7th International Water Resources Management Conference of ICWRS,
18–20 May 2016, Bochum, Germany, IWRM2016-85-1

An Update on Multivariate Return Periods in Hydrology

Benedikt Gräler[1], Andrea Petroselli[2], Salvatore Grimaldi[3], Bernard De Baets[4], and Niko Verhoest[5]

[1]Institute of Hydrology, Ruhr University Bochum, Germany
[2]Dipartimento di scienze e tecnologie per l'agricoltura, le foreste, la natura e l'energia (DAFNE Department), University of Tuscia, Italy
[3]Dipartimento per la innovazione nei sistemi biologici agroalimentari e forestali (DIBAF Department), University of Tuscia, Italy
[4]Department of Mathematical Modelling, Statistics and Bioinformatics, Belgium
[5]Laboratory of Hydrology and Water Management, Ghent University, Belgium

Correspondence to: Benedikt Gräler (benedikt.graeler@rub.de)

Abstract. Many hydrological studies are devoted to the identification of events that are expected to occur on average within a certain time span. While this topic is well established in the univariate case, recent advances focus on a multivariate characterization of events based on copulas. Following a previous study, we show how the definition of the survival Kendall return period fits into the set of multivariate return periods.

Moreover, we preliminary investigate the ability of the multivariate return period definitions to select maximal events from a time series. Starting from a rich simulated data set, we show how similar the selection of events from a data set is. It can be deduced from the study and theoretically underpinned that the strength of correlation in the sample influences the differences between the selection of maximal events.

1 Introduction

Studying extremes in hydrological multivariate time series often aims at getting an estimate of the size of events to be expected in a period of 10, 50 or 100 years. This information is relevant for the construction of many hydrological structures such as dams and dykes. As most of these natural events are characterized by several variables (e.g. peak discharge, volume, duration, ...) and several locations, it is important to understand their dependence structure and which constellations result in an extreme event. Copulas allow to flexibly model the dependence between the variables and add different marginal distribution functions to build a probabilistic multivariate model. The natural ordering in univariate time series does not extend to the multivariate case calling for different tools to identify multivariate extremes.

In a previous study (Gräler et al., 2013), the practical impact of different bivariate multivariate return period definitions have been studied based on a simulated data set. Meanwhile, an additional approach, the survival Kendall return period (SKRP), has been developed (Salvadori et al., 2013). Using the same data as before, the SKRP is calculated and related to the previously studied return periods (AND, OR and Kendall return period).

Currently, multivariate maxima are often selected based on a single driving variable (e.g. peak discharge) and the associated variables (e.g. volume and duration) are studied in a multivariate setting. However, this does not a priori reflect the joint extreme characteristic that is the actual focus of such a study. Different notions of maximality can be defined following the above return period definitions. These allow to calculate the empirical joint extremeness and to select the maxima of multivariate time series.

In this paper, we will only briefly quote the key concepts. The interested reader is referred to the predecessor of this paper, Gräler et al. (2013), for further details. The following section recalls the definitions of the different multivariate return periods under study and puts them into relation. In Section 3, the different maxima selection regimes are presented and their effect is studied. Section 4 provides a discussion and conclusions.

2 Multivariate return periods

The driving tool underlying the multivariate return period definitions are copulas. Copulas are multivariate distribution functions defined on the unit hypercube. Based on Sklar's Theorem (Sklar, 1959), they combine marginal distribution functions $F_1, \ldots, F_d$ into multivariate distributions H via $H(x_1,\ldots,x_d) = C(F_1(x_1),\ldots,F_d(x_d))$ and solely determine the entire dependence structure. For a detailed introduction, see e.g. the book by Nelsen (2006).

Going from univariate to multivariate extremes is not immediate. One major constrain is the lack of a natural ordering for problems of dimension $d \geq 2$. Typical definitions of the multivariate joint return periods include the OR case corresponding to $P(X_1 > x_1,\ldots,X_d > x_d)$ and the AND case defined for $P(X_1 > x_1 \wedge \cdots \wedge X_d > x_d)$. The Kendall return period (KRP) introduced by Salvadori et al. (2011) is an approach that shares a unique property with the univariate return periods: the critical layer separating safe from dangerous events is unique for every design return period. This is not the case for the OR and AND approaches where different regions of safe and dangerous events exist for the same return period. The basis of the Kendall return period, the *Kendall distribution function* is the distribution function of the copula's mass below its level curves.

Salvadori et al. (2013) present the *survival Kendall return period* (SKRP) to overcome limitations of the *Kendall return period* (KRP) described in Salvadori et al. (2011). The drawback of the latter is its unboundedness. The critical layer splits the region into safe and dangerous events in a way, such that one of the margins might tend to infinity (even though with very small probability). This limitation is overcome by the SKRP, as the critical layer is nicely bounded, as for the OR return period, but every point on the critical layer exhibits the same return period, as in the Kendall scenario. In a way, the SKRP combines the best of both worlds. Its mathematical definition reads

$$T_{\text{SKRP}} = \frac{\mu}{1 - \bar{K}(t)}$$

with $\bar{K}$ the survival Kendall distribution function given by

$$\bar{K}(t) = P\big(\bar{F}(x_1,\ldots,x_d) \geq t\big)$$
$$= P\big(\hat{C}(\bar{F}_1(x_1),\ldots,\bar{F}_d(x_d)) \geq t\big)$$

and $\hat{C}$ the survival copula and $\bar{F}_i$ the marginal survival distribution functions. See Salvadori et al. (2013) for the full details.

In order to extend our previous study, we use the same data (simulated using the COSMO4SUB model (Grimaldi et al., 2012), compare Section 4 in Gräler et al. (2013)) and adopt the same parametrization as in Gräler et al. (2013) to also calculate the SKRP for the bivariate approach. Higher dimensional approaches are out of the scope of this follow-up paper. The peak discharges Q_p are said to follow a Weibull distribution while the associated volumes follow an exponential distribution. Recall that the selection of the annual maxima was done based on the peak discharges and the volumes are the ones corresponding to the same event, but as such not necessarily the largest one in the respective year.

Figure 1 depicts the four different MRP definitions. All points indicate the most likely bivariate event on the respective critical layers. It is important to notice that also an ensemble approach could have been taken, where a series of design events is obtained.

3 Maxima selection

In the previous section and study, the annual maxima were selected based on the maximum peak discharge and the volumes were only the corresponding, but not necessarily maximal ones. An alternate approach can be taken either based on the empirical copula or the adoption of multivariate distributions. For these, the same MRP definitions can be applied as quoted above and the largest values per year can be selected. In the following, we will follow this avenue and investigate the differences between these approaches where the copula C might be the empirical copula or an appropriate family.

For ease of notation, we will stick to bivariate events. We say that an event (x_1, x_2) is *OR-maximal*, if

$$F(x_1,x_2) = C(F_1(x_1),F_2(x_2)) \geq C(F_1(y_1),F_2(y_2)) = F(y_1,y_2)$$

for all (y_1, y_2) in the same temporal window (i.e. year) as (x_1, x_2). Analogously, we say that an event (x_1, x_2) is *AND-maximal*, if

$$1 - P(X_1 > x_1, X_2 > x_2) = F_1(x_1) + F_2(x_2) - C(F_1(x_1),F_2(x_2))$$
$$\geq F_1(y_1) + F_2(y_2) - C(F_1(y_1),F_2(y_2))$$

for all (y_1, y_2) in the same temporal window (i.e. year) as (x_1, x_2). Adopting also the Kendall and survival Kenedall return period definitions, we say that an event is *Kendall-maximal*, if

$$K(C(F_1(x_1),F_2(x_2))) \geq K(C(F_1(y_1),F_2(y_2)))$$

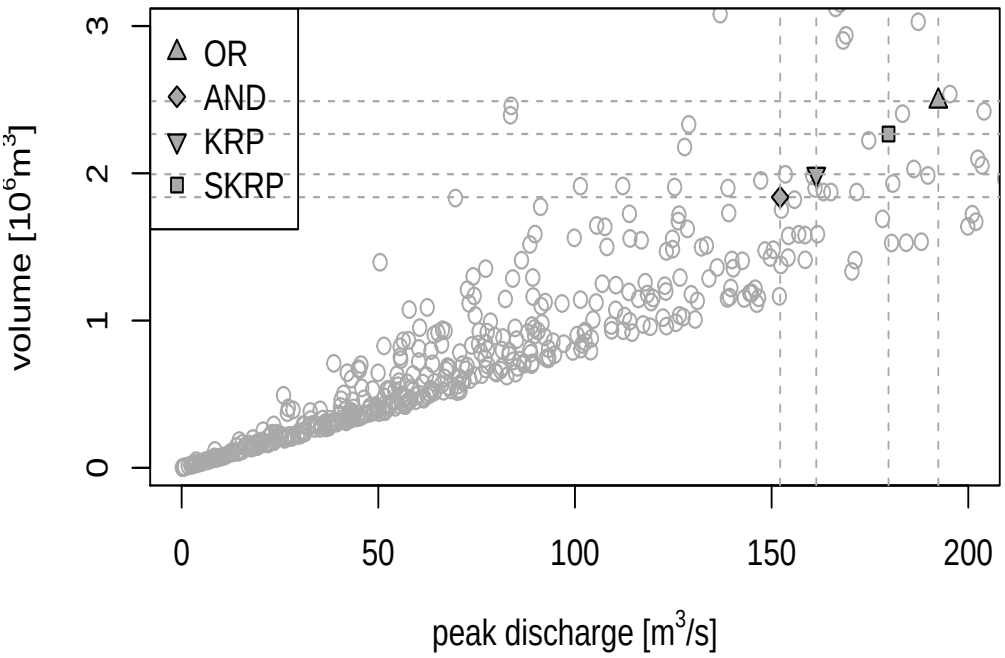

Figure 1. Comparison of different MRP definitions for a return period of 10 years (compare Figure 6 in Gräler et al. (2013)).

for all (y_1, y_2) in the same temporal window (i.e. year) as (x_1, x_2) where K is the Kendall distribution function corresponding to C. Finally, we call an event *survival Kendall maximal*, if

$$\bar{K}(\bar{C}(1-F_1(x_1), 1-F_2(x_2))) \geq \bar{K}(\bar{C}(1-F_1(y_1), 1-F_2(y_2)))$$

for all (y_1, y_2) in the same temporal window (i.e. year) as (x_1, x_2) where $\bar{K}$ is the survival Kendall distribution function with corresponding survival copula $\bar{C}$.

To study the impact of the aforementioned definitions, we use a second run of 500 simulated years of 5.625 minute resolution discharge data that were aggregated to separate rainfall events. This is different from the previous data set where only annual maxima have been used. This second data set contains 12466 events. In order to reduce the effect of autocorrelation within this simulation, we only consider a random subset of 50 % of the data (autocorrelation plots indicate an uncorrelated time series, not shown here). We do not fit any parametric family, and solely use the empirical definitions of the above equations.

In our simulated data set, the largest event in a year often is the same for all four definitions. This is not too surprising, considering that there are on average less than 25 rainfall events in each year. What remains different, is how extreme the event is for each of the four notions. The left plot of Figure 2 illustrates the relationship between the annual maximum value of the different definitions for the studied data set.

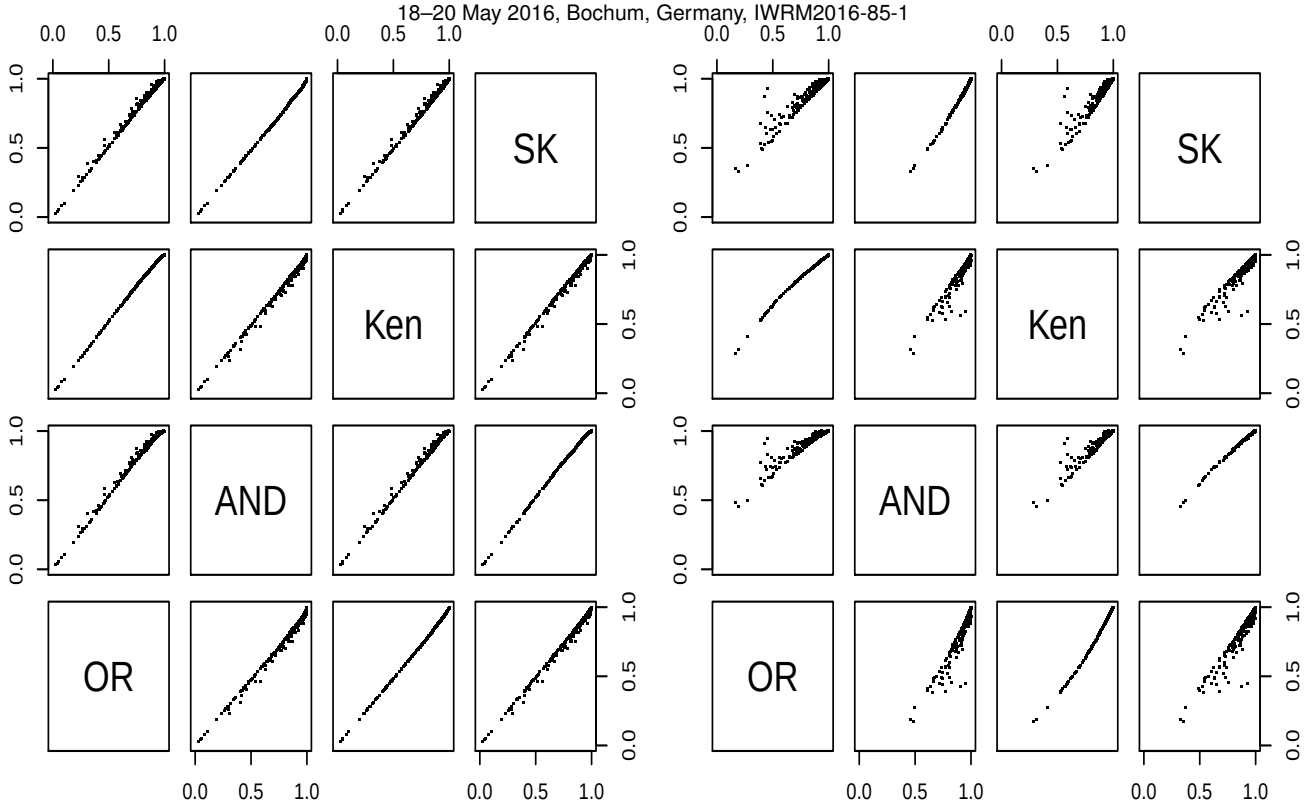

Figure 2. Comparison of the annual maximum value for the four different definitions. Left: based on 500 years of simulated rainfall data. Right: moderately correlated sample of a Gumbel copula.

Identifying the single rainfall events and looking into the marginal distributions, visually reveals identical histograms for peak discharge as well as for volume for the four bivariate and respective univariate maxima selections. An overlay shows only very little variations for discharge values and volumes. Larger values of the margins tend to even better coincide.

As this data set follows a very strong correlation, we draw a sample of a Gumbel copula with a moderate Kendall's tau of 0.6, assign it to the same temporal structure as our previous data set and repeat the above analysis. The right plot of Figure 2 is based on the copula sample and shows a larger variation in the annual maximum values for the four approaches. As in the left plot, the OR and KRP as well as the AND and SKRP approaches seem to be much more alike than the other pairwise combinations. The selected margins show a little more variability, but the histograms remain hardly distinguishable. All the variation appears in the center of the distribution.

4 Discussion and conclusion

The SKRP yields the most reliable separation into safe (sub-critical) and dangerous (super-critical) events. Nevertheless, the selection of a single design event, as often required by subsequent studies, remains an open question. Here, we selected the most probable bivariate event, but any event along the critical layer separates the sub- and super-critical regions.

The differences between the four definitions of maximality were minor in the simulated rainfall time series, but this is also due to the very strong correlation. This strong dependence causes the copula to be close the upper Fréchet-Hoeffding bound where all four definitions coincide. If all points lie close to the diagonal, there is no difference whether the critical layer follows the OR definition (enclosing the lower left rectangle), the SKRP definition (very sharply bend contour lines enclosing the lower left region), the KRP definition (very sharply bend contour lines excluding the top right region) or the AND definition (excluding the top right rectangle).

The temporal structure was not changed, only the dependence structure to investigate the effect. A less dry study area with much more rainfall events in a year will further influence the selection. However, the large extreme values appear to be extreme in each of the definitions.

Here we use the raw definitions of multivariate return periods, but an alternative would be to investigate derived measures. Requena et al. (2013) developed the *routed return period* where the water levels in a dam are used to characterize the return period of bivariate rainfall events. This idea could also be used in the same manner as presented in this paper to initially select the largest events from the original time series.

The investigated data set features a very long time series. Shorter time series might be more sensitive to changes of the maximum selection regime applied, as few events might have a strong influence on the selection of the marginal distributions. The influence of these outer properties needs to be further investigated. An avenue of future research is to consider the joint extremeness for the selection of extremes to be fed into a point over threshold approach.

References

Gräler, B., van den Berg, M., Vandenberghe, S., Petroselli, A., Grimaldi, S., De Baets, B., and Verhoest, N.: Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation, Hydrology and Earth System Sciences, 17, 1281–1296, doi:10.5194/hess-17-1281-2013, 2013.

Grimaldi, S., Petroselli, A., and Serinaldi, F.: A continuous simulation model for design-hydrograph estimation in small and ungauged watersheds, Hydrological Sciences Journal, 57, 1035–1051, doi:10.1080/02626667.2012.702214, 2012.

Nelsen, R. B.: An Introduction to Copulas, Springer Science+Buisness, New York, second edn., 2006.

Requena, A., Mediero Orduña, L., and Garrote de Marcos, L.: A bivariate return period based on copulas for hydrologic dam design: accounting for reservoir routing in risk estimation, Hydrology and Earth System Sciences, 17, 3023–3038, 2013.

Salvadori, G., De Michele, C., and Durante, F.: On the return period and design in a multivariate framework, Hydrology and Earth System Sciences, 15, 3293–3305, doi:10.5194/hess-15-3293-2011, 2011.

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

Sklar, A.: Fonctions de répartition á n dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris, 8, 229–231, 1959.