Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Threshold selection for regional peaks-over-threshold data

M. Roth, G. Jongbloed & T.A. Buishand

To cite this article: M. Roth, G. Jongbloed & T.A. Buishand (2015): Threshold selection for regional peaks-over-threshold data, Journal of Applied Statistics, DOI: 10.1080/02664763.2015.1100589

To link to this article: http://dx.doi.org/10.1080/02664763.2015.1100589

Published online: 03 Nov 2015.

Journal of Applied Statistics, 2015
http://dx.doi.org/10.1080/02664763.2015.1100589

Taylor & Francis
Taylor & Francis Group

Threshold selection for regional peaks-over-threshold data

M. Roth[a,b*], G. Jongbloed[c] and T.A. Buishand[b]

[a]*EURANDOM, Eindhoven University of Technology, Den Dolech 2, Eindhoven, The Netherlands;* [b]*Royal Netherlands Meteorological Institute (KNMI), Utrechtseweg 297, De Bilt, The Netherlands;* [c]*Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, Delft, The Netherlands*

(*Received 17 April 2014; accepted 23 September 2015*)

A hurdle in the peaks-over-threshold approach for analyzing extreme values is the selection of the threshold. A method is developed to reduce this obstacle in the presence of multiple, similar data samples. This is for instance the case in many environmental applications. The idea is to combine threshold selection methods into a regional method. Regionalized versions of the threshold stability and the mean excess plot are presented as graphical tools for threshold selection. Moreover, quantitative approaches based on the bootstrap distribution of the spatially averaged Kolmogorov–Smirnov and Anderson–Darling test statistics are introduced. It is demonstrated that the proposed regional method leads to an increased sensitivity for too low thresholds, compared to methods that do not take into account the regional information. The approach can be used for a wide range of univariate threshold selection methods. We test the methods using simulated data and present an application to rainfall data from the Dutch water board *Vallei en Veluwe*.

Keywords: peaks-over-threshold; regional frequency analysis; threshold selection; GPD; precipitation

1. Introduction

Extreme events, like financial shocks, storms or floods, often affect society disastrously. In order to take the right measures for their regulation, for example, by building a dyke of appropriate height, statistical inference about these extremes is crucial. In particular it might be important to predict which level of future extremes is to be expected. This can be analyzed by extreme value theory, which allows extrapolation beyond the range of observed data, based on asymptotically justified models [6,38].

Two common approaches for the study of extremes of independent and identically distributed (i.i.d.) random variables $X(t)$ $(t = 1, \ldots, T)$ are the block maxima (BM) and the peaks-over-threshold (POT) approach, see, for example, [6]. The BM approach considers the maxima of disjoint, equally sized blocks of the data, which are then usually modeled by a generalized extreme value (GEV) distribution. This is justified by the Fisher–Tippet–Gnedenko Theorem,

*Corresponding author. Email: roth@knmi.nl

see, for example, [12]. The POT approach on the other hand considers all excesses $Y(t) = X(t) - u$ over a previously chosen (high) threshold u. For $x \geq u$ we have, writing $y = x - u$,

$$P(X > x) = P(X > y + u) = P(X > u) \cdot P(X > y + u \,|\, X > u)$$
$$= P(X > u) \cdot P(Y > y \,|\, Y > 0).$$

The Balkema-de Haan-Pickands theorem states that the conditional distribution of the excesses Y can be approximated by a generalized Pareto distribution (GPD), if the threshold u is sufficiently high and certain regularity conditions hold, see, for example, [27]:

$$P(Y \leq y \,|\, Y \geq 0) \approx G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + \dfrac{\xi y}{\sigma}\right)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left(-\dfrac{y}{\sigma}\right), & \xi = 0, \end{cases}$$

for $y \geq 0$ if $\xi \geq 0$ and $0 \leq y \leq -\sigma/\xi$ if $\xi < 0$, where σ and ξ are the scale and the shape parameter. For $\xi = 0$ the GPD reduces to an exponential distribution.

From a theoretical point of view, the POT approach is often preferred over the BM approach, owing to the possibility to consider multiple extremes in a block, and the corresponding reduction in estimation uncertainty. However, in practice where i.i.d. data are the exception, the POT approach is less common than the BM approach. A reason for this is that the choice of the block size in the BM approach is often quite natural, for example, a year in many environmental applications, and block maxima are likely to be independent. For the POT approach the choice of the threshold u requires judgment and expertise. In fact, there is no single threshold choice technique that has been shown to work well in all applications. Further, declustering methods have been used to achieve that peaks are independent.

Threshold selection constitutes a trade-off situation between bias and variance. If the selected threshold is too low, the GPD approximation is poor. This causes bias in the estimated return level. However, with increasing thresholds the variance in the estimated GPD parameters increases, due to the small number of observations effectively used.

The aim of the present study is to guide threshold selection in the presence of multiple *similar* data samples. In this context regional frequency analysis (RFA) has been applied to reduce the variance in the parameter estimates. One commonly used variant of RFA is the index-variable method (or index-flood method, because it was introduced for flood data), which assumes that all site-specific distributions of the quantity of interest, that is, maxima or exceedances, are the same after scaling by an index variable [19]. The index-variable method has been applied mostly in combination with BM [13,16, 35], but there are several studies in combination with POT data [23,29,40] as well. However, to the knowledge of the authors, Roth *et al.* [30] were the first to propose combining multiple *similar* sites for the threshold selection itself.

The paper is organized in the following way. In Section 2 we explain the underlying assumptions of the regional POT model. In Section 3 we present a number of the available univariate threshold selection procedures. Based on these, new regional threshold selection approaches are introduced in Section 4. These approaches are tested in a simulation setting in Section 5, using smooth mixture densities with GPD tail. In Section 6 we apply the proposed methods to high quality precipitation measurements over the area of the *Vallei en Veluwe* water board in the Netherlands. The paper ends with the conclusions in Section 7.

2. Preliminaries

We represent the spatio-temporal data by the matrix $(X_s(t))_{s,t}$, where $s \in \{1, \ldots, S\}$ denotes the site and $t \in \{1, \ldots, T\}$ the time. For each site s the data $X_s(t)$, $1 \leq t \leq T$ are distributed according

to an unknown marginal distribution function F_s. We assume that for all sites s and for some small common exceedance probability $\zeta^0 = 1 - \tau^0$, the distribution of the excesses over the associated quantile $u_s^0 = F_s^{-1}(\tau^0)$ can be approximated by a GPD with scale parameter σ_s and shape parameter ξ_s, that is,

$$P(X_s - u_s^0 \leq y \,|\, X_s \geq u_s^0) \approx G_{\xi_s,\sigma_s}(y), \quad \forall s \in \{1, \ldots, S\}. \tag{H1}$$

Moreover, the following assumptions will be used regularly:

$$\xi_s \equiv \xi, \tag{H2}$$

$$\frac{\sigma_s}{u_s^0} \equiv \gamma, \tag{H3}$$

where γ denotes the dispersion coefficient. The assumption (H1) ensures that about the same amount of data is used for each location, which is always fulfilled in RFA of BM data. A common exceedance probability has often been considered in studies of multiple POT series, see, for example, [2,22,28].

The variance of the estimated shape parameter is in general very large, therefore assumption (H2) is often used in RFA, see, for example, [19]. Hanel *et al.* [16] introduce an index-variable model for BM based on the corresponding assumptions (H2), (H3) for the GEV distributions, that is, a constant shape parameter and dispersion coefficient over the region. Roth *et al.* [29] deduce from this a regional index-variable POT model relying on assumptions (H1)–(H3). This model will be used also in the simulation setting, yet most of the methods developed in the following rely only on assumption (H1).

Spatial data usually exhibit dependence. Therefore, spatial dependence should be taken into account for the quantitative approaches to threshold selection. As for the marginal distributions, we assume that the dependence structure does not change over time, that is, each row of the matrix $(X_s(t))_{s,t}$ has the same dependency structure. To strictly formulate this we rely on the notion of copulas, which are multivariate distributions with standard uniform margins. Sklar's theorem, see, for example, [24], ensures, that every multivariate distribution function can be disentangled into its marginal distributions and a copula. Therefore, we assume, that there exists a copula C, such that $\forall t \in \{1, \ldots, T\}$:

$$P(X_1(t) \leq x_1, \ldots, X_S(t) \leq x_S) = C(F_1(x_1), \ldots, F_S(x_S)). \tag{1}$$

3. Threshold selection

Scarrott and MacDonald [31] provide a review of existing threshold selection methods, here we focus only on two types of commonly used approaches.

3.1 *Approaches based on visual inspection*

The threshold stability (TS) plot and the mean excess (ME) plot (also referred to as mean residual life plot) are widely used graphical tools for the selection of the threshold in POT analyses, see, for example, [6,12]. The TS plot is based on the fact that once the GPD model holds at grid point s for some threshold u_s^0, it holds for every threshold $u \geq u_s^0$. In particular we have that the associated shape parameter $\xi_s(u)$ is the same for u and u_s^0. This property is called TS. The TS can be exploited by estimating the shape parameter for a range of high thresholds and plotting

$$(u, \hat{\xi}_s(u))_{u \geq 0},$$

where $\hat{\xi}_s(u)$ denotes the estimated shape parameter based on the excesses over threshold u. If the GPD model holds for u_s^0 the graph should be (approximately) constant for $u \geq u_s^0$. However,

owing to the decreasing number of excesses above higher thresholds, the plot becomes unstable and the constant behavior is difficult to see. In the following we will use a slightly modified version of the TS plot, that is,

$$(\tau, \hat{\xi}_s(\tau))_{\tau\in[0,1)},$$

where $\hat{\xi}_s(\tau)$ is a short notation for $\hat{\xi}_s(\hat{F}_s^{-1}(\tau))$ and $\hat{F}_s^{-1}(\tau)$ denotes the empirical τ-quantile.

In order to support proper selection of the threshold, it is recommended to also consider other available diagnostic plots. The ME plot relies on a similar consideration but utilizes the mean excess function

$$e_s(u) := \frac{1}{1 - F_s(u)} \int_u^\infty (x - u)\, \mathrm{d}F_s(x).$$

The empirical version of the ME function is given by

$$\hat{e}_s(u) = \frac{\sum_{t=1}^{\mathrm{T}} (x_s(t) - u)\mathbb{I}_{(u,\infty)}(x_s(t))}{\sum_{t=1}^{\mathrm{T}} \mathbb{I}_{(u,\infty)}(x_s(t))},$$

where $\mathbb{I}_A$ is the indicator function for set A, that is, $\mathbb{I}_A(x) = 1$ if $x \in A$ and otherwise zero. If the GPD model holds, $e_s(u)$ is linear in the threshold u, with the slope being determined by the shape parameter ξ, see, for example, [12], and $\hat{e}_s(u)$ becomes approximately linear. However, as described in greater detail in [14], there are some problems associated with the use of the ME plot. These are in particular that the ME function is only well defined for $\xi < 1$, and – as for the TS plot – the empirical ME function becomes unstable for high values of the threshold. In many applications only values of $\xi < 1$ are found, in which case the ME plot is well defined.

3.2 *Approaches based on GoF tests*

A more quantitative way to select the threshold is to use a goodness of fit (GoF) test. Classical tests are the Kolmogorov–Smirnov (KS) test and the Anderson–Darling (AD) test, see, for example, [5,9,11,26]. Both tests can be used to test the composite hypothesis, that the n excesses over a fixed threshold u follow a GPD. The KS statistic is defined as

$$D_n = \sup_y |F_n(y) - \hat{G}(y)|, \tag{2}$$

where $\hat{G}$ is the estimated GPD from the n excesses, and F_n the empirical distribution of the n excesses. The AD statistic is defined by

$$A_n = n \int_{-\infty}^{\infty} \frac{(F_n(y) - \hat{G}(y))^2}{[\hat{G}(y)(1 - \hat{G}(y))]} \mathrm{d}G(y). \tag{3}$$

Because there is typically only a limited amount of data and the sample size n decreases even further with increasing threshold, threshold selection is a real challenge. For the univariate case there exist tables with the critical values, which depend on the (estimated) shape parameter, see, for example, [5]. Notice that their critical values for finite samples were obtained by Monte Carlo experiments. We will also obtain critical values via simulation using two slightly different simulation approaches:

(1) Fix τ^0 and the corresponding threshold u.
(2) Estimate the parameters based on the n excesses above u.
(3a) Simulate n independent random variables from the GPD distribution with the parameters estimated in step 2.

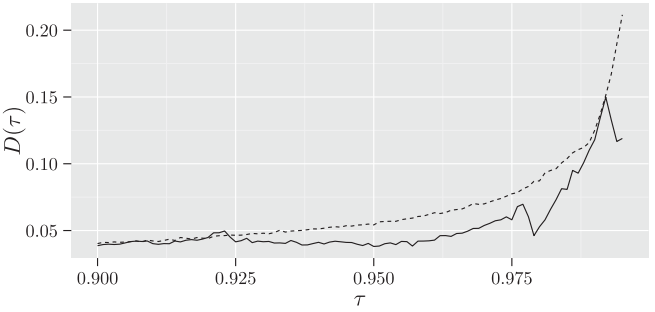

Figure 1. KS statistic for daily rainfall at *Putten* in the Netherlands (see Figure 11) as a function of the probability τ (solid line). The critical values (dashed line) are obtained as the 95 percentage point of the simulated KS test statistic from step 3b in the simulation algorithm.

(3b) Sample $\tilde{n}$ from a binomial distribution of size T and success probability $(1 - \tau^0)$. Simulate $\tilde{n}$ GPD random variables with the estimated parameters.
(4) Calculate for the simulated data the KS and AD statistic as before.
(5) Repeat steps 3 and 4 a thousand times and take the 0.95-quantile of the bootstrapped statistic as critical value for D_n (or A_n) at τ^0.
(6) The procedure can be repeated for every τ in a reasonable range.

In step 3a, we simulate conditional on the observed number of excesses. However, if we would have different samples from the same population the proportion above the threshold would probably be different. This is accounted for in step 3b, which is equivalent to generating T standard uniform random variables and transforming the values exceeding τ^0 to GPD data. In the regional setting this approach is generalized for multiple dimensions. Figure 1 shows a typical plot of the KS statistic. It illustrates the difficulties in the selection of the threshold, owing to the low statistical power of the test given the broad alternative [9]. For instance, the 0.9-quantile is not rejected, but slightly higher quantiles are. The 0.925-quantile might be considered, because no higher quantile is rejected.

4. Threshold selection in the regional POT model

The general idea behind the proposed regional approach is to use spatial averaging in the selection process. Although we demonstrate the regionalization only with the threshold selection methods, described in Section 3, it can be applied to a variety of other threshold selection approaches, for example, the approach proposed in [37].

4.1 *Approaches based on visual inspection*

The following version of the TS plot only assumes (H1). For $\tau \in [0, 1)$ we estimate the corresponding site-specific shape parameters $\xi_s(\tau)$, $s \in \{1, \ldots, S\}$ based on the excesses over the threshold $u_s(\tau) := \hat{F}_s^{-1}(\tau)$. Then, we can consider the spatially averaged TS plot

$$(\tau, \hat{\xi}_{\text{avg}}(\tau))_{\tau \in [0,1)}, \tag{4}$$

where $\hat{\xi}_{\text{avg}}(\tau) = S^{-1} \sum_{s=1}^{S} \hat{\xi}_s(\tau)$. Due to the fact that the individual plots $(\tau, \hat{\xi}_s(\tau))$ should be roughly constant for $\tau \geq \tau^0$, we have that the averaged TS plot in Equation (4) is approximately constant too. Moreover, due to the averaging, small-sample size effects are reduced and the

constant behavior becomes more apparent. Assuming also (H2) leads to a variant by considering a common estimate for the shape parameter instead of $\hat{\xi}_{\text{avg}}$, see [29]. In the case of unequal sample sizes weighting the $\hat{\xi}_s(\tau)$ proportionally to the sample sizes could be considered [19].

The empirical ME function is linear in the threshold itself and not in the corresponding probability τ. Therefore, we average not only the excesses but also the thresholds for a common τ. Here the assumption (H2), which implies that the slopes in the single ME plots are the same, is critical, for the following reason. Consider the two linear functions:

$$y_i^{(1)} = au_i + b \quad \text{and} \quad y_i^{(2)} = cv_i + d.$$

Then we can not say much about their average as function of the average argument, because

$$\frac{(y_i^{(1)} + y_i^{(2)})}{2} = \frac{(au_i + cv_i)}{2} + \frac{(b+d)}{2}.$$

However, with $a = c$ we have that the average is again linear with the same slope:

$$\frac{(y_i^{(1)} + y_i^{(2)})}{2} = \frac{a(u_i + v_i)}{2} + \frac{(b+d)}{2}.$$

This gives rise to the spatially averaged ME plot

$$(u_{\text{avg}}(\tau), \hat{e}_{\text{avg}}(\tau))_{\tau \in [0,1)}, \tag{5}$$

where $u_{\text{avg}}(\tau) = S^{-1} \sum_{s=1}^{S} u_s(\tau), \hat{e}_{\text{avg}}(\tau) = S^{-1} \sum_{s=1}^{S} \hat{e}_s(\tau)$ and $\hat{e}_s(\tau)$ is the ME at site s evaluated at level $u_s(\tau)$. The averaged ME plot should be approximately linear for $\tau \geq \tau^0$, and the slope is determined by the common shape parameter ξ. The strength of the spatially averaged plots lies in the increased detection probability of non-constant behavior in the TS plot and non-linear behavior in the ME plot, when the threshold is too low.

4.2 *Approaches based on GoF tests*

We want to test the hypothesis (H1), that is, that the GPD behavior sets in at $F_s^{-1}(\tau^0)$ for all sites $s \in \{1, \ldots, S\}$. Douglas *et al.* [10] evaluate the significance of trends in flood data across the USA by averaging the test statistic regionally. We follow this idea and average the KS statistic, respectively the AD statistic, computed as in Section 3.2, over the available sites. Also here, weighting could be considered in the case of unequal sample sizes. Averaging reduces the variance, which leads to an increased sensitivity for too low thresholds. However, there is no standard way to obtain critical values for the averaged test statistic and one has to rely on simulation.

For the simulation of the spatial dependence we use copulas. Assume the data-generating copula C from Equation (1) is known, then one can use this copula for the computation of the critical values without knowing the full marginal distribution. Say we want to compute critical values for the exceedance probability $\zeta^0 = 1 - \tau^0 = 0.05$. First we generate marginally uniform data from the copula C, for example, for 4 sites and 5 days

$$\begin{pmatrix} 0.11 & 0.23 & 0.45 & 0.96 \\ 0.97 & 0.98 & 0.11 & 0.05 \\ 0.02 & 0.45 & 0.23 & 0.61 \\ 0.89 & 0.92 & 0.96 & 0.95 \\ 0.93 & 0.96 & 0.98 & 0.76 \end{pmatrix}$$

and keep only those entries exceeding τ^0, that is, in our example

$$\begin{pmatrix} - & - & - & 0.96 \\ 0.97 & 0.98 & - & - \\ - & - & - & - \\ - & - & 0.96 & - \\ - & 0.96 & 0.98 & - \end{pmatrix}.$$

These are then transformed to GPD data using the marginal distribution function with the estimated parameters. For each site s this is equivalent to the unconditioned simulation of GPD random variables in Section 3.2. Note that owing to the construction process, we are only interested in events, where at least one component exceeds the at-site τ^0-quantile, that is,

$$1 - C(\tau^0, \ldots, \tau^0) = P(F_1(X_1) > \tau^0 \text{ or } \cdots \text{ or } F_S(X_S) > \tau^0). \tag{6}$$

Finally, we determine for each site the test statistic and compute the average over all sites. Repeating the procedure, for example, 1000 times, leads to critical values for the averaged test statistic that take the spatial dependence into account. In general the copula is not known and has to be estimated. Because copula estimation lies outside the focus of the present study, we use fixed copulas in the simulation section and use a heuristic estimate in the application section, to describe the dependence at high quantiles.

5. Simulation study

The evaluation of a threshold selection method is challenging. If the bulk distribution and the GPD tail are quite different the threshold can easily be detected, maybe also with univariate methods [39]. In most cases, however, a simple discrimination between bulk and tail distribution is not possible. Using daily rainfall data from 15,029 stations around the world, Papalexiou and Koutsoyiannis [25] showed that the GPD generally described the tail of the distribution better than the Weibull and Gamma distributions. However, in a subsequent study [33], it was observed that the fit of the GPD becomes worse if non-extreme values are progressively incorporated in the POT sample. Several extreme value mixture models have been proposed in the literature, which combine the GPD with a component to capture the non-extreme part of the distribution [31]. To explore the ability to detect a theoretical threshold, we simulate data from a mixture distribution with differentiable density. Differentiability is imposed to get a smooth transition between the bulk of the distribution and the GPD tail.

5.1 *Marginal model*

Any continuous probability distribution function F on $[0, \infty)$ can be written as

$$F(x) = 1 - \exp(-H(x)) = 1 - \exp\left(-\int_0^x h(u)\,\mathrm{d}u\right), \tag{7}$$

where (H) h is the (cumulative) hazard rate. The density is given by

$$f(x) = h(x)\exp(-H(x)),$$

which is of (differentiability) class C^k, whenever h is of class C^k.

A distribution with differentiable density and GPD tail above threshold u can be obtained by defining the following hazard rate

$$h(x) := \eta\left(\frac{x-u}{\varepsilon}\right) h_1(x) + \left(1 - \eta\left(\frac{x-u}{\varepsilon}\right)\right) h_2(x), \tag{8}$$

where $\varepsilon > 0$ is the length of the transition interval, h_1 the hazard rate of a general bulk distribution with differentiable density, and h_2 the hazard rate of a GPD with shape parameter ξ and scale parameter σ, given by

$$h_2(x) = \frac{1}{\sigma + \xi(x-u)}, \quad x \geq u, \tag{9}$$

and $h_2(x) = 0$ otherwise. The use of the transition function

$$\eta(x) = \begin{cases} 1, & x \leq 0, \\ 2x^3 - 3x^2 + 1, & 0 < x < 1, \\ 0, & x \geq 1, \end{cases} \tag{10}$$

see Figure 2, with $\eta(0) = 1$ and $\eta'(0) = \eta'(1) = \eta(1) = 0$, yields a hazard function h with continuous derivative h'. Using a linear transition function would yield only a continuous hazard function. Note that the tail, the part of the distribution to the right of u, is contaminated over the interval $(u, u + \varepsilon)$. This construction of a mixture GPD distribution is similar to that proposed by Holden and Haug [18], based on truncated distribution functions.

For a fixed exceedance probability $\zeta^0 = 1 - \tau^0$, for example, 5%, the corresponding threshold is defined via the following relation:

$$\zeta^0 = \exp(-H_1(u)) \Longleftrightarrow u = H_1^{-1}(-\ln(\zeta^0)). \tag{11}$$

The transition can be smoothed further by considering the following restriction on Mill's ratio $\phi(x) = 1/h(x)$ at $x = u$:

$$\phi'(u) = -\frac{h_1'(u)}{(h_1(u))^2} = \xi. \tag{12}$$

This is motivated by the theory of penultimate approximations [34,38].

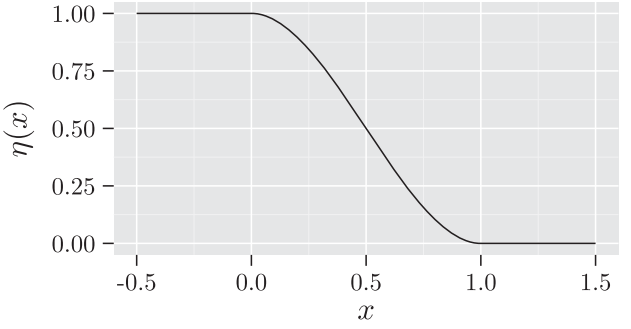

Figure 2. Smooth transition function.

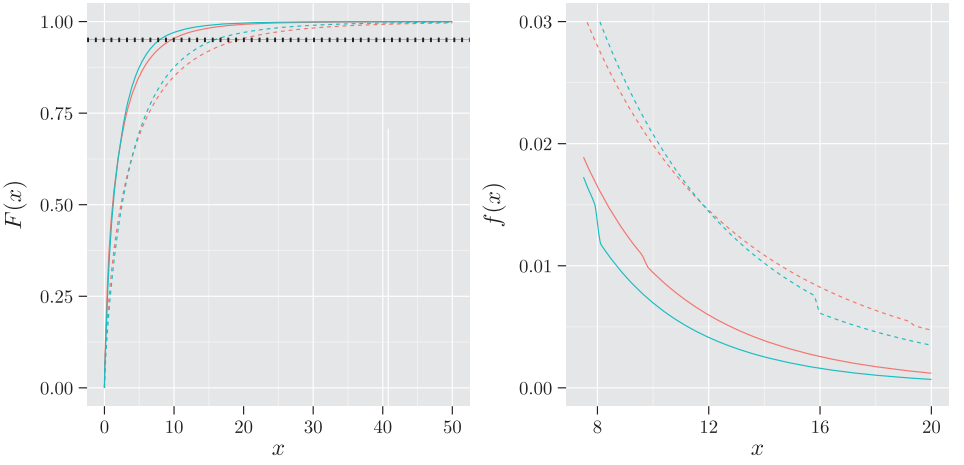

Figure 3. Left: Cumulative distribution function for the hybrid Weibull GPD model. The dotted horizontal line indicates the probability τ^0 where the GPD tail sets in. Right: zoom into the transition part of the density ($\xi = 0.15$, $\gamma = 0.5$, $\zeta^0 = 0.05$, $\varepsilon = 0.25$, $\kappa = 0.7/0.8$ (red/blue), $\beta = \frac{2}{4}$ (solid/dashed)) [colour online].

As bulk distribution we consider the Weibull distribution

$$W_{\kappa,\beta}(x) = 1 - \mathrm{e}^{-(x/\beta)^{\kappa}}, \tag{13}$$

with shape parameter κ and scale parameter β. The hazard rate h_1 of the Weibull distribution is given by

$$h_1(x) = \frac{\kappa}{\beta}\left(\frac{x}{\beta}\right)^{\kappa-1}.$$

For this distribution $u = \beta(-\ln(\zeta^0))^{1/\kappa}$. Restriction (12) implies then the following relationship between the Weibull and GPD shape parameters:

$$\xi = \frac{1-\kappa}{\kappa}\left(\frac{u}{\beta}\right)^{-\kappa} = \frac{1-\kappa}{\kappa}(-\ln(\zeta^0))^{-1}.$$

Figure 3 shows the cumulative distribution and a zoom into the transition part of the density for one set of GPD parameters and varying Weibull parameters. The GPD shape parameter is here 0.15, which is a representative value for daily rainfall extremes [20]. We observe that the transition is much smoother for $\kappa = 0.7$, than for $\kappa = 0.8$. The reason is, that for $\kappa = 0.7$ and $\xi = 0.15$ Equation (12) is approximately true, which is not the case for $\kappa = 0.8$ and the same ξ.

5.2 *Spatially independent data*

After defining the marginal model we first investigate the influence of the number of considered locations on the selected threshold. Therefore, we simulate spatially and serially independent data for 16 locations and 4600 days, for example, 50 summer seasons. We took for all sites $\tau^0 = 1 - \zeta^0 = 0.95$, $\varepsilon = 0.25$, $\gamma = 0.5$, and $\xi = 0.15$. The Weibull scale parameters β_s were obtained as a random sample from the uniform distribution between 2 and 4 and the Weibull shape parameters κ_s as a random sample from the beta distribution, with parameters 2 and 5, shifted by 0.5 (mode at 0.7). We denote this parameter set by I.

Figure 4 shows the TS plot for one site and the spatially averaged version for 8 and 16 sites for the simulated data. The individual TS plot shows no stabilization, it increases sharply after an

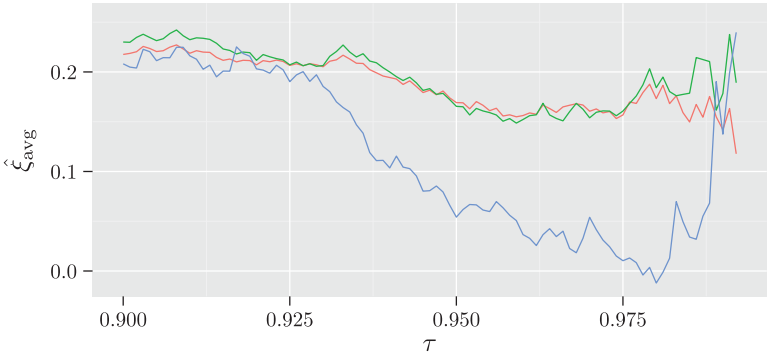

Figure 4. TS plot for one site (blue), 8 sites (green) and 16 sites (red) for simulated data from the hybrid Weibull GPD distribution (parameter set I, no spatial dependence) [colour online].

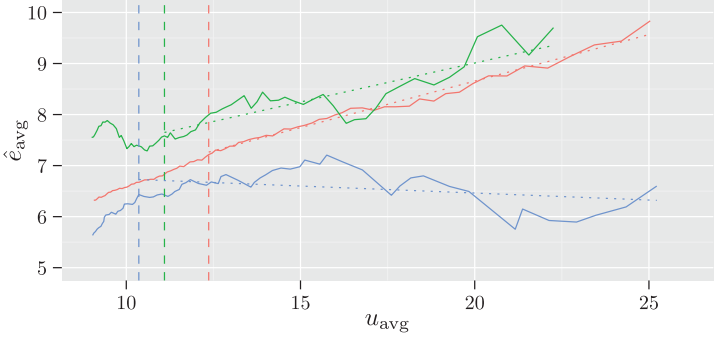

Figure 5. ME plot for one site (blue), 8 sites (green) and 16 sites (red) for simulated data from the hybrid Weibull GPD distribution (parameter set I, no spatial dependence). The dashed vertical lines indicate the corresponding (mean) 0.95-quantile. The dotted lines show the fitted linear regression above the (mean) 0.95-quantile [colour online].

almost linear decrease up to $\tau = 0.98$. The averaged at-site estimate stabilizes around $\tau = 0.96$ for 8 sites and slightly earlier for 16 sites. Moreover, we observe that the variation in the averaged shape parameter estimate reduces substantially from 8 to 16 sites.

Figure 5 shows the (spatially averaged) ME plot for the same data. Here, the reduced variance of the averaged at-site estimates for 8 and in particular for 16 sites becomes even more apparent. One observes only a small change in the slope of the spatially averaged ME plot for 16 sites around the averaged 0.95-quantile, that is, it decreases from 0.25 to 0.19.

For the quantitative approach based on GoF tests, we first have to define a threshold selection criterion. We select the smallest $\tau \in [0.9, 1)$, such that the test statistic is smaller than the 95 percentage point of the bootstrapped test statistic. For instance, for the precipitation data from *Putten* in Figure 1 the selected probability τ would be 0.9, because the value of the KS test statistic is below the corresponding 95 percentage point for $\tau = 0.9$, as shown in Figure 1. The adequacy of the proposed procedure is explored by simulating $B = 1000$ samples for the 16 locations of the same size from the same marginal distributions. For each simulated sample the threshold was determined using the criterion above for single site data and for the averaged KS statistic over 2, 4, 8, and 16 sites. The GPD parameters were estimated for each single site but in the case of averaging over 16 sites, joint estimation was also considered, using the approach described in [29].

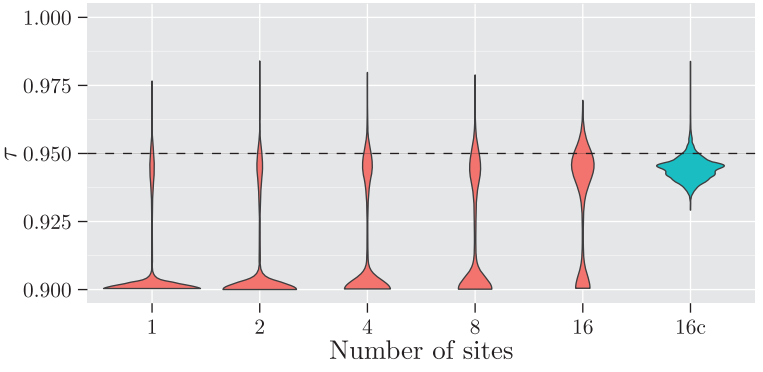

Figure 6. Violin plots [17] of the selected probability τ based on the 95% critical value of the KS test for the simulated data from the hybrid Weibull GPD distribution (parameter set I, no spatial dependence). Plots are shown for single site data and various groups of multiple sites. The right plot is based on joint estimation of the GPD parameters. The dashed horizontal line indicates the true probability τ^0 [colour online].

Figure 6 shows the influence of the number of considered sites on the selected threshold. The more sites are considered the closer the selected threshold is to the *true one*. This can be explained by the increasing power of the GoF test, when using more data. Moreover, the figure shows the advantage of using the regional similarity in the parameter estimation. In the case of regional estimation the mean selected τ is 0.945, that is, the mean exceedance probability is 5.5%, which is a strong improvement compared to the mean exceedance probability obtained for the univariate samples. Note that strong overestimation of the exceedance probability in the univariate case is also reported for different threshold selection methods, see, for example [15,31].

After looking at the influence of the number of locations, we investigate how the threshold selection method performs, when bulk and tail distribution are alike near the threshold level. Therefore, we simulate 1000 samples using the same marginal parameters as before except for the Weibull shape parameter, which we specify such that Equation (12) holds, that is, $\kappa = 0.69$. We denote this parameter set by II. For this parameter set, the 0.9-quantile is selected as threshold in more than 90% of the simulated samples, even when considering all 16 sites. Due to the small differences between bulk and tail distribution the method is not able to select the theoretical threshold.

Focusing on the threshold may not be optimal in terms of bias and variance of estimated parameters or return levels. The bias-variance trade-off was studied for return levels. Denote the vector of the true return levels by $\boldsymbol{r}$ and the estimated return levels associated with non-exceedance probability τ from sample i by $\hat{\boldsymbol{r}}_i(\tau)$. Then we compute for each τ the averaged Euclidean error over the B simulated samples:

$$\mathrm{AEE}(\tau) = \frac{1}{B}\sum_{i=1}^{B} ||\hat{\boldsymbol{r}}_i(\tau) - \boldsymbol{r}||_2, \tag{14}$$

where $||.||_2$ denotes the Euclidean (or l_2) norm. Figure 7 shows $\mathrm{AEE}(\tau)$ for the 50-year return level. For parameter set I $\mathrm{AEE}(\tau)$ reaches a minimum around $\tau = 0.95$. This nicely demonstrates the trade-off, if the threshold is too low a large bias is introduced, and if the threshold is too high the variance dominates. However, for parameter set II the AEE is almost constant up to $\tau = 0.945$ and it increases afterwards. This means there is hardly any bias introduced by selecting lower thresholds, which highlights the effect of the restriction posed in Equation (12) on the smoothness of the transition between bulk and tail distribution. Though, there is large uncertainty in the

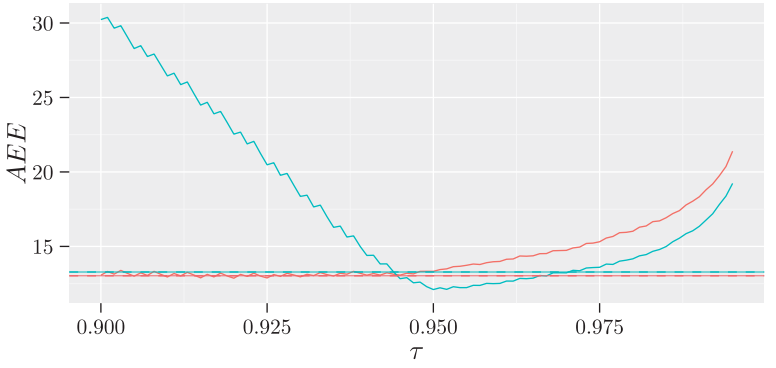

Figure 7. Averaged Euclidean error in the 50-year return level as a function of the probability τ for the simulated data from the hybrid Weibull GPD distribution (no spatial dependence, blue – parameter set I, red – parameter set II). For both parameter sets the GPD parameters were jointly estimated for all 16 sites. The dashed horizontal lines indicate the averaged Euclidean error using the selected threshold from the KS test (AEE^0) [colour online].

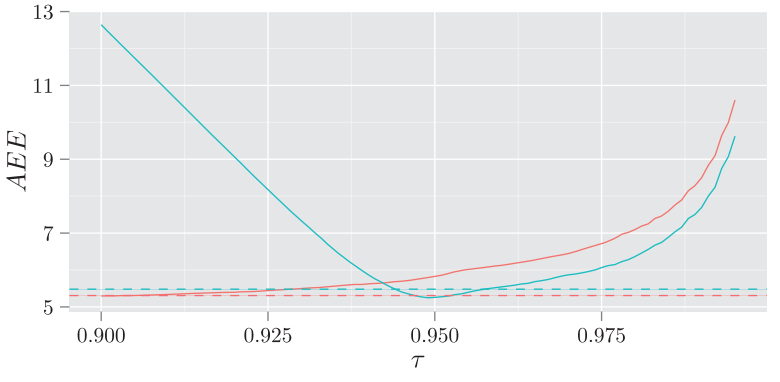

Figure 8. Same as Figure 7 but for the 5-year return level [colour online].

threshold level for parameter set II, the choice of the threshold will probably have not so much influence on the return level estimate and its mean square error for this parameter set. A similar picture, in particular for the AEE in the 5-year return level, is obtained if we generate data from a Weibull distribution with common shape parameter $\kappa = 0.7$ and the same spatial variation of the scale parameter as in parameter sets I and II. The 0.9-quantile is selected in more than 93% of the samples in that case.

We are interested how the averaged Euclidean error AEE^0 of the estimated return level, based on the automatically chosen probability $\tau(i)$, compares to the minimal AEE over all τ. AEE^0 can be computed by replacing $\hat{\boldsymbol{r}}_i(\tau)$ by $\hat{\boldsymbol{r}}_i(\tau(i))$ in Equation (14). In Figure 7, we see that AEE^0 is for both parameter sets close to the minimum value of $AEE(\tau)$. The situation for the 5-year return level is very similar, see Figure 8. The results in Figures 7 and 8 apply to the KS test. The AD test results in slightly smaller values of AEE^0 as well as $AEE(\tau)$ in general, but the observed differences are smaller than 2%.

We compared several significance levels for both tests. The values of AEE^0 are very similar for significance levels between 5 % and 20 % , but increase outside this interval.

Table 1. Gumbel copula parameter θ and normal copula parameter ρ used in the simulation of spatially dependent data (only one data set without spatial dependence was simulated).

$l^u(0.9)$	0.1	0.25	0.5	0.75	0.9
θ	1	1.1514	1.5994	2.9254	6.8769
ρ	0	0.3686	0.7366	0.9358	0.9898

5.3 *Spatially dependent data*

The modeling of spatial dependence in general and in particular that in extremes is a very active field of research, see, for example [8,32]. The dependence between small to medium values should have no influence on the threshold selection. Therefore, classical dependence measures like correlation and Kendall's tau, that are usually estimated from all data, are not appropriate. In extremes the notion of tail dependence arises. A bivariate random vector (X_1, X_2) with marginal distributions F_1 and F_2, is said to be asymptotically dependent if [7]

$$l^u := \lim_{v \uparrow 1} P(X_1 > F_1^{-1}(v) \,|\, X_2 > F_2^{-1}(v)) > 0. \tag{15}$$

Similarly, we say that the bivariate vector is asymptotically independent if $l^u = 0$. For many copulas one can derive l^u directly from the copula parameters. Consider, for example, the Gumbel copula with parameter $\theta \in [1, \infty)$, given by

$$C_\theta(u, v) = \exp(-[(-\ln(u))^\theta + (-\ln(v))^\theta]^{1/\theta}), \tag{16}$$

which can be easily generalized to d dimensions. Here, the tail dependence can be easily computed as $l^u = 2 - 2^{1/\theta}$, which is greater than 0 if $\theta > 1$ [24]. However, it is known that a bivariate normal random vector is asymptotically independent for every correlation coefficient $\rho \in [-1, 1)$. Hence, we conclude that a quantile-dependent measure of dependence, for example,

$$l^u(\tau) := P(X_1 > F_1^{-1}(\tau) \,|\, X_2 > F_2^{-1}(\tau)) = \frac{P(X_1 > F_1^{-1}(\tau), X_2 > F_2^{-1}(\tau))}{P(X_1 > F_1^{-1}(\tau))}, \tag{17}$$

might be more appropriate than a correlation coefficient to describe the influence of spatial dependence on threshold selection. $l^u(\tau)$ can be simply estimated from the data by counting the excesses over the sample τ-quantile.

We simulate spatially dependent data from the hybrid Weibull GPD distribution for 16 sites using parameter set I. To determine the influence of the dependence structure and in particular of tail dependence on the selection of the threshold, we choose to simulate from a 16-dimensional Gumbel copula with parameter θ and a 16-dimensional normal copula with the same correlation ρ for all pairs. To obtain a fair comparison between the two copulas we prescribe fixed values of $l^u(\tau)$ for $\tau = 0.9$, the lowest possible probability in our setting, and select the copula parameters correspondingly. We fix the following values for $l^u(0.9)$: 0.1 (which corresponds to independent data), 0.25, 0.5, 0.75, and 0.9. The corresponding parameters for the Gumbel and normal copula are shown in Table 1.

Figure 9 shows violin plots of the selected non-exceedance probability τ for different levels of dependence. We see that the method is quite able to select a threshold close to the true 0.95-quantile. Moreover, we see that there is no clear dependence of the selected threshold on the amount of spatial dependence in the data. In contrast to the selected threshold, AEE^0 increases strongly with $l^u(0.9)$ and is generally larger for the Gumbel copula than for the normal copula, see Figure 10.

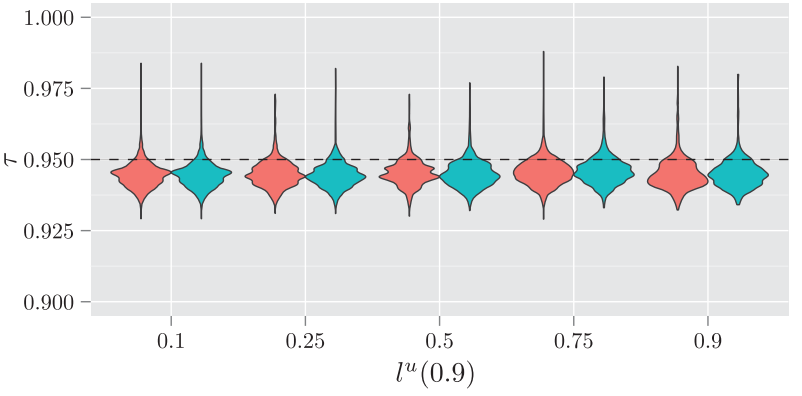

Figure 9. Selected probability τ for different dependence levels $l^u(0.9)$, based on the 95% critical value of the KS test using joint parameter estimation, for simulated data from the hybrid Weibull GPD distribution (red – Gumbel copula, blue – normal copula) [colour online].

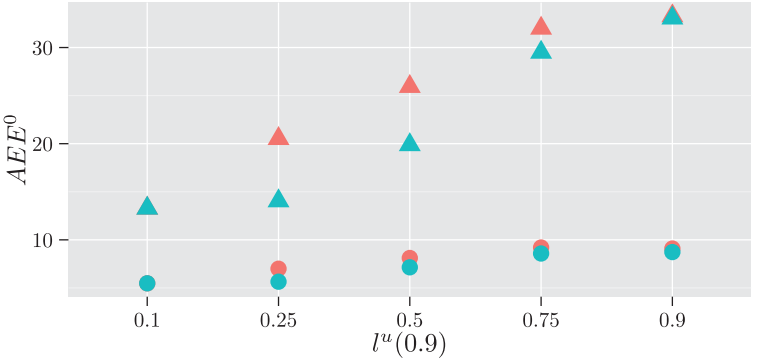

Figure 10. AEE^0 of the 5- (dots) and 50-year (triangles) return level for simulated data from the hybrid Weibull GPD distribution (red – Gumbel copula, blue – normal copula) [colour online].

6. Application to rainfall data

Figure 11 shows the area of the *Vallei en Veluwe* water board in the Netherlands, together with the locations of the 21 rainfall stations in this area. For each station we have summer (JJA) records of daily precipitation available for the period 1951–2009, that were corrected for inhomogeneities owing to changes in local measurement conditions [4]. Thus, we obtain a rainfall matrix $(X_s(t))_{s,t}$ with $s \in \{1, \ldots, 21\}$ and $t \in \{1, \ldots, 5428\}$.

In order to exclude temporal dependence, we apply first the declustering algorithm, described in [30]. Owing to the weak temporal dependence in summer, a separation time of 1 day is taken as sufficient [21]. Figure 12 shows the spatially averaged TS plot for the declustered data. The shape parameter estimate is either determined as the average of the individual maximum likelihood estimates or as common estimate, using the estimation approach in [29]. For the latter approach we rely on assumptions (H2, H3), introduced in Section 2. Both approaches show a strong increase of the estimated shape parameter up to the 0.96-quantile, which indicates that one should consider thresholds above the 0.96-quantile. The common estimate is more stable at large τ than the average at-site estimate, thus allows for a better selection of τ^0. The smaller value of the average at-site estimate of ξ is probably due to the bias in this estimate. At the 0.96-quantile, the at-site estimate is based on 217 excesses. For a sample size of 200 and a shape

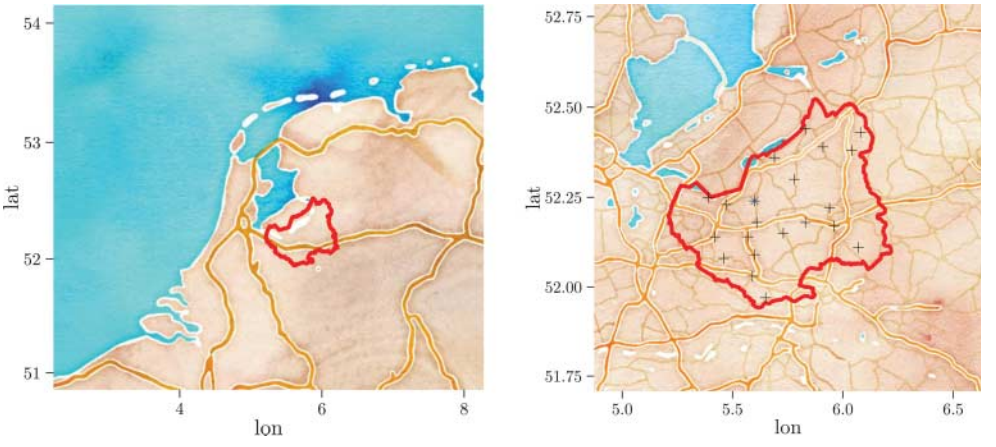

Figure 11. Left: Map of the Netherlands with the area of the *Vallei en Veluwe* water board surrounded by a red line. Right: a zoom into the map, the pluses indicate the individual measurement locations. The blue star is the measurement site *Putten* [colour online].
Source: The map is based on the watercolor map by Stamen Design (http://www.stamen.com).

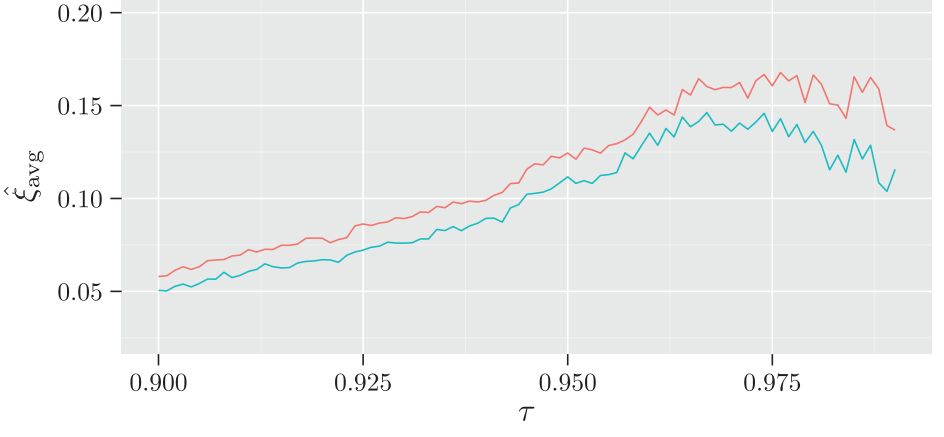

Figure 12. Spatially averaged TS plot for the rainfall data. The estimate of the shape parameter is obtained by averaging maximum likelihood estimates of the GPD parameters for the individual sites (blue) or by using joint estimation of the GPD parameters (red) [colour online].

parameter of 0.25, Zhang and Stephens [41] report a bias of -0.013 for the maximum likelihood estimate.

Using the criterion defined in Section 5.2 and the KS test individually for each site, we obtain $\tau < 0.91$ for all but one site. Figure 13 shows the 5-, 50-, and 500-year return levels based on joint estimation of the parameters as a function of the non-exceedance probability τ. While the 5-year return level stays virtually constant, the 50- and 500-year return levels show an increase up to $\tau = 0.96$, which was chosen using the spatially averaged TS plot. Thus, there is a risk of underestimating high return levels if local thresholds near $\tau = 0.9$ are used.

To apply the quantitative approach to threshold selection in the regional setting, we have to model the spatial dependence. Buishand [3] concludes that annual maximum rainfalls over the Netherlands are asymptotically independent if the measurement locations are separated by more than 30 km. Ancona-Navarrete and Tawn [1] confirm this for daily rainfall records in south-west

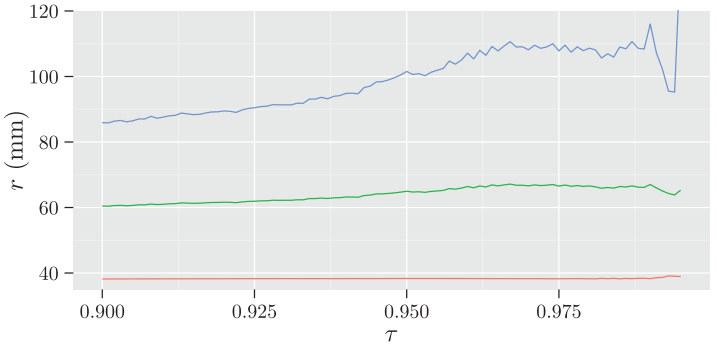

Figure 13. Average return level r of the rainfall data as a function of the non-exceedance probability τ for return periods of 5 (red), 50 (green) and 500 (blue) years [colour online].

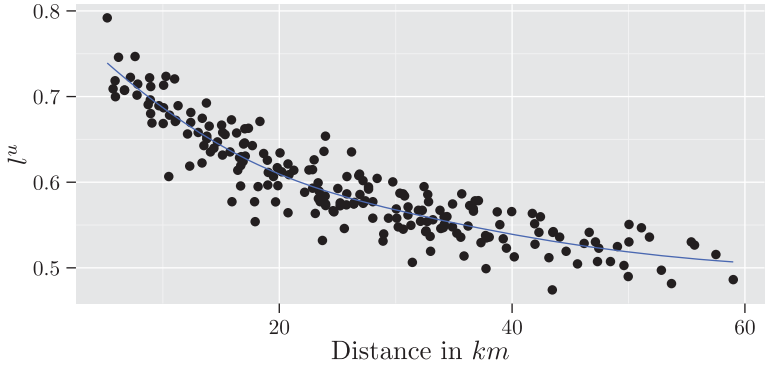

Figure 14. Estimated measure of dependence $l^{\mathrm{u}}(0.9)$ versus distance for the rainfall data. The smooth line is obtained by a loess smoother [colour online].

England, adding that there exists a quite strong association at high levels, which decays only slightly with distance. Thibaud *et al.* [36] find that asymptotically independent models perform better even at very small spatial scales for summer rainfall in Switzerland.

Figure 14 shows the estimated dependence measure $l^{\mathrm{u}}(0.9)$ for each pair of sites versus their distance. The mean of this measure is 0.6. For higher levels this decreases but stays above zero, that is, 0.5 (0.35) at the 0.95(0.99)-quantile. This decay indicates that the normal copula is more appropriate than a asymptotically dependent copula.

To study potential influences of the tail dependence, we model the spatial dependence in three different ways. A multivariate Gumbel copula with parameter $\theta = 1.933$ representing a simple asymptotically dependent model on the one hand and a normal copula with parameter $\rho = 0.8336$ for all pairs representing a simple asymptotically independent model on the other hand. Both copula satisfy $l^{\mathrm{u}}(0.9) = 0.6$. Since, these models do not reproduce the decay of the tail dependence with increasing distance, we also use a multivariate normal copula, with individual correlation parameters ρ_{s_1,s_2} for each pair of sites (s_1, s_2). The correlation parameter ρ_{s_1,s_2} is derived from the loess-fit of $l^{\mathrm{u}}(0.9)$, shown in Figure 14. Figure 15 presents the KS statistic for the rainfall data based on the joint estimation of the GPD parameters, together with the 95 percentage points of the statistic based on three different dependence models: The differences between these dependence models are remarkably small and overall point to a threshold above the 0.96-quantile. In particular the choice to model the decrease of dependence with distance does not seem to influence the selected threshold.

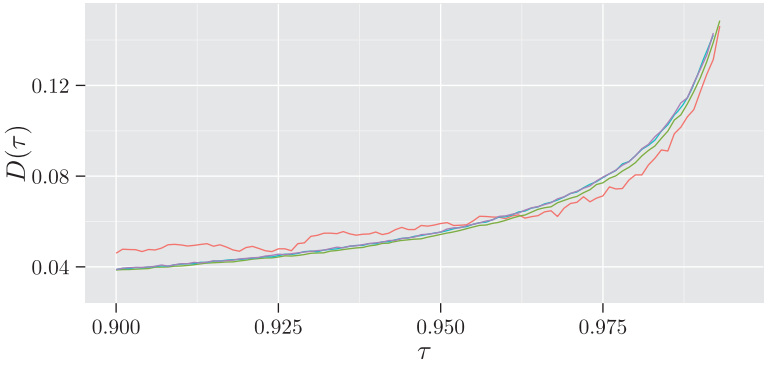

Figure 15. The averaged KS statistic of the rainfall data as a function of τ (red), together with the 95% percentage points based on a Gumbel copula with $\theta = 1.933$ (green), normal copula with $\rho = 0.8336$ (blue), and normal copula with parameters corresponding to the fitted measure of dependence $l^{\mathrm{u}}(0.9)$ (purple) [colour online].

7. Conclusion

This paper has presented a method to incorporate similarity of multiple data samples into the selection of a threshold for a subsequent POT analysis. This setting is often encountered in environmental research but also different stock prices from a similar branch constitute a possible example. We proposed spatially averaged versions of the TS and the ME plots. Moreover, we have introduced spatially averaged KS and AD statistics for assessing the goodness of the GPD fit for the threshold excesses. Our method of threshold selection was validated using a smooth hybrid Weibull GPD distribution. Simulations from this distribution show that the use of common GPD parameters reduces the (negative) bias of the selected threshold substantially. Moreover, the averaged Euclidean distance of the 5- and 50-year return level for the selected threshold is close to the optimal averaged Euclidean distance. The regional method was applied to a set of summer precipitation records in the Netherlands. We found that the 0.96-quantile is a suitable threshold. The choice of the threshold is not sensitive to assumptions regarding the spatial dependence of this data set.

Acknowledgements

All calculations were performed using the R environment (http://www.r-project.org). We are grateful to two reviewers for their helpful comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research was supported by the Dutch research program Knowledge for Climate.

References

[1] M.A. Ancona-Navarrete and J.A. Tawn, *Diagnostics for pairwise extremal dependence in spatial processes*, Extremes 5 (2002), pp. 271–285.

[2] S. Beguería, M. Angulo-Martínez, S.M. Vicente-Serrano, J.I. López-Moreno, and A. El-Kenawy, *Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis: A case study in northeast Spain from 1930 to 2006*, Int. J. Climatol. 31 (2010), pp. 2102–2114.
[3] T.A. Buishand, *Bivariate extreme-value data and the station-year method*, J. Hydrol. 69 (1984), pp. 77–95.
[4] T.A. Buishand, G. De Martino, J.N. Spreeuw, and T. Brandsma, *Homogeneity of precipitation series in the Netherlands and their trends in the past century*, Int. J. Climatol. 33 (2013), pp. 815–833.
[5] V. Choulakian and M.A. Stephens, *Goodness-of-fit tests for the generalized Pareto distribution*, Technometrics 43 (2001), pp. 478–484.
[6] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer, London, 2001.
[7] S. Coles, J. Heffernan, and J.A. Tawn, *Dependence measures for extreme value analyses*, Extremes 2 (1999), pp. 339–365.
[8] A.C. Davison, S.A. Padoan, and M. Ribatet, *Statistical modeling of spatial extremes*, Statist. Sci. 27 (2012), pp. 161–186.
[9] A.C. Davison and R.L. Smith, *Models for exceedances over high thresholds*, J. R. Stat. Soc. B 52 (1990), pp. 393–442.
[10] E.M. Douglas, R.M. Vogel, and C.N. Kroll, *Trends in floods and low flows in the United States: Impact of spatial correlation*, J. Hydrol. 240 (2000), pp. 90–105.
[11] M.-A. El-Aroui and J. Diebolt, *On the use of the peaks over thresholds method for estimating out-of-sample quantiles*, Comput. Statist. Data Anal. 39 (2002), pp. 453–475.
[12] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events*, Springer, Berlin, 1997.
[13] H.J. Fowler, M. Ekström, C.G. Kilsby, and P.D. Jones, *New estimates of future changes in extreme rainfall across the UK using regional climate model integrations. 1. Assessment of control climate*, J. Hydrol. 300 (2005), pp. 212–233.
[14] S. Ghosh and S. Resnick, *A discussion on mean excess plots*, Stochastic Process. Appl. 120 (2010), pp. 1492–1517.
[15] P. Hall and A.H. Welsh, *Adaptive estimates of parameters of regular variation*, Ann. Statist. 13 (1985), pp. 331–341.
[16] M. Hanel, T.A. Buishand, and C.A.T. Ferro, *A nonstationary index flood model for precipitation extremes in transient regional climate model simulations*, J. Geophys. Res. 114 (2009), p. D15107. doi:10.1029/2009JD011712.
[17] J.L. Hintze and R.D. Nelson, *Violin plots: A box plot-density trace synergism*, Amer. Statist. 52 (1998), pp. 181–184.
[18] L. Holden and O. Haug, *A Multidimensial Mixture Model for Unsupervised Tail Estimation*, Samba/09/09, Norwegian Computer Center, 2009.
[19] J.R.M. Hosking and J.R. Wallis, *Regional Frequency Analysis: An Approach Based on L-Moments*, Cambridge University Press, Cambridge, 1997.
[20] D. Koutsoyiannis, *Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records*, Hydrolog. Sci. J. 49 (2004), pp. 591–610.
[21] J. Kyselý and R. Beranová, *Climate-change effects on extreme precipitation in central Europe: Uncertainties of scenarios based on regional climate models*, Theory Appl. Climatol. 95 (2009), pp. 361–374.
[22] J. Kyselý, J. Picek, and R. Beranová, *Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold*, Global Planet. Change 72 (2010), pp. 55–68.
[23] H. Madsen and D. Rosbjerg, *The partial duration series method in regional index-flood modeling*, Water Resour. Res. 33 (1997), pp. 737–746.
[24] R.B. Nelsen, *An Introduction to Copulas*, Springer, New York, 2006.
[25] S.M. Papalexiou and D. Koutsoyiannis, *Battle of extreme value distributions: A global survey on extreme daily rainfall*, Water Resour. Res. 49 (2013), pp. 187–201.
[26] J.I. Pickands, *Statistical inference using extreme order statistics*, Ann. Statist. 3 (1975), pp. 119–131.
[27] R.D. Reiss and M. Thomas, *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*, 3rd ed., Birkhäuser, Basel, 2007.
[28] M. Ribatet, E. Sauquet, J.-M . Grésillon, and T.B.M.J. Ouarda, *A regional Bayesian POT model for flood frequency analysis*, Stoch. Env. Res. Risk A. 21 (2007), pp. 327–339.
[29] M. Roth, T.A. Buishand, G. Jongbloed, A.M.G. Klein Tank, and J.H. van Zanten, *A regional peaks-over-threshold model in a nonstationary climate*, Water Resour. Res. 48 (2012), p. W11533. doi:10.1029/2012WR012214.
[30] M. Roth, T.A. Buishand, G. Jongbloed, A.M.G. Klein Tank, and J.H. van Zanten, *Projections of precipitation extremes based on a regional, non-stationary peaks-over-threshold approach: A case study for the Netherlands and north-western Germany*, Weather Climate Extremes 4 (2014), pp. 1–10.
[31] C. Scarrott and A. MacDonald, *A review of extreme value threshold estimation and uncertainty quantification*, REVSTAT 10 (2012), pp. 33–60.
[32] J. Segers, *Max-stable models for multivariate extremes*, REVSTAT 10 (2012), pp. 61–82.
[33] F. Serinaldi and C.G. Kilsby, *Rainfall extremes: toward reconciliation after the battle of distributions*, Water Resour. Res. 50 (2014), pp. 336–352.
[34] R.L. Smith, *Extreme value theory*, in *Handbook of Applicable Mathematics Supplement*, W. Ledermann, E. Lloyd, S. Vajda, and C. Alexander, eds., Chapter 14, Wiley, Chichester, 1990, pp. 437–472.

[35] C. Svensson and D.A. Jones, *Review of rainfall frequency estimation methods*, J. Flood Risk Manage. 3 (2010), pp. 296–313.
[36] E. Thibaud, R. Mutzner, and A.C. Davison, *Threshold modeling of extreme spatial rainfall*, Water Resour. Res. 49 (2013), pp. 4633–4644.
[37] P. Thompson, Y. Cai, D. Reeve, and J. Stander, *Automated threshold selection methods for extreme wave analysis*, Coast. Eng. 56 (2009), pp. 1013–1021.
[38] J.L. Wadsworth and J.A. Tawn, *Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling*, J. R. Stat. Soc. B 74 (2012), pp. 543–567.
[39] Y. Wang, *Jump and sharp cusp detection by wavelets*, Biometrika 82 (1995), pp. 385–397.
[40] J. Weiss and P. Bernardara, *Comparison of local indices for regional frequency analysis with an application to extreme skew surges*, Water Resour. Res. 49 (2013), pp. 2940–2951.
[41] J. Zhang and M.A. Stephens, *A new and efficient estimation method for the generalized Pareto distribution*, Technometrics 51 (2009), pp. 316–325.