Climate Variability and Change—Hydrological Impacts (Proceedings of the Fifth FRIEND World Conference held at Havana, Cuba, November 2006), IAHS Publ. 308, 2006.

Bayesian priors based on regional information: application to regional flood frequency analysis

MATHIEU RIBATET[1,2], ERIC SAUQUET[2], JEAN-MICHEL GRESILLON[2] & TAHA B. M. J. OUARDA[1]

1 *INRS-ETE, University of Quebec, 490 de la Couronne, Quebec G1K 9A9, Canada*
mathieu.ribatet@ete.inrs.ca

2 *Cemagref, 3 bis quai Chauveau CP 220, F-69336 Lyon Cedex 9, France*

Abstract Flood frequency analysis is usually based on the fitting of an extreme value distribution to the series of local streamflow. However, when the at-site time series is short, frequency analysis results become unreliable. In this work, a regional Bayesian model to estimate flood quantile from a few years of stream flow data is proposed. This model is less restrictive than the Index Flood model while preserving the formalism of "homogeneous regions". Performance of the proposed model is assessed on a set of French gauging stations. The accuracy of quantile estimates as a function of homogeneity level of the pooling group is also analysed. Results indicate that the regional Bayesian model outperforms the Index Flood model and local estimators. Furthermore, it seems that working with relatively large and homogeneous regions may lead to more accurate results than working with smaller and highly homogeneous regions.

Key words regional frequency analysis; Bayesian inference; Index Flood; MCMC

INTRODUCTION

Flood frequency analysis is essential in preliminary studies to design flood defence structures. Methods for estimating design flows usually consist of fitting one of the distributions given by the extreme value theory to a sample of flood events. If modelling exceedence over a threshold is of interest, a theoretical justification exists for the use of the Generalized Pareto distribution (GP):

$$F(x) = 1 - \left[1 + \frac{\zeta(x-\mu)}{\sigma}\right]^{-1/\zeta} \text{ where } 1 + \frac{\zeta(x-\mu)}{\sigma} > 0, \sigma > 0 \quad (1)$$

where μ, σ and ζ are the location, scale and shape parameters, respectively. This distribution is defined for $\zeta \neq 0$, and can be derived by continuity in the case $\zeta = 0$, corresponding to the exponential case.

However, frequency analysis can lead to unreliable flood quantiles when few data are available at the site of interest. A convenient way to improve estimates of flood statistics is to incorporate data from other gauged locations in the estimation procedures. This approach is widely applied in hydrology and is known as Regional Flood Frequency Analysis (RFFA).

One of the most popular and simple approaches favoured by engineers is the Index Flood method (Dalrymple, 1960). However, the assumptions of the Index Flood model need often to be relaxed to suit the observations. We suggest here to carry out a Bayesian approach that encompasses the classical Index Flood model and uses the whole data set in a more efficient manner.

The main goal of this paper is to test the efficiency and robustness of the developed regional Bayesian model when dealing with series of short record length. For this purpose, the suggested regional Bayesian approach will be compared to local analysis and traditional RFFA. The next section presents a brief summary of the classical Index Flood model. Next, the data set used to illustrate the method is described, followed by a description of the procedure used to elicit the prior distribution. In the fifth section the weaknesses and strengths of each approach on a typical homogeneous region are outlined. Finally an analysis of the effect of homogeneity level on quantile estimation is presented.

THE INDEX FLOOD MODEL

The Index Flood method states that flood frequency distributions within a particular region are supposed to be identical when divided by a scale factor – namely the Index Flood. Mathematically,

this assumption is expressed as:

$$Q^{(S)} = C^{(S)} Q^{(R)} \tag{2}$$

where $Q^{(S)}$ is the quantile function at site S, $C^{(S)}$ is the Index Flood at site S and $Q^{(R)}$ is the regional quantile function.

Equation (2) is supposed to be satisfied if all sites are hydrologically and/or statistically similar. Therefore, one of the main aspects of this approach is to identify a homogeneous region, which includes the target site.

The parameters of the regional distribution are usually derived from weighted average of at-site L-moments. Finally, the target site distribution $Q^{(S)}$ is computed from equation (2). It can be seen that the observations of all samples have the same weight. This is debatable since the most relevant information is certainly the at-site one. Thus, in this approach, the available information is not efficiently used.

DATA DESCRIPTION

The selection of the gauging sites was initially based on the 22 regions into which France is divided for the implementation of the European Water Framework Directive. Therefore it seems reasonable to consider this division as a preliminary guide for pooling stations. The pre-regions were subsequently altered to satisfy the heterogeneity test of Hosking & Wallis (1997). Finally, a set of 14 stations was selected for this study.

The record length of time series ranges from a minimum of 22 years to a maximum of 37 years, with a mean value of 32 years. The drainage areas vary from 32 to 792 km^2. Most of the gauging stations monitored first-order streams. Threshold levels were selected to extract on average around two events per year, while meeting the criteria of independence between floods.

Three stations (U4505010, U4635010 and V3015010) were of particular interest because of their extended record length of 37 years. In this case study, the scale factor was set to correspond to the 1-year return flood quantile. Analysing the influence of Index Flood selection is beyond the scope of this work. The main point is to keep the same Index Flood throughout the case study to compare approaches on the same basis.

ELICITING THE PRIOR DISTRIBUTION

In the proposed regional Bayesian model, the regional information is not used to build a regional distribution but to specify a kind of "suspicion'' about the target site distribution. This is easily achieved in the Bayesian framework through the so-called prior distribution. The prior model is usually a multivariate distribution, which must represent beliefs about the distribution of the parameters, i.e. μ, σ and ζ prior to having any information about the data. Consider all sites of a homogeneous region except the target site – say the jth site. A set of "pseudo target site estimates" is computed:

$$\tilde{\mu}^{(i)} = C^{(j)} \mu_*^{(i)} \tag{3}$$

$$\tilde{\sigma}^{(i)} = C^{(j)} \sigma_*^{(i)} \tag{4}$$

$$\tilde{\zeta}^{(i)} = \zeta^{(i)} \tag{5}$$

for all $i \neq j$ where $\mu_*^{(i)}$ and $\sigma_*^{(i)}$ are parameters estimates from rescaled sample. According to the Index Flood model, the set of parameters $(\tilde{\mu}^{(i)}, \tilde{\sigma}^{(i)}, \tilde{\zeta}^{(i)})$ for $i \neq j$ is expected to be distributed as $(\mu^{(i)}, \sigma^{(i)}, \zeta^{(i)})$. Note that, information from the target site sample is not used to elicit the prior distribution. Thus, $C^{(j)}$ in equations (3) and (4) must be estimated without use of the jth site sample. In this case study, $\log C^{(j)}$ is estimated through a Generalized Linear Model defined by:

$$\begin{cases} E\left[\log C^{(j)}\right] = \upsilon \\ \mathrm{var}\left[\log C^{(j)}\right] = \phi V(\upsilon) \end{cases} \qquad \eta = \log(\upsilon) = X\beta \tag{6}$$

where X are basin characteristics, ϕ is the dispersion parameter and V the variance function. As $C^{(j)}$ is estimated without the target site sample, it is important to incorporate uncertainties from the elicitation of the prior distribution. Under the independence assumptions between $C^{(j)}$ and $\mu_*^{(i)}$, $\sigma_*^{(i)}$, the following relations hold:

$$\mathrm{var}\left[\log \tilde{\mu}^{(i)}\right] = \mathrm{var}\left[\log\left(C^{(j)}\mu_*^{(i)}\right)\right] = \mathrm{var}\left[\log C^{(j)}\right] + \mathrm{var}\left[\log \mu_*^{(i)}\right] \quad (7)$$

$$\mathrm{var}\left[\log \tilde{\sigma}^{(i)}\right] = \mathrm{var}\left[\log\left(C^{(j)}\sigma_*^{(i)}\right)\right] = \mathrm{var}\left[\log C^{(j)}\right] + \mathrm{var}\left[\log \sigma_*^{(i)}\right] \quad (8)$$

The independence assumptions between $C^{(j)}$ and $\mu_*^{(i)}$, $\sigma_*^{(i)}$ is not too restrictive as the target site Index Flood $C^{(j)}$ is estimated independently from $\mu_*^{(i)}$, $\sigma_*^{(i)}$.

In this case study, the marginal prior distributions were supposed to be independent lognormal for both the location parameter μ and the scale parameter σ and normal for the shape parameter ζ. The lognormal distribution is justified by a physical and theoretical lower bound. Indeed, (a) discharge data are naturally non-negative; so the location parameter should also be non-negative; and (b) the scale parameter is strictly positive by definition of the GP distribution. Furthermore, as the prior distribution is based on the Index Flood model, equation (2) implies that if the scale parameter is lognormally distributed, the location parameter should also be lognormally distributed. This choice is confirmed by the data.

The prior distribution $\pi(\theta)$ is therefore elicited by means of the following equations:

$$E\left[\log \mu^{(j)}\right] \approx \frac{1}{N-1}\sum_{i\neq j} \log \tilde{\mu}^{(i)} \qquad \mathrm{var}\left[\log \mu^{(j)}\right] \approx \frac{1}{N-1}\sum_{i\neq j} \mathrm{var}\left[\log \tilde{\mu}^{(i)}\right]$$

$$E\left[\log \sigma^{(j)}\right] \approx \frac{1}{N-1}\sum_{i\neq j} \log \tilde{\sigma}^{(i)} \qquad \mathrm{var}\left[\log \sigma^{(j)}\right] \approx \frac{1}{N-1}\sum_{i\neq j} \mathrm{var}\left[\log \tilde{\sigma}^{(i)}\right]$$

$$E\left[\zeta^{(j)}\right] \approx \frac{1}{N-1}\sum_{i\neq j} \zeta^{(i)} = \bar{\zeta} \qquad \mathrm{var}\left[\zeta^{(j)}\right] \approx \frac{1}{N-2}\sum_{i\neq j} \left(\zeta^{(i)} - \bar{\zeta}\right)^2$$

where N is the number of stations within the homogeneous region. Thus:

$$\pi(\theta) \propto J \exp\left[-\frac{1}{2}\left(\theta' - \gamma\right)^T \Sigma^{-1}\left(\theta' - \gamma\right)\right] \quad (9)$$

where γ, Σ are hyper-parameters, $\theta' = (\log\mu, \log\sigma, \zeta)$ and J the Jacobian of the transformation from θ' to θ, namely $J = 1/\mu\sigma$.

Thus, the posterior distribution $\pi(\theta|x)$ is given by the Bayes Theorem:

$$\pi(\theta \mid x) = \frac{\pi(\theta)\pi(\theta; x)}{\int_\Theta \pi(\theta)\pi(\theta; x)\mathrm{d}\theta} \propto \pi(\theta)\pi(\theta; x) \quad (10)$$

where Θ is the space parameter, $\theta = (\mu, \sigma, \zeta)$; $\pi(\theta; x)$ is the likelihood of the GP distribution; and x is the at-site sample. In theory, the posterior distribution is entirely known but is often insolvable. To overcome this problem, the most convenient way is to implement MCMC techniques to sample the posterior distribution.

PERFORMANCE OF THE REGIONAL BAYESIAN POT MODEL ON A HOMOGENEOUS REGION

In this section, three different models are applied. For this purpose, the three stations – U4505010, U4635010 and V3015010 – were selected to assess the robustness and efficiency of the local, regional and Bayesian regional models. These three different approaches correspond to: (a) local: fit the GP distribution to the peaks-over-threshold (POT) data with the maximum likelihood estimator (MLE), unbiased and biased probability weighted moments (PWU, PWB); (b) regional (REG): fit the target site distribution as described in the Index Flood model section; and

Table 1 Benchmark values for 2, 5, 10 and 20 years quantiles and the associated 90% profile likelihood confidence intervals in bracket.

Station	Q_2	Q_5	Q_{10}	Q_{20}
U4505010	10.8 (10.1, 1.7)	15.3 (13.9,17.4)	19.5 (17.2,23.4)	24.4 (20.6,31.5)
U4635010	33.0 (30.0,36.5)	52.2 (45.5,62.5)	72.2 (60.2,95.4)	98.9 (69.2,200.5)
V3015010	7.5 (6.9, 8.3)	11.7 (10.4,13.7)	15.9 (13.6,19.9)	21.3 (17.3,28.8)

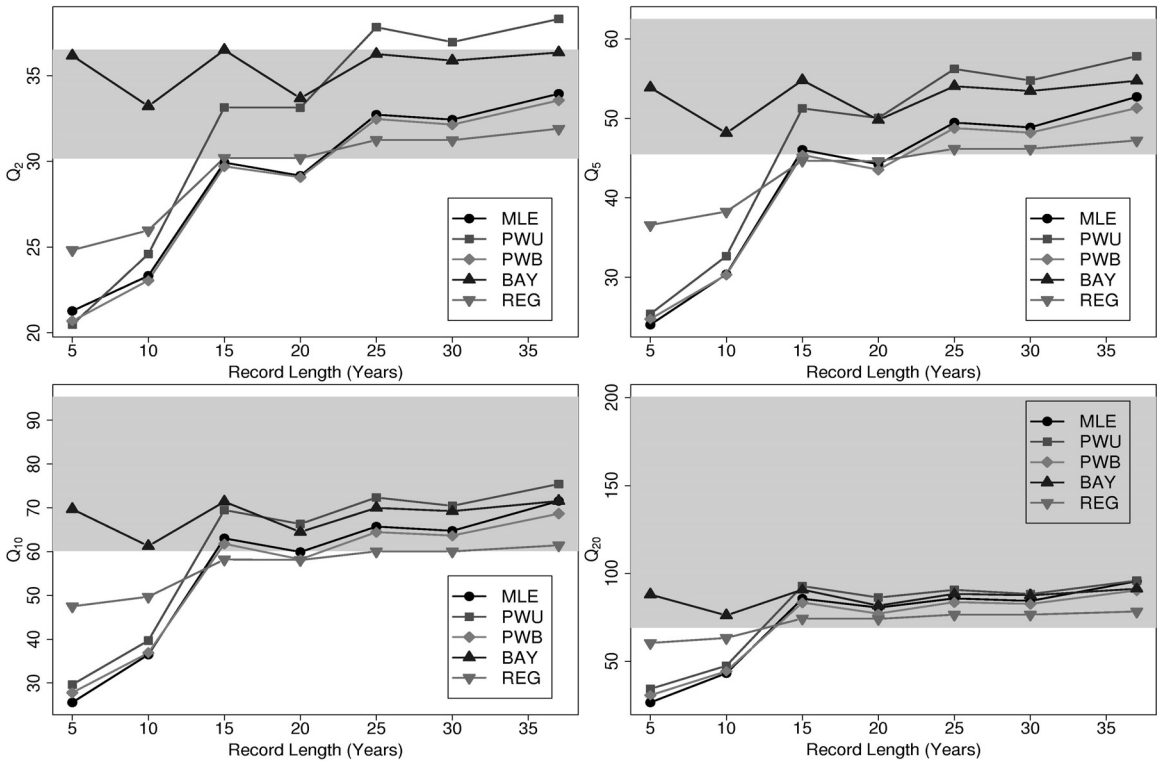

Fig. 1 Evolution of Q_2, Q_5, Q_{10}, Q_{20} estimates as the size increases for the site U4635010 and 90% profile likelihood confidence interval for the benchmark values – grey area.

(c) regional Bayesian (BAY): compute the target site distribution as described in the Eliciting the Prior Distribution section.

As the main goal of this work is to compare models on small samples, efficiency is evaluated on sub-samples from the original data. The MLE on the whole sample is used as a benchmark to assess the performance of each model. For this purpose, the target site sample was truncated to obtain shortened periods of records of m years, $m \in \{5,10,15,20,25,30,37\}$. Quantile estimates corresponding to return period 2, 5, 10 and 20 years are selected – cf. Table 1. Quantiles with return periods greater than 20 years are considered unreliable, as uncertainties on these quantiles are too large with only 37 years of record. The evolution of quantile estimates as a function of the record length period is presented in Fig. 1, considering here only the first m years. Systematic underestimation of benchmark values for local and REG approaches can be noticed. It shows that, on the one hand, for small samples, classical inference models such as MLE, PWB and PWU are too responsive if too many "regular" events occurred. On the other hand, for the REG model, underestimation of quantiles is related to the underestimation of the scale factor $C^{(j)}$ in equation (2) because of these "regular" events. Only the BAY model performs well enough even with record lengths shorter than 15 years. Moreover, it is by far the most robust and accurate model as, on the whole range of record length, and for all benchmark values, estimation lies in the 90% profile likelihood confidence interval. The advantage of incorporating regional information within a Bayesian framework is certainly to define a "restricted space" to which distribution parameters belong. Thus, the impact of a very extreme event – or too many low-level events – should be regarded as an extreme event related to this "restricted space".

EFFECT OF HOMOGENEITY DEGREE ON QUANTILE ESTIMATION

We focus now on the impact of the level of homogeneity of the region. For this purpose, we consider four different regions – denoted He^+, He, Ho and Ho^+ – which correspond to increasingly homogeneous regions according to the test of Hosking & Wallis (1997). The Ho region corresponds to the region analysed in the previous section. All regions have 14 sites, except for the most homogeneous one Ho^+, which contains only eight stations. To evaluate the influence of homogeneity level of a region on quantile estimation, models are assessed using two performance criteria: the Normalized Bias (*NBIAS*) and the Normalized Root Mean Squared Error (*NRMSE*). These indices are defined as follows:

$$NBIAS = \frac{1}{k}\sum_{i=1}^{k}\frac{\hat{Q}_i - Q}{Q} \qquad NRMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\frac{\hat{Q}_i - Q}{Q}\right)^2}$$

where k is the number of estimates of Q, and $\hat{Q}_i$ is the ith estimate of the benchmark value Q. To compute these two indices, we fit all models on all shortened periods of size m years $m \in \{5,10,15,20,25,30\}$. Moreover, the overall performance of each model is evaluated using a rank score $R_S = (pq - R_o)/(pq - q)$, where p is the number of models being considered, q the number of indices and $R_o \in \{1, \ldots, p\}$, 1 corresponding to the best model and p to the worst. A rank score close to 1 – resp. 0 – is associated to a model with a good – resp. poor – performance. Three quantiles are of particular interest: Q_5, Q_{10} and Q_{20}. *NRMSE*, *NBIAS* and the rank score for station U4635010 and a record length of five years are illustrated in Table 2.

From Table 2, it can be seen that the Bayesian model performs quite well independently of the region being considered. However, this model seems to perform even better when applied to a "homogeneous" or "probably heterogeneous" region according to Hosking & Wallis (1997) terminology. In contrast, the overall rank score of the REG model surprisingly decreases with the homogeneity degree of the region. These results may be related to inaccurate estimation of the Index Flood $C^{(j)}$ with only 5 years of recording. Moreover, the overall rank score for the REG model never exceeds the value of 0.6 – reached for the He^+ region. This value remains much lower than the best rank score for the BAY model – i.e. 0.88. These results corroborate the superiority of the Bayesian approach.

From Table 2, two conclusions can be established: on the one hand, for small samples, the Bayesian approach is the most competitive model; on the other, results seem to indicate that there is no need to keep increasing the homogeneity of the region as it increases the risk of being too confident in the "homogeneous region" without increasing significantly the efficiency of the model. These results are in line with similar results obtained for stations U4635010 and V3015010, except for the REG model. Indeed, for the other stations, the REG model score is larger but always lower than the BAY one.

Table 2 Estimation of *NRMSE* and *NBIAS* for station U4505010 with a record length of 5 years.

Model	*NRMSE*:			*NBIAS* :			
	Q_5	Q_{10}	Q_{20}	Q_5	Q_{10}	Q_{20}	Rank Score
MLE	0.35	0.45	0.56	–0.03	–0.10	–0.16	0.17
bHe^+	0.11	0.13	0.17	–0.01	–0.05	–0.09	0.65
rHe^+	0.21	0.20	0.19	0.06	0.03	–0.01	0.59
bHe	0.06	0.08	0.10	–0.03	–0.03	–0.05	0.78
rHe	0.28	0.29	0.29	0.17	0.18	0.19	0.33
bHo	0.06	0.07	0.09	0.02	0.03	0.03	0.88
rHo	0.31	0.33	0.35	0.21	0.24	0.26	0.21
bHo^+	0.10	0.05	0.04	–0.09	–0.04	0.01	0.81
rHo^+	0.34	0.40	0.47	0.24	0.31	0.39	0.07

MLE: maximum likelihood estimator model.
He^+, He, Ho and Ho^+ correspond to increasingly homogeneous regions according to the test of Hosking & Wallis (1997).
b: Bayesian approach.
r: Regional Index Flood.

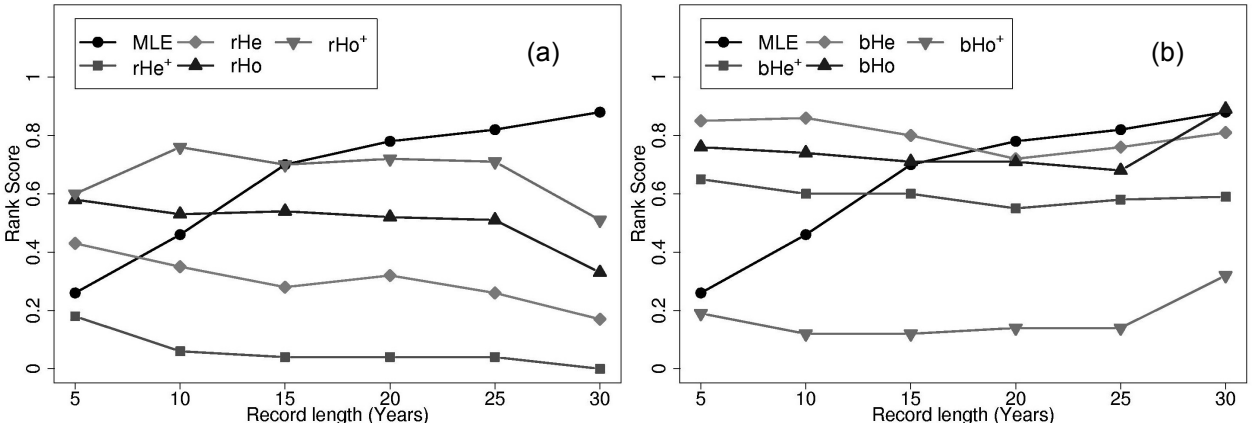

Fig. 2 Score evolution as a function of record length for Station V3015010: (a) REG scores, and (b) BAY scores.

In Fig. 2, the evolution of the overall rank score as a function of the record length is illustrated for station V3015010. The MLE score is also presented. Figure 2 indicates that the evolution of the overall rank score is more stable for regional models – that is REG and BAY models – than for the MLE. Furthermore, the benefit of increasing the homogeneity degree of the region is more relevant for the REG model than for the BAY model. Nevertheless, the worst BAY rank score is always quite close to the best REG rank score. This seems to indicate the superiority of the Bayesian approach. This last point is corroborated with the results corresponding to stations U4505010 and U4635010, except for the bHo^+ model for station U4635010 because of the inaccurate estimation of the scale factor $C^{(j)}$.

As the record length increases, the MLE model becomes more and more efficient. In particular, for record lengths greater than 15 years, it is more effective than rHe^+, rHe and rHo models. For record lengths smaller than 15 years, MLE is always less efficient than Bayesian approaches and even significantly for bHe, bHo and bHo^+ models. This is quite logical as Bayesian estimation can be looked at as a restrictive maximum likelihood estimator – restriction being defined by the prior distribution. So, under the hypothesis that the prior distribution is well defined, the "restrictive estimator" is unbiased and has a smaller variance. On the other hand, for record lengths greater than 15 years, MLE, bHe and bHo seems to be equivalent.

CONCLUSION

A framework to perform a regional Bayesian frequency analysis for partially gauged stations is presented. The proposed model has the advantage of being less restrictive than the most widely used regional model, that is the Index Flood. Several case studies from French sites were analysed to illustrate the superiority of the Bayesian approach in comparison to the traditional Index Flood and to local approaches. The influence of the homogeneity level of the pooling group on quantile estimates was also considered. Results demonstrate that working with quite large and homogeneous regions, rather than small and strongly homogeneous regions, is more efficient. Further work can focus on the regional estimation of other characteristics of the flood hydrograph.

REFERENCES

Dalrymple, T. (1960) Flood frequency analysis. *US Geol. Survey Water Supply Paper 1543A*.

Hosking, J. R. M. & Wallis, J. R. (1997) *Regional Frequency Analysis*. Cambridge University Press, Cambridge, UK.