustness against outliers

This final analysis concerns the robustness against outliers. An outlier can occur from a measurement error but also from a rare event. The first type could be eliminated by simply removing it, but the second one is a valid observation and removing it would alter the data. Due to the normally short period of observations, already a few very high measurements can have a big impact on the analysis of homogeneity. Therefore it is desirable that single events do not change the results completely.

To check this property a small study of the robustness against outliers was done. A homogeneous group of medium skewness and with a Gumbel(1.5) copula describing their intersite dependence was simulated. After generating data, up to ten observations of one single station were changed to the theoretical 99%, 99.9% or 99.99%-quantile of the chosen distribution. This reflects the situation in which an extraordinarily large, but still possible, observation occurs.

The mean rejection rate, in this case equivalent to the type-I-error, was calculated applying L-, TL(0,1)- and TL(0,2)-moments. The results, displayed in Figure 11, show that the use of trimmed L-moments leads to better robustness against extreme observations. A small number of

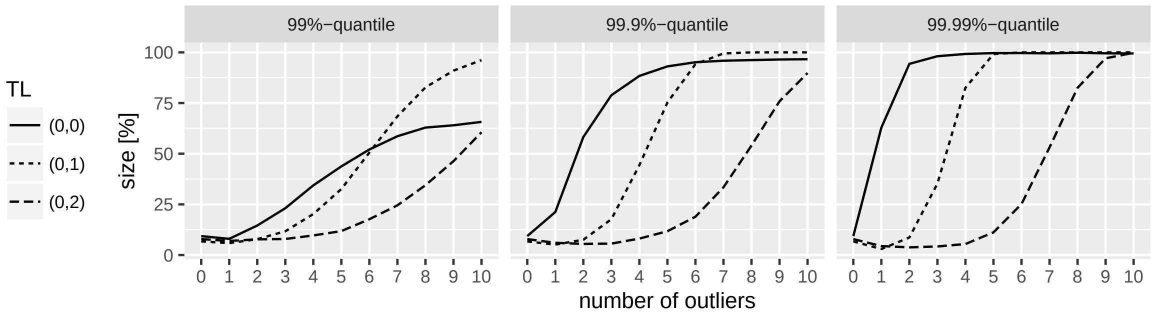

Figure 11: Size of the generalized Hosking-Wallis procedure depending on the number of artificial outliers added to one single site, used TL-moments (linetype) and grouped by the height of outlier.

values equalling the 99%-quantile can be handled by each method, but there are big differences when the very high 99.99%-quantile (corresponding to a return period of 10,000 years) is chosen. Applying L-moments already one of those events leads to a high size, while with TL(0,1)- or TL(0,2)-moments two respectively five events can be handled without a drastic increase of the size.

5 Case study

A small real data example is presented to illustrate the advantage of the modified procedure. For this, maximum flood peaks in summer months (May to October) of five stations in Saxony, Germany are used. The earliest records are from 1910 and the last ones are from 2013. At least 78 years of measurements are available at each station. Table 2 contains basic information about the stations. Besides mean and standard deviation of the summerly maximum flood peak the coefficient of variation is given. The coefficient of variation of the first four stations varies around a value of 1, the last station features a lower value. This discrepancy can be an indication of heterogeneity.

Our task is therefore to assess the homogeneity of the given group of stations, e.g. in order to check if a regional flood frequency analysis is reasonable. A difficulty that arises is the high

Table 2: Summary of maximum flood peaks at stations used in the case study.

Name of station	catchment area [km^2]	years of measurements	n	mean	standard deviation	coefficient of variation
Borstendorf	644	1929 - 2013	78	69.59	71.68	1.03
Hopfgarten	529	1911 - 2013	97	56.69	57.16	1.01
Rothenthal	75	1929 - 2013	82	11.99	11.44	0.95
Streckewalde	206	1921 - 2013	88	22.85	21.24	0.93
Goeritzhain	532	1910 - 2013	101	57.84	44.96	0.78

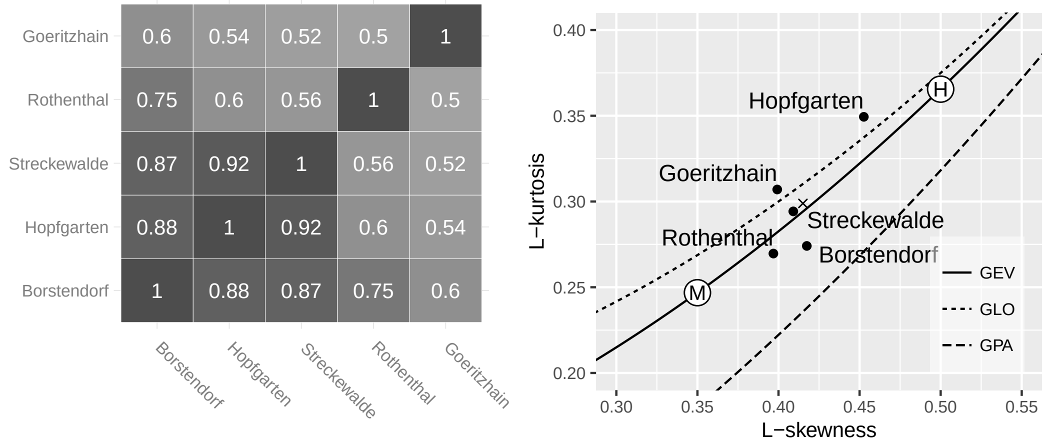

Figure 12: Left: Matrix of pairwise rank correlations between stations. Right: L-moment ratio diagram of L-skewness and L-kurtosis of five stations (dots) and their group mean (cross). Curves give distribution families of generalized extreme-value (`GEV`), generalized logistic (`GLO`), and generalized Pareto (`GPA`). Big circles give distributions employed in the simulation study (`M` and `H` indicating medium and high skewness, respectively).

intersite dependence. Spearman's rank correlation coefficient was calculated for each pair of stations and is displayed in the left panel of Figure 12. Coefficients between 0.5 and 0.92 verify the existence of medium to strong dependencies.

The right panel of Figure 12 depicts a L-moment ratio diagram, which is commonly utilised to assess which distribution is suitable to model the data (Vogel and Fennessey, 1993). The group mean is near the GEV distribution line and inside the area covered by the four-parameter kappa distribution, which contains all pictured combinations below the GLO curve (see Hosking and Wallis, 1993; Hosking, 1994). Usage of the kappa distribution within the bootstrap procedure seems therefore reasonable. Additionally it is observable that all stations lie between the medium skewed and highly skewed setting of the simulation study. Hence, the recommendation derived in this study is to choose asymmetrically trimmed L-moments.

The group was tested with the classical Hosking-Wallis procedure (i.e. with independence copula and L-moments) as well as with the modified procedure using different copula models and TL(0,1)-moments. The simulation studies showed that when trimmed L-moments are applied the procedure is relatively robust against copula misspecification and that simple copulas can be sufficient. To check this, different copula models are fitted: a simple one-parameter Gumbel copula, a mixed Gumbel copula and a more complex pair-copula using a D-vine structure (see Aas et al., 2009). Pair-copulas split the joint density into a combination of marginal densities and bivariate copulas and are popular in financial applications. The pair-copula is added here to compare the results of copula models, whose structure differ completely from our applied models. Table 3 gives information about the fitted copula models. The pair-copula is given in the notation of Aas et al. (2009).

Table 3: Parameters of copula models fitted in the case study.

Model	Parameters
Independence	-
Gumbel	$\theta = 1.77$
Mixed Gumbel	$\theta_1 = 3.45, \theta_2 = 1.71$
(see formula 6)	$a = (0.59, 0.11, 0.72, 0.13, 0.75)^T$
Pair-copula	$c_{12} = G(1.67), c_{23} = J(2.10),$
	$c_{34} = N(0.58), c_{45} = G(1.50)$
	$c_{13\|2} = N(0.78), c_{24\|3} = N(0.23),$
	$c_{35\|4} = G(3.21), c_{14\|23} = G(1.77)$
	$c_{25\|34} = F(-0.82), c_{15\|234} = N(0.42)$

G, J, N, F: Gumbel, Joe, Normal, Frank copula, respectively.

The classical Hosking-Wallis procedure gives us the score $H_1 = -0.30$ and is therefore not able to detect heterogeneity. The corrected version of Castellarin et al. (2008) (see formula (7)) yields $H_{1,\text{adj}} = 0$ putting in the correction coefficient $C = 0.122$ and the average squared correlation $\bar{\rho}^2 = 0.61$. Determination of a customised correction coefficient for this situation might be necessary. Our proposed generalized procedure using TL(0,1)-moments returns $H_1 = 3.78$, $H_1 = 3.36$, and $H_1 = 3.68$ assuming a Gumbel, mixed Gumbel, or pair-copula model, respectively. Thus, all generalized procedures considering intersite dependence yield heterogeneity within the given group of stations, even with quite similar values of the test statistic. Actually, even the procedure using TL(0,1)-moments but assuming intersite independence indicates heterogeneity ($H_1 = 2.16$). Regardless of the chosen copula model, all procedures yield homogeneity of the group only if the station "Goeritzhain" is excluded from the group.

6 Summary and conclusions

We have proposed a generalization of the Hosking-Wallis procedure that uses trimmed L-moments and copula models to overcome the known disadvantages when highly-skewed or cross-correlated data occurs.

First we compared it directly to results of former studies. This showed that the new procedure is capable of improving drawbacks of the original procedure. In simulation studies we investigated the choice of the degree of trimming of TL-moments and the selection of copula models, as well as the robustness to unusual frequently appearance of extreme values. A small case study illustrates that the classic procedure can fail to detect heterogeneity due to intersite dependence and medium to high skewed distributions. The improved procedure, however, is able to detect heterogeneity in this application.

Overall the generalized procedure offers an improvement to the original procedure in many cases. The drawback, in exchange, is the need to specify a copula model and to select trimming parameters of L-moments. The most important observations/recommendations are as follows:

1. In our simulations the test statistic H_1 is superior to H_2 in almost every setting.

2. There is a disparity in the test's power rate depending on the direction in which the discordant site varies. It is important to keep in mind that some variations are not detectable and that the detectable region differs depending on whether L- or TL-moments are used.
3. Asymmetrical trimmed L-moments are beneficial in most settings. We recommend the application of TL(0,1)-moments when analysing moderately skewed data. TL(0,2)-moments could be superior when analysing highly skewed data, but tend to increase the type-I-error.
4. The usage of copula models does not harm when analysing independent data when the independence copula is a special case of the used copula model.
5. When analysing dependent data, simple copula models seem to be sufficient to calculate adequate test statistics, even when the dependence structure is more complex. Trimmed L-moments lead to a robustification against copula misspecification.
6. Application of trimmed L-moments also leads to a more robust behaviour when a station experiences very high values unusually often. The degree of robustness increases with the degree of trimming, which needs to be chosen prior to the analysis.

Based on our results, we can deduce appropriate trimming parameters depending on the empirical regionalized L-CV and L-skewness. The use of these moment ratios is straightforward since we have seen a strong connection to the optimal trimming in Section 4.2. Given a new data set we calculate the regionalized L-CV and L-skewness and deduce the trimming parameter based on Figure 3 (bottem middle and bottom right panel) by selecting the trimming which leads to higher power at the computed position of L-CV and L-skewness. To check this approach we have applied it to the same data sets used previously in Section 4.2. Except for the border region shown in Figure 3 this leads us to the better trimming for most of the considered parameter combinations. Even in the border region the power resulting from this approach never drops below 90% of the better test, so that we recommend such an adaptive choice for practical use.

Acknowledgements

The financial support of the Deutsche Forschungsgemeinschaft (SFB 823, "Statistical modelling of nonlinear dynamic processes" is gratefully acknowledged.

References

Aas, K., Czado, C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. Insurance: Mathematics and economics, 44(2), 182-198. doi:10.1016/j.insmatheco.2007.02.001

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., & Roncalli, T. (2000). Copulas for finance-a reading guide and some applications. Available at SSRN 1032533. doi:10.2139/ssrn.1032533

Castellarin, A., Burn, D. H., & Brath, A. (2008). Homogeneity testing: How homogeneous do heterogeneous cross-correlated regions seem?. Journal of Hydrology, 360(1), 67-76. doi:10.1016/j.jhydrol.2008.07.014

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). An introduction to statistical modeling of extreme values (Vol. 208). London: Springer.

Dalrymple, T. (1960). Flood-frequency analyses, manual of hydrology: Part 3 (No. 1543-A). USGPO.

De Haan, L., & Ferreira, A. (2007). Extreme value theory: an introduction. Springer Science & Business Media.

Durante, F., & Salvadori, G. (2010). On the construction of multivariate extreme value models via copulas. Environmetrics, 21(2), 143-161. doi:10.1002/env.988

Durbin J., & Knott M. (1972). Components of Cramer-von Mises Statistics. I. Journal of the Royal Statistical Society. Series B (methodological), 34(2), 290–307.

Elamir, E. A., & Seheult, A. H. (2003). Trimmed L-moments. Computational Statistics & Data Analysis, 43(3), 299-314. doi:10.1016/S0167-9473(02)00250-5

Genest, C., & Favre, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. Journal of Hydrologic Engineering, 12(4), 347-368. doi:10.1061/(ASCE)1084-0699(2007)12:4(347)

Gudendorf, G., & Segers, J. (2010). Extreme-value copulas. In Copula theory and its applications (pp. 127-145). Springer Berlin Heidelberg.

Hosking, J. R. M. (1990). L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. Journal of the Royal Statistical Society. Series B (methodological), 52(1), 105–124.

Hosking, J. R. M. (1994). The four-parameter kappa distribution. IBM Journal of Research and Development, 38(3), 251-258. doi:10.1147/rd.383.0251

Hosking, J. R. M. (2007). Some theory and practical uses of trimmed L-moments. Journal of Statistical Planning and Inference, 137(9), 3024-3039. doi:10.1016/j.jspi.2006.12.002

Hosking, J. R. M., & Balakrishnan, N. (2015). A uniqueness result for L-estimators, with applications to L-moments. Statistical Methodology, 24, 69-80. doi:10.1016/j.stamet.2014.08.002

Hosking, J. R. M., & Wallis, J. R. (1993). Some statistics useful in regional frequency analysis. Water Resources Research, 29(2), 271-281. doi:10.1029/92WR01980

Hosking, J. R. M., & Wallis, J. R. (2005). Regional frequency analysis: an approach based on L-moments. Cambridge University Press.

Lettenmaier, D. P., Wallis, J. R., & Wood, E. F. (1987). Effect of regional heterogeneity on flood frequency estimation. Water Resources Research, 23(2), 313-323. doi:10.1029/WR023i002p00313

Marshall, A. W., & Olkin, I. (1988). Families of multivariate distributions. Journal of the American Statistical Association, 83(403), 834-841. doi:10.1080/01621459.1988.10478671

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Université de Paris, 8, 229-231.

Viglione, A., Laio, F., & Claps, P. (2007). A comparison of homogeneity tests for regional frequency analysis. Water Resources Research, 43(3). doi:10.1029/2006WR005095

Vogel, R. M., & Fennessey, N. M. (1993). L moment diagrams should replace product moment diagrams. Water Resources Research, 29(6), 1745-1752. doi:10.1029/93WR00341