AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 1, 21–30

Statistical Modelling of Annual Maxima in Hydrology

Johannes Hofrichter[1], Till Harum[2] and Herwig Friedl[3]

[1]Institute of Applied Statistics, Joanneum Research, Austria
[2]Institute of Water Resources Management, Joanneum Research, Austria
[3]Institute of Statistics, Graz University of Technology

Abstract: In this paper conditional modelling of annual maxima for predicting flood water is considered. The aim is to predict flood water of rivers, where no data about discharge but data about properties of the catchment of the rivers are available. A generalized linear mixed model is used to model the annual maxima depending on properties of the catchment and to take the correlation among measurements of one year into account. The fitted means and variances according to this model are plugged into the method of moment estimates of the parameters of the Gumbel distribution to obtain some extreme quantiles. These quantiles are commonly used to predict flood water of rivers. This approach is applied to data from Styria (Austria). The result is a satisfactory model for predicting flood water for rivers, where no data about the discharge are available.

Zusammenfassung: In diesem Beitrag wird das bedingte Modellieren von jährlichen Maxima zur Vorhersage von Hochwasser betrachtet. Das Ziel ist es die Höhe von Hochwasser von Flüssen vorherzusagen, bei denen keine Daten bezüglich des Abflusses aber bezüglich der Eigenschaften der Einzugsgebiete der Flüsse existieren. Es werden generalisierte lineare Mischmodelle verwendet um einerseits die jährlichen Maxima in Abhängigkeit von den Eigenschaften der Einzugsgebiete zu modellieren und anderseits die Korrelation zwischen den jährlichen Maxima verschiedener Flüsse eines Jahres zu berücksichtigen. Die geschätzten Erwartungswerte und Varianzen unter diesem Modell werden in die Momentenschätzer der Parameter der Gumbel Verteilung eingesetzt, um Schätzer für extreme Quantile zu erhalten. Diese Quantile werden häufig verwendet um Hochwasser von Flüssen vorherzusagen. Diese Methode wurde an Flüssen in der Steiermark (Österreich) angewendet. Das Resultat ist ein zufriedenstellendes Modell zur Vorhersage von Hochwasser für Flüsse, bei denen keine Abflussdaten zur Verfügung stehen.

Keywords: Gumbel Distribution, Generalized Linear Mixed Model, Random Effects, Annual Maxima.

1 Introduction

Annual maxima of the discharge of a river are commonly used to predict flood water. In this paper annual maxima from several rivers are modelled to predict flood water for rivers, where no data about the discharge are available.

The analyzed data are annual maxima of discharge from rivers in Styria (Austria). In hydrology it is well known, that the discharge and hence the annual maxima are influenced by the properties of the catchment of a river. Such a catchment is defined as

the area of the landscape, where all the rain falling on this area discharges into the river. These properties can be easily obtained from a Geographic Information System (GIS) and are available for any river in Styria. Thus, the idea was to model annual maxima depending on these properties to obtain an appropriate model for predicting flood water. Because the distribution of such annual maxima is certainly non-normal, we do not consider any linear regression models. However, modelling can be done within the context of quasi-likelihood estimation in the generalized linear model (GLM) framework. To take the correlation among observations of different rivers at one year into account, the model is augmented with a random effect, which leads to the broad class of generalized linear mixed models (GLMMs). First analysis of the annual maxima indicates, that it is reasonable to assume temporal independence over the years. Therefore, only spatial dependency of the data is considered. The fitted values are plugged into the method of moment estimator of the parameters in the extreme value distribution to obtain estimates of certain quantiles. These are often used to predict flood water. This approach provides satisfactory estimates of flood water and can be applied for rivers, where only data about the properties of the catchment are available.

The analysis of extreme values has a long history and goes back to the traditional approach introduced by Gumbel (1958). There is an enormous literature on this topic, especially applied on environmental data such as rainfall and river heights. Davison and Smith (1990) presented a method for modelling univariate extremes in dependency of explanatory variables. Their approach is based on the exceedance over a threshold, with the assumption, that the difference between the observations and the threshold follow a generalized Pareto distribution. An extension to multivariate extremes is given in Coles and Twan (1991) and this technique is later on applied to extremes of areal rainfall in Coles and Twan (1996). In this work the extremes were modelled by means of explanatory variables taking their spatial dependence into account. A critical point of these models is the choice of a suitable threshold. If this value is chosen too large, then only some of the data exceed this value and all of them can be considered as extremes. If the value is, however, small, then we have many observations above it but most of them are not extremes at all. To avoid this problem we develop a method based on the classical extreme value theory, which additionally takes the spatial dependence into account.

2 Classical Extreme Value Theory

By means of classical extreme value distributions we are often able to analyze the statistical behavior of

$$y = \max\{x_1, \ldots, x_n\},$$

where $x_1, \ldots, x_n$ is a sequence of independent identically distributed (i.i.d.) random variables having some distribution function F. Such a sequence usually represents values measured on a regular time scale. In our application, these x_i's ($i = 1, \ldots, n = 365$) represent daily measured maxima of discharge of a river, so that y is the annual maximum of the discharge. As y is the maximum of a block of values, it is often denoted as block maximum. In general F is unknown and therefore the distribution function G of y can

not be calculated exactly. But for large n (as in our case), the block maximum y follows one of three extreme value distributions, known as the Gumbel, Fréchet and Weibull distribution (Coles, 2001). These three distributions can be combined into a single family of distributions, the Generalized Extreme Value (GEV) family, with distribution function

$$G(y|\lambda,\nu,\xi) = \exp\left\{-\left[1+\xi\left(\frac{y-\lambda}{\nu}\right)\right]^{-1/\xi}\right\}$$

defined on the set $\{y \,|\, 1+\xi(y-\lambda)/\nu > 0\}$, where the parameters satisfy $-\infty < \lambda < \infty$, $\nu > 0$ and $-\infty < \xi < \infty$ and are usually referred to as the location, scale and shape parameter, respectively. In this parametrization the cases of $\xi > 0$ and $\xi < 0$ correspond to the Fréchet and Weibull distribution, respectively. The special case $\xi = 0$ is known as the Gumbel distribution with distribution function

$$G(y|\lambda,\nu) = \exp\left\{-\exp\left[-\left(\frac{y-\lambda}{\nu}\right)\right]\right\}. \tag{1}$$

The theoretical moments according to this distribution are

$$\mathrm{E}(Y) = \lambda + \nu\gamma\,, \qquad \text{and} \quad \mathrm{var}(Y) = \nu^2\pi^2/6\,, \tag{2}$$

where $\gamma \simeq 0.57722$ is Euler's constant.

The advantage of this parametrization is, that for estimating the parameters no prior knowledge about the distribution of y is necessary. The data themselves determine, which of these three distributions is appropriate. Thus, we first fit a model allowing for all three parameters and then we subsequently test for the necessity of the shape parameter (Hosking, 1984). Because there is strong evidence that the annual maxima in our application follow a Gumbel distribution, we restrict our attention only onto this special member of the GEV family in the remainder.

Consider now a random sample of n annual maxima $y = (y_1, \ldots, y_n)$ from a Gumbel distribution (1). Then the log likelihood function of this sample is

$$\ell(\lambda,\nu \,|\, y) = -n\log\nu - \sum_{i=1}^{n}\left(\frac{y_i-\lambda}{\nu}\right) - \sum_{i=1}^{n}\exp\left(-\frac{y_i-\lambda}{\nu}\right). \tag{3}$$

To find the maximum likelihood (ML) estimates, the derivatives of (3) with respect to the parameters λ and ν

$$\begin{aligned}
\frac{\partial\ell(\lambda,\nu|y)}{\partial\lambda} &= \frac{n}{\nu} - \frac{1}{\nu}\sum_{i=1}^{n}\exp\left(-\frac{y_i-\lambda}{\nu}\right)\\
\frac{\partial\ell(\lambda,\nu|y)}{\partial\nu} &= -\frac{n}{\nu} + \sum_{i=1}^{n}\frac{y_i-\lambda}{\nu^2}\left[1-\exp\left(-\frac{y_i-\lambda}{\nu}\right)\right],
\end{aligned}$$

have to be solved. There is no analytical solution to the system of these two equations, but standard iterative optimization methods can be applied in a straightforward way. An

alternative method based on the empirical mean and variance is the method of moments. The theoretical moments in (2) yield estimates

$$\hat{\lambda} = \bar{y} - \hat{\nu}\gamma\,, \qquad \text{and} \qquad \hat{\nu} = \sqrt{6s_y^2/\pi^2}\,, \tag{4}$$

where $\bar{y}$ and s_y^2 denotes empirical mean and variance of our sample y.

A common question in hydrology is: What is the highest level the flood water will exceed once every T years? This is often denoted as T-year threshold or return level. In extreme value theory this return level is defined as the quantile y_q with $q = 1 - 1/T$, i.e. $G(y_q|\lambda,\nu) = 1 - 1/T$. Thus, in order to predict extreme flood water of a river for several return periods T, we are very interested in some extreme quantiles of a Gumbel distribution. These quantiles are obtained by inverting (1) resulting in

$$y_q = \lambda - \nu \log\{-\log(1 - 1/T)\}\,, \tag{5}$$

where $G(y_q|\lambda,\nu) = 1 - 1/T$ and T is the return period of interest. For example, if y_i are annual maxima and the given return period is $T = 100$ years, then y_q is the level, which is expected to exceed once every 100 years.

If we are interested in predicting flood water for a specified return period, we first estimate the parameters and then plug them into the quantiles (5) (see e.g. Coles, 2001). Of course this is only possible, if concrete data on such annual maxima are available. Sometimes there is only information about catchment properties available. In such cases it would be very useful to have some knowledge about the relationship between annual maxima of the discharge of the river and properties of the corresponding catchment.

3 Modelling Annual Maxima

Now the focus is on modelling the mean of non-identically distributed annual maxima to later use these estimates in the return levels (5).

Consider block maxima y_{it} of $i = 1, \ldots, n$ rivers observed at time t. This can be annual maxima of several years from different rivers. It is assumed, that the maxima y_{it} of a subject are independent and Gumbel distributed, that is $y_{it} \overset{ind}{\sim} \text{Gumbel}(\lambda_i, \nu_i)$. Furthermore, we assume that the parameters λ_i and ν_i of two different subjects are independent. Additionally, some explanatory variables $x_{it} = (x_{it1}, \ldots, x_{itp})$ are observed. The idea is to estimate the mean and the variance of y_{it} depending on such explanatory variables and plug this estimates into the method of moment estimates of the parameters of the Gumbel distribution. If the mean-variance relationship of y_{it} is known, then it can be modelled by a quasi-likelihood approach (Wedderburn, 1974) within the GLM framework (McCullagh und Nelder, 1989). In such GLMs the mean of a response variable is modelled as a function of the linear predictor, i.e.

$$\mathrm{E}(y_{it}) = \mu_{it} = g^{-1}(x_{it}\beta)\,,$$

where β is the p-dimensional column vector of interest and $g(\cdot)$ is the so called link function, which links the linear predictor to the mean. The variance of y_{it} is assumed to be proportional to a specified function of the mean and is thus given by

$$\mathrm{var}(y_{it}) = \phi V(\mu_{it})\,.$$

We call $V(\cdot)$ the variance function and ϕ the dispersion parameter which may be known or unknown.

Now, we allow observations from different rivers but observed at the same time t to be positively correlated, i.e. $\text{cor}(y_{it}, y_{jt}) > 0$. There are two ways to take this correlation into account. If the focus of the analysis is on the population average, the quasi-likelihood approach can be generalized by allowing a very general form of the variance structure of y_{it} (Fitzmaurice et al., 2004). In this case the marginal mean is modelled and the model can be fitted by Generalized Estimating Equations (GEEs). Here we consider a second way where the subject specific analysis is of main interest. Thus, the mean is conditionally modelled by augmenting the GLM with a random factor for each year, say z_t. This combination of fixed effects β and random effects z_t defines a GLMM. In contrast to GLMs, in GLMMs the conditional mean of y_{it} given the random effect z_t is modelled. Hence, we consider

$$\mathrm{E}(y_{it}|z_t) = g^{-1}(x_{it}\beta + z_t)\,.$$

Here, z_t is a random intercept specific for time t and is assumed to follow a normal distribution with zero mean and variance σ_z^2, i.e. $z_t \overset{iid}{\sim} \mathrm{N}(0, \sigma_z^2)$. Note, that this additional random effect induces correlation between observations from different rivers but made at the same time t. The model can be fitted either by the penalized quasi-likelihood (PQL) approach as discussed in Breslow and Clayton (1993) or by applying the EM-Algorithm of Dempster et al. (1977). If the dispersion parameter ϕ is unknown, it has to be estimated, too. A usual estimator is the mean Pearson statistic.

Once the parameters β and ϕ are estimated, the random effects can be predicted by the best linear unbiased predictor (BLUP) which is

$$\hat{z}_t = \mathrm{E}(z_t|y_{it}, \hat{\beta}, \hat{\phi})$$

and coincides with the empirical Bayes predictor. Thus the estimated conditional mean is

$$\widehat{\mathrm{E}}(y_{it}|z_t) = g^{-1}\left(x_{it}\hat{\beta} + \mathrm{E}(z_t|y_{it}, \hat{\beta}, \hat{\phi})\right)\,.$$

Note, that this estimates the conditional mean given the random effects. The marginal mean is obtained by integrating out the random effect and is given by

$$\mathrm{E}(y_{it}) = \mathrm{E}_z\left[g^{-1}(x_{it}\beta + z_t)\right]\,.$$

In the case of a log-link we have

$$\mathrm{E}(y_{it}) = \exp\left[x_{it}\beta + \log M(e^{z_t})\right]\,, \tag{6}$$

where $M(\cdot)$ is the moment generating function of z_t with $\log M(e^{z_t}) = \sigma_z^2/2$. Substituting this in (6) leads to the marginal mean of y_{it}

$$\mathrm{E}(y_{it}) = \exp\left[x_{it}\beta + \sigma_z^2/2\right]\,.$$

To obtain an estimator of the time invariant mean $\mathrm{E}(y_i)$ for river i an appropriate summary statistic has to be applied. In case of time invariant explanatory variables, i.e. $x_{it_1} = x_{it_2}$ for all $t_1 \neq t_2$, the equation

$$\mu_i = \mathrm{E}(y_i) = \mathrm{E}(y_{it})$$

holds, because the shift in the intercept, $\sigma_z^2/2$, is constant over time.

Now, for a given mean-variance relationship the estimated mean $\hat{\mu}_i$ and variance $\hat{\phi}V(\hat{\mu}_i)$ can be plugged into the method of moment estimates of the parameters of the Gumbel distribution and the return level based on this approach can be estimated. In particular cases, the calculation of the return values can be simplified. Consider a quadratic mean-variance relationship, as it is the case for Gamma distributed responses. Then the estimates (4) based on the fitted mean $\hat{\mu}_i$ and variance $\hat{\phi}V(\hat{\mu}_i)$ are

$$\hat{\lambda} = \hat{\mu}\left(1 - \frac{\gamma}{\pi}\sqrt{6\hat{\phi}}\right), \qquad \text{and} \quad \hat{\nu} = \hat{\mu}\frac{1}{\pi}\sqrt{6\hat{\phi}} \tag{7}$$

and the return levels (5) simplify to

$$\hat{y}_q = \hat{\mu}\left[1 + \frac{1}{\pi}\sqrt{6\hat{\phi}}\ \left[\gamma + \log(-\log(1 - 1/T))\right]\right]. \tag{8}$$

This approach allows to predict return levels based on the estimation results of GLMMs.

4 Application

The data analyzed in this study are annual maxima of discharge measured at 102 rivers in Styria, Austria. The lengths of the observation times are between 10 and 52 years, giving an unbalanced data set. Empirical analysis suggested that the annual maxima follow a Gumbel distribution. For each river, there are various properties of its catchment available. Amongst these properties there is the catchment's `area`, the mean altitude `dtm`, the average amount of `rain` in this area, drainage density `gd` and types of land use, like the proportions of `forest` and no-vegetation (`noveg`). The idea now is to model the annual maxima in terms of these properties, because this information can be easily obtained from a GIS for any river in Styria. Thus, the mean annual maximum is assumed to be a function of these properties. Because the hydrogeological conditions are also relevant, the rivers were grouped into five homogeneous `regions`. Assuming that the annual maxima of a river are uncorrelated over time and that the annual maxima of different rivers in the same year are also uncorrelated, the mean response is analyzed within the framework of ordinary GLMs. For this purpose an appropriate link function and a suitable mean-variance relationship has to be specified. A log-link model is applied because the mean discharge of a river has to be a positive quantity. To get a first glimpse of the true mean-variance relationship, we generate the scatter-plot of river specific empirical means against their variances in Figure 1. The point pattern reveals a quadratic mean-variance relationship, which is estimated by the function $0.215\mu^2$. Because of that, a quasi-likelihood approach is utilized based on a log-link model for the mean and on a quadratic variance function for the responses. A variable selection procedure was then applied to select the best fitting set of explanatory variables. Finally we found the model

$$\log(\mu) = \texttt{region} * \Big(\log(\texttt{area}) + \texttt{noveg} + \texttt{rain} + \texttt{gd} + \texttt{dtm} + \texttt{forest} + \texttt{dtm.forest}\Big),$$

with $\texttt{dtm.forest} = \sqrt{\texttt{dtm} \cdot \texttt{forest}}$. This model results in an estimated dispersion parameter of $\hat{\phi} = 0.267$, which is slightly larger than the estimate from fitting a quadratic curve through the empirical means/variances in Figure 1.

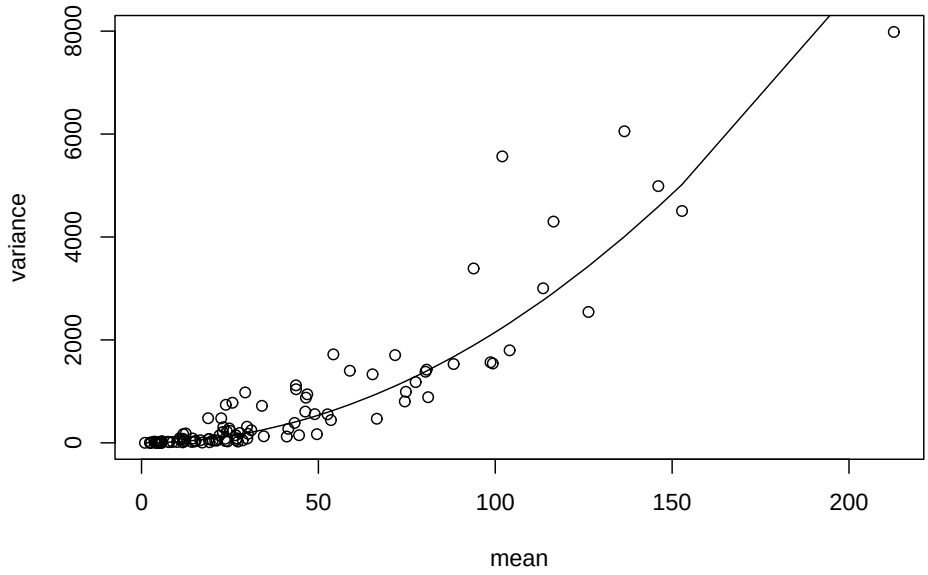

Figure 1: Empirical means and variances of annual maxima from 102 rivers and a fitted quadratic (variance) model through these points.

Table 1: Estimated intercept and standard error for each region.

Region	1	2	3	4	5
β_0	-9.36	19.77	-5.14	0.57	-2.27
s.e.	1.35	5.81	0.57	0.67	0.66

However, as the annual maxima of discharge usually result from heavy rain storm events, we should expect that observations from the same year are presumably positively correlated. This might be due to the fact that heavy rain storms are not restricted to a small area or to a single catchment of a river. Thus, this spatial correlation should be taken into account and considered in the model. We extend the model just found before and incorporate some additional random effects which are specific for each year. Therefore, the considered model includes an intercept for each region, region specific coefficients for all explanatory variables and an random intercept for each year. In Table 1 estimates of the region specific intercepts are listed. The estimated coefficients in the reference region 1 are given in Table 2 and their deviations for regions 2 to 5 are in Table 3.

The estimated dispersion parameter under this model is $\hat{\phi} = 0.199$, which is much closer the value from the model in Figure 1. Of course it is still smaller than the respective fixed effects result, as the random effect now also explains some variability of the responses.

Because the main interest is on estimating return levels and hence on quantiles of the Gumbel distribution, the fitted mean and variance of each river was plugged into (7) to obtain the estimated quantiles (8). This was done for the results of both considered

Table 2: Coefficients and standard errors for the reference region 1.

Predictor	$\hat{\beta}$	s.e.
`log(area)`	0.830	0.067
`noveg`	−0.007	0.017
`dtm`	0.009	0.002
`rain`	0.002	0.0002
`forest`	0.170	0.032
`dtm.forest`	−0.060	0.015
`gd`	0.336	0.209

Table 3: Coefficients and standard errors for the deviations of each region from region 1.

Region	2		3		4		5	
Predictor	$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.
`log(area)`	−0.091	0.118	0.149	0.074	−0.135	0.076	0.126	0.072
`noveg`	−0.122	0.036	−0.024	0.024	0.013	0.059	−0.035	0.026
`dtm`	−0.012	0.002	−0.003	0.003	−0.007	0.005	0.050	0.007
`rain`	−0.012	0.003	0.0002	0.001	−0.001	0.0004	−0.004	0.001
`forest`	−0.338	0.050	−0.077	0.047	−0.148	0.073	0.599	0.096
`dtm.forest`	0.094	0.021	0.017	0.022	0.041	0.036	−0.361	0.052
`gd`	−1.168	0.417	−0.755	0.256	−0.428	0.246	0.792	0.320

models. Then the obtained quantiles are compared to the reference quantiles based on the ML estimates of the parameters of the Gumbel distribution. In Figure 2 the reference return levels are plotted against the return levels based on the fitted mean and variance of the GLMM.

To compare both models we consider respective mean residual sum of squares. The residuals are defined as differences between reference return levels and return levels based on the GLM and on the GLMM. Their values are 4524 for the GLM and 4252 for the GLMM. This confirms the impression when comparing estimates of the dispersion parameter, namely that an additional random effect leads to a better result.

5 Results and Further Investigations

We present a method to predict flood water for rivers, even when no data on the discharge of the river is available. It is based on modelling the annual maxima depending only on some properties of the catchment of the river. This enables to predict flood water for any river, as soon as some characteristics of its catchment are known. A GLMM is utilized to model the mean of these annual maxima, mainly because allowing for random intercepts enhances the goodness-of-fit and it also accounts for an appropriate correlation structure between measurements within the same year.

As this method applied to data on rivers from Styria provides satisfactory prediction of flood water, it has some limitations. First it can be only applied, if we assume that

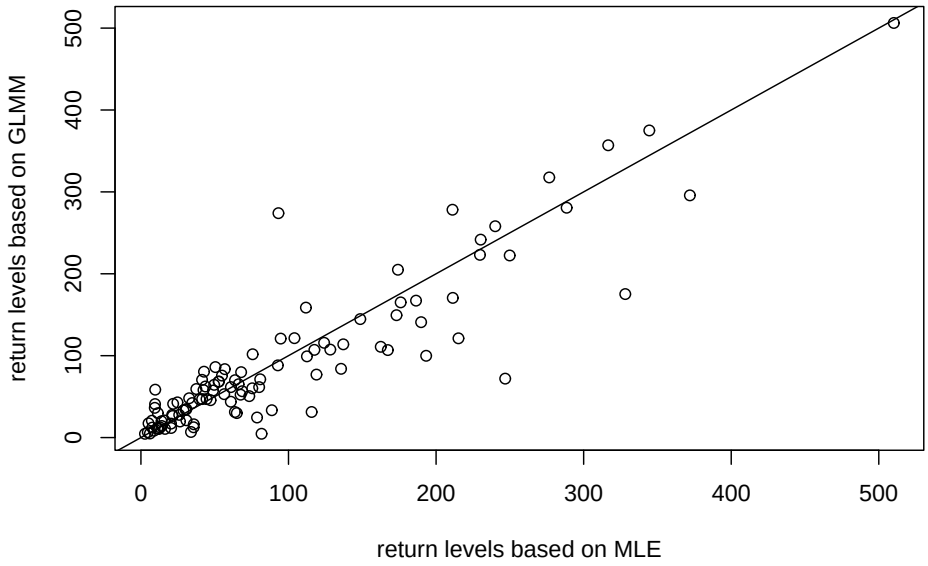

Figure 2: Comparison of the predicted return levels for a return period of 100 years based on the ML estimates and the fitted means and dispersion parameter of the GLMM.

the extremes stem from a Gumbel distribution. Second, as with the classical univariate approach this method is wasteful of data, thus it should be only applied if sufficiently many data is available. Whether this is compensate by analyzing data of several subjects together is an open question. Finally, confidence interval for return levels are not directly available and method for calculating such confidence intervals have to be developed.

As the dispersion parameter plays a main role for predicting return levels, further work has to be done to compare different estimates of this parameter. Second, prediction intervals for the return levels are of interest, thus it should be investigated, how at least approximative prediction intervals can be obtained. And finally, if we are not able to assume, that the maxima follow a Gumbel distribution, how can this method be extended onto the entire GEV family.

Acknowledgements

The authors are grateful for the comments of an anonymous referee which have improved the readability of this manuscript.

References

Breslow, N., und Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9–25.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.

Coles, S., und Twan, J. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society, Series B*, *53*, 377–392.

Coles, S., und Twan, J. (1996). Modelling extremes of areal rainfall process. *Journal of the Royal Statistical Society, Series B*, *58*, 392–347.

Davison, A., und Smith, R. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society, Series B*, *52*, 393–442.

Dempster, A., Laird, N., und Rubin, D. (1977). Maximum likelihood with incomplete data via the E-M algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Fitzmaurice, G. M., Laird, N. M., und Ware, J. H. (2004). *Applied Longitudinal Analysis*. Wiley.

Gumbel, E. (1958). *Statistics of Extremes*. New York: Columbia University Press.

Hosking, J. (1984). Testing wether the shape parameter is zero in the generalized extreme-value distribution. *Biometrica*, *71*(2), 367–374.

McCullagh, P., und Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall.

Wedderburn, R. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrica*, *61*, 439–447.

Author's addresses:

Johannes Hofrichter
Institute of Applied Statistics
JOANNEUM RESEARCH
Steyrergasse 25a
A-8010 Graz

Tel. +43 316 876 1556
Fax +43 316 876 91556
E-mail: johannes.hofrichter@joanneum.at

Till Harum
Institute of Water Resources Management
JOANNEUM RESEARCH
Elisabethstraße 16
A-8010 Graz

Tel. +43 316 876 1372
Fax +43 316 876 91372
E-mail: till.harum@joanneum.at

Herwig Friedl
Institute of Statistics
Graz University of Technology
Steyrergasse 17
A-8010 Graz

Tel. +43 316 873 6477
Fax +43 316 873 6977
E-mail: hfriedl@tugraz.at