

**UNE MÉTHODE RATIONNELLE DE COMPARAISON
DES DISTRIBUTIONS DE CRUE**

**Mémoire présenté comme exigence partielle
pour l'obtention du grade de Maître ès Sciences (M.Sc.)**

**par
Vincent Fortin**

**Institut national de la recherche scientifique, INRS-Eau
Case postale 7500, Sainte-Foy (Québec) CANADA G1V 4C7**

Septembre 1994

© Tous droits réservés, Vincent Fortin 1994

Résumé

Pour dimensionner des ouvrages hydrauliques en rivière, on utilise souvent la valeur de débit dépassée en moyenne une fois toutes les T années, que l'on nomme débit x_T de période de retour T . À un site jaugé, x_T peut être estimé en ajustant une distribution statistique aux mesures disponibles du débit maximum annuel (DMA). Parce que ces séries sont en général assez courtes, il est difficile de déterminer la distribution statistique la plus performante pour estimer x_T .

Nous avons développé une procédure de comparaison de ces distributions qui tient compte du nombre N d'observations disponibles, de la période de retour T d'intérêt, de l'objectif réel de l'estimation de x_T (qui est de réaliser un projet au moindre coût), ainsi que de l'incertitude sur la distribution réelle des DMA. L'application de cette procédure à quelques séries de DMA du Québec et de l'Ontario montre que l'utilisation systématique de distributions à 2 paramètres, spécialement la distribution Gumbel (ou EV1), est préférable à l'utilisation systématique de distributions à 3 paramètres, incluant les distributions des valeurs extrêmes (GEV) et log-Pearson type III (LP3).

L'utilisation d'une politique de sélection tenant compte des coefficients d'asymétrie et de variation donne cependant de meilleurs résultats que l'application systématique de n'importe quelle distribution. De plus cette politique sélectionne fréquemment des distributions à 3 paramètres, telles GEV et LP3, spécialement pour l'extrapolation ($T > N$). Les distributions à 3 paramètres ne doivent donc pas être négligées, mais utilisées dans le cadre d'une procédure de sélection tenant compte des caractéristiques statistiques des séries ajustées.

Remerciements

Je désire remercier messieurs Peter Rasmussen, Jacques Bernier et Bernard Bobée pour leurs judicieux conseils lors de la rédaction de ce mémoire, ainsi que M. Éric Parent pour le temps qu'il a consacré à sa correction. L'aide financière du Conseil de recherches en sciences naturelles et en génie du Canada a été appréciée.

Il me faut mentionner que ce mémoire n'aurait pu voir le jour sans le constant support de ma conjointe, Isabelle Moutquin. Je veux enfin remercier mes parents de m'avoir transmis le goût d'apprendre et la discipline nécessaire pour atteindre mes objectifs.

Table des matières

Résumé	iii
Remerciements.....	v
Liste des figures	x
Liste des tableaux.....	xii
Liste des symboles et sigles.....	xiii
Introduction	1
1 Problématique	3
1.1 Lien entre le risque et la période de retour.....	3
1.2 Estimation du débit de période de retour T	5
1.3 Approximation de la distribution des débits.....	7
1.3.1 Approche non paramétrique.....	7
1.3.2 Approche paramétrique.....	8
1.4 Sélection d'une combinaison D/M	11
1.5 Comparaison des politiques de sélection d'une combinaison D/M.....	15
2 Aspects théoriques.....	17
2.1 Simulation de débits maximums annuels.....	17
2.1.1 Simulation à l'aide d'une distribution parente	18
2.1.2 Simulation par suréchantillonnage	19
2.2 Mesure de la performance d'une politique de sélection d'une combinaison D/M.....	23
2.3 Mesure de la performance d'une politique de sélection appliquée à une région non homogène.....	24
2.4 Une politique de sélection découlant de la procédure de comparaison.....	27
2.4.1 Hypothèses implicites liées à l'analyse Bayésienne.....	30
2.4.2 Optimalité de la politique de sélection.....	31
3 Exemple d'application.....	33
3.1 Base de données.....	33
3.2 Tailles d'échantillon étudiées	34
3.3 Périodes de retour considérées	34
3.4 Politiques de sélection considérées.....	34
3.5 Application de la procédure de comparaison de politiques de sélection.....	38

3.5.1	Simulation de débits à partir des échantillons de référence.....	42
3.5.2	Comparaison de la distribution des séries simulées avec celle des données réelles	42
3.5.3	Détermination des valeurs correctes de x_T	43
3.5.4	Résultats de la comparaison pour $\alpha=1$ et $r=2$ (EQM).....	44
3.5.5	Étude de sensibilité aux paramètres α et r	47
3.6	Application de la politique dynamique de sélection	51
3.7	Validation de la politique dynamique de sélection.....	55
4	Discussion	57
4.1	Distributions statistiques recommandées	57
4.2	Utilisation de d'autres politiques dynamiques	59
4.3	Application au développement d'un système-expert	59
4.4	Estimation du paramètre α	60
4.5	Application à de grandes périodes de retour	60
	Conclusion	63
	Annexe A: Méthodes d'estimation.....	65
A.1	Méthode du maximum de vraisemblance.....	65
A.2	Méthode des moments.....	66
A.2.1	Moments d'une distribution.....	66
A.2.2	Application pratique de la méthode des moments	67
A.3	Méthode des moments pondérés	68
	Annexe B: Comparaison des procédés de simulation par suréchantillonnage.....	71
	Annexe C: Distributions utilisées pour l'estimation des crues de conception	75
	Annexe D: Types de paramètres	87
D.1	Définition des types de paramètres.....	87
D.2	Effet d'une transformation logarithmique	88
	Annexe E: Distributions statistiques utilisées	89
E.1	Distributions normale et log-normale	89
E.1.1	Distribution normale (N)	89
E.1.2	Distribution log-normale à 2 paramètres (2LN)	90
E.2	Famille gamma	90
E.2.1	Distribution gamma à 2 paramètres (G2).....	91
E.2.2	Distribution Pearson type III (P3).....	92
E.2.3	Distribution log-Pearson type III (LP3).....	92

E.2.4	Distribution gamma généralisée à 3 paramètres (GG3).....	94
E.2.5	Distribution Weibull à 2 paramètres (W2).....	95
E.3	Distributions des valeurs extrêmes.....	95
E.3.1	Distribution Gumbel (EV1).....	96
E.3.2	Distribution généralisée des valeurs extrêmes (GEV).....	96
	Bibliographie.....	99

Liste des figures

Figure 1	Exemple de relation entre le débit x_T , la période de retour T et la probabilité au dépassement π	p. 5
Figure 2	Étapes menant à l'estimation d'une crue de conception	p. 10
Figure 3	Estimation de la performance d'une politique de sélection d'une combinaison D/M	p. 14
Figure 4	Algorithme de simulation par suréchantillonnage <i>Bootstrap</i>	p. 20
Figure 5	Algorithme de simulation par tirage de Pólya	p. 21
Figure 6	Relations entre les différentes distributions considérées	p. 35
Figure 7	Distribution des couples (\hat{C}_s, \hat{C}_v) des séries réelles de taille $N=20$	p. 40
Figure 8	Distribution des couples (\hat{C}_s, \hat{C}_v) des séries réelles de taille $N=40$	p. 40
Figure 9	Distribution des couples (\hat{C}_s, \hat{C}_v) des séries simulées de taille $N=20$	p. 40
Figure 10	Distribution des couples (\hat{C}_s, \hat{C}_v) des séries simulées de taille $N=40$	p. 41
Figure 11	Distribution des couples (\hat{C}_s, \hat{C}_v) des séries réelles et simulées de taille $N=20$	p. 41
Figure 12	Distribution des couples (\hat{C}_s, \hat{C}_v) des séries réelles et simulées de taille $N=40$	p. 41
Figure 13	Mesures de performance pour $\alpha = 1$ et différentes valeurs de r	p. 48
Figure 14	Mesures de performance pour $r=2$ et différentes valeurs de α	p. 48
Figure 15	Performance de la politique dynamique de sélection pour les séries de calibration	p. 55
Figure 16	Performance de la politique dynamique de sélection pour les séries de validation	p. 56

Liste des figures (suite...)

Figure B1	Distribution <i>a posteriori</i> de la proportion Q pour $N=20$ et $q=0,25$	p. 72
Figure C1	Observations mesurées à la station 02FC001 tracées sur papier de probabilité normal en utilisant la FPE de Cunnane	p. 80
Figure C2	Observations mesurées à la station 02FC002 tracées sur papier de probabilité normal en utilisant la FPE de Cunnane	p. 81
Figure C3	Observations mesurées à la station 02EA005 tracées sur papier de probabilité normal en utilisant la FPE de Cunnane	p. 82
Figure C4	Observations mesurées à la station 02EC002 tracées sur papier de probabilité normal en utilisant la FPE de Cunnane	p. 83
Figure C5	Observations mesurées à la station 05PB014 tracées sur papier de probabilité normal en utilisant la FPE de Cunnane	p. 84
Figure C6	Observations mesurées à la station 04LJ001 tracées sur papier de probabilité normal en utilisant la FPE de Cunnane	p. 85
Figure C7	Observations mesurées à la station 02PJ007 tracées sur papier de probabilité normal en utilisant la FPE de Cunnane	p. 86

Liste des tableaux

Tableau 1	Résumé des distributions considérées pour estimer les débits x_T	p. 36
Tableau 2	Politiques de sélection d'une combinaison D/M envisagées	p. 39
Tableau 3	Valeur correcte de x_T selon l'échantillon de référence et la FPE considérés	p. 44
Tableau 4	Politique minimisant l'écart quadratique moyen	p. 46
Tableau 5	Politique minimisant l'écart quadratique moyen (EQM) pour la FPE de Cunnane	p. 46
Tableau 6	Politiques sélectionnées pour $\alpha=1$ et différentes valeurs de r	p. 49
Tableau 7	Politiques sélectionnées pour $r=2$ et différentes valeurs de α	p. 50
Tableau 8	Probabilité, en pourcentage, qu'une série provienne d'un échantillon de référence donné, estimée à partir de \hat{C}_v et \hat{C}_s	p. 53
Tableau 9	Politique dynamique de sélection d'une combinaison D/M pour estimer x_T	p. 54
Tableau C1	Séries de données réelles de moins de 40 observations	p. 76
Tableau C2	Séries de données réelles de 40 observations et plus	p. 78
Tableau C3	Détail des observations des échantillons de référence (m^3/s)	p. 79

Liste des symboles et sigles

$\hat{?}$	Indique que l'expression qu'il chapeaute est une estimation
$\bar{?}$	Représente la moyenne de l'expression qu'il chapeaute
$\underline{?}$	Indique que l'expression soulignée est un vecteur
2LN	Distribution log-normale à 2 paramètres
α	Paramètre influençant la symétrie de la fonction $g(.)$
Δx_T	Différence entre l'estimation de x_T et la valeur exacte
φ	Perte nette moyenne encourue par plusieurs estimations de x_T
π	Probabilité au dépassement
θ	Paramètre d'une distribution statistique
a	Constante entrant dans le calcul d'une FPE
C^*	Coût moyen annuel optimal d'un ouvrage hydraulique
C_d	Coût moyen annuel des défaillances d'un ouvrage hydraulique
C_o	Coût moyen annuel d'opération d'un ouvrage hydraulique
C_s	Coefficient d'asymétrie
C_v	Coefficient de variation
D	Distribution statistique
D/M	Combinaison d'une distribution statistique D et d'une méthode d'estimation M
E_k	Sous-espace de S
DMA	Débits maximums annuels
EQM	Écart quadratique moyen
EV1	Distribution Gumbel (<i>Extreme Value type I</i>)
$F(.)$	Fonction de densité cumulée inverse
f.d.p.	Fonction de densité de probabilité
FPE	Formule de probabilité empirique
$g(.)$	Fonction mesurant la performance d'une estimation de x_T
G2	Distribution gamma à 2 paramètres
GEV	Distribution généralisée des valeurs extrêmes (<i>Generalized Extreme Value</i>)
GG3	Distribution gamma généralisée à 3 paramètres
LP3	Distribution log-Pearson type III
M	Méthode d'ajustement des paramètres d'une distribution statistique
MM	Méthode des moments

Liste des symboles et sigles (suite...)

MV	Méthode du maximum de vraisemblance
n	Taille d'échantillon des séries simulées
N	Taille d'échantillon des séries réelles, ou distribution normale
P3	Distribution Pearson type III
PWM	Méthode des moments pondérés (<i>Probability-Weighted Moments</i>)
q	Proportion de zéro dans un échantillon de variables de Bernouilli
Q	Proportion de zéro dans une population infinie de variables de Bernouilli
ρ	Risque de défaillance d'un ouvrage hydraulique
r	Paramètre déterminant le degré de la fonction $g(\cdot)$
R	Nombre d'échantillons simulés
S	Ensemble de caractéristiques statistiques
T	Période de retour
X, Y	Échantillon de mesures du débit maximum annuel
x_T	Débit de période de retour T
Z	Perte nette encourue par l'estimation de x_T

INTRODUCTION

L'un des critères utilisés pour le dimensionnement d'ouvrages hydrauliques, pour la gestion des réservoirs et pour la prévision des inondations est la valeur de débit x_T qui est dépassée en moyenne (sur une longue période) une fois toutes les T années, et qui correspond au débit de période de retour T . À un site où le débit est mesuré depuis un grand nombre d'années, x_T peut être estimé en ajustant une distribution statistique aux mesures disponibles du débit maximum atteint à chaque année.

Depuis le début du siècle, de nombreuses distributions statistiques et méthodes d'ajustement ont été proposées pour estimer x_T . Cependant, aucune des distributions proposées n'est basée sur la modélisation des phénomènes physiques causant les fluctuations du débit: ce sont toutes des approximations supposées acceptables de la distribution réelle, mais inconnue, des débits maximums annuels.

Parce que les séries de mesures du débit maximum annuel disponibles sont en général assez courtes, il est difficile de choisir une distribution statistique et une méthode d'ajustement pour estimer x_T . Au cours des vingt-cinq dernières années, plusieurs études ont été menées pour comparer les différentes distributions statistiques et méthodes d'ajustement proposées. Cependant, un grand nombre de ces études reposent sur des hypothèses discutables et très peu ont été pensées en fonction de l'objectif réel derrière l'estimation de x_T , qui est la réalisation d'un projet au moindre coût (économique, social ou environnemental).

Après avoir confronté les différentes approches utilisées pour comparer la performance des diverses distributions statistiques et méthodes d'ajustement utilisées en pratique, nous élaborerons, en utilisant des éléments de ces approches, une procédure de comparaison basée sur des hypothèses vérifiables, et qui prend en compte l'objectif réel qui nous amène à estimer x_T . À titre d'exemple, nous appliquerons cette procédure à un ensemble de séries de débits maximums annuels du Québec et de l'Ontario.

1 PROBLÉMATIQUE

Les ouvrages construits en rivière sont en général dimensionnés de façon à pouvoir résister à une crue extrême, que l'on nomme crue de conception (*design flood*). Lorsque le débit est supérieur à la crue de conception, l'ouvrage est en situation de défaillance. Par exemple, pour un évacuateur de crue, la crue de conception est le débit à partir duquel il devient impossible de contrôler le niveau d'eau du réservoir en amont. Au delà de ce seuil, l'évacuateur ne remplit plus son rôle. On peut aussi penser à un muret destiné à empêcher une rivière de sortir de son lit pour inonder le voisinage. Lorsque le débit est tel que l'eau dépasse le haut du muret, l'ouvrage est en situation de défaillance.

En général, lorsque l'on construit un ouvrage hydraulique, on doit s'assurer que le risque d'observer une défaillance pendant la durée de vie de l'ouvrage ne dépasse pas un certain niveau. Nous rappellerons dans un premier temps que le risque de défaillance est lié à la période de retour T de la crue de conception. Ensuite, nous montrerons comment le débit x_T correspondant à une période de retour T peut être estimé. Enfin nous évoquerons les difficultés rencontrées lors de l'estimation de x_T ainsi que l'approche que nous proposons pour comparer différentes estimations de x_T .

1.1 Lien entre le risque et la période de retour

Par définition la probabilité d'observer une défaillance durant une année donnée est égale à la probabilité que le débit dépasse la crue de conception cette année là. On peut montrer que le débit ayant une probabilité au dépassement π a une période de retour $T=1/\pi$, c'est à dire qu'il sera dépassé en moyenne, sur une longue période, une fois toutes les T années [BOBÉE *et al.* 1993a]. La probabilité que la valeur maximale X du débit atteinte une année donnée dépasse le débit x_T de période de retour T est donc l'inverse de cette période de retour:

$$\Pr[X > x_T] = \pi = 1/T \quad (1)$$

Lors de la conception d'un ouvrage hydraulique, et selon la nature de l'ouvrage à construire, on doit fixer le risque de défaillance que l'on est prêt à accepter. Ce risque hydrologique

correspond à la probabilité d'observer au moins une défaillance durant la vie de l'ouvrage. Ainsi, si l'on note $\{x_1, x_2, \dots, x_N\}$ un ensemble de réalisations indépendantes de la variable aléatoire X (le débit maximum annuel), le risque (ρ) qu'au moins une réalisation de X dépasse x_T est défini par $\rho = 1 - \prod_{i=1, N} \Pr[X \leq x_T]$, soit $1 - (1 - 1/T)^N$.

L'ingénieur qui conçoit l'ouvrage doit alors s'assurer que celui-ci peut contenir une crue telle que la probabilité qu'il se produise au moins une défaillance au cours de la vie de l'ouvrage est inférieure ou égale au risque fixé. Il existe une relation entre la durée de vie d'un ouvrage (N), le risque de défaillance (ρ), et la période de retour (T) du débit correspondant à la crue de conception [BOBÉE *et al.* 1993a]:

$$T = \frac{1}{1 - (1 - \rho)^{1/N}} \quad (2)$$

Connaissant le risque et la durée de vie de l'ouvrage, on peut donc déterminer la période de retour de la crue de conception. Pour aider à la conception d'ouvrages hydrauliques fiables, le défi de l'hydrologie statistique est donc d'approximer la relation réelle mais inconnue entre le débit x_T d'une rivière et sa période de retour T ou sa probabilité au dépassement $\pi = 1/T$. La figure 1 montre une telle relation. En abscisse on retrouve de façon équivalente la période de retour T ou la probabilité au dépassement $\pi = 1/T$, alors qu'on a en ordonnée une échelle de débits. La relation peut avoir différentes formes mais est toujours une fonction croissante de la période de retour et décroissante de la probabilité au dépassement.

En combinant l'équation 1 à la relation entre le débit x_T et sa période de retour T , on peut estimer l'amplitude de la crue de conception. Par exemple, lors de la conception d'un évacuateur de crue d'une durée de vie prévue de cent ans, si l'on veut limiter le risque de défaillance à 20%, on doit s'assurer que l'évacuateur peut contenir le débit de période de retour $T \approx 450$ ans. Ou encore, si l'on construit un muret d'une durée de vie de 50 ans et que l'on veut limiter le risque de défaillance à 40%, on doit le construire assez haut pour qu'il puisse contenir le débit de période de retour $T \approx 100$ ans, ce qui correspondrait à un débit de près de 380 m³/s, pour l'exemple donné à la figure 1.

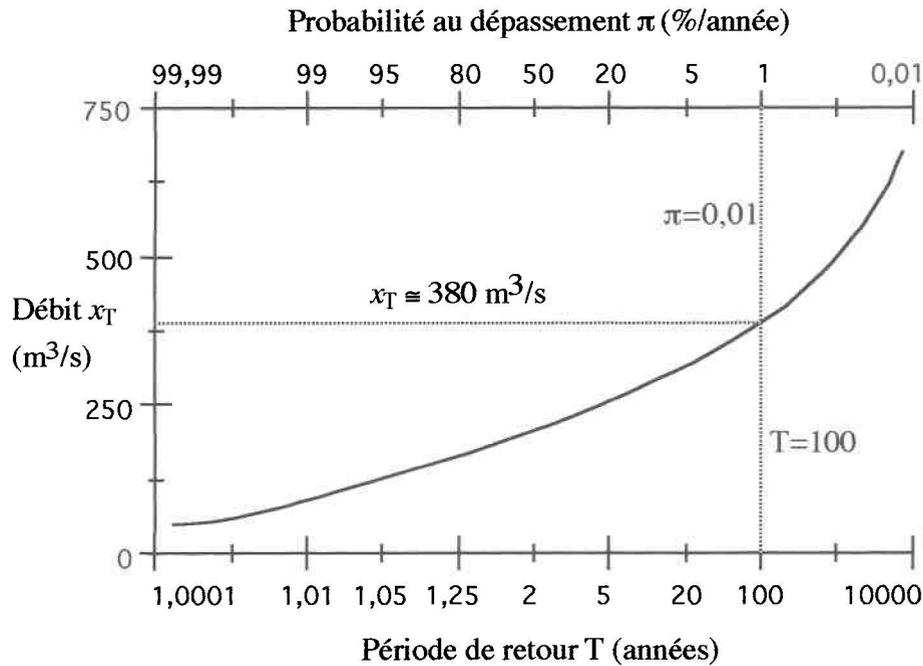


Figure 1: Exemple de relation entre le débit x_T , la période de retour T et la probabilité au dépassement π

1.2 Estimation du débit de période de retour T

Plusieurs modèles ont été développés pour déterminer la relation entre le débit d'une rivière et sa période de retour. Tous ces modèles ont besoin d'être calibrés à l'aide de mesures de l'évolution du débit en fonction du temps. On peut utiliser la série des débits maximums annuels (DMA) pour effectuer la calibration. Évidemment, si d'autres informations sont disponibles, il est possible de les combiner à la série des débits maximums annuels. Par exemple, on pourrait connaître approximativement l'amplitude de certaines crues qui ne font pas partie de la série des crues mesurées, mais qui ont laissé leur marque dans la mémoire des hommes (crues historiques) ou dans la géologie du lit de la rivière (paléocrues). [STEDINGER et COHN 1986] ont montré que l'utilisation de telles informations contribue à améliorer la précision de l'estimation du débit correspondant à une période de retour T . Au Québec cependant, il n'existe que très peu d'information sur les crues historiques.

On peut aussi considérer l'information disponible à des sites avoisinant. Plusieurs méthodes dites régionales permettent d'utiliser ce type d'information pour améliorer la précision de l'estimation du débit de période de retour T . On peut voir à ce sujet entre autres les études de [ACREMAN 1987], [CAVADIAS 1990] ou encore [GINGRAS *et al.* 1994]. De façon à limiter l'ampleur de ce travail, nous supposerons ici que la seule source d'information utilisée pour estimer le débit de période de retour T est une série de débits maximums annuels.

En général les valeurs mesurées de débit maximum annuel peuvent être considérées comme les résultats de la répétition d'une même expérience aléatoire. Étant homogènes, elles ont toutes la même distribution statistique. De plus, vu la courte mémoire des systèmes météorologiques, les valeurs obtenues pour deux années distinctes sont généralement indépendantes. La série des débits maximums annuels mesurés peut donc être vue comme un ensemble de variables aléatoires indépendantes et homogènes. Notons que les hypothèses d'indépendance et d'homogénéité peuvent respectivement être vérifiées par les tests non paramétriques de [WALD et WOLFOWITZ 1943] et de [MANN et WHITNEY 1947].

Selon ces hypothèses, le débit maximum annuel d'une année quelconque est une variable aléatoire X , évidemment continue, issue d'une population P ayant une fonction de densité de probabilité cumulée $F(\cdot)$ telle que:

$$F(x) = \Pr[X \leq x] \quad (3)$$

En combinant les équations (1) et (3), on montre que la relation entre le débit x_T et sa période de retour T est liée à la densité cumulée $F(\cdot)$ de la population des débits maximums annuels:

$$F(x_T) = 1 - 1/T \quad (4)$$

Puisque l'on cherche le débit x_T correspondant à une période de retour T fixée, on devra déterminer l'inverse de la densité cumulée $F(\cdot)$:

$$x_T = F^{-1}(1 - 1/T) \quad (5)$$

L'estimation du débit x_T de période de retour T nécessite donc de connaître la distribution statistique de la population des débits maximums annuels.

1.3 Approximation de la distribution des débits

Nous avons montré que l'estimation de la crue de conception correspondant à un risque de défaillance fixé pour un ouvrage hydraulique est directement liée à l'approximation de la distribution statistique des débits maximums annuels. Nous montrerons maintenant comment la crue de conception peut être approximée à partir d'un échantillon \mathbf{X} de mesures du débit maximum annuel $\{x_1, x_2, \dots, x_N\}$ (où N est le nombre d'années passées pour lesquelles on connaît le débit maximum annuel). Nous étudierons une approche non paramétrique et une approche paramétrique.

1.3.1 Approche non paramétrique

Plusieurs méthodes non paramétriques d'approximation de la distribution des débits maximums annuels sont basées sur l'ordre qu'occupe chaque observation dans l'échantillon classé des débits mesurés. Si l'on note l'échantillon \mathbf{X} ordonné en ordre croissant $\{x_{(1)} < x_{(2)} < \dots < x_{(N)}\}$, on montre que la densité de probabilité cumulée $F(x_{(i)})$ de la population des débits peut être approximée par une fonction de i et de N [CUNNANE 1978]. En particulier, plusieurs approximations non paramétriques sont de la forme suivante:

$$F(x_{(i)}) \approx \frac{i - a}{N + 1 - 2a} \quad (6)$$

où a est une constante. On nomme ce type d'approximation non paramétrique "formule de probabilité empirique", ou FPE. Par exemple, la FPE de Weibull, recommandée par [CHOW 1953] pour les données hydrologiques, est obtenue en posant $a=0$:

$$F(x_{(i)}) \approx \frac{i}{N + 1} \quad (7)$$

L'estimation de $F(\cdot)$ nécessite le choix d'une FPE. La densité cumulée peut ensuite être obtenue par interpolation (habituellement linéaire) de la table donnant la relation entre $x_{(i)}$ et $F(\cdot)$. Nous nommerons cette relation la densité cumulée empirique d'un échantillon.

L'adéquation d'une FPE particulière dépend de la distribution statistique de la population des débits maximums annuels. Conséquemment, même en utilisant une approche non paramétrique pour estimer $F(\cdot)$, on doit faire indirectement une hypothèse sur la distribution statistique de cette population, de façon à choisir la FPE la plus adéquate. Pour éviter d'avoir à identifier la distribution statistique, on peut utiliser la FPE de Cunnane, dont les performances sont peu influencées par la distribution statistique des débits maximums annuels [CUNNANE 1978]. On obtient la FPE de Cunnane en fixant $a=0,4$:

$$F(x_{(i)}) = \frac{i - 0,4}{N + 0,2} \quad (8)$$

L'utilisation des FPE pour estimer $F(\cdot)$ se limite aux cas où $1-1/T$ est inférieur à $F(x_{(N)})$, puisque l'on procède par interpolation pour estimer la valeur de $F^{-1}(1-1/T)$. De plus, les FPE sont imprécises pour les valeurs extrêmes d'un échantillon. Conséquemment, l'estimation non paramétrique du débit de période de retour T est imprécise lorsque T est voisin de N . En raison de la taille réduite des échantillons disponibles, la période de retour de la crue de conception à déterminer est souvent trop élevée pour permettre l'utilisation d'une approche non paramétrique ($T > N$). Notons enfin qu'il existe d'autres approches non paramétriques que les FPE, comme l'utilisation de noyaux paramétriques [ADAMOWSKI 1985]. Aucune approche non paramétrique ne permet cependant d'estimer x_T lorsque T est beaucoup plus grand que N .

1.3.2 Approche paramétrique

Il est aussi possible d'approximer la distribution des débits maximums annuels en ajustant une distribution statistique paramétrique aux observations disponibles du débit maximum annuel. Dès le début du siècle, la distribution normale a été utilisée à cet effet [HORTON 1913]. Par la suite, pour améliorer la qualité des ajustements et tenir compte de l'asymétrie des distributions observées, plusieurs autres distributions ont été utilisées, dont la

distribution log-normale, les distributions Pearson type 3 et log-Pearson type 3 (LP3), la distribution Gumbel, et la distribution généralisée des valeurs extrêmes (*Generalized Extreme Value* ou GEV) [KIRBY et MOSS 1987].

La première étape de cette approche paramétrique consiste à choisir une distribution D pour représenter les débits maximums annuels. Chaque distribution paramétrique est caractérisée par une fonction de densité de probabilité cumulée différente notée $F(x|\underline{\theta})$, où $\underline{\theta}=[\theta_1, \theta_2, \dots, \theta_p]^T$ car elle dépend de un ou plusieurs paramètres. Il faut ensuite obtenir les estimateurs $\hat{\underline{\theta}} = [\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p]^T$ des paramètres $\underline{\theta}$ à partir de l'échantillon des débits maximums annuels mesurés, en utilisant une méthode d'ajustement M . Il est alors possible d'estimer, analytiquement ou numériquement, la fonction de densité de probabilité (f.d.p) cumulée inverse par $F^{-1}(1-1/T|\hat{\underline{\theta}})$, et donc d'approximer le débit x_T correspondant à une période de retour T . Les méthodes d'ajustement classiques, décrites à l'annexe A, sont la méthode du maximum de vraisemblance (MV) et la méthode des moments (MM). Notons que la méthode MV est en général la plus efficace pour de grands échantillons, mais peut se révéler inefficace pour bien des séries de débits maximums annuels, puisque celles-ci sont en général assez courtes.

L'approche paramétrique pour l'estimation du débit x_T de période de retour T est décrite par la figure 2. Le formalisme employé est simple: une boîte rectangulaire représente un processus, alors qu'une boîte aux coins arrondis représente une donnée. Une superposition de boîtes aux coins arrondis représente un vecteur. La figure 2 peut donc être interprétée ainsi: à partir de la période de retour T considérée et de la série des débits maximums annuels (DMA), on sélectionne une combinaison d'une distribution statistique D et d'une méthode M d'ajustement des paramètres de cette distribution en utilisant, potentiellement, divers outils statistiques pour comparer les différentes combinaisons D/M . Cette combinaison est utilisée pour approximer la population des débits maximums annuels par $F(x|\hat{\underline{\theta}})$, permettant ainsi d'estimer le débit x_T de période de retour T .

Il est clair que l'estimation \hat{x}_T du débit x_T de période de retour T obtenue à l'aide de cette approche est influencée non seulement par le choix de la distribution D mais aussi par le choix de la méthode d'ajustement M . Le choix d'un couple D/M est la véritable difficulté

rencontrée lors de l'estimation du débit x_T de période de retour T lorsqu'il est nécessaire d'utiliser un modèle paramétrique.

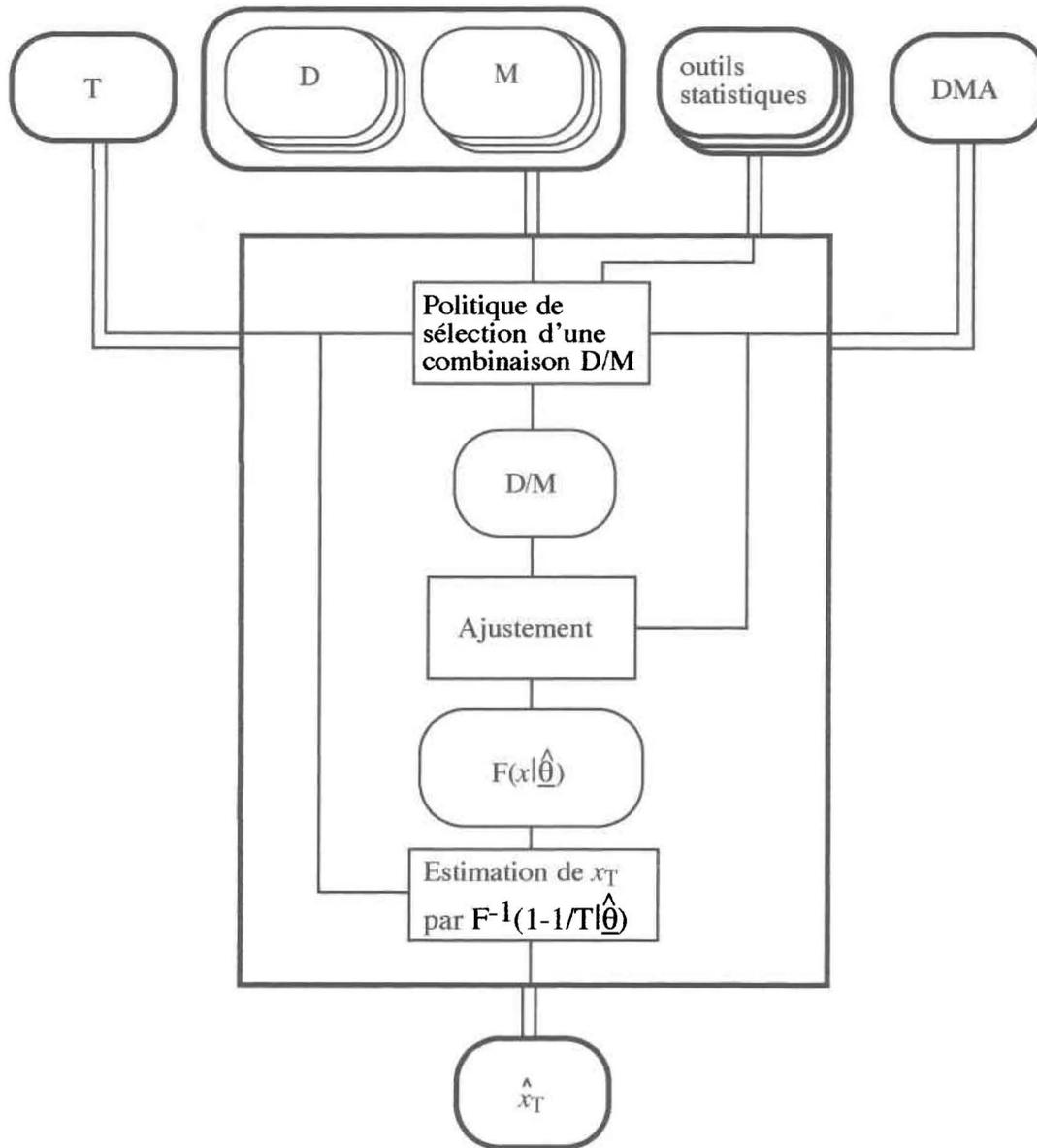


Figure 2: Étapes menant à l'estimation d'une crue de conception

1.4 Sélection d'une combinaison D/M

Le choix de la distribution statistique D et de la méthode d'ajustement des paramètres M à utiliser pour décrire les séries de débits maximums annuels est la difficulté majeure rencontrée lors de l'estimation de x_T . Parce que la crue d'une rivière est la résultante d'une multitude de facteurs et de phénomènes physiques complexes (conditions météorologiques, caractéristiques physiques du bassin versant, contribution des eaux souterraines, aménagement du territoire, ...) il est en pratique impossible de déterminer théoriquement la distribution statistique réelle du débit maximum annuel. Si certaines distributions reposent sur des hypothèses physiques, ces hypothèses ne sont pas vérifiées (ou vérifiables) à partir des observations effectuées jusqu'à maintenant.

Par exemple, [CHOW 1954] a montré que si l'on fait l'hypothèse que les crues sont causées par un grand nombre de phénomènes aléatoires multiplicatifs indépendants, elles doivent suivre une distribution log-normale. Cependant, même si l'on accepte que les crues sont causées par une combinaison de phénomènes aléatoires, aucune observation n'appuie l'hypothèse que ces phénomènes soient indépendants et multiplicatifs. La distribution généralisée des valeurs extrêmes (GEV) a aussi été proposée pour les hypothèses qui la soutiennent. En effet, on montre [GUMBEL 1957] que la valeur maximale d'un échantillon de grande taille d'observations aléatoires indépendantes et identiquement distribuées suit nécessairement une des trois lois des valeurs extrêmes qui ont été regroupées par la distribution GEV [JENKINSON 1969]. Cependant, bien que le débit maximum annuel soit la valeur maximale d'un assez grand nombre (365) de débits journaliers, ceux-ci ne sont ni indépendants, ni identiquement distribués [OMM 1989].

Toutes les distributions statistiques utilisées pour estimer x_T sont donc des modèles empiriques de la distribution statistique réelle inconnue des débits maximums annuels. Conséquemment, l'utilisation d'une combinaison D/M particulière pour estimer x_T à un site donné ne peut être justifiée que si elle s'ajuste bien à la relation entre les observations du débit maximum annuel et leur probabilité empirique (obtenue à l'aide d'une FPE, et que l'on nomme la densité cumulée empirique), spécialement dans la plage de probabilités correspondant à la période de retour T d'intérêt.

Cependant, parce que les séries de mesures du débit maximum annuel sont en général assez courtes et parce que l'on s'intéresse habituellement à de grandes périodes de retour, il est très difficile de comparer la qualité des ajustements de plusieurs combinaisons D/M car dans ces cas la densité cumulée empirique est peu précise et ne couvre pas l'ensemble des probabilités au dépassement d'intérêt. Même les tests d'ajustement du Chi-deux et de Kolmogorov-Smirnov [HAAN 1977] ne sont en général pas assez puissants pour discriminer entre les combinaisons D/M [BOBÉE et ASHKAR 1991].

Le choix d'une combinaison D/M reste donc une procédure partiellement arbitraire. En ingénierie, ce type de situation présente des inconvénients importants [THOMAS 1985, BOBÉE *et al.* 1993b]:

- des ressources considérables peuvent être consacrées au choix de la combinaison D/M à utiliser lors de chaque projet nécessitant l'estimation d'une crue de conception;
- le responsable d'un projet peut être tenté d'utiliser une combinaison D/M qui confirme son idée a priori de l'amplitude de la crue de conception et ainsi camoufler son intuition derrière une démarche qui semble rigoureuse;
- une combinaison D/M surestimant largement la crue de conception peut être choisie pour influencer la répartition des ressources monétaires entre plusieurs sites;
- lors d'une défaillance d'un ouvrage hydraulique, l'estimation de la crue de conception pourrait être mise en doute et le responsable du projet accusé de négligence professionnelle.

Considérant cet état de chose, plusieurs pays ont cru nécessaire d'imposer l'utilisation d'une même distribution pour toutes les situations hydrologiques. Au Royaume-Uni, la distribution GEV a été recommandée [NERC 1975], alors qu'aux États-Unis [USWRC 1967, USWRC 1981] et en Australie [IEA 1977, 1987], la distribution LP3 a été imposée. Au Canada, par contre, une étude du Conseil National de Recherche du Canada [CNRC 1990] a conclu qu'aucune distribution n'était applicable à toutes les situations hydrologiques.

Pour éviter les dangers cités plus haut, il serait utile de disposer, pour l'analyse des séries de débits maximums annuels canadiennes, d'une politique de sélection de la combinaison D/M à utiliser à un site donné pour déterminer la combinaison D/M la plus appropriée. Il est bien important de distinguer 'politique de sélection' et 'combinaison D/M'. Une politique de sélection peut cependant consister en l'application systématique d'une même combinaison D/M. Nous nommerons 'politique statique' ce type de politique. Il est aussi possible de définir des politiques de sélection, que nous nommerons 'politiques dynamiques' tenant compte des caractéristiques statistiques des observations disponibles, comme leur asymétrie. Par exemple, une politique dynamique de sélection pourrait être de sélectionner la combinaison [distribution normale/méthode MV] lorsque $|\hat{C}_s| < 0,1$ et la combinaison [distribution Pearson type III/méthode MV] lorsque $|\hat{C}_s| \geq 0,1$. (Rappelons que \hat{C}_s est l'estimateur du coefficient d'asymétrie, tel que défini à l'annexe A).

Un grand nombre d'auteurs ont proposé de tels modèles dynamiques de sélection d'une combinaison D/M. Ces modèles se distinguent par les combinaisons D/M considérées, par les outils statistiques utilisés pour effectuer la sélection, par les variables de décision retenues, par la structure de l'algorithme de sélection ainsi que par l'approche utilisée pour calibrer le modèle. Par exemple [KUBLER 1979] propose d'utiliser une combinaison de tests de rapport des vraisemblances maximales [DUMONCEAUX et ANTLE 1973] pour comparer certaines distributions à trois paramètres (gamma généralisée, log-normale et Weibull). [CHOW et WATT 1990] discutent de l'utilisation d'un système-expert incorporant entre autres le test de corrélation des probabilités empiriques (*Probability Plot Correlation Coefficient*, ou PPCC) décrit par [VOGEL 1986] et le critère d'information d'Akaike (*Minimum Akaike Information Criterion Estimate*, ou MAICE) [AKAIKE 1974]. [LU et STEDINGER 1992] proposent de différencier les distributions GEV et Gumbel sur la base de l'estimation du coefficient d'asymétrie, de la taille d'échantillon et de la période de retour. Enfin [TAKARA et STEDINGER 1993] proposent d'utiliser l'estimation du coefficient d'asymétrie en combinaison avec des tests d'adéquation pour sélectionner une combinaison D/M.

Le choix d'une politique de sélection adéquate ne peut se faire que si l'on dispose d'un mécanisme pour les comparer. Nous discuterons dans la suite de ce travail de l'élaboration d'une méthode de comparaison de ces politiques de sélection.

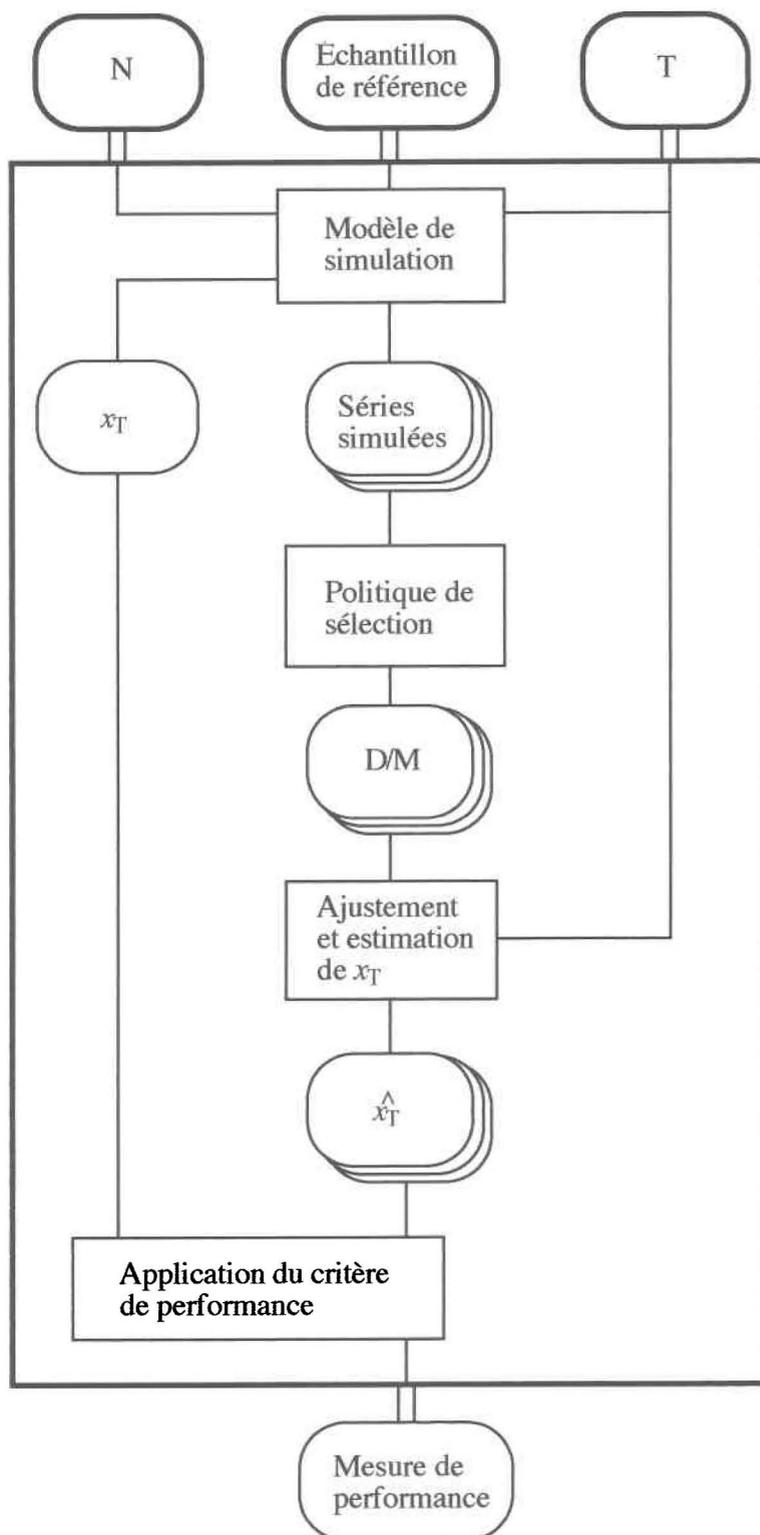


Figure 3: Estimation de la performance d'une politique de sélection d'une combinaison D/M

1.5 Comparaison des politiques de sélection d'une combinaison D/M

Nous avons vu que pour des raisons économiques et légales il serait utile de disposer au Canada d'une politique de sélection de la combinaison D/M à utiliser à un site qui ne nécessiterait pas de décision arbitraire de la part des responsables de projets. Cependant, pour pouvoir choisir la politique de sélection la plus performante, il est nécessaire de disposer d'un moyen de les comparer.

La figure 3 présente une approche générale permettant de mesurer la précision avec laquelle une politique de sélection permet d'estimer le débit x_T de période de retour T à partir de séries de débits maximums annuels de taille N ressemblant statistiquement à une série donnée, que l'on nommera échantillon de référence. Ce type d'approche nécessite de disposer d'un modèle de simulation de débits capable de simuler un grand nombre de séries de taille N ressemblant statistiquement à l'échantillon de référence, et capable de prédire la valeur exacte du débit x_T des séries simulées. Une fois les séries simulées, la politique de sélection leur est appliquée tour à tour, permettant ainsi d'obtenir un ensemble d'estimations de x_T . Ces estimations sont alors comparées à la valeur exacte selon un critère de performance fixé au préalable, permettant ainsi d'obtenir une mesure de la performance de la politique de sélection. En comparant les mesures de performance de différentes politiques de sélection, il est alors possible de déterminer la meilleure.

Les résultats de ce type de comparaison dépendent de tous les intrants au processus: la taille d'échantillon (N), la période de retour (T), l'échantillon de référence, et sont aussi influencés par le modèle de simulation et par le critère de performance utilisé. Pour simuler des données à partir d'un échantillon de référence, deux approches ont été envisagées: la simulation par suréchantillonnage [EFRON 1979, POTTER et LETTENMAIER 1990, BERNIER 1993] et la simulation à l'aide d'une distribution parente [SLACK *et al.* 1975, KUCZERA 1982, WALLIS et WOOD 1985]. Différents critères de performance peuvent aussi être utilisés. On retrouve dans la plupart des études de comparaison des critères classiques tels le biais, l'écart relatif et l'écart quadratique moyen (EQM) [USWRC 1976, KUCZERA 1982, WALLIS et WOOD 1985, HAKTANIR 1992, HAKTANIR et HORLACHER 1993]. D'autres auteurs ont considéré des critères de performance basés sur le coût réel de l'erreur d'estimation de x_T

par \hat{x}_T [SLACK *et al.* 1975, POTTER et LETTENMAIER 1990]. Il n'existe pas en fait de protocole bien établi pour estimer la performance d'une politique de sélection d'une combinaison D/M [BOBÉE *et al.* 1993b].

Alors que de nouvelles distributions et méthodes d'estimation (donc de nouvelles politiques de sélection) sont introduites à un rythme élevé [KLEMES 1986], peu d'efforts sont consacrés à établir leur supériorité sur les méthodes existantes [LETTENMAIER 1988]. Conséquemment, malgré l'effort considérable mis à développer des méthodes statistiques plus performantes, très peu de ces méthodes sont adoptées par les praticiens [PILGRIM 1986].

Pour remédier en partie à cet état de choses, nous étudierons dans ce travail les différentes approches utilisées pour comparer des politiques de sélection et nous composerons, à partir des meilleurs éléments des approches étudiées, un protocole de comparaison basé sur des hypothèses raisonnables et vérifiables. Nous discuterons aussi d'une généralisation de ce protocole dans le cas où l'on cherche à appliquer une politique de sélection à un ensemble de séries de débits maximums annuels ne provenant pas d'une région statistiquement homogène, c'est à dire ne ressemblant pas toutes statistiquement à un même échantillon de référence.

Pour donner un exemple d'application du protocole proposé, nous le mettrons en pratique en utilisant quelques séries de débits maximums annuels du Québec et de l'Ontario. Nous nous limiterons, pour cette application, à comparer des politiques de sélection simples consistant en l'application systématique de la même combinaison D/M indépendamment de l'échantillon considéré.

Nous montrerons aussi qu'en raffinant le protocole de comparaison proposé, il est possible de développer une politique de sélection d'une combinaison D/M suivant la même philosophie. Nous évaluerons enfin la performance de cette nouvelle politique de sélection à l'aide du protocole proposé.

2 ASPECTS THÉORIQUES

Nous avons vu au chapitre précédent qu'il serait utile de standardiser le processus de sélection de la distribution statistique D et de la méthode d'ajustement M à utiliser pour estimer un débit x_T de période de retour T . Parce qu'une multitude de politiques de sélection peuvent être appliquées (de l'emploi systématique de la même combinaison D/M à l'utilisation d'un système expert), il est nécessaire de disposer d'un moyen de les comparer.

L'objectif de ce chapitre est d'asseoir les bases de la méthodologie que nous proposons d'employer pour comparer la performance de diverses politiques de sélection d'une combinaison D/M pour estimer un débit x_T de période de retour T . Comme nous l'avons montré au chapitre précédent (figure 3), il est possible de déterminer quelle politique de sélection est la plus performante pour des échantillons ressemblant statistiquement un échantillon de référence si l'on dispose d'un modèle de simulation de débits ainsi que d'un critère de performance.

Nous discuterons d'abord des différents modèles de simulation existants et examinerons différentes définitions du critère de performance. Nous verrons ensuite comment la performance d'une politique de sélection peut être estimée lorsque l'on considère son application à un ensemble de séries ne ressemblant pas tous statistiquement à un même échantillon de référence, c'est à dire pour des séries non hydrologiquement homogènes. Nous verrons enfin comment la procédure de comparaison proposée peut mener à la conception d'une politique de sélection basée sur la même philosophie.

2.1 Simulation de débits maximums annuels

L'application de la procédure de comparaison de politiques de sélection d'une combinaison D/M pour estimer x_T présentée précédemment nécessite de pouvoir simuler des séries de débits ressemblant statistiquement à un échantillon de référence. Nous étudierons ici deux approches permettant de réaliser cet objectif: l'utilisation d'une distribution parente et le suréchantillonnage.

2.1.1 Simulation à l'aide d'une distribution parente

Cette approche paramétrique permettant de simuler des séries de débits maximums annuels est couramment utilisée en hydrologie. Présentée par [SLACK *et al.* 1975], elle a été utilisée par une multitude d'auteurs, dont [KUCZERA 1982, WALLIS et WOOD 1985, DUCKSTEIN *et al.* 1991, LU et STEDINGER 1992, HAKTANIR 1992, HAKTANIR et HORLACHER 1993]. Pour simuler des séries ressemblant statistiquement à un échantillon de référence à l'aide de cette approche, on sélectionne d'abord un modèle paramétrique assez flexible pour pouvoir représenter de façon adéquate la distribution des observations de l'échantillon de référence. On ajuste ensuite les paramètres de ce modèle de façon à l'adapter à l'échantillon de référence, et l'on génère des observations aléatoires provenant de ce modèle.

Puisque la distribution réelle des débits maximums annuels est en réalité inconnue [BOBÉE *et al.* 1993b], le choix d'un modèle adéquat est une tâche plutôt ardue. Il est cependant reconnu que les distributions utilisées pour ajuster des séries de débits maximums annuels dans le but d'estimer x_T , qui n'ont en général que deux ou trois paramètres, ne sont pas assez générales pour inclure la distribution réelle des débits maximums annuels [HOUGHTON 1978, KUCZERA 1982]. Il est donc nécessaire d'utiliser un modèle plus flexible, donc ayant plus de paramètres.

Les distributions Kappa [HOSKING et WALLIS 1993] et Wakeby [HOUGHTON 1978], qui ont respectivement quatre et cinq paramètres, sont couramment utilisées comme distributions parentes. Cependant, même si ces distributions sont plus générales que les distributions usuelles à deux ou trois paramètres, il n'est en rien assuré que cette généralisation les rend capable de modéliser les observations de débit maximum annuel, puisqu'il existe une variété infinie de façon de généraliser une distribution paramétrique [KUCZERA 1982, STEDINGER et COHN 1986, POTTER et LETTENMAIER 1990].

Même si l'on arrive à déterminer une distribution parente acceptable, il n'est pas aisé de fixer les paramètres pour pouvoir simuler des observations ressemblant statistiquement à celles d'un échantillon de référence. En effet, les paramètres de la distribution parente doivent être

estimés, et vu le grand nombre de paramètres des distributions parentes la précision de ces estimations est évidemment très limitée.

L'utilisation d'une distribution parente pour simuler des débits maximums annuels demande d'effectuer des hypothèses non vérifiables, et est donc difficilement justifiable. Malgré cela, les simulations effectuées en utilisant une distribution parente sont souvent considérées précises et suffisantes pour tirer des conclusions indiscutables. Par exemple, après avoir montré que sur la base de ce type de simulation la politique de sélection recommandée par le *United States Water Resources Council* [USWRC 1967, 1981] démontrait de piètres performances, [WALLIS et WOOD 1985] ont affirmé qu'un ingénieur qui continuerait à utiliser la politique du USWRC pourrait être trouvé coupable de négligence professionnelle. Notons que cet article s'est attiré une critique sévère [BEARD 1987, LANDWEHR *et al.* 1987, BOBÉE *et al.* 1993b].

Parce que notre objectif principal est de développer une méthode de comparaison de politiques de sélection qui soit basée sur des hypothèses raisonnables et vérifiables, nous croyons que la simulation à l'aide d'une distribution parente est une approche à éviter.

2.1.2 Simulation par suréchantillonnage

La simulation par suréchantillonnage, ou *Bootstrap*, introduite par [EFRON 1979], est une technique simple permettant de simuler un grand nombre d'échantillons d'observations d'un phénomène aléatoire à partir d'un échantillon de référence sans qu'il soit nécessaire de spécifier la forme de la distribution statistique du phénomène. Pour simuler un échantillon de taille n en utilisant la technique *Bootstrap*, on tire simplement avec remise n observations dans l'échantillon de référence. L'algorithme suivant permet donc de simuler un échantillon $\mathbf{X}=\{x_1, x_2, \dots, x_n\}$ d'observations d'un phénomène aléatoire à partir d'un échantillon de référence $\mathbf{Y}=\{y_1, y_2, \dots, y_N\}$:

POUR i variant de 1 à n RÉPÉTER les instructions suivantes:
 CHOISIR un nombre j au hasard entre 1 et N
 SOIT $x_i=y_j$

Figure 4: Algorithme de simulation par suréchantillonnage *Bootstrap*

Cette technique est d'une séduisante simplicité, mais sa validité n'a pas été démontrée dans le contexte où nous voulons l'utiliser. L'insuffisance des bases théoriques de cette approche amène les utilisateurs à modifier la technique selon leurs intuitions. Par exemple, [POTTER et LETTENMAIER 1990], qui ont effectué une étude de comparaison par suréchantillonnage de quelques méthodes régionales d'estimation des crues ont choisi de tirer les observations sans remise parce qu'ils ne trouvaient pas sensé qu'une même observation se retrouve plusieurs fois dans le même échantillon.

Il serait utile de disposer d'une interprétation de la technique *Bootstrap* basée sur une théorie plus solide. [BERNIER 1993] a montré qu'une version modifiée la technique *Bootstrap* pouvait se justifier dans un cadre bayésien. Bien qu'il ne soit pas approprié de répéter ici la démonstration, nous croyons utile de présenter quelques concepts menant à cette adaptation de la technique *Bootstrap*, puisque les travaux mentionnés plus haut ne sont pas encore publiés.

Ce qui nous intéresse dans la technique *Bootstrap*, c'est de pouvoir prévoir de nouvelles observations $\mathbf{X}=\{x_1, x_2, \dots, x_n\}$ d'un phénomène aléatoire à partir d'observations préalables constituant un échantillon de référence $\mathbf{Y}=\{y_1, y_2, \dots, y_N\}$. Supposons que le phénomène aléatoire soit distribué selon une f.d.p. $f(x|\underline{\theta})$, où $\underline{\theta}$ est un vecteur de paramètres. Supposons encore qu'à l'aide de l'échantillon de référence on puisse déterminer une f.d.p. subjective sur $\underline{\theta}$, notée $b(\underline{\theta}|\mathbf{Y})$, représentant les incertitudes qui subsistent sur les valeurs des paramètres du phénomène aléatoire après l'analyse de l'information contenue dans l'échantillon de référence. Dans un contexte bayésien, on montre que les nouvelles observations \mathbf{X} devraient être simulées à partir de la distribution $p(\mathbf{X}|\mathbf{Y})$, dite prévisionnelle ou prépostérieure, qui intègre l'incertitude sur la valeur des paramètres [PRESS 1988]:

$$p(\mathbf{X}|\mathbf{Y}) = \int_{\Theta} \{f(x_1|\Theta) \cdot f(x_2|\Theta) \cdot \dots \cdot f(x_n|\Theta)\} b(\Theta|\mathbf{Y}) \cdot d\Theta \quad (9)$$

[BERNIER 1993] a précisément montré que la simulation par tirage de Pólya dans l'échantillon de référence est équivalente à la simulation à partir de la distribution prépostérieure. Rappelons que le tirage de Pólya est un schéma d'urne. Soit une urne contenant N boules de différentes couleurs. Pour construire un échantillon d'observations du contenu de l'urne par tirage de Pólya on procède comme suit. On tire d'abord une première boule dans l'urne et on note sa couleur; c'est la première observation. On replace ensuite la boule dans l'urne *et on ajoute une boule de la même couleur*. On répète ensuite ce procédé de tirage autant de fois que nécessaire. La simulation d'un échantillon $\mathbf{X}=\{x_1, x_2, \dots, x_n\}$ d'observations d'un phénomène aléatoire à partir d'un échantillon de référence $\mathbf{Y}=\{y_1, y_2, \dots, y_N\}$ par tirage de Pólya peut donc être effectuée en appliquant l'algorithme suivant:

POUR i variant de 1 à $N+i-1$ RÉPÉTER les instructions suivantes:
 CHOISIR un nombre j au hasard entre 1 et $N+i-1$
 SOIT $x_i=y_j$
 AJOUTER une observation $y_{N+i-1}=y_j$ à l'échantillon \mathbf{Y}

Figure 5: Algorithme de simulation par tirage de Pólya

Dans la mesure où le nombre N d'observations contenues dans l'échantillon de référence est grand par rapport au nombre n d'observations de l'échantillon simulé, [BERNIER 1993] montre que le tirage de Pólya est approximativement équivalent au *Bootstrap* classique. Cependant, l'implantation du second algorithme n'est guère plus compliquée. Nous utiliserons donc cette dernière méthode pour simuler des échantillons de débits maximums annuels.

Parce que le *Bootstrap* modifié par tirage de Pólya n'a jamais à notre connaissance été appliqué, du moins à des données hydrologiques, nous croyons qu'il est nécessaire, pour en comprendre l'avantage, de le comparer par un exemple simple au *Bootstrap* classique, tel que proposé par [EFRON 1979]. Cette digression, non essentielle à la compréhension de la

suite de ce travail, est exposée à l'annexe B. Nous devons cependant mentionner une hypothèse sous-jacente qui conditionne la méthode de tirage de Pólya. Il s'agit de la forme de la distribution *a priori* $b(\underline{\theta})$, utilisée pour représenter l'incertitude sur les paramètres $\underline{\theta}$ lorsque aucune information n'est disponible pour estimer ces paramètres (c'est à dire une distribution *a priori* sur la classe de modèles englobant le modèle exact du phénomène étudié). On peut montrer que le processus de Dirichlet permet de construire une distribution *a priori* rencontrant un ensemble de conditions de cohérences nécessaires [BERNIER 1993]. Bien qu'il s'agisse d'un choix arbitraire, son importance s'atténue lorsqu'un nombre suffisant d'observations est disponible pour déterminer la distribution prépostérieure.

L'utilisation de la technique *Bootstrap* modifiée par tirage Pólya pour simuler des échantillons de débits maximums annuels a une base solide dans le cas qui nous intéresse. On ne peut cependant appliquer cette technique que pour simuler des séries de taille n inférieure à celle de l'échantillon de référence (N) [BERNIER 1993].

Pour appliquer la procédure de comparaison de politiques de sélection présentée à la figure 3, le modèle de simulation doit aussi permettre de déterminer la valeur exacte de x_T , de façon à permettre l'estimation de la performance de chaque politique. Parce que la simulation par suréchantillonnage ne nécessite pas de spécifier complètement la forme de la distribution du phénomène simulé, il est impossible d'obtenir cette valeur. Si la période de retour T est inférieure à la taille N de l'échantillon de référence, il est cependant possible d'estimer x_T à l'aide d'une approche non paramétrique, telle une formule de probabilité empirique (FPE).

Ainsi, il n'est possible d'utiliser la technique *Bootstrap* pour déterminer la performance de politiques de sélection que pour des périodes de retour T et des tailles n de séries inférieures à celle de l'échantillon de référence. Ceci n'est pas en soi surprenant puisque l'approche utilisée est entièrement non paramétrique. En effet, puisque l'on n'utilise aucune autre information qu'un échantillon de référence pour approximer la distribution des crues, il n'est pas possible d'étudier le comportement des politiques de sélection pour des périodes d'observation dépassant celle de l'échantillon de référence. Puisque nous ne connaissons les processus donnant naissance aux crues que de façon empirique, il est probablement prudent, de toutes façons, d'éviter d'effectuer des prédictions pour des périodes de retour ou des tailles d'échantillon supérieures à la période réelle d'observation du phénomène.

2.2 Mesure de la performance d'une politique de sélection d'une combinaison D/M

Selon le modèle de comparaison détaillé au chapitre précédent (figure 3), la performance d'une politique de sélection d'une combinaison D/M est évaluée en comparant un grand nombre d'estimations de x_T obtenues en appliquant cette politique, avec la valeur prédite par le modèle de simulation, que nous nommerons "valeur correcte". Dans le cas du modèle de simulation par tirage Pólya présenté plus haut, la valeur correcte de x_T est obtenue à l'aide d'une formule de probabilité empirique (FPE) appliquée à l'échantillon de référence.

Différentes mesures de performance peuvent être employées. Des mesures de performance classiques, comme le biais et la variance des estimations ainsi que la somme des écarts au carré entre les estimations de x_T et la valeur correcte, sont utilisées dans la majorité des études [USWRC 1967, KUCZERA 1982, WALLIS et WOOD 1985, HAKTANIR 1992, HAKTANIR et HORLACHER 1993]. Le principal défaut de ces mesures est d'accorder automatiquement le même poids aux sous-estimations ($\hat{x}_T < x_T$) et aux surestimations ($\hat{x}_T > x_T$). Or, ces deux éventualités n'entraînent pas du tout les mêmes conséquences lorsque l'estimation de x_T est utilisée pour dimensionner un ouvrage hydraulique. En effet, une sous-estimation de x_T entraînera une augmentation de la fréquence des défaillances de l'ouvrage alors qu'une surestimation causera un surdimensionnement de l'ouvrage. Il serait préférable d'utiliser une mesure de performance qui tient compte de cette réalité.

Pour mesurer la performance d'une seule estimation \hat{x}_T du débit x_T , [SLACK *et al.* 1975] proposent à cet effet l'utilisation d'une mesure de performance plus générale, justifiée par le raisonnement suivant. Considérons un ouvrage hydraulique qui doit être construit pour résister à un débit critique x_T de période de retour T . Supposons que l'on connaît exactement la valeur de ce débit critique. On peut alors déterminer la façon la moins coûteuse, que l'on nomme solution optimale, de réaliser l'ouvrage. Soit le coût moyen annuel de cette solution égal à C^* . Ce coût peut se diviser en deux composantes: C_o , le coût moyen annuel d'opération de l'ouvrage (incluant l'amortissement des coûts de construction), et C_d , le coût moyen annuel des dégâts causés par les défaillances de l'ouvrage:

$$C^* = C_o + C_d \quad (10)$$

En pratique, on ne connaît pas x_T . On doit dimensionner l'ouvrage à partir d'une estimation du débit critique \hat{x}_T . Si $\hat{x}_T = x_T$, alors le coût réel C sera bien égal à C^* . Par contre, toute erreur d'estimation de x_T entraînera un coût réel $C^* + Z$ supérieur ou égal à C^* , où Z est la perte nette occasionnée par l'erreur d'estimation du débit critique x_T . Alors que C^* est une valeur fixe, Z est relié à l'écart entre x_T et \hat{x}_T .

Supposons que x_T soit surestimé par un facteur Δx_T , tel que $\hat{x}_T = x_T + \Delta x_T$. Cela entraînera un surdimensionnement de l'ouvrage, augmentera le coût moyen annuel d'opération d'un facteur ΔC_o et diminuera le coût moyen annuel des défaillances d'un facteur ΔC_d , pour un coût annuel total de $(C_o + \Delta C_o) + (C_d - \Delta C_d)$. Par contre, une sous-estimation de x_T par un facteur $-\Delta x_T$ entraînera un sous-dimensionnement de l'ouvrage, diminuant le coût moyen annuel d'opération d'un facteur $\Delta' C_o$ mais augmentant cependant le coût moyen annuel des défaillances d'un facteur $\Delta' C_d$, pour un coût annuel total de $(C_o - \Delta' C_o) + (C_d + \Delta' C_d)$. Soit Z la perte nette encourue par l'erreur d'estimation du débit critique. Cette valeur est fonction de la valeur du débit critique, x_T , et de son estimation, \hat{x}_T :

$$Z = \begin{cases} \Delta' C_d - \Delta' C_o & \text{si } \hat{x}_T < x_T \\ 0 & \text{si } \hat{x}_T = x_T \\ \Delta C_o - \Delta C_d & \text{si } \hat{x}_T > x_T \end{cases} \quad (11)$$

La forme de la fonction Z n'est cependant pas connue. Nous proposons de supposer que Z est de la forme suivante. Notons qu'il s'agit, à une constante près, de la forme proposée par [SLACK *et al.* 1975].

$$Z \propto g(x_T, \hat{x}_T) = \begin{cases} \alpha \left(\frac{x_T - \hat{x}_T}{x_T} \right)^r & \text{si } \hat{x}_T < x_T \\ 0 & \text{si } \hat{x}_T = x_T \\ \left(\frac{\hat{x}_T - x_T}{x_T} \right)^r & \text{si } \hat{x}_T > x_T \end{cases} \quad (12)$$

L'avantage de cette formulation est que la fonction $g(x_T, \hat{x}_T)$ est indépendante de l'unité de mesure des débits et ne dépend que de deux paramètres, α et r , qui peuvent être interprétés de la façon suivante. Le paramètre α détermine le poids relatif des pertes causées par la surestimation de x_T , en comparaison des pertes encourues par sa sous-estimation. Ainsi, si la sous-estimation de x_T engendre, pour une erreur de même amplitude, des coûts deux fois plus élevés que la surestimation, on fixera $\alpha=2$. Le paramètre r détermine plutôt l'importance accordée à des erreurs de grande amplitude. Par exemple, si une erreur d'amplitude deux fois plus grande engendre des coûts huit fois plus élevés, on fixera $r=\log_2(8)=3$.

Il est évident que la fonction Z peut ne pas être du tout de la forme (12) et qu'en pratique il est très ardu d'estimer cette fonction. Nous proposons donc d'utiliser comme fonction de perte nette l'écart quadratique moyen (EQM), obtenu à partir de (12) en fixant $\alpha=1$ et $r=2$, et d'étudier la sensibilité du résultat de la comparaison des politiques de sélection en faisant varier les paramètres α et r .

L'équation (12) permet d'évaluer la performance d'une politique de sélection à un site particulier. Elle peut cependant facilement être généralisée pour le cas où l'on désire mesurer la performance pour un ensemble de R séries issues de la même population, pour lesquelles on dispose des estimations $\{\hat{x}_{T_1}, \hat{x}_{T_2}, \dots, \hat{x}_{T_R}\}$. Dans ce cas, la performance d'une politique de sélection peut être mesurée par la perte moyenne occasionnée par l'application de cette politique pour l'ensemble des séries considérées, que nous noterons φ :

$$\varphi = \frac{1}{R} \sum_{j=1}^R g(x_T, \hat{x}_{T_j}) \quad (13)$$

2.3 Mesure de la performance d'une politique de sélection appliquée à une région non homogène

Le procédé de simulation de séries de débits maximums annuels et le critère de performance présentés précédemment permettent, en conjonction avec la procédure de comparaison décrite par la figure 3, de mesurer la performance de diverses politiques de sélection d'une combinaison D/M pour des séries ressemblant statistiquement à un échantillon de référence, à

condition de s'intéresser à une période de retour T et à une taille de série n inférieure à celle de l'échantillon de référence utilisé. Cependant, lorsqu'une politique de sélection d'une combinaison D/M permettant d'estimer x_T est appliquée à un site donné dans une région non statistiquement homogène, on ne sait pas à quel échantillon de référence il ressemble, et différentes politiques de sélections peuvent être retenues pour différents échantillons de références. Nous proposerons dans cette section une solution simple pour contourner cette difficulté.

Supposons qu'une politique de sélection d'une combinaison D/M doit être appliquée à un ensemble E_X de séries de débits maximums annuels de taille n pour estimer le débit de période de retour T contenues dans une région non hydrologiquement homogène mais constituée de sous-régions homogènes. On entend par sous-région homogène un ensemble de sites pour lesquels le débit maximum annuel suit la même distribution à une transformation d'échelle près. Supposons encore que la fonction de perte $g(x_T, \hat{x}_T)$ soit la même pour tous les sites. Supposons enfin que l'on dispose d'un second ensemble $E_Y = \{Y_1, Y_2, \dots, Y_K\}$ de K séries de taille supérieure à n et T , dispersées de la même façon que les séries de l'ensemble E_X sur le territoire.

Si l'on dispose d'un nombre suffisant de ces longues séries dans chaque sous-région homogène, il est alors raisonnable de présumer que les séries simulées à partir de l'ensemble E_Y seront représentatives des séries de l'ensemble E_X , à une transformation d'échelle près. Ainsi, si l'on note φ_i la perte nette moyenne estimée pour une politique donnée en simulant des séries ressemblant statistiquement à l'échantillon de référence Y_i , on pourra estimer, à une constante près, la perte nette moyenne engendrée par l'utilisation de cette politique pour les séries de l'ensemble E_X par:

$$\bar{\varphi} = \frac{1}{K} \sum_{i=1}^K \varphi_i \quad (14)$$

En pratique, si l'on veut comparer la performance d'un ensemble de procédures de sélection pour une taille d'échantillon n et une période de retour T fixée dans une région non homogène donnée, on sélectionnera toutes les stations de cette région ayant assez

d'observations pour être utilisées comme échantillon de référence (i.e. un nombre N d'observations raisonnablement supérieur à n et T).

Le problème principal avec cette approche est que la distribution des échantillons de référence entre les sous-régions homogènes peut différer de celle des sites auxquels on appliquera la politique de sélection. Considérons par exemple les sous-régions homogènes délimitées par [GINGRAS *et al.* 1994] pour le Québec et l'Ontario. Dans ces provinces, les stations possédant les plus longues séries de mesures du débit sont concentrées dans le sud, alors que les stations plus récentes sont assez dispersées sur le territoire [MATHIER *et al.* 1993]. En appliquant l'équation (14) pour choisir une politique de sélection d'une combinaison D/M pour les stations les plus récentes, on donnerait beaucoup plus d'importance à une estimation précise de x_T pour les stations du sud.

Avant d'utiliser l'approche proposée ici pour estimer la performance d'une procédure de sélection pour des séries non homogènes, nous suggérons à l'utilisateur de s'assurer que les échantillons de référence sont bel et bien représentatifs de l'ensemble des séries auxquelles sera appliquée la politique de sélection. Une façon de procéder est de vérifier si la distribution de certaines caractéristiques statistiques de ces séries, comme le coefficient de variation (C_v) et le coefficient d'asymétrie (C_s) (ces coefficients sont définis à l'annexe A), est similaire à la distribution de ces mêmes caractéristiques pour les séries simulées à partir des échantillons de référence.

2.4 Une politique de sélection découlant de la procédure de comparaison

À partir des idées développées à la dernière section, nous montrerons qu'il est possible de transformer, en la raffinant un peu, la procédure de comparaison de politiques de sélection présentée dans ce chapitre en une nouvelle politique de sélection. L'avantage de cette nouvelle politique est d'être, dans un sens que nous définirons plus loin, optimale par rapport à notre procédure de comparaison.

Supposons que l'on désire estimer un débit x_T de période de retour T à l'aide d'un échantillon \mathbf{X} de n observations du débit maximum annuel, et que l'on hésite entre l'application de c combinaisons D/M notées $[D/M]_1, [D/M]_2, \dots, [D/M]_c$. Dans un contexte d'analyse de la décision, il s'agit en fait de décider quelle action parmi cet ensemble doit être effectuée. Supposons encore que l'on sache que la distribution de ces observations est similaire à la distribution de l'un des échantillons de référence \mathbf{Y}^* inclus dans l'ensemble $\mathbf{E}_Y = \{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_J\}$. On notera cette hypothèse $\mathbf{X} \approx \mathbf{Y}^*$. S'il était possible d'identifier cet échantillon de référence (c'est à dire quelle hypothèse est vraie), on pourrait comparer les diverses politiques consistant en l'application systématique des combinaisons $[D/M]_1, [D/M]_2, \dots, [D/M]_c$ en utilisant la procédure de comparaison présentée au début de ce chapitre. En observant les caractéristiques statistiques de l'échantillon \mathbf{X} , il est possible de se donner une idée de la vraisemblance que la distribution de chaque échantillon de référence \mathbf{Y}^* soit similaire à celle de l'échantillon \mathbf{X} en adoptant un point de vue bayésien. C'est à dire qu'il est possible de déterminer la probabilité de chaque hypothèse $\mathbf{X} \approx \mathbf{Y}^*$, conditionnelle à l'information disponible.

Supposons que l'on dispose d'estimations $\hat{\mathbf{S}} = \{\hat{S}_1, \hat{S}_2, \dots, \hat{S}_s\}$, mesurées à l'aide de l'échantillon \mathbf{X} , d'un ensemble $\mathbf{S} = \{S_1, S_2, \dots, S_s\}$ de caractéristiques statistiques, telles le coefficient de variation (Cv) et le coefficient d'asymétrie (Cs), etc. En observant les valeurs des estimations de $\mathbf{S} = \{S_1, S_2, \dots, S_s\}$ pour des séries, de taille n , simulées à partir des différents échantillons de référence $\{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_J\}$, il est possible d'estimer une probabilité subjective que la distribution de \mathbf{X} soit similaire à chaque échantillon de référence.

Une façon pratique d'atteindre cet objectif est de partitionner l'espace $S_1 \times S_2 \times \dots \times S_s$ en sous-espaces E_1, E_2, \dots, E_w , dans lesquels peuvent être classés les estimations de \mathbf{S} obtenues pour chaque série simulée à partir de chaque échantillon de référence de l'ensemble $\{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_J\}$. En comptabilisant le nombre d'estimations de \mathbf{S} rencontrées dans un sous-espace E_k , pour un échantillon de référence \mathbf{Y}^* , on mesure la probabilité qu'une estimation de \mathbf{S} soit élément d'un sous-espace donné, qui est une estimation de la probabilité que $\hat{\mathbf{S}}$ soit compris dans ce sous-espace si la distribution de \mathbf{X} est similaire à celle de \mathbf{Y}^* , et que l'on note $\text{Pr}[\hat{\mathbf{S}} \in E_k | \mathbf{X} \approx \mathbf{Y}^*]$. On peut aussi facilement déterminer dans quel sous-espace tombe la valeur de $\hat{\mathbf{S}}$ mesurée à l'aide de l'échantillon \mathbf{X} .

On peut donc déterminer la probabilité subjective de l'information disponible ($\hat{S} \in E_k$), conditionnelle à une hypothèse donnée ($\mathbf{X} \approx \mathbf{Y}^*$). Cependant, nous sommes intéressés précisément par l'inverse, c'est à dire la probabilité subjective de chaque hypothèse, conditionnelle à l'information disponible, que l'on note $\Pr[\mathbf{X} \approx \mathbf{Y}^* | \hat{S} \in E_k]$. Il s'agit de la probabilité subjective que la distribution de \mathbf{X} soit similaire à celle d'un échantillon de référence donné, étant donné que \hat{S} fait partie d'un sous-espace particulier. Cette transformation peut être obtenue par application du théorème de Bayes [PRESS 1988]:

$$\Pr[\mathbf{X} \approx \mathbf{Y}^* | \hat{S} \in E_k] = \frac{\Pr[\mathbf{X} \approx \mathbf{Y}^*] \cdot \Pr[\hat{S} \in E_k | \mathbf{X} \approx \mathbf{Y}^*]}{\sum_{j=1}^J \Pr[\mathbf{X} \approx \mathbf{Y}_j] \cdot \Pr[\hat{S} \in E_k | \mathbf{X} \approx \mathbf{Y}_j]} \quad (15)$$

où $\Pr[\mathbf{X} \approx \mathbf{Y}^*]$ est la probabilité subjective *a priori* associée à l'hypothèse que la distribution de \mathbf{X} soit similaire à celle de \mathbf{Y}^* , c'est à dire avant d'observer l'information \hat{S} . On utilisera une distribution uniforme dans ce cas ($\forall \mathbf{Y} \in E_Y, \Pr[\mathbf{X} \approx \mathbf{Y}] = 1/J$), puisque lorsque l'on néglige l'information \hat{S} il serait cohérent d'obtenir la procédure de comparaison présentée à la section précédente. Ainsi, l'équation (15) devient:

$$\Pr[\mathbf{X} \approx \mathbf{Y}^* | \hat{S} \in E_k] = \frac{\Pr[\hat{S} \in E_k | \mathbf{X} \approx \mathbf{Y}^*]}{\sum_{j=1}^J \Pr[\hat{S} \in E_k | \mathbf{X} \approx \mathbf{Y}_j]} \quad (16)$$

En plus d'estimer ces probabilités, le classement des échantillons simulés à l'aide de chaque échantillon de référence nous permet d'évaluer la perte moyenne engendrée par une action $[D/M]_i$ pour des séries simulées selon l'hypothèse $\mathbf{X} \approx \mathbf{Y}_j$, lorsque l'information \hat{S} pointe vers un sous-espace E_k , que nous noterons φ_{ijk} . Il est alors naturel d'estimer par l'équation 17 la perte moyenne engendrée par une action $[D/M]_i$ conditionnelle à l'information disponible ($\hat{S} \in E_k$), lorsque aucune hypothèse n'est faite sur \mathbf{Y}_j (excepté l'hypothèse implicite spécifiant que $\mathbf{Y}_j \in E_Y$).

$$\bar{\varphi}_{ik} = \sum_{j=1}^J \varphi_{ijk} \cdot \Pr[\mathbf{X} \approx \mathbf{Y}_j | \hat{\mathbf{S}} \in E_k] \quad (17)$$

En utilisant cette dernière équation, il est possible de déterminer la combinaison D/M minimisant la perte moyenne pour chaque sous-espace E_k . Nous avons donc développé une politique de sélection d'une combinaison D/M basée sur l'estimation $\hat{\mathbf{S}}$ d'un ensemble de caractéristiques et sur son classement dans un sous-espace E_k . L'utilisation comme critère de sélection d'un coût moyen se justifie aisément si cette politique de sélection est appliquée à un grand nombre de sites, car alors les pertes monétaires seront minimisées. Cependant, ce critère n'est pas adapté pour le cas où une politique de sélection doit être appliquée à un faible nombre de sites.

2.4.1 Hypothèses implicites liées à l'analyse Bayésienne

La construction de la politique de sélection présentée dans la section précédente repose sur quelques hypothèses implicites, inhérentes à l'analyse Bayésienne. Premièrement, tous les calculs de probabilité sont conditionnés par l'ensemble des états de la nature considérés. On suppose donc que l'ensemble des hypothèses de travail considérées (c'est à dire l'ensemble des échantillons de référence) permet de mettre en évidence tous les cas de figure qui peuvent se poser en pratique lors de l'application de la politique de sélection. Dans notre cas, cela revient à supposer que l'ensemble des échantillons de référence considéré est représentatif de l'ensemble des échantillons du territoire où pourrait être appliquée la politique de sélection.

Une seconde hypothèse est nécessaire pour construire la distribution *a priori*, ou distribution non informative. En effet, il n'est pas suffisant de savoir si tous les cas de figures sont présents dans l'ensemble des échantillons de référence, car certains cas peuvent se présenter plus fréquemment que d'autres dans l'ensemble des échantillons du territoire. La distribution non informative permet d'associer un poids différent à chaque échantillon de référence. Cependant, il est très difficile de déterminer ces poids, puisque l'on ne peut justement pas déterminer précisément à quel échantillon de référence ressemble un échantillon de plus faible taille.

Pour résoudre cette difficulté, nous avons convenu d'utiliser comme échantillons de référence tous les échantillons de taille suffisante (c'est à dire suffisamment supérieure à la période de retour d'intérêt pour qu'il soit possible d'estimer x_T de façon non paramétrique). Ainsi, l'ensemble des échantillons de référence forme un sous-ensemble dont la distribution est représentative de l'ensemble des échantillons du territoire, à condition que le nombre d'observations disponibles à un site soit indépendant de la forme de la distribution des débits à ce site, et que le nombre d'échantillons de référence obtenus soit suffisant. Sous cette hypothèse, une distribution *a priori* uniforme est justifiable.

2.4.2 Optimalité de la politique de sélection

Dans la mesure où l'on peut calculer (16), i.e. si $\sum_{j=1,J} \Pr[\hat{S} \in E_k | \mathbf{X} \approx \mathbf{Y}_j] \neq 0$, on peut rendre les sous-espaces E_1, E_2, \dots, E_w aussi petits que l'on veut. Cependant, plus ils seront petits, moins il y aura d'estimations de \mathbf{S} , obtenues à partir des séries simulées, classées dans chaque sous-espace. La précision de l'estimation de la fréquence avec laquelle une estimation \hat{S} appartient à un sous-espace particulier E_k étant donné que l'échantillon de référence est \mathbf{Y}_j , $\Pr[\hat{S} \in E_k | \mathbf{X} \approx \mathbf{Y}_j]$, diminuera donc avec la taille du sous-espace E_k . Notons qu'il est par contre aisé d'augmenter le nombre de séries simulées à l'aide de chaque échantillon de référence pour contourner cette difficulté. Ainsi, les sous-espaces E_k peuvent être rendus aussi petit que l'on veut tout en estimant $\Pr[\hat{S} \in E_k | \mathbf{X} \approx \mathbf{Y}_j]$ avec une précision suffisante.

Comme nous l'avons mentionné au début de cette section, cette politique de sélection est en quelque sorte optimale par rapport à la procédure de comparaison présentée au début de ce chapitre. En effet, supposons que l'on choisisse les sous-espaces E_k si petits que la distribution D/M minimisant la fonction de perte $g(x_T, \hat{x}_T)$ soit la même pour toutes les séries simulées dont l'estimation de \mathbf{S} fait partie d'un même sous-espace. Par exemple, on peut choisir de partitionner les sous-espaces de façon à ne retrouver dans chaque sous-espace au maximum une seule estimation de \mathbf{S} par échantillon de référence. Dans ce cas, la politique de sélection permettrait de choisir la combinaison D/M minimisant $g(x_T, \hat{x}_T)$ pour toutes les séries simulées.

Selon la procédure de comparaison développée précédemment, cette politique serait donc nécessairement la meilleure politique de sélection qu'il soit possible de construire.

Évidemment, cette politique de sélection ne sera optimale que pour les séries sur lesquelles elle aura été calibrée. Pour une série de données réelles dont on suppose seulement qu'elle ressemble statistiquement à l'un des échantillons de référence, les résultats peuvent être moins encourageants. Nous suggérons, pour vérifier la performance de cette politique de sélection, d'exclure une portion des échantillons de référence lors de la calibration de la politique (i.e. lors de l'estimation des $\Pr[\hat{S} \in E_k | \mathbf{X} \approx \mathbf{Y}_j]$ et φ_{ijk}). Les séries simulées à partir de ces échantillons de référence pourraient être alors utilisées pour valider cette politique.

Pour fixer les idées, le prochain chapitre sera consacré à un exemple d'application de cette politique de sélection, précédé d'un exemple d'application de la procédure de comparaison de politiques de sélection présentée au début de ce chapitre.

3 EXEMPLE D'APPLICATION

Nous avons développé au chapitre précédent une procédure de comparaison de politiques de sélection d'une combinaison D/M pour estimer x_T , sur la base de laquelle nous avons aussi élaboré une politique de sélection. Dans ce chapitre, nous donnerons un exemple d'application de la procédure de comparaison ainsi que de la politique de sélection.

3.1 Base de données

Nous appliquerons la procédure de comparaison et la politique de sélection développées au chapitre précédent aux provinces du Québec et de l'Ontario. [MATHIER *et al.* 1993] ont examiné l'ensemble des stations de mesure du débit de ces deux provinces jusqu'en 1990. Dans ces deux provinces, ils ont identifié 189 stations de mesure du débit quotidien moyen situées sur des rivières ayant un régime naturel pour lesquelles on dispose d'au moins 20 années complètes d'enregistrement du débit quotidien moyen. Le maximum de chaque année de mesure du débit quotidien moyen donnant en général une mesure acceptable du débit maximum annuel instantané, 189 séries d'au moins 20 mesures du débit maximum annuel ont donc été identifiées pour ces deux provinces.

De ce nombre, 18 ont été éliminées parce que le test d'homogénéité de Mann-Whitney [MANN et WHITNEY 1947] a conclu, au niveau de signification de 5%, que les crues de printemps (de janvier à juin) n'appartenaient pas à la même population que les crues d'automne (de juillet à décembre). Des 171 stations restantes, 9 ont été éliminées parce qu'elles démontraient une tendance ou parce qu'un lac situé en amont causait une autocorrélation des débits maximums annuels [MATHIER 1994]. Ainsi, 162 séries de mesure du débit maximum annuel jugées indépendantes et identiquement distribuées ont donc été identifiées. Ces stations sont énumérées à l'annexe C en ordre décroissant de nombre d'années de mesures. Les mesures de débit maximum annuel ne sont pas énumérées pour chaque station, mais elles sont archivées sur le disque compact HYDAT 3.0 [ENVIRONNEMENT CANADA 1990].

Les simulations par suréchantillonnage, nécessaires pour appliquer la procédure de comparaison et la politique de sélection développées précédemment, seront effectuées à partir

des 7 séries parmi les 162 comportant plus de 60 observations. Toutes les observations de ces 7 séries sont énumérées à l'annexe C. De ces 7 séries, 6 sont situées en Ontario, une seule est située au Québec. Nous utiliserons les 6 premières pour la calibration de la procédure de comparaison et de la politique de sélection, et la dernière pour valider les résultats obtenus. Pour donner une idée de la distribution empirique de ces échantillons de référence, nous avons tracé les observations sur papier de probabilité normal en utilisant la formule de probabilité empirique de Cunnane (cf. équation 8). On retrouve ces graphiques à l'annexe C. Rappelons que le papier de probabilité normal est un papier dont l'abscisse est transformée de façon à ce que les observations forment une droite si elles suivent une distribution normale. [LAW et KELTON 1982]. Remarquons que sur la base de ces tracés il semble difficile d'accepter la normalité des observations, puisque ceux-ci ne sont visiblement pas linéaires.

3.2 Tailles d'échantillon étudiées

Les séries de débits maximums annuels sélectionnées comportent en général (87% des cas) entre 20 et 40 observations, le mode étant de 24, la médiane 25 et la moyenne 30. Nous étudierons la performance des diverses politiques pour deux tailles aux extrêmes de cette distribution: 20 et 40 ans.

3.3 Périodes de retour considérées

Il ne sera possible de comparer la performance des politiques de sélection que pour des périodes de retour suffisamment inférieures à 62 années, puisqu'il s'agit de la taille de l'échantillon de référence utilisé comportant le moins d'observations. Nous considérerons donc des périodes de retour de 20 et 50 années. Ce type de période de retour peut par exemple servir à la conception d'ouvrages de dérivation temporaires [SÉBJ 1987].

3.4 Politiques de sélection considérées

Nous considérerons quelques politiques de sélection simples, consistant en l'application systématique d'une même combinaison D/M. Puisque toutes les combinaisons D/M ne

peuvent être considérées, nous nous limiterons à certaines distributions et méthodes d'ajustement incluses dans le logiciel Ajuste 1.1 [FORTIN *et al.* 1992], conçu pour répondre aux besoins de Hydro-Québec. Une seule distribution sera ajoutée: la distribution log-normale à deux paramètres (2LN) qui est un cas limite des distributions log-Pearson type III (LP3) et gamma généralisée (GG3), toutes deux incluses dans Ajuste 1.1.

Le tableau 1 donne l'abréviation, le domaine de définition de x_T , les paramètres, leur type et leur domaine de définition ainsi que les méthodes d'estimation de chaque distribution que nous considérerons. Trois méthodes d'estimation sont utilisées: la méthode des moments (MM), la méthode du maximum de vraisemblance (MV) et la méthode des moments pondérés (*Probability Weighted Moments*, ou PWM). Les différentes méthodes d'estimation sont expliquées à l'annexe A, alors que les divers types de paramètres mentionnés sont discutés à l'annexe D. On retrouve à l'annexe E une description plus détaillée de chaque distribution comprenant la f.d.p. cumulée inverse de chaque distribution, nécessaire pour pouvoir calculer un débit x_T de période de retour T (cf. équation 5). Nous n'ajouterons ici qu'une figure, adaptée de [BOBÉE et ASHKAR 1991] exprimant les relations qui existent entre les différentes distributions considérées.

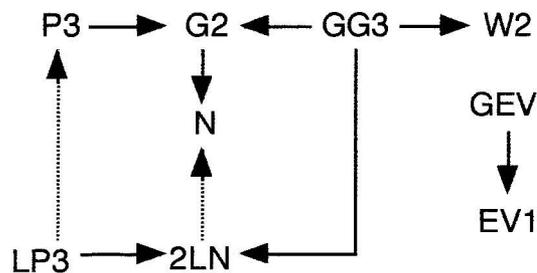


Figure 6: Relations entre les différentes distributions considérées

Les relations entre les distributions exprimées par la figure 6 doivent s'interpréter ainsi: une flèche continue dirigée d'une distribution A vers une distribution B indique que la distribution B est un cas particulier ou un cas limite de la distribution A, alors qu'une flèche pointillée dirigée d'une distribution A vers une distribution B indique que la distribution B est la distribution du logarithme des observations de la distribution A. Les distributions statistiques présentées à la figure 6 peuvent se diviser en quatre familles, regroupant chacune

les cas particuliers et les cas limites d'une distribution plus générale. On distingue les familles Pearson type III (P3,G2,N), log-Pearson type III (LP3,2LN), Gamma généralisée (GG3,W2,G2,N), et la famille des distributions des valeurs extrêmes (GEV,EV1). Il existe d'autres types de liens entre les distributions de la figure 6, comme il existe des liens entre celles-ci et d'autres distributions que nous ne présentons pas ici. [BOBÉE et ASHKAR 1991] donnent une vision plus complète de ces relations.

Tableau 1: Résumé des distributions considérées pour estimer les débits x_T

Distribution	Domaine de définition	Paramètres	Méthodes d'estimation
normale (N)	$x_T \in \mathbb{R}$	μ (position.), σ (échelle)	MV
log-normale (2LN)	$x_T > 0$	μ (échelle), σ (forme)	MV
gamma (G2)	$x_T > 0$	$\alpha \neq 0$ (échelle), $\lambda > 0$ (forme)	MM, MV
Pearson type III (P3)	Si $\alpha > 0$: $x_T > m$ Si $\alpha < 0$: $x_T < m$	m (position), $\alpha \neq 0$ (échelle), $\lambda > 0$ (forme)	MM, MV
log-Pearson type III (LP3)	Si $\alpha > 0$: $x_T > \exp\{m\}$ Si $\alpha < 0$: $0 < x_T < \exp\{m\}$	m (échelle), $\alpha \neq 0$ (forme) $\lambda > 0$ (forme)	BOB*, MV, WRC**
gamma généralisée (GG3)	$x_T > 0$	$\alpha \neq 0$ (échelle), $\lambda > 0$ (forme), $s \neq 0$ (forme)	MM, MV
Weibull (W2)	$x_T > 0$	$\alpha > 0$ (échelle), $c > 0$ (forme)	MM, MV
Gumbel (EV1)	$x_T \in \mathbb{R}$	u (position), $\alpha > 0$ (échelle)	MM, PWM, MV***
Dist. généralisée des valeurs extrêmes (GEV)	Si $k < 0$: $x_T > u + \alpha/k$ Si $k = 0$: $x_T \in \mathbb{R}$ Si $k > 0$: $x_T < u + \alpha/k$	u (position), $\alpha > 0$ (échelle), k (forme)	MM, PWM, MV

* Méthode des moments non centrés (voir détails en annexe E)

** Méthode des moments sur le logarithme des observations (voir détails en annexe E)

*** Avec correction pour le biais (voir détails en annexe E)

L'application d'une combinaison D/M nécessite la résolution d'un système comportant autant d'équations que la distribution D possède de paramètres. De plus, dans la majorité des cas ce

système ne possède pas de solution analytique, et l'on doit employer une méthode numérique pour le résoudre. Pour certaines combinaisons D/M du tableau 1, il arrive que ce système ne possède pas de solution, ou soit tellement mal conditionné que la solution numérique soit impossible à obtenir en pratique. Ces combinaisons sont G2/MV, P3/MV, LP3/BOB, LP3/MV, GG3/MM, GG3/MV, GEV/MM, GEV/MV, GEV/PWM. Les causes précises qui empêchent parfois d'obtenir une solution (donc d'estimer x_T) lorsque l'on utilise ces combinaisons D/M sont discutées séparément pour chacune à l'annexe E.

L'implication pratique de l'impossibilité d'estimer x_T dans certains cas pour certaines combinaisons D/M est que l'on ne peut pas construire de politique statique de sélection utilisant uniquement l'une de ces combinaisons D/M, car une politique de sélection appliquée à une série quelconque de débits maximums annuels doit permettre, en toute logique, de sélectionner une combinaison D/M permettant d'estimer x_T pour cet échantillon.

Si l'on veut bâtir une politique statique de sélection utilisant ces combinaisons D/M, il faut prévoir une combinaison D/M de rechange, qui sera utilisée lorsque la première ne permet pas d'estimer x_T . S'il est en plus parfois impossible d'estimer x_T en utilisant cette seconde combinaison, il faudra en prévoir une troisième, et ainsi de suite.

Nous proposons d'utiliser, lorsqu'une méthode d'estimation d'une distribution ne fonctionne pas, une autre méthode d'estimation de la même distribution. Quand aucune des méthodes d'estimation d'une distribution ne permettra d'estimer x_T , nous considérerons l'utilisation d'une distribution de la même famille. Nous pourrions ainsi définir une politique statique de sélection consistant en l'application successive d'une série de combinaisons D/M, toujours dans le même ordre, jusqu'à ce que l'une d'elles permette d'estimer x_T .

Dans un cadre d'une comparaison de politiques de sélection de combinaisons D/M, il est important de gérer les cas où les combinaisons D/M ne permettent pas d'estimer x_T (surtout lorsque la fréquence de ces cas problèmes est élevée) puisque l'on peut facilement imaginer des scénarios où le résultat de la comparaison en est influencé. Il se peut par exemple que les cas où une combinaison D/M ne permet pas d'estimer x_T correspondent à des échantillons ayant des propriétés particulières telles qu'il est plus difficile *pour toute combinaison D/M* d'estimer x_T avec précision (par exemple, ces échantillons comportent une valeur singulière,

qui semble provenir d'une autre population). Si la performance de chaque combinaison D/M est évaluée, en utilisant l'équation 13, mais en considérant seulement l'ensemble des échantillons pour lesquels la combinaison permet d'estimer x_T , alors les combinaisons D/M dont l'algorithme d'estimation, plus robuste, permet d'estimer x_T pour tout échantillon seront nettement désavantagées. C'est pourquoi nous suggérons de ne considérer que des politiques de sélection permettant d'estimer x_T pour tout échantillon.

Cette approche nous conduit à définir 34 politiques statiques de sélection à partir des 19 combinaisons D/M. Ces politiques, que nous nommerons P_1 à P_{34} , sont résumées par le tableau 2. Le tableau 2 indique, pour chaque politique (P), quelle combinaison D/M seront utilisées et dans quel ordre elles le seront. L'application d'une politique de sélection peut nécessiter l'emploi d'une à quatre combinaisons D/M différentes, le dernier choix étant une combinaison D/M pour laquelle il est toujours possible d'estimer x_T , exception faite des politiques statiques faisant intervenir la distribution GEV. En effet, il arrive que chacune des 3 méthodes d'ajustement de cette distribution ne permette pas d'estimer x_T mais en pratique, pour un échantillon donné, nous n'avons jamais observé de cas où le problème se posait pour les 3 méthodes en même temps.

3.5 Application de la procédure de comparaison de politiques de sélection

Après avoir fixé les tailles d'échantillon, les périodes de retour et les politiques de sélection à étudier, nous pouvons utiliser la procédure de comparaison de politiques présentée au chapitre précédent pour déterminer quelle politique de sélection il est préférable d'utiliser pour estimer x_T au Québec et en Ontario. Il est cependant prudent de vérifier auparavant si les échantillons de référence sont représentatifs de l'ensemble des séries de ces deux provinces, c'est à dire si les séries simulées ressemblent statistiquement aux séries réelles. Après avoir simulé les séries de débits maximums annuels à partir des échantillons de référence et après avoir comparé leur distribution avec celle des données réelles disponibles, nous déterminerons les valeurs correctes de x_T pour les périodes de retour sélectionnées, pour enfin pouvoir estimer, à l'aide des équations 12, 13 et 14, la performance de chaque politique de sélection identifiée précédemment.

Tableau 2: Politiques de sélection d'une combinaison D/M envisagées

P	Choix 1	Choix 2	Choix 3	Choix 4	P	Choix 1	Choix 2	Choix 3	Choix 4
1	N/MV				18	GG3/MV	GG3/MM	G2/MM	
2	EV1/MM				19	GG3/MV	GG3/MM	G2/MV	G2/MM
3	EV1/MV				20	GG3/MV	GG3/MM	2LN/MV	
4	EV1/PWM				21	P3/MM			
5	W2/MM				22	P3/MV	P3/MM		
6	W2/MV				23	GEV/MM	GEV/MV	GEV/PWM	
7	G2/MM				24	GEV/MM	GEV/PWM	GEV/MV	
8	G2/MV	G2/MM			25	GEV/MV	GEV/MM	GEV/PWM	
9	GG3/MM	GG3/MV	N/MV		26	GEV/MV	GEV/PWM	GEV/MM	
10	GG3/MM	GG3/MV	W2/MM		27	GEV/PWM	GEV/MM	GEV/MV	
11	GG3/MM	GG3/MV	W2/MV		28	GEV/PWM	GEV/MV	GEV/MM	
12	GG3/MM	GG3/MV	G2/MM		29	LP3/WRC			
13	GG3/MM	GG3/MV	G2/MV	G2/MM	30	LP3/BOB	LP3/WRC		
14	GG3/MM	GG3/MV	2LN/MV		31	LP3/BOB	LP3/MV	LP3/WRC	
15	GG3/MV	GG3/MM	N/MV		32	LP3/MV	LP3/WRC		
16	GG3/MV	GG3/MM	W2/MM		33	LP3/MV	LP3/BOB	LP3/WRC	
17	GG3/MV	GG3/MM	W2/MV		34	2LN/MV			

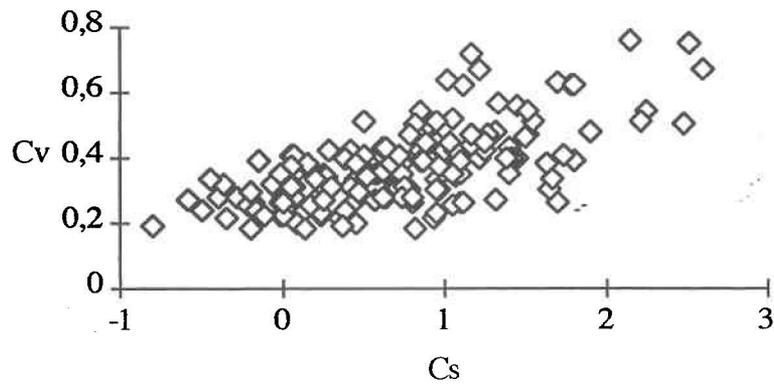


Figure 7: Distribution des couples (\hat{C}_s, \hat{C}_v) des séries réelles de taille $N=20$

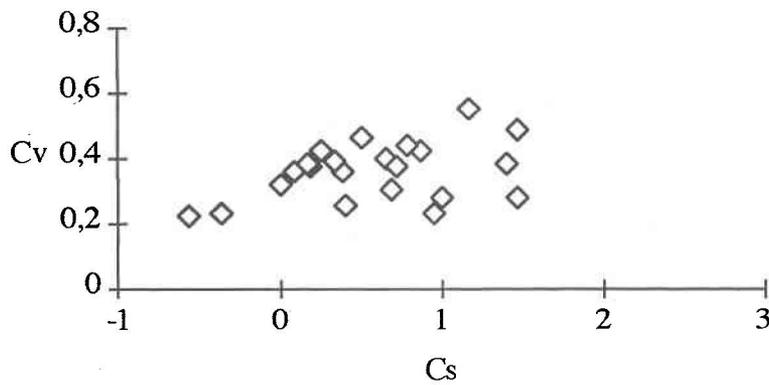


Figure 8: Distribution des couples (\hat{C}_s, \hat{C}_v) des séries réelles de taille $N=40$

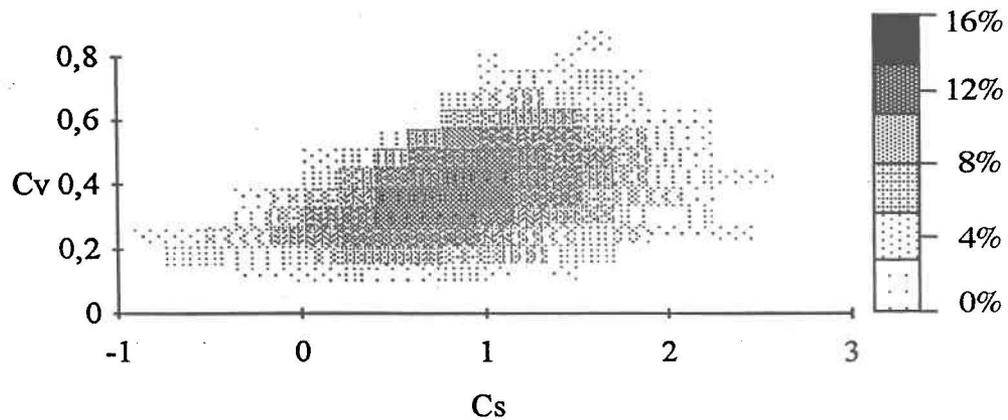


Figure 9: Distribution des couples (\hat{C}_s, \hat{C}_v) des séries simulées de taille $N=20$

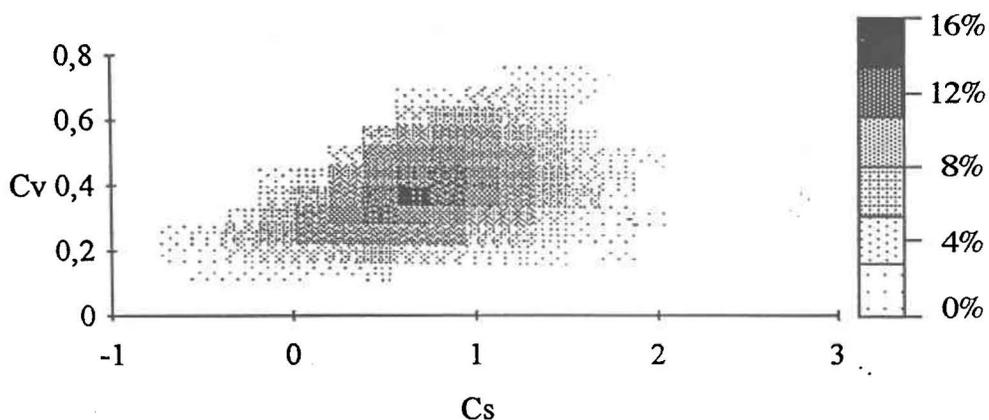


Figure 10: Distribution des couples (\hat{C}_s, \hat{C}_v) des séries simulées de taille $N=40$

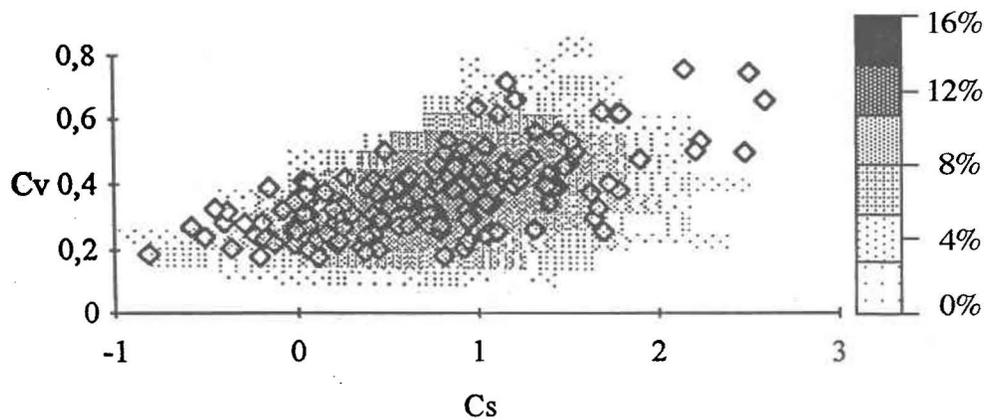


Figure 11: Distribution des couples (\hat{C}_s, \hat{C}_v) des séries réelles et simulées de taille $N=20$

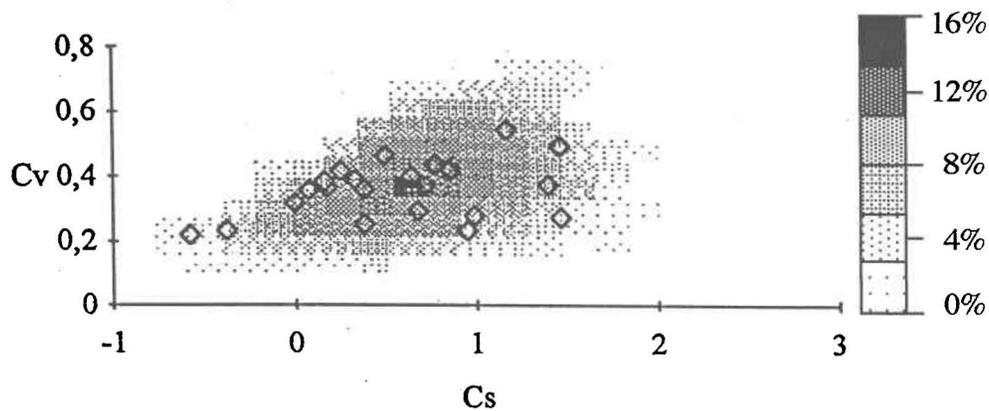


Figure 12: Distribution des couples (\hat{C}_s, \hat{C}_v) des séries réelles et simulées de taille $N=40$

3.5.1 Simulation de débits à partir des échantillons de référence

L'algorithme de simulation par tirage de Pólya, détaillé à la figure 5, est passablement simple. La seule difficulté rencontrée lors de sa programmation est l'automatisation du tirage aléatoire d'un nombre aléatoire entre 1 et $N+i-1$. Nous avons utilisé le générateur pseudo-aléatoire à congruence linéaire (GCM) DRAND, proposé par [LAW et KELTON 1982]. Il faut aussi fixer, au moment de la simulation, le nombre de séries simulées par échantillon de référence. Pour obtenir une précision suffisante, nous avons simulé $R=500$ séries par échantillon de référence.

3.5.2 Comparaison de la distribution des séries simulées avec celle des données réelles

Une façon de vérifier si les données simulées ont une distribution similaire à celle des données réelles disponibles est de comparer la distribution de quelques caractéristiques statistiques des séries simulées et des données réelles. Nous avons choisi de comparer la distribution de l'estimation du coefficient de variation (C_v) et celle du coefficient d'asymétrie (C_s). Comme la distribution de ces coefficients est fonction de la taille d'échantillon, nous avons répété l'expérience suivante pour les tailles $N=20$ et $N=40$:

- 1) C_v et C_s ont été estimés pour les $R=500$ échantillons simulés par suréchantillonnage à partir de chacun des 6 échantillons de référence utilisés pour construire la procédure de comparaison, donnant 3000 couples (\hat{C}_s, \hat{C}_v) simulés.
- 2) Parmi les 162 séries réelles, toutes celles comportant N observations ou plus ont été sélectionnées, puis tronquées pour ne garder que N observations (les observations les plus récentes ont été conservées). C_v et C_s ont été estimés pour ces séries tronquées.

Les figures 7 et 8 montrent la distribution des couples (\hat{C}_s, \hat{C}_v) mesurés à partir des séries réelles, respectivement pour les tailles d'échantillon $N=20$ et $N=40$. On peut comparer ces distributions avec celles obtenues à partir des séries simulées, présentées aux figures 9 et 10. Pour permettre l'interprétation de ces deux figures, nous avons groupé les couples (\hat{C}_s, \hat{C}_v) obtenus pour les 3000 échantillons simulés en classes. La proportion relative p des couples

(\hat{C}_s, \hat{C}_v) retrouvée dans chaque classe a été codée avec un ton de gris variant du blanc ($p=0$) au noir ($p=0,16$). La plage du diagramme (C_s, C_v) couverte par les couples (\hat{C}_s, \hat{C}_v) simulés est à peu près équivalente à celle couverte par les couples (\hat{C}_s, \hat{C}_v) obtenus à partir des données réelles, sauf pour $N=20$ et $\hat{C}_s > 1,5$. Pour mieux visualiser ces résultats, les figures 11 et 12 montrent l'effet de la superposition des données des figures 7 et 9 ainsi que 8 et 10.

Dans l'ensemble, les échantillons de référence semblent assez représentatifs des séries de données réelles, sauf lorsque le coefficient d'asymétrie estimé (\hat{C}_s) est élevé ($>1,5$). En effet, aucun échantillon de référence ne permet de simuler suffisamment de séries ayant un tel coefficient d'asymétrie, alors que le \hat{C}_s des séries réelles dépasse souvent cette marque. Ce manque de représentativité pourrait s'expliquer par le fait que les stations comportant le plus de données au Québec et en Ontario sont situées principalement au sud-ouest de cette région, et que la distribution des débits pour des stations plus au nord diffère suffisamment de celle des échantillons de référence pour produire cet effet. Il est aussi raisonnable de supposer que le nombre limité d'échantillons de référence disponibles (seulement 7) cause en grande partie ce phénomène, indépendamment du fait qu'ils ne sont pas répartis également sur le territoire.

3.5.3 Détermination des valeurs correctes de x_T

Comme expliqué au chapitre précédent, le modèle de simulation doit permettre de déterminer la valeur exacte de x_T des séries simulées. Si les séries étaient simulées à partir d'un modèle paramétrique complètement spécifié, cela serait aisé à réaliser en utilisant l'équation 5. Cependant, nous avons voulu utiliser un modèle de simulation non paramétrique ne nécessitant pas de spécifier complètement la forme de la distribution des débits; il est donc impossible de déterminer la valeur exacte de x_T . Il est par contre possible d'approximer cette valeur en utilisant une formule de probabilité empirique (FPE), à condition que T soit moins élevé que la taille de l'échantillon de référence. Nous nommerons cette approximation *valeur correcte* de x_T . Cependant, chaque FPE conduit à une valeur de x_T différente. Parce que le choix de la FPE peut influencer les résultats, nous considérerons 3 FPE différentes, provenant cependant toutes de l'équation générale 6: les FPE de Weibull ($a=0$), Cunnane ($a=0,4$) et Hazen ($a=0,5$). Puisque en général $0 \leq a \leq 0,5$, les FPE de Weibull et Hazen se situent aux deux extrémités du spectre, alors que la FPE de Cunnane est un choix intermédiaire, peu influencé par la distribution réelle des débits [CUNNANE 1978].

Les valeurs correctes de x_T sont donc calculées de la façon suivante pour chaque FPE, chaque échantillon de référence et chaque période de retour T d'intérêt. Premièrement, la fonction de densité de probabilité (f.d.p.) cumulée $F(x_{(i)})$ est estimée pour chaque observation $x_{(i)}$ à l'aide de l'équation 6. La période de retour correspondant à chaque observation est ensuite estimée par $T_{(i)}=1/[1-F(x_{(i)})]$. Si $T < N$, on peut alors trouver $j < N$ tel que $T_{(j)} \leq T \leq T_{(j+1)}$. Ainsi, l'on sait que la valeur correcte x_T se situe entre $x_{(j)}$ et $x_{(j+1)}$. Une interpolation linéaire peut enfin être utilisée pour déterminer x_T . Les valeurs de x_T obtenues pour chaque échantillon de référence et pour chaque période de retour T considérée sont détaillées au tableau 3 pour chaque FPE. Notons que x_T est une fonction décroissante de α .

Tableau 3: Valeur correcte de x_T selon l'échantillon de référence et la FPE considérés

Station	T=20			T=50		
	Hazen	Cunnane	Weibull	Hazen	Cunnane	Weibull
02FC001	818	818	822	894	905	957
02FC002	531	531	539	628	630	634
02EA005	69,1	69,5	71,9	82,0	83,1	87,5
02EC002	181	182	185	199	202	213
05PB014	233	233	237	258	262	280
04LJ001	1320	1320	1350	1500	1530	1630
02PJ007*	284	286	292	307	309	314

* Station utilisée pour la validation uniquement

3.5.4 Résultats de la comparaison pour $\alpha=1$ et $r=2$ (EQM)

En utilisant l'écart quadratique moyen (EQM) comme fonction de coût (obtenu en fixant $\alpha=1$ et $r=2$ dans l'équation 12), on peut déterminer à l'aide des séries simulées quelle politique, de P_1 à P_{34} , minimise l'équation 14, c'est à dire la politique qui minimise la moyenne des EQM pour toutes les séries simulées à partir de tous les échantillons de référence (sauf celui réservé pour la validation). Le tableau 4 donne pour chaque période de retour, taille d'échantillon et FPE considérées les cinq politiques pour lesquelles la moyenne des EQM est la plus faible. Les choix entre crochets sont considérés *ex æquo*, leur mesure de performance

étant égales à 2 chiffres significatifs. Rappelons que les combinaisons D/M correspondant à chaque politique sont énumérées au tableau 2.

Il faut bien comprendre ce que nous entendons par politiques *ex æquo*. Aucun test statistique n'a été employé pour déterminer si les mesures de performance sont significativement différentes les unes des autres. En effet, comme ces mesures sont déterminées par simulation, une différence non significative entre les résultats impliquerait seulement que le nombre d'échantillons simulés est insuffisant. La seule indication qu'il est possible d'utiliser pour déterminer si les différences sont significative est la précision des calculs. Comme les mesures du débit maximum annuel obtenues n'ont que 3 chiffres significatifs (voir Annexe C), il n'est possible, au maximum, que d'obtenir ce degré de précision. Il est cependant raisonnable, étant donné la propagation d'erreurs d'arrondi, de ne considérer significatifs que les deux premiers chiffres.

En observant les combinaisons D/M utilisées dans les politiques sélectionnées comme premier choix (ou premier choix *ex æquo*), soit N/MV, EV1/MM, EV1/PWM, G2/MM et G2/MV, on remarque d'abord qu'aucune distribution à trois paramètres n'a été retenue. La distribution EV1 est toujours sélectionnée en premier choix, sauf pour les FPE Hazen et Cunnane lorsque $T < N$; dans ce cas, la distribution Gamma est sélectionnée. On remarque aussi que les résultats produits par les FPE de Hazen et de Cunnane sont toujours les mêmes pour les trois premiers choix. La FPE de Weibull produit des résultats quelque peu différents. Cependant, si l'on considère le premier choix lorsque $T \geq N$, les variations se font au niveau de la méthode d'estimation: l'utilisation de la FPE de Weibull amène à sélectionner la combinaison EV1/PWM au lieu de EV1/MM. Le deuxième choix (sur les 34 possibles) lorsque l'on considère la FPE de Weibull est d'ailleurs EV1/MM. Lorsque $T \geq N$, toutes les FPE sont donc à peu près équivalentes. Même dans le seul cas considéré pour lequel $T < N$ ($T=20$ et $N=40$), les différences entre les FPE sont tout de même assez faibles: l'utilisation des FPE de Hazen et Cunnane amènent à sélectionner $P_7=G2/MM$, alors que l'utilisation de la FPE de Weibull amène à sélectionner $P_2=EV1/MM$, mais le second choix des FPE de Hazen et Cunnane, après le choix *ex æquo* P_8 , est aussi $P_2=EV1/MM$, et le troisième choix de la FPE de Weibull est $P_7=G2/MM$.

Tableau 4: Politiques minimisant l'écart quadratique moyen

T	N	Choix	Hazen		Cunnane		Weibull	
			Politique	EQM	Politique	EQM	Politique	EQM
20	20	1	[2]	1,2E-2	[2]	1,2E-2	[2]	1,2E-2
		2	[7]	1,2E-2	[7]	1,2E-2	[7]	1,2E-2
		3	[8]	1,2E-2	[8]	1,2E-2	[4]	1,2E-2
		4	[1]	1,2E-2	[1]	1,2E-2	[8]	1,2E-2
		5	4	1,3E-2	4	1,3E-2	[24]	1,3E-2
20	40	1	[7]	7,9E-3	[7]	7,8E-3	2	7,6E-3
		2	[8]	7,9E-3	[8]	7,8E-3	4	7,8E-3
		3	2	8,0E-3	2	7,9E-3	7	7,9E-3
		4	[24]	8,4E-3	24	8,3E-3	8	8,2E-3
		5	[6]	8,4E-3	6	8,4E-3	24	8,5E-3
50	20	1	[2]	1,6E-2	2	1,5E-2	4	1,4E-2
		2	[7]	1,6E-2	[7]	1,6E-2	2	1,5E-2
		3	[8]	1,6E-2	[8]	1,6E-2	3	1,6E-2
		4	[4]	1,8E-2	4	1,7E-2	7	1,8E-2
		5	[1]	1,8E-2	1	1,9E-2	8	2,0E-2
50	40	1	[2]	1,0E-2	2	9,9E-3	4	9,3E-3
		2	[7]	1,0E-2	[7]	1,6E-2	2	9,7E-3
		3	[8]	1,0E-2	[8]	1,6E-2	3	1,1E-2
		4	21	1,1E-2	4	1,8E-2	7	1,3E-2
		5	4	1,2E-2	21	1,8E-2	8	1,4E-2

Tableau 5: Politique minimisant l'écart quadratique moyen (EQM) pour la FPE de Cunnane

N	T	Politique	Combinaison D/M
20	50	2	EV1/MM
40	50	2	EV1/MM
20	20	2	EV1/MM
40	20	7	G2/MM

Considérant que les différences entre les FPE sont faibles, surtout lorsque $T \geq N$, et considérant que l'utilisation d'un modèle paramétrique pour estimer x_T est surtout utile dans ce cas (on peut utiliser une approche non paramétrique lorsque $T < N$, voir section 1.3.1), nous croyons qu'il est raisonnable ne considérer qu'une seule FPE pour comparer des politiques de sélection. Nous utiliserons la FPE de Cunnane, qui est en général un compromis acceptable [CUNNANE 1978]. Selon cette hypothèse, les politiques recommandées pour minimiser la moyenne des EQM sont résumées par le tableau 5. La combinaison EV1/MM est recommandée pour l'extrapolation ($T \geq N$), et la combinaison G2/MM est recommandée pour l'interpolation ($T < N$).

3.5.5 Étude de sensibilité aux paramètres α et r

Les recommandations effectuées à la section précédente ne sont valables que si la mesure de performance est l'écart quadratique moyen au carré (EQM). Or, nous avons vu à la section 2.2 que rien ne supporte l'utilisation de cette mesure de performance. Nous l'avons donc généralisée par l'ajout de deux paramètres, α (>0) et r (>0), permettant de modifier la forme de cette mesure de performance (voir équation 12). Comme expliqué à la section 2.2, le paramètre α permet de faire varier le poids accordé aux erreurs de sous-estimation ($\hat{x}_T < x_T$) par rapport aux erreurs de surestimation ($\hat{x}_T > x_T$). Si $\alpha < 1$, plus de poids est accordé aux erreurs de surestimation, de sorte que les politiques générant plus d'erreurs de sous-estimation que de surestimation sont favorisées. Au contraire, si $\alpha > 1$, plus de poids est accordé aux erreurs de sous-estimation, et les politiques générant plus d'erreurs de surestimation sont avantagées. Le paramètre r permet plutôt de déterminer le poids des erreurs de grande amplitude, que ce soit des erreurs de sous-estimation ou des erreurs de surestimation: plus r est élevé, plus un poids important est accordé aux erreurs de grande amplitude, défavorisant ainsi les politiques générant parfois de grandes erreurs.

Pour observer l'effet de ces deux paramètres sur le résultat de la procédure de comparaison, nous avons d'abord fait varier le paramètre r autour de 2, en fixant α à 1 (lorsque $r=1$ et $\alpha=1$, la mesure de performance est l'écart moyen au carré). Les mesures de performance, $Z^{\alpha}g(x_T, \hat{x}_T)$, correspondant à chaque valeur de r sont schématisées par la figure 13. Rappelons que $\Delta x_T = \hat{x}_T - x_T$. Les politiques de sélection minimisant $\bar{\varphi}$, pour $r=1,2$ et $r=3$ avec $\alpha=1$ et en utilisant la FPE de Cunnane sont énumérées au tableau 6. Les numéros des 5

meilleures politiques sont donnés dans chaque cas avec la valeur numérique du critère de performance obtenu. Les politiques entre crochets sont considérées *ex aequo*, leur mesure de performance étant égales à 2 chiffres significatifs.

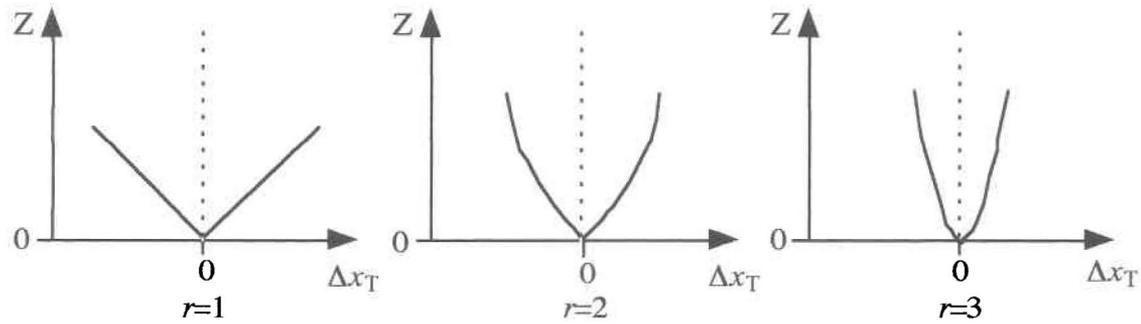


Figure 13: Mesures de performance pour à $\alpha=1$ et différentes valeurs de r

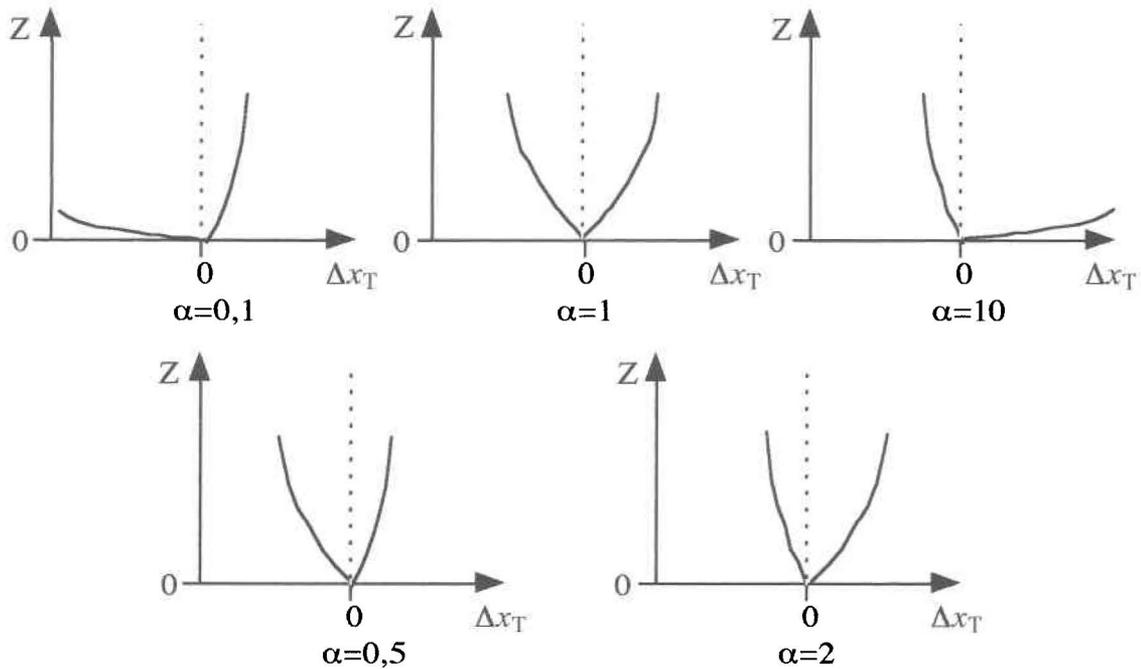


Figure 14: Mesures de performance pour à $r=2$ et différentes valeurs de α

Tableau 6: Politiques sélectionnées pour $\alpha=1$ et différentes valeurs de r

T	N	Choix	$r=1$		$r=2$		$r=3$	
			Politique	$\bar{\varphi}$	Politique	$\bar{\varphi}$	Politique	$\bar{\varphi}$
20	20	1	2	8,7E-2	[2]	1,2E-2	[2]	2,3E-3
		2	[7]	8,8E-2	7	1,2E-2	7	2,3E-3
		3	[8]	8,8E-2	8	1,2E-2	8	2,3E-3
		4	[4]	9,0E-2	[1]	1,2E-2	[1]	2,3E-3
		5	[1]	9,0E-2	4	1,3E-2	5	2,4E-3
20	40	1	[7]	6,9E-2	[7]	7,8E-3	[7]	1,1E-3
		2	[8]	6,9E-2	[8]	7,8E-3	[8]	1,1E-3
		3	[2]	6,9E-2	2	7,9E-3	[2]	1,1E-3
		4	[24]	7,2E-2	24	8,3E-3	[24]	1,2E-3
		5	[6]	7,2E-2	6	8,4E-3	[23]	1,2E-3
50	20	1	2	9,8E-2	2	1,5E-2	2	3,1E-3
		2	[7]	1,0E-1	[7]	1,6E-2	7	3,2E-3
		3	[4]	1,0E-1	[8]	1,6E-2	8	3,4E-3
		4	[8]	1,0E-1	4	1,7E-2	4	3,7E-3
		5	12	1,1E-1	1	1,9E-2	21	4,0E-3
50	40	1	2	7,8E-2	2	9,9E-3	[2]	1,6E-3
		2	7	8,0E-2	[7]	1,6E-2	[7]	1,6E-3
		3	4	8,2E-2	[8]	1,6E-2	8	1,7E-3
		4	8	8,3E-2	4	1,8E-2	4	1,9E-3
		5	21	8,8E-2	21	1,8E-2	21	2,0E-3

Il est clair qu'une variation de $\pm 100\%$ du paramètre r n'a qu'une influence mineure sur le résultat de la comparaison des politiques de sélection, car les deux premiers choix sont toujours les mêmes pour $r=1, 2$ et 3 . Il est donc raisonnable de fixer $r=2$. L'effet du paramètre α peut être étudié de la même façon, en fixant $r=2$, et en faisant varier α . Le tableau 7 présente les 5 meilleures politiques sélectionnées en utilisant la FPE de Cunnane, en fixant $r=2$, et pour $\alpha=1/10, 1/2, 1, 2$ et 10 . Les mesures de performance correspondant à ces valeurs de α sont schématisées par la figure 14.

Tableau 7: Politiques sélectionnées pour $r=2$ et différentes valeurs de α

T	N	Choix	$\alpha=0,1$		$\alpha=0,5$		$\alpha=1$		$\alpha=2$		$\alpha=10$	
			Pol.	$\bar{\varphi}$	Pol.	$\bar{\varphi}$	Pol.	$\bar{\varphi}$	Pol.	$\bar{\varphi}$	Pol.	$\bar{\varphi}$
20	20	1	1	4,0E-3	1	7,9E-3	[2]	1,2E-2	[2]	1,8E-2	[3]	5,1E-2
		2	6	4,6E-3	6	8,6E-3	[7]	1,2E-2	[4]	1,8E-2	[34]	5,4E-2
		3	5	5,1E-3	5	8,8E-3	[8]	1,2E-2	[7]	1,9E-2	4	5,5E-2
		4	[30]	5,6E-3	8	9,0E-3	[1]	1,2E-2	[8]	1,9E-2	2	6,2E-2
		5	[31]	5,6E-3	30	9,4E-3	4	1,3E-2	24	2,1E-2	7	7,0E-2
20	40	1	1	2,2E-3	1	5,0E-3	[7]	7,8E-3	[2]	1,1E-2	[3]	3,1E-2
		2	[6]	2,8E-3	[6]	5,3E-3	[8]	7,8E-3	[4]	1,1E-2	[34]	3,1E-2
		3	[5]	2,8E-3	[5]	5,3E-3	2	7,9E-3	[7]	1,1E-2	4	3,4E-2
		4	[30]	3,6E-3	8	5,7E-3	24	8,3E-3	8	1,2E-2	2	3,8E-2
		5	[31]	3,6E-3	7	5,8E-3	6	8,4E-3	24	1,3E-2	7	4,4E-2
50	20	1	1	3,2E-3	1	1,0E-2	2	1,5E-2	[2]	2,2E-2	3	6,0E-2
		2	6	3,9E-3	[6]	1,1E-2	[7]	1,6E-2	[4]	2,2E-2	4	6,5E-2
		3	5	4,4E-3	[5]	1,1E-2	[8]	1,6E-2	7	2,6E-2	2	7,4E-2
		4	[30]	5,4E-3	[7]	1,1E-2	4	1,7E-2	3	2,7E-2	34	7,9E-2
		5	[31]	5,4E-3	[8]	1,1E-2	1	1,9E-2	8	2,8E-2	7	1,0E-1
50	40	1	1	2,0E-3	7	6,9E-3	2	9,9E-3	2	1,3E-2	3	3,8E-2
		2	6	2,3E-3	8	7,1E-3	[7]	1,6E-2	4	1,4E-2	4	4,1E-2
		3	5	2,5E-3	21	7,4E-3	[8]	1,6E-2	7	1,7E-2	2	4,6E-2
		4	[30]	3,4E-3	[12]	7,7E-3	4	1,8E-2	8	1,8E-2	34	4,8E-2
		5	[31]	3,4E-3	[13]	7,7E-3	21	1,8E-2	3	1,9E-2	7	7,0E-2

Contrairement au paramètre r , le paramètre α a une influence marquée. On remarque que, en général, la politique $P_1=N/MV$ est sélectionnée comme premier choix lorsque $\alpha < 1$, alors que les politiques $P_2=EV1/MM$ et $P_3=EV1/MV$ sont sélectionnées lorsque $\alpha > 1$. L'influence de ce paramètre implique que l'utilisateur de la procédure de comparaison que nous avons développée doit nécessairement décider de l'importance relative qu'il accorde à la sous-estimation et à la surestimation en fixant le paramètre α .

S'il peut être difficile pour un utilisateur de fixer la valeur du paramètre α , nous croyons qu'il lui serait tout de même possible de déterminer si α est supérieur, inférieur ou égal à 1,

en déterminant s'il préfère risquer plus de surestimations ($\alpha > 1$), de sous-estimations ($\alpha < 1$), ou s'il préfère accorder un poids égal à ces deux types d'événements ($\alpha = 1$). Comme le montre le tableau 7, la connaissance de cette information (en plus de T et N) permet dans la majorité des cas de déterminer au moins la distribution statistique à utiliser (N, G2 ou EV1). Pour spécifier complètement la politique à utiliser (c'est à dire la combinaison D/M à utiliser, puisque toutes les politiques sélectionnées comme premier choix ne sont composées que d'une seule combinaison D/M), il faut que l'utilisateur détermine la valeur de α . En fonction de son attitude face au risque d'une sous-estimation, un utilisateur peut fixer α en choisissant, parmi les courbes de la figure 14, la forme de mesure de performance qu'il trouve la plus adéquate.

3.6 Application de la politique dynamique de sélection

Nous avons développé au chapitre 2 une politique de sélection permettant d'employer les caractéristiques statistiques de l'échantillon pour déterminer la combinaison D/M à utiliser pour estimer x_T . Maintenant que nous avons identifié à partir de données réelles les meilleures politiques statiques (i.e. consistant en l'application systématique d'une ou d'un ensemble de combinaisons D/M), nous pouvons mesurer la performance de cette politique dynamique.

Les caractéristiques statistiques que nous avons choisi d'utiliser pour effectuer la sélection de la combinaison D/M à utiliser sont les coefficients de variation (Cv) et d'asymétrie (Cs). Pour des séries de 20 ou 40 observations, nous avons en effet supposé qu'il ne serait que peu utile d'utiliser des coefficients faisant intervenir des moments d'ordre supérieur à 3, vu leur grande variabilité pour de petites tailles d'échantillon.

Comme nous l'avons montré à la section 2.4, il est nécessaire de partitionner l'espace $C_s \times C_v$ pour appliquer cette politique dynamique de sélection, et plus la partition sera fine, meilleure sera la performance de la politique *pour les séries sur lesquelles elle sera calibrée*. Par contre, pour que la politique dynamique donne de bons résultats lors de sa validation, il est important de ne pas dépasser la précision des estimations de Cv et de Cs lors de la partition de l'espace $C_s \times C_v$. On peut avoir une idée de cette précision en estimant l'écart-

type de ces coefficients pour des échantillons d'observations normalement distribuées. En effet, dans ce cas, l'écart-type de C_v est approximativement $C_v\{(1+2C_v^2)/2N\}^{0,5}$, et l'écart-type de C_s est approximativement $\{6N(N-1)/[(N-2)(N+1)(N+3)]\}^{0,5}$ [BOBÉE et ASHKAR 1991]. Pour $N=20$, l'écart-type sur C_v est donc de l'ordre de 0,1 (pour C_v voisin de 0,5), et l'écart-type sur C_s est de l'ordre de 0,5.

Considérant cette variabilité et la distribution des estimations de C_v et C_s (voir figures 7-12), nous avons choisi de partitionner l'espace $C_s \times C_v$ en faisant trois classes pour C_v et quatre pour C_s . On distinguera pour C_v les classes $C_v < 0,3$, $0,3 \leq C_v < 0,5$ et $C_v \geq 0,5$, alors que l'on considérera pour C_s les classes $C_s < 0$, $0 \leq C_s < 0,75$, $0,75 \leq C_s < 1,25$ et $C_s \geq 1,25$. Un couple $\hat{S} = (\hat{C}_s, \hat{C}_v)$ d'estimations de C_v et C_s fera donc toujours partie de l'une des 12 classes E_1, E_2, \dots, E_{12} définies par le produit des classes de C_v et de C_s .

Les coefficients de variation et d'asymétrie ont été estimés pour chacune des $R (=500)$ séries simulées de taille $N=20$ et de taille $N=40$ provenant des 6 échantillons de référence (notés Y_1, Y_2, \dots, Y_6) utilisés pour la calibration. Pour chaque échantillon de référence Y_j et chaque taille d'échantillon N , le nombre ω de couples $\hat{S} = (\hat{C}_s, \hat{C}_v)$ faisant partie de chaque classe E_1, E_2, \dots, E_{12} , a été noté, permettant d'estimer par ω/R la probabilité subjective que le couple (\hat{C}_s, \hat{C}_v) d'une série X fasse partie d'une classe E_k , sachant qu'elle a été simulée à partir d'un échantillon de référence Y_j , et que l'on note $\Pr[(\hat{C}_s, \hat{C}_v) \in E_k | X \approx Y_j]$. En utilisant l'équation 16, il est alors aisé de calculer la probabilité subjective qu'une série X provienne d'un échantillon de référence Y_j étant donné que les estimations de C_v et C_s font partie d'une classe particulière E_k , et que l'on note $\Pr[X \approx Y_j | (\hat{C}_s, \hat{C}_v) \in E_k]$.

Le tableau 8 donne les valeurs, en pourcentage, attribuées à ces probabilités subjectives. On remarque qu'aucune probabilité n'est inscrite dans la case réservée à la classe $C_v \geq 0,5$ et $C_s < 0$, ceci parce qu'aucun couple (\hat{C}_s, \hat{C}_v) ne s'est trouvé à faire partie de cette classe. On note aussi que, pour une même classe, la somme des pourcentages fixés pour les 6 échantillons de référence ne donne pas toujours exactement à 100 (parfois 99, parfois 101), à cause des erreurs d'arrondi.

Tableau 8: Probabilité, en pourcentage, qu'une série provienne d'un échantillon de référence donné, estimée à partir de \hat{C}_v et \hat{C}_s

Éch. de référence		N=20				N=40			
		$\hat{C}_s < 0$	$< 0,75$	$< 1,25$	$\geq 1,25$	$\hat{C}_s < 0$	$< 0,75$	$< 1,25$	$\geq 1,25$
02FC001	$\hat{C}_v \geq 0,5$		4	10	0		3	0	0
	$0,3 \leq \hat{C}_v < 0,5$	38	25	11	10	54	27	10	5
	$\hat{C}_v < 0,3$	9	11	7	4	8	7	7	5
02FC002	$\hat{C}_v \geq 0,5$		26	18	23		18	18	22
	$0,3 \leq \hat{C}_v < 0,5$	14	20	35	39	9	19	36	39
	$\hat{C}_v < 0,3$	3	3	7	9	1	1	3	5
02EA005	$\hat{C}_v \geq 0,5$		0	0	0		0	0	0
	$0,3 \leq \hat{C}_v < 0,5$	4	15	24	26	2	14	28	37
	$\hat{C}_v < 0,3$	11	26	27	15	6	18	26	19
02EC002	$\hat{C}_v \geq 0,5$		0	0	0		0	0	0
	$0,3 \leq \hat{C}_v < 0,5$	6	3	2	1	3	1	1	1
	$\hat{C}_v < 0,3$	50	31	35	51	60	41	45	43
05PB014	$\hat{C}_v \geq 0,5$		71	72	77		79	82	78
	$0,3 \leq \hat{C}_v < 0,5$	26	25	16	14	18	26	17	10
	$\hat{C}_v < 0,3$	2	1	0	0	0	0	0	0
04LJ001	$\hat{C}_v \geq 0,5$		0	0	0		0	0	0
	$0,3 \leq \hat{C}_v < 0,5$	12	13	11	10	14	13	9	8
	$\hat{C}_v < 0,3$	25	28	24	21	25	32	20	29

En combinant les données du tableau 8 et l'équation 17, il est possible de comparer la performance d'un ensemble de politiques statiques pour chaque classe E_k , telles les 34 politiques statiques P_1, P_2, \dots, P_{34} présentées au tableau 2, si l'on connaît φ_{ijk} , la performance de la politique statique P_j appliquée à l'ensemble des séries simulées à partir de l'échantillon de référence Y_j pour lesquelles le couple (\hat{C}_s, \hat{C}_v) fait partie de la classe E_k . Pour calculer φ_{ijk} , Les $R=500$ séries simulées à partir de chaque échantillon de référence ont été classées en sous-ensembles selon les valeurs des estimations de C_v et C_s , et la

performance de chacune des 34 politiques statiques a été évaluée en utilisant comme mesure de performance l'écart quadratique moyen (EQM) pour chaque sous-ensemble.

Tableau 9: Politique dynamique de sélection d'une combinaison D/M pour estimer x_T

N	T	Politique statique sélectionnée				
		$\hat{C}_s < 0$	$0 \leq \hat{C}_s < 0,75$	$0,75 \leq \hat{C}_s < 1,25$	$\hat{C}_s \geq 1,25$	
20	50	$\hat{C}_v \geq 0,5$		N/MV		W2/MV
		$\hat{C}_v < 0,5$	EV1/MM	1) GEV/MM 2) GEV/PWM 3) GEV/MV		1) LP3/BOB 2) LP3/WRC
		$\hat{C}_v < 0,3$	EV1/PWM		EV1/MM	
40	50	$\hat{C}_v \geq 0,5$		N/MV		1) G2/MV 2) G2/MM
		$\hat{C}_v < 0,5$	G2/MM		1) LP3/BOB 2) LP3/MV 3) LP3/WRC	EV1/MM
		$\hat{C}_v < 0,3$	EV1/MM	EV1/PWM		P3/MM
20	20	$\hat{C}_v \geq 0,5$		N/MV		1) GEV/MM 2) GEV/MV 3) GEV/PWM
		$\hat{C}_v < 0,5$	EV1/MM	W2/MM		W2/MV
		$\hat{C}_v < 0,3$	EV1/PWM			N/MV
40	20	$\hat{C}_v \geq 0,5$		N/MV		EV1/MM
		$\hat{C}_v < 0,5$		G2/MM	1) G2/MV 2) G2/MM	2LN/MV
		$\hat{C}_v < 0,3$	EV1/MM		G2/MM	W2/MV

En pondérant les mesures de performance obtenues pour chaque échantillon de référence selon l'équation 17, la meilleure politique statique a donc pu être identifiée pour chaque classe E_k , chaque période de retour T et chaque taille d'échantillon N. Les résultats de cette comparaison, détaillés au tableau 9, établissent une nouvelle politique de sélection qui peut être employée pour déterminer la combinaison D/M à utiliser pour estimer x_T à partir de T, N, \hat{C}_v et \hat{C}_s . Pour faciliter la lecture de ce tableau, les numéros des politiques sélectionnées ont été remplacés par les combinaisons D/M correspondantes, et certaines classes ont été regroupées. On remarque que les distributions à 2 paramètres (N, EV1, G2 et W2) prédominent, mais que les distributions à 3 paramètres GEV, P3 et LP3 sont aussi présentes

lorsque $T > N$. On note aussi que la méthode d'estimation sélectionnée pour ces distributions à 3 paramètres est toujours en premier lieu la méthode des moments.

3.7 Validation de la politique dynamique de sélection

Nous savons déjà que la politique dynamique de sélection présentée à la section précédente donnera toujours de meilleurs résultats que les politiques statiques pour les séries sur lesquelles elle est calibrée, à condition que la partition de l'espace $C_s \times C_v$ soit assez fine. Nous montrerons d'abord que c'est le cas pour la partition que nous avons effectuée. Il est cependant primordial de valider cette politique en utilisant des données sur lesquelles elle n'a pas été calibrée. C'est à cet effet que nous avons réservé l'échantillon de référence 02PJ007.

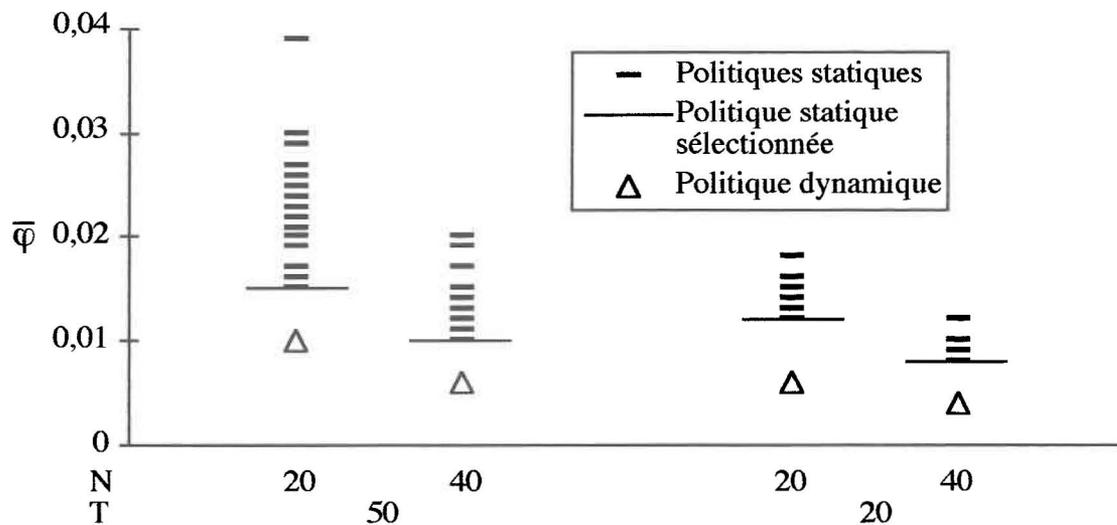


Figure 15: Performance de la politique dynamique de sélection pour les séries de calibration

La figure 15 permet de comparer la performance, mesurée par l'écart quadratique moyen (EQM) en utilisant la FPE de Cunnane, des 34 politiques statiques avec la politique dynamique présentée à la section précédente pour les séries utilisées pour sa calibration. Dans tous les cas, l'utilisation de la politique dynamique de sélection est préférable à l'emploi

de la meilleure politique statique, qui correspond à l'utilisation de EV1/MM pour $T \geq N$ et G2/MM pour $T < N$ (voir tableau 5).

Pour les séries simulées à partir de l'échantillon de référence 02PJ007, l'avantage de la politique dynamique de sélection n'est pas aussi net, comme le montre la figure 16. On remarque que pour $T=50$ et $N=20$ ($T > N$), il existe une politique statique donnant une performance supérieure. Cependant, la politique dynamique de sélection donne toujours des résultats supérieurs à l'application de la politique statique sélectionnée à partir des données de calibration (voir tableau 5), et ce spécialement dans le cas $T > N$. Il est donc préférable de sélectionner une combinaison D/M après avoir estimé C_v et C_s que d'utiliser une politique de sélection qui ne tient compte que de T et de N .

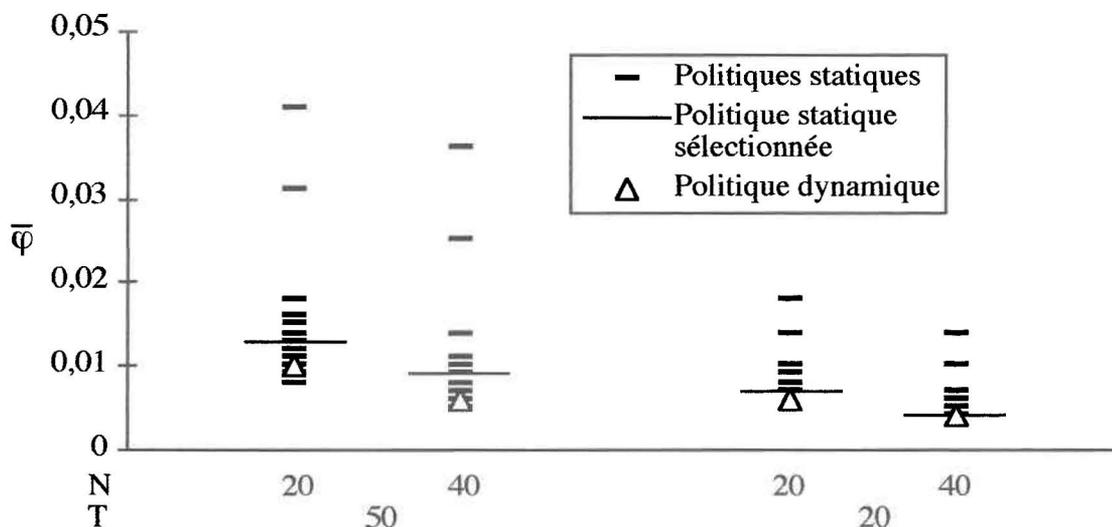


Figure 16: Performance de la politique dynamique de sélection pour les séries de validation

Nous avons donné dans ce chapitre un exemple d'application d'une procédure de comparaison de politiques de sélection d'une combinaison D/M pour estimer un débit x_T , ainsi que d'une politique dynamique de sélection basée sur l'estimation des coefficients d'asymétrie et de variation. Nous discuterons au cours du prochain chapitre des implications des résultats obtenus ainsi que de perspectives d'utilisation ultérieure de ces concepts.

4 DISCUSSION

L'objectif de ce chapitre est d'apporter, en se basant sur les résultats présentés au chapitre précédent, une contribution au débat actuel concernant la sélection d'une distribution statistique pour estimer un débit x_T de période de retour T . Nous discuterons des distributions statistiques qui peuvent être recommandées, ainsi que de l'utilisation de politiques dynamiques autres que celle présentée au chapitre précédent. Nous montrerons aussi comment la procédure de comparaison des politiques de sélection que nous avons développée pourrait être utilisée pour amorcer l'élaboration d'un système-expert permettant de sélectionner une combinaison D/M pour estimer x_T . Nous discuterons ensuite de l'estimation du paramètre de α , qui détermine la forme de la mesure de performance des politiques de sélection, pour terminer par une discussion de l'une des faiblesses de l'approche proposée, soit l'impossibilité de comparer les combinaisons D/M pour de grandes périodes de retour.

4.1 Distributions statistiques recommandées

L'application de la procédure de comparaison a montré (tableaux 4 à 8) que les distributions à deux paramètres, plus spécifiquement les distributions normale (N), gamma (G2) et Gumbel (EV1), donnent de meilleures performances. Il ne faut cependant pas conclure que les distributions à 3 paramètres ne donnent pas de bons résultats, mais plutôt que pour les tailles d'échantillon et les périodes de retour considérées, l'application *systématique* de ces distributions à 2 paramètres est préférable à l'application *systématique* des distributions à 3 paramètres considérées, pour des séries ressemblant statistiquement aux échantillons de référence considérés, selon les différentes mesures de performance utilisées.

D'ailleurs, l'élaboration de la politique dynamique de sélection basée sur les estimations de C_v et C_s a montré (tableau 9) que certaines distributions à 3 paramètres performant mieux que les distributions à 2 paramètres, lorsque les estimations de C_v et C_s appartiennent à certaines classes. Ceci est particulièrement vrai lorsque $T > N$. Il faut donc conclure que lorsque l'on observe les caractéristiques statistiques d'un échantillon avant de sélectionner la combinaison D/M à utiliser, les distributions à 3 paramètres peuvent donner de bons

résultats, surtout pour l'extrapolation. Il faut cependant éviter de les utiliser systématiquement.

Cette dernière affirmation semble en contradiction avec les politiques officielles britanniques [NERC 1975] et américaines [USWRC 1967, 1987], qui recommandent respectivement l'application systématique des distributions à 3 paramètres GEV et LP3. Par contre, on remarque au tableau 9 que ces distributions sont précisément celles qui sont sélectionnées pour l'extrapolation ($T > N$), pour certaines valeurs de \hat{C}_v et \hat{C}_s . Les distributions statistiques recommandées par les autorités britanniques et américaines sont donc efficaces, à condition de ne pas les utiliser systématiquement.

Il est inefficace d'utiliser systématiquement la même distribution pour estimer x_T . Cependant, s'il est nécessaire de procéder ainsi, une distribution à 2 paramètres devrait être utilisée. La sélection de la distribution devrait tenir compte du poids relatif accordé à la sous-estimation et à la surestimation, pris en compte par le paramètre α de l'équation 12. Le poids relatif accordé aux erreurs de grande amplitude, représenté par le paramètre r , semble avoir peu d'importance (voir tableau 6). Comme le montre le tableau 7, en général la distribution normale (N) est préférable si $\alpha < 1$, alors que la distribution Gumbel (EV1) est préférable si $\alpha \geq 1$. Conséquemment la distribution EV1 peut être utilisée si l'on ne veut pas trop risquer de sous-estimer x_T , et la distribution normale peut servir si l'on préfère sous-estimer que surestimer x_T . Il est cependant bien préférable, comme le montrent les figures 15 et 16, d'utiliser une politique dynamique de sélection.

Il est important de noter que les résultats obtenus dans ce travail proviennent uniquement de l'analyse de 7 stations de mesure du débit maximum annuel. Conséquemment, les recommandations présentées dans cette section ne s'appliquent que pour des sites similaires. Il serait hasardeux de prétendre que les résultats obtenus permettent de sélectionner avec assurance une distribution statistique pour estimer la distribution des débits à n'importe quelle site du Québec et de l'Ontario. L'objectif de l'exemple d'application présenté n'était que de présenter de façon concrète la méthodologie, non de déterminer une politique de sélection robuste pour l'ensemble des sites du Québec et de l'Ontario.

Les conclusions générales que l'on peu retenir se limitent à reconnaître que pour les sites étudiés la prise en compte des caractéristiques statistiques des séries de débit maximum annuel (précisément C_v et C_s) permet de diminuer le coût moyen des erreurs d'estimation de x_T . Puisque l'utilisation systématique d'une même politique statique, choisie en minimisant le coût moyen des erreurs d'estimations de x_T sur toute la plage de C_v et C_s permet de construire une politique de sélection indépendante de ces paramètres, nous pouvons en conclure que ces paramètres sont pertinents lors du choix d'une combinaison D/M pour estimer x_T .

4.2 Utilisation de d'autres politiques dynamiques

D'autres politiques dynamiques de sélection que celle présentée à la section 2.4 (reposant sur l'estimation de C_v et C_s) peuvent être utilisées pour choisir la combinaison D/M à utiliser pour estimer x_T . On pourrait par exemple envisager l'utilisation du critère d'Akaike [AKAIKE 1974], qui repose sur l'estimation de la fonction de vraisemblance. Puisque nous avons développé une procédure de comparaison de politiques et non de combinaisons D/M, celle-ci peut être directement utilisée pour évaluer la performance de toute politique de sélection. Ainsi, il serait aisé de vérifier si d'autres politiques dynamiques donnent de meilleurs résultats.

4.3 Application au développement d'un système-expert

La procédure de comparaison de politiques de sélection présentée dans ce travail pourrait aussi servir de base à l'élaboration d'une politique de sélection basée sur les méthodes heuristiques utilisée par des experts pour sélectionner une combinaison D/M. En effet, pour une mesure de performance fixée (i.e. une valeur fixée du paramètre α), la procédure de comparaison pourrait être utilisée pour évaluer la performance de différents experts, permettant ainsi de déterminer quels experts font la meilleure sélection. Une fois ces experts identifiés, un système-expert pourrait être construit pour reproduire leur raisonnement. La procédure de comparaison pourrait également être utilisée pour déterminer si le système-expert est aussi efficace que les experts.

Il serait cependant intéressant de construire un système-expert pouvant s'adapter à différentes mesures de performance, c'est à dire à différentes valeurs du paramètre α . Pour cela, nous proposons de procéder comme suit. Au lieu de déterminer les meilleurs experts pour une valeur fixée de α , on pourrait supposer que chaque expert sélectionne en fait dans chaque cas la meilleure combinaison D/M, selon une mesure de performance qui lui est propre. Si l'on suppose que cette mesure de performance peut être décrite approximativement par l'équation 12, on pourrait trouver la valeur de α maximisant la performance de chaque expert, telle que mesurée par notre procédure de comparaison. Ainsi, on pourrait déterminer le poids relatif que chaque expert accorde aux erreurs de surestimation par rapport aux erreurs de sous-estimation, étudier les méthodes heuristiques utilisées par chaque expert, et ainsi bâtir un ensemble de petits systèmes-experts correspondant aux différentes valeurs de α . L'architecture interne d'un tel système-expert a été discutée par [FORTIN et BOBÉE 1993].

4.4 Estimation du paramètre α

La principale difficulté que présente, pour un utilisateur éventuel, l'application de la procédure de comparaison développée dans ce travail est l'estimation du paramètre α . Si l'utilisation de graphiques tels que ceux présentés à la figure 14 n'est pas suffisante pour permettre d'estimer ce paramètre, une façon de procéder serait d'amener l'utilisateur à identifier quel expert en sélection il préfère imiter, puisque l'on peut déterminer la valeur de α utilisée par différents experts. Différents critères pourraient être utilisés pour déterminer l'expert à imiter, comme la réputation des experts disponibles et l'utilisation faite de l'estimation de x_T . Dans l'hypothèse où un système-expert serait construit pour sélectionner la combinaison D/M à utiliser, une des fonctions de ce système-expert pourrait être de questionner l'utilisateur de façon à identifier l'expert que le système devrait imiter.

4.5 Application à de grandes périodes de retour

Le procédé de simulation par suréchantillonnage utilisé pour construire la procédure de comparaison se distingue à la fois par ses forces et ses faiblesses. Il a l'avantage de ne nécessiter aucune hypothèse sur la distribution réelle des débits maximums annuels, mais nécessite de disposer d'une banque de donnée contenant un nombre suffisant de séries

pouvant être utilisées comme échantillons de référence. La plus grande faiblesse de cette approche est cependant de ne pas permettre d'étudier des périodes de retour T supérieures au nombre d'observations contenues dans les échantillons de référence.

Puisque toutes les distributions statistiques utilisées pour prédire x_T sont des modèles empiriques, il est en fait peu gênant de ne pouvoir étudier la performance de différentes combinaisons D/M pour de grandes périodes de retour puisqu'il est de toutes façons très dangereux de prédire, à l'aide d'un modèle empirique ajustés à des observations, le comportement d'un système dans des conditions non encore observées. Il n'existe par exemple au Canada, et tout probablement dans le monde entier, aucun site auquel le débit a été mesuré systématiquement pendant plusieurs centaines d'années. Il est donc inutile de tenter de prédire le débit x_{100} , qui est le débit qui est dépassé *en moyenne* tous les 100 ans, uniquement à partir de mesures du débit maximum annuel.

Par contre, il peut arriver que l'on dispose à certains sites d'une région d'information complémentaire de type historique, paléohydrologique ou encore météorologique, permettant d'allonger la période d'observation des débits extrêmes de façon à autoriser l'estimation non paramétrique de x_T pour des valeurs de T supérieures à la période d'enregistrement systématique du débit.

Il faudrait, pour permettre de comparer la performance de combinaisons D/M à estimer x_T pour ces grandes périodes de retour, réussir à simuler de façon non paramétrique des séries de débits maximums annuels à partir de l'ensemble des informations disponibles au site. Les procédures de simulation par suréchantillonnage ne s'appliqueraient plus: il faudrait tenir compte des informations autres que la série des mesures du débit maximum annuel. Une façon évidente serait de tenter de reconstituer cette série pour l'ensemble des années pour lesquelles d'autres informations sont disponibles. Une autre approche, suggérée par [BERNIER 1993], consiste à se former une idée *a priori* de la distribution des débits à partir d'informations autres que les mesures de débit maximums annuel, et de simuler des observations en tirant au hasard dans un mélange de cette distribution et de la distribution prépostérieure obtenue uniquement à partir des mesures de débit maximum annuel.

CONCLUSION

En utilisant la simulation par suréchantillonnage, il est possible de construire une procédure de comparaison des distributions de crue reposant sur des hypothèses vérifiables. L'application de cette procédure de comparaison à 7 stations de mesure du débit maximum annuel des provinces de l'Ontario et du Québec a montré que pour ces stations l'utilisation systématique de certaines distributions à 2 paramètres est préférable à l'utilisation systématique de distributions à 3 paramètres, incluant la distribution généralisée des valeurs extrêmes (GEV) et la log-Pearson type III (LP3), souvent recommandées pour l'estimation de x_T . Le choix de la distribution à utiliser dépend cependant du poids relatif accordé aux erreurs de surestimation en comparaison des erreurs de sous-estimation. La distribution normale est en générale préférable si plus d'importance est accordée aux erreurs de surestimation, alors que la distribution Gumbel (EV1) est préférable dans le cas contraire. Lorsque le même poids est accordé à la sous-estimation et à la surestimation, la distribution Gumbel ajustée à l'aide de la méthode de moments (EV1/MM) est recommandée pour l'extrapolation ($T > N$), alors que la distribution Gamma, aussi ajustée à l'aide de la méthode des moments (G2/MM) est recommandée pour l'interpolation ($T < N$).

L'application systématique de la même distribution n'est par contre pas recommandée, car on obtient de meilleurs résultats en tenant compte des caractéristiques statistiques des échantillons, telles les estimations des coefficients d'asymétrie (C_s) et de variation (C_v). Lorsque ces estimations sont employées pour sélectionner la distribution, des distributions à 3 paramètres, incluant les distributions GEV et LP3, sont fréquemment sélectionnées, spécialement pour l'extrapolation (lorsque la période de retour T dépasse le nombre d'observations disponibles). Il ne faut donc pas rejeter ces distributions, mais ne les utiliser que lorsqu'elles sont nécessaires. Notons que les résultats obtenus sont entièrement conditionnés par les données utilisées; l'application de la méthodologie développée dans ce travail à d'autres stations pourrait donner des résultats différents.

La procédure de comparaison développée dans ce travail peut être utilisée pour déterminer la performance de toute politique de sélection, incluant les heuristiques utilisées par les experts. Ainsi, elle pourrait aussi servir à l'élaboration d'un système-expert sélectionnant la distribution statistique à utiliser pour estimer x_T .

ANNEXE A: MÉTHODES D'ESTIMATION

L'estimation des paramètres d'une distribution quelconque à partir d'un échantillon d'observations par trois méthodes différentes est présentée dans cette annexe. Les trois méthodes sont la méthode du maximum de vraisemblance (MV), la méthode des moments (MM) et la méthode des moments pondérés (*Probability weighted moments* ou PWM).

A.1 Méthode du maximum de vraisemblance

Soit un échantillon \mathbf{X} de réalisations $\{x_1, x_2, \dots, x_N\}$ de variables aléatoires $\{X_1, X_2, \dots, X_N\}$ indépendantes et identiquement distribuées (IID) selon une fonction de densité de probabilité (fdp) égale à $f(x|\underline{\theta})$, où $\underline{\theta}$ est un vecteur de paramètres. La probabilité d'observer cet échantillon ou un échantillon semblable est proportionnelle à la fdp jointe de $\{X_1, X_2, \dots, X_N\}$ évaluée en $\{x_1, x_2, \dots, x_N\}$. Parce que les variables aléatoires sont indépendantes, celle-ci est égale au produit des fdp individuelles. On nomme cette quantité $L(\mathbf{X})$ vraisemblance d'un échantillon:

$$L(\mathbf{X}) = \prod_{i=1}^N f(x_i) \quad (\text{A.1})$$

Les estimateurs $\hat{\underline{\theta}}$ du maximum de vraisemblance (MV) des paramètres $\underline{\theta}$ calculés à partir d'un échantillon \mathbf{X} sont les valeurs qui maximisent la vraisemblance de cet échantillon. En pratique, les estimateurs du maximum de vraisemblance sont souvent obtenus en maximisant le logarithme de la fonction de vraisemblance, ce qui est équivalent mais souvent plus simple algébriquement:

$$\hat{\underline{\theta}} = \max_{\underline{\theta}} \{\log L(\mathbf{X})\} = \max_{\underline{\theta}} \left\{ \sum_{i=1}^N \log [f(x_i|\underline{\theta})] \right\} \quad (\text{A.2})$$

La solution $\hat{\underline{\theta}}$ de cette équation est habituellement obtenue en égalant à zéro les dérivées partielles de $L(\cdot)$ ou $\log L(\cdot)$ par rapport à chacun des paramètres, et en vérifiant qu'il s'agit bien d'un maximum. Le système d'équations obtenu est souvent non linéaire et pose parfois

de grandes difficultés numériques. De plus il n'existe quelquefois pas de solution à ce système d'équations.

A.2 Méthode des moments

La méthode des moments consiste à égaler les expressions théoriques des moments de la distribution considérée (qui sont fonctions des paramètres) aux estimations de ces moments obtenues à partir de l'échantillon. Après avoir défini les moments d'une distribution et leurs estimateurs, nous montrerons comment s'applique cette méthode en pratique.

A.2.1 Moments d'une distribution

Le moment non centré d'ordre r (μ'_r) d'une variable aléatoire X est l'espérance mathématique de X^r , alors que le moment centré d'ordre r (μ_r) est l'espérance mathématique de $(X - \mu)^r$. Notons que $\mu_1 = E[X]$ est l'espérance mathématique de X , que l'on note par convention simplement μ .

$$\begin{aligned}\mu'_r &= E[X^r] \\ \mu_r &= E[(X - \mu)^r]\end{aligned}\tag{A.3}$$

Quelques moments et rapports de moments couramment utilisés sont représentés par des symboles différents. En plus de la moyenne qui est notée simplement μ , la variance, ou moment centré d'ordre 2 (μ_2), est habituellement notée $\text{var}(X)$ ou σ^2 . Sa racine carrée, σ , est l'écart-type de X . On utilise aussi couramment les coefficients de variation (Cv) et d'asymétrie (Cs), définis respectivement par le rapport de l'écart-type sur la moyenne ($Cv = \sigma/\mu$) et par le rapport du moment centré d'ordre 3 sur le cube de l'écart-type ($Cs = \mu_3/\sigma^3$).

On note l'estimateur du moment non centré d'ordre r m'_r et m_r l'estimateur du moment centré d'ordre r . On peut les calculer à partir d'un échantillon \mathbf{X} d'observations $\{x_1, x_2, \dots, x_N\}$ en utilisant les relations suivantes:

$$\begin{aligned}
 m'_r &= \frac{1}{N} \sum_{i=1}^N x_i \\
 m_r &= \frac{1}{N} \sum_{i=1}^N (x_i - m'_1)^r
 \end{aligned}
 \tag{A.4}$$

Alors que m'_r constitue un estimateur non biaisé de μ'_r , m_r est par contre un estimateur biaisé de μ_r . Il existe un estimateur non biaisé de μ_r obtenu en appliquant une correction à l'estimateur m_r mais qui a cependant une variance supérieure. Nous noterons cet estimateur \tilde{m}_r et le nommerons estimateur du moment centré d'ordre r corrigé pour le biais. On le calcule ainsi:

$$\tilde{m}_r = \frac{N^r}{(N-1)(N-2)\dots(N-r)} m_r
 \tag{A.5}$$

A.2.2 Application pratique de la méthode des moments

Pour appliquer la méthode des moments, on doit choisir, pour une distribution à p paramètres, p moments indépendants dont la valeur dépend des paramètres. On utilise généralement des moments d'ordre peu élevé parce que la variabilité des estimateurs de ces moments augmente avec r . La méthode habituelle consiste prendre le moment non centré d'ordre 1 ainsi que les moments centrés corrigés pour le biais d'ordre 2 à p . Les distributions que nous utiliserons pour estimer un débit x_T de période de retour T ont toutes deux ou trois paramètres. Dans le cas d'une distribution à deux paramètres, on utilise la moyenne, μ , ainsi que la variance, σ^2 . Dans le cas d'une distribution à trois paramètres, on ajoute généralement le coefficient d'asymétrie, C_s , qui est une fonction de σ^2 et de μ_3 . Les estimateurs de μ, σ^2 et C_s calculés à partir de l'échantillon sont notés respectivement \bar{x}, s^2 et \hat{C}_s . On estime ces estimateurs à partir d'un échantillon \mathbf{X} d'observations $\{x_1, x_2, \dots, x_N\}$ de la façon suivante:

$$\begin{aligned}
 \bar{x} &= m'_1 \\
 s^2 &= \tilde{m}_2 \\
 \hat{C}_s &= \frac{\tilde{m}_3}{s^3}
 \end{aligned}
 \tag{A.6}$$

Notons qu'il existe plusieurs autres formules pour l'estimation de Cs en hydrologie, voir par exemple [YEVJEVICH et OBEYSEKERA 1984]. Elles tentent toutes de diminuer le fort biais que l'on cause en estimant le Cs simplement par $m_3 / m_2^{3/2}$. Celle que nous utilisons est nommée Cs1 par [BOBÉE ET ROBITAILLE 1977]. Elle donne un estimateur biaisé de Cs même si l'estimateur est fonction de deux estimateurs non biaisés, \tilde{m}_2 et \tilde{m}_3 . En effet, le rapport de deux estimateurs non biaisés est généralement biaisé.

La méthode des moments, telle que nous l'appliquerons consiste donc à égaler les valeurs théoriques μ , σ^2 et Cs, qui sont fonctions des paramètres de la distribution, aux estimateurs de ces moments \bar{x} , s^2 et \hat{C}_s obtenus à partir de l'échantillon. Les estimateurs $\hat{\theta}$ de la méthode des moments (MM) des paramètres θ sont conséquemment les valeurs qui résolvent ce système d'équations:

$$\left. \begin{array}{l} \mu(\hat{\theta}) = \bar{x} \\ \mu_2(\hat{\theta}) = s^2 \\ C_s(\hat{\theta}) = \hat{C}_s \end{array} \right\} \begin{array}{l} \text{si } p = 2 \\ \text{si } p = 3 \end{array} \quad (\text{A.7})$$

Il est nécessaire, pour utiliser la méthode des moments, de connaître la relation entre μ , σ^2 et Cs, et les paramètres de la distribution. Notons qu'il arrive que certains moments d'une distribution n'existent pas. Dans ces cas, il n'est pas possible d'utiliser cette méthode. Par exemple, [BOBÉE 1975] a montré que lorsque le paramètre α de la distribution log Pearson type III (LP3) est positif, le moment (centré ou non centré) d'ordre r n'existe que si α est plus grand que r . En pratique, puisque la distribution LP3 comporte trois paramètres et que l'estimation du Cs nécessite l'évaluation du moment centré d'ordre 3, la méthode des moments ne peut être utilisée pour cette distribution que si $\alpha < 0$ ou $\alpha > 3$.

A.3 Méthode des moments pondérés

Le principe de la méthode des moments pondérés est très semblable à celui de la méthode des moments. Cette méthode consiste à égaler les expressions théoriques des moments pondérés à leur estimation obtenue à partir de l'échantillon. Le moment pondéré d'ordre r d'une variable aléatoire X de densité cumulée $F(X)$ est défini ainsi [GREENWOOD *et al.* 1979]:

$$\beta_r = E\left\{X[F(X)]^r\right\} \quad (\text{A.8})$$

Un estimateur non biaisé de β_r est [LANDWEHR *et al.* 1979]:

$$b_r = \frac{1}{N} \sum_{i=1}^N \frac{(i-1)(i-2)\cdots(i-r)}{(N-1)(N-2)\cdots(N-r)} x_{(i)} \quad (\text{A.9})$$

où les $x_{(i)}$ sont les N observations utilisées pour estimer les moments pondérées, rangées en ordre croissant de telle façon que $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$.

On utilise les moments pondérés d'ordre 0 à $p-1$ pour estimer les paramètres d'une distribution comportant p paramètres, à condition que ces moments soient tous indépendants et fonctions des paramètres de la distribution. Comme nous l'avons mentionné précédemment, les distributions que nous utiliserons auront toutes deux ou trois paramètres. Les estimateurs $\hat{\theta}$ de la méthode des moments pondérés (PWM) seront donc les valeurs qui solutionnent ce système d'équations:

$$\left. \begin{aligned} \beta_0(\hat{\theta}) &= \frac{1}{N} \sum_{i=1}^N x_{(i)} \\ \beta_1(\hat{\theta}) &= \frac{1}{N} \sum_{i=1}^N \frac{i-1}{N-1} x_{(i)} \\ \beta_2(\hat{\theta}) &= \frac{1}{N} \sum_{i=1}^N \frac{(i-1)(i-2)}{(N-1)(N-2)} x_{(i)} \end{aligned} \right\} \begin{array}{l} p=2 \\ \\ p=3 \end{array} \quad (\text{A.10})$$

Pour pouvoir utiliser cette méthode, il est nécessaire de pouvoir déterminer le lien entre les moments pondérés et les paramètres de l'échantillon. En pratique, il n'est possible d'obtenir des expressions analytiques simples pour les moments pondérés d'une distribution que s'il existe une forme explicite de la densité cumulée inverse $F^{-1}(\cdot)$ de cette distribution.

ANNEXE B: COMPARAISON DES PROCÉDÉS DE SIMULATION PAR SURÉCHANTILLONNAGE

Comme nous l'avons expliqué au chapitre 2, nous utilisons dans ce travail une technique de suréchantillonnage par tirage Pólya pour simuler des séries de débits maximums annuels. Cette technique n'étant pas couramment utilisée, il est utile de la comparer à deux autres techniques de simulation par suréchantillonnage, le tirage sans remise (exhaustif) [POTTER et LETTENMAIER 1990] et le tirage avec remise (non exhaustif), ou *Bootstrap* [EFRON 1979]. Nous avons déjà expliqué à la section 2.1.2 l'avantage d'utiliser le tirage de Pólya sur le plan théorique. Nous justifierons ici son utilisation par un exemple simple.

Notons premièrement qu'il est évident que le tirage sans remise est l'approche à utiliser pour simuler des échantillons si l'échantillon de référence est exactement la population d'où proviennent les observations, ce qui ne peut arriver que si celle-ci est finie. La population des débits maximums annuels étant conceptuellement infinie, cette approche doit donc être évitée. Nous comparerons donc le *Bootstrap* avec la simulation par tirage de Pólya.

L'utilisation du tirage avec remise, ou *Bootstrap*, demande de supposer que la population complète est une répétition infinie de l'échantillon de référence. Comme dans le cas du tirage sans remise, cela suppose que l'information comprise dans l'échantillon de référence est suffisante pour permettre de connaître exactement la population, mais permet de supposer que cette population est infinie. Comme nous le verrons, l'utilisation du tirage de Pólya ne demande que de supposer que l'échantillon de référence contient assez d'information pour pouvoir quantifier l'incertitude qui reste sur la forme de la distribution réelle de la population une fois que l'échantillon a été utilisé pour l'approximer.

Pour simplifier les choses, supposons que l'on veuille simuler un échantillon de taille n à partir d'un échantillon de référence contenant N réalisations d'une variable aléatoire de Bernoulli, tel que $n < N$. Rappelons qu'une variable aléatoire de Bernoulli ne peut prendre que deux valeurs, 0 ou 1. Ce type d'échantillon de référence est un cas particulier des échantillons de référence que nous utilisons dans ce travail, qui contiennent des observations pouvant prendre n'importe quelle valeur réelle. L'avantage d'utiliser des variables aléatoires

de Bernouilli est que l'information contenue dans l'échantillon de référence est résumée entièrement par N , le nombre d'observations, et q , la proportion (entre 0 et 1) d'observations égales à 0 contenues dans l'échantillon. Supposons enfin que la population d'où proviennent les observations contenues dans l'échantillon de référence est infinie et contient une proportion Q d'observations égales à 0.

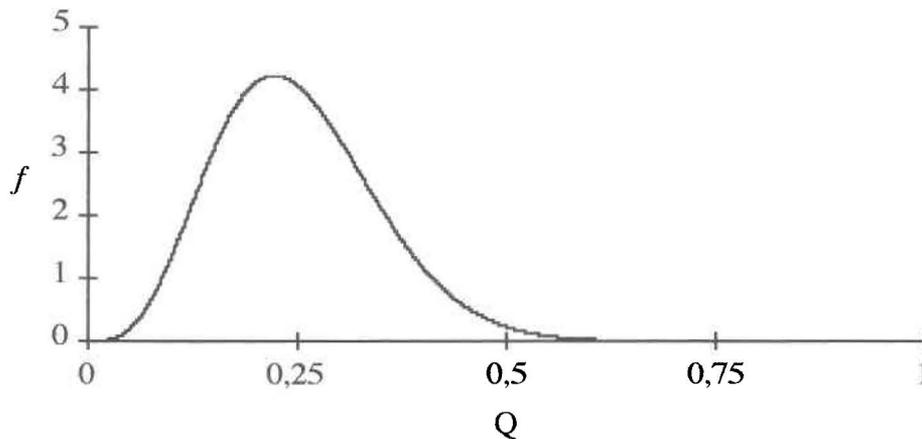


Figure B1: Distribution *a posteriori* de la proportion Q pour $N=20$ et $q=0,25$

Les observations simulées par *Bootstrap* à partir de cet échantillon de référence auront une proportion q d'observations égales à 0, et une proportion $1-q$ d'observations égales à 1. En fait, la simulation par *Bootstrap* revient à supposer que $Q=q$. Pourtant, q n'est qu'une estimation de Q . La simulation par tirage de Pólya tient compte de cette incertitude qui subsiste sur la valeur de Q . [BERNIER 1993] montre que dans un cadre bayésien, l'incertitude peut être représentée par une distribution *a posteriori* $f(Q|q)$ Bêta de paramètres $\alpha_1=Nq$ et $\alpha_2=N(1-q)$. Rappelons que la fonction de densité de probabilité (fdp) de la distribution Bêta, $\beta(x;\alpha_1,\alpha_2)$, est donnée par:

$$\beta(x;\alpha_1,\alpha_2) = \frac{(\alpha_1 + \alpha_2 - 1)!}{(\alpha_1 - 1)!(\alpha_2 - 1)!} \cdot x^{\alpha_1 - 1} (1 - x)^{\alpha_2 - 1} \quad (\text{B.1})$$

Par exemple, pour $N=20$ et $q=0,25$ la figure B1 représente $f(Q|q)=\beta(x;5,15)$, la fonction de densité de probabilité de l'incertitude sur Q . On remarque que la proportion d'observations égales à 0 dans la population peut vraisemblablement se situer entre 0,1 et 0,5. D'un point de vue Bayésien, la façon d'intégrer l'incertitude sur Q lors de la simulation d'observations, est d'utiliser la distribution prévisionnelle ou prépostérieure correspondant à la distribution *a posteriori*, qui est précisément dans ce cas une distribution de Pólya [BERNIER 1993]. La simulation par tirage de Pólya est donc préférable au *Bootstrap*, puisqu'elle prend en compte l'incertitude sur Q .

Il est intéressant de voir l'effet du tirage de Pólya sur les échantillons simulés. Comme l'a montré [BERNIER 1993], l'effet est négligeable lorsque $N \gg n$. Lorsque ce n'est pas le cas, le tirage de Pólya a pour effet d'accroître la probabilité des événements rares. Ainsi, les échantillons simulés comporteront plus d'observations dont la fréquence est faible dans l'échantillon de référence qu'en utilisant la technique *Bootstrap*. Il s'agit donc d'événements rares à l'intérieur de l'échantillon de référence, et non par rapport à une distribution parente inconnue.

Considérons par exemple la probabilité d'obtenir, à partir d'un échantillon de référence \mathbf{Y} constitué de N observations ayant toutes des valeurs distinctes, un échantillon \mathbf{X} constitué de n répétitions de la même observation. Si l'on utilise la technique *Bootstrap*, la probabilité de simuler un tel échantillon est N^{-n} . Par contre, si l'on utilise le tirage de Pólya, la probabilité de simuler un tel échantillon est, par définition, $(1 \cdot 2 \cdot \dots \cdot n) / [N \cdot (N+1) \cdot \dots \cdot (N+n-1)]$, c'est à dire $n!(N-1)! / (N+n-1)!$. Puisque $n!(N-1)! / (N+n-1)! \geq N^{-n}$, la probabilité de simuler l'échantillon \mathbf{X} en utilisant le tirage de Pólya est donc plus grande ou égale à la probabilité de simuler ce même échantillon par *Bootstrap*.

La preuve de cette dernière affirmation est assez directe. Si l'on pose $A(r) = r!(N-1)! / (N+r-1)!$ et $B(r) = N^{-r}$, il faut montrer que $A(n) \geq B(n)$ pour tout n et pour tout N , sachant que $N \geq n \geq 1$. On peut procéder par induction. Supposons que $A(r) \geq B(r)$ pour une certaine valeur de r . Puisque $A(r+1) = (r+1)A(r) / (N+r)$ et que $B(r+1) = B(r) / N$, alors $A(r+1) \geq B(r+1)$ ssi $A(r)/B(r) \geq (N+r)/(N+r+N)$. Puisque $A(r)/B(r) \geq 1$, alors $A(r+1) \geq B(r+1)$ si $(N+r)/(N+r+N) \leq 1$, ce qui se prouve sans difficulté. Conséquemment, si $A(r) \geq B(r)$, alors $A(r+1) \geq B(r+1)$. Puisque de plus $A(1) = B(1) = 1/N$, la preuve par induction est complète.

ANNEXE C: BANQUE DE DONNÉES UTILISÉE

Les séries de données réelles utilisées dans ce travail proviennent de la banque de données HYDAT 3.0 [ENVIRONNEMENT CANADA 1990], qui comprend les données compilées par le ministère fédéral de l'environnement sur les eaux de surfaces jusqu'en 1990. Dans le cadre d'un projet de recherche portant sur la régionalisation des crues (CRSNG-Stratégique) 189 stations de mesure de débit ont été sélectionnées en respectant les trois critères suivants [MATHIER *et al.* 1993]:

- 1) stations situées au Québec ou en Ontario mesurant le débit quotidiennement;
- 2) stations ayant un régime naturel, non influencées par les marées, des barrages ou le refoulement des confluent du cours d'eau;
- 3) stations possédant 20 ans ou plus d'enregistrement effectif du débit journalier moyen.

Comme nous l'avons expliqué au début du chapitre 3, 27 séries pour lesquelles les observations ne semblaient pas identiquement distribuées ou indépendantes ont été par la suite éliminées, laissant une banque de données de 162 stations.

Nous utilisons les 7 plus longues séries de données réelles pour simuler des débits, et les plus courtes pour approximer la distribution des estimateurs des coefficients C_s et C_v . Au lieu d'énumérer toutes les observations des 162 séries, nous nous contenterons donc de donner pour chaque station de mesure son identificateur (donné par Environnement Canada), le nombre d'observations disponibles, les estimations des coefficients C_s et C_v obtenues en utilisant d'abord toutes les observations, ensuite en n'utilisant que les 20 observations les plus récentes, et enfin (pour les séries comportant plus de 40 observations) en n'utilisant que les 40 observations les plus récentes.

On retrouve tout d'abord au tableau C1 les stations de mesure ayant moins de 40 observations, pour lesquelles on ne peut calculer C_s et C_v que pour la série entière et pour les 20 dernières années d'observation. Le tableau C2 donne ensuite les stations de mesure ayant 40 observations et plus. Les observations des 7 stations les plus longues, utilisées comme échantillons de référence, sont par contre toutes énumérées au tableau C3. Notons que ces stations sont aussi incluses dans le tableau C2.

Tableau C1: Séries de données réelles de moins de 40 observations

Station	N	selon série entière		sur 20 dernières observations		Station	N	selon série entière		sur 20 dernières observations	
		Cs	Cv	Cs	Cv			Cs	Cv		
05PC011	39	0,78	0,60	1,17	0,72	02EC011	24	0,19	0,36	0,05	0,37
02GA018	38	0,48	0,41	0,42	0,42	02FF007	24	1,04	0,45	0,91	0,45
02AB008	37	1,15	0,59	1,22	0,67	02GC002	24	1,28	0,49	0,27	0,42
02FC011	37	1,35	0,48	0,89	0,50	02GD019	24	1,46	0,39	1,38	0,40
02HC009	37	1,86	0,61	0,15	0,38	02GG005	24	1,64	0,51	0,94	0,44
04JC003	37	0,59	0,25	0,93	0,21	02GG006	24	1,05	0,58	1,12	0,62
02OE027	36	1,60	0,32	1,63	0,38	02HC027	24	0,81	0,39	0,85	0,39
02OJ001	36	-0,51	0,18	-0,81	0,19	02HC030	24	0,83	0,47	0,62	0,43
03AC001	36	1,69	0,43	2,48	0,50	02JB003	24	1,38	0,25	1,32	0,27
04JA002	36	0,18	0,22	0,22	0,22	02NE011	24	0,71	0,36	0,87	0,39
01BH002	35	0,94	0,28	1,09	0,26	02PD012	24	0,67	0,33	0,73	0,35
02LB008	35	1,18	0,51	0,84	0,54	03ED001	24	0,54	0,21	0,45	0,20
02OA035	35	0,26	0,33	0,56	0,35	04CA002	24	0,44	0,37	0,25	0,35
02QC001	35	1,28	0,40	1,42	0,40	04DA001	24	1,22	0,46	1,51	0,47
05PC010	35	0,90	0,50	0,49	0,51	04DB001	24	0,77	0,42	0,95	0,45
02HA006	34	0,45	0,40	0,06	0,35	04FA003	24	0,85	0,33	1,05	0,35
02FA001	33	-0,11	0,32	0,06	0,31	02BB002	23	-0,24	0,27	-0,20	0,29
02HL004	33	0,74	0,31	0,78	0,30	02BF001	23	1,45	0,40	1,38	0,40
02JB004	33	0,99	0,32	1,09	0,35	02BF002	23	0,80	0,39	1,00	0,42
02VA001	32	3,29	0,69	0,97	0,37	02EA010	23	0,85	0,28	0,69	0,30
02VC001	32	-0,20	0,28	-0,29	0,28	02FE008	23	0,57	0,38	0,49	0,41
02HD006	31	2,17	0,75	2,15	0,76	02GE005	23	1,71	0,51	1,90	0,48
02HD008	31	1,12	0,65	1,01	0,64	02PD013	23	1,06	0,39	0,62	0,36
02CF007	30	1,20	0,41	1,04	0,42	02PG004	23	0,15	0,35	0,26	0,34
02GC010	30	0,78	0,53	0,82	0,50	02QB011	23	0,89	0,30	0,97	0,30
02HC013	30	1,94	0,62	1,78	0,62	02UC002	23	0,01	0,27	-0,51	0,24
02MC001	30	0,02	0,33	-0,19	0,26	03JB001	23	0,33	0,28	0,47	0,29
05QA004	29	0,74	0,48	0,94	0,51	03KA001	23	1,07	0,29	0,95	0,30
02GB009	28	0,58	0,45	0,62	0,41	04CA003	23	1,06	0,45	1,16	0,47
02HC018	28	0,08	0,47	0,01	0,36	04DC001	23	0,81	0,43	0,78	0,47
02HC019	28	0,92	0,49	1,21	0,40	04FC001	23	1,42	0,43	1,49	0,46
02HC023	28	0,29	0,47	0,06	0,41	04JD005	23	-0,46	0,26	-0,59	0,27
02HC025	28	1,24	0,48	0,46	0,37	02EC010	22	0,49	0,36	0,45	0,38
02HJ001	28	1,19	0,42	0,26	0,33	02GA017	22	-0,34	0,37	-0,16	0,39
02NE007	28	0,24	0,24	0,46	0,28	02PA007	22	-0,06	0,21	-0,01	0,22
02RD002	28	0,27	0,21	-0,02	0,22	02PD014	22	0,93	0,43	1,06	0,42
02HC028	27	-0,04	0,34	-0,01	0,28	02PE009	22	1,48	0,65	1,70	0,63
02QA002	27	0,95	0,39	1,39	0,35	03AB002	22	1,81	0,25	1,70	0,26
02RD003	27	0,03	0,24	-0,23	0,25	04CE002	22	0,74	0,36	0,67	0,38
02RF001	27	0,32	0,27	0,19	0,27	04DC002	22	0,67	0,34	0,52	0,34
02RF002	27	0,32	0,26	0,06	0,28	04EA001	22	0,68	0,41	0,63	0,43
02RF006	27	0,33	0,22	0,42	0,25	04JF001	22	1,48	0,41	1,73	0,41
03AC004	27	0,91	0,22	0,95	0,23	02BA002	21	-0,50	0,32	-0,45	0,33

Tableau C1: Séries de données réelles de moins de 40 observations (suite)

Station	N	selon série entière		sur 20 dernières observations		Station	N	selon série entière		sur 20 dernières observations	
		Cs	Cv	Cs	Cv			Cs	Cv		
03DD002	27	0,05	0,22	0,35	0,21	02BB003	21	-0,09	0,34	-0,02	0,35
03LF002	27	0,71	0,37	0,84	0,42	02HC031	21	1,61	0,51	1,55	0,51
03MB002	27	0,36	0,23	0,38	0,24	02HE001	21	0,79	0,41	0,71	0,41
02DD008	26	1,63	0,34	0,79	0,26	02HM005	21	1,68	0,31	1,65	0,30
02GB007	26	-0,01	0,36	-0,38	0,32	02LC043	21	1,05	0,44	1,03	0,45
02GC018	26	1,00	0,47	1,31	0,48	02PL005	21	0,38	0,31	0,30	0,31
02LH002	26	0,36	0,30	0,23	0,23	03AD001	21	1,12	0,25	1,05	0,25
03AC002	26	0,58	0,35	0,73	0,38	05QC003	21	1,30	0,45	1,24	0,45
03BC002	26	0,48	0,17	0,81	0,18	05QE008	21	2,11	0,66	2,59	0,67
03EA001	26	0,40	0,27	0,57	0,27	05QE009	21	1,41	0,52	1,52	0,54
01BF001	25	0,41	0,34	0,45	0,33	01BD002	20	1,67	0,33	1,67	0,33
02EC009	25	1,25	0,48	1,03	0,44	01BH005	20	2,24	0,54	2,24	0,54
02ED007	25	0,98	0,34	0,69	0,35	02AC001	20	1,80	0,62	1,80	0,62
02FC004	25	1,97	0,36	1,79	0,39	02GG004	20	0,88	0,45	0,88	0,45
02GD020	25	1,45	0,47	1,24	0,42	02JB013	20	0,63	0,28	0,63	0,28
02HB012	25	1,45	0,48	1,27	0,47	02JC008	20	-0,01	0,26	-0,01	0,26
02HC032	25	0,54	0,43	0,38	0,40	02LD005	20	-0,13	0,22	-0,13	0,22
02HD002	25	1,13	0,60	1,45	0,56	02LH004	20	0,74	0,31	0,74	0,31
02HD009	25	2,62	0,77	2,51	0,75	02OG026	20	0,20	0,33	0,20	0,33
02HL005	25	0,47	0,26	0,16	0,25	02XC001	20	1,10	0,39	1,10	0,39
02PB006	25	0,40	0,31	0,42	0,31	03BA003	20	0,37	0,19	0,37	0,19
02WA001	25	0,04	0,24	-0,02	0,26	03BE001	20	0,73	0,28	0,73	0,28
03ED004	25	0,43	0,28	0,25	0,27	03CC001	20	0,12	0,18	0,12	0,18
03KC004	25	0,08	0,25	-0,18	0,24	03DC002	20	0,80	0,28	0,80	0,28
04FB001	25	0,82	0,40	1,16	0,44	03DD003	20	1,11	0,26	1,11	0,26
01BH001	24	1,02	0,34	1,04	0,35	03HA001	20	0,62	0,28	0,62	0,28
02CD001	24	0,09	0,32	0,04	0,31	04GA002	20	2,21	0,51	2,21	0,51

Tableau C2: Séries de données réelles de 40 observations et plus

Station	N	selon série entière		sur 20 dernières observations		sur 40 dernières observations	
		Cs	Cv	Cs	Cv	Cs	Cv
02FC001	76	0,39	0,36	0,58	0,35	0,38	0,36
02FC002	76	0,88	0,41	-0,08	0,32	0,18	0,37
02EA005	75	0,87	0,33	0,77	0,31	0,68	0,30
02EC002	75	0,19	0,24	0,10	0,20	-0,38	0,23
05PB014	74	0,74	0,45	0,43	0,41	0,33	0,39
04LJ001	71	0,47	0,30	1,07	0,34	0,99	0,28
02PJ007	62	0,08	0,31	0,11	0,23	0,00	0,32
02FB007	56	0,86	0,42	0,10	0,38	0,16	0,38
04NA001	55	0,98	0,24	-0,21	0,18	0,95	0,23
02OJ007	53	-0,48	0,21	-0,36	0,21	-0,58	0,22
02GA010	52	0,39	0,44	0,04	0,41	0,25	0,42
02LD001	46	1,21	0,28	0,42	0,25	1,47	0,28
02FF002	45	0,51	0,46	0,57	0,40	0,50	0,46
02GD010	44	1,10	0,53	1,33	0,57	1,17	0,55
02PL001	44	0,59	0,41	1,05	0,52	0,86	0,42
02BD003	43	1,37	0,37	1,39	0,43	1,39	0,38
02LB006	43	0,01	0,35	-0,40	0,28	0,08	0,36
02LB007	43	0,83	0,42	0,63	0,35	0,64	0,40
02LD002	43	0,70	0,37	0,94	0,33	0,72	0,37
02ED003	42	1,30	0,50	1,45	0,40	1,46	0,49
02GG002	42	0,75	0,43	0,88	0,45	0,78	0,44
04JC002	40	0,39	0,25	0,07	0,20	0,39	0,25

Pour avoir une idée de la forme de la distribution des observations issues de ces 7 séries utilisées pour simuler des débits, nous les avons tracées sur un papier de probabilité normal, en utilisant la formule de probabilité empirique (FPE) de Cunnane (cf. équation 8). Rappelons que la propriété du papier de probabilité normal est de linéariser la fonction de distribution des observations lorsque celles-ci proviennent d'une loi normale.

Tableau C3: Détail des observations des échantillons de référence (m³/s)

02FC001 N=76		02FC002 N=76		02EA005 N=75		02EC002 N=75		05PB014 N=74		04LJ001 N=71		02PJ007 N=62	
172	501	96,5	262	21,8	42,5	49,8	131	37,5	124	405	903	81,0	200
174	501	106	272	21,9	42,8	60,0	133	39,1	125	411	912	84,7	202
177	507	107	276	22,1	43,6	74,5	133	42,2	126	436	912	85,0	205
210	513	115	277	22,3	43,9	78,4	136	46,7	126	484	924	88,9	207
231	518	126	283	24,2	44,2	78,7	136	50,7	130	487	937	90,6	208
251	527	132	289	24,6	44,2	87,8	138	55,2	131	496	943	96,8	209
257	529	148	289	25,8	44,2	88,3	138	56,4	131	538	946	97,1	216
264	530	163	292	25,9	45,9	93,1	140	58,2	136	561	946	99,1	219
269	532	170	297	26,4	46,2	99,4	140	60,2	139	565	949	116	221
272	537	180	306	28,8	46,4	100	140	61,4	140	569	950	118	221
297	538	180	326	28,9	47,0	101	141	66,5	144	578	977	120	222
309	547	183	328	30,6	47,3	102	143	66,8	150	595	1000	120	226
337	547	190	334	30,9	47,6	102	144	66,8	153	600	1020	133	231
337	548	192	335	31,1	47,6	103	144	70,5	153	623	1030	134	237
351	561	192	345	32,3	50,1	105	144	71,3	154	660	1030	136	243
362	575	197	345	33,0	50,1	106	146	71,4	157	674	1050	142	243
362	586	210	349	33,1	50,2	106	146	73,1	158	699	1060	152	251
368	589	213	352	33,5	50,7	107	147	74,5	159	705	1070	152	256
371	589	215	367	34,3	52,4	107	147	78,2	160	711	1080	153	261
371	595	215	368	34,5	54,1	109	149	80,7	163	725	1110	157	271
372	611	218	368	35,1	55,5	110	151	81,3	163	728	1130	157	278
388	623	220	371	35,4	56,6	112	152	81,8	164	753	1170	161	294
391	629	220	371	35,4	57,5	113	153	83,5	165	762	1190	161	304
425	668	228	374	35,7	57,5	116	153	90,0	168	770	1190	165	317
425	671	234	377	36,0	57,7	117	154	92,3	179	779	1200	176	
442	680	240	385	36,2	59,5	118	157	93,2	183	779	1220	180	
447	680	242	411	37,4	59,5	121	158	93,2	186	799	1250	181	
459	694	246	412	37,9	59,8	121	159	95,7	190	804	1250	187	
459	725	247	413	38,2	64,0	121	163	99,1	194	807	1310	188	
462	733	248	425	38,5	64,5	122	166	100	201	824	1320	188	
467	742	249	428	40,0	64,8	122	167	101	203	867	1390	189	
470	750	253	490	40,5	64,8	123	170	104	229	869	1480	189	
476	818	253	515	40,8	65,4	123	173	104	234	870	1740	189	
479	818	257	527	40,8	70,3	127	184	112	245	872		192	
484	818	260	532	40,8	78,2	128	189	120	257	883		192	
484	844	261	578	41,6	82,0	129	199	120	303	887		195	
490	895	261	629	41,9	93,4	130	229	120		898		197	
490	1030	262	640	42,2		130		123		900		197	

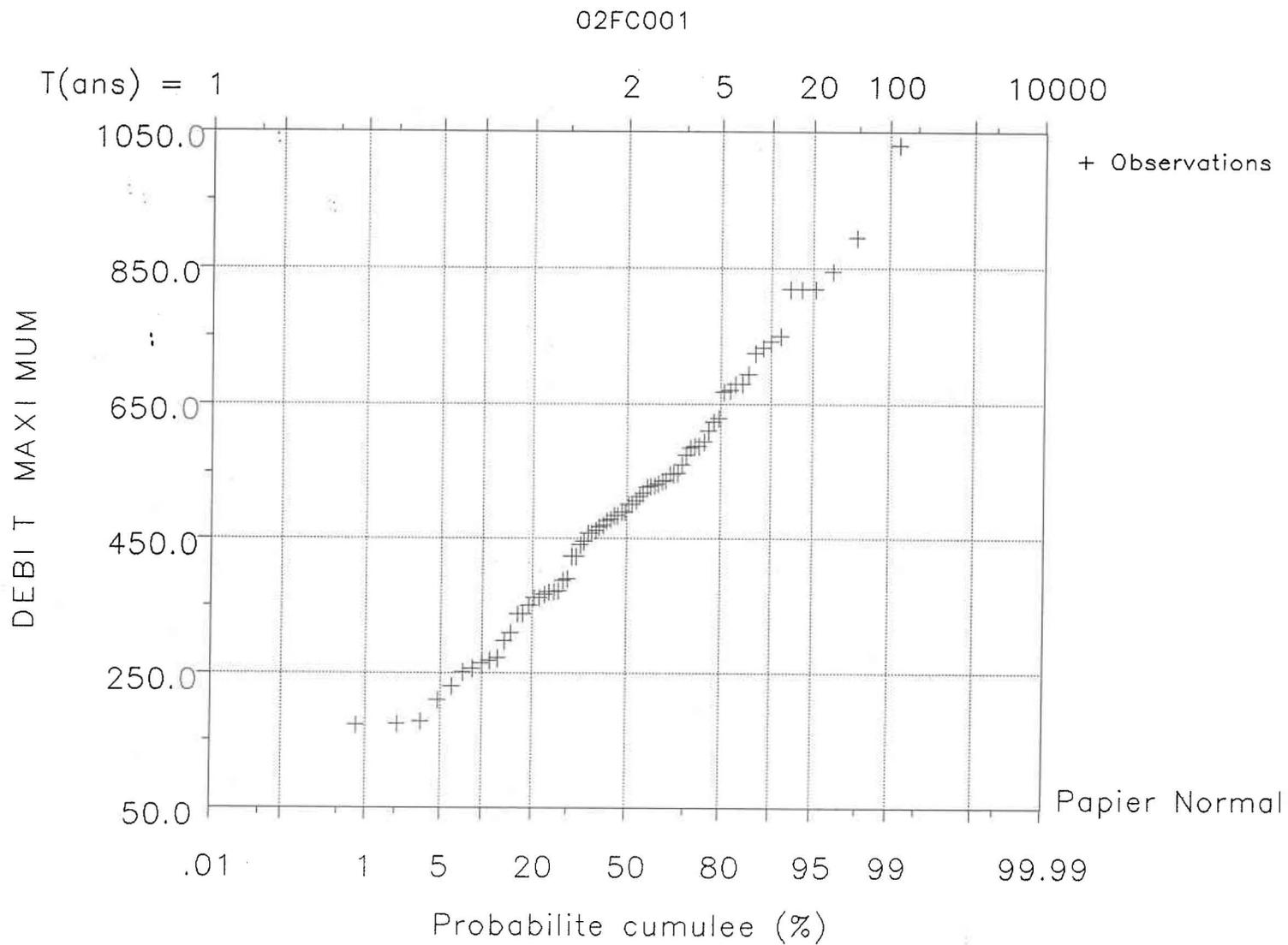


Figure C1: Observations mesurées à la station 02FC001 tracées sur papier de probabilité normal en utilisant la FPE de Cunnane

ANNEXE D: TYPES DE PARAMÈTRES

Les distributions statistiques traitées dans cette étude comportent toutes deux ou trois paramètres. On distingue généralement trois classes de paramètres: les paramètres de position, d'échelle et de forme. Nous définirons ces types de paramètres et étudierons l'effet d'une transformation logarithmique sur ceux-ci.

D.1 Définition des types de paramètres

Soit une variable aléatoire X provenant d'une distribution statistique D ayant p paramètres $(\theta_1, \theta_2, \dots, \theta_p)$. Par définition, le paramètre de position est celui qui n'influence que le moment d'ordre un, la moyenne de X . Le paramètre d'échelle est celui qui n'influence que les moments d'ordre un et deux, la moyenne et la variance. Les paramètres de forme sont tous ceux qui influencent des moments d'ordre supérieur à deux.

En pratique, toutes les distributions utilisées pour estimer x_T possèdent un paramètre d'échelle, mais peuvent ne pas posséder de paramètre de position ou de paramètres de forme. Toute variable aléatoire X provenant d'une telle distribution D ayant un paramètre d'échelle η peut s'écrire ainsi:

$$X = g(\eta) Y \quad (D.1)$$

où $g(\cdot)$ est une fonction quelconque et Y est une variable aléatoire provenant aussi de la distribution D et ayant les mêmes valeurs de paramètres que X , sauf pour η . Si Y a une variance unitaire, on la nomme alors variable de variance standardisée.

Plusieurs des distributions utilisées en pratique ont aussi un paramètre de position qui leur permet de s'adapter à un éventail de moyennes différentes. Toute variable aléatoire X provenant d'une telle distribution D ayant un paramètre de position ω peut s'écrire ainsi:

$$X = h(\omega) + W \quad (D.2)$$

où $h(\cdot)$ est une fonction quelconque et W est une variable aléatoire provenant aussi de la distribution D et ayant les mêmes valeurs de paramètres que X , sauf pour ω . Si W a une moyenne nulle, on la nomme alors variable de moyenne standardisée.

Si une distribution D possède à la fois un paramètre de position ω et un paramètre d'échelle η , alors une variable aléatoire X provenant de cette distribution peut s'écrire ainsi:

$$X = h(\omega) + s(\eta) Z \quad (\text{D.3})$$

où $h(\cdot)$ et $s(\cdot)$ sont des fonctions quelconques et Z est une variable aléatoire provenant de la distribution D et ayant les mêmes valeurs de paramètres que X , sauf pour ω et η . Si Z possède une moyenne nulle et une variance unitaire, on nomme Z variable standardisée.

D.2 Effet d'une transformation logarithmique

Il est facile de voir que la transformation logarithmique d'une variable aléatoire change le type de ses paramètres. Soit par exemple une variable X ayant des paramètres de position et d'échelle ω et η tels que $X = h(\omega) + s(\eta) Z$, où Z est la variable standardisée de moyenne nulle et de variance unitaire. Si l'on considère une variable $Y = \exp\{X\}$, on montre facilement qu'elle peut s'exprimer ainsi:

$$Y = \exp\{h(\omega)\} \exp\{s(\eta) Z\} = g(\omega) W \quad (\text{D.4})$$

où $g(\cdot)$ est une fonction quelconque et W une variable aléatoire provenant de la même distribution que Y et ayant les mêmes valeurs de paramètres que X , sauf pour ω . On remarque que le paramètre ω devient un paramètre d'échelle et qu'il n'est plus possible d'exprimer η comme un paramètre d'échelle ou de localisation. Lorsque l'on prend l'exponentielle d'une variable aléatoire, le paramètre de localisation devient donc un paramètre d'échelle, et le paramètre d'échelle un paramètre de forme.

ANNEXE E: DISTRIBUTIONS STATISTIQUES UTILISÉES

Cet annexe contient une courte description des différentes distributions utilisées pour estimer le débit X_T de période de retour T , ainsi que des références à des ouvrages traitant plus en détail de ces distributions. La fonction de densité de probabilité (fdp) cumulée inverse est donnée pour chaque distribution, ainsi que quelques détails sur les problèmes rencontrés lors de l'estimation des paramètres.

E.1 Distributions normale et log-normale

Les distributions normales et log-normales sont bien connues. Elles peuvent être déduites de la distribution normale standardisée. Une variable aléatoire ayant une distribution normale standardisée possède la fdp suivante:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \quad (\text{E.1})$$

La f.d.p. cumulée inverse de la distribution normale standardisée, que nous nommerons $Z(p)$, peut être estimée à l'aide d'une table de la fdp cumulée ou à l'aide d'une approximation numérique (voir par exemple [ABRAMOWITZ et STEGUN 1970]). Notons que le domaine de définition de la loi normale standardisée est l'ensemble des nombres réels.

E.1.1 Distribution normale (N)

La distribution normale possède deux paramètres: un paramètre de position, μ , et un paramètre d'échelle, σ (voir l'annexe D pour une brève discussion des types de paramètres d'une distribution). La f.d.p. cumulée inverse de la loi normale s'obtient ainsi:

$$F_N^{-1}(p; \mu, \sigma) = \mu + \sigma Z(p) \quad (\text{E.2})$$

Comme pour la distribution normale standardisée, le domaine de définition de la distribution normale est l'ensemble des nombres réels. Notons que la méthode des moments est équivalente à la méthode du maximum de vraisemblance pour la distribution normale [PERREAULT et BOBÉE 1991], et que dans la mesure où l'on possède plus d'une observation il est toujours possible d'estimer les paramètres.

E.1.2 Distribution log-normale à 2 paramètres (2LN)

La distribution log-normale à deux paramètres (2LN) peut être déduite de la distribution normale par transformation logarithmique. Si L est une variable aléatoire ayant une distribution log-normale à deux paramètres, alors le logarithme naturel de cette variable est distribué normalement. La distribution 2LN a donc elle aussi deux paramètres μ et σ . La fdp cumulée inverse est:

$$F_{2LN}^{-1}(p; \mu, \sigma) = \exp\left\{F_N^{-1}(p; \mu, \sigma)\right\} \quad (E.3)$$

La transformation logarithmique, comme nous l'avons expliqué à l'annexe D, change le paramètre μ en paramètre d'échelle et le paramètre σ en paramètre de forme. Le domaine de définition de la distribution 2LN devient quant à lui l'ensemble des nombres positifs non nuls.

On peut montrer que pour la distribution log-normale à deux paramètres la méthode du maximum de vraisemblance est toujours plus efficace que la méthode des moments [PERREAULT et BOBÉE 1994]. Conséquemment, nous ne considérerons que cette méthode dans notre étude. Dans la mesure où l'on possède un ensemble d'observations positives pour estimer les paramètres, il est toujours possible de trouver une solution avec la méthode du maximum de vraisemblance dans le cas de la loi 2LN.

E.2 Famille gamma

L'utilisation des distributions de la famille gamma en hydrologie a été longuement étudiée, voir par exemple [BOBÉE ET ASHKAR 1991]. Les distributions de cette famille sont toutes

déduites de la distribution gamma à un paramètre (G1). La densité de cette distribution dépend d'un seul paramètre de forme, $\lambda (> 0)$ et s'écrit :

$$f(x;\lambda) = \frac{x^{\lambda-1} \exp\{-x\}}{\Gamma(\lambda)} \quad (\text{E.4})$$

où $\Gamma(\cdot)$ est la fonction gamma, définie ainsi:

$$\Gamma(\lambda) = \int_0^{\infty} t^{\lambda-1} \exp\{-t\} dt \quad (\text{E.5})$$

Le domaine de cette distribution est l'ensemble des nombres réels positifs non nuls. Notons que lorsque λ tend vers l'infini la distribution G1 tend vers la distribution normale. La fdp cumulée inverse de la distribution G1, $F_{G1}^{-1}(p;\lambda)$, peut être obtenue de [HARTER 1969], ou à l'aide d'une approximation numérique, telle celle proposée par [PERREAULT *et al.* 1992].

E.2.1 Distribution gamma à 2 paramètres (G2)

La distribution gamma à deux paramètres (G2) est déduite de la distribution G1 par l'ajout d'un paramètre d'échelle, $\alpha (\neq 0)$. La fdp cumulée inverse de cette distribution s'écrit donc ainsi:

$$F_{G2}^{-1}(p;\lambda,\alpha) = \alpha F_{G1}^{-1}(p;\lambda) \quad (\text{E.6})$$

Selon le signe de α le domaine de définition de la distribution G2 sera l'ensemble des nombres réels positifs non nuls (si $\alpha > 0$) ou l'ensemble des nombres réels négatifs non nuls (si $\alpha < 0$). Puisque les débits maximums annuels sont toujours positifs, nous considérerons donc toujours que α est positif.

Il arrive que la méthode du maximum de vraisemblance ne puisse donner de résultat. Cela se produit principalement lorsque λ est trop grand; dans ce cas, la distribution normale, vers laquelle tend la distribution gamma lorsque λ est grand, devrait plutôt être utilisée.

E.2.2 Distribution Pearson type III (P3)

La distribution Pearson type III (P3) provient de l'ajout d'un paramètre de position, m , à la distribution gamma à deux paramètres (G2). Elle possède donc trois paramètres: un paramètre de forme λ (> 0), un paramètre d'échelle α ($\neq 0$), et un paramètre de position m . La fdp cumulée inverse de cette distribution s'écrit conséquemment ainsi:

$$F_{P3}^{-1}(p; \lambda, \alpha, m) = m + F_{G2}^{-1}(p; \lambda, \alpha) \quad (E.7)$$

Selon le signe de α la distribution P3 possède soit une borne inférieure égale à m (si $\alpha > 0$), soit une borne supérieure égale à m (si $\alpha < 0$). Notons que comme la distribution G2, la distribution P3 tend vers la distribution normale lorsque λ tend vers l'infini.

On montre que les estimateurs du maximum de vraisemblance sont biaisés lorsque λ est inférieur à un [BOBÉE et ASHKAR 1991]. De plus, [JOHNSON et KOTZ 1970] ont montré que les estimateurs sont instables pour $\lambda < 2,5$. En pratique, il est donc difficile d'obtenir une solution avec la méthode du maximum de vraisemblance lorsque λ est petit. Le paramètre λ étant un paramètre de forme, il est lié au coefficient d'asymétrie de la distribution P3. On montre en effet que $\lambda = 4/Cs^2$ [BOBÉE et ASHKAR 1991]. Conséquemment, Cs est supérieur à deux lorsque λ est inférieur à un. En accord avec [BOBÉE et ASHKAR 1991], nous avons choisi de ne pas utiliser la méthode du maximum de vraisemblance lorsque l'estimateur du coefficient d'asymétrie (voir annexe A) est supérieur à deux.

E.2.3 Distribution log-Pearson type III (LP3)

La distribution log-Pearson type III (LP3) est déduite de la distribution P3 par transformation logarithmique. Le logarithme Népérien d'une variable aléatoire ayant une distribution log-Pearson type III suit une distribution P3. Conséquemment, la fdp cumulée inverse s'obtient ainsi:

$$F_{LP3}^{-1}(p; \lambda, \alpha, m) = \exp\left\{F_{P3}^{-1}(p; \lambda, \alpha, m)\right\} \quad (E.8)$$

La transformation logarithmique modifie le type des paramètres et le domaine de définition de cette distribution. Le paramètre de position de la distribution P3, m , devient un paramètre d'échelle et le paramètre d'échelle, α , devient un paramètre de forme (voir annexe D). Le paramètre de forme λ de la distribution P3 reste évidemment un paramètre de forme. Ainsi, la distribution LP3 a 1 paramètre d'échelle et 2 paramètres de forme. Le domaine de définition dépend quant à lui du signe de α . Si α est positif, le domaine est borné inférieurement par $\exp\{m\}$, alors que si α est négatif, le domaine est borné inférieurement par zéro et supérieurement par $\exp\{m\}$. Notons que lorsque λ tend vers l'infini, la distribution LP3 tend vers la distribution log-normale à deux paramètres (LN2).

On peut montrer que la méthode du maximum de vraisemblance de la distribution LP3 est équivalente à la méthode du maximum de vraisemblance de la distribution P3 appliquée au logarithme des observations [BOBÉE et ASHKAR 1991]. Conséquemment, son application pose les mêmes problèmes que ceux rencontrés avec la distribution P3 lorsque le coefficient d'asymétrie du logarithme des observations est supérieur à deux.

Nous utiliserons pour la distribution LP3 deux variations de la méthode des moments quelque peu différentes de celle expliquée à l'annexe A, qui sont la méthode des moments non centrés et la méthode des moments sur le logarithme des observations (couramment nommée 'méthode indirecte des moments').

La méthode des moments non centrés, proposée par [BOBÉE 1975] et couramment nommée méthode BOB consiste à égaliser expressions théoriques des moments non centrés d'ordre 1 à 3 (μ'_1, μ'_2, μ'_3) à leurs estimateurs non biaisés (m'_1, m'_2, m'_3). Cette variante de la méthode des moments est équivalente à la méthode des moments classique, telle que présentée à l'annexe A, lorsque l'on effectue aucune correction pour éliminer le biais des estimateurs de l'écart-type et du coefficient d'asymétrie. La méthode BOB ne peut être appliquée lorsque $0 < \alpha < 3$ parce que dans ce cas le moment non centré d'ordre 3, m'_3 , de la distribution LP3 n'existe pas [BOBÉE et ASHKAR 1991].

Nous considérerons aussi pour la distribution LP3 la méthode des moments appliquée sur le logarithme des observations, car son utilisation systématique a été recommandée aux États-Unis [USWRC 1967]. Nous la noterons WRC. Cette méthode revient à appliquer la

méthode des moments de la distribution P3 aux observations transformées. Aucun problème d'estimation ne se pose pour cette méthode à condition que le coefficient d'asymétrie des observations transformées ne soit pas nul.

E.2.4 Distribution gamma généralisée à 3 paramètres (GG3)

La distribution gamma généralisée à trois paramètres (GG3) est obtenue par l'ajout d'un paramètre de forme, s , à la distribution gamma à deux paramètres (G2). Si G_2 est une variable aléatoire provenant d'une distribution G2, alors $G_3 = G_2^s$ suit une distribution GG3. Conséquemment, la fdp cumulée inverse de cette distribution s'écrit ainsi:

$$F_{GG3}^{-1}(p; \lambda, \alpha, s) = \begin{cases} \text{Si } s > 0: & F_{G2}^{-1}(p; \lambda, \alpha)^s \\ \text{Si } s < 0: & F_{G2}^{-1}(1 - p; \lambda, \alpha)^s \end{cases} \quad (\text{E.9})$$

La distribution GG3 comporte donc deux paramètres de forme, $\lambda (> 0)$ et $s (\neq 0)$, et un paramètre d'échelle, $\alpha (\neq 0)$. Le domaine de définition de la distribution GG3 reste le même que celui de la distribution G2, soit l'ensemble des nombres positifs non nuls. Notons que lorsque s tend vers zéro, la distribution GG3 tend vers la distribution LN2, et que lorsque $\lambda=1$, la distribution GG3 est équivalente à la distribution Weibull à deux paramètres (W2) [BOBÉE et ASHKAR 1991].

L'introduction du paramètre s rend la résolution des équations de la méthode du maximum de vraisemblance très difficile, même pour de grandes tailles d'échantillon [HAGER *et al.* 1971]. Conséquemment, il arrive souvent que cette méthode ne donne pas de solution. L'utilisation de la méthode des moments pose aussi un problème parce que le coefficient d'asymétrie de la distribution GG3 n'existe que si $3/s > -\lambda$ [BOBÉE et ASHKAR 1991]. De plus, des difficultés numériques nous empêchent parfois d'obtenir une solution lorsque s est petit, c'est à dire lorsque la distribution GG3 tend vers la distribution LN2.

E.2.5 Distribution Weibull à 2 paramètres (W2)

La distribution Weibull à deux paramètres (W2) est un cas particulier de la distribution GG3 lorsque $\lambda = 1$ et s est positif. Conséquemment, elle comporte un paramètre de forme et un paramètre d'échelle. Par convention, on écrit la densité de la distribution W2 ainsi:

$$f(x) = \frac{c}{\alpha} \left(\frac{x}{\alpha}\right)^{c-1} \exp\left\{-\left(\frac{x}{\alpha}\right)^c\right\} \quad (\text{E.10})$$

Si l'on compare cette paramétrisation avec celle utilisée pour la distribution GG3, on se rend compte que le paramètre de forme s de la GG3 est remplacé par le paramètre c , et que la paramètre d'échelle de la GG3, α , correspond à l'inverse du paramètre d'échelle, aussi nommé α , de la distribution Weibull. Le domaine de définition de cette distribution est, comme pour la distribution GG3, l'ensemble des nombres réels positifs non nuls.

Contrairement aux autres distributions de la famille gamma, il existe pour cette distribution une forme explicite de la fdp cumulée inverse [PERREAULT et BOBÉE 1992b]:

$$F_{W2}^{-1}(p; c, \alpha) = \alpha[-\log(1-p)]^{1/c} \quad (\text{E.11})$$

E.3 Distributions des valeurs extrêmes

Les distributions des valeurs extrêmes sont issues de l'étude de la valeur minimale et de la valeur maximale mesurées dans un échantillon de grande taille d'observations indépendantes et identiquement distribuées. La distribution de ces valeurs extrêmes peut prendre trois formes, nommées EV1, EV2 et EV3 par [FISHER et TIPPETT 1928]. Ces trois distributions ont été regroupées en une seule, plus générale [JENKINSON 1955], et a été nommée loi généralisée des valeurs extrêmes (*Generalized Extreme Value*, ou GEV).

E.3.1 Distribution Gumbel (EV1)

La distribution EV1 porte aussi le nom de distribution Gumbel, qui l'étudia en détail [GUMBEL 1960]. Elle possède un paramètre d'échelle, $\alpha (> 0)$ et un paramètre de position, u . Sa fonction de densité cumulée s'écrit ainsi: [PERREAULT et BOBÉE 1992a]

$$F(x) = \exp\left\{-\exp\left[-\frac{x-u}{\alpha}\right]\right\} \quad (\text{E.12})$$

Le domaine de définition de cette distribution est l'ensemble des nombres réels. Il existe pour cette distribution une formulation explicite de la fdp cumulée inverse:

$$F_{EV1}^{-1}(p; u, \alpha) = u - \alpha \log[-\log(p)] \quad (\text{E.13})$$

Seule l'application de la méthode du maximum de vraisemblance à la distribution EV1 nécessite la résolution numérique d'une équation. Il est de plus très rare qu'il soit impossible de trouver une solution numérique. Les estimateurs obtenus par cette méthode sont cependant biaisés. Nous utiliserons la correction proposée par [FIORENTINO et GABRIELE 1984] pour diminuer ce biais.

E.3.2 Distribution généralisée des valeurs extrêmes (GEV)

La distribution GEV regroupe les trois distributions des valeurs extrêmes. Elle possède un paramètre de forme de plus que la distribution EV1, que l'on note k . Dans le cas où $k=0$, sa densité cumulée est équivalente à celle de la distribution EV1. Lorsque par contre $k \neq 0$ la densité cumulée de la distribution GEV s'écrit ainsi: [PERREAULT et BOBÉE 1992a]

$$F(x) = \exp\left\{-\left[1 - \frac{k}{u}(x-u)\right]^{1/k}\right\} \quad (\text{E.14})$$

Conséquemment, la fdp cumulée inverse de cette distribution s'écrit ainsi:

$$F_{\text{GEV}}^{-1}(p; u, \alpha, k) = u + \frac{\alpha}{k} \left\{ 1 - [-\log(p)]^k \right\} \quad (\text{E.15})$$

Le domaine de la distribution GEV dépend du signe du paramètre de forme, k . Lorsque k est négatif, le domaine est borné inférieurement par $u + \alpha/k$. Par contre, lorsque k est positif, le domaine est borné supérieurement par cette même quantité, $u + \alpha/k$. Évidemment, lorsque k vaut exactement zéro, le domaine comprend l'ensemble des nombres réels, puisque la distribution GEV se confond alors avec la distribution EV1.

L'estimation des paramètres de la distribution GEV par toutes les méthodes d'estimation (MV, MM, PWM) nécessite la résolution numérique d'une équation qui ne dépend que du paramètre k . En pratique, il arrive qu'il ne soit pas possible de trouver une solution lorsque k est près de zéro.

BIBLIOGRAPHIE

- Abramowitz, W. et I. Stegun (1970). **Handbook of mathematical functions**, National Bureau of Standards, Washington D.C., États-Unis.
- Acreman, M.C. (1987). Regional flood frequency analysis in the United Kingdom: recent research, new ideas. **Institute of hydrology**, Wallingford, Royaume-Uni, 1987, 17 pages.
- Adamowski, K. (1985). Nonparametric Kernel Estimation of Flood Frequencies. **Water resources research**, 21(11):1585:1590.
- Akaike, H. (1974). A new look at the statistical model identification. **IEEE transactions on automatic control**, AC-19(6): 716-723.
- Beard, L.R. (1987). Discussion of "Relative accuracy of log-Pearson type III procedures" by J.R. Wallis et E.F. Wood. **Journal of hydraulic engineering**, 113: 1205-1206.
- Bernier, J. (1993). **Simulation, Bayes et Bootstrap en hydrologie statistique**, note interne, INRS-Eau.
- Bobée, B. (1975). The log-Pearson type 3 distribution and its application in hydrology. **Water resources research**, 11(5): 681-689.
- Bobée, B. et F. Ashkar (1991). **The gamma family and derived distributions applied in hydrology**, Water resources publications, Littleton, Colorado, États-Unis.
- Bobée, B., P. Rasmussen, L. Perreault et F. Ashkar (1993a). Risk analysis of hydrologic data: review and new developments concerning the Halphen distributions, **Engineering Risk in Natural Resources Management**, L. Duckstein et É. Parent éditeurs, Kluwer Academic Publishers, Pays-Bas, pp. 177-190.
- Bobée, B., G. Cavadias, F. Ashkar, J. Bernier et P. Rasmussen (1993b). Towards a systematic approach to comparing distributions used in flood frequency analysis. **Journal of hydrology**, 142: 121-136.
- Bobée, B. et R. Robitaille (1977). The use of the Pearson type 3 and log Pearson type 3 distributions revisited. **Water resources research**, 13(2): 427-443.

- Cavadias, G. (1990). The canonical correlation approach to regional flood estimation. **Regionalization in hydrology**, IAHS, publication 191, pp. 171-178.
- Chow, K.C.A. et W.E. Watt (1990). A knowledge-based expert system for flood frequency analysis, **Canadian journal of civil engineering**, 17: 597-609.
- Chow, V.T. (1953). Frequency analysis of hydrologic data with special application to rainfall intensities, **University of Illinois**, Eng. Exper. Stat. Bull.: 414.
- Chow, V.T. (1954). The log-probability law and its engineering applications. **Proceedings of ASCE**, 80(5): 1-25.
- C.N.R.C. (1990). **Hydrologie des crues au Canada**, Comité associé d'hydrologie du Conseil National de Recherche du Canada, Ottawa, Canada, J. Rousselle éd.
- Cunnane, C. (1978). Unbiased plotting position - a review. **Journal of hydrology**, 37: 205-222.
- Duckstein, L., B. Bobée et F. Ashkar (1991). A multiple criteria decision modelling approach to selection of estimation techniques for fitting extreme floods. **Stochastic Hydrology and Hydraulics**, 5:227-238
- Dumonceaux, R., C.E. Antle et G. Haas (1973). Likelihood ratio test for discrimination between two models with unknown location and scale parameter. **Technometrics**, 15(1): 19-27.
- Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. **Society for Industrial and Applied Mathematics**, 21(4).
- Environnement Canada (1990). **HYDAT version 3.0**, Direction des ressources en eau, données sur les eaux de surface.
- Fiorentino, M. et S. Gabriele (1984). A correction for the bias of maximum likelihood estimators of Gumbel parameters. **Journal of hydrology**, 73: 39-49.
- Fisher, R.A. et L.H.C. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. **Proceedings of the Cambridge philosophical society**, 24: 180-190.

- Fortin V. et B. Bobée (1993) Formalization of the Selection of Statistical Distributions Process in Flood Frequency Analysis. **Proceedings of the Eight International Conference on Artificial Intelligence in Engineering, Vol 2: Applications and Techniques**. Computational Mechanics Publications. pp. 387-402.
- Fortin V., M. Lavoie et D. Monaghan (1992). **Ajuste version 1.1 - Manuel technique**. Rapport interne I-109. INRS-Eau.
- Gingras, D., K. Adamowski et P.J. Pilon (1994). Regional flood equations for the provinces of Ontario and Quebec. **Water resources bulletin**, 30(1):55-67.
- Greenwood, J.A., J.M. Landwehr, N.C. Matalas et J.R. Wallis (1979). Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form, **Water resources research**, 15(5): 1049-1054.
- Gumbel, E.J. (1960). **Statistics of extremes**. Columbia University Press, New York, États-Unis.
- Haan, C.T. (1977). **Statistical methods in hydrology**, Iowa State University Press, Ames, Iowa, États-Unis, 378 pages.
- Hager, H.W., L.J. Bain et C.E. Antle (1971). Reliability estimation for the generalized gamma distribution and robustness of the Weibull model. **Technometrics**, 13(3): 547-557.
- Haktanir, T. (1992). Comparison of various flood frequency distributions using annual flood peaks data of rivers in Anatolia, **Journal of hydrology**, 136(1): 1-31.
- Haktanir, T. et H.B. Horlacher (1993). Evaluation of various distributions for flood frequency analysis. **Hydrological sciences - Journal - des sciences hydrologiques**, 38(1): 15-32.
- Harter, H.L. (1969). A new table of percentage points of the Pearson type III distribution. **Technometrics**. 11(1): 177-187.
- Horton, R.E. (1913). Frequency of recurrence of Hudson river floods. **U.S. Weather Bureau Bulletin**, 2: 109-112.

- Hosking J.R.M. et J.R. Wallis (1993). Some statistics useful in regional frequency analysis, **Water resources research**, 29(2):272-281.
- Houghton, J.C. (1978). Birth of a parent: the Wakeby distribution for modeling flood flows. **Water resources research**, 12(6): 1105-1110.
- I.E.A. (1977,1987) **Australian rainfall and runoff flood frequency analysis and design**, Institute of Engineers, Australie, 149 pages.
- Jenkinson, A.F. (1955). The frequency distribution of the annual maximum (or minimum) of meteorological elements. **Quarterly journal of the royal meteorological society**, 81: 158-171.
- Jenkinson, A.F. (1969). Statistics of extremes. **Notes techniques de l'Organisation Météorologique Mondiale**, 98: 183-228.
- Johnson, W.L. et S. Kotz (1970). **Distributions in statistics: continuous univariate distributions vol. 1**, Houghton-Mifflin, Boston, États-Unis.
- Kirby, W. et M. Moss (1987). Summary of flood frequency analysis in the United States. **Journal of hydrology**, 96: 5-14.
- Klemes, V. (1986). Dilettantism in hydrology: transition or destiny? **Water resources research**, 22(9).
- Kübler, H. (1979). On the fitting of the three-parameter distributions lognormal, gamma and Weibull. **Statistische Aefte**, vol. 20.
- Kuczera, G. (1982). Robust flood frequency models. **Water resources research**, 18(2): 315-324.
- Landwehr, J.M., N.C. Matalas et J.R. Wallis (1979). Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. **Water resources research**, 15(5): 1055-1064.
- Landwehr, J.M., G.D. Tasker et R.D. Jarrett (1987). Discussion de l'article "Relative accuracy of log-Pearson III procedures" par J.R. Wallis et E.F. Wood. **Journal of hydraulic engineering**, 113: 1206-1210.

- Law, A.M. et W.D. Kelton (1982). **Simulation modeling and analysis**, McGraw-Hill, New-York, États-Unis, 400 pages.
- Lettenmaier D.P. (1988). Evaluation and testing of flood frequency estimation methods, **Proceedings of the International Conference on Natural Hazards in European-Mediterranean Countries**, Perugia.
- Lu, L.-H. et J.R. Stedinger (1992). Variance of two- and three-parameter GEV/PWM quantile estimators: formulae, confidence intervals, and a comparison. **Journal of hydrology**, 138: 247-267.
- Mann, H.B. et D.R. Whitney (1947). On the test of whether one of two random variables is stochastically larger than the other. **Annals of mathematical statistics**, 18: 50-60.
- Matalas, N.C., J.R. Slack et J.R. Wallis (1975). Regional skew in search for a parent. **Water resources research**, 11(6): 815-826.
- Mathier, L. (1994). **Banques de données, stations à retirer de l'analyse**, Note interne, Projet CRSNG-Stratégique, INRS-Eau.
- Mathier, L., R. Roy, B. Bobée, H. Perron et V. Fortin (1993). **Estimation régionale des débits de crue: description des banques de données hydrométriques, météorologiques et physiographiques pour le Québec et l'Ontario**, Rapport Interne I-123, INRS-Eau.
- N.E.R.C. (1975). **Flood studies report**, Natural Environment Research Council, Londres, Royaume-Uni, 1100 pages.
- O.M.M. (1989). **Statistical distributions for flood frequency analysis**, Organisation Météorologique Mondiale, Hydrologie opérationnelle, publication 718, Genève, Suisse, 73 pages.
- Perreault, L. et B. Bobée (1994). **Propriétés statistiques de la distribution log-normale**, Rapport Scientifique, INRS-Eau, à paraître.
- Perreault, L. et B. Bobée (1992a). **Loi généralisée des valeurs extrêmes: propriétés mathématiques et statistiques, estimation des paramètres et des quantiles X_t de période de retour T**, Rapport Scientifique 350, INRS-Eau.

- Perreault, L. et B. Bobée (1992b). **Loi Weibull à deux paramètres: propriétés mathématiques et statistiques, estimation des paramètres et des quantiles X_t de période de retour T** , Rapport Scientifique 351, INRS-Eau.
- Perreault, L. et B. Bobée (1991). **Propriétés statistiques de la distribution normale**, Rapport Scientifique, INRS-Eau.
- Perreault, L., B. Bobée et V. Fortin (1992). **Approximation des quantiles de la loi Pearson type 3 standardisée par les polynômes de Tchebichef**, Rapport Scientifique 346, INRS-Eau, 36 pages.
- Pilgrim, D.H. (1986). Bridging the gap between flood frequency and design practice. **Water resources research**, 22(9): 165S-176S.
- Potter, K.W. et D.P. Lettenmaier (1990). A comparison of regional flood frequency estimation methods using a resampling method. **Water resources research**, 26(3): 415-424.
- Press, S.J. (1989). **Bayesian statistics: principles, models, and applications**. John Wiley & Sons, New York, États-Unis, 237 pp.
- S.É.B.J. (1987). **Le complexe hydroélectrique de La Grande Rivière: Réalisation de la première phase**, Société d'Énergie de la Baie James, Éditions de la Chenelière.
- Slack, J.R., J.R. Wallis J.R. et N.C. Matalas (1975). On the value of information to flood frequency analysis. **Water resources research**, 11(5): 629-647.
- Stedinger, J.R. et T.A. Cohn (1986). Flood frequency analysis with historical and paleoflood information, **Water resources research**, 22(5): 785-793.
- Takara, K. et J.R. Stedinger (1993). Recent japanese contributions to frequency analysis and quantile lower bound estimators.
- Thomas, W.O. (1985). A uniform technique for flood frequency analysis, **Journal of water resources planning and management**, 111(3): 321-337.
- U.S.W.R.C. (1967, 1976, 1981). **Guidelines for determining flood flow frequency**, United States Water Resources Council, Hydrology Committee, Washington D.C., États-Unis.

- Vogel, R.M. (1986). The probability plot correlation coefficient test for the normal, lognormal, and gumbel distributional hypotheses, **Water resources research**, 22(4): 587-590. Correction (1987), **Water resources research**, 23(10):2013.
- Wald A. et Wolfowitz J. (1943). An exact test for randomness in the non-parametric case based on serial correlation, **Annals of mathematical statistics**, 14: 378-388.
- Wallis, J.R. et E.F. Wood (1985). Relative accuracy of log-Pearson type III procedures. **Journal of hydraulic engineering**, 111(7): 1043-1056.
- Yevjevich, V. et J.T.B. Obeysekera (1984). Estimation of skewness of hydrologic variables. **Water resources research**. 20(7): 935-943.