Nat. Hazards Earth Syst. Sci., 17, 515–531, 2017
www.nat-hazards-earth-syst-sci.net/17/515/2017/
doi:10.5194/nhess-17-515-2017

Natural Hazards and Earth System Sciences Open Access EGU

Extreme weather exposure identification for road networks – a comparative assessment of statistical methods

Matthias Schlögl[1] **and Gregor Laaha**[2]

[1]Transportation Infrastructure Technologies, Center for Mobility Systems, Austrian Institute of Technology (AIT), Vienna, 1210, Austria
[2]Institute of Applied Statistics and Scientific Computing, Department of Landscape, Spatial and Infrastructure Sciences, University of Natural Resources and Life Sciences (BOKU), Vienna, 1190, Austria

Correspondence to: Matthias Schlögl (matthias.schloegl@ait.ac.at)

Received: 18 November 2016 – Discussion started: 5 December 2016
Revised: 7 March 2017 – Accepted: 7 March 2017 – Published: 3 April 2017

Abstract. The assessment of road infrastructure exposure to extreme weather events is of major importance for scientists and practitioners alike. In this study, we compare the different extreme value approaches and fitting methods with respect to their value for assessing the exposure of transport networks to extreme precipitation and temperature impacts. Based on an Austrian data set from 25 meteorological stations representing diverse meteorological conditions, we assess the added value of partial duration series (PDS) over the standardly used annual maxima series (AMS) in order to give recommendations for performing extreme value statistics of meteorological hazards. Results show the merits of the robust L-moment estimation, which yielded better results than maximum likelihood estimation in 62 % of all cases. At the same time, results question the general assumption of the threshold excess approach (employing PDS) being superior to the block maxima approach (employing AMS) due to information gain. For low return periods (non-extreme events) the PDS approach tends to overestimate return levels as compared to the AMS approach, whereas an opposite behavior was found for high return levels (extreme events). In extreme cases, an inappropriate threshold was shown to lead to considerable biases that may outperform the possible gain of information from including additional extreme events by far. This effect was visible from neither the square-root criterion nor standardly used graphical diagnosis (mean residual life plot) but rather from a direct comparison of AMS and PDS in combined quantile plots. We therefore recommend performing AMS and PDS approaches simultaneously in order to select the best-suited approach. This will make the analyses more robust, not only in cases where threshold selection and dependency introduces biases to the PDS approach but also in cases where the AMS contains non-extreme events that may introduce similar biases. For assessing the performance of extreme events we recommend the use of conditional performance measures that focus on rare events only in addition to standardly used unconditional indicators. The findings of the study directly address road and traffic management but can be transferred to a range of other environmental variables including meteorological and hydrological quantities.

1 Introduction

Reliable information about the exposure of road infrastructure networks to extreme weather events is of major concern for road authorities, governmental institutions and safety researchers all over the world (TRB, 2008; Koetse and Rietveld, 2009; Eisenack et al., 2011; Doll et al., 2013; UNECE, 2013; Meyer et al., 2014; Michaelides, 2014; Schweikert et al., 2014a, b; Matulla et al., 2017). In a changing climate (IPCC, 2012) and due to extensive soil sealing (Nestroy, 2006) the impacts of extreme weather events are likely to increase in both frequency and intensity (APCC, 2014). Against this background, the resilience of transport systems with respect to weather hazards has become increasingly important.

A basic requirement for foresightful road infrastructure management is data about both the probability and magnitude of severe weather events. This information can be derived from long-term records of weather quantities such as precipitation and temperature by means of statistical extreme value modeling. While extreme value theory provides a methodological framework that is commonly used in various scientific disciplines, such as hydrology (Katz et al., 2002), finance (Embrechts et al., 2003), engineering (Castillo et al., 2005) and climate sciences (Katz, 2010; Cheng et al., 2014), the application of these tools for road network exposure analysis is a relatively uncharted area. In particular, formal comparative assessments of the various statistical methods that can be applied for estimating return levels of extreme events are rare.

Two basic approaches have been proposed for deriving extreme value series (Coles, 2001), which are both widely applied in studying extreme meteorological events (e.g., Smith, 1989; Davison and Smith, 1990; Parey et al., 2010; Villarini, 2011; Papalexiou and Koutsoyiannis, 2013). On the one hand, the maximum value per year can be used in the block maxima approach, resulting in an annual maxima series (AMS). On the other hand, all values exceeding a certain threshold can be considered extreme, leading to the threshold excess approach based on partial duration series (PDS). Once the extreme value series has been derived, an appropriate distribution function is fitted to these observations by using different parameter estimation methods, such as maximum likelihood estimation, method of moments or Bayesian methods for parameter estimation. Clearly, there are a number of possible combinations of the approaches that may lead to different, often equally plausible results.

Several efforts have been made to compare the performance of block maxima and threshold excess approaches. While some studies only provide a qualitative description of resulting parameter estimates and estimated return levels for both methods (Jarušková and Hanek, 2006), more formal assessment approaches are based on the asymptotic variance of the T-year event estimator (Cunnane, 1973) or on various goodness-of-fit tests and model performance metrics (Madsen et al., 1997a, b; Bezak et al., 2014). Controversial conclusions have been drawn. For instance, Madsen et al. (1997a) found for extreme discharges that the most suitable approach depends on the sample size and the shape parameter of the fitted functions. However, Ben-Zvi (2009) and Bezak et al. (2014) argue that a generalized Pareto (GP) distribution fitted to partial duration series yields the best results for modeling rainfall and discharge extremes. Mkhandi et al. (2005), again, found that AMS and PDS methods result in similar predictions of flood magnitudes. All of these studies document the importance of extreme value analysis in hydrology, but similar studies on temperature extremes, which are equally important as rainfall impacts for road networks, are rare. Based on a literature review, Grotjahn et al. (2016) argue in favor of using PDS for analyzing extremes in large-scale meteorological patterns, but their review did not contain any direct quantitative comparisons based on a common data set. Moreover, studies so far did not specifically assess the performance of methods with respect to rare events, such as 100-year events, which are more relevant for risk assessment than events at the moderate tail of the distribution.

In this study, we compare the different extreme value approaches and fitting methods with respect to their value for assessing the exposure of transport networks to extreme weather impacts. Based on an Austrian data set from 25 meteorological stations representing diverse meteorological conditions, we assess the added value of partial duration series over the standardly used annual maxima series in order to give recommendations for performing extreme value statistics of meteorological hazards.

2 Materials and methods

2.1 Data – meteorological indicators

This study focuses on several meteorological indicators that can be used to assess the exposure of road networks to two main meteorological quantities: precipitation and temperature. These two variables are considered to have the most serious influence on damage to infrastructure (Matulla et al., 2017). They are measured by meteorological services on a regular basis so the data quality is usually high. Nevertheless, the methodology presented in this paper is applicable to various other meteorological quantities (e.g., maximum wind speed) when time series of about 30 years or more are available.

Four meteorological indices are used in this study. Temperature impacts are considered by daily minimum ($T_{\min}$) and daily maximum temperature ($T_{\max}$). In addition, maximum daily temperature difference ($T_{\Delta} = T_{\max} - T_{\min}$) is analyzed, with all temperature indices in °C. Regarding precipitation impacts, the daily precipitation sum mm day^{-1} has been chosen.

In order to identify suitable meteorological stations that represent the main climate features of the highway network in Austria, all monitoring stations operated by the national weather service Zentralanstalt für Meteorologie und Geodynamik (ZAMG) served as a starting point. The selection of suitable stations was carried out in a stepwise procedure with respect to the following considerations. Firstly, the spatial proximity of available measuring stations to the highway network was considered by excluding stations with a distance greater than 10 km from the data set. Secondly, data availability and data quality were considered. As sufficiently long time series are a prerequisite for reliable return level estimation, only stations with more than 30 years of record (i.e., since 1 January 1985) and with less than 5 % missing values were selected. Finally, topographic conditions and regional peculiarities were taken into account for selecting evenly

spread and climatically representative stations. This step was guided by visual inspection of climate maps (Hiebl et al., 2011) and the digital hydrological atlas of Austria (BML-FUW, 2007; Fürst et al., 2009). The data set so obtained consists of 25 hot spots representing climatically homogeneous regions of Austria (Fig. 1).

2.2 Extreme value selection

2.2.1 Block maxima method

The first approach for deriving extreme value series consists in selecting maximum (or similarly minimum) values of the observations within subsequent time intervals (blocks) of constant length. While the block size is freely selectable, a trade-off has to be made between bias (small blocks) and variance (large blocks). Most commonly, the length of the block is chosen to correspond to a calendar year (Coles, 2001), resulting in an annual extreme value series. This was also the case in our study.

Based on the Fisher–Tippett–Gnedenko theorem, a generalized extreme value (GEV) distribution is appropriate for modeling the resulting annual maxima series (Fisher and Tippet, 1928; Gnedenko, 1943). The cumulative distribution function of the GEV is defined by

$$G_{\mu,\sigma,\xi}(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \quad (1)$$

for the set $\left\{z : 1+\xi\left(\frac{z-\mu}{\sigma}\right) > 0\right\}$, where μ is the location parameter, σ is the scale parameter and ξ is the shape parameter. Alternative formulations with inverse sign of ξ are also common (e.g., Hosking, 1990). In both cases, the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$ (Coles, 2001).

The GEV comprises three different types of distributions, which can be distinguished by the sign of their shape parameter: Gumbel, Fréchet and Weibull distribution (Fréchet, 1927; Gumbel, 1958; Coles, 2001; Embrechts et al., 2003; Basrak, 2014). The Gumbel distribution is commonly applied for maxima that are not limited towards an upper bound, whereas the Weibull case is more appropriate for minima that are often limited by a lower bound (Tallaksen and van Lanen, 2004).

2.2.2 Threshold excess method

In some cases, fitting distributions to block maxima data is a wasteful approach as only one value per block is used for modeling. A threshold excess approach potentially provides more information on extremes (Coles, 2001).

Analogous to the choice of the block size in the block maxima approach, the selection of the threshold value in the threshold excess method is also subject to a trade-off between bias (due to selecting non-extreme events if the threshold is low) and variance (due to a small number of exceedances when selecting a high threshold). Hence, the choice of a suitable threshold is important. The basic aim is to select the potentially lowest threshold, given the prerequisite that the extreme value model must provide a reasonable approximation to exceedances above this threshold and shall not contain non-extreme events (Coles, 2001). According to the Pickands–Balkema–de Haan theorem, a GP distribution is suited for modeling the resulting threshold excesses (Balkema and de Haan, 1974; Pickands, 1975): It states that, for some large threshold u, the distribution function of $(X-u)$, conditional on $X > u$ can be well approximated by the GP distribution, which is defined by

$$H_{\xi,\sigma}(z) = \begin{cases} 1-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi} & \text{for } \xi \neq 0 \\ 1-\exp\left(-\frac{z-\mu}{\sigma}\right) & \text{for } \xi = 0, \end{cases} \quad (2)$$

where the support is $z \geq \mu$ in the case $\xi \geq 0$, and $\mu \leq z \leq \mu - \sigma/\xi$ when $\xi < 0$. This is valid for $x_1, x_2, \ldots, x_n$ being a sample of n independent and identically distributed realizations of a random variable x following some common distribution function F (Coles, 2001).

A number of approaches have been proposed for selecting an appropriate threshold. Coles (2001) suggests to let the selection be guided by graphical diagnostics about bias (i.e., mean excess; see Ghosh and Resnick, 2010, for a detailed discussion) and stability of the scale and shape parameter. Despite these criteria being well justified from a theoretical point of view, their application involves substantial elements of subjectivity, leading to ambiguous results (Scarrott and MacDonald, 2012; Northrop and Coleman, 2014). To overcome this problem, we employed the deterministic square-root rule $k = \sqrt{n}$ (Ferreira et al., 2003) for pre-selecting the threshold level in an objective way, using the kth upper-order statistic as a threshold, which is related to the total time series length n. Although this rule does not properly account for threshold uncertainty on subsequent inferences (Scarrot and MacDonald, 2012), it satisfies the intermediate sequence of order statistics that formally ensures tail convergence (Leadbetter et al., 1983). The so-obtained threshold was subsequently validated by the graphical criteria of Coles (2001) for bias and parameter stability.

2.3 Dealing with non-stationarity and dependency

Extreme value theory assumes that data are independent and identically distributed (Coles, 2001; Gilleland and Katz, 2011; Katz, 2010, 2013; Cheng et al., 2014). To test for non-stationarity in the expected value we perform separate Mann–Kendall trend tests (Mann, 1945; Kendall, 1976; Gilbert, 1987) at a significance level of $\alpha = 0.05$ (Zhang et al., 2004) for the extreme value series of each meteorological indicator. In case of significant trends, detrending was performed with respect to the last year of the time series (i.e.,

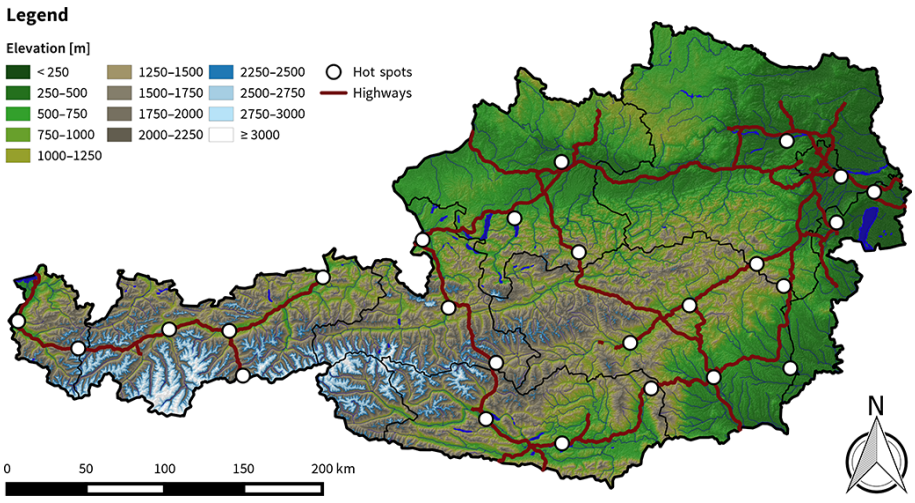

Figure 1. Location of the selected meteorological stations used for extreme value analysis.

2015). The trend-corrected estimation of a meteorological indicator z at time t is obtained as

$$\hat{z}_t = y_t - \hat{y}_t + \hat{y}_{2015}, \tag{3}$$

where y_t is the measurement at time t and $\hat{y}_t$ is the trend at time t obtained from the linear trend model

$$\hat{y}_t = \beta_0 + \beta_1 t, \tag{4}$$

with intercept β_0 and slope β_1, and $\hat{y}_{2015}$ being the trend estimate for 2015.

For climate variables independence of data is usually a minor issue for the annual maxima approach as multi-annual dependencies are usually low for most climates (Madsen et al., 1997a; Katz et al., 2002). Regarding the threshold excess method, threshold exceedances on consecutive days will likely violate the assumption of independence. Dependent values in the threshold excess series are eliminated by a declustering procedure that consists in removing threshold exceedances within the autocorrelation length on both sides of the local maxima (Jarušková and Hanek, 2006). Based on sensitivity analysis an autocorrelation window of 5 days was chosen for the three temperature indicators, while a window of 3 days was chosen for accumulated daily precipitation.

2.4 Parameter estimation

Once the extreme value series is available, a theoretical distribution needs to be fitted. Two different methods of parameter estimation are used within the scope of the present analysis.

The first method, maximum-likelihood estimation (MLE), was formally introduced by Fisher in the early 20th century (Fisher, 1912; Aldrich, 1997; Hald, 1999). Let $x_1, x_2, \ldots, x_n$ be a sample of n independent and identically distributed realizations of a random variable with the unknown probability density function $f(x|\theta_0)$. As the true value of the parameter vector θ_0 is unknown, an estimate $\hat{\theta}$ which is as close to θ_0 as possible is found by maximizing the likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta), \tag{5}$$

i.e., by maximizing the accordance of the extreme value model with the observed data (Coles, 2001).

The second method, L-moment estimation (LMOM), evolved from modifications of probability weighted moments of Greenwood et al. (1979). They are linear combinations of first-order statistics and are hence more robust to measurement errors or sampling uncertainty than conventional moments (Hosking, 1990). The rth population L-moment of a random variable X is defined as

$$\lambda_r \equiv r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \mathrm{EX}_{r-k:r} \quad r = 1, 2, \ldots \tag{6}$$

As compared to MLE, L-moments are superior for fitting GEV distributions in terms of bias and variance, in particular for small sample sizes (Hosking et al., 1985).

As far as reliability of the fitting results is concerned, confidence intervals play a major role for assessing uncertainty. The most common way to derive a $(1-\alpha)$ confidence interval for a particular component θ_i of a parameter vector θ is by using the formula $\hat{\theta}_i \pm z_{\alpha/2} \times \sigma/\sqrt{n}$, with $\hat{\theta}_i$ denoting the estimate for θ_i, $z_{\alpha/2}$ indicating the $\alpha/2$ quantile of the standard normal distribution and $\sigma/\sqrt{n}$ indicating the standard error of the estimate.

The approach assumes Gaussian distributed parameter estimators, which may be inappropriate for extreme value distributions. For LMOM estimators resampling methods have been recommended (Burn, 2003). Thus, nonparametric bootstrapping with 500 iterations was applied in this study. MLE

offers a more accurate method for deriving confidence intervals based on the profile likelihood (Coles, 2001). The profile log-likelihood for θ_i is defined as

$$L_{\mathrm{p}}(\theta_i) = \max_{\delta} L(\theta_i, \delta), \tag{7}$$

where δ denotes all components of parameter vector θ excluding θ_i. That is, for each value of $\theta_i, L_{\mathrm{p}}(\theta_i)$ is the maximized log-likelihood over all remaining elements of θ.

2.5 Assessment method

There are various performance measures that are regularly employed in model evaluation, including the root-mean-square error (RMSE) and the mean absolute error (MAE). These metrics provide a comprehensible and objective basis regarding the assessment of the fitted functions.

However, most events of the extreme value series are only moderate and these will have an overly excessive influence on the performance measure. In order to specifically assess the accuracy of the fitted models for higher quantiles (i.e., for larger return periods), we propose conditional variants of the RMSE (CRMSE_{T^*}) and MAE (CMAE_{T^*}). These metrics are specifically consider the upper tail of the fitted functions above some return period T^*. Using Weibull plotting positions as empirical probability estimator (Weibull, 1939; Makkonen, 2006), these measures are defined as

$$\mathrm{CRMSE}_{T^*} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n_{T^*}}} \forall y_i : \left[-\frac{1}{\ln\left(\frac{m}{N+1}\right)}\right] \geq T^*, \tag{8}$$

$$\mathrm{CMAE}_{T^*} = \frac{\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|}{n_{T^*}} \forall y_i : \left[-\frac{1}{\ln\left(\frac{m}{N+1}\right)}\right] \geq T^*, \tag{9}$$

where $\hat{y}_i$ denotes the model prediction or the ith element of the extreme value series, y_i is its observed value, m is its order statistic (with $m = 1$ for the minimum and $m = N$ for the maximum) and n_{T^*} is the number of elements with an empirical return period greater than T^*. Hence, the conditional performance measures are calculated by using only the residuals of observations and theoretical distribution above some relevant return level T^*. The value for T^* should be chosen depending on the length of the time series available. Since the records at the stations used for this study date back to the period between the world wars in most cases, or even further back to as early as 1895, $T^* = 10$ years has been chosen as the base value of the conditional performance measure, and CRMSE_{10} and CMAE_{10} are calculated accordingly. Similar return periods (about 5–10 years) are often considered as a minimum requirement in storm infrastructure design (e.g., GRCA, 2014; EPA, 2014). Hence, such a level appears well suited to separate extreme and non-extreme events.

Distribution-fitting tests such as Kolmogorov–Smirnov, Anderson–Darling or Cramér–von Mises are not used in this study. Such tests are primarily useful for gaining an appreciation of whether a lack of fit is statistically significant or an effect of sampling uncertainty, but they have little discriminative power to identify the "true" or "best" distribution to use (e.g., Stedinger et al., 1993). Instead, we perform graphical diagnosis of the extreme value series and the fitted distributions in quantile plots, which allow a more complete assessment. Plotting of empirical distributions is straightforward. The return level (i.e., magnitude) z_T of each observed extreme event is plotted against its return period (i.e., recurrence interval) $T = 1/(1-P)$, using Weibull plotting positions as an estimator of empirical recurrence probability P. For AMS, the T-year return level is obtained using the quantile function of the GEV:

$$z_{T,\mathrm{AMS}} = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\ln(P)\}^{-\xi}\right] & \text{for } \xi \neq 0 \\ \mu - \sigma \ln\{-\ln(P)\} & \text{for } \xi = 0, \end{cases} \tag{10}$$

with parameters according to Eq. (1). For PDS, the quantile function provides N-observation return levels rather than T-year events, with N being the number of threshold exceedances. When calculating T-year return levels, the return period T needs to be transformed from an annual scale to an observation scale by taking the ratio of threshold exceedances and years of record into account (Coles, 2001). Hence, the T-year return level is obtained from the quantile function of the GP by

$$z_{T,\mathrm{PDS}} = u + \frac{\sigma}{\xi}\left[(T\lambda)^{\xi} - 1\right], \tag{11}$$

where λ is the mean number of threshold exceedances per year, u is the threshold, and remaining parameters according to Eq. (2). Although Eqs. (10) and (11) yield consistent return levels for both types of extreme value series, the return periods of AMS/GEV and PDS/GP are not fully comparable. As pointed out by Langbein (1949) and Rosbjerg (1977), their relationship can be well approximated by an exponential equation of the form

$$\frac{1}{T_{\mathrm{GEV}}} = 1 - e^{(-1/T_{\mathrm{GP}})}, \tag{12}$$

and the return periods of one approach need to be transformed to obtain consistent plots. Following the convention of the extRemes package (Gilleland and Katz, 2016), the PDS/GP-based T-year event definition is applied in this paper, and we transformed AMS/GEV return periods accordingly. Note, however, that the transformation difference is mostly relevant for small return periods, as differences between T_{GEV} and T_{GP} become negligible for return periods of more than 5 years (Langbein, 1949; Rosbjerg, 1977; WMO, 2009).

3 Results

3.1 Non-stationarity

Extreme value series were checked for stationarity. Most of the temperature hot spots showed a significant trend in at least one of the temperature indicators, but often maxima and minima series were simultaneously affected. All significant trends were incorporated in the model. As illustrated by Fig. 2, the consequence of incorporating a trend model in the analysis is non-stationary return levels that refer to a specific time. We will give results for the end of the observation period.

For precipitation, non-stationarity seems less important than for temperature indicators: about 85 % of the hot spots of our study area showed no trend in the annual extremes. This is consistent with the expectation of the Austrian Panel of Climate Change (APCC, 2014) that climate impacts on precipitation will mainly lead to seasonal shifts rather to changes in total annual precipitation.

3.2 Parameter estimation method

The two approaches have been tested for the four meteorological indicators. In summary, it becomes apparent that the relative performances of MLE and LMOM are strongly situation dependent. For instance, while the return level plots for temperature maxima at Schwechat in the eastern lowlands show that the function fitted on the basis of LMOM behaves more robust, which appears to be beneficial in this case (Fig. 3), return level plots of daily rainfall at Brenner on the Austrian–Italian border indicate that the less robust MLE offers better fit for higher quantiles (Fig. 4).

Table 1 summarizes the overall goodness of fit for the 100 climate records (25 stations × 4 indicators) assessed in this study for the AMS approach. LMOM performed better in 69 % of the cases when assessed by the RMSE and in 94 % when assessed by the MAE (note that for 100 climate records one percent corresponds to one record). Since the MAE favors overall model accuracy and gives little weight to outliers with large errors, the better overall fit achieved by LMOM nicely illustrates the greater robustness of this method. These differences apply to most individual meteorological indicators. The sole exception is daily minimum temperature, which yields similar success rates of MLE and LMOM for both goodness-of-fit measures. This is attributable to several larger residuals in these time series.

The relative performances turned out to be more balanced with respect to the PDS approach. As indicated by Table 2, MLE performed better in 56 and 53 % of the cases when judged by the RMSE and MAE, respectively. Again, daily minimum temperature deviates from the general picture by showing clear advantages in favor of LMOM estimation in this case.

Table 1. Comparison of parameter estimation methods for the AMS approach based on goodness-of-fit measures RMSE and MAE. Numbers indicate success cases of MLE and LMOM.

Indicator	RMSE (MLE)	RMSE (LMOM)	MAE (MLE)	MAE (LMOM)
Precipitation	7	18	4	21
$T_{\min}$	13	12	1	24
$T_{\max}$	5	20	0	25
T_{Δ}	6	19	1	24
Total	31	69	6	94

Table 2. Comparison of parameter estimation methods for the PDS approach based on goodness-of-fit measures RMSE and MAE. Numbers indicate success cases of MLE and LMOM.

Indicator	RMSE (MLE)	RMSE (LMOM)	MAE (MLE)	MAE (LMOM)
Precipitation	14	11	13	12
$T_{\min}$	9	16	12	13
$T_{\max}$	17	8	14	11
T_{Δ}	16	9	14	11
Total	56	44	53	47

Apart from the overall goodness of fit it is interesting to assess how the fit depends on the return period of events. This has been done by visual inspection of the distribution plots, such as the examples shown in Figs. 3 and 4. In most cases there were only minor differences between MLE and LMOM when considering return levels below 10 years, but often considerable differences for larger return periods. For the 100-year events, for example, results of the temperature indicators differed by about 0.5 °C on average and by up to 2 °C for single stations. With maximum differences around 10 mm day^{-1}, the 100-year precipitation events showed even greater variation.

As the objective of extreme value analysis is usually related to return periods of 10 years or more, we specifically assessed the performance of the extreme upper tail of the distribution by the conditional goodness-of-fit measures CRMSE_{10} and CMAE_{10}. Results indicate again a favorable performance of LMOM-method for AMS series (Table 3), when judged by the CRMSE_{10} (58 %) and the CMAE_{10} (62 %).

In contrast, results for the PDS showed, again, a slight advantage of MLE when assessed with the goodness-of-fit measures for the conditional variants. Both measures indicate a preference towards MLE in 58 % of the cases. The better performance of the MLE method is against the expectation based on robustness and will be examined in more detail in the following section.

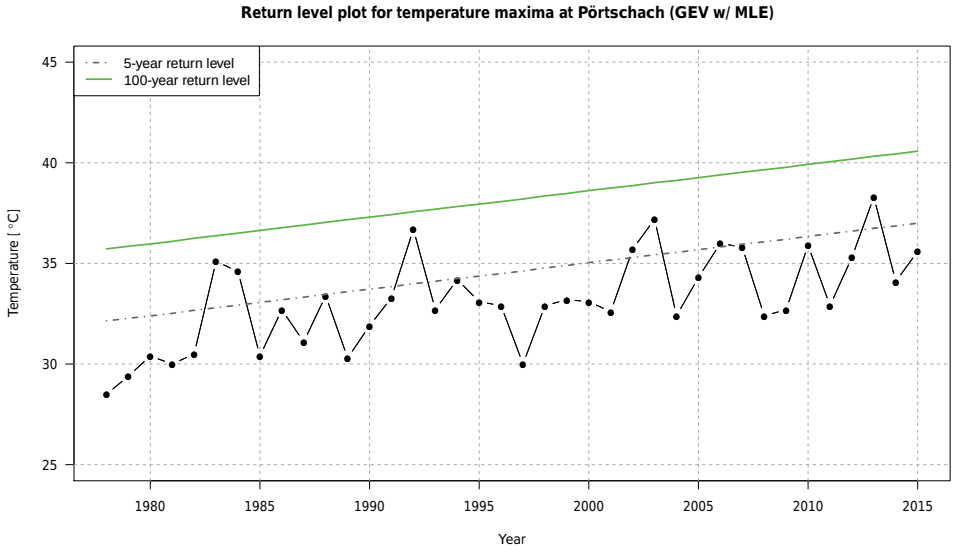

Figure 2. Return level plot of temperature maxima at Pörtschach (Carinthia) with linear trend correction. The trend is visible in the lines depicting the 5-year return level (gray dashed line) and the 100-year return level (green solid line). This is an illustrative example of temperature trends that are commonly observed at the selected stations for both temperature maxima (increasing trend) and temperature minima (decreasing trend).

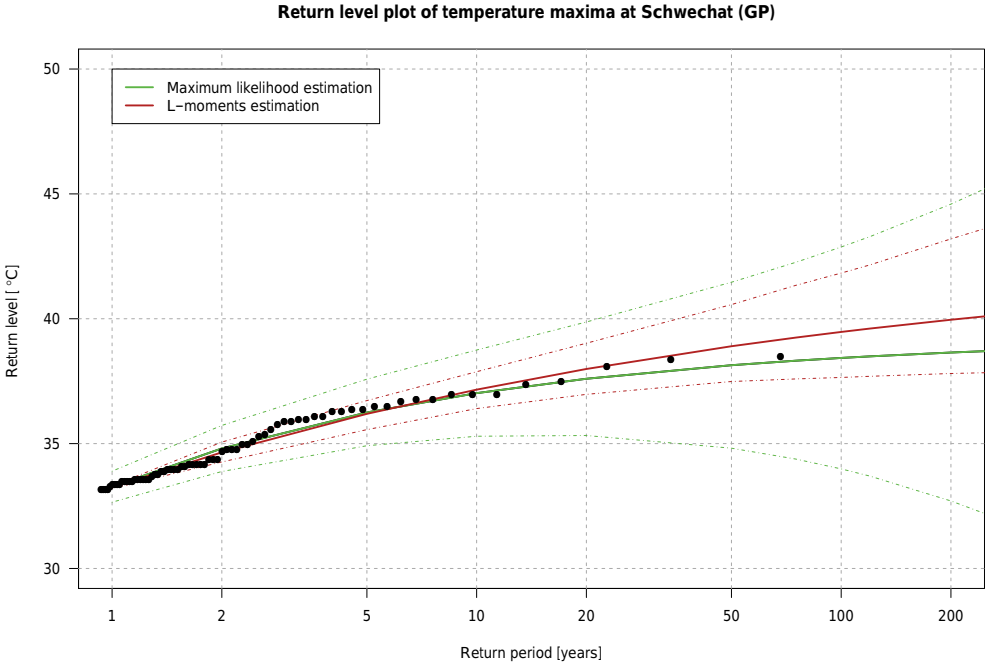

Figure 3. Return level plot of temperature maxima at Schwechat. Return level estimation is based on the threshold excess approach with two different parameter estimation methods (MLE and LMOM estimation). Solid lines show the mean estimate, while dashed lines indicate the 95 % confidence intervals for the fitted functions.

3.3 Extreme value selection

Table 5 presents the relative performances of AMS and PDS approaches based on the two parameter estimation methods. Although overall results show advantages for the AMS approach in terms of goodness of fit for the upper tail of the underlying distributions, results largely depend on the underlying meteorological indicators. While precipitation and daily maximum temperature difference offer a better fit when using GEV distributions of AMS, GP distributions of PDS

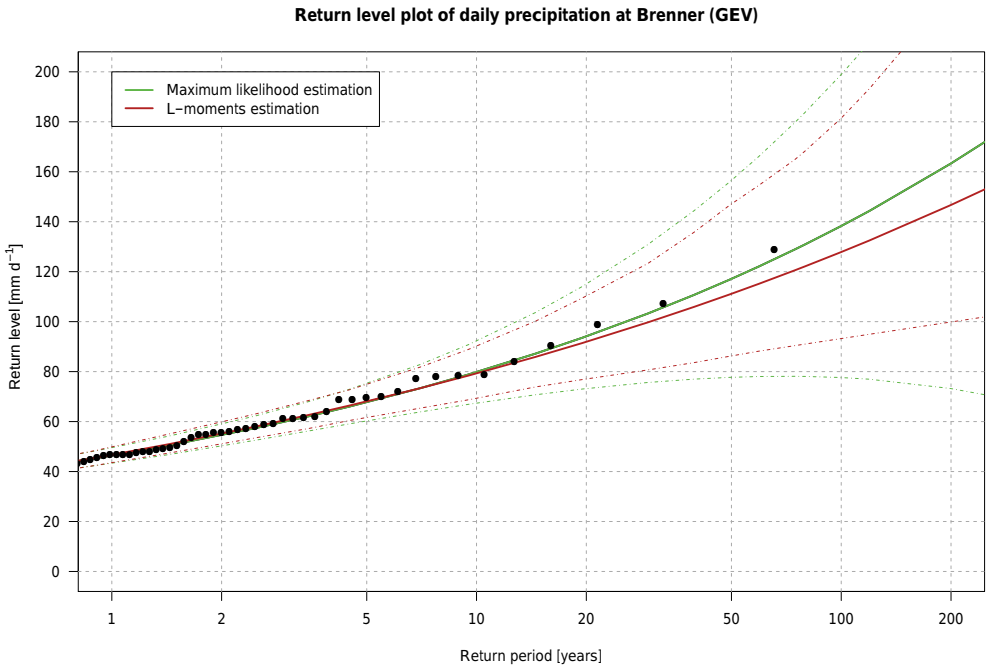

Figure 4. Return level plot of daily rainfall events at Brenner. Return level estimation is based on the block maxima approach with two different parameter estimation methods (MLE and LMOM estimation). Solid lines show the mean estimate, while dashed lines indicate the 95 % confidence intervals for the fitted functions.

Table 3. Comparison of parameter estimation methods for the AMS approach based on conditional goodness-of-fit measures $CRMSE_{10}$ and $CMAE_{10}$. Numbers indicate success cases of MLE and LMOM.

Indicator	$CRMSE_{10}$ (MLE)	$CRMSE_{10}$ (LMOM)	$CMAE_{10}$ (MLE)	$CMAE_{10}$ (LMOM)
Precipitation	11	14	11	14
T_{min}	11	14	10	15
T_{max}	12	13	8	17
T_{Δ}	8	17	9	16
Total	42	58	38	62

Table 4. Comparison of parameter estimation methods for the PDS approach based on conditional goodness-of-fit measures $CRMSE_{10}$ and $CMAE_{10}$. Numbers indicate success cases of MLE and LMOM.

Indicator	$CRMSE_{10}$ (MLE)	$CRMSE_{10}$ (LMOM)	$CMAE_{10}$ (MLE)	$CMAE_{10}$ (LMOM)
Precipitation	14	11	14	11
T_{min}	9	16	10	15
T_{max}	19	6	19	6
T_{Δ}	16	9	15	10
Total	58	42	58	42

Table 5. Comparison of AMS and PDS approach based on conditional goodness-of-fit measures $CRMSE_{10}$ and $CMAE_{10}$ for two parameter estimation methods MLE and LMOM. Numbers indicate success cases of approaches.

Indicator	Fitting method	$CRMSE_{10}$ (GEV)	$CRMSE_{10}$ (GP)	$CMAE_{10}$ (GEV)	$CMAE_{10}$ (GP)
Precipitation	MLE	18	7	19	6
Precipitation	LMOM	19	6	20	5
T_{min}	MLE	9	16	8	17
T_{min}	LMOM	10	15	10	15
T_{max}	MLE	10	15	11	14
T_{max}	LMOM	13	12	14	11
T_{Δ}	MLE	16	9	17	8
T_{Δ}	LMOM	17	8	16	9
Total		112	88	115	85

appear better suited for modeling daily temperature maxima and minima.

To perform a direct comparison, Fig. 5 presents the deviations between return levels derived via AMS and PDS approach for the four meteorological indicators. Interesting patterns regarding the magnitude of the estimated return levels can be observed. For precipitation, PDS/GP estimates result in slightly higher return levels for lower return periods (indicated by negative deviations) and this behavior changes to the opposite for higher return periods. Maximum temperature shows the same tendencies as precipitation, but the PDS/GP always yields higher return levels than the AMS/GEV, sug-

Table 6. Success rates of methods according to $CRMSE_{10}$. The bold value in the center of each field indicates the overall count. The four numbers in the corners display the counts with respect to temperature minima (top left), temperature maxima (top right), temperature difference (bottom left) and precipitation totals (bottom right). Bold values indicate better performance.

		Distribution					
		GEV		GP		Total	
Fitting method	MLE	3	**19**	4	5	**19**	**12**	**38**
		4		8	1		1	
	LMOM	7	**35**	4	**10**	**27**	5	**62**
		12		**12**	8		4	
	Total		**54**			**46**		**100**

gesting that differences mainly occur at higher return periods. Temperature minima, however, show a rather constant overestimation (i.e., underestimation of negative magnitude) of PDS compared to AMS regardless of the frequency of events, with patterns of temperature difference being a combination of the effects of temperature maxima and minima. Overall, the average deviations between methods mostly increase with the return period, and the variability between cases increases as well. This issue will be further explored in the discussion section.

Finally, Table 6 summarizes the success rates of all methods based on $CRMSE_{10}$. Results show an overall advantage of using L-moment estimation as compared to MLE. As far as the two different methods of extreme value selection are concerned, the AMS approach seems to slightly outperform the threshold excess approach in this study. While results are basically quite balanced between all four methods, AMS fitted on the basis of LMOM estimation turned out to yield the best results in about 35 % of all cases.

4 Discussion

We compared the relative merits of the block maxima method and the threshold excess approach. In addition, two different fitting methods have been contrasted. This results in four possible combinations of extreme value model parameter estimation, all of which have certain strengths and weaknesses. Concerning the fit of the distributions to sample, we found a slight advantage of using LMOM instead of MLE, especially in combination with AMS/GEV. For PDS/GP there was a slight advantage of using MLE. However, overall, the differences were not huge.

The conditional assessment of the individual deviation between return levels of AMS/GEV and PDS/GP yielded deeper insight in the relative performances of methods. Most importantly, we found ambiguous systematic deviations between both approaches (Fig. 5), depending on the meteorological indicators under consideration: concerning temperature minima, PDS/GP was found to consistently overestimate return levels compared to the AMS/GEV approach, while results for temperature maxima and – albeit to a lesser extent – temperature differences show just the opposite. Regarding daily rainfall events, the PDS/GP approach tends to slightly overestimate return levels for low return periods (non-extreme events) as compared to the AMS/GEV approach, while an opposite behavior was found for high return levels (extreme events). To assess the reasons for this systematic behavior, we selected four example series that represent extreme cases, where results of approaches differ significantly.

The first two examples are daily precipitation at Sankt Michael (Fig. 6a) and Brenner (Fig. 6b), where extreme value series deviate from the ideal, smooth behavior of a homogeneous extreme value series. These fluctuations point to either measurement errors or process heterogeneity that will introduce uncertainty into extreme value analysis. In the case of Sankt Michael, the most extreme events appear as outliers that deviate from the general behavior of the sample. In general, LMOM will give lower weight to such leverage points but this seems not the case here where the GP fitted by LMOM seems more attracted. A plausible explanation would be that the upper-tail behavior results from the attraction of the distribution at the lower end because of the limited flexibility of the GP. In the case of Brenner, the extreme values seem to follow the same distribution as the remaining sample so one would have more confidence in the validity of these values. However, extreme values are always prone to higher uncertainty than the remaining sample. The MLE estimate gives more weight to these values and shows a better fit at the upper tail in this case, whereas LMOM gives less weight to these values and makes visible that they are not perfectly following the shape of the entire distribution. The choice of the parameter estimation method will finally depend on the weight one tends to give to the extreme values as compared to the remaining sample.

It is also interesting to analyze extreme cases where AMS/GEV and PDS/GP methods yield contrasting results (Fig. 7). When focusing on the empirical distributions, we observe that only the more extreme events (three in the case of Bruck an der Mur and two in the case of Graz) have almost identical empirical probabilities in both extreme value series. At the lower end, we observe that there are several events in the AMS/GEV below the threshold level of PDS, which fit well to the distribution of the higher values so we find no evidence to exclude them from the analysis. The shift in the distribution can therefore be regarded as an effect of threshold level selection, which determines the lower end and therefore the shape of the lower part of the PDS/GP distribution. Between the undisturbed upper part and the disturbed lower part a breakpoint at $T = 15$ years in the PDS/GP is clearly visible from the robustly fitted GP distribution using the LMOM

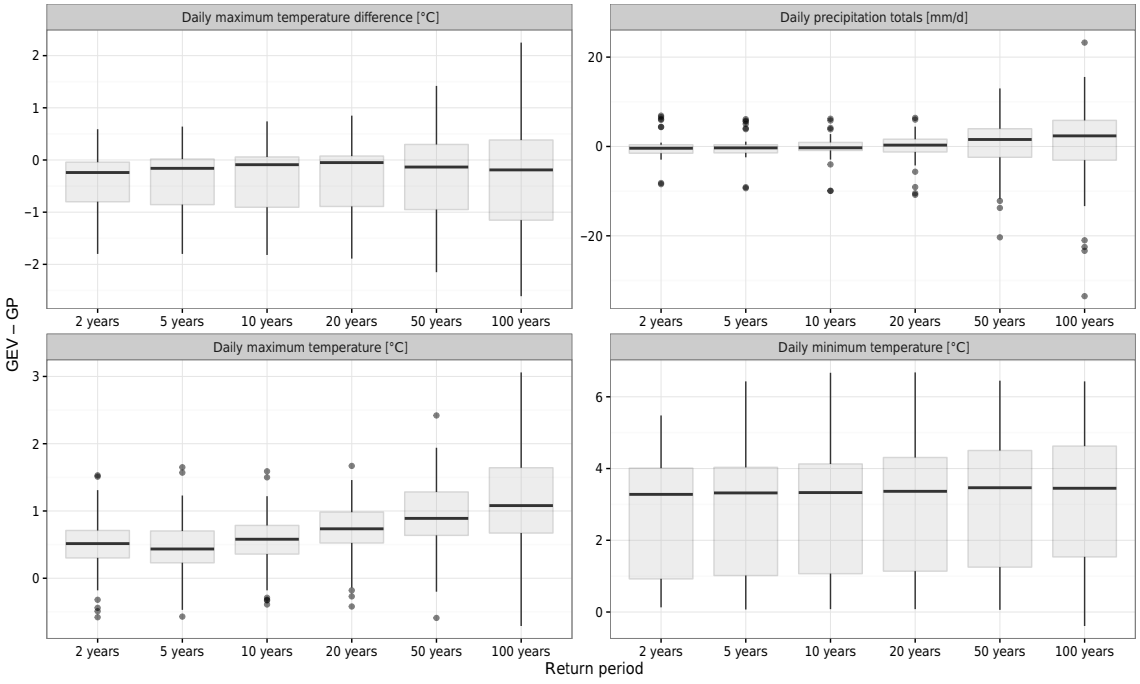

Figure 5. Differences in estimated return levels between GEV and GP models for six selected return periods. These differences are calculated by subtracting the GP estimate from the GEV estimate, given the same parameter estimation method. This results in $n = 50$ observations per box plot.

method. This illustrates an inherent danger of the PDS/GP approach: an inappropriate threshold may entail considerable biases that outperform the possible gain of information by the method by far. This was visible from neither the square-root criterion nor the graphical diagnosis (residual life plot; Fig. 8), which yielded almost no bias in both cases (in the case of Bruck an der Mur, mean excess = 2.99 for the threshold of −17.1 °C; in the case of Graz, mean excess = 1.63 for threshold of 32.6 °C).

Similar shifts may arise if the extreme value series contains dependent events. Non-extreme events are generally more likely to cluster than extreme events because they are generated by exceptional process combinations, which are unlikely to occur more often during one extreme weather situation. Thus, dependencies may possibly affect all parts (but more likely the lower part) of the distribution apart from the maximum, which remains unchanged. In consequence, the empirical distribution is stretched at the lower tail (shifted to the left), with similar consequences on lower and upper tail as described for the case of data uncertainty and leverage points. Such artifacts are difficult to detect in quantile plots of one extreme value series alone but are often visible from direct comparison of AMS/GEV and PDS/GP approaches. Although both AMS/GEV and PDS/GP may be affected by dependency of events, AMS/GEV behaves more robust since it selects only one event per year.

These findings are against our initial expectation and contradict the spirit of most existing studies that aimed to recommend the best-performing method for a variable or situation. Instead of recommending either block maxima or threshold excess method, we recommend performing both approaches, as their combined assessment by means of diagnostic plots together with overall and conditional goodness-of-fit measures offers a more complete diagnosis of the quality of extreme series and the resulting distributions.

Concerning the parameter estimation method, there are also benefits and disadvantages that have to be balanced against each other. MLE has some merit with respect to calculating reliable confidence intervals via profile likelihood. Confidence intervals for estimation via LMOM were derived with non-parametric bootstrapping, which is arguably less trustworthy for indicating the uncertainty of the estimates. However, LMOM estimation has been shown to yield more robust estimation results for small sample sizes (Hosking et al., 1985; Hosking et al., 1987), which can be especially beneficial when analyzing environmental data like temperature or precipitation indicators, which are derived from raw measurements at meteorological measuring stations. Regarding the overall results, LMOM estimation turned out to offer a better fit than MLE, which is consistent with previous findings (Hosking et al., 1985; Hosking et al., 1987; Bezak et al., 2014).

Concerning the comparison based on the goodness of fit of the distributions it shall be noted that a formal comparison of the two extreme value selection approaches is not straightforward. Measures of goodness of fit are not fully conclusive, as

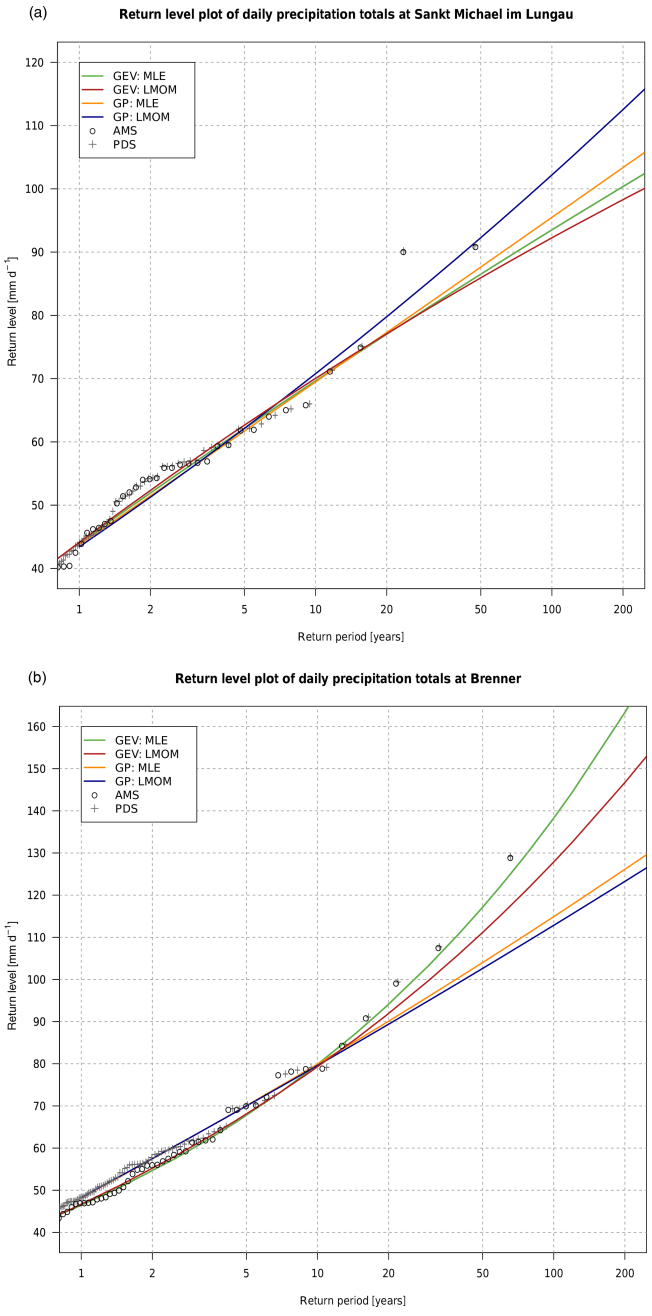

Figure 6. Return level plots of daily rainfall events at the hot spot in **(a)** Sankt Michael im Lungau, which is located in the Central Eastern Alps, and **(b)** Brenner pass, located at the Austro-Italian border. Return level estimation is based on the block maxima approach and on the threshold excess approach with two different parameter estimation methods (MLE and LMOM estimation). Based on the $CRMSE_{10}$, GP fitted on the basis of LMOM estimation was found to be the most appropriate method for Sankt Michael, while GEV with MLE was found to be most suitable at Brenner. Please note that functions are plotted without associated confidence intervals for the sake of clarity.

the underlying extreme value series are derived by different methods and thus are not directly comparable. Our analysis demonstrates that the choice between these approaches has to be based on the statistical properties of the extreme value series, which are related to the indicators under consideration and on data availability. The conditional measures proposed in this paper help to perform a more specific assessment for extreme events, but they are also not a remedy to overcome this problem. They are a way to assess the goodness of fit at the upper tail of the distribution and facilitate the comparison between AMS/GEV and PDS/GP. These metrics can assist, but not substitute, careful analysis of assumptions. We show that contrastive plotting methods can strongly support these analyses.

While the methodology of this study can be easily generalized and extended to cover other environmental variables, four possible limitations have to be discussed. First, the seasonality of temperature and precipitation extremes has not been taken into account. While maximum/minimum temperatures will always occur in the same season, which will factor out any seasonal heterogeneity, this is not genuinely the case for extreme precipitation events, where seasonal occurrence may be associated with diverging processes (Hundecha et al., 2009). In order to account for seasonal effects, a common approach is to split the events into process-homogeneous subsets. This can be based either on seasonality (e.g., Laaha and Blöschl, 2006, for low streamflows) or on a typology of processes (e.g., Merz and Blöschl, 2003, for floods based on rainfall types and catchment preconditions), or a temporal stratification of records is applied (e.g., Méndez et al., 2008, for wave height and Maraun et al., 2009, for heavy precipitation). For each subset extreme value analysis is performed separately, leading to process-specific return levels, such as summer and winter low flows in the case of minimum discharges. These quantities may be combined by a mixed distribution model to yield overall return levels (e.g., Hundecha et al., 2009). For further discussion of modeling dependent and non-stationary time series extremes, the reader is referred to Chavez-Demoulin and Davison (2012).

Secondly, threshold selection in the PDS/GP method is a legitimate subject for debate. In recent years, efforts have been made to overcome the problem of visual threshold selection, e.g., by robust threshold selection (Dupuis, 1999), likelihood-based visual diagnostics (Wadsworth and Tawn, 2012; Wadsworth, 2016), Bayesian approaches (Tancredi et al., 2006; Lee et al., 2014), approaches based on goodness-of-fit tests (Roth et al., 2016) and extreme value mixture models (MacDonald et al., 2011). In addition, attempts were made to develop more automated approaches for extreme value threshold estimation, including the automated threshold selection approach (ATSM) by Thompson et al. (2009), the multiple threshold method (MTM) by Deidda (2010) and the automatic threshold and run parameter selection by Fukutome et al. (2015). While these approaches are appealing from a theoretical perspective, their practical value is of-

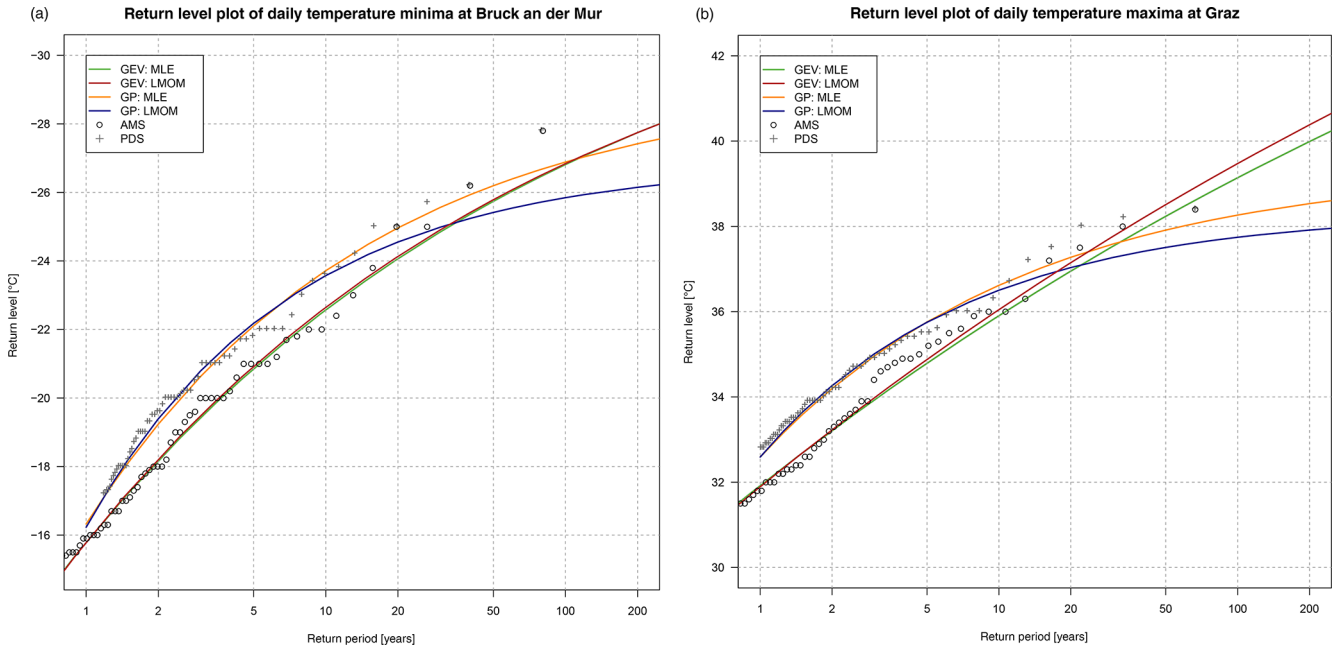

Figure 7. Return level plots of **(a)** temperature minima at Bruck an der Mur and **(b)** temperature maxima at Graz. Return level estimation is based on the block maxima approach and on the threshold excess approach with two different parameter estimation methods (MLE and LMOM estimation). Based on the CRMSE_{10}, GP fitted on the basis of MLE was found to be the most appropriate method for both Bruck an der Mur and Graz.

ten reduced by numerical issues and sampling effects. At least for the time series tested in this study, both the ATSM and MTM yielded inconsistent results: threshold values of similarly distributed time series obtained by ATSM varied considerably, and parameter estimates were depending on range and resolution of the thresholds considered. One could think that automatic threshold selection procedures replace the threshold selection problem with that of selecting an appropriate range and resolution of the thresholds to be tested.

However, the authors are aware that also the semi-supervised method applied in this study may not be optimal in all cases. Rather than performing in-depth analysis of single time series, we have given priority to analyzing a large amount of time series covering a range of environmental conditions. Therefore, the application of the square-root rule in combination with graphical diagnostics is argued to be a feasible approach that led to satisfactory results in the present study.

Thirdly, the conditional performance metrics depend, to some extent, on the chosen plotting position. While the choice of plotting position formula is only of minor importance in many cases, it might be influential in the present case with emphasis on the upper-order statistics. However, a sensitivity analysis based on Beard (i.e., median) plotting positions has shown that effects on results in terms of return level estimates are small in this study, since changes mainly occur in cases where both estimation methods yield very similar parameter estimates.

Finally, it has to be noted that the conditional metrics are of limited robustness, especially if time series are short and the condition is chosen inappropriately. Since the variance of the order statistics strongly increases towards the upper end of the ordered sample, the conditional metrics may be subject to high uncertainty, particularly if inadequate (i.e., too high) return periods are selected. Thus, the authors want to emphasize that an appropriate base value has to be chosen depending on the length of the time series under consideration. For small samples, priority should be given to robust error metrics such as CMAE_{T*}.

5 Conclusion

We compared statistical methods for extreme value analyses based on four climate indicators related to daily precipitation and temperature. While the indicators were selected for studying the exposure of road infrastructure to extreme weather events, the assessments are equally relevant for a range of other environmental variables including meteorological and hydrological quantities. We first analyzed the goodness of fit of distributions to extreme value series consisting of annual maxima (AMS/GEV) and threshold exceedances (PDS/GP) using two parameter estimation methods.

Results for the parameter estimation methods vary considerably between stations and approaches. For the AMS/GEV approach, LMOM yielded, on average, better fitted distribu-

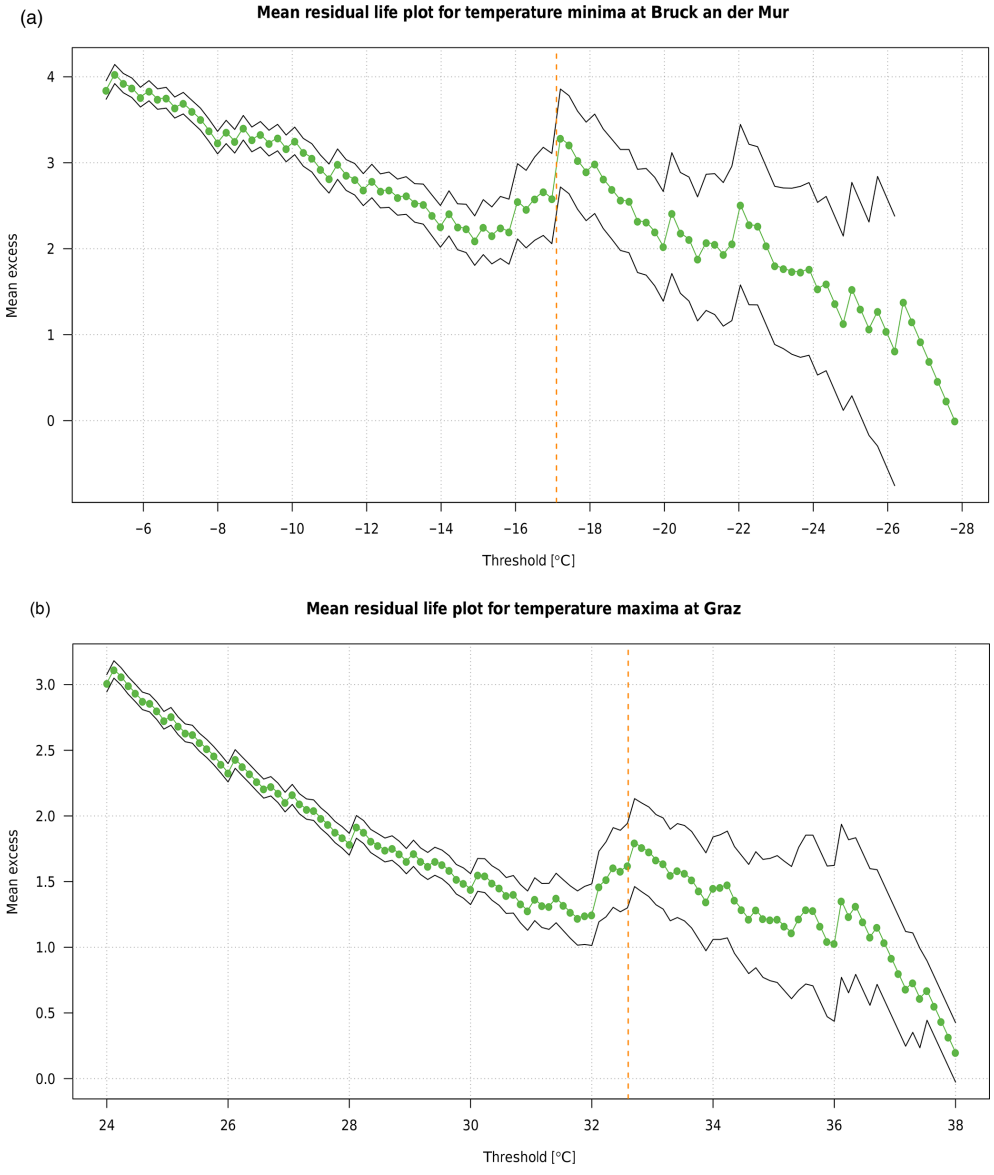

Figure 8. Mean residual life plots of **(a)** temperature minima the hot spot in Bruck an der Mur and **(b)** temperature maxima at Graz. Black lines indicate the 95 % confidence interval for the mean excess and orange lines indicate the threshold selected by means of the square-root rule.

tions than MLE. The goodness of fit turned out to be more balanced with respect to the PDS/GP approach, with a slight advantage of MLE. In most cases there were only minor differences between MLE and LMOM when considering return levels below 10 years, but often considerable differences for larger return periods.

Concerning extreme value selection, the relative performance of AMS/GEV and PDS/GP approaches varies between meteorological indicators. For precipitation and temperature difference the AMS/GEV data outperformed the PDS/GP approach. For temperature maxima and minima the PDS/GP approach appeared better suited.

Regarding goodness of fit for extreme events that are typically used as design values (T of 10 years and more), results show an overall advantage of using L-moment estimation as compared to MLE and that the AMS/GEV approach slightly outperforms the threshold excess approach. The AMS/GEV fitted on the basis of LMOM estimation method performed better than all other combinations of approaches in this study.

We further examined the conditional performances of AMS/GEV and PDS/GP approaches with respect to the return period in more detail. From conditional performance measures and combined plots, we found systematic deviations between AMS/GEV and PDS/GP approaches. For low return periods (non-extreme events) the PDS/GP ap-

proach tends to overestimate return levels as compared to the AMS/GEV approach, whereas an opposite behavior was found for high return levels (extreme events). The assessment of extreme cases where approaches differed significantly suggests that this behavior may be related to two factors: sampling uncertainty and threshold selection.

Regarding sampling uncertainty, we found that outliers may not only attract the distribution at the tail where they occur, but they may also bend the curve at the opposite tail as a consequence of limited flexibility of the extreme value distributions. Such leverage effects can be handled by careful inspection of quantile plots. Regarding threshold selection, the analysis of extreme cases within the data set revealed that an inappropriate threshold may lead to considerable biases that may outperform the possible gain of information from including additional extreme events by far. Selecting a high threshold will determine the lower end of the extreme value distribution whereas the upper tail remains unchanged. This may introduce an inflection point in the distribution, which is against its ideal shape according to extreme value theory, resulting in poor estimates of the theoretical distribution. This effect was not visible from either the square-root criterion or the graphical diagnosis (mean residual life plot), which yielded no atypical biases for the analyzed cases. Similar effects may arise when the extreme value series contains dependent events that may stretch the empirical distribution at the part where they occur. These findings were against our expectations that the estimation of the theoretical distribution will greatly profit from the gain of information that is provided by the PDS/GP approach.

We emphasize that reliable extreme value statistics require controlling for sample effects in order to avoid biased models. In our study, the differences and relative merits of methods were best visible from a direct comparison of AMS/GEV and PDS/GP approaches. We therefore recommend performing both analyses and carefully analyzing the distribution fit relative to the respective sample and relative to each other by means of combined quantile plots. This will make the analyses more robust in cases where threshold selection and dependency introduce biases to the PDS/GP approach as well as in cases where the AMS/GEV contains non-extreme events that may introduce similar biases. For assessing the performance of extreme events we recommend conditional performance measures such as $CRMSE_{10}$ and $CMAE_{10}$ in addition to unconditional indicators.

Data availability. All data used in this study have been obtained from WMO measurement stations in Austria. These stations are operated by the national meteorological and geophysical service of Austria, the Central Institution for Meteorology and Geodynamics (Zentralanstalt für Meteorologie und Geodynamik, ZAMG). Data for the following measurement stations have been kindly provided by ZAMG: Brenner, Bischofshofen, Bruckneudorf, Bruck an der Mur, Eisenstadt, Feldkirch, Fürstenfeld, Gmunden, Graz, Haiming, Hörsching, Innsbruck Universität, Kufstein, Langenlebarn, Langen am Arlberg, Mönichkirchen, Pörtschach, Preitenegg, Salzburg, Sankt Michael im Lungau, Schwechat, Semmering, Spittal an der Drau, Windischgarsten, and Zeltweg. For access to these data, please contact ZAMG.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This paper is a contribution to UNESCO's FRIEND-Water program. The authors would like to thank the Austrian Climate Research Program ACRP for financial support through the project DALF-Pro (GZ B464822). We thank the Central Institution for Meteorology and Geodynamics (ZAMG) for providing meteorological data. We thank Dan Rosbjerg and the anonymous reviewer for providing insightful comments and suggestions on an earlier draft of the manuscript.

Edited by: V. Kotroni
Reviewed by: D. Rosbjerg and one anonymous referee

References

Aldrich, J.: R. A. Fisher and the making of maximum likelihood 1912–1922, Stat. Sci., 12, 162–176, doi:10.1214/ss/1030037906, 1997.

APCC: Österreichischer Sachstandsbericht Klimawandel 2014 (AAR14), Austrian Panel on Climate Change, Austrian Academy of Sciences Press, Vienna, Austria, 1096 pp., ISBN-13: 978-3-7001-7699-2, available at: http://hw.oeaw.ac.at/APCC_AAR2014.pdf (last access: 7 March 2017), 2014.

Balkema, A. and de Haan, L.: Residual life time at great age, Ann. Probab., 2, 792–804, doi:10.1214/aop/1176996548, 1974.

Basrak, B.: Fisher-Tippett Theorem, in: International Encyclopedia of Stat. Sci., edited by: Lovric, M., Berlin, Heidelberg, Springer, 525–526, 2014.

Ben-Zvi, A.: Rainfall intensity–duration–frequency relationships derived from large partial duration series, J. Hydrol., 367, 104–114, doi:10.1016/j.jhydrol.2009.01.007, 2009.

Bezak, N., Brilly, M., and Šraj, M.: Comparison between the peaks-over-threshold method and the annual maximum method for flood frequency analysis, Hydrolog. Sci. J., 59, 959–977, doi:10.1080/02626667.2013.831174, 2014.

BMLFUW: Hydrologischer Atlas Österreichs, Austrian Federal Ministry of Agriculture, Forestry, Environment and Water Management, Vienna, 2007.

Burn, D. H.: The use of resampling for estimating confidence intervals for single site and pooled frequency analysis, Hydrolog. Sci. J., 48, 25–38, doi:10.1623/hysj.48.1.25.43485, 2003.

Castillo, E., Hadi, A. S., Balakrishnan, N., and Sarabia, J. M.: Extreme Value and Related Models with Applications in Engineering and Science, Wiley Series in Probability and Statistics, Wiley, Hoboken, New Jersey, 2005.

Chavez-Demoulin, V. and Davison, A. C.: Modelling time series extremes, Revstat, 10, 109–133, 2012.

Cheng, L., AghaKouchak, A., Gilleland, E., and Katz, R. W.: Non-stationary extreme value analysis in a changing climate, Climatic Change, 127, 353–369, doi:10.1007/s10584-014-1254-5, 2014.

Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Berlin, Heidelberg, Springer, 2001.

Cunnane, C.: A particular comparison of annual maxima and partial duration series methods of flood frequency prediction, J. Hydrol., 18, 257–271, doi:10.1016/0022-1694(73)90051-6, 1973.

Davison, A. C. and Smith, R. L.: Models for exceedances over high thresholds, J. Roy. Stat. Soc. B, 52, 393–442, doi:10.2307/2345667, 1990.

Deidda, R.: A multiple threshold method for fitting the generalized Pareto distribution to rainfall time series, Hydrol. Earth Syst. Sci., 14, 2559–2575, doi:10.5194/hess-14-2559-2010, 2010.

Dupuis, D. J.: Exceedances over high thresholds: a guide to threshold selection, Extremes, 1, 251–261, doi:10.1023/A:1009914915709, 1999.

Doll, C., Trinks, C., Sedlacek, N., Pelikan, V., Comes, T., and Schultmann, F.: Adapting rail and road networks to weather extremes: case studies for southern Germany and Austria, Nat. Hazards, 72, 63–85, doi:10.1007/s11069-013-0969-3, 2013.

Eisenack, K., Stecker, R., Reckien, D., and Hoffmann, E.: Adaptation to Climate Change in the Transport Sector: A Review, PIK-Report, 122, Potsdam, Potsdam Institute for Climate Impact Research, 2011.

Embrechts, P., Klüppelberg, C., and Mikosch, T.: Modelling extremal Events for Insurance and Finance, Berlin-Heidelberg, Springer, 2003.

EPA: Addressing green infrastructure design challenges in the Pittsburgh region – Abundant and frequent rainfall, United States Environmental Protection Agency, Pittsburgh, Pennsylvania, available at: http://www.3riverswetweather.org/sites/default/files/Rainfallwhitepaper.pdf (last access: 7 March 2017), 2014

Ferreira, A., de Haan, L., and Peng, L.: On optimising the estimation of high quantiles of a probability distribution, Statistics, 37, 401–434, doi:10.1080/0233188021000055345, 2003.

Fisher, R. A.: On an absolute criterion for fitting frequency curves, Messenger of Mathematics, 41, 155–160, 1912, reprinted in Stat. Sci., 12, 39–41, 1997.

Fisher, R. A. and Tippett, L. H. C.: Limiting forms of the frequency distribution of the largest and smallest member of a sample, Proc. Camb. Philos. Soc., 24, 180–190, doi:10.1017/s0305004100015681, 1928.

Fréchet, M.: Sur la loi de probabilité de l'écart maximum, Ann. Soc. Polon. Math., 6, 92–116, 1927.

Fukutome, S., Liniger, M. A., and Süveges, M.: Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in Switzerland, Theor. Appl. Climatol., 120, 403–416, doi:10.1007/s00704-014-1180-5, 2015.

Fürst, J., Godina, R., Nachtnebel, H. P., and Nobilis, F.: Der Hydrologische Atlas Österreichs – Grundstock einer hydrologischen Geodateninfrastruktur für Ingenieure, Planer und die Öffentlichkeit, in: Angewandte Geoinformatik, edited by: Strobl, J., Blaschke, T., and Griesebner, G., Heidelberg, Verlag Wichmann, 334–343, ISBN-13: 978-3-87907-4808, 2009.

Gilbert, R. O.: Statistical Methods for Environmental Pollution Monitoring, New York, Wiley, 1987.

Gilleland, E. and Katz, R. W.: New software to analyze how extremes change over time, Eos, 92, 13–14, doi:10.1029/2011EO020001, 2011.

Gilleland, E. and Katz, R. W.: extRemes 2.0: An Extreme Value Analysis Package in R, J. Stat. Soft., 72, 1–39, doi:10.18637/jss.v072.i08, 2016.

Ghosh, S. and Resnick, S.: A discussion on mean excess plots, Stoch. Proc. Appl., 120, 1492–1517, doi:10.1016/j.spa.2010.04.002, 2010.

Gnedenko, B. V.: Sur la distribution limite du terme maximum d'une serie aleatoire, Ann. Math., 44, 423–453, doi:10.2307/1968974, 1943.

GRCA: Technical and engineering guidelines for stormwater management submissions, Ganaraska Region Conservation Authority, Port Hope, Ontario, available at: http://www.grca.on.ca/Guidelines_for_swm_submissions-_FINAL.pdf (last access: 7 March 2017), 2014.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R.: Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form, Water Resour. Res., 15, 1049–1054, doi:10.1029/WR015i005p01049, 1979.

Grotjahn, R., Black, R., Leung, R., Wehner, M. F., Barlow, M., Bosilovich, M., Gershunov, A., Gutowski, W. J., Gyakum, J. R., Katz, R. W., Lee, Y.-Y., Lim, Y.-K., and Prabhat: North American extreme temperature events and related large scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends, Clim. Dynam., 46, 1151–1184, doi:10.1007/s00382-015-2638-6, 2016.

Gumbel, E. J.: Statistics of extremes, New York, Columbia University Press, 1958.

Hald, A.: On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares, Stat. Sci., 14, 214–222, doi:10.1214/ss/1009212248, 1999.

Hiebl, J., Reisenhofer, S., Auer, I., Böhm, R., and Schöner, W.: Multi-methodical realisation of Austrian climate maps for 1971–2000, Adv. Sci. Res., 6, 19–26, doi:10.5194/asr-6-19-2011, 2011.

Hosking, J. R. M.: L-Moments: analysis and estimation of distributions using linear combinations of order statistics, J. Roy. Stat. Soc. B, 52, 105–124, doi:10.2307/2345653, 1990.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F.: Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, Technometrics, 27, 251–261, doi:10.1080/00401706.1985.10488049, 1985.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F.: Parameter and quantile estimation for the generalized Pareto distribution, Technometrics, 29, 339–349, doi:10.2307/1269343, 1987.

Hundecha, Y., Pahlow, M., and Schumann, A.: Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes, Water Resour. Res., 45, W12412, doi:10.1029/2008WR007453, 2009.

IPCC: Managing the Risks of extreme Events and Disasters to Advance Climate Change Adaptation, Special Report of the Intergovernmental Panel on Climate Change, Cambridge, Cambridge University Press, available at: http://www.ipcc.ch/pdf/special-reports/srex/SREX_Full_Report.pdf (last access: 7 March 2017), 2012.

Jarušková, D. and Hanek, M.: Peaks over threshold method in comparison with block-maxima method for estimating

high return levels of several Northern Moravia precipitation and discharges series, J. Hydrol. Hydromech., 54, 309–319, doi:10.2478/v10098-010-0009-x, 2006.

Katz, R. W.: Statistics of extremes in climate change, Climatic Change, 100, 71–76, doi:10.1007/s10584-010-9834-5, 2010.

Katz, R. W.: Statistical methods for nonstationary extremes, in: Extremes in a Changing Climate – Detection, Analysis and Uncertainty, edited by: AghaKouchak, A., Easterling, D., Hsu, K., Schubert, S., and Sorooshian, S., 15–37, Springer, Netherlands, 2013.

Katz, R. W., Parlange, M. B., and Naveau, P.: Statistics of extremes in hydrology, Adv. Water Resour., 25, 1287–1304, doi:10.1016/S0309-1708(02)00056-8, 2002.

Kendall, M. G.: Rank Correlation Methods, 4th Edition, Griffin, London, 1976.

Koetse, M. J. and Rietveld, P.: The impact of climate change and weather on transport: An overview of empirical findings, Transport. Res. D-Tr. E., 14, 205–221, doi:10.1016/j.trd.2008.12.004, 2009.

Laaha, G. and Blöschl, G.: Seasonality indices for regionalizing low flows, Hydrol. Process., 20, 3851–3878, doi:10.1002/hyp.6161, 2006.

Langbein, W. B.: Annual Floods and the partial duration flood series, Am. Geophys. Union Trans., 30, 879–881, 1949.

Leadbetter, M. R., Lindgren, G., and Rootzén, H.: Extremes and Related Properties of Random Sequences and Processes, Springer, London, 1983.

Lee, J., Fan, Y., and Sisson, S. A.: Bayesian threshold selection for extremal models using measures of surprise, Comput. Stat. Data An., 84, 84–99, doi:10.1016/j.csda.2014.12.004, 2014.

MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M., and Russell, G.: A flexible extreme value mixture model, Comput. Stat. Data An., 55, 2137–2157, doi:10.1016/j.csda.2011.01.005, 2011.

Mann, H. B.: Nonparametric tests against trend, Econometrica, 13, 245–259, doi:10.2307/1907187, 1945.

Madsen, H., Rasmussen, P. F., and Rosbjerg, D.: Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. At-site modelling, Water Resour. Res., 33, 747–757, doi:10.1029/96WR03848, 1997a.

Madsen, H., Pearson, C. P., and Rosbjerg, D.: Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 2. Regional modelling, Water Resour. Res., 33, 759–769, doi:10.1029/96WR03849, 1997b.

Maraun, D., Rust, H. W., and Osborn, T. J.: The annual cycle of heavy precipitation across the United Kingdom: a model based on extreme value statistics, Int. J. Climatol., 29, 1731–1744, doi:10.1002/joc.1811, 2009.

Makkonen, L.: Plotting Positions in Extreme Value Analysis, J. Appl. Meteorol., 45, 334–340, doi:10.1175/JAM2349.1, 2006.

Matulla, C., Hollosi, B., Andre, K., Gringinger, J., Chimani, B., Namyslo, J., Fuchs, T., Auerbach, M., Herrmann, C., Sladek, B., Berghold, H., Gschier, R., and Eichinger-Vill, E.: Climate change driven evolution of hazards to Europe's transport infrastructure throughout the 21st century, Theor. Appl. Climatol., in press, 2017.

Méndez, F. J., Menéndez, M., Luceño, A., Medina, R., and Graham, N. E.: Seasonality and duration in extreme value distributions of significant wave height, Ocean Eng., 35, 131–138, doi:10.1016/j.oceaneng.2007.07.012, 2008.

Merz, R. and Blöschl, G.: A process typology of regional floods, Water Resour. Res., 39, 1340–1359, doi:10.1029/2002WR001952, 2003.

Meyer, M. D., Flood, M., Keller, J., Lennon, J., McVoy, G., Dorney, C., Leonard, K., Hyman, R., and Smith, J.: Climate Change, Extreme Weather Events, and the Highway System, Practitioner's Guide and Research Report, NCHRP Report 750, Strategic Issues Facing Transportation, Volume 2, Transportation Research Board, Washington, DC, 2014.

Michaelides, S.: Vulnerability of transportation to extreme weather and climate change, Nat. Hazards, 72, 1–4, doi:10.1007/s11069-013-0975-5, 2014.

Mkhandi, S., Opere, A. O., and Willems, P.: Comparison between annual maximum and peaks over threshold models for flood frequency prediction, Proceedings of the International Conference on UNESCO FRIEND/Nile Project: "Towards a better Cooperation", Sharm-El-Sheikh, Egypt, 2005.

Nestroy, O.: Soil sealing in Austria and its consequences, Ecohydrol. Hydrobiol., 6, 171–173, doi:10.1016/S1642-3593(06)70139-2, 2006.

Northrop, P. N. and Coleman, C. L.: Improved threshold diagnostic plots for extreme value analyses, Extremes, 17, 289–303, doi:10.1007/s10687-014-0183-z, 2014.

Papalexiou, S. M. and Koutsoyiannis, D.: Battle of extreme value distributions: A global survey on extreme daily rainfall, Water Resour. Res., 49, 187–201, doi:10.1029/2012WR012557, 2013.

Parey, S., Hoang, T. T. H., and Dacunha-Castelle, D.: Different ways to compute temperature return levels in the climate change context, Environmetrics, 21, 698–718, doi:10.1002/env.1060, 2010.

Pickands, J.: Statistical inference using extreme order statistics, Ann. Stat., 3, 119–131, doi:10.1214/aos/1176343003, 1975.

Rosbjerg, D.: Return Periods of Hydrological Events, Nord. Hydrol., 8, 57–61, 1977.

Roth, M., Jongbloed, G., and Buishand, T. A.: Threshold selection for regional peaks-over-threshold data, J. Appl. Stat., 43, 1291–1309, doi:10.1080/02664763.2015.1100589, 2016.

Scarrott, C. J. and MacDonald, A.: A review of extreme value threshold estimation and uncertainty quantification, Revstat, 10, 33–59, 2012.

Schweikert, A., Chinowsky, P., Kwiatkowski, K., and Espinet, X.: The infrastructure planning support system: Analyzing the impact of climate change on road infrastructure and development, Transp. Policy, 35, 146–153, doi:10.1016/j.tranpol.2014.05.019, 2014a.

Schweikert, A., Chinowsky, P., Espinet, X., and Tarbert, M.: Climate Change and Infrastructure Impacts: Comparing the Impact on Roads in ten Countries through 2100, Procedia Eng., 78, 306–316, doi:10.1016/j.proeng.2014.07.072, 2014b.

Smith, R. L.: Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone, Stat. Sci., 4, 367–377, 1989.

Stedinger, J. R., Vogel, R. M., and Foufoula-Georgiou, E.: Frequency analysis of extreme events, in: Handbook of Hydrology, edited by: Maidment, D. R., New York, McGraw-Hill, Inc., 1993.

Tallaksen, L. M. and van Lanen, H. A. J.: Hydrological Drought, 1st Edition Processes and Estimation Methods for Streamflow and

Groundwater, Developments in Water Science, 48, Amsterdam, Boston, Elsevier Science, 2004.

Tancredi, A., Anderson, C., and O'Hagan, A.: Accounting for threshold uncertainty in extreme value estimation, Extremes, 9, 86–106, doi:10.1007/s10687-006-0009-8, 2006.

Thompson, P., Cai, Y., Reeve, D., and Stander, J.: Automated threshold selection methods for extreme wave analysis, Coast. Eng., 56, 1013–1021, doi:10.1016/j.coastaleng.2009.06.003, 2009.

TRB: Potential Impacts of Climate Change on US Transportation, Transportation Research Board Special Report 290, Washington, DC, available at: http://onlinepubs.trb.org/onlinepubs/sr/sr290.pdf (last access: 7 March 2017), 2008.

UNECE: Climate Change Impacts and Adaptation for International Transport Networks, United Nations Economic Commission for Europe, New York and Geneva, available at: http://www.unece.org/fileadmin/DAM/trans/main/wp5/publications/climate_change_2014.pdf (last access: 7 March 2017), 2013.

Villarini, G., Smith, J. A., Ntelekos, A. A., and Schwarz, U.: Annual maximum and peaks-over-threshold analyses of daily rainfall accumulations for Austria, J. Geophys. Res., 116, D05103, doi:10.1029/2010JD015038, 2011.

Wadsworth, J. L.: Exploiting structure of maximum likelihood estimators for extreme value threshold selection, Technometrics, 58, 116–126, doi:10.1080/00401706.2014.998345, 2016.

Wadsworth, J. L. and Tawn, J. A.: Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling, J. Roy. Stat. Soc. B, 74, 543–567, doi:10.1111/j.1467-9868.2011.01017.x, 2012.

Weibull, W.: A statistical theory of strength of materials, Ingeniörsvetenskapsakademiens handlingar, 151, 1–45, 1939.

WMO: Guide to Hydrological Practices, Volume II, Management of Water Resources and Application of Hydrological Practices, WMO-No. 168, Geneva, World Meteorological Organization, available at: http://www.whycos.org/chy/guide/168_Vol_II_en.pdf (last access: 7 March 2017), 2009.

Zhang, X., Zwiers, F. W., and Li, G.: Monte Carlo experiments on the detection of trends in extreme values, J. Climate, 17, 1945–1952, doi:10.1175/1520-0442(2004)017<1945:MCEOTD>2.0.CO;2, 2004.