SFB 823

Joint modeling of annual maximum precipitation across different duration levels

Benedikt Gräler, Svenja Fischer, Andreas Schumann

Nr. 83/2016

Discussion Paper

Joint modeling of annual maximum precipitation across different duration levels

Benedikt Gräler
Ruhr University Bochum
Benedikt.Graeler@rub.de

Svenja Fischer
Ruhr University Bochum
Svenja.Fischer@rub.de

Andreas Schumann
Ruhr University Bochum
Andreas.Schumann@rub.de

Abstract

Summarizing a series of rainfall events for different duration levels by their annual maxima provides valuable information. These statistics are e.g. the design base of urban drainage systems. Investigating an entire set of duration levels, the dependence among them has to be taken into account. We propose an approach where a set of generalized extreme value distributions and a D-vine copula are flexibly parameterized by the set of duration levels of interest. A priori, it is not necessary to fix the duration levels nor the number of duration levels. This joint model produces increasing values for both, longer duration levels and larger return periods. In a sample application, we show that this model is flexible enough to capture variations across the duration levels while reproducing the correlation structure of the data. A joint probabilistic model allows to study a new set of design questions where conditional probabilities or joint return periods are of interest. This is for instance the case when nested subbasins are studied. An urban area within a larger catchment will be sensitive to annual maxima of shorter durations due to high intensities while the enclosing catchment is prone to annual maxima of long durations due to huge volumes. A risk analysis of the entire catchment requires a joint study of both and an approach where the duration levels' dependence is taken into account.

1 Introduction

The modeling of rainfall maxima in terms of their intensity for different durations is widely used in hydrological applications, often known as IDF-curves (intensity-duration-frequency curves). These charts present return periods of the total amount of rainfall for a set of durations that are basically cumulative distribution functions (*Eagleson*, 1970; *Chow et al.*, 1988). The curves - or regionalized proxies - are used to derive design storms that serve as input for many hydrological designs (e.g. dams, dikes, sewer systems). Often, this modeling is done separately for each aggregation level and the results are balanced to ensure consistency among the levels of duration. Obviously, the amount of rainfall that fell in the preceding smaller duration level is a lower bound for all larger aggregation levels. Also, large amounts of rainfall in short duration levels will more likely be increased by small steps opposed to small amounts in the first level. This implies a strong relation between duration levels. In order to mimic this dependence, we introduce a new vine copula based approach where duration levels are no longer treated as independent random variables.

Common approaches, as the one by *Willems* (2000), identify a relationship between the aggregation level and the parameter(s) of the duration level's distribution. *Koutsoyiannis et al.* (1998) present and discuss a mathematical framework for the relationship of IDF-curves. *Grimaldi and Serinaldi* (2006) use a trivariate copula to capture the dependence between critical depth, peak and total depth of a hyetograph. A recent approach by *Bezak et al.* (2016) treats duration and intensity as random variables and models their joint bivariate distribution through copulas. All these approaches clearly introduce dependence among the parameters, but the set of duration levels for a single year are still treated independently in a probabilistic sense. The expected value of these distributions will in most cases be increasing with the duration, but the probabilistic model is merely a set of univariate distributions parameterized by duration.

A commonly used distribution to model yearly, daily or hourly precipitation is the generalized extreme value distribution (GEV). For the United states this is the most recommended distribution in the precipitation atlas for almost all duration levels (*Bonnin et al.*, 2004). *Hanson and Vogel* (2008) show that the Pearson-III-distribution performs best when considering daily precipitation, whereas in Germany the Gumbel-distribution, a special case of the GEV, is used to model 5-minute up to 7-day-precipitation sums (*DWA*, 2012). In general, when considering block maxima such as annual maxima, the GEV is an often used distribution function having theoretical validity due to the Fisher-Tippet-Theorem. The presented approach in this manuscript uses the generalized extreme value distribution across all duration levels.

The paper is organized as follows. The newly developed probabilistic model is motivated and explained in Section 2. An application is presented in Section 3. The model and its applicability is discussed in Section 4. Section 5 concludes the manuscript. An Appendix provides additional details.

Implementations of this approach can be found in the corresponding R-package `hydroTools`[1] on GitHub.

2 The probabilistic multivariate model

2.1 Prerequisites

The goal that motivates this research is to be able to identify a probabilistic model across a set of duration levels. This set of duration levels shall not be fixed in advance (neither their actual durations, nor the number of levels). Hence, a parametrization is sought that generates a multivariate distribution $D_{\mathcal{D}}$ based on a set of durations $\mathcal{D} := \{d_1, \dots, d_n\}$:

$$(X_1, \dots, X_n) \sim D_{\mathcal{D}}$$

where X_i denotes the random variable of the annual maximum rainfall of duration d_i. In order to achieve this, we need to understand the relationship between the duration d_i and the marginal distribution of this duration level as well as the dependence among different duration levels d_i and d_j for any $i \neq j$. At the probabilistic core, we will apply the concept of copulas that allows to tackle univariate margins and dependence structure in two steps. For the marginal design, we will jointly, but independent in a probabilistic sense, estimate multiple generalized extreme value distributions under certain auxillary conditions. The copula design will be based on a D-vine copula.

2.2 The marginal design

The task to fit distributions for different duration levels can be seen as seeking a family of distributions (continuously) indexed by the duration level. By the inclusion of shorter duration levels in longer ones, the surface of the joint cumulative distribution function (CDF) needs to be decreasing along the direction of increasing durations. This allows for increasing annual maxima for both, longer duration levels and larger return periods.

We define a distribution family as combination of generalized extreme value distributions (cGEV), with the additional requirement that the parameters location μ and scale σ and the scale-shape ratio ${}^{\sigma}/_{\xi}$ are non-decreasing for increasing durations. These are sufficient but not necessary conditions to achieve a decreasing CDF surface along d. The CDF for a duration d is given by

$$F_{\mathrm{cGEV}}(x; d) := \exp\left(-\left(1 + \frac{\xi(d)(x - \mu(d))}{\sigma(d)}\right)^{-\frac{1}{\xi(d)}}\right)$$

for $x > \mu - {}^{\sigma}/_{\xi}$ if $\xi > 0$, where $\mu : (0, \infty) \to [0, \infty)$, $\sigma : (0, \infty) \to (0, \infty)$ and ${}^{\sigma}/_{\xi} : (0, \infty) \to [0, \infty)$ are non-decreasing functions for the location and scale parameter and the scale-shape ratio respectively. The parameter function ξ can then easily be inferred from the scale and scale-shape ratio functions. The case $\xi \equiv 0$ corresponds to a combination of Gumbel distributions where the scale-shape ratio function does not apply. The formula for the density function f_{cGEV} follows the analogous notation where constant parameters are replaced by the same functions as above. Considering hydrological applications of extreme

[1] Available on GitHub: `http://github.com/BenGraeler/hydroTools`

rainfall, a non-decreasing location parameter for increasing durations is a natural assumption. Additionally, longer durations of rainfall also allow for a larger variability hence typically leading to an increasing scale. A non-decreasing scale-shape ratio can be explained by a stronger change of the variance than the tail index. Hence, the requirements on the parameter functions are more of a theoretical nature than a limitation in applied hydrology. The monotonicity of the cGEV is discussed further in Appendix A.

The estimation can be based on a stepwise approach where for a set of duration levels d_i the GEV distributions are individually optimized. Plotting these optimized parameters location $\hat{\mu}_i$, scale $\hat{\sigma}_i$ and the scale-shape ratio $\hat{\sigma}_i/\hat{\xi}_i$ against the duration levels d_i will help to identify the family of non-decreasing link functions. A first fit of these link functions can be plugged into the cGEV. In a second iteration, the parameters of the link functions might be optimized to find the best overall fit of cGEV using a joint maximum likelihood approach. The cGEV generalizes the approaches where a linear regression is carried out on the parameters (*Willems*, 2000), as the link functions may as well include linear or piecewise linear relationships. Furthermore, the cGEV distribution also includes the Gumbel distribution that is used in several approaches (e.g. *DWA*, 2012). In the following paragraphs on copulas, we assume that a marginal distribution has already been fitted and the data has been transformed to $[0,1]^n$ using the CDF or using the marginal independent rank-order transformation..

2.3 Copulas

Originating from Sklar's Theroem (*Sklar*, 1959), copulas are multivariate distributions defined on the unit hypercube $[0,1]^n$ that allow to decompose any continuous n-variate distribution H into its margins $F_1, \dots, F_n$ and corresponding copula C by:

$$H(x_1, \dots, x_n) = C\big(F_1(x_1), \dots, F_n(x_n)\big).$$

Hence, copulas "couple" the univariate marginal distributions into a multivariate distribution. A thorough introduction to copulas can be found in *Nelsen* (2007). Many parametric bivariate copula families are well studied and also used in hydrological rainfall applications (among others, see: *De Michele and Salvadori* (2003); *Salvadori and De Michele* (2004); *Zhang and Singh* (2007)). The Kendal's tau correlation measure plays an important role in the concept of copulas. As a rank based measure, it does not depend on the margins and measures dependence already on the copula level. Many copula families facilitate a 1-1 relationship between their parameter and Kendall's tau. This allows for a joint parameterization across different copula families.

Switching from a bivariate to a multivariate setting, many copula families lack the necessary flexibility. Using a multivariate copula family poses the restriction that all pairwise copulas belong to the same family. Furthermore, some Archimedean copulas only allow for a single parameter in any dimension further restricting their ability to adopt to the data. In addition to nested Archimedean copulas, vine copulas (*Aas et al.*, 2009; *Bedford and Cooke*, 2002; *Hobæk Haff et al.*, 2010) are a very flexible extension that decompose the multivariate copula into bivariate building blocks. These bivariate building blocks can then take any available bivariate copula without further limitations. Unfortunately, this decomposition is in general not unique in terms of its structure (trees of the vine, see Figure 1) and copula choices increasing the search space of possible models. *Dissmann et al.* (2013) present a heuristic approach to identify the decomposition structure of the multivariate distribution. They suggest to identify the maximum spanning tree based on the absolute Kendall's tau correlations. Vine copulas have only recently advanced to hydrological rainfall modeling (*Gyasi-Agyei*, 2011; *Vernieuwe et al.*, 2015).

2.4 The multivariate duration level model

In our use-case of vine copulas, we predefine the vine structure to D-vines (drawable vines), based on the premise that the dependence of neighboring duration levels is the strongest (conceptually following *Dissmann et al.* (2013)). For a D-vine, the first tree consist of $n-1$ copulas for the pairs (d_1, d_2), (d_2, d_3), $\dots$, (d_{n-1}, d_n). Each conditioning iteration reduces the set of copulas by one, ending up with in general $1/2(n-1)(n)$ copulas for a complete vine copula of dimension n. Often, the strength of dependence reduces on the higher trees motivating the idea of truncated vines (*Brechmann et al.*, 2012) where independence is assumed for all copulas beyond a certain tree. If such a truncation can be validated based on the data, the number of copulas can considerably be reduced easing the estimation and evaluation of the vine copula.

Figure 1 shows the first two trees for a vine copula on 5 duration levels ($d_1 = 1$, $d_2 = 3$, $d_3 = 6$, $d_4 = 12$ and $d_5 = 24$ hours). The ellipses in the first tree indicate bivariate copulas indexed by pairs of durations for the original data (transformed to $[0, 1]$), while the ellipses on the following trees use conditioned data based on the previous tree. As an example, the copula $C_{1,3|2}(u_{1|2}, u_{3|2})$ describes the dependence between the conditional observations $u_{1|2}$ and $u_{3|2}$ where $u_{i|j}$ can be obtained through the partial derivative ${}^{\partial}/_{\partial u_j}$ of the copula $C_{i,j}(u_i, u_j)$ from the previous tree. The density of the entire D-vine copula is then the product of all bivariate copulas involved (*Aas et al.*, 2009).

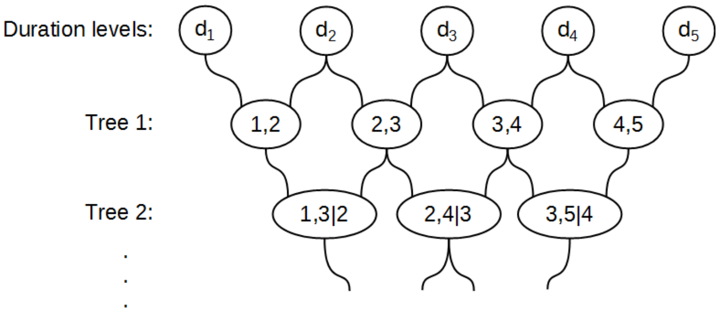

Figure 1: Exemplary D-vine showing the first two trees in the application of duration levels for e.g. $d_1 = 1$, $d_2 = 3$, $d_3 = 6$, $d_4 = 12$ and $d_5 = 24$ hours. The numbers indicate subscripts of the copulas based on the duration levels, e.g. $C_{1,3|2} = C_{d_1,d_3|d_2}$.

Recalling our goal to define a model where a priori neither the number of duration levels nor the durations are known, we need to find a parameterization of the bivariate copulas in the trees of the vine that only depends on the involved duration levels. With such a copula parameterization at hand, any combination of duration levels can be combined into a vine copula. In order to facilitate this, link functions are required that relate the pairs of duration levels directly to a copula parameter or to a proxy like Kendall's tau. At least on the first tree of the vine copula, asymmetries are to be expected, as the rank of the succeeding duration level might considerably increase, but only slightly decrease. Note that the rank of the preceding duration level is not a sharp conceptual lower bound opposed to the observed data of the preceding duration level, as the marginal distributions also increase with the duration level (i.e. the same value receives smaller ranks for larger durations). Nevertheless, these asymmetries are crucial in the design and need to be taken into account for instance by using copulas of the Tawn family (*Tawn*, 1988). Hence, we assume for a set of duration levels $\mathcal{D} = \{d_1, \dots, d_n\}$ a parameterization of the bivariate copulas of the vine such that

$$\begin{aligned} c(u_1, \dots, u_n; \mathcal{D}) &= \prod_{i=1}^{n-1} c_{i,i+1}\big(u_i, u_{i+1}; d_i, d_{i+1}\big) \\ &\quad \cdot \prod_{i=1}^{n-1}\prod_{j=2}^{n-i} c_{i,i+j|i+1,\dots,i+j-1}\big(u_{i|i+1,\dots,i+j-1}, u_{i+j|i+1,\dots,i+j-1}; d_i, \dots, d_{i+j}\big) \end{aligned}$$

where the subscripts of the copulas refer to the subscripts of the duration levels to simplify the notation (e.g. $c_{i,i+1} = c_{d_i,d_{i+1}}$) and recursively:

$$\begin{aligned} u_{i+j|i+1,\dots,i+j-1} &= \frac{\partial}{\partial u_{i+1|i+2,\dots,i+j-1}} C_{i+1,i+j|i+2,\dots,i+j-1}\big(u_{i+1|i+2,\dots,i+j-1}, u_{i+j|i+2,\dots,i+j-1}\big) \\ u_{i|i+1,\dots,i+j-1} &= \frac{\partial}{\partial u_{i+j-1|i+1,\dots,i+j-2}} C_{i,i+j-1|i+1,\dots,i+j-2}\big(u_{i|i+1,\dots,i+j-2}, u_{i+j-1|i+1,\dots,i+j-2}\big). \end{aligned}$$

In general, the bivariate copulas will be of a form where the set of durations $\{d_i, \dots, d_{i+j}\}$ are plugged into a corresponding link function that returns the suitable copula parameters. The copula family might also change for different combinations of durations as implemented for the bivariate spatial copulas (*Gräler*, 2014), but motivated by the application in Section 3 we will here only consider the single family case.

2.5 Estimation of the multivariate duration level model

As the relationship between durations and copula parameters will in general vary between different stations in their parameterization and possibly their general shape, we propose a two stage approach per tree of the vine. In the first stage, the copula parameters are estimated for several combinations of duration levels. A screening of empirical scatter plots and comparison with theoretical copulas[2] might help to identify the appropriate copula family. In addition, a set of likely suitable copula families could be fitted to the same set of pairs of duration levels and the overall best performing family (e.g. in terms of AIC) is selected. The parameter estimates for a copula family results in a (scattered) surface of values for each parameter interpreting the first and second input of the copulas (the durations) as coordinate axes (i.e. each pair of duration levels results in a parameter estimate). Now, a link function has to be identified that is able to describe the relationship between pairs of duration levels and copula parameter(s) (see Section 3 for an applied example). With these link functions, the copulas of the first tree are fully parameterized and their theoretical Kendall's tau value can be evaluated. This measure can be employed in the second stage to optimize the link functions' parameters jointly to achieve the overall best approximation of the empirical Kendall's tau values. Likewise, other measures like Spearman's rho or the likelihood could be used for the joint optimization. A computational advantage of using a dependence measure originates from the fact that it can often be expressed in terms of the parameters and an evaluation based on the entire data takes place only once and not in each optimization step as for a likelihood based approach. Once a satisfying set of link functions and their parameters has been identified, the initial observations can be conditioned based on the partial derivatives of the identified copulas and passed to the second tree. In the second tree, the estimation starts again in the first stage by identifying a copula family, suitable link functions and a joint optimization of their parameters in the second stage. This procedure iterates up to a tree where no more relevant correlations are present resulting in a potentially truncated vine copula. To summarize, the general estimation schema is as follows:

1. calculation of annual maxima for a set of duration levels $d_1, \dots, d_n$
2. estimation of a link function for the marginal distribution
3. transformation of the data set to $(0,1)^n$
4. estimation of the link function for the copula in the first tree
5. conditioning the data to proceed with the next tree in the D-vine
6. estimation of the link function for the copula in the current tree
7. repeat steps 5 and 6 up to the desired truncation level

3 Application

Observational data used in this example are hourly precipitation data (mm) at the station De Bilt (WMO: 06260, 52° 6 N, 5° 11 E, 4 m a.s.l.) in The Netherlands. The data records are obtained from the Royal Netherlands Meteorological Institute (KNMI) and span from 1906 to 2007. From this data set, annual maxima are obtained for each duration level from 1 up to 24 hours (every hour).

The two stage estimation approach as introduced in Section 2.2 is applied to the above set of annual maxima. The location and scale parameters follow (quite) nicely the functions $\mu(d) = \frac{a_\mu}{d^{b_\mu}}$ and $\sigma(d) = \frac{a_\sigma}{d^{b_\sigma}}$ for suitable parameters a_μ, b_μ, a_σ and b_σ respectively. However, the shape parameter is, as in many applications, hard to identify. Hence, we base our fit on the scale-shape ratio function allowing to infer the function of the actual shape parameter. Initially assuming a constant scale-shape ratio function (the mean of the ratio of the independently estimated shape and scale parameters), the joint maximum likelihood optimization is in favor of a slight increasing trend of the scale-shape ratio. See Figure 2 for a graphical representation of the marginal link functions and Figure 3 for histograms with superimposed density curves based on the link functions for 5 selected duration levels..

With the marginal distributions set-up, we focus on the copula parameterization in the following. To minimize the effect of variations in the marginal fit, we apply a rank-order transformation to the data and

[2] An interactive tool can be found at www.copulatheque.org.

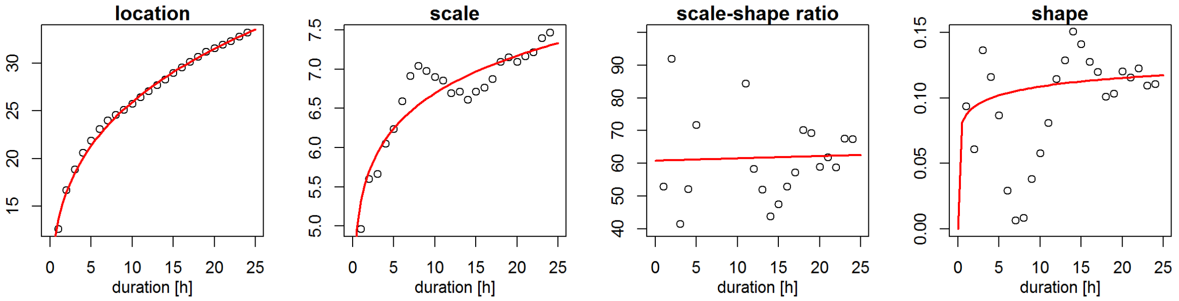

Figure 2: Independently (dots) and modeled (red line) parameters of the cGEV for 1, 2, . . . , 24 hours of duration following the two stage estimation approach outlined in Section 2.2.

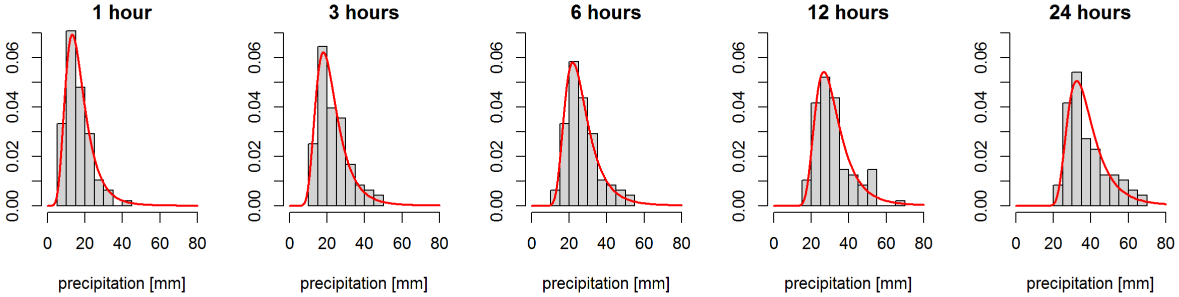

Figure 3: Histograms of annual maximum precipitation with superimposed density curves for five selected duration levels based on the jointly optimized parametrization of the cGEV.

divide by the number of years + 1 to achieve perfectly uniform distributed data on the open unit hypercube $(0, 1)^n$. As initially motived, asymmetries appear in the scatter plots due to the inclusion of smaller duration levels in larger ones. All Archimedean copulas are symmetric about the main diagonal, so are the elliptical copulas, i.e. the Gaussian and Student families. *Tawn* (1988) added additional parameters to the Gumbel copula family to generate a flexible, i.e. asymmetric extreme value copula. Therefore, we restrict the copula families to the Tawn family (see Appendix B for a brief presentation of the Tawn copula family). The data frequently suggests the Tawn type 2 copula (based on the VineCopula package (*Schepsmeier et al.*, 2016), simply referred to as Tawn copula in the following). The set of parameters reveals, as also evident from the scatter plots (not shown) and Figure 4, stronger dependence for pairs of larger duration levels and weaker dependence if the separation of duration levels increases. Also, the degree of asymmetry reduces when the strength of correlation increases as a matter of the definition of the Tawn copula, but also supported by the scatter of the data.

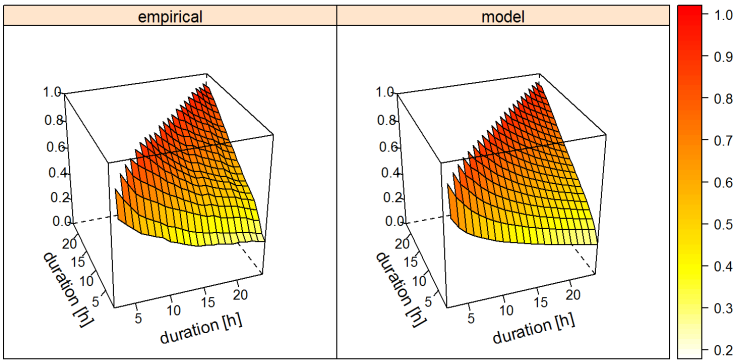

Figure 4: Empirical and modeled surface of Kendall's tau values for pairs of increasing duration levels as used in the first tree of the vine.

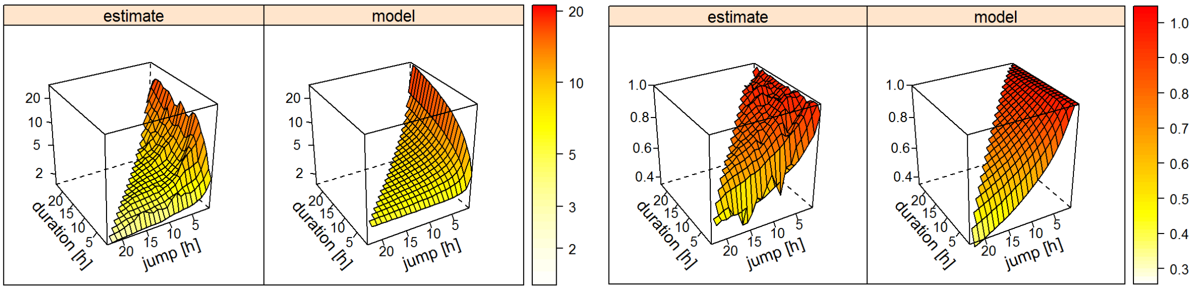

Figure 5: Parameter surfaces based on separate estimates (estimate) and optimized link functions (model) for the first (left panel) and second (right panel) Tawn parameter. The parameter axis in the plots of the first Tawn parameter is on log-scale to ease a visual comparison.

To identify a model that is capable of providing the parameters of the Tawn copula, we look at pairs of duration levels with the same separating distance (referred to as "jumps", e.g. for a jump of two: the pairs for 1 and 3 hours, 2 and 4 hours, ...) and the corresponding parameter estimates. A first set of link functions is fitted to these jumps in terms of the starting duration level and in the following their parameters are modeled across the different jumps. With the fitted link functions for pairs of increasing duration levels, a numerical optimization on the set of parameters is carried out where the theoretical Kendall's tau value of the Tawn copulas is optimized to fit the empirical Kendall's tau values from the data set, see Figure 4 for the surfaces of empirical and modeled Kendall's tau values. Figure 5 presents the parameter surfaces for the first (left panel) and second (right panel) Tawn parameter. The left plot in each panel corresponds to the separate maximum likelihood estimates of the Tawn copula for the pairs of duration levels and jumps, while the right plot in each panel shows the achieved parameter surface based on the link functions. The selected link functions are of the form

$$\begin{aligned}
\mathrm{tawn}_1(d_i, d_j) \quad &:= \quad \left(a_1 + \frac{a_2}{(d_j - d_i)^{a_3}}\right) \cdot d_i^{\,a_4 + \frac{a_5}{(d_j - d_i)^{a_6}}} \\
\mathrm{tawn}_2(d_i, d_j) \quad &:= \quad b_1 + \frac{b_2}{(d_j - d_i)^{b_3}} + b_4 \cdot (d_j - d_i)^{b_5} \left(1 - d_i^{\frac{(d_j - d_i)^{b_7}}{b_6}}\right)
\end{aligned}$$

with suitable parameters $a_1, \ldots, a_6$ and $b_1, \ldots, b_7$, where $d_i < d_j$.

Now, with the fixed parametrization of the copulas of the first tree, we are in the position to generate conditional observations for any combination of duration levels. For the second tree, we do not directly model the copula's parameter, but illustrate the alternative approach of modeling Kendall's tau instead. The copulas on the second tree include the left, center and right duration level of each triple of increasing duration levels. We have to consider that e.g. the triple $(1, 3, 6)$ induces a different relationship than $(6, 8, 11)$ even if both have the same jumps. Hence, we need to look at the spread of Kendall's tau from three perspectives. The empirical Kendall's tau values suggest a reasonable pattern where the absolute correlation is strongest for triples of duration levels where the outer duration levels are direct neighbors of the conditioning duration level (see Figure 6). All correlations on the second tree are negative, indicating that if for a triple of annual maxima corresponding to e.g. 1, 3 and 6 hours of rainfall duration the observed rainfall for 1 hour is "small" given the observed annual maximum for 3 hours (yielding a small value of $u_{1|3}$) it is likely to see a "large" amount for the annual maximum for 6 hours given the amount for 3 hours (indicated by a large value of $u_{6|3}$, compare Figure 10 for an applied example of conditional densities). The link functions receive their initial parameter estimates by a stepwise approach for each conditioning duration level and are in the second stage jointly optimized to overall best approximate the empirical Kendall's tau values. The following link functions representing Kendall's tau in the second tree based on the triple of increasing durations (d_i, d_c, d_j) are of the form

$$\mathrm{ken}(d_i, d_c, d_j) := \frac{c_1}{d_c^{c_2}} \cdot d_i + \left(c_3 + \frac{c_4}{d_c^{c_5}}\right) \cdot d_j$$

for suitable parameters $c_1, \ldots, c_5$.

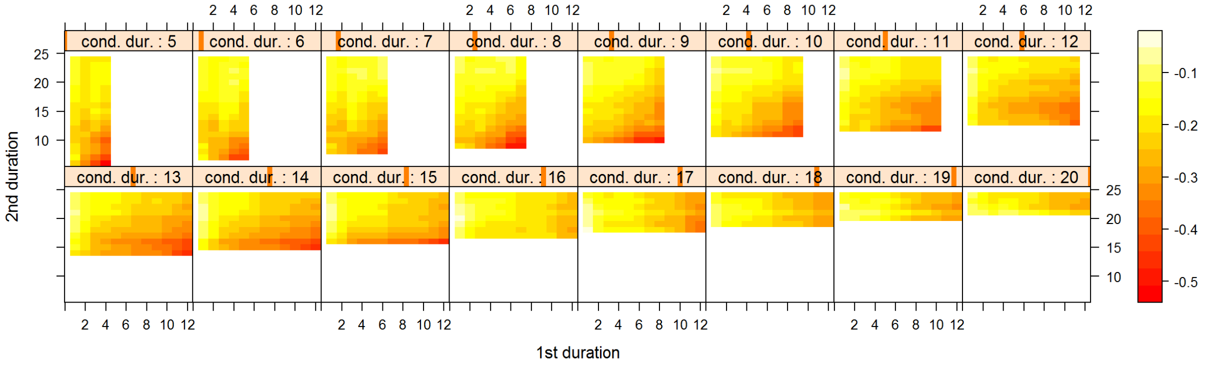

Figure 6: Empirical Kendall's tau values for a subset of the conditioned data pairs in the second tree.

Based on the above link function $\text{ken}(d_i, d_c, d_j)$ we can evaluate the fit of a series of copulas (those that have a 1-1 relationship between their parameter and Kendall's tau) to the datasets for triples of increasing duration levels in the second tree. The copula families in favor showed little variation, but mainly for correlations close to zero, where all investigated families tend to the independence copula. The 270° rotated Joe copula turned out to be the overall best performing family for this dataset. Conditioning the data again using the copula parameterization of the second tree and moving ahead to the third tree did only show small correlations. Therefore, the vine is already truncated after the second tree assuming independence for all larger conditioning sets. Thus, the vine and in conjunction with the cGEV the entire probabilistic model is fully specified for any set of duration levels $\mathcal{D}$.

All calculations have been done using R 3.3.1 (*R Core Team*, 2016). The generalized extreme value distributions are based on the R-package `evd` (*Stephenson*, 2002). The copula functions originate from the packages `copula` (*Kojadinovic et al.*, 2010; *Yan et al.*, 2007), `VineCopula` (*Schepsmeier et al.*, 2016) and `spcopula`[3]. The script reproducing the presented results is contained in the demo `multiDurationLevelModel` of the R-package `hydroTools` on GitHub.

3.1 Simulation based evaluation of the model

Using the fitted model from the above application, we can now simulate from the multivariate distribution. Let us assume that the duration levels 1, 3, 6, 12 and 24 hours are of interest. Repeatedly (1000 times), samples are taken of the same length as the data set (96 years). Inspecting the empirical CDFs of the originally observed and simulated duration levels show only small deviations (see Figure 7 for one exemplary sample). The major effort of this study lies in modeling the dependence structure of the data set. Figure 8 allows to assess the alignment of the model's correlation with the empirical correlation (both using Kendall's tau). Comparing the correlations of the simulated sets with the empirical ones from the data set reveals that for each boxplot approximately 25 % of the correlations are larger or smaller than the empirical ones. Hence, the empirical correlations lie within the central range of the simulated correlations.

As the data consist of only increasing tuples of annual maxima for each year by construction, this is not guaranteed by this model as the marginal distributions have overlapping domains (compare Figure 3). Nevertheless, it is worth to compare the newly developed model with a simulation based on the cGEV only. The samples drawn from the cGEV are independent per duration level. This also becomes evident by their correlation matrix (not shown). Figure 9 shows the distribution of jumps between duration levels for the independent cGEV and the joint cGEV and D-vine model alongside with the original data. The cGEV based simulation typically produces less than 15 % valid tuples while the developed joined vine copula and cGEV model typically achieves a rate of about 80 % valid tuples. Despite the undesired backward jumps, the joint model nicely approximates the distribution of jumps in the original data while the jumps of the unconditioned cGEV case are almost symmetric around 0 and have a considerably different range.

[3] Available on R-forge: `https://r-forge.r-project.org/projects/spcopula/`

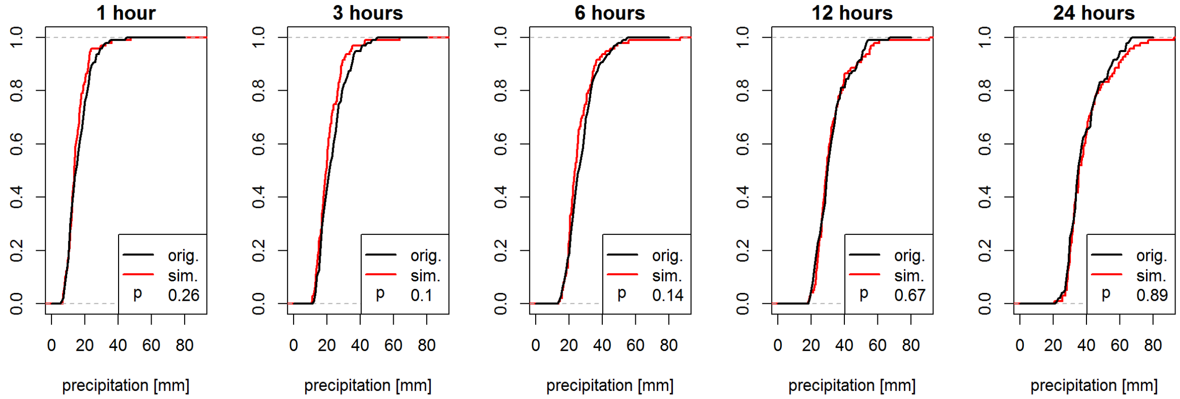

Figure 7: Empirical CDFs of one simulation and the empirical annual maxima for a set of duration levels. The simulated set has the same length (i.e. same number of years) as the empirical one. The legend also quotes the p-value of the Kolmogorov-Smirnov test of the Null-Hypothesis that the two samples stem from the same distribution.

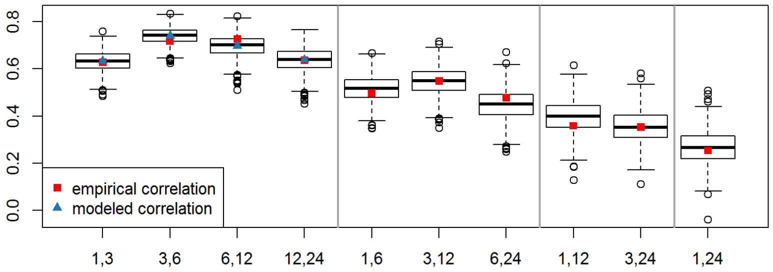

Figure 8: Boxplots of Kendall's tau correlations achieved during 1000 runs. Superimposed are the empirical correlations of the data set (red squares) and the modeled correlations for the directly neighboring duration levels (blue triangles). Note that only the correlations of the first tree in the vine are represented in the correlation matrix (the first off-diagonal). The vertical gray lines separate the boxplots according to the first, second, third and fourth off-diagonal.

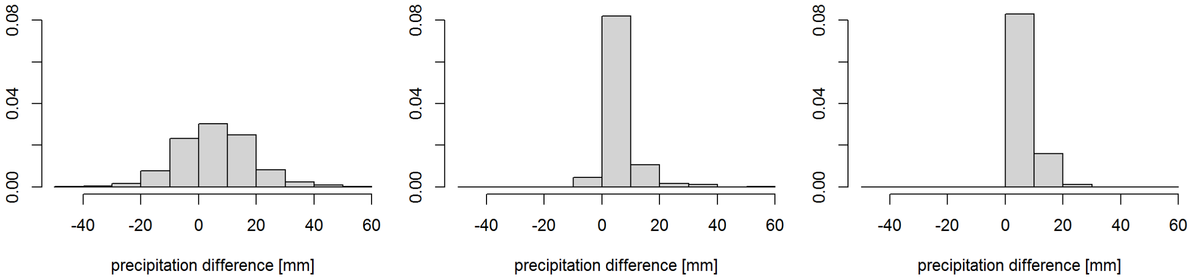

Figure 9: Relative frequencies of jumps in the simulated annual maxima for the pure cGEV approach (left) and the developed joined cGEV and D-vine copula model (center) and the observed data (right).

Now that we have a model describing the dependence of the duration levels, we can investigate the conditioning effect of fixing the annual maximum precipitation for one duration level on another. This is of interest in the scenario where annual maximum precipitation has to be assessed for a heterogeneous catchment containing e.g. urban structures. The left panel of Figure 10 sketches the situation. The urban area is affected the most by high intensity precipitation, hence the annual maximum of short durations is of interest (say 1 hour). The entire catchment has a longer concentration time and the annual maximum of long lasting precipitation is more important (say 12 hours). Hence, one would like to calculate the 95-percentile of the 1 hour annual maximum given that the 12 hours duration level is as well at its 95-percentile. To achieve this, we need the conditional densities of the 1 hour annual maxima $f(x_1|X_4 = x_4) = f_{\text{cGEV}}(x_1; 1)c_{1,4}(F_{\text{cGEV}}(x_1; 1), F_{\text{cGEV}}(x_4; 12))$ for a given value of x_4 (where X_4 is the random variable of duration level d_4 corresponding to 12 hours, as in the previous examples). Note, that this is still the annual maximum of 1 hour and that the extreme 1 hour rainfall might occur during a different event. Nevertheless, it is the joint conditional threat of this heterogeneous catchment within 1 year time.

The effect on the conditional distribution for different conditioning scenarios is illustrated in the right panel of Figure 10. It can be seen that the density curves take considerably different shapes for different conditioning values. The conditional distributions differ by their location, variance and skewness. These scenarios stress the importance of the dependence between duration levels. Noteworthy is the weak bimodal shape of the curve $f(x_1|X_4 = F^{-1}_{\text{cGEV}}(0.95; 12))$ where a tendency for either typical (unconditioned) or large values of X_1 for large values of X_4 can be deduced, going along with an increase in variance. The vertical line segments at the bottom indicate the corresponding 95-percentile of the different scenarios further underpinning the importance of the joint modeling. As previously mentioned, the ranges of the marginal distribution overlap in this model resulting in positive probability for the implausible case that the annual maximum rainfall of 12 hours is smaller than the one for 1 hour. The red dashed line in Figure 10 is the 12 hours annual maximum (X_4) conditional density given $X_1 = 32\,\text{mm}$, but the conditional probability of the larger duration level X_4 being smaller than 32 mm is positive: $F(X_4 \leq 32\,\text{mm}|X_1 = 32\,\text{mm}) \approx 0.06 > 0$. However, the implausible probability mass is considerably reduced compared to the independent treatment of the duration levels where: $F_{\text{cGEV}}(X_4 \leq 32\,\text{mm}; 12) \approx 0.60$. The center and bottom panels of Figure 10 also underline the asymmetry in the dependence. The two scenarios of the 95-percentile can be seen as opponents where only the variables are interchanged. Despite the interchange of variables, symmetric and i.e. elliptical copulas evaluate to the same density altering the unconditioned density the same way in both scenarios (see Appendix C and Figure 12 for a detailed illustration). Similar plots can be obtained for other combinations of duration levels. The evaluation of the conditioned density also works for sets of duration levels, but their visualization is less intuitive.

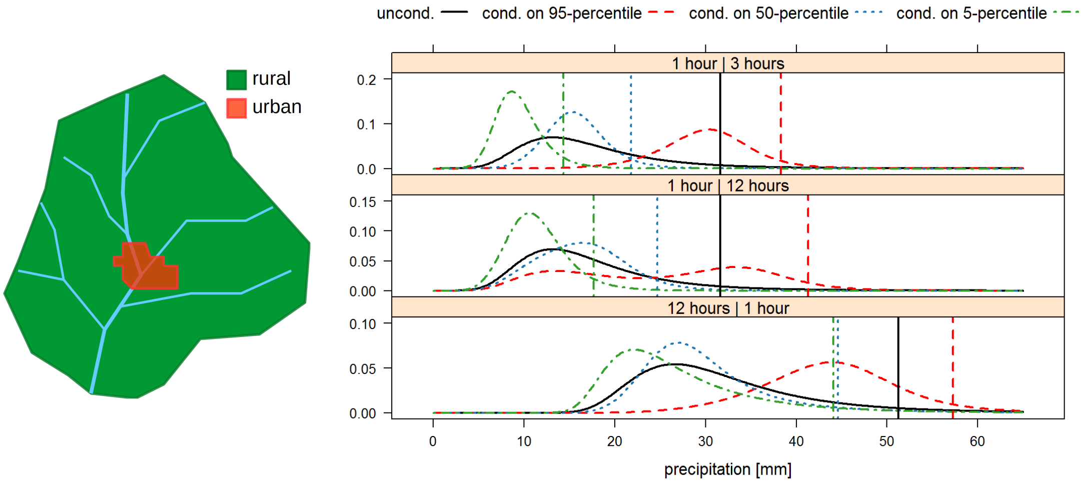

Figure 10: Sketch of a heterogeneous catchment (left) and densities of the annual maximum precipitation for different conditioning scenarios (right). The density curves are shown for the unconditioned case and conditioned under the 95-, 50- and 5-precentile. The corresponding conditioning 95-percentiles are 12 mm, 21 mm and 39 mm for 3 hours and 20 mm, 30 mm and 51 mm for 12 hours and 8 mm, 16 mm and 32 mm for 1 hour. The vertical lines in each panel indicate the resulting 95-percentile of annual maximum 1 hour and 12 hours rainfall respectively. Note that the y-axis uses different ranges to increase the readability of the plot.

4 Discussion

The model is fully specified by the set of duration levels of interest. However, as the estimation is based on a limited range of duration levels, the extrapolation performance of the model beyond these bounds remains subject to the individual quality of the fit of underlying link functions. However, the link functions used in the presented application posses an asymptotic structure that is desirable in an extrapolation scenario.

The sensitivity to outliers has not explicitly been addressed. However, as the estimation of the dependence structure largely relies on Kendall's tau, which is as rank based measure not affected by outliers, the D-vine does not strongly react on outliers. The marginal link functions weaken to some degree the effect of outliers on the margins. Nevertheless, a robust estimation of the marginal distributions might be beneficial in some applications.

Many link functions show systematic deviations around the duration levels of 6 and 15 hours (also the marginal cGEV). This could be due to seasonal effects as most short extreme events will originate from summer events and longer extreme events typically occur during the winter period. Further studies are needed to better understand this behavior. Furthermore, several link functions happen to be of the form $a + \frac{b}{x^c}$. Further data sets need to be investigated to discuss whether this is only a coincidence of the studied data set in De Bilt or if any systematic pattern can be deduced.

The presented model has been designed under the premise that the duration levels of interest are neither fixed in terms of their durations nor their number. If the set of duration levels is known a priori, a static D-vine copula could also be used to model the dependence. Note that this does not necessarily reduce the number of parameters in the model, but would allow to use standard tools that do not require the elaborated steps of identifying and fitting link functions. A switch to a multivariate Gaussian dependence structure would not allow for the asymmetric dependencies present in many scatter plots of the first tree. Hence, the number of non-fully increasingly ordered tuples would likely increase.

With a joint probabilistic model at hand, it is now possible to answer questions based on the distribution of a duration level conditioned on the value of other duration levels (one or more). This enriches the study of annual maximum precipitation for heterogeneous catchments where different duration levels are of simultaneous interest. In the presented application, different conditional bivariate designs have been illustrated, but this approach also allows to consider multivariate return periods for the joint distribution of duration levels (compare *Salvadori and De Michele* (2010); *Salvadori et al.* (2013)). Furthermore, the selection of an ensemble of critical amounts of rainfall for a given return period is possible. This allows to route a set of annual maxima for different duration levels through a design. The presented statistical model only considers the annual maximum precipitation discarding the event structure. An event oriented formulation where shorter duration levels are part of longer ones and their dependence is captured would yield a different model.

5 Conclusions

The introduced approach allows to flexibly parameterize a probabilistic model for a joint modeling of several duration levels based on a D-vine and a set of GEV distributions. The dependence structure of the empirical data set is successfully captured and reproduced during a series of simulations. The model does not guarantee by definition that all tuples are increasing rainfall amounts with increasing duration. Nevertheless, including the dependence considerably increases the amount of well ordered samples from typically below 15 % in the independence case, the current standard, to typically 80 %. Using a joint probabilistic model allows to study a new set of design questions where conditional probabilities or joint return periods are of interest. A situation occurring for heterogeneous catchments where different duration levels are of simultaneous interest. Further efforts have to be made to simplify the estimation process and to invest a seasonality component.

Acknowledgments

The financial support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, SFB 823 "Statistical modelling of nonlinear dynamic processes") is gratefully acknowledged. The data has been made freely available by the Royal Netherlands Meteorological Institute (KNMI) via their website[4].

[4] `http://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi`

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009), Pair-copula constructions of multiple dependence, *Insurance: Mathematics and economics*, *44*(2), 182–198.

Bedford, T., and R. M. Cooke (2002), Vines: A new graphical model for dependent random variables, *Annals of Statistics*, pp. 1031–1068.

Bezak, N., M. Šraj, and M. Mikoš (2016), Copula-based IDF curves and empirical rainfall thresholds for flash floods and rainfall-induced landslides, *Journal of Hydrology*, *541, Part A*, 272–284, doi:10.1016/j.jhydrol.2016.02.058.

Bonnin, G. M., D. Martin, B. Lin, T. Parzybok, M. Yekta, and D. Riley (2004), *Precipitation-Frequency Atlas of the United States*, NOAA, Maryland, United States.

Brechmann, E. C., C. Czado, and K. Aas (2012), Truncated regular vines in high dimensions with application to financial data, *Canadian Journal of Statistics*, *40*(1), 68–85.

Chow, V. T., D. R. Maidment, and L. W. Mays (1988), *Applied hydrology*, McGraw-Hill Series in Water Resources & Environmental Engineering, first ed., McGraw-Hill.

De Michele, C., and G. Salvadori (2003), A generalized pareto intensity-duration model of storm rainfall exploiting 2-copulas, *Journal of Geophysical Research: Atmospheres*, *108*(D2).

Dissmann, J., E. C. Brechmann, C. Czado, and D. Kurowicka (2013), Selecting and estimating regular vine copulae and application to financial returns, *Computational Statistics & Data Analysis*, *59*, 52–69.

DWA (2012), *Arbeitsblatt DWA-A 531: Starkregen in Abhängigkeit von Wiederkehrzeit und Dauer*, DWA working group HW 1.1 ”Niederschlag”, Hennef, Germany.

Eagleson, P. (1970), *Dynamic Hydrology, 462*, McGraw-Hill, New York.

Gräler, B. (2014), Modelling skewed spatial random fields through the spatial vine copula, *Spatial Statistics*, *10*, 87–102.

Grimaldi, S., and F. Serinaldi (2006), Design hyetograph analysis with 3-copula function, *Hydrological Sciences Journal*, *51*(2), 223–238.

Gyasi-Agyei, Y. (2011), Copula-based daily rainfall disaggregation model, *Water Resources Research*, *47*(7).

Hanson, L. S., and R. Vogel (2008), The probability distribution of daily rainfall in the United States, in *World Environmental and Water Resources Congress 2008*, pp. 1–10, American Society of Civil Engineers.

Hobæk Haff, I., K. Aas, and A. Frigessi (2010), On the simplified pair-copula construction – simply useful or too simplistic?, *Journal of Multivariate Analysis*, *101*(5), 1296–1310.

Joe, H. (1997), *Multivariate models and multivariate dependence concepts*, CRC Press.

Kojadinovic, I., J. Yan, et al. (2010), Modeling multivariate distributions with continuous margins using the copula R package, *Journal of Statistical Software*, *34*(9), 1–20.

Koutsoyiannis, D., D. Kozonis, and A. Manetas (1998), A mathematical framework for studying rainfall intensity-duration-frequency relationships, *Journal of Hydrology*, *206*(1), 118–135.

Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Salvadori, G., and C. De Michele (2004), Frequency analysis via copulas: Theoretical aspects and applications to hydrological events, *Water Resources Research*, *40*(12).

Salvadori, G., and C. De Michele (2010), Multivariate multiparameter extreme value models and return periods: A copula approach, *Water Resources Research*, *46*(10).

Salvadori, G., F. Durante, and C. Michele (2013), Multivariate return period calculation via survival functions, *Water Resources Research*, *49*(4), 2308–2311.

Schepsmeier, U., J. Stöber, E. C. Brechmann, B. Gräler, T. Nagler, and T. Erhardt (2016), *VineCopula: Statistical Inference of Vine Copulas*, R package version 2.0.5.

Sklar, A. (1959), Fonctions de répartition à n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, *8*, 229–231.

Stephenson, A. G. (2002), evd: Extreme value distributions, *R News*, *2*(2), 0.

Tawn, J. A. (1988), Bivariate extreme value theory: models and estimation, *Biometrika*, *75*(3), 397–415.

Vernieuwe, H., S. Vandenberghe, B. De Baets, and N. Verhoest (2015), A continuous rainfall model based on vine copulas, *Hydrology and Earth System Sciences*, *19*(6), 2685–2699.

Willems, P. (2000), Compound intensity/duration/frequency-relationships of extreme precipitation for two seasons and two storm types, *Journal of Hydrology*, *233*(1), 189–205.

Yan, J., et al. (2007), Enjoy the joy of copulas: with a package copula, *Journal of Statistical Software*, *21*(4), 1–21.

Zhang, L., and V. P. Singh (2007), Bivariate rainfall frequency distributions using Archimedean copulas, *Journal of Hydrology*, *332*(1), 93–109.

A Monotonicity of the cGEV for increasing duration levels and larger return periods

For a fixed duration d, the quantile function of the cGEV

$$F^{-1}(p) = \mu(d) - \frac{\sigma(d)}{\xi(d)}\Big(1 - (-\log(p))^{-\xi(d)}\Big)$$

is monotone in p by definition and with $p = 1 - {}^{1}\!/\!_{T}$ also monotone for the return period T.

Now, let $d_1 < d_2$ be arbitrarily fixed duration levels. Defining return periods for $T \geq 2$, we have $1 > p = 1 - {}^{1}\!/\!_{T} \geq 0.5$ and hence $0 < -\log(p) < 1$ and:

$$\begin{aligned} 0 > 1 - \big(-\log(p)\big)^{-\xi(d_1)} \quad &\geq \quad 1 - \big(-\log(p)\big)^{-\xi(d_2)} > -\infty \\ 0 < -\Big(1 - \big(-\log(p)\big)^{-\xi(d_1)}\Big) \quad &\leq \quad -\Big(1 - \big(-\log(p)\big)^{-\xi(d_2)}\Big) < \infty. \end{aligned} \tag{1}$$

Based on the positive and non-decreasing scale-shape ratio we get

$$\frac{\sigma(d_1)}{\xi(d_1)} \leq \frac{\sigma(d_2)}{\xi(d_2)}. \tag{2}$$

Multiplying separately left and right parts of (1) and (2) we get

$$0 < -\Big(1 - (-\log(p))^{-\xi(d_1)}\Big) \cdot \frac{\sigma(d_1)}{\xi(d_1)} \leq -\Big(1 - (-\log(p))^{-\xi(d_2)}\Big) \cdot \frac{\sigma(d_2)}{\xi(d_2)}$$

and conclude that adding not more to a smaller $\mu(d_1)$ than to a larger $\mu(d_2)$ yields

$$\mu(d_1) + \left(-\big(1 - \Big(-\log(p)\Big)^{-\xi(d_1)}\Big) \cdot \frac{\sigma(d_1)}{\xi(d_1)}\right) \leq \mu(d_2) + \left(-\big(1 - \Big(-\log(p)\Big)^{-\xi(d_2)}\Big) \cdot \frac{\sigma(d_2)}{\xi(d_2)}\right),$$

completing the proof.

In the Gumbel case where $\xi \equiv 0$ and the restriction on the scale-shape ratio vanishes, we have the quantile function:

$$F^{-1}(p) = \mu - \sigma \log(-\log(p)).$$

As above, it is increasing in p by definition and we also show as above that $0 < -\log(p) < 1$ for $T \geq 2$ with $p = 1 - {}^{1}\!/\!_{T}$ and hence $-\infty < \log(-\log(p)) < 0$. Thus, for any $d_1 < d_2$ and non-decreasing σ we get $0 < -\sigma(d_1) \cdot \log(-\log(p)) \leq -\sigma(d_2) \cdot \log(-\log(p)) < \infty$. Concluding again that adding not more to a smaller $\mu(d_1)$ than to a larger $\mu(d_2)$ yields

$$\mu(d_1) + \big(-\sigma(d_1) \cdot \log\big(-\log(p)\big)\big) \leq \mu(d_2) + \big(-\sigma(d_2) \cdot \log\big(-\log(p)\big)\big),$$

completing the proof. Note that the conditions we give are sufficient for the inequalities to hold, but not necessary. Empirically validating the inequalities for more general link functions might suffice in a specific application.

B The Tawn copula family

Tawn (1988) added additional parameters ψ_1 and ψ_2 to the Gumbel copula family to generate a flexible, i.e. asymmetric extreme value copula. Its Pickands dependence function is given by

$$A(t) = (1-\psi_2)(1-t) + (1-\psi_1)t + \left(\left(\psi_1(1-t)\right)^\theta + (\psi_2 t)^\theta\right)^{\frac{1}{\theta}}$$

for $t \in [0,1]$, $0 \leq \psi_1, \psi_2 \leq 1$ and $\theta \in [1, \infty)$. The extreme value copula is then given by

$$C(u,v) = \exp\left(\log(uv) A\left(\frac{\log(v)}{\log(uv)}\right)\right).$$

The Gumbel copula and its Pickands dependence function occurs when $\psi_1 = \psi_2 = 1$. Note that any extreme value copula can be defined as above for a suitable Pickands dependence function A, where $A : [0,1] \to [1/2, 1]$ is a convex function with $A(0) = A(1) = 1$ and $\max(t, 1-t) \leq A(t) \leq 1$ (see *Joe* (1997) for further details). The Tawn copula (or Tawn Type 2) is a restriction to a two-parameter version of the general three-parameter Tawn copula for $\psi_2 = 1$. Figure 11 illustrates the density for four different pairs of duration levels based on the parameterization of the presented application.

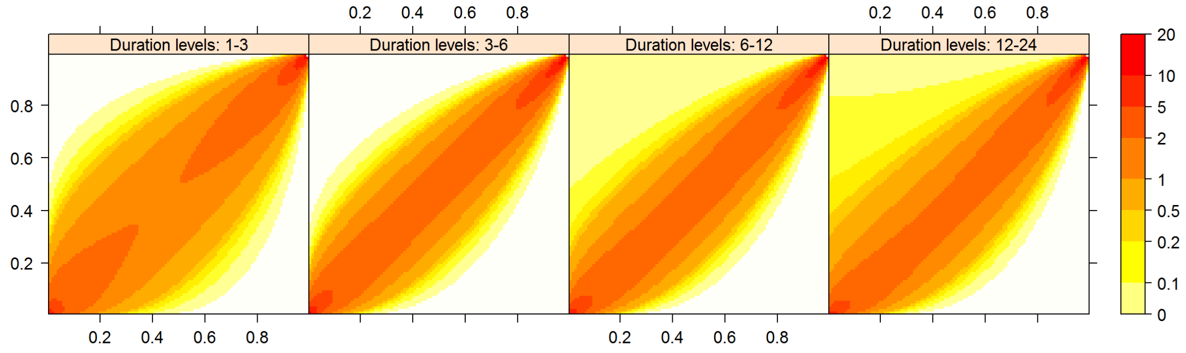

Figure 11: The density of the Tawn type 2 copula for four combinations of duration levels. The colors follow a log-scale to ease visual comparison.

C The asymmetry effect in the conditional densities

The asymmetry of the Tawn copula also has its impact on the conditional densities illustrated in Figure 10. In order to delineate the impact of the symmetry and tail dependence, the Gumbel copula (symmetric with tail dependence) and Gaussian copula (elliptically symmetric, no tail dependence) are selected with the same strength of correlation (Kendall's tau of ≈ 0.37) and the same tail dependence for the Gumbel copula (tail index of ≈ 0.45) as the fitted Tawn copula for 1 hour and 12 hours. Figure 12 shows the unconditioned and conditioned densities using the three different copulas for a conditioning 95-percentile (the continuous black and dashed red lines are the same as in Figure 10). The distortion of the unconditioned densities is the same for the scenario 1 hour—12 hours and 12 hours—1 hour for the Gumbel and Gaussian copula as $c(u, v) = c(v, u)$ holds for symmetric copulas. In contrast, the Tawn copula distorts the unconditional densities considerably different. Furthermore, also the tail dependence has an effect on the distortion as a relatively large quantile is considered. This can be seen in the shift of location of the distributions. The Gumbel and Tawn copulas move the center of mass further to the right than the Gaussian copula that has a tail index of zero by construction.

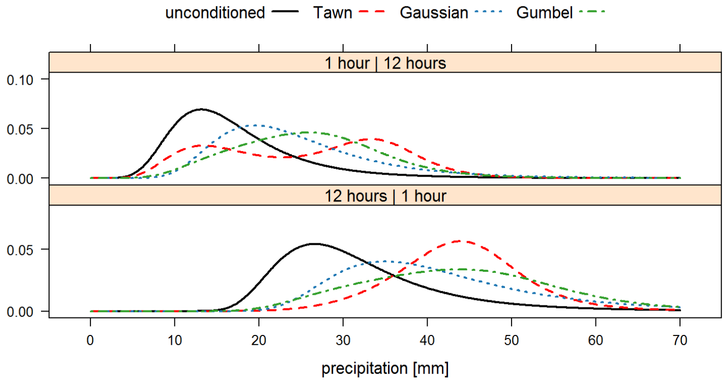

Figure 12: Comparison between the effect of symmetric and asymmetric copulas on the conditional density. The unconditioned (black continuous lines) and conditioned using the Tawn copula (red dashed lines) are identical with the unconditioned and 95-percentile scenarios in the center and bottom panel of Figure 10. The additional curves use the same set-up in terms of conditioning, margins, Kendall's tau (Tawn, Gaussian and Gumbel copula: ≈ 0.37) and tail index (Tawn and Gumbel: ≈ 0.45).