



HAL
open science

An automatic procedure to select a block size in the continuous generalized extreme value model estimation

Pascal Sielenou Dkengne, Stéphane Girard, Samia Ahiad

► **To cite this version:**

Pascal Sielenou Dkengne, Stéphane Girard, Samia Ahiad. An automatic procedure to select a block size in the continuous generalized extreme value model estimation. 2020. hal-02952279

HAL Id: hal-02952279

<https://inria.hal.science/hal-02952279>

Preprint submitted on 29 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An automatic procedure to select a block size in the continuous generalized extreme value model estimation

Pascal Alain Dkengne Sielenou^{a,*}, Stéphane Girard^a, Samia Ahiad^b

^a*Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.*

^b*Valeo Driving Assistance Domain (34 rue St-André, ZI des Vignes. F-93012 BOBIGNY - France)*

Abstract

The block maxima approach is one of the main methodologies in extreme value theory to obtain a suitable distribution to estimate the probability of large values. In this approach, the block size is usually selected in order to reflect the possible intrinsic periodicity of the studied phenomenon. The generalization of this approach to data from non-seasonal phenomena is not straightforward. To address this problem, we propose an automatic data-driven method to identify the block size to use in the generalized extreme value (GEV) distribution for extrapolation. This methodology includes the validation of sufficient theoretical conditions ensuring that the maximum term converges to the GEV distribution. The selected GEV model can be different from the GEV model fitted on a sample of block maxima from arbitrary large block size. This selected GEV model has the special property to associate high values of the underlining variable with the corresponding smallest return periods. Such a model is useful in practice as it allows, for example, a better sizing of certain structures of protection against natural disasters. To illustrate the developed method, we consider two real datasets. The first dataset contains daily observations over several years from some meteorological variables while the second dataset contains data observed at millisecond time scale over several minutes from sensors in the field of vehicle engineering.

Keywords:

Extreme value distribution, Block maxima, Block size selection, Meteorological data, Sensors reliability

1. Introduction

Let X be a random variable (associated with the phenomenon of interest) for which we want to assess the probability of extreme events. Let X_1, \dots, X_n be n independent copies of X . Define the sample maximum by $M_n = \max\{X_1, \dots, X_n\}$. The main goal of extreme value analysis is to appropriately estimate for a large value $x \geq M_n$ the following probability

$$\mathbb{P}\{X > x\}. \tag{1}$$

*Corresponding Author: Pascal Alain Dkengne Sielenou

Email addresses: sielenou_alain@yahoo.fr (Pascal Alain Dkengne Sielenou), stephane.girard@inria.fr (Stéphane Girard), samia.ahiad@valeo.com (Samia Ahiad)

6 The inverse of the probability (1) is defined as the return period T of x . In other words, T is the time
7 period during which X is expected to exceed on average once the value x . It is clear that classical statistical
8 methods are not applicable to solve the above problem. Indeed, for $x \geq M_n$ the empirical estimation of
9 the probability (1) is equal to zero as there is no observation beyond the sample maximum. Moreover,
10 a parametric estimation may not be reliable either since a good fit in the distribution bulk does not
11 necessarily yield a good fit in the tail. For instance, both Gaussian and Student distributions can fit very
12 well a given set of observations whereas the behavior of large values from the fitted Student distribution is
13 significantly different from the behavior of large values from the fitted Gaussian distribution. Extreme
14 value theory provides the solid fundamentals needed for the statistical modeling of extreme events and
15 the computation of probabilities such as (1). The strength of extreme value theory is that, ideally, the
16 original parent distribution function of X needs not to be known, because the maximum term M_n , up
17 to linear normalization, asymptotically follows a distribution nowadays called generalized extreme value
18 (GEV) family (e.g. Fisher and Tippett, 1928; Gnedenko, 1943; Leadbetter et al., 1983; Embrechts et al.,
19 1997; Coles, 2001; Beirlant et al., 2004). Consequently, a sample of M_n (also called block maxima) where
20 the nonnegative integer n (referred to as block size) approaches infinity can be approximated by the GEV
21 distribution as stated in Theorem 1.1 from Coles (2001).

22 **Theorem 1.1.** *If there exist sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that*

$$\mathbb{P} \left\{ \frac{M_n - b_n}{a_n} \leq x \right\} \rightarrow G(x) \quad (2)$$

23 *as $n \rightarrow +\infty$ for a non-degenerate distribution function G , then G belongs to the Generalized Extreme*
24 *Value (GEV) family*

$$G(x) = G(x; \mu, \sigma, \gamma) = \exp \left\{ - \left[1 + \gamma \left(\frac{x - \mu}{\sigma} \right) \right]^{-\frac{1}{\gamma}} \right\}, \quad (3)$$

25 *defined on $\{x \in \mathbb{R} : 1 + \gamma \left(\frac{x - \mu}{\sigma} \right) > 0\}$, where $\gamma, \mu \in \mathbb{R}, \sigma > 0$.*

26 The distribution G includes three parameters: the location parameter μ , the scale parameter σ and
27 the shape parameter γ also referred to as the extreme value index. The GEV family can be divided into
28 three families, namely the Fréchet family, the Weibull family and the Gumbel family. The Fréchet and
29 the Weibull families correspond respectively to the cases where $\gamma > 0$ and $\gamma < 0$. The Gumbel family with
30 $\gamma = 0$ is interpreted as the limit of (3) as $\gamma \rightarrow 0$, leading to the distribution

$$G(x) = \exp \left\{ - \exp \left\{ - \left(\frac{x - \mu}{\sigma} \right) \right\} \right\}, \quad x \in \mathbb{R}. \quad (4)$$

31 By Taylor expansion, one can observe that the Fréchet family has a power law decaying tail whereas the
32 Gumbel family has an exponentially decaying tail (Embrechts et al., 1997). Consequently, the Fréchet
33 family suits well heavy tailed distributions (e.g. the Pareto and the Loggamma distributions) while the
34 Gumbel family characterizes light tailed distributions (e.g. the Gaussian and the Gamma distributions).
35 Finally, the Weibull family is the asymptotic distribution of finite right endpoint distributions such as the
36 Uniform and the Beta distributions.

37 Each of the extreme value models derived so far has been obtained through mathematical arguments
38 that assume an underlying process consisting of a sequence of independent random variables. However,
39 for some data to which extreme value models are commonly applied, temporal independence is usually an
40 unrealistic assumption. Extreme conditions often persist over several consecutive observations, bringing
41 into question the appropriateness of models such as GEV distributions. A detailed investigation of this
42 question is given in [Leadbetter et al. \(1983\)](#). The dependence in stationary series can take many different
43 forms, and it is impossible to develop a general characterization of the behaviors of extremes unless some
44 constraints are imposed. These conditions aim to ensure that the gap to independence between sets of
45 variables that are far enough apart is sufficiently close to zero to have no effect on the limit laws for
46 extremes. A summary of the obtained results is given in Theorem 1.2 from [Coles \(2001\)](#).

Theorem 1.2. *Let X_1, X_2, \dots be a stationary process and X_1^*, X_2^*, \dots be a sequence of independent variables with the same marginal distribution. Define $M_n = \max\{X_1, \dots, X_n\}$ and $M_n^* = \max\{X_1^*, \dots, X_n^*\}$. Under suitable regularity conditions,*

$$\lim_{n \rightarrow +\infty} \mathbb{P} \left\{ \frac{M_n^* - b_n}{a_n} \leq x \right\} = G_1(x)$$

for normalizing sequences $a_n > 0$ and $b_n \in \mathbb{R}$, where G_1 is a non-degenerate distribution function, if and only if

$$\lim_{n \rightarrow +\infty} \mathbb{P} \left\{ \frac{M_n - b_n}{a_n} \leq x \right\} = G_2(x),$$

47 where

$$G_2(x) = G_1^\theta(x) \tag{5}$$

48 for some $\theta \in (0, 1]$.

49 Since the marginal distributions of the X_i and X_i^* are the same, any difference in the limiting
50 distribution of maxima must be attributable to the dependence of the X_i series. The parameter θ defined
51 by (5) is called the extremal index. This quantity summarizes the strength of dependence between
52 extremes in a stationary sequence. Theorem 1.2 implies that, if maxima of a stationary series converge,
53 provided that an appropriate condition is satisfied, the limit distribution is related to the limit distribution
54 of an independent series according to equation (5). The effect of dependence in stationary series is simply
55 a replacement of G_1 as the limit distribution, which would have arisen for the associated independent
56 series with same marginal distribution, with G_1^θ . This is consistent with Theorem 1.1, because if G_1 is a
57 GEV distribution, so is G_1^θ . According to the foregoing, if the limiting distribution of a random sequence
58 $M_n = \max\{X_1, \dots, X_n\}$ from a stationary sequence X_1, X_2, \dots is non degenerate, then the probability
59 distribution of the sample maxima M_n can be approximated by the continuous GEV distribution family
60 for large values of n . One of the practical methodologies for statistical modeling of extreme values consists
61 to apply the block maxima approach. In this method, data are splitted into sequences of observations of
62 length n , for some large value of n , generating a series of m block maxima, $M_{n,1}, M_{n,2}, \dots, M_{n,m}$, say, to

63 which the generalized extreme value distribution can be fitted. The choice of a block size n is equivalent
64 to the choice of the number m of block maxima. The delicate point of this method is the appropriate
65 choice of the time periods defining blocks. Indeed, a too high value of n results in too few block maxima
66 and consequently high variance estimators. For too small n , estimators become biased. A similar issue is
67 the selection of threshold in the peak over threshold (POT) method for fitting the generalized Pareto
68 distribution to excesses (Tancredi et al., 2006; Scarrott and MacDonald, 2012; Wu and Qiu, 2018; Yang
69 et al., 2018).

70 The block maxima method has been widely used in extreme value modeling of seasonal data such
71 as wind speeds, flood and rainfall by setting for example, with a year as block size when data are daily
72 observed. For non seasonal data from other fields such as vehicle engineering, the selection of an optimal
73 block size is still a problem. Some recent studies in the literature have attempted to solve this issue Wang
74 et al. (2016); Esra Ezgi et al. (2018); Özari et al. (2019). The method proposed by Esra Ezgi et al. (2018)
75 and Özari et al. (2019) can be summarized as follows. The last 10% part of the actual data is reserved as
76 test data. GEV models are fitted to different samples of block maxima from the first 90% part of the
77 actual data. The estimated GEV models are used to generate samples (also referred to as predicted data)
78 of size equal to that of test data. The selected block size is associated with the GEV model for which the
79 highest similarity is observed between large values from the predicted and test data. Our main comment
80 about this method is that the use of only one test data may not be enough to guarantee that the resulting
81 GEV model is suitable to characterize large values from future data. To continue reviewing the literature,
82 one can sum up the method developed by Wang et al. (2016) as follows. GEV models are fitted on
83 different samples of block maxima from the actual data. The goodness-of-fit (g.o.f.) of the estimated GEV
84 models is evaluated by means of an entropy based indicator which includes three g.o.f. measures, namely
85 the Kolmogorov Smirnov, the Chi-square and the average deviation in probability density function. The
86 selected block size is associated with the GEV model for which the smallest value of the above mentioned
87 g.o.f. indicator is observed. Our main comment about this method is that the resulting GEV model
88 exhibits a better fitting result. However, the selected GEV distribution does not necessarily have desired
89 property to associate high values of the underlining variable with the corresponding smallest return period.
90 The rest of this study is designed to explain the theoretical and practical aspects of the methodology
91 we propose to achieve this block size selection goal. Section 2 presents the proposed block size selection
92 procedure along with the related theoretical framework. An approach to assess the practical performances
93 of this methodology is described in Section 3. Section 4 illustrates the practical applications of the block
94 size selection procedure on real datasets. Tables, figures and additional results are postponed to the
95 appendix.

96 **2. Block size selection procedure**

97 This section aims at providing an answer to the following natural question which arises in practice:
 98 “Given a continuous stationary sequence X_1, X_2, \dots , how can we choose the value of n which guarantees
 99 that the GEV model fitted to the sample maxima $M_n = \max\{X_1, \dots, X_n\}$ is appropriate for extrapolation?”

100 *2.1. Theoretical foundations*

101 In the sequel, we exploit Theorem 2.1 to provide an heuristic answer to the above question which is
 102 valid for both continuous and discrete random variables.

Theorem 2.1. *Let X_1, X_2, \dots , be a continuous stationary sequence. Let $M_n = \max\{X_1, \dots, X_n\}$. Under suitable regularity conditions, suppose that for large n , there are constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that for all $x \in \mathbb{R}$*

$$\lim_{n \rightarrow +\infty} \mathbb{P}\{M_n \leq a_n x + b_n\} = G(x; \mu, \sigma, \gamma),$$

103 for some constants $\mu \in \mathbb{R}$, $\sigma > 0$ and $\gamma \in \mathbb{R}$, where G is the GEV distribution function. Then for all
 104 non-negative integer $j > 1$, we have

$$\lim_{n \rightarrow +\infty} \mathbb{P}\{M_{j \times n} \leq a_n x + b_n\} = G(x; \mu_j, \sigma_j, \gamma_j), \quad (6)$$

105 where for $\gamma \neq 0$,

$$\mu_j = \mu + \sigma \left(\frac{j^\gamma - 1}{\gamma} \right), \quad \sigma_j = \sigma j^\gamma, \quad \gamma_j = \gamma \quad (7)$$

and for $\gamma = 0$,

$$\mu_j = \mu + \sigma \log(j), \quad \sigma_j = \sigma.$$

Proof of Theorem 2.1. Let X_1^*, X_2^*, \dots be a continuous sequence of independent and identically distributed random variables whose common distribution is the marginal distribution of the stationary sequence X_1, X_2, \dots . Define $M_n^* = \max\{X_1^*, \dots, X_n^*\}$. The idea is to consider $M_{j \times n}^*$, the maximum random variable in a sequence of $j \times n$ variables for some large value of n , as the maximum of j maxima, each of which is the maximum of n observations. From Theorem 1.2, there exists $\theta \in (0, 1]$ such that the following equality holds true for all $j > 1$.

$$\lim_{n \rightarrow +\infty} \mathbb{P}\{M_{j \times n} \leq a_n x + b_n\} = \left[\left(\lim_{n \rightarrow +\infty} \mathbb{P}\{M_n^* \leq a_n x + b_n\} \right)^\theta \right]^j.$$

Hence, one can write

$$\lim_{n \rightarrow +\infty} \mathbb{P}\{M_{j \times n} \leq a_n x + b_n\} = \left(\lim_{n \rightarrow +\infty} \mathbb{P}\{M_n \leq a_n x + b_n\} \right)^j = (G(x; \mu, \sigma, \gamma))^j.$$

106 The conclusion follows from a straightforward algebraic computation of $(G(x; \mu, \sigma, \gamma))^j = G(x; \mu_j, \sigma_j, \gamma)$.

107 □

108 A natural technique to identify potential candidates for the optimal block size consists in fitting the
 109 GEV distribution at a range of block sizes, and to look for stability of parameter estimates. The argument
 110 is as follows. By Theorem 2.1, if a GEV distribution is a reasonable model for block maxima of a block
 111 size n_0 , then block maxima of block size $n_j = j \times n_0$ for any integer $j > 1$, should also follow a GEV
 112 distribution with the same shape parameters. However, the location parameter μ_j and the scale parameter
 113 σ_j are expected to change with j as in formula (6) and (7). By reparametrizing the GEV distribution
 114 parameters when $\gamma \neq 0$ as

$$\mu^* = \mu_j - \sigma_j j^{-\gamma} \left(\frac{j^\gamma - 1}{\gamma} \right), \quad \sigma^* = \sigma_j j^{-\gamma} \quad (8)$$

115 and when $\gamma = 0$ as

$$\mu^* = \mu_j - \sigma_j \log(j), \quad \sigma^* = \sigma_j \quad (9)$$

116 the estimates $\hat{\gamma}$, $\hat{\sigma}^*$ and $\hat{\mu}^*$, of γ , σ^* and μ^* should be constant (up to estimation uncertainty) if n_0 is a
 117 valid block size for sample maxima to follow the GEV distribution. This argument suggests plotting $\hat{\gamma}$, $\hat{\sigma}^*$
 118 and $\hat{\mu}^*$, together with their respective confidence intervals, and selecting for each normalized parameter
 119 an integer n_0 as the lowest value for which these estimates remain approximately constant for almost all
 120 $n_j = j \times n_0$ with $j \geq 1$. Uncertainty in the estimation of the normalized GEV distribution parameters μ^*
 121 and σ^* can be assessed by using the delta method as follows. For $\gamma = 0$, the asymptotic variance of the
 122 rescaled location parameter is

$$\text{Var}(\hat{\mu}^*) = (\nabla \hat{\mu}^*)^T \text{V}(\hat{\mu}_j, \hat{\sigma}_j) \nabla \hat{\mu}^*, \quad (10)$$

where $\text{V}(\hat{\mu}_j, \hat{\sigma}_j)$ is the asymptotic variance-covariance matrix of the joint estimate $(\hat{\mu}_j, \hat{\sigma}_j)$ of the parameter
 (μ_j, σ_j) . Here, the gradient is calculated by the following formula

$$(\nabla \hat{\mu}^*)^T = \left[\frac{\partial \hat{\mu}^*}{\partial \hat{\mu}_j}, \frac{\partial \hat{\mu}^*}{\partial \hat{\sigma}_j} \right] = [1, -\log(j)].$$

123 Similarly, for $\gamma \neq 0$, the asymptotic variances of the rescaled location parameter and the rescaled scale
 124 parameter are

$$\begin{cases} \text{Var}(\hat{\mu}^*) = (\nabla \hat{\mu}^*)^T \text{V}(\hat{\mu}_j, \hat{\sigma}_j, \hat{\gamma}_j) \nabla \hat{\mu}^* \\ \text{Var}(\hat{\sigma}^*) = (\nabla \hat{\sigma}^*)^T \text{V}(\hat{\mu}_j, \hat{\sigma}_j, \hat{\gamma}_j) \nabla \hat{\sigma}^* \end{cases} \quad (11)$$

where $\text{V}(\hat{\mu}_j, \hat{\sigma}_j, \hat{\gamma}_j)$ is the asymptotic variance-covariance matrix of the joint estimate $(\hat{\mu}_j, \hat{\sigma}_j, \hat{\gamma}_j)$ of the
 parameter $(\mu_j, \sigma_j, \gamma_j)$. Here, the gradients are calculated by the following formula in which $\hat{\gamma}_j$ is denoted
 by $\hat{\gamma}$ for the sake of clarity

$$(\nabla \hat{\mu}^*)^T = \left[\frac{\partial \hat{\mu}^*}{\partial \hat{\mu}_j}, \frac{\partial \hat{\mu}^*}{\partial \hat{\sigma}_j}, \frac{\partial \hat{\mu}^*}{\partial \hat{\gamma}} \right] = \left[1, -j^{-\hat{\gamma}} \left(\frac{j^{\hat{\gamma}} - 1}{\hat{\gamma}} \right), \hat{\sigma}_j \left(\frac{1 - j^{-\hat{\gamma}} (\hat{\gamma} \log(j) + 1)}{\hat{\gamma}^2} \right) \right],$$

and

$$(\nabla \hat{\sigma}^*)^T = \left[\frac{\partial \hat{\sigma}^*}{\partial \hat{\mu}_j}, \frac{\partial \hat{\sigma}^*}{\partial \hat{\sigma}_j}, \frac{\partial \hat{\sigma}^*}{\partial \hat{\gamma}} \right] = \left[0, j^{-\hat{\gamma}}, -\hat{\sigma}_j j^{-\hat{\gamma}} \log(j) \right].$$

125 Let $M_{n_j} = (M_{n_j,1}, \dots, M_{n_j,m_j})$ be the sample maxima associated with the block size $n_j = j \times n_0$
 126 with $j \geq 1$, where n_0 is the minimum block size which simultaneously stabilizes the three parameters $\hat{\gamma}$,
 127 $\hat{\sigma}^*$ and $\hat{\mu}^*$. It is easily shown that the rescaled random variable $M_{n_j}^*$ defined by

$$M_{n_j}^* = \frac{M_{n_j} - \mu_{n_j}}{\sigma_{n_j}} \quad (12)$$

128 is expected to follow the GEV model having the distribution function $G(\cdot; \gamma, \sigma = 1, \mu = 0)$. It follows
 129 that the values of the random variable $M_{n_j}^*$ are expected to be large as the shape parameter γ increases.
 130 Making use of transformation (12) together with formula (6), the sample maxima M_{n_j} can be written as

$$M_{n_j} = \sigma_{n_0} j^\gamma M_{n_j}^* + \mu_{n_0} + \sigma_{n_0} j^\gamma \left(\frac{j^\gamma - 1}{\gamma} \right). \quad (13)$$

131 By standard calculations, one can show that the values of the random variable M_{n_j} are also expected
 132 to be large as the shape parameter γ increases. Besides, it is straightforward to see that for all $n'_0 \geq n_0$
 133 and $j \geq 1$, we have $M_{n'_j} \geq M_{n_j}$ where $n'_j = j \times n'_0$ and $n_j = j \times n_0$. Consequently, the GEV distribution
 134 fitted to any sample of block maxima M_n whose block size n is greater than n_0 is expected to also have
 135 shape parameter γ . It results from the foregoing that the desired block size (that is, leading to the largest
 136 extrapolated values) must be greater than n_0 and must be associated with the GEV distribution having
 137 the largest estimated shape parameter γ . In other words, the selected block size for extrapolation is
 138 associated with the heaviest tailed and stable fitted GEV distribution. With such a GEV model, two types
 139 of predictions can be made: the frequency associated with a given intensity phenomenon or the intensity
 140 of a phenomenon having a given frequency. In both cases, this GEV model will provide a prediction
 141 of the greatest possible quantity of interest (frequency or intensity) associated with the phenomenon
 142 under consideration. Such estimates should allow to make decisions that will significantly reduce the risks
 143 associated with increasingly extreme events.

144 2.2. Algorithmic procedure

145 In Section 2.1, we argued that the quality of a fitted GEV model to a sample maxima depends on
 146 the value of the considered block size. We also suggested therein the outline of a block size selection
 147 procedure. The main idea of this procedure consists of the following three stages. In the first stage, fit the
 148 GEV distribution on samples of block maxima from a range of block sizes. In the second stage, identify
 149 the stabilizing block size as the minimum block size which simultaneously stabilizes the shape parameter
 150 $\gamma \neq 0$, the normalized location parameter μ^* and scale parameter σ^* defined by (8). In the third stage,
 151 the selected block size is the one associated with the largest estimated shape parameter which is not
 152 significantly different from the estimated shape parameter at the stabilizing block size. Obviously, the
 153 selected block size is expected to be greater than the stabilizing block size.

154 To fit the GEV models on samples of block maxima, we apply the maximum likelihood estimation
 155 procedure, which is one of the most popular inference method for extreme value models (Hosking, 1985;
 156 Smith, 1985; Coles, 2001; Gilleland and Katz, 2016). To check the stability of the normalized GEV

157 distribution parameters it is sufficient to check if their estimated values are approximately constant when
 158 the block size is large enough so that the corresponding sequence of sample maxima is stationary and is
 159 well fitted by the GEV distribution. To check the goodness of fit, we use the Kolmogorov Simirnov (KS)
 160 test (Chakravarti et al., 1967; Durbin, 1973). In this case, the null hypothesis is that the distribution
 161 function which generates the sample maxima is the fitted GEV distribution. We use two tests to check
 162 the stationarity of a given sequence of block maxima. The first test is the Augmented Dickey-Fuller
 163 (ADF) test (Said and Dickey, 1984; Banerjee et al., 1993; Trapletti and Hornik, 2018). In this case,
 164 the null hypothesis is that the sequence of block maxima is non-stationary. The second test is the
 165 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski et al., 1992; Trapletti and Hornik, 2018).
 166 In this case, the null hypothesis is that the sequence of block maxima is stationary. We validate that a
 167 sequence of block maxima is stationary if at least one of these two tests does not reject this hypothesis at
 168 a given level of significance. Recall that the null hypothesis of a test is rejected if the obtained p -value
 169 is less than the considered value of the test significance level $\alpha \in (0, 1)$. Moreover, the probability of
 170 rejecting the null hypothesis when it is in fact false (a correct decision) is $1 - \alpha$ as the p -value of a test
 171 statistic is the probability of rejecting the null hypothesis when it is in fact true (an incorrect decision).

172 Algorithm 1 describes the main steps of our procedure to select a block size to use in the block
 173 maxima modeling approach. The theoretical justifications are provided in Section 2.1. Consider a data set
 174 $\mathcal{X} = (x_1, \dots, x_n)$ of n observations whose extreme values are to be modeled with the aim to extrapolate
 175 beyond the largest observed value. Algorithm 1 can also be considered as a GEV model determination
 176 procedure. Indeed, the main output of the developed procedure is the heaviest tailed GEV distribution
 177 function $G_{z_{i^*}}$ fitted to a specific sample of block maxima z_{i^*} , where all required theoretical properties are
 178 satisfied at the block size i^* .

Algorithm 1 Block size selection

Stage 1: Obtain I samples of block maxima, denoted by $z_i = (z_{i,1}, \dots, z_{i,m(i)})$ in which $i = i_{\min}, i_{\min} + 1, \dots, I$ is the considered block size. The constants are explained below.

- $m(i) = \lceil \frac{n}{i} \rceil$. Here, $\lceil y \rceil$ is the smallest integer greater than or equal to y , and I is the largest block size, that is the size $m(I)$ of the corresponding sample maxima is the minimum size required for the estimation of GEV distribution parameters by the maximum likelihood method. In this study, we set $m(I)$ to 25.
- i_{\min} is the smallest block size which ensures that all block maxima associated with higher block sizes are strictly greater than the eventual excess of zero-counts from the \mathcal{X}' 's observations. This concerns a continuous random event containing excess zero-count data in unit time (e.g. daily accumulated precipitation or snow amount). We refer to candidate GEV models all GEV models fitted on samples of block maxima associated with block sizes $i \geq i_{\min}$.
- $z_{i,j}$ is the maximum of the \mathcal{X}' 's observations within the j -th block of size i .

Stage 2: For $i = 1, \dots, I$ do the following tasks.

- Carry out the ADF stationary test on the sample maxima z_i and record the p -value, denoted by $p_{i,\text{ADF}}$, of the test statistic.
- Carry out the KPSS stationary test on sample maxima z_i and record the p -value, denoted by $p_{i,\text{KPSS}}$, of the test statistic.

Stage 3: For $i = 1, \dots, I$ do the following tasks.

- Use the maximum likelihood estimation method to fit the GEV distribution with non zero shape parameter to each sample maxima z_i .
 - Carry out the KS test to check the goodness-of-fit of the sample maxima z_i with the GEV distribution. Then record the p -value, denoted by $p_{i,\text{KS}}$, of the test statistic.
 - Construct a $100 \times (1 - \alpha)\%$ -confidence interval for the normalized location parameter μ^* , the normalized scale parameter σ^* and the (normalized) shape parameter $\gamma^* = \gamma \neq 0$, denoted by $C_i(\mu^*)$, $C_i(\sigma^*)$ and $C_i(\gamma^*)$, respectively. To construct such confidence intervals, one can use formula (11) provided in Section 2.1 to approximate the variance of the MLE of μ^* and σ^* .
-

Stage 4: Compute the subset S of block sizes defined by

$$S = \{i = 1, \dots, I : p_{i,\text{KS}} \geq \alpha, \text{ and } (p_{i,\text{ADF}} < \alpha \text{ or } p_{i,\text{KPSS}} \geq \alpha)\},$$

where $\alpha \in (0, 1)$ is the significance level for the tests. The set S contains all block sizes i for which the sample maxima z_i is stationary and is in adequacy with the GEV distribution.

Stage 5: Compute the three subsets $S(\gamma^*)$, $S(\sigma^*)$ and $S(\mu^*)$, where

$$S(\gamma^*) = \{i \in S : C_i(\gamma^*) \cap C_j(\gamma^*) \neq \emptyset, \forall j \in S \setminus \{i\}\},$$

$$S(\sigma^*) = \{i \in S : C_i(\sigma^*) \cap C_j(\sigma^*) \neq \emptyset, \forall j \in S \setminus \{i\}\},$$

$$S(\mu^*) = \{i \in S : C_i(\mu^*) \cap C_j(\mu^*) \neq \emptyset, \forall j \in S \setminus \{i\}\}.$$

It results that $S(\gamma^*)$, $S(\sigma^*)$ and $S(\mu^*)$ are the highest subsets of block sizes for which the confidence intervals $C_i(\gamma^*)$, $C_i(\sigma^*)$ and $C_i(\mu^*)$ respectively satisfy the conditions:

$$\bigcap_{i \in S(\gamma^*)} C_i(\gamma^*) \neq \emptyset, \quad \bigcap_{i \in S(\sigma^*)} C_i(\sigma^*) \neq \emptyset, \quad \bigcap_{i \in S(\mu^*)} C_i(\mu^*) \neq \emptyset.$$

This means that the estimated values of the normalized GEV distribution parameters $\gamma^* \neq 0$, σ^* and μ^* are approximately constant for all block sizes in the sets $S(\gamma^*)$, $S(\sigma^*)$ and $S(\mu^*)$, respectively.

Stage 6: Perform the following tasks.

- i) Find the smallest elements of the sets $S(\gamma^*)$, $S(\sigma^*)$ and $S(\mu^*)$, and denote them by $i(\gamma^*)$, $i(\sigma^*)$ and $i(\mu^*)$, respectively.
 - ii) Set $i_0 = \max\{i(\gamma^*), i(\sigma^*), i(\mu^*)\}$. It is natural to consider i_0 as the block size which simultaneously stabilizes the normalized GEV distribution parameters $\gamma^* \neq 0$, σ^* and μ^* . We refer to equivalent GEV models all GEV models fitted on samples of block maxima associated with block sizes $i \geq i_0$ and $i \in S(\gamma^*)$.
 - iii) Select the block size as $i^* = \arg \max_{i \geq i_0} \{\gamma_i : i \in S(\gamma^*)\}$, where γ_i is the shape parameter of the GEV distribution fitted to the sample maxima z_i .
-

179 **3. Performance assessment**

180 Let (X_1, X_2, \dots) be a continuous stationary sequence. Denote the corresponding sequence of maximum
 181 term by $M_n = \max\{X_1, \dots, X_n\}$, where $n \in \mathbb{N}$. Recall that under some regularity conditions, the limiting
 182 distribution of the random variable M_n is expected to be a member of the GEV distribution family (3).
 183 The quantities of interest are not the GEV distribution parameters themselves, but the quantiles, also
 184 called return levels, of the estimated GEV distribution. The return level $x(T)$ associated with return
 185 period $T > 1$ of the stationary sequence X can be calculated from the GEV distribution as

$$x(T) = \mu - \frac{\sigma}{\gamma} \left\{ 1 - \left[-\log \left(1 - \frac{1}{T} \right) \right]^{-\gamma} \right\}. \quad (14)$$

186 The maximum likelihood estimator (MLE) of the parameter vector $\psi = (\mu, \sigma, \gamma)$ requires a sample
 187 $z = (z_1, \dots, z_m)$ of block maxima, where the block size is sufficiently large. It is worth noticing that
 188 the return period T in (14) is expressed in terms of number blocks. In some fields such as meteorology,
 189 hydrology and glaciology, it is often convenient to express this return period in terms of a unit time
 190 duration (second, minute, hour, day, month or year). Recall that the T -block return level is the level
 191 expected to be exceeded in average once every T blocks of raw observations. One can use the following
 192 relationship $n_b \times T = n_d \times D$ to convert a return period T whose unit is the number of block having size
 193 n_b (also referred to as the number of raw observations per block) to the return period D whose unit is a
 194 given unit time duration in which there are n_d raw observations.

195 It is well known that under certain regularity conditions, the MLE of ψ is normally distributed as
 196 m approaches infinity. In such a case, the distribution of any functions of the MLE of ψ , such as return
 197 levels, can also be approximated by a Gaussian distribution. The performance of Algorithm 1 is assessed
 198 by comparing true return levels from the parent distribution of observations with those estimated from
 199 the selected GEV model. Consider a sample $\mathcal{X} = (x_1, \dots, x_n)$ of n observations from a known parametric
 200 probability distribution with cumulative distribution function $F(\cdot; \phi)$, where $\phi \in \mathbb{R}^d$ for some $d \in \mathbb{N}$ is the
 201 parameter. Set a desired level of significance $\alpha \in (0, 1)$. Our validation approach consists in showing that
 202 the $100 \times (1 - \alpha)\%$ -confidence intervals of all estimated return levels from the selected GEV model contain
 203 the corresponding true return levels from the known distribution $F(\cdot; \phi)$. For the sake of simplicity, we
 204 generate a sample \mathcal{X} of independent observations. Moreover, the verification is performed by means of
 205 a bootstrap scheme to show that the conclusion is not specific to the considered sample \mathcal{X} . The above
 206 procedure can be implemented and evaluated automatically thanks to Algorithm 2 in which we used the
 207 following quantities as input:

- 208 • The sample size $n = 2 \times 10^4$.
- 209 • The significance level $\alpha = 5\%$.
- 210 • The bootstrap parameter $B = 200$.

- 211 • The known distribution $F(\cdot; \phi)$ to generate samples and to compute the true return levels.
- 212 • The discrete set $T = \{T^{(j)}, j = 1, \dots, J\}$ of return periods to estimate the corresponding return
- 213 levels, where $n \leq T^{(j)} \leq 10^{15}$ is the j -th smallest element of the set T . Here, we take 81 equispaced
- 214 values of $T^{(j)}$ between two consecutive powers of ten. Thus, the set T contains $J = 872$ values.

215 Density functions $f(\cdot; \phi)$ associated with the considered known families of distribution functions $F(\cdot; \phi)$
 216 are listed below:

- 217 1. The gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ has probability
- 218 density function defined for $x > 0$ by

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\}, \quad (15)$$

where $\Gamma(\cdot)$ is the gamma function defined by

$$\Gamma(z) = \int_0^{+\infty} x^{z-1} \exp\{-x\} dx, \quad z > 0.$$

219 The validation results are gathered in the top panel of Figures C1-C2 when using this distribution
 220 with the following parameters: $\alpha = 2$ and $\beta = 1$.

- 221 2. The loggamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ has probability
- 222 density function defined for $x > 0$ by

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{(\log x)^{\alpha-1}}{x^{\beta+1}}. \quad (16)$$

223 The validation results are gathered in the center panel of Figures C1-C2 when using this distribution
 224 with the following parameters: $\alpha = 2$ and $\beta = 5$.

- 225 3. The normal distribution with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$ has probability
- 226 density function defined for $x > 0$ by

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \quad (17)$$

227 The validation results are gathered in the bottom panel of Figures C1-C2 when using this distribution
 228 with the following parameters: $\mu = 0$ and $\sigma = 1$.

- 229 4. The generalized extreme value (GEV) distribution with parameters $\mu \in \mathbb{R}$, $\sigma > 0$ and $\gamma \in \mathbb{R}$ has
- 230 probability density function defined for $\gamma \neq 0$ and $x \in \mathbb{R}$ such that $1 + \gamma \left(\frac{x-\mu}{\sigma}\right) > 0$ by

$$f(x; \mu, \sigma, \gamma) = \frac{1}{\sigma} \left[1 + \gamma \left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\gamma-1} \exp\left\{-\left[1 + \gamma \left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\gamma}\right\}. \quad (18)$$

231 The case where $\gamma = 0$ corresponds to the Gumbel distribution whose density is defined for $x \in \mathbb{R}$ by

$$f(x; \mu, \sigma) = \frac{1}{\sigma} \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\} \exp\left\{-\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right\}. \quad (19)$$

232 The validation results are gathered in Figures C3-C4 when using this distribution with the three
 233 following vectors of parameters, namely $(\gamma = -0.2, \sigma = 1, \mu = 0)$, $(\gamma = 0, \sigma = 1, \mu = 0)$ and
 234 $(\gamma = +0.2, \sigma = 1, \mu = 0)$.

235 The methods developed in this work have been implemented in R (R Core Team, 2020). The code
236 is available upon request. The following packages are used: *tseries* (Trapletti and Hornik, 2018) and
237 *extRemes* (Gilleland and Katz, 2016). One can see on Figures C1–C4 (Appendix C) that the true values
238 of the shape parameter as well as the true values of extrapolated return levels belong at least to 95% of
239 their respective 95%-confidence intervals constructed from the selected GEV models. This allows us to
240 validate the procedure described in Algorithm 1 on classical continuous probability distributions.

Algorithm 2 Performance assessment of Algorithm 1

Stage 1: Generate several samples, say $x^{(b)} = (x_1^{(b)}, \dots, x_n^{(b)})$ of n independent observations from the true distribution $F(\cdot; \phi)$, where $b = 1, \dots, B$.

Stage 2: Run Algorithm 1 on the sample $x^{(b)}$ and denote the output by

$$G_{z_{i^*(b)}}(\cdot; \hat{\gamma}_{i^*(b)}, \hat{\sigma}_{i^*(b)}, \hat{\mu}_{i^*(b)})$$

which is the selected GEV distribution estimated on the sample of block maxima $z_{i^*(b)}$ associated with the block size $i^*(b)$.

Stage 3: Estimate all return levels $x_{\text{GEV},j}^{(b)}$ associated with the return periods $T^{(j)}$, namely

$$\hat{x}_{\text{GEV},j}^{(b)} = G_{z_{i^*(b)}}^{-1} \left(1 - \frac{i^*(b)}{T^{(j)}}; \hat{\gamma}_{i^*(b)}, \hat{\sigma}_{i^*(b)}, \hat{\mu}_{i^*(b)} \right)$$

and construct its corresponding $100 \times (1 - \alpha)\%$ -confidence interval, namely

$$\left[x_{\text{GEV},j}^{(b,-)}, x_{\text{GEV},j}^{(b,+)} \right]. \quad (20)$$

To construct confidence interval (20), one can use formula (27) provided in Appendix A to approximate the variance of the estimated return level $\hat{x}_{\text{GEV},j}^{(b)}$.

Stage 4: Compute all quantities $x_{\text{GEV},j,\alpha}^{(+)}$ which are the α -quantiles of the following set of return level upper confidence bounds

$$\left\{ x_{\text{GEV},j}^{(b,+)}, b = 1, \dots, B \right\}.$$

Stage 5: Compute the true return levels $x_{\phi,j}$ from the known distribution $F(\cdot; \phi)$ associated with the return period $T^{(j)}$ by

$$x_{\phi,j} = F^{-1} \left(1 - \frac{1}{T^{(j)}}; \phi \right), \quad j = 1, \dots, J. \quad (21)$$

Stage 6: Evaluate the truthfulness of all inequalities $x_{\phi,j} \leq x_{\text{GEV},j,\alpha}^{(+)}$ for $j = 1, \dots, J$. If all these inequalities hold true, it will follow that

$$\mathbb{P} \left\{ x_{\phi,j} \leq x_{\text{GEV},j,\alpha}^{(+)} \right\} = 1 - \alpha$$

as $100 \times \alpha\%$ of optimal GEV models were discarded. Such a conclusion is exactly the expected guarantee to validate Algorithm 1.

241 4. Applications to real datasets

242 To illustrate the developed method, we consider two types of real datasets. The first dataset contains
243 daily observations over several years from some meteorological variables while the second dataset contains
244 observations at millisecond scale over several minutes from sensors in the field of vehicle engineering.
245 We would like to determine GEV models for extrapolation corresponding to some variables from these
246 datasets. Recall that each of these selected GEV models has the property to associate high values of the
247 underlining variable with the corresponding smallest return periods. This makes it different from the
248 GEV model fitted on a sample of block maxima from arbitrary large block size.

249 4.1. Applications to the assessment of extreme meteorological events

250 In this section, we consider the daily weather data from Fort Collins, Colorado, U.S.A. from January 1,
251 1900 to December 31, 1999. This dataset can be downloaded from [Dkengne Sielenou \(2020\)](#) and it is also
252 available in [Gilleland and Katz \(2016\)](#). In this dataset we consider the following three variables. The
253 first one (MxT) is the daily maximum temperature (degrees Fahrenheit). The second one (Snow) is the
254 daily accumulated snow amount and the third one (Prec) is the daily accumulated precipitation (inches).
255 [Katz et al. \(2002\)](#) showed that the annual maxima of this daily precipitation amount is associated with a
256 heavy tailed GEV distribution having $\hat{\gamma} = 0.174$ as estimate of the shape parameter. The basic summary
257 statistics of the three variables MxT, Snow and Prec can be found in Table B1 (Appendix B). The main
258 goal of this section is to estimate GEV models to characterize suitably extreme values of the three above
259 mentioned weather variables.

260 We address this problem by mean of Algorithms 1 separately applied to each sample of raw observations.
261 The results are gathered in Figures C5-C6 (Appendix C). Figure C5 illustrates the last stage of Algorithm 1
262 to select block sizes. Figure C6 shows the adequacy of sample maxima associated with the optimal block
263 sizes to the GEV distribution as well as the estimated return levels along with their 97.5%-upper confidence
264 bounds from the selected GEV models. One can conclude from the results that at the studied location,
265 the unknown parent distribution of daily maximum temperature has a finite right endpoint whereas the
266 unknown parent distributions of daily accumulated snow amount and precipitation are light and heavy
267 tailed, respectively.

268 4.2. Application to the assessment of sensors reliability

269 In this experiment, two sensors are embedded on the same vehicle. The first one is the sensor of
270 interest. Its measurements are considered as observations from a random variable, say Y . The second one is
271 a high-precision sensor which serves as a reference. Its measurements are considered as observations from a
272 random variable, say Z . The sensors provide approximately 36 measures every second. Ignoring all missing
273 values, the dataset includes $n = 113,133$ observations of the random pair (Y, Z) . These observations are
274 associated with 920 objects identified in time by 21,523 distinct timestamps (in millisecond) ranging

275 between 1,347,571 and 3,329,292. Adding up the differences of consecutive timestamps, it follows that the
 276 time period during which both values of Y and Z collected in the set $\{(y_i, z_i), i = 1, \dots, n\}$ are observed
 277 is 1,981,721 milliseconds, namely 33.03 minutes or 0.55 hour. Define the random variable V of errors
 278 associated with the sensor of interest by $V = Z - Y$. Obviously, the sample $v = (v_1, \dots, v_n)$ of size n ,
 279 where $v_i = z_i - y_i$ contains observations from the random variable V .

280 Consider the magnitude or absolute value of V as the random variable $X = |V|$ also defined by
 281 $X = \max\{V, -V\}$. The random variable V is assumed to be continuous so that $\mathbb{P}\{V = 0\} = 0$. Hence, the
 282 positive part of V is the random variable X_+ defined by $X_+ = \max\{V, 0\}$ whereas the negative part of V
 283 is the random variable X_- defined by $X_- = \max\{-V, 0\}$. Let $\mathcal{X} = (x_1, \dots, x_n)$, $\mathcal{X}_+ = (x_{+,1}, \dots, x_{+,n})$
 284 and $\mathcal{X}_- = (x_{-,1}, \dots, x_{-,n})$ be the samples of $n = 113,133$ observations from the random variables X ,
 285 X_+ and X_- , respectively. Here, the quantities x_i , $x_{+,i}$ and $x_{-,i}$ are defined by $x_i = \max\{v_i, -v_i\}$,
 286 $x_{+,i} = \max\{v_i, 0\}$ and $x_{-,i} = \max\{-v_i, 0\}$ for $i = 1, \dots, n$. It is worth noticing that there are 81,891
 287 nonzero values in the sample \mathcal{X}_+ and 31,242 nonzero values in the sample \mathcal{X}_- . The basic summary
 288 statistics of the four variables V , X , X_+ and X_- are provided in Table B1 (Appendix B). The main goal
 289 of this section is to estimate the selected GEV models to characterize extreme values of the three types of
 290 the above mentioned errors.

291 The magnitudes of errors as well as the negative and positive parts of absolute errors can impact
 292 differently the system under consideration when large critical values are observed. Furthermore, in this
 293 field of vehicle engineering, there is no trivial way to prefer a particular block size to another one. To
 294 overcome this issue, we thus apply Algorithm 1 to the samples \mathcal{X} , \mathcal{X}_+ and \mathcal{X}_- . The results are collected in
 295 Figures C7-C8 (Appendix C). The graphs of Figure C7 illustrate the last stage of Algorithm 1 to identify
 296 optimal block sizes. The graphs of Figure C8 show the adequacy of sample maxima associated with
 297 the selected block sizes to the GEV distribution as well as the estimated return levels along with their
 298 97.5%-upper confidence bounds from the selected GEV models. It follows from the obtained GEV models
 299 that for the studied sensor of interest, the unknown parent distributions of the three types of errors are
 300 heavy tailed.

301 5. Conclusion

302 In the block maxima approach, we showed that, when the goal is to obtain a generalized extreme
 303 value model for extrapolation beyond the sample maximum, the size of blocks can be specified thanks
 304 to an algorithmic procedure. We clearly established and justified the theoretical foundations of this
 305 methodology. We successfully demonstrated the efficiency of the method on several samples from some
 306 classical continuous probability distributions. The proposed scheme has been illustrated on two real
 307 datasets. By definition, the selected GEV model is likely to generate larger values than competing GEV
 308 models. However, large values above an eventual unknown threshold might be unrealistic for the studied
 309 phenomenon. Our next study will focus on the determination of such a threshold which is also termed as

310 the extrapolation limit.

311 **Appendix A Inference for the return levels based on the GEV distribution**

312 Let $p \in (0, 1)$. The quantile z_p of the GEV family is obtained by solving the equation

$$G(z_p) = 1 - p, \quad (22)$$

313 where G is the GEV distribution function. For $\gamma \neq 0$, the solution of equation (22) is

$$z_p = \mu - \frac{\sigma}{\gamma} \{1 - [-\log(1 - p)]^{-\gamma}\} \quad (23)$$

314 and for $\gamma = 0$, the solution of equation (22) is

$$z_p = \mu - \sigma \log \{-\log(1 - p)\}. \quad (24)$$

315 In common terminology, z_p is the return level associated with the return period $T = p^{-1}$. This means
 316 that the level z_p is expected to be exceeded on average once every T blocks of observations. It is easy to
 317 see that the return level z_T is strictly increasing with the return period T . Consequently, one can estimate
 318 the frequency of events associated with values larger than the highest observation of the studied random
 319 variable. Let us denote by $(\hat{\mu}, \hat{\sigma}, \hat{\gamma})$ the maximum likelihood estimate of the discrete GEV distribution
 320 parameters (μ, σ, γ) obtained when fitting a sample of m block maxima z_i , $i = 1, \dots, m$ with a GEV
 321 distribution, where the block size is equal to n . By substituting $(\hat{\mu}, \hat{\sigma}, \hat{\gamma})$ into (23) and (24), the maximum
 322 likelihood estimate of z_p is obtained for $\gamma \neq 0$ as

$$\hat{z}_p = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\gamma}} [1 - y_p^{-\hat{\gamma}}] \quad (25)$$

323 and for $\gamma = 0$, the maximum likelihood estimate of z_p is obtained as

$$\hat{z}_p = \hat{\mu} - \hat{\sigma} \log y_p, \quad (26)$$

324 where $y_p = -\log(1 - p)$. Furthermore, by the delta method,

$$\text{Var}(\hat{z}_p) \approx \nabla z_p^T V \nabla z_p, \quad (27)$$

325 where V is the asymptotic variance-covariance matrix (Coles, 2001) of the joint estimate $(\hat{\mu}, \hat{\sigma}, \hat{\gamma})$ of the
 326 parameter (μ, σ, γ) and

$$\begin{aligned} \nabla z_p^T &= \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \gamma} \right] \\ &= [1, -\gamma^{-1} (1 - y_p^{-\gamma}), \sigma \gamma^{-2} (1 - y_p^{-\gamma}) - \sigma \gamma^{-1} y_p^{-\gamma} \log y_p] \end{aligned}$$

evaluated at $(\hat{\mu}, \hat{\sigma}, \hat{\gamma})$. In the particular case where $\gamma = 0$, V stands for the asymptotic variance-covariance
 matrix (Coles, 2001) of the joint estimate $(\hat{\mu}, \hat{\sigma})$ of the parameter (μ, σ) and

$$\nabla z_p^T = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma} \right] = [1, -\log y_p]$$

327 evaluated at $(\hat{\mu}, \hat{\sigma})$.

328 **Appendix B Basic statistics of variables from the real datasets**

	MxT	Snow	Prec	V	X	X ₊	X ₋
Mean	62.403	4.181	1.349	1.013	2.549	0.768	1.781
Std.Dev	18.816	16.677	7.749	4.309	3.618	2.856	2.769
Min	-10.000	0.000	0.000	-50.341	0.000	0.000	0.000
Q1	49.000	0.000	0.000	-0.141	0.744	0.000	0.000
Median	64.000	0.000	0.000	1.154	1.577	0.000	1.154
Q3	78.000	0.000	0.000	2.155	2.741	0.141	2.155
Max	102.000	463.000	211.000	55.061	55.061	50.341	55.061
MAD	22.239	0.000	0.000	1.726	1.373	0.000	1.711
IQR	29.000	0.000	0.000	2.296	1.997	0.141	2.155
CV	0.302	3.988	5.743	4.255	1.420	3.719	1.555
Skewness	-0.369	8.859	9.685	-1.390	4.684	7.591	4.510
SE.Skewness	0.013	0.013	0.013	0.007	0.007	0.007	0.007
Kurtosis	-0.524	123.329	128.941	24.185	33.274	77.697	35.256
N.Valid	36524.000	36524.000	35794.000	113133.000	113133.000	113133.000	113133.000
Pct.Valid	100.000	100.000	98.001	100.000	100.000	100.000	100.000

Table B1: Basic summary statistics of the variables studied in Sections 4.1-4.2. These common descriptive statistics for numerical variables can be organized into 4 main types of measures, namely the measures of frequency (N.Valid: number of valid observations, Pct.Valid: percentage of valid observations), the measures of central tendency (Mean: average value, Median: middle value), the measures of variability (Min: minimum value, Max: maximum value, Q1: first quartile, Q3: third quartile, IQR: inter quartile range, MAD: mean absolute deviation, Std.Dev: standard deviation, CV: coefficient of variation) and the shape measures (Skewness: degree of asymmetry, Kurtosis: degree of tail heaviness).

329 **Appendix C Graphical results from illustrations on simulated and real datasets**

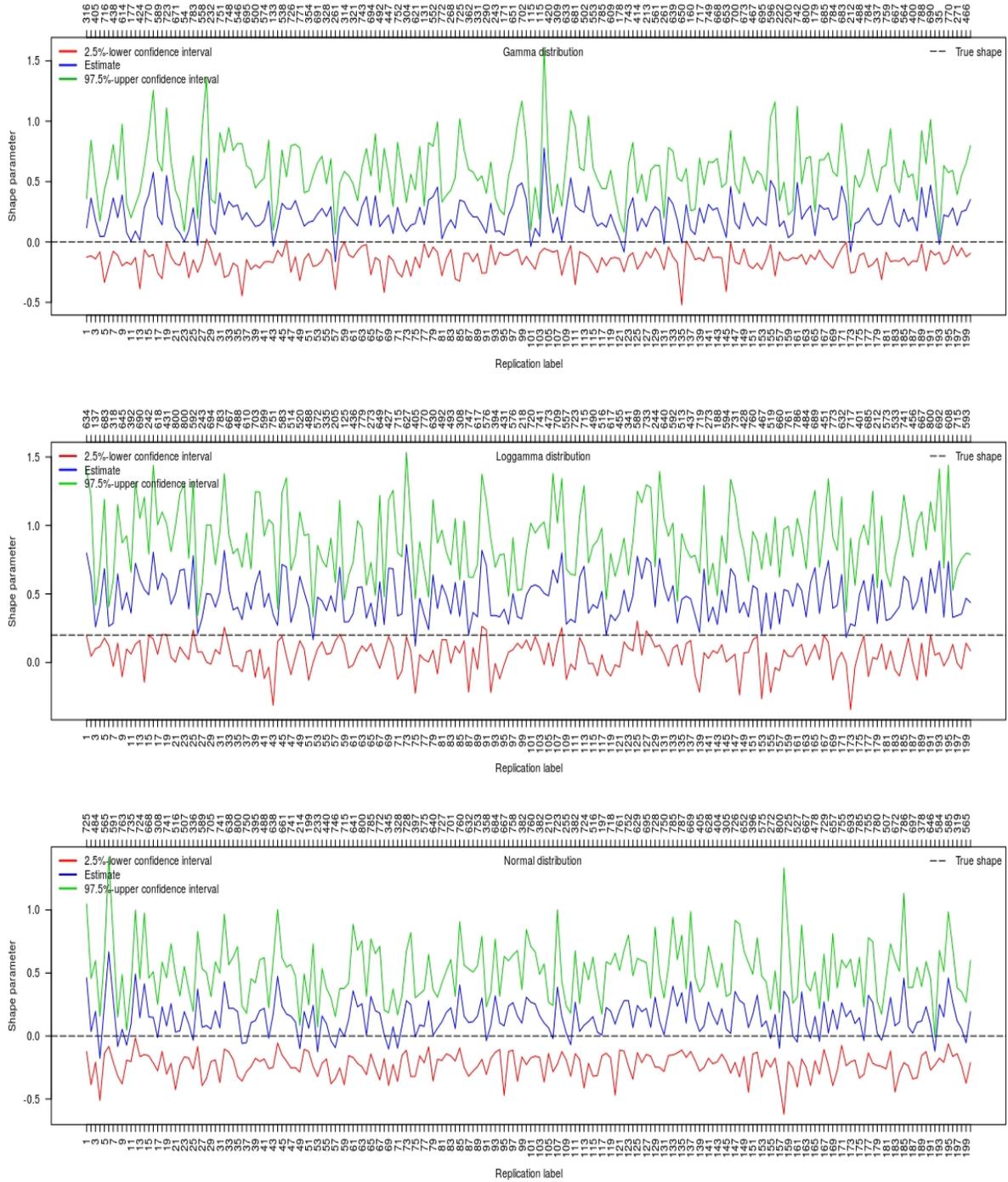


Figure C1: From top to bottom, validation results of Algorithm 2 when using the gamma distribution (15), the loggamma distribution (16) and the normal distribution (17). Each panel displays the 95%-confidence intervals of the estimated shape parameters $\hat{\gamma}_{i^*(b)}$ from the selected GEV models obtained in Stage 2 of Algorithm 2. Selected block sizes $i^*(b)$ are indicated on the top margin.

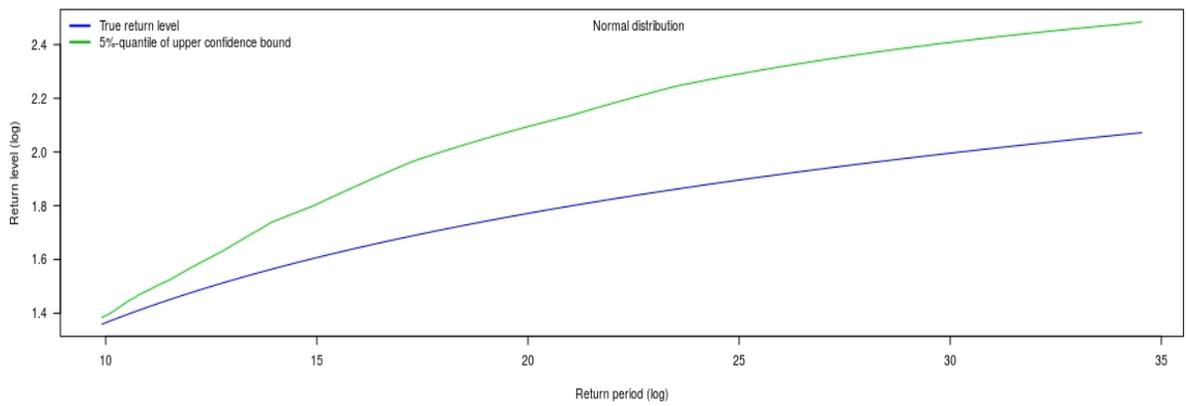
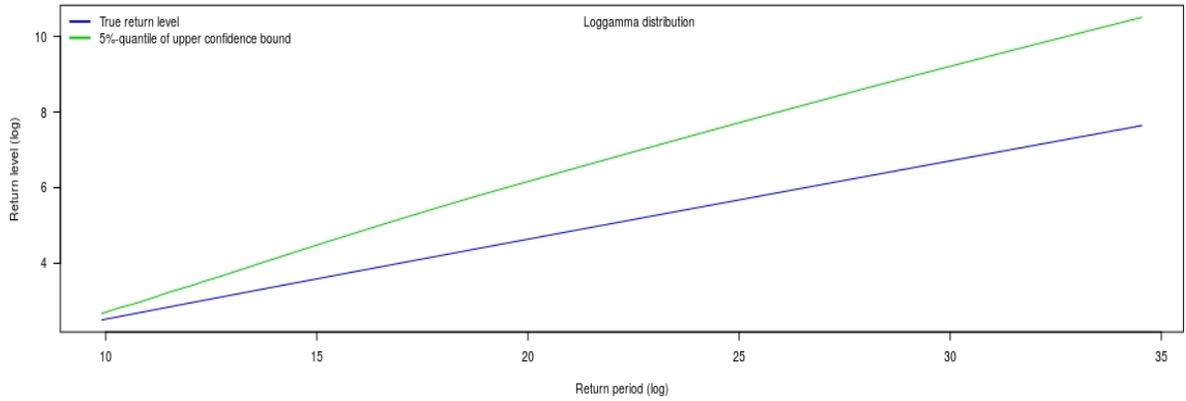
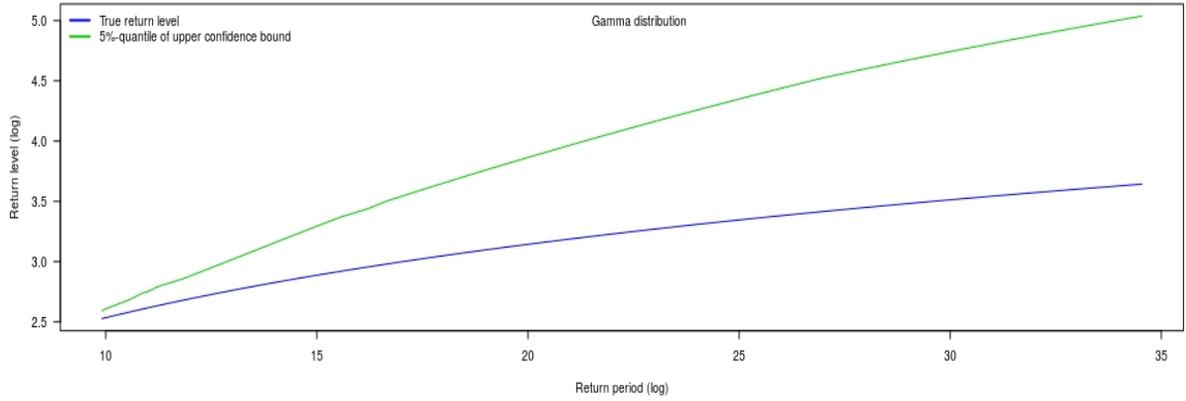


Figure C2: From top to bottom, validation results of Algorithm 2 when using the gamma distribution (15), the loggamma distribution (16) and the normal distribution (17). Graphs display the upper confidence bounds $x_{\text{GEV},j,0.05}^{(+)}$ obtained in Stage 4 of Algorithm 2 as well as the true return levels $x_{\phi,j}$.

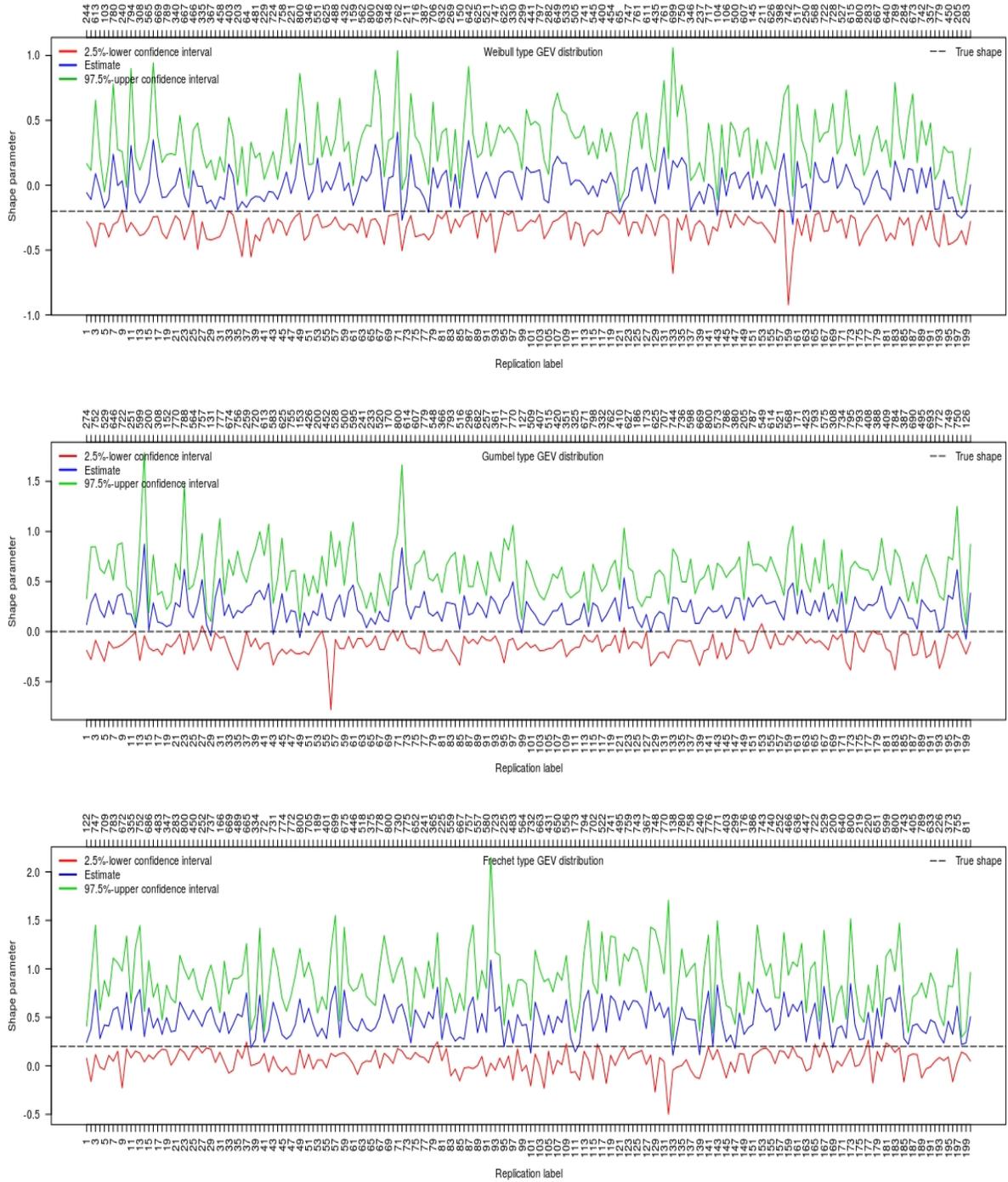


Figure C3: From top to bottom, validation results of Algorithm 2 when using the GEV distributions (18-19) with a negative, zero and positive shape parameter γ . Each panel displays the 95%-confidence intervals of the estimated shape parameters $\hat{\gamma}_{i^*(b)}$ from the selected GEV models obtained in Stage 2 of Algorithm 2. Selected block sizes $i^*(b)$ are indicated on the top margin.

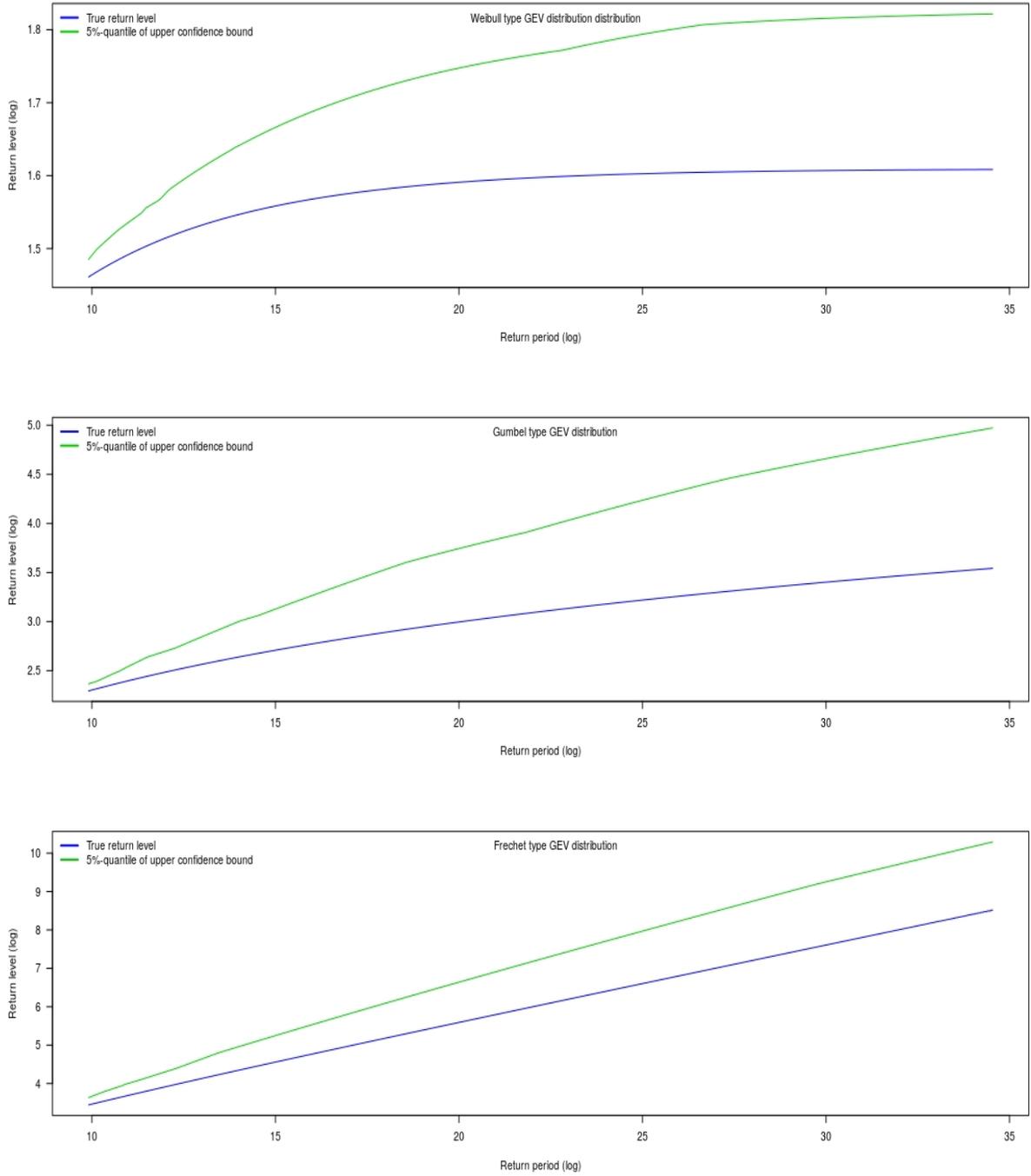


Figure C4: From top to bottom, validation results of Algorithm 2 when using the GEV distributions (18-19) with a negative, zero and positive shape parameter γ . Graphs display the upper confidence bounds $x_{\text{GEV},j,0.05}^{(+)}$ obtained in Stage 4 of Algorithm 2 as well as the true return levels $x_{\phi,j}$.

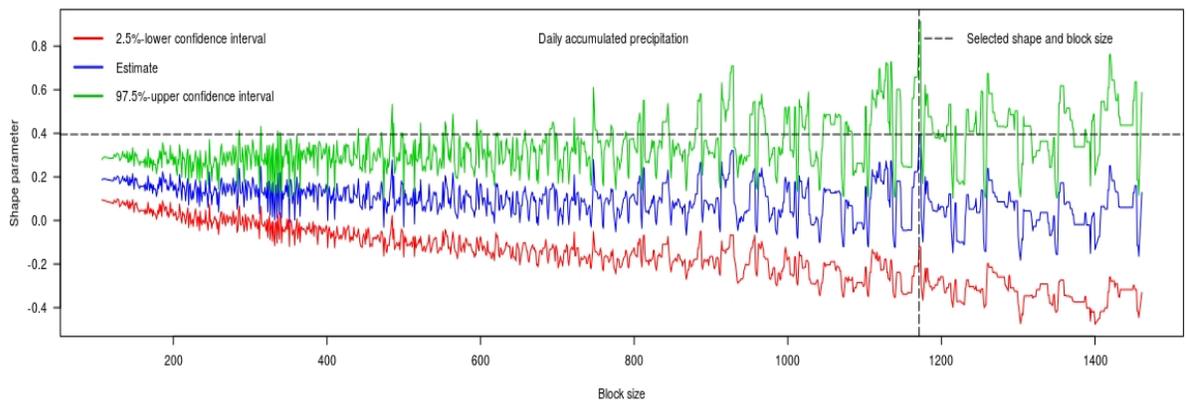
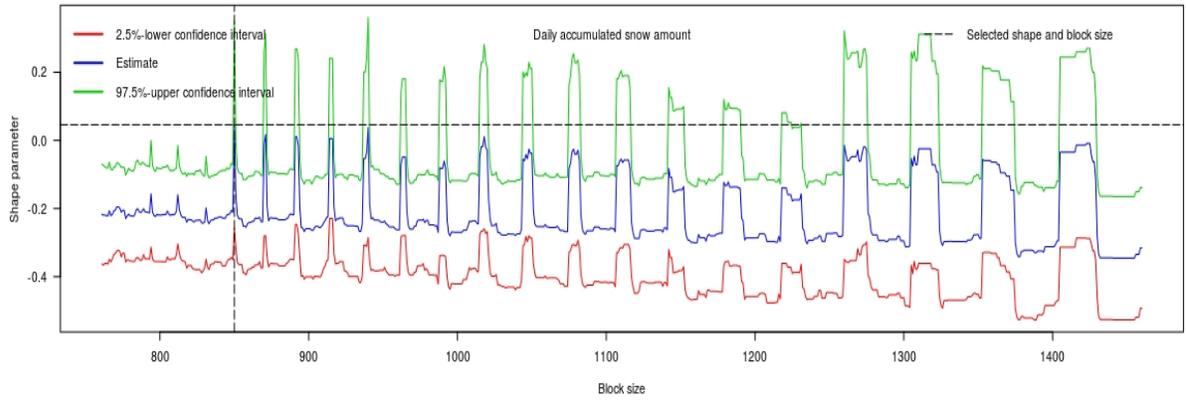
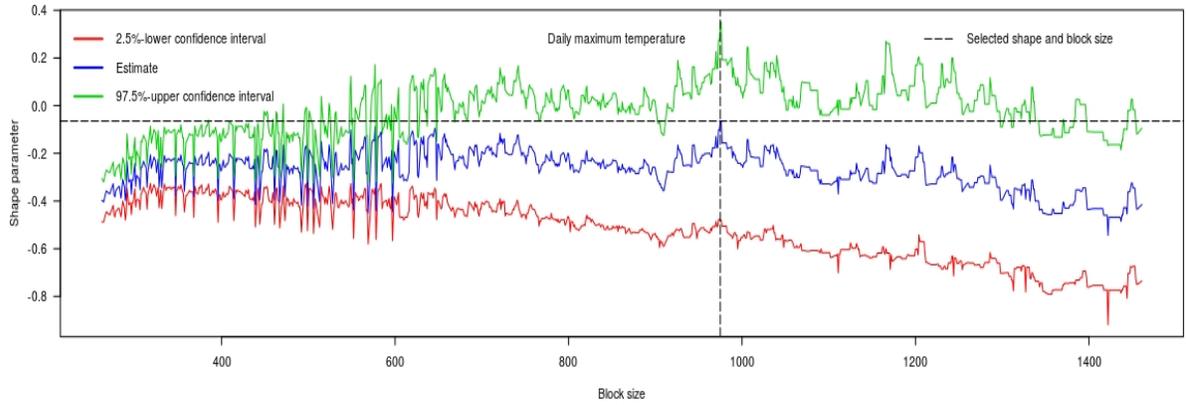


Figure C5: 95%-confidence intervals $C_i(\gamma^*)$ for the shape parameters associated with equivalent GEV models obtained in Stage 6 of Algorithm 1 applied on the real dataset described in Section 4.1. The horizontal dotted line displays the selected GEV distribution shape parameter and the vertical dotted line displays the selected block size i^* .

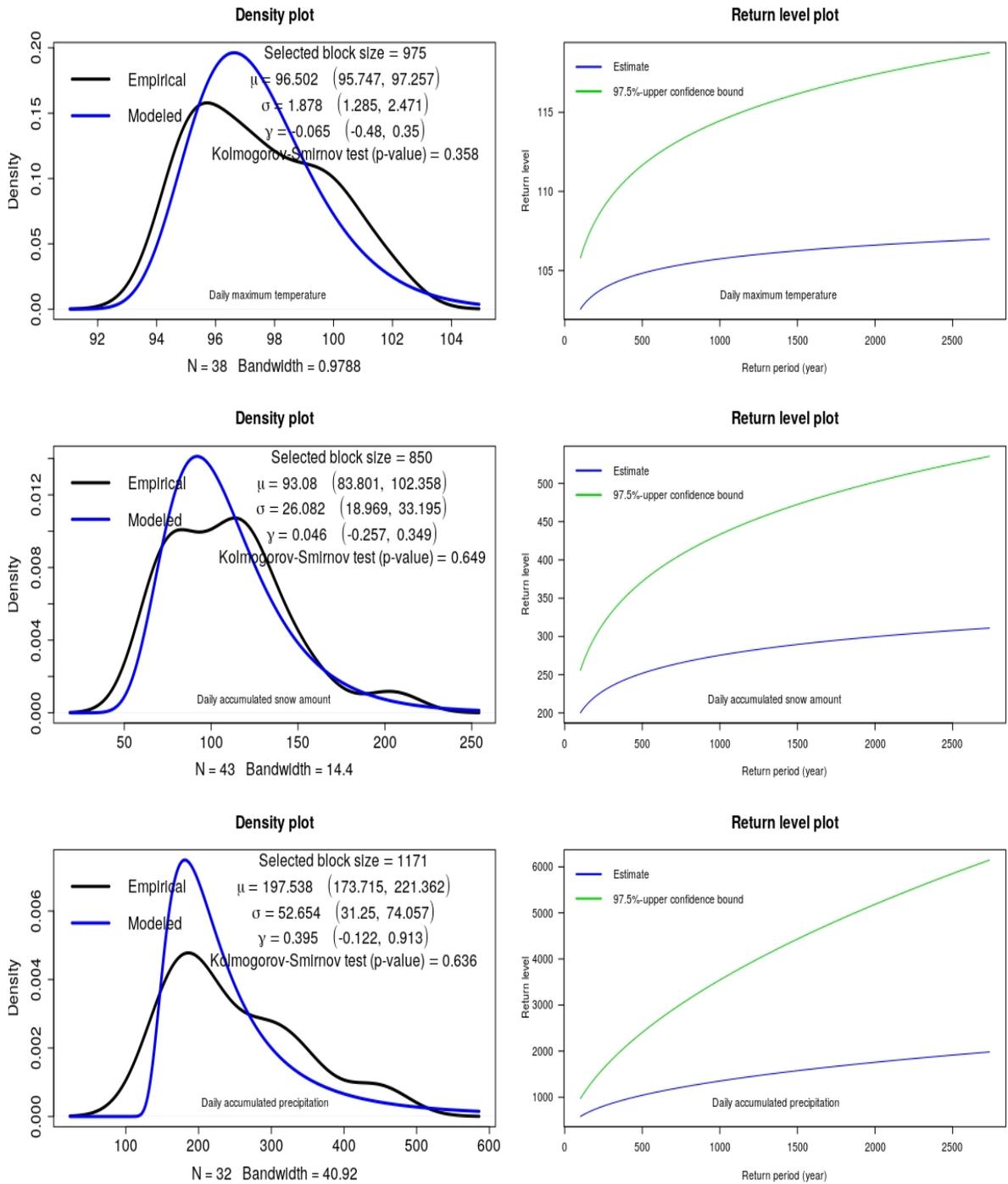


Figure C6: Left panel: graphical display of the goodness of fit of the selected GEV distributions obtained as output of Algorithm 1 applied on the real dataset described in Section 4.1 (blue: fitted GEV density, black: kernel density estimate). Right panel: graphical display of the corresponding estimated return levels (blue) along with their 95%-upper confidence bounds from the estimated selected GEV models (green).

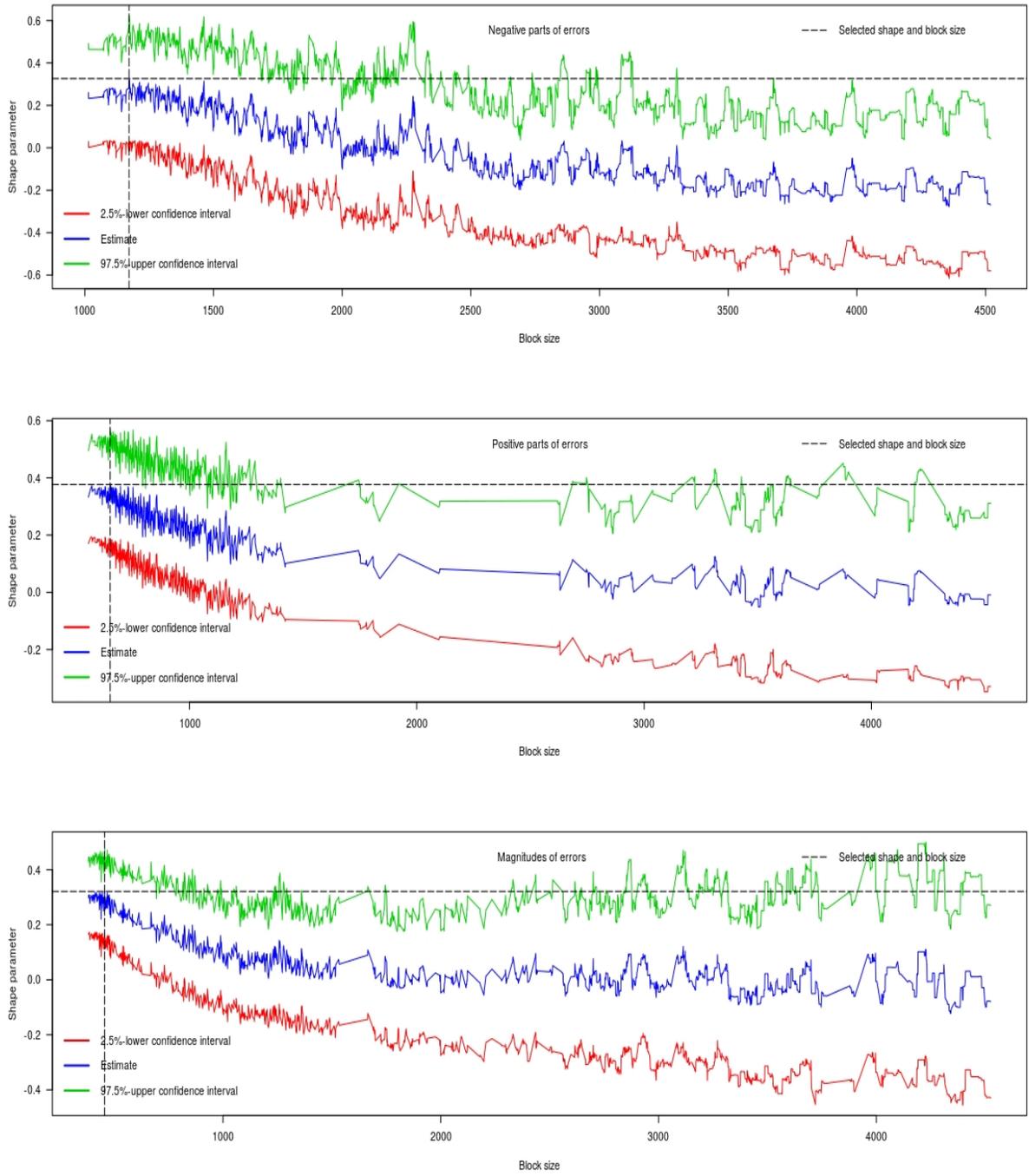


Figure C7: 95%-confidence intervals $C_i(\gamma^*)$ for the shape parameters associated with equivalent GEV models obtained in Stage 6 of Algorithm 1 applied on the real dataset described in Section 4.2. The horizontal dotted line displays the selected GEV distribution shape parameter and the vertical dotted line displays the selected block size i^* .

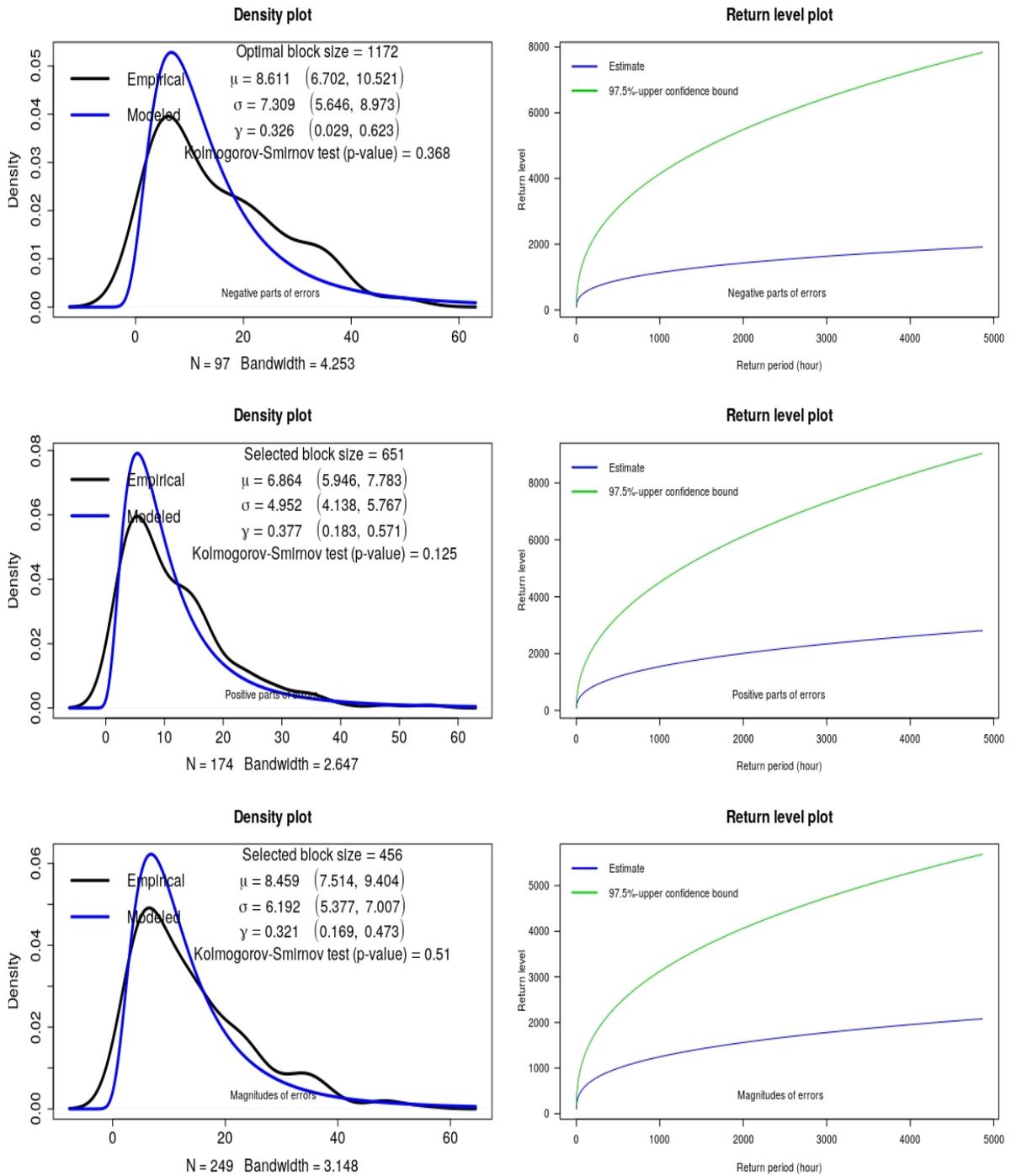


Figure C8: Left panel: graphical display of the goodness of fit of the selected GEV distributions obtained as output of Algorithm 1 applied on the real dataset described in Section 4.2 (blue: fitted GEV density, black: kernel density estimate). Right panel: graphical display of the corresponding estimated return levels (blue) along with their 95%-upper confidence bounds from the estimated selected GEV models (green).

330 Acknowledgements

331 The authors would like to express sincere thanks to Julien Martiniak and Keilatt Andriantavison for
332 their very constructive comments and suggestions to improve the quality of the paper.

333 References

- 334 Banerjee, A., Dolado, J., Galbraith, J., Hendry, D., Press, O.U., 1993. Co-integration, Error Correction,
335 and the Econometric Analysis of Non-stationary Data. Advanced texts in econometrics, Oxford
336 University Press.
- 337 Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., 2004. Statistics of Extremes: Theory and Applications.
338 Wiley Series in Probability and Statistics.
- 339 Chakravarti, I., Laha, R., Roy, J., 1967. Handbook of Methods of Applied Statistics: Techniques of
340 computation, descriptive methods, and statistical inference. Handbook of Methods of Applied Statistics,
341 Wiley.
- 342 Coles, S.G., 2001. An introduction to statistical modeling of extreme values. Springer Series in Statistics.
- 343 Dkengne Sielenou, P.A., 2020. Fcwx.csv. Mendeley Data URL: [http://dx.doi.org/10.17632/
344 32w7fw9ysm.1](http://dx.doi.org/10.17632/32w7fw9ysm.1), doi:10.17632/32w7fw9ysm.1.
- 345 Durbin, J., 1973. Distribution Theory for Tests Based on the Sample Distribution Function. CBMS-NSF
346 Regional Conference Series in Applied Mathematics.
- 347 Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modelling Extremal Events for Insurance and Finance.
348 Springer-Verlag, Berlin.
- 349 Esra Ezgi, E., Eren, O., Özari, C., 2018. A proposal method to select the optimal block size: An application
350 to financial markets. International Journal of Research in Technology and Management 4, 5 pages.
- 351 Fisher, R.A., Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest
352 member of a sample. Mathematical Proceedings of the Cambridge Philosophical Society 24, 180–190.
- 353 Gilleland, E., Katz, R.W., 2016. extRemes 2.0: An extreme value analysis package in R. Journal of
354 Statistical Software 72, 1–39. doi:10.18637/jss.v072.i08.
- 355 Gnedenko, B., 1943. Sur la distribution limite du terme maximum d’une série aléatoire. Annals of
356 Mathematics 44, 423–453.
- 357 Hosking, J.R.M., 1985. Algorithm as 215: Maximum-likelihood estimation of the parameters of the
358 generalized extreme-value distribution. Journal of the Royal Statistical Society. Series C (Applied
359 Statistics) 34, 301–310.

- 360 Katz, R.W., Parlange, M.B., Naveau, P., 2002. Statistics of extremes in hydrology. *Advances in Water*
361 *Resources* 25, 1287–1304.
- 362 Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity
363 against the alternative of a unit root: How sure are we that economic time series have a unit root?
364 *Journal of Econometrics* 54, 159 – 178.
- 365 Leadbetter, M.R., Lindgren, G., Rootzén, H., 1983. *Extremes and related properties of random sequences*
366 *and processes*. Springer-Verlag, New York Inc.
- 367 Özari, C., Eren, O., Saygin, H., 2019. A new methodology for the block maxima approach in selecting the
368 optimal block size. *Tehnički vjesnik* 26, 1292–1296.
- 369 R Core Team, 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for
370 *Statistical Computing*. Vienna, Austria. URL: <https://www.R-project.org/>.
- 371 Said, S.E., Dickey, D.A., 1984. Testing for unit roots in autoregressive-moving average models of unknown
372 order. *Biometrika* 71, 599–607.
- 373 Scarrott, C., MacDonald, A., 2012. A review of extreme value threshold estimation and uncertainty
374 quantification. *REVSTAT - Statistical Journal* 10, 33–60.
- 375 Smith, R.L., 1985. Maximum likelihood estimation in a class of nonregular cases. *Biometrika* 72, 67–90.
- 376 Tancredi, A., Anderson, C., O’Hagan, A., 2006. Accounting for threshold uncertainty in extreme value
377 estimation. *Extremes* 9, 87–106.
- 378 Trapletti, A., Hornik, K., 2018. *tseries: Time Series Analysis and Computational Finance*. URL:
379 <https://CRAN.R-project.org/package=tseries>. R package version 0.10-46.
- 380 Wang, J., You, S., Wu, Y., Zhang, Y., Bin, S., 2016. A method of selecting the block size of BMM for
381 estimating extreme loads in engineering vehicles. *Mathematical Problems in Engineering* 2016, 1–9.
- 382 Wu, G., Qiu, G., 2018. Threshold selection for POT framework in the extreme vehicle loads analysis
383 based on multiple criteria. *Shock and Vibration* 2018, 9 pages.
- 384 Yang, X., Zhang, J., Ren, W.X., 2018. Threshold selection for extreme value estimation of vehicle load
385 effect on bridges. *International Journal of Distributed Sensor Networks* 14, 12 pages.