Statistics and Its Interface Volume 8 (2015) 33–43

Spatial aggregation and high quantile estimation applied to extreme precipitation*

Ana Ferreira

When estimating high quantiles and tail probabilities related to the distribution of a spatially aggregated continuous stochastic process, one needs to account for spatial dependence. A way to tackle this problem uses the areal coefficient recently analysed in [8] Ferreira, de Haan and Zhou (2012). We present new ways to estimate this spatial parameter and obtain asymptotic normality of the resulting quantile and tail probability estimators. Note that only consistency for the tail probability estimator was achieved in [8] mainly due to theoretical difficulties with the estimator of the areal coefficient therein considered.

Moreover, we evaluate the effect of the areal coefficient on return values, by an application to three case studies on precipitation extremes: North Holland, Venice Bay in Italy and Northwest Portugal. The proposed estimators seem to be a compromise, in the sense of being easier at a theoretical level and to apply but seem less effective in their performance when compared to the only existing alternative from [8].

In all we intend to draw attention to the areal coefficient. Though it is a unique number characterizing spatial dependence, it helps to explain in a simple way the differences usually observed when estimating quantiles (or tail probabilities) locally and from spatially aggregated data.

AMS 2000 subject classifications: Primary 60G70, 62G32, 62M30; secondary 62P12 .
Keywords and phrases: Extreme quantile and tail probability estimation, Generalized Pareto distribution, Spatial dependence, Spatial aggregation, Areal coefficient, Extreme precipitation.

1. INTRODUCTION

Let daily rainfall over the space be represented by a continuous stochastic process X on some compact space S, i.e. $X = \{X(s)\}_{s\in S}$. We want to study estimation of extreme events on the basis of independent and identically distributed (i.i.d.) observations of this process, and of special concern are quantities related to the tail probability of the process spatially aggregated. The aggregation of precipitation over space corresponds then to $\int_S X(s)ds$, the total amount of daily rain over S. When estimating (e.g. 100-year) return values one may consider local estimates on the basis of univariate extreme value theory, but also important is the estimation of return values for the total amount of rain over the whole region S.

To tackle the aggregated problem, a parameter accounting for the effect of spatial aggregation, called the areal coefficient, was introduced in [2]. Subsequently, [8] justified it in practical terms, relating it with the appropriate domain of attraction condition, and analysed its estimation. Though with a very practical interpretation, the introduced estimator for the areal coefficient relies heavily on the underlying theoretical framework. Perhaps this explains why it has not been exploited so much. We present an alternative way to estimate this parameter and obtain asymptotic normality of resulting estimators.

We consider extreme value theory. Let X be a stochastic process defined in the space of continuous functions, $X \in C(S)$, with S some compact subset of $\mathbb{R}^2$. Throughout we assume the domain of attraction condition of max-stable processes [10]:

Let $X_1, X_2, \ldots$ be i.i.d. random elements of $C(S)$. Suppose there exist normalizing functions $a_n = \{a_n(s) > 0\}_{s\in S}$ and $b_n = \{b_n(s)\}_{s\in S}$ in $C(S)$ such that,

$$\left\{\max_{1\le i\le n} \frac{X_i(s) - b_n(s)}{a_n(s)}\right\}_{s\in S} \to \eta, \qquad n\to\infty, \tag{1}$$

weakly (or in distribution) and in $C(S)$ with $\eta = \{\eta(s)\}_S \in C(\mathbb{R})$ a stochastic process with non-degenerate marginals. Then it is well known that the limiting process is a max-stable process ([9]; cf. also [11]): for $\eta_1, \eta_2, \ldots$, i.i.d. copies of η, there are real continuous functions $c_n = \{c_n(s) > 0\}_{s\in S}$ and $d_n = \{d_n(s)\}_{s\in S}$ such that,

$$\max_{1\le i\le n} \frac{\eta_i - d_n}{c_n} \overset{d}{=} \eta \qquad \text{for all } \ n = 1, 2, \ldots.$$

The process is called simple if its marginal distributions are standard Fréchet.

Under the domain of attraction condition (and some further natural assumptions) the tail distribution of $\int_S X(s)ds$ is asymptotically generalized Pareto (GP) [8]. One of the

*Research partially supported by FCT Project PTDC /MAT /112770 /2009; EXPL/MAT-STA/0622/2013 and PEst-OE/MAT/UI0006/2014.

important applications of extreme value theory is the estimation of high quantiles and the GP tail is a key approximation. Under our framework one has to additionally deal with the estimation of the areal coefficient. We aim at exploiting new ways to estimate it, and justify the procedure both theoretically and in practice.

The outline of the paper is as follows. In Section 2 we present the theoretical results, namely asymptotic normality of the resulting estimators for the areal coefficient, high quantiles ('high' means, as usual in extreme value theory, that we are specially interested in values beyond the sample information) and tail probabilities. In Section 3, the results are applied to spatial daily precipitation data over three different regions: Northwest Portugal, North Holland and Venice Bay in Italy. We focus mainly on return value estimation as being commonly used when evaluating precipitation extremes. Our results seem to be a compromise with the results from [8], in the sense that they are easier to handle both theoretically and in practice but, on the other hand, show larger variance or bias.

2. ESTIMATORS AND ASYMPTOTIC NORMALITY

2.1 Definition of estimators

Denote by $\bar{\eta} = \{\bar{\eta}(s)\}_{s\in S}$ any simple max-stable process in $C^+(S) = \{f \in C(S) : f \geq 0\}$. Any max-stable process $\eta = \{\eta(s)\}_{s\in S}$ in $C(S)$ can be represented by

$$\eta = \frac{\bar{\eta}^\gamma - 1}{\gamma}, \tag{2}$$

for some $\bar{\eta}$ and a continuous function $\gamma = \{\gamma(s)\}_{s\in S}$ called the extreme value index function.

For any simple max-stable process, there exists a finite measure ρ on $\overline{C}_1^+(S) = \{f \in C(S) \ : \ f \geq 0 \, , \sup_{s\in S} f = 1\}$, called the spectral measure such that

$$\int_{\overline{C}_1^+(S)} f(s) \, d\rho(f) = 1 \tag{3}$$

for all $s \in S$ and for $m = 1, 2, \ldots$, $K_1, K_2, \ldots, K_m$ compact sets in S and $x_1, x_2, \ldots, x_m > 0$

$$-\log P(\eta(s) \leq x_j, \text{ for } s \in K_j, j = 1, 2, \ldots, m) = \int_{\overline{C}_1^+(S)} \max_{1\leq j\leq m} \left(x_j^{-1} \sup_{s\in K_j} g(s) \right) \, d\rho(g);$$

cf. [9].

For completeness we state next a result from [8], establishing the limiting tail probability of $\int_S X(s)ds$, which is our main motivation.

Theorem 2.1. *Suppose (1)–(2) hold with constant $\gamma \equiv \gamma(s)$, $s \in S$, and that for some positive functions a_t and $A(s)$,*

$$\sup_{s\in S} \left| \frac{a_t(s)}{a_t} - A(s) \right| \to 0, \quad as \quad t \to \infty, \tag{4}$$

(then one can take $a_t = \int_S a_t(s)ds$ which implies $\int_S A(s)ds = 1$); for $\gamma = 0$ require $a_t(s) = a_t A(s)$ for all t and $s \in S$. Additionally, if $\gamma \leq 0$ require $\rho\{g \in \bar{C}_1^+(S) : \inf_{s\in S} g(s) = 0\} = 0$; if $\gamma > 0$ require that X is non-negative.

Then,

$$\lim_{t\to\infty} tP\left(\frac{\int_S X(s)ds - \int_S b_t(s)ds}{a_t} > x \right) = \theta_\gamma (1+\gamma x)^{-1/\gamma} \tag{5}$$

for all x with $1 + \gamma x > 0$ where

$$\theta_\gamma = \int_{\bar{C}_1^+(S)} \left(\int_S A(s) \, g^\gamma(s) \, ds \right)^{1/\gamma} d\rho(g). \tag{6}$$

For $\gamma = 0$ the right-hand side of (5) should be read as $\theta_0 e^{-x}$ and the right-hand side of (6) as $\int_{\bar{C}_1^+(S)} \exp(\int_S A(s) \, \log g(s) \, ds) d\rho(g)$.

The extra parameter θ_γ appearing in the limit (5) is the areal coefficient, completely specified in (6). In general it follows that (cf. [8]):

1. $0 < \theta_\gamma \leq 1, \quad$ for $\gamma \leq 1$,
2. $\theta_\gamma \geq 1, \quad$ for $\gamma \geq 1$.

Particular situations are: $\theta_\gamma = 1$ in case of total dependence and $\theta_1 = 1$.

A direct consequence of the above theorem is that the distribution of $\int_S X(s)ds$ is in the domain of attraction of some max-stable distribution. This follows from standard univariate extreme value theory and the fact that

$$\theta_\gamma (1+\gamma x)^{-1/\gamma} = \left(1 + \gamma \frac{x - (\theta_\gamma^\gamma - 1)/\gamma}{\theta_\gamma^\gamma} \right)^{-1/\gamma}.$$

That is, in the right-hand side of (5) we have a GP distribution. Hence,

Corollary 2.1. *Under the conditions of Theorem 2.1,*

$$\lim_{t\to\infty} tP\left(\frac{\int_S X(s)ds - \tilde{b}_t}{\tilde{a}_t} > x \right) = (1+\gamma x)^{-1/\gamma}, \tag{7}$$

for all x with $1 + \gamma x > 0$, $\gamma \in \mathbb{R}$, and the normalizing constants $\tilde{a}_t$ and $\tilde{b}_t$ can be taken as

$$\tilde{a}_t = \theta_\gamma^\gamma a_t \quad \text{and} \quad \tilde{b}_t = \int_S b_t(s)ds + a_t \frac{\theta_\gamma^\gamma - 1}{\gamma}, \tag{8}$$

with $a_t = \int_S a_t(s)ds$ and $b_t(s)$ from (5).

Solving both equations in (8) in θ_γ, motivates the estimators for the areal coefficient:

$$\hat\theta_1 \equiv \hat\theta_\gamma^{(1)} = \left(\frac{\widehat{\tilde a}_{n/k}}{\hat a_{n/k}}\right)^{1/\hat\gamma_{n/k}} \quad \text{and}$$

$$\hat\theta_2 \equiv \hat\theta_\gamma^{(2)} = \left(1+\hat\gamma_{n/k}\frac{\widehat{\tilde b}_{n/k}-\int_S \hat b_{n/k}(s)ds}{\hat a_{n/k}}\right)^{1/\hat\gamma_{n/k}},$$

for suitable estimators $\hat\gamma_{n/k}$, $\widehat{\tilde a}_{n/k}$, $\widehat{\tilde b}_{n/k}$, $\hat a_{n/k}$ and $\hat b_{n/k} = \int_S \hat b_{n/k}(s)ds$ of γ, $\tilde a_{n/k}$, $\tilde b_{n/k}$, $a_{n/k}$ and $b_{n/k} = \int_S b_{n/k}(s)ds$ respectively.

We aim at establishing asymptotic normality of $\hat\theta_1$ and $\hat\theta_2$ and, for that we need asymptotic normality of all the intermediate estimators. Assume throughout the conditions of Theorem 2.1.

2.2 Asymptotic normality of $\hat\gamma_{n/k}$, $\widehat{\tilde a}_{n/k}$, $\widehat{\tilde b}_{n/k}$, $\hat a_{n/k}$ and $\hat b_{n/k}$

As it is usual in extreme value theory, we need second order conditions. We shall impose conditions on $U_\cdot$, the inverse function of $1/P(\int_S X(s)ds > x)$, and on $U_\cdot(s)$, the marginal inverse functions of $1/P(X(s) > x)$, i.e. for each $s \in S$.

We start with the former. Assume there exists a function α_t, positive or negative with $|\alpha_t|$ regularly varying with index $\rho \le 0$ and $\lim_{t\to\infty}\alpha_t = 0$ such that

$$\lim_{t\to\infty} \frac{\frac{U_{tx}-\int_S b_t(s)\,ds}{a_t} - \frac{(\theta_\gamma x)^\gamma - 1}{\gamma}}{\alpha_t} = \theta_\gamma^\gamma H_{\gamma,\rho}(x), \quad x>0 \tag{10}$$

with

$$H_{\gamma,\rho}(x) = \begin{cases} \frac{1}{\rho}\left(\frac{x^{\gamma+\rho}-1}{\gamma+\rho} - \frac{x^\gamma-1}{\gamma}\right), & \rho\neq 0\neq\gamma \\ \frac{1}{\gamma}\left(x^\gamma\log x - \frac{x^\gamma-1}{\gamma}\right), & \rho=0\neq\gamma \\ \frac{1}{\rho}\left(\frac{x^\rho-1}{\rho} - \log x\right), & \rho\neq 0=\gamma \\ \frac{1}{2}(\log x)^2, & \rho=0=\gamma. \end{cases}$$

The above second order condition is the standard one related to the limit (7) ([13]; cf. also [11] Sect. 2.3) but rewritten in terms of the normalizing functions a_t and $b_t(s)$ from (5) and using the relations (8). Note that it provides a relation between the functions U_t and $\int_S U_t(s)ds$.

Then it is well known (cf. [11] Sect. 4.2) that there are estimators $\widehat{\tilde a}_{n/k}$ and $\widehat{\tilde b}_{n/k}$ on the basis of an i.i.d. sample of size n, for which

$$\sqrt{k}\left(\frac{\widehat{\tilde a}_{n/k}}{\tilde a_{n/k}} - 1, \frac{\widehat{\tilde b}_{n/k} - b_{n/k}}{\tilde a_{n/k}}\right) \to^d \left(\mathcal{N}_{\tilde a}, \mathcal{N}_{\tilde b}\right), \quad n\to\infty, \tag{11}$$

with $k = k_n$ an intermediate sequence (i.e. $k\to\infty$ and $k/n \to 0$) such that $\sqrt{k}\alpha_{n/k} \to \lambda \ge 0$, as $n\to\infty$, and $(\mathcal{N}_{\tilde a}, \mathcal{N}_{\tilde b})$ are jointly normal random variables. In the given framework, i.e. from i.i.d. replicates of $\int_S X(s)ds$, estimators for γ are also standard to obtain. Indeed we shall need to estimate γ but a better estimator for γ is introduced next.

On the other hand, similarly as in [8], one can define estimators for γ, $a_t(s)$ and $b_t(s)$ such that for convenient rates

$$\sqrt{k}\left(\hat\gamma_{n/k}-\gamma, \frac{\int_S \hat a_{n/k}(s)}{a_{n/k}} - 1, \frac{\int_S \hat b_{n/k}(s) - b_{n/k}(s)ds}{a_{n/k}}\right) \to^d (\mathcal{N}_\gamma, \mathcal{N}_a, \mathcal{N}_b), \quad n\to\infty, \tag{12}$$

with $(\mathcal{N}_\gamma, \mathcal{N}_a, \mathcal{N}_b)$ jointly normal random variables. This can be obtained under the following marginal second order conditions: there exist functions α_t positive or negative with $|\alpha_t|$ regularly varying of index $\rho \le 0$ and $\lim_{t\to\infty}\alpha_t = 0$ and, $\beta(s)$ positive or negative and bounded, such that

$$\lim_{t\to\infty} \frac{\frac{U_{tx}(s)-U_t(s)}{a_t(s)} - \frac{x^\gamma-1}{\gamma}}{\alpha_t\beta(s)} = H_{\gamma,\rho}(x), \quad x>0 \tag{13}$$

holds uniformly for $s \in S$.

Then, (11) and (12) hold for an intermediate sequence verifying $\sqrt{k}\alpha_{n/k} \to \lambda \ge 0$ and

$$\sqrt{k}\sup_{s\in S}\left|\frac{a_{n/k}(s)}{a_{n/k}} - A(s)\right| \to \Delta \ge 0. \tag{14}$$

The joint limiting distributions in (11) and (12) can be very general. Note that they depend on the underlying structure of the limiting max-stable process. As far as we know very little on this is available, and characterizations of the random variables $\mathcal{N}_\gamma$, $\mathcal{N}_a$, $\mathcal{N}_b$ are fully known only for the moment estimators [7]. More specific characterizations are beyond the scope of this work.

2.3 Asymptotic normality of $\hat\theta_1$ and $\hat\theta_2$

Theorem 2.2. *Let k be an intermediate sequence, i.e. $k = k(n) \to \infty$, $k(n)/n \to 0$, as $n\to\infty$, such that:*

1. *For $\gamma \neq 0$ and for suitable estimators $\hat\gamma_{n/k}$, $\widehat{\tilde a}_{n/k}$ and $\hat a_{n/k}$,*

$$\sqrt{k}\left(\hat\gamma_{n/k}-\gamma, \frac{\widehat{\tilde a}_{n/k}}{\tilde a_{n/k}} - 1, \frac{\hat a_{n/k}}{a_{n/k}} - 1\right) \to^d (\mathcal{N}_\gamma, \mathcal{N}_{\tilde a}, \mathcal{N}_a), \tag{15}$$

as $n\to\infty$, with $(\mathcal{N}_\gamma, \mathcal{N}_{\tilde a}, \mathcal{N}_a)$ jointly normal random variables. Then,

$$\sqrt{k}\left(\frac{\hat\theta_1}{\theta_\gamma} - 1\right) \to^d \frac{1}{\gamma}\left(\mathcal{N}_{\tilde a} - \mathcal{N}_a - (\log\theta_\gamma)\mathcal{N}_\gamma\right), n\to\infty. \tag{16}$$

2. *For suitable estimators* $\hat{\gamma}_{n/k}$, $\widehat{\tilde{a}}_{n/k}$, $\widehat{\tilde{b}}_{n/k}$, $\hat{a}_{n/k}$ *and* $\hat{b}_{n/k} = \int_S \hat{b}_t(s)ds$,

(17)
$$\sqrt{k}\left(\hat{\gamma}_{n/k}-\gamma, \frac{\hat{a}_{n/k}}{a_{n/k}}-1, \frac{\widehat{\tilde{b}}_{n/k}-\tilde{b}_{n/k}}{\tilde{a}_{n/k}}, \frac{\hat{b}_{n/k}-b_{n/k}}{a_{n/k}}\right) \to^d \left(\mathcal{N}_\gamma, \mathcal{N}_a, \mathcal{N}_{\tilde{b}}, \mathcal{N}_b\right), \quad n\to\infty,$$

with $(\mathcal{N}_\gamma, \mathcal{N}_a, \mathcal{N}_{\tilde{b}}, \mathcal{N}_b)$ *jointly normal random variables. Then, for* $\gamma \neq 0$,

(18)
$$\sqrt{k}\left(\frac{\hat{\theta}_2}{\theta_\gamma}-1\right) \to^d \left(\frac{1-\theta_\gamma^{-\gamma}}{\gamma}-\log\theta_\gamma\right)\frac{1}{\gamma}\mathcal{N}_\gamma - \frac{1-\theta_\gamma^{-\gamma}}{\gamma}\frac{1}{\gamma}\mathcal{N}_a + \left(\mathcal{N}_{\tilde{b}} - \theta_\gamma^{-\gamma}\mathcal{N}_b\right),$$

as $n\to\infty$, *and, for* $\gamma = 0$,

(19)
$$\sqrt{k}\left(\frac{\hat{\theta}_2}{\theta_0}-1\right) \to^d \mathcal{N}_{\tilde{b}} - \mathcal{N}_b - (\log\theta_0)\,\mathcal{N}_a,$$

as $n\to\infty$.

Proof. 1. First note that by (9) and (15),

$$\frac{\widehat{\tilde{a}}_{n/k}}{\hat{a}_{n/k}} = \theta_\gamma^\gamma \frac{\widehat{\tilde{a}}_{n/k}}{\theta_\gamma^\gamma a_{n/k}}\frac{a_{n/k}}{\hat{a}_{n/k}} = \theta_\gamma^\gamma\left(1+\frac{\mathcal{N}_{\tilde{a}}-\mathcal{N}_a}{\sqrt{k}}(1+o_p(1))\right).$$

Hence,

$$\begin{aligned}\sqrt{k}\left(\frac{\hat{\theta}_1}{\theta_\gamma}-1\right) &= \sqrt{k}\left(\frac{(\frac{\widehat{\tilde{a}}_{n/k}}{\hat{a}_{n/k}})^{1/\hat{\gamma}_{n/k}}}{\theta_\gamma}-1\right)\\ &= \sqrt{k}\left(\theta^{\frac{\gamma}{\hat{\gamma}_{n/k}}-1}\right.\\ &\quad\left.\times\left\{1+\frac{\mathcal{N}_{\tilde{a}}-\mathcal{N}_a}{\sqrt{k}}(1+o_p(1))\right\}^{1/\hat{\gamma}_{n/k}}-1\right)\\ &= \sqrt{k}\left(\left\{1-\frac{\mathcal{N}_\gamma}{\gamma\sqrt{k}}\log\theta_\gamma(1+o_p(1))\right\}\right.\\ &\quad\left.\times\left\{1+\frac{1}{\gamma}\frac{\mathcal{N}_{\tilde{a}}-\mathcal{N}_a}{\sqrt{k}}(1+o_p(1))\right\}-1\right)\\ &= \frac{1}{\gamma}\left(\mathcal{N}_{\tilde{a}}-\mathcal{N}_a-\mathcal{N}_\gamma\log\theta_\gamma\right)(1+o_p(1)).\end{aligned}$$

2. First note that (10) with $x=1$ implies

$$\frac{U_t - \int_S b_t(s)\,ds}{a_t} = \frac{\theta_\gamma^\gamma - 1}{\gamma} + o(\alpha_t).$$

Then by (17),

$$\begin{aligned}\frac{\widehat{\tilde{b}}_{n/k}-\hat{b}_{n/k}}{\hat{a}_{n/k}} &= \frac{a_{n/k}}{\hat{a}_{n/k}}\\ &\quad\times\left(\frac{\widehat{\tilde{b}}_{n/k}-\tilde{b}_{n/k}}{\tilde{a}_{n/k}}\frac{\tilde{a}_{n/k}}{a_{n/k}} - \frac{\hat{b}_{n/k}-b_{n/k}}{a_{n/k}} + \frac{\tilde{b}_{n/k}-b_{n/k}}{a_{n/k}}\right)\\ &= \left\{\frac{\theta_\gamma^\gamma\mathcal{N}_{\tilde{b}}-\mathcal{N}_b}{\sqrt{k}}(1+o_p(1)) + \frac{\theta_\gamma^\gamma-1}{\gamma}(1+o(\alpha_{n/k}))\right\}\\ &\quad\times\left\{1-\frac{\mathcal{N}_a}{\sqrt{k}}(1+o_p(1))\right\}\\ &= \frac{\theta_\gamma^\gamma-1}{\gamma}(1+o(\alpha_{n/k}))\\ &\quad+\left(\theta_\gamma^\gamma\frac{\mathcal{N}_{\tilde{b}}}{\sqrt{k}} - \frac{\mathcal{N}_b}{\sqrt{k}} - \frac{\theta_\gamma^\gamma-1}{\gamma}\frac{\mathcal{N}_a}{\sqrt{k}}\right)(1+o_p(1)).\end{aligned}$$

Hence,

$$\begin{aligned}1+\hat{\gamma}_{n/k}\frac{\widehat{\tilde{b}}_{n/k}-\int_S\hat{b}_{n/k}(s)ds}{\hat{a}_{n/k}} &= \theta_\gamma^\gamma + \left(\theta_\gamma^\gamma-1\right)o(\alpha_{n/k})\\ &\quad+\frac{1}{\sqrt{k}}\left(\left(\mathcal{N}_\gamma-\mathcal{N}_a\right)\frac{\theta_\gamma^\gamma-1}{\gamma} + \gamma\theta_\gamma^\gamma\mathcal{N}_{\tilde{b}} - \gamma\mathcal{N}_b\right)(1+o_p(1)).\end{aligned}$$

Therefore, by similar calculations as in the previous case, for $\gamma\neq 0$,

$$\sqrt{k}\left(\frac{\hat{\theta}_2}{\theta_\gamma}-1\right) = \left(\frac{1-\theta_\gamma^{-\gamma}}{\gamma}-\log\theta_\gamma\right)\frac{1}{\gamma}\mathcal{N}_\gamma - \frac{1-\theta_\gamma^{-\gamma}}{\gamma}\frac{1}{\gamma}\mathcal{N}_a + \left(\mathcal{N}_{\tilde{b}}-\theta_\gamma^{-\gamma}\mathcal{N}_b\right) + o_p(1).$$

The case $\gamma = 0$ follows similarly. □

2.4 Quantile and tail probability estimation

We proceed with the asymptotic normality of estimators for high quantiles $x_n = U(1/p_n)$ and tail probabilities $p_n = P(\int_S X(s)ds > x_n)$, for given p_n and x_n respectively where, as usual in high quantile estimation, $p_n\to 0$ or $x_n\to U(\infty)$, as the sample size $n\to\infty$. Note that only consistency for a tail probability estimator was obtained in [8], due to theoretical difficulties with the estimator for θ therein considered. Moreover it needed a consistent estimator of the limiting spectral measure whereas we do not need it here.

Define the quantile estimator as

(20)
$$\hat{x}_{p_n} = \hat{b}_{n/k} + \hat{a}_{n/k}\frac{(\frac{\hat{\theta}k}{np_n})^{\hat{\gamma}_{n/k}}-1}{\hat{\gamma}_{n/k}},$$

and the tail probability estimator by

$$\hat{p}_n = \frac{k}{n}\hat{\theta}\left(1+\hat{\gamma}_{n/k}\frac{x_n-\hat{b}_{n/k}}{\hat{a}_{n/k}}\right)_+^{-1/\hat{\gamma}_{n/k}}, \tag{21}$$

with $\hat{\theta}=\hat{\theta}_i$, $i=1$ or 2 from (9) and $x_+=\max(0,x)$.

Theorem 2.3. *Assume the conditions of Theorem 2.1. Suppose for some function α_t, positive or negative with $|\alpha_t|$ regularly varying with index $\rho\le 0$ and $\lim_{t\to\infty}\alpha_t=0$, the second order condition (10) with $\rho<0$, or $\rho=0$ if $\gamma<0$, holds. Let k be an intermediate sequence such that $\sqrt{k}\alpha_{n/k}\to\lambda\in\mathbb{R}$, $np_n=o(k)$, $\log(np_n)=o(\sqrt{k})$ $(n\to\infty)$ and the conditions of Theorem 2.2 are satisfied. Then:*

1. (22)
$$\sqrt{k}\frac{\hat{x}_n-x_n}{a_{n/k}q_\gamma(d_n)} \to^d \mathcal{N}_\gamma+(\gamma_-)^2\mathcal{N}_b-\gamma_-\mathcal{N}_a-\lambda\frac{\theta_\gamma^\gamma\gamma_-}{\gamma_-+\rho};$$

2. *For $\gamma>-1/2$,*
$$\frac{\sqrt{k}}{d_n^{-\gamma}q_\gamma(d_n)}\left(\frac{\hat{p}_n}{p_n}-1\right) \to^d \mathcal{N}_\gamma+(\gamma_-)^2\mathcal{N}_b-\gamma_-\mathcal{N}_a-\lambda\frac{\theta_\gamma^\gamma\gamma_-}{\gamma_-+\rho}; \tag{23}$$

as $n\to\infty$, with $d_n=\theta_\gamma k/(np_n)$, $\gamma_-=\min(0,\gamma)$ and $q_\gamma(t)=\int_1^t s^{\gamma-1}(\log s)ds$ for $t>1$.

The following Lemma is needed for the proof of Theorem 2.3.

Lemma 2.1. *If (10) holds with $\rho<0$ or $\rho=0$ and $\gamma<0$ then*

$$\lim_{\substack{t\to\infty\\ x=x(t)\to\infty}}\frac{\frac{U(tx)-U(t)}{a(t)}\frac{\gamma}{(\theta_\gamma x)^\gamma-1}-1}{\alpha(t)}=-\frac{\theta_\gamma^{\gamma_-}}{\rho+\gamma_-}.$$

Proof. It follows by similar arguments as in the proof of Lemma 4.3.5 [11] and, note that

$$\theta_\gamma^\gamma H_{\gamma,\rho}(x)\frac{\gamma}{(\theta x)^\gamma-1}\sim-\frac{\theta_\gamma^{\gamma_-}}{\rho+\gamma_-},\quad \text{as } x\to\infty. \qquad \square$$

Proof of Theorem 2.3. 1. Note that

$$q_\gamma(d_n)\sim\begin{cases}\frac{1}{\gamma}d_n^\gamma\log d_n\,, & \gamma>0\\ \frac{1}{2}(\log d_n)^2\,, & \gamma=0\\ 1/\gamma^2\,, & \gamma<0.\end{cases} \tag{24}$$

Then, similarly as in the one-dimensional case ([12], cf. [11] Sect. 4.3),

$$\begin{aligned}\sqrt{k}\frac{\hat{x}_n-x_n}{a_{n/k}q_\gamma(d_n)} &= \sqrt{k}\frac{\hat{b}_{n/k}-b_{n/k}}{a_{n/k}}\frac{1}{q_\gamma(d_n)}\\ &\quad+\frac{\hat{a}_{n/k}}{a_{n/k}}\left\{\frac{\sqrt{k}}{q_\gamma(d_n)}\left(\frac{(\frac{\hat{\theta}k}{np_n})^{\hat{\gamma}_{n/k}}-1}{\hat{\gamma}_{n/k}}-\frac{\tilde{d}_n^\gamma-1}{\gamma}\right)\right\}\\ &\quad+\sqrt{k}\left(\frac{\hat{a}_{n/k}}{a_{n/k}}-1\right)\frac{d_n^\gamma-1}{\gamma q_\gamma(d_n)}\\ &\quad-\frac{\sqrt{k}}{q_\gamma(d_n)}\left\{\frac{U\left(\frac{n}{k}d_n\right)-U\left(\frac{n}{k}\right)}{a_{n/k}}-\frac{d_n^\gamma-1}{\gamma}\right\}.\end{aligned}$$

The first and third terms are easily seen to converge to $(\gamma_-)^2\mathcal{N}_b$ and $(\gamma_-)\mathcal{N}_a$, respectively. For the second term, note that

$$\begin{aligned}&\frac{\sqrt{k}}{q_\gamma(d_n)}\frac{(\frac{\hat{\theta}k}{np_n})^{\hat{\gamma}_{n/k}}-d_n^{\hat{\gamma}_{n/k}}}{\hat{\gamma}_{n/k}}\\ &\quad=\frac{d_n^\gamma}{q_\gamma(d_n)\hat{\gamma}_{n/k}}d_n^{\hat{\gamma}_{n/k}-\gamma}\sqrt{k}\left\{\left(\frac{\hat{\theta}}{\theta}\right)^{\hat{\gamma}_{n/k}}-1\right\}\\ &\quad=o_p(1)\end{aligned}$$

and

$$\frac{\sqrt{k}}{q_\gamma(d_n)}=\left(\frac{d_n^{\hat{\gamma}_{n/k}}-1}{\hat{\gamma}_{n/k}}-\frac{d_n^\gamma-1}{\gamma}\right)\sim\sqrt{k}\left(\hat{\gamma}_{n/k}-\gamma\right),$$

hence it converges to $\mathcal{N}_\gamma$. Finally, for the last term, by Lemma 2.1 it converges to $-\lambda\gamma_-\frac{\theta_\gamma^{\gamma_-}}{\rho+\gamma_-}$. Combining all terms the result follows.

2. It follows by the reasoning on tail probability estimation in the one dimensional case ([4], cf. [11] Sect. 4.4) adapted to the inclusion of the estimator $\hat{\theta}$, in a way similar to the previous proof for quantile estimation. $\square$

3. CASE STUDIES

The estimators discussed in the previous sections are applied to three data sets of daily surface precipitation (mm). The data was collected at rain-gauge stations corresponding to three different regions in Europe: Northwest Portugal, Venice Bay in Italy and North Holland. In the Appendix we give a spatial representation of the stations (cf. Figures 7–9), where the stations are identified by their ID-numbers as provided in the original data sets from the Institutes.

Precipitation data usually exhibits some temporal dependence and our methods are developed under the i.i.d. assumption for the stochastic processes in the maximum domain of attraction. There are some univariate approaches

Table 1. Characteristics of observed regions and corresponding data sets

Region	Time Period	N. of Years	Sample size	N. of Stations	Total area (Km^2)
NW Portugal	1950–2008	59	5369	31	3,754.0
Venice Bay	1940–1994	55	5005	24	3,074.3
North Holland	1971–2000	30	2730	32	2,009.6

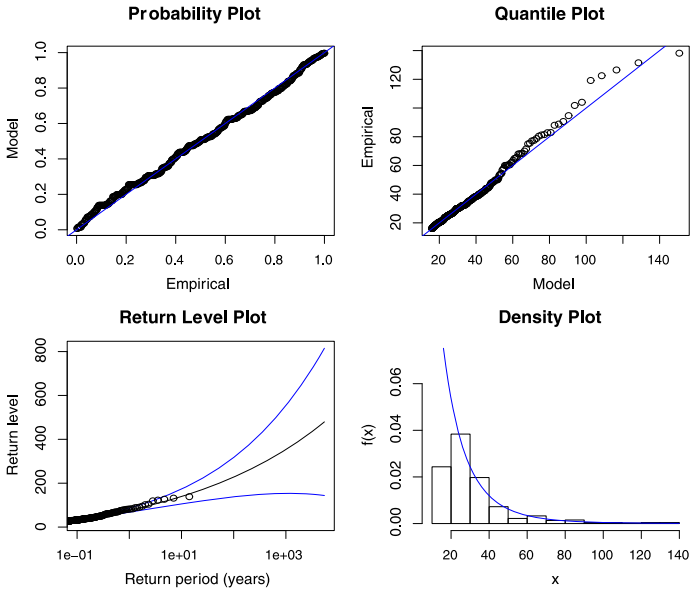

Figure 1. Diagnostic plots (from library ismev, R software) for the marginal GP-fit of observed daily precipitation at Station 56, NW Portugal.

that try to clean univariate serial data for short-range dependence, for example the declustering method but results are frequently not too sensitive in this respect [15]. Additionally it is also known that many of the univariate methods hold under β-mixing dependence conditions though with more complicated asymptotic properties ([5], [6]). On the other hand, our framework is not univariate but spatial, hence more complicated to deal with. Therefore, to our purposes we found it reasonable to consider daily data (as opposed to e.g. hourly data expected to show stronger time dependence) and only the fall season i.e. the observation period September-October-November, the latter to overcome long-range dependence effects giving a total of 91 observations per year and, consequently, treat the data as being (approximately) independent; note that this same approach was taken in [8] and [17].

A brief summary of characteristics of the three data sets is given in Table 1.

Marginal analysis

One has first to proceed with marginal estimation. In general, a reasonable Pareto fit is obtained for the tails (i.e. for the highest values at each station) for all stations and for the three regions. In Figure 1 is a typical fit, with the diagnostic plots from library ismev, from R software.

Throughout we use the moment estimators ([3] and [7]). In Figure 2 are marginal estimates of shape and scale parameters ranked with the number of the stations. In particular, we do not identify any remarkable non-homogeneous pattern from the shape estimates. This is important for the results to be consistent with the theory.

Regarding the results, it is interesting to see similar magnitudes for the shape parameter throughout all the regions but very different magnitudes of the scale estimates. In what regards the shape, NW Portugal has a tendency for slightly lower values and Venice for slightly higher values, with NW Portugal and Venice presenting higher variability (cf. also Figure 6 in Appendix A). Note that the NW Portugal region varies much geographically, especially in altitude. Overall, the majority of the shape estimates correspond to low but positive values, around 0.05–0.2, what is typically observed for precipitation data (e.g. cf. [14], [15], [17]). In what regards the scale (and location) estimates, NW Portugal have the highest values and North Holland the lowest (cf. also Figure 6 in Appendix A).

Return value estimation and the spatial aggregation effect

We proceed with the estimation of the 100-year return value (averaged by the total area) for each of the three regions. Loosely, the N-year return value is that value that is expected to occur once in N-years. Hence, for N large it is basically an extreme quantile of the precipitation distribution and, as discussed earlier, the asymptotic theory suggests the estimator,

$$\hat{rv}_{100} = \int_S \hat{b}_{n/k}(s)ds + \int_S \hat{a}_{n/k}(s)ds\, \frac{(\frac{9100\,k\,\hat{\theta}}{n})^{\hat{\gamma}_{n/k}} - 1}{\hat{\gamma}_{n/k}} \tag{25}$$

with

$$\hat{\gamma}_{n/k} = \frac{1}{|S|}\int_S \hat{\gamma}_{n/k}(s)\,ds, \tag{26}$$

with n the sample size and k such that $k \to \infty$ and $k/n \to 0$, as $n \to \infty$. Recall that we consider 91 observations per year.

In Figure 3 are shown 100-year return value estimates with different estimators for θ ($\hat{\theta}_1$ and $\hat{\theta}_2$ from (9) and the estimator from [8]), and θ estimates, with k. For the performance of the estimators, the ones from [8], in the following FHZ, seem to be the best. In what regards θ estimation, the estimates from $\hat{\theta}_1$ are somehow close to the ones from FHZ,

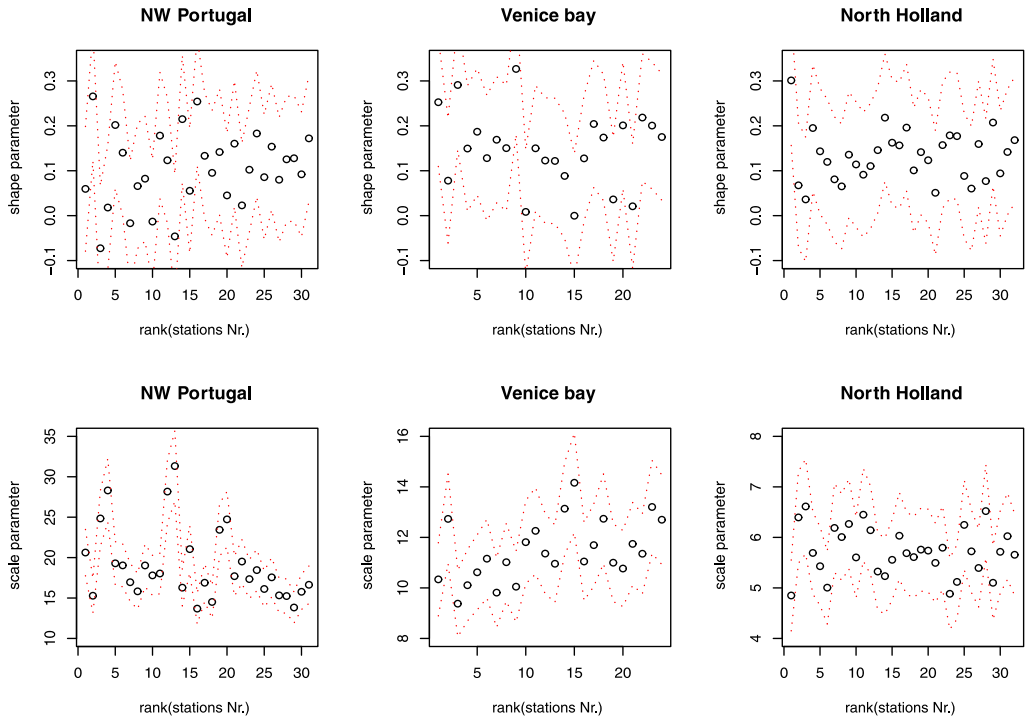

Figure 2. Marginal estimates of shape and scale parameters with $k = 200$ ranked with the number of the stations, with approximate 95% confidence intervals.

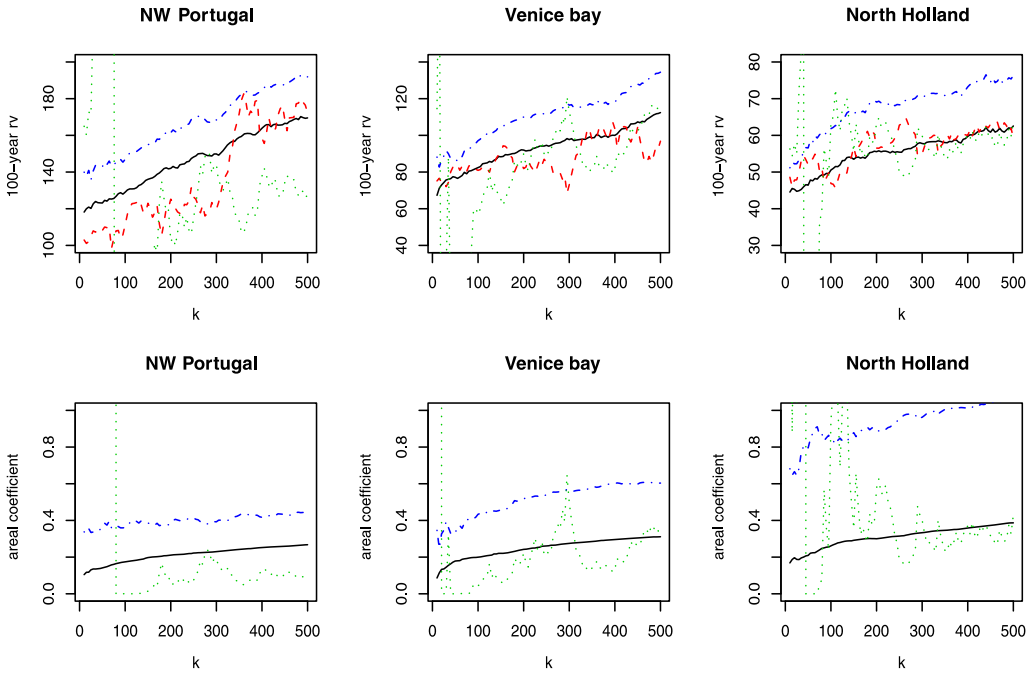

Figure 3. 100-year return value (mm) and θ estimates with k (FHZ – black line, 'simple aggregation' – dashed, $\hat{\theta}_1$ – dotted, $\hat{\theta}_2$ – dashed and dotted).

but with higher variability. The instability of $\hat{\theta}_1$ for small values of k is due to estimates of γ close to zero for small values of k. For $\hat{\theta}_2$, this estimator is systematically biased most notably in the North Holland case. By comparing the results for return values and θ estimation, it is clear the importance of having a good estimation of θ, with a clearly

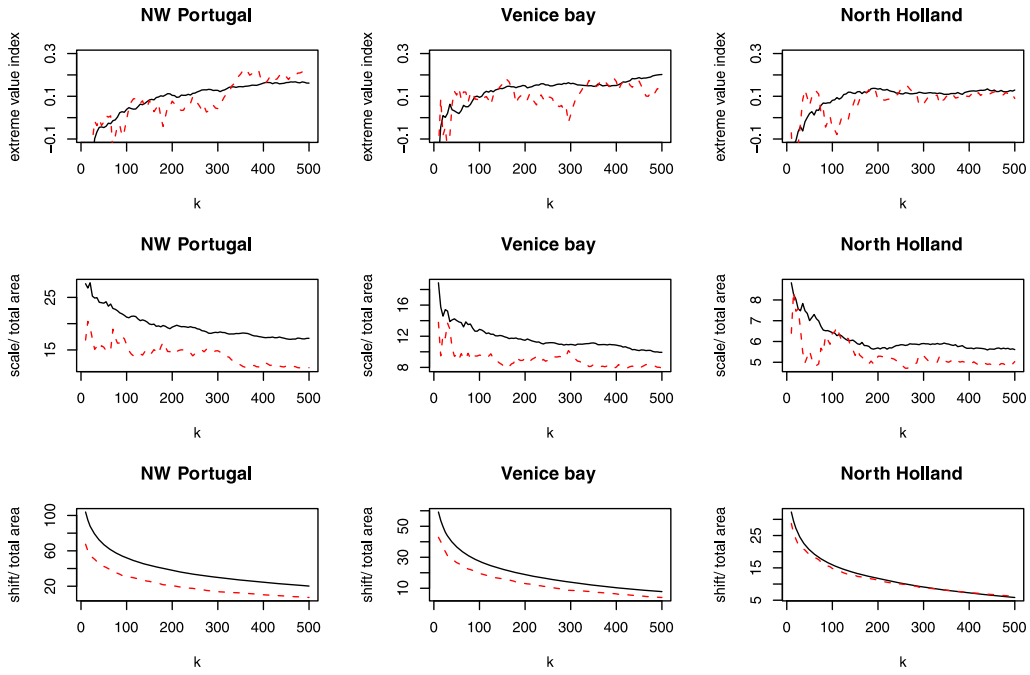

Figure 4. Estimates of the shape, scale and location parameters from (25)–(26) (black line), compared with the ones by 'simply aggregating' the data (dashed), with k.

visible influence in the corresponding return value. We also compare the return value results with the 100-year return value estimates obtained from 'simply aggregating' the data through the region and applying the univariate methods, i.e. discarding any spatial effect.

In Figure 4 are the corresponding shape, scale and location estimates used in both situations, i.e. when taking into account the spatial effect or not. Note that the aggregation smooths out the shape, scale and location estimates. On the contrary in the case of 'simply aggregating' the data (i.e. based on aggregating the data over the region and applying standard univariate extreme value theory) it is visible the influence of the gamma estimates in the corresponding return value estimates. That is, as expected the shape estimates obtained from (26) are very stable in all cases, at least when compared with the 'simple aggregation' procedure. Similar comments apply to the results on scale (and location) estimation though less visible (in the Appendix we add extra boxplots showing the influence of scale and location estimates in the marginal quantile estimates which is in agreement with what should be expected from standard univariate approach, cf. Figure 6).

In Figure 5 are the four curves of 100-year return value estimates, i.e. by using $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}$ from FHZ and the standard univariate approach as seen in Figure 3 but in a wider range of k. The univariate approach (i.e. by 'simply aggregating' the data) shows clearly the typical strong bias for large values of k. The spatial approach shows smoother curves, specially when using the estimates from FHZ and $\hat{\theta}_2$. In all, it is difficult to find a stability zone for picking up 'the best k', a more frequent feature when estimating high quantiles than shape parameter. On the other hand, if some stability zone is found, this is typically for slight higher value of k than when estimating the shape. Altogether from all the curves and graphics and given the many similar patterns throughout the different estimators and data sets, if to choose some stability zone for k we would suggest around 300 to 400, but closer to 300. In Table 2 are the 100-year return value and theta estimates with $k = 300$. Curiously with the same data set but based on a specific max-stable process (hence via a parametric approach) [1] estimated a 100-rv of 58.8 mm for the North Holland region.

In summary, accordingly to what has been observed in the marginal analysis there are substantial differences in the magnitudes of the return values for the three regions: NW Portugal have the highest magnitudes and variability, then Venice Bay and the lowest and more stable results are for North Holland. On the other hand, the magnitudes for theta are very similar throughout the three regions, around 0.2–0.3, and North Holland with a tendency for higher values.

From the quantile estimator (25), estimates for γ and θ positive but small should imply that the aggregation effect lowers the magnitudes of return value estimates. In Figure 5 we compare the 100-year return value estimates represented in Figure 3, with the marginal 100-year return value estimates from the first 4 stations of each region. Indeed, we see the influence of aggregating data and the

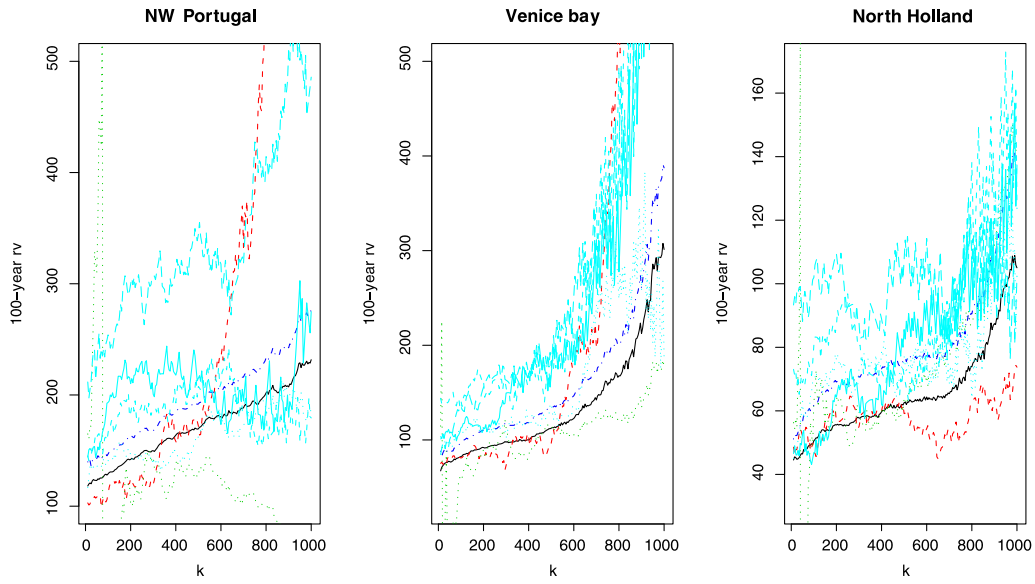

Figure 5. Comparison of marginal 100-year return value estimates (mm) for the first 4 stations of each region (extra 4 grey lines) with the ones from Figure 3, with k.

Table 2. Estimates with $k = 300$: the estimates for 100-year return values (mm) are calculated using the respective estimates of θ

Region	$\hat{rv}_{100}$			$\hat{\gamma}$	$\hat{a}_{n/k}$	$\hat{b}_{n/k}$
NW Portugal	149.8	144.2	168.4	0.13	18.3	30.0
	($\hat{\theta}_{FHZ} = .23$)	($\hat{\theta}_1 = .19$)	($\hat{\theta}_2 = .39$)			
Venice Bay	97.7	112.8	116.4	0.16	10.9	13.9
	($\hat{\theta}_{FHZ} = .27$)	($\hat{\theta}_1 = .50$)	($\hat{\theta}_2 = .57$)			
North Holland	57.8	59.9	70.7	0.11	5.9	9.1
	($\hat{\theta}_{FHZ} = .33$)	($\hat{\theta}_1 = .40$)	($\hat{\theta}_2 = .96$)			

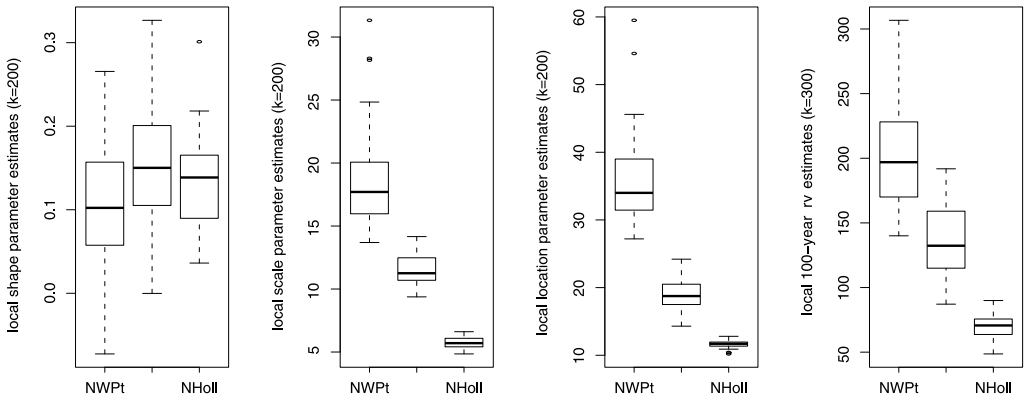

Figure 6. Boxplots of local estimates of the shape, scale and location parameters with $k = 200$, and the corresponding 100-year return values (mm) with $k = 300$.

role of the areal coefficient. Though being a unique number characterizing spatial dependence, the areal coefficient helps to explain the observed differences in quantiles when estimated locally and when based on spatially aggregated data.

ACKNOWLEDGEMENTS

The author would like to thank colleagues at Instituto D. Luiz and Institute of Meteorology and the Portuguese Water Institute (SNIRH, 2010) for providing the Portuguese data

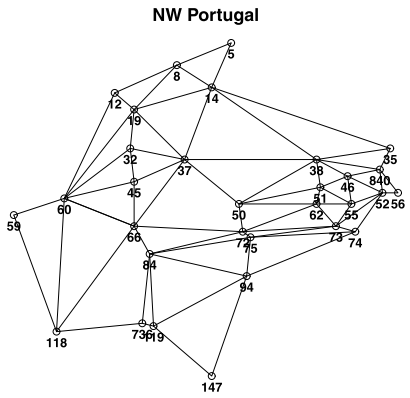

Figure 7. Spatial representation of the 31 precipitation stations from NW Portugal with a total area of approximately 3,754 Km^2.

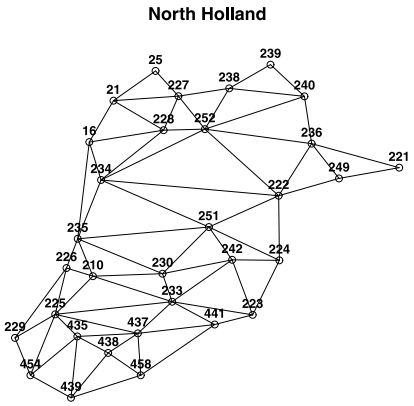

Figure 9. Spatial representation of the 32 precipitation stations from North Holland with a total area of approximately 2,010 Km^2.

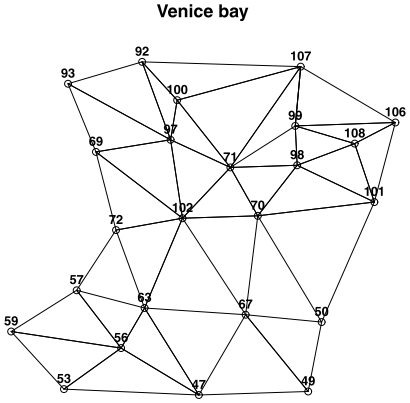

Figure 8. Spatial representation of the 24 precipitation stations from Venice Bay with a total area of approximately 3,074 Km^2.

and, specially to Rita Maria Cardoso and Pedro Soares for our nice discussions on the data, and Carlo Gaetan for the Italian data set. The North Holland data is the same data used in [8].

The author thanks Laurens de Haan for a useful suggestion. The author would also like to thank an anonymous referee for her/his constructive remarks.

APPENDIX A

In Figure 6 are boxplots of the local (or marginal) estimates of the shape, scale and location parameters (for the shape and scale these are the same represented in Figure 2) with $k = 200$ and, additionally, the local estimates of 100-year return values with $k = 300$. That is, basically apply simply univariate estimation by using at each site (20) with $\theta = 1$ and with the corresponding marginal shape, scale and location estimates; note that each 100 year return value estimate uses the marginal estimates with $k = 300$ and not with $k = 200$. With the former there are slight increases in variability and/or bias on the results but the overall picture is very similar. In particular, it is clear the influence of the scale and location estimates (and their variability) on the marginal 100-year rv estimates.

In Figures 7–9 we represent the sations over the space for the three regions, NW Portugal in Figure 7, North Holland in Figure 9 and Venice Bay in Italy in Figure 8. All stations are identified with their ID-numbers as provided in the original data sets from the Institutes.

Received 23 November 2013

REFERENCES

[1] Buishand, A., de Haan, L., and Zhou, C. (2008). On spatial extremes; with application to a rainfall problem. *Annals of Applied Statistics* **2**, 624–642. MR2524349

[2] Coles, S. G. and Tawn, J. A. (1996). Modelling extremes of the areal rainfall process. *J. R. Statist. Soc. B* **58**, 329–347. MR1377836

[3] Dekkers, A. L. M., Einmahl, J. H. J., and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Statist.* **17**, 1833–1855. MR1026315

[4] Dijk, V. and de Haan, L. (1992). On the estimation of the exceedance probability of a high level. In: *Order Statist. and Nonparametrics: Theory and Applications* (P. K. Sen and I. A. Salama, eds.), 79–92. North-Holland, Amsterdam. MR1229349

[5] Drees, H. (2000). Weighted approximations of tail processes for β-mixing random variables. *Ann. Appl. Probab.* **10**, 1274–1301. MR1810875

[6] Drees, H. (2003). Extreme quantile estimation for dependent data with applications to finance. *Bernoulli* **9**, 617–657. MR1996273

[7] Einmahl, J. H. J. and Lin, T. (2006). Asymptotic normality of extreme value estimators on C[0,1]. *Ann. Statist.* **34**, 469–492. MR2275250

[8] Ferreira, A., de Haan, L., and Zhou, C. (2012). Exceedance probability of the integral of a stochastic process. *Journal of Multivariate Analysis* **105**, 241–257. MR2877515

[9] Giné, E., Hahn, M. G., and Vatan, P. (1990). Max-infinitely divisible and max-stable sample continuous processes. *Probab. Th. Rel. Fields* **73**, 139–165. MR1080487

[10] de Haan, L. and Lin, T. (2001). On convergence toward an extreme value distribution in C[0,1]. *Ann. Probab.* **29**, 467–483. MR1825160

[11] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer, Boston. MR2234156

[12] de Haan, L. and Rootzén, H. (1993). On the estimation of high quantiles. *J. Statist. Planning and Inference* **35**, 1–13. MR1220401

[13] de Haan, L. and Stadtmüller, U. (1996). Generalized regular variation of second order. *J. Australian Math. Soc. (Series A)* **61**, 381–395. MR1420345

[14] Katz, R. W., Parlange M. B. and Naveau P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources* **25**, 1287–1304.

[15] Mannshardt-Shamseldin, E. C., Smith, R. L., Sain, S. R., Mearns, L. O., and Cooley, D. (2010). Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data. *Annals of Applied Statistics* **4**, 484–502. MR2758181

[16] Sistema Nacional de Informacão de Recursos Hídricos (2010). Available at http://snirh.pt/.

[17] Thibaud, E., Mutzner R., and Davison A. C. (2013). Threshold modelling of extreme spatial rainfall. *Water Resources Research* **49**, 4633–4644.

Ana Ferreira
ISA-Univ Lisboa and CEAUL
Instituto Superior de Agronomia, DCEB
Tapada da Ajuda
1349-017, Lisbon
Portugal
E-mail address: anafh@isa.utl.pt