Acta Geophysica (2023) 71:385–390
https://doi.org/10.1007/s11600-022-00852-1

SHORT COMMUNICATION - HYDROLOGY

A tale of two stations: a note on rejecting the Gumbel distribution

Jesper Rydén[1]

Received: 2 March 2022 / Accepted: 4 June 2022 / Published online: 27 July 2022

Abstract
The existence of an upper limit for extremes of quantities in the earth sciences, e.g. for river discharge or wind speed, is sometimes suggested. Estimated parameters in extreme-value distributions can assist in interpreting the behaviour of the system. Using simulation, this study investigated how sample size influences the results of statistical tests and related interpretations. Commonly used estimation techniques (maximum likelihood and probability-weighted moments) were employed in a case study; the results were applied in judging time series of annual maximum river flow from two stations on the same river, but with different lengths of observation records. The results revealed that sample size is crucial for determining the existence of an upper bound.

Keywords Extreme values · River discharge · GEV distribution · Sample size · Bootstrap

Introduction

Use of statistical models and methodology is of critical importance in hydrology. A proper statistical model is the basis for further applications, not least when studying extremes, such as estimation of return levels or examination of trends in hydrological variables under investigation. With current access to suitable software, model selection is of interest. On the one hand, this involves selection of a probability distribution. For a recent discussion of modelling extreme river discharges with typical distributions, see Rydén (2022a). On the other hand, it involves considering the methodological basis for choices. Using a maximum-likelihood (ML) based framework, metrics such as Akaike's Information Criterion (AIC) is an option. However, for other estimation techniques (e.g. methods of moments), alternative approaches must be used. In addition, careful examination of the fitted distributions with respect to parameter estimates and resulting related return levels could be good practice (Rydén 2019).

A key factor in statistical analysis of almost any kind is the sample size. Small samples are challenging, and ML approaches may be less suitable for extremes in certain situations (Coles and Dixon 1999). Hosking et al. (1985) present comparisons (theoretical and numerical examples) of methodology for estimation of parameters in the Generalised Extreme Value (GEV) distribution. For the practitioner, the question arises as to whether there are statistical implications of using so-called small samples in estimation of a GEV distribution. This question is one of the subjects of the present analysis.

In fact, properties of the parameters in the GEV distribution can be linked to behaviour in nature. If the so-called shape parameter is negative, there exists (in theory) a finite right end point. It may be of interest from an applied point of view, and for understanding the hydrological system, to assess this parameter. For example, Roden (1967) found that the dimensions of a drainage basin determines the upper limit of discharge. However, negative and positive estimates of the shape parameter may still occur, e.g. Hosking et al. (1985) give examples, based on series of annual maxima. Hence, the estimated parameters, which depend upon the sample and more precisely sample size, have implications for interpretations of the system.

The present study extended the analysis in Hosking et al. (1985) through investigating by simulation studies the percentage of rejections of a Gumbel distribution (a GEV distribution with the shape parameter equal to zero) for various

Edited by Dr. Michael Nones (CO-EDITOR-IN-CHIEF).

✉ Jesper Rydén
jesper.ryden@slu.se

[1] Swedish University of Agricultural Sciences, Uppsala, Sweden

sample sizes and chosen values of the shape parameter. As a case study and illustration, two time series of annual maximum flows from an unregulated river in northern Sweden, were analysed. These time series actually have the same estimated shape parameter, but the decision to reject a Gumbel distribution differs, due to a substantial difference in lengths of the series.

The remainder of this paper is organised as follows: Sect. "Material and methods" provides a brief review of statistical extreme-value analysis and in particular inference with the GEV distribution. Moreover, the data from the two stations are described. Section "Result" outlines the simulation study, and presents the results which are then evaluated in relation to the estimation results for the two stations. Moreover, these results are discussed in relation to the estimations for the two stations. A concluding discussion is given in Sect. "Summary and discussion".

Material and methods

The GEV distribution

Briefly, extreme-value analysis can be said to concern the tails of distributions. A conventional approach, with origins in Gumbel (1958), is to fit a generalised extreme-value (GEV) distribution to a sample of independent annual maxima. This is the limiting distribution of independent maxima. The distribution function for the GEV distribution is given as:

$$\mathrm{P}(X \le x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \qquad (1)$$

defined on $\{x:\ 1+\xi(x-\mu)/\sigma > 0\}$ and where $\mu, \sigma > 0$ and ξ are the location, scale and shape parameters respectively. The location parameter μ is related to the centre of the distribution, while the scale parameter σ describes deviations. The shape parameter ξ is related to the nature of the tail: if $\xi < 0$, the upper tail is bounded; if $\xi = 0$ the tail decays exponentially (the case of the so-called Gumbel distribution); if $\xi > 0$, the tail decays as a power function.

Estimation techniques

When fitting a conventional GEV distribution to data, estimation is usually performed using the ML method. An implementation in R (R Core Team 2021) is provided in the package `extRemes` (Gilleland and Katz 2016).

A previous comparison of methodologies for fitting parameters in the GEV distributions by Hosking et al. (1985) covered probability-weighted moments (PWM) in addition to ML. In simulation studies, those authors found bias and standard deviation for settings of sample sizes and values of the shape parameter ξ. Estimation of quantiles was also considered (Hosking et al. 1985). Concerning parameter estimation, ML estimators were shown to be the least biased, but more variable than the PWM estimator in small samples. In the present study, estimation procedures as implemented in the R package `lmom` (Hosking 2019) were used.

For a review of estimation techniques for statistical extremes, see Coles and Dixon (1999).

Remark: Note that in general, deriving large-sample asymptotics of the ML estimator for a distribution family with varying support is a difficult problem. For a recent treatment of this topic, see Bucher and Segers (2017).

Testing hypotheses

According to Hosking et al. (1985): "It is often useful to test whether a given set of data is generated by a Gumbel rather than a GEV distribution". A simulation study is described in Sect. "Result", while the test methodology is briefly presented below. In a nutshell, one tests the null hypothesis $\xi = 0$ against some alternative hypothesis (typically $\xi \neq 0$). We employed two approaches for the inference: Wald tests and bootstrap.

Wald tests:

When performing hypothesis testing for the shape parameter ξ from an ML estimation, so-called Wald tests can be employed. For instance, to test the null hypothesis $\xi = 0$, the test quantity $\hat{\xi}/\text{s.e.}\ (\hat{\xi})$ is computed, where s.e. $(\hat{\xi})$ is the standard error of the estimated parameter. Computation of p-values follows from asymptotic normality, when $\xi > -0.5$. When $-1 < \xi \le -0.5$, the ML estimate exists, but does not have the standard asymptotic properties (Smith 1985). For situations investigated in this note is chosen $\xi > -0.5$, which is common in practice (Dey et al. 2016).

For estimation using probability-weighted moments, let $\widetilde{\xi}$ be the point estimate. Following Hosking et al. (1985), the test quantity $\widetilde{\xi}(n/0.5633)^{1/2}$ belongs to a standard normal distribution, and conventional statistical inference procedures follow (p-values etc.).

Bootstrap:

In a recent article, Gilleland (2020) finds bootstrap methods appealing in an extreme-value context and gives several numerical illustrations. Simulation studies have been carried out by Caires (2007); for instance, confidence intervals for return levels based on a fitted GEV distribution were computed, investigating various sample sizes as well as bootstrap sample size.

In this paper, we will use adjusted percentile bootstrap intervals, advocated by Caires (2007) and originally proposed by Coles and Simiu (2003). These are obtained by taking the quantile of probability 0.025 of the empirical

Table 1 Descriptions of the two stations

Station	Name	River ID	River	Area (km^2)	Start	End
2357	Abisko	1000	Torneträsk	3345.5	1985	2019
16722	Kukkolankoski övre	1000	Torne	33929.6	1911	2019

distribution of the sample of bootstrap estimates as the lower limit and the quantile of probability 0.975 as the upper limit.

More precisely, from a random sample $x = \{x_i, i = 1, \ldots, n\}$, estimate the shape parameter by an estimator $\xi = \hat{\xi}(x)$. A bootstrap samples x^* is obtained by resampling; by randomly sampling n times, with replacement, from the original sample x. Now, let B be a large integer. In all, B bootstrap samples $x_b^*, b = 1, \ldots, B$, are created and a set of estimates $\hat{\xi}_b^*$ are computed. The adjusted bootstrap estimates, following Coles and Simiu (2003), are given by:

$$\hat{\xi}_b^{*a} = \hat{\xi}_b^* - \frac{1}{B}\sum_{b=1}^{B}\hat{\xi}_b^* + \hat{\xi}, \quad b = 1, \ldots, B. \tag{2}$$

The quality of the coverage of the intervals does not considerably depend on the bootstrap sample size B (Caires 2007).

In the simulation study performed in this paper, the resulting confidence interval based on Eq. 2 is used for testing: If zero is not found within the interval, the null hypothesis $\xi = 0$ is rejected. The bootstrap algorithm as implemented in the routine `boot` in the R package with the same name was employed (Canty and Ripley 2021; Davison and Hinkley 1997).

Data sources

As an illustrative case, this study considered data on annual maximum flow or, more precisely, on the original time series of daily observations of flow (m^3/s) and annual maxima were extracted from these series. The data originated from two stations on the river Torne, an unregulated river in the far north of Sweden. A study of extreme flows at multiple stations in this region was performed by Rydén (2022b), with the focus on investigation of possible trends in the magnitude and timing of extreme floods in northern Sweden. The two stations considered in the present study were selected since they happen to have the same ML estimate of the shape parameter.

Data are provided online by the Swedish Meteorological and Hydrological Institute (SMHI): http://vattenwebb.smhi.se/station/# . Table 1 provides details of the two stations investigated. A map showing the locations of the two stations is presented in Fig. 1, and the corresponding time series are shown in Fig. 2.

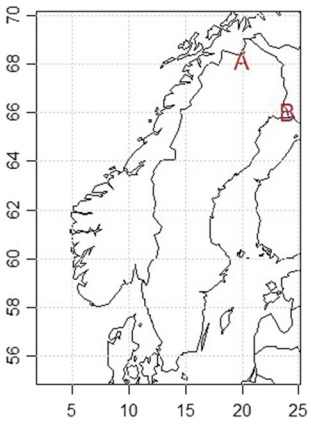

Fig. 1 Map showing locations of the two stations: **A** Station 2357, **B** Station 16722

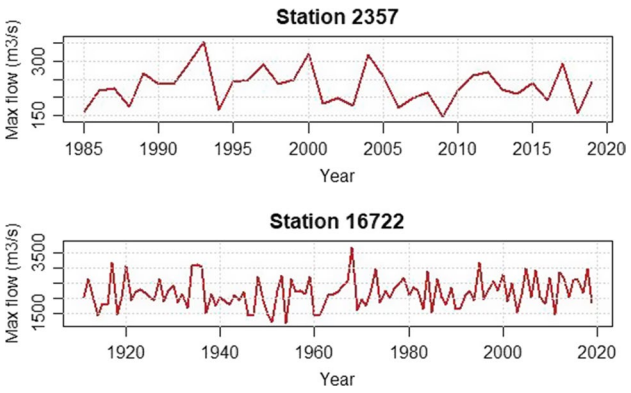

Fig. 2 Time series of annual maximum flow for: Station 2357 (upper panel) and Station 16722 (lower pannel)

Results

Details of the simulation studies, which resulted in plots for assessment, are provided in this section. Furthermore, estimated shape parameters and related tests for the two stations are then considered made with these plots in mind.

Simulation study

The algorithm is outlined as follows:

1: Initiate values of shape parameter ξ and the sample size n.

2: Simulate from a GEV according to Step **1**. Fit parameters by ML and by PWM, respectively.

3: Perform test for the null hypothesis $\xi = 0$ against $\xi \neq 0$ for each of the two estimates (cf. Sect. "Estimation techniques") and test options (Sect. "Testing hypotheses"). The conventional choice of significance level 0.05 is used here. Record whether the null hypothesis is rejected.

4: Repeat steps **2**-**3** a large number of times (2 000) and record, finally, the proportion of rejections for each case.

In step **1**, the values of ξ varies between -0.02 and -0.16 in 0.02 increments. The scale parameter was set here as $\mu = 0$ and the shape parameter as $\sigma = 1$, without loss of generality (the same argument as made by Hosking et al. (1985)). The sample sizes were from the set $\{20, 40, 60, 80, 100, 150, 200, 250, 300\}$. In Step **3**, for the bootstrap analysis, $B = 200$ was chosen.

The results from ML estimation (Wald test for inference) are shown in Fig. 3. The interpretation is straightforward: with increasing sample size, the proportion of rejection of the null hypothesis $\xi = 0$ increases. The smaller the shape parameter, the more evident is this feature. This figure, and the related findings, may be of interest to a practitioner, facing a situation with a specific estimated shape parameter for the sample size at hand. For instance, for a sample size of 75, the difference in rejection behaviour between $\xi = -0.1$ and $\xi = -0.16$ is notable.

The differences in overall rejection behaviour were minor when instead employing PWM, so the corresponding figure showing proportions of rejections versus sample size is omitted. However, for each sample size, differences in proportions of rejections between the estimation methodologies are shown in Fig. 4. Note that an interesting peak arises as ξ moves away from zero. With decreasing values of ξ, the difference becomes more pronounced and the peak moves to smaller sample sizes.

Finally in this section, we illustrate the choice of test procedure (Wald test or bootstrap). In Fig. 5, we present in the left panel results based on simulations from a GEV distribution with $\xi = -0.16$; in the right panel, the corresponding results when $\xi = -0.02$. In the first situation, it is desirable that the null hypothesis $\xi = 0$ is rejected, at least for larger samples. We note that ML estimation with a Wald test works best, and for samples larger than 100, the bootstrap inference based on the ML estimate renders similar, or slightly better, results. In the second situation (right panel), with a shape parameter closer to zero, the inference by bootstrap has a stronger tendency to reject, at least for sample sizes larger than 50, which could be seen as an advantage.

A tale of two stations

In the case study of two stations along the Torne river (see Sect. "Data sources"), GEV distributions were fitted with the ML method, and p-values were calculated (two-sided alternative hypothesis, $\xi \neq 0$). The results are shown in Table 2. Note that from a practical point of view, the estimated values of ξ are identical (-0.16), but, based on the conventional significance level of 0.05, the p-values yield different conclusions: a clear rejection of $\xi = 0$ for Station 16722, but not at all for Station 2357.

The reason might be the substantial difference in time series length. Turning to Fig. 3 and following the curve corresponding to $\xi = -0.16$, a sharp decrease in proportion

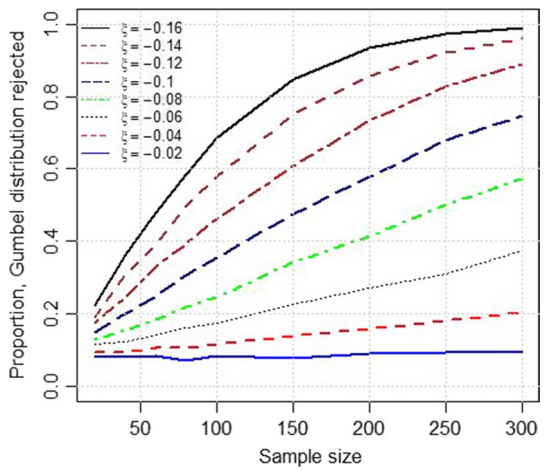

Fig. 3 Proportion of rejections of the Gumbel distribution with increasing sample size (ML estimation)

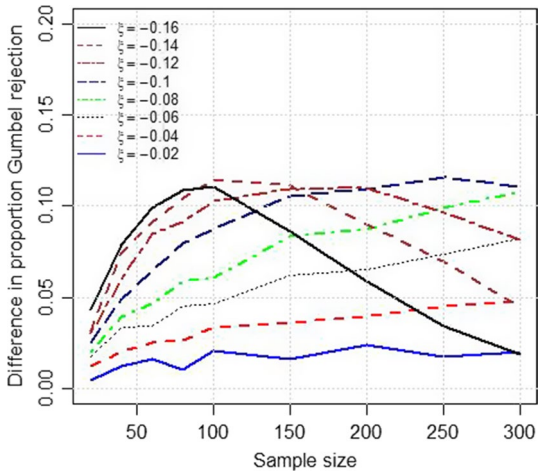

Fig. 4 Difference in proportion of Gumbel rejections by ML and PWM estimation, as a function of sample size

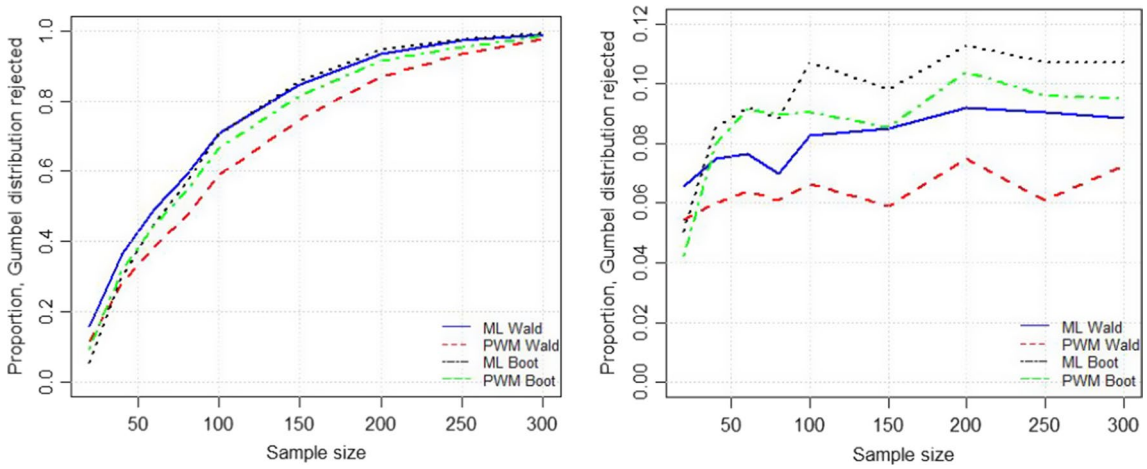

Fig. 5 Simulations studies, comparison of test strategies (Wald or bootstrap): left panel: $\xi = -0.16$; right panel: $\xi = -0.02$

Table 2 Results of ML estimation, for the two case study stations

Station	Period	$\hat{\xi}$	p-value
2357	1985-2019	−0.158	0.24
16722	1911-2019	−0.159	0.018

of rejections of the Gumbel distribution can be seen with decreasing sample size.

Summary and discussion

One aim of this article was to more closely examine possible links between statistical distributions for extremes and their interpretations in earth sciences. The case example concerned river flows, where a negative value of the shape parameter in a GEV distribution implies that an upper bound exists. A previous study examining maximum discharge for the Rhine at Lobith concluded that "there is in fact an upper limit to the discharge at Lobith" (de Vriend et al. 2017). Thus, the question of an upper bound is non-trivial. Issues concerning the existence of an upper limit and its relation to statistical modelling have sometimes even resulted in controversies in the research literature, see, for example, Harris (2005) and Simiu (2007) for a dispute on modelling extreme wind speeds.

The study by Hosking et al. (1985) was extended in the present analysis, which by simulation studies examined the influence of sample size on the tests for a possible Gumbel distribution. In both estimation strategies employed in this paper (ML, PWM), asymptotic normality is the assumption for the Wald tests. In practice, it is not easy to determine when this is attained, and there is no rules of thumbs to rely upon. The bootstrap-based inference seems a plausible choice, cf. the discussion related to Fig. 5.

In the simulation studies, the lowest sample size was set at 20. In the type of application discussed here, there are usually at least some decades of data (i.e. annual maxima) available. In this note, we studied merely inference of the shape parameter, ξ. Cai and Hames (2010) discuss the influence of sample size for estimation of return levels (employing a bootstrap approach).

Another asymptotic argument is to determine the limiting distribution of the extremes. Dey et al. (2016) performed simulation studies for some selected distributions (Pareto, Gamma, Normal, Beta) and investigated the rate of convergence for various sample sets. They used the Kolmogorov-Smirnov goodness-of-fit test to assess the proximity of distributions and empirical rejection percentages observed. They found that heavy-tailed distributions converge faster. Similar studies with typical distribution settings and parameter choices from the earth sciences might be of interest. According to Dey et al. (2016), compared with other sources on extreme-value analysis their paper "provides more statistical flavour through numerical studies". Future work oriented towards applications in, for example, hydrology would thus be appropriate.

Acknowledgements The author is grateful for the comments provided by two anonymous reviewers.

Funding Open access funding provided by Swedish University of Agricultural Sciences.

Declarations

Conflict of interest The author of the manuscript declares no conflict of interest.

References

Bucher A, Segers J (2017) On the maximum likelihood estimator for the generalized extreme-value distribution. Extremes 20:839–872. https://doi.org/10.1007/s10687-017-0292-6

Cai Y, Hames D (2010) Minimum sample size determination for generalized extreme value distribution. Commun Stat Simul Comput 40(1):87–98. https://doi.org/10.1080/03610918.2010.530368

Caires S (2007) Extreme wave statistics: confidence intervals. Report prepared for Rijkswaterstaat, Rijksinstituut voor Kust en Zee (RIKZ). http://resolver.tudelft.nl/uuid:8d38ef9c-ead4-4b9d-850c-d4dd2e71a34f

Canty A, Ripley B (2021) boot: Bootstrap R (S-Plus) Functions. R package version 1.3-28

Coles SG, Dixon MJ (1999) Likelihood-based inference for extreme value models. Extremes 2:5–23. https://doi.org/10.1023/A:1009905222644

Coles S, Simiu E (2003) Estimating uncertainty in the extreme value analysis of data generated by a hurricane simulation model. J Engrg Mech 129:1288–1294. https://doi.org/10.1061/(ASCE)0733-9399(2003)129:11(1288)

Davison AC, Hinkley DV (1997) Bootstrap methods and their applications. Cambridge University Press, Cambridge. 0-521-57391-2. https://doi.org/10.1017/CBO9780511802843

Dey D, Roy D, Yan J (2016) Univariate extreme value analysis. In extreme value modeling and risk analysis. Methods and applications. CRC Press, Chapman & Hall, Boca Raton. https://doi.org/10.1201/b19721

Gilleland E (2020) Bootstrap methods for statistical inference. Part II: extreme-value analysis. J Atmos Ocean Technol 37:2135–2144. https://doi.org/10.1175/JTECH-D-20-0070.1

Gilleland E, Katz, RW (2016) extRemes 2.0: An extreme value analysis package in r. journal of statistical software 72(8), 1–39. https://doi.org/10.18637/jss.v072.i08

Gumbel EJ (1958) Statistics of extremes. Columbia University Press, New York. https://doi.org/10.7312/gumb92958

Harris I (2005) Generalised Pareto methods for wind extremes. Useful tool or mathematical mirage? J Wind Eng Ind Aerodyn 93(5):341-360. https://doi.org/10.1016/j.jweia.2005.02.004

Hosking JRM (2019) L-Moments. R package, version 2.8. https://CRAN.R-project.org/package=lmom

Hosking JRM, Wallis JR, Wood EF (1985) Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics 27(3):251–261. https://doi.org/10.2307/1269706

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Roden GI (1967) On river discharge into the northeastern Pacific Ocean and the Bering Sea. J Geophys Res 72(22):5613–5629. https://doi.org/10.1029/JZ072i022p05613

Rydén J (2019) A note on analysis of extreme minimum temperatures with the GAMLSS framework. Acta Geophys 67:1599–1604. https://doi.org/10.1007/s11600-019-00363-6

Rydén J (2022a) Tales of the Wakeby tail and alternatives when modelling extreme floods. REVSTAT—Statistical Journal. https://revstat.ine.pt/index.php/REVSTAT/article/view/454 (accepted)

Rydén J (2022b) Statistical analysis of possible trends for extreme floods in northern Sweden. River Res Appl. https://doi.org/10.1002/rra.3980 (accepted)

Simiu E (2007) Discussion: Generalized Pareto methods for wind extremes. Useful tool or mathematical mirage? by Ian Harris. J Wind Eng Ind Aerodyn 95(2):133-136. https://doi.org/10.1016/j.jweia.2006.05.002

Smith RL (1985) Maximum likelihood estimation in a class of nonregular cases. Biometrika 72:67–90. https://doi.org/10.2307/2336336

de Vriend HJ, Kok M, Pol J, Hegnauer M (2017) Is there a maximum discharge for the Rhine at Lobith? A publication of the Expertisenetwerk Waterveiligheid. URL: https://www.enwinfo.nl/publish/pages/183541/is-there-a-maximum-discharge-for-the-rhine-at-lobith-march2017.pdf