Probability Distributions of Annual Maximum River Discharges in North-Western and Central Europe

P H A J M van Gelder[1], N M Neykov[2], P Neytchev[2], J K Vrijling[1], H Chbab[3]

1 Delft University of Technology, Faculty of Civil Engineering, P.O. Box 5048, 2600 GA Delft, The Netherlands, E-mail: p.vangelder@ct.tudelft.nl

2 National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, 66 Tsarigradsko shaussee, Sofia 1784, Bulgaria, E-mail: neyko.neykov@meteo.bg

3 Ministry of Water Management, Postbus 17, 8200 AA Lelystad, The Netherlands, E-mail: H.Chbab@RIZA.RWS.minvenw.nl

ABSTRACT: The goal of this study is to evaluate the goodness of fit of alternate PDFs (probability distribution functions) to sequences of annual maximum streamflows in North-Western and Middle Europe. Though we never know with certainty the true population from which observed streamflows arise, studies such as this may provide some guidence on which PDFs provide a reasonable approximation. L-Moment diagrams were constructed for annual maxima streamflows at more than 200 river basins in Germany, Belgium, France, Luxemburg, The Netherlands, Switzerland, Austria, Czech Republic, Poland, Slovakia, Hungary and the UK. Homogeneous regions will be derived on basis of statistical techniques and physical-based considerations. Goodness of fit comparisons will reveal which distribution functions provide the best approximations to the distribution of the annual maxima flood flows for each homogeneous region.

1 INTRODUCTION

It has long been recognized that many annual flood series are too short to allow for a reliable estimation of extreme events. The difficulties are related both to the identification of the appropriate statistical distribution for describing the data and to the estimation of the parameters of a selected distribution. Regionalization provides a means to cope with this problem by assisting in the identification of the shape of potential parent distributions, leaving only a measure of scale to be estimated from the at-site data.

Although generally recognized as a powerful means to improve flood estimates, research in regional flood frequency analysis is hampered by the unwillingness of researchers to deal with problems that cannot be treated mathematically rigourously. In fact, regional flood frequency analysis calls for assumptions, tests, and methods of a somewhat ad hoc nature. It is generally difficult to assess or compare the performance of regional estimation methods, because the degree to which implied assumptions are valid is hard to measure or quantify in practice. This, however, should challenge rather than discourage hydrologists. At present, the index flood method is the most used regional flood frequency procedure. A fundamental assumption of the index flood method is that data at different sites in a region follow the same distribution except for scale.

Regional flood frequency analysis involves two major steps: (1) Grouping of sites into homogeneous regions, and (2) Regional estimation of flood quantiles at the site of interest. The performance of any regional estimation method strongly depends on the grouping of sites into homogeneous regions. Geographically contiguous regions have been used for a long time in hydrology, but have been critized for being of arbitrary character. In fact, the geographical proximity does not guarantee hydrological similarity. During the last five to ten years researchers have attempted to develop methods in which similarity between sites is defined in a multidimensional space of catchment-related characteristics or statistical characteristics.

A significant contribution to solving the delineation issue is the region-of-influence approach. This method dispenses completely with the classical notion of regions in that each site is allowed to have its own region. The site of interest is located at the centre of gravity in a space of relevant flood and/or catchment characteristics, each weighted properly according to its relevance. The method also involves

the choice of a distance threshold; only sites whose distance to the target site (in the weighted attribute space) does not exceed this threshold are included in the region-of-influence. An advantage of the region-of-influence method is that in the estimation of a regional growth curve, each site can be weighted according to its proximity to the site of interest.

In this paper, cluster analysis is used as a first attempt to group sites into homogeneous regions. The delineation of homogeneous regions is closely related to the identification of the common regional distributions that apply within each region. A region can only be considered homogeneous if sufficient evidence can be established that data at different sites in the region are drawn from the same parent distribution (except for the scale parameter). L-moment ratio diagrams have become popular tools for regional distribution identification, and for testing for outlier stations. Hosking and Wallis (1997) developed several tests for use in regional studies. They gave guidelines for judging the degree of homogeneity of a group of sites, and for choosing and estimating a regional distribution.

In Regional Frequency Analysis (RFA) data are assumed to come from homogeneous regions. To aid the presentation, a formal definition is given. Let Q_{ij}, $j=1,...,n_i$, be observed data at N sites of a region, with sample size n_i at site i, and let $Q_i(F)$, $0<F<1$, be the quantile function of the distribution at site i. A region of N sites is called homogeneous if $Q_i(F)=\mu_i q(F)$, $i=1,...,N$, where μ_i is the site-dependent scale factor and q(F) is the quantile of the regional frequency distribution.

Hosking and Wallis (1997) developed a unified robust approach to RFA, based on L-moments described by Hosking (1990), that involves objective and subjective techniques for defining homogeneous regions, of assigning sites to regions, identifying and fitting regional probability distributions to data and testing hypotheses about distributions. By robustness Hosking and Wallis (1997) refer to statistics that work well even if the data are contaminated or the model assumptions are slightly violated. The advantages of their approach over the conventional method of moments and the maximum likelihood method are the smaller effect of outliers and more reliable inference from small samples, as the L-moments are a linear combination of data.

The main stages of the RFA procedure are: (i) screening of the data; (ii) identification of homogeneous regions; (iii) choice of a regional frequency distribution; (iv) estimation of the regional frequency distribution.

Hosking and Wallis (1997) recommend: the standard discordance measure of Wilks to be used for identifying unusual sites in a region in terms of the sample L-moments ratios; a heterogeneity measure, for assessing whether a proposed region is homogeneous; and a goodness-of-fit measure, for assessing whether a candidate distribution provides an adequate fit to the data. The RFA is an iterative procedure. However, Hosking and Wallis (1997), emphasize physical reasoning rather than formal statistical significance in data processing. Let us denote by D_i the Wilks' discordancy measure based on the sample L-moments ratios (L-CV,L-SKEW, L-KURT) for each site as Hosking and Wallis (1997) proposed to be used in the screening of data by discordant sites. A site is declared as discordant if its $D_i \geq 3$. It is well known that the standard measure of Wilks (which is in fact equal to the Mahalonobis distance up to a fixed constant) for detection of multivariate outliers is not robust as it is based on the sample mean and covariance matrix which are themselves affected by outliers. Alternatives to it based on robust estimates of multivariate location and scatter, developed by Rousseeuw and van Zomeren (1990) could be used instead. The Wilks' discordance measure D_i and several of its alternatives denoted by $RD_i(.)$ based on robust estimates of the location and scatter were applied in this study.

This paper is organized as follows. First the contents of the rivers data base will be presented. Some preliminary statistical results will be shown in Section 3, followed by the RFA-results. The main conclusions of the study are summarized in the Section 4. The last section contains a list of references.

2 DATABASE

From the GRDC (Global Runoff Data Centre, Koblenz, Germany) a 18MB database was received with daily river discharges of 213 locations over a period of time varying from location to location from 1 year to almost 200 years. Characteristics of the datasets such as station ID, river name, location name, country, latitude and longitude of the location, size of the basin area, starting year and month of the daily measurements, ending year and month of the daily measurements were given by the GRDC. Statistics of the annual average flows, annual maximum flows, annual minimum flows, and the mean annual precipitation, population density and elevation of the location, were obtained in the screening of the data and the analysis of various maps from BosAtlas (1997). The 213 stations are depicted in figure 1.

Figure 1: Distribution of the measurement stations over North-Western and Central Europe.

In figure 2, the daily river discharges of a river at a certain location can be seen. For every year the maximum, minimum and average discharges can be determined. The maximum values are of interest in a flood analysis.

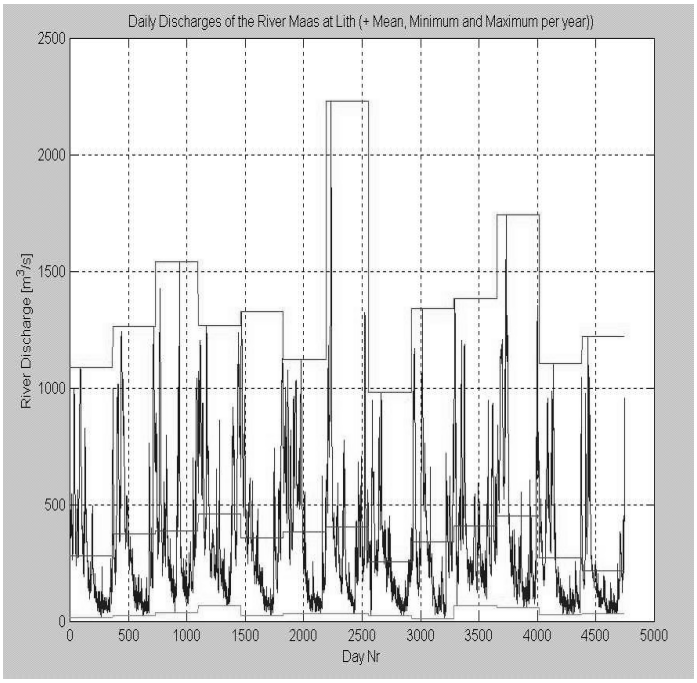

Figure 2: Station Lith along the river Maas.

In total data the annual maxima from 168 rivers were used in the study. Therefore from the 213 rivers, a total of 45 rivers could not be used because of too much missing data values or a too short period of time (< 15 years). In Van Gelder et al. (1999), the results are presented of the at-site frequency analysis of the 168 locations.

3 FORMATION OF REGIONS

A K-means clustering of site characteristics (latitude, longitude, elevation, annual precipitation, population density and the size of basin area) was performed following the algorithm of Hartigan and Wong (1979). Latitude, longitude and size of the basin area were available from the database. The other three characteristics were obtained from geographical maps. The following transformations of the site characteristics were used:

Size of basin area: $Y = 3 \log(X)/\sigma(\log X)$

Elevation: $Y=\sqrt{X} / \sigma(\sqrt{X})$

Latitude and Longitude: $Y=X / \sigma(X)$

Eight clusters were derived from the K-means clustering algorithm (figure 3).

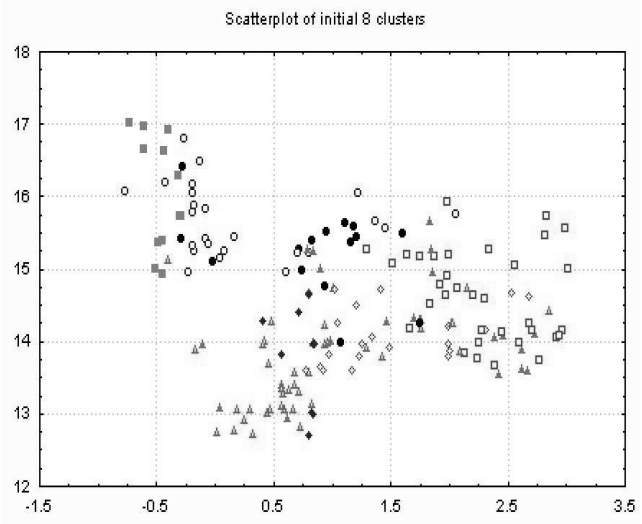

Figure 3: Results of the cluster analysis.

The at-site L-moment ratios are calculated for all sites at each cluster and the corresponding heterogeneity measure H was determined. Hosking and Wallis (1997) regard a region to be acceptably homogeneous if $H<1$, "possibly" heterogeneous if $1 \leq H<2$ and definitely heterogeneous if $H \geq 2$. If the region is not acceptably homogeneous, some redefinition of the region were considered, based on Wilks discordancy test and/or robust discordancy distances. The obtained regions based on the cluster analysis gave definitely heterogeneous regions according to the H-measure. Therefore, the regions were reduced by excluding the discordant sites (according to the robust distances, as the classical distance of Wilks in most of the cases was not significant). This was not unexpected as already shown in

the simulation results of Van Gelder and Neykov (1999). In this way, by several iterations of excluding discordant sites, the following homogeneous regions were derived:

Region 1: England and Northern Europe (containing 10 sites)

Region 4: South France (17 sites)

Region 5: Thames river with 6 other sites (7 sites)

Region 6: Western UK (11 sites)

Region 8: Rhine and Danube rivers (10 sites)

Region h33: Dutch and German sites (12 sites)

In total 67 sites (40% of all sites) are assigned to homogeneous regions.

The regions are shown in figure 4.

It is emphasized that without the use of the robust discordancy measures it would not have been possible to derive the homogeneous regions so easily. If H appeared to be "significant", i.e. H > 1, then it was observed that the robust distances of discordancy were also significant for some of the sites. Eight clusters appeared to be an optimal choice for the forming of homogeneous regions. Fewer clusters did not result in satisfiable regions. It is recommended that the forming of more clusters is investigated since this may result in more homogeneous regions, but on the other hand they may contain very few sites.

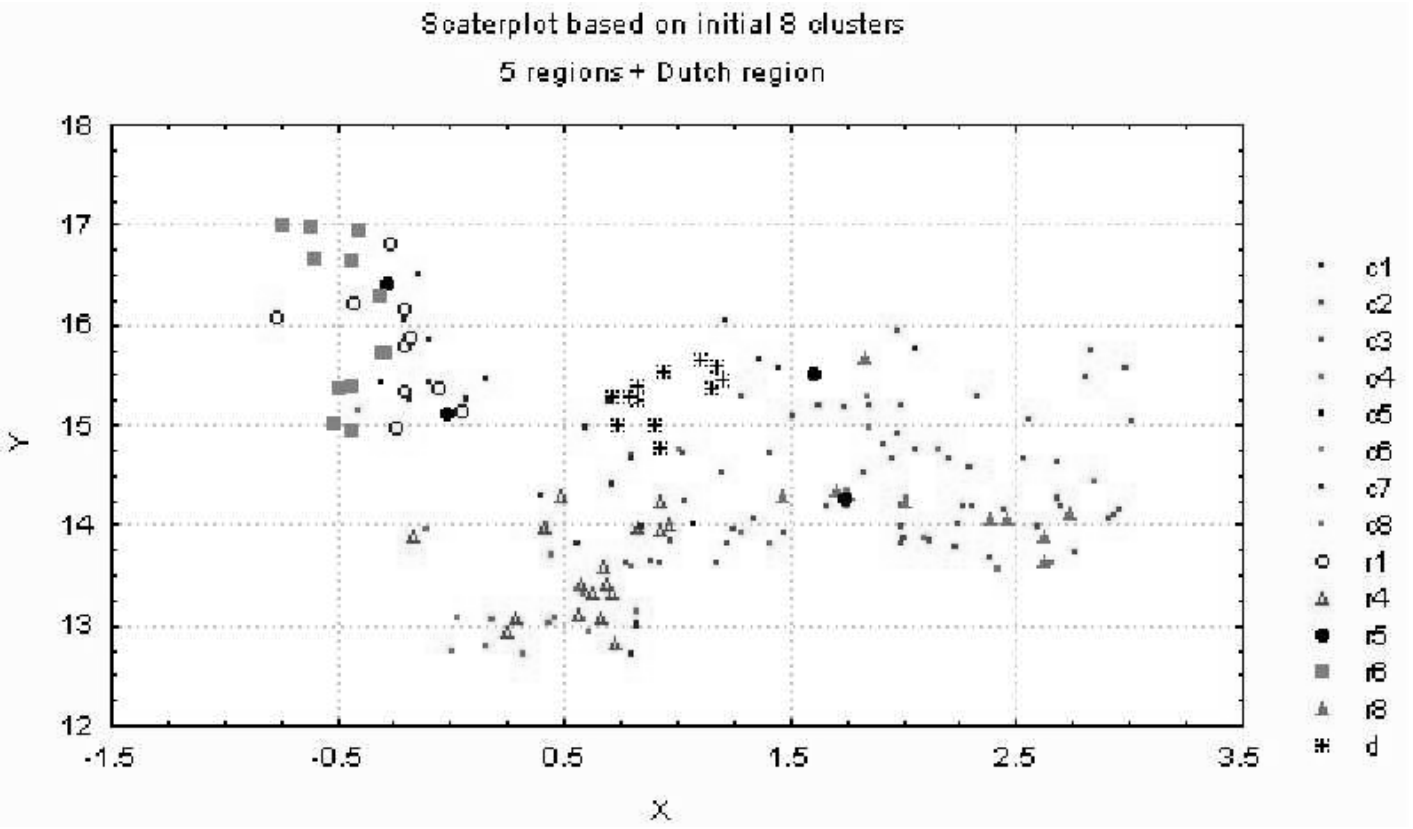

Figure 4: The 6 homogeneous regions

In this way the regional growth curves were estimated, followed by the estimations of the at-site quantiles. For every region also a goodness-of-fit statistic was calculated for the distributions. The results of these goodness-of-fits are presented in Van Gelder et al. (1999). For the regions with a "significant" correlated structure a modified algorithm of Hosking and Wallis (1997) for RFA was used for the purpose of obtaining more reliable estimates of the H-statistics and of the confidence bounds.

4 CONCLUSIONS

Extreme river discharges may have severe consequences for flood protection structures. How frequently the events of extreme river discharges may be expected to occur is of great importance. Design of civil engineering structures and insurance risk calculations, for instance, rely on knowledge of the frequency of these extreme events. Estimation of these frequencies is, however, difficult because extreme events are by definition rare and data records are often short. Regional frequency analysis resolves

this problem by 'trading space for time': it does so by using data from several sites, which are judged to have frequency distributions similar to the site of interest, in estimating event frequencies at that site. L-moments are a recent development within statistics (Hosking and Wallis, 1997). They form the basis of an elegant mathematical theory in their own right, and can be used to facilitate the estimation process in regional frequency analysis. L-moment methods are demonstrably superior to those that have been used previously, and are now being adopted by major organizations worldwide (Hosking and Wallis, 1996).

The largest possible homogeneous region that includes the Dutch rivers contains ten sites. These ten sites are given by 3 Dutch sites, 1 Belgian site and 6 German sites. Trying to add low-lying English and Polish sites led to a heterogeneous region.

It appeared that the number of datapoints appears to be a very sensitive factor in the homogenization process. Our approach was to reduce the datasets to 30 years in order to form homogeneous regions. As pointed out by Hosking and Wallis, datasets with more than 30-40 years can better be analyzed as at-site data. The RFA-method is only benificial for datasets up to 30-40 observations.

The assessment of the accuracy of the regionally estimated quantiles can be performed with Monte Carlo simulations. The simulation algorithm of Hosking and Wallis (1997) has been used for this purpose. Because of the (in some regions) high intersite dependence, its correlation matrices have been included in the algorithm. The proposed modification of Hosking and Wallis (1997) should only be applied in case of strongly correlated regions (average correlation coefficient around 70%).

ACKNOWLEDGEMENTS

The GRDC (Global Runoff Data Centre) of the Bundesanstalt für Gewässerkunde in Koblenz Germany is gratefully acknowledged for providing the datasets. N M Neykov and Pl N Neytchev gratefully acknowledge the financial support for this study provided by the Faculty of Civil Engineering, T.U. Delft, The Netherlands, and the Ministry of Water Management in Lelystad. Furthermore Valentin Todorov of Ericsson Software Design in Vienna Austria is also acknowledged for providing an efficient computer program for MCD-calculation.

REFERENCES

Rousseeuw,P. and B. van Zomeren,(1990) Unmasking Multivariate Outliers and Leverage Point (with discussion), Journal of the American Statistical Association, 85, pp.633 - 651).

Hosking, J.R.M. (1990) L-moments analysis and estimation of distributions using linear combination of order statistics. J. of the Royal Statistical Society, Series B, 52, 105-124.

Hosking, J.R.M. and Wallis, J.R. (1997). *Regional Frequency Analysis: An Approach based on L-Moments*. Cambridge University Press, Cambridge, UK.

Hosking, J. R. M. and J. R. Wallis (1996), The U.S. National Electronic Drought Atlas: Statistical Data Analysis with GIS-based presentation of results. In*: Proceedings of Conference on GIS and Statistics*, Korean Statistical Association, and Korean Association of GIS, Seoul, Korea, July 1996.

Hartigan, J.A. and Wong, M.A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, 28, 100-108.

P H A J M van Gelder, N M Neykov, Pl Neytchev, J K Vrijling, H Chbab (1999), Probability Distributions of Annual Maximum River Discharges in North-Western and Central Europe, TU Delft Report, August 1999.

P.H.A.J.M. van Gelder and N.M. Neykov (1999), ROBUST DISTANCES VERSUS THE WILKS DISCORDANCY MEASURE IN THE REGIONAL FREQUENCY ANALYSIS BASED ON L-MOMENTS OF FORMING HOMOGENEOUS REGIONS: A MONTE CARLO INVESTIGATION, European Geophysical Society, XXIV General Assembly, The Hague, The Netherlands, 19-23 April 1999.