2nd International Conference - Water resources and wetlands. 11-13 September, 2014 Tulcea (Romania); Available online at http://www.limnology.ro/water2014/proceedings.html
Editors: Petre Gâștescu ; Włodzimierz Marszelewski ; Petre Bretcan; ISSN: 2285-7923; Pages: 250-256;

CHOICE OF THE ADEQUATE FREQUENCY MODEL FOR MAXIMUM RAINFALLS BY MINIMIZING THE PENALIZED CRITERIA (AIC AND BIC)

Amel Oucherif
National Polytechnic School, Algiers, Algeria
10 Avenue Hassen Badi, El harrach, Algiers, Algeria, +213 21 521027, Email Amel.oucherif@hotmail.fr

Abstract
One of the main challenges in hydraulics engineering is the determination of the statistical law governing the studied phenomenon. Accordingly, the main aim of our work is to determine the best statistical distribution for modeling maximum rainfalls of the considered chronological series belonging to the two studied Algerian watersheds namely the Soummam and the Chélif. In order to determine the best frequency model, different methods have been used which are: the visual adjustments, Q-Q plot, goodness of fit tests (Khi-2, Kolmogorov-Smirnov and Anderson Darling), the harmonization of the return periods and quantiles and the minimization of Akaike and Bayesian information criteria (AIC and BIC). Compared to the others methods, the information criteria are the best mean to overcome the problem of choosing the "quasi-true" model among the considered competing statistical distributions which are: Normal distribution, Log-Normal distribution, Gumbel distribution, Weibull distribution, Exponential distribution, GEV distribution, Pearson III distribution and Log-Pearson III distribution. In addition, the probability and return period values analysis showed that the Cunnane formula is the best for calculating the empirical frequency among the ones studied namely: Hazen, Weibull, Chegodayev, Cunnane, Gringorten and Tukey. Therefore, the Cunnane formula has been used for the visual adjustments. Otherwise, the consequences generated by the change in the number of individuals composing a long sample were studied by varying the sample size to seek the law of probability, for each data set generated, and compare it to the law obtained for the original data sets.

Keywords: information criteria AIC and BIC, chronological series modeling, statistical distribution, empirical frequency, inferential statistics, goodness of fit tests.

1. INTRODUCTION

In the field of hydraulics engineering, we often use the combination of statistics and hydrology; this has given rise to statistical hydrology. It is the description of hydrological processes such as precipitation or runoff by using statistical techniques based mainly on the analysis of data and the probability theory. One of the main challenges in this field is the determination of the statistical law governing the studied phenomenon.

2. RETURN PERIODS AND EMPIRICAL FREQUENCIES

The aim of this study is to analyze the values of the probabilities and return periods given by the considered empirical frequency formulas, namely: Hazen, Weibull, Chegodayev, Cunnane, Gringorten and Tukey.

3. METHODOLOGY OF QUASI-TRUE MODEL CHOICE

In order to determine the best statistical distribution fitting the maximum precipitation chronological series, we must, firstly, perform a data pre-processing by checking the homogeneity, the independence, the stationarity and presence of outliers values. This is done to ensure the validity of the statistical analysis to be performed. Secondly, we perform a visual adjustment by highlighting confidence intervals and quantile-quantile plots for the considered distributions listed above. Thirdly, we use the goodness of fit tests. Fourthly, the harmonization of return periods and quantiles is applied by comparing the calculated and the observed values of the quantiles and return periods using all the tested statistical distributions, according to the period of observation. Finally, the Akaike and Bayesian information criteria (AIC and BIC) are calculated for each distribution, in order to choose the best one by the Bayesian theory.

4. STUDY OF THE EFFECT OF VARYING THE STUDIED CHRONOLOGICAL SERIES SIZE

The aim is to determine the consequences generated by the change in the number of individuals composing a long sample by varying the sample size to seek the law of probability, for each data set generated, and compare it to the law obtained for the original data sets. Then, we will try to draw possible conclusions. Within the framework of our study, two watersheds in the North of Algeria have been considered, namely: the Chélif and the Soummam. According to the availability and the reliability of the datasets, 5 chronological series of maximum rainfall have been selected.

5. RESULTS AND DISCUSSION

We will present and discuss the various results obtained according to the various analyzes described above. As we arrived to the same conclusions after studying all the above datasets and for reasons of brevity, we present only the results obtained for one station which is: Ain Oussera 011205 (01: Chélif).

5.1. Return periods and empirical frequencies

For the station 011205, the sample size is 60 and the left and right tails are respectively 12.3 mm and 81 mm. The values obtained for the return periods and the empirical frequencies according to the formulas above are mentioned in the following table:

Table 1. Return periods and empirical frequencies for the highest value

P=81 mm	Hazen	Weibull	Chegodayev	Cunnane	Tukey	Gringorten
T (years)	120.00	61.00	86.29	86.00	90.50	107.36
FD (exceedance)	0.008	0.016	0.012	0.012	0.011	0.009
FND (non-exceedance)	0.992	0.984	0.988	0.992	0.989	0.991

We can see that for the highest value (81 mm), the return periods obtained are different according to the formula used even if non-exceedance probability are relatively close. It is the same thing for the largest values, except for other than extreme values where the return period is not too different. Consequently, the difference is not important for the exceedance probabilities but it is for the return periods of the right tails values. Therefore, the issue is at level of the extreme values where the results are very different according to the used formula. However, we found that for all the studied series, the Cunnane formula is the best compromise because it gives the closest values of the empirical frequencies to "1". This is why, we recommend using the Cunnane formula for calculating the empirical frequencies.

5.2. Quasi-true model choice

5.2.1. Pre-processing

The results of the **stationarity (KPSS), homogeneity (Pettitt)**, **independence (Wald-Wolfowitz)** and **outliers (Grubbs and Beck)** tests are shown in the tables below:

Table 2. Pre-processing tests results

Wald-Wolfowitz test	**Pettitt test**	**KPSS test**
Statistic : $\|u\|$**=0.738**	P-value : p**=0.985**	Statistic : $\|K\|$**=0.095**
Decision : **u<1,96** => **Independence verified (CI = 95%)**	Decision : **p>0.05** => **Homogeneity verified (CI = 95%)**	Decision: **K<0.146** => **Stationarity verified (CI = 95%)**

Table 3. Grubbs and Beck test result

Null hypothesis	Test statistic Kn	Maximum value (Xh)	Minimum value (Xl)
H0 : there is outliers : don't belong to [Xl , Xh]	**2.84**	**85.30 mm**	**9.34 mm**

As mentioned before, the extreme values of the studied dataset are 12.3 mm and 81mm. Therefore, according to the Grubbs and Beck test, there are no outliers. Consequently, the pre-processing tests being positive, the statistical analysis is significant.

5.2.2. Visual adjustment

The visual adjustment highlighting the confidence intervals for the different considered distributions are presented in the following graphs:

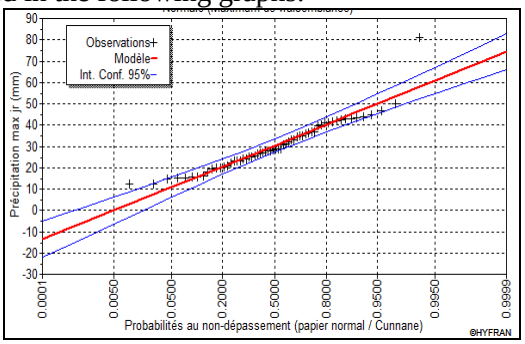

Figure 1. Normal visual adjustment

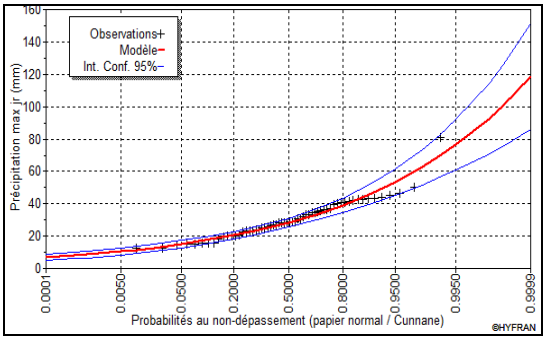

Figure 2. Lognormal visual adjustment

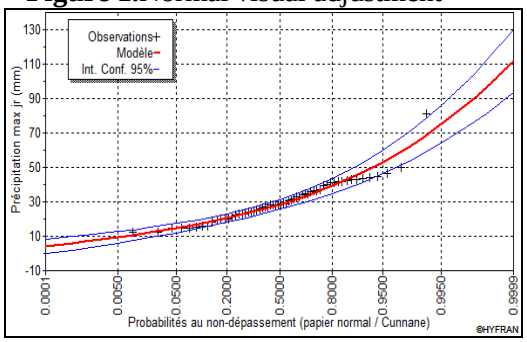

Figure 3. Gumbel visual adjustment

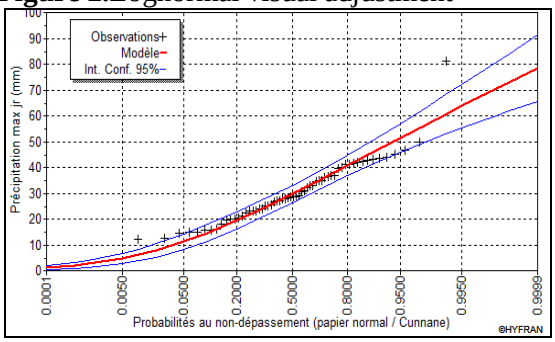

Figure 4. Weibull visual adjustment

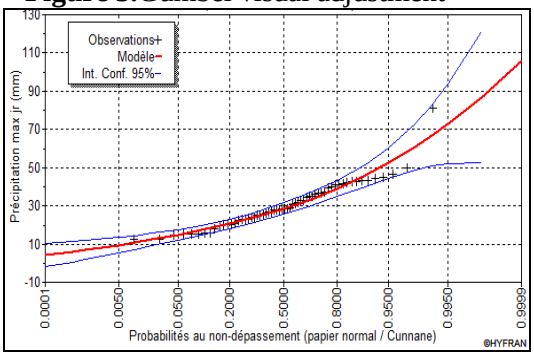

Figure 5.GEV visual adjustment

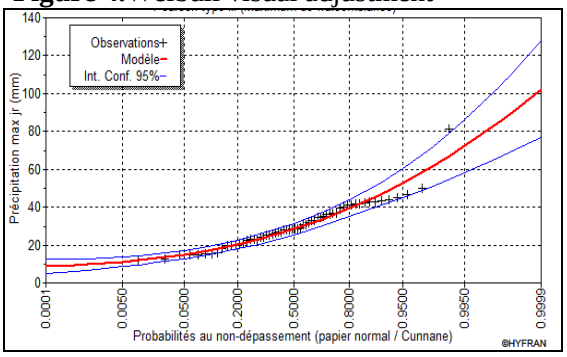

Figure 6. Pearson III visual adjustment

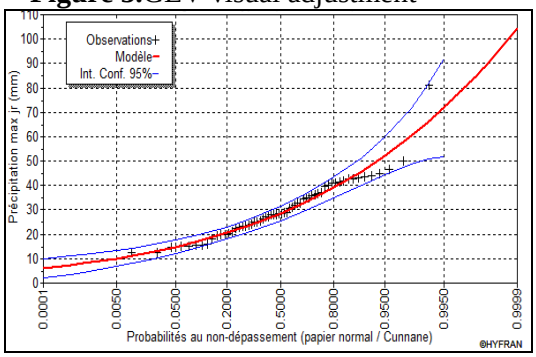

Figure 7. Log-Pearson III visual adjustment

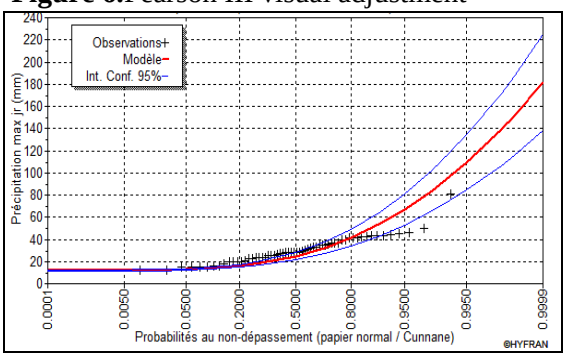

Figure 8. Exponential visual adjustment

Through these charts, it is clear that the exponential distribution is rejected because several points are outside of the envelope curves of the confidence intervals, in addition to the fact that compared to the rest of

the laws, the points do not align with the curve of the model. For the remaining laws, it is difficult to choose among them, that best matches the sample; in fact, they all seem more or less, adjust to it.

Another type of visual adjustment which is the Q-Q plot has been used. The graphs are as follows:

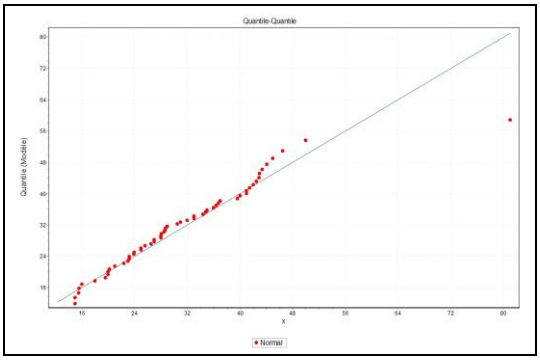

Figure 9. Normal Q-Q plot

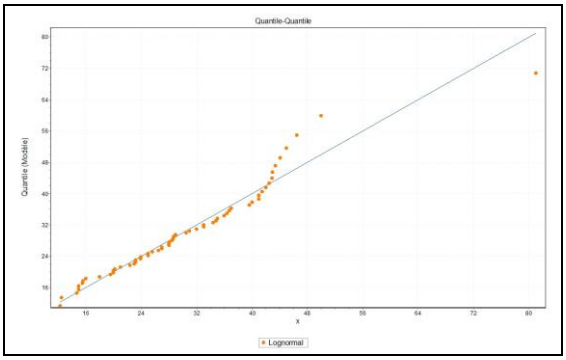

Figure 10. Lognormal Q-Q plot

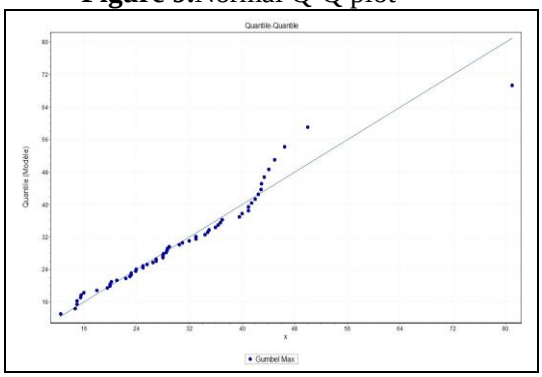

Figure 11. Gumbel Q-Q plot

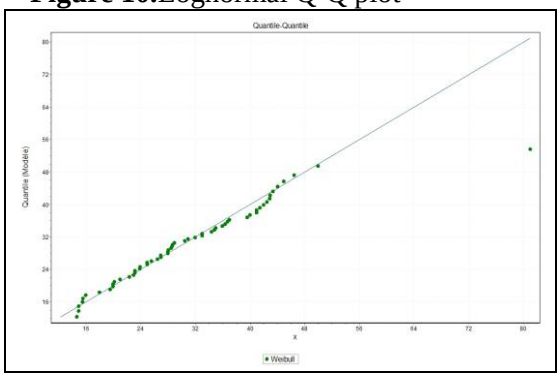

Figure 12. Weibull Q-Q plot

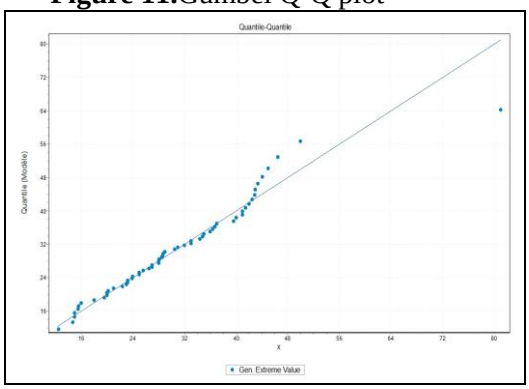

Figure 13.GEV Q-Q plot

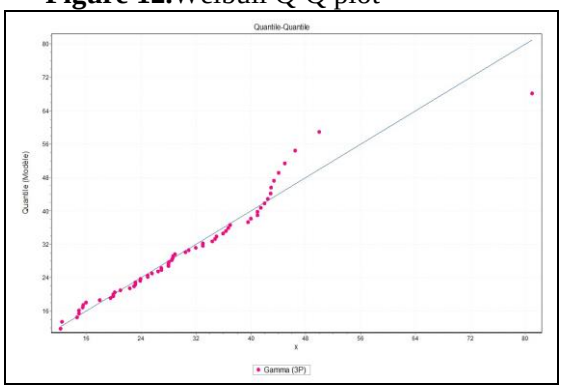

Figure 14. Pearson III Q-Q plot

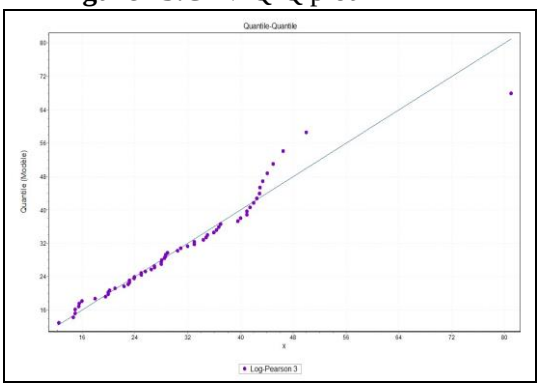

Figure 15. Log-Pearson III Q-Q plot

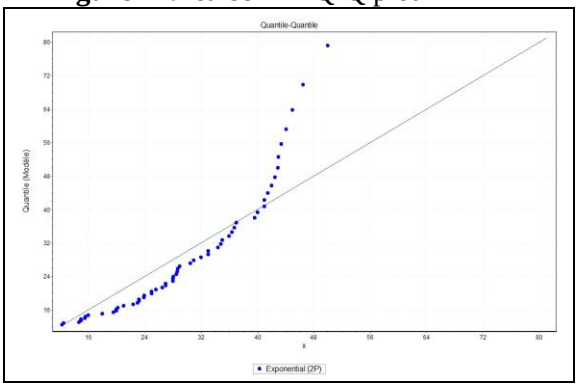

Figure 16. Exponential Q-Q plot

We can see that all of the statistical distributions suggest a scatter relatively aligning around the first bisector. This is not the case for exponential and normal laws.

5.2.3 Goodness of fit tests

The visual adjustments must be supported by the goodness of fit tests indicated previously namely: Khi-2, Kolmogorov-Smirnov ad Anderson-Darling tests.

Table 4.Goodness of fit tests results

Distribution	Tests	Stastistic test : stat	Critical value: c	Decision
Normal	Khi-deux	1.008	11.07	Stat<c => OK
	K-S	0.096	0.172	Stat<c => OK
	A-D	0.632	2.502	Stat<c => OK
Lognormal	Khi-deux	1,156	11.07	Stat<c => OK
	K-S	0,065	0.172	Stat<c => OK
	A-D	0,495	2.502	Stat<c => OK
Gumbel	Khi-deux	2.395	11.07	Stat<c => OK
	K-S	0.064	0.172	Stat<c => OK
	A-D	0.457	2.502	Stat<c => OK
Weibull	Khi-deux	4.599	11.07	Stat<c => OK
	K-S	0.085	0.172	Stat<c => OK
	A-D	0.622	2.502	Stat<c => OK
Exponential	Khi-deux	12.041	9.488	Stat>c => OK
	K-S	0.196	0.172	**Stat>c => rejected**
	A-D	3.872	2.502	**Stat>c => rejected**
GEV	Khi-deux	2.000	11.07	Stat<c => OK
	K-S	0.059	0.172	Stat<c => OK
	A-D	0.355	2.502	Stat<c => OK
Pearson II	Khi-deux	0.633	11.07	Stat<c => OK
	K-S	0.064	0.172	Stat<c => OK
	A-D	0.466	2.502	Stat<c => OK
Log-Pearson III	Khi-deux	2.322	11.07	Stat<c => OK
	K-S	0.061	0.172	Stat<c => OK
	A-D	0.409	2.502	Stat<c => OK

We can note, according to these results, that the three used tests are positive regarding to the adjustment of all the distributions except for the exponential law.

5.2.4. Harmonization of quantiles and return periods

The calculated quantiles and return periods (according to Bobée & Ashkar, 1991) for the studied frequency models for the highest value are shown in the following table:

Table 5. Quantiles and return periods calculation

Distribution	**Quantiles for T= 60 years (mm)**	**Return periods (years) for Pmax=81 mm**
Normal	55.6	3571.43
Lognormal	64.3	357.14
Gumbel	63.6	376.69
Pearson III	62.3	68.41
Log-Pearson III	61.8	66.03
Weibull	57.9	54.55
Exponential	87.3	185.18
GEV	62.3	310.51

According to these results, all the calculated quantiles for a return period equal to 60 years don't match with the highest observed value which is 81mm. In addition, the return periods corresponding to the

right tail value (81 mm) are different from the observation period which is 60 years. Consequently, it is concluded that we could not choose the right model by this method.

5.2.5. Choice of the appropriate model by the information criteria (AIC and BIC)

The Akaike and Bayesian information criteria obtained are as follows:

Table 6. AIC and BIC calculation

Distribution	BIC	AIC
Lognormal	**237.583**	**234.781**
Gumbel	237.839	235.036
Exponential	240.213	237.411
Pearson III	240.640	236.436
Log-Pearson III	240.945	236.741
GEV	241.087	236.884
Weibull	243.736	240.934
Normal	245.534	242.732

We can note that the smallest value for the two criteria is obtained with the **lognormal distribution.** Consequently, the most appropriate frequency model to choose is the **lognormal.**

5.3. Results of the effect of varying the sample size study

For the considered datasets, the best distribution given by AIC and BIC criteria is lognormal till the iteration N=20. For **N<20**, the information criteria indicate the exponential law as the best model. Consequently, we couldn't draw conclusions from these results regarding the limit size concerning the extrapolation for the studied chronological series. However, an interesting result can be noted which is that for the threshold methods, the threshold must not be chosen arbitrarily, because the model best fitting the chronic corresponds to the two-parameter exponential law only from a certain precise value. Furthermore, the approach that it is possible to estimate the quantiles for the short chronological series can be justified because when the number of values of the datasets decreases, the best model tends to the exponential one, knowing that this law is a particular case of the generalized Pareto distribution used for modeling peak over threshold series.

6. CONCLUSIONS

All the results we have achieved through our work lead us to the following conjectures:

By performing an analysis of return periods, we could conclude that the problem essentially arises at the extreme values. This finding has been confirmed when analyzing visual adjustments. In addition, we opted for the Cunnane formula as the empirical frequencies formula as representing the best compromise for the different studied distributions.

Moreover, the graphical adjustment and Quantile-Quantile diagrams are a good criterion which can guide the choice of the model, but they remain insufficient to detect the quasi-true model.

Concerning the goodness of fit tests, we can issue the same point regarding the visual adjustments but they allow us, only, the analysis of the distribution fitting, taken alone, without defining the best distribution. On the other hand, theories encountered in the literature according to which the Khi-2 test take into account, only the average values and that the Anderson-Darling test fits just the extreme values are not highlighted for the studied data.

The method based on the harmonization of quantiles and return periods highlights the fact that it's possible that the highest value of a sample has a return period greater than the observation period because the right tail has a very important variability (Miquel, 1984) as seen previously.

Taking into account the insufficiency of all these methods, we opted for the bayesian theory by the minimization of the Akaike and Bayesian information criteria (AIC and BIC) because according to its mathematical theory (Bertrand & Bertrand, 2012), it is, for the time being, one of the best ways to chose the best model offering the best compromise between the bias and the standard deviation square root for the AIC criterion and the most parsimonious model for the BIC criterion (Lebarbier & Mary-Huard, 2004). Consequently, according to these criteria, we obtained that, for the studied chronological series of maximum rainfalls for **the Soummam and the Chélif watersheds,** the best fitting distribution is the **lognormal**.

From there, we can conclude that we could get rid of the idea that the Gumbel distribution is the best fitting law for modeling extreme values because, it is not always the case as shown in the present study.

Ultimately, it would be beneficial for statistical hydrology specialists to expand their investigation field to develop other methods to ensure the greatest possible convergence to the true frequency model to solve this thorny issue.

REFERENCES

Bobée B & Ashkar F, 1991, The Gamma family and derived distributions applied in hydrology, *Water Resources Publications*;

Bertrand F. & Bertrand, M 2012, Choix du modèle, IRIMA Université de Strasbourg, France, Econométrie Appliquée.

Lebarbier E. & Mary-Huard T., 2004, Le critère BIC: fondements théoriques et interprétation, Rapport de recherche, INRIA;

Miquel, J. 1984, *Guide pratique d'estimation des probabilités de crues*, Editions EYROLLES.